NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA

FRANCISCO CARVÃO MAGARREIRO PROENÇA DA SILVA

IDENTIFICAÇÃO DE OPERAÇÕES DE LIMPEZA EM FAIXAS DE GESTÃO DE COMBUSTÍVEL DE INCÊNDIOS AO REDOR DE HABITAÇÕES E LOCALIDADES

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Julho, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA

IDENTIFICAÇÃO DE OPERAÇÕES DE LIMPEZA EM FAIXAS DE GESTÃO DE COMBUSTÍVEL DE INCÊNDIOS AO REDOR DE HABITAÇÕES E LOCALIDADES

FRANCISCO CARVÃO MAGARREIRO PROENÇA DA SILVA

Orientador: Carlos Augusto Isaac Piló Viegas Damásio
Professor Associado, Faculdade De Ciências e Tecnologia da Universidade Nova de Lisboa

Coorientador: João Carlos Gomes Moura Pires
Professor Associado, Faculdade De Ciências e Tecnologia da Universidade Nova de Lisboa

Júri

Presidente: Carla Maria Gonçalves Ferreira
Professora Associada, Faculdade De Ciências e Tecnologia da Universidade Nova de Lisboa

Arguente: Isabel Franco Trigo
Investigadora Auxiliar, Instituto Português do Mar e da Atmosfera

Orientador: Carlos Augusto Isaac Piló Viegas Damásio
Professor Associado, Faculdade De Ciências e Tecnologia da Universidade Nova de Lisboa

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Julho, 2022

Identificação de operações de limpeza em faixas de gestão de combustível de incêndios ao redor de habitações e localidades

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo NOVAthesis (v6.9.5) [28].

Agradecimentos

O trabalho desenvolvido nesta dissertação não seria possível sem o apoio dos meus orientadores, Carlos Damásio e João Moura Pires, que me guiaram neste processo, oferecendo a orientação que necessitava para abordar este projeto de forma pragmática e para poder superar os vários desafios e frustrações que se apresentaram.

Deixo também um agradecimento à NOVA FCT, por tudo o que me deu, seja na vida académica como nas experiências que tive durante o curso. Este trabalho é apoiado pela NOVA LINCS (PCIF/MOG/0161/2019), com o apoio financeiro da FCT.IP.

Os dados que foram fornecidos pela Câmara Municipal de Mação e de Santarém foram imprescindíveis para a elaboração desta tese, sem eles não seria possível avaliar o trabalho realizado, estou muito grato por estas colaborações.

Finalmente, deixo um enorme agradecimento aos amigos e à família, que sempre me apoiaram em todo o processo desta tese e que me ajudaram a encontrar motivação nos momentos mais difíceis deste trabalho e do curso.

«To finish first, you must first finish.» (Juan Manuel Fangio)

Resumo

As florestas portuguesas são todos os anos ameaçadas por incêndios florestais. Para minimizar esse flagelo, existem planos de defesa com o objectivo de diminuir o risco de incêndio e limitar a sua propagação. Nesses planos incluem-se as faixas de gestão de combustível (FGCI), zonas onde a presença de vegetação deve ser reduzida ou inexistente.

Nesta dissertação é proposto o uso de imagens de satélite, denominadamente da missão Sentinel-2 para monitorizar o estado das FGCI, identificando os instantes de intervenção sobre as FGCI, tal como a sua ausência.

O uso de técnicas de aprendizagem automática como Random Forests são aplicadas para distinguir um instante num contexto normal, e num contexto pós-intervenção. São usados os dados de referência fornecidos e complementados por um processo manual de marcação de séries para treinar os modelos. O contexto temporal fornece informação importante para apoiar esta classificação.

Foi explorada também a hipótese de gerar automaticamente as FGCI, com base na deteção de estruturas permanentes, de forma a identificar faixas adicionais que não estão mapeadas nos planos de defesa.

Palavras-chave: Deteção Remota; Aprendizagem Automática; Séries Temporais; Faixas de Gestão de Combustível; Sentinel

Abstract

Every year, the Portuguese forests are threatened by wild fires. To minimize that danger, there are mitigation plans put in place with the intent of lowering the fire risk and limiting its spread. Those plans include the definition of Fire Management Zones (FMZ), zones in which the vegetation should be routinely trimmed, or removed.

In this thesis it is proposed the usage of satellite imagery, namely from the Sentinel-2 mission, to monitor the state of the FMZs, identifying the moments where an intervention occured, as well as the periods without any action, such as when the vegetation cover evolves naturally.

The usage of machine learning techniques like Random Forests are applied to distinguish a normal context in an FMZ's vegetation cycle and a moment that follows an intervention. These techniques use the ground truth data that was provided, complemented by a manual effort of time series tagging to be able to train these models. The temporal context provides valuable information to support the classification.

The possibility of automatically generating the FMZ's was also explored, relying on the detection of permanent artifical strutctures to generate them. This new map of automatic FMZ's can be used to detect aditional magament zones that are not currently defined in the oficial plans.

Keywords: Remote Sensing; Machine Learning; Time Series; Fire Management Zones; Sentinel

Índice

Índice de Figuras

Índice de Tabelas

Siglas

CMM	Câmara Municipal de Mação
COS	Carta de Uso e Ocupação do Solo
DGT	Direção-Geral do Território
DI	Departamento de Informática
FCT	Faculdade de Ciências e Tecnologias
FGCI	Faixa de Gestão de Combustível de Incêndios
FIC	Faixa de Interrupção de Combustível
FMZ	Fire Management Zones
FRC	Faixa de Redução de Combustível
ICNF	Instituto da Conservação da Natureza e das Florestas
IPMA	Instituto Português do Mar e da Atmosfera
IRECI	Inverted Red Edge Chlorophyll Index
NBR	Normalized Burn Ratio
NDI45	Normalized Difference Index 45
NDVI	Normalized Difference Vegetation Index
NOVA	Universidade NOVA de Lisboa
OSM	Open Street Maps
PMDFCI	Plano Municipal de Defesa da Floresta Contra Incêndios
PNDFCI	Plano Nacional de Defesa da Floresta Contra Incêndios

1 Introdução

1.1 Contexto

Os incêndios florestais são todos os anos responsáveis por enormes perdas a nível ambiental, económico e humano. Os espaços florestais têm uma vasta extensão, cobrindo mais de 6 milhões de hectares em Portugal Continental [12], o que dificulta a implementação de medidas de prevenção e controlo de incêndios.

Visando a proteção das florestas e dos recursos que as rodeiam, em 2006 foi delineado o Plano Nacional de Defesa da Floresta Contra Incêndios (PNDFCI), que envolve a recolha de características relevantes (climáticas, populacionais, topográficas, entre outras) e a definição de planos de ação para reduzir o risco de incêndio. Estes planos são delineados ao nível municipal, nos Planos Municipais de Defesa da Floresta Contra Incêndios (PMDFCI), seguindo as regras definidas pelo PNDFCI [30].

Entre as medidas adotadas nos PMDFCI inclui-se a definição de redes de Faixas de Gestão de Combustível de Incêndios (FGCI). As FGCI são faixas e parcelas que são mantidas, sendo a sua vegetação removida, de forma total ou parcial, com o intuito de reduzir ou impedir a propagação de fogos nessa faixa. O seu objetivo é diminuir a superfície percorrida por fogos, isolar focos de incêndios, e proteger infraestruturas e bens (como redes ferroviárias, linhas elétricas e habitações).

Esta tese segue o trabalho anteriormente realizado por outros alunos da FCT/UNL, que tiveram sucesso a monitorizar remotamente o estado das FGCI em torno das estradas, e a detetar estruturas permanentes de forma remota [32][3]. Esta tese pretende continuar esse trabalho, integrado no projeto Floresta Limpa, que foi financiado pela Fundação para a Ciência e Tecnologia, contando com a parceria de entidades como o Instituto de Engenharia de Sistemas e Computadores (INESC-ID), o Município de Mação, o Município de Santarém, a Navigator Company e o IPMA.

1.2 Problema e Motivação

As FGCI são uma componente muito importante para o combate dos incêndios florestais, isto é evidenciado pelos históricos incêndios de Pedrógão Grande em 2017, que causaram 66 vítimas mortais e centenas de feridos. Para além da trágica perda de vidas, causaram também a destruição de 491 habitações e danos diversos em 48 empresas, que, no seu conjunto, empregavam 372 trabalhadores [8].

Na análise posterior ao incêndio, realizada pelo Centro de Estudos sobre Incêndios Florestais (CEFI), revelou-se que a causa principal foi a má manutenção da vegetação junto a uma linha elétrica [10]. É importante também mencionar que, em 95% das estruturas que sofreram uma destruição total, não existia uma gestão de combustíveis apropriada ao redor das mesmas [4]. O município de Pedrógão Grande também não possuía um PMDFCI validado pelo ICNF, e a taxa de implementação de faixas protetoras ao redor de aglomerados populacionais era quase nula [10]. Estas informações reforçam a importância que a gestão de combustível tem no combate aos incêndios e a sua capacidade de reduzir danos.

A monitorização do estado das FGCI e a sua respetiva manutenção é dificultada pela vasta área que estas ocupam, e pela variedade de entidades responsáveis pela sua limpeza (incluindo entidades e proprietários privados). O crescimento da vegetação é influenciado pela meteorologia e tem variações sazonais, que podem causar alterações relevantes em curtos espaços de tempo. É necessário um enorme esforço para fiscalizar todas as faixas do território nacional, e fazê-lo de forma frequente e completa é virtualmente impraticável.

A identificação automática do estado de limpeza das FGCI e a análise da sua variação vegetativa natural ao longo do ano permitiria acelerar o processo de monitorização e fazer com que essas zonas fossem intervencionadas da forma mais expedita possível, quando tal fosse necessário. Esta monitorização automática avaliaria também, direta, ou indiretamente, o nível de biomassa presente nas faixas, que são indicadores de um perigo de incêndio acrescido. [31, 4, 2]

1.3 Floresta Limpa

Esta dissertação foi desenvolvida no âmbito do projeto Floresta Limpa, enquadrada na tarefa de deteção remota e classificação de ações de limpeza nas FGCI, sendo que existem outros objetivos dentro do projeto para a coleção de dados do estado das FGCI no terreno, para a integração de toda a informação relevante ao projeto, entre outros. Todas estas tarefas têm um objetivo comum que é o de minimizar o risco e impacto dos incêndios florestais através da monitorização das FGCI, e de outras áreas de interesse.

O projeto conta com a colaboração de investigadores e colaboradores de diferentes áreas, desde a área da meteorologia, com o IPMA, de investigação de informática com o INESC-ID, e das instituições que estão diretamente envolvidas nos processos de manutenção de FGCI, como a Câmara Municipal de Mação e a Navigator Company.

1.4 Objetivos e Abordagem

Para fazer frente aos problemas expostos foi tomada a decisão de se usar imagens de satélite para monitorizar remotamente o estado atual das FGCI. Esta abordagem permitiria obter informação atualizada sobre as zonas que estão sobre um maior risco de incêndio e que necessitam de intervenção.

Usando as definições das FGCI e os momentos em que foram limpas, é possível estabelecer uma correlação entre as imagens fornecidas pelos satélites nessa área e o estado da faixa (classificada como intervencionada ou não intervencionada).

A análise das imagens ao longo do tempo permite identificar tendências no crescimento da vegetação nas áreas analisadas, ajudando a detetar alterações significativas, como seria o caso de uma intervenção de limpeza ou de um incêndio. Será também possível correlacionar determinados fatores, como o tipo de vegetação e as condições meteorológicas com o crescimento, ajudando a melhorar os modelos de classificação com base no contexto.

As espécies que compõem a floresta nacional têm características diferentes em termos de combustibilidade (espécies como os eucaliptos são muito inflamáveis, por exemplo) e em termos fenológicos[1]. Como tal, os mapas de ocupação de solo (*land use/land cover* - LULC) também são relevantes para o objetivo em questão. Estes mapas classificam zonas de variadas áreas em classes e respetivas sub-classes, permitindo distinguir entre, por exemplo, zonas florestais, matos, pastagens e estruturas artificiais.

Os satélites da missão Sentinel-2, lançados em 2015 e 2017, mostram-se particularmente úteis para a monitorização das FGCI e da vegetação em geral, devido à sua resolução espacial e frequência de capturas suficientemente alta (a cada 5 dias). Estes satélites disponibilizam imagens captadas em várias bandas eletromagnéticas, valores estes que podem ser usados para calcular índices que se correlacionam com características como a densidade da vegetação. Recorrendo à informação das bandas disponibilizadas e dos índices calculados, pretende-se utilizar algoritmos de aprendizagem automática para avaliar se as faixas foram, ou não, intervencionadas.

O uso de diferentes conjuntos de dados nos processos de treino e diferentes parametrizações irão influenciar os resultados da aprendizagem automática, e a análise posterior dos mesmos permite tirar conclusões sobre as abordagens e os dados que são mais adequados para a tarefa de identificação de instantes de intervenção.

O trabalho realizado tem como área de estudo o concelho de Mação e de Santarém, pois são as regiões nas quais temos informação atualizada sobre o estado e datas de intervenção. Estes conjunto de dados georreferenciados irão permitir-nos criar zonas de análise mais pequenas, para que nelas se possa avaliar a mudança dos níveis de vegetação ao longo do tempo e identificar os instantes de intervenção, caso os haja. É também necessário enriquecer os dados fornecidos, marcando os instantes exatos das ações de intervenção. A análise feita irá focar-se primeiramente nos dados de Mação, dado que

[1] A fenologia refere-se ao estudo do desenvolvimento das plantas ao longo das suas diferentes fases

é o conjunto de dados a que se teve acesso mais cedo e que aparentam ser de melhor qualidade, sendo as conclusões e os modelos serão posteriormente testados para a zona de Santarém, de forma a avaliar se a abordagem escolhida tem uma boa capacidade de generalização.

Além dos dados de referência e das faixas dos PMDFCI existem outros conjuntos de dados que são relevantes para os objetivos. Os mapas do COS, por exemplo, permitem-nos separar as áreas de estudo pela sua ocupação de solo, ajudando a assegurar que estas são o mais homogéneas possíveis.

1.5 Resultados

Resulta desta dissertação um conjunto de métodos que permitem, de forma algo precisa, identificar os instantes que constituem momentos de intervenção, sendo que não foi identificado nenhum trabalho prévio que o fizesse relativamente às faixas em redor de habitações e aglomerados populacionais.

São exploradas várias formas de analisar o valor e variação da vegetação ao longo do tempo, permitindo depois determinar que características são mais importantes para os objetivos em curso.

Foi realizado um trabalho considerável nas fases de pré-processamento, tanto dos ficheiros geográficos, como dos valores de deteção remota, para melhorar a qualidade dos dados usados.

Após a análise da qualidade das FGCI e dos dados de referência foram explorados abordagens que permitissem identificar as faixas em falta, e adicionar informação mais precisa aos dados de referência, enriquecendo os dados e permitindo tirar melhores conclusões.

Para agilizar o processo de marcação de séries temporais criou-se também uma ferramenta versátil de anotação de séries que pode facilmente ser adaptada para registar outros eventos de outros tipos de séries. Esta plataforma cumpre o seu objetivo de tornar o processo de marcação de séries mais eficiente, e de permitir ao utilizador indicar o grau de certeza das suas anotações.

1.6 Estrutura do Documento

Este documento encontra-se estruturado da seguinte forma:

- **Conceitos** - Aqui introduzem-se os conceitos essenciais desta dissertação, como a estrutura das FGCI, conceitos de deteção remota e de índices de vegetação.

- **Dados Relevantes** - Neste capítulo exploram-se os conjuntos de dados mais importantes para o trabalho realizado, a sua composição e a necessidade dos mesmos para os objetivos.

- **Trabalho Relacionado** - Aqui inclui-se um conjunto de trabalhos relacionados que se enquadram diretamente com os objetivos desta dissertação, nomeadamente da avaliação de vegetação e de FGCI.
- **Estado da Arte** - Neste capítulo detalha-se o estado da arte, o conjunto de ferramentas e metodologias que se apresentam mais relevantes para os objetivos da dissertação.
- **Abordagem** - Aqui pretende-se demonstrar a abordagem usada para o processamento e enriquecimento de dados, geográficos, ou de deteção remota, e a análise e métodos de aprendizagem realizados sobre os mesmos.
- **Implementação** - Nesta fase descreve-se os detalhes de implementação que foram mencionados de forma mais geral no capítulo da Abordagem.
- **Trabalho Experimental e Resultados** - Neste capítulo documenta-se as experiências realizadas sobre os dados, e os resultados das mesmas.
- **Conclusão e Trabalho Futuro** - São tiradas conclusões sobre a metodologia usada e os resultados obtidos, fazendo-se uma reflexão sobre as diferentes abordagens que podiam ser tomadas no trabalho futuro.

2

Conceitos

2.1 Faixas de Gestão de Combustível de Incêndios

As FGCI são um conjunto de parcelas onde a vegetação é removida parcial ou totalmente, com o intuito de limitar a propagação de um incêndio e facilitar o combate ao mesmo.

As datas em que estas faixas foram intervencionadas não constam da base de dados pública do ICNF. Esta informação é atualmente disponibilizada pelo Município de Mação ao projeto Floresta Limpa, que inclui detalhes sobre algumas faixas que rodeiam as estradas, habitações, aglomerados populacionais e faixas primárias. O projeto Floresta Limpa pretende explorar outras fontes de informação para dados de terreno, havendo já outras câmaras interessadas em aderir ao projeto, nomeadamente a Câmara Municipal de Santarém.

2.1.1 Níveis

Estão, de acordo com o Decreto-Lei n.º 124/2006 [30], definidas três tipos de redes de FGCI:

1. **Rede Primária** (planeada ao nível distrital), cujas funções são:
   a) Diminuição da superfície percorrida por grandes incêndios;
   b) Proteção das vias de comunicação, infraestruturas, povoamentos, entre outros;
   c) Isolamento de pontenciais focos de ignição de incêndios.
2. **Rede Secundária** (de interesse municipal ou local) - Cumpre as funções b) e c) definidas na Rede Primária.
3. **Rede Terciária** (de interesse local) - Cumpre a função c) definida na Rede Primária.

Podemos também classificar as FGCI de acordo com o tipo de intervenção que é aplicado, sendo definidas as Faixas de Interrupção de Combustível (FIC), nas quais é feita uma remoção total da vegetação, e as Faixas de Redução de Combustível (FRC), onde a remoção é parcial.

2.1.2 Rede Primária

A Rede Primária inclui as faixas de maior importância para a contenção de incêndios florestais. Possuem, no mínimo, uma largura de 125 metros, sendo que os 5 metros interiores constituem uma via, os próximos 40 metros adjacentes são FIC, e os seguintes 100 metros são FRC. As faixas que compõem a Rede Primária têm áreas totais entre 500 e 10000 hectares.

2.1.3 Rede Secundária e Terciária

As faixas da Rede Terciária geralmente incidem sobre a rede viária e elétrica, sendo que as da Rede Secundária abrangem também edificações e aglomerados habitacionais. As faixas que rodeiam as redes viárias, florestais e de energia têm uma largura mínima de 10 metros em cada direção, sendo que no caso de linhas elétricas de média tensão o mínimo é de 7 metros.

No caso das faixas da Rede Secundária ao redor de habitações o raio da faixa é de, no mínimo, 50 metros, e no caso de aglomerados populacionais, áreas industriais e parques de campismo, de 100 metros.

Dentro da definição geral de FRC existem algumas restrições adicionais que se aplicam às faixas que rodeiam edificações e infraestruturas. Estas limitam a altura da vegetação a 50 centímetros no caso de arbustos e a 20 centímetros no caso de ervas (subarbustivo). Obrigam também ao distanciamento de 4 metros entre as copas das árvores presentes (10 metros para pinheiros bravos e eucaliptos), bem como ao distanciamento de, no mínimo, 5 metros entre a propriedade e as copas das árvores e arbustos. A figura 2.1 exemplifica estas regras (excepto a limitação da altura das copas).

Destas redes não é esperado que contenham, por si, um incêndio florestal. O seu foco está em diminuir os eventuais danos materiais e humanos que um incêndio possa causar [4].

2.2 Deteção Remota

A Deteção Remota consiste em observações à distância utilizando instrumentos a bordo de satélites, aviões, ou mesmo em terra, para obter variáveis relevantes para o estudo e compreensão do planeta. Em particular, dados de deteção remota podem ser usado para caracterizar as superfícies terrestres. A captura das imagens a partir dos meios aéreos pode ser feita de forma ativa ou passiva.

Num caso de deteção passiva, o método de atuação é semelhante ao de uma câmara fotográfica convencional. O Sol atua como fonte de energia, emitindo radiação eletromagnética sobre a superfície da terra. Estes raios interagem com a atmosfera e a superfície, sendo depois refletidos e posteriormente medidos pelos sensores.

A deteção ativa, pelo outro lado, envolve o uso de energia própria para propagar sinais para a superfície terrestre. Este tipo de deteção, apesar de requerer um uso de

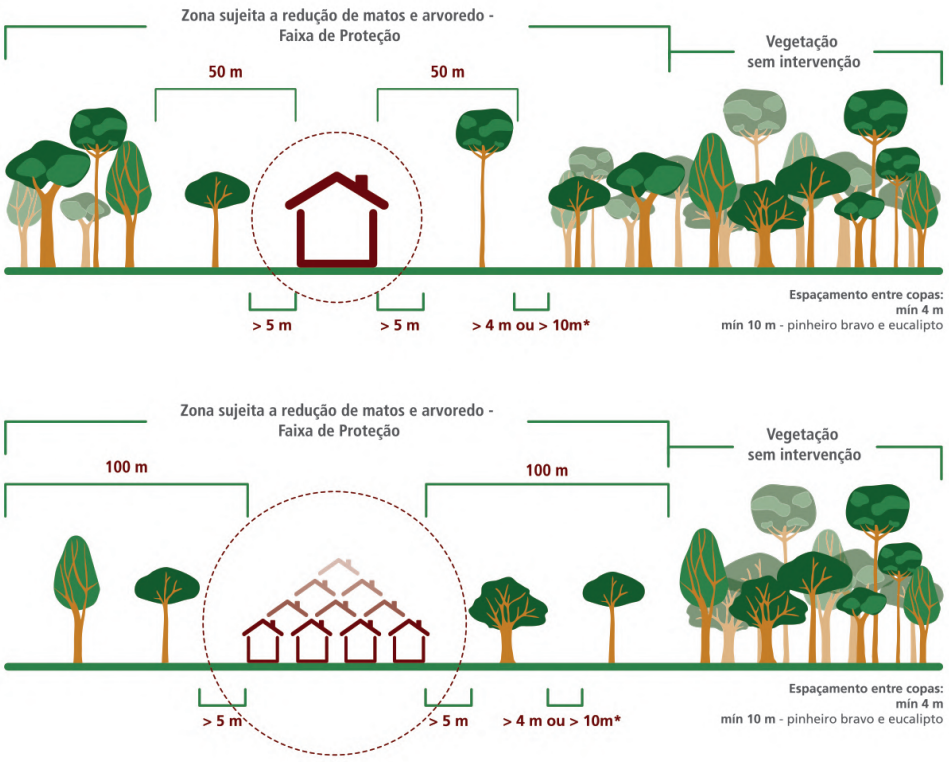


Figura 2.1: Esquema representativo das regras de gestão de FGCI em torno de habitações e localidades - *Retirado de [19]*

energia muito superior pode mostrar-se útil para analisar, por exemplo, faixas do espetro eletromagnético que o Sol emite em baixa quantidade. Como a radiação emitida penetra as camadas de nuvens pode ser usado também para capturar imagens do solo quando as capturas passivas estão ocluídas por nuvens.

Esta dissertação usará tanto satélites de deteção ativa, como de deteção passiva, embora o ênfase se coloque nestes últimos.

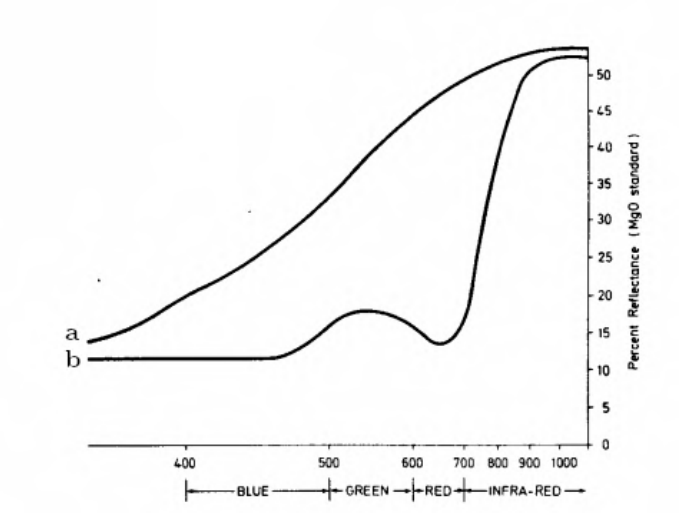
Os satélites que se mostram mais promissores são os da missão Sentinel-2 devido à sua alta resolução espacial, acesso livre aos dados, e ao facto do Sentinel-2 (equipado com um sensor passivo) ser multiespectral, o que permite obter uma grande quantidade de características das suas superfícies.

2.2.1 Fundamentos de Deteção Remota

As ondas eletromagnéticas emitidas, seja pelo Sol ou pelos componentes de um satélite, não se encontram limitadas pelo intervalo de frequências que define a luz visível. Satélites multiespectrais capturam valores de várias bandas, que frequentemente incluem a gama de azul, verde e vermelho que nos são visíveis, e podem também incluir valores de vários intervalos no infra-vermelho, rádio, microondas e demais. Cada objeto reage à

incidência de radiação de forma diferente, as características de absorção, reflexão e transmissão de radiação formam uma espécie de impressão digital que nos permite identificar características sobre os objetos que estamos a analisar.

No caso da vegetação, a forma como esta reflete a radiação é indicativo da sua densidade e do seu estado. Como exemplificado pela figura 2.2, duas zonas de vegetação em fases diferentes têm refletividade diferente em diferentes comprimentos de onda. Na gama do vermelho a vegetação viva tem valores baixos, enquanto a vegetação senescente tem valores altos, sendo que ambas têm o mesmo tipo de comportamento na gama infravermelho. Este tipo de discrepâncias permite-nos identificar a presença e o estado da vegetação numa dada área. Estas características são a base para a criação de índices de vegetação, como o NDVI, que será abordado no capítulo seguinte.



(a) Vegetação senescente, em fase de deterioração

(b) Vegetação viva, das espécies subarbustivas *Holcus*, *Molinia e Juncus*

Figura 2.2: Reflexão espectral de dois tipos de vegetação - *Retirado de Curran et al. 1970*

Existem vários fatores que limitam a informação disponibilizada pelos satélites, além das bandas que são disponibilizadas, há a limitação da resolução espacial, que define o tamanho real do elemento mais pequeno capturado (um pixel na imagem gerada). Os dados são também limitados pela dimensão temporal, sendo que cada satélite tem um período de revisita, fixo ou não, que define a frequência com a qual são geradas imagens para um dado local. A resolução temporal é particularmente importante quando se pretende analisar a variação ao longo do tempo.

2.3 Conclusões

A informação de deteção remota demonstra ser uma forma fiável de avaliar as características da vegetação no solo que, aliado às definições das FGCI, permitirá estimar o estado da vegetação dentro destas faixas. É importante, no entanto, que as leituras feitas pelos

satélites sejam precisas, sendo que a presença de nuvens é um grande impedimento na avaliação das características do terreno.

As FGCI têm regras diferentes consoante o tipo de faixa, e o tipo de vegetação nela presente, como é o caso de espécies como os eucaliptos, é importante então não abordar as faixas como um todo, mas ter em conta as diferentes características das zonas das faixas e que as mesmas podem sofrer tipos de intervenção diferentes consoante a região e as características da área.

3

Dados Relevantes

3.1 Informação Geográfica

3.1.1 PMDFCI

Os dados oficiais dos Planos Municipal da Defesa da Floresta Contra Incêndios (PMDFCI) estão disponibilizados na plataforma da ICNF [25], sendo possível obter esta informação através do website, individualmente, ou através do servidor FTP que disponibiliza todos os planos organizados por Distrito e Concelho.

Estes planos, além da definição das Faixas de Gestão de Combustível, incluem relatórios de diagnóstico, que têm como objetivo caracterizar o terreno, a população e a atividade histórica de incêndios florestais. Nesta secção inclui-se, por exemplo, o declive do terreno, a hidrografia, a densidade populacional, entre outros detalhes. O foco desta dissertação recai sobre a informação do plano de ação, mais concretamente, das faixas de gestão de combustível.

Existem diferenças entre os dados fornecidos por cada concelho, nomeadamente sobre a informação que é fornecida sobre as faixas, e a atualização dos planos, sendo que, à data de escrita, o plano de Peniche tinha sido atualizado pela última vez em 2007, enquanto outros planos, como o de Miranda do Corvo, teriam sido atualizados em 2021.

Podemos, de uma forma geral, descrever a informação fornecida nos dados georreferenciados das Faixas de Gestão de Combustível, ainda que hajam algumas discrepâncias a apontar. Esta informação é fornecida no formato de *shapefiles* que definem a geometria das faixas, tal como os atributos associados, que constam na tabela 3.1.

É importante notar que nem todas as FGCI incluem todos estes atributos, num caso concreto nem existe um atributo identificador da faixa. E pelo outro lado, muitas incluem atributos adicionais que não constam do guia de PMDFCI. Sem nenhum tipo adicional de metadados ou informação das autoridades que geraram estas faixas não é possível considerar o conteúdos destes dados adicionais.

Os atributos mais importantes a notar são o ID_R_FGC, que nos permite discretizar as faixas, e o DESC_FGC que nos permite identificar o tipo de faixa, e consequentemente, as regras de manutenção a que ela está sujeita.

Tabela 3.1: Atributos principais presentes nos atributos das FGCI

Campo	**Descrição**	**Exemplo**
ID_R_FGC	Identificador da FGC	26
ID_S_FGC	ID de secção da FGC	26.12
DATA_ACCAO	Data do levantamento das características	07-07-2018
COD_INE	Código referente a freguesia	131420
DESC_FGC	Código de descrição da FGC	2 (aglomerado populacional)
TIPO_FGC	Tipo de FGC (Redução ou Interrupção)	FRC (faixa de redução)
OBJEC_FUNC	Objetivo da FGC	3 (Isolar focos potenciais de incêndios)
AREA	Área da secção da faixa (em hectares)	0,35
RESP_GC	Identificação do responsável pela gestão de combustível	Privado
INTER_AAAA	Tipo de intervenção a realizar num dado ano	QQQ (Gestão com fogo controlado)
EXEC_AAAA	Meio de execução das faixas num dado ano	1 (Equipa de Sapadores Florestais)

3.1.2 Open Street Maps

O *Open Street Map* é um projeto colaborativo que recorre a contribuições públicas para criar uma base de dados geográfica do mundo. A informação que os mapas da OSM inclui são inúmeras, incluindo habitações, lojas, estradas, redes ferroviárias, e muitas outras, tal como os atributos relativos a cada[1]. A informação que é disponibilizada pela plataforma é disponibilizada sobre uma licença de uso livre, o que é um incentivo para a sua utilização, dado a ausência de custos, e haver um maior grau de liberdade sobre a forma como é possível usar os dados fornecidos. Para o contexto desta tese são particularmente relevantes as definições de estradas, e outras redes viárias, de forma a evitar que as áreas florestais em análise não intersetem com estas zonas, influenciando a leitura dos sensores de deteção remota.

Na figura 3.1 estão representados os vários elementos que compõem o mapa da OSM, incluindo estradas, habitações, zonas verdes, entre outros.

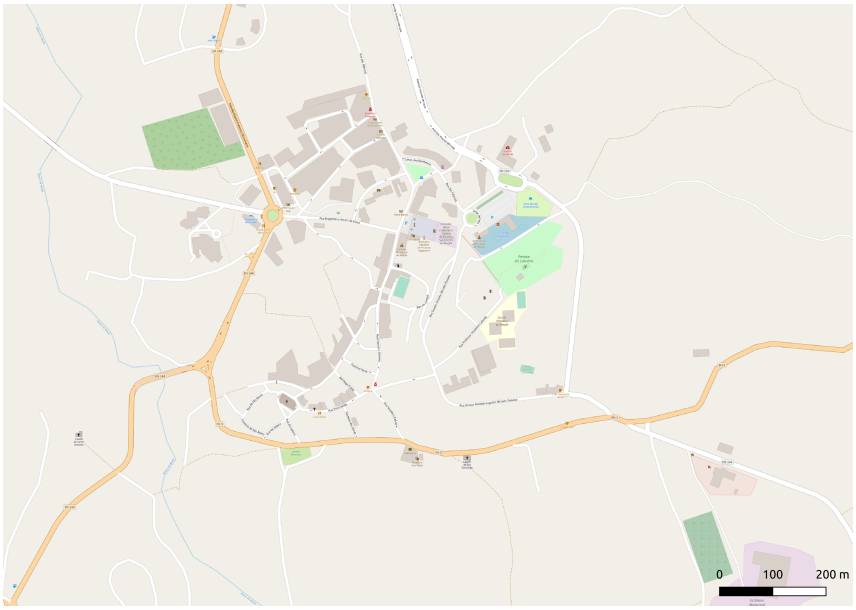


Figura 3.1: Mapa fornecido pelo OSM na região da vila de Mação

3.1.3 Carta de Ocupação do Solo

A COS é produzida pela Direção-Geral do Território, e constitui uma série temporal de cinco anos de referência (1995, 2007, 2010, 2015 e 2018) [41] . Esta carta baseia-se na interpretação de ortofotos de diversas fontes para identificar, em zonas de 1 hectare ou mais, o tipo de ocupação do solo presente. A classificação destas áreas divide-se em 9 classes principais, das quais, territórios artificiais, agricultura, pastagem, florestas, entre outros, e que se subdividem em sub-classes mais específicas, 83 no total. As classes no nível mais detalhado distinguem categorias como olivais, praias ou jardins, por exemplo. [42]

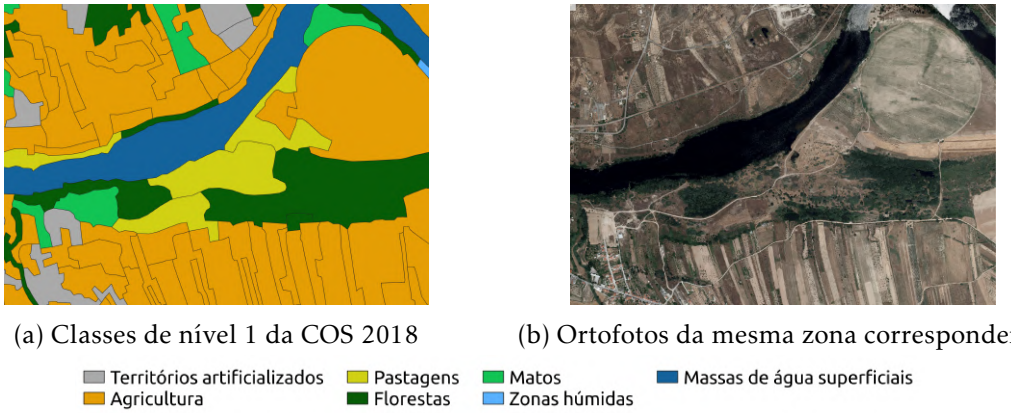
Para o objetivo final desta tese de dissertação considera-se particularmente relevante a classificação de diferentes tipos de vegetação para contextualizar o seu crescimento ao longo do tempo, e ajudar a identificar variações.

