D3.3 DATA PRODUCTS INITIAL REPORT

Project: Monitoring of Environmental Practices for Sustainable Agriculture Supported by Earth Observation

Acronym: ENVISION

This project has received funding from the European Union`s Horizon 2020 research and innovation programme under grant agreement No. 869366.

Document Information

Grant Agreement Number	869366	**Acronym**	ENVISION	
Full Title	Monitoring of Environmental Practices for Sustainable Agriculture Supported by Earth Observation			
Start Date	1st September 2020	**Duration**	36 months	
Project URL	https://envision-h2020.eu/			
Deliverable	D3.3- Data products initial report			
Work Package	WP3 - Earth Observation data products			
Date of Delivery	**Contractual**	M18	**Actual**	M18
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	National Observatory of Athens (NOA)			
Responsible Author	Mr. Iason Tsardanidis (NOA)			
Contributions from	Mr. Vasileios Sitokonstantinou (NOA), Mr. Athanasios Drivas (NOA), Dr. Haris Kontoes (NOA), Ms. Ifigeneia Tsioutsia (AgroApps), Dr. Panos Ilias (ILVO)			

Document History

Version	Issue Date	Stage	Description	Contributor
V0.1	14/2/2022	Draft	Input received from partners	NOA
V0.2	21/2/2022	Draft	Review document	DRAXIS
V1.0	28/2/2022	Final	Integration of input	

Disclaimer

This document and its content reflect only the author's view, therefore the EASME is not responsible for any use that may be made of the information it contains!

CONTENT

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

LIST OF TABLES

LIST OF FIGURES

The ENVISION project has received funding from the European Union’s Horizon 2020 research
and innovation programme under grant agreement No 869366

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

ABBREVIATIONS

Acronym	Full Term
AI4EO	Artificial Intelligence For Earth Observation
AL	Arable Land
AOI	Area Of Interest
AOI	Area Of Interest
API	Application Programming Interface
ARD	Analysis Ready Data
BC	Business Case
BSM	Burnt Scar Mapping
CAP	Common Agricultural Policy
CAPO	Cyprus Agricultural Payments Organization
CB	Control Body
CC	Cross-Compliance
CCTM	Cultivated Crop Type Maps
CD	Crops Diversification
CNN	Convolutional Neural Network
DIAS	Data and Information Access Services
DL	Deep Learning
DP	Data Product
DRXS	Draxis Environmental
EAA	Eligible Agricultural Area
EFA	Ecological Focus Area
EO	Earth Observation
ESA	European Space Agency
FOI	Field Of Interest
GA	Grand Agreement
GAEC	Good Agricultural and Environmental Conditions
GDAL	Geospatial Data Abstraction Library
GIS	Geographic Information System
IACS	Integrated Administration and Control System
ICT	Information and Communication Technologies
ILVO	Instituut voor Landbouw-, Visserij- en Voedingsonderzoek
JRC	Joint Research Centre
LAI	Leaf Area Index
LPIS	Land Parcel Identification System
LSTM	Long Short-Term Memory
LV	Vlammse Gewest
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NOA	National Observatory of Athens

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

NPA	National Paying Agency
OCS	Organic Control System Subotica
ODC	Open Data Cube
OTSC	On-The-Sport-Checks
PA	Paying Agency
PA	Producer Accuracy
PCA	Principal Component Analysis
PSRI	Plant Senescence Reflectance Index
RF	Random Forest
RNN	Recurrent Neural Network
RUSLE	Revised Universal Soil Loss Equation
SAVI	Soil-Adjusted Vegetation Index
SBI	Stubble Burning Identification
SCL	Scene Classification Level
SITS	Satellite Image Time-Series
SMOTE	Synthetic Minority Oversampling Technique
SMR	Statutory Management Requirements
SOC	Soil Organic Carbon
SVM	Support Vector Machines
UA	User Accuracy
UR	User Requirements
USLE	Universal Soil Loss Equation
VHR	Very High Resolution
VM	Virtual Machine
WP	Work Package
NVZ	Nitrate Vulnerable Zone

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

1) Executive Summary

The aim of this deliverable is to define the data products that will ultimately form the services of the ENVISION business cases. This deliverable is related to the following tasks of the “WP3 Earth Observation data products”:

- Task 3.3 Analytics on Vegetation and Soil Index Time-series
- Task 3.4 Cultivated crop type maps
- Task 3.5 Grassland mowing events detection
- Task 3.6 Soil condition monitoring
- Task 3.7 Crop growth Monitoring and identification of organic farming practices

D3.3 is related to the majority of the WP3, which designs and develops the EO data products of the ENVISION platform, which will address all the potential customers’ specific needs.

This deliverable is based on the “D3.2 Catalogue on auxiliary data and available repositories to be incorporated” and will be the basis for the next deliverable of these Tasks, more specifically the D3.7 Data products final report. This deliverable is also strongly connected to deliverable D3.4 which is also due M18. D3.3. reports on the methodology of the data products developed to provide solutions to the business case requirements. It elaborates on the business case user requirements, the specificities of the areas of interest and fully describes how to design the data products to address these. D3.4 complements this deliverable, showcasing the first results and validation outputs of the hereby presented data products. D3.4. also includes detailed sections on limitations and future work that will not be as extensive in this deliverable.

Specifically, the sections of this deliverable are:

Section 2 – ENVISION data products and Business cases: Contains the overview of the data products to be developed matched with the Tasks of the GA and the Business Cases.

Section 3 – User requirements - products: Contains the matching of all the user requirements with the data products

Section 4 – ENVISION Data and Tools: Presents all the business cases and their requirements of each of the end-users along with the methods implemented and data utilized.

Section 5 – Conclusions: Contains the main conclusions of this deliverable

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

2) ENVISION data products and Business cases

This Chapter briefly presents the main characteristics of the ENVISION project related to the data products and the relevant task descriptions, as well as the task connection with the relevant business cases.

2.1 Data products

The ENVISION data products are based on the identification of the state-of-the-art in the field of EO enabled agriculture monitoring through past experiences of ENVISION partners, the relevant literature, and past and current relevant projects. Additionally, are to be further defined after the customer requirements phase. The data products, mentioned in the Grant Agreement, include:

Table 1: ENVISION Data Products

ID	Related Task	Description
DP1	Task 3.3	Analytics on Vegetation and Soil Index Time-series: Analytics and change detection on time-series of vegetation and soil indices
DP2	Task 3.4	Cultivated crop type maps (with Fusion of Sentinel-1 and Sentinel-2 data time-series)
DP3	Task 3.5	Grassland mowing events detection (with Fusion of Sentinel-1 and -2 data time-series)
DP4	Task 3.6	Soil condition monitoring: Soil Organic Carbon: The estimation of the soil organic carbon is based on Sentinel -2 optical data feeding a machine learning algorithm, pre-trained using in-situ data of SOC (soil sampling campaign)
DP5	Task 3.7	Crop growth Monitoring and identification of organic farming practices: Crop growth monitoring will include the crop status, estimation of LAI (Leaf Area Index), above ground biomass and yield estimation of 5 crops (maize, soybean, alfalfa, wheat, sunflower)

2.2 Business cases

Table 2 presents a brief reference to the business cases that are targeted by ENVSION's products, containing the ID, the title, the products as well as the actors involved.

Table 2: Business cases

ID	Business Case	Interested Partner	Data products				
			Vegetation status	Crop type	Grassland mowing	Soil condition	Crop growth
BC1	Lithuania	NPA	T3.3	T3.4	T3.5		
BC2	Cyprus	CAPO	T3.3	T3.4			
BC3	Belgium	LV				T3.6	
BC4	Serbia	OCS					T3.7

The ENVISION project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 869366

3) User requirements - products

This Chapter lists the user requirements collected during WP2 (D2.2) and then focuses on matching them with the data products and the business cases.

The table below presents the user requirements (UR) that will be used for the relevant actor, as well as their description.

Table 3: User requirements gathered in WP2

ID	Actor	Description
UR1	NPA	As a Controller, I would like to receive data of crop type map every two weeks from middle of April to the middle of August (ideally middle of September).
UR2	NPA	As a Controller, I would like grassland mowing and grazing layers every two weeks from June till November with more than 85% accuracy
UR3	NPA	As a Controller, I would like to receive crop type and grassland mowing maps that are at least 95% accurate compared to in situ data
UR4	NPA	As a Controller, I would like to receive vegetation status maps with a priority on EFA catch-crop fields and all fallow land fields
UR5	NPA	As a Controller, I would like to be able to mask layers of interest with information from ENVISION outputs, for example to check parcels which intersect with soil erosion results, or to link crop type maps with grassland mowing layers.
UR6	OCS & CAPO	As an Organisation, we would like to be able to identify and distinguish between organic and conventional crop, and to monitor pesticide use on the declared plots because this is an important objective in many agri-environmental policies.
UR7	CAPO	As an Organisation, we need to receive information about the specific crop types even in very small parcels, or a coarser level of classification with a group of possible crop types.
UR8	OCS	As an Organisation, we want to get ENVISION outputs per parcel, especially for information on yield of each crop
UR9	OCS	As an Organisation, we want to get information once a year about the crops of neighbouring plots that are not involved in organic production (neighbouring to the plots that the organisation inspects)
UR10	OCS	As an Organisation, we would like to get data once a year for the crop types of conventional plots that belong to the same farmers that are involved also in organic production, even if the organisation's primary target is monitoring the farmer's organic crops
UR11	OCS	As an Organisation, we would like to track reductions in the number of plants through several times of the year, because this could be an indication of potential damages to crops that can result to events such as the re-cultivation of different crops on the same parcel, which is illegal
UR12	LV	As an Organisation, we want the system to provide us with errors against legislation that we can communicate to farmers.
UR13	NPA	As an Organisation, we want to have an idea of the accuracy of the output of a service through relevant indicators and sufficient documentation of the

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

		methodology, as well as to receive notifications when the accuracy degrades throughout the cultivation period.
UR14	LV	As an Administrator, I need to know when ENVISION services' outputs are not available so I can warn the respective farmer they need to provide it themselves.
UR15	LV	As an Inspector, I want the results from remote monitoring services to be reliable and verifiable on the spot.
UR16	LV & NPA	As an Organisation, we need to receive outputs both as maps/layers and relevant tables/numeric information, as well as to receive time series of indicators to study changes and emerging problems.
UR17	NPA & CAPO	As a Controller, I would like to receive data from the whole country (declared parcels) and not specific zones.
UR18	CAPO & LV & NPA	As an Administrator, I would like to receive ENVISION outputs from the time of submission and throughout the entire application period to help applicants and explain possible implications of wrong declarations / ineligibility of plots, considering the eligibility criteria / rules for multiple agri-environmental schemes.
UR19	LV	As an IT Expert, I want good quality to characterize the ENVISION platform services.
UR20	NPA & LV	As an Organisation, we want the output of services to be stable and the services set-up for long-term use.

The Table below provides a matrix linking each component with each one of the Business Cases (BC) and User Requirement (UR) presented above.

Table 4: The matrix of services, the business cases, and the user requirements

ID	User Requirements	Business Case	Data Product	Service	Responsible
1	UR1/UR3/UR5/UR7/U13/U17	NMA & CAPO	Cultivated Crop Type Maps	Multi-temporal Crop Type Maps	NOA
2	UR1/UR3/UR18	NMA & CAPO	Cultivated Crop Type Maps	Early Crop Classification and Alert System	NOA
3	UR1/UR3/UR13/UR18/UR19	NMA & CAPO	Cultivated Crop Type Maps	Smart Sampling for OTSC (Traffic Light System)	NOA
4	UR3/UR17/UR18	NMA & CAPO	Cultivated Crop Type Maps	Crop Diversification Compliance Map	NOA
5	UR2/UR3/UR5/UR13/U17	NMA	Grassland Mowing Events Detection	Grassland Mowing Events Monitoring	NOA
6	UR2/UR3/UR5/UR13/U17/UR18	NMA	Grassland Mowing Events Detection	Mowing Compliance Map	NOA

7	UR5/UR17	NMA & CAPO	Analytics on Vegetation and Soil Index Time-Series	Minimum Soil Cover for Soil Erosion	NOA
8	UR5/UR17	NMA & CAPO	Analytics on Vegetation and Soil Index Time-Series	Stubble Burning Identification	NOA
9	UR5/UR17	CAPO	Analytics on Vegetation and Soil Index Time-Series	Natura2000 regions activity hotspot detection (Illegal Land-Clearing)	NOA
10	UR5/UR17	NMA & CAPO	Analytics on Vegetation and Soil Index Time-Series	Runoff Risk assessment for the reduction of water pollution in Nitrate Vulnerable Areas	NOA
11	UR7/UR13/UR16/UR17/ UR18	NMA & CAPO	Analytics on Vegetation and Soil Index Time-Series	Retrieve Analytics and Aggregated Statistics on custom Areas using GIS querying functionalities over DataCube	NOA
12	UR19	LV	SOC Products	Soil Organic Carbon	ILVO
13	UR20	NPA & LV	SOC Products	Soil Organic Carbon	ILVO
14	UR6	OCS & CAPO	Distinction of organic vs conventional	Vegetation indices - phenology	AgroA pps
15	UR8	OCS	Crop monitoring	Vegetation indices	AgroA pps
16	UR9	OCS	Crop monitoring	Vegetation indices	AgroA pps
17	UR10	OCS	Crop monitoring	Vegetation indices	AgroA pps
18	UR11	OCS	Crop monitoring	Vegetation indices	AgroA pps

4) ENVISION data and tools

This Chapter presents the data products per business case, along with the methods implemented and data utilized in order to face them respectively. This is the first iteration of this deliverable and more data products are expected to be added and methods will be fine-tuned. The final version of the data products will be reported in D3.7 and the final version of the results and validation reports in D3.6, which are both due M34. This deliverable is also connected to i) D2.2., where the customer requirements inspire the methods of this section, ii) D1.3. that is the initial data management plan, mentioning the datasets used to implement the data products and iii) D3.2. that reports on the auxiliary data used, mostly received from the customers that were necessary to use as input, but also train and validate our algorithms.

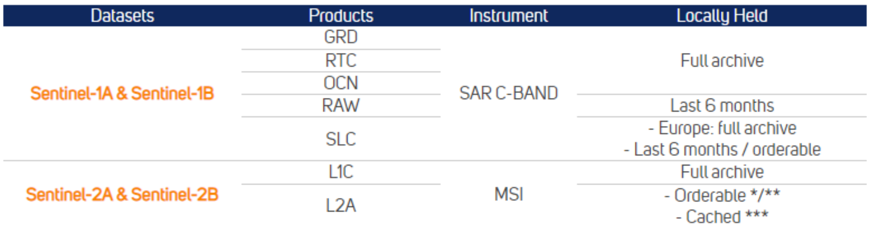
4.1 Analysis Ready Data and Technologies that enable data products

The introduction of the Sentinel missions, the newly available PlanetFusion data but also many more satellite missions that offer their data freely or at low cost, provide enormous amount of high temporal and spatial resolution. These data are of great significance for agriculture monitoring and have enabled numerous applications that were not possible just a few years ago. Nonetheless, there is still a challenge to overcome and that is to unlock the full information potential of this data. The volume of big Earth data, the different modalities of the different sources and sensors, but also the access and pre-processing steps are not easy to handle, particularly for non EO experts. Thus, there is the need for unlocking its power of information via a series of acquisition, indexing and pre-processing steps to ultimately end up with Analysis Ready Data (ARD) that can be used by the rest of the AI4EO community but even non-ICT experts. In order to be characterized as ARD, data has to be processed to a minimum degree fulfilling requirements such as atmospheric correction, geometric calibration, re-projection and resampling. In addition, ARD should be organized in such way to enable an effortless and direct analysis. Moreover, the pre-processing can be enhanced by additional steps such as cloud masking so to simplify and support data analysis. In this direction, NOA is developing automated workflows for generation of ARD, which will be the input of the later stage of data analysis and AI pipelines.

4.1.1 Sentinel Data

The back-end processes are hosted on CREODIAS, data can be accessed directly via the offered S3 object storage directory (/eodata). This kind of data storage ensures a high level of performance. Users have the potential to access the full archive of Sentinel-1 GRD, SLC and Sentinel-2 Level-1C (L1C) for Europe, whereas Sentinel Level-2A (L2A) products are not fully offered (Figure 1). The latter are provided for Lithuania, but not for Cyprus in specific years, therefore Sen2Cor software is used to transform L1C to L2A. Providing such a directory, developers can retrieve and process the data straightforwardly without the need of downloading the data locally or copy it to dedicated VMs. At the same time, NOA has already developed Python scripts for searching and pre-processing products in eodata directory based on several parameters.

The ENVISION project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 869366



Datasets	Products	Instrument	Locally Held
Sentinel-1A & Sentinel-1B	GRD	SAR C-BAND	Full archive
	RTC		
	OCN		
	RAW		Last 6 months
	SLC		- Europe: full archive - Last 6 months / orderable
Sentinel-2A & Sentinel-2B	L1C	MSI	Full archive
	L2A		- Orderable */** - Cached ***

Figure 1: Datasets and products provided by CreoDIAS along with their archive policy

4.1.2 Data Pre-processing

As mentioned above, data pre-processing is an essential step for generating ARD. NOA has developed a series of automated procedures that are executed in the powerful infrastructure of CreoDIAS. These procedures include Sentinel-1 backscatter and coherence generation, Sentinel-2 pre-processing, cloud masking and parcel buffering analysis. In this deliverable we present the latter two, as the sentinel pre-processing have been analysed in D4.1.

Cloud Masking

The existence of cloud and cloud shadows may affect the capability of data analysis as it reduces significantly the detection and monitoring information of surface features captured by the satellites' sensors. Thus, cloud Masking is an essential procedure aiming at detecting clouds along with their shadows. The issue has to be addressed before any remote sensing analysis takes place. Currently, there are a series of tools used for classifying pixels as clouds, from which Sen2cor has been selected. Sentinel 2 Correction is a single-date processor designed for land cover classification and atmospheric correction of top-of-atmosphere Level 1C input data. It creates a scene classification product, which uses a series of spectral reflectance thresholds, ratios and indices (e.g. NDWI, NDVI) to compute cloud probabilities for each pixel. Thresholding is performed on all bands except the water vapor band (Band 9) and two of the three vegetation red edge bands (Bands 6 and 7). Sen2Cor finds less valid pixels due to its class definition of dark area and also it performs poorly in identifying observations affected by clouds and shadows. Nonetheless, it has a very high overall accuracy. ENVISION has also selected the sen2cor solution as CreoDIAS already offers the L2A products via the eodata directory, except from a small number of cases (for example Cyprus for the year 2018). To address the disadvantages of sen2cor, especially the weakness of misclassifying some of the cloudy and shadow pixels, we applied a buffer zone around the cloud objects. As a result, the pixels adjacent to clouds are now classified as cloudy, providing a trade-off between better cloud masking and fewer clear pixel for analysis.

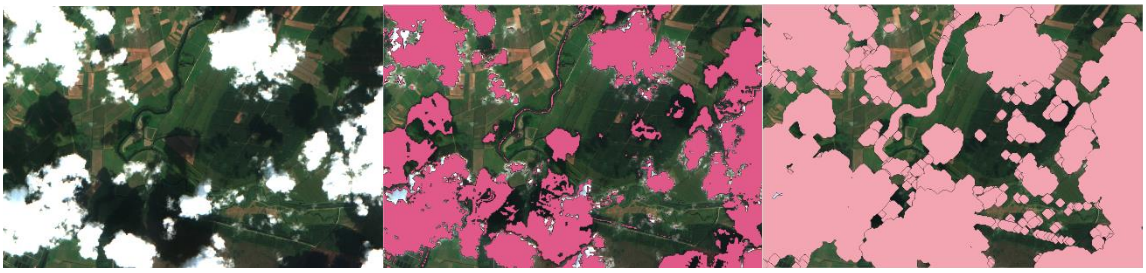


Figure 2: Cloud Masking Buffering

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Buffering Parcels

LPIS include information about parcels where geometry is one the main attributes. This geometry (polygon or multipolygon) is often mixed with pixels of another object, creating mixels (pixels that belong to more than one fields). As the analysis takes into consideration either each pixel or an aggregation of them, it is important to avoid these outliers. Thus, the boundaries of the geometries are reduced through a 5m buffer zone using GDAL. The buffer corresponds to a polygon containing the region within the buffer distance of the original geometry. Afterwards, the polygons are rasterized in order to avoid conflict cases of mixels and acquire samples that are more representative. Below there is an example visualization where with green we depict the area of the field after the buffering routine

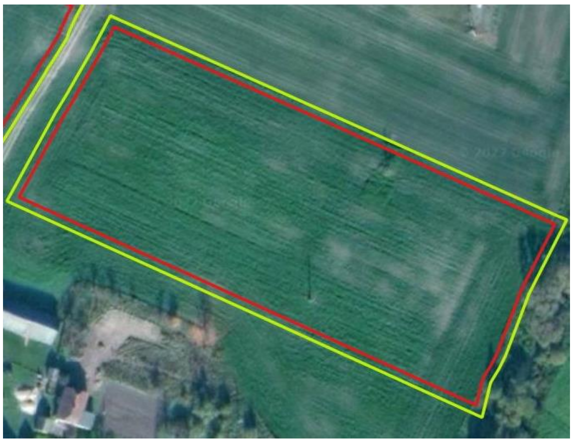


Figure 3: Parcels geometry after buffering

4.1.3 Open DataCube (ODC)

One of ENVISION's goals is the monitoring of agriculture at national scale providing long time-series indices and opening the door for deriving trends and analytics to support decision making process. To this direction, the harnessing of a great amount of EO data requires powerful resources along with dedicated tools and frameworks. Since a series of technologies are available, we chose a mature and open-source framework that has been applied in many use cases. The Open Data Cube (ODC), an extension of the Australian Geoscience Data Cube (AGDC), is an analytical framework offering a series of data structures and tools to organize and analyse effortless EO data. It is available as a free and open-source solution, under Apache 2.0 license and as a suite of applications. Currently, more than 30 countries support and use ODC. DataCube has already been installed, tested and used locally in the DataCAP [1]application developed by NOA (http://62.217.82.91). The latter is a series of automated modules that download, pre-process and index data in the DataCube, which is used as the core component of the application. At the same time DataCAP exploits street level images harvested from mapillary API so to be used during the validation process of any analysis. Thus, DataCAP considered to be a generic tool that leverage big satellite data and crowdsourced street- level images for monitoring CAP. Having the experience of implementing such as application, we can definitely say that one of the main advantages of the ODC framework is the capability of cataloguing massive EO data sets. After this cataloguing, access and manipulation of the data can take place easily via a Python API.

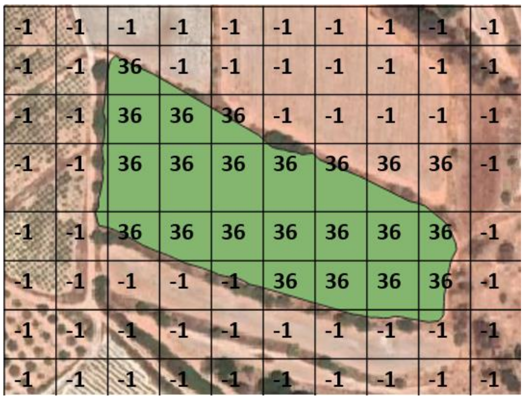
[1] https://github.com/Agri-Hub/datacap, https://github.com/Agri-Hub/Mapillary_Annotation

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

For the cataloguing of data, ODC offers two methods; Indexing (catalogue only metadata) and Ingesting (catalogue the entire data). Currently, indexing seems to be the most efficient method according to the developers of ODC and it is based on the DataCube-core module, which utilizes a PostgreSQL database to write the metadata of the products. The latter are hosted either in a local file system or in the cloud. The implementation of the **ENVISION DataCube** includes the installation of the required environment along with the configuration of several files and the initialization of the database.

ENVISION exploits both eodata catalogue mentioned before and the ODC framework to provide **innovative** and **scalable** solutions at national scale. Currently, we focus on Cyprus and Lithuania, but this can be effortlessly applied to any other area. Furthermore, it offers **both Sentinel-1 and Sentinel-2 pre-processed data** for a complete agriculture monitoring, leaving also room for research and operational outputs such as data fusion, crop classification and analytics. At the time of writing, **Sentinel-1 and Sentinel-2 processed data for 2019 and 2020 related to the entire extent of Lithuania and Cyprus have been indexed.** To make this happen, all the observed products for these two countries are **processed directly from this catalogue** and the generated ARD are automatically indexed to the database, hosted in the dedicated CreoDIAS VM. Before the data indexing, users have to create the products related to each indexed dataset. Products are considered as collection of datasets, which share the same sets of measurements.

In ENVISION we have created **products per country** (e.g., S2PreprocessedLithuania) via yaml files. It becomes apparent, that the previous procedure is **area-based,** which enables the scalability of the system to any other area and year in order to monitor agriculture by creating dedicated ODC products. After the indexing of products, DataCube hosts time series data for the two countries with the same spatial resolution of 10m. In order to request and exploit this data, we can use a Python API. The latter allows the load of metadata or/and data based on parameters such as the product, the time range, the bounding box, the bands etc. The loaded data are stored into X**arrays** of three dimensions (time, latitude, longitude). Nevertheless, we have to deal with the time complexity of a national scale analysis. The processing of millions of parcels is an important factor to be considered in terms of execution time. In that direction, we transformed the vectorized declarations of the farmers to raster format in order for every pixel included in a parcel to carry the id of it. These rasters are indexed into the DataCube and loaded afterwards in the same resolution and extent. The existence of this information stored into the DataCube **allows the optimization of the zonal statistics** calculation for each parcel by using the **group by** function. This function groups the Xarray bands based on the ids of the parcels enabling the aggregation based on the geometry. Figure 4 depicts the stack of two layers: an RGB raster and the ids raster.



-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	36	-1	-1	-1	-1	-1	-1	-1
-1	-1	36	36	36	-1	-1	-1	-1	-1
-1	-1	36	36	36	36	36	36	36	-1
-1	-1	36	36	36	36	36	36	36	-1
-1	-1	-1	-1	-1	36	36	36	36	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Figure 4: Parcel's index rasterization inside DataCube

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

4.1.4 Geospatial Database

The fundamental basis for the generation of the aforementioned EO Big Data Analytics are both the installed Open Data Cube as described in 4.1.3 and a geospatial database. ENVISION utilizes scripts to collect data from farmers' declarations via an API, which is stored to a PostgreSQL / PostGIS database, along with the corresponding satellite metadata. This data enables and populates all the back-end pipelines of T3.3, T3.4 and T3.5. The results of each of these tasks are used to update the database enabling a **back-and-forth communication between ODC and database** as it is shown in Figure 5. Thus, we have datacubes that include and keep on being dynamically populated by Sentinel-1 and Sentinel-2 products, but also auxiliary geospatial data that are used to enable the provision of the data products that in turn populate the cubes as well. This way, we end up with **member state specific knowledge bases** for CAP monitoring. Meanwhile, useful operations such as the computation of distance between two geometries, the calculation of area and buffer analysis, and geospatial queries, can take place exploiting the power of the POSTGIS extension.

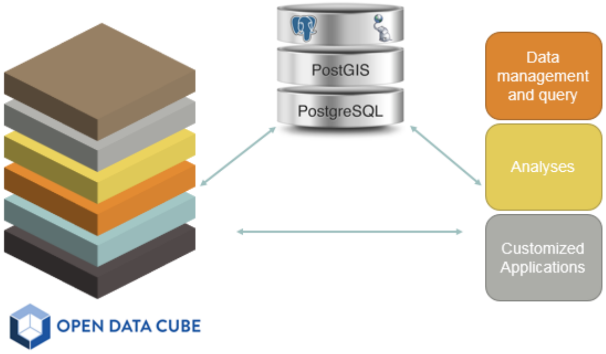


Figure 5: ENVISION products connection scheme

4.2 BC1: Monitoring multiple environmental and climate requirements of CAP – Lithuania

The National Paying Agency (NPA) supports measures for agriculture, rural development and fisheries and is responsible for all Integrated Administration and Control System (IACS) controls in Lithuania. It is an organization that goes in parallel with the current technological flow and it has developed all necessary elements of software and registers (100% geo-spatial applications, e-documents forms, statistical reports etc.) including robotic software for the evaluation of applications. While providing the consultancy services to its clients, NPA has established strong communication links with all the relevant stakeholders, including the farmer unions, municipalities and advisory bodies.

NPA has participated in multiple EU projects (e.g. SEN4CAP, RECAP, NIVA, EO4AGRI) regarding the monitoring of CAP requirements, modernization of IACS and in general the improvement of operational agriculture techniques. Lithuanian customers are continuously showing an active interest in employing EO technologies for monitoring farmers' performance (e.g., participation in RECAP project), as well as other technologies apart from on-farm checks; for example, farmers have the option to provide evidence regarding their activities by using NPA's mobile app, i.e., geotagged photos

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

with captured coordinates, direction, azimuth value, and date stamp. Through this app, 1. farmers can inform NPA about performed activities (e.g., grass mowing, grass removal, catch crop seeding, green fallow ploughing), and 2. all users, including citizens, can inform NPA about bad farming practices (e.g., grassland areas that are not mowed). NPA officers evaluate the data and decide if they will perform on-farm-checks.

Consequently, the development of a “smart” and self-operative service for the entire agricultural land of Lithuania will be of high supportive value for the national paying agency in order to incorporate all the aforementioned into an “all-in-one" integrated platform towards the supervision of CAP measures. In this section will be presented the general picture of the Lithuanian business case and a brief descriptive Statistical Analysis in order to help us understand and answer the pilot’s requirements (see also D.2.2), highlight the limitations and form the methodologies applied with a view to approach and sufficiently answer to the respective problems.

4.2.1 Data products description

Study Site

Lithuania is a North-West European country located in the Baltic region with a total area of 65,300 km². It has a temperate climate with both maritime and continental influences. It is defined as humid continental under the Köppen climate classification scheme [1], but it can be considered that is close to oceanic for a narrow coastal zone. The climate of Lithuania is described as averagely cold with inclement and snowy winters [2]. Based on these climatic peculiarities the climatic regions can be distinguished are the Coastal, the Samogitian, the Middle Lowlands and South-eastern Highlands that can be divided to smaller sub-regions (Figure 6) [3].

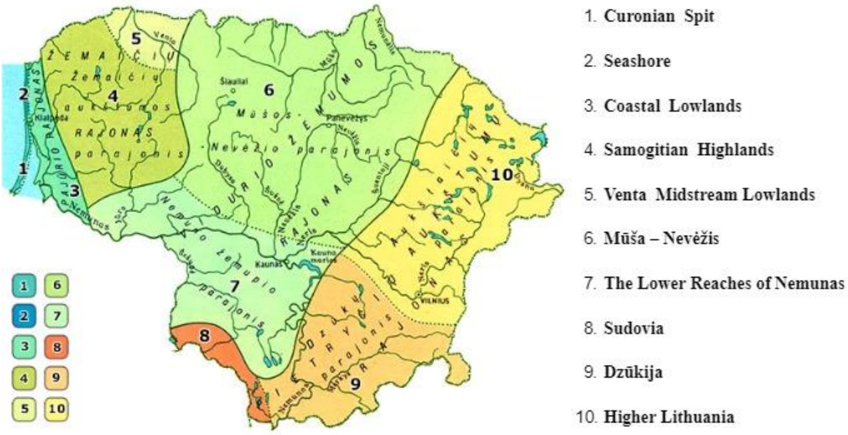


Figure 6: The map of climatic regioning in Lithuania (Climatic regioning 2013) adapted from [3]

All these apparently play a significant role in the profile and the development of agricultural land across it’s the territory, since the local environmental factors can affect directly the growth of the different crops and the application of various rural practices accordingly. With a quick view (Figure 7), we can see an illustration of the different crops across the country based on farmers’ declarations of 2020.

The ENVISION project has received funding from the European Union’s Horizon 2020 research
and innovation programme under grant agreement No 869366

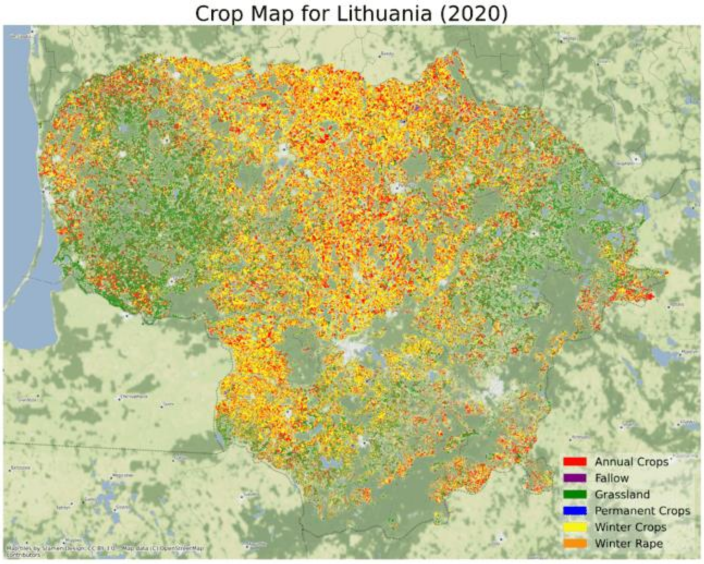


Figure 7: Crop Map of Lithuania (2020)

Taking into account the respective climatic zones, we can observe the different distribution of cultivated crops across Lithuania. Grasslands and Cereals (winter and spring) are spread around the country equally, with grasslands making their appearance more drastic in the western and eastern part of the country, while other cultivations in the central zone. Thus, towards the monitoring of the entire country (UR17), it is essential to divide it into sub-areas and study each one of them individually.

Another thing that should be evaluated is the large variety of crop types and their numerical distribution. As we can see in more detail from the Figure 8, the vast majority of crops (~85%) can be allocated in 5 characteristic agricultural categories (grasses, winter cereals, spring cereals, winter rape and potatoes). However, it is a requirement for the Lithuania business case to include multiple crop types and not just then 5-10 most dominant ones, as it is usually the case.

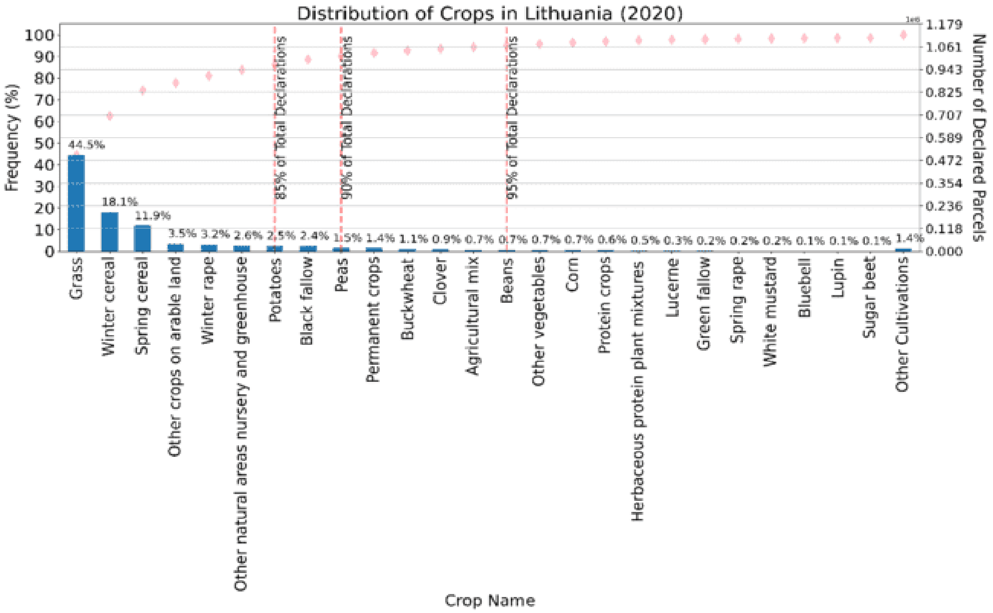


Figure 8: Lithuania Crop Type Distribution (2020)

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Additionally, for the Lithuanian business case, we use the crop taxonomy as it has been formulated in the SEN4CAP[2] project (Figure 9). It is worth mentioning that crop classification in a higher taxonomy is useful and sufficient in order to define most of the wrong declarations.

