



**André Miguel Neves**

Licenciado em Engenharia Informática

## **Deteção Remota de Estruturas Artificiais Permanentes**

Dissertação para obtenção do Grau de Mestre em  
**Engenharia Informática**

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

Co-orientador: João Carlos Gomes Moura Pires,  
Professor Associado,  
Faculdade de Ciências e Tecnologia  
Universidade Nova de Lisboa

Júri

Presidente: Doutor Pedro Abílio Duarte de Medeiros  
Arguente: Doutora Ana Cláudia Moreira Teodoro  
Vogal: Doutor Carlos Augusto Isaac Piló Viegas Damásio



FACULDADE DE  
CIÊNCIAS E TECNOLOGIA  
UNIVERSIDADE NOVA DE LISBOA

**Dezembro, 2019**



## **Deteção Remota de Estruturas Artificiais Permanentes**

Copyright © André Miguel Neves, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade NOVA de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.



*Dedico esta tese à minha mãe, que sacrificou muitas vezes os  
seus desejos pelo meu sucesso.*



## AGRADECIMENTOS

Em primeiro lugar, quero agradecer aos meus orientadores, Professor Carlos Viegas Damásio e o Professor João Moura Pires, por toda a ajuda e orientação durante a elaboração desta dissertação. Dissertação esta financiada por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/CEC/04516/2019 (NOVA LINCS).

Agradeço também ao Professor Fernando Birra e todo o grupo da NOVA LINCS no projeto MORENA, em especial aos meus colegas de investigação: Ricardo Afonso, João Albuquerque, André Lopes e Igor Ruivo. Especialmente ao colega Ricardo Afonso, por me ter literalmente salvo a vida durante uma sessão de trabalho. Um grande obrigado à Faculdade de Ciências e Tecnologia (Universidade Nova de Lisboa), pela oportunidade que me disponibilizou com este curso e também pela fantástica investigação que aqui é feita. Estendo este agradecimento aos meus colegas de curso, Daniel Henriques, Filipe Amador, Ivo Rocha, Didier Dias, Dinis Cabanas e Pedro Almeida, por nos termos acompanhado uns aos outros desde o início e pelo fantástico espírito de camaradagem.

Finalmente, mas não menos importante, quero agradecer à minha família mais próxima, aos meus pais e à minha irmã, pelo tremendo suporte durante os meus cinco anos de estudo e por acreditarem em mim durante esta etapa. Agradeço também à minha namorada, Tatiana Barrigana, por puxar por mim durante o meu percurso académico e por me dar motivação para dar o meu melhor em todos os meus projetos.





## RESUMO

---

Os Incêndios florestais, principalmente em zonas rurais, são uma ameaça muito significativa à qual temos de ter em conta no que toca à proteção de vidas e estruturas. Para uma boa proteção a nível nacional é necessário minimizar o impacto dos incêndios. Esta minimização do risco e do impacto dos incêndios florestais obriga à monitorização das faixas de gestão de combustível (FGC), na envolvente a habitações, estradas nacionais entre outros. Assim, é essencial a identificação das estruturas artificiais permanentes a partir das quais se definem as faixas de gestão de combustível. As cartas existentes que mapeiam estas estruturas possuem uma frequência de geração baixa e a sua resolução poderá não ser adequada para a extração de pequenas estruturas, como pequenos aglomerados ou estradas, que carecem de proteção contra incêndios. Esta investigação pretende então, com base nas imagens dos satélites Sentinel 1 e 2, mapas de elevação de terreno, e dados de referência da carta COS 2015 juntamente com informação OpenStreetMap, recorrer a algoritmos de aprendizagem automática para a deteção de estruturas artificiais permanentes com base nas características de refletância e de índices ao longo do tempo. No final foi concluindo que a utilização de séries temporais e modelos de elevação de terreno, juntamente com o modelo de classificação XGBoost, apresentam os melhores resultados em todas as classificações, classificações estas que possuem uma melhor resolução e deteção de estruturas que os produtos disponíveis estado da arte. Estas classificações possuem também estruturas artificiais permanentes não presentes nos dados de referência. A aplicação desta classificação mais fina beneficia a delimitação das FGC, podendo acelerar o processo de geração e detetar estruturas e aglomerados sem proteção adequada.

**Palavras-chave:** Aprendizagem automática; deteção remota; sistemas de informação geográfica; CORINE; Copernicus; Sentinel-1; Sentinel-2

---



## ABSTRACT

---

Forest fires pose a serious threat not only to our forest but to buildings and people who live and work in rural areas. For good protection at the national level it is necessary the minimization of the impact of those fires, this minimization of the impact and risk of forest fires enforces the monitoring of fuel management networks surrounding housing, national roads and other structures. Therefore, it is essential the identification of artificial permanent structures from which the fuel management networks are defined. The existing permanent structure cartography has a low frequency of updates and their resolution might not be enough for the extraction of small structures, like small population agglomerates and houses, both of which require a strong forest fire protection. Based on sentinel imagery from the satellites Sentinel 1 e 2, terrain elevation models and reference data from COS 2015 together with OpenStreetMap data, it was resorted to machine learning algorithms in order to detect artificial permanent structures based on reflectance characteristics and radiometric indexes. At the end it was concluded that the best classification model was XGBoost, this model presents the best results over all classifications, this classifications have higher resolution and a better detection of permanent structures than available state-of-the-art products. They can also detect structures not present on the reference dataset. The application of those results benefits the delimitation of fuel management networks, with the ability to accelerate the generation process and detect structures and populations without adequate protection.

**Keywords:** Machine learning; remote detection; Copernicus; Sentinel-1; Sentinel-2; CORINE; Geospatial analysis; geographic information systems

---



# ÍNDICE

<b>Lista de Figuras</b>	<b>xvii</b>
<b>Lista de Tabelas</b>	<b>xxi</b>
<b>Siglas</b>	<b>xxiii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Motivação . . . . .	1
1.2 Enquadramento . . . . .	2
1.2.1 Estruturas Artificiais Permanentes . . . . .	2
1.2.2 Detecção Remota . . . . .	3
1.2.3 Faixas de gestão de combustível . . . . .	5
1.3 Problema . . . . .	5
1.4 Objetivos . . . . .	6
1.5 Estrutura do documento . . . . .	7
<b>2 Aprendizagem automática para classificação de ocupação de solo</b>	<b>9</b>
2.1 Algoritmos de Classificação . . . . .	9
2.1.1 Random Forest . . . . .	10
2.1.2 Gradient Boosted Trees . . . . .	11
2.1.3 Support Vector Machines . . . . .	12
2.1.4 Stochastic Gradient Descent . . . . .	13
2.1.5 Artificial Neural Network . . . . .	14
2.2 Avaliação de Modelos . . . . .	15
2.2.1 Matriz de Confusão . . . . .	15
2.2.2 Kappa . . . . .	17
2.3 Redução dimensional dos dados . . . . .	18
2.3.1 Importância dos atributos . . . . .	18
2.3.2 Boruta . . . . .	19
2.3.3 Least Absolute Shrinkage and Selection Operator . . . . .	19
2.3.4 Elastic Net . . . . .	20
2.4 Conclusão . . . . .	20

<b>3</b>	<b>Trabalho relacionado</b>	<b>21</b>
3.1	Cartografia de Ocupação do Solo . . . . .	22
3.1.1	OpenStreetMap . . . . .	24
3.2	Classificação Utilizando Detecção Remota . . . . .	24
3.2.1	Sentinel-1 . . . . .	24
3.2.2	Sentinel-2 . . . . .	26
3.3	Extração de Informação de Textura . . . . .	29
3.4	Índices Radiométricos . . . . .	30
3.4.1	Normalized Difference Water Index . . . . .	30
3.4.2	Normalized Difference Built-up Index . . . . .	31
3.4.3	Índices de vegetação . . . . .	31
3.5	Sinergia Entre Detecção Remota Passiva e Ativa . . . . .	31
3.6	Carta: Global Human Settlement Layer . . . . .	33
3.7	Resumo . . . . .	34
<b>4</b>	<b>Abordagem</b>	<b>37</b>
4.1	Formulação do problema de aprendizagem . . . . .	38
4.2	Problemas e estrutura da solução . . . . .	39
4.2.1	Solução proposta . . . . .	40
4.3	Ambiente experimental . . . . .	41
4.4	Metodologia . . . . .	42
4.4.1	Aprendizagem . . . . .	42
4.5	Conclusão . . . . .	46
<b>5</b>	<b>Estruturação dos Dados</b>	<b>49</b>
5.1	Extração dos dados de referência . . . . .	49
5.2	Obtenção e processamento de dados de treino . . . . .	56
5.2.1	Dados estáticos . . . . .	56
5.2.2	Séries temporais . . . . .	57
5.2.3	Modelo de elevação de terreno . . . . .	59
5.3	Conclusão . . . . .	59
<b>6</b>	<b>Validação experimental e otimização dos modelos</b>	<b>63</b>
6.0.1	Análise baseada em pixels . . . . .	64
6.1	Validação e análise experimental . . . . .	69
6.1.1	Classificação de solo . . . . .	69
6.2	Resultados da otimização dos algoritmos . . . . .	70
6.2.1	Random Forest . . . . .	71
6.2.2	Gradient Tree Boosting . . . . .	71
6.2.3	Artificial Neural Networks . . . . .	71
6.2.4	Support Vector Machines . . . . .	72
6.2.5	Stochastic Gradient Descent . . . . .	72

6.3	Conclusão . . . . .	73
<b>7</b>	<b>Resultados</b>	<b>75</b>
7.1	Metodologia experimental . . . . .	76
7.2	Análise dos resultados . . . . .	76
7.2.1	Análise estática . . . . .	80
7.2.2	Análise de séries temporais . . . . .	85
7.2.3	Adição do modelo de elevação do terreno . . . . .	89
7.3	Validação dos modelos e análise do seu erro . . . . .	95
7.4	Caso de estudo: Faixas de gestão de combustível . . . . .	101
7.5	Resultados da redução dimensional dos dados de treino . . . . .	104
7.6	Confirmação dos resultados do melhor modelo . . . . .	109
7.7	Discussão . . . . .	110
<b>8</b>	<b>Conclusão</b>	<b>111</b>
8.1	Contribuições . . . . .	112
8.2	Trabalhos futuros . . . . .	113
	<b>Bibliografia</b>	<b>115</b>
<b>A</b>	<b>Identificadores dos produtos utilizados</b>	<b>123</b>
<b>B</b>	<b>Tabela de métricas</b>	<b>125</b>
<b>C</b>	<b>Rasters</b>	<b>127</b>
<b>D</b>	<b>Tabelas de atributos</b>	<b>137</b>





## LISTA DE FIGURAS

1.1	Ilustração da diferença entre deteção remota ativa e passiva . . . . .	4
1.2	Redes das FGC na zona da Vila de Rei. . . . .	6
1.3	Região de treino, imagem do Sentinel-2 adquirida em Abril de 2016, Santarém	7
2.1	Exemplo de uma <i>Random Forest</i> . . . . .	10
2.2	Exemplo de um classificador <i>Gradient Boosted Trees</i> . . . . .	11
2.3	Exemplo das margens do classificador Support Vector Machines (SVM) . . .	12
2.4	Arquitetura de um neurónio artificial . . . . .	14
2.5	Estrutura de uma rede neuronal artificial . . . . .	15
3.1	Impacto da UMC na representação das classes de ocupação de solo . . . . .	22
3.2	Classes <i>artificial surfaces</i> da carta CORINE land cover . . . . .	23
3.3	NDWI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.	30
3.4	NDBI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.	31
3.5	NDVI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.	32
3.6	EVI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere. .	32
3.7	Comparação das cartas GHSL Landsat, à esquerda e GHSL Sentinel-1, à direita, sobre a região de Tomar. . . . .	34
4.1	Grupos de classes . . . . .	38
4.2	Cadeia de processamento utilizada nas experiências . . . . .	43
4.3	Estruturação dos dados de treino em 3-fold Cross Validation . . . . .	44
4.4	Cadeia de processamento para a obtenção dos melhores parâmetros de cada modelo . . . . .	44
4.5	Cadeia de processamento para redução de dimensionalidade dos atributos .	46
5.1	Segmento da carta COS 2015 sobre a região de Vila de Rei . . . . .	50
5.2	Classes extraídas da carta COS 2015 sobre a região de testes, grupo 1 . . . .	51
5.3	Distribuição das classes do grupo 1 . . . . .	52
5.4	Classes extraídas da carta COS 2015 sobre a região de testes, grupo 2 . . . .	52
5.5	Distribuição das classes do grupo 2 . . . . .	53
5.6	Dados com estradas extraídas do OSM e da carta COS 2015 sobre a região de testes, grupo 3 . . . . .	54

---

5.7	Distribuição das classes do grupo 3 . . . . .	55
5.8	Distribuição temporal das imagens recolhidas . . . . .	58
5.9	Percentagem de tempo passada em cada faixa de nebulosidade, categorizada pela percentagem de céu encoberto por nuvens. . . . .	58
5.10	Segmento do Models Digital de Elevação (MDE) de uma região após processamento . . . . .	61
6.1	Variação do valor NDBI de 3 píxeis ao longo do tempo . . . . .	66
6.2	Variação do valor NDVI de 3 píxeis ao longo do tempo . . . . .	67
6.3	Variação do valor NDVI de diferentes tipos de estradas ao longo do tempo . . . . .	68
6.4	Resultado final da classificação de uma imagem do Sentinel-2 utilizando a carta CLC como dados de treino . . . . .	69
6.5	Detalhe aumentado dos fenómenos observados . . . . .	70
6.6	Estradas parcialmente detetadas e estradas ausentes da classificação . . . . .	70
6.7	Métricas de avaliação do classificador SGD após otimização . . . . .	73
7.1	<i>Pipeline</i> das experiências. . . . .	77
7.2	Métricas de teste Kappa, por grupo, por dataset, por algoritmo . . . . .	78
7.3	Tempo de treino médio em segundos, por dados de treino, para cada modelo. . . . .	79
7.4	Valores da estatística Kappa para cada modelo, por tarefa de classificação . . . . .	81
7.5	Métricas de avaliação do modelo <i>XGBoost</i> por tarefa de classificação, utilizando dados estáticos . . . . .	82
7.6	Métricas de avaliação do modelo <i>Random Forest</i> por tarefa de classificação, utilizando dados estáticos . . . . .	82
7.7	Métricas de avaliação do modelo <i>Support Vector Machine</i> por tarefa de classificação, utilizando dados estáticos . . . . .	83
7.8	Métricas de avaliação do modelo <i>Stochastic Gradient Descent</i> por tarefa de classificação, utilizando dados estáticos . . . . .	83
7.9	Métricas de avaliação do modelo <i>Artificial Neural Network</i> por tarefa de classificação, utilizando dados estáticos . . . . .	84
7.10	Valores da estatística Kappa para cada modelo, por tarefa de classificação . . . . .	85
7.11	Métricas de avaliação do modelo <i>XGBoost</i> por tarefa de classificação, utilizando séries temporais . . . . .	86
7.12	Métricas de avaliação do modelo <i>Random Forest</i> por tarefa de classificação, utilizando séries temporais . . . . .	86
7.13	Métricas de avaliação do modelo <i>Artificial Neural Network</i> por tarefa de classificação, utilizando séries temporais . . . . .	87
7.14	Métricas de avaliação do modelo <i>Support Vector Machine</i> por tarefa de classificação, utilizando séries temporais . . . . .	87
7.15	Métricas de avaliação do modelo <i>Stochastic Gradient Descent</i> por tarefa de classificação, utilizando séries temporais . . . . .	88

---

7.16	Estatística Kappa de todos os modelos por tarefa de classificação, sem e com o modelo de elevação de terreno . . . . .	89
7.17	Impacto da adição do modelo de elevação de terreno na classificação do grupo 1 . . . . .	90
7.18	Impacto da adição do modelo de elevação de terreno na classificação do grupo 2 . . . . .	91
7.19	Resultados da classificação dos diferentes modelos para o grupo 3. . . . .	92
7.20	Imagem resultante da classificação do grupo 3 com modelo XGBoost utilizando séries temporais, sobre Tomar. . . . .	93
7.21	Imagem resultante da classificação do grupo 3 com modelo XGBoost utilizando séries temporais, sobre Vila de Rei. . . . .	94
7.22	Comparação da carta GHSL, à esquerda, com a carta gerada, à direita. Imagens sobre a região de Tomar. . . . .	95
7.23	Comparação da carta GHSL, à esquerda, com a carta gerada, à direita. Imagens sobre a região do Rio Zêzere. . . . .	96
7.24	Erros detetados sobrepostos à classificação do grupo 1 (em cima) e do grupo 2 (em baixo). Imagem sobre a zona de Tomar. . . . .	97
7.25	Estruturas permanentes detetadas não presentes nos dados de referência; Carta COS 2015 à esquerda, classificação XGBoost à direita. . . . .	98
7.26	Erros detetados sobrepostos à classificação do grupo 1 (em cima) e do grupo 2 (em baixo). Imagem sobre o Rio Zêzere. . . . .	100
7.27	FGC oficiais de Tomar e Ferreira do Zêzere. . . . .	101
7.28	Cadeia de processamento para a criação das Faixas de gestão de combustível	102
7.29	FGC geradas com base no produto final da classificação de estruturas permanentes. . . . .	103
7.30	FGC oficiais de Tomar e Ferreira do Zêzere, à esquerda; FGC definidas a partir da classificação, à direita. . . . .	103
7.31	Métricas de avaliação do método “importances” para cada valor total de importâncias selecionado . . . . .	105
7.32	Impacto da do número de atributos na performance do modo “importances”. . . . .	106
7.33	Magnitude dos coeficientes resultantes da regressão Lasso por atributo . . . . .	107
7.34	Tempos de treino, em segundos, por técnica de redução dimensional . . . . .	108
B.1	Tabela de métricas de todos os testes . . . . .	126
C.1	Raster resultante do classificador XGBoost com séries temporais, grupo 1. . . . .	127
C.2	Raster resultante do classificador XGBoost com séries temporais, grupo 2. . . . .	128
C.3	Raster resultante do classificador XGBoost com séries temporais, grupo 3. . . . .	129
C.4	Raster resultante do classificador Random Forest com séries temporais, grupo 1. . . . .	130

C.5	Raster resultante do classificador Random Forest com séries temporais, grupo 2. . . . .	131
C.6	Raster resultante do classificador Random Forest com dados estáticos, grupo 1.	132
C.7	Raster resultante do classificador Random Forest com dados estáticos, grupo 2.	133
C.8	Raster resultante do classificador DNN com séries temporais, grupo 1. . . . .	134
C.9	Raster resultante do classificador DNN com dados estáticos, grupo 1. . . . .	135
C.10	Raster resultante do classificador DNN com dados estáticos, grupo 1. . . . .	136

## LISTA DE TABELAS

2.1	Exemplo de uma matriz de confusão . . . . .	16
2.2	Tabela de confusão . . . . .	16
2.3	Tabela de confusão de um classificador que diferencia edifícios e florestas . . . . .	17
3.1	Cenários de teste do estudo de Abdikan et al. [1] . . . . .	25
3.2	Dados de entrada dos vários classificadores [5]. . . . .	26
3.3	Informação detalhada das bandas da constelação Sentinel-2 . . . . .	27
4.1	Suporte dos algoritmos para CPU e GPU. . . . .	41
4.2	Bibliotecas e implementações utilizadas . . . . .	41
5.1	Identificadores dos produtos utilizados na análise estática . . . . .	57
6.1	Parâmetros ótimos após otimização do classificador Random Forest . . . . .	71
6.2	Parâmetros ótimos após otimização do classificador XGBoost . . . . .	71
6.3	Parâmetros ótimos após otimização do classificador SVM . . . . .	72
6.4	Parâmetros ótimos, para a tarefa de classificação 1, após otimização do classificador SGD . . . . .	72
6.5	Parâmetros ótimos, para a tarefa de classificação 2, após otimização do classificador SGD . . . . .	72
6.6	Parâmetros ótimos, para a tarefa de classificação 3, após otimização do classificador SGD . . . . .	72
7.1	Mapeamento das classes do grupo 1 e do grupo 2 . . . . .	80
7.2	Métricas de teste finais (grupo 1) para o melhor classificador: XGBoost . . . . .	109
7.3	Métricas de teste finais (grupo 2) para o melhor classificador: XGBoost . . . . .	109
7.4	Métricas de teste finais (grupo 3) para o melhor classificador: XGBoost . . . . .	109
A.1	Identificadores dos produtos que compõem a série temporal Sentinel-1 . . . . .	123
A.2	Identificadores dos produtos que compõem a série temporal Sentinel-2 . . . . .	124
D.1	Tabela de atributos utilizados na análise estática . . . . .	137
D.2	Tabela de atributos utilizados na análise de séries temporais, excluindo dados de elevação de terreno . . . . .	138

D.3 Tabela de atributos utilizados na análise de séries temporais, complementadas com dados de elevação de terreno . . . . .	142
--	-----

## SIGLAS

AAE	Agência Europeia do Ambiente.
AEE	Agência Espacial Europeia.
ANN	Artificial Neural Networks.
CLC	CORINE land cover.
CNNs	Convolutional neural networks.
CORINE	CO-ordination of INFORMATION on the Environment.
COS	Carta de Ocupação do Solo.
DGT	Direção Geral do Território.
DNN	Dense Neural Networks.
EVI	Enhanced Vegetation Index.
FGC	Faixas de Gestão de Combustível.
GDAL	Geospatial Data Abstraction Library.
GHSL	Global Human Settlement Layer.
GLCM	Gray Level Co-occurrence Matrix.
ICNF	Instituto da Conservação da Natureza e das Florestas.
IM	Infravermelho Médio.
IOC	Infravermelho de Ondas Curtas.
IP	Infravermelho Próximo.
LASSO	Least Absolute Shrinkage and Selection Operator.
LULC	Uso e Ocupação do Solo (land use / land cover).
MDE	Models Digital de Elevação.

NDBI	Normalized Difference Built-up Index.
NDVI	Normalized Difference Vegetation Index.
NDWI	Normalized Difference Water Index.
OSM	OpenStreetMap.
RAS	Radar de Abertura Sintética.
RF	Random Forest.
SGD	Stochastic Gradient Descent.
SRC	Sistema de Referência de Coordenadas.
SRTM	Missão Topográfica Radar Shuttle.
SVM	Support Vector Machines.
UMC	Unidade Mínima Cartográfica.



## INTRODUÇÃO

Com o lançamento recente de vários satélites, como o satélite *Sentinel-1* [69] e *Sentinel-2* [22], a disponibilidade de vários tipos de dados sobre a Terra tem vindo a aumentar. Hoje em dia existe uma maior acessibilidade a imagens satélite gratuitamente [29], criando assim novas oportunidades que têm vindo a ser exploradas para uma melhor monitorização terrestre. Uma das áreas exploradas é a geração de cartas de uso / ocupação do solo, este tipo de cartografia possui informação sobre estruturas artificiais permanentes, como aglomerados populacionais, zonas industriais, estradas e muitas outras, e estruturas naturais. Esta informação ajuda no suporte a um bom desenvolvimento urbano, gestão de risco e de desastres [53, 72], criação de máscaras de zonas com estruturas permanentes, planeamento urbano, mapeamento de populações [25], delimitação de zonas de vigilância para prevenção de danos e vítimas em caso incêndio, entre outras.

### 1.1 Motivação

Muitos dos indicadores necessários para a análise de zonas urbanas podem ser derivados através de mapas **Uso e Ocupação do Solo** (land use / land cover) (LULC) [18, 70]. Estes mapas possuem uma resolução espacial não adequada à deteção de aglomerados pequenos de habitações e uma frequência de geração baixa, o que faz com que não acompanhem o rápido crescimento urbano e estruturas pequenas, como edifícios isolados, tornam-se difíceis de detetar [42, 65]. Por exemplo, até á data, a última versão da Carta de Ocupação de Solo é de 2015 e apenas são cartografadas estruturas com área superior a 1 hectare. Abordagens automáticas de deteção surgem como resposta a estas limitações, utilizando dados de grande resolução temporal e espacial.

A evolução de tecnologias de aquisição de dados terrestres, nomeadamente imagens satélite, trazem novos desafios e oportunidades no que toca à monitorização terrestre,

a utilização de imagens de alta resolução espacial e temporal são o foco inicial desta dissertação. Recorrendo a imagens obtidas por satélites com frequência de 5 dias e uma resolução espacial adequada ao problema, junto com a análise de mapas existentes de **LULC**, será possível a criação de um algoritmo, baseado em aprendizagem automática, que junte as características de imagens de satélite com uma análise da ocupação de solo baseada em mapas de **LULC**.

De todas as aplicações das cartas de estruturas permanentes é destacada a delimitação das **Faixas de Gestão de Combustível (FGC)**. As **FGC** são uma componente importante no que toca à prevenção e ao combate a incêndios. A boa manutenção e monitorização das faixas promove a conservação da floresta, bem como a proteção de vidas e bens. Estas faixas são geradas e mantidas por várias entidades oficiais e, de acordo com os critérios utilizados, podem conter diferenças significativas nos seus limites. A utilização de cartografia de estruturas permanentes, quanto utilizado no processo, permite uma maior coerência na definição destas faixas, para este efeito a deteção de edifícios isolados, ou aglomerados pequenos de habitações é crucial para o planeamento e monitorização das **FGC**.

Nesta tese é explorada a utilização a algoritmos de aprendizagem automática para a deteção de estruturas artificiais permanentes e, posteriormente, o estudo da aplicação destes mapas em prol de melhorar o mapeamento urbano e assim ajudar na delimitação de zonas de risco e planeamento da monitorização das **FGC**.

## 1.2 Enquadramento

Nesta secção são apresentados alguns conceitos fundamentais que servem de base para uma melhor compreensão dos temas apresentados nesta dissertação, serão abordadas três definições: **Estruturas Artificiais Permanentes**, **Deteção Remota** e **Faixas de gestão de combustível**.

### 1.2.1 Estruturas Artificiais Permanentes

O objetivo principal desta dissertação é a deteção de estruturas artificiais permanentes. Como tal, estas têm de ser bem definidas a priori. Uma estrutura artificial permanente é uma estrutura constituída por materiais artificiais feitos pelo homem em que a sua posição, ou existência, não é alterada num dado período de tempo definido, variável, desde a sua criação. A variável “tempo” dá alguma liberdade à definição, pois pode ser alterada para excluir (ou incluir) vários tipos de estruturas de interesse, como por exemplo habitações móveis e algumas estruturas temporárias de longo prazo.

Esta definição engloba as classes de nível 1, 2 e 3 da **Carta de Ocupação do Solo (COS)** de 2015 definidas como territórios artificializados. Mais especificamente as seguintes classes pertencem à definição de "estrutura artificial permanente":

- Tecido urbano contínuo

- Tecido urbano descontínuo
- Indústria e comércio
- Estradas
- Áreas portuárias
- Aeroportos
- Minas e áreas de construção
- Instalações desportivas
- Zonas históricas
- Zonas de deposição de resíduos

Estas classes servirão de base para a classificação das estruturas artificiais permanentes.

### 1.2.2 Detecção Remota

Define-se como deteção remota o conjunto de técnicas que possibilitam a obtenção de informações e propriedades de objetos na superfície da Terra usando dados adquiridos via sensores distantes, como sensores colocados em satélites, em aviões ou drones <sup>1</sup>. Este tipo de deteção é então focado em detetar propriedades à distância, como não existe contacto com os objetos de interesse as observações dependem de sinais propagados, como por exemplo sinais acústicos, ondas rádio e sinais óticos. Nesta tese serão abordados sinais ondas rádio, sinais óticos e qualquer outro tipo de informação derivada desses sinais, como por exemplo informação de relevo e textura.

Esta dissertação focar-se-á em dois tipos de deteção remota, deteção passiva e ativa. Na deteção remota passiva, ilustrada na Figura 1.1a, são utilizados vários sensores com sensibilidade a vários comprimentos de onda do espectro eletromagnético, estes sensores recebem ondas refletidas por objetos, onde a fonte principal de energia é o sol. Na sua forma mais simples, a deteção passiva é como tirar uma foto com uma câmara a objetos. Esta forma de deteção usa maioritariamente comprimentos de onda na parte visível do espectro e na porção do **Infravermelho Próximo (IP)** [36].

A deteção remota ativa, ilustrada na Figura 1.1b, utiliza sensores que produzem o sua própria energia. Os sensores emitem a onda a analisar, que por sua vez é refletida quando atinge a superfície terrestre. Quando a reflexão volta aos sensores estes guardam a informação. Utilizando a analogia de uma câmara fotográfica, a utilização do *flash* pode ser considerado um método de captura ativo. No escopo desta tese, vão ser considerados apenas técnicas de radar, nomeadamente o **Radar de Abertura Sintética (RAS)** utilizado pelo satélite Sentinel-1 [36].

---

<sup>1</sup>Definição dada pelo DGT - visitado a 15/01/2019

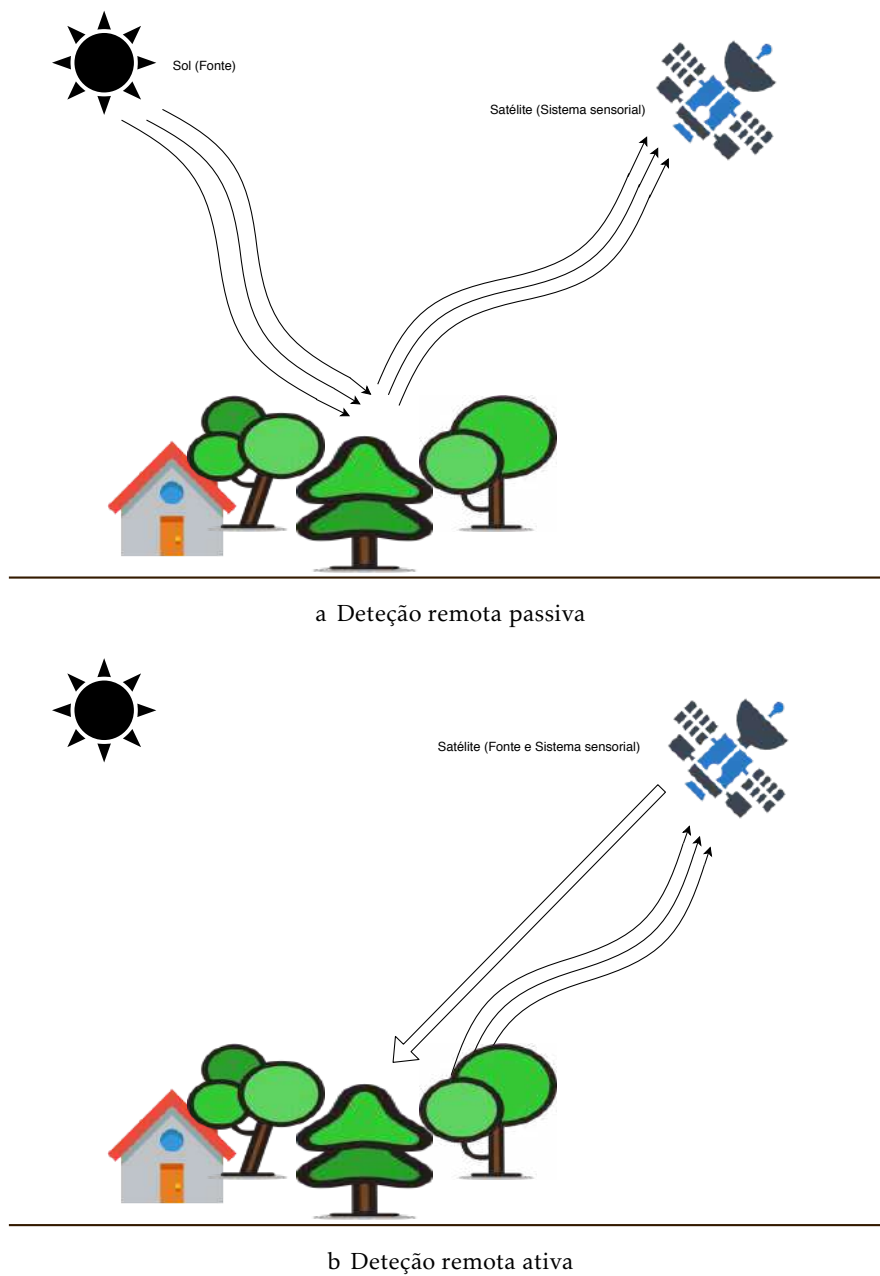


Figura 1.1: Ilustração da diferença entre detecção remota ativa e passiva

Vários sistemas de detecção remota têm sido desenvolvidos de modo a cobrir as necessidades de vários tipos de usos. Estes sistemas oferecem uma variedade de parâmetros, resolução temporal, espacial e espectral.

### 1.2.3 Faixas de gestão de combustível

As FGC são troços onde a vegetação foi reduzida ou até mesmo removida num dado perímetro à sua volta, de modo a reduzir a área percorrida pelos incêndios ou facilitar o combate aos mesmos <sup>2</sup>. Segundo o Decreto-lei n.º 124/2006, <sup>3</sup> existem três funções principais para as FGC:

- F1: Diminuição da superfície percorrida por grandes incêndios.
- F2: Proteção de vias de comunicação, edifícios, equipamentos, etc.
- F3: Isolamento de focos potenciais de ignição de incêndios.

A rede das FGC está dividida em três sub-redes, cada uma com as suas funções. A rede principal a nível sub-regional cumpre a função F1, a rede secundária a nível municipal cumpre as funções F2 e F3, por fim, a rede terciária a nível local cumpre a função F3. A figura 1.2 apresenta o planeamento das FGC sobre uma zona do concelho de Vila de Rei, onde podem ser observadas várias redes envolvendo pequenos aglomerados habitacionais. As faixas observadas que percorrem grandes extensões de terreno pertencem à rede primária.

Um dos objetivos desta dissertação é a possibilidade da delimitação automática das FGC envolvendo edifícios e aglomerados habitacionais, como consequência, as redes secundárias e terciárias, para proteção de vidas e bens, serão o foco principal.

## 1.3 Problema

Como foi apresentado na secção 1.1, o objetivo desta dissertação é, a partir de detecção remota, acelerar o processo de classificação de estruturas permanentes e, posteriormente, estudar a aplicação das cartas geradas à delimitação das FGC. Dito isto, esta tese aborda os seguintes problemas:

- A resolução temporal e espacial das cartas e produtos de LULC existentes, que contêm estruturas artificiais permanentes, não são suficientes para muitas das suas potenciais aplicações.
- Num país com 92.212 km<sup>2</sup> [15], a delimitação das FGC torna-se numa tarefa pesada e morosa, devido aos diferentes critérios dependendo da entidade responsável, as faixas podem registar diferenças significativas nos seus limites.

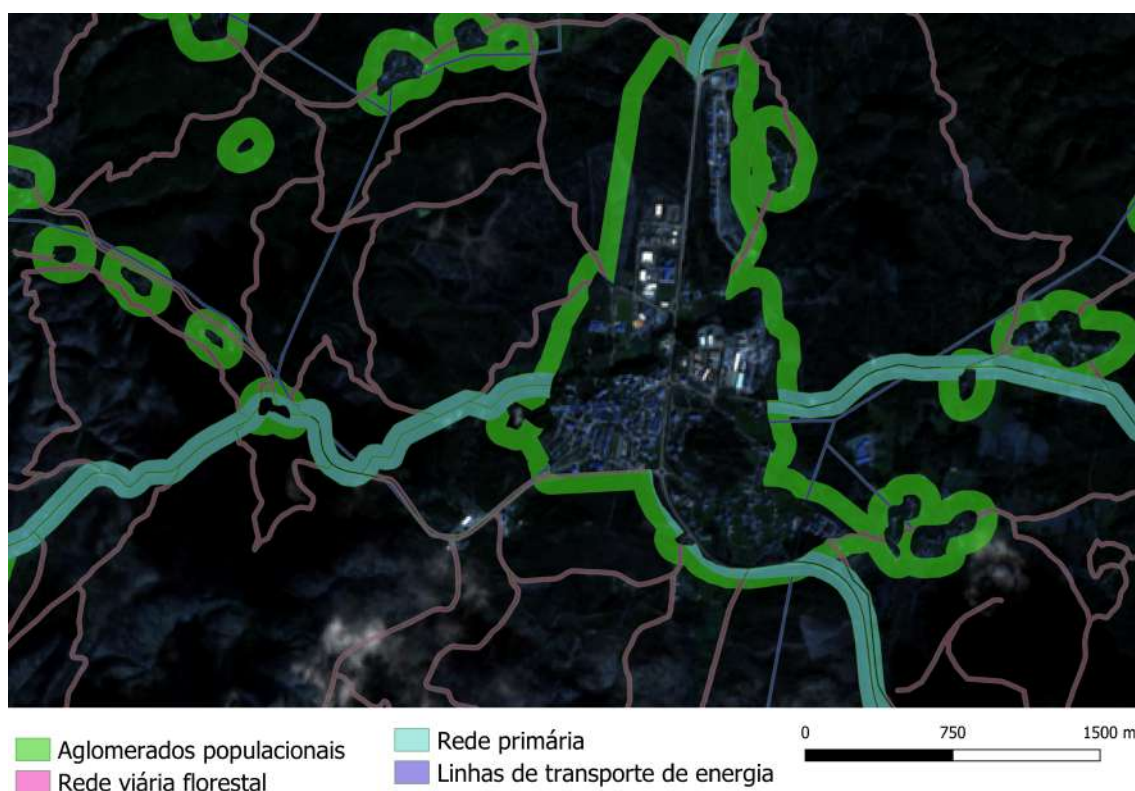


Figura 1.2: Redes das FGC na zona da Vila de Rei.

Para resolver os problemas identificados é preciso propor uma zona de testes, para isso foi selecionada uma zona a Noroeste de Santarém, presente na Figura 1.3, esta zona contém grandes aglomerados populacionais, assim como grandes superfícies com vegetação com vários aglomerados mais pequenos. Sendo maioritariamente composta por vegetação, por ser uma zona rural, esta zona apresenta um desafio aos algoritmos de aprendizagem automática. A região cobre uma área de 50x38 km o que, a uma resolução de 10 por 10 metros, equivale a aproximadamente 19 milhões de pixels. Esta zona foi escolhida pela sua distribuição das faixas de gestão de combustível e por estar sobre apenas um grânulo de ambos os satélites Sentinel-1 e 2, o que simplifica e acelera o processo de processamento das imagens.

## 1.4 Objetivos

Os objetivos desta dissertação estão separados em vários sub problemas, nomeadamente:

- **Classificação automática das estruturas artificiais permanentes:** O principal objetivo é a criação de um algoritmo que identifique as estruturas artificiais permanentes de maneira automática. Espera-se que este algoritmo consiga criar mapas de

<sup>2</sup>Informação retirada de <http://www.drapc.min-agricultura.pt> a 19/02/2019

<sup>3</sup>[http://www.pgdlisboa.pt/leis/lei\\_mostra\\_articulado.php?nid=1931&tabela=leis](http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=1931&tabela=leis) visitado a 19/02/2019

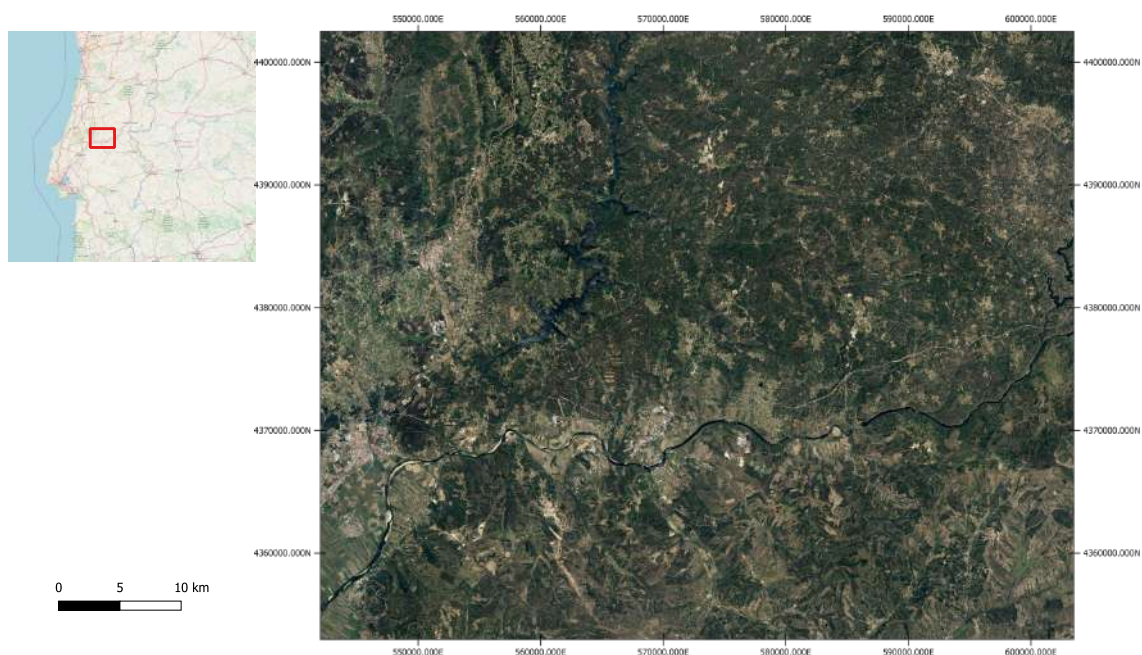


Figura 1.3: Região de treino, imagem do Sentinel-2 adquirida em Abril de 2016, Santarém

LULC com uma resolução adequada que permita a geração das rede das FGC.

- **Comparação de classificadores:** Como consequência do estudo necessário para criar o melhor modelo de classificação de estruturas é necessária a comparação de vários algoritmos de classificação. No final é proposto o melhor modelo e os seus parâmetros.
- **Comparação de conjuntos de dados:** Não existindo uma resposta a quais os melhores conjuntos de dados a utilizar para uma melhor deteção de estruturas permanentes, é necessário investigar os conjuntos com melhor performance na área de modo a obter dados robustos para resolver o problema principal desta tese. São comparados vários conjuntos de dados e proposto o conjunto com melhor performance.
- **Caso de estudo: Delimitação das faixas de gestão de combustível:** É pretendido também estudar o potencial das cartas geradas para a delimitação automática das FGC da rede secundária e terciária a partir das estruturas artificiais detetadas. O produto final será também comparado às FGC oficiais.

## 1.5 Estrutura do documento

Este documento é composto por oito capítulos:

- **Introdução:** o capítulo 1 desta tese contém uma pequena introdução do tema bem como a descrição geral do problema, são apresentadas também as motivações que guiam o projeto, os objetivos. Este capítulo tem também como objetivo fazer uma

introdução a alguns conceitos fundamentais que servem de base para os capítulos seguintes.

- **Aprendizagem automática para classificação de ocupação de solo:** no capítulo 2 são apresentados os algoritmos de aprendizagem com os melhores resultados na classificação de LULC; são apresentados também métricas de avaliação e, no final, técnicas de redução de atributos para os dados de entrada.
- **Trabalho relacionado:** o estado da arte, apresentado no capítulo 3, resume a pesquisa realizada durante a dissertação, aqui são apresentadas as tecnologias, produtos e técnicas mais utilizadas e com melhor desempenho conhecidas no âmbito do tema.
- **Abordagem:** como resultado do período de investigação, o capítulo 4 apresenta os problemas identificados e proposta das suas soluções, bem como uma descrição mais detalhada da metodologia. São apresentados também os algoritmos utilizados e as suas implementações.
- **Estruturação dos dados:** neste capítulo, número 5, são apresentados ambos os dados de treino e os dados de referência. É apresentado não só a sua estrutura e os diferentes grupos de dados gerados, mas também quais foram os produtos iniciais e como estes foram recolhidos.
- **Validação experimental e otimização dos modelos** o capítulo 6 começa com uma análise aos dados de treino, tentando prever os possíveis problemas da investigação, culminando numa validação de conceito. Este capítulo dá ênfase também aos resultados da otimização dos parâmetros dos algoritmos estudados.
- **Resultados:** o capítulo 7 apresenta os testes finais realizados durante a investigação. São enumerados os dados de entrada utilizados e de seguida todas as métricas derivadas bem como a sua comparação. No final é apresentada a combinação do algoritmo e dos dados com métricas mais elevadas.
- **Conclusões:** no final da dissertação, capítulo 8, são apresentadas as conclusões de todo o trabalho efetuado durante a tese. Tendo encontra os objetivos enunciados, são apresentadas as conclusões finais para cada.



## APRENDIZAGEM AUTOMÁTICA PARA CLASSIFICAÇÃO DE OCUPAÇÃO DE SOLO

Visto que o objetivo desta dissertação é a detecção automática de estruturas permanentes, é necessário estudar quais são as metodologias de classificação utilizadas na área. A aprendizagem automática é a ferramenta disponível mais poderosa no que toca a classificação **LULC** e será o maior alvo de estudo durante a dissertação. Este capítulo apresenta os algoritmos mais utilizados e com melhores resultados na área de classificação de **LULC**. O capítulo está dividido da seguinte forma:

- **Algoritmos de Classificação**, secção 2.1: Esta secção apresenta os algoritmos de classificação mais utilizados na classificação **LULC** e com melhores resultados. São também apresentados alguns algoritmos adicionais que possuem um bom potencial para o efeito.
- **Avaliação de modelos**, secção 2.2: Escolhidos os modelos é necessário então avaliar os mesmos. Esta secção apresenta as métricas e estatísticas mais utilizadas e com resultados robustos para a análise, comparação e otimização de algoritmos de aprendizagem automática.
- **Redução dimensional dos dados**, secção 2.3: Finalmente, são apresentadas técnicas para a redução dimensional dos dados de treino dos algoritmos. Estas técnicas visam a reduzir a complexidade do treino e melhorar a interpretabilidade dos resultados.

### 2.1 Algoritmos de Classificação

Dado um conjunto de amostras de treino de tamanho  $N$  na forma  $\{(x_1, y_1), \dots, (x_N, y_N)\}$  de tal modo que  $x_i$  é o vetor de atributos do  $i$ -ésimo exemplo e  $y_i$  a sua classe, um algoritmo

de aprendizagem procura determinar uma função  $f : X \rightarrow Y$ , onde  $X$  é o universo de entrada e  $Y$  o universo de saída. A função  $f$  pertence ao espaço de todas as funções possíveis  $F$ , nomeado também como espaço de hipóteses. Isto significa que o algoritmo de aprendizagem procura a função que melhor representa os dados no espaço de hipóteses  $F$ .

No escopo desta tese, o universo de entrada  $X$  são todas as combinações possíveis de valores para todas as bandas e outras informações derivadas de um pixel, o  $Y$  são as classes dos mapas de cartografia LULC, uma amostra  $(x_i, y_i)$  contém o vetor  $x_i$  que contém toda a informação de um pixel e a sua classe  $y_i$  correspondente.

### 2.1.1 Random Forest

O classificador *Random Forest (RF)* [12] possui, em geral, uma performance mais elevada do que outros classificadores na área de detecção remota [28]. Sobretudo quando aplicado a grandes áreas geográficas onde os dados são multidimensionais [30]. Este classificador é baseado árvores de decisão, cada RF possui  $k$  árvores. Exemplificado na Figura 2.1, quando um exemplo é classificado os seus atributos são passados a todas as  $k$  árvores. Cada  $k$  previsão conta como um voto e a classe mais votada é escolhida como a classe final do exemplo.

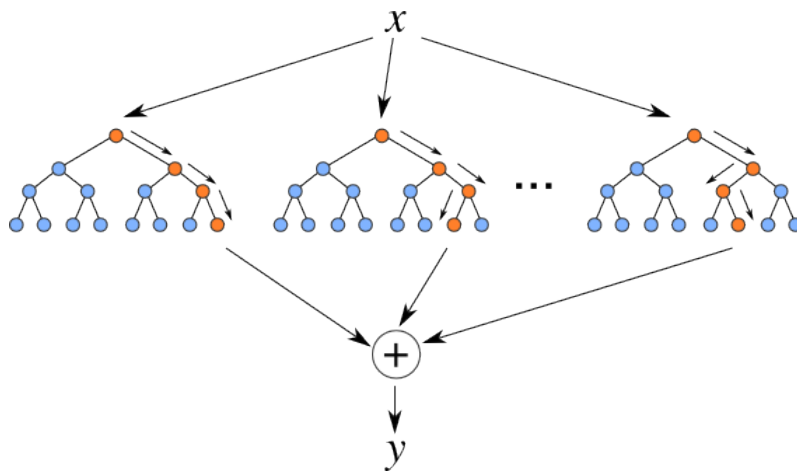


Figura 2.1: Exemplo de uma *Random Forest*; fonte: <https://dsc-spidal.github.io/harp/docs/examples/rf/>

O algoritmo de treino utilizado para as RF aplica uma técnica chamada *bootstrap aggregating* ou *bagging*. *bagging* é um Meta-algoritmo de *ensemble*. Dado um conjunto de dados  $D$  de tamanho  $n$ , esta técnica consiste em gerar  $k$  conjuntos de dados  $D_i$  via amostragem com reposição de  $D$ . Após a amostragem são treinados  $k$  modelos de aprendizagem com os  $k$  conjuntos, finalmente os seus resultados são agregados num só. Ao utilizar árvores de decisão, cada  $k$  árvore é treinada utilizando uma amostra aleatória dos dados de treino, o que resulta em  $k$  árvores diferentes não correlacionadas, ao contrário de treinar as todas

árvores de decisão com os dados de treino completos o que iria gerar árvores correlacionadas. Os atributos das amostras escolhidas também são aleatoriamente limitados. Esta técnica conduz a uma melhor performance pois diminui a variância do modelo. O que significa que apesar de uma única árvore ser sensível a anomalias nos dados de treino a combinação das  $k$  árvores não o é. A amostragem *bootstrap* é uma maneira de gerar árvores de decisão sem correlação ao treinar as mesmas com conjuntos de amostras diferentes.

Ao utilizar *bagging* pode ser calculado o erro *out-of-bag* [9], este é a média do erro para cada exemplo de treino  $X_i$ , utilizando apenas as árvores que não tiverem  $X_i$  na sua amostra de treino. Visto que cada amostra utilizada para treinar uma árvore de decisão deixa de fora uma parte dos dados estes podem ser utilizados para estimar o erro de generalização do modelo e outras métricas. Essencialmente, as amostras não utilizadas servem de conjunto de testes para obter as métricas finais, evitando assim não utilizar grande parte dos dados para posteriormente obter as mesmas métricas.

O método RF é computacionalmente menos intenso que outros métodos [46, 58] e também possui uma parametrização simples. Estes fatores tornam as RF um método apelativo no que toca a classificação rápida de mapas LULC.

Vários estudos apresentam resultados diferentes para o número ótimo de árvores, este apresenta-se entre 10 [79] a 100 árvores [58]. Apesar deste número não estar definido o estudo de Goldblatt et al. 2016 [28] mostra que o desempenho do método RF aumenta com o número de árvores, mas tende a estabilizar por volta das 10. Esta performance mantém-se até 50 ou até 100 árvores. Belgiu e Drăgu 2016 [7] sugerem que o número de árvores de decisão numa RF seja de 500.

A sua resistência a anomalias nos dados de treino e a sua fácil parametrização tornam as RF uma das escolhas principais na abordagem ao problema desta dissertação.

### 2.1.2 Gradient Boosted Trees

As *Gradient Boosted Trees* são um classificador baseado em árvores de decisão. Ao contrário de um classificador RF, que utiliza varias árvores em voto paralelo, este utiliza árvores de decisão de forma sequencial tentando com que cada árvore nova aumente a performance das restantes [16], exemplificado na Figura 2.2.

