One-Stop-Shop Artificial Intelligence and Machine Learning for Official Statistics

Project 101146355 – AIML4OS

Work package 7

Earth Observation

Deliverable	7.1 Preliminary results
Month due	April 2026
Type	R — Document, report
Prepared by	WP7 team

Work package Leader:

Remco Paulussen (CBS, The Netherlands)
e-mail address : r.paulussen@cbs.nl
mobile phone : +31 (0)6 2123 0392

Index

1. Introduction

Earth observation (EO) is potentially a very rich data source and can be used effectively to integrate, supplement and augment datasets obtained from traditional (register) sources, to improve the accuracy, timeliness, and granularity of official statistics, but also to better interpret the official statistics. Using earth observation offers opportunities to reduce survey costs, enhance monitoring capabilities, and provide insights into areas previously difficult to assess, like environmental and climate change impacts.

In October 2021, the Directors General of National Statistical Institutes and of Eurostat acknowledged the importance and potential of earth observation at the DGINS Conference in Warsaw, Poland by signing the Warsaw memorandum. They emphasize the growing importance of integrating earth observation data into official statistics. They highlight the need for collaboration in key initiatives, like the Green Deal and Agricultural Data Spaces, and stress the value of earth observation for monitoring SDGs, agriculture, and maritime sectors. To advance this integration, they call for improved training, development of common methodologies, open and interoperable data infrastructures and closer cooperation among statistical bodies, space agencies, and other stakeholders.

An important aspect of earth observation is that there is a freely available public data source. Copernicus is the earth observation component of the European Union's space program, managed by the European Commission, and the Copernicus Data Space Ecosystem (CDSE) is an open ecosystem that provides free instant access to a wide range of data and services from the Copernicus Sentinel missions and more on our planet's land, oceans and atmosphere. At present time it contains more than 80 petabytes of data.

The Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS) project is funded by Eurostat and led by CSO, Ireland. The project focuses on the application of machine learning methods in official statistics. The Earth Observation Work Package (WP7) is led by Statistics Netherlands (CBS) and involves eleven statistical offices, Eurostat and representatives of Copernicus (CDSE).

Several statistical offices, but also other parties, are developing AI/ML models using earth observation, thereby showing promising results. The goal of WP7 is to investigate whether these AI/ML models using earth observation can be generalized to other countries/regions (space) and/or timeframes (time) and under what conditions, so that they can also be applied by other statistical offices. If this is feasible, methodological and implementation guidelines will be developed. As a side-effect, we will also have results in terms of (validated) predictions for other countries/timeframes than the original study.

2. Inventory and selection of the models

To establish if generalisation of earlier developed 'earth-observation data' based solutions for official statistics indicators over different countries and timeframes produces valid and comparable results, we first selected the AI/ML models using earth observation that would serve as our research objects.

The first task of the project was Task 7.1 – Make inventory. During this task, an inventory was made of the existing AI/ML models that are based on earth observation data and used by the participating statistical offices. In total, 25 models were identified. For these models, more information was gathered, and we asked ourselves the question of whether these models were viable and mature enough to perform our task at hand. After several meetings, we selected six candidate models from this inventory. The six candidate models were:

- Detection of buildings and changes – INSEE, France
- Crop mapping – GUS, Poland
- Yield forecasting – GUS, Poland
- Crop yield – Ministry of Agriculture, Food and Forestry (MAAF/SSP), France
- Land cover (STATeo) – Statistik, Austria
- Land cover and activities (CoSIA) – National Institute of Geographic and Forest Information (IGN), France

For these candidate models, even more information was collected. A template was developed containing the following topics:

- Information about the 'predictor' data sources, which are the data sources that are available on a regular basis and provide physical measurements (e.g. RGB, IR, radar images).
- Information about the 'ground truth' data sources, which are used to model the input data to make predictions about the 'predictor' data sources. Often these data sources are expensive to generate, hence the need for a model.
- Information about the preprocessing that is required on the data sources (both ground truth as well as predictor) to apply the modelling algorithms or to learn and train the models. Examples are the tiling and buffering of aerial images, shadow corrections, ortho corrections, reprojection of grids, etc.
- Information about the models that are used to generate predictions, like deep learning algorithms (convolutional neural networks).
- Information about the infrastructure that is needed to perform the training, testing and application of the models.
- Information about the software needed for the manual and automated process steps.
- Information about the output of the solution, so what kind of 'real life' phenomena are predicted and with which typology (classification).

On 13 and 14 November 2024, WP7 had their first and very successful physical meeting in Paris, France. The main goal of the physical meeting was to select which models we want to use as our study models and to make a high-level plan on how to proceed. In order to be able to select two models to take forward, first the selection criteria were defined: relevance, availability of required training and test data, accessibility (privacy), required infrastructure, interoperability, use of open source, quality measures (metrics to evaluate model), proven pipeline (readiness for production), availability of skills within the team.

In the next step, all six models were presented. After each presentation, the selection criteria were applied. Note that this was not an assessment of these models, but merely a way to select the best models for our task. Based on the assessment we selected two models: the land cover model from IGN, France, and the crop type model from GUS, Poland.

On the second day, we defined for each of these models, a high-level plan on how to proceed until 2026. Furthermore, we defined homework to gather necessary information to use these models and questions we collected for the use of the Copernicus Data Space Ecosystem (CDSE). We want to use the CDSE and, where necessary, combine it with the Onyxia platform of INSEE, France. Finally, we discussed how we would organize ourselves and decided to divide ourselves into two groups. The first group will work on the land cover model from IGN, France, under the supervision of ISTAT. The other group will work on the crop type model from GUS, Poland, under the supervision of Statistics Netherlands (CBS).

3. Land cover model

3.1 Introduction

The French Land Cover from Aerospace ImageRy land cover model from IGN France, from here onwards called FLAIR, is one of the two models selected for WP7.

FLAIR is trained to classify pixels from orthophotos into 15 different land-cover classes (see Table 3.1 below), and it can process orthophotos with different spatial resolutions and spectral band compositions: RGB and RG + Near-Infrared.

The model is a deep learning model based on attention mechanisms that uses a Swin-Transformer encoder backbone and a UPerNet decoder, and it has been re-trained on French orthophotos (see here for a more detailed description of the architecture and training data).

The goal of this project is to assess whether the FLAIR model can be reused on data from other EU countries without compromising on the quality of its output. To do so, we will build a generic pipeline that applies the model on a set of user-supplied orthophotos and that will output accuracy metrics that compare the output to user-provided true values, LUCAS 2018 strata or LUCAS 2022 results. This pipeline will be used on data from several European NUTS2 regions (Region Hovedstaden in Denmark, Salzburg in Austria and Veneto in Italy). Apart from being a tool in the current project, the pipeline is also an intended project output, as it will be disseminated (via GitHub) such that other NSIs can use it on their own data.

Table 3.1: FLAIR land cover classes and output bands

FLAIR class	FLAIR label
0	building
1	greenhouse
2	swimming pool
3	impervious surface
4	pervious surface
5	bare soil
6	water
7	snow
8	herbaceous vegetation
9	agricultural land

FLAIR class	FLAIR label
10	ploughed land
11	vineyard
12	deciduous
13	coniferous
14	brushwood
15	clear cut
16	ligneous
17	mixed
18	undefined

Note: classes 15, 16, 17 and 18 were deactivated during training of the FLAIR model. The output in this project does therefore never assign probability to these classes. A more detailed description of the classes can be found on page 5 in the FLAIR document. Note that class numbers are different in this paper, but labels are the same.

3.2 Pipeline

The pipeline is orchestrated through a coherent set of command-line interfaces (CLIs) that regulate the execution of all processing phases. Each phase is explicitly triggered via a dedicated CLI, ensuring controlled and reproducible activation of the workflow. Monitoring

CLIs provide real-time reporting of processing status by counting the total number of tiles, as well as those pending, running, completed, or failed for each stage. Download CLIs allow retrieval of both raster inference outputs and vector products (GeoPackage files) from object storage to local NSI environments.

The land cover processing pipeline is designed to transform heterogeneous national orthophoto datasets into harmonised, high-resolution products suitable for official statistical production.

The pipeline is built around the FLAIR land cover machine learning algorithm and embeds it within a structured production framework. It is conceived as an end-to-end system governing the entire land cover processing lifecycle — from the ingestion of raw orthophotos to the generation of statistically validated outputs. The objective is to integrate model-based inference within a controlled, reproducible, and traceable statistical production environment. The architecture follows a cloud-oriented design, leveraging scalable computational resources and object storage to ensure robustness and flexibility across national contexts.

The development of the pipeline addresses requirements inherent to the use of Earth Observation data in official statistics.

First, national orthophoto datasets are inherently heterogeneous. They differ in file formats (e.g. GeoTIFF, ECW), band configurations (RGB, RGB+IR, RGBI), tiling schemes, spatial resolution, and coordinate reference systems. Such heterogeneity requires a harmonisation layer capable of standardizing inputs.

Second, production at regional or national scale may involve processing thousands of raster tiles. This requires a robust batch-oriented execution model, isolation of execution errors and the ability to safely restart interrupted processes without duplicating completed work. Scalability and fault tolerance are therefore structural requirements rather than optional enhancements.

Third, outputs intended for statistical use must have coherent quality control procedures at the end of the process.

To address these challenges, the pipeline has been designed as a modular, cloud-oriented, storage-driven architecture in which storage and compute resources are fully decoupled.

At the core of the system, S3-compatible object storage is acting as the single source of truth. All harmonised input imagery, inference outputs, derived vector products, execution state markers and quality assessment results are stored within object storage under a structured project namespace. Computationally intensive operations, including model inference, are executed on GPU-enabled virtual machines that interact exclusively with object storage. This separation between storage and compute ensures scalability, robustness, efficient resource allocation, and infrastructure flexibility.

Each pipeline execution is uniquely identified by a project identifier that defines an immutable namespace in object storage and guarantees full traceability of all intermediate and final products associated with a specific execution. By structuring the workflow around this

identifier, the pipeline ensures consistency, reproducibility of results, and comparability across national implementations.

The pipeline is organized into clearly separated yet interoperable processing phases:

a. Preprocessing & Upload (Local Phase)

Preprocessing is executed locally through a dedicated command-line interface. This phase includes ingestion of heterogeneous orthophoto datasets, harmonization of band configurations, optional band merging (e.g. RGB + IR), conversion to the standard GeoTIFF format and upload of prepared imagery to object storage.

Upload operations are idempotent: existing files are not transferred again. This design ensures that responsibility for data preparation remains at national level, while subsequent cloud-based processing operates on technically consistent inputs.

b. Inference (Cloud Phase)

The FLAIR deep learning algorithm for land cover classification benefits significantly from GPU-based parallel computation. For this reason, inference is executed on GPU-enabled virtual machines within a cloud environment. The pipeline is therefore designed for scalable cloud deployment.

The pipeline does not modify the internal logic of the model; rather, it orchestrates its execution in a controlled and deterministic manner.

Inference operates in batch, tile-based mode. For each tile stored in object storage, runtime configuration files (YAML) are generated dynamically, enabling flexible output modes (e.g. class labels or class probabilities). Errors are isolated at tile level, and previously completed tiles are automatically skipped in case of re-execution. This restart-safe behaviour is achieved through systematic output existence checks and phase-specific execution markers stored in S3-compatible object storage.

The FLAIR model can operate under two different configurations, with a different number of parameters and therefore varying levels of computational complexity during inference. It is possible to choose which configuration to use — small or large — depending on the desired trade-off between computational cost and model capacity. These options can be selected directly from the command line at runtime.

c. Vectorisation

Raster outputs may optionally be transformed into vector products (GeoPackage layers) through an optional vectorisation phase. This phase reads raster outputs directly from object storage and maintains independent execution state markers. Intermediate raster files may optionally be removed after successful completion, ensuring storage efficiency while preserving traceability.

d. Production of aggregated statistical tables

The pipeline includes a dedicated phase for producing aggregated land-cover statistical tables for the FLAIR model classes. Statistics are computed from the pixel counts assigned to each land-cover class and are provided both in terms of area and percentage coverage. Any overlaps generated by the tiling system are handled through clipping operations or predefined priority rules, in order to ensure consistency and avoid duplication in the aggregated results.

e. Integrated Quality Assessment

Quality assessment, integrated within the pipeline, is conducted through comparison with reference points derived from photointerpretation or official statistical sources (e.g. LUCAS). To this end, the pipeline extracts class labels and associated class probabilities at the locations corresponding to the reference points. This approach enables the computation of accuracy metrics consistent with the characteristics of the mapping process, including top-1 accuracy, the Brier score, the Brier Skill Score, and the soft Jaccard index. Since the generation of probability outputs can be computationally demanding, a subset of tiles can be selected from which reference points are sampled.