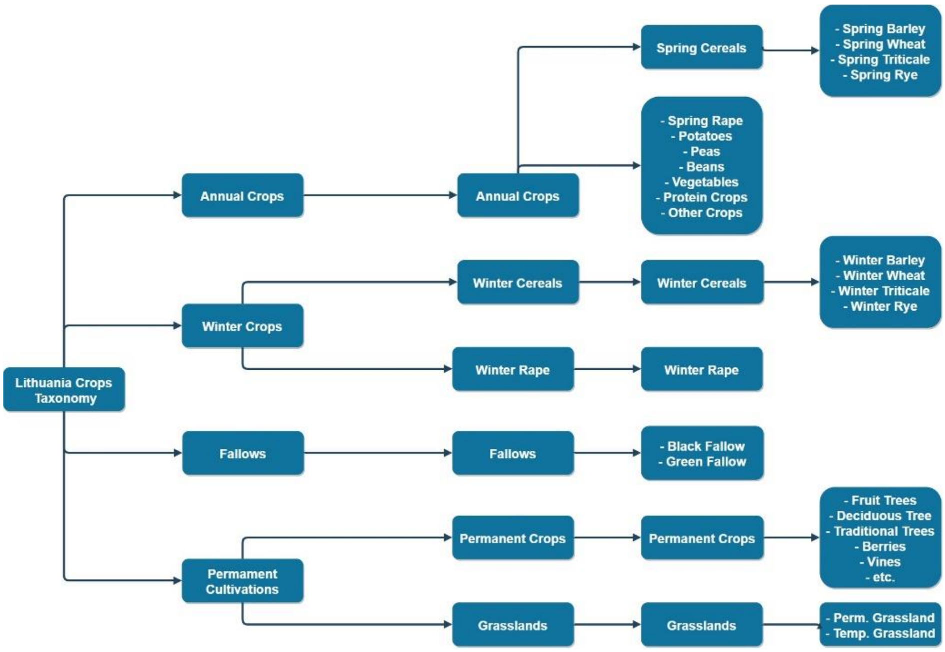


Figure 9: Crop Taxonomy for Lithuania

Cloud Coverage

A major issue for the Lithuania BC is that is suffers from very frequent and extended cloud coverage which can result to sparse Sentinel-2 image time series. Specifically, more than 25% of all the available Sentinel-2 products had at least 90% cloud coverage, for both 2020 and 2021. In some cases, the gap between two clear Sentinel-2 acquisitions can exceed a period of an entire month (see Figure 10). Clouds and their shadows may produce misleading results in analyses, import noise and may have dramatic consequences into the precision of agricultural monitoring. Tasks that are related with very drastic practices, such as the detection of mowing events (UR2), are sensitive to cloud appearance since we have to observe swift changes in the land greenery inside a very brief period of time.

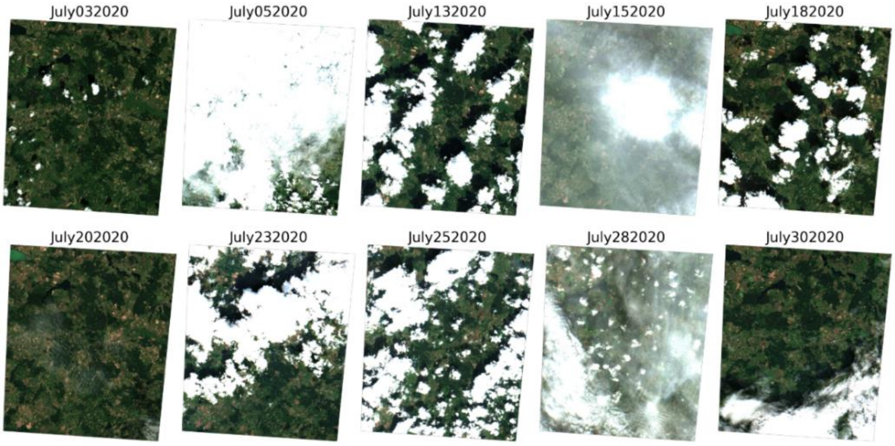


Figure 10: Frequent Cloud Coverage of a sample area in Lithuania

[2] http://esa-sen4cap.org

4.2.1.1 Analytics on Vegetation and Soil Index Time-series

The massive production of Earth observation data has as a result the need for an efficient, stable and smart solution for managing and analysing it. EO Big Data technologies are evolving rapidly making room for progress in big data analytics. Cloud storage and computing key-players such as CreoDIAS provide one stop shop; direct access to high amount of satellite data along with powerful infrastructure and tools so to enable users to working directly on cloud. ENVISION aims at developing advanced analytical techniques in order to extract knowledge from high-resolution images.

Specifically, analytics on vegetation and soil index time-series is one of the microservices described in the GA and D4.1. The processes related to this product aim at providing analytics on several EO indicators to automatically detect trends and flag risk areas and conditions. Thus, the time that researchers and decision-makers spend on analysing and identifying such cases will be decreased. In that direction, long time series of multiple vegetation indices have been created to address the needs of monitoring the agriculture land of Lithuania and Cyprus at national scale. Beyond the vegetation indices, analytics can serve additional information such as statistics (number of declarations per crop type, number of non-compliant parcels per declared crop type, mowing events etc.) in a user-defined area of interest. Additionally, this task deals with soil monitoring. Soil is considered an essential agricultural resource, which provides a strong basis for food production along with additional resources for a circular bio-economy. In addition, it has an impact to carbon sequestration and storage. It becomes apparent that there is a close link between agriculture and soil health. However, this link faces a number of challenges, with soil erosion to be one of them. The CAP aims at addressing these challenges by proposing and enabling a sustainable soil management strategy via a variety of cross-compliance rules, statutory management requirements (SMRs) and good agricultural and environmental conditions (GAECs).

Minimum Soil Cover

Ensuring a minimum soil cover over parcels is one of cross-compliance rules relevant to soil. In particular, the rules of GAEC 4 aim at the protection of soil against erosion after harvest until the end of winter. In Lithuania, the rules focus on the need of growing agricultural crops or keep black fallow on arable land. Black fallow (excluding black fallow in ecological field protection zones) must be sown or planted with agricultural crops before 15th November each year.

Nitrate Vulnerable Zones (NVZ)

In order to answer the statutory management requirement and SMR 1, a runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas has been developed, taking into account the proximity to the closest surface waters. The aim of the rule is to protect water against the runoff of nitrate polluted soil and water that could possibly reach fresh surface waters nearby. The requirements restrict storage, application of fertiliser and pesticides and cultivations along watercourses. In this direction, the usage of nitrogen fertilizer is accepted in parcels located more than 10m away from streams, 50m away from rivers and lakes and finally 300m away from any source used for water supply.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Stubble Burning Identification (SBI)

Remote sensing has been extensively used for the robust, accurate and timely assessment of forest wildfire damages, focusing on their extent, severity and other crucial indicators. The mapping of burnt areas has proved to be of high importance for environmental agencies which are responsible for coping with the aftermath of any relevant crisis.

This type of information is also significant in the monitoring of the agricultural sector. More specifically, Stubbles Burning Identification (SBI) is a requirement described by policies and practices through GAEC 6 in the CAP. Users, like CAPO and NPA in ENVISION's case, have expressed great interest in this stubble burning monitoring requirement. Through the identification of burnt crop parcels and the date they were burnt, paying agencies can monitor GAEC 6 compliance for each parcel.

Stubbles burn between March and May considered to be illicit activity in Lithuania based on NPA's regulations. In Lithuania, stubble burning is not a common practice and as NPA mentions farmers that follow this practice almost never declare this action. For the Lithuanian paying agency, it is important to have a clear overview of the situation of stubble burning and in that direction this task works. The output is a map (shapefile) containing parcels found to be burned, the date of the event and the compliance based on NPA's regulations.

Harvest event detection

The agricultural monitoring practices focus on Ecological Focus Area (EFA) practices and important part of this is the harvest event detection. It is conducted for arable land parcels and the parcels with declared EFA practice according to Land Parcel Identification System (LPIS). The purpose of this task is to locate the parcels that, at some point of the required period got harvested and the time the event took place, in order to conclude in a compliance level based on the rules of the Lithuanian paying agency.

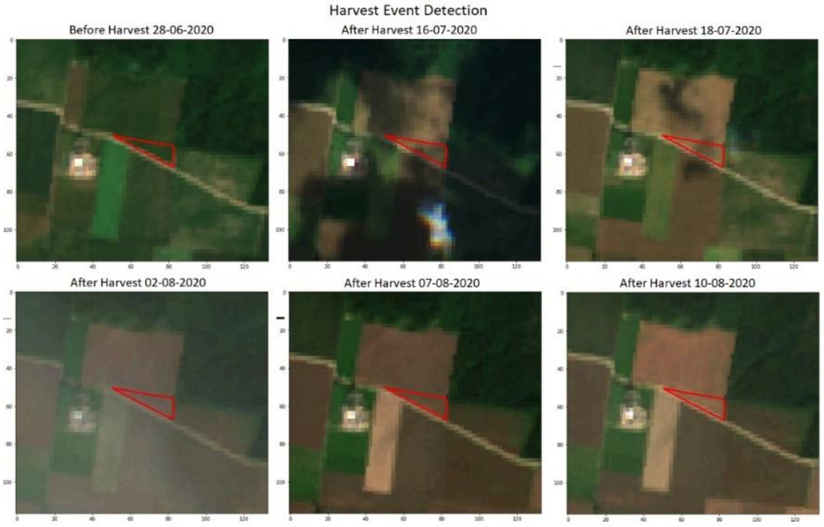


Figure 11: Sample area before and after harvesting

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

4.2.1.2 Cultivated Crop Type Maps (CCTM)

The **Cultivated Crop Type Maps (CCTM)** is an EO based crop classification product that exploits satellite data along with the usage of Machine Learning (ML) techniques in order to provide services related to the validation of the declared crop type by farmers (UR1). This can be used from the Paying Agencies as a tool to **enhance the process of inspecting the farmers' declarations** both during the declaration process and during the validation process (OTSCs) using a smart sampling algorithm. It is of great importance to export results for the entire territory (UR17), as well as for multiple instances throughout the cultivation period (UR1). Overall, the outcomes of this task will be used as a backbone for addressing other **Cross-Compliance and Greening requirements**.

The Lithuanian cultivation period starts from January, but essentially, the sowing of the majority of the crops starts after March and the harvest takes place around July-August. NPA performs their OTSC campaigns from late August till the end of September.

Figure 12 presents the Normalized Difference Vegetation Index (NDVI) signatures for the various crop types. As we can see, there crops that share **very similar growth rates** during the cultivation period (and mostly during the first months), which renders the discrimination of crops in that level of taxonomy, **early in the year,** a **difficult problem**. On the other hand, we observe that **after July** the distinction among categories is much **easier**.

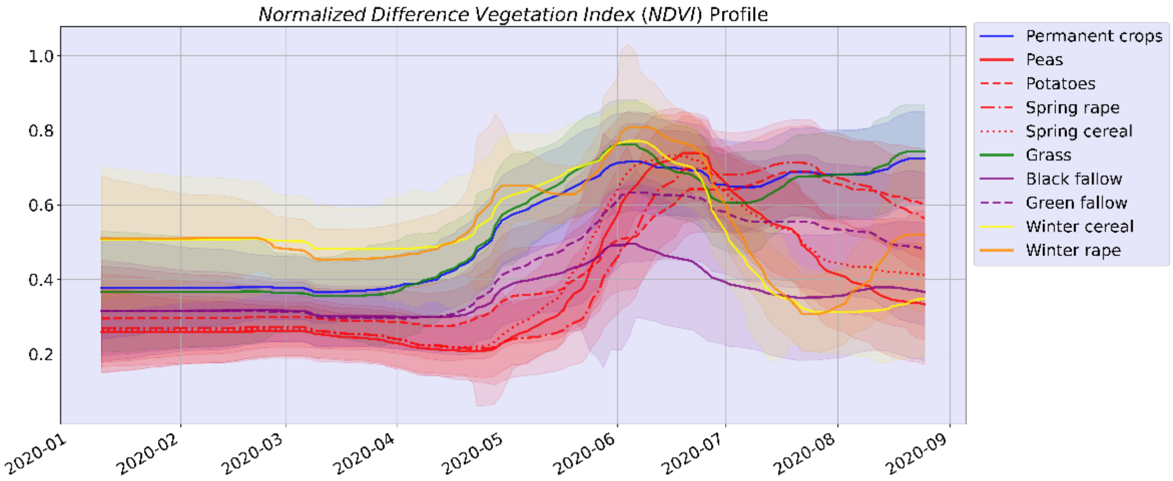


Figure 12: Crops NDVI signatures (Lithuania-2020)

For Lithuania, the proportion of false declared applications is relatively small, around 2-3% (D3.2). Performing some quick explanatory analysis using validation data of previous periods from the respective on-the-spot-checks or assessments through Very High Resolution (VHR) imagery provided by NPA, we can indicate the sources of these errors. In Figure 13 we present confusion matrices for the wrong declarations based on the evaluation of NPA, for the years of 2019 and 2020. Most of the errors in both matrices are related with mistakenly declared grasslands or vice-versa.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

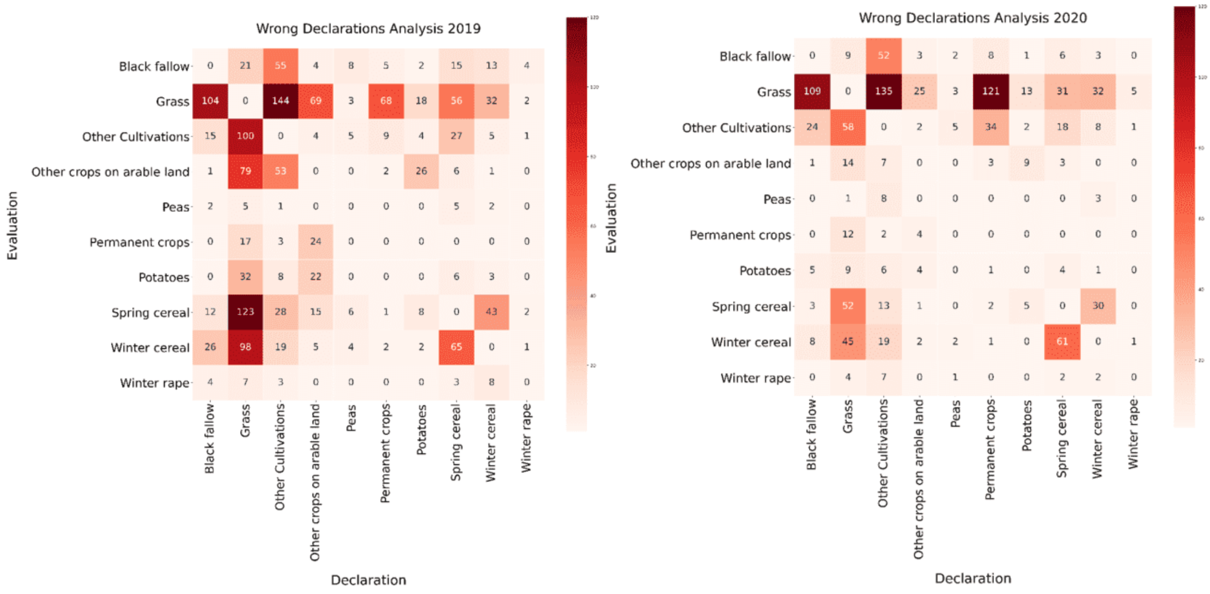


Figure 13: Wrong Declarations-Evaluations Confusion Matrix (Lithuania 2019 and 2020)

Greening Requirements and Crops Diversification (CD)

From 2013, CAP introduced a payment for a compulsory set of 'greening' measures, accounting for 30% of the direct payments budget. The aim of this action is to introduce a more effective strategy towards the delivering of its environmental and climate objectives and ensure the long-term sustainability of EU agriculture. More specifically, we are targeting to rules that focus on i) crop diversification and ii) the maintenance of permanent grasslands between successive cultivation periods. So far, we have focused on i), For the case of Lithuania, CD measures refers only to applicants whose **total declared arable land is larger than 10 hectares** and more specifically:

a. If the total expanse of the arable land declared from the applicant is **between 10 and 30 hectares** it should contain **at least 2 different categories of crops** inside this land and:
    - The major cultivation must cover **no more than the 75%** of this land

b. If the total expanse of the arable land declared from the applicant is **more than 30 hectares** it should contain **at least 3 different categories of crops** inside this land and:
    - The **two major** cultivations must cover together **no more than the 95%** of this land

c. If not less than the 75% of the total expanse of the arable land declared from the applicant is used as a combination of a **Temporal Grassland and/or Fallow Land**:
    - The major cultivation of the remaining arable land must cover no more than the 75% of this land

However, there are some exemptions that not Crop Diversification it is required:

Exemption A

The total declared arable land is less than 10 hectares (as already been mentioned)

Exemption B

If proportion of Arable Land (AL) larger than 75%:

- It is used as Temporal Grassland
- It is declared as Fallow Land
- Any combination of the above

- And the remaining Arable land is less than 30 hectares

Exemption C

If proportion of Eligible Agricultural Area (EAA) larger than 75%:

- It is declared as Permanent Grassland
- And the remaining Arable land is less than 30 hectares
- It is declared as Crop Under Water.
- Any combination of the above
- And the remaining Arable land is less than 30 hectares

Exemption D

If the entire amount of Total Arable land (100%) it is declared as Crop Under Water

4.2.1.3 Grasslands Mowing Events Detection

As it has already been mentioned, grasslands have special value for the Lithuanian business case. In general, the main function of grassland is to provide feed for livestock, but they also serve other functions like the provision of habitat for the regional fauna and flora, the filtering of sediment and pollutants before they reach the water network, the prevention of soil erosion, the storing of greenhouse gases, etc. [4]. In Lithuanian case, grasslands can be discerned into finest categories of permanents and temporal grasslands.

The motivation for developing this product lies on both **pillar 1** and **pillar 2** of the CAP. Regarding the first, we have the direct payments for grassland value maintenance that requires thorough knowledge of the activity of grasslands. For pillar 2, we have the motivation lying upon the conceptual design of targeted agro-ecological and climate focused measures. Moreover, the grassland mowing events detection service aims at compliance checks. The service will provide alerts at various time instances during the growing season and support compliance checks with regards to greening rules. Different compliance regulations and policies regarding time and number of mowing/grazing events (see Table 5) can take place during a cultivation period. Finally, yet importantly, the monitoring of grassland activity should be performed on fields that have already been certified that they are referring to actual grasslands (UR5).

Table 5: Lithuania National Mowing Regulations

National Mowing Regulations National regulations to assess the compliancy of grassland mowing. The regulations are specified for each crop, if different		
Crop type (code, name)	**Mandatory mowing period.**	**Additional rules in case of mowing event is outside the mandatory period (C or NC), depending on 2 cases:** **1) Compliant (C) or not compliant (NC) if a mowing event occurred in the mandatory period** **2) Compliant (C) or not compliant (NC) if a mowing occurred outside the mandatory period and no mowing occurred in the mandatory period**

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

GPŽ, Pasture or meadow, perennial grass up to 5 years	At least 1 mowing or grazing within 1st August	1) C 2) NC
DGP, Perennial pastures or meadows 5 years and more	At least 1 mowing or grazing within 1st August	1) C 2) NC
EPT, Extensive meadows grazing with livestock	At least 1 grazing between 1st May and 30th October	1) C 2) NC
NPT, Natural and semi-natural meadows	At least 1 mowing between 15th July and 30th September	1) C 2) NC
SPT, Specific meadows	At least 1 mowing between 15th July and 15th October	1) C 2) NC
5PT-2, Extensive management of wetlands	At least 1 grazing between 1st May and 30th October	1) C 2) NC
GPA, Pasture or meadow, perennial grass up to 5 years, renewed in the current year	At least 1 mowing within 1st August	1) C 2) NC
MNP, Aquatic warbler habitats storage in raw and semi-natural grasslands	At least 1 mowing between 1th July and 1st October	1) C 2) NC
MNŠ, Aquatic warbler habitats storage in wetlands	At least 1 mowing between 1th August and 1st October	1) C 2) NC

4.2.2 Methodology

4.2.2.1 Analytics on Vegetation and Soil Index Time-series

According to GA and D4.1 the outputs of task 3.3 will deliver a series of analytics consisting of:

- Runoff Risk assessment for the reduction of water pollution in Nitrate Vulnerable Areas.
- Buffer zones for the proximity to waterways nearby
- Minimum soil cover for Soil Erosion
- Stubble Burning Identification
- Other GIS querying functionalities such as buffer analysis, area calculation etc.

The product will provide multiple Analytics reports throughout the year taking advantage of both Sentinel-1 and Sentinel-2.

<u>Input data</u>

I. Satellite Data:
   a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC)
      i. Spectral bands (B01-B12)
      ii. Scene Classification (SCL)
   b. Sentinel-1 GRD (rel. orbits: 58, 131)
      i. Backscattering coefficients (VV-VH)
   c. Sentinel-1 Coherence (rel. orbits: 58, 131)
      i. Coherence coefficients (VV-VH)

II. Auxiliary Data:

a. Annual soil loss
b. Rainfall erosivity factor
c. Soil erodibility factor
d. Slope length factor and slope steepness factor
e. Crop and cover management factor
f. Conservation supporting practices factor
g. Slope

III. Paying agency's:
a. A lookup table for all the available crop type names, codes, families and CD ancillary info
b. Parcels geometries and initial declarations as a shapefile (updated when is necessary)
c. Agricultural Practices Descriptions
d. Hydrographic networks

Output data

The product will provide:

I. Pixel-based Analytics or Aggregated Statistics per parcel over a specific time range
II. Runoff Risk level map of the parcels.
III. Shapefile of CC indicators of compliance such as Traffic light system, GAEC 1 and GAEC 6
IV. Rasters of Stubble Burning Identification for the maintenance of organic matter in soil
V. Rasters of minimum soil cover for Soil Erosion

Retrieve Analytics and Aggregated Statistics on custom Areas using GIS querying on the top of the DataCube

The integration of the temporal dimension into remote sensing analytics gives the potential of creating long and dense time series. Hence, all the Sentinel-1 and Sentinel-2 products for the last two years and for the countries of interest, namely Lithuania and Cypurs, have been downloaded and pre-processed as described in D4.1. The pre-processed products come in GeoTIFF format and with spatial resolution of 10m. These new analysis-ready data are stored in a dedicated file system in CreoDIAS platform and are exploited by the **ENVSION DataCube** that has been also installed in the allocated CreoDIAS virtual machines. The DataCube stores metadata related to backscatter and coherence products for Sentinel-1, all the spectral bands for Sentinel-2 mission, LPIS for each country and additional environmental rasters such as the C-Factor. As a result, **effortless**, **rapid** and **national-scale** analyses can take place combined with calculation of an important number of vegetation indices exploiting the power of xarrays in terms of parallel processing in many dimensions. However, the following series of steps must take place before the analysis:

1. Load data from the DataCube based on four parameters; bounding box coordinates, time range, product and bands.
2. Keep only the cloud-free pixel for Sentinel-2 products. Cloud masks via SCL band have been applied so to eliminate noisy pixels and keep only the pixels considered as clear.

As the data has been loaded and cleared, if needed, an exploratory analysis of time series along with many others services can take place. It must be highlighted that the scripts give the potential for either a **pixel-based** analysis or a **parcel-based** aggregated one within a given time range. The first approach

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

reveals the behaviour of each pixel throughout the time, whereas the second one aggregated all the clear pixels inside a geometry revealing the parcel's trend. Below, there is a list with all the already implemented analytics:

1. **Animation of the temporal evolution of an area.** Raw band values or/and calculated on the fly indices for each time slice are used in order to create an animation using the animatplot. The purpose of it is to reveal clearly all the changes that take place throughout the selected time range.
2. **Temporal Statistics over an area.** In the same way with the previous service, all the required raw data or indices are calculated on a time window. The statistics of the selected bands can be provided either in the form of aggregated values for a parcel or as a plot for a larger area. A great advantage of the DataCube usage is that these statistics can be grouped in any time dimension such as day, month, season, year.
3. **Smoothed time-series.** This calculation is based on temporal statistics analysis, but it generated an aggregated value per time slice, which is presented via a smoothed line. The goal of this analytic is to identify and highlight any possible trend of the monitored area.
4. **Index Anomalies.** The understanding of how the longer-term changes affect the crops requires the calculation NDVI standardized anomaly as follows:

$$\frac{mean(short) - mean(long)}{std(long)}$$

   Mean(short) is the mean value of clear pixels included in a parcel on a short time range (e.g. during a month), whereas mean(long) and std(long) refer to mean and standard deviation values respectively again of clear pixels inside a parcel, but in a long time range (e.g. during a year).
5. **HeatMap of Clear Pixels.** It has already mentioned that cloud pixel considered as an issue especially for Lithuania. Therefore, it is important to generate a metric for calculating the frequency of cloud presence over each pixel in a time range. Having this information, we can exclude frequent cloudy pixels from our analysis or replace their values using interpolating methods.
6. **Spatial queries.** Except from the analytics generated from the Open Data Cube, ENVISION has also room for geospatial queries to the database as the following:
   - **a.** Show the number of declarations per crop type on a specific area
   - **b.** Show all the possible declarations that do not match with the predicted crop type
   - **c.** Show the ids of the parcels that do not comply with at least one GAEC
   - **d.** Show the ids of the parcels that have the maximum run off risk

Runoff Risk Assessment for the Reduction of Water Pollution in Nitrate Vulnerable Areas (GAEC 1/ SMR 1):

In order to answer the SMR1 requirement, a runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas has been developed, taking into account the proximity to the closest water areas. Therefore, distance from every point of parcel's geometry to the closest water surface is calculated. Parcels that are above a certain distance threshold are excluded. Afterwards, according to bibliography, several models have been developed to identify the probability or size of soil erosion. The Universal Soil Loss Equation (USLE) and its revised version Revised Universal Soil Loss Equation [5]

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

are the most widely used and accepted empirical soil erosion models. The later has the following equation as described in D4.1:

$$A=R\times K\times LS\times C\times P,$$

where:
A = annual soil loss (Mg·ha−1·year−1)
R = rainfall erosivity factor (MJ·mm·ha−1·h−1·year−1)
K = soil erodibility factor (Mg·h·MJ−1·mm−1)
LS = slope length factor and slope steepness factor (unitless)
C = crop and cover management factor (unitless)
P = conservation supporting practices factor (unitless)

All the involved parameters are downloaded from the ESDAC and resampled to the Sentinel 2 spatial resolution (10 m), except from the LS and C factor, which were calculated using LPIS and NDVI. Afterwards, these rasters have been indexed to the DataCube, **enhancing the number and diversity of multi-source products** stored. Having calculated RUSLE and the minimum distance from a water surface, every parcel is labelled with a risk category as the following table indicates:

Table 6: The rules for runoff risk assessment

		Water Proximity (meters)			
		<=10	**<=50**	**>50**	**>100**
RUSLE	**<=4**	High	Low	Low	Very Low
	>4 and <=8	High	Moderate	Low	Very Low
	>8 and <=15	High	High	Moderate	Very Low
	>15	Very High	Very High	Moderate	Very Low

Minimum Soil Cover for Soil Erosion

GAEC 4 demands the identification of soil coverage during specific months throughout the year. The identification of soil cover for a parcel takes into consideration multiple parameters. Initially, the average slope for each parcel has been calculated based on a 20m raster Digital Elevation Model. This slope refers to the full polygon, without using any buffer zone. In addition, a two-step pixel-based analysis takes places:

- Firstly, the pixels classified as vegetation, soil and water based on Sen2Cor are considered as “clear” pixels, whereas the other ones are set to null. If there is not even one clear pixel, the parcel gets also the value null.
- Then, the SAVI index is chosen as discussed in D4.1 and calculated for these clear pixels, along with the mean SAVI value of the parcel and the distance of each pixel to this aggregated value.

As the satellite observations do not always coincide with the fully grazed-out status of the parcel, there is the need to identify a minimum percentage of soil existence so to characterize it as bare ground. Currently, this percentage is 20% of the clear pixels. This means that if at least 20% of these pixels’ SAVI value is lower than a certain threshold, we flag this parcel as bare ground. In order to enhance the accuracy of the decision, we keep all the subsequent dates on which each parcel is considered as bare soil. Thus, the final decision for classifying a parcel as bare soil or not requires the parcel to be flagged as one at least two or more times, excluding the null values. Specifically for Lithuania, the

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

mapping of soil cover takes places during the period from July to October as parcels have to be checked for vegetation presence after the harvest of their main crop. The results of the SAVI are aggregated at the parcel level so as to keep the mean value of clear pixels' SAVI.

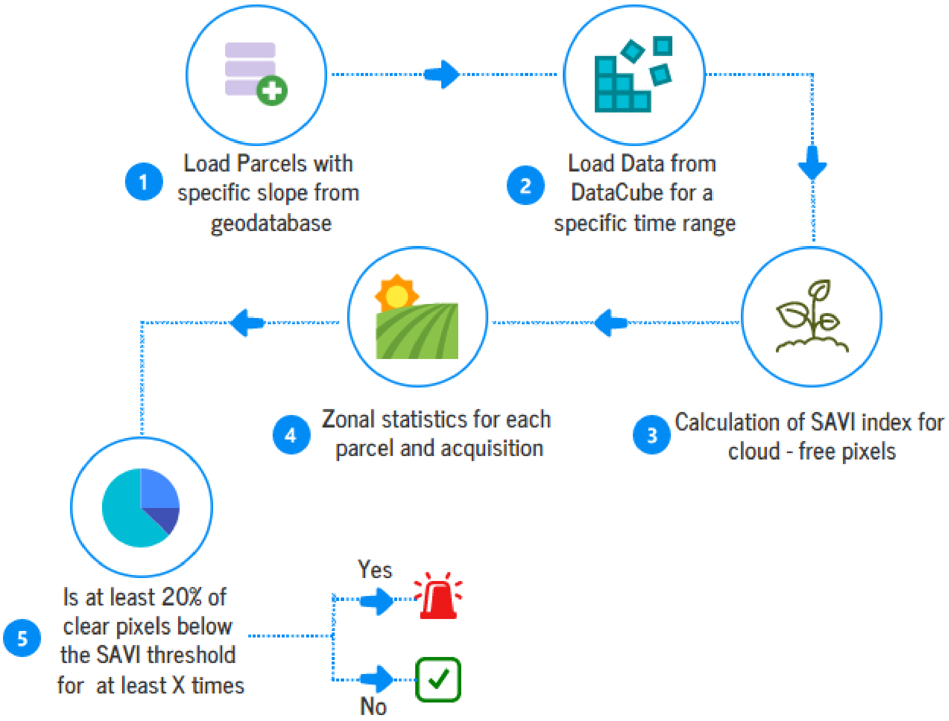

1 Load Parcels with specific slope from geodatabase
2 Load Data from DataCube for a specific time range
3 Calculation of SAVI index for cloud - free pixels
4 Zonal statistics for each parcel and acquisition
5 Is at least 20% of clear pixels below the SAVI threshold for at least X times
Yes
No

Figure 14: Workflow for the minimum soil cover detection

Stubble Burning Identification

The methodology of this approach focuses on data from satellite images and, specifically, Sentinel-2 data. The required Sentinel-2 products for the areas of Lithuania are pre-processed and indexed into the ENVISION DataCube. The algorithm focuses on detecting stubble burning in crops using Normalized Burn Ratio (NBR) and Difference Normalized Burn Ratio (dNBR) of Sentinel-2 satellite data. The approach starts by extracting NBR time series data for each parcel. This part of the methodology consists of calculating the average value of the pixels but only from the parcels in which at least 20% of the pixels were not covered by clouds. If parcels are covered from clouds by 80% or more, the algorithm is not applied. This product is about determining whether and when a parcel has been burnt, which is achieved using a rule-based algorithm. Specifically, a notable increase of dNBR (>150) and respectively very low NBR values (<-110) for at least 5 cloud-free acquisitions, indicate an event. This will be used as a baseline methodology, which will assist in photointerpretation to generate a validation dataset for stubble burnings.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

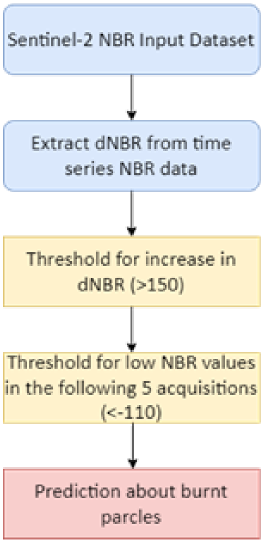

Sentinel-2 NBR Input Dataset
Extract dNBR from time series NBR data
Threshold for increase in dNBR (>150)
Threshold for low NBR values in the following 5 acquisitions (<-110)
Prediction about burnt parcles

Figure 15: Stubble Burning Identification in Lithuania product scheme

Harvest event detection

The harvest detection algorithm is set using NDVI from Sentinel-2 satellite images. The output produced from this algorithm is a map which contains the parcels where a harvest event detected and the date the event happened. This algorithm requires two conditions. Firstly, a land clearing event results in a remarkable drop in NDVI's value caused by loss of vegetation. Secondly, when a land clearing event happens, vegetation requires some time to grow again, so for the following period of time, NDVI is expected to be low. Therefore, this algorithm is based on thresholds and detects events when a remarkable decline in NDVI time-series values (>320) is detected, followed by followed by a mean value less than 350 for the next 30 days The algorithm produces harvest event predictions throughout the year. Finally, this data product refers to arable land, so, grasslands (45% of Lithuania's crops) were removed from the dataset in order to remove noise.

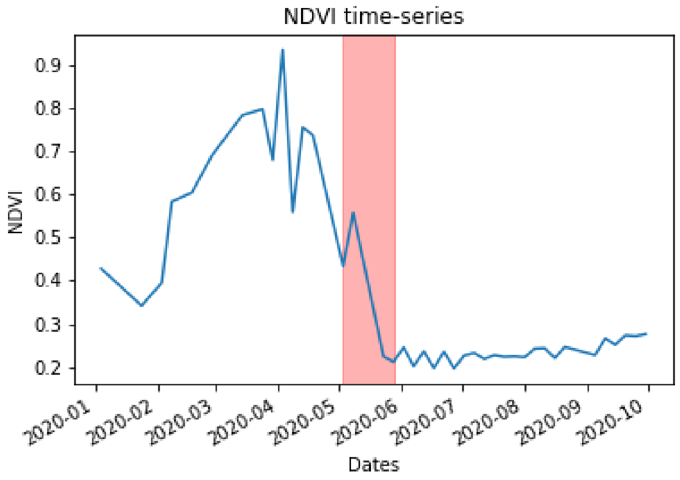


Figure 16 Harvest event detection, loss of vegetation represented in decline of NDVI

4.2.2.2 Cultivated Crop Type Maps (CCTM)

According to D4.1 Cultivated Crop Type Maps (CCTM) product will deliver EO derived outputs for cultivated crop type maps consisting of:

- Dynamic Crop type maps
- Alert mechanism using smart sampling
- Crop compliance with Greening-1 rule

For the business case of Lithuania, the product will provide dynamic crop type maps throughout the cultivation period, for every new or a group of new Sentinel acquisitions, to help in early alert during the application process, optimize the confidence of the classification process and work as the pillar for the monitoring of crop diversification requirements. To address the problem of cloud coverage and thus ensure higher accuracy, the system utilizes both Sentinel-1 and Sentinel-2 data. Data pre-processing follows the routines described in D4.1 and the product will be built on top of the DataCube. The provider of the product is NOA, whereas the results of the product are provided either via a RESTful API or as shapefiles. The smart sampling and compliance with Greening 1 services are based on these dynamic crop type maps, which are generated dynamically from early April until the end of the cultivation season.

<u>Input data</u>

More specifically we utilize the following data:

I. Satellite:
   a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC)
      i. Spectral bands (B01-B12)
      ii. Scene Classification (SCL)
      iii. Vegetation Indices - VIs (NDVI, NDWI, PSRI)
   b. Sentinel-1 GRD (rel. orbits: 58, 131)
      i. Backscattering coefficients (VV-VH)
II. Paying agency's:
   a. A lookup table for all the available crop type names, codes, families and CD ancillary info (see Table 7)
   b. Crop declarations as a shapefile (updated when is necessary)

For the Sentinel-2 bands and indices, we use the sen2cor SCL product as described in section 4.1.2 to mask out all cloudy pixels, i.e. replace them with null values. Consequently, using a linear interpolation method we fill these gaps. However, in order to cover the whole country, we use data from several Sentinel-2 tiles, which can be translated to different acquisition dates for parcels located in separate tiles. Therefore, instead of filling in the null values, we use linear interpolation to generate values that represent the 10th, 20th, 30th, and so forth, day of the year.

Moreover, as it has been already highlighted, since Lithuania suffers from **extensive and frequent cloud coverage**, apart from Sentinel-2 data we also utilize **Sentinel-1** data.Moreover, in D3.2 can be found an extensive description of the auxiliary data for BC1 and T3.4.