Na figura 3.2 podemos comparar as classes de nível 1 do COS na região de Abrantes com as Ortofotos disponibilizadas pela DGT.

3.1.4 Ortofotos

A Direção Geral do Território (DGT) disponibiliza um conjunto de ortofotos geradas através de fotografia aérea que cobrem todo o território nacional[14][13]. Estas ortofotos têm sido disponibilizadas desde 1995, sendo que a última captura foi em 2018, tendo uma resolução de 25 cm, e são particularmente relevantes para o objetivo desta dissertação devido à sua alta resolução, que permite mais facilmente identificar as áreas de estudo.

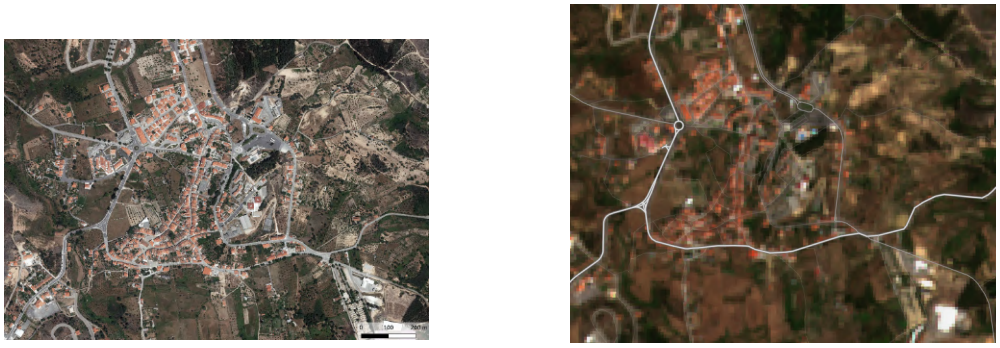
Na figura 3.3 estão duas imagens da mesma região, a vila de Mação, uma foi obtida a partir das ortofotos da DGT, outra a partir das capturas do Sentinel-2. A resolução é



(a) Classes de nível 1 da COS 2018 (b) Ortofotos da mesma zona correspondente

Figura 3.2: Comparação entre ortofotos e da COS, na região de Alvega, Abrantes

muito superior na primeira imagem, e permite discernir muitos mais detalhes sobre o terreno e a ocupação de solo.



(a) Ortofotos de 2018 da vila de Mação (b) Imagem do Sentinel-2 da vila de Mação (estradas marcadas para referência)

Figura 3.3: Comparação entre imagens das ortofotos da DGT e do Sentinel-2 para a mesma região

3.2 Dados de Referência

3.2.1 Mação

A Câmara Municipal de Mação disponibilizou um conjunto de dados referentes ao ano de 2021 que identifica as áreas que sofreram ações de intervenção no concelho de Mação. Esta informação refere-se às intervenções sobre faixas de aglomerados populacionais, de habitações, de redes primárias e viárias. O foco desta dissertação incide sobre as faixas de aglomerados populacionais e de habitações, portanto serão seleccionadas essas faixas para posterior transformação e análise.

A informação foi fornecida em 9 *shapefiles* separados, sendo que cada um tem um conjunto de atributos relativo ao tipo de faixas que contêm, nomeadamente identificadores, ou, por exemplo, a entidade responsável pela manutenção da faixa. É importante notar

Tabela 3.2: Distribuição dos valores do atributo 'exec' pelas faixas

	Faixas de Habitações/Aglomerados	Total
Data de Execução	115	320
Em Execução	50	116
Por Executar	7	22

que estas faixas fornecidas nos dados de referência não correspondem à totalidade das faixas tal como definidas no PMDFCI, podendo representar apenas uma parte das faixas oficiais.

Os diferentes ficheiros georreferenciados são unidos em um só, preservando todos os seus atributos. Pretende-se, a partir deste ponto, criar um novo identificador que separe as diferentes faixas, dado que no seu estado original as geometrias estão fragmentadas. Por cada habitação, ou localidade, pretende-se que apenas seja identificada uma faixa em seu redor, e que cada faixa apenas consista de um tipo de FGCI (por exemplo, não pode haver nenhum fragmento de uma faixa viária nas faixas habitacionais). Foram analisados os atributos, ou conjuntos de atributos, que cumpriam estes requisitos para cada ficheiro, e gerou-se um novo identificador através da concatenação do valor desses atributos.

Depois de unir as faixas por este novo identificador global, contamos com 458 faixas que têm o estado de execução devidamente identificadas, sendo que destas, 172 correspondem a faixas de aglomerados populacionais ou habitações.

O atributo mais relevante desta informação geográfica é o atributo *"exec"*, que indica a data limite para as ações de intervenção, cuja distribuição é apresentada na tabela 3.2.

É importante mencionar que os valores da coluna *exec* não identificam necessariamente a data exata na qual uma zona foi intervencionada. Este atributo define sim a data limite para as ações de intervenção, sendo que as reais datas de intervenção habitualmente ocorrem algumas semanas antes. Existem várias exceções à regra, com ações a serem feitas um pouco depois da data limite, ou mais de um mês antes do limite.

À primeira vista pode-se entender que existe uma pequena quantidade de informação relativamente às faixas por executar, no entanto muitas das faixas têm grandes dimensões (evidenciado pelos valores da tabela 3.3), e portanto podem ser subdivididas em várias áreas de análise, áreas estas que podem ter comportamentos diferentes, tipos de vegetação diferentes, e cujas intervenções podem ser mais, ou menos, acentuadas. Permitindo extrair informação mais detalhada a partir de uma única faixa.

Tabela 3.3: Distribuição da áreas das geometrias dos dados de referência

	Área (m^2)	**Nº de Faixas de Habitações**	**Nº de Faixas de Localidades**	**Nº Total de Faixas**
0	**0-1.000**	0	0	0
1	**1.000-2.000**	5	0	5
2	**2.000-4.000**	24	0	24
3	**4.000-8.000**	26	5	31
4	**8.000-16.000**	18	5	23
5	**16.000-32.000**	6	10	16
6	**32.000-64.000**	0	14	14
7	**64.000-128.000**	0	7	7
8	**128.000+**	0	2	2

3.2.2 Santarém

Foi disponibilizado pela Câmara Municipal de Santarém um conjunto de dados de referência que englobam o concelho de Santarém, relativos aos anos de 2018 a 2021. Representando faixas em zonas junto de e no interior de aglomerados populacionais tal como junto a faixas viárias. O formato destes dados é substancialmente diferente daquele da área de Mação, isto porque a informação relativamente às intervenções apenas indica, para um dado ano, que zonas foram intervencionadas, não sendo nestes dados referido a data de intervenção, ou de limite de conclusão dos trabalhos. Para cada um destes anos foram fornecidos dois *shapefiles*, um para as zonas viárias e outro para as restantes faixas. Cada um dos polígonos que define uma zona intervencionada num dado ano tem um conjunto de atributos que descrevem o tipo de intervenção, algumas observações (por exemplo indicando se a limpeza não foi totalmente concluída por algum impedimento), um identificador.

Para posteriormente fazer uma análise do comportamento destas zonas a serem intervencionadas ao longo do tempo, é preciso definir áreas de estudo para a partir das quais poder extrair séries temporais que descrevam o comportamento da vegetação. À primeira vista os polígonos têm identificadores que nos permitiriam correlacionar os polígonos dos diferentes anos e extrair séries para estes anos. No entanto, os identificadores para polígonos equivalentes entre vários anos são diferentes, o que impossibilita essa abordagem. Assim sendo, é preciso analisar as geometrias das áreas intervencionadas nos vários anos para poder identificar os polígonos equivalentes entre diferentes anos. Para tal, é necessária alguma flexibilidade, dado que existe um desalinhamento dos polígonos entre os vários anos, como demonstra a figura 3.4, e portanto, a sobreposição da mesma zona intervencionada entre vários anos não é perfeita.

Outro fator relevante é o facto de as faixas não terem de ser totalmente intervencionadas todos os anos, por exemplo, num ano podemos ter a faixa totalmente intervencionada, e no seguinte, apenas metade sofrer ação de limpeza.

O exemplo do desalinhamento das faixas entre vários anos da figura 3.4 demonstra como se torna difícil estabelecer uma correlação entre as faixas nos vários anos e como

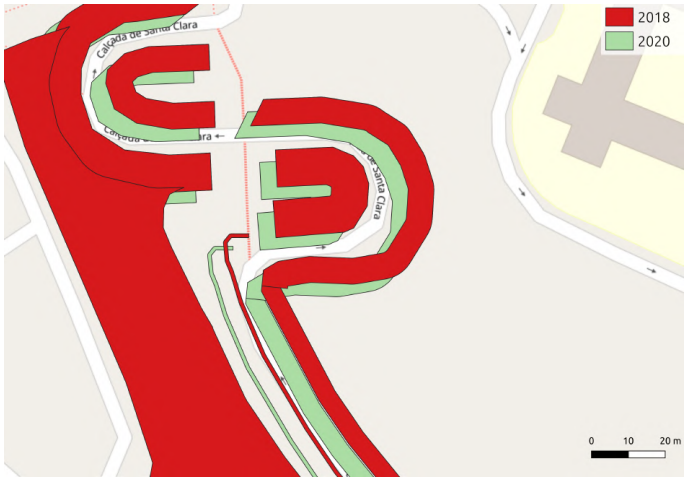
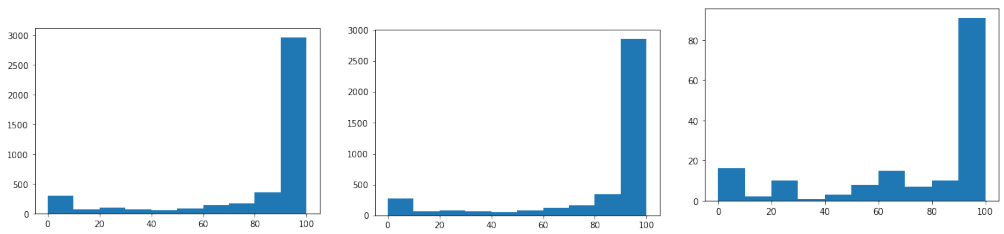


Figura 3.4: Exemplo do desalinhamento de faixas equivalentes entre vários anos

este desalinhamento pode ser prejudicial para as leituras, dado que, para este caso, as faixas de 2020 parecem ter uma forte sobreposição sobre as estradas, o que demonstra que não estão alinhadas com a zona de intervenção.

Nos dados de referência de Mação todos os polígonos que definem uma zona que foi, ou não, intervencionada estão perfeitamente incluídos nas geometrias do PMDFCI do município. Pretendeu-se entender se o mesmo se verificava para os dados de referência de Santarém. Foi, para tal, feita uma análise da sobreposição dos dados de referência com as faixas definidas no PMDFCI, obtendo os seguintes resultados, apresentados na figura 3.5.



(a) Sobreposição de todos os dados (%)

(b) Sobreposição nas faixas viárias (%)

(c) Sobreposição nas faixas de aglomerados populacionais (%)

Figura 3.5: Histogramas da sobreposição entre os dados de referência e as faixas definidas no PMDFCI de Santarém

Conclui-se que existe uma sobreposição bastante significativa entre os dados de referência e as faixas do PMDFCI, sendo esta mais pronunciada nas faixas ao redor de estradas. É importante no entanto ter em conta que as faixas têm uma grande diversidade de áreas, e como tal, pode na verdade existir uma área sem sobreposição maior do que os anteriores histogramas sugerem. Tendo isso em conta gerou-se, para as faixas de aglomerados populacionais, um histograma que usa a área da zona de intervenção como peso, obtendo os resultados apresentados na figura 3.6.

Na figura 3.6 é possível entender que existe uma área bastante significativa que tem uma baixa sobreposição com o PMDFCI, sendo que mais de metade da área das zonas de intervenção corresponde a zonas que têm uma sobreposição com o PMDFCI de menos de 70%. É importante notar também que, mesmo existindo uma dada sobreposição com as faixas do plano de defesa, essa pode corresponder a uma faixa não relacionada com a zona de intervenção, por exemplo uma estrada adjacente a uma zona habitacional.

Não excluímos, no entanto, a possibilidade de problemas do alinhamento entre os polígonos dos vários anos serem devidos a algum erro deste trabalho no processamento das geometrias.

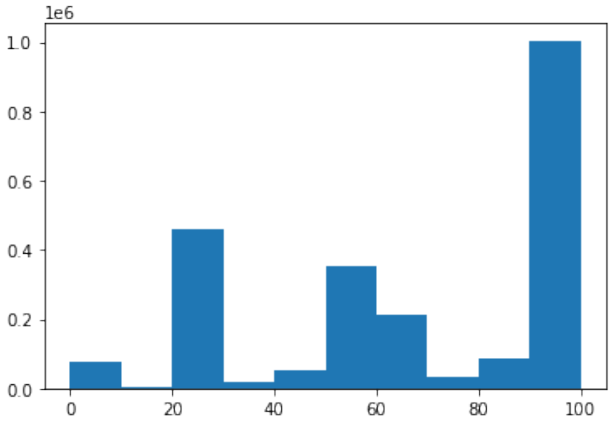


Figura 3.6: Histograma da sobreposição (%) entre os dados de referência e as faixas definidas no PMDFCI de Santarém (tendo em conta a área das faixas)

Como mencionado anteriormente, as zonas de intervenção encontram-se desalinhadas entre os vários anos. Porém, após aplicar uma abordagem que correlaciona os polígonos entre os vários anos, obtemos a seguinte dispersão de valores para as faixas de intervenção, representada na tabela 3.4. Neste contexto faixa indica um polígono nos dados de referência, o que explica o maior número de faixas viárias, dado que as geometrias destas se encontram mais fragmentadas. Cada entrada nesta tabela representa o número de polígonos que, para cada ano, sofreu, ou não, uma intervenção de acordo com os dados de referência.

Tabela 3.4: Distribuição do estado de intervenção das faixas ao longo dos 4 anos do intervalo de estudo

	Faixas Viárias	Restantes Faixas
Faixas-anos executadas	525	104
Faixas-anos por executar	451	90

Após dividir estas faixas em fragmentos mais pequenos obtemos a tabela 3.5, cuja diferença de valores para a tabela anterior é justificada pela área superior das faixas dos aglomerados populacionais. São estes fragmentos, e as suas séries temporais para cada ano que serão utilizados para estimar o estado de limpeza das faixas.

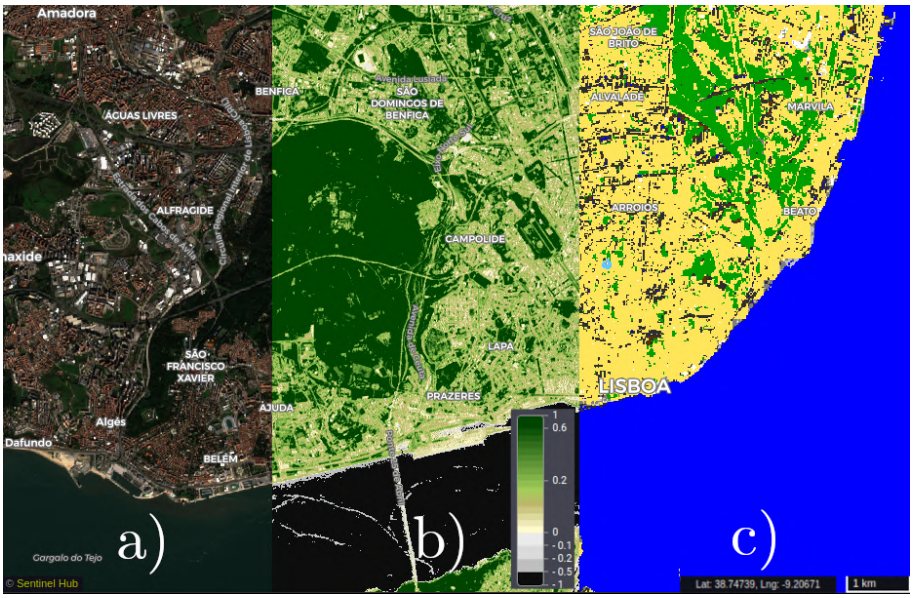
Tabela 3.5: Distribuição do estado de intervenção dos fragmentos das faixas ao longo dos 4 anos

	Faixas Viárias	Restantes Faixas
Faixas-anos executados	3860	2582
Faixas-anos por executar	2130	985

3.3 Produtos de Remote Sensing

Os dados recolhidos pelos sensores presentes nos satélites não são, na maioria dos casos, diretamente disponibilizados ao utilizador, nem tal seria desejado para fazer análise dos mesmos. As entidades que fornecem esta informação oferecem diferentes níveis de processamento sobre as imagens para facilitar o seu uso e para corrigir eventuais erros. Por exemplo, a atmosfera reflete parte da radiação do sol e essa reflexão é capturada pelos sensores, contaminando os valores lidos da superfície terrestre. São disponibilizadas imagens que corrigem estes efeitos atmosféricos para uma leitura mais precisa da superfície. Num nível superior, mapas de cobertura de nuvens, de vapor de água e até da classificação estimada do terreno (distinção entre superfícies de água, vegetação, neve, entre outras), podem ser incluídos.

Na figura 3.8c comparam-se 3 tipos de visualização diferentes, a primeira e última são fornecidos diretamente pela Copernicus como parte do produto de nível 2A do Sentinel-2, e a central é uma visualização calculada *a posteriori* pelo utilizador com base nos níveis das bandas de infra-vermelhos e vermelho.



(a) Luz Visível (bandas *Red*, *Green*, *Blue*)

(b) Indíce de Vegetação (Favorece Infravermelhos e penaliza Vermelho)

(c) Scene Classification Map (SCM)

Figura 3.7: Comparação de 3 visualizações do Sentinel-2

3.3.1 Requisitos

A escolha das imagens de satélites a serem usadas é uma decisão que deve ter em conta o contexto do problema a resolver. As FGCI podem ter larguras relativamente pequenas, de 10 metros em cada lado no caso das estradas, por exemplo. Para fazer uma deteção apropriada do estado da faixa é necessário que a resolução espacial esteja, no mínimo, próxima destas dimensões. Relativamente à componente temporal é preferível que a informação seja atualizada o mais frequentemente possível, no entanto, apesar de não ser o desejável, períodos mais longos, abrangendo por exemplo, mais que uma semana, retornariam resultado úteis ainda assim e teriam um impacto positivo na fiscalização das FGCI. Ao nível espectral e de funcionamento do sensor, existem vários parâmetros que estão associados à presença de vegetação e que poderiam ser usados para detetar intervenções. Dentro das bandas de um sensor multiespectral os valores do espectro infravermelho próximo (*Near Infrared - NIF*) são muito usados para derivar características da vegetação, por exemplo [45, 7]. As imagens de sensores ativos, como os SAR são frequentemente utilizados para complementar a informação multiespectral, detetando características como a forma, textura e humidade, que são mais difíceis de serem extraídas de um sensor passivo [36, 39].

Com base na informação previamente descrita selecionou-se como principal fonte de dados remotos os satélites que melhor se enquadram dentro dos requerimentos, os satélites da missões Sentinel-2

3.3.2 Sentinel 2

Sentinel-2 é uma missão do programa Copernicus, dirigido pela Agência Espacial Europeia (ESA). Esta missão é composta por dois satélites de composição idêntica, o Sentinel-2A e o Sentinel-2B, ambos equipados com sensores multiespectrais. Estes satélites têm uma órbita que é completada a cada 10 dias, e que é síncrona em relação ao sol, isto para garantir que a incidência da radiação solar é semelhante entre as várias capturas, e que o impacto das sombras nas imagens é o mínimo possível. O Sentinel-2A e 2B têm a mesma órbita, mas encontram-se separados por 180 graus, fazendo com que o produto extraído das capturas de ambos tenha um período de 5 dias em vez de 10.

Os sensores multiespectrais desta missão captam dados em 13 bandas, com diferenças mínimas entre cada satélite.

3.3.3 Pré-Processamento

As características da atmosfera, do terreno, dos sensores e das órbitas têm influência nos valores que irão ser capturados, distorcendo os dados que pretendemos ler. Portanto, para fazer uma análise detalhada e precisa da superfície é necessário transformar os dados captados através de algumas etapas de pré-processamento.

Tabela 3.6: Bandas capturadas pelos sensores da missão Sentinel 2

Banda	Comprimento de Onda Central (nm)	Largura da Banda (nm)	Resolução Espacial (m)
1 - Aerosol	442.7	21	60
2 - Azul	492.4	66	10
3 - Verde	559.8	36	10
4 - Vermelho	664.6	31	10
5 - RE1	704.1	15	20
6 - RE2	740.5	15	20
7 - RE3	782.8	20	20
8 - NIR	832.8	106	10
8A - NIR_A	864.7	21	20
9 - Vapor de Água	945.1	20	60
10 - SWIR	1373.5	31	60
11 - SWIR	1613.7	91	20
12 - SWIR	2202.4	175	20

Erros nas leituras por parte dos sensores, como pontos-mortos, diferenças de sensibilidade, ou diferentes incidências do sol em determinadas etapas da órbita contribuem para que exista alguma discrepância entre os valores, tanto na dimensão temporal, como espacial. Este tipo de erros é normalmente corrigido pelas entidades que fornecem os dados de deteção remota, corrigindo e normalizando os valores para assegurar o maior nível de fiabilidade e consistência temporal. Outro problema frequente é a interferência da atmosfera nas leituras, a composição da atmosfera ao longo do tempo, e em cada local de captura varia, o que pode colocar em causa a consistência espacio-temporal dos dados obtidos.

Parte do processamento que é feito, normalmente pelas entidades que fornecem os dados, é a georreferenciação das imagens, permitindo associar cada pixel da imagem a um ponto na superfície terrestre. No caso do projeto em mão, da deteção do estado das FGCI, uma faixa pode ter apenas 10 metros de largura, precisamente o tamanho de um pixel nas bandas de melhor resolução do Sentinel-2. Um erro de georreferenciação de alguns metros é suficiente para que a leitura seja imprecisa e afete a precisão do modelo. No caso do Sentinel-2 é esperado que os erros sejam inferiores a 11 metros num intervalo de confiança de 95% [15]. Para corrigir estes pequenos erros pode ser necessário recorrer a outras imagens de satélite, de melhor resolução para alinhar os produtos e assegurar que a referenciação seja o mais precisa possível.

Outro aspeto do pré-processamento corresponde às transformações que não pretendem colmatar nenhum erro de captura, como é o exemplo, da reamostragem, frequentemente usada para colocar as imagens de todas as bandas numa mesma resolução comparável.

3.3.4 Índices Espectrais

Os índices espectrais são calculados com base nos valores de radiação em diversas bandas de forma a estimarem certas características da vegetação. Abaixo listam-se os índices que

foram frequentemente referidos na literatura e que se correlacionam com as características vegetativas que se pretendem estudar (biomassa, cobertura da vegetação, conteúdo de clorofila, entre outros):

3.3.4.1 Normalized Difference Vegetation Index (NDVI)

O NDVI é o índice mais utilizado nesta área, estando correlacionado com a presença de vegetação saudável, numa escala de -1 a 1. É calculado usando a banda vermelha e o infravermelho próximo, que corresponde, no sentinel-2 às bandas Vermelho e NIR. Calculado da seguinte forma, como é evidenciado pela figura 2.2:

$$NDVI = \frac{NIR - Vermelho}{NIR + Vermelho}$$

Este índice é um dos mais populares índices para avaliar o estado da vegetação, havendo trabalho prévio que explora a sua utilização para a monitoração de FGCI [34, 5, 3], tal como para avaliar a variação natural da vegetação ao longo do tempo[7] e estimar o risco de incêndio[40].

3.3.4.2 Inverted Red-Edge Chlorophyll Index (IRECI)

O IRECI tem uma maior dependência nas bandas *Red-Edge* do Sentinel-2, evitando que o seu peso recaia sobre a banda Vermelha, e evitando assim que atinga saturação mais rápido. Este índice está fortemente correlacionado com a cobertura das copas, com a quantidade de clorofila e com a biomassa total da vegetação, segundo estudos que analisaram áreas florestais e agrícolas na Sicília, em Barrax, na Espanha, nas Filipinas e na China [20, 9, 27, 16].

A fórmula que o descreve usa as bandas do Sentinel-2, da seguinte forma:

$$IRECI = \frac{(RE3 - Vermelho)}{RE1/RE2}$$

3.3.4.3 Normalized Difference Index (NDI45)

O NDI45 tem a mesma fórmula que o NDVI, embora use diferentes bandas do Sentinel-2 para o implementar, usando a banda RE1 (Banda 5) como substituta do NIR. Como tal, a fórmula é a seguinte:

$$NDI45 = \frac{RE1 - Vermelho}{RE1 + Vermelho}$$

Este índice, tem muitas semelhanças de aplicabilidades ao IRECI, estando também fortemente correlacionado com variáveis biofísicas como a cobertura das copas, e a biomassa. [20]

3.3.4.4 Normalized Burn Ratio (NBR)

O NBR é um índice criado para destacar zonas queimadas em incêndios florestais. A fórmula apresenta semelhanças ao NDVI, mas combina tanto as bandas Red-Edge como as bandas SWIR (infravermelho de onda-curta). Vegetação saudável e densa apresenta valores altos e terrenos áridos e queimados um valor tendencialmente mais baixo.

Além da banda SWIR ter a capacidade de melhor penetrar a atmosfera e ser menos perturbadas por interferências como fumo [33], ela é relevante para o caso de estudo pois um dos métodos usados para efetuar a manutenção das FGCI é por meio de queimas controladas. [26]

$$NBR = \frac{NIR - SWIR}{NIR + SWIR}$$

4

Trabalho Relacionado

Existem vários trabalhos prévios na área da deteção remota que estão associados à análise dos níveis de vegetação com recurso a imagens de satélite, sendo que alguns destes estão diretamente associados à monitoração das FGCI.

A análise do estado da vegetação, e das variáveis biofísicas associadas é um tema muito explorado na literatura, sendo que os dados dos satélites Sentinel-2 são frequentemente citados como a fonte de informação remota, mostrando como o uso de índices espectrais permite estimar estas características e identificar as alterações do seu comportamento ao longo do tempo.

O uso do Sentinel-2 e dos seus índices encontra-se também referido em trabalhos prévio que pretendem analisar o estado de vegetação das FGCI, ou criar modelos de classificação automática para identificar ações de intervenção sobre as mesmas. Os resultados destas experimentações são, no geral, encorajadores.

Nesta secção abordamos este conjunto de trabalho prévio que está mais intimamente relacionado com o objetivo desta dissertação.

4.1 Análise de Faixas de Gestão de Combustível

Existem vários trabalhos na literatura com o propósito de avaliar o estado da vegetação [46, 21, 7], sendo que se encontraram poucos exemplos que realizam esta análise no âmbito concreto das FGCI. Estes trabalhos focam-se maioritariamente em Faixas Primárias, zonas habitualmente de floresta mais densa, com uma largura de pelo menos 125 metros.

4.1.1 Avaliação do estado das faixas

Um trabalho recente foca-se nas faixas primárias, cujo objetivo é o de avaliar o estado relativo da vegetação nestas faixas. Entenda-se por estado relativo o estado comparativamente ao nível de vegetação pré-intervenção e pós-intervenção. Relativamente aos dados usados, o índice NDVI do Sentinel-2 foi usado como base para esta análise, tendo sido comparado com os dados da altura de copas da missão GEDI, uma missão de *sampling*, que cobre 4% da superfície terrestre [34].

O processo passou por suavizar a série temporal de NDVI com uma média móvel (com uma janela de três meses), de seguida definir os valores mínimos e máximos de NDVI para atuarem como referência para cada faixa, e finalmente estimar o crescimento de vegetação a um nível mensal.

Este valor mínimo é o valor após a intervenção, antes de a vegetação crescer, e o máximo é calculado com a média dos 6 meses antes do início da intervenção.

O cálculo do crescimento da vegetação é feito em comparação com os valores do ano anterior, sendo esta comparação feita a cada mês. O cálculo do estado da vegetação resulta da soma destes valores de crescimento ao longo dos meses. Limitando os valores entre a referência de mínimo e máximo (0% equivale à referência mínima e 100% à referência máxima).

Os resultados são positivos, refletindo o crescimento da vegetação, contextualizado aos valores que são esperados em cada faixa, no entanto, a análise feita é de longo-prazo, sendo que, por exemplo, nestas faixas primárias é esperado que seja necessária uma nova intervenção a cada 24 a 36 meses, de acordo com o ICNF.

Neste trabalho demonstra-se também uma correlação entre o NDVI e altura das copas medida pelo GEDI, que reforça a utilidade deste índice de vegetação para a estimação do estado das faixas. No entanto, é de notar que o trabalho desenvolvido tem a sua utilidade limitada às faixas primárias, dado que a análise apenas iniciar-se um ano após o fim de uma dada intervenção, e que muitas faixas viárias e habitacionais precisam de uma ou mais intervenções por ano.

4.1.2 Identificação de intervenções em Faixas Primárias

Existem outros estudos cujo foco é a identificação de momentos de intervenção sobre faixas primárias de gestão de combustível de incêndios [5, 35]. Estes têm abordagens diferentes, tanto quanto à forma como os dados são recolhidos, como à modelação dos mesmos.

Resumindo, os aspetos que são comuns a ambos são a comparação com a vegetação adjacente à faixa de interesse, o uso de índices de vegetação (dos quais se destaca o NDVI), e o uso de contexto temporal para fazer a classificação.

Uma abordagem explorada usa redes neuronais para identificar os instantes de intervenção [35]. São extraídos dados de diversas bandas do Sentinel-2, tal como alguns índices de vegetação, como o NDVI, o NDMI e o EVI, entre outros. A forma como é introduzida o contexto temporal nos dados de treino do modelo é através da concatenação do valor do mês anterior a cada mês.

Foi usado o algoritmo de correlação de Pearson para identificar as *features* redundantes e eliminá-las antes do processo de aprendizagem. O modelo de aprendizagem, foi desenhado de forma empírica, tendo sido feito experiências variando o número de camadas escondidas entre uma e duas, e o número total de neurónios da rede. Os resultados foram avaliados com um conjunto de avaliação, que não esteve envolvido no processo

de treino, sendo que o melhor modelo teve um F1-Score de 70% e uma precisão de 57%, sendo mencionado pelos autores que o valor baixo da precisão é um custo aceitável para que os falsos positivos sejam mais raros.

É importante notar que o conjunto de dados deste último trabalho é pequeno, sendo analisadas apenas 10 FGC, todas elas faixas primárias.

Uma outra abordagem é única pelo seu aspeto não-supervisionado [5]. Usando a comparação dos valores do interior da faixa e exteriores, são derivadas duas séries temporais (cujos valores seleccionados são inter-quartis) e a cada instante temporal definem-se duas janelas, uma anterior ao instante e outra posterior (com uma dimensão até 2 meses). Estas janelas são a base para a aplicação do teste welsh-t que permite automaticamente determinar as datas com maior significância (comparando a janela anterior e posterior), sendo que são seleccionadas as datas que se mostram significantes nas séries do interior da faixa, nas séries de diferença entre o interior e exterior, e não nas séries do exterior da faixa.

4.1.3 Trabalho do grupo SI-MORENA

Em relação à averiguação do estado de limpeza das FGCI, o trabalho de Ricardo Afonso, em 2019 [3] é o mais relevante, abordando este problema ao tentar classificar o estado de intervenção sobre as faixas em torno das vias de comunicação. Foram comparadas diferentes abordagens, relativamente aos modelos de aprendizagem, aos dados usados e à incorporação da componente temporal na avaliação do estado de intervenção. Recorreu-se também a métodos de seleção de características na tentativa de reduzir a dimensão dos dados sem comprometer significativamente os resultados.