3.3 Preparation of datasets

At the beginning of the project, the participating countries were asked to complete a questionnaire describing the characteristics of the orthophotos available to them, which are summarised in the table 3.2 below. The choice for the countries and regions included was mostly practical. We chose home-countries from project participants and selected regions from those countries that were relatively small (to reduce the data size) and contained a diversity of land-cover classes. The final selected regions were Region Hovedstaden in Denmark, Salzburg in Austria and Veneto in Italy.

Table 3.2: Description of the orthophotos of the example regions.

Aspect	**Austria**	**Denmark**	**Italy**
Orthophoto bands	RGB + NIR	R, G, B, NIR	RGB+NIR
Bit_dept	8-bit integer	32-bit	8-bit integer
Region	Salzburg (7.154 km²)	Region Hovedstaden (2.559 km²)	Veneto (18.399 km²)
Tiling	Yes (3 tiles)	Yes (3613 tiles)	Yes (2253 tiles)
Available format	WMS, Geotiff	ECW	ECW
Spatial resolution	20 cm	10–12.5 cm	20 cm
Latest available year	2024	2024	2021

Note: the table shows the available data in their heterogeneity in terms of spatial resolution and format, temporal availability, acquisition modality (for Austria via WMS), and tiling scheme (for Italy and Denmark, thousands of tiles).

3.4 Code and infrastructure

The FLAIR code

Developed by the French mapping agency IGN, the FLAIR initiative emerged in the context of efforts to support the production of a national high-resolution land-cover reference, particularly in relation to France's net zero land take (ZAN) objective. Within this framework, FLAIR provides large benchmark datasets and associated pre-trained models focused specifically on AI-based land-cover mapping from Earth observation and aerial imagery.

The core task of the FLAIR models is large-scale semantic segmentation, i.e., predicting a class for each pixel in order to map features and surface types such as buildings, impervious/pervious areas, water, bare soil, greenhouses, and vegetation/crop categories. The models were trained on roughly 2,800 km² of authoritative geospatial data produced or curated by IGN and annotated through expert photointerpretation.

The code related to the FLAIR models is distributed under an Apache-2.0 license. It is built on widely used Python deep-learning libraries (notably *PyTorch*, *PyTorch Lightning*, *segmentation-models-pytorch* and *Transformers* backends) and is configured through YAML files for data paths, model settings, training, prediction, and evaluation. The code supports both patch-level workflows (training, inference, and metric computation) and large-area inference on georeferenced rasters via a dedicated detection module. It also abstracts compute scaling across hardware resources (CPU/GPU, multi-GPU, and multi-node distributed setups).

For inference on large images, the pipeline automatically tiles the input raster into smaller patches, runs the model patch by patch, and supports overlapping windows (via a margin parameter) to reduce boundary artifacts at tile edges. The code can also include a final post-processing step to convert raw model outputs into a vectorized land-cover product. This step typically involves identifying and removing small, isolated pixel clumps below a predefined size threshold; converting the raster predictions into polygons and smoothing their boundaries to reduce vertex count and raster-induced jagged edges.

The Land Cover Pipeline code

As described in Section 3.1, the land cover pipeline is a software application written in Python and distributed on GitHub[1] under the EUPL v1.2 license. The land cover pipeline uses FLAIR as a wrapper, which is called during inference, to generate land cover maps. The land cover pipeline, more generally, orchestrates the entire process of producing land cover maps and related statistics, executing the scripts of preprocessing, inference, vectorization, quality measurement, coverage table production, process monitoring utilities, and output downloading.

The pipeline is currently being tested in the Veneto region, where the large number of tiles used to organize the Italian orthophotos required the development of software capable of ensuring reliable execution at the regional level.

[1] https://github.com/AIML4OS/WP7_landcover

Infrastructure

For the execution, Copernicus Data Space Ecosystem (CDSE) provided us with access to the Earth Observation user platform CREODIAS, built on CloudFerro infrastructure.

For our use case, we created a shared network for the land cover and crop projects and made use of the virtual machine provisioning services and S3-compatible object storage.

In a preliminary study phase, we selected a small area around the city of Venice and estimated execution times and potential costs across different virtual machine configurations. We also tested the FLAIR model by varying several configuration parameters in order to assess their impact on computational performance on which the land cover pipeline saves input, output, and process files. For our purposes, virtual machines can be configured in different sizes. For our pipeline, we opted for Linux OS and GPUs.

Table 3.3 presents the results of the comparative analysis carried out on the different virtual machine configurations. The results show that the vm.l40 series machines, equipped with GPUs, achieve significantly better processing times compared to CPU-only configurations.

Although GPU-enabled machines have a higher hourly cost, the substantial reduction in execution time more than compensates for this difference, resulting in a lower overall processing cost for large-scale tasks.

Spot instances, while offering very competitive pricing, are not recommended unless appropriate backup and checkpointing mechanisms are in place, as they may be terminated unexpectedly, potentially leading to the loss of intermediate results.

Based on the current estimates, the time required to process a region comparable in size to Veneto (18 399 km²) is approximately four days using a GPU-based configuration.

Table 3.3: Comparative analysis of virtual machine configurations

VM Configuration	vCore	RAM (GB)	vGPU RAM (GB)	Price €/h	Time	Workers	Batch	Image Size	Days	Cost NUTS (€)
spot.vm.l40s.2	8	32	12	0.25	00:02:30	8	2	1024	4	23
vm.l40s.4	16	64	24	0.99	00:01:58	1	2	1024	3.1	73
vm.l40s.4	16	64	24	0.99	00:02:32	16	2	2048	4	94
vm.l40s.4	16	64	24	0.99	00:03:00	1	2	512	4.7	111
eo2a.3xlarge	16	64	-	0.67	00:26:02	16	2	4096	41	657
hmad.xlarge	8	64	-	0.46	00:25:24	8	2	4096	40	437
eo2a.2xlarge	8	32	-	0.34	00:37:07	8	2	2048	58	468
hmad.large	4	32	-	0.23	00:44:55	4	1	4096	70	386

For our pipeline, we ultimately selected a configuration consisting of two Linux Ubuntu vm.l40s.4 virtual machines for processing the data for Italy and Austria. Denmark, for processing its region, will be able to use one of these two machines.

The costs reported refer to the prices observed on CloudFerro website in November 2025 and are used for evaluation purposes.

Object Storage is managed on CREODIAS in S3 containers; we loaded several terabytes of input and output data. Adopting an S3-compatible object storage system offers significant advantages for the land cover pipeline. First, it enables a clear separation between storage and processing, allowing virtual machines to be replaced or scaled independently without

compromising data persistence. All input data, intermediate processing outputs, and final results are stored centrally and accessible via a unified namespace.

This architecture facilitates orchestration and remote monitoring from local computers, as processing status, execution markers, and intermediate results can be inspected directly within the object storage environment without requiring direct access to the virtual machines. Furthermore, upload and download operations are completely decoupled from the VM lifecycle: data can be transferred before processing begins, during execution, or after completion, ensuring operational flexibility and robustness.

3.5 Evaluation of the results

In this section, we will evaluate the quality of the results regarding accuracy of the predictions. Evaluation of other quality dimensions will be given in the next section, "evaluation of the generalization".

To evaluate the results the pipeline includes automatic accuracy assessment against user-provided true-values, LUCAS 2018 strata and user-provided LUCAS results (from any year, can be downloaded from the Eurostat website as .xlsx files). The accuracy assessment returns several different metrics: standard accuracy (correct classifications / all classifications), several versions of Brier scores and a Jaccard index. A more elaborate description and their implementation in the pipeline can be found below.

Note that for comparison with LUCAS 2018 strata and LUCAS (2022) results the FLAIR classes have to be mapped to the relevant LUCAS classes. Table 3.4 and 3.5 show the mappings from the FLAIR output bands to the LUCAS 2018 strata (table 3.4) and LUCAS classes (table 3.5) used in this project. Note that we do not provide mappings for FLAIR classes 15-18 since they are deactivated in the FLAIR model which means they are never assigned any probability.

Table 3.4: Mapping: mapping from FLAIR output bands to LUCAS 2018 strata

FLAIR class	FLAIR label	LUCAS 2018 stratum code	LUCAS 2018 stratum label
0	building	7	artificial
1	greenhouse	7	artificial
2	swimming pool	8	inland water
3	impervious surface	7	artificial
4	pervious surface	7	artificial
5	bare soil	6	bare land
6	water	8, 9	inland water, transitional water
7	snow	8	inland water
8	herbaceous vegetation	3	grassland
9	agricultural land	1, 2	arable land, permanent crops
10	plowed land	1	arable land
11	vineyard	2	permanent crops
12	deciduous	4	wooded
13	coniferous	4	wooded
14	brushwood	5	shrubland

Table 3.5: Mapping from FLAIR output bands to LUCAS classes (as used in 2022 results)

FLAIR class	LUCAS class
0	A10, A11, A12, A30
1	A13
2	G11, G12
3	A20, A21, A22, A30
4	A20, A21, A22, A30
5	F10, F20, F30, F40
6	G10, G11, G12, G20, G21, G22, G30, G40
7	G50
8	E10, E20
9	Bx1, Bx2, B10, B11, B12, B13, B14, B15, B16, B17, B18, B19, B21, B21, B22, B23, B30, B31, B32, B33, B34, B35, B36, B37, B40, B41, B42, B43, B44, B45, B50, B51, B52, B53, B54, B55, B70, B71, B72, B73, B74, B75, B76, B77, B80, B81, B83, B84
10	Bx1, Bx2, B10, B11, B12, B13, B14, B15, B16, B17, B18, B19, B21, B21, B22, B23, B30, B31, B32, B33, B34, B35, B36, B37, B40, B41, B42, B43, B44, B45, B50, B51, B52, B53, B54, B55, B70, B71, B72, B73, B74, B75, B76, B77, B80, B81, B83, B84
11	Bx2, B82
12	C10
13	C20, C21, C22, C23
14	D10, D20, E30

3.5.1 Implementation of accuracy measures

Inputs

The script accuracy_measures.py evaluates model outputs against reference labels using two tab-delimited files:

1. Predictions file (user input): contains one column with the observed reference label t_i for each observation i, and a set of probability columns named band_* representing the FLAIR model's output classes. These band_* values are stored on a 0–255 scale and are converted to probabilities by division by 255. The predictions file is generated using the script extract_predictions.py.

2. Mapping file: links each FLAIR output class (column Band) to one or more reference labels (column LUCAS_2018_stratum in the current configuration). The mapping is read into a dictionary $C(k)$ that returns the set of valid reference labels for a given FLAIR class k. There are three mapping files provided in the pipeline: FLAIR_LUCAS_STRATA_MAP (for the LUCAS 2018 strata), FLAIR_LUCAS_CLASS_MAP (for LUCAS 2022 classes) and FLAIR_NO_MAP (for manual photointerpretation using the original FLAIR classes).

Pre-processing and matched observations

Let N_{total} be the number of observations in the predictions file. We first check which reference labels occur in the mapping. Observations with reference labels not present anywhere in the mapping are excluded from all metrics. Let N be the number of remaining matched observations. A list of excluded reference labels (in data but not included in the mapping) is printed to the terminal and written to the output.

Notation

For each matched observation $i \in \{1, \dots, N\}$ and each FLAIR class $k \in 1\{, \dots, K\}$:

- $p_{ik} \in [0,1]$: predicted probability observation i and class k (after scaling to 0–1 and merging)
- t_i: observed reference label for observation i
- $C(k)$: set of reference labels mapped to original FLAIR class k, derived from the mapping file

Mapping-based target vectors and row-normalization

Because a single reference label can correspond to multiple FLAIR classes (many-to-many mapping), the script constructs a mapping-based target indicator vector $y_i \in \{0,1\}^K$ as:

$$y_{ik} = \begin{cases} 1, & if\, t_i \in C(k) \\ 0, & otherwise. \end{cases}$$

If a reference label matches multiple combined classes, the row sum $\sum_k y_{ik}$ can exceed 1. Therefore, we row-normalize the target to sum to 1:

$$\tilde{y}_{ik} = \frac{y_{ik}}{\sum_{j=1}^{K} y_{ij}}$$

with $\sum_j y_{ij} = 0$ handled by setting the divisor to 1 (to avoid division by zero).

