João Celorico Moreira de Albuquerque

Master of Science

Deteção semi-automática de áreas verdes permanentes

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática

Orientador: Fernando Pedro Reino da Silva Birra, Professor Auxiliar,
NOVA University of Lisbon

Co-orientador: Carlos Augusto Isaac Piló Viegas Damásio, Professor Auxiliar, NOVA University of Lisbon

Júri

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2019

Deteção semi-automática de áreas verdes permanentes

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Agradecimentos

Primeiramente gostaria de agradecer aos meus orientadores por todos os conselhos, apoio e ajuda durante o desenvolvimento da dissertação. Gostaria também de agradecer o apoio do Centro de Informação Geo-Espacial do Exército, mais especificamente ao Tenente Coronel Paulo Pires por toda a ajuda disponibilizada.

Por fim gostaria de agradecer à minha família e aos meus amigos, pelo apoio incondicional que me deram durante este período.

Resumo

O Centro de Informação Geoespacial do Exército (CIGeoE) cartografa o território nacional através da interpretação manual de imagens aéreas ortorretificadas com posterior validação no terreno utilizando recursos humanos. Este processo revela-se demasiado moroso, o que pode afectar o desfasamento das cartas do exército, em relação à realidade observada.

Deste problema nasce uma premissa interessante, a de acelerar este processo usando técnicas de detecção remota associadas a métodos de classificação de aprendizagem automática, mantendo todas as normas necessárias para garantir que as cartas topográficas possuem o mesmo grau de detalhe e rigor. Para o âmbito deste projecto escolheu-se a detecção semi-automática de vegetação permanente.

A solução utilizada utiliza de dados de detecção remota conjuntamente com métodos de classificação automática da vegetação.

Os 3 métodos de classificação utilizados são *Random Forests*, *XGBoost e Support Vecto Machines*. Foram utilizadas as metodologias temporal estática e de séries temporais com *feature selection*.

Para um estudo inicial dentro de uma área limitada as metodologias utilizadas obtiveram resultados muito bons para 4 regiões diferentes de Portugal Continental e com vegetação distinta. No entanto esta metodologia de classificação aprensenta fraca aplicabilidade na cadeia de produção do CIGeoE.

No entanto aquando a generalização revelaram-se alguns problemas nomeadamente com o desequilíbrio entre classes e com a heterogeneidade espectral de cada classe para uma região diferente.

Concluiu-se que para este problema é necessário chegar a um compromisso entre desempenho e aplicabilidade na cadeira de produção.

Palavras-chave: Deteção remota Aprendizagem Automática SAR Sentinel Random Forest

Abstract

The Army Center of Geo-spacial Information (CIGeoE) maps the Portuguese territory through the manual interpretation of aerial orthorectified imagery, which will be posteriorly validated in the field using human resources. This process is too time consuming and it takes a lot of resources to complete which seriously increases the time lag between the cartographic products and the current our last years.

From this problem arises a interesting premise which is how can we accelerate this process and too which extent can it be done, using remote sensing techniques and classification based on machine learning classifiers, while meeting all the standards of rigorousness and detail. The scope of this thesis will focus on the semi-automatic detection and mapping of permanent vegetation.

The final solution uses remote detected data with machine learning classification to achieve the desired results.

The 3 classification methods used were Random Forests, XGBoost e Support Vecto Machines. These classifiers were evaluated with single image and timseries features.

Inittialy for a limited area, the algorithm showed really good results for 4 different regions in Portugal with distinctive vegetation. On the other hand it suffers greatly from a harder implementation in the CIGeoE pipeline.

When tests were made to measure the generalization of the algorithm, some problems were encountered, namely the highly number of samples of each class, and the spectral heterogeneity of each class for differente regions.

Our conclusion is that a compromise needs to be made in order to maximize the performance/pipeline implementation relation

Keywords: Remote Detection Machine Learning SAR Sentinel Random Forests

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nas últimas décadas tem-se verificado um progresso muito relevante no desenvolvimento de técnicas de detecção remota eficiente e precisa de cobertura terrestre.

Isto revela-se essencial para a monitorização global da vegetação, permitindo uma melhor aferição das características qualitativas e quantitativas da mesma, numa visão não só à escala global, como também a um escala extraordinariamente pormenorizada.

Aliado ao desenvolvimento nesta área, têm-se disponibilizado cada vez mais ferramentas para o estudo da cobertura terrestre, como repositórios de imagens de sensores multi-espectrais e outras plataformas de dados com informação relevante sobre a ocupação do solo, de livre acesso à população geral, servindo como motivação para o uso cada vez mais regular destas técnicas para uma monitorização mais pessoal da cobertura terrestre.

1.1 Motivação e Contexto

A cartografia é essencial não só para o mapeamento do território mas também para construção de um perfil demográfico, epidemiológico, socioeconómico e ambiental. Para o sucesso desta actividade a representação correta e precisa da cobertura terrestre é indispensável.

A cobertura terrestre é retratada pelas características físicas e químicas dos materiais que residem à superfície da terra. Diferentes coberturas terrestres incluem arvoredo denso, esparso, água, solo árido, solo húmido, etc.

Para os efeitos desta tese a cobertura terrestre que irá ser focada será a vegetação permanente. Este tipo de cobertura terrestre apresenta uma complexidade elevada pois é um tipo de cobertura terrestre com muita diversidade e é muito importante para a gestão de recursos naturais.

A utilização de processos de determinação de cobertura terrestre pode ser aplicada à cartografia nomeadamente ao mapeamento de vegetação permanente de modo a servir de ajuda na elaboração de cartas precisas e actuais.

O aumento do número de plataformas de dados com informação de diferentes satélites, com resoluções cada vez maiores e com informação multi-espectral assim como o desenvolvimento de técnicas de aprendizagem automática com reconhecido sucesso na classificação de cobertura terrestre servem como o ponto de partida essencial desta tese.

Será de interesse a implementação de múltiplas metodologias para a interpretação de imagens com classificação baseada em aprendizagem automática realizando posteriormente uma análise comparativa destas técnicas estudando e identificando o seu comportamento, pontos fortes e pontos fracos. Também será de interesse comparar o desempenho das metodologias desenvolvidas nesta dissertação com o desempenho dos métodos já empregados pelo CIGeoE.

1.2 Descrição

Uma das responsabilidades do Centro de Informação Geoespacial do Exército(CIGeoE) é a produção da Carta Militar de Portugal, Série M888, na escala 1:25 000, que é, de facto, a Carta Base de Portugal. Um dos temas que faz parte da informação Geoespacial da Carta Militar é a vegetação. Este tema é adquirido através de fotografia aérea que depois é interpretado e vetorizado manualmente por operadores de fotogrametria que classificam a informação geográfica pretendidas de acordo com as Normas de Aquisição, num processo designado por restituição. [1]

Devido ao volume de informação que tem que ser processado manualmente pelos operadores, e ao facto de esta informação ser validada no terreno face ao desfasamento temporal entre as fotografias aéreas e realização dos trabalhos de campo, este processo caracteriza-se como sendo bastante demorado e exigente.

Pretende-se então, com a elaboração deste projecto, o desenvolvimento de ferramentas que, recorrendo a algoritmos de aprendizagem automática, permitam a classificação automática da vegetação permanente, a partir de imagens multi-espectrais de Sentinel, de acordo com as Normas de Aquisição do CIGeoE. A inclusão deste processo na cadeia de produção da Carta Militar, visa permitir a redução drástica do tempo de aquisição dos dados, assim como um maior grau de actualização face aos dados de validação provenientes do trabalho de campo.

A vegetação permanente é definida de acordo com estas normas como a vegetação, seja ela natural ou cultivada, que mantém a sua área vegetacional durante a mudança de estações, sem a presença de factores exteriores que possam afectar consideravelmente a sua dimensão (incêndios, desflorestação, etc.). As suas características poderão ser influenciadas ligeiramente com o decorrer do tempo, mas no geral manter-se-ão relativamente

[1] Estas normas são confidencias, e consequentemente de acesso restrito, sendo-me disponibilizadas para efeitos auxiliares na construção de uma metodologia eficiente

consistentes com o decorrer do ano. Estas características tornam este tipo de vegetação relevante para a catalogação cartográfica.

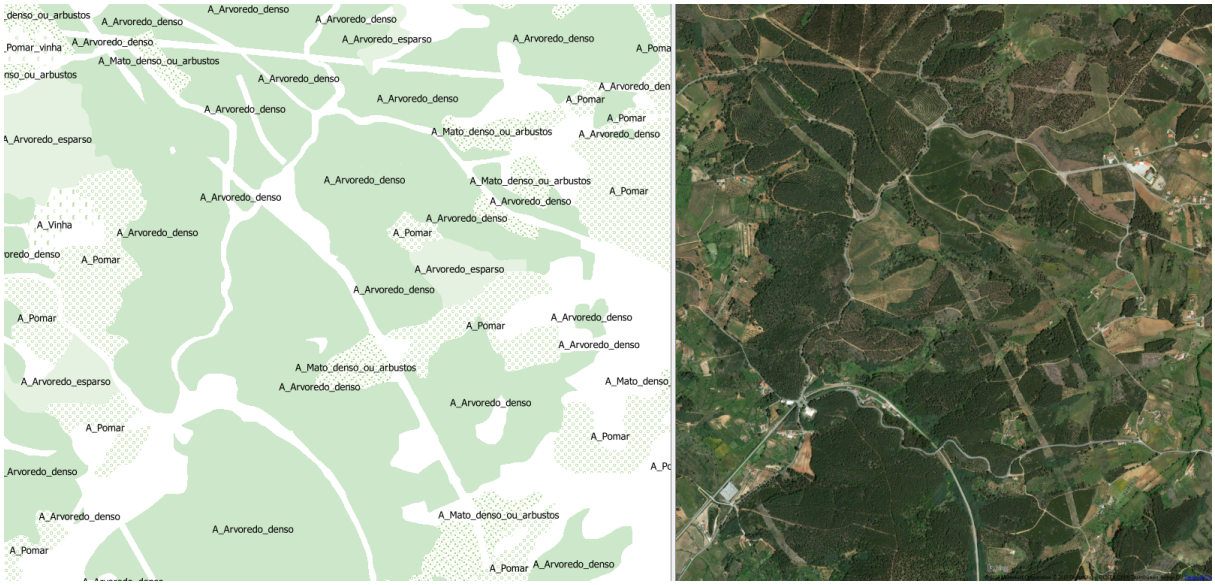

Figura 1.1: Exemplo da classificação da vegetação permanente feita pelo CIGeoE

A figura 1.1 pretende ilustrar o produto final do CIGeoE comparativamente às ortofotos utilizadas para a classificação da vegetação. A vegetação permanente, de acordo com as normas de aquisição do CIGeoE, está divididas em 10 classes principais ilustradas na figura 1.2 que representam os tipos de vegetação permanente mais incidente em Portugal, assim como os tipos de vegetação permanente com maior grau de interesse para o âmbito desta tese:

- **Vinha** - A vinha representa as áreas de cultivo predominante de videiras. Distingue-se por apresentar culturas ordenadas e de copas reduzidas. Por ser um tipo vegetação muito predominante em Portugal, sentiu-se a necessidade de fazer a distinção entre outras culturas aramadas e as vinhas. O limite mínimo de área para a classificação de uma vinha é de 2 ha.

- **Pomar** - As zonas de pomar são caracterizadas por áreas de cultivo metódico e abundância de árvores de fruto, sendo que estas árvores se encontram ordenadas e sistematicamente distribuídas. Estas são distinguíveis das vinhas devido ao seu tamanho superior de copa. Culturas aramadas que não sejam vinhas serão classificadas como pomar como por exemplo a cultura do kiwi. O limite mínimo de área para a classificação de um pomar é de 2 ha.

- **Pomar/Vinha** - Esta classe nasceu da observação de áreas em que as culturas viníferas estão englobadas em pomares ou vice-versa. O limite mínimo de área para a classificação de um pomar/vinha é de 2 ha.

- **Mato Denso ou Arbusto** - O mato denso ou arbusto é caracterizado pela sua cobertura predominante de vegetação arbustiva, com altura superior a 1 m, e de densidade elevada, com a possível existência de clareiras ou áreas de outro tipo de vegetação permanente (áreas de vegetação esparsa). O limite mínimo de área para a classificação de um mato denso é de 2 ha.

- **Mata** - A mata distingue-se do mato denso ou arbusto por apresentar uma cobertura vegetal muito mais diversificada, apresentado vegetação arbustiva mas também arvoredos extensos. Consequentemente pode apresentar uma copa mais densa e mais diversificada.O limite mínimo de área para a classificação da mata é de 2 ha.

- **Jardim ou Horta** - O jardim ou horta representa zonas de cultivo, com vegetação sistematicamente organizada e metodicamente distribuída. Esta vegetação classifica-se como de muito pequeno porte e de alta heterogeneidade.O limite mínimo de área para a classificação de um jardim ou horta é de 2 ha.

- **Arvoredo Esparso** - O arvoredo esparso é caracterizado por arvoredo com elevado grau de espaçamento e que por isso não cobre o solo devidamente, podendo até destacar-se copas singulares. Por norma esta classe é apenas considerada se existir um espaçamento limite entre copas das árvores equivalente a uma dezena de copas, valores consideravelmente superiores não serão adquiridos. Árvores de fruto que se encontrem dispersas e não alinhadas são consideradas como arvoredo esparso. O limite mínimo de área para a classificação de um arvoredo esparso é de 2 ha.

- **Arvoredo Denso** - O arvoredo denso é caracterizado por zonas arborizadas de elevada densidade e de natureza frondente, cobrindo totalmente o solo. Este é considerado quando as copas das árvores se tocam. Plantações de árvores frondosas de grande porte como Pinheiros ou Eucaliptos devem ser consideradas como arvoredo denso. O limite mínimo de área para a classificação de um arvoredo denso é de 2 ha.

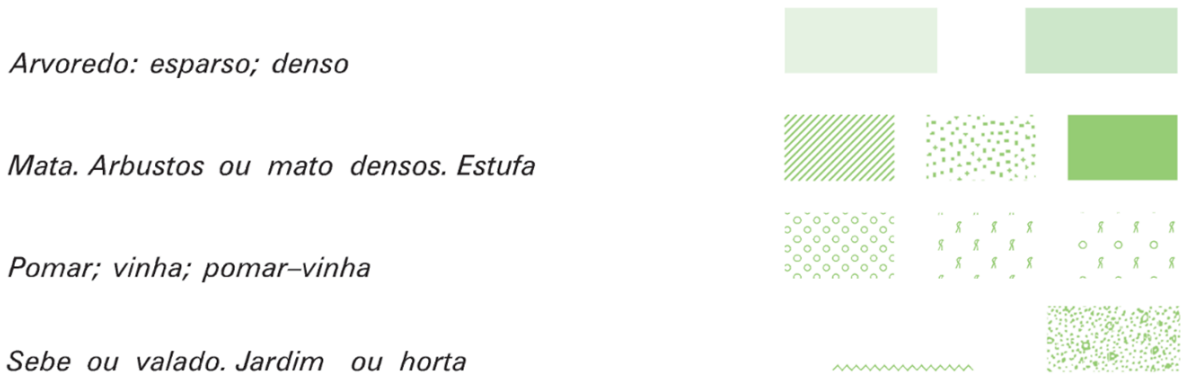

Figura 1.2: As 10 principais classes de vegetação e as suas representações num produto raster do CIGeoE

1.3 Principais Contribuições Previstas

Prevê-se que, no âmbito desta dissertação seja concebida uma metodologia com elevado sucesso na classificação da vegetação permanente, respeitando as normas de aquisição anteriormente referidas.

Como metodologias principais a serem utilizadas prevê-se que se teste uma metodologia temporal estática utilizando apenas uma imagem de cada um dos satélites a utilizar, tendo como classificadores principais, as *random forests*, o *xgboost* e as *support vector machines*. Posteriormente, prevê-se a utilização de diversas imagens de cada satélite numa metodologia de séries temporais, utilizando os mesmos classificadores.

Com base na metodologia concebida, consegue-se o desenvolvimento de uma ferramenta que consiga, precisa e eficazmente, classificar a vegetação permanente, a partir de dados multi-espectrais e de radar.

Por fim o ideal seria a integração total desta ferramenta no cadeia de produção do CIGeoE.

1.4 Organização do Documento

O presente documento está estruturado em 3 capítulos, organizado da seguinte maneira:

- **Capítulo 1 - Introdução** - Neste capítulo estão descritos os objectivos deste projecto assim como o contexto e as motivações por detrás do seu desenvolvimento.
- **Capítulo 2 - Estado da Arte** - No segundo capítulo está descrito trabalho relacionado que poderá ser relevante no contexto do problema. Este capítulo refere as técnicas de detecção remota mais usadas para a resolução de problemas semelhantes, e as metodologias de interpretação de imagens mais utilizadas assim como os métodos de classificação, através da utilização de algoritmos de aprendizagem automática, que apresentam o maior grau de sucesso em problemas semelhantes.
- **Capítulo 3 - Abordagem** - Neste capítulo será descrita extensivamente a abordagem que será utilizada para a construção da ferramenta de classificação de vegetação permanente. O trabalho preliminar executado, as metodologias de classificação utilizadas, a validação de resultados e o pós-processamento vectorial.
- **Capítulo 4 - Discussão de Resultados** - Neste capítulo serão apresentados os resultados da implementação da abordagem descrita no capítulo 3. Posteriormente estes resultados serão alvo de discussão com o objectivo do entendimento do desempenho das metodologias utilizadas.
- **Capítulo 5 - Conclusão e Trabalho Futuro** - Neste capítulo é apresentada a conclusão deste trabalho assim como trabalho futuro que poderá estender o âmbito desta dissertação.

Capítulo 2

Conceitos Básicos e Estado de Arte

Para a execução deste projecto foi necessário realizar-se uma investigação intensa e sistemática nas áreas em que o projecto se insere de modo a capturar as técnicas referidas na literatura com um maior grau de interesse para este ambiente de dissertação.

Neste capítulo pretende-se apresentar uma visão geral do conhecimento presente nestas áreas para efeitos explicativos e contextuais. Primeiramente dos conceitos básicos sobre detecção remota e das técnicas de literatura mais usadas e relevantes para esta dissertação e por fim uma visão geral de alguns conceitos básicos de aprendizagem automática como classificação supervisionada e não-supervisionada assim como os classificadores associados, e das abordagens mais utilizadas aquando a classificação de imagens obtidas remotamente. Serão apresentadas técnicas relevantes para o âmbito deste problema, assim como será feita uma avaliação comparativa entres as técnicas mencionadas. Deste modo pretende-se fornecer uma visão contextual do problema e das áreas envolventes que serão relevantes.

2.1 Detecção Remota

A detecção remota consiste em medições efectuadas sobre um objecto ou fenómeno sem a necessidade de contacto físico ou presença no local onde as medições serão efectuadas. Para contexto da tese, a detecção remota será considerada como a medição de propriedades de um objecto na superfície da terra, à distância, utilizando informação disponibilizada por satélites ou por aeronaves[40].

Devido à ausência de contacto físico é necessário recorrer-se à propagação de sinais tais como luz ou microondas.

Uma das principais características de imagem usadas para detecção remota é a região de comprimento de onda do espectro electromagnético (EEM). Algumas imagens

representam a radiação visível e infravermelha refletida do EEM, outras representam as medições de energia emitida pela superfície da Terra na região de comprimento de onda do infravermelho.

Existem dois métodos principais de detecção remota, detecção activa e passiva. Na detecção activa, a energia que serve como fonte para as medições é emitida por um emissor não natural, construído com o objectivo de posteriormente poderem ser medidas por um sensor. Geralmente, é medida a intensidade do retorno de ondas radar da superfície terrestre. Contrariamente se as medições dependerem de uma fonte externa de energia (radiação solar), estes processos são denominados de sistemas de detecção remota passiva.

Um sistema de detecção remota é um sistema multi-disciplinar que envolve disciplinas como óptica, espectroscopia, fotografia, telecomunicações etc. Este sistema tem por dever a detecção e descriminação de objectos ou características da superfície, detectando por sua vez energia reflectida ou emitida pelo objecto ou objectos do estudo em questão. Diferentes objectos devolvem diferentes quantidades de energia em diferentes bandas do EEM. O processo de detecção remota pode ser dividido nos seguintes passos:

1. Emissão de radiação electromagnética

2. Transmissão de energia do emissor para a superfície da terra com a subsequente absorção e dispersão

3. interacção da radiação electromagnética com a superfície da Terra: reflexão e emissão

4. Transmissão de energia proveniente da superfície terrestre para o sensor

5. Construção do output a partir da informação recebida.

2.1.1 Análise Espectral

A radiação do sol quando incide na superfície terrestre é reflectida, transmitida ou absorvida pela superfície terrestre. Devido a estes fenómenos o espectro electromagnético sofre variações na sua magnitude, direcção, comprimento de onda polarização e fase. Estas variações são subsequentemente detectadas pelo satélite e depois poderão ser interpretadas para a obtenção de informação útil sobre o objecto de estudo. Os dados adquiridos contêm informação não só espacial, como o tamanho forma e orientação, mas também informação espectral. Na figura 2.1 podem ser observados os diferentes meios de propagação de radiação.

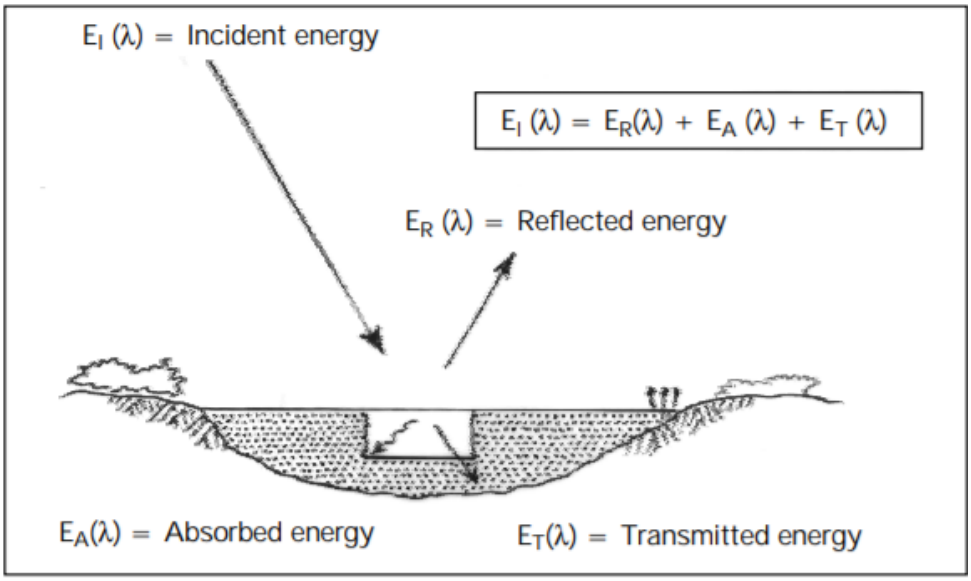

Figura 2.1: Propagação da radiação electromagnética retirada de [2]

Reflexão [2] ocorre quando um raio de luz é devolvido ao meio emissor após atingir um determinado meio de superfície. A sua intensidade depende do ângulo de incidência da radiação, índice de refracção e do coeficiente de absorção da superfície. Esta interacção é bastante útil no campo de detecção remota.

A reflexão é extremamente valiosa para o cálculo da assinatura espectral. A assinatura espectral é a intensidade relativa com que cada corpo reflecte ou emite radiação electromagnética nos seus diversos comprimentos de onda. Tecnicamente é a proporção de energia reflectida sobre energia incidente em função do comprimento de onda. É esta assinatura espectral que confere o tom e cor às fotografias. Cada objecto tem uma assinatura espectral particular, significando que cada objecto tem valores discretos para a sua refletância sobre valores diferentes e bem definidos de comprimentos de onda, possibilitando a sua distinção. É necessário que as características espectrais de cada objecto estejam bem estudadas de modo a possibilitar-se uma classificação coerente dos objectos de estudo.

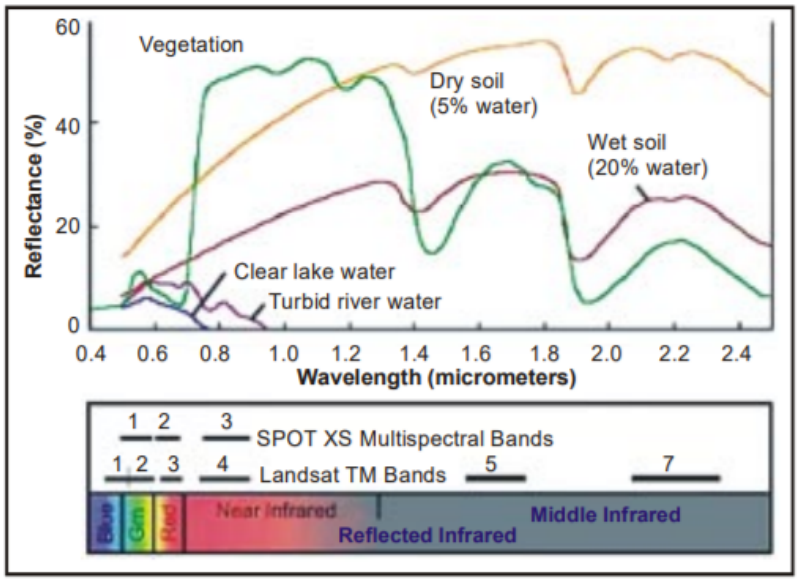

Figura 2.2: Curvas de refletância de diferentes coberturas terrestres

Na Figura 2.2 retirada de [2] verificam-se as curvas de refletância para diversos tipos de ocupação de solo.

Pode-se verificar que para o caso da vegetação esta absorve radiação com o comprimento de onda correspondente ao azul e vermelho do espectro de radiação visível mas reflecte radiação com o comprimento de onda correspondente ao verde do espectro de radiação visível. Na zona de comprimentos de onda *Near InfraRed* (NIR) a radiação é reflectida difusamente devido à estrutura interna de folhagem saudável. A refletância nestas regiões é usada como um indicador de vegetação saudável.

Na água, grande parte da radiação incidente é absorvida ou transmitida, sendo que o maior índice de absorção é verificado nas zonas do vermelho e NIR. Consequentemente a água tende a parecer mais azul ou verde devido a maiores níveis de refletância nestas radiações de menor comprimento de onda. A refletância dos corpos de água é afectada principalmente pela profundidade e por minerais contidos na água.

No solo, a maior parte da radiação é reflectida ou absorvida sendo muito pouca transmitida. A refletância do solo é muito dependente da sua quantidade de humidade, matéria orgânica, textura e estrutura. No geral, os solos apresentam uma distribuição mais uniforme na refletância da radiação de diferentes comprimentos de onda.

Na análise multi-espectral é ainda possível o cálculo de transformações matemáticas propostas de modo a avaliarem o impacto da vegetação em observações multi-espectrais. Estas transformações são designadas de índices de vegetação e realçam as diferenças espectrais da vegetação com base na sua forte absorção de bandas no vermelho e a sua forte refletância de bandas NIR [9, 22]. Vários índices de vegetação foram desenvolvidos como: o índice de percentagem de vegetação (RVI); o índice de diferença de vegetação normalizado (NDVI); o índice de vegetação de clorofila Verde (CGVI); e o índice de água de superfície terrestre (LSWI). As equações destes índices são:

$$RVI = \frac{(NIR)}{RED} \tag{2.1}$$

$$NDVI = \frac{(NIR - RED)}{NIR + RED} \tag{2.2}$$

$$CGVI = \frac{NIR}{GREEN - 1} \tag{2.3}$$

$$LSWI = \frac{(NIR - SWIR1)}{(NIR + SWIR1)} \tag{2.4}$$

A análise da assinatura espectral é fulcral para o ramo de detecção remota pois permite a distinção coerente das várias características da superfície terrestre.

2.1.2 Análise RADAR

Apesar do seu frequente uso, a análise espectral das imagens tem algumas limitações nomeadamente a falta de capacidade de penetração das ondas das radiações usadas, o que dificulta certas medições quando estas são realizadas em sítios com uma copa de árvores muito densa. As imagens são dependentes de condições meteorológicas mais especificamente da percentagem de nuvens que esconde o solo, sendo este um problema muito comum na análise espectral das imagens em locais de humidade muito elevada o que dificulta a obtenção de fotografias com uma percentagem elevada de píxeis sem nuvens. A dependência destas imagens da radiação fornecida pelo sol, também influencia o número de imagens disponíveis pois durante a noite não há radiação suficiente reflectida para se poder capturar as imagens. As interpretações também estarão sempre dependentes da resolução geométrica das imagens pois com resoluções baixas dificulta-se a distinção de objectos na superfície terrestre.

Devido a estes problemas, desenvolveu-se outro método de detecção remota que envolve sensores que utilizam Radar para a detecção da superfície terrestre e das suas características.

Estes sensores Radar baseiam-se no princípio Radar[1, 6, 25]. O radar é um aparelho de medição de frequências de micro-ondas ou ondas de rádio. Um transmissor (TX) irradia um pacote de ondas de frequência curta (fO) e duração τp

$$g(t)cos(2\pi f 0t) \tag{2.5}$$

onde g(t) é o invólucro do pulso e $cos(2\pi f 0t)$ é o veículo das ondas. Posteriormente a onda atinge um objecto e parte da sua energia é disseminada de volta e o sensor irá registar o eco. Para um único ponto de disseminação que esteja a uma distância R de um radar, o eco é uma réplica atenuada da onda transmitida, atrasada pelo tempo de ida e volta $2\frac{R}{c}$.

Este é o princípio radar pelo qual o *Synthetic-Aperture Radar* (SAR) se baseia. Através do processamento de vários ecos registados, o SAR é capaz de gerar imagens de alta resolução relativas à textura da superfície. Quanto maior for a antena do dispositivo SAR maior será a resolução da imagem. O SAR utiliza a distância que o seu dispositivo percorre, num tempo de ida e volta de um pulso à antena, criando assim uma antena sintética *virtual* de comprimento muito superior à física resultando em imagens com resoluções espaciais elevadas, mas mantendo o tamanho da antena física relativamente pequeno. A figura 2.3 ilustra este procedimento.

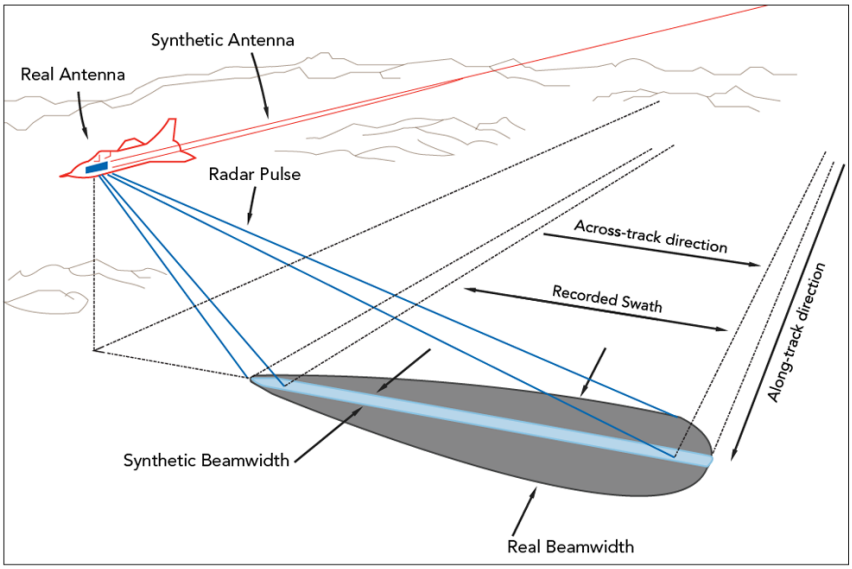

Figura 2.3: Procedimento SAR

Todas estas características permitem ao SAR ter propriedades que lhe concede diversas vantagens sobre sensores ópticos:

- Como é um método de detecção remota activo este é independente do sol.
- As micro-ondas utilizadas pelo SAR permitem a penetração de nuvens e copa de vegetação.
- Como usa ondas rádio polarizadas pode explorar a polarização para obter mais informação sobre a estrutura dos objectos em estudo
- O SAR é um método coerente de obtenção de imagens. Isto indica que o SAR é insensível a ruído.