Os dados a serem usados foram divididos em 4 grupos:

- **CD_TUDO** - Toda a informação abaixo referida
- **CD_ÍNDICES** - Dados Sentinel-1 + Índices de Vegetação (NDVI, NDWI, SAVI, IRECI, SR)
- **CD_FUSÃO** - Dados Sentinel-1 + Índices NDVI e NDWI + Bandas verde, vermelho, *red-edge* e *NIR* (Bandas 3 a 8).
- **CD_BANDAS** - Dados Sentinel-1 + Dados Sentinel-2

Neste trabalho foram comparados diferentes algoritmos de aprendizagem automática, incluindo *Gradient Boosting* (XGBoost), *SVM*, *KNN* e *Random Forests* (RF). Todos estes métodos foram treinado com os quatro datasets apresentados. Os resultados foram bastante promissores, obtendo precisões muito elevadas, com valores Kappa acima de 0.88 em determinadas circunstâncias. Existe, no entanto, uma grande variação de resultados tendo em conta o tipo de FGCI que está a ser avaliada, sendo que ao redor das linhas elétricas o melhor resultado obteve um valor Kappa de 0.74, o que se justifica pelo facto das faixas em seu redor terem uma largura menor.

Os modelos de aprendizagem automática foram treinados em dois contextos temporais. Num era fornecida a informação de um único instante no tempo (análise estática), sendo fornecido o valor médio do interior da faixa, tal como o valor médio do exterior, para motivos de comparação. No outro eram fornecidas as últimas 5 capturas, com o intuito de permitir ao modelo detetar variações significativas (análise temporal). Esta última abordagem obteve os melhores resultados, sendo que o resultado obtido foi essencialmente igual entre os métodos RF, SVM e XGB.

Pode-se concluir que, de uma forma geral, o uso de um conjunto de dados total com o algoritmo XGBoost obteve dos melhores resultados de uma forma consistente, ainda que em muitas ocasiões a diferença de resultados para os restantes algoritmos fosse muito reduzida.

A maior dimensão dos dados exige um maior tempo de computação para treinar os modelos de aprendizagem, e como tal, poderá ser vantajoso usar datasets mais reduzidos, ou recorrer a métodos de seleção de características como o LASSO que identificam as características mais relevantes para um dado objetivo. Este método de seleção foi aplicado, tendo resultado na diminuição de características num fator superior a 3 vezes. Esta modificação do conjunto de dados traduziu-se numa redução no F1-Score das classificações de apenas 0.005 para o melhor classificador.

O trabalho aqui descrito, embora seja promissor, tem pormenores limitadores que merecem referência. A escolha das FGCI a analisar foi reduzida, englobando faixas ao longo de estradas e linhas elétricas, sendo que as faixas que rodeiam as habitações e localidades não constaram da análise por falta de informação das datas de intervenção. O tipo de faixas analisado neste trabalho têm uma peculiaridade de, no geral, permitirem ser comparadas com a vegetação diretamente adjacente à faixa, o que permite normalizar os valores, e gerar séries temporais que melhor refletem comportamentos anormais nas tendências da vegetação. Esta normalização, que também foi usada em outro trabalhos relacionados com a análise de FGCI é mais dificilmente aplicada a faixas de localidades e habitações, pois as áreas das faixas são diferentes, englobando diferentes tipos de ocupação de solo, e as zonas adjacentes são menos homogéneas, havendo menos garantias que são comparáveis às áreas interiores da faixa.

O processo de normalização acima descrito depende da criação de buffers que rodeiam as faixas, como mostra a figura 4.1 para cada zona de análise de uma faixa.

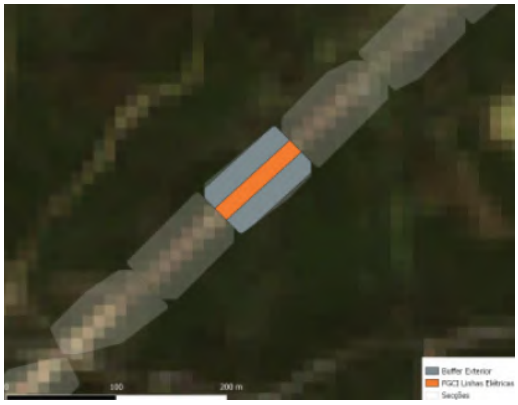


Figura 4.1: FGCI em redor de uma linha elétrica (a laranja) e o buffer adjacente (a cinzento) para efetuar o cálculo de normalização - *Retirado de [3]*

4.2 Deteção automática de estruturas artificiais

Num trabalho desenvolvido em 2019 por André Neves mostrou-se a viabilidade do uso de técnicas de aprendizagem automática para a deteção remota de estruturas artificiais com recurso a imagens de satélite das missões Sentinel-1 e Sentinel-2.

Os métodos de aprendizagem, dos quais se destacou o XGBoost, utilizando a informação de deteção remota e mapas de elevação de terreno são treinadas para identificar cada píxel como pertencendo a uma determinada classe. Foram efetuadas 3 experiências, com o objetivo de identificar diferentes conjuntos de classes, sendo que na sua totalidade existem 6 classes a identificar, **estruturas artificiais urbanas**, **estruturas artificiais rurais**, **estradas**, **natural**, **água** e **outras estruturas**.

Os dados do COS e das estradas, definidas pelo projeto OSM, permitiram definir os dados de referência que serviram de treino para estes modelos de aprendizagem.

Neste trabalho descreve-se a importância da deteção de estruturas artificiais para a prevenção de incêndios, dado que a correta definição das FGCI está dependente da correta identificação de habitações, cidades e estradas em torno das quais se devem implementar faixas de proteção. O facto dos PMDFCI estarem muito frequentemente desatualizados e por vezes não incluírem o detalhe necessário para proteger algumas habitações isoladas serviu como contexto para demonstrar a utilidade do uso deste processo de identificação para a geração automática de faixas de gestão de combustível.

Foi efetuada uma experiência com o objetivo de gerar FGC através desta classificação. Apesar da metodologia ser simples, recorrendo à aplicação direta de buffers sobre as áreas artificiais identificadas, verificou-se uma grande semelhança entre estas faixas geradas e as faixas oficiais de Tomar e Ferreira do Zêzere, como evidenciado pela figura 4.2.

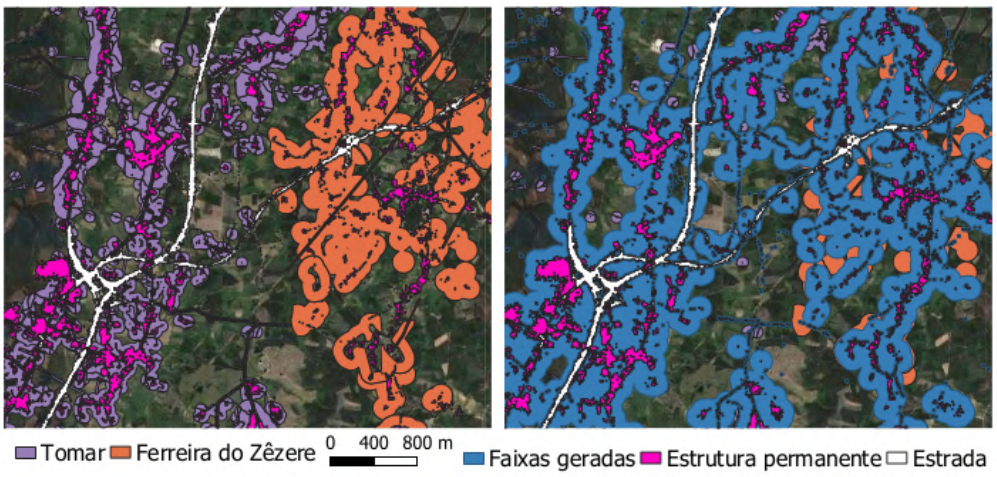


Figura 4.2: FGC oficiais à esquerda; FGC derivadas da classificação à direita - *Retirado de Neves et. al 2019*

4.3 Avaliação do estado de vegetação

Sendo um dos nossos objetivos medir o atual estado vegetativo em determinados pontos no país, é necessário entender de que forma podemos avaliar a vegetação, e de que forma esses indicadores se correlacionam com as imagens obtidas por satélite. As características gerais que parecem ser mais relevantes são a humidade da vegetação e a densidade e quantidade de biomassa presente. Valores altos da primeira reduzem substancialmente a probabilidade de combustão e da segunda alimentam os incêndios e possibilitam a sua propagação [4].

A medição destes parâmetros através de deteção remota resume-se à estimação de variáveis biofísicas. Estas associações são usadas em análises agrícolas, florestais ou ecológicas, e as mais frequentemente citadas no trabalho prévio estudado são as seguintes:

- Leaf Area Index (LAI) - A quantidade de área foliar por unidade de superfície.
- Canopy Chlorophyl Content (CCC) - A quantidade de clorofila (a e b) presentes por unidade de superfície na canópia.
- Canopy Water Content (CWC) - A quantidade de água presente na canópia por unidade de superfície.
- Aboveground Biomass (AGB) - Definido como o peso seco da vegetação, viva ou morta, presente acima da superfície.

Variáveis biofísicas como a *Leaf Area Index (LAI)*, que caracteriza a área foliar numa superfície, e a *Aboveground Biomass (AGB)* já foram usados, com sucesso, para avaliar o risco de incêndio [23][2], e como tal, mostram-se relevantes para o contexto deste projeto, especialmente no âmbito da monitorização da vegetação para o combate a fogos.

Existe trabalho prévio que correlaciona de forma significativa as variáveis biofísicas a determinados índices de vegetação, sendo comum a correlação com índices como o NDVI[9, 7, 24, 20], EVI [24], NDI45[20, 9] e IRECI [20, 9].

Como a criação de um novo modelo para estimar determinadas variáveis biofísicas está fora do contexto deste projeto, recorreu-se aos índices de vegetação que se correlacionam com estas variáveis.

4.4 Conclusões

Foram analisado vários estudos que se consideraram relacionados com o objetivo desta dissertação. Nestes incluem-se estudos com o objetivo de monitorizar o estado de FGCI, de detetar remotamente estruturas artificiais e avaliar o estado atual de vegetação.

O trabalho que tinha como objetivo determinar o estado de intervenção das FGCI mostrou um bom desempenho e um conjunto de abordagens que são comuns. Nestas incluem-se o uso de *buffers* externos à faixa para normalizar os valores e mais facilmente identificar alterações artificiais nas faixas (que neste caso correspondem a intervenções). Verificou-se o uso de índices espectrais para estimar o nível de vegetação das faixas, particularmente o uso do NDVI, e os resultados destes estudos demonstram que é um índice adequado para a tarefa.

A utilidade do NDVI e dos restantes índices espectrais para este caso de uso é apoiada por um conjunto de estudo que pretendem avaliar a correlação entre os mesmos e as variáveis biofísicas que são de interesse para a gestão das FGCI, como a biomassa.

Para podermos gerar faixas automaticamente e identificar aquelas que estão em falta nos planos de defesa é necessário identificar as zonas em redor das quais pretendemos definir as faixas. O trabalho prévio mencionado não só classificou um conjunto de classes de ocupação do solo que é do nosso interesse para identificar casas e localidade, como demonstrou que os resultados obtidos tinham uma qualidade suficiente para estimar as faixas, demonstrando que, mesmo com uma abordagem simples, existe uma grande sobreposição entre faixas geradas e faixas oficiais

5

Estado da Arte em Aprendizagem Automática e Séries Temporais

5.1 Classificação Automática

O objetivo desta tese é classificar o estado das faixas de gestão, e entender como as diferentes características, da vegetação e do seu comportamento ao longo do tempo influenciam o seu estado. Tendo em conta o trabalho prévio realizado na área de deteção remota, e particularmente na análise de séries temporais de vegetação, as técnicas de aprendizagem automática são as mais adequadas para a classificação do estado de intervenção das faixas, especialmente quando aliadas a uma análise do contexto temporal.

Definindo, de uma forma geral, a aprendizagem automática, esta refere-se à criação de um sistema que é capaz de aprender com experiência passada e, autonomamente, melhorar a capacidade de cumprir a tarefa que lhe foi atribuída.

No contexto da deteção remota os algoritmos mais utilizados dividem-se em duas categorias, supervisionados e não-supervisionados.

Aprendizagem supervisionada consiste em aprender a prever um dado atributo a partir de características (*features*) que está previamente classificada. É criado um modelo que classifica esta informação e compara os seus resultados com os de referência de forma a corrigir o modelo e melhorar a sua precisão na próxima iteração.

Na aprendizagem não-supervisionada os dados tratados pelo sistema não estão classificados, sendo que o objetivo do programa é encontrar as características comuns entre os dados e separar os mesmos em classes distintas. Na área da deteção remota um exemplo comum do uso de aprendizagem não-supervisionada prende-se com a separação de diferentes tipos de ocupação de solo, isto porque, por exemplo, existe um maior grau de semelhança entre diferentes área florestais do que entre uma zona florestal e uma área urbana.

Tendo em conta o trabalho prévio no uso de aprendizagem automática para a classificação de imagens remotas, destacaram-se as seguintes técnicas devido à sua utilização em contextos semelhantes [3, 17]

5.1.1 SVM

Uma Máquina de Vetores de Suporte (SVM) quando classifica um conjunto de dados define um hiperplano que separa os elementos de duas classes distintas. Este plano é traçado de forma a que a distância entre o mesmo e os pontos mais próximos de cada classe seja maximizado. Estes pontos mais próximos denominam-se de vetores de suporte.

No exemplo da figura 5.1 o hiperplano definido é a linha central e a margem, a tracejado, contém os dois vetores de suporte, de cada classe.

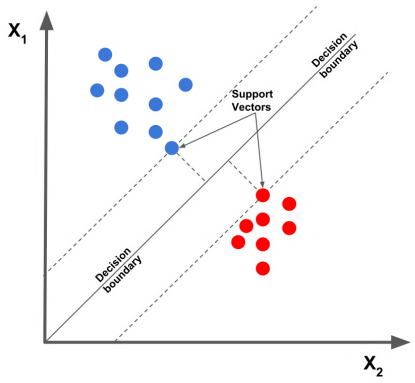


Figura 5.1: Exemplo do funcionamento do algoritmo SVM *Retirado de https://learnopencv.com/support-vector-machines-svm/*

Como se pretende separar duas classes com uma só linha, é impossível definir essa linha sobre um conjunto não-linearmente separável. Para contornar este problema, aplica-se a técnica *Kernel Trick*, que mapeia os dados para uma dimensão superior na qual se pode definir um plano que separe ambas as classes, como exemplificado na figura 5.2, onde um conjunto de dados num espaço bi-dimensional é projetado para 3 dimensões, tornando as classes separáveis.

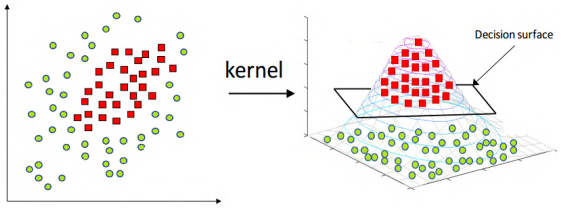


Figura 5.2: Utilização do Kernel-Trick *Retirado de https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d*

Uma grande vantagem deste método é que pode ser treinado com pequenas quantidades de dados e ainda assim obter resultados comparáveis ou superiores aos demais métodos na área de deteção remota [29]. O principal problema é a da classificação de mais de duas classes, e de classes cuja relação com os atributos seja complexa, e não-linear. Para o problema da classificação das FGCI o impacto desta desvantagem parece ser mínimo,

pois a classificação é binária, e a correlação entre as previsões e os atributos é de uma complexidade pouco alta, como demonstrado pelos resultados de [3].

5.1.2 Florestas Aleatórias (RF)

As Florestas Aleatórias são um método *ensemble* de aprendizagem supervisionada, isto quer dizer que usa vários modelos de aprendizagem automática para obter resultados mais precisos do que seria obtido por qualquer um dos modelos sozinhos. Neste caso, a RF utiliza várias árvores de decisão, cada uma delas operando independentemente de todas as outras. Cada uma das árvores gera uma classificação, e a classe mais popular é selecionada como classificação final.

A partir dos dados de treino são gerados subconjuntos, inferiores em dimensão ao grupo total de dados, e amostrados aleatoriamente para cada árvore. Este processo denomina-se de *bagging* e ajuda a melhorar a estabilidade e precisão do modelo.

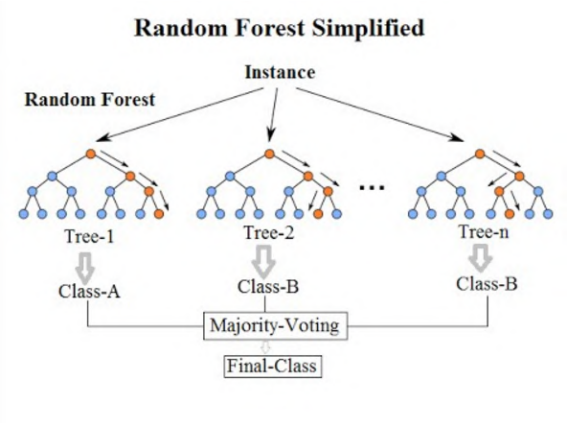


Figura 5.3: Exemplo do funcionamento de uma RF com 3 árvores *Retirado de https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png*

O seu uso é muito frequente na área da deteção remota, e particularmente, na análise de vegetação. Alguns exemplos que melhor se relacionam com os objetivos desta dissertação incluem a estimar variáveis biofísicas, como o *LAI*[17] e a *AGB*[43], a identificar diferentes classes de ocupação de solo[44] e em métodos de *Change Detection* em séries temporais [47].

O exemplo mais importante do uso de *Random Forests* é no trabalho desenvolvido por Ricardo Afonso[3], cujo objetivo é sensivelmente o mesmo desta dissertação. Como previamente mencionado este método foi o que retornou um dos melhores resultados de classificação para a tarefa de identificação de intervenções nas FGCI.

5.1.3 Gradient Boosting

Gradient Boosting é uma técnica de *ensemble*, tal como as Random Forest. No entanto, nesta técnica, as árvores de decisão não são executadas paralelamente, mas em sequência.

A primeira árvore recebe os dados originais e tenta prever a classificação correta, e esta classificação, como a de todas as árvores tem um erro inerente. As seguintes árvores, tentam encontar a função de ajuste dos dados que elimina o erro da árvore anterior. Este processo é repetido recursivamente, sendo que cada árvore recebe um conjunto de dados progressivamente mais ajustado.

O trabalho prévio indica que esta técnica, na generalidade, obtém melhores resultados do que a Random Forest, mas que, no entanto, requer um maior cuidado na configuração do modelo [22, 37, 3]. É importante notar que em [3], um trabalho que pretendia avaliar o estado das FGCI, a vantagem do uso de Gradient Boosting é debatível, sendo que a diferença de valores para o Random Forest era tangencial na maioria dos casos.

Uma implementação popular do conceito de Gradient Boosting é o XGBoost que tem sido usado com sucesso em trabalhos relacionados, incluindo para a deteção de intervenções em FGCI. [11]

5.1.4 Grid Search

Os modelos de aprendizagem habitualmente são acompanhados de um conjunto de *hiper-parâmetros*, definições externas ao modelo, e que portanto não são ajustadas no processo de aprendizagem. Os valores destes parâmetros têm de ser definidos antes do processo de aprendizagem começar. A otimização dos hiper-parâmetros pode ser feita manualmente, seja através de experimentação, ou dedução, tendo em conta o efeito dos parâmetros e o contexto do problema. No entanto, o método Grid Search permite-nos explorar as diferentes combinações de parâmetros, dentro de um determinado intervalo, de forma a definir valores ótimos para o modelo.

O método Grid Search está implementado na biblioteca *scikit learn*, permitindo uma fácil configuração sobre a forma como os parâmetros devem ser otimizados.

5.2 Análise de Séries Temporais

5.2.1 Métricas de dissimilaridade

Dadas duas séries temporais, uma métrica importante a ser calculada é o valor de semelhança, ou diferença, entre as duas. A comparação entre séries é importante tanto para uma fase de análise, de forma a entender as semelhanças entre as séries de uma dada classe, e a forma mais apropriada de as comparar de forma a identificar os comportamentos que se pretendem analisar. Além da análise é importante para definir como agrupar as séries, em processos de clustering, por exemplo.

5.2.1.1 Distância Euclidiana

O método mais simples para comparar duas séries é a **Distância Euclidiana**, consiste em, com duas séries sincronizadas, calcular a distância entre valores, em cada instante

de tempo, e somar as distâncias. É uma métrica rápida e simples, no entanto é sensível à presença de ruído, valores anómalos e a um desfasamento entre as duas séries.

5.2.1.2 Dynamic Time Warping (DTW)

Quando existe algum tipo de distorção temporal entre as séries uma das abordagens mais comuns é o uso de **Dynamic Time Warping** (DTW). Neste processo é criado, para duas séries temporais, um *warping path*, uma sequência de distorções que minimiza a diferença entre as séries dentro de uma dada restrição global. Restrição esta limita a distorção que máxima para cada instante de tempo.

Existem variações sobre o DTW, como o *Derivative DTW* que além de distorcer os valores no eixo do tempo, distorce no eixo do y, de forma a que a distância seja mais focada na forma da série. O *Weighted DTW* que aplica um peso sobre o cálculo da distância tendo em conta a distorção imposta naquele instante temporal.

Relativamente ao contexto de deteção remota e da deteção de instantes de intervenção sobre FGCI, o uso do DTW aparenta ser promissor, dado que os instantes de intervenção e a velocidade com qual eles ocorrem diferem de faixa para faixa, e este tipo de distorção ajuda a colmatar essas diferenças. É importante notar que o nível da vegetação é relevante para avaliar a intervenção, por exemplo, tendo em conta o NDVI, uma descida de um valor de 0.8 para 0.4 (descida de 50%) é muito relevante, enquanto uma descida igualmente íngreme de 0.2 para 0.1 é menos relevante, dado que provavelmente corresponde a um terreno relativamente árido. E no entanto, ambas estas séries teriam a mesma forma e seriam consideradas semelhantes pelo *Derivative DTW*.

5.2.2 Aprendizagem

Uma das formas mais simples e comummente usadas para classificar uma série é classificar os atributos que dela se calculam. Usando as métricas de distância previamente mencionadas, e calculando a distância aos conjuntos de treino de dados podemos criar modelos de aprendizagem com classificadores como **SVM**, **Random Forests** e **XGBoost** que são os algoritmos contemplados para esta fase de aprendizagem.

5.2.3 Suavização

Quando estamos a lidar com séries temporais que são geradas a partir de informação do mundo real, é normal que estas tenham imperfeições, *outliers*, interferências e pequenas divergências da tendência esperada. Um processo usado para colmatar estes erros e gerar uma série temporal de melhor qualidade são processo de suavização ou *smoothing*. Um dos processos mais simples e comummente usados é o de uma média móvel, que consiste em, a partir de uma janela temporal ao redor de um instante, calcular uma média de valores e gerar uma nova série com uma sequência dessas médias. Este método já foi

usado em trabalhos prévios que abordavam precisamente a avaliação do estado de FGCI [34].

Uma outra abordagem que se tem revelado popular é o filtro **Savitzky-Golay**, que foi demonstrado como sendo um método muito eficaz a reconstruir séries temporais de NDVI, um dos índices de vegetação mais relevantes para o trabalho em curso [6].

5.3 Agrupamento

Nesta secção listamos os algoritmos seleccionados para testar o uso de aprendizagem não-supervisionada com o objetivo de separar as séries que têm instantes de intervenção e as séries que não sofreram cortes. Os processos de clustering dependerão das métrica de distância usada, sendo que a distância entre cada série temporal, ou mais especificamente, entre cada série e a série composta de um cluster é que determina o cluster a que ela pertence.

O objetivo geral desta dissertação é a deteção de intervenções nas FGCI, logo, o processo de clustering não nos fornece este tipo de identificação, sendo que o seu funcionamento não se baseia no uso de dados de referência. No entanto, as classes que são geradas a partir dos clusters podem ter algumas semelhanças às classes que se pretendem identificar. Podendo, caso elas sejam significativamente homogéneas, identificar os vários clusters gerados como 'cluster de intervenção', 'cluster de não-intervenção' e 'misto'.

5.3.1 K-Means

O algoritmo k-means é um dos mais populares métodos de clustering, pretendendo particionar um conjunto de dados em k clusters, valor este definido previamente ao processo de clustering.

O seu funcionamento é relativamente simples, inicialmente são atribuídos k centros, um por cada cluster a ser gerado. Cada ponto passará a pertencer ao cluster cujo centro lhe é mais próximo, sendo que de seguida os centros são recalculados usando os centróides de cada cluster. Este processo repete-se até convergir, e o centro de cada cluster não se alterar.

No contexto de séries temporais, o funcionamento é semelhante, cada centro corresponderá a uma série temporal, sendo que as métricas de similaridade serão usadas para calcular a distância entre cada série, seja a distância euclidiana, ou a distância com DTW, por exemplo. O cálculo do centróide é feito pela média das séries temporais de cada cluster.

5.3.2 KShape

O algoritmo KShape é, comparativamente ao K-Means, um método mais recente que se foca na forma das séries temporais. Ainda assim, o seu funcionamento base é bastante semelhante ao K-Means.

Neste caso a métrica de distância é baseada na relação cruzada (ou *cross-correlation*) entre duas séries temporais. Tal como outros métodos, como o DTW, esta métrica faz um ajuste sobre as séries de forma a sincronizar sinais semelhantes. A parte que diferencia este algoritmo é que usa a informação do ajuste que foi feito para sincronizar as séries para recalcular os centróides de cada cluster.

É criada uma nova medida de distância baseado na forma das séries, e tendo como base o princípio de correlação-cruzada.

No processo de cálculo do centróide, as séries temporais são ajustadas relativamente ao último centróide da sua classe (de acordo com o processo de auto-correlação), e daí segue-se um problema de maximização, isto é, gerar uma série que maximize a soma das semelhanças de todas as séries temporais ajustadas do mesmo cluster. A série resultante define o novo centro de cada cluster.

5.4 Avaliação

O processo de avaliação dependerá se estamos a avaliar uma classificação binária, ou um processo de *clustering*.

No caso de uma classificação binária temos 4 tipos de resultados de uma predição do modelo. Estes resultados podem ser contabilizados numa **Matrix de Confusão**. Cada entrada desta matriz corresponde à quantidade de Falsos Positivos (FP), Verdadeiros Positivos (VP), Verdadeiros Negativos (VN) ou Falsos Negativos (FN), como representado na figura 5.1 para N elementos.

Tabela 5.1: Matriz de Confusão de um classificador binário

		Resultado Real		
		Positivo	Negativo	Total
Resultado estimado	Positivo	VP	FP	$VP+FP$
	Negativo	FN	VN	$FN+VN$
	Total	$VP+FN$	$FP+VN$	N

5.4.1 Precisão

A precisão é um valor que que se refere a uma classe só (x), medindo a proporção de elementos que foram corretamente classificados como sendo dessa classe entre aqueles que foram previstos como tal. Fórmula descrita em 5.6

5.4.2 Recall

O Recall, também referente a uma só classe (x), representa a proporção de elementos que foram corretamente classificados para aquela classe face ao total de elementos de referência dessa classe. Fórmula descrita em 5.7

5.4.3 F1-Score

O F1-Score é uma métrica que incorpora a precisão e o recall, correspondendo ao cálculo de uma média harmónica entre os dois valores. O maior valor possível é 1, quando a precisão e o recall são perfeitos e o menos valor é 0, quando ou a precisão ou o recall são zero. Fórmula descrita em 5.3.

$$\text{Precisão} = \frac{V_x}{V_x + F_x} \quad (5.1) \qquad \text{Recall} = \frac{V_x}{V_x + \sum_{y \neq x} F_y} \quad (5.2) \qquad \text{F1} = \frac{2 * \text{precisão} * recall}{\text{precisão} + recall} \quad (5.3)$$

5.4.4 Accuracy

A Accuracy é uma métrica popular que simplesmente mede a proporção entre os elementos que foram corretamente classificado e o total.

$$\text{Accuracy} = \frac{\sum_{x \in X} (V_x)}{\sum_{x \in X} (V_x + F_x)}$$

Quando existe um conjunto de dados cujas classes não estão balanceadas o valor de *accuracy* pode não estar a refletir a verdadeira precisão do modelo, sendo que o resultado da classe dominante tem um maior impacto no valor da métrica. Nestas situações pode ser útil usar uma versão pesada da métrica, definida na fórmula 5.5 para o caso de uma classificação binária.

$$\text{Especificidade} = \frac{VP}{VN + FP} \quad (5.4) \qquad \text{Accuracy Pesada} = \frac{\text{Recall} + \text{Especificidade}}{2} \quad (5.5)$$

5.4.4.1 Kappa

O Kappa compara os resultados obtidos à probabilidade de os mesmos serem atingidos aleatoriamente (designado por p_e).

$$p_e = \frac{1}{N^2} \sum_{k \in X} n_{k1} * n_{k2} \quad (5.6) \qquad \text{Kappa} = \frac{accuracy - p_e}{1 - p_e} \quad (5.7)$$

Sendo que N indica o número de elementos classificados, k corresponde a cada uma das classe existentes, n_{k1} indica o número de elementos que são da classe k e n_{k2} o número de elementos que foram classificados como fazendo parte de k.

5.4.5 Avaliação de Clustering

Relativamente a um processo de clustering existem diferentes formas de avaliar o processo de clustering, dependendo se se pretende comparar com dados de referência ou as características intrínsecas dos clusters.

No caso da existência de um conjunto de dados de referência o processo de clustering pode ser avaliado relativamente à forma como separa uma determinada classe. Um métodos muito comum é o Rand index. Este índice mede a semelhança entre dois clusterings do mesmo conjunto de dados, logo pode ser usado para comparar o resultado de clustering com os dados de referência, sendo que neste contexto cada valor da classe em estudo é comparado a um cluster. A desvantagem deste tipo de métrica é que a sua utilidade é limitada ao caso onde se pretende equiparar os clusters a uma classe, penalizando casos onde uma classe está representada em vários clusters.

No contexto desta dissertação pretende-se avaliar o estado binário das intervenções sobre FGCI, logo existem apenas duas classes para prever. No entanto entende-se que fixar o número de clusters a dois será contraprodutivo, pois cada geometria em análise terá comportamentos fenológicos diferentes, cada intervenção acontecerá em alturas diferentes, e será efetuada usando métodos diferentes, logo, dentro de cada classe (intervencionado ou não-intervencionado) existe uma grande variedade de séries temporais. Um maior número de clusters permite que cada cluster represente, idealmente, o estado de intervenção, dado o contexto da vegetação e do corte efetuado.

A **homogeneidade** calcula, como o nome indica, a homogeneidade de classes dentro de cada cluster. Caso cada cluster contenha uma única classe esse valor será de 1, e no caso do processo de clustering ser aleatório, esse valor tenderá para 0. O seu cálculo é feito com recurso à formula da entropia, comparando o resultado obtido com aquele esperado numa distribuição aleatória.

5.5 Conclusões

Neste capítulo exploraram-se diferentes metodologias de aprendizagem automática, tanto supervisionada, como não-supervisionadas. O trabalho prévio realizado na área da monitoração de FGCI e de vegetação no geral mostram que particularmente, os algoritmos de Florestas Aleatórias são robustos para este tipo de tarefas.

A abordagem tomada para a identificação das intervenções irá ser tomada ao nível das séries temporais, e ao nível de cada instante temporal. Os métodos de agrupamento operam ao nível das séries, utilizando um conjunto de métricas para comparar as mesmas e poder assim separar em conjuntos relativamente homogéneos. Pelo outro lado, os métodos de classificação dependerão das características de cada instante, utilizando essa informação para treinar modelos de aprendizagem.

Tendo em conta os objetivos desta dissertação, determinou-se que o uso da homogeneidade é o mais adequado para avaliar os agrupamentos gerados, sendo que no seu valor máximo, existem grupos exclusivamente compostos por séries intervencionadas e outros de séries não-intervencionadas.

Relativamente aos classificadores binários que serão testados, deu-se destaque ao F1 Score e ao Kappa como as métricas mais relevantes para avaliar os modelos.

6

Abordagem

6.1 Introdução

Neste capítulo apresenta-se a metodologia que foi usada no trabalho para completar as várias etapas de processamento de dados, e para efetuar as análises feitas sobre os mesmos, com o principal objetivo de identificar os instantes de intervenção sobre as FGCI.

É também descrito a abordagem para o objetivo secundário da geração automática de faixas de gestão de combustível e o processo manual de enriquecimento dos dados de referência através da marcação das datas de intervenção.

Os dados de deteção remota precisam de ser pré-processados para que as séries temporais extraídas sejam o mais precisas possíveis e para extrair novas características a partir dos valores das bandas de satélite e do contexto temporal.

A informação vetorial que define as FGCI, seja nos PMDFCI, seja nos dados de referência, são essenciais pois definem a área de estudo deste trabalho. É necessário algum processamento das mesmas, nomeadamente para segmentar as faixas em áreas de estudo mais pequenas.

Finalmente, o conjunto de dados de referência que servirão de base para o treino e avaliação dos modelos de aprendizagem, são separados em grupos, para que após o processo de avaliação, se retirem conclusões sobre a importância de determinadas características para a deteção de ações de intervenção.

6.2 Ingestão e processamento de dados geográficos

6.2.1 FGCI

6.2.1.1 Abordagem para cada classe de faixa

Em relação às faixas oficiais definidas nos PMDFCI, estas são discretizadas através de um identificador, *ID_R_FGC*, único a cada faixa. Existem exceções, em quatro concelhos este identificador, ou o seu equivalente, tem outro nome, e num único caso, não existe nenhum identificador de faixa.

