Desenvolvimento de algoritmos paralelizáveis para a atualização da COSsim

Contrato de cooperação 261/2021 entre a DGT e o ISA celebrado no âmbito do projeto POCI-05-5762-FSE-000368

Início: 23 de julho de 2021

Entregável 4: Fundamentos para um algoritmo de deteção de alterações de ocupação do solo

15 de novembro de 2022

André Alves; Bruno Barbosa; Daniel Moraes; Manuel Campagnolo

Código e outros elementos:
https://github.com/manuelcampagnolo/Contrato_ISA_DGT_261_2021

Índice

Abreviaturas:

BDR_TNE_300	Base de dados de referência, mesmo que DGT300
CCDC	Continuous Change Detection and Classification
COSsim	Carta de ocupação do solo conjuntural
FN	Falso negativo
FP	Falso positivo
GeoDataFrame	Estrutura de dados que representa dados geográficos como uma tabela com uma coluna designada “geometry” (pode conter por exemplo WKT)
DGT300	Base de dados de referência
NDVI	Índice de vegetação
UA	Unidade Amostral
VN	Verdadeiro negativo
VP	Verdadeiro positivo
WKT	Well Known Text: string para representar objetos geométricos vetoriais

1 Enquadramento

O presente relatório corresponde à tarefa 4 do contrato de cooperação entre a DGT e o ISA que prevê os seguintes entregáveis:

- E4.1 – Algoritmo final para deteção de alterações de ocupação do solo
- E4.2 – Algoritmo final para classificação de ocupação do solo

O código dos algoritmos em formato notebook de Python (ficheiro principal “main” e ficheiro com funções auxiliares) está disponível no seguinte repositório, assim como os relatórios anteriores:

https://github.com/manuelcampagnolo/Contrato_ISA_DGT_261_2021

Principais conclusões:

1. Com uma boa parametrização do método CCDC é obtêm-se uma precisão de 80% na deteção de alterações da ocupação do solo, com erros de comissão (13.55%) superiores aos erros de omissão (5.69%).
2. O resultado acima sugere que o algoritmo de classificação pode apenas rever as classificações dos pixels que são identificados como tendo alteração pelo CCDC, com erro de 20%.
3. Os coeficientes CCDC são uma alternativa à utilização dos perfis de NDVI para a caracterização das principais classes de ocupação do solo.

2 Base de dados de referência de polígonos DGT300

Os testes sobre deteção de alterações e classificação da ocupação do solo foram realizados com base num conjunto de dados de referência delineados no quadro do presente projeto e elaborados por uma equipa da DGT.

A base de dados de referência (BDR_TNE_300, que será designada simplesmente por DGT300 ao longo do relatório) consiste em Unidades Amostrais (UAs) recolhidas na área do tile 29TNE dentro um raio de 200m ao redor de 300 UAs centrais. A informação de referência foi preenchida pela equipa da DGT por meio de fotointerpretação. A construção foi feita de forma que metade das UAs centrais correspondesse a exemplos em que há potencial alteração na ocupação e a outra metade a exemplos em que potencialmente não há alteração. Das 150 UAs centrais correspondentes a potenciais alterações, 100 referem-se a potenciais alterações identificadas no período compreendido entre os anos agrícolas de 2018 a 2020 e 50 UAs centrais a potenciais alterações durante o ano agrícola de 2021 (Tabela 1).

Tabela 1: Distribuição das unidades amostrais e respectivos períodos de referência.

	Período de referência	**Nº de UAs centrais**
Potencial alteração na ocupação	2018 - 2020	100
	2021	50
Sem alteração	2018 - 2021	150

O desenho da base de dados de referência visou incluir 150 unidades amostrais com e 150 unidades sem alteração na ocupação. Considerou-se que uma unidade amostral corresponde a uma alteração na ocupação quando há evidência de alteração, obtida através da comparação de produtos cartográficos e/ou do resultado da aplicação do CCDC.
As unidades amostrais que referem-se a potenciais alterações no período entre os anos agrícolas de 2018 a 2020 foram geradas a partir de uma combinação de processos. Em primeiro lugar, foram identificados os pixels em que houve perda de vegetação no período, com auxílio da COSsim de 2018 e 2020. Foram consideradas apenas as classes de ocupação do solo (Eucalipto, Pinheiro Bravo e Matos) e transições (para Matos, Vegetação Herbácea ou Superfície sem Vegetação) de interesse a este projeto, conforme ilustrado na Figura 1. A seguir, foram eliminados os grupos de pixels contiguos com área inferior a 0.5 ha, havendo a seguir um processo de erosão para a remoção de pixels na bordadura dos grupos. Além disso, foi computado o resultado do CCDC para os pixels restantes. Pixels em que o CCDC não indicou alteração na ocupação no período e pixels com magnitude da quebra positiva (i.e. indicando haver ganho de vegetação) foram excluídos. Dentre os pixels restantes foram, então, recolhidas 100 UAs através de uma amostragem aleatória estratificada por faixas de magnitude da quebra referente ao NDVI. Os estratos são apresentados na Tabela 2 abaixo.

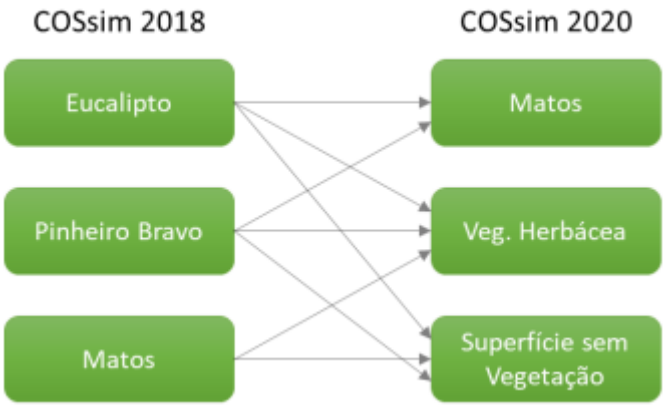

Figura 1 Classes de ocupação e transições consideradas na amostragem referente a 2018-2020.

Tabela 2: Estratos utilizados na amostragem estratificada.

Estrato	**Magnitude da quebra (NDVI x 10.000)**
1	(-10.000, -6.000]
2	(-6.000, -4.000]
3	(-4.000, -2.000]
4	(-2.000, 0)

As unidades amostrais consideradas sem alteração foram obtidas considerando os pixels em que a classe da COSsim de 2018 e 2020 se manteve igual. Além disso, foram excluídos pixels em que o CCDC detetou alteração no período do ano agrícola de 2021 (Outubro/2020 a Setembro/2021). De forma semelhante às amostragens anteriores, foram implementados os processos de remoção de grupos de pixels com área inferior a 0,5 ha e erosão da bordadura. Ao final, foram selecionadas 150 UAs através de uma amostragem aleatória.

A equipa da DGT foi responsável pela interpretação da amostra, a qual foi efetuada através de dados base e auxiliares. A interpretação foi realizada considerando duas componentes: espacial e atributos. Foram desenhados polígonos correspondentes à manchas de alteração dentro da área circundante às UAs centrais (raio de 200m). A cada polígono foi associada informação referente a diversos atributos, visando caracterizar as eventuais alterações na ocupação em termos de data e tipo da alteração e classes de ocupação antes e depois da alteração. A Tabela 3 apresenta a descrição dos atributos. A informação foi disponibilizada em formato vetorial.

Tabela 3: Atributos dos polígonos da base de dados de referência. As “classes COSsim” são 13 classes, (13 classes) , e informação adicional (e.g. “escura”, “clara”) do analísta.

