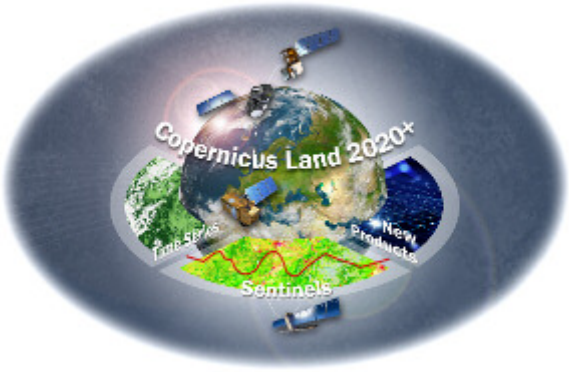
Horizon 2020

Space Call - Earth Observation: EO-3-2016: Evolution of Copernicus services

Grant Agreement No. 730008

ECoLaSS

Evolution of Copernicus Land Services based on Sentinel data



D15.1

"D45.1a – Prototype Report: New LC/LU Products"

Issue/Rev.: 1.0

Date Issued: 03.08.2018

submitted by:

SIRS

in collaboration with the consortium partners:

GAFAG JOANNEUM RESEARCH DIGITAL UCL Université catholique de Louvain DLR

submitted to:

European Commission – Research Executive Agency

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 730008.

CONSORTIUM PARTNERS

NO.	PARTICIPANT ORGANISATION NAME	SHORT NAME	CITY, COUNTRY
1	GAF AG	GAF	Munich, Germany
2	Systèmes d'Information à Référence Spatiale SAS	SIRS	Villeneuve d'Ascq, France
3	JOANNEUM RESEARCH Forschungsgesellschaft mbH	JR	Graz, Austria
4	Université catholique de Louvain, Earth and Life Institute (ELI)	UCL	Louvain-la-Neuve, Belgium
5	German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Wessling	DLR	Wessling, Germany

CONTACT:

GAF AG
Arnulfstr. 199 – D-80634 München – Germany
Phone: ++49 (0)89 121528 0 – FAX: ++49 (0)89 121528 79
E-mail: copernicus@gaf.de – Internet: www.gaf.de

DISCLAIMER:

Document Release Sheet

	Name, Function	Date	Signature
Author(s):	Alice Lhernould (SIRS) Clémence Kenner (SIRS)	18.07.2018	
Review:	Sophie Villerot (SIRS) Katharina Schwab (GAF)	01.08.2018	
Approval:	Linda Moser (GAF)	03.08.2018	
Acceptance:	Massimo Ciscato (REA)		
Distribution:	Public		

Dissemination Level

Dissemination Level		
PU	Public	X
CO	Confidential: only for members of the consortium (including the Commission Services)	

Document Status Sheet

Issue/Rev	Date	Page(s)	Description / Changes
1.0	03.08.2018	25	First version of WP45 Deliverable on the Prototype of a New Land Cover Product

Applicable Documents

ID	Document Name / Issue Date
AD01	Horizon 2020 Work Programme 2016 – 2017, 5 iii. Leadership in Enabling and Industrial Technologies – Space. Call: EO-3-2016: Evolution of Copernicus services. Issued: 13.10.2015
AD02	Guidance Document: Research Needs Of Copernicus Operational Services. Final Version issued: 30.10.2015
AD03	Proposal: Evolution of Copernicus Land Services based on Sentinel data. Proposal acronym: ECoLaSS, Proposal number: 730008. Submitted: 03.03.2016
AD04	Grant Agreement – ECoLaSS. Grant Agreement number: 730008 – ECoLaSS – H2020-EO-2016/H2020-EO-2016, Issued: 18.10.2016
AD05	D6.1: D31.1a - Methods Compendium: Sentinel-1/2/3 Integration Strategies, (Issue 1), Issued: March 2018

ID	DOCUMENT NAME / ISSUE DATE
AD06	D7.1: D32.1a- Methods Compendium: Time Series Preparation, (Issue 1), Issued: February 2018
AD07	D8.1: D33.1a - Methods Compendium: Time Series Analysis for Thematic Classification (Issue 1), Issued: 29.03.2018
AD08	D9.1: D34.1a - Methods Compendium: Time Series Analysis for Change Detection (Issue 1), Issued: 29.03.2018
AD09	D10.1: D35.1a - Methods Compendium: HRL Time Series Consistency for HRL Product (incremental) Updates (Issue 1), Issued: 14.05.2018

EXECUTIVE SUMMARY

The Horizon 2020 (H2020) project, "Evolution of Copernicus Land Services based on Sentinel data" (ECoLaSS) addresses the H2020 Work Programme 5 iii. Leadership in Enabling and Industrial technologies - Space, specifically the Topic EO-3-2016: Evolution of Copernicus services. ECoLaSS is being conducted from 2017–2019 and aims at developing and prototypically demonstrating selected innovative products and methods for future next-generation operational Copernicus Land Monitoring Service (CLMS) products of the pan-European and Global Components. This will contribute to demonstrating operational readiness of the finally selected products, and shall allow the key CLMS stakeholders (i.e. mainly the Entrusted European Entities (EEE) EEA and JRC) to take informed decisions on potential procurement of the next generation of Copernicus Land services from 2020 onwards.

To achieve this goal, ECoLaSS makes full use of dense time series of Sentinel-2 and Sentinel-3 optical data as well as Sentinel-1 Synthetic Aperture Radar (SAR) data. Rapidly evolving scientific as well as user requirements are being analysed in support of a future pan-European roll-out of new/improved CLMS products, and the transfer to global applications.

This report is based on the WP21 (D21.1 Service Evolution Requirements Report) and the WP22 (D22.1 Earth Observations (EO) and other Data Requirements Report) – which both clarify requirements, in terms of service and EO availability – as well on the outputs of all reports related to the Task 3 – the WP31 (D31.1 Methods Compendium: Sentinel-1/2/3 Integration Strategies), the WP32 (D32.1 Methods Compendium: Time Series Preparation), the WP34 (D34.1 Methods Compendium: Time Series Analysis for Change Detection), the WP35 (D35.1 Methods Compendium: Time Series Consistency for High Resolution Layers (HRL) Product (incremental) Updates) and more importantly the WP33 (D33.1 Methods Compendium: Time Series Analysis for Thematic Classification) mainly – whose aim are to test methodologies and select the best ones for the demonstration phase. All four other report contributions for the Task 4 are used to redact this report – namely, the WP41 (D41.1 Prototype Report: Time Series-derived Indicators and Variables), the WP42 (D42.1 Prototype Report: Consistent HR Layer Time Series/Incremental Updates), the WP43 (D43.1 Prototype Report: Improved Permanent Grassland) and for most part, the WP44 (D44.1 Prototype Report: Crop Area and Crop Status/Parameters).

Those inputs represent the foundations to design new Land Cover (LC)/Land Use (LU) products, not only built on existing datasets, but also designed to meet the future CLMS user needs. Those new LC/LU products will therefore be based on Sentinel data, optical and SAR, and will present an increased spatial resolution to fit all actual and future CLMS production requirements.

To do so, state of play for the availability of the required data will be sketched, depending on the production need for new LC/LU. Combined with the refinement on thematic classes definition, possibly following the EAGLE (EIONET Action Group on Land monitoring in Europe) recommendations, those two axes should enable the implementation of a new HR LC layer, that will then be validated at pan-European level.

After the production of a cropland type HRL in the first phase, several developments will be conducted: the setting of aggregation rules to combine individual HRL into a spatially and temporally coherent product, the creation of agri-environmental intensification indicator, based on the outputs of WP41, and finally, the application of all those research to site located outside Europe.

Section 1 of the document details the objectives of this WP45. Section 2 presents the backgrounds of previous LC/LU products and the summary of their limitations. Section 3 is a short presentation of the demonstration sites and section 4 describes all the applied time series analysis methods, based on the testing from Task 3. Section 5 lays out the prototype implementation approaches, from the integration of auxiliary data in the Sentinel workflow, the experimental set-up to the final validation of the prototype.

Table of Contents

List of Figures and Tables

Abbreviations

AOI	Area Of Interest
BRI	Brightness Index
CLMS	Copernicus Land Monitoring Service
CLC	CORINE Land Cover
CORINE	Coordination of Information on the Environment
CT	Classification Tree
ECoLaSS	Evolution of Copernicus Land Services based on Sentinel data
EAGLE	EIONET Action Group for Land monitoring in Europe
EEA	European Environment Agency
EEA-39	39 European Countries
EEE	Entrusted European Entities
EIONET	European Environment Information and Observation Network
EO	Earth Observations
ETM+	Enhanced Thematic Mapper Plus
EU	European Union
HR	High Resolution
HRL	High Resolution Layers
IMD	Imperviousness Degree
INSPIRE	Infrastructure for Spatial Information in the European Community
JRC	Joint Research Centre
LAEA	Lambert Azimuthal Equal-Area (projection)
LC	Land Cover
LPIS	Land Parcel Identification System
LSMSS	Large-Scale Mean Shift Segmentation
LU	Land Use
LZW	Lempel-Ziv-Welch
MMU	Minimal Mapping Unit
MSS	Minimum Segment Size
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near-InfraRed
NLC	New Land Cover
OA	Overall Accuracy
OSM	Open Street Map
PA	Producer Accuracy
RF	Random Forest
S-1	Sentinel-1
S-2	Sentinel-2
S-3	Sentinel-3
SAR	Synthetic Aperture Radar
SWIR	Short Wave InfraRed
TCD	Tree Cover Density
TIFF	Tag Image File Format
TM	Thematic Mapper
UA	User Accuracy
XML	Extensible Markup Language
WAW	Water And Wetness
WP	Work Package

1 Introduction

The Horizon 2020 (H2020) project, "Evolution of Copernicus Land Services based on Sentinel data" (ECoLaSS) addresses the H2020 Work Programme 5 iii. Leadership in Enabling and Industrial technologies - Space, specifically the Topic EO-3-2016: Evolution of Copernicus services. ECoLaSS will be conducted from 2017–2019 and aims at developing and prototypically demonstrating selected innovative products and methods for future next-generation operational Copernicus Land Monitoring Service (CLMS) products of the pan-European and Global Components. This will contribute to demonstrating operational readiness of the finally selected products, and shall allow the key CLMS stakeholders (i.e. mainly the Entrusted European Entities (EEE) EEA and JRC) to take informed decisions on potential procurement of the next generation of Copernicus Land services from 2020 onwards.

