NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA

EDUARDO FILIPE ALMEIDA FERNANDES

Licenciado em Ciências e Engenharia Informática

DETERMINAÇÃO DA IDADE DA FLORESTA NACIONAL USANDO SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
dezembro, 2021

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTAMENTO DE
INFORMÁTICA

DETERMINAÇÃO DA IDADE DA FLORESTA NACIONAL USANDO SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE

EDUARDO FILIPE ALMEIDA FERNANDES

Licenciado em Ciências e Engenharia Informática

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: Doutor Pedro Abílio Duarte Medeiros
Prof. Associado,
Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Orientador: Doutor Carlos Augusto Isaac Piló Viegas Damásio
Prof. Associado,
Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

Vogal: Doutora Célia Marina Pedroso Gouveia
Investigadora Auxiliar,
Nucleo de Observação da Terra, Instituto Português do Mar e Atmosfera, IP

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
dezembro, 2021

Determinação da Idade da Floresta Nacional Usando Séries Temporais de Imagens de Satélite

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeXe o modelo NOVAthesis (v6.9.4) [35].

Agradecimentos

Quero agradecer à Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa pelo seu papel crucial na minha formação académica. Aos meus orientadores, Professor Carlos Viegas Damásio e Professor João Moura Pires, pelo constante acompanhamento e pela ajuda na realização da presente dissertação. O trabalho realizado foi apoiado por NOVA LINCS (UIDB/04516/2020) com o apoio financeira da FCT.IP, e em colaboração com o Projecto Floresta Limpa (PCIF/MOG/0161/2019).

Queria agradecer ainda aos colegas do projeto MORENA, em especial à Marta Carlos pela amizade, momentos de desabafo e inúmeras pausas para café. Aos meus amigos da Confraria das Pizzas Bruno Calapez, Cláudio Pereira, David Silva, Francisco Guerreiro, João Antão, João Mota, Lucas Dias, Manuel Almeida e Rafael Conceição por terem feito parte deste capítulo da minha vida entre trabalho e vida boémia. Agradeço aos meus grandes amigos de sempre Ana Filipe e João Ramos. Um especial agradecimento ao meu estimado colega de trabalho e grande amigo Bruno Anjos, a quem admiro a ética de trabalho, pela dedicação aos projetos de grupo e por todos os momentos partilhados.

Finalmente, quero agradecer à minha família, aos meus pais e irmã pelo afeto e por me terem apoiado ao longo destes anos apesar dos desafios que a distância física colocou. Agradeço ainda aos meus senhorios Sr. Pimenta e Dona Isabel e restante família, pela forma calorosa como me acolheram durante a minha estadia. Agradeço à Exma. Dra. Mariana Caetano, por todos os momentos que partilhamos, pela motivação que me deu, pelas inúmeras correções ortográficas da presente tese e pela paciência que demonstrou para me aturar nos momentos mais complicados da realização da presente dissertação.

Resumo

A floresta é um recurso natural importante a nível ecológico e económico. Dado que a floresta é o principal uso de solo em área ocupada no território português, o seu desenvolvimento merece especial atenção, sendo necessárias ferramentas para a monitorizar. Uma caraterística importante da floresta é a sua idade pois carateriza o seu aproveitamento e capacidade de filtração de dióxido de carbono. Sendo um indicador da quantidade de biomassa presente, poderá fornecer informação adicional no combate a incêndios. Portugal possui fontes de informação florestal, como é o caso do Inventário Florestal Nacional (IFN), da Carta de Uso e Ocupação do Solo (COS) e da cartografia de áreas ardidas. Quer o IFN quer a COS fornecem informação sobre a distribuição geográfica da floresta e das espécies que a compõem. A atualização destas fontes de informação é irregular, tendo no passado demorado 3 a 12 anos entre edições consecutivas.

A determinação da idade da floresta foi conseguida pela deteção de momentos de distúrbio, analisando séries temporais provenientes de imagens de satélite. Este assunto não é trivial, e o estado da arte não oferece soluções concretas, pelo que foi necessário comparar algumas abordagens para selecionar a que melhor se adequava ao problema. No campo da análise de séries temporais de deteção remota destacaram-se dois algoritmos: o LandTrendr e o CCDC. Adicionalmente, no campo da análise de pontos de quebra em séries temporais genéricas, destacou-se o BOCPD. Foram usados dados das missões Landsat, que já contam com um historial de 4 décadas. Devido à escassez de dados sobre a idade da floresta, foi também necessária a criação de um novo conjunto de dados de referência para avaliar os algoritmos. Este conjunto de dados indica de forma exaustiva todos os distúrbios entre os anos de 1986 e 2019 para 664 pontos de floresta portuguesa. Comparando estes dados com as classificações obtidas pelos algoritmos já referidos, foi possível concluir que o BOCPD obtém os melhores resultados com um F1 de 0.717, False Negative Rate de 0.365 e False Discovery Rate de 0.175.

Palavras-chave: Deteção Remota, Idade da Floresta, Séries temporais, Deteção de alterações Florestais

Abstract

The forest is important at an ecological and economical level. Given that the forest is the main land use by area in Portuguese territory, its development deserves special attention, therefore tools are needed to monitor it. An important characteristic of the forest is its age as it characterizes its economical potential and capacity to filter carbon dioxide. As an indicator of the amount of biomass, it can also provide additional information for fire-fighters. Portugal has forest information sources, such as the Inventário Florestal Nacional (IFN), the Carta de Uso e Ocupação do Solo (COS) and the burned areas cartography. Both the IFN and the COS provide information on the geographic distribution of the forest and its composition. Updates on these information sources are irregular, having taken as little as 3 to as much as 12 years between consecutive editions.

Determining the age of the forest was achieved by detecting disturbance moments, while analysing time series from satellite images. This issue is not trivial, and the state of the art does not offer concrete solutions, so it was necessary to compare some approaches and to select the one that best suited the problem. In the field of remote sensing time series analysis, two algorithms stood out: LandTrendr and CCDC. Additionally, in the field of of breakpoint detection in generic time series, the BOCPD stood out. Data from Landsat missions, which already has a history of 4 decades, were used. Due to the scarcity of data on the age of the forest, it was also necessary to create a new reference dataset to evaluate the algorithms. This dataset exhaustively indicates all disturbances between 1986 and 2019 for 664 points of Portuguese forest. Comparing this data with the classifications obtained by the aforementioned algorithms, it was possible to conclude that the BOCPD obtains the best results with an F1 of 0.717, False Negative Rate of 0.365 and False Discovery Rate of 0.175.

Keywords: Remote Sensing, Forest Age, Time Series, Forest Change Detection

Índice

Índice de Figuras

Índice de Tabelas

Siglas

BOCPD Bayesian Online Change Point Detection ix, xi, xii, 4, 24–27, 29, 41, 43–45, 47, 55, 58, 72, 73, 78–84

CCDC *Continuous Change Detection and Classification* ix, xi, xii, 22, 23, 25, 29, 41, 43–45, 47, 55–57, 72, 75, 77, 84

COS Carta de Uso e Ocupação do Solo 2, 4, 6, 10–12

EWMACD Exponentially Weighted Moving Average Change Detection 25, 57

GEE Google Earth Engine 5, 54, 55, 58

ICNF Instituto da Conservação da Natureza e das Florestas 4, 6, 7, 9

IFN Inventário Florestal Nacional xii, 1, 2, 6, 8, 52

NBR Normalized Burn Ratio x, 16, 17, 36, 37, 40, 42, 43, 45, 78, 81, 84

NDVI Normalized Difference Vegetation Index x, 16, 18, 19, 36, 37, 42, 43, 45, 55–57, 84

1

Introdução

A floresta tem um papel determinante a nível ambiental e económico. É, por isso, essencial proteger este recurso natural para que o seu desenvolvimento seja sustentável. É necessário ter instrumentos capazes de nos informar do estado das florestas de modo a ser possível tomar medidas que visem a proteção e conservação deste bioma. No entanto as florestas são vastas, diversificadas e complexas, o que dificulta a recolha de informação. Consequentemente, dados sobre a sua composição, estrutura e historial nem sempre estão disponíveis com a devida celeridade. Surge por isso a necessidade de obter estes dados de uma forma automática para que estejam sempre o mais atualizados possível. Uma possível solução passa por utilizar deteção remota (imagens de satélite) para determinar informações sobre a floresta. Uma vez que temos um acesso contínuo a estes dados, é possível obter estatísticas atualizadas sobre a floresta. Há uma variedade de informações importantes para aferir sobre as florestas, mas este trabalho foca-se principalmente na idade. A idade da floresta tem interesse a vários níveis: a nível económico é um bom indicador da quantidade de biomassa no local e consequentemente o seu potencial aproveitamento [46]; a nível ecológico fornece informação sobre a capacidade de captação de dióxido de carbono [22]. Adicionalmente, também poderá servir como informação complementar sobre o risco de incêndio [37].

1.1 Contexto e Motivação

Portugal é um país que conta com uma vasta área florestal. Segundo o mais recente IFN [24], 36% do território português está ocupado por floresta, sendo por isso o principal uso de solo em Portugal. A floresta conta com uma diversidade de espécies, é possível observar a distribuição percentual das mesmas na Tabela 1.1. A principal espécie presente na floresta portuguesa é o eucalipto, seguida pelo sobreiro e pinheiro-bravo.

As florestas são ecossistemas que sustentam uma grande biodiversidade. Para além da importância ambiental, têm também um grande interesse económico dado que contribuem para a criação de emprego e de rendimentos para o país através das exportações.

Tabela 1.1: Ocupação do solo das florestas em Portugal continental no ano de 2015. Adaptado do relatório final do IFN de 2015[25]

Espécie	Cobertura Florestal (%)
Eucaliptos	26,2
Sobreiro	22,3
Pinheiro-bravo	22,1
Azinheira	10,8
Pinheiro-manso	6,0
Outras folhosas	5,9
Carvalhos	2,5
Outras resinosas	1,6
Castanheiro	1,5
Alfarrobeira	< 1,0
Acácias	< 1,0
Temporariamente desarborizada	< 1,0

Em 2017, a indústria responsável por explorar a floresta empregava cerca de 70 mil pessoas, e exportou 5927 milhões de euros, o que corresponde a 10% do total das exportações portuguesas [17].

Os fogos são uma enorme ameaça à floresta nacional. Todos os anos durante o verão ardem milhares de hectares de floresta, deixando marcas no território com repercussões que duram vários anos. Recentemente, tem havido uma maior preocupação por parte das autoridades competentes em proteger este recurso. Como tal, surge a necessidade de compreender os efeitos que as medidas tomadas têm no terreno, bem como a necessidade de possuir uma ideia geral do estado atual da floresta portuguesa.

1.2 O Problema

O principal problema tratado na presente dissertação passa por criar um cadastro nacional florestal utilizando dados de satélite. Isto é, conseguir determinar a idade da floresta para um qualquer ponto de floresta portuguesa.

Este tipo de informação é crucial para compreender como têm evoluído as florestas portuguesas, qual o impacto a curto-médio prazo de medidas administrativas tomadas, bem como informação adicional no combate aos incêndios.

O principal foco deste trabalho é a idade da floresta, visto que para Portugal já há acesso a informações sobre a localização e classificação das florestas através da Carta de Uso e Ocupação do Solo (COS) [15].

A procura de uma solução a este problema impôs alguns desafios: em primeiro lugar foi necessário processar elevadas quantidades de dados. Isto porque uma imagem de satélite pode ultrapassar 1 giga byte [18] e, adicionalmente, há um grande volume de imagens de satélite para processar dado que cada imagem apenas captura uma parte

do território português. Em segundo lugar, a qualidade dos dados de satélite também se revelou por vezes um desafio, dado que diversos fenómenos atmosféricos podem interferir com as medições - por exemplo, as nuvens, a sombra das mesmas e, no caso do Landsat 7, problemas com o sensor. Em terceiro lugar, a uma escassez de informação de qualidade relacionada com a idade das florestas em Portugal, que limitou a utilização de algumas abordagens que têm necessidade de usar muitos dados e obrigou a construção de um novo conjunto de dados de forma manual. Em quarto e último lugar, salienta-se o desafio de conciliar diferentes fontes de dados para criar uma visão mais completa da floresta em Portugal.

Durante o desenvolvimento da presente dissertação, havia informação disponível relativamente às florestas portuguesas [25, 36, 15]. Esta informação é recolhida de forma manual com uma periodicidade de vários anos. Automatizando este processo seria possível atualizar a informação mais rapidamente. De qualquer forma, a recolha manual mantém a sua importância como meio de avaliar a performance da recolha automática de informação. A solução do problema visa complementar a informação já existente.

1.3 Abordagem

Para resolver o problema apresentado o presente trabalho comparou quatro algoritmos de análise de séries temporais. Os algoritmos analisaram séries temporais resultantes de imagens de satélite dos últimos 30 anos provenientes das missões Landsat. A principal ideia consiste em analisar o desenvolvimento da floresta ao longo do tempo e detetar momentos de grande mudança (abate, incêndios, etc.) para determinar o momento em que uma nova floresta é gerada. A partir destas datas de nascimento o cálculo da idade da floresta é imediato.

Numa primeira fase foi necessário compreender e visualizar o efeito que momentos de distúrbio da floresta têm nas séries temporais, durante este processo foram criadas visualizações úteis para analisar séries temporais.

Posteriormente, numa segunda fase, foi necessário recolher informações disponíveis sobre a floresta portuguesa. Os dados recolhidos variam, desde mapas detalhados de áreas de floresta ardida até pontos de inventários florestais com informação sobre as espécies de árvores presentes no local. Foi necessário compreender estes *datasets* e formatá-los de modo a extrair o máximo de informação possível da sua utilização conjunta, isto é o principal foco passou por compatibilizar as fontes de informação para obter uma visão mais completa da floresta.

Relativamente aos algoritmos comparados, foi necessário fazer tomar decisões de implementação para os algoritmos que não estavam talhados para deteção remota. Para todos os algoritmos foram testadas várias parametrizações dos mesmos com o objetivo de perceber quais as melhores parametrizações e quais as diferenças que estas poderão fazer.

Para avaliar os algoritmos foi necessário construir um novo *dataset* específico para o propósito por não haver um conjunto de dados disponível capaz de detalhar a idade das

florestas. A criação deste *dataset* envolveu a seleção de pontos usando uma amostragem estratificada por zona climática e espécie de árvores. Posteriormente foram interpretadas as séries temporais, imagens de satélite bem como informações adicionais para de forma manual anotar os momentos de distúrbio da floresta.

Finalmente foi necessário comparar os vários algoritmos e respetivas parametrização. Para esse efeito foram criadas visualizações para melhor compreender os pontos fortes e pontos fracos de cada algoritmo, quais os parâmetros mais importantes para a tarefa em mãos e como se comparam os algoritmos entre si.

1.4 Contribuições

O presente trabalho conta com quatro principais contribuições.

Em primeiro lugar, e o contributo mais direto, são as conclusões retiradas da análise dos diversos algoritmos de deteção de alterações florestais e o seu impacto na análise da idade das florestas. Nomeadamente, salientar a utilidade do algoritmo BOCPD e possíveis parametrizações dos seus parâmetros bem como o peso e importância de cada uma.

Em segundo lugar resultam dois produtos derivados de outros conjuntos de dados já existentes, nomeadamente um onde estão calculadas as alterações na COS entre as edições de 1995, 2007, 2010 e 2015. E um outro produto onde os dados da área ardida provenientes do Instituto da Conservação da Natureza e das Florestas (ICNF) foram normalizados e unificados desde 1975 até 2018.

Finalmente, foi produzido de forma manual um conjunto de dados de referência com informações relativas a alterações na floresta de Portugal continental. Este conta com 664 pontos de floresta portuguesa com o ano de todas as alterações manualmente classificadas de forma exaustiva entre 1986 e 2019.

1.5 Estrutura do documento

O Documento está dividido em 7 Capítulos:

- **1 Introdução** : Este primeiro capítulo é o ponto de partida do trabalho, começa por apresentar o contexto do trabalho ao leitor, posteriormente expõe brevemente o problema a resolver e a abordagem tomada para o explorar. Finalmente temos a presente secção que fornece uma descrição do documento.
- **2 Dados e produtos de informação florestal** : O segundo capítulo apresenta as fontes de dados usadas no presente trabalho. Após uma breve introdução para esclarecer conceitos importantes, o capítulo apresenta dois grandes tipos de dados usados: dados sobre a floresta portuguesa e produtos de deteção remota.
- **3 Trabalho Relacionado e Estado da Arte** : O terceiro capítulo põe o leitor a par do trabalho relacionado importante para a elaboração desta dissertação. Numa

primeira fase são apresentados estudos que lidam diretamente com a idade das florestas. Seguidamente são abordados estudos que lidam com deteção de pontos de quebra em séries temporais de dados de deteção remota. Finalmente são analisados alguns estudos que lidam com pontos de quebra mas de âmbito geral.

- **4 Abordagem e Metodologia** : O quarto capítulo merece especial destaque por ser o responsável em dar uma visão global do trabalho desenvolvido durante a elaboração do presente trabalho. Sem entrar em detalhes de implementação, são explicadas as principais tarefas realizadas bem como o seu contributo para o trabalho.

- **5 Implementação** : O quinto capítulo conta com os detalhes de implementação mais relevantes. Neste capítulo são apresentadas as tecnologias e ferramentas usadas para desenvolver o presente trabalho, nomeadamente o Google Earth Engine (GEE) que foi a ferramenta mais utilizada.

- **6 Avaliação e Resultados** : O sexto capítulo reporta os resultados obtidos com as experiências realizadas. Em primeiro lugar são explorados os dados de referência recolhidos. Seguidamente são analisados os tempos de exportação da plataforma GEE. Finalmente são analisados os resultados da experiências de deteção de idade das florestas.

- **7 Conclusões** : O sétimo e último capítulo faz um apanhado das principais conclusões do trabalho. Há também espaço para explicar os principais contributos diretos do trabalho. Finalmente, é feita uma reflexão sobre o trabalho futuro tendo em conta o que foi explorado.

2

Dados e produtos de informação florestal

Dado que o presente trabalho tem como foco principal a floresta portuguesa, é importante ter uma visão do tipo de dados disponíveis sobre a mesma. Numa primeira, fase para confirmar que a informação sobre a idade é escassa; e numa segunda fase, para olhar para os dados disponíveis e compreender que informações relativas à idade é possível retirar.

Este capítulo oferece uma visão dos dados disponíveis, contextualizando a sua importância para a determinação da idade das florestas. Adicionalmente, apresenta alguns conceitos ligados ao domínio dos dados. Em 2.1 definem-se dois conceitos importantes: floresta e idade da floresta. Em 2.2 apresentam-se os tipos e dados disponíveis sobre a floresta portuguesa, salientado as suas diferenças e similaridades e qual a sua importância para o trabalho. Em 2.3 começam-se por esclarecer alguns conceitos sobre deteção remota e apresentam-se fontes de dados de satélite. Finalmente, em 2.4 resume-se os aspetos mais importantes a reter do presente capítulo.

2.1 Introdução

Antes de apresentar os dados disponíveis sobre a floresta é importante definir de forma mais concreta o conceito de floresta.

Para efeitos do IFN, o ICNF define floresta como: "Terreno, com área mínima de 0,5 ha e largura mínima de 20 m, com árvores florestais com uma altura mínima de 5 m e um grau de coberto mínimo de 10%, ou com capacidade para atingir esses limiares in situ."[24]

No entanto, para efeitos de classificação do uso e ocupação do solo, a COS, realizada pela Direção Geral do Território, define floresta como: "Áreas ocupadas por conjuntos de árvores florestais resultantes de regeneração natural, sementeira ou plantação. As árvores devem, em condições climatéricas normais, ter uma altura superior ou igual a 5 m e no seu conjunto constituir uma área com grau de coberto superior ou igual a 30%. O sob-coberto não é dedicado à agricultura nem a atividades recreativas quando inseridas num contexto urbano."[15]

Apesar das diferenças de requisitos, que se devem aos diferentes contextos das duas definições, há pontos em comum que devem ser salientados na definição de floresta. Em primeiro lugar, são áreas com um conjunto de árvores florestais. Em segundo lugar, as árvores presentes devem ter uma altura mínima de 5m. Finalmente, em terceiro lugar, há um limite mínimo de coberto, num caso 30%, noutro caso 10%. É importante salientar que a definição do ICNF salvaguarda que, mesmo que uma zona não cumpra os limites mínimos de altura e coberto, pode ser considerada floresta se tiver condições para atingir estes limites. Esta salvaguarda é importante, pois permite incluir na definição de floresta zonas que tenham ardido recentemente e zonas onde uma replantação de floresta tenha ocorrido. Este conjunto é suficiente para uma definição satisfatória de floresta para o presente trabalho. Surge no entanto a necessidade de especificar de forma mais rigorosa como é determinada a idade da floresta.

Para efeitos do inventário florestal nacional, o ICNF tem um conjunto de definições que ajudam a chegar a uma definição de idade: "Idade do povoamento - Média das idades das árvores dominantes."[24] "Árvores dominantes - Correspondem às árvores com maior DAP da parcela de inventário. É a partir destas árvores que são avaliadas a altura dominante, o diâmetro dominante e a idade do povoamento."[24] "DAP - Diâmetro à Altura do Peito - Diâmetro do tronco da árvore medido sobre a casca a 1,30 metros do solo. (unidades: cm)"[24]

Reformulando este conjunto de definições, é possível afirmar que a idade de uma parcela de floresta é definida pela média das idades das árvores com um maior diâmetro quando medida a uma altura de 1,30 m. Novamente, a definição deixa alguma margem de manobra, mas, no entanto há casos que são fáceis de compreender. É possível afirmar que qualquer tipo de eventos como incêndios e abates florestais têm influência nesta idade, enquanto que a limpeza de parcelas não deverá ter impacto na mesma.

Tendo estes conceitos em mente, é possível realizar uma análise mais cuidada dos dados relevantes para a determinação da idade das florestas. Os dados aqui mencionados dividem-se em dois grandes grupos, cada um com a sua respetiva secção. Por um lado, temos os produtos de deteção remota, que contam com catálogos de imagens de satélite da superfície terrestre; por outro, os produtos de informação sobre a floresta, uma categoria mais abrangente que conta com uma grande variedade de dados sobre a floresta. Nesta segunda categoria, enquadram-se fontes de informação que caraterizam o uso de solo, assinalam áreas ardidas, entre outras.

2.2 Produtos relevantes para extração de informação sobre a floresta portuguesa

Antes de tentar produzir informação sobre a idade das florestas em Portugal, é importante perceber quais as informações já disponíveis sobre a floresta, e quais desses produtos podem ser de interesse para o presente trabalho.

Durante a procura de fontes de informação, houve a necessidade de ter em conta alguns critérios para selecionar produtos relevantes. Em primeiro lugar, como o âmbito geográfico do trabalho é a totalidade de Portugal, é necessário que as fontes de informação sejam georreferenciadas e cubram a totalidade do território de Portugal continental. Em segundo lugar, os dados tem de conter informação relevante para o trabalho, como por exemplo, classificação de florestas e até eventos de destruição florestal (incêndios).

Para além dos dois principais requisitos referidos anteriormente, há também duas preferências, relativas ao detalhe espacial e à especificidade dos dados. São preferíveis fontes de informações específicas a Portugal, estando assim mais adaptadas ao país e suas peculiaridades. Finalmente, por norma, fontes com maior detalhe espacial são preferíveis a fontes com menor detalhe espacial, o mesmo se aplica ao detalhe temporal.

Tendo estes critérios em vista, foram selecionadas as seguintes fontes de informação:

- Inventário Florestal Nacional (IFN)
- Cartografia da área ardida do Instituto da Conservação da Natureza e das Florestas (ICNF)
- Carta de Uso e Ocupação do Solo (COS)
- *CORINE Land Cover* (CLC)

Individualmente, cada conjunto de dados fornece informação sobre o seu domínio específico e, em parte, útil para determinar a idade da floresta. No entanto, nenhuma fonte de informação indica de forma inequívoca a idade da floresta. É importante, por isso, correlacionar informação das várias fontes apresentadas para obter uma imagem mais completa e precisa da evolução da floresta portuguesa ao longo do tempo.

O uso de várias fontes de informação traz consigo um conjunto de desafios, estes derivam principalmente da heterogeneidade dos dados. Para melhor compreender como tirar proveito das várias fontes de informação ao mesmo tempo, primeiro é necessário compreendê-las melhor individualmente e posteriormente compreender como se podem complementar. Segue-se uma apresentação das várias fontes de informação mencionadas.

2.2.1 Inventário Florestal Nacional

Em Portugal é produzido a cada 10 anos o IFN onde consta informação das florestas nacionais. O primeiro inventário foi realizado em 1965 e o mais recente em 2015. A informação proveniente destes inventários é importante, dado que servem como base para realizar o ordenamento de território, compreender como evolui a floresta portuguesa ao longo do tempo, definir zonas de recuperação, entre outras aplicações. O IFN é uma iniciativa de natureza estatística, realizada através da análise de fotografias aéreas, que é complementada com visitas ao terreno para realizar medições.[25]

A partir desta iniciativa, são gerados dois conjuntos de dados: fotopontos e os dados de campo. Os fotopontos consistem numa grelha regular de pontos, que se encontram

espaçados de 500 em 500 metros e cobrem a totalidade de Portugal continental. Associado a cada ponto, há um conjunto de informação das quais se destacam:

- **Uso do solo** - Qual o tipo de uso de solo presente. Ex: Floresta, Urbano, Agricultura, etc.
- **Ocupação principal** - Qual a espécie presente no local. Ex: Eucalipto, Pinheiro, etc.
- **Percentagem do coberto** - Qual a percentagem coberta pela ocupação principal. Ex: [20, 30[%, [60, 70[%, etc.

Tendo uma ideia do que este conjunto de dados descreve, é também importante perceber como foi criado. Segundo o relatório do Inventário Florestal Nacional nº6: "Cada fotoponto foi classificado em função das características foto-interpretadas na mancha de terreno onde o ponto incidiu. Entendeu-se por 'mancha' a porção de terreno, de área igual ou superior a 5000 m^2 , e de largura média igual ou superior a 20 m, que constitui uma unidade homogénea considerando o uso e ocupação do solo."[24]. É possível concluir que o processo é realizado por um interpretador que analisa imagens de satélite em conjunto com outras fontes de dados para extrair informação.

O segundo conjunto de dados gerados para o Inventário Florestal Nacional são os dados de campo. Estes resultam da recolha de uma variedade de dados no local e têm como principal objetivo a recolha de dados com maior detalhe e rigor. Dos três inventários florestais disponíveis no website do ICNF, apenas há acesso aos dados de campo do inventário florestal número 4, que data de 1995. Para os 2336 pontos de amostragem é possível consultar: a sua localização, idade, espécies, data do último incêndio, etc.[12]

2.2.2 Cartografia da área ardida em Portugal

O ICNF é responsável por, todos os anos, realizar um levantamento das áreas ardidas em Portugal. Deste levantamento, resulta um ficheiro, no formato *shapefile*, com polígonos para representar as áreas ardidas. Estes polígonos contêm informações associadas, tais como: data do início e fim do incêndio, causa, entre outras, permitindo assim localizar no espaço e no tempo os incêndios que ocorreram em Portugal nos últimos 30 anos. No entanto, diferentes anos contam com diferentes informações, sendo que anos mais recentes contam com mais detalhes sobre cada área ardida. Na Figura 2.1 é possível observar vários polígonos bem como alguma informação sobre um dos incêndios que ocorreram em Pampilhosa da Serra. [36]

2.2.3 Uso e Ocupação do Solo

Relativamente a fontes de informação que fornecem dados sobre a ocupação e uso do solo, há duas que se destacam: a Carta de Uso e Ocupação do Solo e o *CORINE Land Cover*. Estas fontes de dados informam sobre o uso e a ocupação do solo, por exemplo: se uma dada zona é uma floresta, um oceano, parte de tecido urbano, entre outras. As classes

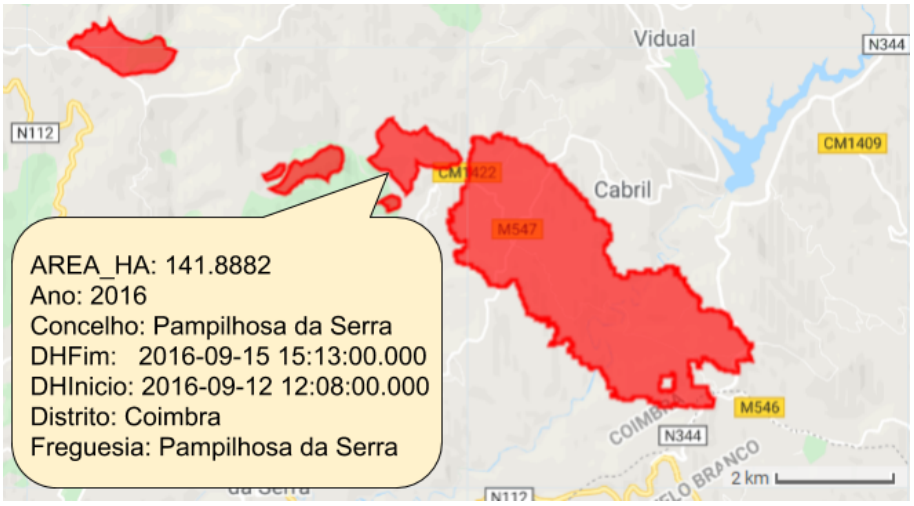


Figura 2.1: Um exemplo com vários polígonos de áreas ardidas na zona de Pampilhosa da Serra e algumas informações associadas como a data e hora de início(DHInicio) e área ardida em hectares (AREA_HA). Localização: 40.096341, -7.923360 WGS 84

utilizadas variam entre os dois conjuntos de dados, e não tendo encontrado nenhuma tradução entre as duas classificações, impossibilitam a comparação entre as duas. Ainda assim, são fontes complementares importantes para ter em mente. Segue-se mais algum detalhe individual sobre ambas.

A Carta de Uso e Ocupação do Solo é realizada pela Direção Geral do Território e conta com cinco anos de referência (1995, 2007, 2010, 2015 e 2018)[7]. Esta iniciativa produz um mapa temático a partir da interpretação de ortofotos onde é possível compreender para cada zona do país qual o tipo de ocupação do solo. Há várias classes de terrenos que podem ser agrupadas caso seja necessário menos detalhe. Por exemplo, na COS 2018 há 4 níveis de detalhe no nível 1 (o menos detalhado) temos florestas, agricultura, territórios artificializados entre outros, e no nível 4 (o mais detalhado) categorias como floresta de sobreiro, floresta de azinheira, aeroportos, etc.

Algumas das classes presentes na COS fornecem informação limitada sobre a idade da floresta. Este tipo de informação é capaz de salientar alterações de uso de território que implicam a replantação de uma floresta e, consequentemente, uma idade para a floresta presente. No entanto, as várias edições lançadas até ao momento usam nomenclaturas distintas. Sem algum tipo de tratamento dos dados, não seria possível fazer comparações entre as várias edições.

Durante a produção da COS (2018) foram detetados erros na COS (2015). Face a estes achados, decorrem agora esforços para corrigir estes erros em versões anteriores. Estas correções tornarão esta série temporal coerente de modo a que seja possível uma melhor compreensão da evolução do território português através da comparação entre os diversos anos.

Esta iniciativa fornece dados sobre o uso das florestas portuguesas, nomeadamente informação sobre o tipo de vegetação que lá podemos encontrar (eucalipto, pinheiro bravo,

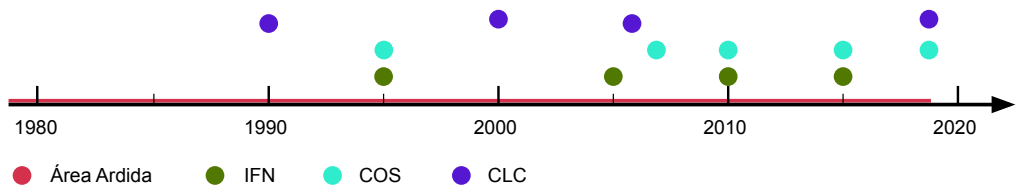


Figura 2.2: Visualização da distribuição temporal das fontes de dados relevantes para a extração de informação sobre a floresta portuguesa.

pinheiro manso, etc.) e se é uma exploração agrícola florestal. Assim, é também uma fonte de informação de referência para o estado das florestas em Portugal. [14]

O *CORINE Land Cover*[23, 10] é um projeto a nível europeu para a construção de um inventário de uso de solo que conta já com 5 edições (1990, 2000, 2006, 2012 e 2018). O inventário contém uma nomenclatura hierárquica que inclui três níveis de detalhe. No nível de detalhe mais baixo, conta com 5 grupos: superfícies artificiais, áreas agrícolas, florestas e zonas seminaturais, zonas húmidas e corpos de água. O nível de maior detalhe conta com um total de 44 classes de ocupação do solo. Para além do tipo de ocupação, também é fornecido um ficheiro das alterações entre anos sucessivos.