Interpretation: if one reference label maps to two classes, the “true” target mass becomes (0.5, 0.5) across those classes rather than (1, 1), preventing those observations from being over-weighted in squared-error metrics.

Implemented accuracy-related measures

Option 1: Top-1 mapping-consistent accuracy

The predicted class for observation i is the most likely combined class:

$$\hat{k}_i = \arg\max_k p_{ik}$$

An observation is counted correct if its reference label is contained in the mapped label set of the predicted combined class:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1, \{t_i \in C(k)\}$$

How to interpret: $Acc \in [0,1]$, higher is better. Unlike standard multiclass accuracy, this is a *mapping-consistent* top-1 accuracy: it treats a prediction as correct if the predicted FLAIR class is compatible with the reference label under the mapping (even when there is no one-to-one correspondence).

Option2: Multiclass Brier score (BS) and Brier Skill Score (BSS)

The multiclass Brier score is computed using the row-normalized mapping-based targets:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - \tilde{y}_{ik})^2$$

How to interpret: lower is better. It measures squared error of the full probability vector and strongly penalizes confident probability assigned to wrong classes. Row-normalization ensures that observations whose reference label maps to multiple classes do not automatically contribute disproportionately large errors. Since the maximum BS can deviate from 2 in our case where FLAIR classes can map to multiple reference labels (in case we use LUCAS data), we normalize the Brier score to a 0-1 scale by computing $\frac{BS}{BS_{max}}$

BS_{max} (worst-case Brier score) is estimated by simulating worst-case predictions by assigning probability 1 to a class outside the target support for each observation and computing a BS for this setting. By normalizing the Brier Score it is easier to compare FLAIRs performance on different test data (manual photointerpretation, LUCAS data, etc.)

Because Brier scores are difficult to interpret on its own, we also report a Brier Skill Score which is a relative metric comparing the predictions from FLAIR to constant baseline reference predictions (p_{ik}^{ref}). The reference predictions are defined by the empirical mean of the normalized targets for each class:

$$p_{ik}^{ref} = \frac{1}{N}\sum_{i=1}^{N}\tilde{y}_{ik}$$

The reference Brier Score (BS_{ref}) and Brier Skill Score are then defined as:

$$BS_{ref} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(p_{ik}^{ref} - \tilde{y}_{ik}\right)^2, BSS = 1 - \frac{BS}{BS_{ref}}, \textit{(with } BSS = 0, \textit{if } BS_{ref} = 0\textit{).}$$

How to interpret $BSS = 1$: is perfect; $BSS = 0$ means no improvement over predicting the empirical mean class frequencies; $BSS < 0$ indicates predictions are worse than baseline.

Option 3: Soft Jaccard index (probability intersection-over-union)

Instead of a hard-label multiclass Jaccard/IoU, the script computes a soft (probability-based)

Jaccard index:

$$J = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} min(p_{ik},\tilde{y}_{ik})}{\sum_{i=1}^{N}\sum_{k=1}^{K} max(p_{ik},\tilde{y}_{ik})}.$$

How to interpret: $J \in [0,1]$, higher is better. These measures overlap between predicted probability mass and the mapping-based target distribution. It gives partial credit when the model assigns nonzero probability to target-compatible classes even if those classes are not top-1, and it naturally handles cases where the reference label maps to multiple FLAIR classes.

Outputs reported

The output includes overall values for $Acc\ BS$, BSS, J, BS_{max}, $\frac{BS}{BS_{max}}$ (normalized Brier scores), and also reports per-class Brier scores and Jaccard values. We also list reference labels present in the predictions file but missing from the mapping (excluded from scoring).

Accuracy results for Denmark

For the Region Hovedstaden in Denmark the overall metrics and the metrics per FLAIR class (Normalized BS and Jaccard Index) have been calculated based on the 2024 data. The results are shown in Table 3.6 and 3.7.

Table 3.6: Overall metrics for Denmark and different validation data

Region Hovedstaden (DK), 2024 data	**#points**	**BS (0-1)**	**BSS**	**Jaccard-index**	**Accuracy**
Manual photointerpretation	993	0.05	0.88	0.72	0.996
LUCAS 2018 strata	643	0.39	-0.46	0.26	0.739
LUCAS 2022 survey	361	0.38	-0.10	0.30	0.648

Table 3.7: Per class metrics (Normalized BS and Jaccard Index) for different validation data for Region Hovedstaden (DK)

	Manual photointerpretation		LUCAS 2018 strata		LUCAS 2022 survey	
FLAIR Class	**BS (0-1)**	**Jaccard-index**	**BS (0-1)**	**Jaccard-index**	**BS (0-1)**	**Jaccard-index**
building	0.0023	0.77	0.027	0.20	0.0029	0.70
greenhouse	*	*	0.0079	0.01	*	*
swimming pool	*	*	0.0035	0.00	0.0065	0.00
impervious surface	0.0032	0.76	0.031	0.24	0.013	0.33
pervious surface	0.0021	0.46	0.016	0.06	0.021	0.08
bare soil	0.00034	0.55	0.0034	0.08	0.030	0.02
water	0.0021	0.75	0.024	0.26	0.019	0.31
snow	*	*	0.0036	0.00	*	*
herbaceous vegetation	0.016	0.69	0.074	0.28	0.088	0.37
agricultural land	0.0052	0.86	0.058	0.41	0.11	0.32
ploughed land	0.0017	0.81	0.057	0.15	0.066	0.15
vineyard	*	*	0.00052	0.00	0.00023	0.00
deciduous	0.0094	0.68	0.032	0.46	0.027	0.49
coniferous	0.0016	0.52	0.037	0.15	0.0065	0.20
brushwood	0.0035	0.27	0.0081	0.04	0.013	0.04

Note: (*) indicate these classes were not present in the validation data (according to the mapping)

3.6 Evaluation of the generalisation

The primary objective of WP7 is to assess whether the FLAIR land cover model, originally trained on French orthophotos, can be transferred to other European regions while maintaining its good performance.

At the current stage of the project, generalization cannot yet be considered fully proven. The land cover pipeline has been technically implemented and tested primarily in the Italian case (Veneto region), where large-scale processing, cloud deployment, and inference for the production of raster and vector land cover maps were successfully performed robustly and efficiently. This confirms the pipeline's operational feasibility under production conditions similar to those in Italy, with orthophotos divided into thousands of tiles.

However, full cross-country validation has not yet been completed. In Denmark, the quality assessment procedure was performed outside the land cover pipeline on local computers, and consolidated quality assessment results for all countries are not yet available. Therefore, a comprehensive statistical comparison across national contexts is not yet possible.

Preliminary results from Denmark indicate strong agreement when validated with manual photointerpretation, while less agreement is observed when comparing the forecasts with LUCAS-based reference data. Several factors may explain these discrepancies:

- Differences in definitions between FLAIR and LUCAS classes, which require many-to-many mapping.
- The inherent difference between field observations and what is observable in orthophotos.
- The temporal discrepancy between the acquisition of Danish orthophotos (2024) and the LUCAS surveys (2018 and 2022), and consequently potential changes in land cover occurring between the reference years. Indeed, some land cover classes can change very rapidly.

Furthermore, harmonised land cover tables for different countries have not yet been produced. Therefore, a comprehensive statistical comparison between countries, including comparison with LUCAS survey data at NUTS2 level, is not yet possible.

3.7 Next steps

The project is now in its third year. The preparation and implementation phase (Task 7.2) is well advanced, and over the next 12 months the focus will be on consolidating the land cover pipeline and making progress towards producing harmonised regional statistics. The inference of the model was done in the cloud, but certain processing steps have been done locally. To process the entire pipeline in the cloud certain adaptations for different countries are necessary, also the quality assessment has to be performed by all countries. In Table 3.8 we summarize the stage at which the countries are in the pipeline execution.

The planned activities for the next period will be:

- Running the entire land cover pipeline in the cloud environment for Denmark at NUTS2 regional level.
- Testing and adapting the pipeline for Austria at NUTS2 level, addressing specific technical constraints related to orthophoto access and processing.
- Producing tables with land cover percentages for the different classes identified by the FLAIR model for the participating countries at NUTS2 regional level.
- Running the quality assessment procedure within the land cover pipeline for all participating countries at NUTS2 level. If, during this quality assessment phase, the results are significantly lower than the performance achieved in France, the pipeline and the model may be adjusted accordingly.

Table 3.8: Status of the land cover pipeline execution per country

Country	Pipeline execution	NUTS2 Quality assessment	NUTS2 Land Cover tables
Italy	Completed (cloud)	In preparation (cloud)	In preparation (cloud)
Denmark	Not yet	Completed (local)	Not yet
Austria	In preparation (cloud)	In preparation (cloud)	In preparation (cloud)

The study could be extended to other regions or to the entire national territory of participating countries in order to further assess scalability, robustness, and usability under broader production conditions.

By applying the FLAIR model outside its original French context and adapting it to three different national contexts (Italy, Denmark, and Austria), the project has demonstrated the feasibility of a harmonized approach to land cover data production.

This harmonization ensures methodological consistency of results across countries, a key element for producing comparable European statistics. Furthermore, the implementation process itself, including the challenges encountered in managing heterogeneous orthophoto formats, tiling schemes, and local infrastructure, has generated valuable technical knowledge that can directly support and guide future implementations by other NSIs.

4. Crop type model

4.1 Introduction

The crop type model investigated in WP7 originates from Statistics Poland (GUS) and was selected as one of the two study models for assessing the generalisability of earth observation–based AI/ML pipelines across countries and timeframes.

The original Polish implementation distinguishes multiple crop types using Sentinel-1 radar time series in combination with advanced object-based classification software.

Within WP7, the objective was not to replicate the full operational system, but to develop a simplified and modular version of the processing chain that could be shared and tested across National Statistical Institutes (NSIs) with a realistic infrastructure and governance context. The focus was therefore placed on determining whether a radar-based temporal signal derived from Sentinel-1 data (specifically or Ground Range Detection data [GRD]) can be used to distinguish broad crop categories in multiple member states.

The conceptual premise underlying the model is illustrated in Figure 4.1 and 4.2. Different crops exhibit distinct phenological development cycles throughout the agricultural year. These phenological differences translate into distinguishable temporal backscatter signatures in radar data. Winter crops, spring crops and rapeseed follow structurally different growth trajectories between autumn and late spring, which are reflected in the amplitude and seasonal data of Sentinel-1 backscatter signals. The model therefore relies primarily on the temporal behaviour of radar reflectance rather than on single-date imagery.