Atributo	**Descrição**
ID	Identificador único para cada polígono
buffer_ID	Identifica qual o *buffer* a que cada polígono está associado
altera	Indicação da presença ou ausência de alterações
tipo_1	Indica o tipo de alteração (corte, fogo, agricultura, água)
classe_0	Classe de uso do solo no momento anterior à primeira alteração, com base nas classes da COSsim
data_0	Corresponde à data do momento anterior à primeira alteração
classe_1	Classe de ocupação do solo no momento posterior à primeira alteração registada, com base nas classes da COSsim
data_1	Corresponde à data do momento posterior à primeira alteração
tipo_2	Indica o tipo da segunda alteração, caso tenha ocorrido
classe_2	Classe de uso do solo no momento anterior a uma segunda alteração, com base nas classes da COSsim
data_2	Corresponde à data do momento anterior a uma segunda alteração
class_3	Classe de uso do solo no momento posterior a uma segunda alteração, com base nas classes da COSsim
data_3	Corresponde à data do momento posterior a uma segunda alteração
classe2018	Corresponde à classe presente no polígono em setembro de 2018 com base nas classes da COSsim
classe2019	Corresponde à classe presente no polígono em setembro de 2019
classe2020	Corresponde à classe presente no polígono em setembro de 2020
classe2021	Corresponde à classe presente no polígono em setembro de 2021

area	Área dos polígonos em metros quadrados
notas	Identifica dúvidas, observações ou casos específicos

3 Bordaduras de polígonos e “join” espacial para cruzar dados de referência com dados de satélite e outros dados auxiliares

Foram combinados os dados derivados de Sentinel-2, e em particular os resultados da análise da sequência temporal de imagens Sentinel-2 através da técnica de análise CCDC, com a informação de referência contida nos polígonos gerados pela DGT que representam o resultado da avaliação feita por analistas nas áreas dos 300 buffers de análise.

Para minimizar efeitos de transição entre as classes de uso e ocupação do solo foi gerado um buffer negativo de 10 m em todas as features existentes. A área que corresponde a este buffer negativo foi chamada de bordadura (valor 1), o que estava fora deste limite foi entendido como centro do polígono (valor 0). Os pixels com valor 0 têm portanto maior fiabilidade de ocupação, pois está mais distante das transições. A Figura 2 apresenta um dos buffers depois de processar o buffer negativo.

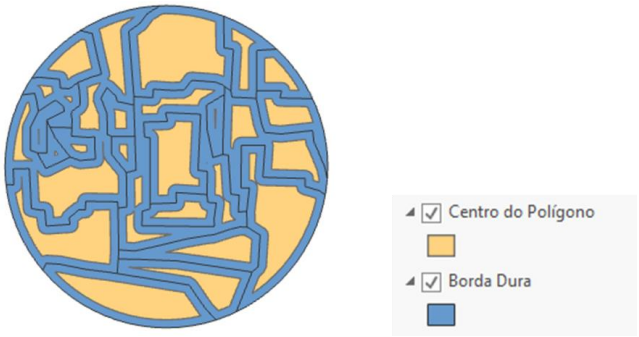

Figura 2. Bordaduras identificadas

Posteriormente foi realizado um ‘spatial join’ que consiste em unir os atributos existentes de uma “feature” a outra com base nas suas relações espaciais, como apresentado na figura abaixo (Figura 3).

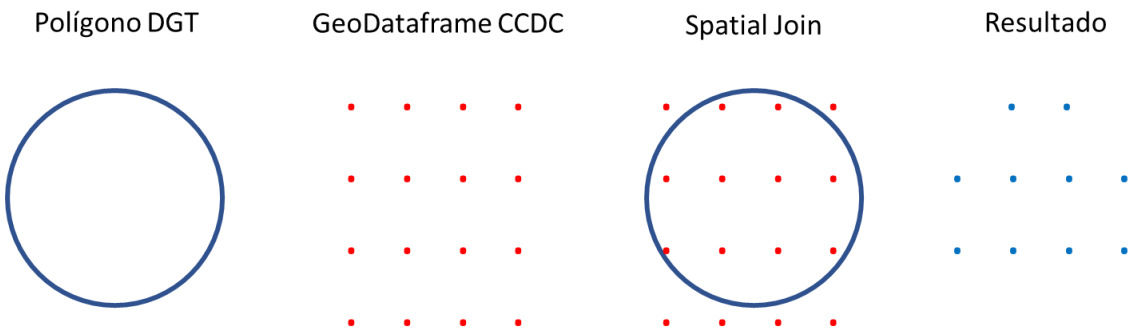

Figura 3. Funcionamento do Spatial Join

O 'Resultado' apresenta apenas as geometrias que estão espacialmente relacionadas, porém contendo todos os atributos das features utilizadas com input. É de realçar que o resultado do CCDC é utilizado em formato GeoDataFrame, que é um data frame (tabela) com uma coluna designada "geometry" que contém informações de geometria e posição espacial. Os valores associados à geometria correspondem ao centróide do pixel do raster. A partir desta informação é possível realizar o 'spatial join' com os polígonos de validação. Neste GeoDataFrame cada linha corresponde a um segmento temporal que se termina ou (1) por um 'break' identificado pelo CCDC com a informação gerada pelo algoritmo, ou (2) pela observação mais recente disponível (segmento final). Por exemplo: 'tBreak' data da identificação de uma alteração, 'probChange' probabilidade de alteração identificada ou 'ndvi_magnitude' a magnitude da variação de uma determinada banda, neste caso NDVI.

Ao realizar o 'spatial join', toda a informação existente nos polígonos é adicionada ao GeoDataFrame do CCDC resultando em um novo GeoDataFrame com mais de 120 colunas.

4 Falsos e verdadeiros positivos e negativos na deteção de alterações pelo CCDC

O objetivo é comparar os resultados do CCDC com os dados de referência dos analistas gerada pela equipa da DGT e, em particular, a data da deteção. Neste processo utilizou-se a informação dos campos 'data_1' e 'data_3' (quando existente) da base de dados e o campo 'tBreak' do CCDC, como forma de comparar a distância temporal em que o modelo e os analistas identificaram uma alteração na ocupação do solo.

A definição de FP, FN, VP, VN, depende de um parâmetro temporal (theta) que indica a diferença de dias que se considera aceitável para uma deteção a partir da série temporal de dados Sentinel-2 ser um verdadeiro positivo.
Desta forma foi possível obter quatro resultados: os verdadeiros positivos 'VP' foram identificados quando a diferença entre a data de alteração identificada pelo analista e a data do 'break' do modelo estava entre +/- 60 dias (para *theta=60*, o que pode ser modificado). Os falsos positivos 'FP' ocorrem quando o CCDC identifica um break, porém este está a mais de *theta* dias da alteração identificada pelo analista ou o analista não identifica uma alteração.

Os falsos negativos ‘FN’ ocorrem quando o CCDC não identifica o break e o analista indica uma alteração e, por fim, os verdadeiros negativos ‘VN’ são quando nem o CCDC nem o analista identificam alteração. Podem ocorrer simultaneamente um FP e um FN quando ambos CCDC e analista indicam alteração, mas com diferença de datas maior do que theta dias.