2.2 Sentinel

O Sentinel é um programa desenvolvido pela Agência Espacial Europeia (ESA) que pretende substituir antigas missões de satélite de observação da Terra que estão a chegar

ao fim do seu ciclo operacional. Assim, é possível um fornecimento contínuo de dados espectrais sobre a superfície e atmosfera terrestre.

As missões de Sentinel existentes são Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-4, Sentinel-5 e Sentinel-5P. No âmbito desta dissertação iremos focar-nos nas missões Sentinel-1 e Sentinel-2.

2.2.1 Sentinel-1

A missão Sentinel-1 é constituída por um grupo de dois satélites com órbitas polares, realizando imagiologia SAR [41]. Estas imagens podem ser adquiridas em quatro modos diferentes, nomeadamente Stripmap (SM), Interfereometric Wide Swath (IW), Extra Wide Swath (EW) e Wave (WV). No âmbito desta tese, iremos focar-nos no modo IW, pois este modo apresenta uma resolução mais semelhante à desejada que será uma resolução 10 metros por 10 metros.

O Sentinel-1 também oferece diferentes tipos de polarização. A polarização caracteriza-se pelo plano ou planos em que este transmite e recebe a onda radar longitudinal. Os dois planos mais usuais na utilização são o vertical e o horizontal. Estes são:

- **HH** - transmissão horizontal e receção horizontal.
- **VV** - transmissão vertical e receção vertical.
- **HV** - transmissão horizontal e receção vertical.
- **VH** - transmissão vertical e receção horizontal.

As imagens de Sentinel-1 também podem ser adquiridas em vários níveis de processamento, desde o nível 0, que não está disponível ao público, até ao nível 2. Os produtos de nível 0 são produtos que consistem em dados SAR não tratados. Estes produtos constituem a base a partir da qual todos os produtos de nível 1 e 2 são produzidos. Estes dados são por isso de pouca relevância para os utilizadores e precisam de processamento significativo para servirem de algum uso aos utilizadores.

Os produtos de nível 0 são posteriormente transformados em produtos de nível 1 por uma instalação especializada no processamento destes produtos (IPF). Os dados de nível 1 podem ser convertidos em produtos *Single Look Complex* (SLC) ou *Ground Range Detected* (GRD). Os produtos SLC consistem em dados tratados de SAR, georreferenciados usando informação sobre a órbita e altitude proveniente do satélite, fornecendo informação sobre o alcance da inclinação do satélite. O alcance da inclinação do satélite é a coordenada que representa a linha de visão do satélite para cada objecto refletor. Os produtos GRD diferem dos SLC na medida em que foram projetados no plano do solo usando o modelo elipsoide da Terra WGS84. Esta projeção permite uma aferição mais realista dos valores contidos em cada píxel, obtendo, no entanto, uma resolução inferior.

Os produtos de nível 2 são produtos que adicionam aos dados tratados dos produtos de nível 1 componentes geofísicos como o campo de vento de oceano, espectros de ondas

do oceano e velocidade radial da superfície. Estes produtos permitem a observação de ventos alísios correntes e ondas e por isso são utilizados para a detecção remota de zonas oceânicas. Estes produtos são denominados *Ocean* (OCN).

Estas imagens são disponibilizadas no formato SENTINEL-SAFE, um formato comum no armazenamento e transmissão de informação por parte de entidades pertencentes à ESA Earth Observation. Este formato permite uma agregação de dados de imagem em formato binário e meta-dados em formato XML. O formato SENTINEL-SAFE pode ser decomposto nos seguintes elementos:

- Um ficheiro manifest.safe que contém a informação geral do produto em XML.
- Uma conjunto de pastas que inclui os datasets de medições sobre os dados da imagem em diferentes formatos binários .
- Um conjunto de pastas contendo *quicklooks* no formato PNG, GoogleEarth, KML e HTML.
- Uma pasta de anotações contendo os meta-dados do produto em XML e os respetivos dados de calibração.
- Uma pasta de suporte que contém os esquemas XML que descrevem o XML do produto.

No âmbito desta tese serão focados os produtos do nível um tanto SLC como GRD.

2.2.2 Sentinel-2

A missão Sentinel-2 é constituída por um grupo de dois satélites com órbitas polares que realizam imagiologia multi-espectral de média resolução. O facto de se usarem dois satélites permite a esta missão ter um período de revisita de 5 dias no equador. Esta missão fornece imagens com 13 bandas espetrais com diferentes resoluções espaciais [42]:

Tabela 2.1: As bandas do Sentinel-2 e as suas características

Banda	Resolução	Comprimento de Onda (nm)	Descrição
B1	60 m	443 nm	Aerosol
B2	10 m	490 nm	Azul
B3	10 m	560 nm	Verde
B4	10 m	665 nm	Vermelho
B5	20 m	705 nm	VNIR
B6	20 m	740 nm	VNIR
B7	20 m	783 nm	VNIR
B8	10 m	842 nm	VNIR
B8a	20 m	865 nm	VNIR
B9	60 m	940 nm	SWIR
B10	60 m	1375 nm	SWIR
B11	20 m	1610 nm	SWIR
B12	20 m	2190 nm	SWIR

As imagens de Sentinel-2 também se encontram em formato SAFE. Uma diretoria SAFE de um produto Sentinel-2 inclui:

- Um ficheiro manifest.safe que contém a informação geral do produto em XML.
- Uma visualização da imagem no formato JPEG2000.
- Um conjunto de pastas que inclui os datasets de medições que afetam os dados da imagem no formato JPEG2000.
- Um conjunto de pastas com informação sobre diversas datastrips.
- Um conjunto de pastas com informação adicional.
- Visualizações HTML

No âmbito desta dissertação o conjunto de sensores da missão espacial Sentinel revela-se o mais apropriado devido à sua elevada disponibilidade (gratuito e com APIs que poderão ser usadas para automatização do descarregamento das imagens) com imagens disponíveis desde Abril de 2014 para Sentinel-1 e Junho de 2015 no caso de Sentinel-2, o que neste caso, é mais que suficiente pois as regiões a classificar têm a sua verdade do terreno classificada em 2015. Poderiam ser usadas outras missões Sentinel como Sentinel-4 e Sentinel-5 no entanto estas missões são missões muito recentes e por isso não encontram imagens disponíveis no ano de 2015. Este sensor apresenta também uma resolução espacial de 10m x 10m para sensores de análise espectral o que ultrapassa as resoluções de outros sensores de satélite como o Landsat. Associadas a este conjunto de satélites estão diversas de ferramentas disponíveis para o processamento das imagens como o SNAP e o sen2cor, que permitem não só uma melhoria dos dados remotamente obtidos como um possível enriquecimento da informação disponível através de dados adicionais como a camada SLC do Sentinel-2 ou camadas de Texturas calculadas a partir de Sentinel-1.

2.3 Aprendizagem Automática

Aprendizagem Automática (*Machine Learning*) é caracterizada pela construção de sistemas que se aperfeiçoam com os dados. Deste modo, o processo de aprendizagem torna-se um passo muito importante para sistemas que têm uma base de aprendizagem automática, pois é deste modo que os sistemas conseguem melhorar o seu desempenho na sua tarefa designada. Para a aprendizagem supervisionada se realizar é necessária a existência de dados pré-classificados que o sistema possa utilizar para *"ensinar"* aos seus classificadores previamente à classificação de dados não classificados. No âmbito desta dissertação já existem dados previamente classificados por operadores do CIGeoE que podem ser usados para o processo de aprendizagem do algoritmo. Os classificadores escolhidos e que serão estudados nesta secção serão as *Random Forests*, as *Support Vector Machines* e o *XGBoost*, pois são algoritmos muito referidos na literatura e apresentam resultados muito razoáveis.

2.3.1 Técnicas de Classificação Não Supervisionada

A classificação não supervisionada distingue-se da supervisionada pois não requer que o utilizador tenha conhecimento prévio das classes em que os píxeis se irão dividir. É usada muito em análise de agrupamento de dados ou clustering. Estes métodos não só classificam e etiquetam os píxeis como determinam a localização das diversas classes e o seu número. As classes são posteriormente identificadas pelo utilizador após a comparação com um conjunto de referência. Estes métodos são desejáveis para a determinação da composição espectral das imagens antes da sua análise detalhada efectuada por métodos supervisionados.

Clustering implica o agrupamento de píxeis num espaço multi-espectral. Isto implica que múltiplos píxeis agrupados no mesmo cluster têm características espectrais semelhantes. Para atingir este efeito é necessário que exista uma medida de semelhança, de modo a possibilitar o agrupamento dos píxeis. As medidas mais usadas nos algoritmos de clustering são simples medidas de distância. As mais frequentes são a distância Euclidiana e a distância inter-ponto.

No entanto esta medida não é suficiente para determinar que todos os píxeis foram agrupados no cluster correto, pois podem existir vários clusters disponíveis que satisfaçam as condições impostas pelo algoritmo. Para a avaliação deste parâmetro é necessário ter-se uma medida de qualidade do clustering. A mais usada é a soma dos erros quadrados, esta medida calcula a distância acumulativa de cada píxel ao centro do seu cluster atribuído e depois soma todos os resultados de cada cluster. Esta medida representa a distância de cada píxel ao seu cluster logo só quando tiver um valor reduzido é que o clustering é aceite.

2.3.2 Técnicas de Classificação Supervisionada

A classificação supervisionada é o procedimento mais usada para uma análise quantitativa dos dados provenientes da detecção remota [38]. Estas técnicas de classificação têm como base a utilização de um algoritmo adequado de modo a classificar os píxeis ou segmentos de imagem como uma superfície específica terrestre. Apesar de existir uma miríade de algoritmos adequados para a classificação supervisionada, os passos que estão presentes numa abordagem de classificação supervisionada costumam incluir:

1. Decidir o conjunto de classes de superfícies terrestres em quais a imagem irá ser segmentada. Este conjunto representa as classes de informação como água, regiões urbanas, terras de cultura etc.

2. Escolher o conjunto de dados de treino. Usualmente em problemas de detecção remota são usados dados provenientes de satélites e que estão guardados em bases de dados acedidas através de sites.

3. Utilizar o conjunto de treino para estimar os parâmetros do algoritmo classificador a utilizar. Estes parâmetros permitem uma adaptação do algoritmo às características dos dados em questão.

4. Usando o classificador treinado, classificar todos os píxeis/objectos da imagem como um dos tipos de cobertura terrestre anteriormente definidos. Aqui toda a imagem é seleccionada, contrariamente à selecção do conjunto de treino que, para efeitos de prevenção de *overfitting*, apenas é seleccionada uma percentagem do conjunto total de dados.

5. Produzir resultados tabulares que poderão sumarizar o sucesso da classificação.

6. Analisar os diversos parâmetros de desempenho usados para a avaliação do classificador.

De acordo com os resultados obtidos, poderá ser necessário fazer alterações ao conjunto de treino, tais como refinar este conjunto de acordo com o resultado obtido dos parâmetros de desempenho.

No contexto desta tese irão ser analisadas as três abordagens mais usadas para a classificação de dados remotamente detectados estas serão as técnicas de *Gradient Boosting*, *Random Forest*, e *Support Vector Machines*.

2.3.3 Random Forest

A classificação baseada em *Random Forest* (RF) tem como princípios as *Decision Trees* (DTs) [7, 8]. As DTs são árvores binárias que recebem uma única variável de input e se divide depois em dois ramos possíveis com base na análise do input recebido (Assumindo o input com numérico). As folhas da DT representam os outputs possíveis. No contexto

da tese os inputs recebidos pela DT serão as características espectrais do píxel/objecto em estudo e os outputs serão as classes (diferentes coberturas de terreno) às quais o píxel irá ser associado.

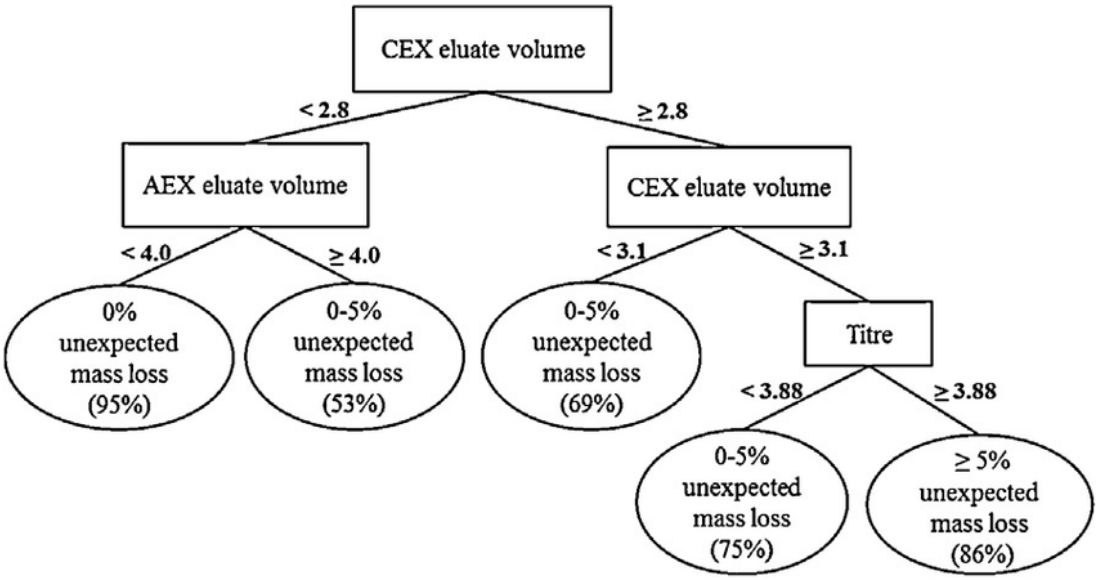

Figura 2.4: Exemplo de uma árvore de decisão (DT).

Na figura 2.4 é possível observar-se uma simples *decision tree*. Com base no input recebido esta árvore irá tentar prever qual a percentagem de massa gorda perdida.

Uma DT pode ser obtida através de sucessivas partições binárias de um conjunto inicial de características. Este processo é repetido para cada um dos sub-conjuntos, até se acreditar que particionar os sub-conjuntos deixa de adicionar valor às variáveis.

O algoritmo Random Forest é um método de combinação de aprendizagem automática para a classificação e regressão, que opera construindo múltiplas DT durante o tempo de treino e devolvendo como output a moda do output das classes de cada DT individual.

Para o sucesso deste classificador é no entanto necessário introduzir-se um certo grau de aleatoriedade.

O algoritmo Random Forest aplica a técnica geral de agregação de bootstrap ou bagging às DTs. Para cada classificador de DT é selecionada uma amostra aleatória e independente das características espetrais em estudo e o classificador é treinado nestas amostras. Após o treino de todos as DTs previsões posteriores são calculadas através de uma média de todas as previsões de cada classificador individual. Desta maneira é possível decrementar-se a variância deste modelo sem se incrementar o bias. Isto significa que recorrendo à lei dos grandes números (a média aritmética dos resultados duma mesma experiência repetida sucessivas vezes tende a aproximar-se do resultado esperado com o aumento de vezes sucessivas que a experiência é realizada) é possível inferir-se que apesar de uma DT ser sensível a valores aberrantes a média das DTs não é desde que a amostragem seja feita de forma independente. A classificação baseada em Random Forest introduz um método que a distingue de um método tradicional de bagging. Para cada

candidato durante o processo de treino é selecionado apenas um subconjunto do conjunto total de características. Isto evita que se criem correlações entre DTs devido à existência de características mais valorizadas na previsão de outputs e consequentemente mais escolhidas nas DTs

O classificador Random Forest é um classificador muito estudado em contexto de classificação de dados de detecção remota não só de vegetação mas como diversos tipos de cobertura terrestre[7]. Este algoritmo é menos sensível à qualidade das amostras de treino e overfitting, que outros algoritmos, pois é um algoritmo que utiliza um conjunto de classificadores individuais (DTs) que são treinados apenas com um subconjunto de treino do conjunto total de treino. É também um conjunto que não é afetado pela alta dimensão dos dados, problema muito comum quando se trata de dados espetrais.

No entanto o desempenho deste algoritmo é altamente dependente do número de árvores usado assim como do tamanho do subconjunto dado para o treino de cada árvore, podendo-se tornar factores com influência negativa aquando a da utilização deste algoritmo.

2.3.4 Support Vector Machines

De um modo muito resumido o objetivo das SVM é encontrar um hiperplano que que, num espaço N-dimensional, consiga classificar os dados, distinguindo-os uns dos outros [16].

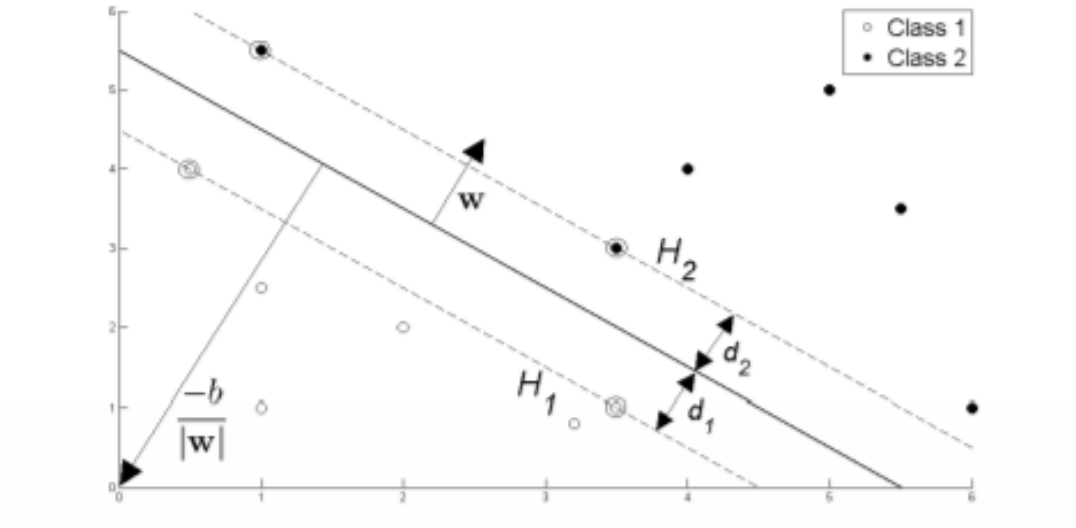

Figura 2.5: Exemplo de um hiperplano a dividir um conjunto de dados.

Na figura 2.5 podemos observar que existe uma recta no referencial que divide o conjunto de pontos em dois conjuntos distintos. Isto acontece porque este conjunto de dados é linearmente separável por um hiperplano. Este hiperplano pode ser descrito

como w * x + b = 0 onde w é a normal para o hiperplano e $\frac{b}{\|W\|}$ é a distância perpendicular do hiperplano à origem.

Os Vetores de Suporte são os pontos mais próximos do hiperplano e o seu papel é orientar este hiperplano de modo a que este esteja o mais longe possível de ambas as classes. Se considerarmos agora apenas estes pontos que estão mais próximos do hiperplano que os separa, os vetores de suporte (representados por círculos à volta), então temos mais dois planos onde estes pontos estão dispostos que podem ser caracterizados por:

Definindo d1 e d2 como a distância de H1 e H2 ao hiperplano podemos observar que o hiperplano é equidistante a H1 e H2 o que por sua vez infere que d1=d2 e estes planos são conhecido como as margens da SVM. De modo a manter o hiperplano o mais longe possível aos Vetores de Suporte o objetivo é a maximização destas margens.

No entanto podem existir conjuntos onde os dados não são separáveis nas dimensões atuais. Para estes conjuntos é necessário uma utilização de uma função *kernel*. Uma função *kernel* permite a transformação de dados de *input* de modo a que estes sejam processados por algoritmos como as SVMs mais facilmente.

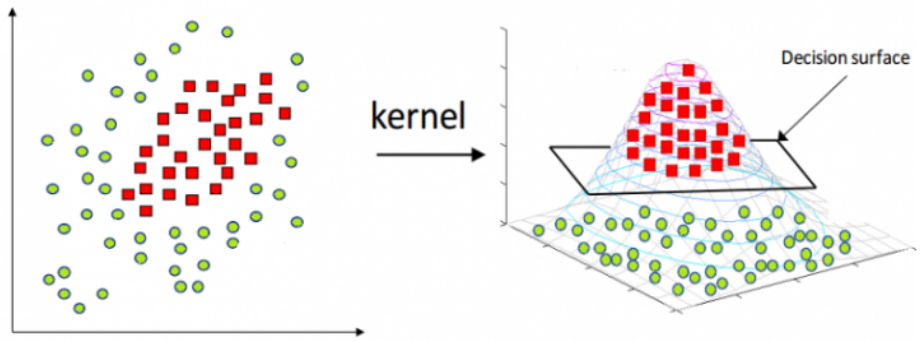

Figura 2.6: Exemplo de uma função *Kernel*. (Imagem extraída de: https://www.hackerearth.com/blog/machine-learning/simple-tutorial-svm-parameter-tuning-python-r/)

Na figura 2.6 podemos observar uma função *kernel* que transforma os dados projectando-os num espaço dimensional superior e tornando o conjunto inicial num conjunto linearmente separável.

Este algoritmo acaba por ser muito popular na vertente de classificação de dados com origem de detecção remota porque estes dados conseguem lidar com conjuntos de treino relativamente pequenos e conseguir produzir resultados aceitáveis, com métricas de avaliação superiores a métodos tradicionais probabilísticos [33]. Com a adição de *kernels* este classificador consegue ainda lidar com problemas de dados com dimensionalidade elevada, aumentando assim a sua flexibilidade e o seu desempenho geral.

No entanto devido a transformações espaciais complexas que mexem com a dimensionalidade do conjunto de treino, este classificador pode ser visto muitas vezes como uma *"blackbox"*. Este classificador também tem problemas em lidar com conjuntos de

dados pouco equilibrados , e a sua dependência de uma função *kernel* para lidar com dimensionalidades diferentes poderá ser o foco de resultados não satisfatórios.

2.3.5 Gradient Boosting

O *Gradient Boosting* é um método que recentemente tem ganho popularidade no âmbito de aprendizagem automática com dados remotos pois além de resultados bastante apelativos devido à sua possível paralelização, possui um tempo de treino e previsão baixos quando comparados com os outros dois classificadores apresentados anteriormente.

O *Gradient Boosting* [26, 29] é uma variação do algoritmo de *AdaBoosting* [17] que, como o nome indica, utiliza *Boosting* de modo a transformar classificadores fracos em classificadores fortes. No caso de *Tree Boosting*, cada árvore nova é ajustada em relação a um dataset que foi modificado consoante os resultados do último classificador a ser avaliado. Em *AdaBoosting* o algoritmo começa por treinar uma Árvore de Decisão (DT) e em que a cada observação (neste caso o píxel) é atribuído um peso inicialmente igual. Após a avaliação de resultados dessa primeira árvore, os pesos são alterados de modo a aumentarem para píxeis mais difíceis de classificar e diminuir para píxeis mais fáceis de classificar. Deste modo a segunda árvore a ser treinada, será treinada com dados onde já estão refletidos estes pesos adicionais no conjunto de píxeis. O objetivo é que com estes pesos acrescido o algoritmo consiga melhorar o seu desempenho. Com este conjunto de duas árvores é outra vez calculado o erro de classificação e subsequentemente criado uma terceira árvore que será treinada com novos pesos no dataset de treino. No final, a previsão final deste conjunto de árvores será a soma das previsões de todos os modelos posteriores de árvores.

A diferença entre o algoritmo *AdaBoost* e o *Gradient Boosting* é o modo como estes algoritmos tentam melhorar o desempenho dos classificadores. Ao contrário do algoritmo *AdaBoost*, o *Gradient Boosting* não usa pesos nos dados de treino mas sim uma função de perda com gradientes. Esta função de perda representa o sucesso do modelo ao se adaptar aos dados subjacentes. Isto permite ao *Gradient Boosting* utilizar uma função de custo/perda definida pelo utilizador.

2.3.6 Classificação Baseada em Píxeis

A classificação tradicional de imagens é baseada em píxeis. Cada píxel individual é analisado de acordo com a informação espectral que contém e, depois, devido ao seu carácter espectral, é lhe atribuída uma classificação [36]. Esta abordagem é ideal visto que o píxel é a unidade espacial fundamental da imagem e por isso representa um classificação natural e fácil de implementar.

A classificação dos píxeis como diferentes coberturas terrestres é feita fundamentalmente recorrendo a técnicas de reconhecimento de padrões [38] nomeadamente as implementações destas técnicas como algoritmos de aprendizagem automática. Fundamentalmente os padrões são os píxeis em si, numa imagem multi-espectral se cada píxel

tiver associado um vetor que contenha os valores de análise espectral do píxel para cada banda medida na imagem, a classificação do píxel envolve alocação (classificação) deste píxel a um determinado grupo espectral (classe) através da análise dos valores presentes no vetor associado ao píxel. Este processo é descrito como mapeamento e é demonstrado na figura 2.7.

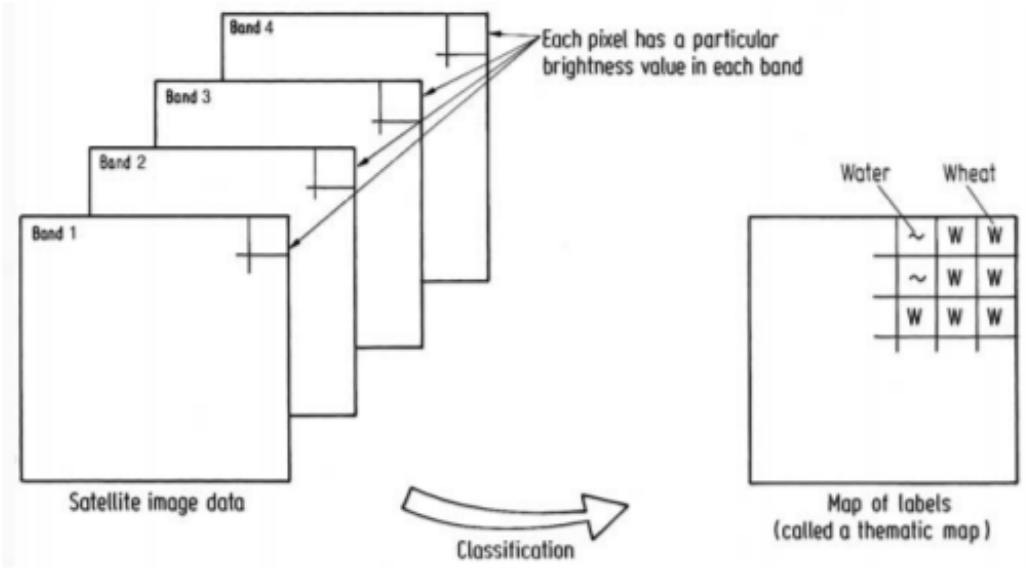

Figura 2.7: Classificação de Píxeis

Apesar desta abordagem ser muito utilizada tem algumas limitações [36]. Uma limitação fundamental desta abordagem é que a informação de píxeis circundantes não é usada na classificação, o que implica perda de informação que poderia ser útil na atribuição de classes. Consequentemente classes muito heterogéneas podem ter os seus píxeis classificados como classes diferentes. Este efeito é recorrente e denominada por efeito sal e pimenta, e em abordagens baseadas em píxeis pode introduzir um grau de erro de classificação significante no mapeamento de cobertura terrestre. Outra limitação desta abordagem é a possível existência de píxeis mistos. Estes píxeis, devido a resoluções mais baixas proveniente de satélites como o MODIS e AVHRR, apresentam diferentes coberturas terrestres dificultando a sua atribuição a uma classe apenas.

2.3.7 Classificação Baseada em objectos

A classificação baseada em objectos não opera directamente sobre píxeis [31, 37]. Em vez disso opera sobre objectos bem definidos dentro da imagem. Estes objectos podem ser definidos como um cluster contíguo de píxeis. É necessário, por isso, segmentar a imagem em regiões homogéneas e espacialmente contíguas. A este processo é dado o nome de segmentação de imagem. Apenas posteriormente à segmentação de imagem é que se implementa qualquer tipo de método de classificação.

A segmentação de imagem é muito importante para o sucesso do processo classificação da imagem. É necessário que a segmentação seja uniforme e homogénea e que esteja concordante com características da cobertura terrestre como a textura. Os interiores de cada região devem ser simples e sem muitos buracos, enquanto que as fronteiras de cada segmento devem ser simples, regulares e espacialmente precisas. Entre regiões adjacentes deve observar-se uma diferença entre os valores referentes às características da cobertura terrestre significativa. É difícil no entanto atingir estas propriedades durante a segmentação devido à própria natureza da cobertura terrestre.

As técnicas de segmentação de imagem podem-se classificar em três ramos principais: limiarização/clustering, técnicas baseadas em região e técnicas baseadas em beiras.

2.3.8 Segmentação Baseada em Limiarização/Clustering

Esta técnica de segmentação usa uma medida previamente determinada denominada de critério de clustering, para definir uma partição no espaço de segmentação [31]. Posteriormente cada píxel é atribuído à partição onde pertence consequentemente rotulando-se com a classe representativa dessa partição. Esta técnica caracteriza os segmentos da imagem como os componente conectados dos píxeis que têm a mesma classe.

Para o sucesso desta técnica é necessário que o processo de medição do espaço de clustering consiga dividir os objectos de interesse da imagem em espaços de clustering diferentes de modo a fornecer uma distinção clara entre as características dos objectos. Das várias técnicas utilizadas para a definição do espaço de clustering a que iremos discutir será a técnica de *histogram mode seeking* pois é a que envolve menos tempo de computação. *histogram mode seeking* é um processo de medição de espaço de clustering que assume que objectos homogéneos na imagem se manifestam como clusters no espaço de clustering. A segmentação é conseguida mapeando os clusters de volta para a imagem, onde os segmentos são constituídos por os componentes máximos dos clusters mapeados. Para imagens com apenas uma banda, o cálculo do histograma num vector é directo. Se os máximos locais do histograma constituem a segmentação da imagem então, de modo a determinar os clusters,é necessário de determinar-se os vales do histograma e os clusters serão os intervalos de valores entre vales. Se um píxel cujo o valor esteja no intervalo n entre vales então este terá o rótulo n e pertencerá ao componente onde todos os píxeis terão o rótulo n.