To achieve this goal, ECoLaSS will make full use of dense time series of Sentinel-2 and Sentinel-3 optical data as well as Sentinel-1 Synthetic Aperture Radar (SAR) data. Rapidly evolving scientific as well as user requirements will be analysed in support of a future pan-European roll-out of new/improved CLMS products, and the transfer to global applications.

This Deliverable **D15.1: "D45.1a – Prototype Report: New LC/LU Products (Issue 1)"** contains the description of produced prototype datasets of new LC/LU products (linked to Deliverable P45.2). It provides the detailed characterization of objectives with added methodology explanation, results and conclusions that have been reached in WP45. It addresses the prototype methodologies for preparation of reference data, validation datasets and Sentinel images in the demonstration site SOUTH-WEST, covering France and a smaller part of Spain, as well as accuracy assessment for those results. As such it is part of **WP 45** of Task 4: "Thematic Proof-of-Concept/Prototype on Continental/Global Scale", which aims at implementing new LC/LU products based on S-2 and S-1 time series, at a global scale, while fitting as far as possible with the current CLMS products. This report will be accompanied by the Deliverable **D15.3: "P45.2a – Data Sets of New LC/LU Products"** constituting the prototype dataset. This report serves as documentation for the prototype dataset.

In the ECoLaSS project a prototype is defined as a prototypic/thematic proof-of-concept implementation of an improved or newly defined potential future Copernicus Land layer, building on the methods and processing lines developed in Task 3. The consortium has selected representative demonstration sites both in Europe and Africa, covering various bio-geographic regions and biomes. All prototype products and services are being prototypically implemented in a selection of these sites in the frame of the Task 4 WPs In ECoLaSS, proofs-of-concept/prototype demonstrations are carried out with respect to five topics of relevance: (i) Time series derived indicators and variables, (ii) Incremental Updates of HR Layers, (iii) Improved permanent grassland identification, (iv) Crop area and crop status / parameters monitoring, and (v) New LC/LU products. This deliverable focuses on the prototype **New Land Cover Product** as part of WP45.

This report is a first issue, covering preliminary results presented at month 18. The final results, obtained at the end of the second iteration, in particular regarding the intensification indicator for agri-environmental purposes that will be closely linked to the results of WP44, will be delivered at the end of phase 2.

This report comprises a chapter on the background to the New Land Cover Product Prototype and associated requirements (Chapter 2); a description of the Demonstration Sites where the prototype is implemented (Chapter 3); an overview of the methodologies carried over from the testing and benchmarking in Task 3 (Chapter 4); followed by a Chapter on the prototype implementation itself including a description of the dataset (Chapter 5); and a summary and outlook (Chapter 6).

2 Background and Summary of Requirements

After first methods have been tested by the Task 3 WPs (AD05, AD06, AD07, AD08, AD09) in various test sites and algorithms have been described, the demonstration activities of Task 4 have commenced to set up the developed processing lines in demonstration sites and derive first prototype versions. This will comprise establishing prototypes for: (i) deriving indicators and variables both for Continental and Global Component products and services from high-volume time series data with high spatial resolution and temporal repeat frequency; (ii) improving one of the main pan-European Copernicus Land products, i.e. the current (2012) and future (2015, 2018) HRLs on Forest and Imperviousness by developing incremental update strategies and ensuring time series consistency; (iii) improved permanent grassland identification targeting the HRL Grassland 2015 improvement; (iv) crop area and crop status/parameters monitoring targeting a potential future Agricultural service; as well as (v) further novel LC/LU products, e.g. as tested in Task 3.

The project is basing all its developments on regularly updated high-priority user requirements, and assess/benchmark all operational product candidates in view of their innovation potential and technical excellence, automation level, potential for roll-out to pan-European level and/or global scale, timeliness for operational implementation, costs versus benefits, etc. (further elaborations will be performed in Task 5).

Various products have been created in attempt to map continent(s) at high or medium spatial resolution. The following examples can be mentioned:

- CORINE Land Cover (CLC): designed to map the pan-European territory at 100m resolution, every ten then six years, it suffers from a known lack of methodological consistency, since the production is done at national scale, hampering the accuracy. The thematic classes are a mixed between LC and LU, disregarding EAGLE recommendations, and sometimes unevenly representing the landscape characteristics depending on the bio-geographical situation.
- Global Land Cover (GLC): the project has produced 3 iterations, FROM-GLC (2013), FROM-GLC-seg, FROM-GLC-agg. The first two used a 30m spatial resolution, while the last was an aggregation of the previous two at a coarser resolution of 1km. Those global covers have been made using Landsat (TM and ETM+), complemented by MODIS EVI time series and other auxiliary datasets, such as Bioclimatic variables, global Dem, and soil-water variables. The lack of temporal features, combined with the use of old input and validation data, sometimes dating back to 20 years, in the classification process drove the accuracy down to 65% at best due to multiple confusion among 9 land-cover types used.
- At national scale, the CESBIO created maps of French territory in 2016 based on Landsat 8 then in 2017 based on Sentinel-2 with 17 land cover classes, at a 30m resolution, then at 10m.

The current known limitations among the available land cover products can be summarized as too low spatial and temporal resolutions, as well as some inconsistencies between the different datasets.

All those summarized issues call for the emergence of new land cover (LC) products, which should exhibit new properties to increase their spatial and temporal consistency. The creation of a pan-European HR LC layer could be obtained by merging together all the currently available layers, in addition to a new agricultural layer. This merge constitutes an opportunity to enforce a logical consistency between the current and upcoming thematic products, which are being produced independently, without requiring post-processing to ensure the spatial and temporal coherence.

3 Demonstration Sites

All prototypes are implemented in selected representative demonstration sites, which cover various bio-geographic regions and biomes. The New Land Cover Product is demonstrated in the the South-West demonstration site.

3.1 ECoLaSS Demonstration Sites

The selected larger prototype sites (60,000/90,000 km² per prototype site) contain the 5 test sites from Task 3. These prototype sites are relevant in Task 4 for demonstrating the proposed candidates for a Copernicus Land Service Evolution roll-out on a larger scale. As shown in Figure 3-1, the pre-selected prototype sites cover the Atlantic and Continental zones (Source EEA: https://www.eea.europa.eu/data-and-maps/data/biogeographical-regions-europe-3#tab-gis-data) of the member and associated states of EEA-39. The selected prototype sites are located in the North of Europe, in the Alpine/Central region, in the West, in the South-West and in the South-East of Europe. All prototype products and services will be prototypically implemented in one or more prototype sites in project phase 1, and in three prototype sites in phase 2.

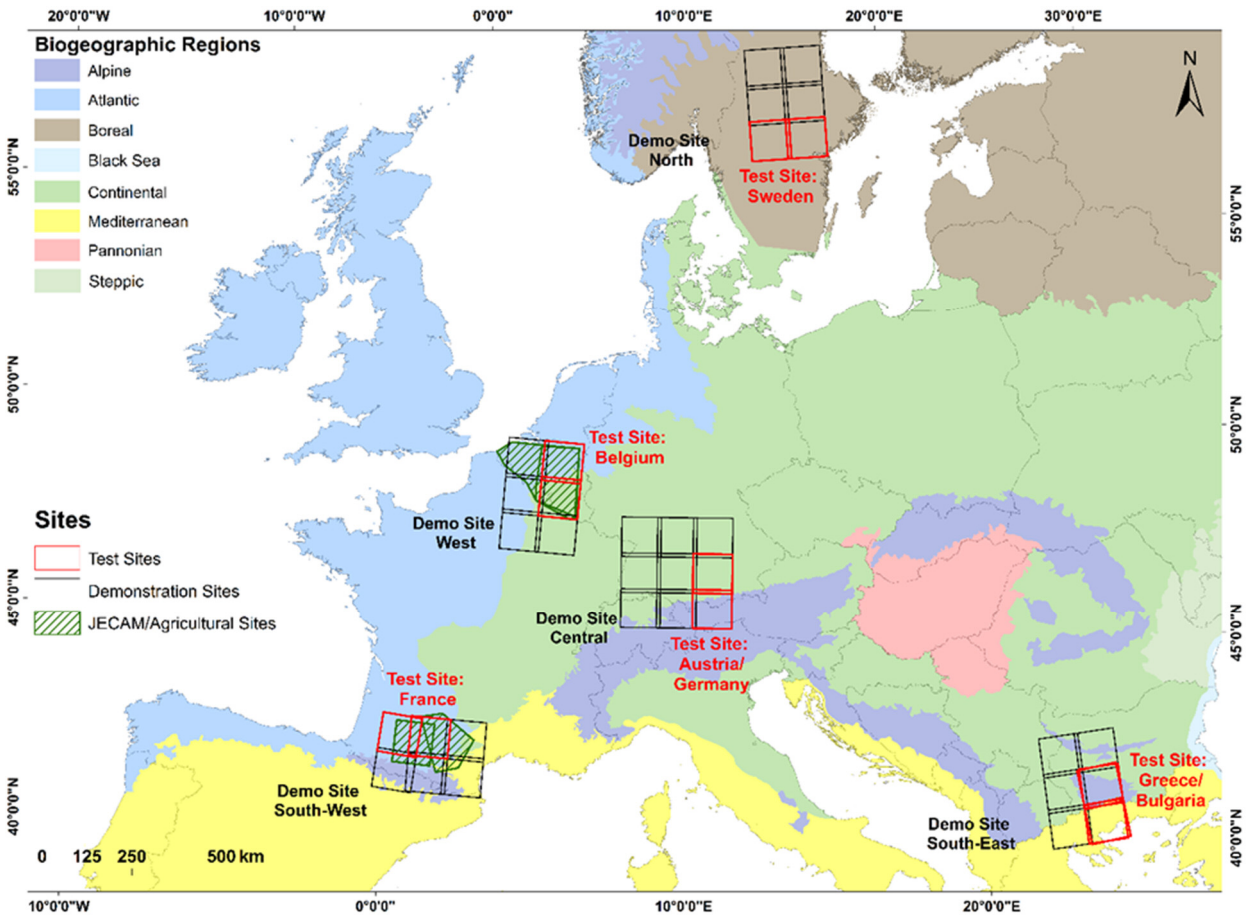


Figure 3-1 - Biogeographic Regions of Europe (2015) and European ECoLaSS Demonstration Sites
(Map: © European Environment Agency; administrative boundaries: ©EuroGeographics)