Ressalva-se que nos casos em que o modelo identificou mais de um ‘break’ foram necessários ajustes na contagem das métricas. No exemplo abaixo o CCDC identificou 3 ‘breaks’ e os analistas indicaram duas alterações. Neste caso, o primeiro ‘break’ identificado estava mais próximo da ‘data_1’ identificada pela DGT (que nomeamos como ‘A’) porém há mais de *theta* dias o que resultou em um FP. O segundo ‘break’ identificado também estava mais próximo da ‘data_1’ com uma diferença de 1 dia, o que gerou um VP. O terceiro ‘break’ identificado estava mais próximo da ‘data_3’ (que nomeamos como ‘B’), porém também com uma distância maior que *theta*, o que resultou em um FP, entretanto aqui também é gerada uma penalização para o modelo por não ter identificado a alteração na data do analista, o que gera um FN neste mesmo ‘break’.

buffer_ID	IDCCDC	tBreak	data1_z	analistas	nome	Valid_breaks	delta_min	VP	FP	FN	VN
232	1762	2018-10-22 11:28:30.000	2020-01-05	2	A	3	439.0	0.0	1.0	0.0	0.0
232	1762	2020-01-05 11:30:09.741	2020-01-05	2	A	3	1.0	1.0	0.0	0.0	0.0
232	1762	2021-03-20 11:30:14.049	2020-08-22	2	B	3	211.0	0.0	1.0	1.0	0.0

Figura 4. VP, FP, FN, e VN na DataFrame

5 Descrição das colunas da data frame que contém toda a informação disponível para todos os pixels de DGT300

A estrutura do dado em formato GeoDataFrame permite que sejam feitas uma série de “queries” para filtragem da informação desejada. Apesar de o número de colunas existentes no resultado ser variável de acordo com os parâmetros de entrada do CCDC, algumas colunas estarão sempre presentes do dado final. A Figura 5 apresenta algumas das colunas que sempre estarão presentes. As colunas apresentadas indicam o ID do buffer em que está presente o break identificado pelo CCDC (para DGT300 terá um valor entre 1 e 300), também contém a coordenada do ponto em questão, a magnitude da queda de NDVI e o erro médio encontrado do segmento (RMSE) e a data da alteração encontrada.

	buffer_ID	IDCCDC	coord_ccdc	ndvi_magnitude	ndvi_rmse	tBreak
0	78	13	(39.86637890970445, -7.83829492734909)	-6411.363501	617.631867	2021-02-23 11:30:14.521
1	78	14	(39.86637890970445, -7.838205095820679)	-6407.675099	597.257390	2020-10-11 11:30:18.570
2	78	15	(39.86637890970445, -7.8381152642922665)	-6226.986696	581.358844	2020-10-11 11:30:18.570
3	78	16	(39.86637890970445, -7.838025432763855)	-5764.501145	604.416572	2021-02-23 11:30:14.521
4	78	17	(39.86637890970445, -7.837935601235443)	-5764.501145	604.416572	2021-02-23 11:30:14.521

Figura 5. Colunas do Data Frame

Para todas as bandas dadas de entrada como parâmetros do CCDC obteremos a magnitude, o erro médio e sete coeficientes que permitem gerar as curvas harmônicas associadas ao ponto. Abaixo a representação destas colunas para o NDVI (Figura 6). O INTP indica o 'intercept' ou coeficiente linear, enquanto o SLP é referente ao 'slope' a declividade da reta, o COS e SIN indicam o cosseno e seno associados a curva gerada para o intervalo anual, os mesmos índices estão posteriormente com os valores 2 e 3 que indicam estes mesmos valores para os intervalos semestral e trimestral, respetivamente.

	buffer_ID	IDCCDC	ndvi_INTP	ndvi_SLP	ndvi_COS	ndvi_SIN	ndvi_COS2	ndvi_SIN2	ndvi_COS3	ndvi_SIN3
0	78	13	9483.326349	-6.863116e-10	482.568637	0.000000	47.286702	0.0	0.0	0.0
1	78	14	8556.324141	0.000000e+00	605.358526	-37.925921	141.209793	0.0	0.0	0.0
2	78	15	8446.487048	0.000000e+00	692.959677	-7.311685	138.181949	0.0	0.0	0.0
3	78	16	7945.255422	0.000000e+00	902.534953	-54.890445	141.861511	0.0	0.0	0.0
4	78	17	7945.255422	0.000000e+00	902.534953	-54.890445	141.861511	0.0	0.0	0.0

Figura 6. Colunas com os coeficientes do Data Frame. Cada linha representa um segmento temporal do sinal, sem quebras segundo o CCDC

Figura 5: Colunas com os coeficientes do Data Frame. Cada linha representa um segmento temporal do sinal, sem quebras segundo o CCDC.

Com o intuito de aprimorar a análise e permitir a progressão para os passos futuros do trabalho é possível de adicionar informação extra à tabela Adicionamos os dados referentes às classes de ocupação das COSsims de 2018, 2020 e 2021 (Figura 7). Isso permite, por exemplo, verificar os pontos que são verdadeiros negativos e as COSsims apresentam a mesma classe nas três versões.

	buffer_ID	IDCCDC	ndvi_magnitude	VN	COSSIM18_H	COSSIM20_H	COSSIM21_H
193005	207	1069	0.0	1.0	312	312	312
508506	250	585	0.0	1.0	321	321	321
470672	137	1431	0.0	1.0	410	410	410

Figura 7. Colunas com valores da COSsim

Outro exemplo seria verificar os locais indicados como verdadeiros positivos, suas respectivas classes nas COSsims e a informação apresentada pelo analista e magnitude de queda do NDVI menor que 6000 (Figura 8).

	buffer_ID	IDCCDC	ndvi_magnitude	tBreak	VP	COSSIM18_H	COSSIM20_H	COSSIM21_H	classeAnterior	classeAtual
185973	263	306	-6532.873092	2021-03-25 11:30:13.148	1.0	312	312	420	Eucalipto	Vegetacao herbacea espontanea
50275	199	526	-6027.169671	2019-12-11 11:30:07.545	1.0	321	500	420	Pinheiro bravo	Superficie sem vegetacao clara
5038	263	600	-7236.881752	2021-04-04 11:30:11.341	1.0	312	312	420	Eucalipto	Vegetacao herbacea espontanea

Figura 8. Colunas com informações da DGT e da COSsim

Desta forma, com apenas alguns comandos simples e filtragens no data frame é possível aceder e explorar os dados de diversas formas e manter a informação completa dentro de uma única tabela que pode ser facilmente pesquisada.

6 Parametrização do CCDC e avaliação da precisão na deteção de alterações

O procedimento da Secção 5 permite identificar os VP, VN, FP e FN para qualquer parametrização de CCDC. Assim, é possível testar quais são os melhores parâmetros do método CCDC, comparando os resultados para diversos conjuntos de parâmetros escolhidos.

Foram realizados testes para avaliar qual combinação de valores dos parâmetros do CCDC resultou em melhores performances. Os experimentos foram feitos considerando o NDVI e as bandas B3 e B12 como as bandas selecionadas para a identificação das quebras (*break point bands*). Entre os diversos parâmetros de entrada do algoritmo, os seguintes foram avaliados:

- Lambda - fator relacionado ao ajuste da regressão LASSO;
- Chi Squared (chi2) - valor limite da probabilidade para detecção de alteração;
- Minimum Years Scaler (minYears) - fator do número mínimo de anos para a aplicação de novos ajustes.

Relativamente ao *lambda*, foram testados os valores de 10, 50, 100, 200 e 300. Para o parâmetro *Chi2*, foram utilizados os valores de 0.99, 0.995 e 0.9975. No que diz respeito ao *Minimum Years Scaler*, os exercícios incluíram os valores de 1 e 1.33. Os diferentes testes e as suas respectivas especificações de parâmetros são exibidos na Tabela 4.

