remote sensing

MDPI

Relatório do trabalho prático realizado no âmbito da disciplina de Deteção Remota Multiespectral

Análise comparativa de técnicas de amostragem e algoritmos de aprendizagem automática para classificação de uso e ocupação do solo com imagens Sentinel-2

Sofia Freire[1], Silvia Mourão[2]

[1] Faculdade de Ciências, Universidade de Lisboa; sbarradasfreire@gmail.com
[2] Faculdade de Ciências, Universidade de Lisboa; silamourao@gmail.com
* Correspondence: silamourao@gmail.com;

Abstract: A elaboração de mapas de uso e ocupação do solo é uma tarefa de grande importância no âmbito da deteção remota. Embora existam vários métodos para classificação de imagens, devido à elevada extensão do território a classificar e à quantidade de classes presentes, a elaboração destes mapas continua ainda muito dependente da foto interpretação, estando as técnicas da deteção remota ainda um pouco aquém das necessidades dos utilizadores. Nesse âmbito, propusemo-nos a testar algumas metodologias na área da aprendizagem automática, com foco na seleção de amostras de treino e nos algoritmos Random Forest, Decision Trees e Naive Bayes com o objetivo de avaliar a classificação da imagem obtida, tendo como objetivo final obter um resultado próximo da carta de uso e ocupação do solo portuguesa de 2018. Os resultados foram bastante variáveis consoante os casos considerados, no entanto, a utilização de amostras de 70% das classes, numa imagem que continha apenas as bandas, classificada pelo algoritmo Naive Bayes obteve, na generalidade, um melhor resultado do que os outros métodos, com Overall Accuracy de 83%, Coeficiente K de 0.8 e um F1 Score médio de 73.58%.

Keywords: ocupação do solo; machine learning; random forests; naive bayes; decision trees; deteção remota multiespetral

Citation: Freire, S.; Mourão, S.; F. Análise comparativa de técnicas de amostragem e algoritmos de aprendizagem automática para classificação de uso e ocupação do solo com imagens Sentinel-2. *Remote Sens.* **2022**, *14*, x.

Received: date
Accepted: date
Published: date

1. Introdução

O mapeamento de ocupação do solo urbano é um método fundamental para reconhecer e localizar utilizações do solo para diferentes propósitos, como monitorização de ambientes urbanos, planeamento e ordenamento do território. Atualmente, os métodos utilizados para atualizar estes mapas dependem ainda da interpretação de fotografia aérea e trabalho de levantamento em campo, ambos dispendiosos e morosos [1].

Com a evolução das imagens de satélite e das ferramentas de deteção remota, procuraram-se novas soluções para conseguir de alguma forma obter resultados aceitáveis utilizando menos recursos humanos e monetários. No entanto, existem ainda várias limitações na análise destas imagens, nomeadamente o facto das imagens de alta resolução não estarem disponíveis gratuitamente ao publico geral e das imagens de baixa resolução serem de difícil interpretação.

Os métodos de classificação de imagens com recurso a machine learning e deep learning têm sofrido um aumento de popularidade exponencial nos últimos anos, devido à sua elevada capacidade de aprendizagem e interpretação de imagens, conseguindo em alguns casos obter resultados bastante bons no âmbito da classificação do solo [2], utilizando algoritmos como Random Forest [3], Decision Trees [4] , Naive Bayes [5], K-Nearest Neighbour

[6], Convolutional Neural Network (CNN) [7], entre muitos outros. No entanto, muitos dos estudos focam-se em áreas de pequenas dimensões, ou imagens com resolução espacial muito elevada o que pode levar a um sobre ajustamento (overfitting) dos resultados ou à definição de classes demasiado descritivas, o que consequentemente pode levar a que estes não consigam ser replicados noutras situações. Outros destes estudos focam-se em análises comparativas de métodos para determinadas regiões, no entanto não investigam a forma como fatores como a escolha das amostras ou dos parâmetros dos algoritmos pode influenciar a solução final.

Embora atualmente os métodos de CNN apresentem os melhores resultados nesta área, no âmbito académico foi relevante estudar a influência de determinados fatores em algoritmos de aprendizagem automática, nomeadamente a forma como as amostras são escolhidas, a influência de classes com pouca expressão e a comparação entre três classificadores diferentes: Random Forest, Decision Tree e Naive Bayes. O projeto foi realizado para uma imagem Sentinel-2 na zona centro-sul de Portugal com uma área de 1210km^2, utilizando também informação da Carta de Uso e Ocupação do Solo 2018 (COS2018) da Direção Geral do Território (DGT) para melhor compreensão do aspeto visual das classes com nível mais detalhado.

2. Materiais e Métodos

2.1 Área de estudo

A área de estudo localiza-se numa região onde se encontram 3 zonas do nível III da Nomenclatura de Unidades Territoriais para Fins Estatísticos (NUTS) sendo essas Área Metropolitana de Lisboa, Alentejo Litoral e Alentejo Central. As coordenadas do seu canto superior esquerdo são (38º 35′ 45.6′′ N, -8º 50′ 9.6′′E) e do seu canto inferior direito (38º 23′38.4′′N, -8′12′ 0′′E), sendo as suas dimensões de 55km de comprimento, 22km de altura e 1210km^2 de área.

A área metropolitana de Lisboa é uma região situada no centro-sul de Portugal Continental. Esta é constituída por 18 municípios e é caracterizada por, na grande generalidade, apresentar áreas planas de baixa altitude (menores que 100 metros). A oeste encontra-se uma área complexa e diversificada a nível morfológico. A área do Alentejo Central é uma parte da região do Alentejo que corresponde por completo o distrito de Évora. Esta é constituída por 14 municípios e é caracterizada por extensas áreas de planícies (altitudes inferiores a 400 m), apresentando algumas áreas com zonas de relevo acentuado, no entanto, sem características montanhosas. Esta área é rica em recursos minerais não metálicos. O clima apresenta características mediterrânicas. Por fim, a área do Alentejo Litoral é, também, uma parte da região do Alentejo que é dividida pelo distrito de Setúbal e o distrito de Beja. Esta é constituída por 5 concelhos. A área em estudo apresenta um clima pré-mediterrânico com uma forte influência marítima, sendo o solo constituído, na sua grande maioria por uma vasta gama de rochas metamórficas, sedimentares e vulcânicas. No entanto, no sistema estuarino do Sado, o solo é essencialmente arenoso.

O enquadramento da área de estudo em Portugal continental e as respetivas regiões NUTS III podem ser vistas na figura 1.