Figure 4.1: Types of radar signal scatter from crops during growing season

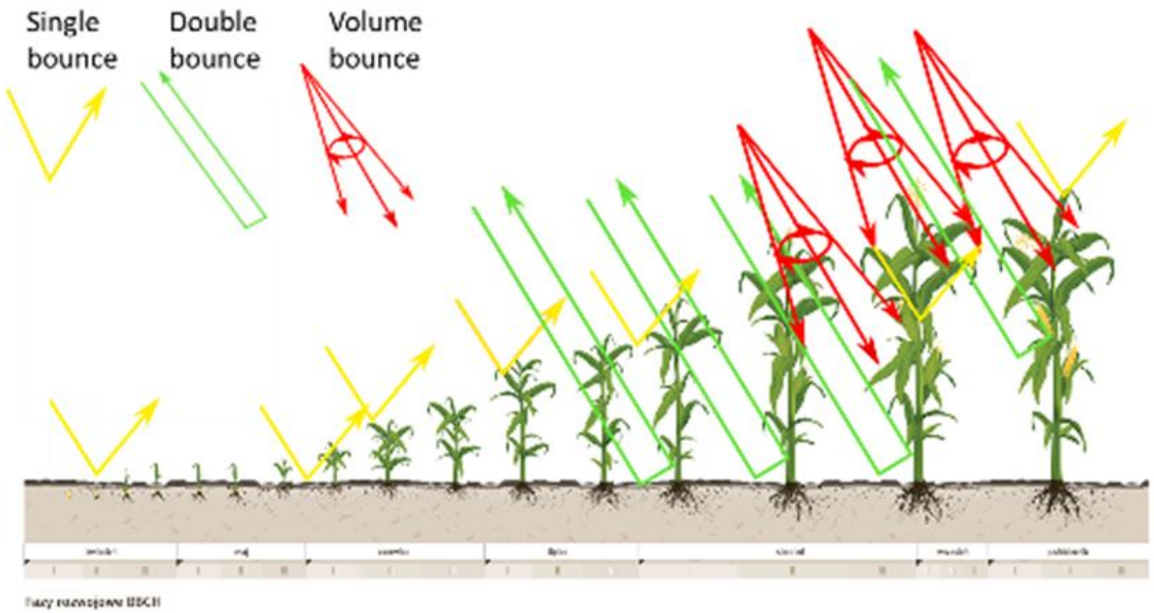
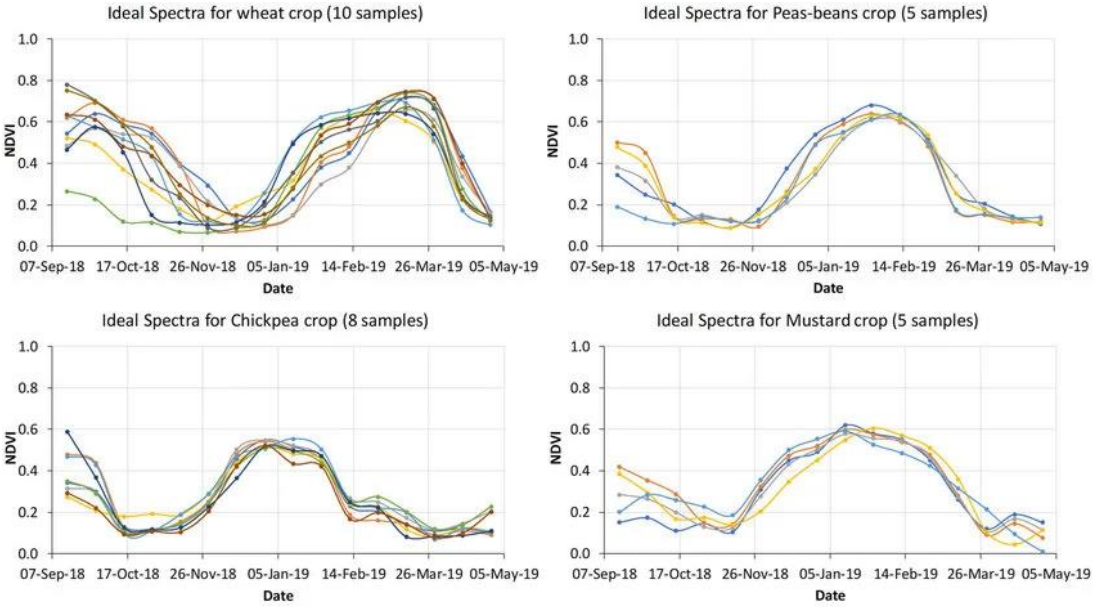
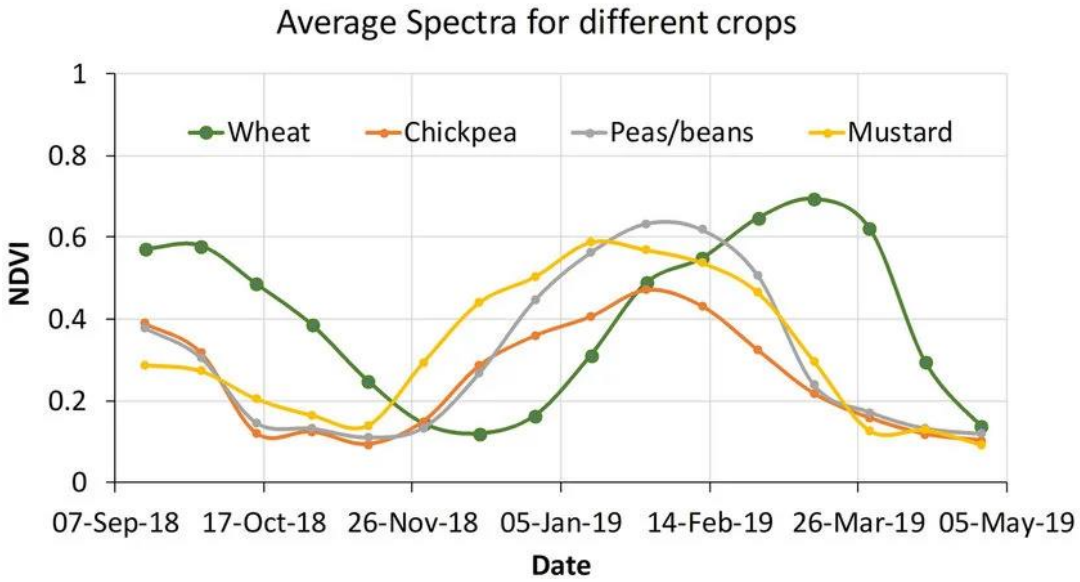


Figure 4.2: Example of differing temporal fingerprints (NDVI-based) from crops (Gumma et al)





Given infrastructure limitations and data-sharing constraints within WP7, classification was restricted to three aggregated classes: winter crops, spring crops and rapeseed. This limitation is pragmatic more than it is conceptual or methodological. Three practical constraints motivated this decision. First, privacy-sensitive training data and detailed parcel-level information cannot be shared across processing infrastructures. Second, the simplified pipeline relies on Sentinel-1 Ground Range Detected (GRD) products rather than Single Look Complex (SLC) data. Third, in line with the open source strategy of the EU, it uses open-source OBIA tools rather than proprietary eCognition software. This reduced signal complexity and the performance of the software limits the separability of more detailed crop classes. Future work will explicitly address these limitations.

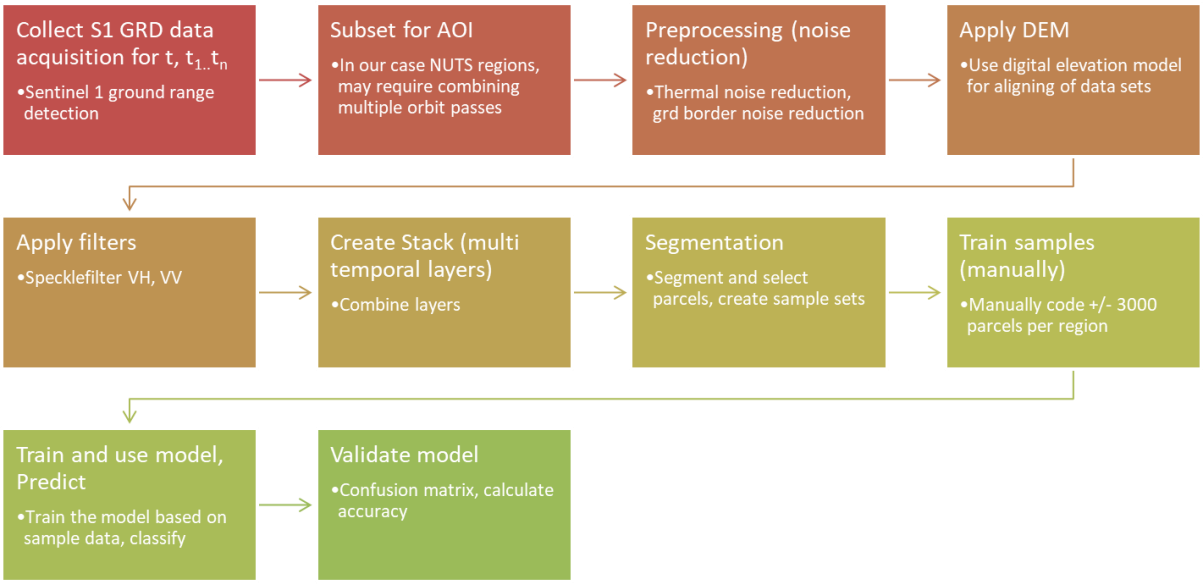
4.2 Pipeline

The overall processing workflow is summarised in Figure 4.3. The pipeline consists of radar data acquisition, preprocessing, multi-temporal stacking, segmentation, training data integration, classification and validation.

Sentinel-1 GRD imagery was collected for the period October 2024 to May 2025. This temporal window captures the most relevant phenological stages needed to distinguish winter crops, spring crops and rapeseed. Data were acquired for Ireland, Portugal, Austria and Poland, while the Netherlands case could not yet be executed due to unresolved domain-specific training data preparation issues.

Preprocessing was implemented using SNAP-based batch schemas, as discussed during the WP7 Warsaw meeting that was held in October 2025.

Figure 4.3: overview crop type classification algorithm based on collection of S1 GRD data



The steps include thermal noise removal, border noise correction, radiometric calibration, terrain correction using a digital elevation model, and multi-temporal coregistration. A multi-temporal speckle filtering approach was applied to reduce radar noise while preserving structural information. The resulting time series were stacked to create temporally aligned radar feature layers representing the full October–May observation window. These steps are demanding in terms of computing and the stack differs depending on the time-window that is relevant for a certain set of crops. It may be a good idea, though, that if we (or other researchers) establish that specific stacks are a base requirement for doing temporal 'fingerprint' research, ESA or Copernicus provides standard algorithms or preprocessed data that is readily available.

Segmentation was performed using open-source Object-Based Image Analysis (OBIA) tools, notably the Orfeo Toolbox. The objective was to delineate homogeneous radar-response units that approximate agricultural parcels. In practice, segmentation frequently produced sub-parcels smaller than the corresponding Land Parcel Identification System (LPIS) parcels.

However, inspection revealed that these subdivisions often correspond to real-world structural artefacts such as hedgerows, tree lines, irrigation features or small landscape elements. While this leads to geometric fragmentation, it may enhance spectral homogeneity within segments. Segmentation parameters can be adjusted to balance over-segmentation against parcel coherence.

Training data were derived from LPIS parcel centre points and manual interpretation of Sentinel-2 RGB mosaics. Sentinel-2 data were used exclusively for labelling and were not part of the predictive feature set. Class harmonisation was necessary to ensure comparability across countries, which resulted in aggregation into the three broad crop classes. Challenges were encountered in Portugal, where incomplete masking and LPIS registration inconsistencies led to confusion between arable land, grassland and permanent crops. These issues underline the importance of high-quality masks. The use of LPIS-based arable masks in combination with Copernicus High Resolution Grassland Layers (HRGL) is recommended to reduce label noise.

Classification was performed using a Random Forest algorithm. This choice reflects robustness, interpretability and moderate computational demand. As expected, the model shows a tendency to favour dominant classes and to underrepresent rare classes. In the present setup, this effect was limited but visible, particularly for crop types that were sparsely represented in the training dataset. Because the classification was restricted to three aggregated classes, imbalance effects were manageable.

4.3 Preparation of datasets

The preparation of training and testing datasets differed slightly across countries, reflecting variations in LPIS structure, data governance and available domain knowledge. Nevertheless, a common methodological framework was followed: selection of a NUTS region, harmonisation of crop classes into three aggregated categories (winter crops, spring crops and rapeseed), and verification of seasonal labels using Sentinel-2 imagery.

In Portugal, the focus region was Oeste e Vale do Tejo (PT11D). LPIS 2025 data were preprocessed to link parcels, land use and crop information layers. The Portuguese team provided labelled training samples derived from LPIS, while model training and prediction were executed by Statistics Poland. Evaluation was subsequently performed by analysing the spatial overlap between predicted segments and LPIS-registered arable land. Particular attention was given to distinguishing arable crops from grasslands, acknowledging that incomplete LPIS coverage of grasslands and crops limits the construction of a comprehensive agricultural mask.

In Ireland, LPIS 2024 data required a more extensive cleaning procedure due to overlapping polygons associated with multiple vegetation activities on single holdings.

Unique parcel identifiers were constructed and the dataset was intersected with the relevant NUTS boundary to ensure spatial consistency. From the cleaned dataset, only parcels belonging to selected EU aggregate crop classes were retained. A balanced training dataset of 900 parcels (300 per class) was randomly selected. Seasonal labels of the parcels of the

training dataset were verified manually using multi-date Sentinel-2 imagery, reducing label uncertainty. The remaining parcels (the parcels from the cleaned dataset that did not go into the training dataset) constituted the testing dataset.

