Journal of Big Data

METHODOLOGY

Open Access

Scalable approach for high-resolution land cover: a case study in the Mediterranean Basin

Antonio Manuel Burgueño[1], José F. Aldana-Martín[1], María Vázquez-Pendón[1], Cristóbal Barba-González[1], Yaiza Jiménez Gómez[2], Virginia García Millán[2] and Ismael Navas-Delgado[1*]

*Correspondence: ismael@uma.es

[1] KHAOS, ITIS Software, Universidad de Málaga Spain, 29071 Málaga, Spain
[2] ETC-UMA, Universidad de Málaga Spain, 29071 Málaga, Spain

Abstract

The production of land cover maps is an everyday use of image classification applications on remote sensing. However, managing Earth observation satellite data for a large region of interest is challenging in the task of creating land cover maps. Since satellite imagery is getting more precise and extensive, Big Data techniques are becoming essential to handle the rising quantity of data. Furthermore, given the complexity of managing and analysing the data, defining a methodology that reduces the complexity of the process into different smaller steps is vital to data processing. This paper presents a Big Data methodology for creating land cover maps employing artificial intelligence algorithms. Machine Learning algorithms are contemplated for remote sensing and geodata classification, supported by explainable artificial intelligence. Furthermore, the process considers aspects related to downloading data from different satellites, Copernicus and ASTER, executing the pre-processing and processing of the data in a distributed environment, and depicting the visualisation of the result. The methodology is validated in a test case for er map of the Mediterranean Basin.

Keywords: Big Data, Land cover, Workflow, Explainable AI, Remote sensing, Multispectral, Machine learning

Introduction

Nowadays, Earth observation based on remote sensing is becoming more significant as it provides a solid technological foundation for creating cutting-edge applications across various fields such as climate change [1], precision agriculture [2], smart urbanism [3], soil degradation, and land cover changes [4]. Particularly, land degradation risk has grown significantly during the past few decades. The natural ecosystem and the socio-economic system are interconnected systems affected by land degradation [5, 6]. Monitoring land use fulfils a vital role in achieving several worldwide strategic objectives such as saving biodiversity [7], reducing carbon emissions and global warming [8, 9], urban planning [10] and agriculture [11]. In this sense, LifeWatch ERIC[1] is a European Research Infrastructure Consortium born in 2017 that offers e-Science research capabilities to researchers looking into the functions and services of biodiversity and

[1] https://www.lifewatch.eu/.

ecosystems to help society address major planetary concerns. This paper is developed through the environmental and biodiversity climate change lab (EnBiC2-Lab[2]) project (in the LifeWatch ERIC infrastructure ecosystem).

A meaningful way to describe the Earth's surface is through its land cover (LC). Various local, national, and international natural resources' management decisions require spatially detailed land cover data. The functional link between topography, climate, and soil is influenced by land cover, which also provides biophysical insights into the environment and change-causing factors [12]. It is hardly a stretch to suggest that the entire planet has gone digital; thus, LC mapping has become a *Big Data* issue [13]. Therefore, the amount of data necessary to manage is an ambitious task due to the significant growth of its volume [14]. The extensive data generated by remote sensing have several distinct and tangible properties, such as being multi-source, multi-scale, high-dimensional, dynamic-state, and non-linear [15]. Besides, satellite remote sensing has long been regarded as the optimal technique and data source for large-area land cover classifications [16].

Moreover, land cover on a large scale is challenging due to spectral heterogeneity and complexity of the terrain [17]. A suitable method for mapping vegetation on a global, regional, or local scale, periodically and repeatedly, is provided by the European *Copernicus* programme,[3] served by the *Sentinel satellite missions*.[4] Since 2013 *Sentinel-2*[5] continuously collects optical imagery and delivers, every 2 to 5 days, high spatial resolution (10–60 m) multispectral images for global monitoring data.

Other National Space Agencies, such as NASA (U.S. National Aerospace Agency), also contribute with missions and data useful for LC mapping. The USGS (U.S. Geological Survey) develops the Land Change Monitoring, Assessment, and Projection (LCMAP) program[6] focus on LC monitoring using Landsat mission data. In addition, Since 1999, NASA and JAXA (Japanese Aerospace Exploration Agency) launched the *Advanced Spaceborne Thermal Emission and Reflection Radiometer* (ASTER) satellite,[7] which provides multispectral satellite images with high spatial and spectral resolutions (15–90 ms) [18]. The satellite is loaded with two cameras with different angles of observation (nadir and 23 degrees), which allows the stereoscopic reconstruction of the Digital Surface Model of the Earth at a nominal pixel resolution of 25 ms.

Besides, private initiatives such as the *Google Cloud Storage* platform[8] allow the download of the data from Sentinel-2 [19]. All of these platforms have free and open services available to their users [20].

In the context of monitoring, mapping, and modelling activities, the satellites mentioned above are used for large-area, multiple-image-based, multiple-sensor land cover mapping [21].

[2] https://enbic2lab.uma.es/.

[3] https://www.copernicus.eu.

[4] https://sentinels.copernicus.eu/web/sentinel/home.

[5] https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2.

[6] (https://www.usgs.gov/special-topics/lcmap).

[7] https://asterweb.jpl.nasa.gov/gdem.asp.

[8] https://cloud.google.com/storage/docs/public-datasets.

Data collection and pre-processing, map legend generation, classification strategy, stratification, the inclusion of auxiliary data, and accuracy evaluation have all been highlighted as standard methodological processes, usually in the form of a *workflow* above all in the context of Big Data [22, 23]. Generally, methods to classify land cover over an enormous region must be reliable and reproducible. This situation poses new challenges beyond those encountered in large-area image classifications [24]. The extensive amount of unprocessed remote sensing data is accompanied by the "four Vs", or volume, variety, velocity, and veracity, which are referred to as the "four problems of Big Data" [25]. Specialised tools and techniques are needed to efficiently mine and extract useful information from such data as well as control its volume [26].

To find a helpful technique for mapping LC patterns, the remote sensing community has attempted to approach the issue of LC classification from various algorithmic technique angles [27]. Machine Learning (ML) algorithms are widely used for land cover classification using remote sensing data [28–31]. In fact, *Random Forest* algorithm [32] has been proved to perform well for LC mapping due to its capacity successfully handle high data dimensionality and multicollinearity, being both fast and insensitive to overfitting [33].

A particular ML technique known as *eXplainable artificial intelligence* (XAI), which allow human users to comprehend and trust the results and outputs of machine learning algorithms, has attracted increasing research interest [34]. In general, XAI lacks a widely accepted definition. Still, it has been shown that it explains essential elements like transparency, justification, and informativeness that are essential, particularly for social or therapeutic applications. ML models frequently turn into *black boxes*, making it difficult to deduce broad guidelines for input attributes and output scores [35]. According to this viewpoint, every method that aims to create an understandable model may come inside the XAI domain. XAI is used to interpret the RF method and analyze the impact of different LC patterns [36].

The challenge in this work is describing a methodology to facilitate land cover classification in large areas such as the Mediterranean Basin. Furthermore, a workflow is developed for implementing the methodology. The Mediterranean Basin is covered by more than 450 Sentinel-2 tiles and around 1200 ASTER tiles. Moreover, three seasons of Sentinel-2 data are used, together with several derived products from Sentinel-2 and ASTER, which multiplies the amount of computation needed to process them. Parts of the workflow have been distributed using a tool for paralleling algorithms, *Dask* [37], to solve this problem. The main contributions of this work are described next:

- Development of an Open Source workflow to support the analysis of classifying land cover in the context of Big Data.
- A versatile tool for generating maps, which can deliver different cartography depending on the input data (images and labelled data). Therefore, this work's results are transferable to other areas, years and thematic maps. The minimum attributes required for each point are latitude, longitude, and label.

- A controlled Big Data storage and management and metadata protocols. A distributed information *S3-compatible object storage* like *MinIO* [9] is used. In the case of metadata, they are stored in a NoSQL database, like *MongoDB* ,[10] seeking flexibility and velocity.

The rest of the paper is structured as follows. "Related work" section presents the most relevant related work to our proposal. "Materials and methods" section details the proposal workflow step by step. "Results and discussion" section presents the application of our proposal to the case study. Conclusions and future work are presented in "Conclusions" section.

Related work

At the European level, the *CORINE* (CLC) program is a standardised data collection and LC mapping to support environmental policy development since 1990. The development of the CLC maps is in the hands of each nation and region in the European Economic Area, following Copernicus standards on photo interpretation, using High Resolution (HR) images such as aerial and satellite images, and ground truth data for labelling. Since the last CLC version (CLC+,[11] 2018), semi-automatic approaches have been used in a few countries, utilising national in-situ data, Sentinel-2 image processing, and geographic information system integration (GIS). The limitations of the CORINE (at least the older versions) are the large amount of human power and time required, besides the bias on human interpretation of polygon size and labelling, which produces differences in the LC maps among regions and nations. By automating the LC map production, using a single source of data (Sentinel-2), it is expected that the CLC+ will be a homogeneous product across Europe by reducing the processing time and human power required.