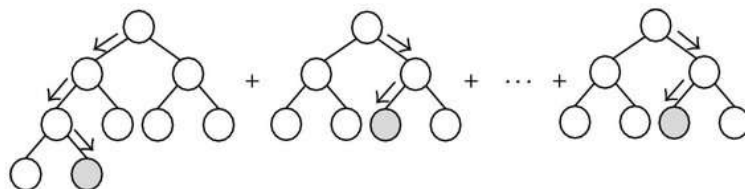


Figura 2.2: Exemplo de um classificador *Gradient Boosted Trees*; (fonte: Shin 2015 [63])

Para isso é utilizado *Boosting*, em que a primeira árvore é treinada com os dados originais e as seguintes iterações, adicionando novas árvores, são treinadas de modo encontrar

o valor residual do erro da iteração anterior. Assumindo que  $F_i$  é um classificador imperfeito então  $F_{i+1}$  tenta corrigir os erros de  $F_i$  tentando encontrar a função residual  $h$  tal que  $F_{i+1}(x) = F_i(x) + h(x) = y$ .

No final é obtido um classificador mais forte, composto por classificadores fracos, em que cada classificador mais fraco apenas se foca nos erros do anterior. Esta tática reduz o viés e a variância nos classificadores [11]. A utilização deste algoritmo mostra resultados que rivalizam a utilização de RF [26, 59] o que o torna um alvo de investigação de interesse.

### 2.1.3 Support Vector Machines

As SVM [20] são algoritmos de aprendizagem utilizados para classificação e regressão. Este algoritmo possui bons resultados no que toca a classificação de LULC [48, 49].

Este classificador tenta separar as classes maximizando a margem entre elas. Para isso é construído um hiperplano acompanhado de pontos nomeados *support vectors*. Estes pontos representam a maior margem possível na separação das classes, a Figura 2.3 mostra um exemplo das margens de uma SVM com três *support vectors*.

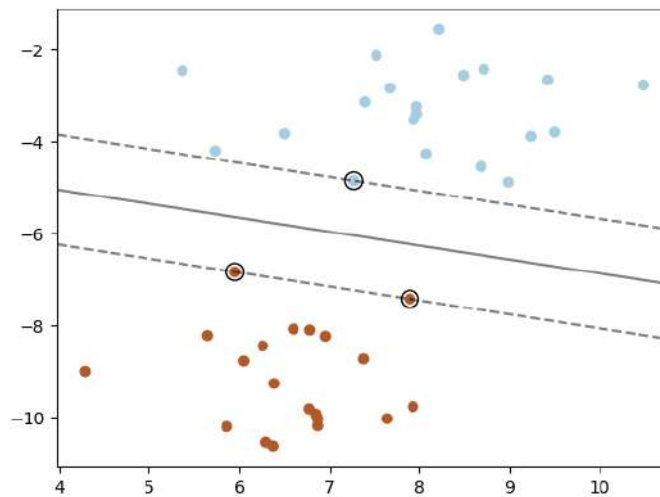


Figura 2.3: Exemplo das margens do classificador SVM; fonte: <https://scikit-learn.org/stable/modules/svm.html>

A técnica de utilizar um hiperplano que maximize a margem apenas tem bons resultados se os dados forem linearmente separáveis, caso contrário não existe nenhum plano que permita a maior margem sem que exista nenhum ponto dentro da mesma. Este problema é resolvido ao inserir uma margem que permita a existência de pontos entre os vetores de suporte (*soft-margin*), é adicionado o parâmetro  $C$  que define a robustez da margem.

Caso a utilização de uma margem deste tipo não seja suficiente é utilizado uma técnica nomeada *kernel trick*. Nesta abordagem os dados são transformados em dados com dimensões superiores utilizando uma função *kernel*, posteriormente os dados são separados por um hiperplano na nova dimensão. O classificador continua a ser um hiperplano no espaço transformado mas no espaço original passa a poder ser não linear.

O método com mais resultados na classificação de LULC é a utilização de uma função *kernel* radial [48, 49], esta função adiciona uma nova variável  $\gamma$ . Um valor  $\gamma$  mais alto transforma a margem de forma a que esta fique mais apertada envolvendo as classes, enquanto um valor mais baixo cria uma fronteira mais suave.

A utilização de SVM é um pouco mais complexa do que a utilização de RF devido aos parâmetros acrescidos, mas os seus bons resultados apresentados na literatura tornam este método um forte candidato para utilizar na solução final.

#### 2.1.4 Stochastic Gradient Descent

Com origens no artigo de Robbins e Monro [57] publicado a 1951, *Stochastic Gradient Descent* (SGD) é um algoritmo bastante eficiente para treinar classificadores lineares, como por exemplo SVM. Um dos seus pontos fortes é a eficiência, este algoritmo é de grande interesse quando utilizado em problemas de grande escala [80], o que vai de encontro com o tema desta dissertação, visto que dados geográficos podem ter milhões de amostras e centenas de atributos. Por exemplo, os dados mais complexos desta dissertação utilizam por volta de 144 atributos e de 2 milhões de amostras. A implementação do algoritmo utilizada nesta tese será a disponível na ferramenta *Scikit Learn* [51], segundo a documentação, este classificador escala facilmente para problemas com mais de  $10^5$  amostras de treino e  $10^5$  atributos.

Dado um conjunto de amostras de treino do tipo  $(x_1, y_1), \dots, (x_n, y_n)$ , com  $x_i \in \mathbb{R}^m$  e  $y_i \in \{-1, 1\}$ , o algoritmo tem como objetivo aprender a função  $f(x) = w^T x + b$ , com  $w \in \mathbb{R}^m$  e  $b \in \mathbb{R}$ . Para aprender estes parâmetros é minimizado o erro regularizado dado na equação 2.1, onde  $L$  é a função de perda e  $R$  é o termo da regularização que penaliza a complexidade do modelo.

$$E(w, b) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \alpha R(w) \quad (2.1)$$

Existem várias escolhas disponíveis para os termos  $L$  e  $R$ , para ir de encontro com o estudado, e visto que possui bons resultados na área de deteção remota, será escolhida a função *Hinge* para  $L$ , o que implica classificação com *soft-margins* (SVM). Existem 3 possibilidades para o termo  $R$ , L2 (*Ridge*), L1 (*Lasso*) e *Elastic Net*.

Visto que este método é bastante eficiente e suporta regularização de maneira nativa, é um bom alvo de estudo. Será de grande interesse a comparação do SGD com a implementação base das SVM.

### 2.1.5 Artificial Neural Network

As *Artificial Neural Networks (ANN)*, como o nome indica, são modelos inspirados em redes neuronais biológicas. Foram criadas com o objetivo de aprender simulando as ligações de um cérebro humano. Uma ANN é composta por um conjunto de neurónios interligados com a estrutura representada na Figura 2.4, cada conexão possui um peso ( $W_i$ ) que representa quão forte é a mesma. Tendo  $m$  como o número de entradas,  $\phi$  a função de ativação,  $x_i$  a  $i$ -ésima entrada,  $w_i$  o  $i$ -ésimo peso, e  $b$  representando o *bias*, um neurónio pode ser representado através a função 2.2.

$$Y = \phi \left( \sum_{i=0}^m w_i x_i + b \right) \quad (2.2)$$

Ao juntar vários neurónios é construída uma camada e a junção das mesmas, ligando as saídas dos neurónios como entradas dos seguintes, compõe uma rede como a apresentada na Figura 2.5. Uma maneira de conectar as camadas é ligar todos os neurónios de uma camada à camada seguinte, esta estrutura designa-se *Dense Neural Networks (DNN)*. As ANN aprendem ao efetuar várias passagens pelos dados de entrada e calcular o erro na camada final, finalmente os pesos e os *bias* são alterados de modo a diminuir o mesmo.

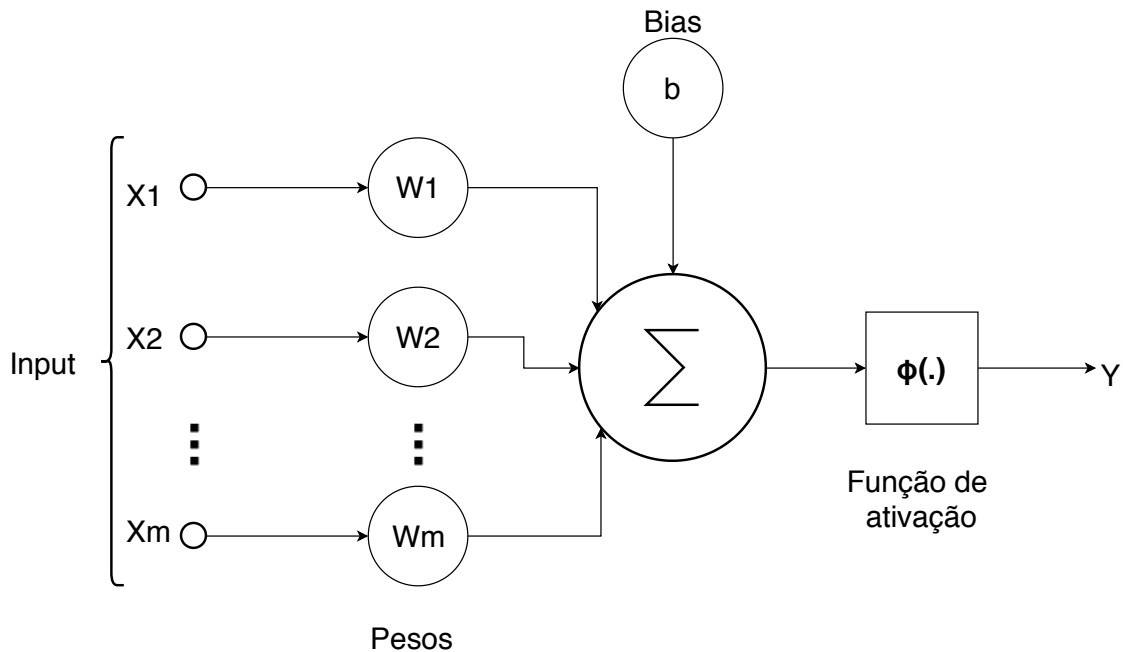


Figura 2.4: Arquitetura de um neurónio artificial

As ANN possuem bons resultados em áreas variadas, sendo as mais relacionadas com o tema da dissertação a segmentação de imagens [60] e reconhecimento de objetos [56]. No que toca à classificação de cartas LULC, as redes neuronais apresentam resultados que rivalizam outros métodos [50, 61].

A construção de uma rede neuronal artificial para classificação é algo mais complexo que os outros modelos, tem de ser escolhido o número de camadas e seus respetivos

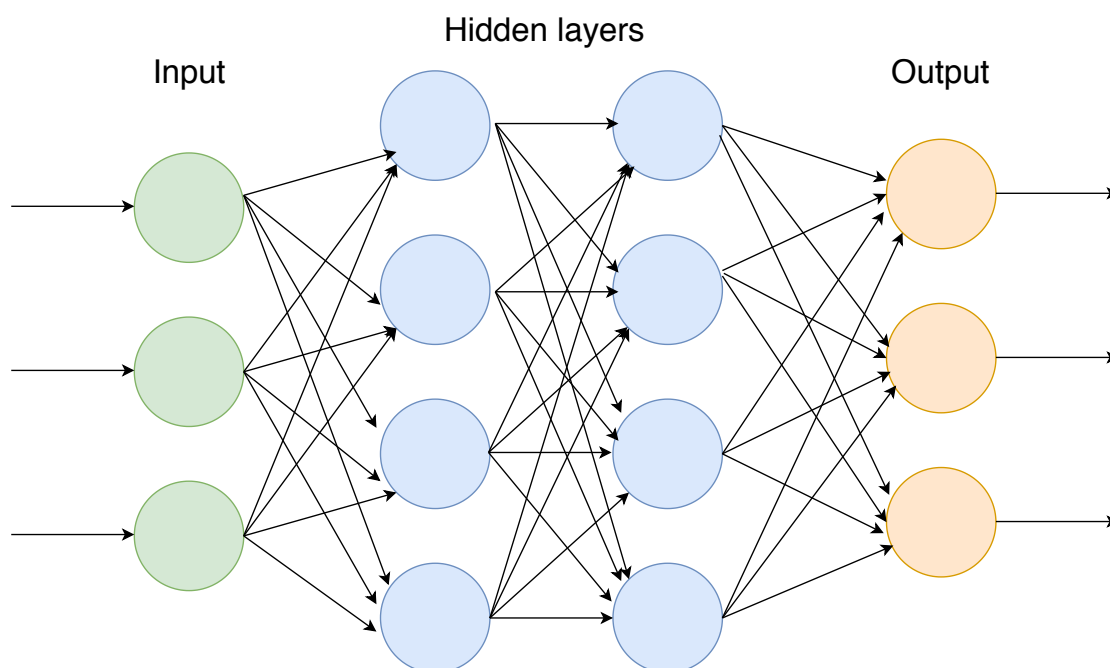


Figura 2.5: Estrutura de uma rede neural artificial

neurónios bem como eles se conectam entre si. Existem várias heurísticas para estes parâmetros [38], apesar disso não existe algo conclusivo para LULC em Portugal Continental. Concluindo, estes modelos possuem estado da arte em quase todas as suas aplicações, o que os torna num alvo de estudo interessante.

## 2.2 Avaliação de Modelos

Uma das componentes mais importantes num problema de aprendizagem automática é a obtenção de métricas para avaliar a performance dos modelos. Estas métricas são utilizadas para comparação de classificadores durante o desenvolvimento de modo a encontrar os parâmetros ótimos ou o melhor modelo para o problema em questão.

Esta secção apresenta várias técnicas e métricas mais utilizadas para validação e comparação de modelos.

### 2.2.1 Matriz de Confusão

Uma matriz de confusão representa numericamente a performance de um modelo, em que cada linha representa a previsão do modelo de uma dada classe e cada coluna representa a classe real.

Por exemplo, um modelo treinado para classificar edifício, florestas e rochedo numa amostra de 40 pontos de numa imagem pode gerar a tabela 2.1. As previsões corretas são apresentadas na diagonal e valores fora da diagonal representam erros de classificação.

Tabela 2.1: Exemplo de uma matriz de confusão

		Classe real		
		Edifício	Floresta	Rochedo
Previsão	Edifício	10	4	0
	Floresta	5	4	2
	Rochedo	1	3	11

É relevante notar que, no contexto desta dissertação, um erro de omissão ocorre quando um algoritmo não classifica algo existente num pixel, como por exemplo uma estrutura existente. Para um erro de comissão o algoritmo classifica algo erradamente num dado pixel, por exemplo a classificação de uma estrutura permanente como sendo água.

Apesar de este tipo de análise mostrar a performance de um classificador poderá ser difícil analisar a mesma apenas olhando para uma matriz devido à sua complexidade, a complexidade aumenta ainda mais quando for necessário comparar duas ou mais matrizes; para resolver este problema são derivadas novas métricas utilizando valores presentes na matriz. Estas métricas podem ser aplicadas a uma classificação multi classe caso necessário, transformando cada variável num cenário um contra todos, em que o evento representado na tabela 2.2 por A será pertencer à classe e o B não pertencer. Neste caso cada métrica será relevante apenas a uma dada classe.

Tabela 2.2: Tabela de confusão

		Classe real	
		A	B
Previsão	A	TP	FP
	B	FN	TN

Reduzindo o problema a uma classificação binária obtemos a tabela 2.2, em que TP representa verdadeiro positivo, TF verdadeiro negativo, FN falso negativo (erros de omissão) e FP falso positivo (erros de comissão). A partir desta tabela podem ser geradas várias métricas comuns [23]:

$$FP\ ratio = \frac{FP}{FP + TN} \quad (2.3)$$

$$TP\ ratio/recall = \frac{TP}{TP + FN} \quad (2.4)$$

$$precision = \frac{TP}{TP + FP} \quad (2.5)$$

$$F1 - score = \frac{1}{\frac{\frac{1}{precision} + \frac{1}{recall}}{2}} \quad (2.6)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.7)$$



Algumas das métricas apresentadas são sensíveis à distribuição dos dados de referência utilizados na classificação, por exemplo se uma classe representar 90% dos dados totais um classificador que classifique tudo como apenas essa classe possuirá métricas na casa dos 90%, como a *accuracy* e a *precision*. Mesmo assim não deixa de ser um mau classificador. É preciso observar as métricas por classe e introduzir outras que tenham isto em conta, como a estatística Kappa, apresentada na próxima secção.

### 2.2.2 Kappa

O Kappa é uma variável estatística mais robusta que as outras métricas no que toca à avaliação quando os classes são bastante desequilibradas. Esta estatística é bastante utilizada na área de deteção remota [1, 24, 35, 42, 58]. O impacto do balanceamento dos dados é algo bastante importante a considerar na avaliação. Um classificador pode tender a classificar apenas as classes com mais representação corretamente, deixando as minorias mal classificadas e mesmo assim ter uma acurácia elevada.

A estatística Kappa é definida com a seguinte função, em que  $p_0$  é a concordância observado (*accuracy*) e  $p_e$  a probabilidade de concordância por acaso:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (2.8)$$

Para as categorias  $k$ , com  $N$  amostras e  $n_{ki}$  o número de vezes que o classificador  $i$  prevê  $k$ , sendo  $n_{k1}$  o classificador 1 e  $n_{k2}$  o classificador 2, então a  $p_e$ :

$$p_e = \frac{1}{N^2} \sum_k n_{k1} * n_{k2} \quad (2.9)$$

Sé existir uma concordância total então Kappa toma o valor 1, caso apenas exista concordância aleatória então Kappa toma o valor 0. O classificador pode tomar valores negativos caso a concordância seja pior que uma classificação aleatória.

Tabela 2.3: Tabela de confusão de um classificador que diferencia edifícios e florestas

		Classe real	
		Edifício	Floresta
Previsão	Edifício	a=50	b=5
	Floresta	c=15	d=25

Utilizando como exemplo a tabela 2.3, resultando de um classificador binário, com  $N = 95$  podemos então calcular:

$$p_0 = \frac{a + d}{N} = \frac{50 + 25}{95} \approx 0,79 \quad (2.10)$$

$$p_e = \frac{1}{N^2} ((a + b) * (a + c)) + ((b + d) * (c + d)) = \frac{1}{95^2} (55 * 65) + (30 * 40) \approx 0,53 \quad (2.11)$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.79 - 0.53}{1 - 0.53} \approx 0,55 \quad (2.12)$$

## 2.3 Redução dimensional dos dados

Na área de aprendizagem automática e estatística existe uma técnica chamada seleção de atributos (*Feature Selection*). Como o nome indica é um processo que seleciona atributos relevantes para a classificação dos atributos originais [14], com a premissa da existência de atributos redundantes ou irrelevantes que podem ter um impacto negativo nos algoritmos de classificação [3]. A remoção destes atributos pode ser então efetuada sem perder demasiada informação.

Em geral estas técnicas tem três vantagens no contexto e tema desta tese:

- **Encurtar os tempos de treino**, no caso desta dissertação, que utiliza imagens de satélite com uma cobertura vasta do terreno, a remoção de um atributo pode significar gigabytes de espaço reduzido, diminuindo assim significativamente os tempos de treino [3, 37].
- **Reduzir *Overfitting***, este fenómeno acontece quando se cria de um modelo com correlação muito grande a um dado grupo de dados, e por isso falha ao ser aplicado num grupo de dados diferente. Um modelo que faça *overfitting* aos dados de treino vai ter uma performance reduzida quando aplicado nos dados reais.
- **Simplificar os modelos**, ao remover atributos desnecessários a facilidade de interpretação do algoritmo aumenta [37].

Como consequência a utilização de seleção de atributos tem o potencial de aumentar significativamente a performance dos algoritmos a utilizar [14, 37].

Esta secção vai explorar 4 métodos para seleção de atributos, importância dos atributos (*feature importance*), Boruta, **Least Absolute Shrinkage and Selection Operator (LASSO)** e Elastic Net.

### 2.3.1 Importância dos atributos

Para um dado atributo é possível atribuir-lhe um valor para a sua importância. Quanto maior este valor mais informação este atributo possui e maior é o seu impacto na classificação final.

As *Random Forests* podem ser utilizadas para extrair a importância dos atributos [10], dado um conjunto de dados  $D = \{(x_i, y_i)\}_{n=1}^i$  com atributos  $V = \{X_1, \dots, X_n\}$  e categorias  $Y$ , é feito primeiramente o treino do algoritmo, guardando o erro out-of-bag ( $E_{OOB}^i$ ) para cada  $(x_i, y_i)$ . Para calcular a importância de cada  $X_i \in V$ , após do treino os valores do atributo são permutados e o erro out-of-bag é calculado novamente ( $E_{OOB}^f$ ). No final os erros são comparados, quanto maior o efeito da permutação de  $X_i$  no erro maior a sua importância.

Uma das limitações desta técnica é que apenas faz uma ordenação dos atributos por importância, a decisão de quais os  $k$  melhores é outro problema a resolver. Uma das táticas possíveis a utilizar é somar as importâncias até obter 0.95 de importância total, ou outro

qualquer valor total definido, e descartar os restantes, atributos com 0 de importância podem ser descartados sem perda de informação.

### 2.3.2 Boruta

Este algoritmo é uma extensão a algoritmos que utilizem conjuntos de árvores [39], em suma, o algoritmo compara a importância do conjunto de atributos original  $V = \{X_1, \dots, X_n\}$  com a importância dos atributos originais fundidos com cópias suas aleatórias durante várias iterações.

Apenas atributos em que a sua importância é mais alta que a dos atributos aleatórios são considerados importantes, para isso é feito um teste estatístico bilateral para todos os atributos com a hipótese nula  $H_0 : Imp(X_i) = IMMA$ , sendo IMMA a importância máxima dos atributos aleatórios.

Para cada atributo é guardado a soma do número de vezes em que a sua importância foi maior que a IMMA (HITS) nas  $N$  iterações, com valor esperado  $E(N) = 0,5N$  e desvio padrão  $S = \sqrt{0,25N}$ . Caso o número de HITS seja suficiente maior que o valor esperado, o atributo é aceite como importante.

A vantagem deste algoritmo à técnica descrita em 2.3.1 é a seleção automática dos atributos a utilizar.

### 2.3.3 Least Absolute Shrinkage and Selection Operator

O método LASSO [68] é um método que efetua regularização e seleção de atributos. Este método adiciona uma restrição, na forma de um limite superior, à soma do valor absoluto dos coeficientes do modelo. Para satisfazer esta restrição o método aplica regularização, onde penaliza os coeficientes reduzindo alguns a zero. Coeficientes com valor superior a zero no final são considerados parte do modelo.

Sendo  $N$  o número de amostras,  $m$  o número de coeficientes,  $y_i$  a classificação e  $x_i$  a  $i$ -ésima amostra, o objetivo final é minimizar o erro utilizando a função:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^N (y_i - x_i^t \hat{\beta})^2 + \lambda \sum_{j=1}^m |\hat{\beta}_j| \quad (2.13)$$

onde  $\lambda \sum_{j=1}^m |\hat{\beta}_j|$  representa a restrição a aplicar aos coeficientes. O parâmetro  $\lambda$  controla a força da penalização, quando  $\lambda$  é suficientemente grande os coeficientes vão ser forçados a ser 0. Quanto maior  $\lambda$  mais coeficientes vão ser diminuídos a 0, e como consequência menor o número de atributos selecionados para o modelo final.

Este método consegue remover coeficientes e reduzir variância sem um aumento substancial no viés [77]. O LASSO tem tendência a apenas selecionar um atributo de um grupo de atributos com correlação elevada, mesmo que estes tenham impacto na classificação final significativa.

### 2.3.4 Elastic Net

*Elastic net* [82] visa a resolver a limitação do **LASSO**, a selecção de apenas um atributo num grupo de atributos com correlação. A solução é adicionar uma penalização quadrática  $\beta^2$ , esta restrição quando utilizada na função de perda sozinha efetua regressão *Ridge*.

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^N (y_i - x_i^t \hat{\beta})^2}{2n} + \lambda \left( \frac{1-\alpha}{2} \sum_{j=1}^m \hat{\beta}_j^2 + \alpha \sum_{j=1}^m |\hat{\beta}_j| \right) \quad (2.14)$$

Ambos *Ridge* ( $\alpha = 0$ ) e **LASSO** ( $\alpha = 1$ ) são casos especiais da técnica *Elastic Net*, o novo parâmetro  $\alpha$  regulariza a mistura entre as duas técnicas.

Esta técnica reduz coeficientes a zero (propriedade do **LASSO**) e combina propriedades do *Ridge* para eliminar o problema da correlação. Este algoritmo é utilizado em conjunto com **SVM** [75].

## 2.4 Conclusão

Existem vários algoritmos de aprendizagem de interesse, isto abre a oportunidade de uma comparação das suas performances para a classificação do território de Portugal Continental. Os algoritmos de *ensemble* mostram performance elevada na classificação de **LULC**, por esta razão é de grande interesse a avaliação do método *Gradient Boosted Trees*, visto que pode ser acelerado por GPU, uma propriedade vantajosa quando se utiliza um grande volume de dados.

A utilização de dados geográficos requer uma análise dos seus atributos. Estes dados são dados de grande volume e de alta resolução, quando combinados com uma área de testes grande pode ser demasiado para alguns algoritmos dado o tempo restrito da investigação. Nem todos os atributos possuem informação suficiente ou benéfica, não é pretendido que os modelos aprendam com ruído nos dados, é esperado que a utilização das técnicas de redução de atributos diminua a complexidade computacional dos modelos sem perder demasiada informação, são aceitáveis diminuições entre 0-5%, caso o tempo de treino seja reduzido, ou até aumentos de performance em alguns casos. A utilização de técnicas automáticas de redução de dimensão dos dados de treino é algo bastante importante, pois vai ser gerado um grande número de atributos e não será prático verificar a correlação dos mesmos manualmente. A redução de atributos sem alterar muito a performance faz com que os algoritmos sejam mais facilmente escaláveis, diminuindo então o seu tempo de treino e de inferência. A redução dos atributos dos dados de treino promovem também a interpretabilidade dos dados, visto que com um número reduzido de atributos será mais fácil analisar manualmente ou graficamente os resultados.

Concluindo, não existe uma clara combinação de algoritmos de classificação e de redução de dimensionalidade que resolvem todos os problemas de classificação **LULC**. Para isso vão ser testadas varias combinações para no final obter a resposta de qual será o melhor para o problema proposto.

## TRABALHO RELACIONADO

Dado a introdução do problema abordado nesta tese é então necessário estudar quais os trabalhos no mesmo âmbito de classificação e avaliar os produtos existentes que contêm informação de estruturas artificiais permanentes. Os trabalhos relacionados servem de base para derivar a metodologia usada, desde os algoritmos a avaliar ao tipo de dados de treino e de referência a utilizar. O capítulo apresenta primeiramente as cartas de ocupação de solo mais utilizadas, uma introdução aos satélites a utilizar, apresenta as fontes de dados mais usadas e pertinentes ao trabalho a realizar na dissertação e conclui com uma síntese de trabalhos relacionados que exploram e demonstram a utilidade destas ferramentas. O capítulo está estruturado da seguinte forma:

- **Cartografia de Ocupação do Solo**, secção 3.1: A primeira secção deste capítulo apresenta a cartografia de ocupação de solo, baseada em cartografia de polígonos. Estas cartas são a base para todo o estudo efetuado.
- **Classificação Utilizando Detecção Remota**, secção 3.2: Esta secção apresenta vários estudos sobre os dois tipos de deteção remota, ativa e passiva, em prol de melhorar classificação LULC. São apresentadas metodologias utilizando deteção remota ativa, Sentinel-1 e deteção remota passiva, Sentinel-2.
- **Índices Radiométricos**, secção 3.4: Aqui são apresentadas as imagens derivadas de produtos de deteção remota passiva e o seu impacto na classificação LULC.
- **Extração de Informação de Textura**, secção 3.3: Seguindo a lógica da secção anterior, é aqui apresentada a extração de informação de textura, desta vez derivada de produtos de deteção remota ativa.

- **Sinergia Entre Detecção Remota Passiva e Ativa**, secção 3.5: A penúltima secção deste capítulo apresenta o potencial da utilização dos dois tipos de deteção remota para problemas de classificação LULC.
- **Carta: Global Human Settlement Layer**, secção 3.6: A secção final apresenta uma carta que visa a criar mapeamentos de áreas urbanas robusto e automático. São apresentadas duas variações das mesmas e o seu papel para a comparação e validação dos resultados desta tese.

### 3.1 Cartografia de Ocupação do Solo

A maneira mais direta da extração de zonas com estruturas artificiais permanentes é a utilização de cartografia LULC. Estes mapas dividem o território português em vários segmentos (polígonos). Toda a área contida num dado polígono partilha a sua classificação LULC.

As principais fontes cartográficas de LULC, atualmente disponíveis para Portugal Continental, são a COS e a carta CORINE land cover (CLC) [52]. Ambas as cartas partilham as definições das classes de solo, sendo a diferença mais significativa entre elas a **Unidade Mínima Cartográfica (UMC)**. Este valor é dado antes da geração dos mapas e define a área mínima de um polígono, por exemplo, com uma UMC de 25 ha apenas podem ser cartografados aglomerados populacionais com área superior num só polígono, para aglomerados mais pequenos estes não são representados e são incluídos numa classe envolvente com uma maior área. Isto significa que quanto maior for a UMC mais objetos e zonas serão classificadas incorretamente. O impacto deste valor pode ser observado na figura 3.1, onde se pode observar que a carta COS 2015, com uma UMC mais pequena, representa estruturas mais detalhadamente, como aglomerados populacionais mais pequenos.

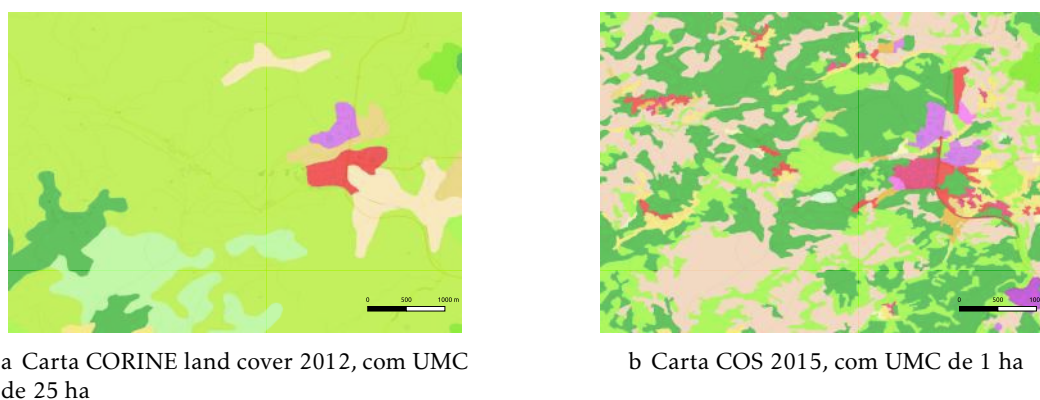


Figura 3.1: Impacto da UMC na representação das classes de ocupação de solo

O programa CO-oRdination of INformation on the Environment (CORINE) foi criado



Figura 3.2: Classes *artificial surfaces* da carta CORINE land cover

em 1985 pelos membros da Comissão Europeia, tendo como objetivo criar uma compilação de informação sobre o estado do ambiente de maneira coordenada entre os seus membros e garantir que estes dados são consistentes e compatíveis. A componente principal do programa são as cartas CLC [66], estas cartas providenciam o único sistema de classificação europeu de cobertura terrestre a longo prazo na Europa, sendo a Direção Geral do Território a entidade responsável pela geração deste mapa em Portugal.

O CLC é composto por uma série temporal de 5 cartas (datadas de 1990, 2000, 2006, 2012 e 2018) <sup>1</sup>. No momento da escrita deste documento a versão de 2018 ainda se encontra em fase de validação <sup>2</sup>.

A nomenclatura possui 44 classes de cobertura de solo agrupadas numa hierarquia de 3 níveis. As classes de nível um são: 1) *artificial surfaces*, 2) *agricultural areas*, 3) *forests and semi-natural areas*, 4) *wetlands*, 5) *water bodies*, os nomes são apresentados em inglês para manter a coerência com a definição das cartas. No escopo desta tese o estudo será focado nas classes de nível 1, 2 e 3 classificadas como *artificial surface*, apresentadas na figura 3.2.

A resolução desta carta não é adequada à extração direta de aglomerados habitacionais com menos de 25 ha de área. Apesar disso, o CLC é frequentemente utilizado em técnicas de classificação de solo baseadas em aprendizagem automática [5, 55]. Como esta dissertação está focada em Portugal Continental, existe a possibilidade da substituição do CLC por um produto mais específico, a carta COS.

A COS é gerada pela mesma entidade responsável pela criação das cartas CLC em Portugal Continental. É essencialmente uma versão mais refinada do CLC, sendo também cartografia de polígonos possuindo uma UMC de 1 ha. Esta carta utiliza a definição de LULC de Caetano et al. 2010 [13], definida como qualquer área de terreno superior ou igual à UMC definida, com distância entre linhas superior ou igual a 20 m e cada polígono classificado seja preenchido a 75% ou superior pela dada classe de ocupação/uso do solo. A carta COS 2015 possui 48 classes no nível 5 e é assegurada a compatibilidade com a

<sup>1</sup>Mapas CLC - <https://land.copernicus.eu/pan-european/corine-land-cover>, visitado a 29/01/2019

<sup>2</sup>Mapa CLC2018 - <https://land.copernicus.eu/pan-european/corine-land-cover/clc2018>, visitado a 29/01/2019

carta CLC. A sua nomenclatura detalhada é apresentada em anexo na sua especificação técnica [21].

### 3.1.1 OpenStreetMap

Devido à UMC das cartas apresentadas anteriormente existem estruturas que não estão presentes, como edifícios singulares e estradas. Dado isto é preciso procurar outras fontes de dados complementares. O projeto OpenStreetMap (OSM)<sup>3</sup> é um projeto que visa a criação de uma plataforma que disponibiliza de dados geográficos grátis e sem restrições de uso. O projeto é baseado em *CrowdSourcing*, onde os seus utilizadores são responsáveis pela criação dos mapas e pela sua qualidade. No âmbito desta dissertação vai ser focado apenas o potencial destes mapas para a extração de estruturas artificiais permanentes, tendo em conta então apenas a informação sobre edifícios e estradas presentes nos mapas.

Como o projeto é baseado nas contribuições dos seus utilizadores, as zonas com menos utilizadores não possuem informação suficiente para a extração de edifícios, como por exemplo zonas rurais. Em contraste, este mapa possui informação sobre estradas de todo o país, até nas zonas rurais, com uma alta confiabilidade [4, 6]. Esta informação de estradas complementarará os dados das cartas de ocupação de solo.

## 3.2 Classificação Utilizando Detecção Remota

Como a extração da informação sobre a localização das estruturas permanentes a partir da cartografia de solo existente não é suficiente para a deteção de estruturas mais pequenas, por exemplo aglomerados e estruturas com menos de 1 hectare de área, tem de ser encontrada outra solução. A técnica mais utilizada e com melhores resultados atualmente é a utilização de algoritmos de aprendizagem automática. Esta secção apresenta várias abordagens ao problema de classificação LULC.

### 3.2.1 Sentinel-1

A missão Sentinel-1 é a primeira constelação de satélites do programa Copernicus conduzido pela Agência Espacial Europeia (AEE), com o primeiro satélite lançado em abril de 2014. Estes satélites captam informação na banda C, utilizando uma técnica ativa de deteção remota. A sua maior vantagem é a penetração das nuvens pelas ondas rádio, assim este tipo de deteção não é afetada pela cobertura de nuvens, chuva ou noite. A constelação, atualmente composta pelos satélites 1A e 1B, produz imagens com uma frequência de 6 dias [64].

O modo de aquisição sobre terra é o *interferometric wide swath*. Este modo adquire dados em polarização dupla (HH+HV e VV+VH) e singular (HH e VV) [2] com uma pegada (*swath*) de 250 km a uma resolução de 5 metros por 25 metros [64]. Os outros

---

<sup>3</sup>OpenStreetMap - <https://www.openstreetmap.org/>



Tabela 3.1: Cenários de teste do estudo de Abdikan et al. [1]

Cenário	Nº de variáveis	Descrição das variáveis
1	1	VV
2	1	VH
3	1	$(VV+VH)/2$
4	2	VV, VH
5	3	VV, VH, VV/VH
6	3	VV, VH, $(VV+VH)/2$
7	3	VV, VH, $(VV-VH)$
8	5	VV, VH, $(VV-VH)$ , $(VV/VH)$ , $(VV+VH)/2$

modos de aquisição, *wave mode*, *wide swath mode* e *strip map mode*, são utilizados para casos de uso fora do escopo da tese, com o modo *strip map mode* apenas usado para casos extremos como gestão de emergências.

A missão Sentinel-1 gera 3 tipos de produtos, produtos de nível-0, 1 e 2. Cada nível apresenta um tipo diferente de processamento dos dados obtidos. Os produtos nível-0 consistem em dados RAS comprimidos e desfocados, este produto é a base para todos os outros de maior nível. Produtos de nível-1 são produtos para a maioria dos utilizadores [69]. Existem vários estudos de exploração do potencial do Sentinel-1 RAS para mapeamento de ocupação de solo [1, 5, 40]. Finalmente, produtos de nível-2 são produtos com informação geofísica derivados do nível-1 que contêm informação sobre vento, ondas e correntes.

Abdikan et al. (2016) [1] concluíram que dados com polarização dupla do Sentinel-1 RAS podem ser explorados com eficiência para produzir mapas de ocupação de solo com grande precisão, mais especificamente mapas que classificam o solo em cinco classes: Água, Urbano (aglomerados artificiais e indústria), Floresta, Agricultura e terra (campos ou minas). Para isso foi utilizado um produto nível-1 sobre a cidade de Istambul, adquirido a dia 16 de setembro de 2016. Utilizando SVM como algoritmo de classificação e dados de treino e validação retirados de imagens de alta resolução do Google Earth, foram criados 8 cenários, apresentados na tabela 3.1. Para cada cenário foi treinado um modelo SVM, após a comparação de resultados os autores concluíram que a polarimetria VV obteve uma precisão mais alta que a VH e utilizar polarimetria dupla aumentou a precisão drasticamente, passando de 70% para valores acima dos 85%. A utilização da diferença de imagens (VV-VH) obteve melhores resultados do que variáveis de média  $(VV+VH)/2$  e rácio  $(VV/VH)$ . Os resultados indicam que imagens com dupla polarização do Sentinel-1 são adequadas para aplicação em extração rápida de informação de LULC.

Baltzer et al, 2015 [5] investigaram o uso de imagens de nível 1 do Sentinel-1 para mapeamento de cobertura terrestre. Para isso foram criadas 27 classes híbridas de nível 2/3 a partir da carta CLC. 17 das quais foram utilizadas nas classificações. Foram criados quatro cenários de classificação utilizando RF como algoritmo. Cada classificação foi efetuada com diferentes tipos de input. Neste estudo foram usados dados da Missão Topográfica

Tabela 3.2: Dados de entrada dos vários classificadores [5].

Classificador	Dados de entrada
RFPOL	HH, HV, VV, VH
RFTEX	HH, HV, VV, VH, textura HH
RFTEXSRTM	HH, HV, VV, VH, textura HH, Mapa de elevação do terreno, declive, orientação
RFSRTM	Mapa de elevação do terreno, declive, orientação

**Radar Shuttle (SRTM)** para complementar a informação do Sentinel-1, nomeadamente um mapa de elevação de terreno e as métricas derivadas declive e orientação. Para estender a informação do Sentinel-1 foi derivada informação de textura da banda HH, a textura possui informação sobre o padrão de variações dos pixels numa imagem, será falado em mais detalhe na secção 3.3. Os quatro cenários a avaliar foram: **RFPOL** (random forest e dados **RAS**), **RFTEX** (random forest, dados **RAS** e de textura), **RFTEXSRTM** (random forest, dados **RAS**, textura e de elevação) e **RFSRTM** (random forest e dados de elevação). Os diferentes tipo de cenários de teste são apresentados na tabela 3.2.

Ambos os classificadores **RFPOL** e **RFTEX** confundem algumas zonas agrícolas com zonas florestais, como esta dissertação vai ser focada na deteção de estruturas artificiais permanentes este erro não é muito relevante.

As classes 111 (tecido urbano contínuo), 112 (tecido urbano descontínuo) e 120 (unidades de transporte, indústria e comercial) possuem valores de textura e banda C-HH mais elevados quando comparados a outras classes. A classe 140 (áreas de vegetação artificial) também possui um valor de textura mais elevado que classes não artificiais. O estudo concluiu que a utilização de dados provenientes do Sentinel-1, informação de textura HH e os quatro produtos derivados da missão **SRTM (RFTEXSRTM)** possuem a melhor precisão de classificação (68,4%). A sinergia entre dados da missão **SRTM** e do Sentinel-1 permitem uma melhor classificação das classes da carta **CLC**. Este estudo utilizou o maior número de classes até à data no que toca a classificação **LULC**.

### 3.2.2 Sentinel-2

A constelação Sentinel-2 foi desenvolvida e está a ser conduzida pela **AEE**, com o primeiro satélite lançado em Junho de 2015. Por contraste ao Sentinel-1, os satélites desta missão possuem sensores passivos. A missão Sentinel-2 captura dados multi-espectrais em 13 bandas nas partes visíveis, **IP** e **Infravermelho de Ondas Curtas (IOC)** do espectro. Estas bandas possuem uma resolução espacial de 10, 20 ou 60 metros [8]. Informação mais detalhada sobre as bandas encontra-se na tabela 3.3.

Uma das grandes desvantagens deste tipo de captura de imagens é a sua dependência das condições climáticas, nomeadamente a cobertura de nuvens. O Sentinel-2 têm dois tipos de produtos acessíveis, produtos nível-1C e 2A. Estes produtos possuem uma granularidade de 100x100km, cada imagem contém toda a informação possível de todas

Tabela 3.3: Informação detalhada das bandas da constelação Sentinel-2

Bandas Sentinel-2	Sentinel-2		
	Comprimento de onda central (nm)	Largura de banda (nm)	Resolução espacial (m)
Banda 1 - Aerosol costeiro	442,7	21	60
Banda 2 - Azul	492,4	66	10
Banda 3 - Verde	559,8	36	10
Banda 4 - Vermelho	664,6	31	10
Banda 5 - Vegetation red edge	704,1	15	20
Banda 6 - Vegetation red edge	740,5	15	20
Banda 7 - Vegetation red edge	782,8	20	20
Banda 8 - NIR	832,8	106	10
Banda 8A - NIR médio	864,7	21	20
Banda 9 - Vapores de água	945,1	20	60
Banda 10 - SWIR - Cirros	1373,5	31	60
Banda 11 - SWIR	1613,7	91	20
Banda 12 - SWIR	2202,4	175	20

as bandas espectrais. Os produtos nível-1C contêm imagens com reflectância topo da atmosfera, estas imagens contêm informação sobre luz refletida da superfície terrestre e da atmosfera, o que lhes dá uma aparência nebulosa e azulada. Para a geração de produtos nível-2A é aplicada correção atmosférica ao produto 1C, o que resulta em imagens com reflectância à superfície terrestre (imagens corrigidas). A correção pode ser aplicada pelo utilizador utilizando software disponibilizado pela [AEE](#).

A boa resolução destes satélites tornam-nos apelativos, como esta dissertação é focada na deteção de estruturas artificiais permanentes. A resolução das imagens utilizadas é algo crítico, pois a abordagem ao problema pode mudar drasticamente dependendo da resolução. Se uma estrutura tiver tamanho inferior a um pixel na imagem capturada pode ser impossível de detetar.

Pesaresi et al. 2016 [55] mostram que a utilização das bandas de grande resolução de 10m do Sentinel-2 aumentam a qualidade da classificação de mapas [LULC](#) devido ao detalhe espacial acrescido. A área de estudo selecionada foi sobre a cidade de Porto Viro na área do delta do rio Pó, foi utilizado um produto nível-2A do Sentinel-2. Como dados de treino para o classificador utilizaram uma combinação de atributos espectrais e texturas, quatro bandas de 10m do Sentinel-2 e quatro bandas de textura geradas a partir das quatro bandas de entrada.

Os dados de referência foram gerados com base na carta [CLC](#) de 2006. Para efeitos de classificação foram criadas duas classes *build-up* e *water*. A classe *build-up* foi extraída desta carta ao juntar as classes *continuous urban fabric*, *discontinuous urban fabric*, e *industrial or commercial units*. A classe *water* foi extraída ao agregar as classes *inland* e *marine waters*.

É importante referenciar a existência de uma diferença relevante entre os dados das classes de treino e dos dados de imagem. A carta *CLC* utilizado tem uma diferença de 9 anos da imagem retirada do Sentinel-2 e também possui uma escala diferente, a *CLC* 2006 tem uma resolução espacial de 100m e a imagem da região de interesse possui uma resolução de 10m. Isto implica que existem diferenças significativas entre a carta *CLC* e as imagens a classificar, o que irá afetar os resultados finais. Por exemplo, a classe *discontinuous urban fabric* inclui áreas com vegetação ou solos sem estruturas artificiais permanentes que serão capturadas pelo Sentinel-2.

Após a classificação, o classificador obteve uma precisão balanceada de 80%. Esta medida é utilizada para prevenir a tendência de outros valores representarem mal os resultados ao lidar com dados desequilibrados. No caso da classificação da classe *built-up* existem poucos exemplos positivos quando comparado com a classe *non-built-up*.

A interpretação visual manual dos autores, após a classificação, concluiu que existe um detalhe acrescido no que toca a informação temática dos dados de treino da carta *CLC*. É notado também que dados da classificação contrariam os dados de treino. Aglomerados populacionais dispersos foram classificados como *built-up* mesmo sem existirem nos dados de referência. Foi observado que o classificador obteve mais erros de classificação nas zonas perto da costa, onde dunas de areia são classificadas como *built-up* devido ao comportamento espectral semelhante a estruturas artificiais. Os erros de omissão do classificador estão relacionados com o edifícios isolados com tamanho médio mais pequeno que a resolução do Sentinel-2 de 10m. Os autores concluíram que a utilização de imagens provenientes do Sentinel-2 são atraentes no que toca ao mapeamento de áreas populacionais devido à sua alta resolução espacial.

O estudo de Marangoz 2017 [43], compara imagens da constelação Sentinel-2 com outros dados de média resolução (Landsat-8) e estudam o impacto das mesmas na performance de uma classificação *LULC*, utilizando classificação baseada em objetos da ferramenta *eCognition*<sup>4</sup>. O estudo foi realizado numa pequena região de Zonguldak, Turquia. Esta zona possui terreno íngreme e acidentado composto maioritariamente por zonas de floresta e agricultura. Devido a esta zona ter uma distribuição parecida à região de testes torna este estudo bastante relevante. Como classes para a tarefa de classificação são utilizadas cinco classes *LULC* gerais, zonas habitacionais, corpos de água, solo, floresta e vegetação. A imagem Sentinel-2 foi adquirida a 6 de abril de 2016 e recortada a uma área de 60 por 60 km, para a classificação foram extraídas apenas as bandas do espectro visível e a banda dos infravermelhos próximos, para comparação os autores extraíram as bandas equivalentes do satélite de média resolução. Estas foram expostas a um processamento para aumentar a sua nitidez, para uma resolução de 15 por 15 metros, contrastando com a resolução nativa da constelação Sentinel-2 de 10 por 10 metros.

No final o classificador que utiliza produtos Sentinel-2 obteve um Kappa de 0,74 versus o Kappa de 0,66 das imagens Landsat-8, concluindo que os produtos Sentinel-2

---

<sup>4</sup>eCognition <http://www.ecognition.com/> - visitado a 13-08-2019

possuem um elevado potencial para a classificação LULC. Curiosamente, a classe zonas habitacionais, obteve uma classificação muito mais baixa ao utilizar produtos Sentinel-2, obtendo um Kappa de 0,5 comparado com um Kappa de 0,91 dos dados de média resolução, esta foi a única classe a apresentar uma descida de métricas com dados de melhor resolução. É concluindo que, as cartas urbanas derivadas de produtos Sentinel-2 apresentam melhores resultados para planeamento urbano e promovem a sustentabilidade.

Num estudo relacionado, de Sekertekin et al. 2017 [62], é investigado também o potencial das imagens de alta resolução da constelação Sentinel-2 quando comparadas a imagens de média resolução. Desta vez os autores utilizam um metodologia de classificação baseada em pixels, onde cada pixel representa uma amostra. O estudo é efetuado sobre a mesma área, Zonguldaque, Turquia, utilizando as mesmas bandas, espectro visível e infravermelho próximo. Em contraste com outros estudo da área os autores utilizam um classificador de máxima verosimilhança.

Quando comparadas, as duas classificações apresentam um Kappa de 0,85 e 0,78, para os dados Sentinel-1 e Landsat-8 respetivamente. Em geral, as métricas aumentam com o aumento da resolução dos dados para todas as classes, contrastando com o estudo anterior, a classe zonas habitacionais sofre um aumento nas suas métricas métricas com a subida de resolução dos dados, chegando a obter uma diferença de 15% na sua exatidão. Apesar dos melhores resultados da constelação Sentinel-2, as imagens Landsat-8 apresentam resultados melhores em algumas zonas, especialmente em zonas sobre o mar.

É possível concluir que, ao utilizar a mesma metodologia de processamento, uma abordagem que considera cada pixel como uma amostra tem um potencial mais elevado para a classificação LULC utilizando dados da constelação Sentinel-2 de alta resolução. Quando comparadas, as duas metodologias obtêm um Kappa de 0,85, baseada em pixels, e 0,74, baseada em objetos.

### 3.3 Extração de Informação de Textura

A textura é descrita como sendo o padrão de variações de intensidade numa imagem, este tipo de análise tem em conta o contexto envolvendo um pixel e não apenas o próprio, o que pode ser uma grande ajuda na classificação de solo devido à alta correlação de pixels vizinhos quando se utiliza informação geográfica. Um dos métodos mais utilizados é a **Gray Level Co-occurrence Matrix (GLCM)**, proposta por Haralick et al. 2007 [32]. A **GLCM** mede a probabilidade da ocorrência de dois níveis de cinza, numa dada imagem, separados por uma certa distancia e direção. Existem vários atributos possíveis para extração. Cada um modela diferentes estatísticas dentro de uma janela deslizante de tamanho fixo, como por exemplo 5 por 5 pixels, que percorre a **GLCM**. Os atributos podem ser divididos em três grupos, contraste, ordem e estatística:

- **Atributos de contraste:** Contrast; Dissimilarity (DIS); Homogeneity (HOM).



Figura 3.3: NDWI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.

- **Atributos de ordem:** Angular Second Moment (ASM); Maximum Probability (MAX); Entropy (ENT).
- **Atributos estatísticos:** GLCM Mean; GLCM Variance; GLCM Correlation.

A extração de informação de textura está descrita mais detalhadamente em Hall-Beyer 2017 [31].

### 3.4 Índices Radiométricos

Para além do uso direto dos valores das diversas bandas capturadas nas imagens de satélite, existem vários índices que podem ser calculados e integrados nos algoritmos de classificação, nomeadamente: *Normalized Difference Water Index (NDWI)*, *Normalized Difference Built-up Index (NDBI)*, *Enhanced Vegetation Index (EVI)* e *Normalized Difference Vegetation Index (NDVI)*. Estes índices espectrais aproveitam as vantagens do comportamento da reflectância dos objetos em comprimentos de onda diferentes e possuem uma escala de -1 a 1. Por exemplo o *NDBI* assume uma maior reflectância na zona *IOC* do espectro do que na zona *IP*, este comportamento é característico de zonas com superfícies artificiais.

#### 3.4.1 Normalized Difference Water Index

O *NWDI* [27, 44], apresentado na Figura 3.3, é usado para identificar águas abertas, ou para medir o teor de água na vegetação. Este índice tem uma correlação negativa com áreas urbanas e superfícies impermeáveis [17, 76]. Na Figura 3.3, pode ser observado que zonas mais perto do valor 1 representam corpos de água. As zonas mais perto do -1 representam zonas de rochedo, vegetação reduzida e aglomerados populacionais. Calculado como:

$$NDWI = \frac{Verde - NIR}{Verde + NIR} \quad (3.1)$$



Figura 3.4: NDBI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.