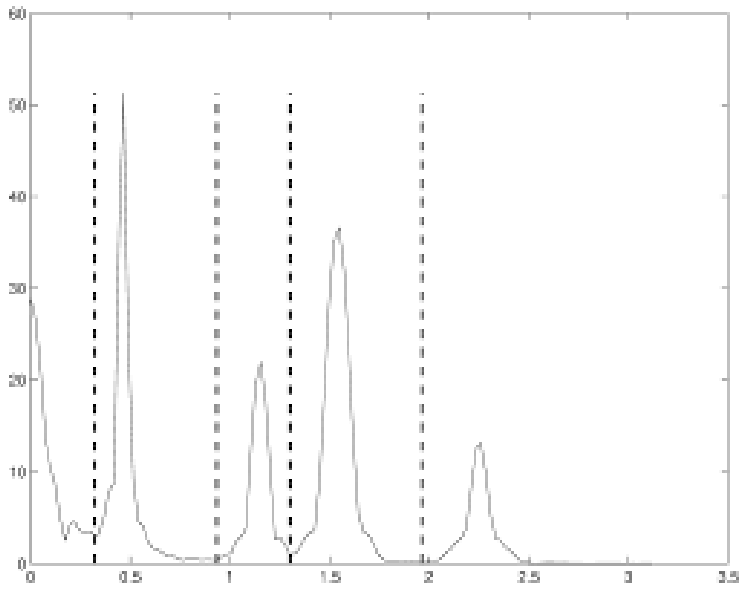

Figura 2.8: Histograma de uma Imagem

Na figura 2.8 pode-se observar o histograma proveniente de uma imagem. Se considerarmos as linhas a tracejado como o ponto central de cada vale podemos facilmente classificar os clusters. O primeiro cluster será do início do gráfico até à primeira linha a tracejado (primeiro vale) o segundo cluster será entre o primeiro vale e o segundo vale, o terceiro cluster entre o segundo e o terceiro, o quarto cluster entre o terceiro e o quarto e finalmente o quinto cluster entre o quarto vale e o final do gráfico.

No entanto nem sempre os histogramas resultantes de uma imagem são assim tão bem definidos e analisáveis.

Se a imagem contiver um objecto claro colocado num fundo escuro e o espaço de clustering for uni-dimensional, o processo de clustering de espaço de medição terá que definir um limiar de modo a que todos os píxeis inferiores ou iguais ao limiar sejam atribuídos a um cluster e todos os outros ponto que serão superiores ao limiar sejam atribuídos ao outro cluster. Idealmente seria apenas necessário a análise do histograma e colocar o limiar no vale entre os dois clusters. No entanto em imagens mais homogéneas, e que porventura terão um histograma mais difícil de analisar e com vales com $y \neq 0$ entre clusters, não será possível usar este procedimento. Para lidar com este problema é necessário técnicas que combinem a informação espacial da imagem com a informação espectral da mesma. Pegando na ideia de ter um objecto claro num fundo preto foi sugerida uma técnica para a resolução deste problema por Panda e Rosenfold [45]. Considerando um histograma de informação espectral para píxeis com gradientes pequenos, se um um píxel tiver um gradiente pequeno não é provável que este píxel esteja numa aresta/borda. Se não for uma aresta/borda então ou faz parte do objecto claro ou do fundo escuro. Neste caso histograma representativo da informação espectral destes píxeis será bimodal (terá dois picos) e será adequado para a definição do limiar. Poderemos usar a técnica dos valores

entre vales de modo a segmentar a imagem. Considerando agora o histograma de informação espectral sobre os píxeis com um gradiente elevado, estes píxeis representam as arestas/bordas a separar o objecto claro do objecto escuro. Se as arestas que separam o objecto do fundo forem relativamente difusas então o histograma será unimodal e adequadas para a atribuição do limiar, separando o objecto claro do fundo escuro. É então adequada a definição de dois limiares, um para píxeis com gradiente elevado e outro para píxeis com gradiente baixo. Este clustering é executado num espaço de medição de duas dimensões uma representado o gradiente e outra a informação espectral.

Para imagens com múltiplas bandas espectrais não é alcançável a determinação do histograma. Por exemplo numa imagem de seis bandas, em que cada banda tenha intensidades de 0 a 99 o vector teria o espaço de 10^{12} valores. Uma imagem que tenha 10,000 píxeis por coluna e 10,000 píxeis por linha apenas teria 10^8 píxeis o que seria uma amostra demasiado pequena para o espaço de 10^{12} valores.

Clustering usando um histograma multi-dimensional é mais difícil do que clustering usando histograma uni-dimensional. Uma técnica proposta [30] é a limiarização do histograma multi-dimensional de modo a seleccionar todos os tuplos situados nos picos mais proeminentes. Posteriormente seria feita uma medição do espaço de clustering de modo a recolher e juntar os tuplos que se situam no topo dos picos mais proeminentes. Este processo conectaria os conjuntos dos núcleos de cluster. Os clusters seriam definidos como o conjunto de todos os tuplos mais próximos de cada núcleo de cluster.

2.3.8.1 Segmentação Baseada em Região

Os esquemas de crescimento de região com ligação simples consideram cada píxel como o nó de um grafo [31]. píxeis vizinhos cujas propriedades sejam semelhantes são ligados por um arco. Os segmentos de imagem são formados consequentemente pelos conjuntos máximos de píxeis que pertencem ao mesmo componente conectado. Esquemas de segmentação de imagem com ligação simples são desejáveis devido à sua simplicidade. No entanto, são extremamente sensíveis a erros de encadeamento. Se uma arco incorrecto se formar entre um nó e outro, pode verificar-se uma fusão de regiões com características muito distintas.

As semelhanças entre píxeis são definidas, da maneira mais simplista possível, pela diferença da informação espectral entre dois píxeis. Se esta diferença for pequena o suficiente os píxeis são considerados similares e juntados por um arco.

Para píxeis com vectores de valores, a solução é realizar a norma do vector resultante da diferença entre os dois valores. Uma sugestão muitas vezes utilizada [5] é a comparação da diferença entre dois píxeis com a media dos valores entre um píxel no centro do grafo e os píxeis vizinhos desse píxel central. Se, comparativamente, a diferença entre os dois píxeis for pequena o suficiente então os dois píxeis são juntados pelo arco.

Para a resolução dos problemas provenientes das técnicas de crescimento de região com ligação simples, foram desenvolvidas técnicas de crescimento de região com ligação

híbridas [31]. Estas técnicas procuram atribuir um vector de propriedades a cada píxel onde o vector de propriedade depende da área da vizinhança de cada píxel. píxeis que são semelhantes, são semelhantes porque as suas vizinhanças são semelhantes. Assim, a semelhança entre píxeis é estabelecida como uma função de valores de píxeis na semelhança sendo esta técnica menos sensível a erros de encadeamento.

Um esquema de crescimento de região com ligação híbrida consiste em utilizar um operador de arestas para denominar cada píxel como aresta ou não. Píxeis vizinhos que não são arestas e que são semelhantes são juntados por um arco. Os segmentos iniciais são são todos os componentes conectados de todos os píxeis que não são arestas. Todos os píxeis arestas podem ser considerados como fundo ou podem ser atribuídos à região espacialmente mais perto.

A qualidade desta técnica está muito dependente do operador de detecção de arestas usado. Operadores mais simples podem fornecer demasiada ligação entre regiões e por isso promover encadeamento descontrolado.

Por fim a última técnica de segmentação de imagens baseada em região é a de divisão e fusão. Esta técnica começa com a imagem inteira como o segmento inicial. Posteriormente e consecutivamente se o segmento não for homogéneo o suficiente este será divido em partes mais pequenas. Homogeneidade pode ser facilmente estabelecida determinando se a amplitude de intensidades espectrais for pequena o suficiente. Algoritmos deste tipo foram sugerido pela primeira vez por Robertson e Klinger [39]. A eficiência desta técnica pode ser aumentada se a imagem for arbitrariamente particionada em regiões quadradas de um determinado tamanho definido pelo utilizador, e, se homogeneidade não tiver sido atingida, dividir estas regiões ainda mais.

Devido à sucessiva divisão de segmentos em segmentos de menor tamanho, as fronteiros entre alguns segmentos tendem a ser demasiado artificiais. O que é indicador de que estes segmentos precisam de ser fundidos em vez de divididos. Se duas ou mais regiões adjacentes tem uma distribuição espectral homogénea e possuem características espectrais semelhantes então estas devem ser fundidas

2.3.8.2 Segmentação Baseada em Fronteiras

Os métodos de segmentação de imagem baseada em fronteiras/arestas [32] transforma as imagens originais em imagens de arestas que beneficiam com as mudanças drásticas de tons na imagem. Em processamento de imagem a detecção de arestas localiza variações espectrais importantes numa imagem assim como a detecção de propriedades físicas e geométricas dos objectos de estudo. É um processo fundamental para a detecção de descontinuidades significantes nos valores de intensidade espectral.

Arestas são mudanças locais na intensidade espectral da imagem. As arestas tipicamente estão localizadas nas fronteiras entre duas regiões.

Existem diversas técnicas de detecção de arestas, nomeadamente a técnica de detecção de arestas *Canny* é das técnicas mais utilizadas e com melhores resultados e brevemente

descrita a seguir.

A técnica *Canny* foi criada por John Canny em 1983 e continua a ultrapassar técnicas mais recentes. Esta técnica baseava-se na separação da imagem de possível ruído antes de recorrer à detecção de arestas. Este método *Canny* consegue evitar a criação de perturbações na imagem após a separação do ruído. O algoritmo de forma geral segue os seguintes passos:

- Convolver a imagem f(r, c) com uma função Gaussiana de modo a a suavizá-la.
- Aplicar a primeira diferença de gradiente de modo a computar a intensidade das arestas depois a sua magnitude e por fim direcção
- Aplicar uma suspensão crítica à magnitude do gradiente
- Aplicar um limiar à imagem de supressão que não é a máxima

2.3.9 Classificação Baseada em Texturas

Nesta secção pretende-se descrever uma abordagem de classificação adicional baseada em texturas. Mais especificamente será descrito o classificador de redes neuronais profundas e como é que este poderá ser uma ferramenta eficaz na classificação de cobertura terrestres nomeadamente vegetação permanente.

2.3.9.1 Redes Neuronais Profundas

As redes neuronais profundas podem ser utilizadas para a classificação de cobertura terrestre [24]. No entanto a implementação destas redes pode ser um desafio devido à natureza dos dados utilizados nestes problemas. Cada píxel contem valores físicos como a refletância em bandas múltiplas bandas espectrais e em casos onde informação SAR é acrescentada pela intensidade do retro-espalhamento das ondas em múltiplas polaridades. No entanto é possível utilizar uma rede neuronal profunda para a classificação de cobertura terrestre usando dados de múltiplas fontes.

Esta rede neuronal profunda é constituída por um conjunto de redes neuronais convolucionais (CNNs). A soma total das bandas usadas pelo sensor utilizado formam um vector que servirá como input de características para cada CNN. CNN tradicionais (de duas dimensões) têm em conta o contexto espacial de uma imagem, o que é uma melhoria em relação à abordagem tradicional baseada em píxeis. Neste contexto isto permite ao algoritmo diferenciar pequenas estradas, ou pequenas tiras isoladas de floresta que seriam mal classificadas ou então perdidas na sua íntegra. A figura 2.9 representa uma típica rede neuronal convolucional [24]

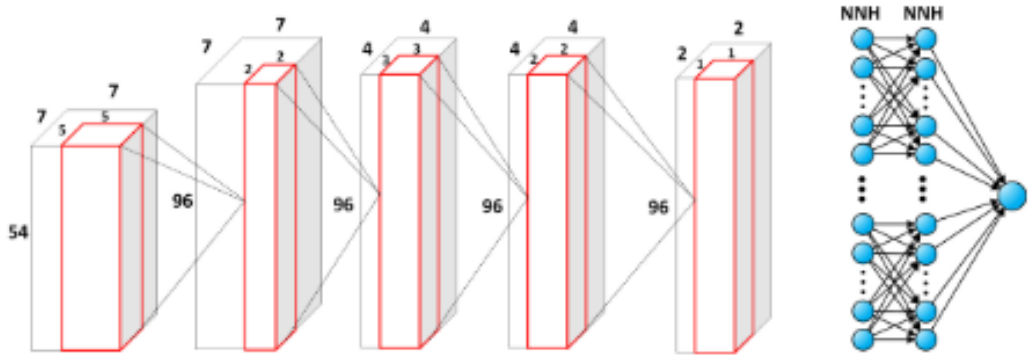

Figura 2.9: Arquitetura de uma Rede Neuronal Convolucional

Como se pode observar na figura 2.9, uma arquitectura tradicional de uma CNN corresponde a duas camadas convolucionais seguidas de uma camada de *pooling* máximo (de modo a a reduzir o tempo de computação e garantir invariância através das diferentes dimensões) e duas camadas totalmente conectadas. A função de activação mais usada para este contexto de detecção remota é a unidade linear retificada (ReLU). A ReLU permite uma computação eficiente e propagação eficiente do gradiente. Consequentemente, neste contexto, a ReLU tem um comportamento mais eficiente e eficaz do que a função sigmóide, pois esta exibe um comportamento que se assemelha mais ao funcionamento de um neurónio físico. Cada CNN tem a mesma estrutura convolucional e de *pooling* máximo diferindo apenas no número de filtros treinados e neurónios na camada escondida.

2.4 Métricas de Avaliação

Após a obtenção dos resultados do processo de classificação, é necessário quantificar-se a importância e o significado de cada valor obtido, deste modo permitindo inferência de conclusões objectivas com base neste valores quantificados. À quantificação dos valores obtidos dá-se o nome de métricas de avaliação e nesta secção pretende-se uma síntese das métricas que serão mais úteis no âmbito desta dissertação, assim como o significado dos seus valores.

2.4.1 Matriz de Confusão

A Matriz de Confusão é uma tabela que, através da comparação dos resultados obtidos das classes previstas com as classes verdadeiras (que foram previamente pré-classificadas manualmente), permite a visualização do desempenho de um determinado algoritmo de classificação.

Nesta tabela as colunas representam cada instância das classes verdadeiras e as linhas representam cada instância das classes previstas, ou vice-versa. Cada entrada nesta tabela significa que um membro que previamente estava classificado como a classe x com este algoritmo de classificação foi classificado como y. Esta tabela permite uma fácil visualização

de erros de classificação e adicionalmente quais as classes mais dificilmente discerníveis.

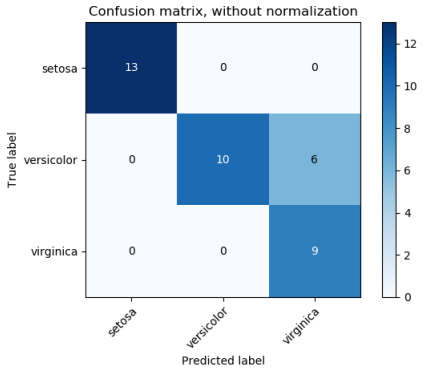

Figura 2.10: Exemplo de uma Matriz de Confusão (Imagem extraída de: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html)

Na figura 2.10 está demonstrada uma matriz de confusão para um sistema de classificação de plantas Íris neste caso em 3 espécies diferentes pertencentes a esta classe de plantas. Existem 38 amostras 13 do tipo *setosa*, 16 do tipo *versicolor* e 9 do tipo *virginica*. Pode-se observar que das 16 amostras de *versicolor* apenas 10 foram classificadas correctamente como *versicolor* e as outras 6 foram classificadas como *virginica*, no entanto em ambas as restantes classes todas as amostras foram previstas correctamente. Pode-se aferir também que todos os resultados correctos estão na diagonal da tabela facilitando assim a procura de erros, sendo estes ocorrências fora da diagonal da tabela.

A partir desta tabela também se podem aferir outras métricas com valores relevantes para o tema da dissertação. Estas métricas são a *accuracy*, a *precision* o *recall* e o *f1-score*.

Destas métricas a mais fácil de compreensão é a *accuracy*. Esta métrica é a fracção de ocorrências classificadas correctamente sobre o número total de ocorrências. Neste caso a *accuracy* total seria 0.842 (32/38). No entanto devido ao desequilíbrio entre classes de muitos datasets, esta métrica por si só não permite a inferência de conclusões porque classes menos numerosas têm um peso inferior o que tendencia a métrica.

A *precision* mede a fracção de ocorrências previstas que foram correctamente classificadas, apresentando assim uma medida de confiança positiva. Neste caso para a classe *virginica* a *precision* é de 0.6 (9/15). Esta métrica é útil pois em problemas com âmbitos semelhantes ao desta dissertação serviu para medir a frequência com que os píxeis eram classificados na mesma área que os polígonos das classes verdadeiras (*ground truth*). A *precision* varia entre 0 e 1 sendo que 1 seria o seu valor ideal.

O *recall* por outro lado mede a fracção de ocorrências corretas que foram classificadas, mostrando assim a frequência relativa de positivos que de facto são apanhados. Neste caso para a classe *versicolor* a *recall* é de 0.625 (10/16). Esta métrica por sua vez é útil em problemas com âmbitos semelhantes ao desta dissertação porque serve para avaliar o grau de preenchimento de cada polígono. O *recall* varia entre 0 e 1 sendo que 1 seria o

seu valor ideal.

O *f1 score* é uma métrica que se calcula fazendo a média harmónica entre a *precision* e o *recall* ($2 * \frac{Precision*Recall}{Precision+Recall}$). Esta medida pretende substituir a *accuracy*, sendo que é sensível a desequilíbrios no dataset e por isso com maior utilidade. O *f1 score* varia entre 0 e 1 sendo que 1 seria o seu valor ideal.

2.4.2 Coeficiente de Cohen's Kappa

O coeficiente de Cohen's Kappa [12] é uma medida estatística de concordância entre avaliadores em casos categóricos (qualitativos). Este mede a concordância entre dois avaliadores que classificam um número determinado de objectos em classes mutuamente exclusivas. É geralmente mais robusto que cálculos simples de percentagem de concordância porque tem em conta o acaso.

A sua fórmula é a seguinte:

$$K = \frac{p0 - pe}{1 - pe}$$

Onde $p0$ é a concordância reactiva observada entre avaliadores e pe é a probabilidade hipotética de concordância por acaso. Se os avaliadores estiverem em perfeita concordância então $k = 1$, por outro lado se não houver concordância entre avaliadores então $K <= 0$.

Previsto/Verdade	Vinha	Não Vinha
Vinha	20	5
Não Vinha	10	15

Figura 2.11: Matriz de Confusão para Arvoredo Denso

Considerando a tabela na figura 2.11 como o resultado de um algoritmo preliminar que pretende diferenciar vinha de outro tipo de vegetação, como o algoritmo previu correctamente 20 classes de vinha e 15 classes de não vinha então:

$$p0 = \frac{20 + 15}{50} = 0.7$$

Como o algoritmo previu metade das ocorrências como vinha e na realidade 60% das ocorrência são vinha então a probabilidade do algoritmo identificar uma vinha verdadeira como vinha é:

$$ps = 0.5 * 0.6 = 0.3$$

Por outro lado aplicando a mesma lógica a classes não vinha:

$$pn = 0.5 * 0.4 = 0.2$$

E portanto pode-se agora calcular *pe*:

$$pe = 0.3 + 0.2 = 0.5$$

E substituindo *pe* na fórmula do Kappa:

$$K = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

2.4.3 Validação dos Modelos

Nesta secção irão ser expostas duas metodologias de validação modelos dos classificadores. Estas metodologias são a *Grid Search* e a *Random Search*. Primeiramente estas metodologias serão expostas e de em seguida serão comparadas.

2.4.3.1 Grid Search

Na metodologia *Grid Search*, mediante uma lista de possíveis valores para os hiper-parâmetros de um modelo, todas as possíveis combinações de valores são utilizadas para treino e avaliadas posteriormente [18].

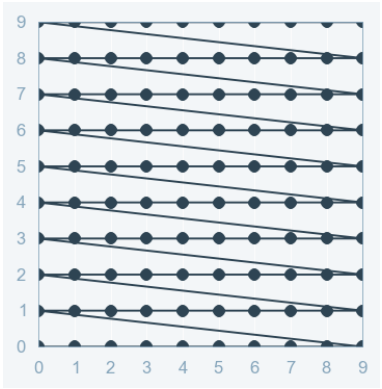

Figura 2.12: Representação Visual de uma *Grid Search* (Imagem retirada de https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318)

Como se pode observar na figura 2.12 este padrão é muito semelhante a uma grade, onde todos os valores formam uma matriz. Nesta figura cada ponto representa uma combinação de valores diferentes dos hiper-parâmetros a optimizar, cada linha as combinações a serem avaliadas para a optimização dos parâmetros e cada coluna representa todos os valores possíveis que um hiper-parâmetro poderá ter. Após a avaliação de todas as combinações, a combinação que apresenta a melhor métrica de validação é considerada o modelo ideal.

Ao considerar todas combinações possíveis esta metodologia garante também que o melhor modelo que encontrar será o modelo óptimo dentro dos valores dos hiper-parâmetros. No entanto para atingir esta invariante torna-se um processo demasiado

pesado para um conjunto de hiper-parâmetros elevado pois o número de iterações necessárias com o aumento de cada parâmetro aumenta exponencialmente revelando-se uma técnica pouco prática para muitos modelos actualmente existentes.

2.4.3.2 Random Search

A *Random Search* em oposição à *Grid Search* considera apenas um subconjunto aleatório de todas as combinações possíveis para encontrar o melhor modelo. Este espaço de combinações possíveis é parametrizável pelo utilizador e no final de cada treino a função é avaliada de maneira idêntica à de *Grid Search* [18].

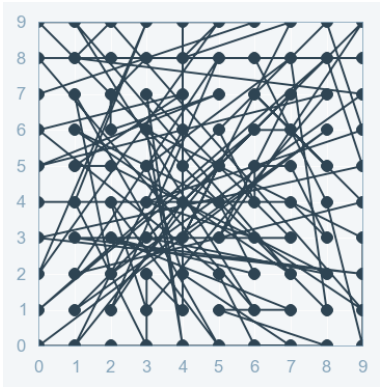

Figura 2.13: Representação Visual de uma *Random Search* (Imagem retirada de https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318)

A figura 2.13 tem uma composição semelhante à figura 2.12. No entanto comparativamente com a figura 2.12 pode observar-se que na figura 2.13 não são consideradas todas as combinações possíveis de valores dos hiper-parâmetros.

Apesar de a *Random Search* não garantir que o melhor modelo encontrado será o modelo óptimo, esta garante uma probabilidade elevada de encontrar o modelo óptimo ou no mínimo um modelo muito próximo do óptimo. Aliado ao facto de não possuir um peso computacional tão elevado como a *Grid Search* a *Random Search* é uma opção adequada para efectuar a validação do modelo especialmente em dimensões baixas de hiper-parâmetros mas que exigiriam um custo computacional elevado por parte da *Grid Search*

2.5 Trabalho Relacionado

Nesta secção pretende-se fazer uma exposição dos trabalhos relacionados com os conceitos descritos neste capítulo, e realizados com o objectivo de classificar correctamente vegetação permanente.

Primeiramente pretende-se salientar a importância da escolha do sensor (Sentinel) e o porquê de se o ter preferido a outros sensores de uso gratuito como o Landsat. De em seguida será referida a importância da utilização de imagens provenientes de SAR

na classificação de cobertura terrestre. Por fim serão comparadas as três abordagens de classificação de dados de detecção remota assim como todos os classificadores referidos.

Abdikan et al. [1] utilizaram imagens de Sentinel-1 SAR no mapeamento de cobertura terrestre. Aliado a estas imagens usaram um classificador de *Support Vector Machines* que foi implementado com uma função de base radial como função de *Kernel*. Na classificação foram usados duas polarimetrias diferentes VV e VH, usando-se em diferentes cenários diferentes combinações destas polarimetrias. No final verificou-se que através de 5 variáveis baseadas nas duas polarimetrias originais foram atingidas precisões superiores a 90%.

Yu et al. [46] criaram um inventário compreensivo de vegetação para uma área no Norte da Califórnia utilizando uma técnica de análise de imagem orientada a objectos e conseguiram demonstrar empiricamente que esta abordagem conseguiu mitigar o efeito sal pimenta verificado em classificação orientada a píxeis.

Heyman et al. [21] usaram uma abordagem orientada a objectos numa tentativa de melhorar a classificação de choupos em *Oregon* central classificando-o em três classes. Esta obteve uma melhoria significativa na *accuracy* de classificação dos choupos, revelando-se muito apta na descriminação de tipos de vegetação diferentes.

Stow et al. [43] conseguiram diferenciar tipos de vegetação arbustiva, utilizando uma técnica de análise de imagem orientada a objectos , numa zona costeira da Califórnia conseguindo posteriormente aferir que os padrões de diminuição de população arbustiva estavam mais relacionados com actividade antropogénica do que com secas longas.

Johansen et al. [23] mapearam a estrutura da vegetação na ilha de *Vancouver*, descriminando os estados estruturais da vegetação ripária e adjacente, utilizando imagens de *QuickBird* e classificação orientada a objectos. Este mapeamento resultou num mapa detalhado dos vários tipos estruturais de vegetação.

Ma et al. [27] fizeram uma avaliação de diversos artigos relacionados com métodos supervisionados de classificação orientada a objectos e chegaram à conclusão que: as imagens de alta resolução espacial continuam a ser a fonte de dados mais usada para a classificação orientada a objectos de cobertura terrestre nomeadamente as imagens de Landsat, no entanto imagens sensores de maior resolução espacial nomeadamente imagens provenientes de UAV obtêm melhores resultados. Técnicas de segmentação de múltiplas escalas são as técnicas mais usadas como algoritmos de segmentação. Grande parte dos estudos são incidentes sobre áreas menores que 300 ha. Dos classificadores supervisionados o classificador que se destaca com melhores resultados é o classificador de *Random Forest*.

Kussul et al. [24] usaram uma classificação baseada em redes neuronais profundas, em que usou duas arquitecturas distintas: uma de apenas uma dimensão; e outra de duas dimensões. Esta abordagem obteve resultados muito bons ultrapassando classificadores de sucesso como *Random Forest* e redes neuronais artificiais.

Oruc et al. [28] utilizaram imagens de Landsat 7 de modo a compararem as abordagens orientadas a píxeis e a objectos na zona de *Zonguldak* na Turquia. Verificou-se que a

classificação orientada a objectos produziu resultados mais precisos do que a classificação orientada a píxeis.

Castillejo-González et al. [10] usaram imagens de QuickBird para comparar classificação orientada a píxeis e classificação orientada a objectos em ambientes agrícolas. Os métodos obtiveram uma *accuracy* semelhante usando o classificador de *Maximum Likelihood*.

Duro et al. [14] comparou a utilização de *Support Vector Machines*, *decision trees* e *Random Forest* em classificações orientadas a píxeis e classificações orientadas a objectos e verificou que as implementações orientadas a objectos foram claramente superiores e que de entre os classificadores o que se destacou foram as *Support Vector Machines*.

Brenning [4] comparou 11 algoritmos de classificação utilizando classificação orientada a píxeis e imagens Landsat e verificou que o classificador com resultados superiores era o discriminante linear penalizado, superando classificadores como *Random Forest* e *Support Vector Machines*.

Otsukei e Blaschke [35] compararam classificadores baseados em *Support Vector Machines*, *Decision Trees* e *Maximum Likelihood* de modo a aferir qual o que demonstrava melhor comportamento na classificação de florestação em Uganda. O classificador que superou os outros foi as decision trees, apesar de todos os classificadores terem apresentado resultados aceitáveis acima dos 85%. Verificou-se que no caso das *Support Vector Machines* a simplificação do espaço vetorial, através da diminuição do número de bandas, aumentou a *accuracy* deste classificador. Nos outros classificadores, este processo não alterou a *accuracy*.

Noi e Kappas [33] utilizaram imagens de Sentinel-2 para comparar o comportamento de três classificadores: *Random Forest*, *Support Vector Machines* e *k-Nearest Neighbor*. Foram usados 14 conjuntos de dados diferentes, com diferentes conjuntos de dados. O classificador de *Support Vector Machines* destacou-se apresentando maior *accuracy* e menor susceptibilidade ao tamanho do conjunto de treino. Para os outros dois classificadores o seu comportamento foi semelhante, apresentando maior susceptibilidade à alteração do tamanho do conjunto de treino.

Fernandes [15] utilizou imagens de *LANDSAT 8* para a comparação de métodos de detecção e classificação de culturas. Com este objectivo utilizou 5 algoritmos sendo estes *Random Forests*, *Support Vector Machines*, *k-Nearest Neighbor*, *Maximum Likelihood Classifier* e *Decisions Trees*, e analisou os resultados, utilizando uma metodologia temporal estática e metodologia de séries temporais. Estes testes tiveram muito bons resultados especialmente com os algoritmos RF e SVM, ambos com séries temporais.

Nunes decidiu expandir o trabalho de Fernandes mencionado anteriormente aumentando o número de classes a classificar e também avaliar o desempenho destes algoritmos utilizando imagens provenientes de sensores diferentes [34]. Os sensores em questão são o *LANDSAT* 8, o Sentinel-1 e o Sentinel-2. A área de estudo também foi aumentada consideravelmente. À semelhança do trabalho anterior os classificadores com melhor desempenho foram as RF e SVMs com séries temporais. Adicionalmente verificou-se que o

melhor sensor de satélite para a detecção de espécies de cultivo é o Sentinel-2. Por fim verificou-se que as diferentes classificações mantiveram-se robustas apesar do aumento significativo da complexidade dos dados a classificar.

Comparativamente a estes trabalhos, observa-se que o número de amostras que serão utilizadas para este trabalho será consideravelmente maior, introduzindo complexidade adicional ao problema. Esta particularidade poderá afectar o desempenho de diversos classificadores devido ao elevado número de amostras a serem consideradas. Também se reflecte que nenhum destes trabalhos tem um sistema de classificação utilizando classes tão abrangentes de diferentes tipos de vegetação. Será interessante ver o comportamento e desempenho dos algoritmos perante um sistema de classificação relativamente subjectivo. Finalmente observa-se que muitos destes estudos tem um número de classes de vegetação semelhante ao que será usado no âmbito desta dissertação.

Capítulo 3

Abordagem

O objectivo desta dissertação é desenvolver uma ferramenta capaz de classificar vegetação permanente no contexto presente nas cartas militares, utilizando dados de detecção remota. Para isso irão ser comparados três classificadores distintos, nomeadamente SVM, RF e *Gradient Boosting* utilizando produtos de Sentinel-1 e Sentinel-2. Esta ferramenta seguirá duas metodologias principais, metodologia temporal estática, e metodologia de série temporal. Dentro da metodologia de série temporal foram desenvolvidas diversas variantes no intento de aferir qual a maneira mais eficaz de se aproveitar a informação adicional proveniente de uma série temporal. De uma maneira geral a arquitectura da abordagem prevista segue um fluxo de execução como o descrito no diagrama da figura 3.1.