In Austria, the preparation process was more straightforward. A NUTS 2 region was selected, and approximately 3,000 samples were manually classified based on Sentinel-2 imagery. These labelled samples were then used by Statistics Poland to train and execute the model on their infrastructure. Unlike Portugal and Ireland, Austria did not perform an extensive parcel-overlap analysis but provided a confusion matrix for evaluation.

Across all cases, harmonisation of class definitions was required to ensure comparability. The restriction to three broad crop classes was driven by both methodological and governance considerations. The simplified feature space derived from Sentinel-1 GRD data, combined with infrastructure and data-sharing constraints, necessitated aggregation of detailed LPIS crop codes into robust seasonal categories.

4.4 Code and infrastructure

An implementation of the preprocessing pipeline was developed on the CREODIAS cloud platform, targeting both Linux and Windows environments. While the Windows implementation proved manageable, the Linux implementation presented significant technical challenges. In particular, SNAP processing workflows did not integrate smoothly with the CREODIAS file system architecture, which differs from conventional local or network-mounted storage environments. Path handling, intermediate file management and batch execution required substantial manual adaptation.

Installation and configuration of the OBIA software components (notably Orfeo Toolbox) also proved complex. Dependency management and environment configuration demanded considerable effort before stable execution could be achieved. These experiences indicate that, although technically feasible, deployment of the pipeline in cloud environments requires detailed technical documentation and potentially containerised solutions to ensure reproducibility.

Despite these implementation efforts, the operational runs for Portugal, Austria and Poland were executed on the infrastructure of Statistics Poland. This decision was motivated by stability considerations and the availability of a functioning end-to-end pipeline. Training and prediction for Portugal and Austria were therefore performed centrally by Statistics Poland, based on nationally provided training samples.

The dual approach — experimental deployment on CREODIAS combined with operational execution on Polish infrastructure — provided valuable insights into portability constraints. It demonstrates that while the methodology itself is transferable, practical implementation requires careful alignment between software dependencies, file system structure and processing environment.

Future work will focus on improving portability, potentially through containerisation or partial migration to the Copernicus Data Space Ecosystem, in line with the broader WP7 objectives.

4.5 Evaluation of the results

Quantitative evaluation results are currently available for Austria (confusion matrix), while for Ireland and Portugal more detailed analysis of the results is available. For the Netherlands, focus was on segmentation results and a first attempt was made to do more detailed classifications based on the OBIA tooling. While segmentation results were encouraging, the classification still left much to improve.

Preliminary qualitative assessment indicates that the temporal radar signal allows for meaningful separation between winter crops and spring crops, while rapeseed displays a particularly distinctive backscatter trajectory during the observation period. Misclassification mainly occurs in regions where masking quality is suboptimal or where parcel heterogeneity is high.

Beyond parcel-level validation, an additional validation layer is planned using official agricultural statistics at NUTS level. Aggregated model predictions will be compared against harvest statistics to assess consistency between remote sensing–based classification and statistical reporting. This cross-validation approach will help evaluate whether parcel-level classification errors propagate to regional agricultural indicators.

4.5.1 Portugal case study – Oeste e Vale do Tejo (PT11D)

The Portuguese case study focuses on the NUTS3 region Oeste e Vale do Tejo (PT11D). Training samples derived from LPIS were provided by Statistics Portugal (INE), while model training and inference were executed by Statistics Poland on their infrastructure. The Portuguese team did not retrain the algorithm locally; their analysis focused on evaluating the spatial overlap between the model predictions and the LPIS registry.

According to the LPIS 2025 dataset for the region, 83,718 parcels were registered, with 79,567 hectares classified as arable crops (see Figure 4.4). The Polish algorithm classified a total area of 98,009 hectares as belonging to one of the crop classes (winter crops, spring crops), see Figure 4.5. This results in a difference of approximately 18,442 hectares between the total predicted area and the officially registered arable area.

The difference between the total predicted area and the LPIS-registered arable area can be explained by a combination of coverage limitations and classification scope. First, the algorithm does not fully capture all LPIS-registered crop areas. This is illustrated by crop-specific comparisons between LPIS and model output. For example, for wheat, approximately 83% of the LPIS-registered area is detected by the algorithm, while for oats this drops to around 46%. These differences indicate that part of the discrepancy is due to *incomplete coverage (false negatives)*, where the algorithm does not assign a crop label to every area from LPIS. Early in the project, this effect was made worse by the use of only single-crop parcels for training and validation, due to limited access to ‘within parcel’ crop information. This restriction has since been resolved, allowing for more accurate representation of mixed parcels.

Second, the algorithm also predicts crop areas outside the LPIS-registered arable land, contributing to the observed overestimation. These *false positives* are primarily linked to

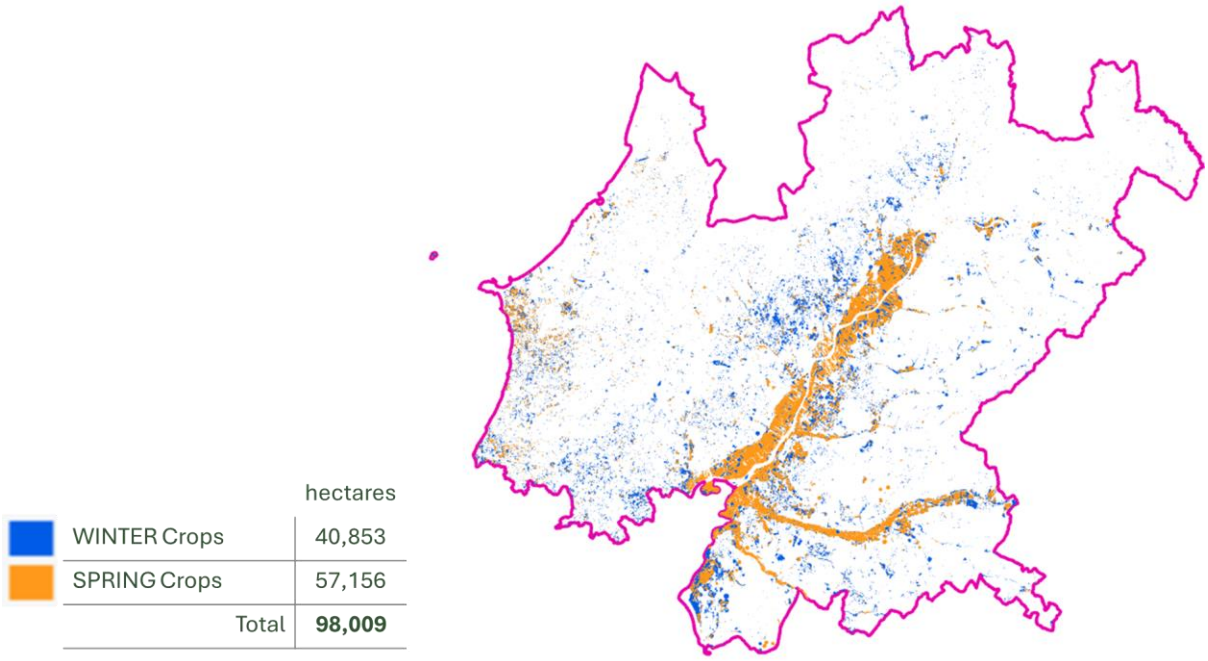
incomplete masking of non-arable land, particularly grasslands and other land-use classes that are not fully captured in LPIS. As a result, areas that are not registered as arable crops may still exhibit radar signatures similar to seasonal crops and are therefore classified as such. Within the area that does overlap with LPIS, however, the classification performance is strong. For instance, in the case of wheat, 87% of the detected area is correctly identified as winter crop, while for oats approximately 85% of the detected area is correctly classified. This indicates that, although spatial coverage is not complete, the model performs reliably in distinguishing between winter and spring crop types once relevant areas are identified.

Figure 4.4: hectares of LPIS based land use classes in Oeste e Vale do Tejo (right) and small example area (right)

	hectares
Permanent Crops	58,758
Arable Crops	79,567
Forest	12,833
Agroforestry	65,503
Grassland	65,824
Others	1,707
Total	284,231

However, total area comparison alone is not sufficient. A spatial overlay analysis was therefore performed. Of the 79,567 hectares of LPIS-registered arable crops, 71,478 hectares overlap with the area classified by the algorithm. This corresponds to approximately 74% coverage of registered arable land. Conversely, 27% of the area classified by the algorithm does not overlap with LPIS arable land and therefore requires further investigation

Figure 4.5: Predicted area of crop type (hectares) and corresponding map



When the overlap between LPIS classes and predicted parcels are considered we see that mainly parcels that are registered as permanent crops and grasslands are predicted to be either winter or spring crops (Table 4.1 and 4.2).

Table 4.1: Within LPIS 'arable crops' areas (in hectares): prediction (rows) vs LPIS classes

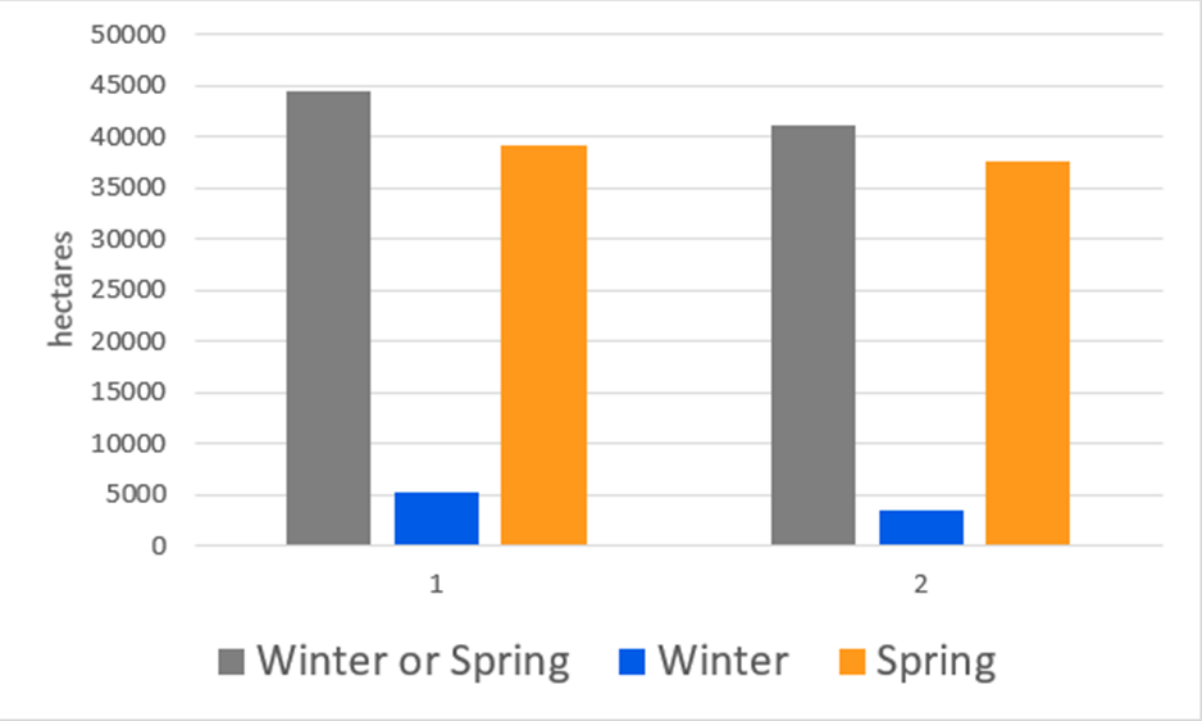
Type of crops	Arable crops	Permanent crops	Forest	Agro forestry	Grassland	No class	Total overlapped area
Winter crops	16,679	2,683	172	131	5,390	144	**25,199**
Spring crops	41,946	2,440	67	27	1,465	334	**46,279**
Total	**58.625**	**5,124**	**239**	**158**	**6,855**	**478**	**71,478**

Table 4.2: Within LPIS 'arable crops' areas (in hectares): overlap with algorithm

Type of crops	LPIS	Overlap Algorithm	Match & no match
Winter & spring crops	44,442	41,179	93%
Winter crops	5,248	3,559	68%
Spring crops	39,194	37,620	96%

A more detailed examination of the overlapping area showed encouraging results. Within the subset of winter and spring crops, the algorithm covers 44,442 hectares. For these categories, 93% of the overlapping area is correctly classified within the winter/spring distinction. Disaggregated results show that 96% of spring crops are successfully identified, while 68% of winter crops are classified correctly (Figure 4.6).

