D3.5 REPORT ON COLLECTED AUXILIARY DATA

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	D3.5 Report on collected auxiliary data			
Work Package	WP3 – Earth Observation data products			
Date of Delivery	**Contractual**	M34	**Actual**	M34
Nature	Report	**Dissemination Level**		Public
Lead Beneficiary	DRAXIS			
Responsible Author	Panagiotis Papadimitriou (DRAXIS)			
Contributions from	Panos Ilias (EV ILVO), Bert Callens (EV ILVO), Iason Tsardanidis (NOA), Thanassis Drivas (NOA), Alexia Tsouni (NOA), Stella Girtsou (NOA), Evangelos Oikonomopoulos (AgroApps)			

Document History

Version	Issue Date	Stage	Description	Contributor
D0.1	22/6/2023	Draft	Draft for review	NOA
F1.0	30/6/2023	Final	Final to be submitted	DRAXIS
F2.0	3/11/2023	Final	Final based on the comments provided	DRAXIS

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 FIGURES

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

Introduction

This document aims to present the collected auxiliary data with regards to services development and it could be considered as complementary documentation to the previous deliverables (D3.2 Catalogue on auxiliary data and available repositories to be incorporated, D3.3 Data products initial report).

The sections of this deliverables are:

- Section 1: Introduction which presents a brief description of the deliverable content.
- Section 2: Sample design which presents a brief description of the collected sample for each business case.
- Section 3: Data Collection Methods which presents the description of the data sources (i.e. form where were the data collected, how do the data from each data source shape the design of the provided service, etc.).
- Section 4: Data Processing Methods which presents the data preparation overview (i.e. merging data sets, selecting a sample subset of data, etc).
- Section 5: Lessons learned which presents the lessons learned from the data collection experience and provides recommendations for improving future similar works.

For your reference, Table 1 below presents a matrix that aligns the developed data products with the corresponding outputs within the service.

Table 1: Summarized outputs per data product

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	

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

			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	

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

1 Sample design

1.1 Monitoring multiple environmental and climate requirements of CAP (NPA – Lithuania and CAPO – Cyprus)

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. Timestamps are usually provided by the respective business cases and represent respective time labels of the events recorded. However, acquiring these timestamps can be a challenge due to their scarcity and difficulty to obtain. In order to obtain the necessary number of training instances, as well as to enhance the accuracy of our assessments, experts from NOA conducted a further annotation label acquisition through photo-interpretation using Satellite imagery. This process involves a meticulous analysis of cloud-free Sentinel-2 satellite images to offer additional validation and expand the dataset used for training and assessment. Notably, this method leverages datacube services hosted in Creodias, a platform developed by NOA. This advanced system significantly streamlines the photo-interpretation process, enabling efficient and rapid analysis of available Sentinel-2 satellite data. By capitalizing on the fields' geometrical characteristics, this scientifically rigorous approach ensures precise and reliable results in our evaluation.
    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.

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

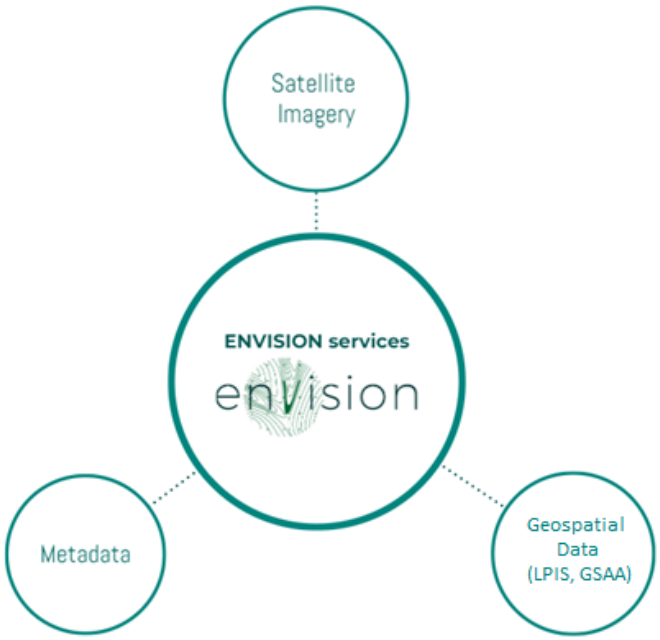


Figure 1: ENVISION services data requirements

Satellite Imagery data are collected using CreoDIAS, in order to extract real-time series data for the monitoring of crops. This data is processed and transformed into Analysis Ready Data (ARD), ensuring its suitability for agricultural monitoring purposes. To construct the datasets, different approaches are employed based on the requirements of the analysis. However, in certain situations, such as with small parcels, pixel values are extracted and subsequently the results provided were aggregated (usually by using the *Majority Voting* criterion) to make a final decision at the level of the parcel. This computation is based on the Geospatial Aid application (GSAA) provided by the users, which contains information about the declared codes and geometries of agricultural parcels. However, in situations where pixel-level precision is necessary, the value of the parcel is annotated at each unique pixel within the parcel's geometry. In addition to satellite imagery and GSAA data, a lookup table is provided by the Business Cases (BCs). This lookup table (see an example in Table 2) serves as a reference to properly match the crop codes provided with their corresponding names and agricultural categories. This ensures consistency and standardization in the analysis and interpretation of the crop data across different regions.

Table 2: Summarized outputs per data product

ID	CROP CODE	CROP NAME	CROP FAMILY	EAA	AL	PGrass	TGrass	Fallow	Cwater	Protein	Cother
1	C1	SOFT WHEAT	CEREAL	1	1	0	0	0	0	0	0
2	C2	BARLEY	CEREAL	1	1	0	0	0	0	0	0
3	C3	COMMON OAT	CEREAL	1	1	0	0	0	0	0	0
4	C4	MAIZE	CEREAL	1	1	0	0	0	0	0	0
5	C5	SORGHUM	CEREAL	1	1	0	0	0	0	0	0
6	C6	FORAGE PEAS	BROADLEAF CROPS	1	1	0	0	0	0	1	0

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

7	C7	CUCUMBERS	VEGETABLES	1	1	0	0	0	0	0	0
8	C8	RYEGRASS	VICIA	1	1	0	0	0	0	1	0
9	C9	ALFALFA	VICIA	1	1	0	0	0	0	1	0
10	C10	CHICKPEA	BROADLEAF CROPS	1	1	0	0	0	0	1	0
11	C11	LENTILS	BROADLEAF CROPS	1	1	0	0	0	0	1	0
12	C12	TOMATOES	VEGETABLES	1	1	0	0	0	0	0	0
13	C13	VINEYARDS	VINES	1	0	0	0	0	0	0	0
14	C14	OLIVE TREES	TREES	1	0	0	0	0	0	0	0

In-situ data, which is extensively described in former documentations, are crucial for validating and creating models. By incorporating actual field code validations and event timestamps, the models can be trained and evaluated more effectively, aligning the predictions with the real-world agricultural practices.

Finally, to form the training, validation, and test datasets, random sampling techniques were employed. The goal was to create representative datasets that capture the diversity of crop types and their proportions accurately. The sampling process was stratified based on the actual declared code distribution, meaning that the samples were drawn in a way that maintains the proportional representation of each crop type. This stratified sampling approach helps ensure that the datasets reflect the true distribution of crops in the target area, allowing for robust model training and evaluation.

Below, we provide a summary of the essential data required for each business case to effectively utilize NOA's services:

<u>Lithuania BC</u>

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 (A)
    b. Rainfall erosivity factor layer (R)
    c. Soil erodibility factor layer (K)
    d. Slope length factor and slope steepness factor layer (LS)
    e. Crop and cover management factor layer (C)
    f. Conservation supporting practices factor layer (P)
    g. Slope - DEM layer
3. Paying agencies:
    a. LPIS and 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

Cyprus BC

2. Satellite Data:
    b. Sentinel-2 L2A (tiles: 36SWD, 36SVD)
        ii. Spectral bands (B01-B12)
        iii. Scene Classification (SCL)
    c. Sentinel-1 GRD (rel. orbits: 94, 167)
        ii. Backscattering coefficients (VV-VH)
3. Auxiliary Data:
    b. Annual soil loss layer (A)
    c. Rainfall erosivity factor layer (R)
    d. Soil erodibility factor layer (K)
    e. Slope length factor and slope steepness factor layer (LS)
    f. Crop and cover management factor layer (C)
    g. Conservation supporting practices factor layer (P)
    h. Slope - DEM
4. Paying agencies:
    b. LPIS and GSAA: Parcels geometries and farmers declarations as a shapefile (updated when is necessary)
    c. A lookup table for all the available crop type names, codes, families and CD ancillary info
    d. Events Timestamps to fine-tune the algorithms (Stubble Burning, Harvest of arable land)
    e. Agricultural Practices Descriptions – National CAP strategic Plans
    f. Hydrographic Network
    g. Natura2000 regions

1.2 Monitoring the condition of soil (LV – Belgium)

The availability of a consistent number of ancillary data (or covariates) allows applying sampling strategies according to the feature space and not only based on the geographical space. Provided that the covariates are strongly related to the target variable, i.e., to the soil organic carbon (SOC), the sampling strategy based on feature space can ensure to collection soil samples representative of the whole range of SOC values within the investigated area. For these purposes, remote sensing data cheaply provides covariates over large areas. The physical link between spectral data in the optical domain and SOC exists and is widely exploited in remote sensing contexts. Thus, the absorbance/reflectance values at a given wavelength can be considered as covariates related to the target variable and consequently, the spectral variability can be exploited for sampling strategies based on feature space. However, some absorption features are quite broad and they can partly overlap with spectral regions related to a different soil property. For this reason, it is desirable using the whole spectrum as covariates instead of a single band or a narrow region. SOC prediction models exploit most

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

of the spectral regions across the electromagnetic spectrum between 400 and 2500 nm and this is due to the large heterogeneity of the components of the organic matter.
In the ENVISION project, we used a soil sampling strategy based on a stratified feature-based approach, in which the feature space consists of the satellite spectral data retrieved by a multi-temporal analysis of the Sentinel-2 (S2) data and the strata are the soil associations of the Soil map of the Flemish region that assure an adequate geographical distribution and a proportioned representation of the all-soil types insisting in the Flemish region.

For Belgium, a collection of images from the Sentinel 2 is used, on the Level-2A data product., The period is between May 2018 and May 2021, and the median is taken, to determine the soil spectrum space for S2. The revisit time of the Sentinel 2 satellites is 5 days. This collection is filtered based on the 2A-Level data product. The region is reduced to Flanders where we have the ground truth dataset. Almost 30% of the images are removed because the cloud percentage was too high (+90%). The pixels in the remaining images are filtered by using the quality layer of the Sentinel 2A product in order to mask the pixels with a high probability of being cloudy or snowy (See Data product report 3.7). Different indexes like NDVI, NBR2, VNSIR are used to mask pixels when they are expected to have been vegetation and not bare soil. This was based on literature and a trial and error evaluation of the synthetic layer (median of the collection in RGB). The times when the pixels match the bare soil are significantly reduced because most of the time fields have crops or catch crops growing on them. The temporal resolution can very a lot in that regard, from maybe 1 valid per pixel per year, in extreme cases maybe for some years none, to more, depending on the combination of Sentinel 2 passing times, the presence of clouds, the crop rotation, the growth of weeds. This huge variation in available information per pixel is one of the challenges to overcome, together with for example moisture levels, crop residues.

The synthetic layers take a median value of al valid values per pixel in the timeframe that was used. We can evaluate that the effectivity of this selection from bare soil by the fact that in most cases when we exclude fields were maybe only a few pixels are predicted, the parcels declared as grassland in the LPIS (Land Parcel identification system) are not predicted in most cases. Some variation can exist because part of the grasslands in the 2021 LPIS are not permanent grassland, so they can give bare soil results in other years.

First, we made a bare soil composite image using a pixel-based multi-temporal analysis using S2 images acquired from 2018 to 2021. The bare soil pixels were selected according to the computation of indices and a cloud mask that can detect green and dry vegetation and high soil moisture content that can affect the soil spectrum shape. The output of this analysis is a synthetic bare soil layer ‘(SBSL) for croplands in each soil association region. Croplands were detected using the parcels of the Land Parcel Identification System (LPIS) provided by LV.