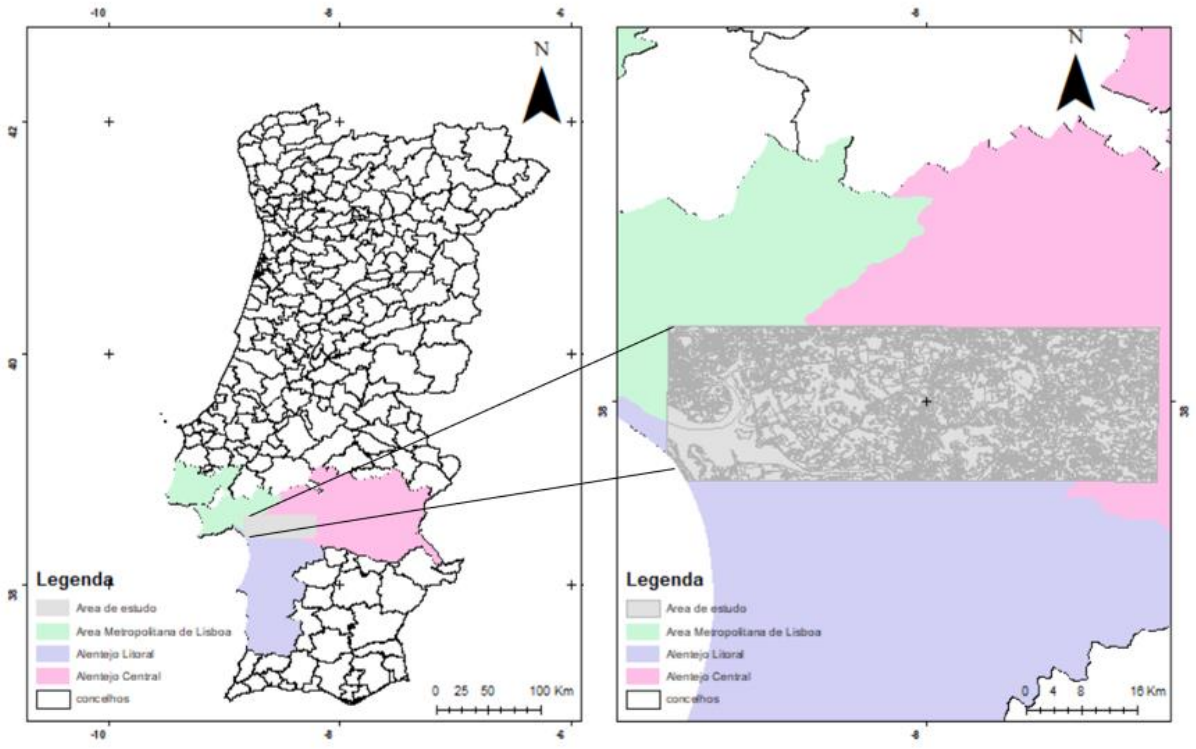

Figura 1: Área de estudo

2.2 Carta de Uso e Ocupação do Solo

Um mapa de uso e ocupação do solo representa informação espacial de vários tipos (classes) de cobertura física da superfície terrestre, como por exemplo florestas, matos, rios e zonas urbanas. Estes mapas têm um tempo de vida limitado pois o solo está constantemente a ser alvo de alterações, sendo necessária a sua atualização em períodos de tempo regulares. Em Portugal, a elaboração da carta de uso e ocupação do solo (COS) está a cargo da DGT, sendo que a versão mais recente disponibilizada ao público foi elaborada para o ano de 2018. A COS consiste numa cartografia de polígonos que representam unidades de ocupação do solo homogéneas, sendo especificado na sua elaboração que a área de terreno mínima representada corresponde a 1ha, com distancia entre linhas superior ou igual a 20 m e cuja percentagem de uma determinada classe de uso do solo seja superior ou igual a 75% da área delimitada (Caetano et al., 2017; Direção-Geral do Território, 2019).

A nomenclatura da COS é constituída por um sistema hierárquico de quatro classes, sendo que para a COS2018 foram consideradas um total de 83 classes no nível mais profundo, tendo esta carta sido elaborada principalmente pela interpretação de ortofotos. A DGT considera ainda uma outra nomenclatura, denominada COSsim, uma simplificação da nomenclatura do COS que contém apenas três níveis, sendo o mais detalhado composto por 15 classes, e que podem ser vistas na figura 2. As classes que serão utilizadas no âmbito deste projeto correspondem a 12 classes de nível III e uma classe de nível II, provenientes do COSsim, e que foram obtidas a partir do COS2018 por generalização.

Nível 1	Nível 2	Nível 3
1 – Artificializado	10 – Artificializado	100 – Artificializado
2 – Agricultura	21 – Agricultura	211 – Culturas anuais de outono/inverno*
		212 – Culturas anuais de primavera/verão*
		213 – Outras áreas agrícolas*
3 - Floresta	31 – Folhosas	311 – Sobreiro e Azinheira
		312 – Eucalipto
		313 – Outras folhosas
	32 - Resinosas	321 – Pinheiro bravo
		322 – Pinheiro manso
		323 – Outras resinosas
4 – Matos e vegetação herbácea espontânea	41 – Matos	410 – Matos
	42 – Vegetação herbácea espontânea	420 – Vegetação herbácea espontânea
5 – Superfícies sem vegetação	50 – Superfícies sem vegetação	500 – Superfícies sem vegetação
6 – Água e zonas húmidas	61 – Zonas húmidas	610 – Zonas húmidas
	62 – Água	620 – Água

* A COSsim2018 não tem estas subclasses da Agricultura.

Figura 2 - Classes do COSsim

2.3 Algoritmos de Machine Learning

2.3.1 Random Forest (RF)

Random Forest é um classificador que é muitas vezes utilizado devido à sua grande precisão na classificação. Este consiste numa combinação de um grande número de árvores classificadoras individuais que vão operar em conjunto. Cada árvore individual, treinada previamente num conjunto de teste, vai gerar uma previsão de classe e a classe mais vezes votada vai ser a escolhida para o modelo de classificação [3].

Neste estudo foram utilizadas 200 árvores de decisão para cada método.

2.3.2 Decision trees

O método de árvores de decisão consiste numa aprendizagem não paramétrica que é amplamente utilizada para processos de classificação. O objetivo é determinar o valor de uma variável com base na criação de uma hierarquia de questões if/else.

