D2.6 MLs classification in Carbon sequestration capacity groups

***MAIL**:* Identifying Marginal Lands in Europe and strengthening their contribution potentialities in a CO2 sequestration strategy

MAIL project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 823805; [H2020 MSCA RISE 2018]

Deliverable D2.6 reflects only the authors' views, and the Agency is not responsible for any use that may be made of the information it contains.

Project title	Identifying Marginal Lands in Europe and strengthening their contribution potentialities in a CO2 sequestration strategy
Call identifier	H2020 MSCA RISE 2018
Project acronym	MAIL
Starting date	01.01.2019
End date	31.12.2021
Funding scheme	Marie Skłodowska-Curie
Contract no.	823805
Deliverable no.	D2.6
Document name	MAIL_D2.6.pdf
Deliverable name	MLs classification in Carbon sequestration capacity groups
Work Package	WP2
Nature[1]	R
Dissemination[2]	PU
Editor	Franscisco Gallego Ciprés, Fernando Bezares Sanfelip (Cesefor), Charalampos Georgiadis (AUTH)
Authors	Pilot site level: Dzhaner Emin, Giulia Molisse, Ragasree Polepally (IABG) European level: Lefteris Mystakidis, Lampros Papalampros (HOMEOTECH)
Contributors	All consortium Partners
Date	28.12.2021

[1] **R** = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

[2] **PU** = Public, **PP** = Restricted to other programme participants (including the Commission Services), **RE** = Restricted to a group specified by the consortium (including the Commission Services), **CO** = Confidential, only for members of the consortium (including the Commission Services).

MAIL CONSORTIUM

Aristotle University of Thessaloniki (AUTH) Greece	Industrieanlagen Betriebsgesellschaft MBH (IABG) Germany
Gounaris N. – Kontos K. OE (HOMEOTECH) Greece	Centrum Badan Kosmicznych Polskiej Akademii Nauk (CBK PAN) Poland
Universitat Politecnica de Valencia (UPV) Spain	Fundacion Centro De Servicios Y Promocion Forestral Y de su Industria De Castilla y Leon (CESEFOR) Spain

ABBREVIATIONS

Term	Explanation
AGB	Above Ground Biomass
AGBC	Above Ground Biomass Carbon
AOI	Area Of Interest
API	Application Programming Interface
CCC	Carbon Carrying Capacity
CCS	Current Carbon Sequestration
CGAT	Geo-Environmental Cartography and Remote Sensing Group
CHM	Canopy Height Model
CPU	Central Processing Unit
CSC	Carbon Sequestration Capacity
CSP	Carbon Sequestration Potential
DBH	Diameter at Breast Height
DTM	Digital Terrain Model
EEA	European Environmental Agency
EFI	European Forest Institute
ESA	European Space Agency
FAO	Food and Agriculture Organization
GEE	Google Earth Engine
GIS	Geographical Information Systems
GLCM	Gray Level Co-occurrence Matrix
IPCC	Intergovernmental Panel on Climate Change

KNN	K-Nearest Neighbor
LIDAR	Light Detection and Ranging
ML	Marginal Land
RADAR	Radio Detection and Ranging
RF	Random Forest
RMSE	Root Mean Squared Error
SNAP	Sentinel Application Platform
SOM	Soil Organic Matter
SWIR	Shortwave Infrared
VI	Vegetation Indices
XGB	Extreme Gradient Boosting

Contents

EXECUTIVE SUMMARY

The objective of this task was the development of a methodology for Marginal Lands classification into Carbon Sequestration Capacity (CSC) groups. In the framework of the ***MAIL*** project, this task used the knowledge acquired in tasks 2.3, 2.5 and 4.2.

In order to meet the objective, two different approaches were developed, due to the different scale levels reported on this task, at pilot site level and at European level.

The methodology implemented regarding pilot site level consisted of defining and estimating Above Ground Biomass (AGB), Current Carbon Sequestration (CCS) and Carbon Sequestration Capacity (CSC). Accurate AGB mapping is crucial for studies on carbon sequestration as they are directly connected, due to this, in the pilot site level, several indicators generated from satellite images tested in order to evaluate which and how they influence the prediction of AGB. This was followed by the integration of results from task 4.2 for the estimation of CSC within the study area and finally the classification of the pilot area into CSC groups.

A different methodology implemented regarding European level, as methodology applied in pilot site level needs very long computation time, beyond the scope of this Task. The methodology based on multicriteria GIS analysis with data including tree species maps, land cover maps and Aboveground Biomass maps. The aim was to estimate potential suitable species for afforestation for each Marginal Land as well species' Above Ground Biomass Carbon (AGBC) and proceed to classification into CSC groups.

From the findings of the task, a tool was developed to assist all possible users (policy makers, stakeholders, students, etc.) by providing a general overview regarding CSC groups and Potential Suitable Species for afforestation of MLs.

1. INTRODUCTION AND GOALS

The following document reports the main findings from the implementation of Task 2.7, namely “Marginal Lands classification in Carbon Sequestration Capacity groups”. Hence, this Chapter illustrates the structure of the entire document as well as its main objectives.

For the completion of task 2.7, the following objectives were defined:

1. Classifying Carbon Sequestration Capacity (CSC) groups;

2. Identifying and developing indicators which help the estimation of the Current Carbon Sequestration (CCS) within an area.

In order to reach the objectives above, the following work proposes different approaches for different levels of detail. That are, at pilot site and at European level.

At the finest scale, a pilot site of around 4,000 hectares (ha) is selected as a case of study. Within the pilot site, Above Ground Biomass (AGB), Current Carbon Sequestration (CCS) and related Carbon Sequestration Capacity (CSC) are defined and estimated. This first part of the report includes the definition of a structured methodology, a thorough analysis of several indicators, and the final creation of a CSC groups map (Chapter 3). Subsequently, the second part of the report concerns a broader scale that of the European Union (Chapter 4).

Serra de Espadan was selected as pilot site to assess whether the use of different indicators that influence the prediction of Above Ground Biomass (AGB). Accurate AGB mapping is a major step for studies on carbon sequestration, as the estimation of the latter is directly connected to the biomass present in the area. Hence, through the testing and evaluation of a variety of indicators – i.e., Vegetation Indices (VI), topographic measures, etc. – generated from satellite images, we aim to identify which of these result in a better predictive capacity for modelling AGB when working with Machine Learning regression models.

This approach is characterized by high-quality outputs, however, complex methodologies could not be easily scaled up due to their long computation time, cost of production, as well as the need for expert professionals. Therefore, classifying MLs into CSC groups at European level was performed by implementing an alternative

approach, which is by estimating potential suitable species for afforestation and their Aboveground Biomass Carbon.

2. DEFINITIONS

The following Chapter highlights the role that forested areas have in the context of Carbon sequestration; hence, it illustrates the concepts of Above Ground Biomass (AGB), Current Carbon Sequestration (CCS), Carbon Carrying Capacity (CCC), and Carbon Sequestration Capacity (CSC).

2.1. Above Ground Biomass and Current Carbon Sequestration

Carbon sequestration is a natural process involving the capture of carbon dioxide from the atmosphere and its long-term storage into 3 major carbon pools; namely, terrestrial, oceanic, and geological pool (IPCC, 2006; Salem et al., 2020). Furthermore, carbon can be stored in either a liquid or a solid state by different means such as trees, soil, ocean, or organic matter (Lackner, 2003). The Intergovernmental Panel on Climate Change (IPCC, 2006) divides the terrestrial pool into 5 main reservoirs: Above Ground Biomass (AGB), Below Ground Biomass (BGB), litter, woody debris, and Soil Organic Matter (SOM). Hence, the carbon sequestrated within the terrestrial pool is the sum of the amounts of carbon in vegetation biomass and soil (Salem et al., 2020).

Forests account for the largest portion of the terrestrial vegetation biomass and are characterized by the highest carbon density compared to other terrestrial environments (Stinson, et al., 2012). Up to 80% of the above-ground carbon stored in the terrestrial pool (IPCC, 2006) is stored in forest ecosystems. However, because forests are affected by many disturbances - i.e., fires, deforestation, parasites, land use change, etc., around 60% of newly stored carbon is cyclically returned to the atmosphere (Salem et al., 2020).

These characteristics make monitoring the dynamics of forest biomass an important step in acquiring up-to-date and reliable information about the state of global carbon budget, particularly in the context of climate change mitigation (Galidaki, et al., 2017). Thus, the estimation of Above Ground Biomass (AGB) has been the subject of extensive research, given its importance in planning carbon emission mitigation strategies and sustainable forest management. There are two main groups of methods regarding ABG estimation. ABG can be estimated either through direct or indirect

methods (Salem et al., 2020; Galidaki, et al., 2017; Lu, 2006). Direct methods, also known as destructive, include the harvesting, separation into components, oven drying, and weighing of the tree components as fractions of the biomass. These methods are highly precise; however, they are also expensive, destructive, and time-consuming, which makes them unsustainable to be used in large scale (Picardet al., 2012; Salem et al., 2020). The outputs from direct methods have been used throughout the years for building allometric equations - i.e., statistical models belonging to the indirect methods category. Indirect methods, also known as non-destructive, do not require the physical destruction of trees. These can be further divided into two main approaches: allometric equations and Remote Sensing-based estimations (Galidaki, et al., 2017), the latter being dependent on the first one for model training.

Allometric equations are statistical models that use forest measurable biophysical characteristics - i.e., height, Diameter at Breast Height (DBH) or crown size - to estimate either tree volume or biomass; such models allow for the estimation of biomass without the need of harvesting (Vashum, 2012). Allometric equations have been developed for many tree species as well as for the most common species combination (Salem et al., 2020). The use of remote sensing and GIS in the context of estimation of forest biomass have received attention due to increase of spatial resolution (Zheng, et al., 2004; Pandit et al., 2019; Gao, et al., 2018; Cairns et al., 1997). Furthermore, previous studies estimate BGB and litter in mature forests to be around 20% and 10-20% of the predicted AGB, respectively (Kankare, et al., 2013; Cairns et al., 1997).

Estimated or measured AGB is used as the basis for the estimation of the Current Carbon Sequestration (CCS). Within a forested area, the CCS can be defined as the amount of carbon stored in forest biomass in the moment of the forest inventory. For this work we will focus on the amount of carbon stored in the above ground portion of a forested area. Therefore, the estimation of CCS from AGB is commonly performed by multiplying the AGB by an average conversion factor of 0.5, which assumes the 50% of the dry biomass to be carbon (Khan et al., 2020). More precise factors have been provided for different ecological zones; the IPCC (2006) suggests a factor of 0.47 for tropical and subtropical forests. However, when the estimation of CS requires higher precision, it should be noticed that the carbon content in dry matter changes between tree species, as well as amongst climate zones, and it is correlated to the species wood density; species with higher wood density store higher quantities of carbon; that

is Mediterranean species have higher carbon content than tropical species, despite their extent and density being more limited (Thomas & Martin, 2012).

2.2. Carbon Carrying Capacity and Carbon Sequestration Capacity

Other than the Current Carbon Sequestration (CCS), two further concepts essential in this project framework are the Carbon Carrying Capacity (CCC) and the Carbon Sequestration Capacity or Potential (CSC or CSP). Henceforth, the latest will be referred to as CSC.

The Carbon Carrying Capacity (CCC) is defined by Keith (2009) as the amount of carbon stored in a forest in a state of dynamic equilibrium and excluding anthropogenic disturbances; this state of saturation is reached when the forest reaches a full-growth, namely old-growth forest. Hence, a forest's CCC, together with its CCS, can be used to estimate its Carbon Sequestration Capacity (CSC). This is defined as the maximum potential quantity of carbon confinement for a forest in the moment being, and it is estimated as the difference between the CCC and the CCS (Liu, 2012; Keith, 2009; Khan et al., 2020).

$$CSC = CCC - CCS$$

Equation 1. Carbon Sequestration Capacity.

However, the estimation of a forest's CCC is not a trivial task. A forest storage capacity tends to rapidly grow during its early development and slows down around 80-100 years, depending on the species or forest types, reaching a dynamic balance with the amount of carbon in the atmosphere, when constant climatic conditions are considered (Zhou et al., 2002). Moreover, forestry Inventories of such old plots are rare to be found. Liu (2012) used a mix of remote sensing and geo-statistics techniques, as well as old-growth forest inventories, to estimate a reference CCC for each one of the level-2 FAO world ecological zones (FAO, 2012; Liu, 2012). However, for the pilot site located in Serra de Espadan, Spain, the CCC was accurately estimated in the framework of task 4.2, within which several scenarios were considered.

3. CARBON SEQUESTRATION CAPACITY MAPPING AT THE PILOT SITE LEVEL

The following Chapter focuses on a study area located in Serra de Espadan Natural Park, Spain. Hence, it covers the main steps of an exploratory workflow for the

estimation of Above Ground Biomass (AGB), related Current Carbon Sequestration (CCS), and the final mapping of Carbon Sequestration Capacity (CSC) groups.

3.1. Objectives

Starting off with the 2 main objectives defined for task T2.7 – that are the (1) classification of CSC groups and (2) the identification and generation of indicators which help the estimation of the CCS –, these were broken down to build a methodology which aims to assess the capabilities of Sentinel-2 derived measures for the estimation AGB. More specifically, we want to identify an optimal Season for satellite data acquisition and a ranking of the most influential satellite-generated indicators using Machine Learning algorithms.

Therefore, the following pilot area-specific objectives were defined:

- Generation and testing of several Vegetation Indices (VI), Biophysical param, texture and topographic measures as CCS indicators;
- Testing satellite images collected in different Seasons: Summer (August 2015) and Autumn (November 2016);
- Incorporating outputs from task T4.2 for the estimation of CSC within the study area;
- Classification of the pilot area into CSC groups.

3.2. Related studies

Above Ground Biomass (AGB) estimation, together with the estimation of CCC, is the foundation for defining CSC groups. This aspect is explained in more detail in section 3.3. Therefore, the following section contains a brief literature review on remote sensing-based studies for the estimation of AGB. More specifically, the focus is on those studies involving the use of the twin satellites Sentinel-2A and 2B.

Remote sensing-based estimation of AGB utilizes a variety of sensors, features and several regression models (Galidaki, et al., 2017; Lu, 2006; Salem et al., 2020). It their extensive literature review, Salem et al. (2020), analysed over 150 peer-reviewed articled on AGB and CS estimation. Only 25% of these studies use imagery data from active sensors, almost equally divided into LiDAR (laser imaging, detection, and ranging) and Radar (radio detection and ranging) technologies; while the majority of them deals with imagery collected from various passive sensors.

Sentinel-2A was launched in 2015 by the European Space Agency (ESA) and since then its potential in AGB estimation has been evaluated by several authors and within a variety of ecological zones (Pandit et al., 2019).