### 3.4.2 Normalized Difference Built-up Index

O NDBI [78] é utilizado para identificar áreas urbanizadas e tem uma correlação positiva com as mesmas bem como superfícies impermeáveis [17]. Calculado com a reflectância na zona IP e na zona IOC do espectro. O índice está representado na Figura 3.4. Zonas a vermelho, representam zonas com alta impermeabilidade do solo, como habitações e aglomerados populacionais.

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (3.2)$$

### 3.4.3 Índices de vegetação

NDVI [71] e a sua variante o EVI [34] produzem índices espectrais que indicam se uma dada área possui vegetação verde ou não, como consequência têm uma correlação negativa com áreas urbanas.

$$NDVI = \frac{NIR - Vermelho}{NIR + Vermelho} \quad (3.3)$$

O EVI é um índice de vegetação otimizado. Este possui uma sensibilidade mais alta em zonas onde o NDVI tende a saturar, nomeadamente zonas com um nível elevado de biomassa [34]. Apresentados nas Figuras 3.5 e 3.6, respectivamente.

$$EVI = G * \frac{(NIR - Vermelho)}{(NIR + C1 * Vermelho - C2 * Azul + L)} \quad (3.4)$$

## 3.5 Sinergia Entre Detecção Remota Passiva e Ativa

Os estudos apresentados na secção 3.2 estudam o potencial de apenas um tipo de deteção remota para classificação de LULC, satélite ativo ou passivo. Apesar destes dados serem complementados com outros tipos de dados, nomeadamente MDE e dados texturais, é de interesse estudar o impacto da utilização de dois tipos de satélite em simultâneo. No



Figura 3.5: NDVI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.

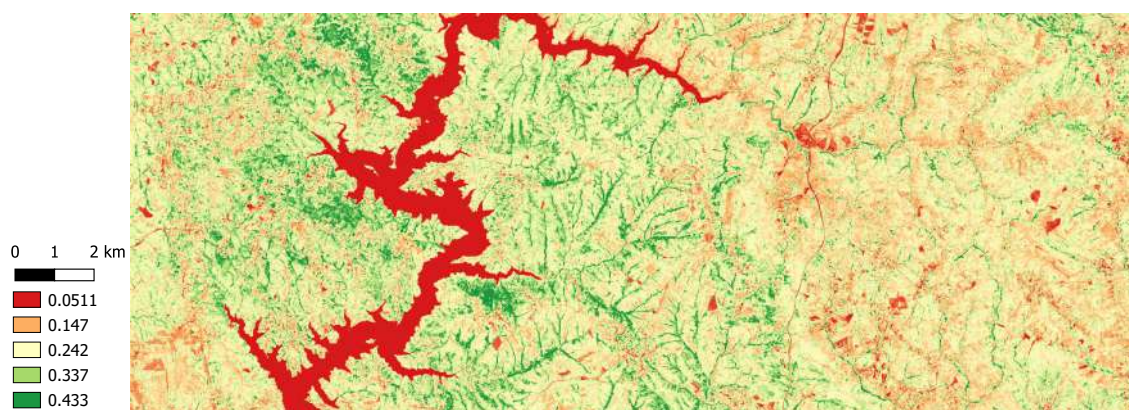


Figura 3.6: EVI de uma imagem Sentinel-2 capturada a 29-07-2016 sobre o Rio Zêzere.

âmbito desta dissertação, os dois satélites com melhor resolução, ou seja mais adequados à deteção de estruturas artificiais permanentes, são o Sentinel-1 e Sentinel-2.

De acordo com um estudo por Clerici et al. [19], a sinergia entre dados das missões Sentinel-1 e 2 permite uma melhor classificação de ocupação de solo. No estudo, realizado sobre uma região com  $2095\text{km}^2$  dentro de Cundinamarca, Colômbia, foram utilizadas dois produtos de cada um dos satélites de modo a cobrir toda a área.

Todos os dados recolhidos foram processados a uma resolução 10 por 10 metros, os autores utilizam também dados de textura derivados dos produtos Sentinel-1. Como objetivo final os dados LULC são divididos nas seguintes classes: floresta, água, estruturas permanentes, cultivo, pastagem e arbustos.

Foi concluído que a metodologia de classificação utilizando Sentinel-1 e Sentinel-2 possui duas vantagens principais: As cadeias de processamento de ambos os satélites estão bem suportadas pelas AEE e apresentam uma cadeia de processamento rápida. Em segundo, a integração de ambos dados de textura e dados espectrais produzem classificações LULC satisfatórias, chegando a ter uma exatidão geral de 88% e um Kappa de 0,86, quando classificados utilizando SVM.

As métricas apresentadas pelos autores, utilizando o classificador SVM, mostram que



dados de Sentinel-2 possuem mais informação para a classificação do que dados Sentinel-1. Quando separados, dados Sentinel-2 apresentam exatidão geral de 72% e Kappa de 0,67, dados Sentinel-1 apresentam exatidão geral de 30% e Kappa de 0,15. Apesar disso a fusão de ambas as fontes de dados aumenta significativamente a performance dos algoritmos.

No final, é concluído também, em semelhança ao estudo de Baltzer et al. [5], que a utilização de dados de séries temporais tem potencial para promover uma melhor discriminação entre classes, e é algo a testar para melhorar a classificação de LULC.

Tavares et al. 2019 [67] mostram que a integração de dados Sentinel-1 e Sentinel-2, obtêm os melhores resultados na classificação LULC. Este estudo foi feito sobre uma área urbana de Belém, Amazônia Oriental, com  $1059,46\text{km}^2$ . Compararam seis tipos de estruturação de dados de ambos os satélites para treinar um classificador *Random Forest*. Sentinel-1 apenas, Sentinel-2 apenas, S-1 com informação de textura, Sentinel-2 com índices radiométricos, Sentinel-1 e Sentinel-2 e finalmente todos os produtos juntos. Para a criação destes dados, foi utilizado um produto Sentinel-2 1C, com 0% de nuvens sobre a área de estudo, e dois produtos Sentinel-2 com polarização dupla (VV + VH). Após a recolha os produtos são então processados para uma resolução de 10 por 10 metros. Os autores dividem a classificação de solo em 12 classes, Agricultura, Aeroportos, Solo, Praias, Estruturas artificiais permanentes, Pastagem, Autoestrada, Minas, Vegetação primária, Vegetação urbana, Água com sedimentos e Água sem sedimentos.

No final, a utilização de dados Sentinel-1 e Sentinel-2 obtêm a melhor performance na classificação, obtendo uma exatidão de 91,07% e um Kappa de 0,8709. Curiosamente, a utilização de todos os produtos recolhidos obteve piores resultados do que a integração dos dados Sentinel-1 e 2 apenas. Os piores resultados foram obtidos ao utilizar apenas dados radar, Sentinel-1. Apesar da vantagem da penetração das nuvens, este tipo de dados não permite uma distinção de classes satisfatória.

### 3.6 Carta: Global Human Settlement Layer

É de grande importância referir também o projeto de classificação de solo, *Global Human Settlement Layer* (GHSL) [54]. O projeto GHSL tem como âmbito fornecer métodos e sistemas para o mapeamento robusto e automático de áreas urbanas [55]. Este possui atualmente dois produtos de interesse, a carta *GHSL Landsat*, gerada utilizando *Symbolic machine learning* treinando em dados Landsat 8, com uma resolução final de 30 por 30 metros e a carta *GHSL Sentinel-1 Built-up*, gerada utilizando o mesmo modelo mas em dados Sentinel-1 e possui uma resolução final de 20 por 20 metros, uma comparação de ambas as cartas está presente na figura 3.7.

Nas cartas *Global Human Settlement Layer* (GHSL) a classe *built-up*, que representa estruturas artificiais permanentes, é definida como a união de todas as unidades espaciais, pixels, contendo uma estrutura artificial permanente ou parte dela. Dito isto estas cartas não possuem nenhuma diferenciação de estruturas, como por exemplo estrutura rural

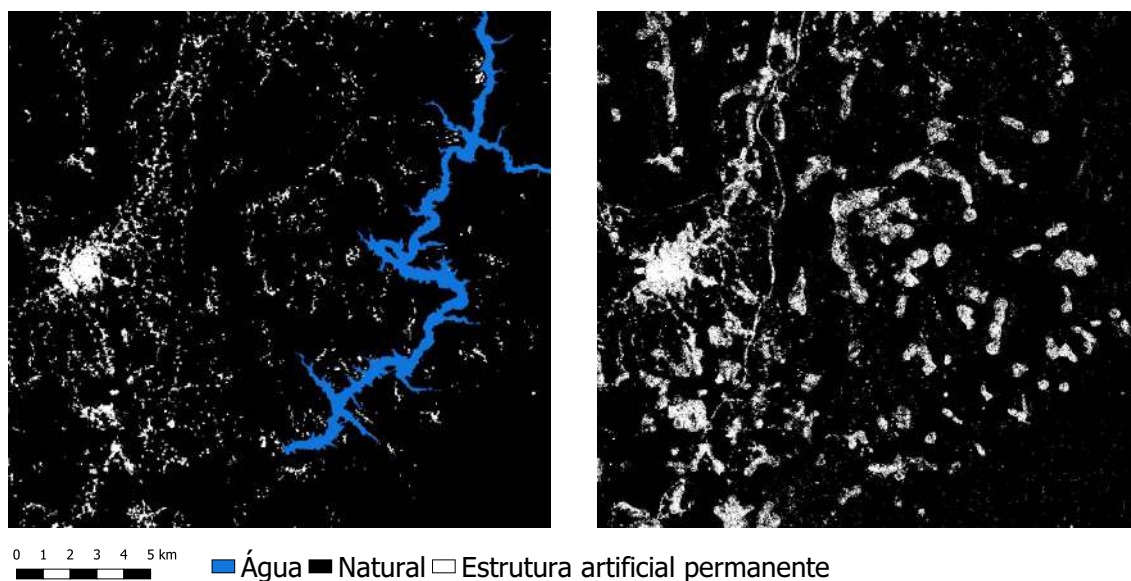


Figura 3.7: Comparação das cartas GHSL Landsat, à esquerda e GHSL Sentinel-1, à direita, sobre a região de Tomar.

ou urbana, alguns tipos de análise feitos sobre estruturas artificiais podem depender do tipo de estrutura, como por exemplo a delimitação das FGC em que a largura das mesmas depende da estrutura em questão. Este tipo de informação normalmente não é representado na maioria das cartas de estruturas artificiais permanentes [74].

Para além dos dados de referência utilizados para treinar os algoritmos, as cartas GHSL serão utilizadas para fazer uma comparação com os resultados obtidos, visto que o objetivo final do projeto GHSL vai de encontro com o objetivo desta dissertação.

### 3.7 Resumo

Neste capítulo foi sintetizado o resultado do estudo sobre as tecnologias, métodos e tipos de dados mais utilizados com impacto na deteção de estruturas artificiais permanentes. As três componentes com maior impacto para o problema exploradas foram **Cartografia de ocupação de solo**, **Classificação utilizando deteção remota** e os seus **Produtos derivados**.

Começando com cartografia de ocupação de solo, estes mapas apresentam a maneira mais direta e simples de extração de edifícios, ao extrair as classes *artificial surfaces* estas podem posteriormente ser utilizadas como uma máscara sendo depois sobrepostas a imagens de satélite. Esta abordagem tem vários problemas. Os mapas não são adequados para uma abordagem em tempo real devido à sua baixa resolução temporal; cada mapa de ocupação de solo pode até demorar vários anos a ser gerado. Como foi referido na secção 3.1, este tipo de cartografia possui também uma resolução espacial não indicada ao problema desta tese devido a existirem estruturas com uma área menor que a unidade mínima cartográfica dos mapas apresentados. Um dos pontos fortes da aplicação das

técnicas estudadas a Portugal continental é a utilização da carta COS. Esta carta irá permitir a construção de modelos mais refinados com a possibilidade de deteção de estruturas mais pequenas.

Para efeitos de classificação de cartas de ocupação de solo são utilizadas técnicas de aprendizagem automática, como dados de treino, é recorrente a utilização de produtos Sentinel-1 e Sentinel-2, em geral, dados derivados de produtos Sentinel-2 possuem mais informação à cerca do solo e uma melhor separabilidade dos atributos das várias classes, o que faz com que as bandas destes produtos sejam mais importantes. Mesmo assim, a utilização de produtos Sentinel-1 é algo importante que pode aumentar a performance do classificador. Regra geral, a utilização de produtos derivados, por exemplo dados de textura extraídos do Sentinel-1 ou índices radiométricos do Sentinel-2, podem aumentar a performance classificação. Apesar disso não é algo conclusivo, podendo até descer a mesma em alguns casos.

Concluindo, existe uma grande variedade de opções e produtos que podem ser explorados de modo a melhorar a deteção de estruturas artificiais permanentes. Esta variedade de produtos de alta resolução disponíveis possui um potencial elevado para a deteção de estruturas artificiais permanentes quando combinados com metodologias de aprendizagem automática.



## ABORDAGEM

É apresentado neste capítulo a abordagem escolhida de modo a atingir os objetivos propostos. A solução final será focada na detecção de estruturas artificiais permanentes e a posterior análise do potencial da mesma para a delimitação das FGC. Esta solução irá explorar o estudado no capítulo 3, ao utilizar dados provenientes do Sentinel-1, Sentinel-2 e modelos de elevação do terreno. Será também dada prioridade à resolução espacial de modo a obter uma melhor detecção de estruturas mais pequenas. O capítulo, dividido em 5 partes, está segmentado da seguinte forma:

- **Formulação do problema de aprendizagem**, secção 4.1: Primeiramente são definidos formalmente, utilizando a formulação de Tom Mitchell [45], os vários problemas de aprendizagem da dissertação.
- **Problemas e estrutura da solução**, secção 4.2: Nesta secção é feita uma descrição da solução e dos problemas a resolver, dividindo os mesmos nos seus componentes individuais.
- **Implementação e Bibliotecas**, secção 4.3: Para dar a entender melhor as experiências a efetuar esta secção apresenta as implementações dos algoritmos utilizados, e enumera as bibliotecas utilizadas durante a dissertação.
- **Metodologia**, secção 4.4: Finalmente, esta secção apresenta em detalhe a metodologia utilizada, desde o processamento dos dados de satélite até à otimização dos algoritmos.
- **Conclusão**, secção 4.5: O capítulo é fechado com uma breve conclusão.

## 4.1 Formulação do problema de aprendizagem

Começando por definir mais formalmente o problema em análise, Tom Mitchell em 1997, [45] define formalmente um problema de aprendizagem da seguinte forma: “Diz-se que um programa aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas em T, como medido por P, melhora com a experiência E.”

Utilizando então esta formalização, podemos definir T, P e E como:

- **Tarefa (T):** Classificação de pixels de imagens como sendo ou não estruturas artificiais permanentes.
- **Experiência (E):** Conjunto de imagens de satélites com sensores ativos e passivos e seus derivados. Complementado com uma carta de ocupação de solo que mapeia estruturas artificiais permanentes.
- **Performance (P):** Estatística Kappa da classificação, Precisão e *Recall* por classe.

Dito isto o problema apresentado nesta dissertação está dividido em três tarefas:

- T1: Classificação dos pixels como sendo estruturas artificiais permanentes, natural ou água.
- T2: Classificação dos pixels como sendo estruturas artificiais urbanas, estruturas artificiais rural, outras estruturas, natural ou água.
- T3: Classificação dos pixels como sendo estruturas artificiais urbanas, estradas, natural ou água.

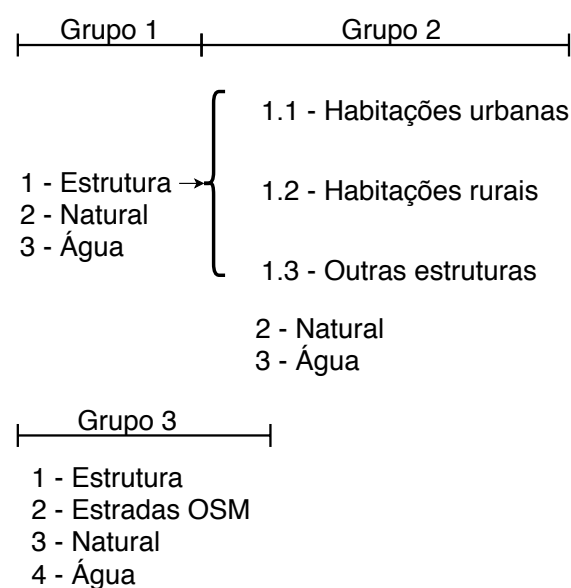


Figura 4.1: Grupos de classes

Estas tarefas correspondem aos grupos dos dados de referência grupo 1, 2 e 3, respectivamente, representados na Figura 4.1. O grupo 1 consiste na separação das estruturas artificiais permanentes do restante, o grupo 2 é uma especialização do grupo 1, dividindo as estruturas em três subclasses, finalmente o grupo 3 adiciona, ao grupo 1, a informação de estradas da carta OSM.

Cada tarefa T terá também variações da experiência E, ou seja, varias estruturações dos dados de treino, sendo o único componente constante ao longo dos testes a performance P.

### 4.2 Problemas e estrutura da solução

Dado a formulação do problema, é agora apresentado o mesmo de maneira mais intuitiva. Após o estudo, foram derivados 5 pontos principais a seguir para obter uma solução robusta a um problema de classificação. Para resolver o dado problema de classificação utilizando metodologias de aprendizagem automática supervisionadas, os seguintes passos principais que têm de ser abordados:

- Selecionar o tipo de dados de treino (amostras). No caso do problema apresentado nesta tese, as amostras utilizadas são conjuntos de pixels em que o vetor de atributos de uma amostra é composto pelas fontes de diversas de dados estudados.
- Extrair os dados de referência. Este passo poderia ser feito manualmente, como em alguns artigos estudados, mas a existência de mapas como o CLC e o COS facilitam este passo. No escopo desta dissertação estes dados foram retirados do mapa COS de 2015 e agrupados em diferentes tarefas de classificação.
- Determinar quais os atributos a utilizar. Estes atributos representam um pixel na classificação. Na fase inicial as bandas de 10m do Sentinel-2 serão o foco principal devido à sua alta resolução espacial. Dito isto, todos os restantes produtos serão reamostrados para uma resolução de 10 por 10 metros.
- Escolher o algoritmo de aprendizagem automática. Este algoritmo será responsável pela classificação do mapa final. Como não existe um algoritmo ótimo, todos os algoritmos serão treinados e comparados.
- Refinar o algoritmo. Nesta fase o algoritmo é corrido nos dados de treino e avaliado. Será também necessário ajustar os parâmetros dos modelos de modo a obter uma boa adaptação dos mesmos aos dados.

É esperado também que a solução seja escalável a Portugal Continental. Para isso serão utilizadas também tecnologias assente em GPUs. Problemas de aprendizagem automática possuem um requisito computacional elevado e para isso é importante utilizar estas tecnologias para acelerar o processo de aprendizagem. Existem várias ferramentas

disponíveis para este efeito, como por exemplo o Google Colab <sup>1</sup>, Tensorflow GPU <sup>2</sup> e o Google Earth Engine<sup>3</sup>. É importante notar que o requerimento da utilização de GPUs para aceleração do desenvolvimento é independente da solução proposta.

### 4.2.1 Solução proposta

A solução proposta, vem de encontro a resolver os problemas observados e responder aos objetivos da dissertação. Esta está dividida em 4 componentes gerais, **pré-processamento**, **aprendizagem**, **avaliação** e **pós-processamento**, cada componente terá a seu objetivo e estão divididos em sub-tarefas a completar. Foi escolhida uma implementação utilizando uma metodologia iterativa: a fase de aprendizagem e avaliação serão iteradas de modo a obter os melhores parâmetros e a melhor performance possível. Apresentado em mais detalhe, a solução proposta possui os seguintes componentes:

- **Pré-processamento:** Esta componente será focada no processamento de imagens e geração de outros índices e informação a partir das imagens de satélite originais. O tratamento das imagens tem como objetivo remover efeitos de dispersão causados pela atmosfera, aplicar cadeias de processamento padrão para alguns produtos e remover produtos com contaminação de nuvens. Serão feitas também todas as re-estruturações dos dados necessárias de modo a estarem prontos para consumo dos algoritmos de aprendizagem.
- **Aprendizagem:** A componente de aprendizagem é a componente mais importante e mais pesada desta solução, pois é aqui que os modelos vão ser treinados com várias combinações dos dados extraídos, experiência E, em prol de resolver as tarefas T apresentadas.
- **Avaliação:** Uma das componentes mais importantes da solução. É aqui que a performance P dos modelos será avaliada utilizando métricas estudadas. Será também feita uma interpretação manual do raster da classificação de cada modelo. No final do cálculo das métricas e da avaliação manual será dado um resultado binário: a solução tem a performance suficiente para a solução do problema proposto, ou volta à fase de aprendizagem.
- **Pós-processamento:** No final, após o mapa de cobertura de solo de estruturas artificiais ser gerado, este pode ser alvo de várias transformações, como por exemplo reduzir a sua fragmentação, transformar o raster gerado num produto vetorial. Com o processamento feito entra então na fase final, a geração das FGC envolvendo as estruturas.

---

<sup>1</sup>Google Colab - <https://colab.research.google.com> visitado a 13/03/2019

<sup>2</sup>Tensorflow GPU - <https://www.tensorflow.org/install/gpu> visitado a 13/03/2019

<sup>3</sup>Google Earth Engine - <https://earthengine.google.com/> visitado a 13/03/2019



### 4.3 Ambiente experimental

Antes de entrar em maior detalhe na solução proposta, é preciso descrever alguns detalhes de implementação, como bibliotecas utilizadas e quais as implementações dos algoritmos. Isto leva a um melhor entendimento do processo e possibilita a replicabilidade dos resultados obtidos. Para toda a implementação foi escolhida a linguagem de programação Python. Esta linguagem apresenta um ambiente de alto nível adequado a projetos de investigação e é também muito utilizada na área de aprendizagem automática, possuindo várias bibliotecas com a implementação de todos os algoritmos necessários para esta tese.

A maior parte dos algoritmos utilizados está implementado na biblioteca scikit-learn. Esta é a biblioteca escolhida para todos os algoritmos que não tenham implementações em GPU. A tabela 4.1 apresenta os algoritmos que possuem implementações em CPU versus os que tem implementações em GPU.

Tabela 4.1: Suporte dos algoritmos para CPU e GPU.

CPU	GPU
Support Vector Machine	Gradient Tree Boosting
Stochastic Gradient Descent	Dense Neural Network
Random Forest	

Para os algoritmos com implementação em CPU foram escolhidas as seguintes implementações da biblioteca scikit-learn:

- Support Vector Machine - `sklearn.svm.SVC`;
- Stochastic Gradient Descent - `sklearn.linear_model.SGDClassifier`;
- Random Forest - `sklearn.ensemble.RandomForestClassifier`.

Em contraste, para os algoritmos acelerados com GPUs, foram escolhidas as seguintes implementações:

- Dense Neural Network - `tensorflow.keras`;
- Gradient Tree Boosting - `xgboost.xgb`.

Para efeitos de replicação as bibliotecas mais importantes utilizadas nos testes são as presentes na Tabela 4.2.

Tabela 4.2: Bibliotecas e implementações utilizadas

scikit-learn 0.21.2	Boruta 0.1.5
tensorflow-gpu 1.13.1	GDAL 2.3.3
xgboost 0.82	numpy 1.16.4
CUDA Toolkit 10.1	scipy 1.2.1

Finalmente, de modo a justificar e adicionar robustez aos resultados, principalmente resultados que envolvam tempos de execução, é necessário apresentar a máquina de testes usada nesta dissertação, possuindo os seguintes componentes:

- Intel Core i5 6600K @ 4.20 GHz
- NVIDIA GTX1070 8GB
- 16GB RAM DDR4 @ 2400MHz

### 4.4 Metodologia

Entrando agora em mais detalhe na estruturação da solução proposta, a Figura 4.2 apresenta a cadeia de processamento utilizada durante todo o processo de testes, desde o tratamento dos dados recolhidos até ao treino dos algoritmos. Começando com a captura dos dados, apenas os produtos Sentinel-1 foram processados, utilizando uma cadeia de processamento padrão. Após o processamento, todas as imagens de todas as fontes de dados são modificadas de modo a obter uma resolução de 10 por 10 metros no [Sistema de Referência de Coordenadas \(SRC\) EPSG:32629](#). Finalmente, as imagens resultantes da cadeia de processamento são agregadas para consumo final do algoritmo, a fase de agregação depende da natureza dos dados, na abordagem estática esta agregação é apenas um empilhamento de todas as imagens recolhidas, na abordagem que utiliza séries temporais, a agregação é uma série de dados estatísticos sobre cada série de produtos.

#### 4.4.1 Aprendizagem

É necessário agora virar a atenção para os algoritmos de aprendizagem. Qualquer treino de um algoritmo de aprendizagem precisa de repartir os dados de entrada. Para evitar enviesar as métricas finais é necessário utilizar amostras do conjunto de dados não utilizadas no treino. Primeiramente, são extraídos 20% dos pixels da região de interesse, este é o conjunto de dados principal visto que utilizar todos os pixels da imagem da região escolhida é algo demasiado pesado para a máquina de testes utilizada. Ao utilizar o conjunto de dados principal, para o treino e otimização dos modelos de classificação este conjunto é dividido em dois, dados de treino (80%) e dados de teste (20%).

##### 4.4.1.1 Otimização dos parâmetros dos classificadores

Cada classificador possui vários hiperparâmetros que devem ser ajustados de modo aos modelos se adaptarem aos dados. Apesar de existirem várias sugestões de como escolher um bom parâmetro inicial, é necessário fazer uma pesquisa no espaço de valores possíveis de modo a encontrar os parâmetros mais adequados ao problema. Foram considerados dois tipos de pesquisa, pesquisa aleatória e pesquisa utilizando força bruta. Na primeira, o algoritmo de pesquisa itera varias permutações de valores aleatórios e no final são apresentados os parâmetros com melhores métricas de avaliação. Com pesquisa utilizando

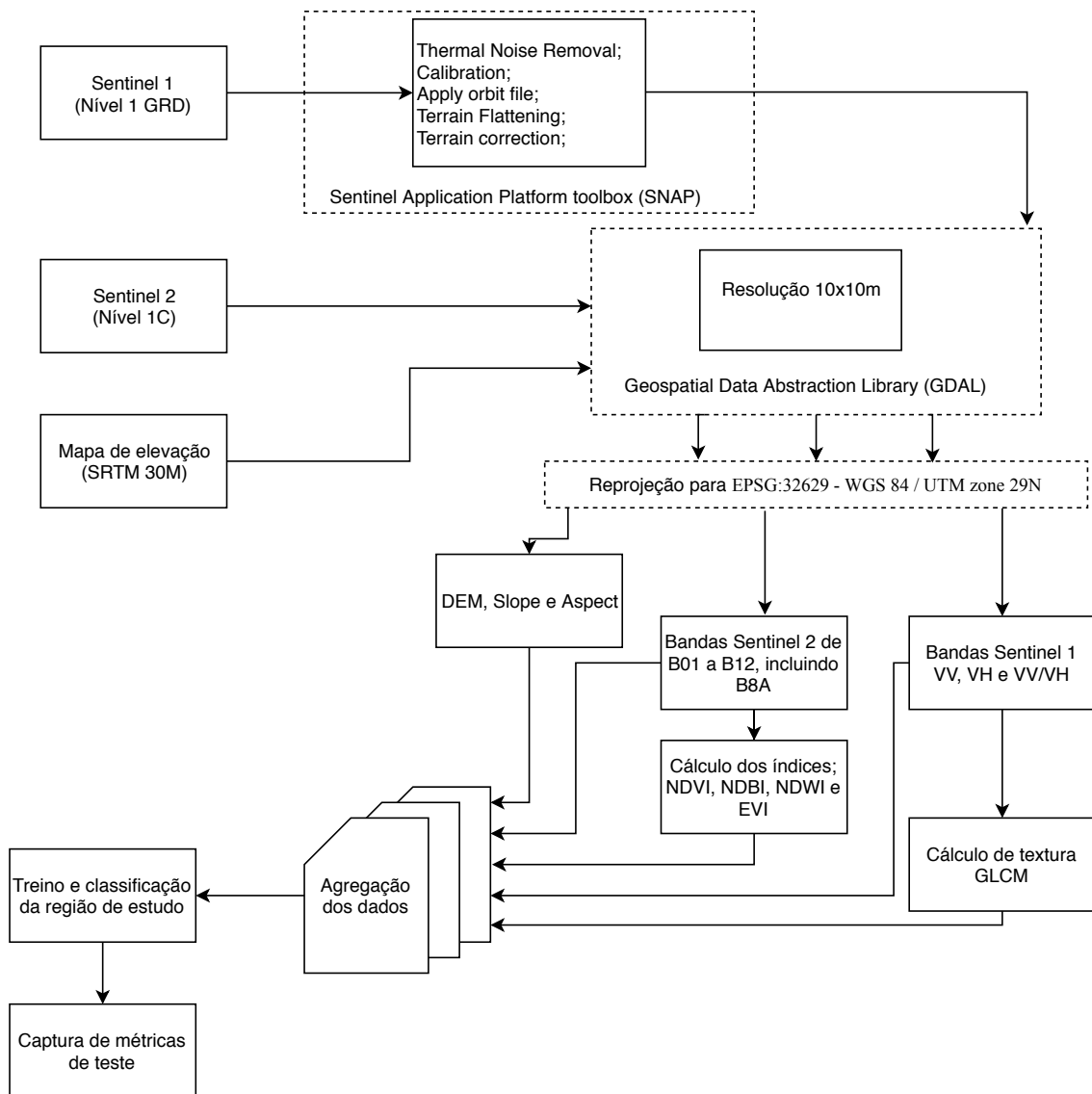


Figura 4.2: Cadeia de processamento utilizada nas experiências

força bruta, todas as combinações de parâmetros consideradas são testadas, este método tem uma complexidade computacional e temporal significativamente mais alta.

Devido ao número de treinos que têm de ser feitos para encontrar os melhores parâmetros, apenas são utilizados 5% dos pixels das imagens da região de testes. A pesquisa dos parâmetros utiliza também uma técnica nomeada *k-fold Cross Validation*, ilustrada na Figura 4.3, dividindo o conjunto de treino em  $k$  conjuntos (dobras). Após a divisão o modelo é treinado  $k$  vezes deixando uma parte de fora, treinando com  $k - 1$ , este processo é repetido iterando todas as dobras deixando as mesmas de parte. Esta técnica tem como objetivo estimar a performance do algoritmo para amostras não presentes nos dados de treino. No final os resultados dos treinos são agregados para a criação das métricas, por exemplo, utilizando a média.

A Figura 4.4, apresenta a cadeia de processamento implementada de modo a obter

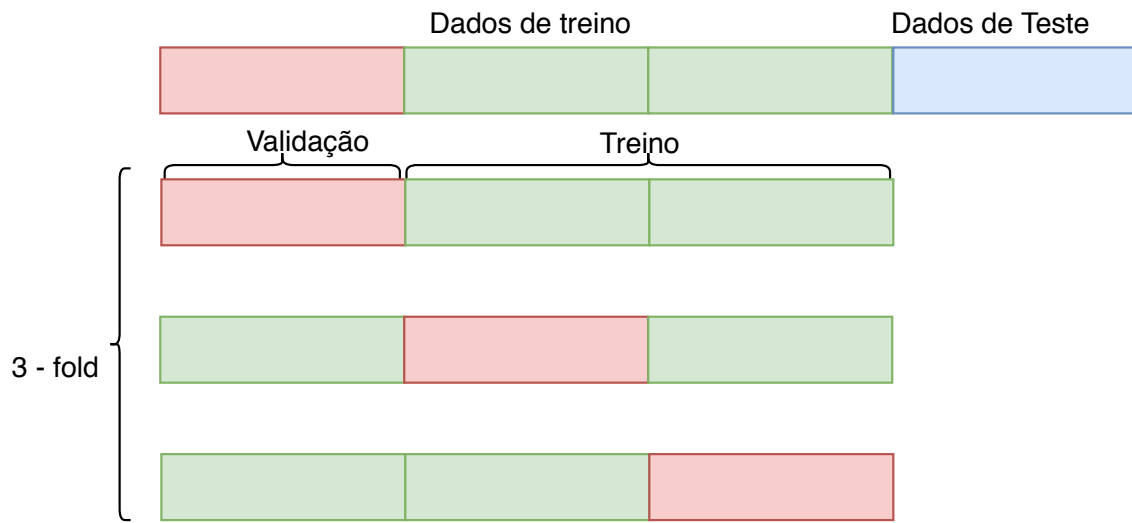


Figura 4.3: Estruturação dos dados de treino em 3-fold Cross Validation

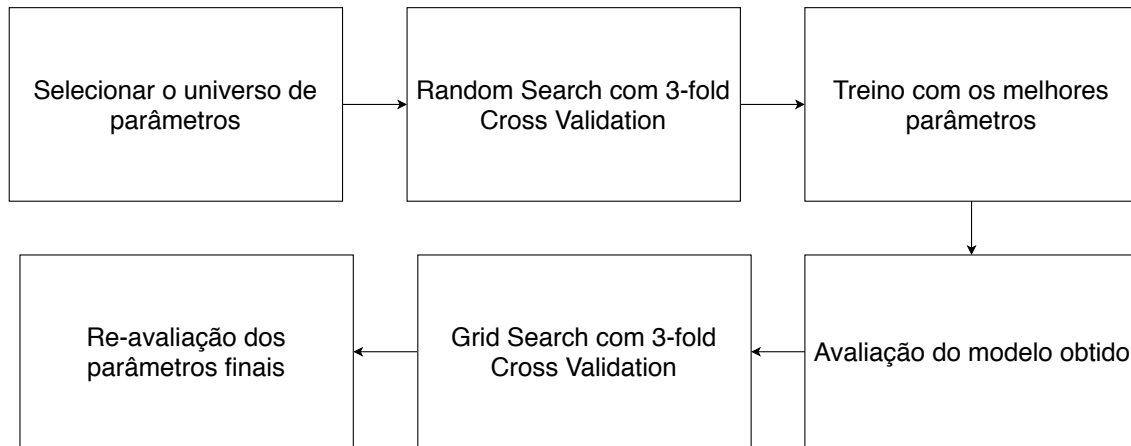


Figura 4.4: Cadeia de processamento para a obtenção dos melhores parâmetros de cada modelo

os parâmetros ótimos para o problema. Primeiro são selecionados quais os atributos a otimizar, e qual o intervalo de valores, depois desta seleção é feito uma pesquisa aleatória sobre permutações dos mesmos. Depois da obtenção dos melhores parâmetros nesta pesquisa, o modelo em questão é avaliado, é feita uma pesquisa força bruta num intervalo mais restrito, em volta do valor ótimo dos parâmetros, para procurar outros parâmetros potencialmente melhores. Finalmente os modelos são avaliados de novo, de modo a gerar as suas métricas de teste.

A metodologia de pesquisa apresentada foi implementada, de maneira semi-automática, para cada modelo de classificação único. É de notar que a metodologia adotada não garante os parâmetros ótimos, pois durante a pesquisa aleatória a melhor combinação pode não ser selecionada e assim, no final, são obtidos os parâmetros de um máximo local na pesquisa. Mesmo assim, o algoritmo garante a convergência do modelo para uma adaptação aos dados aceitável e perto do ótimo.

Os parâmetros escolhidos para otimização para cada classificador, apresentados por ordem de importância ditada pela sua documentação, são os seguintes:

- **Support Vector Machines:** C, gamma, class\_weight;
- **Stochastic Gradient Descent:** penalty, alpha, l1\_ratio, max\_iter;
- **Random Forest:** max\_depth, min\_samples\_leaf, min\_samples\_split, n\_estimators, class\_weight;
- **XGBoost:** learning\_rate, max\_depth, min\_child\_weight, gamma, colsample\_bytree, n\_estimators, max\_delta\_step;
- **Dense Neural Network:** n\_hidden\_layers, n\_neurons\_per\_layer.

Quando ao intervalo de valores dos parâmetros, para parâmetros contínuos, com learning\_rate do *XGBoost*, foi dado ao algoritmo de otimização uma distribuição uniforme entre o valor máximo e mínimo desse mesmo parâmetro. Para parâmetros descontínuos, como o n\_estimators das *Random Forest*, é dado uma lista de valores, espaçados de maneira a cobrir a maioria dos valores possíveis.

#### 4.4.1.2 Redução dimensional dos dados

Uma das particularidades dos dados utilizados é o seu grande volume, isto tem um peso na complexidade computacional dos algoritmos bastante alto. Existem algumas técnicas para reduzir o número de atributos dos dados, as quatro abordadas são *feature importance*, *elastic net*, *boruta* e **LASSO**. Devido à quantidade de testes necessários de efetuar, para testar todas as técnicas, estes foram automatizados, resultando num módulo Python para redução de atributos automática utilizando as quatro metodologias apresentadas na secção 2.3. A sua estrutura de processamento geral é apresentada na Figura 4.5.

O módulo funciona em três modos, “importances”, “elastic” e “boruta”, sendo o algoritmo “Lasso” um caso especial do modo “elastic”. O modo “elastic” é um pouco diferente dos outros modos, este aplica *3-fold cross validation*, para achar os melhores parâmetros e o termo de regularização L1 (Lasso) e L2 (Ridge), no final devolve o melhor transformador.

Com o modo “importances”, o módulo recebe um classificador, com o requerimento de ser um modelo de *ensemble*, e recebe um limite de importâncias no intervalo de 0 a 1. Este soma as importâncias de cada atributo, por ordem decrescente, até a soma total ser igual ou superior ao limite dado, descartando os restantes atributos.

O modo “boruta”, como o nome indica, aplica diretamente o algoritmo de redução de atributos Boruta, descrito na secção 2.3.

Os transformadores de dados resultantes podem ser aplicados a qualquer estrutura de dados. O único requisito é ter os mesmos atributos que os dados originais e pela mesma ordem. Exemplos da criação dos diferentes transformadores apresentado na listagem 4.1.

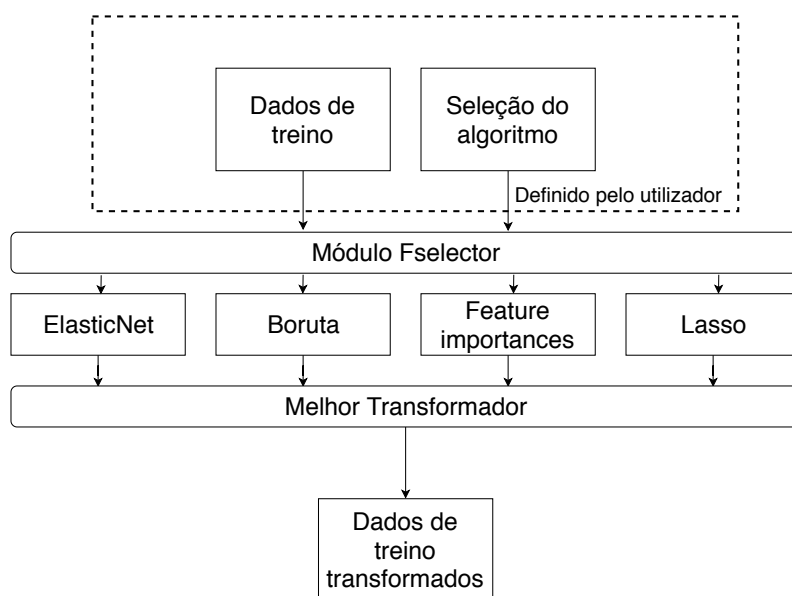


Figura 4.5: Cadeia de processamento para redução de dimensionalidade dos atributos, implementada no módulo Fselector

Listagem 4.1: Exemplos de utilização do módulo Fselector

```

1  x_train_feature, y_train_feature = carregar dados de treino;
2
3  # redutor de atributos utilizando "feature importance" com limite de 0.95
4  selector = fselector.Fselector(clf, mode="importances", thold=0.95)
5  transformer = selector.select(x_train_feature, y_train_feature)
6
7  # redutor de atributos utilizando "elastic net" com limite de 0.25
8  selector = fselector.Fselector(clf, mode="elastic", thold=0.25)
9  transformer = selector.select(x_train_feature, y_train_feature)
10
11 # redutor de atributos utilizando "boruta" com limite de 0.95
12 selector = fselector.Fselector(clf, mode="boruta", thold=0.95)
13 transformer = selector.select(x_train_feature, y_train_feature)
14
15 x_train_feature = transformer.transform(x_train_feature)
  
```

Este módulo não foi testado para produção e apenas garante resultados robustos para os dados utilizados neste estudo. Será preciso trabalho futuro para transformar o módulo apresentado num módulo oficial robusto.

## 4.5 Conclusão

Este capítulo aborda pontos bastante importantes neste estudo. É nele que é definida a cadeia de processamento da solução proposta, sendo os seus dois componentes principais, antes do treino dos algoritmos, a otimização e a redução de atributos dos dados de

treino. A otimização foi implementada de maneira a diminuir a complexidade de uma procura de força bruta pelos parâmetros, o que resulta numa abordagem mais leve que, ao mesmo tempo, produz modelos robustos. A abordagem implementada para a otimização não garante os parâmetros ótimos, para isso teriam de ser testadas todas as combinações possíveis, que poderiam ser potencialmente infinitas, mesmo assim apresenta uma otimização robusta com resultados aceitáveis.

Devido à complexidade temporal que este estudo envolve, todos os processos foram automatizados até um certo ponto, isto tem um impacto na discussão de resultados final, visto não existe acesso aos resultados intermédios. Um exemplo são as métricas de validação durante a otimização, não será possível apresentar a evolução dos algoritmos com a alteração dos seus parâmetros, mesmo assim é garantido que os parâmetros obtidos são robustos e confiáveis.

Dado a informação aqui apresentada, o próximo capítulo apresenta a recolha dos dados, de treino e de referência.





## ESTRUTURAÇÃO DOS DADOS

Neste capítulo são apresentados todos os dados utilizados durante a investigação. Os dados de treino, são separados em dois grupos principais, dados estáticos e séries temporais e, ambos os grupos podem ter pequenas variações, como a remoção dos índices ou a adição de um atributo extra. Quanto aos dados de referência, que representam as classes, são apresentados os três diferentes grupos de classes e uma breve análise dos mesmos. O capítulo esta dividido nas seguintes três secções:

- **Extração dos dados de referência**, secção 5.1: São aqui apresentados os dados de referência que servem de base para o treino dos algoritmos, mais especificamente é apresentada a fonte de onde são extraídos e os vários mapeamentos para a criação dos grupos de teste.
- **Obtenção e processamento de dados de treino**, secção 5.2: Esta secção apresenta em detalhe os dados de treino, são representadas as duas estruturas principais, dados estáticos e as séries temporais.
- **Conclusão**, secção 5.3: O capítulo é concluído com uma pequena síntese.

### 5.1 Extração dos dados de referência

A carta mais importante de toda a dissertação é a carta [COS 2015](#), é dela que são extraídos os dados de referencia principais, complementados posteriormente com dados sobre estradas do projeto [OSM](#). As estruturas artificiais permanentes principais são extraídas da carta [COS](#) de 2015, para Portugal Continental, até à data esta é a carta mais atualizada e com melhor resolução espacial sobre a região. Os dados de referência foram extraídos a partir do Ficheiro vetorial Modelos de Combustível, providenciado pelo [Instituto da Conservação da Natureza e das Florestas \(ICNF\)](#), este modelo é baseado na carta [COS](#) de

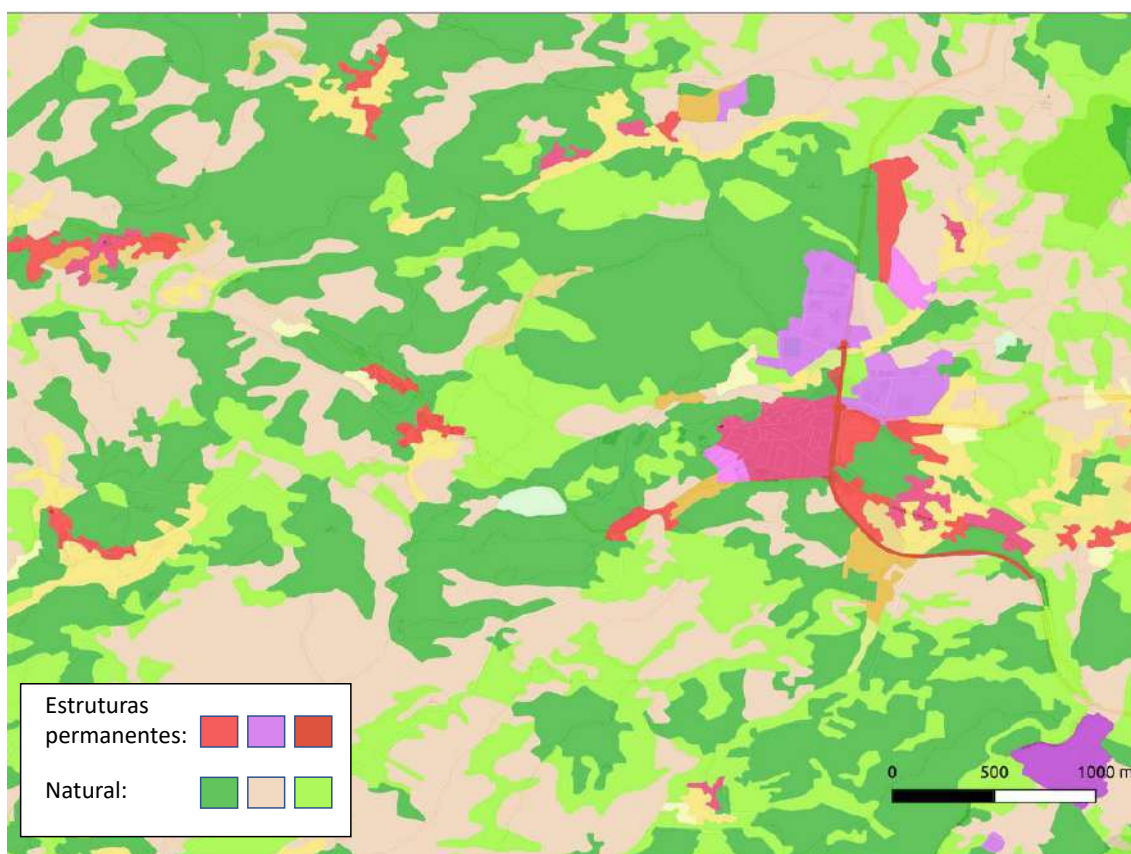


Figura 5.1: Segmento da carta COS 2015 sobre a região de Vila de Rei

2015 possuindo algumas classes extra para vegetação.<sup>1</sup> Um exemplo de uma pequena região, dentro da área de estudo, da carta COS 2015 é apresentado na Figura 5.1.

Após a obtenção dos dados de treino, da carta COS 2015, em formato vetorial, foi necessário extrair estruturas artificiais permanentes. Para isso foi adicionada uma coluna extra ao ficheiro vetorial, esta coluna mapeia as classes originais da carta para um identificador único por classe, simplificando o processamento posterior. O mapeamento final, de classes COS para as classes de interesse que representam estruturas permanentes é feito em tempo de execução, isto adiciona flexibilidade para testar vários tipos de diferenciação de classes sem ter de alterar a carta de entrada.

As estradas são retiradas do projeto OSM. Para a adição das estradas, visto que a informação proveniente do OSM é em formato de vetorial, foi necessário a criação de um *buffer* dinâmico envolvendo as mesmas, sendo a largura deste *buffer*, de entre 10 a 5 metros dependendo do tipo de estrada, pois tipos de estrada diferentes possuem diferentes largura. Para encontrar o melhor tamanho do *buffer* para cada tipo de estrada, tendo em conta também o tamanho dos pixels de 10 por 10 metros, foi escolhida uma metodologia manual. Para cada tipo de estradas foram testados vários valores para a largura do *buffer*. Após a geração do mesmo é então verificado para ver se engloba bem as

<sup>1</sup>Fonte da carta COS 2015 em vetor: <http://www2.icnf.pt/portal/florestas/dfci/cartografia-dfci> - Visitado a 01/08/2019

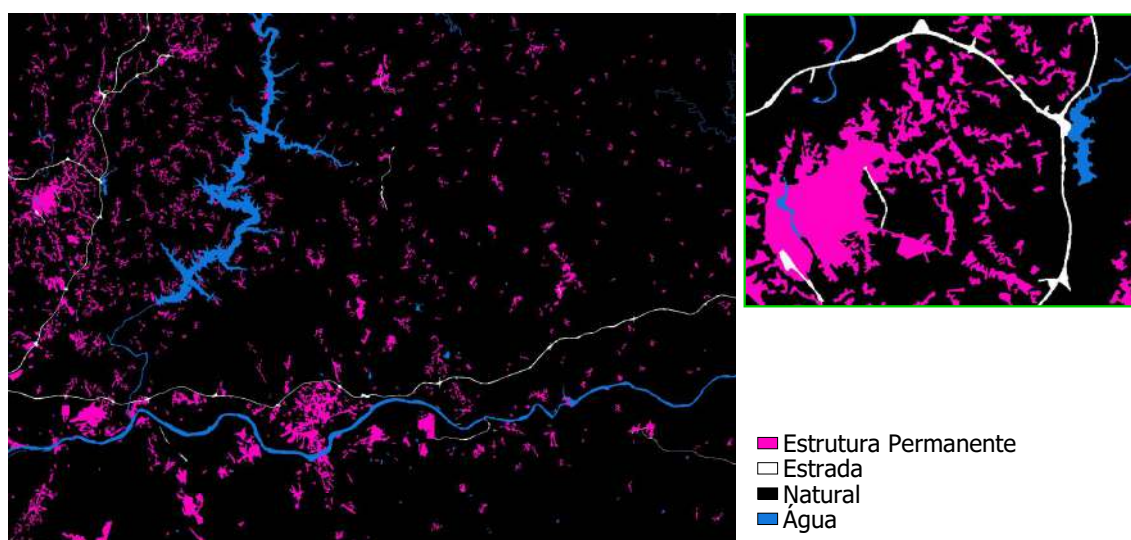


Figura 5.2: Classes extraídas da carta COS 2015 sobre a região de testes, grupo 1

estradas presentes nas imagens Sentinel-2 recolhidas.

As classes são divididas em vários grupos. Um dos objetivos da dissertação é avaliar o potencial dos classificadores para distinguirem vários tipos de estruturas, como por exemplo, diferenciar zonas rurais de zonas urbanas para estudar o potencial da extração de estradas. Como existem vários tipos de estruturas permanentes, é de interesse observar o comportamento dos classificadores a diferentes definições das classes. Relembrando a Figura 4.1 da página 38, existem três grupos de classificação, o grupo 1, 2 e 3.

O grupo 1, que diferencia estruturas artificiais permanentes do restante é o grupo principal e foi o grupo utilizado para os testes preliminares, o mesmo está representado na Figura 5.2. Neste grupo não constam as estradas do OSM, como resultado apenas são observáveis na figura as autoestradas e algumas estradas principais. As estradas representadas na Figura 5.2 (a branco) servem apenas de ilustração. Durante a classificação estas são consideradas estruturas permanentes. Como se pode observar na Figura 5.3, existem muitas mais amostras da classe “Natural” do que as restantes, devido à região de testes ser uma região predominantemente rural composta por dois aglomerados urbanos principais e vários aglomerados de estruturas rurais, como pode ser observado na Figura 5.2.