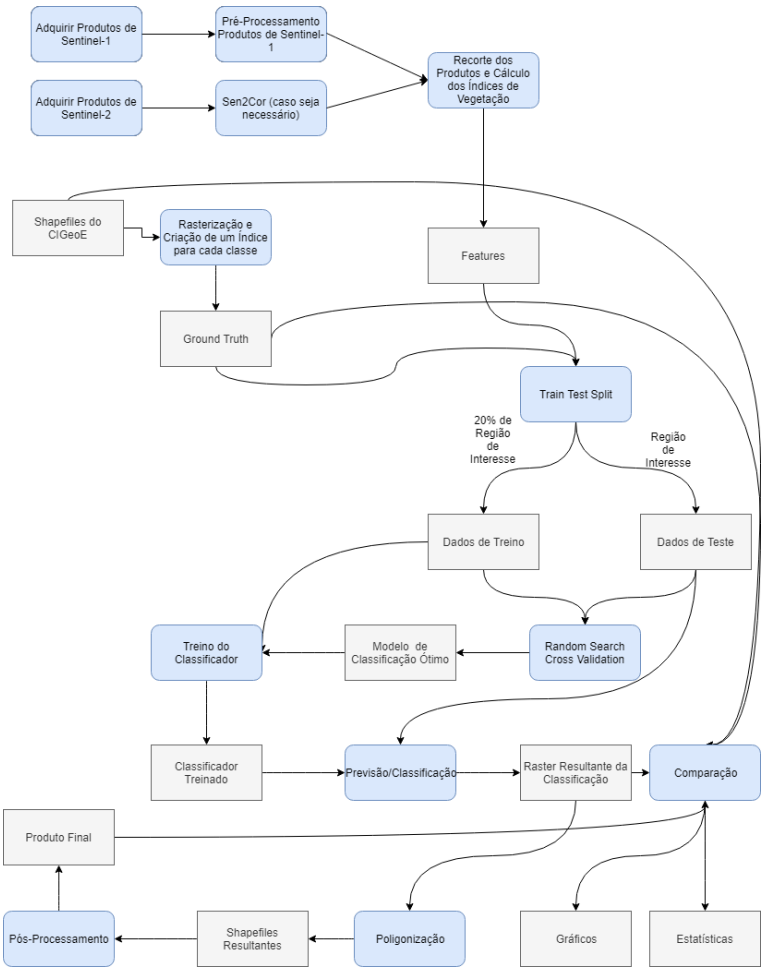

Figura 3.1: *Flow Chart* da Arquitectura da Abordagem Mencionada

3.1 Trabalho Preliminar

Foi realizado trabalho preliminar no que consta ao estudo de outros produtos de classificação de cobertura terrestre e de que maneira é que estes se podem relacionar com a classificação pretendida pelo CIGeoE. Nesta secção pretendemos dar uma visão geral de como é que este objectivo foi atingido.

3.1.1 Estudo de Impacto das Cartas de Ocupação de Solo

A direcção geral do território (DGT) está responsável pela produção das cartas de ocupação (COS), estas cartas possuem 48 classes e têm como unidade mínima cartográfica de solo de um hectare. Apesar da diferença de resolução destas cartas para os produtos obtidos pelo CIGeoE é interessante saber se poderá existir alguma correlação entre elas.

Para isso foi criada uma matriz de incidências entre as duas classificações, em que as linhas são as classes presentes nas cartas COS e nas colunas são as classes dos vegetação de CIGeoE. Esta matriz de confusão é criada píxel a píxel com a resolução do CIGeoE

contando as ocorrências em que um píxel de uma determinada classe de vegetação do CIGeoE é classificada como uma determinada classe nas cartas COS.

Após a criação dessa matriz é então criado um vector que faz a correspondência entre as classes COS, sendo estas o índice do vector, e a classe CIGeoE que teve o maior número de incidências nessa mesma classe, correspondendo estas classes ao que está contido em cada posição do vector.

Este vector é usado então para criar um raster baseado na classificação COS mas em que, para cada píxel, é atribuído o valor da classe de CIGeoE obtida a partir vector referido.

Por fim são calculadas as métricas de validação, permitindo assim uma boa ideia da intensidade de uma possível correlação entre estes dois sistemas de classificação distintos.

3.2 Preparação dos Dados Ground Truth

Como *ground truth* o CIGeoE forneceu 4 *shapefiles* respectivos a 4 regiões distintas de Portugal. Estes *shapefiles* têm 10000 por 16000 metros de área e detêm já uma classificação da vegetação permanente de acordo com o contexto da carta militar. Todos estes *shapefiles* têm a sua classificação proveniente de ortofotos tiradas em Maio de 2015. Estas regiões são Fundão, Monchique, Sendim do Douro e Praia da Tocha.

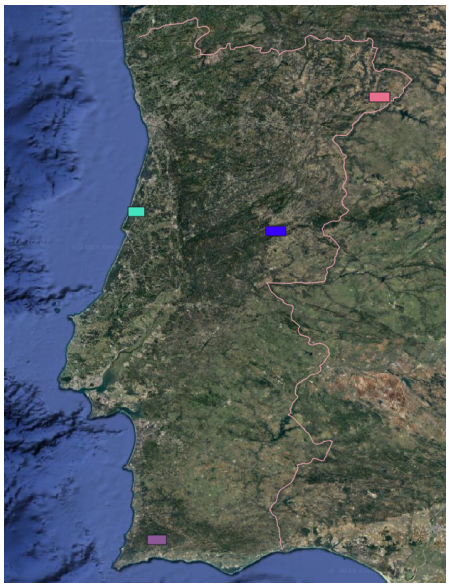

Figura 3.2: Áreas de Interesse

Para a utilização desta informação como dados de input de treino e para uma classificação píxel a píxel foi necessário adicionar-se uma *feature* numérica a estes *shapefiles* que correlaciona o tipo de vegetação a um algarismo, funcionando assim com um *id* único de cada classe. Assim foi possível rasterizar-se estes *shapefiles* utilizando o seu *Id*, no *CRS*

EPSG:32629, em que cada píxel representa uma área de 10 por 10 metros. Assim garantimos que a resolução espacial é igual à dos produtos de Sentinel-2 e que ambos os dados se encontram alinhados.

3.3 Produtos de Sentinel

Para aquisição de produtos de Sentinel foram usados métodos diferentes para produtos de Sentinel-1 e Sentinel-2.

Os produtos de Sentinel-1 foram obtidos através de do *Alaska Satellite Facility* (ASF) [3], devido a muitos dos produtos mais antigos de Sentinel-1 estarem *offline* no *Copernicus Open Acess Hub*, e apenas disponibilizados após um pedido ao *Hub* que poderá demorar dias e por isso se revela um processo demasiado moroso. Após a aquisição dos produtos de Sentinel-1 é necessário realizar-se um pré-processamento intensivo a estas imagens de modo a estas produzirem benefícios no processo de classificação.

Este pré-processamento foi efectuado na ferramenta Sentinel Application Platform (SNAP) [44] para *Desktop* onde através das ferramentas de *Graph Builder* e de *Batch Processing* foram aplicados os seguintes passos de processamento aos produtos:

- **Calibração** A calibração radiométrica é aplicada de modo a cada píxel reflectir o valor correto de *backscattering* medido pelo sensor e assim poder ser utilizado com *feature* para input de classificação.
- **Aplicação de *Orbit File*** Este passo garante que a geo-codificação da imagem de Sentinel-1 é óptima, melhorando o desempenho dos passos de processamento posteriores.
- **Achatamento do Terreno** Tenta corrigir distorções provocadas por diferentes ângulos de incidência, utilizando Modelos Digitais do Terreno (DEM).
- **Correcção do Terreno** Pretende remover a interferência de topografia nos valores de *backscatter*. Estas distorções são provocadas por medição de lado e não medição directamente por cima. Ao sofrerem correcção os píxeis são movidos para a sua localização correta e as relações espaciais entre píxeis são corrigidas.
- **Filtragem de *Speckle*** Por fim a filtragem de *Speckle* pretende remover ruído/efeito sal pimenta na imagem.

Nas imagens 3.3 e 3.4 pode-se observar o resultado do pré-processamento efectuado nas imagens de Sentinel-1. As bandas que foram utilizadas para o treino dos algoritmos, foram as bandas VV e VH. Foram também calculadas através do SNAP, *features* adicionais que são relevantes para este contexto como a *Grey Level CO-ocurrence Matrix* (GLCM). A GLCM foi calculada a partir dos produtos de Sentinel-1 e foi gerada através da ferramenta de *Texture Analysis* presente no SNAP. A GLCM enriquece as *features* com

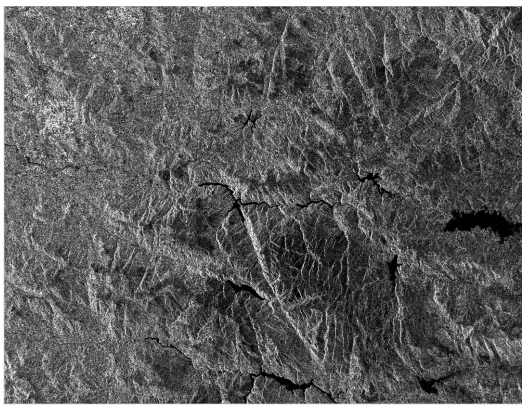

Figura 3.3: Imagem de Sentinel-1 antes do pré-processamento

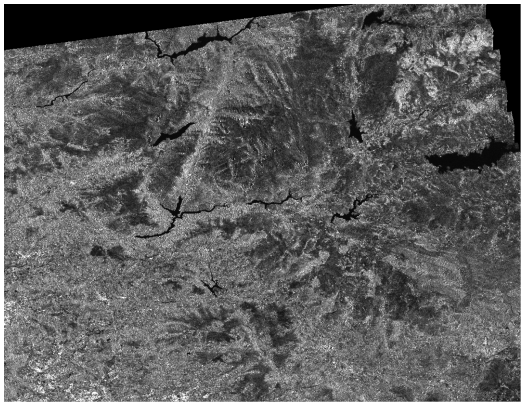

Figura 3.4: Imagem de Sentinel-1 após o pré-processamento

informação referente à textura da área referente a cada produto [19]. Com este objectivo a GLCM quantifica diversas ocorrências nas imagens como o contraste (diferenças de tom cinzento), o tamanho da área de cada contraste e a direcção ou a sua falta de. O processamento desta análise resultou em 10 *features* adicionais: Contraste, Homogeneidade, Dessemelhança, *Angular Second Movement* (ASM), Energia, Máximo, Entropia, Correlação, Média e por fim Variância.

No que consta a produtos de Sentinel-2, estes foram adquiridos através do *Copernicus Open Acess Hub* [13]. Estas imagens foram descarregadas automaticamente através de um *script* de *python* que, fornecendo uma data de início e fim, região e percentagem de cobertura de nuvens, descarrega todas os produtos de Sentinel-2 existentes entre as datas especificadas, que contenham a região de interesse e com uma percentagem de cobertura de nuvens dentro dos limites fornecidos. Após a aquisição destes produtos foi necessário avaliar-se se a cobertura de nuvens presente nos produtos adquiridos teria um impacto negativo no processo de classificação ou não, se sim o produto respectivo era removido. Alguns produtos também foram processados com a ferramenta sen2cor que transforma produtos do tipo 1-C do Sentinel 2 em produtos do tipo 2-A. Para produtos deste tipo, numa fase inicial, foram usadas todas as 13 bandas disponíveis.

Após a aquisição destes produtos foram criados então os respectivos índices de vegetação: o NDVI e o GCVI. Estes índices foram criados de forma automatizada utilizando os produtos de Sentinel-2 adquiridos e recorrendo a um *script* de *Python* para o cálculo dos índices.

Após a aquisição dos produtos e a criação dos índices foi necessário recortarem-se as imagens de cada banda de cada produto às áreas de interesse. O recorte foi feito com as extensão dos produtos raster criados para a *ground truth*, com uma resolução espacial de 10 metros obrigando todas as imagens a coincidirem na resolução espacial. Este processo é o mais automatizado possível recorrendo a um *script* de *Python* para o recorte de todas as imagens disponíveis.

3.4 Pré-Classificação

Anteriormente à classificação foi estudada a distribuição das classes nos produtos referentes à *ground truth*.

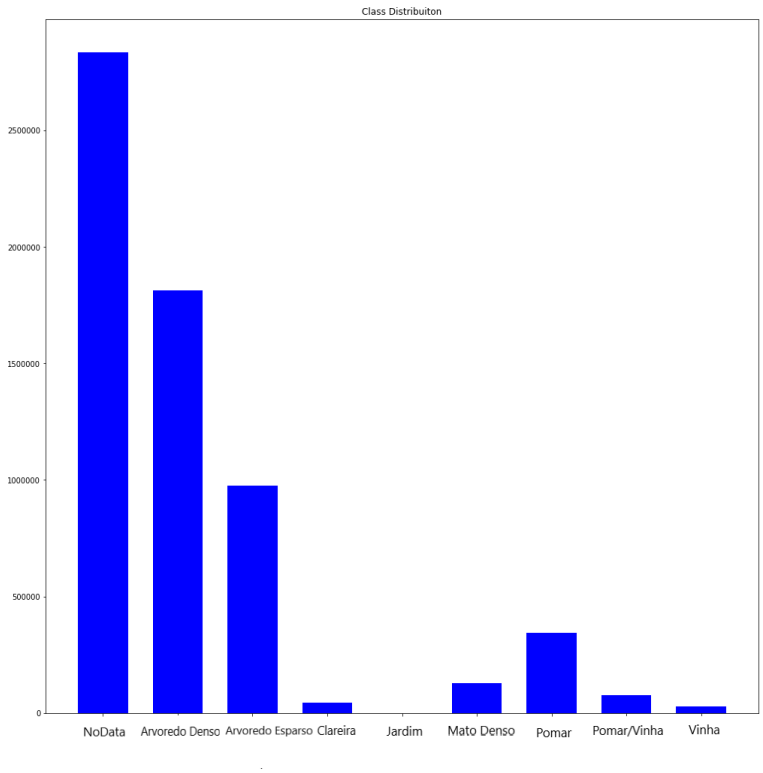

Figura 3.5: Distribuição de Classes Pelas 4 Folhas de Classificação

Na figura 3.5 está representada a distribuição de classes através das quatro regiões de interesse. Como se pode observar a distribuição destas classes é muito heterogénea, com uma larga incidência nas classes de *NoData* (referente a píxeis que, ou não são vegetação, ou são vegetação que sai fora do contexto de classificação do CIGeoE) e também nas classes de Arvoredo Denso e Arvoredo Esparso. Por outro lado as classes Mata, Estufa e Sebe ou Valado que inicialmente estava presente nas classes a classificar, não têm qualquer número de amostras nestas folhas de modo que será ignorada neste trabalho. No entanto, contrariamente ao ponto anterior surge aqui uma classe previamente desconhecida, a Clareira, que, de acordo com as especificações do CIGeoE, representa áreas de vegetação rasteira que se encontram circundadas de Arvoredo. A classe Jardim ou Horta, em mais de quatro milhões de amostras apenas tem representação em cerca de uma centena dessas amostras e, consequentemente, foi também desconsiderada para o algoritmo de classificação.

Classes Principais do CIGeoE	Classes a Utilizar
Sem Representação	No Data
Arvoredo Denso	Arvoredo Denso
Arvoredo Esparso	Arvoredo Esparso
Sem Representação	Clareira
Mata	Sem Representação
Mato Denso ou Arbustos	Mato Denso ou Arbustos
Estufa	Sem Representação
Pomar	Pomar
Vinha	Vinha
Pomar/Vinha	Pomar/Vinha
Sebe ou Valado	Sebe ou Valado
Jardim ou Horta	Representação Muito Diminuta (Descartada)

Tabela 3.1: Classes Principais do CIGeoE e a sua Representação no Algoritmo

A tabela 3.1 mostra qual é a representação das classes principais do CIGeoE nas amostras contidas na *ground truth*. Consequentemente das 10 classes principais da classificação do CIGeoE serão trabalhadas 7 classes com uma classe adicional, sendo esta a Clareira, perfazendo um total de 8 classes a classificar.

3.5 Metodologia Temporal Estática

Numa fase inicial desta dissertação, decidiu-se fazer uma análise temporal estática, para ver quais eram os resultados do algoritmo com apenas um produto de Sentinel-1 e Sentinel-2, da data da classificação (Maio de 2015), como *features*. Para esta metodologia utilizaram-se não só, as bandas disponíveis de Sentinel-1 e Sentinel-2 e os respectivos índices, mas também todas as *features* de texturas calculadas a partir dos produtos de Sentinel-1. Nesta fase, todos os algoritmos de classificação foram utilizados. Cada região de interesse constava com um número de amostras perto de um milhão e meio. O algoritmo de classificação de modo geral segue os seguintes passos:

- **Carregamento e Processamento de Dados** Os dados, tanto as *features* como a *ground truth* são carregados em memória utilizando o GDAL, sendos as imagens carregadas a partir do ficheiro e em seguida transformadas em matrizes que cada valor é representativo de um píxel dessa mesma imagem. No caso das *features* cada entrada na matriz representava todos as bandas associadas ao píxel respectivo disponíveis. Cada imagem é carregada individualmente.

- **Validação dos Modelos** Inicialmente todos os classificadores foram optimizados utilizando o algoritmo de validação de modelos *Random Search*. O algoritmo de *Random Seach* foi implementado com 100 iterações, e utilizando o Cohen Kappa com métrica de validação, foi-lhe também transmitido o conjunto total de dados a

utilizar pois este algoritmo internamente utilizado *Stratified K-Fold* para dividir o conjunto de dados. Após esta optimização, os modelos óptimos de cada classificador são mantidos para os testes posteriores, evitando assim que se tenha de correr a optimização deste algoritmo sempre que seja preciso proceder à classificação de dados tornando o processo de classificação mais rápido e menos pesado.

- **Treino** Posteriormente à obtenção dos parâmetros óptimos para cada classificador, 20% dos dados relativos a cada imagem são utilizados para treino do modelo de classificação. Estes dados estão estratificados, o que significa que o modelo de treino vai receber 20% dos dados de cada classe, mantendo uma distribuição idêntica à do conjunto total, evitando assim a ausência de classes com fraca representação no conjunto de input.

- **Previsão da Imagem** Por fim, todos os dados respectivos a cada imagem são, individualmente, fornecidos ao algoritmo para a previsão da imagem na totalidade. Após a previsão, uma nova imagem raster é gerada com as previsões geradas pelo algoritmo. São também calculadas métricas de validação para perceber melhor o desempenho geral de cada algoritmo.

3.5.1 Estudos Complementares

Foram realizados alguns estudos complementares para perceber qual a influência que estes poderiam ter no desempenho dos classificadores.

- **Variação do tamanho do conjunto de treino** O tamanho do conjunto de treino foi alterado de modo a estudar o desempenho do algoritmo face ao número de dados utilizados no treino. É também interessante porque caso o algoritmo precise de uma percentagem reduzida dos dados de treino isto significa que para futuras utilizações deste algoritmo por parte do CIGeoE, caso não existisse uma folha actualizada para treino, de uma determinada região. Os operadores apenas precisariam de classificar manualmente uma parte diminuta da folha.

- **Oversampling** Foi também estudado o impacto de algoritmos de *oversampling* como o *SMOOTE* e o *ADASYN* [11, 20], no desempenho dos classificadores, visto que a distribuição de classes é altamente desequilibrada, os dados de treino passariam pelo processo de oversampling de modo a existir uma maior representação de classes menos frequentes.

- **Sentinel SCL e *Features* baseadas nas Cartas COS** Na fase temporal estática foi ainda estudado o impacto que *features* adicionais poderiam ter no desempenho dos algoritmos de classificação. Nomeadamente as *features* estudadas foram : a banda SCL de produtos de Sentinel 2-A e uma *feature* criada a partir de cartas COS denominada de *COSBool*. A banda SCL de Sentinel é uma banda existente em

produtos do tipo 2-A, que realize uma pré-classificação das imagens de Sentinel em 9 classes diferentes. Essas classes são: sem dados, saturado ou com defeito, píxel de região escura, sombra de nuvem, vegetação, não vegetação, água, sem classificação, média probabilidade de nuvens, alta probabilidade de nuvens, nuvem *cirrus* fina e neve. Por outro lado a *feature COSBool* é criada utilizando as cartas COS em que classifica os píxeis como vegetação ou não vegetação de acordo com o que as cartas COS classificaram.

3.6 Metodologia de Série Temporal

Apesar da metodologia temporal estática ser muito utilizada na classificação de cobertura terrestre, decidiu-se que seria interessante implementar uma solução que utilizasse mais do que uma imagem de Sentinel para a classificação.

Assim para cada região de interesse foram seleccionados todas os produtos de Sentinel-2 entre a data das imagens e Dezembro de 2016, e de Sentinel-1 foram seleccionados produtos entre as mesmas datas de modo a que existisse no mínimo um produto por mês. Consequentemente os classificadores foram treinados utilizando este novo conjunto de *features* (mantendo a mesma percentagem de dados relegados para treino). No entanto devido ao aumento brusco do tamanho do conjunto de *features*, foi necessário retirarem-se as *features* de textura do conjunto de input, pois este conjunto ficaria demasiado pesado para o algoritmo de classificação neste ambiente de programação. O produto raster resultante da classificação foi gerado de maneira congruente à descrita em 3.5, assim como a aplicação das métricas de avaliação.

Nesta experiência apenas os classificadores RF e XGBoost foram usados pois as SVMs não conseguiram lidar com o aumento do número de *features*.

Para a implementação desta metodologia foram criadas 4 variantes de implementações de séries temporais, construindo assim uma melhor base de comparação, procurando uma implementação ideal com resultados excepcionais que pudesse ser implementada na linha de produção de cartas militares do CIGeoE. Estas variantes serão de em seguida explicadas onde serão expostos os seus benefícios e as suas possíveis limitações

3.6.1 Série Temporal com Métricas Estatísticas

A primeira implementação da metodologia de série temporal, foi a série temporal com métricas estatísticas. Pensou-se inicialmente que esta iria ser a implementação com maior sucesso e por isso foi a primeira variante a ser implementada.

Esta implementação, em vez de usar as *features* fornecidas pelos produtos de Sentinel como input para treino, agrupa as *features* por grupos, grupos estes que são as respectivas bandas ou índices. Deste modo ficamos com um input dividido por bandas e índices de vegetação. Após o agrupamento de *features*, para cada grupo são calculadas 7 métricas estatísticas: o quantil 100, o quantil 75, o quantil 50 (equivalente à mediana), o quantil 25,

o quantil 0, a média e por fim a variância. Após o cálculo destas métricas para cada grupo de *features*, estas métricas são todas agrupadas e utilizadas como input para o treino dos classificadores.

Esta metodologia previne que um número absurdo de *features* seja usado com input tentando resumir as suas características em métricas estatísticas, tentando mascarar *outliers* que possam enviesar o treino.

3.6.2 Série Temporal Convencional

Após a implementação da variante descrita anteriormente, decidiu comparar-se o seu desempenho, não só com o desempenho da implementação estática mas também com uma série temporal que use *features* mais tradicionais para o treino dos classificadores.

Consequentemente, foi implementada uma série temporal que apenas se limita a utilizar todas as *features* obtidas como *input*, à semelhança de uma metodologia estática.

Esta implementação mantém as características dos dados o mais próximo do original possível, contrariamente à implementação em 3.6.1. Deste modo, garante que nenhuma banda ou índice sejam obsoletos e que se mantém um número elevado de dados para se treinar os classificadores.

3.6.3 Série Temporal com Informação Temporal Incorporada

Posteriormente às implementações referidas em 3.6.1 e 3.6.2, pensou-se em enriquecer a informação temporal disponível. Com esse objectivo, utilizou-se a informação temporal disponível dos próprios produtos de Sentinel e fez-se uma tentativa de incorporação desses dados no conjunto de *features* já obtidos.

Desta tentativa surgiram duas variações. Uma que incorpora os dados temporais como um tuplo:

```
(Val_Banda, Data)
```
(3.1)

E outra que incorpora os dados temporais como peso:

```
(Val_Banda * Data)
```
(3.2)

Na primeira variação cada amostra presente nas *features* é então acompanhada pela data em que o produto de Sentinel foi gerado, pretendendo conferir uma presença temporal mais forte a cada *feature*.

No entanto, pensou-se que adicionar o mesmo valor para todas as amostras de uma determinada *feature* era redundante e poderia comprometer o desempenho dos classificadores, desenvolvendo assim uma segunda implementação, que apenas multiplicaria o valor do mês em que o produto de Sentinel foi obtido ao valor da banda, evitando assim repetições do mesmo valor para cada amostra de uma mesma *feature*.

3.7 Validação da Classificação

Para avaliação dos classificadores, foram utilizadas as métricas descritas em 2.4. No entanto, determinou-se que uma implementação destas métricas píxel a píxel seria insuficiente para a quantificação da qualidade do desempenho dos algoritmos de classificação, razão pela qual foram implementadas duas versões adicionais. Uma que avalia o desempenho do classificador com base na maioria de píxeis classificados por polígono da classificação original , a nossa *ground truth*, e outro que implementa estas métricas polígono a polígono.

Por conseguinte, a avaliação da classificação de vegetação permanente segue três implementações diferentes:

- **Píxel a Píxel** Uma implementação convencional das métricas de avaliação em que a unidade básica considerada para cada amostra é o píxel. Para cada métrica a *ground truth* e a classificação obtida são comparadas píxel a píxel, e as métricas são inferidas com base nessa comparação. Esta métrica é fácil e intuitiva de implementar, pois o treino e a previsão são efectuados píxel a píxel, e fornece uma ideia geral do comportamento do nosso algoritmo na classificação da vegetação permanente.

- **Maioria de Píxeis Classificados por Polígono de Classificação Original** Nesta implementação, para cada polígono resultante da classificação original efectuada por operadores do CIGeoE, é realizada uma contagem das classes presentes nos píxeis do resultado da classificação, inseridos na área desse polígono. Após esta contagem, cada píxel do polígono da classificação original, é atribuído à classe maioritária resultante da contagem anterior. A contagem pode ser efectuada de maneira simplista em que cada contagem de um píxel vale um, ou feita com pesos inversamente proporcionais à frequência total de cada classe (à semelhança do parâmetro *balanced* de RF). Esta implementação tem como objectivo perceber se a maioria dos polígonos originais do CIGeoE está bem classificada, pretendendo fornecer uma visão semelhante às métricas *precision* e *recall*.

- **Polígono a Polígono** Como o título indica esta implementação utiliza como unidade básica o polígono. Com esse objectivo é necessário transformar a classificação raster num *shapefile* com uma classificação perceptível (metodologia descrita em 3.10). Como o número de polígonos existentes no *shapefile* original poderá não ser o mesmo que o número de polígonos do *shapefile* da classificação, nem que cada polígono seja linearmente traduzível de um *shapefile* para o outro, é necessário ter um método eficaz de comparação de polígonos. Este método tenta encontrar polígonos entre os dois *shapefiles* que se interceptem significativamente, e cria uma matriz de confusão com base nestas intersecções. Após a geração da matriz de confusão, são calculadas as métricas de avaliação. Mais aprofundadamente para a comparação de dois polígonos ser considerada relevante, estes têm que ter uma intercepção não

nula, caso a intercepção não seja nula, a diferença entre os polígonos responsáveis tem que ser menor do que 10% da área de ambos os polígonos. Deste modo garante-se que caso existem polígonos que se toquem apenas na fronteira mas com áreas muito discrepantes, esta intercepção não irá ser considerada relevante devido à reduzida área de um dos polígonos. Se a diferença se mantiver dentro destes limites, então posteriormente terá de se verificar se algum dos polígonos está contido no outro. Caso ocorra a intercepção apenas será considerada relevante se o polígono não estiver contido num dos buracos do polígono maior, se estiver esta intercepção será descartada. Após a consideração destes parâmetros se a intercepção os respeitar então será considerada uma intercepção válida e será contada para a matriz de confusão. Na figura 3.6 pode-se observar um exemplo simplificado desta avaliação. O polígono original a vermelho aquando a classificação fornecida pelos classificadores é interceptado por 5 polígonos onde as intercepções estão representadas a laranja. Destas 5 intercepções, apenas 3 são válidas pois a diferença entre estes polígonos é inferior a 10% de uma das áreas dos polígonos a serem avaliados (a diferença é representada pela área a amarelo). Os polígonos A, B e E contêm mais de 90% da sua área incluída na área do polígono original, enquanto que, por outro lado, os polígonos D e C constituem uma intercepção não válida pois grande parte da sua área não está incluída na área do polígono original. Esta implementação fornece uma perspectiva a nível do objecto (polígono) do desempenho deste algoritmo.

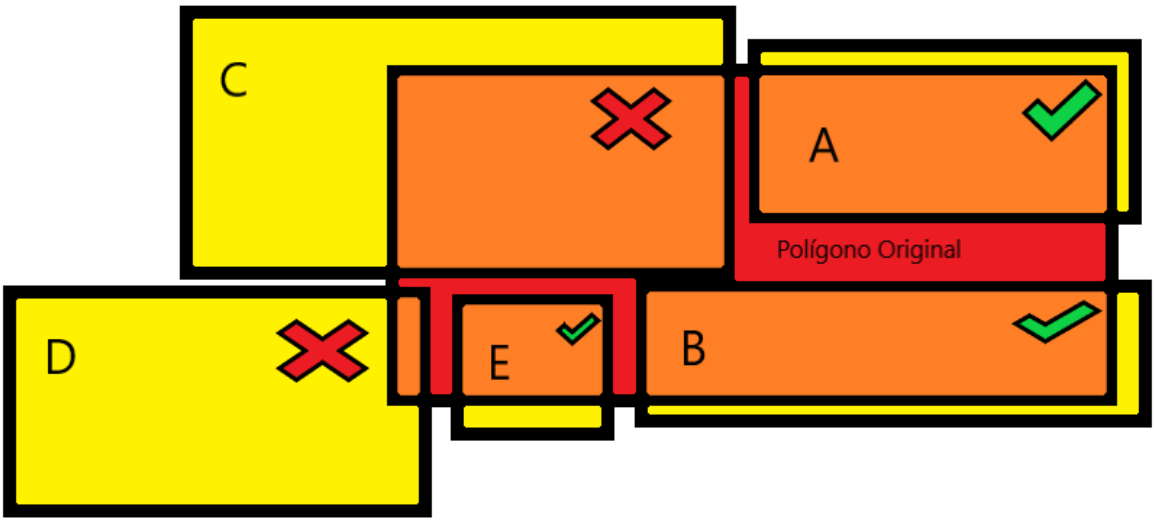

Figura 3.6: Exemplo de Validação Polígono a Polígono

3.8 *Feature Selection*

Com o uso de inúmeras *features* espectrais, principalmente na metodologia temporal estática, a existência de *features* redundantes ou irrelevantes é comum, o que pode ter impactos negativos na classificação. Consequentemente, é necessário remover ou desconsiderar as *features* que poderão ter mais impacto negativo do que positivo. Deste modo, é possível não só melhorar o desempenho do algoritmo, como reduzir o peso do algoritmo.

De modo a suprimir este problema, foi utilizada *feature selection* onde as *features* mais relevantes são seleccionadas descartando as *features* irrelevantes.

A cada *feature* é atribuído um grau de importância, e apenas as *features* com importância superior a um limite irão ser consideradas para o processo de classificação. Este processo de importância de *features* já se encontra incorporado nos classificadores de *Random Forest* e *XGBoost*.

Após a selecção das *features* mais importantes, o classificador é treinado apenas com essas *features* e consequentemente os resultados são comparados com os resultados obtidos com o treino utilizando o número total de *features*.

O objectivo deste processo é encontrar o equilíbrio entre *features* relevantes e irrelevantes de modo a encontrar a combinação de *features* que poderá ser óptima para este processo de classificação.

3.9 Tempo de execução

O tempo foi medido em segundos, e fornece outra perspectiva em termos de qualidade do desempenho de cada classificador. Até que ponto é que melhores resultados justificam um aumento no tempo de treino e de previsão de um classificador. Neste caso, o tempo medido foi dividido em duas fases: a fase de treino e a fase de previsão.