In this line of thinking, at the global level, the Copernicus Global Land Service systematically monitors global bio-geophysical characteristics by the yearly production of the *Copernicus Global Land Cover Map* (GLC)) since 2015. More than 20 institutes are involved in processing low-to medium-resolution optical and radar data, data validation, data supply, and product quality control of the GLC [4, 38]. GLC is created through a processing workflow, which is not sensor-specific and can be applied to any satellite data. Initially, the GLC product was developed using PROBA-V sensor data [39], although the workflow has also been tested on Sentinel-2 and Landsat data.[12] However, the accuracy of the LC maps based on automatic techniques strongly depends on the quality and quantity of ground-truth data, which is heterogeneous across the planet at the moment. A technical report by the European Environmental Agency [40] evaluated the accuracy of the GLC and other cartography between the European, North African and Middle East areas of the Mediterranean. The conclusion was that map accuracy (particularly

[9] https://min.io/.

[10] https://www.mongodb.com/.

[11] https://land.copernicus.eu/pan-european/clc-plus.

[12] https://land.copernicus.eu/global/products/lc.

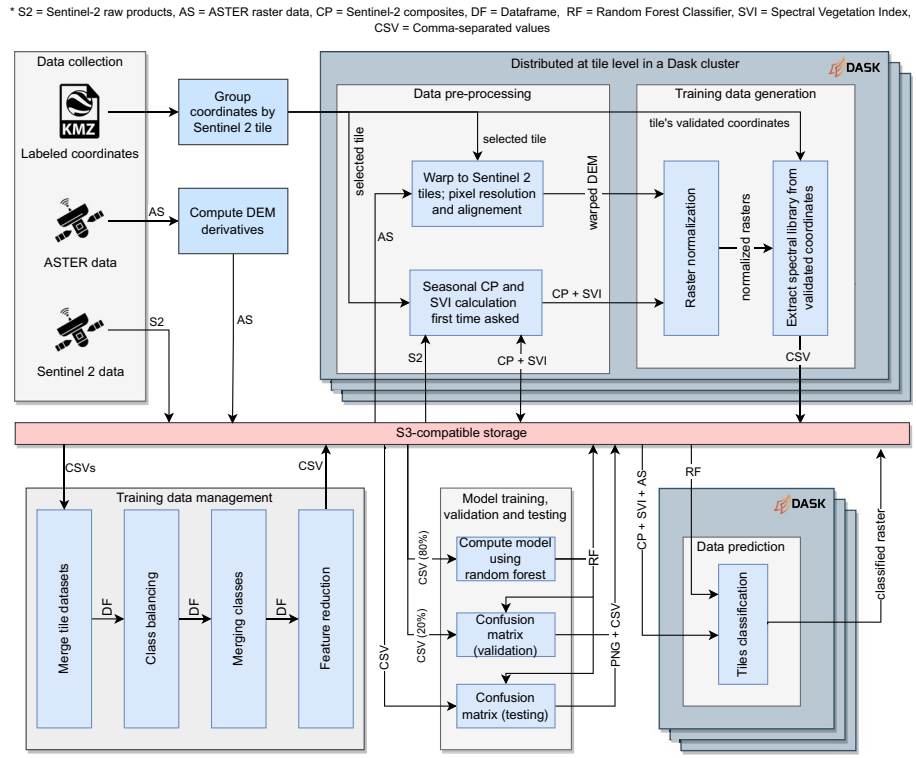


Fig. 1 High-level diagram of the proposed workflow. Processes inside grey background boxes are executed in a distributed Dask cluster

forest class accuracy) is low in non-European countries due to scarce ground-truth data on those regions.

Materials and methods

This section describes the methodology proposed to build a Land Cover map from satellite images utilising Machine Learning algorithms to train predictive models using a dataset composed of labelled coordinates, taking advantage of Big Data technologies to deal with the complexity of working with the data volume of satellite images for large areas.

Figure 1 shows a general view of the proposed methodology that has been converted into a functional software workflow for the automatic land cover classification. This methodology is divided into different steps described in detail in the following subsections.

The first part of the workflow is composed of methods related to data collection, such as downloading raster data and collecting the labelled geolocated information ("Data collection" subsection). Then, raster data has to be harmonised and cleaned using preprocessing data methods, such as removing clouds from the S2 images, denoising the ASTER DEM, generating S2 monthly composites, calculating spectral indices [41] and normalizing variables ("Data pre-processing" subsection). The third step aims at preparing the data to be used in the ML algorithms. Thus, this step transforms the raster data in combination with the ground-truth point dataset into tabular datasets (a spectral

library) for training the ML algorithms ("Training data generation" subsection). The training (ground truth) dataset has to be as well pre-processed to adapt the raw dataset into the LC classes of the map and to balance the classes statistically, based on representativeness and overfitting ("Training data management" subsection). Having a spectral library (obtained from the intersection of the image data and the ground-truth data) prepared through the designed workflow, a ML model is ready to be trained and validated ("Model training and validation" subsection) and XAI is used to analyse the ML model ("Explainability and feature reduction" subsection). Finally, satellite images are classified using the trained model by iterating the model through all tiles ("Data prediction" subsection).

Data collection

The first step in the data collection stage is to collect the labelled geolocated data for training the ML algorithms. The inputs required by the designed algorithms must be based on point coordinates. Thus, the data collected consists of a set of coordinates and a label for each one. This input may be provided as a KMZ or GeoJSON file. For the case of the Land Cover Map of the Mediterranean region, the training data consists of ground-truth data of the LC classes. The data comes from different sources, such as LC maps or National LC inventories of several countries, such as Lebanon, Spain[13] and Tunisia [42]. Besides, it has been collected data from eleven other countries through the photo interpretation of high-resolution images from Google Earth in 2021. The data can be a wall-to-wall LC map in polygon shapefile format or tabular point data. As sometimes the ground-truth data are polygons, an algorithm called *polylabel* [43] has been used for parsing polygons to points. The points computed by this method are the most distant internal coordinate from the polygon outline, which is always a coordinate contained in the region. Another alternative would be the centroid, but it has not been considered as it may be located outside the polygon region (which can belong to another class).

The definitions of the different ground-truth data sources were harmonised to the 9 main LC classes of Copernicus Global Land Cover [44] from 2019[14] (excluding snow & ice class).

Sentinel 2 L2A reflectance products were used as a basis to create the LC map. Sentinel-2 products can be downloaded from the Copernicus API Hub[15] and from Google Cloud Storage public Sentinel-2 dataset. Google Cloud Storage was found to be more stable than downloading from the Copernicus API. Sentinel-2 provides multispectral data in 13 bands covering the wavelength range between 440-2190 nm to a nominal pixel resolution of 10, 20 and 60 m every 3–5 days. Reflectance is known to be related to plant phenology [45], which can be helpful for the discrimination of vegetation LC classes. Three months of 2021 were selected to represent the yearly plant cycle: March, June and November. Each month selected defines a year's season. Winter was not considered

[13] https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/biodiversidad/mfe.aspx.

[14] https://lcviewer.vito.be/2019.

[15] https://scihub.copernicus.eu/dhus.

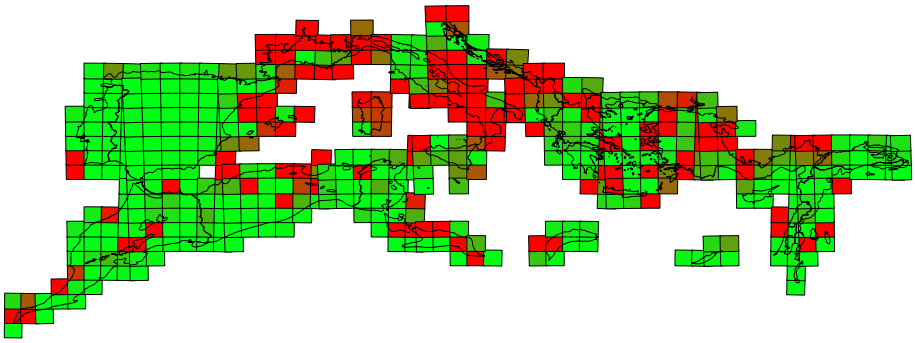


Fig. 2 Visual representation of the quality of the input data for the Mediterranean basin use case. The colour of each polygon displayed represents the percentage of missing pixels in a Sentinel tile, marked as nodata. If the nodata pixels occupy more than 20% of the final image, the colour shown is red. Tiles with a missing data percentage between 5% and 20% are graded from green to red

because the snow could cover large surfaces, and its high reflectance could be a problem for the analysis [46].