A short description of the different prototype sites is given in the following Table 3-1 below:

Table 3-1 - Description of the selected Prototype Sites

Location	Biogeographical region(s)	Countries	Distribution of CORINE land cover classes 2012 (Level 1) per prototype site *
Northern Europe	Boreal	Sweden	Artificial areas: 1.38 %, Agricultural areas: 10.54 %, Forest and semi-natural areas: 70.61 %, Wetlands: 4.25 %, Water bodies: 13.22 %
Alpine / Central Europe	Continental, Alpine	Germany, Austria, Switzerland, Italy and France	Artificial areas: 5.86 %, Agricultural areas: 42.96 %, Forest and semi-natural areas: 49.95 %, Wetlands: 0.24 %, Water bodies: 0.98 %
West Europe	Atlantic, Continental	Belgium, France, Luxembourg, Netherlands	Artificial areas: 7.81 %, Agricultural areas: 53.75 %, Forest and semi-natural areas: 13.15 %, Wetlands: 0.25 %, Water bodies: 25.04 %
South-East Europe	Mediterranean, Continental, Alpine	Serbia, Macedonia, Greece, Bulgaria	Artificial areas: 3.03%, Agricultural areas: 37.00 %, Forest and semi-natural areas: 53.95 %, Wetlands: 0.17 %, Water bodies: 5.71 %
South-West Europe	Atlantic, Mediterranean, Alpine	France, Spain	Artificial areas: 3.26 %, Agricultural areas: 46.73 %, Forest and semi-natural areas: 49.20 %, Wetlands: 0.01 %, Water bodies: 0.40 %

3.2 Demo-Site South-West for new LC/LU Product

In the frame of the Task 4 where all prototype products and services are implemented, the South-West demonstration site has been chosen as the primary demonstration site for the improved HRL Imperviousness (IMP) as well as a potential future Copernicus New Land Cover Layer Product at high resolution. The demonstration site contains the test site “France” that has been studied in Task 3.

The South-West demonstration site is used as a prototypic / thematic proof-of-concept implementation of the outcome of the methods and processing lines developed in Task 3. The demonstration site covers the south of France, small parts of Spain and the Principality of Andorra, in the shape of 6 Sentinel-2 tiles. Landscapes are divided between different biogeographic regions such as Mediterranean, Alpine and Atlantic. Three Sentinel-2 tiles are dominated by mountain landscapes, a mix of bare soils and natural grasslands, due to the presence of the Pyrénées. The 31TCJ tile is dominated by a strong proportion of impervious soil, because of Toulouse, a major French city. The plains surrounding the city are mainly composed of croplands mixed with grassland and an increasing amount of forest with the proximity of the coastal region. In general, the Mediterranean area in the East is a patchwork of cropland, dry grassland and vineyard. We also can notice a small portion of the Landes forest in the Nord-West of the demonstration site. Major landscapes can be visualized on this **Fehler! Verweisquelle konnte nicht gefunden werden.** of the CLC 2012 product over the area where Andorra is wrongly coded as waterbody, since it does not belong to the EU.

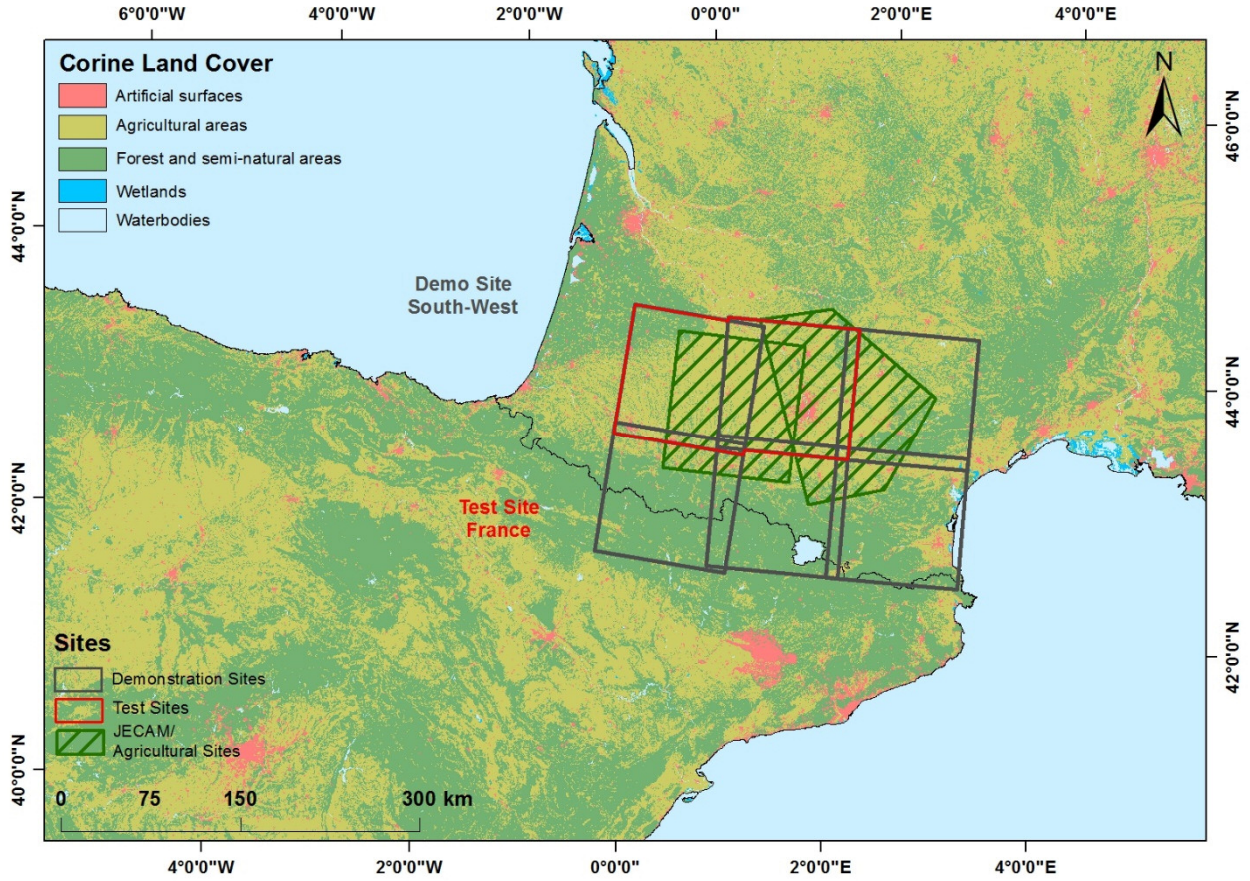


Figure 3-2 - South-West Prototype Site, draped over CLC dataset (2012)

4 Overview of applied methods

Multiple classes, exhibiting strongly different variability characteristics, are studied here. Time series indicators are essential, especially to outline various phenological cycles, but also to find a balance between datasets of dense time series, rich in spectral and textural information, and the time consumed by the whole processing chain.

4.1 Method for new LC/LU Product – South West Demonstration site

This section shows an overview of the methods implemented for the New Land Cover Product prototype following the outcomes of the Task 3, from the preparation of the data to the classification algorithm.

4.1.1 Spectral indices and temporal metrics

Many spectral indices have already been defined in the previous work packages. Only those used for the New Land Cover Product prototype generation will be specified in this paragraph.

Three indices have been calculated, for their abilities to discriminate between several classes at the same time.

NDVI – NORMALIZED DIFFERENCE VEGETATION INDEX:

As already described in the WP31 (AD05), the "Normalized Difference Vegetation Index" (Rouse Jr. et al., 1974; Tucker, 1979) is used as an indicator to monitor vegetation health, and can be used as a proxy for photosynthetic activity and primary production from vegetation biomass. It is calculated as the difference

in the reflectance between those two spectral regions, normalized by the sum of the reflectance measurements:

$$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}$$

NDWI – NORMALIZED DIFFERENCE WATER INDEX:

The "Normalized Difference Water Index" (Gao, 1996) is defined as the ratio

$$NDWI = \frac{\rho_{NIR} - \rho_{SWIR}}{\rho_{NIR} + \rho_{SWIR}}$$

where ρ is the radiance in reflectance unit. Both wavelengths are localized in the part of the spectrum reflected by vegetation canopies. The NIR channel is linked to a negligible absorption of light by the water content present in the vegetal, while the SWIR channel present a weak liquid absorption.
The NDWI is therefore sensitive to slight changes in the liquid water absorbed by vegetation canopies, giving an indication on the vegetation water stress.

BRI – BRIGHTNESS INDEX:

The Brightness Index (BI) can expressed as (Mathieu et al., 1998):

$$BI = \sqrt{\frac{\rho_{Red} + \rho_{Green} + \rho_{Blue}}{3}}$$

where ρ is a measure of the average reflectance magnitude in the visible bands, used to quantify the soil color effect.

TEMPORAL FEATURES:

Furthermore, temporal statistics on those derived indices have been applied, based on the seasonal time intervals. They are especially useful for determining vegetated classes (Esch et al., 2018). On the prototype generation only nine of them have been generated, for each index, on each considered period, at pixel-level:

- maximum value,
- mean value,
- minimal value,
- standard deviation,
  10^{th}, 25^{th}, 50^{th} (i.e. the median), 75^{th} and 90^{th} percentiles.

4.1.2 Pre-classification steps

Before launching the classification algorithm details, it should be noticed that further manipulation is required on the datasets: a temporal split to enhance the seasonal variations of the phenology and the preparation of the sample data, used for the training of the classifier as well as its validation.

4.1.2.1 Image generation (composite of 81 bands)

The New Land Cover Product prototype generation was established with three different time-windows on Sentinel-2 data dated 2017/01/03 to 2017/11/14. The approach was to make the land cover identification possible thanks to spectral signatures. The optimized time windows on the demonstration site South-West were then: January to March, April to June and July to November.
To discern the phenological dynamics of the land cover present on the study area, three indices have been generated: NDVI, NDWI and Brightness. Those indicators have been chosen for their capacity to highlight the phenological distinction of the vegetation.