3.10 Pós-Classificação

Visto que o CIGeoE essencialmente trabalha com produtos vectoriais e o objectivo desta dissertação é a apresentação de um produto vectorial que contenha a vegetação permanente classificada, após a classificação, é necessário vetorizar-se/poligonizar-se o ficheiro raster para um *shapefile*. Ademais, dados vectoriais fornecem outras opções de processamento de imagens que poderão ser de uso benéfico para esta dissertação. Para este efeito, recorreu-se à ferramenta *PostGis*

Por isso mesmo, após a obtenção do *shapefile* resultante da classificação de vegetação os seguinte passos de processamento foram executados:

- **Limpeza da Geometria** Limpa pequenas auto-intersecções e aberturas nos polígonos, garantindo que estes não interfiram com as execuções de passos posteriores.

- **Simplificação e Suavização da Geometria dos Polígonos** Os polígonos resultantes da classificação píxel a píxel, apresentam formas grosseiras devido à classificação utilizando uma unidade fundamental mais pequena. Muitas vezes para efeitos cartográficos estas geometrias são desnecessárias sendo por isso necessário um processo de suavização e simplificação das mesmas. Os métodos utilizados para este fim, foram o *ST_SimplifyPreserveTopology* e o *ST_ChaikinSmoothing*. O método *ST_SimplifyPreserveTopology* tenta simplificar a geometria simplificando a informação de cada polígono enquanto que o método *ST_ChaikinSmoothing* tenta suavizar os vértices ou arestas de cada Polígono de modo a contrariar a geometria grosseria resultante de uma classificação píxel a píxel.

- **Limpeza de Polígonos Insignificantes** A classificação píxel a píxel também introduz muito ruído nos produtos resultantes, resultando em inúmeros polígonos de área insignificante no *shapefile*. Estes polígonos são filtrados e limpados do *shapefile* resultante.

Capítulo 4

Experimentação e Análise de Resultados

4.1 Ambiente de Experimentação

Todos os testes realizados foram feitos em ambientes idênticos utilizando Python versão 3.7 da distribuição de *Anaconda*, com o suporte de diversas *packages* adicionais e produtos Sentinel descritos em 3.3. Os dados utilizados têm as seguintes características principais:

- Cerca de 6.400.000 amostras divididas pelas 4 regiões de interesse do Fundão, Monchique, Sendim do Douro e Tocha.

- 8 classes principais sendo estas *No Data*, Arvoredo Esparso, Arvoredo Denso, Clareira, Mato Denso ou Arbustos, Pomar, Pomar/Vinha e Vinha.

A tabela 3.1 mostra qual é a representação das classes principais do CIGeoE nas amostras contidas na *ground truth*. Consequentemente das 10 classes principais da classificação do CIGeoE serão trabalhadas 7 classes com uma classe adicional, sendo esta a Clareira.

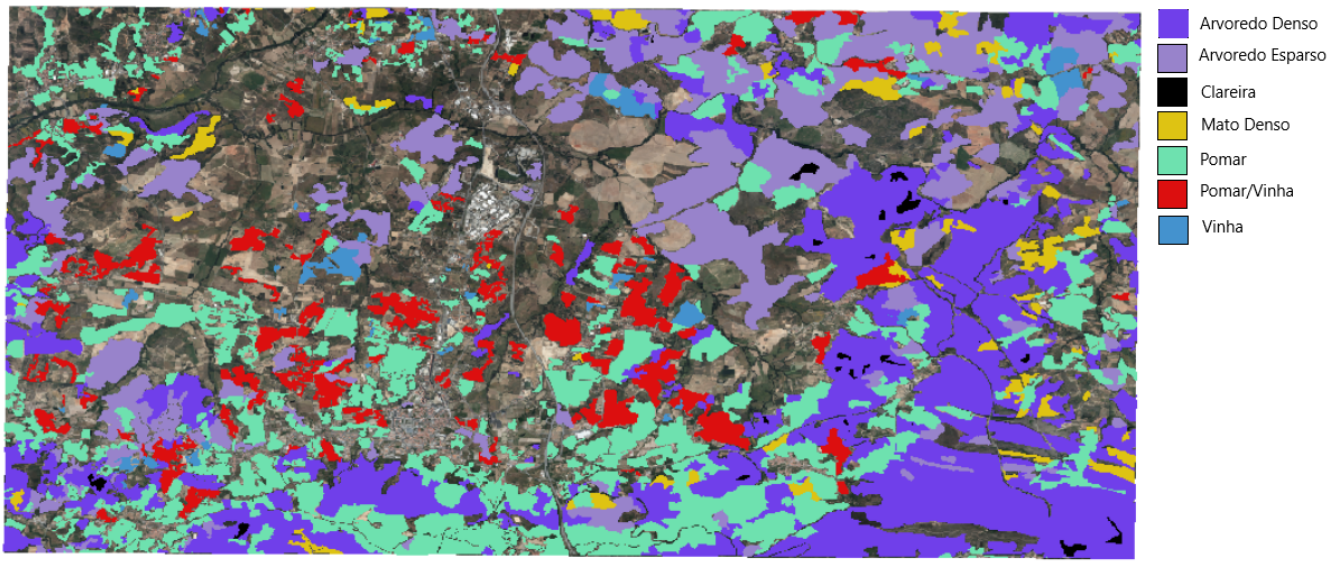

Figura 4.1: Classificação do CiGeoE na zona do Fundão

Na figura 4.1 pode-se observar um exemplo da *groundtruth* que será utilizada na classificação.

4.1.1 Especificações de Hardware

A máquina utilizada tem as seguintes especificações:

- Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz/2.81GHz
- 16 GB de RAM
- Placa Gráfica GeForce GTX 1050 4 GB

4.1.2 Bibliotecas Python

As bibliotecas principais utilizadas foram:

Tuning

sklearn.model_selection.RandomizedSearchCV

Classificação

sklearn.emsemble.RandomForestClassifier

sklearn.svm.LinearSVC

xgboost

sklearn.model_selection.train_test_split

Feature Selection

sklearn.feature_selection.SelectFromModel

Validação

sklearn.metrics.classification_report

sklearn.metrics.accuracy_score

sklearn.metrics.confusion_matrix

sklearn.metrics.cohen_kappa_score

sklearn.metrics.recall_score

sklearn.metrics.precision_score

Geral

numpy

osgeo.gdal

fiona

shapely.geometry.shape

4.2 Discussão de Resultados

Nesta secção, serão discutidos e avaliados os resultados obtidos em todos os ambientes de testes desenvolvidos.

Inicialmente será brevemente discutido quais são os melhores hiper-parâmetros para cada classificador utilizado.

De seguida serão mostrados os resultados para a metodologia temporal estática para cada um dos classificadores, assim como para cada uma das quatro regiões de interesse.

Posteriormente será feita uma análise de todas as variações da metodologia de séries temporais e após a determinação da variação mias apropriada serão mostrados os resultados para esta variação.

Por fim serão analisados os métodos de avaliação baseados em informação vectorial, abrangendo assim a avaliação do desempenho não considerando apenas abordagens píxel a píxel.

4.2.1 Metodologia Temporal Estática

4.2.1.1 Random Forest

Tuning

Para o *tuning* deste classificador consideraram-se os seguintes parâmetros:

- ***n_estimators*** Este parâmetros reflecte o número de árvores a serem utilizados no algoritmo RF.
- ***class_weight*** Peso associado a cada classe. Poderá ser nulo ou então ser ajustado inversamente proporcional à frequência das classes
- ***min_samples_split*** O número de amostras necessário para se dividir um nó da árvore.
- ***min_samples_leaf*** O número de amostras para ser considerado um nó folha.
- ***max_depth*** A profundidade máxima da árvore.
- ***criterion*** A função quantificadora da qualidade de uma divisão.

Após o *tuning* do classificador de *random forest*, o modelo óptimo obteve os seguintes parâmetros:

- ***n_estimators*** = 1000
- ***class_weight*** = "Balanced"
- ***min_samples_split*** = 2
- ***min_samples_leaf*** = 1
- ***max_depth <= min_samples_split***
- ***criterion*** = *gini*

Classificação

Após o *tuning* do classificador e após o seu treino com o número de amostras apropriado procedeu-se à classificação total das regiões de interesse. O classificador RF obteve uma boa classificação com uma *accuracy* média de 87.1% e um valor de *kappa* médio de 83.325% .

Nas tabelas 4.1 e 4.2 estão as métricas de avaliação do classificador RF. Consegue-se discernir que existe alguma confusão significativa entre as classes de vegetação.

Esta confusão incide principalmente entre classes Clareira, Mato Denso ou Arbustos, Pomar, Pomar/Vinha e Vinha com as classes Arvoredo Denso, Arvoredo Esparso e No Data.

Classes	Precision	Recall	F1-Score	Samples
No Data	86.5%	96%	91%	807862
	83.4%	90.0%	87%	641699
	89.3%	96.5%	93%	914704
	95.9%	94.1%	95%	471119
Arvoredo Denso	87.6%	91.1%	89%	285900
	81.8%	92.1%	87%	599266
	82.7%	85%	85%	377754
	89.5%	93.8%	92%	551286
Arvoredo Esparso	91.1%	70.4%	79%	183354
	94.6%	63.5%	76%	311478
	94.1%	66.3%	78%	168528
	90%	86.9 %	88%	311452
Clareira	99.2%	53.7%	70%	4635
	98.3%	55.1%	71%	18616
	99.3%	51.4%	68%	6221
	99.1%	56.2%	72%	15425
Mato Denso ou Arbustos	97.2%	61.1%	75%	29014
	98.5%	56.1%	71%	37160
	96.3%	62.6%	76%	63202
	Na	Na	Na	Na
Pomar	84.9%	79.1%	82%	2231190
	94.7%	64.3 %	77%	22182
	90.7%	71%	80%	88582
	99.8%	56.3 %	95%	914
Pomar/Vinha	95.5%	61.1%	75%	72192
	Na	Na	Na	Na
	98.2%	64%	78%	4702
	96.3%	49.1%	65%	53
Vinha	97%	73.5%	84%	16958
	98.6%	64.9 %	78%	111
	95.9%	67.9%	80%	12067
	Na	Na	Na	Na

Tabela 4.1: Classificação RF nas 4 Regiões de Estudo (Ordem das entradas é Fundão, Monchique, Sendim e Tocha)

Região	Overall Accuracy	Overall Recall	Overall Precision	Overall F1 Score	Overall Kappa
Fundão	87%	73%	92%	81%	81%
Monchique	85%	70%	93%	78%	76%
Sendim	88%	71%	93%	80%	80%
Tocha	92%	73%	88%	81%	88%

Tabela 4.2: Métricas Gerais Resultantes da Classificação RF nas 4 Regiões de Estudo.

As classes de Clareira, Mato Denso, Pomar, Pomar/Vinha e Vinha possuem um número de amostras relativamente reduzido quando comparado ao número de amostras das classes de Arvoredo Denso, Esparso e No Data. Consequentemente, existe uma predisposição para classificar amostras destas classes minoritárias como classes com um maior número de amostras explicando assim o porquê de o *recall* baixar significativamente para estas classes, e a *precision* das classes maioritárias ser inferior devido ao elevado número de falsos positivos. No entanto, é de notar que existe pouca confusão entre classes contextualmente muito semelhantes como o Arvoredo Denso e Esparso e as classes de Pomar, Pomar/Vinha e Vinha o que mostra que o classificador consegue fazer a distinção entre classes muito semelhantes, partindo do princípio que o número de amostras entre elas seja relativamente idêntico. Também é importante referir que no caso da região da Tocha as métricas sobem todas consideravelmente pois existe uma menor diversidade entre as classes de vegetação existindo a falta de classes como Mato Denso e Vinha, aumentando assim a proporção de amostras de classes maioritárias como No Data e as classes de Arvoredo.

No entanto, no geral o classificador consegue atingir resultados aceitáveis onde consegue com uma grande maioria classificar correctamente a maioria dos píxeis de cada classe.

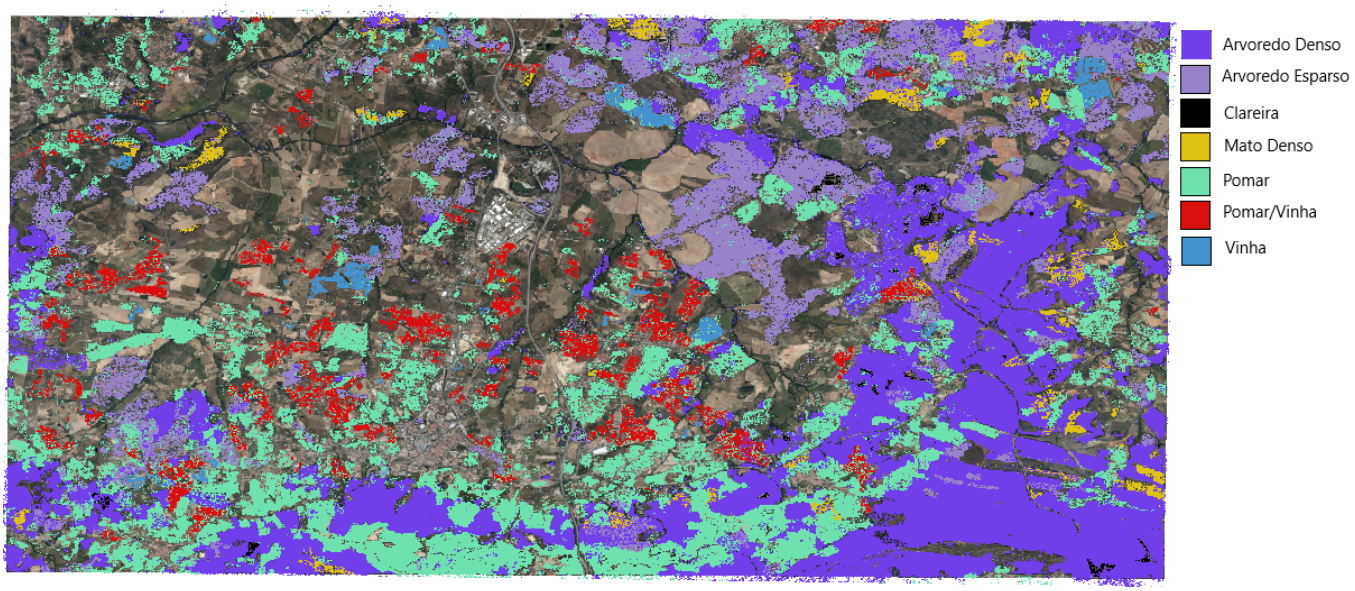

Figura 4.2: Produto *Raster* Resultante da Classificação RF na região do Fundão

Feature Selection

Para a *feature selection* deste classificador, foram primeiramente, calculadas as importâncias de cada *feature* que foi utilizada para a classificação. No caso do classificador RF recorreu-se ao atributo *features_importances_* inerente ao classificador.

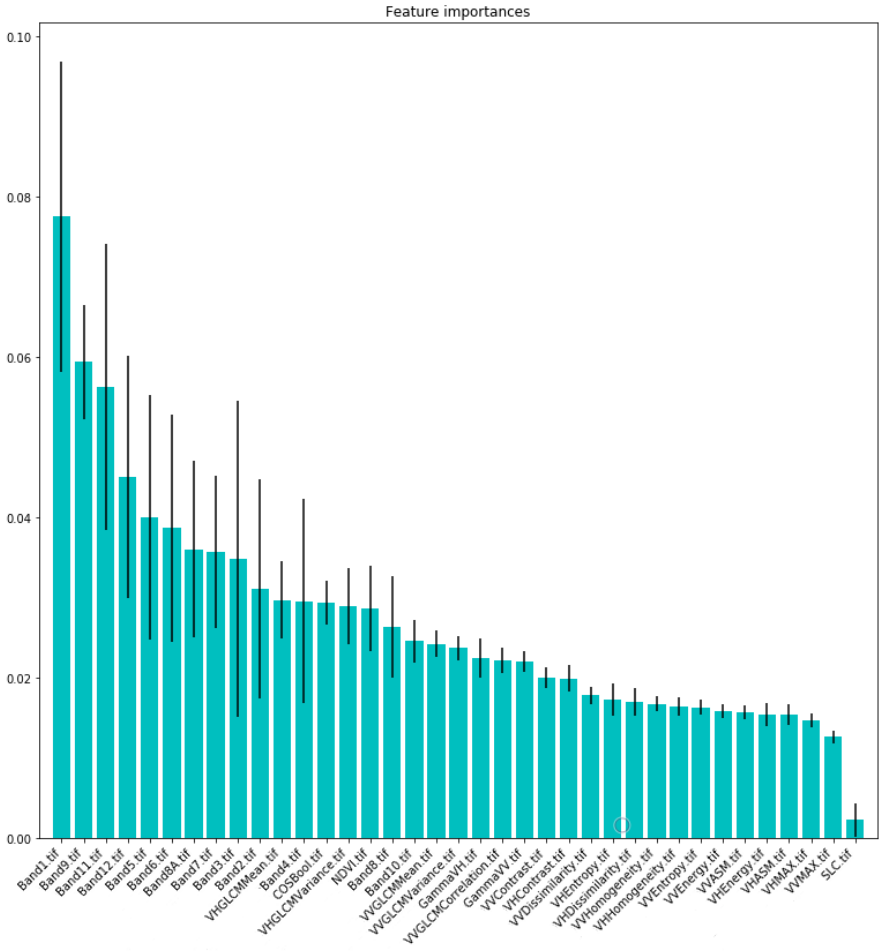

Figura 4.3: Importâncias das *Features* utilizadas pelo classificador RF

Como se pode observar no gráfico presente na figura 4.3 as *features* mais relevantes são as bandas de Sentinel-2. Adicionalmente observa-se que as bandas com menor resolução espacial como as bandas 1 e 9 que se encontram como as *features* com maior importância para este classificador na classificação de vegetação. Por outro lado as bandas com maior resolução (Bandas 2, 3, 4, 8), revelam-se as *features* de Sentinel-2 com menor grau de importância.

As bandas de Sentinel-1 mantêm um grau de importância médio mantendo-se à frente da maioria das *features* de textura mas acabando por ter menor influência na classificação que as bandas de Sentinel-2. No entanto, pode-se inferir que estas constituem uma influência positiva adicional na classificação de vegetação.

As *features* de textura encontram-se na sua grande maioria na parte inferior do espectro de importância de *features*, obtendo todas uma grau de importância muito semelhante e atrás das bandas de Sentinel-1 e 2. No entanto existem algumas excepções no caso das bandas *VHGLCMMean e VHGLCMVariance* que se equiparam a algumas bandas de Sentinel-2 no seu grau de importância.

Por fim as *features* criadas artificialmente mostraram resultados muito diferentes. A

feature COSBool revelou-se uma mais valia nesta classificação ultrapassando algumas bandas de Sentinel-2 e todas as bandas de Sentinel-1. Contrariamente a este comportamento a *feature SLC* apresenta uma importância quase insignificativa sendo a *feature* redundante para esta classificação.

Após a determinação da importância de cada *feature* foi feito o treino e classificação na região do Fundão apenas com as *features* com importância superior à mediana da importância de todas as *features*. Esta selecção reduziu o número de *features* para cerca de metade e apresentou os seguintes resultados:

Classes	Precision	Recall	F1-Score	Samples		
No Data	88%	96%	92%	807862		
Arvoredo Denso	89%	92%	90%	285900		
Arvoredo Esparso	91%	74%	80%	183354		
Clareira	99%	60%	75%	4635	Overall Accuracy	89%
Mato Denso ou Arbustos	97%	65%	78%	29014	Overall Recall	76%
Pomar	86%	81%	84%	2231190	Overall Precision	93%
Pomar/-Vinha	95%	66%	78%	72192	Overall F1-Score	83%
Vinha	97%	78%	87%	16958	Overall Kappa	83%

Tabela 4.3: Métricas resultantes da *Feature Selection* no Algoritmo RF na região do Fundão

Comparando a tabela 4.3 com a tabela presente na figura 4.1 observa-se que se obtém um acréscimo insignificante na *accuracy, precision e kappa* mas um ganho significativo no *recall*. Ao se reduzir o número de *features*, conseguiu-se obter uma classificação com menos confusão entre as classes maioritárias e minoritárias. Devido ao elevado número de *features* e o desequilíbrio acentuado entre classes nos dados obtidos, as classes maioritárias com a ajuda de *features* menos construtivas, estavam a ser demasiado generalizadas acabando por provocar um ruído significativo na classificação geral. Com a ajuda da *feature selection* foi-se possível mitigar de certa forma este problema.

4.2.1.2 XGBoost

Tuning

Para o tuning deste classificador os seguintes parâmetros foram considerados:

- ***n_estimators*** Número de árvores a utilizar neste algoritmo.

- ***min_child_weight*** Soma máxima de peso necessária numa nodo folha.
- ***max_depth*** Profundida máxima de cada árvore.
- ***learning_rate*** Índice de aprendizagem a utilizar.
- ***gamma*** Perda mínima para se particionar um nodo folha.
- ***colsample_by_tree*** Denota a fração de colunas para serem amostras para cada árvore.

Após o *tuning* do classificador de *xgboost*, o modelo ótimo obteve os seguintes parâmetros:

- ***n_estimators*** = 1000
- ***min_child_weight*** = 1
- ***max_depth*** = 12
- ***learning_rate*** = 0.05
- ***gamma*** = 0.1
- ***colsample_by_tree*** = 0.7

Classificação

Após o *tuning* do classificador e após o seu treino com o número de amostras apropriado procedeu-se à classificação total das regiões de interesse. Para a classificação utilizaram-se dois modos diferentes do classificador XGBoost, um que utiliza *multithreading* recorrendo a GPU, e o modo convencional que utiliza *multiprocessing* recorrendo a CPU. O classificador *XGBoost* obteve uma boa classificação com uma *accuracy* média de 86.775% e um valor de *kappa* médio de 79.525% .

Classes	Precision	Recall	F1-Score	Samples
No Data	87%	93%	90%	807862
	81.5%	88.2%	85%	641699
	89.7%	94.7%	92%	914704
	96.8%	93.6%	95%	471119
Arvoredo Denso	87.4%	90.7%	89%	285900
	79%	91%	85%	599266
	80%	87.8%	84%	377754
	90.5%	93.2%	92%	551286
Arvoredo Esparso	83.8%	70.6%	77%	183354
	89.6%	55.5%	69%	311478
	88.3%	65%	77%	168528
	88.7%	89.8%	89%	311452
Clareira	98.9%	66.4%	79%	4635
	98.5%	63.2%	77%	18616
	98.8%	65%	78%	6221
	98.7%	70.5%	82%	15425
Mato Denso ou Arbustos	97.2%	65.6%	78%	29014
	98%	63.6%	77%	37160
	95.2%	64.6%	77%	63202
	Na	Na	Na	Na
Pomar	77.4%	77%	77%	2231190
	95%	68.3%	79%	22182
	86.3%	70%	77%	88582
	98.9%	70.4%	82%	914
Pomar/Vinha	91.4%	59.8%	72%	72192
	Na	Na	Na	Na
	98%	71.6%	83%	4702
	96.4%	50.9%	67%	53
Vinha	96.9%	78%	86%	16958
	96.3%	71.2%	82%	111
	95.2%	72.7%	87%	12067
	Na	Na	Na	Na

Tabela 4.4: Classificação *XGBoost* nas 4 Regiões de Estudo (Ordem das entradas é Fundão, Monchique, Sendim e Tocha)

Região	Overall Accuracy	Overall Recall	Overall Precision	Overall F1 Score	Overall Kappa
Fundão	86%	75%	90%	81%	79%
Monchique	82%	72%	91%	79%	72%
Sendim	87%	74%	91%	81%	79%
Tocha	92%	78%	95%	85%	88%

Tabela 4.5: Métricas Gerais Resultantes da Classificação *XGBoost* nas 4 Regiões de Estudo.

Com a matrizes de confusão e métricas gerais representadas nas tabelas 4.4 e 4.5

pode-se inferir que o classificador *XGBoost* tem um desempenho muito semelhante ao do classificador RF. Repete-se a confusão entre classes minoritárias e classes minoritárias. No entanto é importante referir que o classificador *XGBoost* apresenta maior tolerância ao desequilíbrio entre as classes apresentando menor confusão entre classes maioritárias e classes minoritárias e por isso, apesar de no geral as métricas serem inferiores, este classificador apresenta um *recall* ligeiramente melhor que o classificador RF (e por conseguinte em alguns casos o F1-Score). Será interessante estudar este comportamento aquando a utilização da metodologia de série temporal pois o número de *features* será superior. Esta característica é salientada na região da Tocha onde a amplitude de amostras entre classes maioritárias e minoritárias é superior, e onde este classificador supera o algoritmo RF de uma maneira geral em todas as métricas.

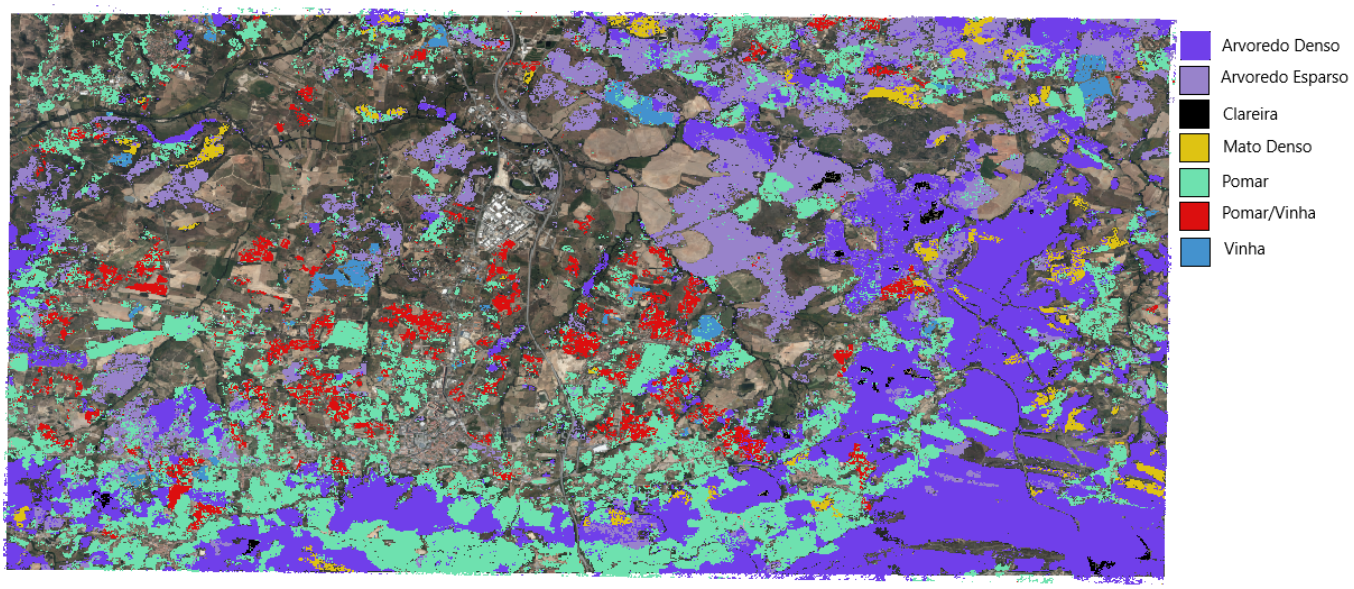

Figura 4.4: Produto *Raster* Resultante da Classificação *XGBoost* na região do Fundão

Feature Selection

Na *feature selection* deste classificador utilizou-se o método inerente ao *XGBoost* para determinar a importância das *features* utilizadas. As importâncias foram calculadas directamente através do método *plot_importance* proveniente do *XGBoost*.

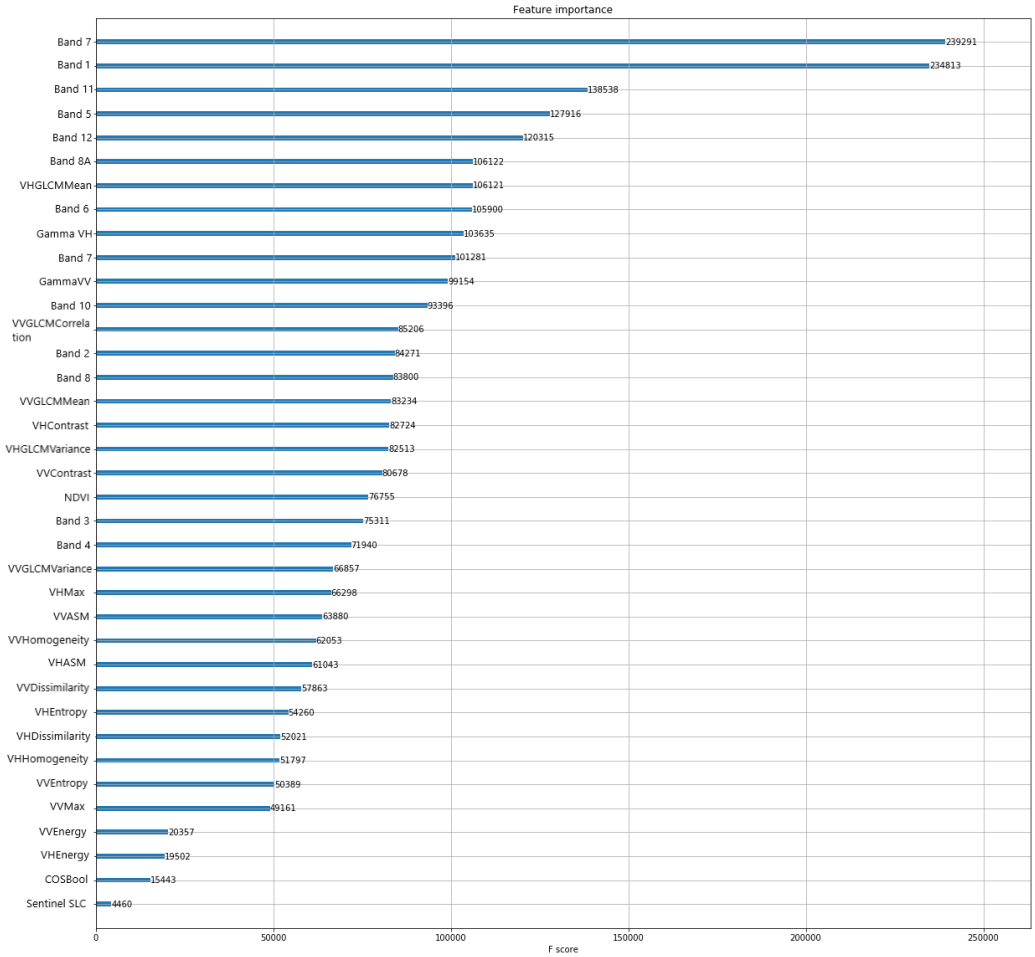

Figura 4.5: Importâncias das *Features* utilizadas pelo classificador *XGBoost*

Como se pode observar na figura 4.5 o classificador *XGBoost* é mais influenciado pelas *features* adquiridas do Sentinel-2. Curiosamente nenhuma destas *features* mais importantes do Sentinel-2 constituem as bandas de radiação visível do verde azul e vermelho. Adicionalmente, as bandas de 10 metros obtiveram o menor grau de importância o que é surpreendente dada a resolução superior destas bandas.

Relativamente às *features* relacionadas com as texturas, a grande maioria obteve um grau de importância semelhante às bandas de 10 metros de Sentinel-2, sendo largamente ultrapassadas pelas *features* de Sentinel-2 de bandas de 20 e 60 metros. Apenas a *features VHGLCMMean* obteve um grau de importância semelhante às bandas de Sentinel-2.

As bandas de Sentinel-1 *GammaVH e GammaVV* conseguiram manter um grau de importância superior ou muito semelhante a 10000 mostrando que a utilização adicional do Sentinel-1 na classificação de vegetação tem uma influência positiva.

