D3.7 DATA PRODUCTS FINAL 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	D3.7- Data products final report			
Work Package	WP3 - Earth Observation data products			
Date of Delivery	**Contractual**	M34	**Actual**	M34
Nature	Report	**Dissemination Level**		Public
Lead Beneficiary	National Observatory of Athens (NOA)			
Responsible Author	Mr. Iason Tsardanidis (NOA)			
Contributions from	Mr. Athanasios Drivas (NOA), Dr. Vasileios Sitokonstantinou (NOA), Ms. Alexia Tsouni (NOA), Ms. Ifigeneia Tsioutsia (AgroApps), Dr. Panos Ilias (ILVO)			

Document History

Version	Issue Date	Stage	Description	Contributor
V0.1	03/06/2023	Draft	Input received from partners	NOA, AgroApps, EV ILVO
V0.2	27/06/2023	Draft	Comments from review	DRAXIS
F1.0	30/06/2023	Final	Final Version	NOA
F2.0	2/11/2023	Final	Final version after review comments	NOA

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!

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

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

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

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
GSAA	Geospatial Aid applications
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

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

NDWI	Normalized Difference Water Index
NOA	National Observatory of Athens
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 objective of this deliverable is to provide a comprehensive description of the data products that were developed and used in the ENVISION business cases. Specifically, it 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: Monitoring organic farming requirements

This deliverable is closely related to the following reports and documents:

1. "**D1.6 Final Data Management Plan**": This document provides a comprehensive list of the datasets used in the implementation of the final version of the data products.
2. "**D2.2 Report of Customer Requirements from ENVISION Services**": This report outlines the specific requirements and needs expressed by the customers, which served as a basis for developing the methods employed in this section.
3. "**D3.3 Data Products Initial Report**": This report presents the initial version of the data products that will be utilized in the business cases. A former version of the current deliverable.
4. "**D3.4 Data Product Validation Report**": This report constitutes the initial version of the validation process for the Earth Observation (EO) data products.
5. "**D3.5 Report on Collected Auxiliary Data**": This final report provides detailed information on the auxiliary data utilized in the algorithms. These auxiliary data, primarily sourced from the customers, were crucial as inputs for our algorithms and played a vital role in the training, fine-tuning, and validation of our methodologies.
6. It will be also linked to the upcoming "**D3.6 Data product validation report (final version)**" on M38, that presents the performance and validation outputs of the presented data products.

This final report provides an overview of the business case user requirements the specificities of the areas of interest, and a complete description of how the data products were designed to meet these requirements. It also takes into account the feedback received during the validation process in the second half of the ENVISION project and incorporates improvements made as a result. The report aims to address the comments and feedback provided by the reviewers during the first review meeting.

Overall, this deliverable serves as the culmination of the work done in WP3 to design and develop the EO data products for the ENVISION project, which aim to meet the specific needs of its customers. The report is organized into the following sections:

- **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 – Risks Mitigated:** Presents a concise overview is presented regarding the risks that have been deliberated.

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

- throughout the project's duration.
- **Section 6 – ENVISION's competitive advantage to Paying Agencies:** Outlines the benefits ENVISION offers to PAs.
- **Section 7 – Conclusions:** Contains the main conclusions of this deliverable and the users' feedback.

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

Five ENVISION data products have been designed based on an extensive analysis of the state-of-the-art in EO-enabled agriculture monitoring. This analysis incorporates insights from ENVISION partners, relevant literature, and past and ongoing projects. These data products are designed to meet customers' initial requirements and further refined during project implementation phase. The data products, mentioned in the Grant Agreement (GA), include:

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

Table 1: ENVISION Data Products

ID	Related Task	Data Product	Business Case	Services	Service Provider
DP1	Task 3.3	Analytics on Vegetation and Soil Index Time-series	NMA	Harvest events detection	NOA
			NMA & CAPO	Stubble burning identification on arable land	
			CAPO	Detection of illegal land clearing in Natura2000 protection areas	
			NMA & CAPO	Minimum soil cover for soil erosion	
			NMA & CAPO	Runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas	
DP2	Task 3.4	Cultivated crop type maps	NMA & CAPO	Confirmation of GSAA	NOA
				Smart sampling for OTSC inspections	
				Crops diversification compliance	
DP3	Task 3.5	Grassland mowing events detection	NMA	Grassland activity monitoring and management	NOA
DP4	Task 3.6	Soil condition monitoring	LV	Top-soil qualitative soil organic carbon estimations	EV ILVO
DP5	Task 3.7	Crop growth Monitoring and identification of organic farming practices	OCS	Distinction of organic farming practices	AgroApps
				Crop growth monitoring/ Crop phenology monitoring	

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

Task 3.3: Analytics on Vegetation and Soil Index Time-series: The objective of this task is to provide customized, user-defined analytics based on Earth observation (EO) data, focusing on vegetation status, phenological stages, soil condition, soil exposure, and Common Agricultural Policy (CAP) practices monitoring. By analyzing long time series of vegetation indices, we can identify crop-specific growth trends, detect changes, and track phenological phases. The task also includes delivering soil-specific indicators to assess soil health and address challenges like soil erosion, aligning with CAP objectives and complying with statutory management requirements (SMRs) and good agricultural and environmental conditions (GAECs). The generated analytics aim to support sustainable soil management strategies and facilitate compliance with CAP regulations. Furthermore, the product provides multidimensional time-space statistics and insightful plots/graphics for effective monitoring and evaluation purposes.

Task 3.4: Cultivated crop type maps: The objective of this task is to develop and test machine learning-based processing chains for accurately classifying different crop types in specific AoIs using the LPIS data provided by users. This classification will be used as the basis for monitoring products related to crop diversification, permanent grassland identification, and more. The cultivated crop type maps will be continuously updated and delivered dynamically to users with new satellite acquisitions.

Task 3.5: Grassland mowing events detection: The objective of this task is to develop the technical framework for detecting mowing events, which is a crucial component for monitoring farmers' compliance with CAP's environmental and climate requirements, as well as the maintenance of permanent grasslands. This solution will enable the quantification of grassland activity at a national scale, providing valuable insights into agricultural practices.

Task 3.6: Soil condition monitoring: The objective of this task focuses on addressing the challenge of soil condition monitoring, specifically targeting the estimation of Soil Organic Carbon (SOC) levels. The service aims to provide accurate estimations of top-soil qualitative SOC at the parcel level.

Task 3.7: Monitoring organic farming requirements: The Organic crop monitoring product offers an automated service to accurately distinguish between organic and conventional crops, providing valuable insights into farming practices. In addition, this product enables crop growth monitoring, including the assessment of crop status, estimation of Leaf Area Index (LAI), above-ground biomass, and yield estimation.

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				
			Analytics on Vegetation TS	Crop Type Maps	Grassland Mowing	Soil Condition Monitoring	Monitoring organic farming

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

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 3 below presents the user requirements (UR) that will be used for the relevant actor, as well as their description. This table has been also displayed in D3.3. The Table 4 provides a matrix linking each service component with each one of the Business Cases (BC) in order to satisfy the User Requirement (UR) presented above.

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 the 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.

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

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 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.

Table 4: ENVISION data products and the final services provided

ID	Related Task	Data Product	Business Case	Services	Service Provider
DP1	Task 3.3	Analytics on Vegetation and Soil Index Time-series	NMA	Harvest events detection	NOA
			NMA & CAPO	Stubble burning identification on arable land	
			CAPO	Detection of illegal land clearing in Natura2000 protection areas	
			NMA & CAPO	Minimum soil cover for soil erosion	
			NMA & CAPO	Runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas	
DP2	Task 3.4	Cultivated crop type maps	NMA & CAPO	Confirmation of GSAA	NOA
				Smart sampling for OTSC inspections	
				Crops diversification compliance	
DP3	Task 3.5	Grassland mowing events detection	NMA	Grassland activity monitoring and management	NOA

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

DP4	Task 3.6	Soil condition monitoring	LV	Top-soil qualitative soil organic carbon estimations	EV ILVO
DP5	Task 3.7	Crop growth Monitoring and identification of organic farming practices	OCS	Distinction of organic farming practices	AgroApps
				Crop growth monitoring	

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

4 ENVISION data and tools

This chapter provides the final version of the data products developed for each business case. This version was achieved after a thorough communication with the users to ensure the products meet their specific requirements. The methods implemented and data utilized to address the user needs are also discussed.

4.1 Analysis Ready Data and Technologies that enable data products

The introduction of the Sentinel missions has revolutionized agriculture monitoring by providing freely accessible, high-resolution data with both spatial and temporal variability. However, harnessing the full potential of these data presents challenges, including the processing and organization of the data into Analysis Ready Data (ARD). ARD undergoes essential steps such as atmospheric correction, geometric calibration, and resampling to facilitate easy analysis. NOA has developed a series of automated workflows for generating ARD. A more detailed documentation of data collection methods and preprocessing can be found in D3.3 and D3.5.

The provision of ENVISION products requires several key data sources:

1. **Satellite Imagery**: Sentinel-1 and Sentinel-2 is essential as the primary input for developing and enhancing the services. This imagery enables the extraction of valuable information and the creation of additional features.
2. **Geospatial Data -**: Usage of LPIS, a system that digitally identifies and maps agricultural parcels, helping to manage agricultural subsidies, land use, and related data. This is combined with farmers' declarations (GSAA) providing essential information for the development of the data products.
3. **Metadata**:
    a. Crop Validations (OTSCs, RS) help ensure accurate crop identification and calibrate the ENVISION products.
    b. Event Timestamps (Mowing, Harvest, Stubble Burning etc.) serve as crucial temporal markers for monitoring agricultural practices and optimizing the corresponding algorithms.
    c. National CAP Strategic Plans (GAECs, SMRs), contribute to the comprehensive understanding and implementation of the ENVISION products. The integration of these diverse data sources enables the generation of valuable insights and supports effective agricultural monitoring and evaluation.

4.1.1 Sentinel Data

NOA's back-end processes are hosted on the CREODIAS platform, providing direct access to EO data through a scalable object storage directory. The storage system offers high-performance access to the full archive of Sentinel-1 GRD, SLC, and Sentinel-2 L1C data for Europe (see Table 5). To overcome the limited availability of Sentinel-2 L2A products, NOA utilizes Sen2Cor software [1] for transforming L1C to L2A for specific cases. Python scripts have been developed to enable efficient searching and pre-processing of products in the eodata directory based on user-defined parameters.

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

Table 5: Datasets and products provided by CreoDIAS along with their archive policy

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 ***

4.1.2 Data Pre-processing

As mentioned above, data pre-processing is an essential step for generating ARD. To achieve this, NOA has developed a comprehensive set of automated procedures that are executed within the robust infrastructure of CreoDIAS. These procedures encompass various tasks such as Sentinel-1 backscatter generation, Sentinel-2 pre-processing, cloud masking, and parcel buffering analysis (important step to handle the problem of mixed pixels information). The details and findings of these procedures have been extensively analyzed and documented in D3.3 and D3.5.

Figure 1: Cloud Masking and Cloud Masking buffering in order to tackle cases of undetected clouds/shadows on the boundaries of the produced masks.

4.1.3 Open DataCube (ODC)

ENVISION utilizes the Open Data Cube (ODC) to effectively manage and analyze a large volume of Earth observation (EO) data for nationwide agriculture monitoring. ODC offers robust data structures and tools for efficient organization and analysis of EO data. NOA has successfully implemented and utilized ODC in the past in the context of DataCAP application (Figure 2), which includes automated modules for data download, pre-processing, and indexing. ODC's key advantage is its ability to catalog extensive EO datasets, providing convenient access and manipulation through a Python API.

ODC offers two methods for cataloging data: indexing and ingesting. Currently, indexing is considered the more efficient method, utilizing the DataCube-core module and a PostgreSQL database to store metadata. The ENVISION DataCube implementation involves setting up the required environment, configuring files, and initializing the database.

By combining eodata catalog and ODC framework we are able to deliver scalable solutions for nationwide agriculture monitoring. While initially focused on Cyprus and Lithuania, the methodology can be applied to other regions as well. The platform provides pre-processed data from Sentinel-1 and Sentinel-2 satellites, enabling comprehensive agriculture monitoring and supporting research and operational outputs. The indexed data covers the entire extent of Lithuania and Cyprus for multiple

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

cultivation periods, and all observed products are automatically indexed to the database hosted on the CreoDIAS VM. Users are required to create products associated with each indexed dataset, with a product defined as a collection of datasets sharing the same measurements.

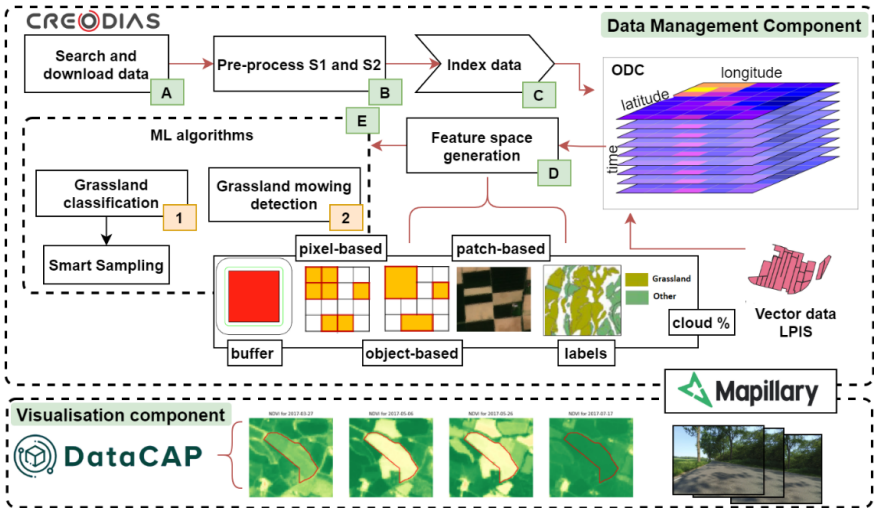


Figure 2: DataCAP architecture

In ENVISION, country-specific products are created and indexed in the DataCube using YAML files. These products provide time series data with a consistent spatial resolution of 10 meters. We utilize a Python API to request and access the data based on parameters such as product, time range, bounding box, and bands of interest. The data is loaded into Xarrays, which have dimensions of time, latitude, and longitude. However, analyzing data at a national scale introduces challenges due to the time complexity involved. Processing millions of parcels can significantly impact execution time. To address this, we transformed farmers' vectorized declarations (GSAA) into raster format (Figure 3). Each pixel within a parcel carries the corresponding parcel ID. Rasters are generated for eight different buffer variations (+10m, +5m, +3m, 0m, -1m, -2m, -3m, -5m).

Figure 3: Produced rasters for the case of Cyprus based on the farmers' declarations shapefile (blue line) for no buffering (red colour) and 5m inward buffering (green colour).

These rasters are indexed into the DataCube and loaded at the same resolution and extent. Storing this information in the DataCube allows us to optimize the calculation of zonal statistics for each parcel by utilizing the "*group by*" function. This function groups Xarray bands based on parcel IDs, enabling aggregation based on the parcel geometry. Figure 4 illustrates the stack of two layers: an RGB raster and the IDs raster.

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

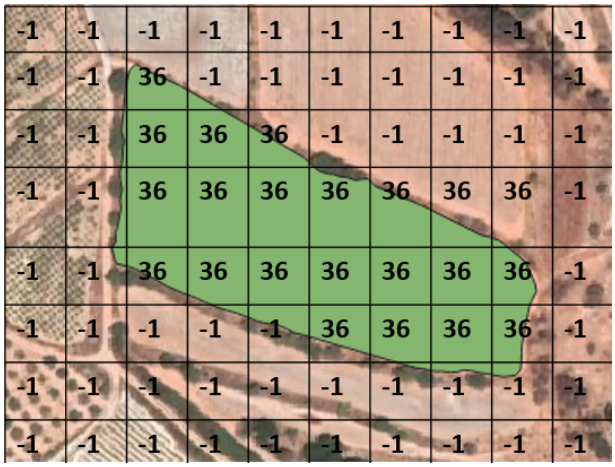


Figure 4: Parcel's index rasterization inside DataCube

The calculation and import of Sentinel-2 bands B02-B12 and Sentinel-1 VV, VH backscatter, and SLC data into the DataCube are crucial steps for the agricultural monitoring tasks. Sentinel-2 bands provide valuable multispectral information, allowing for the analysis of vegetation health, land cover classification, and crop identification. The Sentinel-1 VV and VH backscatter data offer insights into soil moisture, vegetation structure, and surface roughness. Additionally, the SLC data provides high-resolution radar imagery for detecting changes in land surface features. These datasets are processed, transformed into Analysis Ready Data (ARD), and indexed in the DataCube, ensuring easy access and efficient analysis for agricultural applications. After importing the Sentinel bands into the datacube, the next step in the processing pipeline involves the calculation of various vegetation and burning indices. The necessary formulas and band combinations (see D3.5) are applied to derive indices such as NDVI, NDWI, NDMI, PSRI, SAVI, EVI, DVI, VIgreen, VARIGreen, GDVI, SIPI, BSI, NBR, NBR2 and MIRBI for Sentinel-2 data. Similarly, for Sentinel-1 data, indices like VV/VH Ratio, VH/VV Ratio, RVI, and Cross-Ratio are computed. These calculated indices provide valuable insights into vegetation health, water content, burn severity, and other relevant parameters, enabling users to make informed decisions in various fields such as agriculture, forestry, and environmental monitoring.

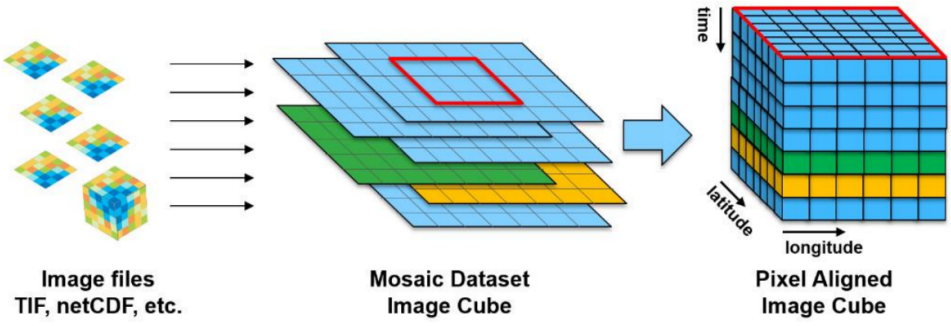


Figure 5: Datacube population with Sentinel products

4.1.4 Geospatial Database

The foundation of ENVISION's EO Big Data Analytics lies in the ODC and a geospatial database. The collection of data from farmers' declarations through an API is performed using scripts, which are then stored in 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 outcomes of these tasks

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

are used to update the database, establishing a continuous back-and-forth communication between ODC and database as it is shown in Figure 6. Consequently, we have a DataCube that are dynamically populated with Sentinel-1 and Sentinel-2 products, as well as auxiliary geospatial data, enabling the generation of data products that further enrich the cubes. As a result, member state-specific knowledge bases for CAP monitoring are established. Additionally, the power of the POSTGIS extension facilitates various operations such as computing distances between geometries, calculating areas, conducting buffer analyses, and performing geospatial queries.

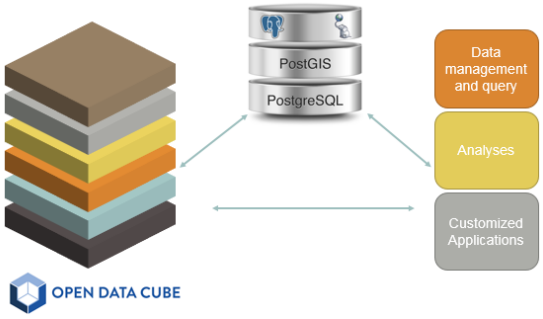


Figure 6: ENVISION products connection scheme

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

The National Paying Agency (NPA) is the key organization responsible for the implementation and monitoring of agricultural, rural development, and fisheries measures in Lithuania. The NPA has embraced technological advancements by developing software, registers, and geo-spatial applications. Additionally, it maintains strong communication channels with stakeholders such as farmer unions, municipalities, and advisory bodies. It has actively participated in several EU projects (e.g., DIONE, SEN4CAP, RECAP, NIVA and EO4AGRI) focused on CAP monitoring, Integrated Administration and Control System (IACS) modernization, and agricultural operational improvements. Lithuanian customers have shown a strong interest in utilizing Earth observation (EO) technologies for monitoring farmers' performance, with the NPA's mobile app allowing farmers to provide geotagged photos as evidence of their activities. Incorporating ENVISION outcomes into their operations will greatly enhance the NPA's supervision of CAP measures. This section provides an overview of the Lithuanian business case, including user requirements, the methodological approach, services provided, and current limitations. For more detailed and descriptive statistical analysis of the study site properties, please refer to document D.3.3.

4.2.1 Data products description

Lithuania is located in the Baltic region of North-West Europe and covers an area of 65,300 km². The country benefits from a temperate climate influenced by both maritime and continental factors. Its diverse agricultural landscape comprises a wide range of crops, with approximately 85% falling into five key agricultural categories: grasses, winter and spring cereals, rape, and potatoes. Grasslands are particularly prominent in the western and eastern parts of the country, while other cultivations thrive in the central zone. The remaining crop types encompass a range of agricultural products, including fallows, permanent cultivations (e.g. trees and vines), peas, beans, alfalfa, clover, vegetables, protein crops, agricultural mixes and herbs. This diverse agricultural landscape presents a unique opportunity

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

for ENVISION to provide comprehensive monitoring and analysis services for multiple crop types, which is not usually the case. Furthermore, with around 1.1 million farmers' declarations submitted annually to the NPA systems (Figure 7), there is a substantial volume of data that needs to be processed, underscoring the need for efficient and scalable solutions offered by ENVISION. One of the major challenges faced by the Lithuanian business case is the high occurrence of frequent and extended cloud coverage, leading to sparse Sentinel-2 image time series. Over 25% of the available Sentinel-2 products in Lithuania exhibit cloud coverage of at least 90%. This often results in significant temporal gaps between clear acquisitions, sometimes spanning an entire month. The presence of clouds and their shadows can introduce inaccuracies, noise, and negatively impact the precision of agricultural monitoring. Particularly, tasks involving the detection of rapid changes, such as mowing events, are sensitive to cloud interference, as they require observing swift variations in land greenery within a short time frame.

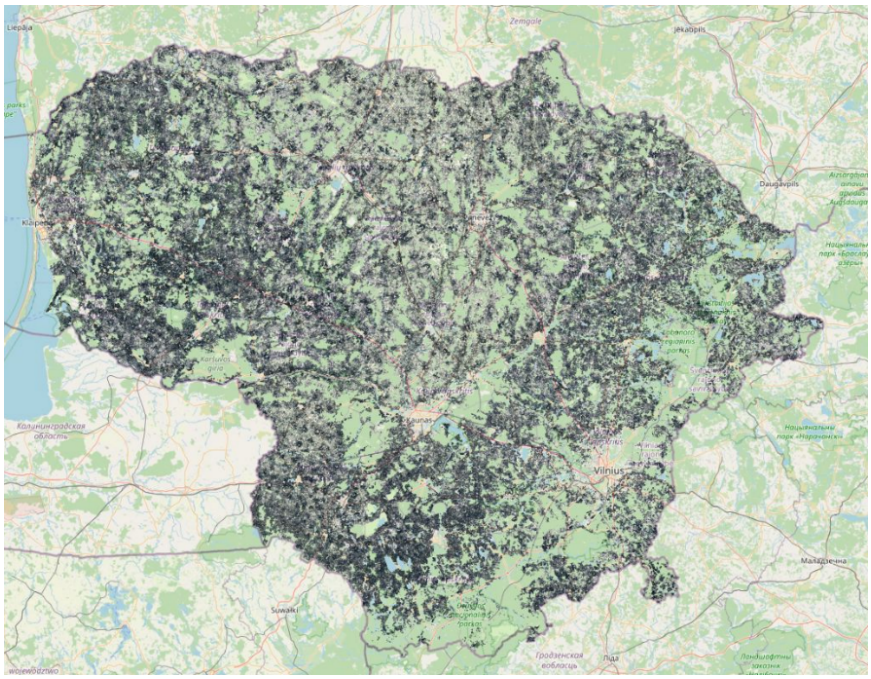


Figure 7: Lithuania Applicants Declarations (GSAA-2022)

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

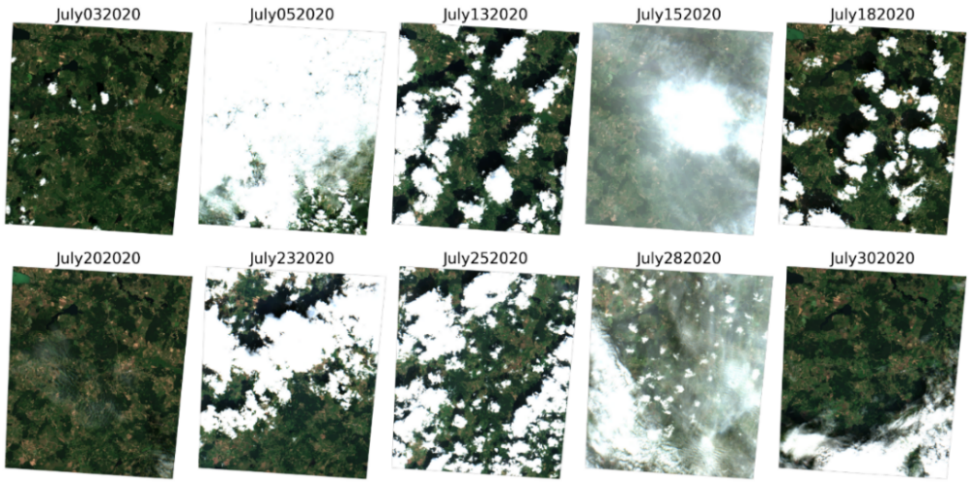


Figure 8: Frequent Cloud Coverage of a sample area in Lithuania

The distribution of crop types in Lithuania exhibits a wide variety, with numerous agricultural categories represented across the country. In the context of Crop Type Mapping, it is noteworthy that approximately 90% of the declared crops have been successfully included in the analysis. Figure 9 provides a detailed visualization of this distribution, highlighting the prevalence of different crop types. Interestingly, it is observed that more than 85% of the crops can be classified into five main agricultural categories, namely grasses, winter cereals, spring cereals, winter rape, and potatoes. While these dominant categories play a significant role in Lithuania's agricultural landscape, it is crucial to consider the inclusion of multiple crop types beyond the most prevalent ones.

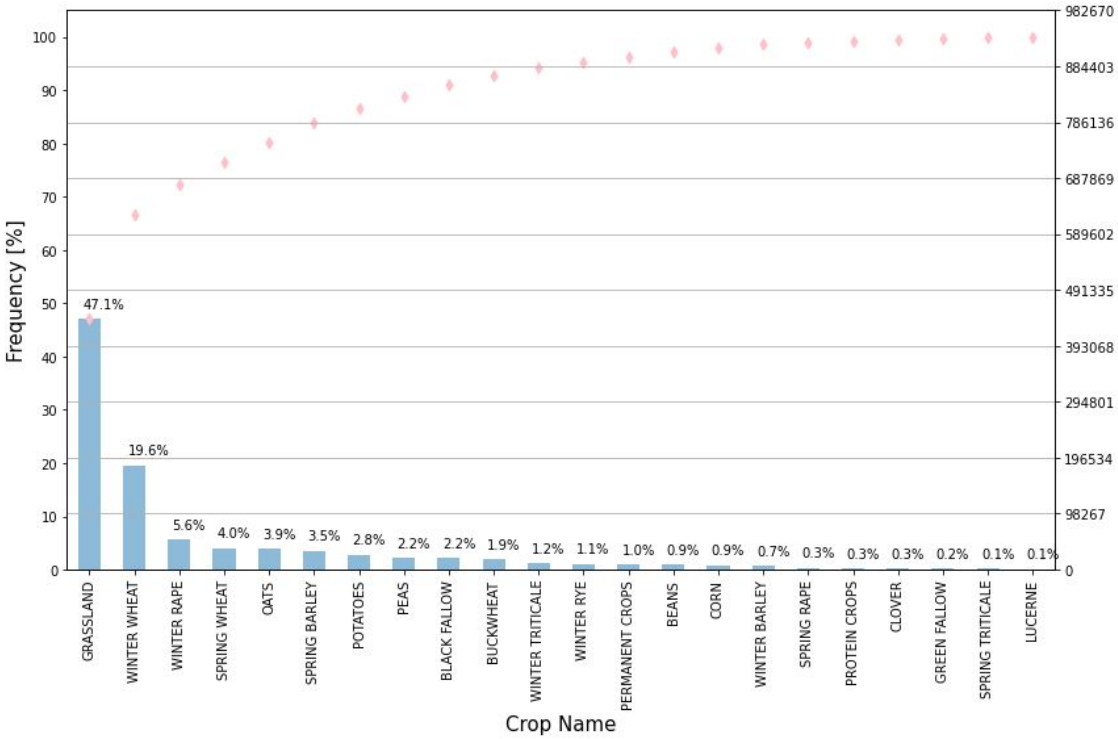


Figure 9: Lithuania (2022) – Distribution of Crops Predicted

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

Additionally, a crop taxonomy scheme (Figure 10) is employed in collaboration with agricultural experts from the NPA to facilitate crop classification. This taxonomy scheme organizes crops into higher-level categories, providing a useful and effective approach for identifying incorrect declarations and improving the classification of finer crop categories. By utilizing a higher-level taxonomy, the classification process of finer crop categories benefits from a more comprehensive understanding of the relationships and characteristics of different crop types. This approach enables the identification of misclassified or inaccurately declared crops, leading to improved accuracy and reliability in the crop classification results.

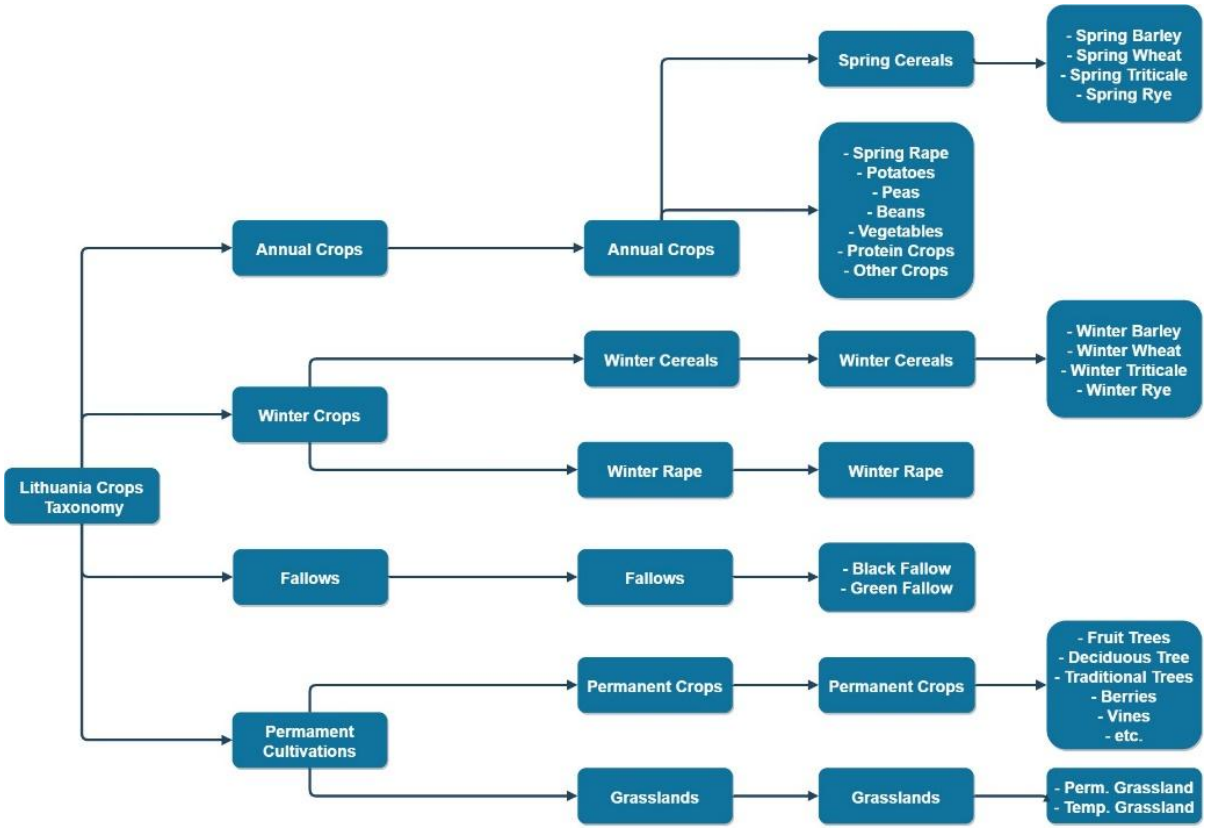


Figure 10: Crop Taxonomy for Lithuania

Within the BC1 of Lithuania, three data products have been developed to meet the specific monitoring needs. Throughout the ENVISION project, user requests and EU regulations have been taken into account, resulting in the development of specific EO services to address the current CAP requirements. The following is a list of the current data products providing:

1. Analytics on Vegetation and Soil Index Time-series:
    a. Minimum Soil Cover for Soil Erosion: Provides information on soil percentage and minimum soil cover to assess soil erosion risk.
    b. Runoff Risk Assessment for the Reduction of Water Pollution in Nitrate Vulnerable Areas: Helps assess the risk of runoff and water pollution in nitrate vulnerable areas.
    c. Harvest Event Detection: Identifies the occurrence of harvest events in agricultural areas.

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

   d. Stubble Burning Identification on arable lands: Detects and identifies instances of stubble burning.
2. Cultivated Crop Type Maps:
   a. Confirmation of applicants' declarations (based on GSAA): Utilizes *machine learning* techniques to classify crop types dynamically throughout the entire cultivation period or confirm the declared crop types.
   b. Smart Sampling for OTSC: Implements a traffic light system to assess the risk of incorrect declarations from applicants.
   c. Crop Diversification (CD) Compliance Map: Generates a map indicating compliance with *Greening I* requirements.
3. Grassland Mowing Events Detection:
   a. Grassland Activity Monitoring and Management: Detects and identifies mowing events in grassland areas, facilitating grassland management activity and compliance checks of farmers.

Please refer to document D3.3 for an additional analysis of these data products.

4.2.2 Methodology

4.2.2.1 Analytics on Vegetation and Soil Index Time-series

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

Input data

1. Satellite Data:
   a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC, 35VMC)
      i. Spectral bands (B01-B12)
      ii. Scene Classification (SCL)
   b. Sentinel-1 GRD (rel. orbits: 29, 58, 131, 160)
      i. Backscattering coefficients (VV-VH)
2. Auxiliary Data:
   a. Annual soil loss layer
   b. Rainfall erosivity factor layer
   c. Soil erodibility factor layer
   d. Slope length factor and slope steepness factor layer
   e. Crop and cover management factor layer
   f. Conservation supporting practices factor layer
   g. Slope - DEM

3. Paying agencies:
   a. Geospatial data (LPIS, GSAA): Parcels geometries and farmers declarations as a shapefile (updated when is necessary)

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

b. A lookup table for all the available crop type names, codes, families and CD ancillary info
c. Events Timestamps to fine-tune the algorithms (Stubble Burning, Harvest of arable land)
d. Agricultural Practices Descriptions – National CAP strategic Plans
e. Hydrographic Network

Output data

The product offers four distinct output components based on the specific delivery requirements of end users:

1. A shapefile for indicating bare soil percentage and minimum soil cover alerts for soil erosion.
2. A shapefile of Runoff Risk level map of parcels within nitrate vulnerable zones.
3. A shapefile for identifying stubble burning events on arable land, including dates of the events detected at the parcel level.
4. A shapefile for tracking harvest events on arable cultivations, including dates of the events detected at the parcel level.

Minimum Soil Cover for Soil Erosion

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. The identification of soil cover for a parcel takes into a series of masking rules most of them based on the Geospatial Soil Sensing System (GEOS3):

- 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, 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 calculation of the following band algebra and indices generation takes place: SAVI, NDVI, NBR2, B3-B2 and B2-B1 for each of the clear pixels.
- Finally, a binary mask is created based on the mutual fulfilment of the following conditions:

i. $0 < NDVI < 0.3$
ii. $0 < SAVI < 0.35$
iii. $B2 - B1 > 0$
iv. $B3 - B2 > 0$
v. $NBR2 < 0.35$

The thresholds in the equation can be adjusted and fine-tuned according to the desired level of strictness for soil masking. This flexibility allows for customization based on specific requirements and preferences.

- 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

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

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 after the desired date, excluding the null values. Specifically for Lithuania, the 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. Into this direction and taken into consideration the high percentage of clouds, the algorithm utilises cloudless images up to the end of January.

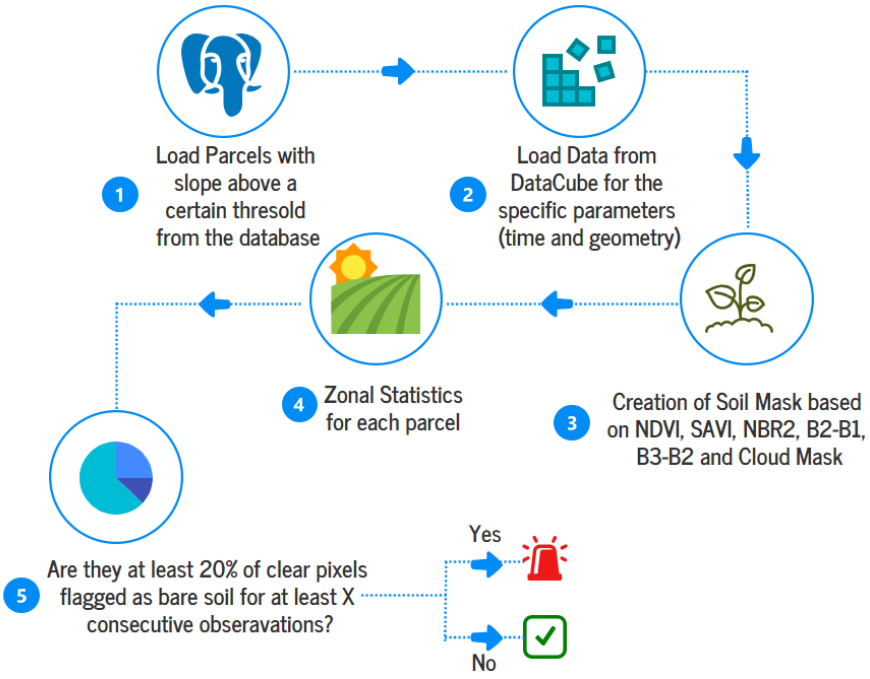


Figure 11: Workflow for the minimum soil cover detection

This module output shapefile contains the following fields:

- **UNIQUE_ID**: This field contains a unique identification number provided by NPA, serving as a distinct identifier for each parcel or record within the dataset.
- **APPL_ID**: It stores the ID associated with the applicant responsible for the specific parcel or record, providing a reference to the applicant's identity.
- **PIXELS_S2:** This field records the count of Sentinel-2 pixels considered during the analysis.
- **DECL_N:** Here, you can find the crop name declared by the applicant for the analyzed parcel. It represents the crop type as reported by the applicant.
- **DECL_C:** This field contains the unique ID or code corresponding to the crop declared by the applicant. It helps in associating a specific code with the declared crop.
- **DATE:** This field presents the predicted date of minimum soil cover. It is a crucial piece of information for understanding and managing soil conditions over time.
- **dSAVI:** This field measures the change in the SAVI between the detected date and its previous date as a confidence proxy.

