UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Classificação de culturas agrícolas de Inverno com recurso à plataforma *Google Earth Engine* e imagens dos satélites Sentinel-1 e Sentinel-2

Maria João Gonçalves dos Santos

Mestrado em Sistemas de Informação Geográfica – Tecnologias e Aplicações

Trabalho de Projeto orientado por:
Prof. Doutor João Catalão Fernandes

2021

Resumo

A área de deteção remota é uma área em desenvolvimento devido à sua capacidade em adquirir remotamente dados da superfície e da atmosfera da Terra ou de qualquer outro planeta. A quantidade massiva de dados e de imagens de satélite existentes, tem contribuído para uma melhoria e para um aumento da qualidade de informação de observação da Terra. As metodologias científicas tais como a classificação supervisionada por imagens de satélite têm sido aplicadas na análise de dados de deteção remota. O presente estudo tem como objetivo avaliar os resultados obtidos dos processos de classificação por diferentes algoritmos, na utilização de diferentes tipos de dados e otimizá-los através da fusão dos mesmos (*SAR* e multiespectrais), com recurso à plataforma *cloud Google Earth Engine.*

Recorreu-se à classificação supervisionada de imagens de satélite por cinco abordagens distintas, para a classificação de culturas de Inverno, com seis classificadores: *Classification and regression trees (CART), Random Forest (RF), Support Vector Machine (SVM), Maximum Entropy (MAXE), Minimum Distance (MD) e Naive Bayes (NB);* uma delas, a abordagem com recurso à *fusão de dados* SAR e dados multiespectrais.

De todas as abordagens testadas, os melhores resultados foram obtidos com o classificador *Random Forest* (RF), na fusão de dados, com um valor de 76, 2% de exatidão global e de 68,3% de coeficiente kappa com 301 bandas, 102 pertencentes às imagens Sentinel-*1 (SAR)* com polarização *VV* e *VH* e 189 bandas de imagens do Sentinel-2 (multiespectrais) com nove bandas por imagem. Verificou-se que a fusão de dados multiespectrais e *SAR* beneficiam claramente a classificação efetuada, em parte pelo número de imagens utilizadas, fazendo com que as imagens *SAR* beneficiem os sistemas óticos principalmente na época de Inverno, onde as imagens óticas são mais limitadas, devido a nebulosidade presente; obtendo-se os valores de exatidão global de 76,2% comparativamente aos resultados individuais de 73% e 70,8%, um da coleção de imagens *SAR* com 112 bandas*,* e outro dos sistemas óticos com 189 bandas, respetivamente. O conjunto das imagens *SAR* em Sentinel-*1 (*com número de órbita 147 e 52*)* revelam resultados mais elevados do que as imagens individuais dos sistemas multiespectrais em Sentinel-*2*, tendo em conta a época de Inverno.

A classificação de culturas de Inverno foi efetuada com dados fornecidos pelo Instituto de Financiamento de Agricultura e Pescas (IFAP) com informação geográfica das parcelas correspondente à área do Baixo Alentejo e com identificação da classe e área de cada cultura, num método de classificação supervisionado, por repartição em dados de treino e teste.

Este projeto foi realizado com a utilização da plataforma *Google Earth Engine* (*GEE)* pelo processamento computacional demonstrado para a análise de um grande volume de dados como as coleções de imagens de satélite por séries temporais prontas a usar existentes no catálogo de dados; e pela vasta oferta de funções presentes, entre elas os algoritmos de classificação. Revela-se uma plataforma excecional no processamento, análise e classificação, pela sua versatilidade e performance, tornando-se uma ferramenta imprescindível na área de deteção remota que potencializa a utilização de uma quantidade massiva e heterogénea de dados.

Palavras-chave: *Google Earth Engine*, Sentinel-1 e Sentinel- 2, classificação supervisionada, aprendizagem automática, *Big Data*

Abstract

The remote sensing area is an area under development, due to the amount of existing geospatial data. The massive amount of data and existing satellite images has contributed to an improvement and an increase in the quality of Earth observation information. Scientific methodologies such as the classification supervised by satellite images have been applied in the analysis of remote sensing data. This project aimed to evaluate the results obtained from the classification processes by different algorithms, in the use of different types of data and to optimize them through the fusion of them (SAR and multispectral), using the Google Earth Engine cloud platform.

Supervised classification of satellite images is done by five different approaches, using the GEE cloud platform for the classification of winter crops, with six classifiers: Classification and regression trees (CART), Random Forest (RF), Support Vector Machine (SVM), Maximum Entropy (MAXE), Minimum Distance (MD) and Naive Bayes (NB); in one of them, approach A5 using fusion of SAR and multispectral data.

Of all the approaches, RF demonstrated to have the best results in data fusion with a value of 76, 2% of global accuracy and 68.3% of kappa coefficient, reaching peaks in the A5 approach, in the use of 301 bands, 102 belonging to Sentinel-1 (SAR) images with VV and VH polarization and 189 Sentinel-2 image bands (multispectral) with nine bands per image; it was found that the fusion of optical and SAR data clearly benefits the classification made, in part by the number of images used, making SAR images benefit optical systems mainly in the winter season, where optical images are more limited, due the present cloudiness; obtaining the global accuracy values of 76.2% compared to the individual results of 73% and 70.8%, one from the collection of SAR images with 112 bands, and another from the optical systems with 189 bands, respectively. The set of SAR images in Sentinel-1 (with orbit number 147 and 52) show higher results than the individual images of the multispectral systems in Sentinel-2, taking into account the winter season.

The classification of winter crops was carried out using data provided by IFAP with geographic information of the plots corresponding to the area of the lower Alentejo and with identification of the class and area of each crop, in a supervised classification method, by distribution in training and test data.

This project was carried out using the Google earth Engine (GEE) platform for the computational processing demonstrated for the analysis of a large volume of data such as collections of satellite images by ready-to-use time series existing in the data catalog, and for the wide range of functions present, among them the classification algorithms. It proves to be an exceptional platform in the processing and analysis of the classification of satellite images, due to its versatility and performance, making it an essential tool in the area of remote sensing, leveraging the use of a massive and heterogeneous amount of data.

Keywords: Google Earth Engine, Sentinel-1 and Sentinel-2, supervised classification, machine learning, big data

Agradecimentos

Venho agradecer pelo desenvolvimento e conclusão deste projeto ao Prof. João Catalão, pela sua atenção e disponibilidade; ao Instituto de Financiamento de Agricultura e Pesca (*IFAP*), pela valiosa cedência dos dados; e à minha mãe, por tudo.

Índice

Índice de Tabelas

Índice de Figuras

Acrónimos

CART	*Classification and Regression Trees*
CI	Culturas de Inverno
DR	Deteção Remota
EG	Exatidão Global
ETRS89/PT-TM06	*European Terrestrial Reference System* 1989/Portugal Transverse Mercator 2006
F1	F1- score
Fq.	Frequência
GEE	*Google Earth Engine*
GRD	*Ground Range Detected*
IFAP	Instituto de Financiamento de Agricultura e Pescas
IW	*Interferometric Wide Swath*
MAXE	*Maximum Entropy*
MD	*Minimum Distance*
MSI	*MultiSpectral Instrument*
NB	*Naive Bayes*
PP	Precisão do Produtor
PU	Precisão do utilizador
QGIS	*Quantum GIS* (*Geographic Information System*)
Re.	Revocação
RF	*Random Forest*
S1	Sentinel-1
S2	Sentinel-2
S1/A	Sentinel-1/A (satélite A)
S1/B	Sentinel-1/B (satélite B)
SAR	*Synthetic Aperture Radar*
SC	Sistema de Coordenadas
SNAP	*Sentinel Application Platform*
SVM	*Support Vector Machine*
WGS84 / UTM Fuso 29N	*World Geodetic System 1984* / *Universal Transverse Mercator Fuso 29*

1 Introdução

1.1 Enquadramento

A nível global, tem aumentado o interesse de muitos países pela observação da Terra como instrumento de monitorização ambiental e social, para um planeamento sustentável dos recursos naturais, bem como pela sua afirmação como país com capacidade espacial. Antecipa-se para as próximas décadas um crescimento da área do espaço quer ao nível do posicionamento e navegação quer ao nível da observação da Terra, com o interesse de entidades de capital privado na operação e exploração da tecnologia espacial e deteção remota. A agência espacial europeia (*ESA*) tem em curso um programa espacial de observação da Terra destinado a monitorizar o ambiente e a segurança do espaço europeu, com disponibilização de dados de satélite e *in situ* em tempo quase real ao nível global para suporte a ações de investigação sobre o nosso planeta e gestão sustentável do ambiente. O programa é financiado pela Comissão Europeia e designa-se por Programa *Copernicus*. No âmbito deste programa, com presentemente seis constelações de satélites, são disponibilizadas diariamente milhares de imagens e dados sobre todo o planeta. Em particular, e com interesse para este projeto, são disponibilizadas imagens dos satélites Sentinel-1 e Sentinel-2, imagens radar de abertura sintética (*SAR*) e imagens multiespectrais, respetivamente.

Este interesse na observação da Terra tem como primeira consequência a produção diária de *terabytes* de dados sobre a Terra que requerem o seu armazenamento e processamento. Este é um grande desafio da comunidade científica, que dispõe neste momento um enorme volume de dados, e que não tem meios computacionais para o seu processamento e análise, devido ao elevado número de satélites existentes.

Para lidar com estes grandes volumes de dados, há a necessidade de ferramentas específicas. Entre elas, o *Google Earth Engine (GEE)* [28, 37], uma plataforma *web* que, com recurso a tecnologias *cloud computing (*de computação na nuvem*),* oferece capacidade no complexo processamento dos dados de satélite, reduzindo a barreira de infraestruturas de alta performance outrora necessárias, permitindo o acesso a uma vasta série de estudos [31].

O *GEE* é, assim, uma plataforma de escala planetária, com oferta de um vasto catálogo e repositório de dados geoespaciais, que apresenta elevada performance para várias aplicações na área da deteção remota, onde todos os dados são pré-processados e prontos a usar. Dentro das muitas funções oferecidas está incluída a classificação supervisionada [27, 63], por algoritmos de aprendizagem automática tais como: *Classification and regression trees (CART)*, *Random Forest (RF), Support Vector Machine (SVM)*, *GMO Maximum Entropy (MAXE), Minimum Distance (MD), e Naive Bayes* (*NB*). Esta plataforma está em crescente popularidade na sua utilização*,* pela sua proximidade com o utilizador.

A crescente geração de dados bem como a maior acessibilidade aos mesmos através de plataformas como o GEE tem, essencialmente na área da deteção remota, contribuído para o enorme volume de dados geoespaciais existentes, fazendo com que a deteção remota seja uma das áreas que lida com um grande volume de dados (*Big Data*) [7, 29], pela sua aplicação, formato e diversidade existente [33]. Desde a variação na estrutura de dados, em *raster* (imagens que contêm a descrição de cada píxel) ou vetor e o facto de refletirem um estado dinâmico da observação do planeta.

A utilização de diferentes equipamentos, sensores, tendo em conta as dimensões espetrais e temporais na aquisição dos dados, faz com que imagens de satélite, que outrora eram analisadas singularmente ou em pequenas sequências temporais, possam agora ser utilizadas massivamente em longas séries temporais ou por extensas áreas da superfície da Terra, devido à acumulação de grandes volumes de dados pelo aumento do número das constelações de satélites existentes, avanços técnicos e capacidade de processamento, reduzindo a lacuna no acesso a estes.

Neste trabalho de projeto, pretendeu-se explorar a plataforma *GEE* para classificar as culturas agrícolas de Inverno, por aprendizagem automática (*machine learning*), com base numa série temporal de imagens dos satélites *S*entinel-1 e Sentinel-2.

1.2 Motivação e Objetivos

Um dos problemas na área da deteção remota, refere-se à capacidade computacional necessária para lidar com todo o processamento de imagens de satélite e às limitações que isso implica, incluindo o armazenamento local necessário em disco perante os *gigabytes* de informação provenientes de uma só imagem de satélite. As plataformas em nuvem, como o *GEE* visam combater isso, pela oferta da capacidade computacional, por um vasto catálogo de dados inclusive de coleções de imagens de satélite e várias funções para análise, observação e visualização desses dados [72].

No projeto "Avaliação de metodologias de aprendizagem automática na classificação de culturas agrícolas com base em imagens do Sentinel-*2*", Silva, I. (2020) refere a necessidade em enriquecer as amostras de treino pelo impacto que algumas imagens em falta, devido à cobertura nebulosa, tiveram na qualidade dos resultados obtidos [60]. O objetivo da fusão das imagens *SAR* do satélite Sentinel-1 com as imagens multiespectrais do satélite Sentinel-2, pretende dar resposta a essa necessidade pelo facto de estas poderem ser utilizadas em condições atmosféricas menos favoráveis, logo vantajosas no período de inverno, quando ocorre uma limitação das imagens existentes dos sistemas passivos multiespectrais (pela presença de nuvens); para além de fornecer informação distinta de retrodispersão para cada cultura, com a contribuição de mais imagens para uma melhoria do *training set*.

Este trabalho de projeto tem dois objetivos: a) estudar a fusão de imagens *SAR* e imagens multiespectrais (sistemas óticos) para classificação de culturas de inverno e b) estudar as potencialidades de processamento na *cloud* através da plataforma *Google Earth Engine*.

Os dados a serem utilizados no processo de classificação, foram fornecidos pelo *IFAP* com informação sobre culturas agrícolas existentes na área do Baixo Alentejo; com a localização, identificação geográfica e área de cada parcela foram processados de modo a obter somente as culturas de Inverno de interesse.

A avaliação da classificação de culturas de Inverno, como apoio à atividade agrícola, por imagens de satélite Sentinel-2 e Sentinel-1 foi realizada com recurso à plataforma *GEE* por meio da importação de dados em formato *shapefile* [44]. Pretendeu-se avaliar este processo pela utilização dos classificadores disponibilizados na plataforma, com recurso à linguagem de programação *JavaScript* e medir o impacto da plataforma no mesmo.

Seguindo os procedimentos de uma classificação supervisionada, com a repartição dos dados em treino e teste e com a utilização das duas séries temporais de imagens efetuou-se a extração da média dos valores dos píxeis por parcela com a criação de uma imagem multibanda. Para cada algoritmo de classificação foi generalizado o modelo através dos dados de treino com o conhecimento da ocupação cultural declarada em cada parcela e a avaliação do modelo pela classificação dos dados de teste.

Tendo em conta o impacto da utilização de imagens *SAR* (Sentinel-1) no período de Inverno, procurou-se integrar, nos métodos de classificação supervisionada [54], a sua utilização com imagens multiespectrais (Sentinel-*2*), na busca dos melhores valores de exatidão global. Esta busca, está dependente não só dos classificadores utilizados, mas na definição dos parâmetros dos mesmos e na resolução espacial, no tipo de dados e época de análise das imagens utilizadas para obtenção dos valores dos píxeis.

1.3 Organização do trabalho

O presente trabalho de projeto encontra-se estruturado em cinco capítulos, divididos em subcapítulos. O primeiro introduz e contextualiza, referindo os objetivos do tema em estudo; no segundo revela-se o estado da arte, que visa caraterizar os tópicos abordados em literatura e metodologias, fundamentando-os com uma abordagem aos conceitos teóricos; o terceiro carateriza a área de estudo, pela particularidade dos dados e a metodologia utilizada; no quarto apresenta-se os resultados obtidos e a sua explicação; no quinto e último capítulo a conclusão, com algumas justificações, motivos e sugestões de novas abordagens, perspetivas e aplicações futuras.

2 Estado da Arte

2.1 Mega Dados (*Big Data*)

A recolha de dados na área de deteção remota é realizada por dois tipos de sensores: passivos, que dependem da radiação solar como fonte de iluminação, e ativos em que a energia é emitida pelo próprio sensor. Exemplos destes sistemas são os sensores multiespectrais *MSI* a bordo do satélite Sentinel-2 e sistemas *radar* como o *SAR* a bordo do satélite Sentinel-1, respetivamente [39].

A quantidade massiva de dados de diferentes sensores é recolhida, essencialmente pelo número crescente de satélites e missões existentes, contribuindo para várias aplicações na área de observação da Terra. A aplicação no uso e ocupação do solo, pela classificação de culturas é uma das áreas de elevado interesse na qual tem emergido a fusão de dados [2] de diferentes sensores.

Esta quantidade massiva de dados, os mega dados (*big data*) [40], referem-se normalmente a dados que pelo seu volume e conjunto, acabam por ser difíceis de armazenar e de processar. Caraterizam-se pelo seu volume na enorme quantidade de dados geoespaciais existentes de diferentes fontes cuja dificuldade é o armazenamento; pela variedade devido à heterogeneidade presente e nos formatos com estruturas complexas, sejam estes *raster* ou vetor; e pela velocidade, no qual estes dados são obtidos ou transmitidos. Estas caraterísticas fazem com que os dados utilizados em deteção remota (DR) e na observação da terra se definam como mega dados [31].

A quantidade de dados em DR, recolhida por um único satélite está na ordem de vários *terabytes* por dia. Neste momento, centenas de sensores de satélites em órbita providenciam informação atualizada de uma contínua observação da terra, através de dados multi-espaciais e multi-temporais. A complexidade dos mega dados deve-se à sua diversidade, à elevada dimensionalidade e à intrincada organização dos metadados. Pois o crescimento de dados de DR traz consigo o presente aumento de informação de metadados (um exemplo é a descrição e o tipo de informação da imagem, a sua localização geográfica, a projeção existente, entre outros) [72].

2.1.1 Tecnologias para armazenamento dos mega dados

Um modo de lidar com os enormes desafios no processamento e análise dos mega dados é através de sistemas de plataforma de nuvem (*cloud plataforms*) e supercomputadores (conhecidos como *High-Performance Computing, HPC systems*). Nestes últimos, múltiplos computadores apresentam um "sistema singular de imagem" com enorme poder computacional, embora tenham que efetuar todo o processamento necessário e o carregamento de quantidades massivas de dados, onde apresentam fraca performance. Já as *cloud plataforms*, como o *GEE,* oferecem maior acessibilidade e empenho, com processadores, memória e disco como um verdadeiro computador físico por infraestruturas de tecnologias virtualizadas [*65*]. São muito menos dispendiosas e disponibilizam como serviço, a plataforma, o *software,* a infraestrutura e ainda espaço de armazenamento na nuvem. São das técnicas mais robustas para os mega dados.

Para lidar com a disponibilidade e localização dos dados, os melhores sistemas são as bases de dados e os sistemas paralelos (*PFS - Parallel file system*), pois aceder e armazenar *terabytes* ou até mesmo

exabytes de dados tornou-se um desafio extremo. Deste modo, o sistema paralelo lida com esta situação distribuindo o enorme volume de dados em repartições por diferentes discos, reduzindo o tempo de *input e output* de informação. Assim o acesso aos dados pode ser efetuado simultaneamente em diferentes discos. Como sistema de base de dados usa o *NoSQL (Not only SQL) Big Data Database Service* pois é ideal na utilização e na distribuição de mega dados não estruturados e não relacionados, oferecendo elevada capacidade de armazenamento [65].

É maioritariamente com estas tecnologias e sistemas, entre ferramentas de gestão de dados e de tarefas que os métodos de computação na nuvem apresentam plataformas, *software,* infraestruturas e armazenamento, dos quais o *GEE* é um exemplo de acesso livre para dados de observação da Terra.

2.2 *Google Earth Engine* (*GEE*)

O GEE é uma plataforma que permite a análise de informação geoespacial em nuvem (*cloud based*) para observação da terra com o objetivo científico de monitorização, análise e visualização de dados. Fornece como catálogo de dados vários *petabytes* de imagens de satélites de observação da Terra; uma quantidade massiva de informação pronta a usar. Possui, como referido anteriormente processamento paralelo com elevada velocidade, onde toda a infraestrutura computacional pertence à *Google* [28].

A sua utilização é gratuita se for fundamentada como método educativo sem fins lucrativos, onde o acesso é limitado por uma quota e cada utilizador tem uma conta de serviço; a quota individual por pedido é a mesma para cada um dos utilizadores. O apoio técnico, é oferecido através da *Google Developers email list* e *Developers Q&A site.*

O catálogo de dados existente na plataforma consiste num repositório de conjuntos de dados geoespaciais públicos. Neste encontram-se imagens multiespectrais e imagens *SAR* de uma variedade de satélites, incluindo imagens das constelações Sentinel [6, 40]; variáveis ambientais e de cobertura do solo; dados topográficos e socioeconómicos, continuamente atualizados. Só neste aspeto, ter uma vasta gama de dados pronta a usar é das maiores vantagens, onde qualquer utilizador pode aceder conjugando a utilização do catálogo de dados com dados privados, ao importá-los para a plataforma. Um exemplo disso está neste projeto pela combinação de dados, em que dados de ocupação do solo foram importados para a plataforma e outros foram obtidos do catálogo de dados pela escolha de coleções de imagens Sentinel-1 e Sentinel-2.

Relativamente à informação em *raster*, o *GEE* usa um modelo de dados baseado em grelhas de 2D, onde as bandas singulares precisam de uniformidade no tipo de dados, na sua resolução e projeção; contudo, cada imagem pode conter um vasto número de bandas e estas serem diferentes entre si. Neste caso as imagens produzidas por um único sensor, são agrupadas e fazem parte de uma coleção. Estas coleções são passíveis de serem filtradas, de modo a obter os produtos necessários ou com interesse, tanto em critérios temporais, espaciais, ou até mesmo pela cobertura mínima de nuvens desejadas. As imagens são pré-processadas por cortes em *tiles* de *256 x 256* píxeis, e armazenadas numa base de dados para um acesso eficiente garantindo a preservação da informação, com a mesma resolução e projeção da imagem original. Outra particularidade é que a rápida visualização se deve a um esquema pirâmide de redução, onde cada nível é criado pelo *downsampling* do nível anterior. Deste modo, quando é feito um pedido de uma parte da imagem, somente os *tiles* necessários de um determinado nível são obtidos.