For all tiles in the working area, all non-cloudy products for the selected months are downloaded. As a general rule, the downloaded products do not contain more than 20% of the cloud pixels, as reported in the metadata. Nevertheless, there is a possibility that not enough products will be available in some months (e.g. in continuous cloudy weather). In this case, they were downloaded from the nearest available date from April (spring), July (summer) and October (autumn) until at least two non-partial products were obtained for each season.

Additionally to the Sentinel products, the ASTER Digital Elevation Model (DEM), from the Japan Space Systems services,[16] was also downloaded to obtain variables on altitude, slope and aspect of the ground, as they can also influence the distribution of LC classes on the terrain, especially natural vegetation [47]. ASTER DEM has a nominal pixel resolution of 25 ms. Since raster data are generally big files, some kind of object-based storage is recommended since downloading them at runtime is not feasible. For our use case, we utilised MinIO, an open-source S3-compatible object storage [48, 49].

Due to several incomplete Sentinel-2 products (either covering only partial tiles or with a high percentage of cloudy-covered area), a quality map has been calculated. This map allows visually checking each tile's product availability status; therefore, it checks that there is no lack of data because it may cause the land cover prediction to be incomplete. Figure 2 showcases an example of the quality of the data we could obtain for our use case. In addition, for each tile of the Mediterranean basin, a list of all the products downloaded with their cloud coverage percentage and no data percentage has been calculated to check the quality and completeness of the results.

Data pre-processing

This section describes the different pre-processing techniques applied to the raw data collected from Sentinel-2 and ASTER. These steps are executed only once for each

[16] https://www.jspacesystems.or.jp/ersdac/GDEM/E.

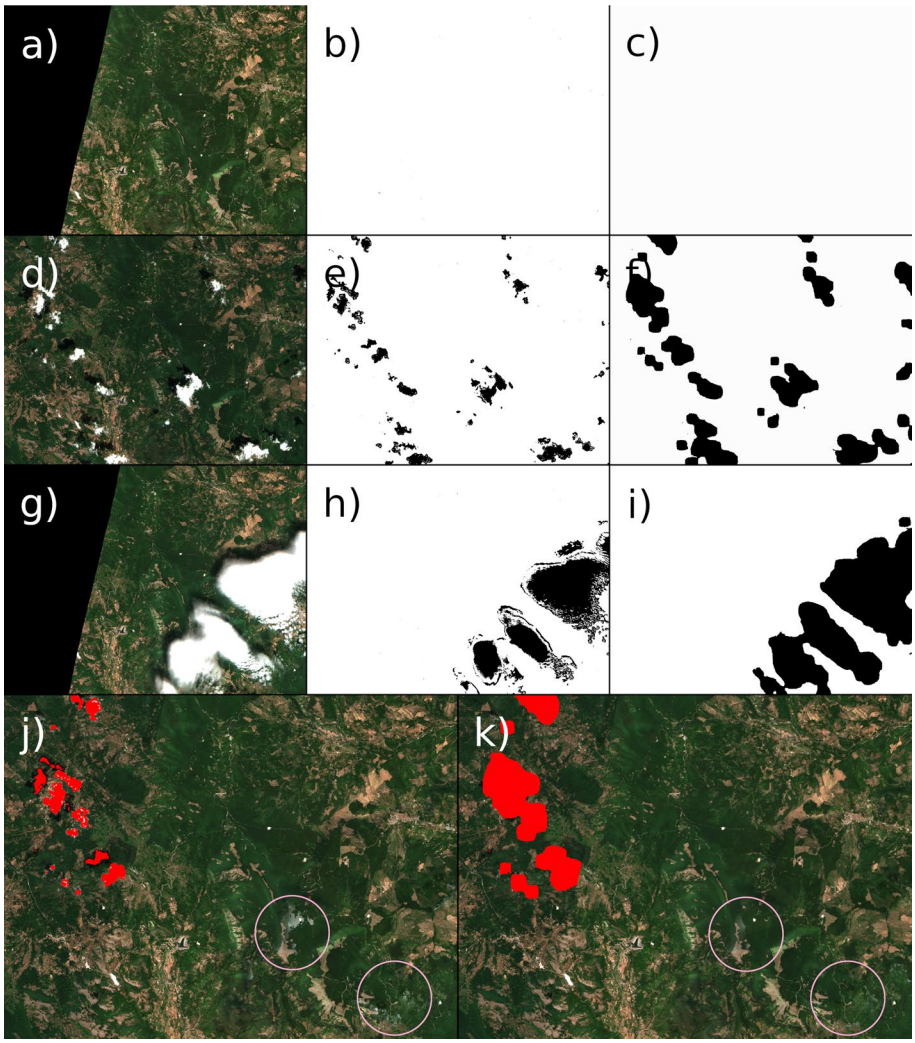


Fig. 3 **a**, **d**, **g** True Colour images (TCI) of similar dates, **b**, **e**, **h** original cloud masks using Sentinel-2 SCL cloud-related pixels, **c**, **f**, **i** post-processed cloud masks. j) is the composite of the three images using the original S-2 cloud masks, and **k** is the composite using the dilated cloud masks. Both composites show in red the no-data pixels, where data is not available in any of all original images. Marked with circles are some examples where the post-processed masks improve the composite. Note that **j** is problematic to use to obtain validated data, as pixels near no-data are probably corrupted. This problem is solved in **k** as corrupted pixels are not used for the arithmetic median. Also, the resulting composite **k** is much more precise than **j**, more noticeable when there are few products (2–3) available for making the composite

satellite data product. The results are stored for the rest of the steps in the S3-compatible object storage system, MinIO. Due to the high requirements on time, access to memory and computational efforts, it is recommendable to avoid performing these tasks during run-time.

Sentinel-2 composites A well-known problem in remote sensing is the corruption of individual products [50], either by the existence of clouds and their shadows (in multispectral imagery) or by the possibility of obtaining a partially covered tile (due to the Sentinel-2 tiling grid system). These problems are usually solved by composing each band using different products captured on a similar date [51]. In the proposed methodology, the composited bands are created using the pixel-wise median employing

Table 1 The Land Cover classifier uses a list of spectral vegetation indices as input

Index	Formula	References
Carotenoid Reflectance Index 1 (CRI1)	$\frac{B03}{B02}$	[52]
Enhanced Vegetation Index 2 (EVI2)	$2.4 * \frac{B08-B04}{B08+B04+1.0}$	[53]
Global Vegetation Moisture Index (GVMI)	$\frac{B08A-B11}{B08A+B11}$	[54]
Normalised Difference Water Index (MNDWI)	$\frac{B03-B11}{B03+B11}$	[55]
Normalised Difference Red-Edge (NDRE)	$\frac{B09-B05}{B09+B05}$	[56]
Normalised Difference Vegetation Index (NDVI)	$\frac{B08-B04}{B08+B04}$	[57]
Normalised Difference Yellowness Index (NDYI)	$\frac{B03-B02}{B03+B02}$	[58]
Optimised Soil Adjusted Vegetation Index (OSAVI)	$(1+Y) * \frac{B08-B04}{B08+B04+Y}$	[59]
Normalised Difference Red/Green Redness Index (RI)	$\frac{B04-B03}{B04+B03}$	[60]

All indices are calculated using Sentinel-2 bands. On the OSAVI index, the community accepted value for Y is 0.16

between 2 and 5 low-cloud ($\leq$ 20%) products for each season. It is worth noting that the Sentinel-2 SCL (Scene Classification Layer) is used to create a cloud mask (even including cirrus and cloud shadows) for each product. Thus, the median will only consider the cloud-free pixels.

The SCL has been improved to reduce the false negatives in the Sentinel-2 cloud mask, as it is prone to false positives in cloud surroundings (see Fig. 3). The algorithm post-processes every cloud mask, expanding only the dense trunks (usual clouds) and not expanding isolated valid pixels (usually cloud false positives). It is based on a convolution operation using an average filter with a kernel size equivalent to $600 \times 600\, m^2$ (e.g., if expanding a cloud mask with a $60\, m$ spatial resolution, the kernel will have a (10, 10) shape). The result of the convolution per pixel is a number between 0 and 1, representing the probability of finding a cloud in the defined neighbour area. That number is mapped to true when it is more significant than 0.075, transforming cloud proximity from false negatives to true positives (isolated false positives will also disappear). Both variables were defined empirically until desired results were achieved. It is noteworthy that the convolution separability property has been employed to reduce computation complexity, having a low impact on execution time even with huge bands.

As the whole operation of making a composite product is time-consuming and computationally expensive, the composites are stored in the data storage once they have been calculated. To allow reusing the composites during training and prediction stages or execution of new cases (with a different classification goal) if they share the same geographic area.

