Masters Program in Geospatial Technologies

LANDCOVER AND CROP TYPE CLASSIFICATION with intra-annual times series of sentinel-2 and machine learning at Central Portugal

Itzá Alejandra Hernández Sequeira

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

LANDCOVER AND CROP TYPE CLASSIFICATION

with intra-annual times series of sentinel-2 and machine learning at Central Portugal

Dissertation supervised by:

Ph.D. Mário Sílvio Rochinha de Andrade Caetano

Associate Professor, Nova Information Management School (NOVA IMS)

University of Nova, Lisbon, Portugal

Ph.D. Filiberto Pla Bañón

Professor, Institute of New Imaging Technologies (INIT)

University of Jaume I (UJI)

Castellón, Spain

Ph.D. Hugo Alexandre Gomes da Costa

Invited Lecturer, Nova Information Management School (NOVA IMS)

University of Nova, Lisbon, Portugal

February 24, 2020

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisboa, February 24, 2020

Itzá Alejandra Hernández Sequeira

[the signed original has been archived by the NOVA IMS services]

ACKNOWLEDGMENTS

First, I would like to thank my supervisor Mário Caetano for allowing Geotech master students to develop research in the Direção-Geral do Território (DGT); for visualizing the project and encouraging me to take this challenge. Additionally, I would like to thank my co-supervisor Hugo Costa and (my un-official co-supervisor) Pedro Benevides from the Divisão de Informação Geográfica (DIG) for their helpful guidance and all the constant feedback provided to this work; and to my co-supervisor Filiberto Plá for his enthusiasm on the thesis topic and constructive suggestions provided. Special thanks to William Martínez, because his applied coding knowledge to geospatial information was essential to develop many of the scripts. I learned a lot from their insights in the remote sensing and land cover mapping field; working alongside them was a great professional experience.

Also, I express my gratitude towards the Erasmus Mundus Program for financing students from all over the world to pursue their academic development with the Master in Geospatial Technologies. The multi-cultural environment we are surrounded by makes us more tolerant and open towards others. Many thanks to Marco Painho, Christoph Brox, and Joaquín Huerta for coordinating this program through the three universities. Special thanks to Joel Silva from NOVA IMS for the codes he provided during the remote sensing course as they were a fundamental input to this work.

Additionally, I am grateful for my GeoTech family, which is an enthusiastic and talented group of people. I was encouraged by the quality of their work in every project they delivered. Especially during the Geomundus Conference that we organized, that is an excellent memory of outstanding teamwork. Special thanks to the wonderful people that supported me until the very end of the master Carlos, Francisco, Mihail, and Vicente; it was a pleasure sharing many Lisbon memories with you.

Finally, I cannot thank enough my family in Nicaragua; they are the real strength behind every project I achieve.

This thesis was developed under the framework of the project foRESTER (PCIF/SSI/0102/2017) funded by Fundação para a Ciência e Tecnologia

LANDCOVER AND CROP TYPE CLASSIFICATION

with intra-annual times series of sentinel-2 and machine learning at Central Portugal

ABSTRACT

Land cover and crop type mapping have benefited from a daily revisiting period of sensors such as MODIS, SPOT-VGT, NOAA-AVHRR that contains long time-series archive. However, they have low accuracy in an Area of Interest (ROI) due to their coarse spatial resolution (i.e., pixel size > 250m). The Copernicus Sentinel-2 mission from the European Spatial Agency (ESA) provides free data access for Sentinel 2-A(S2a) and B (S2b). This satellite constellation guarantees a high temporal (5-day revisit cycle) and high spatial resolution (10m), allowing frequent updates on land cover products through supervised classification. Nevertheless, this requires training samples that are traditionally collected manually via fieldwork or image interpretation. This thesis aims to implement an automatic workflow to classify land cover and crop types at 10m resolution in central Portugal using existing databases, intra-annual time series of S2a and S2b, and Random Forest, a supervised machine learning algorithm. The agricultural classes such as temporary and permanent crops as well as agricultural grasslands were extracted from the Portuguese Land Parcel Identification System (LPIS) of the Instituto de Financiamento da Agricultura e Pescas (IFAP); land cover classes like urban, forest and water were trained from the Carta de Ocupação do Solo (COS) that is the national Land Use and Land Cover (LULC) map of Portugal; and lastly, the burned areas are identified from the corresponding national map of the Instituto da Conservação da Natureza e das Florestas (ICNF). Also, a set of preprocessing steps were defined based on the implementation of ancillary data allowing to avoid the inclusion of mislabeled pixels to the classifier. Mislabeling of pixels can occur due to errors in digitalization, generalization, and differences in the Minimum Mapping Unit (MMU) between datasets. An inner buffer was applied to all datasets to reduce border overlap among classes; the mask from the ICNF was applied to remove burned areas, and NDVI rule based on Landsat 8 allowed to erase recent clear-cuts in the forest. Also, the Copernicus High-Resolution Layers (HRL) datasets from 2015 (latest available), namely Dominant Leaf Type (DLT) and Tree Cover Density (TCD) are used to distinguish

between forest with more than 60% coverage (coniferous and broadleaf) such as Holm Oak and Stone Pine and between 10 and 60% (coniferous) for instance Open Maritime Pine. Next, temporal gap-filled monthly composites were created for the agricultural period in Portugal, ranging from October 2017 till September 2018. The composites provided data free of missing values in opposition to single date acquisition images. Finally, a pixel-based approach classification was carried out in the "Tejo and Sado" region of Portugal using Random Forest (RF). The resulting map achieves a 76% overall accuracy for 31 classes (17 land cover and 14 crop types). The RF algorithm captured the most relevant features for the classification from the cloud-free composites, mainly during the spring and summer and in the bands on the Red Edge, NIR and SWIR. Overall, the classification was more successful on the irrigated temporary crops whereas the grasslands presented the most complexity to classify as they were confused with other rainfed crops and burned areas.

KEYWORDS

Sentinel-2
Land Cover Mapping
Crop type Mapping
Random Forest
Automatic Sample Extraction
Intra-annual time series
Portugal

ACRONYMS

CAP – Common Agricultural Policy

COS – Carta de Uso e Ocupação do Solo (Portuguese of LCLU)

DGT – Direção-Geral do Território - General Directorate for Territorial Management

FCT - Fundação para a Ciência e Tecnologia - Foundation for Science and Technology.

foRESTER – Rede de sensores combinada com modelação da propagação do fogo integrado num sistema de apoio à decisão para o combate a incêndios florestais - Data fusion of sensor networks and fire spread modelling for decision support in forest fire suppression.

EU – European Union

HRL – High-Resolution Layers

ICNF - Instituto da Conservação da Natureza e das Florestas - Institute for Nature Conservation and Forests

IFAP – Instituto de Financiamento da Agricultura e Pescas - Agriculture and Fisheries Financing Institute

INSPIRE– Infrastructure for spatial information in Europe

IPSTERS - IPSentinel Terrestrial Enhanced Recognition System

LCLU – Land Cover Land Use

LPIS - Land Parcel Identification System

NCPA – National Control and Paying Agency

OTS – On-The-Spot

SCAPE FIRE - A sustainable landSCAPE planning model for rural FIREs prevention

INDEX OF THE TEXT

INDEX OF TABLES

INDEX OF FIGURES

1 INTRODUCTION

1.1 Background

There is a need to quantify land cover and its changes over time in a precise and timely way for monitoring human and physical environments [1] as well as for providing information to support studies, research, and sustainable development policies [2]. The constant changes in land cover dynamics and the seasonality of crops demand a spatial and temporal continuity in the mapping of the areas of interest. Nowadays, it is possible to produce robust large-scale land cover mapping automatically using supervised classification, time series of high-resolution optical imagery, and existing databases for data training and validation [3]. The new paradigm in land cover production -Land Cover 2.0- takes advantage on the developments in computer hardware and software; increased spatial, spectral, and temporal resolutions of satellite imagery; open-access data and automated data processing using classification algorithms to generate timely, reproducible and accurate land cover maps [4]. Currently, it is possible to classify large geographic areas over multiple decades at an annual time step, as reported by Hermosilla et al. (2018) [5] that generated a 29-year data cube of land cover for the years 1984 to 2012. Moreover, automated systems as the Sen2-Agri can ingest and process multi-sensor imagery (Sentinel-2 and Landsat 8 time series) for operational agriculture monitoring systems [6].

1.2 Problem Statement and Motivation

The General Directorate for Territorial Management (DGT) in Portugal is the entity responsible for producing two land-use maps for mainland Portugal: the CORINE Land Cover (CLC) and the Carta de Uso e Ocupação do Solo (COS) that is the official Land Cover Land Use (LCLU) of the country. From one side, the CLC is a European project with a minimum mapping unit of 25 ha and 44 thematic classes with five years of reference (1990, 2000, 2006, 2012 and 2018) while the COS is a national product with a minimum mapping unit of 1 ha, 88 classes in 2018 and 6 years of reference (1990, 1995, 2007, 2010, 2015 and 2018) [7]. Mapping 88 classes at a spatial resolution of 1 ha require very high-resolution orthophoto maps and rely mainly on visual interpretation for its competition. Despite the significant improvement in the reduction of production time from 10 years in 2000 to 3 years in 2018, COS remains a product that takes time and human effort.

During the year 2017, Portugal registered an extreme wildfire season with a record of 500,000 ha burned and more than 100 human lives lost. These natural hazards, along with droughts and heatwaves, are intensifying in the Mediterranean basin due to climate change [8]. Wildfires represent a severe hazard that can have negative impacts on society and the environment; they can become a disaster when a significant number of people in vulnerability are exposed, consequentially human lives are lost and livelihoods damaged [9]. Characterizing and predicting fire spread and behavior is applied to determine higher risk areas and firefighting strategies to minimize damage [10]. In order to predict fire-spread and behavior, fire simulation models use gridded geospatial information as input data for fire simulation. This data can comprise elements such as topography (i.e., elevation, aspect, and slope), weather conditions, and fuel types (i.e., surface fuel type and canopy metrics) [11].

As part of a decision support system for firefighting, DGT aims to provide an annual Land Cover (LC) map for fire propagation models in 2020 with fewer thematic classes than the COS for central Portugal. This map is intended to be prepared before the fire season and will enable the updated characterization of the terrain, as well as areas that burned and vegetation cuts. It will be produced in raster format (10m pixel size) and based on supervised classification over the satellite time series of Sentinel-2. The realization of this annual LC map is part of three projects: the "IPSentinel Terrestrial Enhanced Recognition System" (IPSTERS) whose primary goal is the implementation of AI algorithms in the digestion of Big Data for remote sensing in order to derive LCLU maps [12]; the "Data fusion of sensor networks and fire spread modeling for decision support in forest fire suppression" (foRESTER) that intends to derive LC maps from satellite imagery and ground data for near-real-time (NRT) fire spread predictions (FSP) [13]; and the Sustainable landSCAPE planning model for rural FIREs prevention (SCAPE FIRE). Nonetheless, the production of these LC maps is dependent on the availability of sample data, and typically, training samples are acquired manually through visual interpretation or fieldwork. The challenge remains to train supervised algorithms without human intervention in sample labeling. Instead, to acquire training samples from pre-existing datasets filtered with auxiliary information to discard possible data mislabeling.

The automatic sample extraction from existing datasets for supervised classification is ongoing research at DGT, using Central Portugal as a study case. Lüdtke, D. (2018) [14] implemented the EUROSTAT's Land Use/Cover Area Statistical Survey (LUCAS) database as training data and Sentinel-2 time series for monthly and annual

classification. She concluded that the leading cause of the low Overall Accuracy (OA) achieved (58% for six classes) was the uncertainties of the LUCAS database. Yet, a compelling finding was that the simultaneous use of the bands for the period of analysis (November 2016 to October 2017) resulted in higher accuracy than using monthly data. Later, Blanco, W. (2019) [15] used training data from an old map of COS 2015 to classify imagery of 2017 and using Sentinel-2 seasonal composites following a Best Available Pixel (BAP). The overall accuracy achieved using 13 features resulted in and 73% for six 61% in nine classes for the baseline. He concluded that although COS is a valuable source for sample extraction, it was not possible to increase the OA after refinements on the training data. Still, the BAP composites provided a free-cloud efficient input for a seasonal LC mapping. At present, DGT is implementing some other approaches to extract consistently labeled training samples from outdated maps; a novel-approach tested is the implementation of unsupervised clustering methods based on the methodology of Paris, C. (2019) [16]. In addition, training samples have been obtained from the visual interpretation of orthophoto maps and on auxiliary data [17]. This approach was carried out for Continental Portugal includes using Landsat Time-Series to derive LCLU maps from 2010 to 2015 achieving accuracies of 87.5% for the 2010 map using 15 classes.

This thesis was developed under the framework of the project foRESTER, and it investigates the possible results of implementing current research at DGT for automatically deriving samples from ancillary data for supervised classification. The methodology corresponds to the protocol for Land Cover and Crop Mapping 2018 for Tejo and Sado.

1.3 Research Question

When performing supervised image classification, several algorithms can be applied, and different data sources can be utilized for training the classifier. Depending on the number of target classes, the overall accuracy of the map fluctuates. The more classes are added to the classification, the higher the probability of misclassifications and, therefore, the reduction in the ability of the classifier to map the classes accurately. To create an automatic map, the main task is to generate a stable workflow for classifying satellite imagery in a reproducible way. Developing the present research at DGT and the availability of time series of Sentinel-2 and the up to date ancillary datasets (COS 2018, IFAP parcels 2018 and ICNF burned areas 2018), three research questions are proposed:

1. How accurate is it to classify 31 classes of land cover and crop types at 10m resolution?
2. Which are the most important features/variables to consider when using intra-annual time series?
3. When performing automatic sample extraction, can a set of pre-processing rules allow us to extract spectral signatures of the classes suitable for image classification?

1.4 Aim

This investigation project aims to generate an automatic land cover and crop type map in raster format using in situ and up-to-date data, satellite imagery, and machine learning algorithms. The automatization method relies on the retrieval of the spectral signatures of the land cover and crop types from intra-annual time-series imagery of Sentinel-2 at the pixel level taking advantage of the availability from COS 2018 land cover dataset, and IFAP 2018 monitored agricultural parcels as well as ICNF 2018 burned areas. These areas will serve as training and testing input to the Random Forest classifier, allowing to implement the supervised machine learning method for land cover and crop type classification.

1.5 Objectives

- Contribute to an automatic supervised classification workflow to produce land-cover and crop type maps.
- Review the existing state-of-the-art in machine learning and multi-temporal optical imagery for classifying land cover and crop type areas.
- Classify land cover and agricultural areas using the Random Forest algorithm and the features extracted from the training datasets and Sentinel-2 time series.

1.6 Thesis structure

- **The literature review** presents the core concepts for the development of the research focusing on the state-of-the-art in land cover and crop type mapping, use of sentinel-2 intra-annual time series, and random forest classifier.
- **The methodology** comprehensively describes the study area that is the focus of this research, the preprocessing of the primary datasets, the practical steps to the sample selection and spectral signature extraction, the selection of the best parameters for the random forest model and the challenges of training a machine-learning algorithm to classify large study areas and generate a final map.
- **Results and discussion** describe the results of the classification and contextualize the goodness-of-fit based on accuracy metrics. Critically analyze the results and relate them to literature.
- **Conclusions, limitations, and recommendations**: summary of the research, present the main findings and contribution as well as the limitations and recommendations for future steps in the automatic annual classification of land cover and crop type mapping.

2 LITERATURE REVIEW

This chapter focus on four sections that will allow contextualizing the framework for this study. Section 2.1 is dedicated to the primary considerations for land cover mapping, whereas section 2.2 presents the current agricultural monitoring systems based on remotely sensed data and how these efforts benefit from existing datasets such as the Land Parcel Information System (LPIS). Then, the concept of time series of satellite imagery is introduced in section 2.3, covering the differences between intra-annual and inter-annual time series and the need to produce gap-filled composites before using the data. Finally, section 2.4 comprises the principles of Machine Learning, also called Statistical Learning, and how these algorithms are different from the most commonly employed in remote sensing (i.e., Maximum Likelihood). It also describes the traditional machine learning workflow and its key constituents, the differences between supervised and unsupervised learning as well as regression, classification and clustering. At last, it introduces the algorithm that will be implemented through the thesis that is Random Forest and put it into the context of remote sensing and image classification.

2.1 Land cover mapping

Land cover analyses have evolved from studying a small geographic region at a determined period to global studies using smaller spatial resolution and higher temporal periods [18]. It remains, however, an intricate process, and in supervised land cover approaches, the critical component is the availability of training data (ground truth or reference data) for the signature generation [5]. According to the meta-analysis on supervised pixel-based land-cover image classification [19] that compared 266 articles between 1998 and 2012 the most relevant features considered when performing a classification process were texture, ancillary data (e.g., topographic, active sensors such as radar or LiDAR and passive sensors), multi-time imagery (e.g., fusion of images for the same area captured at different times), multi-angle imagery, image pre-processing (e.g., radiometric correction, atmospheric correction, pan sharpening, and geometric corrections), spectral indices (e.g., NDVI -arithmetic combinations of different spectral bands) and feature extraction (e.g., dimensionality reduction).

However, training data collection is delicate in large jurisdictions and over remote areas [3]. Ongoing research on the extraction of automatic training data includes the majority rule approach in polygon level source maps [16].

2.2 Crop type mapping

When using Earth Observation data for monitoring agriculture, there are recognized frameworks such as the GEOGLAM initiative. Currently, several main global and regional scale agricultural monitoring systems are in place, some of them are the Global Information And Early Warning System (GIEWS), the Famine Early Warning Systems Network (FEWS NET) and the Crop Watch for China [20]. The ESA "Sentinel 2 for Agriculture" (Sen2Agri) that started in 2015 aims to create operational crop types maps and dynamic cropland masks [21] that are required as input for global agricultural monitoring systems. Differences have been established between cropland maps, crop calendars, cropping intensity, crop type, growing calendar, crop condition indicators, and crop yield [20]. In crop type classification, the classifiers yielding the best performances are Random Forest, followed by the gradient boosted trees and then SVM [22]. RF has also been implemented in binary operations (cropland/non-cropland) systems [23]. Nevertheless, the key to differentiating individual crop types is the availability of temporal information [24]. For the calibration and correlation of the spectral signal to the various crops, information at a parcel-level is also a crucial element [24].

Several studies, including the Sen2-Agri system, have reported the use of the Land Parcel Identification System (LPIS) to extract samples from agricultural parcels [6], [17], [24]. The LPIS is an IT system based on aerial photographs of agricultural parcels employed to check payments made under the Common Agricultural Policy (CAP) of approximately 45.5 billion euro in 2015 [25]. In Portugal, the Instituto de Financiamento da Agricultura e Pescas (IFAP) ensures the financing, implementation, and control mechanisms of the measures defined at the national level in agriculture and fisheries. It acts as National Control and Paying Agency (NCPA) designated by the European Union (EU) under the Common Agricultural Policy (CAP) and is responsible for the administration and control of the subsidies in this sector. For applying to financial support, farmers are required to submit an application to the NCPA and declare the precise location and area of the agricultural parcels and the crop type [26]. For this, landowners use an online Geographic Information System (GIS) to digitize their parcels on orthophotos or very high-resolution satellite imagery [24]. The NCPA controls at least 5% of the declarations by performing an On-The-Spot (OTS) check, penalizing the farmers that submitted incorrect information [26].

2.3 Sentinel-2 time series

2.3.1 Time series imagery

The coarse (i.e., pixel size > 250 m) to medium resolution optical instruments on board of SPOT-Vegetation, MODIS, and PROBA-V have a daily revisit cycle, global coverage, and long-term archive [6]. This data can be exploited in long time-series research at regional or global scales, but often suffer from low local accuracy in land cover products [27] and high mixtures of crop types [28]. The revisiting period of 16 days for the Landsat 8 satellite of the U.S. Geological Survey (USGS) allows us to describe spatial details of land cover but cannot capture changes in crop phenology and growth due to low temporal repeat cycles and frequent cloud contamination [28]. Sentinel-2A (S2a) satellite of the European Spatial Agency (ESA) provides a revisit time of 10 days, and the Sentinel-2B (S2B) ensures a 5-days revisit time allowing the collection of high-quality spatial and temporal data [6]. More generally, time series algorithms have emerged over the last decade that can exploit dense, multi-sensor time series to derive improved land cover classifications [29]. Recent country-scale studies have demonstrated the added value of multi-sensor time series from Landsat and Sentinel-2 to differentiate crop types and grasslands [24].

2.3.2 Intra-annual and Inter-annual

In the review of time series analysis for Land Cover mapping [27] Gómez et al. (2016) pointed out the temporal relevance for images collected over intervals in the same year (intra-annual) or over some years (inter-annual). The intra-annual imagery allows monitoring the subtle differences and variations over the growing period by calculating an averaged phenology while the inter-annual imagery allows to compute a unique spectral profile that makes more visible when abrupt changes occur in the land cover. Mapping landcover is complex, time-series spectral data (intra-annual for phenology and inter-annual for land cover dynamics) provide more information for increasing the classification. For a specific class of interest (e.g., crop type mapping), it is necessary to incorporate the knowledge of other underlying processes (e.g., phenology, disturbance, succession) [27].

2.3.3 Gap-filled image composites

The availability of operational imaging satellites that covers all lands frequently, such as Landsat 8 and Sentinel-2, providing free and open access to these data has prompted new applications based on time series of images covering vast territories [3], [5], [24].

However, when processing optical satellite images on land surfaces, the detection of cloud and cloud shadows is one of the first issues. Clouds can frequently be mistaken with bright landscapes, semi-transparent clouds observed reflectance contains a mixture of cloud and land signals, and cloud shadows can be confused with water pixels, burnt areas or topographic shadows [30]. As for now, cloud and cloud shadow masking algorithms for Landsat 8 and Sentinel 2 include the MAJA algorithm by the French Space Agency (CNES), Sen2Cor, from the European Space Agency (ESA) and FMask of the United States Geological Survey (USGS).