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

Runoff Risk Assessment for the Reduction of Water Pollution in Nitrate Vulnerable Areas.

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.

Therefore, distance from every point of the 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 Revised Universal Soil Loss Equation [2] are the most widely used and accepted empirical soil erosion models. The later has the following equation:

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

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 6 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

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

Figure 12: Visualization of the run-off risk for a subset of parcels along with the water surfaces around them

This module output shapefile contains the following fields:

- **UNIQUE_ID**: This field contains a unique identification number provided by NPA, serving as a distinct identifier for each parcel or record within the dataset. It helps maintain data integrity and individual record identification.
- **APPL_ID**: It stores the ID associated with the applicant responsible for the specific parcel or record, providing a reference to the applicant's identity. This field links the analysis to the applicant.
- **PIXELS_S2**: This field records the count of Sentinel-2 (S2) pixels considered during the analysis. Sentinel-2 pixels are units of measurement from the Sentinel-2 satellite imagery, often used for land monitoring and assessment. It reflects the spatial resolution of the analysis.
- **DECL_N**: Here, you can find the crop name declared by the applicant for the analyzed parcel. It represents the crop type as reported by the applicant, providing insight into the intended land use.
- **DECL_C**: This field contains the unique ID or code corresponding to the crop declared by the applicant. It helps in associating a specific code with the declared crop, simplifying data management and analysis.
- **DISTANCE_MIN**: This field presents the minimum distance for water proximity in hectares. It indicates how close the parcel is to water bodies or water-related features.
- **RUSLE_MEAN**: This field presents the average RUSLE calculated from all the points included in the analysis.
- **RUNOFF_RISK**: This field represents the inferred risk based on water proximity and RUSLE parameters, as detailed in Table 6.

Stubble Burning Identification

The detection and mapping of burnt areas are crucial for environmental agencies and agricultural monitoring. Stubble Burning Identification is a key requirement for compliance monitoring, particularly for the Lithuanian paying agency. This task focuses on utilizing satellite imagery, specifically Sentinel-2 data, to detect stubble burning events in arable areas. The algorithm analyzes time series data,

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

including NBR, dNBR, slope of NBR, MIRBI, and NDWI, applying specific threshold conditions to identify potential burning events at the pixel level (Figure 13). This approach provides valuable insights for monitoring stubble burning activities and ensuring compliance in Lithuania's agricultural sector.

```
detect_burning_event_A

Parameters:
- time_series_nbr
- time_series_dnbr
- time_series_slope_nbr
- time_series_mirbi
- time_series_ndwi

Variables:
- burning_events

Algorithm:
for index = 1 to length(time_series_nbr) - 1:
    nbr = time_series_nbr[index]
    dnbr = time_series_dnbr[index]
    slope_nbr = time_series_slope_nbr[index]
    mirbi = time_series_mirbi[index]
    ndwi = time_series_ndwi[index]

    if (nbr < threshold_nbr) and
       (dnbr < threshold_dnbr) and
       (slope_nbr < threshold_slope_nbr) and
       (mirbi > threshold_mirbi) and
       (ndwi < threshold_ndwi):
        add index to burning_events

return burning_events
```

Figure 13: Stubble Burning Events Identification "Pseudocode A" for Lithuania Pilot

If at least 10% of the pixels within a parcel are identified as burned, the entire parcel is considered potentially burned. However, detecting burning events in Lithuania poses a significant challenge. Many of these events are often misidentified as simple cases of tillage or plowing due to similar behaviour observed in the aforementioned indices. To address this, a secondary routine has been developed to filter and retain only genuine burn cases. Through careful analysis of the region and reference to actual burn examples provided by the NPA, it has been observed that stubble burning events typically exhibit significant spectral discrepancies across the affected area and the evaluated pixels, with only a small portion of the parcel actually being burned. Therefore, a secondary routine focuses on identifying concentrated regions of burned pixels. This is achieved by calculating the mean and standard deviation of NBR. By checking that the mean value is lower than a specific threshold (*NBR_mean<th_1*), while

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

the standard deviation (*NBR_std>th_2*) remains considerably high, we can ensure the accurate identification of true stubble burning events in Lithuania.

Finally, the output is a map (shapefile) highlighting burned parcels, along with the event date, necessary for the compliance information based on NPA regulations.

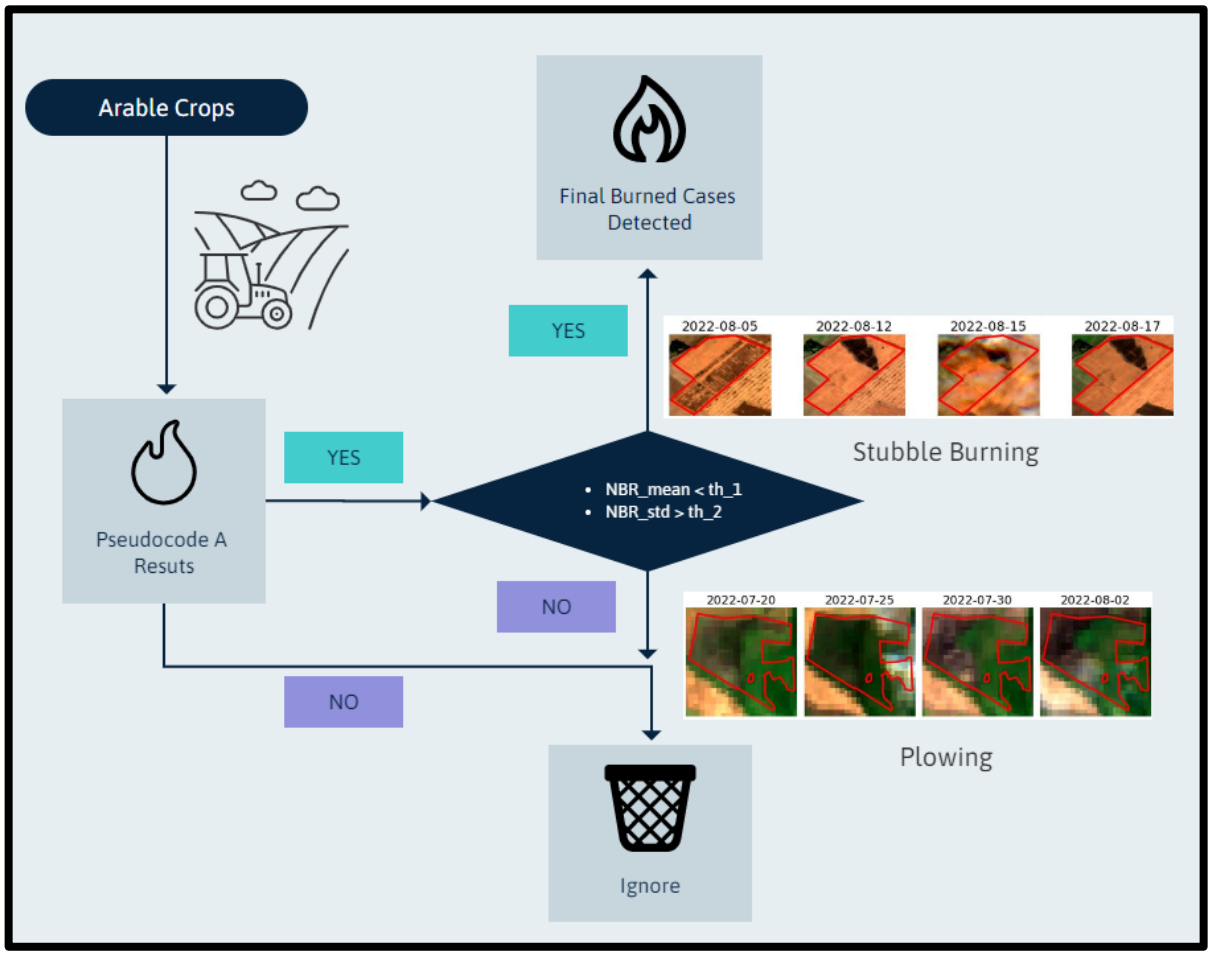


Figure 14: Total workflow of Stubble Burning Identification for Lithuanian BC

This module output shapefile contains the following fields:

- **UNIQUE_ID**: This field contains a unique identification number provided by NPA, serving as a distinct identifier for each parcel or record within the dataset.
- **APPL_ID**: It stores the ID associated with the applicant responsible for the specific parcel or record, providing a reference to the applicant's identity.
- **PIXELS_S2:** This field records the count of Sentinel-2 pixels considered during the analysis.
- **DECL_N:** Here, you can find the crop name declared by the applicant for the analyzed parcel. It represents the crop type as reported by the applicant.
- **DECL_C:** This field contains the unique ID or code corresponding to the crop declared by the applicant. It helps in associating a specific code with the declared crop.
- **DATE:** This field presents the predicted date of minimum soil cover. It is a crucial piece of information for understanding and managing soil conditions over time.

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

- **dNBR:** This field measures the change in the NBR between the detected date and its previous date as a confidence proxy.

Harvest event detection

The agricultural monitoring practices prioritize the detection of harvest events, particularly for arable land parcels designated for Ecological Focus Area (EFA) practices as per the GSAA. The objective of this task is to identify the parcels that have undergone harvesting within the cultivation period and determine the compliance level based on the regulations of the Lithuanian paying agency. To achieve this, a harvest detection algorithm is employed (Figure 16), leveraging a combination of VIs derived from Sentinel-2 SITS. The algorithm generates a map that highlights the parcels where harvest events have been detected, along with the corresponding dates of these events. The harvest detection algorithm incorporates specific conditions based on significant changes observed in key indices, such as NDVI, NDMI, PSRI, and BSI, to enhance the accuracy of predictions. These indices have been selected after thorough analysis to capture substantial vegetation loss and soil emergence. The thresholds for these indices have been calibrated to optimize the performance of the algorithm. By employing this algorithm, prediction dates for harvest events can be obtained throughout the cultivation period. It is important to note that this data product focuses exclusively on arable land, with the removal of grasslands (accounting for 45% of Lithuania's crops), fallow land fields, permanent cultivations and greenhouses from the dataset. This step ensures the elimination of noise and enhances the accuracy of the harvest event detection process. Furthermore, by monitoring the entire cultivation period, the algorithm can identify more than one harvest event per parcel. This capability can be used to infer the existence of secondary, auxiliary cultivations or catch-crops following the main crop's harvest. Finally, the output is a map (shapefile) highlighting harvested parcels, along with the event date, necessary for the compliance information based on NPA regulations.

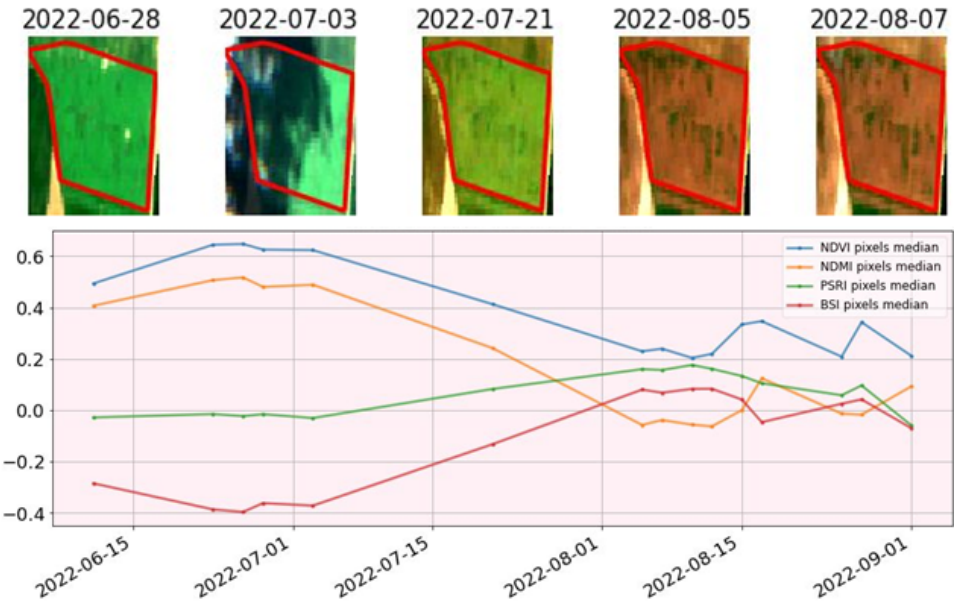


Figure 15: Harvest Event Detection example case (Lithuania)

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

```
detect_harvest_event

Parameters:
- time_series_ndvi
- time_series_ndmi
- time_series_psri
- time_series_bsi

Variables:
- harvest_events

Algorithm:
for index = 1 to length(time_series_ndvi) - 1:
    ndvi = time_series_ndvi[index]
    ndmi = time_series_ndmi[index]
    psri = time_series_psri[index]
    bsi = time_series_bsi[index]

    if (ndvi[t-1] - ndvi[t] > ndvi_diff_th) and
       (ndvi[t-1] - ndvi[t] > d * ndvi_rate) and
       (ndmi < ndmi_th) and
       (ndmi[t-1] - ndmi[t] > ndmi_diff_th) and
       (ndmi[t-1] - ndmi[t] > d * ndmi_rate) and
       (psri > psri_th) and
       (psri - psri[t-1] > psri_diff_th) and
       (bsi > bsi_th) and
       (bsi - bsi[t-1] > bsi_diff_th):
        add index to harvest_events

return harvest_events
```

Figure 16: Harvest Events Detection "Pseudocode" for Lithuania Pilot

This module output shapefile contains the following fields:

- **UNIQUE_ID**: This field contains a unique identification number provided by NPA, serving as a distinct identifier for each parcel or record within the dataset.
- **APPL_ID**: It stores the ID associated with the applicant responsible for the specific parcel or record, providing a reference to the applicant's identity.
- **PIXELS_S2:** This field records the count of Sentinel-2 pixels considered during the analysis.
- **DECL_N:** Here, you can find the crop name declared by the applicant for the analyzed parcel. It represents the crop type as reported by the applicant.
- **DECL_C:** This field contains the unique ID or code corresponding to the crop declared by the applicant. It helps in associating a specific code with the declared crop.
- **N_EVENTS:** This field contains the total number of harvest events detected through the evaluated period.
- **DATE_1/DATE_2/DATE_3:** This field presents the predicted dates of harvest events. It is crucial for understanding and managing arable fields' practices over time.
- **dNDVI_1/dNDVI_2/dNDVI_3:** This field measures the change in the NDVI between the associated detected date and its previous date per event as a confidence proxy.

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

4.2.2.2 Cultivated Crop Type Maps (CCTM)

The Cultivated Crop Type Maps (CCTM) product used in the Lithuanian pilot case combines data from Sentinel-1 and Sentinel-2 satellites to generate maps of cultivated crop types. The CCTM product includes the following features:

- Confirmation of applicants' declarations (based on GSAA): Utilizes machine learning techniques to classify crop types dynamically throughout the entire cultivation period or confirm the declared crop types.
- A traffic light alert system to identify potential false declarations, using smart sampling algorithms
- Assessment of crop compliance with the Greening-1 rule for Crop Diversification, ensuring adherence to regulations.

Input data

More specifically we utilize the following data:

1. Satellite:
    a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC, 35VMC)
        i. Spectral bands (B01-B12)
        ii. Scene Classification (SCL)
    b. Sentinel-1 GRD (rel. orbits: 29, 58, 131, 160)
        i. Backscattering coefficients (VV-VH)
2. Paying agencies:
    a. Geospatial data (LPIS, GSAA): Parcels geometries and farmers declarations as a shapefile (updated when is necessary)
    b. A lookup table for all the available crop type names, codes, families and CD ancillary info (Table 7)
    c. Agricultural Practices Descriptions – National CAP strategic Plans

Table 7: Lithuania Look-up Table sample schema

ID	CROP CODE	CROP NAME	CROP FAMILY	EAA	AL	PGrass	TGrass	Fallow	Cwater	Protein	Cother
1	PDJ	BLACK FALLOW	FALLOW	1	1	0	0	1	0	0	0
2	MIV	SPRING BARLEY	ANNUAL CROPS	1	1	0	0	0	0	1	0
3	KVZ	WINTER WHEAT	WINTER CEREAL	1	1	0	0	0	0	0	0
4	GPZ	PASTURE	GRASSLAND	1	1	0	1	0	0	0	0
5	RST	WALNUT	PERMANENT CROPS	1	0	0	0	0	0	0	0
6	RAZ	WINTER RAPE	WINTER RAPE	1	1	0	0	0	0	0	0

Output data

The product is providing an output shapefile consist of:

1. Dynamic crop type maps over the registered parcels for every new or group of new Sentinel acquisitions.

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

2. Traffic light maps over the registered parcels for smart sampling of on-the-spot inspections and early alert of the users.
3. Crops Diversification (Greening-1) compliance over the registered parcels at the end of the cultivation period.

Dynamic Crop Type Classification

The primary objective of this product is to generate crop type maps by classifying a set of polygons provided by the PAs. Specifically, we produce multiple crop type maps, for every group of new sentinel acquisition, which are acquired approximately every 15 days. To achieve this, we iteratively employ supervised machine learning (ML) algorithms and make the fundamental assumption that the majority of farmers' declarations correspond to the actual crop types, even though this may not be true for all cases. The Random Forest (RF) [3] algorithm is selected due to its widespread use in similar operational scenarios and its ability to provide fast and efficient results. It is important to note that alternative algorithms can also be considered for the classification task.

An ML pipeline is trained and provides results using a range of features derived from Sentinel-2 bands (B02-B12), Sentinel-1 backscatter coefficients (VV, VH), and calculated vegetation indices (e.g., NDVI, NDMI, PSRI, etc., as described in the relevant section of D3.5) extracted until the current date of routine's execution. The training of the algorithm is performed with parcel time series (mean series of all the pixels included in the referred geometries). On the contrary, during the inference phase, the results are provided (i) initially at the pixel level and then aggregated at the parcel level for small parcels (less than 10 clear s2 pixels after buffering, usually an area of 0.1 hectares) or (ii) directly at the parcel level for larger ones. The selection of the training data time series for the RF is selected based on the presence of at least 10 clear pixels per unique parcel to ensure representative information from each observation. The remaining cases are exclusively used for inference purposes. Cases with not any clear S2 pixel after buffering (approximately 0.01 hectares), are excluded from the model's estimation.

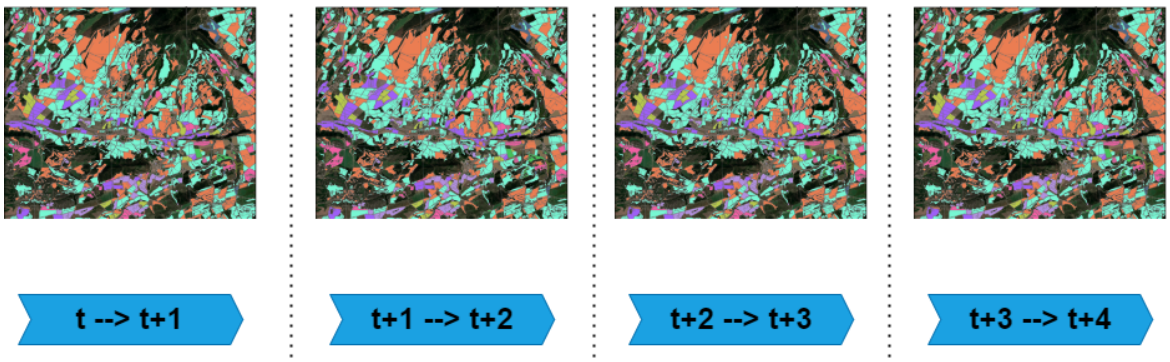


Figure 17: Multi-temporal Crop Type Mapping

Addressing the high class imbalance in the dataset, we focus on significant crop type categories that represent more than 0.1% of the total cases. Among these, crop types with less than 1000 samples are resampled using the Synthetic Minority Oversampling Technique (SMOTE) [4]. The classification weights parameter, inversely proportional to crop frequencies, is incorporated into the algorithm to handle class imbalance. Moreover, to further improve the accuracy of the predicted maps, a hierarchical classification scheme is designed, exploiting information from higher hierarchical levels where discrimination is definitely easier. Stacking ensembles are also employed to train different models on different samples extracted from the initial population and aggregate the results at the end.

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

These strategies contribute to enhancing the accuracy and robustness of the crop type maps generated by the ML algorithm.

Given that Lithuania has an area of approximately 65,300 km² and over 1.1 million cases (e.g. billions of pixels) to analyze, performing parcel-based or pixel-based analysis (when is necessary) and classification inference on such a large scale can be computationally expensive. To address this, parallel DataCube indexing routines are employed to significantly reduce the time required for extracting a national-scale crop type map. The indexing process involves dividing the entire region into smaller rectangular bounding boxes of a predefined size. These bounding boxes are then sequentially scanned, and their respective pipelines are executed in parallel in order to distribute and optimize the computational workload (see Figure 18). This parallel processing approach allows for individual box processing, providing results much faster for each pixel and subsequently each parcel within the bounding box.

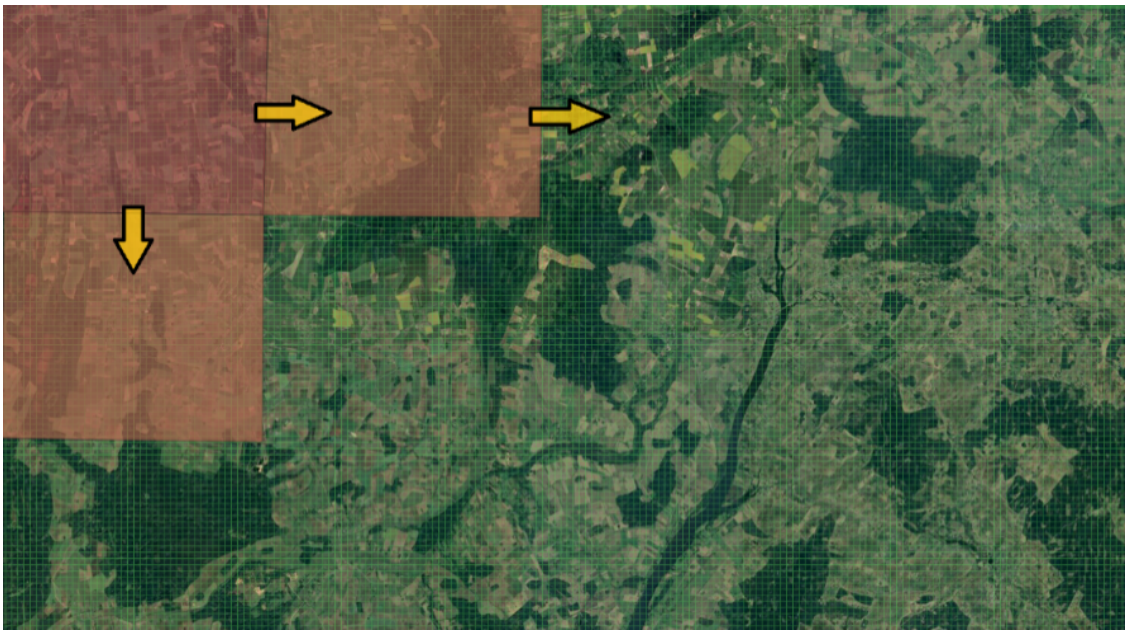


Figure 18: Shifting window visualization for inference of the results

The parcel-level produced maps, along with the associated confidence level, obtained from each iteration, will serve as input for the subsequent Traffic Light Alert System and the Crops Diversification Compliance Map services. Figure 23 illustrates a general scheme of the entire pipeline.

Hierarchical Classification

The general concept of the hierarchical scheme is to exploit the information coming from higher level of taxonomy between the various crops since classes at these stages are easier to be distinguished between them. That way, classification models trained on the different level of taxonomy (land use and crop family) are able to impose extra knowledge to the finest level and construct classification models for the various tier level. Furthermore, hierarchical classification strategies are suggested as a solution to alleviate the problem of imbalanced classes. Figure 19 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 RF models are trained. Each higher-level model provides its

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

outputs to the rest of the lower-level ones (e.g., predictions from L_1 model are provided as input to the training of L_2 and L_3 models, while predictions from L_2 are provided as input into the lowest level L_3). Finally, the L_3 model is applied to the entire training dataset to acquire predictions on the crop type level. Overall, this hierarchical approach allows for the integration of information from multiple levels, improving the classification performance and enabling more accurate identification of crop types.

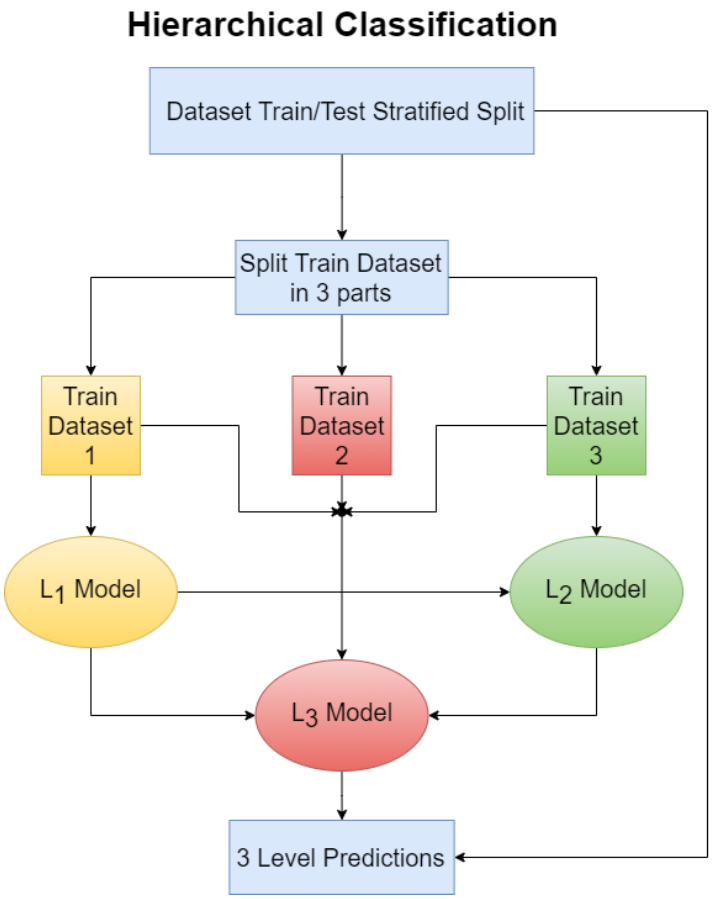


Figure 19: Hierarchical classification scheme

Stacking Ensembles

The proposed stacking ensemble methodology involves training k different base hierarchical models (for our case k is chosen to be equal with 3), each trained on a different individual sample from the initial dataset. The goal is to take advantage of the diversity of these models and their individual strengths in capturing different aspects of the data. During the training phase, each base hierarchical model receives a subset of the initial dataset and is trained independently. These models are designed to exploit the hierarchical structure of the crop classification task, utilizing information from higher-level taxonomy to inform predictions at lower-level crop types. Once the base hierarchical models are trained, the stacking ensemble combines their predictions to make final predictions. The ensemble uses a majority voting approach, where each base model casts its prediction for a given sample, and the final prediction is determined by the majority vote among the models.

By aggregating the predictions of the base hierarchical models through majority voting, the stacking ensemble can benefit from the collective wisdom of these models, capturing a more comprehensive representation of the data. This approach enhances the overall accuracy and robustness of the final

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

predictions, as it takes into account the diverse perspectives provided by the individual base models. Overall, the stacking ensemble methodology trained on k different base hierarchical models, with final predictions determined by majority voting, allows for a powerful integration of the models' insights, resulting in improved performance and more reliable predictions for crop classification.

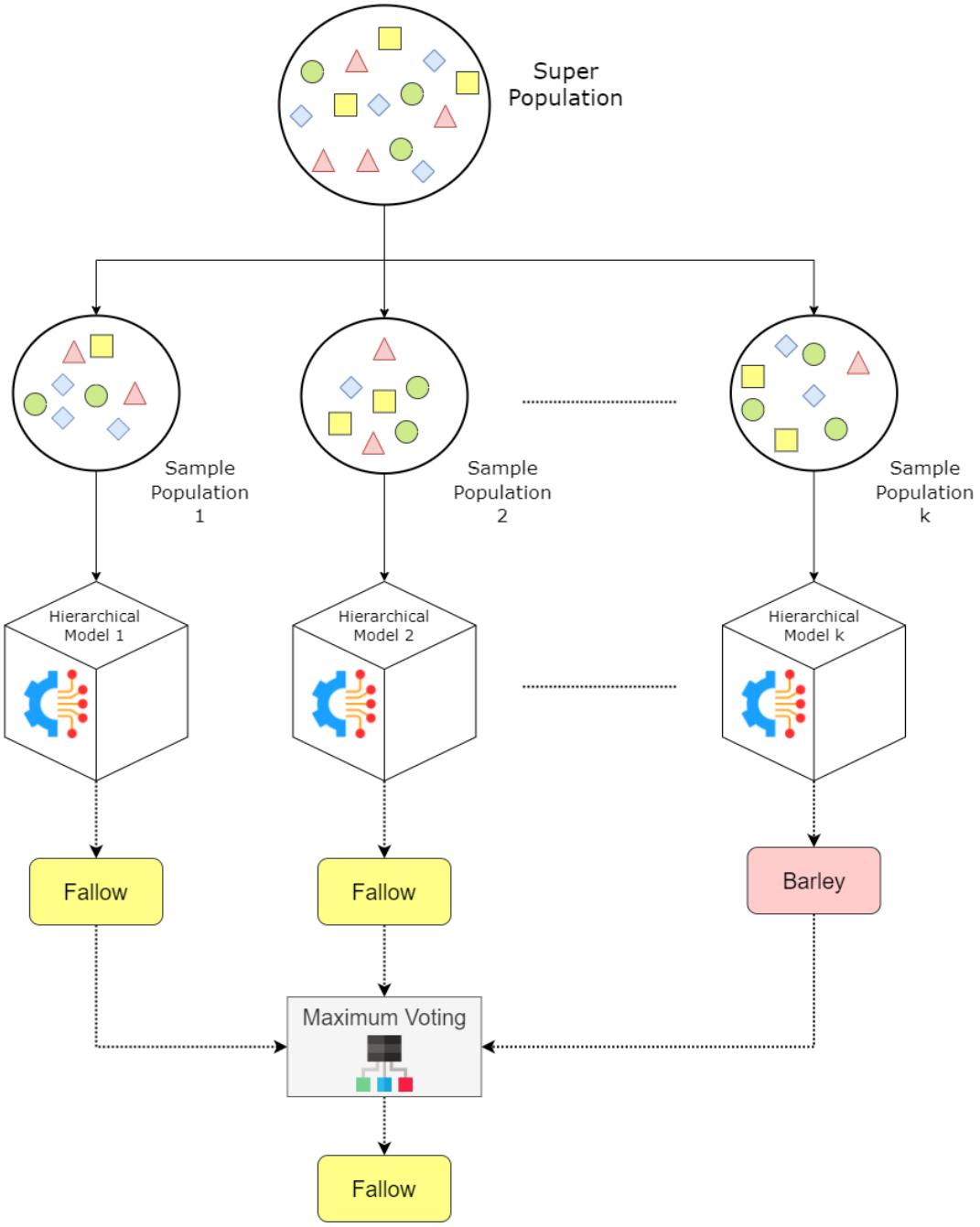


Figure 20: Stacking Ensembles Scheme

Confirmation of Declaration:

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

A declaration's conformity assessment is determined based on the alignment between the crop type declared by the farmer and the classification results. This classification system helps determine whether the parcel's crop type declaration matches the model's predictions with the necessary level of confidence, making it clear when the classification is reliable, unreliable, or falls in between. Here's how it works:

- "Conform" Assessment: A parcel is considered "conform" when the crop type declared by the farmer matches one of the two highest-confidence predictions made by the classification model. The declaration that agrees with the classification should have a confidence level above a specified critical threshold (this is usually set to 0.1).
- "Not Conform" Assessment: A parcel is marked as "not conform" when both of the two major predictions from the classification model do not match the farmer's declaration. In this case, both of the non-matching predictions should have confidence levels above the specific critical threshold.
- "Ambiguous" Assessment: In all other cases, the prediction is characterized as "ambiguous." This means that if the farmer's declaration doesn't align with either of the two highest-confidence predictions, and the conditions for "not conform" aren't met, the assessment is considered ambiguous.

Smart Sampling for OTSC (traffic light alert system)

Smart sampling is a sophisticated algorithm developed by NOA [5] 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 a "*traffic light" alert 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 (2 first crop categories predicted). 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) during the ENVISION project, we can estimate that the percentage of false declarations in Lithuania is roughly 3%, and the vast majority of

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

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.
- **Moderate 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.
- **No Risk 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 beginning of the farmer application submissions.

In Figure 21 below, we can see how the percentage of correctly detected cases (precision score) and the percentage of the actual wrongly declared cases predicted (recall score) is increasing during the cultivation period. This progressive enhancement reflects the iterative nature of the methodology and the continuous refinement of the crops classification process throughout this period, resulting in improved reliability and effectiveness in capturing accurate crop information.

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

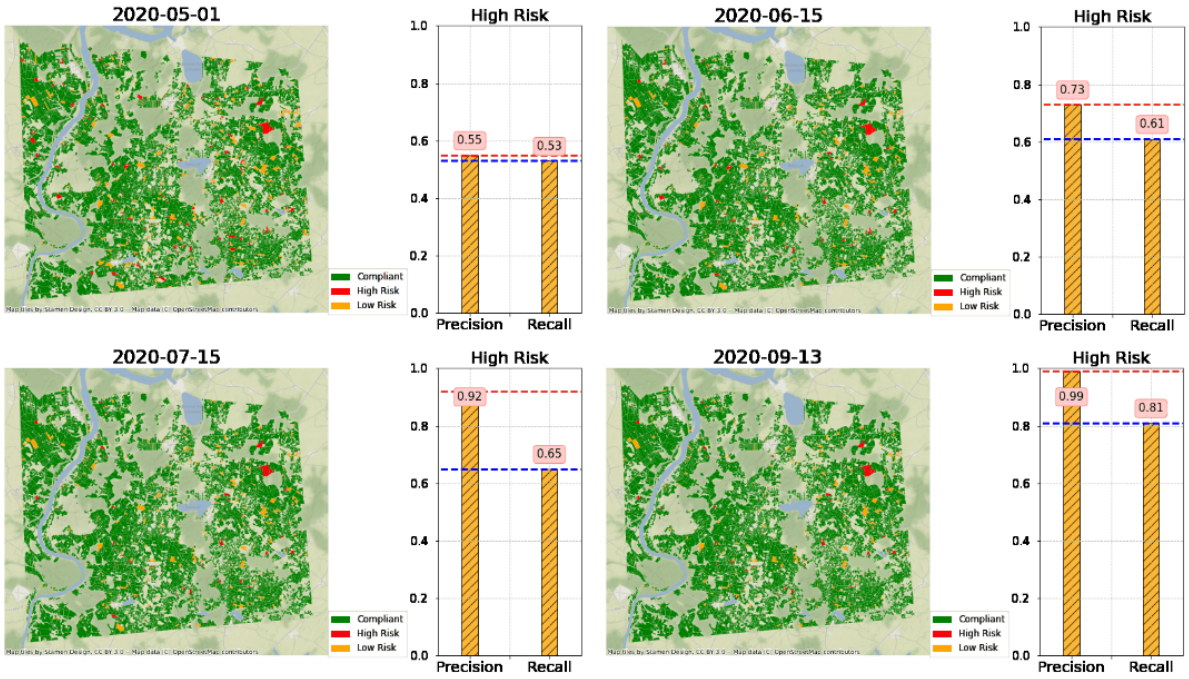


Figure 21: Declarations Traffic Light Alert Maps using NOA's Smart Sampling. Precision and Recall are progressively improved throughout the cultivation period.

Crop Diversification (CD) Compliance Map

The Crop Diversification exploits the Land Parcel Information System (LPIS) and the Geospatial Aid applications (GSAA). 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.

The crop codes used for the service correspond to the declarations, but they will be replaced with predicted ones from the Crop Type Mapping (CTM) module if flagged as potentially incorrect by the traffic light system. The results will be visualized via a shapefile map (Figure 22), 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.