De uma forma geral, o processamento essencial cinge-se à separação das faixas através do seu identificador único, unindo os vários fragmentos geográficos que compõem cada faixa, corrigindo qualquer artefacto presente, e posteriormente, dividindo as faixas entre 4 ficheiros, referentes às **faixa primárias**, às **faixas de habitações e aglomerados populacionais**, às **faixas de redes viárias** e às **faixas da rede elétrica**.

O foco desta dissertação é na análise das faixas de habitações e aglomerados, no entanto o processo geral pode aplicar-se a qualquer tipo de faixa. A fase inicial deste processamento corresponde apenas à correção automática das geometrias, pois alguns dos ficheiro geográficos incluem erros geométricos que impossibilitam algumas operações, e da união das geometrias que possuem o mesmo identificador. Posteriormente é efetuada uma análise de sobreposição com o COS da região, sendo adicionado aos atributos a classe COS predominante em cada faixa.

6.2.1.2 Segmentação de faixas

Após termos estas faixas "limpas", segmentadas por classe, podemos passar à próxima fase, que corresponde à segmentação de cada faixa de habitações ou aglomerados populacionais em áreas de estudo mais pequenas, posteriormente chamadas de *fragmentos*.

Os critérios relevantes para a criação dos fragmentos são a classe do COS de determinada área, a distância ao centro da faixa (no caso dos aglomerados e habitações), e a sua área.

Relativamente à classe do COS, é particularmente pertinente pois cada tipo de vegetação tem um comportamento diferente, e como tal, é preferível ter áreas de estudo, e consequentemente, séries temporais que reflitam um único tipo de vegetação. Em segundo lugar, a distância ao centro das faixas. Esta informação mostrou-se relevante após uma conversa com o Vice-Presidente da Câmara de Mação, Engenheiro António Louro, que informou que, de uma forma geral, as características da vegetação no interior das cidades era tendencialmente agrícola, enquanto as da periferia eram tendencialmente florestais. Como tal, decidiu-se usar esta característica para segmentar as áreas de análise, esperando-se aumentar a homogeneidade de cada fragmento. Finalmente, passando à área, é importante que os fragmentos tenham um tamanho grande o suficiente para que possa ser feita uma leitura de qualidade, no entanto, pequenas o suficiente para que a área, e o seu comportamento, seja minimamente homogéneo. Um valor médio de 18 píxeis considera-se adequado, sendo que abaixo de 6 píxeis se considera que a área é pequena demais para se fazer uma amostragem adequada.

Para melhor entender a relevância do COS na segmentação das faixas podemos visualizar a diversidade de classes do COS nas faixas, como demonstrado no seguinte histograma 6.1, referente às faixas de aglomerados populacionais e habitações. Esta figura permite confirmar que as faixas não são homogéneas e que é importante separar os diferentes tipos de vegetação antes de fazer algum tipo de análise.

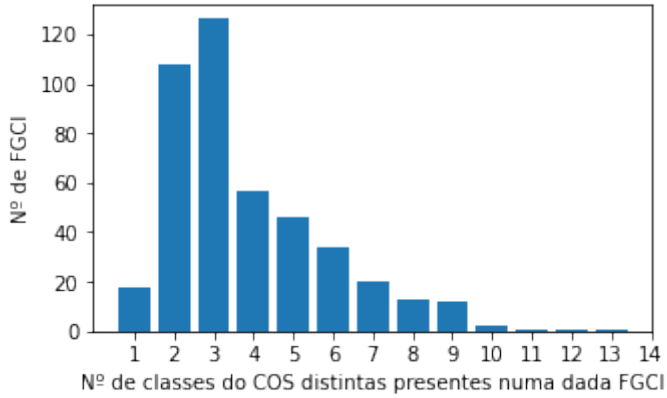


Figura 6.1: Histograma do nº de classes distintas do COS em cada faixa no concelho de Mação

6.2.1.3 Análise de qualidade

Após os processos de integração e transformação das faixas é importante que seja feita uma análise para melhor entender a qualidade das mesmas. Enquanto a maioria dos problemas encontrados nas faixas não são críticos, como algumas pequenas geometrias inválidas, ou pequenas sobreposições entre faixas, e com zonas artificiais, houve um detalhe que sobressaiu, relativamente às habitações isoladas. Estas habitações estão mapeadas nas FGCI disponibilizadas, no entanto existem várias habitações cujas faixas não estão definidas, apesar de se encontrarem em zonas florestais. Este tipo de faixas é crucial para preservar habitações e vidas humanas, e como tal, considerou-se muito importante que estejam propriamente definidas nos planos de proteção.

Tendo em conta esta falta de faixas em algumas zonas decidiu-se explorar a possibilidade de gerar automaticamente as faixas de gestão de combustível teóricas tendo em conta as zonas habitacionais, para posteriormente se poder comparar estas faixas teóricas com as oficiais e identificar as FGCI em falta.

6.2.2 Dados de Referência

6.2.2.1 Contexto

O objetivo da obtenção destes dados de referência é usar a informação relativa ao estado de manutenção para fazer uma análise sobre os comportamentos que caracterizam as ações de corte, tal como usar estes dados para treinar e avaliar os modelos que nos permitem identificar estas ações de intervenção.

Como previamente mencionado na tabela 3.2 existem 3 tipos de estados manutenção para as faixas de Mação, **executado** (indicando a data limite de execução), **por executar** e **em execução**, sendo que pretendemos distinguir séries intervencionadas e não-intervencionadas, as faixas que se encontram **em execução** não nos indicam de forma clara em que estado da manutenção se encontra a faixa, podendo estar nas fases finais, ou

iniciais. Como tal, decidiu-se não utilizar as faixas que estivessem em execução, usando apenas aquelas onde existe certeza sobre o estado de manutenção.

Para os dados de referência de Santarém a informação fornecida é diferente, não indicando a data limite de execução, e usando as geometrias em si como indicação de uma intervenção num dado ano (com raras exceções). Há quatro ficheiros geográficos, para cada ano entre 2018 e 2021, sendo possível identificar as zonas que **não** foram intervencionadas pela ausência de polígonos numa dada área. Acontece que, entre os vários anos, as mesmas faixas e zonas de intervenção têm geometrias ligeiramente diferentes o que dificulta uma análise ao longo dos vários anos.

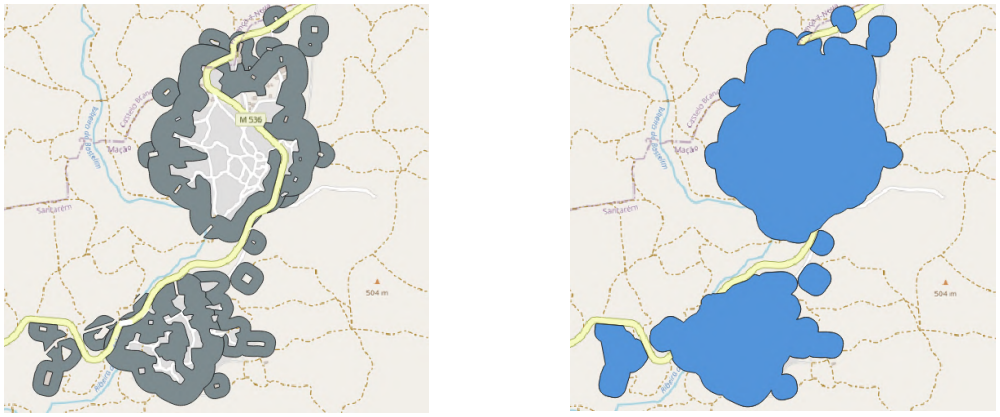
6.2.2.2 Enriquecimento da informação em Mação

De forma a enriquecer a informação disponibilizada, depois de unir as geometrias de acordo com os atributos identificadores, foi realizada uma interseção com os dois parâmetros que pretendemos usar para criar os fragmentos para análise, nomeadamente a distância ao centro das faixas (figura 6.3) e a classe maioritária do COS. Após esta interseção ficamos com umas geometrias mais segmentadas, sendo necessário depois, garantir que o tamanho das mesma é suficiente para ser feita uma captura fiável, mas não grande o suficiente que não nos permita garantir alguma homogeneidade dentro da geometria, sendo que ao haver diferentes tipos de vegetação, existem diferentes comportamentos, que causa uma interferência de sinal, e consequentemente uma série temporal da qual seja mais difícil tirar conclusões.

Como mencionado, uma das características para a criação dos fragmentos é a distância de um dado ponto da faixa à periferia da mesma. Para poder calcular esta distância foi criado um mapa de distâncias para todas as faixas relativas a habitações e aglomerados populacionais. A partir da definição das faixas originais (figura 6.2a) geram-se as geometrias que representam o interior das faixas (figura 6.2b), com recurso ao algoritmo Concave-Hull, e a partir destas geometrias são aplicados sucessivos *buffers* negativos para gerar as várias camadas que se podem ver na figura 6.3

Após a interseção destes mapas de distâncias com as diferentes classes do COS obtemos novos fragmentos que, idealmente, são mais homogéneos que a faixa como um todo. No entanto, estas novas zonas continuam a poder ter áreas muito consideráveis, e pretende-se que os fragmentos em análise sejam relativamente pequeno, pois mesmo que a COS divida as diferentes classes de ocupação de forma perfeita continuam a haver variações no terreno, na densidade da floresta, e na qualidade das imagens capturadas na área. Caso exista a presença de nuvens numa secção de um fragmento grande toda a área terá de ser desconsiderada.

Como tal, decidiu-se além da distância à periferia dos aglomerados e da classe do COS aplicar uma fragmentação adicional para restringir o tamanho de cada fragmento. De forma geral, foi gerada uma grelha de pontos sobre as faixas, sendo aplicado um clustering K-Means com um número de classes tal que cada classe tivesse em média 20



(a) FGCI junto a uma localidade (b) Área interior às FGCI

Figura 6.2: Fases intermédias para a criação de um mapa de distâncias à periferia das faixas. Este exemplo está localizado no concelho de Mação.

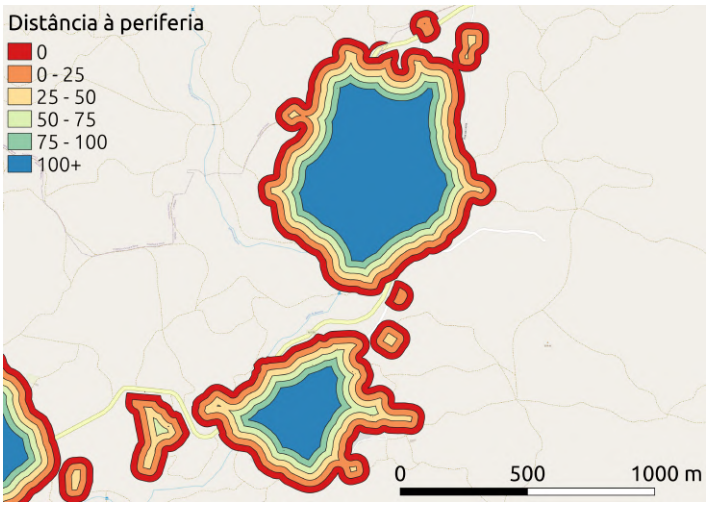


Figura 6.3: Distância calculada à periferia da faixa de uma localidade do concelho de Mação.

pontos.

Esta grelha de pontos usada para gerar os clusters tem um espaçamento de 10 metros entre cada ponto, a mesma resolução que aquela fornecida pelas bandas utilizadas do Sentinel-2. Considerou-se que seria benéfico alinhar as fronteiras de cada cluster às fronteiras dos píxeis do Sentinel-2, isto para diminuir qualquer influência de zonas externas a um dado fragmento nas leituras feitas do mesmo. Para tal, foi gerada a grelha de pontos que definem o centro de cada píxel do Sentinel-2. A estes pontos foi aplicado um buffer para que os polígonos resultantes representassem as fronteiras dos píxeis do Sentinel-2, sendo depois estas fronteiras usadas para delimitar os diferentes fragmentos gerados. Este processo foi feito recorrendo ao Google Earth Engine, exportando uma imagem da zona em análise, em formato GeoTIFF, e gerando um ponto para o centro de cada píxel.

Na figura 6.4 podemos ver o exemplo de uma faixa, cujas cores identificam os fragmentos para ela gerados antes de ser aplicado o método de clustering. Os pontos centrais dos píxeis do Sentinel-2 estão sobrepostos à imagem, a cinzento, e é possível ver como eles definem a fronteira dos fragmentos após o processo de clustering.

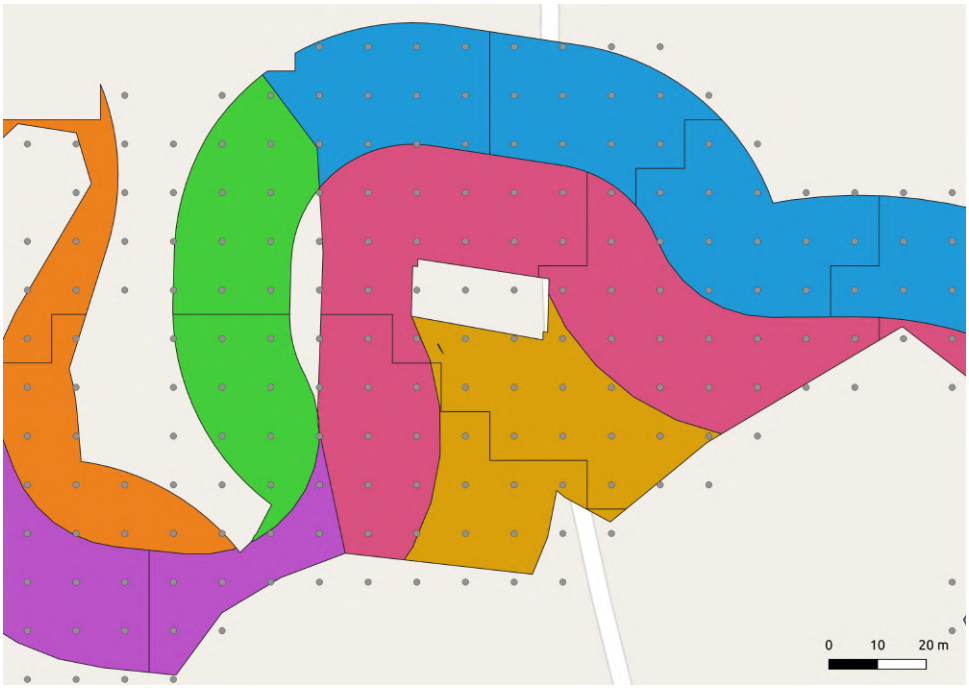


Figura 6.4: Fragmentos gerados para uma faixa, sobrepostos com os pontos centrais dos píxeis do Sentinel-2

É também realizada uma interseção das geometrias de referência com as faixas do PMDFCI, de forma a ter mais informação e contexto caso seja necessário no processo de análise.

6.2.2.3 Enriquecimento da informação em Santarém

Para o caso das faixas de Santarém, o processamento tem de ser feito de forma ligeiramente diferente às de Mação. Primeiramente foi feita uma análise para verificar se seria possível fazer o mesmo tipo de processamento em "anel"feito para as faixas de Mação. Para tal, foi analisada a sobreposição das zonas intervencionadas com as faixas assinaladas no PMDFCI (como descrito nas figuras 3.5). Para determinar a distância ao interior de um aglomerado populacional e fazer a divisão dos fragmentos dessa forma seria necessário uma muito boa sobreposição das zonas de intervenção com as faixas do plano de defesa, o que não se verificou. Como tal, e sabendo que a área média destas faixas é inferior às de Mação, decidiu-se fragmentar estas zonas apenas pelo COS e pelo método de clustering K-Means.

Como anteriormente mencionado, as zonas de intervenção estão ligeiramente desalinhadas entre os vários anos, além de que, uma zona pode não estar presente na geometria num dado ano, ou apenas parcialmente presente, indicando que essa zona não foi totalmente intervencionada. Como tal, será necessário antes de iniciar o processo de fragmentação das séries, criar geometrias que combinem a informação dos 4 anos a que temos acesso. Este método pode não ser preciso a gerar as zonas que efetivamente foram sujeitas a intervenção, dado que algumas das faixas estão muito desalinhadas (como no exemplo da figura 6.5), não representando a área que foi sujeita a intervenção. Este problema vem da definição das geometrias nos dados que nos foram fornecidos, não sendo possível corrigir estes defeitos de forma automática.

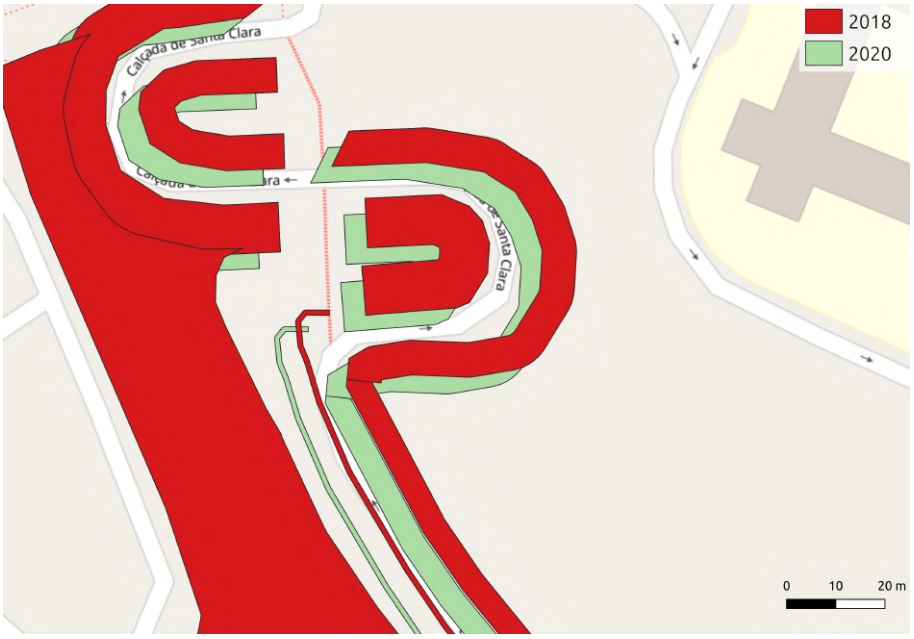


Figura 6.5: Exemplo do desalinhamento de faixas equivalentes entre vários anos

As 8 camadas, dos 4 anos e dos 2 tipos de faixas (aglomerados populacionais e estradas), foram unidas, assegurando que não havia nenhuma geometria que unisse dois tipos de faixas diferentes. Estas geometrias unidas definem, para o âmbito deste trabalho, as faixas de intervenção, e são a base para a criação dos fragmentos que serão posteriormente analisados para determinar o seu estado.

A estas novas faixas unidas foi atribuído um identificador, um por cada geometria, foi feita uma interseção com o COS da região de Santarém, gerando um novo identificador, e por fim, foi aplicado o método K-Means para gerar zonas de análise mais pequenas. Este método é idêntico àquele aplicado para as faixas de Mação, excluindo a separação pela distância ao interior de um aglomerado populacional.

Para cada um destes fragmentos, gerados pelos processos de interseção e *clustering*, foi criado um atributo que identifica se foi intervencionado num dado ano ou não. Esta identificação fez-se através da análise da sobreposição do fragmento com a camada dos dados de referência para um dado ano. Isto é, caso 75% do fragmento esteja incluído nas

faixas dos dados de referência de um dado ano, conclui-se que esse fragmento foi intervencionado neste ano. O valor limite de 75% foi definido para dar alguma flexibilidade a este processo, pois sabe-se que os dados de referência estão frequentemente desalinhados.

Concluindo, temos os fragmentos das zonas intervencionadas entre 2018 e 2021, e a informação dos anos em que estes fragmentos foram sujeitos a ações de manutenção (de acordo com os dados de referência).

6.3 Extração de dados de Remote Sensing

6.3.1 Google Earth Engine

A parte deste projeto que se encarrega do acesso a dados de deteção remota foi conduzida recorrendo à plataforma Google Earth Engine (GEE). Esta plataforma de computação na cloud encarrega-se do armazenamento de dados históricos de imagens de satélite, de missões como as Landsat, Sentinel, MODIS, entre muitas outras. O GEE disponibiliza também um conjunto de ferramentas para a análise de dados geográficos, permitindo adicionar e manipular múltiplas bandas disponibilizadas pelos satélites, de forma a visualizar rácios, índices ou outros tipos de valores deduzíveis. A plataforma permite-nos fazer *upload* de informação geográfica relevante para extrair características e posteriormente fazer uma análise.

Os dados que se pretendem exportar através desta plataforma referem-se às séries temporais de cada fragmento de faixa em análise, extraindo-se da plataforma os índices de vegetação que se consideram relevantes, tal como os metadados referentes à captura e à qualidade dos dados.

Além de séries temporais, permite-nos também exportar dados raster, tanto imagens de um único instante, visualizadas sobre as bandas desejadas, tal como uma sequência de capturas, em formato de vídeo para facilitar a análise de determinadas áreas ao longo do tempo.

A plataforma GEE disponibiliza uma interface online, com recurso à linguagem *Javascript* para a execução de scripts, e visualização dos resultados. É também disponibilizada uma interface para Python que requer alguma configuração local adicional, no entanto foi optado o uso da plataforma web, tanto pelas ferramentas de interação disponibilizadas no mapa da plataforma, na presença de documentação extensiva dentro do editor e da facilidade de uso da linguagem.

6.3.2 Objetivos

Os dados de deteção remota são a base que possibilita a análise das FGCI e a sua mudança de comportamento ao longo do tempo. No entanto, para poder tirar conclusões a partir dos dados remotos é necessário garantir que os valores extraídos são fiáveis, livres de interferência e temporalmente consistentes. Além disso é importante que os dados que iremos extrair explicitem as características em análise (presença de vegetação e momentos

de corte), tornando-as fáceis de identificar, tanto por processos automáticos como por processos manuais.

Relativamente às imagens de satélite, o principal problema que inviabiliza uma leitura de qualidade do solo é a presença de nuvens e de outras interferências atmosféricas. Será necessário identificar e marcar a presença das nuvens de forma a que se possam seleccionar apenas as capturas com leituras válidas.

As características que pretendemos analisar prendem-se com os níveis de vegetação. Como previamente mencionado, existe um conjunto de índices espectrais que se correlacionam com as variáveis biofísicas são de interesse para o projeto, e como tal, devem ser calculadas a partir das capturas de satélite.

Num processo de extração de séries temporais os processos anteriormente referidos serão aplicados à área e ao intervalo temporal de interesse, sendo possível extrair os resultados para cada geometria, sendo essa informação posteriormente transformada, analisada e modelada localmente.

Além do processo de extração de informação textual (no formato de tabelas CSV) é importante, para o fator humano, a visualização e exportação de informação visual, seja no formato de imagens ou vídeos, de forma a facilitar a análise num contexto temporal e espacial. Isto permitirá melhor entender e resolver problemas nas anteriores fases de processamento, adicionando a componente espacial que pode melhor contextualizar os valores extraídos.

6.3.3 Extração de séries temporais

Foi criado no Google Earth Engine uma coleção de imagens de Sentinel-2 que incluísse as várias características e informações que são relevantes para esta dissertação. Isto inclui a identificação dos píxeis obscurecidos por nuvens e os valores calculados dos 4 índices de vegetação eleitos para cada píxel. Para gerar esta coleção também se renomeou os nomes originais de cada banda do Sentinel-2 para nomes mais legíveis, sendo que os nomes originais seguiam a convenção B1, B2, B3, e por aí diante, tendo sido renomeados para "Red", "Green, "Blue", por exemplo.

Para o processo de extração de séries temporais foi criada uma função que exporta um conjunto de séries temporais, em formato CSV, de acordo com os parâmetros introduzidos. Estes incluem **o intervalo de tempo das capturas**, **um conjunto de geometrias**, **um identificador sobre o qual agrupar essas geometrias** e **os atributos que se pretendem exportar**.

Para que cada geometria seja reduzida a um único valor para cada instante e por cada atributo, terá de ser escolhida uma função redutora. As 3 opções ponderadas são a média, a mediana, e a mediana inter-quartil. Esta última refere-se à mediana dentro de os valores do percentil 25 e 75, após uma análise de dispersão.

Efetuou-se uma análise mais detalhada de forma a escolher um destes métodos, sendo

extraídos os valores de todos os píxeis dentro de cada geometria e sendo analisado, localmente, o efeito que cada método redutor tinha sobre a série temporal das geometrias. Verificou-se que a mediana inter-quartil (IQM) produzia as séries mais suaves, com o menor número de *outliers*, tendo sido eleita para a redução dos valores das geometrias na plataforma GEE.

Finalmente, estando incluídos os dados e atributos relevantes, tal como os métodos para reduzir cada geometria a um único valor, procede-se à criação das séries temporais, iterando-se por cada captura e geometria, sendo incorporadas as bandas e índices que se pretende extrair, tal como alguns metadados, como a presença de nuvens, a data da captura, o identificador da captura, e o número de píxeis que a geometria engloba.

Após ser extraída a informação de todas as geometrias para cada instante de tempo, os resultados são guardados numa tabela em formato CSV que é posteriormente exportada para a plataforma Google Drive.

6.3.4 Identificação de cobertura de nuvens

Uma parte crucial deste trabalho prende-se na obtenção de dados de remote sensing. Como previamente discutido, estes dados são provenientes da missão Sentinel-2, sendo que se pretende, destes produtos, extrair os valores que são considerados úteis para a análise em curso, e extrair o máximo de pontos temporais relevantes, assegurando que os mesmos são pontos válidos e não representam valores de interferências atmosféricas.

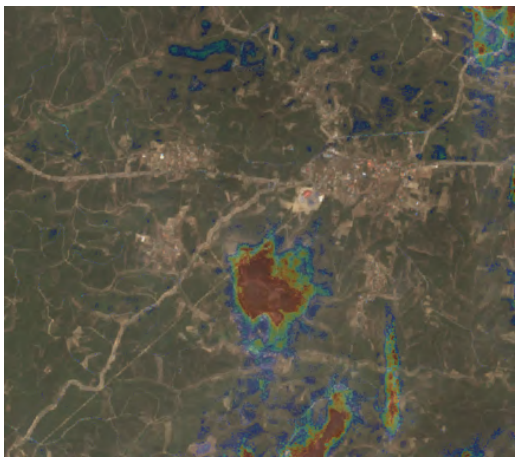
A maior causa de valores inválidos nas séries temporais dos produtos de remote sensing do Sentinel-2, e de outras observações no domínio do infra-vermelho e do visível, é a presença de nuvens. Pretende-se portanto criar uma máscara de nuvens que permite excluir estes píxeis das capturas que estão em análise. Existe um conjunto de informação que podemos usar para avaliar a presença de nuvens numa dada captura, como previamente mencionado. Dos métodos listados seleccionou-se o produto s2cloudless computado pela Sentinel Hub. Este produto tem uma resolução de 10 metros, sendo que cada píxel contém um atributo que define a probabilidade de o mesmo pertencer a uma nuvem. É importante notar que desta classificação resulta uma imagem com um grau relativamente elevado de ruído. Píxeis que individualmente foram classificados como nuvens, mas cujos vizinhos têm uma muito baixa probabilidade de o serem, provavelmente são píxeis mal classificados. É necessário que haja uma densidade suficiente de píxeis com uma probabilidade elevada para que se possa considerar que uma determinada área de uma captura tem uma alta probabilidade de ter a presença de nuvens.

A plataforma Google Earth Engine fornece num tutorial um exemplo de como criar esta máscara, tendo em conta este conceito de classificar um píxel como nuvem tendo em conta o seu valor e o da sua vizinhança [38]. No entanto, notou-se que estas máscaras continuavam a ter problemas em classificar nuvens cirrus, que são mais translúcidas, e que eram criados buffers excessivos que mascaravam zonas nas quais não havia qualquer

presença de nuvens. Tendo em conta estes problemas tornou-se clara a necessidade de reformular o método fornecido pela plataforma, de forma a garantir o máximo de cobertura das nuvens, com a mínima área de máscara possível.

As figuras 6.6 demonstram o produto s2cloudless em comparação à imagem do Sentinel-2. Neste caso de exemplo existe uma cobertura significativa por nuvens cirrus que, apesar de por vezes serem difíceis de ver, têm a capacidade de causar interferência na leitura dos valores do Sentinel-2.

Nas figuras 6.7 é possível ver de que forma as diferentes máscaras, a da Google e esta nova proposta lidam com este exemplo. Neste caso, a nova proposta tem uma área maior mas é capaz de detetar, especialmente na parte inferior da imagem, uma maior área que está coberta por estas nuvens, ainda que estas sejam muito translúcidas.



(a) Mapa de probabilidade de nuvens s2cloudless imposta na área de exemplo (cor de maior temperatura indica uma probabilidade maior)

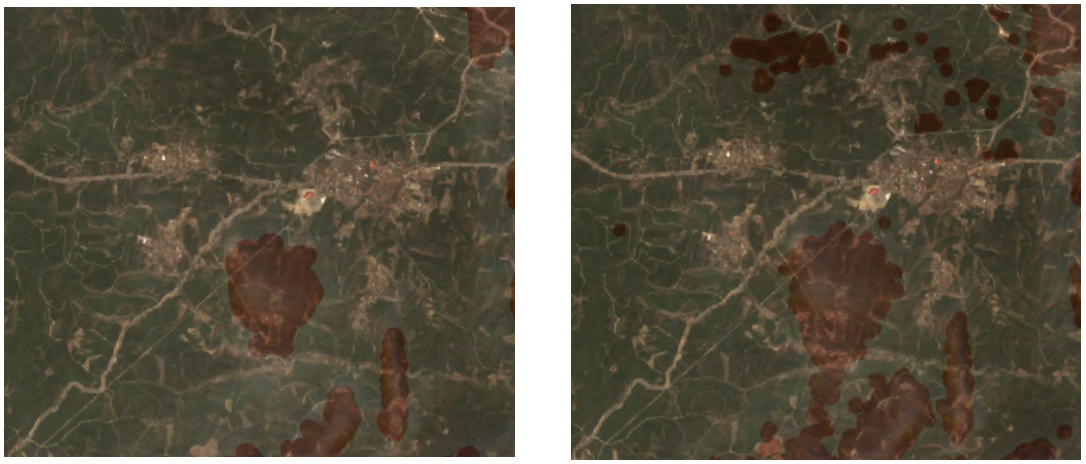
(b) Área de exemplo com a presença de nuvens cirrus

Figura 6.6: Comparação da imagem do Sentinel-2 da área exemplo e do mapa de probabilidades da *s2cloudless*

Para poder avaliar as máscaras de identificação de nuvens foi escolhida uma abordagem que usa o filtro Savitzky-Golay, previamente mencionado, para estabelecer uma referência de uma série suave e com interferência mínima de *outliers*, usando a comparação desta série suave com a série original para avaliar o sucesso que os filtros aplicados têm em remover *outliers*, neste caso nuvens.

O processo segue o seguintes passos:

- Escolher o filtro de nuvens a ser analisado e filtrar as séries
- Gerar uma série "suave"com o filtro Savitzky-Golay para cada série
- Comparar cada série com a sua versão "suave"
- Usar o somatório desta diferença como métrica para avaliar os filtros



(a) A máscara de nuvens proposta pela Google no seu tutorial (área a vermelho)

(b) A máscara de nuvens criada nesta dissertação (área a vermelho)

Figura 6.7: Comparação entre máscaras de nuvens

Analisando as séries das faixas de gestão de combustível de mação para as quais temos dados de referência, obtemos os seguintes resultados: Ao longo de uma série do índice NDVI de um dado grupo existe uma diferença acumulada entre a série suave e a série filtrada com as máscaras de **0.883** para as máscaras da Google e de **0.835** para as máscaras novas desenvolvidas nesta tese, uma melhoria de aproximadamente 5.7%.

É importante ver a quantidade de instantes espacio-temporais que foram ignorados para gerar cada uma destas séries. Para cada máscara os resultados foram os seguintes:

- **Máscara da Google**: 4.90% dos valores filtrados
- **Máscara nova**: 4.62% dos valores filtrados

Pode-se concluir portanto que a nova máscara é mais específica, filtrando um número de valores menores, mas apresentado séries mais suaves, e, crê-se, com menos dados obstruídos pela presença de nuvens.

Além do processo automático de deteção de nuvens, decidiu-se também marcar manualmente as capturas do Sentinel-2 que, para o ano de 2021, aparentam ter a presença de nuvens. A lista de identificadores de capturas do satélite foi incorporada nos *scripts* da plataforma GEE para que os valores dessas capturas de baixa qualidade possam ser marcados como inválidos. Este trabalho manual é importante, pois permite melhorar a qualidade dos dados de deteção remota nos casos onde a deteção automática é imprecisa.