Este inventário é similar ao COS, a principal diferença é a unidade mínima de área representada em cada um, isto é, qual a menor área representada. Enquanto que na COS a unidade mínima é de 1 ha, no CLC a unidade mínima é 25 ha, o que torna o COS numa representação com maior granularidade espacial.

2.2.4 Conclusão

Tendo apresentado individualmente cada conjunto de dados, é agora possível compreender que nenhum destes é capaz de indicar a idade da floresta para um qualquer ponto em Portugal. Olhando para os *datasets*, a informação que podemos retirar não é a idade, mas sim conjuntos de eventos que têm impacto na idade da floresta. No caso dos incêndios, é fácil deduzir que um incêndio que destrua a floresta marca o momento em que a idade da floresta passa a ser zero. O mesmo se aplica a uma zona que numa dada altura tenha um tipo de floresta e que, numa altura posterior, tenha um tipo distinto - algures no meio é de esperar um corte. Por exemplo, uma zona de pinheiros que passa a ter uma exploração de eucaliptos. A melhor forma de aproveitar ao máximo as fontes de informação disponíveis para compreender a história de uma dada zona é olhando para as fontes como eventos que alteram a idade da floresta.

No entanto, também é importante compreender que a história que estes dados contam não é exaustiva. Existem duas grandes lacunas nos dados apresentados: a primeira lacuna está relacionada com a resolução temporal, que, para além de ser bastante superior a um ano em quase todos os conjuntos de dados, também é irregular, dificultando a sua análise.

É possível observar este facto na imagem 2.2. Por exemplo, entre edições consecutivas da COS podem passar-se entre 3 a 12 anos. A segunda lacuna tem que ver com a falta de informação sobre um tipo de evento importante nas alterações de uma floresta, que é o abate de árvores. Este tipo muito relevante de alterações não está presente em nenhum conjunto de dados. Em [9] os dados de referência indicam que 30% das alterações à floresta são devidas ao abate floresta; ainda que o estudo seja numa área diferente e tenha em conta outros tipos de alteração, é possível concluir que este tipo de alterações é pertinente quando se analisam alterações florestais.

Em suma, há uma grande variedade de informação sobre as florestas portuguesas dispersa por várias fontes de informação. No entanto, nenhuma fonte é capaz de fornecer uma história completa de uma dada zona de floresta. Para melhor aproveitar todos os recursos apresentados, é necessário uma visão que seja capaz de compatibilizar as diferentes fontes de modo a obter uma perspetiva mais completa da história da floresta, ainda que incompleta.

2.3 Dados e produtos de deteção remota

2.3.1 Introdução

Segundo Campbell e Wynne [6], o termo deteção remota já foi definido várias vezes. Ainda que as ideias principais sejam bem claras, os limites são mais complicados de definir. O autor propõe a seguinte definição: "Deteção remota é a prática de derivar informação sobre a superfície terrestre e os seus corpos de água através de imagens adquiridas de uma perspetiva aérea, usando radiação eletromagnética numa ou mais regiões do espectro eletromagnético, refletida ou emitida pela superfície terrestre."

Tal como é referido na descrição de Deteção Remota, a informação é adquirida no espectro eletromagnético, pelo que é necessário ter algumas bases para melhor compreender o trabalho realizado nesta área. Algumas manifestações eletromagnéticas são familiares, tais como a luz no visível, os raios ultra-violeta, o infravermelho ou até mesmo os raios-X. Estas manifestações de energia enquadram-se no espectro eletromagnético, cada um com comprimento de onda caraterístico. Ao atingir um determinado alvo a radiação pode ser, absorvida, refletida ou transmitida dependendo do comprimento de onda e do material atingido. No caso das plantas, estas contêm clorofila que é um composto químico que absorve radiação vermelha e azul e reflete o verde dando a aparência verde às mesmas. Já no caso da água, esta tende a absorver radiações com maior comprimento de onda, como é o caso do infravermelho; consequentemente, apresenta uma cor azulada. Conclui-se, portanto, que a partir da assinatura eletromagnética é possível deduzir informação sobre o alvo observado[30].

Os sensores usados em deteção remota capturam a intensidade em várias bandas do espectro eletromagnético e apresentam-nas na forma de uma imagem. Segundo Köhl,

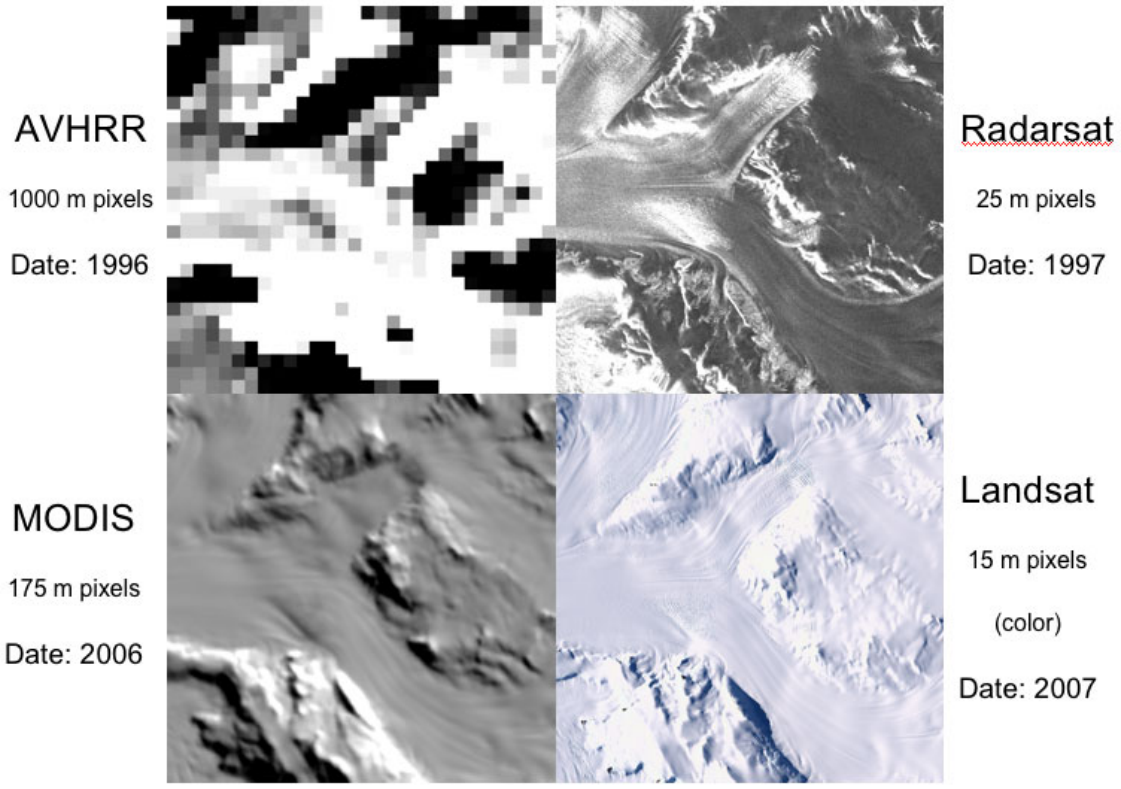


Figura 2.3: Quatro imagens provenientes de sensores distintos com diferentes resoluções espaciais. Imagem retirada de: https://www.nasa.gov/vision/earth/lookingatearth/lima_feature.html

Magnussen e Marchetti [30] cada pixel destas imagens representa uma área da superfície terrestre à qual chamamos resolução espacial. Na figura 2.3 é possível observar que as imagens com maior resolução espacial apresentam também mais detalhe. Em [45] é dito que, apesar de instintivamente se poder pensar que é sempre melhor imagens mais detalhadas, na prática nem sempre é o caso. A resolução espacial adequada depende do objetivo a alcançar. Para melhor provar este ponto, o autor dá um exemplo do dia a dia, dizendo que reconhecemos uma cara pela combinação de caraterísticas físicas tais como os olhos, nariz e boca. Segue-se que a resolução indicada para este tipo de tarefa seria uma que, por um lado, permitisse identificar estes elementos e por outro localizá-los no contexto de uma cara.

Além da resolução espacial, também há outro tipo de resolução importante: a resolução temporal, que indica a capacidade de deteção de mudança e é usada para uma sequência de imagens de um mesmo sensor. É comum um satélite ter uma órbita com um período fixo de 10, 16 ou 24 dias[33], pelo que esse número de dias seria a sua resolução temporal. Com visitas mais frequentes, serão obtidos mais dados, o que, por um lado, representa de forma mais clara as alterações que ocorrem ao longo do tempo, e, por outro, necessita de mais tempo de processamento.

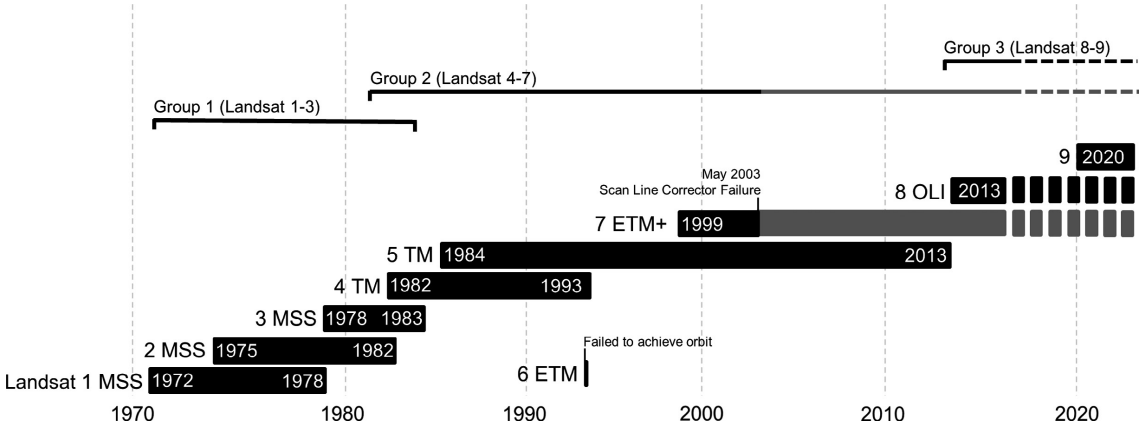


Figura 2.4: Cronologia do projeto Landsat. Imagem retirada de [48]

Os sensores que capturam bandas no espectro do radar têm algumas vantagens, nomeadamente: é possível a observação noturna e não está limitado pelas condições atmosféricas, já que este tipo de radiação não é afetado pelas nuvens, ao contrário do espetro visível. Porém, os sensores radar também têm algumas desvantagens, nomeadamente o ruído presente nas imagens (conhecido como *speckle*), que pode ser corrigido com algum processamento adicional.

Dado que este tipo de sensores capturam comprimentos de onda maiores, conseguem captar caraterísticas como a textura do terreno. Infelizmente, não há um catálogo tão vasto deste tipo de dados do mesmo modo que há de imagens na região do visível.

2.3.2 Missões Landsat

O Landsat é o projeto de recolha de imagens de satélite com o maior catálogo histórico. O primeiro satélite foi lançado em 1972, na altura ainda com o nome Earth Resources Technology Satellite (ERTS). Desde então, mais 6 satélites estiveram ao serviço. O mais recente (Landsat 8) iniciou o processo de observação terrestre no início de 2013. No momento da escrita a missão Landsat 9 já tem data estimada de lançamento para o setembro de 2021[32]. Garantindo a continuação deste projeto. É possível observar a cronologia do projeto Landsat na figura 2.4, salienta-se a missão 6 por não ter atingido a órbita.

As imagens de satélite recolhidas por este projeto contam com várias bandas do espectro eletromagnético. Diferentes satélites contam com diferentes sensores que se enquadram em um de três grupos(ver figura 2.4). O primeiro grupo de satélites conta com as missões 1,2 e 3. Este grupo usa o *Multispectral Scanner* (MSS), que captura quatro bandas como é possível observar na tabela 2.1. O segundo grupo é constituído pelas missões 4 até à 7. As missões 4 e 5 utilizam o sensor *Thematic Mapper* (TM) juntamente com MSS, a missão 7 utiliza o sensor o *Enhanced Thematic Mapper* (ETM+). Os sensores deste segundo grupo contam com uma maior resolução espacial. Para além disso também capturam um maior número de bandas. Finalmente o terceiro grupo conta com a missão 8 e utiliza dois

Tabela 2.1: A tabela mostra para cada zona do espectro eletromagnético a banda correspondente de entre os vários sensores usados pelas missões Landsat. * - dados capturados a uma resolução diferente mas posteriormente processados para uma resolução superior. Tabela adaptada de [48]

Zona do espectro	LS 1–5 MSS	LS 4–5 TM	LS 7 ETM+	LS 8 OLI/TIRS	Pixel size (m)
Coastal aerosol				B1	30
Blue		B1	B1	B2	30
Green	B1	B2	B2	B3	30 (60 for MSS)
Red	B2	B3	B3	B4	30 (60 for MSS)
NIR 1	B3				60
NIR	B4	B4	B4	B5	30 (60 for MSS)
SWIR 1		B5	B5	B6	30
SWIR 2		B7	B7	B7	30
Thermal		B6	B6	B10 B11	30*
Pan- Chromatic			B8	B8	15
Cirrus				B9	30

sensores: *Operational Land Imager* (OLI) e *Thermal Infrared Sensor* (TIRS). Estes sensores adicionam ainda mais bandas e contêm informação para o processo de calibração das imagens de satélite.

Relativamente à resolução temporal destas missões, quando consideradas de forma individual a resolução temporal é de 16 dias, ou seja, 16 é o número mínimo de dias entre imagens consecutivas de uma dada missão Landsat. No entanto, como é possível observar na imagem 2.4 há várias missões que decorrem em simultâneo, tirando proveito deste facto é possível, em alturas onde duas missões estão ativas, obter uma resolução temporal de 8 dias.

2.3.3 Moderate-resolution Imaging Spectroradiometer (MODIS)

O Moderate-resolution Imaging Spectroradiometer (MODIS) é um instrumento de observação terrestre presente a bordo de dois satélites: Terra e Aqua. Este par de satélites é capaz de obter observações para a totalidade da superfície terrestre em 1 ou 2 dias. Adicionalmente, o sensor captura um total de 36 bandas do espectro eletromagnético com resolução espacial que varia entre os 250 m, 500 m e 1000 m. Os dados capturados por este sensor são determinantes para compreender mudanças a nível planetário[38]. Observações do satélite Terra estão disponíveis desde o ano 2000 e para o satélite Aqua desde 2002. Ambos os satélites continuam a produzir observações até à data de escrita do presente documento [44, 3].

2.3.4 Sentinel-2

O Sentinel-2 é uma missão de recolha de imagens multi-espectrais de alta resolução e frequência composta por dois satélites. O tempo de revisita é de 10 dias se apenas um satélite for considerado, passando para metade (5 dias) quando ambos os satélites são usados. O primeiro satélite foi lançado em 2015 e o segundo dois anos depois, em

2017 [1]. As imagens capturadas contam com treze bandas do espectro eletromagnético: quatro com a resolução de 10 m, seis com 20m e, finalmente, três com 60m. O Sentinel-2 foi criado tendo em mente complementar o papel que as missões Landsat desempenham [13].

2.3.5 Transformações e Índices Espectrais

Há bastante informação em cada banda de uma imagem de satélite, ainda que nem sempre seja claro. Para salientar esta informação são usadas transformações nas bandas. Usar este tipo de transformações é por vezes importante para reduzir a quantidade de dados a processar mantendo a informação necessária. De seguida serão apresentadas algumas transformações usadas em deteção remota.

O NDVI é um índice usado em deteção remota. Este índice é usado em estudos que se focam em florestas por salientar a vegetação. O NDVI é calculado com base em duas bandas infravermelho (IV) e vermelho (V). A equação 2.1 indica como o calcular usando estas duas bandas.

$$\text{NDVI} = \frac{IV - V}{IV + V} \tag{2.1}$$

Este índice varia entre -1 e 1. Quando o valor da banda IV é superior a V o NDVI apresenta um valor positivo, caso contrário, o NDVI é negativo. Como a vegetação absorve luz vermelha e reflete o infravermelho apresenta um NDVI positivo.

O NBR apresentado em [34], é um índice especialmente útil para detetar área ardida. Este é similar ao NDVI, no entanto em vez do vermelho usa os infravermelhos de onda curta(SWIR)

$$\text{NBR} = \frac{IV - SWIR}{IV + SWIR} \tag{2.2}$$

O comportamento do NBR é similar ao NDVI, porém este índice tem a particularidade de acentuar zonas devastadas por incêndios.

O Structural Index (SI) é definido em Fiorella e Ripple [20] como o rácio entre as bandas 4 e 5 do instrumento TM abordo das missões Landsat.

$$\text{SI} = \frac{TM4}{TM5} \tag{2.3}$$

A transformação *Tasseled Cap* é apresentada em [11] como uma transformação usada para reduzir a dimensão dos dados dos instrumentos MSS e TM das missões Landsat. Isto é esta transformação processa as 7 bandas do instrumento TM e compacta a informação em 3 dimensões ortogonais às quais é possível associar um significado físico. As 3 dimensões resultantes são: brilho 2.4, verde (*greenness*) 2.5 e humidade (*wetness*) 2.6. Esta transformação preserva a distância euclidiana e captura 95% da variabilidade dos dados. São apresentadas mais 3 dimensões ortogonais, porém, a maior parte da informação já se encontra nas três principais dimensões. Esta transformação também está disponível para

outras missões como o NBR e o Sentinel-2 porém os parâmetros usados são diferentes [41, 47].

$$\begin{aligned} \text{brilho} = {} & 0.3037(TM1) + 0.2793(TM2) + 0.4743(TM3) \\ & + 0.5582(TM4) + 0.5082(TM5) + 0.1863(TM7) \end{aligned} \tag{2.4}$$

$$\begin{aligned} \text{verde} = {} & (-0.2848(TM1)) + (-0.2435(TM2)) + (-0.5436(TM3)) \\ & + 0.7243(TM4) + 0.0840(TM5) + (-0.1800(TM7)) \end{aligned} \tag{2.5}$$

$$\begin{aligned} \text{humidade} = {} & 0.1509(TM1) + 0.1973(TM2) + 0.3279(TM3) \\ & + 0.3406(TM4) + (-0.7112(TM5)) + (-0.4572(TM7)) \end{aligned} \tag{2.6}$$

O índice ângulo *Tasseled Cap* (TCA) é definido em [39] e usa dois valores resultantes da transformação *Tasseled Cap*: brilho 2.4 e verde 2.5. O índice TCA está definido pela equação 2.7

$$\text{TCA} = arctan(\frac{verde}{brilho}) \tag{2.7}$$

2.3.6 Conclusão

Tendo em conta os três satélites, é importante compreender qual o mais adequado para o trabalho em questão. Dado que o trabalho em questão lida com séries temporais, o historial dos satélites tem um grande peso na decisão; isto porque só serão detetadas alterações durante o período de disponibilidade de imagens. Este critério penaliza o Sentinel-2 por ser um projeto bem mais recente do que o Landsat e o NBR. Comparando as duas fontes de imagens de satélite, deparamo-nos com um dilema entre resolução temporal e resolução espacial. Por um lado, as missões Landsat oferecem uma resolução espacial 30m x 30m e uma resolução temporal de 16 dias. Por outro lado, o NBR apresenta uma resolução espacial de 500m x 500m e uma resolução temporal de 2 dias. Tendo apenas estes aspetos em consideração, não há uma escolha clara, mas, tendo em conta a longevidade das duas missões, há uma clara vantagem para as missões Landsat, mesmo excluindo as primeiras três missões por uma questão de continuidade. Finalmente há que ter em conta que os estudos apresentados no capítulo 3, os quais foram importantes para o desenvolvimento deste trabalho, usam imagens Landsat. Assim sendo, torna-se mais claro que as imagens provenientes das missões Landsat são as mais indicadas a usar durante a análise realizada.

2.4 Conclusão

Os conjuntos de dados apresentados estão divididos em dois grupos: por um lado produtos relevantes para extração de informação sobre a floresta portuguesa 2.2 e, por outro, dados e produtos de deteção remota 2.3.

Relativamente ao primeiro conjunto de dados, que conta com o Inventário Florestal Nacional (IFN), áreas ardidas, Carta de Uso e Ocupação do Solo (COS) e *CORINE Land Cover* (CLC), é possível afirmar que têm uma diversidade de fontes de informação, mas nenhuma delas só por si é capaz de apontar a idade das florestas portuguesas. No entanto, a informação que disponibilizam em conjunto complementa-se de modo a obter uma história mais coesa e completa da floresta portuguesa. Os incêndios e as alterações de uso de solo são eventos possíveis de detetar nestes conjuntos de dados que têm influência na idade da floresta presente no local. No entanto, não há nenhuma fonte de informação que garanta a estabilidade de uma zona de floresta, isto porque uma zona que é constantemente declarada como floresta de eucaliptos pode ter vários cortes ao longo dos anos sem que haja qualquer tipo de repercussão no que está assinalado nos conjuntos de dados. Para poder criar um dataset da idade das florestas é preciso mais do que este conjunto de dados disponibiliza.

Relativamente ao segundo conjunto de dados, das imagens de satélite é possível concluir que há uma grande disponibilidade de imagens de satélite, das quais se destaca, para o presente trabalho, as missões Landsat. Esta fonte de imagens de satélite conta com um largo historial de imagens que se enquadra perfeitamente para a análise da floresta ao longo do tempo. Para além das imagens de satélite, também é importante salientar a relevância que os índices radiométricos têm na análise de imagens. Em vez de se trabalhar com uma imagem para cada banda, é possível utilizar um índice como, por exemplo, o NDVI, que saliente o que desejamos, neste caso em particular a vegetação e o seu vigor. A análise da evolução destas imagens de satélite é uma parte importante do presente trabalho, e terá espaço para discussão no próximo capítulo.

3

Trabalho Relacionado e Estado da Arte

O presente capítulo serve para por o leitor a par da literatura essencial para desenvolver o presente trabalho. O capítulo está dividido em cinco secções. Na secção 3.1 são apresentados estudos que lidam diretamente com a deteção da idade da floresta. Estes estudos usam abordagens que apenas têm em conta dados relativos a um momento no tempo, sejam imagens de satélite ou produtos derivados. Na secção 3.2 são apresentados estudos que usam a análise de séries temporais para determinar alterações na floresta, os estudos aqui apresentados são específicos a deteção remota tendo já sido testados nesse contexto. Na secção 3.3 são apresentados estudos que não se focam especificamente em deteção remota, lidam com deteção de alterações em contextos genéricos. Na secção 3.4 é realizado um apanhado dos métodos de avaliação usados com especial foco nas métricas e no seu significado. Finalmente na secção 3.5 conclui-se com as principais ideias apresentadas no capítulo e a forma como se enquadram no trabalho.

3.1 Métodos para deteção da idade das florestas

Determinar a idade de florestas quando há informação relativa ao abate e plantação de zonas geridas é relativamente fácil. Quando não há acesso a este tipo de dados, a informação da idade das árvores pode ser determinada através da contagem dos anéis que se formam anualmente[30]. Estas abordagens não aparentam ser promissoras para obter uma solução escalável ao país, uma vez que, por um lado, não há dados de todo o território e, por outro, a colheita de dados manualmente é trabalhosa e demorada. Olhando para a literatura há uma variedade de abordagens. Seguem-se alguns exemplos pertinentes para a presente dissertação.

Fiorella e Ripple [19] estuda a correlação entre a idade das florestas e os valores espectrais das sete bandas do instrumento *Thematic Maper* abordo das missões Landsat. A zona de estudo é a *H.J. Andrews Experimental Forest* situada no Oregon, Estados Unidos da América e conta com floresta boreal. A partir de uma imagem retirada no dia 30 de julho de 1988, são calculadas 3 transformações: NDVI, SI, e as 3 componentes extraídas da transformação *Tasseled Cap*. Foram estudados três tipos de correlação: linear, log-linear

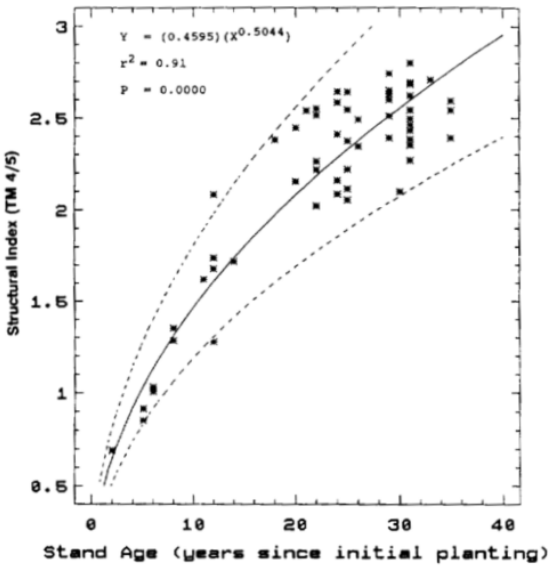


Figura 3.1: Relação entre o índice SI e a Idade para zonas com regeneração normal. Imagem retirada de [19]

e log-log. Os resultados indicam que, à exceção da banda 4, todas as outras bandas apresentam uma correlação inversa forte com a idade, quando esta é inferior a 35 anos. Para os valores das bandas, a relação log-linear apresentou os melhores resultados; enquanto que os índices calculados com base nestas bandas apresentaram melhores resultados usando a relação log-log. É possível observar o resultado da relação log-log entre a idade e o índice SI na figura 3.1 em casos de regeneração florestal normal. Esta relação foi selecionada como a melhor.

Com os conhecimentos adquiridos em [19], Kimes et al. [29] explora a possibilidade de usar redes neuronais para extrair a idade das florestas a partir de dados Landsat. A zona de estudo é também a *H.J. Andrews Experimental Forest* situada no Oregon, Estados Unidos da América e é usada uma imagem do dia 7 de julho de 1991, bem como informação adicional relativa à topologia do terreno. Os dados de referência, os quais contam com cerca de dez mil pontos, são divididos em treino(67%) e teste(33%). Foram testadas várias arquiteturas com 3 camadas: input, camada escondida e output. A camada de input contém tantos neurónios quanto o número de inputs; na camada escondida o número é variável; e na camada final apenas um neurónio, onde surgirá a idade do ponto analisado. Após treinar várias redes neuronais com diferentes inputs concluiu-se que as bandas 1, 2, 6 e 7 não eram significativamente relevantes para a determinação da idade. O melhor resultado foi obtido por uma arquitetura com 6 neurónios de input, 5 escondidos e 1 de output obtendo um erro médio quadrático de 5.59 nos dados de teste. Porém, o autor salienta que outros estudos obtiveram diferentes estruturas favoráveis pelo que é necessário realizar testes para cada região em específico. Em conclusão, é dito que as redes neuronais podem ser usadas para extrair a idade de florestas jovens (idade inferior

a 50 anos) apresentando resultados superiores aos conseguidos através de regressões lineares. O autor acrescenta ainda que informação topográfica como elevação e o declive mostraram-se capazes de melhorar o desempenho deste tipo de abordagem.

Em Drezet e Quegan [16] são utilizadas dados da coerência SAR da missão ERS em conjunção com dados meteorológicos e dados de referência para estimar a idade das florestas britânicas. Os dados meteorológicos têm como principal objetivo a seleção de imagens onde não haja condições adversas, como por exemplo chuva. Os dados de referência são usados para calcular os parâmetros de um modelo matemático desenvolvido para este estudo. O modelo correlaciona a coerência SAR com a idade das florestas. A avaliação deste modelo é feita de duas formas através da análise qualitativa de dois gráficos: num deles pode observar-se que o modelo se assemelha à distribuição dos dados de referência; no outro podem analisar-se gráficos de probabilidade condicionada resultantes.

Zhang et al. [49] tem como principal objetivo o mapeamento da idade das florestas na China. O estudo usa um mapa da altura das florestas com a resolução de 1 km , medições anuais de temperatura e precipitação, bem como informação sobre o tipo de floresta existente. Como modelo, assumem uma relação exponencial entre a altura e a idade de uma floresta e constroem uma equação com vários parâmetros, os quais são posteriormente calculados através de um método de otimização chamado *swarm*. Não são apresentadas métricas para validação dos resultados; como tal, assume-se que o maior erro será proveniente do mapa da altura das florestas e, consequentemente, o mapa de incerteza desse produto é selecionado como uma aproximação da incerteza do produto.

O conjunto de estudos apresentados descreve algumas abordagens para determinar a idade da floresta. Os estudos apresentam algumas limitações no que toca à precisão da idade detetada e à necessidade de dados específicos nem sempre disponíveis, como a altura da floresta. Tendo uma ideia geral dos algoritmos, é possível identificar que todos utilizam informação de apenas um momento no tempo. Para os dois primeiros, estudos são usadas imagens Landsat; para o terceiro, imagens SAR; e, para o último, um mapa da altura das árvores obtido a partir de dados LiDar. Esta abordagem é análoga a tentar determinar a idade de uma pessoa usando apenas uma fotografia. Dado que há acesso contínuo a novas imagens de satélite, é possível tomar uma abordagem diferente onde a análise recai sobre a série temporal criada pelo conjunto de imagens de satélite. O seguinte capítulo apresenta alguns estudos que usam esta abordagem.

3.2 Métodos para a deteção de mudanças florestais

Há um conjunto de estudos que se focam na deteção de distúrbios na floresta. Como olhar para uma só imagem não dá uma visão completa do seu historial, esta abordagem acaba por usar as imagens disponíveis de forma a construir uma série temporal do comportamento. Desta forma é possível observar o historial de uma determinada zona de floresta e reconhecer momentos de distúrbio (cortes, fogos, etc). Estes distúrbios marcam

momentos onde a floresta anterior desaparece e uma nova nasce, sendo assim possível determinar a sua idade. Vejamos agora alguns estudos que adotam esta abordagem.

Em Cohen et al. [9] é feita a comparação entre 7 algoritmos que tiram proveito de séries temporais para detetar distúrbios nas florestas. O tipo de distúrbios tidos em conta variam entre grandes impactos com consequências imediatas, desde incêndios e o abate de árvores, até distúrbios subtis a longo prazo, como os gerados por uma seca.

Todos os algoritmos apresentados usam imagens da missão Landsat entre 1984 e 2012 e têm como unidade base de análise o pixel. Há bastantes diferenças entre as várias abordagens, uma das quais os dados usados para a deteção de mudança. A maioria usa uma imagem por ano para todos os anos em estudo, de modo a obter uma série temporal; no entanto, dois algoritmos processam todas as imagens disponíveis. Estes dois algoritmos são: *Continuous Change Detection and Classification* (CCDC) [50] e Exponentially Weighted Moving Average Change Detection (EWMACD) [4]. Relativamente aos algoritmos que apenas usam uma imagem por ano destaca-se o LandTrendr [27] por estar presente no Google Earth Engine.

O estudo comparativo destes algoritmos de deteção de mudanças [9] conclui que não há um algoritmo inerentemente melhor que outro para a deteção de alterações florestais. É, por isso, necessário compreender as caraterísticas de cada um relativamente a erros de omissão e comissão de modo a escolher o que melhor se enquadra com o objetivo do utilizador. Adicionalmente, algoritmos com o objetivo de detetar um maior espectro de alterações tendem a cometer mais erros de comissão e menos erros de omissão quando comparados com algoritmos com o objetivo de detetar um grupo mais restrito de alterações. Na subsecção 3.4 são fornecidos mais detalhes sobre erros de comissão e omissão. Segue-se uma explicação de cada um, necessária para compreender como funcionam.

3.2.1 Continuous Change Detection and Classification

O CCDC[50] é um algoritmo que deteta alterações do uso de solo. Para este propósito usa uma sequência de imagens Landsat. Estas passam por um processo de remoção de artefactos anómalos como nuvens, sombra e neve de modo a reduzir erros em fases posteriores. A descrição que se segue refere-se a um processamento pixel a pixel, isto é, cada sequência temporal de pixeis é analisada de forma independente das adjacentes.