Spectral indices For the downloaded Sentinel-2 products, several *Spectral Indices*, which can reveal the relative abundance of plant greenness and moisture [41], has been computed for each product at the highest resolution available for the bands used. The final list of Spectral indices is described in Table 1.

ASTER-gdem ASTER DEM products have been denoised to reduce artefacts passing to later stages of the processing [61]. The products have been denoised using the System for Automated Geoscientific Analyses (SAGA) [62], with a value of sigma 0.85 and 60 iterations. The rest of the parameters are left to default. The sigma (or threshold on the original paper) values have been chosen following the guidelines described on [63], which suggest higher values for both parameters as the number of sharp edges increases.

Table 2 Possible values that each raster can contain, along with values predefined for normalisation

Raster name	Value range	Normalisation values
Sentinel-2 bands	(0-65535)	(0,7000)
Slope	(0-90)	(0,70)
Aspect	(0-360)	(0,360)
DEM	(0-8849)	(0,2000)

Those values have been considered typical values found in our study area

The derived products' *slope* (maximum angle of elevation between a point and its neighbours) and *aspect* (orientation of slope) have been calculated with GRASS [64] version 7.8 from the denoised products.

ASTER DEM images have to be aligned to the corresponding Sentinel-2 tile, which is done using a joint product with all the DEM tiles intersecting the Sentinel-2 one. Afterwards, this merged product is re-projected to the Sentinel-2 coordinate reference system and cropped to match the Sentinel-2 tile area. In the process, the ASTER DEM pixel size (25 ms) is also warped to the Sentinel-2 pixel resolution (10 ms). Among the different raster variables, there are different spatial resolutions (e.g., the DEM has a 30 m spatial resolution, S-2 B02 10 m and S-2 B09 60 ms). For harmonisation, all features are re-projected to 10 ms of spatial resolution. This resolution is selected because it is the finer spatial resolution and the nominal resolution of most bands of the Sentinel-2 products.

It is widespread in ML to normalise every feature of the dataset to reduce the disparities in units among different variables, which could affect the ML model. Most spectral indexes are normalised by default from -1 to 1. The most used normalisation methods are *min-max* and *Z-score* [65], which normalise each feature using the minimum and maximum values as limits, and global mean and standard deviation, respectively. However, these are designed for normalising datasets where all data can be loaded simultaneously. This is not feasible when dealing with several high-resolution rasters in terms of Random Access Memory (RAM). As commented above, data is split into tiles, which are normalised independently. Still, it is necessary to establish a normalisation range for every feature to preserve consistency across tiles.

The used vegetation indexes are already normalised, so the features to be normalised are Sentinel-2 bands (B1 to B12, aerosol optical thickness (AOT) and scene average water vapour (WVP)), and ASTER-related rasters (aspect, slope, and DEM). Table 2 shows the range of possible values and normalisation values for each feature. The normalisation values have two values (x, y), meaning that x will be mapped to -1 and y to 1. An example to help illustrate the selection of ranges per variable is the values used for the DEM; the value range is between 0 (sea level) and 8849 (highest recorded value). However, in our area of interest, there is no elevation value above 3000 ms, being the range 0–2000 the usual range of values.

Training data generation

This section focuses on describing how the data for the training stage is generated. This data combines the labelled data (ground truth data) and previously pre-processed satellite images. This is a complex process, as several challenges have been faced related to

the computation requirements for the use case. Time optimisation is done to read more than 10,000 Sentinel-2 products covering the Mediterranean Basin.

The training process involves the generation of a spectral library by intersecting the input labelled dataset against the raster variables and writing the values into a table sorted by the coordinates of each location. All labelled points are grouped per Sentinel-2 tile they fall into to make the process parallel; therefore, each tile is only read once. All points are converted to the *Military Grid Reference System* (MGRS) [66] with a precision level of 100 km to obtain the Grid Square ID used to reference Sentinel-2 product tiles.

Each Sentinel-2 raster is read as a matrix with geographical metadata associated. Thus, each matrix pixel is related to a geographical coordinate (and vice versa). This process is done employing *Rasterio* Python package [67], which provides several methods to work with raster data.

A mask, including the class annotations of each pixel, is created to process the points of the different LC classes. This way, the classes are separated afterwards, only reading the product to mask it once. In addition, to increase the data diversity for the training process, each class point is raised by a radius of one in all directions to all adjacent cells, leaving 9 samples for each point in the original database.

The resulting training data (spectral library) is organised in a dataframe where each column is a feature (ground truth sample) and each row corresponds to the coordinates of a pixel in the target spatial resolution (10 ms in our use case) containing $n\_validated\_pixels \times 9$ rows and 78 columns (raster variables), for each tile. Those columns are composed of:

- Data from ASTER-related bands (DEM, slope, and aspect).
- The point coordinates (latitude and longitude) along with its label.
- 24 columns for each of the observed seasons (summer, autumn, and spring): the identifier of the composite used, 14 Sentinel-2 bands (B01-09, B11, B12, B8A, AOT, and WVP), and the 9 indexes shown in Table 1.

Training data management

This section covers the data operations that are made to the training dataset composed in "Training data generation" section, aiming to achieve a balanced and accurate training dataset to feed the model, reducing the confusion between classes:

Class balancing. It is essential to check the histogram of the labelled data, as class over- or under- representation can lead to a lousy classification. The number of samples per class must be related to their representative in reality. In our use case, as 9 neighbouring pixels are taken for each validated coordinate, the way of balancing an over-represented class is to drop some adjacent pixels from the dataset. This way, the class histogram can be balanced without losing data heterogeneity.

Data shuffle. In ML, it is usually necessary to shuffle the data, as it reduces variance and helps prevent the model from overfitting. Shuffling the dataset is crucial in the proposed methodology as pixels are added nine by nine (central pixel and 8 neighbours). These pixels have the same label and share a similar spectral signature, causing overfitting to the model if they are not shuffled.

Merging classes. Sometimes, when two labels are very similar, the model may be unable to separate them. In this case, those classes could be merged into a more general one, encapsulating both. In remote sensing, several ways exist to identify non-spectrally separable classes, such as plotting the class spectral signatures or computing the Jeffries-Matusita [68] distance between them.

Model training and validation

Random forest and neural networks are the first ML models considered for our proposal. On the one hand, relating to the multi-layer neural network, the best-performing architecture found is a 2-layer with 40 perceptrons each using a sigmoid activation function, achieving a tolerable accuracy in the test dataset. Yet results are not promising, especially in spectrally heterogeneous classes, such as agriculture or cities. On the other hand, the best-performing random forest is composed of 100 trees, achieving very high accuracy in the training set, clearly outperforming the best neural network found. The rest of the RF parameters, implemented in the Python library *sklearn* [69],[17] were left as default. Both models are trained using the same data, which consists of 80% of the entire spectral library (from now on, training dataset). The other 20% of the dataset is used for validating the models' performance (from now on, validation dataset). The splitting percentage depends on the quantity of data available and the number of labels to predict. Before splitting the dataset in training and validation, rows with nodata features are dropped. Models are then compared by a confusion matrix: Each pixel in the validation dataset is predicted by the trained model, obtaining a set of labels (predicted). Then, the classes are compared against their actual label, which was validated by an expert (true). The correct targets are summarised along the diagonal of the confusion matrix (overall accuracy), while errors are spread in off-diagonal cells. The errors are accounted per class as errors of omission (last row), and errors of commission (last column) [70].

This classification accuracy difference is due to the random forest being an ensemble of models and the neural network being a single one. The ensemble has the benefit of predicting spectrally heterogeneous classes. For example, the agriculture class contains validated spectral data from different cereals, vegetables, and fruits, each with a different spectral signature. Each tree in the ensemble would specialise in one of those parts of the agriculture class in the training phase, facilitating its prediction in the final class. Only the random forest model is used based on the validation phase results.

Explainability and feature reduction

A tree explainer is used, from the Artificial Intelligence Explanaible method SHAP (SHapley Additive exPlanation) [71], to provide an insight on the feature importance for our model. Explainability techniques are applied to Land Cover classification to identify classes that are not being separated correctly by the model and to help improve the model by, for example, showing how the model arrived at a wrong prediction. Figure 4 shows the results of the explainer in our use case. The essential features in the general classification are represented on the Y-axis in descending order (top-bottom). The

[17] https://scikit-learn.org/stable/.