4.1.2.2 Reference Data

Several input datasets have been used to produce this prototype:

- Open Street Map (OSM) data, which contains roads and railways;
- EU-Hydro, a database of river networks and coastal lines – in beta version for the time being;
- Water & Wetness (WAW) product, which is one of the 2015 HRL layer, characterizing permanent waters and wetlands;
- Tree Cover Density (TCD), one of the layers of the HRL 2015 forest dataset;
- Imperviousness Degree (IMD) layer from the HRL 2015, representing the pan-European sealed areas;
- Finally, the CLC dataset from 2012 that has been partially used to discriminate cropland types.

More details on their usage can be found in section 5.1.1.2.

4.1.2.3 Hardbone

Using the listed reference data above, the first step of the prototype implementation is the creation of a "hardbone". This could be defined as a fixed skeleton which represents persistent object borders in the landscape. It has been done with an automated approach. This hardbone is based on vector data, mainly, such as roads, railways and river networks. A priority order has been set between those datasets to ensure that each polygon has only one label. A more detailed description of the process can be found in section 5.1.3.2.

4.1.2.4 Softbone

The second step is the creation of a "softbone", a process in which persistent landscape objects, represented by polygons, are set through the segmentation of the multi-temporal Sentinel images. This approach is reviewed in more details in section 5.1.3.3.

4.1.3 Classification algorithm

As already explained in the WP33 (AD07), Random Forest (RF) classification combines many decision trees to obtain better predictive performance. Each decision tree is calibrated on a selection of random subset. Algorithms such as RF have recently received increasing interest (Rodriguez-Galiano et al., 2012) because they have proved to be more accurate and more robust to noise than single classifiers (Shang & Breiman, 1996). Ensemble classifier, like the multitudes of decision trees in RF, are known to perform better than an individual classifier can. Breiman (Breiman, 2001) introduced RF in 2001 which presents many advantages for its application in remote sensing:

- Efficiency on large data bases;
- Thousands of input variables without variable deletion;
- Estimation of which variables are important in the classification;
- Relative robustness to outliers and noise;
- Computational lightness compared to other tree ensemble methods (e.g. Boosting);
- Much less sensitivity to overtraining or overfitting.

A RF consists of a combination of classifiers where each one of those contributes with a single vote to the assignation of the most frequent class detected for the input vector. This grant RF special characteristics which make it substantially different to traditional classification trees (CT). In fact, a RF increases the diversity of the trees by making them grow from different training data subsets created through the process.

5 Prototype Implementation

This chapter shows the implementation for the New Land Cover prototype at a 10m spatial resolution and with a minimum mapping unit (MMU) of 10m by 10m (NLC_2017_010m_SW_03035_prototype_v01). All thematic classes present in the HRL are present on the demonstration site, as well as various cropland types – in this implementation, several types of land cover characterization have been tested.

The section 5.1 tackles the data integration and the resulting processing setup. The following section 5.2 presents the results and the validation and accuracy assessment of the and cover prototype and lastly, the dataset and its metadata, referring to P45.2 – Data Sets of New LC/LU Products (Issue 1) are described in the section 5.3.

5.1 Data and Processing Setup

The integration of the EO input data as well as ancillary data are described in section 5.1.1. The pre-processing of the Sentinel input data is explained in section 5.1.2 and the experimental set-up for preparing and performing the classification is detailed in section 5.1.3.

5.1.1 Input Data and Data Integration

In this first phase, only optical S-2 data from 2017 have been used, complemented by several reference data for the year 2015.

5.1.1.1 Image data

The raw input dataset for the creation of this prototype over the South-West demonstration site is composed by Sentinel-2 data from 2017/01/03 to 2017/11/14. When the cloud presence was too important, images were not used. The Table 5-1 shows the number of images used by S-2 tiles over the all demonstration site.

Table 5-1 - Number of S-2 images used in the composite input dataset for each tile

	30TYP	30TYN	31TCJ	31TCH	31TDJ	31TDH	total
Sentinel-2	21	25	13	25	40	26	150

It has been decided to split the data into three three-month periods (January to Mars, April to June and July to September). The aim of this approach was indeed to establish a variability in the spectral signature over the study year for each different class, through the combined use of several indices such as NDVI, NDWI or BRI (see 4.1.1 for a description of the indices and their time features). For all those reasons, the dataset used was very restricted (only 150 S-2 images were used). In order to increase the accuracy of the results for further classification, it should be useful to integrate S-1 or fused images from S-3 in the next iteration of the task 4.

5.1.1.2 Reference data

Several reference datasets were required for the creation of the hardbone. Those datasets were then also used for the calibration and validation of the classification model.

The hardbone is built upon:

- OSM data over France, Andorra and Spain: lines of road and railways, depending on their importance. It has been decided to keep only the primary, secondary and tertiary roads and the associated links;
- EU-Hydro beta version: canals and rivers have been integrated without distinction;
- Water & Wetness product from HRL 2015: only Permanent Water Bodies (class 1) were used;

- Tree Cover Density product from HRL 2015: areas with a minimum of 30% of tree cover density were integrated;
- IMD from HRL 2015: areas with a minimum of 10% of imperviousness degree at least were considered;
- CLC from 2012: a few CLC classes were selected in order to have a schematic cropland area map. They are listed in the section 5.1.3.2.

The data used for the calibration and validation of the classification model are:

- High-Resolution Layers 2015 that have been validated, to collect and generate samples. Imperviousness product was used to create urban samples, the TCD layer for forest samples, WAW product for water samples and the Grassland mask for grassland samples.
- Land Parcel Identification System (LPIS) 2016 over France was used as a basis to find cropland samples in the collection of 2017 images – most fields remained the same from 2016 to 2017. Only the most representative croplands were selected: the first six categories represent 78.23% of the overall number of agricultural parcels for this particular region as describe in the Table 5-2.

Table 5-2 - Main cropland types over the demonstration site from French LPIS

Crop code	Label	Number of parcels	%
VRC	Vineyard : Wine grape	57 311	24.09%
BTH	Common Winter Wheat	37 683	15.84%
TRN	Sunflower	30 559	12.84%
MIS	Maize	25 715	10.81%
BDH	Durum Winter Wheat	20 316	8.54%
ORH	Winter Barley	14 554	6.12%
Total for these 6 classes		186 138	78.23%
Number total of parcels		237 942	100%

5.1.2 Pre-processing

The ECoLaSS South-West demonstration site in France and Spain is comprised of the footprints of six adjacent Sentinel-2 tiles (30TYN, 30TYP, 31TCH, 31TDH, 31 TCJ and 31TDJ) for which Sentinel-2 data were processed as explained in the deliverables of the WP32. The methods to obtain spatio-temporally consistent optical images with top of atmosphere reflectance values are:

- Atmospheric correction;
- Topographic normalization;
- Clouds, cloud shadow and snow masking.

ATMOSPHERIC CORRECTION

The Copernicus Sentinel-2 data, processed at level L2A by CNES for THEIA Land data center are available for download and corrected atmospheric effects, including adjacency effects. These atmospheric corrections include the light absorption by air molecules and the light scattering by molecules and aerosols (Hagolle et al., 2015).

Several models may be used to perform atmospheric corrections. In the case of the MAJA software, the MACCS processor is the model used. It pre-computes look-up tables using an accurate radiative transfer code (with successive orders of scattering), that simulates the light propagation through the atmosphere. The MACCS/MAJA method combines different approaches to obtain robust estimates of aerosol optical thickness.

TOPOGRAPHIC NORMALISATION

A topographic correction is necessary if the sites are characterized by mountainous terrain as it is for the South-West Demonstration site. The topography can significantly influence the radiometric properties of the signal received from the satellite (see Wulder & Franklin, 2012). This effect is caused by the different lighting angles resulting from the topography (see Gallaun et al., 2007). The aim of a topographical correction is to compensate for the differences in reflectance intensity between the areas with varying slope, exposure and inclination and to obtain the radiation values that the sensor would have in the case of a flat surface.

The Sentinel-2 data using the MAJA software and available for download are corrected from the topographic effects at the online THEIA data center.

CLOUD, CLOUD SHADOW AND SNOW MASKING

The MAJA cloud detection method (Hagolle et al., 2010) is based on a number of threshold tests including the cirrus band (B10). Additionally, multi-temporal tests are carried out to detect clouds by measuring a steep increase of the blue surface reflectance. Finally, the correlation of the pixel neighbourhood with previous images is calculated to avoid over detections based on the assumption that two different clouds at the same location on successive dates will not have the same shape. If a large correlation is observed, the pixel is excluded from the cloud mask.

5.1.3 Experimental Setup

The developed processing chain is able to process a large amount of input data within a reasonable amount of time to provide the classification results. The achieved level of automation ensures the effective application of the process to map different cropland types of almost the entirety of Europe.

The workflow for the production of the New Land Cover Prototype is listed hereafter:

1. The images processing, subdivided in:
    a. Stacking images into multiband
    b. Applying the cloud mask on selected image data
    c. Regrouping the data per 3-month period (i.e. three three-months periods)
    d. Computing the spectral indices (NDVI, NDWI ad BRI)
    e. Computing temporal features of those spectral indices (maximum, mean, minimum, 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles and standard deviation)
    f. Concatenating all the previous data
2. The generation of the hardbone based on ancillary data
3. The Large-Scale Mean Shift Segmentation (LSMSS) computation to create the softbone
4. The merge of the skeleton and this latter segmentation, with a MMU of 5ha
5. The production of an independent pixel based classification:
    a. Computing of the image spatial statistics
    b. Generating the sampling based on reference database (50% of the samples for calibration and 50% of the samples for validation)
    c. Launching the RF classification
    d. Applying a majority filter to harmonize the results
    e. Calculating automatically the associated confusion matrix
    f. Reprocessing the classification if needed (optional)
6. The aggregation of the results with the softbone and the hardbone.

The results could be treated further by mosaicking of the results obtained per tiles into a single map, by eliminating isolated pixels (which are merged into the surrounding surfaces).

5.1.3.1 Image Processing

All these time features, described in sections 4.1.1 and 4.1.2.1, that have been computed for each spectral index and each three-month period are stacked over a S-2 tile. The choice of the indices enforce a clear discrimination between – here, the cloud presence does not pose any trouble insofar as there is at least one cloudless pixel per temporal period for each considered tile - and the three periods create a time series for the spectral signature of each land cover type.