The work of Pandit et all (2018) explores the performance of spectrally derived indices from Sentinel-2A as inputs in a Random Forest (RF) model in a subtropical forest in Nepal. Field-based AGB values were estimated by applying an allometric equation using forestry inventory data from 113 measured plots with a radius of 12 m. The model performance was assessed by using a 10-fold cross-validation. The predicted AGB ranged from a minimum of 35.42 t/ha to a maximum of 276.92 t/ha, with a mean of 160 t/ha; the final model resulted in a Root Mean Squared Error (RMSE) of 25.32 t/ha between the observed and the predicted biomass.

Khan et al. (2020) explored the use of Sentinel-2 images in a mountainous temperate forest in Pakistan. Their study examines the performance of 3 categories of spectrally derived VIs - i.e., Broadband, Canopy Water Content, and Narrow band red-edge VIs. Out of 25 indices, only 11 were used as inputs in the linear regression model, most of which were red-edge VIs. The predicted AGB ranged from a minimum of 46.45 t/ha to a maximum of 279.59 t/ha, and a mean of 148.79 t/ha. The final biomass map was validated using 10 out of the 55 plots, with a radius of 17 m, and resulted in a RMSE of 35.23 t/ha between observed and predicted biomass.

Additionally, Sentinel-2 single bands have been used in boreal areas for estimating AGB over the entire Norwegian territory. In their work, Puliti et al. (2020), used the whole Norwegian forestry inventory, composed of 7,710 plots with a radius of 9 m. Furthermore, a Canopy Height Model (CHM) was included as model input; this was generated by normalising a freely available 2 m resolution DEM covering the entire area north of 60 degrees of latitude with a freely available 10 m resolution Digital Terrain Model (DTM). The final maps were evaluated using cross-validation, resulting in a RMSE of 45.8 t/ha when using solely Sentinel-2 single bands; a similar performance was achieved from using solely the CHM, with a RMSE OF 47.7 t/ha; finally, a noticeable synergy was found when using both, with a decrease of the error down to 41.4 t/ha.

In conclusion, studies involving the twin satellites Sentinel-2A and 2B show higher performance in AGB estimation compared to working with Landsat images. This is attributed to its higher spectral, spatial, and temporal resolution compared to the

Landsat mission. Specially, the presence of red-edge bands makes these images highly valuable for vegetation analysis (Pandit, Tsuyuki, & Dube, 2018).

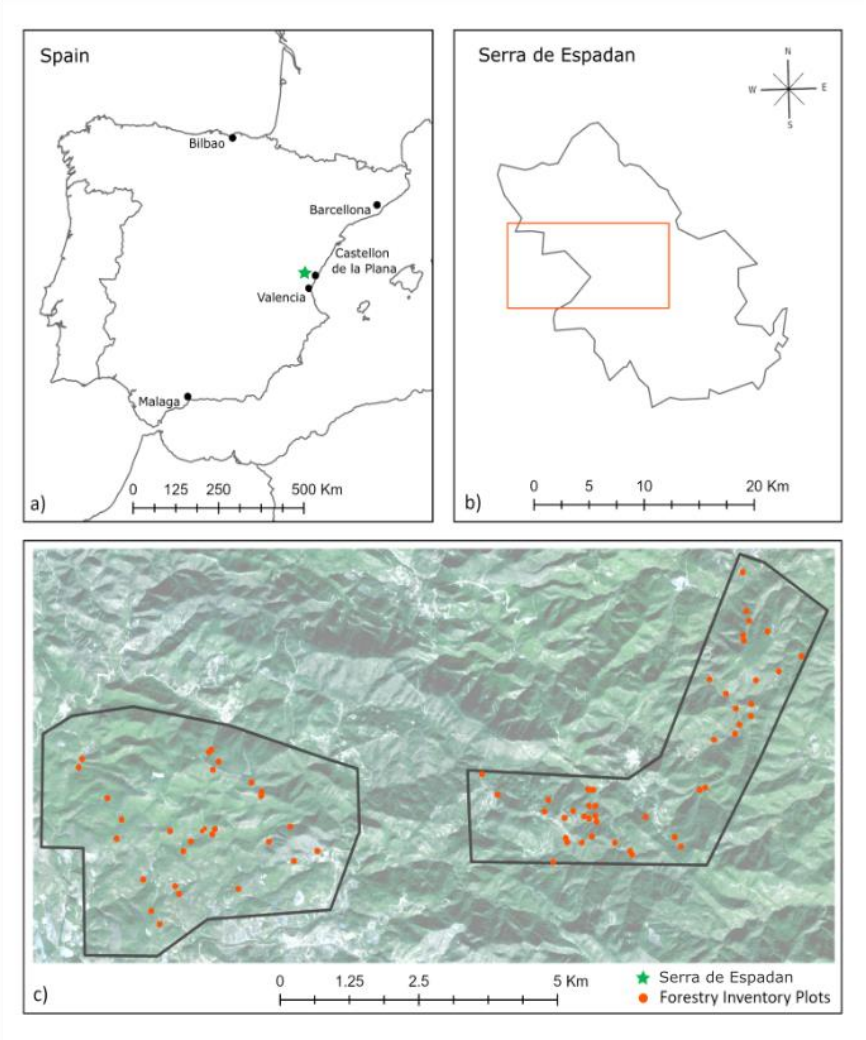
3.3. Materials & Methodology

This Section contains the description of the pilot site, as well as data collection and data pre-processing. Moreover, it illustrates the proposed methodological workflow.

3.3.1. Study area

The area was identified as Marginal Land (ML) in a previous ***MAIL*** task, which aimed to the identification of MLs in Europe. The following description of the study area was provided by a study from Torralba and Crespo-Peremarch (2018). The area object of study covers a total of 3,741.5 ha and is located in the Natural Park of Serra de Espadan, in the eastern Spain province of Castellon (Figure 1). This natural park is a Mediterranean forest with soft and rounded hills, presence of abandoned farming with artificial terraces, and mountain peaks up to 1100 meters of altitude.

A European Environment Agency report from 2017 classified this area as a semi-natural forest with a natural function, composition, and structure, but modified by human activities throughout history (EEA, 2017). Forest types and conditions have been influenced by human needs and changes in land use, as well as reforestation of single species policies from the last century. This area displays a heterogeneous landscape dominated by pure and mixed native coniferous and deciduous forests, with species of the genera Pinus and Quercus. In accordance with the global ecological zones described by FAO and to the Koppen-Trewartha Climatic groups, the Mediterranean climate is a variety of the subtropical climate, together with the Oceanic, the Humid subtropical, the Semi-desert, and desert climate (FAO, 2012). Hence, since many of the authors mentioned in Chapter 2 refer to their area of interest by using the climate domain - i.e., tropical, subtropical, temperate, boreal, polar - from now on, our area of study will be referred to by the name of its major climate domain, that is subtropical.



Location of Serra de Espadan Natural Park, Spain (a). Location of the study area within Serra de Espadan Natural Park (b). Distribution of the forestry inventory plots within the area (c).

Figure 1. Study area and location of Forestry Inventory plots.

3.3.2. Field data and field based AGB

A field inventory with measured Above Ground Biomass (AGB) at the plot level was provided by the Geo-Environmental Cartography and Remote Sensing Group (CGAT) at the Universitat Politecnica de Valencia (UPV); the collection of this forestry inventory was funded by the Spanish Ministerio de Economıa y Competitividad, in the framework of the project CGL2016-80705-R. The field data was collected in September 2015 for a total of 73 circular plots with a radius of 15 m distributed throughout the study area. Diameter at Breast Height (DBH) and height were measured for trees with a DBH above 5 cm. For each species or forest type within a plot, AGB was estimated in t/plots using species-specific and forest type-specific allometric equations from (Gregorioet

al., 2005). As showed in Equation 2, the provided field based AGB was then converted from t/plots into t/ha.

$$Area\ plot\ _{(ha)} = \frac{\pi r^2}{10000}$$

$$AGB\ _{(t/ha)} = \frac{AGB\ _{(t/plot)}}{Area\ plot\ _{(ha)}}$$

Equation 2. Above Ground Biomass.

The field-based AGB ranges from a minimum of 0.35 t/ha to a maximum of 274.50 t/ha, with a mean value of 92.49 t/ha; the distribution appears right-skewed (Figure 2). It must be noted that the values on the vertical axis in Figure 2 are referring to the status of data after preprocessing, that is once the forestry plots - now represented as polygons - are transformed into a feature point class where each plot results in 6 to 7 points; more information on this can be found in the following section.

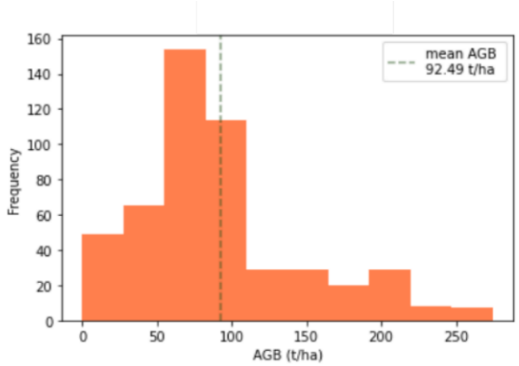


Figure 2. Distribution of measures Above Ground Biomass.

3.3.3. Sentinel-2 collection and pre-processing

From the Copernicus Open Access Hub[3], 2 images covering 2 seasons were downloaded: Summer (August 2015) and Autumn (November 2016). Both dates are close to the time the forestry's inventory measures and have a cloud coverage of less than 5% with clouds absent in the spatial subset where the study area is located. A single Sentinel 2A Level 1C product is an ortho-image provided in the UTM/WGS84

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

projection, and composed of 13 spectral bands: visible, NIR, red-edge, and shortwave infrared (SWIR), which spatial resolution varying from 10, 20 and 60 m.

The first preprocessing steps were carried out using license free tools provided by the European Spatial Agency (ESA). First, to adjust the images from Top of the Atmosphere (ToA, L1C) to Bottom of the Atmosphere (BoA, L2A) reflectance, Sen2Cor was used. This plugin allows for atmospheric, terrain, and cirrus correction. Second, using the Sentinel Application Platform (SNAP), red-edge and SWIR bands were re-sampled from 20 to 10 m using the nearest neighbor method. The three bands with a spatial resolution of 60 m (band 1, 9 and 10) were excluded from the analysis by using the subset tool in SNAP, as these are mostly used for climate and atmospheric related studies. Finally, the remaining bands were clipped to the extent of the study area.

3.3.4. Vegetation Indices and biophysical parameters generation

Vegetation Indices (VI) and biophysical parameters have been proved to increase the performance of regression algorithms for AGB estimation in different ecological zones and forest types (Forkuor, et al., 2020; Pandit, et all, 2018; Galidaki, et al., 2017). With the purpose of evaluating the performance of several feature selection methods, a wide range of VIs was generated using the calculator in SNAP. Care was taken in including VIs which required SWIR and red-edge bands, as shown in Table 1. The VIs were calculated for both dates, that is August 2015 and November 2016.

Furthermore, 5 biophysical parameters (Table 2) - Leaf Area Index (LAI), Canopy Water Content (LAI cwc), Canopy Chlorophyll Content (LAI cab), Fraction of absorbed photosynthetically active radiation (FAPAR) and Fraction of vegetation cover (FCOVER) - were calculated for each image by using the biophysical processor in SNAP. Such variables have been found to enhance the estimation of biomass by describing spatial distribution and dynamics of vegetation (Forkuor, et al., 2020).

Table 1. Sentinel-2 generated Vegetation Indices.

Index	Definition	Reference
NDVI	(B8-B4)/(B8+B4)	(Rouse et al., 1973)
GNDVI	(B8-B3)/(B8+B3)	(Gitelson et al., 1996)
SAVI	[(1+L*)(B8-B4)]/(B8+B4+L*)	(Huete, 1988)
MSAVI	[B8+1-$\sqrt{(2B8+1)^2-8(B8-B4)}$]/2	(Huete, 1994)
GEMI	n**(1-0.25n**)-(B4-0.125)/(1-B4)	(Pinty and Verstraete, 1992)
NDVIre1	(B8-B5)/(B8+B5)	(Delegido et al., 2011)
NDVIre2	(B8-B6)/(B8+B6)	(Delegido et al., 2011)
NDVIre3	(B8-B7)/(B8+B7)	(Delegido et al., 2011)
Clre	(B7/B8)-1	(Gitelson et al., 2003)
NDWI	(B8-B12)/(B8+B12)	(Bo-cai, 1996)

$L^* = 0.5$, $n^{**} = (2(B8^2 - B4^2) + 1.5B8 + 0.5B4)/(B8 + B4 + 0.5)$

Table 2. Sentinel-2 generated biophysical parameters.

Biophysical parameter	Definition
LAI	Leaf Area Index
LAIcw	Canopy Water Content
LAIcb	Canopy Chlorophyll Content
FAPAR	Fraction of Absorbed Radiation
FCOVER	Fraction of vegetation Cover

3.3.5. Texture measures generation

Once several spectral variables were generated - VIs and biophysical parameters -, further spatial predictors were included. This allows to consider not only the spectral response of different surfaces, but also the spatial relationships among these surfaces.

For this purpose, texture measures derived from the Gray Level Co-occurrence Matrix (GLCM) were included, as they have been widely used for enhancing remote sensing-based classification and regression forestry-related problems (Pandit et al., 2019;

Kelsey & Neff, 2014). As Hall-Beyer (2017a) points out in her tutorial on GLCM texture, a GLCM is not an image, it is rather a tabulation expressing how often different combinations of Digital Numbers (Gray Levels) occur in an image band, showing all the possible combinations of value pairs and their frequency. This table is constructed by using each and every pixel of the image (reference pixel) and considering its neighboring pixel or pixels (neighbor pixel).

Texture measures derived from the GLCM consider the relationship between 2 pixels (GLCM) in the original image, as opposite to first order texture measures which are calculated directly from the original image pixels values without considering how this are related to one another (Haralick et al., 1973); most of GLCM derived measures used in Remote Sensing come from a series of papers of Haralick and colleagues in the 60s. GLCM derived texture measures do not have a single way to be classified, hence, we are going to use the division used by Hall-Beyer in her tutorial (Hall-Beyer, 2017a); in this work, these measures are divided into 3 categories depending on the weights in the equations: measures related to contrast, measures related to orderliness and GLCM descriptive statistics.

Pandit (2019) explored the use of Sentinel-2 extracted GLCM texture measures in AGB estimation, and concluded that GLCM mean, variance, and dissimilarity from band 2 (blue), with a window size of 7x7, yield the best AGB predictor for tropical and subtropical forests dominated by *Shorea robusta* and *Pinus roxburghii*, located in Parsa National Park, Nepal. In another study, AGB was estimated in the San Juan National Forest located in the southwest of Colorado; the area is characterized by *Ponderosa Pine woodlands*, Warm-Dry Mixed Conifer forests, Cool-Moist Mixed Conifer forests, and *Spruce-Fir* forests. In order to estimate AGB in such a context, a GLCM was derived from Landsat TM band 2; entropy, mean, and correlation were found to be the best predictors for that region (Kelsey & Neff, 2014). Thus, the importance of GLCM measures can vary based on sensor and study area, however, to minimize correlation among measures, Hall-Beyer suggests using one measure from each category.