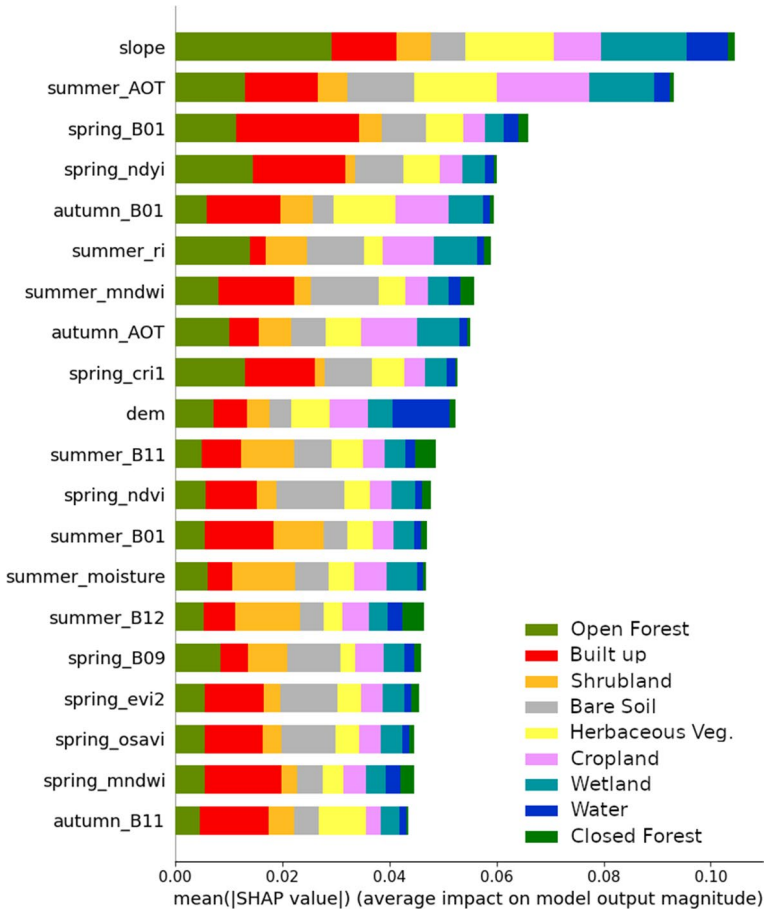


Fig. 4 Feature importance summary of the model, the Y axis list the most important bands in descending order (each class importance corresponds to a different colour) while the X axis shows the mean of the absolute value of SHAP values

impact of each feature in the classification of each class is defined in the X-axis in absolute value on a scale from 0 to the maximum contribution (0.1 in this case) as SHAP values. SHAP values are relative values that attribute to each band change in the expected model prediction when conditioning on that feature [72]. A higher SHAP value means this feature impacts the model output towards that class, while a lower value impacts the output towards a different class.

The explainability analysis can be helpful in case an improvement in time and resource usage is needed to make a feature reduction. For the case of the study, it can be beneficial since a large part of the execution time is given by loading rasters to memory, which can be reduced proportionally to the amount of discarded features. The memory needed to train the model or classify the terrain would also be reduced.

Recursive feature elimination (RFE) has been used in this project to remove the least relevant features identified in the explainability analysis. RFE is a feature selection method that trains a model and drops the feature that contributes the least to the model until the specified number of features is reached. The model is elastic net [73], a state-of-the-art regularised regression method. The RFE has been used to select the 30 most relevant features that have been used to train and test the random forest algorithm.

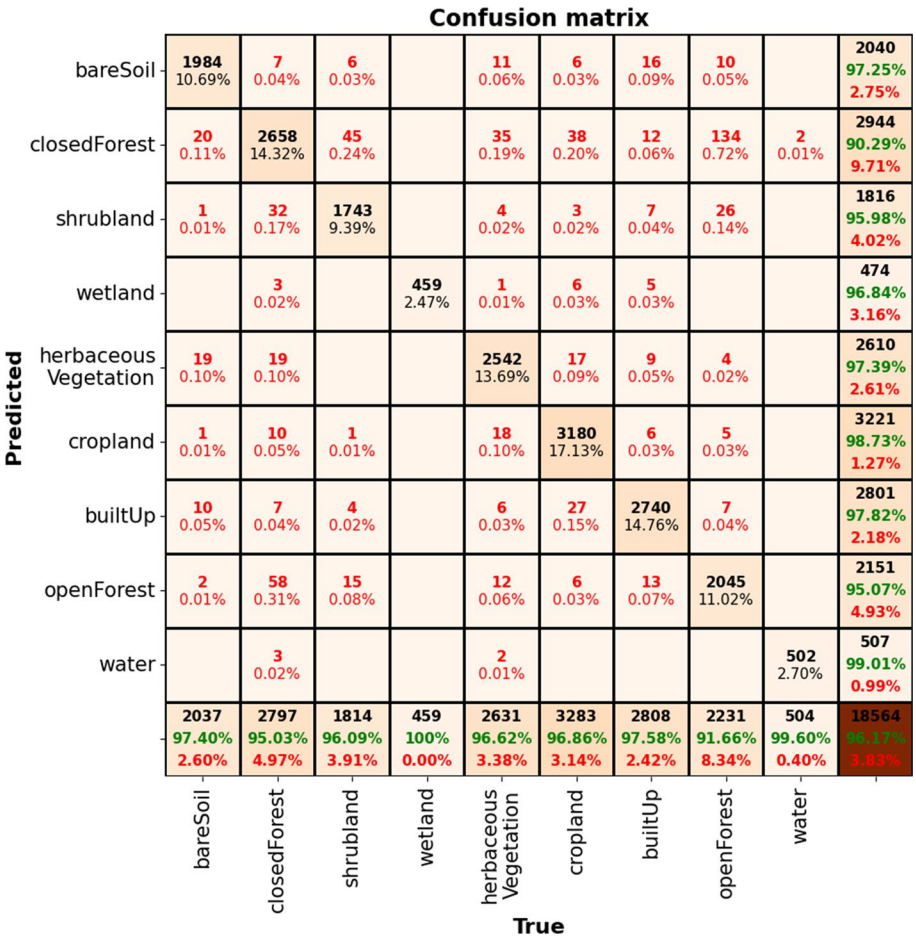


Fig. 5 Confusion matrix obtains using a Random Forest model in our use case (LC of the Mediterranean Basin). Overall map accuracy is near 96%

Results are not as precise as using all the available features. Still, the time and resources are noticeably reduced in the same proportion as the number of bands used while keeping a similar performance. Furthermore, the number of features chosen can be adapted to the resources available.

Data prediction

Data prediction is the final step of a trained ML model. The machine applies the model to each raster tile stack pixel-by-pixel in this step. The prediction results in a matrix of cells of unsigned integers (0–255), which describe the LC classes numerically. This matrix will have the exact dimensions as the Sentinel-2 tile with 10 m spatial resolution (120 MB size each).

Results and discussion

Model validation, testing and comparison with Copernicus GLC

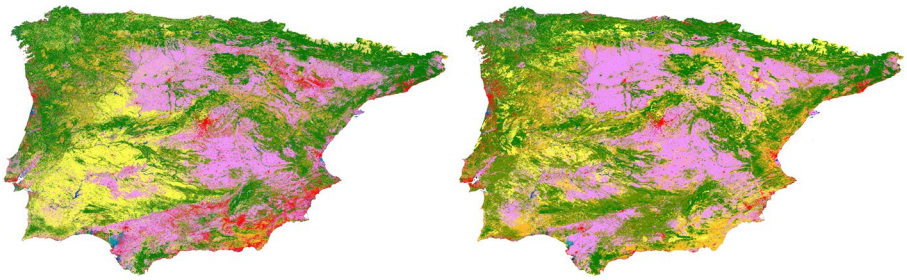
Once the whole area of interest has been predicted using the obtained model, it has to be checked if it is making correct predictions. It is usual to assess a ML model using two different datasets, the validation set and the test set.

As explained in "Model training and validation" section, the model is validated using the validation dataset, which is used to overview the correctness of the proposed model in early phases and detect coarse problems in the design of the model and selection of variables (Fig. 5). The validation dataset is a 20% of records from the training dataset, for which it is expected certain correlation and an optimistic accuracy results, if the model is correct. The test dataset is predominately used to evaluate the final product and is an independent dataset from the training dataset. Indeed, good results in the validation set (training model) do not guarantee good results with production data (entire study area). In our case, the training and validation datasets belong to only three countries (Lebanon, Spain and Tunisia), portraying the most representative landscapes of the Mediterranean basin; while the testing dataset was not used for training and belongs to 14 countries of the basin (including Lebanon, Spain and Tunisia).

The results of the validation dataset showed that the overall accuracy of the map is 96.2%, with producers' and users' accuracy per class always above the 90% (errors of omission and commission lower than 10% for all classes) (Fig. 5). These results confirm the methodological workflow is successful in the task of generating a LC map for a large area, such as the Mediterranean basin, by processing Big Data from satellite images and large ground-truth databases.