A avaliação da performance levou em conta os acertos (verdadeiros positivos e verdadeiros negativos) e erros (comissão e omissão) do algoritmo na identificação de alterações em cada pixel, tendo como base a informação de referência apresentada anteriormente neste relatório. Para computar os erros e acertos, considerou-se uma margem de theta=60 dias entre a informação de referência e o resultado do CCDC, de modo que diferenças maiores do que 60 dias foram consideradas como erros. Foram excluídos os pixels em regiões de bordadura, uma vez que estes representam regiões de elevada incerteza.

Os resultados dos testes (Tabela 4) indicaram que, relativamente ao parâmetro *minYears*, foram obtidos consistentemente resultados melhores utilizando o valor 1. Optou-se por não testar valores inferiores a 1 uma vez que isto implicaria na possibilidade de haver mais de uma quebra por ano na série temporal. No que diz respeito ao *lambda* (Figura 8) e ao *chi2* (Figura 9), a adoção de valores mais elevados em ambos parâmetros exibiu uma tendência

de redução dos erros de comissão, acompanhada de um aumento dos erros de omissão. Por isso, buscou-se identificar combinações dos referidos parâmetros que resultassem numa menor taxa de erro. De acordo com a tabela, as menores taxas de erro foram obtidas com valores de *lambda* elevados (testes número 6 e 7). Entretanto, apesar da maior exatidão global (percentual de acertos), tais testes apresentaram um aumento significativo no número de omissões. Portanto, buscou-se identificar combinações dos parâmetros *lambda* e *chi* que permitissem alcançar valores satisfatórios em termos de exatidão global, mantendo a taxa de erros de omissão controlada. Dentro desse cenário, os testes número 4 e 8 foram identificados como as alternativas de combinações de parâmetros mais adequadas.

Tabela 4: Especificação dos testes e resultados obtidos sobre os parâmetros de CCDC.

#	chi2	lambda	minYears	VP	FP	FN	VN	Total	Acertos	Erros	Comissão	Omissão
1	0,99	50	1	63.631	57.550	19.269	204.255	344.705	77,71%	22,29%	16,70%	5,59%
2	0,99	100	1	63.507	50.747	19.393	210.783	344.430	79,64%	20,36%	14,73%	5,63%
3	0,995	10	1	62.338	64.549	20.562	197.020	344.469	75,29%	24,71%	18,74%	5,97%
4	**0,995**	**50**	**1**	**63.459**	**51.186**	**19.441**	**209.232**	**343.318**	**79,43%**	**20,57%**	**14,91%**	**5,66%**
5	0,995	100	1	63.015	44.811	19.885	215.449	343.160	81,15%	18,85%	13,06%	5,79%
6	0,995	200	1	60.475	38.992	22.425	221.084	342.976	82,09%	17,91%	11,37%	6,54%
7	0,995	300	1	58.266	35.234	24.634	224.912	343.046	82,55%	17,45%	10,27%	7,18%
8	**0,9975**	**50**	**1**	**63.419**	**46.365**	**19.481**	**212.987**	**342.252**	**80,76%**	**19,24%**	**13,55%**	**5,69%**
9	0,99	50	1,33	63.520	82.168	19.380	188.860	353.928	71,31%	28,69%	23,22%	5,48%
10	0,99	100	1,33	63.786	73.020	19.115	197.467	353.388	73,93%	26,07%	20,66%	5,41%
11	0,995	50	1,33	63.666	74.615	19.235	194.704	352.220	73,35%	26,65%	21,18%	5,46%
12	0,995	100	1,33	63.696	66.389	19.206	202.602	351.893	75,68%	24,32%	18,87%	5,46%

A Tabela 4 indica nomeadamente que o desempenho do CCDC tem uma precisão de 80% na identificação de alterações ou de não alterações em comparação com os dados de referência dos analistas. Com a parametrização escolhida, os erros de comissão (13.55%) são superiores aos erros de omissão (5.69%).

A Figura 9 e Figura 10 mostram como o erro de omissão estimado varia com os restantes dois parâmetros. Nota-se alguma influência do parâmetro lambda e uma influência quase nula do parâmetro chi2.

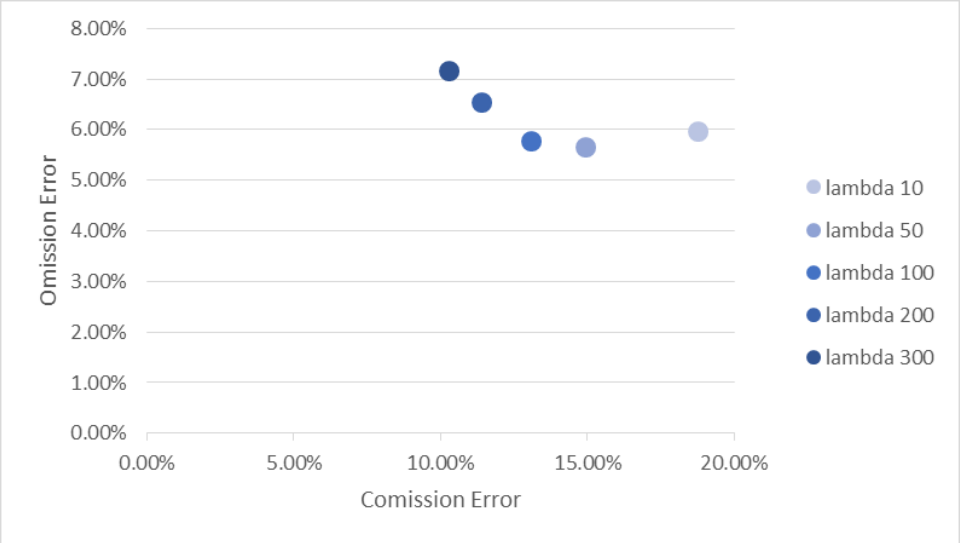

Figura 9. Variação no valor do lambda, com chi2 e minYears constantes

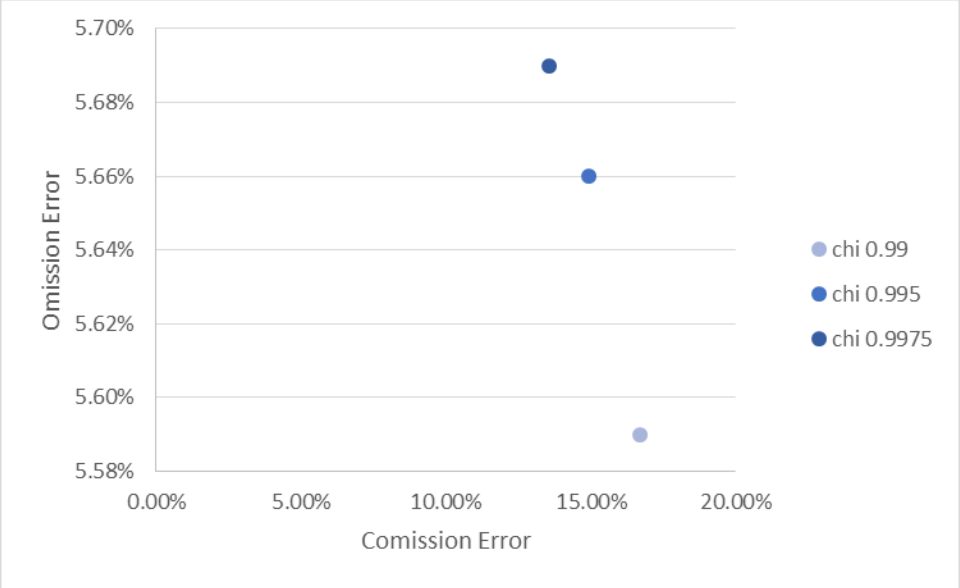

Figura 10. Variação no chi2, com lambda e minYears constantes.