Toda a arquitetura do *GEE* é desenvolvida num vasto conjunto de tecnologias informáticas. Um dos benefícios é que o utilizador está protegido de todos os detalhes de funcionamento do ambiente *back end* de processamento paralelo, alocação de recursos, computação, distribuição e armazenamento de dados. O que os utilizadores usam na realidade são bibliotecas de cliente (*client libraries)* com linguagens de programação como *Python* ou *JavaScrip, (*neste caso em específico no desenvolvimento deste projeto foi *JavaScript)* e que providenciam objetos *proxy* para imagens, coleções e outros tipos de dados num *IDE (Interactive development environment),* como o editor de código da plataforma do *GEE* que por sua vez é uma *API (Application programming interface) [11].*

A biblioteca do *GEE* possui uma enorme variedade de funções, desde funções matemáticas, geoestatística, aprendizagem automática e operações no processamento de imagens, entre outras. Entre variadíssimas operações, só na área de aprendizagem automática, oferece dezenas de tipos de classificação, regressão e de operações. Estas podem ser conjugadas e compostas de modo a satisfazer cada uma das necessidades do utilizador. Um dos exemplos referidos nesta dissertação, para além da classificação supervisionada, são as operações de redução realizadas nas coleções de imagens, de um ponto de vista de análise temporal em série, por modo de agregações estatísticas onde as estatísticas por píxel são calculadas no conjunto de todas as imagens. São operações que envolvem *tiling* e agregação. Dependendo do tipo de operação necessária, estas agregações podem ser mais rápidas e eficientes, ou ter um custo intensivo de memória.

O *GEE* usa um modelo de cálculo lento para calcular somente as partes do resultado que dão resposta ao pedido, para facilitar e suportar uma exploração rápida e interativa na análise de dados. É capaz de suportar intensas e grandes computações, e parte da sua eficiência consta na otimização do código e na distribuição eficiente de computações complexas.

A infraestrutura é então baseada num modelo de programação cliente-servidor onde a biblioteca de cliente proporciona um *script* para a escrita do código. Mas, na realidade são os mecanismos por parte de objetos *proxy earth engine,* começados por "*ee."* (diferentes de objetos de *JavaScript)* existentes na biblioteca no lado do cliente que fazem os pedidos para o servidor para serem executados [28].

Com inúmeras aplicações, desde a análise do uso e ocupação do solo [3, 30], à classificação [25] [47][59], mapeamento urbano e análise de desflorestação [52], o *GEE* pretende ser uma ferramenta útil e pragmática no uso de informação geoespacial, mais especificamente na área de deteção remota, para benefício, estudo, análise e monitorização do planeta, providenciando poder computacional e acesso gratuito a uma grande quantidade de dados [70].

2.2.1 Aprendizagem automática no *GEE*

Duas áreas com elevado interesse e impacto têm contribuindo para a área de deteção remota e observação da terra: a área dos mega dados e a de aprendizagem automática (*machine learning*); esta última, tem como objetivo a deteção de padrões entre os dados, e é na DR que é fundamentalmente usada para prever e compreender uma vasta gama de aplicações, entre elas, a classificação. É uma área que permite um estudo emergente para futuras aplicações nas ciências da terra [43].

Com recurso à aprendizagem automática com algoritmos de classificação supervisionada no *GEE*, pela escolha de classes e de dados de treino [67], algoritmos de classificação paramétricos e não paramétricos podem ser utilizados para classificação do uso do solo em vastas áreas do território. Esta

classificação é realizada pela atribuição de valores de píxeis a determinadas classes pela sua assinatura espetral e informação contextual. Os algoritmos de classificação, em geral, são muito sensíveis à escolha dos dados de treino pelo seu tamanho e pela representatividade presente por classe.

Dos classificadores não paramétricos existentes na plataforma [43] consideraram-se o *CART (Classification and Regression Trees), RF (Random Forest), SVM (Support Vector Machine) e MAXE (Maximum Entropy)*. Estes classificadores dependem fortemente da amostra de dados de treino e embora tenham elevado poder e performance apresentam um risco quanto ao sobreajustamento dos dados. Os algoritmos paramétricos, como *MD (Minimum Distance)* e *NB (Naive Bayes)* embora sejam muito mais simples e rápidos na obtenção dos resultados, em parte pela independência dos dados de treino, demonstram em geral uma complexidade muito limitada pela má generalização dos modelos [59].

O ponto seguinte, apresenta uma breve caraterização dos classificadores paramétricos e não paramétricos disponíveis na plataforma que puderam ser usados na classificação supervisionada.

2.2.1.1 Classification and Regression Trees (CART)

CART é um algoritmo de classificação que se baseia numa escolha hierárquica de árvores de decisão binárias a partir dos dados de treino, onde cada árvore é composta e conectada por nós, e em cada nó é efetuado uma divisão (*split*) em dois ramos levando a nós singulares chamados folhas que representam a escolha da classe obtida. Usa o *Gini Impurity Index*, de modo a escolher qual o atributo ideal a efetuar o *split* em cada nó. O processo de divisão ocorre, até ser admitido o limite de critério. É um classificador simples, do qual são interpretáveis as relações e processos de escolha executados pelo classificador, pelo seu sistema *white box.* É um algoritmo com rápido processamento e com boas referências em vários estudos de deteção remota [59]. Em parte, por ser altamente sensível à amostra de dados em cada classe e à elevada dimensão dos dados. Apesar da rápida performance, tende a efetuar o sobreajustamento dos dados. Um modo de lidar com esta situação é pelo *prunning (*redução dos ramos da árvore*),* tornando o modelo mais robusto aos dados de teste [10].

Desenvolvido por Breiman, Freidman, Olshen, Stone, Chapman and Hall em 1984, possui na plataforma dois parâmetros: o número mínimo de folhas e o número máximo de nós [24, 59].

2.2.1.2 Random Forest (RF)

RF é um dos classificadores com elevada exatidão que combina resultados de múltiplas árvores de decisão pela seleção de subconjuntos aleatórios dos dados de treino para a construção de múltiplas árvores singulares. Usa a técnica de *bagging (bootstrap aggregating)* que faz com que diferentes árvores possam ter acesso à mesma amostra, enquanto outros dados poderão nem ser selecionados. A árvore é construída progressivamente, pela escolha em cada nó das variáveis do subconjunto pelo melhor valor de *split.* A escolha do valor de *split*, é realizada pelo *Gini Impurity Index,* pela medição do grau de impureza. É um classificador robusto ao ruído sem necessidade de métodos de *pruning* nas árvores, ao contrário do *CART [38]*.

Definido por vários parâmetros, entre eles a escolha do número de árvores que afeta os valores de exatidão obtidos; e que, embora de acordo com a literatura um número aceitável são cerca de 500 árvores, pois acima de um certo número de árvores a exatidão obtida não é melhorada; para esta

dissertação em específico os valores de exatidão com 150 árvores eram semelhantes a valores de exatidão com números muito mais elevados de árvores. O valor de *split* também de acordo com a literatura é obtido pela raiz quadrada dos dados a usar; valor este pré-definido na plataforma. É um *ensemble classifier* pois resulta da combinação de múltiplas árvores tipo *CART*, onde cada árvore independentemente classifica os dados e vota pela classe mais popular. Em DR tem a particularidade de ajudar a perceber a importância de diferentes variáveis pela escolha das bandas de satélite utilizadas. A exatidão é obtida pelo erro *out of the bag* (com a utilização dos dados de teste) e do *Gini Impurity Index.*

O *RF* é dos classificadores mais populares em *DR,* pelo seu rápido processamento tem a particularidade de lidar bem com dados de elevada dimensionalidade. Os classificadores *ensemble* com métodos de *bagging* permitem obter melhores valores de exatidão do que classificadores singulares, pois acabam por ser mais estáveis e robustos ao ruído.

É muito versátil e favorável na aplicação em vários estudos em *DR*, principalmente em classificação de culturas, em dados multi-temporais e multi-espetrais provenientes de diversas fontes. É um classificador robusto ao fenómeno de sobreajustamento, mas sensível aos dados de treino. De acordo com a literatura, a diversidade de conjunto de dados, pela fusão de diferentes bandas de coleções diferentes, pode e foi usada para melhorar a exatidão obtida. A particularidade de demonstrar quais as variáveis com maior importância oferecem a possibilidade de reduzir o esforço computacional, pela escolha das bandas de maior interesse, mantendo os valores de exatidão.

De acordo com Belgiu e Drăguţ [4] outra situação particular para o *RF* é na escolha de dados, onde a seleção de um número igual de amostras para cada classe permite alcançar os melhores resultados de exatidão. Isto é algo que não ocorre nos dados utilizados para os resultados nesta dissertação, onde as classes de aveia e cevada apresentam um maior número de dados acabando por favorecer as classes mais representativas. A redução de erros de omissão e de comissão poderão ser obtidos com a utilização de uma amostra proporcional de dados. *RF* é um classificador estável e rápido que apresenta melhores resultados do que o *SVM* em dados com elevada dimensionalidade.

2.2.1.3 Suporte Vector Machine

O SVM é um classificador popular muito usado em deteção remota, pelo facto de proporcionar elevada exatidão. Seleciona um pequeno conjunto de dados de treino, para a generalização do modelo, e parte dum princípio de que dados de treino com valores aproximados perto do limite de uma determinada classe discriminam melhor essa classe do que outros dados [4]. É um classificador fortemente sensível à escolha do *kernel,* ao parâmetro de custo *C*, e ao valor *gamma* na função do *kernel*, mas do qual a literatura existente não é muito explícita na melhor opção dos mesmos, embora estes tenham uma forte influência na performance e nos valores de exatidão obtidos. Do ponto de vista geral, sabe-se que elevados valores de *C* levam a um sobre-ajustamento do modelo e valores baixos a píxeis mal classificados.

O *SVM* é um classificador com bons resultados, com boa generalização do modelo, mesmo em dados de elevada dimensionalidade, algo que é benéfico em DR devido à variedade de dados multi-espetrais e multi-temporais existentes. Contudo, nesta dissertação não foi o classificador com os melhores resultados obtidos. Apresenta-se como um classificador distante do utilizador, pelo seu modelo *black-box* na imprevisibilidade dos híper-parâmetros [59, 64].

2.2.1.4 GMO Maximum Entropy

O algoritmo de *maximum entropy [45]*, só mais recentemente é que começou a ser ponderado para a utilização na classificação singular do uso do solo. Porém, o algoritmo usado na plataforma corresponde ao *GMO maximum entropy (MAXE),* muito usado em problemas de classificação multi-classe *(*com o objetivo de identificar e classificar todas as classes de interesse aqui representadas*)*; combinando com algumas das caraterísticas clássicas do *maximum entropy.*

A sua aplicação na área de classificação faz com que as probabilidades das classes possam ser aprendidas a partir dos dados de treino e posteriormente usadas para classificar os dados de teste. O treino dos modelos com grandes conjuntos de dados requer elevado poder computacional. Na plataforma os híper-parâmetros a considerar são os pesos, uma variável *epsilon* de otimização de paragem e o valor de máxima e mínima iteração. De todos os classificadores é o que tem um custo computacional elevado para grandes volumes de dados, embora seja atraente pela sua base teórica e resultados demonstrados.

2.2.1.5 Minimum Distance

O *MD* tem como parâmetro a escolha da métrica de distância a usar (euclidiana, *mahalanobis*, *cosine*); é um algoritmo usado para classificar dados de teste em classes que minimizam a distância entre os dados de treino e as etiquetas de classe. Onde a distância é definida como um índice de similaridade, para que a distância mínima seja idêntica à similaridade máxima.

Apresenta uma caraterização reduzida por ser um classificador muito simples, que tem como parâmetro somente a distância a considerar.

2.2.1.6 Naive Bayes

NB tem como argumento um valor *lambda*, de modo a evitar que a probabilidade zero seja destacada a classes nunca vistas durante o treino. É um algoritmo de classificação baseado no teorema de *Bayes*, para prever a classe dos dados de teste e tem como princípio a independência condicional das variáveis.

É um modelo de fácil classificação, rápido e útil a um elevado conjunto de dados, não requerendo muitos dados de treino para a generalização do modelo [48].

2.3 Fusão de Dados – *SAR* e multiespectrais

Os diferentes satélites na área da DR adquirem a radiação eletromagnética refletida pela superfície da Terra, cobrindo o espetro eletromagnético desde os ultravioletas às micro-ondas. A possibilidade de fusão de dados de diferentes sensores [39] tem sido investigada por diversos autores tendo em vista a melhoria do processo de classificação de imagem [41 – 42, 47, 71]. Em parte, devido à limitação inerente das imagens multiespectrais, por dependerem de iluminação solar, e de boas condições meteorológicas, inclusive ausência de nuvens, enquanto que as imagens *SAR* estão disponíveis em quaisquer condições meteorológicas, de dia e noite, e embora sejam ricas em informação espacial, não contêm a informação espetral dos sensores multiespectrais [39].

Esta fusão consiste na combinação de diferentes imagens provenientes de diferentes sensores numa só imagem multibanda com todas as bandas, com o intuito de obter uma melhoria significativa dos resultados, pela discriminação das classes nas diferentes caraterísticas espetrais de cada cultura. Em vários artigos, comprova-se que os métodos de classificação realizados a partir de fusão de dados, oferecem uma exatidão global mais elevada; a exatidão obtida dos métodos de classificação utilizados foi significativamente melhorada com a fusão dos dados *SAR* (Sentinel-1*)* [5,47].

A aplicação desta abordagem, torna-se muito mais simples e eficiente com a utilização de *cloud plataforms* como o *GEE* pela abundância de dados proveniente de diferentes sensores. Devido aos avanços nas técnicas de processamento de dados, a fusão de dados [53] traz consigo vários assuntos em voga, entre eles dois referidos anteriormente: os mega-dados e a computação em nuvem.

A correção do *speckle* é considerada na literatura como tendo um efeito positivo nos resultados de classificação, pela redução do ruído das imagens de S1 referindo também que o *SAR* é uma escolha relevante e adequada para monitorização de atividades agrícolas. Gaetano [25] refere que a combinação de fusão de dados multiespectrais do satélite Sentinel-2 com as polarizações *VH* e *VV* do Sentinel-1 obtiveram os melhores valores de exatidão.

A abordagem seguida neste projeto consistiu também na fusão de dados multiespectrais (óticos) provenientes do Sentinel-2 com as polarizações *VH* e *VV* do Sentinel-1 para uma melhoria dos valores de exatidão global obtidos. A correção do *speckle* também foi abordada, com o objetivo da redução do ruído das imagens S1, mas perante resultados ambíguos [25].

Vários artigos revelam a versatilidade da utilização do *GEE* em diferentes aplicações, principalmente como fonte de acesso às coleções de imagens de satélite existentes. Entre eles: o estudo da fenologia da superfície do solo no Ártico com imagens do Sentinel-2 pela sua elevada resolução no suporte à monitorização das mudanças de vegetação, com recurso a índices de vegetação [13]**;** a deteção das mudanças de vegetação urbana, onde o *GEE* foi usado para efetuar o mapeamento de toda a vegetação dum ponto de vista global nas cidades mundiais com o objetivo de a manter e expandir em futuras urbanizações onde a classificação das áreas vegetativas foi identificada com recurso ao algoritmo *RF* [52]; a redução de dimensionalidade e a seleção de *features* dos dados na comparação dos resultados com recurso a *LDA (Linear Discriminant Analysis), MI (Mutual Information)* e *F-score (Fisher's Criterion)* para a classificação de culturas com o algoritmo *SVM* na fusão de imagens Sentinel-1 e Sentinel-2; e a utilização de imagens *SAR [*62*]* e multiespectrais no estudo fenológico, com recurso ao índice de *NDVI* e *cross ratio* para deteção do crescimento das culturas de Inverno pela sua sensibilidade ao crescimento da vegetação antes e depois do Inverno [47].

Todos estes artigos tiveram em comum a utilização da plataforma *GEE*, para acesso às coleções de imagens de satélite e alguns pela classificação com recurso a algoritmos oferecidos pela plataforma como o *SVM* e *RF*, entre outras técnicas. Demonstraram resultados extremamente positivos e facilitadores pelo uso da plataforma no seu desenvolvimento, onde a fusão de dados tem demonstrado melhorias significativas.

3 Dados e Métodos

3.1 Área de Estudo

A área de teste usada neste projeto está localizada no Baixo Alentejo, Portugal, *Figura 3.1*, com clima mediterrânico e de invernos amenos, abrangendo parte dos concelhos de Alvito, Vidigueira, Beja e Aljustrel, e limitada pelas coordenadas cartográficas M (-198000: -158000) e P (-37000: 23000) no sistema de coordenadas ETRS89/PT-TM06.

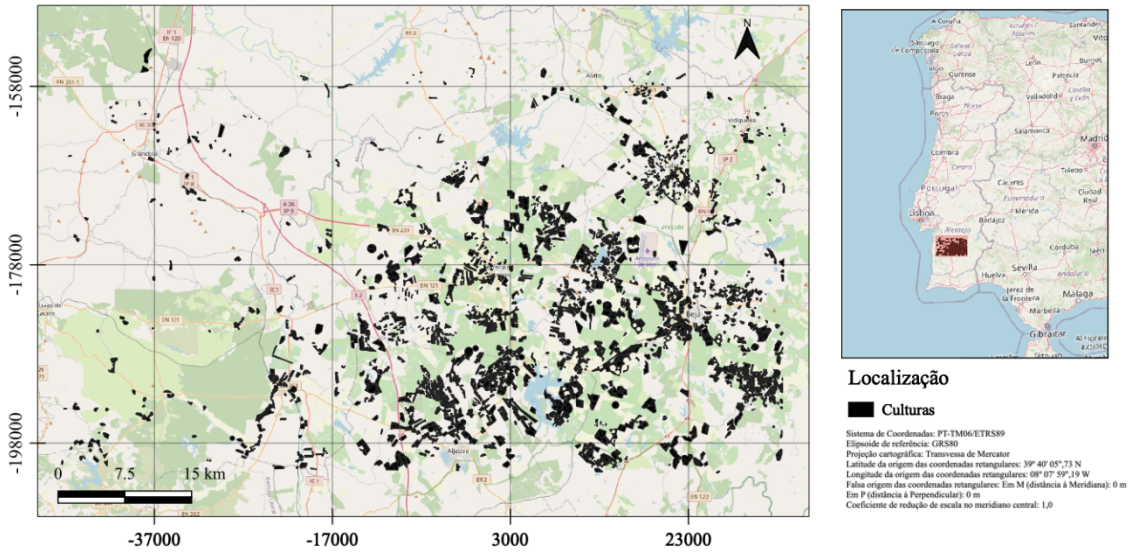

Figura 3.1 Localização das parcelas agrícolas com ocupação cultural conhecida

3.2 Dados

3.2.1 Dados da ocupação do solo

Os dados da ocupação cultural das parcelas foram fornecidos pelo IFAP (Instituto de Financiamento de Agricultura e Pescas) em que os dados de ocupação do solo foram declarados pelos agricultores nos pedidos de apoio à atividade agrícola na campanha de 2018/2019. No ficheiro disponibilizado, constavam culturas temporárias, de verão e de inverno. Foram eliminadas todas as parcelas com declaração de culturas de verão reduzindo-se o ficheiro a apenas culturas temporárias de inverno. Com o objetivo de treinar os modelos de classificação, na tabela de atributos, considerou-se a declaração da ocupação do solo, isto é, o tipo de cultivo, e a área pela delimitação geográfica de cada parcela agrícola. No ficheiro, cada linha representa uma *feature [16]*, identificando cada parcela agrícola. Devido ao tamanho deste e para que pudesse ser importado para o *Google Earth Engine* (*GEE*), foi necessário reduzir o número de parcelas a 38108 parcelas. De notar que neste projeto foi usada uma inscrição com justificação académica para utilização do *GEE* com limitações ao nível do volume de dados pela restrição de quota individual existente na plataforma; esta restrição garante os recursos computacionais a toda a comunidade *GEE.* Poderá ser levantada perante casos especiais [26].

A localização das parcelas apresenta-se coberta pela quadrícula UTM "NC" correspondente às imagens de Sentinel-2, e por imagens com dois *relative orbit number* distintos para Sentinel-1, o 147

e o 52. Das culturas temporárias mais relevantes e com enfoque nas culturas de inverno, consideraram-se nove classes de cultivo para classificação: aveia, azevém, cevada, colza, ervilha, fava, tremocilho, trigo e triticale (Tabela 3.1) e com área maior e igual a 1000 m^2 e representatividade geográfica maior que vinte polígonos. Foram eliminadas todas as parcelas com menor representatividade (Tabela 3.2).

Após a redução de tamanho, este ficheiro foi validado, de modo a obter polígonos e elementos válidos. Definiu-se o sistema de coordenadas WGS84 / UTM Fuso 29N (EPSG: 32629), para a importação das 38108 parcelas.

Tabela 3.1 Identificação das classes das parcelas por valores numéricos

Categórico	Aveia	Azevém	Cevada	Colza	Ervilha	Fava	Tremocilha	Trigo	Triticale
Classe	0	1	2	3	4	5	6	7	8

Após o processamento, e como em *GEE* a classificação é realizada em valores numéricos, foi necessária a reconversão dos valores categóricos do tipo de cultivo (com a utilização do comando *remap*) para valores de classe numéricos (Tabela 3.1).

Tabela 3.2 Identificação e seleção das parcelas de interesse

Nº total de parcelas	38108
Nº parcelas com culturas de Inverno selecionadas	6466
Culturas de Inverno eliminadas	Batata Centeio Morangos Trevo

Ao se obterem os dados relevantes às culturas de Inverno, foram identificados o número de parcelas agrícolas por cultura. A proporção das diferentes culturas de inverno, apresenta-se na Figura *3.2*; as culturas com maior predominância são a cevada, o trigo e a aveia. Caracteriza-se por uma amostra não homogénea na proporção de dados para o treino dos classificadores.