Figure 4.6: within LPIS arable land class prediction match

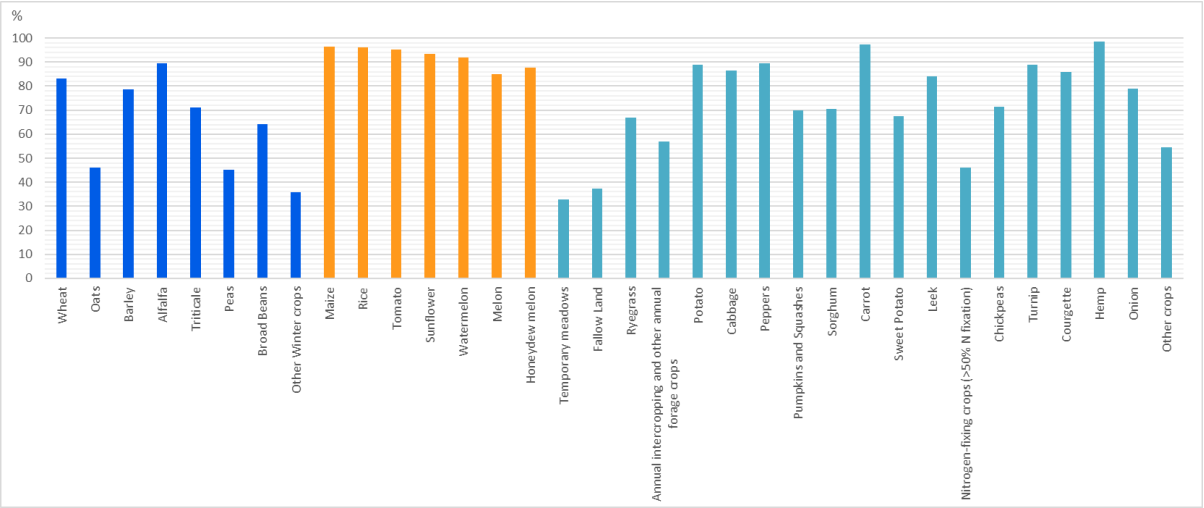


The overall accuracy for distinguishing winter versus spring crops in the analysed region is reported as 89%.

The research shows that 82% of the overlapping area corresponds to LPIS arable crops, indicating that the algorithm mostly identifies agricultural land when spatial overlap exists. Crop-specific charts (mapped to winter and spring predictions) confirm that the most representative crop types are classified consistently, including oats, which initially appeared problematic but were correctly captured after label correction (Figure 4.7 and 4.8). Note that the 'peas' class only covers a very small fraction of the total arable land, which we consider the main reason for the lower accuracy scores of this class.

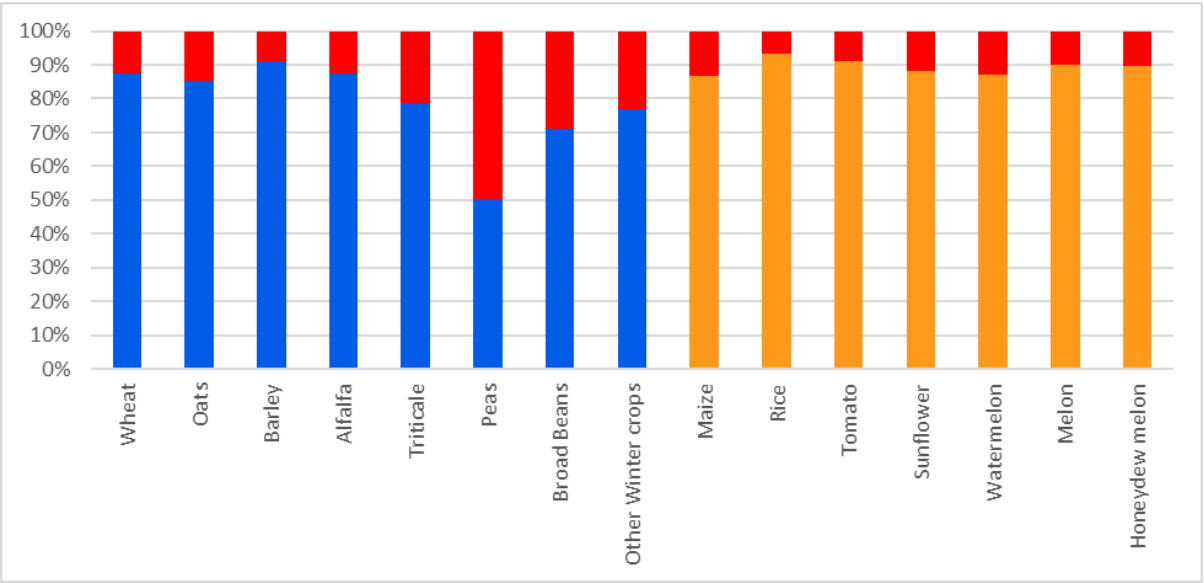
In Figure 4.7 the LPIS area by crop as classified by the algorithm is shown. The area for each crop is the actual figure; we have access to the specific area of each crop within a parcel, even when a parcel contains multiple crops. For example, for wheat, there are 1,764 hectares, but the algorithm classified only 1,470, meaning 83% of the area was classified.

Figure 4.7: LPIS area by crop as classified by the algorithm



In figure 4.8 the area classified by the algorithm and its performance in differentiating between winter and spring crops is shown. For example, for wheat, 83% of the LPIS area was classified by the algorithm. Of this 83%, 87% was correctly identified as a winter crop.

Figure 4.8: Area classified by the algorithm and its performance



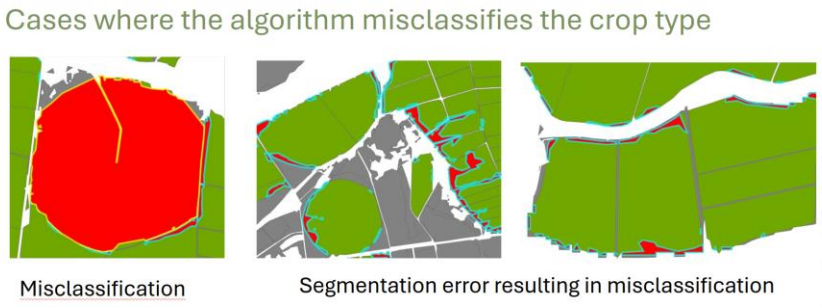
The observed differences between the LPIS-registered arable area and the area classified by the algorithm can be explained by a combination of incomplete spatial coverage and classification outside the target domain. First, the algorithm does not fully capture all LPIS-registered crop areas. Crop-specific comparisons indicate that, for example, approximately 83% of the LPIS-registered wheat area is detected by the algorithm, while for oats this is around 46%. This indicates that part of the discrepancy is due to incomplete coverage (false

negatives), where not all arable land parcels are assigned a crop label. Earlier in the project, this effect was reinforced by the restriction to parcels with a single crop type, due to limited availability of within-parcel crop information. This limitation has since been resolved.

Second, the algorithm predicts crop classes outside LPIS-registered arable land. The overlap analysis shows that a substantial share of predicted winter and spring crops falls within areas classified in LPIS as grassland or permanent crops. This is primarily due to incomplete masking of non-arable land. In Portugal, not all grasslands are included in LPIS, as registration is linked to subsidy eligibility. As a result, areas with radar signatures similar to seasonal crops may be classified as such, leading to overestimation (false positives).

Within the area that does overlap with LPIS arable land, the classification performance is strong. The model correctly distinguishes between winter and spring crops for 93% of the overlapping area, with 96% accuracy for spring crops and 68% for winter crops. This indicates that, once relevant agricultural areas are identified, the model performs reliably in capturing seasonal crop dynamics. Remaining misclassification patterns are partly related to segmentation effects, where small segments within parcels may be assigned different labels due to local heterogeneity or boundary artefacts.

Figure 4.9: Classification errors cases (often small subsets of parcels)



Average land-use area of arable crops per parcel = **19.500 m²**

Polygons Size Classes (ha)

Class size	MATCH	NO MATCH
<100 m²	15	38
<1.000 m²	79	226
< 10.000 m²	1,757	836
<100.000 m²	19,584	2,280
>100.000 m²	15,231	1,134

The Portuguese team also explored alternative aggregation approaches. One promising strategy discussed during the meeting is to select, within each parcel, the segment with the highest proportional coverage and highest prediction confidence as the representative crop label. Such a parcel-level aggregation rule may improve alignment between segment-based classification and parcel-based statistical reporting. Figure 4.9 shows that the smaller segments that are misclassified tend to be part of a bigger parcel (with some notable exceptions). Dissolving these smaller sections into bigger (LPIS based) parcels may be a way to accommodate for these ‘small-scale’ errors.

Finally, the potential use of the Copernicus High Resolution Layer (HRL) Grassland product as an auxiliary mask was discussed. Its suitability for Portuguese conditions requires further evaluation, but it may provide a consistent, non-privacy-sensitive complement to LPIS data for distinguishing grasslands from winter crops.

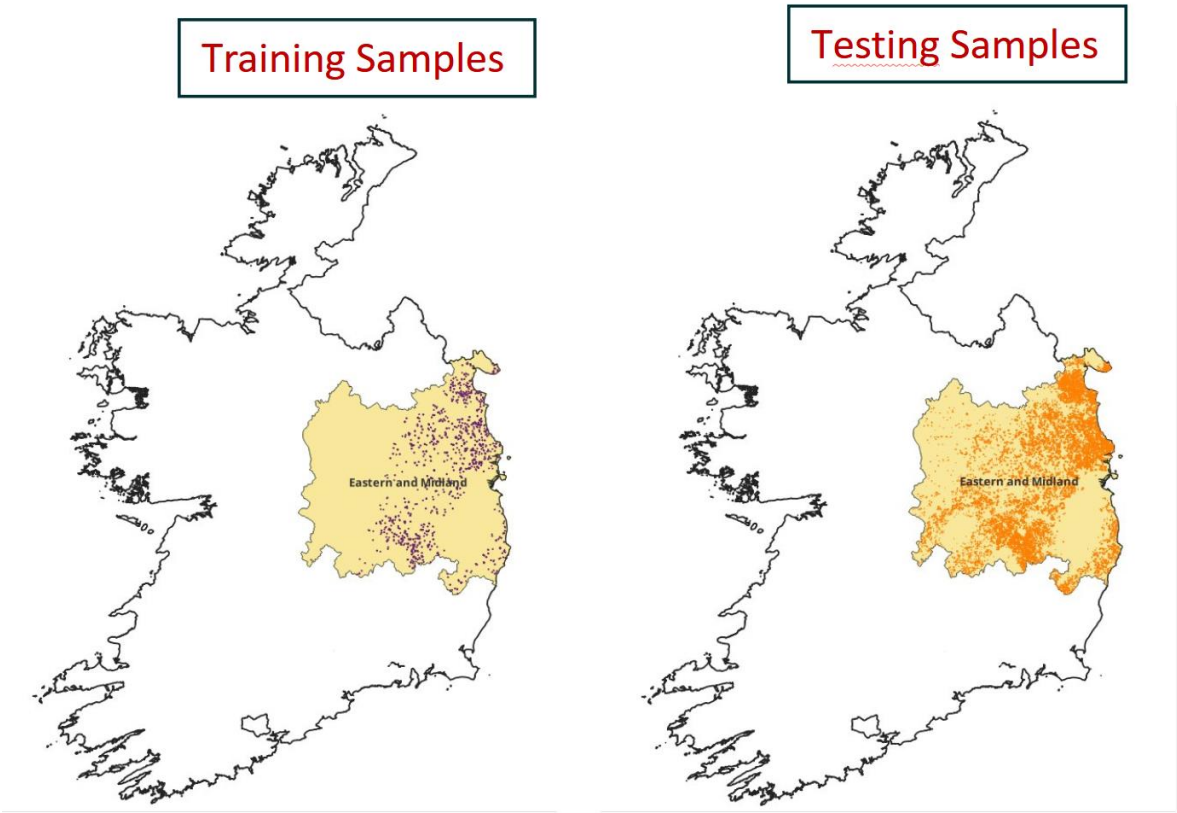
4.5.2 Ireland case study

The Irish case study was conducted using LPIS 2024 data and focused on a selected NUTS 2 region (Eastern and Midland, see Figure 4.10). The dataset initially consisted of 1,547,508 polygons.