The ENVISION project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 869366

Table 7: Lithuania Look-up Table sample schema

Crop Code	L0	L1num	L1	L2num	L2	L3num	L3	CDnum	CD	EAA	AL	Grass	EFA	Cwater
BUL	1	1	Arable Land	109	Potatoes	51	Potatoes	109	Potatoes	1	1	0	0	0
GPA	1	3	Grass	3000	Grass	171	Pastures or meadow	3100	Grass_temp	1	1	1	1	0
DGP	1	3	Grass	3000	Grass	130	Perennial Pastures	3200	Grass_perm	1	0	1	0	0
EPT	1	3	Grass	3000	Grass	131	Extensive meadows	3200	Grass_perm	1	0	1	0	0
PDJ	1	4	Fallow	18	Black Fallow	72	Black Fallow	4000	Fallow	1	1	0	1	0

Output data

The product will provide:

i. Dynamic crop type maps as a shape file over the registered parcels for every new or group of new Sentinel acquisitions.

ii. Traffic light maps as a shapefile over the registered parcels for smart sampling of on-the-spot inspections and early alert of the users.

iii. Greening-1 compliance map as a shape file over the registered parcels at the end of the cultivation period.

Dynamic Crop Type Classification

The target of this product is to produce a crop type map, by classifying the set of polygons given by the PA. Specifically, we produce **multiple crop type maps**, for every group of new sentinel acquisition, which are acquired approximately every 15 days. For that reason, training of supervised machine learning (ML) algorithms is essential; and to do so we make a fundamental assumption that the vast majority of the farmers' declarations correspond to the reality, even though this is not the case for all of them. In this case, we have used the Random Forest [6] algorithm since it has been widely used in operational scenarios.

Since in Lithuania the average fields size is considerably large, results of CCTM will be evaluated directly at the **parcel level**, taking as inputs the statistical mean, median and std of all the pixels encompassed inside each parcel's buffered geometries. Given that, the training of the RF considers only cases that contain more than 10 clear pixels per parcel, in order to acquire representative information from each observation. The rest of the cases will be utilized only for inference. Moreover, we have excluded a test dataset, based on the OTSCs performed by the NPA, which is around 5% of the total parcels. Consequently, the dataset is divided into training and validation sets (30% and 70% respectively), in a stratified fashion, which means that we take into account the crop types distribution. Furthermore, in order to deal with the **high imbalance** of the dataset, we select only crop types that appear in more than 100 samples, and from those, the ones with less than 1000 samples are resampled using the Synthetic Minority Oversampling Technique (SMOTE) [7]. Finally, we excluded also some classes that presented very poor performance on the evaluation metrics, such as other crops on arable land, agricultural mix, protein crops and other vegetables. Lastly, the classification weights parameter, which is inversely proportional to the crop frequencies, is imported into the algorithm in order to deal with class imbalance.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

It is worth mentioning that the training dataset contains samples from areas across the entire Lithuania territory in order to exploit efficiently the various local peculiarities.

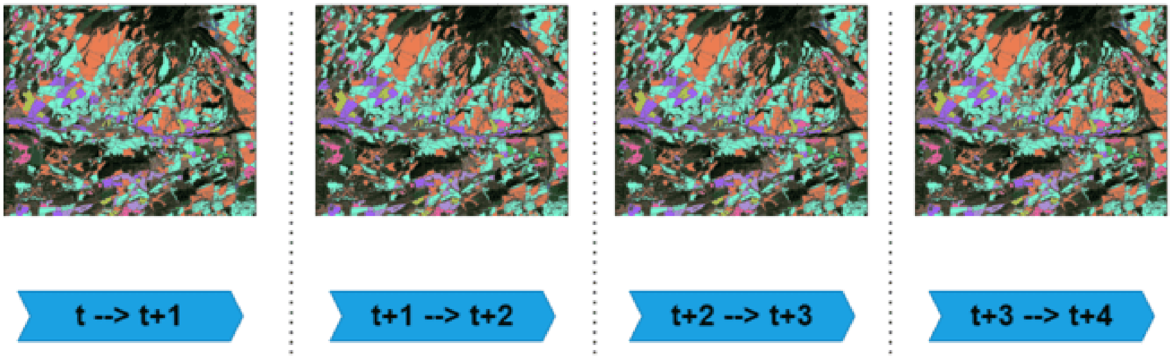


Figure 17: Multi-temporal Crop Type Mapping

Smart Sampling for OTSC (traffic light system)

Smart sampling is a sophisticated algorithm developed by NOA [8] that is evolving dynamically throughout the cultivation period by taking advantage of both the current and the previously generated Crop Type Maps {t, t-1, t-2, t-3, …}, in order to identify the most confident misclassifications. In essence, by taking into account the confidence level, which is based on the logit probabilities, as well as the logits' difference between the two most probable categories indicated by the ML classifier, we can identify misclassifications that we except to be false declarations rather than false predictions. The basis of the methodology is what we call **traffic light system**. This system enables the categorization of parcels into four classes, based on the RF algorithm ranking scores. Specifically, parcels are labelled as green, yellow, red and unreliable by taking into consideration the calculated difference of the two highest scores. For example, **green class** includes all the parcels that have high difference between the two scores and represent the class with the **highest confidence**. We focus on the later class by recording the parcels that have been systematically mislabelled during the multiple executions of crop classification algorithm throughout the cultivating season. This is achieved through the use of a user-defined threshold, which is dynamically increasing over time, and the total number of misclassifications of a given parcel. If the number of misclassifications for that parcel are above the threshold, it is considered as an alert.

There are two important parameters that we tune here. The probability difference above which a parcel is classified as green and the dynamic threshold which is used to define the alerts.

Based on the validation data provided from the PA (D3.2), we can estimate that the percentage of false declarations in Lithuania is roughly 3%, and the vast majority of them is related with Grassland cases. For this reason, we selected values for the aforementioned parameters aiming to approach this 3% of false declarations.

Finally, we have also used the confidence scores of a trained RF model with the labels of the highest level of taxonomy (see Figure 9), in order to enhance even more the accuracy of alert identification system. Based on the level of confidence and the level of disagreement already mentioned, we will distribute all the cases into three alert categories:

- **High-Risk Alerts**: These are alerts that the predictions and the respective declarations disagree on the highest level of crops hierarchy and we are strongly confident that they have been declared erroneously.
- **Medium-Risk Alerts**: These are alerts that the predictions and the respective declarations disagree on a lower level of crops hierarchy but agree on the highest level of crops hierarchy, and thus we are less confident that they have been declared erroneously.
- **Low-Risk Alerts**: These are cases that we are not confident regarding the outcome of the prediction.
- **Compliant Cases**: These are cases that the predictions agree with the declarations.

Last but not least, early results of smart sampling at the beginning of the cultivation period will be used in order to indicate **Early Alert** cases into the platform based on the alerts raised at the very start of the farmer application submissions.

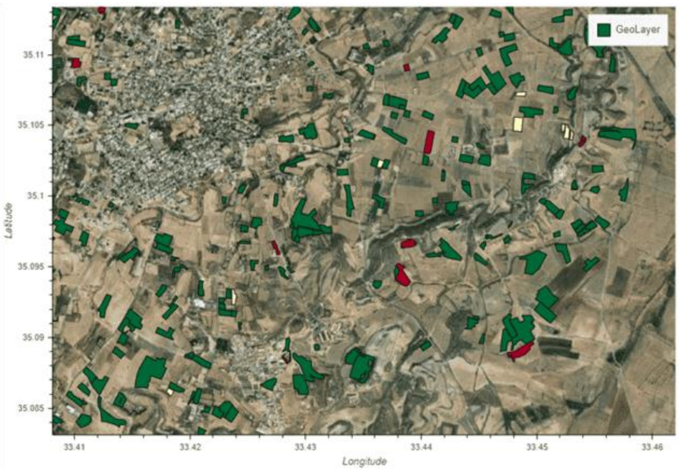


Figure 18: Declarations Alert Map

Crop Diversification (CD) Compliance Map

The Crop Diversification Service exploits the Land Parcel Information System (LPIS) and the declarations of the farmers. CD compliance map is a compilation of ***if conditions*** according to the **greening 1** set of rules listed in the description section above and "worst case scenario" approach (presented by JRC, MARS conference, November 2018) which examines the hypothetical impacts between an actual truth and crop label mapped. For the area estimation, we will use the area size of the fields calculated (in hectares) using the GIS geometries before the buffering and the rasterization of the input shapefiles. More specifically, by using the lookup table (see Table 8) and the collection of CD if conditions mentioned earlier, we can infer the compliance or not of the holders. For the case of Lithuania, the crop codes to be used for the service are that of declarations, and it will change to the predicted ones from the CTM module only if they have been indicated from the traffic light system as potentially wrong. The results will be visualized via a shapefile map, illustrating the compliance or not of the applicants, or any exemptions from the respective regulations if it is necessary. The information will be accompanied with a relative comment box describing the exact category of the cases or any other form of problem may arise in case we cannot infer any result.

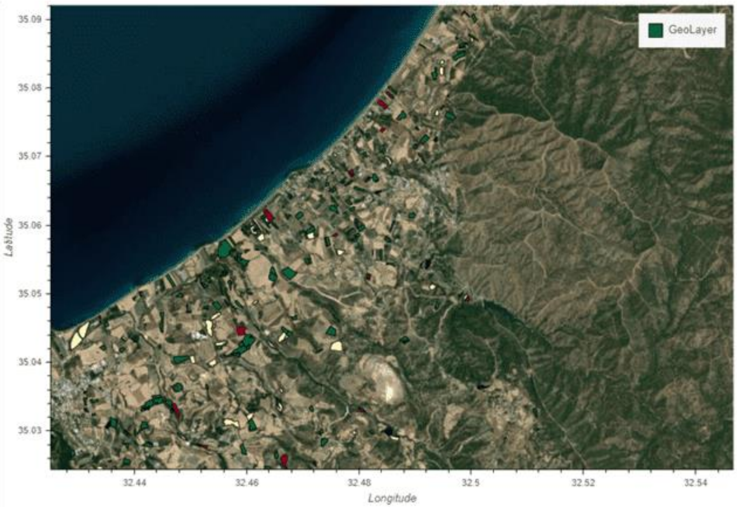


Figure 19: Crop Diversification Compliance Map

4.2.2.3 Grassland Mowing Events Detection

Grassland Mowing Detection is another ENVISION data product which consists of 3 individual pre-processing steps (D4.1):

- Data Fusion
- Mowing Events Detection Algorithm
- Mowing Compliance

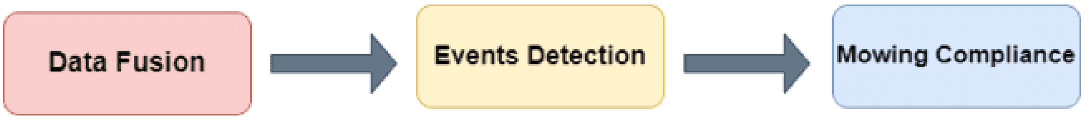


Figure 20: Mowing Events Detection product scheme

The Data Fusion (DFS) workflow combines Sentinel-1 and Sentinel-2 data in order to "increase” the number of cloud-free observations. Optical sensors are sensitive to clouds resulting in gaps in the time series. Sentinel-1 data are weather independent and not affected by clouds, therefore they can assist in predicting the Sentinel-2 values in the case of cloud obstructions. On the other hand, Sentinel-1 data are sensitive to water element, especially when meteorological phenomena of precipitation or high humidity are taking place. In ENVISION we use **Sentinel-1 and Sentinel-2 as input in Deep Neural Networks that generate cloud-free Sentinel-2 time series.** The results are ingested in the ENVISION DataCube and are used as input to the rest of the services, mainly the Grassland Mowing Events Detection. The service will provide complete S2 time series for the NDVI of every pixel, acting as an ancillary interpolation routine in order to replace the respective cloudy observations [9].

The Grassland Mowing Events Detection micro-service is an EO change detection module that exploits satellite data along with the usage of Deep Learning algorithms. The main scope is to efficiently monitor grassland activity (of grassland fields indicated by the CCTM) and precisely track the key dates of those cultivation events. Specifically, it combines Sentinel-1 data (VV, VH and VV/VH backscatter polarization coefficients and VV, VH coherences) and the reconstructed Sentinel-2 NDVI time series (see Data Fusion). The product is updated with every new Sentinel-2 acquisition. Finally, the grassland event detection service provides dynamically updated maps, accompanied with the respective confidence level. The grassland mowing detection processing chain is built on top of the DataCube, allowing for large scale and timely application. This information can be used from the PAs to estimate the compliancy of the farmers with regards to mowing grassland regulations (Table 5). Overall, the mowing event maps are populating the CAP monitoring ENVISION DataCube as soon as they are produced. Thus, we can also create specific geo-queries and extract summary statistics for particular spatial and temporal conditions. One example of such a query would be to retrieve the mowing intensity profile of a predefined area (Figure 21).

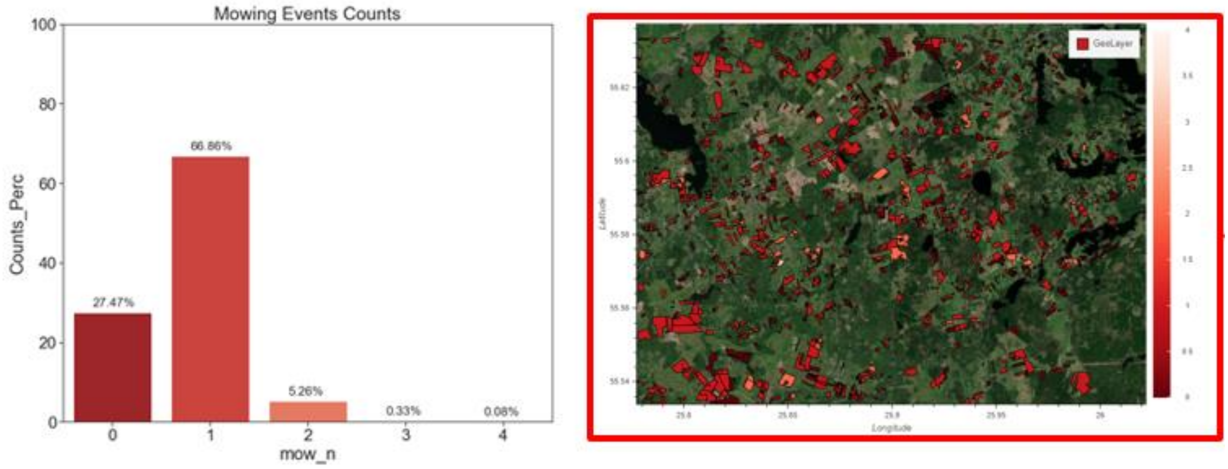


Figure 21: Mowing Summary Statistics of a Predefined Area

<u>Input data</u>

I. Satellite:
   a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC)
      i. Vegetation Indices - VIs (NDVI)
      ii. Scene Classification (SCL)
   b. Sentinel-1 GRD (rel. orbits: 58, 131)
      i. Backscattering coefficients (VV-VH-VV/VH)
   c. Sentinel-1 SLC (rel. orbits: 58, 131)
      i. Coherence coefficients (VV-VH-VV/VH)

II. Paying agencies:
   a. Exact regulations characterize the grassland mowing or grazing policies. These files contain, among others, the maximum number of allowed mowing events and the exact period, during these events can take place.
   b. Grassland's polygons to be check after CCTM as a shapefile

The boundaries of the geometries are reduced through a 5m buffer zone and then rasterized in order to avoid conflict cases of mixels and acquire samples that are more representative.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Output data

The service will provide:

I. Dynamic Events Map as a shapefile of grassland mowing detection per parcel encapsulating all the extracted information regarding the detected events, their confidence levels and their compliance into the respective mowing regulations (Figure 22).

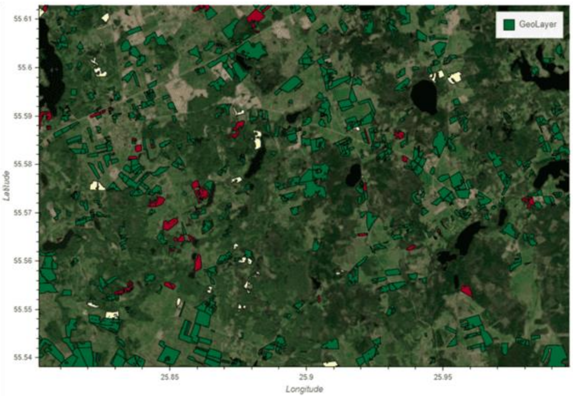


Figure 22: Mowing Compliancy Map

Data Fusion

The main goal of this product is to tackle the problem of cloudy observations of Sentinel-2 images, taking into consideration information coming from Sentinel-1 and the rest of cloudless Sentinel-2 cases. A very sophisticated Deep Neural Network architecture was designed, taking into consideration the current state-of-the-art, in order to **provide complete NDVI time series without gaps**. The DNN module is trained and applied to every **pixel** individually. The inputs of the model are the VV, VH and VV/VH backscatter coefficients and the VV and VH coherence, as well as the cloud-free NDVI observations at disposal until that time. In order to take advantage of the 6-day revisit time of Sentinel-1, the Sentinel-2 data are shifted and assigned to the closest S1 transition date. This way, we create a new more robust and resilient feature space of a consistent time step, which will transfer a S2 measurement at most 3 days earlier or later. In order to train the model, dense NDVI time series are temporal resampled (through daily linear interpolation) and smoothed, in order to assume the lowest possible deviation from the actual ground truth of the missing values and then they are used as training labels.

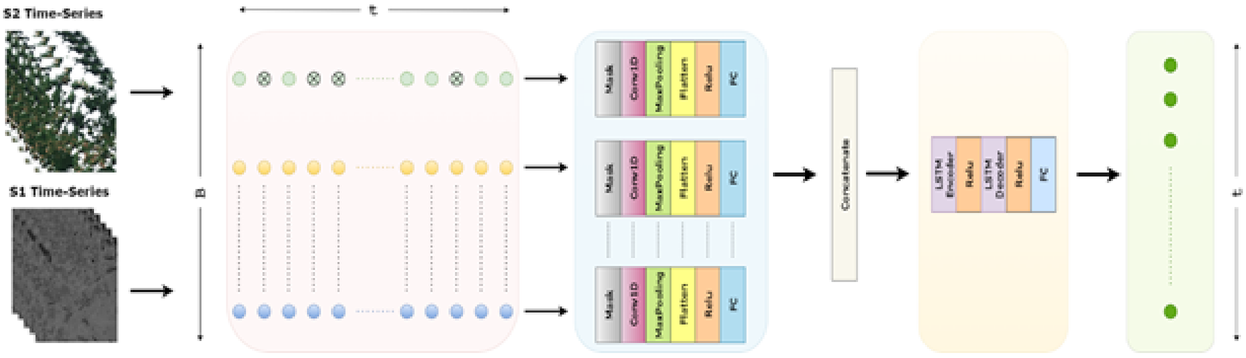


Figure 23: Data Fusion DNN architecture

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The respective NDVI time series that are used as inputs are subjected to an artificially hiding of several random time steps in order to simulate the existence of cloudy observations. These values are replaced with a value of –1 and they are masked inside the model. At the end, the model will use the rest of the available S2 observations of the adjacent time steps and the current SAR and InSAR observations in order to infer the missing values and generate the final complete NDVI time series. The architecture of the model (Figure 23) is composed from a succeeding series of 1-D convolutional and encoder-decoder blocks of modules. The convolution module in the beginning is trying to minimize the percentage of noise coming from every band of the input data (NDVI_input, InSAR VV-VH, SAR VV-VH) and it comprises subsequent masking, convolution and pooling layers that acts in parallel. Afterwards, the encoder-decoder module consisting from LSTM layers, is trying to identify any temporal correlation and patterns between the different time-steps and synthesize the final continuous NDVI output.

All in all, **at the end the model will be able to provide results for every pixel inside a parcel geometry and alleviate the problem of cloudy observation and sparse NDVI time series**. Even more S1/S2 fusion product is at place to provide very smooth time series and eliminate the noise coming from cloudy cases that pre-processing cloud masks were not able to detect, something that, especially for the mowing detection task, consists a major problem since these abrupt changes can be identified mistakenly as potential mowing events.

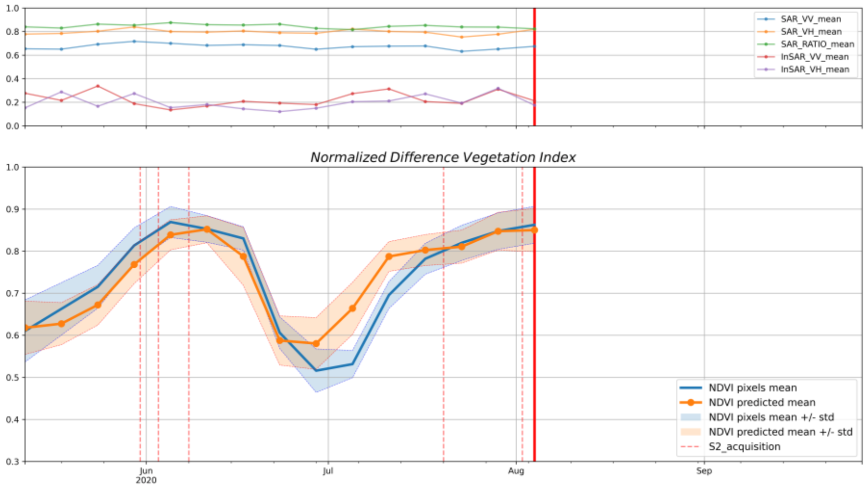


Figure 24: Data Fusion Results over time

Mowing Events Detection Algorithm and Compliance Check

The grassland mowing events detection is based on monitoring Sentinel-1 and Sentinel-2 images over time and identifying abrupt changes that are then characterized as a mowing event. The algorithm is dynamic and provides a new mowing events layer with every new acquisition. Here we implemented a **novel DL architecture,** similar to the S1/S1 fusion one (Figure 23), that takes as input the new artificially created NDVI time series of a fixed time step along with the S1 data (backscatter and coherence) and aims at identifying the 6-day timeframe during which an event took place (Figure 25). However, Lithuanian PA was able to provide us with labels referring only to the compliance of the parcels – D3.2 (if at least one mowing event took place during the mandatory mowing period – see Table 6). Hence, in order to generate annotated training and validation event instances and train the model there are two ways to accomplish it: I) either through a photo-interpretation process II) or by using results provided from other similar solutions (e.g., SEN4CAP). We used both the aforementioned

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

in order to acquire as many training instances as possible and create a model based on samples from the different regions of the AOI.

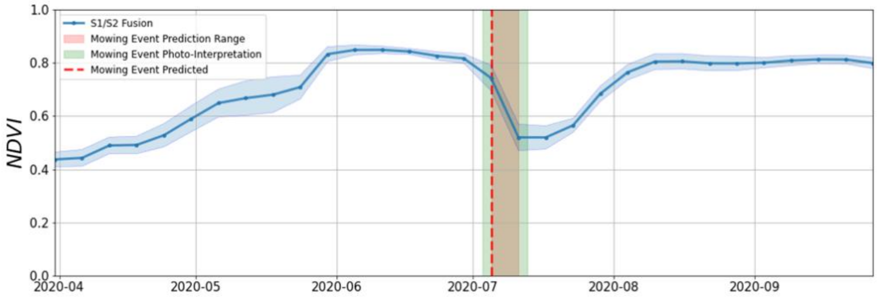


Figure 25: Mowing Event Identification

Analysis on each pixel individually and aggregated statistics can provide us with a representative level of confidence regarding the extent and the exact time instance that a mowing event took place.
Taking into consideration the type of grassland (see Table 6), and the exact instance that a mowing event is detected from the model, logits probabilities are used to prioritize inspections and quantify the risk of non-compliance. This way we can move a step closer to exhaustive monitoring. For Lithuania, the percentage of non-compliance it is expected around 5%.
Moreover, considering that parcels are usually mowed gradually through time [10], one big advantage of working with pixel level mowing markers is that it is feasible to evaluate the spatial extend of an event taking place (Figure 26). That way we can respond directly to many CAP requirements related to the proportion of the mowed area w.r.t the total parcel area declared.

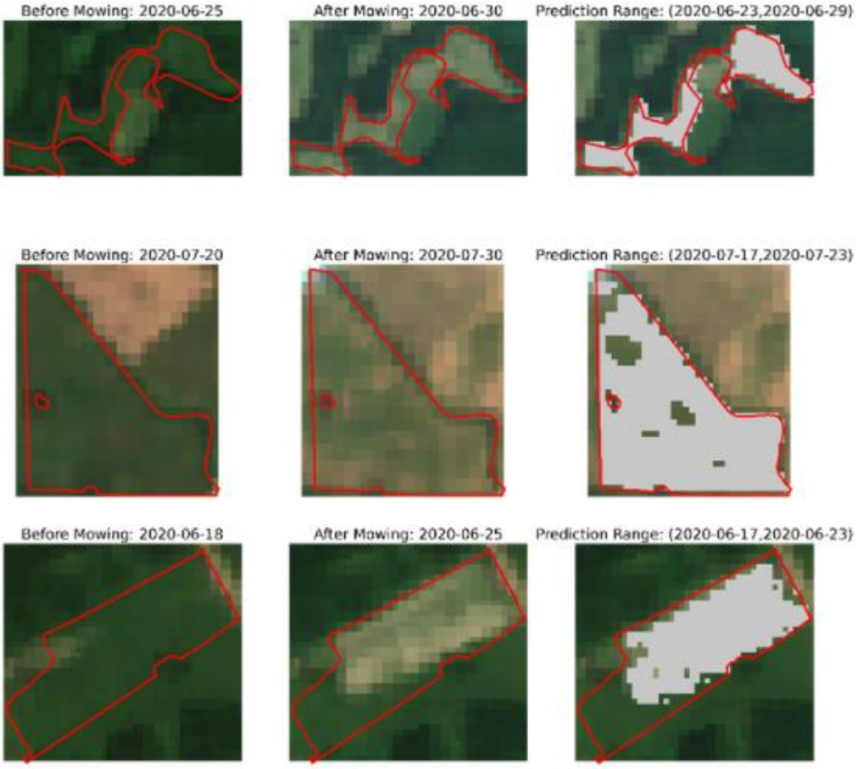

Before Mowing: 2020-06-25
After Mowing: 2020-06-30
Prediction Range: (2020-06-23,2020-06-29)
Before Mowing: 2020-07-20
After Mowing: 2020-07-30
Prediction Range: (2020-07-17,2020-07-23)
Before Mowing: 2020-06-18
After Mowing: 2020-06-25
Prediction Range: (2020-06-17,2020-06-23)

Figure 26: Mowing events expanse over parcels

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

4.3 BC2: Monitoring multiple environmental and climate requirements of CAP – Cyprus

The Cyprus Agricultural Payments Organization (CAPO) is responsible for the Cypriot business case which includes tasks of on-the-spot controls, maintenance of the LPIS, definition of the eligibility criteria of the declared parcels, verification process for the parcel eligibility according to the relevant legislation of the EU standards and registration of the parameters that witness the eligibility of each parcel, in an electronic fashion. Until today, in Cyprus, traditional practices of monitoring and inspection of the farmers is taking place, consisted from 4 individually steps: 1. the submission of the farmer's electronic application. 2. the on-the-spot checks combined with some remotely sensed controls (using VHR data that are very expensive), 3. administrative controls and 4. submission of additional documents from farmers when required. The decision on the cases to be checked from CAPO is based on criteria resulted from an annual *risk analysis* according to some informative factors (e.g. area of interest, age of the applicant, crop type, etc.) and a smaller sample of randomly selected parcels. It becomes apparent that a "smart" and self-operative service for the entire agricultural land of Cyprus will be of high supportive value for the national paying agency, in order to plan their OTSC expeditions and fulfil the requirements related to the supervision of Common Agricultural Policy (CAP) measures. Ideally, the process of monitoring and verification of farmers' applications should be converted to a fully-automated routine via remote sensing procedures, significantly decreasing the amount of money spent and time or effort consumed for that purpose, during the OTSC period of step 2. Finally, it is crucial for CAPO to be **alerted in time** for possible non-compliances of CAP requirements, to assist in subsidies allocation decisions or future warnings related to step 3. Taking the aforementioned into consideration, this section presents the general picture of agricultural land of the business case of Cyprus accompanied with a brief descriptive Statistical Analysis to highlight the possible outcomes and limitations there, a thorough explanation of the present pilot's requirements (see also D.2.2) and a detailed description of the methodologies applied with a view to sufficiently answer the respective problems and generate useful product and services.

4.3.1 Data products description

Study Site

Cyprus is an island country located in the eastern Mediterranean Sea. It is characterized from a subtropical climate, according to Köppen climate classification scheme [1], with mild winters and hot

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

summers. Rainfalls are taking place during November until March and summers can be characterized as mostly dry with absence of rain that are accountable for barren lands for extensive long periods [11]. According to [12], Cyprus can be divided in four distinct climatic zones. The mountainous regions, the semi-mountainous, the inland plain and coastal areas.

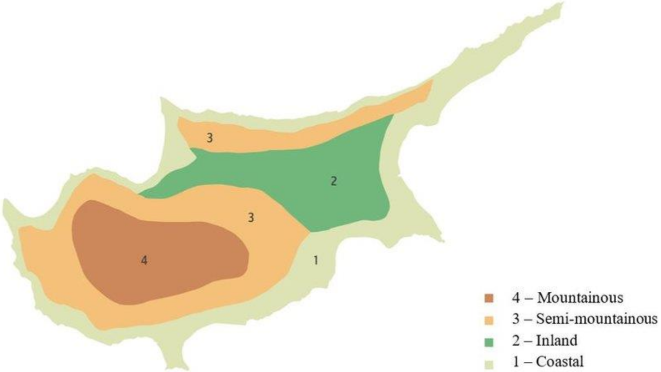


Figure 27: Climate zones in Cyprus, adapted from [11]

All these apparently play significant role in the profile and the development of agricultural land across its territory, since the local environmental factors can affect directly the raise of the different crops and the application of various rural practices accordingly. In Figure 28 we can see an illustration of the different crops across the country based on farmers' declarations of 2020. Taking into account also the respective climatic and morphological zones already mentioned (Figure 27), we can observe the different distribution of cultivated crop families across Cyprus. Cereals, potatoes, vegetables and generally the vast majority of seasonal crops are cultivated in the Eastern inland and coastal part of the country, while vineyards, grasslands and most of the permanent cultivations (olive trees, carob trees, walnut trees, etc.) can be found in the western and central mountainous part of the island. Hence, to address the requirement related to the provision of results for all products and services across the entire territory the country (UR17), we should examine the various morphological parts individually.

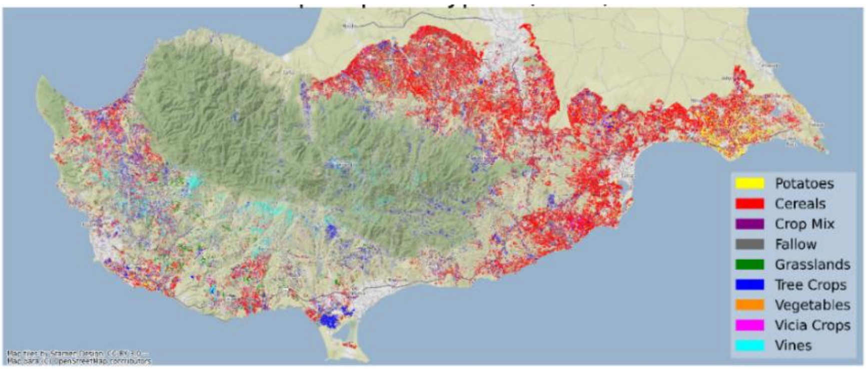

Potatoes
Cereals
Crop Mix
Fallow
Grasslands
Tree Crops
Vegetables
Vicia Crops
Vines
Map tiles by Stamen Design, CC BY 3.0 —
Map data (C) OpenStreetMap contributors

Figure 28: Crop map of Cyprus (2020)

Figure 29 presents the numerical distribution of the crop declarations across the country for the year of 2020. Almost the 90% of the total declaration can be described from less than 20 classes. Specifically, the most frequent crops are Barley, Olive Trees and Fallow lands (an abstract and particular case that is going to be discussed later in this section).

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

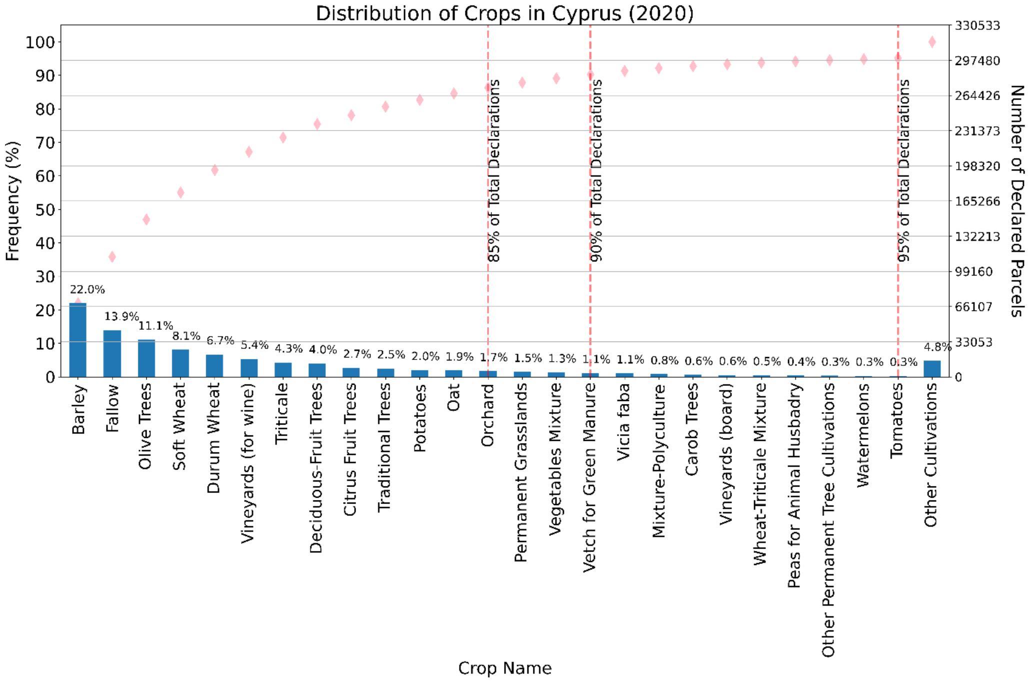


Figure 29: Cyprus Crop Type Distribution (2020)

Moreover, by assigning the finest level of granularity crop codes to the more general crop families, according to their agro-botanical peculiarities, and successively to an even more generic semantic categorization according to the local demands and practices applied, we can form the different levels of crop taxonomy for Cyprus (Figure 30). It should be mentioned that in the near future it is planned to arrange workshops with CAPO in order to form the most suitable classes that explain the dataset in the best way.

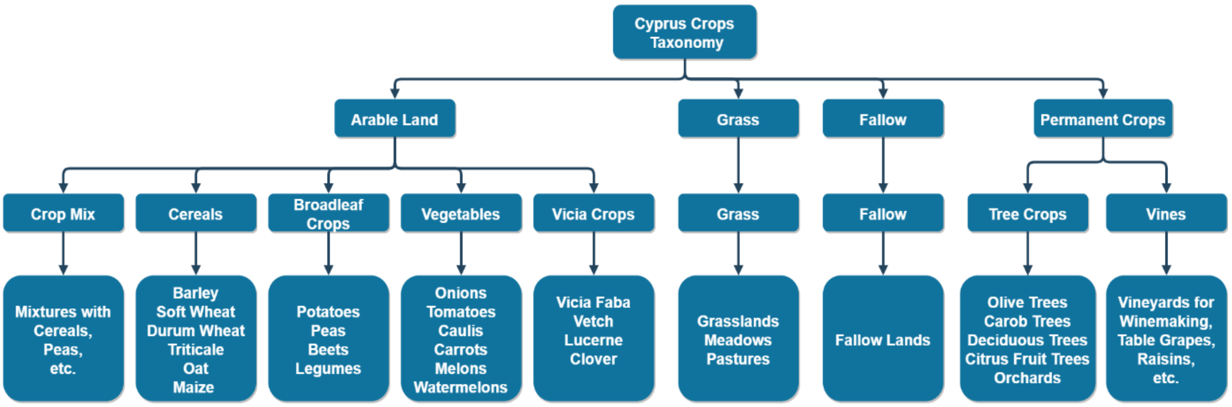

Cyprus Crops Taxonomy
Arable Land
Grass
Fallow
Permanent Crops
Crop Mix
Cereals
Broadleaf Crops
Vegetables
Vicia Crops
Grass
Fallow
Tree Crops
Vines
Mixtures with Cereals, Peas, etc.
Barley Soft Wheat Durum Wheat Triticale Oat Maize
Potatoes Peas Beets Legumes
Onions Tomatoes Caulis Carrots Melons Watermelons
Vicia Faba Vetch Lucerne Clover
Grasslands Meadows Pastures
Fallow Lands
Olive Trees Carob Trees Deciduous Trees Citrus Fruit Trees Orchards
Vineyards for Winemaking, Table Grapes, Raisins, etc.