For each soil association region, we extracted the S2 spectra from each pixel of the SBSL that will form the feature space to assess the geographical position of the soil samples by the Kennard-Stone algorithm. This algorithm allowed to select n samples uniformly distributed over the feature space from all the spectra of the SBSL within each soil association region, thus optimizing the coverage of the

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

spectral variability. First, the algorithm finds the two spectra that are furthest apart based on Euclidean distance assigning them to the calibration dataset and removing them from the input matrix. Then, the procedure is repeated until the number of samples within the calibration dataset is equal to n. Sample suggested points (outputs)

We collected samples from 171 locations, 21 more than the original estimation, overcoming budget restrictions. To increase farmers' awareness and respect their rights to decide if they want to participate in the Soil Campaign, EV ILVO worked together with the LV, and LV sent a formal request to the Farmers. We identify which availability is from the selected points. This way, the geographical distribution of the collected samples ensures the right proportion among soil associations, which in turn allows to proper cover the SOC variability in the Flanders region.

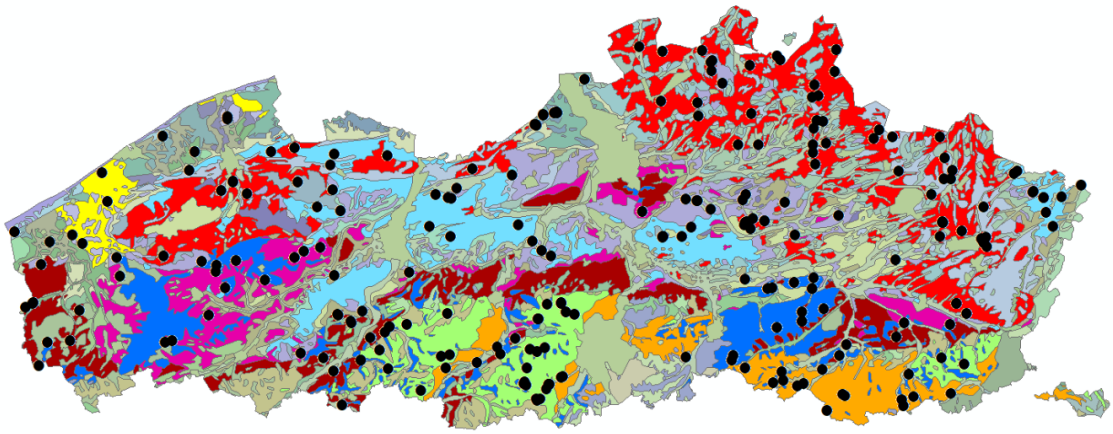


Figure 2: Kennard-Stone algorithm selected points (total 245) to deliver the needed flexibility on final selection.

1.3 Monitoring organic farming requirements (OCS – Serbia)

Dataset Creation – Sampling

The dataset is comprised of in-situ and Earth Observation data. Regarding the in-situ data, they are described in detail in Chapter 3.3. In-situ data sampling in EO data is required for the creation of models. At this point, there are two approaches referring to the data sampling. The first approach is taking into account the average value per polygon while the second one is the random point sampling. Considering the fact that the spatial heterogeneity resulting from the GLCM Texture Features should be taken into account as well, the approach to conduct the sampling in pixel basis was chosen (second approach).

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.

The user interface accepts csv files of GSA data. The information provided in the csv files is combined with the data available on the GeoSerbia web API before they are stored in the PostGIS database. The

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

import component is backed by an HTTP/RESTful API that is secured through the OpenID Connect layer on top of the OAuth 2.0 protocol, provided by the Authorisation Server.
LPIS postgres tables/shapefiles/geojson files and GSA tables should be provided by the user with a harmonized content, which means that they should both share a primary key field of the same data type configuration. This would assure that the SQL join of the spatial and attribute tables would be successful.

The table schema of the GSA should be proposed to the user to include the following fields:

- Parcel ID (unique id - Primary Key)
- Applicant ID
- Year of Declaration
- Farming Practices: The declared type of farming practices (organic or conventional)
- Declared Crop Code
- Crop Variety
- Crop Type Category

There are no minimum requirements about the quantity and quality of the input data, but definitely these issues have a big effect on the algorithm performance. Therefore, there are data size recommendations that relate with the spatial extent of the dataset (how localized is the dataset in regards with soil & climatic conditions and their effect on crop growth) and its sub-stratification within crop varieties, with a “lose” rule of thumb of at least “50 times the number of involved crop varieties” parcel samples. The data quality is influenced by the balance between the number of samples within organic/conventional classes, and the total samples for each variety, aiming for a uniform data distribution.

In general, the concept behind the data service of the product 5, is that each end user could provide regional-or national parcel data and farming practice declarations, employ EO features, Train/Tune-Validate/Evaluate a ML model and finally infere on a parcel level assessing a value from the traffic-light system that relates with its farming practice conformance. During the Envision project this conceptualization was further developed. What spatial manipulation of LPIS should be made? What attributes of the GSA are mandatory? What EO features could yield best results? Are they enough for a very low error performance? What family of algorithms is more effective? Does Data Augmentation aid in tuning of the model? Is Data Cleaning mandatory? What classification threshold to use? All these issues led to the design and implementation of processing components and tools that were incorporated to the service in order to improve its functions. However, the initial concept remains the same. Any user could train and predict on different regions, preferably geographically localized.

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

2 Data Collection Methods

2.1 Monitoring multiple environmental and climate requirements of CAP (NPA – Lithuania and CAPO – Cyprus)

The introduction of the Sentinel missions has brought about a significant impact on agriculture monitoring, providing high-resolution data with both spatial and temporal variability at no cost. These freely accessible data have enabled numerous applications that were not possible before. However, unlocking the full potential of these data still presents a challenge. The volume of big Earth data, the different modalities of the sources and sensors, and the complexity of accessing and pre-processing the data can be daunting, especially for non-experts in Earth observation. Thus, a series of acquisition, indexing, and pre-processing steps is required to transform the data into Analysis Ready Data (ARD), which can be used by the broader AI4EO community, including non-ICT experts. To be classified as ARD, data must undergo a minimum processing level that includes atmospheric correction, geometric calibration, re-projection, and resampling. Furthermore, ARD should be organized to facilitate easy and direct analysis. Additional pre-processing steps such as cloud masking can further simplify and support data analysis. To this end, NOA is developing automated workflows for generating ARD, which will serve as input for the subsequent data analysis and AI pipelines.

Sentinel Data

NOA's back-end processes are hosted on CREODIAS platform that offers direct access to EO data via a scalable object storage directory (/eodata). This storage system is designed to provide high-performance access to data by using buckets. Users have access to the full archive of Sentinel-1 GRD, SLC, and Sentinel-2 Level-1C (L1C) data for Europe. However, Sentinel Level-2A (L2A) products are not fully offered (Table 3), so NOA uses Sen2Cor software to transform L1C to L2A for specific years (e.g. case of Cyprus). This approach provides developers with straightforward access to retrieve and process data without the need to download it locally or copy it to dedicated VMs. Additionally, NOA has developed Python scripts that allow searching and pre-processing of products in the eodata directory based on specific parameters.

Table 3: 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 ***

For the case of Lithuania is expected that around of 2.6 TB (S2: 2.1, S1 0.5) of data per year are download and stored in CreoDIAS infrastructures. While for the case of Cyprus the amount is significantly reduced into 0.9 TB (S2: 0.6, S1: 0.1).

Geospatial Data

To complete the functionalities of ENVISION, NOA collects shapefile data from the pilot cases, which includes the declared codes and parcel geometries provided by farmers. In addition to this, NOA

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

collaborates with business case partners to obtain additional data layers that will be overlaid on the ENVISION maps. These data layers can be manipulated and analyzed using GIS tools, enabling users to perform various spatial operations and analysis. The specific data layers included in the ENVISION system are determined based on the needs of end-users and the availability of data. Overall, the following data layers are currently incorporated:

- Farmers Declarations (GSAA): This data includes information provided by farmers, such as the declared codes and parcel geometries, which are essential for monitoring and analysis.
- Land use/Land cover Maps and LPIS: Detailed information about the distribution of different land use and land cover categories, enabling analysis of land use patterns, changes over time and geometries.
- Natura 2000 regions: Geospatial data representing designated Natura 2000 areas, which are protected sites of high ecological value in Europe.
- Nitrate Vulnerable Zones: Spatial delineation of areas designated as nitrate vulnerable zones, which are subject to specific regulations to prevent nitrate pollution of water bodies from agricultural activities.
- Hydrographic networks: Detailed information about water bodies, such as rivers, lakes, and streams, including watershed delineation, to support water-related analysis and monitoring.
- Digital Elevation Model (DEM): A high-resolution representation of the terrain, providing elevation information that aids in topographic analysis and modelling.
- RUSLE components: it's crucial to clarify that the components on which RUSLE depends on (P, R, K, LS, and C) are spatially distributed variables. These factors are not represented by single, uniform values for an entire country but are calculated at a pixel level (10m spatial resolution by resampling process). The spatial variability accounts for differences in land use, terrain, and other localized factors, making them more accurate and context-specific. All the involved parameters are downloaded for the whole Europe from the ESDAC, cropped to each country borders and resampled to the Sentinel 2 spatial resolution (10 m), except from the LS and C factor, which were calculated using LPIS and NDVI again at 10m resolution.

2.2 Monitoring the condition of soil (LV – Belgium)

From the 245 sampling points that were selected, 5 points were not made available by the farmers (who refused to participate), while 13 points could not be sampled due to the presence of temporary grassland (I.e., parcels with low or null possibility to observe the Soil at bare conditions) and farm buildings. From the remaining 227 points, 171 were sampled within the project, ensuring a good distribution across the different soil association classes identified for Flanders.
In the field, the location of the sampling point was identified by means of a Stonex S10 RTK GPS (centimetre accurate). Further, an area with a radius of 5m around the sampling point was marked by red sticks after which 16 subsamples from the topsoil (0-10cm) were collected randomly within the sampling area by means of an auger with a diameter of 2.5cm. The subsamples were thoroughly mixed and stored in a labelled plastic bag for transportation to the laboratory. A picture was taken at each sampling site and some general field characteristics were monitored (e.g., land cover, soil conditions, tillage).

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

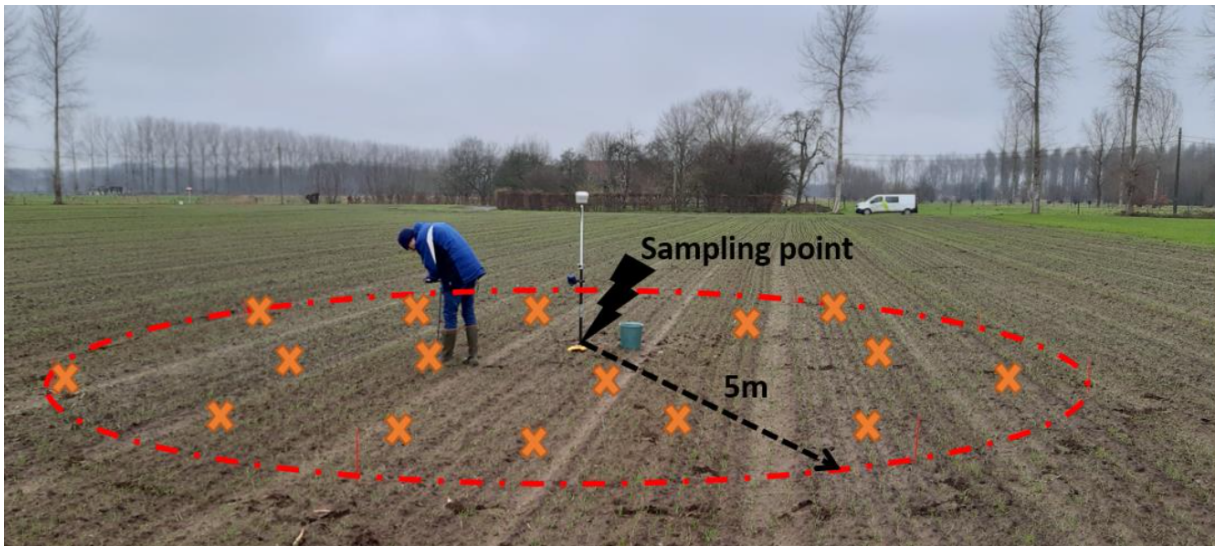


Figure 3: Sampling points and the subsamples area.