Tendo acesso a uma série temporal maioritariamente livre de artefactos indesejados, são usadas as primeiras 15 observações para realizar uma regressão de forma a encontrar os parâmetros para um modelo harmónico. Este modelo tem em atenção as alterações que ocorrem num dado ano e as alterações que ocorrem ao longo dos anos. As alterações que ocorrem num dado ano são modeladas usando uma componente harmónica compensando as diferenças existentes entre que os valores encontrados durante o verão dos encontrados durante o inverno. As alterações intra anuais são modeladas usando uma equação linear; desta forma é possível descrever comportamentos como estabilidade, declínio e crescimento. Este modelo é capaz de descrever o comportamento de zonas onde

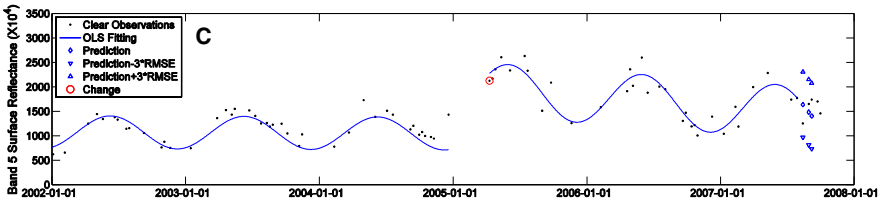


Figura 3.2: Exemplo da deteção de uma alteração por parte do algoritmo CCDC

não há alterações abruptas.

Tendo acesso ao modelo capaz de descrever o comportamento de uma dada zona, desvios significativos deste modelo são indicadores de mudança. Os autores do algoritmo definiram que após 3 observações consecutivas com uma diferença de 3 vezes o desvio padrão será considerada uma mudança. Após a mudança ser detetada são usadas novamente 15 observações para calcular os novos parâmetros do modelo e o processo de procura de anomalias repete-se.

Na imagem 3.2 é possível observar o que foi até ao momento discutido relativamente ao funcionamento do modelo. A imagem retrata uma série temporal com um momento de quebra marcado com um círculo vermelho. Os pontos são as observações obtidas e a linha mostra o modelo gerado para a série temporal em questão. É importante notar que estão presentes duas linhas diferentes, uma antes do ponto de quebra e outra após o ponto de quebra. Na zona mais à direita é possível também observar triângulos a demarcar o intervalo de 3 desvios padrão para três datas distintas; neste caso as observações não excedem o intervalo definido, pelo que nenhuma alteração é detetada.

Relativamente ao desempenho do algoritmo, o estudo original apresenta as seguintes métricas(ver secção 3.4): accuracy de 0.9772, Positive Predictive Value de 0.9772 e True Positive Ratio de 0.8560. Com estes valores é possível também determinar que o F1 toma o valor 0.9126.

No entanto, num outro estudo que compara vários algoritmos, o desempenho medido já é diferente. Neste caso são obtidos os seguintes resultados: False Discovery Rate de 0.22, False Negative Rate de 0.85. Também é possível calcular o F1 com estes resultados, este assume o valor 0.25. É importante ter em atenção que os distúrbios que constam de cada conjunto de referência são diferentes e isso tem impacto nos resultados apresentados, como é possível observar.

3.2.2 LandTrendr

O LandTrendr[27] é um outro algoritmo mas que, ao contrário do CCDC, apenas usa uma imagem por ano. As imagens Landsat são selecionadas entre julho e agosto. O pré processamento consiste na ortorretificação seguida de normalização radiométrica, transformação *tasseled-cap* e, finalmente, deteção de nuvens, neve e sombras. À semelhança do CCDC, o processamento também é realizado pixel a pixel.

A figura 3.3 fornece um diagrama do algoritmo. O primeiro passo, identificado com a letra a), consiste na remoção de alguns picos anómalos que tenham passado a filtragem inicial. Este processo é realizado de forma iterativa, em cada iteração é selecionado o pico mais saliente.

Seguidamente, em b) são identificados potenciais vértices que indiquem mudança. Novamente estamos perante um processo iterativo onde é realizada uma regressão linear para os dados disponíveis. O ponto com maior erro residual é selecionado como ponto de mudança, gerando assim dois grupos sobre os quais volta a ser realizada regressão linear. Este processo é repetido um número definido de vezes pelo utilizador, gerando tantos segmentos de reta quanto os definidos.

Em c) ocorre uma simplificação do modelo criado que consiste em iterativamente remover os vértices que geram um maior ângulo até um número pré determinado pelo utilizador. Este passo acaba por unir segmentos de reta adjacentes e similares.

O passo d) foca-se nas transições entre retas e otimiza-as. Durante este passo são usadas outras técnicas de regressão para aprimorar as transições, tendo como objetivos a redução de ruído e captura de alterações abruptas.

Em e) ocorre um último passo de simplificação onde são removidos os vários vértices iterativamente de forma a minimizar o erro. Em f) compara-se todos os modelos gerados em e) e é selecionado o modelo que apresenta menor erro.

Quanto à performance do algoritmo, no estudo original [28] apenas é feita uma análise superficial ao impacto dos vários parâmetros, tentando compreender quais os efeitos que estes têm nas métricas analisadas. A análise é superficial devido à quantidade limitada de dados de referência; assim, não seria fácil avaliar de forma rigorosa o algoritmo. Adicionalmente, a avaliação realizada não é específica à deteção de quebras mas sim à forma como o algoritmo modela a série temporal.

No entanto, no estudo [9] que compara vários algoritmos na deteção de alterações florestais, é dito que o LandTrendr obtém um False Discovery Rate (equação 3.3) de 0.83 e um False Negative Rate (equação 3.2) de 0.57. A partir destes números é possível determinar que o F1 (equação 3.4) para este conjunto de dados é de 0.6.

3.3 Métodos para deteção de mudanças em séries temporais

Dados os resultados pouco conclusivos das alternativas exploradas até ao momento, procurou-se mais informação sobre algoritmos que não fossem dirigidos especificamente à análise de mudanças florestais. A ideia passou por encontrar outras abordagens que lidassem com a deteção de mudanças em séries temporais de forma genérica para posteriormente se aplicarem na deteção de mudanças florestais. Para este propósito, num estudo de Burg e Williams [5], é feita a comparação de 13 algoritmos de deteção de pontos de alteração.

Do conjunto de algoritmos estudados, destacou-se para o presente trabalho o BOCPD por três razões: em primeiro lugar, apresenta bons resultados em [5] obtendo o melhor F1

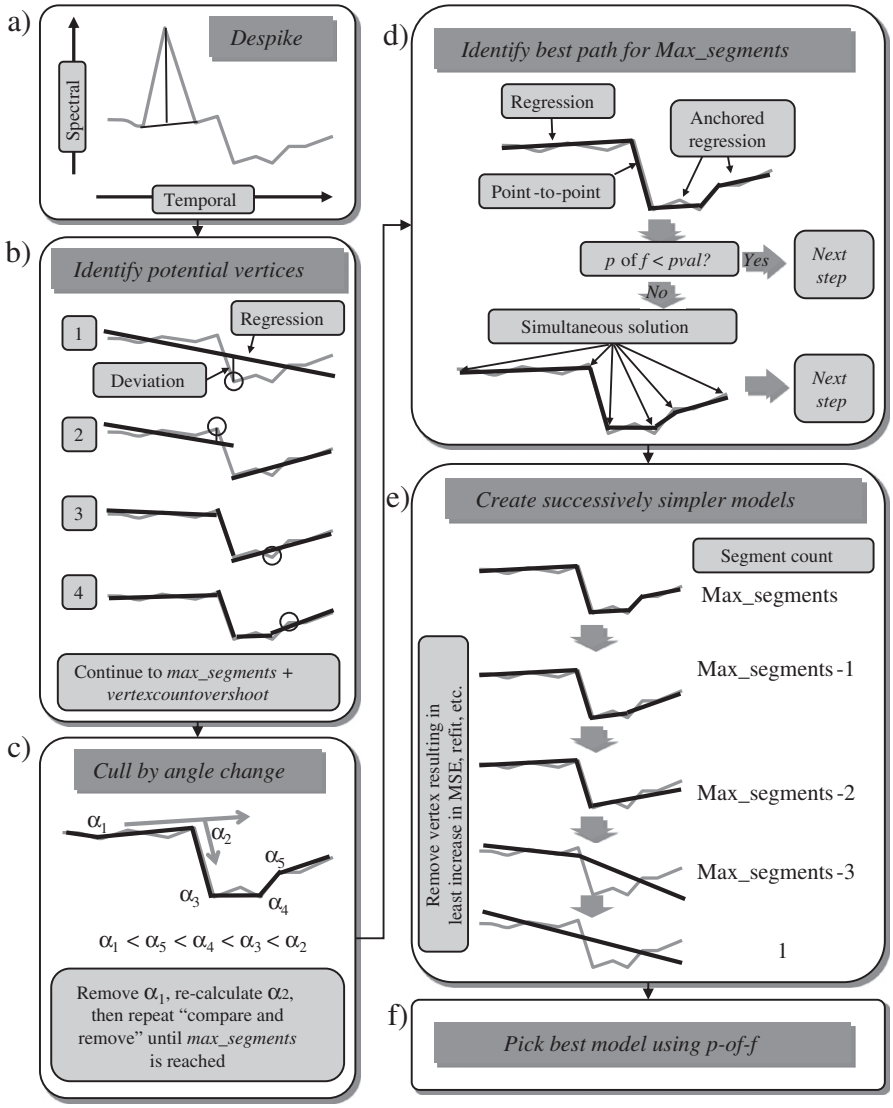


Figura 3.3: Diagrama representativo do processo de deteção de mudanças usado pelo LandTrendr. Imagem retirada de [28]

em média após uma pesquisa em grelha dos parâmetros. Em segundo lugar, encontra-se disponível numa biblioteca python, não havendo assim necessidade de implementação. Finalmente, o algoritmo funciona de forma online tal como o CCDC e o Exponentially Weighted Moving Average Change Detection (EWMACD), dois algoritmos talhados para deteção remota. Esta caraterística permite a contínua ingestão de novos dados por ordem cronológica, não havendo necessidade de ter presente todo o conjunto de dados.

3.3.1 Bayesian Online Changepoint Detection

O BOCPD[2] é um algoritmo que lida com sequências de dados genéricos sendo assim possível aplicá-lo no contexto da deteção remota. Uma vez que não é específico para aplicações de deteção remota, ao contrário dos algoritmos apresentados até ao momento,

não tem quaisquer requisitos de pré-processamento. No entanto, no presente contexto, será importante incorporar ideias de outros algoritmos de deteção remota e remover nuvens, sombras e neve.

Dada uma sequência de observações, o algoritmo calcula para cada observação a distribuição de probabilidades do comprimento da presente sequência desde a última alteração. Olhando para um exemplo em concreto na imagem 3.4. Os três gráficos apresentados partilham o mesmo eixo horizontal que representa o tempo. O gráfico no topo representa um exemplo artificial de input com 3 segmentos distintos com diferente média e variância. O segundo gráfico é o resultado do algoritmo BOCPD e, para cada momento, mostra a distribuição de probabilidade do comprimento (a probabilidade está representada por uma escala logarítmica de cor; a soma das probabilidades para cada momento é 1). Neste segundo gráfico, destacam-se três linhas diagonais com origem nos pontos de mudança; isto quer dizer que dentro de cada uma das sequências, à medida que novos dados chegam, há um consenso sobre o ponto de quebra. A linha vermelha que é possível observar neste gráfico marca probabilidades de estarmos perante sequências de tamanho 20 e está relacionada com o último gráfico (o número 20 não tem nenhuma importância especial, pelo que poderia ter sido escolhido outro). Nesse último gráfico, é possível observar uma "fatia"do gráfico anterior, estando representada com a linha vermelha sólida a probabilidade de haver uma sequência de tamanho 20 para cada observação. Os três picos possíveis de observar estão localizados na zona onde a linha vermelha e as linhas diagonais pretas se intersetam no segundo gráfico. Com o devido desfasamento de 20 observações, é possível encontrar os pontos de quebra (isto está representado no gráfico pela linha pontilhada). Uma forma de olhar para esta linha pontilhada é: tendo em conta o conjunto de observações $x_1, .., x_n, .., x_{n+20}$ qual a probabilidade de haver uma alteração durante a observação n. Desta forma, a deteção de um ponto de mudança pode ter em conta um dado número de observações posteriores, aumentando a certeza de um ponto de quebra. Este número de observações posteriores a analisar foi tomado como um parâmetro ao qual deu-se o nome de **kw**. Um número baixo de **kw** implica pouca certeza sobre um ponto de mudança pois esta confiança cresce com o número de novas observações. Por outro lado, este número também limita o tamanho mínimo do comprimentos de sequências detetáveis, por outras palavras, se duas alterações ocorrerem espaçadas de **kw** observações a primeira não será detetada.

Relativamente à performance do algoritmo, não há valores concretos no estudo onde o algoritmo é apresentado. No entanto, em Burg e Williams [5], o algoritmo é avaliado, juntamente com outros algoritmos, num conjunto de 42 séries temporais de contextos diferentes. Os dados vão desde séries sintéticas, como a apresentada na imagem 3.4, passando por emissões de CO_2 até o número mensal de passageiros que transitaram pelo aeroporto Keflavik na Islândia. Após a otimização dos parâmetros usando uma pesquisa em grelha para todos os algoritmos, foram retiradas as seguintes conclusões: o BOCPD obteve o melhor F1 em 27 das 42 séries temporais, atingindo assim o melhor F1 médio entre todos os algoritmos. Adicionalmente, o pior resultado obtido pelo BOCPD foi um

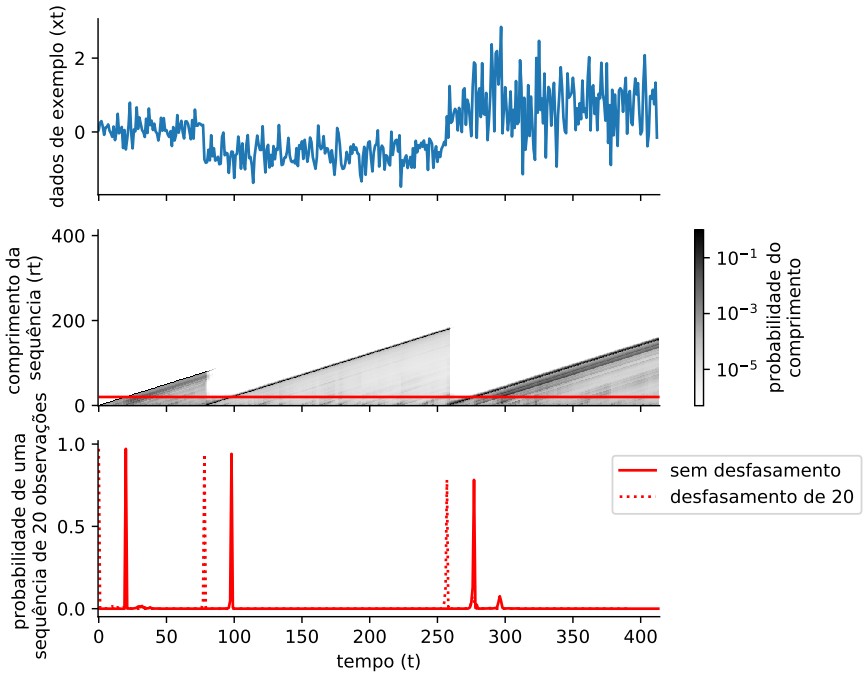


Figura 3.4: Exemplo do funcionamento do algoritmo BOCPD. No primeiro gráfico, é possível observar uma distribuição sintética de dados com dois pontos de quebra. No segundo gráfico, podemos observar o resultado do algoritmo após processar os dados, onde, para cada momento, se obtém uma distribuição de probabilidades do comprimento da sequência de dados. Finalmente, o último gráfico mostra uma parte do segundo gráfico, apresentando para todas as observações a probabilidade de pertencerem a uma sequência de tamanho 20.

F1 de 0.609; no entanto, é importante notar que, nesse conjunto de dados, o melhor algoritmo obteve um F1 de 0.670. Apesar disso, não é possível retirar conclusões sobre como o algoritmo se comportará num contexto distinto como a deteção remota.

3.4 Avaliação

Até ao momento foram referidas algumas métricas de avaliação sem a devida atenção ao seu significado. Esta secção apresenta com maior detalhe as métricas usadas no contexto de deteção de alterações. Em Burg e Williams [5] é referido que para avaliar segmentação temporal é possível usar dois pontos de vista: classificação binária ou *clustering*. O primeiro ponto de vista classifica cada observação como mudança ou não mudança. O segundo ponto de vista olha para as várias secções como geradas de forma distinta. Quando no contexto de deteção de alterações abruptas em florestas, o ponto de vista de classificação binária é o usado na totalidade dos estudos aqui apresentados [9, 27].

A tabela 3.1 mapeia os conceitos necessários para avaliar um classificador binário

Tabela 3.1: Classificação binária da perturbação de uma observação.

		Referência	
		Alterado	Não Alterado
Classificação	Alterado	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não Alterado	Falso Positivo (FP)	Verdadeiro Negativo (VN)

tendo em conta os conceitos específicos do problema de deteção de alterações.

A partir da tabela de confusão é possível calcular um grande conjunto de métricas de avaliação. Em Burg e Williams [5] é dito que, devido à grande disparidade de frequências entre alterações e não alterações, há um conjunto de métricas como a *accuracy* que acabam por estar enviesadas. Por este motivo, são preferidas métricas como *positive predictive rate*, *true positive rate* e F1. Ainda assim, alguns estudos reportam a *accuracy* [9, 27] complementando a mesma com outras métricas não sensíveis ao enviesamento que esta apresenta.

Voltando a atenção agora para as métricas, chega a altura de compreender melhor cada uma e perceber como se calculam. Começando pela *accuracy*, definida pela equação 3.1, esta métrica é intuitiva e simples, pois apenas divide o número de classificações corretas pelo número total de classificações. No entanto, como já foi referido, há métricas mais informativas no contexto de deteção de alterações.

$$accuracy = \frac{VP + VN}{VP + FP + FN + VN} \tag{3.1}$$

O False Negative Rate (FNR), também conhecido por miss rate ou erro de omissão, está definido na equação 3.2. Esta métrica divide o número de falsos negativos pelo número de alterações de referência, fornecendo a probabilidade de omitir um ponto de mudança. O FNR é o complementar do True Positive Rate (TPR), isto porque o TPR fornece a probabilidade de acertar num ponto de mudança, daí que $FNR = 1 - TPR$. Ambos os valores variam entre 0 e 1, sendo que um resultado perfeito seria $FNR = 0$ e $TPR = 1$.

$$FNR = \frac{FN}{VP + FN} \tag{3.2}$$

O False Discovery Rate (FDR), também conhecido por erro de comissão, está definido na equação 3.3. Esta métrica divide o número de falsos positivos pelo número de alterações classificadas, fornecendo a probabilidade de errar quando uma observação é classificada como alterada. Esta métrica também tem uma complementar, o Positive Predictive Value (PPV), dado que o PPV fornece a probabilidade de uma classificação de alteração estar certa, pelo que $FDR = 1 - PPV$. Novamente, ambos os valores variam entre 0 e 1, e um resultado perfeito seria $FDR = 0$ e $PPV = 1$.

$$FDR = \frac{FP}{VP + FP} \tag{3.3}$$

Finalmente, o F1 combina o PPV e o TPR numa única métrica, dando igual peso às duas métricas. O seu cálculo está definido na equação 3.4 e pode ser feito diretamente com os valores TRP e PPV ou, alternativamente, usando valores da matriz de confusão.

$$F1 = 2\frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \tag{3.4}$$

Ainda que mais métricas estejam disponíveis, as que foram apresentadas mostraram-se particularmente úteis no contexto de deteção de alterações.

3.5 Conclusão

Os primeiros estudos apresentados contam com várias abordagens para determinar a idade da floresta, todas elas usando dados relativos a um momento no tempo. Como já foi referido, esta abordagem é análoga a usar uma fotografia de uma pessoa para determinar a sua idade. Adicionalmente os dados necessários para as abordagens mais recentes e sofisticadas não estão disponíveis para o território português.

Posteriormente, foram apresentados estudos que usam séries temporais para determinar distúrbios na floresta. A partir da data destes distúrbios que devastam o ecossistema é trivial calcular a data da floresta presente no local. Os resultados encontrados para os algoritmos apresentados nesta secção são bastante díspares. Tal deve-se ao facto de usarem dados de referência distintos com diferentes tipos de alterações. Por este motivo, é sugerido que a melhor opção é comparar os diferentes algoritmos no contexto pretendido; só assim será possível escolher o melhor algoritmo para cada situação. Deste conjunto de estudos destaca-se o LandTrendr e o CCDC.

Para além dos estudos que lidam com séries temporais de deteção remota, foram também exploradas alternativas que lidassem de forma genérica com deteção de pontos de quebra em séries temporais. Nesta categoria, destacou-se o BOCPD como tendo bons resultados num diversificado conjunto de dados de referência.

Finalmente, relativamente à avaliação dos vários algoritmos, foi apresentado um conjunto de métricas que nos vários estudos se mostraram úteis para avaliar a prestação dos algoritmos de deteção de alterações. Destaque para o F1 como métrica singular para avaliar os resultados.

4

Abordagem e Metodologia

Neste momento já é possível: compreender o problema abordado pelo presente trabalho e o contexto em que se insere; dados disponíveis sobre a floresta portuguesa bem como as imagens de satélite; e, finalmente, estar a par com aquilo que a literatura indica como propostas para problemas semelhantes. Torna-se possível então começar a delinear uma estratégia para explorar a problemática da criação de um cadastro florestal nacional. O presente capítulo apresenta a abordagem tomada durante a investigação sem detalhes de implementação, descritos no capitulo 5.

A secção 4.1 explica o ponto de vista usado para determinar a idade da floresta e justifica a metodologia.

A secção 4.2 revela os passos necessários para processar os dados disponíveis sobre a floresta portuguesa e a forma como foram compatibilizados. Para além da ingestão, os dados são ainda transformados de forma a criar informação complementar sobre mudanças entre edições consecutivas do mesmo projeto. No final da secção, é apresentada a coleção de dados Landsat usada, bem como o processamento necessário para a obter.

A secção 4.3 descreve o primeiro contacto com a análise de séries temporais. É descrito o processo de criação de uma ferramenta de visualização de séries temporais com o propósito de melhor compreender o impacto que as alterações têm nas mesmas.

A secção 4.5 apresenta os algoritmos testados para deteção de alterações florestais. Para cada um são apresentados detalhes sobre o pré-processamento necessário, os conjuntos de parametrizações testadas, e finalmente o processamento necessário para extrair do output de cada algoritmo os pontos de quebra que assinalam alterações florestais.

A secção 4.4 explica como foi criado o conjunto de dados de referência usados para avaliar os algoritmos listados na secção 4.5. São apresentados os dados complementares escolhidos e a metodologia usada para gerar os pontos de referência e as alterações florestais associadas a cada um.

A secção 4.6 descreve as técnicas usadas para avaliar os algoritmos consoante o conjunto de dados de referência criado descrito na secção 4.4. O capítulo está dividido em duas partes, a primeira fala sobre métricas para avaliar o desempenho dos algoritmos, a segunda fala sobre as visualizações usadas para os avaliar e determinar pontos fortes e

pontos fracos de cada algoritmo.

Finalmente, a secção 4.7 faz um breve apanhado do trabalho feito, reforçando as ideias principais e as questões a avaliar no seguinte capítulo.

4.1 Reformulação do problema segundo o ponto de vista de deteção de alterações florestais

Relembrando, o problema inicial passa por conseguir determinar qual a idade de cada ponto de floresta no país. Para atingir este objetivo, e com base no que foi discutido em 3, decidiu-se utilizar algoritmos de deteção de alterações florestais, bem como algoritmos de deteção de mudança em séries temporais. Estes algoritmos, apresentados em 3.2 e em 3.3, usam séries temporais de forma a detetar alterações no comportamento dos pixeis ao longo do tempo: uns são especificamente talhados para deteção remota, enquanto outros foram desenvolvidos para aplicações mais genéricas de deteção de alterações noutros tipos de séries temporais. É possível inferir a idade da floresta assumindo que as alterações identificadas traduzem eventos como incêndios e desflorestações, os quais implicam uma nova floresta.

Para melhor compreender se é possível inferir a idade da floresta a partir das alterações detetadas, primeiro é importante perceber que tipo de alterações são detetadas. Em [9] são comparados vários algoritmos de deteção de alterações que são divididas em 5 grupos:

- 61,5% Abate - Corte de árvores
- 17,9% Declínio - Processo longo (mais do que 1 ano) de degradação da qualidade da floresta (ex: seca)
- 9,0% Outros - Alterações que não se enquadram nos outros grupos
- 6,2% Ventos - Ventos fortes (tornados) que derrubam árvores
- 5,4% Fogos - Fogo que afeta a área

As alterações apresentadas são sem dúvida úteis para determinar a idade da floresta, à exceção do declínio, que, pela descrição fornecida, não aparenta ser um fator determinante na idade da floresta.

Tendo tudo isto em conta, o presente trabalho foca-se na deteção de alterações florestais como forma de determinar a idade das mesmas.

4.2 Ingestão e processamento de dados florestais e de satélite

As fontes de dados descritas em 2.2 têm um papel determinante no trabalho, no entanto nem todas estavam completamente prontas para serem analisadas como séries

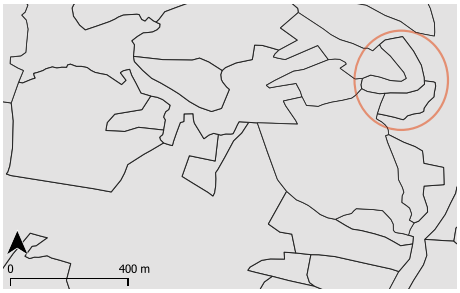
temporais. À parte do Corine Land Cover, todas necessitaram de algum tipo de processamento para facilitar a análise. A presente secção descreve os métodos utilizados, com especial atenção para a Carta de Uso e Ocupação do Solo e as parcelas de campo do Inventário Florestal Nacional 4, pois precisaram de uma maior atenção.

4.2.1 Carta de Uso e Ocupação do Solo

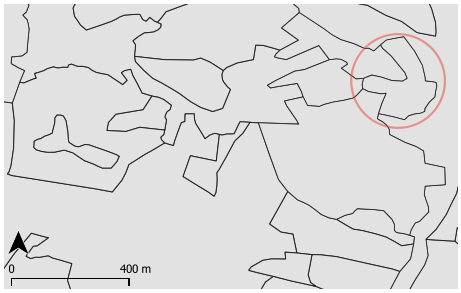
Os dados usados da COS provêm das edições de 1997, 2007, 2010 e 2015, não tendo sido usada a edição de 2018. Segundo [7], não era possível a comparação das outras edições com esta devido à diferença de nomenclaturas usadas. Apesar de serem anunciadas novas versões das edições anteriores compatíveis com a COS de 2018, até à data da escrita não foram ainda publicadas.

As edições usadas, no entanto, não deixam de trazer consigo alguns desafios devido às nomenclaturas. Relativamente às quatro edições usadas, há um total de três nomenclaturas: uma para 1997, outra para 2007 e 2010 e finalmente uma para 2015. Desta forma, apenas seria possível comparar as edições de 2007 e 2010, no entanto a DGT disponibiliza em [42] duas tabelas que tornam possível as restantes comparações. A primeira tabela fornece uma correspondência entre a nomenclatura 2007/2010 e a nomenclatura 1997, a segunda tabela fornece uma correspondência entre a nomenclatura 2007/2010 e a nomenclatura 2015. É importante salientar o sentido da correspondência fornecida: o ponto de partida para ambas as tabelas é a nomenclatura de 2007/2010. Podemos ver estas tabelas como duas funções que, dada a nomenclatura de 2007/2010, nos conseguem obter as outras nomenclaturas. No entanto, estas funções não são injetivas (diferentes códigos 2007/2010 podem ter o mesmo código na nomenclatura 2015 ou 1997), consequentemente, as funções não são invertíveis. Isto faz com que não seja possível passar da nomenclatura 1997 para 2007/2010 nem da nomenclatura 2015 para 2007/2010.

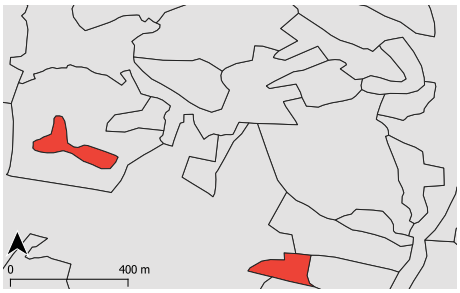
As alterações à COS são relativamente simples de calcular após compreender como funcionam as tabelas de correspondência. Realizando uma diferença entre polígonos que intersectam, e comparando os códigos de uso de duas edições consecutivas tendo em conta as correspondências, obtém-se as diferenças entre as duas edições. Em 4.1 é possível ver um exemplo do cálculo das diferenças, estando representadas três imagens da mesma zona: 4.1(a) mostra os polígonos que definem as várias zonas COS de 2010, 4.1(b) mostra os polígonos que definem as várias zonas COS de 2015 e, finalmente, 4.1(c) mostra os polígonos calculados da diferença entre as duas edições. Os polígonos a vermelho indicam as zonas onde alterações foram detetadas - por exemplo, no canto inferior direito o código passou de 1.3.3.00.0 (Áreas em construção) para 1.2.1.00.0 (Indústria, comércio e equipamentos gerais). Também é possível ver polígonos que são visualmente diferentes, no entanto, referem-se a códigos compatíveis. Por exemplo, no canto superior direito assinalado com um círculo vermelho, na edição 2015 um dos polígonos está dividido em 2 na edição 2010 , no entanto ambos os códigos de uso de solo de 2010 mapeiam para o mesmo código de 2015, pelo que nenhuma diferença é encontrada.



(a) COS 2010



(b) COS 2015



(c) Alterações COS 2010 2015

Figura 4.1: Figuras ilustrativas do ponto de partida e resultado da determinação das alterações entre edições COS. Localização: 38.88943, -9.25083 WGS 84

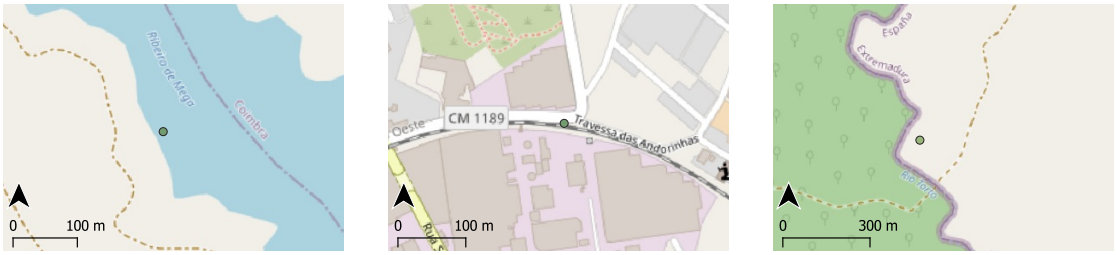
Importa portanto retirar que é possível comparar as várias edições COS e que desta comparação podemos salientar alterações do uso de solo que alteram a idade da floresta.

4.2.2 Parcelas de Campo do Inventário Florestal Nacional 4

Os dados das parcelas de campo aparentam ser os mais promissores, dado que contêm informação muito detalhada, inclusive a idade da floresta. No entanto, uma análise básica dos pontos revelou que o detalhe espacial dos dados era bastante baixo e incompatível com as imagens de satélite.

Os dados estavam disponibilizados num ficheiro de base de dados Access, pelo que foi necessário primeiro compreender a estrutura dessa base de dados. Os dados estavam divididos em várias tabelas; após compreender a organização, foram selecionados atributos que poderiam ter interesse. Finalmente, com base nestes atributos, foi criada uma query para extrair uma única tabela com toda a informação selecionada. Entre estes encontravam-se a classe da idade, a espécie de floresta presente no local e as coordenadas do ponto no sistema de coordenadas EPSG:20790. Os dados foram exportados para um ficheiro CSV e um Shapefile de forma a facilitar a sua análise.

A análise destes pontos revelou alguns problemas com a resolução espacial dos dados. É possível visualizar os três tipos de problemas encontrados em 4.2. Em primeiro lugar, dois pontos foram encontrados em corpos de água um deles pode ser visto em 4.2(a)- este ponto em particular encontra-se a 15 metros da margem definida pelo Open Street



(a) Ponto num rio (Coordenadas: 39.9647571, -8.1118110)

(b) Ponto numa cidade (Coordenadas: 39.7379517, -8.9208564)

(c) Ponto em Espanha (Coordenadas: 40.1642177, -7.0190643)

Figura 4.2: Exemplos de pontos que mostram problemas com a resolução espacial dos pontos de campo do inventário florestal nacional de 1995

Maps e o mesmo se confirma usando imagens de satélite pelo Google Maps. Em segundo lugar, foram encontrados 120 pontos em zonas que a COS 1995 classifica como Territórios artificializados. É possível visualizar em 4.2(b) um ponto no meio de uma zona urbana. Finalmente, foi encontrado um ponto que se situava em solo espanhol de acordo com o Open Street Maps e a COS 1995. Este ponto situa-se a uma distância de 60 metros da fronteira com Portugal.

O conjunto de dados é por isso limitado no que toca a precisão espacial, pelo que qualquer análise feita aos mesmos deve ter este aspeto em consideração.