In the following study, 3 different GLCM texture measures were chosen, for a total of 12 new features to be tested. These are divided as follows: contrast, entropy, and GLCM-mean were derived from both Sentinel-2 band-2 (blue) and NDVI generated for August 2015 and November 2016 to assess whether images from a certain season might provide better results (Table 3).

Table 3. Sentinel-2 generated texture measures.

Texture measure	Definition	Original band
Entropy	$\sum_{i,j=0}^{N-1} iP_{i,j}(-lnP_{i,j})$	Band 2 NDVI
Contrast	$\sum_{i,j=0}^{N-1} iP_{i,j}(i-j)^2$	Band 2 NDVI
GLCM-Mean	$\sum_{i,j=0}^{N-1} iP_{i,j}$	Band 2 NDVI

P (i, j) is the normalized co-occurrence matrix such that SUM (i, j = 0, N-1) (P(i, j)) = 1

(Pandit et al., 2019)

3.3.6. Topographic data collection and pre-processing

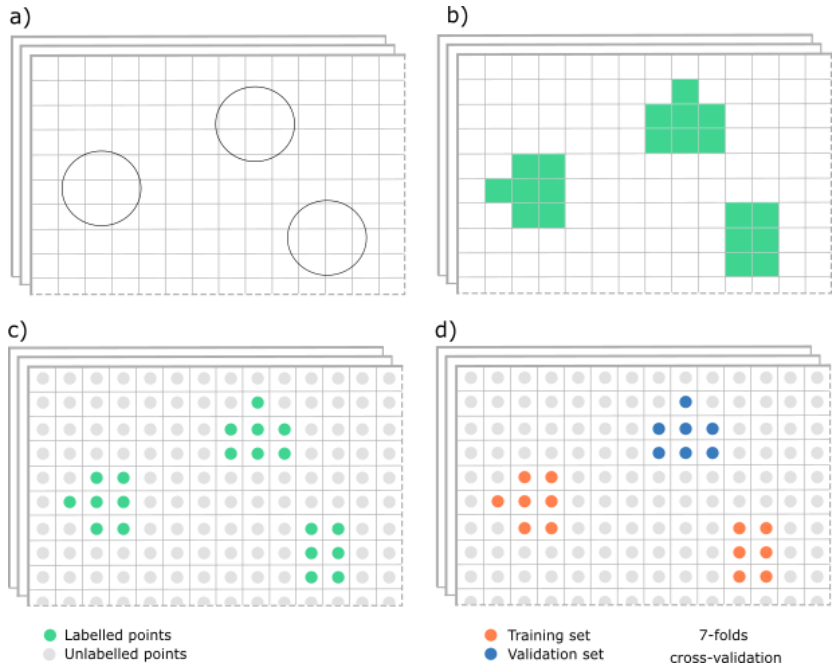
The European Digital Elevation Model (DEM) and derived slope were downloaded in the section Imagery and Reference Data of the Copernicus website. These products have a spatial resolution of 25m. DEM and slope were re-sampled to the same spatial resolution as the Sentinel-2 images (10m) by making sure that cell size and cell positioning matched. DEM up-sampling is carried out in ArcGIS Pro using the nearest neighbor assignment method, since it does not alter the input cell value. Furthermore, DEM and slope are clipped to the study area extent.

3.3.7. Methodology workflow

The predictor feature space consists of 63 features from the following categories: VIs, biophysical parameters, texture measures, spectral bands, DEM and slope. All features were transformed from their original raster format to vector format (points) using the Raster to Point (Conversion) tool in ArcGIS Pro. Similarly, the circular forestry plots were rasterized to the Sentinel-2 10 m grid, as represented in Figure 3 a) and b), and they were given pixel values corresponding the plot AGB. Subsequently, the Extract multiple values to points tool was used to create a final feature dataset with associated table containing a total of 65 columns, of which 63 represent each of the generated features, 1 represents the field based AGB values, and the last one represents the geometry field. This last column allows for the feature point to be transformed back into a raster format to subsequently generate the final maps.

The 63 predictor features were normalized using the RobustScaler from Scikit-learn[4]. The RobustScaler normalizes the dataset according to the interquartile range, such an approach offers a good handle of outliers compared to mean- and variance-based normalization methods (Pedregosa, et al., 2011). This step is necessary as the features differ in both range and unit of measure and certain regression models and feature selection methods tend to give more importance to high cardinality and continuous features.

A 7-folds cross-validation constraining samples from single plot to be either only in the training or in the validation set (Figure 3, c) and d)) was performed Secondly, only the training set is used to fit the scaler, this allows validation data to remain unseen throughout the process. The cross-validation is used before each step, that is feature selection, AGB modelling, and hyper-parameters fine-tuning.



Circular forestry plots in vector format (polygon) (a). Rasterization of the forestry plots (b), vectorization to points of each layer including forestry plots (c), and division in training and validation set (d).

Figure 3. Rasterization and cross-validation.

[4] https://scikit-learn.org/stable/index.html

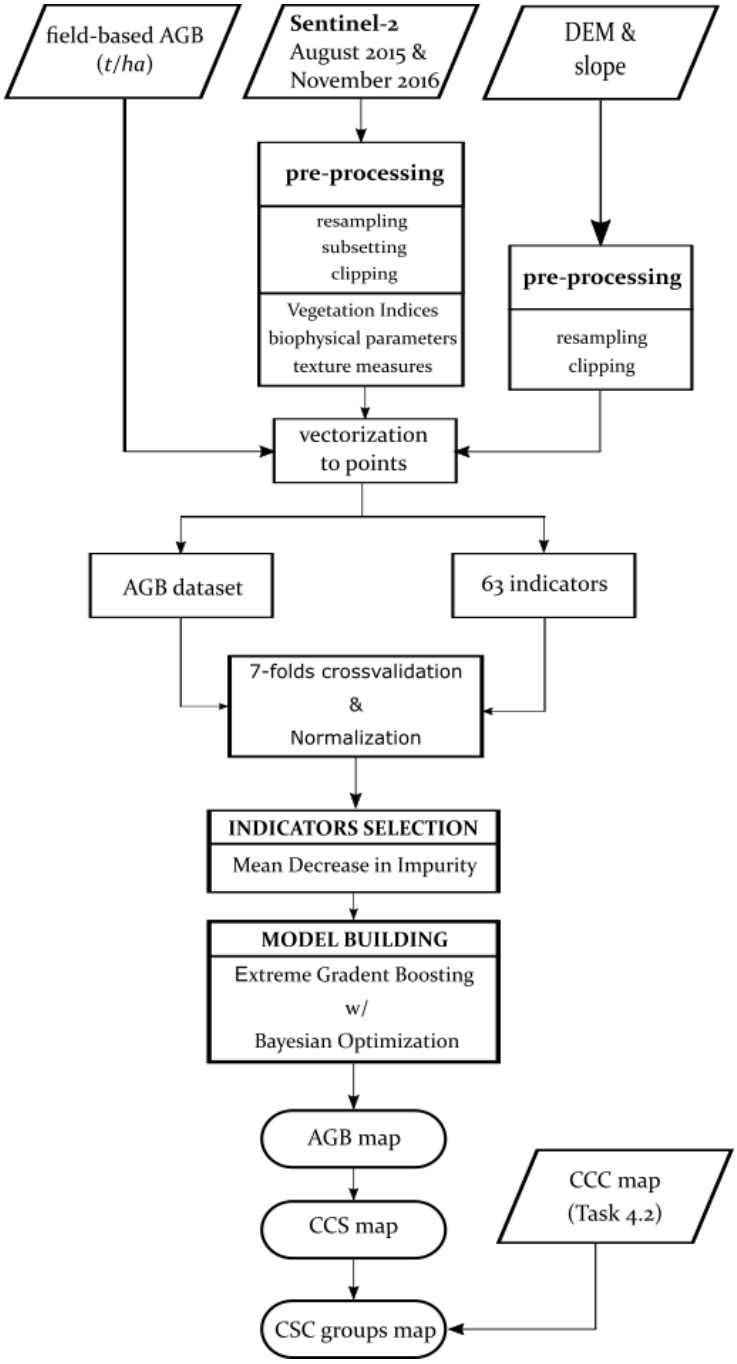


Figure 4. Methodology workflow.

Subsequently, an indicator (or feature) selection method is introduced to generate a ranking for the indicators. For this, the Mean Decrease in Impurity (MDI) measure is implemented; this is a supervised and simple to implement feature selection method derived from the Random Forest algorithm. Furthermore, the influence of the generated features is going to also be analyzed through a frequency table and a model explainer.

Therefore, once the indicators are ranked, a predictive model is going to be built. For predicting AGB within the pilot site, the Extreme Gradient Boosting algorithm is implemented and finetuned using a Bayesian Optimization method. Finally, the model performance is evaluated through its Root Mean Squared Error (RMSE). The RMSE is calculated using the Equation 3, where y_p is the predicted AGB of a n_i point, y_o is the observed AGB of the n_i point, and n is the number of validation points.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_p - y_o\right)^2}$$

Equation 3. Root Mean Squared Error.

Once the Above Ground Biomass (AGB) is estimated, the Current Carbon Sequestration (CCS) is calculated using a conversion coefficient of 0.47, as suggested by the IPCC (2006) for our forest type. Finally, the Carbon Sequestration Capacity (CSC) is being evaluated as showed in Equation 1 (Liu, 2012; Keith, 2009; Khan et al., 2020); hence, the value for the Carbon Carrying Capacity (CCC) of our study area was estimated in task 4.2 of the ***MAIL*** project.

3.4. Implementation & Results

3.4.1. Relevance of Indicators

One of the objectives of this task is to identify effective indicators (or features) which can help enhance the prediction of Above Ground Biomass (AGB). Therefore, a deeper analysis on the performance of the selected indicators is done. The following section contains 3 rankings of the indicators. The first represents the feature importance given by a feature selection method (Figure 4), the second one illustrates the frequency of selection by 4 predictive models (Table 4). Finally, the last ranking shows the impact that the indicators selected by Extreme Gradient Boosting (XGB) algorithm have on the model outputs (Figure 6Figure 6. SHAP summary plot.).

The Random Forest (RF) algorithm provides 3 measures of impurity: Gini index, entropy, and variance. Variance, or residual sum of squares, is used to measure node impurity in regression problems, and it represents the total reduction of the variance of the target variable due to the split of a certain feature at the node (Lewinson, 2019). Impurity measures can be used for feature selection by evaluating the extent to which

each feature contributes to decreasing the averaged impurity in each tree composing the forest (Lewinson, 2019), so as to calculate the Mean Decrease in Impurity (MDI) for each feature. The feature able to account for more variance decrease is going to be at the top of the ranking (Lewinson, 2019). Therefore, the MDI can be seen as the total decrease in node impurity from splitting on the variable, averaged over all trees (Hong et al., 2016).

Figure 5 shows the MDI that each indicator brings to the model. The first 9 selected indicators, apart from the DEM, were all extracted from the summer image. Furthermore, biophysical parameters were often selected as top-ranked indicators, specifically Canopy Chlorophyll Content (LAIcb), Canopy Water Content (LAIcw), and a chlorophyll index calculated using red-edge bands (Clre).

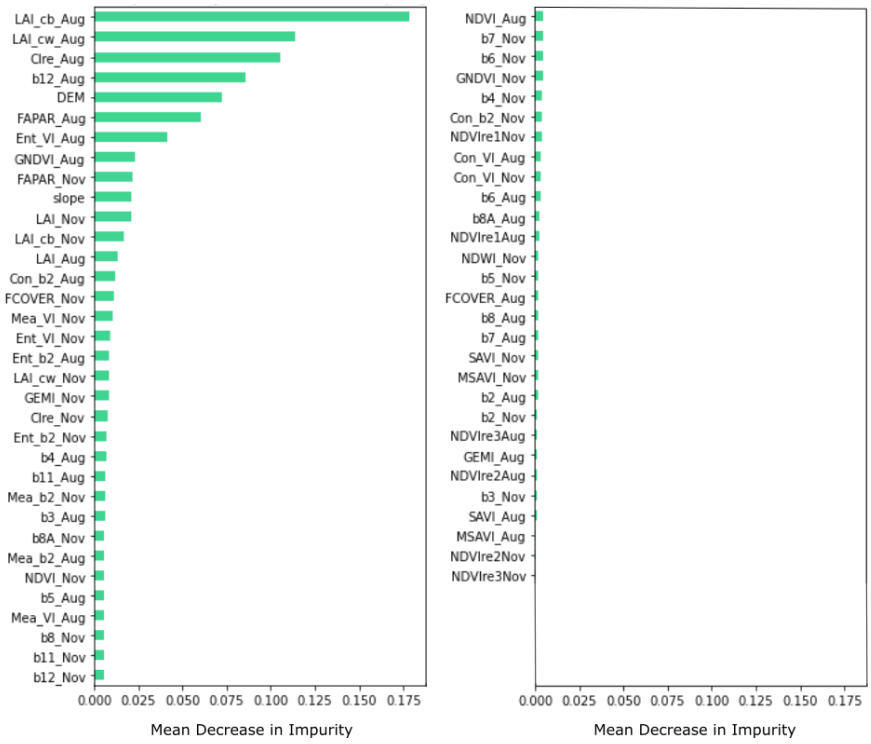


Figure 5. Indicators selection with Mean Decrease in Impurity.

Therefore, a total of 4 Machine Learning models were tested to evaluate their performance in predicting AGB when different indicators are used as model inputs. The indicators used to build the best 4 models are showed in Table 4. Hence, the frequency of selection of each indicator is illustrated. A total of 3 out of 63 indicators were chosen by all the models, these are the Digital Elevation Model (DEM), the SWIR band (B12) and the biophysical parameter (LAI cb), both extracted from the summer month; 75% of

the models also included other two indicators from the summer month, which are another biophysical parameter (LAI cw) and a vegetation index based on red-edge bands (Clre). Also, 16 indicators were included in the making of the best models for the Extreme Gradient Boosting (XGB) and K-Nearest Neighbor (kNN) algorithms. Further 5 indicators were chosen by only 25% of the models. Finally, a total of 37 out of 63 generated indicators were chosen by none of the tested models.

Table 4. Frequency of selection.