Neste estudo foi utilizada uma árvore com 10 ramificações.

2.3.3 Naive Bayes Classifier

O Naive Bayes Classifier, também conhecido como Normal Bayes Classifier, é um classificador probabilístico simples que se baseia na aplicação do teorema de Bayes, que descreve a probabilidade de um evento se relacionar com outro evento, com suposições de independência fortes.

2.4 Dados e Processamento

Neste estudo, foram utilizadas quatro imagens referentes ao satélite. Sentinel-2 do nível Ap, retiradas do site Onda Dias, entre os anos 2017 e 2018, para as quatro estações - Verão, Primavera, Outono e Inverno. As datas em questão foram:

1. 2 de outubro de 2017 - Outono
2. 15 de janeiro de 2018 - Inverno
3. 15 de maio de 2018 - Primavera
4. 29 de julho de 2018 - Verão

Na aquisição das imagens, o principal cuidado que foi tido foi a verificação da não existência de nuvens na área de estudo.

Foram ainda utilizados dados da carta de uso e ocupação do Solo – 2018, retirados do Sistema Nacional de Informação Geográfica (SNIG), no formato shapefile para as zonas de estudo - área metropolitana de Lisboa, Alentejo Litoral e Alentejo Central.

Tendo as imagens sido extraídas do Onda Dias, realizou-se um pré-processamento. Para tal, recorreu-se ao software SNAP. Aqui realizou-se um Resample para uniformizar o tamanho das bandas das quatro imagens. De seguida, realizou-se um Subset de 1/8 da imagem original – tal deveu-se ao facto de uma imagem maior exigir um esforço computacional muito elevado ao qual o grupo não estava preparado. Por fim, foram criadas três técnicas para a redução de dimensionalidade – BandMaths, Principal Component Analysis (PCA) e Texturas.

A primeira consistiu em determinar novos valores de amostra derivadas de bandas já existentes. Foram determinados os índices Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Soil Adjusted Vegetation Index (SAVI) e Normalized Difference Built-up Index (NDBI). O primeiro consiste no índice de vegetação que aproveita o facto da vegetação "verde" interagir de forma característica com a radiação eletromagnética. O NDWI é conhecido por estar relacionado com teor de água da vegetação. O SAVI é, também, um índice de vegetação, no entanto, tem como função minimizar o brilho do solo, utilizando um fator de correção de brilho, sendo muito comum a sua utilização em zonas onde a vegetação é baixa. Por fim, o NDBI, é conhecido por enfatizar áreas onde existem construções. As equações correspondentes a estes índices e as bandas utilizadas para o seu cálculo encontram-se descritas nas equações (1)-(4).

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \quad (1)$$

$$NDWI = \frac{(Green - NIR)}{(Green + NIR)} \quad (2)$$

$$SAVI = \frac{1.5 * (NIR - Red)}{(NIR + Red + 0.5)} \quad (3)$$

$$NDBI = \frac{(SWIR1 - NIR)}{(SWIR1 + NIR)} \quad (4)$$

O PCA teve como objetivo remapear as informações da imagem formando um novo conjunto de imagens, sendo dimensionadas de forma a evitar valores negativos nos pixels. Já as texturas consistem em variações das intensidades numa imagem.

Por fim, para cada técnica, sobrepuseram-se espacialmente as quatro imagens de forma a gerar uma única imagem, sendo exportada uma única imagem no formato GE-Otiff para cada método.

As imagens que irão ser consideradas no estudo são compostas pelas bandas das quatro imagens em estudo, dispostas da seguinte forma:

5. Bandas
6. Índices
7. PCA
8. Texturas
9. Bandas + Índices
10. Bandas + Índices + Texturas

Num passo encadeado ao processamento no SNAP, foram importadas para o QGIS os dados no formato shapefile das três regiões NUTS sobre as quais a nossa área de estudo se encontra. Estas shapefiles foram juntas e depois recortadas para cobrir apenas a área de interesse do projeto. A classificação do COS foi, depois, generalizada de forma a garantir que existia apenas informação correspondente as classes do COSsim, sendo que das 9 classes de nível 1 presentes no COS, 7 tem correspondência direta no COSsim e as duas classes sem correspondência (Pastagens e Superfícies Agroflorestais) foram divididas pelas classes Florestas, Agricultura e Mato consoante a que foi considerada como melhor correspondência.

Utilizando um dos ficheiros GEOtiff resultantes do processamento em SNAP no QGIS e sobrepondo a shapefile resultante com a classificação do COSsim foram desenhados polígonos para cada uma das 13 classes consideradas, com a seguinte distribuição, visível na tabela 1:

Tabela 1 - Classes para classificação da imagem

Classe	Número de Polígonos
100 - Artificializado	100
21 - Agricultura	100
311 – Sobreiro e Azinheira	100
312 – Eucalipto	100
313 – Outras folhosas	100
321 – Pinheiro bravo	100
322 – Pinheiro Manso	100
323 – Outras Resinosas	4
410 – Matos	100
420 – Vegetação herbácea espontânea	35
500 – Superfícies sem vegetação	35
610 – Zonas húmidas	100
620 - Água	100

Os polígonos desenhados foram depois separados através das ferramentas Random Extract Within Subset e Difference do QGIS num conjunto de polígonos de treino, que correspondem a 70% do total e polígonos de teste, que correspondem aos restantes 30%. Devido à quantidade extremamente limitada de polígonos existentes para a classe 323

foram consideradas algumas abordagens diferentes que incluíram ou eliminaram esta classe e que são descritas numa fase seguinte.

Com os polígonos prontos, o projeto pode então passar para o software Orfeo Toolbox

No Orfeo Toolbox foram realizados diversos processamentos de forma a obter uma imagem classificada – utilizando algoritmos de Machine Learning – com a sua respetiva Confusion Matrix. Numa primeira fase, realizou-se o PolygonClassStatistics no qual teve como objetivo gerar as estatísticas do conjunto de treino – criado no QGIS - sendo estas determinadas em função da imagem gerada no SNAP. De seguida realizou-se o SampleSelection que consiste em selecionar amostras do conjunto de dados do treino. Aqui foram utilizadas quatro estratégias diferentes. A primeira consistiu em utilizar todas as classes consideradas para o projeto, incluindo a classe 323 com pouca expressão e a segunda consistiu em eliminar os polígonos da classe 323 e utilizar apenas as outras 12 classes para o treino e classificação. De seguida, para cada uma destas abordagens, foram consideradas duas estratégias de amostragem, a primeira onde foram utilizadas 70% das amostras de treino e a segunda consistiu em definir o mesmo número de amostras, com a menor classe totalmente amostrada.