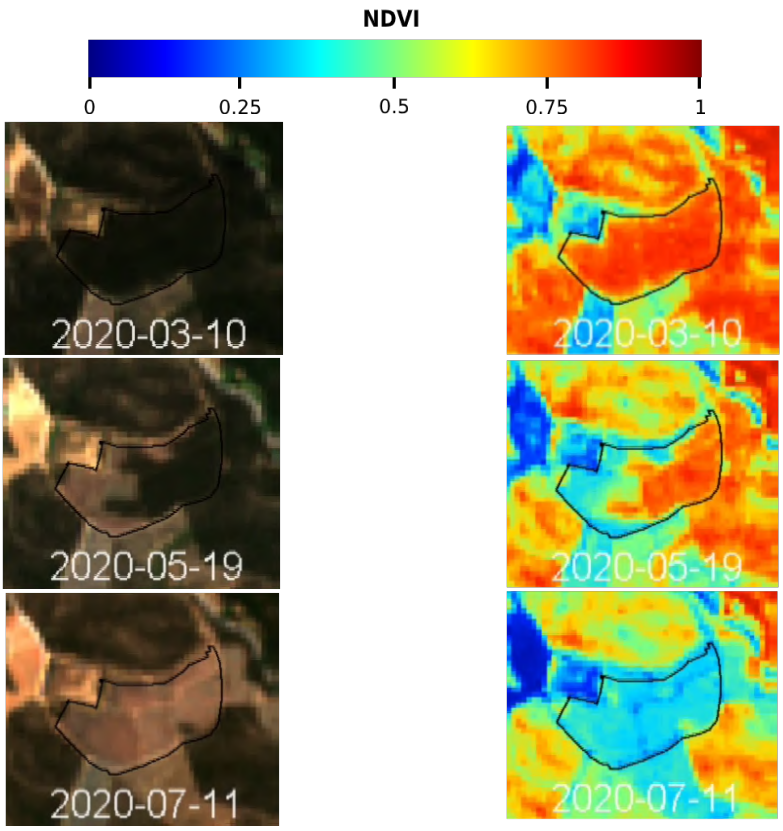
6.3.5 Funcionalidades adicionais

Para melhor entender os produtos de deteção remota foram criadas ferramentas simples que facilitassem a exploração dos dados. Entre os quais destacam-se os seguintes:

- Paletas de cores apropriadas para cada visualização - Foi criado um conjunto de paletas, associado aos valores dos índices mais relevantes de forma a facilitar a visualização dos seus valores de forma mais intuitiva, ou expressiva.

- Ferramenta para extrair vídeos que demonstrem a progressão de uma determinada área ao longo do tempo, sendo sobreposta à imagem a data de cada frame, e a região em análise, que para o contexto desta tese consistia maioritariamente em FGCIs.

Na figura 6.8 podem-se visualizar algumas frames resultantes do vídeo exportado pelo *script* criado no Google Earth Engine, sendo possível entender a progressão da vegetação na zona ao longo do tempo e como isso se reflete nos valores do NDVI, mapeados aqui para uma paleta de temperatura, onde as temperaturas mais quentes indicam um maior teor de vegetação.



(a) Evolução de uma área ao longo do tempo em cores reais

(b) Evolução de uma área ao longo do tempo no índice NDVI

Figura 6.8: Evolução de uma área ao longo do tempo como captado pelo vídeo exportado do Google Earth Engine

6.4 Processamento de séries temporais

6.4.1 Objetivos

Após extrair uma série temporal de uma geometria com recurso ao GEE é necessário proceder ao processamento da mesma, de forma a garantir que a série que vai ser analisada tem uma qualidade superior e que os valores que apresenta são precisos. Além de limpar as séries extraídas pode ser necessário extrair determinadas características adicionais sobre as séries, isto inclui por exemplo a periodicidade das mesmas e a suavidade.

Os principais fatores que influenciam a qualidade da série são interferências atmosféricas, principalmente na forma de nuvens. Como previamente mencionado, é criada uma máscara que cobre as nuvens presentes em cada captura, sendo que as geometrias que intersetam com esta máscara são marcadas como tal nos metadados da série extraída. Com esta informação, e da *blacklist* de capturas criada manualmente, podem-se eliminar os instantes das séries onde existe interferência na forma de presença de nuvens e, portanto, têm valores imprecisos.

Como as nuvens não são o único tipo de evento que invalida as leituras de deteção remota é preciso ter uma abordagem mais geral que permite identificar os pontos cujo comportamento é anormal no contexto da série.

Finalmente, de forma a facilitar o trabalho de análise e aprendizagem, as séries são interpoladas linearmente, garantido que todas as séries têm a mesma dimensão e resolução temporal, permitindo fazer comparações diretas.

6.4.2 Cálculo de novas features

Para facilitar o processo de análise de aprendizagem automática considera-se importante o cálculo de novas características sobre as séries temporais e sobre os instantes destas séries que representam características relevantes para os objetivos deste projeto. Estas novas características podem-se dividir em características ao nível das séries temporais, avaliando características das séries como um todo, e características ao nível do instante temporal, sendo calculado para cada ponto da série, podendo correlacionar-se com os valores na sua vizinhança.

6.4.2.1 Características ao nível do instante temporal

Para os objetivos desta dissertação a característica mais importante é a variação ao longo do tempo, sendo que se identifica um momento de intervenção com uma descida súbita do nível de vegetação. Pretende-se portanto adicionar este contexto temporal a cada instante, sendo feito através do cálculo da derivada da série a cada instante. Esta derivada é calculada de duas formas, primeiramente subtraindo os valores atuais e da última captura, e depois dividindo esse valor pelo tempo decorrido entre cada captura. Se este cálculo for efetuado após o processo de interpolação da série, basta usar a diferença de valores entre cada ponto dado que o intervalo de tempo será sempre o mesmo.

A derivada é importante para determinar a direção e o grau da mudança de valores. No entanto não é suficiente para descrever mudanças que ocorrem ao longo de várias semanas, como é o caso de alguns processos de intervenção. Para poder descrever uma tendência ao longo de um intervalo de tempo superior ao intervalo entre as capturas decidiu-se calcular um atributo que chamaremos de "derivada acumulada", isto é, a diferença entre o valor atual e o valor máximo do mês anterior.

Para além da análise dos momentos anteriores ao instante pode também ser relevante olhar para a frente e avaliar como a vegetação recuperou após uma descida de valor, isto porque uma subida súbita após um instante pode indicar que a descida na verdade correspondia por exemplo, a um erro na série, isto porque o objetivo concreto das operações de limpeza das faixas é reduzir a biomassa presente e garantir que ela se mantenha nesse nível durante o futuro próximo. Esta variável foi chamada de "rise".

A figura 6.9 demonstra os valores destas duas novas características, e a forma como elas se apresentam numa série intervencionada. No instante da intervenção, marcado por uma linha vertical, o valor do declive acumulado é alto, e do *rise* é próximo de zero, como seria esperado.

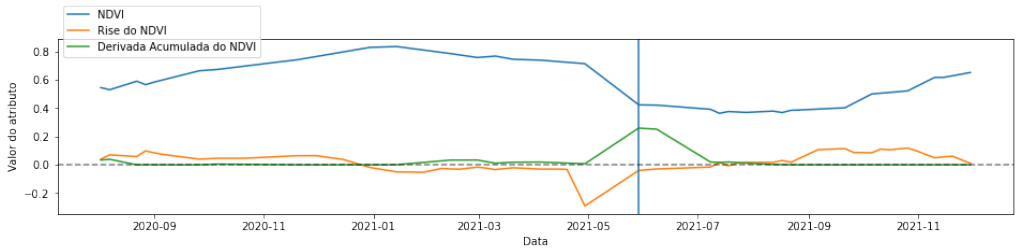


Figura 6.9: Para uma série intervencionada, os valores de NDVI e os valores calculados da derivada acumulada e do *rise*. A data de intervenção encontra-se marcada com uma linha vertical.

Estas duas últimas características, que se relacionam com o comportamento da série num determinado intervalo de tempo serão geradas sobre várias configurações, isto é, o "rise"e o "declive acumulado"serão calculados para intervalos de tempos que precedem ou antecedem, respetivamente, o ponto em análise em 10, 20 ou 30 dias.

6.4.2.2 Características ao nível da série

No processo de análise e marcação manual de instantes de intervenção notou-se que determinadas características da série temporal têm influência para decidir se determinado instante corresponde a um corte ou não. A periodicidade de uma série, isto é, quão semelhante é o seu comportamento entre vários anos é determinante, pois uma série com uma alta periodicidade normalmente corresponde a uma zona agrícola, que não só não costuma ser alvo de intervenções, como a descida dos índices vegetativos corresponde a uma senescência natural e não a uma intervenção humana.

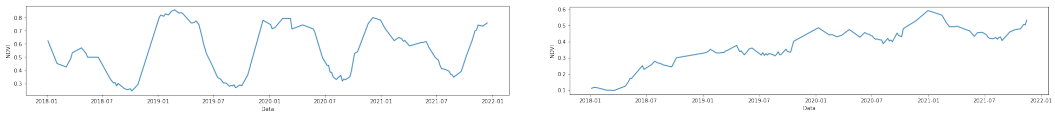
Outro fator a ter em conta é a suavidade da série, isto é, o quão a série segue uma dada tendência. Se uma série de um determinado índice de vegetação tem vários instantes com mudanças abruptas e frequentes de valores, isto pode ser indicativo de uma série com má qualidade, e como tal, tornando mais difícil avaliar um instante de intervenção.

Para determinar a sazonalidade de uma série procedeu-se à separação de cada série nos diferentes anos que ela engloba, no caso em análise foram considerados os anos de 2018 a 2021. De seguida para cada combinação de dois anos é avaliada a diferença entre as série desses dois anos, fazendo no final uma média sobre as diferenças das várias combinações de anos. Para avaliar a diferença entre as séries escolheu-se a métrica DTW, previamente mencionada, pois dá alguma flexibilidade na variação das séries no eixo temporal, isto é, caso um evento fenológico ocorra com algum atraso num dado ano, este processo alinha as duas séries e compara-as em instantes equivalentes do seu desenvolvimento anual.

Para avaliar a suavidade de uma série testaram-se duas abordagens, numa é aplicado o filtro *Savitzky-Golay* previamente mencionado sobre as séries e são comparadas a série original e a série suavizada, usando a métrica de distância euclidiana. A segunda abordagem passa pela análise da derivada da série, calculando a amplitude interquartil da mesma, o que avalia a dispersão de valores da derivada, uma métrica que não tem em conta o contexto temporal. Esta última característica será referida como *IQR* daqui em diante.

Estas novas *features* foram calculadas para todas as séries e exportadas para um ficheiro CSV à parte com o valor final de cada métrica e o id do fragmento de faixa correspondente.

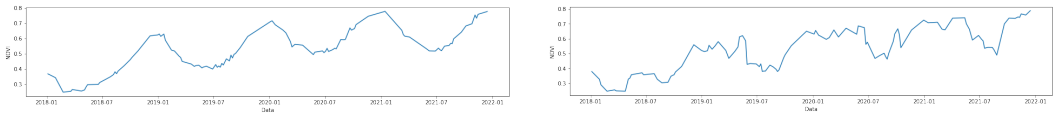
As seguintes figuras demonstram exemplos de séries com valores altos e baixos destas novas características. No caso da periodicidade (figura 6.10), vemos que um valor alto corresponde a uma série com um comportamento muito semelhante e distinto ao longo dos anos, enquanto a de valor baixo tem uma amplitude de valores e variação diferente consoante o ano. Relativamente à medida de suavidade, na figura 6.11 vemos que um valor alto corresponde a uma série relativamente estável, com poucos pontos a distanciarem-se da tendência da série, enquanto um valor baixo é marcado por vários pontos que têm desvios abruptos de valor, sendo mais difícil enquadrar os valores desta série numa tendência.



(a) Exemplo de uma série com um valor alto de periodicidade

(b) Exemplo de uma série com um valor baixo de periodicidade

Figura 6.10: Comparação de séries temporais com valores altos e baixos de periodicidade, respetivamente.



(a) Exemplo de uma série com um valor de suavidade alto

(b) Exemplo de uma série com um valor de suavidade baixo

Figura 6.11: Comparação de séries temporais com valores altos e baixos de suavidade, respetivamente.

6.4.3 Deteção de outliers

Como previamente mencionado existem instantes nas séries temporais em que os valores são inválidos, seja por algum tipo de interferência atmosférica ou por exemplo por a área em análise ser muito pequena e não permitir uma amostragem significante. É necessário identificar estes instantes e removê-los das séries, posteriormente interpolando os valores intermédios.

O problema cinge-se em definir o que podemos considerar um valor anómalo (*outlier*). Está documentado que a interferência atmosférica tem tendência a diminuir os valores de índices de vegetação com o NDVI. Além disso sabe-se que é incomum que num intervalo curto de tempo, digamos 5 dias, o nível de vegetação aumente de forma substancial. Com esta duas informações podemos definir um ponto *outlier* como os pontos com um valor de índice anormalmente baixo e que são seguidos por uma subida súbita para o real valor de vegetação.

É possível que existam pontos outlier consecutivos, causado por exemplo pela presença de nuvens em várias capturas seguidas. Nestes casos os pontos outlier serão seguidos por outro ponto anormalmente baixo e não a subida súbita para um valor "normal"de vegetação. É importante portanto aplicar o método de deteção e eliminação de outliers de forma recursiva de forma a eliminar grupos de valores inválidos das séries temporais.

O processo de deteção de outliers foi aplicado não só ao nível de cada série temporal, mas ao nível global. Isto é, foi calculada uma série composta pela média de valores a cada instante, e posteriormente foi aplicada o mesmo processo de deteção de outliers a esta série. Caso o algoritmo detete um outlier nesta nova série é porque houve um tipo de interferência significativo o suficiente para afetar uma vasta quantidade das áreas em análise. Nestes casos a data correspondente a essa interferência é marcada como uma outlier em todas as séries.

Este processo de deteção global é feito pois caso a interferência tenha uma dimensão relativamente pequena mas seja generalizada, será mais dificilmente detetada numa análise série-a-série.

É apresentado na figura 6.12 um exemplo de uma série com alguns pontos anómalos identificados após um processo de deteção ao nível da série.

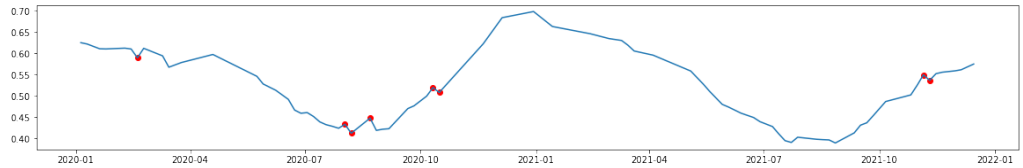


Figura 6.12: Exemplo de uma série com os seus pontos anómalos detetados a vermelho

6.4.4 Processamento adicional

6.4.4.1 Limitação do intervalo temporal

Como previamente mencionado é útil que as diferentes séries tenham a mesma dimensão temporal, e que o intervalo de tempo entre cada ponto seja constante. Para tal é sempre realizado um processamento final de interpolação que preenche os valores em falta em toda a série temporal, sendo que o intervalo entre cada ponto será de 5 dias, o equivalente ao período das capturas na missão Sentinel-2. Os valores contínuos, como os valores dos índices são interpolados, os valores derivados (como a derivada) são recalculados após a interpolação, e os valores discretos e metadados das séries (que incluem o identificador de grupo e a presença de nuvens por exemplo) são definidos como o último valor não-interpolado.

Pode acontecer que as séries das diferentes geometrias em análise, devido à presença de nuvens, e consequente desvalorização desses pontos, tenham um intervalo temporal diferente. Para que se possam comparar todas as séries no mesmo intervalo consideraram-se duas opções.

A primeira opção limita todas as séries ao intervalo temporal que é comum a todas elas, isto permite que todas as áreas em estudo sejam consideradas, ainda que é a opção que elimina a maior quantidade de datas.

A segunda opção ignora as séries *outlier*, que abrangem o intervalo temporal mais curto, e aplica de seguida o processo anterior, limitando as séries ao intervalo temporal em comum das séries restantes. A seleção das séries outlier é feita com a comparação entre a dimensão temporal da série e da dimensão máxima entre todas as séries. As séries cuja data de início ou fim distarem mais de 15 dias do valor da dimensão máxima são excluídas.

No processo de modelação e aprendizagem automática foi usada a primeira opção, pois o intervalo de tempo a ser analisado terminava no fim de 2021. Isto quer dizer que no processo de extração das séries temporais do GEE foram requisitados valores adicionais, de janeiro de 2022, para que os valores em falta no fim de 2021 pudessem ser completados por interpolação com os valores do ano seguinte.

6.5 Enriquecimento de dados de referência

6.5.1 Contexto e Problema

Na informação de referência cedida pela Câmara Municipal de Mação relativamente ao ano de 2021 uma das informações refere-se à data limite de intervenção. Quer isto dizer que esta informação não nos permite identificar o instante exato no qual uma faixa foi intervencionada, apenas um limite superior, assumindo que estes dados estão sempre corretos. Outro detalhe é que a análise feita será relativamente a segmentos de faixas, não à faixa completa, sendo possível, portanto, que dentro de uma mesma faixa os seus segmentos tenham datas de intervenção ligeiramente diferentes, sendo que, por erro de definição da faixa, ou por falta de necessidade de intervenção, pequenos fragmentos das faixas podem não ter sido intervencionados, ainda que a faixa no geral tenha sido.

Para poder criar um modelo que identifique precisamente a data da intervenção, e para a partir de dado modelo podermos deduzir as características que são intrínsecas a um momento de corte, é necessário aumentar os dados de referência que nos foram disponibilizados.

Para comentar estes dados será necessária um trabalho manual de identificação dos instantes que são considerados atos de intervenção, sendo que o conhecimento para os definir advém da análise previamente feita sobre as séries temporais.

6.5.2 Plataforma e Interface

Para o efeito de marcar os instantes de intervenção foi desenhada uma plataforma na qual o utilizador pode marcar manualmente os momentos de corte.

A plataforma visualiza as séries temporais de determinados fragmentos de faixas, permitindo ao utilizador ver a evolução temporal destas áreas em 4 índices espectrais, tal como a derivada dos mesmos. Para contextualizar a informação que está a ser visualizada os pontos que foram detetados como sendo *outliers* ou estando invalidados pela presença de nuvens, têm uma cor diferente, a laranja. Estes pontos não foram eliminados da série para salvaguardar o caso onde um ponto importante seja erroneamente identificado como *outlier*. No fundo de cada gráfico é também possível ver a probabilidade média da presença de nuvens em cada instante temporal.

Ao entrar na plataforma um conjunto de séries é automaticamente carregada, sendo possível, ao usar o botão "Carregar Séries"no canto inferior direito, carregar um novo ficheiro que descreve um conjunto de séries novas. Estes ficheiros resultam de um ligeiro processamento sobre as séries processadas. Neste processamento são apenas seleccionados os atributos relevantes para o processo de marcação, que, para cada instante espacio-temporal são os seguintes:

- O valor dos índices espectrais
- As derivadas dos índices

- O identificador do fragmento
- O dia em questão
- Se o ponto resulta de interpolação
- A probabilidade estimada da presença de nuvens nesse ponto
- Se o ponto é um outlier

Relativamente à informação que indica se o ponto é um *outlier*, este atributo combina o que anteriormente eram 3 atributos, 2 deles que identificavam se o ponto estava obstruído por nuvens e o terceiro que resultava da aplicação do método previamente descrito de deteção de *outliers*.

Após ter seleccionado o conjunto de séries que se pretende analisar pode-se começar o processo de marcação manual das séries. Para marcar os instantes de intervenção basta clicar no instante relevante, sendo possível, usando as teclas "1"e "2", escolher a certeza relativa desse instante de corte, como "certa", ou "incerta", respetivamente.

Cada índice espectral tem um botão correspondente na barra superior, sendo que os índices visíveis estão a azul. Basta clicar nos índices que pretendemos visualizar, ou esconder, para os gráficos na área de trabalho serem atualizados. Além de permitir visualizar diferentes conjuntos de índices em simultâneo, podemos também alternar entre a visualização dos valores originais e da derivada, através de um botão da barra superior que alterna ao clicar entre "Values"e "Delta"para valores e derivada respetivamente.

Para poder visualizar de forma mais detalhada um determinado intervalo de tempo, podemos seleccionar o intervalo de datas pretendidas na barra superior.

Caso se pretenda saltar para uma série específica para fazer a sua marcação podemos usar as caixas de texto para introduzir o identificador dessa série ou a sua posição na lista de séries em análise.

Após marcar os diferentes momentos de intervenção pode-se avaliar a série como um todo, como existindo uma certeza da existência de momentos de corte, uma incerteza geral, ou uma certeza da não existência de cortes. O objetivo desta distinção é para se poder, por exemplo, distinguir uma série na qual se tem a certeza que não existem instantes de corte de outra série na qual não foram marcados cortes mas em que existe alguma incerteza.

Após seleccionar o grau de certeza da série, a próxima série é automaticamente carregada e visualizada, repetindo-se o processo até o utilizador desejar parar. Nesse momento é possível extrair os dados de marcação para uma análise posterior.

Finalmente, após ser feita a marcação das séries que se pretende, e de indicar, para cada uma, o grau de certeza de cada ponto e da série como um todo, podemos exportar os dados, usando um botão da barra inferior, que irá mostrar o JSON que detalha as marcações registadas, cuja estrutura está detalhada no anexo.

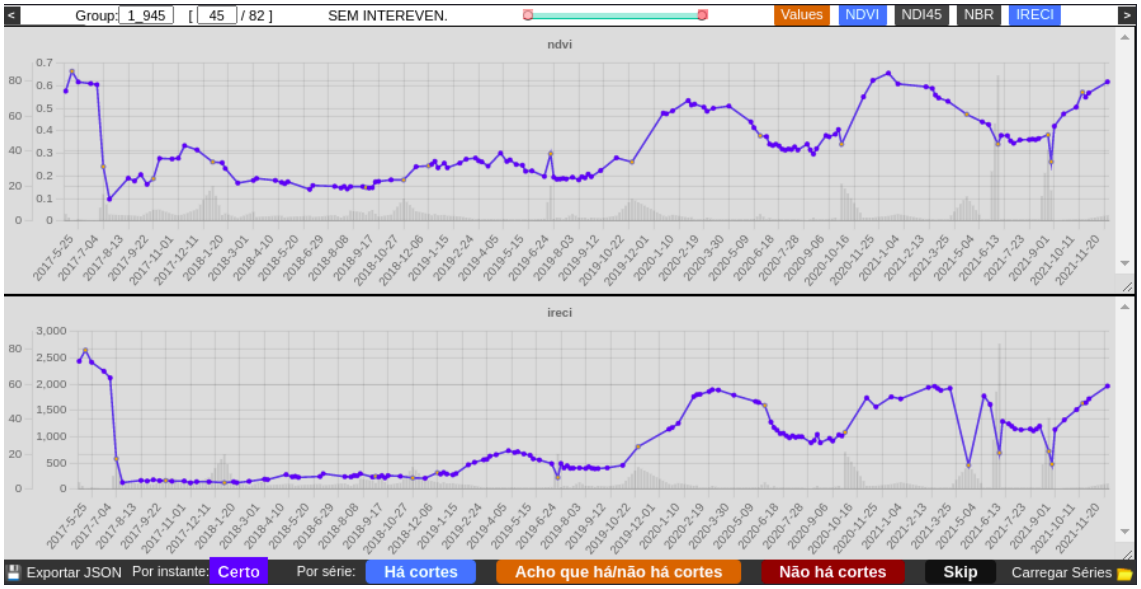


Figura 6.13: Plataforma desenhada para a marcação de instantes de corte

6.5.3 Análise manual de séries temporais

Ao longo do trabalho desenvolvido nesta tese analisei uma grande quantidade de séries temporais, correspondente a faixas intervencionadas e não-intervencionadas, de faixas de classes diferentes, de tipos de vegetação diferente, e com diferentes qualidades. Adicionando a esta experiência o processo de marcar, individualmente, cada momento de intervenção de um conjunto de fragmento de faixa (havendo no total mais de 2 mil fragmentos), acumulou-se um conhecimento intuitivo daquilo que define um instante de intervenção numa série temporal. De uma forma geral pode-se separar em quatro critérios, descritos abaixo. Cada critério influenciam a decisão final, sendo que cada um tem um peso diferente:

- **Concordância**:

Temos vários valores de índices representados, considero que se um evento na série é visível em vários dos índices torna-se mais provável esse evento ser relevante. Não querendo dizer que é necessário que haja uma queda em vários índices para ser considerado um corte. Por vezes uns dos índices têm uma série mais suave que as outras, e considero essas como mais fiáveis e relevantes. É dada alguma prioridade ao NDVI devido ao trabalho prévio que demonstra ser um índice fiável para este tipo de área de estudo.

- **Dados de referência:**

Os dados estão associados a uma ground truth que deverá representar se houve ou não intervenção e a data máxima onde esta ocorreu. Em casos onde há incerteza sobre se existe ou não um evento de corte, esta informação pode servir de desempate.

- **Derivada:**

De longe, o critério mais importante. Define-se uma intervenção como um intervalo de, no máximo um mês (normalmente menos), onde ocorreu uma queda abrupta e minimamente acentuada do valor de um índice de vegetação, sendo que esta queda de valor deverá manter-se durante os instantes após o dito "corte". Caso suba subitamente para valores próximos àqueles antes do "corte"o mais provável é que os valores baixos tenham sido resultado de algum tipo de interferência, nuvens ou algo do género, e não de um corte.

O delta que se considera abrupto tem de ser contextualizado dentro dos valores do resto da série

- **Sazonalidade:**

Determinadas séries, especialmente de terrenos agrícolas têm ciclos muito bem definidos, anuais, ou seja, mesmo que localmente haja uma queda mais abrupta, se for igual ao dos anos anteriores, e se se enquadrar nesse ciclo não considero um corte.

Claro que podem haver faixas que todos os anos são intervencionadas por volta da mesma altura, não é por ter havido uma quebra na mesma altura nos anos passados que se elimina a possibilidade de ser um corte, tem a ver com quão "limpa"e "suave"é a série e quão semelhante é o comportamentos da série entre os vários anos. Quão mais semelhante é, mais provável é que eu considere uma descida que ocorre anualmente como sendo só parte do ciclo natural desse tipo de vegetação.

6.6 Protótipo para geração de faixas de gestão de combustível

6.6.1 Trabalho Prévio

Como previamente mencionado, o trabalho efetuado dentro do âmbito da tese de mestrado de André Santos [32] demonstra que é possível, a partir de imagens de satélite definir de forma algo fiável estruturas artificiais permanentes. A implicação desta conclusão é que será possível com os dados resultantes deste estudo definir as áreas correspondentes a cidades e casas isoladas.

6.6.2 Objetivo

Com base nos resultados na identificação de estruturas permanentes pretende-se isolar apenas as estruturas referentes a cidades e casas isoladas para a partir dessa definição calcular as faixas de gestão de combustível que deveriam, em teoria, existir em seu redor. Isto possibilita a comparação entre as faixas oficialmente disponibilizadas no PMDFCI de cada município com estas faixas geradas, para poder avaliar a completude das faixas oficiais, especialmente das faixas correspondentes a casas isoladas.

6.6.3 Metodologia

Após seleccionar as classes de classificação que nos interessam é necessário remover as estradas, deixando apenas os aglomerados populacionais e habitações. Este processo pode ser feito recorrendo à informação do Open Street Map, recolhendo a informação geográfica da área em estudo, de forma a remover da classificação automática as zonas pertencentes a estradas, posteriormente eliminando qualquer artefacto resultante desta exclusão. A informação do Open Street Map será também usada, através da informação das localidades para ajudar a definir que zonas correspondem a aglomerados populacionais.

Tendo agora um ficheiro geográfico que corresponde às áreas do nosso interesse, pretendemos distinguir as casas isoladas dos aglomerados populacionais, de forma a poder aplicar as regras definidas no PMDFCI respetivamente. Esta distinção é feita através de 3 critérios, a proximidade entre vários fragmentos, a área dos mesmos, e a interseção com a zona correspondente a uma localidade, tal como definida no Open Street Map.

Após este processo de identificação são criados os *buffers* finais, correspondentes à largura das faixas de cada classe. Sendo depois aplicado um processamento final para remover faixas inválidas, e interseções com estradas.

6.7 Construção dos dados de treino

O processo de marcação manual dos momentos de intervenção tinha como objetivo gerar dados para o processo de aprendizagem. Fica-se assim com dados de treino mais precisos, com a data exata da ação de intervenção, que são necessários para treinar os modelos de aprendizagem, e para posteriormente refletir sobre a qualidade das marcações e os critérios usados nesse processo.

Os dados que serão usados no processo de treino do modelo são aqueles do concelho de Mação, sendo que o modelo que será treinado para esta região do país será posteriormente testado para a região de Santarém de forma a avaliar a capacidade de generalização do modelo.

Antes de começar o processo de modelação de dados é necessário criar um conjunto de dados de treino. Este conjunto é composto por pontos positivos, correspondentes a um instante de corte, e negativos, onde não ocorreu intervenção.

Os pontos positivos são seleccionados a partir dos instantes de corte marcados manualmente, sendo apenas escolhidos os pontos nos quais não havia incerteza no momento de marcação.

Os pontos negativos são amostrados dos restantes pontos das séries, que deve obedecer a dois critérios:

- Pertence a uma série que não foi marcada como incerta

- O instante temporal está a mais de um mês de distância de uma marcação de corte (seja certo, ou incerto)

A razão pela qual apenas se consideram pontos negativos aqueles que distam um mês de instantes de corte é porque os processos de manutenção podem durar mais do que os cinco dias que distam cada captura. Isto quer dizer que, além do instante de intervenção marcado podem existir ao seu redor outros pontos que seriam corretamente classificados como "intervencionados", até porque os pontos vizinhos são mais prováveis de ter as mesmas características do ponto de corte. Para evitar estes pontos sejam considerados são apenas seleccionados os pontos negativos que distam mais de um mês de qualquer ponto positivo.

6.8 Modelação

Relativamente aos processos de aprendizagem automática, estes dividem-se em dois, os de aprendizagem supervisionada e não-supervisionada. Relativamente aos algoritmos supervisionados, foram considerados classificadores para detetar o estado de intervenção de um dado fragmento num dado instante de tempo, sendo retornando um valor booleano. Os processos não-supervisionados foram usados para automaticamente separar as diferentes séries em *clusters*, que se esperam reter algumas características fundamentais das séries, nomeadamente o seu tipo de vegetação e, mais importante, se foram intervencionadas ou não.

Além dos parâmetros internos do modelo, existem hiper-parâmetros e configurações de input do modelo que não são ajustados automaticamente no processo de aprendizagem. Estes hiper-parâmetros são definidos pelo utilizador, sendo reavaliada cada mudança de parâmetros de acordo com o seu efeito na precisão do modelo.

Para facilitar a definição destes hiper-parâmetros é usado o Grid Search, previamente mencionado em 5.1.4. Após a afinação destes parâmetros o modelo é avaliado recorrendo aos dados de teste.

As técnicas seleccionadas para esta fase de aprendizagem foram **Support Vector Machines**, **Random Forests** e **XGBoost** para classificação, e para a aprendizagem não-supervisionada, **KMeans**, **Linkage Clustering** e **KShape**.

Os hiper-parâmetros que serão testados são os seguintes:

- **Support Vector Machines** - kernel={'linear','sigmoid','poly','rbf'}, C={1,10,50,200}
- **Random Forests** - criterion={'gini','entropy'}
- **XGBoost** - booster={'gbtree','gblinear','dart'}, max_depth={4,8,12,16}
- **KMeans** - distance_measure={dtw, softdtw ou euclidiana}, n_clusters={2,3,5,7,9,12}
- **KShape** - n_clusters={2,3,5,7,9,12}

Tendo sido seleccionados 3 principais algoritmos para a tarefa de clustering, serão efetuadas experiências com intuito de avaliar se estes métodos se adequam para separar as séries intervencionadas de não intervencionadas, e como os parâmetros e dados de input influenciam o sucesso do modelo.

Os diferentes parâmetros dos métodos serão comparados, de forma a determinar que parametrização obtém os melhores resultados, sendo que a métrica de avaliação escolhida foi a homogeneidade. Além dos parâmetros dos algoritmos de clustering, também serão variados os dados usados para o processo de clustering, dentro dos índices espectrais escolhidos (**NDVI**, **NDI45**, **IRECI** e **NBR**), e características derivadas dos mesmos, neste caso o **declive** e o **declive acumulado**.

Para os processos de classificação é necessário dividir os dados em grupos de treino e de teste. Isto para que, os dados usados para ajustar o modelo não sejam os mesmos a ser usados na classificação, prevenindo o enviesamento do modelo. Decidiu-se que 80% dos dados seriam dados de treino e 20% dados de teste, sendo que os dados de treino são treinados em cross-validation, como na figura 6.14.

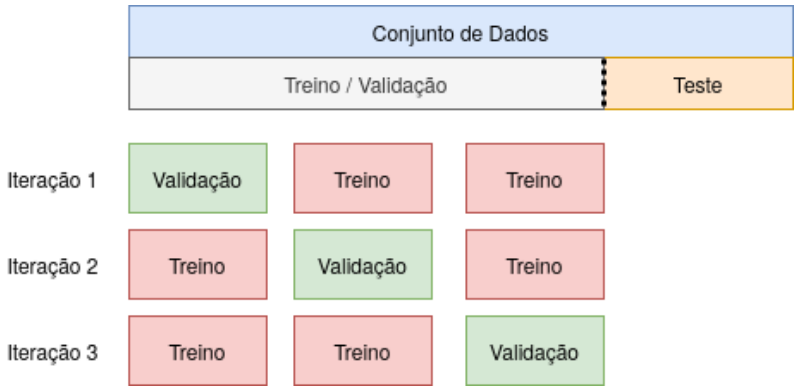


Figura 6.14: Iterações de cross-validation com uma fração de validação de 1/3

6.9 Avaliação

6.9.1 Seleção de Ground Truth

A *Ground Truth* usada para alimentar os modelos de classificação automática são gerados manualmente, para as séries da zona de Mação, guiados pelo conhecimento acumulado sobre os efeitos dos processos de intervenção nos índices espectrais, e das diferenças gerais entre uma série que representa um crescimento ou decréscimo natural de vegetação e uma mudança artificial de comportamento da vegetação.

É importante notar que este tipo de análise irá, inevitavelmente ter erros, seja porque um decréscimo natural foi tão abrupto que foi interpretado como intervenção, seja porque algum erro ou interferência do sinal obscureceu o instante de um corte, ou por simples erro humano, causado por distração, como um clique acidental num instante temporal errado.