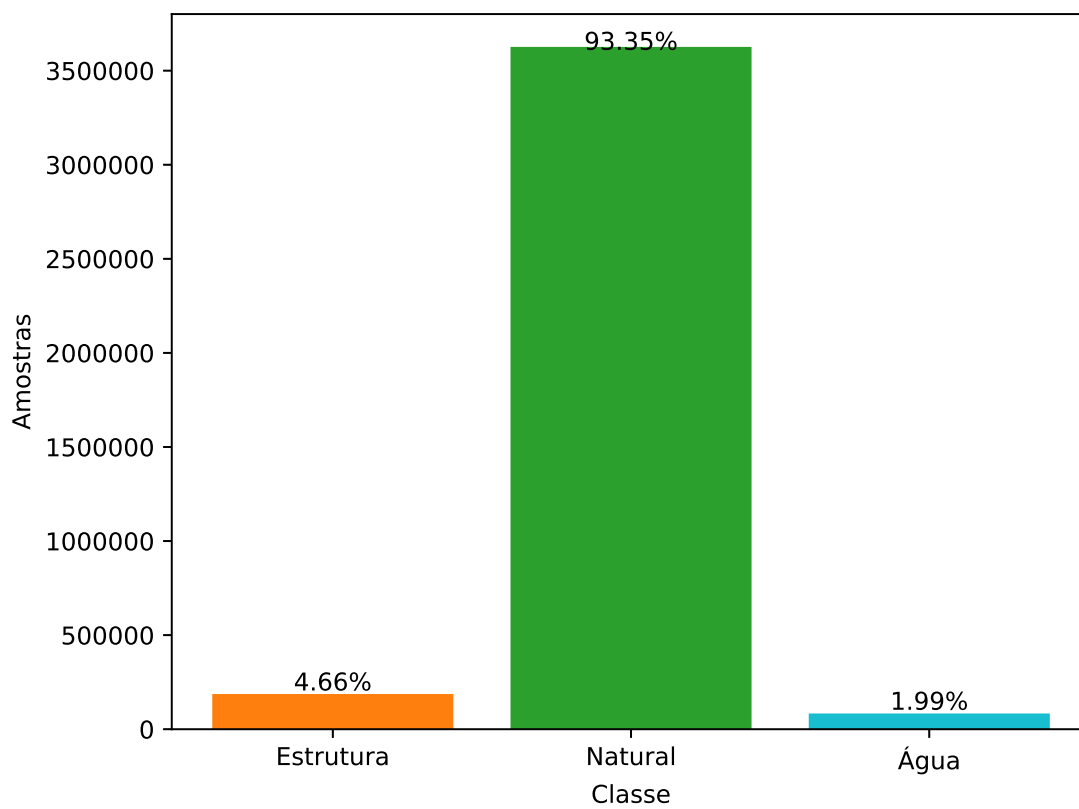


Figura 5.3: Distribuição das classes do grupo 1

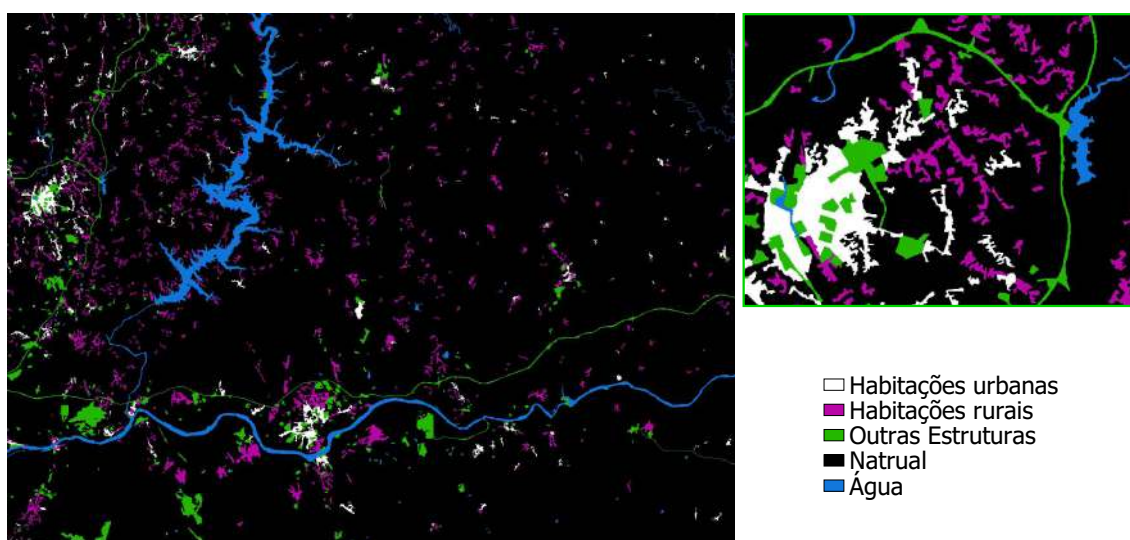


Figura 5.4: Classes extraídas da carta COS 2015 sobre a região de testes, grupo 2

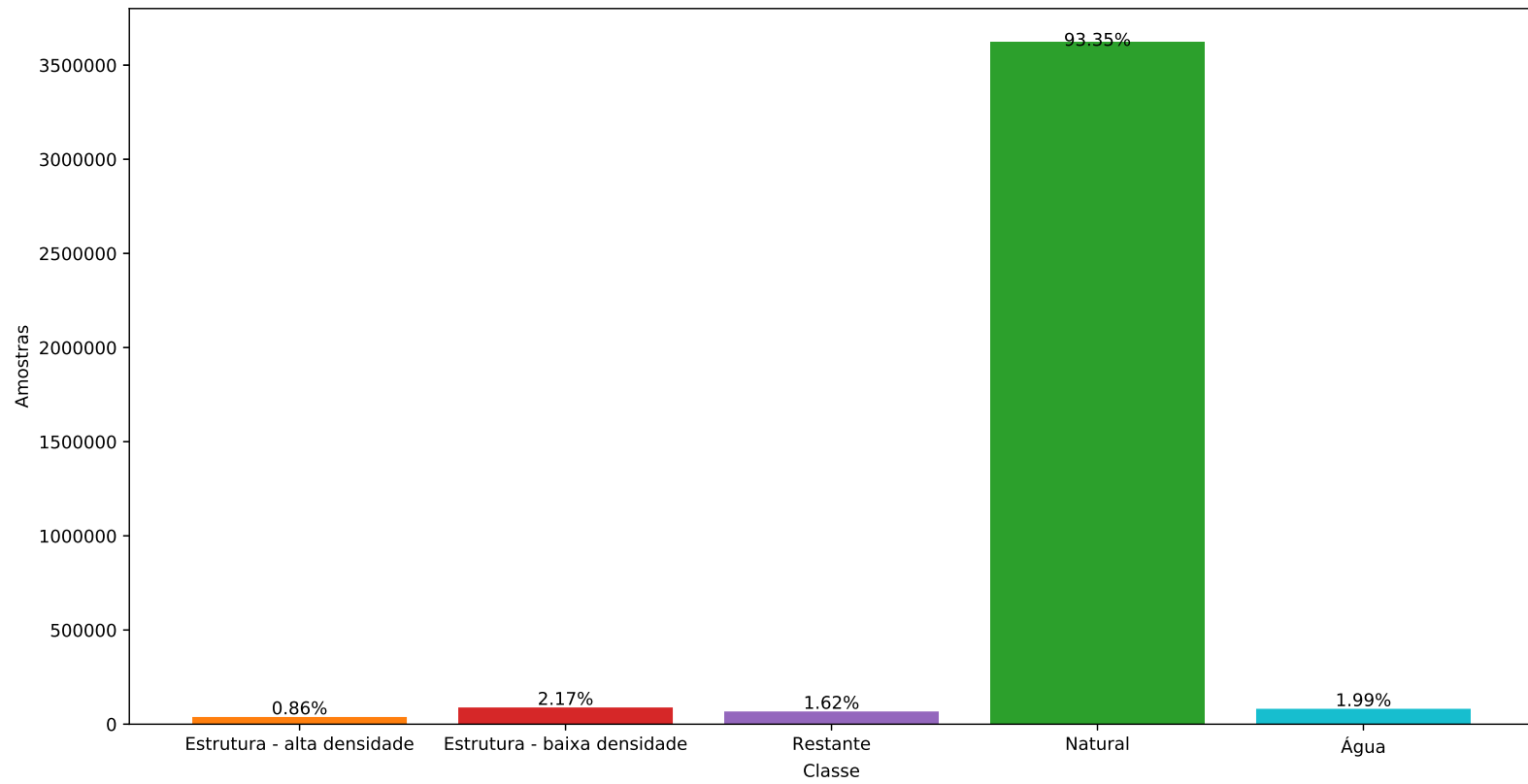


Figura 5.5: Distribuição das classes do grupo 2

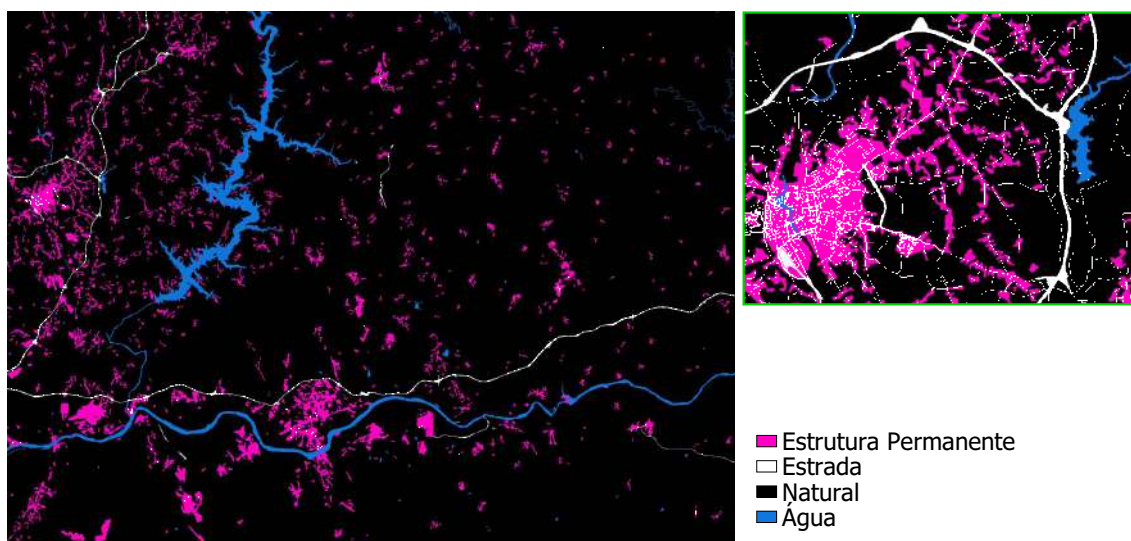


Figura 5.6: Dados com estradas extraídas do OSM e da carta COS 2015 sobre a região de testes, grupo 3

Representado na Figura 5.4, o grupo 2 refina as estruturas permanentes dando uma dicotomia de habitações urbanas (alta densidade de estruturas) ou rurais (baixa densidade de estruturas). Este grupo é um dos mais importantes pois as análises que utilizam cartográfica de estruturas muitas vezes carecem desta diferenciação [73]. As restantes estruturas são agrupadas numa classe adicional, contendo, por exemplo, zonas industriais, autoestradas, minas e as restantes estruturas artificiais extraídas da carta COS 2015. A distribuição deste grupo tem uma distribuição semelhante à do grupo 1, a classe “Natural” é a classe com mais amostras, pode ser observar também que existe um número maior de estruturas pertencentes a zonas rurais do que zonas urbanas de alta densidade. Ao separar também as restantes estruturas, as estruturas de alta densidade ficam agora a classe minoritária. Isto poderá afetar a distinção das mesmas e baixa a performance dos algoritmos.

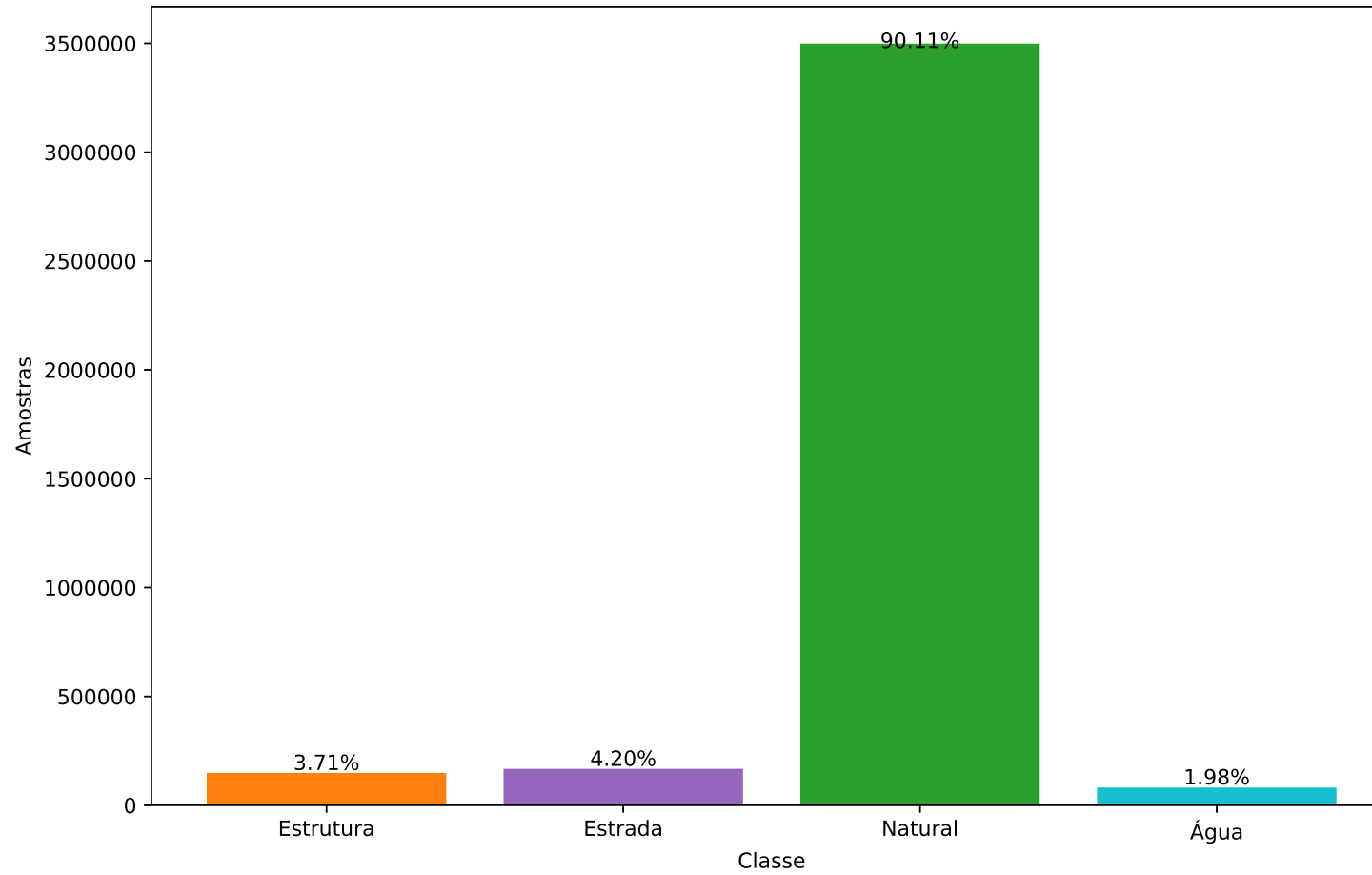


Figura 5.7: Distribuição das classes do grupo 3

Finalmente, o grupo 3 apresenta uma refinação das estradas utilizando dados provenientes do **OSM**. Estes dados são sobrepostos à carta **COS**, construindo assim uma carta que melhor representa o mundo real. O mesmo é apresentado na Figura 5.6. É de notar que existem estradas com uma largura inferior à resolução de 10 por 10 metros, o que faz com que este grupo de classes esteja sujeito a maiores erros de classificação. A Figura 5.7 mostra então a distribuição do grupo final. A adição das estradas **OSM** cria uma classe com número um pouco superior às estruturas. Ao observar a Figura 5.6 em detalhe, podemos concluir que todas as estruturas e aglomerados são conectados por estradas, e dentro dos mesmos são também intersectados pelas mesmas. Isto adiciona dificuldade extra a esta tarefa de classificação, o que faz com que a separabilidade das classes seja muito mais difícil do que as restantes.

## 5.2 Obtenção e processamento de dados de treino

Para um melhor entendimento da metodologia utilizada são então apresentados os dados de treino e como os mesmos foram recolhidos. Os produtos recolhidos são provenientes das constelações Sentinel-1 e Sentinel-2 e do modelo de elevação de terreno da missão **SRTM**. Para as fontes de dados multi-temporais são deixadas de parte alguns produtos derivados devido à complexidade computacional necessária para os incluir.

### 5.2.1 Dados estáticos

O primeiro grupo de dados de treino, dados estáticos, é o grupo mais simples e mais rápido de processar, pois são dados uni-temporais de 2015, como consequência estes dados possuem menos informação acerca do solo da Terra, quando comparados com uma série temporal. Mesmo assim é interessante ver a diferença de performance versus séries temporais, pois a utilização de dados estáticos é facilmente escalável.

Os dados das constelações Sentinel-1 e Sentinel-2, deteção remota ativa e passiva respetivamente, foram extraídos do *Copernicus Open Access Hub*<sup>2</sup>. Esta plataforma fornece dados livres completos de Sentinel-1, Sentinel-2, Sentinel-3 e Sentinel-5P. Para limitar erros derivados de mudanças no terreno, devido à diferença de tempo entre os dados e as classificações do mapa **COS** de 2015, foram extraídos dados de 2015 da plataforma, Sentinel-1 do dia 30/07/2015 e Sentinel-2 do dia 15/07/2015.

Para os produtos Sentinel-2, apenas foram recolhidos produtos com cobertura de nuvens de 0% de modo a anular erros por contaminação da imagem. Os produtos Sentinel-1, devido a não serem afetados por este efeito, o único critério utilizado na escolha foi a proximidade da datas dos mesmos aos produtos Sentinel-2. Os produtos obtidos bem como os seus identificadores são apresentados na Tabela 5.1.

---

<sup>2</sup>Visitado a 15/03/2019 - <https://scihub.copernicus.eu/>



Tabela 5.1: Identificadores dos produtos utilizados na análise estática

Produto
S2A_MSIL1C_20150715T112846_N0204_R037_T29SNC_20150715T112845
S1A_IW_GRDH_1SDV_20150730T182707_20150730T182732_007044_00993A_C15B

### 5.2.2 Séries temporais

O processo de obtenção das imagens da série temporal é distinto do estático, pois a plataforma utilizada apenas possui dados Sentinel-1, disponíveis de imediato, desde de 01-Agosto-2018, para dados anteriores é necessário fazer o pedido dos dados e esperar até que eles estejam disponíveis, o que pode demorar um dia completo. Isto que faz com que não seja possível recolher dados de grandes regiões ou séries temporais, que contém vários produtos, em tempo aceitável.

Para a recolha das imagens Sentinel-1, foi utilizada a plataforma de dados do Alaska, Vertex <sup>3</sup>. Esta plataforma persiste todos os dados provenientes do Sentinel-1, o que faz com que seja possível a recolha rápida dos mesmos. Após a seleção manual dos produtos pretendidos, a plataforma gera um *script* Python para a sua recolha automática. Visto que dados de deteção remota passiva não são afetados pelas condições atmosféricas, foi selecionada a trajetória da constelação com melhor cobertura sobre a área de estudo e recolhido, pelo menos, um produto por cada mês.

Os produtos provenientes do Sentinel-2 foram escolhidos e recolhidos de maneira automática. Foi implementado um pequeno programa que, dado duas datas, percorre todos os produtos no *Copernicus Open Access Hub* que intersectem a área de estudo, neste caso definida com um *shapefile*. Para cada produto é testado se o mesmo tem uma cobertura de nuvens inferior a 30%, caso passe os testes o mesmo é aceite e é feita a sua recolha. Apesar da cobertura de nuvens de um produto ser de 30%, isto não significa que a região de testes possua alguma nuvem, visto que um produto tem uma área bastante maior. Após a recolha cada produto, os mesmos são avaliados manualmente, dentro da região de testes, e caso exista alguma nuvem num dado produto esse mesmo produto é descartado.

A Figura 5.8, apresenta a distribuição de todos os produtos recolhidos ao longo de 2016. O critério para a escolha deste ano foi a proximidade aos dados de referencia, carta COS e o facto da constelação Sentinel-2 possuir uma cobertura anual completa a partir de 2016 apenas.

Pode ser observado a ausência de dados da constelação Sentinel-2 no início do ano, e a concentração dos mesmos entre junho e setembro. Este fenómeno é explicado simplesmente pela cobertura de nuvens ao longo do ano no clima de Portugal sobre a região. A plataforma *Weather Spark* <sup>4</sup> possui dados meteorológicos de Portugal, em especial os dados sobre região de estudo que são gerados com base numa análise estatística de relatórios horários históricos e modelos de 1 de janeiro de 1980 a 31 de dezembro de 2016.

<sup>3</sup><https://vertex.daac.asf.alaska.edu/-visitadoa03-08-2019>

<sup>4</sup><https://pt.weatherspark.com> - visitado a 07/08/2019

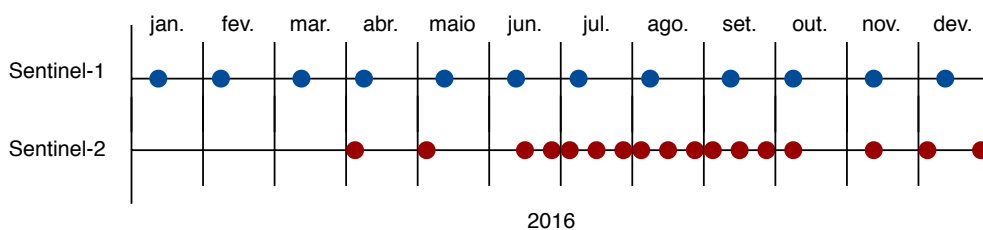


Figura 5.8: Distribuição temporal das imagens recolhidas

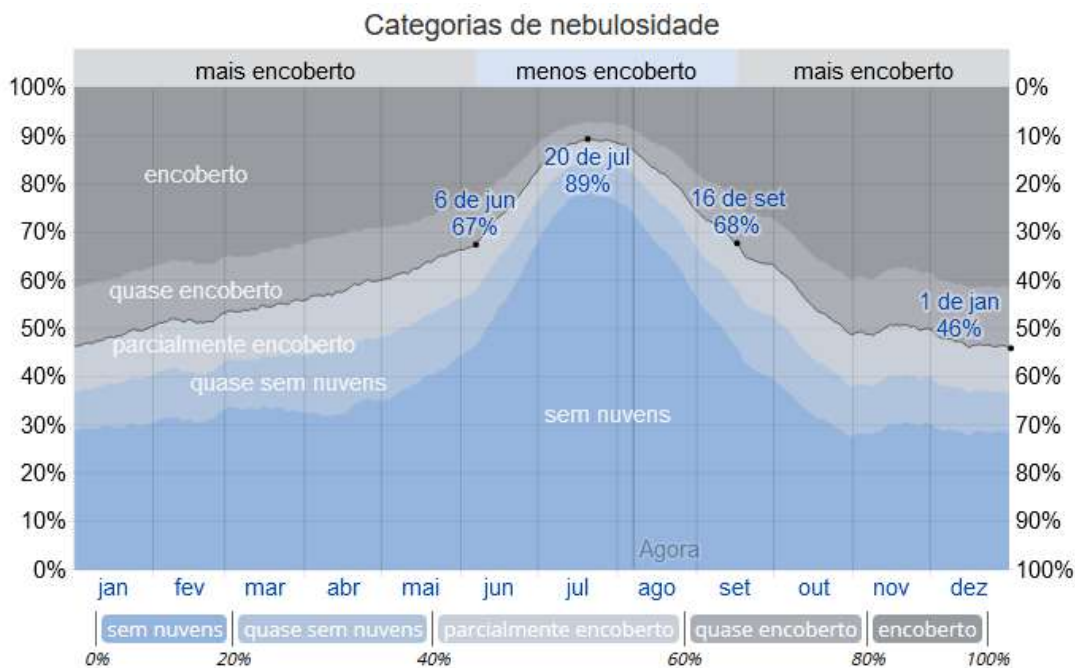


Figura 5.9: Percentagem de tempo passada em cada faixa de nebulosidade, categorizada pela percentagem de céu encoberto por nuvens. Retirado de <https://pt.weatherspark.com>

Segundo os dados desta plataforma, a distribuição média de nuvens ao longo do ano, apresentada na Figura 5.9, vai de encontro com a falta de dados sem nuvens, A época menos encoberta do ano em Mação, região que está contida na área de estudo, dura pelo menos 3 meses em média, começando por volta de 6 de junho e terminando em 16 de setembro.

Para motivos de validação e reprodutibilidade dos resultados os identificadores dos produtos Sentinel-1 e Sentinel-2 recolhidos estão presentes no anexo D.

### 5.2.2.1 Estruturação das séries temporais

Os algoritmos de aprendizagem utilizados são sensíveis quanto ao seu *input*, cada atributo utilizado têm de pertencer ao mesmo domínio e o número de atributos, após o treino, não pode ser alterado. Para a utilização de séries temporais com estes algoritmos foram consideradas duas opções: utilizar cada imagem da série temporal como um único atributo ou

derivar estatísticas temporais, como a média, o desvio padrão e alguns percentis.

A primeira abordagem tem alguns problemas, primeiramente a geração de um número elevado de atributos, que potencialmente não cabem em memória. Seguidamente, a corrupção de apenas uma imagem impede o algoritmo de obter resultados. Isto significa que, se for necessário a utilização do mesmo modelo para previsão de dados de outras datas é necessário a recolha de imagens o mais próximo possível dos mesmos dias, para evitar erros. Caso exista um erro, um produto Sentinel-2 não foi gerado, ou exista demasiada cobertura de nuvens num certo dia, o modelo gerado pode ser inutilizável.

Para resolver estes problemas, foram então extraídas estatísticas das séries temporais. Assim cada atributo ganha outra simbologia, por exemplo o primeiro atributo pode significar o mínimo de uma certa banda ao longo da série temporal, o que é algo calculável mesmo com uma potencial falha de algum produto. Para cada tipo de dados, bandas Sentinel-2 e índices radiométricos, bandas Sentinel-1 e textura. As estatísticas escolhidas foram a média, a variância, percentil 0% (mínimo), percentil 25%, percentil 50% (mediana), percentil 75% e percentil 100% (máximo).

### 5.2.3 Modelo de elevação de terreno

O modelo de elevação de terreno, que complementa os dados espectrais e de textura, foi obtido a partir dos servidores da NASA, da missão [SRTM](#)<sup>5</sup>. Como este tipo de dados não é multi-temporal foram selecionados manualmente todos os segmentos que cobrem Portugal Continental. Este modelo vem subdividido em várias secções em regiões de de 1 arco-segundo, o que resulta uma resolução de 3601x3601 pixeis por segmento. Estes segmentos são disponibilizados com uma projeção EPSG:4326. Cada segmento é processado a uma resolução de 10 por 10 metros e reprojetoado para EPSG:32629. Após a obtenção, as variantes declive e orientação foram calculadas com a ferramenta [Geospatial Data Abstraction Library \(GDAL\)](#), os resultados finais do processamento estão representados na [Figura 5.10](#). Sendo o declive o grau de inclinação do terreno e a orientação a orientação da inclinação. Como os modelos de elevação de terreno não possuem muita variância ao longo do tempo, este modelo é utilizado em ambas as abordagens estáticas e em conjunto com as séries temporais.

## 5.3 Conclusão

Este capítulo apresentou como foram recolhidos e tratados os dados de treino utilizados e apresentou uma breve análise dos mesmos. Durante a apresentação dos resultados obtidos, será apresentada mais detalhadamente a estrutura dos dados de treino utilizados, como por exemplo quais as bandas e informação derivada selecionadas para um dado teste. São dados também os identificadores dos produtos utilizados de modo a promover a replicabilidade das experiências. Uma breve comparação dos dados de séries temporais

<sup>5</sup>DEM da missão SRTM - <https://dwtkns.com/srtm30m/> - <https://gcmd.nasa.gov/>

provenientes da constelação Sentinel-1 e 2, mostra que dados Sentinel-2 não cobrem o ano de 2016 todo, devido à contaminação de nuvens. Alguns produtos tiveram de ser descartados pois caso contrário seriam utilizados dados que não representariam com fiabilidade a ocupação de solo real.

Acompanhando os dados de treino, são apresentados os dados de referência, ou seja as classes, que foram divididas em três grupos. Relembrando, o grupo 1 possui três classes: estruturas artificiais permanentes, natural e águas. O grupo 2 possui cinco classes: habitações urbanas, habitações rurais, outras estruturas, natural e água. Finalmente, o grupo 3 possui quatro classes: estruturas artificiais permanentes, estradas, natural e águas. Como foi visto, todos os grupos apresentam um desequilíbrio bastante grande das classes, sendo a classe “Natural” a classe mais representada em todos os casos.

Com a obtenção dos dados é possível passar à fase final, a fase de testes e de análise, apresentada no próximo capítulo.

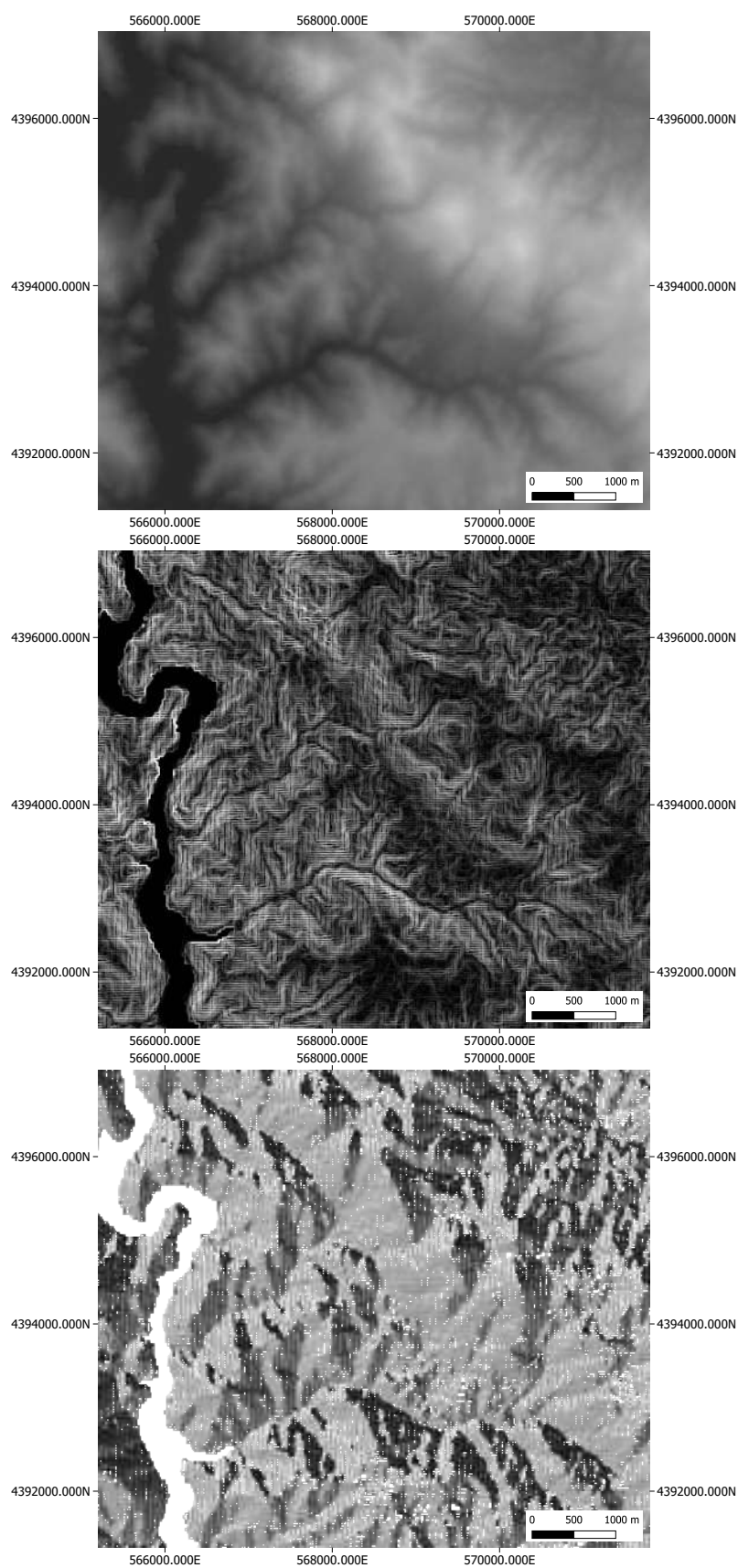


Figura 5.10: Segmento do MDE de uma região após processamento; Em primeiro a elevação, segundo declive e finalmente a orientação



## VALIDAÇÃO EXPERIMENTAL E OTIMIZAÇÃO DOS MODELOS

Este capítulo é focado primeiramente numa análise exploratória dos dados tendo em conta alguns problemas previstos, como a dificuldade da deteção de estruturas mais pequenas. Posteriormente a esta análise, são apresentados os resultados da fase de otimização, fase esta precursora dos resultados finais. O capítulo 6 está dividido da seguinte forma:

- **Validação e análise experimental**, secção 6: A primeira análise realizada, apresentada nesta secção, teve como âmbito explorar o potencial da metodologia apresentada. Começando com uma análise da separabilidade das amostras e uma classificação preliminar, que visa a confirmar fenómenos estudados no estado da arte.
- **Resultados da otimização dos algoritmos**, secção 6.2: Aqui são apresentados os resultados da primeira fase da cadeia de processamento da solução proposta, são mostrados os parâmetros ótimos por algoritmo que serão utilizados para todas as experiências seguintes.

No início da investigação, foram realizadas algumas experiências exploratórias tendo em conta o estado de arte e os problemas derivados. As experiências preliminares têm um papel de validação experimental e visam avaliar o potencial de métodos de aprendizagem automática em resposta aos problemas encontrados. Estas têm um papel fundamental na dissertação pois abrem caminho às experiências finais mais robustas e será possível prever um pouco o comportamento das mesmas, de modo a não existirem artefactos inesperados. As experiências foram maioritariamente feitas com apoio do Google Earth Engine, o programa QGIS e alguns *scripts* Python.

### 6.0.1 Análise baseada em pixels

Esta experiência consiste na análise temporal do comportamento de vários tipos de pixel e vem responder às seguintes perguntas:

- Qual é o impacto do conteúdo de um pixel, por exemplo casas, vegetação ou outras estruturas, no comportamento da sua reflectância?
- Qual é a tarefa T mais difícil e se a mesma é factível pela metodologia apresentada?

Para responder à primeira pergunta proposta foram analisados os valores do **NDBI** médio de 3 tipos de pixel, um pixel dentro de uma vila, pixels que contêm apenas vegetação e por fim, pixels mistos com vegetação e edifícios pequenos. Estes tipos de pixel são os principais que apresentarão comportamentos diferentes na classificação das estruturas dentro dos mesmos. Os resultados da análise temporal são apresentados na figura 6.1. A primeira observação a constatar é a falta de dados entre Dezembro de 2015 e Abril 2016, este comportamento não foi previsto e é importante realçar. Este foi causado pela cobertura de nuvens nas imagens obtidas, para efeitos de teste foram seleccionadas imagens com baixa percentagem de nuvens presentes.

O comportamento das bandas é trivial devido à dinâmica da vegetação, pode ser observado que um pixel numa cidade tem um comportamento mais constante do que um pixel com vegetação, este tipo de pixel tem comportamento espectral totalmente diferente dos outros, enquanto os restantes são mais difíceis de distinguir, isto pode apresentar um problema na deteção de edifícios isolados. A Figura 6.2 mostra a mesma análise mas desta vez com os valores do **NDVI** dos mesmos tipos de pixels, são tiradas as mesmas conclusões.

A deteção de estruturas pequenas, como casas ou pequenos aglomerados, com áreas menores que um pixel é algo que pode não ser possível utilizando a metodologia apresentada. Em geral, zonas com pixels mais difusos, ou seja contendo uma mistura de vegetação e outras classes de classificação, apresentam comportamento espectral semelhante a zonas que apenas contenham vegetação, é aqui que o erro final tem o potencial de ser mais elevado. Dito isto, os limites de aglomerados populacionais, onde se transita de estruturas artificiais para vegetação ou outras estruturas naturais, serão também alvo do maior erro de classificação devido à baixa separabilidade espectral das classes.

As estradas são um subtipo de estrutura artificial permanente. É de interesse avaliar o potencial da deteção das mesmas. Vários tipos de análise necessitam deste tipo de informação, como por exemplo a delimitação das **FGC**, em que a largura das faixas depende do tipo de estrutura. A Figura 6.3 mostra a variação do **NDVI** de três regiões contendo estradas diferentes, enquanto estradas largas, como autoestradas, possuem um valor do **NDVI** baixo estradas mais estreitas tem um comportamento espectral semelhante a vegetação, isto acontece pois estas estradas, em geral, são rodeadas por árvores e outros tipos de vegetação. Entre Outubro e Dezembro de 2016 o valor do **NDVI** da estrada analisada



---

chega a ser maior do que uma zona contendo apenas vegetação. As estradas são potencialmente as estruturas artificiais mais difíceis de detetar, em geral são também ausentes dos estudos focados em deteção de estruturas, mesmo assim será analisado a performance dos modelos apresentados para a deteção de estradas.

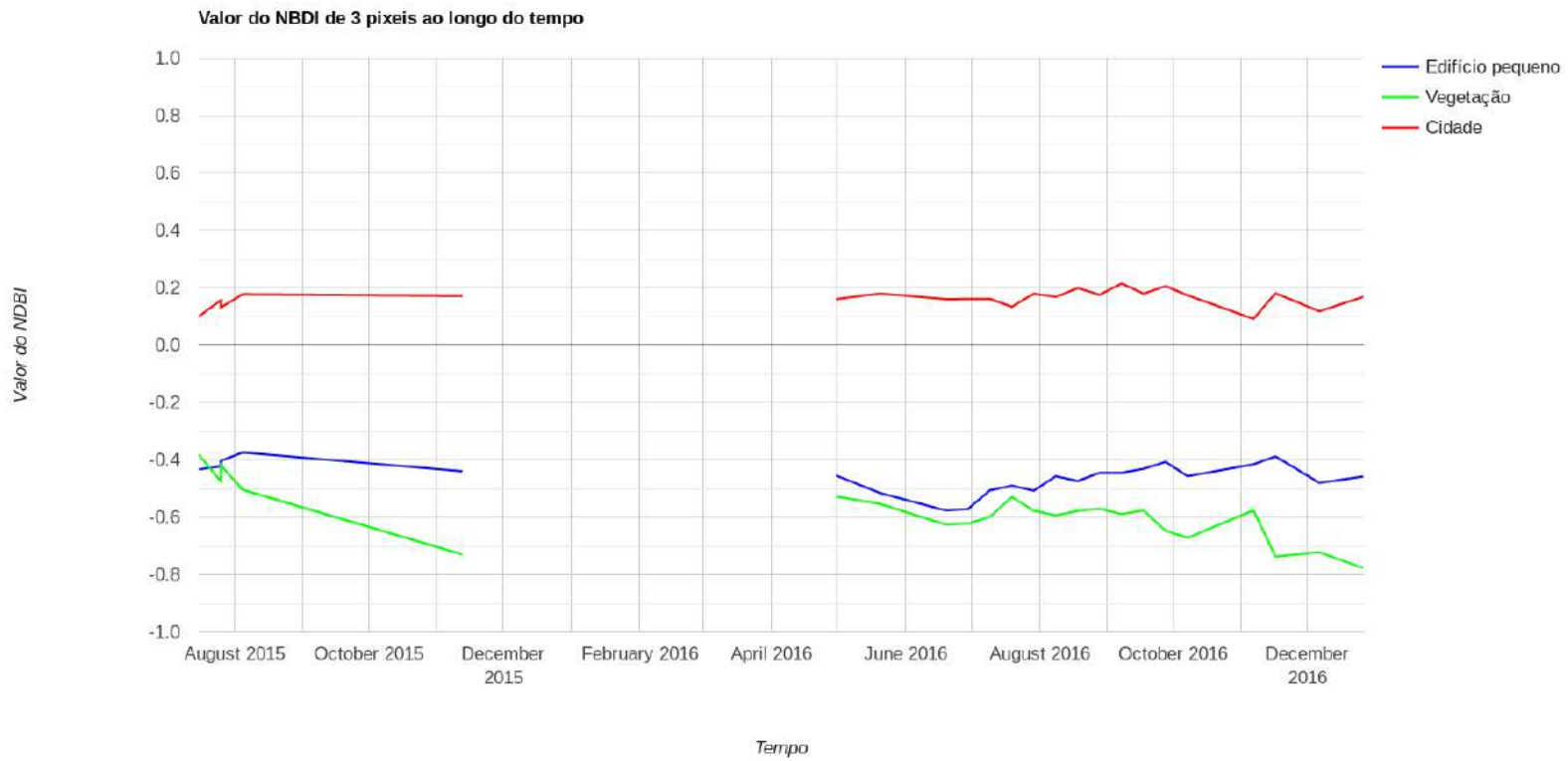


Figura 6.1: Variação do valor NBDI de 3 pixels ao longo do tempo

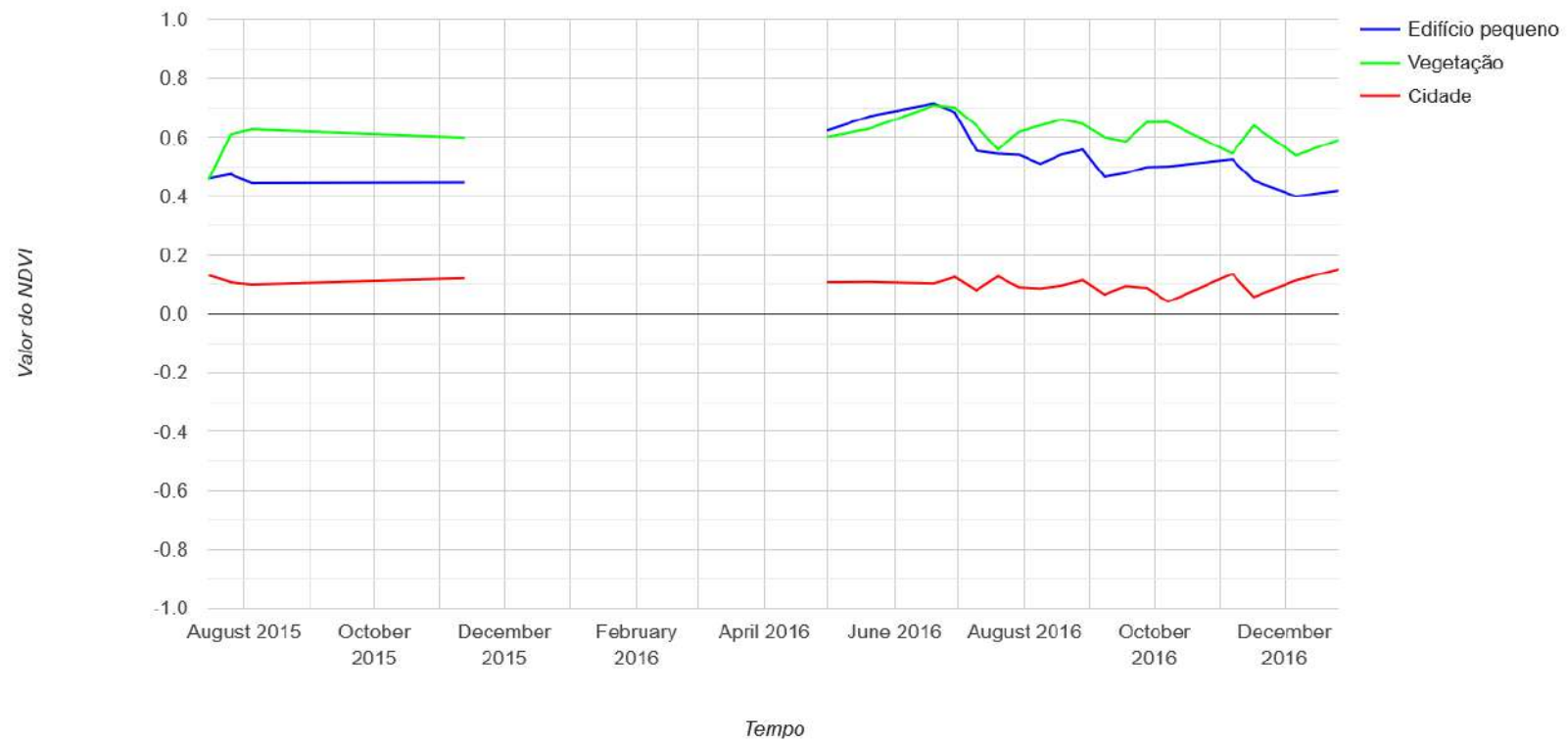


Figura 6.2: Variação do valor NDVI de 3 píxeis ao longo do tempo

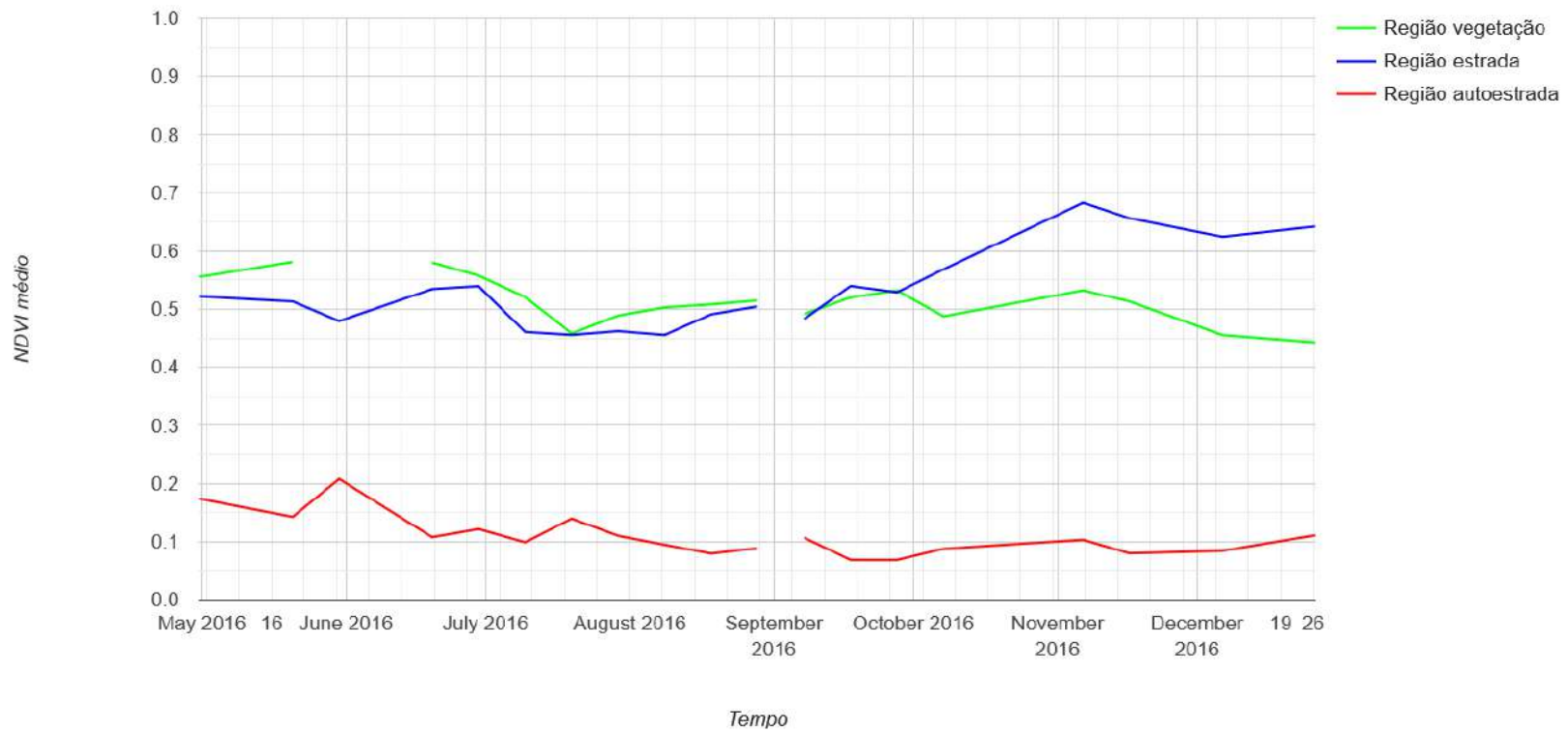


Figura 6.3: Variação do valor NDVI de diferentes tipos de estradas ao longo do tempo

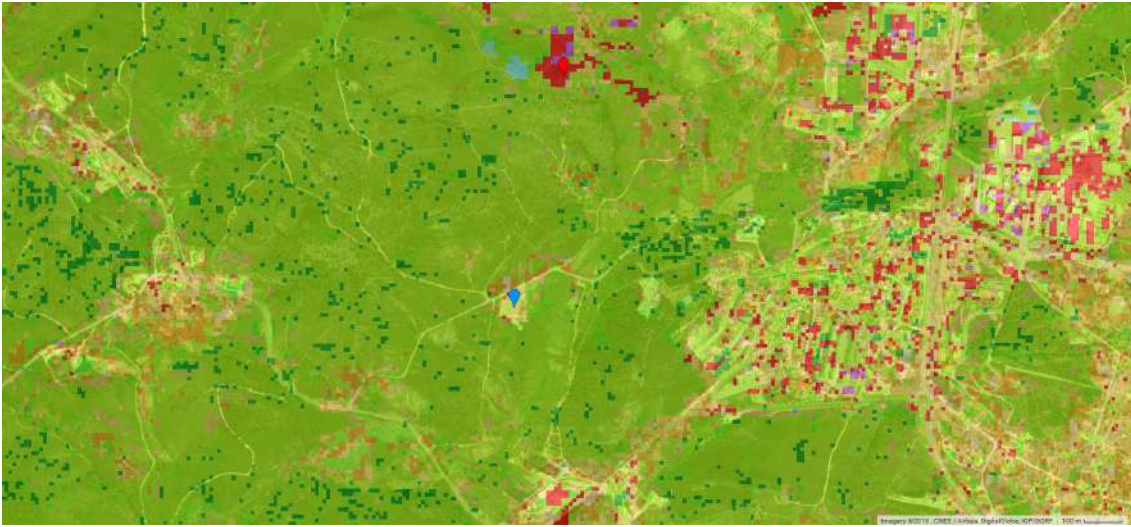


Figura 6.4: Resultado final da classificação de uma imagem do Sentinel-2 utilizando a carta CLC como dados de treino

## 6.1 Validação e análise experimental

### 6.1.1 Classificação de solo

Como experiência inicial, para validar o âmbito da dissertação, foi efetuada uma classificação utilizando todas as bandas do satélite Sentinel-2 sobre a área estudada. A classificação utilizou uma imagem sobre a área de Vila de Rei com todas as bandas de um produto de uma só data da constelação Sentinel-2 e 44 classes extraídas do mapa CLC. Estes dados são dados como input a um algoritmo *Random Forest*. Neste caso não foi feito um mapeamento das classes CLC para as classes das tarefas de classificação finais, visto que o objetivo principal desta experiência é observar o comportamento de uma classificação com base no estudado.

A figura 6.4 apresenta o resultado final sobreposto a uma imagem de muito alta resolução retirada do Google Earth. Os tons de verde pertencem às classes que representam vegetação e os tons de vermelho e roxo a classes que representam estruturas artificiais. Os resultados estão de acordo com o estudado. Pode ser observado que pequenos aglomerados populacionais e estruturas artificiais pequenas são representadas na classificação mesmo sem estarem representadas no mapa CLC. Segundo Peseresi et al. [55], isto acontece se existirem dados suficientes na imagem do Sentinel-2 que contrariem os dados do mapa CLC.