Because certain parcels were subdivided into overlapping polygons due to multiple vegetation-related land use activities on a single holding, a cleaning procedure was required. Unique parcel identifiers were constructed, overlaps were removed, and the LPIS dataset was intersected with the NUTS boundary to ensure geographic consistency

The cleaned dataset consisted of 143,203 distinct-unique crop type polygons and 4,185 non-distinct/non-unique crop type polygons within the selected region (the latter were removed from the working set).

Figure 4.10: splitting the training and testing samples for Ireland



From this dataset, only parcels classified under relevant EU aggregate categories (including cereals, pulses, industrial crops, leguminous plants harvested green, root crops and vegetables) were retained. For the final modelling exercise, the focus was restricted to parcels classified as spring crops, winter crops and winter rapeseed. Crops planted in both seasons were excluded to reduce ambiguity.

A clear separation between training and testing datasets was established. A total of 900 training samples were randomly selected, comprising 300 samples for each class (spring crops, winter crops and winter rapeseed). The seasonal status of these samples was verified using

Sentinel-2 imagery at multiple time points (December, January, April and May), ensuring correct seasonal labelling

The remaining parcels formed the testing dataset, consisting of approximately 21,068 parcels after filtering for the relevant crop categories.

The Irish team applied a parcel-oriented validation logic similar to that what was suggested as a follow up to the Portugal case. Rather than evaluating segment-level predictions in isolation, predictions were aggregated at parcel level. In cases where multiple segments were produced within a single parcel, the segment with the largest proportional coverage and strongest prediction signal was used to assign the parcel label. This approach reduces the impact of segmentation artefacts and aligns the classification output more closely with administrative parcel structures.

Quantitative evaluation of the Ireland case

The Irish evaluation provides results at three complementary levels: parcel-level accuracy, area-weighted confusion analysis and crop-level performance within aggregated classes.

At parcel level (Table 4.3 and 4.4), the testing dataset consisted of 21,068 parcels. Of these, 20,546 parcels were correctly classified, resulting in an overall accuracy of 97.5%. Class-specific accuracies were 98.1% for spring crops, 96.5% for winter crops and 98.7% for winter rapeseed. These results indicate consistent performance across all three aggregated classes.

Table 4.3: parcel-level analysis of prediction results (number of polygons)

Class	TRUE	FALSE	Total_LPIS_Polygons
1-Spring	11578	226	11804
2-Winter	7602	278	7880
3- Winter Rapeseed	1366	18	1384
Grand Total	**20546**	**522**	**21068**

Table 4.4: parcel-level analysis of prediction results (percentages)

Class	TRUE	FALSE	Total_LPIS_Polygons
1-Spring	98.1%	1.9%	100.0%
2-Winter	96.5%	3.5%	100.0%
3- Winter Rapeseed	98.7%	1.3%	100.0%
Grand Total	**97.5%**	**2.5%**	**100.0%**

To account for differences in parcel size, an area-weighted confusion matrix was computed using LPIS parcel areas expressed in square metres (Table 4.5 and 4.6). For LPIS-classified spring crops, 95.6% of the total area was correctly predicted as spring, with 3.6% misclassified as winter crops. For winter crops, 92.7% of the area was correctly identified, with 6.0% misclassified as spring crops. Winter rapeseed showed the highest separability, with 97.3% of the area correctly classified. The slightly higher confusion between winter and spring crops is

consistent with partial overlap in phenological development during the October–May observation window.

Table 4.5: Area-weighted confusion matrices (hectares)

Class	LPIS_areaha	RF_Areasha_Class1	RF_Areaha_Class2	RF_Areaha_Class3	RF_Areaha
1-Spring	60.105	57.468	2.186	450	60.105
2-Winter	48.912	2.942	45.344	626	48.912
3- Winter Rapeseed	9.281	111	144	9.026	9.281
Grand Total	118.297	60.521	47.674	10.102	118.297

Table 4.6: Area-weighted confusion matrices (percentages)

Class	LPIS_areaha	RF_Areasqm_Class1	RF_Areasqm_Class2	RF_Areasqm_Class3	RF_Areasqm
1-Spring	60.105	95,61%	3,64%	0,75%	100,00%
2-Winter	48.912	6,02%	92,71%	1,28%	100,00%
3- Winter Rapeseed	9.281	1,20%	1,55%	97,26%	100,00%
Grand Total	118.297				

A more detailed crop-level analysis (Table 4.7 and 4.8) within the aggregated classes confirms the robustness of the seasonal classification. Spring barley (98.8%), spring beans (98.3%), spring oats (95.8%) and spring wheat (92.5%) all show high parcel-level accuracy. Similarly, winter barley (98.0%), winter wheat (96.0%) and winter oats (94.3%) are classified reliably. Winter oilseed rape achieves 98.7% correct classification.

Table 4.7: Prediction performance by crop type (from LPIS data) (number of polygons)

Crops	Class	TRUE	FALSE	Total_LPIS_Polygons
Barley - Spring	1-Spring	8253	100	8353
Beans - Spring	1-Spring	1389	24	1413
Oats - Spring	1-Spring	1332	58	1390
Oilseed Rape – Spring	1-Spring	91	8	99
Peas	1-Spring	91	2	93
Wheat - Spring	1-Spring	422	34	456
Barley - Winter	2-Winter	3381	70	3451
Beans - Winter	2-Winter	8	22	30
Oats - Winter	2-Winter	517	31	548
Wheat - Winter	2-Winter	3696	155	3851
Oilseed Rape – Winter	3- Winter Rapeseed	1366	18	1384
Grand Total	**Grand Total**	**20546**	**522**	**21068**

Table 4.8: Prediction performance by crop type (from LPIS data) (percentages)

Crops	Class	TRUE	FALSE	Total_LPIS_Polygons
Barley - Spring	1-Spring	98.8%	1.2%	100.0%
Beans - Spring	1-Spring	98.3%	1.7%	100.0%
Oats - Spring	1-Spring	95.8%	4.2%	100.0%
Oilseed Rape – Spring	1-Spring	91.9%	8.1%	100.0%
Peas	1-Spring	97.8%	2.2%	100.0%
Wheat - Spring	1-Spring	92.5%	7.5%	100.0%
Barley - Winter	2-Winter	98.0%	2.0%	100.0%
Beans - Winter	2-Winter	26.7%	73.3%	100.0%
Oats - Winter	2-Winter	94.3%	5.7%	100.0%
Wheat - Winter	2-Winter	96.0%	4.0%	100.0%
Oilseed Rape – Winter	3- Winter Rapeseed	98.7%	1.3%	100.0%

An exception is winter beans, where only 26.7% of parcels were correctly classified. However, this class consists of only 30 parcels in the testing dataset. The instability therefore reflects extreme class imbalance rather than systematic model failure. Such behaviour is consistent with known properties of Random Forest classifiers when trained on limited samples for minority classes.

The strong consistency between parcel-level and area-weighted metrics indicates that performance is not driven by a large number of small parcels but reflects that real phenomena can be extracted from timeseries radar data. These results confirm that the October–May Sentinel-1 GRD time series provides sufficient discriminatory power to distinguish spring crops, winter crops and winter rapeseed under Irish agricultural conditions.

4.5.3 Netherlands case study

In the Netherlands case study, the primary focus was placed on analysing segmentation behaviour and evaluating the structural properties of the object-based workflow prior to full classification.

An analysis of the number of segments per parcel showed that although segmentation frequently subdivides parcels, the majority of agricultural area is represented by parcels containing a limited number of segments. Parcels with up to five segments account for more than half of the total analysed area, while highly fragmented parcels (more than ten segments) contribute only marginal additional surface area. This indicates that segmentation intensity remains manageable and supports the feasibility of parcel-level aggregation strategies.

Figure 4.11: Cropmapper segments (in red) and official parcel boundaries (blue).

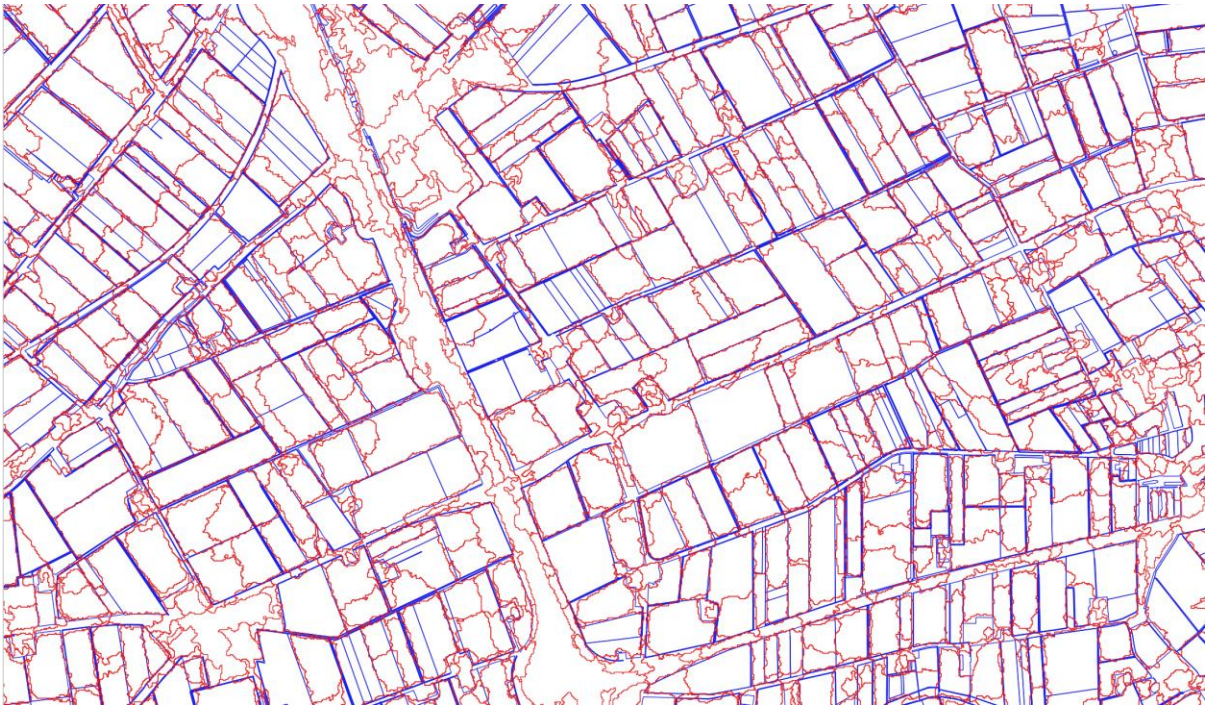
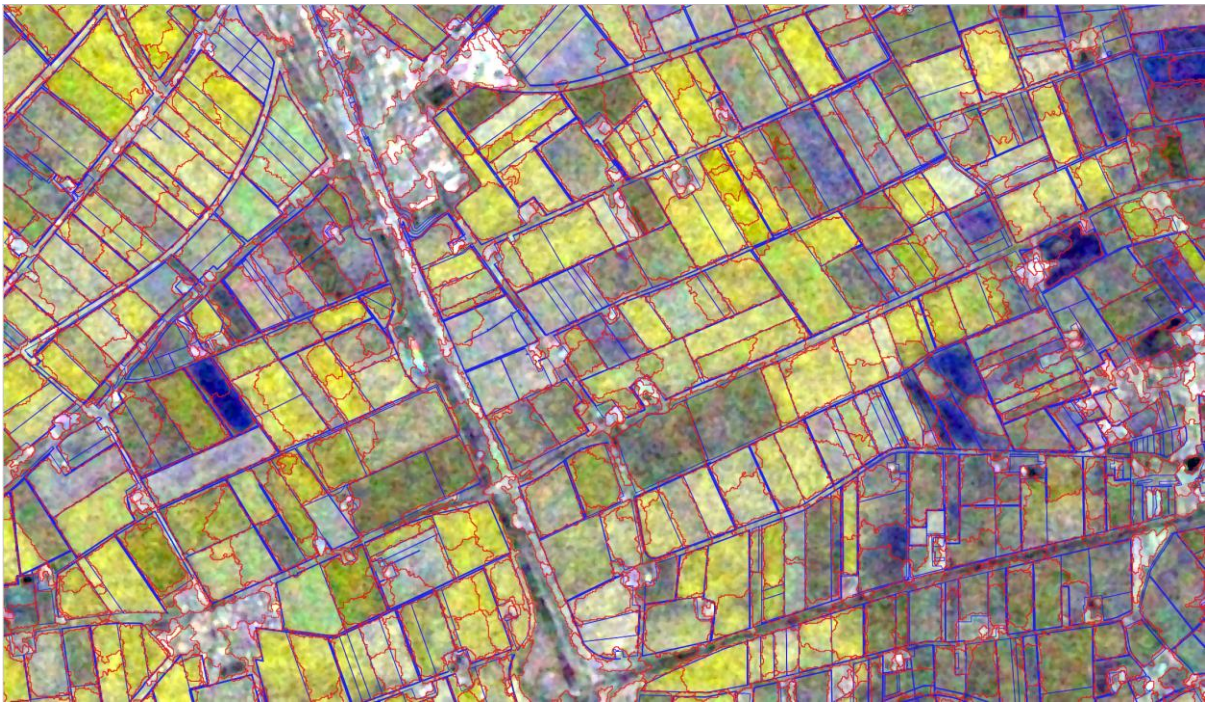


Figure 4.12: Cropmapper segments (in red) and official boundaries (blue), with false colour rendition of GRD signal data for one date.