5.1.3.2 Hardbone

The hardbone can be generated separately from the image processing. It consists on, first, collecting data that were previously listed in section 5.1.1.2, then, extracting only relevant information following some geometric rules and finally merge all these data with a priority order.

OSM DATA:

OSM data is easily available on the net. The grids are classified by countries or regions, and can be selected according to specific requirements. For the realisation of the hardbone, the focus has been put on the road and railway networks that can be found in the shapefile gis_osm_roads_free_1.shp and gis_osm_railways_free_1.shp.
The overall road network is too dense to be used at the highest resolution - therefore only primary, secondary and tertiary roads (and their associated links) have been selected - and it should be noted that even in mountainous areas, the resulting network is dense enough.
In order to work with polygons, a generic 5m buffer along these linear features has been applied. At the time of this study, no information about the width of each road could be found. However, should this information become available, the accuracy of the hardbone could be further improved.
The OSM data also contain information about buildings or waterways. However, their extraction and integration into the hardbone have not been as straightforward as for the road networks. What is more, those features would be redundant or even conflicting with the use of the HRL products such as IMD 2015 and WAW 2015 - it has therefore been decided to exclude them, in order to simplify the aggregation rules.

EU-HYDRO:

EU-Hydro data beta version is available on the Land Copernicus website. Several shapefiles have been used, depending on their content:

- the files called River_Net_p.shp and Canals_p.shp have been used to generate surface rivers, greater than 50 meters;
- the file called River_Net_l.shp and Canals_p.shp for linear rivers whose surface is smaller than 50 meters.

A 5m buffer has also been applied to create polygons, since no information on the real width of the rivers is available at the moment.

HRL PRODUCTS 2015

HRL Products has been useful for generalization of large urban, forest and water landscape. They provide information for the whole EEA-39 area on specific land covers. These products are computed automatically from time series satellite imagery from 2015 (+/- one year) and validated, so they are accurate enough to create as such a first-hand estimate of the European landscape that will be further polished by the classification results.
From the Water & Wetness Product, only Permanent Water Bodies (class 1) have been extracted. This simple rule is set to avoid temporary fluctuation of the water levels.
From the Imperviousness Product, sealed soils with a minimum of 10% of Imperviousness degree were extracted to create coherent and large urban limits without missing small isolated buildings.

From the Tree Cover Density Layer, the threshold was fixed at a minimum of 30% of tree cover to avoid sparse forest.
Finally, a MMU of 0.5ha (corresponding to 5 by 5 pixel-wide polygon) has been applied.

CLC (2012):

At the time this work has been realized, a validated Cropland High Resolution Layer Prototype was not available, it was then decided to select a few CLC codes in order to have a schematic cropland area map. They are listed in the Table 5-3 below. A significant MMU of 5ha has been applied.

Table 5-3 - Selection of CLC codes for cropland areas

CLC_CODE	LABEL2	LABEL3
211	Arable land	Non-irrigated arable land
212	Arable land	Permanently irrigated land
213	Arable land	Rice fields
221	Permanent crops	Vineyards
222	Permanent crops	Fruit trees and berry plantations
223	Permanent crops	Olive groves
241	Heterogeneous agricultural areas	Annual crops associated with permanent crops
242	Heterogeneous agricultural areas	Complex cultivation patterns
243	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation

GENERATION OF THE FINAL HARDBONE

All these data are finally merged into a unique shapefile, in a specific priority order:

- OSM road and railway network;
- EU-Hydro river network;
- Permanent Water Bodies;
- Tree Cover Density;
- Imperviousness;
- Cropland areas from CLC.

This automated approach might be very time-consuming due to the large number of polygons. It should also be noticed that the shapefile format might be not very efficient to carry the remaining part of the processing.

The final result can be seen on Figure 5-1. Some areas are devoid of any hardbone information - mainly in the Pyrénées area and punctually nearby urban areas. This can be linked to potential complex LC/LU presence. However, this result already sketches an interesting idea of the main landscape over the six S-2 tiles.

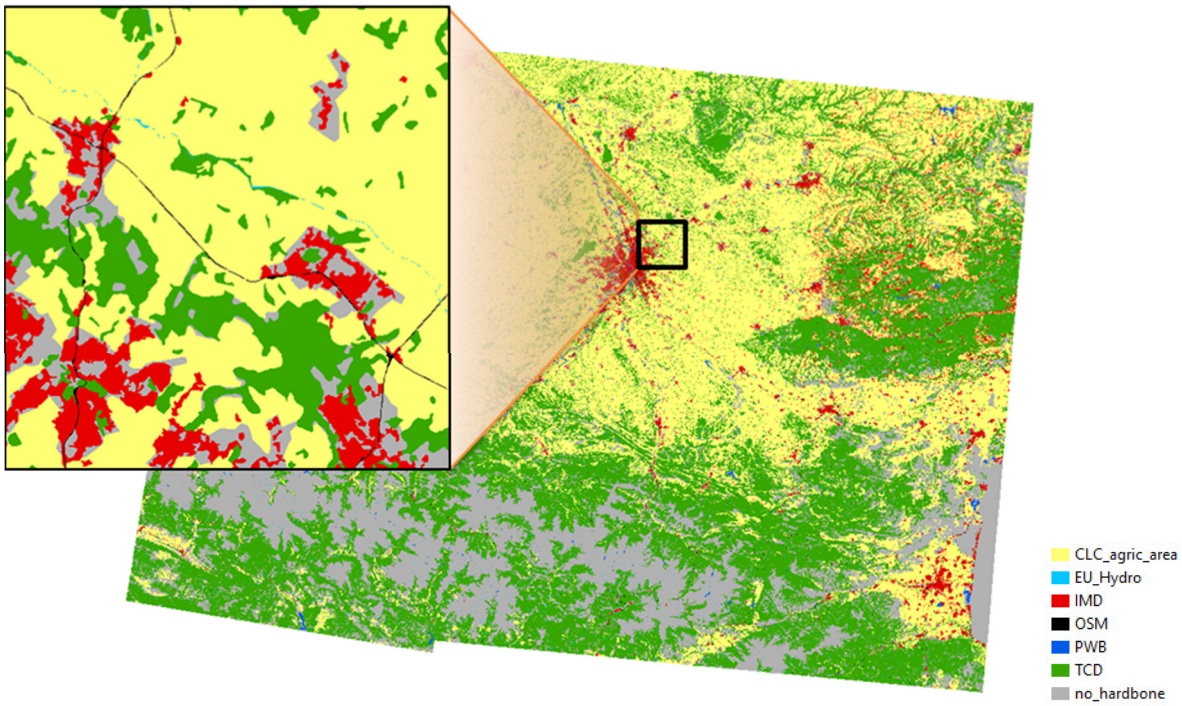


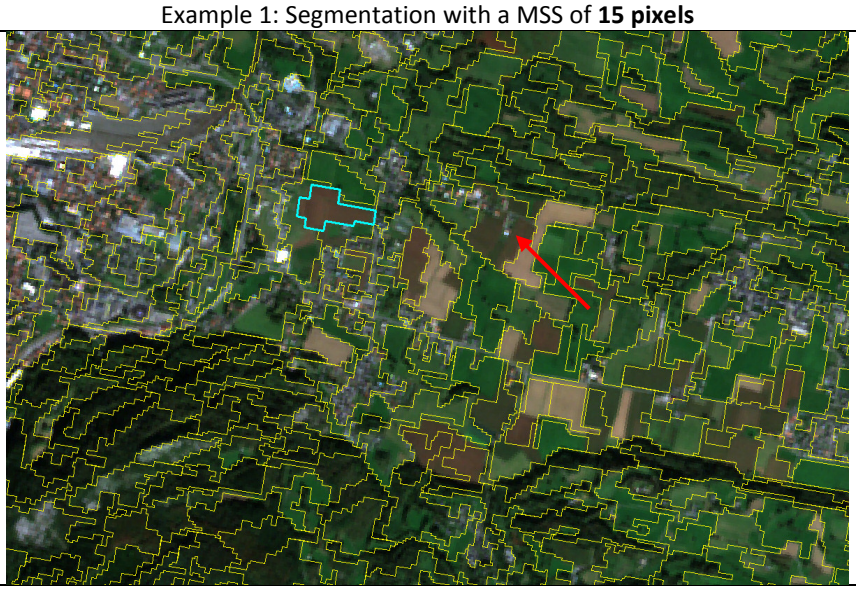
Figure 5-1 - Hardbone over the South-West demonstration site

5.1.3.3 Softbone

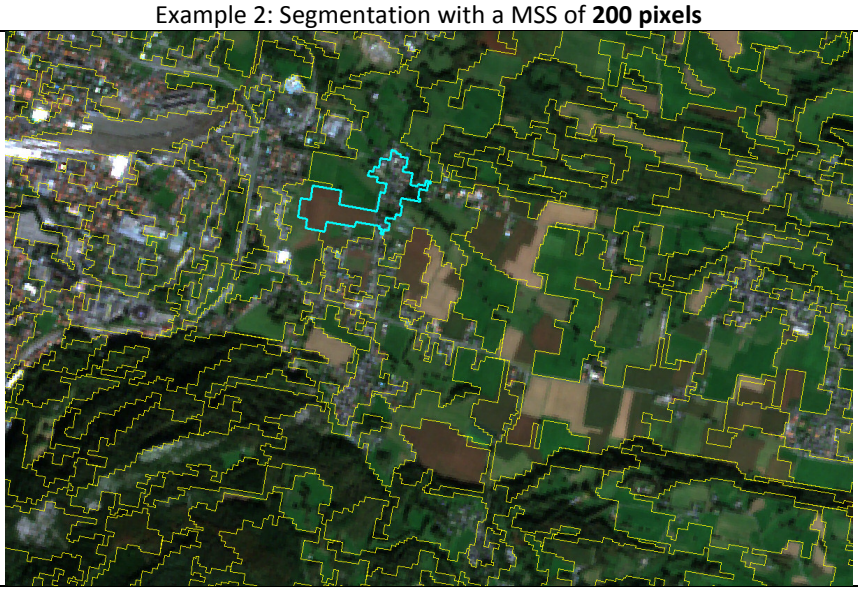
The second step is the creation of a "softbone": a collection of persistent landscape objects, modelled in the shape of polygons, through image segmentation of multi-temporal Sentinel data sets.