As bandas artificialmente criadas do *COSBool* e Sentinel SLC revelam-se um detrimento para a classificação de vegetação, sendo as duas bandas com o menor grau de importância.

Após o estudo da importância das *features* foi então feita a *feature selection* com todas

as *features* acima da mediana do conjunto total. Os resultados foram os seguintes:

Classes	Precision	Recall	F1-Score	Samples		
No Data	87%	93%	90%	807862		
Arvoredo Denso	87%	90%	89%	285900		
Arvoredo Esparso	83%	70%	76%	183354		
Clareira	98%	70%	81%	4635	Overall Accuracy	85%
Mato Denso ou Arbustos	97%	66%	77%	29014	Overall Recall	75%
Pomar	77%	77%	77%	2231190	Overall Precision	89%
Pomar/-Vinha	90%	57%	70%	72192	Overall F1-Score	82%
Vinha	96%	80%	87%	16958	Overall Kappa	78%

Tabela 4.6: Métricas resultantes da *Feature Selection* no Algoritmo *XGBoost* na região do Fundão

Consultando a tabela 4.6, conclui-se que a *feature selection* não tem tanto impacto na classificação como no classificador RF. As métricas mantêm-se praticamente inalteradas verificando-se uma diminuição redundante na *precision e accuracy*. No entanto no *recall* a distinção das classes Vinha e Clareira sobe enquanto a classe Pomar/Vinha é alvo de maior confusão. Estes resultados indicam que o classificador *XGBoost* consegue ser menos influenciado por *features* menos benéficas do que o classificador RF, o que significa que não beneficia tanto de *feature selection*. Não obstante, a utilização de menos *features* neste classificador reduz o tempo de treino significativamente confirmando a sua utilidade para este algoritmo apesar da preservação das suas métricas.

4.2.1.3 SVM

Tuning

Para o tuning deste classificador os seguintes parâmetros foram considerados:

- ***loss*** A função de perda utilizada.
- ***C*** Parâmetro de penálti do termo de erro.
- ***tol*** Tolerância para paragem.
- ***multi_class*** Parâmetro que determina a estratégia de multi-classe.

- ***dual*** Opção que indica ao algoritmo para resolver os problemas de otimização primais ou duais.

Após o *tuning* do classificador SVM, o modelo óptimo obteve os seguintes parâmetros:

Classificação

- ***loss*** = squared_hinge.

- ***C*** = 0.1.

- ***tol*** = $1x10^{-07}$

- ***multi_class*** = ovr

- ***dual*** = False

Após o *tuning* do classificador procedeu-se à classificação das folhas obtidas. O classificador SVM obteve resultados muito pobres com métricas muito más com uma *accuracy* média de 65.615% e um valor de *kappa* médio de 42.988% .

Classes	Precision	Recall	F1-Score	Samples
No Data	68.5%	88.6%	77%	807862
	56.9%	77.3%	66%	641699
	73.1%	91.6%	81%	914704
	84.7%	86.1%	85%	471119
Arvoredo Denso	62%	80.1%	70%	285900
	58.5%	73.7%	65%	599266
	54.1%	67.5%	60%	377754
	67.3%	84.2%	75%	551286
Arvoredo Esparso	50.1%	16%	24%	183354
	42.9%	0.5%	0.01%	311478
	44.4%	0.2%	0%	168528
	63.4%	37%	47%	311452
Clareira	0%	0%	0%	4635
	14.3%	0%	0%	18616
	0%	0%	0%	6221
	0%	0%	0%	15425
Mato Denso ou Arbustos	0%	0%	0%	29014
	0%	0%	0%	37160
	0%	0%	0%	63202
	Na	Na	Na	Na
Pomar	49.4%	33.6%	40%	2231190
	50%	0%	0%	22182
	54.2%	10.3%	17%	88582
	0%	0%	0%	914
Pomar/Vinha	0%	0%	0%	72192
	Na	Na	Na	Na
	0%	0%	0%	4702
	0%	0%	0%	53
Vinha	0%	0%	0%	16958
	0%	0%	0%	111
	0%	0%	0%	12067
	Na	Na	Na	Na

Tabela 4.7: Classificação SVM nas 4 Regiões de Estudo (Ordem das entradas é Fundão, Monchique, Sendim e Tocha)

Região	Overall Accuracy	Overall Recall	Overall Precision	Overall F1 Score	Overall Kappa
Fundão	64%	27%	29%	26%	43%
Monchique	58%	22%	32%	19%	32%
Sendim	67%	21%	28%	20%	40%
Tocha	73%	35%	36%	34%	58%

Tabela 4.8: Métricas Gerais Resultantes da Classificação SVM nas 4 Regiões de Estudo.

A matriz de confusão e métricas gerias representadas nas tabelas 4.7 e 4.8 revelam

que este classificador tem imenso problemas em distinguir classes com menos amostras de classes com mais amostras. As classes minoritárias nem sequer têm amostras previstas revelando que este classificador não é capaz de lidar com a disparidade entre o número de amostras de classes diferentes. Apenas as classes com um número considerável de amostras é que possui amostras previstas na classificação final resultando numa classificação muito pobre em que se verifica apenas a presença de 3 ou 4 classes, tornando todas as restantes classes obsoletas.

Algoritmo	Treino(secs/mins)	Classificação(secs)
RF	1229.137/20.486	293.546
XGBoost (CPU)	1308.625/21.810	469.285
XGBoost (GPU)	831.776/13.863	469.285
SVM	1730.335/28.84	0.23785

Tabela 4.9: Tempos de Execução dos Algoritmos para a Metodologia Estática

Olhando para a tabela 4.9 consegue aferir-se que o classificador que se destaca pelo sua rapidez de previsão é o classificador *XGBoost* recorrendo a *multithreading* com GPU. Apesar disso os classificadores RF e *XGBoost* utilizando CPU apresentam tempos muito aceitáveis. Por outro lado respectivamente ao classificador RF nota-se que o seu tempo de previsão é quase metade do tempo de previsão do classificador *XGBoost* (GPU), que quando aliado ao seu tempo de treino, perfaz um tempo 3 minutos mais lento na execução total do que o classificador *XGBoost* usando GPU. Consequentemente qualquer um destes classificadores se revela adequado para a classificação de vegetação utilizando uma abordagem temporal estática. O classificador SVM além dos seus fracos resultados possui o maior tempo de execução, confirmando as expectativas de que este classificador não é adequado para a classificação com um número tão elevado de amostras.

É importante referir que a classe No Data é uma classe muito abrangente que engloba não só píxeis referentes a não-vegetação como píxeis referentes a vegetação não utilizada no contexto de classificação do CIGeoE. Por isto mesmo é introduzido ruído na *ground truth* que poderá ser prejudicial para a classificação.

Outra observação importante é que a *groud truth* é de Maio de 2015 e só existem imagens de Sentinel-2 a partir de Julho de 2015, outra causa de inconsistência entre *features* e *ground truth* que poderá ter influência no desempenho da classificação.

Como tentativa de combate a estas inconsistências optou-se por se estudar também as implementações de séries temporais.

4.2.2 Metodologia de Série Temporal

Como foi descrito na secção 3.6 foram implementadas várias variações da metodologia de série temporal de modo a aproveitar o máximo de informação possível ganha com a utilização de múltiplas imagens com *features*.

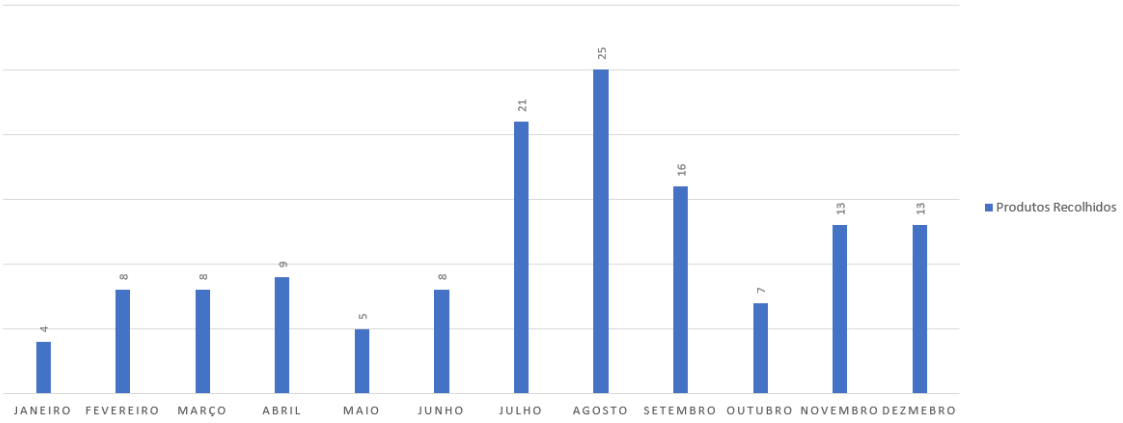

Figura 4.6: Frequência Absoluta de Produtos Utilizados por Mês

Como se pode observar na figura 4.6 a distribuição dos produtos ao longo do ano não é uniforme. Isto é devido não só, à amplitude temporal escolhida para a recolha de produtos (Julho de 2015 até Dezembro de 2016) onde existem meses alvo de recolha de produtos de 2015 e 2016, mas também devido à cobertura de nuvens mais proeminente nalguns meses do que outros.

Consequentemente pode-se observar que a grande maioria dos produtos de Sentinel provêm apenas dos meses de Julho e Agosto, contrariamente a meses como Janeiro que somente detêm produtos de Sentinel-1 (1 por região de interesse). Esta particularidade poderá ter impacto na classificação e por isso procurou-se incorporar dados temporais, neste caso mensais, nas *features* que irão ser utilizadas nas séries temporais.

Estes modelos foram testados utilizando o classificador RF na região do Fundão, pois esta região é a que apresenta maior heterogeneidade de vegetação, existindo uma diferença menor entre as classes maioritárias e as classes minoritárias e, deste modo, revelando-se região de interesse mais apta a ser utilizada para o a validação dos diferentes modelos de séries temporais.

Métricas	S.T. Convencional	S.T. com Métricas Estatísticas	S.T. Informação Temporal como Tuplo	S.T. Informação Temporal como Peso
Accuracy	94%	93%	94%	94%
Precision	97%	96%	96%	97%
Recall	86%	83%	85%	87%
F1Score	91%	89%	90%	91%
Kappa	91%	88%	90%	91%

Tabela 4.10: Validação dos Diferentes Modelos de Série Temporal

Como se pode observar na tabela 4.10 todas estas abordagens conferem uma melhoria significativa na classificação de vegetação, no entanto nota-se que o modelo que usa métricas artificias tem um desempenho notavelmente inferior ao desempenho dos outros

modelos. Este comportamento pode ser devido ao uso exclusivo de métricas estatísticas que apesar, de poderem conferir alguma tolerância a *outliers*, o algoritmo perde informação vital contida nas *features* originais e não contem informação temporal nenhuma. Por outro lado todos os outros modelos obtiveram resultados muito semelhantes com o modelo de série temporal com informação temporal como peso a ter uma ligeira vantagem sobre os outros. Isto deve-se á informação temporal adicional que contem em relação ao modelo de série temporal convencional e por outro lado esta informação não está armazenada de forma redundante como o modelo de série temporal com informação temporal em tuplo em que cada píxel do mesmo produto contem a mesma informação temporal armazenada. Pode-se entaõ aferir que o modelo com melhores resultados é o modelo de s´rei temporal com informação temporal como peso.

4.2.2.1 Random Forest

Com a utilização de série temporal o classificador RF obteve uma melhoria muito significativa. A *accuracy* média subiu 94.6% para e o Kappa para 92-05%

Classes	Precision	Recall	F1-Score	Samples
No Data	91.6%	98.1%	95%	807862
	93.3%	96.8%	95%	641699
	92.3%	98.2%	95%	914704
	98.2%	97.1%	98%	471119
Arvoredo Denso	93.5%	95.4%	94%	285900
	94.1%	96.7%	95%	599266
	92.7%	91.5%	92%	377754
	97.2%	98.5%	98%	551286
Arvoredo Esparso	96.5%	86.6%	91%	183354
	97.1%	87.9%	92%	311478
	97.1%	83.2%	90%	168528
	98.4%	98.5%	98%	311452
Clareira	98%	77.3%	86%	4635
	98.5%	82%	90%	18616
	99.1%	76.3%	87%	6221
	98.1%	86.6%	92%	15425
Mato Denso ou Arbustos	98.3%	79.8%	88%	29014
	97.9%	86.1%	92%	37160
	98%	83.2%	90%	63202
	Na	Na	Na	Na
Pomar	95.9%	87%	91%	2231190
	96.3%	82.2%	89%	22182
	97.3%	80.4%	88%	88582
	98.6%	83.5%	90%	914
Pomar/Vinha	98.9%	78.9%	87%	72192
	Na	Na	Na	Na
	99%	79%	88%	4702
	97.6%	77.4%	85%	53
Vinha	98.4%	89%	94%	16958
	93.5%	90.1%	92%	111
	98%	81.5%	89%	12067
	Na	Na	Na	Na

Tabela 4.11: Classificação RF com séries temporais nas 4 Regiões de Estudo (Ordem das entradas é Fundão, Monchique, Sendim e Tocha)

Região	Overall Accuracy	Overall Recall	Overall Precision	Overall F1 Score	Overall Kappa
Fundão	93%	87%	96%	91%	90%
Monchique	94%	89%	96%	92%	92%
Sendim	93%	84%	97%	90%	89%
Tocha	98%	90%	98%	94%	97%

Tabela 4.12: Métricas Gerais Resultantes da Classificação RF com séries temporais nas 4 Regiões de Estudo.

Como se pode observar nas tabelas 4.11 e 4.12 a utilização de série temporal tem um impacto muito positivo na classificação de vegetação pelo algoritmo RF, a confusão entre classes minoritárias e maioritárias, apesar de presente, atenua-se com o aumento de *features* utilizadas para treino e classificação. O maior aumento nas métricas verifica-se no *recall* onde este em média subiu em média 16.75% o que revela uma melhoria muito significativa no desempenho deste algoritmo. No entanto este classificador continua a ter alguma dificuldade na distinção entre classes maioritárias e classes minoritárias revelando que o desequilíbrio entre classes continua a ser o maior obstáculo deste classificador.

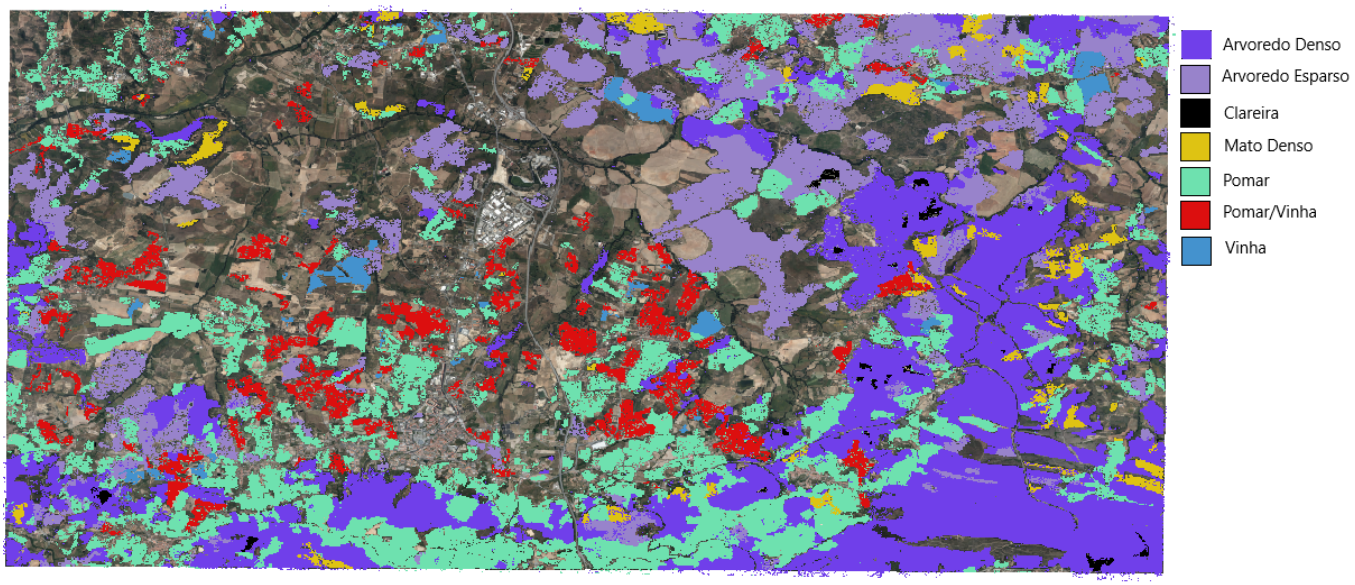

Figura 4.7: Produto *Raster* Resultante da Classificação RF com Metodologia Timeseries

Feature Selection

Devido à melhoria dos resultados na fase estática de classificação provocada pela utilização de *feature selection*, decidiu-se implementá-la na metodologia sde série temporal.

Classes	Precision	Recall	F1-Score	Samples		
No Data	94%	98%	96%	807862		
Arvoredo Denso	95%	97%	96%	285900		
Arvoredo Esparso	97%	91%	94%	183354		
Clareira	98%	82%	89%	4635	Overall Accuracy	95%
Mato Denso ou Arbustos	98%	86%	91%	29014	Overall Recall	90%
Pomar	96%	91%	94%	2231190	Overall Precision	97%
Pomar/-Vinha	97%	86%	91%	72192	Overall F1-Score	95%
Vinha	98%	92%	95%	16958	Overall Kappa	93%

Tabela 4.13: Métricas resultantes da *Feature Selection* no Algoritmo RF com séries temporais na região do Fundão

Como se pode ver na tabela 4.13 com a utilização de *feature selection* o algoritmo de RF tem uma melhoria muito significativa na classificação de vegetação. Notavelmente, e à semelhança do comportamento na metodologia estática, a métrica *recall* tem o maior aumento ultrapassando os 90%. Por outro lado o *Kappa* também tem uma subida notável. A confusão nas classes minoritárias desce consideravelmente. A utilização da *feature selection* revela-se um método muito eficaz de melhorar o desempenho do classificador RF, não só em metodologias estáticas como de série temporal.

4.2.2.2 XGBoost

Para a abordagem baseada em séries temporais, mantiveram-se os mesmos parâmetros para o classificador *XGBoost*. Este classificador foi treinado outra vez com recurso a *multithreading* utilizando GPU e *multiprocessing* recorrendo a CPU. Neste classificador foi onde se revelou a maior melhoria no desempenho.

Classes	Precision	Recall	F1-Score	Samples
No Data	97%	98.3%	98%	807862
	97.1%	97.8%	97%	641699
	96.7%	98.5%	98%	914704
	98.8%	98%	98%	471119
Arvoredo Denso	97.3%	97.7%	98%	285900
	97.1%	97.8%	97%	599266
	96.4%	95.5%	96%	377754
	98.4%	98.9%	98%	551286
Arvoredo Esparso	97.5%	96.5%	97%	183354
	97%	95.7%	96%	311478
	97.2%	95.4%	96%	168528
	99%	99.5%	99%	311452
Clareira	96.9%	90.5%	94%	4635
	97.5%	93.9%	96%	18616
	97.5%	93.2%	95%	6221
	97.6%	93.1%	95%	15425
Mato Denso ou Arbustos	98%	92.1%	95%	29014
	97.8%	93.9%	96%	37160
	97.8%	93.9%	96%	63202
	Na	Na	Na	Na
Pomar	96.4%	94.9%	96%	2231190
	97.1%	86.4%	91%	22182
	96.3%	89.7%	93%	88582
	97.7%	91.2%	95%	914
Pomar/Vinha	97.2%	92.8%	95%	72192
	Na	Na	Na	Na
	98.3%	89.7%	94%	4702
	100%	83%	90%	53
Vinha	98.3%	93.9%	96%	16958
	96.2%	90.1%	93%	111
	98.2%	87.3%	92%	12067
	Na	Na	Na	Na

Tabela 4.14: Classificação *XGBoost* com séries temporais nas 4 Regiões de Estudo (Ordem das entradas é Fundão, Monchique, Sendim e Tocha)

Região	Overall Accuracy	Overall Recall	Overall Precision	Overall F1 Score	Overall Kappa
Fundão	93%	87%	96%	91%	90%
Monchique	94%	89%	96%	92%	92%
Sendim	93%	84%	97%	90%	89%
Tocha	98%	90%	98%	94%	97%

Tabela 4.15: Métricas Gerais Resultantes da Classificação RF com séries temporais nas 4 Regiões de Estudo.

Com a introdução de *features* de uma série temporal o classificador *XGBoost* tem um desempenho excelente, com uma melhoria gigantesca. Como está demonstrado nas tabelas 4.14 e 4.15, com séries temporais este classificador ultrapassa distintamente o classificador RF e apresenta uma confusão muito diminuta entre as classes. Como foi dito na secção 4.2.1.2 o classificador *Xgboost* detinha uma maior capacidade de distinguir classes minoritárias de classes maioritárias do que os outros classificadores, e com o acréscimo do número de *features*, este classificador consegue eficientemente distinguir as classes apresentando muito pouca confusão entre elas apresentando em grande maioria dos casos, *precision e recall* superiores a 90%.

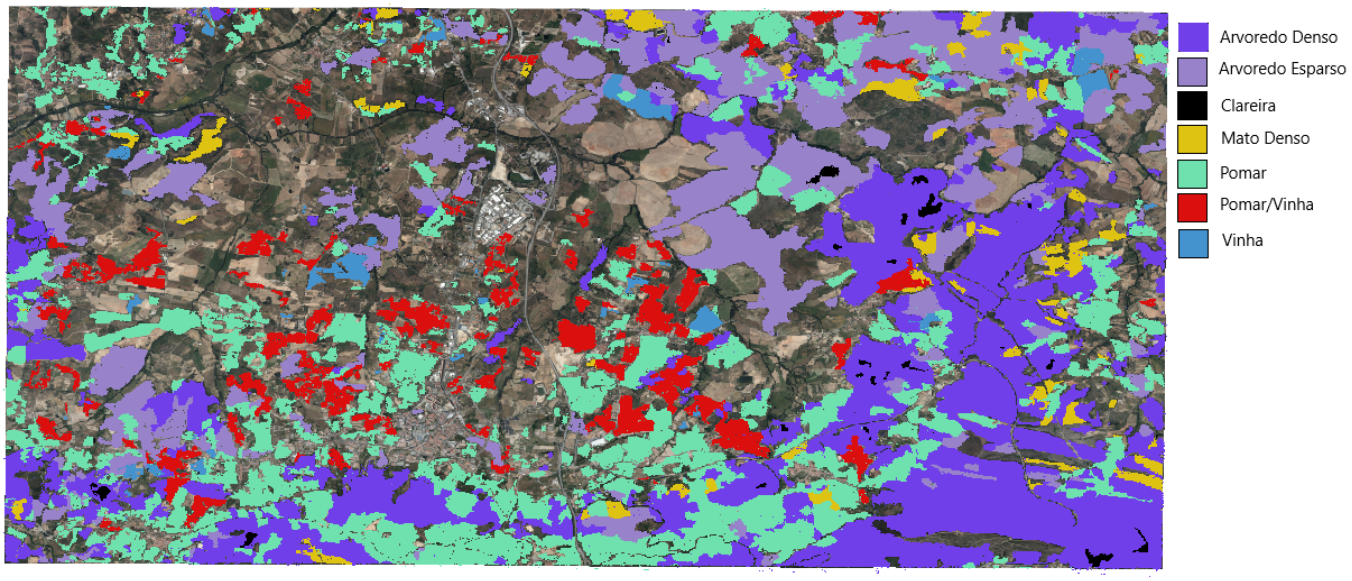

Figura 4.8: Produto *Raster* Resultante da Classificação *XGBoost* com metodologia timeseries

Feature Selection

Após o excelente desempenho deste algoritmo utilizando séries temporais decidiu-se implementar *feature selection* na execução deste algoritmo. Apesar desta adição não ter apresentado nenhuma melhoria significativa na implementação do algoritmo *XGBoost* numa metodologia estática, a utilização de um número superior de *features* pode vir a mudar o comportamento desta adição.

Classes	Precision	Recall	F1-Score	Samples		
No Data	97%	98%	98%	807862		
Arvoredo Denso	97%	98%	97%	285900		
Arvoredo Esparso	97%	97%	97%	183354		
Clareira	96%	91%	94%	4635	Overall Accuracy	97%
Mato Denso ou Arbustos	98%	93%	95%	29014	Overall Recall	95%
Pomar	96%	95%	96%	2231190	Overall Precision	97%
Pomar/-Vinha	97%	93%	95%	72192	Overall F1-Score	97%
Vinha	98%	95%	96%	16958	Overall Kappa	96%

Tabela 4.16: Métricas resultantes da *Feature Selection* no Algoritmo *XGBoost* com séries temporais na região do Fundão

A tabela 4.16 mostra que, à semelhança da sua implementação estática, a implementação de séries temporais não tem nenhuma melhoria significativa. O valor do *F1-Score* sobe ligeiramente, no entanto não contribui nenhuma melhoria significativa para a classificação geral. No entanto a utilização de menos *features* ajuda a encurtar o tempo de treino para cerca de um quarto do tempo original de treino.

Através da introdução de séries temporais foi possível amenizar-se os problemas referentes ao desequilíbrio nas amostras fornecidas. Ao se utilizar diversas imagens impede-se assim que os classificadores estejam sujeitos a interferências de *outliers* assim como a influência previamente falada de apenas se ter imagens de Sentinel-1 da data da classificação do CIGeoE e as imagens de Sentinel-2 serem de dois meses posteriores. Percebe-se assim que a utilização de séries temporais será muito benéfica para a classificação de vegetação permanente.

Algoritmo	Treino(secs/mins)	Classificação(secs)
RF	3459.03/57.017	380.60
XGBoost (CPU)	16136.97/268.95	520
XGBoost (GPU)	5316.17 / 88.6	520

Tabela 4.17: Tempos de Execução dos Algoritmos para a Metodologia Estática

Apesar do classificador *XGBoost* obter resultados notavelmente melhores que o classificador RF, o tempo de execução deste algoritmo é claramente superior ao classificador RF mesmo usando GPU com *multi-threading*. Consequentemente será relevante analisar

se o acréscimo no desempenho de classificação valerá a pena o aumento de tempo de execução (cerca de 3 horas para *XGBoost* CPU e 20 minutos para XGBoost GPU). Para futuras utilizações poderá ser mais benéfico utilizar-se o classificador de RF para uma classificação rápida mas consistente, do que a obtenção de resultados excelentes mas com tempos de execução possivelmente demasiado elevados. Adicionalmente testes feitos com apenas 100 árvores no classififcador *Random Forest* apresentam os mesmo resultados, com um tempo 10 vezes inferior (cerca de 5 min). No entanto conclui-se então que para resultados óptimos a utilização do algoritmo de XGBoost recorrendo a GPU é com uma grande margem o classificador a utilizar.

4.3 Validação Vetorial

Na validação vectorial serão analisados os resultados das duas abordagens referidas em 3.7 nomeadamente a abordagem por maioria e a abordagem por polígono.

Estas abordagens foram realizadas com os produtos resultantes da classificação utilizando *features* de série temporal utilizando o modelo com os melhores resultados (informação temporal como peso), na zona do Fundão. Será avaliada a classificação dos classificadores RF e *XGBoost*.

4.3.1 Validação por Maioria

Classes	Precision	Recall	F1-Score	Samples		
No Data	98%	91.8%	95%	807862		
Arvoredo Denso	95.5%	93.4%	94%	285900		
Arvoredo Esparso	86.7%	96.5%	91%	183354		
Clareira	77.1%	98.1%	86%	4635	Overall Accuracy	93%
Mato Denso ou Arbustos	79.8%	98.3%	88%	29014	Overall Recall	96%
Pomar	87.2%	95.8%	91%	2231190	Overall Precision	87%
Pomar/-Vinha	79.2%	97.5%	87%	72192	Overall F1-Score	91%
Vinha	89.6%	98.2%	94%	16958	Overall Kappa	90%

Tabela 4.18: Classificação RF com séries temporais na região de Fundão por maioria com contagem simples

Classes	Precision	Recall	F1-Score	Samples		
No Data	98.1%	91.7%	95%	807862		
Arvoredo Denso	95.5%	93.3%	94%	285900		
Arvoredo Esparso	86.7%	96.5%	91%	183354		
Clareira	76.9%	98.4%	86%	4635	Overall Accuracy	93%
Mato Denso ou Arbustos	79.7%	98.4%	88%	29014	Overall Recall	96%
Pomar	87.1%	95.8%	91%	2231190	Overall Precision	86%
Pomar/-Vinha	79%	97.9%	87%	72192	Overall F1-Score	91%
Vinha	85.9%	98.8%	92%	16958	Overall Kappa	90%

Tabela 4.19: Classificação RF com séries temporais na região de Fundão por maioria com contagem por peso de classe

Classes	Precision	Recall	F1-Score	Samples		
No Data	98.3%	97%	98%	807862		
Arvoredo Denso	97.7%	97.3%	97%	285900		
Arvoredo Esparso	96.6%	97.5%	97%	183354		
Clareira	90.7%	96.7%	94%	4635	Overall Accuracy	97%
Mato Denso ou Arbustos	92.2%	98%	95%	29014	Overall Recall	97%
Pomar	94.9%	96.4%	96%	2231190	Overall Precision	95%
Pomar/-Vinha	92.8%	97.2%	95%	72192	Overall F1-Score	96%
Vinha	94%	98.2%	96%	16958	Overall Kappa	96%

Tabela 4.20: Classificação *XGBoost* com séries temporais na região de Fundão por maioria com contagem simples

Classes	Precision	Recall	F1-Score	Samples		
No Data	98.3%	97%	98%	807862		
Arvoredo Denso	97.7%	97.3%	97%	285900		
Arvoredo Esparso	96.6%	97.5%	97%	183354		
Clareira	90.5%	96.9%	94%	4635	Overall Accuracy	97%
Mato Denso ou Arbustos	92.1%	98%	95%	29014	Overall Recall	97%
Pomar	94.9%	96.4%	96%	2231190	Overall Precision	95%
Pomar/-Vinha	92.8%	97.2%	95%	72192	Overall F1-Score	96%
Vinha	92.9%	98.4%	96%	16958	Overall Kappa	96%

Tabela 4.21: Classificação *XGBoost* com séries temporais na região de Fundão por maioria com contagem por peso de classe

Consequentemente com os resultados descritos nas tabelas 4.18 e 4.19, observa-se um aumento de cerca de 10% no *recall* e um decréscimo semelhante na *precision* relativamente ao classificador RF, onde estas métricas trocam praticamente de valores em relação à validação a píxel convencional. Pode-se concluir também que, utilizando pesos de cada classe na contagem por maioria não tem influência na validação final.

Como cada polígono é classificado de acordo com a classe maioritária dos píxeis que o constituem, polígonos mais pequenos de classes minoritárias, que na classificação original possuem muito ruído de classes maioritárias são inteiramente mal classificados, constituindo assim um aumento de falsos negativos, influenciando a *precision* negativamente.

Em relação ao classificador *XGBoost* (tabelas 4.20 e 4.21) as métricas mantêm-se completamente inalteradas pois este classificador já tem resultados muito aceitáveis e apenas polígonos com uma área muito pequena são mal classificados. Por outro lado o ruído é inexistente cancelando a influência negativa dos polígonos mal classificados. Este processo de validação apenas confirma a robustez da classificação do *XGBoost*.