Figure 30: Crop Taxonomy for Cyprus

Nevertheless, by exploiting that information, we can also see the numerical distribution of crop families. It is worth mentioning that the category of arable land, which is composed from cereals, vegetables and the leguminous plants like vicia (these are extents of land that are mainly utilized as manures or for fertilizing purposes, since they can provide light amounts of macronutrients, like nitroge) concerns more than 50% of total of the farmers’ declarations.

Small parcel size

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

Fields size is something that plays a significant role in remote sensing and as a result may arise future problems due to the fixed resolution of Sentinel satellites. Therefore, **Cyprus is a challenging occasion** since it is characterized by a very small average parcels' size. The **mean parcel size is roughly 0.4 hectares** and, in most cases, the declared fields **do not even exceed the size of 1 hectare** (Figure 31).

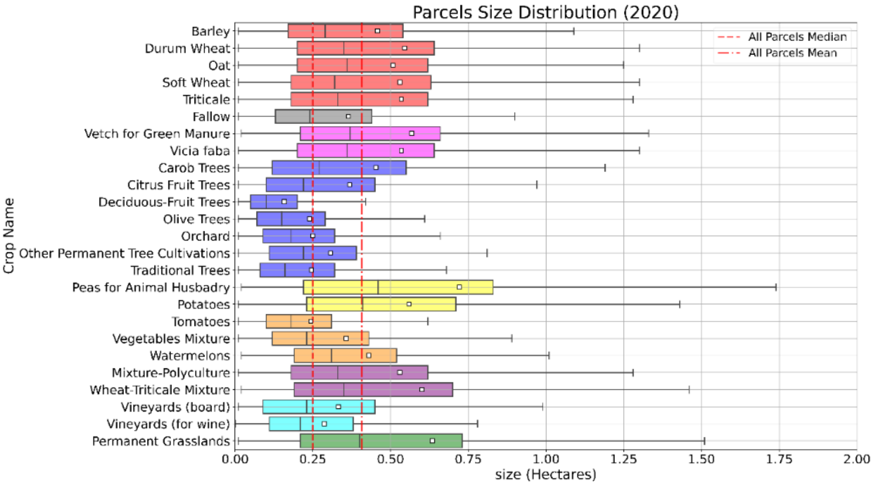


Figure 31: Cyprus Parcels' size Distribution (2020)

In addition, despite the small size of the parcels, their shape is usually **narrow and lengthy**, forming thin strips of land, something that **makes the monitoring task even more demanding**. As a result, Cyprus is a very special case since the average parcel-size is comparable to the sentinels' pixel-size resolution and narrow-shaped geometry of the fields allows for only a trivial number of pixels, which in most cases are mixels (pixels that belong to more than one fields) providing vague information (see Figure 32). Only few of them have enough number of pixels available and can be considered representative.

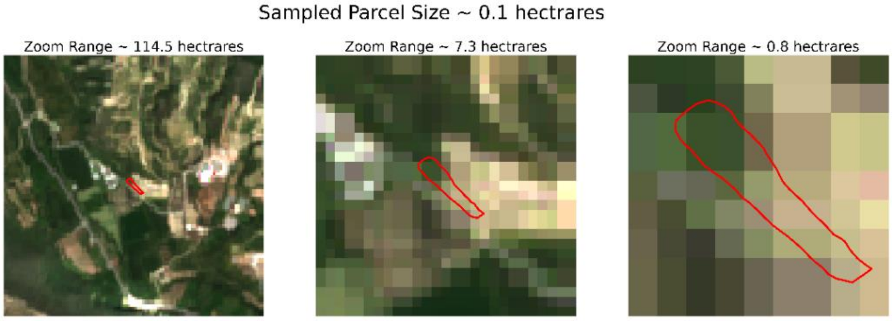

Sampled Parcel Size ~ 0.1 hectrares
Zoom Range ~ 114.5 hectrares
Zoom Range ~ 7.3 hectrares
Zoom Range ~ 0.8 hectrares

Figure 32: Sample parcel of Cyprus from S2

4.3.1.1 Analytics on Vegetation and Soil Index Time-series

Minimum Soil Cover

The rules regarding the existence of minimum soil cover in Cyprus focus on monitoring soil conditions from January to February. In this period of consideration, Cyprus records the greatest measurement of rainfall. At the same time, parcels with slope lower than 10% are excluded from monitoring as the risk for soil erosion is low.

NVR

Regarding the runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas, the requirements for Cyprus are the same as the one described before for Lithuania.

Natura 2000 Hotspot Detection

Natura 2000 is a network of protected areas in the European Union aiming to assure the long-term survival of Europe’s most valuable and threatened species and habitats. While each member state is responsible for managing its own sites, at the same time, they must comply with the requirements of the directives of the European Union [13]. In Cyprus, any agricultural intervention in land inside Natura 2000 sites is prohibited, except special permission has been given. Additionally, Natura policies are not applicable to parcels declared as arable land inside these protected sites, before Natura 2000 was put into effect.

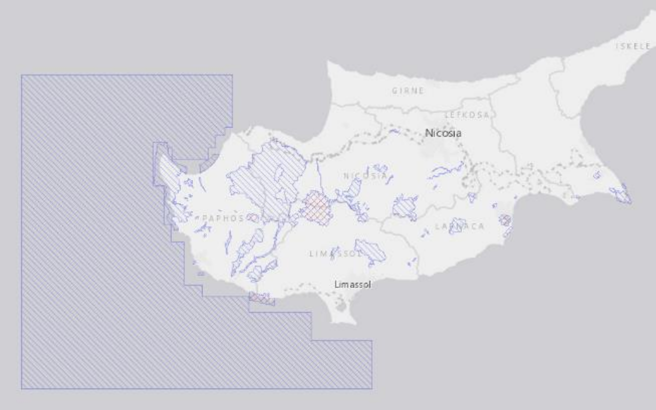


Figure 33: Natura 2000 Network sites in Cyprus

For the time being, Cyprus Agricultural Payments Organization (CAPO) performs random on-the-spot controls in order to locate illicit agricultural practices inside Natura protected sites. Natura hotspot detection data product will satisfy the requirement of CAPO (UR5/UR17) to have a detection mechanism in order to operate their on-the-spot controls efficiently, checking only parcels where illegal activity was spotted.

Stubble Burning Identification

As stated above in the case of Lithuania, mapping of burnt areas has proved to be of high importance for paying agencies in the agricultural sector. In Cyprus as in most countries in the Mediterranean, wildfires are very common and frequently caused by stubble burning. For CAPO, it is significant to have an overview of the frequent stubble burning in order to perform targeted OTSC (UR5/UR17), and discriminate them from the wildfire cases.

4.3.1.2 Cultivated Crop Type Maps (CCTM)

Cyprus cultivation period starts from mid-October where we have some terrain preparation and ploughing of the agricultural land and practically finishes at early June (with some exceptions of vegetables and greenhouses) since almost the total area has been transformed to a barren land due to high temperatures and scarcity of water for irrigation purposes. This is also the period when the holder's subsidy applications terminate and the control deputies schedule their checking campaigns. In the Figures 34, below, we see the NDVI signatures of various crops, in the area of Cyprus. Similar to the BC of Lithuania, **the distinction between the different crop type is quite a challenging task**, since crops share very common growth rates. On the contrary, the separation in the family level is much easier than in the finest level of the unique crop types. Furthermore, we can note that the degree of standard deviation and the range of crop signatures between the separate classes is quite broad, resulted to cases among the same crop types that is difficult to be distinguished.

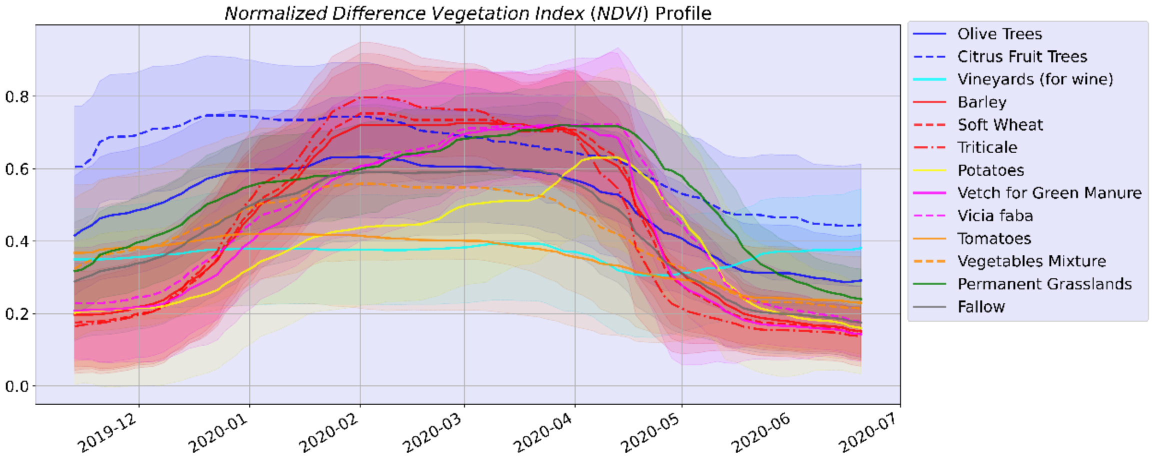


Figure 34: Crops NDVI signatures (Cyprus-2020)

Several of the requests (UR7/UR17/UR18) are strongly aiming on **crop type prediction** for the entire country, during the cultivation period and **as early in the year as** possible. Hence, there is a need of **dynamically, accurately and efficiently** classifying the different crop types in the area of interest. Similar to BC1, the results are provided for multiple periods and levels of granularity at parcel-based level to answer the requirements related to crop identification, early warning of the applicants and monitoring of several cross-compliance policies.

Since it is the **first time that Cyprus is participating** in such a project and there is no relevant legacy on crop classification projects of this area, and a preliminary analysis is necessary to pinpoint the local particularities.

Small intra-parcel coherence and crops vague nomenclature

Agricultural fields coherence and delineation do not present a very smooth picture due to several factors, which are related to the mixture of alternate crop types in the same parcel, crop residues from previous cultivation periods and intense natural vegetation emerging on the fields background. Moreover, there are other occasions where separate adjacent crop fields might be declared in the same parcel, provoking high intra-parcel variation.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

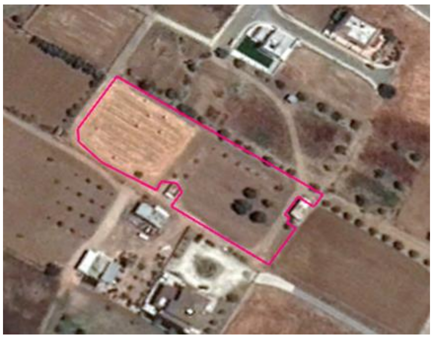


Figure 35: Multiple crop cultivations declared in the same parcel

Likewise, it is not trivial the number of cases that the delineation of a parcel does not coincide with its actual geometry, emphasizing one more time the importance and the utility of a proper boundary detection routine that will indefinitely assist in the proper definition of parcels borders. Moreover, crops residues and background vegetation can notably affect the spectral signatur, especially in olive trees and other sparse permanent cultivations cases, where the distance between the trees is quite large for the tree crowns in order to cover the ground greenery.

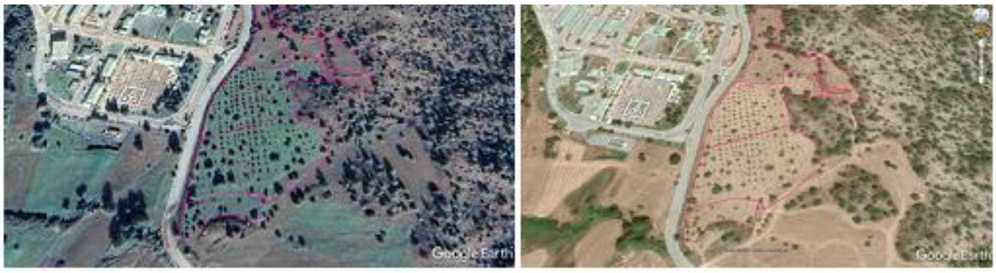


Figure 36: Background vegetation during cultivation period provokes distortion on spectral signatures

Overall, it is critical to review the nomenclature of the different crop codes provided from the users into the service, creating more informative and explicit categories, since crops that are almost alike and present common spectral characteristics are sometimes declared with different names or codes. For example, vetch and vicia, share almost identical spectral signatures, rendering the discrimination between them impossible. Thus, they could be considered as one category.

__Fallow Land for Cyprus__

Fallow lands usually are unsown arable lands for one or more vegetative cycles that have great ecological importance since they are included in the crop rotation and crop diversification system (subsidies under greening regulations), as well as other beneficial agricultural and environmental conditions (GAEC) of EU. However, they are a very vague issue in Cyprus because of their **ambiguous nature and their very loose definition**. In Cyprus, we have two distinct cases of fallows:

- *Land Under Fallow*: Generally, as fallow land is considered the expanse that sustains the minimum agriculture activity (an artificial milling of the ground with at least a ploughing event during the spring of the cultivation period, before pests' insemination) and there is no yield production for no less than 6 months from 1st of February until the 31st of July.
- *Expanses with sparse trees inside*: This is the case of areas with scattered trees (usually olive trees, carob trees or other traditional trees of the region) that block the agricultural activity. Permanent

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

tree cultivations may be considered as potential fallow lands, under the special condition that their distribution is rather sparse, namely no less than 4 trees and no more than 10 trees per 0.1 hectares of agricultural land.

Obviously, fallow lands constitute a very abstract matter, with a confusing defined terminology and as a result high variance of spectral signatures may be detected. In the first scenario, the most frequent one, fallow lands are spectrally similar to the rest arable land (cereals, legumes, etc.) due to the existence of crop residues from the previous cultivation periods and the high existence of natural vegetation (Figure 37a), while in the second one, the visual profile is more alike to that of permanent trees (Figure 37b). Here, we identify the necessity of the definition of more explicit crops denomination in Cyprus, since the same declaration code is being applied for different type of lands that are subjected to different rules and policies.

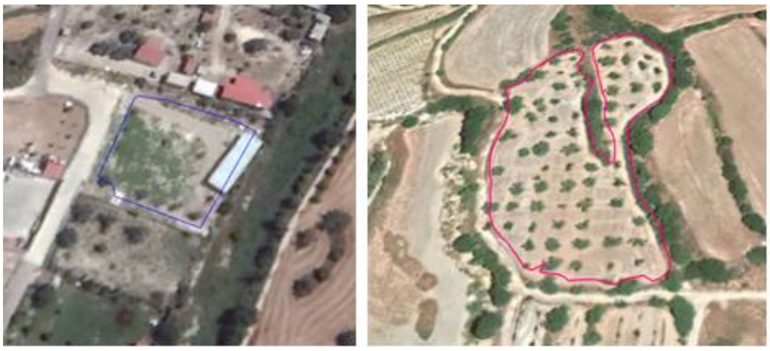


Figure 37 a. Land under fallow (left), b. Expanse with sparse trees inside (right)

<u>False Declarations</u>

Finally, another problem is that of erroneous crop code registrations during declaration period. For Cyprus, the proportion of the erroneous declarations is relatively high (around 10%). One reason for this is the negligence or carelessness of farmers during application period. Usually, holders of large expanses of land tend to repeat the exact same declarations, unaltered from the previous cultivation period. They omit to update them and they submit them into the system without any change. Another reason though, is that of personal benefit. According to CAPO, there are holders declare fields as fallow lands, even though they clearly contain tillable areas, in order to satisfy the minimum EFA or greening requirements and be eligible of the respective allowances. These are clearly cases of fraud and is useful for CAPO to be detected as soon as it is possible in order to make the necessary warnings and recommendations (UR18). We can see some examples in the Figure 38 below provided by CAPO, where we have fields that are clearly cases of cultivated arable land, while in the system are registered as fallow lands.

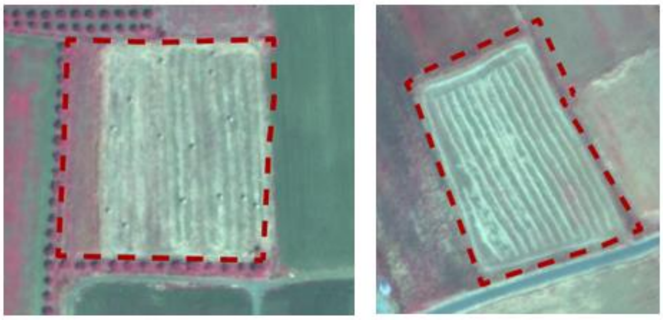
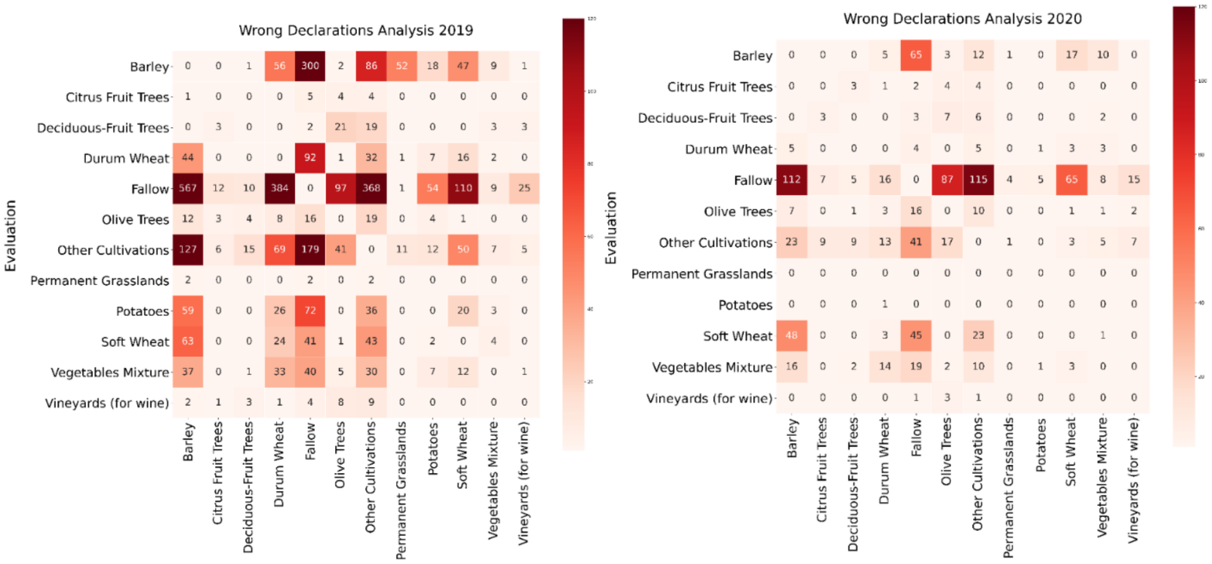


Figure 38: Cases of fields that are false declared as fallows

In case of Cyprus, one of the situations that we need to deal with in the general context of UR18, is the timely warning of the applicants for potentially wrong declarations, situations that many times end up with disputes between CAPO and holders or even trials.

Figure 39 presents the confusion matrices of the wrong declaration, acquired from the validation data of the OTSC (D3.2) or assessments through Very High Resolution (VHR) imagery for the years of 2019 and 2020. As expected based on what said so far, **the vast majority of wrongly declared cases** (more the 80%) are related with **fallow lands**.



Wrong Declarations Analysis 2019

	Barley	Citrus Fruit Trees	Deciduous-Fruit Trees	Durum Wheat	Fallow	Olive Trees	Other Cultivations	Permanent Grasslands	Potatoes	Soft Wheat	Vegetables Mixture	Vineyards (for wine)
Barley	0	0	1	56	300	2	86	52	18	47	9	1
Citrus Fruit Trees	1	0	0	0	5	4	4	0	0	0	0	0
Deciduous-Fruit Trees	0	3	0	0	2	21	19	0	0	0	3	3
Durum Wheat	44	0	0	0	92	1	32	1	7	16	2	0
Fallow	567	12	10	384	0	97	368	1	54	110	9	25
Olive Trees	12	3	4	8	16	0	19	0	4	1	0	0
Other Cultivations	127	6	15	69	179	41	0	11	12	50	7	5
Permanent Grasslands	2	0	0	0	2	0	2	0	0	0	0	0
Potatoes	59	0	0	26	72	0	36	0	0	20	3	0
Soft Wheat	63	0	0	24	41	1	43	0	2	0	4	0
Vegetables Mixture	37	0	1	33	40	5	30	0	7	12	0	1
Vineyards (for wine)	2	1	3	1	4	8	9	0	0	0	0	0

Wrong Declarations Analysis 2020

	Barley	Citrus Fruit Trees	Deciduous-Fruit Trees	Durum Wheat	Fallow	Olive Trees	Other Cultivations	Permanent Grasslands	Potatoes	Soft Wheat	Vegetables Mixture	Vineyards (for wine)
Barley	0	0	0	5	65	3	12	1	0	17	10	0
Citrus Fruit Trees	0	0	3	1	2	4	4	0	0	0	0	0
Deciduous-Fruit Trees	0	3	0	0	3	7	6	0	0	0	2	0
Durum Wheat	5	0	0	0	4	0	5	0	1	3	3	0
Fallow	112	7	5	16	0	87	115	4	5	65	8	15
Olive Trees	7	0	1	3	16	0	10	0	0	1	1	2
Other Cultivations	23	9	9	13	41	17	0	1	0	3	5	7
Permanent Grasslands	0	0	0	0	0	0	0	0	0	0	0	0
Potatoes	0	0	0	1	0	0	0	0	0	0	0	0
Soft Wheat	48	0	0	3	45	0	23	0	0	0	1	0
Vegetables Mixture	16	0	2	14	19	2	10	0	1	3	0	0
Vineyards (for wine)	0	0	0	0	1	3	1	0	0	0	0	0

Figure 39: Wrong Declarations-Evaluations Confusion Matrix (Cyprus 2019 and 2020)

<u>Greening Requirements and Crops Diversification (CD)</u>

Similar to the business case of Lithuania (BC1), also in Cyprus we are interested in the greening measures with regards to crop diversification, namely the maintenance of permanent grasslands between successive cultivation periods (5% margin of flexibility at national or regional level), as well as the sustainability of dedicated proportions of arable land to expanses beneficial for biodiversity and creation of Ecological Focus Areas (EFA). UR7 prerequisite a classification scheme in the lowest possible level of detail since the requirements regarding Crops Diversification (CD) are a set of rules regarding the greater variety of crops that makes soil and ecosystems more resilient. In the very specific case of Cyprus CD measures should be satisfied in the period between mid-March until mid-June and only

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

from applicants whose **total declared arable land is larger than 10 hectares** (cases of Total Arable Land smaller than 10 hectares are exempted) and more specifically:

a. If the total expanse of the arable land declared from the applicant is **between 10 and 30 hectares** it should contain **at least 2 different categories of crops** inside this land and:
   - The major cultivation must cover no more than the 75% of this land

b. If the total expanse of the arable land declared from the applicant is **more than 30 hectares** it should contain **at least 3 different categories of crops** inside this land and:
   - The two major cultivations must cover together no more than the 95% of this land

However, there are again some cases (again of total arable land larger than 10 hectares) that are exempted from the practices above and the applicant is **eligible of greening subsidies**:

Exemption A

If proportion of Arable Land (AL) larger than 75%:

- It is used for the production of grasses or other herbaceous fodder plants.
- It is cultivated with Leguminous Plants
- It is declared as fallow land
- Any combination of the above

Exemption B

If proportion of Eligible Agricultural Area (EAA) larger than 75%:

- It is declared as Permanent Grassland
- It is used for the production of grasses or other herbaceous fodder plants.
- Any combination of the above

Exemption C

If proportion of the declared Arable Land (AL) larger than 50% it is not declared from the farmer into the previous year declaration and as long as the entire amount of this arable land is cultivated with a crop type different from the previous year.

Remarks

- The special case of **Fallow Land with sparse trees** inside, **cannot be considered as Arable Land for greening purposes**.
- The greening rules do not apply to farmers who opted for the small farmer's scheme, for administrative and proportionality reasons.
- Organic farmers automatically receive a greening payment for their farm, as they are considered to provide environmental benefits through the nature of their work.

4.3.2 Methodology

4.3.2.1 Analytics on Vegetation and Soil Index Time-series

Based on the GA and D4.1 the services of task 3.3 will deliver also for Cyprus a series of analytics consisting of:

- Runoff Risk assessment for the reduction of water pollution in Nitrate Vulnerable Areas.
- Buffer zones for the proximity to waterways nearby.
- Minimum soil cover for Soil Erosion
- Natura2000 Hotspot Detection

- Stubble Burning Identification (SBI)
- Other GIS querying functionalities such as buffer analysis, area calculation etc.

The service will provide multiple Analytics reports throughout the year taking advantage of both Sentinel-1 and Sentinel-2.

Input data

I. Satellite:
   a. Sentinel-2 L2A (tiles: 36SWD, 36SVD)
      i. Spectral bands (B01-B12)
      ii. Scene Classification (SCL)
   b. Sentinel-1 GRD (rel. orbits: 167)
      i. Backscattering coefficients (VV-VH)
   c. Sentinel-1 Coherence (rel. orbits: 167)
      i. Coherence coefficients (VV-VH)

II. Products:
   a. Annual soil loss
   b. Rainfall erosivity factor
   c. Soil erodibility factor
   d. Slope length factor and slope steepness factor
   e. Crop and cover management factor
   f. Conservation supporting practices factor
   g. Slope

III. Paying agencies:
   a. A lookup table for all the available crop type names, codes, families and CD ancillary info
   b. Parcels geometries and initial declarations as a shapefile (updated when is necessary)
   c. Agricultural Practices Descriptions
   d. hydrographic networks
   e. Natura2000 regions

Output data

The service will provide:

I. Pixel-based Analytics or Aggregated Statistics per parcel over a specific time range
II. Runoff Risk level map of the parcels.
III. Shapefile of CC indicators of compliance such as Traffic light system, GAEC 1 and GAEC 6
IV. Rasters of Stubble Burning Identification (SBI) for the maintenance of organic matter in soil
V. Rasters of minimum soil cover for Soil Erosion
VI. Rasters of Natura2000 Hotspot Detection

Generation of long time series vegetation products

Long time series vegetation products will be generated the same way as Lithuania, described in the corresponding section before.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Mean, Median and Std values per index aggregated for each parcel

The automated pipelines for the determination of aggregated statistics per parcel and per index are implemented as described for the Lithuanian case.

Runoff Risk Assessment for the Reduction of Water Pollution in Nitrate Vulnerable Areas (GAEC 1/ SMR 1):

In order to answer the statuary monitor requirement and GAEC 1, a runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas has been developed, taking into account the proximity into the closest water areas. Therefore, distance from every point of parcel's geometry to the closes water surface is calculated. Parcels that are above a certain distance threshold are excluded. Afterwards, according to bibliography, several models have been developed to identify the probability or size of soil erosion. The Universal Soil Loss Equation (USLE) and its revised version Revised Universal Soil Loss Equation [5] are the most widely used and accepted empirical soil erosion models.

Monitoring of soil cover

The methodology for monitor soil cover stays almost the same as the one for Lithuania. What differentiates the one for Cyprus is the time period, as the monitoring has to take place during months January and February. In addition, there is no monitoring process for parcel with slope less than 10%. Again, the results of the SAVI calculation are aggregated in parcel level so to keep the mean value of clear pixels' SAVI.

Natura 2000 Hotspot Detection

The methodology here is the same with the harvest event detection of BC1. The only difference is that because of the customers' policy, which allows agricultural intervention in arable land declared before the Natura 2000 characterisation, LPIS declarations are being used in order to exclude permitted actions, inside Natura 2000 sites. Again, this will be used as a **baseline methodology** to assist in the process of generating **validation data**.

Stubble Burning Identification

As already mentioned in Lithuanian case, for the business case of Cyprus, this data product focuses on providing a stubble burning detection map (UR5/UR17). The methodology is the same with that of BC1.

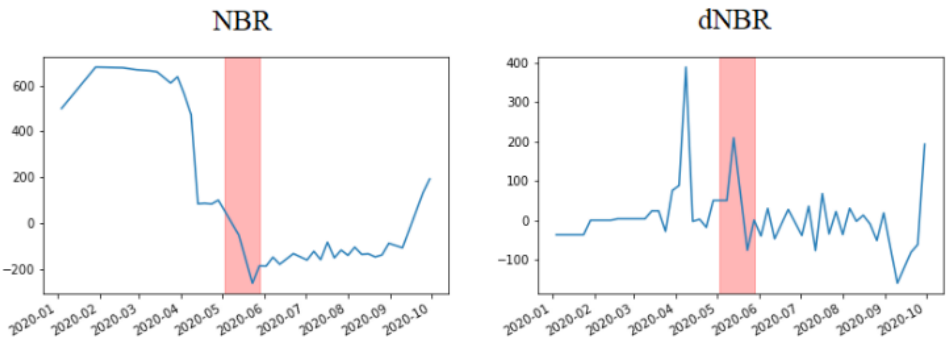


Figure 40: Burning Event detection period highlighted in pink

4.3.2.2 Cultivated Crop Type Maps (CCTM)

As already mentioned in Lithuanian case, for the pilot of Cyprus, the service will provide multiple crop type classifications throughout the cultivation period, for every new group of Sentinel acquisitions, to assist in **early alert** during the application process, optimize the confidence of the classification process and work as the basic pillar for the monitoring of crop diversification requirements. Data pre-processing will follow the routines described in D4.1 and the product will be built on top of the DataCube. The provider of the product is NOA, whereas the results of the product will be provided either via a RESTful API or as shapefiles.

Input data

More specifically we utilize the following data:

I. Satellite:
   a. Sentinel-2 L2A (tiles: 36SWD, 36SVD)
      i. Spectral bands (B01-B12)
      ii. Scene Classification (SCL)
      iii. Vegetation Indices - VIs (NDVI, NDWI, PSRI, etc.)
   b. Sentinel-1 GRD (rel. orbits: 167)
      i. Backscattering coefficients (VV-VH)
II. Paying agencies:
   a. A lookup table for all the available crop type names, codes, families and CD ancillary info (see table 8)
   b. Crop declarations as a shapefile (updated when is necessary)

Similar pre-processing routines mentioned with Lithuanian case is followed in order mask out cloudy measurements and fill gaps using linear interpolation. Additionally, geographical coordinates of each pixel are used as inputs in order to include any geo-spatial information.

In D3.2 can be found an extensive description of the auxiliary data for BC2 and T3.4.

Table 8: Cyprus Look-up Table sample schema

Crop Code	L0	L1num	L1	L2num	L2	L3num	L3	CDnum	CD	EAA	AL	Grass	EFA	Cwater
0	0	0	Greenhouses	0	Greenhouses	0	Greenhouses	0	Greenhouses	0	0	0	0	0
1	1	1	Arable Land	1	Cereals	1	Durum Wheat	1226	Soft Wheat	1	1	0	0	0
3	1	1	Arable Land	1	Cereals	3	Barley	1003	Barley	1	1	0	0	0
4	1	1	Arable Land	1	Cereals	4	Oat	1004	Oat	1	1	0	0	0
40	1	1	Arable Land	3	Broadleaf Crops	40	Potatoes	1040	Potatoes	1	1	0	0	0
68	1	2	Permanent Cultivations	7	Tree Crops	68	Citrus-Fruit Trees	2000	Citrus-Fruit Trees	1	0	0	0	0

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

70	1	2	Permanent Cultivations	8	Vines	70	Vineyards (wines)	2000	Vineyards (wines)	1	0	0	0	0
75	1	4	Fallow	9	Fallow	75	Fallow Land	4000	Fallow Land	1	1	0	1	0
76	1	3	Grass	6	Grass	76	Permanent Grasslands	3000	Permanent Grasslands	1	0	1	0	0
150	1	1	Arable Land	4	Vicia	150	Vetch for green manure	1177	Vicia Faba	1	1	0	1	0
170	1	3	Grass	6	Grass	170	Grasses or other herbaceous fodder plants	1170	Grasses or other herbaceous fodder plants	1	0	0	0	1
177	1	1	Arable Land	4	Vicia	177	Vicia Faba	1177	Vicia Faba	1	1	0	1	0

Output data

The service will provide:

I. **Dynamic crop type maps** as a shape file over the registered parcels for every new or group of new Sentinel acquisitions.

II. Traffic **light maps as a shapefile** over the registered parcels for **smart sampling** of on-the-spot inspections and **early alert of the users.**

III. **Greening-1 compliance map** as a shape file over the registered parcels at the end of the cultivation period.

Dynamic Crop Type Classification

The procedure for the generation of **dynamic crop type maps** is similar to that of the BC of Lithuania (BC1). The results are provided for the major 15 categories, which describe over the 85% of the total number of parcels. In this case, we used the Support Vector Machines (SVM) [14] algorithm, which has been also used widely in operational scenarios and provides optimal results comparable to RF. In order to reduce the dimension of the input feature space and lessen the time complexity that characterizes SVMs, Principal Component Analysis (PCA) [15] was also used.

Furthermore, since Cyprus average field size is significantly small and the corresponding geometry does not allow for obtaining a representative number of clear pixels (not mixels) for each parcel, results of CCTM will be initially evaluated on the **pixel level** and then will be aggregated into a parcel level, by using the median of the pixels' logits. Training of the model is based on cases that contain at least 10 clear pixels, while the rest will be used only for the inference of the results. Moreover, we have excluded a test dataset, based on the OTSCs performed by the CAPO, which is around 10-15% of the total parcels. Consequently, the dataset is divided into training and validation sets (30% and 70% respectively), in a stratified fashion, which means that we take into account the crop types distribution. Opposite to the BC1, here, we adapt a hierarchical based classification scheme in order to exploit information coming for the higher level of taxonomy, as well as class weights inversely proportional to the crop frequencies, for the training of the SVM. Lastly, the special case of fallows is excluded from the main model and they are examined individually, using an outlier detection module based on an extra binary SVM.

Hierarchical Classification and Fallow Outliers Detection

The general concept of the hierarchical scheme is to exploit the information coming from higher level of taxonomy between the various crops (see Figure 30) since classes at these stages are easier to be distinguished. That way, SVM models trained on the different level of information (land use and crop family) are able to impose extra knowledge to the finest level and construct classification models for the various level. Furthermore, hierarchical classification strategies are suggested as a solution to alleviate the problem of imbalanced classes.

Figure 40 showcases the overall methodology of the hierarchical model. Initially, the training dataset is split into three separate datasets, in a stratified fashion and then three different SVM models are trained. Each higher-level model provides its outputs to the rest of the lower-level ones (e.g., predictions from L_1 SVM are provided as input to the training of L_2 and L_3 SVM models, while predictions from L_2 are provided as input into the lowest level L_3). Finally, the L_3 SVM model is applied to the entire training dataset to acquire predictions on the crop type level.

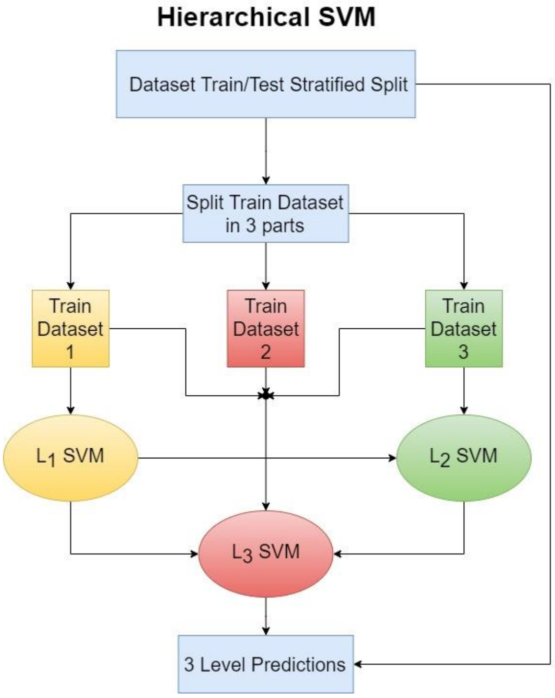


Figure 40: Hierarchical SVM classification scheme

As mentioned earlier, fallow land is a very abstract crop type, even for CAPO, that sometimes presents characteristic similar to arable lands or permanent trees. Additionally, it is the most frequent case in wrongly declarations-validation dataset (see figure 39), either as a wrongly declared fallow land or vice versa. This might be the reason why the hierarchical classification scheme performed very poorly in this class, and therefore were excluded from the procedure. To address this, an outlier detection module is produced, which discriminates fallow lands from the rest cases. Specifically, we train a binary SVM model, and then identify the most confident wrongly declared fallows. We introduce a user-defined parameter ρ, which refers to the expected ratio of wrongly declared fallow lands. Based on the OTSC from CAPO, we set this to 0.1. Eventually, these cases are flagged as “potentially not fallow”

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

and are added to the final crop type map. For the training of the binary SVM, we keep all fallow cases included on the training dataset together with a random sample of the rest crop types.