Table 8: Crops Diversification explanatory table

-	Category	Description	CD Rules
1	Non-Compliant	Applicant's failure to meet the "greening-1" requirements.	-

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

2	Compliant (Category 1)	TAL[1] between 10 and 30 ha.	• At least 2 different crop types • Main crop ≤ 75% TAL
3	Compliant (Category 2)	TAL greater than 30 ha.	• At least 3 different crop types • Main crop ≤ 75% TAL • 2 main crops ≤ 95% TAL
4	Compliant (Category 3)	Temp. Grass and Fallow greater than 75% of TAL.	Main crop ≤ 75% of remaining AL
5	Exemption (Category 1)	TAL less than 10 ha.	No crop diversification required
6	Exemption (Category 2)	Temp. Grass and Fallow greater than 75% of TAL and remaining AL less than 30 ha.	No crop diversification required
7	Exemption (Category 3)	Perm. Grass, Temp. Grass and Cwater greater than 75% of EAA and remaining AL less than 30 ha.	No crop diversification required
8	Exemption (Category 4)	Cwater[2] = TAL	No crop diversification required

Figure 22: Crop Diversification Compliance Map

[1] TAL = Total Arable Land

[2] Cwater = Crop Under Water

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

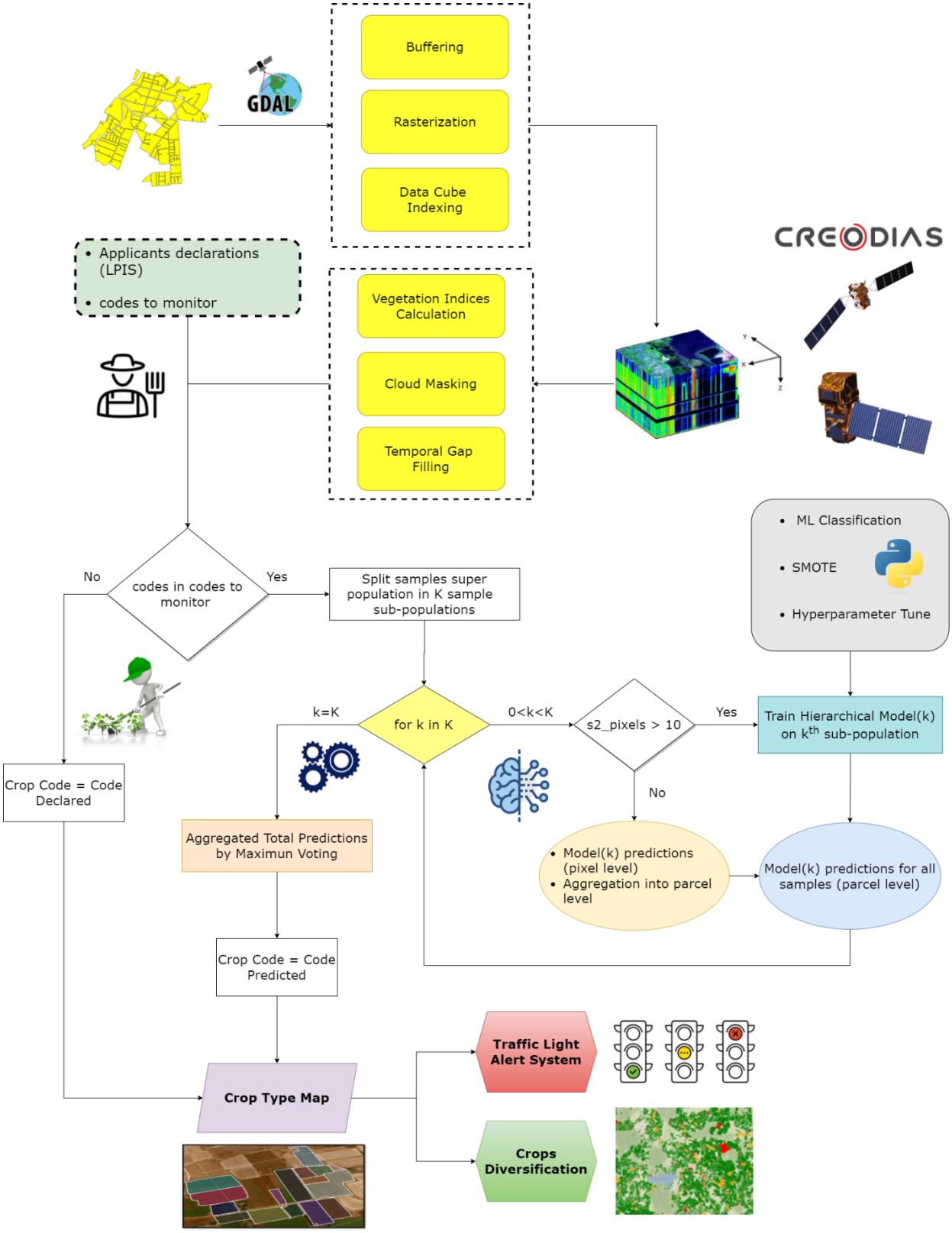


Figure 23: Cultivated Crop Type Maps General Scheme

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

The result shapefile of CCTM products contain the following fields:

- **UNIQUE_ID**: unique_id number provided by NPA.
- **APPL_ID**: id of the applicant.
- **D_AREA**: field's area declared by applicant (in hectares).
- **C_AREA**: field's area calculated by NOA (in hectares).
- **PIXELS_S2**: number of S2 pixels included into the analysis.
- **PIXELS_S1**: number of S1 pixels included into the analysis.
- **DECL_N**: crop name declared.
- **DECL_C**: id of the code declared.
- **PRED_N1**: crop name predicted (1st most confident prediction).
- **PRED_C1**: id of the code predicted.
- **CONF_1**: prediction's confidence.
- **PRED_N2/PRED_C2/CONF_2:** results for the 2nd most confident prediction.
- **PRED_N3/PRED_C3/CONF_3**: results for the 3rd most confident prediction.
- **CONFIRM**: confirmation of declaration.
- **ALERT**: level of alert - false declaration (based on “traffic light alert system”).
- **CD**: applicant’s crop_diversification category based on Table 8.

4.2.2.3 Grassland Mowing Events Detection

Grassland Mowing Detection is another ENVISION data product which consists of 2 individual pre-processing steps:

- Data Fusion
- Mowing Events Detection Algorithms and Management of Activity (compliance check)

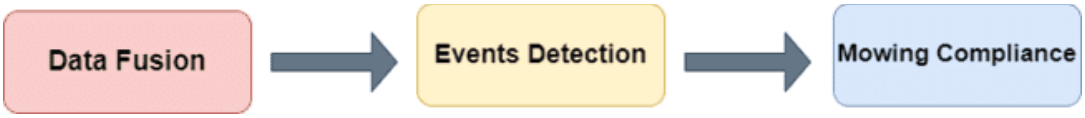


Figure 24: Mowing Events Detection product scheme

The Data Fusion (DFS) workflow combines Sentinel-1 and Sentinel-2 data in order to "increase” the availability of cloud-free observations. Optical sensors, like Sentinel-2, are sensitive to clouds resulting in gaps in time series data. On the other hand, Sentinel-1 data are not affected by clouds and can provide valuable information in the presence of cloud obstructions. Hence, we exploit Sentinel-1 and Sentinel-2 complementary nature to generate cloud-free Sentinel-2 time series through a sophisticated Deep Neural Network architecture**.** The cloud-free time series are then integrated into the ENVISION DataCube and utilized as input for various services, including 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.

Then, the Grassland Mowing Events Detection module exploits the newly constructed data obtained from the Data Fusion service, utilizing either Deep Learning algorithms (*if label data are available*) or a threshold-based alternative approach. When labeled data is provided, a Deep Learning algorithm is employed to analyze the satellite data and identify specific patterns and features associated with

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

grassland mowing events. The algorithm is learning from the labeled examples to make predictions and detect the time ranges where mowing events have occurred. In the absence of labeled data, a threshold-based approach is available, similar to the methodology employed for harvest events detection. By defining specific thresholds or criteria based on relevant parameters, such as steep changes in artificially created NDVI. Both approaches aim to accurately detect and track grassland mowing events.

This information provides valuable insights into the temporal dynamics of grassland activity, enabling effective monitoring and management of agricultural practices. This information can be used from the PAs to estimate the compliancy of the farmers with regards to the national mowing regulations (Table 5) or for the quantification of the grasslands activity for sustainable management. Similarly, 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 25).

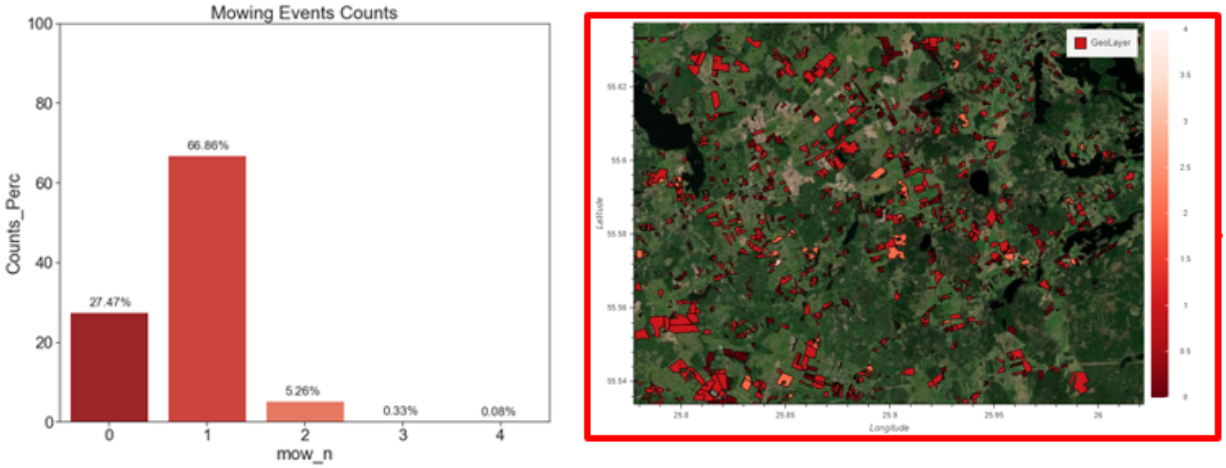


Figure 25: Mowing Summary Statistics of a Predefined Area

Input data

1. Satellite:
    a. Sentinel-2 L2A (tiles: 34UEG, 34UFE, 34UFF, 34UFG, 34UGE, 34VEH, 34VFH, 35ULA, 35ULB, 35ULV, 35UMA, 35UMB, 35VLC, 35VMC)
        i. Spectral bands (B01-B12)
        ii. Scene Classification (SCL)
    b. Sentinel-1 GRD (rel. orbits: 29, 58, 131, 160)
        i. Backscattering coefficients (VV-VH)
2. Paying agencies:
    a. Geospatial data (LPIS, GSAA): Parcels geometries and farmers declarations as a shapefile (updated when is necessary)
    b. 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.

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

c. Grassland's polygons to be check after CCTM as a shapefile
d. Agricultural Practices Descriptions – National CAP strategic Plans

Output data

The product is providing an output shapefile of:

1. Dynamic Events Map of grassland mowing detection per parcel encapsulating all the extracted information regarding the detected events, their confidence levels.

Data Fusion

The main goal of this product is to address the issue of cloudy observations in Sentinel-2 images, taking into consideration information coming from Sentinel-1 and cloudless Sentinel-2 cases. To achieve this, a sophisticated Deep Neural Network (DNN) architecture has been developed, incorporating the latest advancements in the field. The aim is to generate complete NDVI time series without any gaps [6]. The DNN module is trained and applied to each pixel individually. The inputs of the model are the VV, VH and VV/VH backscatter coefficients, as well as the available cloud-free NDVI observations up to the current point in time. To align the temporal information between Sentinel-1 and Sentinel-2, the Sentinel-2 data are shifted and assigned to the nearest transition date of the Sentinel-1 acquisitions. This process creates a consistent time step, ensuring that the Sentinel-2 measurements correspond to the closest Sentinel-1 observations within a 3-day window. To train the DNN model, dense NDVI time series are temporal resampled using daily linear interpolation and then smoothed. This preprocessing step aims to minimize any deviation from the actual ground truth caused by missing values.

To simulate the presence of cloudy observations and their temporal distribution, the NDVI time series used as inputs undergo an artificial hiding process. Random time steps are selected and their values are replaced with a value of -1, indicating missing data. These masked values are then handled within the model. The model takes into account the available Sentinel-2 observations from adjacent time steps and the current Sentinel-1 SAR observations to infer the missing values and generate the final complete NDVI time series. The architecture of the model, as shown in Figure 26, consists of a series of 1-D convolutional and encoder-decoder blocks. The initial convolution module aims to minimize noise in each band of the input data (NDVI_input, SAR VV-VH). It includes masking, convolution, and pooling layers that work in parallel to process the data. The encoder-decoder module, composed of LSTM layers, focuses on identifying temporal correlations and patterns among the different time steps. This module synthesizes the final continuous NDVI output by capturing and incorporating the temporal information. By combining convolutional and encoder-decoder modules, the model can effectively handle the missing data caused by cloud cover, enhance the quality of the NDVI time series, and provide a more comprehensive and continuous representation of vegetation dynamics.

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

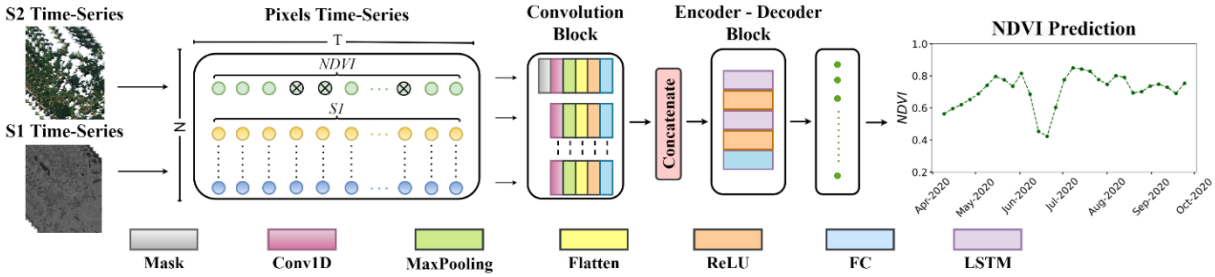


Figure 26: Data Fusion DNN architecture

Overall, at the end the model will be able to provide results for every pixel individually inside a parcel geometry to 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.

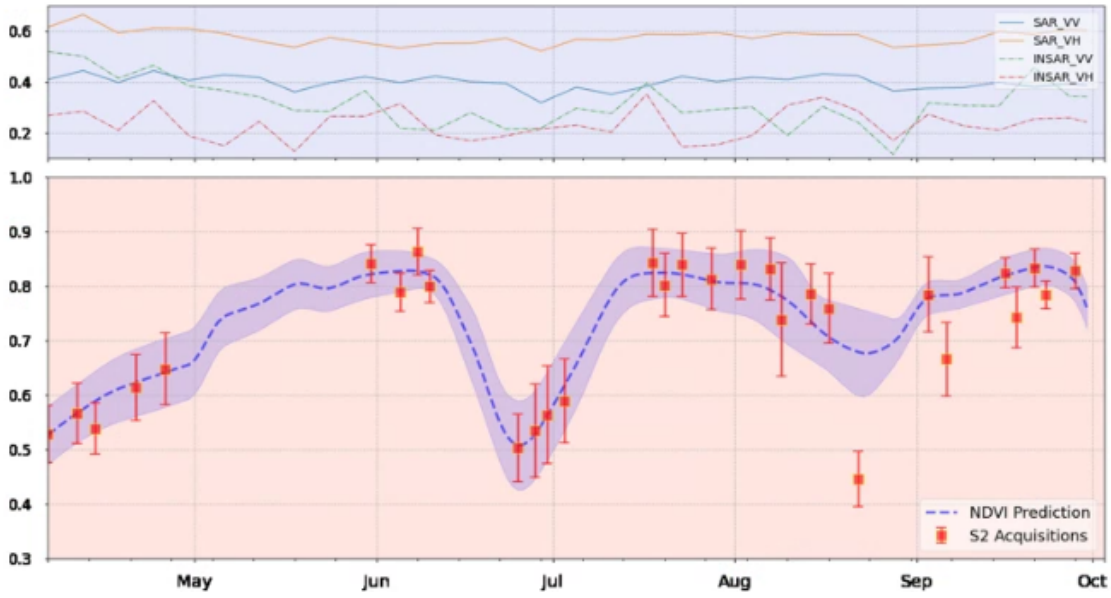


Figure 27: NDVI reconstruction over time exploiting S1 data and the available S2 measurements

Mowing Events Detection Algorithm (Deep Learning Approach)

The grassland mowing events detection algorithm utilizes Sentinel-1 and Sentinel-2 images to monitor changes over time and detect mowing events in grassland areas. The algorithm is designed to identify abrupt changes in vegetation patterns that are indicative of mowing activities. For this purpose, a novel deep learning architecture, similar to the one used for the Sentinel-1/Sentinel-2 fusion (Figure 26), is implemented. The algorithm takes as input the newly created NDVI time series with a fixed time step and aims to identify the specific 6-day timeframe during which a mowing event occurred (Figure 28).

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

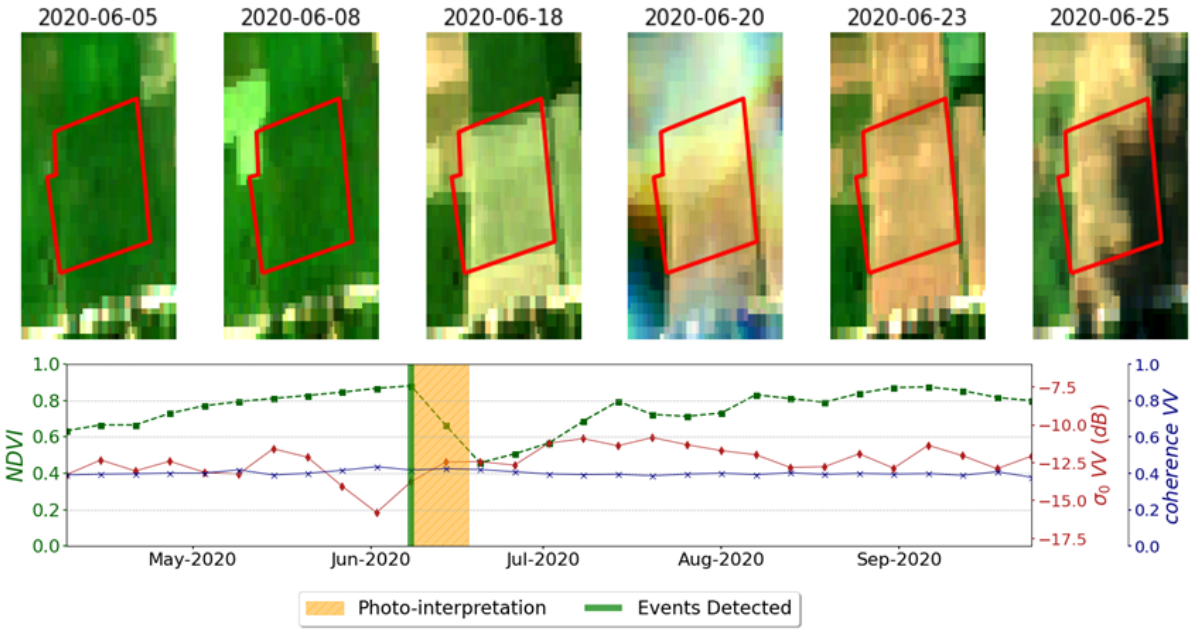


Figure 28: Mowing Events Identification

In the case of Lithuania, we were fortunate to have access to labels provided by the NPA, such as mowing event timestamps. To facilitate the training process, the provided labels were transformed into binary target values consisting of 0s and 1s. The resulting binary series had a size corresponding to the number of evaluated dates. The value 1 corresponds to the closest timestamp of the actual mowing event annotated, allowing the algorithm to effectively learn patterns associated with mowing practices and predict the occurrence of an event at each specific time instance. By leveraging the power of deep learning and utilizing labelled data, the algorithm is able to dynamically detect and track mowing events with each new acquisition of Sentinel-1 and Sentinel-2 images. At the end, the algorithm produces binary time series with values of 0s and 1s for each evaluated grassland pixel, representing the absence or presence of a predicted mowing event, respectively. Additionally, it provides probability logits that indicate the likelihood of a mowing event occurring at each time instance.

Mowing Events Detection Algorithm (threshold-based Approach)

In cases where training labels for mowing events are not available (e.g. Lighthouse customer case in Flanders), an alternative threshold-based approach, initially designed for the SEN4CAP project [7], is modified and implemented. This approach utilizes the reconstructed NDVI time series to detect significant and sharp decreases between consecutive NDVI observations. The algorithm analyzes the difference in NDVI values and the time elapsed between the capture dates (expressed as Days of Year, DOY) of two consecutive observations. By setting specific thresholds for the NDVI difference (*th*) and a minimum decreasing rate (*r*), potential grassland mowing events can be identified.

$$NDVI_t - NDVI_{t-1} \geq th$$
$$NDVI_{t-1} - NDVI_t \geq r \cdot d$$
$$d = DOY_t - DOY_{t-1}$$

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

This threshold-based algorithm has been widely used for mowing event detection and is considered a reliable method. A recent study by Devroey et al. (2022) [8] evaluated its performance in multiple European countries for the year 2019. In the Netherlands, the algorithm achieved an f1-score of 79% over 150 parcels. However, in Lithuania, the evaluation on 118 cases showed a lower f1-score of 60%, primarily due to the lower precision (49%) of the algorithm in identifying true mowing events.

Aggregated Results

The analysis is performed on each pixel individually and aggregated statistics is used to provide us with a representative level of confidence regarding the extent and the exact time instance that a mowing event took place for each parcel (Figure 29). Taking into consideration the type of grassland and the time range that a mowing event is detected, logits probabilities are used to prioritize inspections and quantify the risk of non-compliance for every evaluated field. Finally, in order to reduce computational cost, we employed the parallel Datacube indexing processing chain, using smaller rectangular bounding boxes as described in CCTM section (Figure 18).

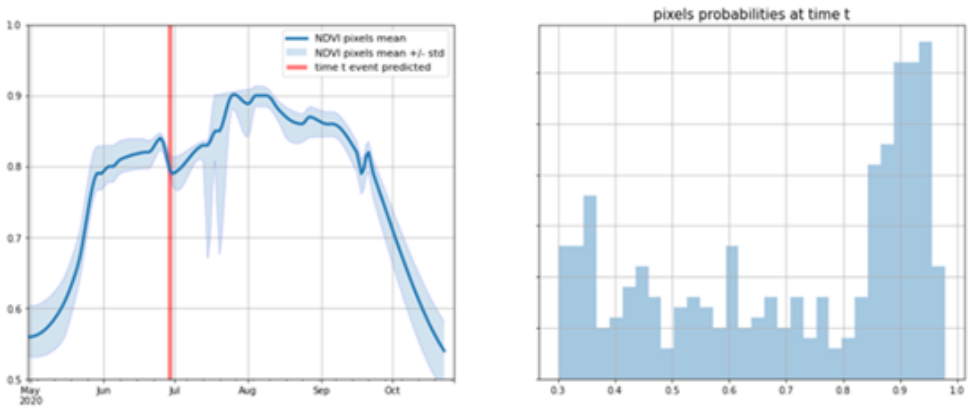


Figure 29: From pixel results to parcel decision

Implementing monitoring techniques brings us closer to achieving exhaustive monitoring of grassland mowing events. In the case of Lithuania, it is anticipated that the percentage of non-compliance with mowing regulations will be approximately 5%. The national regulations require farmers to complete all mowing activities by August (prior to 2023) or by September (from 2023 onwards). By accurately detecting and tracking mowing events, the monitoring system enables better oversight and enforcement of these regulations, contributing to effective grassland management practices.

Finally, considering that parcels are usually mowed gradually through time [9], 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 30). That way we can respond directly to many future CAP requirements related to the proportion of the mowed area with respect to the total parcel area declared.

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

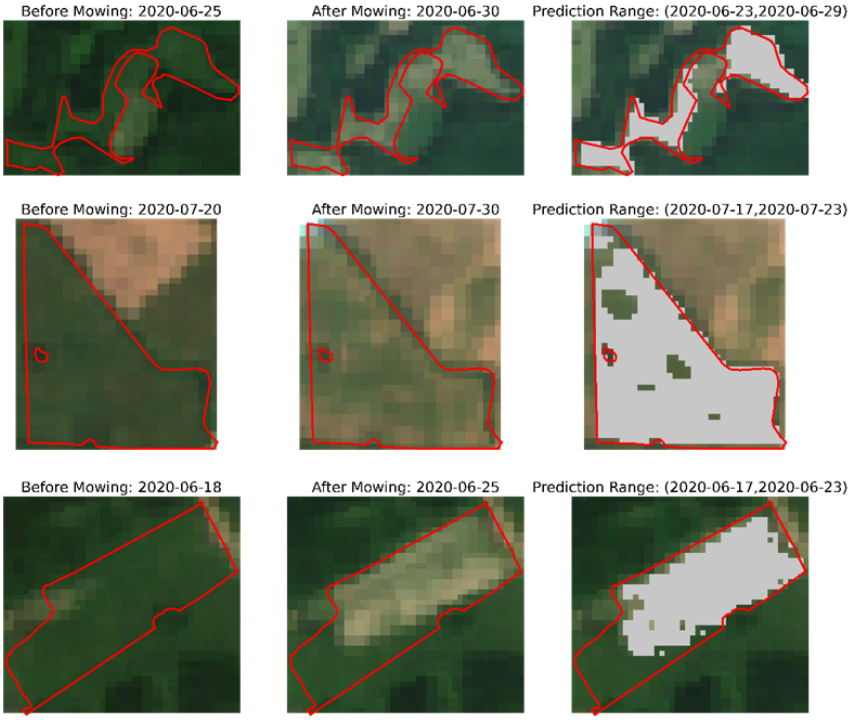


Figure 30: Mowing events expanse over parcels

The result shapefile of Grassland Mowing Events Detection product contains the following fields:

- **UNIQUE_ID**: This field contains a unique identification number provided by NPA, serving as a distinct identifier for each parcel or record within the dataset.
- **APPL_ID**: It stores the ID associated with the applicant responsible for the specific parcel or record, providing a reference to the applicant's identity.
- **PIXELS_S2:** This field records the count of Sentinel-2 pixels considered during the analysis.
- **DECL_N:** Here, you can find the crop name declared by the applicant for the analyzed parcel. It represents the crop type as reported by the applicant.
- **DECL_C:** This field contains the unique ID or code corresponding to the crop declared by the applicant. It helps in associating a specific code with the declared crop.
- **N_EVENTS:** This field contains the total number of mowing events detected through the evaluated period.
- **DATE_1/DATE_2/DATE_3/DATE_4/DATE_5:** This field presents the predicted dates of mowing events. It is crucial for understanding and managing grassland activity over time.
- **CONF_1/CONF_2/CONF_3/CONF_4/CONF_5:** This field indicates the algorithm's confidence for the respective events. In the case of the alternative threshold based approach, as confidence proxy it is depicted the change in the NDVI between the associated detected date and its previous date.

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 tasks such as subsidies allocation, on-the-spot controls, LPIS maintenance, eligibility criteria definition, verification process, and registration of parcel parameters. The current monitoring and inspection process involves electronic applications, on-the-spot checks, administrative controls and submission of additional documents. Usually, CAPO selects those cases for inspection based on annual risk analysis and a random sample. This process includes field visits and remote sensing using high-resolution satellite images, which can be slow and costly. In order to improve the efficiency, the scalability and the accuracy of the monitoring of farmers compliance to the respective CAP measures, CAPO seeks a smart and automated service. Remote sensing procedures can reduce costs and effort during inspections. Precise and timely alerts for non-compliance are necessary for correct subsidy allocation decisions. Integrating ENVISION outcomes into CAPO's operations will enhance their supervision of CAP measures. However, the Cyprus business case a relatively new partner on EU projects aiming on the modernization of agricultural practices monitoring, including ENVISION and DIONE. This section provides an overview of the Cyprus case, user requirements, methodology, services, and limitations. More detailed statistical analysis is available in document D.3.3.

4.3.1 Data products description

Cyprus is an island country in the eastern Mediterranean Sea. It has a subtropical climate with mild winters (with rainfalls during November) and hot summers. Each year, approximately 325,000 farmers' declarations are submitted into the CAPO systems. The eastern inland and coastal areas are primarily used for cultivating cereals, potatoes, vegetables, and seasonal crops. In contrast, the western and central mountainous regions are characterized by vineyards, grasslands, and permanent cultivations such as olive trees, carob trees, and walnut trees.

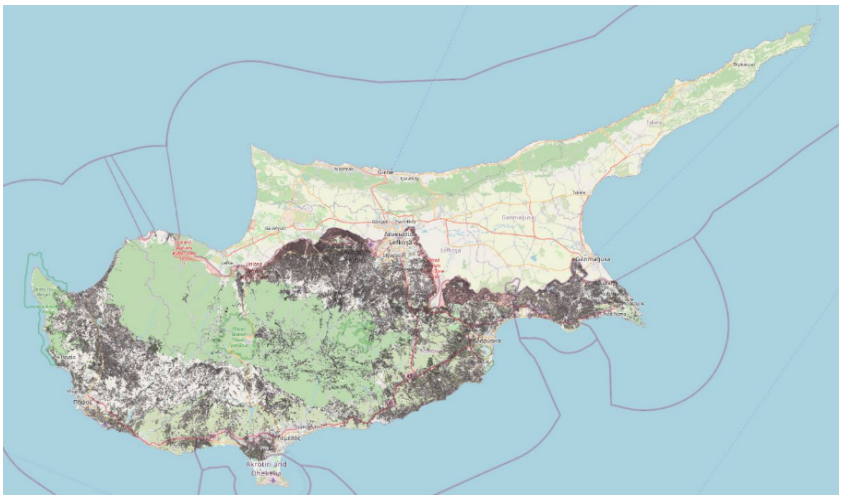


Figure 31: Cyprus Applicants Declarations (GSAA-2022)

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

Specifically, the Cyprus business case faces significant challenges due to the small size of declared fields and abstract geometries, with nearly half of them being below 0.25 hectares. More than 25% of the total cases have a total area less than or equal to 0.1 hectares, resulting in a limited number of available Sentinel-2 pixels per parcel (Figure 32). Additionally, the shape of the parcels is often narrow and lengthy, forming thin strips of land, which further complicates the monitoring task. This situation makes Cyprus a unique case where the average parcel size is comparable to the fixed resolution of Sentinel-2 pixels (10 m.), and the narrow shape of the fields leads to a limited number of pixels that often provide vague or mixed information (mixels). Only a few parcels have a sufficient number of representative pixels available for accurate monitoring and analysis.

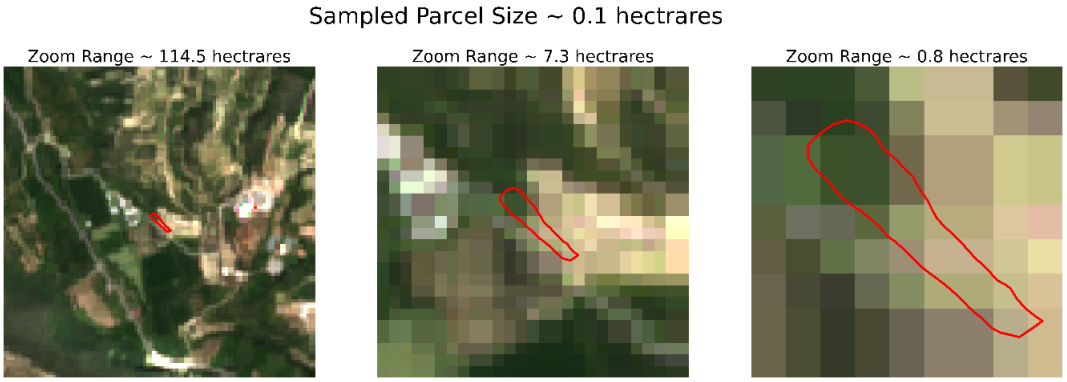


Figure 32: Sample parcel of Cyprus from S2

In addition, a relatively high proportion of erroneous crop code registration during the declaration period pose a significant problem, since it is estimated to be more than 10% of the total cases submitted into the system. This issue arises due to negligence or carelessness on the part of farmers during the application process. Large landowners often submit the same unchanged declarations from previous cultivation periods without updating them. Furthermore, some farmers declare fields as fallow lands, even though they contain tillable areas, in order to meet the minimum EFA (Ecological Focus Area) or greening requirements and qualify for corresponding allowances. These cases can be considered fraudulent and it is important for CAPO to detect them promptly in order to issue appropriate warnings and recommendations.

Last but not least, land lying fallows presents a unique case in Cyprus due to its abstract nature and the ambiguous definition associated with it. It encompasses both arable land with minimal agricultural activity or areas of permanent cultivation with sparse scattered trees inside. As a result, there is a high variance in spectral signatures observed for fallow land. This variability can make it difficult to accurately classify and monitor these areas using remote sensing techniques.
Overall, the distribution of crop types in Lithuania exhibits a wide variety, with numerous agricultural categories represented across the country. The most frequently declared crops include Barley, Wheat, Olive Trees, Fallow lands and Vineyards, which results to more than 70% of the total cases submitted from the applicants. It is important to note that approximately 95% of the declared crops have been successfully included in the Crop Type Mapping analysis product, indicating a high level of coverage

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

and accuracy in capturing the agricultural landscape. Figure 33 provides a detailed visualization of this distribution, highlighting the prevalence of different crop types.

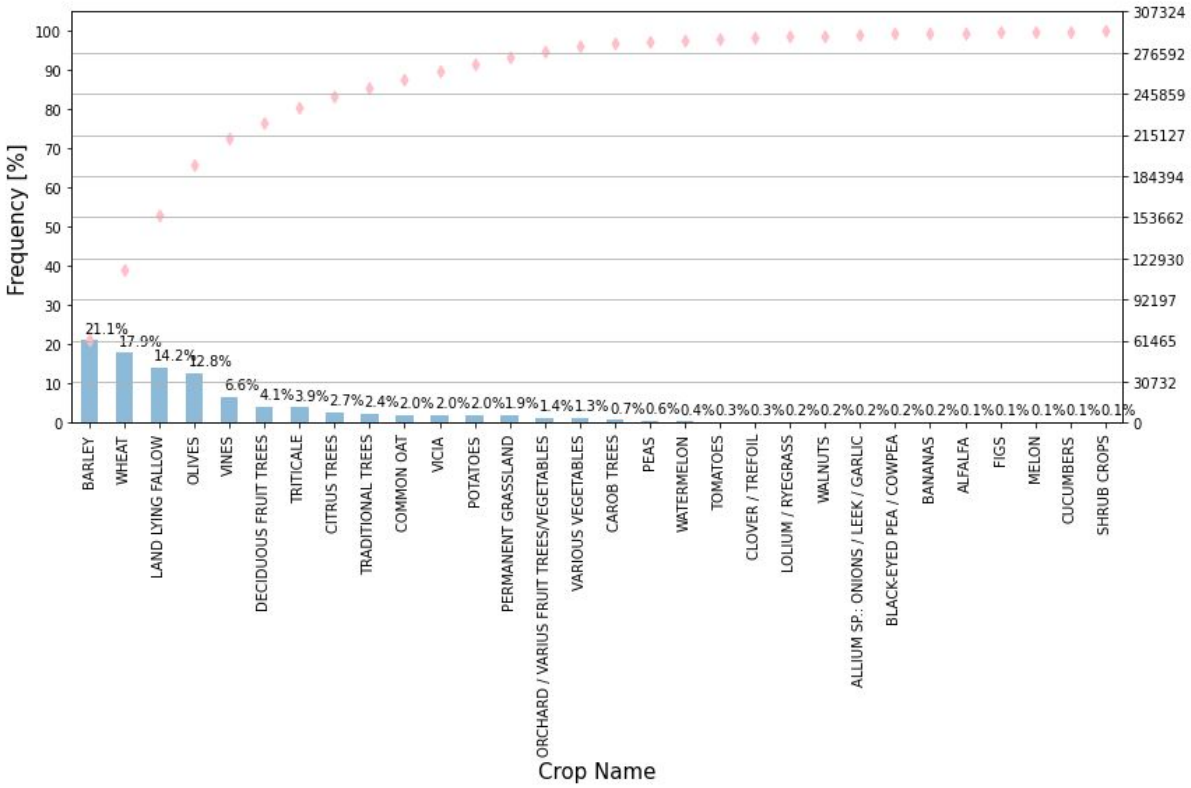


Figure 33: Cyprus (2022) – Distribution of Crops Predicted

Additionally, a crop taxonomy scheme (Figure 34) is employed in collaboration with agricultural experts from the CAPO to facilitate crop classification. This taxonomy scheme organizes crops into higher-level categories according to their agro-botanical peculiarities, providing a useful and effective approach for identifying incorrect declarations and improving the classification of finer crop categories.

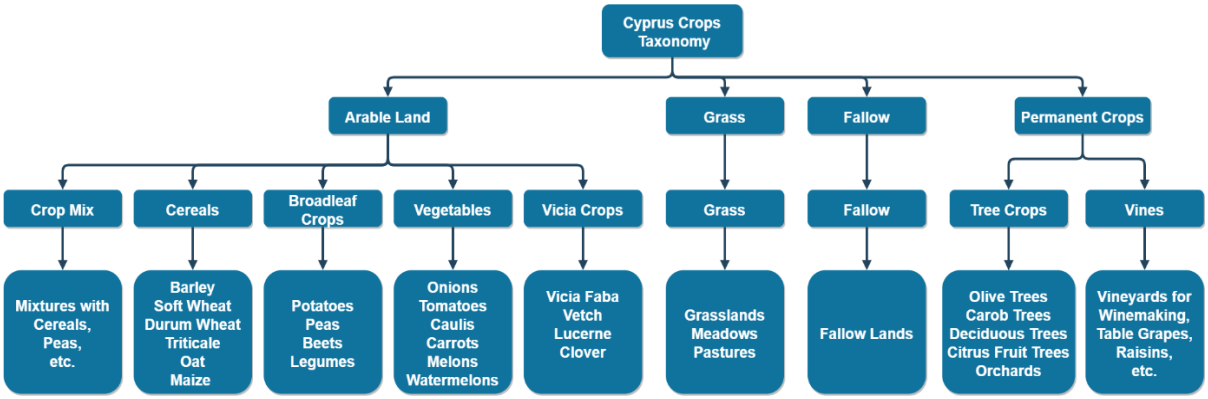


Figure 34: Crop Taxonomy for Cyprus

Within the BC2 of Cyprus, two data products have been developed to meet the specific monitoring needs. Throughout the ENVISION project, user requests and EU regulations have been taken into account, resulting in the development of specific EO services to address the current CAP requirements. The following is a list of the current data products providing:

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

1. Analytics on Vegetation and Soil Index Time-series:
    a. Minimum Soil Cover for Soil Erosion: Provides information on soil percentage and minimum soil cover to assess soil erosion risk.
    b. Runoff Risk Assessment for the Reduction of Water Pollution in Nitrate Vulnerable Areas: Helps assess the risk of runoff and water pollution in nitrate vulnerable areas.
    c. Natura2000 regions activity hotspot detection: Detection of illegal land clearing in nature protection areas.
    d. Stubble Burning Identification: Detects and identifies instances of stubble burning.
2. Cultivated Crop Type Maps:
    a. Confirmation of applicants' declarations (based on GSAA): Utilizes *machine learning* techniques to classify crop types dynamically throughout the entire cultivation period or confirm the declared crop types. Additionally, the provision of a homogeneity marker can assist on the detection of multi-cultures and the presence of multiple cultivations within a single field.
    b. Smart Sampling for OTSC: Implements a traffic light system to assess the risk of incorrect declarations from applicants.
    c. Crop Diversification (CD) Compliance Map: Generates a map indicating compliance with *Greening I* requirements.

Please refer to document D3.3 for an additional analysis of these data products.

4.3.2 Methodology

4.3.2.1 Analytics on Vegetation and Soil Index Time-series

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

Input data

1. Satellite:
    a. Sentinel-2 L2A (tiles: 36SWD, 36SVD)
        i. Spectral bands (B01-B12)
        ii. Scene Classification (SCL)
    b. Sentinel-1 GRD (rel. orbits: 94,167)
        i. Backscattering coefficients (VV-VH)
2. Products:
    a. Annual soil loss layer
    b. Rainfall erosivity factor layer
    c. Soil erodibility factor layer
    d. Slope length factor and slope steepness factor layer
    e. Crop and cover management factor layer
    f. Conservation supporting practices factor layer
    g. Slope - DEM
3. Paying agencies:

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

a. Geospatial data (LPIS, GSAA): Parcels geometries and farmers declarations as a shapefile (updated when is necessary)
b. A lookup table for all the available crop type names, codes, families and CD ancillary info
c. Events Timestamps to fine-tune the algorithms (Stubble Burning, Harvest of arable land)
d. Agricultural Practices Descriptions – National CAP strategic Plans
e. Hydrographic Network
f. Natura2000 regions

Output data

The product offers four distinct output components based on the specific delivery requirements of end users (their fields structure is similar to the respective DP1 output for Cypriot case):

1. A shapefile indicating bare soil percentage and minimum soil cover alerts for soil erosion.
2. A shapefile of Runoff Risk level map of parcels within nitrate vulnerable zones.
3. A Shapefile identifying stubble burning events on arable land, including dates of the events detected at the parcel level.
4. A shapefile for Natura2000 Hotspot Detection in geometrical points based showing the alert pixels detected

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

In order to answer the statutory 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 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

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. As expected, 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. For the time being, CAPO performs

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

random on-the-spot controls in order to locate illicit agricultural practices inside Natura protected zones.