Em terceiro e quarto lugar, realizaram-se o SampleExtraction e o ComputeStatisctis que consiste em extrair os valores das amostras da imagem e determinar a média global e o desvio padrão para cada banda da imagem em questão, respetivamente. De seguida, realizou-se a etapa TrainVectorClassifier que consiste em treinar o conjunto de treino com base no algoritmo de Machine Learning escolhido, para mais tarde, no ImageClassifier se preceder à classificação da imagem de entrada. Deste método, resultam a imagem classificada e ainda dois mapas: Confidence Map e o Probability Map, que exprimem a confiança do algoritmo na classificação da classe da imagem e a probabilidade que foi atribuída à classe que prevaleceu.

Finalmente, procedeu-se ao ClassificationMapRegularization no qual se filtrou a imagem gerada no processo anterior. Utilizando um ficheiro de referência, atribuiu-se cor à imagem filtrada (ColorMapping) e utilizando o conjunto de amostras definidas para validação gerou-se a Confusion Matrix para analisar os resultados da classificação.

2.5 Metodologia

O processamento deste projeto dividiu-se nas fases descritas na figura 3:

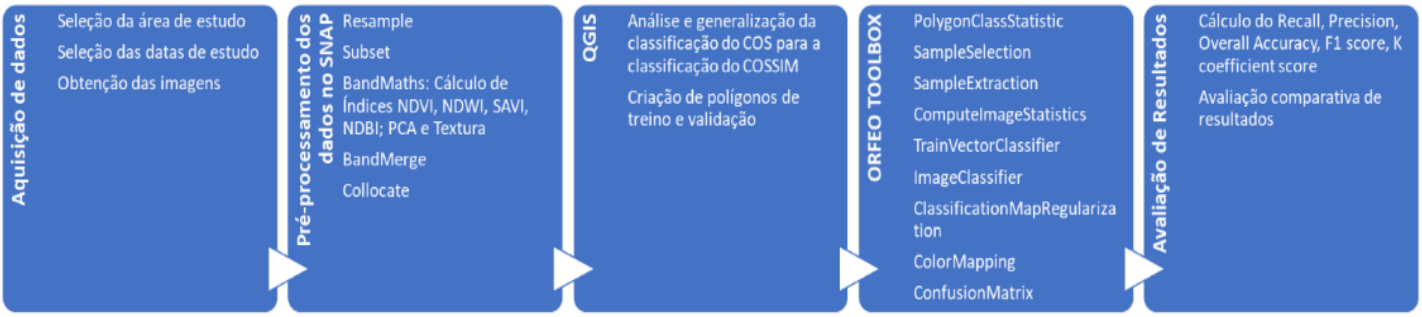

Figura 3 - Metodologia

O processamento no Orfeo Toolbox, é apresentado no seguinte esquema (figura 4) de uma forma mais detalhada:

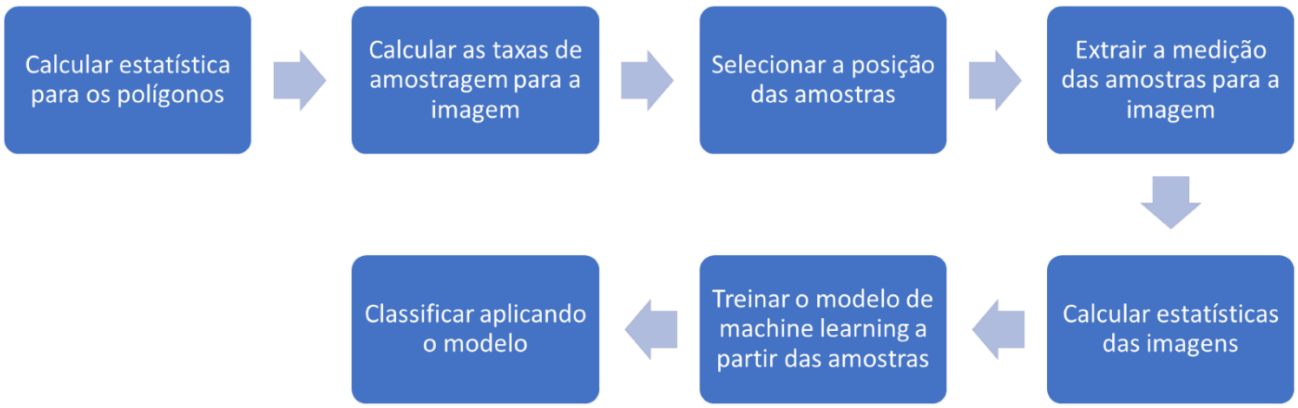

Figura 4 - Processamento no OrfeoToolbox

2.6 Avaliação da precisão

Para verificar qual o método e a técnica mais precisa, recorremos à Confusion Matrix que compara, numa amostra de pixels, o resultado da classificação efetuada pelo algoritmo com a verdadeira ocupação do solo. Aqui, introduzimos o conceito de precisão - determina a proporção de verdadeiros positivos no universo de todos os positivos detetados pelo modelo, incluindo os falsos positivos – e a revocação – determina a proporção dos verdadeiros positivos no universo de positivos que deviam ter sido identificados, incluindo falsos positivos.

Como métodos de avaliação utilizou-se o overall accuracy que consiste na percentagem de pixéis corretamente classificados. (5)

$$overall\ accuracy = \frac{CC}{TS}$$

Sendo CC o número de amostras corretamente classificadas e TS o número total de amostras.

Foram ainda utilizados o K Coefficient e o F1 score. O primeiro consiste numa medida de concordância global da matriz, que, contrariamente ao overall accuracy, não tem apenas em conta as diagonais da matriz como tem em conta, todos os elementos à sua volta. (6)

$$K = \frac{N\sum_{i=1}^{n} m_{i,i} - \sum_{i=1}^{n} G_i C_i}{N^2 - \sum_{i=1}^{n} G_i C_i}$$

Já o segundo, consiste na média ponderada entre a precisão e a revocação. Este método, ao contrário dos restantes é determinado para todas as classes individualmente. (7)

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

3. Resultados

Em primeiro lugar foram avaliados os resultados da classificação por Random Forest através dos valores de overall accuracy e coeficiente k obtidos nas matrizes de confusão de cada um dos casos, visível na tabela 2.