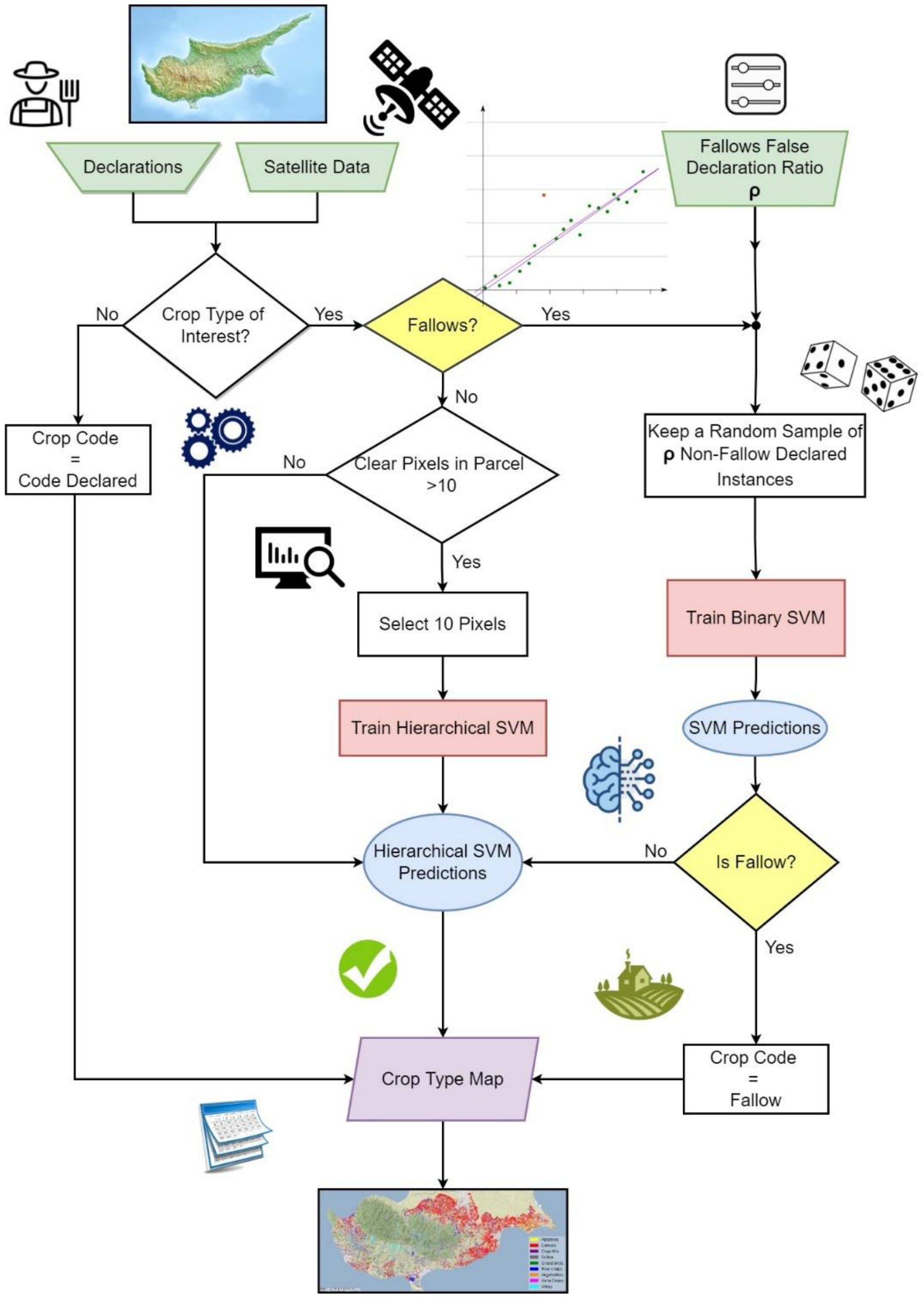


Figure 41: Crop Type Map Model for Cyprus

Smart Sampling for OTSC (traffic light system)

Similar to Lithuanian case, smart sampling will take benefit of the current total of the Crop Type Maps produced in order to identify the most probable wrong declarations. On this scenario again, will be based mainly on misclassification on the various taxonomy level of the predictions. However, we have not yet implemented the particular algorithm for the very special case of Cyprus. It is in our future

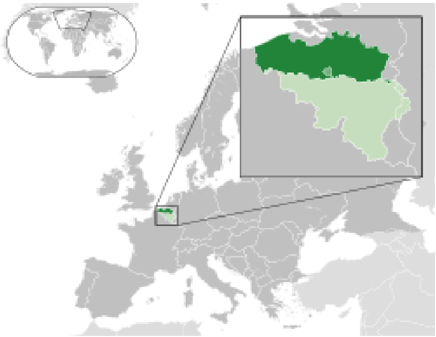
The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

plans to start evaluating the respective methodology here and form the algorithm parameters accordingly.

Crop Diversification Compliance Map

Similar to Lithuanian case, the Crop Diversification Service exploits the Land Parcel Information System (LPIS), the declarations of the farmers and a combination of ***if conditions*** based on the respective lookup table (see Table 8). Again, the crop codes to be used for the service are that of declarations, and they will change to the predicted ones from the CCTM module only if they have been indicated from the traffic light system as potentially wrongly declared. However, in order to follow the set of greening rules of BC2, we suggest to allocate the very specific code of *grasses or other herbaceous fodder plants* into the Cwater field category of the relative lookup imported table in order to identify the cases belonging into *Exemption B* scenario*,* mentioned in the Crop Diversification sub-section.

4.4 BC3: Monitoring the condition of soil – Belgium



DEPARTEMENT LANDBOUW & VISSERIJ

The flemish business case focuses on deploying ENVISION service for topsoil Soil Organic Carbon Monitoring in Flanders, Belgium. Currently, the state of agricultural soils is checked by performing soil samplings and conducting laboratory examinations. However, these methods do not provide a continuous overview of soils' state and require significant effort, time, and resources to be committed. Consequently, these types of controls have to be significantly reduced and replaced with a more automated process.

The business case is implemented in Belgium, within the Flemish region, involving LV, the Flemish Department of Agriculture and Fisheries and Paying Agency, which is in Flanders' the official PA in charge of the financial support for agriculture and the implementation of CAP. The Department of Agriculture and Fisheries is the Flemish Paying Agency and, together with the minister, outlines the policy on agriculture, horticulture, sea fishing and the countryside. The department implements this policy, monitors, and evaluates it. The Department is responsible for providing services to 35.850 farmers for 500.000 agricultural parcels that cover 680.000 ha. Every year ~25.000 On The Spot Controls are performed. The payment entitlements are close to 500.000 and € 245,30 Million.

EV ILVO will assist LV, a scientific institute specialized in service provision in all fields related to agriculture, fisheries, and food in Flanders. LV at various meetings has defined the service requirements (see D2.2 Report of customer requirements from ENVISION services). EV ILVO is responsible for developing the data products that can support the provision of the services. Additionally, in close collaboration with EV ILVO to perform the soil campaign and assess the soil organic carbon model performance. The parameters that control the soil organic carbon model performance are:

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

- The accuracy of the SOC model (validation set)
- The ability to deliver updates that captures the SOC changes (per year)
- The ability to deliver SOC for the majority of agricultural parcels in Flanders

All the parameters will define the service logic and answer critical questions. For example:

- Which is the optimal/suggested frequency for updates?
- How are we going to present the SOC values?
- How are we going to estimate, deliver and show the accuracy of the assessment?
- How can we accurately and without risk identify SOC degradation parcels?
- May we use the SOC monitoring to motivate or award Farmers?

4.4.1 Data product description.

Study site

The study area is the Flemish region (Figure 42), and the data products should cover at least the crop land and grassland areas (Figure 43).

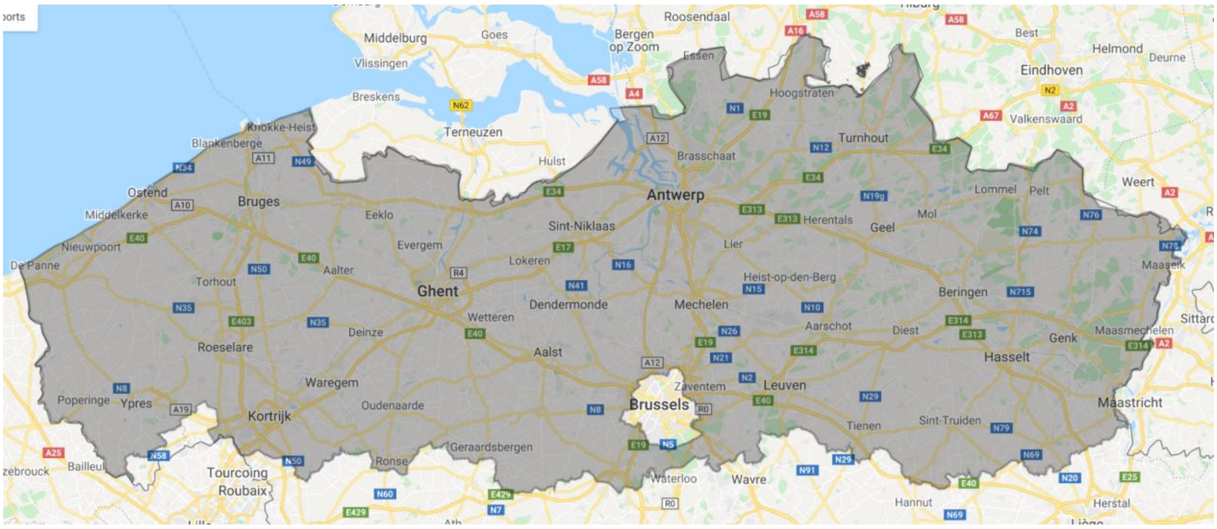
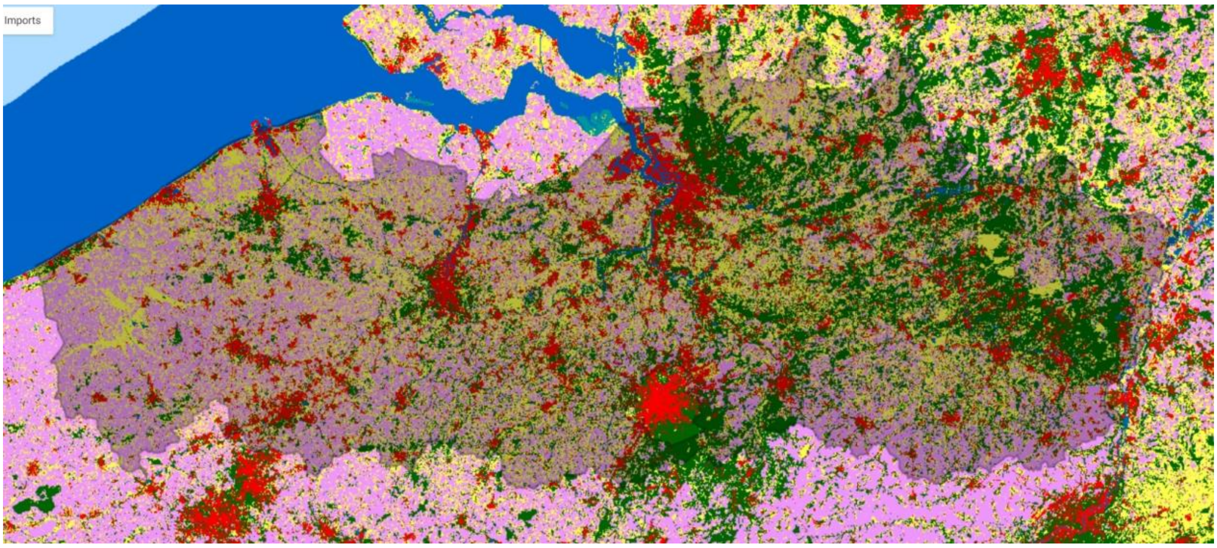


Figure 42: The study area covers 1368207 ha. Within the study are the agricultural parcels that cover 680.000 ha.



The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Figure 43: A land cover map of Flanders using the European Space Agency (ESA) WorldCover 10 m 2020 product provides a global land cover map for 2020 at 10 m resolution based on Sentinel-1 and Sentinel-2 data. The WorldCover product comes with 11 land cover classes and can support the identification of grassland (yellow), cropland (pink) and sparse vegetation (grey).

Cloud Coverage

A significant problem that characterizes the area and we need to overcome is widespread cloud coverage of the atmosphere. Especially in northern European counties like Belgium, it is widespread, substantial fractions of the sky obscured by clouds, resulting in very sparse image time series. The existence of clouds and cloud shadows are, without a doubt, a crucial problem in the acquisition process of optical imagery as they indefinitely alter the spectral signatures captured from satellite data. Sentinels 2 are susceptible to such phenomena since they are a multispectral constellation of satellites that acquires data in the visible, near-infrared, and short-wave infrared parts of the spectrum. This intervention may produce misleading results in analyses of import noise and may have dramatic consequences on the precision of agricultural monitoring. Tasks related to the identification of bare soil are sensitive to cloud appearance. Even if the revisit time of 5 days of S2 is not very rare, we have a gap of almost a month between two successive clear from cloud satellite images.

Error! Reference source not found. below presents the number of products with at least 90% cloud c overage compared to total products above Lithuania for the last two years, almost 1/3. It becomes apparent that in these cases, products generate missing values at pixel level that have to be identified and then dropped.

Table 9: Estimated products with at least 90% cloud coverage in Flanders for 2019 until 2021

Year	Number of products with at least 90% cloud coverage	Total Products
2019	475	1541
2020	397	1539
2021	505	1511

Soil conditions and data

To support the development of the SOC products, a soil sampling campaign was performed at the beginning of 2021. The soil sampling campaign collected samples trying to cover most of the SOC variability in croplands of the Flanders region. Therefore, the soil samples have been collected within the different soil regions insisting on agricultural parcels to ensure a large variability in soil types and SOC content. The SOC variability is necessary to build an effective prediction model to map SOC at a regional scale. For this purpose, we exploited the link between SOC content and spectral behaviour in the optical domain: the Sentinel-2 bands were used as feature space to determine where to collect samples by the Kennard – Stone algorithm. To ensure the proper quantity of soil samples for each soil type, we carried out a stratified feature-based approach for the sampling selection. The strata are 11 soil association regions based on the Soil Association Map of Flanders.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

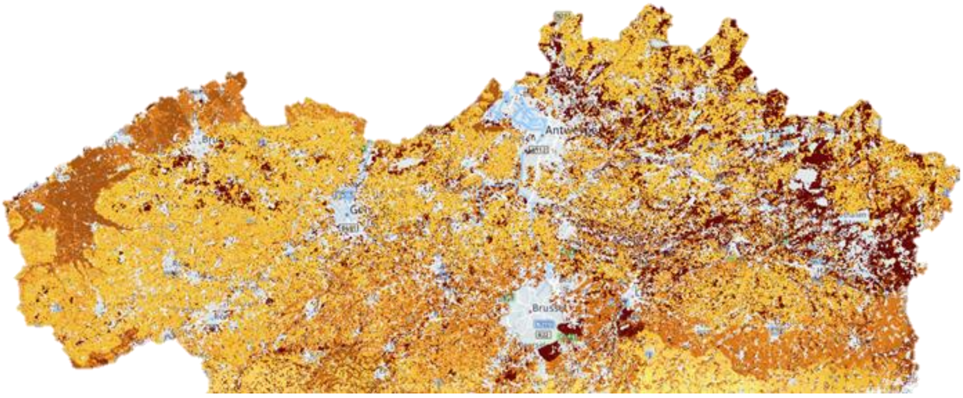


Figure 44: Existing Top Soil Organic Carbon stock map for topsoil[3] (0-30cm) with a mean 40m grid (10m for Flanders and 40m for Wallonia region).

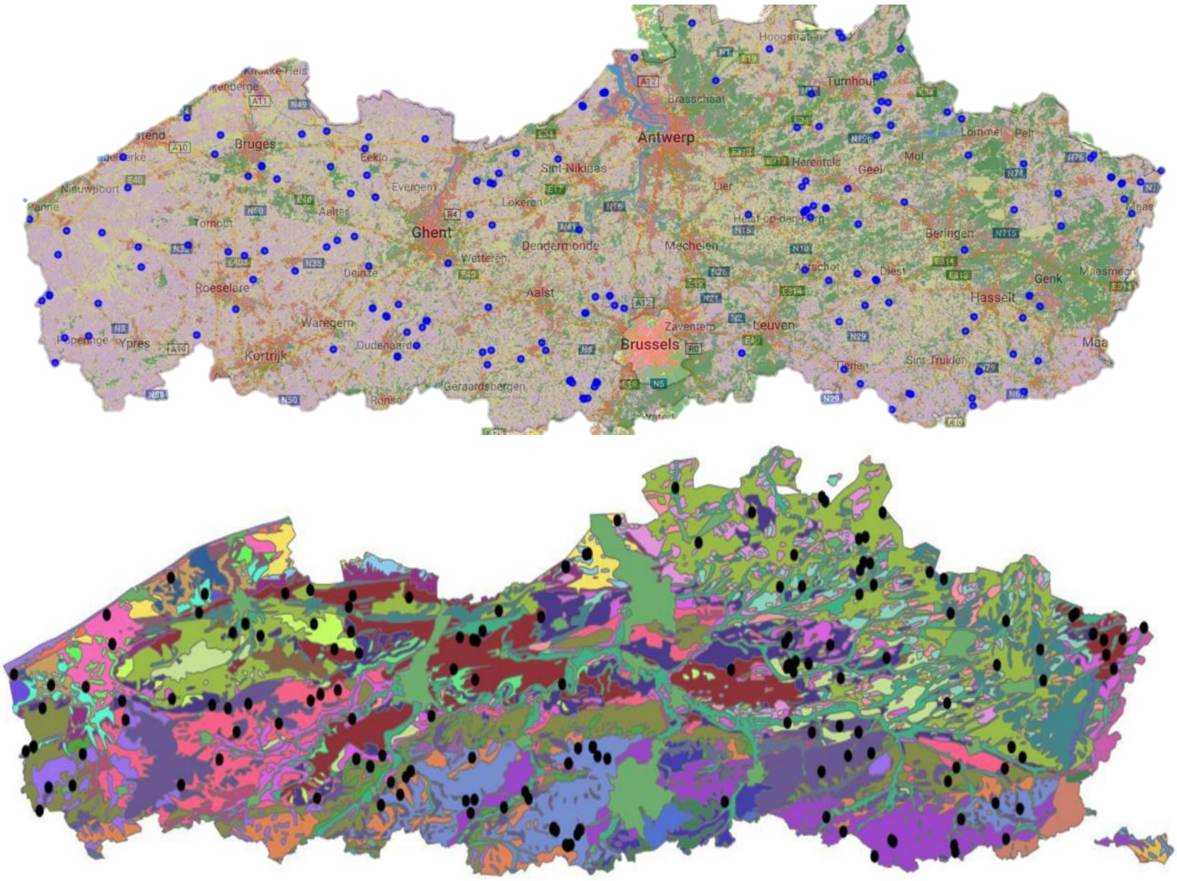
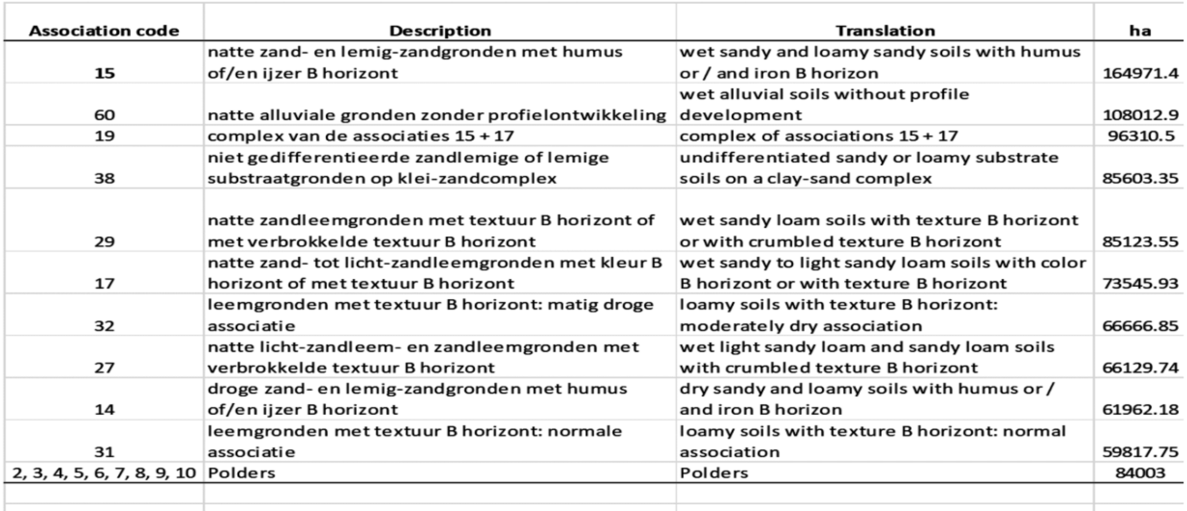


Figure 45: Location of the sampling points with background maps, the land classes (the upper map, the sampling points are with blue points) and soil associations in Flanders[4] (bottom map, the

[3] https://metadata.vlaanderen.be/srv/dut/catalog.search#/metadata/c41ae3d2-7975-457f-94d7-4ab83d659432

[4] Bodemassociaties dataset https://www.geopunt.be/catalogus/datasetfolder/c4f51b28-51bf-4189-8e98-72b94ae8da13

sampling points are with black points). A soil association is a substantive and spatial grouping of soil series.



Association code	Description	Translation	ha
15	natte zand- en lemig-zandgronden met humus of/en ijzer B horizont	wet sandy and loamy sandy soils with humus or / and iron B horizon	164971.4
60	natte alluviale gronden zonder profielontwikkeling	wet alluvial soils without profile development	108012.9
19	complex van de associaties 15 + 17	complex of associations 15 + 17	96310.5
38	niet gedifferentieerde zandlemige of lemige substraatgronden op klei-zandcomplex	undifferentiated sandy or loamy substrate soils on a clay-sand complex	85603.35
29	natte zandleemgronden met textuur B horizont of met verbrokkelde textuur B horizont	wet sandy loam soils with texture B horizont or with crumbled texture B horizont	85123.55
17	natte zand- tot licht-zandleemgronden met kleur B horizont of met textuur B horizont	wet sandy to light sandy loam soils with color B horizont or with texture B horizont	73545.93
32	leemgronden met textuur B horizont: matig droge associatie	loamy soils with texture B horizont: moderately dry association	66666.85
27	natte licht-zandleem- en zandleemgronden met verbrokkelde textuur B horizont	wet light sandy loam and sandy loam soils with crumbled texture B horizont	66129.74
14	droge zand- en lemig-zandgronden met humus of/en ijzer B horizont	dry sandy and loamy sandy soils with humus or / and iron B horizon	61962.18
31	leemgronden met textuur B horizont: normale associatie	loamy soils with texture B horizont: normal association	59817.75
2, 3, 4, 5, 6, 7, 8, 9, 10	Polders	Polders	84003

Figure 46: Area info per soil association in the Flemish Region

The soil organic carbon content (SOC) of the soil samples is displayed in the scatter plot below. The SOC in the dataset ranged from 0.29% to 12.40% (not shown in the figure). 88% of the samples contained a SOC content between 0,5 and 2,0 %.

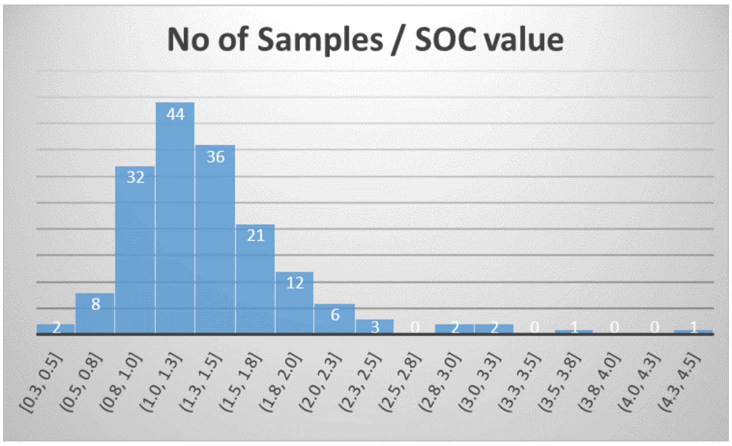


Figure 47: No of samples (y-axis) and the estimated SOC value (x-axis). From the 171 samples, the majority takes SOC values between 0.8 - 1.8 (%/dry soil).

Analytics on various indices and bare soil identification

To develop the Top Soil Organic Carbon products, we follow a methodology that develops a cloudless bare soil composite collection. A bare soil composite is an extensive collection of multispectral satellite data that can be used to map topsoil attributes to a large extent [5]. A composite collection can represent the reflectance of bare fields only if it consists of bare soil pixels. To identify and select the bare soil pixels, first, a set of indices needs to be generated, perform analytics to estimate the upper and down limits and use those limits to mask. The indices need to detect green and dry vegetation and high soil moisture content that can affect the soil spectrum shape and other existing S2-L2 bands as masks.

[5] J.L. Safanelli "Multispectral Models from Bare Soil Composites for Mapping Topsoil Properties over Europe" https://www.mdpi.com/2072-4292/12/9/1369/htm

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

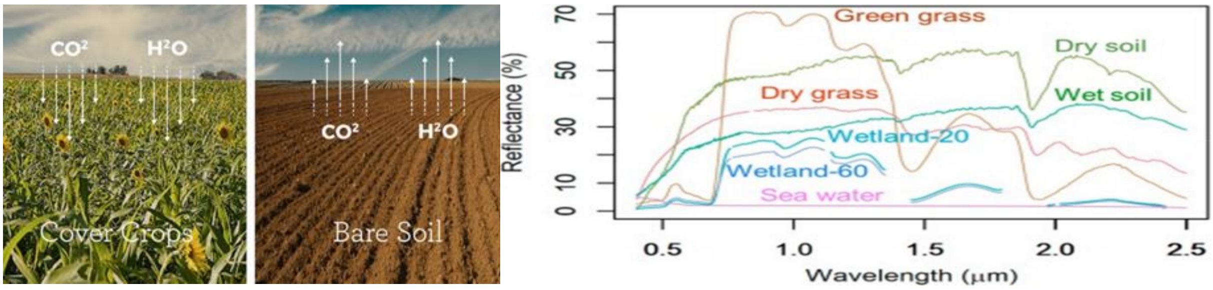
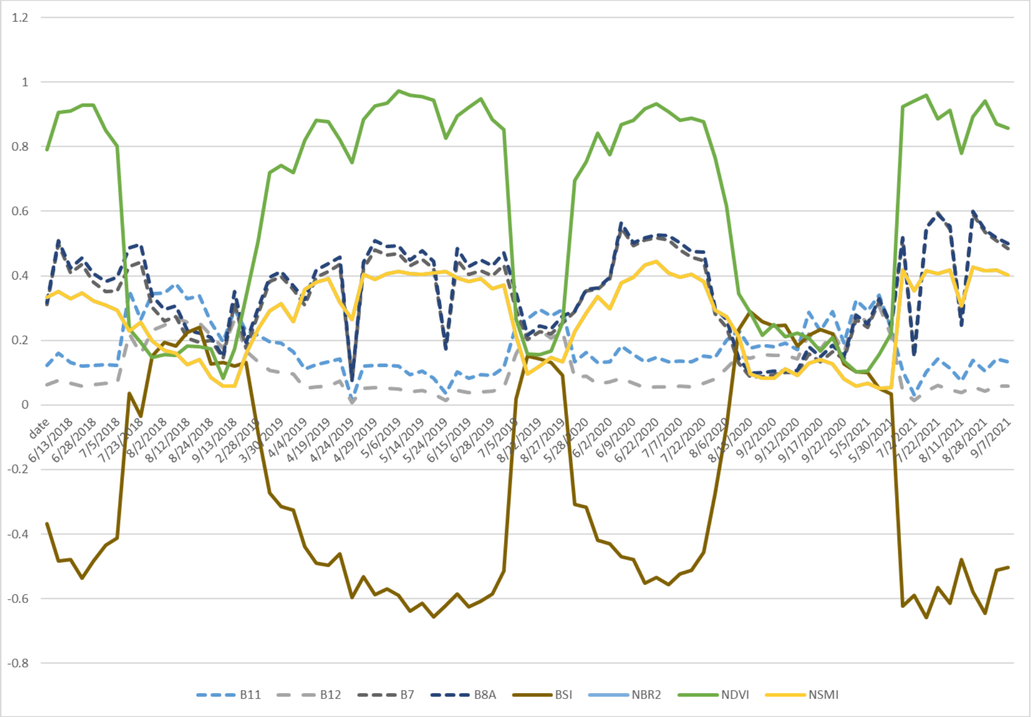


Figure 48: The identification of bare soil pixels in extensive image collection is a significant task supported by vegetation, bare soil and soil moisture indices. Sentinel 2 bands in NIR and SWIR can support the identification of Dry and Wet Soil.

By using our technological tools, it’s possible to generate a large collection of Sentinel 2 images and generate the needed indices for all or just for selected pixels.



The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

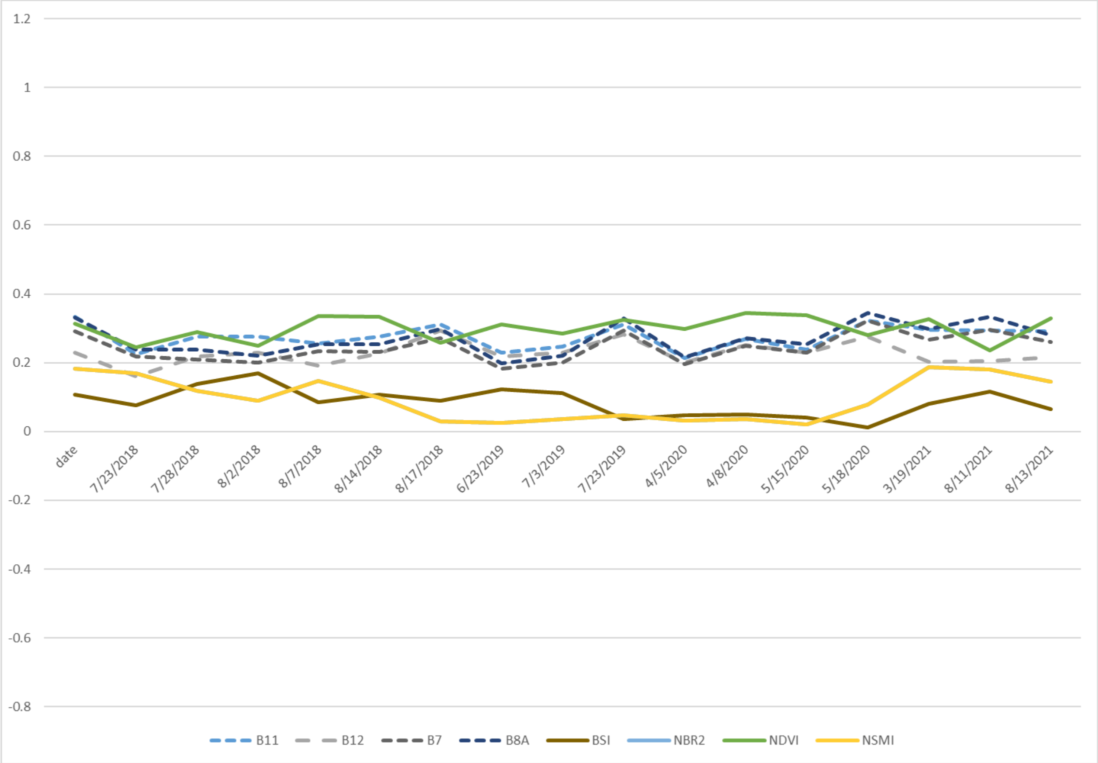
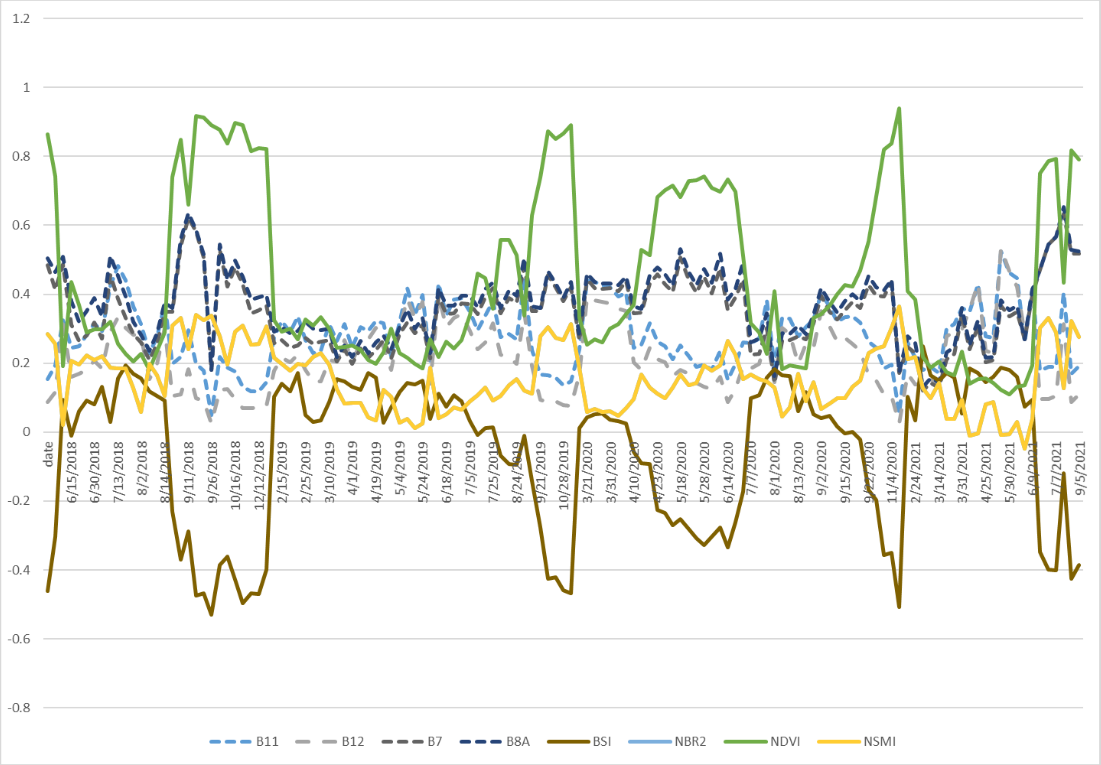


Figure 49: Time series of S2 reflection bands together with bare soil, soil moisture and vegetation indices for **sampling point 12** of the soil campaign for 2018 and 2021 (upper). After applying an NDVI filter of <0.35 reduces the times series points, generating significant time gaps (from a few weeks to a few months).