For this purpose, an intensive activity detection routine has been developed to locate areas of intense activity within Natura 2000 regions using vegetation signatures evaluation. This routine utilizes satellite imagery and evaluates vegetation and soil indices to identify pixels that show significant changes associated with intense activity. The routine requires users to provide the specific Natura 2000 regions of interest for analysis (see Figure 35).

Figure 35: Natura 2000 network sites in Cyprus

The methodology used for intensive activity detection within Natura 2000 regions is similar to the harvest event detection method used in BC1. It involves analyzing a combination of vegetation indices such as NDVI, NDMI, PSRI, and BSI over time. A threshold-based routine, similar to the one described in BC1, is used to identify significant changes in these indices, indicating illicit activity. To ensure compliance with customer policies, Eligible Agricultural Areas defined from the LPIS are excluded from this analysis within Natura 2000 sites. This helps distinguish between authorized interventions and potentially unauthorized activities. Lastly, in order to minimize noise from large forestry regions, where vegetation indices can change significantly due to seasonal variations, specific guidelines are followed. The analysis focuses on checking only the boundaries of these large forestry areas, reducing the impact of their vegetation index variations on the intensive activity detection process.

The output of the routine is provided in the form of point shapefiles, indicating the precise locations of the detected alerts. These alerts serve as valuable indicators of areas where intense activity may be occurring within Natura 2000 regions, allowing for targeted monitoring and management efforts

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

Figure 36: Cyprus Natura2000 Alert Pixels Detected Example

Stubble Burning Identification

As stated earlier 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 and discriminate them from the wildfire cases. This information is crucial for monitoring and managing agricultural practices, as well as for assessing the environmental impact of these activities.

By applying only Stubble Burning Events Identification "Pseudocode A" described in BC.1, the identification of such cases has become much easier. By analysing various indices derived from satellite data, such as NDWI, NDVI, etc., the product is able to effectively detect and map areas where stubble burning activities have taken place.

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

Figure 37: Stubble Burning Event Detection example case (Cyprus)

4.3.2.2 Cultivated Crop Type Maps (CCTM)

As already mentioned in the Lithuanian case, for the pilot of Cyprus, the Cultivated Crop Type Maps (CCTM) product combines data from Sentinel-1 and Sentinel-2 satellites to generate maps of cultivated crop types. The CCTM product includes the following features:

- Dynamic crop type map predictions, providing current information on crop distribution throughout the cultivation period.
- A homogeneity marker for the detection of multi-cultures and the presence of multiple cultivations within a single field.
- A traffic light alert system to identify potential false declarations, using smart sampling algorithms
- Assessment of crop compliance with the Greening-1 rule for Crop Diversification, ensuring adherence to regulations.

Input data

More specifically, we utilize the following data:

1. Satellite:
    a. Sentinel-2 L2A (tiles: 36SWD, 36SVD)
        i. Spectral bands (B01-B12)
        ii. Scene Classification (SCL)
    b. Sentinel-1 GRD (rel. orbits: 94, 167)
        i. Backscattering coefficients (VV-VH)

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

2. Paying agencies:
    a. Geospatial data (LPIS, GSAA): Parcels geometries and farmers declarations as a shapefile (updated when is necessary)
    b. A lookup table for all the available crop type names, codes, families and CD ancillary info (Table 9)
    c. Agricultural Practices Descriptions – National CAP strategic Plans

Table 9: Cyprus Look-up Table sample schema

ID	CROP CODE	CROP NAME	CROP FAMILY	EAA	AL	PGrass	TGrass	Fallow	Cwater	Protein	Cother
1	1	DURUM WHEAT	CEREAL	1	1	0	0	0	0	0	0
2	3	BARLEY	CEREAL	1	1	0	0	0	0	0	0
3	4	COMMON OAT	CEREAL	1	1	0	0	0	0	0	0
4	5	MAIZE	CEREAL	1	1	0	0	0	0	0	0
5	6	SORGHUM	CEREAL	1	1	0	0	0	0	0	0
6	8	FORAGE PEAS	BROADLEAF CROPS	1	1	0	0	0	0	1	0
7	13	LOLIUM/RYEGRASS	VICIA	1	1	0	0	0	0	1	0
8	15	ALFALFA	VICIA	1	1	0	0	0	0	1	0
9	18	CHICKPEA	BROADLEAF CROPS	1	1	0	0	0	0	1	0
10	19	LENTILS	BROADLEAF CROPS	1	1	0	0	0	0	1	0
11	25	TOMATOES	VEGETABLES	1	1	0	0	0	0	0	0
12	26	CUCUMBERS	VEGETABLES	1	1	0	0	0	0	0	0
13	42	OLIVES	TREES	1	0	0	0	0	0	0	0
14	70	WINE VINEYARDS	VINES	1	0	0	0	0	0	0	0

Output data

The product is providing an output shapefile consist of:

1. Dynamic crop type maps over the registered parcels for every new or group of new Sentinel acquisitions.
2. A homogeneity marker for the detection of multi-cultures and the presence of multiple cultivations within a single field.
3. Traffic light maps over the registered parcels for smart sampling of on-the-spot inspections and early alert of the users.
4. Crops Diversification (Greening-1) compliance map over the registered parcels at the end of the cultivation period.

For the outputs 1, 3, and 4, the procedure followed in Lithuania has been replicated. However, in order to accommodate the smaller size of declared parcels in Cyprus, a pixel-based classification approach has been adopted for many cases since the average parcel size of Cyprus is approximately 0.25 hectares.

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

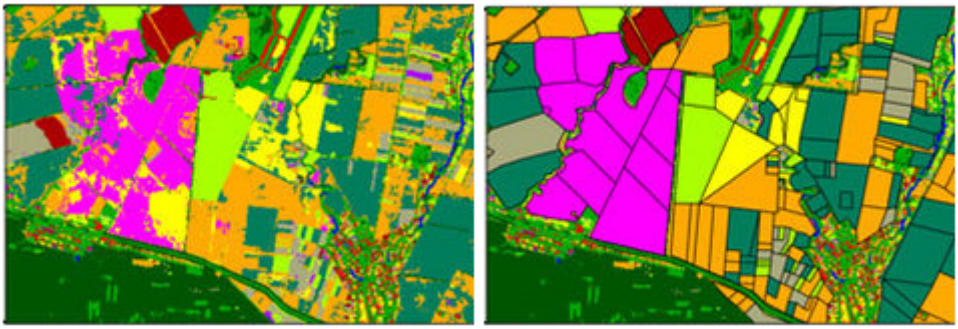


Figure 38: From Pixel-level to Parcel-level classification

Homogeneity Marker

One of the common challenges faced by CAPO is the issue of multiple crop types being cultivated within the same parcel. Resolving this issue is crucial for confirmation of farmers' applications (GSAA) and ensuring eligibility verification, and compliance with agricultural policies. Moreover, this practice can create difficulties in accurate crop mapping and monitoring of agricultural activities, as it becomes challenging to determine the specific crop types and their respective areas within the parcel. To address this, an algorithm has been developed (based on the respective work of Sentinel-Hub[10] and DIONE project (https://zenodo.org/records/7116922)) to create a supplementary homogeneity indicator that can highlight such cases. The presence of multiple crops within a single field often leads to significant spectral variation within the parcel. As a result, the algorithm aims to quantify this intra-parcel variation and provide an indicator of homogeneity.

The routine for calculating the homogeneity (Figure 39) of a parcel involves analyzing the NDVI time series data of the pixels within the parcel's geometry. First, the pixels included in the parcel are identified. Then, for each pixel, the spatial standard deviation is computed from its corresponding time series data. These standard deviation values are collected and used to calculate the temporal mean. Finally, the homogeneity of the parcel is obtained by subtracting this NDVI temporal mean (heterogeneity) from 1. This approach allows for the assessment of the temporal consistency and uniformity of the pixel values within the parcel, providing an indication of the parcel's homogeneity. By quantifying the homogeneity, this routine enables the evaluation of the uniformity of the pixel values over time, which can be useful for identification of polycultures.

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

```
calculate_homogeneity

Parameters:
- parcel_geometry
- NDVI_time_series_data

Variables:
- parcel_pixels
- std_values
- temporal_mean
- homogeneity

Algorithm:
parcel_pixels = get_pixels_within_geometry(parcel_geometry)
std_values = []

for each pixel in parcel_pixels:
    NDVI_time_series = get_time_series_data(pixel)
    std_value = calculate_spatial_std(NDVI_time_series)
    std_values.append(std_value)

temporal_mean = calculate_temporal_mean(std_values)
homogeneity = 1 - temporal_mean

return homogeneity
```

Figure 39: Homogeneity marker calculation "Pseudocode" for Cyprus Pilot

Based on the distribution of homogeneity values (Figure 40), we recommend that CAPO considers the parcels with lower homogeneity (most heterogeneous) as potential cases of multiple cultivation. These parcels should be further investigated and checked for polycultures, where multiple crop types are being cultivated.

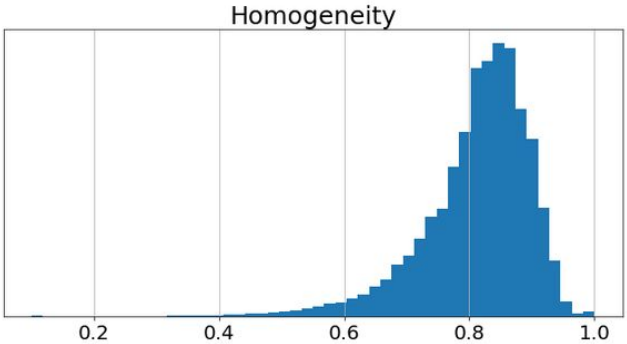


Figure 40: Homogeneity Distribution for all cases in Cyprus

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

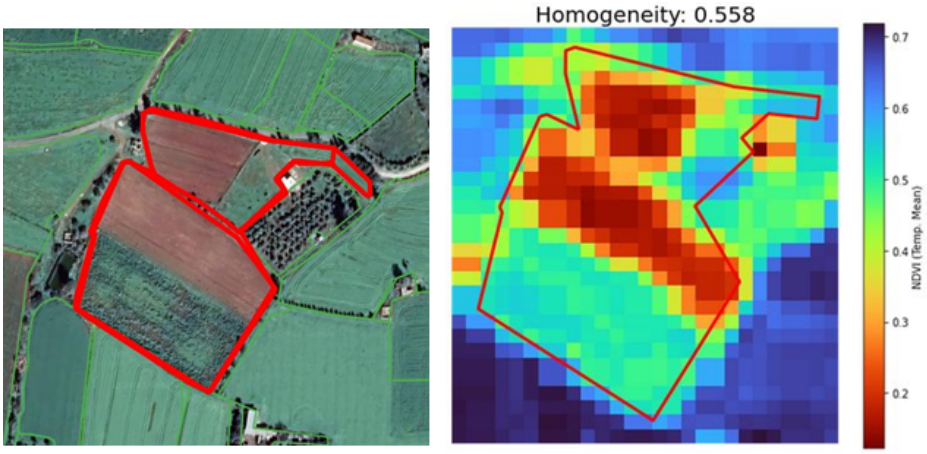


Figure 41: Example of Polyculture

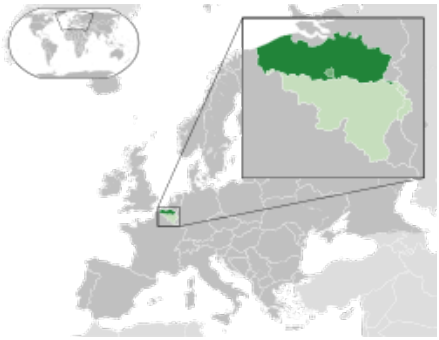
The result shapefile of CCTM products contain the following fields:

- **UNIQUE_ID**: unique_id number provided by NPA.
- **APPL_ID**: id of the applicant.
- **D_AREA**: field's area declared by applicant (in hectares).
- **C_AREA**: field's area calculated by NOA (in hectares).
- **PIXELS_S2**: number of S2 pixels included into the analysis.
- **PIXELS_S1**: number of S1 pixels included into the analysis.
- **DECL_N**: crop name declared.
- **DECL_C**: id of the code declared.
- **PRED_N1**: crop name predicted (1st most confident prediction).
- **PRED_C1**: id of the code predicted.
- **CONF_1**: prediction's confidence.
- **PRED_N2/PRED_C2/CONF_2:** results for the 2nd most confident prediction.
- **PRED_N3/PRED_C3/CONF_3**: results for the 3rd most confident prediction.
- **CONFIRM**: confirmation of declaration.
- **ALERT**: level of alert - false declaration (based on “traffic light alert system”).
- **CD**: applicant’s crop_diversification category based on Table 8.
- **HOMOGENEITY MARKER:** parcel's homogeneity for multi-crops guidance identification.

4.4 BC3: Monitoring the condition of the soil – Belgium

The Flemish business case focuses on deploying ENVISION service for topsoil Soil Organic Carbon Monitoring in Flanders, Belgium. 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 the soil's state and require significant effort, time, and resources to be committed.

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



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, and 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 OnTheSpot Controls are performed. The payment entitlements are close to 500.000 and € 245,30 Million.

EV ILVO will assist LV, a scientific institute specialised in service provision in all fields related to agriculture, fisheries, and food in Flanders. At various meetings, LV has defined the service requirements (see D2.2 Report of customer requirements from ENVISION services).

EV ILVO is responsible for the following:

a) To perform the soil campaign together with LV.
b) Design and develop a process that delivers SOC products to the Envision platform.
c) to assess in which way and under which conditions those products can support the provision of CAP monitoring services.

To develop the data products, we are building EO-based ML models that predict the topsoil organic carbon at the pixel level. After we create indicators that present the soil quality, we need to define thresholds, considering the soil-pedological conditions.

Critical challenges we have tackled until now are related to the bare soil identification, the collection of an adequate number of soil samples covering the different soil conditions in Flanders (for example, texture), the improvement of the accuracy of the top SOC models, the assessment of the optimal period that allows assessing top SOC changes[3], the ability to deliver SOC predictions for the majority of agricultural parcels in Flanders, the way we need to define the soil quality indicators considering the Flemish soil-pedoclimatic conditions but also the ability to develop soil quality products at EU level.

In the following sections, we present the work done until now, adding some extra sections to explain how we have formulated the Soil Quality indicators and the approaches we have followed. We also provided info for the new data products and some additional explanations to address comments coming from the reviewers.

4.4.1 Data product description

Study site

[3] Repetition period between 3 to 5 years, similar to the Monitoring Report and Verification systems

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

The study area is the Flemish region and the data products should cover the cropland areas (Figures 42-43).

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

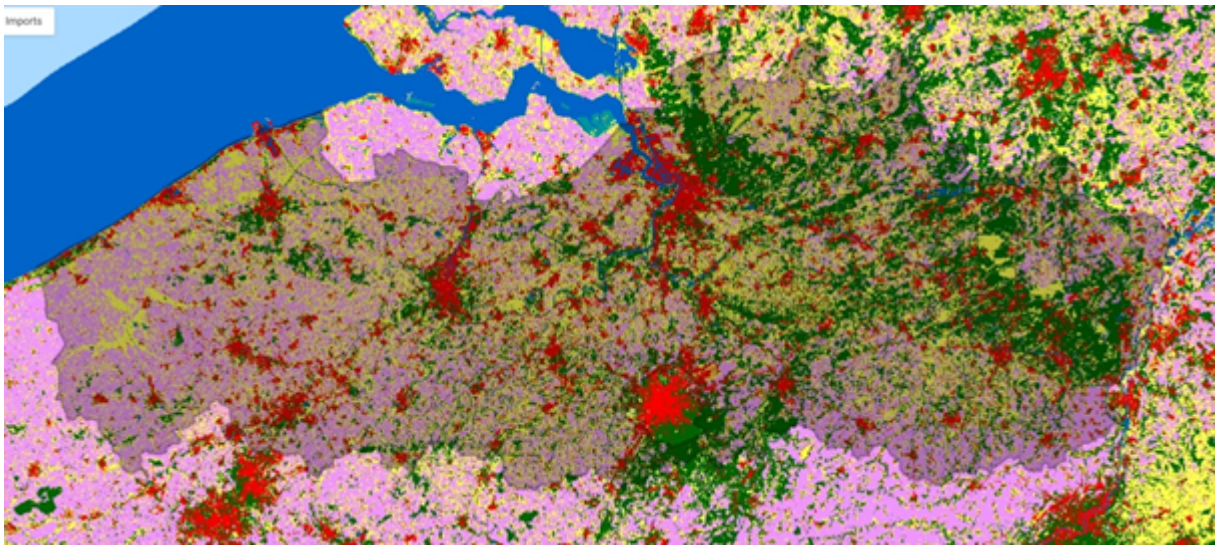


Figure 43: A land cover map of Flanders using the European Space Agency (ESA) WorldCover 10 m 2020 product provides a global land cover map for 2020 at 10 m resolution based on Sentinel-1 and Sentinel-2 data. The WorldCover product comes with 11 land cover classe

Cloud Coverage

A significant problem that characterizes the area and that 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 is, 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 spectrum's visible, near-infrared, and short-wave infrared parts. This intervention may produce misleading results in analyses of import noise and may have dramatic

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

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. Table 10 below presents the number of Sentinel 2 images (L2 products) with at least 90% cloud coverage. Compared to total products, they represent almost 30%. It becomes apparent that in these cases, products generate missing values at the pixel level that must be identified and then dropped.

Table 10: Estimated products with at least 90% cloud coverage in Flanders for 2019 until 2022

Year	Number of products with at least 90% cloud coverage	Total Products
2019	475	1541
2020	397	1539
2021	505	1511
2022	482	1532

Soil conditions and modelling data

To support the development of the SOC products, a soil sampling campaign was performed at the beginning of 2019. 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.

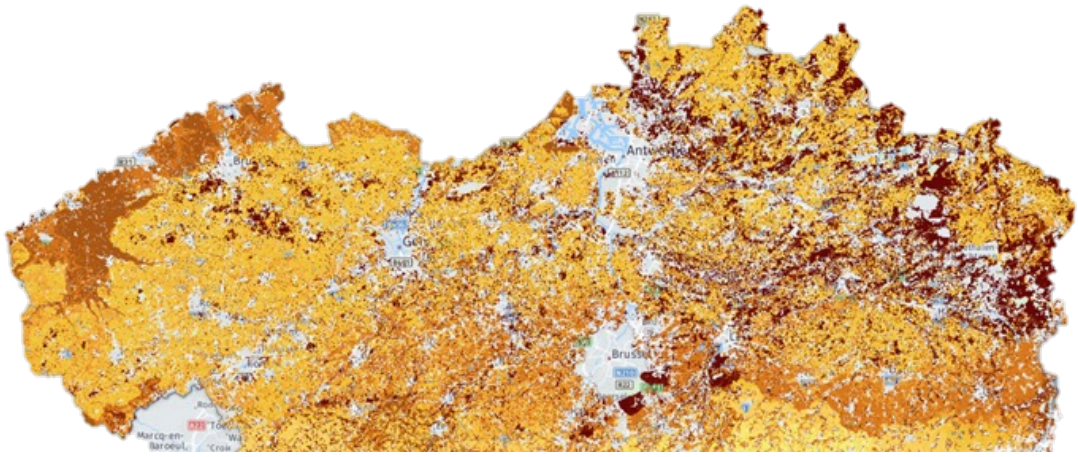


Figure 44: Existing Existing Top Soil Organic Carbon stock map for topsoil (0-30cm) with a mean 40m grid (10m for Flanders and 40m for Wallonia region). The maps are based on digital soil mapping approaches using

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

empirical models calibrated to predict the SOC stock and using covariates available at a sufficient resolution at the regional scale. All maps are strongly dependent on the Belgian Soil Map (texture and drainage parameters).

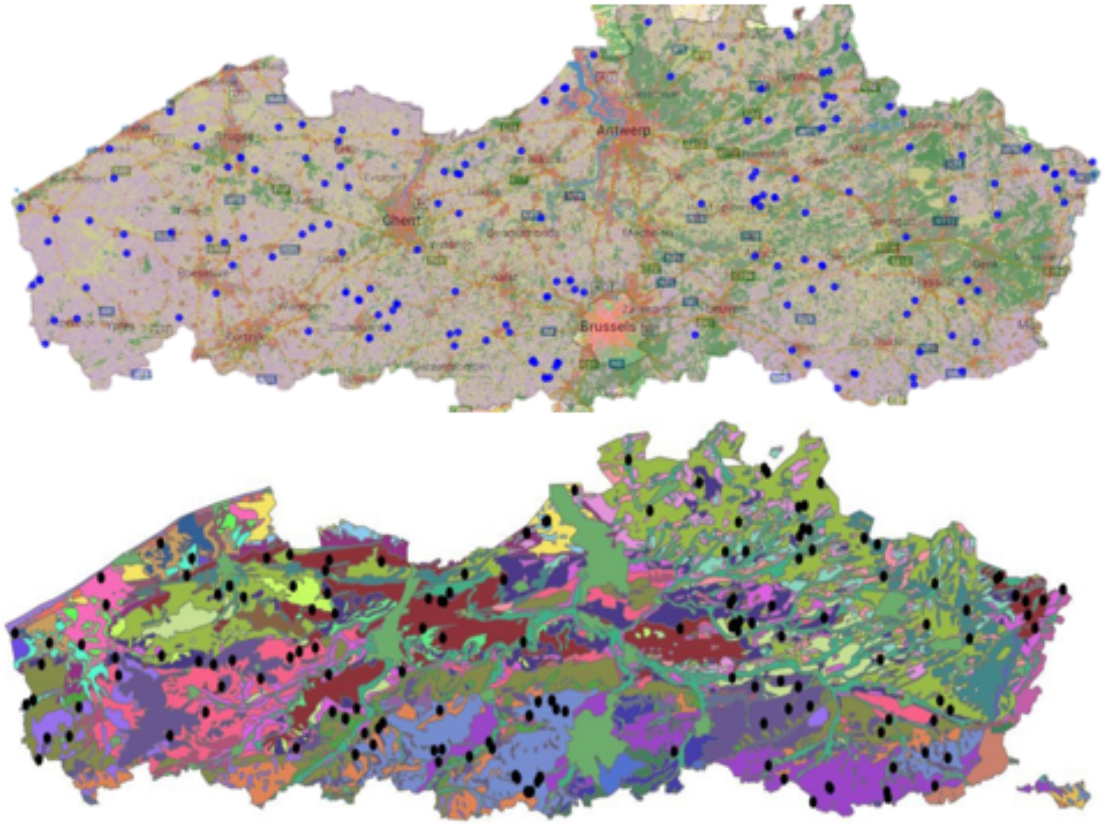


Figure 45: Location of the Envision campaign sampling points with background maps, the land classes (the upper map, the sampling points are with blue points) and soil associations in Flanders (bottom map, the sampling points are with black points). A soil association is a substantive and spatial grouping 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 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 ENVISION project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869366

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 47). 88% of the samples contained a SOC content between 0,5 and 2,0 %.

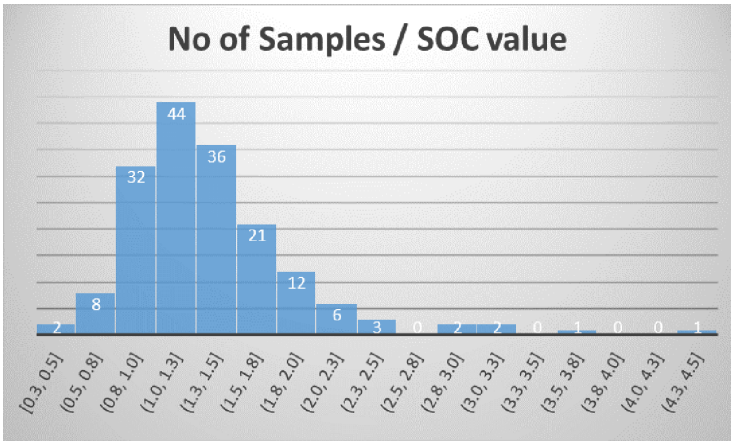


Figure 47: No of samples (y-axis) and the estimated SOC value (x-axis). From the 171 samples, the majority takes SOC values between 0.8 - 1.8 (%/dry soil).

Bare soil identification

We follow a methodology that develops a cloudless bare soil composite collection to develop the Top Soil Organic Carbon products. A bare soil composite is an extensive collection of multispectral satellite data that can be used to map topsoil attributes to a large extent [11]. 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 must 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.

Using STAC services, it's possible to generate a large collection of Sentinel 2 images and generate the needed indices for all or selected pixels.

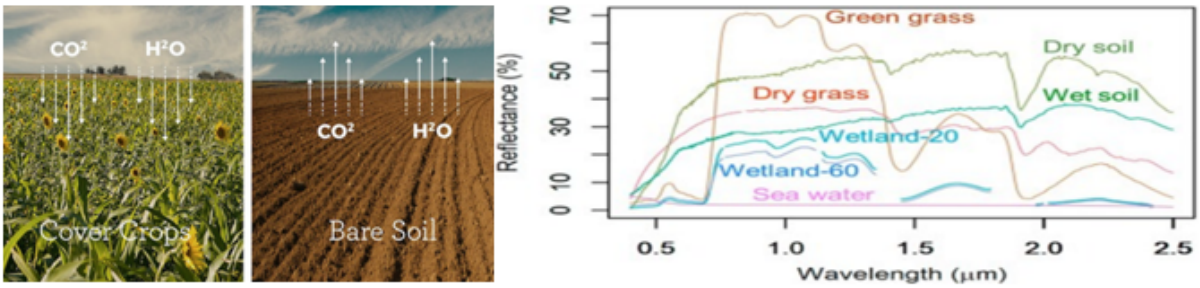


Figure 48: The identification of bare soil pixels in extensive image collection is a significant task supported by vegetation, bare soil and soil moisture indices. Sentinel 2 bands in NIR and SWIR can support the identification of Dry and Wet Soil.

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

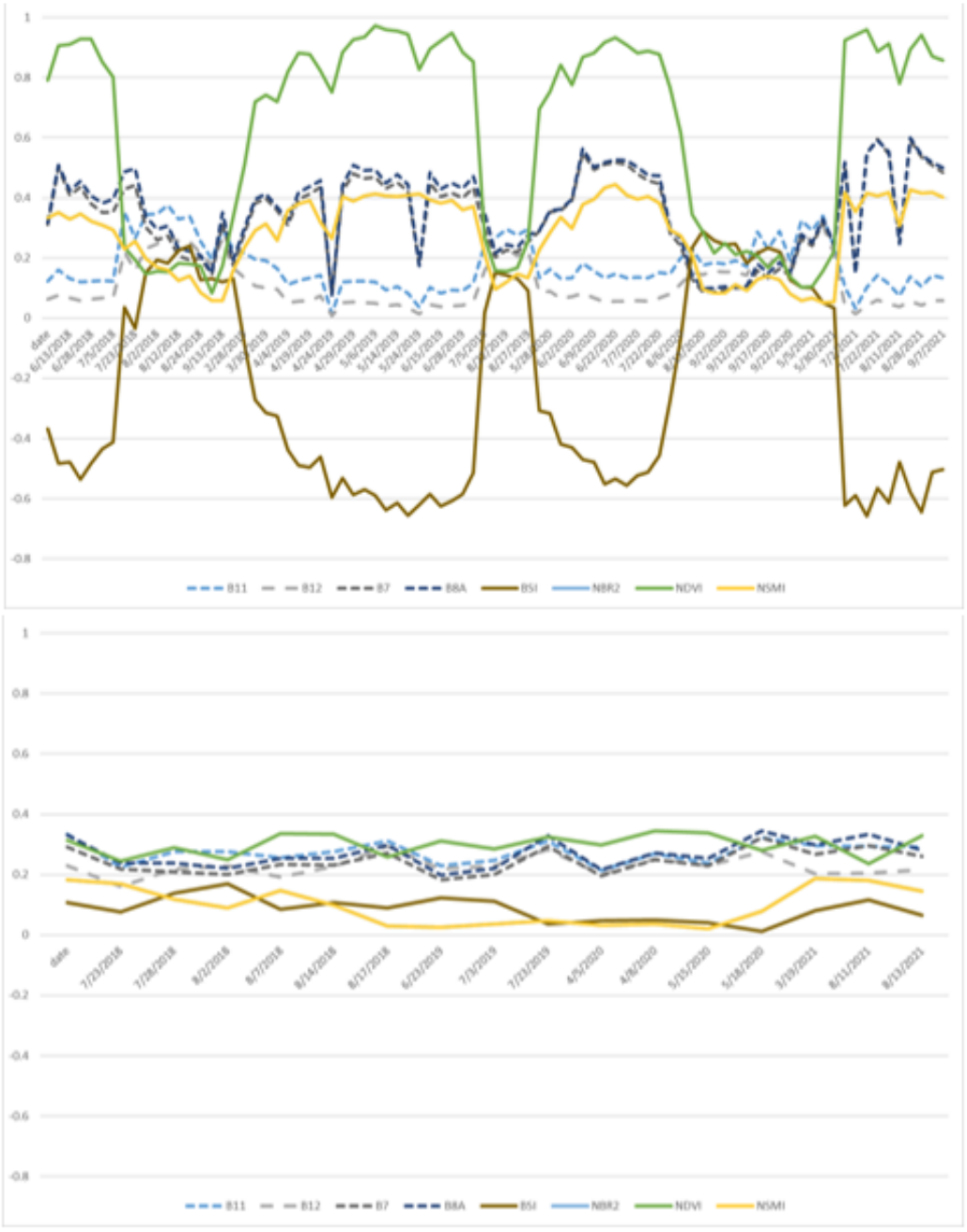


Figure 49: Time series of S2 reflection bands together with bare soil, soil moisture and vegetation indices for sampling point No 12 of the soil campaign for 2018 and 2021 (upper). After applying an NDVI filter of <0.35 reduces the times series points, generating significant time gaps (from a few weeks 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

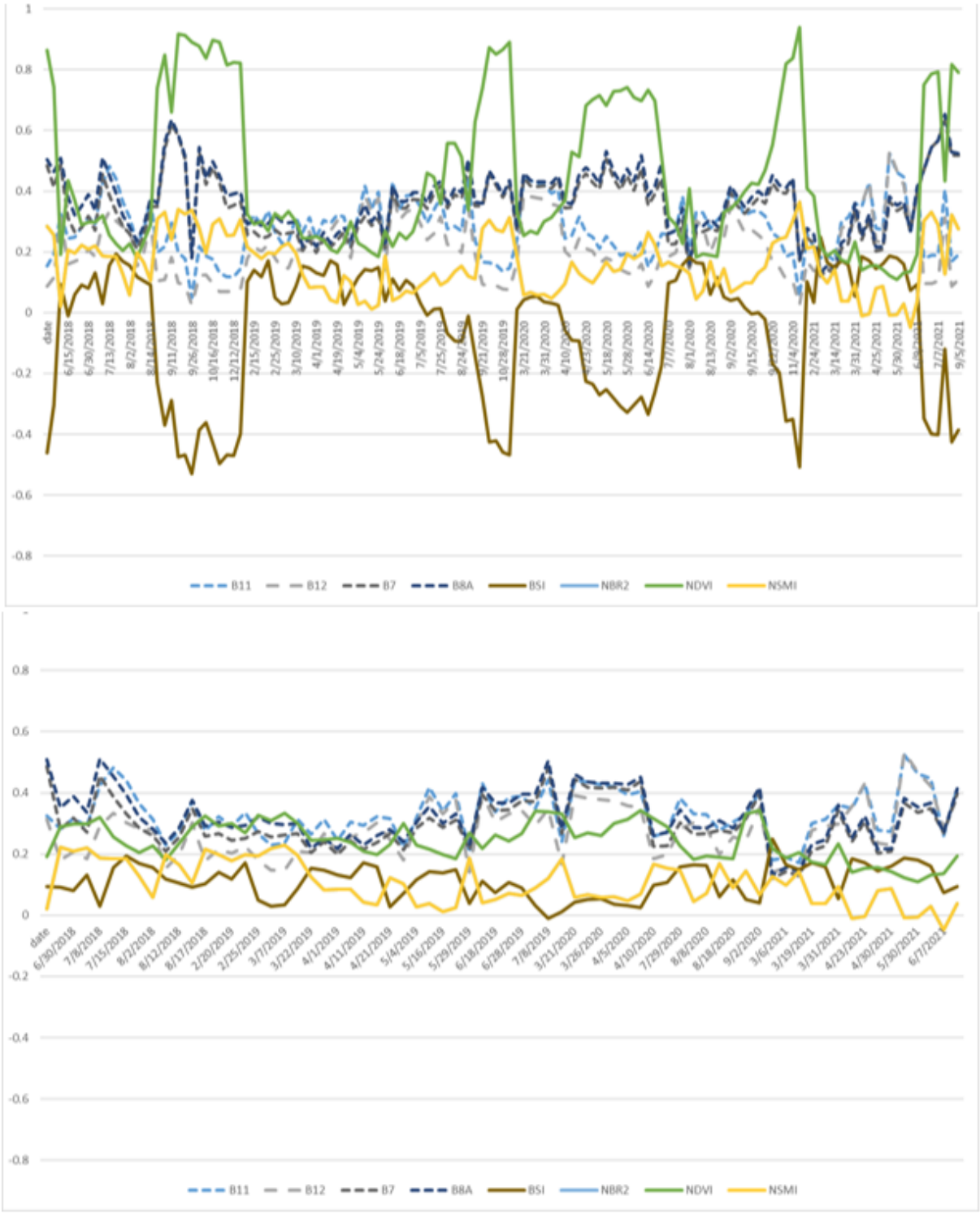


Figure 50: Time series of S2 reflection bands together with bare soil, soil moisture and vegetation indices for sampling point No 2 of the soil campaign for 2018 and 2021 (upper). By applying an NDVI filter of <0.35, we reduce the times series points (from 124 to 70, almost 45%), generating significant time gaps (from a few weeks to a few months). In this process, the goal is to identify bare soil and first 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.

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

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 visualisation 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.

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

Soil Health and Soil Quality

Soil health and soil quality are not identical, and according to the EJP Soil SIREN[45] and SERENA projects, the definitions are the following:

- **Soil Health** is the current capacity of soil to function as a vital living system within natural or managed ecosystem boundaries and land-use boundaries, to sustain plant and animal productivity and health, maintain or enhance water and air quality, and to further provide ecosystem services (Figure 52) on the long-term without (increased) trade-offs between ecosystem services.
- **Soil fertility** is the ability of a soil to sustain plant growth by providing essential plant nutrients, water and favourable chemical, physical and biological properties as a habitat for plant growth.
- **Soil Quality** is the capacity of soil to function within ecosystem and land-use boundaries to sustain biological productivity, maintain environmental quality, and promote plant and animal health (Doran and Parkin, 1994).

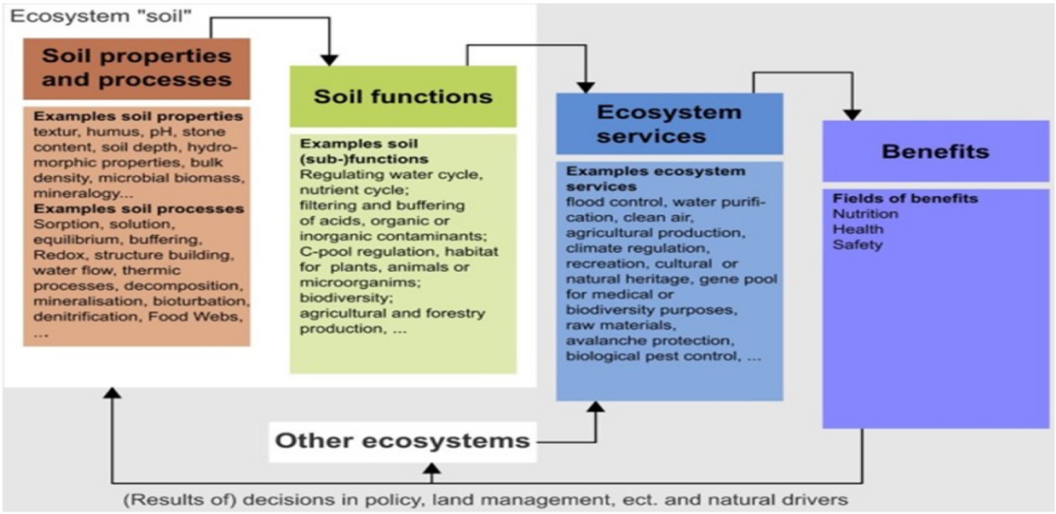


Figure 52: Contributions of soil functions to ecosystem services in the cascading framework developed by Haines-Young and Potschin (2008).

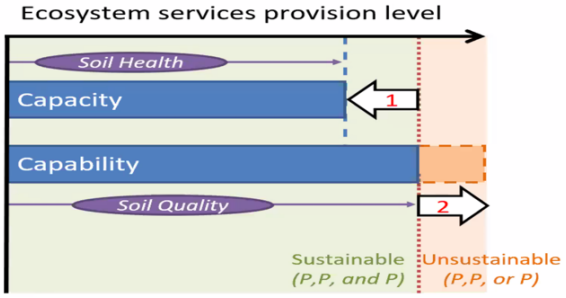


Figure 53: Soil health compared to soil quality. Explanations: 1. Current soil degradation, management practices, climate change, etc., limit Ecosystem Services provision 2. Context properties (e.g., soil type and land use) define potential. An increase in ecosystem services provision is possible by using fertilisers, pesticides,

4

5 https://ejpsoil.eu/soil-research/siren

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

intensive tillage and other management practices, but it leads to increased trade-offs to other services, to other people, elsewhere or later. Land use sustainability in terms of people (P), planet (P) and profit (P).