Staalnameformulier ENVISION 2021

Algemene info

PLOT ID: ……	**Plaats (hoofdgemeente)** ……
Datum:……	
Staalnemer(s): ……	
Beschrijving bereikbaarheid en ligging proefvlak:	

Terreinkenmerken *(Omcirkel of vul aan)*

Landgebruik zoals verwacht?	ja/neen, specifeer……
Landbedekking	Braak - Stoppel - Gewas, specifeer……
Bodemtoestand	Normaal - Waterverzadigd - Bevroren - Overstroomd - Uitgedroogd
Macroreliëf:	Vlak - Depressie (droog/nat) - Helling - Top - Plateau
Bodemverstoring	Geen
	Berijding door landbouwmachines
	Erosie
	Andere, specifeer: ……
Landbouwactiviteit	Recente bewerking: nee/ja, specifeer……
	Recent bemest: nee/ja, specifeer……
	Andere, specifeer: ……

Figure 4: Document template used for the soil campaign.

2.3 Monitoring organic farming requirements (OCS – Serbia)

The data that were collected consisted of both organic and conventional farming practice ground truth parcels, emerging from the Business Case of Serbia (Doo Organic Control System Subotica – OCS).

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

Practice Type was the response variable (**Y** class vector) whereas crop type variable was used to stratify crop specific models. Summarizing the provided ground truth data:

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

- 2335 conventional parcels
- 1866 organic parcels

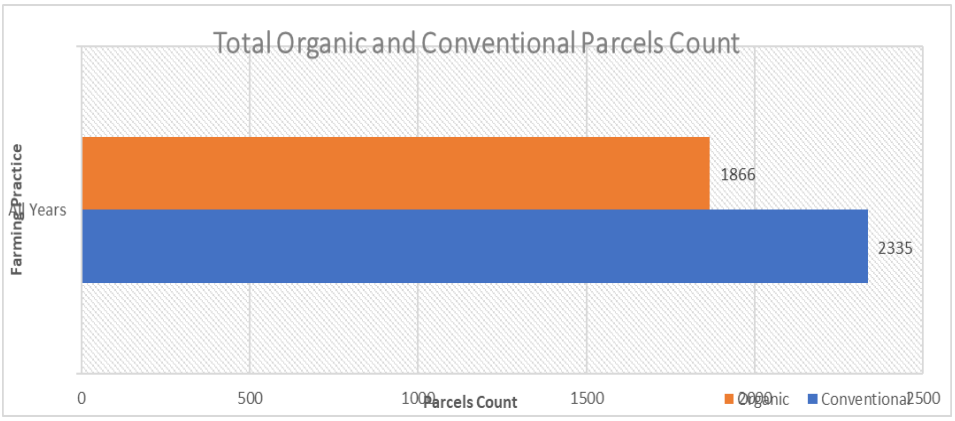


Figure 5: Total Organic and Conventional Parcels Count.

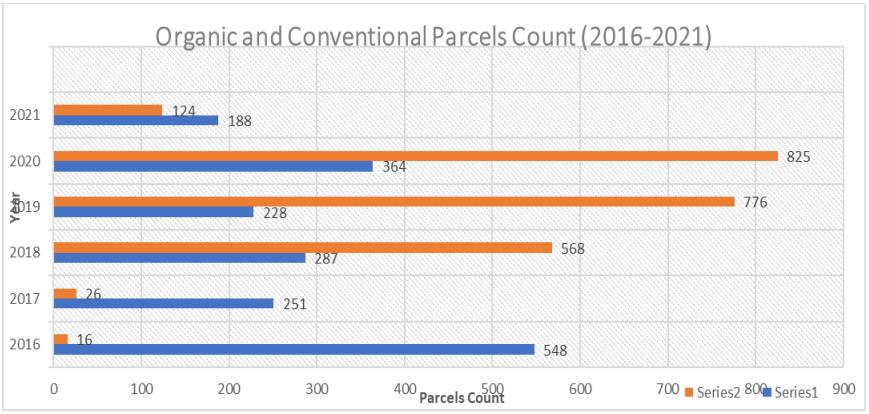


Figure 6: Organic/ Conventional parcel count distribution through the ground truth years.

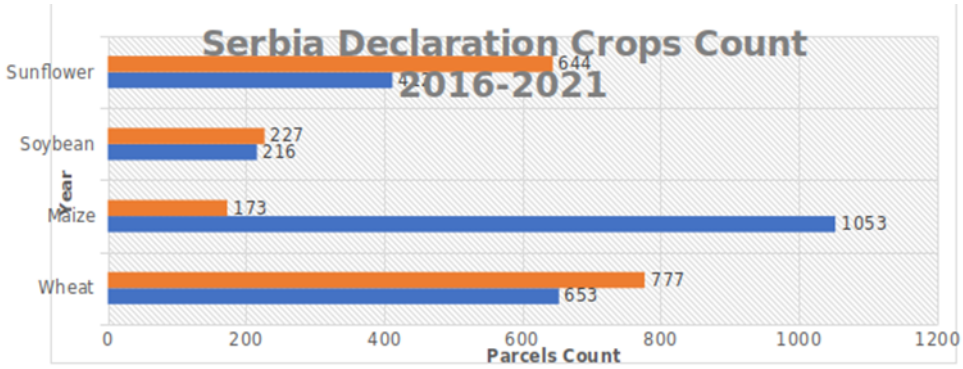


Figure 7: Organic/ Conventional parcel count distribution among crop types.

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

In order to 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.

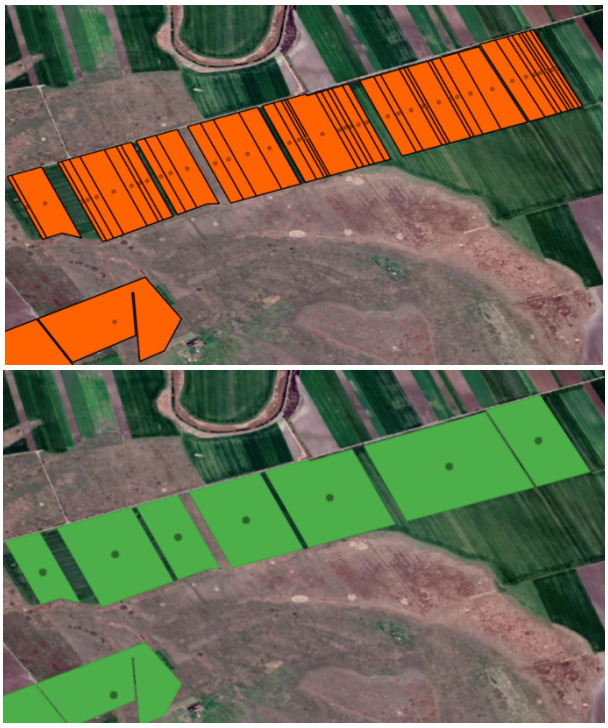


Figure 8: Parcel geometry dissolve.

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

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

Table 4: Number of parcels per category in comparison to the target.

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

Parcel Dispersion & Relevance

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

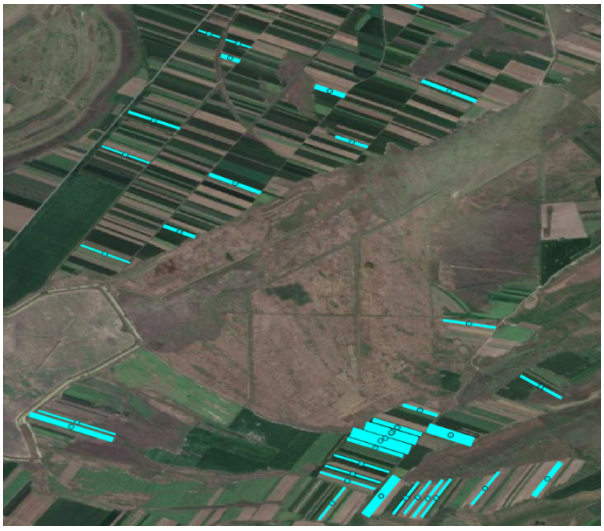


Figure 9: Parcel geometry dissolve.

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

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

Figure 10: Parcel geometry dissolve.

Regarding the Earth Observation data, Sentinel-2 Level-2A data are going to be 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 data will be processed by the exploitation of the Copernicus Data Information Access Services (DIAS) infrastructure, and specifically the CREODIAS platform.

Table 5: 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

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

3 Data Processing Methods

3.1 Monitoring multiple environmental and climate requirements of CAP (NPA – Lithuania and CAPO – Cyprus)

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).

Sentinel-2 Data and Cloud Masking

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

Figure 11: Cloud Masking and Cloud Masking buffering in order to tackle cases of undetected clouds/ shadows on the boundaries of the produced masks.

Sentinel-1 Data

Level-1 Ground-Range-Detected (GRD) products obtained from Sentinel-1 satellite imagery in Interferometric Wide (IW) swath mode to derive valuable information. These GRD products underwent a series of pre-processing steps using the snappy library to ensure data accuracy and usability. This included (i) area clipping to focus on our specific region of interest, (ii) radiometric calibration, (iii) application of the Refined-Lee speckle filter for noise reduction, and (iv) terrain correction using the

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

Shuttle Radar Topography Mission (SRTM) 10-m dataset. Additionally, we transformed the backscatter coefficients into decibels (dB) for better interpretation.

Buffering Parcels

LPIS includes information about parcels geometry. This geometry (polygon or multipolygon) is often mixed with pixels of another object, creating mixels (pixels that belong to more than one field). As the analysis takes into consideration either each pixel or an aggregation of them, it is important to avoid these outliers. Thus, the boundaries of the geometries are reduced through a buffer zone using GDAL. The buffer corresponds to a polygon containing the region within the buffer distance of the original geometry. Afterwards, the polygons are rasterized in order to avoid conflict cases of mixels and acquire samples that are more representative. Below there is an example visualisation (Figure 12) where with green we depict the area of the field after the buffering routine. This methodology is similar to the one implemented by SEN4CAP (following JRC solution) and inward buffer of -5 m has been chosen as default for all data products.

Figure 12: Parcels geometry after buffering.

Open DataCube (ODC)

ENVISION aims to monitor agriculture at a national scale by providing long time-series indices and analytics to support decision-making processes. To achieve this, harnessing a large amount of Earth observation (EO) data requires powerful resources and dedicated tools. The Open Data Cube (ODC) has been chosen due to its maturity, widespread adoption, and cost-effectiveness as it is available for free under the Apache 2.0 license. ODC provides data structures and tools that enable efficient organization and analysis of Earth observation (EO) data, with support from more than 30 countries. NOA has already installed, tested, and used ODC locally in the DataCAP[1] application (Figure 13), which includes automated modules for data download, pre-processing, and indexing in the DataCube. It also incorporates street-level images from the mapillary API for validation purposes. ODC's main advantage lies in its ability to catalogue massive EO datasets, enabling easy access and manipulation through a Python API.