Tabela 2 - Comparação dos resultados dos diferentes modelos

			Random Forest 200 Trees		Comparação dos modelos	
			OA	K		
Bandas	Com 323	0.7	82%	0.77		
		Total	70%	0.65		
	Sem 323	0.7	82%	0.78	0	0.01
		Total	69%	0.63	-1%	-0.02
Índices	Com 323	0.7	76%	0.7		
		Total	73%	0.68		
	Sem 323	0.7	75%	0.69	-1%	-0.01
		Total	73%	0.68	0	0
PCA	Com 323	0.7	75%	0.68		
		Total	64%	0.58		
	Sem 323	0.7	74%	0.68	-1%	0
		Total	70%	0.64	6%	0.06
Textura	Com 323	0.7	82%	0.78		
		Total	68%	0.62		
	Sem 323	0.7	82%	0.78	0	0
		Total	66%	0.6	-2%	-0.02
Bandas + Índices	Com 323	0.7	81%	0.77		
		Total	70%	0.64		
	Sem 323	0.7	82%	0.78	1%	0.01
		Total	67%	0.62	-0.03	-0.02
Bandas + Índices + Texturas	Com 323	0.7	82%	0.78		
		Total	65%	0.59		
	Sem 323	0.7	83%	0.79	1%	0.01
		Total	67%	0.62	2%	0.03

Na tabela 2 foi feita uma comparação entre os modelos, subtraindo os resultados da classificação com [323] aos da classificação sem [323]. Analisando estes resultados é possível verificar que a presença da classe [323] não tem grande impacto nos resultados gerais dos classificadores, sendo, no entanto, de notar, que o classificador que utiliza a amostra de 0.7 nem sempre consegue identificar todas as classes, como é visível na tabela 3. A influência da classe 323 varia nos resultados entre -3% e +6% na overall accuracy e no coeficiente k. Atendendo a este facto, decidimos utilizar apenas os modelos que não

continham a classe [323] para a fase seguinte, sendo que foram escolhidos dois modelos para a amostra de 0.7 e dois modelos para a amostra total da menor classe.

Os modelos que apresentaram melhores resultados nestes casos foram, para a amostra total da menor classe, a imagem dos índices e a imagem do PCA, e para a amostra de 0.7 do total dos polígonos, a imagem das bandas, índices e texturas e também a imagem que contém apenas as bandas. No caso desta última, embora existissem 3 imagens com valores semelhantes de overall accuracy e coeficiente k, foi escolhida a imagem das bandas pois o caso das bandas e índices era já uma imagem composta tal como o outro caso escolhido e a imagem das bandas apresentava melhores resultados quando comparado com o caso que incluía a classe [323].

A performance dos modelos foi também avaliada em termos do F1 score médio de cada uma das classes, sendo que é possível ver nas tabelas 3 e 4 os casos para os quais cada tipo de imagem e cada tipo de amostragem consegue obter melhores resultados.

Tabela 3 - F1Score médio para os diferentes tipos de amostragem

Classe	0.7 com 323	Total com 323	0.7 sem 323	Total sem 323
[21]	89.59%	83.02%	89.57%	85.26%
[100]	69.87%	76.12%	69.69%	73.20%
[311]	50.07%	22.76%	51.31%	17.48%
[312]	28.27%	55.62%	34.58%	49.26%
[313]	0.00%	2.00%	0.00%	1.73%
[321]	59.46%	69.83%	59.36%	68.85%
[322]	60.09%	12.48%	61.07%	5.47%
[323]	0.00%	16.80%		
[410]	10.89%	27.20%	12.05%	25.32%
[420]	0.00%	13.19%	0.00%	17.39%
[500]	48.89%	68.20%	47.29%	67.77%
[610]	94.85%	91.59%	95.08%	90.74%
[620]	98.06%	97.24%	98.07%	97.51%

Aqui podemos verificar que os melhores resultados gerais foram os que utilizaram a estratégia do mesmo número de amostras, com a menor classe totalmente amostrada nas imagens que incluíam a classe 323. É também visível nesta tabela que o modelo que utiliza 0.7 da amostra tem dificuldade em classificar algumas classes, nomeadamente [313], [323] e [420].

Tabela 4 – F1Score médio dos modelos por tipo de imagem

Classe	Bandas	PCA	Índice	Textura	Bandas + Índice	Bandas + Índice + Textura
[21]	86.01%	84.27%	87.80%	90.42%	86.06%	86.59%
[100]	77.52%	67.68%	66.73%	68.47%	76.57%	76.35%
[311]	32.83%	33.19%	58.36%	27.14%	35.38%	25.52%
[312]	37.12%	32.12%	53.45%	62.83%	17.70%	48.36%
[313]	0.89%	0.00%	0.19%	1.78%	1.24%	1.50%
[321]	76.29%	57.42%	33.47%	72.55%	73.70%	72.81%
[322]	35.35%	31.84%	40.86%	33.47%	32.85%	34.29%
[323]	0.05%	23.17%	27.19%	0.00%	0.00%	0.00%
[410]	27.09%	8.90%	11.57%	18.82%	24.77%	22.04%
[420]	10.23%	0.00%	3.49%	18.90%	5.84%	7.40%
[500]	67.73%	35.51%	30.73%	71.49%	69.89%	72.89%
[610]	98.22%	88.52%	87.08%	88.56%	97.92%	98.07%
[620]	99.75%	94.49%	94.55%	97.87%	99.82%	99.85%

Por análise da tabela 4, podemos verificar para quais das imagens obtemos o melhor resultado por classe. Nela é de notar que a classe [323], que tinha muito poucas amostras, não é identificada de todo nas imagens da textura, ou nas imagens de juncão de dois ou mais tipos, no entanto consegue ser identificada no PCA e no Índice. A classe [313] embora tenha um grande número de amostras não é facilmente identificada em nenhum dos modelos. As classes [323], [410] e [420] são também de difícil identificação.

Os quatro casos que foram escolhidos por análise da tabela 2, foram depois utilizados com os algoritmos de machine learning Naive Bayes e Decision Tree e os seus resultados em termos de coeficiente k e overall accuracy visíveis na tabela 4:

Tabela 5 - Overall accuracy e coeficiente K obtidos para cada método e para cada tipo de imagem.