No geral pode-se aferir que a grande maioria dos polígonos é classificada correctamente, significando que as áreas de vegetação relevante têm pelo menos a maioria da sua área bem classificada, e após a vetorização dos produtos *raster* e subsequente tratamento dos *shapefiles*, as métricas de validação terão resultados muito aceitáveis.

4.3.2 Comparação de Polígonos

Na comparação de Polígonos foram considerados os *shapefiles* da região do Fundão resultantes da classificação após algum processamento que envolveu a remoção de polígonos

insignificantes, buracos nos polígonos relevantes e a suavização e simplificação da sua geometria (processos descritos detalhadamente na secção 3.10.

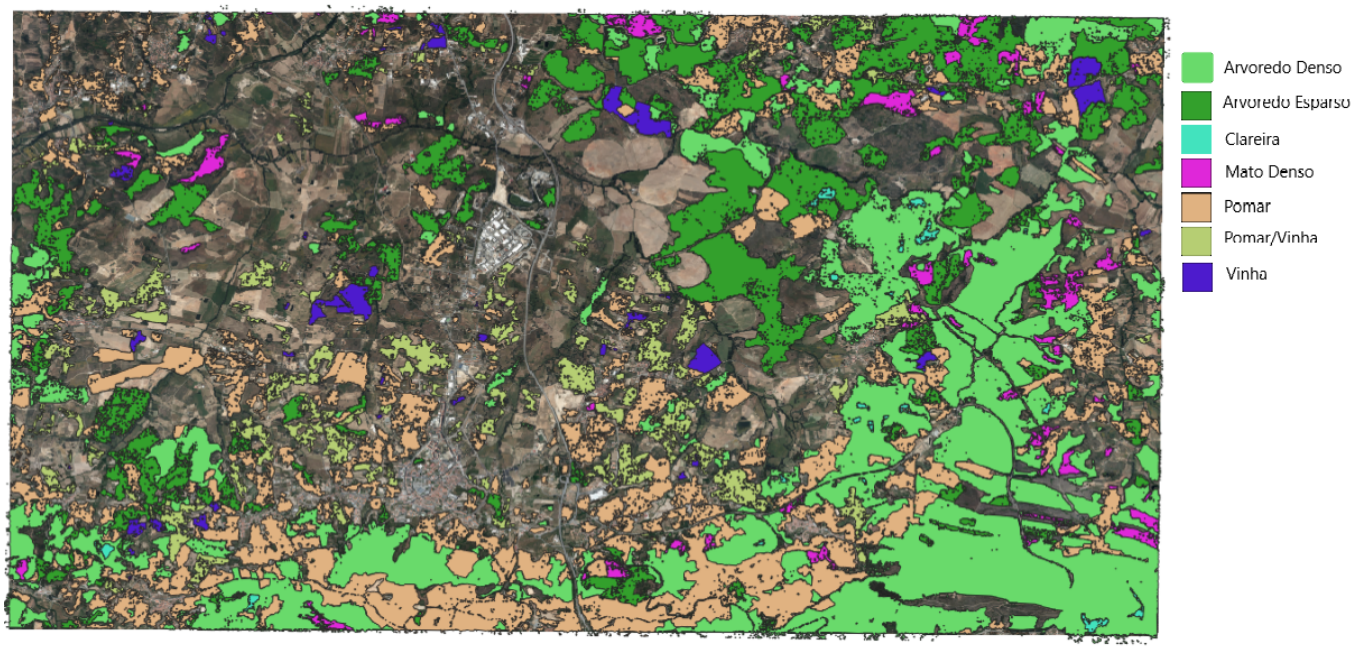

Figura 4.9: Shapefile resultante da classificação RF na região do Fundão

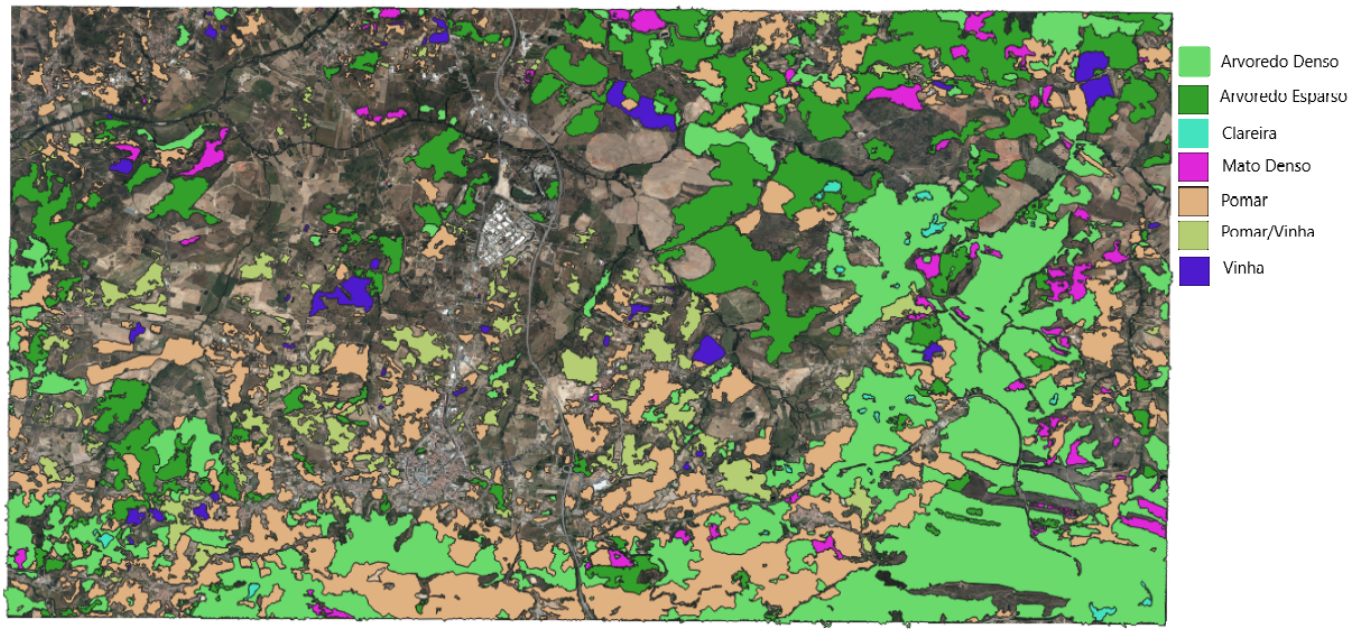

Figura 4.10: Shapefile resultante da classificação RF e processamento na região do Fundão

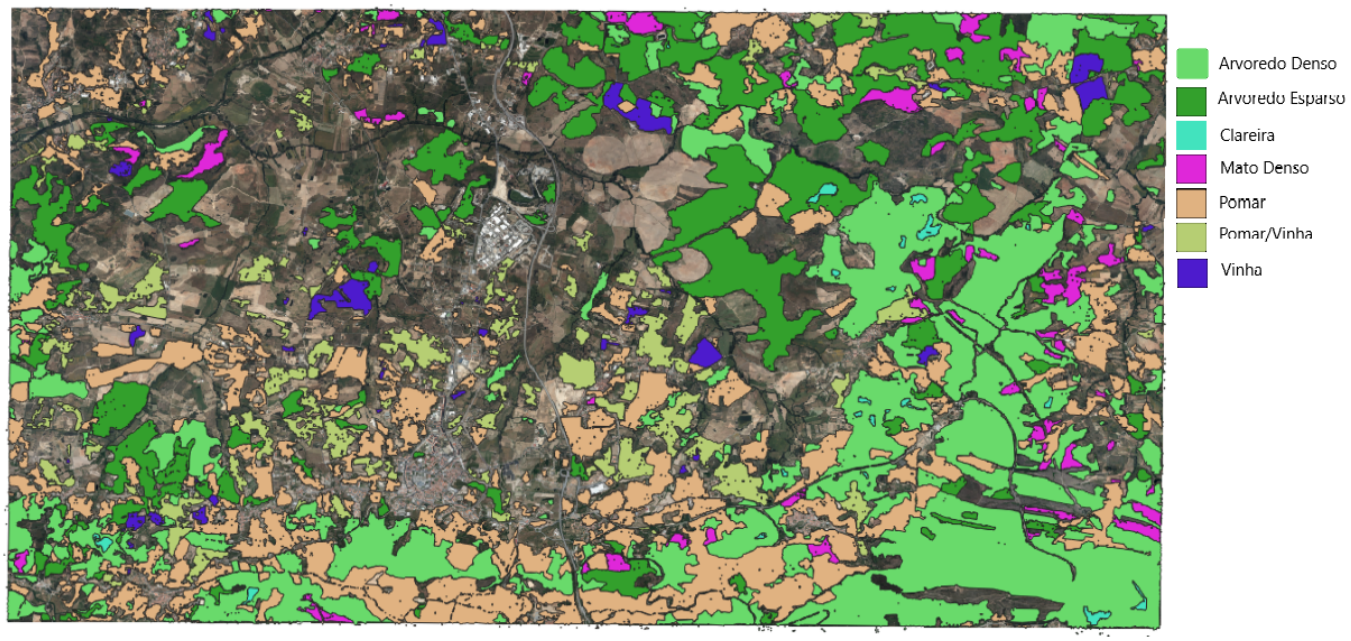

Figura 4.11: Shapefile resultante da classificação *XGBoost* na região do Fundão

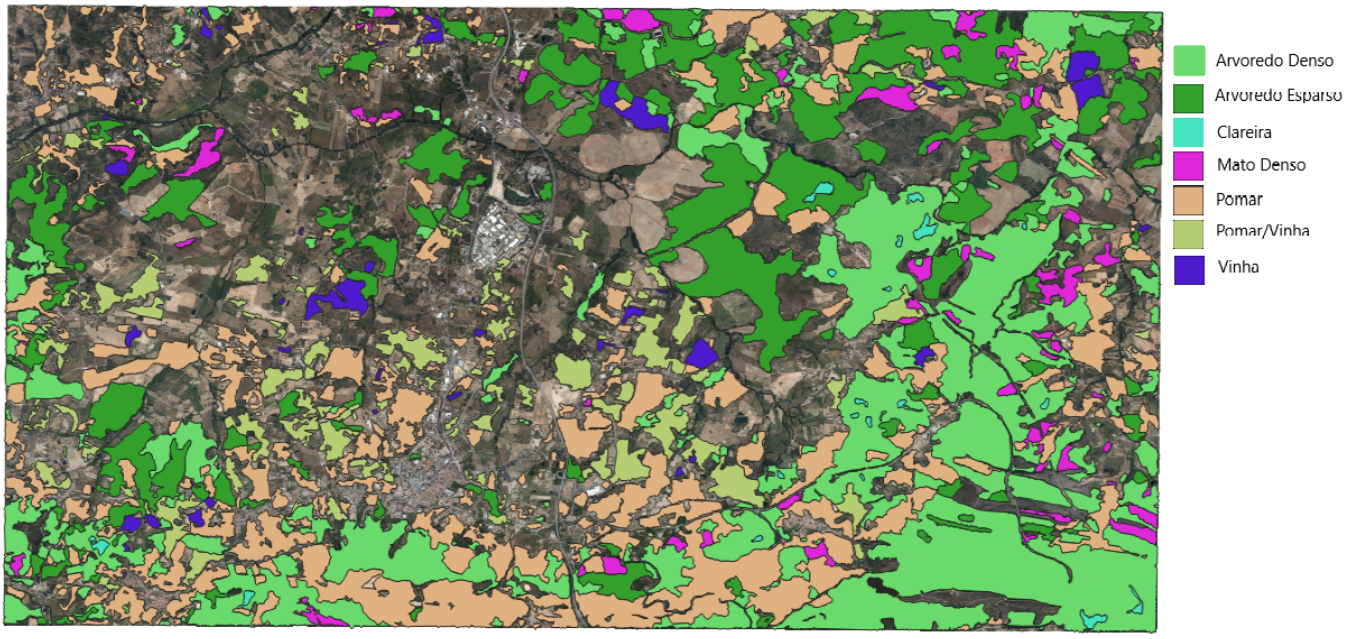

Figura 4.12: Shapefile resultante da classificação *XGBoost* e processamento na região do Fundão

Figura 4.13: Polígono antes da remoção de buracos e suavização de geometria

Figura 4.14: Polígono após a remoção de buracos e suavização de geometria

Nas figuras 4.9, 4.10, 4.11, 4.12 pode-se observar o resultado do processamento efetuado nos *shapefiles* correspondentes à classificação RF e *XGBoost* na região do Fundão. O ruído na é removido na sua totalidade, assim como polígonos com áreas insignificantes para a classificação de vegetação permanente. Adicionalmente nas figuras 4.13 e 4.14 pode-se observar com mais detalhe o efeito que a suavização e simplificação da geometria num polígono do *shapefile*. O objectivo desta simplificação da geometria é evitar o extensivo número de vértices em cada polígono provocado pela classificação píxel a píxel. No mínimo este processamento tenta suavizar os vértices de cada polígono de modo a que estes tenham uma geometria e topologia mais credíveis, e que facilite a análise por parte do CIGeoE.

Por fim, para avaliar não só o desempenho da classificação, mas também a influência

do processamento vectorial efectuado, foram calculadas as matrizes de confusão descritas em detalhe na secção 3.7

Classes	Precision	Recall	F1-Score		
Arvoredo Denso	92%	99%	96%		
Arvoredo Esparso	99%	100%	100%	Overall Accuracy	99%
Clareira	100%	100%	100%	Overall Recall	100%
Pomar	100%	100%	100%	Overall Precision	99%
Pomar/-Vinha	100%	100%	100%	Overall F1-Score	99%
Vinha	100%	100%	100%	Overall Kappa	94%

Tabela 4.22: Métricas resultantes da Classificação RF em Fundão por Polígono

Classes	Recall	Precision	F1-Score		
Arvoredo Denso	100%	100%	100%		
Arvoredo Esparso	100%	100%	100%	Overall Accuracy	100%
Clareira	100%	100%	100%	Overall Recall	100%
Pomar	100%	100%	100%	Overall Precision	100%
Pomar/-Vinha	100%	100%	100%	Overall F1-Score	100%
Vinha	100%	100%	100%	Overall Kappa	100%

Tabela 4.23: Métricas resultantes da Classificação *XGBoost* em Fundão por Polígono

Os resultados desta avaliação foram excelentes revelando que os polígonos do *shapefile* original e os polígonos resultantes de ambas as classificações são, essencialmente, os mesmos. Apesar de a morfologia não ser exactamente a mesma e de no caso de o classificador RF os polígonos poderem ter buracos, os polígonos encontram-se na mesma região e classificados com o mesmo tipo de vegetação. É importante referir que estas métricas de validação poderão ser diferentes se os parâmetros e comparação forem relaxados, os parâmetros de comparação foram descritos na secção 3.7, e por isso mesmo foram feitos testes onde a percentagem da área sobre a qual a diferença entre os polígonos é considerada uma intercepção relevante, foi aumentada até 50% da área original dos polígonos em questão.

Classificadores	Percentagem da Área Limite	Accuracy	Recall	Precision	F1-Score	Kappa
Random Forest	20% da Área Original	99%	100%	99%	99%	92%
	30% da Área Original	98%	100%	99%	99%	90%
	40% da Área Original	98%	100%	99%	99%	89%
	50% da Área Original	98%	100%	99%	99%	89%
XGBoost	20% da Área Original	100%	100%	100%	100%	100%
	30% da Área Original	100%	100%	100%	100%	100%
	40% da Área Original	100%	100%	100%	100%	100%
	50% da Área Original	100%	100%	100%	100%	100%

Tabela 4.24: Resultados dos Testes Feitos com o Limite de Área Usado para Assumir uma Intercepção como Válida

Na tabela 4.24 é possível perceber que com o relaxamento da percentagem de área de limite entre intersecções os valores das métricas mantêm-se relativamente idênticos. À excepção do *Kappa*, no *shapefile* resultante da classificação RF, variação nas métricas mantém-se praticamente nula, e mesmo no *Kappa* a variação é reduzida, revelando que os valores excelentes calculados anteriormente, não são enviesados, e que existe uma elevada robustez na classificação de vegetação e o seu subsequente processamento vectorial, onde a grande maioria dos polígonos se encontram sobrepostos com os polígonos da *ground truth* com classes iguais.

Apesar deste excelentes resultados é importante referir que esta suavização dos Polígonos só deverá ser feita quando se tem uma classificação raster com métricas muito boas (acima ou muito perto de 90%). Com classificações de menor qualidade o ruído será superior e existirão muitos polígonos que ficaram cortados em diversos polígonos de áreas pequenas. Consequentemente este polígonos poderão ser eliminados na fase de suavização da geometria e remoção de áreas pequenas o que poderá deteriorar os resultados

obtidos.

Esta avaliação prova que a classificação píxel a píxel, após a remoção do ruído obtém um grau de classificação de vegetação robusto e aceitável onde todos os polígonos estão representados com uma geometria semelhante onde todos os polígonos do *shapefile* original estão representados no *shapefile* ou por um polígono que cobre uma área muito semelhante á área coberta pelo polígono original ou por uma combinação de polígonos que acabam por perfazer a área original calculada pelo CIGeoE. Além disso não existe nenhuma sobreposição maioritária de polígonos com classificações diferentes entre os *shapefiles* originais e os resultantes da classificação.

Após esta análise pode-se reflectir também na melhoria que o pós-processamento dos *shapefiles* tem na classificação de vegetação permanente. Com este pós processamento eliminou-se o problema de preenchimento de polígonos devido ao desequilíbrio entre as classes. Consequentemente determinou-se que a classificação píxel a píxel aliada a este processamento vectorial constituem uma abordagem à classificação de vegetação permanente robusta eficiente e consistente.

4.4 Validação entre 9 Folhas

Após os excelentes resultados obtidos na classificação de uma folha, decidiu-se estudar o comportamento dos classificadores aquando uma situação de treino com uma folha inteira e a classificação de folhas circundantes.

Para este estudo seleccionou-se uma área constituída por 9 folhas adjacentes perto de vila verde em Portugal. Esta área foi seleccionada por ter as suas 9 todas classificadas na mesma altura e com uma classificação suficientemente recente que seja compatível com a utilização de imagens do satélite Sentinel.

Figura 4.15: As 9 Folhas Adjacentes e a sua numeração.

A figura 4.15 mostra a disposição espacial das folhas e que a cada folha tem uma numeração associada para ser possível a distinção entre elas.

O objectivo deste estudo é perceber se o algoritmo consegue generalizar, e se com o treino de apenas uma das 9 folhas, o algoritmo seria capaz de produzir uma classificação robusta e precisa da vegetação das outras 8 folhas.

Se possível, seria uma muito mais valia na aplicabilidade de cadeia de produção porque o CIGeoE apenas necessitaria de classificar uma folha em 9, ou no mínimo obter a classificação de vegetação correspondente a uma folha para poder classificar 8 folhas, o que encurtaria imenso o tempo de produção da vegetação de uma folha.

Para este efeito os testes realizados foram executados recorrendo ao algoritmo RF com séries temporais, devido aos seus resultados muito aceitáveis com um tempo de treino relativamente reduzido.

À semelhança das áreas utilizadas nos testes anteriores nem todas as áreas têm amostras provenientes de todas as classes de vegetação por isso neste apenas foram realizados testes com o treino de folhas com amostras de todas as classes presentes. As folhas que foram utilizadas para os testes seguintes foram a 28, 29, 42 e 56. As folhas apresentam uma distribuição de classes altamente não balanceada com as classes *NoData* e Arvoredo Denso constituindo cerca de 90% das amostras para cada folha.

Os resultados de folha para folha foram guardados e após a execução de todos os testes pretendidos para folhas com mais do que uma classificação, foi calculada a mediana da classificação de todos os testes feitos.

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
28	70%	18%	30%	17%	34%
29	71%	19%	25%	17%	36%
30	82%	22%	20%	21%	46%
41	77%	20%	33%	19%	53%
42	76%	20%	18%	19%	47%
43	72%	19%	18%	18%	44%
55	80%	23%	23%	23%	60%
56	78%	20%	18%	19%	45%
57	66%	19%	16%	17%	36%

Tabela 4.25: Resultados do Treino com uma Folha com a Classificação das 8 Folhas Circundantes

Como se pode observar na tabela 4.25 os resultados obtidos são muito maus, onde as métricas *precision* e *recall* muito raramente passam o limite de 25% e onde a única métrica que apresenta valores algo aceitáveis é a *accuracy*. No entanto este valor distingue-se dos outros porque estes classificadores apenas conseguem classificar com alguma eficácia as classes *NoData* e Arvoredo Denso. Como estas classes constituem cerca de 90% das amostras de cada folha a *accuracy* tem tendência a subir. As classes que não são Arvoredo Denso e *NoData* não têm qualquer tipo de presença significativa na classificação final.

Após estes resultados muito negativos decidiu-se aumentar o conjunto de treino para 2, 3 e por fim 4 folhas utilizando o mesmo conjunto de folhas anteriores com treino. Para estes testes foram utilizadas todas as combinações possíveis do número desejado de folhas para treino, ou seja para o treino com duas folhas fizeram-se 6 testes diferentes, para o treino com 3 folhas fizeram-se 4 testes diferentes, e por fim para o treino com 4 folhas fez-se apenas um teste. O número de folhas utilizadas para treino não foi aumentado mais porque o carregamento mais folhas seria impossibilitado por limites de memória RAM e porque se determinou que apenas 4 folhas têm uma distribuição de amostras pelas diferentes classes que se revela útil para o treino. Com o aumento de amostras utilizadas para treino, seria de esperar que o algoritmo tivesse mais facilidade em classificar as regiões.

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
28	69%	18%	31%	17%	33%
29	67%	19%	16%	17%	25%
30	82%	17%	26%	18%	23%
41	76%	19%	23%	18%	45%
42	75%	19%	26%	19%	41%
43	65%	15%	19%	14%	20%
55	76%	21%	22%	21%	49%
56	77%	18%	21%	18%	38%
57	58%	14%	20%	12%	12%

Tabela 4.26: Resultados do Treino com 2 Folhas com a Classificação das Restantes Folhas circundantes

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
28	69%	18%	31%	17%	33%
28	69%	19%	18%	17%	27%
30	82%	17%	27%	18%	26%
41	78%	20%	23%	19%	52%
41	75%	20%	26%	20%	43%
43	68%	16%	19%	16%	29%
55	80%	23%	23%	23%	58%
56	78%	19%	19%	19%	42%
57	61%	16%	21%	15%	20%

Tabela 4.27: Resultados do Treino com 3 Folhas com a Classificação das Restantes Folhas circundantes

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
30	84%	20%	25%	21%	45%
41	79%	20%	29%	20%	54%
43	75%	19%	20%	19%	47%
55	80%	23%	37%	23%	61%
57	62%	17%	40%	17%	23%

Tabela 4.28: Resultados do Treino com 4 Folhas com a Classificação das Restantes Folhas circundantes

Os resultados apresentados nas tabelas 4.26, 4.27 e 4.28 mostram que com o aumento de amostras utilizadas para treino, os resultados mantêm-se virtualmente idênticos. Adicionalmente nalgumas folhas com o aumento do número de amostras utilizadas para treino, o número de amostras da classe *NoData* aumenta, o que acaba por diminuir as métricas de validação.

Após a mediocridade destes resultados, o treino com 3 folhas foi replicado mas a classe *NoData* foi retirada da *groundtruth*. O objectivo deste teste é perceber qual é a influência que a classe *NoData* no treino e classificação das diferentes folhas. Como é uma classe muito heterogénea, representando todo o tipo de cobertura terrestre que não é relevante no contexto de vegetação permanente do CIGeoE indo desde água, a tecido urbano etc., e tem distintamente o maior número de amostras por folha poderá ser o culpado principal na enorme confusão de classificação destas folhas. Neste teste utilizaram-se apenas as folhas 29, 42 e 56 correspondendo estas a uma região central do conjunto de 9 folhas e apresentando todas um número de amostras aceitáveis de todas as classes.

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
28	89%	33%	39%	32%	80%
30	96%	31%	40%	32%	88%
41	96%	33%	42%	34%	91%
43	91%	32%	35%	32%	84%
55	97%	37%	41%	38%	94%
57	81%	34%	50%	33%	68%

Tabela 4.29: Resultados do Treino com 3 Folhas com a Classificação das Restantes Folhas circundantes sem a classe *NoData*

Com os resultados obtidos na tabela 4.29 percebe-se que a remoção da classe *NoData* para treino melhora o desempenho do algoritmo, no entanto não é suficiente para este algoritmo apresentar bons resultados. A *accuracy* e o *Kappa* têm uma subida muito acentuada, no entanto isto é devido ao aumento que o peso da classe Arvoredo Denso tem na classificação. Com a remoção da classe *NoData*, a classe Arvoredo Denso passa a ter cerca de 90% das amostras de cada região, e a sua classificação têm uma melhoria muito significativa. Por outro lado a classe Arvoredo Esparso começa a ter uma presença algo positiva nas classificações com um aceitável número de amostras bem classificado. No entanto as outras classes continuam com um desempenho muito fraco e daí a *precision e recall* baixas. Percebe-se que o classificador apenas classifica as amostras como sendo das suas duas classes maioritárias.

Após a replicação destes testes com o classificador *XGBoost* e chegando aos mesmos resultados, pode-se aferir que o algoritmo tem muita dificuldade em generalizar a classificação de uma folha para outras folhas.

Para perceber qual a causa deste problema de generalização decidiu-se avaliar qual era a resposta espectral mediana de cada banda em relação a cada classe. Para cada folha mediu-se a resposta espectral mediana de cada banda em relação a cada classe, e posteriormente a comparação destas respostas. Os resultados chegados foram muito interessantes.

Os resultados revelaram que as mesmas classes em folhas diferentes têm respostas espectrais diferentes para as mesmas bandas. Adicionalmente as classes maioritárias

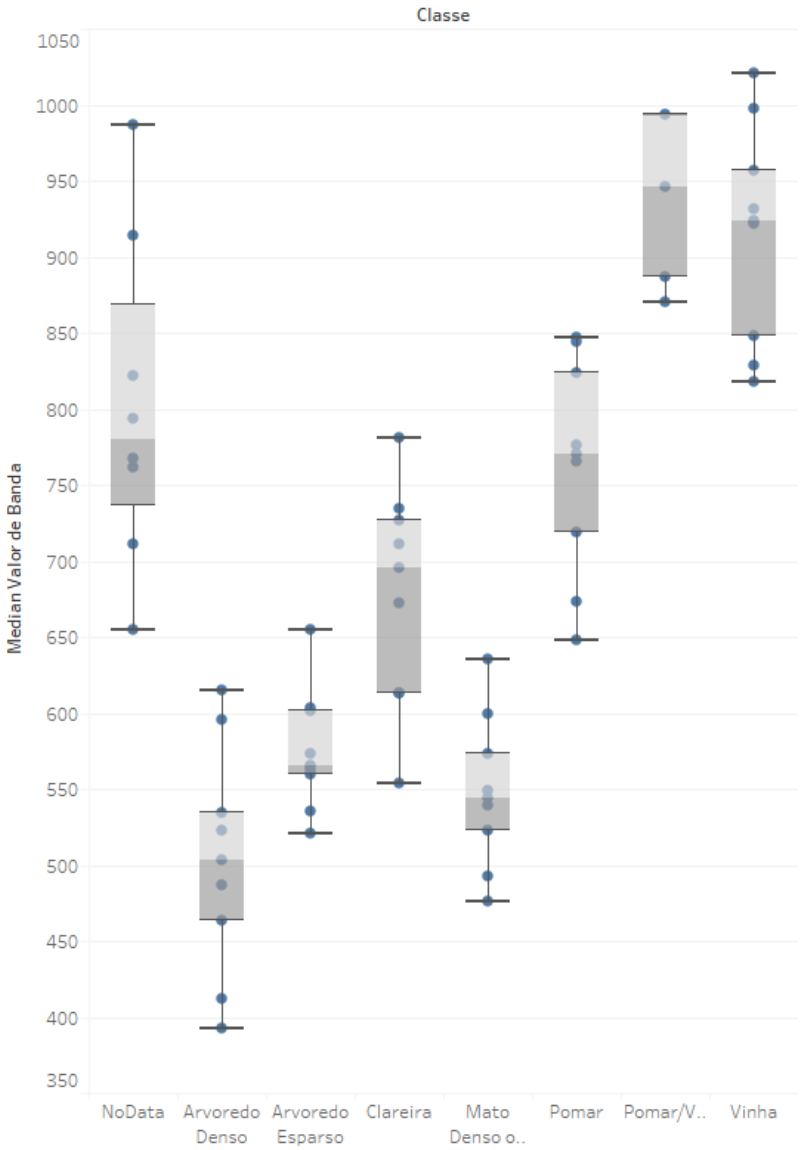

Figura 4.16: *Boxplot* correspondente à resposta Espectral da Banda 3 de Sentinel para cada Classe

apresentam das maiores diferenças entre os valores.

Na figura 4.16 estão as diferentes respostas espectrais das classes em relação à banda 3 de Sentinel (que corresponde à zona do verde na radiação visível) por folha. As outras bandas apresentam comportamentos semelhantes. Como se pode observar, as classes em geral têm respostas espectrais variadas de folha para folha. Sendo os valores calculados valores medianos, podemos assumir que nas folhas esta disparidade será ainda maior. Esta diferença nos valores das bandas é devida às classes usadas no contexto de classificação do CIGeoE são classes muito heterogéneas. A classe Arvoredo Denso pode corresponder a muitos tipos de floresta diferente e uma floresta de eucaliptos terá respostas espectrais diferentes de uma floresta de pinheiros. Isto por sua vez introduz imensa confusão na classificação da vegetação. Aliado ao alto desequilíbrio entre o número de amostras

das diferentes classes estes dois factores têm imensa influência na incompatibilidade de classificação entre diferentes folhas. Consequentemente os classificadores têm muita dificuldade em conseguir generalizar a classificação de folha para folha.

Adicionalmente a esta análise da resposta espectral de cada banda para a mesma classe em folhas diferentes, foi comparada com a *ground truth* do CIGeoE as cartas ocupação de solo (COS) de 2015. Estas cartas apresentam um número muito mais elevado de classes, e estas classes já se apresentam num estado indivisível e uma maior homogeneidade do que as classes do CIGeoE. A comparação destes dois produtos cartográficos baseou-se em avaliar quantas classes COS é que as classes do CIGeoE se dividiam e analisar essa divisão.