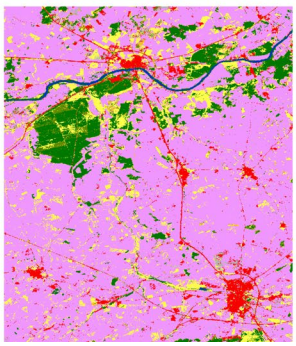
A realistic evaluation of the map is done by generating a testing dataset. The testing dataset should be composed of data validated by experts unaffiliated with the project. Following this criterion, nearly 8000 random points inside our working area were inspected to create a test dataset. As expected, the results of predicting the test dataset are not as good as those of the validation dataset, obtaining an overall accuracy of 72%. However, some classes do classify well. For example, the "closedForest" and "builtUp" classes have a 95% and 91% hit rate, respectively. The fact that there is a significant difference in accuracy in certain classes between the validation and the testing dataset may indicate that the data of this class in the training dataset does not fully represent it. This can be corrected in new iterations of the workflow by adding new data on elements of the landscape that are not represented in the first training dataset, by balancing classes and by merging classes that are not able to be spectrally separated. The "Wetland" class is an example that needs to be reviewed, since it achieves a 96.8% accuracy in the validation but a 63.3% accuracy in the testing. The number of samples reveals it is likely that wetlands are under-represented in the original training dataset.

The Copernicus GLC has inspired the selection of classes for our map; therefore, a visual comparison is also interesting. Despite the time gap between the latest GLC product (2019) and our product (2021), both maps look similar on a large scale. A significant difference occurs in the classification of dehesas and montados (in the Iberian Peninsula), which are defined in GLC as "open forest", while they are described as "Grassland" in our map (Fig. 6). On a smaller scale (Fig. 6c-d-e), our LC map shows more details because of the 10 m pixel resolution, especially linear structures such as rivers and roads. In addition, as observed in the Sentinel-2 true colour composition, the class accuracy is better in our product; for example, the main forest patch in the scene is classified partially as shrubland in GLC, as well as many cropland and grassland areas.



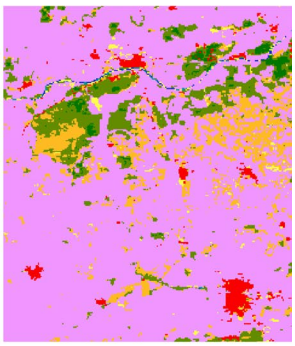
(a) 10 meters Land cover of the Iberian Peninsula in 2021

(b) Copernicus global 100 meters Land Cover of the Iberian Peninsula in 2019.



(c) 10 meters Land cover of Tordesillas, Spain in 2021

(d) Sentinel-2 10 meters true colour composite image of Tordesillas, Spain in March 2021.



(e) Copernicus global 100 meters Land Cover of Tordesillas, Spain in 2019.

Fig. 6 **a**, **b** compares the Iberian Peninsula, while **c**–**e** compares a local region in Tordesillas, Spain. All the land cover maps have the same labels: Closed and open forests (green and olive), shrubland (orange), herbaceous vegetation (yellow), herbaceous wetland (cyan), bare vegetation (grey), cropland (pink), built-up (red) and permanent water bodies (blue)

Evaluation of the explainable artificial intelligence (XAI) analysis

The results of the explainability analysis (Fig. 4) show the 20 most important variables in the generation of the LC map. The slope band (derived from the ASTER DEM) is the most important in the general classification process. This feature has the most impact on the open forest, herbaceous vegetation and wetlands classes. Besides, the outcome depicts that the AOT (Aerosol Optical Thickness) band in summer and autumn is quite relevant in most classes. This is derived from the correlation of shortwave infrared (SWIR) with the blue and red bands [74] as introduced by Kaufman et al. [75]. Among the Sentinel-2 bands, the most relevant ones are B01 (443 nm), B11 (1613 nm), B12 (2202 nm) and B09 (945 nm), which correspond to the ultraviolet, SWIR and red edge spectra, respectively. They are related to light scattering (which relates to AOT importance), plant pigment and water content. They seem to be relevant to distinguish vegetation from non-photosynthetic objects, as they are influencing the classification or built-up areas, as well as different plant types, as they are important for the separability of shrublands. in relation, the most remarkable spectral indices are NDYI, RI, MNDWI, CRI1, NDVI, EVI2, GVMI and OSAVI, especially in spring, which are also related to plant pigments and water content. In contrast with the most relevant Sentinel-2 spectral band (B01, B09 and B11), these indices are built mostly on bands B02 (green, 493 nm),

Table 3 Execution time of the whole process of creating a composite, including product download from MinIO to RAM

Used products	Average execution time	Deviation
2	08′ 23″	0′ 38″
3	11′ 06″	1′ 00″
4	12′ 08″	0′ 33″
5	13′ 29″	1′ 03″

There were 5 executions for each number of products used

B03 (red, 560 nm) and B08 (NIR, 834 nm), showing that the model is using orthogonal variables.

Execution time

The proposed methodology comprises many different tasks with different computational uses (some of them are time-consuming, others almost instantaneous). Thus, the execution time should be defined by the execution time of its smaller tasks. This section will analyse the execution time of the most relevant tasks. Note that those times depend on different variables (e.g., labelled points used, hardware and Sentinel-2 tiles to be processed, etc.). Experimentation has used a Dask cluster with 9 workers, as the methodology has been implemented to be executed in a distributed environment splitting the tasks by tile, being able to process one tile per worker at the cluster in parallel. Each worker has available 2 cores from an Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz processor, 128GB RAM, and a 100 Gbps of data transfer rate to a distributed RAID 6 made of NVMEs disks, where the S3 storage is deployed, and the persistent volumes (local disks) of the workers are available.

Generating a composite is one of the most time-consuming tasks, as all bands from all products used in the composite have to be moved from the S3 storage into RAM. Consider that composites are formed by 2 to 5 products, composed of 14 bands each. This also implies performing 14 times the composition operation, pixel-wise medians with huge matrices. The times for generating composites depending on the products used are shown in Table 3. Note that this time includes uploading the composite to the S3 storage and its metadata to the database. If the composite is needed during the execution of the workflow or following executions, this time will be skipped.

The next step of the methodology is generating the training dataset from labelled data points distributed across one or several tiles. The execution time of this phase depends on the number of tiles in which data is available. All S-2 bands and spectral indexes from the three seasons used have to be read into memory and organised into a dataset, storing that data for the labelled data points. The mean time needed to generate a training dataset from a single tile is 4' 12", with a standard deviation of 12". Whether this operation is executed on a Dask cluster, it is possible to process as many tiles as workers the cluster has in parallel, obtaining a considerable improvement in the total execution time of the workflow.

Also, the random forest model has to be trained with a dataset built by merging the ones generated in the previous step. In our case, a dataframe with $\approx$ 70000 rows have

been generated from validated data. Using that data, 62" are needed to train the model (RFE has not been used, worst-case scenario).

Predicting the land cover of a tile is the most time-consuming task. At 10 m spatial resolution a raster over a Sentinel-2 tile has 120.5 million pixels, and each of them has to be predicted using the random forest model. In the same way as generating the training datasets, the prediction is conducted in a distributed manner. In any case, due to the vast numbers of forecasts made in a tile, the mean time needed is 40', with a deviation of 2' 40".

Having a measurement of those execution times, it is possible to estimate the total execution time depending on the number of tiles used in the training phase, tiles predicted, and the number of workers available in the Dask cluster where the workflow is executed:

$$t \approx t_{ml} + \frac{t_{tt} \cdot |TT| + t_{pt} \cdot |PT| + 3 \cdot t_c \cdot |TT \cup PT|}{n_w}$$

Where t is the total execution time, t_{ml} the mean time for training the machine learning model, t_{tt} the time spent in generating the training dataset of one tile, TT the set of tiles where the training dataset is split, t_{pt} the time spent predicting a tile, PT the set of tiles predicted, t_c the mean time for generating a composite, and n_w the number of workers in the Dask cluster.

Limitations and future research

The whole workflow is designed and optimised to work at the Sentinel-2 tile level, which carries advantages and disadvantages. On the one hand, it allows the distribution of most tasks due to the independence of each tile, which is significantly advantageous when the area of interest is large. Even if computation time is not a problem, merging all rasters into a big one is not feasible because 10 ms spatial resolution rasters demand high memory for managing them together. On the other hand, some others problems and limitations could appear when the tiles are merged. Firstly, all the data from other satellites must be reprojected to Sentinel-2 tiles, e.g. harmonising the elevation and slope rasters. In this work, we have used ASTER data for that purpose; nevertheless, in the case of needing other satellite data (e.g. Sentinel 1 or Landsat-8), this process could be intensive in memory and execution time if it is done in oversized tiles instead of paralleling the task in smaller ones. Secondly, each Sentinel-2 tile has an overlapping area with adjacent tiles. Since this work is creating monthly compositions of each tile using different dates, the reflectance in the overlapping regions may not be consistent (there is more than one reflectance value in the same area). Therefore the predictions of the model could be inconsistent because the model predicts the same areas with different values. However, large labelled data will reduce the effect of such inconsistencies. A possible open solution is to merge all tiles once predicted to create a land cover map, and then post-processing should be done to harmonise predictions in the overlapped area. In the future, we will study how to resolve this issue during the pre-processing stages.