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

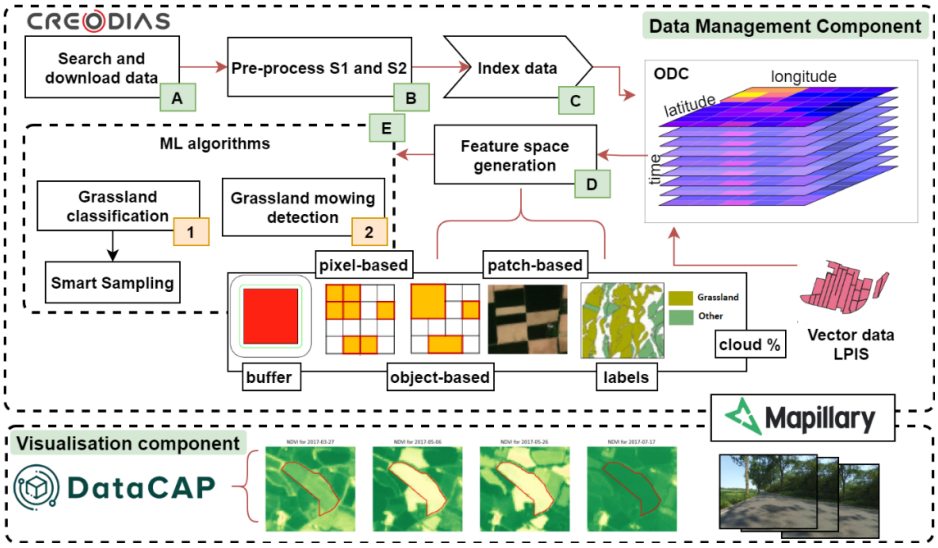


Figure 13: DataCAP architecture.

For the cataloguing of data, ODC offers two methods; Indexing (catalogue only metadata) and Ingesting (catalogue the entire data). Currently, indexing seems to be the most efficient method according to the developers of ODC and it is based on the DataCube-core module, which utilizes a PostgreSQL database to write the metadata of the products. The latter are hosted either in a local file system or in the cloud. The implementation of the **ENVISION DataCube** includes the installation of the required environment along with the configuration of several files and the initialization of the database.
ENVISION utilizes both the eodata catalogue and the ODC framework to deliver innovative and scalable solutions for agriculture monitoring at a national scale. While the initial focus is on Cyprus and Lithuania pilot cases, the methodology can be easily applied to other regions as well (e.g. case of Flanders). The platform provides pre-processed data from both Sentinel-1 and Sentinel-2 satellites, enabling comprehensive agriculture monitoring. It also allows for further research and operational outputs such as data fusion, crop classification, and analytics. Currently, the indexed data covers the entire extent of Lithuania and Cyprus for at least two cultivation periods, including the current (2023) and previous ones. All the observed products for these two countries are **processed directly from this catalogue** and the generated ARD are automatically indexed to the database, hosted in the dedicated CreoDIAS VM. Prior the data indexing, users are required to create products associated with each indexed dataset. A product is defined as a collection of datasets that share the same sets of measurements.

In ENVISION, we have created **country-specific products** (e.g., S2PreprocessedLithuania) using YAML files. Once these products are indexed, the DataCube contains time series data with a consistent spatial resolution of 10 meters. To request and utilize this data, we utilize a Python API. This API allows us to load metadata or data based on various parameters such as the product, time range, bounding box, and specific bands. The loaded data is stored in **Xarrays**, which have three dimensions: 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 14). 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).

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

Figure 14: 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 15 illustrates the stack of two layers: an RGB raster and the IDs raster.

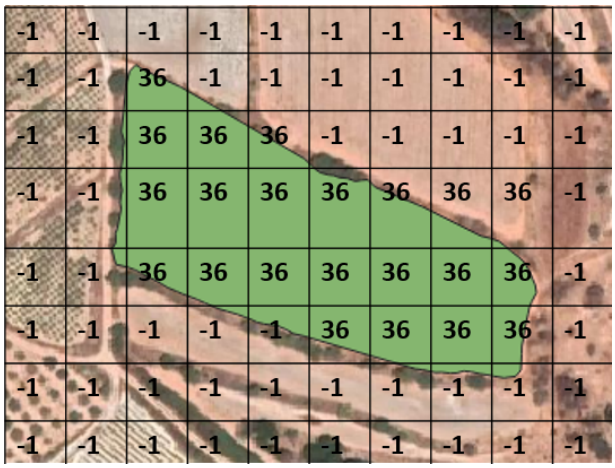


Figure 15: Parcel's index rasterisation inside DataCube.

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 are used to update the database, establishing a continuous **back-and-forth communication between ODC and database** as it is shown in Figure 16. 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.

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

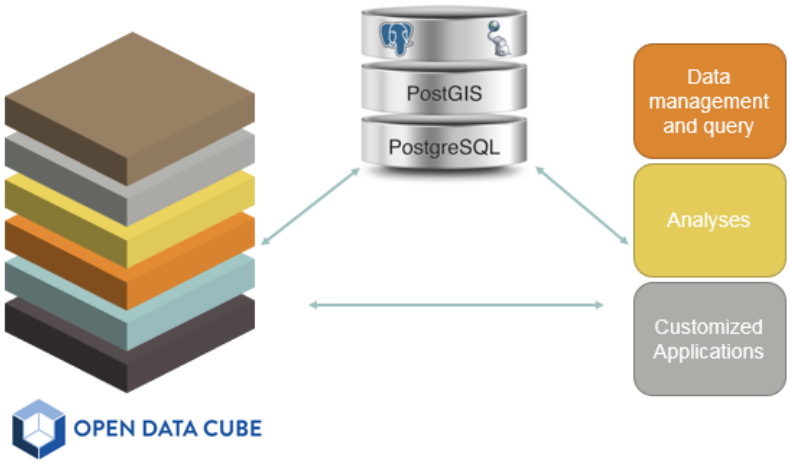


Figure 16: ENVISION products connection scheme.

Index Calculation

After importing the Sentinel bands into the DataCube, the next step in the processing pipeline involves the calculation of various vegetation and burning indices. These indices are essential for extracting valuable information from the satellite imagery and enabling more advanced analysis and monitoring. By referring to the below table 6 and table 7, the necessary formulas and band combinations 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.

Table 6 : Sentinel-2 derived indices

Index	Calculation	Sentinel-2 Bands
NDVI (Normalized Difference Vegetation Index)	(B8 - B4) / (B8 + B4)	Band 8 (Near Infrared), Band 4 (Red)
NDWI (Normalized Difference Water Index)	(B3 - B8) / (B3 + B8)	Band 3 (Green), Band 8 (Near Infrared)

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

NDMI (Normalized Difference Moisture Index)	(B8 - B11) / (B8 + B11)	Band 8 (Near Infrared), Band 11 (Shortwave Infrared)
PSRI (Plant Senescence Reflectance Index)	(B4 – B2) / B6	Band 4 (Red), Band 2 (Blue), Band 6 (Shortwave Infrared)
SAVI (Soil-Adjusted Vegetation Index)	((B8 - B4) / (B8 + B4 + L)) x (1 + L)	Band 8 (Near Infrared), Band 4 (Red), L (soil adjustment factor)
EVI (Enhanced Vegetation Index)	2.5 x ((B8 - B4) / (B8 + 6 x B4 - 7.5 x B2 + 1))	Band 8 (Near Infrared), Band 4 (Red), Band 2 (Blue)
DVI (Difference Vegetation Index)	B8 - B4	Band 8 (Near Infrared), Band 4 (Red)
VIgreen (Vegetation Index Green)	(B3 – B4) / (B3 + B4)	Band 3 (Green), Band 4 (Red)
VARIGreen (Variant of VIgreen)	(B3 – B4) / (B3 + B4 – B2)	Band 3 (Green), Band 4 (Red), Band 2 (Blue)
GDVI (Green Difference Vegetation Index)	(B8 – B3)	Band 3 (Green), Band 8 (Near Infrared)
SIPI (Structure-Insensitive Pigment Index)	(B8 – B2) / (B8 + B4)	Band 8 (Near Infrared), Band 4 (Red), Band 2 (Blue)

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

BSI (Bare Soil Index)	((B11 + B4) – (B8 + B2))/ ((B11 + B4) + (B8 + B2))	Band 11 (Shortwave Infrared), Band 2 (Blue), Band 8 (Near Infrared), Band 4 (Red)
NBR (Normalized Burn Ratio)	(B8 - B12) / (B8 + B12)	Band 8 (Near Infrared), Band 12 (Shortwave Infrared)
NBR2 (Normalized Burn Ratio 2)	(B11 - B12) / (B11 + B12)	Band 11 (Shortwave Infrared), Band 12 (Shortwave Infrared)
MIRBI (Mid-Infrared Burn Index)	(B12 - B11) / (B12 + B11)	Band 12 (Shortwave Infrared), Band 11 (Shortwave Infrared)

Table 7: Sentinel-1 derived indicesV

Index	Calculation	Sentinel-1 Bands
VV/VH Ratio	VV / VH	VV (Vertical transmit, Vertical receive), VH (Vertical transmit, Horizontal receive)
VH/VV Ratio	VH / VV	VH (Vertical transmit, Horizontal receive), VV (Vertical transmit, Vertical receive)
RVI (Ratio Vegetation Index)	VH / VV	VH (Vertical transmit, Horizontal receive), VV (Vertical transmit, Vertical receive)

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

Cross-Ratio	(VV x VV) / (VH x VH)	VV (Vertical transmit, Vertical receive), VH (Vertical transmit, Horizontal receive)

Further Processing Steps

Depending on the service provided, further processing is often required to ensure the quality and usability of the datasets. Hence, several techniques and methods are employed for data refinement and enhancement. These include:

- Cleaning from NaN Values: NaN (Not a Number) values may arise due to various reasons such as sensor errors or data corruption. These NaN values need to be identified and removed or replaced with appropriate values to prevent their negative impact on subsequent analysis.
- Linear Interpolation: In cases where data gaps or missing values are present within the time series, linear interpolation can be used to estimate the missing values based on the neighbouring data points. This helps to maintain a continuous and complete dataset for further analysis. The default temporal resolution chosen of the final time-series is usually that of 6 days.
- Time Series Smoothing: Time series data can exhibit fluctuations and noise that may obscure underlying patterns or trends. Smoothing techniques, such as moving averages or Savitzky-Golay filters, are applied to reduce noise and highlight the overall trends or patterns within the data.
- Resampling Methods: In some cases, it may be necessary to adjust the temporal resolution of the data. Resampling techniques, such as upsampling or downsampling, are employed to modify the time intervals between observations while preserving the overall characteristics of the data.
- Feature Extraction: Feature extraction techniques are used to identify and extract relevant information or characteristics from the data. This can involve the calculation of statistical measures, texture analysis, or the extraction of specific spectral or spatial features.
- Dimensionality Reduction: Dimensionality reduction methods are utilized to reduce the number of variables or features in the dataset while preserving important information. Techniques like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) are often employed to reduce the dimensionality of high-dimensional data.
- Data Augmentation: This helps address challenges such as limited data availability, data imbalance, and overfitting. By generating new synthetic data points, it provides a larger and more diverse training set, enabling machine learning models to generalize better and make more accurate predictions on unseen data.

Creation of Training and Validation Datasets

In the development of ENVISION's data products, it is essential to establish robust training and validation datasets to ensure the accuracy and reliability of our analytical algorithms. To achieve this, we have meticulously split the data for each of the three key data products:

Data Product 1: Analytics on Vegetation and Soil Index Time-series

The dataset used for the fine-tuning and validation of the respective component are stratified based on the distribution of available label events provided (OTSC or photo-interpretation). Stratified sampling ensures that the training and validation sets represent a proportional and unbiased selection of events across various regions and timestamps. This approach helps us capture the diversity and

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

dynamics of vegetation and soil index time-series modules, thereby enhancing the models' ability to analyse and interpret this critical data.

Data Product 2: Cultivated Crop Type Maps

In the context of developing a crop type map service based on crops, a meticulous and stratified approach is adopted for the training-validation split to ensure the model's proficiency in classifying a wide spectrum of crop types. The dataset is thoughtfully organised based on the various crop categories considered, ensuring that each category is well-represented in both the training and validation sets. This not only mitigates biases but also reinforces the model's adaptability to the intricacies of real-world agricultural landscapes. However, recognizing the challenge of high class imbalance within the dataset, particular attention is devoted to significant crop type categories that constitute more than 0.1% of the total cases. For those with fewer than 1000 samples, a Synthetic Minority Oversampling Technique (SMOTE) is implemented to rebalance the representation. Additionally, to address class imbalance effectively, the classification weights parameter, inversely proportional to crop frequencies, is seamlessly integrated into the algorithm, ensuring that the model accurately accounts for the differing prevalence of crop types in its classification process.