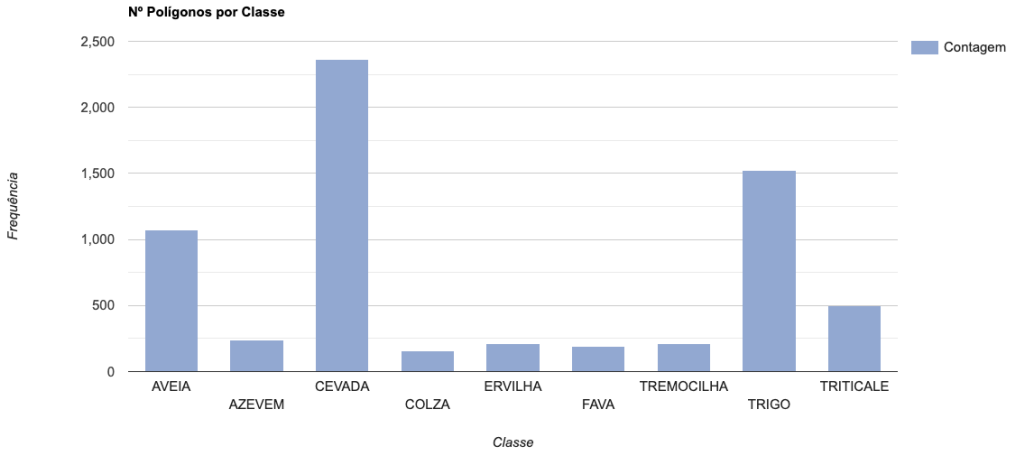

Figura 3.2 Número de parcelas por classe [15]

Outro modo de visualizar a proporção de culturas, é pela agregação da área de cada parcela pelo tipo de cultura (Figura *3.3*); onde as culturas de fava e ervilha representam cerca de 2%, a colza com ~3% e a tremocilha com 4%; a cevada e o trigo representam as culturas com maior área parcelar [20].

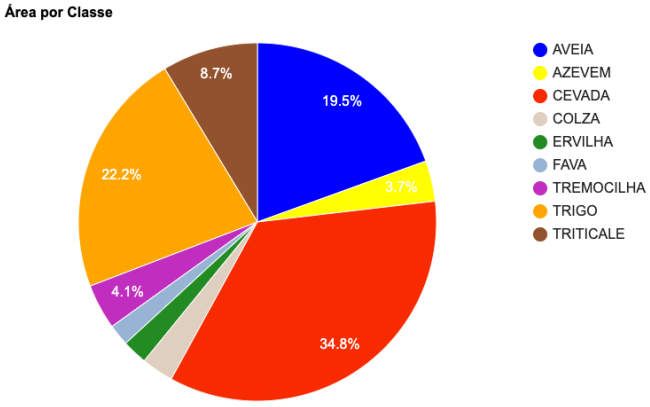

Figura 3.3 Representatividade das classes por área no conjunto de dados

3.2.2 Dados de satélite

A facilidade deste processo recai na utilização do *GEE* para adquirir as coleções de imagens de satélite a serem utilizadas, onde as imagens dos satélites Sentinel atualizadas, estão presentes para serem importadas. Mantendo-se o foco nas culturas de Inverno, e para seguimento do ciclo agrícola, o período selecionado para obtenção das imagens recaiu entre 1 de outubro de 2018 e 15 de maio de 2019. Como limite geográfico, considerou-se o limite da região de interesse (*roi*), tanto para a coleção de imagens de Sentinel-1, como para Sentinel-2. Cada uma das coleções é representada pela sua série temporal de refletância e retrodispersão.

3.2.2.1 Sentinel-1

Na plataforma *GEE*, a coleção de imagens Sentinel-1 *SAR GRD* (*Ground Range Detected*) fornecida, é previamente processada de modo a gerar imagens ortoretificadas e calibradas. Foi importada e filtrada [17], de acordo com a informação existente nos metadados, para o modo do instrumento *IW*

(*Interferometric Wide Swath*), para o tipo de polarização *dual band cross VH* e *VV,* para o número de órbita, 147 e 52, que por sua vez dita, respetivamente, o tipo passagem de órbita ascendente e descendente (Figura 3.4). A resolução das imagens considerada foi de 10 metros [56 - 57].

Provenientes de um instrumento de dupla polarização C-band *Synthetic Aperture Radar (SAR)*, são recolhidas com uma variedade de polarizações e resoluções, daí os filtros serem necessários para obter o produto final desejado. Os valores dos píxeis, são dados em *dB* pelo processamento do coeficiente de retrodispersão registado pelo sensor. É assim obtido um conjunto de imagens da área de interesse tendo em conta o período temporal selecionado.

Com a coleção de imagens Sentinel-1, efetuaram-se duas análises, cada uma com 37 elementos importados. Para a primeira, considerou-se o número de órbita 147, que, cobrindo toda a área de interesse, incluía os satélites *S1*-A e *S1*-B; na segunda, considerou-se o número de órbita 52 e 147 em simultâneo, com filtro para "*platform number*" igual a B, de modo a obter somente elementos *S1-B*, devido ao facto de os elementos *S1-A* (para o número de órbita 52) se encontrarem corrompidos, onde, posteriormente ao serem utilizados nos algoritmos de classificação, deram origem a erro. Na realidade, ambos Sentinel-1A e Sentinel-1B partilham o mesmo plano de órbita, mas com uma diferença de fase orbital de 180º [55].

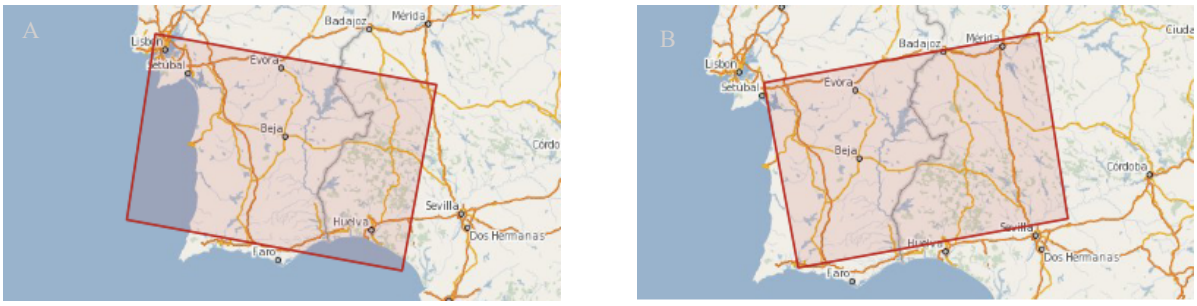

Figura 3.4 Identificação da área coberta pelas imagens Sentinel-1 com número de órbita 52 (A) e 147 (B) respetivamente

Deste modo, obtiveram-se todas as imagens que incluíam estes números de órbita e cobriam a área para a época mencionada na coleção:

a) 37 imagens do Sentinel-1, nº órbita 147 do satélite A e B com valores *dB* para 2 bandas por imagem. Obteve-se uma imagem multibanda com 74 bandas;
b) 19 imagens do Sentinel-1, nº órbita 52 do satélite B. Obteve-se uma imagem multibanda com 38 bandas.

3.2.2.2 Sentinel-2

As imagens Sentinel-2 (S2) são produtos multiespectrais (*MSI*) de elevada resolução [49]. Em *GEE*, importou-se a coleção de imagens S2, de nível 2A (*bottom of atmosphere*),com uma cobertura de nuvens inferior a 15% e que cobria a área de interesse, quadrícula T29SNC. Cada imagem tem 12 bandas no espetro do visível, infravermelho próximo (*VNIR*) e infravermelho de onda curta (*SWIR*) consideraram-se as bandas: B2 (azul), B3 (verde), B4 (vermelho), B5, B6, B7 (limiar do vermelho), B8 (infravermelho próximo), B11 e B12 (*SWIR*) para classificação das parcelas agrícolas. Embora as diferentes bandas apresentem diferentes resoluções obtidas por um sensor multiespectral (*MSI*) em *GEE*, a reamostragem é um processo quase automático pela escolha da escala de 10m, na obtenção dos valores dos píxeis [58].

Neste projeto foram selecionadas 21 imagens de Sentinel-2, com valores de refletância para 9 bandas por imagem. Criou-se uma única imagem multibanda com 189 bandas.

3.3 Metodologia

Considerou-se, dos diversos elementos presentes: *i*) os algoritmos de classificação supervisionada disponíveis na plataforma (*GEE*) [27, 63]; *ii*) o pré-processamento dos dados facultados e respetiva utilização; *iii*) as diferentes abordagens no uso dos dados para construção dos modelos de classificação (Tabela 3.5); e *iv*) a avaliação dos resultados obtidos. Com o objetivo de classificar as culturas, seguiram-se os procedimentos para uma classificação supervisionada. Repartição dos dados em treino (70%) e teste (30%) após o pré-processamento dos mesmos. Recorreu-se aos algoritmos de classificação pré-definidos na plataforma *GEE* com recurso à linguagem *JavaScript*. O código desenvolvido mais extenso tem aproximadamente 1092 linhas; com 3 *scripts* elaborados separadamente para o processamento das imagens Sentinel-1, Sentinel-2, e para ambas as imagens em conjunto.

O esquema da Figura 3.5 representa os processos e os resultados obtidos realizados na plataforma *GEE,* começando pela importação das coleções de imagens *S1* e *S2,* onde em cada coleção se aplicaram os parâmetros de seleção de imagens referidos de modo a obter os produtos desejados [46]; de seguida realizou-se o empilhamento das imagens obtendo uma imagem multibanda com todas as bandas, este processo foi efetuado para cada uma das abordagens referidas posteriormente; à imagem multibanda obtida*,* efetuou-se o *clip* da área de interesse (*roi*) pela *featureCollection [18 - 19]* e com recurso à função *ReduceRegions [34, 61]*, calculou-se a média do valor do píxel por parcela; adicionou-se também uma coluna *random* com valores pseudo-aleatórios a variar entre 0.0 e 1.0 que permitiu efetuar a repartição dos dados em treino (valores menores que 0.7) e teste (valores maiores ou iguais a 0.7). De modo a verificar os dados obtidos exportou-se esse ficheiro em formato *csv.* Neste é possível identificar o valor dos píxeis por parcela para cada uma das bandas existentes na imagem e o valor atribuído por essa coluna *random* a variar entre 0.0 e 1.0 a cada *feature*;

A partir dos dados de treino é realizado o treino dos modelos e os mapas de classificação para cada um dos classificadores. Os dados de teste são usados para a avaliação do modelo pela comparação da classificação efetuada com os dados de referência. Daí resulta a matriz de erro, onde é possível calcular algumas das métricas usadas para avaliação do modelo, como a exatidão global, *F1-score* e *kappa coefficient.*

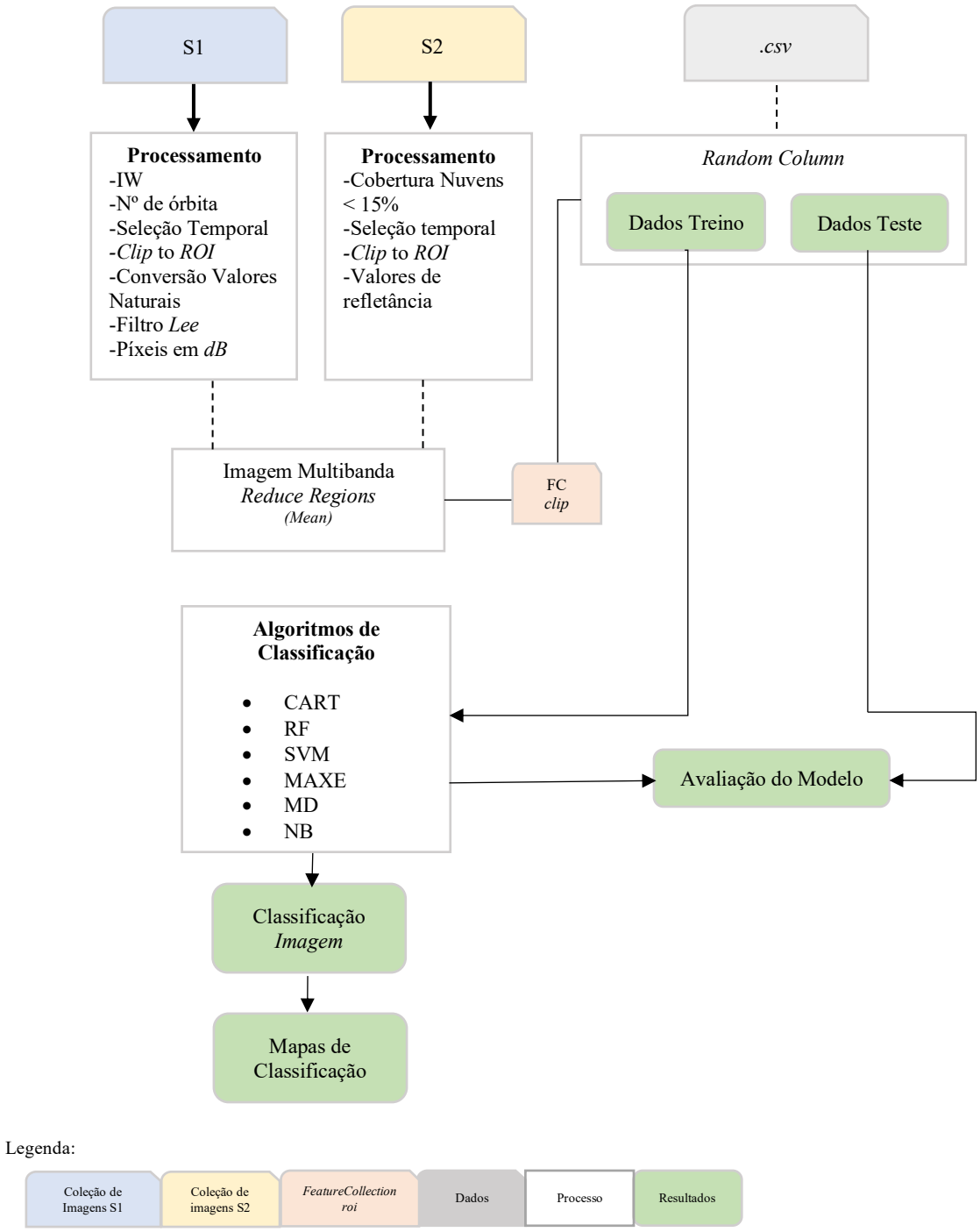

Figura 3.5 Fluxograma do processo computacional na seleção, processamento e classificação dos dados [2]

A plataforma *GEE* oferece um conjunto de algoritmos para classificação [2] e apresenta capacidade no processamento dos dados, bem como disponibiliza formatos de visualização dos resultados e na exportação dos mesmos.

Como estudo comparativo na construção dos modelos de classificação para os diferentes algoritmos, testaram-se várias combinações na inserção dos dados:

i) primeiro para Sentinel-2;

ii) para Sentinel-1, para o número de órbita 147, considerando produtos do satélite A e B (S1-A e S1-B);
iii) para Sentinel-1, com o número de órbita 52, somente com produtos do satélite B (S1-B);
iv) para Sentinel-1, conjugando o número de órbita 147 e 52;
v) Em última análise, a conjugação das imagens Sentinel-2 e Sentinel-1 combinada com os dois números de órbita (Tabela 3.5).

Com o objetivo de tentar obter uma elevada exatidão global nos resultados obtidos, testou-se a plataforma no processo de aprendizagem da classificação de culturas.

3.3.1 *Software* utilizado

Mesmo tendo demonstrado ser uma ferramenta poderosa, houve necessidade da utilização de outros *software*. Algumas das aplicações informáticas utilizadas encontram-se na (Tabela 3.3);

Tabela 3.3 Software utilizado

Software	Aplicação
QGIS	Preparação inicial dos dados: definição do Sistema de coordenadas, corte e validação da *shapefile (*funções *check validity* e *fix geometries)* Visualização e análise dos resultados
GEE	Importação das imagens Sentinel-1 e Sentinel-2 Preparação dos dados: tratamento dos dados vetoriais, com a seleção das culturas de Inverno e atribuição de classes numéricas às respetivas parcelas Extração da média dos valores dos píxeis por parcela e por banda; conversão da informação em formato tubular *.csv* Classificação e validação resultados Visualização e análise dos resultados Exportação dos resultados: *Geotiff* e *.csv [15 - 16]*
Google Sheets	Visualização: dados tubulares (*.csv*)
Google Drive	Armazenamento: imagens classificadas em formato *Geotiff* e dos dados *.csv*
Excel 365	Visualização: dados tubulares (*.csv*)

3.3.2 Tratamento dos dados

O tratamento de dados teve como objetivo obter a média dos píxeis de cada parcela por banda, isto é, a média dos valores de refletância por parcela de cada banda da coleção de imagens Sentinel-2 e a média dos valores de retrodispersão em *dB* por parcela de cada banda da coleção de imagens de Sentinel-1, como preparação dos dados para classificação. A preparação inicial dos dados consistiu na definição do sistema de coordenadas, na validação e do corte das parcelas agrícolas (Figura *3.6*), isto é, na validação de polígonos, obtendo somente parcelas agrícolas válidas e na redução do tamanho devido a requisitos de importação do *GEE*, devido a tempos de cálculo.

Na plataforma *GEE* e com recurso à linguagem *JavaScript*, o tratamento inicial dos dados foi realizado somente pela seleção das culturas de Inverno, e pela alteração dos valores categóricos na identificação da classe de cada parcela para valores numéricos (função *REMAP*).

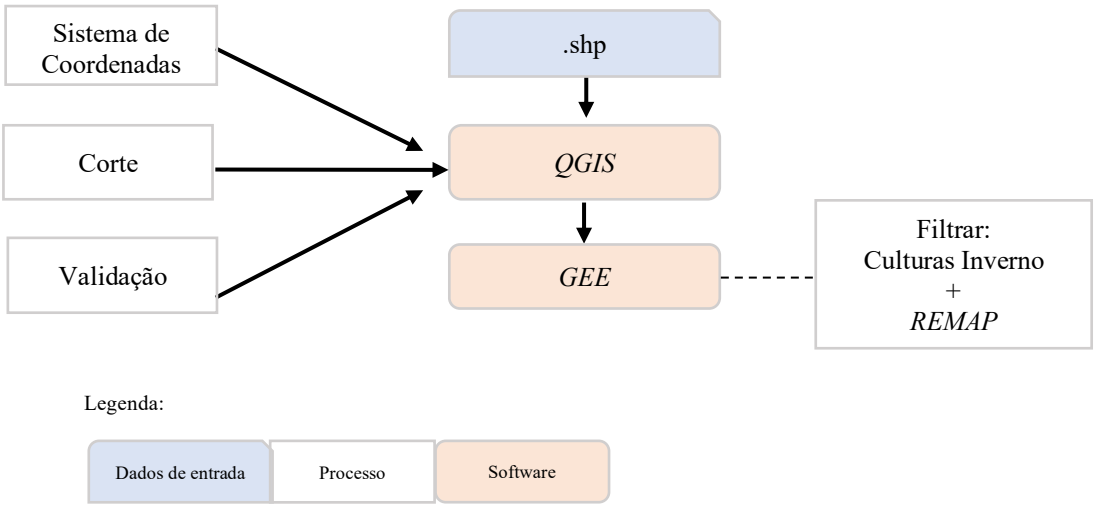

Figura 3.6 Fluxograma do tratamento inicial dos dados

De seguida, foram obtidas as coleções de imagens necessárias para conhecimento dos valores dos píxeis por parcela e filtradas [17] de modo a obter os produtos desejados. A partir de cada coleção, foi obtida uma imagem multibanda que incluía todas as bandas existentes das respetivas coleções. Para o Sentinel-2, obteve-se uma imagem multibanda com valores de refletância e, para Sentinel-1, outra com valores de retrodispersão em *dB* (na combinação de dados a fusão das coleções é efetuada de modo a obter uma multibanda com bandas de ambas as coleções). A partir de uma função específica em *GEE*, extraíram-se os valores da média dos píxeis por parcela em cada uma das imagens multibanda. Obteve-se, adicionando ao ficheiro existente, isto é, à identificação das parcelas geográficas, ao tipo de cultura e à área, o resultado dessa função: as bandas e os valores dos píxeis por parcela. Foi também adicionada uma nova coluna com valores *random* entre 0 e 1 para cada parcela.

Um dos algoritmos de classificação presentes na plataforma, o *Naive Bayes* (NB), funciona somente para valores normalizados; visto que os valores de Sentinel-1 estão presentes em *dB* foi necessária a conversão para valores naturais. As imagens S1 em formato *GRD*, de nível 1 [56] [57], representam o coeficiente de retroespalhamento (*backscatter coefficient*) σ°, sem unidade, que é convertido em *dB*, pela seguinte fórmula (3.1)

$$\sigma^\circ_{dB} = 10\, log_{10}\, \sigma^\circ \tag{3.1}$$

Mas a utilização destes valores no algoritmo de classificação *NB*, não é reconhecida pelo mesmo, sendo necessária a conversão dos valores em *dB* para valores naturais (nível 0), calculados pela propriedade do logaritmo (3.2).

$$\sigma^\circ = 10^{\left(\frac{\sigma_{dB}}{10}\right)} \tag{3.2}$$

Como os resultados obtidos pelo *NB* foram insignificantes, esta abordagem de conversão de valores foi ignorada, bem como a classificação efetuada por este algoritmo.

3.3.3 Dados de Treino

A coluna *random* permitiu a repartição dos dados de modo a ter 70% de dados de treino para o classificador e 30% de teste. Os dados de treino são assim obtidos pelo cálculo da média do valor dos píxeis por parcela a partir da função *Reduce Regions*; ao ser adicionada a coluna *random* (com valores a variar entre 0.0 e 1.0) ao ficheiro é possível efetuar a repartição dos dados, obtendo-se, assim, os dados de treino e teste [68] necessários aos classificadores (Figura *3.7*). Esta *random column* tem um valor de *seed*, que altera a sequência pseudoaleatória da distribuição dos valores. Este foi alterado sete vezes de modo a ajustar os modelos com diferentes dados de treino, e obter um intervalo de variações dos valores de exatidão global, prevenindo o fenómeno de sobreajustamento (*overfitting*).

Posteriormente, o processo foi repetido com outra repartição dos dados (80% para treino e 20% para teste) de modo a verificar se haveria um aumento significativo na exatidão global, que não ocorreu.

Figura 3.7 Fluxograma do processo de extração dos valores píxeis a repartição dos dados

3.3.4 Classificação

Após a divisão dos dados, efetuou-se a classificação das imagens, pela implementação de seis algoritmos [63]: *Classification and Regression Trees* (*CART*), *Random Forest (RF*), *Support Vector Machines (SVM), GMO Maximum Entropy (MAXE), Minimum Distance (MD)* e *Naive Bayes (NB)* disponíveis na plataforma *GEE*.