7 Classificação e avaliação do desempenho

Depois de se ter caracterizado o desempenho de CCDC para deteção de alterações (ver secções anteriores), é analisada o potencial do método CCDC para identificar a classe de ocupação do solo após alteração.

Uma forma de avaliar o potencial da proposta de classificador desenvolvido no quadro do presente contrato de cooperação DGT/ISA é compará-lo com a metodologia atual para produção da COSsim que requer a aplicação semi-automática de regras para melhorar o produto intermédio COSsimA e obter o produto final COSsimH, como ilustrado na Figura 11. Nessa comparação podem ser considerados duas componentes: (1) a precisão temática do resultado; e (2) a automatização do processo.

Nas secções seguintes são explorados algumas características do problema de classificação, nomeadamente as transições entre classes de ocupação do solo (Secção 8) e a discriminação das classes no espaços de representação considerados (Secção 9).

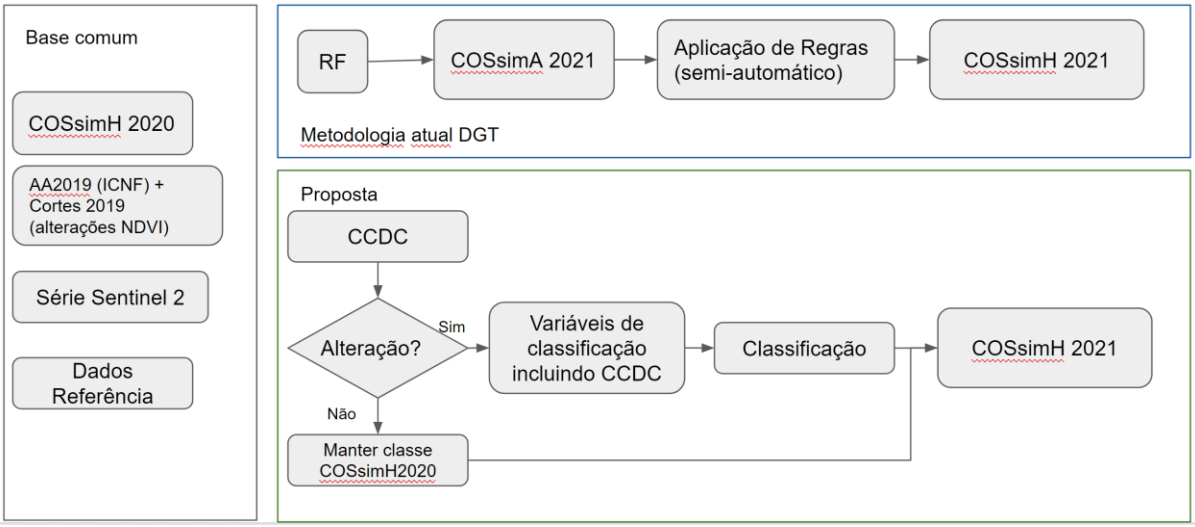

Figura 11. Diagrama simplificado para comparação da abordagem atual e da abordagem proposta.

8 Caracterização de transições de ocupação do solo

8.1 A relação entre alterações de classe na COSsimR e a incerteza da classificação - Portugal Continental

De forma a avaliar a informação associada à incerteza da classificação para cada pixel, realizou-se o levantamento das correções entre a COSsimA2021 (produto da classificação automática de imagem de satélite por *RandomForest*) e a COSsimR2021 (produto resultante da aplicação de regras *if-then-else* combinando conhecimento pericial e dados auxiliares) para Portugal Continental. Por exemplo 16.25% significa que 16.25% da área corrigida de A para R corresponde à transição 710 (sem veg.) para 219 (agri.). As classes são 711 (sem veg), 611 (matos), 312 (herbácea), etc. O valor 62.82 é a média da incerteza do RF.

Tabela 5 - Estatísticas descritivas da incerteza proveniente do Random Forest por tipo de transição (transições até 1% da área)

Alterações de classe entre a COSsimA 2021 e a COSsimR 2021 – Portugal Continental					
Tipo de transição	% área	Média	Desvio-padrão	Percentil 75	Percentil 25
710-210	16.25	62.82	29.94	89	39

710-312	14.12	64.38	28.12	88	45
611-210	11.17	68.90	26.56	91	52
611-312	10.76	69.35	22.95	89	53
312-611	9.07	63.99	28.27	88	44
521-611	7.10	81.71	15.78	94	73
513-611	6.04	80.23	16.97	93	72
312-710	3.18	35.76	32.24	63	5
522-611	2.71	79.89	18.08	94	71
514-611	2.55	78.29	19.70	94	69
210-312	2.42	85.21	15.10	96	79
210-611	1.57	84.82	15.48	96	79
710-611	1.50	78.35	21.65	94	70
611-514	1.12	59.02	31.00	87	33
513-312	1.04	85.72	13.19	96	80
611-513	1.02	68.79	24.09	88	54
x̄	5.73	71.70	22.45	90.00	57.38

As principais correções na ocupação do solo causadas pela COSsimR reportaram-se à alteração do tipo de vegetação ou discriminação entre existência ou ausência de vegetação. **Alterações entre (1) sem vegetação e agricultura, (2) sem vegetação e herbácea, (3) matos e agricultura, (4) matos e herbácea e (5) herbácea para matos totalizam mais de 60% das correções entre a COSsimA 2021 e a respetiva COSsimR.** Estes exemplos tinham uma média de incerteza na classificação (*output* do classificador *RandomForest* de 66%).

Algumas das correções que provocaram uma alteração díspar de ocupação (por exemplo de uma espécie florestal para sem vegetação) registaram incertezas na ordem dos 90%.

A análise considerando as espécies florestais que necessitaram de ser corrigidas tiveram uma média de incerteza de 80.71%. Os casos de maior incerteza ocorreram entre outras resinosas e eucalipto (94.06%) e outras folhosas e sobreiro/azinheira (90.77%). Pelo contrário, menores incertezas, porém com necessidade de correção, reportaram-se entre pinheiro manso e pinheiro-bravo (69.24%) e outras folhosas e outras resinosas (68.49%).

A confusão existente entre agricultura e herbácea é outra das alterações mais comuns. No caso das alterações de herbácea para agricultura a incerteza média cifrou-se em 46.17% enquanto a de agricultura para herbácea nos 85.21%.

No global, as áreas com alteração de classe face à ocupação classificada tinham uma média de incerteza de 78.99% e um desvio-padrão de 16.81. Destaque para a média do percentil 75 (92.28%) e do percentil 25 (69.40%).

Face ao exposto, a informação associada à incerteza de classificação de cada pixel é relevante para ter em conta na melhoria de consistência da COSsim. De acordo com os resultados, conjuntos de células com incerteza reduzida não necessitaram de correções associadas à sua classe e, neste sentido, podem ser excluídas do processo de correção por regras.

A desagregação por unidade de paisagem permitirá melhor compreender os limiares de incerteza.