4.2.3 Áreas Ardidas do Instituto da Conservação da Natureza e das Florestas

Os dados das áreas ardidas contam com a maior periodicidade de edições e com o maior historial. As informações relativas a cada área ardida variam consoante o ano em que ocorreu, com uma clara tendência para haver menos informação quanto mais antiga for a ocorrência. Após uma análise preliminar dos dados[43], e com o objetivo de abranger a totalidade do historial de áreas ardidas, foi possível extrair de todas: o seu identificador original, o ano do incêndio e um polígono georreferenciado que define a sua extensão. Adicionalmente, para áreas ardidas após 2012, foi possível extrair, para além do ano, o dia e o mês do início do incêndio que gerou a área ardida para um conjunto dos registos existentes. A tabela 4.1 mostra a distribuição da percentagem de datas nulas a partir de 2012.

4.2.4 Fotopontos do Inventário Florestal nacional

Como já foi referido em 2.2.1, para além das parcelas de campo, o IFN conta também com fotopontos obtidos a partir de fotointerpretação de imagens de satélite. Os pontos encontram-se perfeitamente alinhados entre edições do inventário, o que simplifica o cálculo das alterações entre edições. Este foi o único processamento realizado.

Tabela 4.1: Percentagem de datas nulas presentes no conjunto de dados de áreas ardidas de Portugal fornecidas pelo ICNF

ano	nº Registos	nº datas nulas	% datas nulas
2012	2979	1859	62.4
2013	3400	2082	61.2
2014	1141	200	17.5
2015	1653	198	12.0
2016	2846	358	12.6
2017	2800	184	6.6
2018	546	43	7.9

4.2.5 Imagens de Satélite Landsat

Finalmente, é necessário abordar as imagens de satélite Landsat utilizadas e do processamento realizado. Foi construída uma coleção de imagens no Google Earth Engine (GEE) contendo todas as imagens Landsat desde a missão 4 até à 8. As coleções de imagens usadas são todas da Tier 1 (imagens com a melhor qualidade), de forma a diminuir o ruído das séries temporais. Foi usado o pré-processamento que fornece refletância no topo da atmosfera para diminuir as irregularidades entre imagens existentes quando se usam os valores diretos do sensor e para não introduzir enviesamento resultantes do processamento da refletância à superfície devido aos diferentes sensores [48].

Após estas considerações da escolha de fontes de imagens de satélite, para gerar uma só coleção de imagens foram necessárias duas coisas: compatibilizar a meta-informação associada a cada imagem e compatibilizar as bandas usadas.

Cada conjunto de imagens fornece um conjunto de meta-informação distinta pelo que foi necessário comparar os dados fornecidos por cada missão e selecionar os dados comuns a todas. Desta análise sobraram 47 propriedades, como por exemplo: o identificador do satélite, data, informação relativa a localização, informação relativa à qualidade da imagem entre outras.

Relativamente às bandas das imagens, os nomes originais referiam-se ao número da banda que era diferente entre as missões. Passou a usar-se ,por isso, a descrição da banda como nome. Por exemplo a banda "B1" da missão 5 e a banda "B2" da missão 8 passaram a chamar-se "Blue" por capturarem informação nesta zona do espectro eletromagnético. Esta nomenclatura facilita o cálculo de índices espectrais entre as várias missões.

Finalmente, foram extraídos indicadores de qualidade que estavam na banda de qualidade das respetivas missões. Deste processo, resultaram três bandas que indicam a presença de neve, sombra e nuvens. Estas bandas são úteis para remover estas contaminações que, na sua maioria, não são desejados durante a análise de imagens.

4.3 Impacto de alterações abruptas em índices de deteção remota

A primeira tarefa realizada passou pela construção de visualizações de séries temporais. Por um lado, a visualização de séries temporais permite uma primeira compreensão intuitiva do comportamento de vários índices perante vários tipos de alterações bem como períodos de relativa estabilidade. Por outro lado, os conhecimentos adquiridos durante a construção destas visualizações serviram como base para a criação de outras para ajudar na interpretação dos resultados dos algoritmos de deteção de mudanças.

O primeiro passo para criar esta visualização foi selecionar zonas para análise. Foram selecionados todos os polígonos COS95 que intersetavam os dados de campo do IFN95. A partir destes polígonos foram geradas séries temporais, que de todos os pixeis que intersetavam o polígono são extraídas algumas medições. Por outras palavras, deste processamento resulta uma tabela com uma linha por cada polígono COS por cada imagem de satélite. As medições extraídas são: NDVI (máximo, mínimo, média e desvio padrão), NBR (máximo, mínimo, média e desvio padrão), total de pixeis processados, satélite usado, identificador da imagem.

Tendo acesso a estes dados foi construido o dashboard possível de visualizar na imagem 4.3. Esta visualização interativa mostra dois índices NBR e NDVI. Para cada um desses índices estão presentes dois gráficos, os gráficos com uma linha mostram a média do índice enquanto que os gráficos de barras mostram a diferença entre duas observações consecutivas. A cor destes gráficos representa a percentagem de pixeis usados em cada observação. Diretamente abaixo destes gráficos é possível observar o uso de solo COS 1995 e 2007. Do lado direito, no canto inferior é possível observar a lista de incêndios que tiveram impacto no polígono. Do lado direito, na parte superior estão um conjunto de filtros para a visualização. Em primeiro lugar é possível selecionar diferentes polígonos COS. Seguidamente o intervalo de datas sobre as quais apresentar as séries temporais. Posteriormente dois filtros sobre as observações, um define um número mínimo de pixeis enquanto que outro define uma percentagem mínima de pixeis. Após a escala de cores da percentagem de pixeis observados é possível ver um controlo sobre a média móvel aplicada nas séries temporais. Finalmente os últimos dois controlos definem um limite mínimo da diferença dos dois índices para ser apresentados no gráfico de barras.

Olhando agora com um espírito crítico para o gráfico para perceber que tipo de informação se pode retirar. É possível observar em ambos os índices uma grande quebra antes de 2002 . Quando reparamos no uso de solo é possível observar que houve uma alteração, a floresta de pinheiro bravo existente passou a ser uma floresta de eucalipto ente os anos de 1995 e 2007. Adicionalmente na lista de incêndios conta um incêndio em 2000.

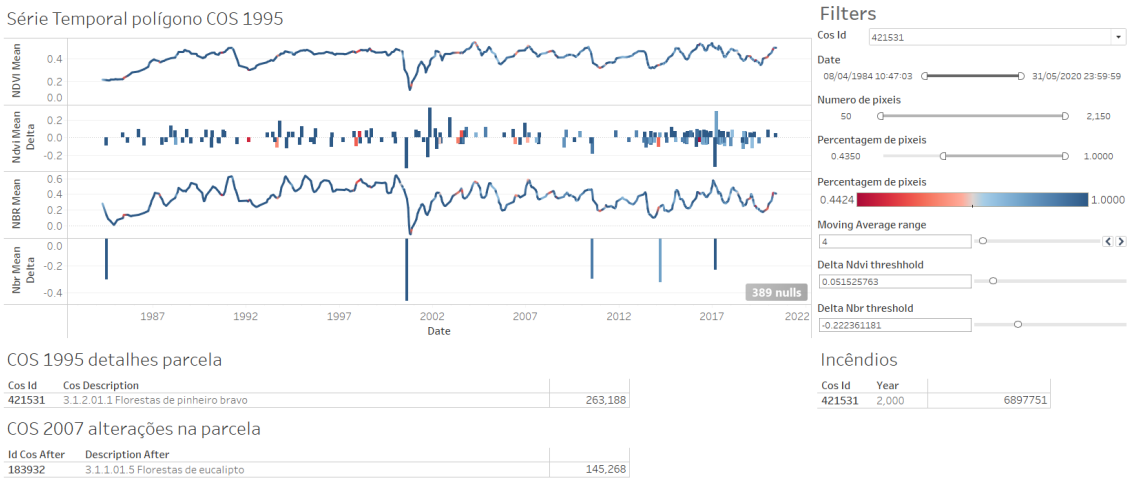


Figura 4.3: Dashboard construído para visualizar o impacto de alterações abruptas em índices de deteção remota. É possível visualizar duas linhas que representam dois índices (NBR e NDVI). Por baixo de cada uma das linhas está representado a diferença entre duas observações consecutivas. Do lado direito estão presentes um conjunto de controlos sobre o dashboard.

4.4 Construção do conjunto de dados de referência

Para criar o conjunto de dados de referência, foi decidido que seria necessária uma avaliação manual de um conjunto de pontos. Esta decisão deriva de três principais fatores: falta de detalhe temporal, conjunto não representativo de alterações e falta de momentos de estabilidade.

Relativamente à falta de detalhe temporal: apesar de estar à disposição um conjunto de dados relativos à ocupação do solo, estes contam com largos anos de intervalo entre edições mesmo tendo em conta várias fontes de informação. Não sendo assim possível atingir a resolução de um ano atingida pelos algoritmos, o melhor que se poderia dizer é que entre o ano X e o ano Y houve uma alteração no uso de solo.

Relativamente aos tipos de alterações que estão presentes nos conjuntos de dados, estes não têm uma distribuição representativa dos tipos de alterações possíveis na floresta. Isto porque nenhum dos conjuntos de dados contém informação sobre o abate de árvores com sucessiva regeneração da floresta. Ainda que não haja certeza de qual a percentagem de alterações que este tipo de alteração representa, é certamente superior a zero.

Em último lugar, no que toca aos momentos de estabilidade, nenhum conjunto de dados fornece informações sobre momentos em que zonas florestais não sofrem qualquer tipo de alterações. Não havendo uma lista exaustiva de todas as alterações, é possível que um ponto onde nenhuma alteração foi detetada na realidade tenha alterações.

A avaliação manual já foi usada em [9, 27, 5], pelo que se enquadra perfeitamente no problema em questão. Em dois dos artigos em questão, é utilizado o TimeSync [8], uma ferramenta desenvolvida para avaliar séries temporais e classificar alterações; o outro

usa uma ferramenta própria descrita no artigo. No entanto, para o presente trabalho foi utilizado um script construído no Google Earth Engine que disponibiliza ferramentas de visualização equivalentes às utilizadas no TimeSync. As observações para cada ponto foram anotadas manualmente num ficheiro excel.

4.4.1 Seleção dos pontos de referência

A primeira parte para criar um conjunto de referência passou por selecionar um conjunto de pontos a serem analisados manualmente por um interpretador. O número de pontos tem de ser estatisticamente significativo para poder comparar os vários algoritmos e, ao mesmo tempo, possível de ser analisado por um interpretador em tempo útil. Adicionalmente, os pontos devem ser representativos da totalidade da área de estudo, no caso em concreto Portugal, não só em termos da diversidade das espécies de floresta existentes mas também da diversidade de zonas climáticas. É importante também dividir o conjunto em teste e validação para não obter resultados sobreajustados ao ajuste de parâmetros.

Com o objetivo de integrar o maior número de fontes de informação, decidiu usar-se um subconjunto dos pontos do inventário florestal como pontos para serem avaliados. Esta escolha deveu-se ao facto de os fotopontos do IFN serem a única fonte de informação que utiliza pontos, facilitando assim a integração com as outras através de uma simples interseção entre os pontos e os polígonos das outras fontes de informação.

Partindo dos pontos do IFN 2015 e de um mapa de regiões climáticas, foi realizada uma amostragem estratificada tendo em conta a espécie indicada no IFN e a zona climática de cada ponto. Foram selecionados na totalidade 998 pontos dos quais 662 se destinam a validação e 336 para teste.

4.4.2 Adição de informações relevantes aos pontos de referência

Tendo acesso a um conjunto de pontos a ser analisados, foi possível realizar a interseção dos mesmos com os conjuntos de dados disponíveis. Para cada conjunto de dados foi realizada a interseção com o ponto e extraídas informações relevantes para cada uma. Segue-se a lista de informações retiradas de cada conjunto de dados:

- **Área Ardida** - id, ano
- **COS** - id, ano, uso do solo
- **CLC** - id, ano, uso do solo
- **IFN** - id, ano, uso do solo, ocupação principal, tipo de povoamento
- **Regiões Climáticas** - distrito, concelho, região climática

Relembrado que, à exceção das Regiões Climáticas, os restantes conjuntos de dados contêm polígonos que se sobrepõem representando diferentes anos, foi extraída uma lista de atributos para estes.

4.4.3 Metodologia de análise de um ponto

Para interpretar devidamente a história de um dado ponto são necessárias várias ferramentas. Em primeiro lugar, é preciso visualizar a série temporal do ponto em questão; para além disso, é necessário algum contexto para o interpretador, pelo que uma imagem de satélite do local é importante para melhor compreender o contexto onde o ponto se enquadra. Finalmente, é importante ter acesso a outro tipo de dados como incêndios e classificações do uso de solo para ajudar em caso de dúvida e confirmar observações da série temporal. Tendo isto em conta, para a interpretação de cada ponto foram usadas duas ferramentas: script de visualização de séries temporais no Google Earth Engine e o Google Earth Pro. O script de visualização contém a totalidade das ferramentas descritas; o Google Earth Pro é utilizado quando é necessário mais contexto por disponibilizar imagens de alta resolução ao longo de alguns anos.

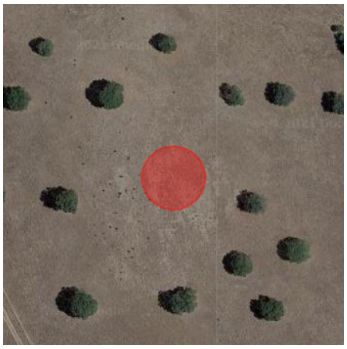
O script de visualização de séries temporais produz um gráfico interativo para um dado ponto de referência; adicionalmente, apresenta também um conjunto de dados associados ao ponto em questão, e, por último, uma imagem de satélite da localização do ponto.

Para analisar um ponto são necessários os seguintes passos:

- Validar a área em volta do ponto
- Selecionar os anos candidatos a pontos de quebra com base na visualização da série temporal
- Cruzar informação com fontes de dados auxiliares
- Usar imagens de satélite para confirmar pontos candidatos em dúvida
- Anotar os resultados da análise

O primeiro passo consiste em confirmar se a área em torno do ponto é fiável para análise. A área em torno do ponto tem de cumprir três critérios: não ser uma zona artificial, não ter uma baixa densidade de árvores, e, finalmente, ser uma zona de floresta uniforme. Este passo é necessário para confirmar que apenas pontos de floresta úteis para análise são avaliados. Relativamente ao primeiro critério, foram descartados todos os pontos que se encontravam numa zona artificializada, como estradas e edifícios. É possível ver um exemplo de um destes casos em 4.4(a). O segundo critério refere-se a pontos que designam lugares com muito baixa densidade de árvores, 4.4(b) é um exemplo de um ponto descartado devido à baixa densidade de árvores presentes no local. Finalmente, os pontos que em seu redor não contam com uma zona uniforme de floresta também são descartados. Estes pontos normalmente encontram-se na fronteira de duas parcelas distintas de floresta. É possível ver um exemplo deste caso em 4.4(c). Apesar de todos os critérios serem cumpridos, há ainda a possibilidade de o ponto ser invalidado se a análise da série temporal for inconclusiva ou de difícil análise.

(a) Área artificializada Localização: -8.214694, 37.130057 WGS84



(b) Baixa densidade de árvores Localização: -7.159203, 39.094502 WGS84

(c) Transição de parcelas Localização: -8.773462, 40.542304 WGS84

Figura 4.4: Exemplos de pontos três pontos de referência invalidados

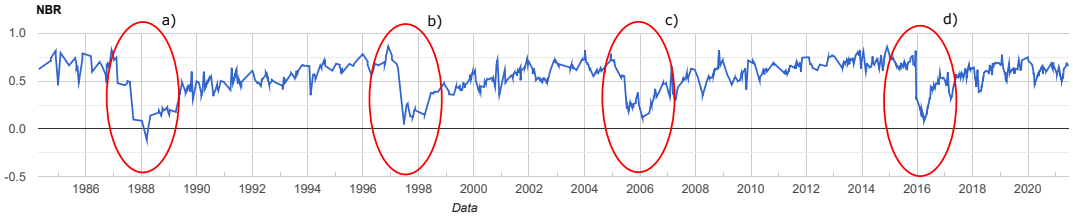


Figura 4.5: Série temporal de uma zona com quatro alterações salientadas.

Estando confirmada a validade da zona de análise, o passo seguinte passa por analisar a série temporal. Uma primeira análise visual permite selecionar candidatos a pontos de quebra. Olhando para o exemplo 4.5 é possível observar claramente quatro quebras abruptas salientadas com elipses. Fora destas elipses a tendência do NBR é de subida ou estabilidade; no entanto, nos lugares assinalados, é possível observar mudanças abruptas e de grande magnitude. Uma análise mais cuidada permite atribuir o ano em que cada alteração ocorreu, a) 1987, b) 1997, c) 2005, d) 2015. Neste caso os pontos são bem claros, no entanto, em situações menos claras, alterações em dúvida deve ser anotadas para serem analisadas com mais detalhe nas fases seguintes.

Tendo um conjunto de possíveis alterações, é altura de cruzar esta informação com informações de outros conjuntos de dados para reforçar a confiança nos resultados e eliminar dúvidas que tenham surgido. Neste caso em concreto, há registo de dois incêndios na área, um em 1987 e outro em 2005 o que está de acordo com duas alterações encontradas, sendo assim possível atribuir-lhes uma causa. Quer o COS quer o IFN afirmam que a zona em questão ao longo do tempo se tem mantido como uma floresta de eucalipto, pelo que as alterações não se devem a alteração do uso do solo. Resta ainda confirmar imagens de satélite: por um lado, o script do GEE permite clicar em qualquer ponto da série temporal e visualizar a imagem Landsat; por outro lado, no Google Earth Pro é possível visualizar imagens com maior resolução. A observação destas imagens permitiu

concluir que a alteração d) se deve a um corte. No entanto, não havendo imagens de alta resolução disponíveis para a alteração b), não foi possível garantir que se tratava de um corte, apesar de ser o mais provável por se tratar de uma zona de eucaliptos dado que esta espécie costuma ser alvo frequente de cortes.

Finalmente, é necessário anotar a informação recolhida durante a análise. Quando um ponto é válido são recolhidos os anos das alterações, que variam entre 1985 e 2020, e a causa associada, que pode ser uma entre abate, fogo, desconhecido e outro. As causas abate e fogo são auto explicativas, outro refere-se a outras causas como, por exemplo, ventos fortes, e desconhecido refere-se a uma causa desconhecida, tendo sido usado quando não havia certeza sobre o tipo de causa que deu origem à alteração.

4.5 Algoritmos de deteção de alterações florestais e deteção de pontos de mudança

Os vários algoritmos têm as suas diferenças, como já explicado em 3.2, e há duas importantes a ter em conta: a sensibilidade e a resolução temporal.

Relativamente à sensibilidade, em [9] é dito que certos algoritmos têm a capacidade de detetar alterações subtis a longo prazo, como secas ou doenças. Outros são destinados para detetarem eventos pontuais e de grande alteração, como incêndios ou abates florestais. Para ser possível determinar a idade das florestas, é necessário que apenas os eventos que alterem a idade da mesma sejam detetados. É importante, por isso, ajustar os parâmetros dos vários algoritmos para este tipo de aplicação.

Relativamente à resolução temporal dos algoritmos, há alguns que apenas detetam o ano da alteração, enquanto que outros são mais precisos e indicam a data da imagem onde detetam a alteração. Para tornar viável a comparação entre os vários algoritmos, a unidade de comparação será o ano. Esta também é a abordagem tomada em [9].

Tendo acesso ao catálogo de imagens Landsat descrito em 4.2.4, foi possível correr 3 algoritmos: LandTrend e CCDC (apropriados para detetar alterações florestais), bem como Bayesian Online Changepoint Detection (algoritmo mais genérico na deteção de quebras em séries temporais). O LandTrend e CCDC foram escolhidos por serem bons candidatos para a deteção de pontos de mudança no comportamento de florestas [9] , por estarem disponíveis no GEE e ainda por serem estáveis na plataforma, isto é, não ocorrerem erros durante a sua execução. O BOCPD foi escolhido por terem um bom desempenho numa variedade de séries temporais em diferentes contextos [5], e ainda por estar facilmente disponível numa biblioteca Python [31].

Tendo isto em conta, chega a altura de compreender as especificidades de cada algoritmo; nomeadamente, se é necessário pré-processamento adicional, quais os parâmetros ajustados durante as experiências e o processamento adicional do output para um formato comum a todos algoritmos.

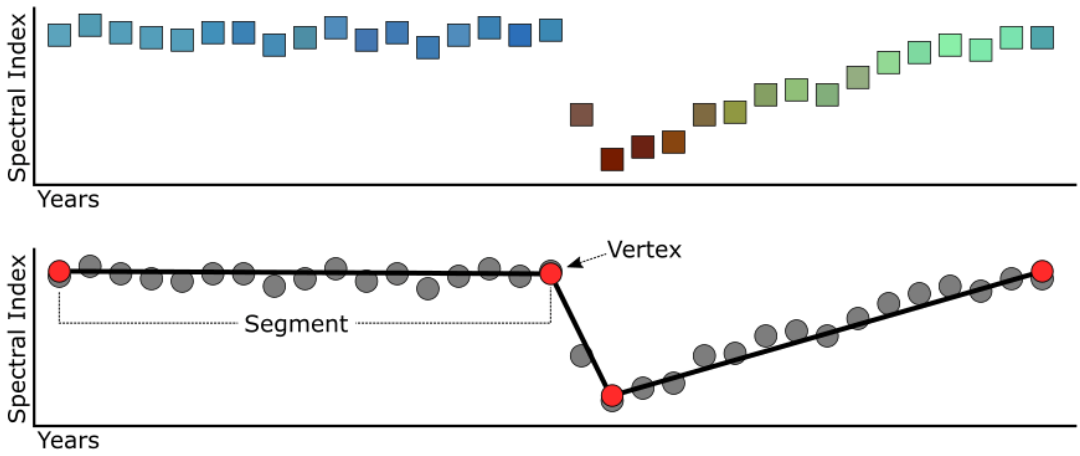


Figura 4.6: Representação do input do LandTrendr e do seu resultado, os vértices salientados são os potências pontos de interesse. Imagem retirada de: https://emapr.github.io/LT-GEE/landtrendr.html

4.5.1 LandTrendr

O LandTrender contém algumas particularidades, como já explicado em 3.2.2, focando primeiro no pré-processamento.

Relembrando que este algoritmo usa apenas uma imagem por cada ano, durante o pré-processamento foi necessário converter o conjunto de imagens de um ano para uma só imagem. Esta conversão foi realizada segundo as instruções dadas em [28], que consiste em usar pixeis válidos de imagens entre meio de julho e final de agosto. Usando múltiplas imagens é possível criar uma única imagem composta sem nuvens, sombras e outro tipo de contaminações indesejados. O último passo de pré-processamento consiste em calcular um índice radiométrico; no artigo que descreve este algoritmo são sugeridos o NDVI e o NBR, e em [9] apenas é testado o NBR pelo que se decidiu apenas usar este índice.

Relativamente aos parâmetros do algoritmo foi realizada uma pesquisa completa dos parâmetros descritos na tabela 4.2. Os valores escolhidos são os mesmos testados em [27], o artigo onde o algoritmo é apresentado. Os parâmetros testados são três para permitir executar os testes em tempo útil. Foram escolhidos os três parâmetros em questão por estarem disponíveis na implementação oferecida pelo GEE e por se revelarem importantes na avaliação realizada pelos autores.

O último detalhe importante relativo a este algoritmo está relacionado com o resultado, que conta com um conjunto de segmentos de reta. Dado que o interesse é na análise de alterações, a informação importante deste resultado não são os segmentos em si, mas sim as transições entre os mesmos e o ano associado. É possível visualizar na imagem 4.6 o resultado do algoritmo, e os pontos de interesse a laranja.

Tabela 4.2: Parâmetros e valores testados na pesquisa exaustiva para otimização do LandTrendr

Parâmetro	Valores
maxSegments	4, 5, 6
recoveryThreshold	1, 0.5, 0.25
pvalThreshold	0.05, 0.1, 0.2

Tabela 4.3: Parâmetros e valores testados na pesquisa exaustiva para otimização do CCDC

Parâmetro	Valores
minObservations	4, 6, 8
chiSquareProbability	0.985, 0.990, 0.995
minNumOfYearsScaler	1.33, 2.00

4.5.2 Continuous Change Detection and Classification

O CCDC é um algoritmo que não requer tanta atenção quer a nível de pré-processamento quer a nível de formatação do resultado. Relativamente ao pré-processamento, o algoritmo tem como input o conjunto de imagens de satélite descrito em 4.2.5. O output também é bastante próximo do desejado, contando com um conjunto de datas onde foram detetadas alterações. Foi apenas necessário das datas extrair o ano e remover eventuais repetições do mesmo ano.

Finalmente, no que toca à parametrização deste algoritmo, foram usados os parâmetros e valores definidos na tabela 4.3. Para os valores usados, devido à falta de valores de referência, foram usados os valores pré-definidos pela implementação do GEE com ligeiras variações. Por exemplo, o parâmetro minObservations, que define o número de observações irregulares necessárias para detetar uma alteração, tinha como valor pré-definido 6. Foram adicionados os valores 4 e 8 para compreender se seria necessário um valor mais baixo ou mais alto neste parâmetro, e compreender qual a importância deste parâmetro quando comparado com os outros.

4.5.3 BOCPD

O algoritmo BOCPD não é um algoritmo desenvolvido especificamente para o propósito de detetar alterações florestais, pelo que fornece uma maior flexibilidade. A primeira decisão importante acaba por ser: qual o input a fornecer ao algoritmo? À semelhança do que é realizado em algoritmos de deteção de alterações florestais, a utilização de índices radiométricos será uma boa opção. Não sabendo qual o índice mais proveitoso, optou-se por analisar duas alternativas: NDVI e NBR, devido à sua utilização no LandTrender. Relativamente ao pré-processamento, este algoritmo necessita de uma série temporal contínua e com intervalos de tempo regulares entre observações. Devido a um conjunto de fatores como nuvens, as séries temporais resultantes de imagens de satélite acabam

por ter algumas lacunas. Desta forma, foi necessário fazer um *resample* da série temporal e preencher eventuais lacunas. Não havendo certeza no *resample*, foram testadas as seguintes hipóteses:

- Sem Resample
- Resample 16 dias
- Resample 24 dias
- Resample 32 dias

Para estabelecer uma base de comparação, foi testada a opção de não realizar qualquer tipo de *resample*. Esta abordagem conta com o problema da frequência irregular de imagens de satélite, quer devido ao número de satélites em órbita, quer devido à falta de imagens devido a fenómenos atmosféricos. Foi testado *resample* de 16 dias por ser a resolução temporal de um satélite apenas, bem como 32 dias por ser o dobro dessa resolução, poupando assim a quantidade de dados a processar pelo algoritmo. Finalmente, dado que uma grande porção do catálogo Landsat conta com 2 satélites em simultâneo, o que resulta numa resolução temporal de 8 dias, foi também testado um *resample* de 24 dias por ser um múltiplo de 8.

Para além desta abordagem mais similar ao CCDC tirando proveito da resolução, foi testado também um caso com bastantes menos dados usando apenas uma observação por ano, mais similar à abordagem do LandTrendr. Neste caso foram testadas duas opções: mínimo anual, e o máximo anual entre os meses de outubro e dezembro. Esta diferença é muito marcante e até leva a parametrizações diferentes e mais à frente até resultados bastante diferentes. Por este motivo estas duas abordagens serão tratadas com nomes distintos, para a abordagem de maior resolução será usado o nome original de BOCPD, para a abordagem que usa uma resolução de 1 ano será usado o nome BOCPD Y.

Para obter resultados a partir das distribuições de probabilidade resultantes do algoritmo foi necessário escolher um parâmetro kw que indica o comprimento de sequência que procuramos. Isto corresponde a dar kw observações ao algoritmo para ele se determinar se houve ou não alteração. Valores de kw baixos levam a uma maior incerteza isto porque, a certeza é maior à medida que mais observações vão chegando após o ponto de quebra.

Olhando agora para a parametrização, o algoritmo requer dois parâmetros: hazard function e observation likelihood. A hazard function usada foi: $H(r) = \frac{1}{lam}$ e tal como é possível observar por sua vez requer parametrizarão do parâmetro lam. Relativamente ao observation likelihood este foi definido como uma distribuição t de student com parâmetros alpha e beta.

Tendo isto em conta as parametrizações testadas no BOCPD estão listadas na tabela 4.4. Os parâmetros escolhidos são baseados no exemplo fornecido em [26] pelo autor da biblioteca que implementa o BOCPD tendo sido usada uma hazard function constante.

Tabela 4.4: Parâmetros e valores testados na pesquisa exaustiva para otimização do BOCPD com maior resolução temporal

Parâmetro	Valores
índice radiométrico	NBR, NDVI
resample(dias)	nenhum, 16, 24, 32
lam	50, 100, 150, 200, 250
kw	10, 15, 20
alpha	0.1, 0.5
beta	0.01, 0.05

Tabela 4.5: Parâmetros e valores testados na pesquisa exaustiva para otimização do BOCPD com resolução temporal anual

Parâmetro	Valores
índice radiométrico	NDVI
redução	mínimo, máximo(inverno)
lam	6, 10, 12, 16
kw	1, 2, 3, 4
alpha	0.1, 0.5
beta	0.01, 0.05

As parametrizações testadas no BOCPD Y, estão listadas na tabela 4.5 e foram também elas derivadas da parametrização do exemplo fornecido em [26].

Relativamente ao processamento do output, foi necessário converter as probabilidades de quebra em pontos exatos de quebra, pelo que foram extraídos os picos de probabilidade. Das datas relativas a esses picos, foi extraído o ano e, novamente, removidos duplicados, tal como para o CCDC.

4.6 Metodologia de avaliação

A presente secção entra em detalhe sobre a metodologia de avaliação dos resultados obtidos pelos algoritmos de deteção de alterações. Para realizar esta avaliação, três principais questões se levantam:

- Como se comparam os algoritmos entre si?
- Para um mesmo algoritmo como é que as várias parametrizações afetam o desempenho?
- Quais os tipos de alterações difíceis de detetar?

Para ajudar a responder a estas questões foi criado o conjunto de dados de referência descrito em 4.4. Relembrando, neste conjunto de dados, para cada ponto geográfico, estão associadas as alterações que o afetaram. Estas alterações indicam o ano da sua ocorrência

e a causa. Dado que estes dados são tomados como referência e as alterações estão listadas de forma exaustiva, segue-se que os anos não presentes na listagem são anos onde não há qualquer tipo de alteração. Os dados, como já foi referido, estão divididos em dois grupos validação e teste. Isto deve-se ao facto de os algoritmos não usarem o treino prévio de modelos como acontece em algoritmos de aprendizagem automática.

Para avaliar a prestação destes algoritmos serão utilizadas algumas métricas já estabelecidas na literatura, com especial atenção para as que melhor descrevem a deteção de mudanças. Posteriormente, serão apresentadas as visualizações que usam as métricas indicadas e como estas ajudam a responder às questões condutoras da avaliação.

4.6.1 Métricas

Como já foi explicado em 4.1 estamos perante um problema onde, para um dado ponto geográfico de um dado ano, temos a classificação de alterado ou não alterado. Este ponto de vista torna-o num problema de classificação binária; consequentemente, é possível utilizar as métricas já exploradas na secção 3.4 para avaliar este tipo de classificação.

Há um detalhe a ter em conta: a ocorrência de uma alteração é muito mais rara quando comparada com a estabilidade. Por outras palavras, é mais provável que uma dada parcela não sofra qualquer tipo de alterações, o que leva a um conjunto de dados assimétrico. Os dados de referência apontam que, para uma observação aleatória, há uma probabilidade de 3% de ser uma alteração. Esta assimetria leva a que métricas que tenham em conta o número de verdadeiros negativos sejam muito inflacionadas e dificultem a avaliação. Vejamos o caso da accuracy; para um classificador ingénuo que classifique qualquer observação como não alteração, usando uma amostra de 100 observações de entre as quais 3 são alterações, iremos obter 97 verdadeiros negativos e 0 verdadeiros positivos e consequentemente uma accuracy de 97%. Tendo isto em conta, as métricas mais apropriadas são: false negative rate, false discovery rate, true positive rate, positive predictive value, f1 score.