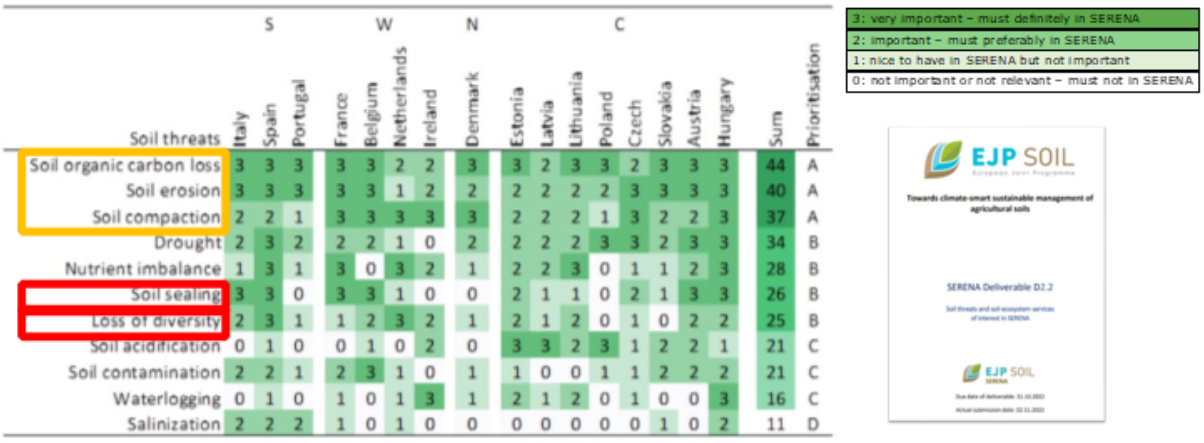
Soil Quality Indicators criteria and approaches

To represent or infer a specific aspect of soil quality, we are using Soil Quality Indicators. An indicator is a Parameter used to quantify and evaluate the impacts of agricultural soil practices on soil quality and the environment to conclude the farming practice or agricultural policy (modified after Piorr, 2003)[6]. Indicators can be measured using analytical protocols, estimated through modelling or expert-based approaches, and quantitative, semi-quantitative or qualitative.

To develop an indicator, we are using **evaluation criteria**, which are:

- **Reference value:** A value for an indicator representing its normal background value for defined local circumstances (ecological conditions). Considered as equivalent to "normal operating range".
- **Target value:** Represents the desired status for a particular indicator or set of indicators given specific ecological conditions, land use and objectives for use by authorities and other stakeholders.
- **Threshold value:** Value above/below which soil health/quality is considered degraded.

EJP Soil, on T2.4.2, performed a survey on Soil Quality Indicators used in Member States. 68 indicators are used to characterise soil quality, for example, the C consertation for the organic carbon or the Soil organic Matter quality or decrease. The same survey highlighted that soil organic carbon loss ranked among the most important threats for most member states (Figure 54). The C concentration and the total N, the P,K and pH as soil quality indicators can be used as **policy indicators** representing **Soil fertility**.



	S			W				N	C								
Soil threats	Italy	Spain	Portugal	France	Belgium	Netherlands	Ireland	Denmark	Estonia	Latvia	Lithuania	Poland	Czech	Slovakia	Austria	Hungary	Sum	Prioritisation
Soil organic carbon loss	3	3	3	3	3	2	2	3	3	2	3	3	2	3	3	3	44	A
Soil erosion	3	3	3	3	3	1	2	2	2	2	2	2	3	3	3	3	40	A
Soil compaction	2	2	1	3	3	3	3	3	2	2	2	1	3	2	2	3	37	A
Drought	2	3	2	2	2	1	0	2	2	2	2	3	3	2	3	3	34	B
Nutrient imbalance	1	3	1	3	0	3	2	1	2	2	3	0	1	1	2	3	28	B
Soil sealing	3	3	0	3	3	1	0	0	2	1	1	0	2	1	3	3	26	B
Loss of diversity	2	3	1	1	2	3	2	1	2	1	2	0	1	0	2	2	25	B
Soil acidification	0	1	0	0	1	0	2	0	3	3	2	3	1	2	2	1	21	C
Soil contamination	2	2	1	2	3	1	0	1	1	0	0	1	1	2	2	2	21	C
Waterlogging	0	1	0	1	0	1	3	1	2	1	2	0	1	0	0	3	16	C
Salinization	2	2	2	1	0	1	0	0	0	0	0	0	0	1	0	2	11	D

3: very important – must definitely in SERENA
2: important – must preferably in SERENA
1: nice to have in SERENA but not important
0: not important or not relevant – must not in SERENA

Figure 6-1 Prioritisation of soil threats by project member states, listed by regions, S = southern Europe, W = western, N= northern and C = central Europe according to the environmental zones after Metzger et al. (2005).

Figure 54: Prioritisation of soil threats by EJP Soil project member states (SERENA project, D2.2)

Another important aspect we need to mention, as it is related to the approach we are using for the development of the Soil Quality products, is the different approaches that can be used to evaluate the soil health indicators. As presented in Figure 55, there are 4 main approaches.

[6] European Commission (EC): Communication from the Commis- sion to the Council, the European Parliament, the European Eco- nomic and Social Committee and the Committee of the Regions, Thematic Strategy for Soil Protection, COM 231 Final, Brussels, 2006.

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

FIXED VALUE	RELATIVE TO NATURAL LAND USES	RELATIVE CHANGES	DISTRIBUTION
One static value, based on best available research/knowledge, stratified as required	One static value, calculated as a percentage of what would be found under 'natural' land uses (stratified as required). Based on modelling	A value calculated based on the local state of the soil and static only for a given period of time (i.e. target is an increase/decrease of y% of current value within x years), after which the value may change)	A value calculated based on the regional state of the soil (i.e. target/threshold is a certain percentile of the current range of values), static only for a given period of time (until distribution is re-measured), after which the target/threshold may change

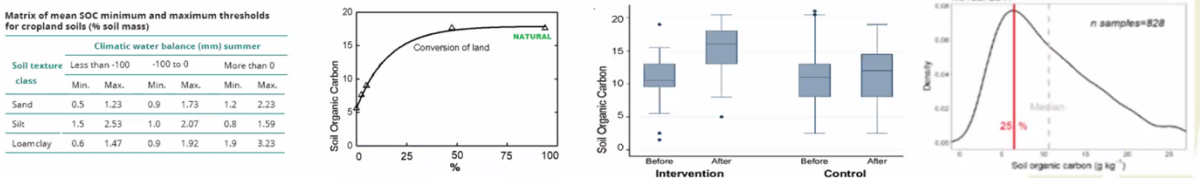


Figure 55: Approaches used for the evaluation of soil health/condition indicators.

Table 11: Pros and cons of different approaches for evaluating the indicators.

Approach	FIXED VALUE	RELATIVE TO NATURAL LAND USES	RELATIVE CHANGES	DISTRIBUTION
Pros	Simple for non-scientists as well	If the modelling is properly elaborated it could work well to fix target values.	A quick way to start evaluating trends. Allows for differentiation due to diverse pedoclimatic conditions. Can be used by advisory services at field scale.	Thresholds adapted to soil districts – pedoclimatic conditions
Cons	Needs stratifications: thresholds must be adapted to specific pedoclimatic conditions. A lot of information is needed.	Few natural lands in Europe that can be used as a reference: most forests and rangelands are managed. Difficult to explain. Requires proper models	May result in problems to credit farmers that have already done well. The mapping for aggregation at smaller scales needs a temporal analysis.	A lot of information is needed to have statistical distributions and must be stratified. If the area is already degraded then the information is biased.

Flemish Soil Quality Indicators, soil-pedoclimatic conditions and approaches tested.

At Envision, for the development of the Soil Quality Indicators, we applied in phase 3, in the beginning, the fixed value approach using as a reference the Belgian Soil Fertility advisory system (Table 12) and the step was to proceed with the Distribution approach. The thresholds on both approaches adapted to the Flemish (Belgium) **soil-pedoclimatic conditions**, using a recently updated soil texture map of the Flemish region, which uses the international soil classification system World Reference Base (Figure 56).

The next step will be to apply and test the Relative Changes approach. For this, we need to generate a new topsoil organic carbon prediction map using a new bare soil collection for the period of May 2022 until May 2023 (optimal until May 2024, to generate 3 years of bare soil collection again) and compare the results at pixel and parcel level.

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

Table 12: The Belgian Soil fertility advice system uses the evaluation classes below for soil organic carbon content for arable land. This table delivers information useful for the definition of a fixed value

Classification / Soil type	Sand	Loam and Sandy-Loam	Clay
	% C	% C	% C
Very low	< 1.2	< 0.8	< 1.0
Low	1.2 – 1.4	0.8 – 0.9	1.0 – 1.2
Moderate low	1.5 – 1.7	1.0 – 1.1	1.3 – 1.5
Normal	1.8 – 2.8	1.2 – 1.6	1.6 - -2.6
Moderate high	2.5 – 4.5	1.7 – 3.0	2.7 – 4.5
High	4.6 – 10.0	3.1 – 7.0	4.6 – 10.0
Peaty	> 10.0	> 7.0	> 10.0

Source: BDB

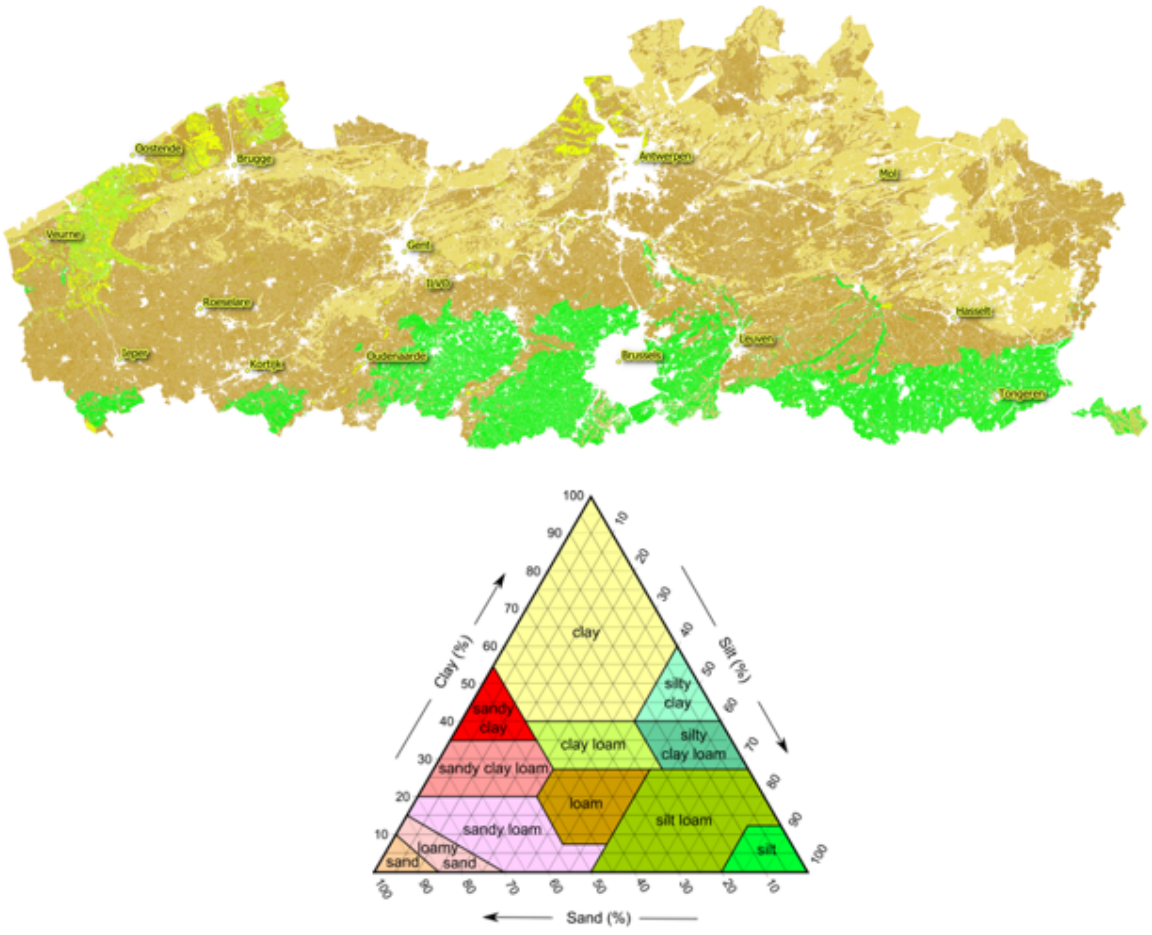


Figure 56: Soil texture map of the Flemish Region according to the international soil classification system World Reference Base on a scale of 1:40,000. Visualisation by EV ILVO using QGIS.

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

Soil Quality data products

To develop the products, we applied the approach described in the section "
Flemish Soil Quality Indicators, soil-pedoclimatic conditions and approaches tested". We developed two products for the Flemish Business Case:

- **A soil quality map at a pixel level**, using an indicator that informs if the pixels have Topsoil Organic Carbon value below the average, around the average, above the average and far above the average, considering **soil-pedoclimatic conditions**. Indicator values have been assigned to 10m by 10m pixels, using as a mask arable crops parcels
- **A soil quality map at a parcel level**, using an indicator that informs if a parcel has a median Topsoil Organic Carbon value below the average, around the average, above the average and far above the average, considering **soil-pedoclimatic conditions**. Indicator values have been assigned to the arable crops parcels by performing a spatial aggregation, using the 10m by 10m pixels as a source.

A soil quality map at a pixel level

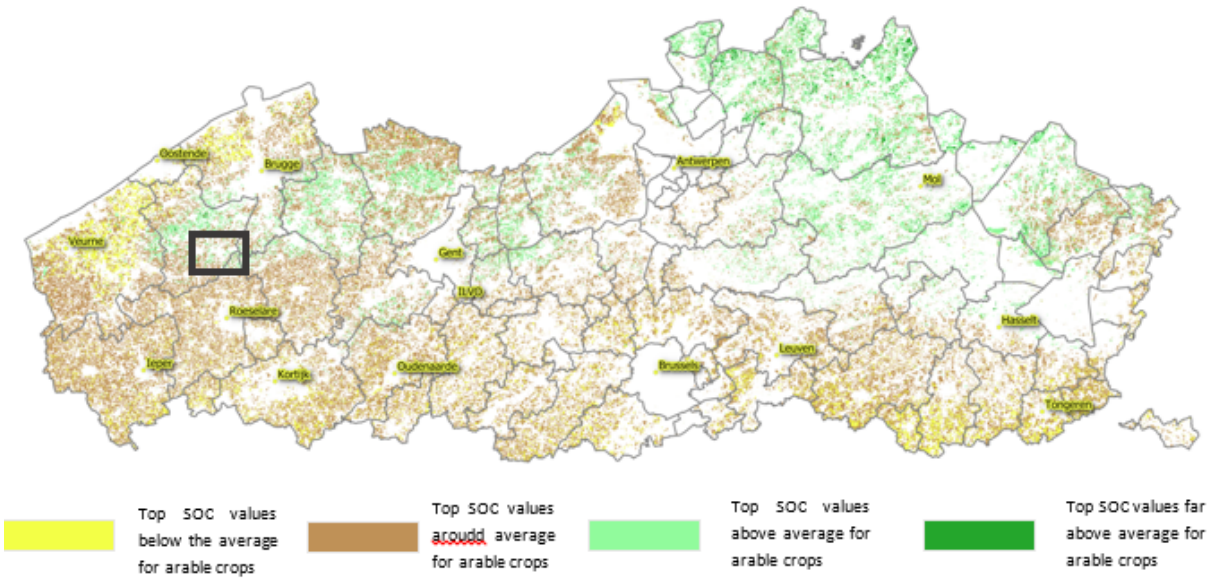


Figure 57: A Soil Quality map 10 by 10 m, pixel size using an indicator representing pixels below the average, around the average, above the average and far above the average.

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

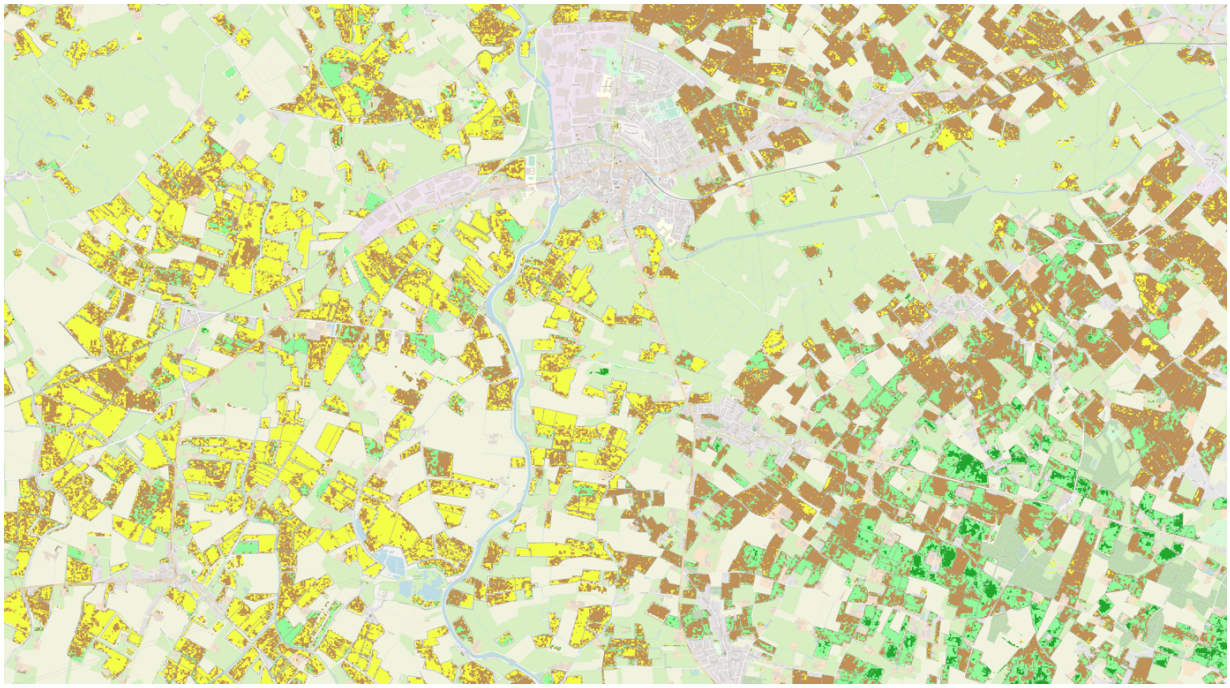


Figure 58: Zoom into the selected window of Figure 16, to visualise the intra-parcel variability of the Soil Quality Indicator.

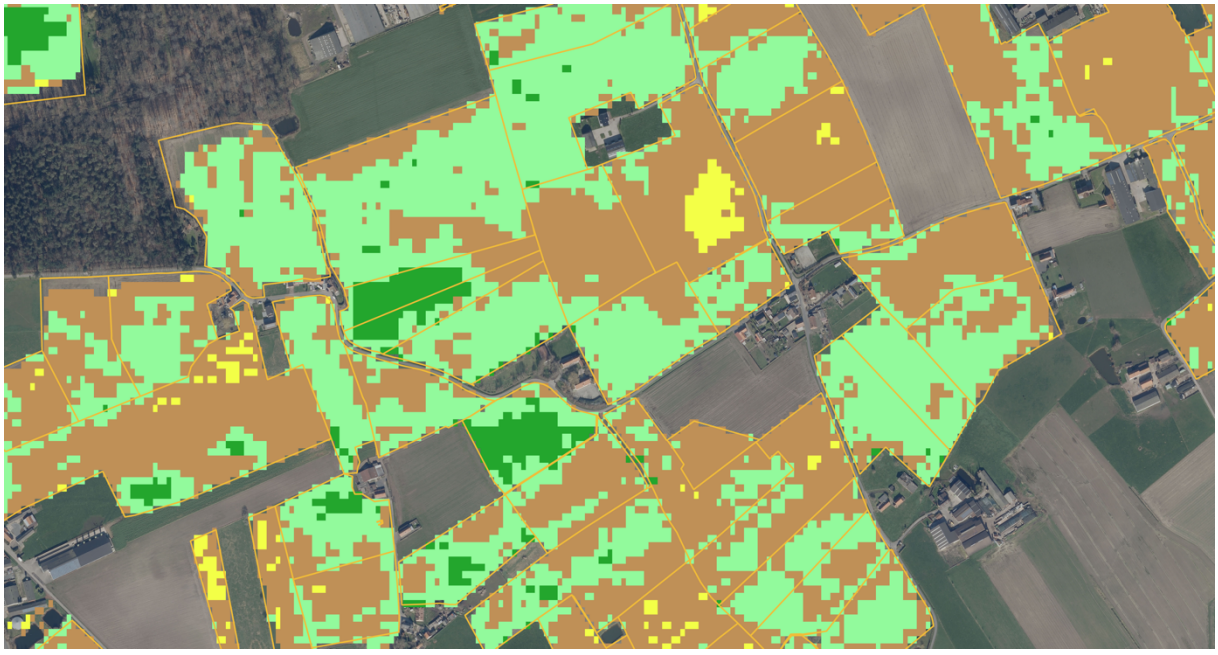


Figure 59: Detail zoom into the selected window of Figure 16, to visualise the intra-parcel variability of the Soil Quality Indicator, using very high-resolution orthophoto maps as a background layer.

The metadata of the raster file provides info for the model, the accuracy of the SOC modelling by using the RMSE (Root Mean Square Errors)[7] 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 and the approach and threholds used for development of the indicators.

[7] 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

A soil quality map at a parcel level

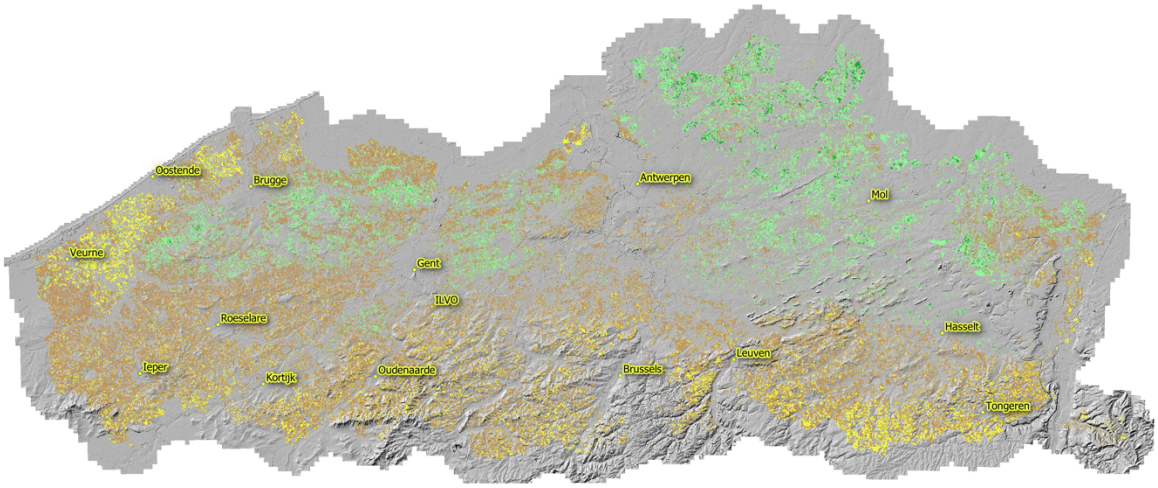


Figure 60: A soil quality map at a parcel level, using an indicator that informs if a parcel has a median Topsoil Organic Carbon value below the average, around the average, above the average and far above the average, considering soil-pedoclimatic conditions. As a background map, we have used the hill shade of Flanders.

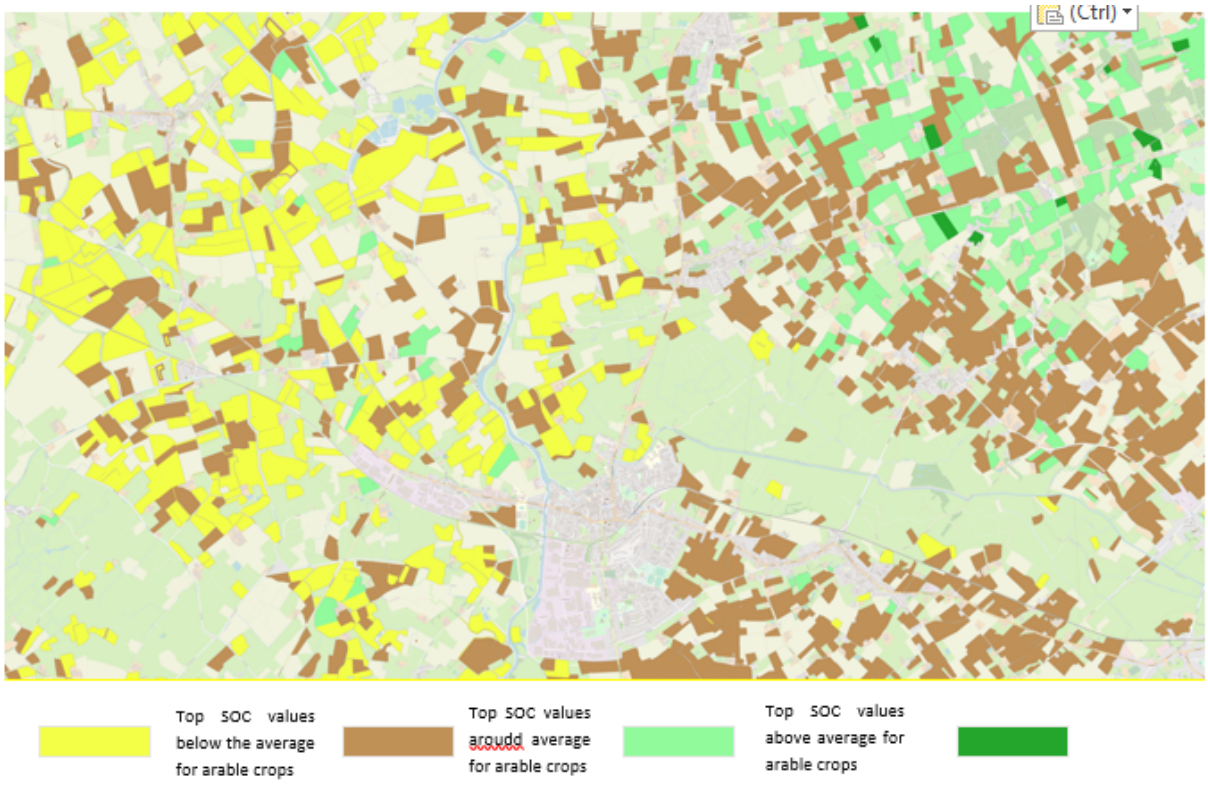


Figure 61: Zoom close to the selected window of Figure 16 to visualise the parcel variability of the Soil Quality Indicator.

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

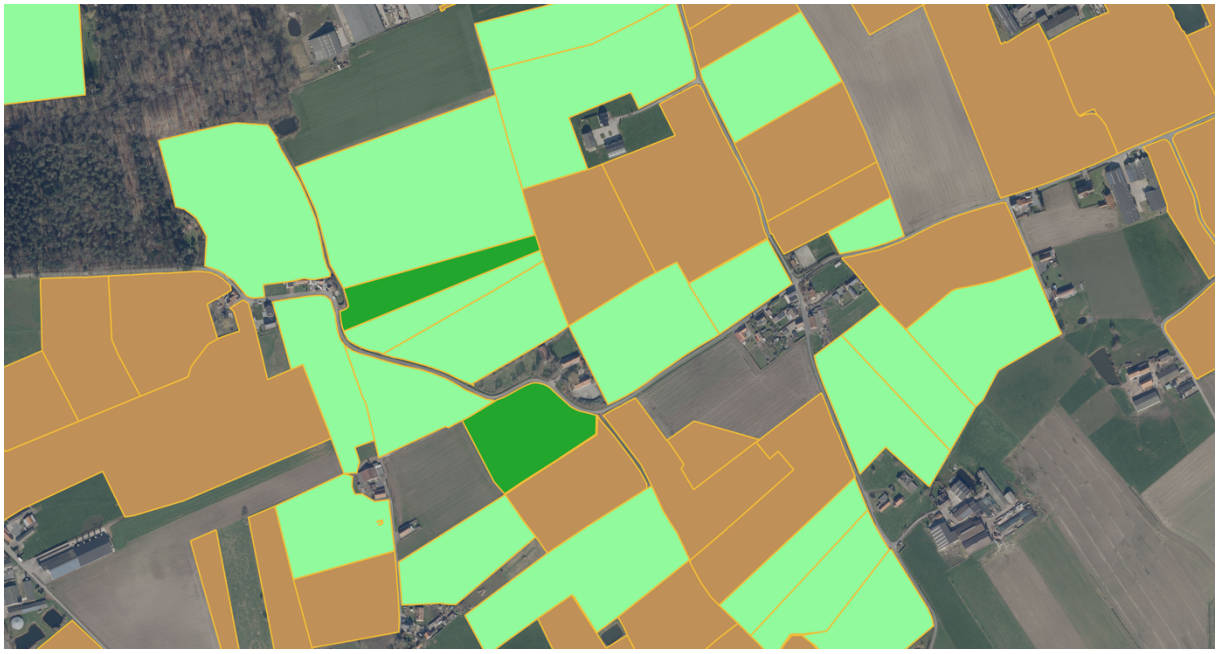


Figure 62: Detail zoom into the selected window of Figure 16, to visualise the the Soil Quality Indicator at parcel level, using very high-resolution orthophoto maps as a background layer.

All products are delivered regularly to the Envision platform repository and visualised at UI using mapping services developed by AgroApps.

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 data development phases and the technological tools we use at each phase (Figure 65).

Data Product 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 to ensure the needed agility on product development.

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 5398 images of L2A that covered the Flemish region from 2018 until the end of May 2022. The first step was to create a collection of S2 images using the GEE Python APIs[8] 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. As an alternative to this other STAC[9] services have been tested

[8] https://stacindex.org/catalogs/google-earth-engine-openeo#/
[9] https://stacspec.org/en

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

to untilize the load of satellite data resources using STAC API[10] from data cubes like EODC[11] or static catalogs like CREODIAS[12].

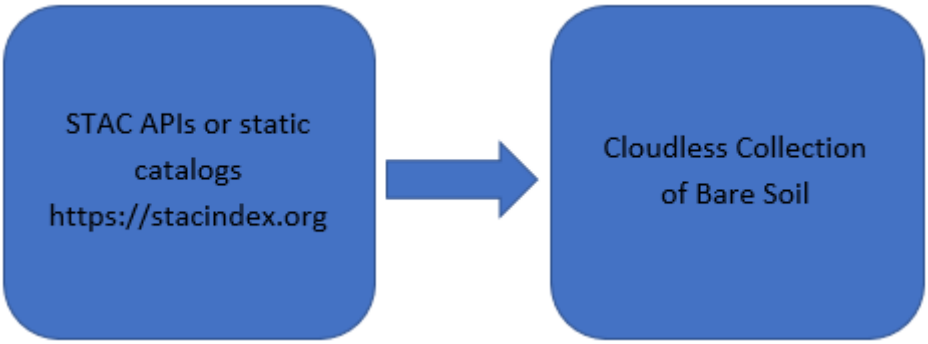


Figure 63: Utilization of STAC services for the development of a bare soil collection.

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 [13]

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 equal the existing crop lands and the grasslands. In Table 13, we provide the indices, formulas, and final threshold values.

[10] https://github.com/radiantearth/stac-api-spec
[11] https://stacindex.org/catalogs/eodc-openeo#/
[12] https://stacindex.org/catalogs/creodias#/
[13] 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

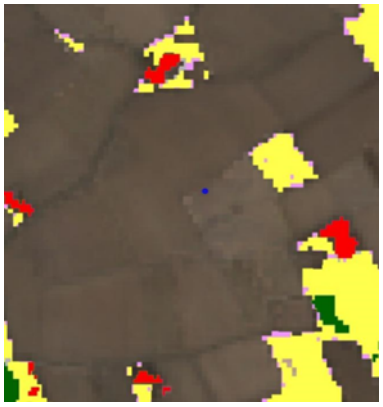


Figure 64: The masking works very well with croplands; however, most of the grasslands areas (yellow) do not belong to the Cloudless bare soil collection. The NDVI values remain high during the whole period, which means it's impossible to receive bare soil reflections.

Table 13: 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.

So different indexes are used in Envision to select the bare soil layer. The most important one is NDVI (Normalized Difference Vegetation Index), but NBR2 (Normalized Burnt Ratio 2) is also used to further fine-tune/ filter for the bare soil layer, and also VNSIR (Visible to Shortwave Infrared Tendency Index). These are al not soil Indices, but are indices to differentiate vegetation and moisture.

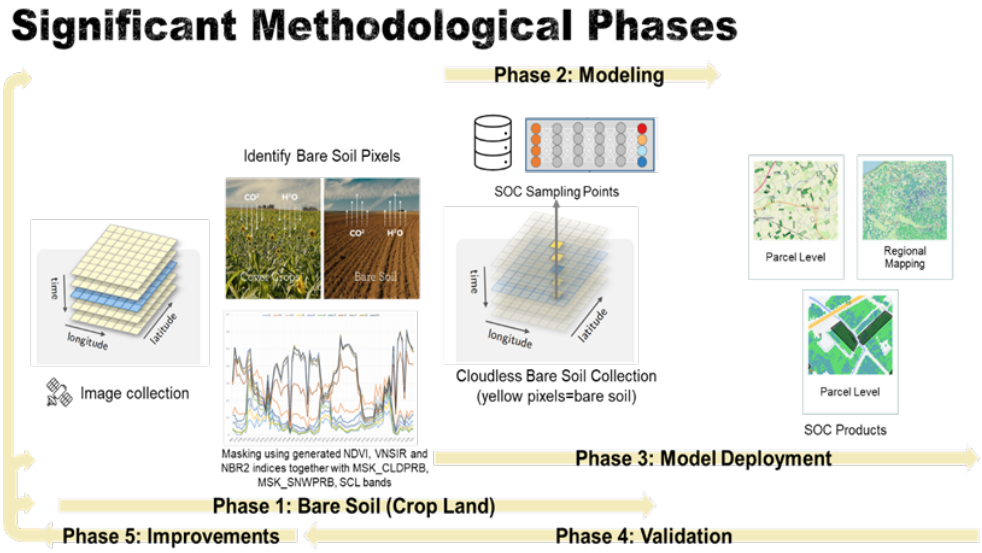
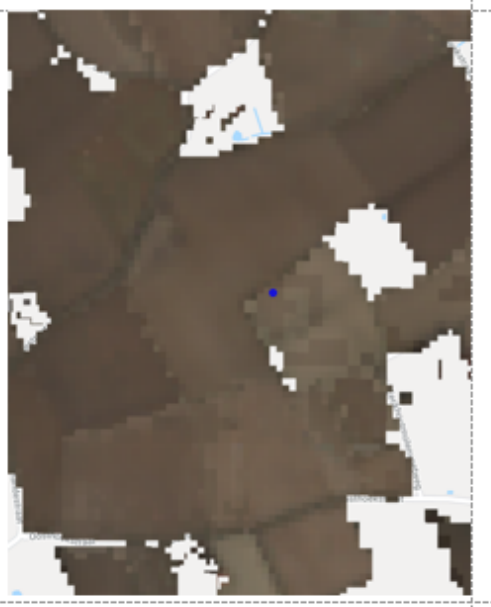


Figure 65: Significant Methodological Phases

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.

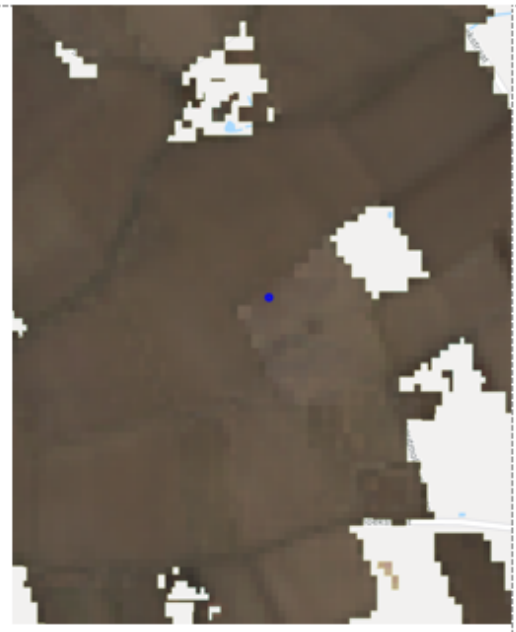


RGB visualization of a cloudless bare soil collection from May-2019 until May-2020.

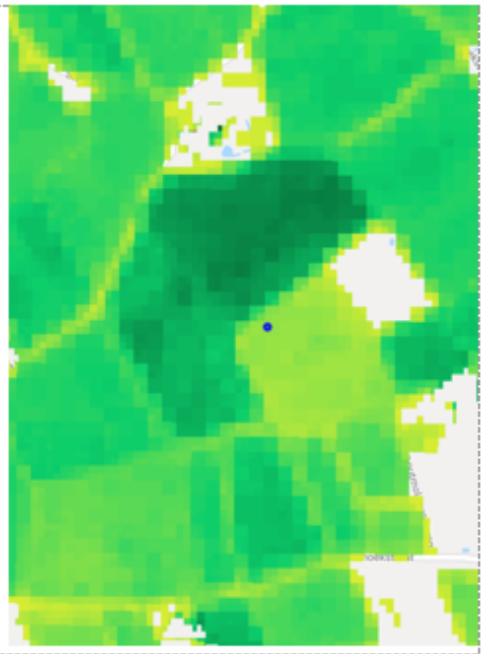
RGB visualization of a cloudless bare soil collection from May-2020 until May-2021. Mainly due to clouds, the cloudless bare soil collection does not cover the sampling point area.

Figure 66: RGB visualisation of continuous-time period stacks of the cloudless bare soil collection area around a soil sampling collection point (point ID 33).

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. 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 67: 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 as it is presented to the right picture (example for the period of the first 3 years). 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).