Os hiperparâmetros dos algoritmos selecionados foram pré-selecionados por defeito (*default*). Exceto para o *RF*, onde houve a obrigatoriedade e a necessidade da escolha de um número de árvores (Tabela 3.4). Deste modo, pretendeu-se que os algoritmos através dos diferentes dados, tanto para Sentinel-1 em imagem *SAR*, como para Sentinel-2 com valores de refletância espectral, pudessem classificar robustamente as culturas de Inverno.

A classificação foi realizada com 5 abordagens referidas anteriormente no ponto 3.3.5.4, na Tabela 3.5. A avaliação e escolha dos melhores classificadores recaíram, de entre as várias métricas existentes, na exatidão global (*accuracy*) obtida por cada classificador (Tabela 4.6). Esta avaliação do modelo deve-se essencialmente às matrizes de erro obtidas a partir do conjunto de dados pré-selecionado para teste em cada abordagem.

Tabela 3.4 Valores *default* dos hiperparâmetros para cada algoritmo de classificação

Algoritmo de Classificação	Valores dos principais hiperparâmetros		Fonte
CART	maxNodes = null	minLeafPopulation = 1	Documentação GEE
RF	numberOfTres = 150 variablesPerSplit = null minLeafPopulation = 1	bagFraction = 0.5 maxNodes = Null seed = 0	
SVM	decisionProcedure = Voting svmType = "C_SVC" kernelType = LINEAR coef0 = null nu = null terminationEpsilon = null lossEpsilon = null	shrinking = true degree = null gamma = null cost = null oneClass = null	
MAXE	weight1 = 0 weight2 = 0.000009999999747378752 (*default*) epsilon = 0.000009999999747378752 (*default*) maxIterations = 100 minIterations = 0		
MD	Metric = euclidean		
NB	Lambda = 0.000001		

A validação foi realizada com o conhecido método *hold-out (*Figura *3.8)*, através da separação pseudoaleatória dos dados numa proporção de 70% para treino e 30% para teste. O método é facilmente recriável na plataforma (ponto 3.3.3). O problema deste método ocorre no fenómeno de sobre-ajustamento (*overfitting*) dos dados [1][67]. Daí, surge a necessidade de alteração da *seed* na escolha dos valores da *Random Column*; para que os dados de treino e teste sejam diferentes (em cada iteração). Este procedimento ocorreu várias vezes de modo a treinar e testar os modelos com diferentes dados.

Os dados de treino servem, assim, para o ajustamento do modelo, e os de teste como dados de referência para analisar a resposta do modelo a novos dados [56]. A exatidão é estimada pela comparação da classificação dos dados de teste (dados de referência), com as decisões do modelo, ao analisar como o modelo executa os dados ainda não vistos, os de teste (chamado *holdout dataset*).

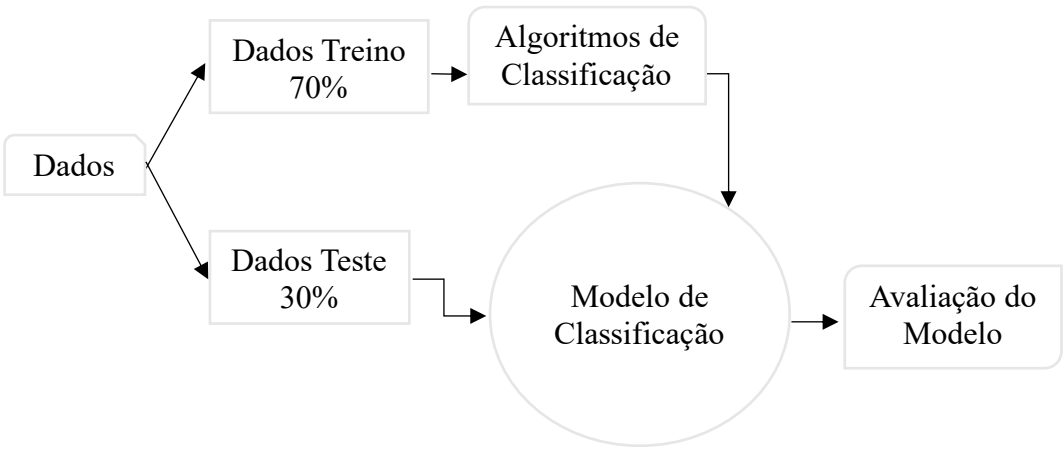

Figura 3.8 Modo de repartição dos dados para avaliação dos modelos de classificação – *hold-out Method*

Efetuou-se este procedimento (Figura *3.8*) para cada abordagem realizada. Como os melhores valores de exatidão foram obtidos para *RF* e *SVM*, a análise de resultados foi efetuada para estes dois classificadores. A validação cruzada [69] é um método ótimo de generalização do modelo. Existem outros métodos de aplicação sem a utilização de dados de teste, mas não se encontram implementados na plataforma *GEE*, para além de requererem mais tempo e poder de computação, algo que, por sua vez, é limitado em *GEE*, como referido anteriormente, pois cada utilizador está limitado por uma quota.

3.3.5 Pós-Classificação

3.3.5.1 Avaliação dos Modelos

a) Matrizes de Confusão e métricas de avaliação dos classificadores

Obtiveram-se as matrizes de confusão (também chamadas matrizes de erro) para cada um dos modelos executados (Tabela 4.7 e Tabela 4.9). Em *GEE,* as matrizes são estruturadas com os dados de referência (as classes reais das culturas no solo) nas linhas e as classes previstas pelo classificador nas colunas. A partir das matrizes, foi quantificada a uniformidade entre a classificação prevista e a realidade nas culturas existentes no solo, pelas métricas de avaliação do desempenho dos classificadores.

A estrutura apresentada de acordo com as células, representa que na linha 0, coluna 0, o número de parcelas que foram identificadas no terreno como classe 0 e que acabaram classificadas com classe 0. Na linha 0, coluna 1 são o número de parcelas que são identificadas no terreno como classe 0 e acabaram classificados com classe 1 e assim sucessivamente. Deste modo, os valores que se encontram fora da diagonal da matriz, em cada linha, são identificados como os erros de classificação por omissão em cada classe. São inversamente proporcionais à exatidão do produtor. E os erros de comissão são, assim, os valores existentes fora da diagonal em cada coluna. São inversamente proporcionais à exatidão do utilizador.

Construído o modelo de classificação, é necessário avaliar quão boas as previsões são. O *GEE*, oferece algumas métricas da avaliação dos classificadores: *accuracy* (exatidão global), *kappa coefficient*, *users accuracy* (exatidão do utilizador, EU), *producers accuracy* (exatidão do produtor, EP) e *F1-score*. A exatidão global (3.3) fornece, assim, a proporção de predições corretas para as predições totais realizadas pelo classificador. Para cada fórmula apresentada, *n* representa o número de parcelas e *r* o número de classes.

$$EG = \left(\frac{\sum_{i=1}^{r} n_{ii}}{\sum_{i=1}^{r} \sum_{j=1}^{r}} \right) * 100 \qquad (3.3)$$

A exatidão do produtor (EP) (3.4) representa a quantidade de dados de referência, isto é, parcelas usadas para teste, que são corretamente classificados. Resulta na divisão do número de parcelas corretamente classificados em cada classe (presentes na diagonal da matriz), pelo número total de dados de referência dessa mesma classe, isto é, o total de cada linha.

$$EP_i = \sum_{j=1}^{r} \frac{n_{ii}}{n_{ij}} \quad (3.4)$$

A exatidão do utilizador (EU) (3.5), por sua vez, representa a proporção de parcelas classificadas numa dada classe ser representada por essa mesma classe no solo. Este valor é computado pela divisão de parcelas corretamente classificadas em cada classe (diagonal da matriz), pelo número total de parcelas que foram classificadas nessa categoria, isto é, o total de cada coluna.

$$EU_i = \sum_{j=1}^{r} \frac{n_{ii}}{n_{fi}} \quad (3.5)$$

F1-Score (3.6) é um valor obtido para cada classe, e calculado a partir dos valores da exatidão do produtor e da exatidão do utilizador; varia entre 0 e 1. É uma métrica que tem em conta como os dados se encontram distribuídos. Para cada um dos classificadores, foi calculada a média aritmética simples do valor de F1, de modo a obter um valor singular [49].

$$F1 = 2 * \frac{EU_i \ * \ EP_i}{EU_i \ + \ EP_i} \quad (3.6)$$

O coeficiente *Kappa* (3.7*)* [9] permite demonstrar a concordância do desempenho de um classificador, tendo em conta a frequência de cada classe, isto é, se os resultados obtidos referem uma correta representação dos dados. Varia entre -1 e 1; onde valores negativos demonstram nenhuma concordância; valores iguais a zero demonstram uma concordância aleatória e valores iguais a 1, uma concordância perfeita.

$$K = \frac{N \ \sum_{i=1}^{r} n_{ii} \ - \sum_{i=1}^{r} (G_i C_i)}{N^2 \ - \ \sum_{i=1}^{r} (G_i C_i)} \quad (3.7)$$

Por ser uma métrica relativa, nem sempre é a melhor a ter em conta para a avaliação de um classificador. Poderá apresentar alguma incoerência na interpretação dos índices de concordância, quando um mau classificador obtém um valor de *kappa* elevado.

b) Representação visual dos resultados – Mapas de Cobertura do Solo

Para além da importância de toda a parte estatística, a representação visual dos resultados da classificação das culturas é fornecida sobre mapas de ocupação do solo. O próprio *IDE* da Google no *code editor* [14] permite a visualização dos resultados. Para cada classificador, foi representado o mapa da classificação de parcelas obtido, de acordo com as cores definidas para cada classe (cultura). Este resultado foi também exportado [10], em formato *Geotiff*.

Como a classificação é realizada ao nível do píxel, e devido a algum ruído obtido nas imagens, recorreu-se à extração (recorrendo a função *ReduceRegions [34]*) do valor da moda de cada parcela; isto é, o valor de classificação com maior representatividade para cada parcela. Este, inicialmente, foi exportado em formato *KML*. A opção de exportar com o formato *shapefile (.shp)* [15 - 16] é fornecida pela plataforma *GEE*, mas dava erro, devido a dois tipos diferentes de geometrias detetados para os polígonos existentes. Uns encontravam-se em *Polygon* e outros em *Geometry Collection*. Isto, por

identificação de geometrias mistas pelo *GEE*. O erro acabou por ser corrigido na plataforma, com a realização de um pequeno *buffer,* que permitiu a exportação do resultado em formato *shapefile* e a correção dos polígonos que se encontravam em *GeometryCollection* e que acabariam por não ser visualizados mesmo em formato *KML,* onde se perderia informação.

De modo a ser visualizada (Figura 4.10), importou-se o ficheiro *.shp* para *QGIS*. Neste, efetuou-se o tratamento dos dados para visualização dos resultados, recorrendo essencialmente à tabela de atributos, pela seleção das colunas de interesse, à identificação das culturas "C1", pelos valores de remapeamento das classes (entre 0 e 8), à área, à coluna "C2", onde é representado o tipo de cultura por valor categórico e à "mode", a coluna com o valor da moda obtido de cada parcela. Seguindo a mesma representação de cores por classe existente na plataforma, selecionou-se a *shapefile* e categorizaram-se as classes pelo mesmo código de cor *html*. Obteve-se uma imagem que revela a classificação obtida com as cores desejadas e estabelecidas em *palette*, onde se obteve as parcelas corretamente classificadas e as não corretamente classificadas identificadas com outro estilo (Figura *4.10*). Esta visualização permite ter uma boa representação visual das parcelas incorretamente classificadas.

O problema inicial destes mapas de classificação, ocorreu na exportação [11] em formato vetorial *KML,* onde, posteriormente, eram convertidos para *shapefile* em QGIS. Neste, todas as parcelas que estavam identificadas por *Geometry Collection* (isto é, com geometrias mistas) não eram incluídas na conversão, por não ser um tipo de geometria reconhecido. Embora este ficheiro fosse válido (com geometrias válidas) em *QGIS*, o mesmo não acontecia em *GEE.* Estes polígonos identificados com geometrias mistas apresentavam linhas, afetando 642 parcelas das 6646. Embora estas geometrias mistas afetem a exportação dos dados, a classificação original efetuada na plataforma não foi afetada.

3.3.5.2 Melhoria dos modelos de classificação

Após os classificadores serem testados, a avaliação do modelo é realizada pela observação dos resultados. Procurou-se melhorar o desempenho: i) primeiro pela ligeira alteração dos parâmetros; ii) pela alteração da abordagem realizada com a inserção dos dados obtidos nos classificadores (Tabela 3.5), pois o uso e aplicação das bandas espetrais é uma boa troca entre exatidão e tempo computacional pelas diferentes abordagens apresentadas na combinação de parâmetros óticos (multiespectrais) e *SAR*. Uma das hipóteses de melhoria do conjunto inicial de dados passaria por conseguir obter uma amostra maior e uniforme dos dados, isto é, com representatividade semelhante em todas as classes; algo que é um pouco difícil, visto que a análise é limitada somente às culturas de inverno e que por limites de importação do tamanho do ficheiro [36], este teve de ser reduzido e limitado à área de interesse de modo a que as culturas fossem cobertas pelos números de órbita previamente referidos (ponto 3.2.2.1), tanto para a coleção S1*,* como para S2*.*

3.3.5.3 Calibração dos Parâmetros

Todos os algoritmos de classificação foram inicialmente usados com os parâmetros definidos por defeito, oferecidos pela plataforma. Selecionaram-se os modelos que apresentavam os melhores resultados: o *RF* e *SVM*. De um ponto de vista prático, e tendo como critério o valor computacional, o classificador *SVM* apresentou maior tempo de computação e de treino. Procuraram-se alterar alguns parâmetros dos algoritmos propostos, com o intuito de melhorar a classificação obtida. Após a alteração, correu-se cada algoritmo sete vezes com a variação do valor da *seed* a cada iteração; para que, a cada iteração, os dados de treino e teste fossem escolhidos aleatoriamente.

De acordo com a literatura, e seguindo o método sugerido [4], para o *RF*, procurou-se aumentar progressivamente o número de árvores. O número de variáveis para *split*, onde é escolhida a raiz quadrada do número de variáveis, é o valor por defeito da plataforma. Os restantes parâmetros, não sendo tão influenciáveis [4], mantiveram-se também definidos por defeito. É um algoritmo robusto e eficiente para classificação, insensível ao fenómeno de sobre ajustamento, mas que requer uma amostra equilibrada de dados. O *SVM*, embora mais sensível à seleção de dados e com maior tempo de computação, é um modelo *black box*, apresenta vários parâmetros críticos não transparentes ao utilizador [4]; nesta análise alterou-se somente o *cost* e o *kernel type*. Posteriormente, manteve-se o *kernel type* em linear e o *cost* definido por defeito. A exatidão obtida era mais elevada.

3.3.5.4 Diferentes abordagens realizadas

Embora estes sejam conhecidos como os classificadores com maiores valores de exatidão na área da deteção remota [4], [49], e sabendo que qualquer alteração aos dados das culturas utilizados para treino do modelo era limitada, pretendeu-se melhorar as classificações obtidas através do uso e aplicação das bandas espetrais por cinco abordagens distintas apresentadas na *Tabela 3.5* com os seguintes dados:

- Imagens multiespectrais provenientes do Sentinel -2;
- Imagens *SAR* com polarização *VV e VH* do Sentinel-1 da órbita 147, do satélite A e B; com e sem correção do *speckle*;
- Imagens *SAR* com polarização *VV e VH* do Sentinel-1 da órbita 52, satélite B; com e sem correção do *speckle;*

Tabela 3.5 Resumo das abordagens realizadas para melhoria dos modelos de classificação

Abordagem	Procedimento	Caraterísticas	Objetivo
A1	Classificação das bandas espetrais provenientes do S2;	Imagem multibanda com 189 bandas; com valores de refletância;	Avaliar o desempenho dos classificadores pelos valores de refletância;
A2	Imagens *SAR* com polarização *VV e VH* do Sentinel-1 da órbita 147; com e sem correção de *speckle;*	Imagem multibanda com 74 bandas do satélite A e B; Dados são os valores em *dB;*	Avaliar o desempenho dos classificadores; se a exatidão melhorava com as imagens corrigidas do *speckle*;
A3	Imagens *SAR* com polarização *VV e VH* do Sentinel-1 da órbita 52; com e sem correção de *speckle*;	Imagem multibanda com 38 bandas do satélite B; dados em *dB*;	Avaliar a exatidão dos classificadores com outra imagem; e se o desempenho melhora com a correção do *speckle*;
A4	Classificação das imagens *SAR* do *S1* – nº de órbita 147 + 52; com e sem correção de *speckle*;	Imagem multibanda com 112 bandas, 74 do satélite A e B; e 38 do satélite B;	Junção das bandas de *S1* para melhoria do desempenho dos classificadores;
A5	Classificação das imagens de *S2* e *S1* (147+52);	Imagem multibanda com 301 bandas;	Junção das bandas de *S2* e *S1* para melhoria da exatidão dos classificadores

Em todas as abordagens realizadas para S1, os modelos foram treinados e testados com e sem correção de *speckle*, de modo a verificar se havia melhoria do modelo. A escolha da correção do *speckle [*51*]*, pretendia reduzir o ruído, contribuindo para uma melhoria visual das mesmas, e nos valores de exatidão obtidos [54].

As cinco abordagens foram realizadas para todos os classificadores. Com uma pequena alteração, pois os valores de S1 teriam de estar normalizados, assumindo a sua forma natural e não em *dB* somente para o classificador *NB*. Como, de todos os classificadores, era o que apresentava muito fraca exatidão, foi desconsiderado logo de início, tanto para a classificação de imagens S2 e S1.

4 Resultados/Discussão

Neste capítulo, apresentam-se todos os resultados para os classificadores nas diferentes abordagens e as matrizes de confusão para os classificadores com os maiores valores de exatidão na abordagem com os melhores resultados. Pretende-se analisar as diferenças dos resultados e visualizar os mapas de classificação; explicando a escolha de certos procedimentos que resultaram na melhoria do desempenho dos classificadores. O valor de *F1* (consta nas matrizes de erro) obtido para cada classe, foi calculado aqui como a média aritmética para cada algoritmo.

4.1 Série Temporal

A visualização das coleções de imagem, pelo ciclo temporal em estudo, encontra-se representada para as imagens do S2 (ponto 3.2.2.2) e S1 (ponto 3.2.2.1), Figura 4.1. Cada uma das coleções é representada pela sua série temporal de refletância e retrodispersão, tendo em conta o número de órbita correspondente, pelo valor derivado (a média) de cada uma das coleções por banda tendo em conta o período temporal do ciclo agrícola e a região em estudo, para a coleção Sentinel-2, na Figura 4.1, e para a coleção Sentinel-1, nº de órbita 147 e 52, na Figura 4.2 e Figura 4.3, respetivamente.

Para a coleção de S2, os valores de refletância apresentam-se em geral mais elevados na época de Inverno com algumas variações, exceto para as bandas B8, B7 e B6, na zona do infravermelho, que apresentam valores de refletância mais elevados na Primavera. Na coleção de S1, os valores de retrodispersão apresentam mínimos e variações entre o mês de novembro e dezembro e valores mais altos nos restantes. Em ambas as figuras a polarização VH atinge os valores menos intensos de retrodispersão na época de inverno com alguns picos e variações.

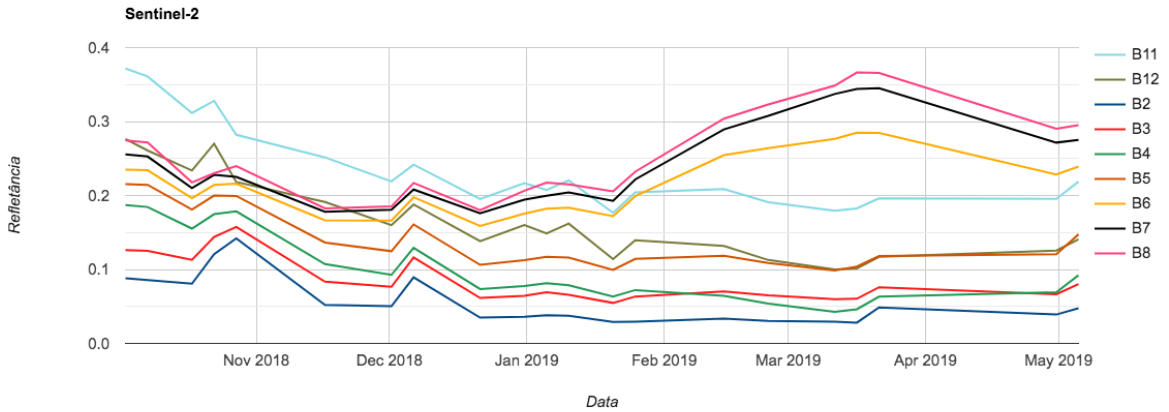

Figura 4.1 Valor médio da refletância de cada banda da série temporal obtida da coleção de imagens de Sentinel-2 para a região em estudo

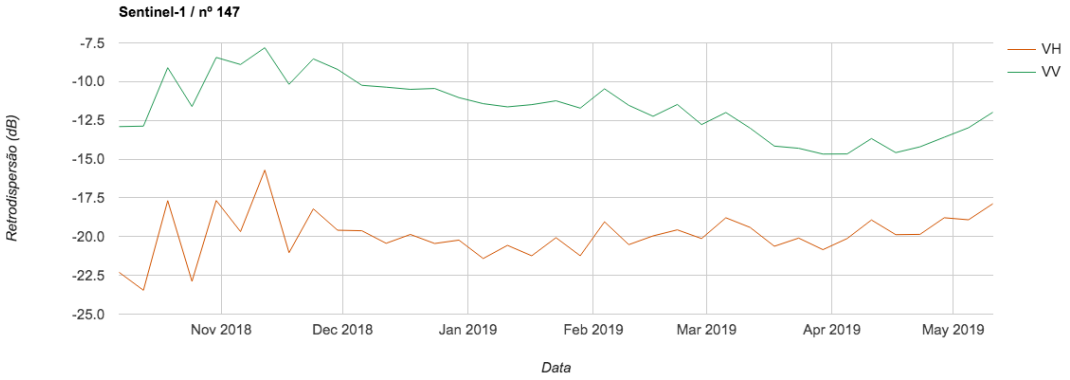

Figura 4.2 Valor médio de retrodispersão de cada banda da série temporal obtida para o nº de órbita 147 da coleção de imagens de Sentinel-1 para a região em estudo

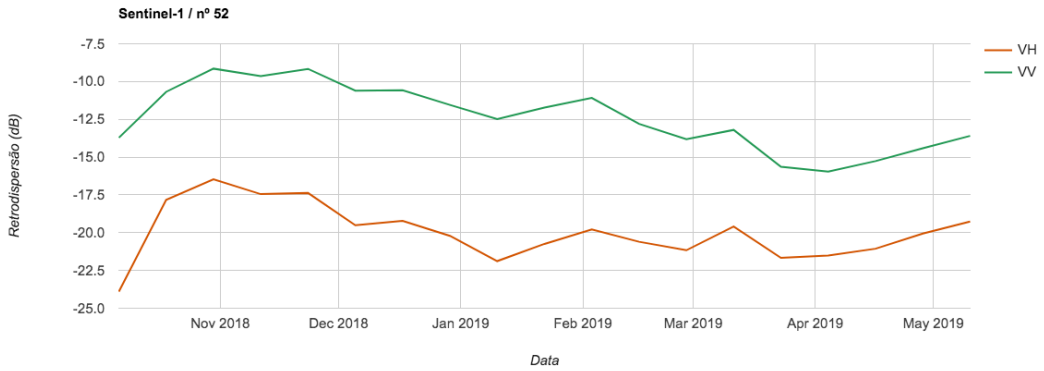

Figura 4.3 Valor médio de retrodispersão de cada banda da série temporal obtida para o nº de órbita 52 da coleção de imagens de Sentinel-1 para a região em estudo

4.2 A1 - Classificação das imagens do Sentinel – 2

Na tabela 4.1 são apresentados os resultados da classificação com a série temporal das imagens S2. Como mencionado anteriormente, consideram-se 189 bandas para a classificação com S2. O *Random Forest* destaca-se na classificação. Este e o *SVM* apresentam os melhores valores de exatidão embora a exatidão global do *MAXE* se aproxime do *SVM*. De entre os algoritmos de classificação, o *SVM* e o *MAXE* são os que apresentam um tempo de computação mais elevado.