Durante uma análise preliminar dos resultados foi possível observar que havia um número significativo de deteções de alterações com um ano de diferença em relação à referência. Ainda que estas deteções não sejam completamente precisas, foi decidido avaliar os resultados permitindo uma margem de erro de um ano. É certo que com esta margem de erro, para uma dada classificação, ou os resultados se mantêm iguais ou melhoram. No entanto, comparar duas classificações distintas tendo em conta a margem de erro pode oferecer uma visão diferente do desempenho das duas classificações. A margem de erro de um ano aparenta ser aceitável para a tarefa em questão por ser a menor margem que se pode dar. Este tipo de margem de erro é mencionado em [5] como útil na avaliação de deteção de pontos de alterações em séries temporais, desde que a margem usada seja adequada.

4.7 Conclusão

De momento já é possível ter uma ideia completa da abordagem do presente trabalho.

Após o problema inicial de detetar a idade da floresta ter sido relembrado, foi proposta uma solução: usar algoritmos de deteção de alterações florestais e alterações em séries temporais para detetar eventos que desbastam a floresta. Tendo acesso a estes pontos a idade da floresta é calculada com base no evento mais recente. Na lista de possíveis eventos encontram-se fogos, abate, desenvolvimento entre outros.

Seguidamente foi explicado como as fontes de dados foram processadas de modo a serem mais facilmente comparadas entre elas e consequentemente fornecer uma imagem mais completa da floresta. Para o COS foi necessário ter em conta as alterações de nomenclatura para uma comparação correta entre edições. As parcelas de campo do IFN4 revelaram alguns problemas relativos a precisão, tendo sido encontrados pontos em cidades, lagos e até em Espanha. Relativamente às áreas ardidas do ICNF, a variedade de formatos e de informação disponibilizada requereu alguma atenção no que toca a uniformizar e retirar apenas informação comum ao longo das várias edições. Os fotopontos do INF foram simples de processar por as várias edições serem muito similares. Finalmente foi ainda descrito como foi criado o conjunto de imagens de satélite a analisar da coleção Landsat. Mais concretamente foram usadas imagens das missões 4,5,7 e 8, com a melhor qualidade (Tier I) e com pré-processamento de forma a obter refletância no topo da atmosfera.

No que diz respeito à análise preliminar sobre o impacto de alterações abruptas em índices de deteção remota, foi possível desenvolver um conjunto de competências úteis para o restante trabalho. Entre estas competências salienta-se a extração de séries temporais e criação de visualizações. Adicionalmente o exemplo fornecido também ofereceu uma ideia sobre a importância de correlacionar várias fontes de informação para melhor compreender o historial da floresta.

Relativamente aos algoritmos foram explicadas as razões da escolha de cada um, que estão relacionadas com o desempenho dos algoritmos noutros estudos e a disponibilidade de implementações. Cada algoritmo conta também com as suas especificidades como o tipo de pré-processamento necessário, formatação do output e quais as parametrizações testadas. Sendo que o algoritmo BOCPD necessitara de uma maior atenção por ser uma algoritmo genérico de deteção de pontos de mudança em séries temporais. Importante referir que o output de todos estes algoritmos para uma série temporal é um conjunto de anos onde há quebras na série temporal. Este formato comum facilita a comparação entre os vários algoritmos.

Relativamente aos algoritmos utilizados, foram selecionados 4 algoritmos: CCDC, LandTrendr e BOCPD. Os dois primeiros desenvolvidos especificamente para deteção de alterações no contexto de deteção remota e os dois últimos por se destacarem num

conjunto variado de deteção de pontos de quebra. Cada algoritmo precisa de um pré-processamento específico de modo a formatar os dados de entrada para o formato esperado, bem como pós-processamento para todos extraírem dados no mesmo formato e facilitar a sua análise. É ainda importante referir que para cada algoritmo foi preciso testar a sua parametrização.

Para avaliar os algoritmos testados é necessário um conjunto de dados de referência, e esse foi mais um dos temas abordados durante este capítulo. A falta de fontes de dados que pudessem fornecer informações sobre a idade das florestas, levou à construção de raiz de um conjunto de dados sobre alterações abruptas na floresta. Os dados foram gerados através da interpretação de séries temporais Landsat, imagens de satélite e dados adicionais sobre a floresta. Este conjunto de dados conta com 664 pontos georreferenciados selecionados a partir de uma amostragem estratificada por zona climática e por espécie de árvores presente no local. Para cada ponto estão listadas as alterações caso estas existam bem como a sua causa.

Finalmente relativamente à metodologia de avaliação é importante relembrar que devido ao ponto de vista usado estamos perante um problema de classificação binária onde todos os anos são classificados como alteração ou não alteração. Por este motivo é possível utilizar métricas como o F1, False Negative Rate e False Discovery Rate.

5

Implementação

O presente capítulo apresenta detalhes de implementação do trabalho desenvolvido. Está dividido em 3 secções que apresentam três partes importantes da implementação: 5.1 armazenamento e gestão de dados, 5.2 processamento e 5.3 visualização de dados. Em seguida são apresentados os desafios encontrados em cada uma destas partes e as ferramentas que permitiram ultrapassar estes desafios.

5.1 Armazenamento e gestão de dados

Relativamente aos dados usados no trabalho, como já foi referido no capítulo 2, estes dividem-se essencialmente em dois grupos: as imagens de satélite e dados vetoriais georreferenciados; cada um com as suas peculiaridades e desafios associados, pelo que foram usadas soluções diferentes para cada um destes grupos. Neste componente da implementação os principais desafios encontrados foram: por um lado, a grande diversidade de dados georreferenciados e, por outro, a massiva quantidade de imagens de satélite. Enquanto que os dados vetoriais foram mantidos localmente numa base de dados Postgres, as imagens de satélite foram mantidas em ambiente cloud, não havendo desta forma elevados requisitos de largura de banda nem espaço local para os descarregar. Nas duas subsecções seguintes serão detalhados estes conjuntos de dados.

5.1.1 Dados georreferenciados

Com a grande quantidade e variedade de dados para analisar e processar, foi necessário encontrar uma ferramenta que ajudasse com o processo. Para tal, foi selecionado o PostgresSQL 12.2 para gerir bases de dados, em conjunto com PGAdmin para facilitar a interação com o PostgresSQL 4.21. Finalmente, foi ainda usado o PostGIS para facilitar a gestão de dados georreferenciados.

A utilização de uma base de dados trouxe vários benefícios: em primeiro lugar, uma fácil integração com outras ferramentas como Tableau e QGIS para analisar e visualizar os dados; em segundo lugar, a centralização e organização dos dados que facilitou a criação de conjuntos de dados derivados; em terceiro lugar, o processamento eficiente dos dados

bem como as ferramentas para os processar; finalmente, o processamento dos dados ficou completamente detalhado nos scripts SQL facilitando a compreensão da sua criação.

Para simplificar a criação deste ambiente foi criado um script docker para lançar um container com Postgres e outro com o PGAdmin. Desta forma, o ambiente é controlado e fácil de recriar noutra máquina.

Para o presente trabalho, nem a atualização contínua dos dados nem o espaço ocupado pelas tabelas foram prioritários no desenho da base de dados. O foco esteve em ter rapidamente toda a informação disponível para consulta e filtragem. Por esse motivo, foram criadas tabelas desnormalizadas, removendo assim a necessidade de realizar junções dispendiosas aquando de uma consulta. O processo foi facilitado pois todos os conjuntos de dados originais já se encontravam desnormalizados, como é normal para dados exportados.

Na imagem 5.1 é possível observar o diagrama das tabelas criadas na base de dados. Algumas tabelas de auxílio foram omitidas e alguns nomes também foram alterados no diagrama para uma melhor compreensão do leitor. É possível observar a desnormalização referida, por exemplo, as tabelas **COS** e **COS_changes**. Todas as propriedades da tabela COS estão repetidas duas vezes na tabela **COS_changes** para representar a alteração com os dois estados. Certamente seria possível remover esta duplicação, porém o interesse principal está em consultar esta tabela com os atributos completos e não estar constantemente a recriá-la. Vejamos agora com mais detalhe cada uma das tabelas presente na imagem 5.1.

A tabela **COS** contém todos os polígonos COS das edições 1995, 2007, 2010 e 2015. Dos atributos salienta-se a **geom** como sendo o polígono que define uma dada área com o uso de solo descrito no atributo **description** e com o código de uso **use_code**. Como é possível observar, a tabela tem dois IDs: o **id**, que é chave primária e um identificador interno à base de dados gerado quando um novo polígono é introduzido; e o **id_cos**, que é o identificador original de cada edição COS respetiva. Finalmente, para possibilitar a tradução entre várias nomenclaturas, o atributo **naming** indica qual a nomenclatura usada para cada linha da tabela, tal como referido na subsecção 4.2.1. Este atributo, juntamente com o **use_code**, torna possível a junção com a tabela **cos_translation**. Esta tabela foi gerada a partir de duas tabelas presentes em [42], uma que possibilita a tradução entre a nomenclatura de 2010 e a de 2015 e outra que possibilita a tradução entre a nomenclatura de 2007 e 1995. É possível observar um excerto desta tabela original na tabela 5.1, e, na tabela 5.2, é possível observar as linhas geradas para a tabela **cos_translation**. Ao realizar uma junção com esta tabela, e filtrando o **naming_to**, é possível fazer a tradução entre duas nomenclaturas.

Sendo possível a tradução entre nomenclaturas, foi possível calcular as alterações entre edições COS. Deste cálculo, resulta a tabela **COS_changes**. Esta tabela tem os mesmos atributos que a tabela COS mas em duplicado de forma a ser possível comparar todos os atributos entre edições consecutivas. É importante salientar um detalhe: para a criação desta tabela, foi necessário calcular todas as interseções entre os polígonos de edições

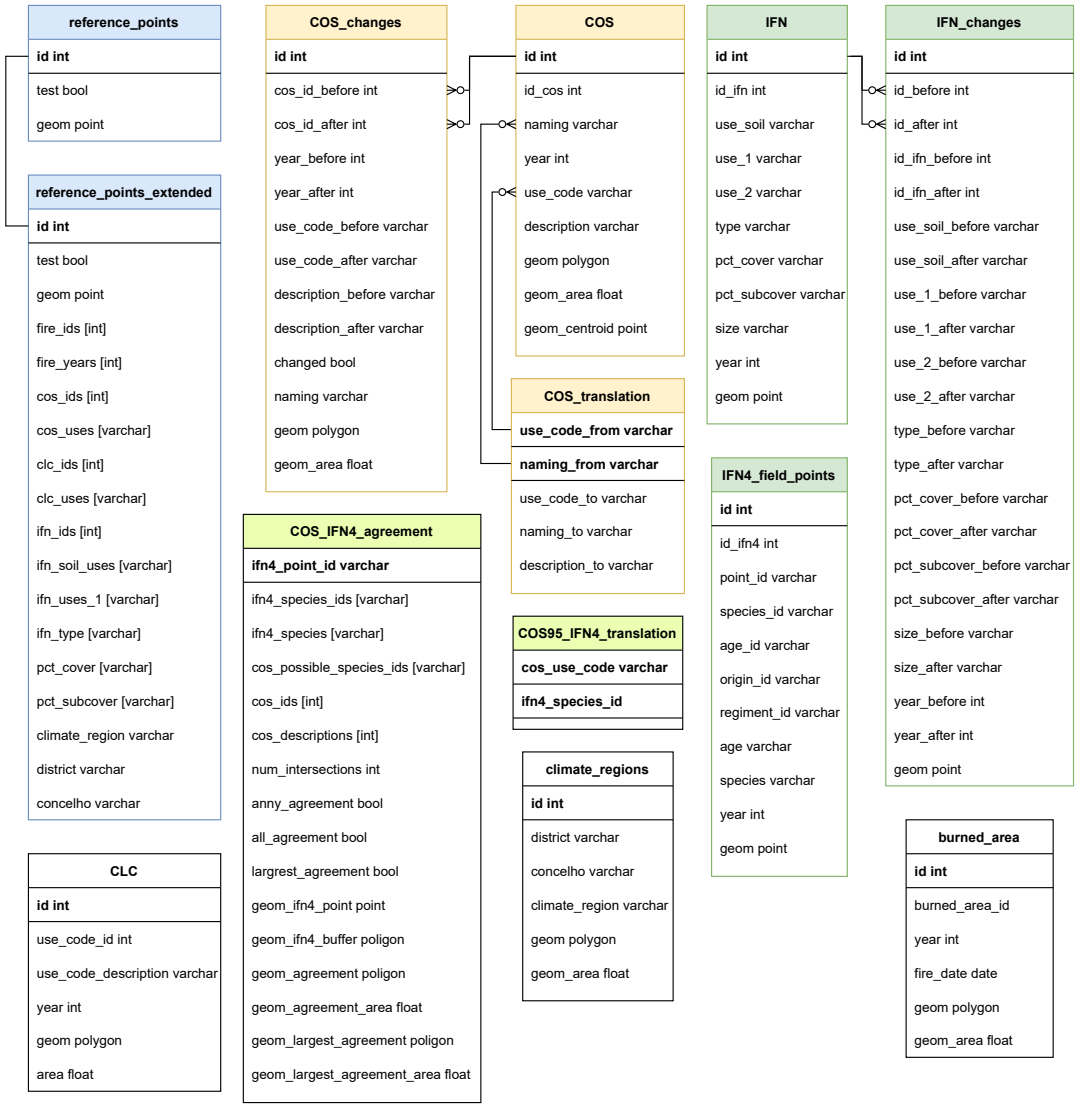


Figura 5.1: Diagrama de tabelas da base de dados criada para acomodar os dados geográficos e criação dos pontos de referência

Tabela 5.1: Pequena parte da tabela de tradução entre nomenclaturas COS disponibilizada em [42].

Nomenclatura COS2010	Nomenclatura COS2015
3.2.4.06.3 Florestas abertas de outras resinosas com folhosas	3.1.2.00.3 Florestas de outras resinosas
3.2.4.06.4 Florestas abertas de misturas de resinosas com folhosas	3.1.2.00.3 Florestas de outras resinosas
3.2.4.07.1 Outras formações lenhosas	3.2.2.00.0 Matos
3.2.4.08.1 Cortes rasos de florestas de sobreiro	3.1.1.00.1 Florestas de sobreiro
3.2.4.08.2 Cortes rasos de florestas de azinheira	3.1.1.00.2 Florestas de azinheira

Tabela 5.2: Excerto da tabela **cos_translation** derivada a partir da tabela 5.1.

naming_from	use_code_from	naming_to	use_code_to	description_to
cos0710	3.2.4.06.3	cos15	3.1.2.00.3	Florestas de outras resinosas
cos0710	3.2.4.06.4	cos15	3.1.2.00.3	Florestas de outras resinosas
cos0710	3.2.4.07.1	cos15	3.2.2.00.0	Matos
cos0710	3.2.4.08.1	cos15	3.1.1.00.1	Florestas de sobreiro
cos0710	3.2.4.08.2	cos15	3.1.1.00.2	Florestas de azinheira

consecutivas; no entanto, este cálculo é dispendioso sem a utilização de índices espaciais, pelo que foi necessário criá-los para acelerar o processo. Ainda assim, só este passo não se provou suficiente para calcular as alterações em tempo útil devido a polígonos de grandes dimensões como as redes de autoestradas, pelo que foi necessário subdividir polígonos com um número de vértices elevado.

A tabela **IFN** contém os vários pontos analisados no IFN ao longo das quatro edições disponíveis. Dos atributos presentes, salientam-se **use_soil** que indica o uso de solo (ex: Agricultura, Floresta, etc.) e **use_1**, que indica qual a espécie dominante presente no local (ex: Castanheiro, Sobreiro, Eucalipto). À semelhança da tabela COS, também conta com dois identificadores: **id**, identificador interno, e o **id_ifn**, identificador original de cada edição do IFN. Com esta tabela foi possível calcular as diferenças entre edições IFN de forma semelhante às alterações COS mas, uma vez que as várias edições são coerentes na nomenclatura, não foi necessária qualquer tradução, apenas comparação direta dos atributos.

A tabela **IFN4_Field_points** contém os dados dos pontos de campo do IFN 4, já apresentado na subsecção 4.2.2. Desta tabela destaca-se o atributo **species_id** que identifica uma espécie presente no ponto em questão. Salienta-se que o mesmo ponto poderá ter mais do que uma espécie presente, pelo que uma nova linha na tabela existirá neste caso. Esta tabela está presente na base de dados para realizar uma comparação entre as espécies indicadas como presentes com o COS do mesmo ano, de modo a tentar esclarecer as questões de precisão levantadas durante a análise preliminar do conjunto de dados. Para tal, foi necessário realizar uma tradução entre a nomenclatura usada pela COS 1995 e pelos pontos de campo IFN. Não estando disponível uma tradução oficial, foi necessário criar uma, interpretando as definições das duas nomenclaturas e selecionando o código respetivo para cada classificação. Desta análise manual resulta a tabela **COS95_IFN4_translation** com dois atributos: **cos_use_code**, que indica um código COS da nomenclatura de 1995, e **ifn4_species_id**, que indica um ou mais códigos de espécie que podem estar presentes no respetivo código COS. Assim sendo, com o auxílio desta tabela é possível, a partir de um código COS, obter os códigos das possíveis espécies IFN. Uma vez que as espécies presentes no IFN são muito mais detalhadas, nem sempre é possível escolher apenas uma espécie. Por exemplo, a descrição COS "3.1.1.01.5 Florestas de eucalipto"apenas pode levar a tradução para o código "EC"indicativo de Eucaliptos; no

entanto, a descrição COS "3.1.1.01.7 Florestas de outras folhosas"já engloba um conjunto de possíveis espécies "AC"(Acácias), "BT"(Bétula), etc.

A tabela **COS_IFN4_agreement** é o resultado da comparação dos pontos de campo do IFN 4 com o COS de 1995. A tabela **COS95_IFN4_translation** tornou esta comparação possível . Usando todos os pontos IFN4, foi criado um buffer de 17m em volta dos pontos por ser a área de análise, transformando todos os pontos em círculos de raio 17m . Seguidamente, foi realizada a interseção com os polígonos COS e consequente tradução dos códigos COS para os códigos de espécies IFN. Neste momento, para cada ponto IFN é possível saber quais as espécies que o IFN indica estarem presentes no local e quais as espécies que a COS apresenta como possíveis estarem no local. A partir destes dados são calculados os atributos: **num_intersections** - quantas parcelas COS a zona de análise IFN interseta; **any_agreement** - se alguma das espécies IFN se encontra nas possíveis espécies COS; **all_agreement** - se todas as espécies IFN se encontram nas possíveis espécies COS; **largest_agreement** - se todas as espécies IFN se encontram nas possíveis espécies da maior a parcela COS .

A tabela **burned_area** contém todos os incêndios presentes nos conjuntos de dados do ICNF. Para cada incêndio é apresentada a geometria da área ardida, o ano e a data da ocorrência.

A tabela **climate_regions** contém todos os concelhos do país com a indicação da região climática a que pertencem. Estes dados foram fornecidos pelo colega Henrique Coelho. É possível na imagem 5.2 observar os vários concelhos e as respetivas regiões climáticas indicadas pela cor. A separação em cinco zonas tem por base o clima das diferentes regiões.

A tabela **CLC** contém todos os polígonos CLC das edições 1990, 2000, 2006, 2012.

Finalmente as tabelas **reference_points** e **reference_points_extended** contêm os pontos que foram usados na avaliação. Estes pontos são um subconjunto dos pontos presentes na tabela **INF** após uma amostragem estratificada tendo em conta a espécie e a região climática indicada pela tabela **climate_regions**. A tabela **reference_points** indica apenas os pontos e o seu identificador enquanto que a tabela **reference_points_extended** conta com informações adicionais sobre a história do ponto segundo as fontes de dados disponíveis. A primeira tabela foi usada para extrair as séries temporais enquanto que a segunda foi usada para auxiliar realizar a avaliação manual dos pontos oferecendo um contexto mais completo do historial do ponto.

5.1.2 Imagens de Satélite

As imagens de satélite nunca chegaram a ser guardadas localmente pois seria necessário muito espaço local bem como largura de banda para descarregar o catálogo de imagens Landsat. A solução para este problema passou pela utilização do Google Earth Engine (GEE) [21] para realizar o processamento necessário na cloud e apenas descarregar os resultados estritamente necessários para serem analisados. Uma vez que a plataforma

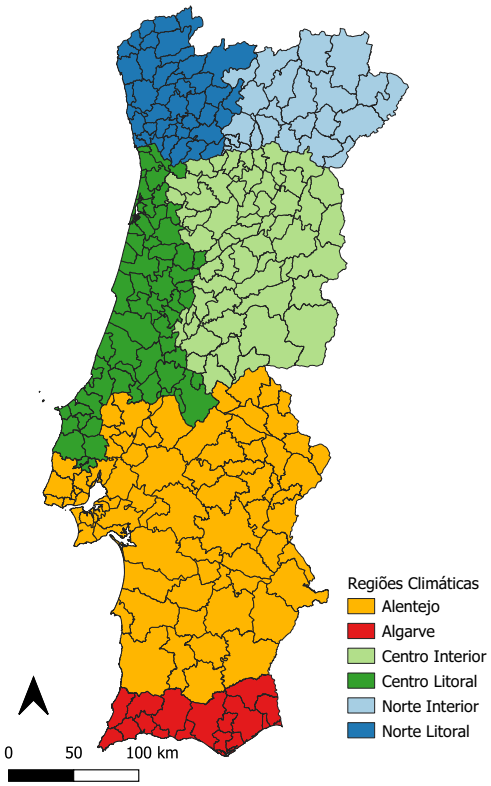


Figura 5.2: Regiões climáticas de Portugal por concelho.

oferece um grande conjunto de imagens de satélite, bem como as ferramentas necessárias para as processar, esta foi a solução mais prática para implementar as partes do trabalho que necessitavam de dados de satélite. Será fornecida uma descrição mais detalhada sobre o GEE em 5.2.1.

5.2 Processamento

Tal como os dados, o processamento ficou dividido em dois: uma parte ocorreu no GEE, enquanto que outra correu no ambiente local. O fluxograma presente na imagem 5.3 mostra claramente esta divisão de tarefas e oferece uma visão global das tarefas de processamento realizadas a as suas dependências. É possível observar duas grandes divisões, do lado esquerdo o ambiente local e do lado direito o GEE. O fluxo diverge à partida para duas tarefas de processamento de dados. No GEE foi feita a criação da coleção contígua de imagens Landsat (4,5,7,8) já descrita na subsecção 4.2.5, e, a partir desta foi criada uma outra coleção de imagens anuais para o algoritmo LandTrendr detalhado no início da subsecção 4.5.1. Já no ambiente local os dados florestais são processados tal como descrito nas subsecções 4.2.1 4.2.2, 4.2.3, 4.2.4 para compatibilizar diferentes edições e calcular diferenças entre elas quando necessário. Com base nestes dados é gerado o conjunto de pontos de referência a serem analisados tal como está detalhado nas subsecções 4.4.1,

4.4.2. Estando o processamento de dados concluído foi possível passar à realização das experiências bem como à análise manual dos pontos de referência. Nesta fase, tal como é possível observar no fluxograma, o algoritmo BOCPD foi executado no ambiente local enquanto que os algoritmos CCDC e LandTrendr foram executados no GEE. Dada esta separação foi necessário importar para o GEE os pontos de referência usando um ficheiro shapefile; e exportar as séries temporais para o ambiente local usando um ficheiro csv com uma observação por linha, isto é, cada linha representa um ponto numa determinada data e está reportado o NDVI, NBR, TasseldCap Angle. Esta movimentação de dados entre os dois ambiente está salientada com setas azuis no diagrama, para mais informação sobre os ficheiros em anexo. A análise manual dos pontos de referência já detalhada na secção 4.4 foi feita parte no GEE, parte no ambiente local; a visualização das séries temporais foi feita no GEE, no entanto os resultados foram anotados no ambiente local. Tendo as classificações de todas as experiências, bem como uma classificação de referência foi possível passar ao último passo a avaliação dos resultados. Os ficheiros com as classificações encontram-se no formato csv e em cada linha indicam indicam de forma exaustiva todos os pontos e o ano onde foi detetada uma alteração.

5.2.1 A plataforma Google Earth Engine

O Google Earth Engine (GEE) [21] é uma plataforma na cloud que tem como objetivo a análise de dados geoespaciais. Esta análise é realizada criando scripts, em Python ou em Javascript, que em parte correm no lado do cliente e em parte no lado dos servidores. O desenvolvimento de scripts em Python requer a instalação do interpretador de Pyhton, bem como todas as dependências necessárias para utilizar o Google Earth Engine. Para desenvolver em JavaScript, é apenas necessário um browser; isto porque a Google disponibiliza um ambiente de desenvolvimento interativo baseado em tecnologias web. A utilização do GEE requer que a inscrição na plataforma seja primeiro aceite; só depois disso poder-se-á usufruir gratuitamente deste serviço.

O GEE permite o acesso a vários conjuntos de dados, dos quais se destaca o projeto LandSat. Também é possível importar dados para a plataforma, o que foi necessário para acomodar os dados de referência.

Existem essencialmente dois tipos de dados no GEE: imagens (dados raster) e features (dados vetoriais). Estes dois tipos de dados encontram-se agrupados em coleções que podem ser filtradas por data e zona e até reduzidas a uma só imagem, calculando valores como o valor médio ou máximo num pixel. As imagens são conjuntos de bandas. As bandas podem representar inúmeros conceitos, como a intensidade de luz no infravermelho, um índice como o NDVI ou uma classificação (floresta jovem/ floresta adulta/ floresta antiga) isto porque, numa mesma imagem, diferentes bandas podem ter diferentes resoluções e tipos de dados. A uma imagem é possível adicionar e remover bandas, realizar cálculos pixel a pixel com bandas para obtermos índices, somar, subtrair e dividir por outras imagens, e recortar segundo um polígono, entre muitas outras. A plataforma

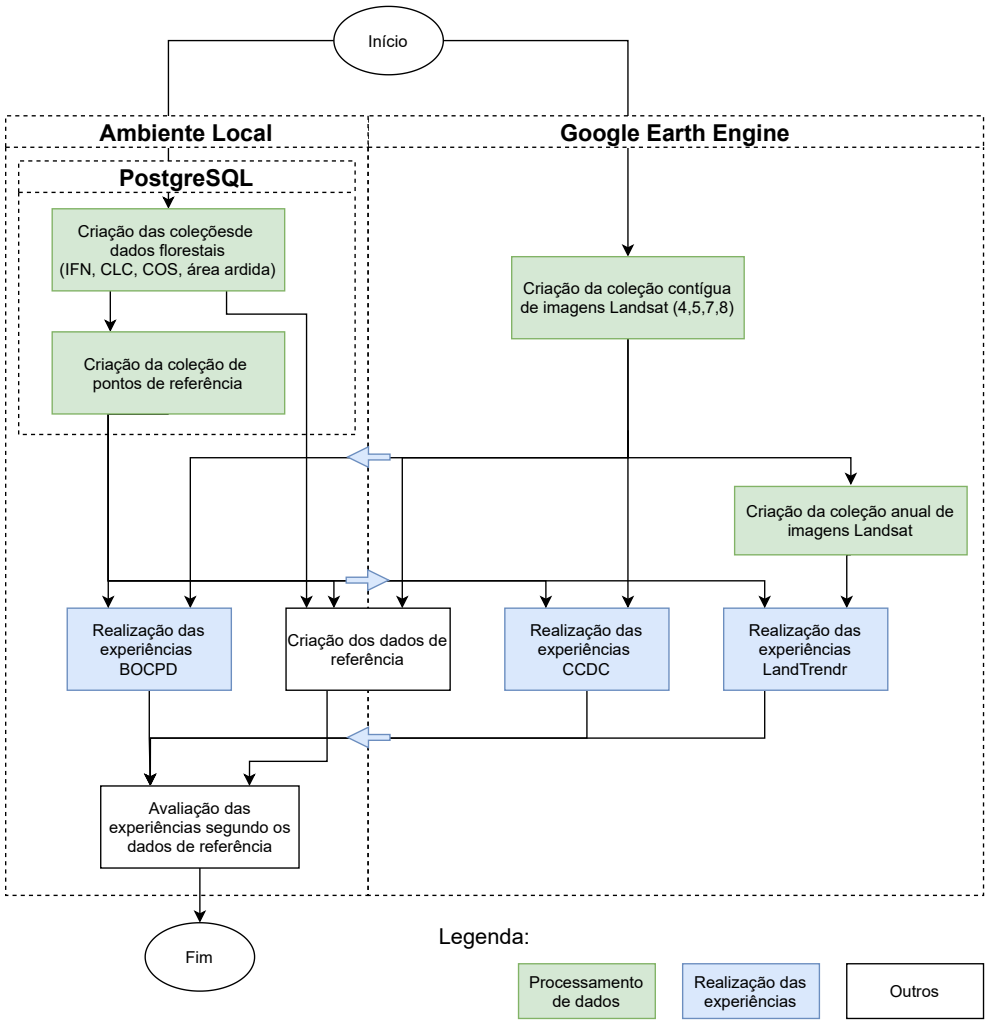


Figura 5.3: Fluxograma com as principais tarefas de processamento realizada, as suas interdependências e qual o ambiente onde foram executadas

disponibiliza implementações para 6 algoritmos de deteção de mudanças entre os quais se destacam: LandTrendr, CCDC e EWACD.

Durante o trabalho, o GEE serviu essencialmente 3 propósitos:

- Visualização de séries temporais e imagens de satélite
- Extração de séries temporais
- Execução de algoritmos de deteção de pontos de quebra

Um dos primeiros scripts produzidos durante o trabalho consiste numa visualização interativa de séries temporais do índice NDVI e das imagens de satélite que a compõem. É possível observar o resultado na imagem 5.4. Após iniciar o script, o utilizador tem acesso a um mapa do planeta, no qual pode clicar em qualquer ponto (onde estejam disponíveis

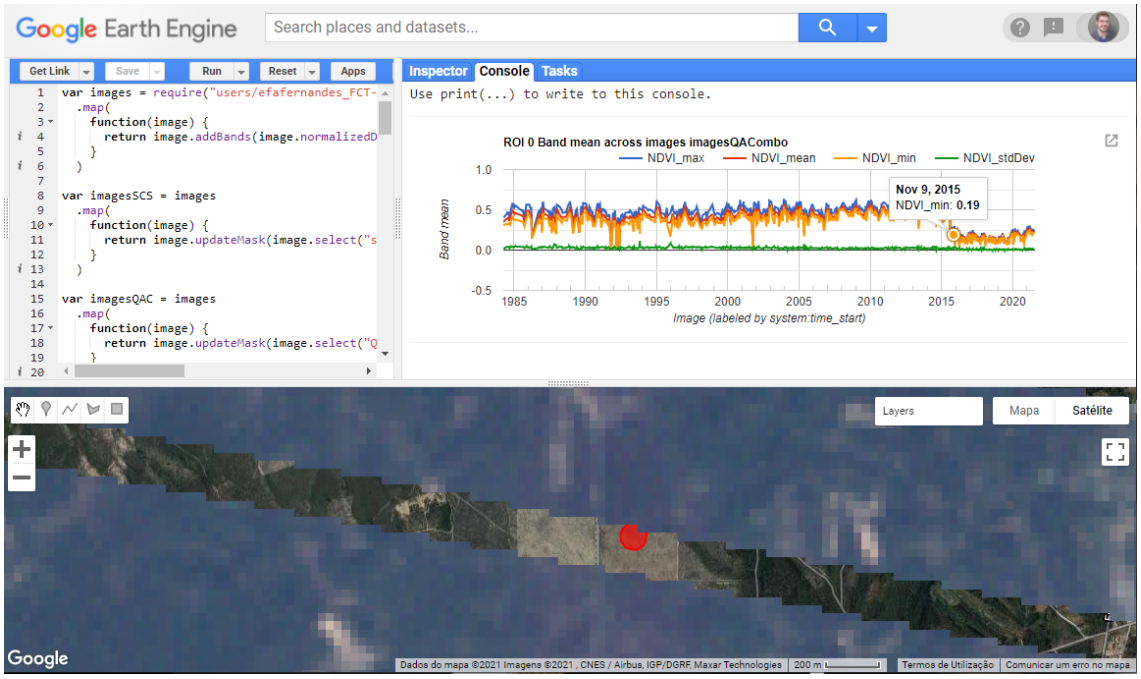


Figura 5.4: Script de visualização de séries temporais de NDVI e imagens de satélite. No quadrante superior esquerdo, observa-se a janela de código. No quadrante superior direito, mostra-se a série temporal do NDVI. Na metade inferior, vê-se a vermelho a zona de onde foi exportada a série temporal e a imagem de satélite da data salientada na série temporal. (É possível visualizar uma linha diagonal pixelizada, este artefacto deve-se a falhas do sensor do satélite Landsat 7).

imagens landsat), e, após algum tempo será apresentada a série temporal do NDVI da zona num raio de 50m do ponto em questão apresentando o valor máximo, mínimo, médio e o desvio padrão. É possível passar com o rato sobre a série temporal e obter informações precisas sobre os valores de um determinado dia como é possível observar no quadrante superior direito na imagem 5.4. É ainda possível clicar numa observação da série temporal para visualizar a imagem que lhe deu origem, tal como é possível observar na metade inferior da imagem 5.4.