When integrating temporal time series of imagery, most approaches follow a best-pixel selection strategy that allows exploiting all the imagery available [24]. The Best Available Pixel (BAP) enables the computation of periodic image composites free of haze, clouds, or shadows over large areas [31]. White, J. C. et al. (2014) [32] proposed three unique types of pixel-based image composites: annual (single-year) composites, multi-year composites, and proxy-value composites. Wherein, Defourny, P. et al. (2019) [6] generates monthly composites using a weighted average algorithm, that averages cloud-free surface reflectance values over the given period. Interpolation over surface reflectance values to fill missing values due to the presence of clouds and clouds shadows has also been implemented for operational systems [3].

2.4 Machine Learning – Statistical Learning

With the rapid growth of "Big data," machine learning, also referred to as statistical learning, a broad set of tools for analyzing and understanding the data emerged. Several models can be built and require a set of input data to predict or estimate output data [33]. An advantage of Machine-learning algorithms is that they do not make assumptions about the data distribution (i.e., non-parametric), can handle data of high dimensionality, and can efficiently classify remotely sensed imagery [34].

2.4.1 Machine Learning Process

A traditional machine learning workflow (Figure 1) requires key constituents: data collection, feature engineering (cleaning and feature selection), model learning (training, validating and testing), and model evaluation [35].

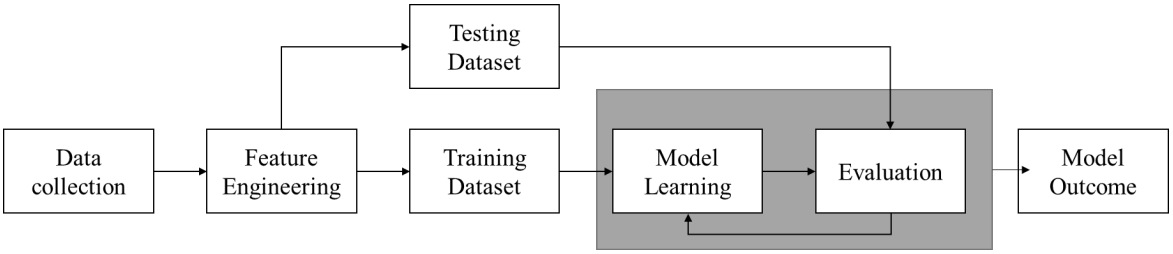


Figure 1 Machine Learning Workflow

The collected raw data may be noisy, incomplete, or inconsistent, and before using the data as input in the model, it is required to pre-process it by removing errors and outliers and fill missing values. Other tasks include the integration of multiple datasets and transform the data into an appropriate format, so it is readable depending on the tool deployed to perform the machine learning process. Feature selection and extraction are utilized to reduce dimensionality in voluminous data, allowing to remove irrelevant or redundant features that promote over-fitting and to reduce computational requirements. Some techniques for dimension reduction include entropy, Fourier transform, and Principal Component Analysis (PCA) [35].

The training dataset is implemented to teach the model how to estimate the function that will be able to predict output for any new observation [33]. A validation dataset is applied to choose a suitable architecture for the model. If the architecture is pre-selected, there is no need for a validation set. Finally, the testing dataset allows the model to iterate and tune the different parameters until the model is ready to be deployed. The main decompositions of the dataset are 60/20/20% if training, validation, and test datasets or 70/30% if validation is not required [35]. The training data selection is relevant, because large, and accurate training datasets result in increased classification accuracy. It is being suggested that the minimum number of training samples should be ten or preferably 100 times the number of variables [34].

The evaluation focusses on the predictive efficacy of the model and on the computational requirements (training and testing time) for its application [36]. A high bias refers to a simple ML model that poorly maps the relations between features and outcomes (under-fitting) while a high variance implies an ML model that fits the training data but does not generalize well to predict new data (over-fitting) [35]. Techniques for experimental algorithm evaluation include bootstrap sampling, cross-validation, and holdout evaluation [36].

2.4.2 Supervised and Unsupervised Learning

Supervised learning uses labeled training datasets to create models, and typically, this approach is used to solve classification and regression problems [35]. Therefore, the algorithm uses patterns to predict the values of the labeled data on additional unlabeled data, and by comparing the actual output with the correct output, it finds the errors, learns, and modifies the model accordingly. Unsupervised learning uses unlabeled training datasets to create models that can discriminate between patterns in the data. This approach is most suited for clustering problems [35]. The algorithm explores the data and finds patterns for grouping together values based on their features.

2.4.3 Regression, Classification, and Clustering

In data clustering, the aim is to partition objects into groups such that similar objects are grouped while dissimilar objects are grouped separately. Categorical clustering views the data as a set of a two-dimensional matrix of data objects and attributes (a set of discrete values that are not comparable) and attempts to partition the set of objects into groups with similar attributes [36]. Well-known clustering algorithms are K-means and Kohonen Self-Organizing Maps (SOM). Whereas in classification and regression problems, the goal is to map a set of new input data to a set of discrete or continuous-valued outputs [35]. Some classification algorithms include Decision Trees (DT), Neural Networks (NN), K-Nearest Neighbors (k-NN), Bayesian Networks (BN), and Support Vector Machines (SVM). There are also ways of combining them into ensemble classifiers such as boosting, bagging, and the ensemble DT - Random Forest (RF). While known regression algorithms are mainly linear models such as Least Squares that include specific techniques such as OLS, MaxEnt, Logistic Regression [35], [36], LASSO Regression, SVM and Multivariate Regression algorithm.

2.4.4 Random Forest (RF)

Random Forest algorithm specifications for classification

The RF classifier is an ensemble classifier that uses multiple Classification and Regression Trees (CART) and combines their outputs to make a prediction, treating them as a "committee" of decision-makers [36], [37]. It combines the Bagging algorithm to reduce variance by the random selection of samples and the Random Subspace method to reduce bias by the random selection of the features employed at each split [36]. When operated for classification, each tree "votes" for a class and then classify using the "majority vote" of the forest [38]. Findings for Random Forest is that it does not overfit as more

trees are added, it is relatively robust to outliers and noise, gives useful internal estimates of error, strength, correlation and variable importance and is easily parallelized [37].

1. For b = 1 to B *(n° of trees in the forest)*:
    a. Draw a bootstrap sample **Z*** of the size N from the training data.
    b. Grow a random-forest tree Tb to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
        i. Select m variables at random from the p variables.
        ii. Pick the best variable/split-point among the m.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{Tb\}_1^B$.

To make a prediction at a new point x:

Classification: Let $\hat{C}b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}b_{rf}^B(x)$ = *majority vote* $\{\hat{C}b(x)\}_1^B$.

Table 1 Algorithm: Random Forest for Classification [37], [38]

The first step in the RF algorithm in Table 1 (a) is to extract a "bootstrap sample" from the training dataset; bootstrapping allows to select the same sample more than once and include it in the subset dataset while other samples may not be selected at all. Bagging (acronym derived from Bootstrap AGGregatING) allows that each member of the ensemble is constructed from a different training set, each dataset being a bootstrap sample from the original [36]. About two-thirds of the samples (in-bag samples) are used to train the trees with the remaining one third (out-of-the-bag) are employed in an internal cross-validation technique for performance estimation of the model [37], [39].

Then, each tree is grown using samples from the bootstrapped dataset (b); however, it will select a random variable m from the full set of variables p available and pick the best one for the top split [38]. The random subspace principle is to increase diversity between members of the ensemble by restricting classifiers to work on different random subsets of the full feature space [36]. This procedure is repeated for the number of trees in the forest; bagging seems to enhance accuracy when random samples are utilized, this is also the case when using a single randomly chosen input variable to split on at each node [37].

Random forest classification of remote sensing datasets

ML algorithms have user-defined parameters that may improve classification accuracy when running parameter optimization. One of the benefits of the RF algorithm is that it is considered easy to optimize in comparison to more complex models such as ANN. Their stability concerning the choice of parameters makes them excellent candidates for operational processing chains, yielding classification accuracies as high as more sophisticated algorithms such as SVM but with much lower computational complexity [3].

RF has been exploited in time series analysis for creating multi-temporal cloud mask for Sentinel-2 imagery [30]; in supervised classification for producing land cover maps at a country scale for France [3], as well as crop type and land cover maps for Germany [24] and in global operational systems such as Sen2Agri for crop type maps [6].

RF algorithm requires only two user-defined parameters: the number of Decision Trees in the ensemble and the number of random variables at each node [34]. RF is computationally efficient and does not overfit [39]; the number of trees does not impact accuracy as long as it is large enough, being 500 a very conservative value [37]. The estimated error rate can be plotted for each ensemble size to determine when the performance stabilizes [34], [37].

Another advantage of RF is that the algorithm itself generates additional information [37]. The out-of-bag (OOB) error provides an unbiased estimate of generalization error and resembles the error estimate obtained by N-fold cross-validation [38]. Also, the Variable Importance (VI) estimation ranks the variables based on the predictive capabilities for discriminating between the target classes [37], [39]. The VI has been exploited in remote sensing to reduce the number of dimensions of hyperspectral data (i.e., the contribution of bands), to identify relevant ancillary data (i.e., topography) and to select the suitable season to classify target classes [34], [39]. This allows addressing the challenges of mitigating the Hughes phenomenon (i.e., the curse of dimensionality) that occurs when the number of variables is much larger than the number of training samples [39].

The main drawback is that it is sensitive to sampling design when imbalanced data is used; the final classification will under-predict the minority class [34]. To reduce misclassification, this sensitivity to sampling design needs to be considered by ensuring that training and testing are independent, establish balance and representativity of each class, and have an extensive training sample to deal with the number of data dimensions [39].

3 METHODOLOGY

The methodology section defines the study area (3.1), enumerates the datasets that were employed (3.2), and describes the methods implemented (3.3).

3.1 Study area

The Region of Interest (ROI) corresponds to the strata 214 in level 3 of the stratification of Continental Portugal (Figure 2). According to the classification, this area is about 1,223,890 ha in the low interior lands of the south of Portugal, which covers most of the valleys of the rivers Tejo and Sado and contains a great diversity of land uses as well as multiple crop types [40].

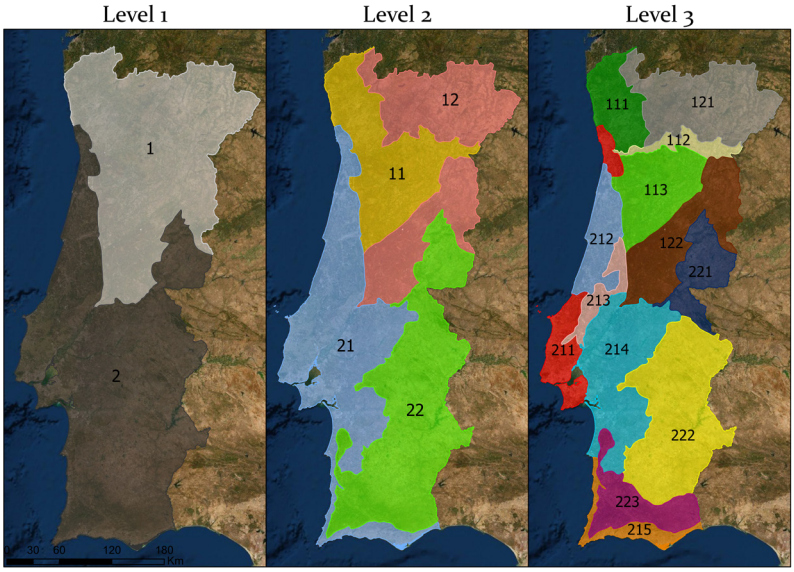


Figure 2 Stratification of Continental Portugal [40]. Scale 1:3,000,000.

This stratification considers the spectral diversity of the surface features, respective landcover, and geographic space [40]. When producing maps over large areas, stratification-based classification is recommended to avoid intra-class variability and has proven to yield better results for all the classification metrics than a tile-based approach [3].

3.2 Data

The research is based on the use of the ancillary data described in section 3.2.1, which allows the extraction of labeled points randomly; these serve as training data for producing a land cover and crop type map in raster format at 10m spatial resolution for 2018. The class nomenclature of the map is composed of 31 classes as can be appreciated in Table 2; the land cover classes like urban, forest and water are derived from the official LULC map of Continental Portugal (COS 2018) while the agriculture comprises annual and permanent crop as well as agricultural pastures obtained from the Land Parcel Identification System (LPIS) of the Instituto de Financiamento da Agricultura e Pescas (IFAP 2018), and finally, the burned areas are identified from the maps of the Institute for Nature Conservation and Forests (ICNF) from 2018. To avoid the inclusion of mislabeled pixels that can occur due to differences in the Minimum Mapping Unit (MMU) between datasets, the classes extracted were filtered with auxiliary information. These filters include the removal of burned areas (ICNF 2015-2018) and alerts based on the decrease of the Normalized Difference Vegetation Index (NDVI) between two dates acquired from Landsat 8 imagery (2015-2018); allowing to remove potential clear cuts [17]. Likewise, the Copernicus High-Resolution Layers (HRL) datasets from 2015 (latest available), particularly Dominant Leaf Type (DLT) and Tree Cover Density (TCD), is used distinguish between forest types, forest density and to eliminate non-forest pixels. All the previous datasets mentioned were provided by DGT, including the simplified COS nomenclature (COSsim) in Table 2 that is under constant improvements for the foRESTER project. The editable version of the Land Cover and Crop Type nomenclature and corresponding RGB color scheme to symbolize the different classes can be found in annexes 7.2 and 7.3, respectively.

Also, a crucial input in this research is the intra-annual time series of Sentinel-2a and b from ESA for the period of October 2017 to September 2018 3.2.2. It is fundamental to mention that all the procedures in section 3.3.3, namely acquisition and pre-processing, indices calculation, production of monthly composites and filling of missing values was done by Hugo Costa and Pedro Benevides under the IPSTER project at DGT, using the R software and the computer specifications provided in section 3.3.1.

3.2.1 Ancillary data

The ancillary data used as a reference for training and validating, as well as the filters applied to preprocess the reference datasets, is described in this section.

COS 2018

The official LULC of Continental Portugal (COS) is a vectorial map with an MMU of 1 hectare, a minimum distance between lines of 20m and produced through visual interpretation of orthophoto maps (25 cm pixel size) and auxiliary data. Each polygon contains only one LCLU code selected from the most detailed hierarchical level of the nomenclature, and this class must occupy equal or more than 75% of the entire delimited area. The COS 2018 contains a total of 83 classes in the fourth level of detail (LV4) that can be aggregated to a first level (LV1) containing 9 LCLU mega classes such as artificialized territories, agriculture, pastures, agroforestry surfaces, forests, open spaces or with sparse vegetation, wetlands, and surface water masses. As an example, in section 7.3 the Cork Oak forest in LV3 corresponds to a Broadleaf forest in LV2 and to the mega class Forests in LV1. The technical specifications are available in the official documentation of COS 2018 for Continental Portugal [7].

The version under current development COS2018v1 is the one being used; therefore, some nomenclatures might change during the writing of this document. A total of 16 LC classes were derived from the 83 classes available in LV4 of the COS nomenclature; all of them will be used for training and testing as it can be appreciated in Table 2, the classes from COS 2442 (Holm oak agroforestry system) and class 3112 (pure forest of Holm oak) were combined into the class 5121 (Holm oak forest) as both represent the same LC but have different uses. The class 6111 corresponding to shrubland corresponds to areas that remained shrubland through the COS series, meaning that shrubland was present in 1990, 1995, 2007, 2010, 2015 and 2018.

IFAP 2018

The Portuguese Land Parcel Identification System (LPIS) of the Instituto de Financiamento da Agricultura e Pescas (IFAP) is comprised of two independent datasets. The "national parcel registry" that will be used for training and the "controlled parcels" for testing. The first consists of the parcels reported by the farmers that applied for agricultural subsidies in the frame of CAP. The second dataset is the controlled parcels containing polygons with rectified edges through the visual interpretation of orthophotos

and field verification to assess the crop type planted. Overall, the LPIS is a very reliable product used in countrywide crop mapping studies; however, it can contain errors such as false claims or digitization errors [24].

The IFAP 2018 is composed of more than 175 types of crops, mapping such a number of classes at 10 m resolution can be challenging. Instead, the ten most abundant temporary crops (5 rainfed and 5 irrigated), three permanent crops, and the agricultural grasslands were selected for the analysis for a total of 14 crop types classes.

The IFAP also provides a crop calendar (in attachment 7.3) that illustrates the growing period for the crops monitored in Portugal. The early stages of the crops correspond to the flooding-only for rice-, seed, and crop development (germination and tillering) where the area is not covered yet by the vegetation. Then, the peak of greenness occurs during the flowering, fruit, and ripening. Finally, during the harvest, depending on the farmer's practices, the soil can remain clear, with stubble or left to natural regeneration.

Burned areas (ICNF 2015-2018)

The Institute for Nature Conservation and Forests (ICNF) is responsible for the realization of an annual map of burned areas for Portugal based on visual interpretation of Landsat TM/ETM. The institute publishes on its website at the end of each fire season a vectorial dataset containing burned areas larger than five hectares [17], [41]. The polygons used contains the information for the areas that burned by wildfires during the years 2015 to 2018.

After a wildfire, it is likely that the LC type changes, as forest and shrubs, would not be present anymore in scorched areas. The ICNF mask is implemented as the first filter for the COS dataset to avoid extracting samples of vegetation from burned areas. This allowed erasing the pixels that correspond to grasslands, forests, or shrubland classes in COS 2018 but fall inside the burned areas. Whereas the year 2018 was used to extract training and testing points for this class, corresponding to the last class in the nomenclature.

NDVI mask

Furthermore, it is not possible to sample broadleaf or coniferous forests from areas where trees have been uniformly cut down (i.e., clear cuts). According to Costa et al. (2018) [17], land cover changes can potentially be detected by monitoring if the inter-annual values of NDVI decrease between two successive years over a certain threshold. Therefore, the NDVI mask (derived from Landsat 8) can help to identify clear cuts in a forest, allowing to exclude these areas from the training samples for broadleaf or coniferous forests in COS 2018.

High-Resolution Layers (HRL 2015)

Delivered at a Pan-European level, the HRL is a product available in the Land Monitoring Service of Copernicus. These layers are complementary to the production of CLC, and it is available for continental Portugal [42]. The HRL for the thematic class forest of 2015 was used in the preprocessing of forests and shrublands of COS 2018. Two of the forest products are used, the Tree Cover Density (TCD) representing the percentage that a pixel is covered by trees and Dominant Leaf Type (DLT) that allows distinguishing between broadleaf or coniferous majority.

CODE	CLASS	Description	Reclassification
1.1.1.1	Build up	this class includes all the artificial or landscaped surfaces intended for activities related to human societies such as urban fabric, road network, and associated spaces.	COS 2018 (1111, 1112, 1221)
2.1.1.1	Wheat	agricultural class corresponding to a temporary rainfed crop, this cereal grows during the autumn and winter.	IFAP 2018 (001)
2.1.1.2	Barley	rainfed temporary cereal.	IFAP 2018 (004)
2.1.1.3	Oatmeal	rainfed temporary cereal.	IFAP 2018 (005)
2.1.1.4	Ryegrass	rainfed temporary cereal (forage).	IFAP 2018 (067)
2.1.1.5	Lupin	rainfed temporary pulses (nitrogen fixer).	IFAP 2018 (240)
2.1.2.1	Maize	agricultural class corresponding to an irrigated temporary crop, this cereal grows during the spring and summer.	IFAP 2018 (006)
2.1.2.2	Sorghum	irrigated temporary cereal.	IFAP 2018 (008)
2.1.2.3	Rice	irrigated temporary cereal.	IFAP 2018 (024)
2.1.2.4	Tomato	irrigated temporary vegetable.	IFAP 2018 (033)
2.1.2.5	Potato	irrigated temporary vegetable.	IFAP 2018 (103)
2.2.1.1	Vineyards	areas where vineyards are dominant over other types of permanent crops such as orchards or olive trees.	IFAP 2018 (034
2.2.2.1	Orchards	cultivated plots with trees intended for fruit production, this class combines 17 types of trees from figs and oranges to walnuts and hazelnuts.	IFAP 2018 (085, 093, 094, 096, 097, 105, 107, 108, 109, 112, 116, 118, 119, 157, 208, 209, 211)
2.2.3.1	Olive Trees	areas with olive tree plantations (*Olea europea var. europea*) for olive production.	IFAP 2018 (083);
3.1.1.1	Agricultural grassland	areas permanently occupied with cultivated herbaceous vegetation.	IFAP 2018 (143);
3.1.2.1	Natural grassland	areas with 25% or more of the surface occupied by herbaceous vegetations growing without fertilization, cultivation, sowing, or drainage.	COS 2018 (321);
5.1.1.1	Cork oak forest	Agroforestry Systems or pure forest of Cork oak (*Quercus suber*).	COS 2018 (2441, 3111);
5.1.2.1	Holm oak forest	Agroforestry Systems or pure forest of Holm oak (*Quercus rotundifolia*).	COS 2018 (2442, 3112);
5.1.3.1	Eucalyptus forest	Broadleaf forest where the angiosperm trees represent 75% or more of the forest cover.	COS 2018 (3115);

CODE	CLASS	Description	Reclassification
5.1.4.1	Other broadleaf forest	Agroforestry Systems or pure forests of oak species other than cork oak and holm oak. These include chestnut trees (*Castanea sativa*), walnut trees (*Juglans regia*), and forests of invasive species.	COS 2018 (2443, 3113, 3114, 3116, 3117);
5.2.1.1	Closed Maritime pine forest	Coniferous forest where the gymnosperm species represent 75% or more of the forest cover.	COS 2018 (3121);
5.2.1.2	Open maritime pine forest	this class is derived from class 5.2.1.1 after the crossing with the High-Resolution Layers (HRL) process described in section 3.2.1.	COS 2018 (3121);
5.2.2.1	Stone pine forest	Agroforestry Systems or pure forest of Pine (*Pinus pinea*).	COS 2018 (2444, 3122);
5.2.3.1	Other coniferous forest	pure forests of other coniferous species not included in the previous classes. (e.g., *Pinus sylvestris*, *Larix spp.*, *Cryptomeria japonica*.).	COS 2018 (3123);
6.1.1.1	Shrubland	natural areas of spontaneous vegetation, little or very dense where shrub cover is 25% or more.	areas that remained shrubland from COS 1990 to COS 2015;
7.1.1.1	Baresoil	areas of open-air mineral extraction, sand exploitation areas, banks of rivers, and coastal sands, including ante-dune vegetal formations.	COS 2018 (1311, 1312, 3311, 3312);
7.1.2.1	Bare Rock	areas where the surface covered by rock is higher than 90%, also included areas of abandoned mineral extraction.	COS 2018 (332);
7.1.3.1	Sparse vegetation	areas where the herbaceous vegetation is between 10% and 25% only.	COS 2018 (333);
8.1.1.1	Wetlands	lowlands flooded in winter, less saturated with water all year round or shore areas submerged during high tide at some point in the cycle of the annual sea.	COS 2018 (411, 421);
9.1.1.1	Water	natural and artificial freshwater surfaces, oceans and surfaces, and coastal lagoons and river mouths.	COS 2018 (5111, 5121, 5122, 5123, 5124, 5125, 521, 522);
9999.	Burned Areas	areas that burned in 2018 and detected by the ICNF.	ICNF (2018)

Table 2 Nomenclature for Land Cover and Crop type

3.2.2 Remote Sensing Data

Sentinel 2

The Copernicus Sentinel-2 mission is a constellation of two polar-orbiting satellites that operate simultaneously, phased at 180° to each other at a mean altitude of 768 km. This allows a and high revisit time (10 days for S2-A and 5 days for S2A/B), and its wide swath width (290 km) provides a high coverage [43] being ideal for the proposed study. The imagery is acquired by the Multispectral Instrument (MSI) on-board Sentinel-2 and contains 13 spectral bands from Visible/Near Infrared (VNIR) to Short Wave Infrared (SWIR) and comes in three spatial resolutions (10, 20 and 60m) as seen in Table 3.