Feature	Month	kNN	RF	XGB	DNN	Frequency of Selection (%)
LAI cb	Aug	x	x	x	x	100
B12	Aug	x	x	x	x	100
DEM	-	x	x	x	x	100
LAI cw	Aug	x		x	x	75
Clre	Aug	x		x	x	75
FAPAR	Aug	x		x		50
Ent VI	Aug	x		x		50
GNDVI	Aug	x		x		50
FAPAR	Nov	x		x		50
slope	-	x		x		50
LAI	Nov	x		x		50
LAI cb	Nov	x		x		50
LAI	Aug	x		x		50
Con b2	Aug	x		x		50
FCOVER	Nov	x		x		50
Mean VI	Nov	x		x		50
Ent VI	Nov	x		x		50
Ent b2	Aug	x		x		50
LAI cw	Nov	x		x		50
GEMI	Nov	x		x		50
Ent b2	Nov	x		x		50
Clre	Nov			x		25
B3	Aug	x				25
B4	Aug			x		25
B11	Aug	x				25
Mean b2	Aug	x				25
Others	-					0

The SHapley Additive exPlanations (SHAP) package was used as model explainer for the Extreme Gradient Boosting (XGB) model. SHAP is a local feature importance method, meaning that a local feature importance is calculated for every observation. This is performed by holding out the feature value before predicting each instance, which is then repeated for each feature and each instance of the entire training set so as to measure the local importance of each feature. The computation of each instance

for a big dataset can become time consuming, however, in 2020 a fast and precise algorithm was created for tree-based models (Lundberg, 2020).

A global feature importance can be obtained by aggregating the local feature importance of each instance, as shown in the summary plot in Figure 6, where each dot represents an observation of the dataset. This plot illustrates how the model predictions were influenced by each feature. More specifically, it shows the features contribution, for each instance, in pushing the model output from a base value to the output value; where the base value is defined as the average model output over the training set (Lundberg, 2020). When a feature has negative SHAP value, the dot representing that instance is found on the left-side of the plot and it means that that feature value pushed the prediction for that instance (or dot) to be lower than the base value; on the other hand, if a dot is found on the right-hand side of the plot, the observed feature value pushed the model output to be higher than the base value. Furthermore, for each feature, overlapping points are visualized in the y-axis direction so to give an idea of the feature values distribution. Additionally, the color of the dots refers to the features value.

By observing the summary plot in Figure 6 the indicators can be divided into 2 groups:

1. Indicators which increasing in values pushed the model to output AGB higher than the base value – belonging to the first group are: The Chlorophyll index based on red-edge bands from the summer month (ClreAug), the Fraction of absorbed radiation from both months (FAPARAug and FAPARNov), the Canopy Chlorophyll Content from the Summer month (LAIcbAug), and the slope;
2. Indicators which increasing in values would push the model to output AGB lower than the base value – belonging to this second group are: band 12 from the summer image (b12Aug) and the DEM belong to the second group.

The remaining indicators do not show any clear visual pattern in the way they impacted the model output.

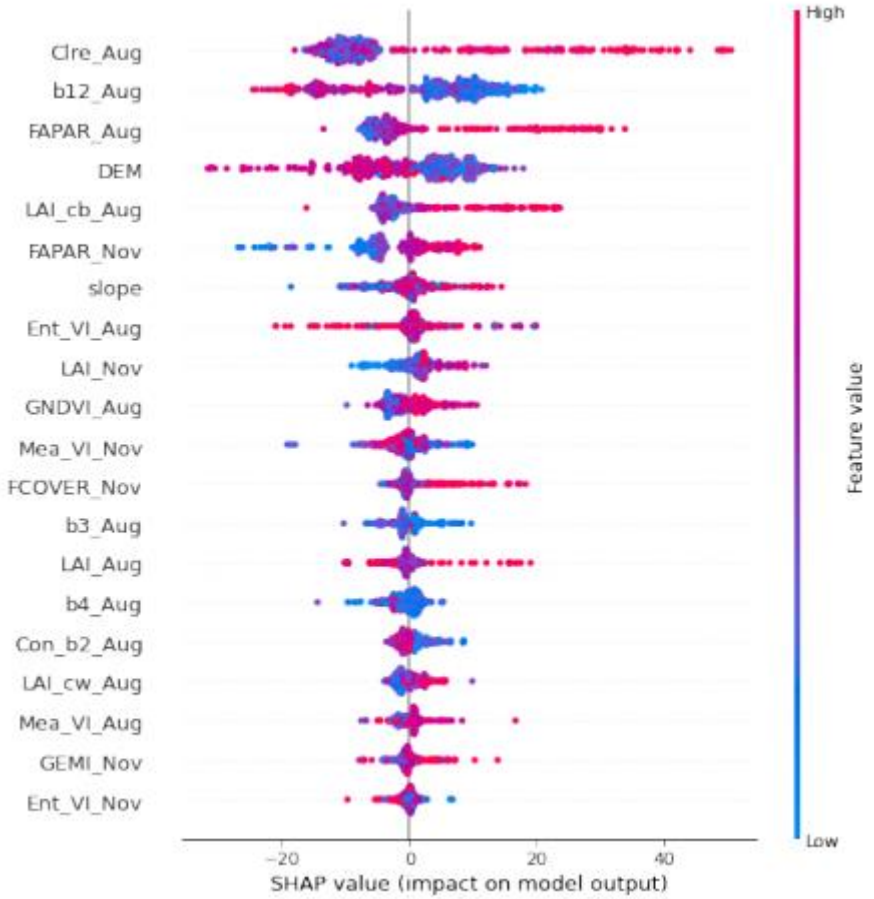


Figure 6. SHAP summary plot.

In conclusion, the 3 implemented approaches provided insights on which of the generated indicators are important for biomass prediction. Hence, these findings will help guide the making of National and European carbon maps in the framework of the ***MAIL*** project.

3.4.2. Mapping of AGB, CCS and CSC groups

The following Section presents the Above Ground Biomass (AGB) map, the Current Carbon Sequestration (CCS) map, and a final Carbon Sequestration Capacity (CSC) groups map.

The AGB map depicted in Figure 8 was generated by using the prediction from the XGB algorithm. This model was built by using the 23 top-ranked features selected by the MDI measure, and hyper-parameters optimization with Bayesian Search. The model was evaluated using 7-folds cross-validation and led to an error of 37.79 t/ha, an

estimated average AGB value of 83 t/ha, a minimum of 0 t/ha, a maximum of 346.56 t/ha and a standard deviation of 51.3 t/ha (Figure 9).

Moreover, the goodness of fit was visually evaluated. In Figure 7 the measured AGB values were plotted together with the best model predictions. Values lower than 40 t/ha and higher than 160 t/ha were over- and underestimated, respectively.

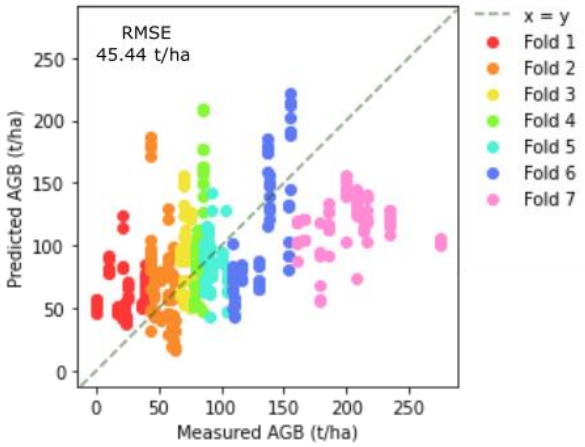


Figure 7: Goodness of fit.

The AGB map was generated by using 6 manual classes; the majority of the territory located on the south-west of map is covered by AGB values ranging from 0 t/ha to 100 t/ha, with spread high values ranging from 101 t/ha to 200 t/ha. On the other hand, the second portion of the study area, located on the north-east side of the map, is characterized by higher AGB values, with areas reaching over 250 t/ha.

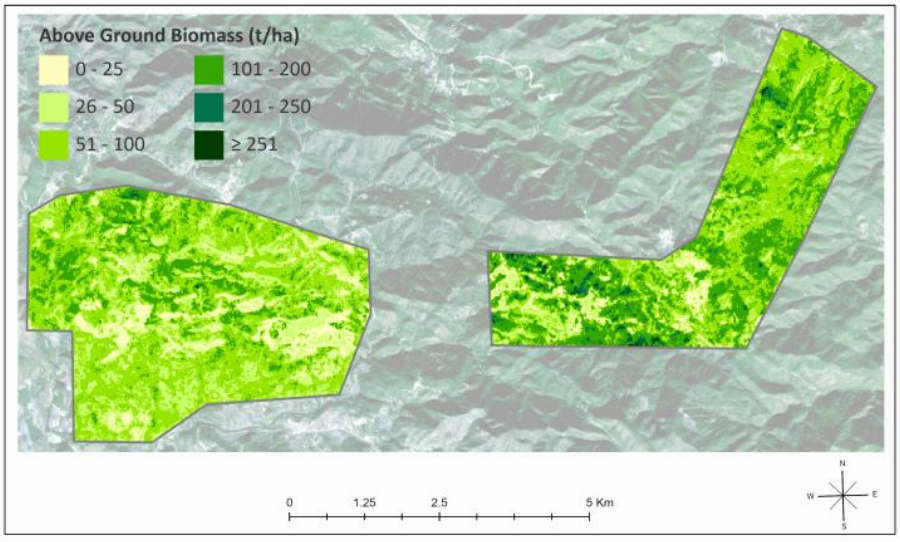


Figure 8: Above Ground Biomass map.

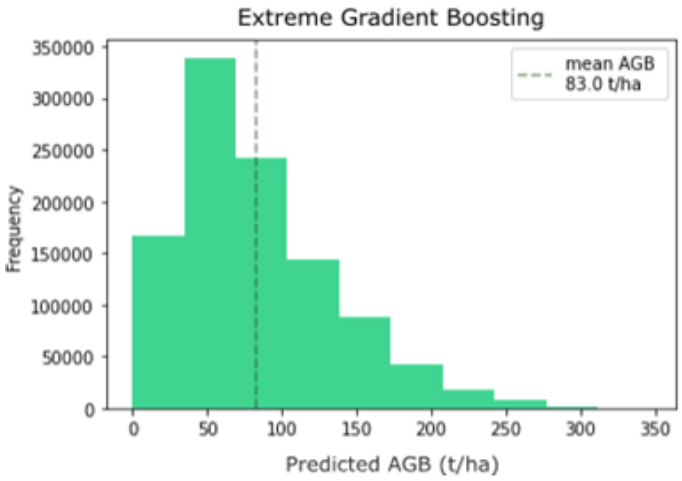


Figure 9: Distribution of predictions.

Hence, a CCS map was generated by using the conversion factor suggested by the IPCC (2006), that is multiplying the predicted AGB values by 0.47. In Figure 10, the portion of the study area located on the south-west of map is sequestrating 0 t/ha to 40 t/ha, with spread higher values ranging from 41 t/ha to 80 t/ha. On the other hand, the second portion of the study area, located on the north-east side of the map, is characterized by higher CS, with areas storing over 120 t/ha.

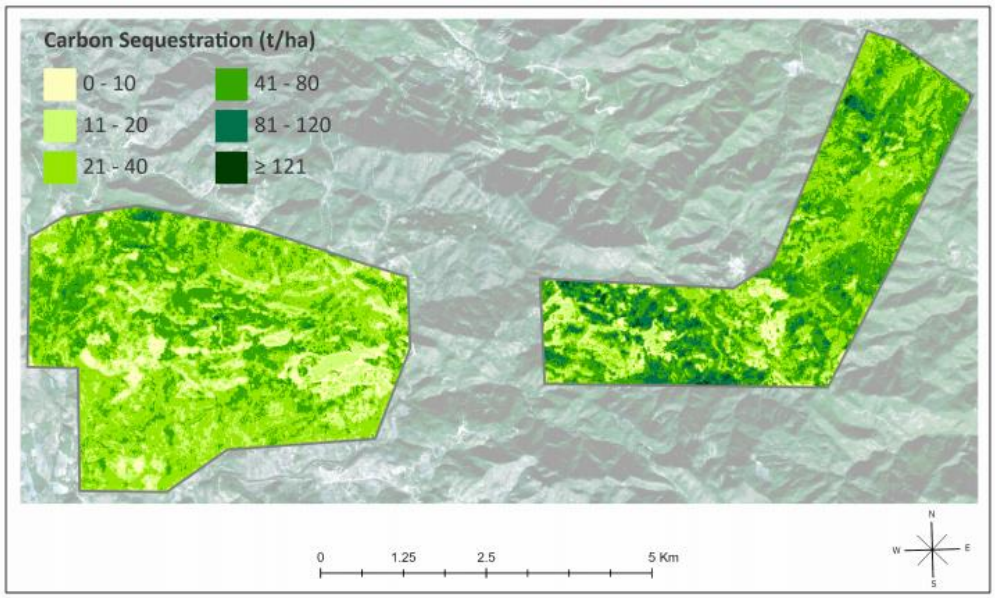


Figure 10: Current Carbon Sequestration map.

Deliverable 2.3 defined 3 levels of marginality. These are Marginal Lands with high plantation suitability (ML1), Marginal Lands with low plantation suitability (ML2), and potentially unsuitable lands (ML3). Hence, within Serra de Espadan, all 3 types of marginality were identified. As illustrated in Figure 11, most of the study area is considered unsuitable for plantation, therefore its carbon sequestration could not be increased through reforestation practices. Hence, small patches were identified as of low suitable for reforestation, while a south-west portion of the area was considered to be highly suitable. Such levels of marginality are due to environmental limitations rather than economic or social.

In the making of Deliverable 4.2, a reforestation scenario for Serra de Espadan was planned out, and the Carbon Carrying Capacity (CCC) for highly suitable (ML1) and low suitable (ML2) areas was estimated. Hence, the species selection prioritized the most resistant species to adverse ecological factors over more resource-demanding species. For this reason, coniferous were chosen for the reforestation proposal. Specifically, a mixture of *Pinus pinaster* in 70-80% and *Pinus halepensis* in 30-20% was suggested for both levels of marginality. Hence, the CCC was estimated using forest growth tables and values of future biomass and carbon capacity of the selected species after 50 years from plantation. For ML1, the CCC after 50 years was estimated to be 94.2 t/ha, whilst in ML2, this value was predicted to be 55.2 t/ha.

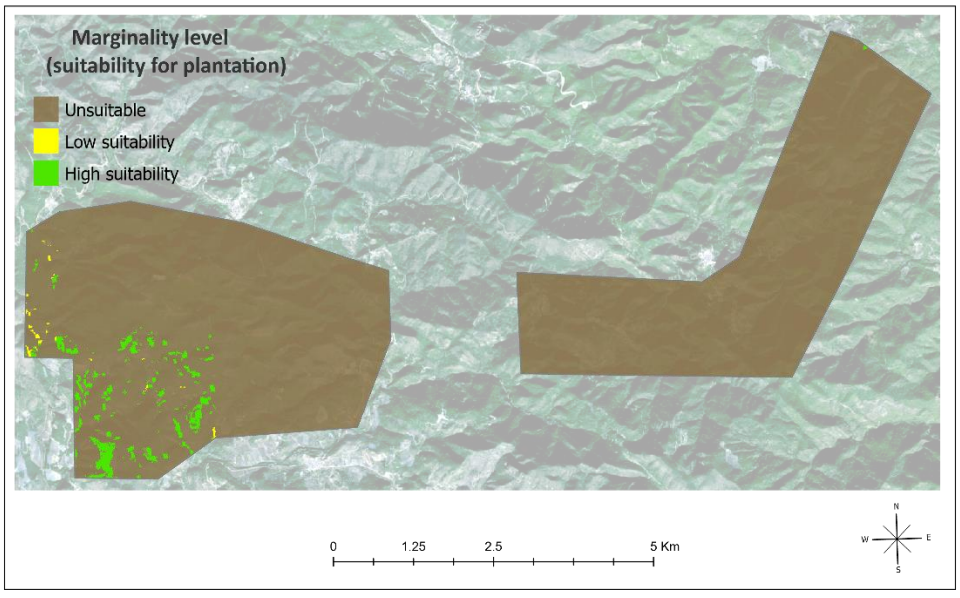


Figure 11: Marginality levels.