O segundo script usa a mesma fonte de imagens Landsat para exportar séries temporais de um conjunto de zonas. Os resultados são extraídos para uma pasta da Google Drive do utilizador no formato csv. Usando este script foi possível extrair as séries temporais precisas para realizar visualizações sem ser com o GEE e correr os algoritmos de deteção de alterações não presentes no GEE.

Finalmente, foi desenvolvido ainda um script que usa as implementações de deteção de alterações do GEE. Durante a implementação foram testados quatro algoritmos: CCDC, Landtrendr, EWMACD e Verdet. Porém, apenas foi possível usar os dois primeiros para áreas mais extensas. Os dois últimos quando aplicados a certas zonas apresentaram erros durante a execução que até à data não foram diagnosticados. Após alguma pesquisa no

fórum de programadores do GEE, verificou-se que outros utilizadores da plataforma também já se cruzaram com os mesmos problemas. Os resultados deste script são exportados como ficheiros csv para uma pasta da Google Drive, como um conjunto de anos onde há quebras por cada ponto.

É importante salientar alguns detalhes de implementação deste script que executa algoritmos de deteção de alterações; todos os algoritmos executados têm como input várias imagens e como output uma imagem. Pelo que, mesmo apenas havendo interesse num conjunto limitado de pontos espalhados pelo país, é necessário correr a computação para todas as imagens do país. De forma a reduzir a quantidade de processamento realizado as imagens de input recebem uma máscara para apenas haver processamento nos pixeis de interesse. Assim sendo, a imagem resultante é praticamente toda vazia à exceção dos pontos de referência. Após esta imagem ser criada é então possível extrair apenas os dados referentes aos pontos de interesse.

O GEE tem um modelo de programação particular, isto porque parte da computação é feita no servidor enquanto que outra parte é corrida localmente. O servidor recebe queries na forma de grafos dirigidos acíclicos que descrevem toda a computação de forma funcional. Localmente um conjunto de primitivas permitem criar as queries de forma relativamente simples e quase transparente. Para tomar partido total das potencialidades do GEE foi necessário não só usar o ambiente interativo no browser mas também criar tarefas de exportação isto porque a primeira via contem limitações de tempo de processamento. Já quando se exportam dados o tempo de processamento não está à partida limitado. Desta forma o código criado consiste em vários scripts que ou exportam resultados para a Google Drive ou apresentam de forma interativa informação no browser.

5.2.2 Ambiente local

Por muitas vantagens que o GEE ofereça é sempre bom poder complementar o ambiente cloud com a maior flexibilidade que um ambiente local oferece. Para este propósito foi usada uma máquina com as seguintes caraterísticas:

- Processador Intel Core i7-8650U
- Memória 32GB DDR4
- Armazenamento 512 GB Samsung M.2 PCIe NVMe SSD

A base de dados Postgres já referida na subsecção 5.1.1 realizou a maior parte do processamento necessário dos dados georreferenciados. Como exemplos do processamento realizado destacam-se: a criação do mapa de alterações COS, a unificação das áreas ardidas do ICNF, a comparação entre o COS 95 e o IFN 95, extração dos pontos de referência bem como informações associadas.

No que diz respeito ao processamento de séries temporais no ambiente local foi usado Python 3.7 juntamente com uma biblioteca que disponibiliza o algoritmo BOCPD [31].

Adicionalmente foram utilizadas as seguintes bibliotecas para completar o processamento e formatação de dados:

- pandas
- geopandas
- numpy
- matplotlib

Finalmente o processamento dos resultados das segmentações temporais também foi realizado num ambiente local com um script python.

5.3 Visualização de dados

Durante o desenvolvimento do trabalho, foi necessário muitas vezes visualizar os dados a trabalhar e os resultados finais. Para tal, foram essencialmente usadas três ferramentas com propósitos distintos:

- Google Earth Engine
- QGIS
- Tableau

Cada ferramenta tem os seus pontos fortes e limitações, por isso foi necessário usar a ferramenta certa para cada visualização.

O Google Earth Engine destaca-se pela quantidade de imagens de satélite disponíveis bem como a sua capacidade de processamento. Assim sendo, foi usado principalmente para visualizações interativas de imagens e séries temporais, tal como já foi explicado em 5.2.1, onde é possível ver a imagem 5.4 que mostra um exemplo de uma visualização criada.

O QGIS[40] é um sistema de informação geográfica com código fonte aberto capaz de visualizar informação georreferenciada como, por exemplo, imagens de satélite ou dados vetoriais. Dentro do conjunto de possibilidades que o software oferece, este foi utilizado em duas principais tarefas: em primeiro lugar, serviu para importar e exportar dados georreferenciados para a base de dados Postgres; e, em segundo lugar, permitiu a rápida análise visual de dados georreferenciados, como é o caso do COS, CLC, IFN etc, quer estejam presentes num ficheiro ou na base de dados usada no trabalho. É possível ver um exemplo da sua utilização na imagem 4.2 na subsecção 4.2.2.

Finalmente, o Tableau Desktop é uma ferramenta de visualização e análise de dados. No contexto do presente trabalho foi essencialmente usado para gerar visualizações de forma a melhor compreender resultados e os dados disponíveis. As principais vantagens

da utilização do software estão relacionadas com a rápida criação de visualizações interativas e completas, e com a fácil interação com a base de dados Postgres e outros ficheiros. As tarefas realizadas incluem:

- Visualizações para compreensão de fontes de dados
- Visualização de classificações de séries temporais
- Visualização de diferentes parametrizações nas métricas de um algoritmo
- Visualização de diferentes algoritmos e as suas métricas

5.4 Conclusão

Durante este capítulo foram apresentadas as principais ferramentas que possibilitaram a implementação do presente trabalho. Dividindo a implementação em três componentes (armazenamento e gestão de dados, processamento e visualização) foram explorados os contributos de cada ferramenta. Deu-se especial atenção ao GEE pelas suas valências nas três componentes. A plataforma da Google distingue-se em dois aspetos: disponibilização de grandes quantidades de dados de satélites prontos para análise, bem como a infraestrutura e capacidade de os processar. O software de gestão de bases de dados provou ser útil no armazenamento de dados georreferenciados, rápidas transformações dos mesmos e interoperabilidade com outras ferramentas. Finalmente, no campo da visualização o Tableau destacou-se na visualização interativa e exploratória de dados enquanto que o QGIS revelou-se importante para visualizar várias fontes de dados georreferenciados.

6

Avaliação e Resultados

O presente capítulo apresenta os resultados das experiências realizadas. Em primeiro lugar na secção 6.1 é feita a comparação entre a edição de 1995 do COS e o IFN do mesmo ano, com o objetivo de avaliar a consistência dos dados. A secção 6.2 explora o conjunto de dados de referência criado manualmente e usado para a comparação dos vários algoritmos. A secção 6.3 comparação o tempo de execução de vários tipos de processamento na plataforma GEE. A secção 6.4 descreve os resultados dos vários algoritmos analisados. A primeira comparação é generalista e toma em conta todos os algoritmos. Posteriormente cada algoritmo é analisado individualmente de forma a compreender o impacto que cada parâmetro têm na sua avaliação. Finalmente na secção 6.5 é feito um pequeno sumário do capítulo.

6.1 Comparação COS95 pontos de campo IFN95

A comparação do COS95 com os pontos de campo do IFN95 tem como objetivo de confirmar a consistência e qualidade das fontes de dados. Tendo em conta os dados fornecidos pelos dois conjuntos de dados é possível comparar os pontos de floresta de ambos bem como as espécies florestais. No entanto há um grande obstáculo pelo caminho, sem informação adicional não é possível comparar os dois conjuntos de dados visto que usam nomenclaturas diferentes. De forma a superar este obstáculo foi necessário criar uma tabela de tradução, que, para cada código do COS indicasse qual a espécie presente no local. Este processo foi feito de forma manual tendo em atenção as especificações de ambos os conjuntos de dados.

Tendo uma forma de compatibilizar as duas nomenclaturas há mais um problema a resolver: apesar de os fotopontos IFN serem ,tal como o nome indica, pontos, a sua classificação resulta da análise de uma área com um raio de 19m em torno do ponto indicado. Isto leva a que um ponto IFN possa intersetar mais do que um polígono COS. Desta forma será necessário analisar todos os polígonos COS que intersetam a área em torno de um dado ponto IFN.

Tabela 6.1: Resultados da comparação das espécies classificadas entre o COS95 e o IFN95

	Concordância	
	Sim	Não
Algum	1705	494
Maior Área	838	1361
Todos	1037	1162

O principal problema é compreender se a classificação COS vai ao encontro da classificação IFN. Durante a avaliação foi necessário ter em conta que para esta comparação um ponto IFN poderá ter associado mais do que um polígono COS. Pelo que foi testado: se algum dos polígonos concordava com a classificação COS, se o polígono com a maior área concordava com a classificação COS e ainda se todos os polígonos concordavam com a classificação. Os resultados encontram-se na tabela 6.1, onde é possível perceber que há uma descrepância significativa mesmo usando a comparação mais permissiva de confirmar se alguma das classificações dos poligonos COS coincide com a classificação IFN. Havendo desacordo 22% das vezes.

6.2 Dados de Referência

A presente secção tem como objetivo analisar os dados de referência obtidos usando a metodologia descrita na secção 4.4. A análise foca-se em quatro questões: 1) Qual o efeito da invalidação de pontos nos dados de referência? 2) Quais são os motivos de invalidação? e, , 3) Como está distribuído o número de pontos de quebra? 4) Como estão distribuídas as causas dos pontos de quebra? Para responder a estas questões foram criados três visualizações que ajudaram a guiar a análise.

6.2.1 Análise do efeito da invalidação de pontos

Olhando para a imagem 6.1 é possível retirar algumas conclusões quanto ao efeito da invalidação de pontos de análise. Em primeiro lugar, tendo em conta a totalidade dos pontos de referência, 33% foram tomados como inválidos. Desta forma passou-se de um total de 998 pontos de referência para 664 válidos.

Em segundo lugar, olhando para a distribuição geográfica dos pontos válidos e inválidos é possível observar uma ligeira concentração de pontos inválidos na zona do Alentejo.

Finalmente, prestando atenção à percentagem de invalidação por espécie, é possível observar que a percentagem de invalidação do eucalipto e do pinheiro bravo foram abaixo da média de 33%. Por outro lado, o sobreiro, azinheira, pinheiro manso e outras folhosas apresentam uma proporção de invalidação superior à média. As restantes espécies também contêm alterações, no entanto é difícil retirar conclusões definitivas devido ao número baixo de observações.

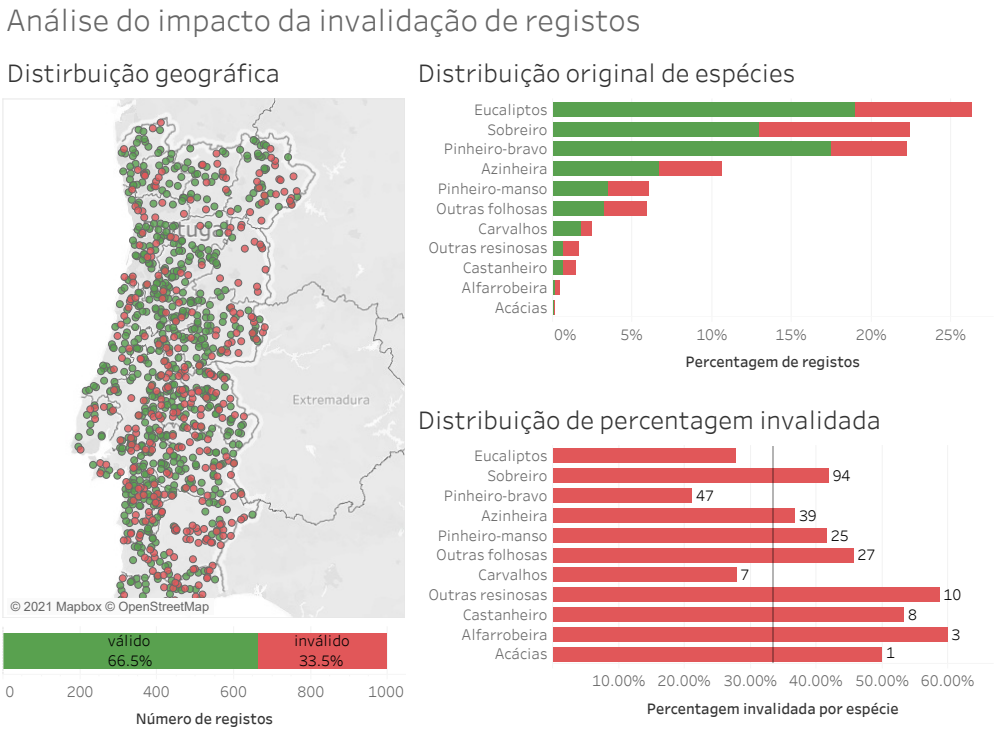
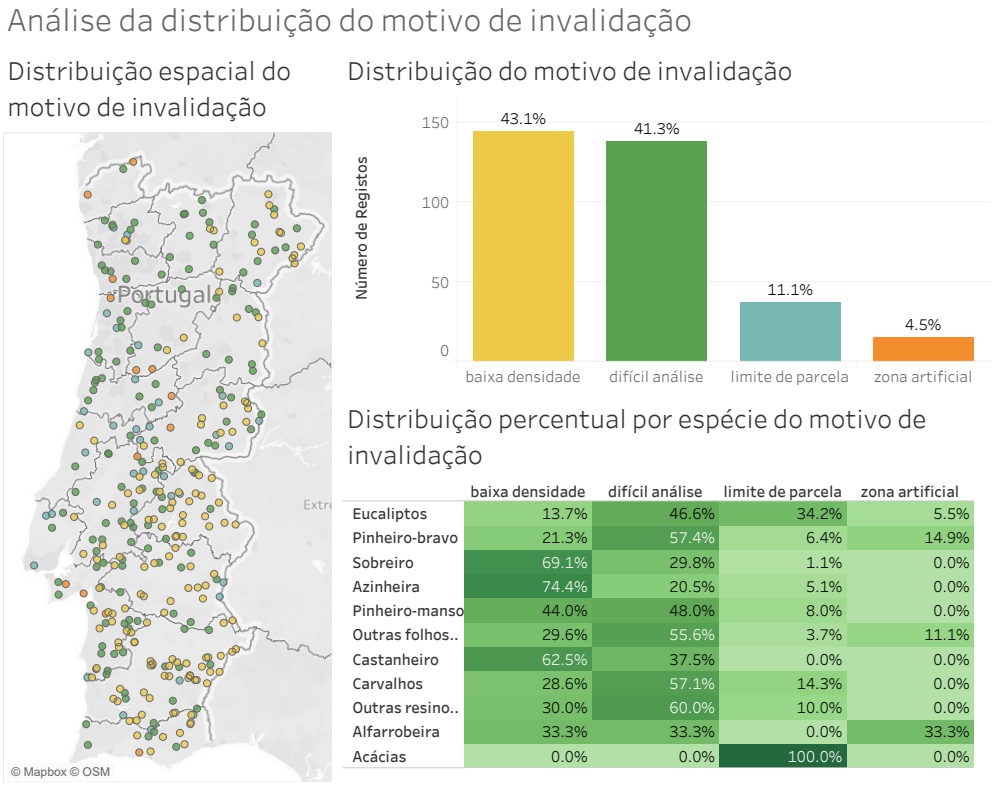


Figura 6.1: Imagem de auxílio à análise das invalidações. No lado esquerdo é visível a distribuição geográfica dos pontos válidos e inválidos. Diretamente abaixo as percentagens de pontos válidos e inválidos. No lado direito é possível observar a distribuição original de espécies. Diretamente abaixo encontra-se a percentagem de registos removidos por espécie.

6.2.2 Análise do motivo de invalidação

Para analisar o motivo de invalidação também foi criada a imagem 6.2. Em primeiro lugar, destaca-se que os principais fatores de invalidação são a baixa densidade de árvores e a difícil análise, enquanto que limites de parcela e zonas artificiais estão em clara minoria. Relativamente à distribuição geográfica destes motivos, há uma maior concentração de baixa densidade de árvores na zona do Alentejo, o que não é surpreendente tendo em conta as paisagens que se encontram nesta zona. Esta distribuição explica a maior concentração de pontos invalidados na zona do Alentejo observada em 6.2.1.

Relativamente à distribuição de motivos por espécie é importante destacar: azinheira, castanheiro e sobreiro. Para estas espécies a percentagem de invalidação devido a baixa densidade é superior quando comparada com a de outras espécies. No entanto, são as que menos sofrem com a difícil análise. Já para as espécies: eucalipto, pinheiro-bravo e outras folhosas verifica-se exatamente o contrário, havendo uma baixa percentagem de invalidações por baixa densidade e uma alta percentagem de invalidações por difícil



	baixa densidade	difícil análise	limite de parcela	zona artificial
Eucaliptos	13.7%	46.6%	34.2%	5.5%
Pinheiro-bravo	21.3%	57.4%	6.4%	14.9%
Sobreiro	69.1%	29.8%	1.1%	0.0%
Azinheira	74.4%	20.5%	5.1%	0.0%
Pinheiro-manso	44.0%	48.0%	8.0%	0.0%
Outras folhos..	29.6%	55.6%	3.7%	11.1%
Castanheiro	62.5%	37.5%	0.0%	0.0%
Carvalhos	28.6%	57.1%	14.3%	0.0%
Outras resino..	30.0%	60.0%	10.0%	0.0%
Alfarrobeira	33.3%	33.3%	0.0%	33.3%
Acácias	0.0%	0.0%	100.0%	0.0%

Figura 6.2: Imagem de auxílio à análise do motivo das invalidações. Do lado esquerdo, é possível observar a distribuição geográfica dos motivos. Do lado direito, na parte superior, encontra-se a distribuição das invalidações. Finalmente, no lado direito inferior, é visível a distribuição percentual por espécie florestal.

análise. O eucalipto tem também destaque na coluna de limite de parcela com sendo esta responsável por 34% das invalidações da espécie.

6.2.3 Análise do número de alterações

Estando terminada a análise sobre a invalidação de pontos e os motivos subjacentes a essa invalidação, muda-se o foco para os pontos válidos, mais concretamente para a distribuição do número de alterações por cada ponto. Tal como nas secções anteriores, a imagem 6.3 servirá como base para a análise.

Olhando primeiro para a distribuição total de pontos, é possível observar que a maior parte dos pontos tem zero alterações; mais do que isso, para uma determinada localização há uma maior probabilidade de haver nenhuma ou poucas alterações. Relativamente à distribuição geográfica, é possível observar que na zona do Alentejo se encontra uma grande concentração de pontos com zero alterações; já o centro do país conta com a maior

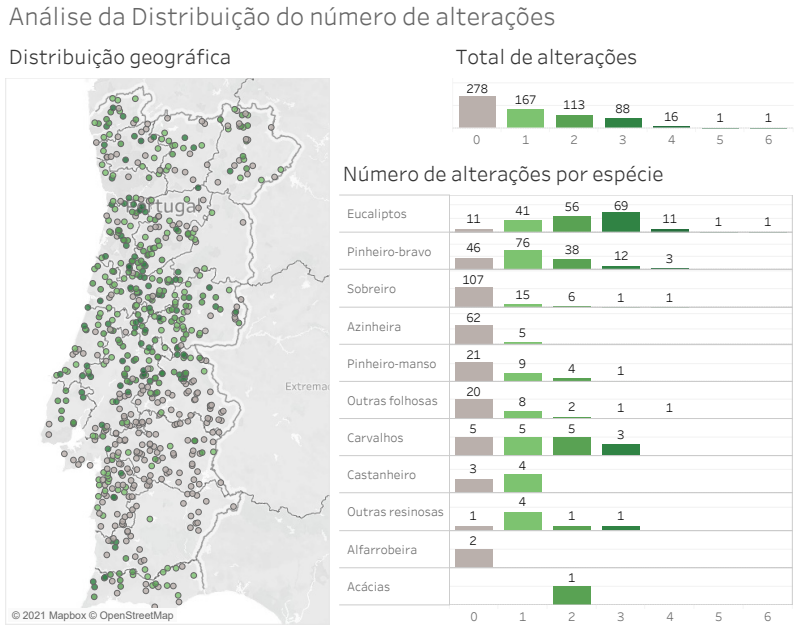


Figura 6.3: Imagem de auxílio à análise do número de alterações por ponto. Do lado esquerdo, é possível observar a distribuição geográfica do número de alterações por ponto. Do lado direito, no canto superior, a encontra-se a distribuição do número de alterações por ponto. Finalmente, no lado direito inferior, é visível a distribuição do número de alterações por espécie.

concentração de número de alterações.

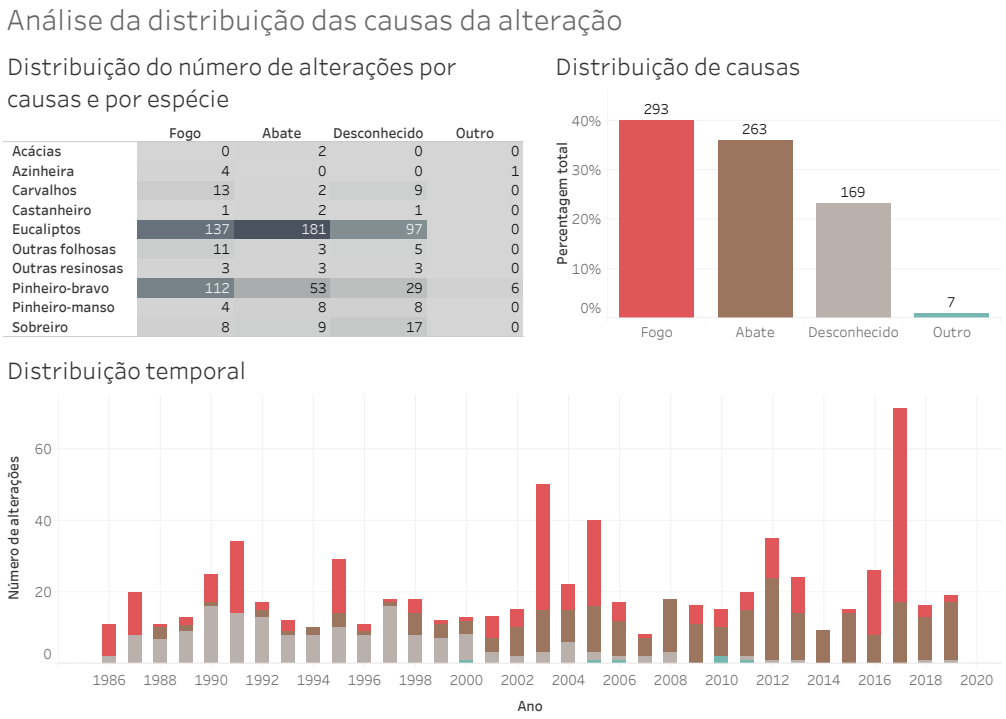
Ao virar a atenção para o número de alterações por espécie salientam-se duas: eucaliptos e pinheiro bravo por não confirmarem a tendência global de haver mais pontos com menos alterações. As causas mais prováveis para justificar o comportamento destas espécies serão as práticas culturais (cortes regulares) e os incêndios.

6.2.4 Análise das causas de alterações

Para além do número de alterações, como foi guardada informação sobre a sua causa, é possível analisar a distribuição das causas. A imagem 6.4 serve de base para a análise.

Em primeiro lugar é importante salientar que a maior parte de alterações têm como causa os incêndios, e em segundo lugar o abate. No entanto, a percentagem de causas desconhecidas é muito significativa, pelo que se torna difícil concluir se na realidade o fogo é o principal culpado pelo número de alterações na floresta.

Tomando atenção à distribuição temporal é bem claro que o número de causas desconhecidas é significativamente mais alto para os anos antes de 2000. Pelo contrário o abate é mais numeroso após o ano 2000. Estes fatores estão relacionados, isto porque, ao



	Fogo	Abate	Desconhecido	Outro
Acácias	0	2	0	0
Azinheira	4	0	0	1
Carvalhos	13	2	9	0
Castanheiro	1	2	1	0
Eucaliptos	137	181	97	0
Outras folhosas	11	3	5	0
Outras resinosas	3	3	3	0
Pinheiro-bravo	112	53	29	6
Pinheiro-manso	4	8	8	0
Sobreiro	8	9	17	0

Figura 6.4: Imagem de auxílio à análise da distribuição das causas de alteração. No canto superior direito encontra-se a distribuição percentual das causas. No canto superior esquerdo uma tabela com o número de alterações por causa por espécie. Finalmente na metade inferior a distribuição do número de alterações ao longo dos anos por tipo de causa.

contrário dos incêndios que estão registados, não há acesso a informação sobre o abate de árvores. Consequentemente, a verificação desta causa apenas foi possível usando imagens de satélite de alta resolução do Google Earth Pro. Por sua vez, a disponibilidade destas imagens está restrita a datas mais recentes.

6.2.5 Conclusão

Fazendo um apanhado do que foi discutido na presente secção, foram declarados inválidos 33% dos pontos iniciais, sobrando um total de 664 para análise, havendo desta forma uma perda mais significativa de sobreiros quando comparados com as restantes espécies.

As duas principais razões para a invalidação de pontos foram baixa densidade de árvores e difícil análise. Espécies como o eucalipto e o pinheiro bravo destacaram-se com maiores percentagens de difícil análise e menores percentagens de baixa densidade de

árvores. Em sentido contrário, o sobreiro e a azinheira destacaram-se por baixas percentagens de difícil análise e altas percentagens de baixa densidade. A mais provável causa para este comportamento é o contexto em que as espécies estão inseridas. A densidade de florestas de sobreiro e azinheira tende a ser mais baixa do que florestas de eucalipto e pinheiro bravo.

Relativamente aos pontos válidos há uma tendência global para haver menos pontos com mais mudanças. Esta tendência é quebrada por duas espécies: eucaliptos e pinheiro bravo.

6.3 Tempos de execução GEE

Com o objetivo de criar um conjunto de dados que não fosse demasiado complexo de processar, decidiu realizar-se alguns testes do tempo de execução do processamento de distúrbios florestais no GEE. Estes testes têm como objetivo compreender como o tempo de execução é afetado por três condicionantes: área a analisar, localização geográfica e dispersão geográfica. A primeira condicionante serve para compreendermos como varia o tempo de processamento com o aumento da área a analisar. A segunda condicionante fornece mais informação sobre um conjunto de zonas geográficas especiais onde há sobreposição de imagens de satélite. Finalmente, a última condicionante serve para informar sobre o tipo de impacto que espalhar a análise de uma mesma área total por diversas zonas tem.

Durante o processo de ingestão de imagens para o GEE, estas são submetidas a vários passos de processamento, um dos quais as subdivide em unidades de 256 por 256 pixeis. O processamento definido pelo utilizador ocorre posteriormente sobre estas subdivisões. Esta foi a unidade de base tomada nas experiências realizadas. Ainda que não garanta que se consiga enquadrar exatamente com as subdivisões, é capaz de fornecer a informação que desejamos.

Traduzindo os pixeis para distâncias de forma a definir as áreas para o Landsat, temos que: usando mosaicos de 265 px por 256 px, com imagens Landsar de resolução 30 m por 30 m, obtém-se um mosaico de 7680 m * 7680m. Para as experiências realizadas foram criados quadrados com as dimensões de 7600m * 7600m na projeção EPSG:3763.

As experiências consistem na execução do algoritmo Landtrendr para um conjunto diverso de áreas. Todas as experiências foram corridas três vezes, atingindo um balanço entre tempo de execução das experiências e consistência dos resultados.

6.3.1 Impacto da área total no tempo de execução

Para examinar a primeira condicionante da área de análise foram tomados em conta sete quadrados com áreas distintas: 0.25, 1, 4, 16, 57.76, 231.04 e 519.84 Km^2. Na imagem 6.5 é possível observar os tempos de execução em função da área processada. Destaca-se que o eixo onde a área está representada usa uma escala logarítmica. Olhando para a

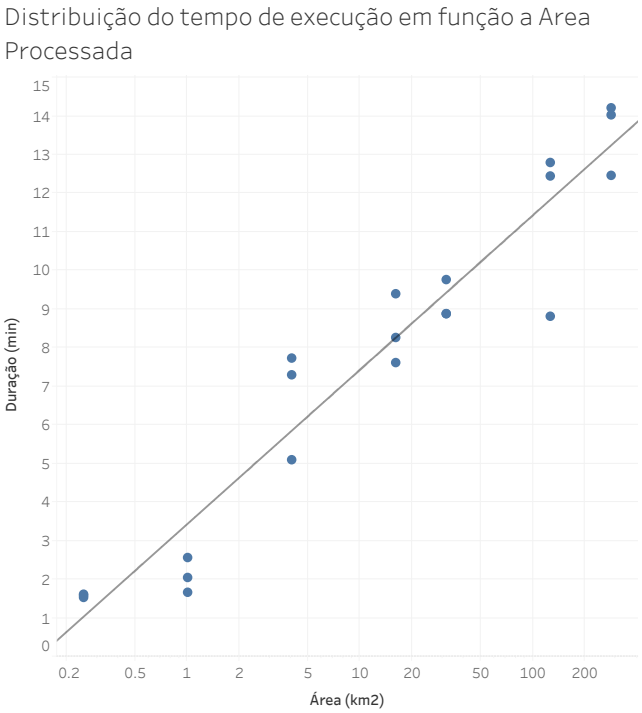


Figura 6.5: Distribuição dos tempos de execução com a linha de regressão encontrada. Equação: Duração = $1.73708 * ln(\text{Área}) + 3.41717$; $R^2 = 0.926231$

disposição dos pontos, aparenta haver uma tendência a estes disporem-se numa linha no gráfico, indicando desta forma uma relação logarítmica entre o tempo de execução e a área processada.

6.3.2 Impacto da localização no tempo de execução

Para examinar a condicionante da localização das áreas foram comparadas três zonas possíveis de observar na imagem 6.6. A zona marcada com a letra a) encontra-se totalmente dentro de apenas uma imagem Landsat; a zona marcada com a letra b) encontra-se na interseção entre duas imagens Landsat; finalmente, a zona marcada com a letra c) encontra-se na interseção de 4 imagens Landsat. Estes testes tentam perceber qual o impacto que este tipo de zonas tem nos tempos de execução.

É possível observar os resultados na imagem 6.7, onde se pode confirma que: a zona a) conta com um tempo de execução entre 13 e 15 minutos, a zona b) apresenta tempos de execução entre 12 e 17 minutos e, finalmente, a zona c) conclui as experiências com um tempo de execução entre 22 e 24 minutos. Destes resultados é possível observar que não há uma diferença significativa entre as áreas a) e b); no entanto, há uma diferença significativa entre os tempos de execução da área c) quando comparada com as restantes.