In addition, the workflow could be easily adapted to re-classify other thematic classes. For example, using a training dataset into the model of specific forest types (pine, oak, beech, etc.) will generate a forest types' map. It also would be interesting to add an

alternative to the *polylabel* algorithm used to parse polygons into points, allowing a way of sampling the polygon to several points instead of just one. This would be useful to users that want to use the classifier in small regions with few validated data.

Conclusions

Environmental protection requires constant monitoring of the status and dynamics of ecosystems. In a scenario of human impacts and climate warming, the need for timely information becomes more acute to make informed land planning and environmental conservation decisions. The current satellite missions provide continuous and frequent data for the purpose. However, the challenge is how to process enormous amounts of data efficiently. Artificial Intelligence and Big Data science are giving solutions to that challenge. In the current paper, a methodological workflow is proposed for the generation of Land Cover Maps at a large scale (in the case of the Mediterranean region) to support the management of ecosystems in the area. The project aims to develop a method and a Machine Learning model for mapping that is versatile and reproducible in other regions and uses cases, depending on the input data. Our results prove that the developed workflow handles large amounts of satellite image data (more than 30,000 rasters have been processed at 10 m pixel resolution and a database with more than 45,000 records) in a short time. The spectral library generated during the training of the model can be reused for data from different years (the current map has been proved on images of 2021), which boosts the usability of the method, allowing the generation of yearly LC maps to analyse the dynamics of the studied LC classes along the time. Additionally, our methodology has been implemented in Python 3.10 and it is distributed under a free and open source license.[18]

Moreover, the use of Artificial Intelligence algorithms improves the accuracy of the maps by transferring ground-truth information from some regions to the entire territory, as the data from only three countries of the Mediterranean basin allowed the prediction of the LC on the rest of the 16 countries that compose the area. In addition, the use of homogeneous satellite data and technique enables the creation of an LC map of comparable quality for all Mediterranean nations, constituting a reference at the regional level in terms of geographical, temporal, and thematic resolution, addressing a significant information vacuum in the region. This motivates our future research agenda, which entails the first phase to provide data from more countries to improve the LC for all Mediterranean areas-furthermore, enriching the ML algorithm with new classes for training the model and predicting different elements in the LC, such as more specific kinds of forests regarding their trees or specific crops. Finally, this would lead us to create new LC maps related to climate change or monitoring affected zones by human behaviour or natural disasters.

Acknowledgements

This work has been partially funded by grants PID2020-112540RB-C41 funded by MCIN/AEI/10.13039/501100011033, and the e-infrastructure LifeWatch ERIC Project "EnBiC2-Lab" LIFEWATCH-2019-11-UMA-4, co-funded by the ERDF (Spain's Pluri-regional Operative Programme 2014-2020 for activities related to LifeWatch ERIC) through the Spanish Ministry for Research and Innovation. José F. Aldana-Martín is supported by Grant PRE2021-098594 (Spanish Ministry of Science, Innovation and Universities). Funding for open access charge: "EnBiC2-Lab" project/LIFEWATCH-2019-11-UMA-4.

[18] https://doi.org/10.5281/zenodo.7462308.

Author contributions
JFA-M: Conceptualisation, Methodology, Software, Writing—Original Draft, Writing—Review & Editing, Supervision, Visualisation. AMB: Conceptualisation, Methodology, Software, Validation, Writing—Original Draft, Writing—Review & Editing, Visualisation. MV-P: Methodology, Software, Validation, Writing—Original Draft. CB-G: Conceptualisation, Methodology, Writing—Original Draft, Writing—Review & Editing, Supervision, Project administration. YJ: Data Curation. VG: Conceptualisation, Methodology, Investigation, Data Curation, Writing—Review & Editing, Project administration. IN-D: Writing—Review & Editing, Funding acquisition. All authors read and approved theipt.

Funding
This study received no outside funding.

Availability of data and materials
The data used for this study will be made available upon request to the authors. And the methodology has been implemented in Python 3.10 and it is distributed under a free and open source license https://doi.org/10.5281/zenodo.7462308.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they do not have any conflicts of interest with regard to this work.

Received: 20 January 2023 Accepted: 17 May 2023
Published online: 02 June 2023

References

1. Plummer S, Lecomte P, Doherty M. The ESA climate change initiative (CCI): a European contribution to the generation of the global climate observing system. Remote Sens Environ. 2017;203:2–8.
2. Weiss M, Jacob F, Duveiller G. Remote sensing for agricultural applications: a meta-review. Remote Sens Environ. 2020;236: 111402.
3. Reba M, Seto KC. A systematic review and assessment of algorithms to detect, characterize, and monitor urban land change. Remote Sens Environ. 2020;242: 111739.
4. Buchhorn M, Lesiv M, Tsendbazar N-E, Herold M, Bertels L, Smets B. Copernicus global land cover layers–collection 2. Remote Sens. 2020;12(6):1044.
5. Di Gregorio A. Land cover classification system: classification concepts and user manual: LCCS. Rome: Food & Agriculture Organization; 2005.
6. Bajocco S, De Angelis A, Perini L, Ferrara A, Salvati L. The impact of land use/land cover changes on land degradation dynamics: a Mediterranean case study. Environ Manag. 2012;49(5):980–9.
7. Potapov P, Yaroshenko A, Turubanova S, Dubinin M, Laestadius L, Thies C, Aksenov D, Egorov A, Yesipova Y, Glushkov I, et al. Mapping the world's intact forest landscapes by remote sensing. Ecol Soc. 2008. https://doi.org/10.5751/ES-02670-130251.
8. Vitousek PM. Beyond global warming: ecology and global change. Ecology. 1994;75(7):1861–76.
9. Houghton RA, House JI, Pongratz J, Van Der Werf GR, Defries RS, Hansen MC, Le Quéré C, Ramankutty N. Carbon emissions from land use and land-cover change. Biogeosciences. 2012;9(12):5125–42.
10. Hashem N, Balakrishnan P. Change analysis of land use/land cover and modelling urban growth in greater Doha, Gatar. Ann GIS. 2015;21(3):233–47.
11. Smith P, Clark H, Dong H, Elsiddig E, Haberl H, Harper R, House J, Jafari M, Masera O, Mbow C, et al. Agriculture, forestry and other land use (Afolu). 2014.
12. Wulder MA, Coops NC, Roy DP, White JC, Hermosilla T. Land cover 2.0. Int J Remote Sens. 2018;39(12):4254–84.
13. Comber A, Wulder M. Considering spatiotemporal processes in big data analysis: insights from remote sensing of land cover and land use. Hoboken: Wiley Online Library; 2019.
14. Chi M, Plaza A, Benediktsson JA, Sun Z, Shen J, Zhu Y. Big data for remote sensing: challenges and opportunities. Proc IEEE. 2016;104(11):2207–19.
15. Liu P. A survey of remote-sensing big data. Front Environ Sci. 2015;3:45.
16. Iverson LR, Graham RL, Cook EA. Applications of satellite remote sensing to forested ecosystems. Landsc Ecol. 1989;3(2):131–43.
17. Carlson TN, Arthur ST. The impact of land use–land cover changes due to urbanization on surface microclimate and hydrology: a satellite perspective. Global Planet Change. 2000;25(1–2):49–65.
18. Abrams M. The advanced spaceborne thermal emission and reflection radiometer (Aster): data products for the high spatial resolution imager on nasa's terra platform. Int J Remote Sens. 2000;21(5):847–59.
19. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017;202:18–27.