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

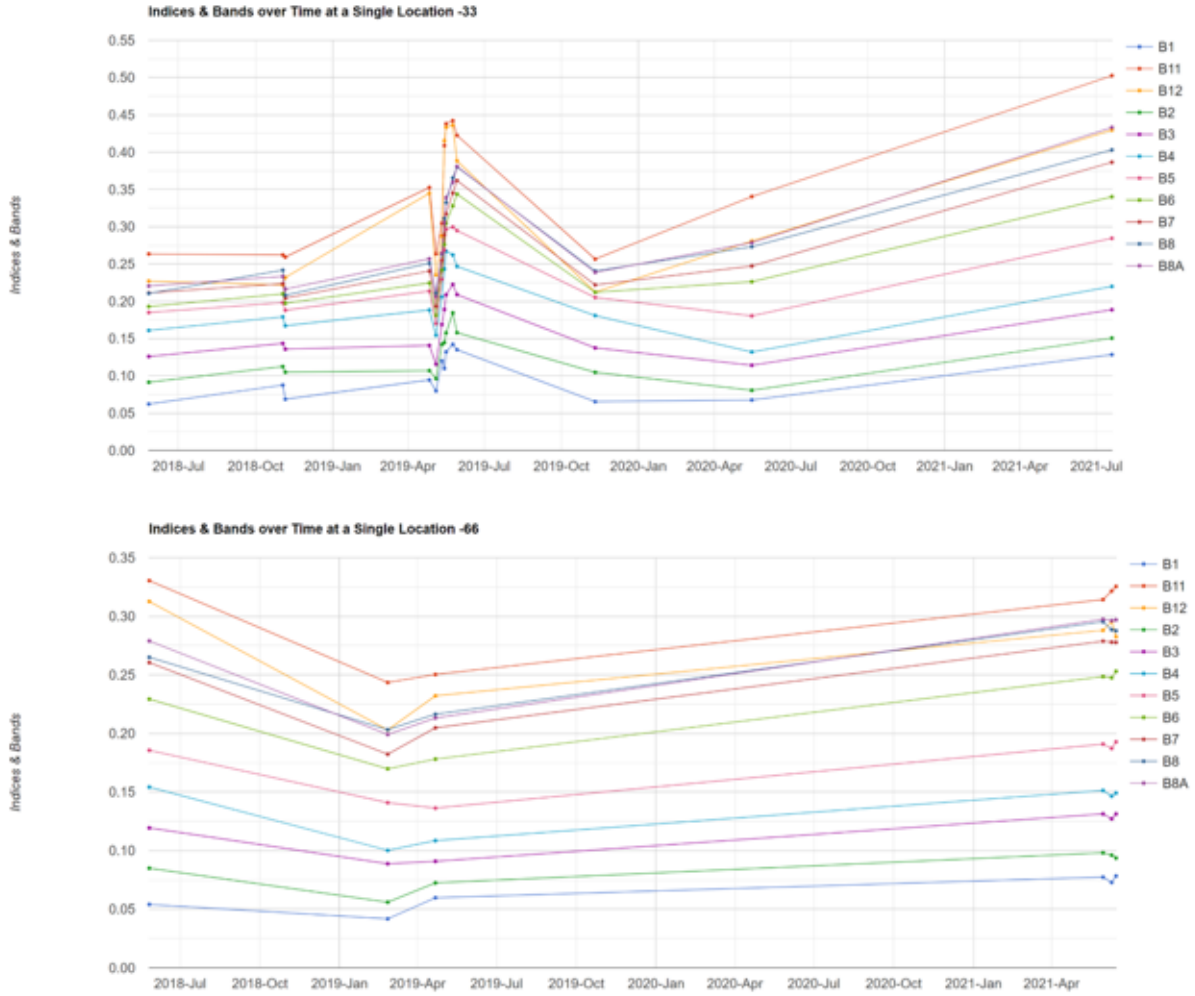


Figure 68: Location 66 has a measure SOC of 1.82%, much higher than location 33 (0.72%). Graphs present the reflection per S2 band for the period of May- 2018 until 2021.

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

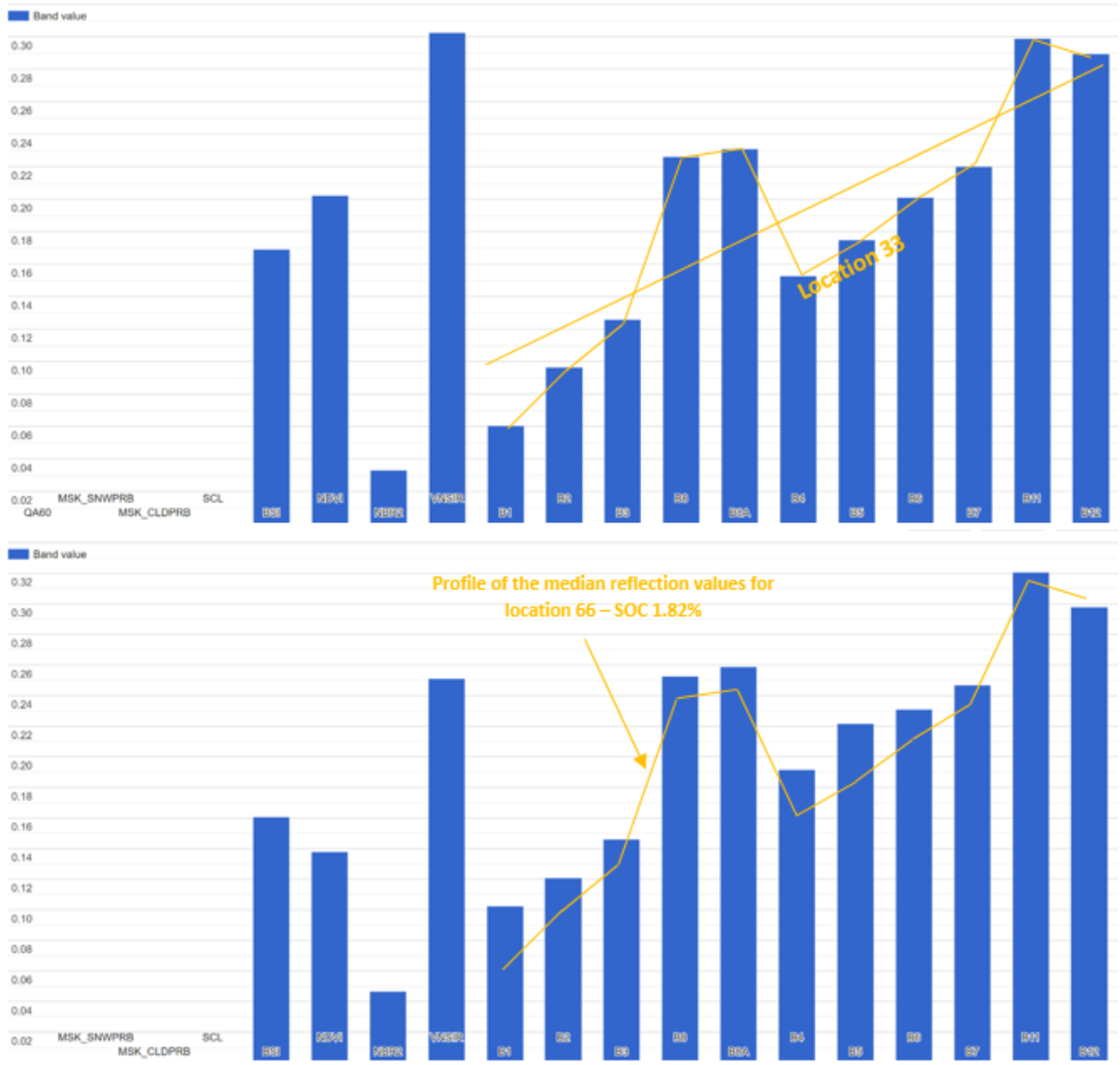


Figure 69: Median reflection values per band from May 2018 until May 2021 for sampling points 66 (upper) and 33 (down). Location 33 has a measure SOC of 0.72%.

Phase Two: Modelling

In **Phase Two**, we perform the modelling and the goal is to create a mapping between the extracted reflection signatures from the Bare Soil Collection and the SOC measurements. Using the GEE Reducers[14] and the Exporting Data[15] ability, we generate raw data per sampling point (Figure 70).

[14] 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)

[15] 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

For the training, we use the sample points' reflection signatures coming from the S2 bare soil collection. Those points (locations) have been selected within arable crop parcels, cover the whole Flemish region.

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

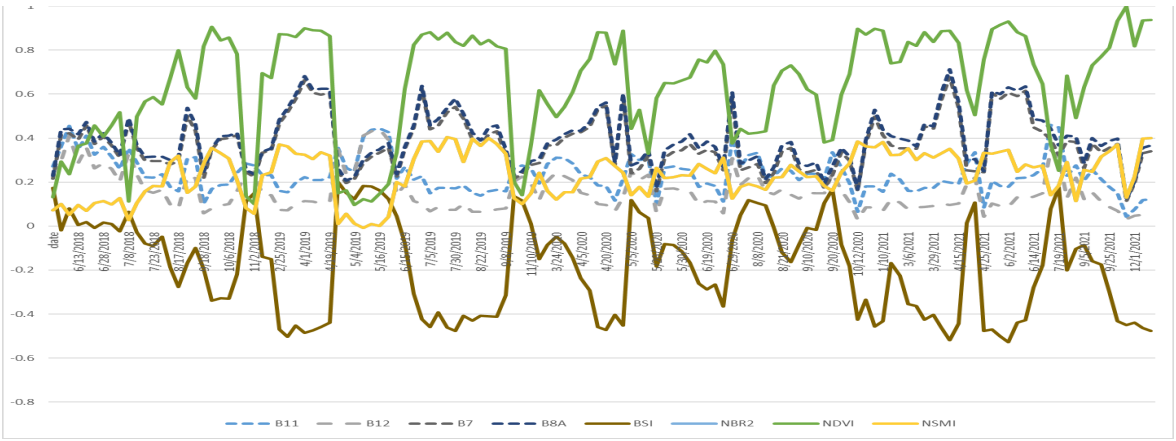


Figure 71: Visualization of reflection bands and indices for the sampling point 33. In total, we have 131 reflection signatures for the period of May- 2018 until the end of 2021. Only 13 reflection signatures correspond to bare soil (10%).

The next step in the modelling process is to link the reflection signatures of each sampling point, which covers a pixel area, 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.

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

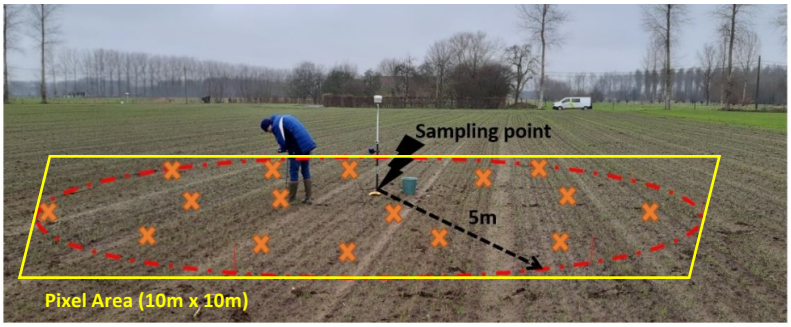


Figure 72: The soil sampling collection area should be within a pixel area. Otherwise is not logical to link the reflection signatures of bare soil pixels with the lab SOC measurements.

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 (Figure 73).
- Make use of sampling points with more than five bare soil pixels (quality hreshold) with the bare soil collection layer.

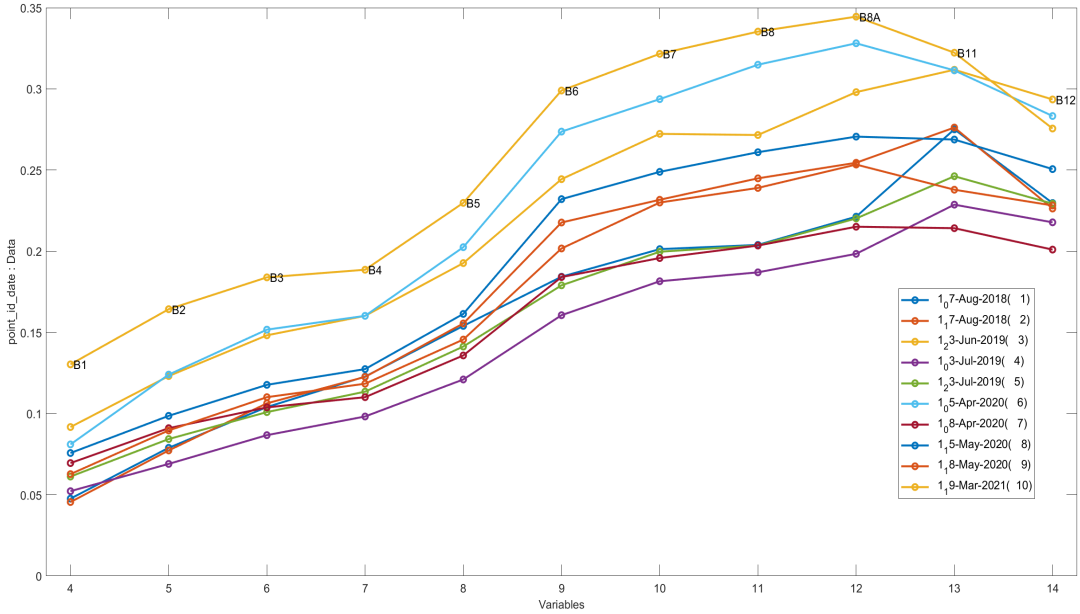


Figure 73: Visualization of all reflection signature for a single point, for a set of available satellite imageries.

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,

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

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 74).

Figure 74: PYCARET allows collaboration, ensures scalability, and supports productivity.

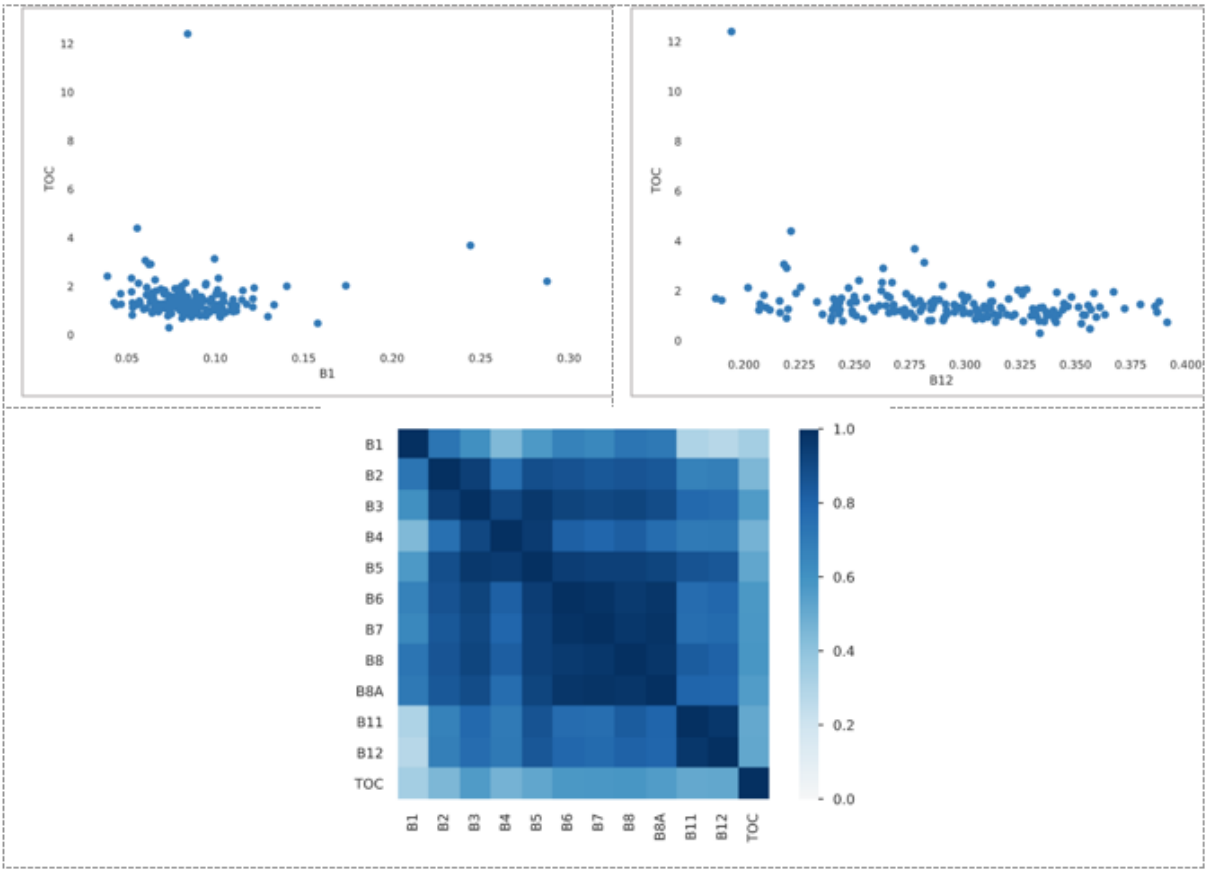


Figure 75: Part of the profiling report presents the Interactions and correlations (Phik) between input and output parameters.

Phase Three: Model Deployment

In **Phase Three,** we apply the deployed model to all pixels belonging to larger areas at the regional (Envision, BC3) or even national scale. The deployment of machine learning models makes models available in production.

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

Web applications, enterprise software, and APIs can use the trained model and generate predictions for new pixels. 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. For the needs of the project we have select to apply the batch process and the outcomes are geotiffs with one band that contains the model predictions per pixel size of 10m by 10m.

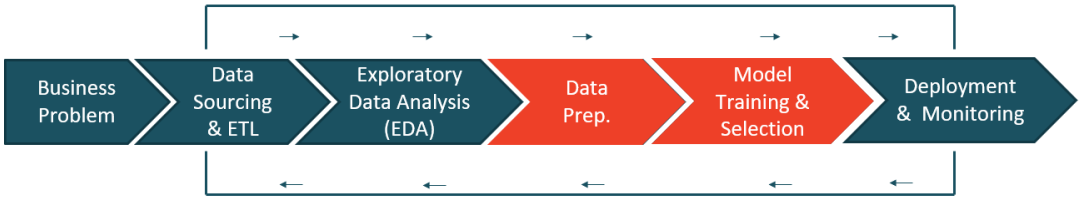


Figure 76: PyCaret — Machine Learning High-Level Workflow

Figure 77: 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.

In this phase, critical decisions are also made on how to present the information to the service consumers or the end-users. At the Envision project the goal is provide Soil Quality Indicator for Top Soil Organic Carbon, considering also the model accuracy and the specific CAP needs for monitoring coming from LV.

The basic model is a regression model, with absolute values. For the product, see previously in the deliverable, is relative based on the distribution approach compared to expected OC values for certain soil textures. This value can change dependingon the monitoring (AMS) area, and the soil texture.

At the beginning of the project, we classify the top Soil SOC predictions (float values between 0 and 4, presenting the % of top SOC) to classes using as reference the suggested classes presented in Table 12 and ignoring the texture info, 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.

The next step, and after the evaluation of the results by LV, was to follow the Value base and the distribution approach as explained in the sections "Soil Health and Soil Quality" and "Flemish Soil Quality Indicators, soil-pedoclimatic conditions and approaches tested.", for the development of the data products.

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

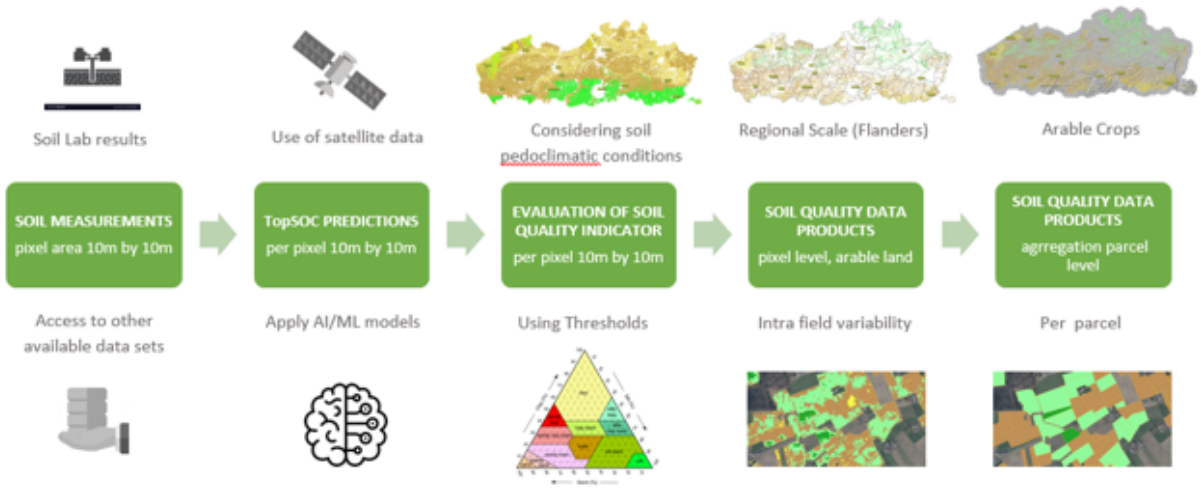


Figure 78: Development of Soil Quality Data Products

We also aggregate the SOC from pixel level to parcel-level at this phase (see Soil Quality data products). LV has provided the needed data sets to develop the data products, consisting of the Flemish Agricultural Parcels. This file aggregates the Top Soil Organic Carbon Assessments at the parcel level. At least 50% of the surface area of the field polygons need covered by pixels that give an OC prediction in order to withhold the prediction for a parcel. But this parameter can be changed depending on conditions and specific needs of the business customer.

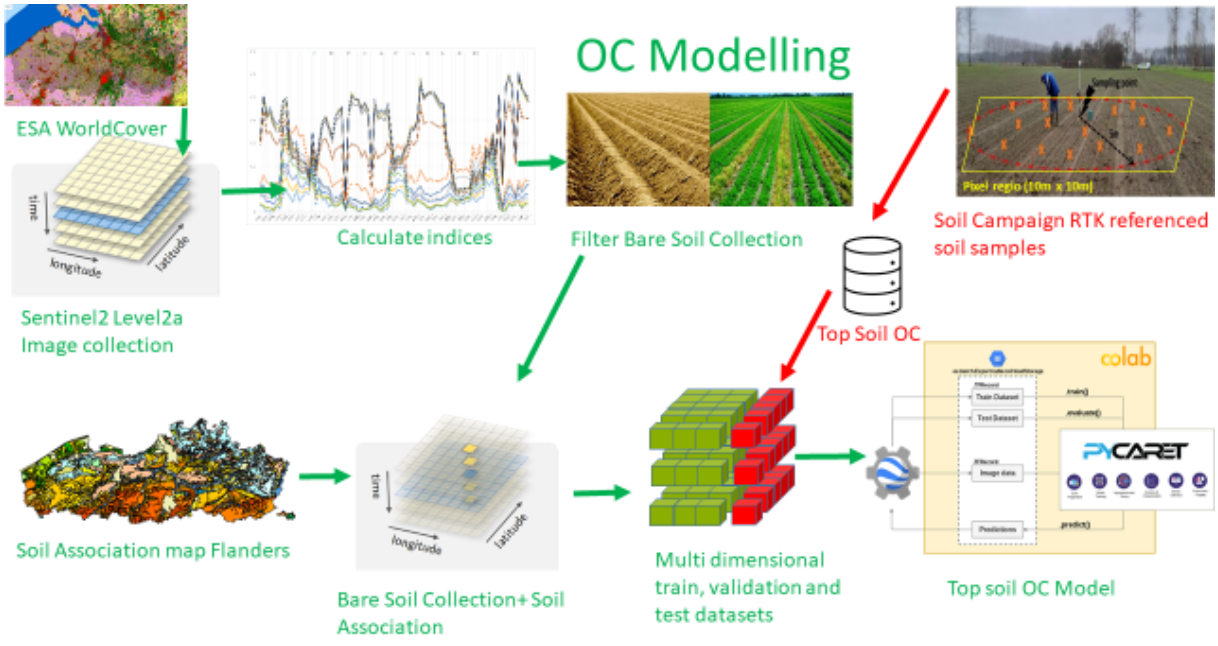


Figure 79: Input-Output schedule for OC modelling, general workflow

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

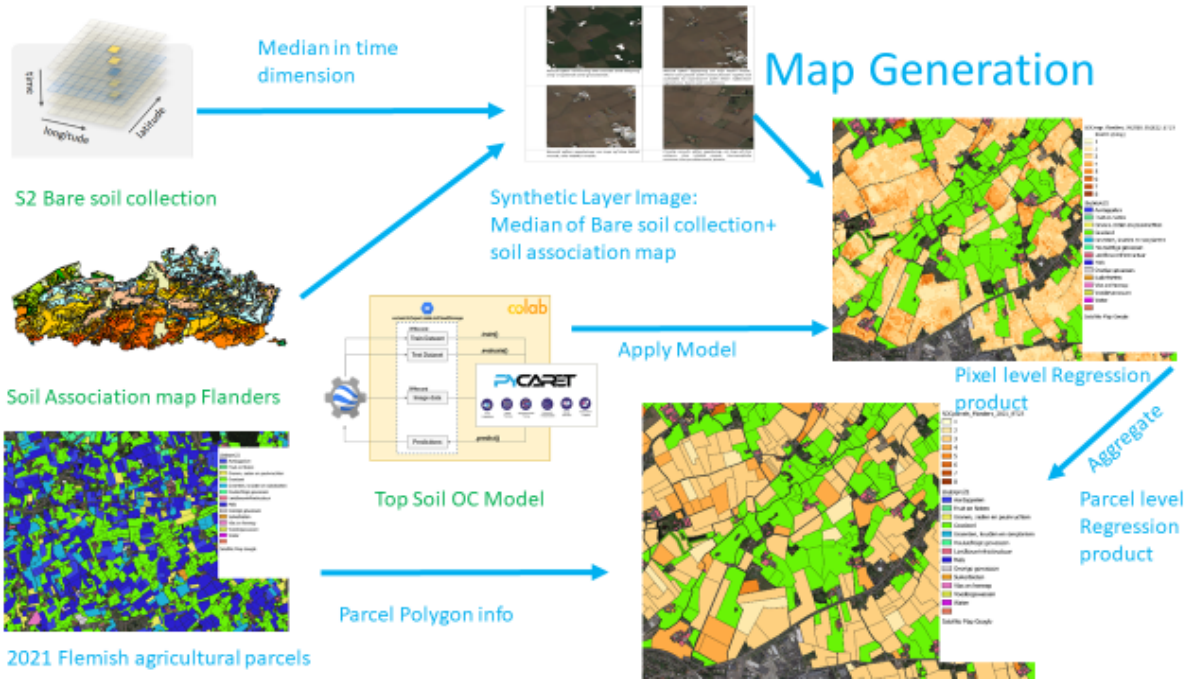


Figure 80: Map generation, input and outputs, general workflow

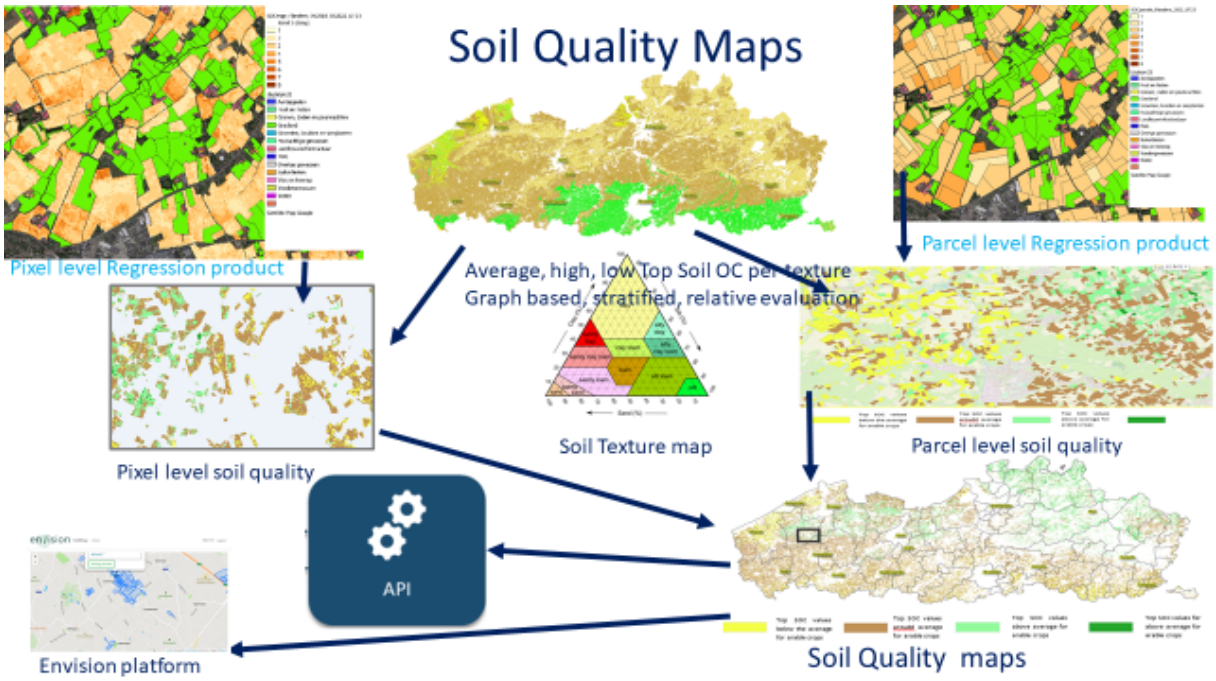


Figure 81: Data quality product, inputs and output, general workflow

Phase Four: 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.

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

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.
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.

Figure 82: sensitivity analysis results for the soil type to evaluate model performance.

It is possible to deploy this model worldwide, but it is now that spectral based models are not generic, it's a datadrive model, that depends on the specific region. The soil organic carbon product was not tested in other regions for validation and upscaling posiblities because you would need a ground truth dataset to verify it. Multiple top soil samples linked to organic carbon, linked to a RTK GPS points, are most of the time costly or protected data, and we did not have access to such a dataset. You could in theory apply the models build in Flanders to a new regions, but since models build on spectral information are always product and region specific, depending on which soil associations, general agricultural practices and the influence of specific terrains and weather, we expect that it will give more accurate results to use the local ground dataset to train new organic carbon models for that region using cross-validation , and hold out as we did in Flanders part of the dataset from the modeling as a test set. The same workflow as described in the different deliverables can be applied, and the thresholding for the bare soil layer can most likely be used again, especially in neighboring countries, although some checks of the synthetic bare soil layer should be taken since the exact optimal threshold values also varied in literature, so there is a chance they vary per region.

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

Use case definition, data processing logic, for other regions: A lot of the data layers are available world-wide (S2, Esa WorldCover, Soil Texture map). Some information will also be available in slightly different formats for other regions (LPIS, Soil association maps,) But the ML can take this differences in soil association into account, and from the LPIS layer only the polygons of the Agricultural fields need to be extracted.
The main necessary component is a reliable ground truth dataset for the new region.

Our approach to developing soil quality data products prioritizes efficiency and adaptability. We aimed to create a product capable of providing estimations at regional, EU, and even global levels from the outset. We achieve this by leveraging Earth Observation and reflection signatures, enabling us to directly predict soil quality conditions on a larger scale and adapt our models to varying regional conditions.

This emphasis on adaptability streamlines production and enhances operational efficiency, recognizing the diversity of agricultural practices and environmental conditions across regions. Whether to reuse existing models in similar regions or retrain models for differing conditions depends on the specific pedoclimatic and farming practices, and this decision is a key part of our approach

Roadmap to implement top soil OC monitoring in other area:
Ground truth dataset:

- Define area of interest (AOI).
- Collect available data sets:
    - Soil association map
    - Soil texture map
    - Recent Agricultural Parcels map of the Region (LPIS)
- Generate bare soil layer for the new AOI using existing infrastructure:
    - using S2 images
    - using cover map (World ESA)
    - using indicies
- Perfrom a control of bare soil layer: Look at indexes, synthetic layer median layer soil, effect on singular images.
- Evaluate if the existing numbers of sampling points is adequate under the condition that for the collection they have use the same procedure as exists in D3.5. More sample points are better if possible.
- Apply OC model pipeline. Check on unseen/ independent data, simular spectral signatures/ soil associations should be present in the train set.
- Generate OC Regression map.
- Generate Soil Quality Data product using indicators considering pedoclimate conditions.

Limits of the algorithm: Depends on existing correlation for OC in a certain soil association with the spectral signature, OC variation in measurements, number of samples per soil association.

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

These kind of models also have problems with values of organic carbon that are outside of the calibration range. Enough examples of OC values of a certain range and with certain soil types are necessary in order to develop models applicable for that OC range and those soil types, and the whole range should include enough variation.

Accuracy measures Regression product: See data Validation 3.6 en 3.4

Phase Five: 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.

Lessons learned: For soil organic carbon, the implementation did give actionable results, but the accuracy was lower the requirements needed for a CAP monitoring in time using only the regression model, and for accurate quantification. So the focus was on visualizing the results and giving information linked to which zones in the fields are probably higher , and the differences between fields in the aggregated data product, using the distribution method en the data quality product. This can assist targeted sample taking, or give an idea in precision agriculture which fields or which zones in the field require more care in order to increase the organic carbon content, or give insight in changes ove time.

Significantly more data will be needed, in the case that an accurate quantification is needed, and this combined with evaluating additional data sources, using also hyperspectral data and other parameters. EV ILVO will follow this up in the future projects like the New Horizon project ScaleAgData.

Automatisation:

The Pycaret framework was used to evaluate on an equal basis different models and parameters for these models.

Automation was done by running all the scripts in a Colab, notebook environment, using parallel processing, dividing the Flemish region into several sub-units, and combing them later in 1 map. Manual optimization was used to determine the thresholds for bare soil, but they can afterward be applied for the whole region, in a fully automated workflow. Manual sample taking is also necessary for the ground truth dataset. In the follow-up projects Scale-AgData, the workflow was dockerized and fully automated in Google cloud.

To summarize:

- We managed to reduce the needed time to generate the model by 80%.
- We managed to reduce the time needed to developed the data products by 70%.

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

Organic Control System Subotica (OCS), established in June 2003, holds the distinction of being the first domestic control organization dedicated to organic production in Serbia. In its inaugural year, the organization sought expertise from a renowned control house in Germany, recruiting their expert to train and equip its employees with the necessary skills for conducting control and certification tasks aligned with the standards set by the EU and USDA/NOP. Since its inception, OCS has diligently adhered to domestic legislation, prioritizing the promotion and growth of organic production in Serbia. A key objective has been to foster trust among domestic producers and consumers in organic products, contributing to the expansion and sustainability of this sector.

4.5.1 Data product description

This product provides an 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 benefits 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 high-resolution optical satellite images and the derivation of image features depicting the phenological status of the cultivated parcels. Machine learning classifiers (MLC) have been trained to understand the temporal and spectral signatures of conventional and organic crops. The Predictor Layers emerge from Sentinel-2 MSI multispectral bands and include.

- Vegetation Indices – Quality masked - Gap Filled NDVI 6d step timeseries
- NDVI timeseries temporal derivatives
- GLCM Image Texture Features
- Crop Phenology features

To support the creation of classification predictor layers, the service relies on a cloud-based processing framework of EO data 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 has been used as the primary resource of retrieving Sentinel data, as according to the Deliverable 3.1 Cost-benefit analysis, seemed to be currently the best-fit solution for ENVISION in terms of budget and the offered services. Specifically, Sentinel-2 Level-2A have been exploited. The Sentinel-2 Level-2A products are offered in most of the cases as Bottom of Atmosphere (BOA) reflectance images derived from the associated Level-1C products.
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

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

has been integrated and delivered as an earth observation component of the ENVISION platform with geographical coverage across the Serbia region.

The general contribution of the product to ENVISION, is towards the replacement of direct and guide on-field checks for priority control and resulting 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.

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 is characterized with the confidence of its classification decision (red, green, blue). These smart inspections methodology identifies potential breaches of compliance and assign the appropriate colour to suspicious parcels declared as organic, based on their deviation from the classification decision.

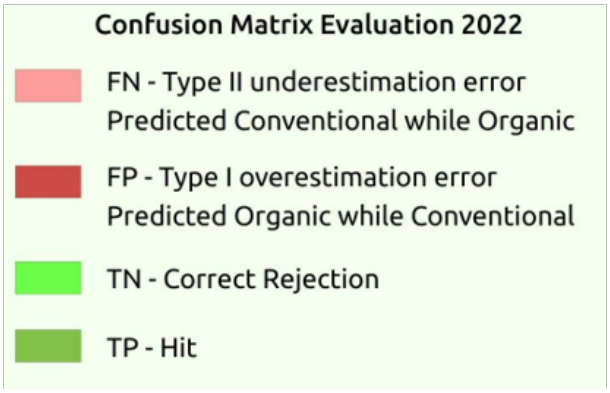


Figure 83: Confusion Matrix Evaluation

The data product D5 which regards the Distinction of Organic Farming Practices" is a vector geospatial feature which is served through the ENVISION platform, or alternatively could be served via WFS to the user, in a shapefile format. It contains the parcel polygon boundary geometries,an as far as

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

attributes, the evaluation of the Classification as Organic/Conventional farming practice, in the representation of a traffic light system. Its values are given in a standardized confusion matrix terminology, and depict the result of the prediction in regards with what was initially declared. The prediction is decided by a configured threshold value on the classification probability , which is the actual output of the algorithm. If the spatial average within the parcel bounds is higher than 0.5 the parcel is inferred as organic. The traffic light values are the following:

- False Negative (FN) also known as type II underestimation error if a parcel was predicted conventional while being organic
- False Positive (FP) also known as type I overestimation error if a parcel was predicted organic while being conventional
- True Negative (TN) if a parcel was predicted correctly as conventional
- True Positive (TP) if a parcel was predicted correctly as organic

Considering the above-mentioned traffic light value definition, a parcel is not conforming when it is predicted as organic while being conventional (FP).

The classification probability threshold was decided by optimization after visual inspection and analysis analysis of the ROC curves of the Internal Cross Validation. It was set to a value that minimized False Positive Rate and maximized True Positive Rate, that approximated 0.55. This value was hard coded in the training and inference scripts, and at the moment it is not a free parameter to be determined by the user.

The service works with no free parameters set by the user.

- Data import is a standardized procedure with specific data schema requirements
- Regarding EO Feature Extraction, the envision experience proposed the creation of specific indices. This step could be parameterized, allowing the used to define himself a subset of the proposed features, or the temporal interpolation timestep
- Outlier Detection : An option to upload in situ field visit data could be included , in order to bypass the visual inspection step
- ML Training / Prediction : The threshold of classification probability could be parameterized, to enable the user to fine tune his decision making

The level of automatization varies among the different components of the data processing flow of the service.

- Field Data Import : Fully Automatized LPIS+GSA data subset import to the database
- Spatial Data POSTGIS processing + Descriptive Stats : Partially Automatized (internal sub-processes need to be interconnected)
- EO Data Import : Fully Automatized import from CreoDIAS and Copernicus Dataspace APIs + Atmospheric Correction for L1C to L2A
- SoilGrids Import : Fully Automatized
- EO Feature Engineering : Fully Automatized
- Training Dataset Creation : Fully Automatized
- Data Outlier/Anomaly Detection : Not Automatized – Involves Visual Interpretation of NDVI Profiles. The import of Ancillary Vector Data of In Situ Field Visits (RFVs) could bypass this manual optimization, which is however much important, so that Data Anomaly detection could be trained in a fully automatized way.

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

- ML Classification Model Training : Fully Automatized
- ML Model Inference on EO Raster Features : Fully Automatized
- Retrieval of ML Inferance results to parcels : Partially Automatized
- OGC WMS to user : Fully Automatized

The validation outcome on every parcel is mapped as a traffic light system symbology that makes confusion matrix terminology regarding the relation of predicted vs declared classes (TP/TN/FP/FN) (more information are available on D3.6).

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.

Within this context, the section addressing the external assessment of the trained model, using an independent test set, highlights the primary evaluation metrics extracted from the confusion matrix. These metrics undergo further scrutiny in deliverable D3.6, where they are tested against novel data from the pilot business cases.

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 [12] and PhenOTB [13] remote module
- R Libraries : mlr, caret, tidyverse, raster,sp,rgdal,tiff,ggplot2, maptools, zoo, signal, timeSeries, doParallel, dplyr [14-15]
- Python libraries :rasterio, numpy, pandas,seaborn, matplotlib, imblearn, scikit-learn, xgboost

The general methodological framework for the training of ML models for organic practice identification is presented on the following flowchart (Figure 80).