Os processos de aprendizagem não servem só para avaliar o modelo que identifica os momentos de intervenção, serve também para refletir sobre as pressupostos que foram tomados no processo de marcar as séries. Previamente foi definido um conjunto de características que servem de guia para identificar um instante de corte, como o decréscimo súbito de vegetação e a periodicidade da série por exemplo. Ao analisar as séries marcadas podemos verificar a relevância que cada uma dessas características realmente teve na decisão de identificar um instante de corte.

Ao alternar o conjunto de dados que é usado para modelar o classificador, podemos identificar o conjunto de atributos que melhor se adequa à tarefa. O resultado das classificações pode também ser útil para detetar e corrigir a *ground truth* gerada. A análise dos falsos positivos pode identificar exemplos que na verdade estão corretamente identificados, por exemplo, que caso ocorra, demonstrará que o modelo está bem ajustado.

6.9.2 Clustering

Como previamente mencionado a métrica destacada para a avaliação dos clusters será a homogeneidade, tendo em conta a classe do estado de execução das faixas.

No entanto, este valor não é suficiente para expressar a distribuição dos valores entre os clusters. Com cada experiência de clustering será exportado uma tabela resumo que, para cada cluster, identificará o nº de elementos, e a percentagem dos mesmos que corresponde a uma série intervencionada.

Estes valores precisam de ser contextualizados, tendo em conta a distribuição de séries intervencionadas e não-intervencionadas no conjunto de dados total, mas ainda assim permite verificar de que forma o algoritmo de clustering está a separar as diferentes classes.

Parte da avaliação consiste na visualização dos clusters com melhor performance, de forma a melhor entender a distribuição das séries. Isto é importante para podermos avaliar o processo de clustering pois permite-nos melhor entender que características estão a ser separadas pelo processo, e em que situações é que existe uma maior, ou menor, homogeneidade.

6.9.3 Aprendizagem

Como foi descrito na secção do Estado da Arte, existe um grupo de métricas que devem ser aplicados para avaliar o desempenho dos algoritmos de classificação automática, nomeadamente o **F1 Score** e os valores da **Matriz de Confusão**. Relativamente à matriz, é preciso ter em conta o objetivo da classificação para melhor poder avaliar os resultados. Uma das formas de abordar este problema é como uma tentativa de identificar as áreas onde não houve intervenção, para que se possa averiguar o estado da vegetação em pessoa. Nesta situação, um Falso Negativo é mais danoso que um Falso Positivo, isto porque será preferível que se verifique o estado de uma faixa que se acredita não estar tratada quando na verdade o está, do que se acreditar que determinada área está adequadamente mantida,

quando na verdade não sofreu intervenções e pode estar a representar um risco para as habitações junto das faixas.

Como foi previamente mencionado em 6.7, o processo que elimina os pontos ao redor de um instante de corte para treino é lógico, no entanto tem consequências no processo de aprendizagem. Isto porque, devido a este processo, cada intervenção apenas será representada num único ponto. Mas no caso de um modelo detetar um falso positivo esse erro provavelmente será repetido em vários instantes, dado que, como previamente mencionado, os pontos vizinhos provavelmente terão características semelhantes, e portanto, seriam igualmente classificados como falsos positivos. Este desequilíbrio entre falsos positivos e verdadeiros positivos pode influenciar as métricas de avaliação a classificar um modelo como pior do que ele realmente é.

Para colmatar este último problema aplicamos a mesma lógica que foi usado nos pontos de intervenção nos falsos positivos resultantes da classificação. Isto é, todas as classificações positivas numa janela de um mês são agregadas num único ponto. Este método é exemplificado na figura 6.15, na qual o modelo identificou duas intervenções, e como ambas estão num intervalo de tempo inferior a um mês, a primeira (marcada a verde) é seleccionada e a segunda (a vermelho) é eliminada.

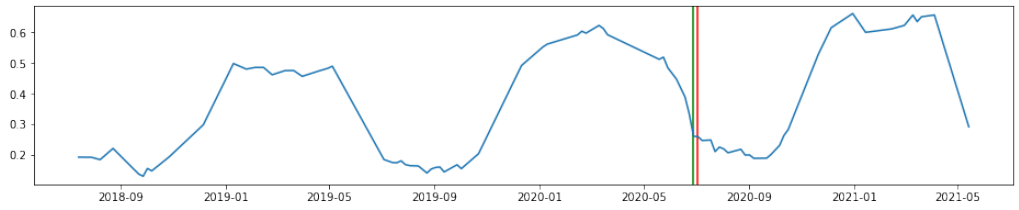


Figura 6.15: Série temporal de uma faixa com duas classificações falso positivas de intervenção. Para efeitos de avaliação a data a vermelho irá ser eliminada e a data a verde mantida

6.9.4 Comparação com dados de referência

Os modelos acima mencionados serão treinados e avaliados com base na marcação manual feita sobre as séries dos diferentes fragmentos do concelho de Mação. No entanto, é preciso validar este modelo com recurso aos dados de referência que nos foram fornecidos relativamente aos concelhos de Mação e de Santarém, pois assumir que os dados manualmente gerados estão inerentemente corretos seria um erro na metodologia de avaliação.

Como tal, os resultados de avaliação dos diferentes fragmentos serão comparados com o resultado esperado da faixa como um todo. No caso do exemplo da figura 6.16, estão representadas 5 faixas, que, de acordo com os dados de referência não sofreram intervenção no ano de 2021. Os fragmentos analisados de quase todas elas confirmam que não ocorreu nenhum intervenção no ano de 2021. Em uma delas, dois fragmentos dos dez que compõem a faixa foram avaliados como tendo sofrido uma intervenção nesse

ano. Nesta situação a forma como se compara a avaliação dos fragmentos com os dados de referência é através da agregação dos dados de avaliação ao nível das faixas, assumindo a avaliação (intervencionado ou não-intervencionado) que for maioritária na faixa.

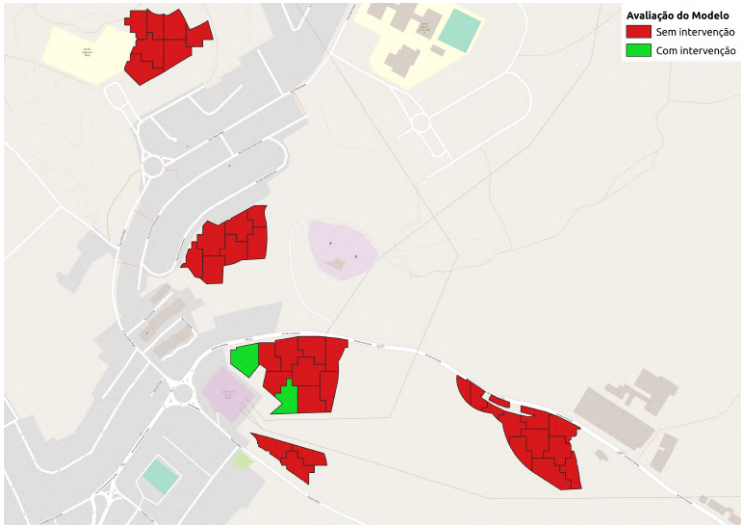


Figura 6.16: Exemplo do resultado do modelo de avaliação em vários fragmentos na área de Santarém

A abordagem descrita para o exemplo anterior é aquela que será aplicada para avaliar o modelo tanto no concelho de Mação como no de Santarém.

6.9.4.1 Dados de Mação

Relativamente aos dados de Mação, é possível não só comparar os resultados com os dados de referência como é possível ainda comparar com as marcações feitas manualmente na região, permitindo averiguar se a metodologia seguida para as mesmas é ou não apropriada.

6.10 Conclusões

Neste capítulo definiu-se toda a metodologia do trabalho, desde o processamento dos dados, sejam georreferenciados ou de deteção remota, ao enriquecimento dos dados de referência através da marcação manual das séries, passando pelo desenvolvimento de um protótipo para a geração automática de faixas e finalmente pelos métodos usados para treinar e avaliar os modelos de aprendizagem, sejam eles supervisionados ou não-supervisionados.

Definiu-se a forma como processar as FGCI, separando as diferentes classes de faixa, enriquecendo a sua informação e unindo as geometrias pelo identificador.

Estando estas faixas processadas explicamos de seguida a forma como elas devem ser fragmentadas em áreas de análise mais pequenas. Dividindo as faixas de acordo com a

ocupação do solo, a distância ao interior da faixa e garantido que estes fragmentos têm uma dimensão adequada.

Além do processamento das FGCI definidas nos planos de defesa, também os dados de referência (dados georreferenciados) necessitam de ser transformados. É criado um identificador adequado a partir do qual são unidas as geometrias, e seleccionam-se as faixas que têm informação útil em relação à data limite da sua intervenção. Além disso, explica-se em detalhe a forma como irão ser criados os fragmentos que servem como áreas de estudo.

Dado que a base deste trabalho é a deteção remota, e o uso de informação de satélites, é mencionado o processo necessário para extrair essa informação remota através da plataforma Google Earth Engine, sendo detalhados as principais problemáticas associadas, como a presença de nuvens nas capturas, e as abordagens tomadas para minimizar os seus efeitos.

Após serem extraído os dados de deteção remota da plataforma é necessário processar estes dados, sendo explicado a forma como se identificam os pontos anómalos das séries temporais, e o tipo de características que são calculadas a partir dos dados exportados, como a derivada e a periodicidade da série.

Como os dados de referência não indicam o instante exato de uma intervenção essa marcação tem de ser feita manualmente, que é um processo moroso. Explica-se portanto os passos tomados para acelerar este processo, criando-se uma plataforma de anotação de séries temporais. É também explicado as características das séries que se consideraram mais importantes para classificar um instante de intervenção.

Tendo sido notado que haviam algumas faixas em falta nos planos de defesa aborda-se aqui a forma como se podem gerar automaticamente FGCI a partir da definição de zonas artificiais.

Sendo o principal objetivo desta dissertação a criação de modelos de aprendizagem automática define-se no fim deste capítulo a forma como serão criados os conjuntos de dados para os modelos seleccionados, a configuração dos modelos e finalmente, a forma como são avaliados os resultados tanto dos classificadores como dos algoritmos de agrupamento.

7

Implementação

7.1 Gestão e transformação dos dados geográficos

Os dados geográficos que são utilizados neste trabalho incluem os seguintes:

- Faixas de Gestão de Combustível, tal como definidas nos PMDFCI
- As secções de faixas marcadas com a data limite de intervenção (relativamente ao concelho de Mação)
- As áreas intervencionadas num dado ano (relativamente ao concelho de Santarém)
- Os fragmentos de FGCI que servirão de base para os processos de análise de séries temporais
- A classificação automática de estruturas artificiais na região da Guarda.
- As estradas e localidades, como definidas pelo projeto OSM
- A COS da zona de interesse

Numa primeira fase foram exportados os dados dos PMDFCI de todo o país, seleccionando-se apenas os ficheiros geográficos que definiam as FGCI. Esta tarefa foi efetuada com recurso ao servidor FTP disponibilizado pelo ICNF, sendo feita uma pesquisa em profundida pelos ficheiros relevantes e de seguida guardando-se localmente os *shapefiles* organizados pelo distrito e concelho a que pertencem.

As FGCI de Mação e da Guarda, os concelhos que são alvo de estudo, são incorporadas numa base de dados PostgresSQL, processadas pelos métodos descritos em 6.2.1.1, sendo guardada uma nova versão dos ficheiros georreferenciados na mesma base de dados.

A utilização da base de dados Postgres, e a sua extensão geoespacial PostGIS, deveu-se ao conjunto de ferramentas de manipulação de dados geográficos que disponibiliza, facilitando o trabalho de manipulação das geometrias e permitindo guardar os vários passos do processamento de forma organizada numa única base de dados. Ao integrar várias fontes

de dados numa única base de dados permite correlacionar várias informações diferentes e facilmente manipular diferentes conjuntos de dados geográficos simultaneamente.

Os dados de referência disponibilizados do concelho de Mação e Santarém serão fragmentados como previamente descritos em áreas de análise mais pequenas. A interoperabilidade entre os vários dados de informação facilita este processo, dado que tanto as faixas oficiais (para o cálculo da distância ao exterior da faixas de Mação), como a COS da região já estão integradas na base de dados.

De uma forma geral, pretendeu-se que cada passo dos processos de transformação de dados geográficos fosse guardado na base de dados, isto para facilitar a identificação e correção de erros, evitando ter de refazer todo o processamento.

7.2 Visualização de dados

Para melhor entender os dados trabalhados e os resultados obtidos foram usadas ferramentas que permitem visualizar e interagir com informação geográfica e estatística. Isto é importante para poder, por exemplo, visualizar o tipo de terreno associado a determinadas áreas (com recurso às ortofotos e ao OSM) e ver a evolução destas áreas ao longo do tempo, permitindo uma análise espácio-temporal das áreas de estudo.

7.2.1 QGIS

O QGIS é uma ferramenta de sistema de informação geográfica grátis, de código fonte aberto, que permite visualizar e manipular informação georreferenciada. Foi usada para visualizar informação geográfica relativamente às FGCI, permitindo contextualizá-las com recurso aos dados do OSM e das Ortofotos disponibilizadas pela DGT.

Este software possui também a possibilidade de se conectar a uma base de dados, tendo sido usada para importar dados geográficos e executar queries na base de dados Postgis, permitindo visualizar rapidamente o resultado das mesmas.

7.2.2 Tableau

O Tableau é um software que permite visualizar e analisar dados em variados formatos de visualização, possibilitando também calcular novos atributos e estatísticas que sejam relevantes.

Esta ferramenta suporta também dados geográficos, e visualizações interativas, o que permite, por exemplo, importar os dados geográficos das faixas, após serem processadas, e visualizar, com um esquema de cores, o seu estado de intervenção, ou ao seleccionar determinada faixa ver a série temporal de um determinado índice de vegetação.

Finalmente, permite-nos visualizar os resultados das experimentações, e correlacionar os diferentes parâmetros dos modelos com as métricas de avaliação.

7.3 Análise e modelação de dados

Após os dados geográficos terem sido corretamente processados e a informação temporal ter sido extraída da plataforma Google Earth Engine, resta fazer a análise desta informação num ambiente local.

Um ambiente local dá uma maior flexibilidade sobre as ferramentas que disponho para analisar e modelar os dados, ao contrário do ambiente do GEE, que tem uma interface muito limitada.

Enquanto as características da máquina local são suficientes para os objetivos pretendidos, foi necessário ter um maior cuidado relativamente à performance das operações de forma a acelerar o processo de análise.

7.3.1 Python e Bibliotecas relevantes

Para otimizar os processos de aprendizagem, modelação e transformação de dados usaram-se bibliotecas que fornecem interfaces para operações comuns, com boa performance.

Das bibliotecas para a transformação de dados destacam-se a **pandas** e **numpy** para a manipulação de dados em formato de tabela e **geopandas** e **osgeo** para a importação de dados geográficos para alguma análise dos dados geográficos e dos seus atributos.

Relativamente à visualização de dados, **matplotlib** e **seaborn** disponibilizam métodos para diversos tipos de visualização.

Para criar e treinar modelos de aprendizagem e clustering, foram usadas as bibliotecas comuns **scipy** e **scikit-learn**, tal como algumas bibliotecas específicas para a análise de séries temporais **pyts**, **sktime** e **tslearn**.

Foram também criados módulos de python que incluem a definição de funções que são frequentemente usadas no processo deste projeto, como a interpolação de séries temporais, a suavização de séries, a deteção de outliers, entre muitos outros. Os *scripts* responsáveis pelo processamento e exportação das séries recorrem a estes módulos para fazer grande parte do processamento.

7.4 Fluxo de processamento de séries temporais

Nesta secção irá descrever-se a metodologia usada e o passos que envolvem o processamento das séries temporais, desde a extração dos valores de deteção remota à visualização dos resultados. Para contextualizar o processo, de forma muito geral, os passos essenciais para a criação dos dados de referência e de treino são os descritos pela figura 7.1

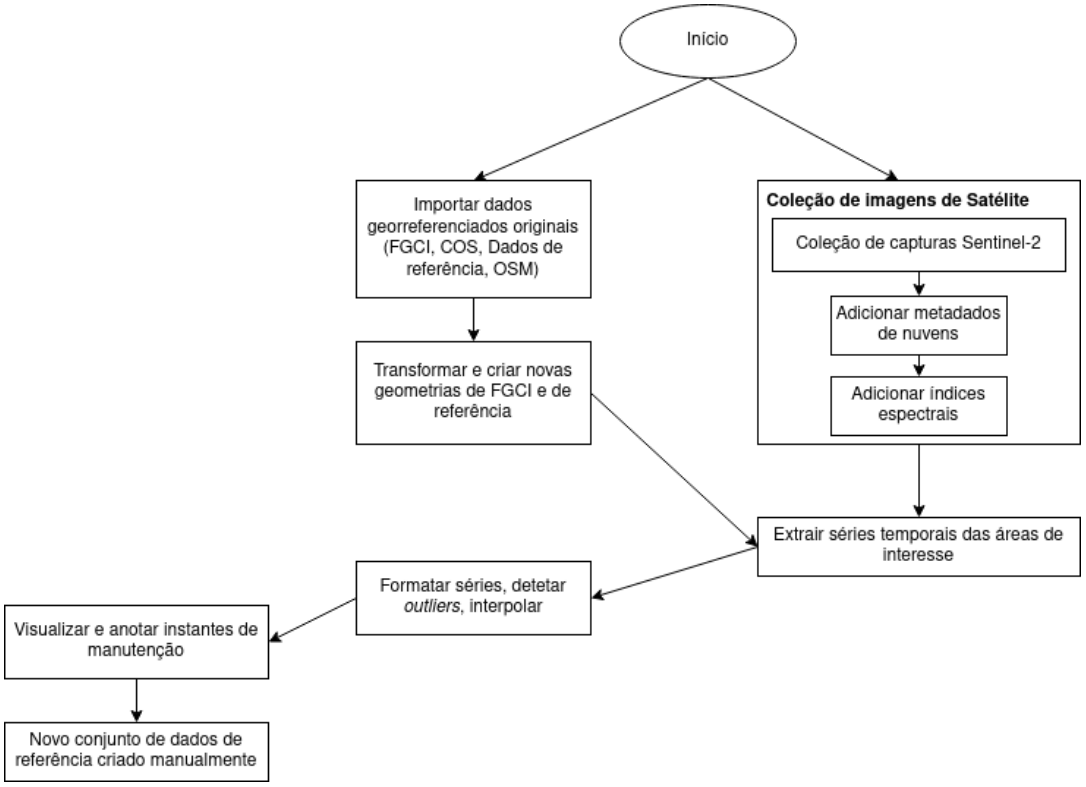


Figura 7.1: Metodologia geral para o processamento da informação geográfica e de deteção remota

7.4.1 Exportação de séries temporais

Toda a manipulação direta com imagens de satélites e extração de informação de deteção remota é feita na plataforma Google Earth Engine. Aí são introduzidas as geometrias que foram previamente processadas com o intuito de exportar, para cada geometria, e para cada instante de tempo, um valor por atributo. Esta informação é agregada numa tabela, sendo exportada para uma pasta do Google Drive.

Notou-se experimentalmente que exportar uma grande quantidade de séries temporais demorava muito tempo, e aumentava a probabilidade de ocorrer algum erro no processamento, que caso ocorresse o processo teria de ser repetido desde o início, desperdicando todas as séries que já tinham sido processadas. Decidiu-se assim criar conjuntos das geometrias para as exportar separadamente. Foram criados grupos de 100 fragmentos, gerados aleatoriamente, mas com uma *seed* fixa, para caso seja necessário resumir a exportação noutro instante. Estes conjuntos de 100 fragmentos são iterativamente exportados, através da função desenvolvida para o efeito. Sendo depois, em ambiente local, unidos de novo num único ficheiro que agrupa todas as séries temporais.

7.4.2 Integração dos dados e processamento

Localmente trabalhou-se em ambiente Python para processar os dados que foram obtidos do GEE. Numa fase inicial o objetivo é, como já foi mencionado, eliminar os *outliers*, calcular os atributos novos e guardar esses dados processados.

A fase inicial consiste em calcular o valor da derivada a cada instante. Neste contexto ainda não foi feito nenhum processo de interpolação e a distância entra as capturas não é uniforme, logo a derivada que aqui se pretende calcular resulta simplesmente de, para cada grupo, a subtração do valor anterior, dividido pelo intervalo temporal.

Depois de obter esta derivada começa o processo de deteção de outliers, primeiramente é efetuada a deteção global de outliers, e apenas depois a deteção ao nível da série. Este processo depende do valor da derivada do índice NDVI calculada anteriormente, sendo que, a deteção de um *outlier* consiste em detetar a subida íngreme de um valor com interferência para um valor normal. Isto porque as interferências atmosféricas, nomeadamente a oclusão por nuvens, têm um efeito negativo nos valores dos índices como o NDVI. Para determinar o que é uma subida abrupta, calcula-se o desvio padrão dos valores da derivada, e são excluídos os elementos que excedem um produto deste desvio padrão, definido experimentalmente a **0,75**.

Após ser detetado os outliers globais, isto é, da série composta pela média de todas as séries, passa-se a uma análise grupo-a-grupo, detetando os instantes *outliers* em cada série, sendo que o processo é sensivelmente o mesmo do processo global.

Após serem detetados os *outliers*, estes não são eliminados, mas sim marcados como tal num atributo novo do instante.

O processo de deteção de outliers é repetido recursivamente, sendo ignorados os outliers anteriores no processo, até já não ser detetado mais nenhum instante.

A estes dados são adicionados um conjunto de novas características, como o declive, o declive acumulado e o "rise", já previamente descritos em 6.4.2.1. As fórmulas que permitem descrever estes novos atributos estão exemplificados nas equações 7.1, 7.2 e 7.3, para o exemplo do NDVI e calculando o declive e o "rise"para um intervalo de 30 dias.

$$\text{Declive} = \frac{NDVI - NDVI_{anterior}}{\Delta dias} \tag{7.1}$$

$$\text{Declive Acumulado} = NDVI_{atual} - Max(NDVI_{-30\ dias} ... NDVI_{atual}) \tag{7.2}$$

$$\text{Rise} = Max(NDVI_{atual} ... NDVI_{+30\ dias}) - NDVI_{atual} \tag{7.3}$$

Os dados finalmente são interpolados, garantindo que entre cada instante existe um intervalo de 5 dias, igual para todas as séries, e finalmente são recalculados os atributos da derivada e do declive acumulado para cada índice e exportados os dados para tabelas CSV.

Relativamente às características ao nível da série, como a periodicidade da séries, estas são calculadas após o processamento das séries e com base numa série que exclui os pontos erróneos.

Para o valor da periodicidade, este é calculado usando a comparação das séries dos vários anos em análise. No caso desta dissertação referimo-nos aos anos de 2018 a 2021. Em alguns casos, foi ignorado o ano de 2018, pois algumas faixas foram vítimas de incêndios florestais neste ano, quebrando a variação de vegetação que seria esperada num ano normal. O cálculo da periodicidade começa por separar as séries nos vários anos e aplicar um filtro de *savitsky-golay* para as suavizar. Obtendo estas séries foi, para cada combinação de anos, aplicada a métrica DTW para definir a distância entre cada par de séries anuais. O inverso da média destas diferenças resulta na métrica de periodicidade.

7.4.3 Exportar séries processadas

Os dados são exportados para 3 ficheiros diferentes, um inclui toda a informação, com os dados detetados como sendo de nuvens, e os outliers, e todos os metadados relativamente à captura, chamado de **extra**. Outro ficheiro, chamado de **clean** inclui toda a informação, mas apenas os instantes de tempo que não são *outliers* nem estão contaminados por nuvens, sendo os restantes valores interpolados. Um último conjunto de dados chamado de **essential** exporta apenas os dados da deteção remota e os seus valores derivados, combinando os metadados de deteção de outliers e nuvens em um só, e ignorando, por exemplo, os metadados que identificam a imagem original da captura do Sentinel-2.

7.4.4 Importação de dados e análise

Para facilitar a análise dos dados e evitar ter de repetir código para operações comuns foi criado um módulo que incorpora um conjunto de ações que facilitam a visualização e exploração dos dados previamente exportados.

Os dados são estruturados numa classe, sendo possível facilmente filtrar os grupos (também chamados de fragmentos), que se pretendem analisar, seleccionar séries pelo seu estado de intervenção e visualizar as séries com diferentes parametrizações. O excerto de código abaixo, por exemplo, visualiza as séries pelo índice ndi45, marcando as séries intervencionadas a vermelho e não-intervencionadas a azul e aplicando um filtro *default* de Savitsky-Golay sobre a série para a suavizar.

A chamada da função anterior resulta no gráfico representado na figura 7.3.

```
series = homes.get_series(measure="ndi45",smooth=True)
options = {
    "plot_params":{"alpha":0.2},
    "group_colors":{"red":homes.cut_groups,"blue":homes.non_groups}
}
graphSeries(series,homes,options)
```

Figura 7.2: Excerto de código que permite visualizar séries temporais de várias FGCI com uma parametrização específica

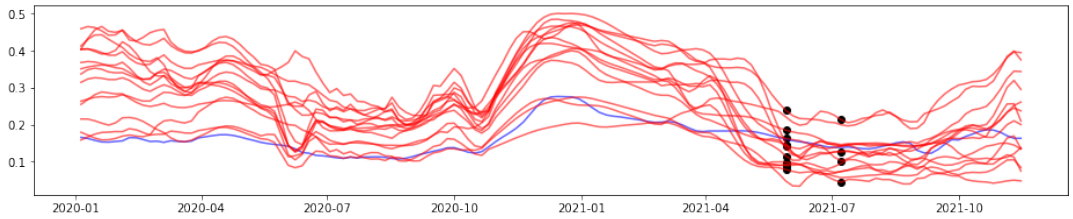


Figura 7.3: Exemplo da visualização de um conjunto de séries de acordo com a parametrização anterior. Data limite de intervenção identificada pelos pontos a preto.

8

Trabalho Experimental e Resultados

8.1 Introdução

Na fase experimental foram usadas as séries temporais extraídas dos produtos Sentinel-2 para identificar intervenções sobre as FGCI em análise. Consideraram-se duas abordagens: a primeira, recorrendo a métodos de classificação, pretende identificar o instante exato no qual ocorreu um corte sobre a faixa. A segunda abordagem, recorrendo a métodos não-supervisionados de *clustering* pretendia automaticamente gerar *clusters* de séries que, idealmente, conteriam apenas faixas intervencionadas, ou não intervencionadas. Ou seja, uma análise ao nível da série como um todo. Os métodos de aprendizagem automática foram aplicados sobre os dados relativos à zona de Mação, sendo que os modelos de classificação supervisionada foram posteriormente testados para faixas na região de Santarém, de forma a avaliar a sua capacidade de generalização.

Para cada abordagem, e para cada algoritmo foram testados um conjunto de parametrizações e dados de treino de forma a encontrar a configuração que permitia obter os melhores resultados.

Os métodos de aprendizagem foram aplicados às faixas da região de Mação e testado na região de Santarém para avaliar a capacidade de generalização das abordagens escolhidas.

8.2 Dados Relevantes

A informação das séries temporais foi extraída através da plataforma Google Earth Engine para cada fragmento em análise dos dados de referência de Mação e de Santarém. Em ambos os casos foram usadas capturas desde o **início de 2018 até ao fim de 2021**, sendo que os dados de referência cedidos pela CMM apenas são referentes a 2021, mas que as datas de intervenção manualmente marcadas abrangem todo este intervalo temporal.

Para o processo de clustering foram usados os dados de referência da CMM para a avaliação, e como tal, limitou-se a série apenas ao ano de interesse, de 2021.

Como previamente referido, para os dados de Mação, apenas foram considerados os

valores de referência que tinham um estado de execução claro, indicando, ou que não se encontrava intervencionada, ou continha a data limite para efetuar a intervenção. Após segmentar as faixas de referência originais de acordo com os critérios de ocupação de solo, distância à periferia e área, obtiveram-se a seguinte distribuição de fragmentos:

- Série intervencionada - **82 fragmentos**
- Série não-intervencionada - **703 fragmentos**

Relativamente às datas limite de intervenção, estes valores distribuem-se da seguinte forma sobre os fragmentos em análise, exemplificado na tabela 8.1.

Tabela 8.1: Distribuição das datas limite de execução pelos fragmentos gerados

Data Limite de Execução	**Nº de Fragmentos**
2021-05-02	7
2021-05-30	151
2021-07-05	104
2021-08-02	71
2021-09-01	97
2021-10-13	170
2021-10-27	103

De uma forma geral, estas séries perfazem **261.720** pontos temporais, para todos os fragmentos em análise e todo o intervalo temporal. Considerando apenas o ano de 2021, que será o intervalo usado para o *clustering*, temos **54.908** pontos.

8.2.1 Clustering

Para o processo de clustering, cada índice espectral vai ser testado, tanto o valor original, como os atributos derivados. Serão testados **12** séries, resultante da interseção dos índices *[NDVI, NDI45, IRECI, NBR]* e os valores *[valor, derivada, derivada acumulada]*. Além desta combinação, cada uma destas 12 séries será testada com todas as parametrizações de cada algoritmo, perfazendo **72 experiências** para o algoritmo KShape e **216 experiências** para o algoritmo K-Means.

8.2.2 Classificação Automática

Para o processo de classificação automática foram criados 4 conjuntos de dados para serem avaliados, de forma a entender que conjunto de características nos permite melhor determinar o estado de intervenção de uma faixa num dado instante.

Nos exemplos seguintes os termos **"value"** irá referir-se ao valor original de um índice, **"delta"** à sua derivada, **"roll"** à derivada acumulada do mês anterior e **"rise"** à subida do valor 1 mês após o instante.

Além da informação ao nível do instante temporal, são consideradas 3 features ao nível da série, o **iqr**, o **smooth_diff** e a **periodicidade**.

Os conjuntos que serão alvo de treino são os seguintes:

- TODOS_ATRIBUTOS - [iqr,smooth_diff,periodicidade] + [value,delta,roll,rise] para cada índice
- NOVOS_ATRIBUTOS - [iqr,smooth_diff,periodicidade] + [roll,rise] para cada índice
- APENAS_NDVI - [iqr,smooth_diff,periodicidade] + [value,delta,roll,rise] para o NDVI
- APENAS_INDEX - [value,delta,roll,rise] para cada índice

É preciso referir que, após o processo de seleção prévia de pontos a usar nos modelos de aprendizagem, a proporção entre pontos negativos e positivos é de aproximadamente 100 vezes, o que coloca um desequilíbrio e uma carga substancial nos modelos de aprendizagem e na máquina que os treina. Por essa razão decidiu-se limitar os pontos negativos a um número máximo de 20 vezes o número de pontos positivos.

Os dados usados para treinar os modelos de classificação automática foram gerados manualmente através de uma ferramente criada para o propósito, apresentado na secção 6.5, tendo sido marcadas 300 séries para o efeito.

8.3 Geração automática de FGCI

Como previamente mencionado o objetivo desta auto-geração de faixas é testar a viabilidade de identificar automaticamente faixas de proteção que estão em falta nos PMDFCI oficiais. Para tal, foi fornecido uma imagem raster que continha o resultado do processo de identificação automática de áreas artificiais (nas quais se incluem habitações e localidades). Este raster engloba a região da Guarda e ocupa uma área de 1811km^2.

Começando por entender os dados, em cada ficheiro raster um píxel pode ter um de cinco valores, que correspondem a diferentes classes, **alta intensidade urbana**, **baixa intensidade urbana (meios rurais)**, **estradas, minas, entre outros**, **zonas não-artificiais** e **zonas de água**. As classes que nos interessam para este estudo são as duas primeiras, e como tal, foram isolados os píxeis que pertencem a essas classes.

Este raster com as classes relevante seleccionadas é posteriormente poligonizado para facilitar os passos seguintes.

É necessário incluir outros dados externos, nos quais se incluem as definições das estradas do OSM e as FGCI oficiais das áreas de estudo. A informação geográfica do OSM foi extraída com o apoio da ferramenta *QuickOSM* que, após seleccionar a zona de interesse e as classes de objetos que se pretendem extrair (neste caso relativas a estradas e delimitações de localidades), nos fornece um ficheiro geográfico com todas as estradas e localidades mapeadas naquelas zonas. Relativamente às faixas oficiais foram incluídas as faixas do distrito da Guarda.

A informação que nos é fornecida pelo OSM das estradas corresponde apenas a vetores das mesmas, sem a dimensão de largura das estradas, sendo que o propósito desta informação é permitir ignorar os elementos de classificação que correspondem a estradas é necessário gerar novas geometrias com a largura apropriada. Ao comparar as estradas marcadas no OSM com as suas larguras, recorrendo às ortofotos, determinou-se que uma largura de 40 metros englobaria a vasta marioria das estradas. Foi portanto aplicado um buffer de 20 metros a estas geometrias.

Os polígonos resultantes da classificação automática são então recortados pelo polígono gerado da rede viária, são depois gerados buffers de 50 metros que rodeiam todos os fragmentos identificados pelo modelo de aprendizagem automática e aplicada a classe de "casa" a todos os polígonos de forma temporária.

São agora adicionados aos polígonos 3 novas características, o número de polígonos com qual eles se intersetam (após a aplicação do buffer), se estão localizados dentro de uma localidade (de acordo com o OSM), e a área dos mesmos.

Através das faixas dos PMDFCI dos municípios foi possível delimitar a área de polígonos em que, aproximadamente, se pode considerar que uma faixa corresponde a um aglomerado, que corresponde a aproximadamente 75km^2.