Band	Spatial resolution (m)	Central wavelength(nm)	Bandwidth (nm)	Purpose
B01	60	443	20	Aerosol detection
B02	10	490	65	Blue
B03	10	560	35	Green
B04	10	665	30	Red
B05	20	705	15	Red Edge
B06	20	740	15	Red Edge
B07	20	783	20	Red Edge
B08	10	842	115	Near Infrared (NIR)
B08A	20	865	20	NIR
B09	60	945	20	Water vapor
B10	60	1375	30	Cirrus
B11	20	1610	90	Snow/ice/cloud discrimination (SWIR)
B12	20	2190	180	Snow/ice/cloud discrimination (SWIR)

Table 3 Specifications of the Sentinel-2 bands

The Sentinel-2 data was obtained from the French Theia Land Data Centre (THEIA). The data is in the Coordinate Reference System (CRS) of Universal Transverse Mercator (UTM) Zone 29N, and it is tiled in the Military Grid Reference System (MGRS) allowing all the images to have the same size (100x100 km2) and a code (e.g., a tile in Portugal is T29SND).

The biogeographical region 214 "Tejo and Sado" is covered by four tiles of Sentinel-2 that correspond to 29SNB, 29SNC, 29SND, and 29SPD. The classification will be done for the strata 214 that is within the tiles 29SNC and 29SND. Most of the study area is comprised within the same orbit; however, a slight corner in the 29SND tile has swath overlap with the adjacent orbit, accounting for more imagery collected within the same period than tile 29SNC. Although the re-visitation period is the same, the dates in the collection of the orbits vary from left to right in adjacent land.

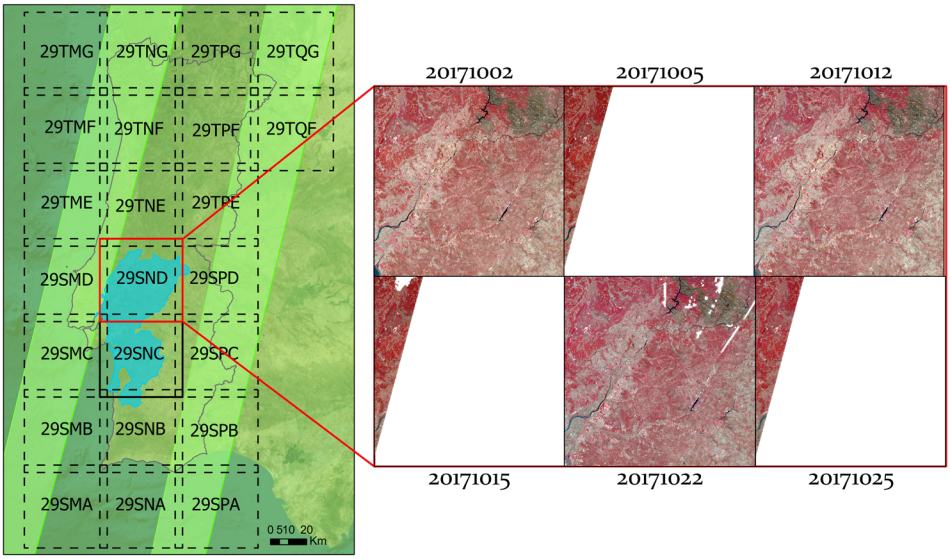


Figure 3 Sentinel-2 orbit, swath, and tilling for the study area and images acquired for October 2017 in the tile 29SND.

Orthophotos for Continental Portugal

For all the visualizations, the orthophotos available as Web Map Service (WMS) from DGT were used. The imagery has a spatial resolution of 25 cm, and it is available for Continental Portugal.

3.3 Methods

The proposed methodology corresponds to an automatic supervised classification procedure using the random forest classifier, intra-annual time-series of Sentinel-2, and filtered auxiliary data to extract the labels for land cover and crop types automatically (Figure 4). First, the reference datasets (COS2018 and IFAP 2018) are reclassified using the nomenclature from, then a set of preprocessing rules is applied to the datasets to remove the pixels that do not match the class label (3.3.2).

Next, section 3.3.3 illustrates the preprocessing of the Sentinel-2 intra-annual time series. Initially, the imagery is downloaded for the period of October 2017 to September 2018; a mask to remove clouds and cloud shadows are applied, and all the bands are resampled to 10m. Later, five spectral indices are calculated, and all the imagery is aggregated to monthly composites. The potential missing values (pixels with no data during a month) are filled using linear interpolation in time to ensure continuity of information during the period.

Afterward, the supervised learning procedure is presented in section 3.3.4. This section starts with the automatic extraction of samples by class from the pre-processed datasets. Two independent sample datasets are acquired, one for training and one for testing with varying percentages 80/20 or 75/25 depending on the number of pixels available. For each sample, the spectral signatures are retrieved at the pixel level from all the bands of the composites and the spectral indices. Then, a grid search is used to determine the best hyperparameters for the RF; the models are fit to the training dataset and assessed with 10-split cross-validation.

At last, the performance of the best model is quantified following the metrics in (3.3.5) based on the predicted labels in the testing dataset. Then, the model is applied to unlabeled data allowing to generate the final map for the biogeographical region.

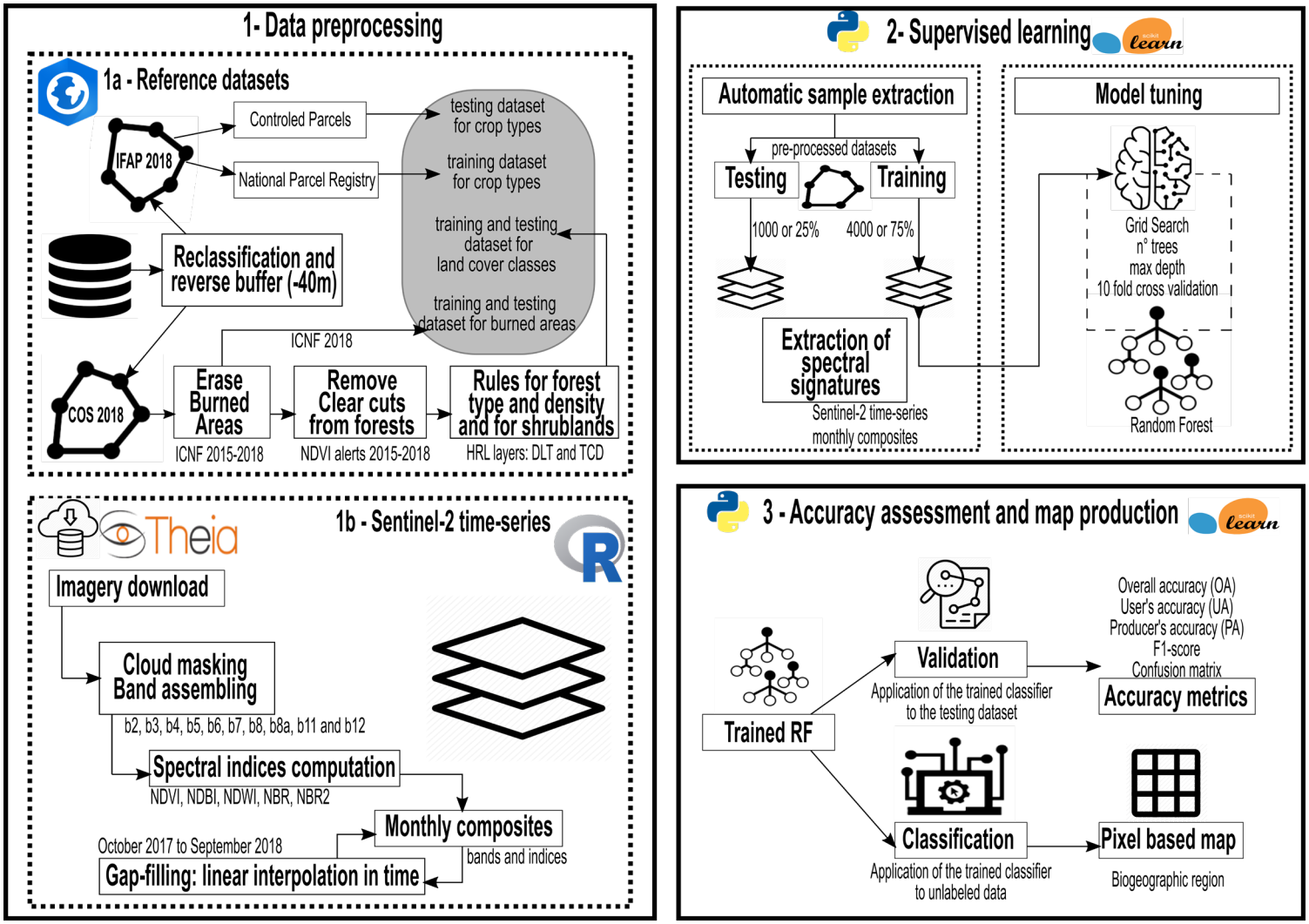


Figure 4 Flowchart for the automatic production of a land cover and crop type map in central Portugal

3.3.1 Software and device specifications

The data pre-preprocessing was done using the software ArcGIS Pro 2.4.0 from ESRI. A geoprocessing workflow was developed with the visual programming language of Model Builder that can later be exported as a python script.

The sample extraction, model training, and classification were done using Anaconda Distribution that is an open-source platform to perform data science and machine learning. The version installed corresponds to Anaconda3-4.4.0-Windows-x86_64 that contains Python (3.5.4) and the required libraries such as NumPy (1.13.1) [44] and Pandas (0.20.3) [45] for data structures, Seaborn and Matplotlib (2.0) [46] for data visualization and Scikit-Learn [47] for conducting machine learning analysis since it includes the random forest classification algorithm. Other libraries installed comprise GeoPandas for its spatial functionality with geospatial data and the Geospatial Data Abstraction Library (GDAL), which is a translator library for raster and vector geospatial data formats [48].

The feature extraction, classification, and the elaboration of the final map were done using the computers of DGT. The computers have an installed RAM of 64.0 GB with a processor Intel (R) Xeon (R) Gold 6140 CPU @ 2.3GHz 2.29 GHz. For all the other procedures, a personal computer was used, with a processor Intel (R) Core (TM) i7-7500U CPU @ 2.70 GHz 2.90 GHz and installed RAM of 8.00 GB.

3.3.2 Preprocessing of the reference datasets

The workflow for pre-processing the reference datasets is detailed in Figure 5. First, the IFAP 2018 dataset was reclassified from 175 crop types to 14, and a buffer of -40m was applied. Next, from the 83 classes available in COS 2018 dataset, a total of 15 were extracted; likewise, an inner buffer was used. Then, the remaining polygons were crossed with the auxiliary data; this includes the ICNF burned areas 2015-2018, NDVI alerts of clear cuts 2015-2018, and HRL layers 2015 (DLT and TCD). Finally, if IFAP had overlapping areas with COS, these were removed from the latter; the final dataset comprises IFAP 2018, COS 2018 and ICNF 2018. As the IFAP controlled parcels were used for testing while the national parcel registry was used for training, both datasets were kept spatially independent. This is not the case for COS nor ICNF, being the whole dataset used both for training and testing.

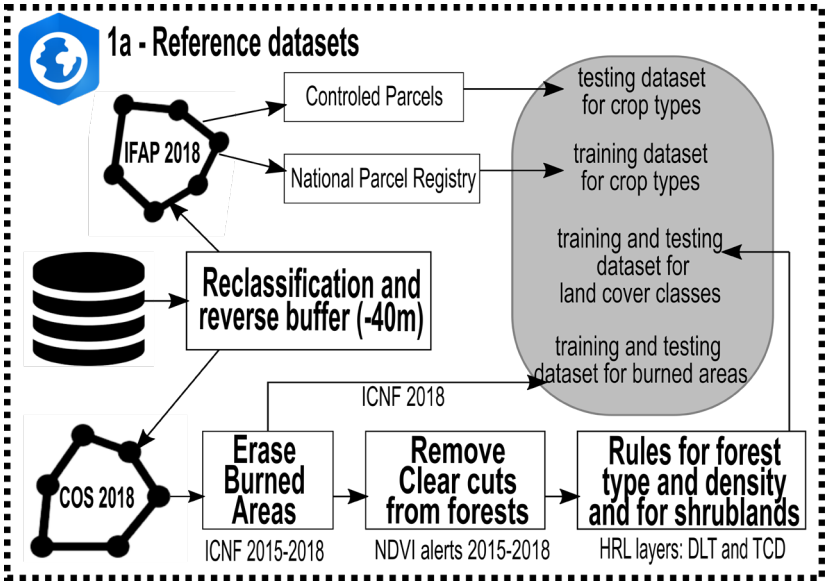


Figure 5 Data preprocessing workflow for reference datasets

Crop type dataset (IFAP 2018)

From the IFAP dataset, the ten most abundant temporary crops (5 rainfed and 5 irrigated), three permanent crops, and the agricultural grasslands were selected for the analysis for a total of 14 crop types classes. An Exploratory Spatial Data Analysis (ESDA) was performed to identify the ten most abundant temporary crops in the study area for the classification (Figure 6); these corresponds to maize (24,012ha), rice (21,595 ha), tomato (12,742ha), ryegrass (5,472ha), oatmeal (4,163ha), wheat (2,723ha), sorghum (2,104ha), barley (1,844ha), lupin (1,762ha) and potato (1,748ha). As for the permanent crops, olive trees, vineyards, and orchards were considered due to their importance in Portugal's agriculture. The orchard class combines 17 types of trees from figs and oranges to walnuts and hazelnuts.

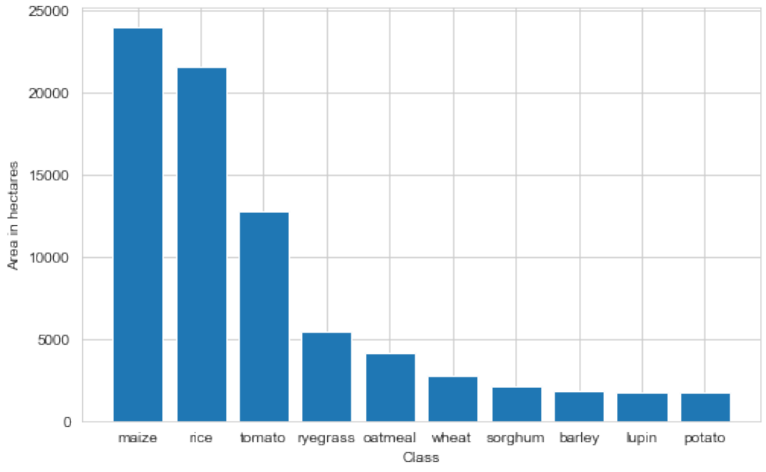


Figure 6 Area covered by the 10 most abundant crops in hectares

The number of parcels is relevant; the more parcels are distributed within the study area, the more representativity is possible to obtain. According to the distribution of the parcels (Figure 7), most of them are less than 10 ha. The average area for all the classes is 3.35 ha, being the tomato parcels with the higher mean area (5.38ha) and oatmeal the lower mean area (2.35ha).

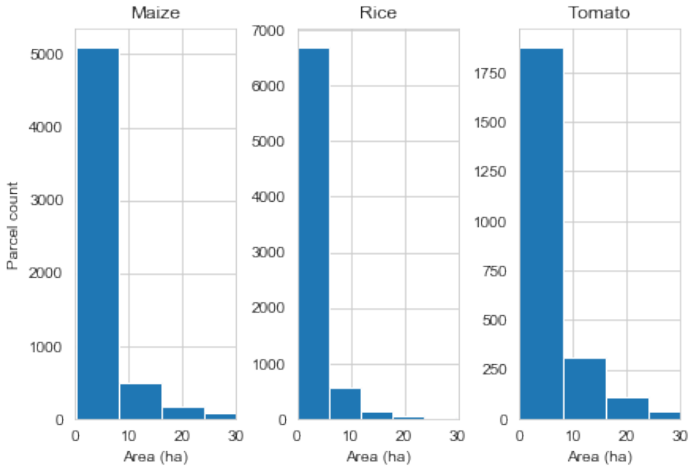


Figure 7 Distribution of the tree main crop parcels by area (ha)

An inverse buffer of -40 m was performed to the original parcels to avoid selecting pixels for which the spectral signature does not match the class label, as can be seen in Figure 8. During the buffering process, it can occur that the smallest parcels are removed from the dataset, reducing the number of available pixels for training. No other crossing with ancillary data was applied to this dataset, training (national parcel registry) and testing (controlled parcels) were kept independent. This was ensured by performing an intersection between the datasets and removing the controlled parcels from the national parcel registry guaranteeing that all the polygons are spatially disjointed.

Figure 8 IFAP parcels pre-processing: inverse buffer (- 40m)

Land Cover dataset (COS 2018)

A total of 16 LC classes were derived from the 83 classes available in COS 2018, and a -40 m buffer was applied to the remaining polygons. Yet, it is critical to emphasize that the MMU of COS (1ha) entails a reduction in detail to better model the reality; and many times, it requires the generalization of polygons. This means that areas smaller than 1ha (paths, edifications and other objects) will be aggregated with the predominant class up to 25% of the total area of the polygon [7]. Classification at the pixel level for Sentinel-2 contemplates a 10 m MMU; therefore, some pre-processing steps are required to prevent the selection of pixels with spectral information that mismatch the class label inherited from COS which has a larger MMU (and potential thematic errors).

The first step was to intersect all COS polygons with the ICNF burned mask for the years 2015- 2018. The mask allowed to create holes in the polygons by eliminating the scorched areas; consequently, no automatic sample will be extracted from these areas. A total of 392 ha in 2015, 2565 in 2016, 6002 ha in 2017, and 99 ha in 2018 were removed from the dataset. In Figure 9, it is possible to recognize a blackened area inside a Maritime Pine class in 2016 and inside a Eucalyptus class in 2017. Still, the burned mask does not cover the polygon extensively, as it can be appreciated in the Maritime pine where two holes remained in the polygon corresponding to edifications and in Eucalyptus where the mask does not cover the total extent of the area. The forest type most reduced in the area after applying by the burned mask correspond mainly to cork oak (263 772 ha) and eucalyptus (80 323 ha). In Portugal, the eucalyptus forest is industrially grown to supply pulp fiber for the paper industry, although they are highly flammable [49].

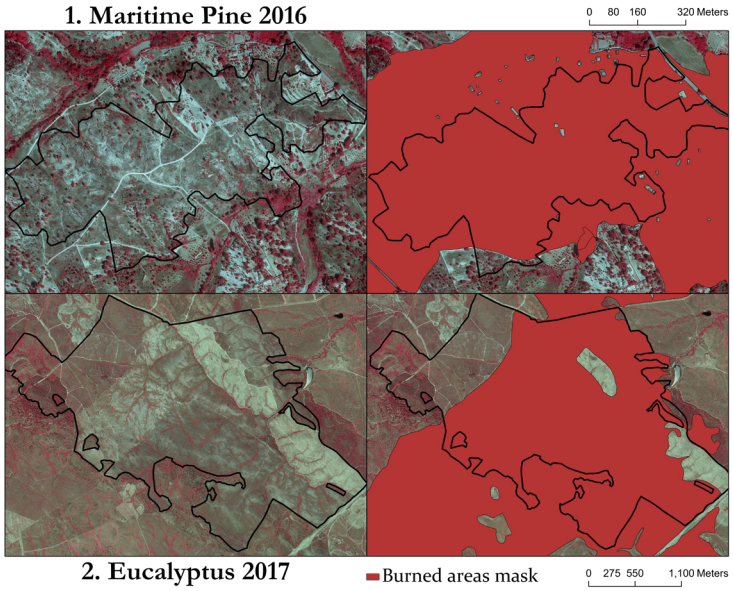


Figure 9 COS2018 polygon overlaid with the burned mask, Maritime Pine OBJECTID: 493848 and Eucalyptus OBJECTID: 415864. Scale 1:7,000 (1) and 1:25,000 (2).