		Random Forest		**Decision Tree**		**Naïve Bayes**	
		OA	**K**	**OA**	**K**	**OA**	**K**
Bandas	0.7	82%	0.78	75%	0.70	83%	0.8
Índices	Total	73%	0.68	68%	0.62	74%	0.69
PCA	Total	70%	0.64	69%	0.63	67%	0.62
Bandas + Índices + Texturas	0.7	83%	0.79	83%	0.80	83%	0.79

Ao observar a tabela em cima, verifica-se que o Método Naïve Bayes é um melhor classificador para as imagens que apresentam apenas Bandas e Índices. No entanto, o Random Forest é um melhor classificador para imagens apenas com o PCA e a imagem constituída pelas Bandas, Índices e Texturas apresenta resultados muito próximos em todos os métodos, sendo o Decision Trees ligeiramente mais preciso.

Tabela 6 - F1 scores obtidos para diferentes métodos e amostragens.

	Bandas + Índices + Texturas			Índices		
	DT	**RF**	**Bayes**	**DT**	**RF**	**Bayes**
[21]	89.93%	91.67%	88.70%	76.73%	89.18%	75.19%
[100]	79.88%	71.04%	78.27%	67.60%	54.08%	72.21%
[311]	63.83%	53.39%	59.57%	53.37%	58.32%	53.93%
[312]	43.71%	35.22%	88.66%	44.20%	42.33%	81.66%
[313]	1.19%	0.00%	2.50%	0.59%	0.00%	1.43%
[321]	77.67%	75.33%	86.87%	73.82%	60.84%	73.75%
[322]	73.95%	69.89%	66.35%	47.27%	30.29%	66.29%
[410]	30.51%	20.22%	59.87%	19.02%	21.48%	28.30%
[420]	0.00%	0.00%	0.00%	16.97%	0.00%	19.74%
[500]	56.66%	65.70%	66.12%	71.15%	52.49%	61.77%
[610]	97.93%	98.90%	94.13%	81.96%	82.42%	89.84%
[620]	99.24%	99.89%	97.32%	90.86%	94.16%	95.43%
Média	59.54%	56.77%	65.70%	53.63%	48.80%	59.96%

	Bandas			PCA		
	DT	**RF**	**Bayes**	**DT**	**RF**	**Bayes**
[21]	73.58%	88.63%	85.39%	78.39%	80.08%	60.84%
[100]	73.72%	75.02%	76.39%	69.29%	69.56%	76.25%
[311]	52.14%	53.11%	54.46%	50.77%	46.43%	44.42%
[312]	0.00%	52.35%	86.95%	0.00%	57.01%	64.40%
[313]	0.00%	0.00%	0.58%	0.00%	0.00%	0.28%
[321]	67.95%	79.90%	89.48%	73.30%	68.86%	75.22%
[322]	73.60%	66.94%	76.50%	52.44%	2.52%	65.69%
[410]	38.40%	12.15%	64.12%	31.04%	16.45%	0.00%
[420]	0.00%	0.00%	64.55%	3.96%	0.00%	0.00%
[500]	44.46%	60.61%	90.59%	35.44%	43.98%	49.75%
[610]	99.50%	98.69%	96.98%	83.05%	87.20%	79.81%
[620]	99.52%	99.74%	96.94%	90.31%	93.94%	0.00%
Média	51.90%	57.26%	73.58%	47.33%	47.17%	43.06%

Ao observar a tabela 6, verifica-se que o algoritmo de Bayes é o que apresenta melhor média de classificação em três dos quatro casos, para a classificação das classes, no entanto, à semelhança dos outros algoritmos, não consegue classificar todas as classes. Aqui verifica-se que RF é, em geral, um bom classificador para a agricultura e para a água.

As imagens resultantes das classificações foram posteriormente coloridas com recurso a uma tabela de correspondência, e os resultados de algumas zonas de interesse na região do estuário do Sado estão apresentadas nas figuras 5-7.

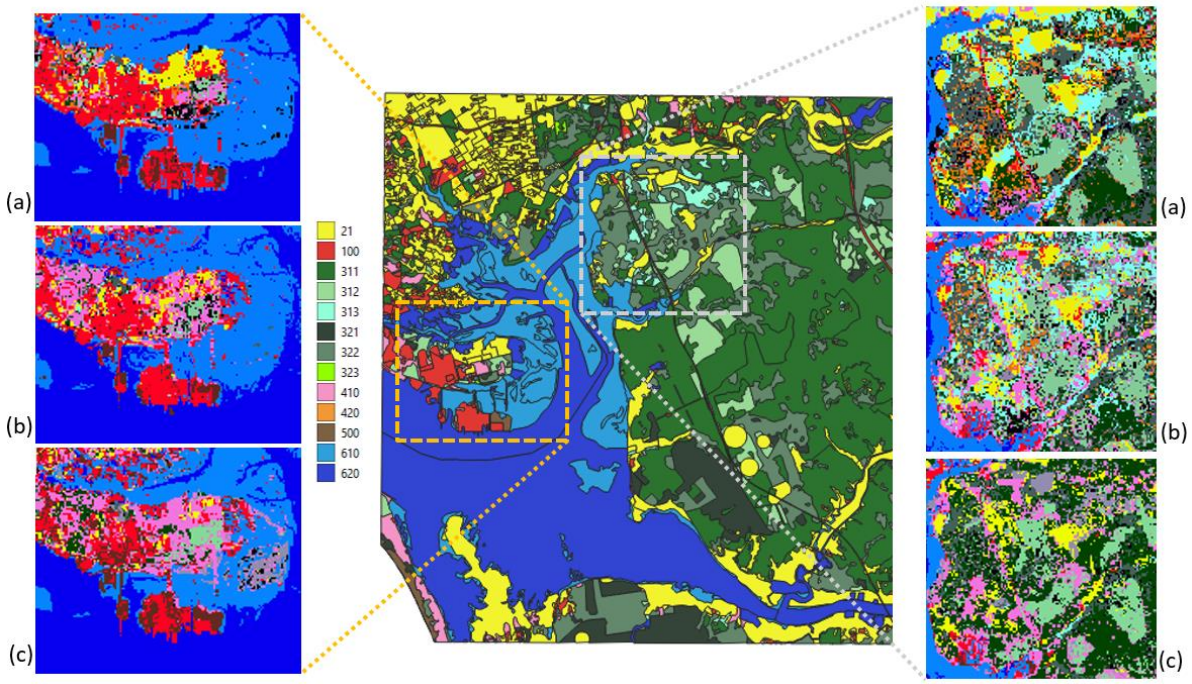

Figura 5 - Classificação da imagem de índices, onde (a) - Bayes, (b) - Decision Tree, (c) – Random Forest