8.2 A relação entre alterações de classe na COSsimR e a incerteza da classificação - Unidades de Paisagem

Repetindo a mesma análise para todos os anos existentes da COSsim (2018, 2020 e 2021) identificou-se por unidade de paisagem as correções mais comuns. Introduziu-se também a incerteza do classificar no caso da COSsim 2021, já que esta informação não se encontra disponível para os anos anteriores.

Dos resultados destaca-se a confusão existente na fase de classificação entre **sem vegetação e herbácea** que é posteriormente alterada, assim como a necessidade de melhor discriminar **herbácea de matos**. Quanto às correções mais comuns relativamente à confusão entre espécies florestais e matos destacam-se as classes de pinheiros mansos assim como de outras folhosas. A correção entre matos e pinheiro-bravo, e vice-versa, não surgiu no top 3 de correções de nenhuma unidade de paisagem, embora espectralmente estas classes sejam difíceis de discriminar. Correções entre espécies florestais mais comum de eucalipto para pinheiro-bravo e vice-versa.

Considerando o caso da COSsim2021, as áreas corrigidas tinham em média uma incerteza de classificação de 64%. Desta forma, a generalidade das correções dadas pela COSsimR afetaram pixéis em que o classificador tinha menos de 1/3 de certeza.

Tabela 6 - Transições mais comuns corrigidas pela COSsimR por unidade de paisagem e média da incerteza para 2021

Unidade de Paisagem	**COSsimA » COSsimR**		
	2018	**2020**	**2021 (x̄ incerteza de classificação RF)**
111 - Minho	312 – 210 611 – 513 521 - 513	710 – 312 611 – 513 210 – 514	611 – 210 (77%) 710 – 312 (54%) 710 – 210 (77%)
112 - Douro	210 – 312 111 – 210 312 - 210	710 – 210 611 – 210 514 - 611	710 – 210 (60%) 312 – 611 (54%) 514 – 611 (70%)
113 - Viseu	312 – 210 312 – 611 611 – 513	312 – 611 513 – 611 521 – 611	521 – 611 (81%) 513 – 611 (77%) 312 – 611 (65%)

114 - Área Metropolitana do Porto	611 – 513 611 – 312 513 – 521	210 – 611 210 – 312 521 – 513	210 – 611 (71%) 611 – 312 (78%) 210 – 312 (77%)
121 - Trás-os-Montes	312 – 210 312 – 611 611 - 514	710 – 210 312 – 611 710 - 312	710 – 210 (59%) 710 – 312 (71%) 611 – 312 (53%)
122 - Serra da Estrela	710 – 312 312 – 210 522 – 521	710 – 312 611 – 312 611 - 513	611 – 210 (69%) 710 – 210 (64%) 312 – 710 (25%)
211 - Área Metropolitana de Lisboa e Oeste	312 – 210 312 – 611 111 – 210	611 – 210 710 – 210 513 - 611	611 – 210 (65%) 710 – 210 (66%) 210 – 312 (86%)
212 - Beira Litoral	312 – 210 522 – 521 522 – 513	710 – 312 611 – 210 710 – 210	312 – 611 (55%) 710 – 312 (45%) 611 – 210 (60%)
213 - Serras calcárias	312 – 210 312 – 611 522 – 513	611 – 210 710 – 210 522 – 611	522 – 611 (72%) 611 – 210 (64%) 710 – 210 (70%)
214 - Tejo e Sado	312 – 210 312 – 510 210 - 522	710 – 210 210 – 611 210 – 312	710 – 210 (63%) 312 – 710 (25%) 611 – 210 (73%)
215 - Algarve Litoral	312 – 210 312 – 611 710 – 911	611 – 210 710 – 210 312 – 611	611 – 210 (69%) 710 – 210 (64%) 312 – 710 (25%)
221 - Castelo Branco	312 – 210 312 – 510 710 – 312	710 – 210 611 – 210 710 – 312	611 – 312 (71%) 710 – 210 (77%) 312 – 710 (31%)
222 - Alentejo	312 – 210 710 – 611 510 – 210	710 – 210 611 – 210 710 – 312	611 – 210 (69%) 710 – 210 (62%) 611 – 312 (66%)
223 - Serra do Algarve	312 – 210 312 – 611 510 – 312	312 – 611 710 – 312 522 – 611	710 – 312 (66%) 611 – 312 (68%) 522 – 611 (82%)

Estes resultados reforçam a potencial utilidade da incerteza de classificação de cada pixel enquanto informação para o melhoramento de consistência da classificação e respectivas transições. A existência desta informação ao nível do pixel constitui a possibilidade de ser utilizada num processo de classificação ou conjugada com probabilidades de transição (e.g., Markov).

8.3 Análise de perfis de NDVI para discriminação de classes de ocupação - Áreas dos buffers

Na sequência dos exercícios anteriores avaliou-se a adequabilidade de utilizar o perfil fenológico baseado num índice de vegetação das classes de ocupação do solo como forma de as discriminar ao longo do tempo.

A identificação de sucessão e recuperação de classes de uso/ocupação do solo após perdas de vegetação continua a ser um desafio da cartografia com origem na classificação de

imagens de satélite (Zhu et al., 2022). Neste sentido, para identificar taxas de recuperação é usual conjugar informação espectral (Kennedy et al., 2010), perfil fenológico de índices de vegetação (Veraverbeke et al., 2012), assim como outras métricas construídas a partir destes (Chu et al., 2016).

Como forma de diagnosticar a possibilidade de identificação de limiares interanuais e intranuais de NDVI para automatização da COSsimR, procedeu-se a uma análise da média mensal de NDVI para um conjunto de ocupações e transições. O período considerado foi de janeiro de 2018 a dezembro de 2021. A área de estudo corresponde a DGT300.

Considerando ocupações que se mantiveram ao longo das três COSsim, a Figura 12 descreve o perfil médio de NDVI. Evidencia-se a semelhança das curvas em termos de trajetória de evolução, embora seja possível discriminar as ocupações a partir do valor de NDVI. Destaca-se a diferença de NDVI médio entre as classes sem vegetação, herbácea e matos, sendo as duas últimas ocupações comummente alvo de correções na COSsimR devido a confusão espectral. Perante o gráfico, é possível perceber que os valores médios são suficientemente díspares para automatizar condições de correção com base em limiares intranuais. Quanto às espécies florestais, a semelhança entre eucalipto e pinheiro-bravo não permite a sua discriminação. Esta semelhança não é confirmada pela literatura que afirma que tendencialmente as folhosas, como o eucalipto, atingem valores de NDVI superiores ao longo do ano, tendo uma amplitude superior, enquanto as resinosas (como o pinheiro-bravo) apesar de registarem NDVI inferior, devido à suas estruturas foliares de menor dimensão, registam amplitude inferior (Caparros-Santiago & Rodríguez-Galiano, 2020). Das espécies florestais testadas, apenas a classe outras folhosas revela um comportamento interanual suficientemente diferenciado para discriminação face ao eucalipto e pinheiro-bravo.