Therefore, the Carbon Sequestration Capacity (CSC) was estimated for both, highly suitable (ML1) and low suitable areas (ML2), by following Equation 1, and using the previously predicted Current Carbon Sequestration (CCS), as well as the 2 Carbon Carrying Capacity (CCC) values proposed in Deliverable 4.2.

Figure 12 shows the CSC groups. Group I (high suitability for plantation) is concentrated on the south-west of the pilot site, and their CSC varies from 9.64 to 94.14 t/ha, with an average of 72.01 and a standard deviation of 9.62 t/ha. Group II represents a very small part of the area and its CSC ranges from 0 to 53 t/ha, with an average of 29.32 and a standard deviation of 11.95 t/ha. Hence, the CSC for the remaining territory (Group III) was not estimated, as the rest of the pilot site was classified as unsuitable for reforestation in the context of Deliverable 2.3.

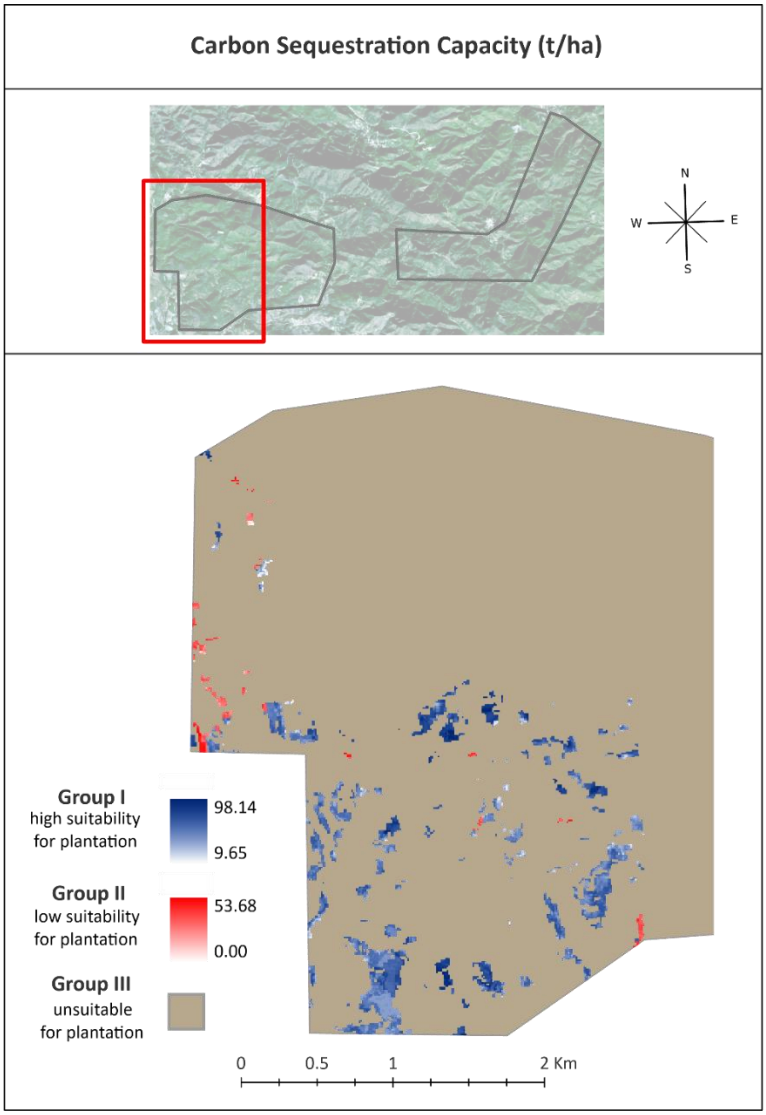


Figure 12: Carbon Sequestration Capacity Groups map.

3.5. Discussion & Conclusions

This Chapter aimed to assess the presence of Carbon Sequestration Capacity (CSC) groups in a Serra de Espadan, a Spanish pilot case which was identified as Marginal Land (ML) in the context of Task 2.3. Furthermore, within the ***MAIL*** project, this area was chosen as a study case to investigate the potential of certain indicators (Vegetations Indices, topographic measures, etc.) in improving the estimation of Above Ground Biomass (AGB). For this task, a Remote Sensing and Machine Learning based approach was proposed.

Overall, indicators or features extracted from the Sentinel-2 Summer image were selected more often than Winter features by all the tested models. Out of the 63 initially generated variables, only 23 features selected by Mean Decrease in Impurity (MDI) were used for the final model. The highest-ranking features by their importance are: biophysical parameters (LAI, LAI cw, FAPAR) from the summer month, the Digital Elevation Model (DEM), a SWIR band (band 12) and a Chlorophyll index generated from the red-edge bands (Clre), both from the summer month.

Model explanation with SHapley Additive exPlanations (SHAP) helped gathering more insights on the input features and their influence on the model output. It was found that high values of a Vegetation Index generated from the red-edge bands and biophysical parameters extracted from the summer season, lead to an increase in predicted AGB. The same positive behavior was found for certain biophysical parameters extracted from the Autumn Season. In contrast, high values of the summer SWIR band and high elevation pushed the predicted AGB to values lower than the baseline. Those findings agree with the scientific consensus on the relationship between elevation and biomass. The insights gathered with SHAP showed the utility of a model explainer for the scientific community.

Concerning the prediction of AGB, a recurrent limitation was found. For all the tested models, AGB values lower than 40-50 t/ha were slightly overpredicted whereas values higher than 150-160 t/ha were underpredicted. As it was confirmed by previous studies (Galidaki, et al., 2017; Salem et al., 2020; Forkuor, et al., 2020), this is a typical issue when estimating AGB with the use of Machine Learning and satellite images, and it is exacerbated by a limited number of representative samples for low and high values of the forestry inventory.

For those areas of the pilot case identified as suitable for plantation by Task 2.3, three groups were proposed: lands with high suitability for plantation (ML1), lands with low suitability for plantation (ML2), and unsuitable lands (ML3). Therefore, Task 4.2 proposed a reforestation scenario for the first 2 groups by estimating their Carbon Carrying Capacity (CCC) 50 years after plantation. Hence, this report delineates the Carbon Sequestration Capacity (CSC) for those areas identified as ML1 and ML2. Finally, we suggest that future reforestation projects should focus on highly suitable areas, because these show a higher CSC compared to ML2, reaching an increasing in carbon sequestration up to almost 100 t/ha after reforestation.

4. CARBON SEQUESTRATION CAPACITY GROUPS MAPPING AT EUROPEAN LEVEL

This chapter presents a new approach for classifying MLs into Carbon Sequestration Capacity groups at European scale that was developed along with a tool in GEE to be embedded in a Decision Support System in the ***MAIL*** geoportal. It also presents the datasets selected for the implementation of the task and the pre-processing required.

4.1. Methodology and Approach

The methodology that presented in this section relates in how MLs can be classified in Carbon Sequestration Capacity (CSC) groups. In order to estimate CSC for MLs and classify in CSC groups, it is crucial to estimate potential suitable species for afforestation and their Aboveground Biomass Carbon. The MLs as calculated on Task 2.3 is the basemap, where the most frequent species from neighbor forested areas, both dominant 1 and 2 species, and species' Aboveground Biomass Carbon values are assigned. Dominant 1 and 2 species of neighbor forested areas are adapted to the ecological and climatological conditions and therefore are considered to be the most suitable for afforestation projects. Also, species' Aboveground Biomass Carbon is used as indicator of the potential maximum capacity, which MLs never acquire as by default have productivity restrictions.

4.2. Google Earth Engine implementation

For the implementation of the task, due to the extent, size, and resolution of the data, as well as for compatibility with the ***MAIL*** geoportal, this task was implemented on Google Earth Engine (GEE).

GEE is a cloud-based platform for planetary-scale geospatial analysis that brings Google's massive computational capabilities to bear on a variety of high-impact societal issues including deforestation, drought, disaster, disease, food security, water management, climate monitoring and environmental protection. It is unique in the field as an integrated platform designed to empower not only traditional remote sensing scientists, but also a much wider audience that lacks the technical capacity needed to utilize traditional supercomputers or large-scale commodity cloud computing resources (Gorelick, et al., 2017). Some of the main benefits of GEE are the large data catalog in combination with massive CPU and its speed and ease of use. GEE was launched in 2010 by Google as a proprietary system, but it is free to non-commercial educational, research, and nonprofit use.

The service utilizes cloud computing to enable different formats of data to be accessed, shared and integrated. This has entailed creating not only an infrastructure with petabyte-scale capability, but also APIs, using JavaScript and Python, that enable the addition and manipulation of various data. By placing multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities and the tools needed to access, filter, perform, and export analyses in the same easy to use application, users are able to explore and scale up analyses in both space and time without any of the hassles traditionally encountered with big data analysis. Constant development and refinement have propelled GEE into one of the most advanced and accessible cloud-based geospatial analysis platforms available.

4.3. Dataset selection

4.3.1 Tree species in Europe according to European Forest Institute (EFI)

Brus et al. (2011) publish a statistical map of tree species in Europe. This map represents the spatial distribution of twenty tree species groups over Europe at 1x1Km resolution where the ICP-Forest Level-I plot data were extended with the National Forest Inventory (NFI) plot data of eighteen countries.

Basic dendrometric data were gathered for 260,000 national forest inventory plot locations from 17 countries. In areas with national forest inventory data, area proportions covered by the 20 species were obtained by compositional kriging. For the rest of Europe, a multinomial logistic regression model was fitted to ICP-level-I plots using various abiotic factors as predictors (soil, biogeographical zones, bioindicators derived from temperature and precipitation data). The regression results were

iteratively scaled to fit NUTS-II forest inventory statistics and the European Forest Map. The predictions for the twenty tree species were validated using 230 plot data separated from the calibration. Figure 13 demonstrates the aggregated results.

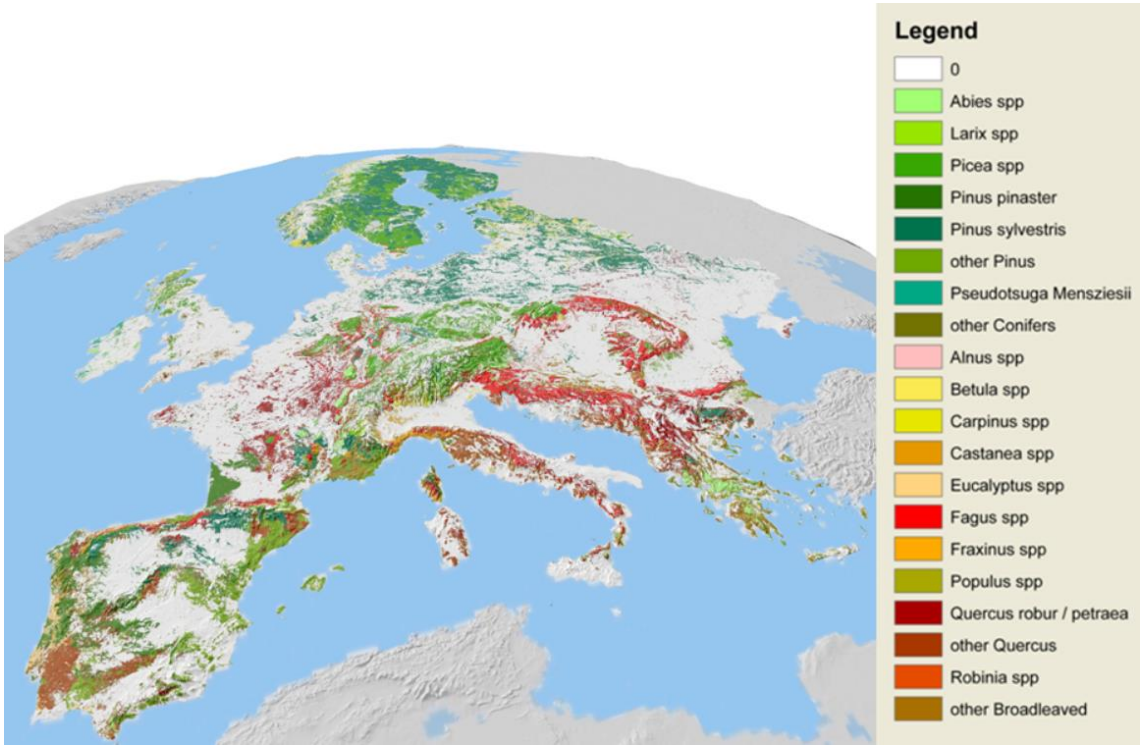


Figure 13: Aggregated results showing the dominant species at 1x1km.

(Source: https://www.efi.int/knowledge/maps/treespecies)

Table 5: Technical specifications of Tree species maps for European forests dataset.

Specification
File name: Tree species in Europe according to European Forest Institute (EFI)
Coordinate system: ETRS89 LAEA
Production date: 2012
Spatial Coverage: Europe
Spatial Resolution: 1km
Completeness: Complete
File type, format: Raster, TIFF image

4.3.2. Global Aboveground and Belowground Biomass Carbon Density Maps

Global Aboveground and Belowground Biomass Carbon Density Maps are available in Google Earth Engine and allow to distinguish between Aboveground Biomass Carbon Density (AGBC) and Belowground Biomass Carbon Density (BGBC). To harmonize the biomass for all the ecosystems around the world, layers listed in Table 1 “Data sources used to generate harmonized global maps of above and belowground biomass carbon density” used as published on Harmonized global maps of above and belowground biomass carbon density in the year 2010 (Spawn et al., 2020).

This dataset provides temporally consistent and harmonized global maps of aboveground and belowground biomass carbon density for the year 2010 at a 300 m spatial resolution. The aboveground biomass map integrates land-cover specific, remotely sensed maps of woody, grassland, cropland, and tundra biomass. Input maps were amassed from the published literature and, where necessary, updated to cover the focal extent or time period. The belowground biomass map similarly integrates matching maps derived from each aboveground biomass map and land-cover specific empirical models. Aboveground and belowground maps were then integrated separately using ancillary maps of percent tree cover and landcover and a rule-based decision tree. Maps reporting the accumulated uncertainty of pixel-level estimates are also provided (Spawn et al.,, 2020).