The second step applies only for the forest areas since training and testing cannot be sampled from forest cuts. NDVI differencing techniques allow discriminating between real changes and seasonal or inter-annual variability of forests [50]. This technique has been implemented in Portugal [51] to detect vegetation loss that occurred between 2015-2018 in forests, so-called NDVI alerts. The forest polygons were crossed with the NDVI alerts mask to remove the areas where there have been changes, and hence the class label of COS does not correspond to the pixel spectral signature. After applying the mask, the most affected forests are eucalyptus with 13785 ha reduced, followed by 7076 ha in stone pine and 3498 ha in maritime pine. This forest fragmentation (i.e., breaking of large, contiguous forested areas into smaller pieces of a forest) is due to some extent to road construction, fires, logging and conversion to agriculture. In the case of forest plantations like eucalyptus, clear-cuts are part of the forest management cycle; as a new forest is expected to follow, the land use remains a forest [51]. However, in a strictly land cover map derived from supervised classification, these changes in vegetation can result in misclassifications when implementing the model and therefore require to be removed from sample extraction.

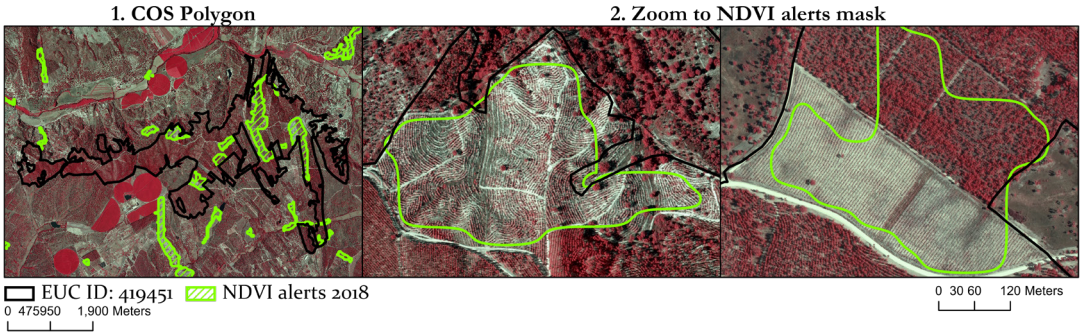


Figure 10 NDVI alerts in a Eucalyptus plantation (COS 2018 OBJECTID: 382944). Scale 1:80,000 (1) and 1:6,000 (2).

The final step was to cross the forest areas, and the shrublands with the High-Resolution Layers (HRL) masks created following the rules in Table 4. The Dominant Leaf Type (DLT) allows separating broadleaf or coniferous majority, while the Tree Cover Density (TCD) ranges from 0 to 100%. For cork oak, holm oak, other broadleaf, and eucalyptus, it is required that they correspond to broadleaf with a tree cover higher than 60% of the pixel in the HLR. For coniferous, if more than 60% of the pixel is covered by

trees, then the classes stone pine and other coniferous are defined. However, for the class maritime pine, if the coverage is more than 60%, it is considered a closed maritime pine. Although, if the pixel coverage is between 10% and 60%, a new class is derived, and it is considered an open maritime pine. A shrub is a type of vegetation that is included in many classes as a percentage in the area, making it challenging to identify. For the shrubland class, the rule is to remove from the class all the areas with broadleaf or coniferous cover.

Dominant Leaf Type (DLT)	Tree Cover Density (TCD)	Class
Broadleaf	> 60%	Cork oak forest
		Holm oak forest
		Other broadleaf forest
		Eucalyptus forest
Coniferous	> 60%	Stone pine forest
		Closed Maritime pine forest
		Other coniferous
	> 10% and < 60%	Open Maritime pine forest
Broadleaf and Coniferous	0%	Shrubland

Table 4 Rules for the crossing of COS polygons with HRL

Following the application of the mask, there is a dramatic reduction in the area for all the classes. Cork oak presented the highest reduction of 260000 ha, followed by eucalyptus with 64000 ha and stone pine with 51000 ha. Figure 11 exemplifies the filtering using the HRL layers in shrubland. It is possible to visualize that the areas with broadleaf and forest containing more than 0% of tree cover density are masked out the shrubland polygon.

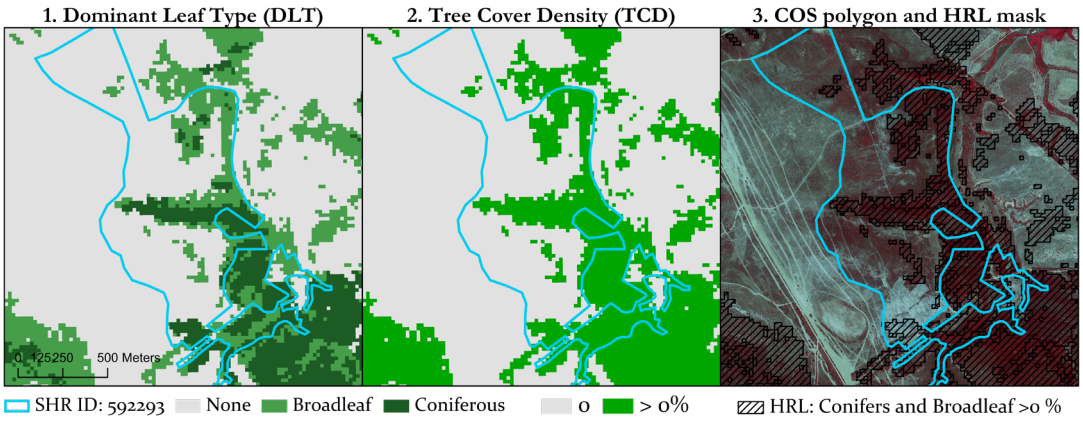


Figure 11 DLT (1) and TCD (2) rule for shrublands (COS OBJECTID: 592293). Scale 1:20,000.

At last, all the polygons within classes must be mutually exclusive; therefore, the polygons from the IFAP dataset were intersected with the polygons from COS. The overlapping areas were erased from the COS land cover dataset, giving priority to IFAP. When removing the areas that overlapped between IFAP and COS dataset, a significant conversion from land cover to crop types were found. Three main classes were reduced in the area: 114 ha of natural grassland, 72ha of open maritime pine, and 121ha of cork oak were reclassified to agriculture.

3.3.3 Preprocessing of the intra-annual time series of Sentinel 2

The following descriptions are summarized from the technical specifications for the generation of multi-temporal Sentinel-2 composites for mainland Portugal [52]. The workflow includes acquisition and preprocessing, indices computation, generation of the monthly composites, and filling of missing values, as illustrated in Figure 12.

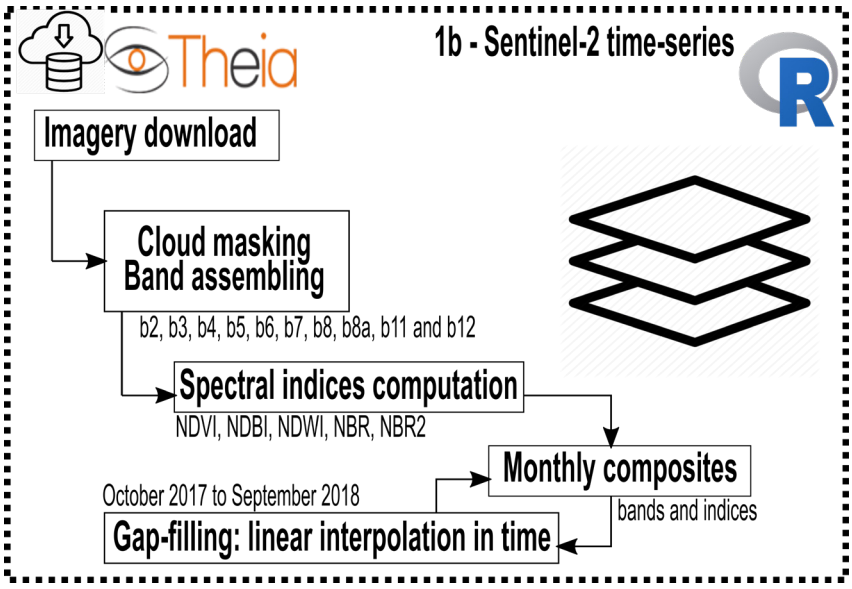


Figure 12 Preprocessing workflow for Sentinel-2 intra-annual time series

Acquisition and preprocessing

The Sentinel-2 images downloaded from THEIA for the agricultural year of 2018 comprise a cloud coverage < 50%, each tile contains around 81 images and occupy 51 GB per tile. The THEIA images available for download are already pre-processed with an algorithm named MAJA and have a more efficient cloud masking algorithm when compared with the original ESA Sen2Cor Sentinel-2 processor [30]. The MAJA algorithm provides atmospheric correction to the bottom of the atmosphere (BOA), a mask for

clouds and cloud shadows, water and snow, and contains a slope effect correction allowing the images to be seen from a flat surface.

The preprocessing of the L2A products at DGT comprises the use of the cloud/cloud shadow mask Tiff available for each product to convert all the pixels contaminated to "missing data" that corresponds to "65535". Then, the bands 5, 6, 7, 8A, 11, and 12 are disaggregated from 20m to 10 m and then assembled in the following sequence: B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12 resulting in a raster with a cell size of 10 m and with ten bands. Finally, the output images are saved as a TIFF file in the projected CRS of WGS 84/UTM zone 29N (EPSG: 32629) and as 16 bits unsigned integer where the floating values were multiplied by 10 000 to save space on the disk.

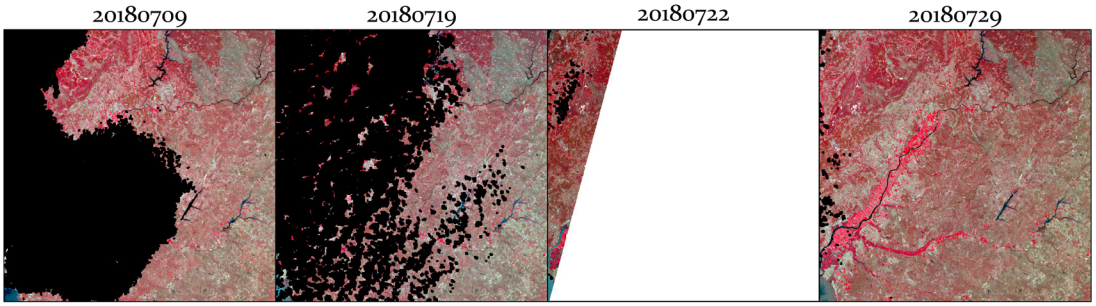


Figure 13 Image acquisition with less than 50% cloud cover for tile 29SND in July 2018

Derived indices

The bands contained in the MSI of SENTINEL-2 allow the calculation of several spectral indices by combining the spectral bands to enhance vegetation, soil, water, and built-up areas. After the imagery pre-processing, five spectral indices are calculated for each image and are summarized in Table 5.

Index	Band combination	Reference
Normalized Vegetation Index (**NDVI**)	**(b8-b4)/(b8+b4)** (NIR-Red)/(NIR+Red)	to enhance vegetation [53]
Normalized Difference Build up Index (**NDBI**)	**(b11-b8a)/(b11+b8a)** (SWIR1-NIR2)/(SWIR1+NIR2)	to map urban built-up area [54]
Normalized Difference Water Index (**NDWI** or NDMI)	**(b3-b8)/(b3+b8)** (Green-NIR)/(Green+NIR)	to detect water bodies [55], [56]
Normalized Burn Ratio (**NBR**)	**(b8a-b12)/(b8a+b12)** (NIR-SWIR2)/(NIR+SWIR2)	to highlight burned areas [57]
Normalized Burn Ratio 2 (**NBR2** or NDMIR)	**(b11-b12)/(b11+b12)** (SWIR1-SWIR2)/(SWIR1+SWIR2)	variation of NBR [58]

Table 5 Spectral indices derived from the monthly composites

Monthly composites and gap filling

All the imagery was used to create a composite for every month. For that time interval, for the images acquired in the same month (e.g., October 2017), the median is calculated at the pixel level. This is done for all the bands in the images, allowing to reduce the number of missing values because it is possible that between the acquisitions, there is a clear sky. However, this might not be the case for all pixels during a month, and this can show as "missing values" in the synthetic composites as well. A linear interpolation method was applied to the pixel with missing value using the previous and following months to fill in the gaps [24]. All the monthly composites and the indices were interpolated to create a pixel-level consistent reflectance composite that can capture field level phenologies [24], as seen in Figure 14.

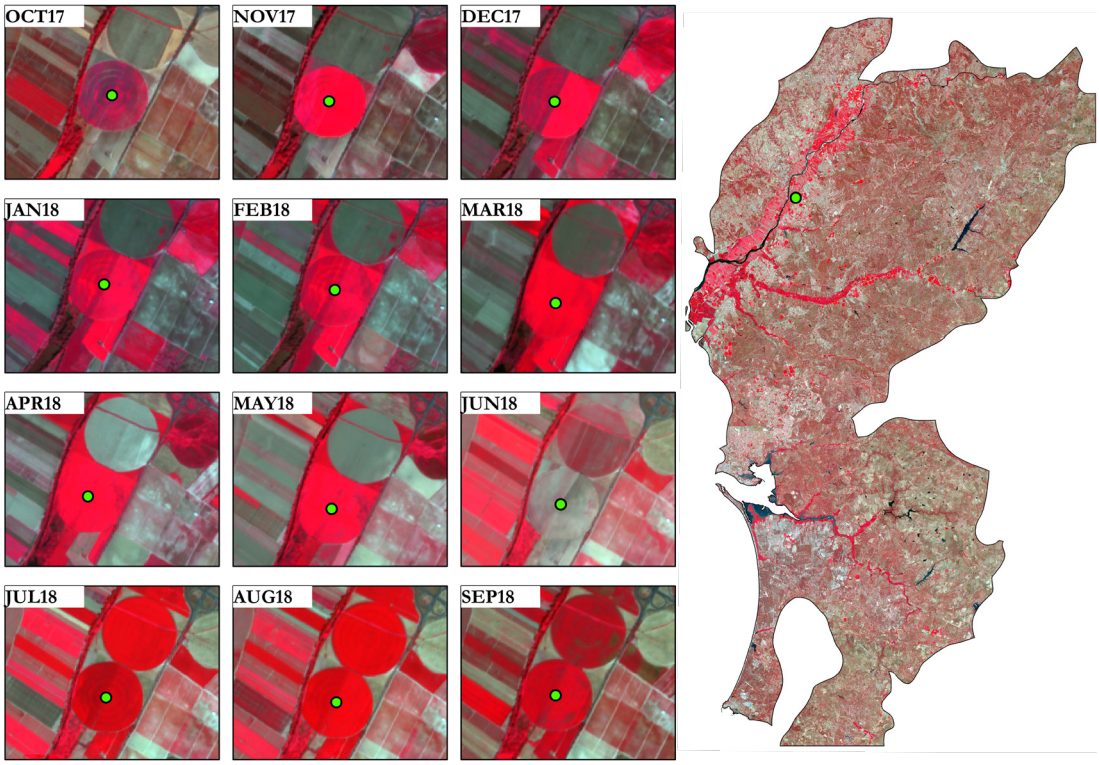


Figure 14 Series of monthly cloud-free reflectance composites at 10m resolution (October 2017 to September 2018) the pointed area corresponds to an agricultural area.

3.3.4 Supervised learning

This subsection is comprised of the automatic extraction of random samples per class from the preprocessed datasets. Next, the spectral signatures are retrieved from the

monthly composites. And finally, different hyperparameters combinations are tried to select the best model for classification, as can be appreciated in Figure 15.

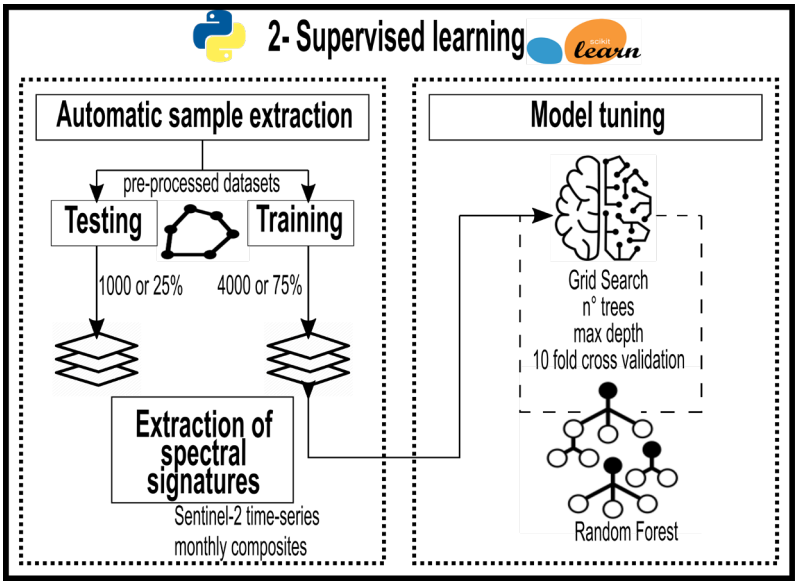


Figure 15 Supervised learning workflow

Automatic sample extraction

After the pre-processing steps mentioned in section 3.3.2, there was a decrease in the area available for automatic sample collection. In the case of IFAP, 58% of the original area was reduced for both training and testing datasets. Similarly, the landcover dataset had a significant decrease in the area; it diminished 94% from its original area. Therefore, if the number of pixels available per class was higher than 5000 samples, the dataset is divided into 4000 training and 1000 for testing. However, some classes did not meet this requirement, and for these, the dataset was divided into the proportion of 75% for training and 25% for testing. A total of 115,880 samples were retrieved for training, and 29,150 samples for testing as can be appreciated in Table 6.

Class	Training	Testing	Total
Class > 5000 samples	4000	1000	5000
Class < 5000 samples	75%	25%	100%
Total number of samples	115 880	29 150	145 030

Table 6 Training and testing samples for the area

Only four classes are imbalanced (Table 7), and they represent 6.9% of the total dataset. A possibility for dealing with class imbalance is to under-sample the majority class at the disadvantage of reducing the overall accuracy; an alternative is to oversample the minority class by duplicating the records [34]. Dealing with imbalance, it is out of the scope of this research; for this class imbalance, the User's and Producer's accuracy and the f-1 score are considered to complement Overall Accuracy.

Class	Training	Testing	Total
Barley	4000	856	4856
Holm oak forest	3408	1136	4544
Other coniferous forests	278	93	371
Bare Rock	194	65	259
Total	7880	2150	10030
Percentage of the dataset	6.8	7.3	6.9

Table 7 Training and testing samples for the four classes with imbalance

The vector dataset was rasterized to 10 m cell size using the sentinel imagery as a reference to extract the samples. Then, the resulting raster was converted to points; these correspond to the centroids of each pixel contained within the raster. This step permitted to create a point grid for random selection of samples, as seen in layout number 2 of Figure 16 and Figure 17. For the land cover classes, the training and testing samples come from the same polygons; this can incur somewhat optimistic accuracies [3]. Whereas, for the crop type classes, the availability of a verified dataset (controlled parcels) permits to have disjoint polygons for training and validation.

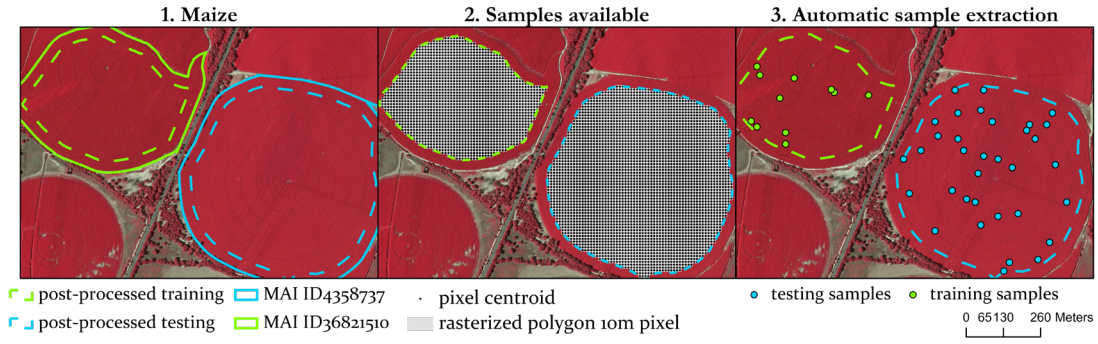


Figure 16 Automatic sample extraction for IFAP 2018 (Maize OSAID: 4358737-training; 36821510-testing). Scale 1:12,500.

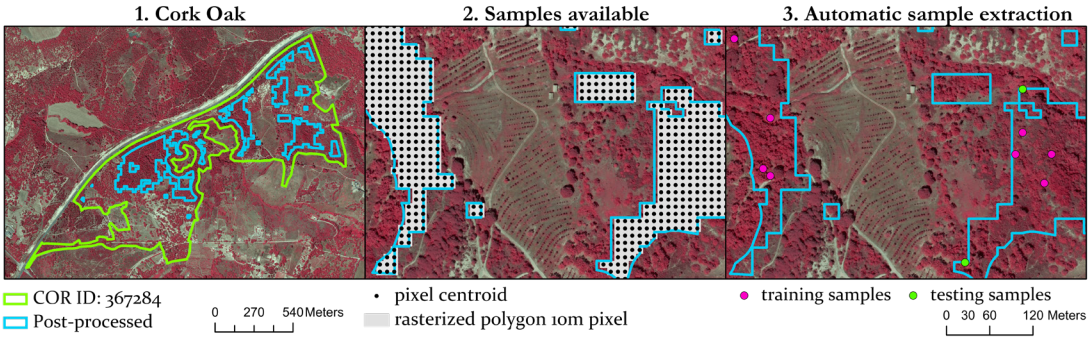


Figure 17 Automatic sample extraction for COS 2018 (Cork Oak OBJECTID: 382944). Scale 1:20,000 (1) and 1:5,000 (2) and (3).