Tabela 4.1 Métricas de avaliação do desempenho para a abordagem A1

Algoritmo de Classificação	Dados S2	EG (%)	Coeficiente *Kappa* (%)	F1
CART	189 bandas	55.7	42.2	45.2
RF		70.8	60.4	61.4
SVM		64.3	50.8	45.9
MAXE		64.1	51.4	48.3
MD		39.4	27.5	30.3
NB		15.5	0	2.9

4.3 A2 - Classificação das Imagens *SAR* do S1 – nº de órbita 147

Na segunda abordagem, a classificação foi realizada com 74 bandas para a coleção de imagens com número de orbita 147, número de plataforma A e B com a polarização *VH* e *VV*. Na tabela 4.2 são apresentados os resultados com e sem aplicação do filtro *Refined Lee [51]*, para correção do *speckle*.

Tabela 4.2 Métricas de avaliação do desempenho para a abordagem A2, com e sem filtro de *speckle*

Algoritmo de Classificação	Dados S1 Nº147	EG (%)	Coeficiente *Kappa* (%)	F1
CART	c/filtro	55.7	42.8	45.4
	s/filtro	54.1	40.9	43.8
RF	c/filtro	69.4	59.1	58.3
	s/filtro	69.1	58.6	58.7
SVM	c/filtro	68.1	58.2	58.8
	s/filtro	65.6	55.1	55.6
MAXE	c/filtro	65.2	53.7	51.7
	s/filtro	65.2	54.1	53.2
MD	c/filtro	48.5	36.8	43.5
	s/filtro	48.2	37.2	43.1

Esta escolha de efetuar classificações com e sem valor de *speckle* deve-se ao facto de tentar reduzir a influência negativa de ruído nas imagens *SAR*, mas não há nenhuma função específica implementada no *GEE* para a correção das imagens; esta correção usada refere-se ao filtro *Refined Lee speckle* baseada na *toolbox* do Sentinel-1 do *SNAP*, com código alternativo, fornecido e readaptado, elaborado pelos criadores da plataforma [51]. Foi adaptado para poder correr as coleções de imagens. Como o próprio código só aceita uma banda em singular, leva algum tempo a processar, pois as coleções têm de ser filtradas em separado somente com uma banda. Os resultados têm de ser todos exportados devido ao elevado tempo de computação, não sendo visíveis na consola.

A outra situação deve-se ao facto do filtro de *speckle*, ser normalmente efetuado *a priori* das correções de terreno realizadas às imagens S1; correções que já são fornecidas pela *GEE* na importação da coleção. Faz sentido a correção do *speckle* para melhorias de visualização da imagem e para classificação, mas, neste caso, como as melhorias eram ínfimas, exceto para as abordagens de Sentinel-1 individuais, como tal, foi desconsiderado; essencialmente pelo:

i) tempo de computação necessário;
ii) por ainda não ser uma função específica implementada na plataforma;
iii) por norma, ser efetuado *a priori* das correções de terreno; é possível também que o *kernel* que estivesse a ser considerado não fosse o mais adequado.

4.4 A3 – Classificação das Imagens *SAR* do S1 – nº de órbita 52

A terceira abordagem consistiu noutra escolha de imagens S1 com outros metadados, com número de órbita (*relative orbit number*) 52; neste foram consideradas 38 bandas e somente imagens com número de plataforma B porque as imagens do satélite A estavam corrompidas. Na tabela 4.3 são apresentados os resultados.

Tabela 4.3 Métricas de avaliação de desempenho para A3 com e sem correção de *speckle*

Algoritmo de Classificação	Dados S1 Nº52/B	EG (%)	Coeficiente *Kappa* (%)	F1
CART	c/filtro	54.7	41.2	45.6
	s/filtro	50.6	36.7	43.2
RF	c/filtro	67.2	56.4	59.1
	s/filtro	65.7	55.1	56.9
SVM	c/filtro	65.4	54.7	58.1
	s/filtro	63.8	53.1	55.9
MAXE	c/filtro	62.2	51.1	51.4
	s/filtro	59.6	47.3	48.2
MD	c/filtro	50.9	39.5	44.8
	s/filtro	49.2	38.1	43.9

De todas as abordagens, esta revela-se aquela que apresenta menor valor de exatidão, mesmo para os melhores classificadores, devido à quantidade e ao tipo de bandas existentes.

4.5 A4 - Classificação do conjunto das imagens *SAR* de S1

De todas as abordagens iniciais, a abordagem pela seleção das 112 bandas combinadas da coleção de imagens de S1, de dois tipos de número de órbita diferentes (147 e 52), resultou numa melhoria na classificação previamente obtida, com polarização *VV* e *VH*. A correção do *speckle* não é significativa para uma melhoria nos resultados de classificação, ao contrário do que acontece anteriormente nos resultados de S1 individuais.

Tabela 4.4 Métricas de avaliação do desempenho para a abordagem A4 com 112 bandas

Algoritmo de Classificação	Dados S1 147 + 52	EG (%)	Coeficiente *Kappa* (%)	F1
CART	c/filtro	57.6	45.2	45.2
	s/filtro	58.2	45.5	46.9
RF	c/filtro	71.4	62.1	61.4
	s/filtro	73.0	64.1	63.9
SVM	c/filtro	69.6	60.6	59.9
	s/filtro	72.1	63.7	63.7
MAXE	c/filtro	67.1	56.4	54.2
	s/filtro	68.6	57.7	56.1
MD	c/filtro	51.2	40.1	45.4
	s/filtro	53.4	42.3	46.7

4.6 A5 - Classificação das imagens de S2 com imagens *SAR* do S1

Com o objetivo de ter ganhos nos valores de exatidão dos classificadores e sabendo que diferentes coberturas temporais, ao combinar diferentes bandas poderiam possibilitar esse aumento [25], testou-se, combinando as bandas consideradas de S2, previamente referidas, juntamente com as bandas utilizadas para S1 na abordagem anterior (nº de órbita 147 + 52), com um total de 301 bandas. De todas as abordagens realizadas, esta foi a que atingiu maiores valores de exatidão, Tabela 4.5.

Tabela 4.5 Métricas de avaliação de desempenho para A5 com 301 bandas

Algoritmo de Classificação	Dados S1+S2	EG (%)	Coeficiente *Kappa* (%)	F1
CART	c/filtro	59.7	48.1	49.1
	s/filtro	61.6	50.3	51.7
RF	c/filtro	74.6	66.2	65.3
	s/filtro	76.2	68.3	70.4
SVM	c/filtro	72.1	63.8	61.9
	s/filtro	71.8	63.4	62.3
MAXE	c/filtro	66.7	56.3	52.9
	s/filtro	68.6	59.1	58.7
MD	c/filtro	51.4	40.4	45.4
	s/filtro	51.1	40.1	45.8

Visto que na última abordagem, A5 consideraram-se 301 bandas para classificação e sabendo que as melhorias com a correção pelo filtro de *Lee* não foram tão significativas, essencialmente para as abordagens A4 e A5, optou-se por não considerar estes valores para apresentação dos resultados; para além do elevado tempo de computação necessário para exportação dos mesmos. Embora a exatidão tenha sido melhorada nas abordagens anteriores, é nas abordagens A4 e A5, pela combinação de diferentes bandas, onde a classificação com filtro não demonstrou melhorias significativas.

Devido à elevada exatidão (acima de 70%), obtiveram-se mapas de classificação com menor ruído. Considerou-se a abordagem A5, para apresentação dos resultados completos, com matrizes de confusão e mapas de classificação para os valores que recaem nos classificadores *RF* e *SVM*.

4.6.1 Matrizes de Confusão

Em resumo, após várias iterações, escolheu-se como resultado final os seguintes algoritmos, da abordagem A4 e A5, por serem os que apresentam melhores resultados.

Tabela 4.6 Melhores resultados de classificação obtidos para a A4 e A5 para os algoritmos de classificação *RF* e *SVM* sem filtro de correção de *speckle*

		A4			A5		
Algoritmo de Classificação	Dados	EG (%)	Coeficiente Kappa (%)	F1	EG (%)	Coeficiente Kappa (%)	F1
RF	s/filtro	73.0	64.1	63.9	76.2	68.3	70.4
SVM	s/filtro	72.1	63.7	63.7	71.8	63.4	62.3

Os resultados das cinco abordagens indicam que a combinação de bandas de diferentes sensores, imagens dos satélites S2 e S1, contribui para uma melhor classificação. Observa-se uma melhoria significativa na abordagem A5 para o classificador *RF*. E que a abordagem A4, demonstra por si só, que as imagens *SAR* têm impacto na classificação de culturas. Embora nas análises individuais de S1 o filtro de *Lee* seja benéfico, numa análise mista não mostrou melhorias evidentes, em parte, possivelmente, por não ser o filtro adequado à situação.

Na Figura *4.4* são apresentados os resultados gerais da abordagem A5, pela comparação dos valores de exatidão global e o coeficiente de *kappa* para o algoritmo de classificação *Random Forest* (*RF*) e Suport Vector Machine (*SVM*).

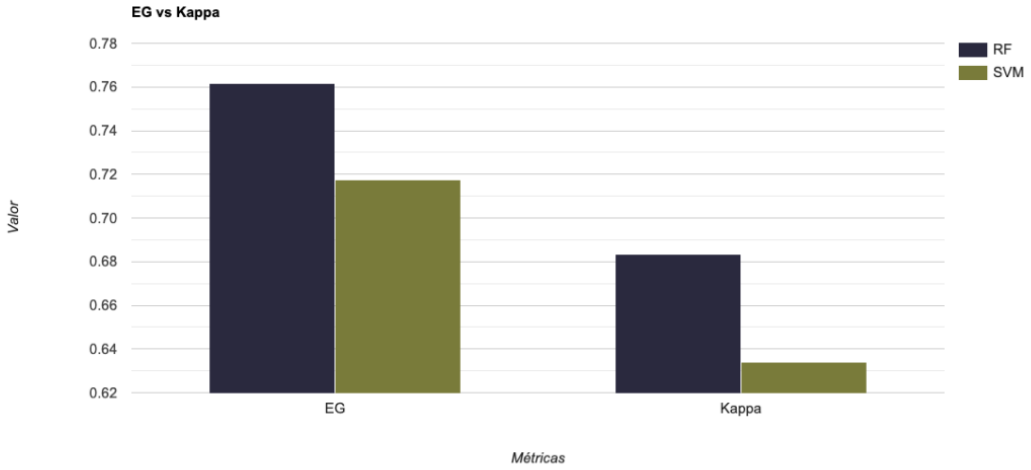

Figura 4.4 Valores de exatidão global e coeficiente *Kappa* para os algoritmos de classificação *RF* e *SVM* na abordagem A5

Tendo em conta a abordagem A5 e aos melhores valores de exatidão obtidos para os classificadores *RF* e *SVM*, apresentam-se nas próximas secções a matriz de confusão para *RF* e *SVM* respetivamente. Cada matriz é representada pelas suas classes e pelas métricas que a compõem. O valor de F1 é representado para cada classe e calculado também como a média aritmética simples dos seus valores. Para uma interpretação visual mais rápida também são apresentados os valores de revocação (*PA*, exatidão do produtor) e precisão (*UA*, e exatidão do utilizador) correspondentes aos valores existentes nas matrizes de erro, tanto do *RF*, como do *SVM* respetivamente.

4.6.1.1 A5 – Matriz de Confusão – *Random Forest (RF)*

Para o *RF*, obteve-se o valor mais alto de exatidão global (*EG*). Na matriz de erro, Tabela 4.7, foi também calculada a frequência (*Fq%*) de cada classe. Como referido anteriormente, o total das colunas representa a classificação efetuada e, as linhas, os dados de referência, isto é, a verdade no terreno.

Tabela 4.7 Matriz de confusão obtida para o algoritmo de classificação *RF* na A5

	Dados classificados												
	0	1	2	3	4	5	6	7	8	Total	Fq%	Re%	F1%
0 Aveia	**203**	5	46	1	1	2	1	49	3	311	16.8	65.3	61.9
1 Azevém	23	**39**	7	1	0	1	4	2	2	79	4.26	49.4	60.5
2 Cevada	22	0	**640**	0	3	1	0	9	2	677	36.5	94.5	86.2
3 Colza	0	0	4	**36**	1	0	1	0	0	42	2.26	85.7	88.9
4 Ervilha	5	0	2	0	**41**	0	1	0	0	49	2.64	83.6	78.8
5 Fava	10	0	9	1	4	**31**	2	1	0	58	3.12	53.4	63.9
6 Tremocilha	23	3	2	0	4	3	**39**	2	2	78	6.22	50	61.9
7 Trigo	33	3	55	0	1	0	0	**333**	1	426	22.9	78.2	79.7
8 Triticale	26	0	43	0	0	1	0	14	**51**	135	7.28	37.8	52.1
Total	345	50	808	39	55	39	48	410	61	1855	-	-	-
Precisão	58.8	78	79.2	92.3	74.5	79.5	81.2	81.2	83.6	-	-	-	-

EG: 76.2% **Coeficiente *Kappa*: 68.3%** ***F1- score*: 70.4%**

A matriz de erro é revelada essencialmente pela diagonal, onde são representadas as parcelas corretamente classificadas. Todos os elementos fora da diagonal são os erros de classificação; onde erros de comissão, são os erros fora da diagonal em cada coluna e os de omissão, fora da diagonal em cada linha. Os erros de comissão estão relacionados com os resultados de classificação. A colza com ~8% de erro de comissão, a cevada, fava, tremocilha, trigo e triticale com cerca de 20% e um erro muito mais elevado ~ 40% para a aveia, revela as parcelas das culturas classificadas como tal foram falsamente classificadas por excesso nas percentagens referidas. São inversamente proporcionais à precisão, à exatidão do utilizador.

Os erros de omissão, valores inversamente proporcionais à revocação (*Re%*), exatidão do produtor, encontram-se mais elevados para a fava e tremocilha com ~50% e para o triticale com ~62%, revelando que estas classes foram erroneamente classificadas em 50% e 60% dos casos respetivamente. Apresenta valores mais baixos, perto dos 5% para a cevada e perto dos 15% para a colza. Os valores mais altos de F1 recaem na cevada, colza e ervilha, onde simultaneamente são obtidos maiores valores de revocação e precisão [49]. Como demonstrado previamente na Figura *3.2*, sabendo que a distribuição das parcelas é maior nas culturas de aveia, cevada e trigo, a frequência calculada também se revela superior nas mesmas.

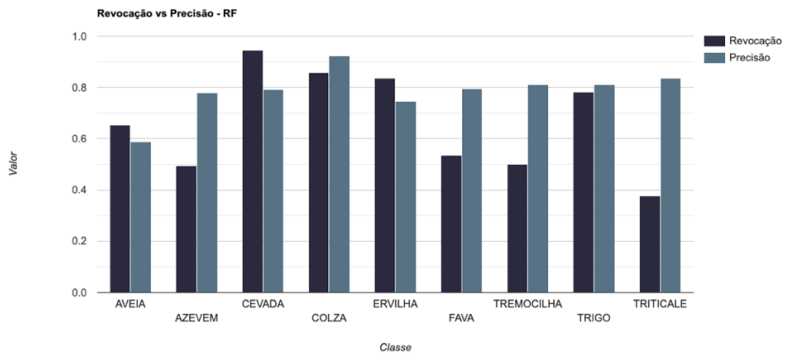

Figura 4.5 Valores de revocação e precisão obtidos para cada uma das classes, a partir da matriz de confusão do algoritmo de classificação *RF (A5)*

A maioria das classes apresenta os valores de precisão maiores que os valores de revocação, respetivamente para o azevém, colza, fava, tremocilha, trigo e triticale, com a colza a atingir valores acima dos 90% (*Figura 4.5*), explicável pela baixa representatividade destas classes nos dados, com a colza a representar 2.5% dos dados (Figura *3.3*). A aveia atinge os valores mais baixos de precisão, demonstrando alguma confusão na classificação das parcelas de aveia pelas restantes classes. E azevém e triticale com os valores mais baixos de revocação (abaixo de 50%), revelando a baixa probabilidade de os dados classificados estarem corretos, tendo em conta a verdade no terreno.

Importância de Variáveis

Nesta abordagem, a importância é calculada para as 302 bandas utilizadas. Como o nome das bandas é composto por um identificador e tipo de coleção, acaba por não ser visível o nome total da banda no gráfico para identificação das bandas com maior e menor importância. Mas, como os gráficos obtidos na plataforma são interativos, foi possível visualizar pela seleção de cada variável, o nome completo da banda e o valor, dados que foram registados na *Tabela 4.8*. Na plataforma e somente para o algoritmo *RF*, obteve-se um gráfico (Figura 7.1- Anexos) com a importância de variáveis para o conjunto total de bandas a ser utilizado para classificação.

Tabela 4.8 Importância de variáveis obtidas para o algoritmo de classificação *RF;* revela as variáveis com maior e menor valor de importância utilizado na classificação do algoritmo

***Variáveis** – A5 – 302 bandas*	*Importância*
S1A_IW_GRDH_1SDV_20190510_VV	122.01
S1A_IW_GRDH_1SDV_20190510_VH	104.01
S1B_IW_GRDH_1SDV_20190504_VV	95.27
S1B_IW_GRDH_1SDV_20190428_VV	93.36
S1B_IW_GRDH_1SDV_20190416_VV	91.53
S1B_IW_GRDH_1SDV_20190416_VH	90.39
S2_SR_20190430_B6	86.29
...	...
S2_SR_20181017_B12	50.16
S2_SR_20190214_B2	47.55
S2_SR_20181231_B4	47.12
S2_SR_20181116_B11	44.36

De todas as variáveis, as bandas do satélite S1 são as que apresentam maior valor correspondente ao período de finais de Abril e início de Maio de 2019, com polarizações *VV* a obter maior importância. Com menor valor, estão as bandas provenientes do S2 do ano de 2018 para o período de Novembro e Dezembro.

Visto que a maior importância é detetada nos meses da primavera, sendo que as culturas em estudo são do período de Inverno, poderia demonstrar resultados dúbios, mas as imagens de S1, *SAR,* oferecem a particularidade de não dependerem de condições atmosféricas, acabando por serem sensíveis às estruturas geométricas e às propriedades das culturas em estudo [42], pela obtenção de dados de dia e noite [54]. Deste modo, as culturas ao apresentarem um crescimento mais lento na época de Inverno [23], permitem que a deteção de variáveis com maior importância ocorra na época da primavera. Outra hipótese de análise seria encurtar o período temporal e limitá-lo somente aos meses de Inverno, mas, com isso, haveria a possibilidade de uma redução da exatidão global e a um maior ruido presente nos resultados dos mapas de classificação.

4.6.1.2 A5 – Matriz de Confusão – *Suport Vector Machine (SVM)*

O *SVM* é o segundo classificador que demonstra bons resultados na classificação. Os valores de frequência calculados mantêm a relevância detetada anteriormente. Os valores mais elevados de F1 recaem nas culturas de cevada, colza e trigo; apresentando os maiores valores de revocação e precisão para as mesmas parcelas (*Tabela 4.9*).