Aboveground living biomass carbon density includes carbon stored in living plant tissues located above the earth’s surface (stems, bark, branches, twigs). It does not include leaf litter or coarse woody debris that was once attached to living plants but have since been deposited and are no longer living. Belowground living biomass carbon density includes carbon stored in living plant tissues located below the earth’s surface (roots). This does not include dead and/or dislocated root tissue, nor does it include soil organic matter. Woody cover includes any vegetation whose biomass is primarily composed woody biomass (e.g., trees and shrubs). Herbaceous cover includes any vegetation whose biomass is primarily composed of leaf-like matter (e.g., grasses and many crops) (Spawn et al., 2020).

Table 6: Technical specifications of Global Aboveground and Belowground Biomass Carbon Density Maps.

Specification	Properties of the GeoTIFFs
File name: Global Aboveground and Belowground Biomass Carbon Density Maps	Bands: 1
Coordinate system: EPSG: 4326	Scaling: 0.1
Spatial Coverage: Global	Data units: Mg C/ha
Spatial Resolution: ~300 m (0.002777778 degree)	Data type: UInt16
Temporal Coverage: 2010-01-01 to 2010-12-31	No data value: 65536
Temporal Resolution: Annual	Map units: degree
File type, format: GeoTIFF (.tif) format	

Table 7: File names and descriptions of Global Aboveground and Belowground Biomass Carbon Density Maps.

File name	Units	Description
aboveground_biomass_carbon_2010.tif	Mg C/ha	Aboveground living biomass carbon stock density of combined woody and herbaceous cover in 2010. This includes carbon stored in living plant tissues that are located above the earth's surface (stems, bark, branches, twigs). This does not include leaf litter or coarse woody debris that were once attached to living plants but have since been deposited and are no longer living.
belowground_biomass_carbon_2010.tif	Mg C/ha	Belowground living biomass carbon stock density of combined woody and herbaceous cover in 2010. This includes carbon stored in living plant tissues that are located below the earth's surface (roots). This does not include dead and/or dislocated root tissue, nor does it include soil organic matter.

File name	Units	Description
aboveground_biomass_carbon_2010_uncertainty.tif	Mg C/ha	Uncertainty of estimated aboveground living biomass carbon density of combined woody and herbaceous cover in 2010. Uncertainty represents the cumulative standard error that has been propagated through the harmonization process using summation in quadrature.
belowground_biomass_carbon_2010_uncertainty.tif	Mg C/ha	Uncertainty of estimated belowground living biomass carbon density of combined woody and herbaceous cover in 2010. Uncertainty represents the cumulative standard error that has been propagated through the harmonization process using summation in quadrature.

4.3.3. S2GLC

S2GLC land cover map is one the most detailed pan-European land cover products. It was produced using automatic classification approach and Sentinel-2 images from 2017. It contains 13 classes with MMU equal to Sentinel-2 pixel which is 10×10 m. The overall accuracy is 86%. The product is available in two forms: mosaic for the whole Europe, Sentinel-2 tiles.

Table 8: Technical specifications of the S2CLG layer.

Specification	Source data specification
File names: 2GLC_Europe_2017_v1.2_grey.tif	Sensor: Sentinel-2
Coordinate system: ETRS89 LAEA	Data type: Thematic mapper
Production date: 2017	Sensor resolution:10 m
Coverage (top L, BR coordinates): Europe	Acquisition date:
Grid size: 10 x 10 m	Grid size: -
Position accuracy: -	Positional accuracy: -
Vertical accuracy: -	Vertical accuracy: -
Completeness: Complete	
File type, format: Raster	

Table 9: Land Cover Classes of the S2CLG layer.

BASEMAP LAYER	Land Cover Classes	
	ID	Name
Land Cover Map of Europe 2017 (S2GLC)	0	Clouds
	62	Artificial surfaces and constructions
	73	Cultivated areas
	75	Vineyards
	82	Broadleaf tree cover
	83	Coniferous tree cover
	102	Herbaceous vegetation
	103	Moors and Heathland
	104	Sclerophyllous vegetation
	105	Marshes
	106	Peatbogs
	121	Natural material surfaces
	123	Permanent snow-covered surfaces
	162	Water bodies

4.4. Dataset pre-processing

In order to apply the methodology for mapping CSC groups at European level, all three datasets needed necessary data preprocessing, such as handling missing or null values, remove errors and data transformation to make the various data consistent.

The datasets imported as assets in Google Earth Engine, reprojected in EPSG:3035 projection and resampled to 10m resolution, to match soft layers as produced on Task 2.3.

Tree species maps for European forests, consist of 20 tree species rasters with values from 0 to 100 which is the percentage share of the respective tree species from land area. These rasters were reclassified by a 5% interval, combined in one raster and processed in order to detect the dominant 1 and 2 species. Finally, 4 singleband rasters were produced: Dominant 1 Species, Dominant 1 Percent, Dominant 2 Species and Dominant 2 Percent as shown on Figure 14.

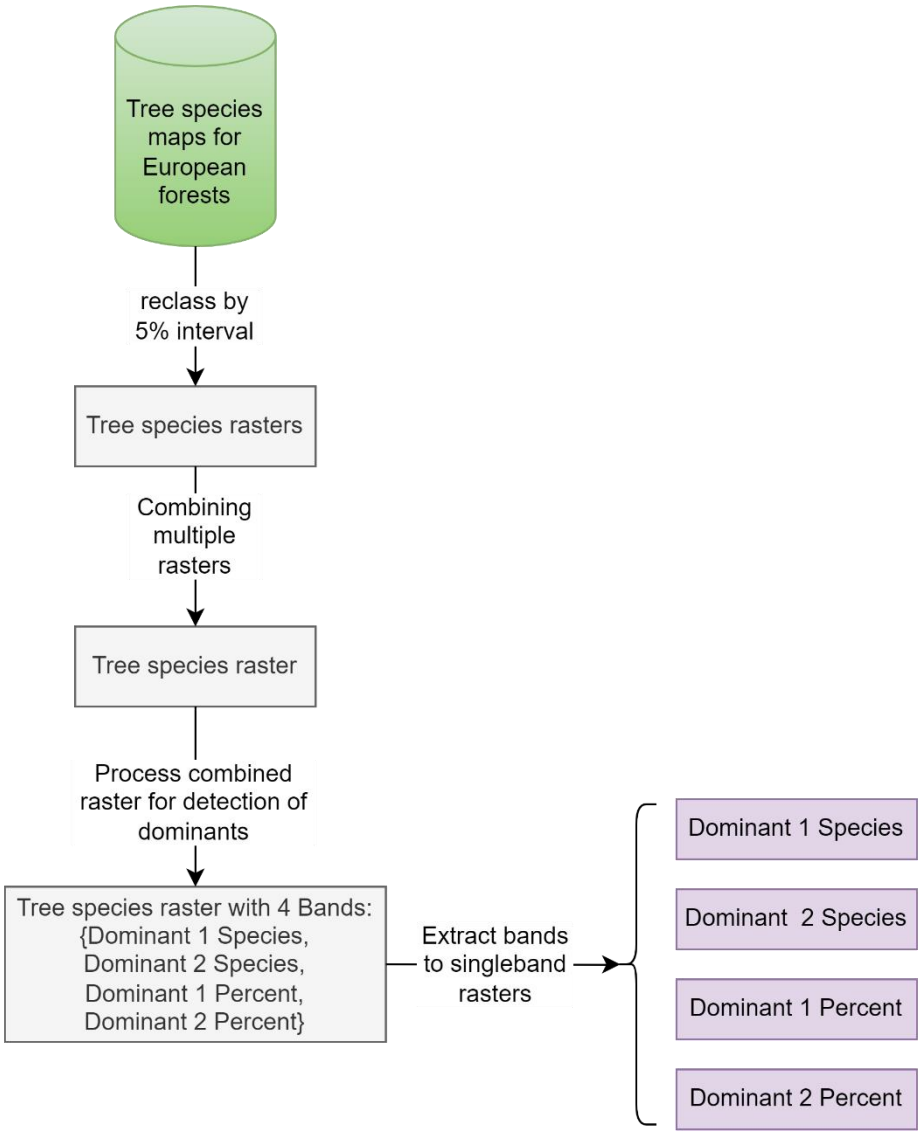


Figure 14: Tree species maps for European forests workflow

Because of the original resolution of Tree species maps for European forests and the limited extent, in order to cover all forested areas, `focal mode` was applied, a morphological operation implemented on Google Earth Engine, configured and repeated in ranges from 500m to 30km, both for Dominant 1 and 2 Species. The results from both operations were composed into new images, using mosaic, with calculated images arranged from 30km to 500m, to preserve accuracy, as the last image is on top.

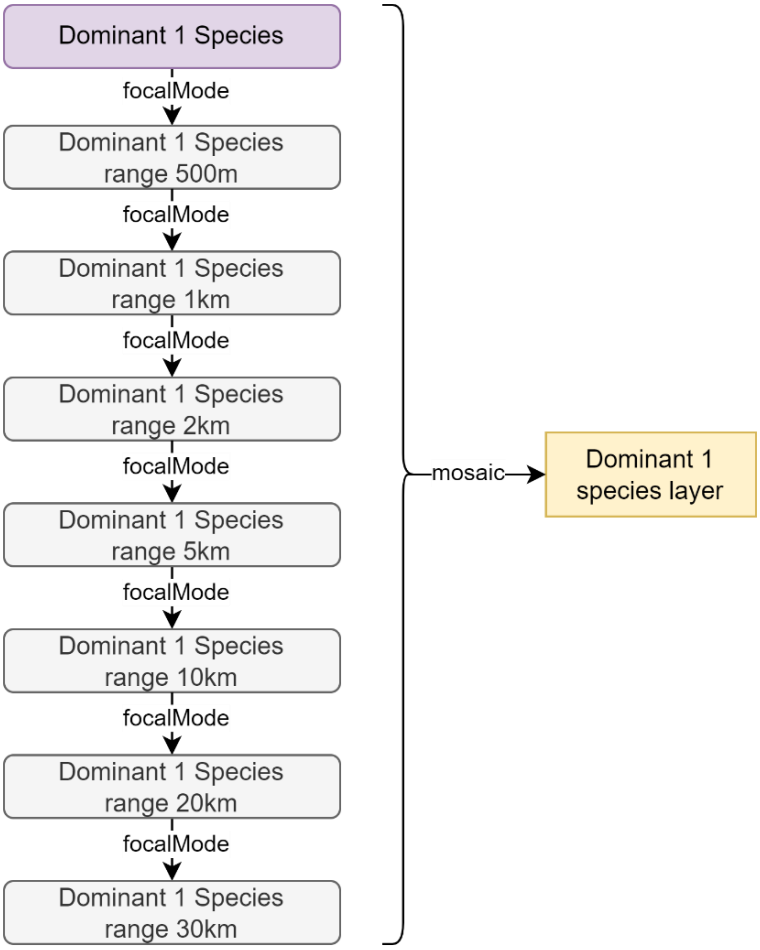


Figure 15: Dominant 1 Species workflow

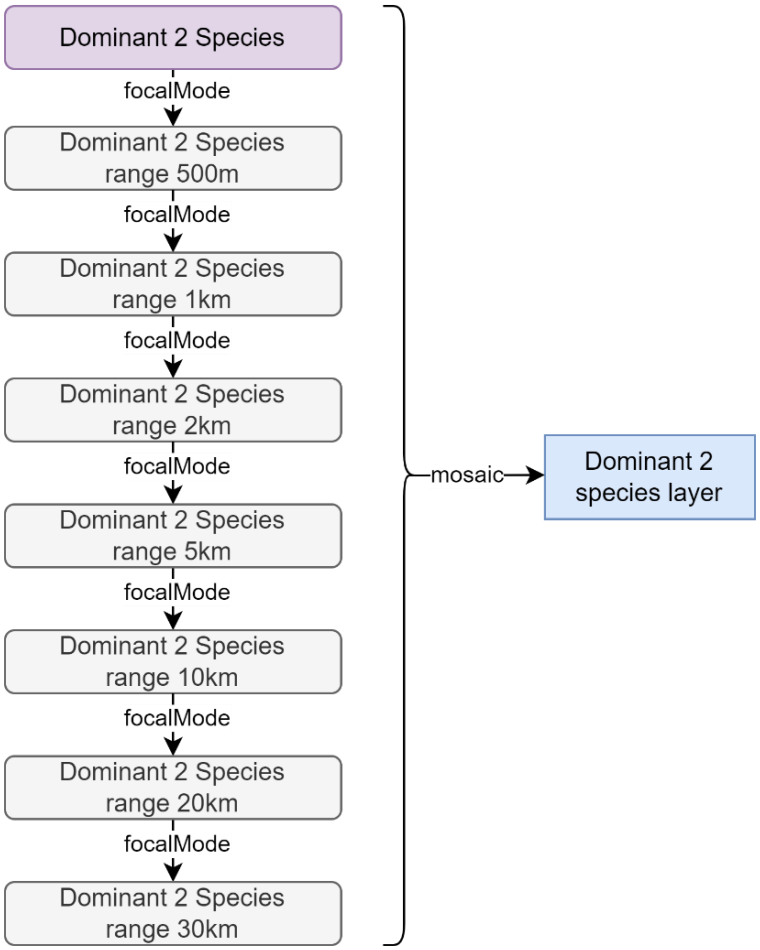


Figure 16: Dominant 2 Species workflow

The same workflow applied also on S2GLC data in order to combine it with Dominant 1 species layer and provide info for forested areas with no data in the EFI Tree species maps like areas in Cyprus, areas near coastline and some other islands.

Firstly, the tree cover pixel values selected, conifers and broadleaves, by masking the layer with these pixel values and then `focal mode` was applied in ranges from 100m to 100km. The results were composed into a new image, using mosaic, with calculated images arranged from 100km to 100m and then reclassified to match Dominant 1 species classes. Finally, Tree species maps were combined with S2GLC data in order to produce a tree species map for all European forested areas.