Os buffers que se intersetam com outros polígonos, que se encontram dentro de uma localidade (de acordo com o OSM) ou que têm uma área superior a 75km^2 são classificados como "aglomerados". De seguida os polígonos são unidos com aqueles com que se intersetam, e aqueles que estão classificados como "aglomerados" é-lhes imposto um buffer adicional de 50m, para perfazer os 100m previstos pela legislação para aglomerados populacionais. A estes polígonos é seleccionada apenas os 50 ou 100 metros exteriores, que corresponderiam às faixas. São também, de novo, excluídas as áreas que intersetam estradas.

É verificada para cada faixa a proporção da mesma que interseta as estradas, sendo eliminadas as faixas cuja interseção excede os 70%. A razão pela definição deste valor relativamente elevado é que existem várias habitações isoladas que se encontram junto a estradas, e como tal, têm grande parte da área das suas faixas intersetadas pelas estradas.

Depois de ter as faixas geradas é necessário avaliar a sua semelhança com as faixas oficiais, e mostrar se este processo cumpriu os objetivos de identificar zonas onde as faixas não estavam adequadamente identificadas no PMDFCI, especialmente no caso das habitações isoladas.

Para avaliar a sobreposição das faixas originais com as geradas, foi calculada a sobreposição de área entre a faixa gerada e a faixa oficial. Ao dividir a área de interseção pela área da faixa oficial obtemos uma percentagem de precisão relativamente às faixas oficiais.

Segmentando esta precisão em intervalos de 20% obtemos os seguintes resultados:

	Precisão relativamente a faixas oficiais				
	0-20%	20%-40%	40%-60%	60-80%	80%-100%
Todas as faixas	1042	97	106	106	595
Apenas Habitações	769	48	79	78	574
Apenas Aglomerados Pop.	273	49	27	28	21

O detalhe que salta à vista é a diferença de dispersão de valores entre as faixas de habitações e de aglomerados. Isto ocorre porque as faixas de aglomerados têm geometrias mais complexas, e como tal a sobreposição com as faixas geradas será sempre menor, ainda que se localizem aproximadamente na mesma zona. As faixas de habitações são mais pequenas, e simples, e como tal uma correta identificação normalmente corresponde a uma sobreposição de pelo menos 60%. As 1042 faixas com menos de 20% de sobreposição são todas potenciais faixas por identificar.

Para entender se o objetivo de detetar faixas em falta nos PMDFCI foi feita uma análise sobre as faixas habitacionais com menos de 20% de sobreposição. Foi feita uma amostragem aleatória de 60 faixas e os resultados foram os seguintes:

- Identificação correta de faixa em falta - 31 (51.7%)
- Identificação incorreta (vegetação) - 9 (15.0%)
- Identificação incorreta (outros) - 9 (15.0%)
- Identificação incorreta (estradas) - 6 (10.0%)
- Caso especial (já incorporado na faixa de uma localidade) - 1 (1.7%)
- Caso especial (parque eólico não mapeado no PMDFCI) - 4 (6.7%)

Os elementos identificados como "outros"na lista acima correspondem a terrenos áridos, rochosos ou minas.

No processo de análise desta amostra de resultados houve o caso inesperado da identificação de parques eólicos, que não estava contemplada nos objetivos originais. Os parque eólicos devem, de acordo com o guia dos PMDFCI, ter uma faixa protetiva em seu redor, incorporando uma rede de defesa de Mosaicos de Parcelas de Gestão de Combustível (MPGC). Após comparar os parques identificados com os MPGC do conselho da Guarda verificou-se que nenhum destes 4 parques estava contemplado no plano de defesa.

8.3.1 Conclusão

As faixas automáticas foram comparadas com as FGCI oficiais, sendo calculada a semelhança entre as mesmas. Existem muitas faixas com uma semelhança muito elevada, valores entre os 80% e 100%, particularmente nas faixas ao redor das habitações, o que demonstra a qualidade dos dados de classificação automática do terreno e da metodologia usada para gerar as faixas.

(a) Identificação errada, correspondente a uma zona de vegetação.

(b) Identificação correta, corresponde à Quinta do Ribeiro, um espaço com vários edifícios usado para eventos.

Figura 8.1: Exemplos de resultados do processo de autogeração de faixas de gestão de combustível. as geometrias a vermelho representam as faixas geradas automaticamente e a imagem de fundo corresponde às ortofotos de 2018

Pretendeu-se também avaliar a capacidade deste método para detetar faixas em falta nos planos de defesa. Ao fazer uma amostragem das faixas geradas e comparando-as com as oficiais notou-se que, nos casos onde não existia uma correspondência significativa entre as duas (semelhança entre 0 a 20%), mais de metade das faixas geradas estavam corretas e efetivamente identificam habitações que não estavam protegidas pelo PMDFCI da região.

Esta metodologia mostra-se promissora para a identificação de faixas em falta nos PMDFCI, e, no geral, para a criação de FGCI automaticamente, especialmente nos casos onde a rede está muito desatualizada ou mal-definida.

8.4 Clustering de Séries Temporais

Tendo sido seleccionados 2 principais algoritmos para a tarefa de clustering, foram efetuadas experiências com intuito de avaliar se estes métodos se adequam para separar as séries intervencionadas de não intervencionadas. A análise destes processos avaliará a forma como as duas classes de intervenção se distribuiram entre os clusters.

Os diferentes parâmetros dos métodos serão comparados, de forma a determinar que parametrização obtém os melhores resultados, sendo que a métrica de avaliação escolhida foi a homogeneidade. Além dos parâmetros dos algoritmos de clustering, também serão variados os dados usados para o processo de clustering, dentro dos índices espectrais escolhidos (**NDVI**, **NDI45**, **IRECI** e **NBR**), e características derivadas dos mesmos, neste caso o declive e o declive acumulado.

Usando a métrica de homogeneidade pode-se concluir que o algoritmo K-Means produziu os melhores resultados, ainda que não estejam muito longe daqueles obtidos pelo KShape, como comprovado pela figura 8.2.

Para o K-Means o melhor resultado foi a obtenção de um *score* de homogeneidade

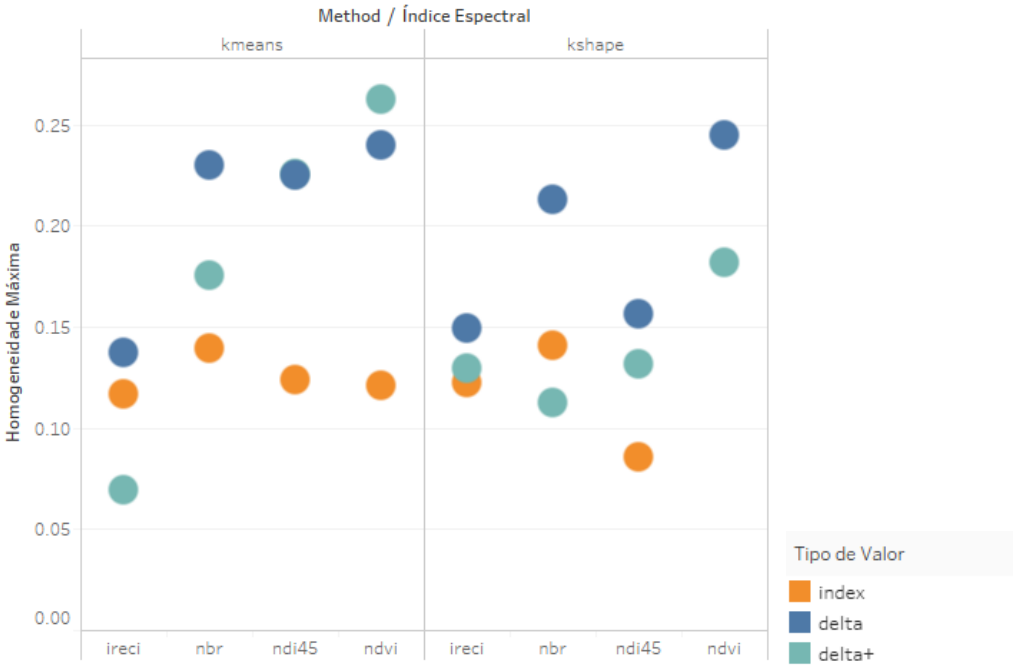


Figura 8.2: Melhores resultados dos algoritmos K-Means e KShape

de 0.263, correspondente ao uso do **declive acumulado** no índice **NDVI**, sendo que as parametrizações do modelo indicam o uso de **DTW** como métrica de distância e definindo **12 clusters**, como demonstrado na figura 8.3.

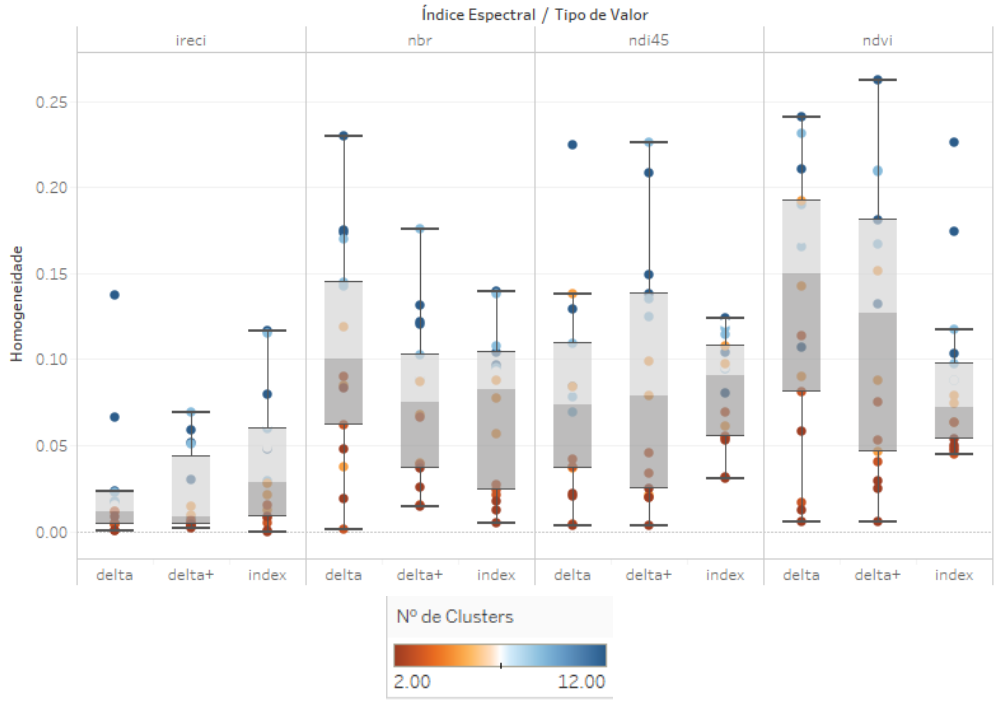


Figura 8.3: Distribuição da homogeneidade relativamente ao índice, tipo de valor, e número de clusters (mapeado nas cores). 'index' corresponde ao valor do índice, 'delta' à derivada e 'delta+' à derivada acumulada

Na figura 8.3 podemos tirar várias conclusões relativamente aos dados que foram usados no processo de *clustering*, nomeadamente que o NDVI apresenta os melhores resultados, enquanto o IRECI consistentemente tem valores muito baixo. Que a feature de declive acumulado tem melhores resultados nos índices NDVI e NDI45, que são índices muito semelhante. A combinação de índice e tipo de valor é relevante, sendo que por exemplo o NBR teve melhores resultados usando valores de declive.

Podemos também analisar os resultados relativamente à métrica de distância utilizada, correlacionando-a com os índices e o tipo de valor (mapeado na cor) e o número de clusters (mapeado no tamanho), como mostra a figura 8.4.

Figura 8.4: Distribuição da homogeneidade relativamente ao índice, tipo de valor, número de cluster e métrica de distância. A cor está mapeada ao tipo de valor e o tamanho ao número de clusters

Como previamente mencionado, além dos *scores* são também extraídos as percentagens de cada classe para cada cluster gerado. É importante contextualizar os resultados ao analisar a distribuição das classes pelos clusters, o que foi feito para o clustering com melhor avaliação. No entanto, o facto de 89% das séries ser intervencionada dificulta a fácil compreensão destes valores, logo, foi feita uma transformação sobre os dados, para

que a distribuição seja equivalente àquela que se encontraria numa situação onde exatamente metade das séries é intervencionada, cujo resultado se pode visualizar na figura 8.5.

Neste caso, um processo de clustering totalmente aleatório resultaria num valor que tenderia para 0.5 em todos os clusters.

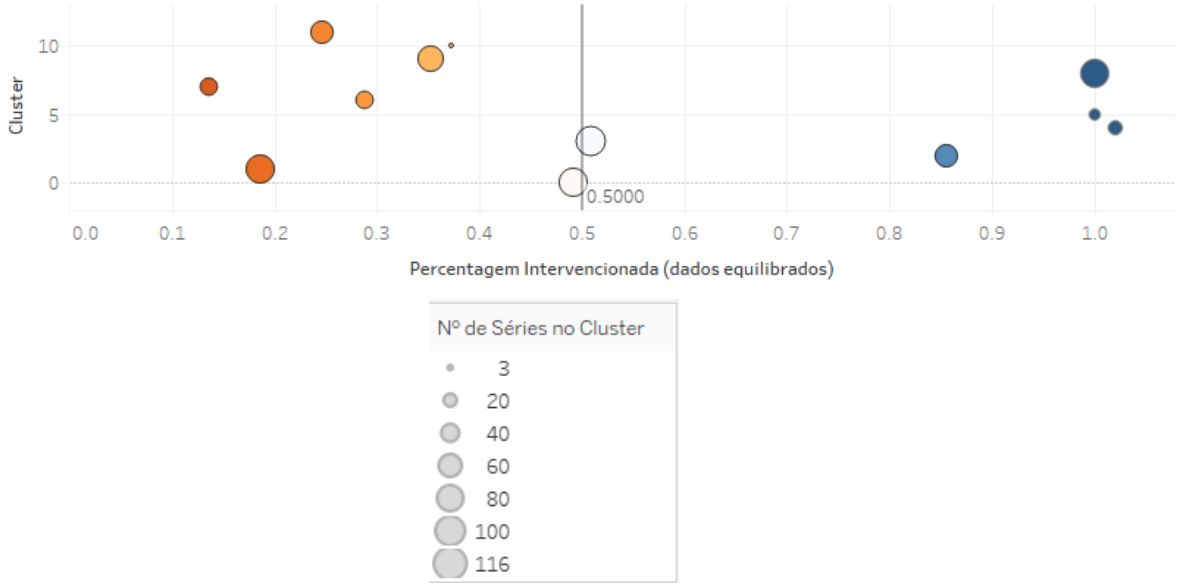


Figura 8.5: Distribuição das classes por clusters num cenário de classes equilibradas. A cor está mapeada à percentagem de séries intervencionadas, o tamanho representa o número de séries.

O que se pode concluir com esta figura é que existem clusters de dimensões significativas, neste caso dois com valores à volta de 0.5, que não discernem nenhuma das classes. Pelo outro lado temos vários clusters que são exclusivamente, ou quase exclusivamente, compostos por séries intervencionadas, sendo que os clusters com uma maioria não-intervencionado têm uma fração significante de séries de corte. Podemos olhar para algumas séries temporais de clusters para melhor entender os resultados, como a figura 8.6 e a figura 8.7

A partir das figuras anteriores podemos ver que, por exemplo, o cluster que é representado na sua maioria por faixas não-intervencionadas, tem algumas faixas intervencionadas, no entanto, grande parte destas tem um comportamento ao longo do tempo que não parece conter nenhum instante de intervenção. Pode acontecer que, para estes casos, os dados de referência estão errados, ou que o fragmento em análise é uma pequena área que não necessitou de corte, ou até que o índice que estamos a visualizar não representa bem o tipo de corte que foi feito.

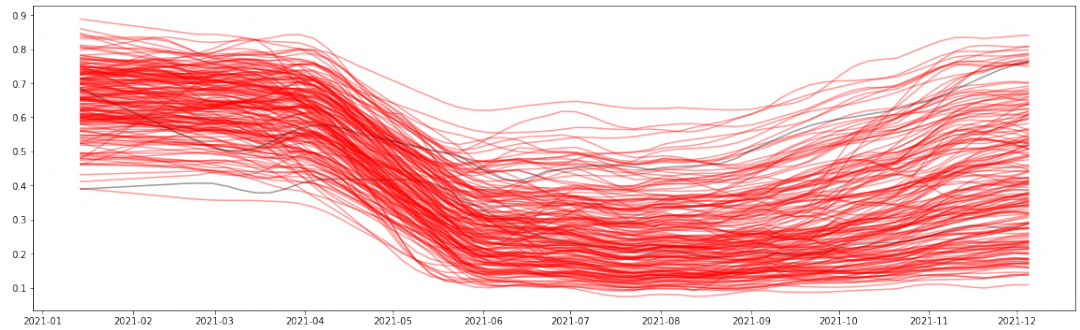


Figura 8.6: Cluster visualizado sobre o índice NDVI, maioritariamente de faixas intervencionadas, marcadas a vermelho (182 séries). Faixas sem intervenção estão marcadas a preto (2 séries).

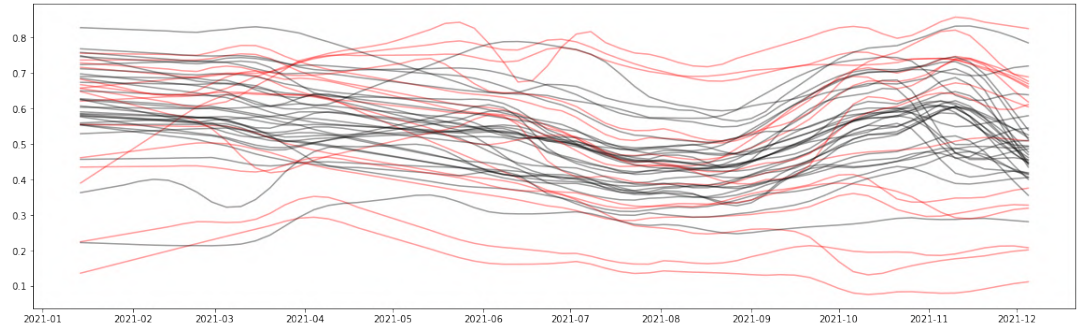


Figura 8.7: Cluster visualizado sobre o índice NDVI, maioritariamente de faixas não-intervencionadas, marcadas a preto (31 séries). Faixas com intervenção estão marcadas a vermelho (16 séries).

8.5 Aprendizagem sobre Séries Temporais

Com os algoritmos de classificação já referidos foram testadas uma combinação de parâmetros e de diferentes conjuntos de inputs para determinar o modelo que melhor distingue um instante de intervenção e um de não-intervenção.

A análise dos diferentes conjuntos de dados irá permitir-nos entender que características são mais importantes para uma boa avaliação de um instante de intervenção, além de validar, ou não, os critérios previamente definidos para a marcação manual dos pontos de corte.

Após testar as diferentes parametrizações e seleccionar o melhor modelo para cada método, podemos avaliar o F1-score de cada modelo, e verificar como este valor varia dependendo do conjunto de dados que é utilizado no treino, como presente na figura8.8.

Em relação ao conjunto de dados é possível verificar que a utilização do conjunto **NOVOS_ATRIBUTOS**, que inclui apenas os valores derivados na janela de um mês, e os atributos ao nível da série, resulta na melhor classificação no melhor modelo, que neste caso é o método **Random Forest**. Para este conjunto de dados o F1-Score varia entre **0.68 e 0.81**. Os piores resultados concentraram-se nos testes feitos com o algoritmo **SVM**, com um F1-score entre **0.65** e **0.69**, sendo a pior classificação com a utilização do conjunto de

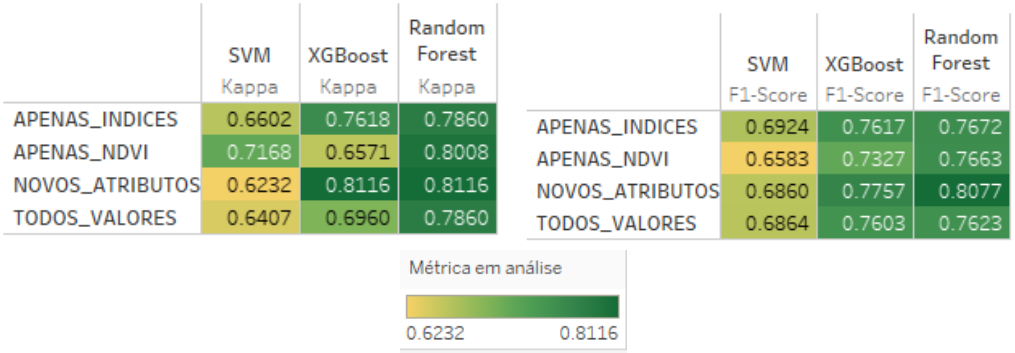


Figura 8.8: Valores de F1-Score e Kappa por método usado e conjuntos de dados

dados **APENAS_NDVI**.

As conclusões são relativamente semelhantes quando usamos o kappa como métrica principal para avaliar a performance do modelo, continuando a favorecer os resultados do Random Forest e do XGBoost, obtendo um valor máximo de **0.81** tanto com o XGBoost como na RF quando se utiliza o conjunto de dados NOVOS_ATRIBUTOS.

Analisando com maior detalhe o melhor modelo, o uso de **Random Forests** com o conjunto de dados **NOVOS_ATRIBUTOS**, e comparando com as restantes experimentações, verificamos que por detrás do valor de **0.81** de F1-Score temos uma distribuição diferente entre a proporção de falsos negativos e falsos positivos.

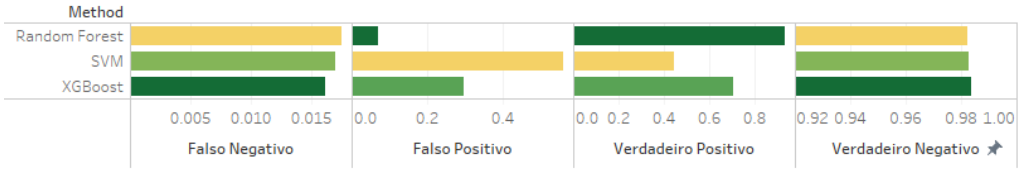


Figura 8.9: Valores da Matriz de Confusão por cada método usado

Comparando, por exemplo, as Random Forests e o XGBoost verificamos que o seu desempenho relativamente aos Falsos Negativos (FN) difere, sendo que o XGBoost tem uma melhor performance neste quadrante. As diferenças entre os valores de FN podem parecer pequenas, mas têm de ser contextualizada com o facto dos valores negativos na fase de treino serem 20 vezes superiores, ou seja, a deteção adicional de alguns falsos negativos tem um efeito pequeno no valor registado. Neste caso concreto, verificamos que, apesar de ter um nível mais alto de falsos negativos, a performance das Random Forests a detetar valores positivos é muito superior à dos restantes métodos, e aparenta justificar a eleição deste método como o melhor daqueles testados.

Como foi previamente mencionado, o processo de classificação não serve apenas para criar um modelo que detete corretamente momentos de intervenção, serve também para validar a marcação manual, e alertar para pontos que foram mal marcados. Tomando como exemplo o fragmento *0_859*, foi uma série temporal que foi marcada como sendo não-intervencionada, pois as descidas do nível de vegetação deviam-se à variação cíclica

que é normal desta série temporal. No entanto o classificador identificou um momento de intervenção, como demonstrado na figura 8.10 a vermelho.

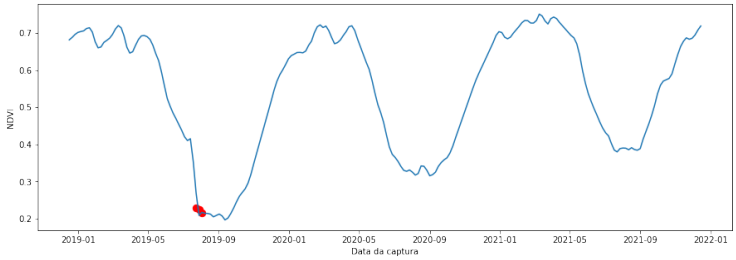


Figura 8.10: Índice NDVI ao longo do tempo para o fragmento 0_859 com uma intervenção identificada pelo classificador Random Forests a vermelho

Para a figura 8.10 podemos concluir que apesar da primeira impressão no processo de marcação de séries ter sido que a variação se enquadrava dentro de um ciclo normal, o classificador notou que em um dos anos o decréscimo foi substancialmente mais acentuado, e descendo para valores menores do que a restante série, o que é um forte indicador de uma intervenção sobre a faixa. Ou seja, a marcação inicial aparenta estar errada, e o classificador demonstrou que 'interiorizou' os conceitos necessários para a identificar intervenções ainda que os dados de referência sejam imperfeitos.

Outros exemplos mais concretos permitem-nos demonstrar o tipo de falsos negativos e falsos positivos que são habituais, como demonstrado nas figuras 8.11 e 8.12. As linhas verdes identificam intervenções identificadas pelos dados de referência manuais, as linhas vermelhas indicam intervenções identificadas apenas pelo modelo, e a castanho estão identificadas pelo modelo e pelos dados de referência manuais.

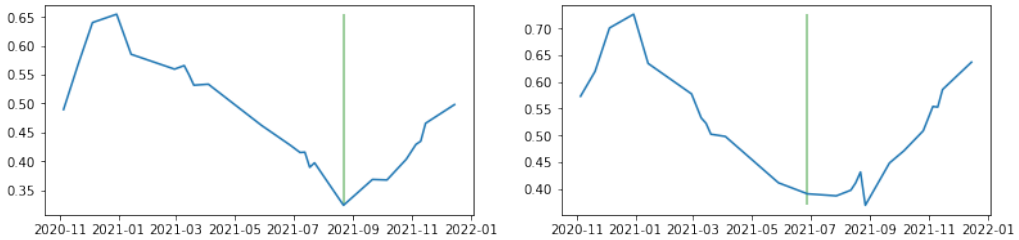


Figura 8.11: Exemplos de falsos negativos do modelo de classificação

Figura 8.12: Exemplos de falsos positivos do modelo de classificação

Nas figuras 8.11, os falsos negativos aparentam ser casos onde a descida marcada como intervenção se prolongou durante vários meses, sendo que na primeira figura existe uma descida mais acentuada mas em que a diferença de NDVI é mínima, de cerca de 0.05, para uma faixa que tem uma variação relativamente grande. No segundo exemplo de falsos negativos. Nestas situações crê-se que o modelo fez a decisão correta, sendo que os dados marcados manualmente parecem estar incorretos, apontando descidas demasiado prolongadas e pequenas diferenças de vegetação como intervenção.

Nas figuras 8.12, os falsos positivos aparentam corresponder a instantes com uma descida súbita de vegetação ao longo de um prazo relativamente curto. Em ambas as situações se podem identificar várias marcações sucessivas de intervenções por parte do modelo, isto deve-se ao facto dos instantes temporais que rodeiam o fim da intervenção terem características semelhantes, isto é, apresentarem um declive acumulado semelhante entre eles. O primeiro caso, à esquerda, corresponde a uma descida mais significativa, que se pode considerar um instante de intervenção que não foi identificado pela marcação manual das séries. A segunda figura, à direita, apresenta no entanto uma descida que se prolonga ao longo de um mês mas que corresponde a uma diferença de vegetação relativamente pequena, é debatível se esta situação se pode considerar um instante de intervenção ou não.

8.5.1 Análise sobre diferentes intervalos temporais

Como previamente mencionado, pretende-se entender o efeito que a escolha dos intervalos temporais usados para medir a recuperação da vegetação após um dado instante tem no desempenho do modelo. Isto é, pretende-se avaliar quão cedo se pode determinar que uma zona foi ou não intervencionada de acordo com o modelo criado, ao comparar, por exemplo, a avaliação de um ponto capturado no instante presente, a um capturado há 30 dias. Os resultados estão representados na figura 8.13, e com eles podemos concluir que a informação da subida de vegetação entre o instante e 20 dias tem um efeito negligenciável no desempenho do modelo, mas que o contexto a 30 dias tem um efeito positivo nos resultados do modelo.

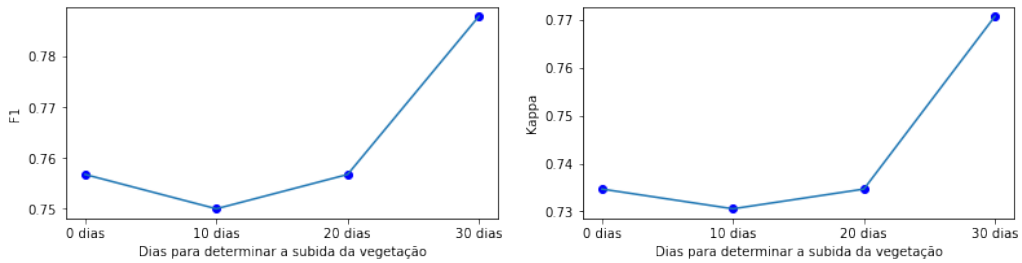


Figura 8.13: Valores de F1-Score e Kappa por intervalo de tempo considerado para a subida de vegetação

A análise do estado de intervenção de uma faixa no mesmo instante em que esta é capturada parece ser viável, no entanto, o contexto do estado da vegetação, a médio prazo

(um mês depois), é relevante para o modelo, confirmando um dos critérios usados para a marcação manual das séries temporais.

8.5.2 Comparação com dados de referência de Mação

Como previamente mencionado, os dados de referência para o concelho de Mação que foram fornecidos têm indicada uma data limite de intervenção, que não nos permite deduzir exatamente o instante da intervenção, mas que é suficiente para indicar se, num dado ano, ocorreram ou não ações de intervenção numa dada faixa.

A avaliação do modelo em Mação previamente discutida reflete a comparação dos resultados do modelo com os dados de referência manualmente gerados. Pretende-se portanto verificar se o modelo treinado nessas marcações manuais é capaz de identificar intervenções numa faixa inteira e se os resultados do modelo para a faixa são semelhantes àqueles que os dados de referência da Câmara de Mação nos indicam.

Esta comparação entre o modelo e os dados de referência é feita em duas etapas, na primeira comparamos o modelo com a referência para cada fragmento individualmente e, na segunda, agregamos a estimativa dos vários fragmentos por cada faixa, classificando-a como um todo. Após esse processo de agregação é possível comparar diretamente a classificação do modelo e os dados de referência para cada faixa.

Na figura 8.14 podemos ver exemplos onde o modelo concorda com os dados de referência fornecidos e na figura 8.15, exemplos onde o modelo não tem uma avaliação homogénea da faixa e é menos concordante com os dados de referência. Cada fragmento está assinalado a verde ou vermelho, indicando se foi intervencionado ou não, respetivamente, e ao redor da faixa está identificada, com as mesmas cores, o estado de execução da faixa de acordo com os dados de referência.

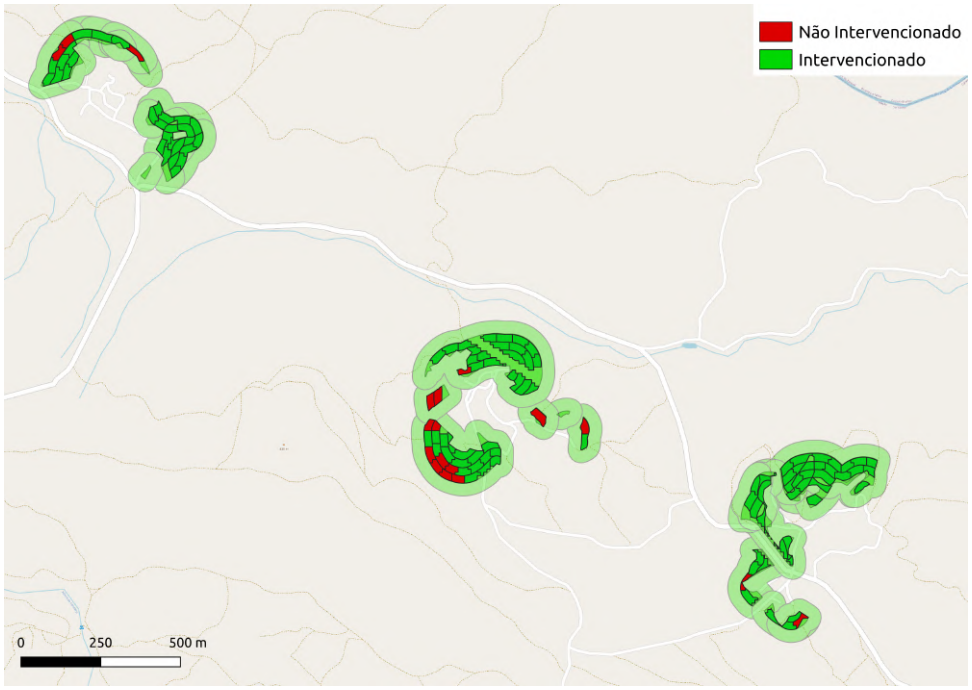


Figura 8.14: Exemplo para o qual o modelo tem resultados semelhantes à referência

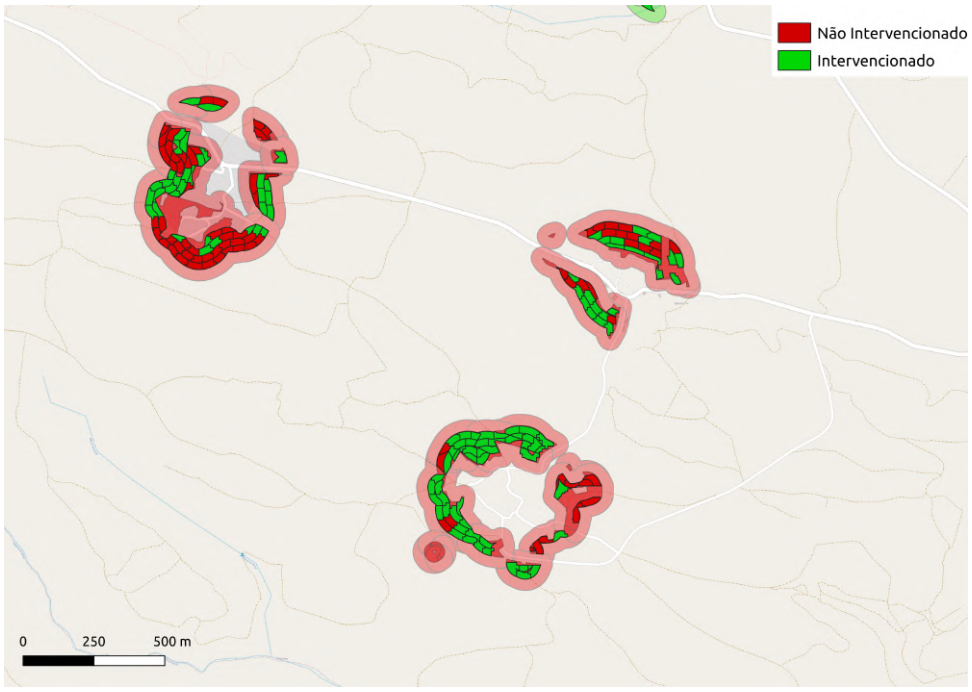


Figura 8.15: Exemplo para o qual o modelo tem resultados semelhantes à referência

Relativamente à análise de cada fragmento, os resultado obtidos estão indicados na tabela 8.2 e para a analise do modelo ao nível da faixa os resultados estão na tabela 8.3.