The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

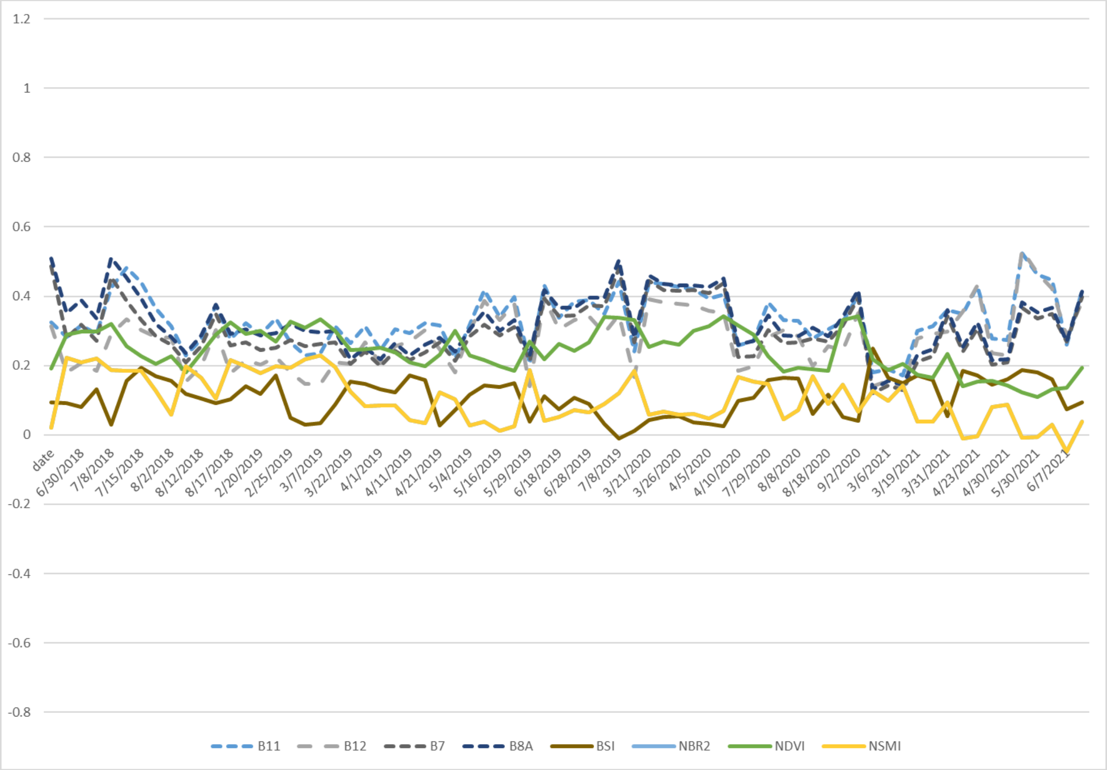
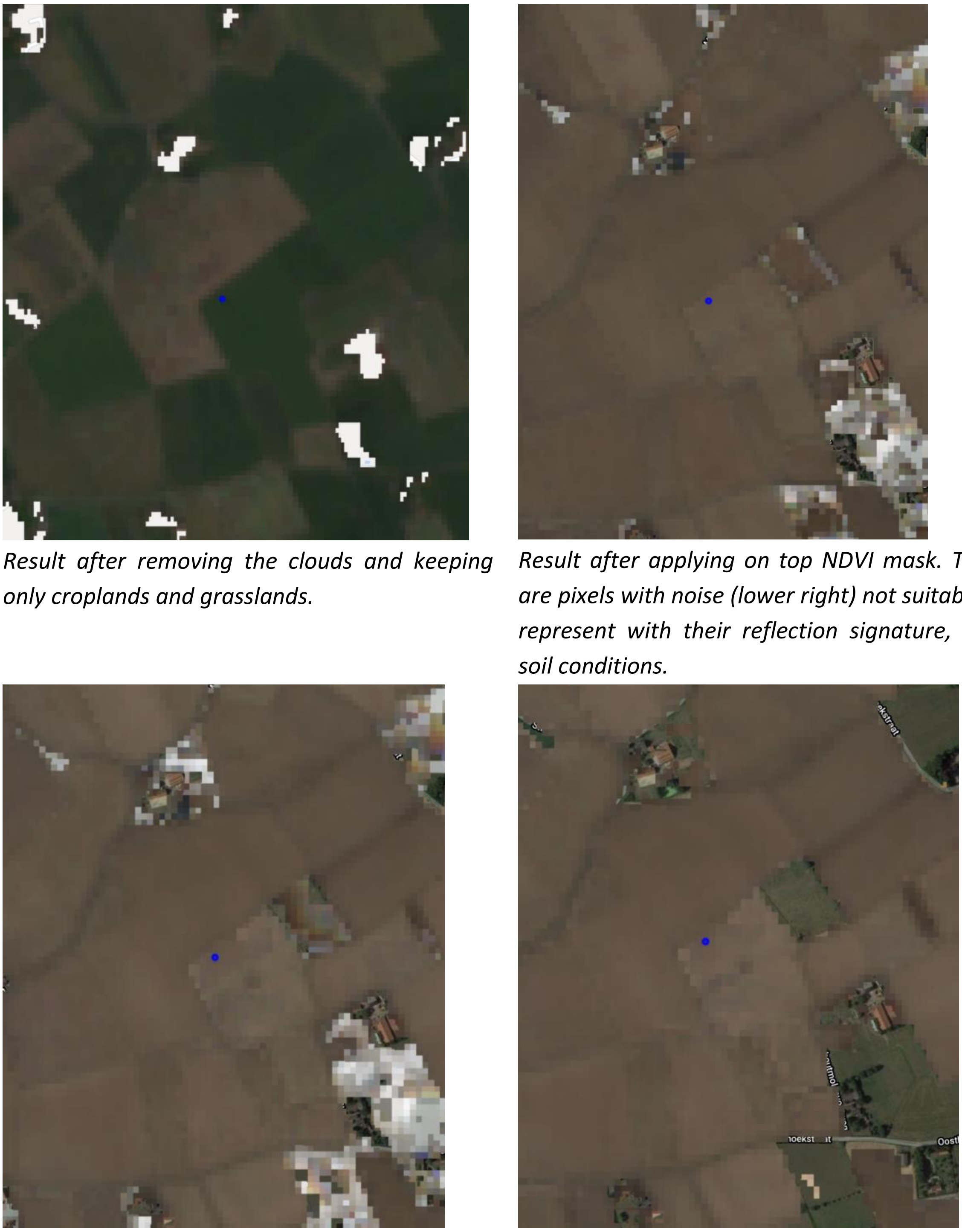


Figure 50: Time series of S2 reflection bands together with bare soil, soil moisture and vegetation indices for **sampling point 2** of the soil campaign for 2018 and 2021 (upper). By applying an NDVI filter of <0.35, we reduce the times series points (from 124 to 70, almost 45%), generating significant time gaps (from a few weeks to a few months).

The ENVISION project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 869366

In this process, it is necessary first to deal with the cloud issue and apply masking techniques at a pixel level to ensure that the indices and the reflections correspond to cloudless pixels.



Result after removing the clouds and keeping only croplands and grasslands.

Result after applying on top NDVI mask. There are pixels with noise (lower right) not suitable to represent with their reflection signature, bare soil conditions.

Result after applying, on top of the NDVI mask, the NBR2 mask.

Finally result, after applying, on top of the others, the VNSIR mask. Successfully remove the problematic pixels.

Figure 51: RGB visualization of the synthetic composite (period May 2018 until the end of 2021) using the median function. The blue spot represents a sampling point of the soil campaign.

SOC products

To develop the Top Soil Organic Carbon products, we follow a methodology that relies scientifically on developing a cloudless bare soil composite collection covering the Flemish region from 2018 until 2021 (Figure 52). The data products are produced in Phase 3, at which we apply the SOC models of Phase 2 to the Cloudless Bare Soil Collection of Phase One. In Phase 4, we validate the products and evaluate them together with the SOC service.

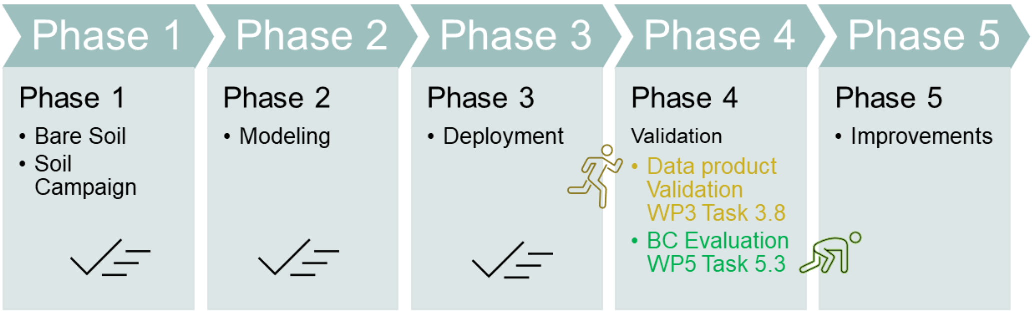


Figure 52: Data product development phases. After deploying the data products to the Envision platform, starts the validation phase, which includes the Data product Validation (running) and the BC

The BC3 data products of Phase 3 are:

- A raster file with pixel spatial resolution of 20 m by 20 m contains top-soil Soil Organic Carbon estimations (% of SOC). This file is presented in the Envision Platform as a background layer map (Figure 53).

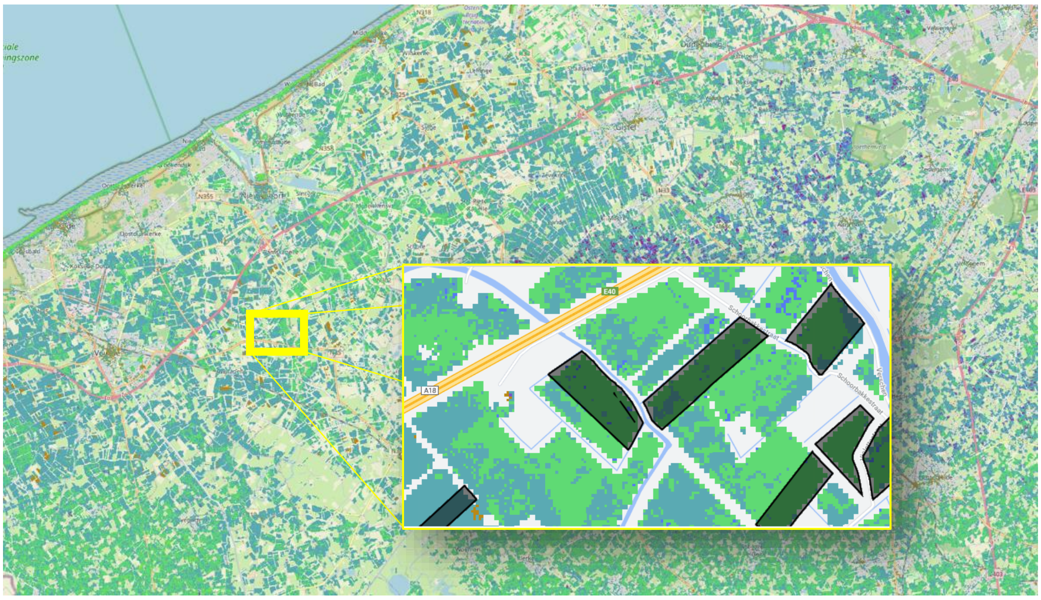


Figure 53: SOC map covering West Flanders. Zoom window overlays a sample of agricultural parcels.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

- The metadata of the raster file provides the accuracy of the SOC modelling by using the RMSE (Root Mean Square Errors)[6] together with the sample point locations, the lab measurements results and the methodology/protocol we have followed to collect the sample data and perform the lab measurements.
- A vector file with the Flemish LPIS agricultural parcels that has the average value of the top-soil Soil Organic Carbon within each parcel (Figure 54). The parcels aggregate the Top Soil Organic Carbon Information using the average value, including pixels coming from the raster file. The vector file will be published in the Envision platform as a vector layer.

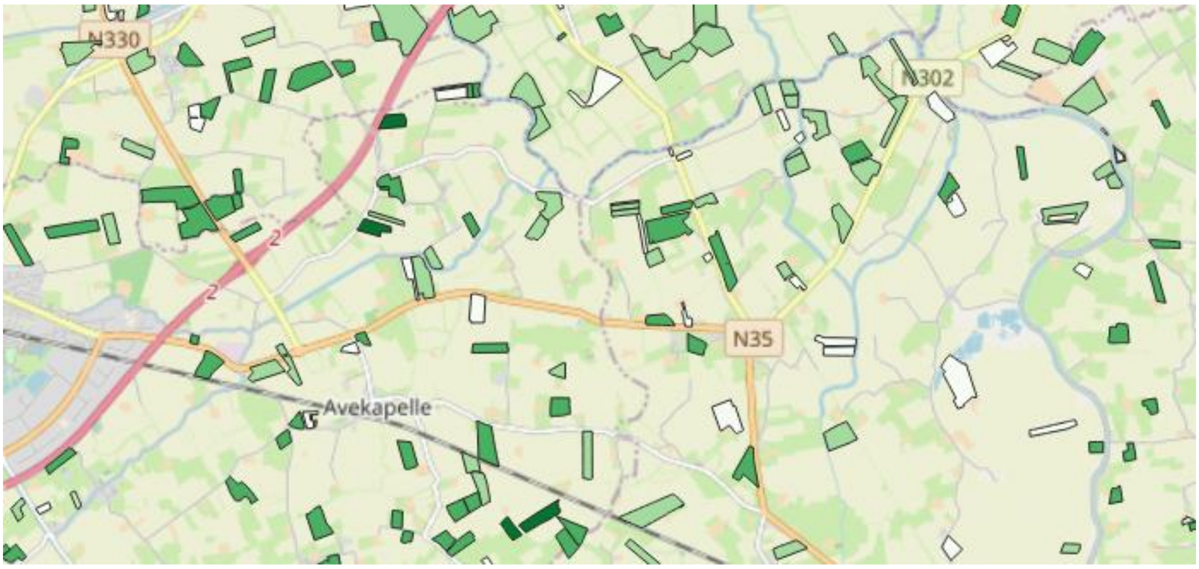


Figure 54: Aggregated SOC information at the parcel level.

All products are delivered to a repository of the Envision platform and are visualized at UI using mapping services developed by AgroApps.

[6] Expected for the calibration RMSEC, cross-validation RMSECV and prediction set RMSEP. RPD and R2 are also used to evaluate the accurancy of the model.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

4.4.2 Methodology

To achieve user requirements and other non-functional requirements related to service scalability, we define a methodology that can enable current scientific research outcomes and deliver soil organic carbon products on a large scale. This section will describe the development phases, together with the technological tools we use at each phase (Figure 57).

4.4.2.1 Development phases

To develop the Top Soil Organic Carbon products, we follow a methodology that develops a cloudless bare soil composite collection of Sentinel 2-L2 images. We have applied this methodological approach using five major phases as described in Figure 56 to ensure the needed agility on product development.

4.4.2.1.1 Phase One: Bare Soil Identification

In **Phase One**, the main goal is to develop a Cloudless Collection of Bare Soil Pixels. The collection consists of 4598 images of L2A that covered the Flemish region from 2018 until the end of April 2021. The first step was to create a collection of S2 images using the GEE Python APIs to access Data Catalogue products (Sentinel 2 MSI, Level 2A) following the latest Legal notice on the use of Copernicus Sentinel Data and Service Information.

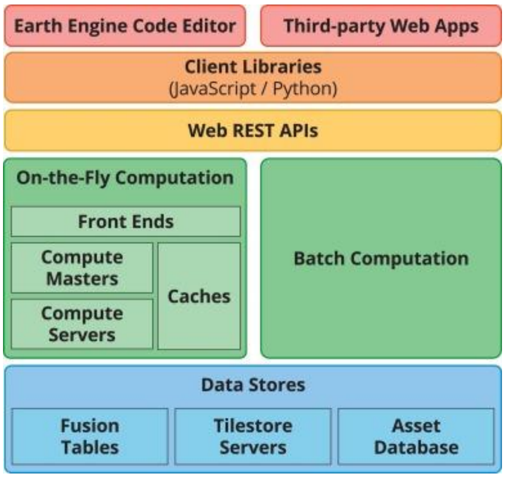


Figure 55: Google Earth Engine (GEE) system architecture diagram. GEE relies on a Client/server programming model.

The next step was to apply cloud masking, and for that, we make use of the:

- MSK_CLDPRB 20 meters Cloud Probability Map (missing in some products)
- MSK_SNWPRB 10 meters Snow Probability Map (missing in some products)
- QA60 60m Cloud mask [7]

After applying, calculate vegetation and moisture indices that can detect green and dry vegetation and high soil moisture content that can affect the soil spectrum shape and other existing S2-L2 bands. We use these indices to mask the collection layer, testing different upper and low threshold values. We also filter the collection layer by using the ESA Worldwide land cover mapping. Alternative we can use the parcels of the Land Parcel Identification System (LPIS) provided by LV, in the masking process, however that requires more computational power and the ESA Worldwide land cover mapping covers

[7] https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-1c/cloud-masks

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

equal the existing crop lands and the grasslands. In Table 10, we provide the indices, formulas, and final threshold values.

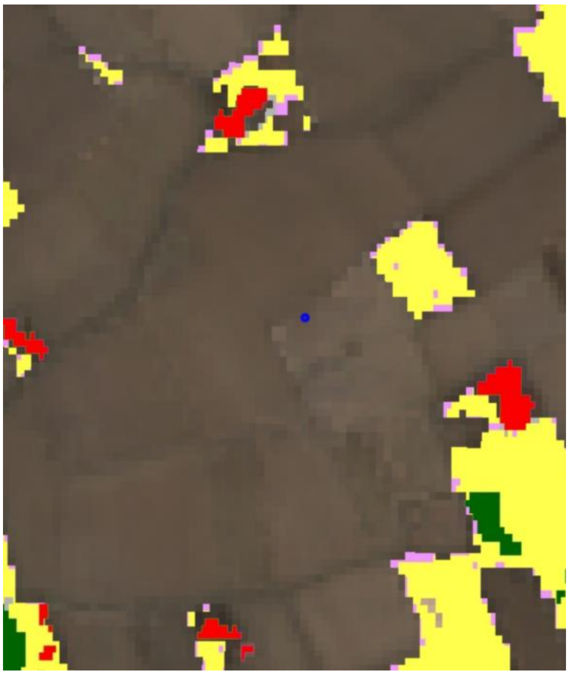


Figure 56: The masking works very well with croplands; however, most of the grasslands areas (yellow) do not belong to the Cloudless bare soil collection. The NDVI values remain high during the whole period, which means it's impossible to receive bare soil reflections.

Table 10: To identify the bare soil layer, we created and applied a set of extra masks using the NDVI, VNSIR and NBR2 indices.

Indices	Formulas	Upper and Down thresholds suitable to identify for Bare Soil
NDVI	(B08-B04)/(B08+B04)	>-0.25 and <0.35
NBR2	(B11-B12)/(B11+B12)	>0 and <0.1
VNSIR	(2 * RED) - GREEN - BLUE) + (3 *(SWIR2 - NIR)	>0.1

The output of this process is a Cloudless Bare Soil Collection covering the Flemish croplands in each soil association region.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

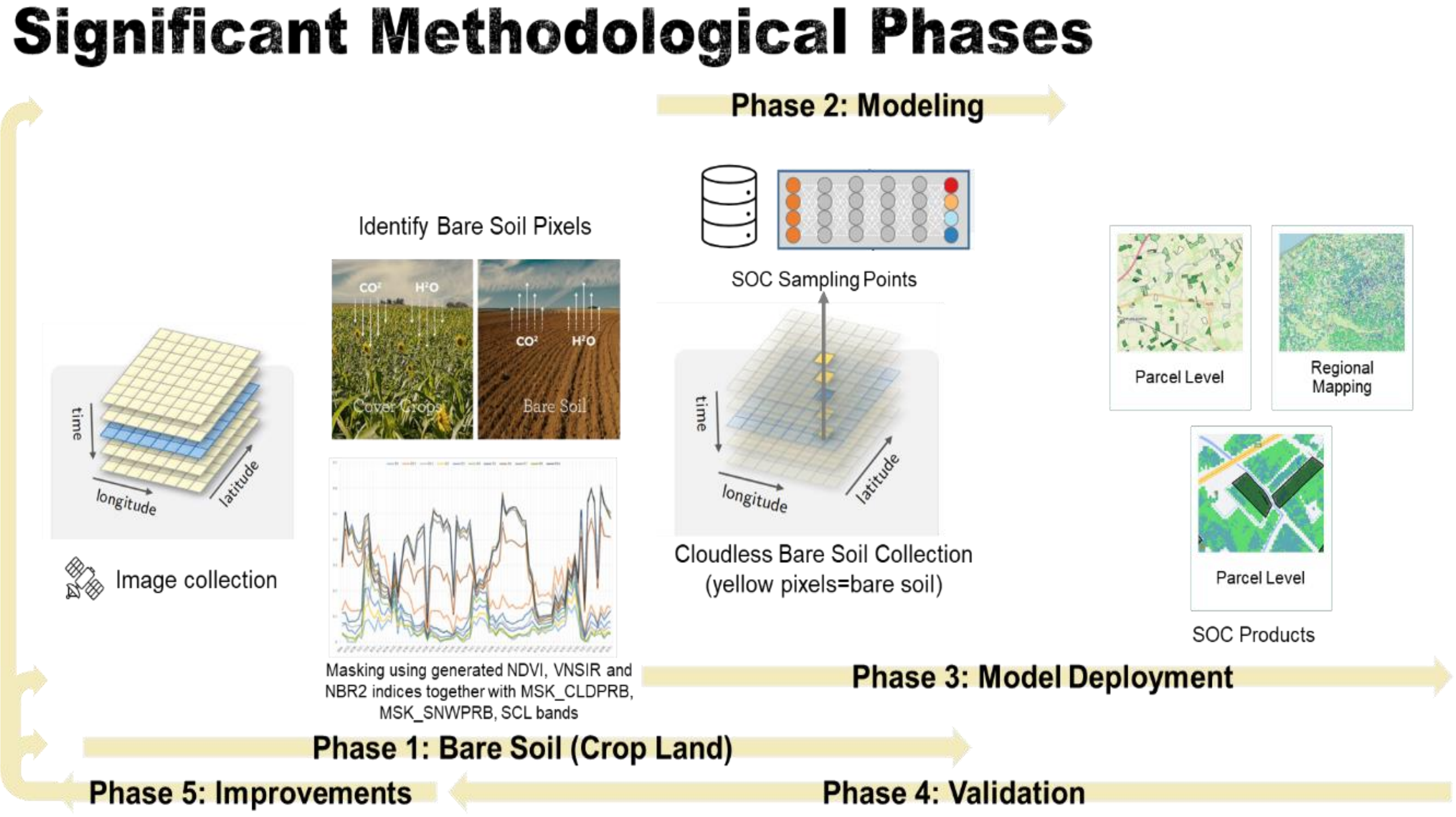
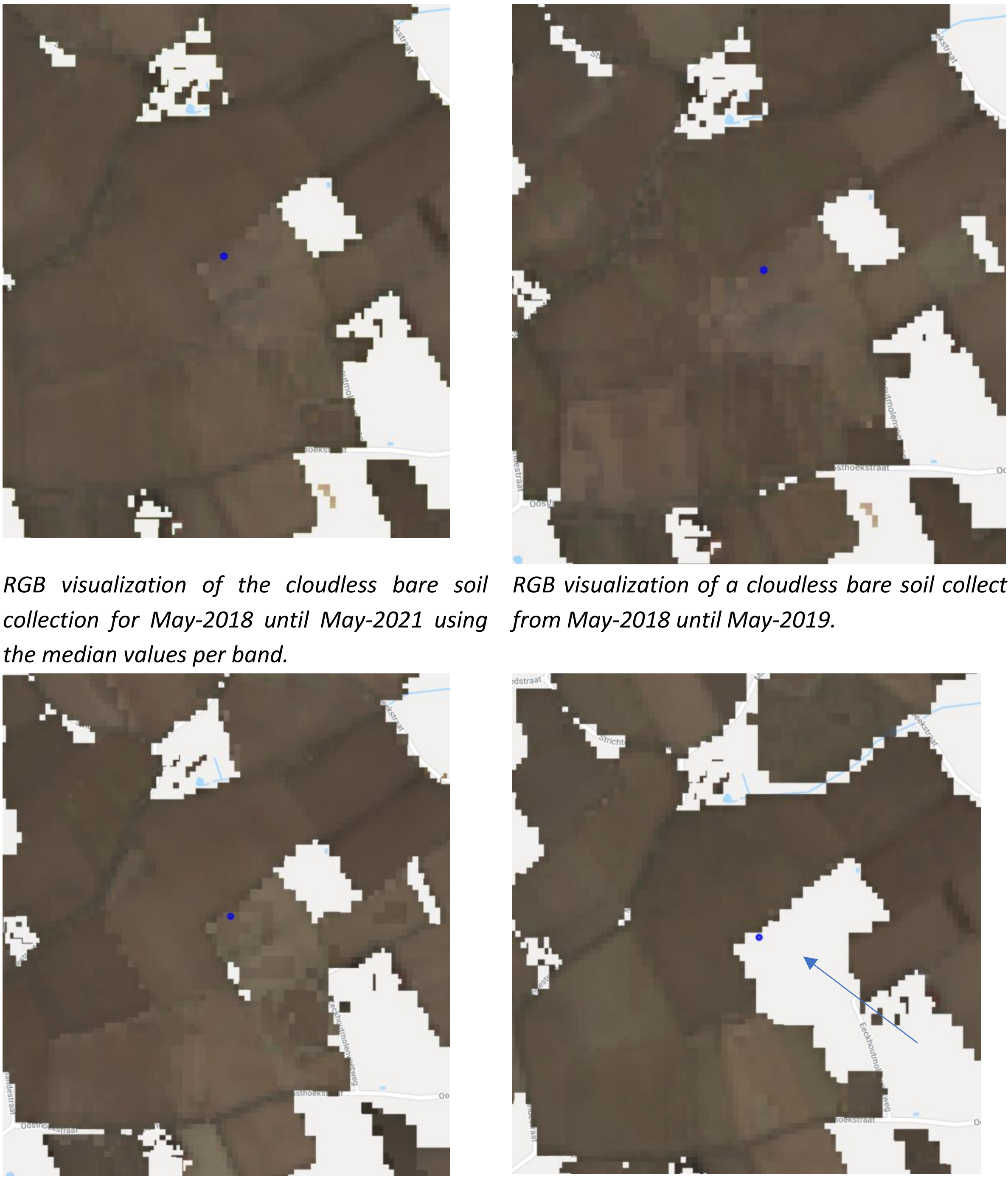

Significant Methodological Phases
Phase 2: Modeling
Identify Bare Soil Pixels
CO²
H²O
Cover Crops
Bare Soil
SOC Sampling Points
time
longitude
latitude
Image collection
Masking using generated NDVI, VNSIR and NBR2 indices together with MSK_CLDPRB, MSK_SNWPRB, SCL bands
Cloudless Bare Soil Collection (yellow pixels=bare soil)
Parcel Level
Regional Mapping
Parcel Level
SOC Products
Phase 3: Model Deployment
Phase 1: Bare Soil (Crop Land)
Phase 5: Improvements
Phase 4: Validation

Figure 57: Basic flow for the development of the synthetic bare soil layer

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366



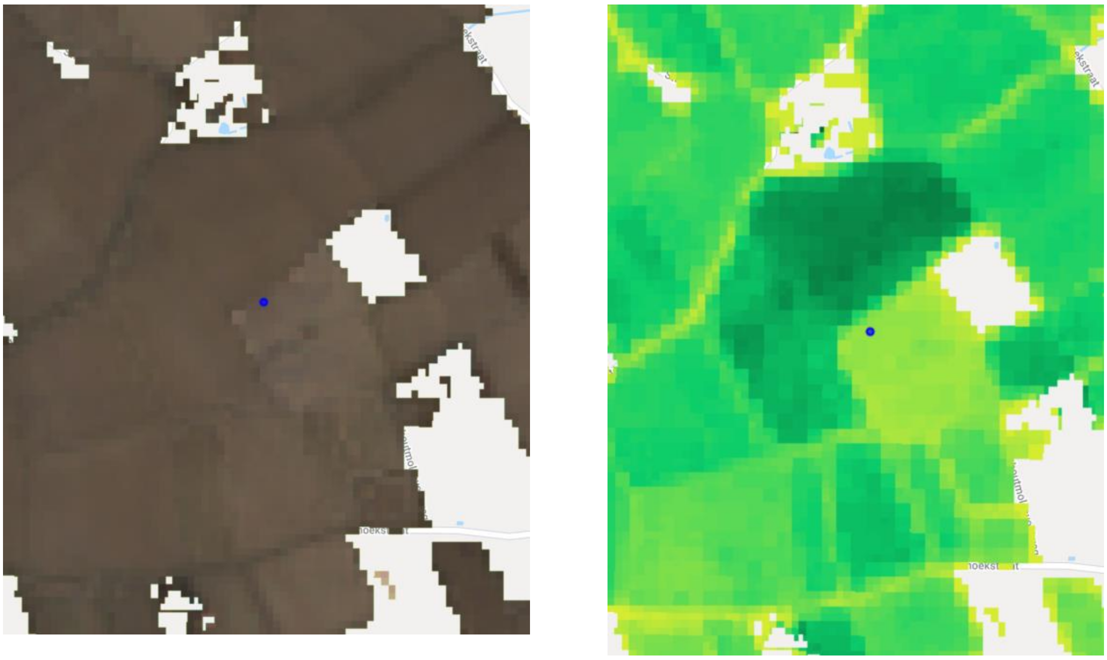
RGB visualization of the cloudless bare soil collection for May-2018 until May-2021 using the median values per band.

RGB visualization of a cloudless bare soil collection from May-2018 until May-2019.

RGB visualization of a cloudless bare soil collection from May-2019 until May-2020.

RGB visualization of a cloudless bare soil collection from May-2020 until May-2021. Mainly due to clouds, the cloudless bare soil collection does not cover the sampling point area.

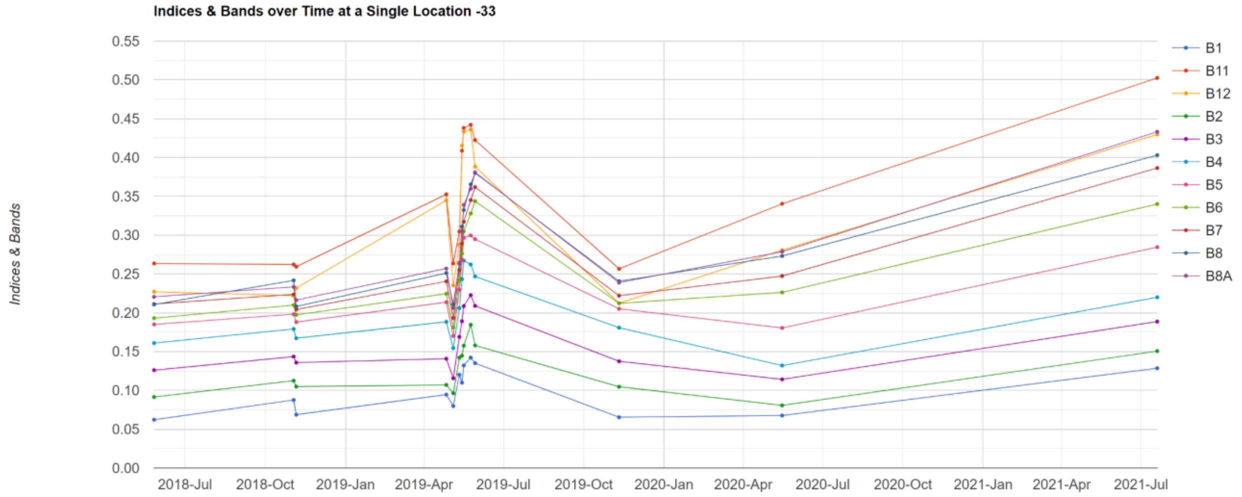
Figure 58: RGB visualization of continuous-time period stacks of the cloudless bare soil collection area around a soil sampling collection point (point ID 33).



RGB visualization of the cloudless bare soil collection for May-2018 until May-2021 using the median values per band. From the lab measurements, the sampling point has SOC 0.72%, which considering very low. The area around the sampling point has light brown colour, which corresponds to lower SOC levels.

Visualization of the number of images per pixel area. As presented in Figure 78, clouds existence or vegetation conditions varies in time and from one period to another period, some pixel areas are not within a bare collection. Each pixel area of the cloudless bare soil collection from May-2018 until May-2021 consists of a number of pixels. The number of pixels per pixel area, difference per pixel area. The sampling point has 12 pixels (light green). Dark green pixels have 50 or more pixels.

Figure 59: In phase one, we develop functions within the scripting code to support the assessment and visualization of various parameters of the cloudless soil collection. One parameter is the **number of pixels within the cloudless bare soil collection per pixel area**. The number of pixels per pixel area can be used as an indicator of trustworthiness if the median values are used in the modelling process (Phase 2).



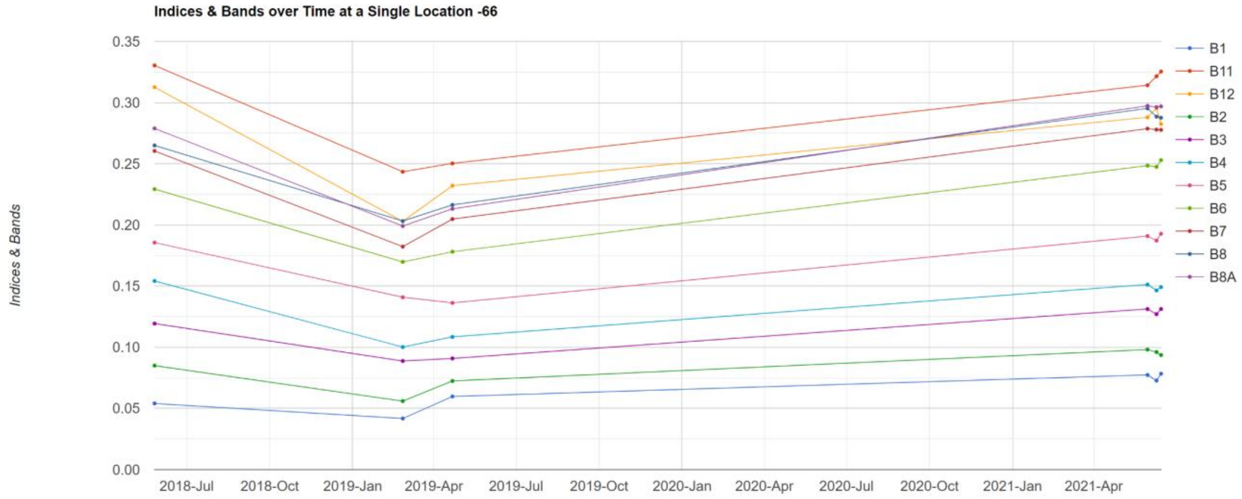


Figure 60: Location 66 has a measure SOC of 1.82%, much higher than location 33 (0.72%). Graphs present the reflection per S2 band for the period of May- 2018 until 2021.

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

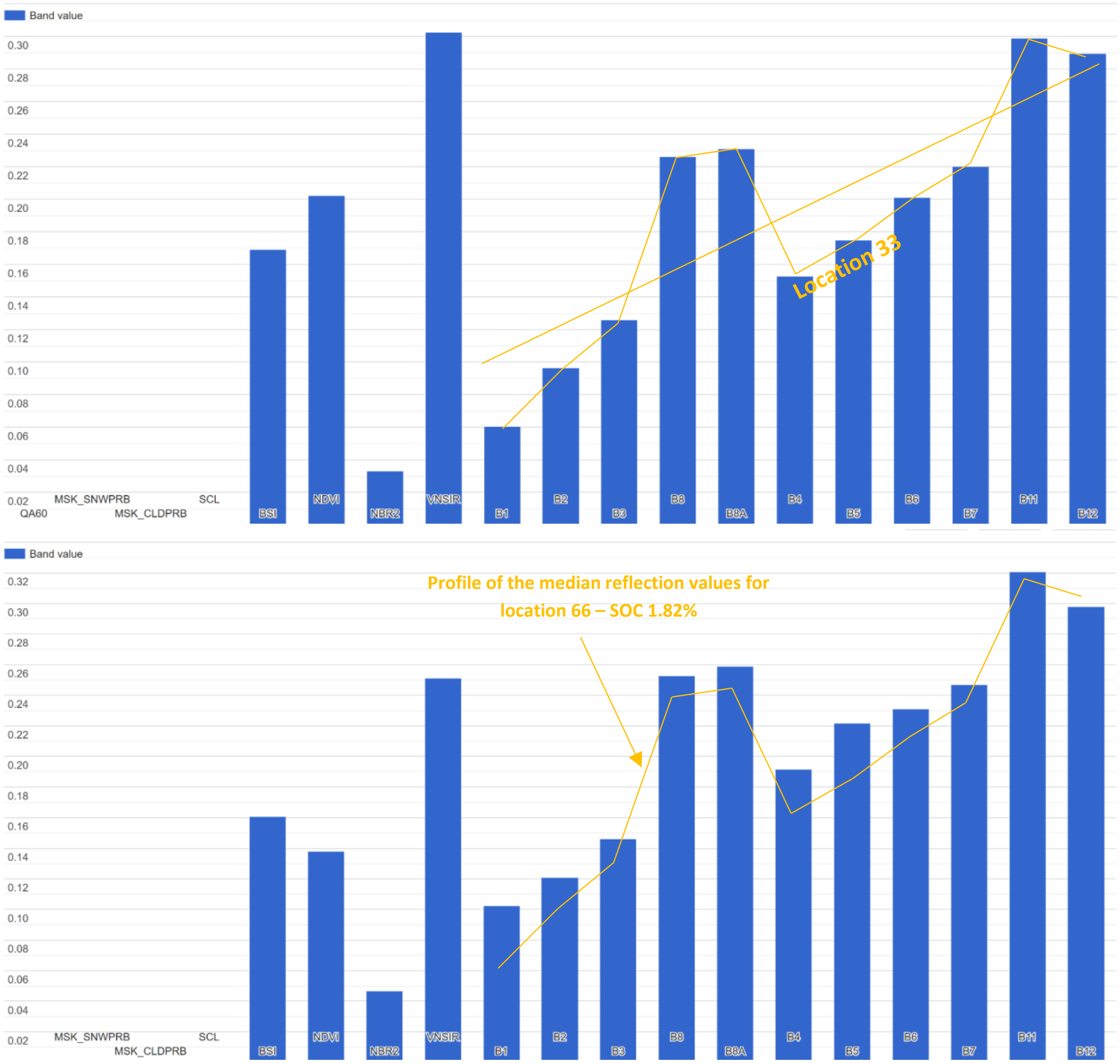
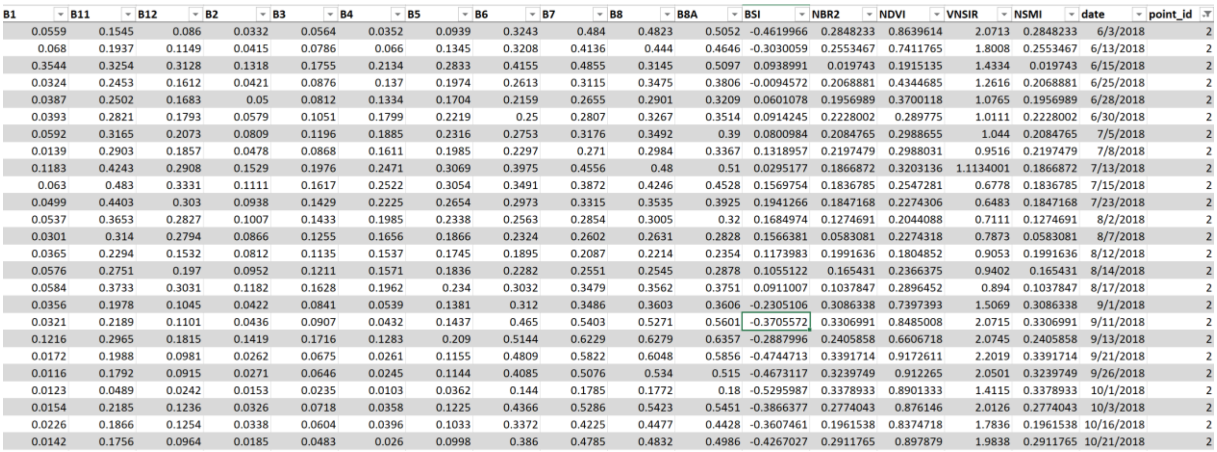


Figure 61: Median reflection values per band for May 2018 until May 2021, for the sampling point 66 (upper) and 33 (down). Location 33 has a measure SOC of 0.72%.