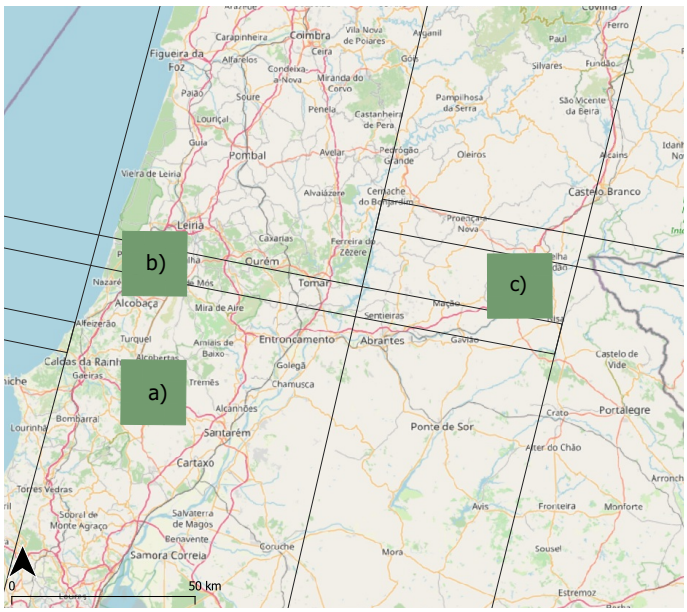


Figura 6.6: Áreas de estudo usado no estudo do impacto da localização no tempo de execução

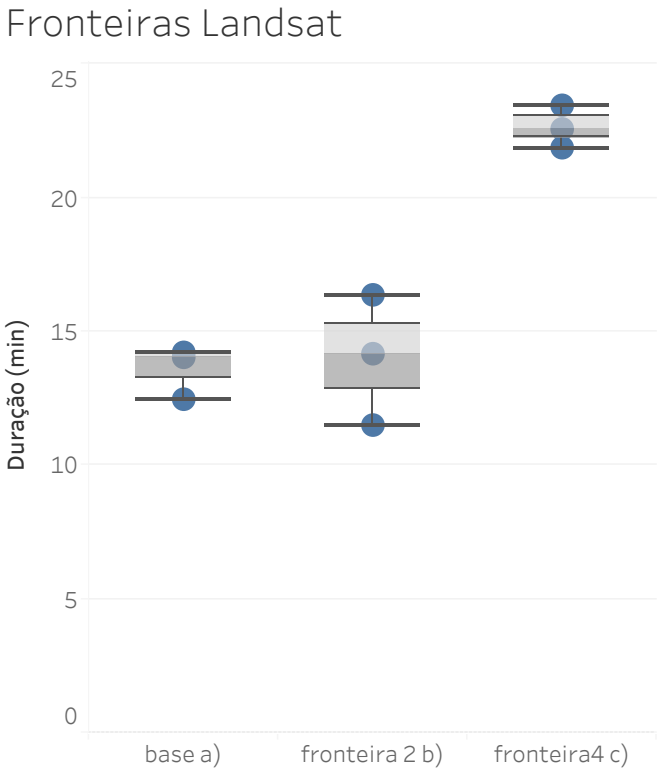


Figura 6.7: Gráfico com os tempos de execução do impacto da localização nos tempos de execução

6.3.3 Impacto da dispersão das áreas de análise no tempo de execução

Finalmente, para examinar a condicionante da dispersão espacial da área total, foram testadas três alternativas, possíveis de visualizar na imagem 6.8. A primeira alternativa, assinalada com a letra a), é uma área contínua; a segunda alternativa, identificada com a letra b), representa uma área não contígua mas aglomerada; finalmente, a terceira alternativa, marcada com a letra c), representa uma situação mais extrema onde quatro áreas estão espalhadas pelo território. É importante notar que o total de área analisada é o mesmo para estes três casos, $125km^2$.

Os resultados podem ser observados na imagem 6.8. Para a zona a), o tempo de execução varia entre os 9 e 12 minutos; para as áreas b), o tempo de execução varia entre os 12 e 15 minutos; finalmente, para as zonas c), o tempo de execução varia entre 34 e 50 minutos. Dos resultados é possível observar que há uma ligeira diferença entre a) e b) e uma grande diferença entre estas áreas e as áreas c). Após alguma ponderação sobre os resultados obtidos, a principal suspeita é que o tempo de guardar os resultados se esteja a sobrepor ao tempo de processamento. Isto porque a imagem resultante da análise da zona c) é superior à resultante da zona b) que por sua vez é superior à imagem resultante da zona a). É possível por isso concluir que a dispersão da área processada tem um impacto no tempo de execução das experiências.

Relativamente ao presente trabalho, estes resultados tiveram impacto na construção do conjunto de dados. Dado que o trabalho tem como escopo geográfico Portugal continental seria necessário um conjunto de dados de referência representativo do país. Adicionalmente os dados teriam de ser processados em tempo útil de modo a ser possível realizar um conjunto de experiências. Tendo estas duas restrições em mente selecionar apenas uma zona do país não seria de todo representativo da realidade. Usar a totalidade do país seria demorado mas representativo, é preferível à opção anterior anterior. No entanto haveria a hipótese de realizar pequenos aglomerados ao longo do país como um compromisso entre tempo e representatividade. No entanto os resultados indicam que aglomerados não escalam bem na plataforma. Portanto levou a decisão de selecionar pontos da totalidade do território.

6.4 Deteção de alterações florestais

Tendo executado os algoritmos com as parametrizações definidas na secção 4.5, chega a altura de comparar estes resultados com os dados de referência recolhidos tal como descrito na na secção 4.4. A análise será realizada de um ponto de vista mais genérico para o mais particular. Em primeiro lugar, serão comparadas as prestações dos algoritmos entre si; em segundo lugar, cada algoritmo será explorado individualmente de forma a compreender o impacto que os vários parâmetros têm no desempenho dos algoritmos; finalmente, será feita uma análise superficial com o objetivo de tentar compreender se é possível perceber quais os pontos fracos de cada análise olhando para séries temporais. A

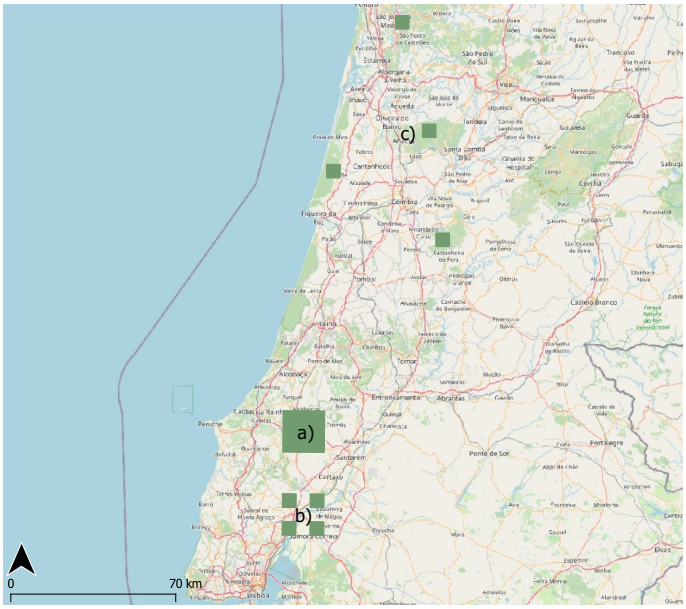


Figura 6.8: Áreas usadas no estudo da dispersão das áreas no tempo de execução.

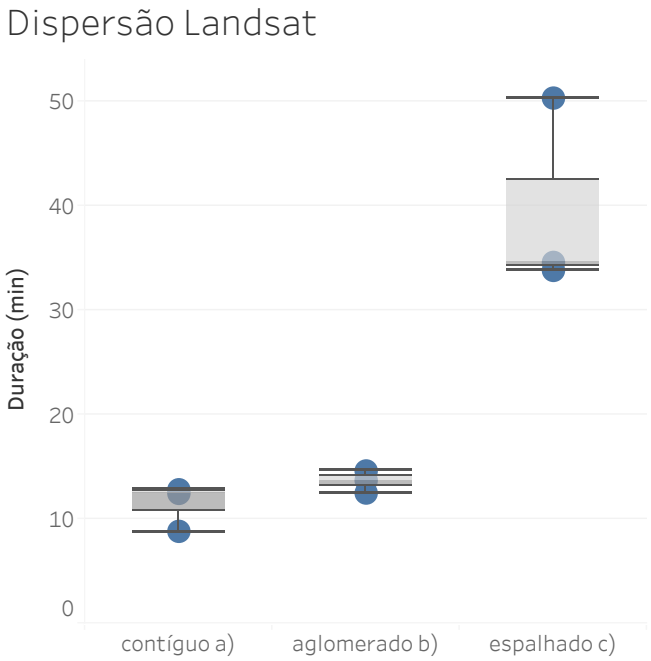


Figura 6.9: Gráfico com os tempos de execução do impacto da dispersão das áreas nos tempos de execução

Figura 6.10: Gráfico de barras com o número de experiências realizado por cada algoritmo.

análise realizada usa o conjunto de dados de validação que conta com 446 pontos.

6.4.1 Comparação do desempenho dos vários algoritmos

Para comparar os vários algoritmos, em primeiro lugar é necessário ter em mente que o número de experiências realizadas por cada algoritmo é diferente. Com o algoritmo BOCPD foi possível realizar mais experiências, isto porque o tempo de execução era de uma ordem de grandeza inferior quando comparado com os outros. Na imagem 6.10 é possível observar a distribuição do número de experiências realizadas por cada algoritmo e perceber que o BOCPD tem claramente um número muito superior de experiências realizadas.

Tendo em mente o número de experiências por algoritmo, é possível olhar agora para a distribuição da métrica F1 nos diferentes algoritmos na imagem 6.11. O BOCPD claramente destaca-se por ter obtido os melhores resultados e, curiosamente, também conta com um conjunto de resultados fora do normal onde o resultado do F1 é zero; posteriormente, será realizada uma análise mais cuidada para melhor compreender a origem deste fenómeno. O BOCPD Y apresenta resultados piores do que o BOCPD, o que seria de esperar tendo em conta que a resolução temporal a que trabalha é muito inferior. Ainda assim obtém melhores resultados do que o CCDC que usa todas as observações disponíveis. Curiosamente, o CCDC apresenta uma melhoria significativa quando a margem de erro é de um ano.Em último lugar, o LandTrendr apresenta os resultados mais baixos em

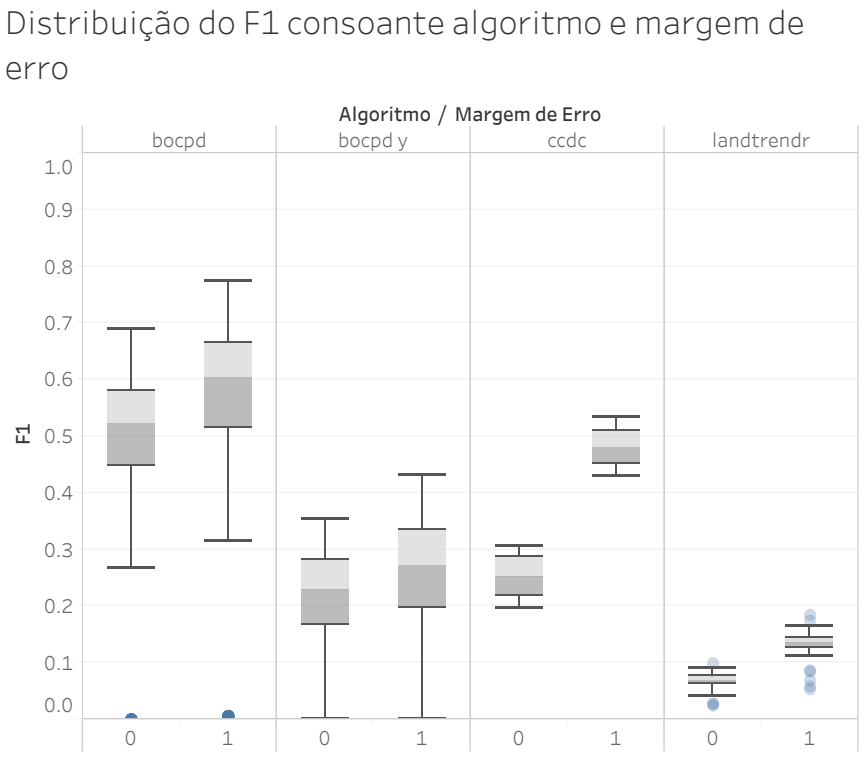


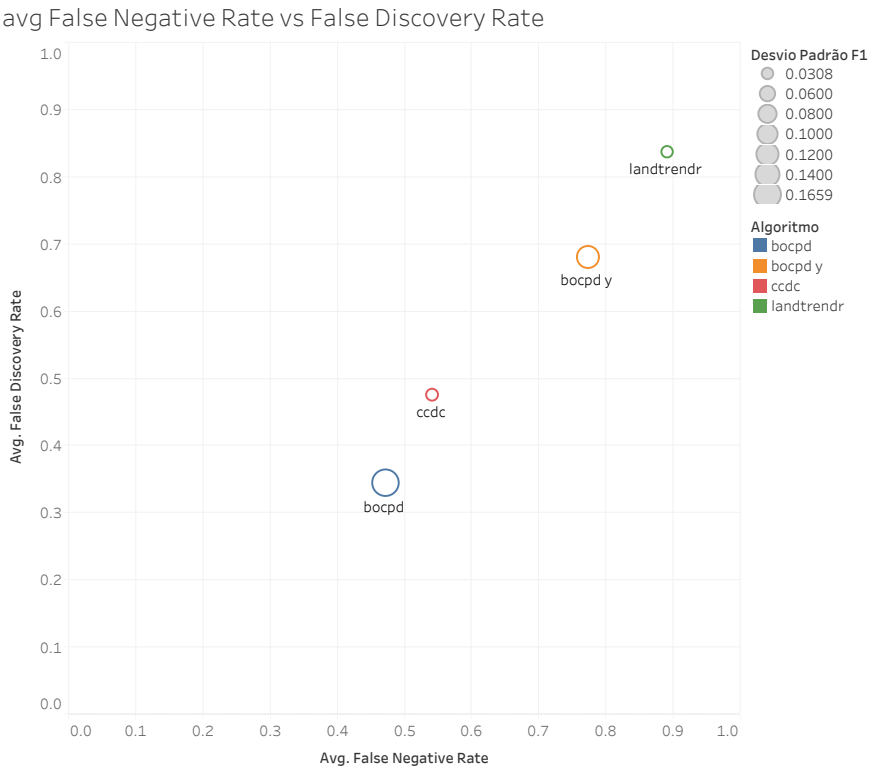
Figura 6.11: Gráfico com a distribuição da métrica F1 por algoritmo para as margens de erro 0 e 1 ano.

média para ambas as margens de erro.

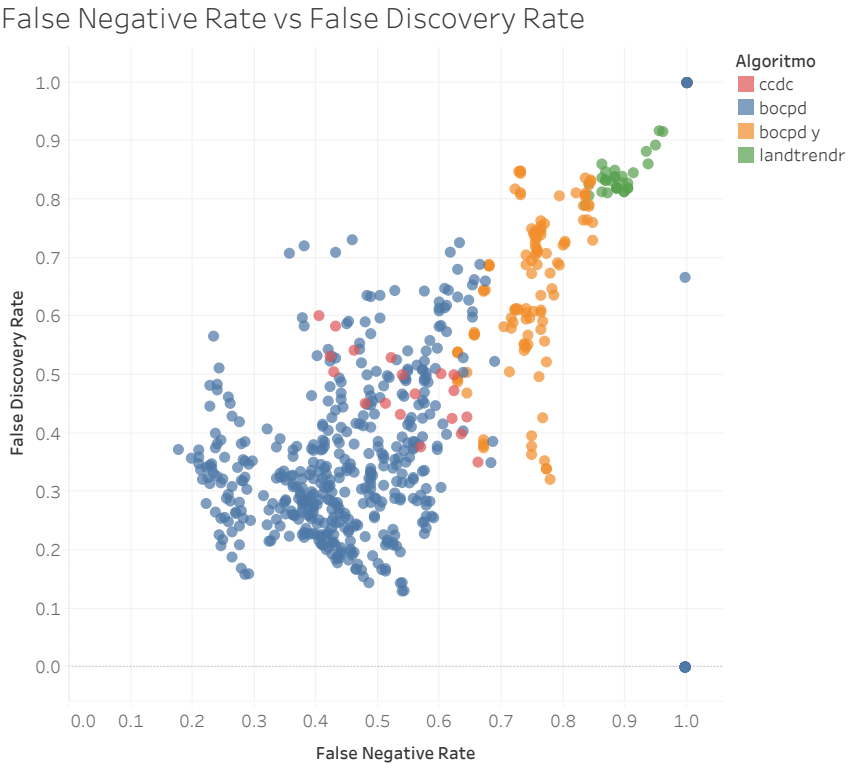
Olhando agora para uma representação diferente dos resultados na imagem 6.12 onde se dividiu o F1 nas suas duas componentes false discovery rate e false negative rate. A imagem 6.12(a) apresenta a média dos valores das métricas para cada algoritmo. É possível confirmar que em média o LandTrender é o pior nas duas métricas. Conclui-se também que não é possível destacar apenas uma métrica como responsável pelo melhor desempenho, pois todos os algoritmos apresentam um equilíbrio entre as duas componentes representadas.

Desagregando os resultados das experiências resulta a imagem 6.12(b). Aqui é possível confirmar que o LandTrendr se restringe ao canto superior direito. Já o BOCPD, devido ao elevado número de experiências, apresenta um grande número de pontos dispersos pelo gráfico; no entanto, é possível observar uma mancha significativa no quadrante inferior esquerdo. Uma análise mais focada no algoritmo ajudará a compreender se é possível atingir de forma mais consistente os melhores resultados. A próxima subsecção fará uma análise mais minuciosa destes resultados ao entrar em detalhe sobre a parametrização dos algoritmos.

Tendo uma ideia do desempenho relativo dos algoritmos, chega agora a altura de entrar em detalhe sobre cada um deles, de modo a compreender como a parametrização afeta os resultados.



(a) Média do false discovery rate e do false negative rate por algoritmo



(b) False discovery rate e false negative rate por experiência realizada

Figura 6.12: Duas visões dos resultados dos algoritmos testados com uma margem de erro de 1 ano.

6.4.2 Parametrização LandTrendr

Em primeiro lugar, o LandTrendr apresenta os piores resultados para a tarefa em questão; ainda assim, impõe-se a questão: será que é possível retirar algumas conclusões sobre a parametrização do algoritmo ?

Na imagem 6.13(a) é possível ver que os piores resultados são obtidos quando o **recovery threshold** tem o valor 0; adicionalmente, aparenta haver uma relação entre o **Pval Threshold** e o F1, sendo que são obtidos melhores resultados para valores mais altos do **Pval**. No entanto, não é claro que tipo de impacto estes valores têm nos resultados. É necessário um ponto de vista diferente.

Observando imagem 6.13(b) confirma-se que os resultados de um **Recovery Threshold** igual a 0 (representados com o icon de uma cruz) são os piores. Adicionalmente, torna-se mais clara a relação que o **Pval** tem com o desempenho do algoritmo, um **Pval** maior apresenta valores de False Negative Rate mais baixos. Já quando o parâmetro **Max Segments** assume o valor 6, apresenta valores ligeiramente mais altos de False Discovery Rate .

Ainda que o algoritmo não seja o melhor para a tarefa em questão, foi possível observar que, quando o **Recovery Threshold** toma o valor 0, apresenta os piores resultados. Conclui-se ainda que aumentar o **Pval** ajuda a reduzir o número de falsos negativos.

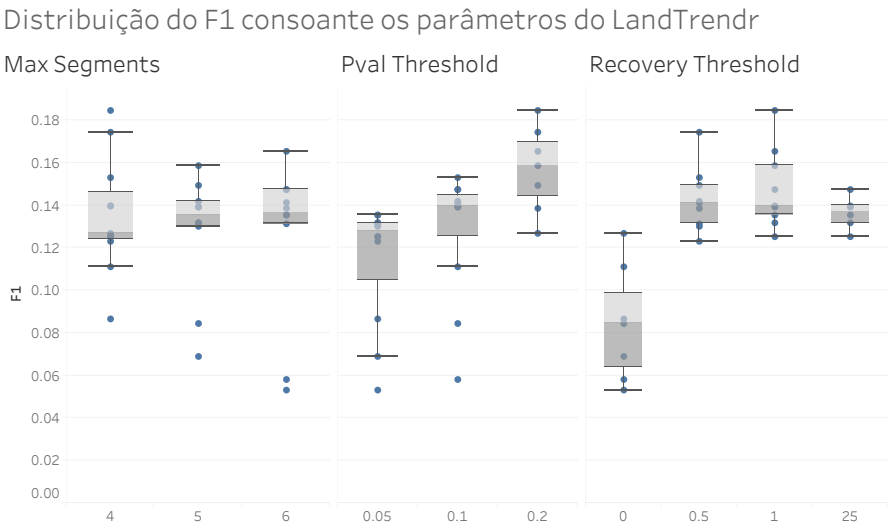
6.4.3 Parametrização Continuous Change Detection and Classification

O CCDC apresenta os segundos melhores resultados em média para uma margem de erro de um ano. Olhemos com mais atenção para este algoritmo de forma a compreender o impacto que a parametrização tem no desempenho. A imagem 6.14(b) apresenta o desempenho do algoritmo para um conjunto de parametrizações.

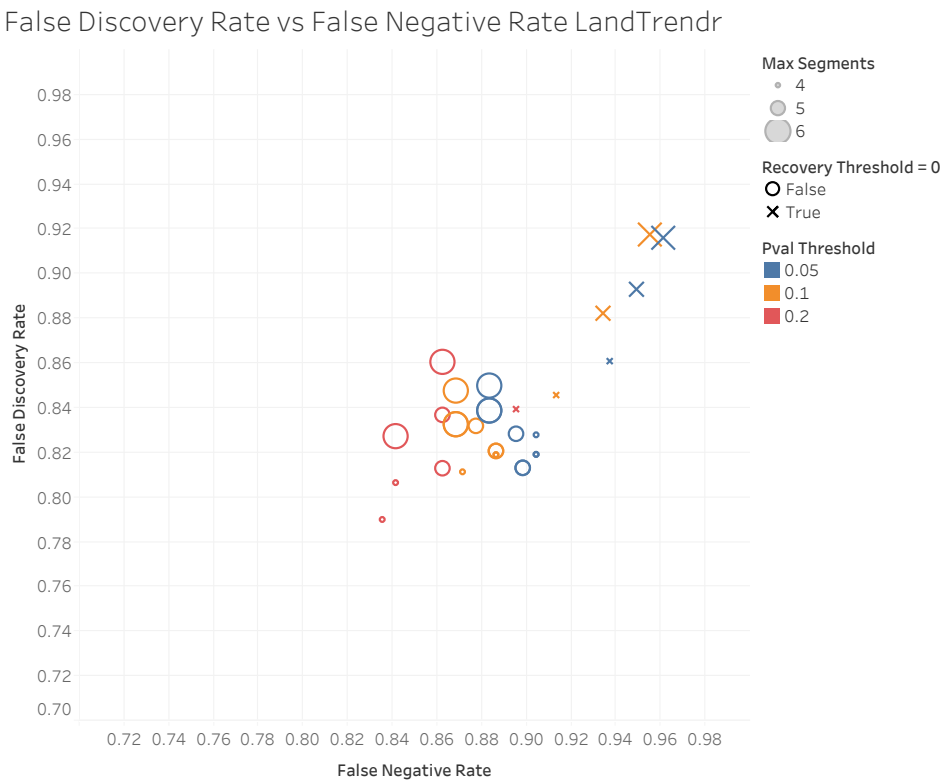
Em primeiro lugar, o parâmetro **min observations** salienta-se por ter a maior influência no desempenho, sendo que a escolha de um valor mais alto leva, por um lado, a um aumento do False Negative Rate e, por outro, a uma diminuição do False Discovery Rate. Este parâmetro controla o número de observações inesperadas até ser declarada uma alteração, pelo que faz sentido que o seu aumento leve a um menor número de deteções alteradas, e, consequentemente, maior False Negative Rate e menor False Discovery Rate.

Em segundo lugar, o parâmetro **min num of years**, para os dois valores testados, apresenta inequivocamente melhores resultados para o valor de 1.33. A utilização de 1.33 leva a uma diminuição significativa de False Discovery Rate aparentando gerar um ligeiro aumento no False Negative Rate.

Finalmente, o parâmetro **Chi Square Probability** tem um comportamento semelhante ao min observations mas com uma influência menor para os valores testados. Os valores mais altos de **Chi Square Probability** resultam em valores mais altos de False Negative Rate e valores mais baixos de False Discovery Rate.

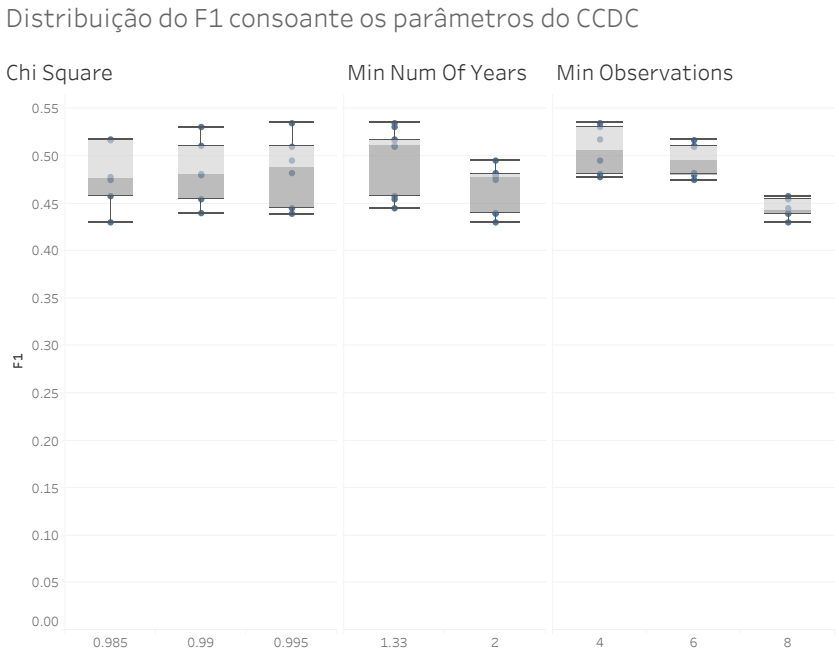


(a) Distribuição da métrica F1 para os três parâmetros testados.

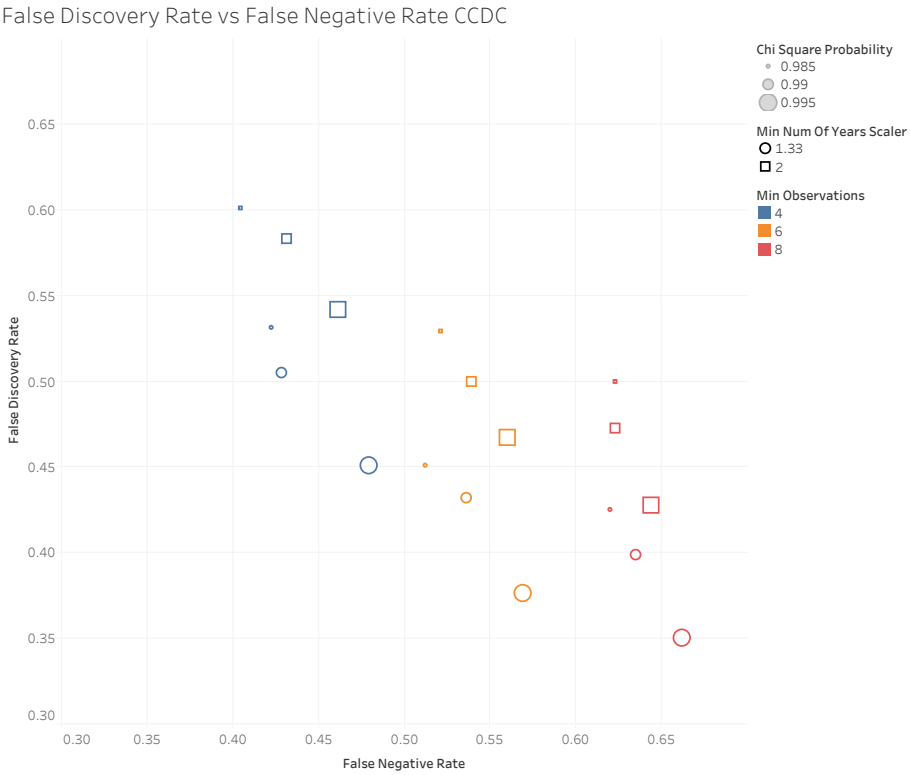


(b) False discovery rate e False Negative Rate por cada parametrização testada.

Figura 6.13: Resultados da avaliação do algoritmo LandTrendr



(a) Distribuição da métrica F1 para os três parâmetros testados.



(b) False discovery rate e False Negative Rate por cada parametrização testada.

Figura 6.14: Resultados da avaliação do algoritmo CCDC

6.4.4 Parametrização Bayesian Online Change Point Detection

Olhando para os resultados do BOCPD na imagem 6.15(a) é possível identificar que o parâmetro que mais influência tem nos resultados é o **resample**. Este parâmetro mostra melhores valores de F1 quando o **resample** usado é de 24 ou 32 dias. Para os parâmetros beta e índice é possível perceber que, quando estes tomam os valores 0.05 e NBR respetivamente, surgem outliers onde o F1 toma valores perto do zero. Para os restantes valores é difícil retirar conclusões perante esta imagem.

Tomando atenção à imagem 6.15(b) é possível ter uma ideia do impacto de cada parâmetro. Em primeiro lugar, olhando para o gráfico do parâmetro **alpha**, é possível observar que em média o valor de **alpha** 0.1 melhora os resultados obtidos quando comparado com 0.5, ainda que para um **resample** de 24 dias aumente o False Discovery Rate. Em segundo lugar, o parâmetro beta apresenta melhores resultados para 0.05 em média para todos os tipos de **resample**. Em terceiro lugar, o parâmetro **kw** apresenta comportamentos diferentes para os diferentes algoritmos; o mais fácil de observar é a diferença entre **resample** de 16 dias e sem qualquer resample: o primeiro beneficia de valores de **kw** mais altos enquanto que o segundo toma proveito de valores mais baixos. Desta forma, não é possível retirar uma conclusão abrangente deste parâmetro. Finalmente, o parâmetro **lam** apresenta as variações mais pequenas nos resultados, sendo que apresenta uma melhoria muito ligeira com o aumento do seu valor.

6.4.5 Parametrização Bayesian Online Change Point Detection Y

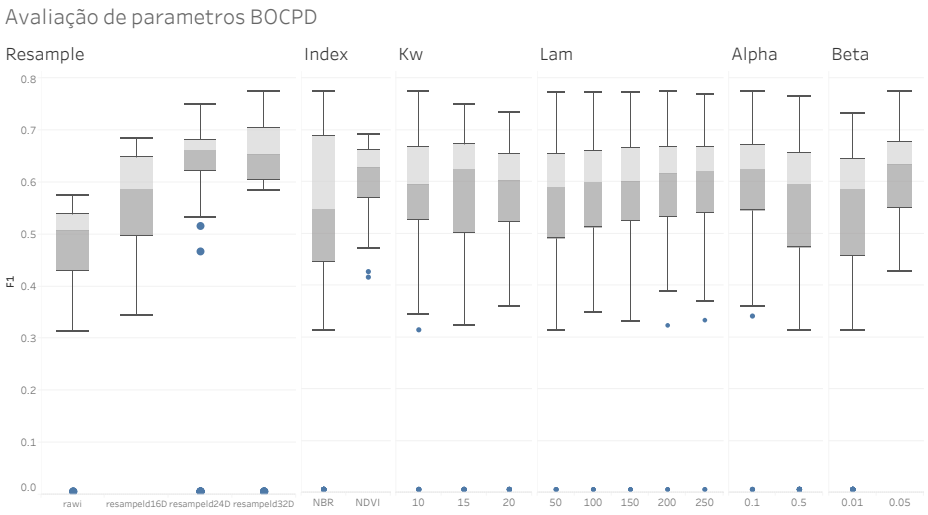
Em último lugar, temos novamente o BOCPD mas usando apenas uma imagem por ano, denominado como BOCPD Y. Verifica-se na imagem 6.16(a) que o parâmetro **kw** apresenta claras melhorias para valores mais baixos. Adicionalmente o **resample** que usa o máximo apresenta uma média similar à do mínimo mas os valores estão mais dispersos. É ainda possível confirmar que em média o algoritmo beneficia quando o **alpha** é 0.5 e o beta toma o valor 0.01. Já a prestação do o parâmetro **lam** aparenta melhorar com valores mais baixos, ainda que, à semelhança do que foi notado na subsecção 6.4.4 não haja um grande diferença entre os valores testados.

A imagem 6.16(b) apresenta as mesmas conclusões retiradas mas com uma pequena surpresa. A tendência de valores baixos de **kw** ajudarem no desempenho do algoritmo é quebrada para quando o **kw** toma o valor 1 e o **resample** usado é o mínimo anual.