A complementary analysis examined the number of official LPIS parcels intersected by each algorithm-generated segment. The results show that many segments intersect more than one administrative parcel, with two parcels per segment being the most frequent case. A long tail distribution is observed, with some segments intersecting a substantially higher number of

parcels. This confirms that the segmentation algorithm operates independently of administrative parcel geometry and instead follows radar homogeneity patterns.

These findings highlight the necessity of applying parcel-level aggregation rules when linking segmentation outputs to official statistical units. Approaches such as selecting the dominant segment (largest area or highest prediction confidence) per parcel are statistically defensible, given the observed fragmentation structure.

In addition, an exploratory classification experiment was conducted using 276 detailed crop classes. The resulting confusion matrix exhibited substantial dispersion and sparsity, indicating that the simplified GRD-based feature space is insufficient to reliably discriminate such a high number of classes. The experiment demonstrates that fine-grained crop classification requires either richer input data (e.g., Sentinel-1 SLC products with phase information), more advanced modelling approaches, or hierarchical rule-based refinement mechanisms.

The Netherlands case therefore provided important methodological insights: while broad seasonal classification is feasible and robust, fine-grained crop typologies require enhanced preprocessing and modelling strategies beyond the current simplified pipeline.

4.5.4 Austria case study

The Austrian case study focused on evaluating the transferability of the simplified crop type pipeline under minimal preprocessing complexity. A NUTS2 region was selected, and approximately 3,000 samples were manually classified using multi-date Sentinel-2 imagery to determine seasonal crop type (spring crops, winter crops and winter rapeseed).

Unlike Ireland and Portugal, no extensive LPIS overlap analysis was conducted. Instead, the manually labelled samples were provided to Statistics Poland, where model training and inference were executed on the established Polish infrastructure. The Austrian team subsequently evaluated the results using a confusion matrix.

This setup differs from the Irish and Portuguese approaches in two important ways. First, training samples were manually generated rather than derived directly from LPIS attribute tables. Second, the evaluation focused on classification accuracy at sample level rather than on spatial overlap with registered arable land.

Despite these methodological differences, the Austrian results confirm the feasibility of distinguishing winter crops, spring crops and rapeseed using the October–May Sentinel-1 GRD time series. The confusion matrix indicates that most misclassifications occur between winter and spring crops, while rapeseed remains clearly separable — a pattern consistent with the Irish and Portuguese findings.

The Austrian case therefore provides independent evidence that the simplified radar-based pipeline is transferable across different agricultural contexts and data preparation strategies. Importantly, the model was trained and executed externally (on Polish infrastructure), demonstrating that country-specific implementation of the full preprocessing chain is not a prerequisite for obtaining meaningful classification results, provided that harmonised training samples are available.

4.6 Evaluation of the generalisation

The intermediate results across Portugal, Ireland, Austria and the Netherlands provide a structured basis for assessing the generalisability of the simplified crop type pipeline. The evidence suggests that generalisation is technically feasible, but conditional upon data preparation quality, masking consistency and harmonised training protocols.

First, the seasonal distinction between winter crops, spring crops and rapeseed appears robust across countries. In Ireland, both parcel-level and area-weighted metrics demonstrate high classification accuracy, with consistent separability of rapeseed and limited confusion between winter and spring crops. In Portugal, although overall spatial overlap with LPIS was affected by masking limitations and incomplete grassland registration, the classification performance (for two classes) within overlapping arable areas was strong. Austria provided independent validation under a different training preparation strategy, confirming that the seasonal radar signal remains discriminatory even when training data are manually derived rather than directly linked to LPIS attributes.

Second, the Netherlands segmentation study demonstrated that the object-based preprocessing component behaves in a structurally stable manner. Although segmentation frequently subdivides parcels, most agricultural area is represented by parcels containing a moderate number of segments. This supports the statistical defensibility of parcel-level aggregation strategies and indicates that segmentation artefacts are unlikely to introduce large-scale bias.

At the same time, the exploratory 276-class experiment in the Netherlands clearly showed that the simplified GRD-based feature space does not support fine-grained crop classification at high typological resolution. This confirms that the current setup is suitable for broad seasonal categories but not for detailed crop taxonomies. Limitations are primarily linked to the use of GRD amplitude data, the absence of phase information available in SLC products, and the use of Random Forest without hierarchical rule-based refinement.

Taken together, the results indicate that generalisation of the seasonal classification concept is feasible across space, provided that:

1. Temporal windows are harmonised,
2. Adequate agricultural masking is applied,
3. Training data are balanced and verified,
4. Parcel-level aggregation rules are consistently defined.

However, generalisation in the sense of a fully transferable, country-independent production model has not yet been conclusively demonstrated. Two further experiments are planned to assess this more rigorously.

First, a model trained exclusively on Polish data will be applied to another country (the Netherlands) without retraining. This will test cross-country transferability of model parameters and feature importance structures. Second, a new model will be trained using combined training data from Portugal, Ireland, Austria, the Netherlands and Poland. This joint

model will allow evaluation of whether a harmonised, multi-country training dataset improves robustness and reduces country-specific bias.

These forthcoming experiments will provide stronger evidence regarding whether generalisation is primarily driven by the radar signal itself or by country-specific training characteristics. They will also help determine whether a single European-scale model is feasible or whether regionally adapted models remain necessary.

At this intermediate stage, the evidence supports cautious optimism: the seasonal radar signature is consistent across diverse agricultural systems, and the simplified pipeline can be transferred between countries under defined methodological conditions. Further work will determine the extent to which this transferability can be operationalised within an integrated European statistical framework.

4.7 Next steps

The intermediate results demonstrate that seasonal crop classification based on Sentinel-1 GRD time series is technically feasible and transferable across countries, provided that preprocessing, masking and training protocols are harmonised. The next phase of the work will therefore move beyond proof of concept and focus on generalisation, increasing classification detail and aligning the methodology more closely with statistical production.

A first priority is to further assess cross-country transferability. Two structured experiments are planned. In the first, a model trained exclusively on Polish data will be applied to the Netherlands without retraining. This will test whether model parameters and learned feature structures are robust across agricultural systems. In the second, a harmonised model will be trained using combined training data from Portugal, Ireland, Austria, the Netherlands and Poland. This joint model will allow evaluation of whether a multi-country training base improves robustness and reduces country-specific bias. Together, these experiments will clarify whether a single European-scale seasonal model is achievable or whether regionally adapted models remain necessary. The same should be done for multi-year transferability (although from experience we believe that retraining will be required).

At the same time, attention will shift towards enabling more detailed crop classification. The exploratory 276-class experiment clearly demonstrated that the current simplified configuration — based on Sentinel-1 GRD amplitude features and Random Forest classification — does not support fine-grained typologies. Achieving higher thematic resolution will require methodological upgrades. In particular, preprocessing of Sentinel-1 Single Look Complex (SLC) data will be explored. SLC products retain phase information and enable the derivation of coherence and more advanced polarimetric features, which may enhance separability between crop types with similar amplitude signatures. In addition, greater integration of agronomic domain knowledge will be necessary. This includes explicit modelling of crop-specific phenological trajectories, seasonal timing differences and crop succession patterns within parcels. Such knowledge-driven refinement is likely essential for distinguishing crops that follow similar seasonal amplitude patterns but differ in growth dynamics or rotation structure.

These methodological enhancements will inevitably increase computational demands. Processing SLC data and generating coherence-based features is substantially more resource-intensive than working with GRD amplitude stacks. Future work will therefore also focus on improving computational efficiency, optimising cloud-based workflows and exploring containerised deployment solutions to ensure portability and scalability.

From a statistical perspective, it may not be necessary to process all available Sentinel-1 data exhaustively in order to produce reliable agricultural indicators. While the current work emphasises full spatial coverage, future research will examine whether statistically designed sampling strategies could achieve comparable aggregate estimates with reduced computational burden. Potential approaches include stratified sampling of parcels, targeted temporal subsampling that preserves phenological signal, or hybrid designs combining full segmentation with sample-based classification. Such approaches would align the methodology more closely with established statistical principles, balancing precision, cost and feasibility.

Validation procedures will also be strengthened. In addition to parcel-level confusion matrices and LPIS overlap analyses, future work will include comparison of aggregated predictions with NUTS-level harvest statistics, assessment of bias and variance in regional crop area estimates and systematic sensitivity analysis of segmentation and masking parameters. This will improve the linkage between earth observation–based classification and official agricultural statistics.

Overall, the next two years will focus on transitioning from a transferable seasonal prototype to a more detailed, scalable and statistically grounded crop classification framework. The results so far indicate that the seasonal radar signature is stable across diverse European agricultural contexts. The forthcoming work will determine to what extent this signal can be exploited for finer thematic resolution and operational integration within the European Statistical System.

Literature

d'Andrimont, R., et al. (2020), *Crop type mapping using Sentinel-1 and Sentinel-2 time series: A review of current practices and challenges*, Remote Sensing, 12(23), 3855. https://doi.org/10.3390/rs12233855

Gumma, M. K., Tummala, K., Dixit, S., Collivignarelli, F., Holecz, F., Kolli, N., & Whitbread, A. (2020), *Crop type identification and spatial mapping using Sentinel-2 satellite data with focus on field-level information*, Geocarto International, 37(1), pp. 1–14. https://doi.org/10.1080/10106049.2020.1805029

Van Tricht, K., et al. (2023), *WorldCereal: A dynamic open system for global cropland and crop type mapping*, Remote Sensing of Environment, 290, 113513. https://doi.org/10.1016/j.rse.2023.113513

Veloso, A., et al. (2017), *Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications*, Remote Sensing of Environment, 199, pp. 415–426. https://doi.org/10.1016/j.rse.2017.07.015

Woźniak, E., Rybicki, M., Kofman, W., Aleksandrowicz, S., Wojtkowski, C., Lewiński, S., Bojanowski, J., Musiał, J., Milewski, T., Ślesiński, P., & Łączyński, A. (2022), *Multi-temporal phenological indices derived from time series Sentinel-1 images to country-wide crop classification*, International Journal of Applied Earth Observation and Geoinformation, 107, 102683. https://doi.org/10.1016/j.jag.2022.102683

Zhang, C., Kerner, H., Wang, S., Hao, P., Li, Z., Hunt, K. A., Abernethy, J., Zhao, H., Gao, F., Di, L., Guo, C., Liu, Z., Yang, Z., Mueller, R., Boryan, C., Chen, Q., Beeson, P. C., Zhang, H. K., & Shen, Y. (2025), *Remote sensing for crop mapping: A perspective on current and future crop-specific land cover data products*, Remote Sensing of Environment, 330, 114995. https://doi.org/10.1016/j.rse.2025.114995