20. Amani M, Ghorbanian A, Ahmadi SA, Kakooei M, Moghimi A, Mirmazloumi SM, Moghaddam SHA, Mahdavi S, Ghahremanloo M, Parsian S, et al. Google earth engine cloud computing platform for remote sensing big data applications: a comprehensive review. IEEE J Sel Top Appl Earth Obs Remote Sens. 2020;13:5326–50.
21. Franklin S, Wulder M. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. Prog Phys Geogr. 2002;26(2):173–205.
22. Gašparović M, Jogun T. The effect of fusing sentinel-2 bands on land-cover classification. Int J Remote Sens. 2018;39(3):822–41.
23. Kussul N, Lavreniuk M, Kolotii A, Skakun S, Rakoid O, Shumilo L. A workflow for sustainable development goals indicators assessment based on high-resolution satellite data. Int J Dig Earth. 2019. https://doi.org/10.1080/17538947.2019.1610807.
24. Ghorbanian A, Kakooei M, Amani M, Mahdavi S, Mohammadzadeh A, Hasanlou M. Improved land cover map of Iran using sentinel imagery within google earth engine and a novel automatic workflow for land cover classification using migrated training samples. ISPRS J Photogramm Remote Sens. 2020;167:276–88.
25. Barba-González C, García-Nieto J, del Mar Roldán-García M, Navas-Delgado I, Nebro AJ, Aldana-Montes JF. Bigowl: knowledge centered big data analytics. Expert Syst Appl. 2019;115:543–56.
26. Vali A, Comai S, Matteucci M. Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. Remote Sens. 2020;12(15):2495.
27. Rwanga SS, Ndambuki JM, et al. Accuracy assessment of land use/land cover classification using remote sensing and GIS. Int J Geosci. 2017;8(04):611.
28. Solano F, Di Fazio S, Modica G. A methodology based on Geobia and worldview-3 imagery to derive vegetation indices at tree crown detail in olive orchards. Int J Appl Earth Obs Geoinf. 2019;83: 101912.
29. Phan TN, Kuch V, Lehnert LW. Land cover classification using google earth engine and random forest classifier–the role of image composition. Remote Sens. 2020;12(15):2411.
30. Modica G, Messina G, De Luca G, Fiozzo V, Praticò S. Monitoring the vegetation vigor in heterogeneous citrus and olive orchards. a multiscale object-based approach to extract trees' crowns from uav multispectral imagery. Comput Electron Agric. 2020;175: 105500.
31. Modica G, De Luca G, Messina G, Praticò S. Comparison and assessment of different object-based classifications using machine learning algorithms and UAVS multispectral imagery: A case study in a citrus orchard and an onion crop. Eur J Remote Sens. 2021;54(1):431–60.
32. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
33. Talukdar S, Singha P, Mahato S, Pal S, Liou Y-A, Rahman A. Land-use land-cover classification by machine learning classifiers for satellite observations–a review. Remote Sens. 2020;12(7):1135.
34. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.
35. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv (CSUR). 2018;51(5):1–42.
36. Kolevatova A, Riegler MA, Cherubini F, Hu X, Hammer HL. Unraveling the impact of land cover changes on climate using machine learning and explainable artificial intelligence. Big Data Cognit Comput. 2021;5(4):55.
37. Rocklin M. Dask: Parallel computation with blocked algorithms and task scheduling. In: Proceedings of the 14th Python in Science Conference, Citeseer, vol. 130, 2015; p. 136.
38. Vogt P, Caudullo G, et al. Global analysis of forest attribute layers for the EU observatory on deforestation and forest degradation. 2022.
39. Dierckx W, Sterckx S, Benhadj I, Livens S, Duhoux G, Van Achteren T, Francois M, Mellab K, Saint G. Proba-v mission for global vegetation monitoring: standard products and image quality. Int J Remote Sens. 2014;35(7):2589–614.
40. European Topic Centre on Urban L, Systems S. Forest fragmentation in the mediterranean basin: progress report on theoretical approach and implementation steps. Technical report, 2020. https://forum.eionet.europa.eu/etc-urban-land-and-soil-systems/library/c.2_ap-2020/1.7.8.1-forest-cooperation-regional-conventions/m2/milestone-report-2-part-ii.-mediterranean-forest.
41. Xue J, Su B. Significant remote sensing vegetation indices: a review of developments and applications. J Sens. 2017. https://doi.org/10.1155/2017/1353691.
42. Selmi K, Tissaoui M, Bacha S, Chok M, Salem A. Inventaire des forêts par télédétection: Résultats du deuxième inventaire forestier et pastoral national. In: Direction Générale des Forêts et Le Centre National de Cartographie et de Télédetection, Media Horizon, 2010.
43. Agafonkin V. Polylabel: a fast algorithm for finding the pole of inaccessibility of a polygon 2016. https://github.com/mapbox/polylabel.
44. Buchhorn M, Lesiv M, Tsendbazar N-E, Herold M, Bertels L, Smets B. Copernicus global land cover layers-collection 2. Remote Sens. 2020. https://doi.org/10.3390/rs12061044.
45. Peñuelas J, Filella I. Visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends Plant Sci. 1998;3(4):151–6.
46. Zhu Z, Woodcock CE. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens Environ. 2014;144:152–71.
47. Rivas-Martínez S. Pisos bioclimáticos de españa. Lazaroa. 1983;5(1983):33–43.
48. Peng C, Jiang Z. Building a cloud storage service system. Procedia Environ Sci. 2011;10:691–6.
49. Kumar A, Lee H, Singh RP. Efficient and secure cloud storage for handling big data. In: 2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012), IEEE, 2012, pp. 162–166.
50. Teng CM. Dealing with data corruption in remote sensing. In: International Symposium on Intelligent Data Analysis, Springer. 2005; pp. 452–463.
51. Roerink G, Menenti M, Verhoef W. Reconstructing cloudfree NDVi composites using fourier analysis of time series. Int J Remote Sens. 2000;21(9):1911–7.

52. Gitelson AA, Zur Y, Chivkunova OB, Merzlyak MN. Assessing carotenoid content in plant leaves with reflectance spectroscopy. Photochem Photobiol. 2002;75(3):272–81.
53. Jiang Z, Huete AR, Didan K, Miura T. Development of a two-band enhanced vegetation index without a blue band. Remote Sens Environ. 2008;112(10):3833–45.
54. Ceccato P, Gobron N, Flasse S, Pinty B, Tarantola S. Designing a spectral index to estimate vegetation water content from remote sensing data: part 1: Theoretical approach. Remote Sens Environ. 2002;82(2–3):188–97.
55. Xu H. A study on information extraction of water body with the modified normalized difference water index (MNDWT). J Remote Sens. 2005;9(5):589–95.
56. Fitzgerald G, Rodriguez D, Christensen L, Belford R, Sadras V, Clarke T. Spectral and thermal sensing for nitrogen and water status in rainfed and irrigated wheat environments. Precision Agric. 2006;7(4):233–48.
57. DeFries RS, Townshend J. NDVI-derived land cover classifications at a global scale. Int J Remote Sens. 1994;15(17):3567–86.
58. Sulik JJ, Long DS. Spectral considerations for modeling yield of canola. Remote Sens Environ. 2016;184:161–74.
59. Rondeaux G, Steven M, Baret F. Optimization of soil-adjusted vegetation indices. Remote Sens Environ. 1996;55(2):95–107.
60. Madeira J, Bedidi A, Cervelle B, Pouget M, Flay N. Visible spectrometric indices of hematite (HM) and goethite (GT) content in lateritic soils: the application of a thematic mapper (TM) image for soil-mapping in Brasilia, Brazil. Int J Remote Sens. 1997;18(13):2835–52.
61. Stevenson JA, Sun X, Mitchell NC. Despeckling SRTM and other topographic data with a denoising algorithm. Geomorphology. 2010;114(3):238–52. https://doi.org/10.1016/j.geomorph.2009.07.006.
62. Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J. System for automated geoscientific analyses (saga) v. 2.1.4. Geosci Model Dev. 2015;8(7):1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.
63. Sun X, Rosin PL, Martin R, Langbein F. Fast and effective feature-preserving mesh denoising. IEEE Trans Visual Comput Graphics. 2007;13(5):925–38. https://doi.org/10.1109/TVCG.2007.1065.
64. Neteler M, Bowman MH, Landa M, Metz M. GRASS GIS: a multi-purpose open source GIS. Environ Model Softw. 2012;31:124–30. https://doi.org/10.1016/j.envsoft.2011.11.014.
65. Pandey A, Jain A. Comparative analysis of KNN algorithm using various normalization techniques. IJCNIS. 2017;9(11):36.
66. Mercator UT. The Military Grid Reference System (MGRS), and the Universal Polar Stereographic (UPS), 2014.
67. Gillies S et al. Rasterio: geospatial raster I/O for Python programmers (2013–). https://github.com/rasterio/rasterio.
68. Wicaksono P, Aryaguna PA. Analyses of inter-class spectral separability and classification accuracy of benthic habitat mapping using multispectral image. RSASE. 2020;19: 100335. https://doi.org/10.1016/j.rsase.2020.100335.
69. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
70. Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ. 1991;37(1):35–46.
71. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
72. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017. pp. 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
73. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67(2):301–20.
74. Obregón MA, Rodrigues G, Costa MJ, Potes M, Silva AM. Validation of ESA sentinel-2 l2a aerosol optical thickness and columnar water vapour during 2017–2018. Remote Sens. 2019. https://doi.org/10.3390/rs11141649.
75. Kaufman YJ, Sendra C. Algorithm for automatic atmospheric corrections to visible and near-IR satellite imagery. Int J Remote Sens. 1988;9(8):1357–81. https://doi.org/10.1080/01431168808954942.

Publisher's Note