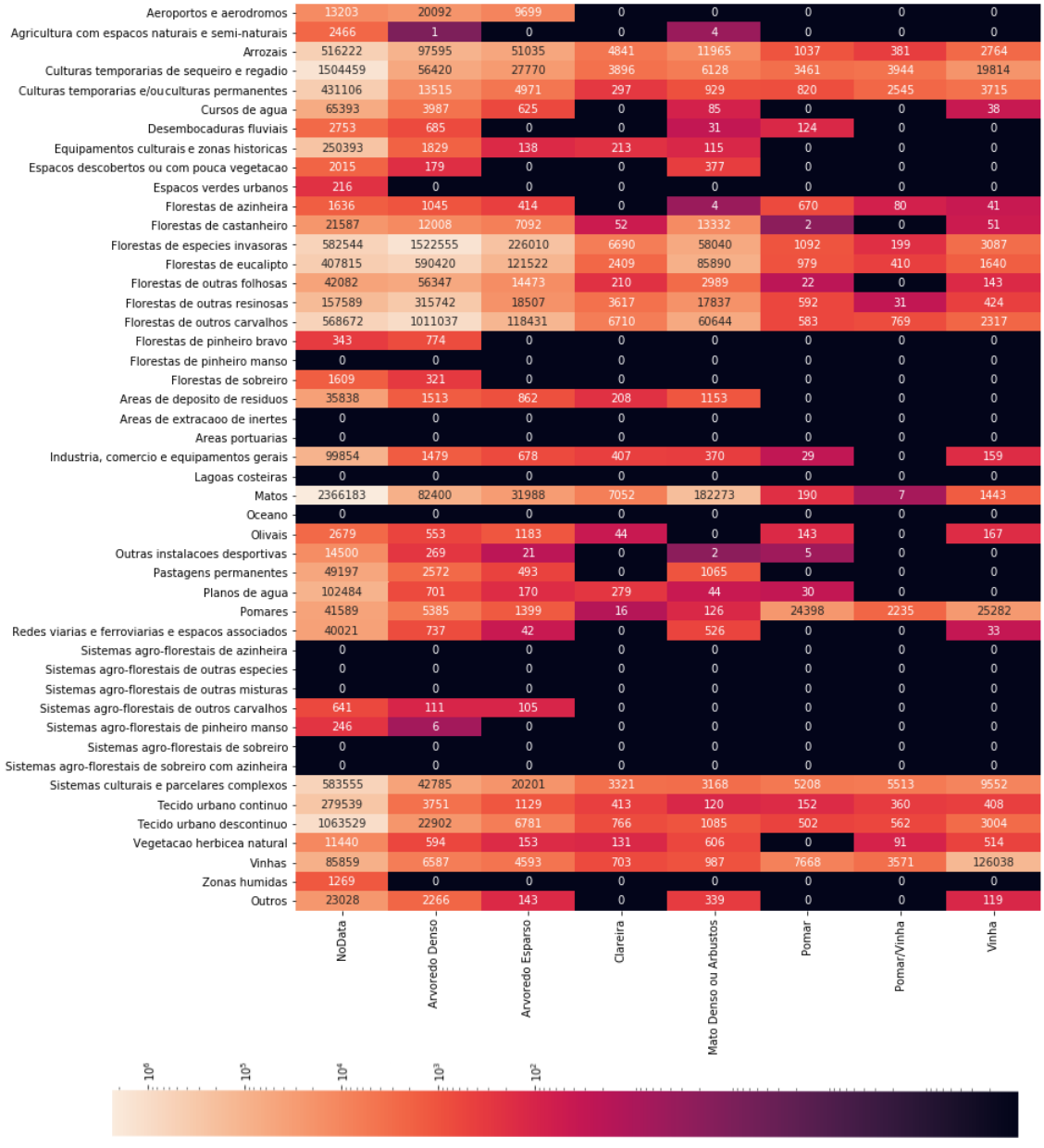

Figura 4.17: *Heatmap* que mostra a relação entre as classes COS e CIGeoE

O *heatmap* representado pela figura 4.17 mostra como é que as classes COS se distribuem pelas classes CIGeoE nas 9 folhas a estudar. Como se pode observar as classes com um maior número de incerteza são as classes *NoData* e Arvoredo Denso. A classe *NoData* é, como de esperado, a classe mais espalhada pelas classes COS. Esta classe tem uma grande incidência, na classe COS Matos assim com diversos tipos de Floresta, o que é surpreendente porque mostra que a classe *NoData* também vai ter certas respostas espectrais muito semelhantes às outras classes de vegetação do CIGeoE, criando assim conflito entre as características das classes a classificar, o que se traduz por um aumento de confusão na classificação automática. Por outro lado a classe Arvoredo Denso distribui-se essencialmente por classes de floresta diferentes. Isto provocará respostas espectrais diferentes por parte da mesma classe reafirmando o estudo que foi feito anteriormente com as respostas espectrais de cada banda. O Arvoredo Esparso apresenta um comportamento muito semelhante ao Arvoredo Denso, dividindo-se essencialmente por classes floresta diferentes. No geral para as outras classes minoritárias elas partilham sempre alguma confusão entre a sua classe correspondente na classificação COS, e outras classes COS que introduzirão ruído na aprendizagem do algoritmo. Esta análise revela que as classes do CIGeoE são classes muito heterogéneas não conseguindo ter uma característica espectral distinta. À falta desta característica torna-se muito difícil a generalização da classificação de uma folha para a outra, explicando assim os resultados muito medíocres verificados anteriormente.

Por fim decidiu-se estudar os resultados obtidos com o treino de amostras de 4 folhas e a posterior classificação dessas mesmas folhas.

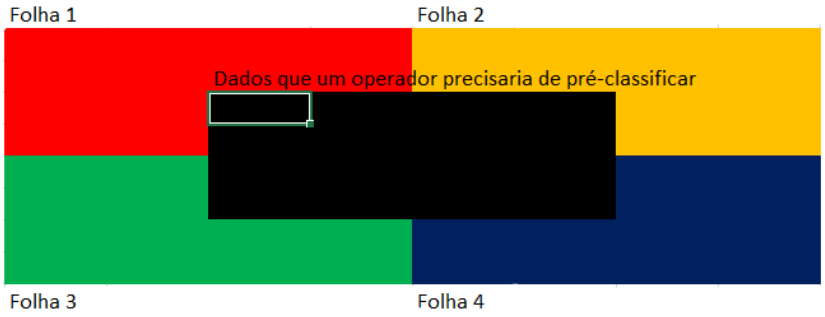

Figura 4.18: Modo de Obtenção das amostras para treino.

A figura 4.18 mostra como é que se obtêm as amostras das folhas. Para cada 4 folhas é manualmente classificada uma folha no meio das 4 folhas correspondendo a 25% de amostras de cada folha garantindo assim, que o classificador tem uma contextualização da vegetação das folhas que posteriormente classificará por inteiro.

Estes teste foram repetidos para as 9 folhas dividindo este conjunto em 4 sub-conjuntos de 4 folhas, ficando com os subconjuntos das folhas 28, 29, 41 e 42; 29, 30, 42 e 43; 41,

42, 55 e 56 e por fim o sub-conjunto das folhas 42, 43, 56 e 57. Para folhas que são classificadas mais do que uma vez, como é o caso da folha 42) foi calculada a mediana das métricas.

Por fim este teste pretende averiguar se, apesar de a não estratificação dos dados e a utilização de dados de diferentes folhas para treino, a classificação ainda se mantém relativamente robusta, e caso se mantenha pode corresponder a uma maneira útil de implementação na cadeia de produção do CIGeoE.

Folhas	Accuracy	Recall	Precision	F1-Score	Kappa
28	95%	82%	97%	88%	91%
29	95%	83%	97%	89%	90%
30	98%	86%	97%	91%	94%
41	96%	83%	97%	90%	93%
42	94%	77%	86%	81%	89%
43	96%	85%	96%	90%	92%
55	97%	82%	97%	89%	94%
56	95%	79%	97%	86%	88%
57	93%	83%	95%	88%	88%

Tabela 4.30: Resultados do Treino com uma Folha correspondente à Área Central de 4 folhas

Como se pode ver na tabela 4.30 esta abordagem de aquisição de amostras para treino tem algum sucesso aquando este conjunto de 9 folhas. As métricas sobem muito significativamente em relação aos testes anteriores, e mantêm valores aceitáveis.

No entanto existe alguma variação nas métricas, mais especificamente no *recall*, de folha para folha. Isto é devido à falta de estratificação das amostras de treino, o que significa cada folha não recebe a mesma proporção de amostras de treino de cada classe. Mesmo assim a classificação mantém-se geralmente robusta.

Porém este método de classificação é altamente dependente da qualidade dos dados de treino, e, apesar da sua consistência na classificação das 9 folhas, pode apresentar resultados muito inferiores caso a folha central não apresente suficientes dados de treino.

Não obstante este é um método de classificação com alguma aplicabilidade na cadeia de produção do CIGeoE.

Capítulo 5

Conclusão e Trabalho Futuro

Neste capítulo pretende-se retirar as conclusões do trabalho feito nesta dissertação assim como deixar recomendações para uma futura continuação deste trabalho e como melhorar esta abordagem de classificação.

5.1 Conclusão

Neste trabalho pretendeu-se o desenvolvimento de uma ferramenta que conseguisse classificar vegetação permanente no contexto da carta militar do CIGeoE.

Esta vegetação consiste em 8 classes principais distribuídas através de quatro folhas correspondendo às regiões de interesse Fundão, Monchique, Sendim do Douro e Tocha. As classes têm uma distribuição altamente desequilibrada constituindo um obstáculo importante no treino e classificação desta vegetação.

Com esse objetivo adquiriram-se produtos de Sentinel-1 e Sentinel-2 e desenvolveram-se três classificadores que apresentam muito bons resultados nas técnicas de literatura, nomeadamente *XGBoost*, *Random Forest* e *Suppor Vector Machines*, utilizando diferentes abordagens de classificação.

Foram implementadas duas metodologias principais de classificação. Uma metodologia envolvendo apenas um produto de Sentinel-1 e 2 da mesma data da classificação original ou muito perto desta. A segunda metodologia envolve o uso de séries temporais com *features* para a classificação, para isso foram utilizados diverso produtos de Sentinel-1 e 2 de diversas datas ao longo dos anos de 2015 e 2016.

O classificador SVM revelou-se incapaz de lidar com um número tão elevado de dados, e apresentou métricas muito inferiores aos seus classificadores concorrentes. Consequentemente com o acréscimo significativo de dados na metodologia de séries temporais o classificador SVM foi desconsiderado.

A metodologia de séries temporais constituiu uma melhoria tremenda na classificação da vegetação, ultrapassando largamente a metodologia temporal estática, constituindo um importante passo no desenvolvimento desta ferramenta

Foram testadas diversas implementações da abordagem *Timeseries*, e após a comparação do sucesso na classificação de cada uma destas implementações concluído que ao se incluir os dados temporais dos produtos de Sentinel na informação espectral teria resultados melhores do que métodos convencionais ou utilizando métricas estatísticas.

Os melhores resultados foram obtidos pelos classificadores RF e *XGBoost*, com ênfase no desempenho do classificador *XGBoost* que apresentou resultados excelentes.

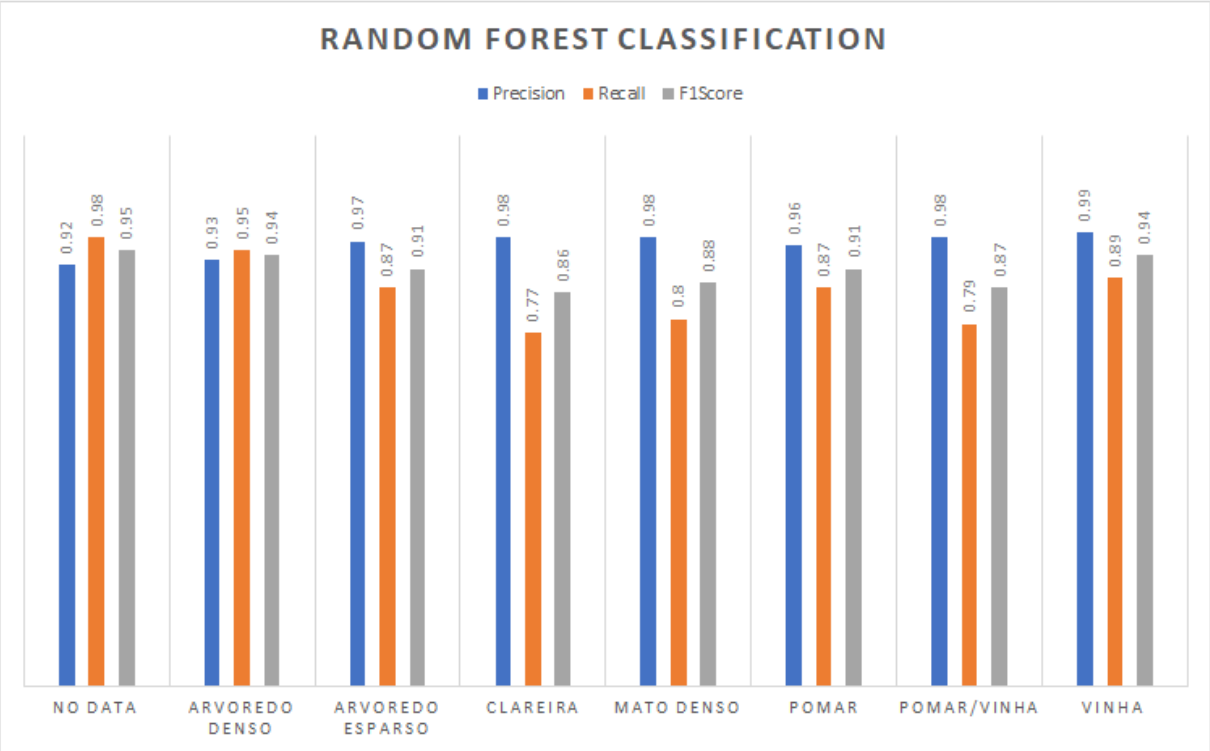

Figura 5.1: Classificação por Classe Random Forest

As figuras 5.1 e 5.2 apresentam as métricas de classificação que obtiveram os melhores resultados nos classificadores RF e *XGBoost*.

Exceptuando o classificador SVM, os resultados obtidos para qualquer um dos classificadores foram resultados muito aceitáveis, salientando a capacidade excelente do classificador *XGBoost* de conseguir classes maioritárias (No Data, Arvoredos) das classes minoritárias (Clareira, Mato Denso, Pomar, Pomar/Vinha, Vinha). Este provou-se o maior obstáculo nesta classificação causando confusão notável na classificação. No entanto o classificador *XGBoost* consegue em grande parte ultrapassar este obstáculo com uma implementação de série temporal.

Apesar deste excelentes resultados por parte do classificador *XGBoost*, o classificador RF desta-se pela sua rapidez de execução constituindo uma alternativa sólida caso seja

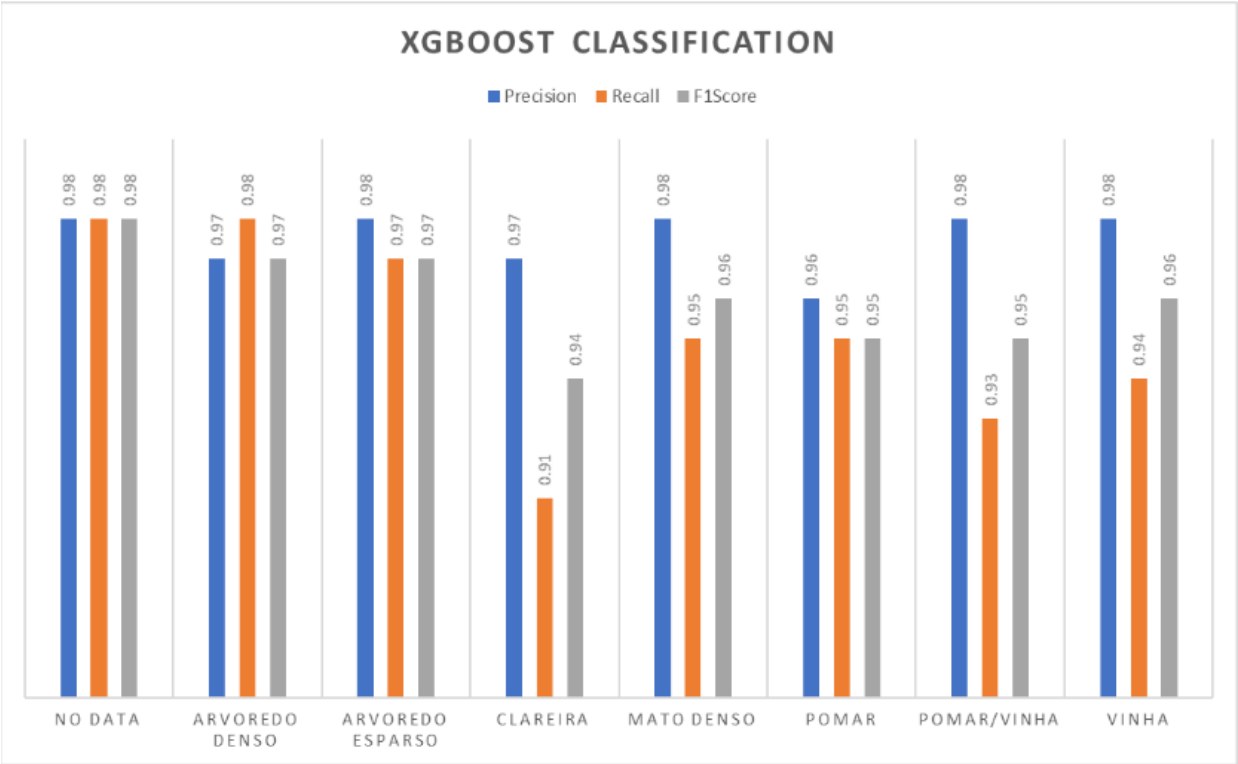

Figura 5.2: Classificação por Classe XGBoost

necessário uma classificação menos robusta mas mais rápida, enquanto ao mesmo tempo mantendo valores de métricas muito respeitáveis.

Comparativamente à classificação executada pelo CIGeoE pode-se determinar que este processo constitui uma melhoria no tempo de classificação de vegetação pois uma folha correspondente a uma das 4 regiões de estudo, demora cerca de 4 meses a classificar, contrariamente a esta ferramenta que utilizando o seu classificador mais lento demora cerca de 4 horas.

Por outro lado o classificador SVM tem um desempenho muito inferior devido ao desequilíbrio das classes. Este classificador não consegue fazer a distinção entre classes minoritárias e maioritárias devido à falta de amostras das classes maioritárias e à heterogeneidade da classe *NoData*. Além do mais que este classificador não lida bem com um número elevado de amostras para treino constituindo um aumento significativo no tempo de treino. Com todos estes dados pode-se concluir que este classificador não é adequado para este problema de classificação.

Por fim decidiu-se testar a capacidade de generalização espacial dos classificadores de modo a se desenvolver uma implementação com alguma aplicabilidade na cadeia de produção do CIGeoE. Deste modo adquiriu-se *ground truth* e imagens sensoriais de mais 9 folhas da região de Vila Verde. Os classificadores apresentaram muitos problemas na generalização da classificação. Após sucessivos testes onde se pretendia treinar com uma

folha e classificar outra folha adjacente, verificou-se que os resultados obtidos com qualquer um dos classificadores utilizados anteriormente eram muito fracos. Após a medição das respostas espectrais de cada classe e a comparação com o sistema de classificação COS, conclui-se que o desequilíbrio entre o número de amostras de cada classe, aliado à elevada heterogeneidade presente nas classes do CIGeoE impossibilitam uma generalização robusta da classificação. Este impedimento coloca em causa uma implementação com aplicabilidade na cadeia de produção do CIGeoE. No entanto apresenta resultados interessantes de interesse futuro para uma melhoria da ferramenta de classificação de vegetação.

Conclui-se assim que o método que apresenta um melhor *trade-off* entre robustez de classificação e aplicabilidade na cadeia de produção do CIGeoE é a pré-classificação manual de uma região central a 4 folhas e com uma dimensão semelhante a uma folha, que será posteriormente utilizada como treino do algoritmo. Com a obtenção desta "folha de treino"consegue-se obter cerca de 25% de amostras não estratificadas de cada uma das 4 folhas. Após o treino com a folha pré-classificada, procede-se à classificação de cada uma das 4 folhas. Adicionalmente a eliminação da classe *NoData* do conjunto de amostras de treino e de classificação, terá um aumento significativo no desempenho do classificador.

Aquando comparado com outros estudos, observa-se a inexistência de estudos com um contexto de vegetação tão distinto. Enquanto que a maior parte dos estudos se baseia em distinguir espécies diferentes de vegetação, esta dissertação baseia-se em distinguir conceitos mais abrangentes como Arvoredo Denso e Pomares e com classes de vegetação tão semelhantes (Pomar, Vinha, Pomar/Vinha). Os resultados obtidos revelam-se bons para a classificação de uma folha, existindo poucos estudos no âmbito de classificação de vegetação, com resultados tão altos ultrapassando o limite de 0.9 na *accuracy*, *recall*, *precision*, *f1-score e kappa*.

No entanto estes resultados apresentam um decréscimo muito elevado quando se tenta generalizar a classificação ficando muito aquém das convenções de uma boa classificação.

Um número de estudos complementares foram executados no intuito de obter uma avaliação vectorial dos resultados assim como numa perspectiva de melhorar a qualidade dos shapefiles obtidos. Com estes estudos foi possível eliminar o ruído dos resultados, resultante da classificação orientada a pixeis, e, ao mesmo tempo, foi possível suavizar-se e simplificar-se a geometria dos polígonos de classificação. Estes estudos obtiveram resultados muito positivos conseguindo obter uma geometria geral muito semelhante à geometria suavizada proveniente da classificação do CIGeoE. Por outro lado grande parte da confusão presente na classificação é eliminada devido à eliminação dos polígonos pequenos resultantes da introdução do ruído.

Após a obtenção do *shapefile* final podemos observar que os resultados obtidos são muito semelhantes à classificação original fruto de trabalho intensivo de 4 meses por parte do CIGeoE.

5.2 Recomendações para trabalho futuro

Para melhoria desta classificação como trabalho futuro existem algumas melhorias ou funcionalidades adicionais que poderão ser implementadas:

- **Segmentação de Imagem** Com segmentação de imagem pode-se, não só mitigar o problema de desequilíbrio de entre as classes, mas também reduzir o ruído provocado por uma classificação pixel a pixel. Por isto mesmo estudar um processo de segmentação das imagens é recomendado.
- **Produtos de Satélite com Maior Resolução Espacial** Com maior resolução espacial será possível aumentar o número de amostras no geral, e consequentemente possivelmente diminuir a percentagem de dados para treino, como obter informação mais rica e mais fácil de catalogar e classificar.
- **Distinguir Estruturas Urbanas** Será interessante utilizar estes classificadores para diferentes problemas ainda no contexto da Carta Militar como a distinção de estruturas urbanas ou artificias, com o intuito de automatizar o máximo possível o processo de criação de cartas militares.
- **Utilização de *Features* Adicionais** Incluir *features* adicionais como texturas e modelos digitais do terreno no processo de classificação utilizando séries temporais. Devido à falta de memória RAM da máquina utilizada, não foi possível incluir estas *features* nesta classificação, no entanto estas *features* tiveram uma influência muito positiva na classificação temporal estática de vegetação e por isso será interessante o estudo destas mesmas *features* numa classificação de séries temporais.
- **Estudar as capacidades de generalização temporal dos classificadores** Nesta dissertação estudou-se a capacidade de generalização espacial dos classificadores. Face aos maus resultados, seria interessantes estudar a capacidade que os classificadores teriam em generalizar a classificação treinando com imagens de um ano e classificando imagens de anos a seguir.

Bibliografia

[1] S. Abdikan, F. B. Sanli, M. Ustuner e F. Calò. "Land cover mapping using sentinel-1 SAR data". Em: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 41.July (2016), pp. 757–761. ISSN: 16821750. DOI: 10.5194/isprsarchives-XLI-B7-757-2016.

[2] S. Aggarwal. "Principles of Remote Sensing". Em: *Photogrammetry and Remote Sensing Division* (2003).

[3] *Alaska Satellite Facility*. https://vertex-retired.daac.asf.alaska.edu/#.

[4] B. Alexander. "Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection". Em: *Remote Sensing of Environment* 113.1 (2009), pp. 239–247. ISSN: 0034-4257. DOI: 10.1016/j.rse.2008.09.005. URL: http://www.sciencedirect.com/science/article/pii/S0034425708002733.

[5] T. Asano e N.Yokoya. "Image Segmentation Scheme for low-level computer vision". Em: *Pattern Recognition 14* (1981).

[6] R. Bamler. "Principles of Synthetic Aperture Radar". Em: *Surveys in Geophysics* (2000).

[7] M. Belgiu e L. Drăgu. "Random forest in remote sensing: A review of applications and future directions". Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (2016), pp. 24–31. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2016.01.011.

[8] L. Breiman. "Random Forests". Em: *Machine Learning 45, 5-32, 2001* (2001).

[9] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow e Z. Li. "A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach". Em: *Remote Sensing of Environment* 210.January (2018), pp. 35–47. ISSN: 00344257. DOI: 10.1016/j.rse.2018.02.045. URL: https://doi.org/10.1016/j.rse.2018.02.045.

[10] I. Castillejo-González, F. López-Granados, A. García-Ferrer e J. Pena-Barragán. "Object- and Pixel-Based Analysis for Mapping Crops and Their Agro- Environmental Associated Measures Using QuickBird Imagery". Em: *Computers and Electronics in Agriculture* 68 (2009), pp. 207–215.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer. "SMOTE : Synthetic Minority Over-sampling Technique". Em: 16 (2002), pp. 321–357.

[12] J. Cohen. "A Coefficient of Agreement for Nominal Scales". Em: *Educational and Psychological Measurement* 20.1 (1960), p. 37.

[13] *Copernicus Open Access Hub*. https://scihub.copernicus.eu/dhus/#/home.

[14] D. C. Duro, S. E. Franklin e M. G. Dub®. "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery". Em: *Remote Sensing of Environment* 118 (2012), pp. 259 –272. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2011.11.020. URL: http://www.sciencedirect.com/science/article/pii/S0034425711004172.

[15] J. L. Fernandes. "Analysis of Classification Algorithms for Crop Detection using LANDSAT 8 images". Tese de mestrado. Faculdade e Ciências e Tecnologia da Universidade Nova, 2015.

[16] T. Fletcher. *Support Vector Machines Explained*. http://www0.cs.ucl.ac.uk/staff/T.Fletcher/. 2009.

[17] Y. Freund e R. E. Schapire. "A Short Introduction to Boosting". Em: *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 1401–1406.

[18] *Grid Search vs Random Search*. https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318.

[19] M. Hall-beyer. "GLCM Texture : A Tutorial". Em: March (2017).

[20] H. He, Y. Bai, E. A. Garcia e S. Li. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning". Em: 3 (2008), pp. 1322–1328.

[21] O Heyman, G. G. Gaston, A. J. Kimerling e J. T. Campbell. "A per-segment approach to improving Aspen mapping from high-resolution remote sensing imagery". Em: *Journal of Forestry* 101.4 (2003), pp. 29–33. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0037493771{\&}partnerID=40{\&}md5=66e5e9e6b1283d6625da7da4b48bbdbc.

[22] M. Hussain, D. Chen, A. Cheng, H. Wei e D. Stanley. "Change detection from remotely sensed images: From pixel-based to object-based approaches". Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 80 (2013), pp. 91–106. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2013.03.006. URL: http://dx.doi.org/10.1016/j.isprsjprs.2013.03.006.

[23] K. Johansen, N. C. Coops, S. E. Gergel e Y. Stange. "Application of high spatial resolution satellite imagery for riparian and forest ecosystem classification". Em: *Remote Sensing of Environment* 110.1 (2007), pp. 29–44. ISSN: 00344257. DOI: 10.1016/j.rse.2007.02.014.

[24] N. Kussul, M. Lavreniuk, S. Skakun e A. Shelestov. "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data". Em: *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017), pp. 778–782. ISSN: 1545598X. DOI: 10.1109/LGRS.2017.2681128.

[25] E. A. Lehmann, P. Caccetta, K. Lowell, A. Mitchell, Z. S. Zhou, A. Held, T. Milne e I. Tapley. "SAR and optical remote sensing: Assessment of complementarity and interoperability in the context of a large-scale operational forest monitoring system". Em: *Remote Sensing of Environment* 156 (2015), pp. 335–348. ISSN: 00344257. DOI: 10.1016/j.rse.2014.09.034. URL: http://dx.doi.org/10.1016/j.rse.2014.09.034.

[26] C. Li e I. Science. "A Gentle Introduction to Gradient Boosting". Em: (1999).

[27] L. Ma, M. Li, X. Ma, L. Cheng, P. Du e Y. Liu. "A review of supervised object-based land-cover image classification". Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 130 (2017), pp. 277–293. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2017.06.001. URL: https://doi.org/10.1016/j.isprsjprs.2017.06.001.

[28] A. Marangoz, M. Oruç, S. Karakış e H. Şahin. "Comparison of Pixel-Based and Object-Oriented Classification Using Ikonos Imagery for Automatic Building Extraction – Safranbolu Testfield". Em: *"Fifth International Symposium"Turkish-German Joint Geodetic Days* November 2015 (2006), pp. 28–31.

[29] L. Mason, J. Baxter, P. Bartlett e M. Frean. "Boosting Algorithms as Gradient Descent". Em: *In Advances in Neural Information Processing Systems 12*. MIT Press, 2000, pp. 512–518.

[30] M.Goldberg e S. Shlien. "A clustering scheme for multispectral image". Em: *IEEE Trans. Systems Man Cybernet* (1978).

[31] R. M.Haralick e L. G.Shapiro. "Survey: Image Segmentation Techniques". Em: *Machine Vision International* (1984).

[32] Muthukrishnan.R e M.Radha. "Edge Detection Techniques For Image Segmentation". Em: *International Journal of Computer Science Information Techonolgy* (2011).

[33] P. T. Noi e M. Kappas. "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery". Em: *Sensors (Switzerland)* 18.1 (2018). ISSN: 14248220. DOI: 10.3390/s18010018.

[34] R. C. Nunes. "Automatic Crop Classification in Alentejo Region using Landasat-8 vs Sentinel Imagery". Tese de mestrado. Faculdade e Ciências e Tecnologia da Universidade Nova, 2017.

[35] J. R. Otukei e T. Blaschke. "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms". Em: *Int. J. Applied Earth Observation and Geoinformation* 12 (2010), S27–S31.

[36] *Pixel Based Classification*. https://www.stars-project.org/en/knowledgeportal/magazine/image-analysis/algorithmic-approaches/classification-approaches/pixel-based-classification/.

[37] J. Quian, Q. Zhou e Q. Hou. "Comparison Of Pixel-Based And Object-Oriented Classification Methods for Extracting Built-Up Areas in Aridzone". Em: *ISPRS Workshop on Updating Geo-Spatial Databases with Imagery The 5th ISPRS Workshop on DMGISs* (2005).

[38] J. A. Richards e X. Jia. *IRemote Sensing Digital Image Analysis*. Fourth. Springer, 2006.

[39] T. Robertson. "Extraction and Classification Of Objects In Multispectral Images". Em: *Machine Processing of Remotely Sensed Data* (1973).

[40] R. A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Third. Elsevier, 2006.

[41] *Sentinel-1 SAR User Guide Introduction*. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar.

[42] *Sentinel-2 MSI Introduction*. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi.

[43] D. Stow, Y. Hamada, L. Coulter e Z. Anguelova. "Monitoring shrubland habitat changes through object-based change identification with airborne multispectral imagery". Em: *Remote Sensing of Environment* 112 (mar. de 2008), pp. 1051–1061. DOI: 10.1016/j.rse.2007.07.011.

[44] *The Sentinel-1 Toolbox*. https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-1.

[45] J. Weszka, R. Nagel e A.Rosenfeld. "A threshold selection technique". Em: *IEEE Trans. Comput.* (1974).

[46] Q. Yu, Q. Yu, P. Gong, N. Clinton, G. Biging, M. Kelly e D. Schirokauer. "Objectbased detailed vegetation classification with airborne high spatial resolution remote sensing imagery. Photogrammetric Engineering and Remote". Em: *Sensing* 72.7 (2006), pp. 799–811. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.211.3508.