Como seria de esperar, a avaliação ao nível da faixa aparenta ter melhores resultados que a avaliação ao nível do fragmento. São comuns os casos em que dentro de cada faixa

Tabela 8.2: Avaliação em Mação ao nível do fragmento

F1-Score: 0.708	
Kappa: 0.173	
VP:77.3%	**FP**:53.4%
FN:22.7%	**VN**:46.6%

Tabela 8.3: Avaliação em Mação ao nível da faixa

F1-Score: 0.786	
Kappa: 0.318	
VP:86.7%	**FP**:40.0%
FN:13.3%	**VN**:60.0%

existem fragmentos que o modelo avalia de forma errada, é também importante ter em conta que para uma dada faixa que foi intervencionada, podem haver zonas onde na verdade não foi necessário fazer nenhuma ação, fazendo com que os fragmentos correspondentes a essas áreas se apresentem como errados.

8.5.3 Generalização do modelo em Santarém

Como previamente descrito, o modelo treinado com os dados manualmente gerados para as faixas de Mação será testado para a região de Santarém, sendo avaliado com base nos dados de referência fornecidos para esta zona. Esta análise tem duas facetas, permitindo avaliar se o modelo generaliza bem para outras regiões do país e para, ao usar um conjunto de dados totalmente diferente, averiguar se o modelo gerado não está sobreajustado.

A metodologia escolhida consiste em agregar o resultado da avaliação do modelo ao nível das faixas e comparar essa avaliação com os dados de referência. Um exemplo do resultado do modelo antes dessa agregação pode ser visto na figura 8.16, referente ao ano de 2021, para diversas faixas no interior da cidade de Santarém, sendo que a vermelho estão as zonas não-intervencionadas, e a verde as intervencionadas. Pode-se verificar que, na sua maioria, o resultado do modelo é semelhante àquele dos dados de referência, sendo que o estado de referência está marcado pela cor da área exterior às faixas.

No decorrer desta avaliação e da análise dos dados de referência e das séries temporais que representam as zonas de intervenção foi possível concluir que provavelmente existe um conjunto de erros na informação fornecida nos dados de referência. Esta informação serve para contextualizar os resultados abaixo descritos, e será melhor detalhada posteriormente.

Após seleccionar o modelo com melhor desempenho na anterior análise em Mação,

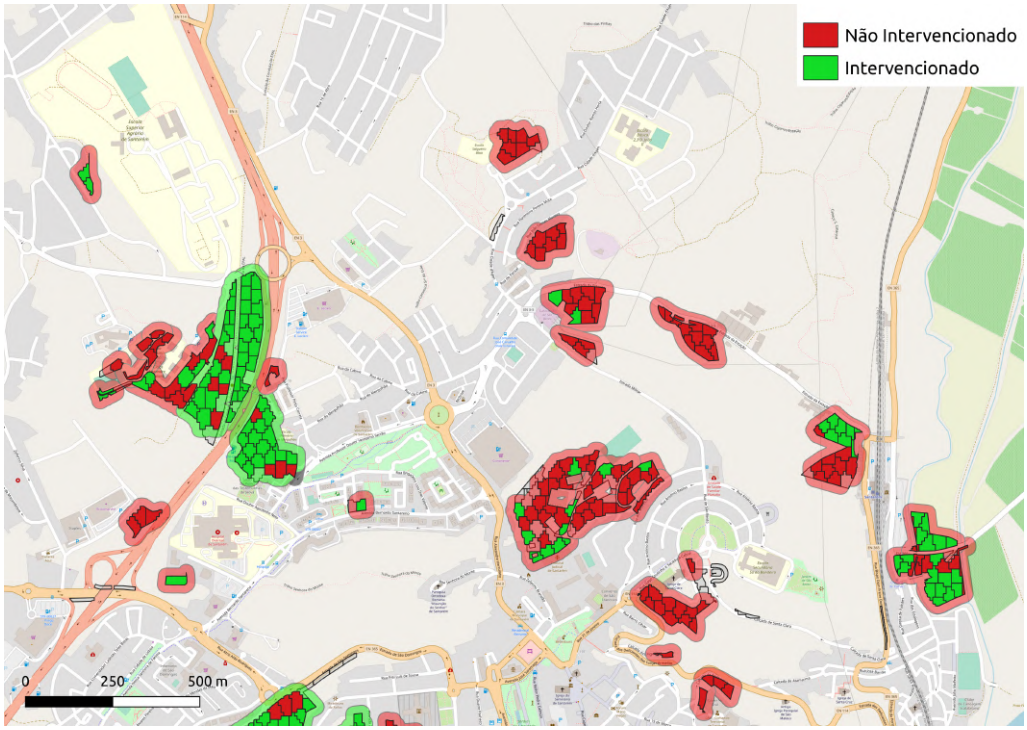


Figura 8.16: Resultado da classificação automática de faixas em Santarém

esse mesmo modelo foi aplicado às séries temporais de cada fragmento gerado em Santarém. Os resultados, no geral foram os seguintes, representados na tabela 8.4

Tabela 8.4: Avaliação em Santarém entre 2018 e 2021

F1-Score: 0.7680	
Kappa: 0.3843	
VP:67.0%	**FP**:33.0%
FN:20.1%	**VN**:79.9%

Separando os resultados por anos obtemos a seguinte informação, descrita nas tabelas das tabelas 8.5.

É possível ver que os resultados entre os vários anos mudam substancialmente, sendo que certos anos, como 2019 tem uma taxa de verdadeiros positivos de 82% enquanto 2020 tem uma de 49%, um resultado pior do que aquele obtido por um sorteio aleatório.

Para tentar entender esta descrepância analisaram-se as faixas e os resultados, recorrendo às séries temporais dos índices espectrais, tais como as imagens de satélite, animadas ao longo do tempo. Começou-se por analisar as faixas onde havia a maior discórdia entre os dados de referência e o modelo de aprendizagem, como quando a maioria dos fragmentos numa faixa são avaliados como intervencionados, mas a referência indica que não o foram.

Tornou-se claro que existiam vários casos onde os dados de referência aparentavam

Tabela 8.5: Avaliação em Santarém para cada ano entre 2018 e 2021

2018	
F1-Score: 0.7909	
Kappa: 0.4492	
VP:73.3%	**FP**:26.7%
FN:24.9%	**VN**:75.1%

2019	
F1-Score: 0.8915	
Kappa: 0.1410	
VP:82.2%	**FP**:17.8%
FN:46.2%	**VN**:53.8%

2020	
F1-Score: 0.6584	
Kappa: 0.1054	
VP:49.01%	**FP**:50.98%
FN:0.00%	**VN**:100.00%

2021	
F1-Score: 0.6506	
Kappa: 0.4606	
VP:64.0%	**FP**:36.0%
FN:18.25%	**VN**:81.75%

estar errados, indicando intervenções onde não se registava nenhuma mudança relevante e omitindo-as em casos onde havia uma alteração da vegetação muito substancial. As figuras 8.17 mostram alguns desses casos, para uma faixa em concreto. Nesta faixa os dados de referência indicavam uma intervenção apenas em 2019, enquanto o modelo de aprendizagem automática não identificava em 2019, mas sim em 2020 e 2021. Na figura 8.17 é aparente que houve uma grande ação de intervenção, ao longo de várias semanas, tendo sido identificada pelo modelo, mas não estando identificada nos dados de referência. Esta conclusão é corroborada pelas imagens no índice NDVI, da figura 8.18, sendo que nesta paleta de cores, as cores mais quentes correspondem a uma maior densidade de vegetação.

Figura 8.17: Imagens de satélite ao longo do tempo para uma faixa em 2020 (delineada a preto)

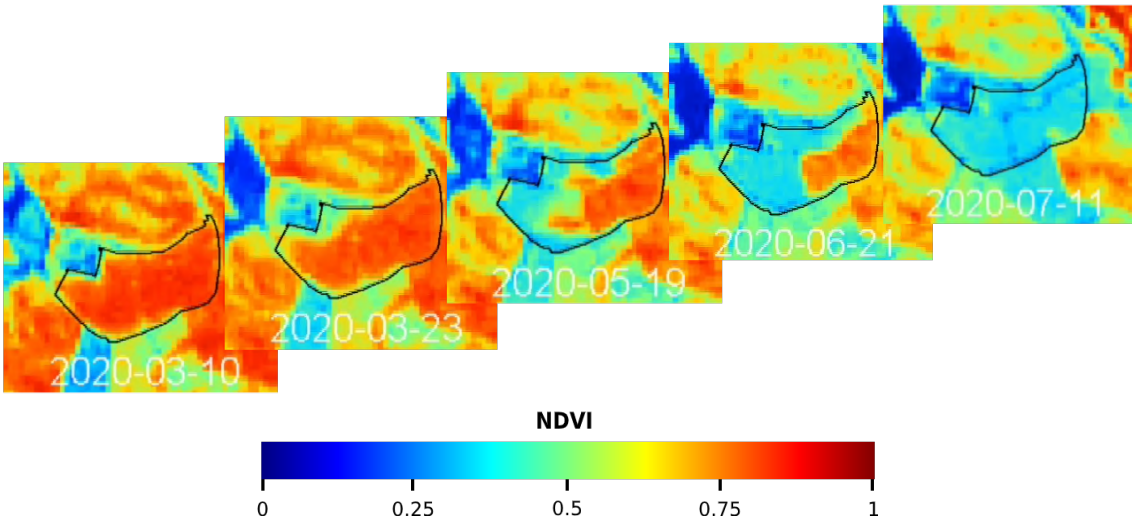


Figura 8.18: Imagens de satélite do índice NDVI ao longo do tempo para uma faixa em 2020 (delineada a preto)

8.6 Discussão dos resultados

Os resultados de um modo geral foram positivos, demonstrando a viabilidade de, com informação de deteção remota, usar métodos de aprendizagem automática para classificar o estado de intervenção de uma faixa de gestão de combustível. A informação extraída das capturas dos satélites Sentinel-2 mostrou-se suficiente para atingir uma precisão satisfatória, sendo que o uso de índices de vegetação como a principal característica para o processo de aprendizagem não só teve bons resultados como permitiu uma melhor compreensão dos comportamentos das zonas de vegetação em análise.

O modelo treinado nas faixas de Mação mostrou alguma promessa na capacidade de ser generalizado para outra zonas do país, no entanto a presença de alguns defeitos nos dados de referência dificultam essa análise.

Relativamente ao processo de clustering, a separabilidade das classes entre os clusters não atingiu o nível que se desejava, no entanto, permitiu-nos tirar conclusões sobre os índices de vegetação, e validar os novos atributos calculados para representar as alterações vegetativas nas faixas, como é o caso do declive acumulado.

Após a análise das FGCI definidas nos PMDFCI verificou-se que existia a possibilidade de haverem muitas habitações não protegidas pelas faixas de gestão de combustível. Recorrendo ao trabalho prévio na área da deteção de estruturas permanentes e com a incorporação de novos dados da OSM permitiu-se gerar faixas bastante precisas comparativamente às faixas oficiais, e detetar dezenas de casas sem a proteção adequada. Essa análise foi feita de uma amostragem de 60 casas identificadas, num único distrito, logo o potencial deste tipo de abordagem é muito significativo.

9

Conclusões e Trabalho Futuro

9.1 Conclusões

Nesta dissertação pretendeu-se abordar a problemática de verificar o estado das intervenções das FGCI, usando para o efeito mecanismos automáticos, com recurso a imagens de satélite e técnicas de deteção remota e aprendizagem automática.

A vasta extensão das faixas impossibilita uma fiscalização efetiva das faixas de forma manual. Um método que detete automaticamente o seu estado e indique as ações de intervenção prévias permitiria focar os esforços de fiscalização nas faixas e zonas nas quais existe uma maior suspeita de a vegetação não estar adequadamente mantida, podendo assim ter um impacto muito positivo na fiscalização e manutenção das FGCI.

Foram analisadas duas metodologias, uma não-supervisionada e outra supervisionada, ambas com recurso aos dados de deteção remota da missão Sentinel-2. A primeira foi testada entre dois diferentes métodos de clustering e com os diferentes valores em estudo, índices vegetativos e alguns atributos calculados ao longo tempo. Apesar do processo não permitir dividir de forma precisa todas as séries intervencionadas das não-intervencionadas, nota-se que existem, em determinadas experiências vários clusters que agrupam apenas as faixas que sofreram cortes, sendo que os resultados em termos da métrica de homogeneidade não excederam os 0.3, sendo que o melhor clustering obteve o valor de 0.263.

A segunda metodologia recorre a técnicas supervisionadas de classificação, tendo sido seleccionados 3 modelos, SVM, RF e XGBoost. Os algoritmos que obtiveram as melhores classificações foram o RF e o XGBoost, sendo que os melhores resultados foram obtidos com o conjunto de dados *NOVOS_ATRIBUTOS* (que inclui características ao nível da série e o declive acumulado e subida posterior para cada índice), o que valida as escolhas feitas na criação de novos atributos.

Os melhores resultados no geral foram obtidos com o modelo Random Forest, sendo que, com o uso do conjunto *NOVOS_ATRIBUTOS* obteve um valor kappa máximo de 0.811 e um F1-Score de 0.807.

No processo da dissertação foram explorados diferentes objetivos que não estavam inicialmente planeados, como a geração automática de FGCI, e uma plataforma que permite marcar, de forma eficiente, os instantes de interesse num conjunto de séries temporal.

Relativamente ao protótipo da geração automática de FGCI, os resultados comprovam que existe uma grande quantidade de casas cujas faixas não estão definidas e que esta abordagem tem resultados bons o suficiente para ajudar a detetar essas faixas em falta, sendo que a maioria dos casos onde se suspeitava haver uma faixa habitacional em falta estavam corretos.

Mostrou-se que o modelo treinado em Mação tem uma boa capacidade de generalização para outras regiões do país, neste caso para Santarém. É, no entanto, importante mencionar que os dados de referência fornecidos, em ambos os casos, têm erros e particularidades que obrigam a processamento adicional e impossibilitam a avaliação precisa dos modelos treinados.

9.2 Trabalho Futuro

Apesar dos resultados mostrarem alguma promessa existem vários aspetos que podem ser melhorados e novos objetivos que podem ser adicionados ao projeto.

Relativamente à informação de deteção remota usada, a incorporação dos dados Sentinel-1 seriam uma mais valia, já que em trabalho prévio se demonstrou que tem um efeito positivo na classificação de características de vegetação, além de que permite gerar séries mais atualizadas nos meses de inverno pois a radiação emitida penetra com facilidade as nuvens e as camadas da atmosfera. Existem outras missões que podem ser úteis para os objetivos desta dissertação, como a BIOMASS, com um lançamento previsto para 2023, que tem o objetivo de estimar a biomassa e altura das copas das florestas, com um tempo de revisita de 3 dias [18].

A região testada nesta dissertação foi muito limitada, sendo que para treinar os modelos apenas se consideraram as faixas habitacionais e de localidades, e apenas referentes ao concelho de Mação, sendo este modelo posteriormente avaliado em Mação e Santarém. A inclusão de novos conjuntos de dados de referência, de outras regiões do país permitiria fazer um estudo mais abrangente, além de que, a análise de diferentes tipos de faixa seria importante para validar a utilidade dos modelos desenvolvidos para as FGCI como um todo. Além de permitir fazer uma comparação direta com os métodos desenvolvidos em trabalho prévio.

O uso de redes neuronais profundas ficou por explorar, sendo interessante a sua testagem, tal como a incorporação de uma maior quantidade de dados de deteção remota, incluindo as bandas originais, um contexto temporal maior, e possivelmente valores que incorporem a dimensão espacial, como texturas.

Bibliografia

[1] *About OpenStreetMap - OpenStreetMap Wiki*. URL: https://wiki.openstreetmap.org/wiki/About_OpenStreetMap (acedido em 2022-02-23) (ver p. 12).

[2] N. Adaktylou, D. Stratoulias e R. Landenberger. «Wildfire Risk Assessment Based on Geospatial Open Data: Application on Chios, Greece». en. Em: *ISPRS International Journal of Geo-Information* 9.9 (2020-09). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 516. DOI: 10.3390/ijgi9090516. URL: https://www.mdpi.com/2220-9964/9/9/516 (acedido em 2021-01-22) (ver pp. 2, 29).

[3] R. F. S. Afonso. «Avaliação por Deteção Remota do Efeito das Operações de Limpeza nas Faixas de Gestão de Combustível de Incêndios». por. Em: (2019-11). Accepted: 2020-03-04T10:55:48Z. URL: https://run.unl.pt/handle/10362/93767 (acedido em 2021-02-24) (ver pp. 1, 22, 26, 28, 31, 33, 34).

[4] M. J. Antunes, D. Lopes e C. Oliveira. «Florestas e Legislação: Planos Municipais da Defesa da Floresta contra Incêndios». Em: Instituto Jurídico, 2020-10. ISBN: 978-989-8891-88-4 (ver pp. 2, 7, 29).

[5] V. Aubard et al. «Fully Automated Countrywide Monitoring of Fuel Break Maintenance Operations». en. Em: *Remote Sensing* 12.18 (2020-01). 4 citations (Crossref) [2022-03-31] Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, p. 2879. ISSN: 2072-4292. DOI: 10.3390/rs12182879. URL: https://www.mdpi.com/2072-4292/12/18/2879 (acedido em 2022-03-31) (ver pp. 22, 25, 26).

[6] R. Cao et al. «A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter». en. Em: *Remote Sensing of Environment* 217 (2018-11). 86 citations (Crossref) [2022-03-15], pp. 244–257. ISSN: 0034-4257. DOI: 10.1016/j.rse.2018.08.022. URL: https://www.sciencedirect.com/science/article/pii/S0034425718303985 (acedido em 2022-03-15) (ver p. 36).

[7] T. N. Carlson e D. A. Ripley. «On the relation between NDVI, fractional vegetation cover, and leaf area index». en. Em: *Remote Sensing of Environment* 62.3 (1997-12). 1360 citations (Crossref) [2021-02-06], pp. 241–252. ISSN: 0034-4257. DOI: 10.1016/S0034-4257(97)00104-1. URL: https://www.sciencedirect.com/science/article/pii/S0034425797001041 (acedido em 2021-02-06) (ver pp. 20, 22, 24, 30).

[8] M. Carriço. *491 casas e 48 empresas afetadas pelos incêndios. Ministro reúne com autarcas segunda-feira*. pt-PT. URL: https://observador.pt/2017/07/01/491-casas-e-48-empresas-afetados-pelos-incendios-ministro-reune-com-autarcas-segunda-feira/ (acedido em 2021-01-31) (ver p. 2).

[9] J. A. A. Castillo et al. «Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the Philippines using Sentinel imagery». en. Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 134 (2017-12). 73 citations (Crossref) [2021-02-15], pp. 70–85. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2017.10.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271617303362 (acedido em 2021-02-15) (ver pp. 22, 30).

[10] Centro de Estudos sobre Incêndios Florestais. *O Complexo de Incêndios de Pedrógão Grande e Concelhos*. 2017-10 (ver p. 2).

[11] T. Chen e C. Guestrin. «XGBoost: A Scalable Tree Boosting System». Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016-08). 7626 citations (Crossref) [2022-03-31] arXiv: 1603.02754, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: http://arxiv.org/abs/1603.02754 (acedido em 2022-03-31) (ver p. 34).

[12] I. da Conservação da Natureza e das Florestas. *6º Inventário Florestal Nacional*. 2018 (ver p. 1).

[13] Direção Geral do Território. *Fotografia aérea e imagens de satélite | DGT*. URL: https://www.dgterritorio.gov.pt/cartografia/fotografia-aerea (acedido em 2022-03-13) (ver p. 13).

[14] Direção Geral do Território. *Ortofotos digitais | DGT*. URL: https://www.dgterritorio.gov.pt/cartografia/cartografia-topografica/ortofotos/ortofotos-digitais (acedido em 2022-03-13) (ver p. 13).

[15] ESA. *Sentinel-2 MPC - L1C Data Quality Report*. 2021-02. URL: https://sentinel.esa.int/documents/247904/685211/Sentinel-2-L1C-Data-Quality-Report-February-2021.pdf/336cc953-28fc-8de5-bb0a-9ce5ed002df0?t=1612525904085 (acedido em 2021-02-23) (ver p. 21).

[16] ESA. *Sentinel-3 Experimental Campaign - Final Report*. 2011-01. URL: https://earth.esa.int/eogateway/documents/20142/37627/SEN3EXP-FinalReport-1.1.pdf (acedido em 2021-02-20) (ver p. 22).

[17] *Estimating grassland LAI using the Random Forests approach and Landsat imagery in the meadow steppe of Hulunber, China | Elsevier Enhanced Reader*. en. DOI: 10.1016/S2095-3119(15)61303-X. URL: https://reader.elsevier.com/reader/sd/pii/S2095953119156130X?token=71F550043DC3E960691531D7EA9B9D5424ADBF42AF1326DEEF2124F433A96DE8EB603554E7AEF6827E81F12B2DD566FF (acedido em 2021-01-05) (ver pp. 31, 33).

[18] European Space Agency. *Biomass*. en. URL: https://www.esa.int/Applications/Observing_the_Earth/FutureEO/Biomass (acedido em 2022-07-15) (ver p. 96).

[19] *Faixas de Gestão de Combustível nos aglomerados populacionais confinantes com espaços florestais*. pt-PT. URL: https://www.cm-olb.pt/pages/855?news_id=1207 (acedido em 2022-03-30) (ver p. 8).

[20] W. J. Frampton et al. «Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation». en. Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 82 (2013-08). 200 citations (Crossref) [2021-02-20], pp. 83–92. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2013.04.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S092427161300107X (acedido em 2021-02-20) (ver pp. 22, 30).

[21] G. Galidaki et al. «Vegetation biomass estimation with remote sensing: focus on forest and other wooded land over the Mediterranean ecosystem». en. Em: *International Journal of Remote Sensing* 38.7 (2017-04). 42 citations (Crossref) [2021-02-03], pp. 1940–1966. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431161.2016.1266113. URL: https://www.tandfonline.com/doi/full/10.1080/01431161.2016.1266113 (acedido em 2021-02-03) (ver p. 24).

[22] S. Georganos et al. «Less is more: optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application». en. Em: *GIScience & Remote Sensing* 55.2 (2018-03). 65 citations (Crossref) [2021-02-24], pp. 221–242. ISSN: 1548-1603, 1943-7226. DOI: 10.1080/15481603.2017.1408892. URL: https://www.tandfonline.com/doi/full/10.1080/15481603.2017.1408892 (acedido em 2021-02-24) (ver p. 34).

[23] D. Gianelle, L. Vescovo e F. Mason. «Estimation of grassland biophysical parameters using hyperspectral reflectance for fire risk map prediction». en. Em: *International Journal of Wildland Fire* 18.7 (2009-11). Publisher: CSIRO PUBLISHING 1 citations (Crossref) [2021-02-20], pp. 815–824. ISSN: 1448-5516. DOI: 10.1071/WF08005. URL: https://www.publish.csiro.au/wf/WF08005 (acedido em 2021-02-20) (ver p. 29).

[24] E. P. Glenn et al. «Relationship Between Remotely-sensed Vegetation Indices, Canopy Attributes and Plant Physiological Processes: What Vegetation Indices Can and Cannot Tell Us About the Landscape». eng. Em: *Sensors (Basel, Switzerland)* 8.4

(2008-03). 348 citations (Crossref) [2021-02-12], pp. 2136–2160. ISSN: 1424-8220. DOI: 10.3390/s8042136 (ver p. 30).

[25] Instituto da Conservação da Natureza e das Florestas. *Consulta PMDFCI*. URL: https://fogos.icnf.pt/infoPMDFCI/PMDFCI_PUBLICOlist.asp (acedido em 2022-03-31) (ver p. 11).

[26] Instituto da Conservação da Natureza e das Florestas. *Guia técnico do Plano Municipal de Defesa De Floresta Contra Incêndios (PMDFCI)*. URL: https://www.icnf.pt/api/file/doc/d6a7ab8782f71698 (acedido em 2021-01-25) (ver p. 23).

[27] L. Li et al. «Estimating Urban Vegetation Biomass from Sentinel-2A Image Data». en. Em: *Forests* 11.2 (2020-02). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute 2 citations (Crossref) [2021-02-20], p. 125. DOI: 10.3390/f11020125. URL: https://www.mdpi.com/1999-4907/11/2/125 (acedido em 2021-02-20) (ver p. 22).

[28] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (ver p. ii).

[29] P. M. Mather e M. Koch. *Computer processing of remotely-sensed images: an introduction*. 4th ed. OCLC: ocn614987867. Chichester, West Sussex, UK ; Hoboken, NJ: Wiley-Blackwell, 2011. ISBN: 978-0-470-74239-6 978-0-470-74238-9 (ver p. 32).

[30] *Medidas e acções a desenvolver no âmbito do Sistema Nacional de Defesa da Floresta contra Incêndios*. pt. URL: https://dre.pt/web/guest/legislacao-consolidada/-/lc/132506093/202010091049/73820205/diploma/indice (acedido em 2021-01-14) (ver pp. 1, 6).

[31] O. Molaudzi e S. Adelabu. «Review of the use of remote sensing for monitoring wildfire risk conditions to support fire risk assessment in protected areas». Em: *South African Journal of Geomatics* 7 (2019-02). 1 citations (Crossref) [2021-02-20], p. 222. DOI: 10.4314/sajg.v7i3.2 (ver p. 2).

[32] A. M. Neves. «Deteção Remota de Estruturas Artificiais Permanentes». por. Em: (2019-11). Accepted: 2020-03-26T10:50:36Z. URL: https://run.unl.pt/handle/10362/95073 (acedido em 2021-02-24) (ver pp. 1, 61).

[33] *Normalized Burn Ratio (NBR) | UN-SPIDER Knowledge Portal*. URL: https://un-spider.org/advisory-support/recommended-practices/recommended-practice-burn-severity/in-detail/normalized-burn-ratio (acedido em 2022-03-29) (ver p. 23).

[34] J. E. Pereira-Pires et al. «Fuel Break Vegetation Monitoring with Sentinel-2 NDVI Robust to Phenology and Environmental Conditions». Em: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 0 citations (Crossref) [2022-03-31] ISSN: 2153-7003. 2021-07, pp. 6264–6267. DOI: 10.1109/IGARSS47720.2021.9554943 (ver pp. 22, 24, 36).

[35] J. E. Pereira-Pires et al. «Semi-Automatic Methodology for Fire Break Maintenance Operations Detection with Sentinel-2 Imagery and Artificial Neural Network». en. Em: *Remote Sensing* 12.6 (2020-01). 12 citations (Crossref) [2022-03-31] Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 909. ISSN: 2072-4292. DOI: 10.3390/rs12060909. URL: https://www.mdpi.com/2072-4292/12/6/909 (acedido em 2022-03-31) (ver p. 25).

[36] P. Rajah et al. «The utility of Sentinel-2 Vegetation Indices (VIs) and Sentinel-1 Synthetic Aperture Radar (SAR) for invasive alien species detection and mapping». Em: *Nature Conservation* 35 (2019-06). 7 citations (Crossref) [2021-02-06], pp. 41–61. DOI: 10.3897/natureconservation.35.29588 (ver p. 20).

[37] E. K. Sahin. «Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest». en. Em: *SN Applied Sciences* 2.7 (2020-06). 5 citations (Crossref) [2021-02-24], p. 1308. ISSN: 2523-3971. DOI: 10.1007/s42452-020-3060-1. URL: https://doi.org/10.1007/s42452-020-3060-1 (acedido em 2021-02-24) (ver p. 34).

[38] *Sentinel-2 Cloud Masking with s2cloudless*. URL: https://developers.google.com/earth-engine/tutorials/community/sentinel-2-s2cloudless (acedido em 2022-03-31) (ver p. 49).

[39] S. M. Shupe e S. E. Marsh. «Cover- and density-based vegetation classifications of the Sonoran Desert using Landsat TM and ERS-1 SAR imagery». en. Em: *Remote Sensing of Environment* 93.1-2 (2004-10). 29 citations (Crossref) [2021-02-06], pp. 131–149. ISSN: 00344257. DOI: 10.1016/j.rse.2004.07.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0034425704002093 (acedido em 2021-02-06) (ver p. 20).

[40] D. M. Szpakowski e J. L. R. Jensen. «A Review of the Applications of Remote Sensing in Fire Ecology». en. Em: *Remote Sensing* 11.22 (2019-01). Number: 22 Publisher: Multidisciplinary Digital Publishing Institute 9 citations (Semantic Scholar/DOI) [2021-02-01], p. 2638. DOI: 10.3390/rs11222638. URL: https://www.mdpi.com/2072-4292/11/22/2638 (acedido em 2021-01-19) (ver p. 22).

[41] D.-G. do Território. *Cartografia de Uso e Ocupação do Solo (COS, CLC e Copernicus) | DGT*. URL: https://www.dgterritorio.gov.pt/cartografia/cartografia-tematica/COS-CLC-COPERNICUS?language=en (acedido em 2021-02-23) (ver p. 13).

[42] D.-G. do Território. *COS - Especificação 2018*. 2018. URL: https://www.dgterritorio.gov.pt/sites/default/files/documentos-publicos/2019-12-26-11-47-32-0__ET-COS-2018_v1.pdf (acedido em 2021-01-25) (ver p. 13).

[43] S. Vafaei et al. «Improving Accuracy Estimation of Forest Aboveground Biomass Based on Incorporation of ALOS-2 PALSAR-2 and Sentinel-2A Imagery and Machine Learning: A Case Study of the Hyrcanian Forest Area (Iran)». en. Em: *Remote Sensing* 10.2 (2018-02). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute 74 citations (Crossref) [2021-02-03], p. 172. DOI: 10.3390/rs10020172. URL: https://www.mdpi.com/2072-4292/10/2/172 (acedido em 2021-02-03) (ver p. 33).

[44] B. T. Wilson, J. F. Knight e R. E. McRoberts. «Harmonic regression of Landsat time series for modeling attributes from national forest inventory data». en. Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 137 (2018-03). 27 citations (Crossref) [2021-02-12], pp. 29–46. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2018.01.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271618300066 (acedido em 2021-02-12) (ver p. 33).

[45] J. Xue e B. Su. «Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications». en. Em: *Journal of Sensors* 2017 (2017). 286 citations (Crossref) [2021-02-02], pp. 1–17. ISSN: 1687-725X, 1687-7268. DOI: 10.1155/2017/1353691. URL: https://www.hindawi.com/journals/js/2017/1353691/ (acedido em 2021-02-02) (ver p. 20).

[46] Z. Yang et al. «Vegetation condition indices for crop vegetation condition monitoring». Em: *2011 IEEE International Geoscience and Remote Sensing Symposium*. ISSN: 2153-7003 11 citations (Crossref) [2021-02-02]. 2011-07, pp. 3534–3537. DOI: 10.1109/IGARSS.2011.6049984 (ver p. 24).

[47] Z. Zhu e C. E. Woodcock. «Continuous change detection and classification of land cover using all available Landsat data». en. Em: *Remote Sensing of Environment* 144 (2014-03). 531 citations (Crossref) [2021-02-22], pp. 152–171. ISSN: 0034-4257. DOI: 10.1016/j.rse.2014.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0034425714000248 (acedido em 2021-02-22) (ver p. 33).

I

Anexos

```
{
  "points": {
    "0_209": [],
    "1_692": [
      {
        "date": "2019-3-21", //data da suposta intervenção
        "measure": "ndvi", //índice usado para a marcação
        "value": 0.3928843029, //valor do índice naquele instante
        "certainty": "uncertain" //certeza desta marcação
      },
    ],
    "1_696": [
      {
        "date": "2021-1-14",
        "measure": "ireci",
        "value": 0.5140597661,
        "certainty": "certain"
      }
    ],
  },
  "actions": {
    "1_425": "noncut", //incerteza sobre a ausência de intervenção
    "1_692": "unsure", //incerteza sobre a série
    "1_696": "cut", //certeza sobre a presença de intervenção
  },
  "groups": [
    "0_209",
    "1_692",
    "1_696",
  ],
}
```

Figura I.1: Exemplo da estrutura geral da árvore JSON exportada pela plataforma de marcação de séries