Tabela 4.9 Matriz de confusão obtida para o algoritmo de classificação *SVM* na A5

	Dados classificados												
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**Fq%**	**Re%**	**F1%**
0 Aveia	**201**	10	28	1	3	12	8	31	17	311	16.8	64.6	61.1
1 Azevém	18	**40**	7	1	0	5	1	4	3	79	4.25	50.6	54.4
2 Cevada	33	5	**578**	2	2	12	1	32	12	677	36.5	85.4	84.9
3 Colza	1	0	4	**34**	0	2	0	1	0	42	2.26	80.9	82.9
4 Ervilha	4	0	3	1	**34**	4	3	0	0	49	2.64	69.4	66.1
5 Fava	9	1	8	1	6	**25**	4	3	1	58	3.12	43.1	37.1
6 Tremocilha	20	5	4	0	5	9	**29**	0	6	78	4.20	37.2	45.6
7 Trigo	44	2	35	0	2	7	0	**325**	11	426	22.9	76.3	76.7
8 Triticale	17	5	17	0	2	1	3	25	**65**	135	7.28	48.1	52
Total	347	68	684	40	54	77	49	421	115	1855	-	-	-
Precisão	57.9	58.8	84.5	85	62.9	32.4	59.2	77.2	56.5	-	-	-	-

EG:71.8% **Coeficiente *Kappa*: 63.4%** ***F1-score*: 62.3%**

O erro de comissão apresenta-se para o triticale com 43.5%, para a aveia com 42% e para a fava com ~68%. Já, os erros de omissão mais elevados ocorrem para tremocilha, fava e triticale, na ordem dos 63%, 57% e ~52% respetivamente. Comparativamente com *RF,* as culturas com maior erro de comissão mantêm-se. Os erros de omissão mantêm-se semelhantes para as classes referidas demonstrando que estas representam-se em minoria nos dados, respetivamente 3.2%, 3% e 7% (Figura *3.3*)

Na Figura 4.6Figura 4.6, são apresentados os valores de revocação e precisão para o algoritmo *SVM.* Comparativamente, entre ambos os algoritmos (*RF vs. SVM*) os valores de F1 são elevados para ambas as classes de cevada e colza.

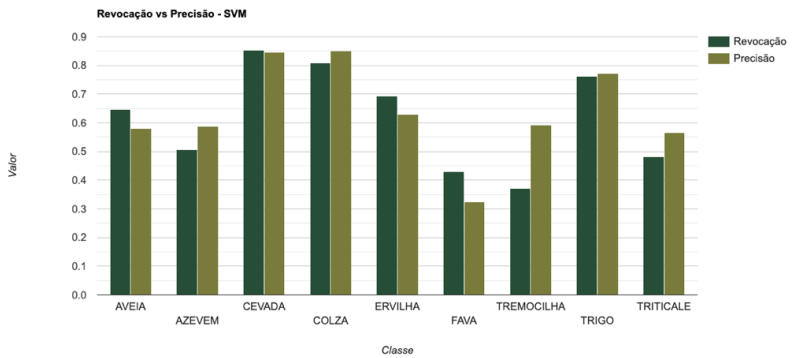

Figura 4.6 Valores de revocação e precisão, obtidos a partir da matriz de confusão do algoritmo de classificação *SVM*

O valor de precisão mantém-se mais elevado para o azevém, colza, tremocilha e triticale. Com valores de 85% de revocação para a cevada e de 85% de precisão para a colza. Existe alguma confusão na classificação da aveia e especialmente na classificação da fava e triticale como se fosse aveia ou cevada.

4.6.2 Refletância e Retrodispersão

O *GEE* oferece dentro de muitas funções, funções de visualização tanto para *FeatureCollection [19]* como para *Image* e *ImageCollection*. Uma das visualizações consiste em derivar os valores das bandas em regiões classificadas [9]. Onde em *xx*, temos as bandas de cada coleção; no eixo dos *yy* os valores de refletância das bandas, tendo em conta o redutor aplicado [50], que foi a média; e a série considerada são os valores por classe. A imagem usada para gerar cada um dos gráficos recaiu na imagem classificada pelos algoritmos *RF* e *SVM.* Analisou-se o comportamento das variáveis pela média dos valores de cada banda para píxeis dentro de regiões classificadas, a partir de uma imagem classificada, obtendo os valores médios de refletância e retrodispersão por banda (Figura 4.7 e Figura 4.8).

Na abordagem, A5, constam as 302 bandas, de diferentes coleções, Sentinel-1 e Sentinel-2 onde cada coleção tem identificadores diferentes. Para a realização dos gráficos seguintes, foi necessário considerar as coleções à parte, devido aos diferentes tipos de bandas envolvidas. Isto é, a visualização foi realizada para as imagens proveniente da coleção Sentinel-*2*; e separadamente para as imagens da coleção Sentinel-1.

Derivou-se os valores das bandas em regiões classificadas para as imagens Sentinel-2 com 189 bandas (abordagem A1, Figura 4.7) e para as 112 bandas da coleção de Sentinel-1 (abordagem A4, Figura 4.8), para ambos os classificadores *Random Forest e SVM.* Esta visualização permite assim analisar o comportamento dos valores médios de refletância e de retrodispersão de cada série temporal por banda e por classe [8].

4.6.2.1 Refletância e Retrodispersão – *RF vs. SVM*

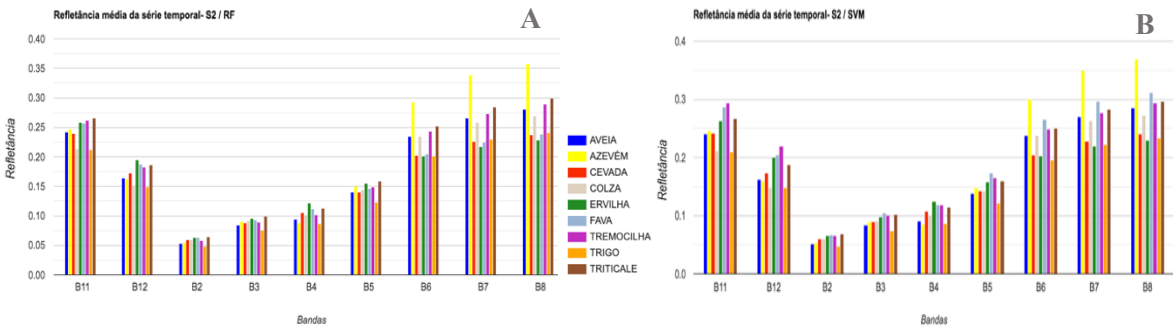

Figura 4.7 Refletância média da série temporal por banda e por classe da coleção Sentinel-2, para os algoritmos de classificação *RF (A)* e *SVM (B)*

Para o *RF,* Figura 4.7-A, o azevém apresenta para as bandas B6, B7 e B8, na região do infravermelho, elevados valores de refletância e a ervilha dos valores mais baixos; enquanto que nas restantes bandas o triticale e a ervilha apresentam os valores de refletância mais elevados.

A média da banda B2, é a que apresenta os menores valores de refletância por classe e a B8 os maiores. O resto das bandas quando assume valores semelhantes por classe, traduzem-se na confusão da

classificação das culturas, mais especificamente, no caso da cevada, trigo e ervilha nas bandas B6, B7 e B8 de ambos os gráficos apresentados, enquanto o azevém e o triticale destacam-se.

Já para o *SVM*, Figura 4.7 - B ocorre uma ligeira alteração no comportamento das classes; o azevém mantém-se na classe com os maiores valores de refletância para as bandas B6, B7 e B8, mas seguido da fava com maior valor. Nas restantes bandas, é a fava que apresenta os maiores valores de refletância, seguido da tremocilha.

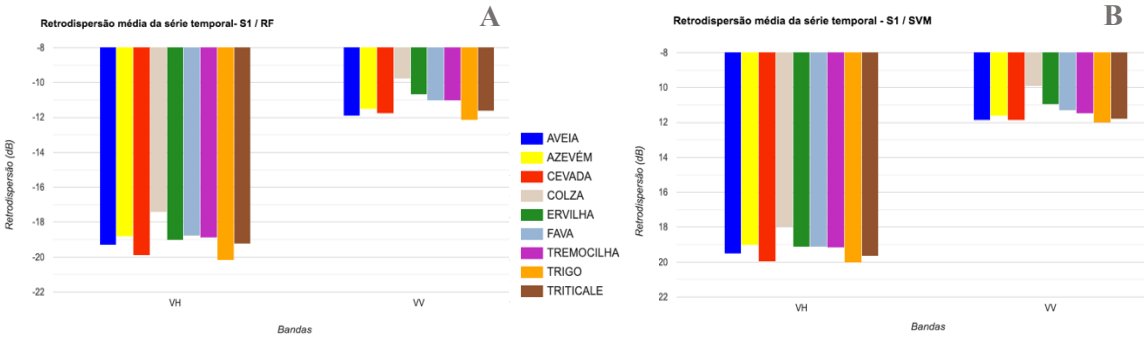

Figura 4.8 Retrodispersão média da série temporal por banda e por classe, da coleção Sentinel-1 (Abordagem A4), para os algoritmos de classificação *RF (A) e SVM (B)*

Analisando os gráficos com os a média dos valores de retrodispersão para a abordagem A4, onde são consideradas as 102 bandas (Figura 4.8) , nota-se uma elevada semelhança, tanto para o *RF (*Figura 4.8 - A*)*, como para o *SVM (*Figura 4.8 - B*)*. Os valores da polarização *VH*, distinguem-se dos *VV*, por muito menor intensidade, com menores valores de retrodispersão nas culturas de cevada e trigo. O comportamento das classes consoantes os valores de retrodispersão por banda são semelhantes; os valores semelhantes entre azevém, ervilha, fava e tremocilha poderão revelar a confusão da classificação nestas culturas.

O *SVM,* Figura 4.8 - B, com comportamento e valores semelhantes ao gráfico dos valores de retrodispersão do *RF,* revela ainda maior confusão do classificador para as culturas previamente referidas, azevém, ervilha, fava e triticale pela semelhança dos valores obtidos na polarização *VH*. Dentro destas mesmas classes, evidencia um comportamento mais discriminatório para a polarização *VV* pela diferença da média dos valores de retrodispersão.

4.6.3 Mapas de Classificação – Abordagem A5

Escolheu-se apresentar os mapas de classificação obtidos para a abordagem A5 para os algoritmos de RF e SVM (Figura 4.9). Nestas figuras é visível a variabilidade no resultado da classificação no interior de algumas parcelas. Esta variabilidade na classificação decorre da abordagem de classificação ao nível do píxel no interior de cada polígono, pois embora ao nível do treino do classificador seja utilizada a média dos píxeis por polígono, a classificação é aplicada a cada píxel da imagem. Devido a isto, propôs-se uma nova abordagem, com a mesma função usada anteriormente (*ReduceRegions*), calcular o valor da moda de cada polígono e exportar esse resultado em *shapefile* para uma melhoria visual (*Figura 4.10*).

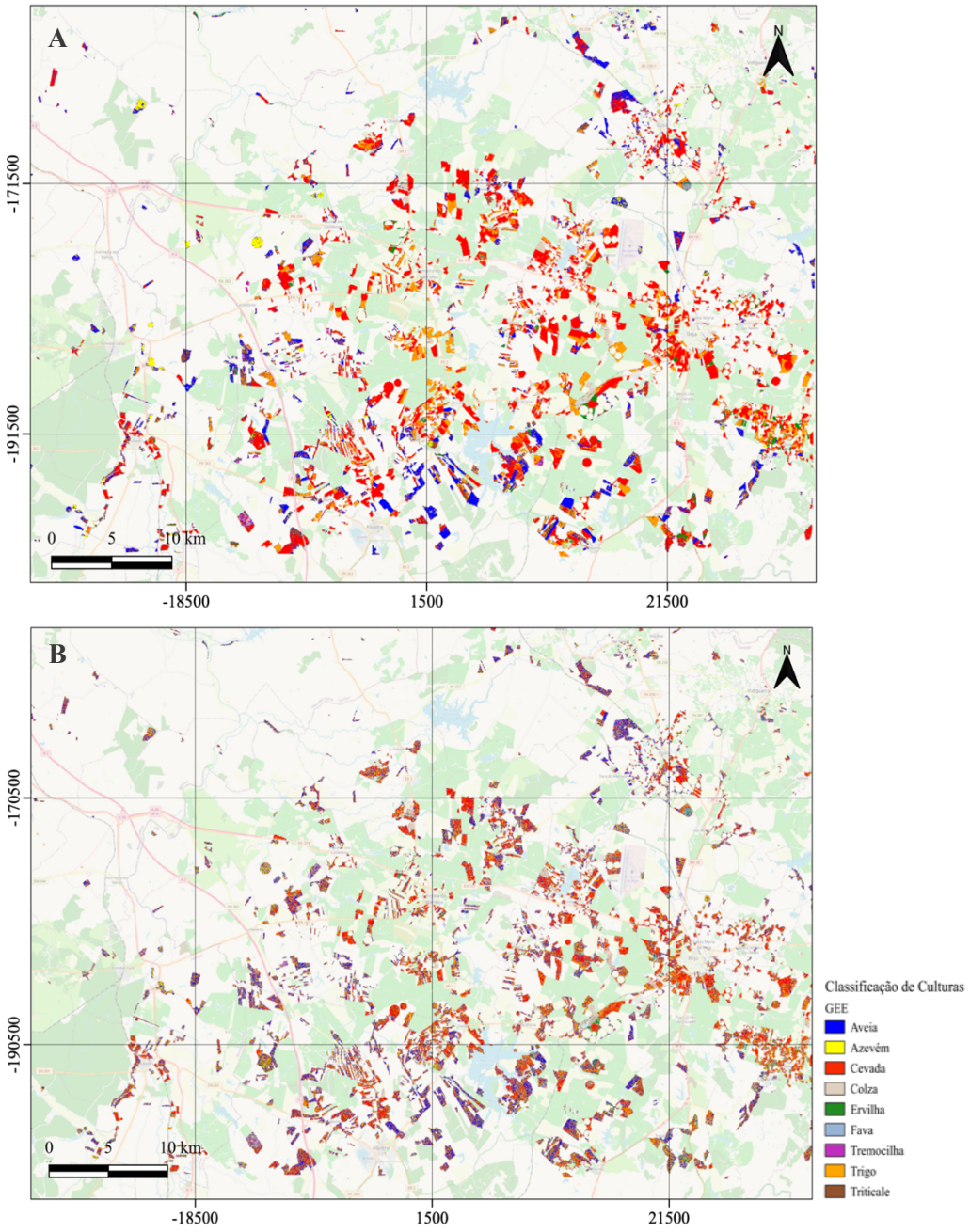

Figura 4.9 Classificação original de culturas obtida na plataforma *GEE* para o algoritmo de classificação *RF* (A) e *SVM* (B)

Como referido anteriormente na exportação das parcelas foram perdidas 642 parcelas restando 5824 parcelas com o resultado da classificação. Deste modo, podia-se ter somente uma perspetiva da classificação ponderada para cada polígono pelo valor de moda, um pouco incomparável com o resultado original obtido da plataforma. Pelo que se pôde apurar, este parece ser um erro comum na importação de *shapefiles* [36] para a plataforma e que, embora tenham geometrias válidas em *QGIS*, as mesmas não são válidas em *GEE,* devido às geometrias mistas. A correção efetuada pela criação de um *buffer* na plataforma regularizou as geometrias presentes no ficheiro, fazendo com que a exportação pudesse ser efetuada em *.shp* e todos os polígonos pudessem ser visualizados.

De futuro e para outras análises, é expectável que, com um aumento do valor de exatidão global, o ruído existente nas imagens tenda a diminuir consideravelmente (não sendo necessária esta abordagem pelo valor da moda. Os mapas da Figura 4.9 dizem respeito à classificação original obtida da plataforma, e os mapas da Figura 4.10 aos valores da moda calculado para cada polígono, ou seja, foi considerada a classe com maior número de píxeis no interior de cada parcela. Os mapas da abordagem A4 estão apresentados no Anexo, Figura 7.6 e Figura 7.7. Todos os mapas encontram-se no sistema de coordenadas ETRS89/PT-TM06.

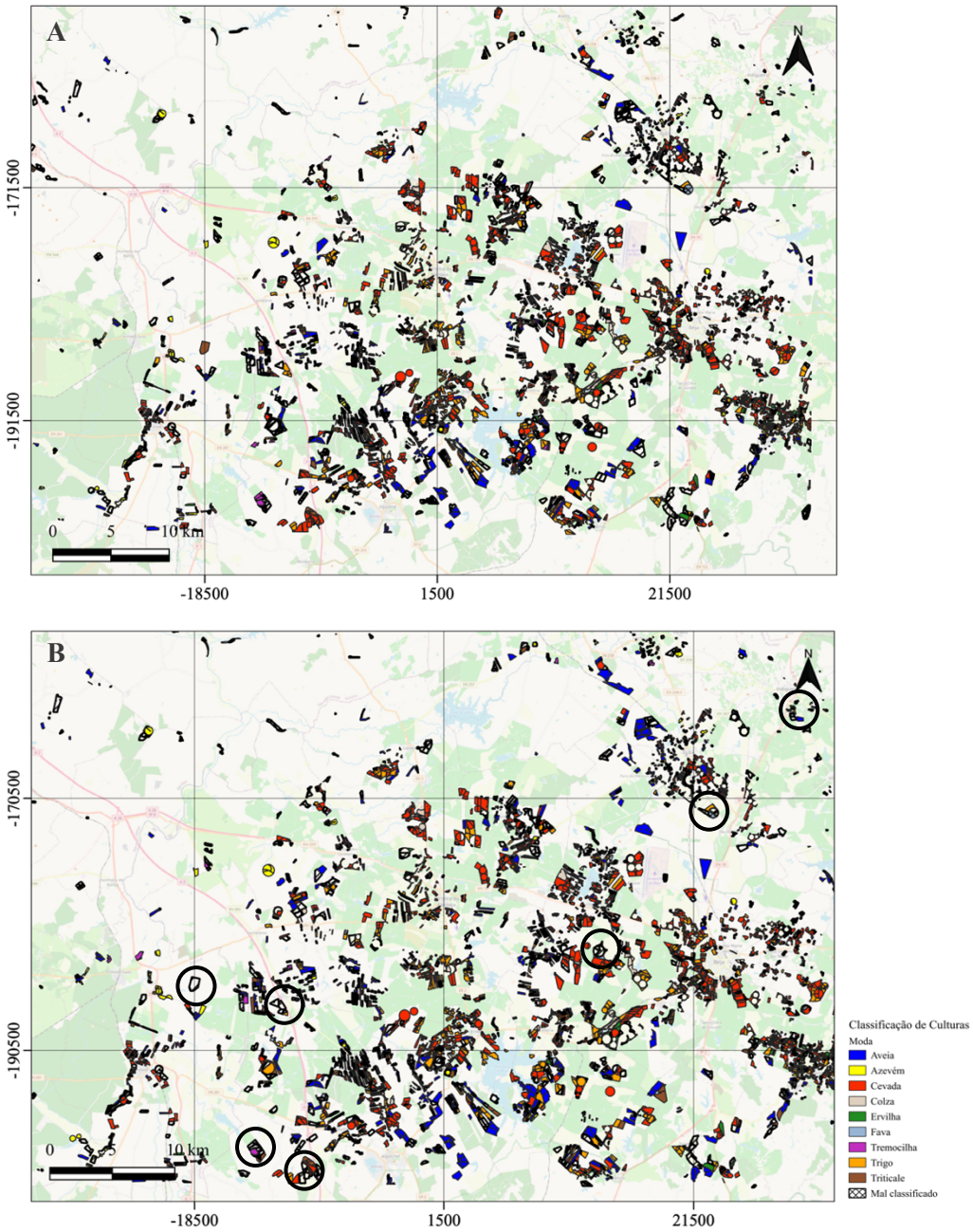

Figura 4.10 Classificação de culturas pelo valor da moda para o algoritmo de classificação *RF* (A) e *SVM* (B)

Ao compararmos os resultados obtidos pela classificação do *SVM (*Figura 4.10 - B*)* com a classificação do *RF (*Figura 4.10- *A)*, é observável algumas parcelas incorretamente classificadas pelo algoritmo de *SVM*, indicado pelas circunferências que foram corretamente classificados pelo algoritmo *RF*. É observável que o *RF* classificou o triticale, trigo e cevada corretamente, enquanto que o *SVM*

classificou mal (parcelas assinaladas) tendo considerado a classe de aveia para estas parcelas, de acordo com o valor da moda pela predominância de píxeis de azuis na Figura 4.9 - B. Como a área em análise é extensa e a maioria das parcelas tem uma representação geográfica pequena, é aconselhável efetuar *zoom* nas imagens para uma melhoria da visualização. A Figura 4.11 corresponde a parte de um grande plano da figura anterior, para uma melhor visualização das parcelas irrisoriamente classificadas.