Data Product 3: Grassland Mowing Events Detection

To train and validate the S1/S2 Fusion component, we have collected data from distinct regions (see Figure 17) across the entire Lithuania. These regions were chosen accordingly to encapsulate a wide range of morphological and meteorological characteristics, as well as different cloud coverage profiles and scenarios. This diverse dataset helps our model adapt to various environmental conditions and ensures robust detection of mowing events in different contexts.

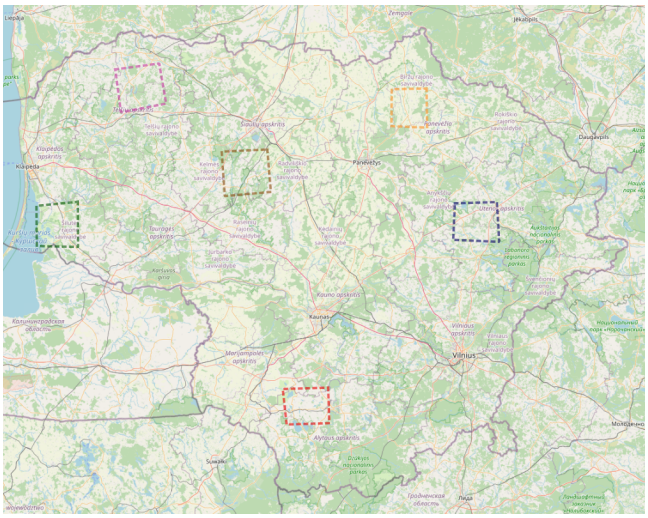


Figure 17: Map of Lithuania. The boxes correspond to the relevant regions for the training and the subsequent evaluation of S1/S2 fusion model for NDVI reconstruction.

In the case of the mowing detection component, similar to Data Product 1, we also take into account the date of the annotated mowing events during the dataset split. This consideration ensures fairness and an accurate representation of mowing events across time. By incorporating temporal information, our model becomes proficient in detecting mowing events as they occur, improving its real-time applicability.

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

3.2 Monitoring the condition of soil (LV – Belgium)

3.2.1 Lab measurements

Prior to analysis, the soil samples were oven-dried at 40°C for 7 days, ground in a mortar and passed through a 250 µm sieve. Soil organic carbon (SOC) and Total N were measured by dry combustion using a Skalar Primacs SNC-1002. 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% till 12.40%. 88% of the samples contained a SOC content between 0,5 and 2,0 %.

Lab results are processed based on provided Excel files. To load in the in datapoints in the notebook Colab environment and Google Earth engine, sometimes a csv of the file was used, sometimes a shape file was created based on the coordinates in the Excel, projected to the WGS 84 coordinate system.

3.2.2 Data from catalogues used

For the study area, we use a set of available data products like the World Cover map of Figure 18 or the Soil Association Map of Flanders or the soil texture map of Flanders (Figure 19). Those data products are available for direct use either to support the identification of crop areas or the assignment of soil parameters to each pixel area or for the development of the Soil Quality Indicators. Since our methodology is EO-based and we target for large-scale coverage at EU level, we prefer to use data products that are harmonized and available in digital formats or, even better in web services or STAC catalogues. The also aim for a min processing effort, which can be done again in an automated way using gdal python scripts or if that is not possible in a QGIS environment.

World Cover Soil map is extracted from Google Earth Engine. The format that is used there is the Cloud Optimised GeoTiff.

Figure 18: The study area covers 1368207 ha. Within the study are the agricultural parcels that cover 680.000 ha.

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

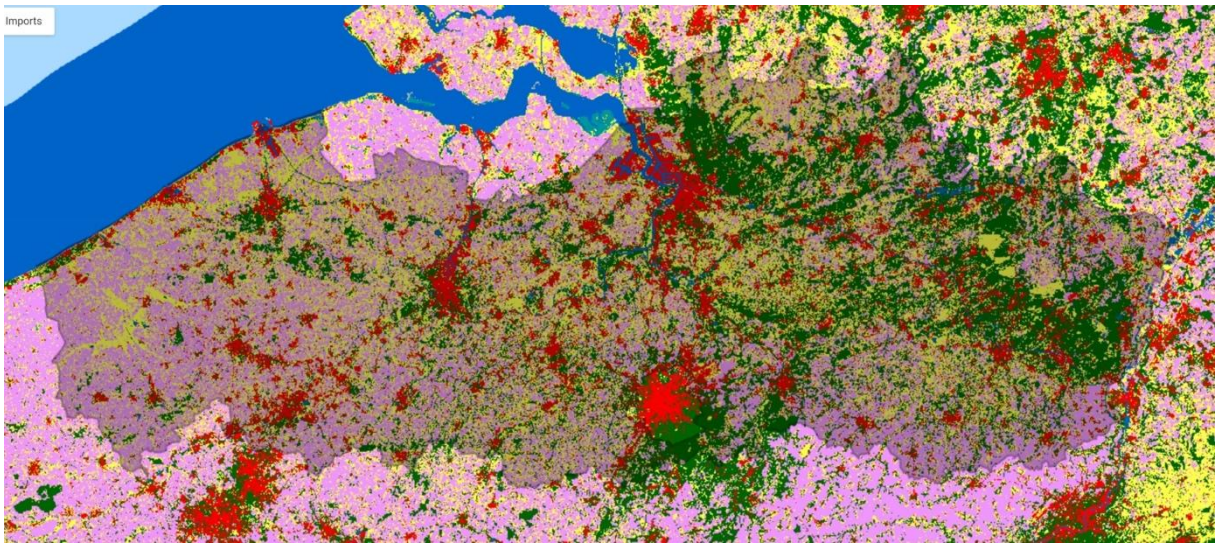


Figure 19: A land cover map of Flanders using the European Space Agency (ESA) WorldCover 10 m 2020 product provides a global land cover map for 2020 at 10 m resolution based on Sentinel-1 and Sentinel-2 data. The WorldCover product comes with 11 land cover classe and can support the identification of grassland (yellow), cropland (pink) and sparse vegetation (grey).

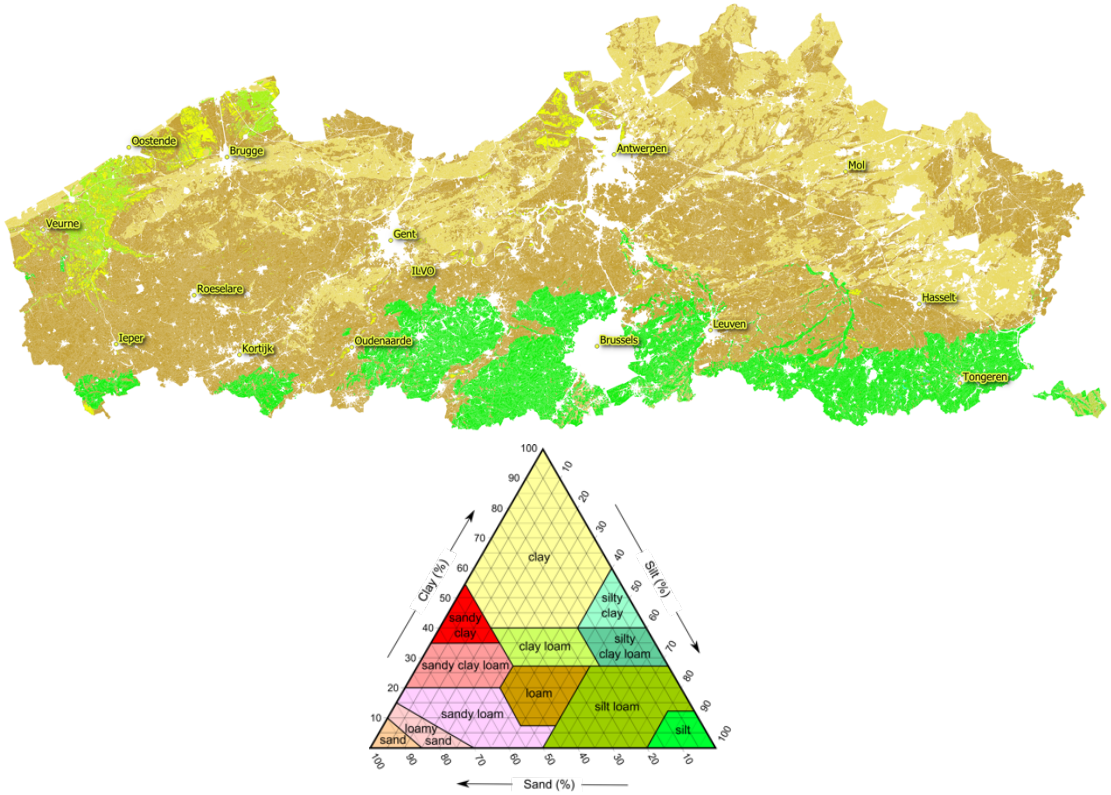


Figure 20: Soil texture map of the Flemish Region according to the international soil classification system World Reference Base on a scale of 1:40,000. Visualisation by EV ILVO using QGIS.

AMS input data definition for soil monitoring: Monitor large areas, over time, and see if the meet the conditions of Common Agricultural Policy (CAP). As we have provided in deliverable 3.7, we have

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

developed a Soil Quality Data Product indicator in order that allows the relative monitoring of agricultural areas considering pedoclimates conditions using the distribution methods.

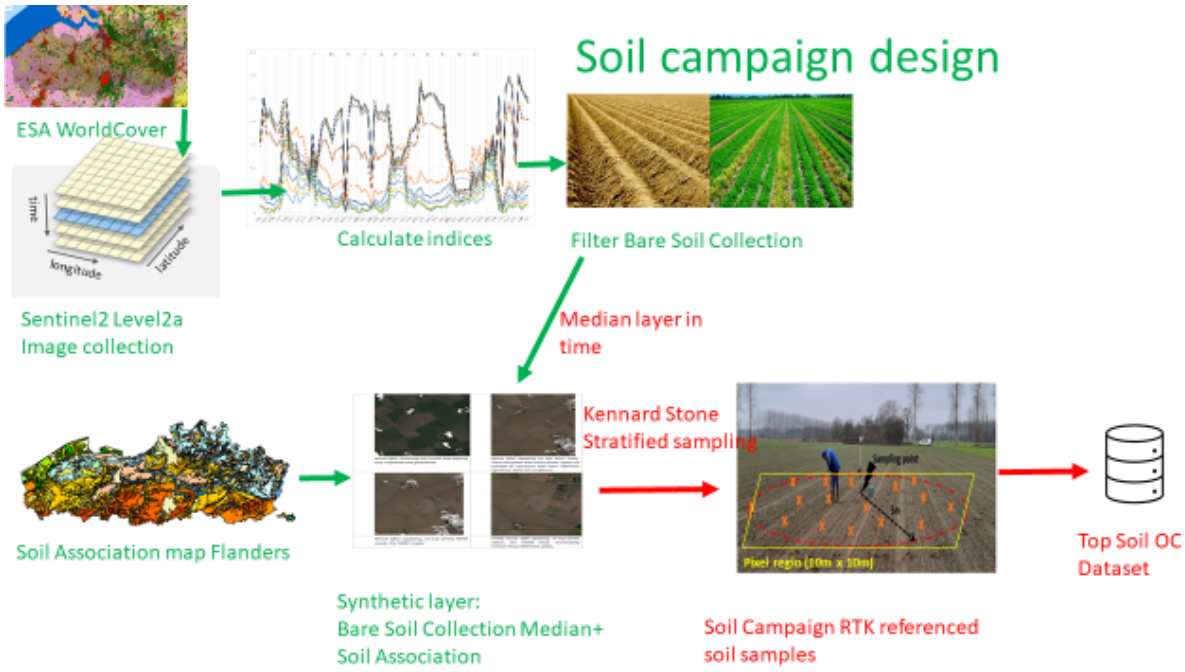


Figure 21: Input output schedules for soil campaign.

3.3 Monitoring organic farming requirements (OCS – Serbia)

The description of the data processing methodology for the creation of the Organic crop identification service is described on the current chapter.

Data processing for Crop practice Identification

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

- ESA SNAP Graph Processing Toolbox
- SAGA GIS
- Orfeo Toolbox and PhenOTB remote module
- R Libraries: mlr, caret, tidyverse, raster, sp, rgdal, tiff, ggplot2, maptools, zoo, signal, timeSeries, doParallel, dplyr
- 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.