A selection of images, with a cloudfree surface represented at least 25% of the AOI, are layer-stacked and fed to a LSMSS. The generated shapefiles contain mean and variance information for polygons regarding the possible membership to each layer. The determining parameter, the Minimum Segment Size (MSS) in pixels, has to be adjusted by trials and errors, and may be depending on the considered region. As shown on Figure 5-2, a segmentation with a minimum segment size of 15 pixels is too dense, too precise and the corresponding shapefile is very huge (4.318.908 polygons for one S-2 tile for example). A minimum size of 200 pixels seems to be a good compromise - but too many details, such as small urban areas, are lost in the process. On the second image, the selected polygon (in blue) corresponds to a cropland field. However, the red arrow shows small sealed areas merged with a field. The third image created with a minimum segment size of 500 exhibits huge concatenations of segments with different land covers, and is visibly not adapted. For this prototype, the segmentation with a minimum segment size of 15 pixels was finally used in order to avoid generalisation of different land covers.

Example 1: Segmentation with a MSS of **15 pixels**



Example 2: Segmentation with a MSS of **200 pixels**

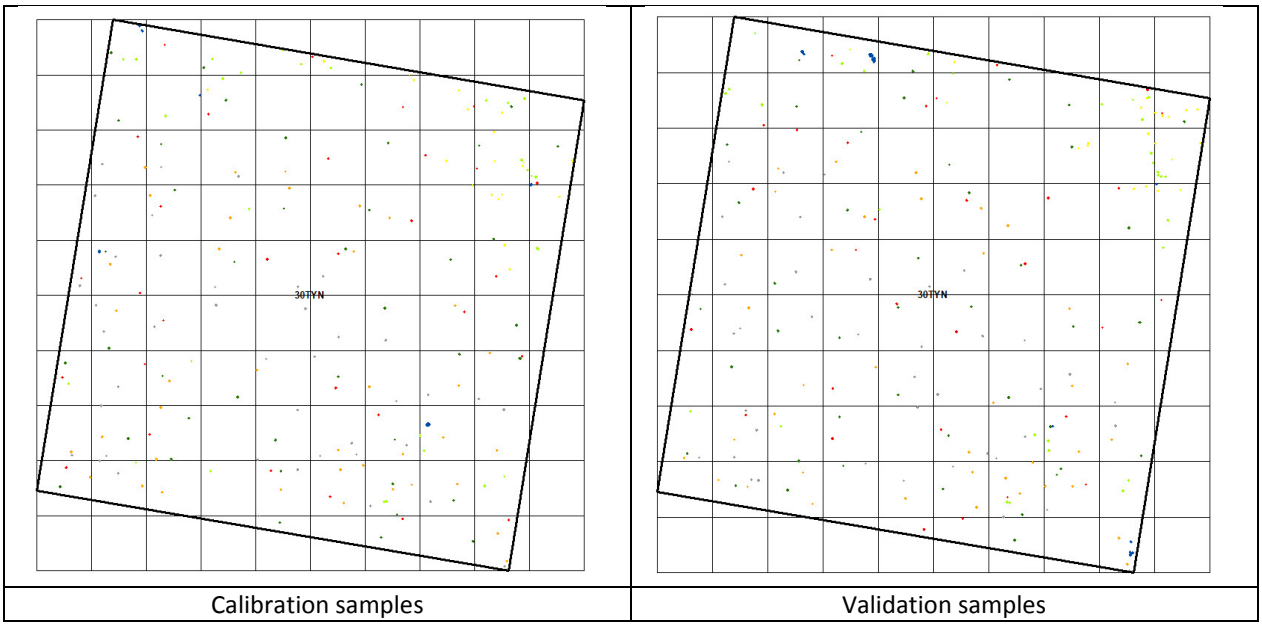


Example 3: Segmentation with a MSS of **500 pixels**

Figure 5-2 - Example of 3 different segmentations/softbones

5.1.3.4 Reference datasets generation

The creation of a dataset of samples based on the validated HRL 2015 products is required for the calibration and for the validation of the result of the classification. As input for the supervised machine learning algorithms, a clear set of training data is required. Those samples must be exempt from anomalies and must be a suitable statistical representation of the considered area. They must be representative of the whole study site by covering all the reflectance variations of the classes. They were selected thanks to a fishnet for each S-2 tile for a better and smoother distribution over the whole area (see example in Figure 5-3). 50% of the collected samples were used for calibration whereas 50% were used for validation.



Calibration samples | Validation samples

Figure 5-3 - Example of distribution of calibration and validation samples

To this end, reliable samples have been derived from the historical HRL 2015 IMD, Forest, Grassland and Water. An automated random selection has been processed on each Sentinel-2 tile. To characterize cropland types, additional samples have been selected in the LPIS data. As mentioned in section 5.1.1.2, 6 main cropland types (Vineyard, Common Winter Wheat, Sunflower, Maize, Durum Winter Wheat and Winter Barley) were selected on the French LPIS for their regional representability. As seen in Figure 5-4, the distribution of cropland parcels is not homogeneous.

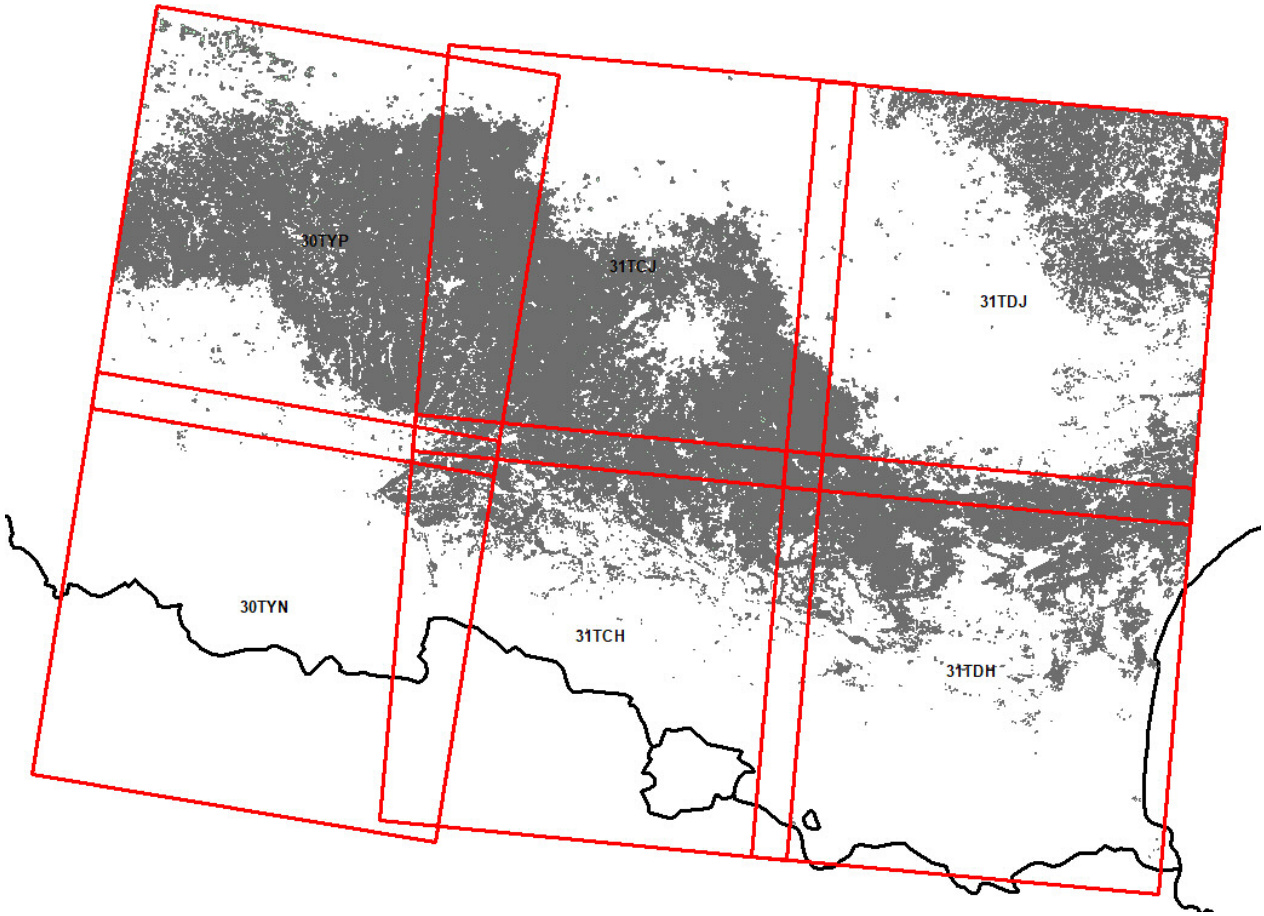


Figure 5-4 - Distribution of the overall LPIS data (grassland excluded)
© RPG France – Graphic Parcel Register, Institut Géographique National (IGN), 2016

5.1.3.5 Classification

The next step consists in a random forest algorithm to classify the different classes - according to the tests conducted in WP33.

A majority filter is also applied after the classification - this presents two advantages:

- To harmonize the results, as it smooths the pixel-based classification and lets appear a more natural landscape;
- To merge isolated pixels into greater ensembles, even if some of them will reappear in the post-processing during the mosaicking.

Finally, once the classification has been launched and a result produced, a first confusion matrix is automatically computed based on the validation samples to estimate the accuracy of the results per S-2 tile. Finally, the last part of the classification consists in ensuring that there are no isolated pixels, especially on the corridors between two tiles.

5.1.3.6 Validation analysis procedure

A confusion matrix is automatically generated after the classification ran, based on the sampling data detailed at section 5.1.3.4, allowing the adjustment of the number of classes (by fusion, suppression or creation of new classes). Indeed, if the overall accuracy (OA) is lower than expected, or if the omission/commission rates are too high, or if the look and feel simply does not seem googd enough, the classification is re-processed.

During the creation of this prototype, the classes "bare-soil" and "natural grassland" were added after a first run of the classification. A final nomenclature of 7 classes has been adapted and can be seen in Table 5-4.