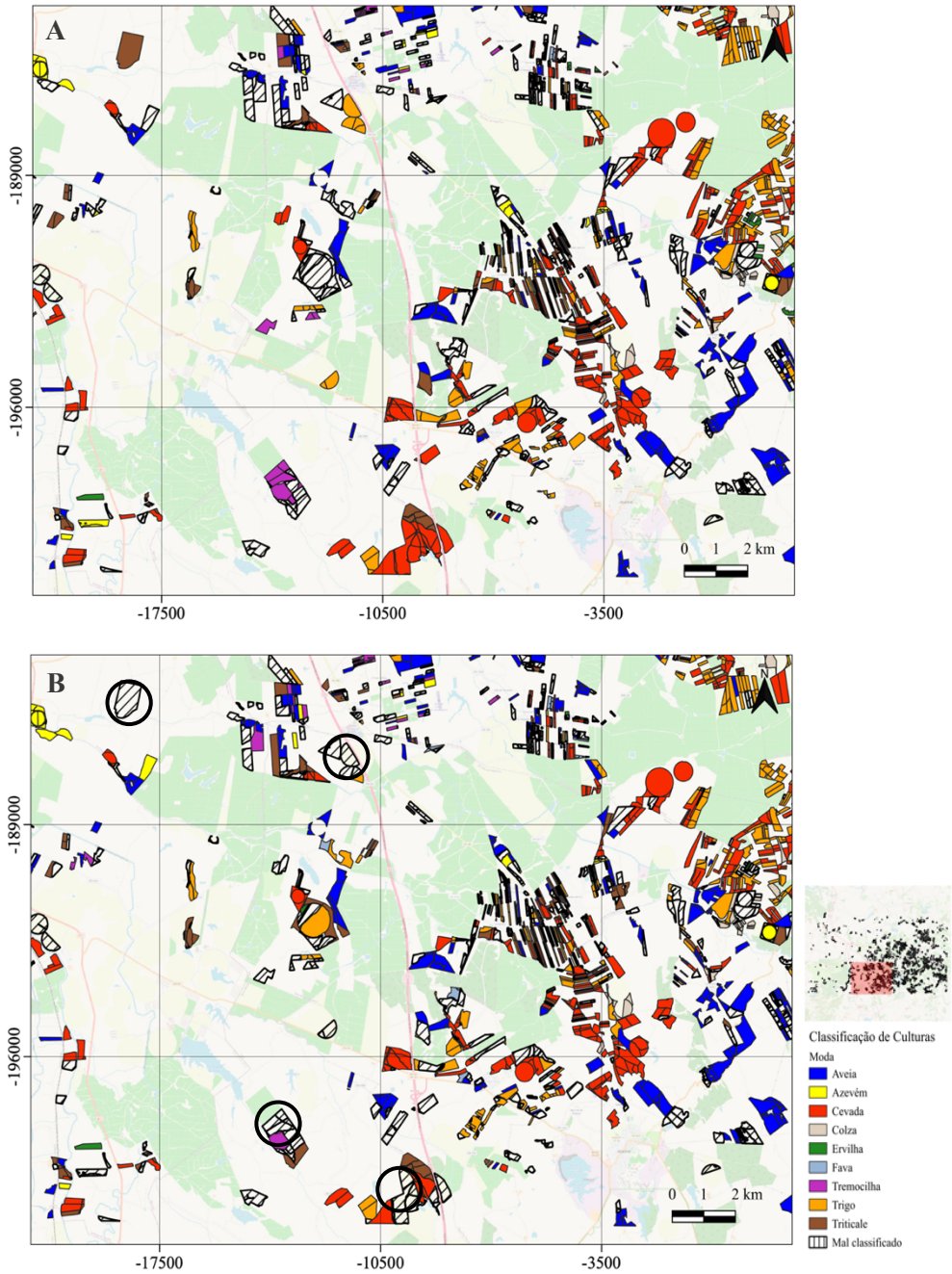

Figura 4.11 Visualização pormenorizada das parcelas erroneamente classificadas do gráfico anterior – para os algoritmos de classificação *RF* (A) e *SVM* (B)

Existe uma predominância das classes de aveia, cevada e trigo no conjunto de dados inicial. O *RF* e o *SVM* de acordo com as parcelas mal classificadas demonstra predominância nos erros de classificação com as classes de cevada, e de aveia. Algo que é visível nas imagens originais a predominância dos píxeis vermelhos e azuis, respetivamente, e demonstrado pelas matrizes de erro que revelam que os

erros de comissão são elevados, para a classe de aveia (~40% e ~30%), e embora não tão elevados para cevada (~15% e ~20%) demonstra que há um número elevado de parcelas classificadas com estas classes que foram erroneamente classificadas, devido à baixa precisão. A classe com melhor desempenho recai no trigo para ambos os classificadores, com parcelas em comum bem classificadas, tendo em conta a visualização pormenorizada (Figura 4.11) e erros de comissão semelhantes para ambos os classificadores.

Os algoritmos de classificação *RF* e *SVM* obtiveram os melhores valores de exatidão global em todos as abordagens nos processos de classificação. O *RF* demonstrou-se como o melhor classificador em todas as abordagens realizadas, facto que é evidenciado para a abordagem A5 nos mapas de classificação pela redução do ruído presente quando comprado com o algoritmo *SVM* (Figura 4.9)*;* tendo a fusão de dados (*SAR* e multiespectrais) contribuído para tal.

Face à literatura existente e referenciada ao longo do projeto, estes resultados confirmam em parte a melhoria dos valores esperados. Apresentam como elemento novo, a utilização de imagens *SAR* combinadas com dois números de órbitas distintos (abordagem A4), sem correção de *speckle* a obter melhores valores de exatidão global do que as imagens multiespectrais utilizadas individualmente na abordagem A1; mas justificado pelo facto de esta classificação ser efetuada em época de Inverno, onde ocorre uma limitação das imagens multiespectrais devido às condições climatéricas de nebulosidade.

Algumas limitações foram encontradas, na utilização da plataforma através da quota existente para cada utilizador pelo tempo de computação máximo permitido (5’ de modo a ter resultados visíveis na consola) e pelos algoritmos de classificação disponibilizados, alguns com pouca relevância; e ainda a desproporcionalidade presente por classe nos dados parcelares.

Como direções futuras a serem propostas nesta área de investigação, assumindo uma melhoria da proporção das classes dos dados parcelares iniciais, a classificação por outros algoritmos, recorrendo à área de *deep learning* pela utilização de redes neuronais; algo que de momento não é oferecido pelo *GEE.* E possivelmente a melhoria dos dados de treino, pela redução da dimensionalidade, reduzindo a redundância que poderá ser existente entre bandas e na seleção das variáveis de maior importância.

5 Conclusão

Este projeto teve como principal objetivo avaliar os resultados de classificação de culturas por cinco abordagens distintas, utilizando a plataforma *cloud GEE*: A1, com a utilização de imagens Sentinel-2; A2, na utilização de imagens Sentinel-1 número de órbita 147; A3, imagens Sentinel-1, número de órbita 52 do satélite B; A4, o conjunto de imagens Sentinel-1, com número de órbita 147 e 52; e A5, pela fusão de imagens *SAR* com multiespectrais (óticas), avaliando as suas potencialidades para processamento, análise e visualização de dados

Dos algoritmos de classificação considerados e testados, o *RF* demonstrou ter melhores resultados na generalização dos modelos, em todas as abordagens realizadas na classificação de culturas de Inverno, atingindo o valor máximo de 76.2% de exatidão global para a abordagem A5, que consta na fusão de dados *SAR* e óticos num total de 301 bandas, 112 de imagens do satélite Sentinel-1 com polarização *VV e VH* e 189 de imagens Sentinel-2 com nove bandas cada, usadas para caraterizar os perfis de cada cultura. O *RF* oferece, na sua análise, a discriminação das variáveis utilizadas pela importância das mesmas, de modo a uma maior compreensão dos dados e melhoria da qualidade dos modelos. É um classificador rápido e robusto, mas sensível à amostra de dados.

As imagens obtidas de diferentes sensores fornecem informação distinta, e embora as imagens *SAR* sejam ricas em informação espacial não contêm a informação espetral favorável na classificação de culturas que os sistemas óticos dispõem. Esta fusão visa beneficiar, principalmente em época de Inverno, a existência de imagens viáveis em condições atmosféricas adversas, que acaba por ser limitante para os sistemas óticos devido a nebulosidade presente. Faz com que a informação oferecida seja muito mais útil do que a informação obtida de imagens de um só sensor.

O conjunto de imagens *SAR* (número de órbita 147 e 52 – abordagem A4), revelaram bons resultados pela deteção de caraterísticas da fenologia das culturas de inverno, e têm demonstrado principalmente nos métodos de classificação supervisionada uma melhoria proveitosa e interessante da performance dos modelos; os sistemas multiespectrais com imagens Sentinel-2, por si só, não têm tão bons valores de exatidão como verificado na abordagem A4. Na abordagem A2 e A3, onde são consideradas as imagens SAR, número de órbita 147 e 52 respetivamente, o filtro de *Lee* para correção do *speckle* beneficia um pouco a melhoria dos resultados; a questão é o tempo de computação necessário para a utilização do mesmo na plataforma por código alternativo.

GEE demonstrou ser uma ferramenta com elevado desempenho e performance, tendo como principal objetivo, neste projeto, a análise e classificação de imagens de satélite. Esta, permite dar uma resposta coerente e objetiva nos problemas e necessidades dos utilizadores para a abordagem da análise de imagens de satélite e nos estudos de deteção remota, pela cedência de uma plataforma específica que não está dependente de armazenamento e capacidade de processamento local para lidar com mega dados e que visa permitir aos utilizadores uma ferramenta de análise, processamento e visualização de dados sem que estes necessitem de infraestruturas específicas e dispendiosas.

Com uma elevada oferta de funções, dispõe de uma biblioteca que com recurso à linguagem *JavaScript* permite lidar com uma elevada quantidade de dados geospaciais, fornecendo imagens de satélite prontas a usar, sem a necessidade de efetuar o seu pré-processamento outrora necessário (com recurso

a equipamento computacional de elevada performance e com espaço de disco). Apresenta algumas falhas, enquanto não existem funções específicas para algumas situações. Uma delas é relativamente ao filtro de correção do ruído do *speckle* presente nas imagens de S1 (em que a melhoria deste ruído beneficia em parte os modelos de classificação); e outra, quanto ao tempo máximo de processamento aceite pela plataforma na consola, em que qualquer resultado fora desse tempo computacional, os resultados têm de ser exportados, mas que se justifica tendo em conta o acesso comunitário da plataforma por quota por vários utilizadores.

As imagens *SAR* demonstram forte poder de análise, tendo em conta a polarização, frequência e caraterísticas temporais selecionadas, sensíveis à estrutura fisiológica e às propriedades geométricas apesar de, em geral, haver uma dormência das culturas em período de Inverno, enquanto que os dados óticos baseiam-se nas caraterísticas espetrais dos alvos. Com elevada resolução e um frequente tempo de revisita, permite ter um elevado número de imagens, com um papel importante na classificação de culturas.

A melhoria dos resultados provavelmente seria conseguida, com uma amostra proporcional por classe dos dados iniciais, pois a proporção equilibrada dos dados de treino visa a melhoria dos classificadores; isto é visível devido à confusão dos classificadores a discriminar as culturas de fava, ervilha e triticale com baixo valor de *F1-score*. A redução da dimensionalidade ou a seleção de *features* com maior importância poderia ser também uma abordagem a reter para a melhoria dos resultados, embora dependente da amostra proporcional dos dados. Posteriormente uma análise com *deep learning* através de redes neuronais pela utilização do *TensorFlow (*biblioteca pertencente à *Google)* [43] seria uma perspetiva interessante para a classificação dos dados mantendo as mesmas abordagens.

Considerando a multiplicidade de sensores e o grande volume de dados existente hoje em dia, na área da deteção remota, o *GEE* revela-se como uma plataforma extraordinária pela oferta de um extenso catálogo de dados, capaz de elevado processamento para análise, visualização e classificação de dados.

6 Referências Bibliográficas

[1] Allibhai, E., 2018. Hold out vs. Cross validation in Machine learning. [Online]. Disponível em: https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f. [Acedido:10- Fevereiro-2020]

[2] Aparício, S., 2018. A Walk-Through on Machine Learning Techniques for Sentinel Big Data Fusion. [Online]. Disponível em: http://phiweek2018.esa.int/agenda/files/presentation260.pdf. [Acedido:10- Fevereiro-2020]

[3] Becker, W., Ló, T., Johann, J., Mercante, E. 2020. Statistical features for land use and land cover classification in Google Earth Engine, *Remote Sensing Applications: Society and Environment*, vol. 21

[4] Belgiu, M., Drăguţ, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24-31

[5] Big Data Fusion. [Online]. Disponível em: https://blogs.esa.int/philab/2019/01/18/big-data-fusion/. [Acedido: 10-Janeiro- 2021]

[6] Brooke, S., D'Arcy, M., Mason, P., Whittaker, A. 2020. Rapid multispectral data sampling using Google Earth Engine. *Computers & Geosciences,* vol. 135 - 6

[7] Casu, F., Manunta, M., Agram, P., Crippen, R. 2017. Big Remotely Sensed Data: tools, applications and experiences. *Remote Sensing of Environment*, vol.202, pp. 1-2

[8] Charts by Image Classes – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/charts_image_by_class. [Acedido:10- Fevereiro-2020]

[9] Chen, Y., Interpretation of Kappa Value. [Online]. Disponível em: https://towardsdatascience.com/interpretation-of-kappa-values-2acd1ca7b18f. [Acedido:10- Fevereiro-2020]

[10] Classification and Regression Trees for Machine Learning. [Online]. Disponível em: https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

[11] Client Server» [Online]. Disponível em: https://developers.google.com/earth-engine/guides/client_server. [Acedido: 5-Outubro- 2020]

[12] Copernicus Open Access Hub - ESA. [Online]. Disponível em: https://scihub.copernicus.eu/dhus/#/home. [Acedido:1- Junho-2020]

[13] Descals, A., Verger, A., Yin, G., Peñuelas, J. 2020. Improved Estimates of Arctic Land Surface Phenology Using Sentinel-2 Time Series, *Remote Sensing*, vol.12

[14] Earth Engine Code Editor – Google Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/playground. [Acedido:10-Abril-2020]

[15] Exporting charts and images – Google Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/tutorials/tutorial_api_07. [Acedido:11-Abril-2020]

[16] Exporting Data – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/exporting. [Acedido:10- Fevereiro-2020]

[17] Filtering an Image Collection – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/ic_filtering. [Acedido: 7-Janeiro-2020]

[18] FeatureCollection Information and Metadata – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/feature_collection_info. [Acedido:10-Fevereiro-2020]

[19] «FeatureCollection Overview». [Online]. Disponível em: https://developers.google.com/earth-engine/guides/feature_collections. [Acedido:10 Fevereiro-2020]

[20] Feature and FeatureCollection Visualization. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/feature_collections_visualizing. [Acedido:10-Fevereiro-2020]

[21] Filtering a FeatureCollection – Google Earth Engine Guides. [Online]. Disponível em:https://developers.google.com/earth-engine/guides/feature_collection_filtering. [Acedido:10-Fevereiro-2020]

[22] Feature Property Charts – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/charts_feature_by. [Acedido:10- Fevereiro-2020]

[23] Fontaneli, R., Santos, H., Fontaneli, R., Oliveira, J., Lehemen, R., Dreon, G. Gramíneas Forrageiras Anuais de Inverno [Em linha]. Disponível em: http://www.cnpt.embrapa.br/biblio/li/li01-forrageiras/cap4.pdf. [Acedido: 20 – Dezembro-2020]

[24] Fundamentals of Classification and Regression Trees. [Online]. Disponível em: https://mathanrajsharma.medium.com/fundamentals-of-classification-and-regression-trees-cart-e9af0b152503. [Acedido: 5-Janeiro- 2021]

[25] Gaetano, R., Cozzolino, D., D'Amiano, L., Verdoliva, L., Poggi, G. 2017. Fusion Of SAR-Optical Data For Land Cover Monitoring, vol.978, pp.1-4

[26] Get started with Earth Engine – Google Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/getstarted. [Acedido:10 -Abril -2020]

[27] Gorelick, N., Classification and Clustering. [Online]. https://docs.google.com/presentation/d/1BFZVhUVKiANHvSi5ApbmjPqiOEHNBV_NizDVzJJ05tI/htmlpresent .[Acedido:10- Fevereiro-2020]

[28] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment,* vol. 202, pp.18-27

[29] Guo, H., Wang, L., Liang, D. 2016. Big Earth Data from space: a new engine for Earth science, Science Bulletin, vol. 61, pp. 505-513

[30] Hao, P., Tang, H., Chen, Z., Yu, L., Wu, M. 2019. High resolution crop intensity mapping using harmonized Landsat-8 and Sentinel-2 data. *Journal of Integrative Agriculture*, vol. 18, pp 2883-2897

[31] Hajjaji, Y., Boulila, W., Farah, I., Romdhani, I., Hussain, A. 2020. Big data and IoT-based applications in smart environments: A systematic review». *Computer Science Review*, vol. 39

[32] Histograms – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/charts_image_histogram. [Acedido:10- Fevereiro-2020]

[33] ImageCollection Information and Metadata - Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/ic_info. [Acedido:10- Fevereiro-2020]

[34] ImageCollection Reductions – Google Earth Engine Guides. [Online]. Disponível em:https://developers.google.com/earth-engine/guides/reducers_image_collection. [Acedido:10-Fevereiro-2020]

[35] Image Regions Charts – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/charts_image_regions. [Acedido:10- Fevereiro-2020]

[36] Importing Table Data – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/importing. [Acedido:10- Fevereiro-2020]

[37] Introduction to Google Earth Engine. [Online]. Disponível em: https://geohackweek.github.io/GoogleEarthEngine/01-introduction/. [Acedido:10- Fevereiro-2020]

[38] Izquierdo-Verdiguier, E., Zurita-Milla, R. 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, vol. 88

[39] Kulkarni, S., Rege, P. 2020. Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion*, vol. 59, pp.13-29

[40] Li, Y., Ma, J., Zhang, Y. 2020. Image retrieval from remote sensing big data: A survey. *Information Fusion*, vol. 67, pp. 94-115

[41] Liu, C., Analysis of Sentinel-1 SAR data for mapping standing water in the Twente region. [Online]. Disponível em: https://webapps.itc.utwente.nl/librarywww/papers_2016/msc/wrem/cliu.pdf. [Acedido:10- Fevereiro-2020]

[42] Liu, C., Chen, Z., Shao, Y., Chen, J., Hasi, T., PAN, H. 2019. Research advances of SAR remote sensing for agriculture applications: A review. *Journal of Integrative Agriculture*, vol. 18, pp. 506-525

[43] Machine Learning in Earth Engine – Google Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/machine-learning. [Acedido: 11 – Abril-2020]

[44] Managing Assets – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/asset_manager. [Acedido:10- Fevereiro-2020]

[45] Mann, G., McDonald, R., Silberman, N., Mohri, M., Walker, D. 2009. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models»

[46] Mapping over an Image Collection – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/ic_mapping. [Acedido: 7-Janeiro-2020]

[47] Meroni, M., D'Andrimont, R., Vrieling, A., Fasbender, D., Lemoine, G., Rembold, F., Seguini, L., Verhegghen, A. 2020. Comparing land surface phenology of major European crops as derived from SAR and multispectral data of Sentinel-1 and -2. *Remote Sensing of Environment*, vol. 253

[48] Rish, I. 2001. An Empirical study of the Naïve bayes Classifier. *In Proceedings of the IJCAI Workshop on Empirical Methods in Artificial Intelligence*, vol 3, pp. 41-46

[49] Gouette, C., Gaussier, E. 2005. A Probabilistic Interpretation of Precision Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science*, vol. 3408, pp. 345–359

[50] Reducing an ImageCollection - Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/ic_reducing.
[Acedido:10- Fevereiro-2020]

[51] Refined Lee speckle filter [for S1] – Google Earth Engine Developers Groups. [Online]. Disponível em: https://groups.google.com/g/google-earth-engine-developers/c/ExepnAmP-hQ/m/8oE8kIoiCAAJ. [Acedido:10- Novembro-2020]

[52] Richards, D., R., Belcher, R., N., Global Changes in Urban Vegetation Cover, *Remote Sensing*, vol.12, pp. 23

[53] Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., Camps-Valls, G. 2020. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, vol. 63, pp. 256-272

[54] Schlund, M., Erasmi, S. 2020. Sentinel-1 time series data for monitoring the phenology of winter wheat. *Remote Sensing of Environment*, vol.246

[55] Sentinel-1 Algorithms – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/sentinel1. [Acedido:10- Fevereiro-2020]

[56] Sentinel-1 Preprocessing – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/sentinel1#sentinel-1-preprocessing. [Acedido:10-Fevereiro-2020]

[57] Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected, log scaling – Google Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD. [Acedido: 5-Janeiro-2020]

[58] Sentinel-2 MSI: MultiSpectral Instrument, Level 2-A – Earth Engine Data Catalog. [Online]. Disponível em: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR. [Acedido: 5-Janeiro-2020]

[59] Shetty, S., 2019. Analyses of Machine Learning Classifiers for LULC Classification on Google Earth Engine, Master Thesis, Faculty of Geo-Information Science and Earth Observation of the University of Twente.

[60] Silva, I. 2020. Avaliação de metodologias de aprendizagem automática na classificação de culturas agrícolas com base em imagens do Sentinel-2. Tese de Mestrado, Faculdade de Ciências da Universidade de Lisboa

[61] Statistics of Image Regions – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/reducers_reduce_regions. [Acedido:10-Fevereiro-2020]

[62] Stromann, O., Nascetti, A., Yousif, O., Ban, Y. 2019. Dimensionality Reduction and Feature Selection for Object-Based Land Cover Classification based on Sentinel-1 and Sentinel-2 Time Series Using Google Earth Engine, *Remote Sensing*, vol.12, pp.76

[62] Synthetic Aperture Radar (SAR) Basics. [Online]. Disponível em: https://developers.google.com/earth-engine/tutorials/community/sar-basics. [Acedido:10-Fevereiro-2020]

[63] Supervised Classification – Earth Engine guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/classification. [Acedido: 5-Janeiro-2020]

[64] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua,L., Lopez, A. 2020. A comprehensive Survey on Support Vector Machine Classification: Applications, challenges and trends. *Neurocomputing*, vol. 408, pp. 189-215

[65] Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B. 2020. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 164, pp.152-170

[66] Time Series Charts – Google Earth Engine Guides. [Online]. Disponível em: https://developers.google.com/earth-engine/guides/charts_image_series.
[Acedido:10- Fevereiro-2020]

[67] Faraway, J. 2013. Does Data Splitting Improve Prediction?. *Statistics and Computing*, vol. 26.

[68] Training and Test sets: Splitting Data. [Online]. Disponível em: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data.
[Acedido: 15-Fevereiro-2020]

[69] Berrar, D. 2019. Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology,* vol.1, pp 542-545

[70] You, N. Dong, J. 2020. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 161, pp.109-123

[71] Zhao, W., Qu, Y., Chen, J., Yuan, Z. 2020. Deeply synergistic optical and SAR time series for crop dynamic monitoring. *Remote Sensing of Environment*, vol. 247

[72] Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., Jie, W. (2014). Remote sensing big data computing: Challenges and opportunities, *Future Generation Computer Systems*, vol.51, pp.47-60

7 Anexos

A.