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

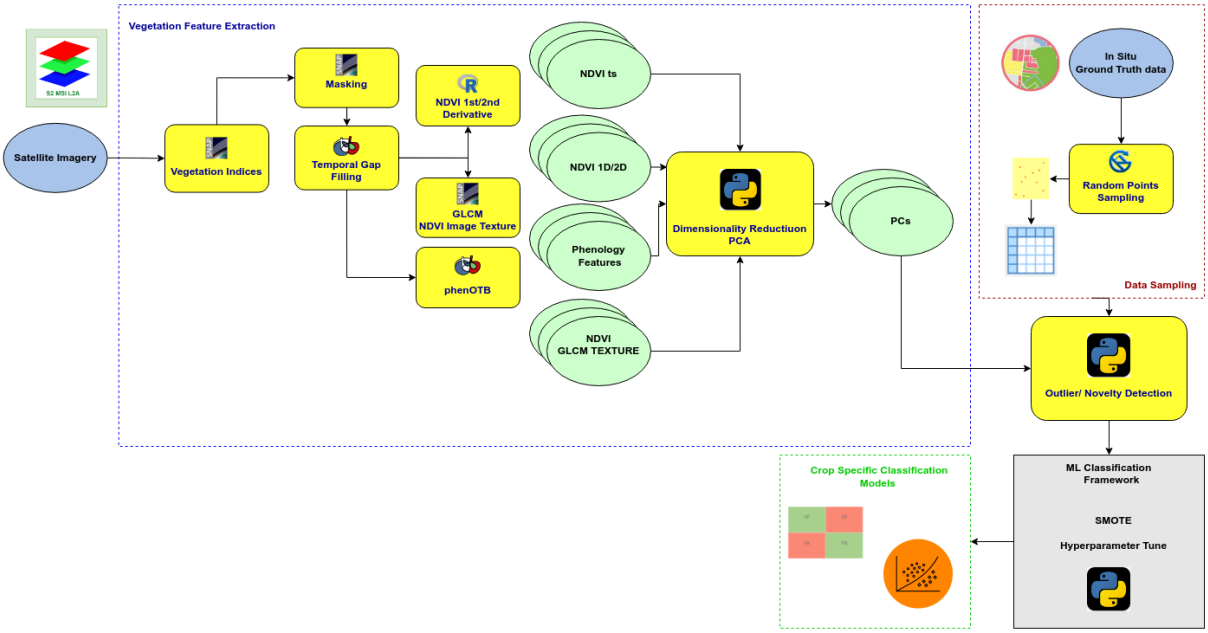


Figure 22: Methodological framework for the training of ML models for organic practice identification.

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 in order to reject pixels belonging to unwanted land cover for the specified classification task. As a result, only pixels denoted as vegetation or barren land were preferred, and all other SCL classes were omitted. 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 that would constitute the predictors of the classification models was therefore some kind of pre-processing that included:

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

- **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:** The creation of a layered raster archive containing all layers of the NDVI time series.
- **Temporal Interpolation** was applied on the masked NDVI layer stack, in order to fill the gaps created from image masking, as well as to create a regular temporal 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:

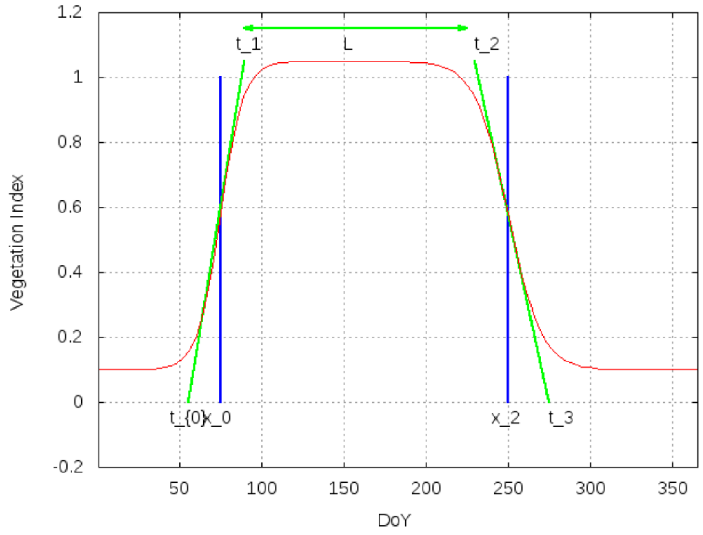


Figure 23: PhenOTB double logistic function fitting and associated parameters.

- Sowing date: t_0

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

- Date of Maximum Positive Gradient: x_0
- Maximum Positive Gradient – Crop Growth Slope: D_{Growth} - $g'(x_0)$
- Parameter related with logistic growth rate: x_1
- NDVI Plateu Initialization Date: t_1
- Plateu Termination Date: t_2
- Parameter related with logistic growth rate: X_3
- Date of Maximum Negative Gradient: x_2
- 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 $g_0(x_0)$, t_0, t_1, t_2, t_3, $g_0(x_2)$. 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= x_0 - t_0: Early Growth Stage Length
- CGS2= t_1 - x_0: Late Growth Stage Length
- CGS3= t_2 - t_1: Plateu Stage Length
- CGS4= x_2 - t_2: Early Senecence Stage Length
- CGS5= t_3 - x_2: Late Senecence Stage Length

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 in the field data and appeared to be valid at some point.
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 in the field data and appeared to be valid at some point.

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

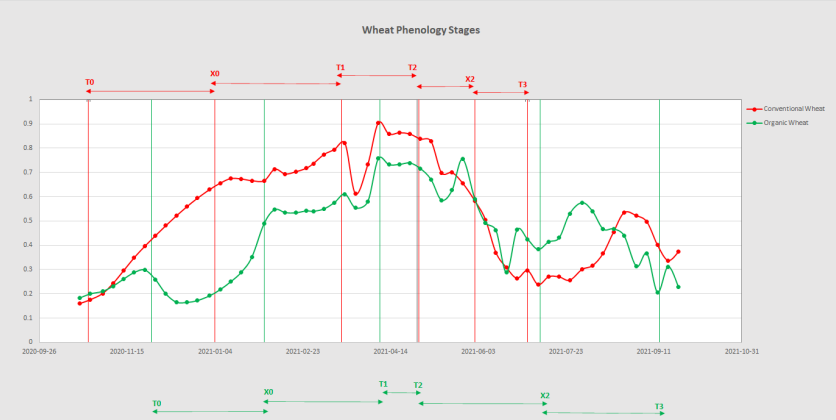
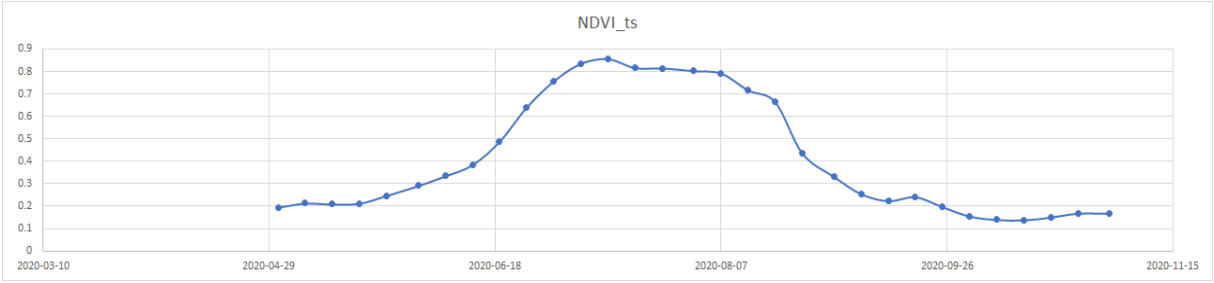


Figure 24: Comparative graph of phenology NDVI curves of a conventional vs an organic crop in relation with their crop growth stages.

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

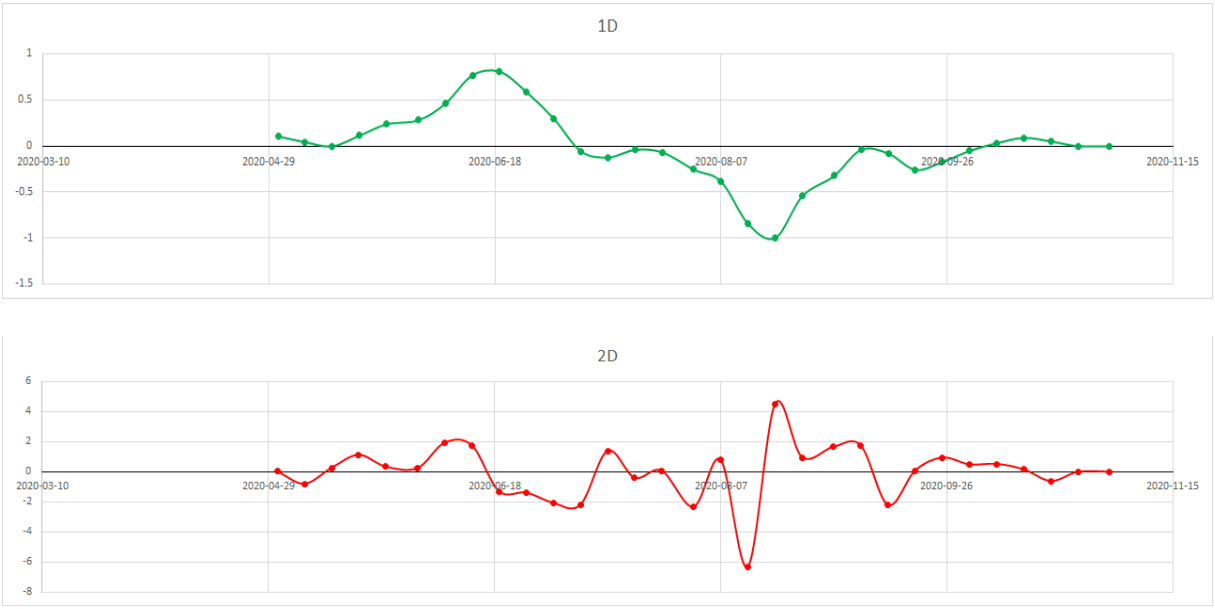


Figure 25: Comparative graph of a phenology NDVI curve of a conventional crop, along with its 1st and 2nd Derivative.

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, 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, would improve the classification results.

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 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

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

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

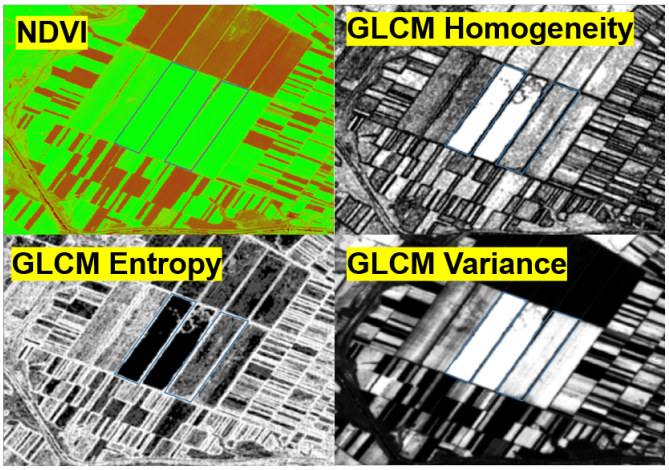


Figure 26: GLCM Texture Features of an NDVI Image. Homogenity, Entropy and Variance.

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 favored 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.

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

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. Organic carbon content can serve as an indirect determination of organic matter through the use of 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 (%).