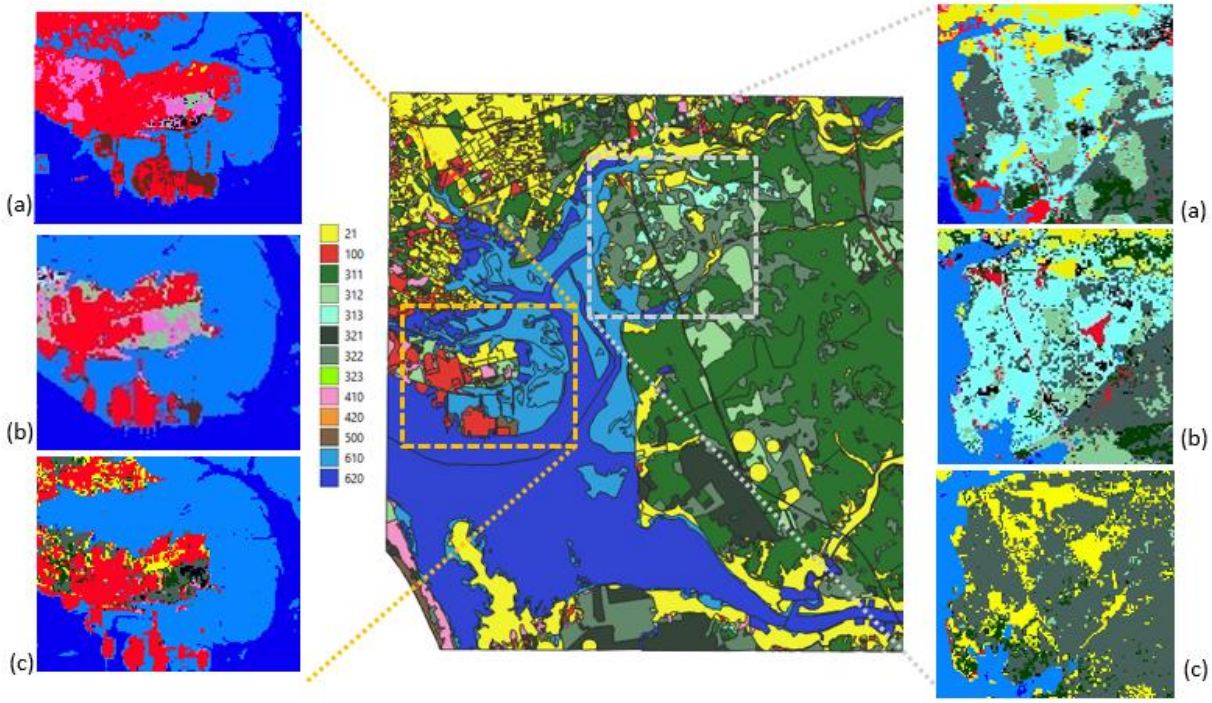

Figura 6 – Classificação da imagem de bandas, onde (a) - Bayes, (b) - Decision Tree, (c) – Random Forest

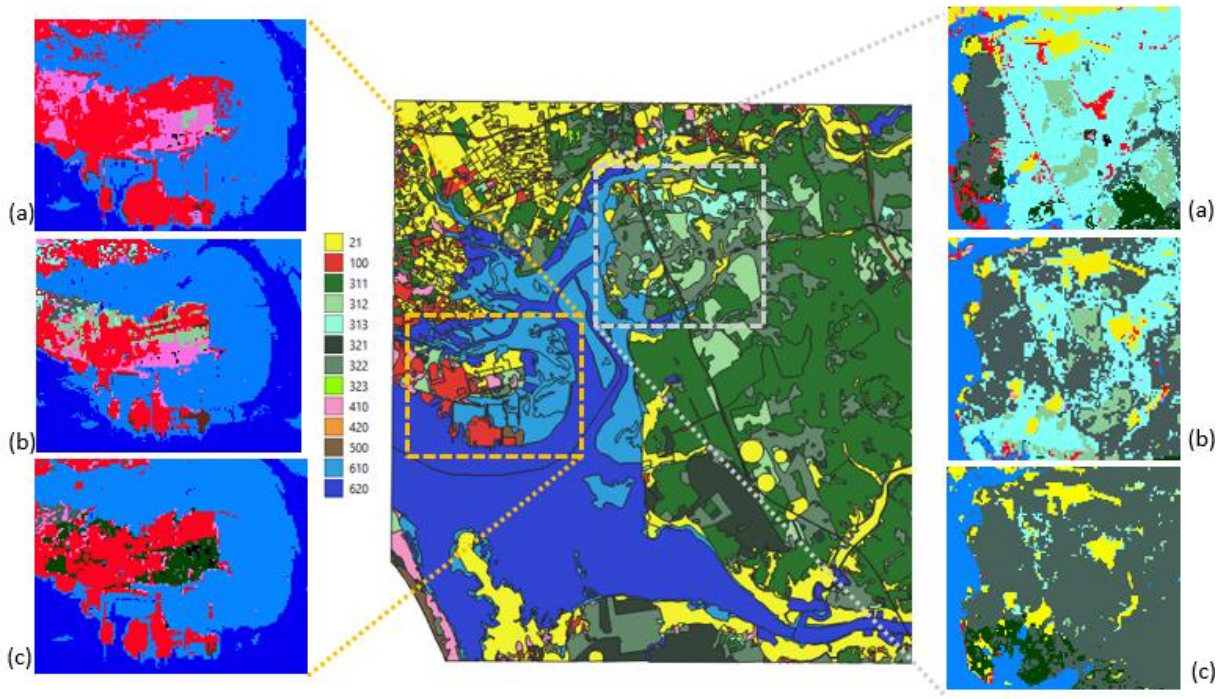

Figura 7 - Classificação da imagem de bandas + índices + texturas, onde (a) - Bayes, (b) - Decision Tree, (c) – Random Forest

Os vários métodos de classificação apresentam resultados variados, sendo que alguns deles apresentam perdas de informação, por exemplo na região do lado esquerdo existe a norte uma região agrícola (21), que é bem classificada apenas com o algoritmo de Bayes para os índices e com o algoritmo de Random Forest para as bandas, sendo que esta classe desaparece completamente na imagem que combina bandas, índices e texturas. As imagens classificadas poderiam necessitar ainda de alguma suavização para terem melhor correspondência com os polígonos do COS, pois estes são gerados com regras de áreas que não são observadas aquando da classificação.