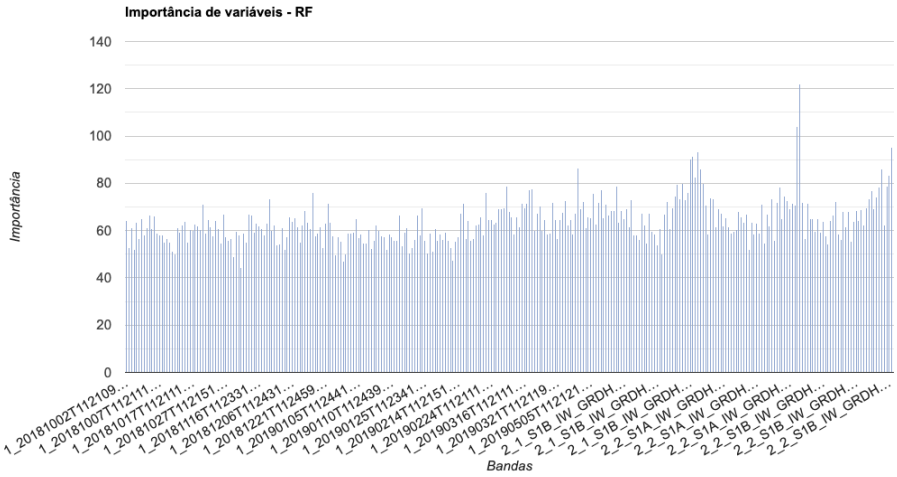

Figura 7.1 Gráfico da importância de variáveis para as 302 bandas usadas no algoritmo de classificação *RF* para a abordagem A5

- Abordagem A4 – Algoritmo de classificação *RF*

Tabela 7.1 Matriz de confusão do algoritmo de classificação *RF* para a abordagem – A4

	Dados classificados												
	0	1	2	3	4	5	6	7	8	Total	Fq %	Re %	F1%
0 Aveia	**217**	3	43	1	2	3	8	45	6	328	16.9	66.2	60.9
1 Azevém	23	**21**	17	0	3	0	5	2	2	73	3.77	29.7	42.0
2 Cevada	22	0	**651**	2	0	1	1	24	5	706	36.5	92.2	85.2
3 Colza	0	1	1	**33**	5	1	0	0	0	41	2.12	80.5	82.5
4 Ervilha	7	0	6	2	**43**	4	0	0	1	63	3.26	68.3	68.8
5 Fava	11	0	4	0	6	**27**	6	1	1	56	2.90	48.2	55.1
6 Tremocilha	16	0	1	1	2	5	**37**	1	0	63	3.26	58.7	59.2
7 Trigo	58	1	50	0	1	1	3	**331**	11	456	23.6	72.6	75.7
8 Triticale	30	1	49	0	0	0	2	14	**52**	148	7.65	35.1	46.1
Total	384	27	822	39	62	42	62	418	78	1934	-	-	-
Precisão %	56.5	77.7	79.2	84.6	69.4	64.3	59.7	79.2	66.6	-	-	-	-

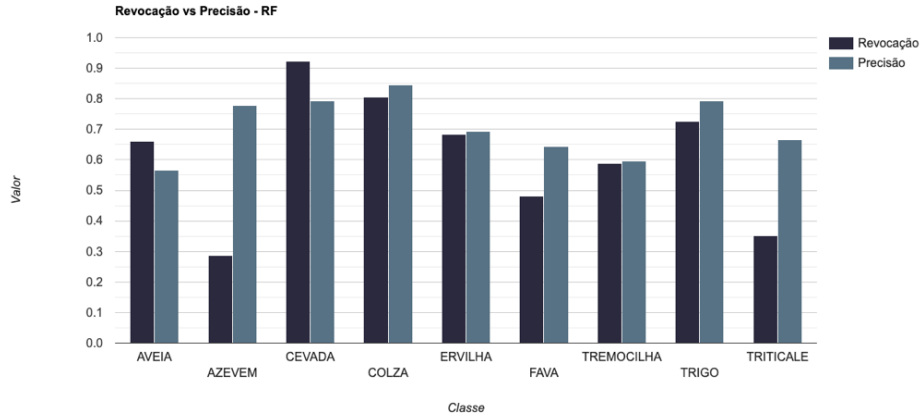

Figura 7.2 Análise dos valores de revocação e precisão obtidos a partir da matriz de confusão do classificador *RF*; revocação a azul escuro e precisão a azul claro – A4*;*

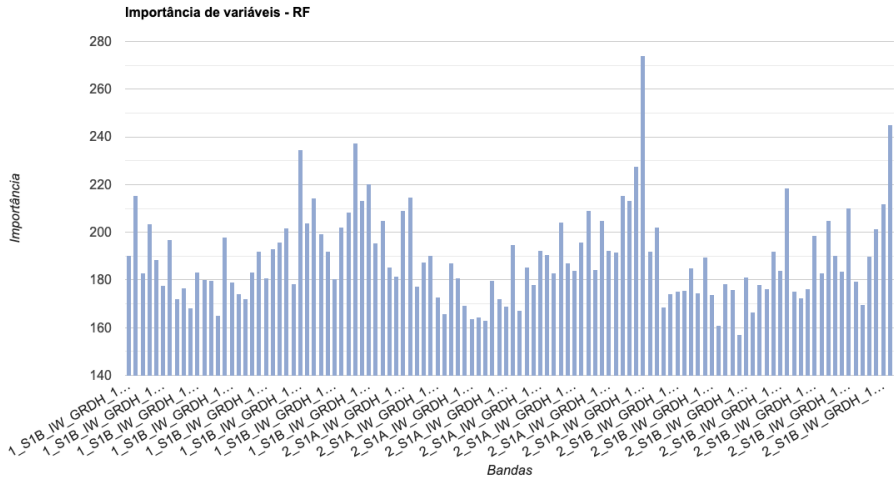

Figura 7.3 Importância de variáveis das 112 bandas usadas no algoritmo de classificação *SVM* para a abordagem A4

Tabela 7.2 Seleção das cinco melhores e piores variáveis com o valor de importância, obtidas a partir do gráfico anterior

Variáveis *– A4 – 112 bandas*	*Importância*
S1A_IW_GRDH_1SDV_20190510_VV	273.94
S1B_IW_GRDH_1SDV_20190227_VV	234.50
S1B_IW_GRDH_1SDV_20190416_VV	237.33
S1A_IW_GRDH_1SDV_20190510_VH	227.74
S1B_IW_GRDH_1SDV_20190504_VV	245.21
S1B_IW_GRDH_1SDV_20181217_VV	165.17
S1A_IW_GRDH_1SDV_20181217_VH	163.67
S1A_IW_GRDH_1SDV_20181229_VH	163.02
S1B_IW_GRDH_1SDV_20181211_VH	160.96
S1B_IW_GRDH_1SDV_20181223_VV	157.02

- Abordagem A4 – Algoritmo de classificação *SVM*

Tabela 7.3 Matriz de Confusão dos resultados do algoritmo de classificação de *SVM* para A4

	Dados classificados												
	0	1	2	3	4	5	6	7	8	Total	Fq %	Re %	F1%
0 Aveia	**227**	8	31	3	6	7	9	29	8	328	16.9	69.2	63.2
1 Azevém	12	**34**	7	0	3	2	4	7	4	73	3.77	46.6	49.2
2 Cevada	35	3	**602**	2	5	6	2	41	10	706	36.5	85.3	84.6
3 Colza	2	0	1	**30**	3	3	2	0	0	41	2.1	73.2	77.9
4 Ervilha	5	0	1	1	**42**	9	3	1	1	63	3.3	66.6	65.1
5 Fava	6	4	9	0	1	**33**	2	1	0	56	2.9	58.9	51.5
6 Tremocilha	13	8	2	0	2	3	**33**	0	2	63	3.3	52.4	54.5
7 Trigo	64	4	38	0	3	7	2	**327**	11	456	23.6	71.7	73.9
8 Triticale	26	4	26	0	1	2	1	22	**66**	148	7.7	44.6	52.8
Total	390	65	717	36	66	72	58	428	102	1934	-	-	-
Precisão %	58.2	52.3	83.9	83.3	63.6	45.8	56.8	76.4	64.7	-	-	-	-

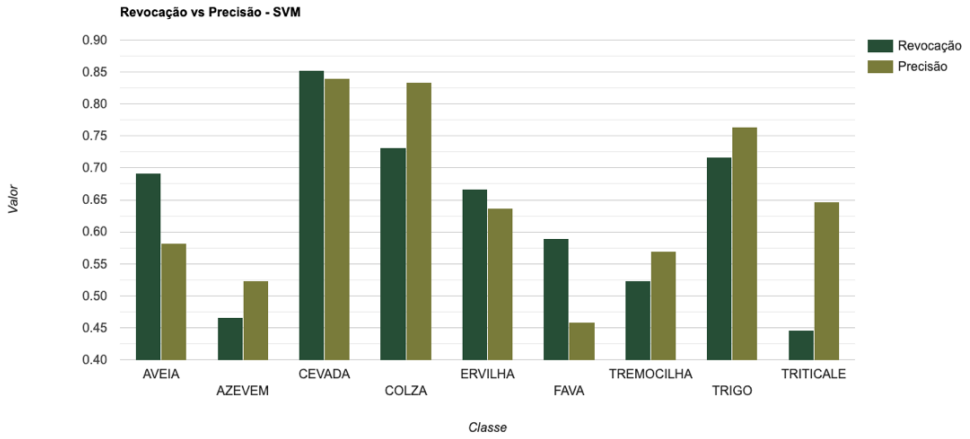

Figura 7.4 Análise dos valores de revocação e precisão obtidos a partir da matriz de confusão do algoritmo de classificação SVM

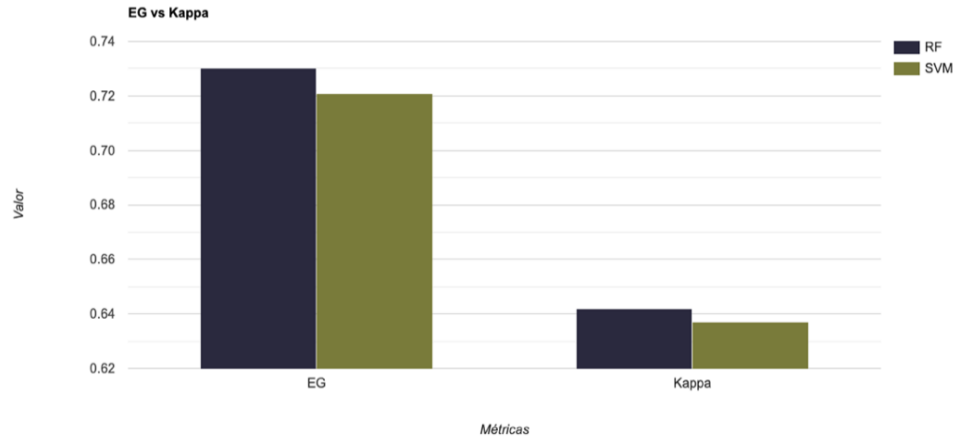

Figura 7.5 Comparação entre os valores de Exatidão Global (EG) e *Kappa* Coefficient para cada um dos classificadores, RF (a azul), e SVM (a verde);

- Mapas de Classificação – *RF vs. SVM*

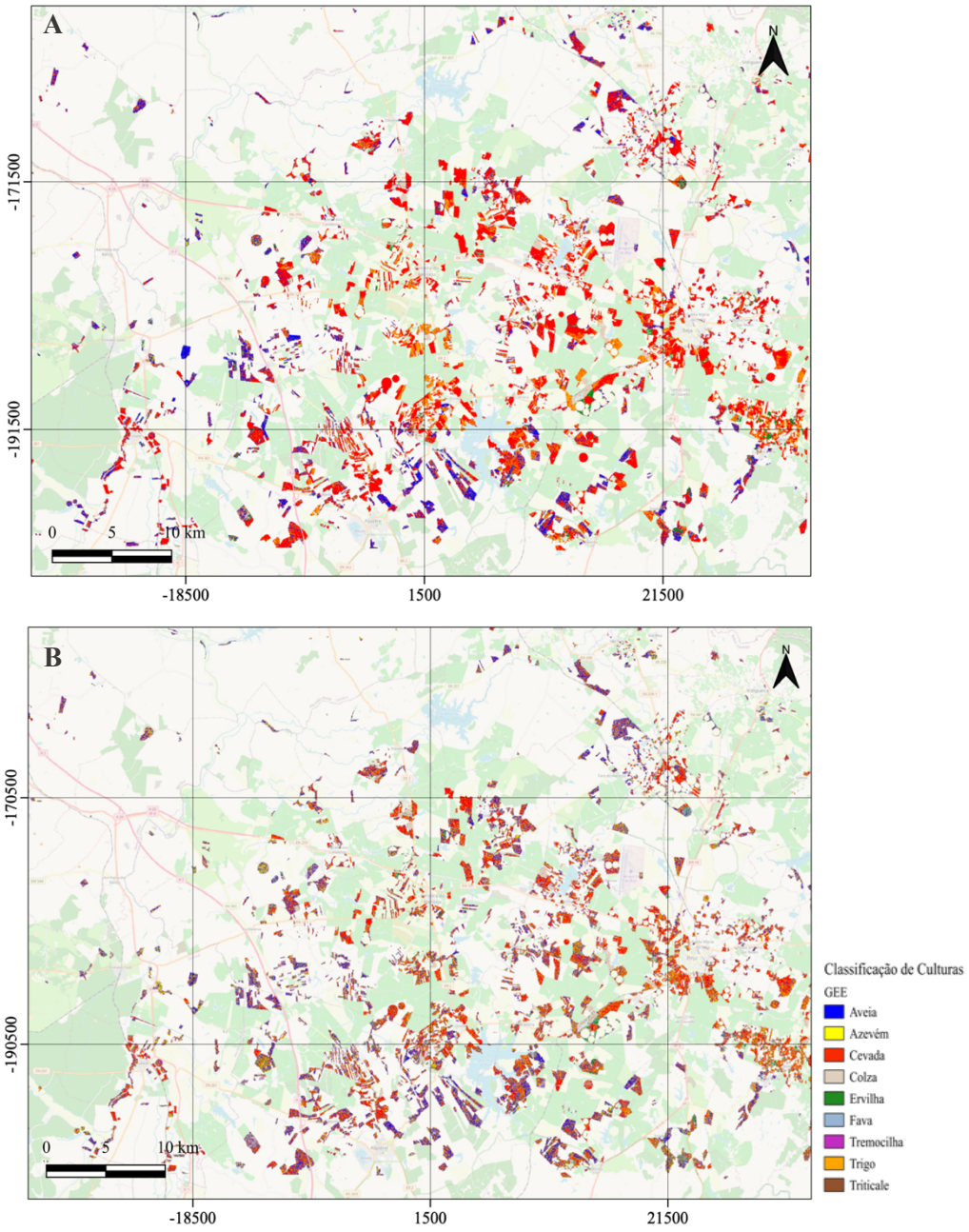

Figura 7.6 Classificação original de culturas obtida em GEE para o RF (A) e SVM (B) para a abordagem A4; visível o ruído presente na classificação pelo SVM;

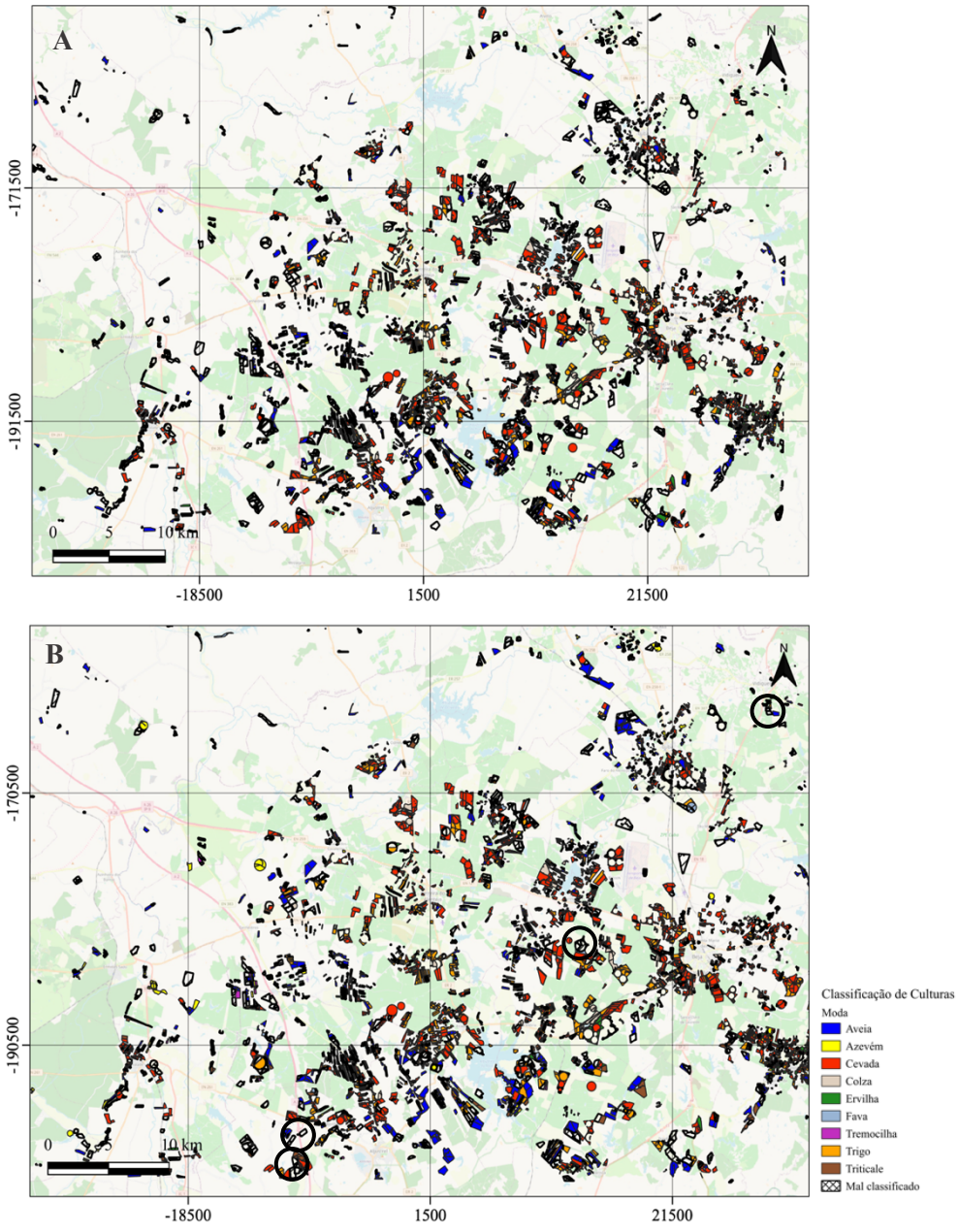

Figura 7.7 Classificação de culturas pelo valor da moda obtido por parcela para o *RF* (A) e o *SVM* (B)

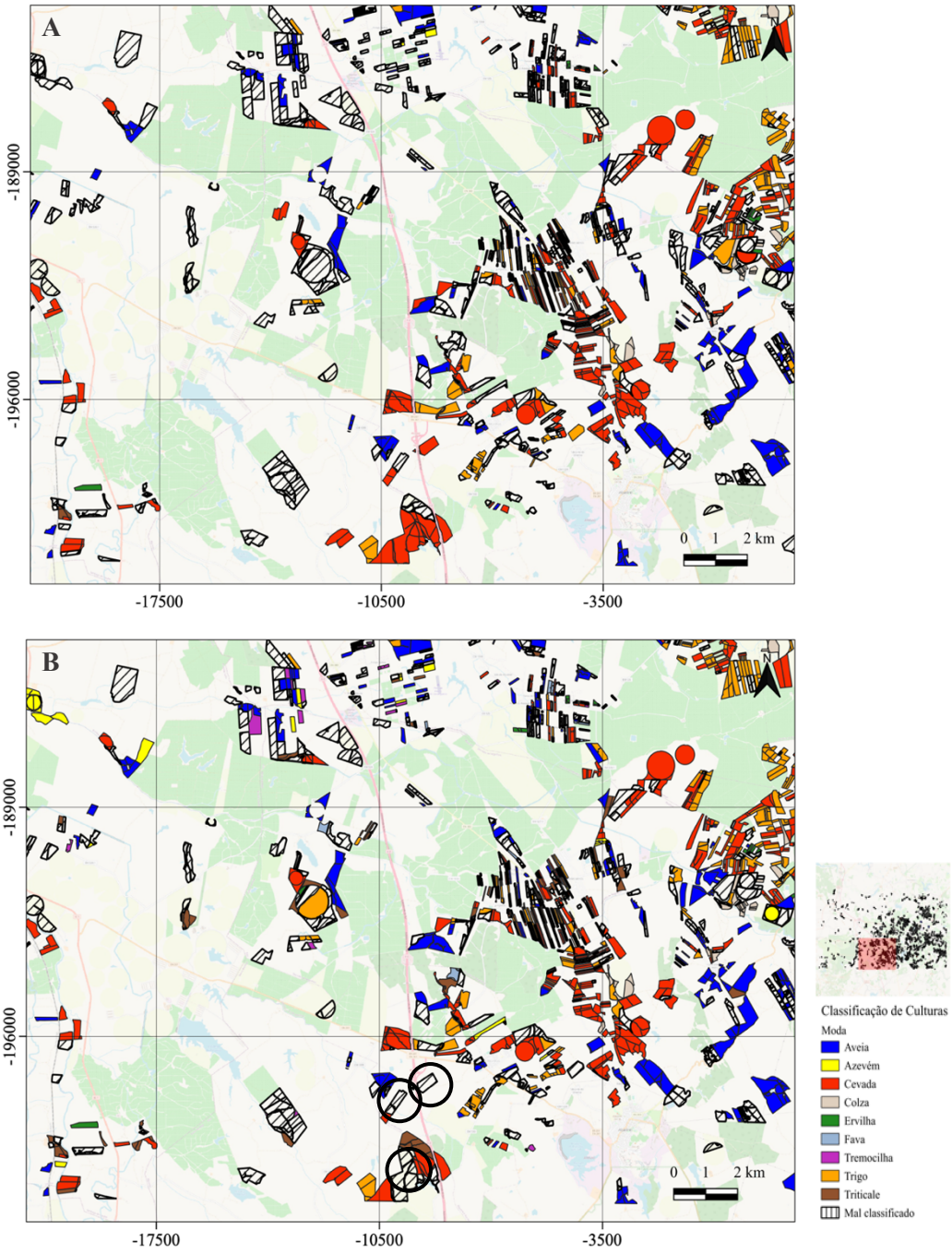

Figura 7.8 Visualização pormenorizada das parcelas mal classificadas a partir da figura anterior, na A4; *RF* (A) e *SVM* (B)