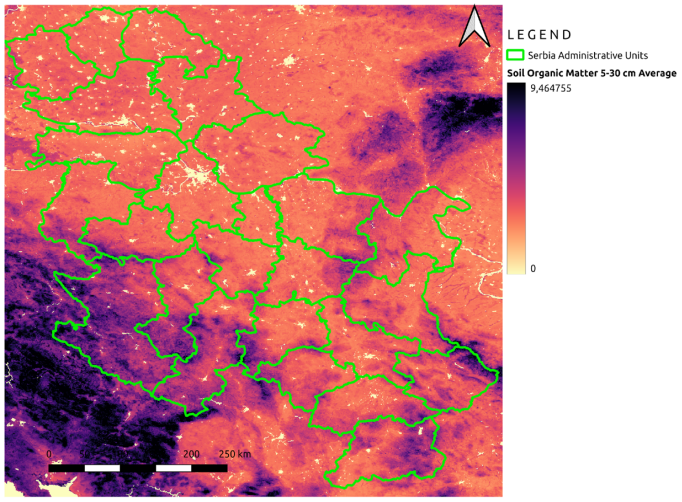


Figure 27: Soil Organic Matter map of Serbia (topsoil).

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.

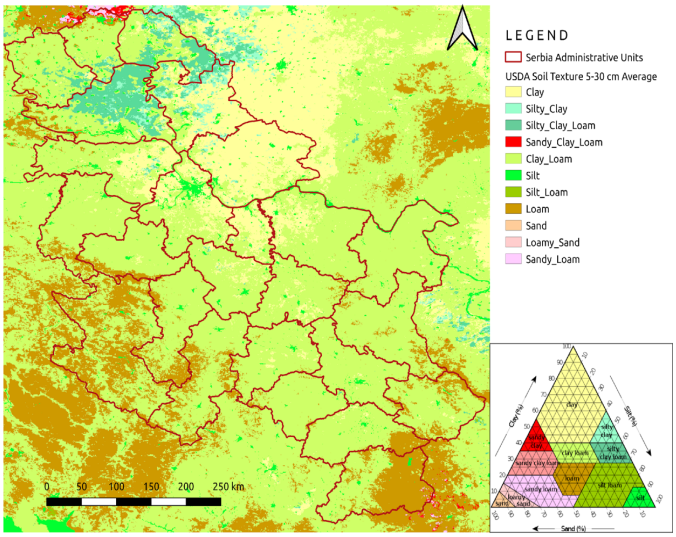


Figure 28: USDA Soil map of Serbia (topsoil).

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

Exploratory data analysis- 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.

A hybrid methodology was chosen to follow for the "cleaning" of the dataset including visual interpretation 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.

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 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.

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).

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

		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		

Figure 29: Formulation of the in-situ dataset after the Outlier/Novelty Detection Analysis.

The result of the Outlier/Novelty detection actions was a drastic reduction in the number of data to be analysed. 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

Data Anomaly Detection procedures using a hybrid approach of visual inspection and machine learning outlier/novelty detection proved to be very important when training the classification models. What was seen in many conventional/organic declarations was that there were incorrect entries in the crop type attribute. A fully automated process using only unsupervised methods did not yield good results. The hybrid approach was far more precise, but had the disadvantage of requiring the involvement of expert knowledge. Visual inspection is a relatively simple process that requires the user to have a fairly basic knowledge, which could possibly be trained, of what a timeseries NDVI profile displays in a crop, as well as the sowing/cutting dates in the area of interest. The user observes the profile of each field-sample in the training set, and scores it as to the correctness of its statement. A critical issue is scale up and the impact on human resources needed for the process. Ideally, to avoid any bias that may arise from continuous photo-interpretation, it would be legitimate to involve more than one user, in overlapping subsets. Considering that expert estimation could be provided by a limited number of users, it is implied that scale up is not linear, but on the contrary, a large increase in the training set disproportionately increases the human resources required. As an example, for the training set of ENVISION DP6, 7 man-days were needed for 3,500 samples, and the process has to be carried out/repeated in the early and full season classification model.

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

An alternative could be a concurrent crop classification system, such as the one that was also developed in the project, or similar ones (Sen4Cap etc), and the selection of a training set with high confidence in crop type estimation (high p-value). But also in this case the presence of outliers in LPIS is significant and this has its impact on the quality of the estimation.
In conclusion, an important data requirement that would enhance the business case would be the possibility for the user to upload a second auxiliary set of large-scale field visits data, even of older years, to train the novelty detection algorithm and improve the training dataset without the need for visual inspection of the NDVI profile.

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

4 Method per product tables

Table 8: Methods for DP1, DP2 and DP3

Product	Service	Data Input		Data Format	Spatial Distribution	Data Processing	Data Output	Data Format
DP1: Analytics on Vegetation and Soil Index Time-series	Harvest events detection	Sentinel 2 MSI L1C L2A	Farmers' Declarations (GSAA)	Shapefile & GeoTIFF	Lithuania	Atmospheric Correction & Cloud Masking (sen2cor) Inner Buffering (5m inward buffering) Datacube Indexing Spatial Aggregation Feature Extraction (vegetation and soil indices calculation) Cleaning Nan Temporal Gap Fill (linear interpolation)	Predicted harvest events on arable cultivations, including dates of the events detected at the parcel level	Shapefile
	Stubble burning identification on arable land		Farmers' Declarations (GSAA)		Lithuania and Cyprus		Identified stubble burning events on arable land, including dates of the events detected at the parcel level	

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

	Detection of illegal land clearing in Natura2000 protection areas		Land Parcel Identification System (LPIS) Natura2000 regions		Cyprus		Natura2000 Hotspot Detection in geometrical points based showing the alert pixels detected	
	Minimum soil cover for soil erosion		Farmers' Declarations (GSAA)		Lithuania and Cyprus		Bare soil percentage and minimum soil cover alerts for soil erosion	
	Runoff risk assessment for the reduction of water pollution in nitrate vulnerable areas		Land Parcel Identification System (LPIS) Nitrate Vulnerable Zones Hydrographic Network Rainfall erosivity factor layer (R) Soil erodibility factor layer (K) Slope length factor and slope steepness factor layer (LS) Crop and cover management factor layer (C) Conservation supporting practices factor layer (P) Digital Elevation Model (DEM)		Lithuanua and Cyprus		Runoff Risk level map of parcels within nitrate vulnerable zones	

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

DP2: Cultivated Crop Type Maps	Confirmation of GSAA	Sentinel 2 MSI L1C L2A Sentinel-1 GRD	Farmers' Declarations (GSAA) Lookup table	Shapefile & csv	Lithuanua and Cyprus	Atmospheric Correction & Cloud Masking (sen2cor) Inner Buffering (5m inward buffering) Datacube Indexing Spatial Aggregation Feature Extraction (vegetation indices calculation) Cleaning Nan Time Series Smoothing (moving average) Temporal Gap Fill (linear interpolation) Data Augmentation (SMOTE)	Dynamic crop type maps over the registered parcels for every new or group of new Sentinel acquisitions	Shapefile
	Smart sampling for OTSC inspections						Traffic light maps over the registered parcels for smart sampling of on-the-spot inspections and early alert of the users	
	Crops diversification compliance						Crops Diversification (Greening-1) compliance over the registered parcels at the end of the	

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

							cultivation period	
DP3: Grassland Mowing Events Detection	Grassland activity monitoring and management	Sentinel 2 MSI L1C L2A Sentinel-1 GRD	Farmers' Declarations (GSAA)	shapefile	Lithuania	Atmospheric Correction & Cloud Masking (sen2cor) Inner Buffering (5m inward buffering) Datacube Indexing Spatial Aggregation Feature Extraction (vegetation indices calculation) Cleaning Nan Time Series Smoothing (moving average) Temporal Gap Fill (linear interpolation) Data Fusion (S1/S2 CNN-RNN Model)	Dynamic Events Map of grassland mowing detection per parcel encapsulating all the extracted information regarding the detected events, their confidence levels	Shapefile

Table 9: Methods for DP4

Product	Service	Group	Data Input	Data Format	Thematic Content	Spatial Distribution	Data Processing	Method	Abandoned Method	Data Output	Data Format

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

DP 4	Top Soil Organic Carbon condition monitoring	Regional& soil texture information	**Landbouwgebruikspercelen LV, 2021, LPIS 2021**	Shapefile	Land Parcel Identification System	User defined across Flanders	Comparison to check performance with bare soil layer with crop classes				
							Extraction from field polygons, take averages	Aggregation		Parcel Level OC	Shapefile
			Organic Carbon Dataset EV ILVO	Excel, csv, shapefile	RTK GPS points linked to top soil organic carbon measurements	Flanders	Used as the basis for point sampling with Google Earth Engine reducers	Campaign designed Kennard Stone + Stratified sampling on soil association+ median reflection map		Table	csv
							Convert csv to shapefile	Data transformation		shapefile	Shapefile
							Remove high OC outliers	Pandas Thresholding based on histogram		Multidimensional dataset	Pandas table
			Soil Association Map Flanders	Shapefile	Soil association classes of Flanders	Flanders	Reclassification: some soil associations with similar characteristics are combined	Conditional replacing of features GEE		Soil Association Raster	GeoTiff
							Rasterization	ReduceToImage GEE		Soil Association Raster	GeoTiff

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

			Soil texture map of the Flemish Region according to the international soil classification system World Reference	Shapefile	Soil texture map of Flanders	Flanders	Link classes to average OC, compare with OC model	Adding data linked to classes, threshold compared to average OC, Rasterization		Raster with average OC per texture	GeoTiff
		Satellite and derived information	**Sentinel 2 MSI**	**Cloud optimised GeoTiffs**	VIS-NIR-SWIR reflectance bands	World, Flanders is selected for use case	Extract indexes derived from spectral bands	Google Earh Engine filtering, reducers		Image collection	Cloud-optimized Geotiff
			Level 2A				Atmospheric correction	Google Earh Engine filtering, reducers		Image collection	Cloud-optimized Geotiff
					Quality indicators (Cloud mask, Cirrus mask, Cloud probability, Snow Probability)		Removing clouds, snow	Google Earh Engine filtering, reducers		Image collection	Cloud-optimized Geotiff

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

						Feature extraction	Calculate indices	Google Earh Engine filtering, reducers		Image collection	Cloud-optimized Geotiff
			ESA World cover	Cloud optimized GeoTiff	Land use derived from S2	World, Flanders is selected for use case	Selected for 40, Cropland, 60, bare/ sparse vegetation, and 30 Grassland	Google Earh Engine filtering, reducers		Image collection	Cloud-optimized Geotiff
						World, Flanders is selected for use case					
						World, Flanders is selected for use case	Thresholding image collection based on NDVI, VNSIR, NBR2 indices	Google Earh Engine filtering, reducers, export	Threshold to select seed bed condition based on timeseries of vegetation indices like NDVI, dataset does not fit theoretic seasonal curves	Bare soil layer	Cloud-optimized Geotiff
			Bare Soil Image collection	Based on filtered	Bare soil pixels, in S2 image collection	World, Flanders is selected for use case	Extract features for OC sampled Points, Per Point	Google earth Engine Batch Export, reducers for points	Feature extraction Point collection in	Multidimensional dataset	Pandas table, csv

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

				S2 data, ESA					shapefile, computation limitations	for modelling	
			Bare Soil Image collection	Based on filtered S2 data, ESA	Bare soil pixels, in S2 image collection	World, Flanders is selected for use case	Median in time dimension	Google earth Engine Median on image collection	Map based on different outputs bare soil layer without taking median not possible computationally	Median Bare Soil Layer, Syntethic Layer	GeoTiff
		Modelling information	**Multidimensional dataset for modelling**	Pandas	Based on bare soil layer+ Soil association map reduced to image, linked to OC Points	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Split in Train, Validation, Test datasets	GroupShuffleSplit sklearn for select test set, Validation selection Pycaret to select on remaining data train-validation		Pycaret model	Pycaret model
			Multidimensional dataset for modelling, train dataset	Pandas	Based on bare soil layer+ Soil association map reduced to image, linked to OC Points	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Regression with Cross-Validation (Pycaret)	Extra Trees Regressor		Pycaret model	Pycaret model
								Support Vector Regression			

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

								Catboost			
								Random Forest Regressor			
								K Neighbors Regressor			
								Extreme Gradient Boosting (XGB)			
								Bayesian Ridge			
								Gradient Boosting Regression			
								Light Gradient Boosting Machine			
								Linear Regression			
								Huber Regression			
								Kernel Ridge			
								Ridge Regression			
								MLP Regressor			
								AdaBoost Regressor			
								Orthogonal Matching Pursuit			
								Random Sampling Consensus			
								Lasso Regression			
								Automatic Relevance Determination			
								Elastic net			
								Lassa Least Angle Regression			