4.4.2.1.2 Phase Two: Modelling

In **Phase Two**, we perform the modelling and there, the goal is to create a mapping between the extracted reflection signatures coming from the Bare Soil Collection and the SOC measurements. Using the GEE Reducers[8] and the Exporting Data[9] ability, we generate raw data per sampling point (Figure 62). It's possible to automatically extract the completed value set without applying masking using the vegetation and moisture indices (by activating only the cloud mask function), supporting the data analysis without using the median values per band and sampling point.



B1	B11	B12	B2	B3	B4	B5	B6	B7	B8	B8A	BSI	NBR2	NDVI	VNSIR	NSMI	date	point_id
0.0559	0.1545	0.086	0.0332	0.0564	0.0352	0.0939	0.3243	0.484	0.4823	0.5052	-0.4619966	0.2848233	0.8639614	2.0713	0.2848233	6/3/2018	2
0.068	0.1937	0.1149	0.0415	0.0786	0.066	0.1345	0.3208	0.4136	0.444	0.4646	-0.3030059	0.2553467	0.7411765	1.8008	0.2553467	6/13/2018	2
0.3544	0.3254	0.3128	0.1318	0.1755	0.2134	0.2833	0.4155	0.4855	0.3145	0.5097	0.0938991	0.019743	0.1915135	1.4334	0.019743	6/15/2018	2
0.0324	0.2453	0.1612	0.0421	0.0876	0.137	0.1974	0.2613	0.3115	0.3475	0.3806	-0.0094572	0.2068881	0.4344685	1.2616	0.2068881	6/25/2018	2
0.0387	0.2502	0.1683	0.05	0.0812	0.1334	0.1704	0.2159	0.2655	0.2901	0.3209	0.0601078	0.1956989	0.3700118	1.0765	0.1956989	6/28/2018	2
0.0393	0.2821	0.1793	0.0579	0.1051	0.1799	0.2219	0.25	0.2807	0.3267	0.3514	0.0914245	0.2228002	0.289775	1.0111	0.2228002	6/30/2018	2
0.0592	0.3165	0.2073	0.0809	0.1196	0.1885	0.2316	0.2753	0.3176	0.3492	0.39	0.0800984	0.2084765	0.2988655	1.044	0.2084765	7/5/2018	2
0.0139	0.2903	0.1857	0.0478	0.0868	0.1611	0.1985	0.2297	0.271	0.2984	0.3367	0.1318957	0.2197479	0.2988031	0.9516	0.2197479	7/8/2018	2
0.1183	0.4243	0.2908	0.1529	0.1976	0.2471	0.3069	0.3975	0.4556	0.48	0.51	0.0295177	0.1866872	0.3203136	1.1134001	0.1866872	7/13/2018	2
0.063	0.483	0.3331	0.1111	0.1617	0.2522	0.3054	0.3491	0.3872	0.4246	0.4528	0.1569754	0.1836785	0.2547281	0.6778	0.1836785	7/15/2018	2
0.0499	0.4403	0.303	0.0938	0.1429	0.2225	0.2654	0.2973	0.3315	0.3535	0.3925	0.1941266	0.1847168	0.2274306	0.6483	0.1847168	7/23/2018	2
0.0537	0.3653	0.2827	0.1007	0.1433	0.1985	0.2338	0.2563	0.2854	0.3005	0.32	0.1684974	0.1274691	0.2044088	0.7111	0.1274691	8/2/2018	2
0.0301	0.314	0.2794	0.0866	0.1255	0.1656	0.1866	0.2324	0.2602	0.2631	0.2828	0.1566381	0.0583081	0.2274318	0.7873	0.0583081	8/7/2018	2
0.0365	0.2294	0.1532	0.0812	0.1135	0.1537	0.1745	0.1895	0.2087	0.2214	0.2354	0.1173983	0.1991636	0.1804852	0.9053	0.1991636	8/12/2018	2
0.0576	0.2751	0.197	0.0952	0.1211	0.1571	0.1836	0.2282	0.2551	0.2545	0.2878	0.1055122	0.165431	0.2366375	0.9402	0.165431	8/14/2018	2
0.0584	0.3733	0.3031	0.1182	0.1628	0.1962	0.234	0.3032	0.3479	0.3562	0.3751	0.0911007	0.1037847	0.2896452	0.894	0.1037847	8/17/2018	2
0.0356	0.1978	0.1045	0.0422	0.0841	0.0539	0.1381	0.312	0.3486	0.3603	0.3606	-0.2305106	0.3086338	0.7397393	1.5069	0.3086338	9/1/2018	2
0.0321	0.2189	0.1101	0.0436	0.0907	0.0432	0.1437	0.465	0.5403	0.5271	0.5601	-0.3705572	0.3306991	0.8485008	2.0715	0.3306991	9/11/2018	2
0.1216	0.2965	0.1815	0.1419	0.1716	0.1283	0.209	0.5144	0.6229	0.6279	0.6357	-0.2887996	0.2405858	0.6606718	2.0745	0.2405858	9/13/2018	2
0.0172	0.1988	0.0981	0.0262	0.0675	0.0261	0.1155	0.4809	0.5822	0.6048	0.5856	-0.4744713	0.3391714	0.9172611	2.2019	0.3391714	9/21/2018	2
0.0116	0.1792	0.0915	0.0271	0.0646	0.0245	0.1144	0.4085	0.5076	0.534	0.515	-0.4673117	0.3239749	0.912265	2.0501	0.3239749	9/26/2018	2
0.0123	0.0489	0.0242	0.0153	0.0235	0.0103	0.0362	0.144	0.1785	0.1772	0.18	-0.5295987	0.3378933	0.8901333	1.4115	0.3378933	10/1/2018	2
0.0154	0.2185	0.1236	0.0326	0.0718	0.0358	0.1225	0.4366	0.5286	0.5423	0.5451	-0.3866377	0.2774043	0.876146	2.0126	0.2774043	10/3/2018	2
0.0226	0.1866	0.1254	0.0338	0.0604	0.0396	0.1033	0.3372	0.4225	0.4477	0.4428	-0.3607461	0.1961538	0.8374718	1.7836	0.1961538	10/16/2018	2
0.0142	0.1756	0.0964	0.0185	0.0483	0.026	0.0998	0.386	0.4785	0.4832	0.4986	-0.4267027	0.2911765	0.897879	1.9838	0.2911765	10/21/2018	2

Figure 62: Reflection values per Sentinel 2 band, together with the computed indices and the image data. Sampling point 2.

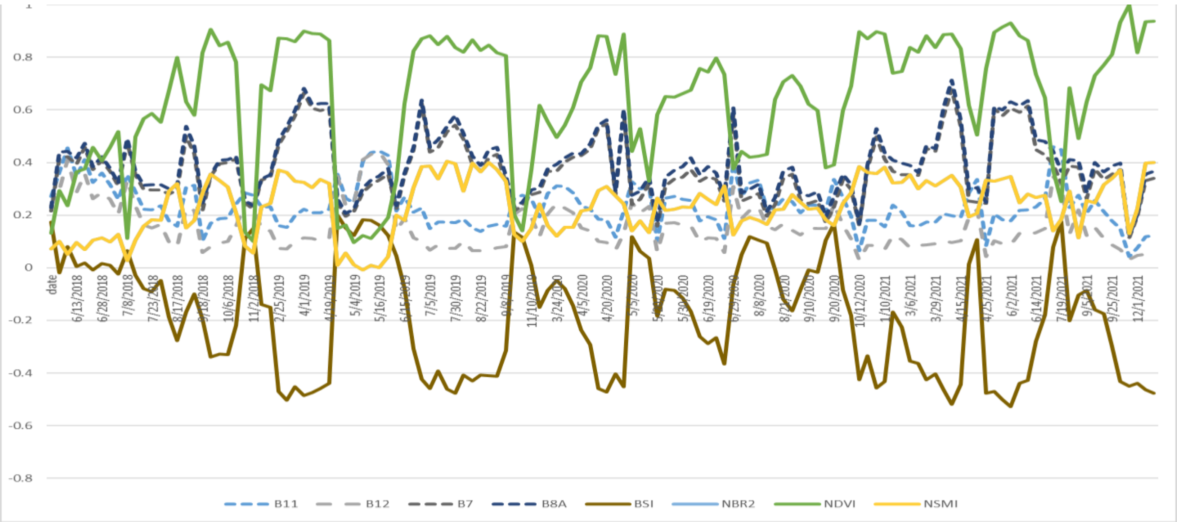


Figure 63: Visualization of reflection bands and indices for the sampling point 33. In total, we have 131 reflection signatures for the period of May- 2018 until the end of 2021. Only 13 reflection signatures correspond to bare soil (10%).

[8] Reducers are the way to aggregate data over time, space, bands, arrays and other data structures in Earth Engine. The ee.Reducer class specifies how data is aggregated. The reducers in this class can specify a simple statistic to use for the aggregation (e.g. minimum, maximum, mean, median, standard deviation, etc.), or a more complex summary of the input data (e.g. histogram, linear regression, list)

[9] You can export images, map tiles, tables and video from Earth Engine. The exports can be sent to your Google Drive account, to Google Cloud Storage or to a new Earth Engine asset.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The next step in the modelling process is to link the reflection signatures of each sampling point with the top Soil Organic Carbon Measurements. This link is only possible if the sampling points locations correspond to a specific pixel area (within). We have ensured this by using a specific soil sample collection protocol described in Deliverable 3.2 Catalogue on auxiliary data and available repositories to be incorporated.

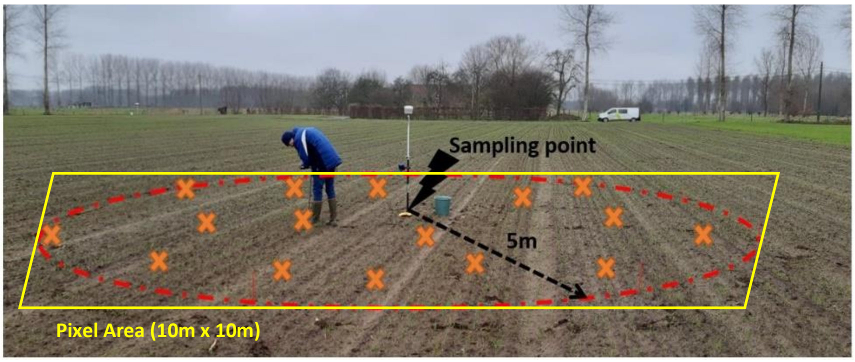


Figure 64: The soil sampling collection area should be within a pixel area. Otherwise is not logical to link the reflection signatures of bare soil pixels with the lab SOC measurements.

After linking the reflection signatures with the measure topsoil organic carbon values, it's possible to develop training, testing and validation data sets under different scenarios. The scenario we created:

- Make use of median values per band and per sampling point.
- Make use of all bands as input data.
- Make use of sampling points with more than five bare soil pixels (quality thershold) with the bare soil collection layer.

We use the Colab notebook as a collaboration environment that harnesses the full power of popular Python libraries and analyze and visualize data. With Colab, or "Collaboratory", you to write and execute Python in your browser, with:

- Zero configuration required
- Free access to GPUs
- Easy sharing

To perform data preparation, model training, hyperparameter tuning, analysis and interpretability, and model selection, we use PyCaret, an open-source, low-code machine learning library in Python that automates machine learning workflows. With PyCaret, within your notebook, you train your model, analyse it, iterate faster than ever before, and deploy it instantaneously as a REST API or even build a simple front-end ML app.

We tested both regression and classification models within the first iteration of our product developments (Phase 2, see Figure 65).

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

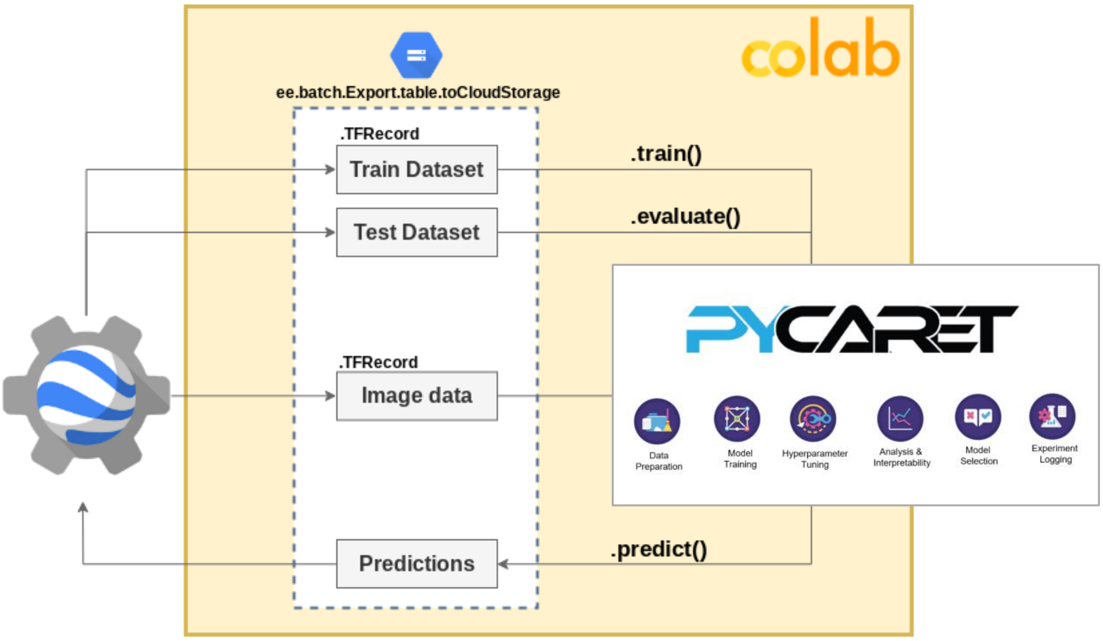


Figure 65: Together with Colab and Pycaret, GEE creates a robust technological framework that allows collaboration, ensures scalability, and supports productivity.

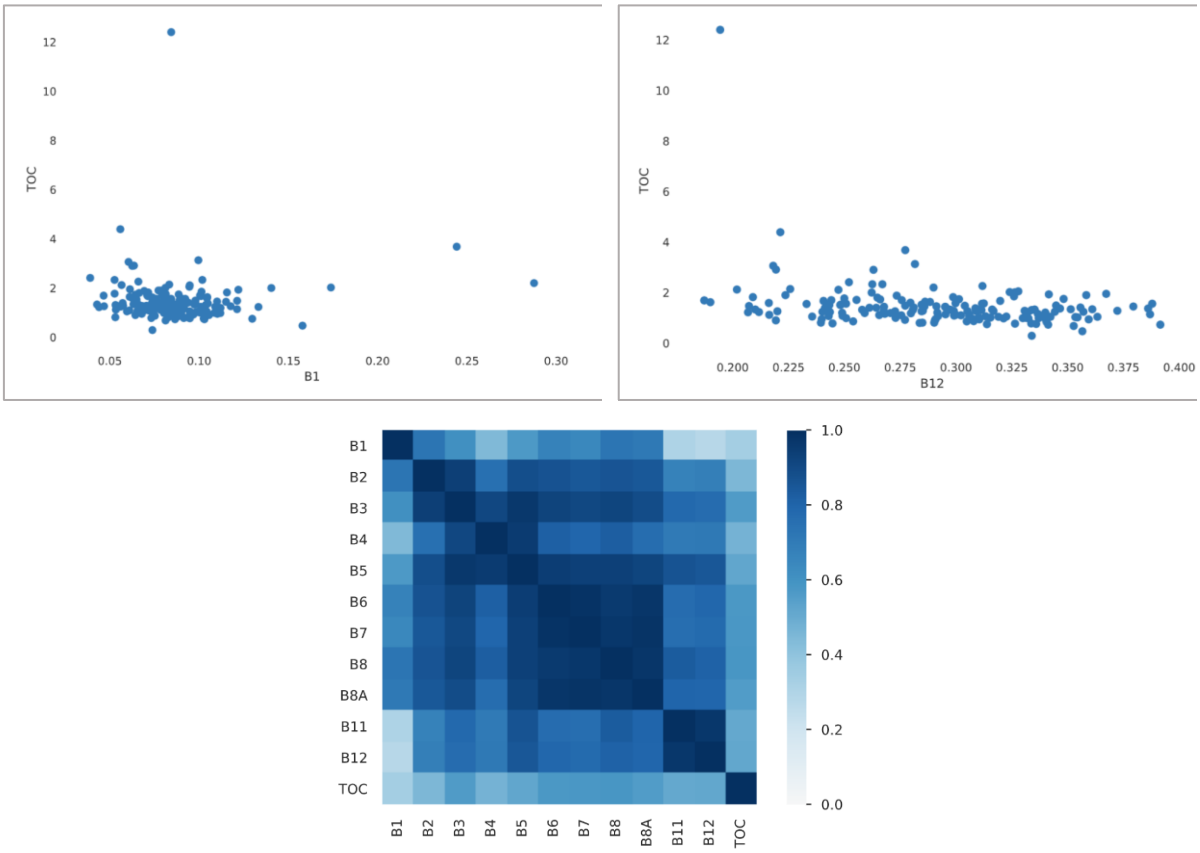


Figure 66: Part of the profiling report presents the Interactions and correlations (Phik) between input and output parameters.

4.4.2.1.3 Phase Three: Model deployment

At **Phase Three,** we apply the deployed model to all pixels belonging to larger areas, at regional (Envision, BC3) or even national scale. The deployment of machine learning models is the process of making models available in production where web applications, enterprise software, and APIs can consume the trained model by providing new data points and generating predictions. Normally machine learning models are built so that they can be used to predict an outcome (binary value i.e. 1 or 0 for Classification, continuous values for Regression, labels for Clustering, etc. There are two broad ways of generating predictions (i) predict by batch; and (ii) predict in real-time.

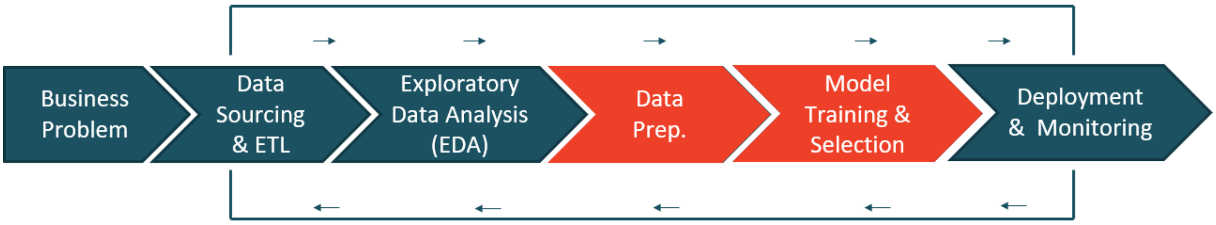


Figure 67: PyCaret — Machine Learning High-Level Workflow

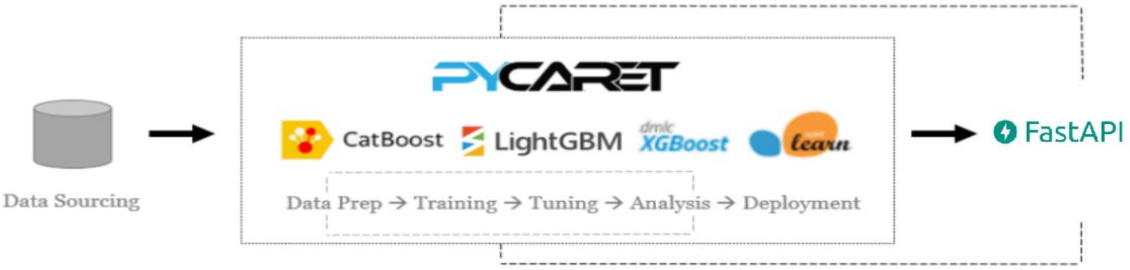


Figure 68: Using PYCAREt and web frameworks for building APIs with Python like FastAPI makes it possible to generate machine learning pipelines for batch or real-time predictions.

We also aggregate the SOC from pixel level to parcel-level at this phase. LV has provided the needed data sets to develop the first version of the data products, consisting of the Flemish Agricultural Parcels (Landbouwgebruikspercelen_LV_2020_GewVLA_Shapefile). This file was used to aggregate the Top Soil Organic Carbon Assessments at the parcel level.

In this phase, critical decisions are also made on how you will present the information to the service consumers or the end-users. In this decision, we need to consider the model accuracy and the specific CAP needs for monitoring coming from LV. We tested the visualization of the results in classes, for example, Low, Medium and High. The use of classes supports better the definitions of rules and logic similar to the logic of the Traffic lights in CAP monitoring.

4.4.2.1.4 Phase 4: Validation and Evaluation

In phase 4, we perform the technical validation and service evaluation, which means validating a complete solution or a segment of a solution that is about to be or has already been implemented to determine how well a solution meets the business needs and delivers value to the organization.

Within the ENVISION project, we perform the technical validation within WP3, mainly in Task 3.8 and the service evaluation within WP5 and Task 5.3. Useful inputs in Phase 4 are the end user requirements from WP2 and end-user evaluation feedback from WP5.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Technical validation is an important activity that focuses on the accuracy of the models and investigates scenarios from improvements built and test them. For example, one scenario for improvements is related to the use or not of the median value to generate the reflectance signatures. Even if this approach is well described and

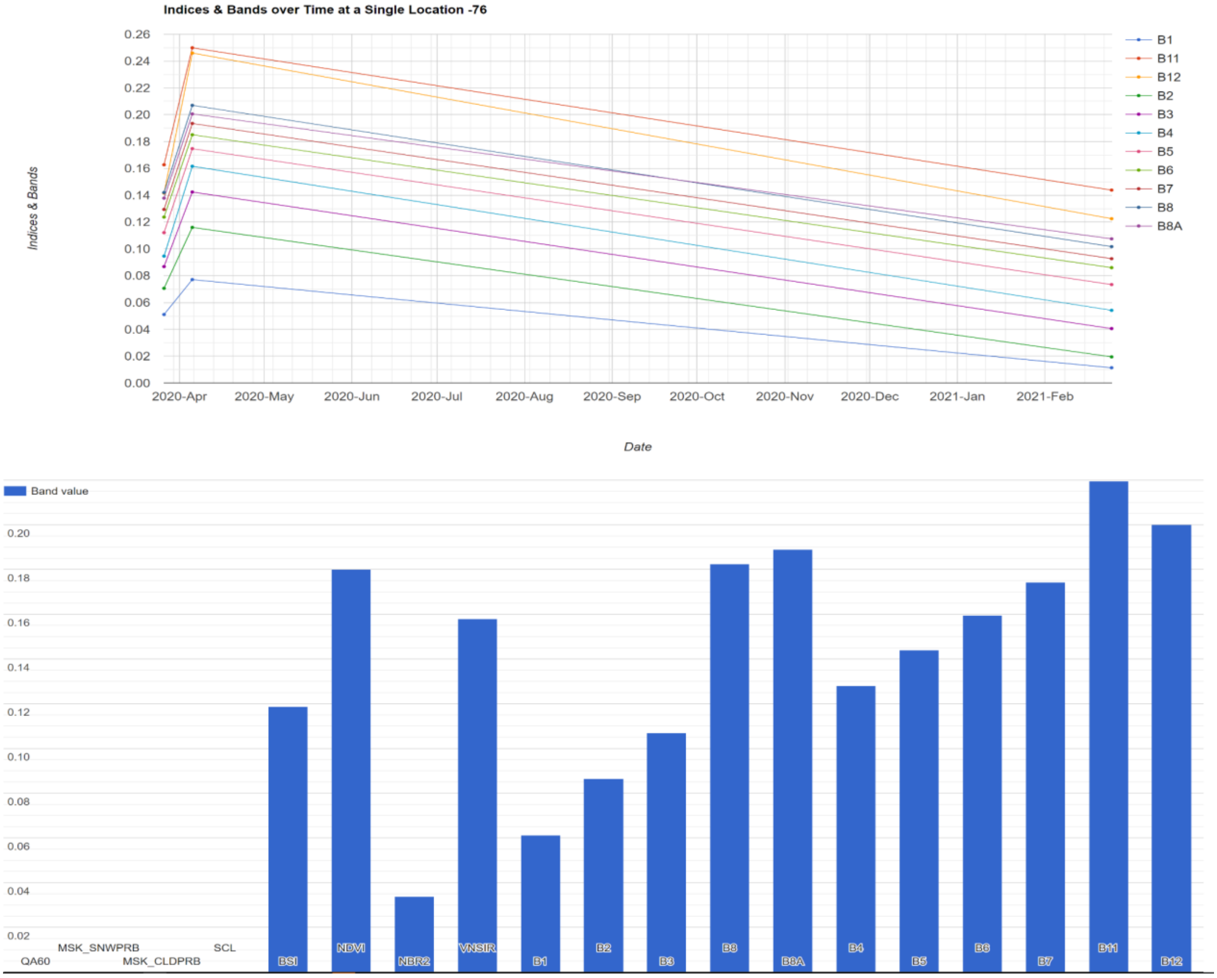


Figure 69: Sampling location point 76 has a measured SOC value of 3.06%, and it’s the only. Only four sampling points exist within the range of 2.8-3.5 (see Figure 47). Additionally, for 2018 until 2021 the median value for point 76 is extracted from only 3 pixels values (April 2020 and Feb 2021). The variation of reflection values per band is significant. For example, for Band 1, the mean value is 0.06, and the reflection in Feb 2021 is 0.01.

4.4.2.1.5 Phase 5: Improvements

In Phase five, you need to evaluate and make improvements, considering how the changes at each phase affect the product in other phases and the service itself. It’s a critical phase because it supports traceability and monitoring, which means approving and assessing changes to product information to manage it throughout the business analysis effort.

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

4.5 BC4: Monitoring organic farming requirements – Serbia

4.5.1 Data product description

This product provides a fully-automated Organic crop identification service, which aims at identifying whether a particular crop type declared as organic is classified as such, based on a traffic light system. Plants cultivated under organic and conventional farming principles present bio-chemico-physical differences that can be detectable by satellite imagery, especially during the vegetative and reproductive growth stages. The Identification of organic farming practices service will benefit from these differences to discriminate between organic and non-organic (conventional) crops. The logic behind the service is to identify distinct patterns characterizing the growth and evolution of organic and conventional crops during the growing season, through the use of both high resolution optical and radar satellite images depicting the phenological status of the cultivated parcels. Machine learning classifiers (MLC) will be trained to understand the temporal and spectral signatures of conventional and organic crops. The Predictor Layers will include

- Sentinel-2 MSI multispectral bands
- Vegetation Indices
- Crop biophysical parameters
- Crop Phenology features

To support the creation of classification predictor layers, the service relies on a cloud-based processing framework of EO data in order to derive vegetation indices and phenology features, that subsequently feeds them as input to a trained classification algorithm. Cloud processing is achieved by the exploitation of the Copernicus Data Information Access Services (DIAS) infrastructure, and specifically the CREODIAS platform. CreoDIAS will be used as the primary resource of retrieving Sentinel data, as according the Deliverable 3.1 Cost-benefit analysis, seems to be currently the best-fit solution for ENVISION in terms of budget and the offered services. Specifically, Sentinel-2 Level-2A data are going to be exploited. The Sentinel-2 Level-2A products are offered in the most of the cases as Bottom of Atmosphere (BOA) reflectance images derived from the associated Level-1C products.

The output form of the Organic crop identification product is a traffic light system with the cultivation method classification at parcel level (vector data). It is set up operationally on the ENVISION Platform to identify the cultivation practices by the end of the growing season. The traffic light system enables a smart sampling technique for the inspections. Each parcel will be characterized with the confidence of its classification decision (red, green, blue). These smart inspections methodology will identify potential breaches of compliance and assign the appropriate color to suspicious parcels declared as organic, based on their deviation from the classification decision

The provider of the Identification of organic farming practices service is AgroApps, and the outcome product addresses to the Serbian Certification Body (CB), which is the relevant authority. The service will be integrated and delivered as an earth observation component of the ENVISION platform with geographical coverage across the Serbia region.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The general contribution of the product to ENVISION, is towards the replacement of direct and guide on-field checks for priority control and will result in the reduction of inspections costs of the Certification Bodies (CBs) administrative burden, thus ensuring targeted and efficient controls and faster delivery of payments/organic certifications to farmers. Regarding more specific user requirements, the product addresses:

- The ability to identify and distinguish between organic and conventional crop, and to monitor pesticide use on the declared plots because this is an important objective in many agri-environmental policies. The product offers a distinct classification/ categorization on the platform between organic and conventional parcels that have been imported. Each of them is colored with a different color based on its category (green- organic, purple – conventional). Furthermore, vegetation indices are provided to the end-users as a monitoring tool.
- The ability to get ENVISION outputs per parcel, especially for information on yield of each crop. The traffic light system is a parcel-based solution, as well as the yield monitoring offered to the client by OCTOPUSH. ENVISION platform offers the possibility to export the outputs.
- The provision of accuracy of the service through relevant indicators and sufficient documentation of the methodology. The Organic crop classification service provides all the relevant accuracy metrics of the trained algorithm, for each crop. Such info is given to the end-user, as a metadata record on the traffic light system attributes, through the ENVISION platform.

4.5.2 Methodology

The description of the methodology for the creation of the Organic crop identification service on the current deliverable is divided on two subchapters, one being the methodology that was followed for the training of the classification models with the combined use of in situ and EO derived data, and a second one regarding the deployment of the classification models, to supply the traffic light system for the Organic crop identification on an operational mode.

Machine Learning models for Crop practice Identification

The methodology process flow for the training of ML models for organic practice identification consisted of the successive preliminary steps of Vegetation Feature extraction and Ground truth data sampling of the EO derived products, which resulted to the creation of the training -validation dataset, and the resultant application of a machine learning framework for the creation of crop specific models. The framework approach was implemented on the CREODIAS platform environment with the aid of the following software and libraries

- ESA SNAP Graph Processing Toolbox
- SAGA GIS
- Orfeo Toolbox and PhenOTB remote module
- R Libraries : mlr, caret, tidyverse
- Python libraries : scikit-learn

The ENVISION project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 869366

The general methodological framework for the training of ML models for organic practice identification is presented on the following flowchart.

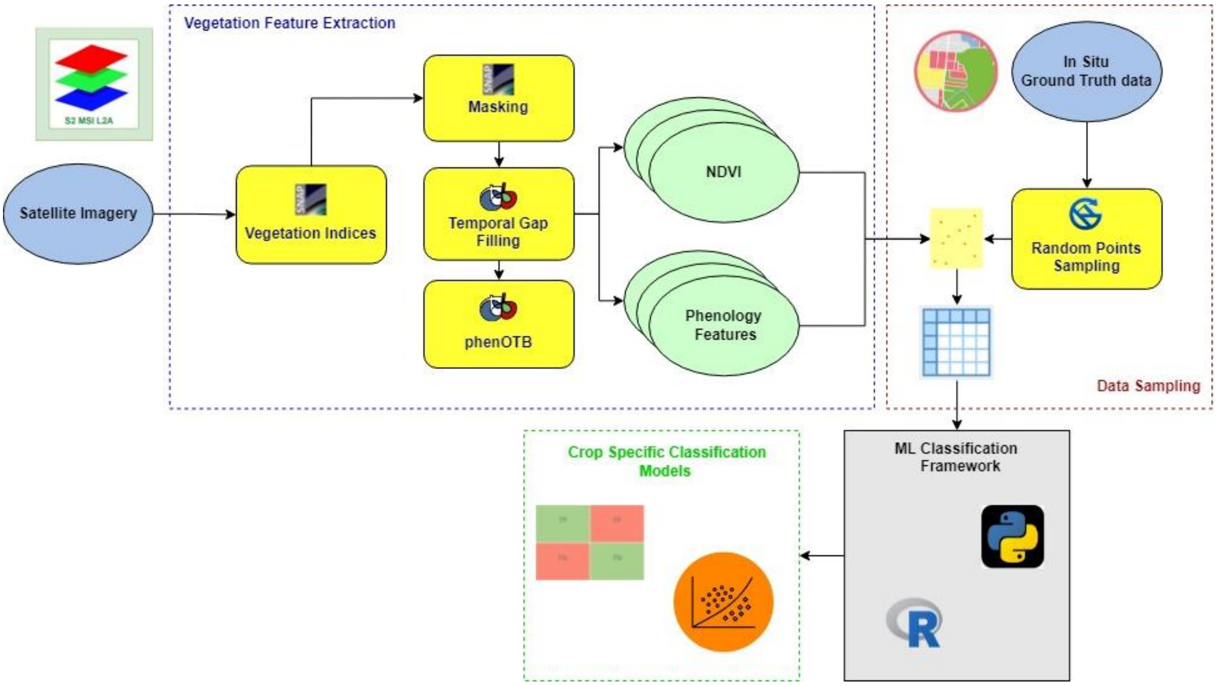


Figure 70: Methodological framework for the training of ML models for organic practice identification

Training Data- Predictor Variables (X)

EO Feature Extraction

The rationale of this specific step was the creation of a dense timeseries of image features that would focus on vegetation optical properties and phenology status, as the predictor variables of the crop classification models.

- **Vegetation Indices Features**: The Vegetation Feature Extraction step received Sentinel-2 L-2A images data as input and involved the calculation of an NDVI timeseries layer stack, and a subsequent processing with image masking and temporal interpolation for gap-filling purposes.
  Image masking was based on the L2A Scene Classification (SC) layer, which provides a pixel classification map (cloud, cloud shadows, vegetation, soils/deserts, water, snow, etc.), and it was decided in order to reject pixels belonging to unwanted land cover for the specified classification task. As a result, only pixels denoted as vegetation or barren land were preferred, and all other SCL classes were omitted.
  Temporal Interpolation was applied on the masked NDVI layer stack, in order to fill the gaps created from image masking, as well as to create a regular temporal 5-day step on the timeseries. For this purpose, the Orfeo Toolbox, ImageTimeSeriesGapFilling, library was used, which replaced invalid pixels (as designated by a mask) by an interpolation using the valid dates of the series. The

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Interpolation technique is based on Spline polynomials and depending on the number of valid dates in the temporal profile, the interpolation will be performed differently.

- With Less than 3 valid dates the algorithm applies linear interpolation

- With 3 or 4 valid dates, cubic splines with natural boundary conditions are used. The resulting curve is piecewise cubic on each interval, with matching first and second derivatives at the supplied data-points. The second derivative is chosen to be zero at the first point and last point.
- With more than 4 valid dates, a non-rounded Akima spline with natural boundary conditions is used.

- **Phenology Features:** The incorporation of phenology features on the classification models was based on the assumption that organic crops would showcase slower vegetation growth and lower yields than conventional crops and this fact could be reflected on lower rates of crop growth curve and lower plateau values on the NDVI temporal profile. For this specific purpose the Orfeo Toolbox remote module, phenOTB, was used. This module implements a several algorithms allowing to extract phenological information from time profiles. These time profiles should represent vegetation status as for instance NDVI, LAI, etc.