Na figura 6.5a o marcador vermelho indica uma zona de erro, esta zona está classificada como estrutura artificial mas apenas possui algum rochedo. Este tipo de erro é também presente no estudo de Peseresi et al. [55]. Estruturas artificiais e zonas em que rochas são predominantes têm um características espectrais semelhantes [55], este comportamento pode ser mitigado utilizando mais informação, como informação de textura ou a utilização de series temporais.

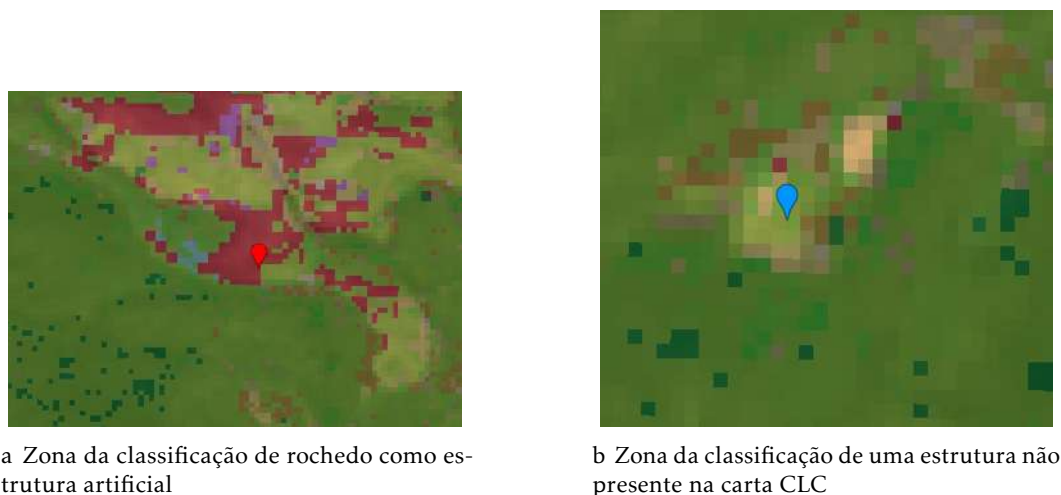


Figura 6.5: Detalhe aumentado dos fenómenos observados



Figura 6.6: Estradas parcialmente detetadas e estradas ausentes da classificação

Estradas com largura inferior a 10 metros são classificadas como sendo vegetação, a sua largura não tem impacto suficiente nos valores de bandas de um pixel para que a sua deteção seja possível, como pode ser visto na figura 6.6. Tendo em conta que o Sentinel-2 possui uma das resoluções mais altas dos dados a utilizar na solução estas estradas não ser detetadas.

Concluindo, os resultados foram de acordo com o esperado e confirmam que existe a possibilidade de refinar os mapas LULC de modo a detetar estruturas artificiais permanentes.

## 6.2 Resultados da otimização dos algoritmos

Para a fase de otimização dos classificadores, foi utilizada a metodologia descrita na sub-seção 4.4.1.1. A fase de pesquisa aleatória foi limitada a 1000 iterações para cada

teste, por algoritmo. Os dados de treino utilizados foram uma combinação de todas as bandas Sentinel-1, todas as bandas Sentinel-2, índices radiométricos respetivos e, por fim, o modelo de elevação de terreno e seus derivados. Para efeitos de otimização apenas foram utilizados 5% dos dados para diminuir a carga computacional. Esta secção apresenta os resultados desta otimização para cada algoritmo juntamente com os seus parâmetros ótimos, sendo a métrica de otimização o Kappa médio de todas as classificações.

### 6.2.1 Random Forest

A melhor combinação de parâmetros está presente na Tabela 6.1, obtendo um Kappa 0,70.

Tabela 6.1: Parâmetros ótimos após otimização do classificador Random Forest

Parâmetro	Valor ótimo
n_estimators	500
min_samples_leaf	4
min_samples_split	2
max_depth	130
class_weight	'balanced'

### 6.2.2 Gradient Tree Boosting

Os parâmetros escolhidos, e seus valores após a otimização, estão presentes na Tabela 6.2. Estes obtiveram um Kappa 0,77.

Tabela 6.2: Parâmetros ótimos após otimização do classificador XGBoost

Parâmetro	Valor ótimo
n_estimators	1500
colsample_bytree	0.7553707061597048
gamma	5
learning_rate	0.2049732654267658
max_depth	8
min_child_weight	1
max_delta_step	9.075685204314162

### 6.2.3 Artificial Neural Networks

As redes neuronais possuem uma interface e estrutura bastante diferente dos outros classificadores, a mesma foi otimizada utilizando apenas pesquisa de parâmetros por força bruta, visto que o intervalo de valores escolhido é mais reduzido do que os outros modelos. O numero de camadas intermédias tomou os valores 1, 2, e 3, com n\_neurons\_per\_layer apenas com valores em potências de 2 no intervalo  $[2^5; 2^6; 2^7; 2^8; 2^9; 2^{10}]$ . Após a otimização a rede neuronal ótima possui a seguinte estrutura (n\_input, 256, 512, 512 ,n\_classes), com um Kappa 0,66.

### 6.2.4 Support Vector Machines

Para o modelo **SVM** os melhores parâmetros após otimização estão presentes na Tabela 6.3. Obtendo um Kappa de 0,71 e F1 score de 0,59.

Tabela 6.3: Parâmetros ótimos após otimização do classificador SVM

Parâmetro	Valor ótimo
C	6.685338321430641
gamma	6.507029881541734
kernel	'rbf'
class_weigth	'balanced'

### 6.2.5 Stochastic Gradient Descent

Os parâmetro ótimos para o modelo **SGD** estão presentes nas Tabelas 6.4, 6.5 e 6.6, para as tarefas de classificação 1, 2 e 3, respetivamente. Com um valor Kappa médio de 0,56.

Tabela 6.4: Parâmetros ótimos, para a tarefa de classificação 1, após otimização do classificador SGD

Parâmetro	Valor ótimo
alpha	1e-05
l1_ratio	0.3879508123619403
max_iter	'1000'
penalty	'elasticnet'
class_weigth	'balanced'

Tabela 6.5: Parâmetros ótimos, para a tarefa de classificação 2, após otimização do classificador SGD

Parâmetro	Valor ótimo
alpha	1e-06
l1_ratio	0.5611522829923167
max_iter	'500'
penalty	'L2'
class_weigth	'balanced'

Tabela 6.6: Parâmetros ótimos, para a tarefa de classificação 3, após otimização do classificador SGD

Parâmetro	Valor ótimo
alpha	1e-05
l1_ratio	0.7751072005346229
max_iter	'500'
penalty	'elasticnet'
class_weigth	'balanced'



Labels	Dataset	Class	Model SGD		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série temporal S1-S2	1	0.5198	0.4215	0.6779
		2	0.8881	0.9917	0.8040
		3	0.2094	0.1171	0.9905
Grupo 2	Série temporal S1-S2	1	0.3258	0.2558	0.4484
		2	0.3550	0.4003	0.3190
		3	0.3990	0.3070	0.5698
		4	0.9743	0.9821	0.9666
		5	0.9130	0.9425	0.8853

Figura 6.7: Métricas de avaliação do classificador SGD após otimização, comparando duas tarefas de avaliação diferentes para os mesmos dados de treino.

Este modelo possui três diferentes parâmetros, por tarefa de classificação, pois durante o desenvolvimento foi observado que este modelo é muito sensível a pequenas mudanças nos dados de referencia. A Figura 6.7, mostra as métricas de avaliação após otimização do modelo, F1 score, precisão e *recall*, respectivamente, e mostra este fenómeno. Por exemplo a classe “Água”, representada no grupo 1 pelo identificador 3 e no grupo 2 pelo identificador 5, sofre a maior descida de performance.

### 6.3 Conclusão

Neste capítulo, foram abordados dois tópicos principais: uma análise preliminar, que serve como prova de conceito antes da apresentação dos resultados finais, e uma otimização dos algoritmos propostos que é utilizada na fase de treino e para a obtenção dos resultados finais.

A análise preliminar teve lugar no início do desenvolvimento da dissertação, com o âmbito de validar conceitos estudados na literatura e ter uma prova que existe potencial na metodologia apresentada. Notou-se que, as estruturas mais pequenas, como estradas e edifícios singulares, apresentam a maior probabilidade de erro. Isto é algo de esperar, visto que com uma resolução máxima de 10 por 10 metros é difícil a obtenção de informação sobre estruturas mais estreitas e pequenas, pode-se dizer que, na perspectiva dos algoritmos, as imagens apresentam pouco detalhe nos limites das estruturas.

Logo após esta análise foi efetuada uma classificação utilizando todas as bandas disponíveis de uma única imagens da constelação Sentinel-2. Concluindo com duas observações importantes, observadas também na literatura, em Pesaresi et al. 2016 [55]: O potencial para a deteção de estruturas artificiais permanentes não presentes nos dados de referencia e a confusão espectral de algumas estruturas naturais com artificiais, nomeadamente em zonas rochosas. Os resultados da classificação corroboram também o apresentado no parágrafo anterior, a falta da representação das estradas no resultado final. O comportamento das estradas neste caso pode ser melhorado aplicando o classificador a novos

dados de referência com uma melhor representação de estradas, dado que a resolução da carta utilizada penaliza as mesmas pois são estruturas pequenas.

Finalmente, os resultados obtidos da otimização, ou seja os parâmetros para cada algoritmo, é algo direto de analisar. Visto que a cadeia de processamento para a obtenção dos mesmos é baseada em mecanismos automáticos, os parâmetros finais a utilizar são os resultantes dos testes efetuados. É de importância referir que os mesmos não garantem o ótimo mas os parâmetros convergem para uma adaptação robusta aos dados.

No capítulo seguinte, utilizando então os resultados da otimização, é feita a análise final: a comparação de todos os conjuntos de dados e algoritmos de classificação.

## RESULTADOS

Após uma breve análise e validação, apresentada no capítulo anterior, são agora analisados os resultados finais, que respondem aos objetivos propostos inicialmente. Ao utilizar os parâmetros ótimos dos algoritmos são então corridos os testes finais e agregados todos os resultados, serão apresentadas todas as métricas necessárias com as devidas interpretações e conclusões que foram feitas no final de cada teste. O capítulo, dividido em 6 partes, está repartido da seguinte forma:

- **Metodologia experimental**, secção 7.1: Esta secção tem como função objetivo uma introdução aos testes efetuados durante a dissertação. São apresentadas as combinações de dados de treino, para cada algoritmo e para cada tarefa de classificação T.
- **Análise dos resultados**, secção 7.2: A primeira experiência que visa a resolver o problema principal da dissertação é apresentada nesta secção. Aqui são sintetizados os resultados do treino com dados estáticos, por algoritmo. A experiência mais importante da dissertação é também apresentada nesta secção, efetuada com séries temporais. A mesma visa a melhorar significativamente os resultados para cada modelo.
- **Validação dos modelos e do seu erro**, secção 7.3: Com as imagens e as métricas obtidas na secção anterior, é agora necessário proceder a uma análise do erro e interpretar manualmente os resultados obtidos. Esta análise compara o resultado com produtos estado da arte e ajuda a combater o impacto do balanceamento das classes nas métricas.
- **Caso de estudo: Faixas de gestão de combustível**, secção 7.4: Após a geração das cartas que mapeiam estruturas artificiais permanentes, provenientes das experiências anteriores, é estudado então o potencial das mesmas para a delimitação das

faixas de gestão de combustível.

- **Resultados da redução dimensional dos dados de treino**, secção 7.5: Estudo do impacto de uma redução dimensional no melhor modelo e no melhor conjunto de dados, esta secção mostra os resultados da redução dimensional dos dados de treino com a metodologia proposta.
- **Discussão**, secção 7.7: O capítulo é concluído com uma discussão de resultados que abre caminho para a conclusão final e às respostas dos objetivos propostos inicialmente.

## 7.1 Metodologia experimental

Nas secções anteriores foram mostrados os resultados obtidos da primeira fase da cadeia de processamento, incluindo os resultados da otimização dos algoritmos e da redução dimensional dos dados. Estes resultados abrem caminho para as análises mais importantes desta dissertação: o treino dos algoritmos com as diversas estruturas dos dados de treino e dos dados de referência.

A Figura 7.1 resume a cadeia de testes efetuados para responder aos problemas desta dissertação. São utilizados três tipos de dados de treino, dados estáticos, séries temporais e a complementação das mesmas com dados de elevação de terreno, em conjunção com as três tarefas de classificação que depois são corridas nos variados algoritmos de classificação. Presentes no apêndice D está uma enumeração dos atributos utilizados para cada conjunto de dados, dados estáticos na Tabela D.1, séries temporais na Tabela D.2 e séries temporais complementadas com dados de elevação de terreno na Tabela D.3. No final são obtidos resultados de 35 testes independentes.

## 7.2 Análise dos resultados

Os resultados de todos os testes são representados pela estatística Kappa na Figura 7.2. estes estão divididos por grupo, por tipos de dados e por modelo. Pode ser observado que o modelo XGBoost apresenta os melhores resultados em todos os grupos de classificação, os melhores resultados pertencem aos métodos de *ensemble*, incluindo o classificador *Random Forest*. À medida que a complexidade dos dados aumenta, de dados estáticos para séries temporais e finalmente a adição do modelo de elevação de terreno, a performance dos algoritmos também aumenta. Analisando os grupos de classificação, pode ser observado que, apesar de existir uma descida geral das métricas do grupo 1 para o grupo 2, esta não é muito acentuada. Concluindo, apesar da tarefa de classificação se tornar mais complexa os modelos adaptam-se bem à mesma, gerando resultados que rivalizam com os do grupo 1. Finalizando com o grupo 3, este grupo foi o grupo mais complexo de classificar. Como foi previsto, a classificação de estradas é algo difícil para os modelos.

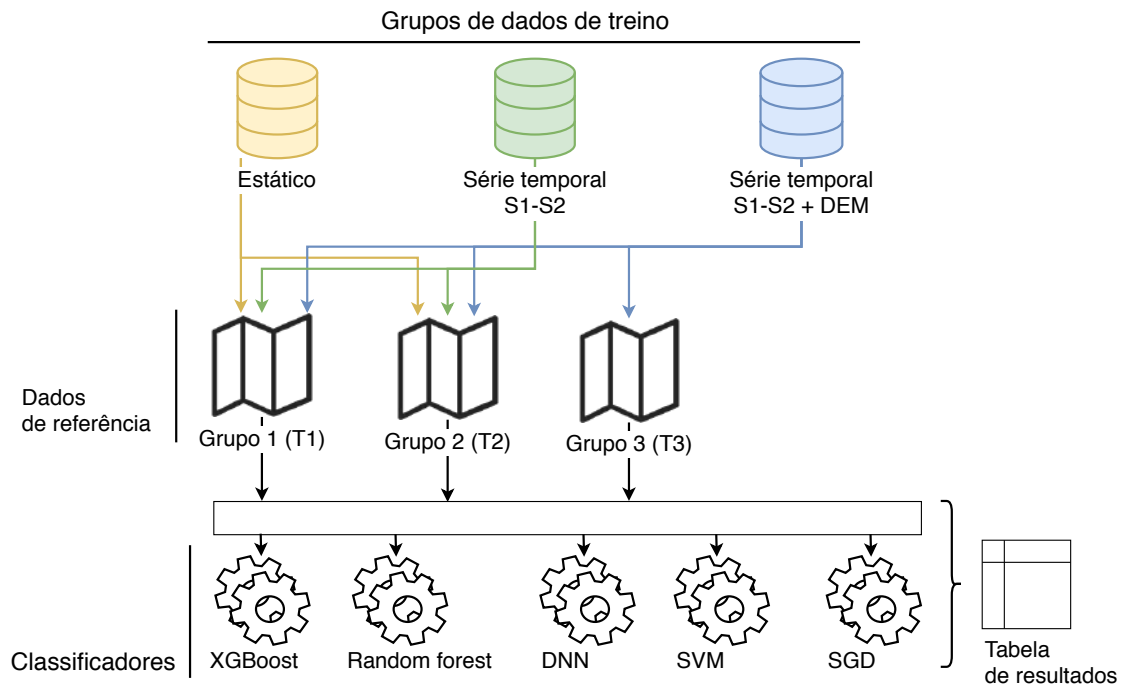


Figura 7.1: Pipeline das experiências.

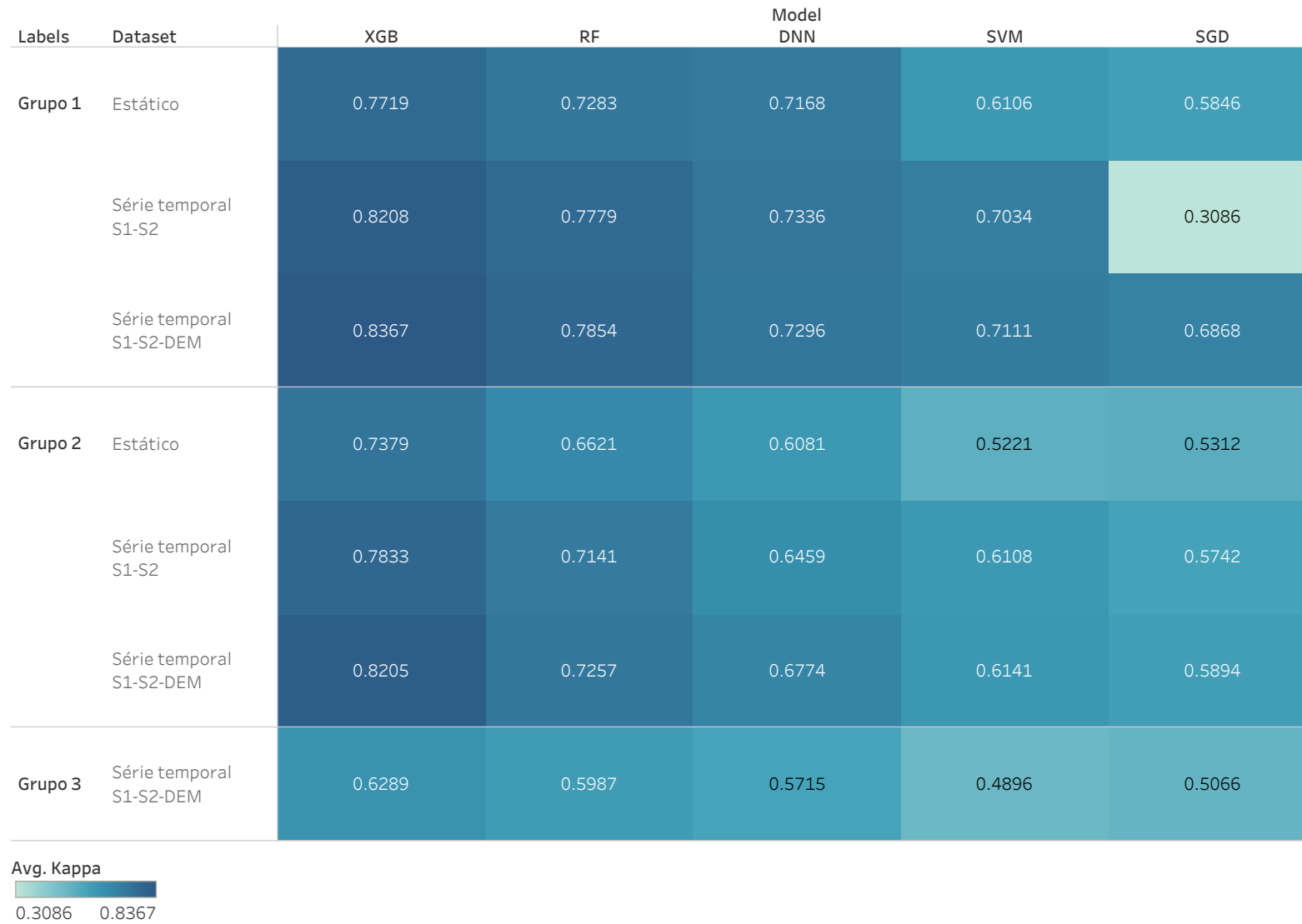


Figura 7.2: Métricas de teste Kappa, por grupo, por dataset, por modelo.

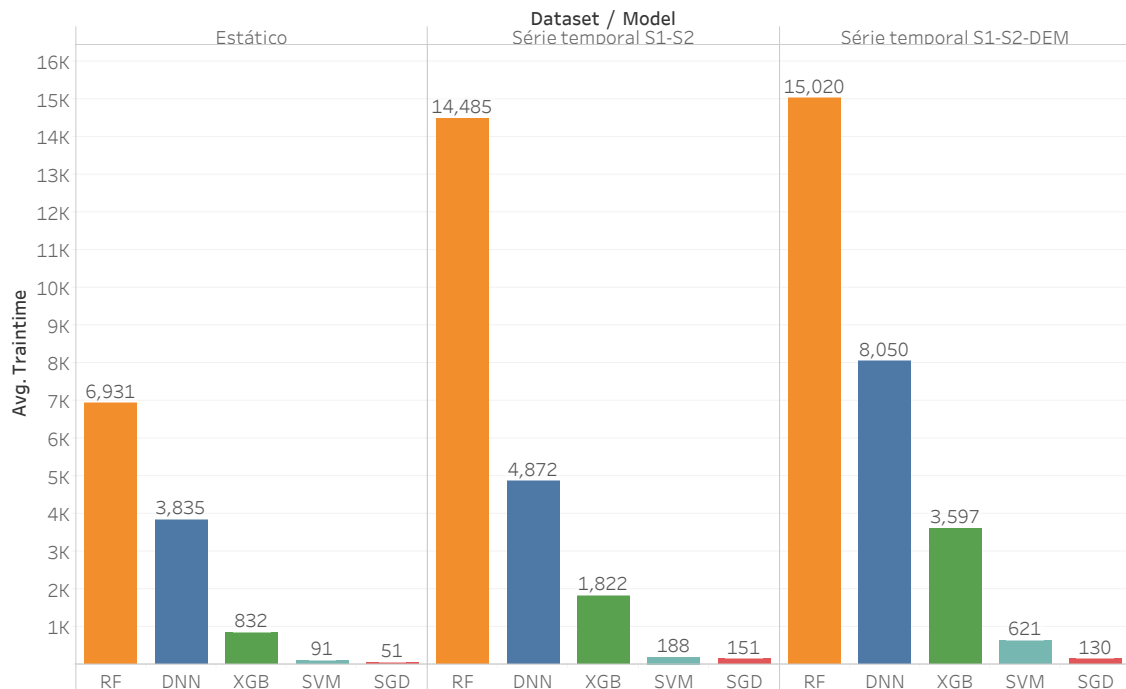


Figura 7.3: Tempo de treino médio em segundos, por dados de treino, para cada modelo.

Agora que a performance dos algoritmos foi apresentada existe uma outra métrica bastante importante: o tempo de treino. A Figura 7.3 apresenta o tempo de treino médio (em segundos) dos testes, para cada conjunto de dados de treino e algoritmo. É interessante observar que quando se fala sobre tempo de treino, o modelo XGBoost não só tem melhores resultados mas também tem o melhor tempo de treino dos melhores algoritmos. Os classificadores SVM e SGD, apesar de terem os piores resultados, apenas demoram alguns minutos, o que é algo a favor dos mesmos. Mesmo assim, a troca entre tempo de treino e performance geral é demasiado grande e não é justificável, visto que o melhor algoritmo termina o treino em duas horas.

Como observação final sobre o tempo de treino existe um comportamento interessante entre a utilização de séries temporais e a adição do modelo de elevação de terreno. Apesar da adição do modelo de elevação ser algo leve, adicionando apenas MDE, declive e orientação, o tempo de treino tem tendência a duplicar em dois dos melhores modelos, DNN e XGBoost. Estes são os dois algoritmos acelerados com GPUs, parte da justificação é o volume dos dados pré adição. A série temporal está no limiar do que o GPU da máquina de testes consegue correr e a adição de novos atributos pode ter impacto negativo. Esta conclusão não é definitiva e carece de mais testes dedicados aos comportamentos dos GPUs com os algoritmos, algo que está fora do escopo desta dissertação.

As imagens resultantes dos classificadores com melhores resultados estão presentes no apêndice C. São omitidas as imagens de classificadores com pouca performance, como os modelos SGD e SVM, e também omitidos grupos de classificação de alguns modelos pois não possuem resultados relevantes, quando comparados com os melhores.

Tabela 7.1: Mapeamento das classes do grupo 1 e do grupo 2

Grupo 1		Grupo 2	
1	Estrutura	1	Habitações urbanas
2	Natural	2	Habitações rurais
3	Água	3	Outras estruturas
		4	Natural
		5	Água

Nas próximas secções será analisado em mais detalhe o comportamento dos modelos. Começando com a análise da utilização de dados estáticos e séries temporais. Finalmente será também analisado em detalhe o comportamento dos algoritmos para a classificação do grupo 3, pois é o grupo com mais variância nos seus resultados.

### 7.2.1 Análise estática

Nesta secção são apresentados em mais detalhe os resultados dos testes utilizando dados de apenas uma data, imagens Sentinel-1 do dia 30/07/2015 e Sentinel-2 do dia 15/07/2015. Estes dados foram utilizados para treinar classificadores sobre o grupo 1 e o grupo 2 de classificação. Para efeitos de será utilizado o mapeamento presente na Tabela 7.1, quando forem referidas as classes de cada grupo de dados de referência.

Os resultados vão ser apresentados em duas parte, primeiramente uma comparação global de todos os algoritmos e posteriormente os resultados por algoritmo mais detalhadamente. Apresentado na Figura 7.4 está o valor da estatística Kappa para cada algoritmo. Os algoritmos de *ensemble*, XGBoost e *Random Forest*, prevalecem aos restantes obtendo valores de Kappa acima dos 0,70 no grupo 1, com o modelo *Random Forest* a descer para 0,66 face ao grupo 2. Os dois modelos baseados em *soft-margins*, SVM e SGD, apresentam os resultados mais baixos em ambos os grupos, isto foi algo inesperado pois na literatura os classificadores SVM obtêm bons resultados [48, 49]. O modelo XGBoost possui a melhor performance nos dois grupos e a maior consistência nos seus resultados, tendo o seu valor Kappa final apresentando uma pequena descida apenas para o grupo 2. As redes neuronais, apesar de ter sido inserido nos testes inicialmente como prova de conceito acabou por obter resultados que rivalizam com os melhores modelos. Estes modelos requerem mais tempo para o seu treino.

Com os resultados globais definidos é agora necessário verificar se o comportamento dos mesmos se estenda às classes individuais apresentadas. Os resultados para cada algoritmo por tarefa de classificação estão presentes nas seguintes figuras: 7.5 (XGBoost), 7.6 (Random Forest), 7.9 (Artificial Neural Network), 7.7 (Support Vector Machine) e 7.8 (Stochastic Gradient Descent).

Começando com o modelo de melhor performance, XGBoost, os seus resultados são representados na Figura 7.5. Para o grupo 1, a métrica mais baixa é o *recall* da classe estrutura permanente (classe 1) com um valor de 0,60. O que significa que o classificador está



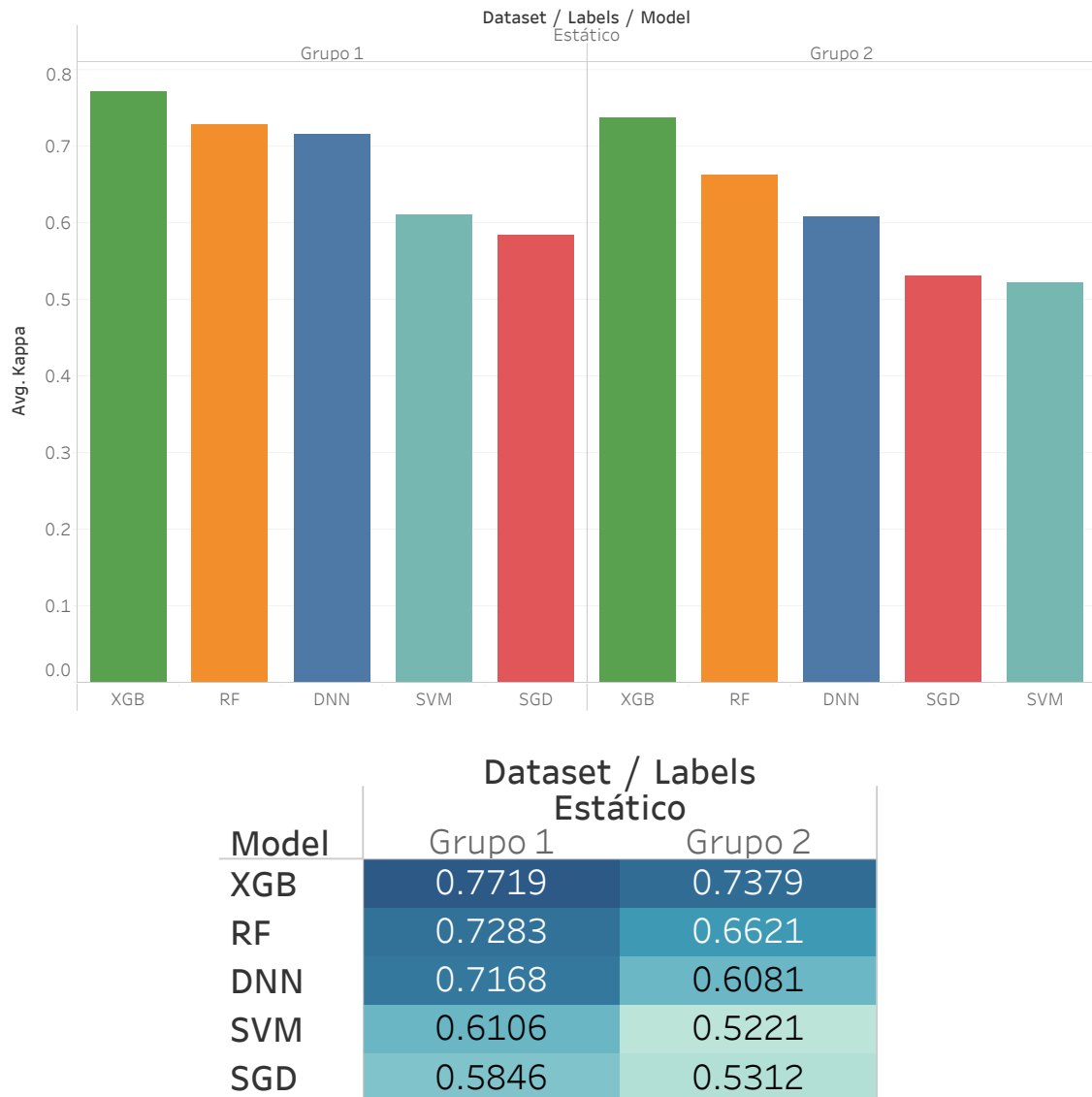


Figura 7.4: Valores da estatística Kappa para cada modelo, por tarefa de classificação

a omitir aproximadamente 40% das estruturas presentes nos dados de referência. Os resultados do grupo 2 apresentam a descida de métricas observada inicialmente. Quando se expande as estruturas artificiais permanentes do grupo 1 para as classes mais específicas (classes 1, 2 e 3), em geral todas as métricas descem.

Este tipo de comportamento é algo constante para a maior parte dos modelos. Ao observar os resultados do classificador *Random Forest*, presentes na Figura 7.6, é possível observar a mesma tendência nas descidas, sendo estas mais acentuadas. O mesmo é verdade nos modelos *SVM*, Figura 7.7, *SGD*, Figura 7.8 e *DNN*, Figura 7.9. Todos os modelos possuem os valores da precisão mais altos que o resto das métricas em todos os casos, com a exceção do classificador mais sensível, o *SGD*.

Existem apenas dois comportamentos fora do padrão analisado: o classificador *SGD*

Labels	Dataset	Class	Model XGB		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.7021	0.8313	0.6077
		2	0.9862	0.9793	0.9932
		3	0.9460	0.9655	0.9273
Grupo 2	Estático	1	0.5893	0.7748	0.4755
		2	0.5269	0.6927	0.4251
		3	0.7083	0.8418	0.6113
		4	0.9862	0.9774	0.9952
		5	0.9440	0.9656	0.9235

Figura 7.5: Métricas de avaliação do modelo *XGBoost* por tarefa de classificação, utilizando dados estáticos

Labels	Dataset	Class	Model RF		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.6534	0.7213	0.5972
		2	0.9829	0.9786	0.9873
		3	0.9267	0.9329	0.9206
Grupo 2	Estático	1	0.4021	0.5816	0.3073
		2	0.3955	0.4833	0.3346
		3	0.6041	0.6658	0.5528
		4	0.9827	0.9754	0.9901
		5	0.9260	0.9379	0.9145

Figura 7.6: Métricas de avaliação do modelo *Random Forest* por tarefa de classificação, utilizando dados estáticos

apresenta a mesma descida de métricas comparando o grupo 1 e 2, sendo esta mais grave na precisão do que no *Recall*; o modelo *DNN*, este apresenta a descida detetada mas muito mais acentuada, chegando a valores de *Recall* perto dos 10% descendo um quinto do seu valor na classe equivalente do grupo 1.

Concluindo, os melhores resultados foram observados no modelo *XGBoost*, apresentando não só as melhores métricas mas também a melhor consistência das mesmas. Terminando esta análise, a próxima secção analisará de forma semelhante os resultados obtidos com dados de séries temporais.

Labels	Dataset	Class	Model SVM		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.4690	0.8263	0.3276
		2	0.9802	0.9643	0.9965
		3	0.8834	0.9657	0.8141
Grupo 2	Estático	1	0.2773	0.5077	0.1908
		2	0.0851	0.5556	0.0461
		3	0.4298	0.7424	0.3025
		4	0.9786	0.9593	0.9986
		5	0.8834	0.9728	0.8090

Figura 7.7: Métricas de avaliação do modelo *Support Vector Machine* por tarefa de classificação, utilizando dados estáticos

Labels	Dataset	Class	Model SGD		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.5055	0.5247	0.5202
		2	0.9707	0.9755	0.9662
		3	0.7753	0.6997	0.8936
Grupo 2	Estático	1	0.2821	0.2071	0.4422
		2	0.2308	0.2875	0.1928
		3	0.4089	0.3484	0.4949
		4	0.9726	0.9772	0.9681
		5	0.8579	0.8752	0.8412

Figura 7.8: Métricas de avaliação do modelo *Stochastic Gradient Descent* por tarefa de classificação, utilizando dados estáticos

Labels	Dataset	Class	Model DNN		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.6381	0.7536	0.5534
		2	0.9829	0.9760	0.9899
		3	0.9153	0.9374	0.8942
Grupo 2	Estático	1	0.1935	0.5883	0.1158
		2	0.3387	0.4877	0.2595
		3	0.4897	0.7590	0.3614
		4	0.9820	0.9681	0.9962
		5	0.9070	0.9445	0.8724

Figura 7.9: Métricas de avaliação do modelo *Artificial Neural Network* por tarefa de classificação, utilizando dados estáticos

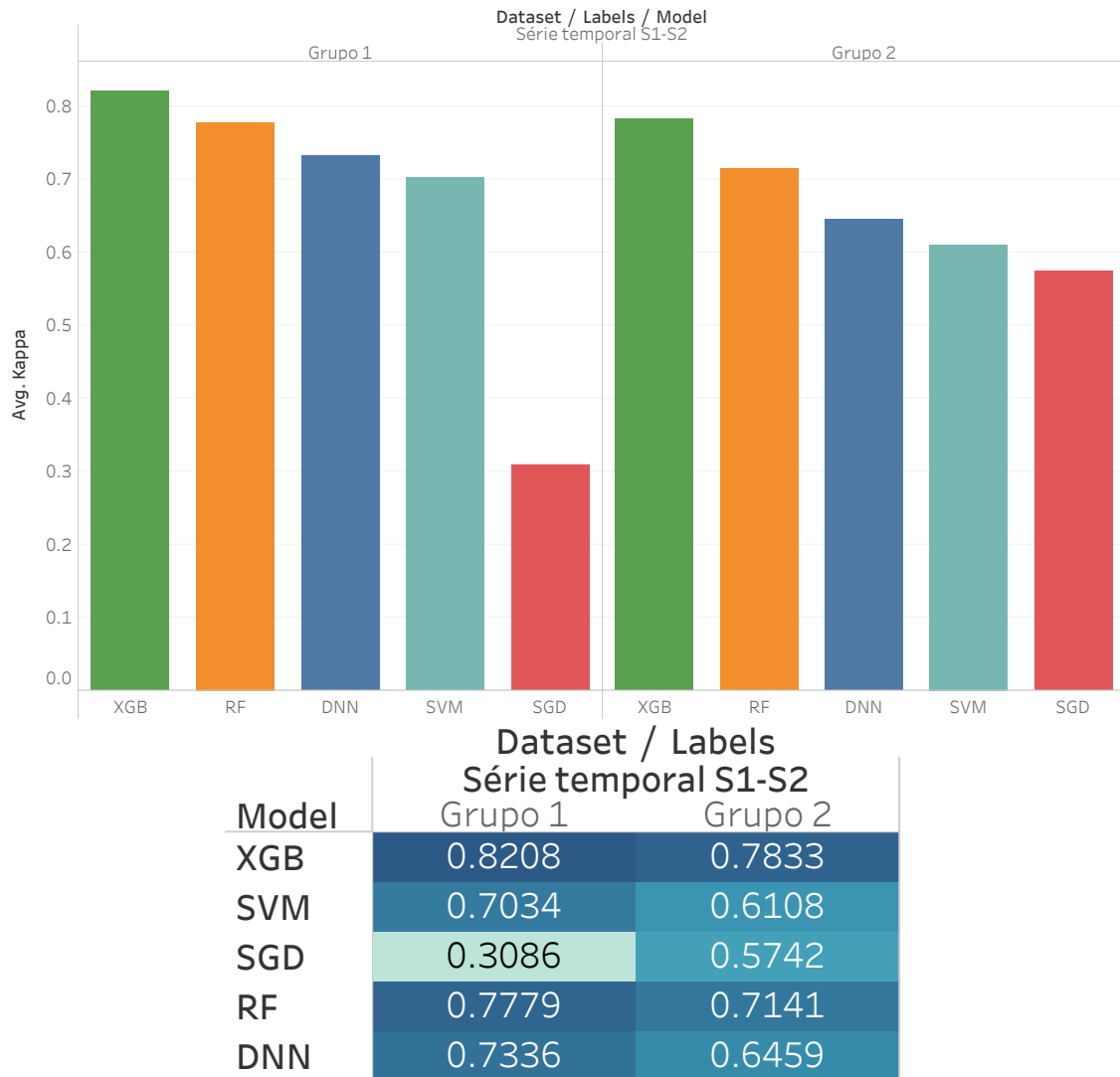


Figura 7.10: Valores da estatística Kappa para cada modelo, por tarefa de classificação

### 7.2.2 Análise de séries temporais

Em paralelo ao apresentado na secção anterior, esta secção apresenta os resultados da utilização de séries temporais em lugar dos dados estáticos. Em geral, é observado na Figura 7.10 o mesmo comportamento do que na análise anterior, sendo o melhor classificador o XGBoost seguido dos outros modelos pela mesma ordem, *Random Forest*, *DNN* e *SVM*. É observado também uma descida de todas as métricas quando se classifica o grupo 2 devido à complexidade acrescida da tarefa.

Com a exceção do modelo *SGD*, que apresenta uma descida de métricas quando classifica o grupo 1. Como visto anteriormente o classificador *SGD* é sensível a mudanças nos dados de referência e treino. Esta mudança causa então o resultado da sua performance do grupo 1 utilizando séries temporais, tendo um valor de 0,3 em contraste com 0,57 no grupo 2. Uma otimização nova para este teste em específico fará com que a performance do mesmo seja perto da do modelo *SVM*. Como já foi concluído que o melhor modelo é o

Labels	Dataset	Class	Model XGB		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série	1	0.7670	0.8493	0.6993
	temporal	2	0.9888	0.9841	0.9934
	S1-S2	3	0.9626	0.9770	0.9486
Grupo 2	Série	1	0.6218	0.8045	0.5067
	temporal	2	0.6017	0.6814	0.5387
	S1-S2	3	0.7595	0.8717	0.6728
		4	0.9887	0.9825	0.9951
		5	0.9616	0.9759	0.9478

Figura 7.11: Métricas de avaliação do modelo *XGBoost* por tarefa de classificação, utilizando séries temporais

Labels	Dataset	Class	Model RF		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série	1	0.7191	0.7492	0.6913
	temporal	2	0.9856	0.9834	0.9878
	S1-S2	3	0.9438	0.9562	0.9317
Grupo 2	Série	1	0.4521	0.6250	0.3542
	temporal	2	0.4980	0.4989	0.4971
	S1-S2	3	0.6499	0.6860	0.6174
		4	0.9854	0.9822	0.9886
		5	0.9420	0.9588	0.9258

Figura 7.12: Métricas de avaliação do modelo *Random Forest* por tarefa de classificação, utilizando séries temporais

XGBoost esta otimização não foi feita pois apresentaria piores resultados que máximo já atingido.

Os resultados dos testes individuais estão presentes nas Figuras: 7.11 (XGBoost), 7.12 (Random Forest), 7.13 (Artificial Neural Network), 7.14 (Support Vector Machine) e 7.15 (Stochastic Gradient Descent).

É concluído o mesmo que na secção anterior utilizando dados estáticos. O melhor algoritmo é o XGBoost, seguido do outro método *ensemble Random forest*. O comportamento dos algoritmos é semelhante, existindo sempre a descida de performance com o aumento da complexidade da tarefa e mantém-se a métrica mais alta, a precisão. O único comportamento inesperado foi o do modelo SGD, em que existe uma anomalia na classificação

Labels	Dataset	Class	Model DNN		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série	1	0.6447	0.8496	0.5194
	temporal	2	0.9846	0.9755	0.9939
	S1-S2	3	0.9243	0.9130	0.9359
Grupo 2	Série	1	0.2016	0.5119	0.1255
	temporal	2	0.4027	0.5453	0.3192
	S1-S2	3	0.5408	0.7623	0.4190
		4	0.9837	0.9713	0.9964
		5	0.9243	0.9573	0.8934

Figura 7.13: Métricas de avaliação do modelo *Artificial Neural Network* por tarefa de classificação, utilizando séries temporais

Labels	Dataset	Class	Model SVM		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série	1	0.6041	0.8238	0.4769
	temporal	2	0.9836	0.9723	0.9951
	S1-S2	3	0.9193	0.9787	0.8668
Grupo 2	Série	1	0.0942	0.5000	0.0520
	temporal	2	0.3767	0.5362	0.2903
	S1-S2	3	0.4612	0.7643	0.3302
		4	0.9820	0.9675	0.9971
		5	0.9305	0.9726	0.8920

Figura 7.14: Métricas de avaliação do modelo *Support Vector Machine* por tarefa de classificação, utilizando séries temporais

do grupo 1, este comportamento carece de validação, mas como a performance geral do algoritmo tem mostrado ser sempre mais baixo que o modelo XGBoost esta não foi feita pois não atingirá valores mais altos, tornando a mesma desnecessária.

Labels	Dataset	Class	Model SGD		
			Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Série	1	0.5198	0.4215	0.6779
	temporal	2	0.8881	0.9917	0.8040
	S1-S2	3	0.2094	0.1171	0.9905
Grupo 2	Série	1	0.3258	0.2558	0.4484
	temporal	2	0.3550	0.4003	0.3190
	S1-S2	3	0.3990	0.3070	0.5698
		4	0.9743	0.9821	0.9666
		5	0.9130	0.9425	0.8853

Figura 7.15: Métricas de avaliação do modelo *Stochastic Gradient Descent* por tarefa de classificação, utilizando séries temporais



Model	Dataset / Labels			
	Série temporal S1-S2		Série temporal S1-S2-DEM	
	Grupo 1	Grupo 2	Grupo 1	Grupo 2
XGB	0.8208	0.7833	0.8367	0.8205
RF	0.7779	0.7141	0.7854	0.7257
DNN	0.7336	0.6459	0.7296	0.6774
SVM	0.7034	0.6108	0.7111	0.6141
SGD	0.3086	0.5742	0.6868	0.5894

Figura 7.16: Estatística Kappa de todos os modelos por tarefa de classificação, sem e com o modelo de elevação de terreno

### 7.2.3 Adição do modelo de elevação do terreno

Após ambas as análises anteriores, esta secção apresenta os resultados dos testes finais, correspondendo a treinar os algoritmos com séries temporais complementadas com o modelo de elevação de terreno e os dois derivados, declive e orientação. É aqui que é estudado a performance dos algoritmos para a extração de estradas, visto que este é o conjunto de dados com performance mais elevada.

Começando pelo impacto da adição do modelo de elevação de terreno nas classificações do grupo 1 e do grupo 2, a Figura 7.16 apresenta a estatística Kappa tipo de dados e por modelo. Observa-se que em geral a adição de dados de elevação aumentam a performance dos classificadores, sendo esta mais destacada no grupo 2. Para validar melhor esta conclusão foram criados dois gráficos, a Figura 7.17, para o grupo 1, e a Figura 7.18, para o grupo 2, foram selecionados apenas os três melhores modelos, XGBoost, DNN e *Random Forest*. Estes gráficos comparam o valor das métricas por classe para os dois tipos de dados, com elevação de terreno e sem, sendo uma mudança mais significativa observada quando a linha que representa a série temporal com o modelo de elevação de terreno está por cima da série temporal apenas. Concluindo das figura 7.17 e 7.18, a adição do atributo adicional tem pouco impacto no grupo 1, quanto ao grupo 2 esta adição melhora ligeiramente os resultados do classificador XGBoost em todas as classes. O mesmo é observado no classificador *Random Forest*, mas este possui uma mudança menos acentuada. Quando ao modelo DNN, a adição provoca uma subida no *Recall* das classes mas existe uma descida na sua precisão. Tendo em conta que o XGBoost é o modelo com melhores resultados até agora, esta adição dos dados de elevação de terreno é benéfica para a classificação final.

Dado o impacto da adição dos dados de elevação de terreno no grupo 1 e 2, é necessário explorar o grupo de classificação final, o grupo 3, este grupo contém a informação de estradas do OSM sobreposta à carta COS 2015, resultando nas seguintes classes: 1 -> Estrutura, 2 -> Estrada, 3 -> Natural e 4 -> Água.

Apresentados na Figura 7.19 estão os resultados da classificação do grupo 3 de todos

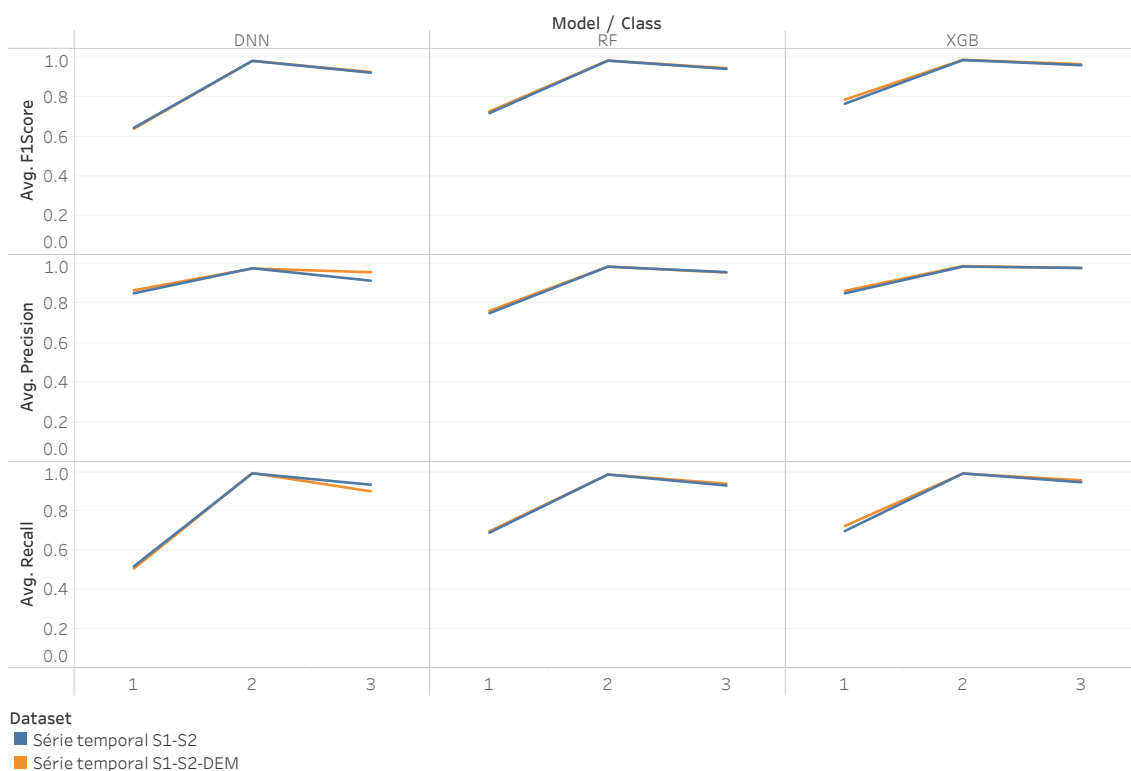


Figura 7.17: Impacto da adição do modelo de elevação de terreno na classificação do grupo 1

os algoritmos testados, em paralelo aos resultados anteriores o modelo com melhor performance continua a ser o XGBoost, seguido do classificador *Random Forest*. Virando a atenção para a coluna da classe 2 (estradas) da tabela da Figura 7.19, observa-se que o *Recall* é baixo, variando de 0,03 a 0,13, e a precisão é muito mais alta, ou seja, os classificadores têm dificuldade em detetar estradas mas as que detetam possuem uma alta probabilidade de serem estradas.

A Figura 7.20 mostra a imagem resultante da classificação do grupo 3 utilizando o modelo XGBoost. Dentro das cidade, representada na imagem com o aglomerado grande branco, podem ser observados vários pixels dispersos de estradas, as estradas dentro de cidades são difíceis de detetar por duas razões: o seu tamanho não é suficiente para as mesmas terem impacto no valor das bandas dos pixels e a sua assinatura espectral é semelhante à dos edifícios. Mesmo com o valor baixo das métricas apresentadas o modelo consegue classificar bem autoestradas e estradas principais.

Virando a análise para uma zona mais rural, a zona de Vila de Rei, obtém-se o resultado na Figura 7.21. As estradas dentro de cidades e aglomerados possuem os mesmos problemas enumerados anteriormente, as restantes estradas apresentam-se segmentadas, com consequência de um valor de *Recall* baixo, o algoritmo classifica erradamente vários segmentos de estradas, ou até estradas por completo.