The tiles 29SND and 29SNC have overlapping areas; to avoid that the same sample extract features from both tiles, the samples were divided. The priority was given to tile 29SND as it contains more images available for the period; the training and testing datasets were cropped to the whole extent whereas, for tile 29SNC, the overlapping area was not considered (Figure 18). The green dots indicate the samples that will extract spectral information from tile 29SND, and the blue ones will retrieve the information from tile 29SNC.

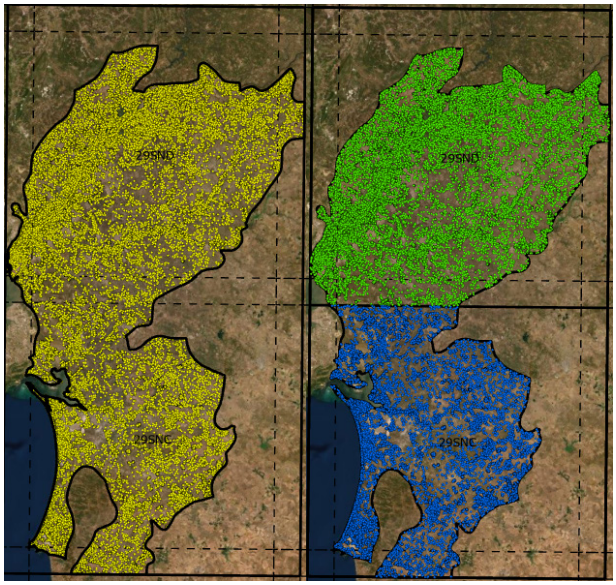


Figure 18 Training and testing samples for the biogeographical region divided between tiles 29SND and 29SNC.

Extraction of the spectral features

Before inputting the data into the model, the training and testing dataset must contain the required features/variables. The information that is provided to the classifier corresponds to the surface reflectance values of the Sentinel-2 composites and the spectral indices derived. A total of 180 features are extracted, equivalent to 10 bands and five spectral indices for each of the 12 months (October 2017 to September 2018), as shown in Figure 19. This process was done with python, adapting the code rs-util in section 7.1. The samples were used as a mask to extract the spectral information at the pixel level from all the imagery. The data retrieved is saved as an array of 180 features; all the arrays were converted to a pandas DataFrame, a two-dimensional tabular data structure in .csv format.

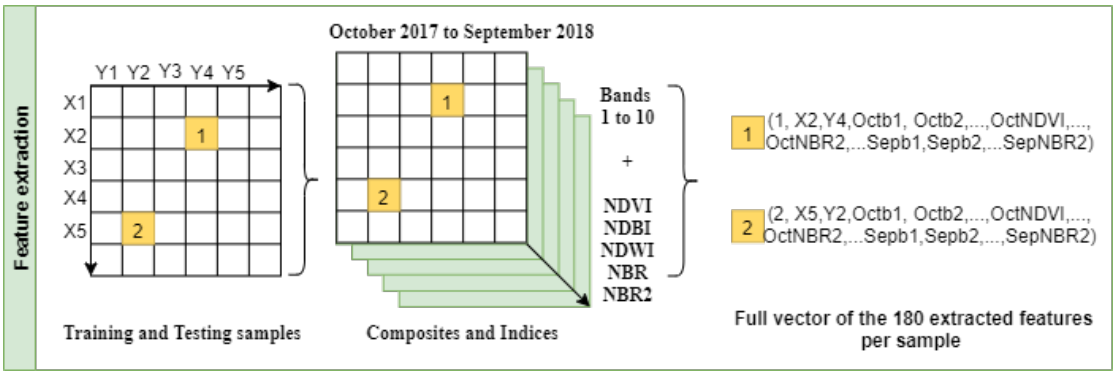


Figure 19 Extraction of the spectral features for training and testing datasets

The resulting training DataFrame contains a total of 115,880 samples with 180 features extracted (Figure 20), while the testing DataFrame contains 29,150 samples. The rows represent the samples used to extract the data at the pixel level, and the columns contain the features retrieved from the Sentinel-2 imagery. The 'CLASS' column contains the labels corresponding to the land cover and crop type target classes; however, the RF model does not accept string variables and therefore, the numerical codes 'LV4' are the ones used as input in the classification. The total time for feature extraction was approximately 2 hours using the computers at DGT.

	LV1	LV2	LV3	LV4	CLASS	OCTb2	OCTb3	OCTb4	...	SEPNBR	SEPNDBI	SEPNDMIR	SEPNDVI	SEPNDWIF
0	3	31	311	3111	Agricultural grassland	697.0	879.0	1198.0	...	1292.0	1207.0	2460.0	4264.0	-5171.0
1	5	51	514	5141	Other broadleaf forest	253.0	416.0	321.0	...	6538.0	-3369.0	4064.0	8117.0	-7410.0
2	5	51	513	5131	Eucalyptus forest	242.0	344.0	322.0	...	6222.0	-3247.0	3727.0	7255.0	-7071.0
3	2	21	212	2125	Potato	1692.0	2226.0	2810.0	...	-490.0	1193.0	708.0	1340.0	-2468.0
4	2	21	212	2125	Potato	1300.0	1580.0	1989.0	...	-628.0	1415.0	793.0	1546.0	-2659.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
115875	3	31	312	3121	Natural grassland	1620.0	2428.0	3094.0	...	519.0	805.0	1319.0	1176.0	-2566.0
115876	2	21	212	2125	Potato	1183.0	1559.0	2000.0	...	-795.0	1607.0	822.0	1670.0	-2837.0
115877	5	51	513	5131	Eucalyptus forest	244.0	352.0	335.0	...	6370.0	-3306.0	3881.0	7462.0	-7295.0
115878	2	21	212	2125	Potato	1673.0	2316.0	2989.0	...	-557.0	1233.0	681.0	1058.0	-2340.0
115879	5	51	514	5141	Other broadleaf forest	327.0	538.0	482.0	...	5994.0	-2790.0	3847.0	7389.0	-6628.0

115880 rows × 185 columns

Figure 20 Dataframe of the training dataset containing the 115,880 samples with 180 extracted features

One of the benefits of the DataFrame tabular structure is that it allows queries and arithmetic operations along both rows and columns. For visualization purposes, the values of only one month were acquired; in this case, October 2018 (Figure 21). From the features obtained, we can appreciate that there are high correlations between the visible spectrum (b2, b3, b4) and the red-edge bands and NIR for vegetation discrimination (b6 to b8a) and between the bands 11 and 12 used for snow/ice and cloud discrimination (SWIR). The bands 1 for aerosol detection and 9 of water vapor were not included in this analysis as they would not provide information about surface reflectance's values

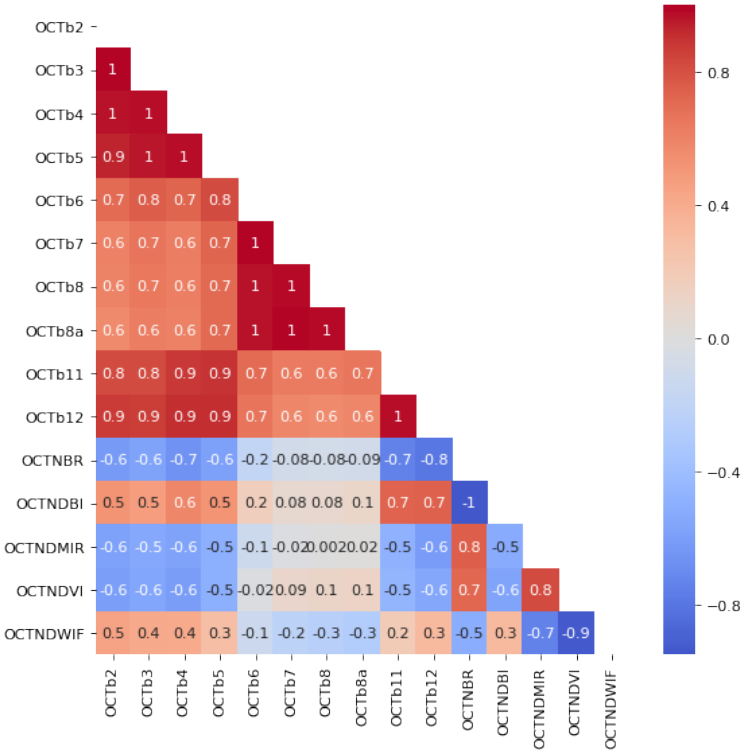


Figure 21 Correlation between spectral signatures for bands and indices for October 2017

Model tuning and training

Model evaluation is a required task in classification using machine learning algorithms; a traditional way to validate the performances of classification is to use part of the available samples for training and another for validation [30]. The 10-fold cross-validation method [59] allowed to randomly split the training dataset into 10 parts and create 10 validation experiments using each time a different part to validate and 9 other parts for training the classifier. This was done for each of the hyperparameters in the Grid Search Table 8.

The RF model was built using the open-source Scikit-learn library for machine learning that is available to deploy using `sklearn.ensemble.RandomForestClassifier` [47].

The `sklearn.model_selection.GridSearchCV` allowed testing various hyperparameters combinations of RF to achieve the optimal classification performance for the training dataset; a summary is provided in Table 8.

Algorithm	User-defined parameters	Values
Random forest	Criterion	Gini
	Number of trees	100, 200, 300, 400, 500
	Maximum depth	None, 2, 16

Table 8 User-defined parameters for Grid Search

The Gini criterion is considered to extract the variable importance, and the number of estimators is tested from 100 that is the default value in Scikit learn to 500, which is the recommended value of RF [39]. Although RF does not overfit, it is possible not to prune the trees; nevertheless, the test is considered from 'none' until 16 splits. By leaving the max features parameter in 'auto,' the software will consider the number of features at each split (m) to be $\sqrt{p}$ (being p the total number of variables); this the default for RF [38]. The training dataset (80% total data for classes with 5000 samples and 75% for classes below 5000 samples) will be used to train the model parameter and adjust the parameters using 10-fold cross-validation.

The best performing model was selected by the ranking achieved in the training dataset, and the 10 cross-validations results for that model are presented in Table 10. A total of 500 trees without pruning outperformed the other models tested; this is consistent with most literature on RF reporting that the error stabilizes before reaching 500 trees and that is recommended to let each tree overfit until the node reaches purity [34], [37], [39]. A total of 15 models were tested using 10 cores and 32 GB of RAM of the DGT servers, the total training time took 93 minutes, for 100 trees as estimator the average training time is 2 minutes while for 500 trees it takes 10 min (Table 9).

Mean accuracy	Number of trees				
Max depth	100	200	300	400	500
None	5	4	2	3	1
4	15	14	11	13	12
16	10	9	6	8	7
Mean fit time (min)	2	4	6	8	10

Table 9 Ranking of the hyperparameters grid using 10-fold cross-validation

1	2	3	4	5	6	7	8	9	10	Mean	Std
0.609	0.78	0.737	0.703	0.740	0.753	0.693	0.64	0.63	0.57	0.687	0.065

Table 10 Cross-validation results for the best model (10 folds, mean and standard deviation)

3.3.5 Accuracy Assessment of the model performance and map production

This final subsection focuses on validating the performance of the model and classifying both tiles to produce a final map in raster format; the biogeographical region is used as a mask to extract the ROI (Figure 22).

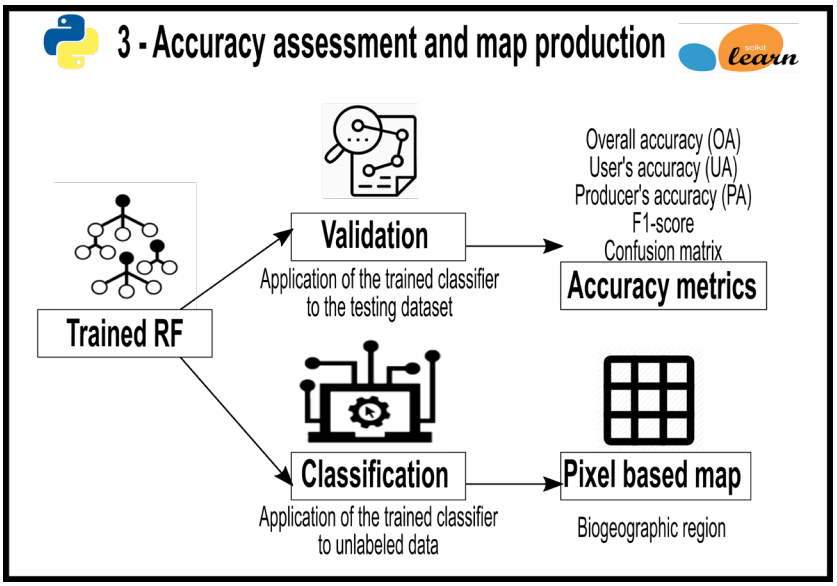


Figure 22 Accuracy assessment and map production workflow

Validation

When using classification models in remote sensing, it is required to quantify the number of times the model predictions match the reality being modeled [36]. Accuracy assessment compares the pixels or polygons from a map classified using ML algorithms to a reference test dataset (ground truth) for which the labels are known [3]. A confusion matrix summarizes the classification performance, where the row entries are the actual classes (reference data), and the column entries contain the number of pixels predicted by the classifier belonging to the column class. In a two-class problem, the confusion matrix is a two-dimensional matrix, one designated the positive class and the other the negative class (Table 11). True positives (TP) are the positive samples correctly classified, and False positives (FP) are a negative class incorrectly classified as positive. Whereas True negatives

(TN) are the negative samples correctly classified, and False Negatives (FN) are a positive class incorrectly classified as negative [36].

		Assigned Class	
	Total population	Positive	Negative
Actual class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 11 Binary confusion matrix

From the binary confusion matrix, several indices can be calculated for each class, and the average of all the values across classes serve for the multiclass purposes. The evaluation metrics considered in this study are based on the python implementation of the metrics is available in the Scikit-learn documentation [47]. A summary is shown below:

- Accuracy (`sklearn.metrics.accuracy_score`): the number of correctly classified samples/total number of samples. The Error rate is directly related to the accuracy, being error rate = 1.0 – accuracy.
- User's accuracy (`sklearn.metrics.precision_score`): a ratio between true positives/total number of positives predicted ($\frac{TP}{TP+FP}$). It is the ability of the classifier not to label as positive a sample that is negative.
- Producers accuracy (`sklearn.metrics.recall_score`): a ratio between true positives/total number of actual positives ($\frac{TP}{TP+FN}$). It is the ability of the classifier to find all the positive samples.
- F1-score: harmonic mean of precision and recall ($2\,\frac{P\,x\,R}{P+R}$).

The Kappa coefficient is the proportion of agreement after the chance agreement is removed [60]; however, it is not reported in this study. Kappa can provide information on assessing the performance of a classifier, but it does not provide information on assessing a map because it is not possible to identify actual pixels classified correctly by random chance hence random classification is not a realistic alternative to create a map [18]. Several authors have explained the unsuitability of the kappa coefficient in accuracy assessment of image classification and encourage researchers to provide more straightforward metrics such as estimates or per-class accuracy and confusion matrices [18], [61], [62].

Aside from the accuracy metrics, the main misclassified classes will undergo a visual interpretation for understanding the confusion in the classifier.

Map production

For the map production, the trained RF model was applied to each unclassified pixel assigning the land cover or crop type that got the most votes in the ensemble.

The first attempt to make the map was to extract the features per tile, each tile containing 10980 x 10980 pixels (approx. 120 million pixels); however, it proved to be computationally demanding. The next approach was to classify subsamples of tiles containing 2.5 million pixels; the method worked; however, it was not efficient as this would have required to classify 48 subsamples for 2 hours each (approx. 4 full computing days). Lastly, with the help of Pedro Benevides and Hugo Costa at DGT, it was possible to implement a classification approach using multicore processing (18 cores/36 threads). Each tile was classified independently with the same trained model; the total computing time was 4 hours (2 hours per tile).

4 RESULTS AND DISCUSSION

In this section, the best hyperparameters selected are fitted to the training data then the trained model is used to classify the testing data allowing to extract the essential variables in the classification 4.1. The final map is presented in section 4.2, and it is evaluated based on the metrics proposed in section 3.3.5. Then, a particular emphasis on crop phenology follows (4.2.2) that will compare the performance of the use of time series for specific band reflectance's based on the crop calendar for Portugal. And finally, a visual assessment of the map (4.2.3) is necessary to comprehend the extent of the accurate classifications but also the misclassifications related to the pre-processing steps.

4.1 Variable importance

After choosing the best parameter, the random RF was trained accordingly with 500 trees and no pruning; a random state of 101 was used to ensure reproducibility. The model was applied to the testing data to retrieve the variable importance, from the 180 variables used in the model. The 10 most informative variables and the 10 least informative variables are displayed in Table 12, along with their scores. The variables are ranked from 0 to 1, meaning that the closer they are to 1, the more information they provided during the split of the decision trees.

The most critical features correspond to the months of spring (April) and summer (June and August), whereas the least important is during autumn (November and December) and winter (January and February). While the most influential bands are in the Red edge (b5, b6, and b7), NIR (b8a) and SWIR (b11 and b12) wherein the least important are in the visible spectrum (b2, b3, and b4) and the NIR (b8).

This correlates with the availability of spectral information, some of the months are entirely cloud-free for all the mainland Portugal (October 2017, November 2017 and August 2018), but there are also some critical periods where the number of missing values is significant (e.g., February, April or July) [52]. This is the main reason why the monthly composites are interpolated in time, as some of the months can have pixels with large spans of missing data.

Also, the Mediterranean type of climate in Portugal is characterized by warm and dry summers and cool and wet winters [63]; the vegetation utilizes the precipitation that is accumulated from November to April during spring and summer for its photosynthetic activities. The photosynthetic phenology is captured by the different bands in the MSI,

mainly in the Red Edge and NIR, whereas the SWIR allows penetrating thin clouds for moisture discrimination on soils. The high correlation between bands 8 and 8a influences the information gain in the split selection, reducing the importance for the band 8 and assigning more weight to band 8a. The inclusion of the five spectral indices appears to provide a minor information gain in the classification, making them not a predominant variable [19].

10 most important variables in the model									
JUNb8a	AGOb11	JUNb11	AGOb5	JUNb6	APRb8a	JUNb7	ABRb7	AGOb8a	JUNb12
0.01146	0.01060	0.01053	0.00998	0.00924	0.00913	0.00882	0.00869	0.00851	0.00849
10 least relevant variables in the model									
JANb3	DICb8a	FEBb4	DICb6	JANb8	DICb7	FEBb2	FEBb8	NOVb8	DICb8
0.00297	0.00294	0.00286	0.00276	0.00270	0.00267	0.00265	0.00253	0.00239	0.00228

Table 12 Extraction of the 10 most important feature in the classification and the 10 least relevant for the selected model

4.2 Land Cover and Crop Type Classification

The land cover and crop type map in raster format (10m pixel size) is presented in (Figure 23) based on the methodology proposed in section 3. The map contains 31 classes, from which 14 correspond to agricultural classes from wheat to agricultural grasslands (in the legend), it also includes burned areas and 16 land cover classes. The quality of the map will be assessed quantitatively (4.2.1) based on the accuracy metrics from section 3.3.5 and discussed using literature.

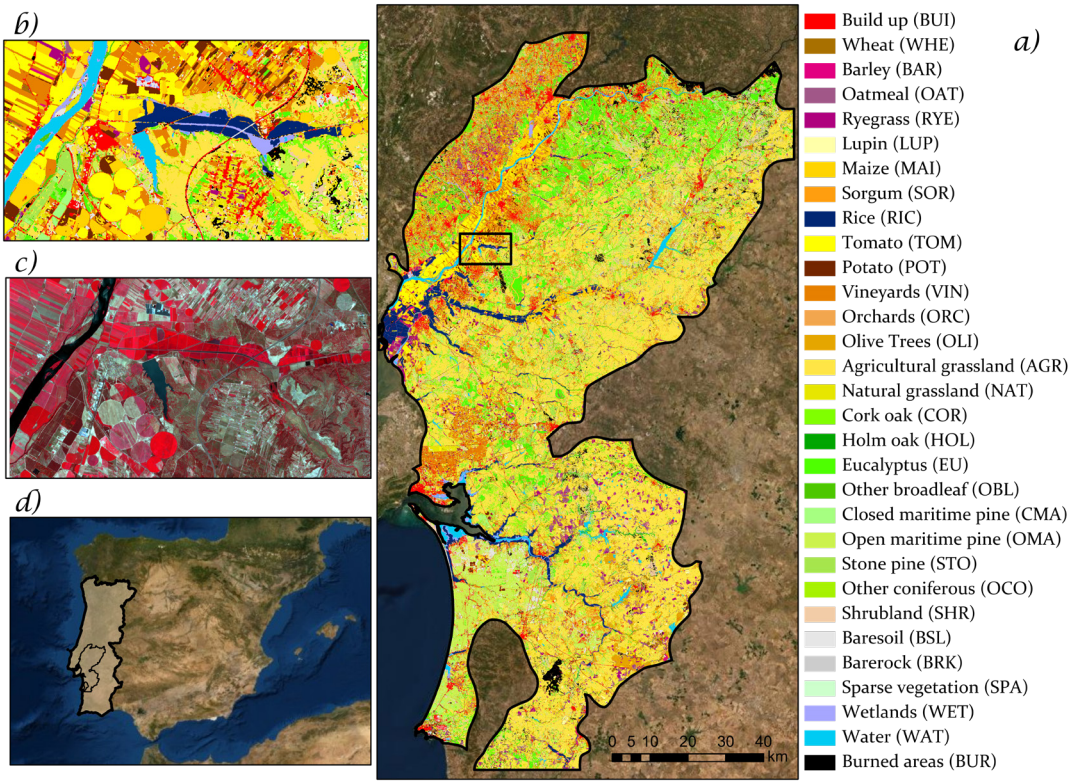


Figure 23 (a) Land Cover and Crop Type in raster format, (b) detail of the map, (c) false-color (RGB: b8, b4, and b3) for august 2018 Sentinel-2a composite, (d) the Iberian Peninsula with Portugal and Stata 214 highlighted.