  The library provides tools for fitting parametric double logistics models to time profiles. From the double logistic fitting, some key parameters can be obtained. The parameters of the model can be used to define the following phenological metrics:

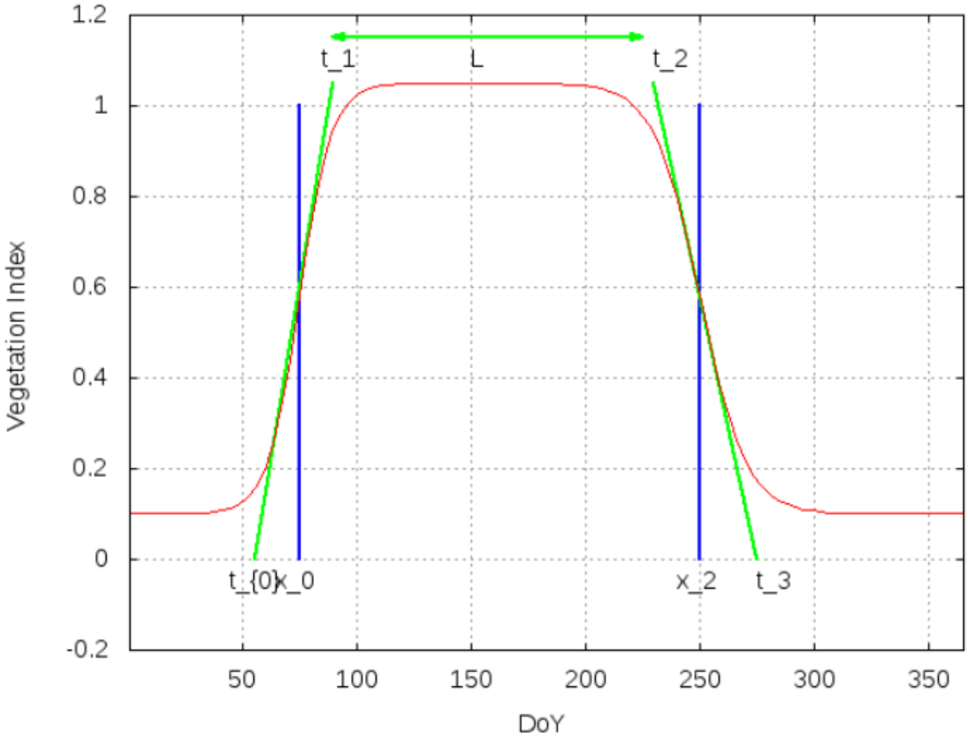


Figure 71: PhenOTB double logistic function fitting and associated parameters

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

- **Sowing date**: t_0
- **Crop Growth slope:** $g'(x_0)$
- **Length of the plateau:** t_2-t_1
- **Senescence slope:** $g'(x_2)$
- **Harvest date:** t_3

It allows fitting double logistics to each pixel of an image time series. The output contains 2 double logistics, one for the main phenological cycle and another one for a secondary cycle. This secondary cycle may not be present in the input data. This should not have any impact in the estimation of the main cycle. The application can output an image where each band is one of the phenological metrics for the 2 cycles. The order of the metrics is g0(x0), t0, t1, t2, t3, g0(x2). For the implementation of the classification tasks the Crop Growth slope, Length of the plateau and Senescence slope layers were used.

Training Data – Response Variable (Y)

The training data consisted of both organic and conventional farming practice ground truth parcels, emerging from the Business Case of Serbia (Doo Organic Control System Subotica – OCS). Practice Type was the response variable (**Y** class vector) whereas crop type variable was used to stratify crop specific models. Summarizing the provided ground truth data:

Out of 5191 parcel records for which crop information has been received, 4201 were successfully imported to the database having all related information including the field of Geometry. Those 4201 parcel records refer to parcels of different crops, years and farming practices as follows:

- 2335 conventional parcels
- 1866 organic parcels

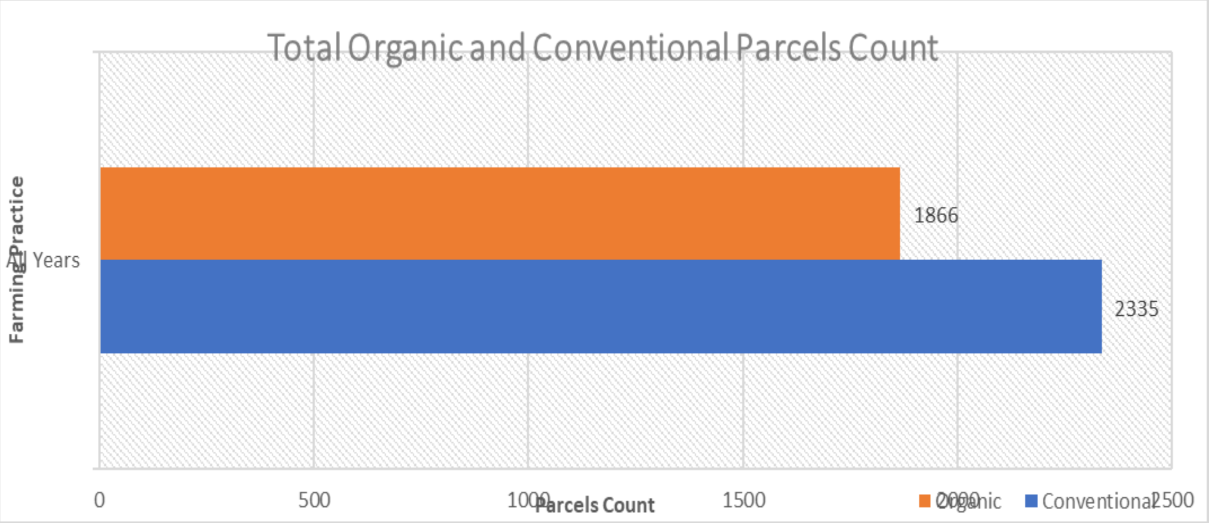


Figure 72: Total Organic and Conventional Parcels Count

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

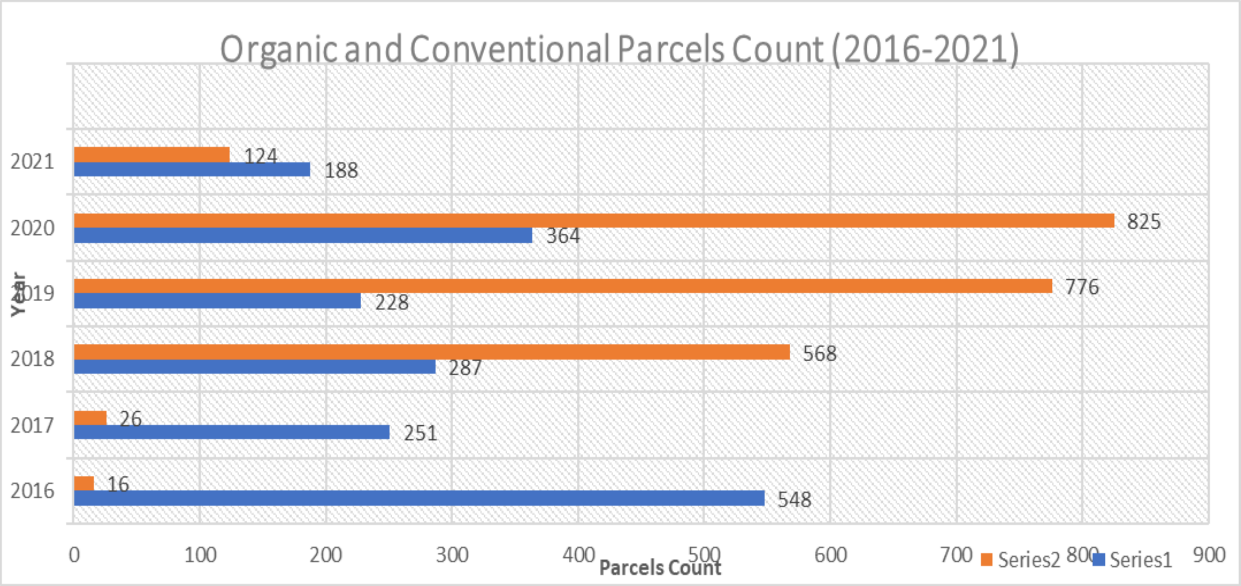


Figure 73: Organic/Conventional parcel count distribution through the ground truth years

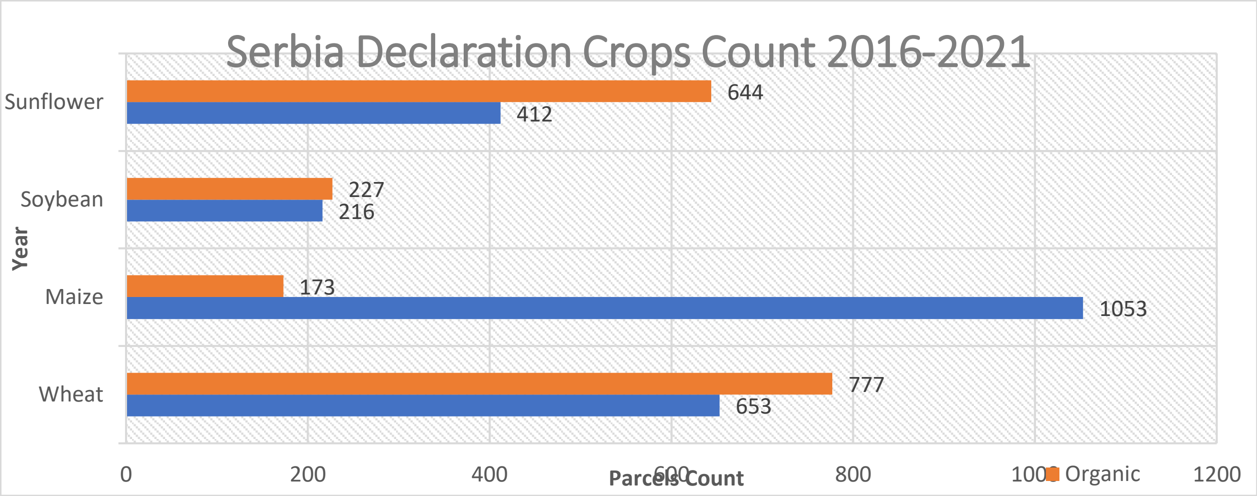


Figure 74: *Organic/Conventional parcel count distribution among crop types*

In order to achieve a fairly successful discrimination between Organic and Conventional crops, a sufficient number of representative pixels was required. Those pixels can be identified since they are located inside parcels of known crop characteristics. Since the pixel size is given (10m*10m), the size and the shape of the parcels should be sufficiently large, so that it totally contains pixels and consequently those pixels are representative of the crop type and practice. Consequently, there are two key-factors regarding the usefulness of the parcel data stemming both from the need to have sufficient number of representative pixels; the size & shape of each parcel, the number of parcels available.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

Parcel Geometry Characteristics

The geometry characteristics analysis of the received parcels showed that

- In general, the average parcel size is small, meaning that despite the number of parcels might be sufficient (which is not), the number of contained useful pixels per parcel is small and so is the total number of pixels.
- 344 / 4201 are very small to have any chance to include an entire pixel
  Parcel_area < 0.2ha, given the pixel size (10m*10m = 100 m^2 = 0.01ha)
- At least 1500 / 4201 have elongated shape (ratio: perimeter / area > very high values)

However, in many cases the long parcels are located next to each other. Therefore, a further step was carried out in order to unify (dissolve) neighbouring parcels of the same category. The following example demonstrates how the unification worked; parcels of the same category and season (in this case Wheat Organic 2016) that have common boarders (direct neighbouring) are unified to form one large parcel.

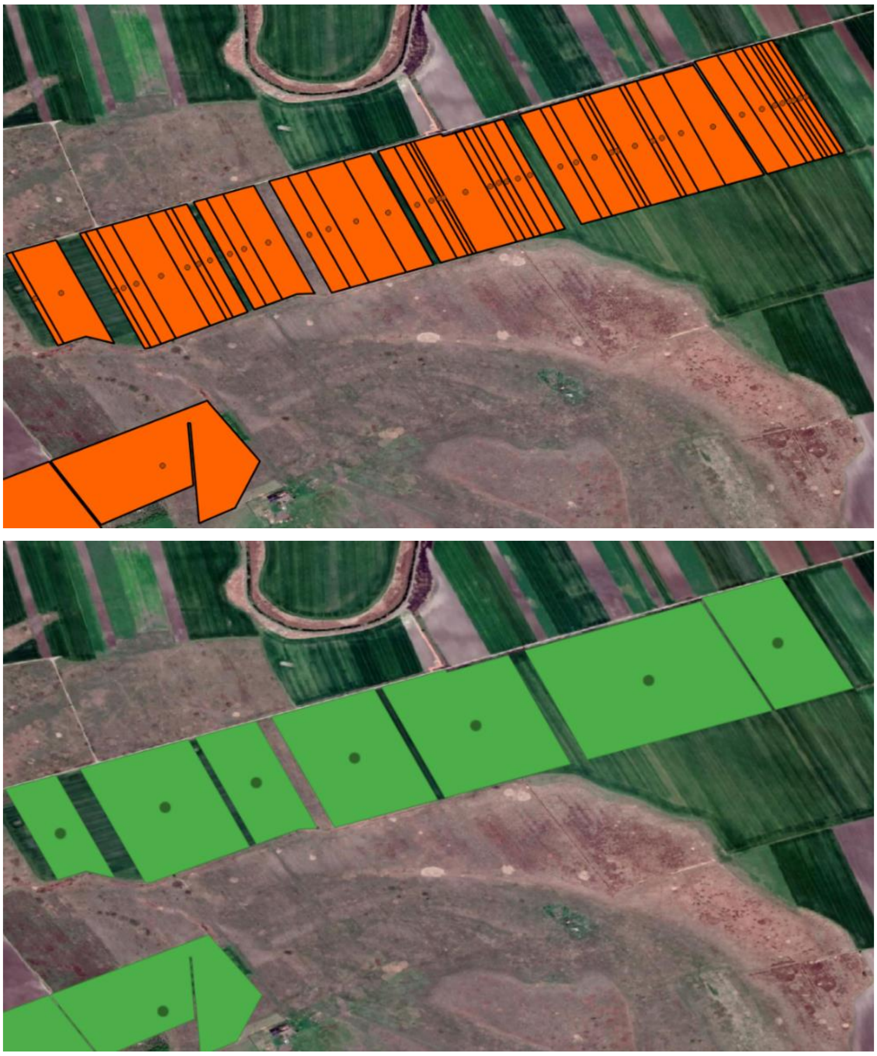


Figure 75: Parcel geometry dissolve

After geometry dissolve the total number of parcels was eventually reduced from 4201 to 1830 but the average parcel size increased.

Table 11: Number of parcels per Category in comparison to the Target

Category	Initially available	After unification (useful parcels)	Target
Wheat Organic	776	**220**	600
Wheat Conventional	653	**395**	600
Maize Organic	172	**73**	600
Maize Conventional	1053	**517**	600
Sunflower Organic	643	**168**	600
Sunflower Conventional	412	**258**	600
Soybean Organic	213	**69**	600
Soybean Conventional	216	**130**	600

Parcel Dispersion & Relevance

Another issue that should be noted here is that in many cases, **elongated single parcels are located scattered** an area making it impossible to unify them with neighbouring ones, making uncertain any possibility of usefulness.

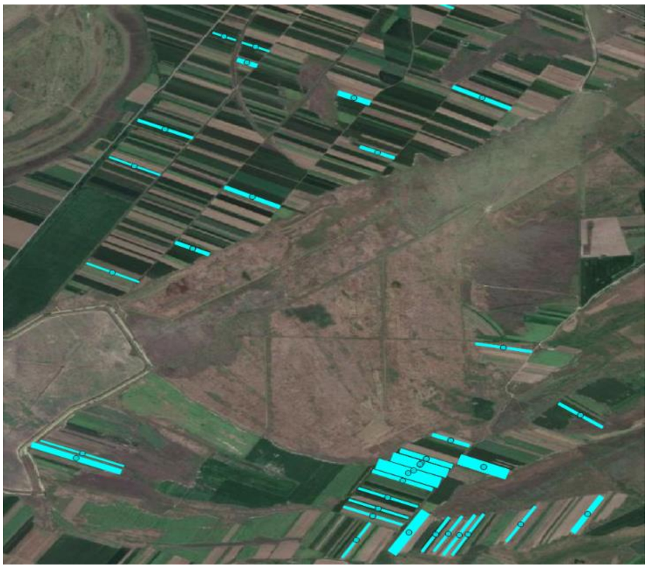


Figure 76: Elongated parcels scattered around an area

Finally, there were cases of parcels that contained land cover not relevant with the crops, like bush/tree boundaries or roads.

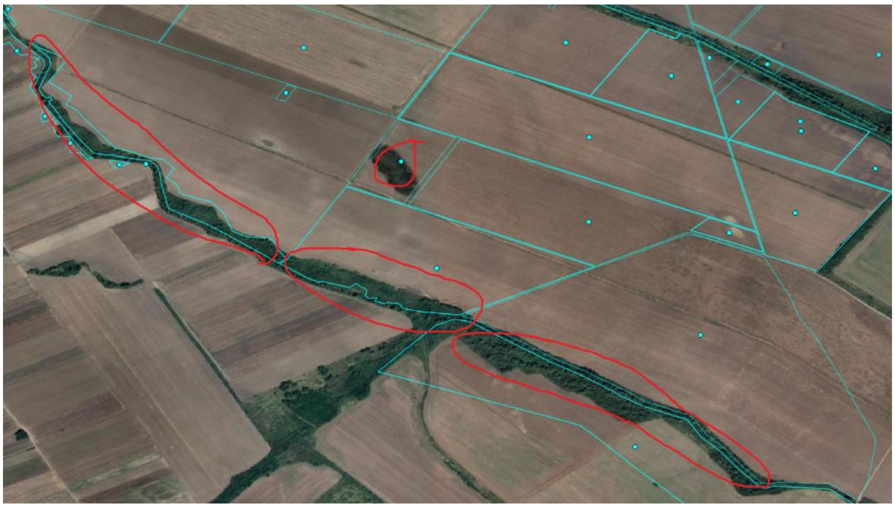


Figure 77: Natural vegetation objects inside crop parcels

Dataset Creation – Sampling

The creation of the training-validation-test datasets was created through random point sampling of the EO extracted features, inside the geometry borders of the ground truth parcel polygons. Initially, a buffer zone of 20m radius was clipped off the parcel geometries, in order to assure that outermost non reliable pixels of the crop parcels would not be included as training sites. A complete spatial random sampling strategy was followed, with a minimum distance of 14m and a sampling density of 60 points per ha. Regarding the vegetation feature timeseries, NDVI was sampled with a 5-day timestep, starting from the earliest sowing date of the ground truth data declarations, to the latest harvesting date. Phenology layers of Crop Growth slope, Length of the plateau and Senescence slope were joined to the dataset.

Machine Learning Framework

Outlier Analysis

Outlier Analysis was performed through PCA (Leave One Out Cross Validation), on the transformed Y-X Feature Space, in order to possibly reject data points belonging to non-relevant land cover. The application of the appropriate distance threshold values on F-Residuals, as indicated by the Residual and Influence Plots, through the use of the Hotelling T^2 criteria.

ML Classification Model Training

This procedure involved the training of crop specific binary classification models, for **Wheat, Maize, Sunflower, Soybean** crops, using the **Support Vector Machine** algorithm, using the libSVM imlementation. The validation strategy, aiming to improve the generalization of the models, had to consider a relatively small dataset and a hyperparameter tuning subroutine, and therefore the Nested Cross Validation approach was preferred.

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

In order to obtain honest performance estimates, all parts of the model building like pre-processing and model selection steps should be included in the resampling, i.e., repeated for every pair of training/test data. For steps that themselves require resampling like parameter tuning this results in two nested resampling loops. The graphic below illustrates nested resampling for parameter tuning with 3-fold cross-validation in the outer and 4-fold cross-validation in the inner loop

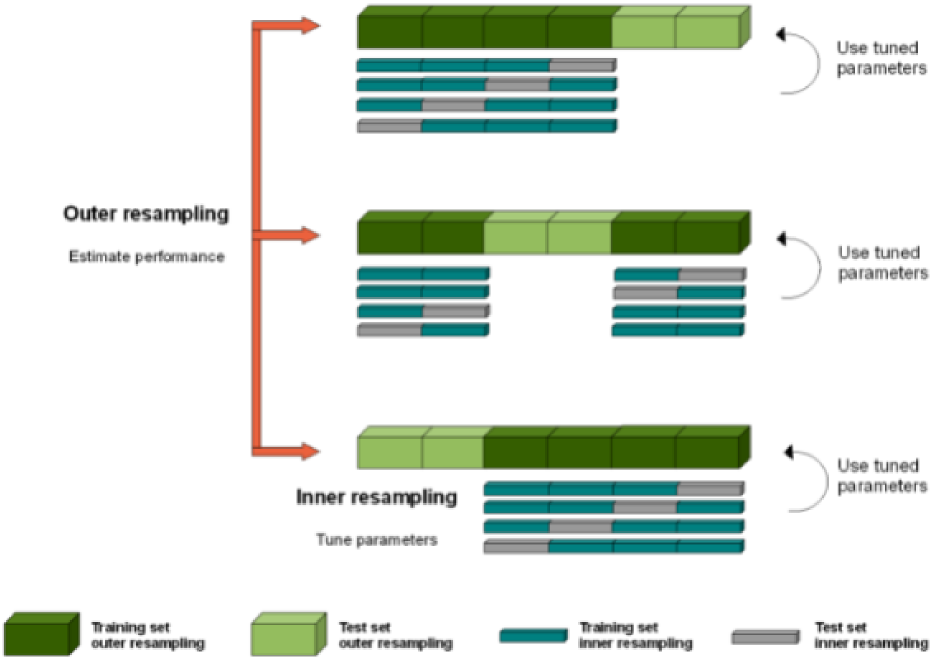


Figure 78: Nested Cross Validation schematic

In the outer resampling loop, three pairs of training/test sets exist. On each of these outer training sets parameter tuning is done, thereby executing the inner resampling loop. This way, one set of selected hyperparameters for each outer training set is tested. Then the learner is fitted on each outer training set using the corresponding selected hyperparameters and its performance is evaluated on the outer test sets. A data set partition of 80 – 20% ratio was decided to the cross-validation resampling.
Regarding the classifier algorithm, a C-SVM was trained for each crop data subset with a data center & scale pre-processing. A Radial Basis function (RBF) was chosen as the classifier kernel and the following hyperparameters were tuned via Grid Search optimization

- **Cost (C)**
- **Epsilon (ε)**
- **Gamma (γ)**

ML Classification Model Evaluation

To assess the accuracy in classification schemes, a comparison is usually presented in a confusion/error matrix where predicted classes, are compared with the actual classes. A confusion matrix includes different aspects of classification that refer to classified cases, here pixels, and they are necessary for

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

the calculation of various evaluation metrics. There are four such aspects of classification and they are described as:

- true positives (TP): number of pixels classified as class “A”, and in reality they belong in class “A”.
- true negatives (TN): number of pixels correctly not classified in class “A” since in reality they do not belong there.
- false positives (FP): number of pixels classified as class “A”, but actually they do not belong in this class. Also known as a "Type I error" or “commission error”.
- false negatives (FN): number of pixels that in reality belong to class “A” but they are classified to other classes. Also known as a "Type II error” or “omission error”.

Overall accuracy, precision, recall, F1 score and Cohen’s Kappa are evaluation metrics and characterize the actual performance of the classification.

- Overall accuracy refers to the number of correctly classified pixels (which in a confusion matrix appear in the diagonal) divided by the total number of pixels.

$$\text{Overall Accuracy} = \frac{\sum(\text{True Positives})}{\sum(\text{Pixels})}$$

(3.9)

- Precision is the proportion of true positives divided by the total number of pixels classified in this class (true positives + false positives).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall refers to the proportion of the true positives to the total number of pixels that actually belong to this class (true positives + false negatives).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

(3.11)

- F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. It is very useful since the balance between precision and recall is expressed that is indicative for the classifier’s performance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

(3.12)

- Cohen’s Kappa is generated to evaluate the accuracy of a classification. Kappa essentially evaluates how well the classification performed as compared to just randomly assigning values. It can range from -1 to 1 with any positive value indicating that the classification is better than random.

$$\text{Kappa} = \frac{\text{Overall Accuracy} - \text{Overall Expected Agreement}}{1 - \text{Overall Expected Agreement}}$$

(3.13)

The confusion matrix with all evaluation metrics are essential for the classification accuracy assessment necessary for the validation of a methodology.

Organic crop identification service operational mode

A general description of the methodology that leads to the data products, on an operational mode, is showcased on the following flowchart, which highlights the succession of processes throughout the services' Classification and Traffic Light components that were described above, to the web mapping interface of the ENVISION platform.

As far as the service input data requirements, the product refers to:

1. Crop data: The uploaded LPIS parcel polygon data of farm area, provided in the reference World Geodetic System (WGS84). The farm must be considered to comply with organic farming practices and will be monitored throughout the growing season to verify its eligibility and compliance. Specific attributes that are handled by the Classification and Traffic Light components are the crop type/ sowing-harvesting date fields.

2. Spectral data. The service can acquire Sentinel-2, satellite images from any available service provider. Based on the imagery used, the appropriate bands and products will be assimilated for the calculation of the indices that will feed the ML classifier, the results will be crop specific, notifications will be produced based on the decisions of the object-based (parcel) analysis, and visualization (graphs, reports, widgets) will be populated based on the results.

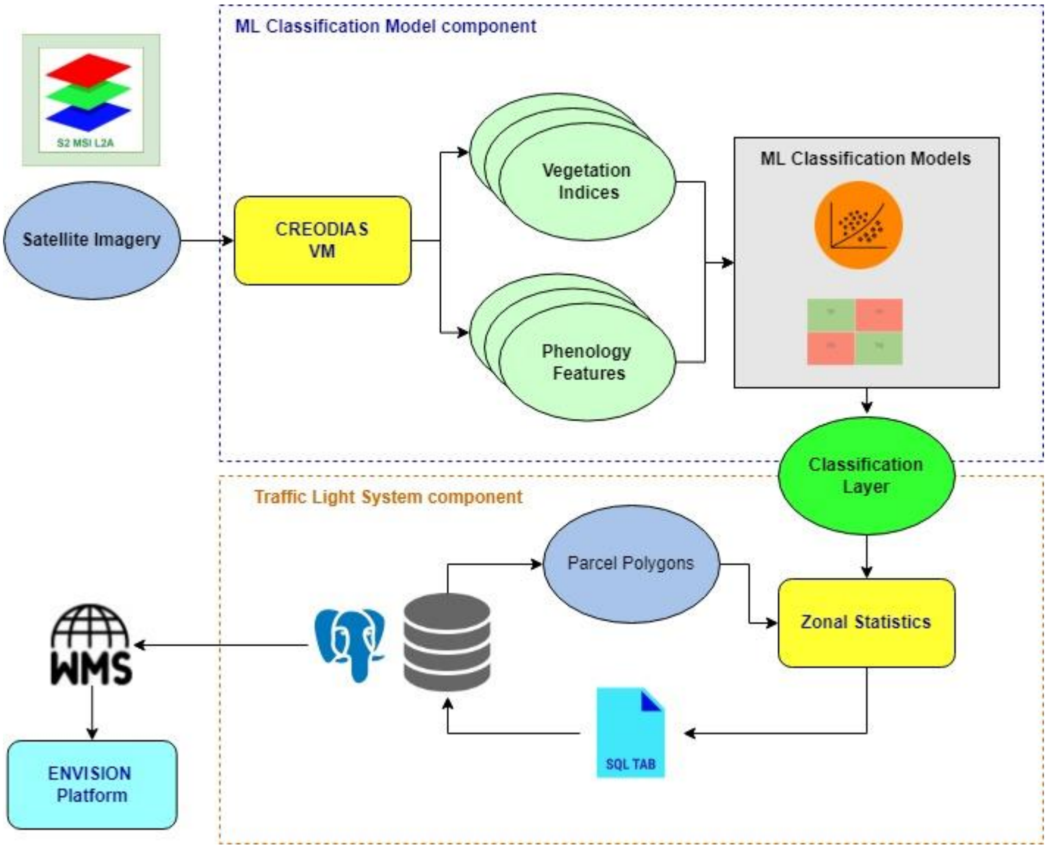


Figure 79: Organic crop identification service operational process flowchart

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

The table below presents the required data sources used for the operation of the Identification of organic farming practices service, as well as the spatial resolution of the data, the derived parameters and the update frequency. The polygon – parcel data, are made available by the CBs, and contribute the geospatial input for crop delineation and farmers' crop declarations, and will be employed i. for the object partitioning of the images, ii. the supervised classifier's training, and iii. to provide the classification decisions.

Table 12: The data required for the development and operation of the service

Source	Required Data	Spatial resolution	Derived Parameters	Update Frequency
Sentinel-2 mission	Sentinel-2 L-2A L-1C, optical multispectral	10 m, 20 m	Spectral Bands, VIs, biophysical parameters	4-6 days
LPIS (Land-Parcel Identification System)	Parcels vector data acquired	Polygon Data Crop Type	Parcel Geometry	Yearly
CBs	Parcels cropping data	Polygon attributes Farmer's declaration of the cultivation method	Parcel Crop Type	Yearly

Table 13: Sentinel-2 bands for the calculation of vegetation indices and texture Analysis Features

Band number	Central wavelength (nm)	Spatial resolution (m)
2	490	10
4	665	10
5	705	20
6	740	20
7	783	20
8	842	10
11	1610	20
12	2190	20

Table 14: Phenology Features applied at the end of crop cycle

Analytics	Parameter
Starting date	Date
End Date	Date
Growth Slope	Number
Senescence Slope	Number
Length of Plateau	Number

Table 15: Data from CBs

Type of Data	Parameter	Source	Units
Crop Data	Crop Type Sowing Date Polygon Data	Farmer	selection date

Regarding the output product, the service provides maps of decision on the cultivated practices and whether these are organic or conventional over a registered parcel by the end of the growing period. The product is accompanied with a legend showing the values of “organic”, “non-organic”, “not classified” (when the decision’s accuracy is lower than an acceptable value).

Table 16: Traffic Light System Output- Table of variables

Type of Data	Parameter	Spatial Resolution	Temporal Resolution
Vector	Decision (Y/N/NC) Organic – Non Organic Not classified	Object-Based (Parcel-Based)	Annually – at harvest

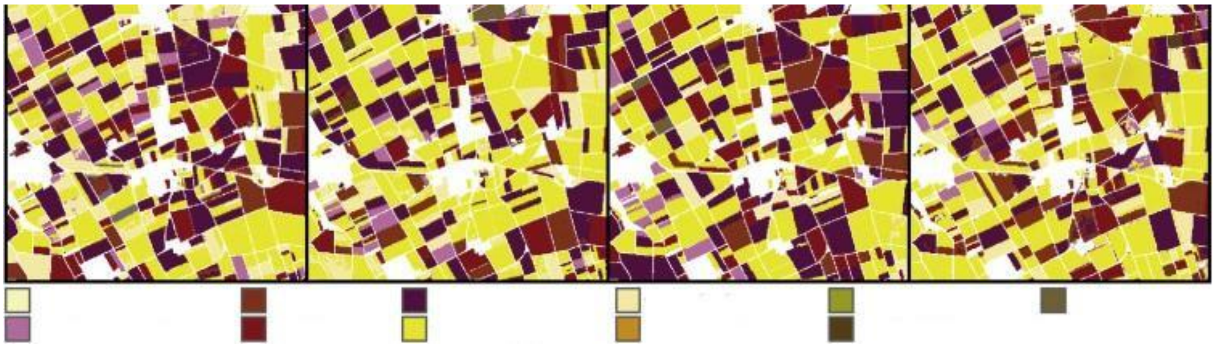


Figure 80: Visualisation of output data

Table 17: Output data for the monitoring of organic farming practices

Short description	File type	Expected size	Frequency
NDVI	GeoTIFF	1-2MB/file	~4/month
Yield estimation	GeoTIFF	1-2MB/file	~4/month
Parcel based Decision (Binary)	GeoTIFF	-	~2/growing season

Some of the specificities of the business cases that were accounted for, in the design of the methodology, were the following:

- Two or more crop types have been associated with the same parcel number
- Small size of a series of parcels
- Wrongly declared parcels cultivations in order to comply with local subsidy regulations
- Intense natural vegetation may confuse algorithms
- Several cultivations in the same parcel declared as one
- Very High Resolution (VHR) data may be needed so to map or/and validate the existence of more complex structures inside the parcels.

5) Conclusion/Final remarks

In this deliverable we report on the initial status of the ENVISION data products. Specifically, this deliverable focuses on i) translating how the user requirements (WP2) correspond to the data products described in the GA, and how these data products can be tailored and then be formed into services that address the user requirements; ii) what are the business case specifics and particularities and what is the exact strategy to address each of the; iii) data collection and pre-processing routines; iv) the methods employed to implement the data products; v) the limitations that emerged for each business case and how they were resolved or will be resolved in the future.

It should be noted that this deliverable is closely related to D3.4 that reports on the results and initial validation of the data products. More information on the limitations and the pertinent future work can be found there. It should also be noted that this is the initial report on the methods employed to develop the data products. A final report both for the methodology D3.7 and the validation of the data products D3.6 is expected by month 34.

Data products are or will be developed on ENVISION Datacube integrated in CREODIAS and OCTOPUSH by DRAXIS, and through Geospatial database all individual data products will be fed into the ENVISION platform. In the following months, appropriate changes will be implemented in order to finalize the data products and optimize the results with thorough evaluation of the methods described. The goal is to have data products that will facilitate useful services that fully address the user requirements. Moreover, we will explore the integration of additional data sources and synergies between similar projects for further improvements. Overall, we can claim that the majority of the data products presented are already mature enough and ready to be incorporated into ENVISION platform in order to initialize testing phase and collect the user’s feedback.

The ENVISION project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 869366

References

[1] Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007, 2007.

[2] Laurinavičius, A.; Juknevičiūtė-Žilinskienė, L. (2011). Eleven Years of RWIS Operation in Lithuania: Possibilities for the Use of the Data Collected. 8th International Conference "Environmental Engineering", May 19-20, 2011, Vilnius, Lithuania: Selected Papers. Vol. 3. Sustainable Urban development. Roads and Railways. Vilnius : Technika, 2011. ISSN 2029-7106. ISBN 9789955288299.

[3] Klimato rajonavimas (2013). Lietuvos hidrometeorologijos tarnyba prie Aplinkos ministerijos (Climatic Regioning of Lithuania. 2013. Lithuanian Hydrometeorological Service under the Ministry of Environment of Lithuania) [cited January 2013]. Available on Internet: http://www.meteo.lt/klim_rajonavimas.php

[4] Ignatavičius, Gytautas & Sinkevičius, Stanislovas & Ložytė, Aurelija. (2013). Effects of grassland management on plant communities. Ekologija. 59. 99-110. 10.6001/ekologija.v59i2.2713.

[5] Wischmeier, W.H. and Smith, D.D. (1978) Predicting Rainfall Erosion Losses: A Guide to Conservation Planning. Science, US Department of Agriculture Handbook, No. 537, Washington DC.

[6] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

[7] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

[8] Rousi, Maria & Sitokonstantinou, Vasileios & Meditskos, Georgios & Papoutsis, Ioannis & Gialampoukidis, Ilias & Koukos, Alkis & Karathanassi, Vassilia & Drivas, Thanassis & Vrochidis, Stefanos & Charalabos, Kontoes & Kompatsiaris, Ioannis. (2020). Semantically Enriched Crop Type Classification and Linked Earth Observation Data to Support the Common Agricultural Policy Monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. PP. 1-1. 10.1109/JSTARS.2020.3038152.

[9] Kontoes, Charalampos; Tsardanidis, Iasonas; et al. "Deep Learning for Event Detection on Grasslands", B42C-07 presented at 2021 AGU Fall Meeting, 13-17 Dec. https://doi.org/10.5281/zenodo.5995583

[10] Matic Lubej, Sentinelhub - Area Monitoring — Pixel-level Mowing Marker, Nov 4 2021: https://medium.com/sentinel-hub/area-monitoring-pixel-level-mowing-marker-968402a8579b

[11] Department of Meteorology (Ministry of Agriculture), The Climate of Cyprus, (n.d.).

[12] M. Katafygiotou, D. Serghides, Bioclimatic chart analysis in three climate zones in Cyprus, Indoor Built Environ. 24 (2015) 746–760. doi:10.1177/1420326X14526909.

[13] Anderson, E., Mammides, C. Changes in land-cover within high nature value farmlands inside and outside Natura 2000 sites in Europe: A preliminary assessment. Ambio 49, 1958–1971 (2020). https://doi.org/10.1007/s13280-020-01330-y

[14] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

The ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

End of Document

This project has received funding from the European Union`s Horizon 2020 research and innovation programme under grant agreement No. 869366.