A performance dos algoritmos para a deteção de estradas é problemática, as estradas são estruturas importantes para alguns tipos de análise. Dito isto são também as

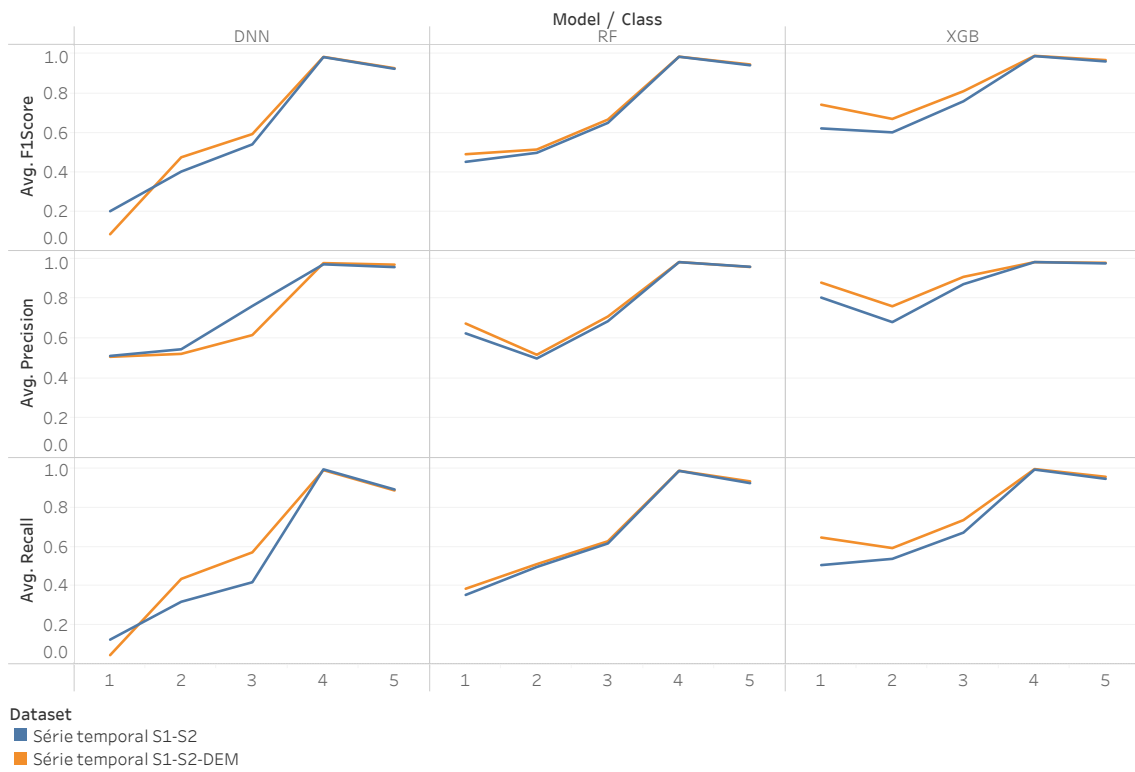
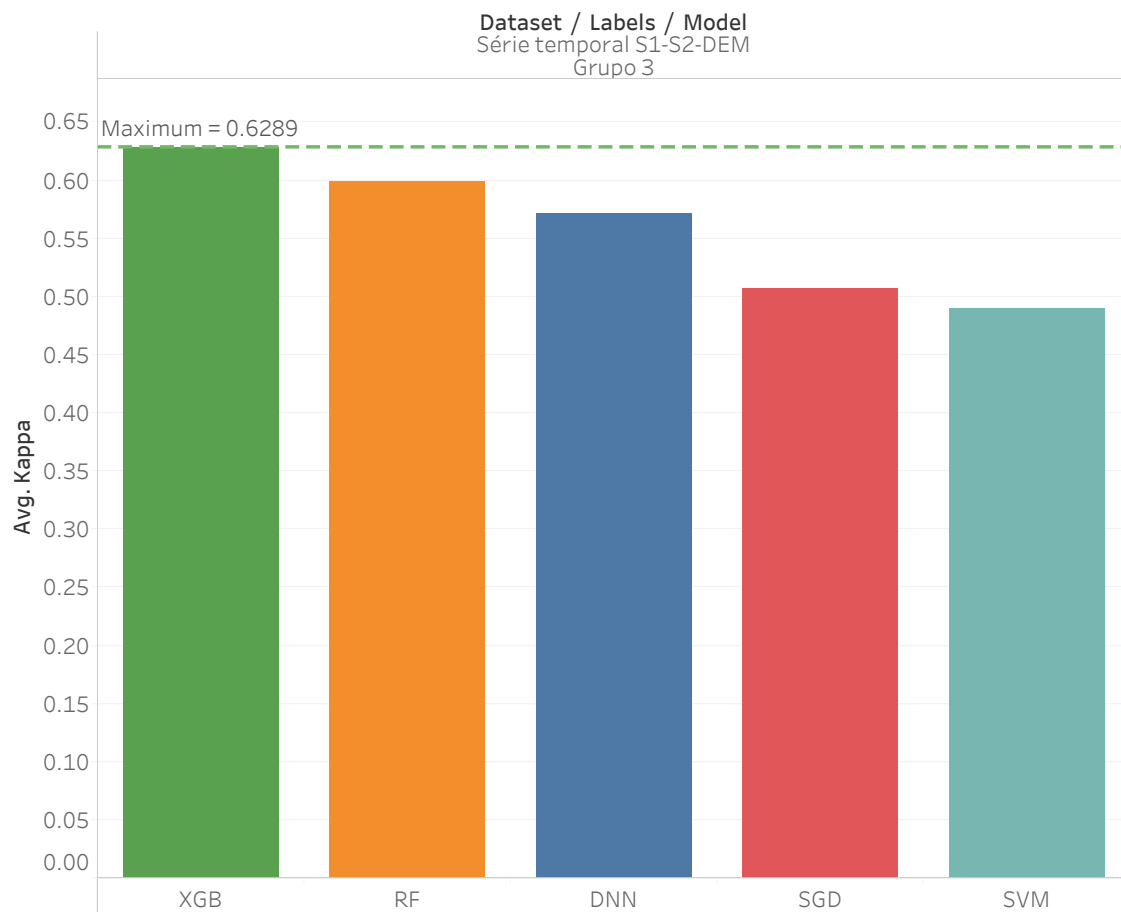


Figura 7.18: Impacto da adição do modelo de elevação de terreno na classificação do grupo 2

estruturas artificiais permanentes mais difíceis de detetar, seja pelo seu tamanho ou pela confusão das mesmas com outras classes. Para resolver este comportamento, será necessário utilizar outras metodologias focadas em estradas ou até encontrar fontes de dados de treino com resolução mais elevada.

CAPÍTULO 7. RESULTADOS



Model		Labels / Dataset / Class Grupo 3 Série temporal S1-S2-DEM			
		1	2	3	4
<b>XGB</b>	Avg. F1Score	0.6964	0.2201	0.9724	0.9639
	Avg. Precision	0.7326	0.6553	0.9530	0.9735
	Avg. Recall	0.6635	0.1323	0.9926	0.9545
<b>SVM</b>	Avg. F1Score	0.4938	0.0684	0.9657	0.9200
	Avg. Precision	0.6796	0.8277	0.9364	0.9733
	Avg. Recall	0.3880	0.0357	0.9968	0.8725
<b>RF</b>	Avg. F1Score	0.6457	0.1938	0.9688	0.9440
	Avg. Precision	0.6128	0.6153	0.9532	0.9471
	Avg. Recall	0.6823	0.1150	0.9850	0.9409
<b>DNN</b>	Avg. F1Score	0.6177	0.1199	0.9685	0.9271
	Avg. Precision	0.6116	0.6805	0.9488	0.9534
	Avg. Recall	0.6239	0.0658	0.9891	0.9023

Figura 7.19: Resultados da classificação dos diferentes modelos para o grupo 3.

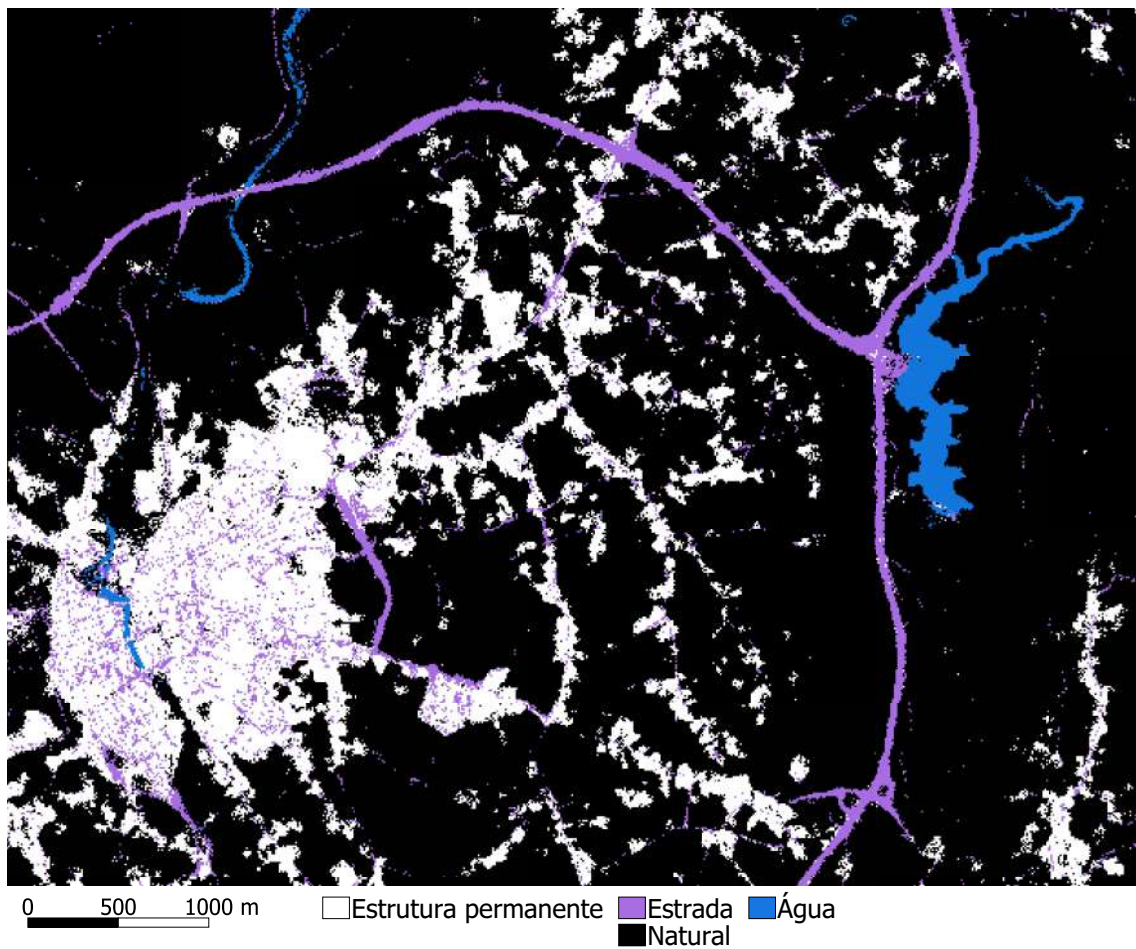


Figura 7.20: Imagem resultante da classificação do grupo 3 com modelo XGBoost utilizando séries temporais, sobre Tomar.

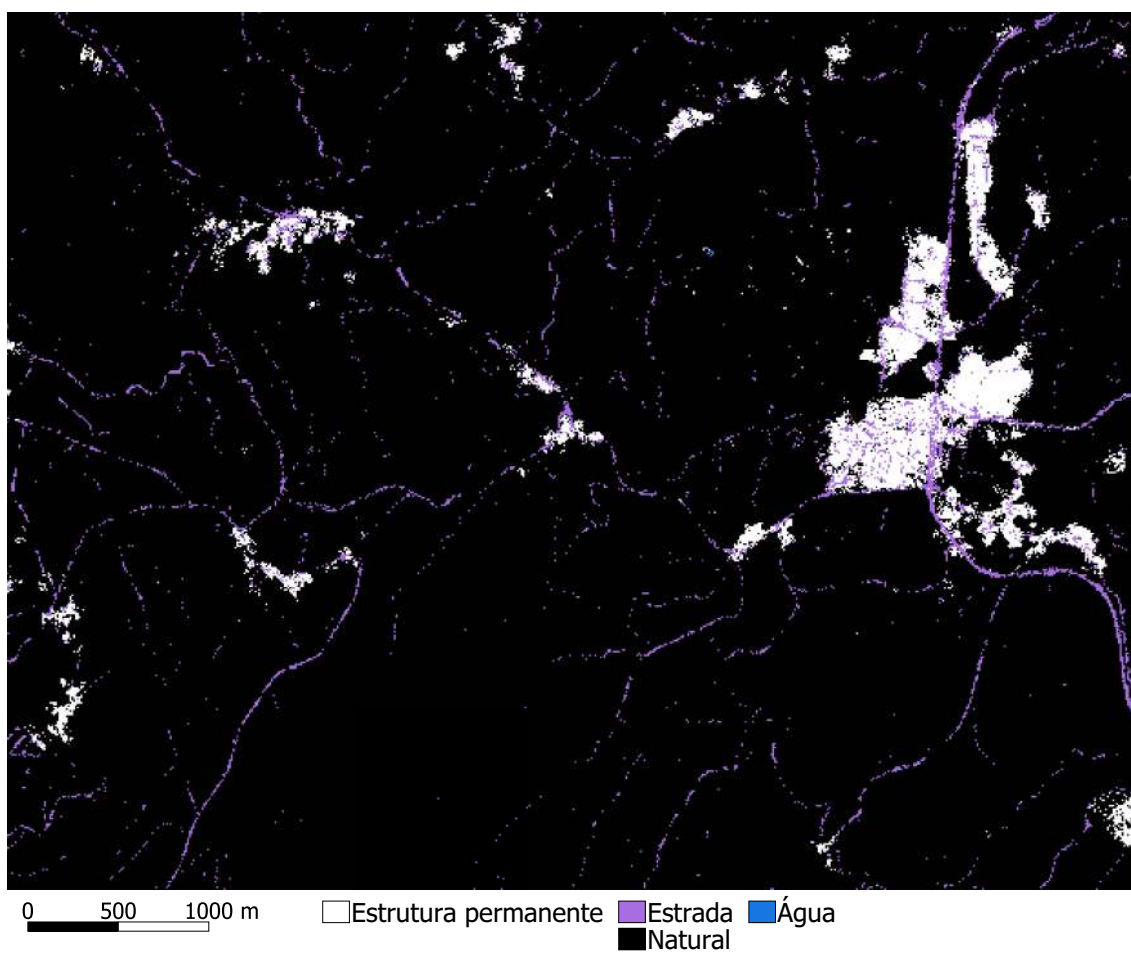


Figura 7.21: Imagem resultante da classificação do grupo 3 com modelo XGBoost utilizando séries temporais, sobre Vila de Rei.

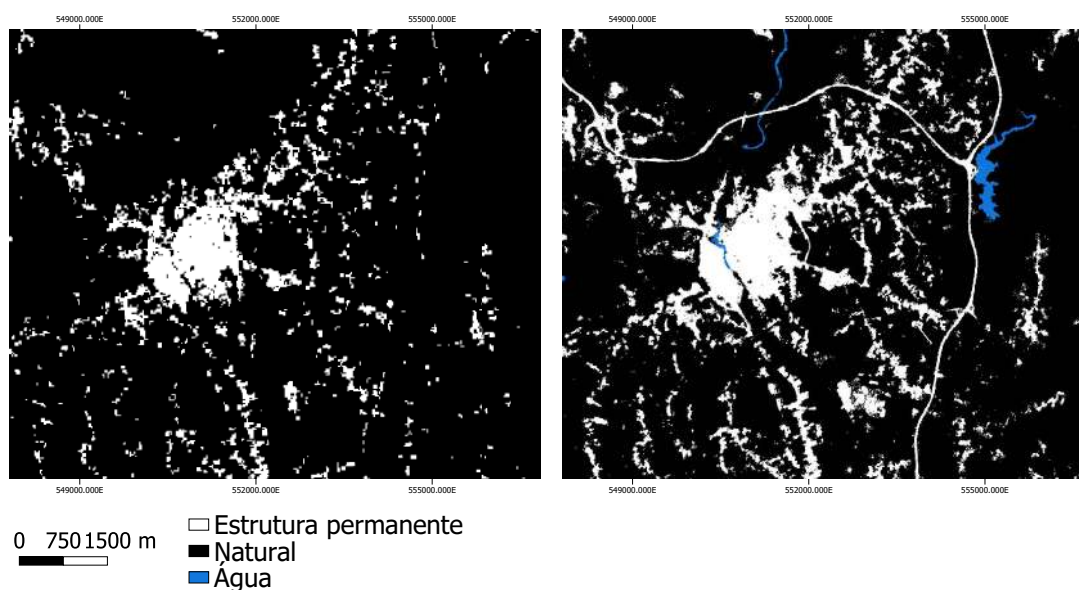


Figura 7.22: Comparação da carta GHSL, à esquerda, com a carta gerada, à direita. Imagens sobre a região de Tomar.

### 7.3 Validação dos modelos e análise do seu erro

Apesar das métricas apresentadas anteriormente serem satisfatórias, as mesmas apresentam um obstáculo à análise. Estas métricas correspondem à performance dos modelos tendo em conta a carta COS 2015. Assim é necessário avaliar o potencial do algoritmo para generalizar, ou seja detetar estruturas artificiais permanentes não existentes nos dados de referência. Para este efeito tem de ser feita uma avaliação manual dos resultados. Para a análise em questão foi escolhido o melhor algoritmo, XGBoost. Esta secção tem os seguintes objetivos:

- Comparação da classificação do grupo 1, com a carta GHSL Landsat 8.
- Análise da distribuição do erro nas classificações do grupo 1 e 2.

O primeiro objetivo vem em resposta ao problema dos dados de referência. Estes apenas possuem apenas estruturas artificiais e aglomerados com mais do que 1 hectare, por isso será necessário comparar os resultados com o estado da arte. Uma das melhores cartas disponíveis para o mapeamento de estruturas é a carta GHSL built-up. Esta carta vai ser utilizada com referência. Para esta comparação foram escolhidas duas zonas dentro de região de interesse, uma zona urbana e uma zona rural. A Figura 7.22 apresenta a comparação da carta GHSL build-up com a classificação do grupo 1 sobre Tomar. A carta gerada possui estruturas mais finas como estradas principais e alguns aglomerados populacionais mais pequenos. É de notar também que a carta gerada define melhor os limites reais dos corpos de água. Apesar das diferenças enumeradas anteriormente, ambas as cartas têm uma semelhança elevada. Olhando agora para a Figura 7.23, esta apresenta a mesma comparação mas desta vez em ambiente rural, onde as diferenças são

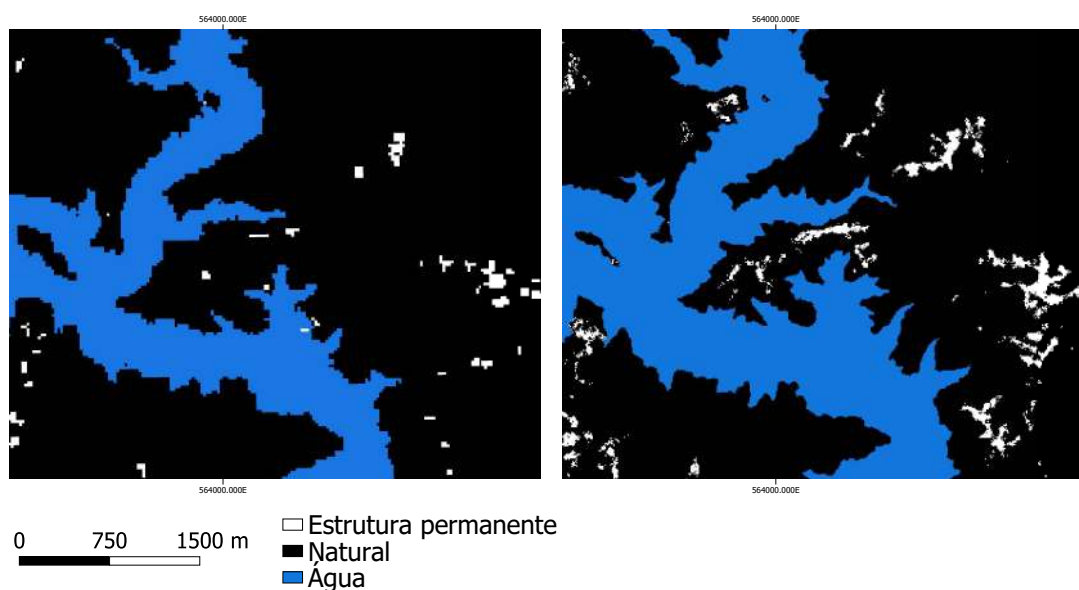


Figura 7.23: Comparação da carta GHSL, à esquerda, com a carta gerada, à direita. Imagens sobre a região do Rio Zêzere.

mais evidentes. A carta gerada tem uma maior definição nos aglomerados e também nos limites do rio, a utilização de dados de alta resolução dão aos resultados uma vantagem técnica, o que permite a deteção destas estruturas mais finas.

Finalmente, o último objetivo desta secção é a análise do erro em si. Utilizando as mesmas regiões mencionadas anteriormente, foram criadas imagens onde se pode observar os erros de omissão e de comissão em relação à carta COS 2015. Começando pela zona urbana, apresentada na Figura 7.24, esta figura apresenta o erro da classificação do grupo 1 e do grupo 2, em geral pode ser observado que o erro está condensado nas periferias das estruturas. Nestas zonas existe uma transição entre classes. A resolução utilizada torna a diferenciação das classes em zonas limites mais difícil resultando em maioritariamente erros de omissão.

Quando aos erros de comissão, é interessante observar que estão situados fora do centro da cidade. Estes erros correspondem a pequenos aglomerados não presentes nos dados de referência, mas os mesmos foram detetados pelo algoritmo. Este tipo de erro baixa a precisão das métricas do algoritmo mas é algo benéfico, visto que prova a possibilidade da deteção de estruturas não presentes nas cartas de referência. Este fenómeno é observável na Figura 7.25, onde estão representadas quatro áreas, a vermelho, em que se observa a deteção de estruturas reais não presentes nos dados de referência.



### 7.3. VALIDAÇÃO DOS MODELOS E ANÁLISE DO SEU ERRO

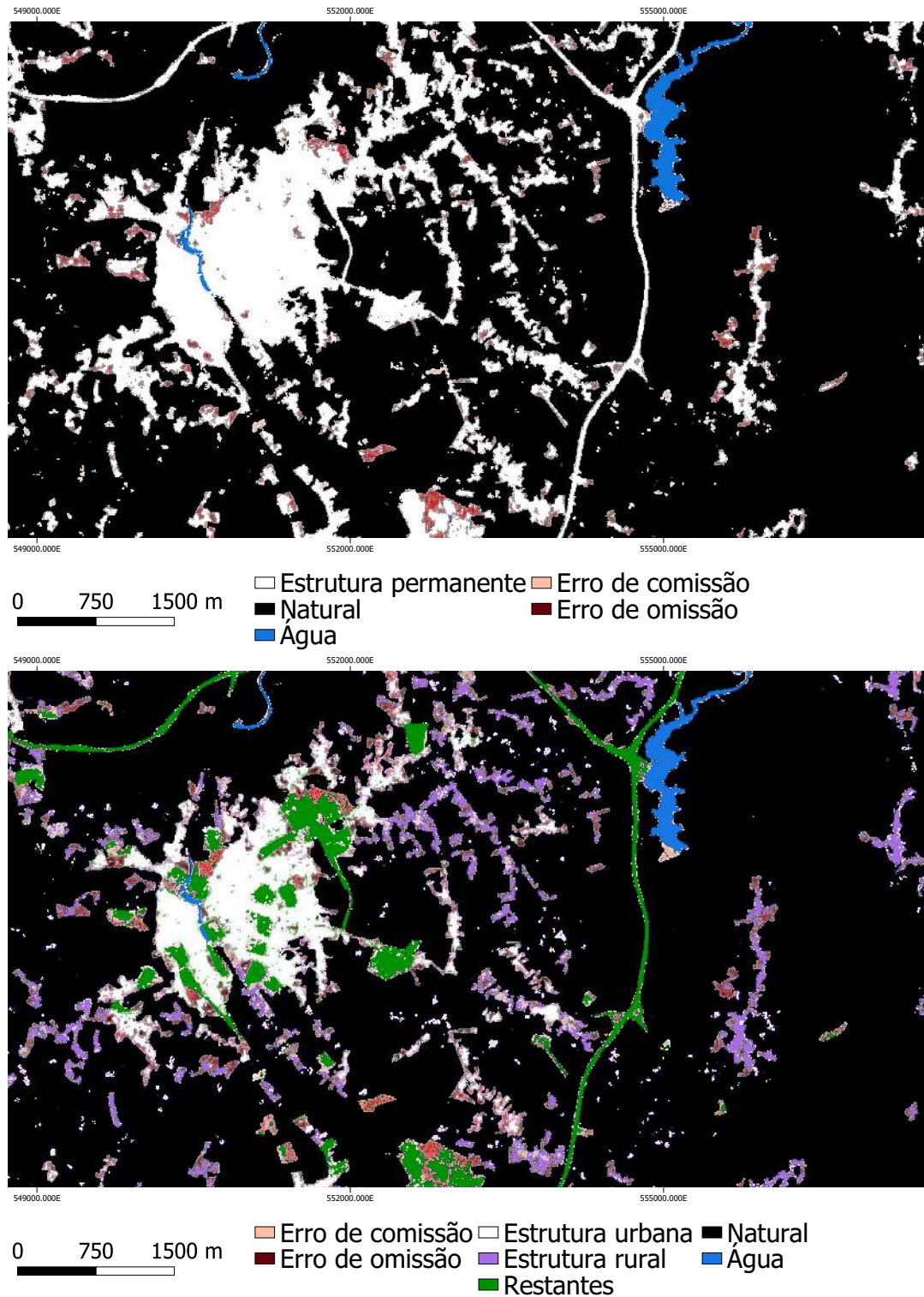


Figura 7.24: Erros detetados sobrepostos à classificação do grupo 1 (em cima) e do grupo 2 (em baixo). Imagem sobre a zona de Tomar.

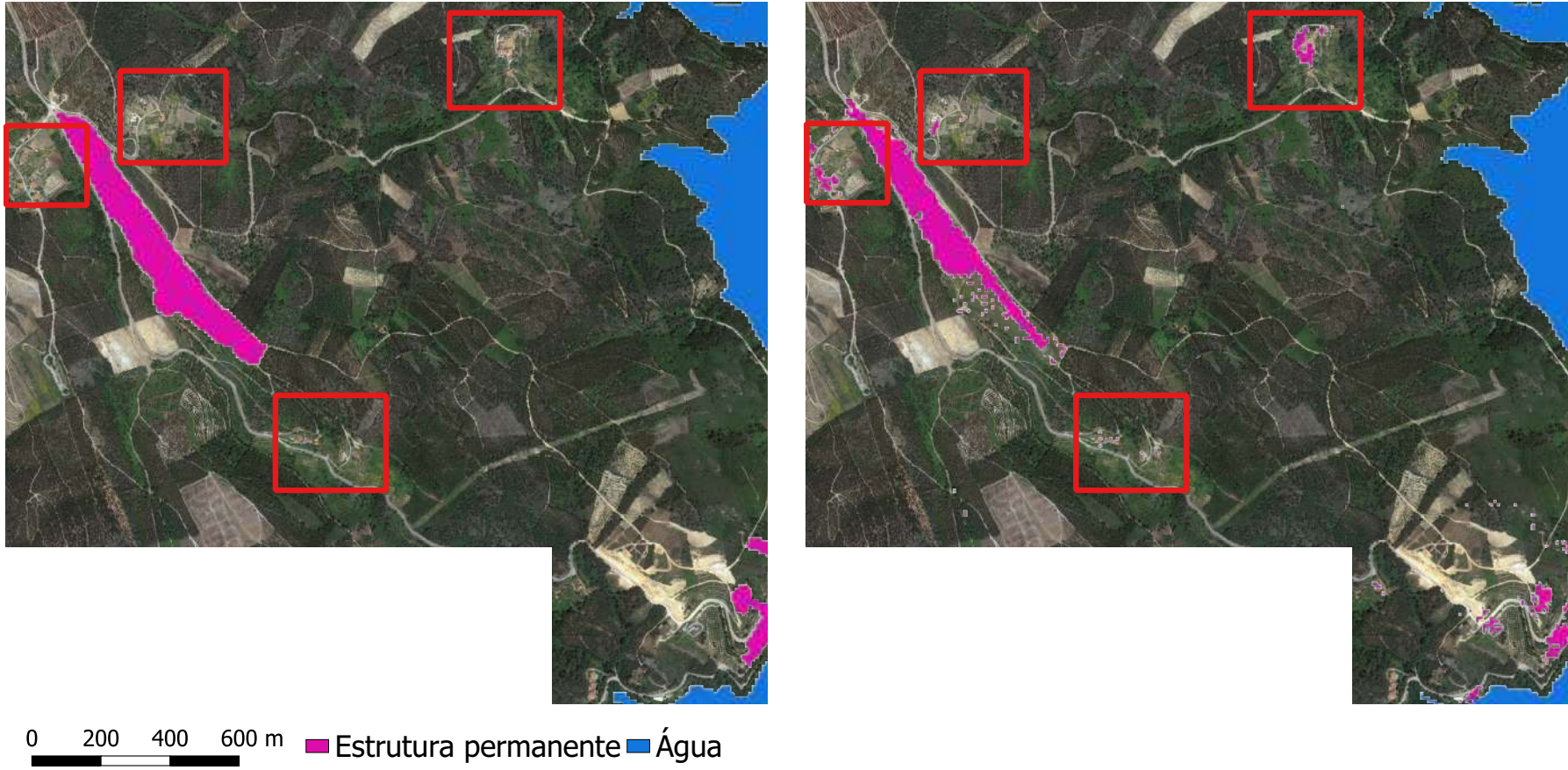


Figura 7.25: Estruturas permanentes detetadas não presentes nos dados de referência; Carta COS 2015 à esquerda, classificação XGBoost à direita.

Ao comparar o erro da classificação do grupo 1 com o do grupo 2, o último possui mais erros nos limites das estruturas detetadas. O algoritmo tem mais dificuldade em diferenciar os diferentes tipos de estruturas, visto que estas possuem um comportamento espectral semelhante, por exemplo é mais fácil para o algoritmo a diferenciação de uma estrutura urbana de vegetação, do que uma estrutura urbana de uma rural.

O erro em zonas rurais apresenta um comportamento semelhante ao apresentado na zona urbana, sendo o mesmo mais presente nos limites exteriores das estruturas e na diferenciação entre as estruturas do grupo 2. Este erro pode ser observado na Figura 7.26.

Concluindo, esta secção mostra que os resultados do melhor algoritmo não só obtêm métricas elevadas quando comparados aos dados de referência, mas também superam as cartas mais recentes com o mesmo âmbito: a deteção de estruturas artificiais permanentes.

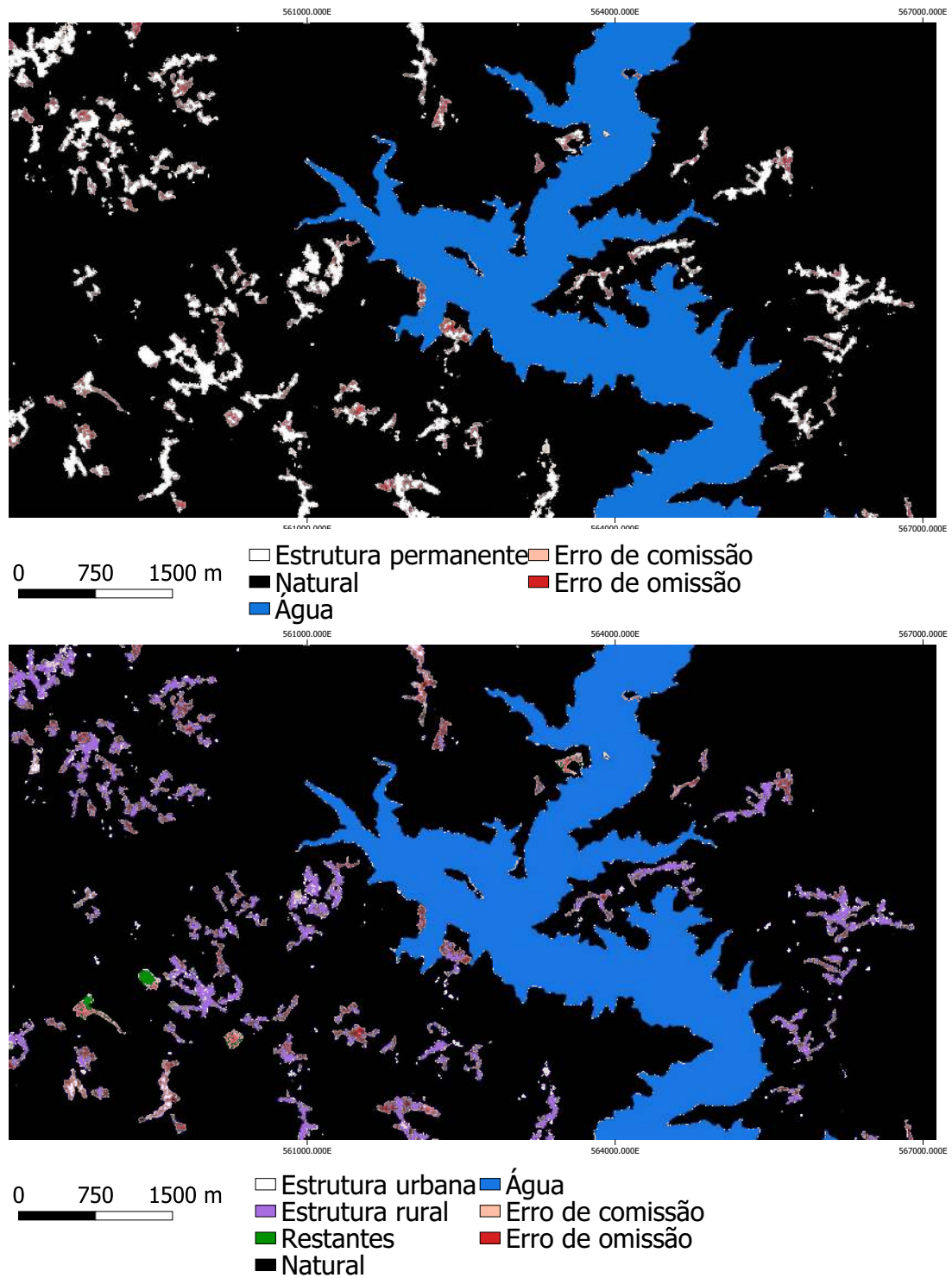


Figura 7.26: Erros detetados sobrepostos à classificação do grupo 1 (em cima) e do grupo 2 (em baixo). Imagem sobre o Rio Zêzere.

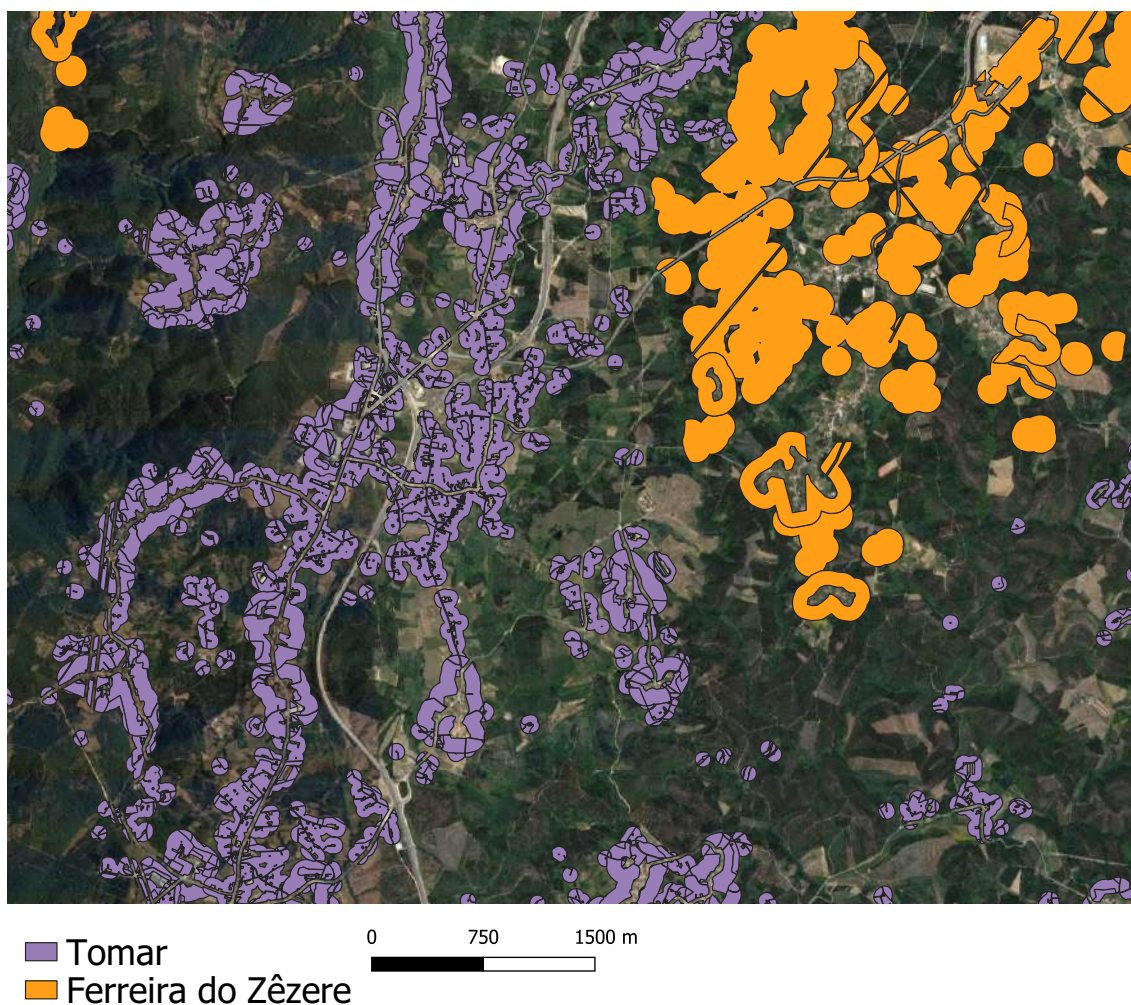


Figura 7.27: FGC oficiais de Tomar e Ferreira do Zêzere.

## 7.4 Caso de estudo: Faixas de gestão de combustível

Com os resultados previamente apresentados, e as suas cartas de estruturas artificiais permanentes geradas, é possível estudar a sua aplicação a um caso de uso real: a delimitação de FGC. Nesta secção são apresentados os resultados dessa aplicação, utilizando as cartas resultantes das classificações anteriores aplicadas à geração automática das FGC.

A Figura 7.27, apresenta as faixas de gestão de combustível oficiais do município de Tomar e Ferreira do Zêzere. Em roxo podem ser visualizadas as FGC de Tomar, este município apresenta faixas mais compactas, envolvendo as estruturas permanentes. O município de Ferreira do Zêzere, representado a laranja, apresenta faixas maiores. Estas faixas possuem nem sempre envolvem aglomerados populacionais na sua totalidade, visto que o interior desses aglomerados já estará posteriormente protegido pelas faixas exteriores. Ambos os municípios apresentam diferenças significativas na sua definição das faixas, isto dependerá das legislações dos próprios ou da metodologia utilizada para a definição das suas FGC oficiais.

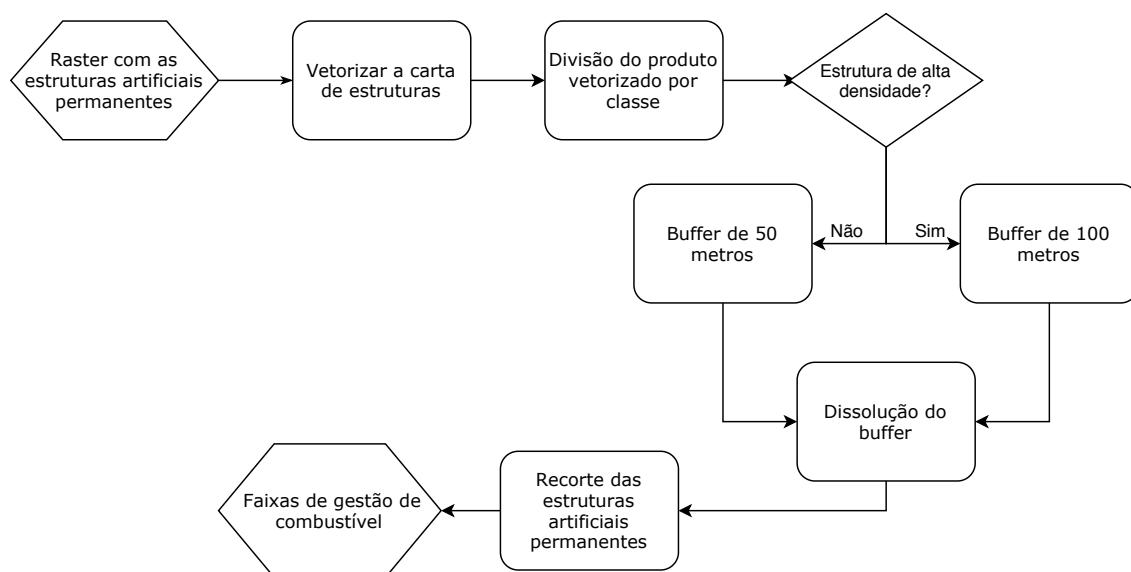


Figura 7.28: Cadeia de processamento para a criação das Faixas de gestão de combustível

O produto final é obtido através da seguinte cadeia de processamento representada na figura 7.28. Resultando no produto presente na Figura 7.29. Comparando agora diretamente as FGC geradas com as oficiais, presente na Figura 7.30, é possível retirar duas conclusões principais: as estruturas artificiais permanentes, geradas pelo melhor método de classificação preenchem quase perfeitamente as FGC oficiais de Tomar e estão contidas entre as FGC de Ferreira do Zêzere, o que indica uma robustez dos resultados obtidos. Finalmente, as FGC geradas, quando sobrepostas às faixas oficiais dos dois municípios, cobrem quase toda a sua extensão, possuindo maioritariamente erros de omissão das faixas oficiais junto a habitações. Como estas estruturas não estão presentes no resultado do classificação dos modelos, que é utilizada para a geração das FGC, não existe faixas no produto final gerado envolvendo as mesmas.

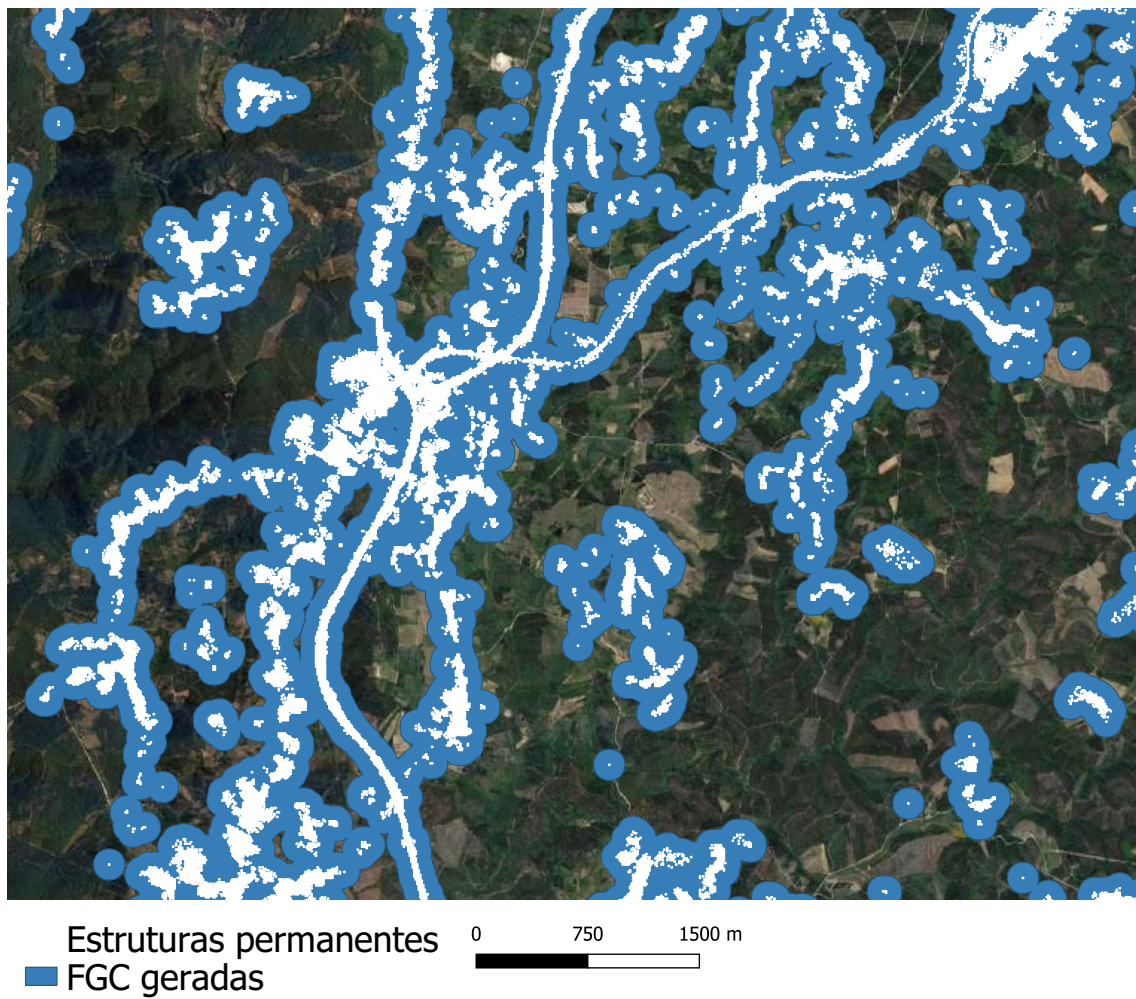


Figura 7.29: FGC geradas com base no produto final da classificação de estruturas permanentes.

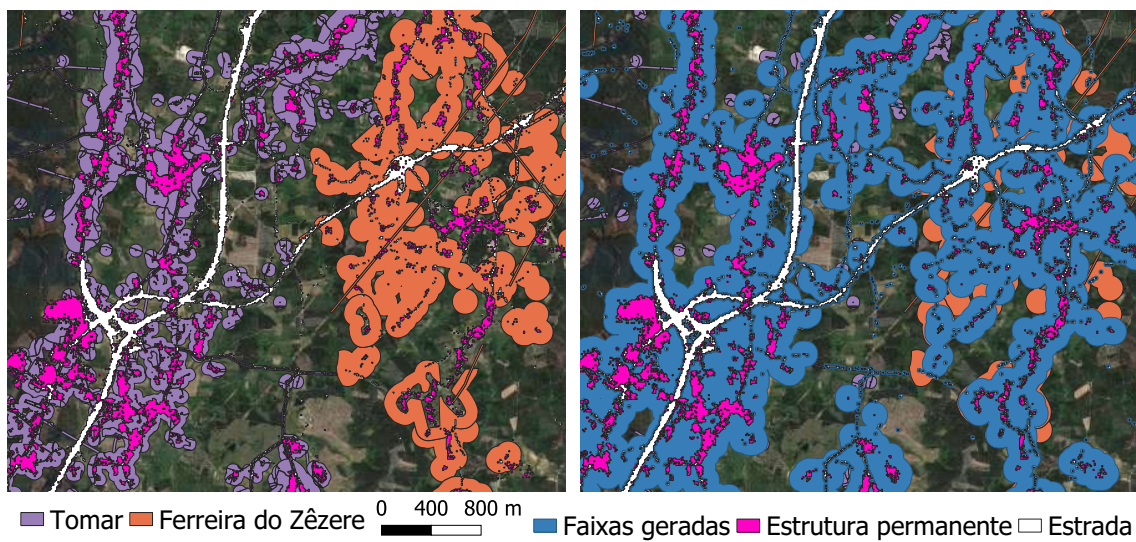


Figura 7.30: FGC oficiais de Tomar e Ferreira do Zêzere, à esquerda; FGC definidas a partir da classificação, à direita. Fonte: Neves et al. 2019 [47].

## 7.5 Resultados da redução dimensional dos dados de treino

Após a obtenção dos resultados da fase de treino, e como consequência o melhor algoritmo e melhor conjunto de dados, é interessante estudar o impacto da redução dimensional dos dados de treino na performance do algoritmo. Para esse efeito foi escolhido o melhor algoritmo, XGBoost e o melhor conjunto de dados, as séries temporais, complementadas com dados de elevação de terreno. Nesta secção é apresentado o resultado dos testes efetuados ao módulo de redução de dimensionalidade de dados, fselector, apresentado na secção 4.4.1.2. O modo “Lasso” não foi incluído pois a otimização do modo “elasticnet” já inclui o mesmo, como é explicado na secção 2.3. A redução de dimensionalidade foi efetuada em dados das séries temporais, utilizando 20% dos pixels da região de testes. Esta estruturação totaliza 143 atributos e possui o maior número de informação possível para resolver os problemas de classificação.

Para o modo “importances”, os valores das importâncias dos atributos foram extraídos do classificador XGBoost. Foi preciso testar vários valores para o limite superior da soma de todas das importâncias dos atributos, para isso foi selecionada a seguinte lista de valores: [0,5; 0,6; 0,7; 0,8; 0,9; 0,95; 0,99]. A Figura 7.31 mostra o valor do Kappa e F1 score para cada valor de importâncias, sendo as importâncias ordenados por ordem crescente. Os resultados mostram que o limite 0,8 obtém os melhores resultados em ambas as métricas avaliadas, com um Kappa de 0,84 e um F1 score de 0,92, reduzindo os 143 atributos para 53.

Organizando agora os resultados por número de atributos, é obtida a Figura 7.32, este gráfico apresenta os valores de F1 score e tempo de treino por número de atributos reduzidos. Nela está marcado o ponto ótimo, 53 atributos com 0,92 de F1 score e tempo de treino de 787 segundos. Como se pode observar na figura, o tempo de treino do algoritmo sofre uma descida de 47,5%, de 1500, ao utilizar todos os atributos, para 787 segundos no melhor caso. A partir deste ponto adicionar mais atributos apenas aumenta a complexidade dos dados e o tempo de treino, descendo o valor do F1 score ligeiramente.