4.2.1 Quantitative Map Evaluation

The accuracy assessment assumes that the map and reference labels represent hard classification and that the samples used during training are not included in the testing. The aim is to compare if the predicted map class label matched the actual label observed on the ground expressed on an overall and per-class basis [18]. First, the proportion of area correctly classified (overall accuracy) is discussed, however, as it does not provide class-specific information, the user's accuracy (UA), producer's accuracy (PA), f1-score and the number of testing samples per class are also showed in Table 13. Then, the error matrix is included to visualize the off-diagonal cells that indicate which classes are confused [18].

The overall accuracy (OA) of the land cover and crop map using monthly composites features and derived indices is 76% for the 31 classes (Table 13).

In crop types, the best-performing classes overall are maize, rice, and tomato with UA and PA values above 90%, excepting sorghum whose PA accuracy is 64% being excluded by omission and assigned to permanent crops such as vineyards, orchards, and olive trees. In general, the irrigated crops (summer crops) are more stable in their

classifications, whereas the rainfed crops (winter crops) have much confusion within themselves, as can be appreciated in the confusion matrix Table 14. In terms of permanent crops, Vineyards achieved UA and PA higher than 75%; this finding is also stated by Schmedtmann, J., Campagnolo, M. (2015) [26] that achieved 85% in parcels classified as maize, rice, wheat or vineyard in the same study area. On the other side, orchards and Olive trees have the lowest PA (35% and 46% respectively) being incorrectly classified primarily as agricultural and natural grasslands as well as other crop types. The high temporal variability of the temporary crops is detected by the different spectral signature depending on the months and the phenological state, as exemplified in section 4.2.2, allowing them to increase their classification accuracy if compared to permanent crops. Crop mapping in Central Portugal can benefit from the use of time series of Sentinel-2 and machine learning for their classification as their average size of the parcels is between 2 and 3 ha. However, for very fragmented landscapes where the agricultural parcels are comparatively smaller, it would require higher resolution imagery to meet the same accuracy [20].

For the land cover classes, the highest in UA and PA (> 90% for both) are water, bare rock, holm oak, and wetlands, whereas the lowest PA (54%) corresponds to the other coniferous class, that was more associated with Stone pine. In general, there are many confusions between the forest classes, for example, the cork oak has a low UA of 65% meaning that the commission error is high and as seen in the matrix (red bounding box), this class can be mistaken with the broadleaf forest as well as coniferous. The shrubland is often mixed with build-up, grasslands, bare soil, and sparse vegetation; the direct mapping of the latter class is challenging as its spectral signal is composed of green vegetation and non-photosynthetic vegetation as well as varying fractions of soil, grass, and shadow [64].

CODE	CLASS	ABREVIATION	UA%	PA%	F1	N°
1111	Build up	BUI	86	88	87	1000
2111	Wheat	WHE	70	67	68	1000
2112	Barley	BAR	85	54	66	856
2113	Oatmeal	OAT	50	49	49	1000
2114	Ryegrass	RYE	57	62	59	1000
2115	Lupin	LUP	62	55	59	1000
2121	Maize	MAI	99	99	99	1000
2122	Sorghum	SOR	84	64	72	1000
2123	Rice	RIC	100	98	99	1000
2124	Tomato	TOM	93	100	96	1000
2125	Potato	POT	78	84	81	1000
2211	Vineyards	VIN	76	94	84	1000
2221	Orchards	ORC	73	35	47	1000
2231	Olive Trees	OLI	61	46	52	1000
3111	Agricultural grassland	AGR	48	72	58	1000
3121	Natural grassland	NAT	56	68	61	1000
5111	Cork oak forest	COR	64	70	67	1000
5121	Holm oak forest	HOL	91	94	92	1136
5131	Eucalyptus forest	EUC	88	82	85	1000
5141	Other broadleaf forest	OBL	65	81	72	1000
5211	Closed Maritime pine forest	CMA	75	75	75	1000
5212	Open maritime pine forest	OMA	78	78	78	1000
5221	Stone pine forest	STO	84	84	84	1000
5231	Other coniferous forest	OCO	100	54	70	93
6111	Shrubland	SHR	70	63	67	1000
7111	Baresoil	BSL	82	80	81	1000
7121	Bare Rock	BRK	95	92	94	65
7131	Sparse vegetation	SPA	88	94	91	1000
8111	Wetlands	WET	90	94	92	1000
9111	Water	WAT	97	95	96	1000
9999	Burned Areas	BUR	68	72	70	1000
OA%			**76**			29150

Table 13 Land Cover and Crop Type results of the classification. The Overall Accuracy (OA%), User's Accuracy (UA%), Producer's Accuracy (PA%), F1-SCORE (F1%), and the number of testing samples (N°) are reported for the RF model with 500 trees.

Table 14 presents the confusion matrix for 31 classes using the RF model with 500 trees and the testing data set. The left column represents the ground data, the upper row corresponds to the predicted labels, and the diagonal represents the correctly classified samples per each class. The correctly classified land cover and crop type are highlighted in bold and red, whereas the classification errors higher than ten are highlighted in yellow. The bounding boxes correspond to the classes that can be aggregated from LV4 to LV3 according to the hierarchical nomenclature in attachments 7.3; these correspond to Rainfed Temporary Crops, Irrigated Temporary Crops, and Maritime Pine Forest, respectively. The red bounding box corresponds to all the forest classes. This is an image for illustrative purposes; the original table can be found in attachments 7.5.

LV4	BUI	WHE	BAR	OAT	RYE	LUP	MAI	SOR	RIC	TOM	POT	VIN	ORC	OLI	AGR	NAT	COR	HOL	EUC	OBL	CMA	OMA	STO	OCO	SHR	BSL	BRK	SPA	WET	WAT	BUR	LV4
BUI	880	1	0	0	0	0	0	0	0	0	0	15	6	9	7	18	1	0	0	2	0	2	0	0	7	25	0	5	2	0	20	BUI
WHE	1	667	10	223	10	15	0	0	0	0	3	0	1	1	6	13	0	0	0	0	0	0	0	0	0	1	0	26	0	0	23	WHE
BAR	0	226	464	0	100	0	0	27	0	0	0	1	0	3	1	7	0	0	0	0	0	0	0	0	21	0	0	6	0	0	0	BAR
OAT	0	47	27	492	65	129	0	0	1	0	0	6	0	16	207	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	OAT
RYE	0	2	0	80	619	75	0	68	0	0	3	3	10	3	32	26	0	0	0	0	0	0	0	0	0	0	0	2	0	0	77	RYE
LUP	4	0	0	122	198	552	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	LUP
MAI	0	0	0	0	2	0	990	1	0	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	MAI
SOR	0	0	2	18	8	6	0	637	2	36	0	126	16	73	46	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	SOR
RIC	0	0	0	0	0	0	10	5	985	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	RIC
TOM	0	0	0	0	0	0	0	2	0	998	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	TOM
POT	5	0	28	0	0	4	0	0	0	31	845	1	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	POT
VIN	8	0	0	0	0	0	0	0	0	0	5	944	4	3	0	12	0	0	0	0	0	0	0	0	1	17	0	1	0	0	5	VIN
ORC	0	0	0	1	25	0	0	1	0	1	59	43	350	48	42	201	11	0	0	171	0	0	0	0	1	0	0	0	45	0	1	ORC
OLI	0	0	11	18	20	24	0	6	0	0	54	33	21	460	174	89	0	3	1	1	0	4	3	0	23	7	0	1	0	0	47	OLI
AGR	2	3	0	13	11	34	0	6	0	0	0	14	3	41	720	37	45	1	2	5	0	15	15	0	15	3	0	3	1	0	11	AGR
NAT	14	1	3	13	20	16	1	2	0	0	0	18	5	26	75	679	7	1	2	11	0	6	0	0	65	17	0	3	4	0	11	NAT
COR	0	0	0	0	0	2	0	0	0	0	0	10	4	3	19	2	700	37	32	92	30	18	10	0	19	1	0	0	2	6	13	COR
HOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	29	1063	3	28	1	1	1	0	3	0	0	0	1	0	1	HOL
EUC	1	0	0	0	0	1	0	0	0	0	0	1	2	3	2	0	78	10	821	28	22	15	6	0	8	2	0	0	0	0	0	EUC
OBL	0	0	0	0	0	0	0	0	1	0	0	0	2	1	4	3	93	31	11	808	12	1	11	0	13	0	0	0	7	0	2	OBL
CMA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	58	5	28	19	751	69	60	0	3	0	0	0	1	1	2	CMA
OMA	6	0	0	0	0	2	0	0	0	0	0	0	0	2	12	0	16	3	5	2	113	781	25	0	21	5	0	5	0	0	2	OMA
STO	0	0	0	0	0	2	0	0	0	0	0	0	0	3	3	0	16	8	12	20	64	24	843	0	3	0	0	0	1	0	1	STO
OCO	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	4	9	1	20	50	0	0	0	0	0	0	0	OCO
SHR	23	1	0	4	2	4	0	0	0	0	0	9	8	16	47	42	18	3	2	28	3	25	1	0	634	69	0	40	1	2	18	SHR
BSL	36	0	0	1	0	3	0	1	0	0	1	11	7	7	14	14	4	0	0	4	0	14	0	0	22	804	2	27	5	9	14	BSL
BRK	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	60	4	0	0	0	BRK
SPA	11	0	0	0	0	0	0	0	0	1	0	1	0	0	8	3	5	0	0	1	0	5	1	0	12	8	1	942	0	0	1	SPA
WET	7	0	0	0	0	1	0	0	0	0	0	2	1	2	4	7	3	1	1	10	0	0	0	0	4	2	0	0	944	11	0	WET
WAT	1	0	0	1	0	1	0	0	0	0	0	0	0	0	3	6	0	1	0	0	0	0	0	0	2	11	0	1	24	948	1	WAT
BUR	17	0	0	6	7	13	0	3	0	0	0	8	9	37	54	35	10	0	7	8	1	14	4	0	24	7	0	4	7	2	723	BUR
LV4	BUI	WHE	BAR	OAT	RYE	LUP	MAI	SOR	RIC	TOM	POT	VIN	ORC	OLI	AGR	NAT	COR	HOL	EUC	OBL	CMA	OMA	STO	OCO	SHR	BSL	BRK	SPA	WET	WAT	BUR	LV4

Table 14 Confusion matrix for the Land Cover and Crop Type classification at LV4

4.2.2 The relevance of time series in crop phenology

The availability of cloud-free monthly composites permitted to compute a smoothed spectro-temporal profile from the averaged reflectance values in the testing dataset, allowing to capture the phenology of the monitored crops. The Figure 24 present the averaged spectro-temporal profiles for a) wheat that is a temporary rainfed crop grown during autumn/winter; b) rice mainly grown during the spring/summer (irrigated temporary) and c) vineyards that is a permanent crop; the illustrations in false color are for a specific parcel for visualization purposes. The bands displayed correspond to the Red Edge (b5) mainly absorbed by the chlorophyll present in leaves for photosynthetic activity, NIR (b8a) that, on the other hand, is strongly reflected by leaves [65] and SWIR (b11) sensible to water and soil moisture.

It is possible to envisage the growing window for wheat from January to May, characterized by the noticeable increasing values in reflectance on the NIR. On average, in the information extracted from the testing set, there is regrowth after the harvest in June. This is not the case for the specific parcel used as visualization example, as farmers can decide to grow multiple crops during the year, implement leguminous plants for soil recovery or leave the plot as fallow land. This variation on the plot usage for the rainfed crops entrains several confusions for the classifier. In rice, the SWIR band captures the rice flooding period (March) and the NIR the flowering period (July to September); possibly, these peaks allow to characterize the crop accurately for the classifier to achieve high accuracy values. For vineyards, there is no much variation through the year as it is a permanent crop; nevertheless, during the pruning (December and January), there is a slight decrease as the crop loses some of their leaves. The multi-temporal information permitted to capture the phenological variation of the crops that cannot be distinguished from single-date acquisitions, justifying the relevance of intra-annual time series in crop type classification for a specific year.

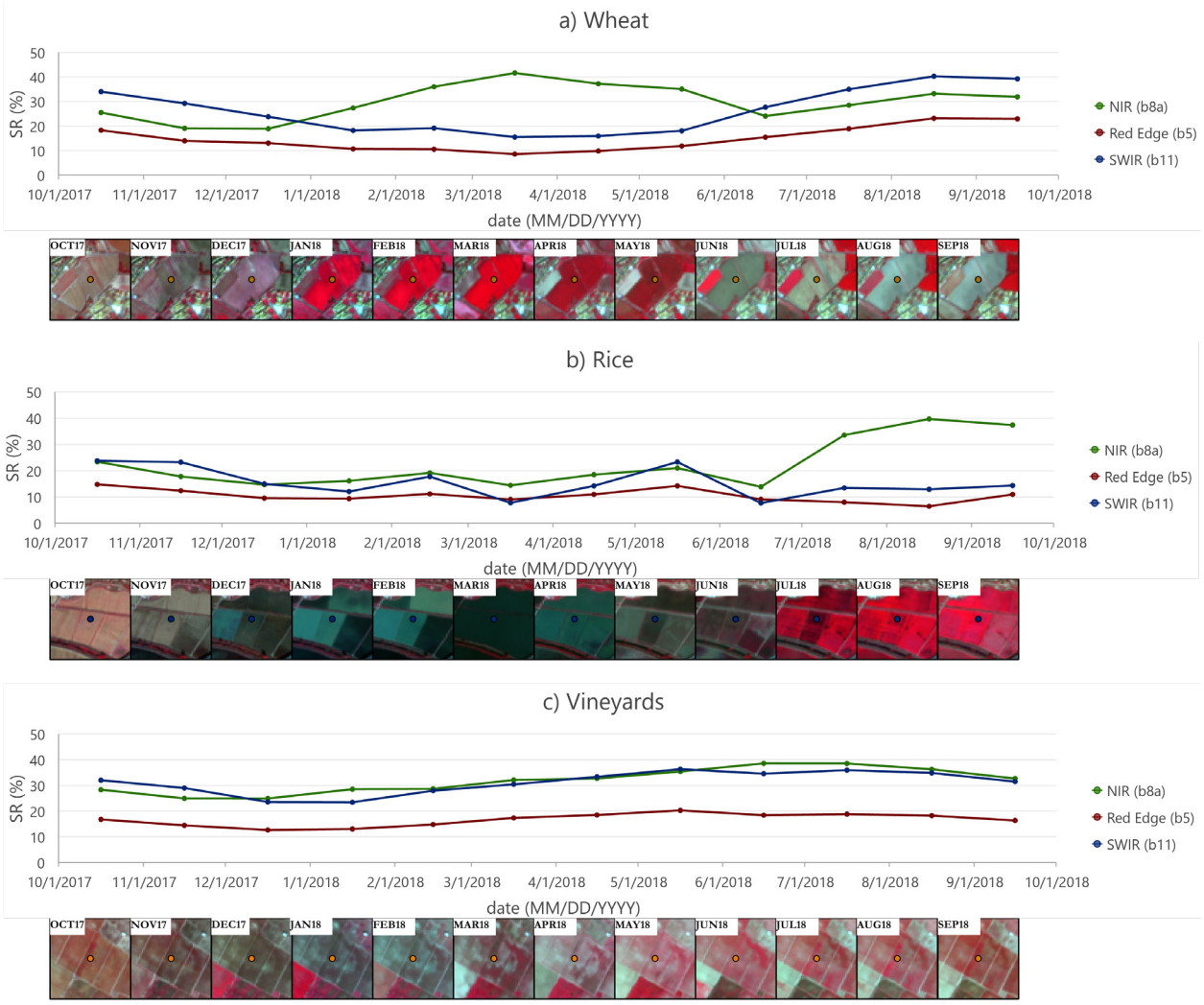


Figure 24 Average surface reflectance in the Red Edge (band 5), NIR (band 8a), and SWIR (b11) for wheat (a), rice (b), and vineyards (c) from October17 to September18.

4.2.3 Visual inspection

In order to find if what was classified in the map correspond to reality (and vice versa), three examples were selected to undergo visual inspection using as base map the orthophotos available as Web Map Service (WMS) at DGT for 2018 in False Color (RGB: b8, b4, b3 -NIR, Red, Green). The aim is to find if the pre-processing rules allowed to select the samples correctly for accurately predicting the class. Also, a close inspection of the tile transitions is performed to see if the use of information from the whole biogeographic region ensured continuity in the classification.

Class correctly predicted on the map

The first example corresponds to w*here "Open Maritime Pine" (PA=78) is equal to "Open Maritime Pine",* and it illustrates when the labels allow classifying the class correctly. In Figure 25, the COS 2018 polygon corresponding to the CODE 3121 "Florestas de Pinheiro bravo" was reclassified to Maritime Pine. According to the COS guidelines [7], the polygon contains 75% or more of the total area covered by forest. Wherein, it contains a regular network of service roads inside the polygon that can have the same probability of being selected as "Maritime Pine". Nevertheless, after the application of the pre-processing steps in section 3.3.2, the area was classified as "Open Maritime Pine Forest" because it contains more than 10% and less than 60% of coniferous tree cover according to the HRL. When including the samples for this class, neither the services roads nor logged areas are taken into consideration for feature extraction, reducing the possible misclassifications. As it is possible to visualize, the final classification coincides with the class; however, the service road network is classified as Baresoil and Urban in some areas, being Baresoil more appropriate.

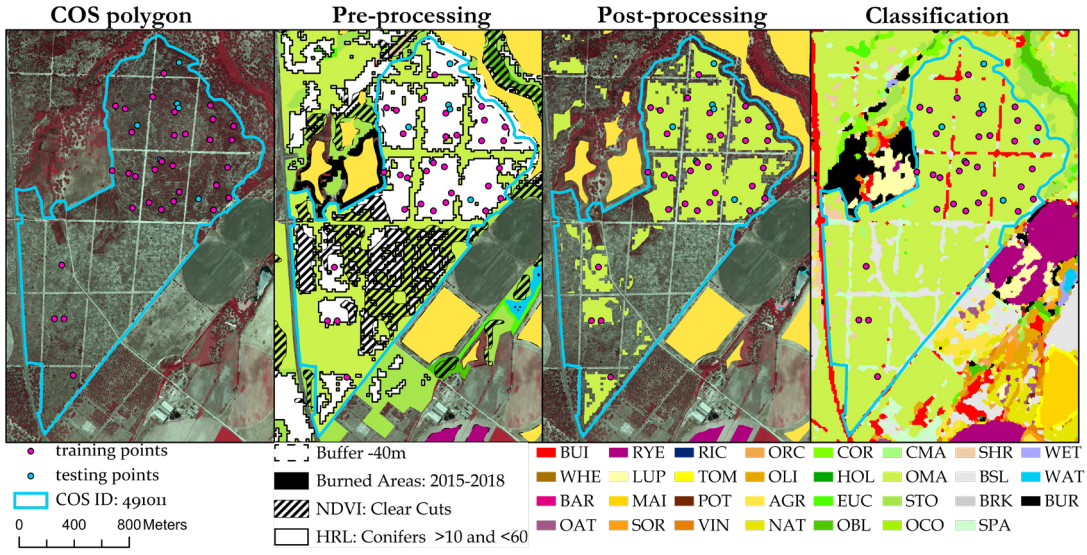


Figure 25 COS 2018 (OBJECTID: 491011) polygon pre and post-processed comparison to predictions for the class Open Maritime Pine Forest. Scale: 1:30.000

Next, it is possible to appreciate *Where "Holm Oak" (PA=94) is equal to "Holm Oak"*; the COS 2018 polygon is labeled as a pure forest of Holm Oak. The HRL mask allowed to allocate training and testing points only where the tree cover density was higher than 60%. However, when overlaying the IFAP 2018 dataset, as this dataset is prioritized over

COS, it ends up removing some Holm Oak areas in favor of agricultural grasslands. The final classification is a combination of Holm Oak and agricultural grasslands for the polygon.