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

								Passive Aggressive Regressor			
								TheilSen Regressor			
								Least Angle Regression			
			Pycaret Models, Train, Validation an Test Dataset	Pycaret Model, Pandas	Top soil OC models+ independent top soil oc validation, test dataset	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Validation results	Pycaret apply model on validation and test set		Model results	html, table, figures
			Pycaret Models, Train, Validation an Test Dataset	Pycaret Model, Pandas	Top soil OC models+ independent top soil oc validation, test dataset	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Sensitivity Analysis	Pycaret check fairness per model		html report, figures	html, images
			Pycaret Models, Train, Validation an Test Dataset	Pycaret Model, Pandas	Top soil OC models+ independent top soil oc validation, test dataset	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Feature Importance	Pycaret interpret model (Shapley, Reason, tSNE, …)		html report, figures	html, images

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

			Pycaret Models, Train, Validation an Test Dataset	Pycaret Model, Pandas	Top soil OC models+ independent top soil oc validation, test dataset	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Dataset features/ statistics	Pycaret Data Drift report between subsets		html report, figures	html, images
		Map Generation	Median Layer Bare Soil, or Synthetic layer + soil association info	GeoTiff	Bare soil pixels median+ soil association info	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Apply models on median layer bare soil	Pycaret predict_model		OC map Pixel Level	GeoTiff or shape file
			OC Pixel level Layer	GeoTiff	OC Predictions Flanders	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Aggregate info based on LPIS 2021 Polygons	Rasterstats, rasterization		OC map Parcel Level	GeoTiff
			OC Parcel level layer	Shapefile	Aggregated OC Predictions Flanders	Flanders, can be applied in other regions with new RTK linked	Results prediction on median image aggregated to LPIS layer, using polygon surface and pixel count, mean from rasterstats	Rasterstats, rasterization, thresholds	Absolute thresholds not adjusted to soil texture	Soil Quality Map Pixel Level	GeoTiff

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

						Top Soil OC Data					
			OC Pixel level Layer	GeoTiff	OC Predictions Flanders	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	Compare with rasterized avere OC based on texture	Rasterstats, rasterization, thresholds	Absolute thresholds not adjusted to soil texture	Soil Quality Map Parcel Level	GeoTiff
			SOC Quality map Pixel Level	GeoTiff	Compare OC results to Flanders Texture Map average OC per class	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	OGC compatible restfull API	Make available trough API		ogc geopackage	ogc
			SOC Quality map Parcel Level	Shapefile	Compare OC results to Flanders Texture Map average OC per class	Flanders, can be applied in other regions with new RTK linked Top Soil OC Data	OGC compatible restful API	Make available through API		ogc geopackage	ogc
		Webservices	API access to data product,	Using Djust	Open restful API		Making datamap available trough API	Google Cloud platform, Djust connect		image in	.jpeg or .tiff

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

			open restful API	Connect						browser	
			Envision Platform, webmap services		wms service		Making datamap available trough API			image in browser	.jpeg or .tiff

Table 10: Methods for DP5

Product	Service	Data Input	Data Format	Thematic Content	Spatial Distribution	Data Processing	Applied Method	Data Output	Data Format
DP5	**Distinction of Organic Farming Practices**	**LPIS (GeoSerbia)**	SQL table, Shapefile ,Geojson	Land Parcel Identification System	User Defined, across Serbian Adminitration Units	*Spatial Aggregation*	*Dissolve*		
		GSAA (OCS)	SQL table, csv	Farming Practice Declarations		*Spatial Proximity*	*Inner Buffering*		
							Compute Parcel Geometry Area		

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

						Compute Parcel Geometry Elongation	
					Data Descriptive Statistics	*Frequency Distribution*	
					Spatial Sampling	*Random Points*	
		SoilGrids TM	Raster Grids (.tif)	Soil Organic Carbon	User Defined Areas of Interest	*Raster Calculation*	*Compute Soil Organic Matter*
				Sand/Silt/Clay content		*Raster Reclassification, Conditional*	*Compute USDA Soil Texture*
		Sentinel 2 MSI	Raster Grids (.SAFE)	VIS-NIR-SWIR reclectance	User Defined Areas of Interest		
		L1C				*Atmospheric*	*Sen2Cor*

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

				Correction		
		L2A		*Feature Extraction*	*Vegetation Indices*	
					Quality Masking	
				Temporal Gap Fill	*Spline Interpolation*	
				Smoothing – Temporal Derivatives	*Savitzky – Golay Filters*	
				Image Texture	*GLCM Metrics*	
				Vegetation Phenology	*Double Sigmoid Curve Fitting*	
				Dimensionality Reduction	*PCA*	
				Outlier	*Isolation*	

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

	Detection	*Forest*	
	Novelty Detection	*One-Class SVM*	
	ML Classification Algorithm	*XG-Boost (XGB)*	
	Data Augmentation		
	Zonal Functions	*Zonal Parcel Area Tabulate*	
		Zonal Parcel Statistics (Mean, Standard Deviation)	

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

			Traffic Light – Confusion Metrics	Shapefile

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

5 Lessons learned

During the development of the automated pipeline for searching, preprocessing, and indexing Sentinel data (and metadata) into the Open Data Cube (ODC) on the CreoDIAS cloud platform, several key lessons were learned. These insights provide valuable knowledge for future projects and initiatives:

- Utilizing CreoDIAS' hosting of the **full archive** of Sentinel data eliminates the need for manual data downloads. Accessing the data in close proximity to the processing environment significantly reduces data transfer time and allows for more efficient and timely data processing. Leveraging the proximity to data minimizes latency and enables real-time or near-real-time data analysis, enhancing the overall pipeline's effectiveness.
- **Efficient data searching techniques** are crucial when dealing with large datasets like Sentinel data. By utilizing metadata and search filters provided by CreoDIAS APIs and objects storage repositories, the pipeline can achieve faster and more accurate data retrieval, enhancing the overall efficiency of the process.
- Preprocessing Sentinel data can be complex due to various formats and metadata. The development of a modular and flexible preprocessing pipeline allows for efficient handling of different data types and performing essential operations such as atmospheric correction, cloud masking, and geometric corrections. By **streamlining the preprocessing steps**, we have ensured data consistency, the improvement of data quality, and the facilitation of downstream analysis and visualization.
- Working in a cloud environment like CreoDIAS offers **scalability and flexibility** in terms of computational resources. It is essential to plan and optimize resource allocation based on workload demands to ensure cost-effectiveness and efficient resource utilization especially in the case of national scale monitoring. Automation of resource provisioning and management helps in dynamically scaling resources, ensuring the pipeline's smooth operation while keeping costs under control.
- Indexing the processed Sentinel data, as well as other metadata (e.g. GSAA, LPIS, meteorological etc.) and integrating it with the Open Data Cube provides a robust platform for data analysis and visualization. **Effective organization, indexing, and management of metadata** are essential for streamlined data exploration. By appropriately indexing the data, it significantly enhances query performance, enabling efficient spatial and temporal analysis, zonal statistics, and integration of data from diverse sources. Moreover, it facilitates data discovery within the broader Open Data Cube ecosystem.
- A crucial lesson we learned is the importance of clear guidelines that outline the workflow step-by-step, from data searching to pre-processing and indexing. By providing explicit instructions and explanations, we have saved a substantial amount of time and effort for the future since it enhances the transferability and reproducibility, that would otherwise be spent on figuring out the workflow implementation.
- **Customization and adaptability** have also emerged as key factors. The workflow is designed to accommodate various data formats, sources, and specific requirements of different

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

business cases. It employs modular and configurable components, enabling customization of the pipelines to fit individual datasets and integration needs.

Overall, the successful implementation of the data download and processing workflow using CreoDIAS has provided valuable lessons for the provision of national-scale datasets. These lessons have deepened our understanding of how to effectively integrate CreoDIAS into data acquisition and processing pipelines, contributing to the success and advancement of the ENVISION project and future projects alike.

Furthermore, a few lessons learned have been concluded with regards to soil organic carbon, such as:

- Soil sampling campaigns, aiming to support the development of TopSOC prediction models with the provision of training and validation data sets, should be carefully designed, considering various parameters that control the quality of the resulting digital signature libraries. Before the fieldwork, a bare soil collection needs to exist to be used together with other parameters like the soil type or the land use to identify soil sampling locations that vary in terms of spectral reflection signatures. This variation needs to be able to support the mapping of different TopSOC values at the full possible TopSOC value range and per different soil texture conditions. By following the aforementioned guidelines, we achieved, judging from the Lab results, within the Envision project, we developed a training set that covers the different expected TopSOC conditions. However, the distribution of the TopSOC values was not optimal to support the development of a TopSOC model that predicts values ranges above 2.5% with the same accuracy. The suggestion is to use, as input in this analysis, existing available information (maps) that have identified, even with low accuracy, different zones of TopSOC and, after, use these zones to cluster the reflection signatures further. Additionally, if feasible, the soil campaign should be performed at different periods and seasons to capture the bare soil reflection in different sunlight conditions and always under a clear sky. Moreover, the soil sampling campaign results control the performance of the model as they can prevent or generate overfitting prediction models.
- EO-based methodologies that target to predict the soil quality conditions at a larger scale directly by using the reflection signatures should always consider the need for reuse of the models or retraining of the models in other regions. That is related to the easiness of production and the operationality. The suggestion is for the models to rely on available harmonized data sources with more extensive coverage that can be easily consumed in a data product pipeline. This may affect the accuracy because, in some regions, the availability of data or the accuracy is higher; however, if the data are not harmonized, or the local conditions are too specific, then the reusability to other regions is eliminated. Ultimately, cost–benefit upscaling decisions need to be taken since the goal is the provision of commercial services. The use of Spatial-Temporal Asset Catalog services and the consumption of L3 data products can reduce the data acquisition and preparation costs, giving more time to generate data pipelines which allow the continued monitoring at a large scale, feed this way, microscale soil sampling efforts which is critical in smart farming.

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

Lastly, in the process of monitoring organic farming practices in Serbia, we have gleaned several valuable insights that will undoubtedly shape our future endeavours in this field. One of the most critical lessons learned is the paramount importance of data quality. The accuracy, completeness, and timeliness of the data collected directly impact the success of the monitoring process. Inaccurate or incomplete data can lead to incorrect conclusions and ineffective strategies. Therefore, we must prioritize ensuring the quality of our data in all future projects.

Another key insight is the importance of local knowledge. Our understanding of the local context and farming practices in Serbia was instrumental in interpreting the data correctly and developing relevant and effective strategies. This local knowledge is not something that can be overlooked or underestimated in its importance.

We also recognized the need for regular monitoring. Organic farming practices are not static; they evolve over time in response to various factors such as changes in weather patterns, market demands, or government policies. Regular monitoring allows us to keep track of these changes and adjust our strategies accordingly.

Our collaboration with local stakeholders was another significant aspect of our project. Working closely with local farmers, agricultural organizations, and government agencies enriched our work and made it more impactful. These stakeholders provided valuable insights, assisted in data collection, and were crucial for the successful implementation of our monitoring strategies.
The use of technology, such as remote sensing and machine learning, was a game-changer in our monitoring process. However, we learned that these technologies must be used judiciously, and the results interpreted correctly to avoid missteps.

Training and capacity building emerged as a crucial need among local stakeholders in Serbia, particularly regarding the use of new technologies for data collection and analysis. This training not only improved the quality of the data collected but also increased the effectiveness of the monitoring process.

Our project's adaptability was tested several times due to unexpected challenges such as changes in weather patterns or farming practices. We learned that flexibility and the ability to adapt quickly are key to the success of such projects.

Finally, we learned the importance of considering the sustainability of the monitoring process. This includes considering the environmental impact of the monitoring activities, as well as the long-term viability of the monitoring strategies. Sustainability is a factor that will be at the forefront of our future projects.

In conclusion, the lessons learned from the organic farming monitoring project in Serbia have provided us with valuable insights that will guide our future work in this field. We look forward to applying these lessons in our future projects and continuing to improve our monitoring processes.

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.