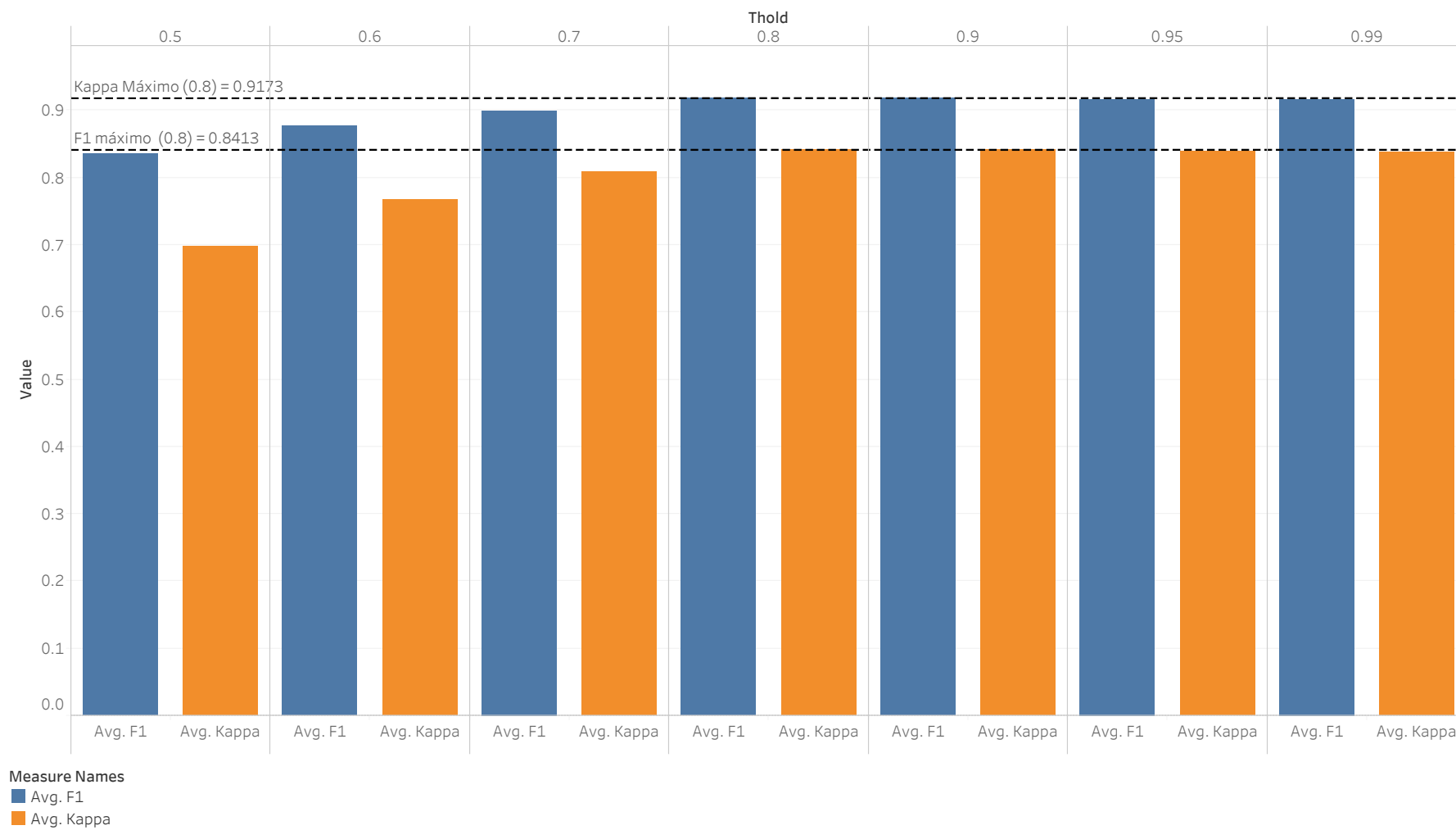


Figura 7.31: Métricas de avaliação do método “importances” para cada valor total de importâncias selecionado

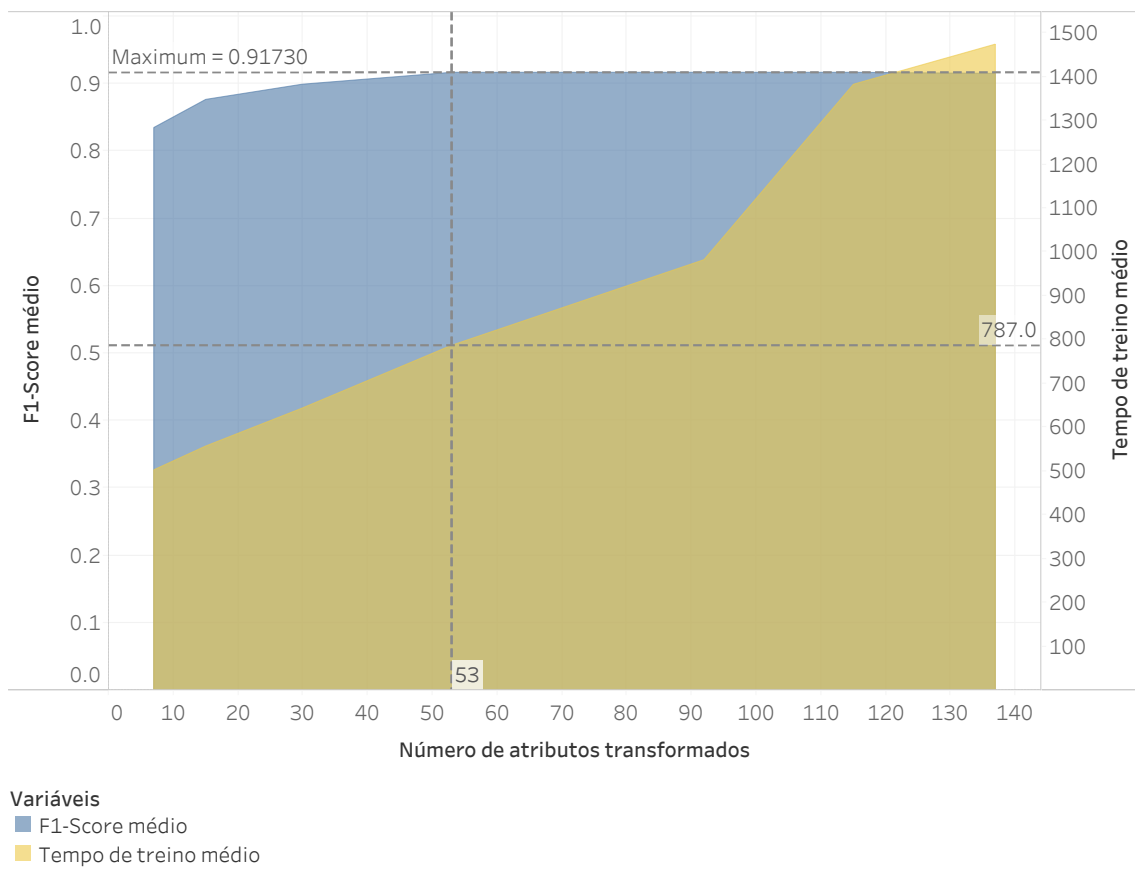


Figura 7.32: Impacto da do número de atributos na performance do modo “importances”.

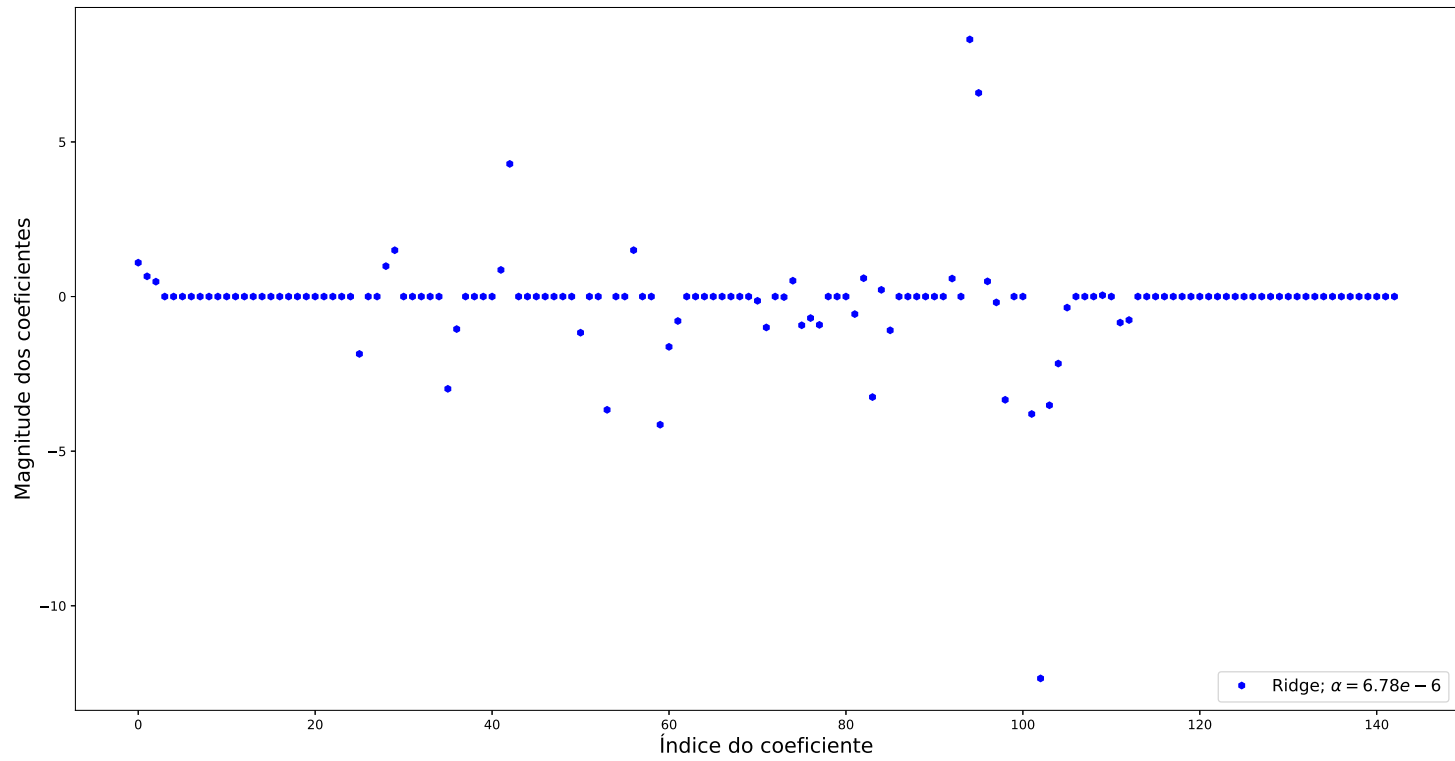


Figura 7.33: Magnitude dos coeficientes resultantes da regressão Lasso por atributo

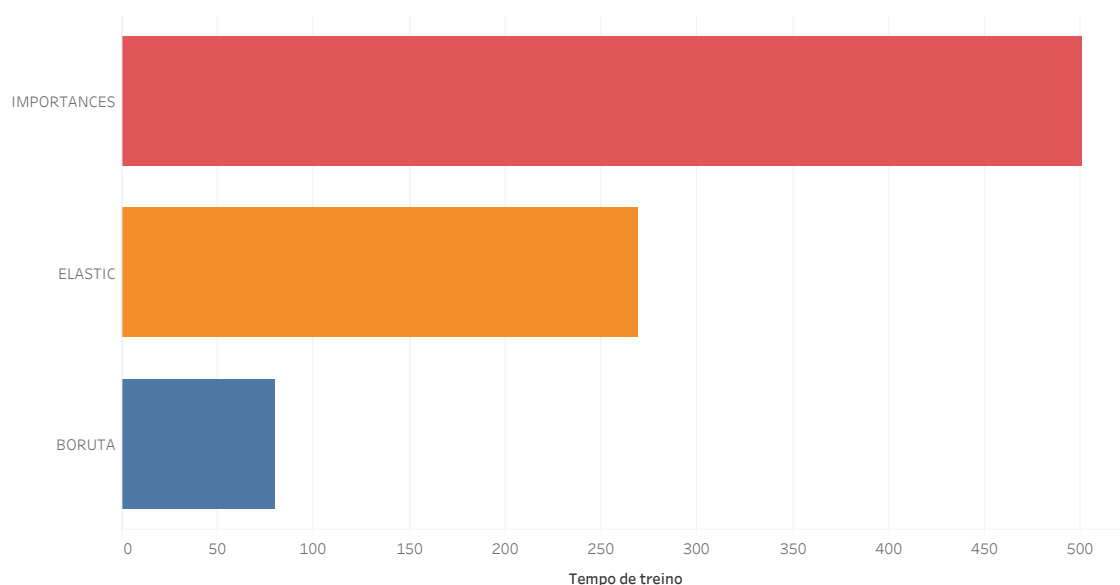


Figura 7.34: Tempos de treino, em segundos, por técnica de redução dimensional

O modo “elasticnet” efetua automaticamente a otimização do mesmo, tentando achar os melhores parâmetros para o  $\alpha$ ,  $L1$  ratio e o número de iterações do algoritmo utilizando *3-fold Cross validation*. Como resultado, foi obtido um  $\alpha$  de  $6,7783e-06$  e um  $L1$  ratio 1,0. Essencialmente a melhor redução utilizando este método é uma regressão Lasso com um  $\alpha$  de  $6,7783e-06$ . Após a descoberta dos melhores valores é corrido então a regressão Lasso, extraídos os coeficientes e é criado um transformador com um valor limite de  $1e-5$ , este valor representa apenas selecionar coeficientes com valor diferente a 0. Os resultados da redução, apresentados na Figura 7.33, mostram os valores dos coeficientes, sendo descartados os coeficientes com valor 0. Esta redução obtém um Kappa de 0,79 e um F1 score de 0,89, reduzindo de 143 atributos para 42. Os atributos utilizados nesta redução, e os seus índices, estão presentes na Tabela D.3 do apêndice D.

Finalmente, o modo “boruta” é o método mais automatizado para a redução de dimensionalidade. Este algoritmo apenas requer um classificador *ensemble*. Após correr o algoritmo, o número de atributos é reduzido de 142 atributos para 30, obtendo um Kappa de 0,79 e um F1 score de 0,89.

Concluindo, o modo “importances” obteve os melhores resultados com uma soma total de importâncias de 0,80, o que reduz a série temporal de 143 atributos para 53 com um ligeiro aumento de performance. Este modo reduz também o tempo de treino de 1500 para 787 segundos. A Figura 7.34 mostra que apesar deste modo ter a melhor performance os outros algoritmos são capazes de reduzir ainda mais o tempo de treino, tendo isto um ligeiro impacto negativo na sua performance.

Tabela 7.2: Métricas de teste finais (grupo 1) para o melhor classificador: XGBoost

	Precision	Recall	F1 score
Estrutura artificial permanente	0,86	0,72	0,78
Restante	0,98	0,99	0,98
Água	0,97	0,95	0,96
Kappa	0,83		

Tabela 7.3: Métricas de teste finais (grupo 2) para o melhor classificador: XGBoost

	Precision	Recall	F1 score
Estrutura artificial permanente	0,73	0,66	0,69
Estrada	0,65	0,13	0,22
Restante	0,95	0,99	0,97
Água	0,97	0,95	0,96
Kappa	0,63		

Tabela 7.4: Métricas de teste finais (grupo 3) para o melhor classificador: XGBoost

	Precision	Recall	F1 score
Habitações Urbanas	0,87	0,65	0,75
Habitações Rurais	0,75	0,60	0,66
Outras estruturas	0,91	0,73	0,71
Restante	0,98	0,99	0,99
Água	0,97	0,95	0,96
Kappa	0,82		

## 7.6 Confirmação dos resultados do melhor modelo

Nas secções anteriores foi concluído que o melhor modelo de classificação é o XGBoost. Dado esta escolha existe agora um viés nos dados, sendo necessário classificar um novo conjunto de dados independente dos utilizados. Para isso foi deixado de parte, desde o início da dissertação, um conjunto independente com 20% do tamanho da região interesse. Este conjunto tem o objetivo de validar os resultados obtidos e eliminar o viés da escolha do melhor modelo. As Tabelas 7.2, 7.3 e 7.4 mostram as métricas finais do classificador ótimo, o XGBoost. Observa-se que o comportamento deste modelo mantém-se e o mesmo generaliza bem para novas amostras não observadas durante a cadeia de processamento do treino, nem durante a escolha do modelo. Este teste é a fase final do treino e análise dos modelos, e confirma a performance do melhor classificador escolhido. A partir deste ponto não se pode fazer mais escolhas sobre os parâmetros do modelo analisar outros modelos com os mesmos dados, caso isso seja necessário teria de se deixar de parte outro conjunto de dados independente.

## 7.7 Discussão

Em geral, observou-se que o algoritmo Gradient Tree Boosting, com a implementação XGBoost, obteve os melhores resultados com todas as métricas utilizadas. Para além disso, este algoritmo mostra não só os melhores resultados mas também apresenta pouca variação dos mesmos ao classificar tarefas mais complexas, como a diferenciação entre estruturas permanentes. Em paralelo ao melhor algoritmo, o melhor conjunto de dados de treino é a combinação de séries temporais com modelos de elevação de terreno.

Os testes mostram uma limitação na metodologia utilizada: a deteção de estradas. Esta é limitada pela resolução, estradas abaixo das primárias e autoestradas apresentam-se demasiado segmentadas impossibilitando a sua utilização. Consequentemente, é mais benéfico a remoção das estradas em prol de melhorar os resultados das outras estruturas a classificar, visto que as estradas são bem mapeadas em projetos como o [OSM](#). Mesmo assim será de interesse estudar em mais detalhe a deteção de estradas.

Dando ênfase agora ao caso de estudo da delimitação das [FGC](#), os resultados mostraram que a utilização de cartas resultantes da metodologia tem a capacidade de mapear quase perfeitamente as [FGC](#) oficiais, promovendo a proteção de estruturas artificiais permanentes. Será também possível a geração das faixas para qualquer altura do ano não sendo necessário esperar pela disponibilização das [FGC](#) oficiais, acelerando então o processo da análise pretendida. Finalmente e bastante importante, existe a possibilidade da deteção de zonas sem proteção adequada, ou seja, é possível comparar as faixas geradas com as oficiais e assim proceder à deteção de aglomerados ou até edifícios singulares em que a sua faixa não foi cartografada no ficheiro oficial.

Terminam-se os testes feitos neste capítulo com a análise da redução dimensional dos dados. Esta análise foi feita para ver o impacto da utilização de várias técnicas, como a utilização da regressão [LASSO](#), o algoritmo *Boruta*, a regressão *ElasticNet* e uma implementação utilizando *Feature Importances*. Concluindo, o método de limite superior no valor da soma das importâncias implementado reduz bastante os atributos dos classificadores com um ligeiro aumento da performance.

Em suma, metodologias com aprendizagem automática possuem vários benefícios para o estudo da ocupação de solo, nomeadamente, é possível combinar dados provenientes de fontes de deteção remota para a obtenção mais rápida de cartas atualizadas, que podem ser posteriormente utilizadas para casos de uso reais, como a delimitação das [FGC](#), criação de máscaras de estruturas permanentes, derivar mapas de risco, como potenciais danos de desastres naturais e muitas outras aplicações. Isto tudo com uma resolução e detalhe acrescido, quando comparado com cartas existentes, e sem esperar pela atualização dessas mesmas cartas.

## CONCLUSÃO

Para a obtenção de cartas de ocupação de solo robustas em tempo aceitável é necessário a integração da cadeia de geração das mesmas com metodologias que utilizem aprendizagem automática. Nesta tese foi investigado o potencial da utilização de dados provenientes de fontes de detecção remota, principalmente dados das constelações Sentinel-1 e Sentinel-2, para melhorar a identificação de estruturas artificiais permanentes em Portugal continental. Como consequência deste objetivo principal são derivados outros objetivos secundários de modo a responder às seguintes perguntas:

- Qual o melhor classificador para a detecção de estruturas?
- Qual o melhor conjunto de dados de treino?
- Caso de estudo: Qual o impacto da utilização dos resultados da metodologia apresentada na delimitação das FGC?

As primeiras duas perguntas estão interligadas, pois para testar qual o melhor algoritmo são precisos dados de treino, ora isto introduz uma nova variável aos testes: o tipo de dados. Dado isto, não se pode escolher o melhor algoritmo sem antes analisar o impacto dos dados de treino no mesmo. Após todos os testes efetuados foi concluído que o algoritmo de aprendizagem *Gradient Tree Boosting*, com a implementação da biblioteca *XGBoost*, apresenta os melhores resultados. Este modelo para além dos melhores resultados também é o mais robusto para todas as tarefas de classificação, sendo a sua variância de resultados mais baixa que todos os outros modelos.

Em geral, respondendo ao objetivo de qual o melhor conjunto de dados de treino para a identificação de estruturas artificiais permanentes, a utilização de dados provenientes de séries temporais têm métricas e resultados finais superiores, quando comparados com a utilização de imagens de uma só data. A utilização de dados deste tipo promovem

a robustez das metodologias onde são utilizados, visto que uma série temporal é mais resistente a anomalias nas imagens, pois em vez de utilizar uma imagem só agrega várias estatísticas temporais de um certo intervalo de tempo.

Apesar da solução final proposta apresentar bons resultados esta tem as seguintes limitações:

- **Limite da resolução dos dados:** Ambos os dados de referência, as classes, e os dados de treino limitam a metodologia em termos de performance. Sem dados de referência robustos, ou seja dados que mapeiam estruturas artificiais permanentes ao nível do edifício singular, não será possível ter uma boa avaliação dos modelos em termos de métricas, necessitando de uma interpretação manual dos resultados. Em paralelo, a utilização de dados com uma resolução de 10 por 10 metros apenas garante a deteção robusta de estruturas com pelo menos o dobro do tamanho.
- **Baixa performance na deteção de estradas:** Como foi discutido no capítulo anterior, a metodologia apresentada deteta estradas robustamente. Estradas mais estreitas, ou seja tudo o que seja mais pequeno que autoestradas e estradas principais, possuem erros de omissão demasiado altos para um bom mapeamento das mesmas.

Concluindo, para uma melhor deteção de estruturas artificiais permanentes é recomendado a utilização de séries temporais complementadas com modelos de elevação de terreno. Como algoritmo de classificação é recomendado a utilização do algoritmo *Gradient Tree Boosting*, pelos seus excelentes e robustos resultados. Mesmo assim, este classificador apresenta bons resultados também com dados estáticos, caso a complexidade temporal do seu treino e do processamento de dados seja muito elevada para a aplicação em questão. A utilização deste tipo de dados consegue produzir bons resultados quando comparado com cartas já existentes.

O ultimo objetivo desta dissertação, a aplicação dos resultados obtidos do algoritmo *XGBoost* à delimitação das *FGC*, a carta resultante do melhor classificador mapeia as *FGC* oficiais com alta fiabilidade. Apresentando assim grande potencial para a sua integração no processo de delimitação das mesmas, podendo acelerar a criação das faixas e detetar zonas sem proteção adequada.

### 8.1 Contribuições

Durante toda a dissertação foram estudadas e desenvolvidas várias metodologias. Estas aplicam algoritmos de aprendizagem que visam a resolver os problemas presentes nas cartas de estruturas permanentes existentes, em Portugal Continental. Os resultados obtidos são uma adição importante a todos os estudos que visam a utilizar este tipo de cartas.

No final desta dissertação foram alcançadas as seguintes contribuições:



- Algoritmo semi-automático, que gera modelos para classificação de estruturas artificiais permanentes. Complementado com uma cadeia de processamento de dados, otimização de modelos e redução dimensional de atributos, esta metodologia de classificação é escalável para Portugal Continental, podendo resultar num único modelo para todo o território que consiga criar uma carta de estruturas permanentes para qualquer altura do ano.
- Definição de duas metodologias: dados estáticos e séries temporais, que mostram resultados melhorados quando comparados com o estado da arte.
- Proposta do melhor modelo e do melhor conjunto de dados, dentro do universo estudado, para classificação de estruturas artificiais permanentes;
- Estudo do impacto de cartas de estruturas artificiais permanentes para aceleração da delimitação das FGC, não sendo necessário esperar pelas faixas oficiais para uma análise atempada. É também concluído que a utilização das cartas de estruturas permanentes permite criar um processo de deteção de zonas sem proteção adequada.
- Artigo publicado na conferência: INForum 2019<sup>1</sup> - Neves et al. 2019 [47]. O artigo em questão visa a apresentação dos resultados intermédios, obtidos pelo treino do algoritmo *Gradient Tree Boosting* com séries temporais Sentinel-1 e Sentinel-2 apenas, utilizando a carta COS 2015 como dados de referência com estradas OSM sobrepostas.

## 8.2 Trabalhos futuros

Apesar da metodologia utilizada apresentar resultados bastante interessantes, existem métodos que não foram explorados durante a dissertação mas que têm o potencial de melhorar os resultados apresentados. Em particular, o estudo apresentado foi restrito à análise por pixels com dados de deteção remota das constelações Sentinel-1 e Sentinel-2, complementados com dados adicionais. A região de interesse foi também restrita devido aos recursos computacionais da máquina de teste. Assim, com o aumento dos recursos disponíveis, um trabalho futuro resultante do concluído nesta dissertação seria então a criação de um novo produto a nível de Portugal Continental, ao escalar a solução proposta. Este produto cobrirá todo o território representado as estruturas artificiais permanentes com, para isso será necessário escalar a metodologia final.

A utilização de redes neuronais artificiais é algo a refinar, com estruturas de redes mais complexas será possível o melhoramento dos resultados obtidos por estas redes, utilizando redes neuronais com arquiteturas mais complexas ou redes neuronais convolucionais. As *Convolutional neural networks (CNNs)* apresentam resultados estado da

---

<sup>1</sup>Ata INForum 2019 - visitado a 06/07/2019 <http://inforum.org.pt/INForum2019/docs/atas-do-inforum2019>

arte em várias áreas classificação de imagens. Este tipo de redes mostra potencial quando combinadas com dados de séries temporais [41], investigação futura deveria estudar o potencial da transferências destas metodologias para a deteção de estruturas artificiais permanentes. Utilizando estas técnicas será de interesse ver o seu impacto na deteção de estradas [81]. Com a utilização deste tipo de modelos vem a utilização de outro tipo de estruturação de dados, o conjunto de dados EuroSAT é algo a investigar pois consegue atingir performances elevadas quando combinados com CNNs [33].

Um dos grandes problemas abordados durante a tese foi a ausência de dados de referência, no que toca ao mapeamento de estruturas artificiais permanentes não existindo nenhuma carta que represente as mesmas até à estrutura individual. Uma área a explorar seria a utilização de uma metodologia que incorpore *crowdsourcing* com a deteção de estruturas ao nível de Portugal Continental, e dividir o mesmo em várias zonas e dividir tarefas entre membros exteriores à investigação. Este trabalho tem o maior interesse de todos os apresentados, pois o seu resultado seria uma das melhores cartas existentes até à data que, quando incorporada com algoritmos de aprendizagem automática, tem o potencial de impulsionar estudos na área de deteção remota.

## BIBLIOGRAFIA

- [1] S. Abdikan, F. B. Sanli, M. Ustuner e F. Calò. “Land cover mapping using sentinel-1 SAR data”. Em: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 41.July (2016), pp. 757–761. ISSN: 16821750. DOI: [10.5194/isprsarchives-XLI-B7-757-2016](https://doi.org/10.5194/isprsarchives-XLI-B7-757-2016).
- [2] N. S. d. O. Andrade, A. N. d. C. Santa Rosa e P. C. d. C. Faria. “Fundamentos de Polarimetria SAR”. Em: *Anais XIII Simpósio Brasileiro de Sensoriamento Remoto Simpósio Brasileiro de Sensoriamento Remoto* (2007), pp. 4775–4782.
- [3] A. Arauzo-Azofra, J. L. Aznarte e J. M. Benítez. “Empirical study of feature selection methods based on individual feature evaluation for classification problems”. Em: *Expert Systems with Applications* 38.7 (2011), pp. 8170–8177. ISSN: 09574174. DOI: [10.1016/j.eswa.2010.12.160](https://doi.org/10.1016/j.eswa.2010.12.160). URL: <http://dx.doi.org/10.1016/j.eswa.2010.12.160>.
- [4] A. Ather. “A quality analysis of openstreetmap data”. Em: *ME Thesis, University College London* May (2009), p. 81. URL: <ftp://s-she-junco.cits.rncan.gc.ca/pub/cartonat/Reference/VGI/Dissertation-OpenStreepMap-Quality-Aather-2009.pdf>.
- [5] H. Balzter, B. Cole, C. Thiel e C. Schmullius. “Mapping CORINE land cover from Sentinel-1A SAR and SRTM digital elevation model data using random forests”. Em: *Remote Sensing* 7.11 (2015), pp. 14876–14898. ISSN: 20724292. DOI: [10.3390/rs71114876](https://doi.org/10.3390/rs71114876).
- [6] C. Barrington-Leigh e A. Millard-Ball. “The world’s user-generated road map is more than 80% complete”. Em: *PLOS ONE* 12.8 (ago. de 2017), pp. 1–20. DOI: [10.1371/journal.pone.0180698](https://doi.org/10.1371/journal.pone.0180698). URL: <https://doi.org/10.1371/journal.pone.0180698>.
- [7] M. Belgiu e L. Drăgu. “Random forest in remote sensing: A review of applications and future directions”. Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (2016), pp. 24–31. ISSN: 09242716. DOI: [10.1016/j.isprsjprs.2016.01.011](https://doi.org/10.1016/j.isprsjprs.2016.01.011).
- [8] F Bertini, O Brand, S Carlier, U. Del Bello, M. Drusch, R Duca, V Fernandez, C Ferrario, M. H. Ferreira, C Isola, V Kirschner, P Laberinti, M Lambert, G Mandorlo, P Marcos, P. Martimort, S Moon, P Oldeman, M Palomba e J Pineiro. *Sentinel-2*

- ESA's Optical High-Resolution Mission for GMES Operational Services*. Vol. SP-1322. March. 2012. ISBN: 9789290922711. DOI: [10.1016/j.rse.2011.11.026](https://doi.org/10.1016/j.rse.2011.11.026).
- [9] L. Breiman. "Out-of-bag estimation". Em: ().
- [10] L. Breiman. "Random Forests". Em: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324). URL: <https://doi.org/10.1023/A:1010933404324>.
- [11] L. Breiman e C. Berkeley. "Bias, variance, and arcing classifier". Em: April (1996).
- [12] L. Breiman, J. H. Friedman, R. A. Olshen e C. J. Stone. "Classification and regression trees". Em: *Classification and Regression Trees* 1. February (2017), pp. 1–358. ISSN: 19424787. DOI: [10.1201/9781315139470](https://doi.org/10.1201/9781315139470). arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- [13] M. Caetano, A. Nunes, J. Dinis, M. Pereira, P. Marrecas e V. Nunes. "Carta de Uso e Ocupação do Solo de Portugal Continental para 2007 (COS2007v2.0): memória descritiva". Em: (2010).
- [14] J. Cai, J. Luo, S. Wang e S. Yang. "Feature selection in machine learning: A new perspective". Em: *Neurocomputing* 300 (2018), pp. 70–79. ISSN: 18728286. DOI: [10.1016/j.neucom.2017.11.077](https://doi.org/10.1016/j.neucom.2017.11.077). URL: <https://doi.org/10.1016/j.neucom.2017.11.077>.
- [15] "Carta Administrativa Oficial e Portugal". Em: (2016), p. 3091.
- [16] T. Chen e C. Guestrin. "XGBoost: A Scalable Tree Boosting System". Em: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785). URL: <http://doi.acm.org/10.1145/2939672.2939785>.
- [17] X. L. Chen, H. M. Zhao, P. X. Li e Z. Y. Yin. "Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes". Em: *Remote Sensing of Environment* 104.2 (2006), pp. 133–146. ISSN: 00344257. DOI: [10.1016/j.rse.2005.11.016](https://doi.org/10.1016/j.rse.2005.11.016).
- [18] N. Chrysoulakis, C. Feigenwinter, D. Triantakonstantis, I. Penyevskiy, A. Tal, E. Parlow, G. Fleishman, S. Düzgün, T. Esch e M. Marconcini. "A Conceptual List of Indicators for Urban Planning and Management Based on Earth Observation". Em: *ISPRS International Journal of Geo-Information* 3.3 (2014), pp. 980–1002. ISSN: 2220-9964. DOI: [10.3390/ijgi3030980](https://doi.org/10.3390/ijgi3030980). URL: <http://www.mdpi.com/2220-9964/3/3/980>  
<http://www.mdpi.com/2220-9964/3/3/980/>.
- [19] N. Clerici, C. A. Valbuena Calderón e J. M. Posada. "Fusion of sentinel-1a and sentinel-2A data for land cover mapping: A case study in the lower Magdalena region, Colombia". Em: *Journal of Maps* 13.2 (2017), pp. 718–726. ISSN: 17445647. DOI: [10.1080/17445647.2017.1372316](https://doi.org/10.1080/17445647.2017.1372316).

- [20] C. Cortes e V. Vapnik. “Support-vector networks”. Em: *Assembly* 44.13 (1995), p. 97. ISSN: 10508171. DOI: 10.1111/j.1747-0285.2009.00840.x. arXiv: 1011.1669.
- [21] Direção-Geral do Território. “Especificações técnicas da Carta de Uso e Ocupação do Solo (COS) de Portugal Continental para 1995, 2007, 2010 e 2015. Relatório Técnico.” Em: (2018), p. 103. URL: <http://mapas.dgterritorio.pt/atom-dgt/pdf-cous/COS2015/ET-COS-1995-2007-2010-2015.pdf>.
- [22] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese e P. Bargellini. “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services”. Em: *Remote Sensing of Environment* 120 (2012), pp. 25–36. ISSN: 00344257. DOI: 10.1016/j.rse.2011.11.026. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0034425712000636>.
- [23] T. Fawcett. “An introduction to ROC analysis”. Em: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010. arXiv: [/dx.doi.org/10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010) [http: ].
- [24] V. I. Fedorov. “Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal”. Em: *Uspekhi Fizicheskikh Nauk* 29.4 (1998), pp. 49–54. ISSN: 03011798. DOI: 10.1007/978-3-319-14280-7.
- [25] S. Freire, T. Kemper, M. Pesaresi, A. Florczyk e V. Syrris. “Combining GHSL and GPW to improve global population mapping”. Em: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 2015, pp. 2541–2543. DOI: 10.1109/IGARSS.2015.7326329.
- [26] M. A. Friedl, D. K. McIver, J. C. Hodges, X. Y. Zhang, D. Muchoney, A. H. Strahler, C. E. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao e C. Schaaf. “Global land cover mapping from MODIS: Algorithms and early results”. Em: *Remote Sensing of Environment* 83.1-2 (2002), pp. 287–302. ISSN: 00344257. DOI: 10.1016/S0034-4257(02)00078-0. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0034425702000780>.
- [27] B.-c. Gao. “A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space”. Em: *Remote Sensing of Environment* 266.April (1996), pp. 257–266. ISSN: 00344257. DOI: 10.1016/S0034-4257(96)00067-3. arXiv: [bhmic00033](https://arxiv.org/abs/bhmic00033).
- [28] R. Goldblatt, W. You, G. Hanson e A. K. Khandelwal. “Detecting the Boundaries of Urban Areas in India : A Dataset for Pixel-Based Image Classification in Google Earth Engine”. Em: (2016). DOI: 10.3390/rs8080634.

- [29] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau e R. Moore. “Google Earth Engine: Planetary-scale geospatial analysis for everyone”. Em: *Remote Sensing of Environment* (2017). DOI: [10.1016/j.rse.2017.06.031](https://doi.org/10.1016/j.rse.2017.06.031). URL: <https://doi.org/10.1016/j.rse.2017.06.031>.
- [30] H. Guan e J. Li. “International Journal of Remote Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests”. Em: July (2013). DOI: [10.1080/01431161.2013.788261](https://doi.org/10.1080/01431161.2013.788261).
- [31] M. Hall-Beyer. “GLCM Texture: a Tutorial”. Em: *17th International Symposium on Ballistics* 2.March (2017), pp. 267–274.
- [32] R. M. Haralick, K. Shanmugam e I. Dinstein. *Textural Features for Image Classification*. 2007. DOI: [10.1109/tsmc.1973.4309314](https://doi.org/10.1109/tsmc.1973.4309314).
- [33] P. Helber, B. Bischke, A. Dengel e D. Borth. “Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification”. Em: *International Geoscience and Remote Sensing Symposium (IGARSS) 2018-July* (2018), pp. 204–207. DOI: [10.1109/IGARSS.2018.8519248](https://doi.org/10.1109/IGARSS.2018.8519248). arXiv: [arXiv:1709.00029v2](https://arxiv.org/abs/1709.00029v2).
- [34] A Huete, K Didan, T Miura, E. P. Rodriguez, X Gao e L. G. Ferreira. “Overview of the radiometric and biophysical performance of the MODIS vegetation indices”. Em: 83 (2002), pp. 195–213. ISSN: 00344257. DOI: [10.1016/S0034-4257\(02\)00096-2](https://doi.org/10.1016/S0034-4257(02)00096-2).
- [35] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin e I. Rodes. “Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series”. Em: *Remote Sensing* 9.1 (2017), p. 95. ISSN: 2072-4292. DOI: [10.3390/rs9010095](https://doi.org/10.3390/rs9010095). URL: <http://www.mdpi.com/2072-4292/9/1/95>.
- [36] R. H. James B. *Introduction to Remote Sensing. Fifth Edition*. Guilford Press, 2011, 1–58. ISBN: 978-1-60918-177-2.
- [37] E. M. Karabulut, S. A. Özel e T. İbrikçi. “A comparative study on the effect of feature selection on classification accuracy”. Em: *Procedia Technology* 1 (2012), pp. 323–327. ISSN: 22120173. DOI: [10.1016/j.protcy.2012.02.068](https://doi.org/10.1016/j.protcy.2012.02.068).
- [38] T Kavzoglu e P. M. Mather. “The use of backpropagating artificial neural networks in land cover classification”. Em: 1161.September (2015). DOI: [10.1080/0143116031000114851](https://doi.org/10.1080/0143116031000114851).
- [39] M. B. Kursu, A. Jankowski e W. R. Rudnicki. “Boruta - A system for feature selection”. Em: *Fundamenta Informaticae* 101.4 (2010), pp. 271–285. ISSN: 01692968. DOI: [10.3233/FI-2010-288](https://doi.org/10.3233/FI-2010-288).
- [40] N. Kussul, M. Lavreniuk, S. Skakun e A. Shelestov. “Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data”. Em: *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017), pp. 778–782. ISSN: 1545598X. DOI: [10.1109/LGRS.2017.2681128](https://doi.org/10.1109/LGRS.2017.2681128).

- [41] N. Kussul, M. Lavreniuk, S. Skakun e A. Shelestov. “Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data”. Em: *a* 14.5 (2017), pp. 778–782. DOI: [10.1109/LGRS.2017.2681128](https://doi.org/10.1109/LGRS.2017.2681128).
- [42] A. Lefebvre, C. Sannier e T. Corpetti. “Monitoring urban areas with Sentinel-2A data: Application to the update of the Copernicus High Resolution Layer ImperVIOUSNESS Degree”. Em: *Remote Sensing* 8.7 (2016), p. 606. ISSN: 20724292. DOI: [10.3390/rs8070606](https://doi.org/10.3390/rs8070606). URL: <http://www.mdpi.com/2072-4292/8/7/606>.
- [43] A. M. Marangoz. “Analysis of Land Use/Land Cover Classification Results Derived From Sentinel-2 Image”. Em: *17th International Multidisciplinary Scientific GeoConference SGEM2017, Informatics, Geoinformatics and Remote Sensing 23* (2017). DOI: [10.5593/sgem2017/23/s10.004](https://doi.org/10.5593/sgem2017/23/s10.004).
- [44] S. K. McFEETERS. “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”. Em: *International Journal of Remote Sensing* 17.7 (1996), pp. 1425–1432. DOI: [10.1080/01431169608948714](https://doi.org/10.1080/01431169608948714). eprint: <https://doi.org/10.1080/01431169608948714>. URL: <https://doi.org/10.1080/01431169608948714>.
- [45] T. M. Mitchell. *Machine Learning*. 1ª ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077, 9780070428072.
- [46] S. E. Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell. “Combining satellite imagery and machine learning to predict poverty”. Em: 353.6301 (2016).
- [47] A. Neves, C. Damásio, J. Pires e F. Birra. “Deteção de estruturas permanentes a partir de dados de séries temporais Sentinel 1 e 2”. Em: *INForum 1* (2019), pp. 449–460. URL: <http://inforum.org.pt/INForum2019/docs/atas-do-inforum2019>.
- [48] P. T. Noi e M. Kappas. “Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery”. Em: *Sensors (Switzerland)* 18.1 (2018). ISSN: 14248220. DOI: [10.3390/s18010018](https://doi.org/10.3390/s18010018).
- [49] J. R. Otukei e T. Blaschke. “Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms”. Em: *International Journal of Applied Earth Observation and Geoinformation* 12.SUPPL. 1 (2010), pp. 27–31. ISSN: 15698432. DOI: [10.1016/j.jag.2009.11.002](https://doi.org/10.1016/j.jag.2009.11.002).
- [50] F. Pacifici, F. Del Frate, C. Solimini e W. Emery. “Neural Networks for Land Cover Applications”. Em: *Studies in Comp Intell* (jan. de 2008). DOI: [10.1007/978-3-540-79353-3\\_11](https://doi.org/10.1007/978-3-540-79353-3_11).

- [51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay. “Scikit-learn: Machine Learning in Python”. Em: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [52] M. J. Pereira. “Cartografia temática de ocupação / uso do solo do Instituto Geográfico Português”. Em: January (2008).
- [53] M. Pesaresi, D. Ehrlich, S. Ferri, A. Florczyk, S. Freire, F. Haag, M. Halkia, A. M. Julea, T. Kemper e P. Soille. “Global human settlement analysis for disaster risk reduction”. Em: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. Vol. 40. 7W3. 2015, pp. 837–843. DOI: 10.5194/isprsarchives-XL-7-W3-837-2015. URL: <http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-7-W3/837/2015/>.
- [54] M. Pesaresi, D. Ehrlich, A. Florczyk, S. Freire, A. Julea, T. Kemper, P. Soille e V. Syrris. *GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014)*. 2015. URL: [http://data.europa.eu/89h/jrc-ghsl-ghs\\_built\\_ldsmt\\_globe\\_r2015b](http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmt_globe_r2015b).
- [55] M. Pesaresi, C. Corbane, A. Julea, A. J. Florczyk, V. Syrris e P. Soille. “Assessment of the added-value of sentinel-2 for detecting built-up areas”. Em: *Remote Sensing* 8.4 (2016). ISSN: 20724292. DOI: 10.3390/rs8040299.
- [56] J. Redmon. “YOLOv3: An Incremental Improvement”. Em: (). arXiv: [arXiv:1804.02767v1](https://arxiv.org/abs/1804.02767v1).
- [57] H. Robbins e S. Monro. “A Stochastic Approximation Method”. Em: *IEEE Transactions on Systems, Man and Cybernetics* 1.4 (1951), pp. 338–344. ISSN: 21682909. DOI: 10.1109/TSMC.1971.4308316.
- [58] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo e J. P. Rigol-Sanchez. “An assessment of the effectiveness of a random forest classifier for land-cover classification”. Em: *ISPRS Journal of Photogrammetry and Remote Sensing* 67.1 (2012), pp. 93–104. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2011.11.002. arXiv: [Rodriguez-G](https://arxiv.org/abs/1111.002). URL: <http://dx.doi.org/10.1016/j.isprsjprs.2011.11.002>.
- [59] J. Rogan, N. Neeti, P. Panday, V. R. Galiano e B. Ghimire. “An Evaluation of Bagging, Boosting, and Random Forests for Land-Cover Classification in Cape Cod, Massachusetts, USA”. Em: *GIScience & Remote Sensing* 49.5 (2012), pp. 623–643. ISSN: 1548-1603. DOI: 10.2747/1548-1603.49.5.623.
- [60] O. Ronneberger, P. Fischer e T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Em: (), pp. 1–8. arXiv: [arXiv:1505.04597v1](https://arxiv.org/abs/1505.04597v1).



- [61] R. Schwaiger, R. Huber e H. A. Mayer. “Land Cover Classification of Landsat Images Using Problem-Adapted Artificial Neural Networks Land Cover Classification of Landsat Images Using Problem – Adapted Artificial Neural Networks”. Em: January (1995).
- [62] A. Sekertekin, A. M. Marangoz e H. Akcin. “Pixel-based classification analysis of land use land cover using Sentinel-2 and Landsat-8 data”. Em: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 42.4W6 (2017), pp. 91–93. ISSN: 16821750. DOI: [10.5194/isprs-archives-XLII-4-W6-91-2017](https://doi.org/10.5194/isprs-archives-XLII-4-W6-91-2017).
- [63] Y. Shin. “Application of boosting regression trees to preliminary cost estimation in building construction projects”. Em: *Computational Intelligence and Neuroscience* 2015.November (2015). ISSN: 16875273. DOI: [10.1155/2015/149702](https://doi.org/10.1155/2015/149702).
- [64] P. Snoeij e N. Miranda. “GMES Sentinel-1 Mission and System Performance Sentinel-1 Mission Objectives”. Em: April (2009).
- [65] G Sohn e I. J. Dowman. “Extraction of buildings from high resolution satellite data”. Em: *Automated Extraction of Man-Made Objects from Aerial and Space Images (III)*. (2001), pp. 345–355. URL: <https://pdfs.semanticscholar.org/9471/787e049f9308d84591c58196bb74b01c015d.pdf>.
- [66] T. Soukup, A. Sousa e T. Langanke. “CLC2018 Technical Guidelines”. Em: 3436 (2017), pp. 0–60.
- [67] P. A. Tavares, N. E. S. Beltrão, U. S. Guimarães e A. C. Teodoro. “Integration of sentinel-1 and sentinel-2 for classification and LULC mapping in the urban area of Belém, eastern Brazilian Amazon”. Em: *Sensors (Switzerland)* 19.5 (2019). ISSN: 14248220. DOI: [10.3390/s19051140](https://doi.org/10.3390/s19051140).
- [68] R. Tibshirani. “Regression shrinkage and selection via the lasso”. Em: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.
- [69] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Ö. Rommen, N. Floury, M. Brown, I. N. Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L’Abbate, R. Croci, A. Pietropaolo, M. Huchler e F. Rostan. “GMES Sentinel-1 mission”. Em: *Remote Sensing of Environment* 120 (2012), pp. 9–24. ISSN: 00344257. DOI: [10.1016/j.rse.2011.05.028](https://doi.org/10.1016/j.rse.2011.05.028). arXiv: [arXiv: 1011.1669v3](https://arxiv.org/abs/1011.1669v3). URL: <https://www.sciencedirect.com/science/article/abs/pii/S0034425712000600>.
- [70] D. Triantakoustantis, N. Chrysoulakis, A. Sazonova, T. Esch, C. Feigenwinter, S. Dzgn, E. Parlow, M. Marconcini e A. Tal. “On-line evaluation of Earth Observation Derived Indicators for Urban Planning and Management”. Em: *Urban Planning and Design Research* 3.0 (2015), p. 17. ISSN: 2328-1626. DOI: [10.14355/updr.2015.03.003](https://doi.org/10.14355/updr.2015.03.003). URL: <http://dx.doi.org/10.14355/updr.2015.03.003>.

- [71] Tucker C. J. “Red and Photographic Infrared Linear Combinations for Monitoring Vegetation”. Em: *Remote Sensing of Environment* 8.2 (1979), pp. 127–150. DOI: [10.1016/0034-4257\(79\)90013-0](https://doi.org/10.1016/0034-4257(79)90013-0).
- [72] J. Verde e J. Zêzere. “AVALIAÇÃO DA PERIGOSIDADE DE INCÊNDIO FLORESTAL”. Em: 49.1 (2009), pp. 1–23.
- [73] J. van Vliet, P. H. Verburg, S. R. Grădinaru e A. M. Hersperger. “Beyond the urban-rural dichotomy: Towards a more nuanced analysis of changes in built-up land”. Em: *Computers, Environment and Urban Systems* 74.November 2018 (2019), pp. 41–49. ISSN: 01989715. DOI: [10.1016/j.compenurbsys.2018.12.002](https://doi.org/10.1016/j.compenurbsys.2018.12.002). URL: <https://doi.org/10.1016/j.compenurbsys.2018.12.002>.
- [74] J. V. Vliet, P. H. Verburg, S. R. Grădinaru e A. M. Hersperger. “Computers , Environment and Urban Systems Beyond the urban-rural dichotomy : Towards a more nuanced analysis of changes in built-up land”. Em: *Computers, Environment and Urban Systems* 74.September 2018 (2019), pp. 41–49. ISSN: 0198-9715. DOI: [10.1016/j.compenurbsys.2018.12.002](https://doi.org/10.1016/j.compenurbsys.2018.12.002). URL: <https://doi.org/10.1016/j.compenurbsys.2018.12.002>.
- [75] L Wang, J Zhu e H Zou. “The doubly regularized support vector machine”. Em: *Statistica Sinica* 16.2 (2006), pp. 589–615. ISSN: 1017-0405.
- [76] H. Xu. “A new index for delineating built-up land features in satellite imagery”. Em: *International Journal of Remote Sensing* 29.14 (2008), pp. 4269–4276. ISSN: 13665901. DOI: [10.1080/01431160802039957](https://doi.org/10.1080/01431160802039957).
- [77] H. Zeng e H. J. Trussell. “Feature selection using LASSO”. Em: *IAENG International Journal of Computer Science* 33.February (2006), pp. 997–1000. ISSN: 1819-656X.
- [78] Y Zha, J Gao e S Ni. “Use of normalized difference built-up index in automatically mapping urban areas from TM imagery”. Em: (2003), pp. 583–594. ISSN: 0143-1161. DOI: [10.1080/01431160110104728](https://doi.org/10.1080/01431160110104728).
- [79] H. Zhang, Y. Zhang e H. Lin. “Urban land cover mapping using random forest combined with optical and SAR data”. Em: jul. de 2012, pp. 6809–6812. ISBN: 978-1-4673-1160-1. DOI: [10.1109/IGARSS.2012.6352600](https://doi.org/10.1109/IGARSS.2012.6352600).
- [80] T. Zhang. “Solving large scale linear prediction problems using stochastic gradient descent algorithms”. Em: (2004), p. 116. DOI: [10.1145/1015330.1015332](https://doi.org/10.1145/1015330.1015332).
- [81] Z. Zhang, Q. Liu e Y. Wang. “Road Extraction by Deep Residual U-Net”. Em: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753. ISSN: 1545598X. DOI: [10.1109/LGRS.2018.2802944](https://doi.org/10.1109/LGRS.2018.2802944). arXiv: [arXiv:1711.10684v1](https://arxiv.org/abs/1711.10684v1).
- [82] H. Zou e T. Hastie. “Regularization and variable selection via the elastic net”. Em: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.