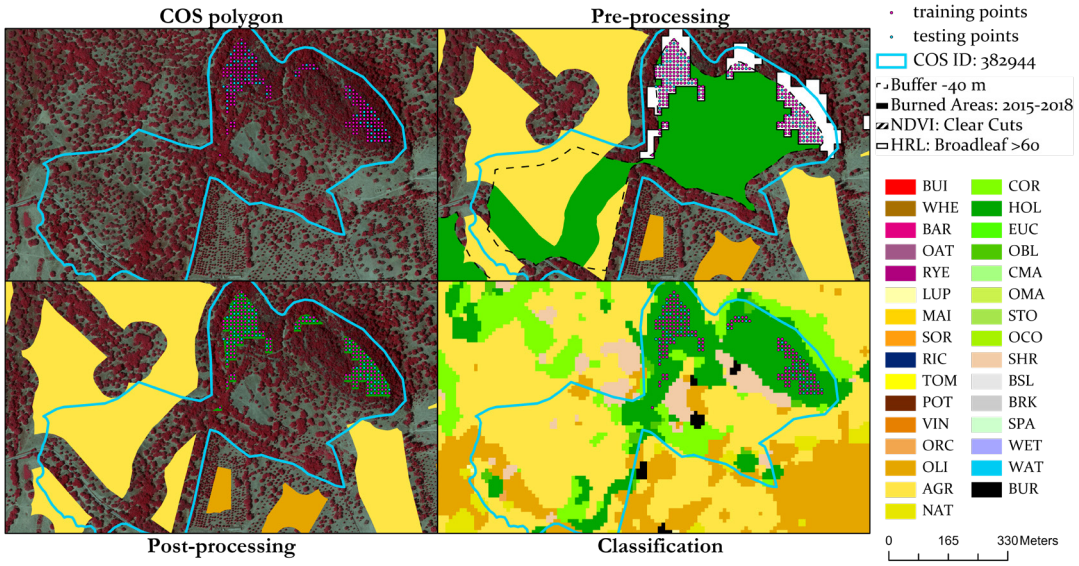


Figure 26 COS 2018 (OBJECTID: 382944) polygon pre and post-processed comparison to predictions for the class Holm Oak. Scale: 1:10.000

Class incorrectly predicted on the map

In the confusion matrix, one of the classes with the lowest accuracy is Orchards; in this example *Where "Orchards" (PA=35) is equal to "Natural grassland"* it is possible to appreciate a classification issue related to plantations. This class was assigned 201 times by commission error to natural grasslands. One of the difficulties in classifying this class is that it contains 17 types of trees ranging from citrus to almonds. Also, orchards are usually planted in 2m separation, meaning that the surface reflectance values captured correspond to a mixture of the crop and soil as the MMU of the Sentinel-2 is 10m. In the first square of Figure 27, it is possible to visualize that half of the polygon contains more vegetation intra rows at the soil level, this can correspond to creeping vegetation (i.e., close to the ground). The final classification dictated that the polygon is considered as olive trees and natural grassland; nothing was classified into the class they genuinely belong.

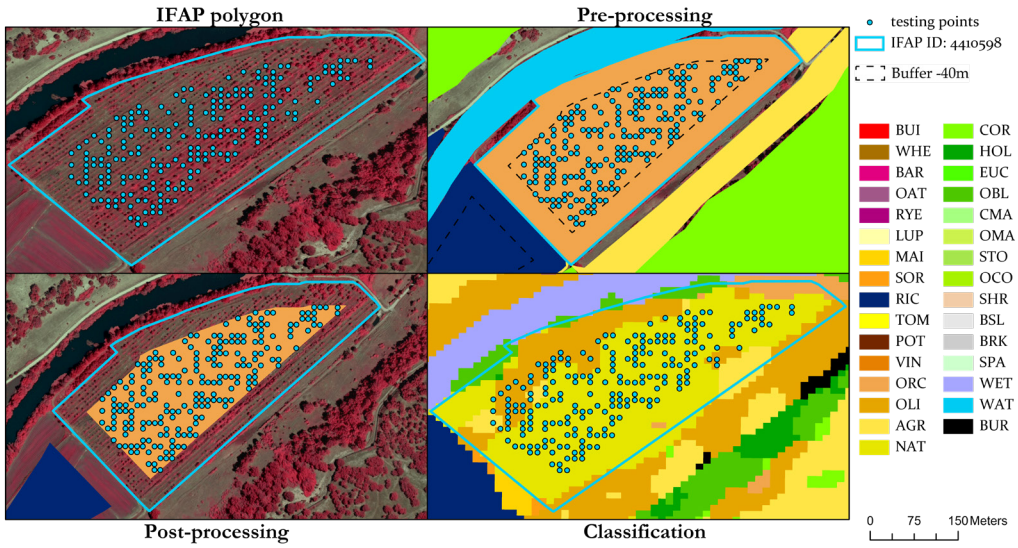


Figure 27 IFAP 2018 (OSAID: 4410598) polygon pre and post-processed comparison to predictions for the class Orchards. Scale: 1:6000

Tile transitions

The aim was to classify the biogeographic region corresponding to the strata 214 in Figure 2. Though this area was covered by two separate Sentinel-2 tiles, which could entrain discontinuities in their limits [3]. Yet, the approach was to automatically extract the samples for the whole study area and retrieving the features from both tiles. Hence, the classifier contained the information for the overall strata, allowing adjacent pixels in the borders to be assigned in the same class as it can be appreciated in Figure 28 that portrays the Land Cover and Crop Type map in three locations (a). The first location (b) corresponds to the Tejo estuary, preserving the continuity of the river and wetlands. The other two locations that are in the border of Santarém and Sétubal (c) and near Évora (d) kept the continuity of classes such as forests and rice fields along the tiles.

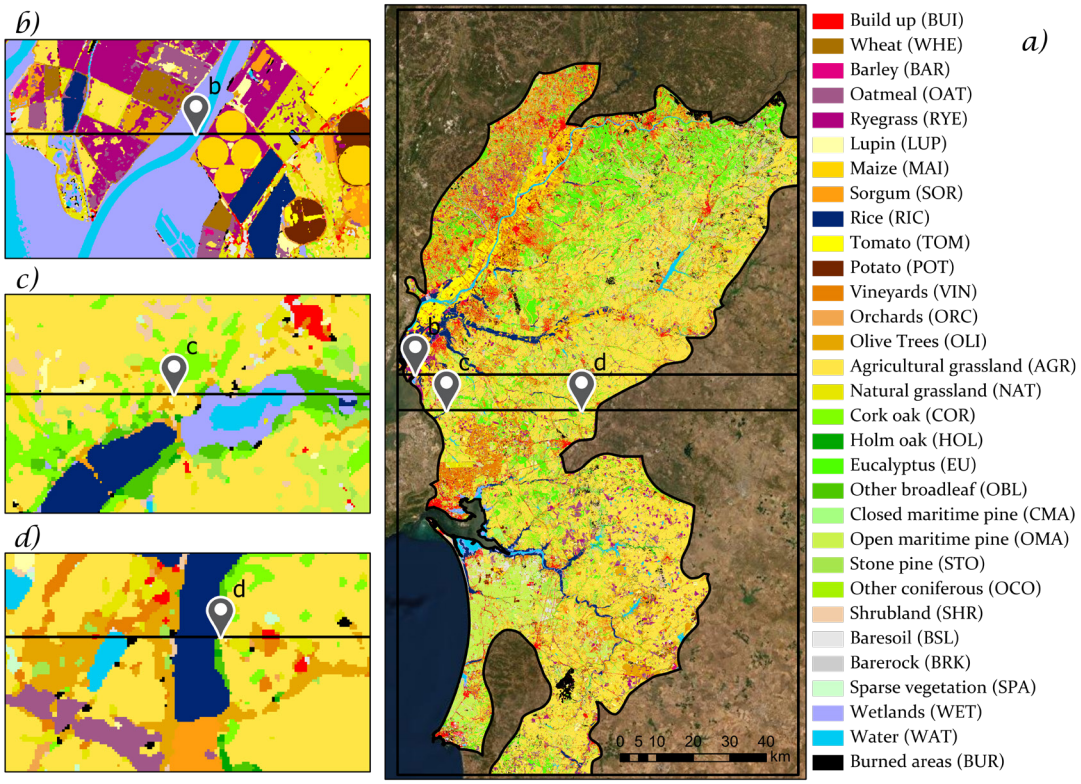


Figure 28 (a) Land Cover and Crop Type in raster format with three locations on the border of the Sentinel-2 tiles 29SND (upper) and 29SNC (lower), (b) Tejo estuary (c) border of Santarém and Sétubal, (d) location near Évora.

5 CONCLUSIONS

Up-to-date land cover and crop type information play an essential role in commercial and environmental monitoring and planning. For its updating, they have benefited from remote sensing imagery at a national, continental and global level. However, many challenges remain to produce accurate and timely land cover and crop type maps. This thesis focused on the use of intra-annual composites of Sentinel-2, supervised classification with random forest, and automatic sample extraction based on a pre-processing set of rules. The overall accuracy of 76% was achieved for 31 land cover and crop type classes.

The use of monthly composites of L2 Sentinel-2 data allowed having cloud-free data in contrast to single acquisitions that have missing values due to cloud cover or cloud shadows. Also, since the classification is done at the pixel level, having missing data would affect the spectral signature extraction and incompletely characterize the classes with missing data. Likewise, the composites represent an excellent opportunity for dimensionality reduction as the number of features would correspond to 10 bands per month. For single acquisitions, each acquisition would contain 10 bands, and the number of features would increase based on acquisitions during the period.

The Random Forest classifier required few hyperparameters to tune as opposed to other classifiers and proved to be computationally efficient as it was possible to parallelize it (multi-core processing) to classify the whole area. Also, it allowed extracting the most important features during the classification. As expected, the most relevant features from the time series correspond to the spring and summer months and the bands on the Red Edge (b5, b6, b7), NIR (b8a) and SWIR (b11 and b12). The inclusion of spectral indices slightly improved the accuracy but was not a predominant variable.

One of the purposes of this research was to test if a pre-defined set of rules could remove possible sources of misclassification, allowing us to extract samples for training and testing automatically. This would permit the classifier to adequately characterize the spectral signature for each class and make an accurate prediction. The data sources (IFAP 2018 and COS 2018) themselves are a product of visual interpretation of high-resolution imagery; in the case of the LPIS, the yearly update of the product and the MMU of a parcel allowed to characterize the types of crops. Nevertheless, the agricultural grassland class coverage was over-optimistic in this dataset and sometimes would mask out forest areas causing several mix-ups within classes.

Regarding the filtering rules, the application of the burned mask allowed to remove from the dataset the areas that ignited with wildfires. Though, in some cases, the burn mask includes build-up areas and water leading to confusion within classes. Therefore, the burn mask can benefit from a set of pre-processing rules before sample extraction, such as build-up, cannot be part of the burned mask and neither water. The following filter was the NDVI alerts; these were produced for the year 2015-2018 from Landsat 8 images at 30m resolution, containing an omission error of 33% [51]. This implicates that some changes are not detected. Also, the difference in pixel size between satellites reduces the precision in the detection of these areas; the same approach yet implementing Sentinel-2 imagery might improve the identification of clear-cuts for this study. Lastly, the HRL rules reduced the number of samples available per class dramatically. By removing many of the forest pixels, the spectral signature was not precisely characterized, and the model could not classify the whole area accurately. Forest in the Portuguese landscape is not as dense as trees are sparsely distributed in space; therefore, a decrease in the tree cover density is encouraged for detecting the forest types.

5.1 Limitations and Recommendations

The limitations of the study can be summarized in the requirement of an independent verification dataset and considerations for increasing accuracy; then, recommendations are provided for potential enhancement of the methodology.

For the IFAP 2018, the availability of an independent dataset ensured a proper verification of the classification for the agricultural classes. However, in the case of land cover, the unavailability of a verified testing dataset for COS2018 raised three main issues. First, the testing dataset underwent the same pre-processing as the training dataset; in consequence, many of the potential testing pixels were removed. Also, the training and testing polygons were not spatially disjoint, meaning that pixels coming from the same polygon can be used as training and testing. This can induce positive values in the accuracies as the spectral signature can be similar for both datasets [3]. The final concern of not having a validated dataset is that the model can correctly predict a class; however, if the class is incorrectly labeled, the class accuracy decreases. The traditional method for validation is visual inspection. However, this requires knowledge of the landscape to identify the different classes correctly.

In terms of increasing accuracy, this study can benefit from a reduction in the number of classes. If the focus is in cropland areas, a binary cropland mask [23] can allow

restricting the classification; this approach has been implemented in operational systems [3]. For land cover classes, the use of the most detailed level in the hierarchical nomenclature (i.e., level 4 in the nomenclature in section 7.3) is useful for the purpose of vegetation characterization for fire modeling. However, this detailed nomenclature adds noise to the classification. A reduction from 16 to the 11 classes (i.e., level 2 in the nomenclature in section 7.3) provides reliable information and can increase the classification accuracy. For the elaboration of the final map, in order to reduce the salt-and-pepper effect, it is possible the implementation object segmentation algorithms where pixels can be aggregated in homogeneous boundaries [17], [66] however this approach requires more complex analysis. At last, Random Forest has been used to classify hyperspectral datasets [39], demonstrating its capabilities to deal with an increasing number of dimensions. More features can be added to this model to improve the accuracy, some of them are the spectral, temporal metrics to describe the distribution of a spectral band or index over a specific period [67] and texture metrics [19].

6 BIBLIOGRAPHIC REFERENCES

[1] GAETANO, R. et al. *A Two-Branch CNN Architecture for Land Cover Classification of PAN and MS Imagery*. Remote Sensing, 2018, 10(11), 1746. Retrieved from: https://doi.org/10.3390/rs10111746

[2] LATHAM, J. et al. *Global Land Cover SHARE (GLC-SHARE) - database Beta-Release Version 1.0-2014*. FAO: Rome, 2014.

[3] INGLADA, J. et al. *Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. Remote Sensing*, 2017, 9(1), 95. Retrieved from: https://doi.org/10.3390/rs9010095

[4] WULDER, M. A. et al. *Land cover 2.0*. International Journal of Remote Sensing, 2018, 39(12), 4254–4284. Retreived from: https://doi.org/10.1080/01431161.2018.1452075.

[5] HERMOSILLA, T. et al. *Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series*. Canadian Journal of Remote Sensing, 2018, 44(1), 67–87. Retrieved from: https://doi.org/10.1080/07038992.2018.1437719

[6] DEFOURNY, P. *et al. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world*. Remote Sensing of Environment, 2019, 221, 551–568. Retrieved from: https://doi.org/10.1016/j.rse.2018.11.007

[7] DIREÇÃO-GERAL DO TERRITÓRIO (DGT). *Especificações técnicas da Carta de Uso e Ocupação do Solo (COS) de Portugal Continental para 2018. Relatório Técnico*. 2019.

[8] TURCO, M. et al. *Climate drivers of the 2017 devastating fires in Portugal*. Scientific Reports, 2019, 9, 13886. Retrieved from: https://doi.org/10.1038/s41598-019-50281-2

[9] WISNER, B.; ADAMS, J. *Environmental health in emergencies and disasters*. World Health Organization, 2002.

[10] BODROZIC, L.; MARASOVIC, J.; STIPANICEV, D. *Fire modeling in forest fire management*. CEEPUS Spring School, 2005.

[11] JAHDI, R. *et al. Evaluating fire modelling systems in recent wildfires of the Golestan National Park, Iran. Forestry: An International Journal of Forest Research,* 2016, 89(2), 136–149. Retrieved from: https://doi.org/10.1093/forestry/cpv045

[12] UNINOVA; NOVAIMS; DGT. *IPSTERS (IPSentinel Terrestrial Enhanced Recognition*

System). 2019. URL: https://www.ca3-uninova.org/project_ipsters [Accessed: 26-Jan-2020].

[13] UNINOVA; NOVA IMS; UA; ISA; DGT; IT. *FoRESTER (Data fusion of sensor networks and fire spread modelling for decision support in forest fire suppression)*. 2019. URL: https://www.ca3-uninova.org/project_forester [Accessed: 26-Jan-2020].

[14] LÜDTKE, D. *Land cover mapping with random forest using intra-annual sentinel 2 data in central Portugal: a comparative analysis*. NIMS - MSc Dissertations Geospatial Technologies (Erasmus-Mundus), 2018. URL: https://run.unl.pt/handle/10362/33648 [Accessed: 15-Feb-2020].

[15] BLANCO-MARTÍNEZ, W.A. *Intra-Annual land cover mapping: Automatic training sample extraction from old maps for intra-annual land cover mapping at central of Portugal*. NIMS - MSc Dissertations Geospatial Technologies (Erasmus-Mundus), 2019. URL: https://run.unl.pt/handle/10362/63946 [Accessed: 15-Feb-2020].

[16] PARIS, C.; BRUZZONE, L.; FERNÁNDEZ-PRIETO, D. *A Novel Approach to the Unsupervised Update of Land-Cover Maps by Classification of Time Series of Multispectral Images. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(7)*. 4259-4277. Retrieved from: https://doi.org/10.1109/TGRS.2018.2890404

[17] COSTA, H. et al. *Land Cover Mapping from Remotely Sensed and Auxiliary Data for Harmonized Official Statistics*. International Journal of Geo-Information, 2018, 7(4), 157. Retreived from: https://doi.org/10.3390/ijgi7040157.

[18] STEHMAN, S. V.; FOODY, G. M. *Key issues in rigorous accuracy assessment of land cover products*. Remote Sensing of Environment, 2019, 231, 111199. Retrieved from: https://doi.org/10.1016/j.rse.2019.05.018

[19] KHATAMI, R.; MOUNTRAKIS, G.; STEHMAN, S. V. *A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research*. Remote Sensing of Environment, 2016, 177, 89–100. Retrieved from: https://doi.org/10.1016/j.rse.2016.02.028

[20] FRITZ, S. *et al. A comparison of global agricultural monitoring systems and current gaps*. Agricultural Systems, 2019, 168, 258–272. Retrieved from: https://doi.org/10.1016/j.agsy.2018.05.010

[21] EUROPEAN SPACE AGENCY (ESA). *Sen2-Agri*. 2018. URL: http://www.esa-sen2agri.org/operational-system/system-description/ [Accessed: 08-Sep-2019].

[22] INGLADA, J. *et al. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery*. Remote Sensing,

2015, 7(9), 12356–12379. Retrieved from: https://doi.org/10.3390/rs70912356

[23] VALERO S. *et al. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions.* Remote Sensing, 2016, 8(1), 55. Retrieved from: https://doi.org/10.3390/rs8010055

[24] GRIFFITHS, P.; NENDEL, C.; HOSTERT, P. *Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping.* Remote Sensing of Environment, 2019, 220, 135–151. Retrieved from: https://doi.org/10.1016/j.rse.2018.10.031

[25] EUROPEAN COURT OF AUDITORS (ECA). *The Land Parcel Identification System A useful tool to determine the eligibility of agricultural land – but its management could be further improved.* Special report no. 25, 2016. URL: *https://www.eca.europa.eu/en/Pages/DocItem.aspx?did=38180* [Accessed: 08-Sep-2019].

[26] SCHMEDTMANN, J.; CAMPAGNOLO, M. L. *Reliable crop identification with satellite imagery in the context of Common Agriculture Policy subsidy control.* Remote Sensing, 2015, 7(7), 9325–9346. Retrieved from: https://doi.org/10.3390/rs70709325

[27] GÓMEZ, C.; WHITE, J. C.; WULDER, M. A. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, 116, 55–72. Retrieved from: https://doi.org/10.1016/j.isprsjprs.2016.03.008

[28] SONG, Q. *et al. In-Season Crop Mapping with GF-1/WFV Databy Combining Object-Based Image Analysisand Random Forest.* Remote Sensing, 2017, 9(11), 1184. Retreived from: https://doi.org/10.3390/rs9111184.

[29] ZHU, Z.; WOODCOCK, C. E. Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment, 2014, 144, 152–171. Retreived from: https://doi.org/10.1016/j.rse.2014.01.011.

[30] BAETENS, L.; DESJARDINS, C.; HAGOLLE, O. *Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure.* Remote Sensing, 2019 11(4), 433. Retrieved from: https://doi.org/10.3390/rs11040433

[31] HERMOSILLA, T. et al. *An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites.* Remote Sensing of Environment, 2015, 158, 220–234. Retrieved from: https://doi.org/10.1016/j.rse.2014.11.005

[32] WHITE, J. C. *et al.* Pixel-Based Image Compositing for Large-Area Dense Time

Series Applications and Science. Canadian Journal of Remote Sensing, 2014, 40(3), 192-212. Retrieved from: https://doi.org/10.1080/07038992.2014.945827

[33] JAMES, G. et al. *An Introduction to Statistical Learning with Applications in R*, vol. 11, no. 4. 2013. Retreived from: 10.1007/978-1-4614-7138-7.

[34] MAXWELL, A. E.; WARNER, T. A.; FANG, F. *Implementation of machine-learning classification in remote sensing: an applied review*. International Journal of Remote Sensing, 2018, 39(9), 2784–2817. Retrieved from: https://doi.org/10.1080/01431161.2018.1433343

[35] BOUTABA, R. et al. *A comprehensive survey on machine learning for networking: evolution, applications and research opportunities*. Journal of Internet Services and Applications, 2018, 9, 16. Retrieved from: https://doi.org/10.1186/s13174-018-0087-2

[36] SAMMUT, C.; WEBB, G. I. (Eds.), *Encyclopedia of Machine Learning*. MA: Springer US. Boston, USA, 2010.

[37] BREIMAN, L. *Random Forests*. Mach. Learn. 45, 1 (October 2001), 5–32. Retreived from: https://doi.org/10.1023/A:1010933404324.

[38] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer Science + Business Media: New York, USA, 2009.

[39] BELGIU, M.; DRĂGU, L. *Random forest in remote sensing: A review of applications and future directions*. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114, 24–31. Retrieved from: https://doi.org/10.1016/j.isprsjprs.2016.01.011

[40] COSTA, H. *Estratégia multi-temporal para produção automática de cartografia de ocupação do solo com imagens AWiFS*. NIMS - Dissertações de Mestrado em Ciência e Sistemas de Informação Geográfica, 2009. URL: https://run.unl.pt/handle/10362/2347 [Accessed: 10-Feb-2020]

[41] INSTITUTO DA CONSERVAÇÃO DA NATUREZA E DAS FLORESTAS (ICNF). *Areas ardidas*. 2018. URL: http://www2.icnf.pt/portal/florestas/dfci/inc/mapas [Accessed: 12-Feb-2020].

[42] MARCELINO, F.; GIRÃO, I.; CAETANO, M. *Séries multitemporais dos Temas de Grande Resolução (HRLs) do programa Copernicus para Portugal Continental 2006-2015*. iGEO - Informação Geográfica, 2018. URL: http://www.igeo.pt/DadosAbertos/Docs/Relatorio-HRL-2015-PT.pdf [Accessed: 12-Feb-2020].

[43] EUROPEAN SPACE AGENCY (ESA). *User Guides - Sentinel-2 MSI - Revisit and*

Coverage - Sentinel Online. 2020. URL: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/revisit-coverage [Accessed: 26-Jan-2020].