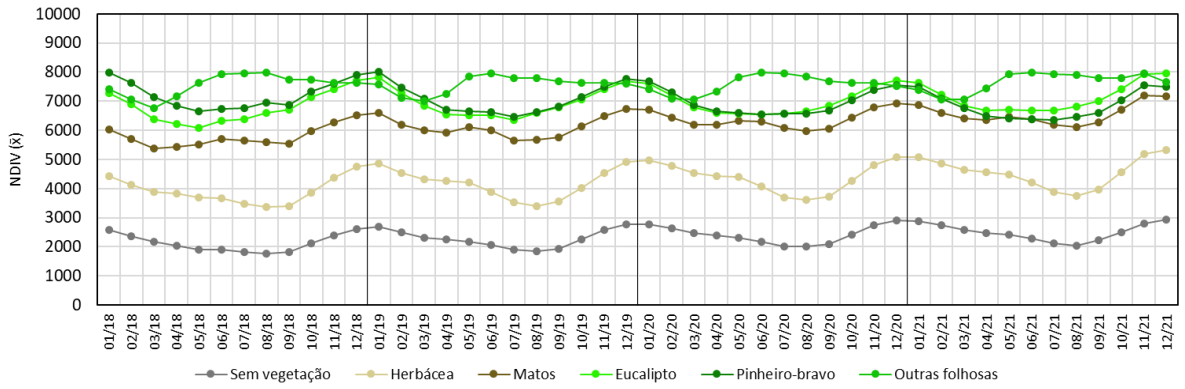

Figura 12. Perfil médio de NDVI (x̄) de classes de ocupação do solo entre janeiro de 2018 e dezembro de 2021 m. Dados de referência DGT300. NDVI está multiplicado por 10000.

Os pixéis considerados podem não ser representativos do real comportamento fenológico das ocupações, já que a área de estudo corresponde apenas aos 300 *buffers*. Além disso, as amostras foram selecionadas a partir de pixéis que mantiveram a ocupação desde a COSsim18 até à COSsim21 podendo existir algumas alterações de classe não identificadas, assim como estas ocupações terem evoluído de estado (e.g., de floresta jovem para madura).

Considerando transições de classe, os perfis são mais informativos. A Figura 13 demonstra a transição de pixéis de sem vegetação para herbácea e posteriormente para matos. É evidente o crescimento dos valores médios de NDVI na transição para herbácea, de

ligeiramente superior a 2000 para perto de 5000 com uma amplitude anual com valores inferiores a 3000. Por sua vez, a transição para matos eleva os valores acima dos 5000 sendo a amplitude menor, já que nos meses de menor NDVI a média mantém-se acima dos 4000. Quanto ao desvio-padrão revelou-se uniforme ao longo do tempo com uma dispersão dos dados à média em torno de 1000.

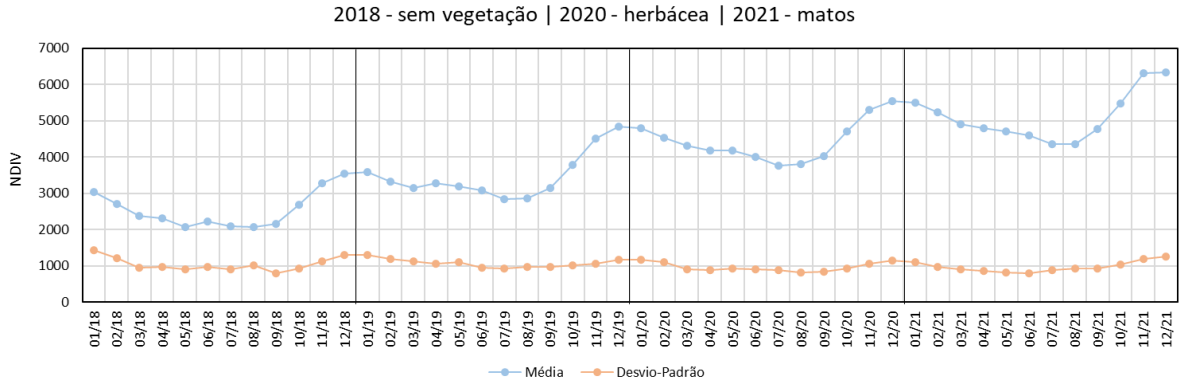

Figura 13. Média e desvio-padrão do NDVI para uma sequência de ocupações (258 pixels)

Testaram-se outras sequências, nomeadamente a transição de herbáceas e de matos para espécies florestais. De forma geral, a partir da Figura 14 identifica-se que a transição de matos para eucalipto é o tipo de perfil que mantém praticamente ao longo do tempo valores de NDVI mais elevados, precisamente porque parte de uma ocupação com maior densidade de vegetação. Pelo contrário, a transição de herbácea para matos manteve geralmente valores de NDVI inferiores aos perfis com transição de herbácea para espécies florestais.

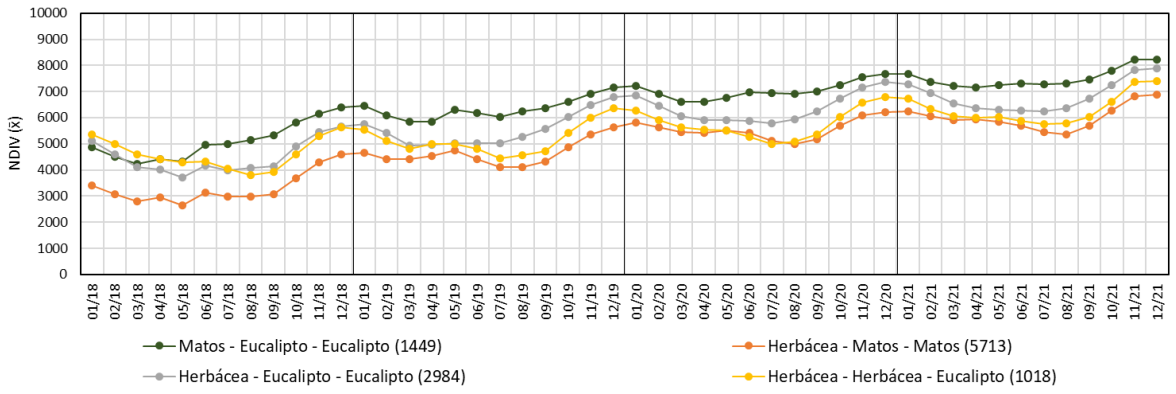

Figura 14. Perfil de NDVI (x̄) de sequências de transição de classes de ocupação do solo entre janeiro de 2018 e dezembro de 2021 (entre parêntesis o número de pixels)

Nestes dois últimos casos (Figura 13 e Figura 14) é preciso ter em consideração a potencial insuficiência de representatividade dos pixéis, assim como o efeito que potenciais erros de classificação podem gerar nos perfis de NDVI.

9 Análise exploratória do output de CCDC para classificação

De forma a responder à problemática de automatização e melhoria de consistência das transições da COSsimR avaliou-se a possibilidade de tirar partido dos coeficientes resultantes da análise de deteção por parte do algoritmo do CCDC para uma classificação. A utilização dos coeficientes do CCDC para classificação de ocupação do solo é utilizado como *input* em outros produtos cartográficos resultantes de classificação de imagem de satélite, como é o caso do Land Change Monitoring, Assessment, and Projection (LCMAP) (Xian *e*t al*.,* 2022).

Para a caracterização das classes com base no output do método CCDC, consideraram-se os pixels de DGT300 que cumprissem cumulativamente as seguintes condições:

- A mesma classe nas COSsim 2018-2020-2021;
- Não corresponder à bordadura dos polígonos;
- Verdadeiro negativo nos resultados do CCDC, ou seja, pixéis que nem o algoritmo nem os analistas identificaram uma perda de vegetação;
- Não ter sido detetado uma perda pelos analistas mesmo se este tivesse ocorrido com mais de 60 dias de diferença em relação ao CCDC.

Recorreu-se a uma análise linear discriminante considerando como variáveis vários coeficientes resultantes da análise de deteção de alterações:

- 8 coeficientes das bandas B2, B3, B4, B8, B11, B12 e NDVI
    - COS - Coeficiente do primeiro cosseno (periodicidade anual)
    - COS2 - Coeficiente do segundo cosseno (periodicidade semestral)
    - CO3 - Coeficiente do terceiro cosseno (periodicidade trimestral)
    - SIN - Coeficiente do primeiro seno (periodicidade anual)
    - SIN2 - Coeficiente do segundo seno (periodicidade semestral)
    - SIN3 - Coeficiente do terceiro seno (periodicidade trimestral)
    - INTP - Intercept (coeficiente linear - onde toca o eixo y)
    - SLP - Slope (coeficiente angular)

Além destes coeficientes incluiu-se o erro médio quadrático para cada uma das bandas e do NDVI.

A análise linear discriminante é um método de classificação e redução de dimensionalidade que visa encontrar p - 1 componentes que melhor discriminam entre as classes alvo. Ao identificar os componentes/eixos discriminantes automaticamente reduz a dimensionalidade dos dados minimizando potenciais efeitos de colinearidade entre as variáveis preditoras. Caracteriza e separa classes calculando combinações lineares ponderadas.

A Figura 15 evidencia a capacidade de discriminação dos eixos discriminantes face às classes de ocupação. Consideraram-se apenas classes processadas na COSsim intercalar excluindo aquelas cuja representatividade na área de estudo era reduzida, como foi o caso da classe “Outras Resinosas”. A Figura 15 compara a discriminação sobre os coeficientes acima referidos a discriminação os compósitos mensais de NDVI reforçando a análise da Secção 8.3.

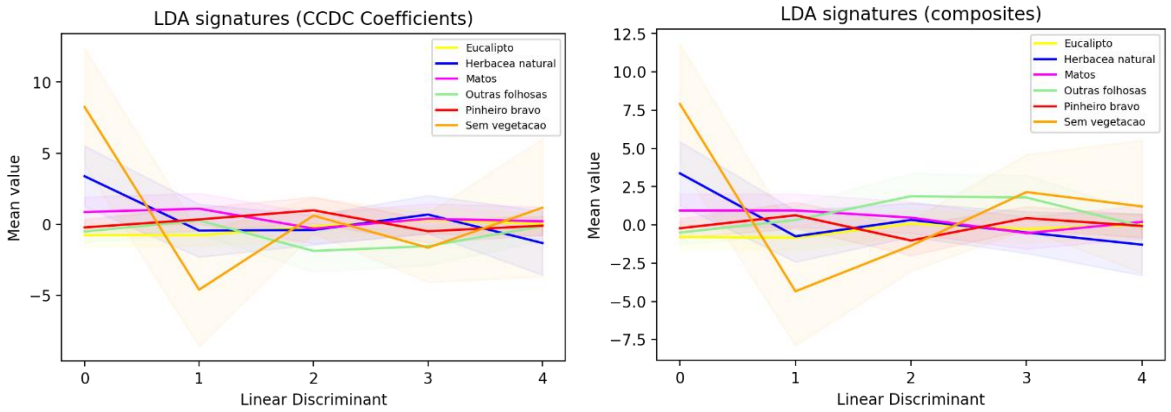

Figura 15. Assinaturas espectrais do eixos discriminantes: a) coeficientes resultantes do CCDC; b) compósitos mensais de NDVI

É evidente a capacidade de discriminação de algumas classes problemáticas em termos de semelhança espectral e de problemas de transição na COSsim. Para tal os eixos 0, 1 e 2 foram utilizados para uma análise mais pormenorizada patente na Figura 16. É notória a capacidade de separação das classes representadas apenas com estes três eixos.

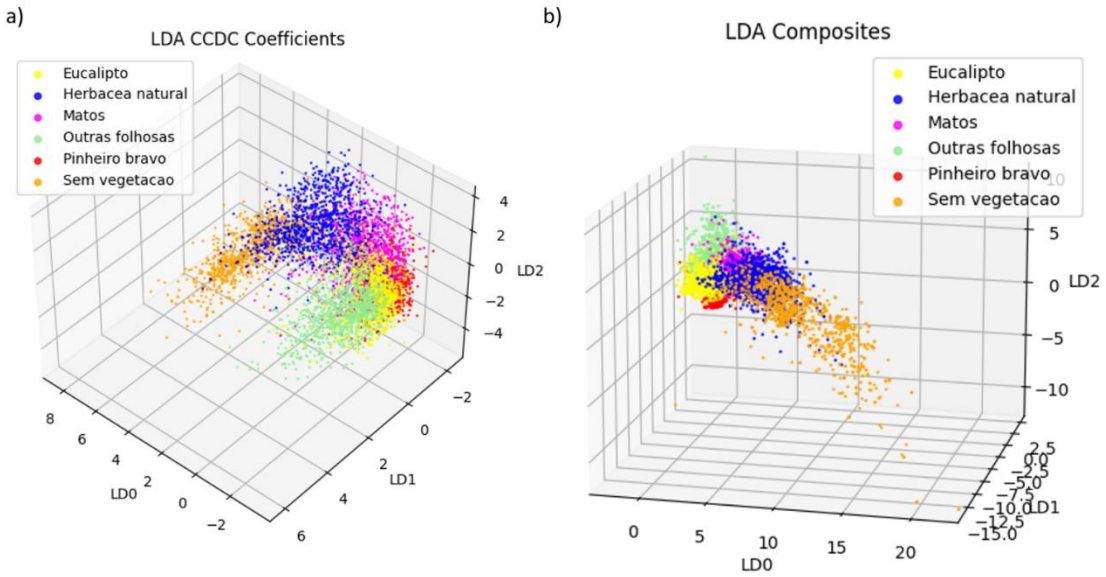

Figura 16. Distribuição das classes utilizando os eixos 0, 1 e 2: a) coeficientes resultantes do CCDC; b) compósitos mensais de NDVI

Referências bibliográficas

Caparros-Santiago, J. A., & Rodríguez-Galiano, V. F. (2020). Estimación de la fenología de la vegetación a partir de imágenes de satélite: El caso de la península ibérica e islas Baleares (2001-2017). *Revista de Teledetección*, *57*, 25. https://doi.org/10.4995/raet.2020.13632

Chu, T., Guo, X., & Takeda, K. (2016). Remote sensing approach to detect post-fire vegetation regrowth in Siberian boreal larch forest. *Ecological Indicators*, *62*, 32–46. https://doi.org/10.1016/j.ecolind.2015.11.026

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

Veraverbeke, S., Gitas, I., Katagis, T., Polychronaki, A., Somers, B., & Goossens, R. (2012). Assessing post-fire vegetation recovery using red–near infrared vegetation indices: Accounting for background and vegetation variability. *ISPRS Journal of Photogrammetry and Remote Sensing*, *68*, 28–39. https://doi.org/10.1016/j.isprsjprs.2011.12.007

Xian, G. Z., Smith, K., Wellington, D., Horton, J., Zhou, Q., Li, C., Auch, R., Brown, J. F., Zhu, Z., & Reker, R. R. (2022). Implementation of the CCDC algorithm to produce the LCMAP Collection 1.0 annual land surface change product. Earth System Science Data, 14(1), 143–162. https://doi.org/10.5194/essd-14-143-2022

Zhu, Z., Qiu, S., & Ye, S. (2022). Remote sensing of land change: A multifaceted perspective. *Remote Sensing of Environment*, *282*, 113266. https://doi.org/10.1016/j.rse.2022.113266