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

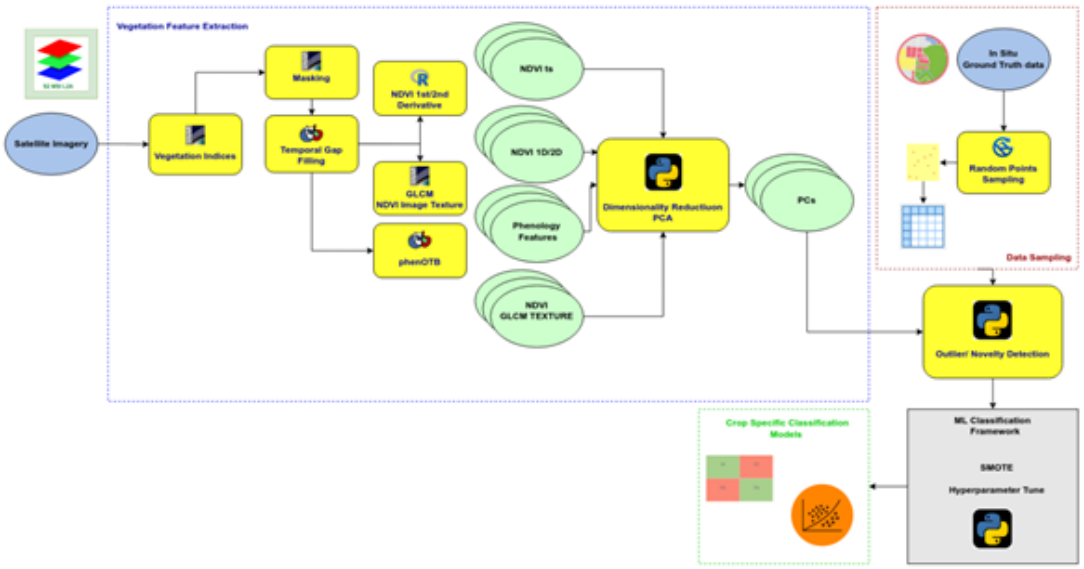


Figure 84: Methodological framework for the training of ML models for organic practice identification

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. For the creation of the NDVI datasets for seasons 2016 and 2017 L2 images were not readily available and for that reason L1 Scenes were preprocessed for atmospheric correction using the default Sen2Cor configuration.

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 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. Quality Mask file of every acquisition date was saved as well, for pixel quality evaluation for the whole timeseries.

The search and acquisition of field data from fields, with organic and conventional agricultural practices, by the certification bodies, yielded a polygonal data distribution with a very large dispersion in time and space, which made it impossible to manage them at the tile level since the space required for primary images and generated features exceeded the available storage resources. It was therefore decided that the creation of the necessary predictor layers should be done on a piecemeal basis and for each area of interest (AOIs). A prerequisite for the creation of the time series of image features

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

that would constitute the predictors of the classification models was therefore some kind of pre-processing that included:

- ✓ **Image Mosaicking**: the creation of a mosaic for NDVI and Quality Mask levels that had the same acquisition date.
- ✓ **Spatial Subset** in the AOI regions of interest.
- ✓ **Layer Stacking**: Creation of a layered raster archive containing all layers of NDVI time series.
- ✓ **Temporal Interpolation** was applied on the masked NDVI layer stack, to fill the gaps created from image masking, as well as to create a regular temporal 6-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 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 and parameters, as shown on figure 85**:**

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

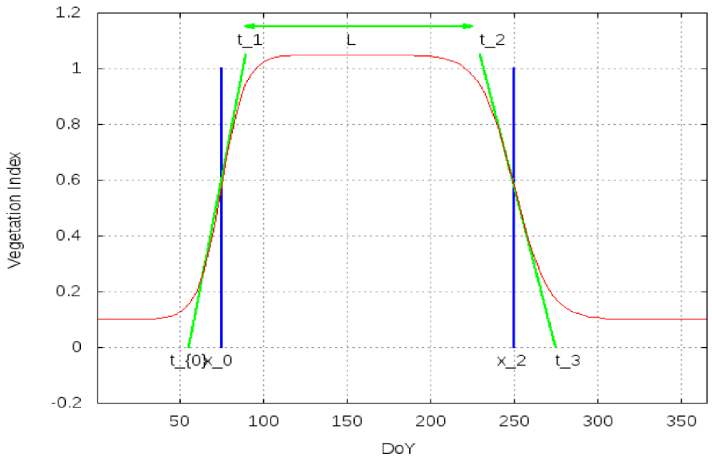


Figure 85:Sigmoid fitting on NDVI profiles and assessment of curve parameters

- Sowing date: t_0
- Date of Maximum Positive Gradient: x0
- Maximum Positive Gradient – Crop Growth Slope: D_{Growth} - $g'(x_0)$
- Parameter related with logistic growth rate: x1
- NDVI Plateu Initialization Date: t1
- Plateu Termination Date: t2
- Parameter related with logistic growth rate: X3
- Date of Maximum Negative Gradient: x2
- Maximum Negative Gradient - Crop Senescence slope: $D_{Senescence}$ -$g'(x_2)$
- Harvesting date: t_3

PhenOTB library works by 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.

As it is obvious from the above description of the phenOTB tool, many of the phenology parameters and metrics that are derived as features from the above analysis are actually dates in the form of DOY (Day Of Year) and cannot be used directly in training classification algorithms. They can, however, be used to calculate the duration in days of some broader phenological stages, in the primary crop growth cycle. It is therefore possible to distinguish 5 growth stages and calculate their duration in days.

- CGS1=x0 - t0: Early Growth Stage Lenght
- CGS2=t1 - x0: Late Growth Stage Lenght
- CGS3=t2 - t1: Plateu Stage Lenght
- CGS4=x2 - t2: Early Senecence Stage Lenght

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

- CGS5=t3 - x2: Late Senecence Stage Lenght

The use of phenology features in the classification models, therefore, was done through the calculation of these 5 growth stages, under the hypothesis that between organic/conventional farming parcels these time intervals are distinguished into individual clusters. The test of this hypothesis was carried out on individual cases (Figure 86) in the field data and appeared to be valid at some point.

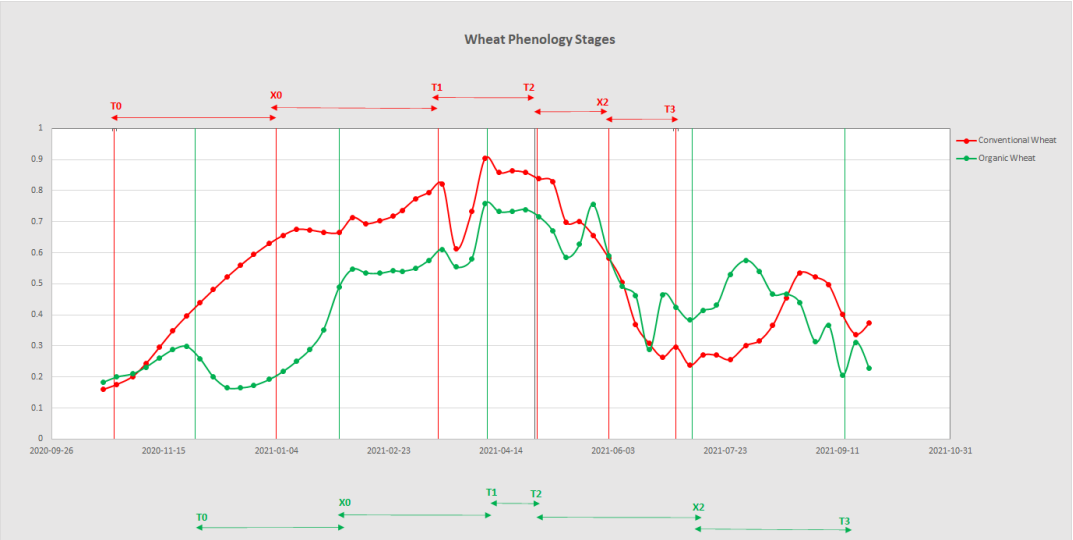


Figure 86: NDVI profiles of Organic/Conventional wheat and phenology stages duration

NDVI Derivatives:

During the early attempts made in the project to train classification models for the discrimination of organic from conventional agricultural practices, the following conclusions were drawn:

- The completeness of the spectral signature in the training data is not as important as it is e.g. in the case of land cover classification, and therefore the use of a generic vegetation index such as NDVI, related to the biomass status of the crop could be sufficient. Thus, it was not considered appropriate to create additional vegetation indices covering other regions of the electromagnetic spectrum.
- As in the general problem of crop identification, the temporal variation of NDVI is more important information to distinguish between the two practices. However, it became apparent that the temporal function of NDVI may "hide" additional information related to the rate of vegetation change in a pixel. An organic crop, not assisted by conventional agricultural practices, may show a different rate of change in NDVI throughout its growth phases. The information concerning the growth rate, the location of the extremes and inflection points of the NDVI, can be captured in the temporal 1st and 2nd derivatives of the specific vegetation index.

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

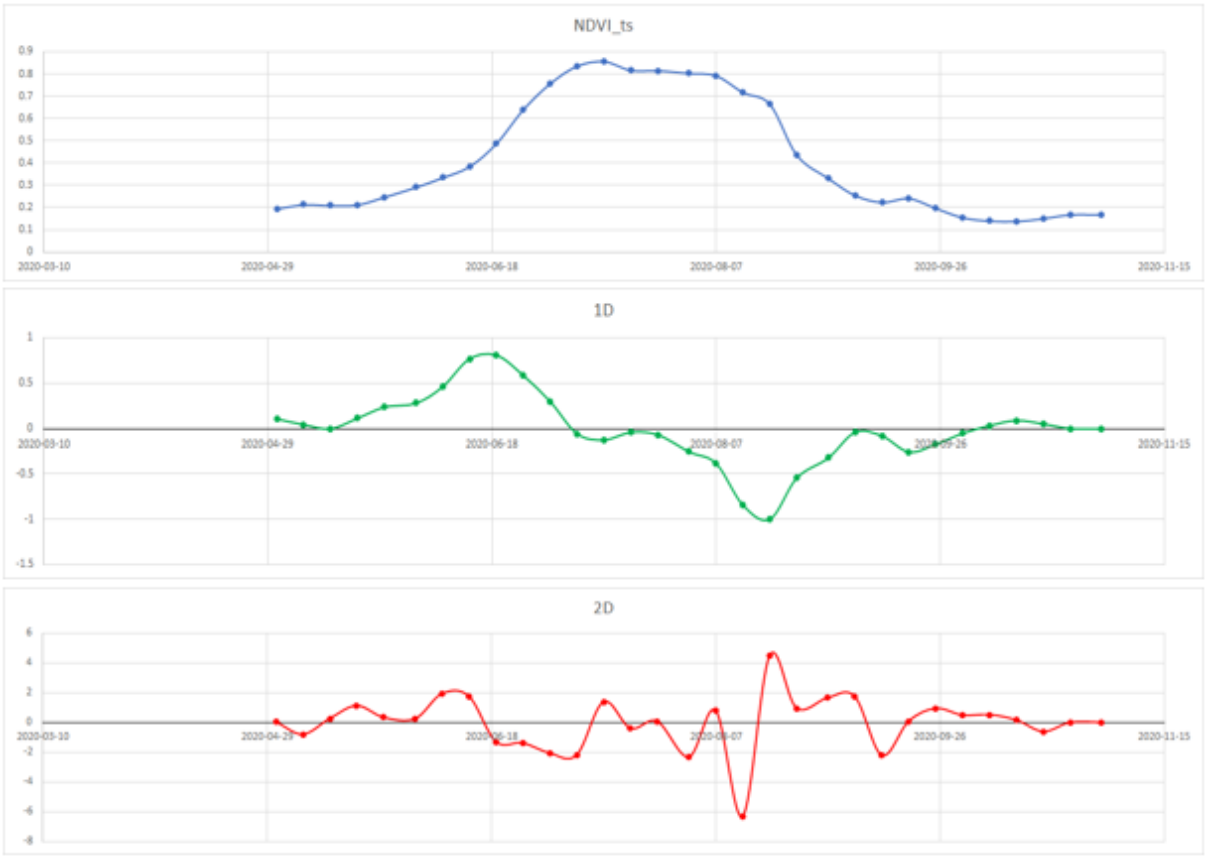


Figure 87: Application of Derivative Filters on NDVI profiles

Due to the high noise contained in the NDVI index curve signal, the derivation was performed in combination with the use of a smoothing filter. Specifically, the extraction of NDVI 1st and 2nd derivative layers with the use of Savitzky- Golay moving window filtering algorithm [16] (Figure 87), could accentuate the rates of change throughout the profile, and help on the classifier improvement, in a significant manner.

One of the most commonly used and frequently cited moving average filters in signal processing is the Savitzky-Golay smoothing and differentiation filter. It is often used as a preprocessing in spectroscopy and can be used to reduce high frequency noise in a signal due to its smoothing properties and reduce low frequency signal (e.g., due to offsets and slopes) using differentiation.

GLCM Texture Features on NDVI

During the evaluation of the 1st iteration of results it was observed that in organic crops there may be spatial heterogeneity of NDVI values, within the boundaries of the plot. This could be attributed due to the way fertilization is applied on organic farming. An assumption has been made that Organic vs Conventional farming practice may imprint significant spatial patterns and context of NDVI values regarding the homogeneity of radiometric values across different spatial lags. It was assumed that the assimilation of GLCM image texture features, such as Homogenity, Entropy and Variance, derived from the NDVI layers (Figure 88), would improve the classification results.

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

Spatial pattern information in the form of texture features could be useful for image classification. Texture measures provide new image features by making use of spatial information inherent in the image. Texture is the pattern of intensity variations in an image and can be a valuable tool in improving land-cover classification accuracy. Texture information involves the information from neighbouring pixels which is important to characterize the identified objects or regions of interest in an image.

The Gray Level Co-occurrence Matrix (GLCM) proposed by Haralik [17] is one of the most widely used methods to compute second order texture measures. By second order metrics, a relationship between groups of two pixels in the original image, is considered. Several texture features can be computed from the GLCM matrix, e.g., angular second moment, contrast, correlation, entropy, variance, inverse difference moment, difference average, difference variance, difference entropy, sum average, sum variance and sum entropy. Each feature models different properties of the statistical relation of pixels co-occurrence estimated within a given moving window and along predefined directions and inter-pixel distances.

The Grey Level Co-Occurance Matrix is a measure of the probability of occurrence of two grey levels separated by a given distance in a given direction. The features can be categorized into three groups, i.e. contrast group, orderliness group and statistics group. Thus, GLCM is a tabulation of how often different combinations of pixel radiometric values (grey levels) occur in an image, at different spatial lags. For the task of organic farming identification, the following GLCM metrics were calculated:

- Homogeneity (HOM) – related with Image Contrast Features
- Entropy (ENT) – related with Image Orderliness Features (how regular,"orderly", the pixel value differences are within the GLCM moving window)
- GLCM Variance (VAR) – related with Image Statistic Features

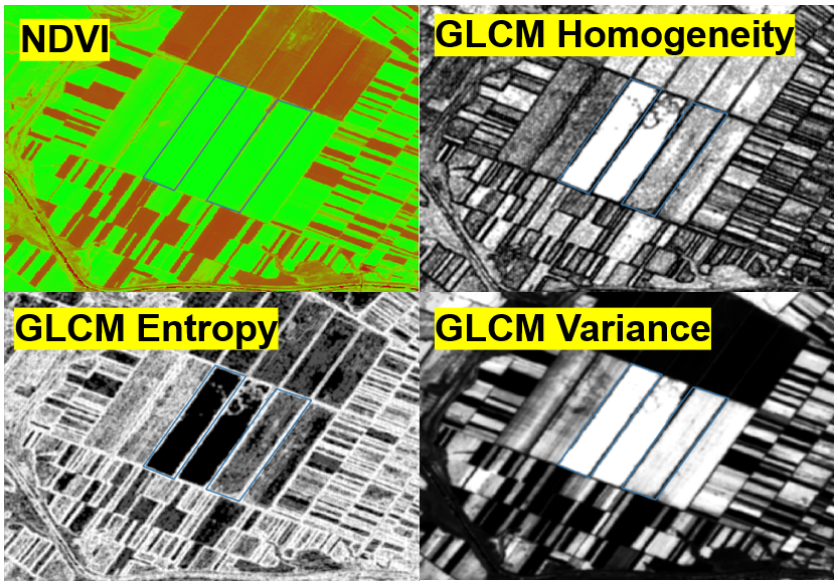


Figure 88: NDVI image texture from GLCM metrics. Homogenity, Entropy and Variance

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

Soil Properties

The use of soil property data as ancillary predictor variables in the classification models was based on the rationale that organic farming practices are favoured in soils belonging to specific soil texture classes and containing high organic matter content. The data used in the training of the algorithms came from the Soil Grids dataset.

SoilGridsTM (hereafter SoilGrids) is a system for global digital soil mapping that uses state-of-the-art machine learning methods to map the spatial distribution of soil properties across the globe. SoilGrids prediction models are fitted using over 230.000 soil profile observations from the WoSIS database and a series of environmental covariates. Covariates were selected from a pool of over 400 environmental layers from Earth observation derived products and other environmental information including climate, land cover and terrain morphology. The outputs of SoilGrids are global soil property maps at six standard depth intervals (according to the GlobalSoilMap IUSS working group and its specifications) at a spatial resolution of 250 meters. Prediction uncertainty is quantified by the lower and upper limits of a 90% prediction interval. The additional uncertainty layer displayed at soilgrids.org is the ratio between the inter-quantile range and the median. The SoilGrids maps are publicly available under the CC-BY 4.0 License.

Maps of the following soil properties are available: pH, soil organic carbon content, bulk density, coarse fragments content, sand content, silt content, clay content, cation exchange capacity (CEC), total nitrogen as well as soil organic carbon density and soil organic carbon (SOC) stock. The classification task involved the creation of the following soil parameter layers:

Soil Organic Matter: The calculation of the Soil Organic Matter (SOM) content was performed on the SOC SoilGrids layer (horizons 5 -30 cm averaged), using a conversion factor **(Figure 89).** Organic carbon content can serve as an indirect determination of organic matter using an approximate correction factor. The "Van Bemmelen factor" of 1.724 has been used for many years and is based on the assumption that organic matter contains 58 percent organic carbon. The literature indicates that the proportion of organic C in soil organic matter for a range of soils is highly variable. Any constant factor that is selected is only an approximation. The equation for the estimation of the organic matter according to this factor is the following one: OM (%) = 1.724 x OC (%).

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

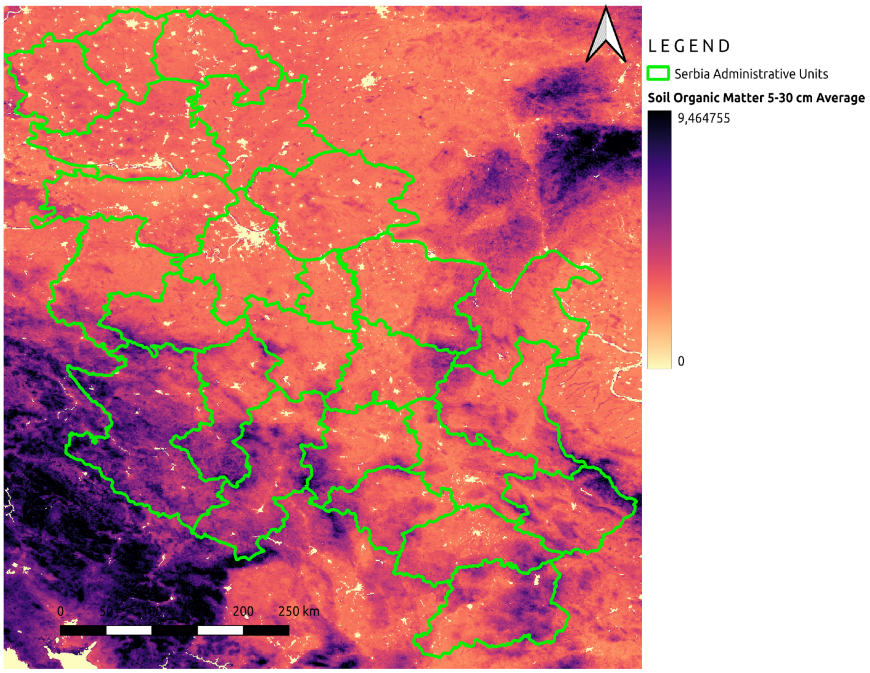


Figure 89: Topsoil Soil Organic Matter mapping of Serbia, as derived from SoilGrids Soil Carbon layers

USDA Soil Texture: Ranking into soil texture categories is performed at the Sand, Silt, Clay layers (horizons 5 -30 cm averaged), using techniques for reclassifying values in raster data, and guided by the USDA classification system (Figure 90).

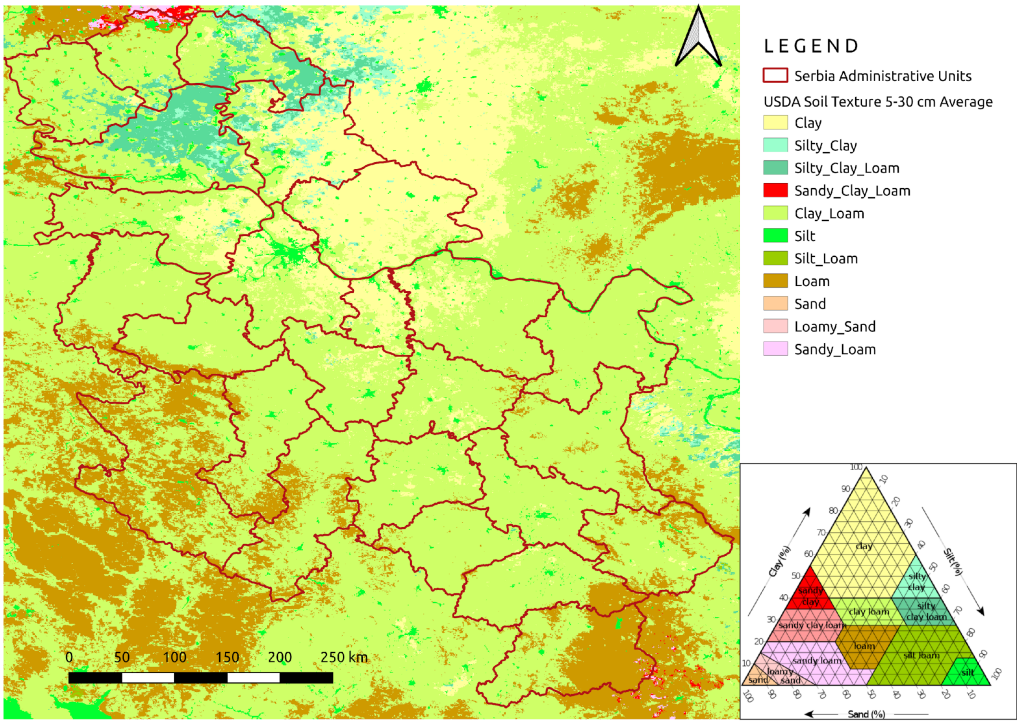


Figure 90: USDA Soil Texture mapping of Serbia, as derived from SoilGrids Sand,Silt and Clay layers

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

Dimensionality Reduction – PCA

From the evidence acquired from the previous attempts in the task of discriminating organic vs conventional farming practices with remote sensing, the need to extract more features that more faithfully describe the two categories to be discriminated, became apparent. In fact, according to Bellman, 1961 [18], in order to estimate an arbitrary discriminative function with a certain accuracy the number of features (or dimensionality) required for estimation, grows exponentially. In the case of the task in progress, dense time series of Layerstacks had to be created, leading to a very extensive and multidimensional feature space.

Issues that arise with high dimensional data are:

- Running a risk of overfitting the machine learning model.
- Difficulty in clustering similar features.
- Increased space and computational time complexity.

The situation where a huge number of feature predictor variables is used to train a ML classification model is problematic. The higher the number of features, the more difficult it is to model them. This is known as the curse of dimensionality in data science. Additionally, some of these features can be quite redundant, adding noise to the dataset and it makes no sense to have them in the training data. This is where feature space needs to be reduced.
The process of dimensionality reduction essentially transforms data from high-dimensional feature space to a low-dimensional feature space. Simultaneously, it is also important that meaningful properties present in the data are not lost during the transformation. To tackle the curse of dimensionality, methods like dimensionality reduction are used. These techniques are very useful to transform sparse features to dense features. Furthermore, dimensionality reduction is also used to clean the data and feature extraction.

To address the increased size of the feature space created for the classification task, Principal Component Analysis (PCA), was used for signal decomposition and dimensionality reduction. PCA helped to the identification of patterns in data based on the correlation between features. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one. The orthogonal axes (principal components) of the new subspace can be interpreted as the directions of maximum variance given the constraint that the new feature axes are orthogonal to each other.

Through the PCA process that preceeded ML training (and prediction), the initial features were standardized columnwise, and their covariance matrix was constructed, to be decomposed into eigenvalues and ranked eigenvectors. The Singular Value Decomposition (SVD) algorithm was used for the computation of the eigenvalue problem. Finally, a subset of the eigenvectors space, explaining 95 % of the initial feature variance, was finally obtained, providing the dimensions of the transformed, low dimension, feature space.

Training Data – Response Variable (Y)

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

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. The distribution of data samples across the class labels is shown on Figures 91, 92, 93. Those 4201 parcel records refer to parcels of different crops, years and farming practices as follows:

- 2335 conventional parcels
- 1866 organic parcels

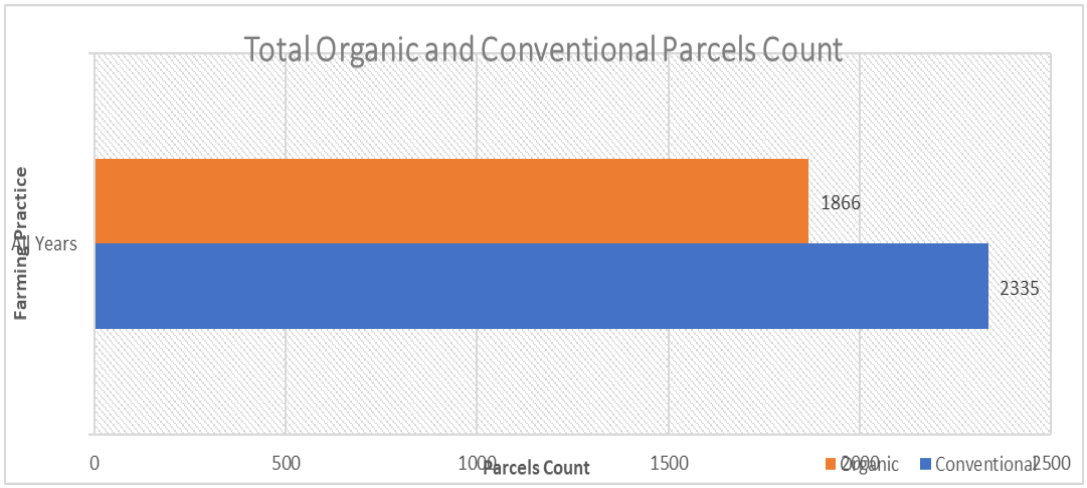


Figure 91: Total Organic and Conventional Parcels Count

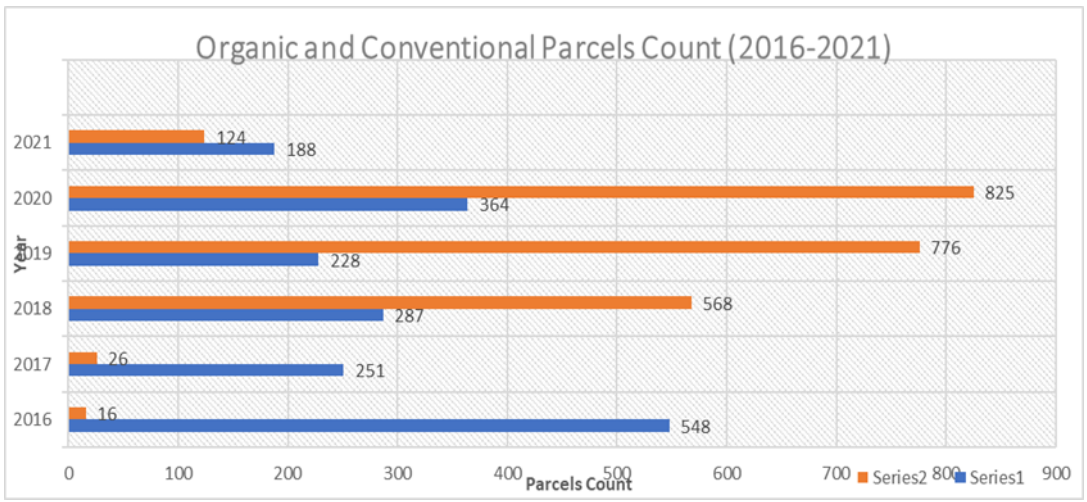


Figure 92: Organic vs Conventional class imbalance over the yearly (2016-2021) in situ data gathered

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

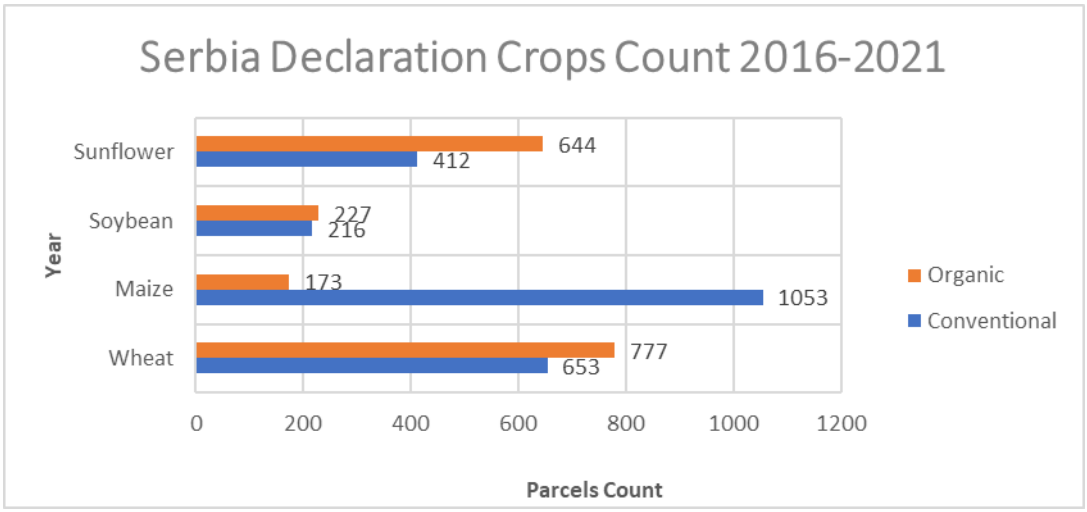


Figure 93: Organic vs Conventional class imbalance on the total dataset

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.

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 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 94).

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

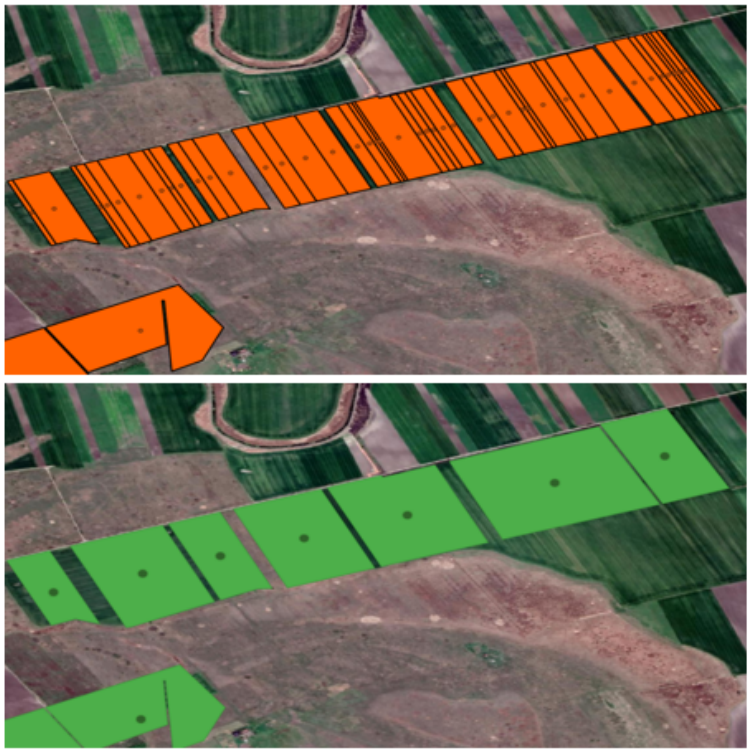


Figure 94: Dissolve preprocess on in situ data belonging to the same crop type/practice

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

Table 14: 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

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

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 95).

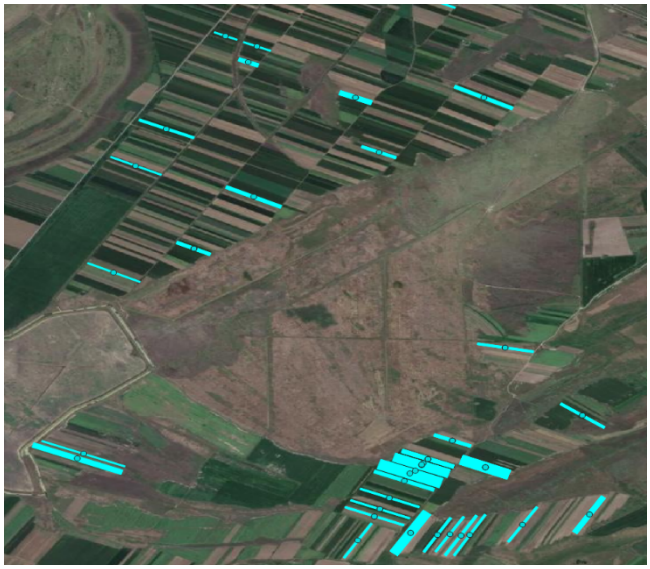


Figure 95: Scattered parcels with elongated geometry

Finally, there were cases of parcels that contained land cover not relevant with the crops, like bush/tree boundaries or roads (Figure 96).

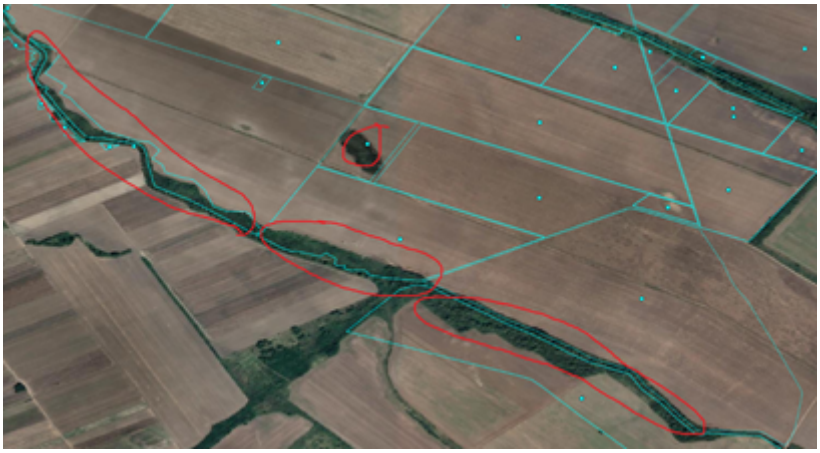


Figure 96: Data Cleaning

The training dataset is emerged by a random point sampling procedure inside parcel bounds, after the exclusion of an inner buffer zone of 15 meters. The minimum distance between sampling points is set, considering a 3x3 window as a region of influence of each point, to 40 meters. On the classification inference, the result is given on a parcel level, using zonal aggregate statistics (mean classification probability and majority of predicted class). In order to consider the result representative, a minimum parcel area of 10 pixels count is defined.

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

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.

Machine Learning - Data Anomaly Detection

One of the main concerns, related to the quality of the training data, was their fidelity on the declaration of farming practice, organic or conventional. For this issue, an attempt was made in a previous iteration to improve the quality of the data, by performing outlier analysis. PCA was performed , 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 T2 criteria. The results of the procedure were unsatisfactory as shown by the evaluation of the trained ML models, since the dataset remained contaminated with highly noisy data that were essentially false statements. The reason behind this was due to the purely unsupervised nature of the method, as well as even the subjectivity involved in imposing a distance threshold on a transformed feature space.

Thus, a hybrid methodology was chosen to follow for the "cleaning" of the dataset including visual interpratation in combination with novelty/outlier detection using ML algorithms. In the broader procedure of data anomaly detection, outlier analysis is used the case where training data contains outliers which are defined as observations that are far from the others. Outlier detection estimators thus try to fit the regions where the training data is the most concentrated, ignoring the deviant observations. In novelty detection the training data is not polluted by outliers and we are interested in detecting whether a new observation is an outlier. In this context an outlier is also called a novelty.

Outlier detection and novelty detection are both used for anomaly detection, where one is interested in detecting abnormal or unusual observations. Outlier detection is then also known as unsupervised anomaly detection and novelty detection as semi-supervised anomaly detection. In the context of outlier detection, the outliers/anomalies cannot form a dense cluster as available estimators assume that the outliers/anomalies are located in low density regions. On the contrary, in the context of novelty detection, novelties/anomalies can form a dense cluster as long as they are in a low density region of the training data, considered as normal in this context.

For this exploratory analysis, the parcel average values of the NDVI timeseries where used, along with the average phenology metrics of sowing/harvesting dates. Crucial work was made on the visual inspection and interpretation of the NDVI profiles (Figures 97, 98, 99, 100). It was a great surprise and disappointment to find out how noisy the data were. In the end, in many cases it was decided that it was a false statement of the crop itself and not just of the farming practice.