## IDENTIFICADORES DOS PRODUTOS UTILIZADOS

Tabela A.1: Identificadores dos produtos que compõem a série temporal Sentinel-1

Sentinel-1
S1B_IW_GRDH_1SDV_20161209T182625_20161209T182650_003323_005ABC_17FF
S1A_IW_GRDH_1SDV_20161109T182713_20161109T182738_013869_0164EF_0E05
S1A_IW_GRDH_1SDV_20161004T182713_20161004T182738_013344_015478_0783
S1A_IW_GRDH_1SDV_20160910T182713_20160910T182738_012994_014917_267F
S1A_IW_GRDH_1SDV_20160805T182711_20160805T182736_012469_0137C7_AC5C
S1A_IW_GRDH_1SDV_20160712T182710_20160712T182735_012119_012C46_3477
S1A_IW_GRDH_1SDV_20160606T182708_20160606T182733_011594_011B77_29FE
S1A_IW_GRDH_1SDV_20160501T182703_20160501T182728_011069_010AA3_4743
S1A_IW_GRDH_1SDV_20160407T182702_20160407T182727_010719_00FFD5_16E2
S1A_IW_GRDH_1SDV_20160314T182701_20160314T182726_010369_00F5DA_F4AA
S1A_IW_GRDH_1SDV_20160207T182701_20160207T182726_009844_00E6A3_7F35
S1A_IW_GRDH_1SDV_20160126T182701_20160126T182726_009669_00E19B_1B4C

Tabela A.2: Identificadores dos produtos que compõem a série temporal Sentinel-2

Sentinel-2
S2A_MSIL1C_20160430T112122_N0201_R037_T29SND_20160430T112639
S2A_MSIL1C_20160520T112122_N0202_R037_T29SND_20160520T112904
S2A_MSIL1C_20160619T112112_N0204_R037_T29SND_20160619T112423
S2A_MSIL1C_20160629T112112_N0204_R037_T29SND_20160629T112737
S2A_MSIL1C_20160709T112122_N0204_R037_T29SND_20160709T112403
S2A_MSIL1C_20160719T112122_N0204_R037_T29SND_20160719T112942
S2A_MSIL1C_20160729T112122_N0204_R037_T29SND_20160729T112407
S2A_MSIL1C_20160808T112122_N0204_R037_T29SND_20160808T112920
S2A_MSIL1C_20160818T112122_N0204_R037_T29SND_20160818T112554
S2A_MSIL1C_20160828T112112_N0204_R037_T29SND_20160828T112114
S2A_MSIL1C_20160907T112112_N0204_R037_T29SND_20160907T112622
S2A_MSIL1C_20160917T112112_N0204_R037_T29SND_20160917T112420
S2A_MSIL1C_20160927T112112_N0204_R037_T29SND_20160927T112525
S2A_MSIL1C_20161007T112112_N0204_R037_T29SND_20161007T112902
S2A_MSIL1C_20161116T112332_N0204_R037_T29SND_20161116T112335
S2A_MSIL1C_20161206T112442_N0204_R037_T29SND_20161206T112747
S2A_MSIL1C_20161226T112452_N0204_R037_T29SND_20161226T112957

APÊNDICE



## TABELA DE MÉTRICAS

Labels	Dataset	Class	XGB			RF			Model DNN			SVM			SGD		
			Avg. F1Score	Avg. Precision	Avg. Recall	Avg. F1Score	Avg. Precision	Avg. Recall	Avg. F1Score	Avg. Precision	Avg. Recall	Avg. F1Score	Avg. Precision	Avg. Recall	Avg. F1Score	Avg. Precision	Avg. Recall
Grupo 1	Estático	1	0.7021	0.8313	0.6077	0.6534	0.7213	0.5972	0.6381	0.7536	0.5534	0.4690	0.8263	0.3276	0.5055	0.5247	0.5202
		2	0.9862	0.9793	0.9932	0.9829	0.9786	0.9873	0.9829	0.9760	0.9899	0.9802	0.9643	0.9965	0.9707	0.9755	0.9662
		3	0.9460	0.9655	0.9273	0.9267	0.9329	0.9206	0.9153	0.9374	0.8942	0.8834	0.9657	0.8141	0.7753	0.6997	0.8936
	Série temporal S1-S2	1	0.7670	0.8493	0.6993	0.7191	0.7492	0.6913	0.6447	0.8496	0.5194	0.6041	0.8238	0.4769	0.5198	0.4215	0.6779
		2	0.9888	0.9841	0.9934	0.9856	0.9834	0.9878	0.9846	0.9755	0.9939	0.9836	0.9723	0.9951	0.8881	0.9917	0.8040
		3	0.9626	0.9770	0.9486	0.9438	0.9562	0.9317	0.9243	0.9130	0.9359	0.9193	0.9787	0.8668	0.2094	0.1171	0.9905
	Série temporal S1-S2-DEM	1	0.7877	0.8617	0.7253	0.7281	0.7609	0.6980	0.6403	0.8647	0.5084	0.6166	0.8449	0.4855	0.6220	0.5789	0.6758
		2	0.9897	0.9866	0.9928	0.9860	0.9839	0.9882	0.9846	0.9742	0.9954	0.9839	0.9725	0.9956	0.9777	0.9828	0.9727
		3	0.9677	0.9770	0.9584	0.9474	0.9541	0.9408	0.9281	0.9556	0.9022	0.9227	0.9830	0.8693	0.8868	0.8596	0.9183
Grupo 2	Estático	1	0.5893	0.7748	0.4755	0.4021	0.5816	0.3073	0.1935	0.5883	0.1158	0.2773	0.5077	0.1908	0.2821	0.2071	0.4422
		2	0.5269	0.6927	0.4251	0.3955	0.4833	0.3346	0.3387	0.4877	0.2595	0.0851	0.5556	0.0461	0.2308	0.2875	0.1928
		3	0.7083	0.8418	0.6113	0.6041	0.6658	0.5528	0.4897	0.7590	0.3614	0.4298	0.7424	0.3025	0.4089	0.3484	0.4949
		4	0.9862	0.9774	0.9952	0.9827	0.9754	0.9901	0.9820	0.9681	0.9962	0.9786	0.9593	0.9986	0.9726	0.9772	0.9681
		5	0.9440	0.9656	0.9235	0.9260	0.9379	0.9145	0.9070	0.9445	0.8724	0.8834	0.9728	0.8090	0.8579	0.8752	0.8412
	Série temporal S1-S2	1	0.6218	0.8045	0.5067	0.4521	0.6250	0.3542	0.2016	0.5119	0.1255	0.0942	0.5000	0.0520	0.3258	0.2558	0.4484
		2	0.6017	0.6814	0.5387	0.4980	0.4989	0.4971	0.4027	0.5453	0.3192	0.3767	0.5362	0.2903	0.3550	0.4003	0.3190
		3	0.7595	0.8717	0.6728	0.6499	0.6860	0.6174	0.5408	0.7623	0.4190	0.4612	0.7643	0.3302	0.3990	0.3070	0.5698
		4	0.9887	0.9825	0.9951	0.9854	0.9822	0.9886	0.9837	0.9713	0.9964	0.9820	0.9675	0.9971	0.9743	0.9821	0.9666
		5	0.9616	0.9759	0.9478	0.9420	0.9588	0.9258	0.9243	0.9573	0.8934	0.9305	0.9726	0.8920	0.9130	0.9425	0.8853
	Série temporal S1-S2-DEM	1	0.7421	0.8795	0.6477	0.4908	0.6744	0.3858	0.0848	0.5074	0.0463	0.0635	0.3750	0.0347	0.3186	0.2685	0.3918
		2	0.6699	0.7609	0.5938	0.5146	0.5177	0.5116	0.4753	0.5222	0.4360	0.3505	0.5088	0.2673	0.3730	0.3948	0.3534
		3	0.8105	0.9084	0.7368	0.6666	0.7094	0.6287	0.5933	0.6164	0.5719	0.4878	0.7143	0.3704	0.4218	0.3467	0.5386
		4	0.9900	0.9823	0.9978	0.9859	0.9826	0.9892	0.9848	0.9774	0.9923	0.9826	0.9684	0.9972	0.9763	0.9813	0.9712
		5	0.9685	0.9789	0.9582	0.9462	0.9579	0.9348	0.9272	0.9696	0.8884	0.9289	0.9751	0.8869	0.9100	0.9356	0.8857
Grupo 3	Série temporal S1-S2-DEM	1	0.6964	0.7326	0.6635	0.6457	0.6128	0.6823	0.6177	0.6116	0.6239	0.4938	0.6796	0.3880	0.5011	0.4079	0.6507
		2	0.2201	0.6553	0.1323	0.1938	0.6153	0.1150	0.1199	0.6805	0.0658	0.0684	0.8277	0.0357	0.0385	0.3941	0.0203
		3	0.9724	0.9530	0.9926	0.9688	0.9532	0.9850	0.9685	0.9488	0.9891	0.9657	0.9364	0.9968	0.9600	0.9509	0.9693
		4	0.9639	0.9735	0.9545	0.9440	0.9471	0.9409	0.9271	0.9534	0.9023	0.9200	0.9733	0.8725	0.9003	0.8969	0.9037

Figura B.1: Tabela de métricas de todos os testes

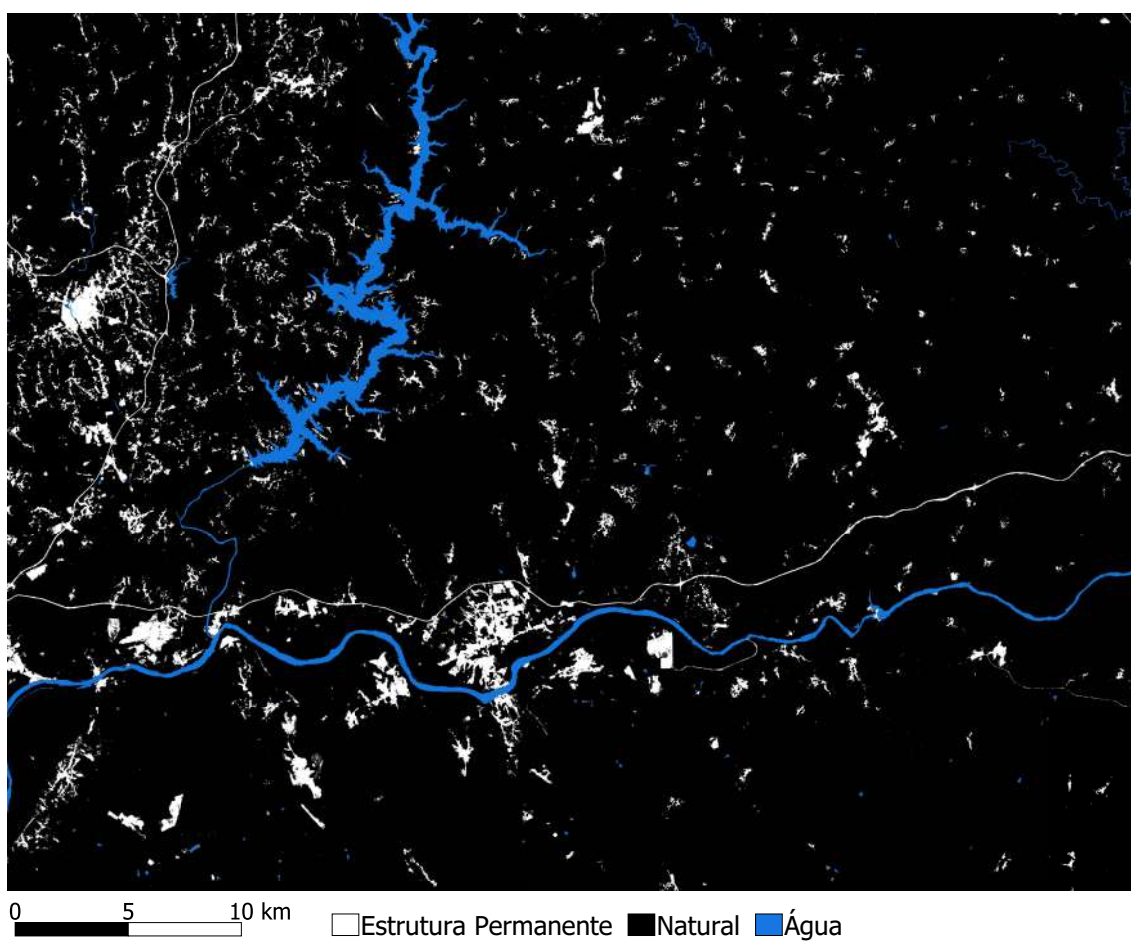


Figura C.1: Raster resultante do classificador XGBoost com séries temporais, grupo 1.

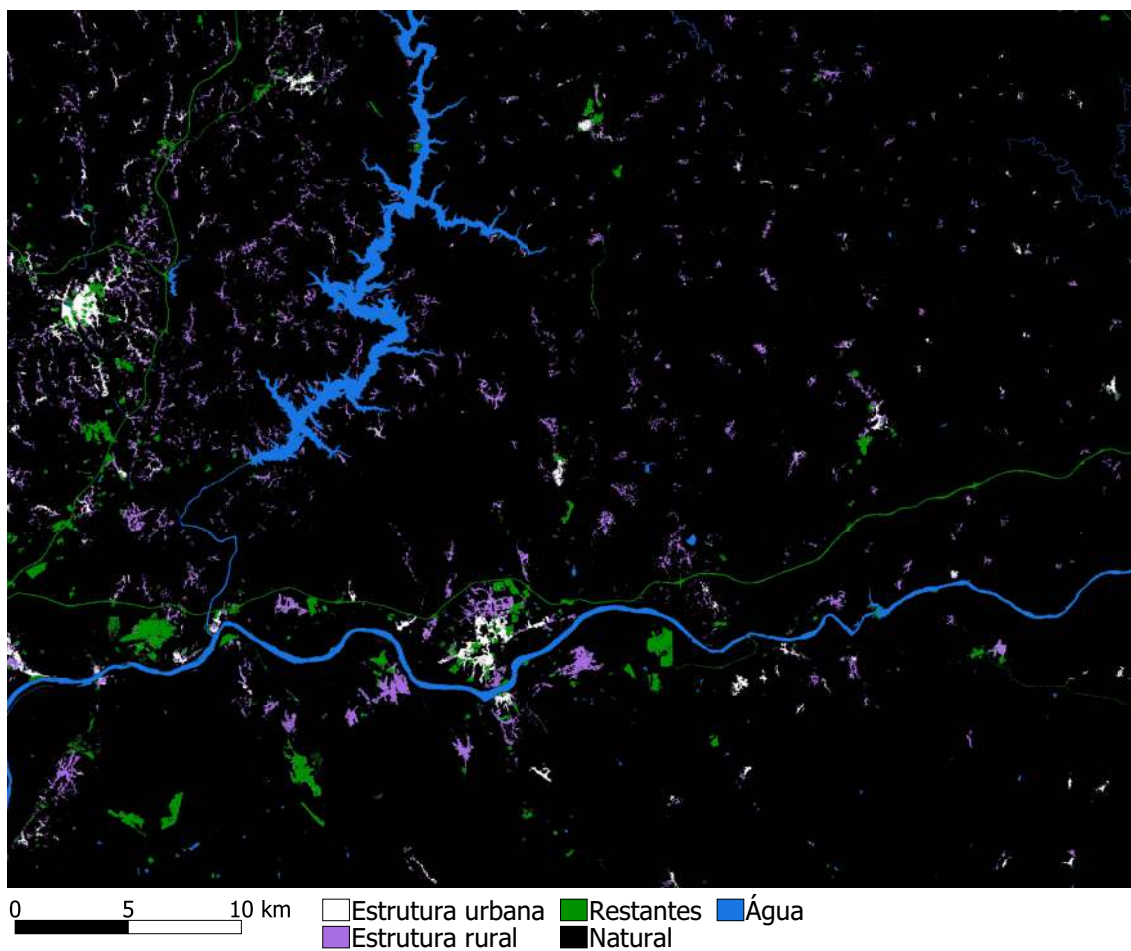


Figura C.2: Raster resultante do classificador XGBoost com séries temporais, grupo 2.



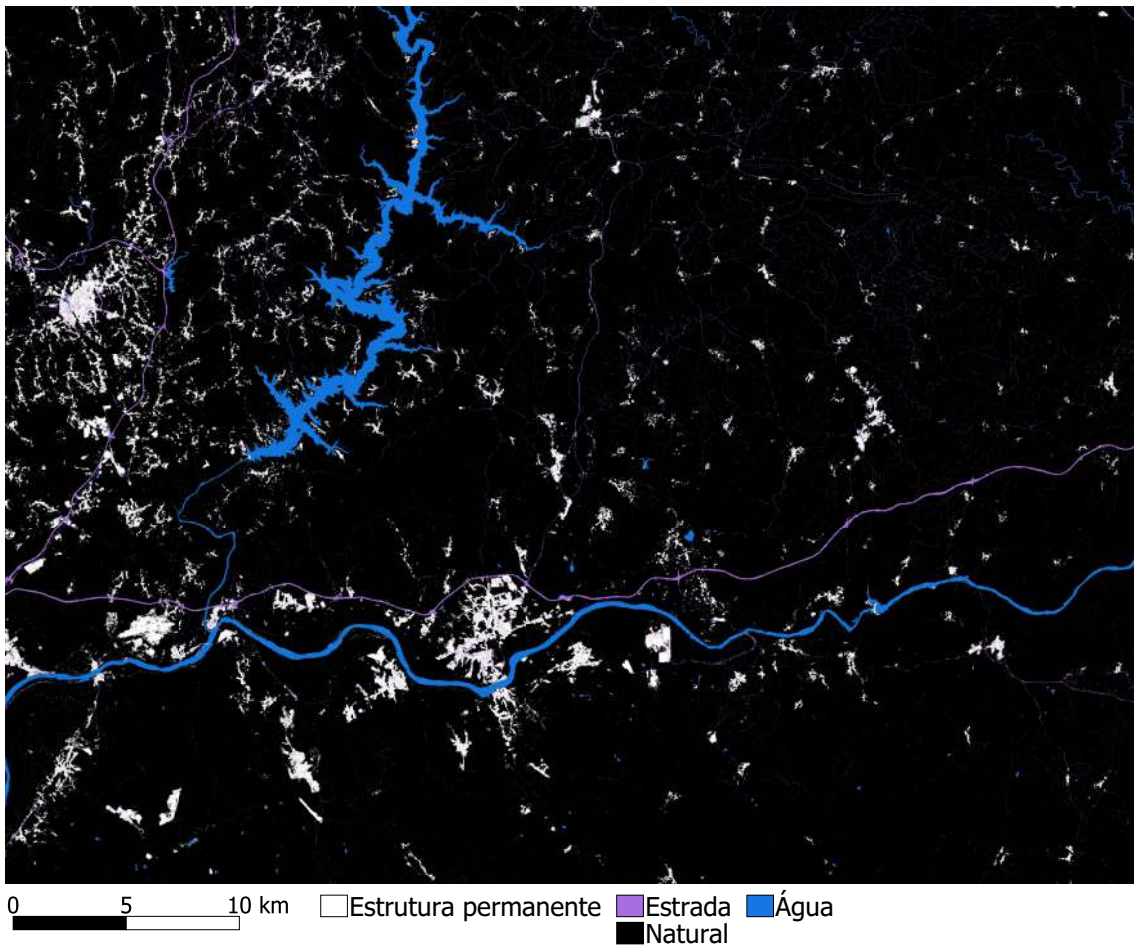


Figura C.3: Raster resultante do classificador XGBoost com séries temporais, grupo 3.

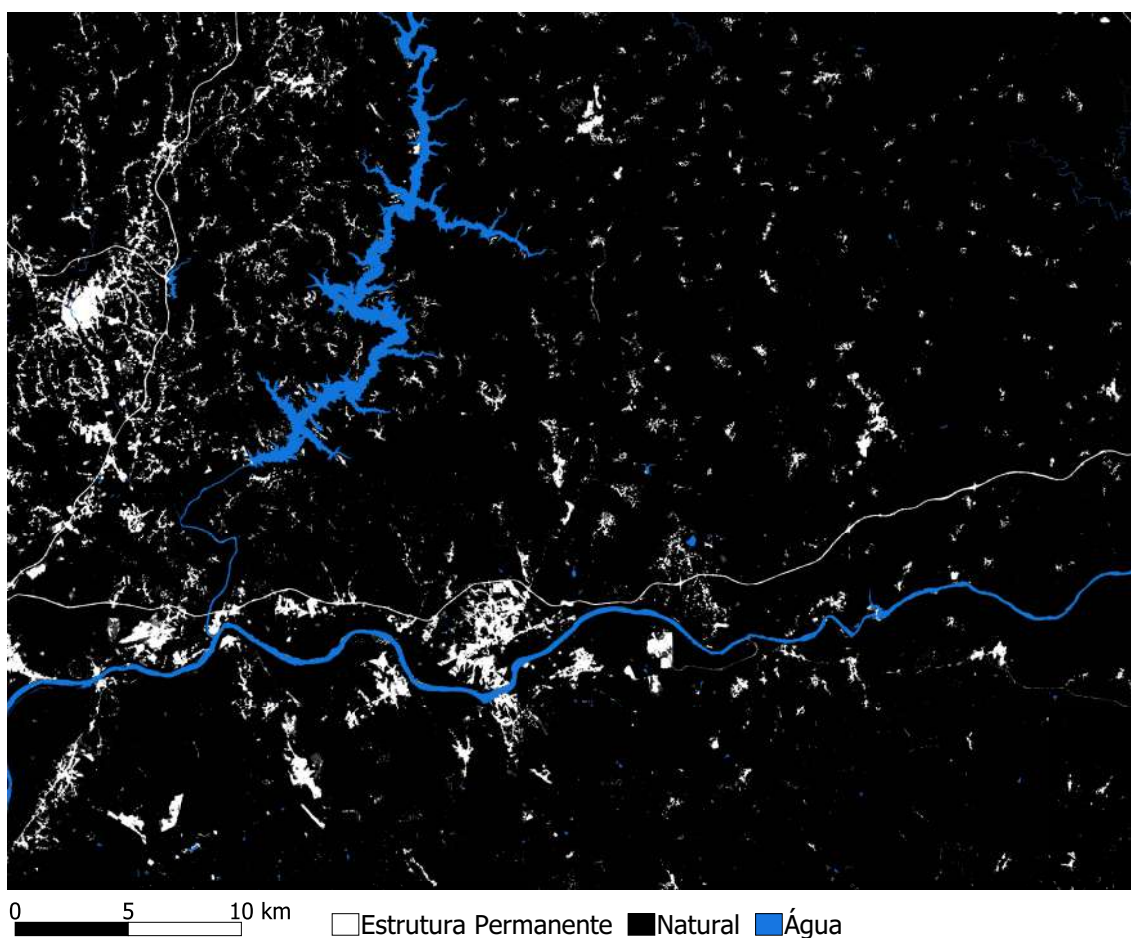


Figura C.4: Raster resultante do classificador Random Forest com séries temporais, grupo 1.

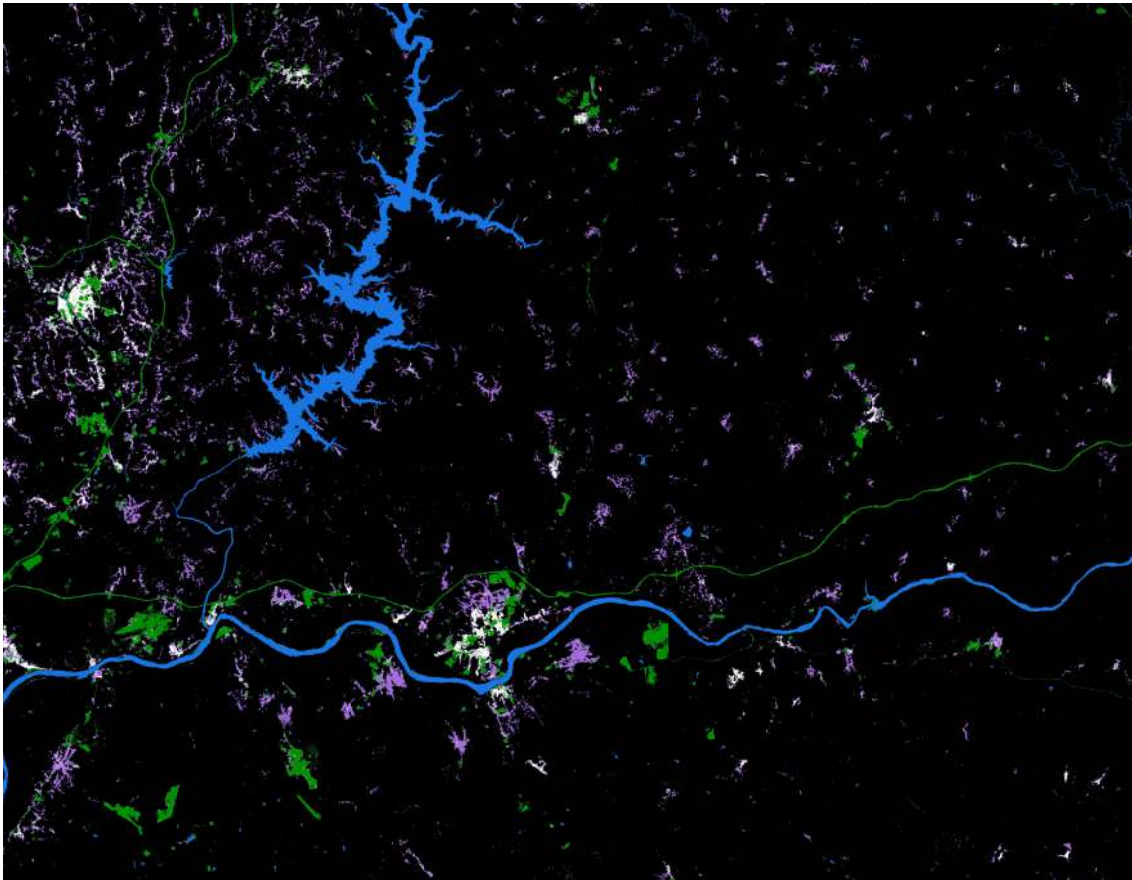


Figura C.5: Raster resultante do classificador Random Forest com séries temporais, grupo 2.

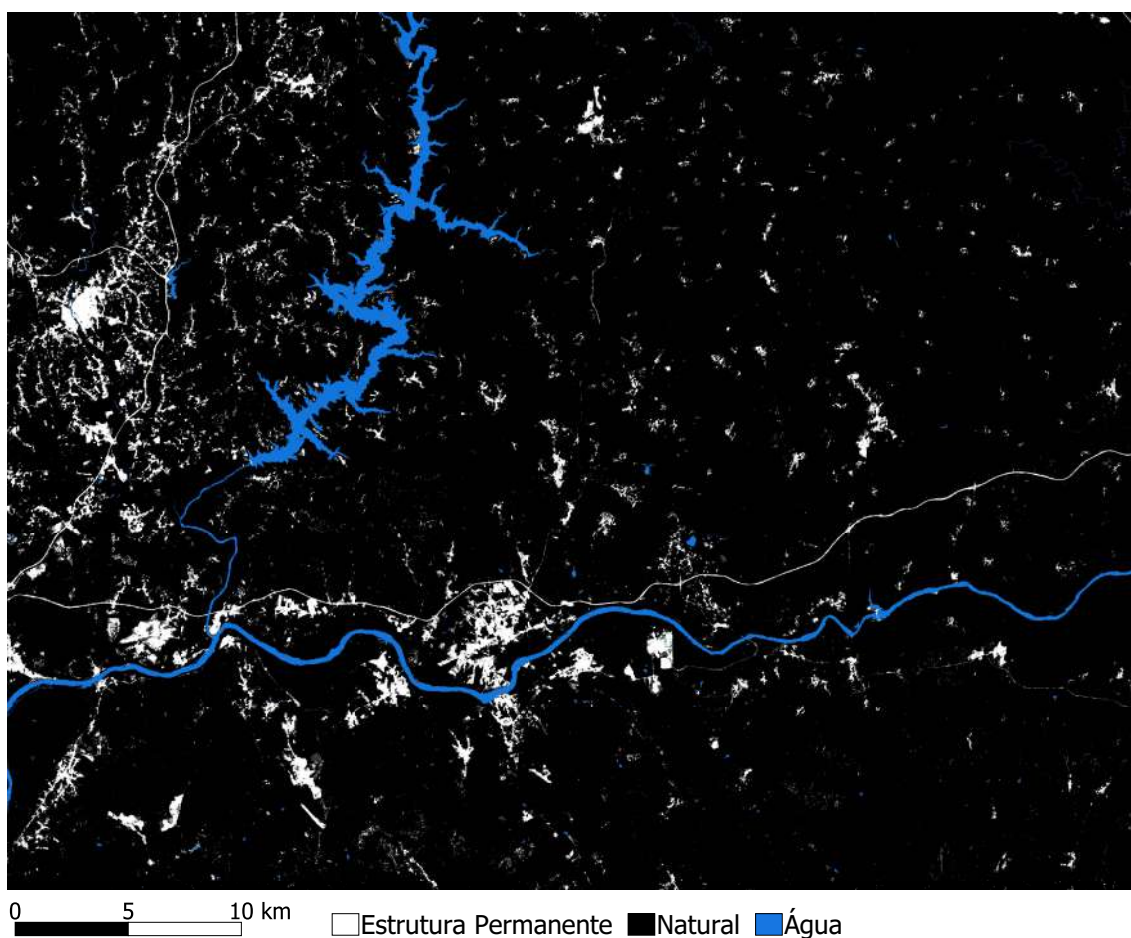
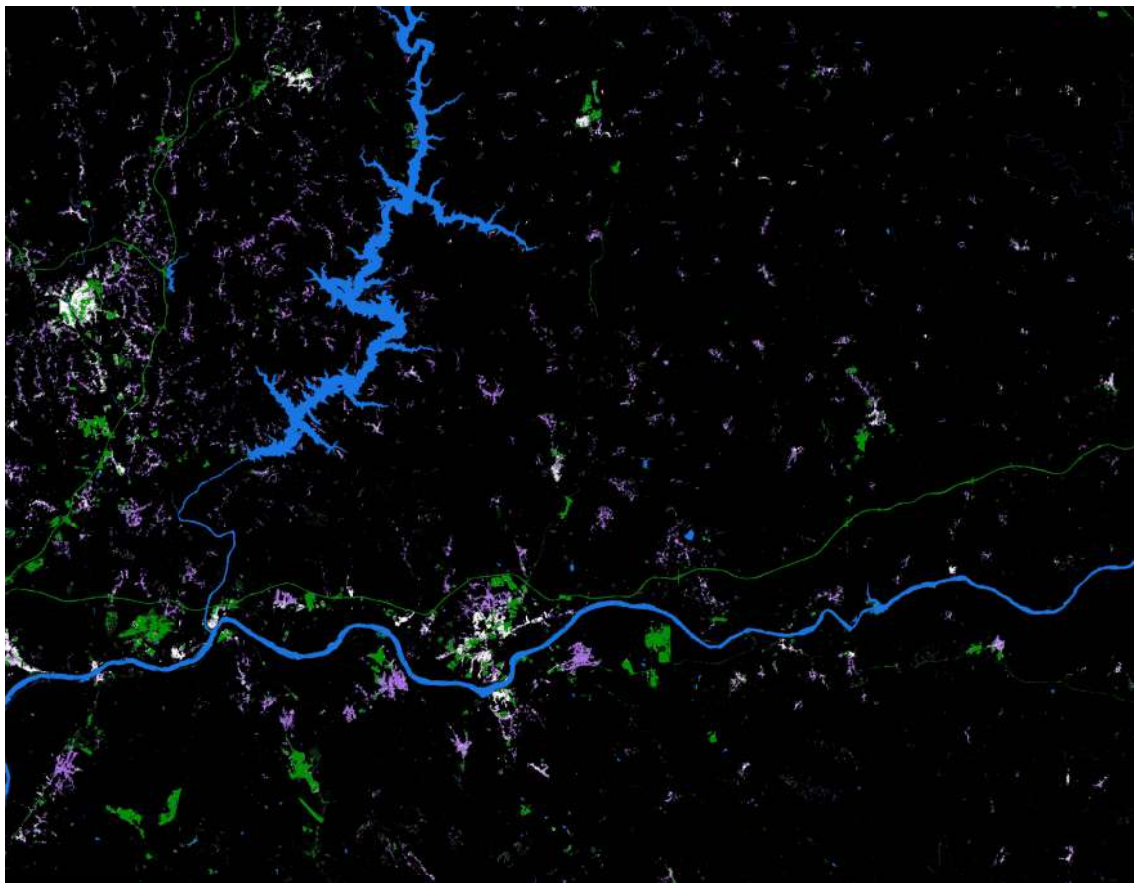


Figura C.6: Raster resultante do classificador Random Forest com dados estáticos, grupo 1.



0 5 10 km    Estrutura urbana   Restantes   Água  
Estrutura rural   Natural

Figura C.7: Raster resultante do classificador Random Forest com dados estáticos, grupo 2.

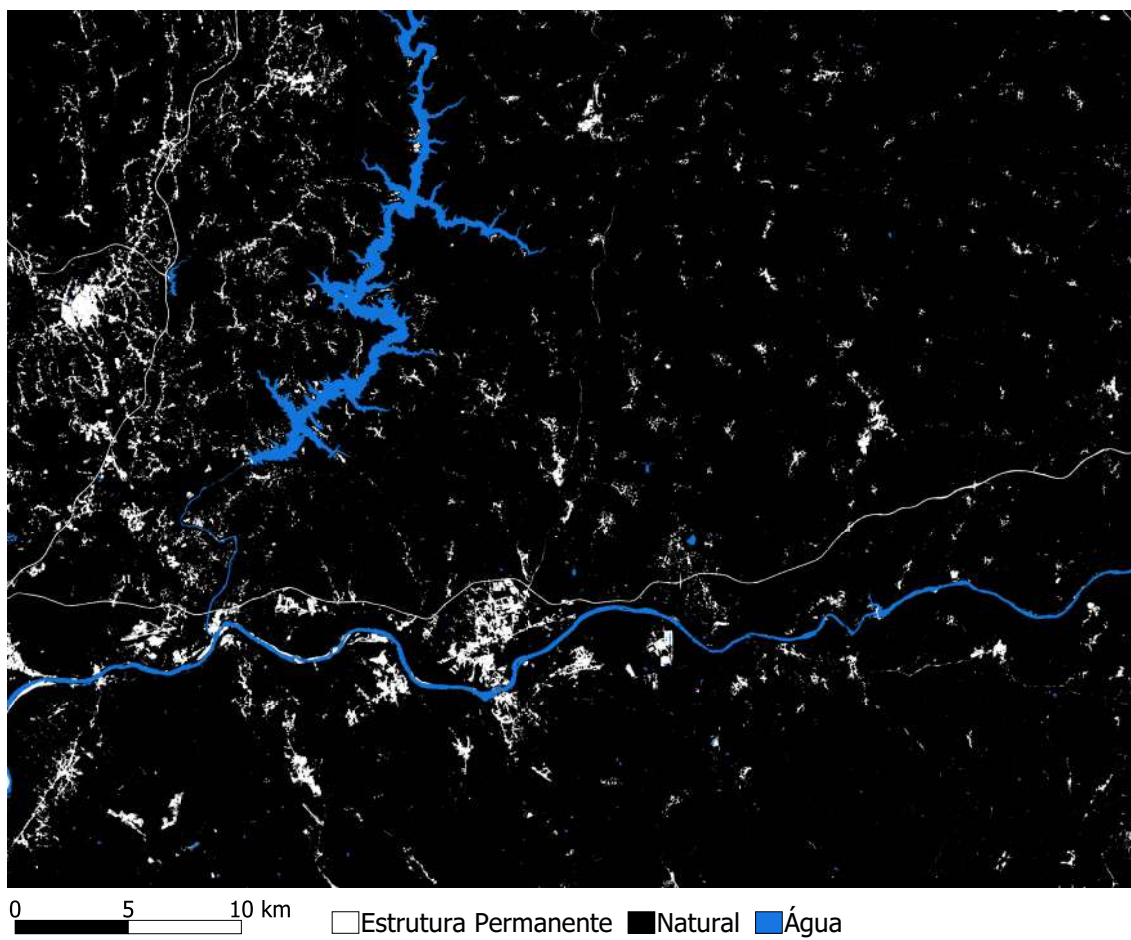


Figura C.8: Raster resultante do classificador DNN com séries temporais, grupo 1.

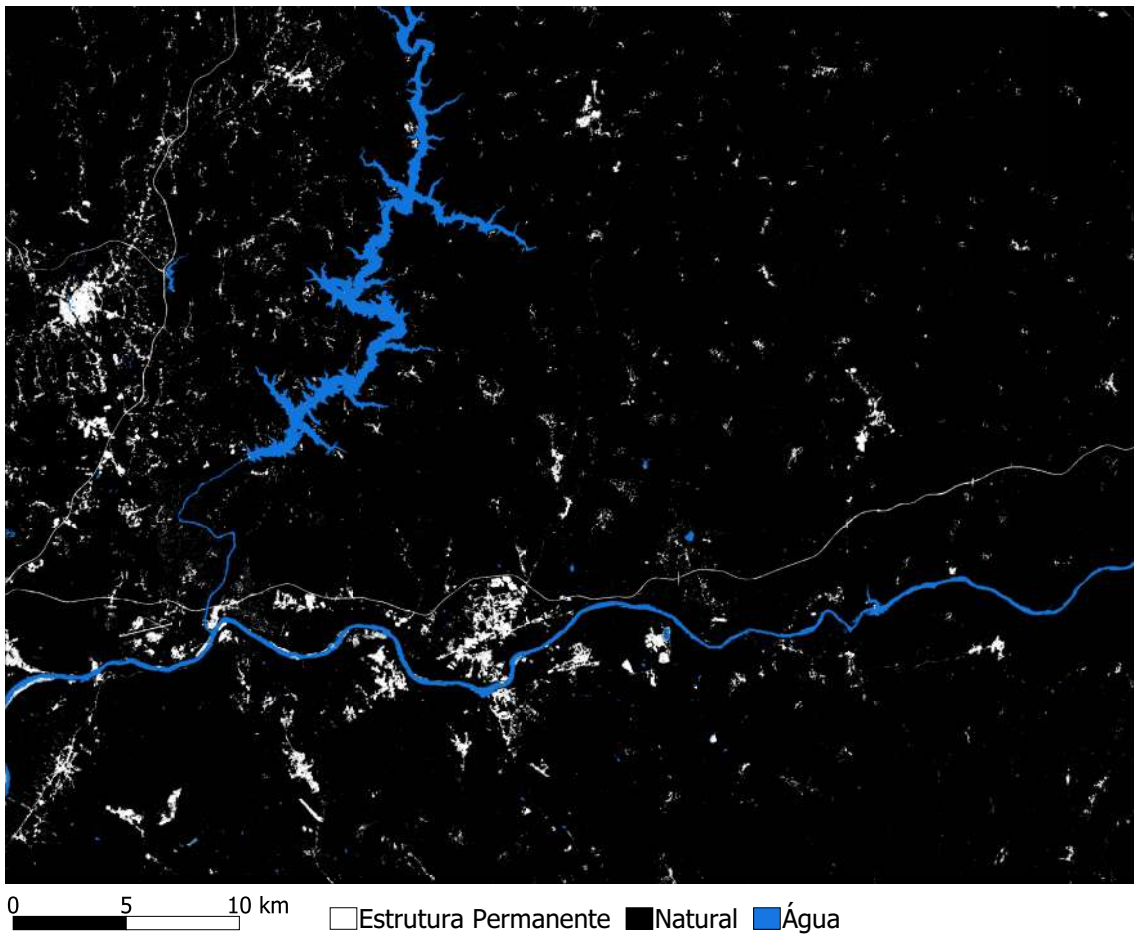


Figura C.9: Raster resultante do classificador DNN com dados estáticos, grupo 1.

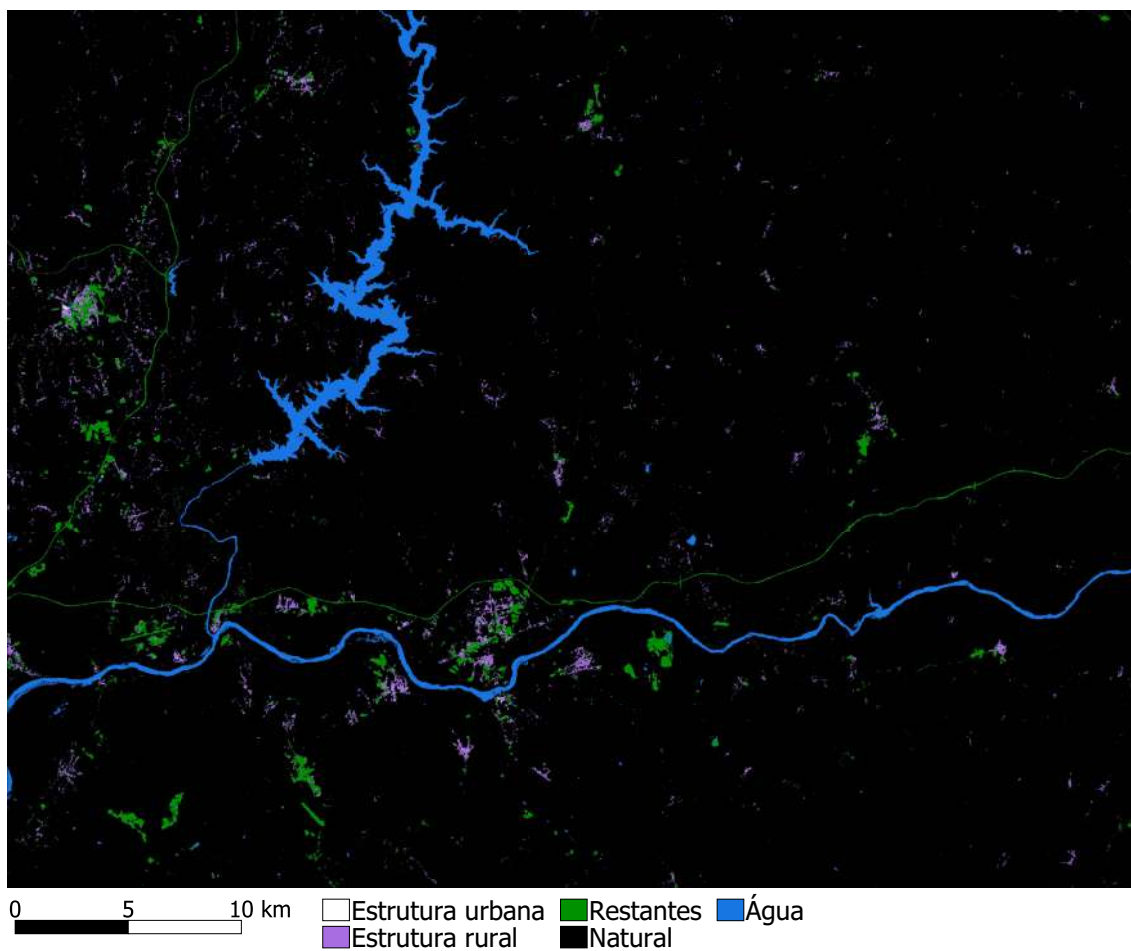


Figura C.10: Raster resultante do classificador DNN com dados estáticos, grupo 1.





## TABELAS DE ATRIBUTOS

Este apêndice consiste nas três tabelas de atributos para cada conjunto de dados utilizados durante o teste, sendo as mesmas extraídas diretamente do código. Nomenclatura utilizada:

- **Estatísticas temporais:** q0 (percentil 0%), q1 (percentil 25%), q2 (percentil 50%), q3 (percentil 74%), q4 (percentil 100%), var (variância), mean (média).
- **Atributos:** dem (modelo de elevação digital), VVVH ( $VV/VH$ ).

Tabela D.1: Tabela de atributos utilizados na análise estática

Índice	Nome
0	clipped_dem.vrt
1	clipped_dem_aspect.tif
2	clipped_dem_slope.tif
3	clipped_evi.tif
4	clipped_ndbi.tif
5	clipped_ndvi.tif
6	clipped_ndwi.tif
7	clipped_sentinel1_Gamma0_VH.vrt
8	clipped_sentinel1_Gamma0_VH_ASM.vrt
9	clipped_sentinel1_Gamma0_VH_Contrast.vrt
10	clipped_sentinel1_Gamma0_VH_Dissimilarity.vrt
11	clipped_sentinel1_Gamma0_VH_Energy.vrt
12	clipped_sentinel1_Gamma0_VH_Entropy.vrt
13	clipped_sentinel1_Gamma0_VH_GLCMCorrelation.vrt

Índice	Nome
14	clipped_sentinel1_Gamma0_VH_GLCMMean.vrt
15	clipped_sentinel1_Gamma0_VH_GLCMVariance.vrt
16	clipped_sentinel1_Gamma0_VH_Homogeneity.vrt
17	clipped_sentinel1_Gamma0_VH_MAX.vrt
18	clipped_sentinel1_Gamma0_VV.vrt
19	clipped_sentinel1_Gamma0_VV_ASM.vrt
20	clipped_sentinel1_Gamma0_VV_Contrast.vrt
21	clipped_sentinel1_Gamma0_VV_Dissimilarity.vrt
22	clipped_sentinel1_Gamma0_VV_Energy.vrt
23	clipped_sentinel1_Gamma0_VV_Entropy.vrt
24	clipped_sentinel1_Gamma0_VV_GLCMCorrelation.vrt
25	clipped_sentinel1_Gamma0_VV_GLCMMean.vrt
26	clipped_sentinel1_Gamma0_VV_GLCMVariance.vrt
27	clipped_sentinel1_Gamma0_VV_Homogeneity.vrt
28	clipped_sentinel1_Gamma0_VV_MAX.vrt
29	clipped_sentinel2_B01.vrt
30	clipped_sentinel2_B02.vrt
31	clipped_sentinel2_B03.vrt
32	clipped_sentinel2_B04.vrt
33	clipped_sentinel2_B05.vrt
34	clipped_sentinel2_B06.vrt
35	clipped_sentinel2_B07.vrt
36	clipped_sentinel2_B08.vrt
37	clipped_sentinel2_B09.vrt
38	clipped_sentinel2_B11.vrt
39	clipped_sentinel2_B12.vrt
40	clipped_sentinel2_B8A.vrt

Tabela D.2: Tabela de atributos utilizados na análise de séries temporais, excluindo dados de elevação de terreno

Índice	Nome
0	B01_mean.tiff
1	B01_q0.tiff
2	B01_q1.tiff
3	B01_q2.tiff
4	B01_q3.tiff
5	B01_q4.tiff
6	B01_var.tiff
7	B02_mean.tiff

---

Índice	Nome
8	B02_q0.tiff
9	B02_q1.tiff
10	B02_q2.tiff
11	B02_q3.tiff
12	B02_q4.tiff
13	B02_var.tiff
14	B03_mean.tiff
15	B03_q0.tiff
16	B03_q1.tiff
17	B03_q2.tiff
18	B03_q3.tiff
19	B03_q4.tiff
20	B03_var.tiff
21	B04_mean.tiff
22	B04_q0.tiff
23	B04_q1.tiff
24	B04_q2.tiff
25	B04_q3.tiff
26	B04_q4.tiff
27	B04_var.tiff
28	B05_mean.tiff
29	B05_q0.tiff
30	B05_q1.tiff
31	B05_q2.tiff
32	B05_q3.tiff
33	B05_q4.tiff
34	B05_var.tiff
35	B06_mean.tiff
36	B06_q0.tiff
37	B06_q1.tiff
38	B06_q2.tiff
39	B06_q3.tiff
40	B06_q4.tiff
41	B06_var.tiff
42	B07_mean.tiff
43	B07_q0.tiff
44	B07_q1.tiff
45	B07_q2.tiff
46	B07_q3.tiff

APÊNDICE D. TABELAS DE ATRIBUTOS

---

Índice	Nome
47	B07_q4.tiff
48	B07_var.tiff
49	B08_mean.tiff
50	B08_q0.tiff
51	B08_q1.tiff
52	B08_q2.tiff
53	B08_q3.tiff
54	B08_q4.tiff
55	B08_var.tiff
56	B09_mean.tiff
57	B09_q0.tiff
58	B09_q1.tiff
59	B09_q2.tiff
60	B09_q3.tiff
61	B09_q4.tiff
62	B09_var.tiff
63	B10_mean.tiff
64	B10_q0.tiff
65	B10_q1.tiff
66	B10_q2.tiff
67	B10_q3.tiff
68	B10_q4.tiff
69	B10_var.tiff
70	B11_mean.tiff
71	B11_q0.tiff
72	B11_q1.tiff
73	B11_q2.tiff
74	B11_q3.tiff
75	B11_q4.tiff
76	B11_var.tiff
77	B12_mean.tiff
78	B12_q0.tiff
79	B12_q1.tiff
80	B12_q2.tiff
81	B12_q3.tiff
82	B12_q4.tiff
83	B12_var.tiff
84	B8A_mean.tiff
85	B8A_q0.tiff

---

Índice	Nome
86	B8A_q1.tiff
87	B8A_q2.tiff
88	B8A_q3.tiff
89	B8A_q4.tiff
90	B8A_var.tiff
91	VH_mean.tiff
92	VH_q0.tiff
93	VH_q1.tiff
94	VH_q2.tiff
95	VH_q3.tiff
96	VH_q4.tiff
97	VH_var.tiff
98	VVVH_mean.tiff
99	VVVH_q0.tiff
100	VVVH_q1.tiff
101	VVVH_q2.tiff
102	VVVH_q3.tiff
103	VVVH_q4.tiff
104	VVVH_var.tiff
105	VV_mean.tiff
106	VV_q0.tiff
107	VV_q1.tiff
108	VV_q2.tiff
109	VV_q3.tiff
110	VV_q4.tiff
111	VV_var.tiff
112	evi_mean.tiff
113	evi_q0.tiff
114	evi_q1.tiff
115	evi_q2.tiff
116	evi_q3.tiff
117	evi_q4.tiff
118	evi_var.tiff
119	ndbi_mean.tiff
120	ndbi_q0.tiff
121	ndbi_q1.tiff
122	ndbi_q2.tiff
123	ndbi_q3.tiff
124	ndbi_q4.tiff

Índice	Nome
125	ndbi_var.tiff
126	ndvi_mean.tiff
127	ndvi_q0.tiff
128	ndvi_q1.tiff
129	ndvi_q2.tiff
130	ndvi_q3.tiff
131	ndvi_q4.tiff
132	ndvi_var.tiff
133	ndwi_mean.tiff
134	ndwi_q0.tiff
135	ndwi_q1.tiff
136	ndwi_q2.tiff
137	ndwi_q3.tiff
138	ndwi_q4.tiff
139	ndwi_var.tiff

Tabela D.3: Tabela de atributos utilizados na análise de séries temporais, complementadas com dados de elevação de terreno

Índice	Nome
0	B01_mean.tiff
1	B01_q0.tiff
2	B01_q1.tiff
3	B01_q2.tiff
4	B01_q3.tiff
5	B01_q4.tiff
6	B01_var.tiff
7	B02_mean.tiff
8	B02_q0.tiff
9	B02_q1.tiff
10	B02_q2.tiff
11	B02_q3.tiff
12	B02_q4.tiff
13	B02_var.tiff
14	B03_mean.tiff
15	B03_q0.tiff
16	B03_q1.tiff
17	B03_q2.tiff
18	B03_q3.tiff
19	B03_q4.tiff

---

Índice	Nome
20	B03_var.tiff
21	B04_mean.tiff
22	B04_q0.tiff
23	B04_q1.tiff
24	B04_q2.tiff
25	B04_q3.tiff
26	B04_q4.tiff
27	B04_var.tiff
28	B05_mean.tiff
29	B05_q0.tiff
30	B05_q1.tiff
31	B05_q2.tiff
32	B05_q3.tiff
33	B05_q4.tiff
34	B05_var.tiff
35	B06_mean.tiff
36	B06_q0.tiff
37	B06_q1.tiff
38	B06_q2.tiff
39	B06_q3.tiff
40	B06_q4.tiff
41	B06_var.tiff
42	B07_mean.tiff
43	B07_q0.tiff
44	B07_q1.tiff
45	B07_q2.tiff
46	B07_q3.tiff
47	B07_q4.tiff
48	B07_var.tiff
49	B08_mean.tiff
50	B08_q0.tiff
51	B08_q1.tiff
52	B08_q2.tiff
53	B08_q3.tiff
54	B08_q4.tiff
55	B08_var.tiff
56	B09_mean.tiff
57	B09_q0.tiff
58	B09_q1.tiff

APÊNDICE D. TABELAS DE ATRIBUTOS

---

Índice	Nome
59	B09_q2.tiff
60	B09_q3.tiff
61	B09_q4.tiff
62	B09_var.tiff
63	B10_mean.tiff
64	B10_q0.tiff
65	B10_q1.tiff
66	B10_q2.tiff
67	B10_q3.tiff
68	B10_q4.tiff
69	B10_var.tiff
70	B11_mean.tiff
71	B11_q0.tiff
72	B11_q1.tiff
73	B11_q2.tiff
74	B11_q3.tiff
75	B11_q4.tiff
76	B11_var.tiff
77	B12_mean.tiff
78	B12_q0.tiff
79	B12_q1.tiff
80	B12_q2.tiff
81	B12_q3.tiff
82	B12_q4.tiff
83	B12_var.tiff
84	B8A_mean.tiff
85	B8A_q0.tiff
86	B8A_q1.tiff
87	B8A_q2.tiff
88	B8A_q3.tiff
89	B8A_q4.tiff
90	B8A_var.tiff
91	VH_mean.tiff
92	VH_q0.tiff
93	VH_q1.tiff
94	VH_q2.tiff
95	VH_q3.tiff
96	VH_q4.tiff
97	VH_var.tiff



---

Índice	Nome
98	VVVH_mean.tiff
99	VVVH_q0.tiff
100	VVVH_q1.tiff
101	VVVH_q2.tiff
102	VVVH_q3.tiff
103	VVVH_q4.tiff
104	VVVH_var.tiff
105	VV_mean.tiff
106	VV_q0.tiff
107	VV_q1.tiff
108	VV_q2.tiff
109	VV_q3.tiff
110	VV_q4.tiff
111	VV_var.tiff
112	clipped_dem.vrt
113	clipped_dem_aspect.tif
114	clipped_dem_slope.tif
115	evi_mean.tiff
116	evi_q0.tiff
117	evi_q1.tiff
118	evi_q2.tiff
119	evi_q3.tiff
120	evi_q4.tiff
121	evi_var.tiff
122	ndbi_mean.tiff
123	ndbi_q0.tiff
124	ndbi_q1.tiff
125	ndbi_q2.tiff
126	ndbi_q3.tiff
127	ndbi_q4.tiff
128	ndbi_var.tiff
129	ndvi_mean.tiff
130	ndvi_q0.tiff
131	ndvi_q1.tiff
132	ndvi_q2.tiff
133	ndvi_q3.tiff
134	ndvi_q4.tiff
135	ndvi_var.tiff
136	ndwi_mean.tiff

## APÊNDICE D. TABELAS DE ATRIBUTOS

---

Índice	Nome
137	ndwi_q0.tiff
138	ndwi_q1.tiff
139	ndwi_q2.tiff
140	ndwi_q3.tiff
141	ndwi_q4.tiff
142	ndwi_var.tiff