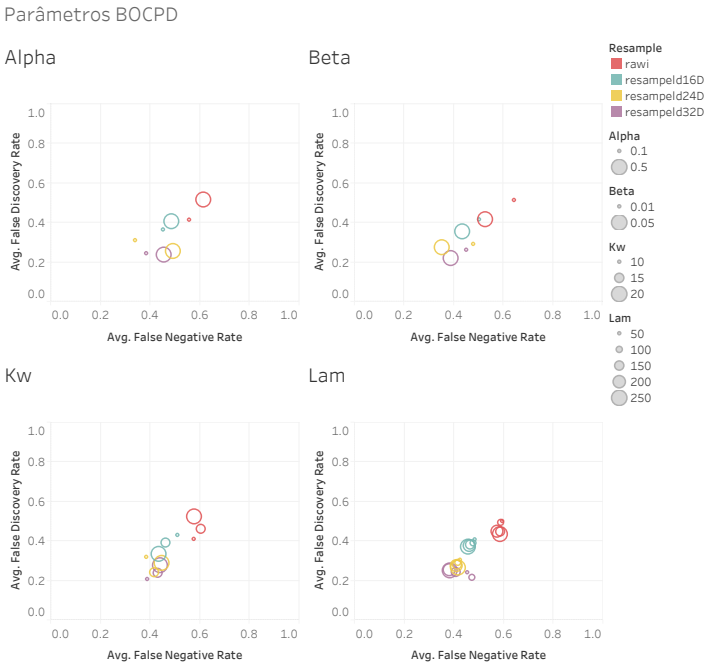
6.4.6 Conclusão

Tendo em conta os resultados apresentados não será difícil perceber que o BOCPD é o algoritmo certo para detetar alterações florestais. Mas qual a melhor parametrização? Tendo em conta os dados de validação, e usando o F1 como métrica, a melhor prestação foi obtida com a seguinte parametrização:

- resample 32dias

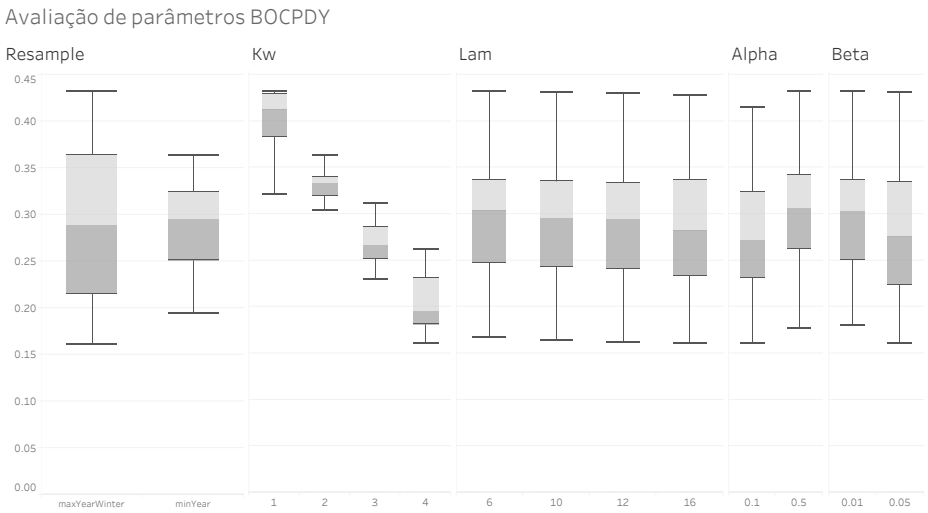


(a) Distribuição da métrica F1 para os seis parâmetros testados.

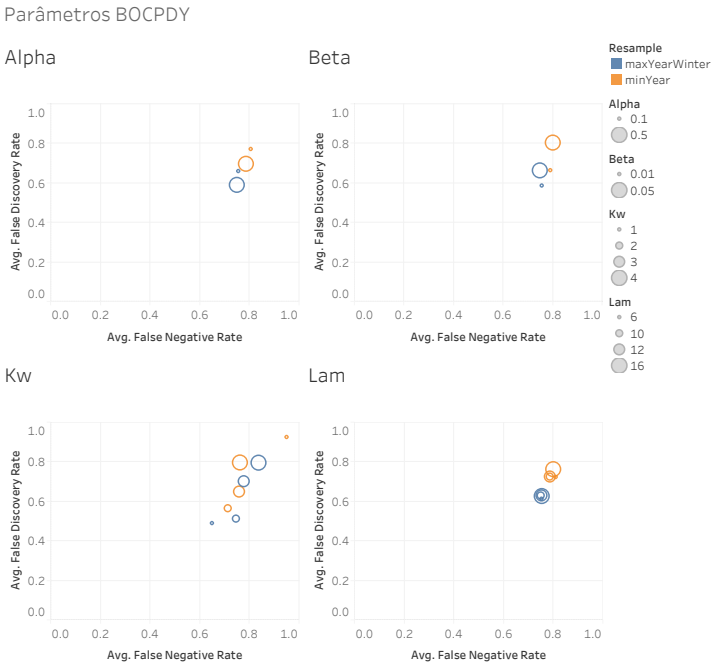


(b) Quatro diferentes visualizações da média do False discovery rate e média False Negative Rate para os parâmetros: **Alpha**, **Beta**, **Kw** e **Lam**. Cada ponto não representa uma experiência, representa a média dos resultados das experiências que enquadram a descrição do ponto.

Figura 6.15: Resultados da avaliação do algoritmo BOCPD



(a) Distribuição da métrica F1 para os cinco parâmetros testados.



(b) Quatro diferentes visualizações da média do False discovery rate e média False Negative Rate para os parâmetros: **Alpha**, **Beta**, **Kw** e **Lam**. Cada ponto não representa uma experiência, representa a média dos resultados das experiências que enquadram a descrição do ponto.

Figura 6.16: Resultados da avaliação do algoritmo BOCPD Y

Tabela 6.2: Comparação de métricas de teste e validação para a melhor parametrização do BOCPD

Métrica	Margem de Erro	Validação	Teste
F1	0	0.690	0.717
	1	0.774	0.778
FDR	0	0.256	0.175
	1	0.158	0.096
FNR	0	0.356	0.365
	1	0.284	0.317

- index NBR
- kw 10
- alpha 0.1
- beta 0.05
- lam 200

Este é o caso quer para quando a margem de erro é de 0 anos ou de 1 ano. Esta parametrização vai ao encontro das conclusões retiradas na subsecção 6.4.4. O único parâmetro para o qual não havia uma certeza do comportamento era o **kw**, que tomou o valor 10. É possível observar os resultados na tabela 6.2. Na tabela, é possível observar que os valores de validação estão alinhados com os valores de teste para as métricas apresentadas. É possível observar que o F1 aumenta ligeiramente para os valores de teste e que o False Disovery Rate diminui enquanto que o False Negative Rate aumenta. As pequenas diferenças entre validação e teste mostram que não está a ocorrer um sobre ajuste da parametrização ao conjunto de dados de validação.

6.5 Conclusão

Relativamente à comparação do COS com o IFN os resultados mostraram-se pouco sólidos havendo no melhor dos casos um acordo de 77%. Reforçando assim a dúvida nos dados de campo do IFN como já tinha sido apontado na secção 4.2.2.

Relativamente aos dados de referência recolhidos, foi possível constatar que 33% dos pontos originais foram declarados como invalidados. Este facto reduziu significativamente o conjunto de dados. O impacto da invalidação não foi igual para todas as espécies. Os sobreiros, azinheiras e pinheiros mansos foram alvo de uma maior percentagem de invalidação devido à menor densidade. No que diz respeito aos dados válidos não houve nenhuma surpresa de maior seguiram tudo o que era esperado.

No que se refere aos tempos de execução do GEE, para uma mesma zona contígua foi possível observar uma relação logarítmica entre a área a processar e o tempo de processamento. Dividir a área total de processamento em pequenos agrupamentos espalhados pelo território mostrou um aumento acentuado do tempo de processamento.

Finalmente relativamente aos algoritmos de deteção de alteração florestal, o BOCPD destacou-se pela positiva tendo obtido para os dados de teste um F1 de 0.717, bastante a cima dos restantes algoritmos. Este tipo de performance é bastante satisfatória.

7

Conclusões

7.1 Contribuições

Foi possível concluir que para a tarefa de detetar alterações na floresta o BOCPD é capaz de obter resultados muito bons quando comparado com outros algoritmos especializados em deteção remota. Aplicando um algoritmo genérico a um problema de deteção remota foi possível obter melhores resultados do que os obtidos através algoritmos específicos da área.

Foi gerado um novo dataset de forma manual que lista de forma exaustiva todas as alterações florestais existentes entre os anos 1986 e 2019. Para cada ponto há um registo de todas as alterações bem como a causa da alteração. Mais do que os pontos de alteração este registo é importante por ser sistemático e dessa forma ser possível encontrar os momentos sem qualquer tipo de alteração significativa e a principal razão para a criação do conjunto de dados.

Foi derivado um dataset único a partir dos dados fornecidos pelo ICNF sobre as áreas ardidas. Este dataset junta todas as áreas ardidas disponibilizadas pelo ICNF num só lugar e com uma só nomenclatura para uma análise mais conveniente.

Foi derivado um dataset de de alterações COS a partir das COS existentes. Para cada ponto do país é possível saber qual o tipo de uso e ocupação anterior e o posterior entre duas datas COS para as edições 1995, 2007, 2010 e 2015. Para além do processamento das interseções de polígonos também foi necessário ter em conta as diferente nomenclaturas e as correspondências entre elas.

7.2 Trabalho Futuro

Tendo agora uma visão completa do trabalho, chega a altura de olhar para o que está feito e compreender o que ficou por fazer. O trabalho futuro pode ser dividido em 3 distintos objetivos:

- Melhoramentos nos resultados da avaliação

- Melhoramentos na avaliação
- Gerar dados sobre o país

Se o objetivo for melhorar os resultados obtidos, ainda há espaço para procurar melhores parametrizações do algoritmo BOCPD. Nesta área destaca-se que, para os testes realizados, os parâmetros hazzard function e observation likelihood usaram sempre os mesmos modelos. Por exemplo, para o observation likelihood foi usada a distribuição t student; no entanto, outro tipo de distribuição poderá ser mais adequada para a tarefa em questão. Adicionalmente, o BOCPD apenas foi testado com séries temporais de índices como o NBR e NDVI. Dado que este algoritmo é capaz de analisar séries temporais com múltiplas variáveis, é possível analisar diretamente todas as bandas à semelhança do que o CCDC faz. Finalmente, olhando para além do BOCPD, Burg e Williams [5] apresentam um conjunto de algoritmos genéricos de segmentação de séries temporais genéricas que se podem mostrar úteis. A vantagem desta abordagem é que os autores fornecem acesso à ferramenta [1] usada para a realização do estudo. Desta forma, apenas é necessário formatar o conjunto de dados de distúrbios florestais aqui criado de forma a ser ingerido pela ferramenta de avaliação.

Se o objetivo for melhorar a avaliação realizada, é sempre possível adicionar mais pontos de referência ao conjunto de dados criado; no entanto, há diferentes formas de aprimorar o processo de avaliação. Seria interessante realizar experiências em zonas de floresta gerida com registos completos de abate, incêndio e limpezas, à semelhança do que é realizado por Brooks et al. [4]. Esta abordagem permitirá compreender como o algoritmo lida com uma área onde o comportamento deverá ser idêntico. Por outro lado, seria interessante obter dados para outras zonas do globo com diferentes tipos de vegetação e compreender se os resultados aqui obtidos se mantêm em diferentes condições.

Finalmente, se o objetivo for usar as conclusões sobre os algoritmos para construir um mapa de distúrbios florestais em Portugal, será ainda necessário implementar o BOCPD de forma eficiente. Uma forma de atingir este objetivo seria implementar o algoritmo no GEE, sendo que o uso desta plataforma também possibilitaria o acesso ao poder de computação e aos dados que disponibiliza.

[1] Repositório da fermenta de avaliação de séries temporais https://github.com/alan-turing-institute/TCPDBench

Bibliografia

[1] *About the launch*. en. URL: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/About_the_launch (acedido em 15/09/2021) (ver p. 16).

[2] R. P. Adams e D. J. C. MacKay. "Bayesian Online Changepoint Detection". Em: *arXiv:0710.3742 [stat]* (out. de 2007). arXiv: 0710.3742 version: 1. URL: http://arxiv.org/abs/0710.3742 (acedido em 26/08/2021) (ver p. 25).

[3] *Aqua Earth-observing satellite mission | Aqua Project Science*. URL: https://aqua.nasa.gov/ (acedido em 16/02/2020) (ver p. 15).

[4] E. B. Brooks et al. "On-the-Fly Massively Multitemporal Change Detection Using Statistical Quality Control Charts and Landsat Data". Em: *IEEE Transactions on Geoscience and Remote Sensing* 52.6 (jun. de 2014), pp. 3316–3332. ISSN: 1558-0644. DOI: 10.1109/TGRS.2013.2272545 (ver pp. 22, 84).

[5] G. J. J. v. d. Burg e C. K. I. Williams. "An Evaluation of Change Point Detection Algorithms". Em: *arXiv:2003.06222 [cs, stat]* (mai. de 2020). arXiv: 2003.06222. URL: http://arxiv.org/abs/2003.06222 (acedido em 15/03/2021) (ver pp. 24, 26–28, 37, 41, 46, 84).

[6] J. B. Campbell e R. H. Wynne. *Introduction to Remote Sensing, Fifth Edition*. en. Google-Books-ID: NkLmDjSS8TsC. Guilford Press, jun. de 2011. ISBN: 978-1-60918-177-2 (ver p. 12).

[7] *Cartografia de Uso e Ocupação do Solo (COS, CLC e Copernicus) | DGT*. URL: https://www.dgterritorio.gov.pt/cartografia/cartografia-tematica/COS-CLC-COPERNICUS (acedido em 29/07/2021) (ver pp. 10, 32).

[8] W. B. Cohen, Z. Yang e R. Kennedy. "Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation". en. Em: *Remote Sensing of Environment* 114.12 (dez. de 2010), pp. 2911–2924. ISSN: 00344257. DOI: 10.1016/j.rse.2010.07.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0034425710002269 (acedido em 02/06/2021) (ver p. 37).

[9] W. B. Cohen et al. "How similar are forest disturbance maps derived from different Landsat time series algorithms?" en. Em: *Forests. 8: 98.* 8 (2017), p. 98. DOI: 10.3390/f8040098. URL: https://www.fs.usda.gov/treesearch/pubs/54976 (acedido em 22/01/2020) (ver pp. 12, 22, 24, 27, 28, 31, 37, 41, 42).

[10] *CORINE Land Cover — Copernicus Land Monitoring Service*. en. Land Section. URL: https://land.copernicus.eu/pan-european/corine-land-cover (acedido em 29/07/2021) (ver p. 11).

[11] E. P. Crist e R. C. Cicone. "A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap". Em: *IEEE Transactions on Geoscience and Remote sensing* 3 (1984), pp. 256–263 (ver p. 16).

[12] *Dados de base de 1995-98 — ICNF*. pt. Página. URL: http://www2.icnf.pt/portal/florestas/ifn/ifn4/dad-base95-98 (acedido em 17/02/2020) (ver p. 9).

[13] S. Delwart. "ESA Standard Document". en. Em: 1 (), p. 64 (ver p. 16).

[14] *Despacho 10248/2018, 2018-11-06*. pt. URL: https://dre.pt/web/guest/home/-/dre/116894722/details/4/maximized?serie=II&at=c&print_preview=print-preview&day=2018-11-06&date=2018-11-01&dreId=116858055 (acedido em 19/01/2020) (ver p. 11).

[15] *DGTerritório - Carta de Uso e Ocupação do Solo de Portugal Continental (COS)*. URL: http://www.dgterritorio.pt/dados_abertos/cos/ (acedido em 18/02/2020) (ver pp. 2, 3, 6).

[16] P. M. L. Drezet e S. Quegan. "Satellite-based radar mapping of British forest age and Net Ecosystem Exchange using ERS tandem coherence". en. Em: *Forest Ecology and Management* 238.1 (jan. de 2007), pp. 65–80. ISSN: 0378-1127. DOI: 10.1016/j.foreco.2006.09.088. URL: http://www.sciencedirect.com/science/article/pii/S0378112706009236 (acedido em 30/12/2019) (ver p. 21).

[17] D.-D.-G. d. A. Económicas. *Indústrias da Fileira Florestal*. PT. URL: https://www.dgae.gov.pt/servicos/politica-empresarial/setores-industriais/industrias-de-base-florestal.aspx (acedido em 28/01/2020) (ver p. 2).

[18] *Fact Sheet*. en. Fact Sheet. 2018. URL: https://www.usgs.gov/faqs/what-are-landsat-collection-1-level-1-data-product-file-sizes?qt-news_science_products=0#qt-news_science_products (acedido em 28/01/2020) (ver p. 2).

[19] M. Fiorella e W. J. Ripple. "Analysis of conifer forest regeneration using Landsat Thematic Mapper data". en. Em: (jan. de 1995). ISSN: 0099-1112. URL: http://ntrs.nasa.gov/search.jsp?R=19950017681 (acedido em 05/02/2020) (ver pp. 19, 20).

[20] M. Fiorella e W. J. Ripple. "Determining successional stage of temperate coniferous forests with Landsat satellite data". Em: (1995) (ver p. 16).

[21] N. Gorelick et al. "Google Earth Engine: Planetary-scale geospatial analysis for everyone". en. Em: *Remote Sensing of Environment*. Big Remotely Sensed Data: tools, applications and experiences 202 (dez. de 2017), pp. 18–27. ISSN: 0034-4257. DOI: 10.1016/j.rse.2017.06.031. URL: http://www.sciencedirect.com/science/article/pii/S0034425717302900 (acedido em 02/03/2020) (ver pp. 53, 55).

[22] M. E. Harmon, W. K. Ferrell e J. F. Franklin. "Effects on Carbon Storage of Conversion of Old-Growth Forests to Young Forests". en. Em: *Science, New Series* 247.4943 (1990), pp. 699–702. URL: http://www.jstor.org/stable/2873679 (ver p. 1).

[23] Y. Heymann, ed. *CORINE land cover: guide technique*. fre. EUR 12585. OCLC: 59839202. Luxembourg: Office des publ. officiellles des Communautés Européennes, 1993. ISBN: 978-92-826-2579-8 (ver p. 11).

[24] *IFN6 — ICNF*. pt. Página. URL: http://www2.icnf.pt/portal/florestas/ifn/ifn6 (acedido em 28/01/2020) (ver pp. 1, 6, 7, 9).

[25] *Inventário Florestal Nacional — ICNF*. pt. Página. URL: http://www2.icnf.pt/portal/florestas/ifn (acedido em 26/01/2020) (ver pp. 2, 3, 8).

[26] *Jupyter Notebook Viewer*. URL: https://nbviewer.jupyter.org/github/hildensia/bayesian_changepoint_detection/blob/master/Example%20Code.ipynb (acedido em 18/09/2021) (ver pp. 44, 45).

[27] R. E. Kennedy, Z. Yang e W. B. Cohen. "Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms". en. Em: *Remote Sensing of Environment* 114.12 (dez. de 2010), pp. 2897–2910. ISSN: 0034-4257. DOI: 10.1016/j.rse.2010.07.008. URL: http://www.sciencedirect.com/science/article/pii/S0034425710002245 (acedido em 22/01/2020) (ver pp. 22, 23, 27, 28, 37, 42).

[28] R. E. Kennedy, Z. Yang e W. B. Cohen. "Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms". en. Em: *Remote Sensing of Environment* 114.12 (dez. de 2010), pp. 2897–2910. ISSN: 0034-4257. DOI: 10.1016/j.rse.2010.07.008. URL: http://www.sciencedirect.com/science/article/pii/S0034425710002245 (acedido em 22/01/2020) (ver pp. 24, 25, 42).

[29] D. S. Kimes et al. "Extracting forest age in a Pacific Northwest forest from Thematic Mapper and topographic data". en. Em: *Remote Sensing of Environment* 56.2 (mai. de 1996), pp. 133–140. ISSN: 0034-4257. DOI: 10.1016/0034-4257(95)00230-8. URL: http://www.sciencedirect.com/science/article/pii/0034425795002308 (acedido em 07/02/2020) (ver p. 20).

[30] M. Köhl, S. Magnussen e M. Marchetti. *Sampling methods, remote sensing and GIS multiresource forest inventory: with 27 tables*. en. Tropical forestry. OCLC: 180902351. Berlin: Springer, 2006. ISBN: 978-3-540-32571-0 (ver pp. 12, 19).

[31] J. Kulick. *Bayesian Changepoint Detection*. original-date: 2014-04-11T08:29:58Z. Jul. de 2021. URL: https://github.com/hildensia/bayesian_changepoint_detection (acedido em 29/07/2021) (ver pp. 41, 58).

[32] *Landsat 9 « Landsat Science*. URL: https://landsat.gsfc.nasa.gov/landsat-9/ (acedido em 28/05/2021) (ver p. 14).

[33] T. Lillesand, R. W. Kiefer e J. Chipman. *Remote Sensing and Image Interpretation*. en. Google-Books-ID: AFHDCAAAQBAJ. John Wiley & Sons, fev. de 2015. ISBN: 978-1-118-34328-9 (ver p. 13).

[34] M. J. López García e V. Caselles. "Mapping burns and natural reforestation using Thematic Mapper data". Em: *Geocarto International* 6 (mar. de 1991), pp. 31–37. DOI: 10.1080/10106049109354290 (ver p. 16).

[35] 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).

[36] *Mapas — ICNF*. pt. Página. URL: http://www2.icnf.pt/portal/florestas/dfci/inc/mapas (acedido em 26/01/2020) (ver pp. 3, 9).

[37] M. A. McCarthy, A. Malcolm Gill e D. B. Lindenmayer. "Fire regimes in mountain ash forest: evidence from forest age structure, extinction models and wildlife habitat". en. Em: *Forest Ecology and Management* 124.2 (dez. de 1999), pp. 193–203. ISSN: 0378-1127. DOI: 10.1016/S0378-1127(99)00066-3. URL: http://www.sciencedirect.com/science/article/pii/S0378112799000663 (acedido em 19/02/2020) (ver p. 1).

[38] *MODIS Web*. URL: https://modis.gsfc.nasa.gov/about/ (acedido em 16/02/2020) (ver p. 15).

[39] S. L. Powell et al. "Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches". en. Em: *Remote Sensing of Environment* 114.5 (mai. de 2010), pp. 1053–1068. ISSN: 0034-4257. DOI: 10.1016/j.rse.2009.12.018. URL: http://www.sciencedirect.com/science/article/pii/S0034425709003745 (acedido em 17/02/2020) (ver p. 17).

[40] QGIS Development Team. *QGIS Geographic Information System*. QGIS Association. 2021. URL: https://www.qgis.org (ver p. 59).

[41] Qingsheng Liu et al. "A tasseled cap transformation for Landsat 8 OLI TOA reflectance images". en. Em: *2014 IEEE Geoscience and Remote Sensing Symposium*. Quebec City, QC: IEEE, jul. de 2014, pp. 541–544. ISBN: 978-1-4799-5775-0. DOI: 10.1109/IGARSS.2014.6946479. URL: http://ieeexplore.ieee.org/document/6946479/ (acedido em 17/09/2021) (ver p. 17).

[42] D.-G. do Territóri et al. "Especificações Técnicas da Carta de Uso e Ocupação do Solo (COS) de Portugal Continental". pt. Em: (2018), p. 103. URL: https://www.dgterritorio.gov.pt/sites/default/files/documentos-publicos/ET-COS-1995-2007-2010-2015.pdf (ver pp. 32, 50, 51).

[43] *Territórios ardidos*. URL: https://sig.icnf.pt/portal/home/item.html?id=983c4e6c4d5b4666b258a3ad5f3ea5af (acedido em 29/07/2021) (ver p. 34).

[44] K. Thome. *About Terra | Terra*. URL: https://terra.nasa.gov/about (acedido em 16/02/2020) (ver p. 15).

[45] B. Tso e P. M. Mather. *Classification methods for remotely sensed data*. en. Second edition. Environmental engineering. OCLC: 845399077. Boca Raton, Fla. London New York: CRC Press, Taylor & FrancisGroup, 2009. ISBN: 978-1-4200-9072-7 (ver p. 13).

[46] C. Uhl, R. Buschbacher e E. A. S. Serrao. "Abandoned Pastures in Eastern Amazonia. I. Patterns of Plant Succession". Em: *Journal of Ecology* 76.3 (1988), pp. 663–681. ISSN: 0022-0477. DOI: 10.2307/2260566. URL: https://www.jstor.org/stable/2260566 (acedido em 19/02/2020) (ver p. 1).

[47] Xiaoyang Zhang et al. "MODIS tasseled cap transformation and its utility". en. Em: *IEEE International Geoscience and Remote Sensing Symposium*. Vol. 2. Toronto, Ont., Canada: IEEE, 2002, pp. 1063–1065. ISBN: 978-0-7803-7536-9. DOI: 10.1109/IGARSS.2002.1025776. URL: http://ieeexplore.ieee.org/document/1025776/ (acedido em 17/09/2021) (ver p. 17).

[48] N. E. Young et al. "A survival guide to Landsat preprocessing". en. Em: *Ecology* 98.4 (2017), pp. 920–932. ISSN: 1939-9170. DOI: 10.1002/ecy.1730. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.1730 (acedido em 11/02/2020) (ver pp. 14, 15, 35).

[49] Y. Zhang et al. "Mapping spatial distribution of forest age in China". en. Em: *Earth and Space Science* 4.3 (2017), pp. 108–116. ISSN: 2333-5084. DOI: 10.1002/2016EA000177. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016EA000177 (acedido em 30/12/2019) (ver p. 21).

[50] 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 (mar. de 2014), pp. 152–171. ISSN: 0034-4257. DOI: 10.1016/j.rse.2014.01.011. URL: http://www.sciencedirect.com/science/article/pii/S0034425714000248 (acedido em 10/01/2020) (ver p. 22).

I

Ficheiros Importantes e Formatos

I.1 Pontos de Referência

Ficheiro com o conjunto de pontos de referência e informações adicionais para serem importados para o GEE. O ficheiro foi usado para extrair séries temporais do GEE e como fonte de informação para a avaliação manual dos pontos de referência. O ficheiro encontra-se no formato shapefile e é possível observar na tabela I.1 os atributos com as respetivas descrições e exemplos.

Tabela I.1: Lista de atributos e respetiva descrição do ficheiro de pontos de referência

Atributo	Descrição	Exemplo
point_id	Identificador único do ponto	112
test	Ponto de teste(1) ou de treino(0)	0
fire_ids	Lista de identificadores de fogos no ponto	[]
fire_years	Lista de anos dos fogos que afetaram o ponto	[]
cos_ids	Lista de identificadores cos dos polígonos cos	[1555535, 853303, 693804, 1619582]
cos_uses	Lista de uso de solo definido nos polígonos cos	["3.1.1.01.5 Florestas de eucalipto", "3.2.4.10.5 Novas ...
clc_ids	Lista de identificadores clc dos polígonos	[145216, 153315, 153544, 155212]
clc_uses	Lista de uso de solo definido nos polígonos clc	["Florestas mistas", "Florestas mistas", "Florestas ...
ifn_ids	Lista de identificadores ifn ao qual o ponto pertence	[853651, 1128715, 609588, 229624]
ifn_ususol	Lista de uso de solo definido pelo ifn	["Floresta", "Floresta", "Floresta", "Floresta"]
inf_ocuppr	Lista de ocupação principal definida pelo ifn	["Eucaliptos", "Eucaliptos", "Eucaliptos", "Eucaliptos"]
inf_tipo	Lista do tipo de ocupação definida pelo ifn	["Povoamento em pé", "Povoamento em pé", ...
inf_percco	Lista da percentagem do coberto definida pelo ifn	["[70, 80[%", "[70, 80[%", "[70, 80[%", "[60, 70[%"]
inf_sobcob	Lista da percentagem do sobcoberto definida pelo ifn	["não identificável", "não identificável", "não ...
inf_dimens	Lista da dimensão da mancha do povoamento definida pelo ifn	["[10; 50[ha", "[10; 50[ha", "[10; 50[ha", "[10; 50[ha"]
climate_re	Região climática à qual o ponto pertence	alentejo
district	Distrito ao qual o ponto pertence	Santarém
concelho	Concelho ao qual o ponto pertence	Santarém

I.2 Séries temporais

Ficheiro das séries temporais extraídas do GEE a partir dos pontos de referência. O script que gerou estas séries temporais é capaz de fazer séries temporais a partir de polígonos, pelo que os atributos que terminam em "_count"não trazem grande informação quando são extraídos pontos. O ficheiro encontra-se no formato csv e é possível observar na tabela I.2 os atributos com as respetivas descrições e exemplos.

Tabela I.2: Lista de atributos, descrição e exemplos do ficheiro de séries temporais

Atributo	Descrição	Exemplo
system:index	Índex interno ao GEE único para cada linha	1_1_1_2_LT04_203031_19881209_00...
DateL	Data da observação	1988-12-09T10:39:21
NBR_count	Número de pixeis de NBR na observação (sempre 1)	1
NBR_first	Valor do NBR da observação	-0.30487133006505057
NDVI_count	Número de pixeis de NDVI na observação (sempre 1)	1
NDVI_first	Valor do NDVI da observação	0.09762431455998122
Satelite	Satélite usado para recolher a observação	LANDSAT_4
Scene_identifier	Identificador único para a imagem de satélite	LT04_L1TP_203031_19881209_20170205_01_T1
TCAngle_count	Número de pixeis de TCAngle na observação (sempre 1)	1
TCAngle_first	Valor do TCAngle da observação	0.06577225512522482
TCBrightness_count	Número de pixeis de TCBrightness na observação (sempre 1)	1
TCBrightness_first	Valor do TCBrightness da observação	0.3398624062538147
TCGreenness_count	Número de pixeis de TCGreenness na observação (sempre 1)	1
TCGreenness_first	Valor do TCGreenness da observação	0.0223535168915987
TCWetness_count	Número de pixeis de TCWetness na observação (sempre 1)	1
TCWetness_first	Valor do TCWetness da observação	-0.23822960257530212
Timestamp	Timestamp UNIX em segundos da imagem	597667161
point_id	Identificador único do ponto observado	806
.geo	Atributo inerente à exportação GEE (sempre vazio)	

I.3 Classificação de pontos de quebra

Ficheiro obtido após a classificação de pontos de quebra por parte dos algoritmos de deteção dos mesmos. Cada linha representa um ponto geográfico e identifica os anos onde foram detetados pontos de quebra. O ficheiro encontra-se no formato csv e é possível observar na tabela I.3 os atributos com as respetivas descrições e exemplos.

Tabela I.3: Lista de atributos, descrição e exemplos do ficheiro de classificação de pontos de quebra

Atributo	Descrição	Exemplo
`point_id`	Identificador do ponto	8
year	Lista de anos onde foram detetadas quebras	[1997, 1999, 2001]

I.4 Avaliação agregada de algoritmos de deteção de pontos de quebra

Ficheiro resultante da avaliação de um algoritmo de deteção de pontos de quebra. Cada linha do ficheiro representa uma parametrização do algoritmo tendo consigo a indicação de qual a parametrização usada. As colunas das parametrizações são especificas ao ficheiro de cada algoritmo. Cada linha conta com um conjunto de métricas de avaliação. O ficheiro encontra-se no formato csv e é possível observar na tabela I.4 os atributos com as respetivas descrições e exemplos.

Tabela I.4: Lista de atributos, descrição e exemplos do ficheiro de classificação de avaliação agregada de algoritmos de deteção de pontos de quebra

Atributo	Descrição	Exemplo
id	Identificador da experiência	genericAlg_p1_NDVI_p2_10
algorithm	Nome do algoritmo	genericAlg
param_1	Primeiro parâmetro (exemplo)	NDVI
param_2	Segundo parâmetro (exemplo)	10
slack	Margem de erro permitida em anos	1
accuracy	Métricas de avaliação	0.964789902
balanced_accuracy		0.65322089
f1		0.3375
f2		0.328867235
false_discovery_rate		0.647058824
false_negative		226
false_negative_rate		0.676646707
false_positive		198
false_positive_rate		0.016911513
kappa		0.319449909
matthews_correlation_coefficient		0.319773008
negative_predictive_value		0.980743013
positive_predictive_value		0.352941176
true_negative		11510
true_negative_rate		0.983088487
true_positive		108
true_positive_rate		0.323353293

I.5 Avaliação desagregada de algoritmos de deteção de pontos de quebra

Ficheiro com uma linha por ponto por cada ano desde 1990 até 2016. Neste nível de detalhe é possível diferenciar todas as classificações anuais. Em cada linha há informações respetivas à parametrização do algoritmo que a gerou, pelo que todos os algoritmos contam com um conjunto diferente de colunas com informações sobre parametrização. O ficheiro encontra-se no formato csv e é possível observar na tabela I.5 os atributos com as respetivas descrições e exemplos.

Tabela I.5: Lista de atributos, descrição e exemplos do ficheiro de classificação de avaliação desagregada de algoritmos de deteção de pontos de quebra.

Atributo	Descrição	Exemplo
point_id	Identificador único do ponto	50
year	Ano da avaliação	2004
cause	Causa de quebra de referência (pode ser nulo)	fire
reference_year	Ano do ponto de referência (pode ser nulo)	2004
result_year	Ano de classificação de quebra (pode ser nulo)	NULL
delta	Diferença entre reference_year e result_year (pode ser nulo)	NULL
has_reference	Há quebra de referência	TRUE
has_result	Há quebra classificada	FALSE
true_positive	Verdadeiro Positivo	FALSE
false_positive	Falso Positivo	FALSE
false_negative	Falso Negativo	TRUE
true_negative	Verdadeiro Negativo	FALSE
id	Identificador da experiência	genericAlg_p1_NDVI_p2_10
algorithm	Nome do algoritmo	genericAlg
param_1	Primeiro parâmetro (exemplo)	NDVI
param_2	Segundo parâmetro (exemplo)	10
slack	Margem de erro permitida em anos	1