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

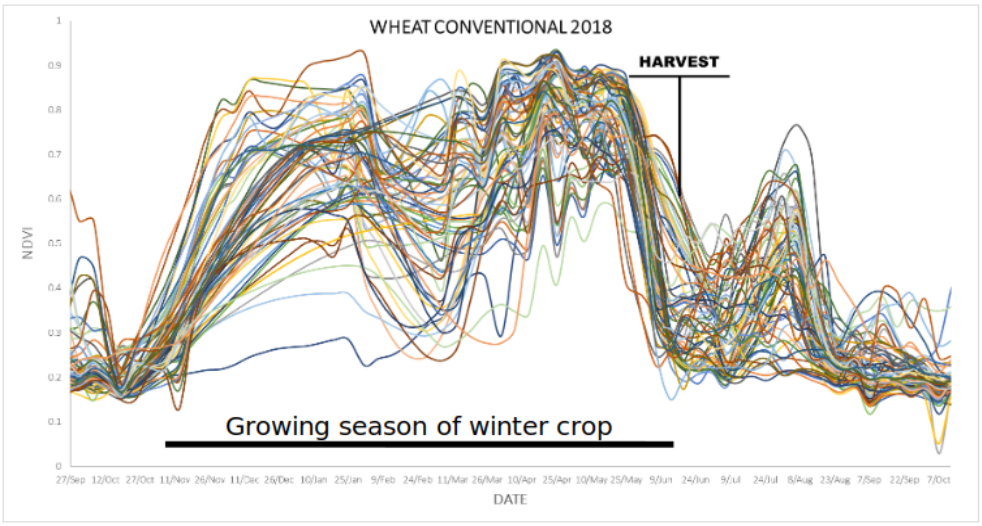


Figure 97: Expert opinion on outlier detection. Identification of true cases of Crop Declaration for conventional winter wheat

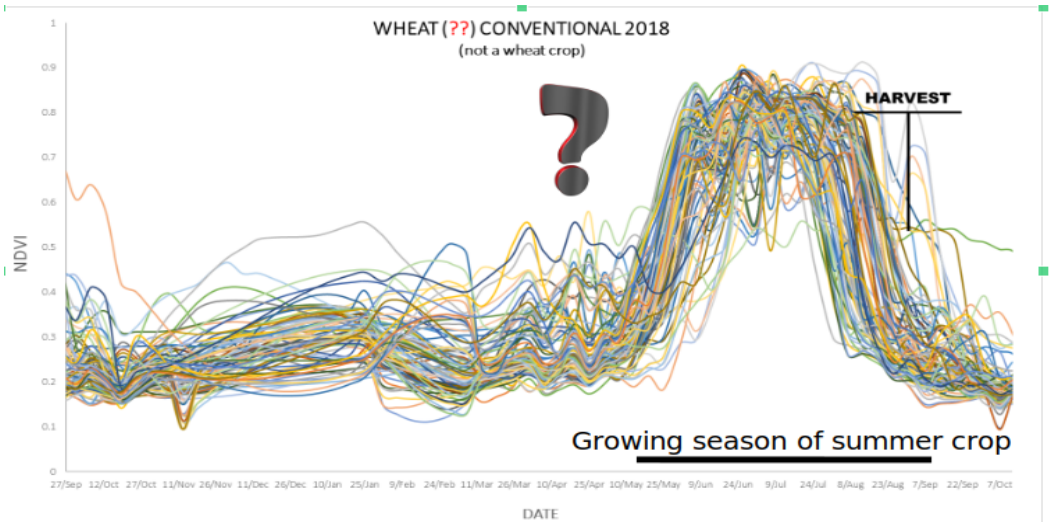


Figure 98: Expert opinion on outlier detection. Identification of false cases of Crop Declaration for conventional

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

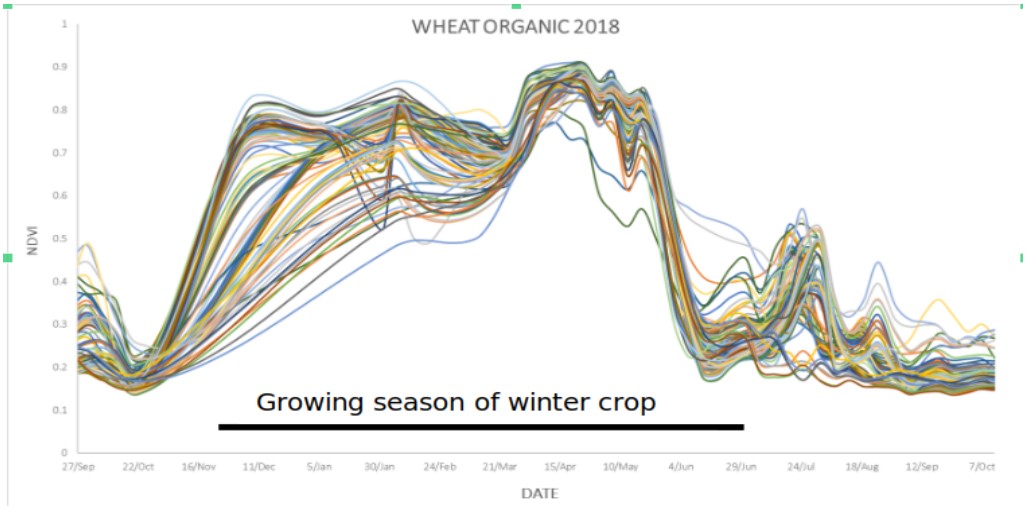


Figure 99: Expert opinion on outlier detection. Identification of true cases of Crop Declaration for organic winter wheat

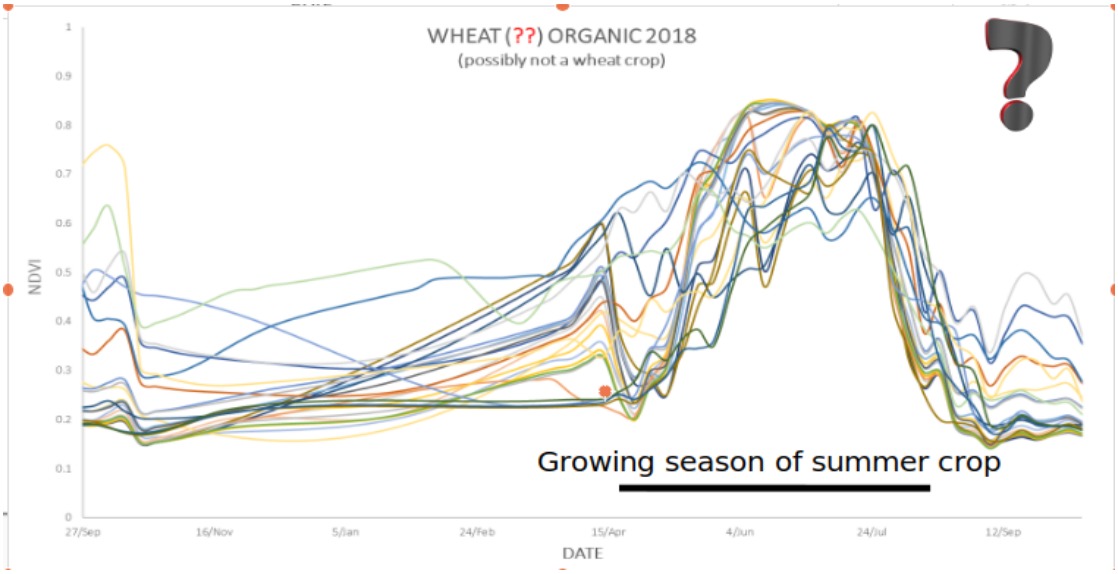


Figure 100: Expert opinion on Outllier detection. Identification of false cases of Crop Declaration for organic winter wheat

Knowing, the typical sowing/harvesting dates in the region of interest and the corresponding ones derived from the phenology metrics, the visual evaluator checked the NDVI profiles, by units first, and in clusters afterwards, spanning the whole 2016-2020 dataset. Too many clear cases were found where a spring crop was reported as a winter crop or the reverse. During the visual examination - assessment of the NDVI profiles, each declaration was assigned the classification : false (0), probably true (2), true (1).

1) **Outlier Detection**

As a first step outlier detection was performed on the dataset subsets belonging to class 1 or 2 of the visual inspection procedure. One efficient way of performing outlier detection in high-dimensional datasets is to use the Random Forest algorithm. The IsolationForest algorithm variant 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the

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

maximum and minimum values of the selected feature. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.

For the implementation of the algorithm, the contamination parameter and the number of estimators had to be set. These parameters controlled the amount of contamination of the data set, i.e. the proportion of outliers in the data set. Outliers detection predicted the outliers among the data that passed the visual inspection.

2) **Novelty Detection**

Having a dataset presumed to be cleaned from outliers, at a high confidence level, was suitable for a semi-supervised novelty detection method in order to check the data categorized as false, or possible false declarations from the visual inspection (classes 0 and 2).

For this task the One-Class SVM algorithm was used. It required the choice of a kernel and a scalar parameter to define a frontier. The Radial Basis Function (RBF) kernel was chosen although there exists no exact formula or algorithm to set its bandwidth parameter. The nu parameter, also known as the margin of the One-Class SVM, corresponds to the probability of finding a new, but regular, observation outside the frontier, and was set to 0.01. The gamma kernel coefficient was set to 1/(number of features).

Table 15: Distribution of ground truth data – Parcels after Outlier/Novelty Detection

		Initial Dataset		After Anomaly Detection	
		Organic	Conventional	Organic	Conventional
Wheat	2016	87	2		
	2017	65	4		
	2018	198	187	159	77
	2019	77	213		
	2020	217	219	191	103
	2021	132	28		
Maize	2016	79	4	26	4
	2017	15	10		
	2018	8	242		
	2019	35	355	7	129
	2020	8	365		
	2021	27	77		
Soybean	2016	89	6	84	6
	2017	71	7	65	
	2018	13	51		
	2019	17	78		
	2020	5	68		
	2021	18	6		
Sunflower	2016	288	4	210	3
	2017	89	4		
	2018	58	88		
	2019	96	130	53	104
	2020	101	173	93	85
	2021	11	13		

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

The result of the Outlier/Novelty detection actions was a drastic reduction in the number of data to be analysed (Table 15). A decision was made to select specific seasons for each crop to provide a minimum number of samples and a relative uniformity in the distribution of organic/conventional crops. For the crop/year specific models the following were therefore chosen to be trained:

- Maize 2016
- Soybean 2016
- Sunflower 2019
- Sunflower 2020
- Wheat 2018
- Wheat 2020

ML Classification Model Training

This procedure involved the training of crop specific binary classification models, for Wheat, Maize, Sunflower, Soybean crops, using the XGBoost algorithm. 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.

Learning Algorithm - Decision Tree Classifiers - XGBoost

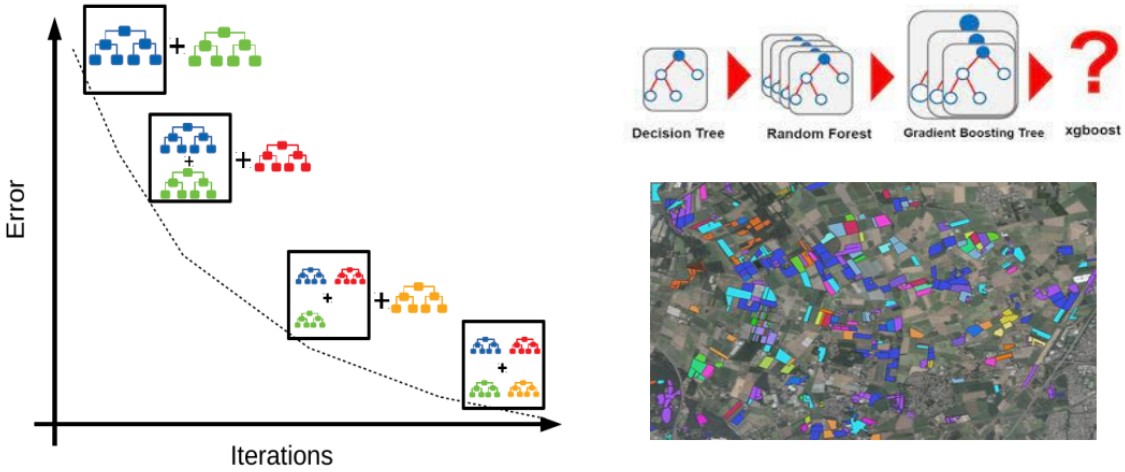


Figure 101:Boosting methods on ensemble classifiers as an evolution from simple CART and Bagging methods. Additive models on CART residuals for the minimization of classification loss.

The Gradient Boosting Trees (Figure 101) algorithm is one of the most powerful machine learning techniques for creating predictive models. This technique includes 3 elements:

- A loss function, which tends to be optimized.
- A decision tree classifier (decision trees) to make predictions.
- An additive model (gradient descent) to sum up the successive classifiers that were generated, in such a way that the error function is minimised.

Gradient boosting works by creating simpler, weak, prediction models successively, where each model tries to predict the error left over from the previous one. Because of this, the algorithm tends to overfit

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

quickly to the training data and should therefore be used in conjunction with an an unbiased method of validation

A weak learner is a model that does a little better than random predictions. The principle on which boosting works is error correction of a previous learner through the next learner. It focuses on the sequential summation of these of weak learners and filtering the observations whose classifier finds the correct at each step. Because of the sequential summation, the algorithm is usually slow in learning but still has high accuracy.

The ML preprocessing involved the PCA, mentioned earlier in the text, and the integration of the transformation to the modelling. The main training of the algorithm, had to deal with the optimization of XGBoost hyperparameters and the handling of a highly imbalanced dataset through data augmentation. The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important. One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE.

Internal & External Validation for Model Training

Initially the dataset was split into Training and Independent Testing subsets, via stratified sampling, following a 80:20 ratio. The training subset would be used for all the data augmentation and hyperparameter tuning, through a nested resampling method (internal cross validation) to avoid data leakage and overall bias, while the testing dataset would be used for the overall assessment of the trained model.

In order to obtain honest performance estimates, all parts of the model building like data augmentation, hyperparameter tuning 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 102).

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 with 4 folds in the outer loop and 10 folds in the inner loop was decided for the cross-validation resampling.

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

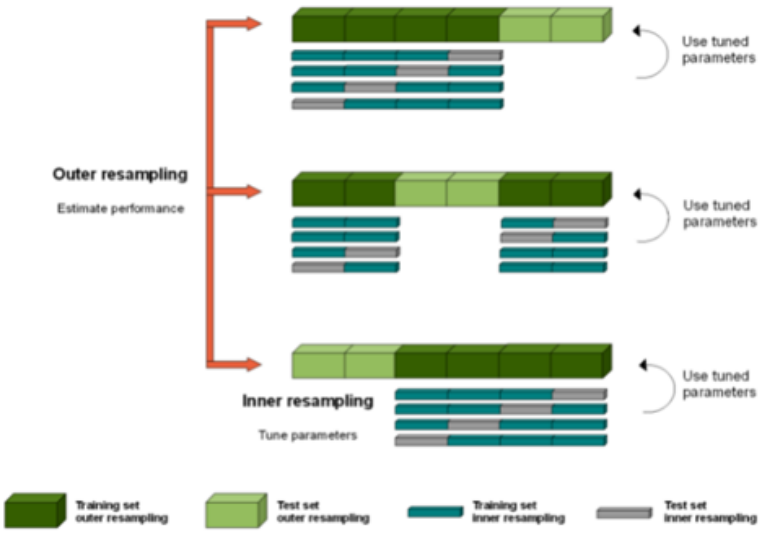


Figure 102: Nested Cross Validation implementation

Data Augmentation

Through the application of data augmentation, the SMOTE technique was used in each fold of the inner loop. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbours for that example are found (typically k=5). A randomly selected neighbour is chosen, and a synthetic example is created at a randomly selected point between the two examples in feature space. The chosen approach was SMOTE With Selective Synthetic Sample Generation, experimenting with the Borderline SMOTE and ADASYN variants.

The Borderline SMOTE algorithm, a popular extension to SMOTE involves selecting those instances of the minority class that are misclassified, such as with a k-nearest neighbour classification model. Those difficult instances, were oversampled, providing more resolution only where it may be required. Instead of generating new synthetic examples for the minority class blindly, the Borderline-SMOTE method to only created synthetic examples along the decision boundary between the two classes.

Another approach that was examined, involved the generation of synthetic samples inversely proportional to the density of the examples in the minority class. That is, generating more synthetic examples in regions of the feature space where the density of minority examples is low, and fewer or none where the density is high. This modification to SMOTE is referred to as the Adaptive Synthetic Sampling Method, or ADASYN.

Hyperparameter Tuning

Regarding the classifier algorithm, a XG-Boost was trained for each crop data subset. Binary logistic was used as the objective parameter, and AUC as the evaluation parameter for each fold. The following hyperparameters were tuned via Grid Search optimization, with an “early stopping” option, preventing model from overfitting:

- n_estimators: [2,3]

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

- learning_rate: [0.001, 0.01, 0.05, 0.1]
- colsample_bytree: [0.1, 0.2]

After finding the best parameters of the model, the training was repeated using all data from "training" partition. For Data Augmentation, the Borderline SMOTE algorithm was applied.

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 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 are evaluation metrics drawn from the independent test set validation (external data) 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.

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

- 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).

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

- 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.

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

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

- 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 leaded 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 (Figure 103).

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 is 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 have been assimilated for the calculation of the indices that fed the ML classifier, the results are crop specific, notifications are produced based on the decisions of the object-based (parcel) analysis, and visualization (graphs, reports, widgets) is populated based on the results.

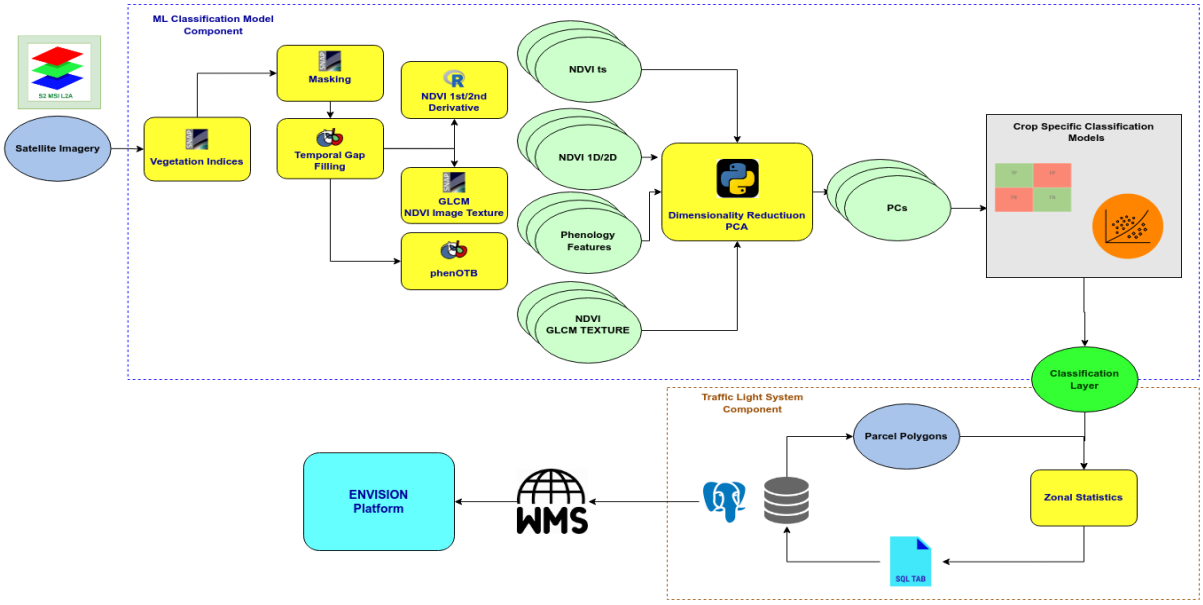


Figure 103: ML Model Prediction as a part of the ENVISION service

The tables below (Tables 16-17-18-19) present 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 is

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

employed i. for the object partitioning of the images, ii. the supervised classifier's training, and iii. to provide the classification decisions.

Table 16: 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 17: 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 18: 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 19: Data from CBs

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

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

Regarding the output product (Table 20), 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 20: 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

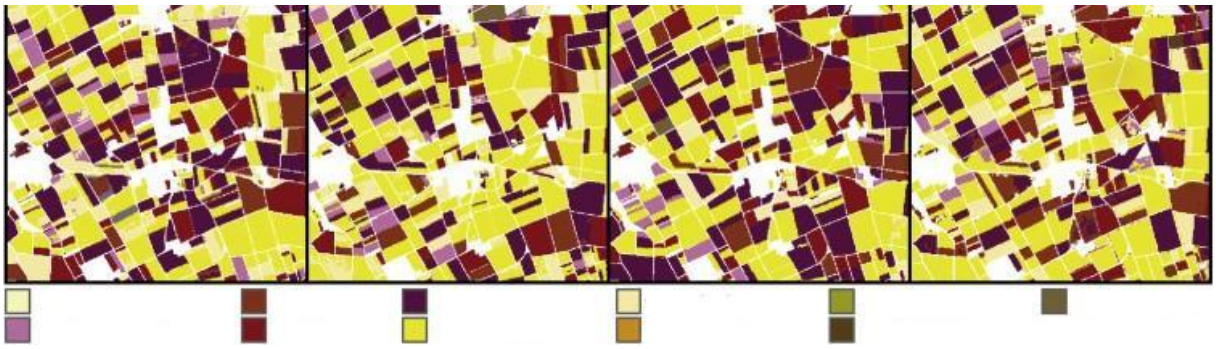


Figure 104: Visualisation of output data

Table 21: 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.

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

4.6 Lighthouse Customer Case: Grassland Mowing Events Detection – Flanders

The grassland mowing event detection algorithm has been developed and implemented for a pilot sub-area in Flanders for the cultivation period of 2022. However, unlike the case in Lithuania where samples with mowing event dates were available, we did not have labeled data for training the AI component for mowing event detection. As a result, a data fusion step for NDVI reconstruction and a subsequent threshold-based event detection approach were employed (for more details of the methodology see BC.1.). Although the threshold-based approach provides a practical solution in the absence of training labels, the evaluation results may highlight the importance of further refining the algorithm's precision that can be acquired via ML, especially in specific regions or conditions where mowing events may exhibit different characteristics.

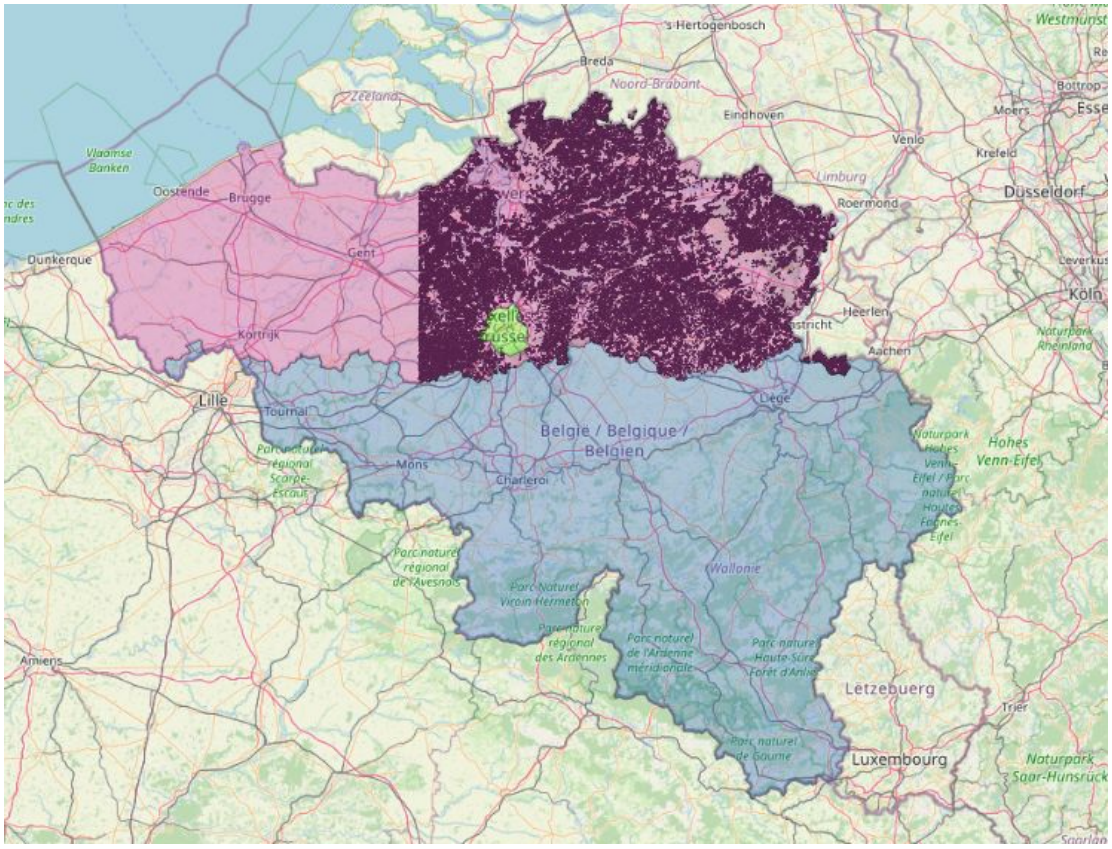


Figure 105: Output of Grassland Mowing Detection in Flanders (pink area)

While the algorithm has been deployed, the evaluation of its performance is currently underway. Through thorough analysis and assessment, the effectiveness and accuracy of the grassland mowing detection routine will be determined. This evaluation will provide valuable insights into the reliability and potential of the algorithm for broader implementation. By meeting the necessary standards and requirements, the algorithm will enable informed decision-making and improved monitoring of grassland management practices in the region of Flanders.

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

5 Risks Mitigated

In this chapter, a concise overview is presented regarding the risks that have been deliberated throughout the project's duration.

Sentinel-1B anomaly

In December 2021, an anomaly occurred with Copernicus Sentinel-1B, specifically affecting its instrument electronics power supply. This anomaly resulted in the satellite's inability to provide radar data, leading to the termination of its mission. Consequently, the SAR temporal resolution experienced a degradation from 6 days to 12 days, necessitating adjustments in the ENVISION S2 data reconstruction based on S1 data. Fortunately, Copernicus Sentinel-1A remains operational, and preparations are underway for the launch of Sentinel-1C. Although objectively quantifying the exact impact of the anomaly is challenging since ENVISION services began actively providing since 2022, the very high performance of the Grassland Mowing Events detection products reported by NPA for 2022 cultivation period indicates that the effects of the anomaly may be less severe than initially anticipated.

Coherence Data

After careful consideration and analysis, we have made the decision to focus solely on utilizing backscatter Sentinel-1 data in ENVISION's algorithms, rather than also incorporating coherence coefficients as initially planned. This decision is based on several factors. Firstly, the calculation of coherence data is a complex and time-consuming process, particularly when working with national scale datasets that can be extremely large, reaching sizes of several terabytes for a cultivation year. The acquisition of heavy Single-Look-Complex (SLC) products is slow and the processing necessary for extracting interferometric coherence coefficients add significantly to the overall time and resource requirements. Additionally, findings from an ablation study conducted by Garioud et al. in 2021 [19] revealed that the importance and contribution of sigma0 backscatter data and coherence features is relatively similar on NDVI reconstruction (especially for the case of grasslands), with a slightly superior importance placed on backscatter data. These insights indicate that by utilizing backscatter data alone, we can achieve valuable results while minimizing the computational cost associated with coherence data calculation and optimize the available storage capacity within Creodias. Based on this assessment, we have made the informed decision to prioritize the usage of backscatter data, ensuring an efficient and effective approach for our project deliverable and services provided.

Discriminating factor with other projects

The development of the ENVISION project has been significantly influenced by its predecessors, the RECAP and SEN4CAP projects, both of which have served as essential benchmarks and sources of knowledge transfer. To be specific:

RECAP laid the foundation for ENVISION by providing initial algorithms for crop classification and smart sampling for OTSC, which were further refined to achieve higher accuracy. Drawing on the experience gained in RECAP, ENVISION expanded its scope to a national level, delivering more precise results. Additionally, ENVISION integrated data cube technology, setting it apart from RECAP, to enable efficient storage and retrieval of agricultural data for monitoring purposes.

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

Similarly, SEN4CAP played a vital role in shaping ENVISION's development as a benchmark for performance and methodology. ENVISION's algorithms underwent refinement and adaptation based on the valuable insights gained from SEN4CAP's results. To address SEN4CAP's main weakness (a one-size-fits-all approach designed for all regions) ENVISION adopted a distinct approach, crafting tailored solutions that considered each area's unique characteristics. For instance, Cyprus, characterized by small parcels near the limits of Sentinel-2's potential due to 10m resolution, and Lithuania, suffering from extended cloud coverage resulting in limited number of available clear-sky images. According to CAPO, SEN4CAP's initial implementation in Cyprus for crop classification yielded sub-optimal results, prompting ENVISION to fine-tune its algorithms, eventually achieving accuracy rates exceeding 80%. Moreover, one of the ENVISION's distinctive attributes, consists the implementation of the traffic light system to quantify the risk of applicants' false declarations and guide inspection campaigns more effectively. Moreover, the sophisticated deep learning mowing event detection algorithm, based on S1/S2 data fusion components to create continuous time series and overcome extended cloud coverage challenges, has produced exceptional results and is currently integrated into NPA's checking infrastructures.

The consortium has already performed an analysis to differentiate itself from other similar projects. NPA conducted a comparison exercise (Table 22) to identify the unique features of ENVISION in relation to DIONE. Also, NPA will proceed in assessing the performance of the similar services and compare accuracies that can be achieved.

Table 22: DIONE – ENVISION Services Provided Comparison by NPA

DIONE	ENVISION
Farmer	
Agronomist - consultant	
inspector at Paying Agency	
official at the Paying Agency	Paying Agency
Researcher -scientist	
Official at the Ministry	
Organic inspector	Certification Bodies
Compliance monitoring tool	Envision platform
Input data: S2, GSAA parcels via feature service	**Input dada: S1 and S2, GSAA parcels via shape file**
Crop type markers (19 crops)	Cultivated Crop Type Maps (22 crops)
Bare-soil marker	
Similarity marker	
Homogeneity marker	
Mowing marker (3 crop codes)	Grasslands Mowing Events Detection (9 crop codes)
Mean NDVI marker	
Non-productive EFAs detection	
	Stubble Burning Identification
	Nitrate Vulnerable Zones (NVZ)
	Harvest event detection

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

Greening compliance check (old rules)	Greening-1 rule (old rules)
Task submision to ask geotagged photos	
Geotagged photos task result review	
Environmental Performance Tool	
Indicators:	
Greenhouse Gas	
Organic farming	Organic farming indentification
Land cover change	
Natura 2000	
HNV Farmlands	
Land irrigation	
Water Quality	
Soil erosion	Minimum Soil Cover (Black fallow)
Soil organic Matter (VNIR sprectrometry)	Soil organic carbon monitoring
DIONE geotag app	Envsion geotag app
Navigation to the point of interest	
AR based photo capture	
DIONE EO	ENVISION EO
Sentinel Data Hub extraction	ENVISION DataCube
Super-resolution	
Drones and VHR data	
Data fusion	S1 and S2 fusion

Training of data for distinction of organic farming practices appears very difficult

The task of distinguishing organic from conventional farming practices with the use of EO data was very challenging. Initially, regarding the strategy, decisions had to be made about what EO derived classification features to use for the discrimination. Data dimensionality and more practical reasons regarding the spatial data extent and the available data space posed a certain limit as to how many features to use. The discussion was about whether to focus more on the spectral or the spatiotemporal content of the EO data. Clear scientific evidence about “a defined spectral signature” of an organic farming practice wasn’t found in the literature, rather than some few experimental cases that focused solely on crop/leaf canopy nutrient content. These studies used very high resolution multispectral and hyperspectral data, questioned the problem on specific crop varieties and on a highly local scale experimental plots, relying on abundant ground truth data about nutrient NPK inputs. On ENVISION,it was known from the start, that such in situ data were not available at a national scale. It was finally resolved , to focus more on the spatiotemporal aspects of vegetation phenology in the EO signal.

In the successive benchmarks and iterations carried out to train the ML algorithms, the central issue was the "Bias-Variance Trade-off" that was noted. During the 1st iteration of ML model training, error values greater than the acceptable threshold were observed. In other words, the model used was not strong enough to produce an accurate prediction. Therefore, this was a case of high bias, which was addressed by increasing the complexity of the models. The number of features was significantly increased by including NDVI Derivatives and Image Texture metrics . A dimensionality reduction was performed and a Boosting Trees classification algorithm was chosen instead of the original SVM, which is suitable for cases of high bias.

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

The generalization of the models, regarding prediction on other years/seasons, also an issue pointed out by the reviewers, was showcased on the pilot validation (D3.6). The inference of models trained on data of other years yielded even worse results. Unfortunately, ground truth data on pilot years were scarce, unequally distributed, and with many outliers regarding the crop type. Thus, they were not enough to train year specific models over the pilot seasons.

At the 2nd iteration (D3.5, D3.7, D1.9_2nd progress report), the behavior of the models was indicative of a case of high variance and a high tendency for over-fitting. The regularization parameters were chosen in a range of values that made the modeling less "aggressive" in memorizing the dataset. Unfortunately, the lack of samples with uniformity of distribution by crop variety and also by geographical distribution became apparent. However, the performance of the models on unseen datasets was not the same across all crops. In the case of Sunflower for 2019 and 2020, the results showed quite good performance and qualities in relation to their over/under estimation (type I & II errors). Their efficiency was quite stable, regardless of the year, both in early and full season prediction (D1.9_2nd progress reported - Chapter "Evaluation of ML Models").

This is an important finding because it shows that when the algorithms are trained in a year specific manner, with enough data, diverse and balanced in terms of the underlying crop variety, and localized in a limited geographical space, they can give promising perspectives and good results, contributing with their predictions to the specific business case, be it mid or full season planning. This seems to indicate that the resulting product could be viable under certain assumptions regarding the input data it receives, and the accepted error thresholds set by its design.

<u>The risk of trying to test a very difficult topic in a single area - identifying organic parcels contra non-organic - was not taken into account.</u>

The approach to constructing the training dataset encompassed sampling across numerous Areas of Interest (AOIs) throughout Serbia. It was observed that utilizing a dataset localized to a region with consistent climate, soil conditions, and crop varieties positively influenced the model's attributes. However, individual AOIs lacked adequate training data, necessitating the merging of datasets.

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

6 ENVISION's competitive advantage to Paying Agencies

ENVISION offers a range of distinct advantages that empower PAs to effectively monitor and manage agricultural activities. Implementation of advanced technologies and innovative approaches through ENVISION, revolutionize the way to access, analyze, and utilize national-scale data. With ENVISION, PAs gain the following key benefits:

- Scalability: Seamlessly obtain precise results at any scale, whether you need insights for small Areas of Interest or comprehensive coverage of entire countries.
- Advanced Algorithms: Harness the power of sophisticated Machine Learning and Deep Learning-based routines, ensuring highly accurate and reliable results that drive informed decision-making.
- Customizable Analysis: Tailor the analysis to meet your specific needs and requirements, accessing a suite of customizable tools that provide actionable insights aligned with your goals.
- Continuous Direction: Receive continuous guidance throughout the cultivation period, enabling real-time monitoring of mowing events and empowering you to make timely and informed decisions.
- Generalization Performance: Gain access to reliable information that transcends regional boundaries, empowering you to make informed decisions about agriculture management across diverse areas.
- Cloud Coverage Resilient: Overcome the challenges posed by cloud coverage and adverse weather conditions through the integration of data from both Sentinel-1 and Sentinel-2 satellites, ensuring high accuracy and comprehensive coverage.
- Cost-Effective: Reduce the reliance on costly manual field visits and optimize resource allocation by leveraging the efficiency and automation offered by ENVISION.
- Enhanced Monitoring: Benefit from continuous monitoring of vegetation and soil over time, enabling early detection of potential issues, supporting effective decision-making, and providing robust validation capabilities.

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

7 Conclusion/Final Remarks

The successful implementation of the data download and processing workflow using CreoDIAS has had a profound impact on the ENVISION project, as moving on into the new era of national-scale data provision and exhaustive monitoring. By exploiting the capabilities of the ENVISION DataCube, which serves as the cornerstone of the project, we have been able to develop sophisticated algorithmic routines and ML/AI pipelines that effectively meet the specific requirements of business case users. This achievement not only advances the objectives of the ENVISION project but also sets a precedent for future initiatives in the field. The development of these algorithmic routines represents a significant milestone, as they empower us with powerful tools for data analysis and accurate predictions. The positive feedback received from our users further validates the effectiveness of these high-quality data products, with a majority expressing high levels of satisfaction and acknowledging the successful fulfillment of their requirements. The insights we have gathered can serve as a valuable reference for other projects aiming to enhance monitoring frameworks and support sustainable agricultural practices. ENVISION's advanced monitoring capabilities contribute to the ongoing evolution of CAP, empowering stakeholders with the necessary tools to make informed decisions, improve agricultural practices, and drive sustainability.

Overall, this deliverable provides a technical overview of the ENVISION data products and their transformation into services that meet user requirements. It discusses the data collection and pre-processing routines involved in generating the data products. Additionally, the methods used to implement the data products are described. It is important to note that a final report for the validation of the data products (D3.6) is anticipated to be completed by month 38, providing a comprehensive assessment of their accuracy and reliability.

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

References

[1] Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-1221, Wilm, U., Cadau, E., Gascon, F., 2016. Sentinel-2 sen2cor: L2a processor1222 for users, in: Ouwehand, L. (Ed.), ESA Living Planet Symposium 2016, Spacebooks Online. pp. 1–8.

[2] 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.

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

[4] 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.

[5] 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. https://doi.org/10.1109/JSTARS.2020.3038152.

[6] 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

[7] Bontemps, S., Bajec, K., Cara, C., Defourny, P., de Vendictis, L., Heymans, D., Kucera, L., Malcorps, P., Milcinski, G., Nicola, L., Slacikova, J., Taymans, M., Tutunaru, F., Udroiu, C., 2022. Sen4cap – sentinels for common agricultural policy. http://esa-sen4cap.org/content/download-package-description

[8] Mathilde De Vroey, Laura de Vendictis, Massimo Zavagli, Sophie Bontemps, Diane Heymans, Julien Radoux, Benjamin Koetz, Pierre Defourny, Mowing detection using Sentinel-1 and Sentinel-2 time series for large scale grassland monitoring, Remote Sensing of Environment, Volume 280, 2022, 113145, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2022.113145.

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

[10] Jernej Puc, Sentinelhub - Area Monitoring — Homogeneity Marker, Oct 13 2020: https://medium.com/sentinel-hub/area-monitoring-homogeneity-marker-742047b834dc

[11] Safanelli, J.L.; Chabrillat, S.; Ben-Dor, E.; Demattê, J.A.M. Multispectral Models from Bare Soil Composites for Mapping Topsoil Properties over Europe. Remote Sens. 2020, 12, 1369. https://doi.org/10.3390/rs12091369

[12] Grizonnet, M., Michel, J., Poughon, V. et al., (2017). Orfeo ToolBox: open source processing of remote sensing images. Open geospatial data, softw. Stand. 2, 15

[13] Inglanda, J., (2016). PhenOTB, Phenological analysis for image time series. DOI:10.5281/zenodo.45573

[14] Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software, 4(44), 1903.

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

[15]Pedregosa et al., (2011).Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830,
[16]Gallagher, N., (2020). Savitzky-Golay Smoothing and Differentiation Filter.
[17]Haralick, R.M., Shanmugam, K., Denstien, I.,(1973). Textural features for image classification. IEEE Trans Syst Man Cybern, vol. 3, no. 6, pp.610–621.
[18]Bellman Richard (1961). Adaptive control processes: A guided tour (A RAND Corporation Research Study). Zbl 0103.12901 Princeton, N. J.: Princeton University Press, XVI, 255 p.
[19]Anatol Garioud, Silvia Valero, Sébastien Giordano, Clément Mallet, Recurrent-based regression of Sentinel time series for continuous vegetation monitoring, Remote Sensing of Environment, Volume 263, 2021, 112419, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2021.112419.

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

envision

End of Document

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