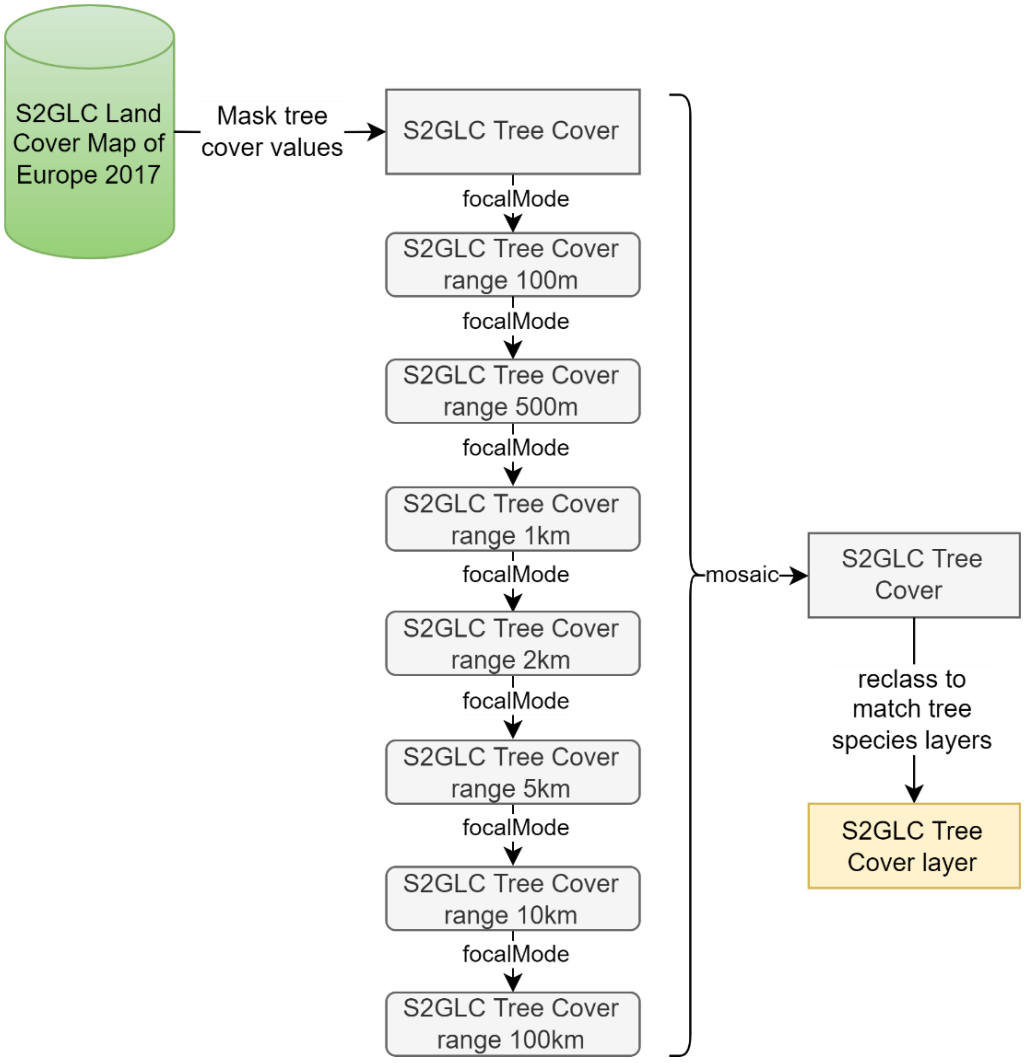


Figure 17: S2GLC Land Cover Map of Europe 2017 workflow

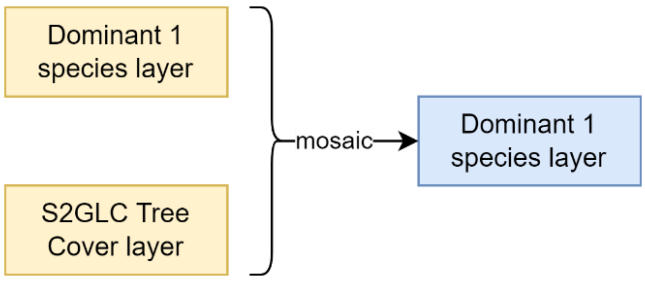
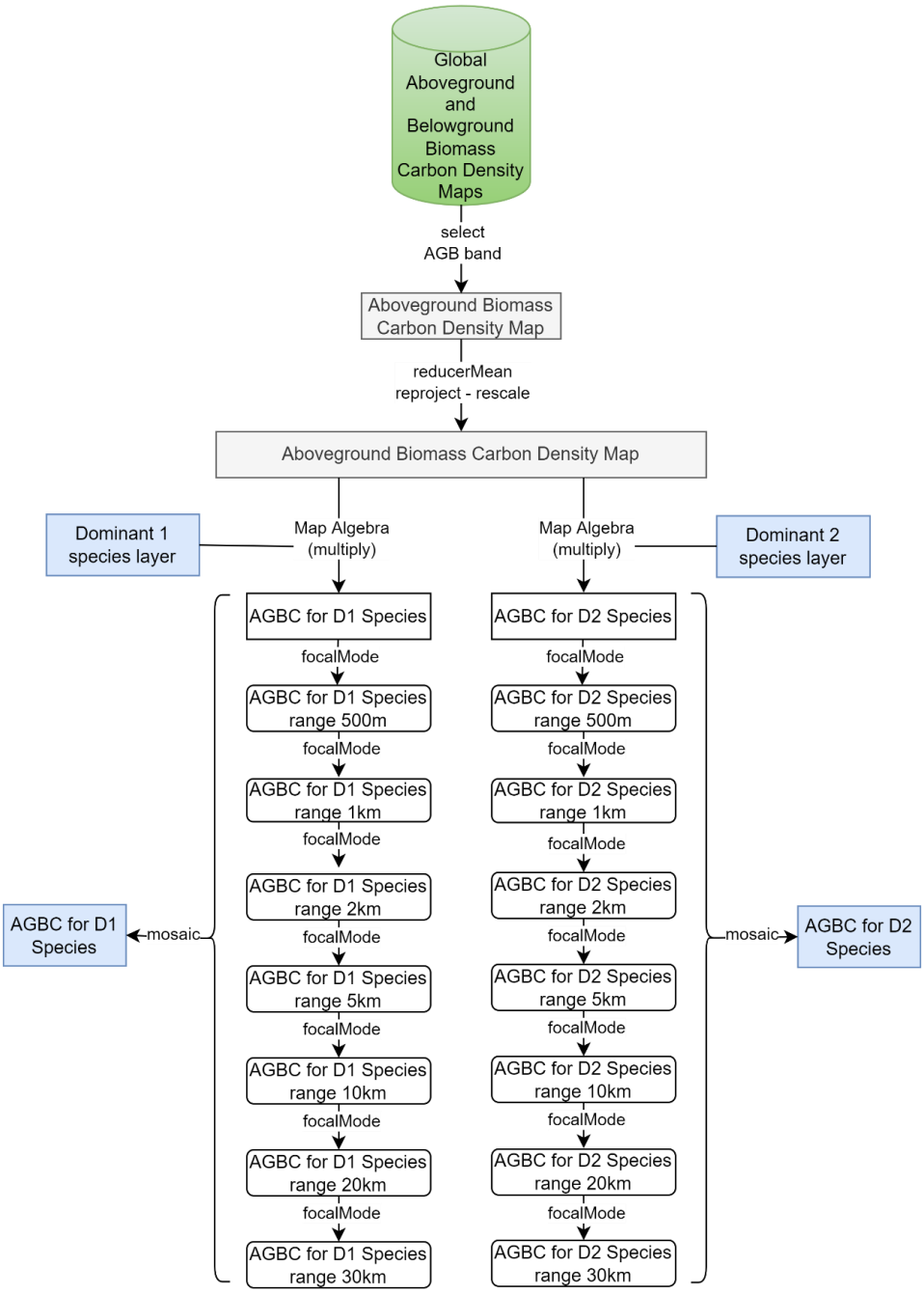


Figure 18: Dominant 1 Species layer workflow

Since Global Aboveground and Belowground Biomass Carbon Density Maps allows to distinguish between AGBC and BGBC, the AGB band was selected in order to calculate through map algebra the AGBC of Dominant 1 and 2 species. For this case, map algebra used by multiplying AGBC with Dominant 1 Species layer and with Dominant 2 Species layer separately and then `focal` mode was applied for ranges from 500m to 30km, same as previously. Finally, mosaic was used to produce AGBC for Dominant 1 Species and AGBC for Dominant 2 Species.



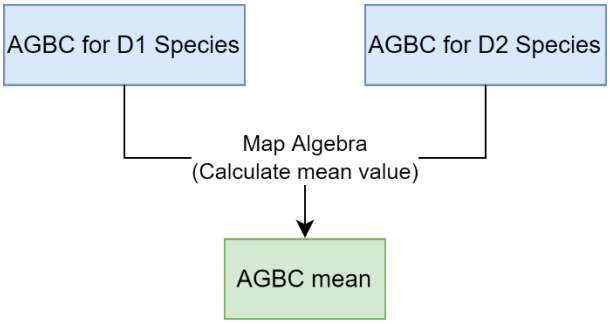


Figure 19: Aboveground Biomass Carbon workflow

4.5. Classification into CSC groups

The results from processing the workflows previously analyzed are multiple rasters. These rasters were combined with layers of previous tasks, Hard and Soft, composed to a multiband raster with 10m spatial resolution providing bands with information as shown on Table 10.

Table 10: Final ML raster

Final ML raster (multiband)	
Bands	**Values**
Hard	1: Marginal Land
Soft	0.41 - 10.3: low to high marginality
Dominant 1 species	1: Abies spp 2: Alnus spp 3: Betula spp 4: Carpinus spp 5: Castanea spp 6: Eucalyptus 7: Fagus spp 8: Fraxinus spp 9: Larix spp 10: Broadleaves 11: Conifers 12: Pines misc 13: Quercus misc 14: Picea spp 15: Pinus pinaster 16: Pinus sylvestris 17: Populus spp 18: Pseudotsuga menziessi 19: Quercus robur & Quercus petraea
Dominant 2 species	

Final ML raster (multiband)	
Bands	**Values**
	20: Robinia spp
Dominant 1 percent	0 - 100: Percentage share of the respective tree species from land area
Dominant 2 percent	0 - 50: Percentage share of the respective tree species from land area
AGBC	0 - 236: Mg C/ha
AGBC for D1 species	0 - 236: Mg C/ha
AGBC for D2 species	0 - 236: Mg C/ha
AGBC mean (AGBC for D1 species + AGBC for D2 species)/2	0 - 236: Mg C/ha

For the estimation of Carbon Sequestration Capacity, a formula used in which AGBC mean, productivity value and weight factor of 70% multiplied. From these three variables, AGBC mean is the mean value calculated from AGBC for D1 species - AGBC for D2 species and productivity value is calculated on task 2.3 for each ML. The weight factor of 70% is defined as the upper limit of a ML's CSC after being afforested, compared to neighbor forested areas, because MLs by default have productivity restrictions.

$$\boldsymbol{AGBC\ mean} \times \boldsymbol{Productivity\ value} \times \boldsymbol{Weight\ factor}\ \mathbf{70\%}$$

Equation 4. Carbon Sequestration Capacity (European level)

Because of the extent of the study area, covering the whole Europe, the available data regarding AGBC which refer to 2010 and the fact that CSC of MLs may change from year to year due to various reasons and factors, the formula's results do not represent the maximum C stock that can be obtained. The results of the formula serve classification purposes only. Through classification into CSC groups, we get a better understanding regarding the relative interconnections between groups and each one's potential trend.

In order to discover and present the frequency distribution of the formula's results, a histogram plotted. Classification into CSC groups was done by manually defining

classes ranges, in such a way so each class to cover approximately the same area across Europe, with the exception of higher and lower sequestration groups, Group A and Group E respectively. Group A represents higher sequestration MLs, covering 5% of Europe's total MLs and on the other side Group E represents lower sequestration MLs covering 31% of Europe's MLs.

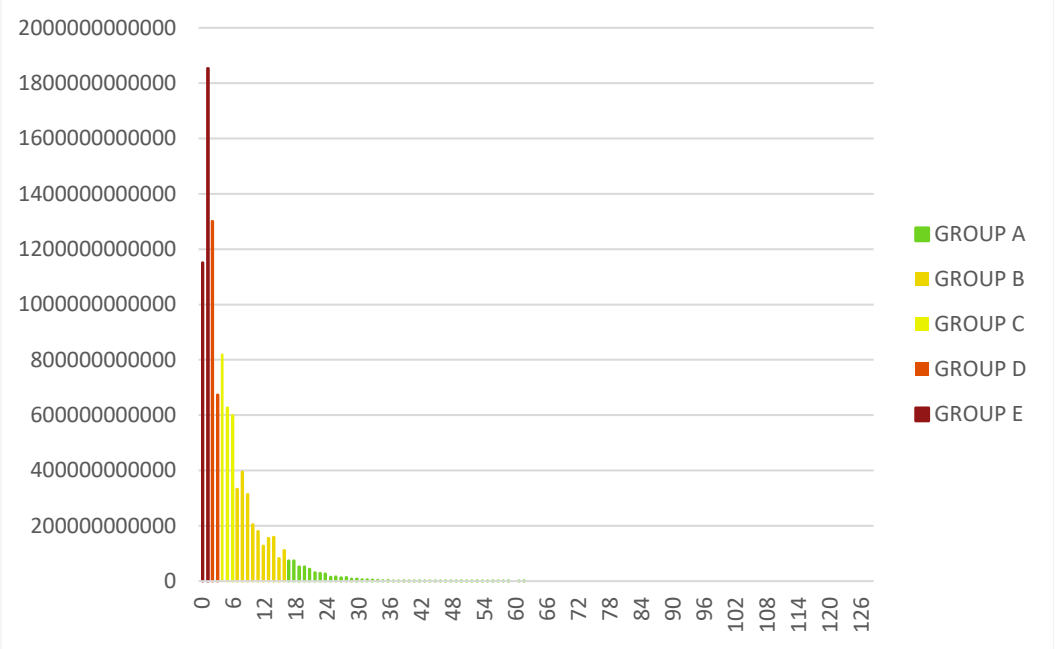


Figure 20: Histogram of CSC values for European level

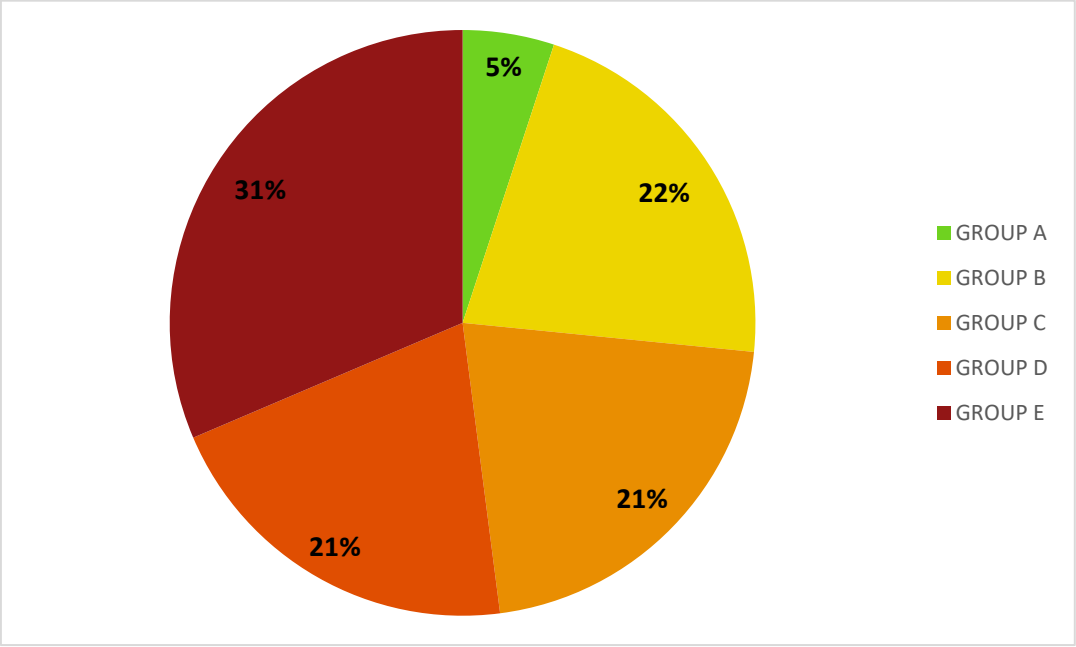


Figure 21: Groups' distribution in percentages for European cover

4.6. Google Earth Engine Tool

From the findings of this chapter, a tool was also developed to be accommodated to the ***MAIL*** geoportal, into the Decision Support System under the name “Potential Suitable Species”.