Table 5-4 – Final nomenclature for the New Land Cover Prototype

Class	Label
1	Grassland
2	Cropland
3	Forest
4	Water
5	Urban area
6	Bare soil
7	Natural grassland

Considering the time schedule for the Task 4, less than ten different LC classes have been considered. However, the results reached a good quality. It should be noticed that for the time being, two S-2 tiles (30TYN and 31TCH) were more difficult to classify than the others because of the complexity of their respective landscapes. The differences between Atlantic, Alpine and Mediterranean bioregions rendered the classifications not homogeneous from the north to the south.
For the second iteration of this task, it could be interesting to classify each bioregion individually instead of using the artificial geometry created by the S-2 tile mesh.

5.2 Classification Results and Validation

To be able to determine the global accuracy of the New Land Cover layer, several validations were conducted. The results are presented in this section.

The first analysis is based on the confusion matrices automatically generated by the classification algorithm on each S-2 tile that are displayed in the Table 5-5 to the Table 5-10. As said in the previous section (5.1.3.6), the first classification attempts made with various cropland types exhibited quite a lot of confusion between the 6 varieties that had been chosen. Since the differentiation between cropland types is not the purpose of this prototype, it has then been decided to merge all those types in a class later named 'all cropland' for the final result.

Table 5-5 – Automatically generated confusion matrix for the S-2 tile 30TYN

30TYN		Reference labels							total	User Accuracy (UA)
		1 - grassland	3 - forest	4 - water	5 - urban	6 - bare soil	7 - natural grassland	2 - all cropland		
Produced labels	1 - grassland	2158	36	0	0	0	309	124	2627	82.15%
	3 - forest	257	2166	0	34	0	252	0	2709	79.96%
	4 - water	0	16	2780	1	96	62	1	2956	94.05%
	5 - urban	53	0	0	1187	3	115	96	1454	81.64%
	6 - bare soil	0	4	75	537	1077	188	0	1881	57.26%
	7 - natural grassland	83	218	0	45	40	1713	3	2102	81.49%
	2 - all cropland	209	117	0	0	0	28	561	915	61.31%
total		2760	2557	2855	1804	1216	2667	785	14644	
Producer Accuracy (PA)		78.19%	84.71%	97.37%	65.80%	88.57%	64.23%	71.46%		79.50%

Table 5-6 – Automatically generated confusion matrix for the S-2 tile 30TYP

30TYP		Reference labels					total	UA
		1 - grassland	3 - forest	4 - water	5 - urban	2 - all cropland		
Produced labels	1 - grassland	4287	57	0	504	1048	5896	72.71%
	3 - forest	3	2934	0	15	1	2953	99.36%
	4 - water	0	7	3057	0	0	3064	99.77%
	5 - urban	18	2	0	1704	32	1756	97.04%
	2 - all cropland	634	1	0	348	8557	9540	89.70%
total		4942	3001	3057	2571	9638	23209	
PA		86.75%	97.77%	100.00%	66.28%	88.78%		88.50%

Table 5-7 – Automatically generated confusion matrix for the S-2 tile 31TCH

31TCH		Reference labels							total	UA
		1 - grassland	3 - forest	4 - water	5 - urban	6 - bare soil	7 - natural grassland	2 - all cropland		
Produced labels	1 - grassland	3718	297	0	135	56	15	1211	5432	68.45%
	3 - forest	0	1986	5	90	0	44	119	2244	88.50%
	4 - water	125	0	884	0	0	0	43	1052	84.03%
	5 - urban	0	15	0	1683	0	5	0	1703	98.83%
	6 - bare soil	0	1	0	74	1053	19	0	1147	91.80%
	7 - natural grassland	0	88	0	22	0	1335	0	1445	92.39%
	2 - all cropland	451	57	0	151	16	1	4042	4718	85.67%
total		4294	2444	889	2155	1125	1419	5415	17741	
PA		86.59%	81.26%	99.44%	78.10%	93.60%	94.08%	74.64%		82.86%

Table 5-8 – Automatically generated confusion matrix for the S-2 tile 31TCJ

31TCJ		Reference labels					total	UA
		1 - grassland	3 - forest	4 - water	5 - urban	2 - all cropland		
Produced labels	1 - grassland	3863	147	0	6	828	4844	79.75%
	3 - forest	0	1217	0	36	8	1261	96.51%
	4 - water	0	0	1640	0	0	1640	100.00%
	5 - urban	61	38	0	1842	70	2011	91.60%
	2 - all cropland	437	351	0	2	10049	10839	92.71%
total		4361	1753	1640	1886	10955	20595	
PA		88.58%	69.42%	100.00%	97.67%	91.73%		90.37%

Table 5-9 – Automatically generated confusion matrix for the S-2 tile 31TDH

31TDH		Reference labels							total	UA
		1 - grassland	3 - forest	4 - water	5 - urban	6 - bare soil	7 - natural grassland	2 - all cropland		
Produced labels	1 - grassland	6489	62	0	365	0	126	241	7283	89.10%
	3 - forest	4	1459	0	68	0	92	75	1698	85.92%
	4 - water	0	5	3735	0	0	0	0	3740	99.87%
	5 - urban	5	10	0	1847	32	13	18	1925	95.95%
	6 - bare soil	0	16	0	101	1539	106	26	1788	86.07%
	7 - natural grassland	162	50	0	180	92	2577	49	3110	82.86%
	2 - all cropland	50	6	0	2	0	22	5024	5104	98.43%
total		6710	1608	3735	2563	1663	2936	5433	24648	
PA		96.71%	90.73%	100.00%	72.06%	92.54%	87.77%	92.47%		91.98%

Table 5-10 – Automatically generated confusion matrix for the S-2 tile 31TDJ

31TDJ		Reference labels					total	UA
		1 - grassland	3 - forest	4 - water	5 - urban	2 - all cropland		
Produced labels	1 - grassland	4811	16	0	1047	1264	7138	67.40%
	3 - forest	7	3896	49	10	26	3988	97.69%
	4 - water	0	0	1804	0	0	1804	100.00%
	5 - urban	91	6	0	2117	21	2235	94.72%
	2 - all cropland	981	277	0	514	3763	5535	67.99%
total		5890	4195	1853	3688	5074	20700	
PA		81.68%	92.87%	97.36%	57.40%	74.16%		79.18%

These results are quite promising, especially for the S-2 tile 31TDH for which the OA is above 91%. However, some tiles exhibit a slightly less sharpened accuracy: the OA is around 79% for the tiles 31TDJ and 30TYN, even if the look and feel aspect does point to any blatant anomaly.

Once the results are deemed precise enough (according to the confusion matrix OA for all tiles as well as the look and feel), the same analyse has been made at a global scale as seen in the Table 5-11.

Table 5-11 – Automatically generated confusion matrix over the South-West demonstration site

DEMOSITE SOUTH WEST		Reference labels							total	user accuracy
		1 - grassland	3 - forest	4 - water	5 - urban	6 - bare soil	7 - natural grassland	2 - all cropland		
Produced labels	1 - grassland	25326	615	0	2057	56	450	4716	33220	76.24%
	3 - forest	271	13658	54	253	0	388	229	14853	91.95%
	4 - water	125	28	13900	1	96	62	44	14256	97.50%
	5 - urban	228	71	0	10380	35	133	237	11084	93.65%
	6 - bare soil	0	21	75	712	3669	313	26	4816	76.18%
	7 - natural grassland	245	356	0	247	132	5625	869	7474	75.26%
	2 - all cropland	2762	809	0	1017	16	51	15446	20101	76.84%
total		28957	15558	14029	14667	4004	7022	21567	105804	
producer accuracy		87.46%	87.79%	99.08%	70.77%	91.63%	80.11%	71.62%		83.18%

According to this final automatically generated confusion matrix, some classes are very well identified such as the forest and water classes. On the other hand, some classes still show confusion errors – for instance, there are omissions on urban and cropland classes, and commissions on grassland, bare soil, natural grassland and cropland classes.

However, the OA remains quite good compared to the preliminary results obtained with the first classification on this study area:

- OA equals to 83%;
- PA ranges from 71 to 99%;
- UA ranges from 75 to 97%;
- Kappa equals to 83%;
- F-Score is between 0.37 and 0.48.

The prototype could certainly be improved with different approach such as various modification of the classes, a refinement of the samples used in the training as well as in the validation, the use of textural indices or even the use of S-1 data. All of these possibilities will be tested in the next phase of this task 4.

Once the confusion matrix presents satisfying results for the classification step, the pixels composing the softbone are aggregated with the classification outputs for S-2 tiles. The resulting products are then mosaicked and isolated pixels are merged into larger ensembles on the final prototype. The relative look and feel is shown below in the Figure 5-5. The different land cover classes are quite well identified: a real "landscape coherence" can be seen.