4. Discussão

4.1 Discussão dos resultados

Observando as tabelas anteriores para os resultados obtidos por Random Forest, verifica-se que as precisões obtidas com a classe 323 e as precisões obtidas sem a mesma classe são muito semelhantes, no entanto há uma ligeira melhoria quando não existe a classe 323. O método de escolha de amostragem é bastante significativo pois utilizando a estratégia do mesmo número de amostras, com a menor classe totalmente amostrada, apresenta, na sua grande maioria, resultados mais satisfatórios nas classificações das classes, avaliadas pelo F1 score, mas apresenta piores resultados a nível global, sendo os valores de OA e K menores do que para a amostragem de 0.7 dos polígonos. Já o conteúdo das imagens é de extrema relevância pois o facto de a imagem ser constituída por Banda-Índices-Texturas apresenta melhores resultados quando comparados com as restantes imagens, no entanto é de destacar que esta imagem não consegue classificar as classes [313] e [420].

Aplicando os dois outros algoritmos – Decision Trees e Bayes – às imagens que obtém melhor classificação com o algoritmo Random Forest verifica-se que o Bayes consegue classificar, numa forma geral, melhor as classes quando comparado com os outros. O melhor resultado foi obtido utilizando o algoritmo de Bayes para a imagem que continha

apenas as bandas, com um OA de 83% e K de 0.8, sendo o seu F1 score médio de 73.58%. No entanto, o grande fator que influencia o F1 score neste caso é a classe [313], onde este valor é apenas de 0.58%. Tendo em conta que existiam 100 polígonos de treino para esta classe, a sua fraca identificação é provavelmente fruto do facto de se tratar de uma classe agrupadora ("Outras folhosas") que poderá ter composições radiométricas diferentes juntas numa mesma classe e por isso seja difícil para o algoritmo calcular um modelo para esta classe.

Apesar de, para cada tipo imagem, existir um método com precisão de 83% que classifica bem os pixéis, continuam a existir classes que não são classificadas ou que apresentam uma baixa taxa de classificação, fazendo com que o algoritmo não seja muito fiável. Contudo, há classes que são sempre classificadas pelo mesmo algoritmo com melhor precisão, nomeadamente a classe da agricultura e da água pelo Random Forest.

4.2 Razões para melhoria

Após a elaboração deste estudo averiguámos que se poderia ter aplicado outros algoritmos de Machine Learning nomeadamente o Convolutional Neural Network (CNN) uma vez que em estudos anteriores, este método é o que apresenta melhores resultados.

Seria interessante, ainda, avaliar os resultados das mudanças entre as características de cada modelo nomeadamente, diferentes árvores de decisão, aumento ou diminuição do número de árvores e também profundidade das árvores no Random Forest. A nível das técnicas de amostragem, poderiam ter sido ainda utilizadas técnicas diferentes, como por exemplo reduzindo as percentagens de polígonos amostrados ou um número fixo com distribuição proporcional ao conteúdo da imagem.

4.3 Limitações do estudo

A principal limitação neste estudo foi o esforço computacional. Devido a esta limitação, o grupo teve de reduzir a sua área de estudo o que, por consequência, fez com que existisse pouca significância de determinadas classes – sendo o caso da 323. Outro problema relacionado com o esforço computacional foi a limitação na escolha dos algoritmos. Foi o caso do Support Vector Machine (SVM) e do K-Nearest Neighbor (KNN) que não foi possível correr devido ao número de horas de processamento.

O modelo que apresenta melhor precisão não foi aplicado a outra área de estudo, como tal, não se consegue prever o seu comportamento em áreas com características diferentes.

5. Conclusão

Com este estudo pretendeu realizar-se a classificação do solo através do nível III do COS2018 combinando diversos modelos de classificação com diferentes formas de amostragem. No entanto, dado o facto do esforço computacional ter sido extremamente elevado, foi necessário reduzir a área de estudo, fazendo com que os métodos não classificassem todas as classes dada a pequena amostragem. Como tal, pressupõe-se que os modelos seriam mais eficientes para o COS2018 nível II pelo facto de ser mais geral.

Dos resultados que foram possíveis obter, a melhor classificação foi conseguida através do algoritmo Naive Bayes para uma imagem contendo só as bandas, que atingiu um overall accuracy de 83%, K de 0.8 e F1 score médio de 73.58%.

.

Referências

(1) Huang, B.; Zhao, B.; Song, Y. Urban Land-Use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery. *Remote Sensing of Environment* **2018**, *214*, 73–86. https://doi.org/10.1016/j.rse.2018.04.050.

(2) Alem, A.; Kumar, S. Deep Learning Methods for Land Cover and Land Use Classification in Remote Sensing: A Review. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* **2020**, 903–908. https://doi.org/10.1109/ICRITO48877.2020.9197824.

(3) Pal, M. Random Forest Classifier for Remote Sensing Classification. *http://dx.doi.org/10.1080/01431160412331269698* **2007**, *26* (1), 217–222. https://doi.org/10.1080/01431160412331269698.

(4) Punia, M.; Joshi, P. K.; Porwal, M. C. Decision Tree Classification of Land Use Land Cover for Delhi, India Using IRS-P6 AWiFS Data. *Expert Systems with Applications* **2011**, *38* (5), 5577–5583. https://doi.org/10.1016/J.ESWA.2010.10.078.

(5) Diengdoh, V. L.; Ondei, S.; Hunt, M.; Brook, B. W. A Validated Ensemble Method for Multinomial Land-Cover Classification. *Ecological Informatics* **2020**, *56*, 101065. https://doi.org/10.1016/J.ECOINF.2020.101065.

(6) Upadhyay, A.; Shetty, A.; Kumar Singh, S.; Siddiqui, Z. *Land Use and Land Cover Classification of LISS-III Satellite Image Using KNN and Decision Tree; Land Use and Land Cover Classification of LISS-III Satellite Image Using KNN and Decision Tree*; 2016.

(7) Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. **2015**. https://doi.org/10.48550/arxiv.1508.00092.

(8) Direção-Geral do Território. *Especificações Técnicas Da Carta de Uso e Ocupação Do Solo (COS) de Portugal Continental Para 2018*; 2019.

(9) Caetano, M.; Igreja, C.; Marcelino, F.; Costa, H. Estatísticas e Dinâmicas Territoriais Multiescala de Portugal Com Base Na Carta de Uso e Ocupação Do Solo (COS). **2017**.