The purpose is to provide a general overview regarding Carbon Sequestration Capacity Groups (CSC Groups) and to suggest Potential Suitable Species for afforestation. The CSC groups are calculated based on the methodology applied for the whole Europe and Potential Suitable Species on presence frequency in the neighbor forested areas, ranked according to dominance.

The analysis occurs at a user’s defined level (student, stakeholder, etc.) by drawing or inserting a specified Area Of Interest (AOI; **.geojson*). The AOI information is displayed on three relative pie charts, one for CSC Groups and another two for species, dominant 1 and dominant 2. In each relative pie chart, the results illustrate the participation percentages in the AOI.

There is no specified limit in the AOI extend. However, the tool designed for parcel scale analysis. Therefore, it is suggested not to exceed 100,000 ha, as the accuracy is inversely proportional to the AOI.

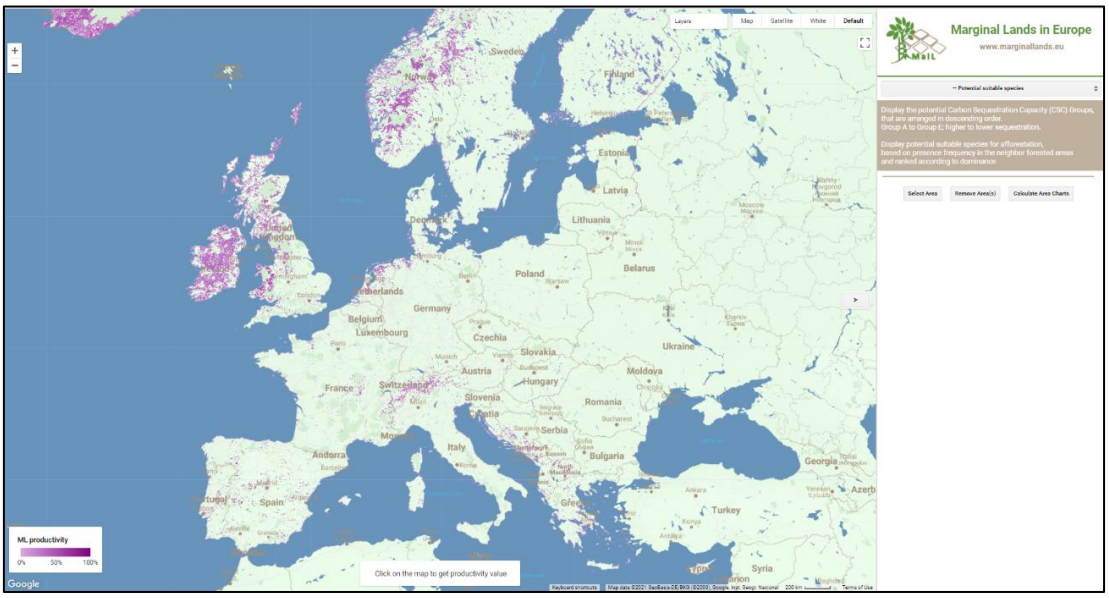


Figure 22: Potential Suitable Species tool menu

The tool scope is a broader approach in European level. Thus, by no means can substitute an in-situ analysis, that takes into account more aspects such as micro-climate, ecological zone, elevation, soil attributes, plant indicators, etc.

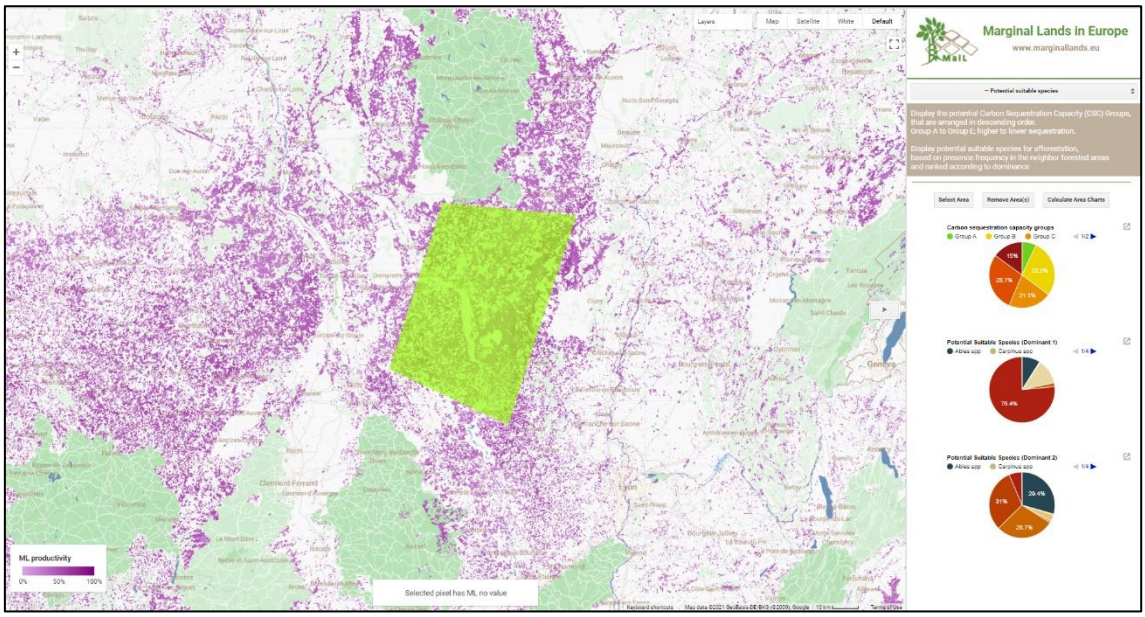


Figure 23: Potential Suitable Species tool results

4.7. Discussion & Conclusions

In this chapter a methodology for classifying MLs into CSC groups was developed. It would not be possible to apply the same methodology applied at pilot site level. Due to the extent of the study area and the time frame of the task, field measurements or remote sensing wouldn't be feasible.

The classification performed by developing a new methodology and using freely available data. Estimating potential suitable species for afforestation and their AGBC were crucial in order to estimate MLs' CSC. Detailed workflows were created that describe every step of the preprocessing of the datasets that used for the MLs classification into CSC groups as well as for estimating potential suitable species for afforestation.

Following the collection and combination of all the appropriate datasets required for the methodology, a formula developed in which AGBC mean, productivity value and a weight factor of 70% are multiplied to estimate CSC ($\boldsymbol{AGBC\ mean} \times \boldsymbol{Productivity\ value} \times \boldsymbol{Weight\ factor}\ \mathbf{70\%}$

Equation 4). MLs were classified in five groups by manually defining classes' ranges, Group A to E, from higher to lower sequestration groups. MLs classified into CSC groups with 2010 as reference year and each group does not represent the maximum C stock that can be obtained, but classification purpose is to give a better understanding regarding relative interconnections between groups. MLs' CSC and the groups may change from year to year due to various reasons and factors.

The GEE platform used for the implementation of the MLs classification into CSC groups and also for the development of a tool to be embedded in ***MAIL*** geoportal, in the Decision Support System. The tool provides the stakeholders with the option to display CSC groups and potential suitable species for afforestation for the areas of their interest.

REFERENCES

[1] Brus, D. J., Hengeveld, G. M., Walvoort, D. J., Goedhart, P. W., Heidema, A. H., Nabuurs, G. J., & Gunia, K. (2011). *Statistical mapping of tree species over Europe. Special Issue European Journal of Forest Research.*

[2] Cairns, M., Brown, S., Helmer, E., & Baumgardner, G. (1997). Root biomass allocation in the world's upland forests. *Oelogia*, 111, 1–11.

[3] EEA. (2017). *European environmental agency. climate change, impacts and vulnerability in europe 2016, an indicator-based report.* European Environment Agency.

[4] FAO. (2012). Global ecological zones for FAO forest reporting: 2010 update. *Food and Agriculture Organization*.

[5] Forkuor, G., Zoungrana, B., Dimobe, K., Ouattara, B., Vadrevu, K., & Tondoh, E. (2020). Above-ground biomass mapping in West African dryland forest using Sentinel-1 and 2 datasets - A case study. *Remote Sensing of Environment*, 236.

[6] Galidaki, G., Zianis, D., Gitas, I., Radoglou, K., Karathanassi, V., Tsakiri–Strati, M., . . . Mallinis, G. (2017). Vegetation biomass estimation with remote sensing: focus on forest and other wooded land over the Mediterranean ecosystem. *International Journal of Remote Sensing*, 38(7), 1940–1966.

[7] Gao, Y., Lu, D., Li, G., Wang, G., Chen, Q., Liu, L., & Li, D. (2018). Comparative analysis of modeling algorithms for forest aboveground biomass estimation in a subtropical region. *Remote Sensing*, 10, 627.

[8] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). *Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment.*

[9] Gregorio, M., Ricardo, R., & Marta, M. (2005). *Produccion de biomasa y fijacion de CO2 por los bosques espanoles.*

[10] Hall-Beyer, M. (2017a). Glcm texture: A tutorial .

[11] Haralick, R., Shanmugam, K., & Dinstein. (1973). Textural features for image classification. *IEEE Trans Syst Man Cybern*, 610–621.

[12] Hong, H., Xiaoling, G., & Hua, Y. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. *IEEE International Conference on Software Engineering and Service Science*, 219–224.

[13] IPCC. (2006). *guidelines for national greenhouse gas inventories volume 4: agriculture, forestry and other land use.* Intergovernmental Panel on Climate Change.

[14] Kankare, V., Vastaranta, M., Holopainen, M., Raty, M., Yu, X., Hyyppa, J., . . . Viitala, R. (2013). Retrieval of forest aboveground biomass and stem volume with airborne scanning lidar. *Remote Sensing*, 5, 2257–2274.

[15] Keith, H. M. (2009). Re-evaluation of forest biomass carbon stocks and lessons from the world's most carbon-dense forests. *National Academy of Sciences*, 106: 11635-11640.

[16] Kelsey, K., & Neff, J. (2014). Estimates of aboveground biomass from texture analysis of Landsat imagery. *Remote Sensing*, 6, 6407–6422.

[17] Khan, K., Iqbal, J., Ali, A., & Khan, S. N. (2020). Assessment of sentinel-2-derived vegetation indices for the estimation of above-ground biomass/carbon stock, temporal deforestation and carbon emissions estimation in the moist temperate forests of Pakistan. *Applied Ecology and Environmental Research*, 18, 783–815.

[18] Lackner, K. S. (2003). A guide to CO2 sequestration. *Science*, 300 (5626) 1677–1678.

[19] Lewinson, E. (2019). *Explaining feature importance by example of a Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

[20] Liu, Y. G.-F. (2012). Huge carbon sequestration potential in global forests. *Journal of Resources and Ecology* , 3, 193–201.

[21] Lu, D. (2006). The potential and challenge of remote sensing-based biomass estimation. *International Journal of Remote Sensing*, 27(7), 1297–1328.

[22] Lundberg, S. E.-I. (2020). 'From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*.

[23] Malinowski, R., Lewiński, S., Rybicki, M., Gromny, E., Jenerowicz, M., Krupiński, M., . . . Schauer, P. (2020). *Automated Production of a Land Cover/Use Map of Europe Based on Sentinel-2 Imagery.* Remote Sensing.

[24] Pandit, S., Tsuyuki, S., & Dube, T. (2018). Estimating above-ground biomass in subtropical buffer zone community forests, Nepal, using sentinel-2 data. *Remote Sensing*, 10, 601.

[25] Pandit, S., Tsuyuki, S., & Dube, T. (2019). Exploring the inclusion of Sentinel-2 MSI texture metrics in above-ground biomass estimation in the community forest of Nepal. *Geocarto International*.

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[27] Picard, N., Saint-Andre, L., & Henry, M. (2012). *Manual for building tree volume and biomass allometric equations: from field measurement to prediction.*

[28] Puliti, S., Hauglin, M., Breidenbach, J., Montesano, P., Neigh, C., Rahlf, J., . . . Astrup, R. (2020). Modelling above-ground biomass stock over norway using national forest inventory data with arcticdem and Sentinel-2 data. *Remote Sensing of Environment*, 236.

[29] Salem, I., Basam, D., Taoufik, K., & Nazmi, S. (2020). A review of terrestrial carbon assessment methods using geo-spatial technologies with emphasis on arid lands. *Remote Sensing*.

[30] Spawn, S. A., Sullivan, C. C., Lark, T. J., & Gibbs, H. K. (2020). *Harmonized global maps of above and belowground biomass carbon density in the year 2010.* Scientific Data.

[31] Stinson, G., Kurz, W. A., Smyth, C. E., Neilson, E. T., Dymond, C. C., & Metsaranta, J. (2012). An inventory-based analysis of canada's managed forest carbon dynamics, 1990 to 2008. *Global Change Biology*, 17(6), 2227–2244.

[32] Thomas, S., & Martin, A. R. (2012). Carbon content of tree tissues: A synthesis. *Forests*, 3, 332–352.

[33] Torralba, J., Crespo-Peremarch, P., & Ruiz, L. A. (2018). Assessing the use of discrete , full-waveform LiDAR and TLS to classify Mediterranean forest species composition. *Revista de Teledetección*, 52, 27-40.

[34] Vashum, K. (2012). Methods to estimate above-ground biomass and carbon stock in natural forests - a review. *Journal of Ecosystem Ecography*.

[35] Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M. K., Moine, J. L., & Ryu, S. R. (2004). Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. *Remote Sensing of Environment*, 93, 402–411.

[36] Zhou, G., Wang, Y., Jiang, Y., & Yang, Z. (2002). Estimating biomass and net primary production from forest inventory data: A case study of China's Larix forests. *Forest Ecology and Management*, 169, 149-157.

ANNEX I: TABLE OF FIGURES

ANNEX II: LIST OF TABLES