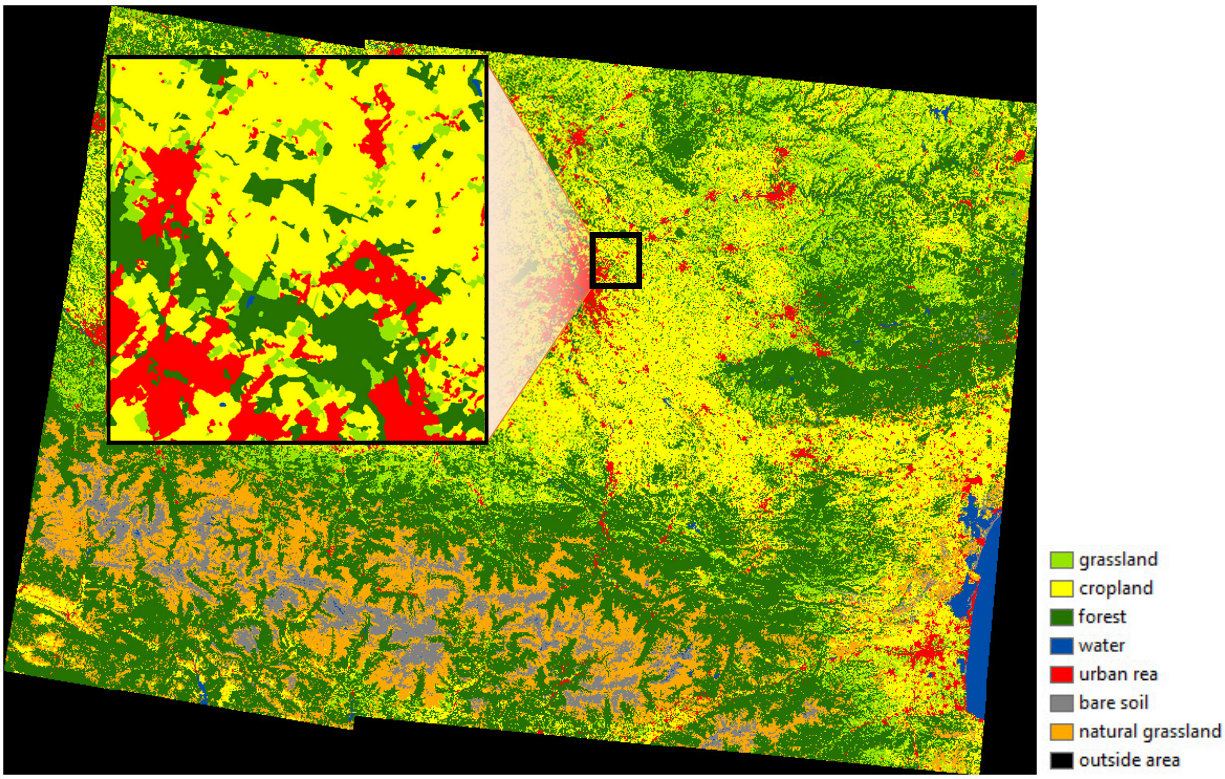


Figure 5-5 - NLC_2017_010m_SW_03035_prototype_v01

An independent analysis, based on other validation samples independently created from the entire process described previously, has been made in order to have a more objective analysis. The confusion matrix, seen in Table 5-12, confirms the results automatically acquired through the classification algorithm.

Table 5-12 – Independently established confusion matrix for the South-West demonstration site on the final New Land Cover prototype

NLC v01		REFERENCE								
		1 - grassland	2 - cropland	3 - forest	4 - water	5 - urban	6 - bare soil	7 - natural grassland	total	user accuracy
PRODUCT	1 - grassland	362.5922	0.2049	8.7881	0.0000	0.6597	0.0000	5.7663	378.0112	95.92%
	2 - cropland	61.2638	1147.6622	17.5761	0.0000	2.3090	1.5869	7.9873	1238.3853	92.67%
	3 - forest	10.7914	0.0000	1256.0608	0.0000	1.3195	1.0579	18.7406	1287.9702	97.52%
	4 - water	0.0000	0.0000	2.9294	45.4819	0.0000	1.0579	0.0000	49.4692	91.94%
	5 - urban	1.5587	0.0000	0.0000	0.0000	129.1757	2.1158	0.0000	132.8503	97.23%
	6 - bare soil	1.5587	0.0000	0.0000	0.0000	0.0000	97.7125	10.0911	109.3623	89.35%
	7 - natural grassland	2.3381	0.0000	26.3642	0.2049	0.0000	9.5213	349.5230	387.9515	90.09%
	total	440.1030	1147.8671	1311.7185	45.6868	133.4640	113.0524	392.1083	3584.0000	
	producer accuracy	82.39%	99.98%	95.76%	99.55%	96.79%	86.43%	89.14%		94.54%

5.3 Prototype Specifications

This section provides a description of the dataset properties and metadata for the implemented prototypes, also referring to "*P45.2a - Data Sets of New LC/LU Products*".

Within ECoLaSS, a standardised and harmonised product file naming convention for all prototypes has been developed which is oriented along the already existing naming convention of the CLMS High Resolution Layers. The naming convention consists of the following 7 descriptors:

THEME YEAR RESOLUTION EXTENT EPSG TYPE VERSION
as follows:

THEME

3 letter abbreviation for main products (DLT (dominant leaf type), TCC (tree cover change), GRA (grassland), IMD (imperviousness degree), IMC (imperviousness change classified), CRT (crop type), CRM (crop mask) and NLC (new land cover products).

REFERENCE YEAR

2017 in four digits; change products in four digits (e.g. 1517)

RESOLUTION

Four-digit (020m and 010m)

EXTENT

2-digit code for demonstration-sites (CE (central), NO (north), WE (west), SW (southwest), SE (southeast), SA (South Africa), ML (Mali))

EPSG

5-digit EPSG code (geodetic parameter dataset code by the European Petroleum Survey Group) "03035" for the European LAEA projection

TYPE

prototype

VERSION

3-digit code “v01”

EXAMPLE:

”DLT_2017_010m_NO_03035_prototype_v01.tif” stands for: Dominant Leaf Type, 2017 reference year, 10m, Demonstration-site North, European projection (EPSG: 3035), prototype, version 01

”TCC_1517_020m_NO_03035_prototype_v01.tif” stands for: Tree Cover Change, 2015-2017 change period, 20m, Demonstration-site North, European projection (EPSG: 3035), prototype, version 01

The raster products of the ECoLaSS prototypes are delivered as GeoTIFF (*.tif) with world file (*.tfw), pyramids (*.ovr), attribute table (*.dbf) and statistics (*.aux.xml), enabling an instant illustration and analysis of the products within Geographic Information System (GIS) software. Each product is accompanied with a product-specific colour table (*.clr) and INSPIRE-compliant metadata in XML format.

Metadata are provided together with the products as INSPIRE-compliant XML files according to the EEA Metadata Standard for Geographic Information (EEA-MSGI). EEA-MSGI has been developed by EEA to meet needs and demands for inter-operability of metadata. EEA’s standard for metadata is a profile of the ISO 19115 standard for geographic metadata and contains more elements than the minimum required to comply with the INSPIRE metadata rules. Detailed conceptual specifications of the EEA-MSGI and other relevant information on metadata can be found at http://www.eionet.europa.eu/gis.

The prototype specifications for the NLC layer are listed in Table 5-13.

Table 5-13 – Detailed specifications for primary 10m NLC status layer

Products
New Land Cover 2017 – NLC_2017_10m
Extent Demo site South-West
Geometric resolution Pixel resolution 10m, grid to fully conform to the EEA Reference Grid.
Coordinate Reference System European ETRS89 LAEA projection
Geometric accuracy (positioning scale) Less than one pixel (According to ortho-rectified satellite image base delivered by Theia.)
Thematic accuracy 90.00% of overall accuracy
Data type 8bit unsigned Raster, compressed with LZW
Minimum mapping unit (MMU) One pixel (10m*10m)
Necessary attributes Raster value, count, class name

Raster coding (Thematic pixel values)
New Land Cover 2017 – NLC_2017_10m 1: grassland 2: cropland 3: forest 4: water 5: urban area 6: bare soil 7: natural grassland 255: outside Area
Metadata XML metadata files are to be produced according to INSPIRE metadata standards
Delivery format GeoTIFF

The associated color palette can be found in the Table 5-14.

Table 5-14 – Palette used for primary 10m NLC status layer

New Land Cover 2017 – NLC_2017_10m					
Class Code	**Class Name**	**Red**	**Green**	**Blue**	
1	Grassland	152	230	0	
2	Cropland	255	255	0	
3	Forest	38	115	0	
4	Water	0	77	168	
5	Urban area	255	0	0	
6	Bare soil	130	130	130	
7	Natural grassland	255	170	0	
255	Outside Area	0	0	0	

6 Conclusion and Outlook

The proposed New Land Cover Prototype, without fully reusing the HRL products, already shows a good accuracy of 94.5% on the entire South-West site, as demonstrated by the independent validation analysis – even with the presence of mountainous regions and the different bioregional characteristics.

A few commission and omission errors remain, which could certainly be improved with the reinforcement of textural indices beside the existing spectral ones. The spatial correlation emphases by the texture analysis will constitute a strong discriminant tool between classes such as cropland and natural grassland, or urban areas and bare soils.

The NLC prototype is still based on validated HRL products from 2015 – and the processing chain used to create it has enforced a priority order in the making of the hardbone to maintain a spatial coherence in the final product.

New methodologies to generate such a product could be set up on a fusion of the 5 existing HRL products, as well as the newly produced cropland type prototype, resulting from WP44 and could be potentially explored in the second phase of task 4.

Regarding the current methodology exposed in this report, several axes emerged as future improvement directions. The integration of S-1 images or fused S-3 images will definitively densify the time series, all the while bringing richer spatial and phenological information as input to the classifier.

Another aspect for a possible enhancement lies in the softbone process. It has been shown that the segmentation requires a fixed MSS as input parameter. Too small, this parameter drags the process and generates too many details that will ultimately be merged into one class, such as the diverse cropland types that are all fused into a single ‘cropland’ class. Too big, this parameter leads the segmentation to lose small details of the landscape, such as isolated urban areas.

The diversity of the size of those persistent objects can lead to either the production of two segmentations at different scale, with a MSS chosen to highlight only field-level objects, and another one chosen to enhance small details. Another way to achieve a similar result would be to classify the landscape without supervision, such as with a K-Means classification, based on phenological and textural criteria, and then use the resulting classes to perform separate segmentation on each of those classes, with a class-tailored MSS. The diversity of the objects present in the landscape would then be preserved by this multi-scale segmentation. The final softbone file would be a merge of those segmentations.

Those ideas will be explored in the second iteration of Task 4.

7 References

Breiman, L. (2001). Random Forests. *Machine Learning, 4*(1), 5-32.

Gallaun, H., Schardt, M. & Linser, S (2007): Remote Sensing Based Forest Map of Austria and Derived Environmental Indicators. *Proceedings of the ForestSAT 2007 Scientific Workshop, Montpellier, France.*

Hagolle, O., Huc, M., Villa Pascual, D., & Dedieu, G. (2010, August). A multi-temporal method for cloud detection, applied to FORMOSAT-2, Venμs, LANDSAT and SENTINEL-2 images. *Remote Sensing of Environment, 8*, 1747-1755.

Hagolle, O., Huc, M., Villa Pascual, D., & Dedieu, G. (2015, March). A Multi-Temporal and Multi-Spectral Method to Estimate Aerosol Optical Thickness over Land, for the Atmospheric Correction of FormoSat-2, Landsat, VENμS and Sentinel-2 images. *Remote Sensing, 7*(3), 2668-2691.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photgrammetry and Remote Sensing, 67*, 93-104.

Shang, N., & Breiman, L. (1996). Distribution based trees are more accurate. *Ionosphere, 33*(2), 351.

Wulder, M. A. & Franklin, S. E (2012): Remote sensing of forest environments: concepts and case studies. Springer.