[44] OLIPHANT, T. *Guide to NumPy*. 2006.

[45] MCKINNEY, W. D*ata Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 2010. URL: https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf [Accessed: 15-Feb-2020].

[46] HUNTER, J. D. *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 2007, 9(3), 99–104.

[47] PEDREGOSA, F. et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011, 12, 2825-2830.

[48] GDAL/OGR-CONTRIBUTORS. "GDAL/OGR Geospatial Data Abstraction software Library." Open Source Geospatial Foundation, 2019.

[49] MIRRA, I. M. et al. *Fuel dynamics following fire hazard reduction treatments in blue gum (Eucalyptus globulus) plantations in Portugal*. Forest Ecology and Management, 2017, 398, 185-195. Retrieved from: https://doi.org/10.1016/j.foreco.2017.05.016

[50] MANCINO, G. et al. *Landsat TM imagery and NDVI differencing to detect vegetation change: Assessing natural forest expansion in Basilicata, southern Italy*. iForest - Biogeosciences and Forestry, 2014, 7, 76-85. Retreived from: 10.3832/ifor0909-007.

[51] COSTA, H.; BENEVIDES, P.; MARCELINO, F.; CAETANO, M. *Introducing automatic satellite image processing into land cover mapping by photo-interpretation of airborne data*. Unpublished work. ISRSE-38, Oct 2019, Baltimore, Maryland, USA, 2019.

[52] BENEVIDES, P.; COSTA, H.; CAETANO, M. *Multi-temporal Sentinel-2 composite technical specifications*. 2019.

[53] TUCKER, C. J. *Red and photographic infrared linear combinations for monitoring vegetation*. Remote Sensing of Environment, 1979, 8(2), 127–150. Retrieved from: https://doi.org/10.1016/0034-4257(79)90013-0

[54] ZHA, Y.; GAO, J.; NI, S. *Use of normalized difference built-up index in automatically mapping urban areas from TM imagery*. International Journal of Remote Sensing, 24(3), 583-594. Retrieved from: https://doi.org/10.1080/01431160304987

[55] MCFEETERS, S. K. *The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features*. International Journal of Remote Sensing, 1996 17(7), 1425-1432. Retrieved from: https://doi.org/10.1080/01431169608948714

[56] MCFEETERS, S. K. *Using the normalized difference water index (ndwi) within a geographic information system to detect swimming pools for mosquito abatement: A practical approach.* Remote Sensing, 2013, 5(7), 3544–3561. Retrieved from: https://doi.org/10.3390/rs5073544

[57] KEY, C. H.; BENSON, N. C. *Landscape Assessment (LA) Sampling and Analysis Methods.* In: Lutes, D.C. et al. Eds.; *FIREMON: Fire effects monitoring and inventory system.* : U.S.Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, USA, 2006; LA-1-55, 164.

[58] HISLOP, S. et al. *Using Landsat Spectral Indices in Time-Series to Assess Wildfire Disturbance and Recovery.* Remote Sensing, 2018, 10(3), 460. Retrieved from: https://doi.org/10.3390/rs10030460

[59] KOHAVI, R. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* International Joint Conference on Artificial Intelligence, 1995. URL: http://ai.stanford.edu/~ronnyk/accEst.pdf [Accessed: 15-Feb-2020].

[60] LIU, C.; FRAZIER, P.; KUMAR, L. *Comparative assessment of the measures of thematic classification accuracy.* Remote Sensing of Environment, 2007, 107(4), 606–616. Retrieved from: https://doi.org/10.1016/j.rse.2006.10.010

[61] PONTIUS, R. G.; MILLONES, M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 2011, 32(15), 4407-4429. Retrieved from: https://doi.org/10.1080/01431161.2011.552923

[62] FOODY, G. M. *Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification.* Remote Sensing of Environment, 2020, 239, 111630. Retrieved from: https://doi.org/10.1016/j.rse.2019.111630

[63] CARVALHO, A. et al. C*limate change research and policy in Portugal.* WIREs Climate Change, 2014, 5, 199-217. Retrieved from: https://doi.org/10.1002/wcc.258

[64] SUESS S. et al. *Characterizing 32 years of shrub cover dynamics in southern Portugal using annual Landsat composites and machine learning regression modeling.* Remote Sensing of Environment, 2018, 219, 353–364. Retrieved from: https://doi.org/10.1016/j.rse.2018.10.004

[65] NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA). *Measuring Vegetation (NDVI and EVI).* URL: https://earthobservatory.nasa.gov/features/MeasuringVegetation. [Accessed: 16-

Fec-2020].

[66] BELGIU, M.; CSILLIK, O. *Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis*. Remote Sensing of Environment, 204, 509–523. Retrieved from: https://doi.org/10.1016/j.rse.2017.10.005

[67] PFLUGMACHER, D. et al. *Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey*. Remote Sensing of Environment, 2019, 221, 583–595. Retrieved from: https://doi.org/10.1016/j.rse.2018.12.001

7 ANNEXES

7.1 External scripts

Spectral feature extraction script adapted from: https://github.com/jdbfsilva/rs-util

7.2 Land Cover and Crop Type nomenclature

CODE	CLASS	Description	Reclassification
1.1.1.1	Build up	this class includes all the artificial or landscaped surfaces intended for activities related to human societies such as urban fabric, road network, and associated spaces.	COS 2018 (1111, 1112, 1221)
2.1.1.1	Wheat	agricultural class corresponding to a temporary rainfed crop, this cereal grows during the autumn and winter.	IFAP 2018 (001)
2.1.1.2	Barley	rainfed temporary cereal.	IFAP 2018 (004)
2.1.1.3	Oatmeal	rainfed temporary cereal.	IFAP 2018 (005)
2.1.1.4	Ryegrass	rainfed temporary cereal (forage).	IFAP 2018 (067)
2.1.1.5	Lupin	rainfed temporary pulses (nitrogen fixer).	IFAP 2018 (240)
2.1.2.1	Maize	agricultural class corresponding to an irrigated temporary crop, this cereal grows during the spring and summer.	IFAP 2018 (006)
2.1.2.2	Sorghum	irrigated temporary cereal.	IFAP 2018 (008)
2.1.2.3	Rice	irrigated temporary cereal.	IFAP 2018 (024)
2.1.2.4	Tomato	irrigated temporary vegetable.	IFAP 2018 (033)
2.1.2.5	Potato	irrigated temporary vegetable.	IFAP 2018 (103)
2.2.1.1	Vineyards	areas where vineyards are dominant over other types of permanent crops such as orchards or olive trees.	IFAP 2018 (034
2.2.2.1	Orchards	cultivated plots with trees intended for fruit production, this class combines 17 types of trees from figs and oranges to walnuts and hazelnuts.	IFAP 2018 (085, 093, 094, 096, 097, 105, 107, 108, 109, 112, 116, 118, 119, 157, 208, 209, 211)
2.2.3.1	Olive Trees	areas with olive tree plantations (*Olea europea var. europea*) for olive production.	IFAP 2018 (083);

3.1.1.1	Agricultural grassland	areas permanently occupied with cultivated herbaceous vegetation.	IFAP 2018 (143);
3.1.2.1	Natural grassland	areas with 25% or more of the surface occupied by herbaceous vegetations growing without fertilization, cultivation, sowing, or drainage.	COS 2018 (321);
5.1.1.1	Cork oak forest	Agroforestry Systems or pure forest of Cork oak (*Quercus suber*).	COS 2018 (2441, 3111);
5.1.2.1	Holm oak forest	Agroforestry Systems or pure forest of Holm oak (*Quercus rotundifolia*).	COS 2018 (2442, 3112);
5.1.3.1	Eucalyptus forest	Broadleaf forest where the angiosperm trees represent 75% or more of the forest cover.	COS 2018 (3115);
5.1.4.1	Other broadleaf forest	Agroforestry Systems or pure forests of oak species other than cork oak and holm oak. These include chestnut trees (*Castanea sativa*), walnut trees (*Juglans regia*), and forests of invasive species.	COS 2018 (2443, 3113, 3114, 3116, 3117);
5.2.1.1	Closed Maritime pine forest	Coniferous forest where the gymnosperm species represent 75% or more of the forest cover.	COS 2018 (3121);
5.2.1.2	Open maritime pine forest	this class is derived from class 5.2.1.1 after the crossing with the High-Resolution Layers (HRL) process described in section 3.2.1.	COS 2018 (3121);
5.2.2.1	Stone pine forest	Agroforestry Systems or pure forest of Pine (*Pinus pinea*).	COS 2018 (2444, 3122);
5.2.3.1	Other coniferous forest	pure forests of other coniferous species not included in the previous classes. (e.g., *Pinus sylvestris*, *Larix spp.*, *Cryptomeria japonica*.).	COS 2018 (3123);
6.1.1.1	Shrubland	natural areas of spontaneous vegetation, little or very dense where shrub cover is 25% or more.	areas that remained shrubland from COS 1990 to COS 2015;
7.1.1.1	Baresoil	areas of open-air mineral extraction, sand exploitation areas, banks of rivers, and coastal sands, including ante-dune vegetal formations.	COS 2018 (1311, 1312, 3311, 3312);
7.1.2.1	Bare Rock	areas where the surface covered by rock is higher than 90%, also included areas of abandoned mineral extraction.	COS 2018 (332);
7.1.3.1	Sparse vegetation	areas where the herbaceous vegetation is between 10% and 25% only.	COS 2018 (333);
8.1.1.1	Wetlands	lowlands flooded in winter, less saturated with water all year round or shore areas submerged during high tide at some point in the cycle of the annual sea.	COS 2018 (411, 421);

9.1.1.1	Water	natural and artificial freshwater surfaces, oceans and surfaces, and coastal lagoons and river mouths.	COS 2018 (5111, 5121, 5122, 5123, 5124, 5125, 521, 522);
9999.	Burned Areas	areas that burned in 2018 and detected by the ICNF.	ICNF (2018)

7.3 RGB color ramp for the Land Cover and Crop Type Classes

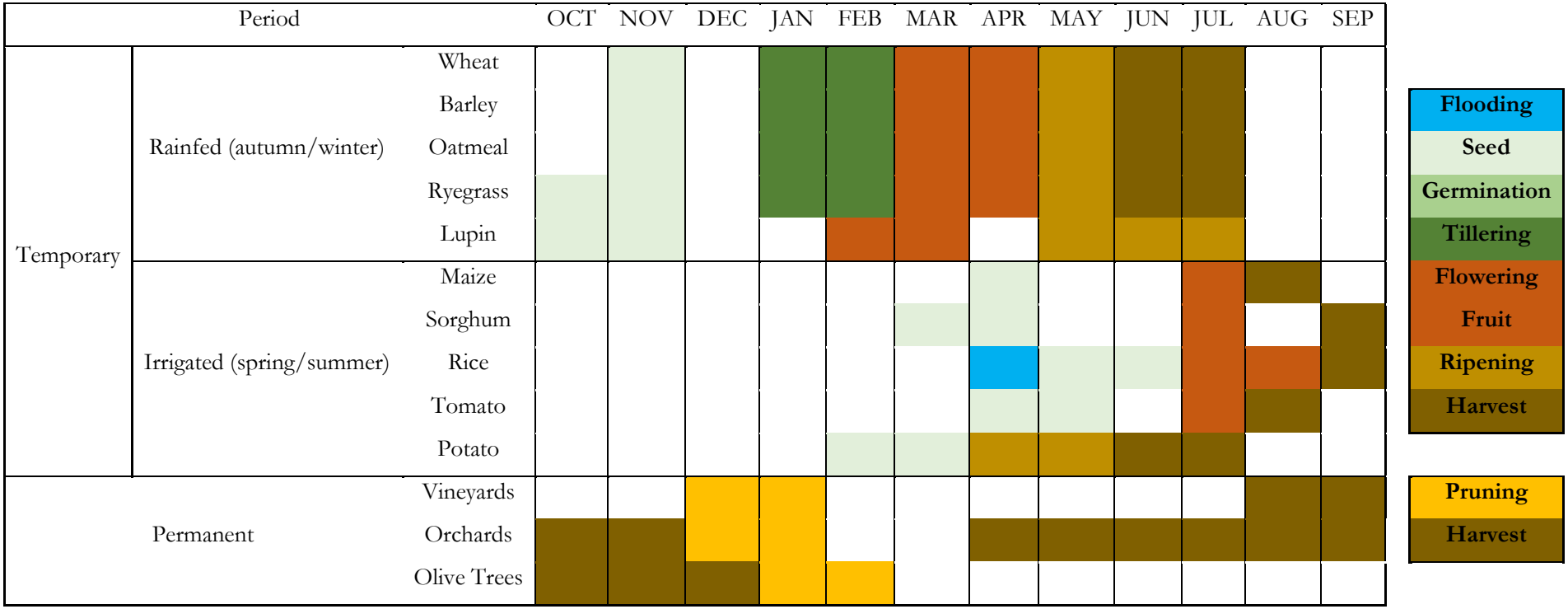
The following color ramp is a combination of the CLC 2018 RGB that can be found in the European Environment Agency (EEA) website and the CropScape RGB available on the website of the United States Department of Agriculture - National Agricultural Statistics Service (USDA-NASS).

LV1	LV2	LV3	LV4	RGB	CODE
1.Build up (BUI)	**1.1 Build up (BUI)**	1.1.1 Build up (BUI)	1.1.1.1 Build up (BUI)	255-000-00	1111
2.Agriculture (AGR)	**2.1 Temporary crops (TCO)**	2.1.1 Rainfed temporary crops (RAI)	2.1.1.1 Wheat (WHE)	168-112-0	2111
			2.1.1.2 Barley (BAR)	226-0-127	2112
			2.1.1.3 Oatmeal (OAT)	161-88-137	2113
			2.1.1.4 Ryegrass (RYE)	174-1-126	2114
			2.1.1.5 Lupin (LUP)	255-255-168	2115
		2.1.2 Irrigated temporary crops (IRR)	2.1.2.1 Maize (MAI)	255-212-0	2121
			2.1.2.2 Sorghum (SOR)	255-158-15	2122
			2.1.2.3 Rice (RIC)	0-38-115	2123
			2.1.2.4 Tomato (TOM)	255-255-0	2124
			2.1.2.5 Potato (POT)	115-38-0	2125
	2.2 Permanent crops (PCO)	2.2.1 Vineyards (VIN)	2.2.1.1 Vineyards (VIN)	230-128-000	2211
		2.2.2 Orchards (ORC)	2.2.2.1 Orchards (ORC)	242-166-077	2221
		2.2.3 Olive Trees (OLI)	2.2.3.1 Olive Trees (OLI)	230-166-000	2231
3.Grassland (GRA)	**3.1 Grassland (GRA)**	3.1.1 Agricultural grassland (AGR)	3.1.1.1 Agricultural grassland (AGR)	255-230-077	3111
		3.1.2 Natural grassland (NAT)	3.1.2.1 Natural grassland (NAT)	230-230-000	3121

5.Forests (FOR)	**5.1 Broadleaf forest (BOF)**	5.1.1 Cork oak forest (COR)	5.1.1.1 Cork oak forest (COR)	128-255-000	5111
		5.1.2 Holm oak forest (HOL)	5.1.2.1 Holm oak forest (HOL)	000-166-000	5121
		5.1.3 Eucalyptus forest (EUC)	5.1.3.1 Eucalyptus forest (EUC)	077-255-000	5131
		5.1.4 Other broadleaf forest (OBL)	5.1.4.1 Other broadleaf forest (OBL)	077-200-0	5141
	5.2 Coniferous forest (COF)	5.2.1 Maritime pine forest (MAR)	5.2.1.1 Closed Maritime pine forest (CMA)	166-255-128	5211
			5.2.1.2 Open maritime pine forest (OMA)	204-242-077	5212
		5.2.2 Stone pine forest (STO)	5.2.2.1 Stone pine forest (STO)	166-230-077	5221
		5.2.3 Other coniferous forest (OCO)	5.2.3.1 Other coniferous forest (OCO)	166-242-000	5231
6. Shrubland (SHR)	**6.1 Shrubland (SHR)**	6.1.1 Shrubland (SHR)	6.1.1.1 Shrubland (SHR)	242-204-166	6111
7. Open spaces with little or no vegetation (OPE)	**7.1 Open spaces with little or no vegetation (OPE)**	7.1.1 Baresoil (BSL)	7.1.1.1 Baresoil (BSL)	230-230-230	7111
		7.1.2 Bare Rock (BRK)	7.1.2.1 Bare Rock (BRK)	204-204-204	7121
		7.1.3 Sparse vegetation (SPA)	7.1.3.1 Sparse vegetation (SPA)	204-255-204	7131
8.Wetlands (WET)	**8.1 Wetlands (WET)**	8.1.1 Wetlands (WET)	8.1.1.1 Wetlands (WET)	166-166-255	8111
9. Water (WAT)	**9.1 Water (WAT)**	9.1.1 Water (WAT)	9.1.1.1 Water (WAT)	000-204-242	9111
9. Burned areas (BUR)	**9.9 Burned areas (BUR)**	9.9.9 Burned areas (BUR)	9.9.9.9 Burned areas (BUR)	000-000-000	9999

7.4 Crop Calendar

This corresponds to the crop calendar for the monitored IFAP parcels, each calendar is defined based on the Portuguese agricultural cycle (October to September). Note: For orchards, it corresponds to an average of 17 different types of trees.



7.5 Confusion matrix

LV4	BUI	WHE	BAR	OAT	RYE	LUP	MAI	SOR	RIC	TOM	POT	VIN	ORC	OLI	AGR	NAT	COR	HOL	EUC	OBL	CMA	OMA	STO	OCO	SHR	BSL	BRK	SPA	WET	WAT	BUR
BUI	880	1	0	0	0	0	0	0	0	0	0	15	6	9	7	18	1	0	0	2	0	2	0	0	7	25	0	5	2	0	20
WHE	1	667	10	223	10	15	0	0	0	0	3	0	1	1	6	13	0	0	0	0	0	0	0	0	0	1	0	26	0	0	23
BAR	0	226	464	0	100	0	0	27	0	0	0	1	0	3	1	7	0	0	0	0	0	0	0	0	21	0	0	6	0	0	0
OAT	0	47	27	492	65	129	0	0	1	0	0	6	0	16	207	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
RYE	0	2	0	80	619	75	0	68	0	0	3	3	10	3	32	26	0	0	0	0	0	0	0	0	0	0	0	2	0	0	77
LUP	4	0	0	122	198	552	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16
MAI	0	0	0	0	2	0	990	1	0	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SOR	0	0	2	18	8	6	0	637	2	36	0	126	16	73	46	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5
RIC	0	0	0	0	0	0	10	5	985	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOM	0	0	0	0	0	0	0	2	0	998	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
POT	5	0	28	0	0	4	0	0	0	31	845	1	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58
VIN	8	0	0	0	0	0	0	0	0	0	5	944	4	3	0	12	0	0	0	0	0	0	0	0	1	17	0	1	0	0	5
ORC	0	0	0	1	25	0	0	1	0	1	59	43	350	48	42	201	11	0	0	171	0	0	0	0	1	0	0	0	45	0	1
OLI	0	0	11	18	20	24	0	6	0	0	54	33	21	460	174	89	0	3	1	1	0	4	3	0	23	7	0	1	0	0	47
AGR	2	3	0	13	11	34	0	6	0	0	0	14	3	41	720	37	45	1	2	5	0	15	15	0	15	3	0	3	1	0	11
NAT	14	1	3	13	20	16	1	2	0	0	0	18	5	26	75	679	7	1	2	11	0	6	0	0	65	17	0	3	4	0	11
COR	0	0	0	0	0	2	0	0	0	0	0	10	4	3	19	2	700	37	32	92	30	18	10	0	19	1	0	0	2	6	13
HOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	29	1063	3	28	1	1	1	0	3	0	0	0	1	0	1
EUC	1	0	0	0	0	1	0	0	0	0	0	1	2	3	2	0	78	10	821	28	22	15	6	0	8	2	0	0	0	0	0
OBL	0	0	0	0	0	0	0	0	1	0	0	0	2	1	4	3	93	31	11	808	12	1	11	0	13	0	0	0	7	0	2
CMA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	58	5	28	19	751	69	60	0	3	0	0	0	1	1	2
OMA	6	0	0	0	0	2	0	0	0	0	0	0	0	2	12	0	16	3	5	2	113	781	25	0	21	5	0	5	0	0	2
STO	0	0	0	0	0	2	0	0	0	0	0	0	0	3	3	0	16	8	12	20	64	24	843	0	3	0	0	0	1	0	1

OCO	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	4	9	1	20	50	0	0	0	0	0	0	0
SHR	23	1	0	4	2	4	0	0	0	0	0	9	8	16	47	42	18	3	2	28	3	25	1	0	634	69	0	40	1	2	18
BSL	36	0	0	1	0	3	0	1	0	0	1	11	7	7	14	14	4	0	0	4	0	14	0	0	22	804	2	27	5	9	14
BRK	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	60	4	0	0	0
SPA	11	0	0	0	0	0	0	0	0	1	0	1	0	0	8	3	5	0	0	1	0	5	1	0	12	8	1	942	0	0	1
WET	7	0	0	0	0	1	0	0	0	0	0	2	1	2	4	7	3	1	1	10	0	0	0	0	4	2	0	0	944	11	0
WAT	1	0	0	1	0	1	0	0	0	0	0	0	0	0	3	6	0	1	0	0	0	0	0	0	2	11	0	1	24	948	1
BUR	17	0	0	6	7	13	0	3	0	0	0	8	9	37	54	35	10	0	7	8	1	14	4	0	24	7	0	4	7	2	723
LV4	BUI	WHE	BAR	OAT	RYE	LUP	MAI	SOR	RIC	TOM	POT	VIN	ORC	OLI	AGR	NAT	COR	HOL	EUC	OBL	CMA	OMA	STO	OCO	SHR	BSL	BRK	SPA	WET	WAT	BUR

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS