ESA Sen4Stat

Sentinels for Agricultural Statistics

D17.0 v1.1 – FR

“Final Report”

Issue/Rev: 1.1

Date Issued: 28-03-25

Milestone: 6

Submitted to European Space Agency

Consortium Partners

Participant Organisation Name	Acronym	City, Country
Université catholique de Louvain	UCLouvain	Louvain-la-Neuve, Belgium
CS Romania	CS RO	Craiova, Romania
Systèmes d'Information à Référence Spatiale SAS	SIRS	Villeneuve d'Ascq, France
Universidad Polytecnica de Madrid	UPM	Madrid, Spain

Contact

Université catholique de Louvain – Earth and Life Institute
Place de l'Université, 1 – B-1348 Louvain-la-Neuve – Belgium
Email : Sophie.Bontemps@uclouvain.be
Internet : https://uclouvain.be/en/research-institutes/eli/elie

Disclaimer

Document sheet

Authors and Distribution

Authors	Sophie BONTEMPS, Nicolas DEFFENSE, Boris NORGAARD, Pierre DEFOURNY
Distribution	ESA - Zoltan SZANTOI

Document Status

Issue/Rev.	Date	Reason
0.1	01/02/2024	Initial version of the deliverable
1.0	30/10/2024	First version submitted to ESA
1.1	28/03/2024	Updated first version based on ESA RIDs

Detailed record sheet

From version 1.0 to 1.1

Comment	Section	Problem description	Change
ESA RID	5.1.1	Blurry text in figure	Figures 5-5, 5-7, 5-8, 5-10, 5-14, 5-15: quality of the figure improved
ESA RID	5.2.1	Blurry text in figure	Figures 5-28, 5-31: quality of the figure improved
ESA RID	5.2.2.1	Blurry text in figure	Figures 5-34 and 5-35: quality of the figure improved

Table of contents

List of figures

List of tables

1 Introduction

1.1 Purpose and scope

This document is the Final Report (FR) of the Sentinels for Agricultural Statistics (Sen4Stat) project funded by the European Space Agency (ESA).

The overall objective for the Sen4Stat project is to facilitate the uptake of Earth Observation (EO) information in the National Statistical Offices (NSO) supporting the agricultural statistics. Special attention shall be given to develop and demonstrate EO products and best practices for agriculture monitoring relevant for Sustainable Development Goals (SDG) reporting and monitoring their progress at national scale

The FR is the key outputs of the Task 6 (WP 6000) of the Sen4Stat project, named "Conclusions and Recommendations" (Figure 1-1). It aims at summarizing the data set, algorithms, products and final service achieved within Sen4Stat project.

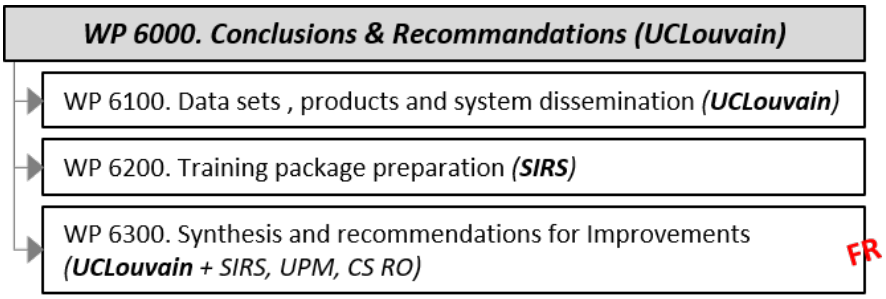


Figure 1-1. Organization of the Task 5 activities (from [AD.2])

1.2 Structure of the document

After this introduction, this document contains ***5*** sections:

- Section 2, presenting the pilot countries and their expectations in terms of EO data; it also presents the use cases defined by the project;
- Section 3, showing the data that were used by the project;
- Section 4, explaining the Sen4Stat system that was developed and demonstrated;
- Section 5, focusing on the demonstration and the successful outcomes in terms of use cases;
- Section 6, concluding the deliverable with the system uptake.

1.3 References

1.3.1 Applicable documents

ID	Title	Reference	Issue/Rev.	Date
AD.1	Statement of Work for ESA Sentinels for Agricultural Statistics	EOEP-EOPS-SW-17-015	1.0	15/03/2017
AD.2	Sen4Stat Implementation Proposal - Chapter 5		1.0	12/05/2017
AD.3	Sen4Stat Concept Paper – Satellite EO for Agricultural Statistics			
AD.4	Sen4Stat User Requirement Document	Sen4Stat_URD_V2.1	2.1	26/11/2020
AD.5	Sen4Stat National Dataset Document	Sen4Stat_NDS_v1.2	1.2	26/11/2020
AD.6	Sen4Stat ATBD for pre-processing	Sen4Stat_ATBD-EO_Data_Pre_Processing_v1.0	1.0	25/08/2021
AD.7	Sen4Stat ATBD for in situ data preparation	Sen4Stat_ATBD-In_Situ_Data_Preparation_v1.0	1.0	27/08/2021
AD.8	Sen4Stat ATBD for compositing	Sen4Stat_ATBD-Compositing_v1.0	1.0	25/08/2021
AD.9	Sen4Stat ATBD for biophysical indicators	Sen4Stat_ATBD-SpectralIndices-BiophysicIndic_v1.0	1.0	25/08/2021
AD.10	Sen4Stat ATBD for crop mapping	Sen4Stat_ATBD-Crop_Mapping_v1.0	1.0	30/08/2021
AD.11	Sen4Stat ATBD for yield estimation	Sen4Stat_ATBD-Yield-Estimation_v1.0	1.0	30/04/2022

Table 1-1. Applicable documents

1.3.2 Acronyms and abbreviations

Acronym	Definition
AD	Applicable Document
ATBD	Algorithm Theoretical Basis Document
BOA	Bottom of Atmosphere
DAPSA	Direction de l'Analyse, de la Prévision et des Statistiques Agricoles
EO	Earth Observation
ESA	European Space Agency

ESU	Elementary Sampling Unit
FAO	Food and Agriculture Organization
FAPAR	fraction of Absorbed Photosynthetically Active Radiation
FCover	fraction of Vegetation Cover
FR	Final Report
GPS	Global Positioning System
INRA	Institut National de la Recherche Agronomique
L1, L2	Level 1, Level 2
LAI	Leaf Area Index
NASD	National Agency for Statistics and Demography (Senegal)
NBS	National Bureau of Statistics (Tanzania)
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NSO	National Statistical Office
ODK	Open Data Kit Collect
RS	Remote Sensing
SAR	Synthetic Aperture Radar
SDG	Sustainable Development Goal
Sen4Stat	Sentinels for Agricultural Statistics
SLC	Single Look Complex
SMOTE	Synthetic Minority Oversampling Technique

Table 1-2. List of acronyms and abbreviations

2 Users' requirements

2.1 Pilot Countries

The selection of the Pilot Countries relied on three main types of criteria:

1) **Institutional** criteria, taking into account the official mandate that has the NSO (and the existence of possible conflicts between the NSO and the Ministry of Agriculture), its technical capacities in statistics and remote sensing as well as the resources available for a possible data collection as part of the Sen4Stat project;
2) **Technical** criteria, looking at the type of sampling frame (area or list) already in place for data collection, the expected value that Sen4Stat could add to the current system, the EO data availability (cloud coverage) and the "compatibility" between the agricultural landscape and the EO resolution (field size, topography, crop complexity, etc.);
3) **Agency uptake** criteria, which depends on the country's actual needs, the political willingness as well as its human and technical capacities.

The following 5 (+1) countries were selected with the Steering Committee and ESA to be the pilot countries of the Sen4Stat project: Spain, Ecuador, Senegal, Malawi and Tanzania. Angola was added as pilot country during the course of the project given the fact that no in situ data were available neither in Malawi nor in Tanzania (see section 3.2).

The contact points and the main characteristics of these countries are given here below:

Spain

The contact point is Mr. Gonzalo Eiriz Gervas, from the Ministry of Agriculture, Fisheries and Food - Department of Analysis and Agricultural Statistics. The responsibilities of his department are:

- Collecting agriculture information supporting the calculation of statistics;
- Processing, compiling and disseminating statistics;
- Data analysis

Ecuador

The contact point is Mr. David Armando Salazar, from the National Institute of Statistics and Censuses; Directorate of Agricultural and Environmental Statistics. The Directorate is directly linked with the Presidency. The responsibilities of his institute are:

- Collecting agriculture information supporting the calculation of statistics;
- Processing statistics;
- Disseminating statistics.

Senegal

The contact point is Mr. Babacar Ndir, assisted by Mr. Kande Cissé, from the National Agency for Statistics and Demography (NASD). NASD is the central body of the National Statistical System.

It provides a technical coordination under the supervision of the National Statistical Council. Agricultural statistics are embedded in its "Statistical Information Management".

NASD's responsibilities are:

- Designing the sampling frame;
- Processing, compiling and disseminating statistics.

NASD is not in charge of the national surveys. The collection of agriculture information that will support the calculation of statistics is under the responsibility of the "Direction de l'Analyse, de la Prévision et des Statistiques Agricoles" (DAPSA), which is part of the Ministry of Agriculture. DAPSA is therefore responsible for collecting the data and they would also be the ones who would own the EO technology.

Malawi

The contact point is Mr. Readwell Musopole, assisted by Mr. Emmanuel Mwanaleza, from the Ministry of Agriculture, Irrigation and Water Development - Department of Agricultural Planning Services. The responsibilities of his department are:

- Collecting agriculture information supporting the calculation of statistics;
- Processing, compiling and disseminating statistics.

The collection of agriculture information supporting the calculation of statistics is carried out by the Ministry of Agriculture in close collaboration with NSO. Surveys organization is the mandate of the Ministry of Agriculture and the coordination of these surveys is done by the Statistics Section of the Ministry. Statisticians belonging to this specific section are from the NSO but being seconded to the Ministry.

Tanzania

The contact point is Dr. Albina Chuwa, assisted by Mr. Titus Mwisomba and Mr. Jerve Gasto, from the National Bureau of Statistics (NBS), belonging to the Ministry of Finance and Planning. The NBS coordinates all related-activities to statistics at national scale, with other Ministries supporting them (Ministry of Agriculture, Ministry of Livestock and Fisheries, Vice – President's and President's Offices, Union and Environment, Ministry of Industrial Trade, Regional Administration and Local Government). Its responsibilities are:

- Collecting agriculture information supporting the calculation of statistics
- Processing, compiling and disseminating statistics.

Each ministry has its own role to play in the data collection.

For the moment, processed statistics are mainly used for national reporting and socio-economical analyses as well as for SDG's reporting (goal 2 – zero hunger), but private sector might be a new user in the future.

Angola

The contact point is Mr. Luciano Lupedia, from the Angola Space Office, named "GGPEN" for "Gabinete de Gestão do Programa Espacial Nacional". GGPEN is not a NSO but is in contact with the NSO and the Ministry of Agriculture in Angola.

2.2 User requirements as use cases

Before the start of the project, a concept paper [AD.3] was already developed with the active input from the steering group and a consultation meeting hosted by FAO. The concept paper laid out the background, drivers and a general roadmap how EO information can be further integrated in agricultural statistics. After the selection of the pilot countries, theses initial requirements were consolidated through (i) a broad review of the Steering Committee initiatives in terms of EO data supporting agricultural statistics and (ii) a survey shared among the pilot countries to better capture their expectations from the project.

The answers of the pilot countries to our survey were synthetized as follows:

1) **Senegal's** expectations were very clear:
    a. *Improving design-based* crop acreage estimators by *optimizing* the integration of EO and ground data in the calculation of crop acreage statistics *to reduce the standard error while providing unbiased estimates* (therefore reducing the coefficient of variation *without increasing the number of samples*);
    b. *Improving design-based* crop production estimators by *optimizing* the integration of EO and ground data in the calculation of crop production statistics *to reduce the standard error of the estimate while not increasing the number of samples*
2) **Ecuador, Malawi and Tanzania** agreed with **Senegal** on the expectation of improving the sampling design, with the aim of *improving/optimizing the spatial allocation of the samples*. **Ecuador** expects, in addition, to reduce the standard error (they say: *uncertainty*) by *balancing/optimizing the allocation of the samples* in such a way that the sample size could be increased in crops where the *uncertainty* is higher (*without increasing the total number of samples*).
3) **Ecuador** agreed with **Senegal** on the expectation of improving design-based estimators *to reduce the standard error while providing unbiased estimates* (therefore reducing the coefficient of variation *without increasing the number of samples*).
4) **Senegal, Spain and Tanzania** agreed on the expectation of improving design-based estimators but, with the aim of reducing the survey costs (without increasing the standard error) instead of reducing the standard error (without increasing the survey cost: number of samples).
5) **Senegal and Spain** expected further reductions in survey costs in the estimation phase, replacing survey data with EO data on specific crops. This expectation is not aligned with the statistical framework sensus stricto. Indeed, in a strict statistical framework, the quality of the survey data and auxiliary data is not the same. In addition, it would be an arbitrary change in the probabilistic scheme used to select the samples and, as a result, the sampling distribution would be invalidated and could not be used for uncertainty assessment. For these reasons, these expectations were not considered as a priority.

6) **Spain and Ecuador** agreed on the expectation of improving design-based estimators to disaggregate the estimates obtained at the national level, at the level of minor administrative units (region/province/canton/county).
7) **Spain, Ecuador, Malawi and Tanzania** expected that EO data will help improve timeliness of estimates.
8) **Spain, Ecuador, Senegal, Malawi and Tanzania** all expected that EO data can be used to increase the quality of the survey data (with different procedures proposed).
9) NSOs from **Ecuador, Malawi and Tanzania** were responsible for SDG's reporting and were thus interested in investigating how EO data can support this activity.

Some NSOs also expressed additional expectations not directly related to the agricultural statistics. These expectations are outside of the scope of this project and might not be considered as a priority.

Various statistical methods developed to integrate EO ancillary data with survey data exist in the sampling survey literature. The statistical methods that can answer NSO's expectations are **cost-efficiency**, **domain and small area estimation**, **timeliness** and **optimizing the sample design**. These use cases are briefly presented below. More details can be found in the Users Requirements Document [AD.4].

2.2.1 Cost-Efficiency

Official statistics is expensive, mainly because they must be based on high quality unbiased and reliable data. As a result, the survey cost is a key criterion for choosing one among a set of sampling techniques. The other key criterion is accuracy, which includes both bias and sampling variance: from now on, we limit ourselves to design-consistent sampling surveys which are unbiased (or approximately unbiased) so that the accuracy measure will be the sampling variance. Cost-efficiency can therefore be calculated as the survey cost multiplied by the sampling variance, so that it integrates the two key criteria in only one.

In order to improve design-based estimators, it is convenient to differentiate between EO contributions to (i) the sampling design, and to (ii) the estimation. This use case focus on the estimation while the EO contribution for optimizing the sampling design is treated in a different use case (section 2.2.4).

In this cost-efficiency use case, what we do is:

a) Integrating EO data with the statistical survey provided by the NSO,
b) Evaluating the effect on the cost-efficiency of the sampling design currently used.

To evaluate this effect, we compare the cost-efficiency of the current sampling design where only ground data (without EO data) are used, with the cost-efficiency of the current sampling design where both ground data and EO data are used.

2.2.2 Domain and small area estimation

In most of the pilot countries, the current sample is designed to achieve the required estimates accuracy at the national or regional level. However, reliable estimates over minor administrative

areas, such as province and county, are also required without increasing the sample size – this has been stated by the NSO.

In a minor administrative area, the sample size will be always lower than at the national level and, as a result, the estimators' accuracy will be also lower. In the literature on sampling survey:

- a domain is a part of the population (say a region/province) where the sample size is big enough for the design-based estimator to be sufficiently precise for most uses;
- a small area is a part of the population (say a county) where, due to the small sample size, the design-based estimator is not sufficiently precise for most uses.

In this use case, we will consider an alternative GREG estimator, which is design-based and more accurate for domain estimation than the projective estimator used at national-level. For small area estimation, we will use a model-based estimator to "borrow strength" from related small areas in order to obtain precise estimates for a given small area. This estimator makes optimal use of the available data, according to statistical criteria, and allows for providing estimates even in counties where the sample size is null. The use of EO data is key for this application.

2.2.3 Timeliness

Official statistics are published a long time after the end of the campaign, thus being not available at the right time to take decisions. NSOs expect that EO data will contribute improving this timeliness. Two main applications were considered here: getting crop acreage forecasts at the mid-season and crop yield forecast one month before the harvest and supporting a more rapid publication of consolidated statistics. Of course, the demonstration of the second application is not in our hands. It will be discussed with the NSO during our iterations but providing a clear demonstration in the framework of the project is not feasible.

2.2.4 Optimizing the sample design

The procedure for elaborating agricultural statistics begins with the design of the sample for ground data collection, and it finishes with the computation of the required estimates. In all the other use cases, the starting point is the sample already existing in the pilot country, and we focus on how the integration of ground and EO data improves the estimators.

However, *optimizing* the sample design is also key. The specific expectation here is to generate a specific maps (e.g. cropland vs non-cropland map, irrigation map) that would be used to update the stratification that serves as a basis for them to define their samples allocation.

2.2.5 Survey protocol and uncertainty reduction in survey data

Statistics estimation relies on a strong assumption, which is that the ground data is of high quality, i.e. unbiased and reliable. In practice, this might not be totally the case. This use case, which is not a statistical use case sensus stricto, investigates how EO data can contribute increasing the quality of the ground data. More precisely, we will propose a two-fold approach:

1) integrate EO data in the ground data collection protocol:

a. Providing satellite images from the current season, to support the interviews and the delineation of the plot outlines;

b. Improving the protocol to collect in situ data (using numerical support and GPS to get parcel-based information);

c. Implementing a quality control procedure of the ground data collected on the field in near real time, during the campaign;

2) using EO data after the survey:

a. Interpreting samples twice independently, on the field and from EO data;

b. Implementing a quality control procedure of the samples collected on the field (ground database) after the survey, to remove systematic errors (bias);

c. Collecting information even for samples where field visits are not possible due to insecurity reasons (conflict areas) thanks to direct interpretation of EO time series.

3 Data sets

3.1 EO dataset

EO data used in the project were from optical Sentinel-2 and Synthetic Aperture Radar (SAR) Sentinel-1 sensors. Dense time series of both SAR and optical images are expected to enable to handle different data flow dynamics providing more complete information all along the growing season.

S2 were acquired as atmospherically-corrected Bottom of Atmosphere (BOA) reflectance products, corresponding to the Level 2A provided by ESA. An additional cloud mask is systematically generated using the FMask algorithm, using the Sentinel-2 Level 1C product. This FMask cloud mask is then combined with the cloud layer of the L2A products to generate a more comprehensive validity mask.

As for the Sentinel-1 data, the Sen4Stat pre-processing generates time series of SAR amplitude/phase and coherences at 10-meter spatial resolution. The pre-processing starts with the Single Look Complex (SLC) L1 of the S1 IW data (Terrain Observation with Progressive Scans (TOPS) mode) in order to generate (i) one stack of calibrated, co-registered and projected amplitudes and (ii) one stack of co-registered and projected 12-days coherences data.

3.2 Statistical surveys

The collection of statistical surveys was a critical step of the project because it provided us with in situ data supporting all R&D developments in Phase 1 and allowing the demonstration in Phase 2. The data collection was carried out from email exchanges, which started in fall 2019. It has to be mentioned that the COVID-19 lockdown had a very strong negative impact on this task (no physical meeting possible) and significantly delayed the project.

Spain and Ecuador were able to provide us rapidly ground survey meeting our requirements. Malawi was also well responsive but did not have ground survey with GPS coordinates at plot level. Existing database in Malawi are only administrative statistics data and survey with GPS coordinates at household level. In fall 2020, Senegal delivered to the project data from the year 2018, with GPS coordinates both at household- and parcel-level. From Tanzania, no was received. This is why the country was replaced by Angola. Unfortunately, in Angola, the data shared did not come from NSO and did not follow a statistical sampling design.

The main datasest (Spain, Ecuador and Angola) are briefly described below; more details can be found in [AD.4] and [AD.5].

- **Spain**

The statistical dataset in Spain is named “ESYRCE”. ESYRCE is an integrated list and area frame survey over all Spain, with a master frame that is the same since 2006-2007. From this date, the sampling frame has not evolved; only the way of collecting the information might have evolved.

Surveys take place each year. Information is collected about crop type, crop area, yield (estimated by crop cutting or visual estimation if the crop is still in place or by farmers interviews if the crop is already harvested and the farmer is in the field at the moment of the information collection), production system (irrigation/not; seeding procedure, soil maintenance, permanent culture age and density).

Elementary Sampling Units (ESU) of the survey are generally of 49 hectares (700x700 m) and exceptionally of 25 ha (500x500 m). They have no link with the Land Parcel Identification System dataset used in the Common Agricultural Policy context. All crops present in the ESUs are identified and the yield is estimated over 1/3 of the ESU. GPS coordinates are recorded at parcel-level, including plot outlines. Figure 3-1 provides an overview of the dataset for one year. Each segment is composed of plots (or polygons) representing agricultural parcels. Crops are described in the attributes table and a separate table with key field describe fruit tree plantations.

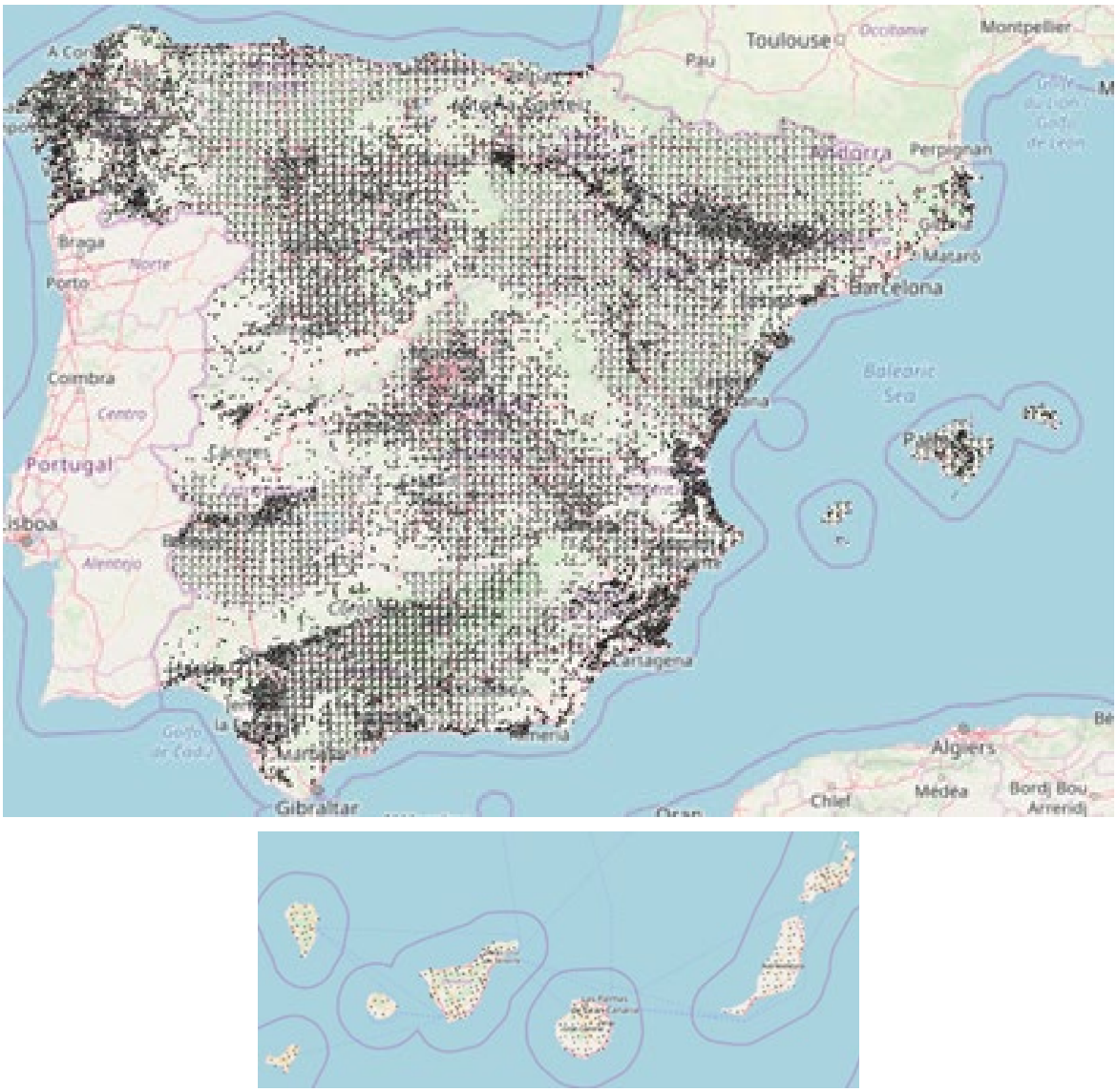


Figure 3-1. Overview of Spain dataset.

- **Senegal**

The survey in place in Senegal is a list frame survey over all country, with a master frame that is the same since 2013. In 2013, an agriculture census took place, which allowed listing all active farmers (identifying new ones and removing the ones who stopped since the previous census). In parallel, a mapping exercise aimed at listing the active farmers by village, resulting in a map of agricultural households by village. Master frames are usually updated every 5 years, based on a new land cover map. The next update should take place soon.

2000 holdings are selected thorough a stratified sampling from the 526.000 holdings in Senegal, which corresponds to ~ 0.4%. These 2000 holdings are spread in all the Senegalese departments, in direct ratio to the size of these departments. The same holdings are visited during 2 consecutive years and then, a new sample of holdings is drawn.

In each holding, farmers are interviewed and GPS measurements are done in all fields belonging to the household. GPS coordinates are recorded at the parcel-level and the plot outlines are also recorded to get the parcels area. For the main season crops (not off-season crops), the Information is collected about crop type, crop area and production (no crop cutting, only farmers' estimates). For the production, information from the past and the current years is collected: from the past year, the crop and the production and from the current year, the expected production. Farmers estimate their production using a variety of units, which are then translated into regular units thanks to conversion tables from the Ministry of Agriculture.

Surveys are conducted annually, during the second half of the season (i.e. starting in August) and in any case, before the harvest. The surveys are carried out using a decentralized approach, through regional offices.

The survey that was shared is from 2018. It contains 16861 lines which correspond to the parcels belonging to about the 4693 households surveyed. Each line is thus dedicated to a single parcel for which the geographic coordinates are provided. The households and parcels are distributed all over Senegal (Figure 3-2).

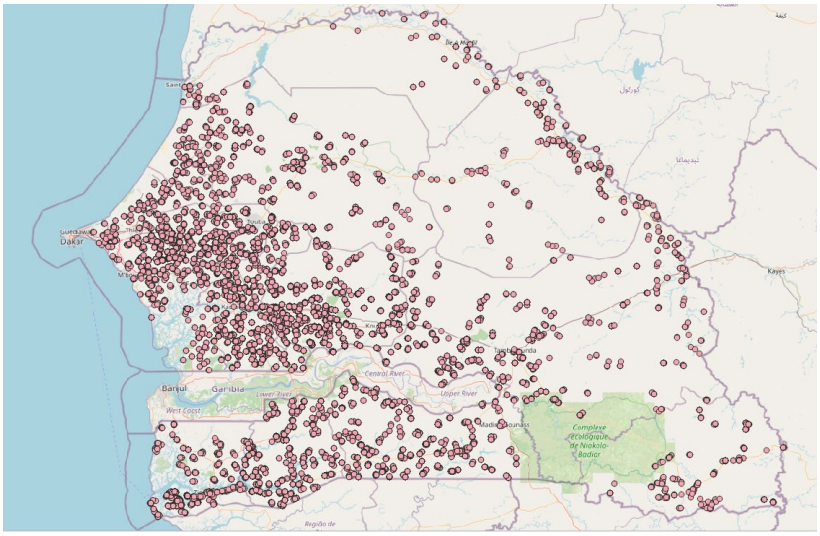


Figure 3-2. Household locations covered by the agricultural survey

- **Angola**

The data shared with the project covered 3 regions - Malanje, Bié and Huambo - and spread over 3 different years - 2019, 2020 and 2021. Their localization and associated year is presented in Figure 3-3. Most of the points were from 2020. In 2019 and 2020, the points are spread over the 3 provinces while in 2021, the points are all included in the Bié province.

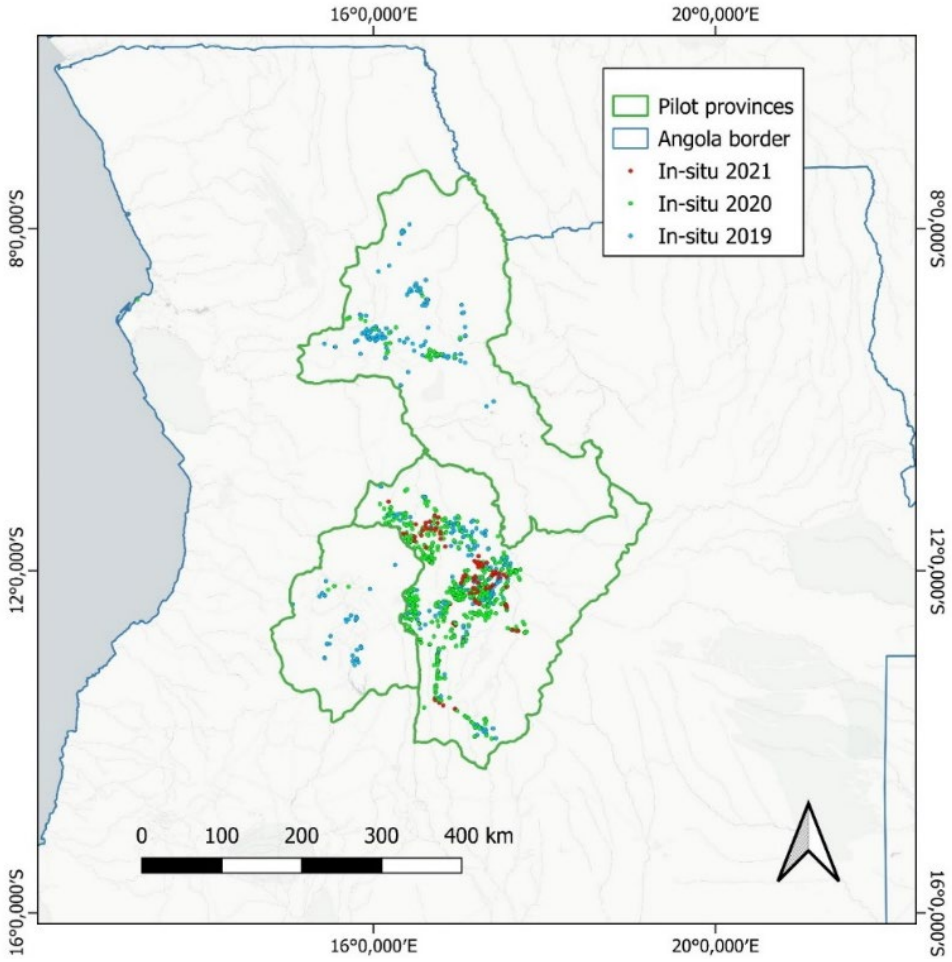


Figure 3-3. Localization of the in-situ data shared by GGPEN over the Area of Interest

4 Overall approach and Sen4Stat system

4.1 Requirements and challenges for coupling statistical survey and EO data

The Sen4Stat objective is certainly not to use EO data to replace the agricultural statistical survey, but rather to leverage EO data to complement the survey in order to produce more accurate, more disaggregated and more timely statistics.

This EO integration in the agricultural statistics workflow faces several challenges:

1) In order to be used to train and validate classification and yield estimation algorithms, information contained in the agricultural surveys need to be georeferenced at parcel-level: crop type, crop area and crop yield estimation need to be associated with a specific parcel and not with the household;
2) The number of crop samples included in the agricultural survey might not be sufficient to efficiently train classification algorithms. In this case, additional crop data need to be collected, which can be done in an opportunistic way: the data collection protocol does not need to follow a strict statistical design as they will only be used to train the algorithm and not to support the acreage estimations.
3) The number of non-crop samples included in the agricultural survey is not sufficient. Parallel strategies need to be implemented to collect training and validation samples for the non-crop classes present in the area of interest (data collection on the ground, visual interpretation, use of existing thematic products). In any case, this effort should be done only once as theses classes are expected to be relatively stable from one year to the other;
4) The concept of seasonal or annual EO products is extremely important because what matters for the NSO is the total production at the end of the year, taking into account all the cropping cycles that took place during the year.
5) A significant agro-climatic gradient spanning over the national territory gradually shifts the cropping calendar of most crops, as well as the crop type distribution, making the scaling up to national level rather challenging.

4.2 Key principles

As a contribution to properly answer these challenges, a combination of elements was inherited from the Sen2-Agri and Sen4CAP systems or specifically developed and implemented in the Sen4Stat system:

1) Like in Sen2-Agri and/or Sen4CAP:
    a. The methods should make a maximal use of the EO time series information, and should not rely on single date images nor on seasonal composites;
    b. In order to have homogeneous Sentinel-2 time series, a regular sampling of observations is necessary. Temporally interpolated surface reflectance values (with a time step equivalent to the sensor revisit cycle) could maintain the information

content, deal with the spatial heterogeneity of time series density and observation date, and fill the gaps due to clouds or missing values. A high-quality cloud mask and atmospheric correction is necessary to get consistent and smooth time series;

c. While the EO data near-real-time processing should be based on tiles to be fully scalable and run in parallel, the machine learning classification models (for crop type maps) must be trained over larger areas corresponding to the country or to smaller agro-climatic zones rather homogeneous in terms of climate, agro-ecological conditions (relief, soil, etc.), cropping systems and agricultural practices. Such training over large areas avoids requiring a complete set of in situ data for each tile and therefore ensures the validity of the trained models over large areas. Stratifying the country into smaller homogeneous regions also allows coping with agro-climatic gradients inducing a very diversity of crop calendars and growing conditions;
d. The machine-learning algorithm to be selected for the crop type map must be able to cope with the diversity of spectro-temporal signatures for a given crop due to various planting dates, cultivars, and weather conditions, still present in any agro-climatic zone;
e. The methods implemented in the Sen4Stat system should integrate the SAR Sentinel-1 time series to ensure robustness against the gaps in the Sentinel-2 times series due to clouds and allow a continuous monitoring along the season;

2) New in Sen4Stat:
   a. The methods implemented in the Sen4Stat system should rely on statistical survey (for training and/or validation) and the generated crop type map and yield estimates should then be used jointly with the agricultural survey for an improved estimation of acreage and yield/production statistics;
   b. The Sen4Stat system needs to perform for both list and area sampling frames;
   c. The crop type mapping and yield estimation methodologies need to target the main crop as a priority;
   d. Nevertheless, for the crop classification algorithm, a specific step is needed to ensure a significant representation of the minor crops and of the non-crop classes in the training dataset;
   e. By-default, the yield estimation algorithm has to rely on yield ground data collected during the survey but in the case these data are not available, an alternative has to be functional based on the statistics from the past years.

4.3 The Sen4Stat system

The Sen4Stat system consists of an open source EO processing system linked with (i) a module for in situ datasets quality control, (ii) a visualization tool and (iii) a set of tools for higher-level statistical analyses. Being open source, it allows any user to generate, at his own premises and in an operational way, products tailored to his needs.

The EO processing chain is a standalone operational processing chain which generates a set of agriculture monitoring products for facilitating the uptake of EO information by the NSO. It relies on Sentinel-2 L1C and/or L2A, Sentinel-1 SLC and Landsat 8 L1T time series to generate

agriculture monitoring products and support the agricultural statistics estimation. These agriculture monitoring products are:

- EO-derived pre-processed reflectance / backscatter / coherence time series;
- EO-derived spectral indices and biophysical indicators, e.g. NDVI or LAI;
- EO-derived crop growth metrics at segment-level;
- Cloud-free colour composites;
- EO-derived crop maps (cropland – non-cropland, annual vs permanent cropland, crop type groups and crop type);
- EO-derived crop yield estimates at the level of the reporting unit.

The logical data flow and the main interfaces of the Sen4Stat EO operational system is provided in Figure 4-1.

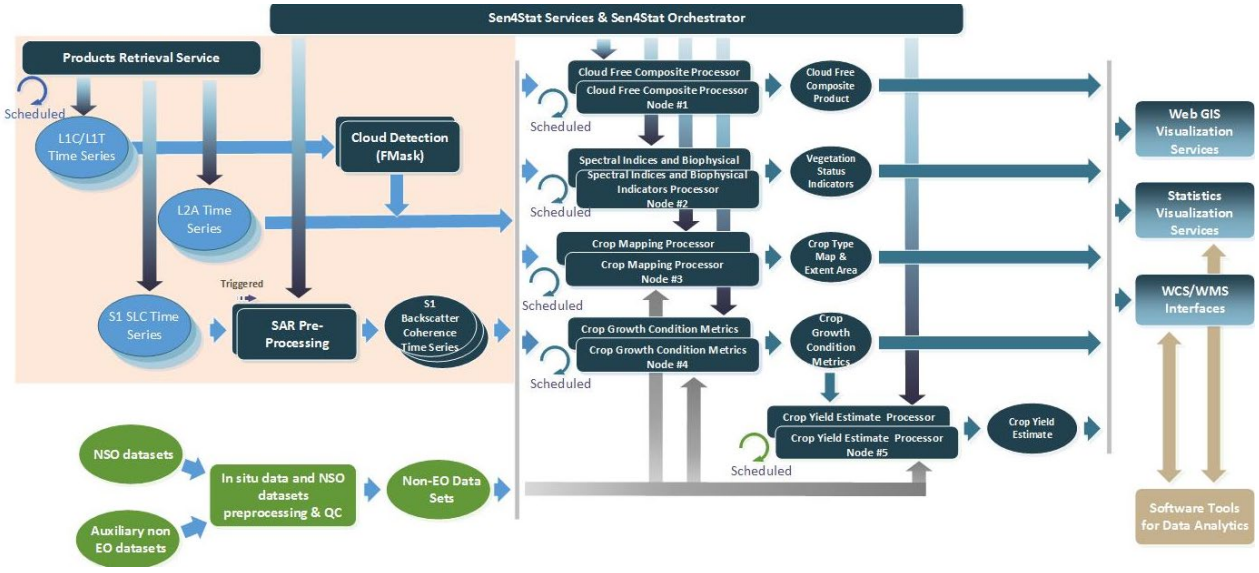


Figure 4-1. Logical data flow of the Sen4Stat EO processing system and its link with the external tools and modules

4.4 Sen4Stat modular approach

The Sen4Stat system is composed of a set of independent processing modules orchestrated by a data-driven approach. These modules, named “processors”, are based on a set of tools which can be re-used outside of the entire Sen4Stat system. They take care of the EO data pre-processing and they transform the pre-processed time series into relevant agriculture products. The methods implemented in each of these processors are briefly described in the sub-sections below. The reader is referred to the corresponding Algorithm Theoretical Basis Documents (ATBDs) for a complete description [AD.6, AD.7, AD.8, AD.9, AD.10, AD.11].

4.4.1 EO Data Pre-Processing processors (SAR and optical)

These processors carry out the pre-processing for all EO data supported by the Sen4Stat system: Sentinel-2, Sentinel-1 and Landsat 8.

For Sentinel-2, users can choose to use directly the Sen2Cor L2A images automatically produced by ESA and available on the ESA Copernicus Data Space Ecosystem and on most of the cloud providers. In this case, the pre-processing processor offers the option to generate an additional cloud mask using FMask. Alternatively, users can decide to work with L1C products. In this case, the processor applies both atmospheric correction and cloud mask algorithms.

The same strategy is implemented for Landsat 8.

For the SAR sensors, the processor transforms the Level 1 (L1) products into backscatter and coherence products.

4.4.2 Spectral Indices & Biophysical Indicators processor

This processor provides three Spectral Indices and three Biophysical Indicators informing about the evolution of the green vegetation:

- The ***Normalized Difference Vegetation Index (NDVI),*** the most popular indicator operationally used for vegetation monitoring, provided to ensure continuity with existing long-term time series and thus, allowing anomalies detection;
- The ***Normalized Difference Water Index (NDWI)***, introduced for the first time in 1996 and reflecting moisture content in plants and soil;
- The ***Brightness***, defined as the Euclidean norm of the surface reflectance values in green, red, NIR and SWIR;
- The ***Leaf Area Index*** (***LAI)***, an intrinsic canopy primary variable that should not depend on observation conditions, which determines the size of leaf interface for exchange of energy and mass between the canopy and the atmosphere;
- The ***fraction of Vegetation Cover (FCOVER)***, corresponding to the fraction of ground covered by green vegetation. It quantifies the spatial extent of the vegetation;
- The ***fraction of Absorbed Photosynthetically Active Radiation (FAPAR)*** by the green and alive elements of the canopy. The FAPAR depends on the canopy structure, vegetation element optical properties, atmospheric conditions and angular configuration.

The NDVI is computed using a standard formulation applied to the Sentinel-2 red (B4) and narrow Near InfraRed (NIR) (B8a) bands. The NDWI also relies on a standard formulation applied to the Sentinel-2 narrow NIR (B8a) and Short-Wave InfraRed) (B11) bands. As already mentioned, the Brightness computation makes uses of the Sentinel-2 bands in the green (B3), red (B4), narrow NIR (B8a) and SWIR (B11).

The LAI retrieval is performed from the bands 3, 4, 5, 6, 7, 8, 9, 12, 13 using machine learning to build a non-linear regression model. For the LAI, FCOVER and FAPAR, the implementation is derived from the one already proposed in the frame of the ESA Sentinel-2 toolbox. The LAI, FCOVER and FAPAR retrieval is performed by applying a global Artificial Neural Network (ANN) on each pixel considering the reflectance values of all the available bands pre-processed at

the L2A and some geometric configuration as input. The training of the ANN, which consists in generating the training database, defining the neural network architecture and calibrating the network, is not performed within the Sen4Stat system. Instead, the Sen4Stat system benefits from an already trained ANN, made openly and freely available by the Institut National de la Recherche Agronomique (INRA) which developed the algorithm. From the system implementation point of view, this trained ANN is given as auxiliary data to the processor.

4.4.3 Cloud-free temporal syntheses processor

This processor provides a cloud-free composite of surface reflectance values in the 10 S2 bands designed for land observation and keeping their native spatial resolution (10 or 20 meters). The processor is based on the weighted average composite approach, includes the correction of directional effects to consider changes in observation angles and therefore in reflectance values among the different images that are stitched to create the product.

4.4.4 In-situ data preparation processor

In situ data (about crop type and crop yield) are mandatory to run the crop type mapping and the crop yield estimation processors. As the Sen4Stat system aims at facilitating the uptake of EO information by the NSO, it is assumed that in situ data will come from agricultural surveys conducted by the NSOs to estimate their crop acreage and yield statistics. Nevertheless, the system can use any other source of in situ data providing that they are in the good format.

In situ data are quality-controlled and formatted before being used to produce the Sen4Stat EO products. This is the objective of the “In Situ Data Preparation” processor. This processor assumes that the in situ data are provided **as polygons with a given set of attributes**. It aims at aims at qualifying each polygon with a set of indicators or flags related to its geometry, area, quality and stratum. The analysis of the geometries allows to:

- Determine if the geometry is valid (i.e., the geometry is not empty nor overlapping itself);
- Determine if the geometry is unique;
- Determine if the geometry is composed of a multipart polygon;
- Identify polygons overlapping their neighbours.

The in-situ data preparation also includes the rasterization of the polygons, allowing to count the number of underlying pixels for each parcel.

Finally, a negative 10m buffer is applied to the geometries and they are reprojected into the WGS 84 / UTM zone coordinate systems that correspond to the Sentinel tiles underlying the parcels.

4.4.5 Crop mapping processor

This processor relies on per-pixel machine learning and deep learning algorithms to generate various crop maps. Two types of data are used to feed classifier algorithms: in situ data (pre-processed by the “In Situ Data Preparation” processor presented above) and EO data.

Three classifiers will be included in the "Crop Mapping" processor: Random Forest (OpenCV and Ranger implementations), Neural Network (not yet implemented in the Sen4Stat 1.1 version) and Broceliande.

A detailed crop type legend is used to train the classifiers and at the end, the detailed legend of the crop map can be simplified into different products: binary cropland - non cropland map, binary annual vs permanent crops, main crop type groups, detailed crop type map.

The "Crop Mapping" processor is designed to be efficient at national scale. As the amount of calibration data is limited and probably not distributed uniformly over the country, the processor offers the possibility to stratify the area of interest, i.e. to split it into multiple agro-climatic regions - called strata - which are homogeneous in terms of climate, agro-ecological conditions (relief, soil, etc.), cropping systems and agricultural practices. The use of such stratification allows reducing the natural variability existing when working at national scale by coping with agro-climatic gradients inducing a very diversity of crop calendars and growing conditions. Each stratum is classified independently, i.e. with his own set of calibration pixels and his own classification model.

4.4.6 Crop growth condition metrics and yield estimation processor

The yield estimation approach implemented in Sen4Stat relies on two steps, which are implemented in two different processors: yield features extraction and yield model design and application.

- **Crop growth condition metrics**

This first yield processor aims at extracting metrics that are representative of the crop growing and that will be further used as proxy variables of the yield in the training and validation of statistical models in the next processor.

The main EO input of the processor is the LAI time series, impacting directly or indirectly all yield features. Climate data extracted from the ERA5-Land database are also used. The processor also needs in situ data with field boundaries. The unit of processing is defined as the elementary area used to extract the yield features and it might thus vary according to the country.

- **Crop yield estimation**

The metrics calculated by the crop growth condition metrics processors are used as input by this second yield processor. Those yield features are used as proxy variables of the yield in the training and validation of statistical models.

5 Sen4Stat system demonstration

During the second phase of the project, the developed Sen4Stat system was run in our pilot countries. This section focuses on the most representative outcomes of the project, which are from Spain and Senegal.

5.1 Demonstration in Spain

5.1.1 EO products

The first cycle of the demonstration focuses on two provinces: Castilla y Leon and Andalusia (Figure 5-1). These two provinces have been selected by the NSO because they are both important from the economic point of view and they have very different agro-climatic conditions and agricultural practices: Castilla y Leon is one of the major winter cereals productor in Europe, is quite flat and has big fields while Andalusia is dominated by olive trees with an arid climate. The area of Castilla y Leon is of 94.226 km² and the area of Andalusia is of 87.599 km².

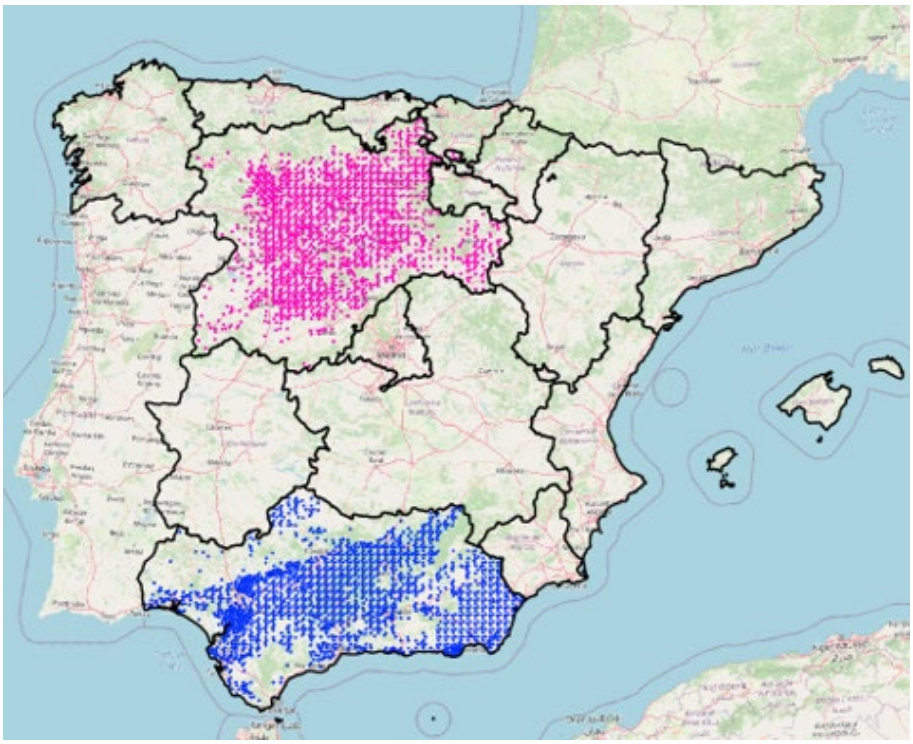


Figure 5-1. Area of Interest in Spain (cycle 1)

The distributions of crops in the ESYRCE samples for the regions of Castilla y Leon and Andalusia (expressed in terms of surface) are shown in Figure 5-2 and Figure 5-3, showing the specificities of each region. There are 60.666 polygons in Castilla y Leon, which are mainly annual crops while Andalusia counts 58.583 polygons which are mainly permanent crops.

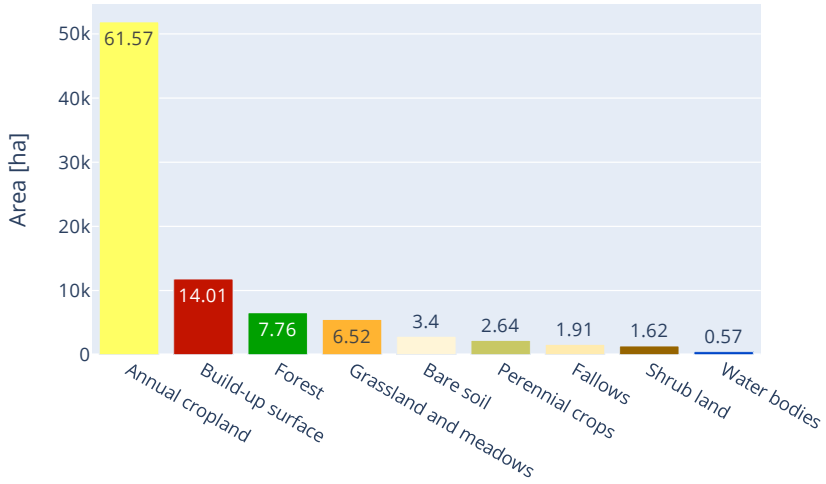


Figure 5-2. Land cover classes representation in the ESYRCE dataset in Castilla y Leon (2020)

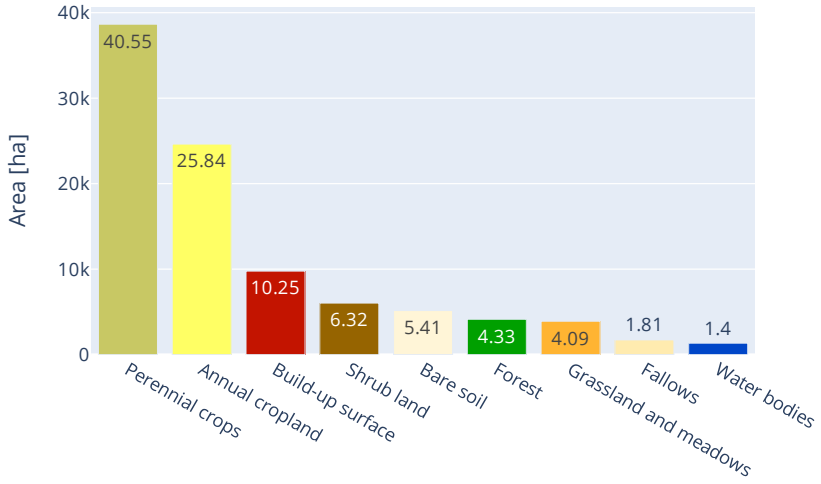


Figure 5-3. Land cover classes representation in the ESYRCE dataset in Andalusia (2020)

The second phase of the demonstration extends the study area to the whole country, using also the 2020 ESYRCE survey. Over the whole Spain, ESYRCE counts 496.182 polygons. The different classes represented in the ESYRCE dataset at national scale is shown in Figure 5-4 (distribution expressed in terms of surface).

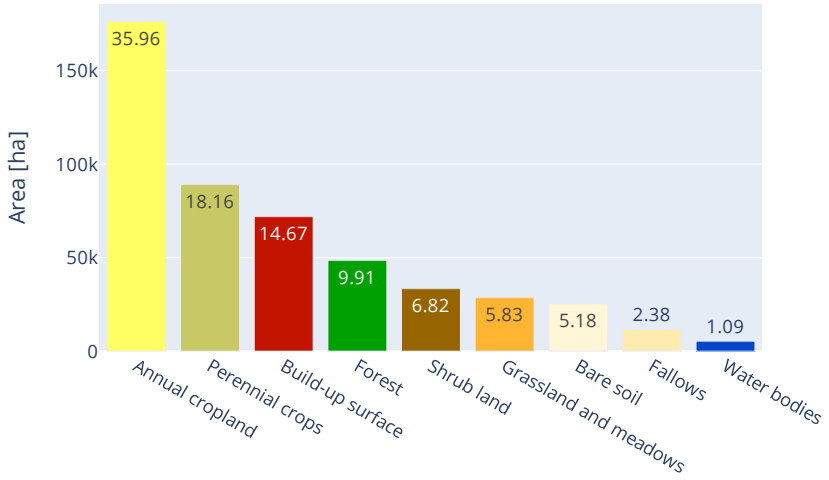


Figure 5-4. Land cover classes representation in the ESYRCE national dataset in 2020

The crop type map over the region of Castilla y Leon is shown in Figure 5-5. The legend counts 28 different crop types. The map is based on Sentinel-2 time series from January to December. A Random Forest algorithm was applied on the S2 bands B03-04-05-06-07-08-11-12 and on the NDVI, NDWI and Brightness. A Synthetic Minority Oversampling Technique (SMOTE) algorithm was used to increase the number of samples of the minor crops and therefore increase their representativeness. The overall accuracy of the map is of 76%.

Figure 5-6 presents a zoom of the crop type map, showing that the map is very smooth despite the fact that this is a per-pixel classification: no a posteriori filtering was applied and the parcels are clearly visible on the map. In addition, the same kind of performance is obtained both for small (right) and larger (left) parcels.

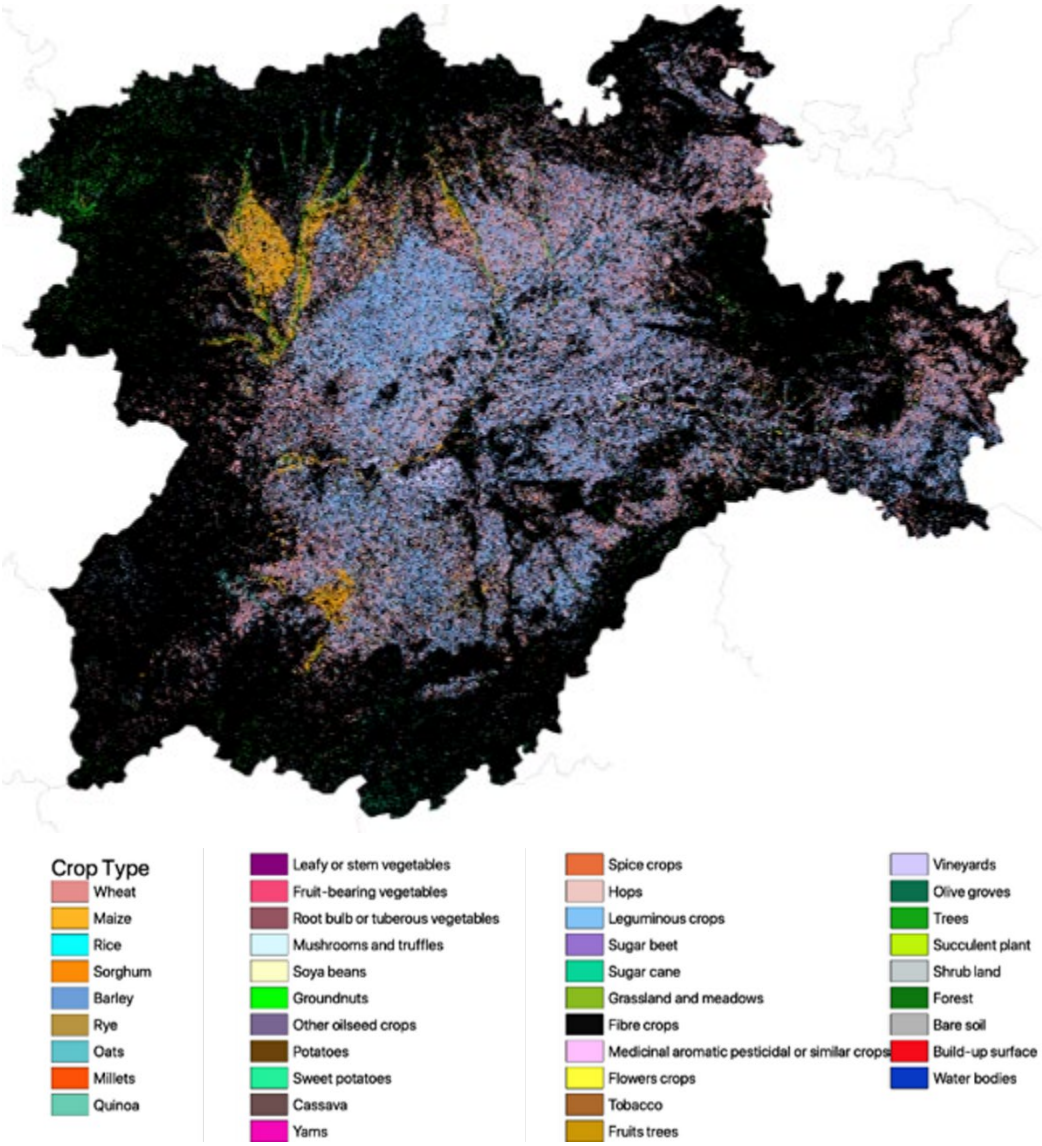


Figure 5-5. Crop type map in Castilla y Leon

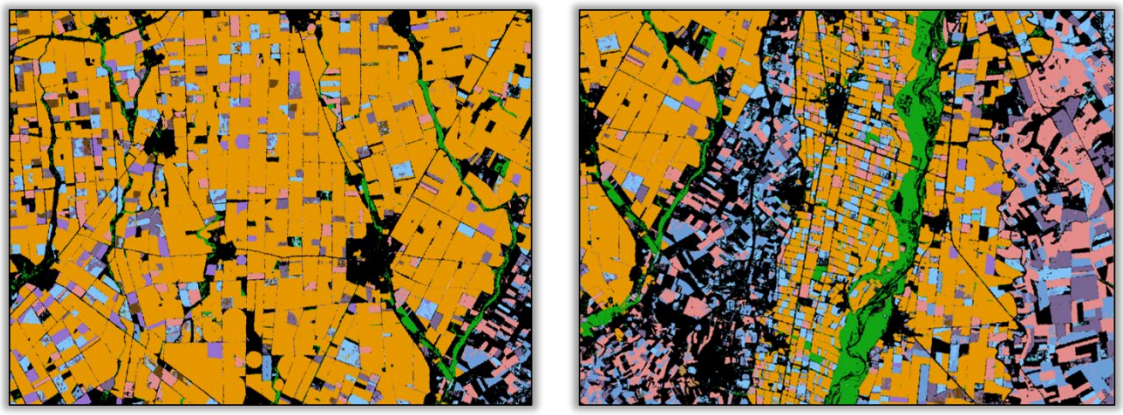


Figure 5-6. Zoom of the crop type map in Castilla y Leon

The confusion matrix of the crop type map revealed that the highest confusion between crops is between wheat and barley, which are two classes very similar, having a low thematic distance. The discrimination between maize and sunflower is very good. The F-Score by class are presented in Figure 5-7. All main crops (barley two row, soft wheat, sunflower and maize) have a F-Score higher than 0.8. Logically, the accuracy increases when grouping the different varieties of barleys and of wheats.

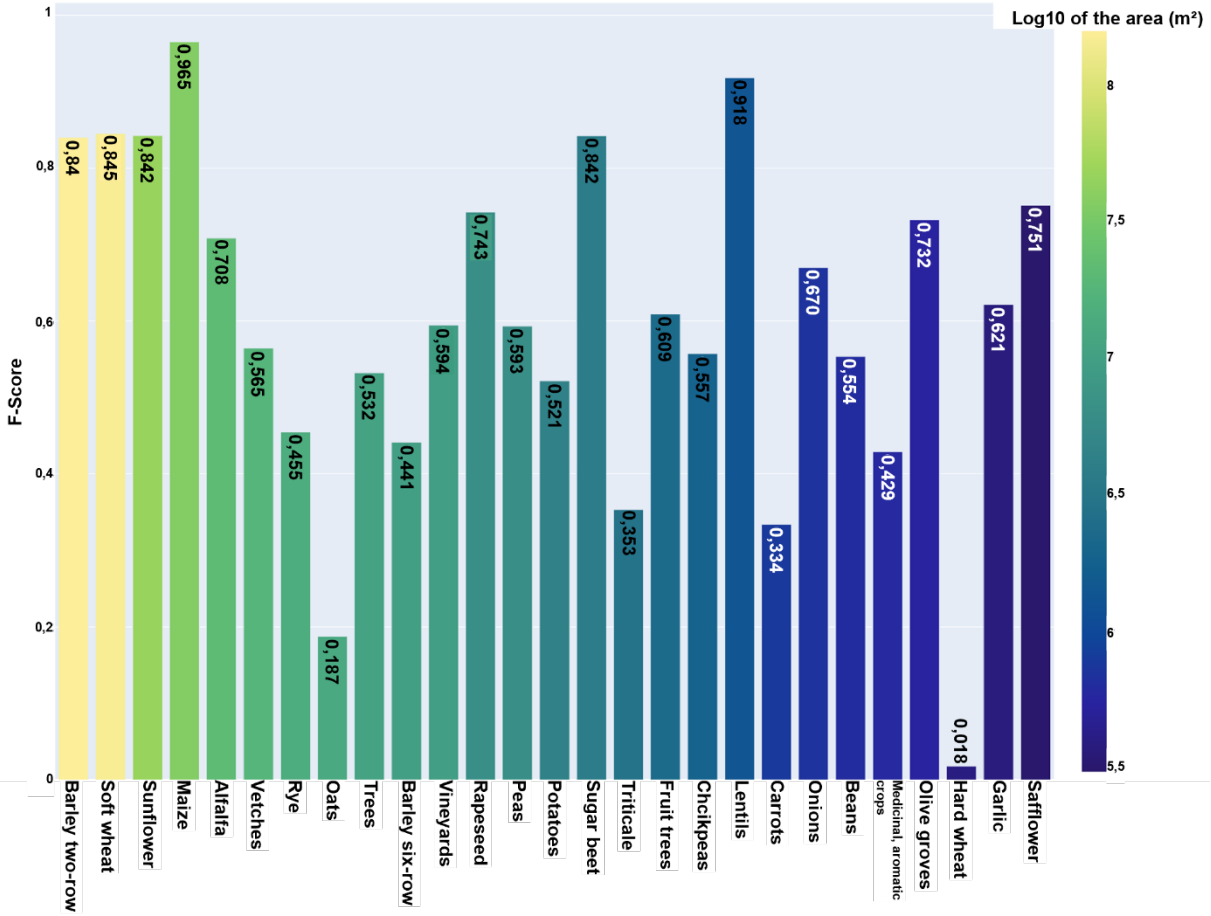


Figure 5-7. F-Score by class sorted by area (largest to smallest) of the crop type map in Castilla y Leon

Similarly, the crop type map obtained over the region of Andalusia, which counts 34 different crop types, is shown in Figure 5-8. The map is based on Sentinel-2 time series from January to December. A Random Forest algorithm was applied on the S2 bands B03-04-05-06-07-08-11-12 and on the NDVI, NDWI and Brightness. A SMOTE algorithm was also used to increase the number of samples of the minor crops and therefore increase their representativeness. The overall accuracy of the map is of 73%.

Figure 5-9 presents a zoom of the crop type map, showing like in Castilla y Leon that the parcels are clearly visible on the map, without too much noise, despite the fact that this is a per-pixel classification without a posteriori filtering. The map is quite good both in very intensive areas with small adjacent parcels having different crops (left illustration) and in more extensive areas fully covered by olive groves (right illustration).

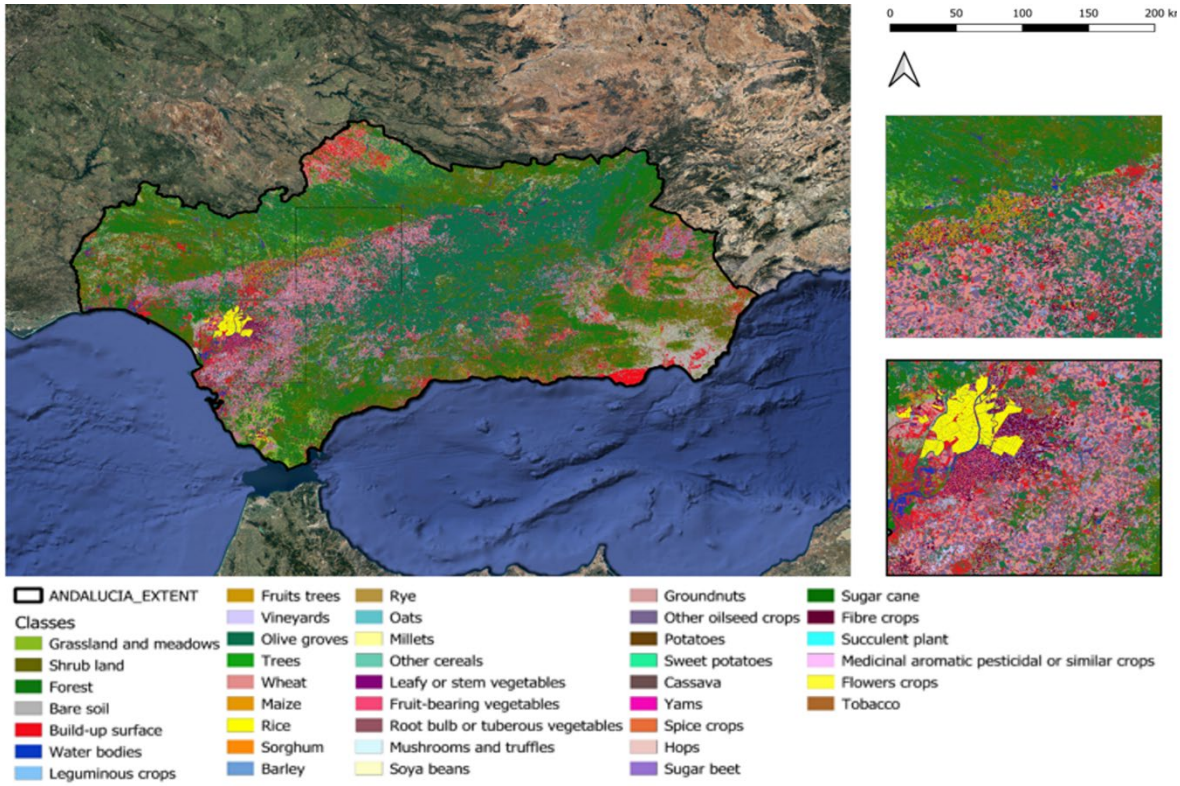


Figure 5-8. Crop type map in Andalusia

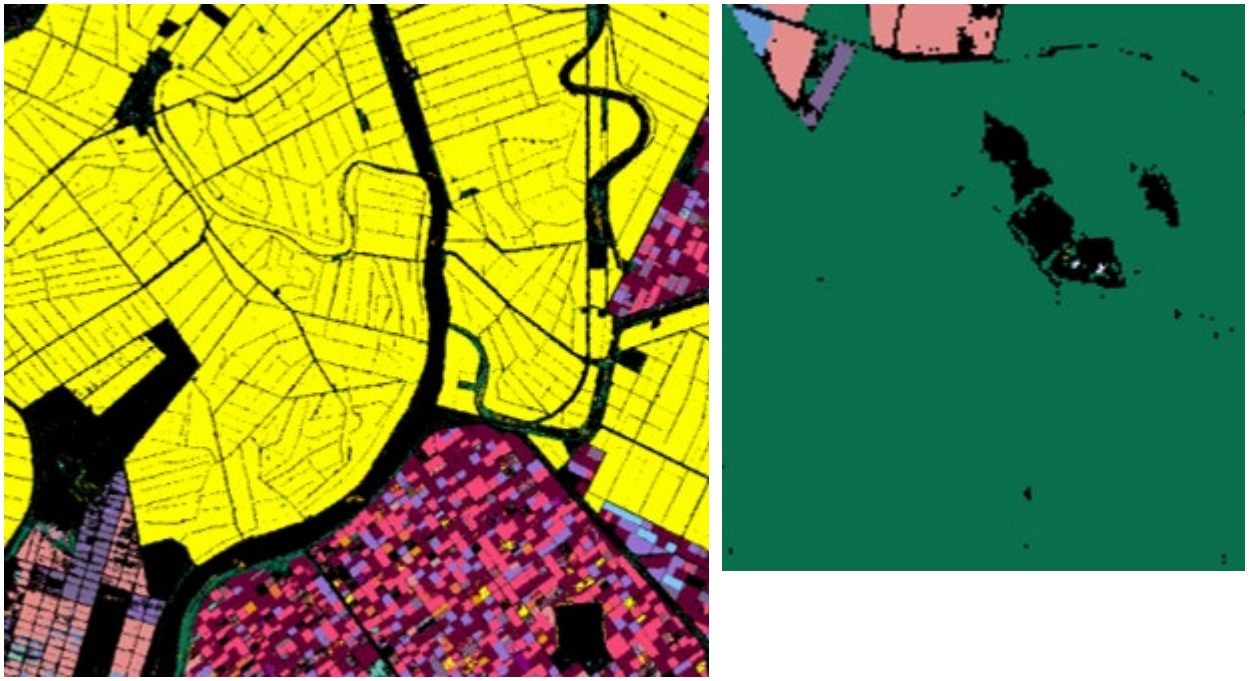


Figure 5-9. Zoom of the crop type map in Andalusia

The confusion matrix showed that the annual crops were well discriminated, especially the maize and sunflower which are the main ones. To some extent, there was a small confusion between olive groves and fruit trees. There existed also some confusion between the non-cropland and crop classes: bare soil is sometimes confused with cereals, grassland and built-up are also confused with the different crop types. The F-Scores by classes are shown in Figure 5-10. Olive groves, sunflower, cotton and rice are well identified and the metric is lower for fruit trees, hard wheat and soft wheat. The accuracy of the fruit trees and vineyards is significantly lower than the olive one; but this can be explained by the fact that there are minor classes in the permanent crops group.

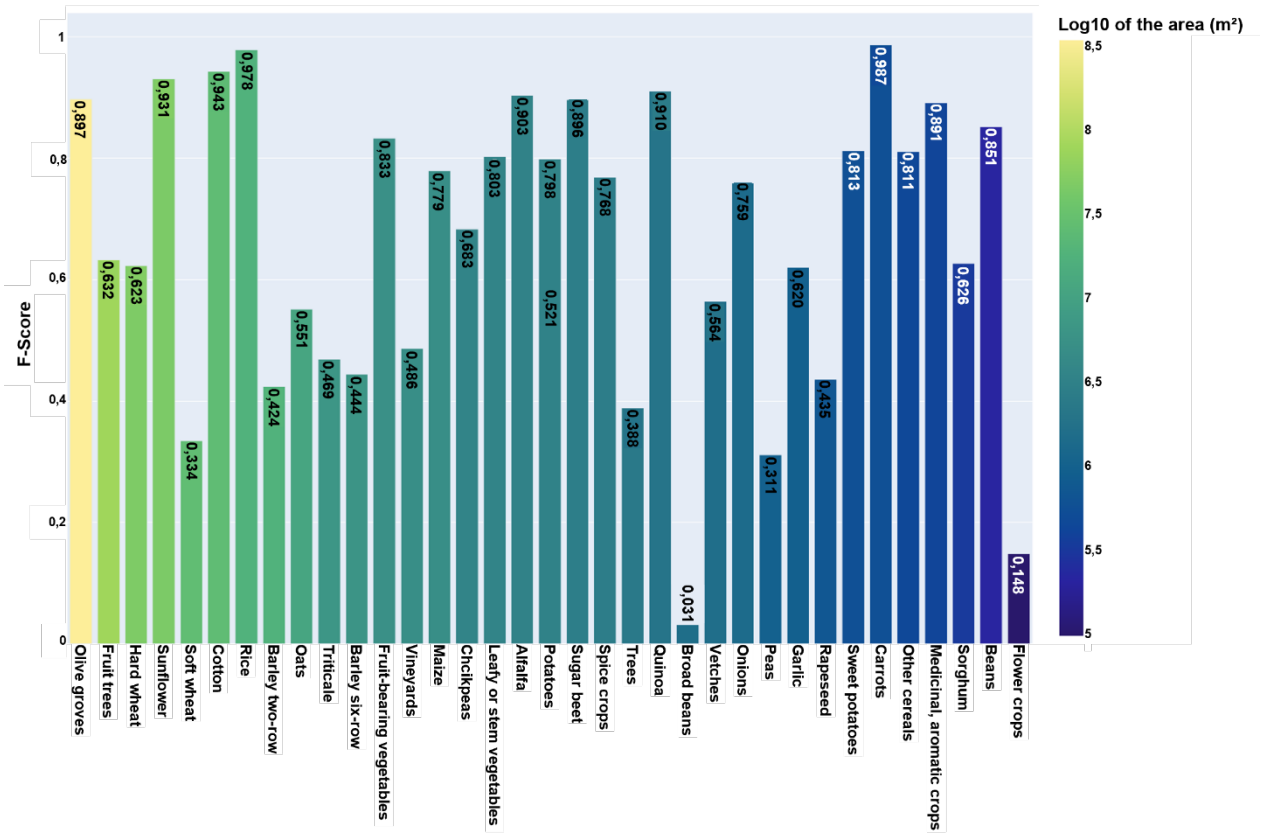


Figure 5-10. F-Score by class sorted by area (largest to smallest) of the crop type map in Andalusia

When moving to national scale, the crop type map counts 38 different classes, including distinc non-crop classes. The map is presented in Figure 5-11. The national crop type map was obtained following the same method as the regional maps, except that a stratification of 4 distinct strata was used.

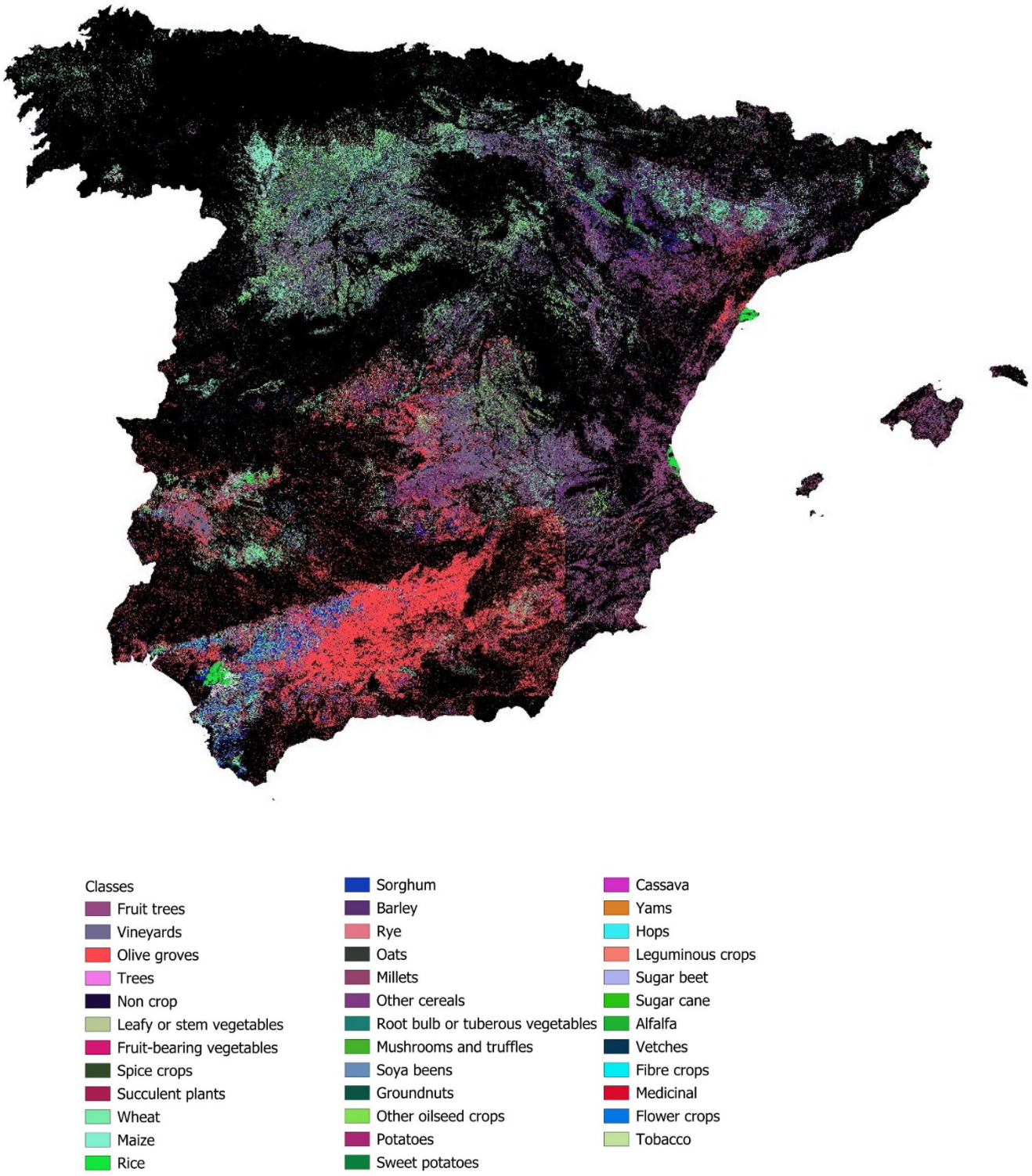


Figure 5-11. National scale crop type map in Spain (non-distinctive non crop class)

A quantitative assessment of the classification was conducted for each stratum, based on the confusion matrix of the model and on the accuracy metrics derived from it. The F-score sorted by

class prevalence and confusion matrix are shown for each stratum below. ***Stratum 1*** is situated along the Atlantic coastline and is characterized by a landscape that is notably dominated by maize, wheat and fruit trees, which were well classified. The remaining crops were found to be of significantly minor proportion and therefore were less classified. ***Stratum 2*** spans over Castilla-y-Leon, Aragon and Catalonia. These autonomous communities are notable for their cereal, perennial fruit tree, fodder and oilseed crop production. The F-scores for each crop show that the primary source of confusion is observed between barley and wheat and between olive groves, vineyards and orchards. These confusions are expected when considering the thematic proximity of these classes. Except for rice, the minor classes were less accurately classified, the confusions often taking place between similar classes (e.g. between hard wheat and soft wheat, two-rows barley and six-rows barley, and so forth). ***Stratum 3*** spans over Castilla-la-Mancha, Murcia, and Valencian Community, where the major crops are barley and perennial fruit trees. The same type of thematic confusion between barley and wheat as in stratum 2 is observed and is geographically located over the overlapping areas between the northern part of stratum 2 and stratum 3. Commissions are also observed for oat, which is mixed with barley. ***Stratum 4*** spans over Andalucia and Extremadura and its agricultural landscape is largely dominated by perennial fruit trees (mainly olive groves). Omissions of fruit trees classified as olive groves were noted as well as expected thematic classification errors between soft and hard wheat. Olive groves are well classified but generate the most confusion with minor classes due to the proportion of crops in the stratum.

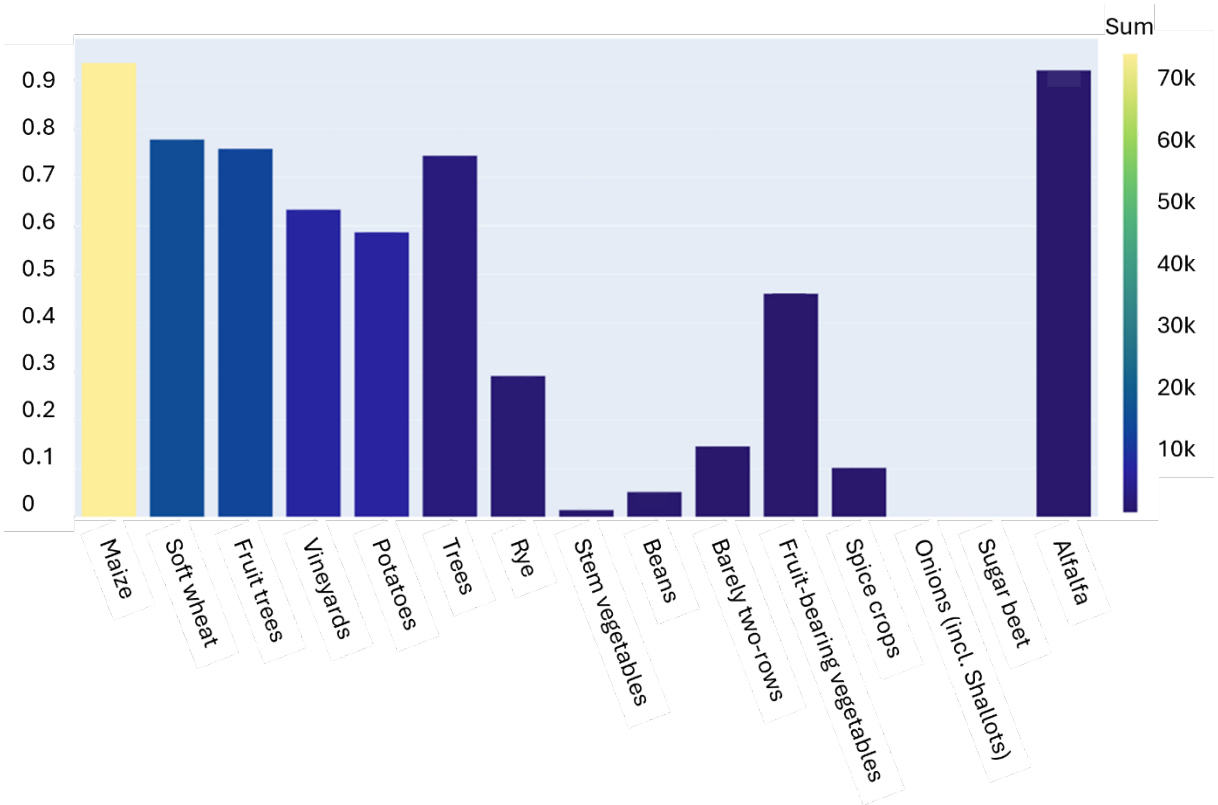


Figure 5-12. F-Score by class sorted by area (largest to smallest) of the crop type map in stratum 1

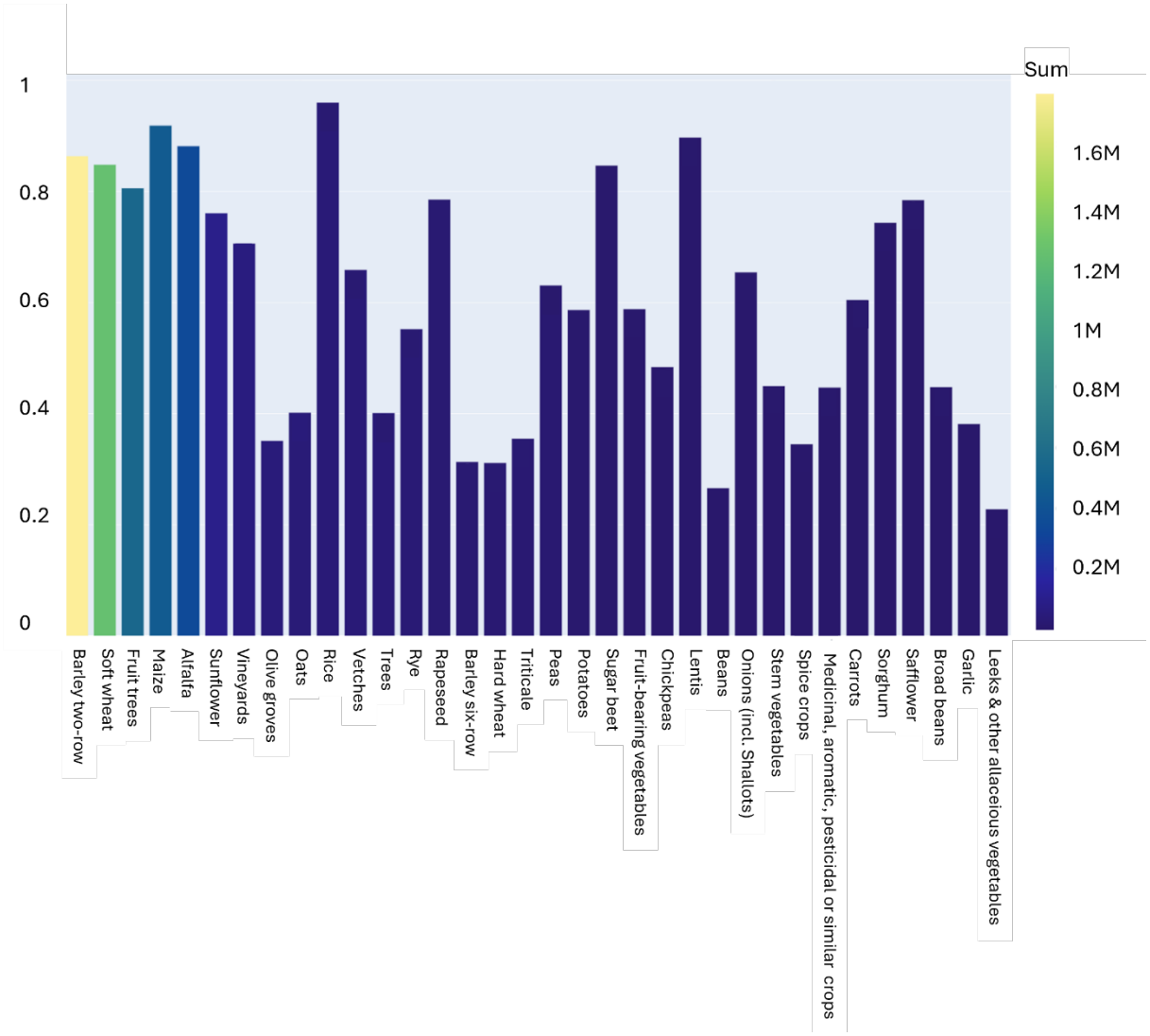


Figure 5-13. F-Score by class sorted by area (largest to smallest) of the crop type map in stratum 2

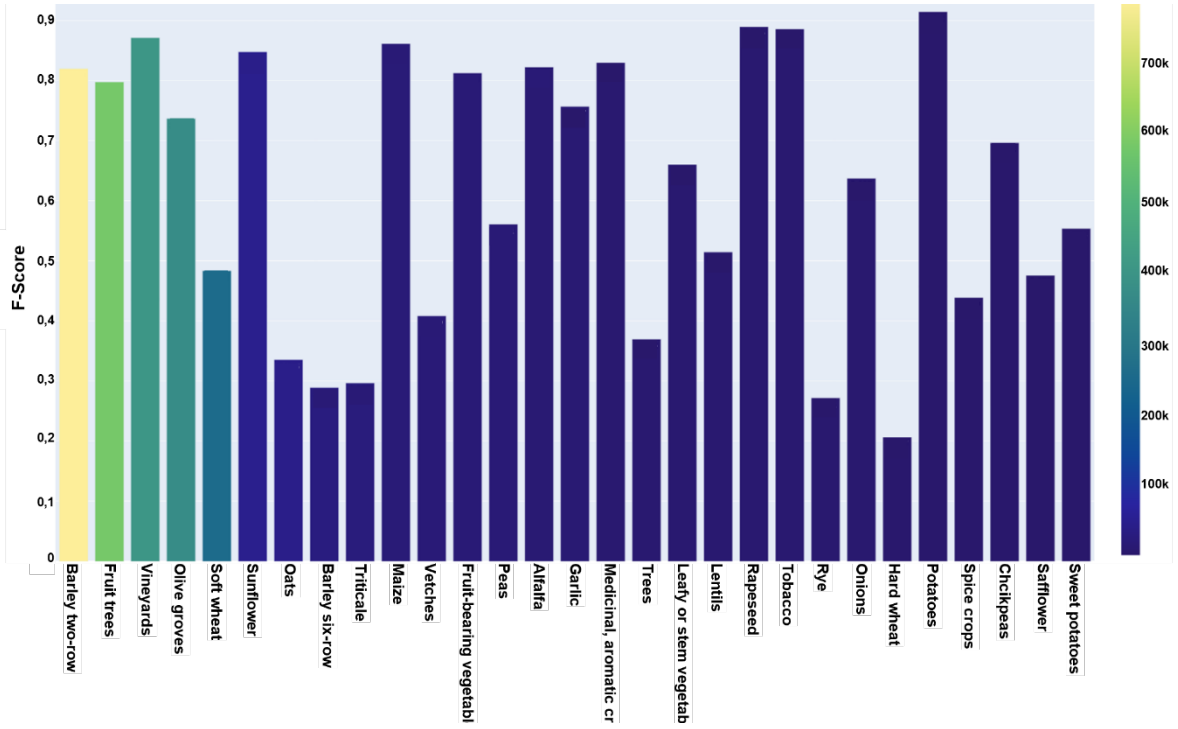


Figure 5-14. F-Score by aggregated class sorted by area (largest to smallest) of the crop type map in stratum 3

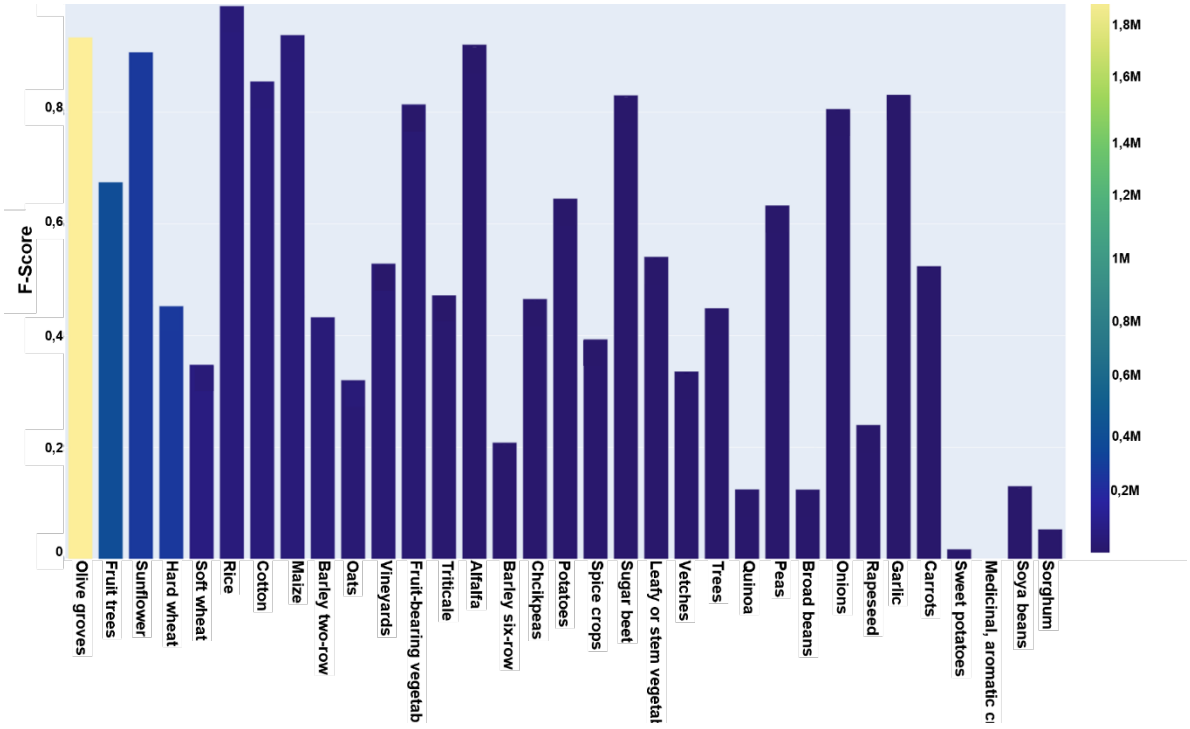


Figure 5-15. F-Score by aggregated class sorted by area (largest to smallest) of the crop type map in stratum 4

In terms of yield estimation, the demonstration in Spain focused on the two-row barley plots recorded in the ESYRCE survey in the region of Castilla y Leon.

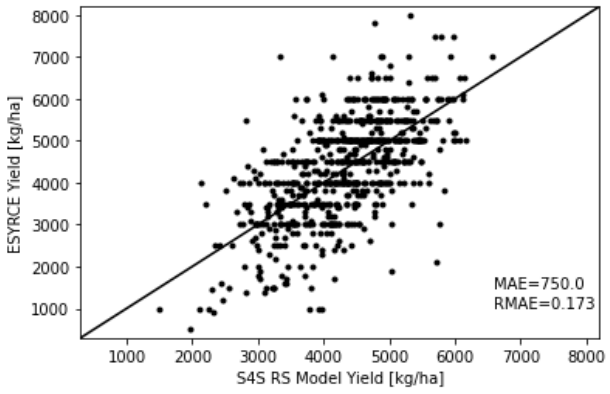
30% of the ESYRCE parcels containing a yield value, randomly selected, are removed from the dataset. Two estimation models based on the remaining 70% are then compared.

The first model (null model), currently used by the NSO, is based on the yield value measured in the field during the incomplete survey. The aggregation at the provincial level is done by weighting the area. In this way, the agricultural production of each field visited in each province is summed. Their respective average yield is equal to the sum of the production of their field divided by the total area of the visited fields in the province.

The second model (RS model) uses the Sen4Stat field-level yield estimation module. The retained data (70%) from ESYRCE are used to train a regression model using the yield explanatory variables derived from the Yield Characteristics module. The regression model was then applied to all fields (70%+30%) to increase the amount of data used for aggregation. The estimates on the training plots (70%) provide information on any biases in the estimates for each province. These biases were used to correct the model estimates.

The selected algorithm was gradient boosting regressor (sklearn default setting) and all Sen4Stat features were used without pre-selection. 10 repetitions of this method were carried out, reiterating the 70-30 split. Table of Figure 5-16 shows the average performance of the 10 regressions and their standard deviations. Graph of Figure 5-16 displays the performance of a randomly selected model from the 10 replications. In both cases, the performance is assessed on the basis of the 30% of plots not used for calibration.

Finally, all yield data referenced in ESYRCE were also aggregated (weighted by area) at the provincial level and used as a baseline for comparing the estimation models in the study.



	Mean	sd
MAE	744.5	24.6
RMAE	0.172	0.004

Figure 5-16. Performance of the Sen4Stat RS model estimation produced by the 10 repetitions of the 70/30 dataset partition, graphical comparison between one set of estimation and the ESYRCE reference yield.

The average provincial yields calculated with both models, their standard deviations and the MAE of the ten repetitions are presented in Table 3-2 and compared with the baseline model.

Table 5-1. Yield estimation of the provinces of Castilla-y-Lèon (kg/ha) given by ESYRCE and both models (Null and RS). The average yield, the standard deviation, and the mean absolute error computed on the ten repetitions of estimation are presented.

	ESYRCE		Null Model (10x)				S4S RS Model			
	N	Yield	N	Mean	Sd	MAE	N	Mean	Sd	MAE
Àvila	151	4250.2	107	4241.5	83.0	84.7	150	4232.4	34.9	37.9
Burgos	446	4852.4	315	4826.8	64.9	69.6	446	4764.3	38.2	88.1
Leòn	52	3792.7	37	3822.0	103.8	109.7	52	3817.5	57.0	59.2
Palencia	304	4585.6	211	4602.1	32.3	39.2	302	4557.5	17.0	29.9
Salamanca	122	4204.3	87	4155.8	63.1	81.5	122	4155.8	57.9	72.3
Segovia	294	4169.5	206	4168.0	52.5	52.8	294	4134.1	35.4	50.1
Soria	275	3617.5	192	3640.1	35.2	40.3	275	3542.6	26.8	74.9
Valladolid	460	4588.2	320	4574.6	37.8	41.4	459	4531.1	26.5	57.1
Zamora	206	4600.0	142	4586.8	65.0	67.2	204	4569.1	54.7	60.4
Castilla Y Leòn	2310	4437.2	1617	4426.5	16.5	20.8	2304	4391.9	14.0	45.3

Finally, the last EO product generated for the demonstration in Spain is a map of irrigation. To do so, we developed a classification algorithm which combined the ESYRCE dataset irrigation attribute (Figure 5-17) with farmers' yearly parcel declarations and parcel delineations. The national scale irrigation map presented in Figure 5-18.

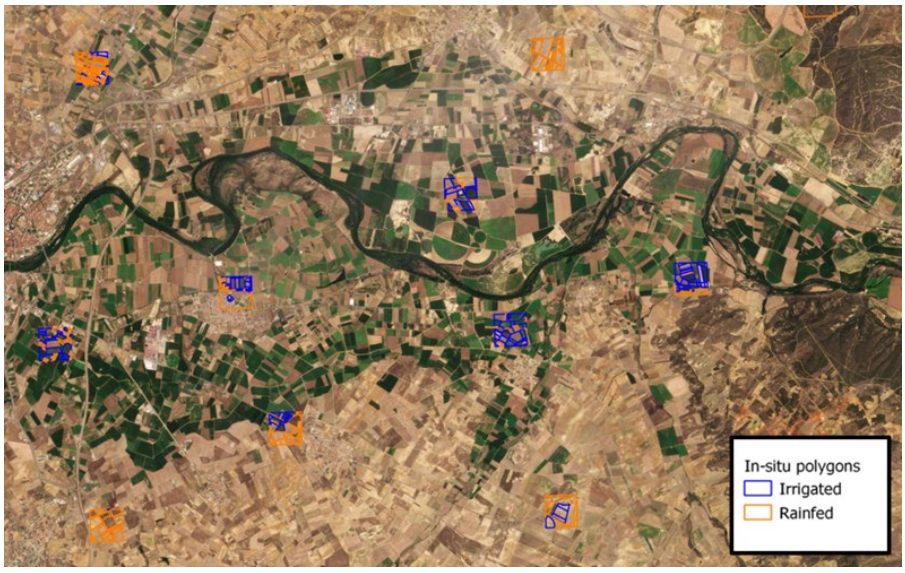


Figure 5-17. Sample of polygons with rainfed and irrigated attributes in the ESYRCE dataset in Castilla-y-Leon (2020)

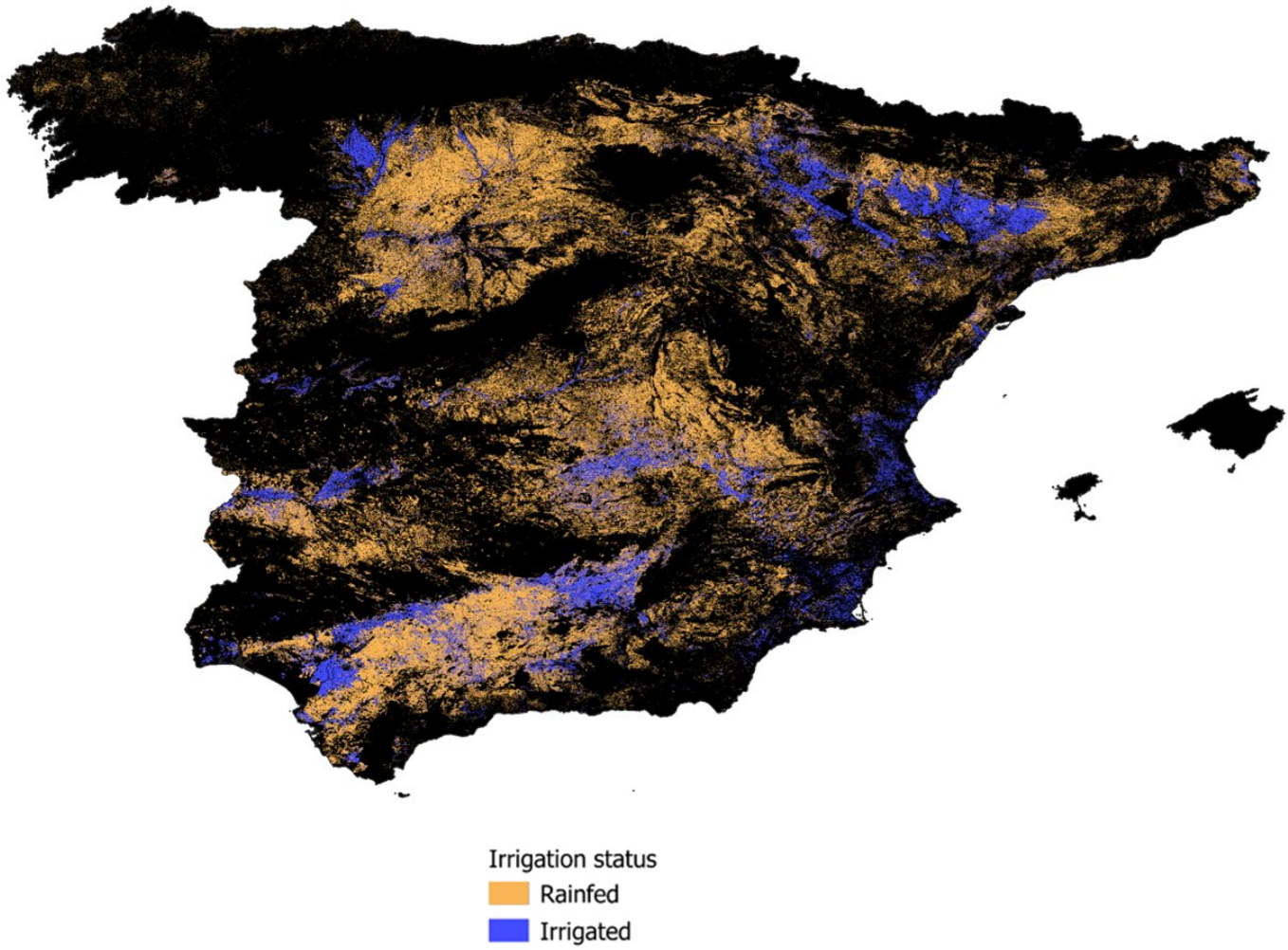


Figure 5-18. Irrigation map in Spain

Accuracy metrics of the classification were computed separately for each crop type in Spain (Figure 5-19). Good F-scores were attained for the major classes, including those with irrigation proportions that were neither notably abundant nor scarce.

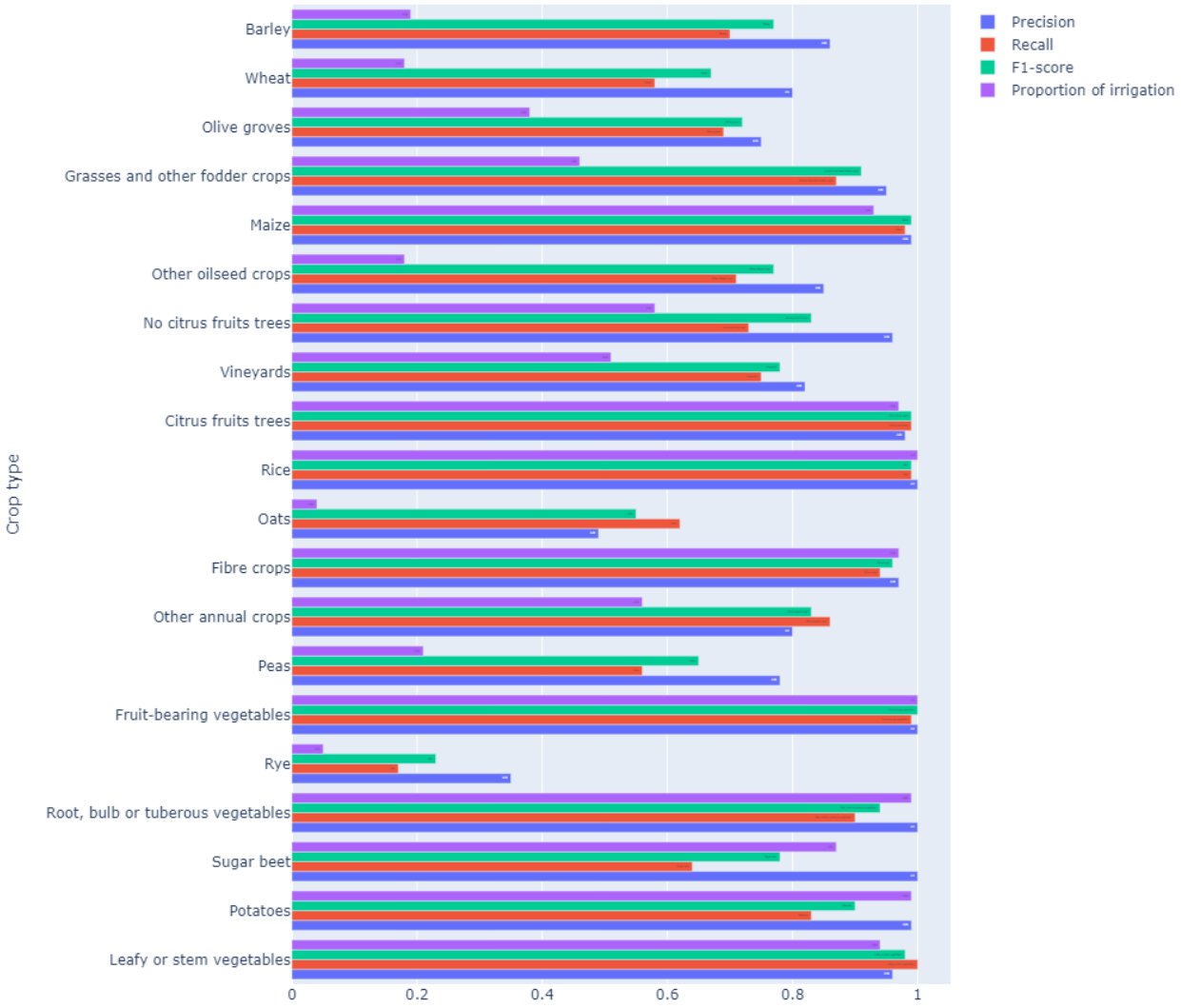


Figure 5-19. F-Score, Precision, Recall and irrigation proportion in validation data by class sorted by area (largest on top to smallest on the bottom) of the irrigation map in Spain.

5.1.2 Use cases

5.1.2.1 Cost-efficiency

Figure 5-20 presents the results for the cost-efficiency use case in Castilla y Leon, for the three main crops which are wheat, maize and sunflower. For each crop, the figure shows the acreage estimate based on ground data only (ESYRCE being the name of the agricultural survey) and on the coupling of ESYRCE with EO data. More interestingly, the figure also provides the confidence intervals around these estimates. The EO impact is a systematic reduction of the interval, and thus

of the sampling error (highlighted in Figure 5-21). As a result, the relative efficiency of the coupling between EO and ESYRCE datasets is high.

	Data	Acreage (hectares)	Uncertainty			Relative efficiency
			95% Confidence Interval (hectares)		Sampling Error (CV%)	
			Limits	Amplitude		
Barley (F-Score: 0,875)	**Ground (ESYRCE)**	980 081	Lw: 924 173 Up: 1 037 461	113 288	2,95	-----
	Gound+RS	923 026	Lw: 899 364 Up: 946 689	47 325	1,31	5,73
Maize (F-score: 0,970)	**Ground (ESYRCE)**	210 558	Lw: 176 727 Up: 244 389	67 662	3,18	-----
	Ground+RS	132 901	Lw: 124 606 Up: 141 195	16 589	8,2	16,64
Wheat (F-score: 0,880)	**Ground (ESYRCE)**	877 717	Lw: 827 030 Up: 928 404	101 374	2,94	-----
	Ground+RS	812 088	Lw: 787 877 Up: 836 299	48 422	1,52	4,38

Figure 5-20. Cost-efficiency use case in Castilla y Leon (Spain, 2020)

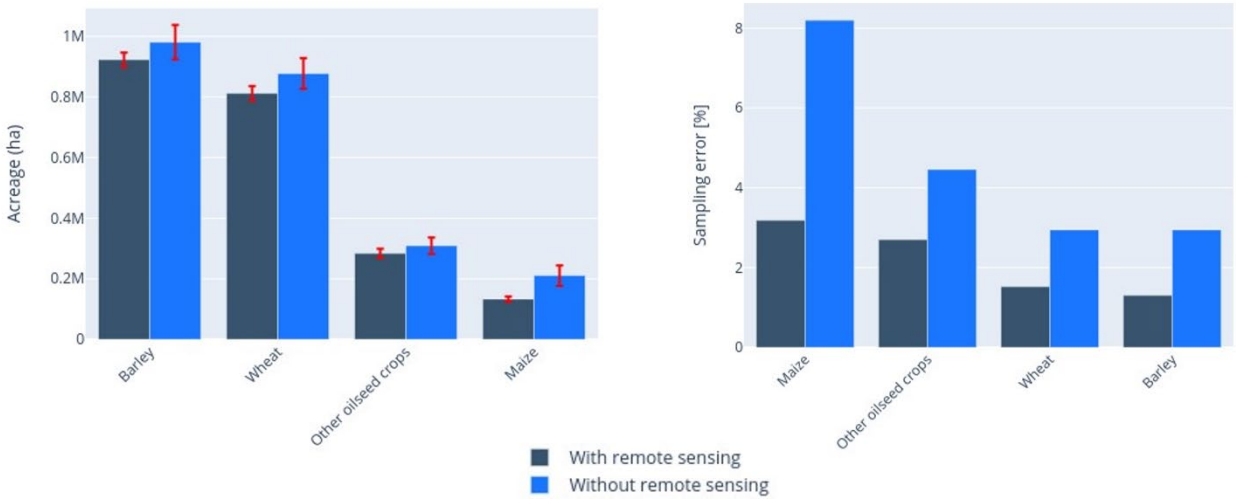


Figure 5-21. Acreage estimation of the predominant classes in Castilla y Leon (2020) presented with their confidence interval (left) and sampling error (right), with (grey) and without (blue) EO data

Similarly, Figure 5-22 presents the results for the cost-efficiency use case in Andalusia, for the three main crops which are olive groves, wheat and sunflower. Here also, the EO impact is a systematic reduction of the interval, and thus of the sampling error, as shown in Figure 5-23, and it can be concluded that the relative efficiency of the coupling between EO and ESYRCE datasets is high.

	Data	Acreage (hectares)	Uncertainty			Relative efficiency
			95% Confidence Interval (hectares)		Sampling Error (CV%)	
			Limits	Amplitude		
Olive groves (F-Score: 0,897)	Ground (ESYRCE)	1 624 187	Lw: 1 529 848 Up: 1 718 525	188 677	2,96	-----
	Ground+RS	1 740 863	Lw: 1 693 061 Up: 1 788 664	95063	1,40	3,896
Wheat (F-score: 0,834)	Ground (ESYRCE)	398 357	Lw: 358 596 Up: 438 119	79523	5,09	-----
	Ground+RS	391 056	Lw: 369 952 Up: 412 161	42209	2,75	3,55
Oilseed crops (F-score: 0,931)	Ground (ESYRCE)	229 049	Lw: 197 700 Up: 260 399	62699	6,98	-----
	Ground+RS	210 232	Lw: 196 088 Up: 224 377	28289	3,43	4,91

Figure 5-22. Cost-efficiency use case in Andalusia (Spain, 2020)

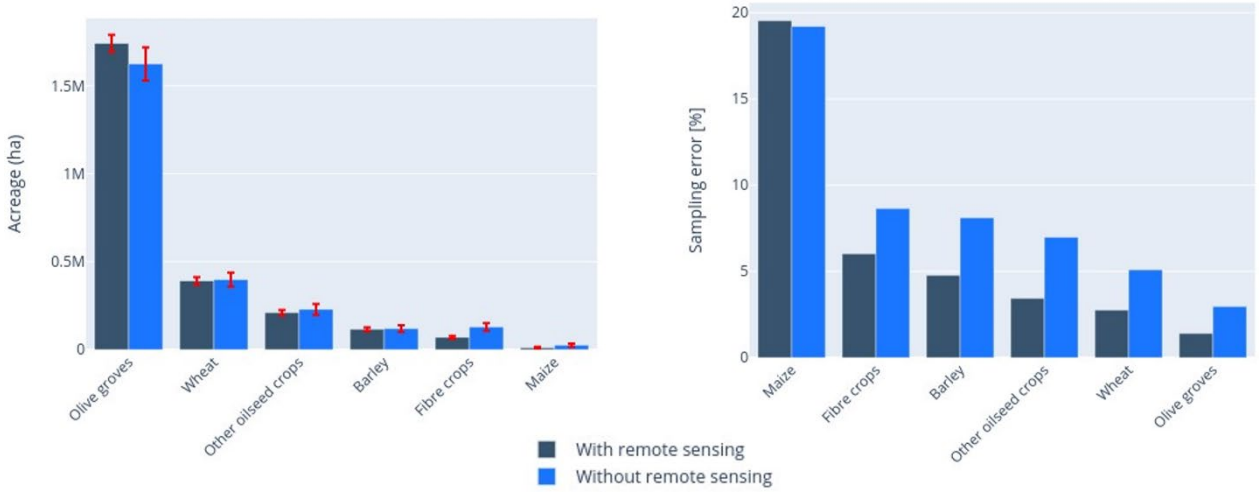


Figure 5-23. Acreage estimation of the predominant classes in Andalusia (2020) presented with their confidence interval (left) and sampling error (right), with (grey) and without (blue) EO data

It is planned to do the same analysis with the national-scale land cover map but the quality at the current moment is not good enough, mainly due to artefacts. This exercise will be carried out during the project extension.

The cost-efficiency use case was also considered for the yield estimation, but in a slightly different way. Indeed, the estimation of the yield was set up to show that estimating yield on a larger sample of data (i.e. with EO data) can improve confidence in aggregate statistics by virtually increasing the number of data points collected in the survey. We showed that the model incorporating the remote sensing variables, although not capable of accurately estimating yields at the plot scale, can be used to synthetically augment data in poorly represented statistical units and thus improve the robustness of estimates at this scale. Since the use of the RS model greatly reduces the standard deviation of the estimates, it is likely that improving the performance of the estimation model at

the field level would allow the number of samples to be measured in the field to be reduced while maintaining the same confidence in the estimates.

5.1.2.2 Domain and small area estimation use case

The test case was demonstrated in Castilla y Leon. Figure 5-24 shows the estimates of barley acreage for four provinces in Castilla y Leon. It can be seen that the estimates are quite similar without and with EO data, but that the sampling error is significantly reduced when EO data is integrated. This is illustrated in a different way in Figure 5-25, emphasizing the decrease of the sampling error.

REGION	ESYRCE		ESYRCE+EO		Relative Efficiency
	Acreage (has.)	Error (CV%)	Acreage (has.)	Error (CV%)	
León	6853.2	24.15	6834.5	16.20	2.3
Palencia	88602.0	7.36	90535.3	3.33	4.7
Valladolid	128209.5	5.57	119707.4	2.66	5.1
Zamora	12324.2	17.71	10948.2	8.16	6.4
TOTAL AREA	235989.1	4.37	228028.5	1.98	5.2

Figure 5-24. Acreage estimates for barley in 4 provinces in the region of Castilla y Leon (2020), obtained without (ESYRCE columns) and with (ESYRCE+EO columns) EO data

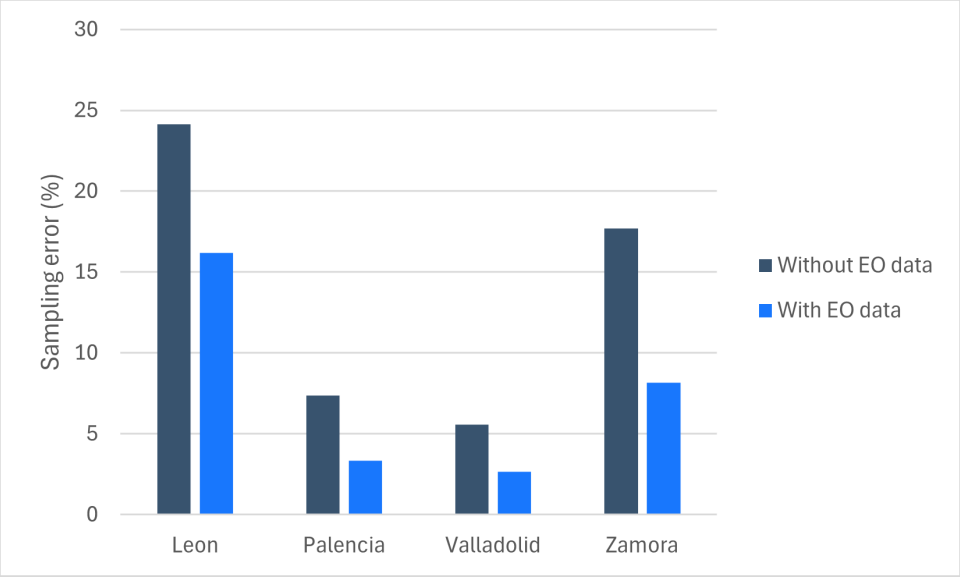


Figure 5-25. Reduction of the sampling error thanks to EO data for acreage estimates at provincial level – the case of barley in Castilla y Leon (2020)

The positive impact of EO data has also been proven in the estimation of barley acreage statistics at the level of the municipalities. For these administrative units, it is not possible to obtain statistics

using only ground data: the very low amount of samples would induce very low accuracy of the statistics. But using EO data, acreage estimates can be obtained with a sampling error remaining reasonable (i.e. less than 20%). This is shown in Figure 5-26.

Municipality		Acreage	
		Has.	Error (CV%)
49020	Belver de los Montes	212.96	29.1
49043	Castroverde	2914.22	8.0
49156	Pinilla de Toro	963.30	10.0
49168	Quintanilla del Monte	466.65	20.3
49219	Toro	615.91	14.0
49235	Vezdemarbán	1358.22	12.6
49250	Villalpando	560.05	39.1
49252	Villamayor de Campos	1056.23	11.1
49260	Villanueva del Campo	784.03	13.2
49263	Villar de Fallaves	844.16	11.0
49267	Villardondiego	516.40	11.5
49270	Villavendimio	656.07	10.4
Total Zamora		10948.2	8.16

Figure 5-26. Acreage estimates for barley in the municipalities of the Zamora province (2020), obtained thanks to the integration of EO data

As for the yield, the yield estimation was also applied at provincial level, with conclusive results as it can be shown in Table 5-2.

Table 5-2. Comparison of ESYRCE Yield estimation given by Province with the S4S RS yield estimation model applied on all the barley fields of the survey.

	ESYRCE		S4S RS Model	
	N	Yield [kg/ha]	N	Yield [kg/ha]
Àvila	151	4250.2	330	4297.8
Burgos	446	4852.4	2530	4678.7
Leòn	52	3792.7	276	4077.1
Palencia	304	4585.6	1068	4541.2
Salamanca	122	4204.3	279	4193.0
Segovia	294	4169.5	775	4327.8
Soria	275	3617.5	662	3611.5
Valladolid	460	4588.2	1556	4676.4
Zamora	206	4600.0	624	4462.6
Castilla Y Leòn	2310	4437.2	8100	4483.0

5.1.2.3 Sampling design use case

The irrigation map aims at updating the stratification that is used by the NSO for defining the sampling frame.

5.2 Demonstration in Senegal

5.2.1 EO products

The first cycle of the demonstration focuses on a regional pilot area, which is the department of Nioro du Rip (Figure 5-27). The department of Nioro du Rip is one of the 46 departments of Senegal and one of the 3 departments of the Kaolack region. Its area is of 2302 km².

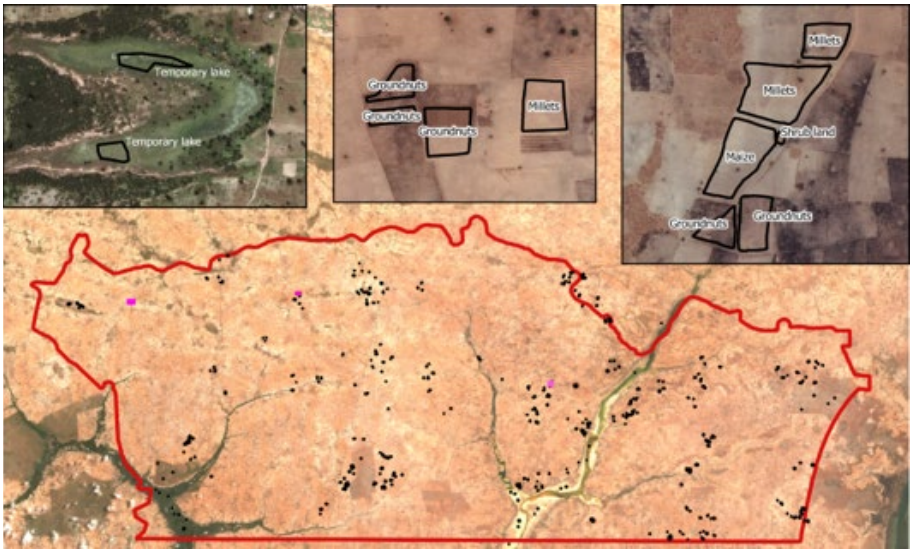


Figure 5-27. Area of Interest in Senegal (cycle 1)

The reference year for this first cycle of the demonstration is 2021, during which a dedicated field data campaign is implemented between August and November 2021. This field campaign corresponds to a first use case, which is the adjustment of survey protocols and the reduction in the uncertainty of the collected data. The objective of this field campaign was to show the added-value of registering parcel boundaries instead of points and to compare the accuracy of these boundaries registered with a tablet and with a GPS device. A total of 247 plots remained after the data collection and the quality control. 50 additional polygons of non-crop classes were created by photo-interpretation and the whole dataset was split between calibration and validation (Figure 5-28). The distribution of observations is very uneven for the different crops and insufficient for maize.

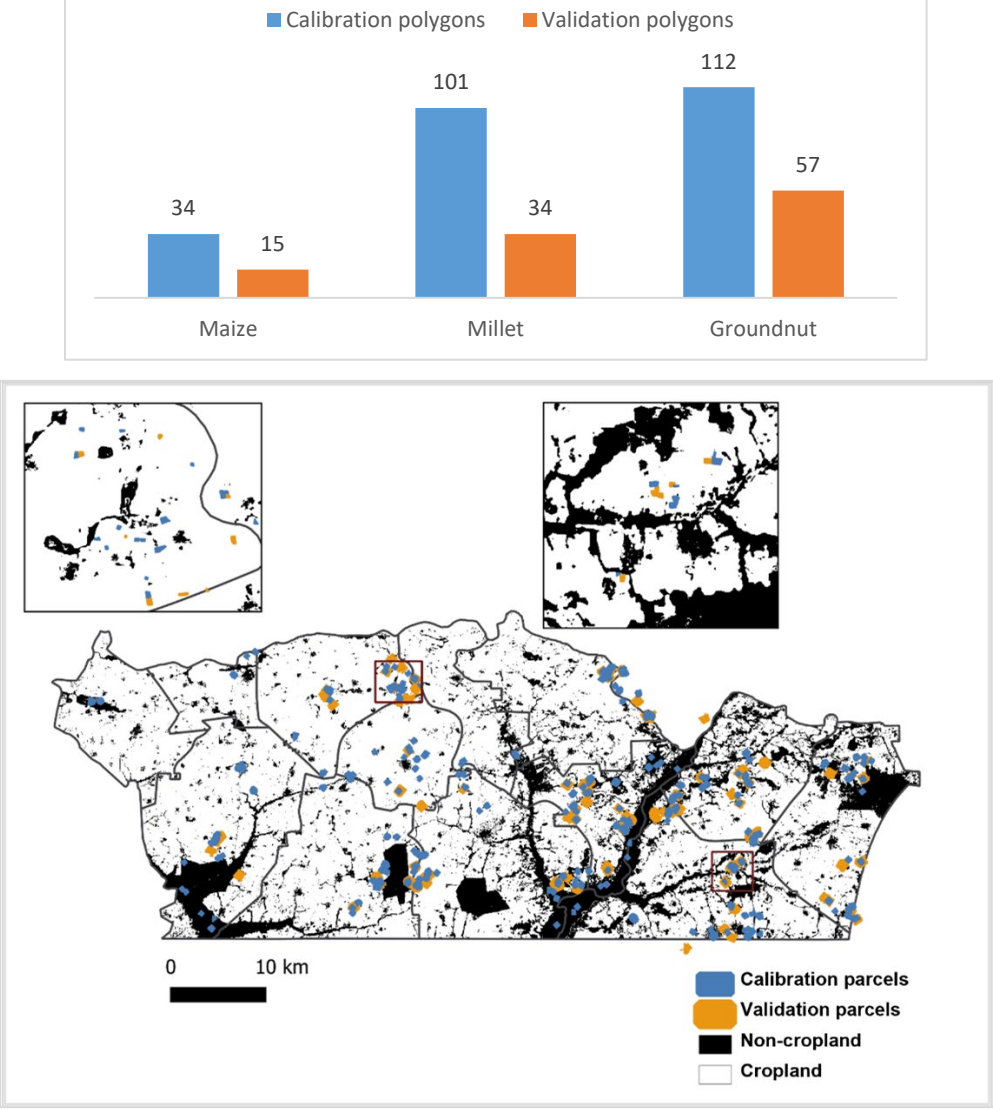


Figure 5-28. Distribution of in-situ data into a calibration data set and a validation data set

The crop type map was obtained using both S2 and Sentinel-1 (S1) time series (Figure 5-29).

The confusion matrix is presented in Figure 5-30. The overall accuracy of the crop mask is 97.1% and it is of 88.2% for the crop type. The F-Score values for cropland and non-cropland are 98% and 95% respectively. The F-score values are 54.8% for maize, 83.8% for millet, and 95.2% for groundnut. There is a very strong omission of maize. Millet is both contaminating and omitted.

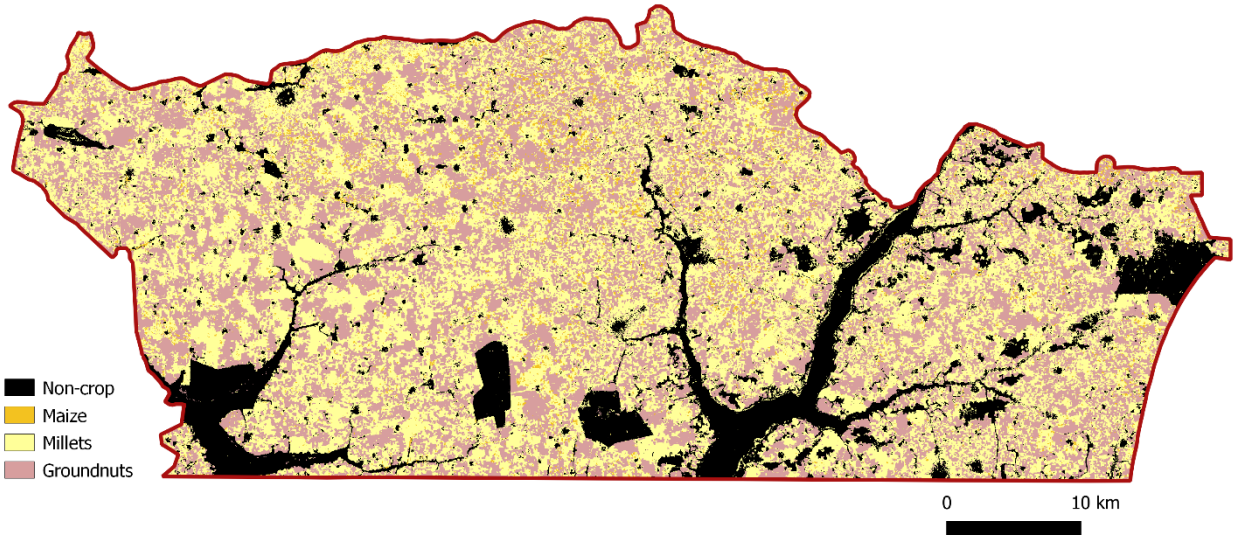


Figure 5-29. Crop type classification of the Nioro department based on S1 and S2 time series from May 1, 2021 to December 31, 2021

Expressed in number of pixels		Field survey: Non-crop	Maize	Millet	Groundnut	UA	Contaminations (%)	Omissions (%)
Crop type map	Non-crop	2205	0	34	17	97.7	2.3	9.3
	Maize	0	325	10	0	97.0	3.0	47.9
	Millet	202	265	2755	3268	80.5	19.5	12.6
	Groundnut	25	34	354	3487	88.8	11.2	6.3
	PA	90.7	52.1	87.4	93.7			

Figure 5-30. Confusion matrix (expressed in number of pixels) for the crop type map, with contamination and omission values for each crop, UA as user accuracy and PA as producer accuracy

The same situation happened for the second cycle of demonstration, which focused on a larger area of interest, corresponding to 6 distinct administrative units: the regions of Kolda and Tambacounda and the departments of Nioro, Mbacke, Koungheul and Dagana. Here again, the field campaign corresponds to one use case. This adjusted protocol aimed at enhancing the compatibility of collected data with EO data. First, field data collection (based on Garmin 64 GPS devices) involved the delineation of parcel boundaries, with additional GPS points recorded via SurveySolutions software in each parcel to ensure data consistency between the GPS traces and the statistics database. Second, GPS points were taken in the yield crop cutting subplots. Compared to the first cycle, the use case went one step further because the adjusted protocol was applied in autonomy by the NSO's enumerators and the quality control of the data was also carried out by the NSO (and by us in parallel).

Table 5-3 provides a quantitative insight about this field campaign, providing the collected data (in SurveySolution and in the form of GPX) and the remaining data after the quality control.

	Number of samples in AAS	Number of GPX	Number of samples after quality control
Total	12827	3925	2215
Dagana		10	8
Kolda		785	430
Kongheul		1231	598
Mbacke		1264	678
Nioro		462	334
Tambacounda		173	167

Table 5-3. Number of data collected and after quality control

The distribution of the crops within the collected data is shown in Figure 5-31.

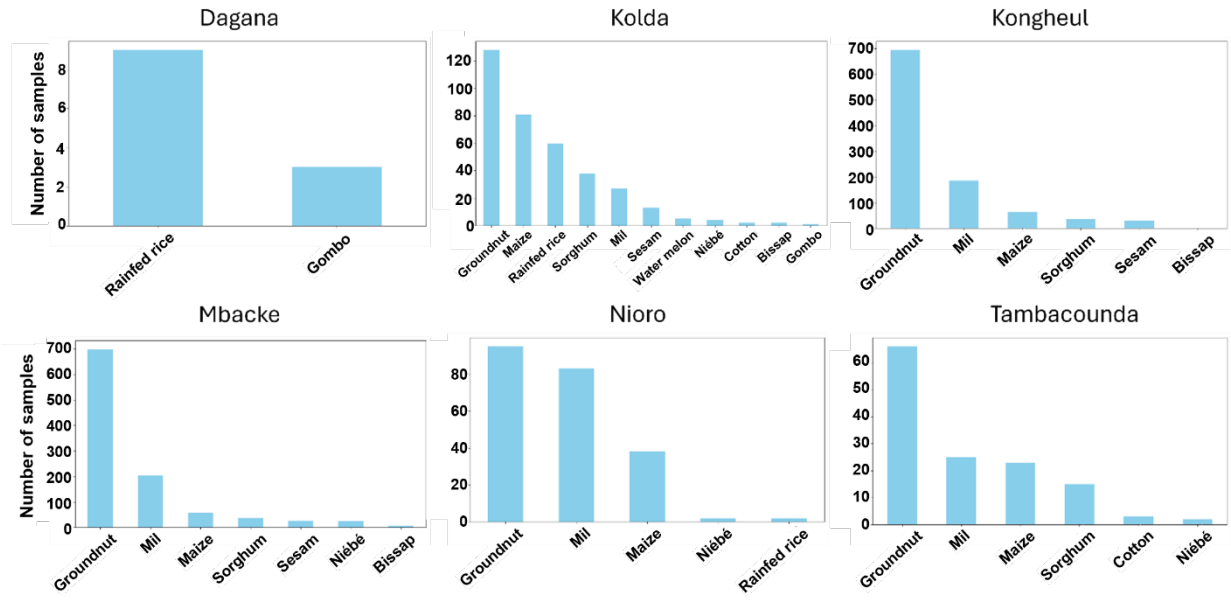


Figure 5-31. Crop distribution in the dataset collected in each department during the 2023 field campaign

The collected in situ data were used to calibrate the random forest algorithm included in the Sen4Stat toolbox.

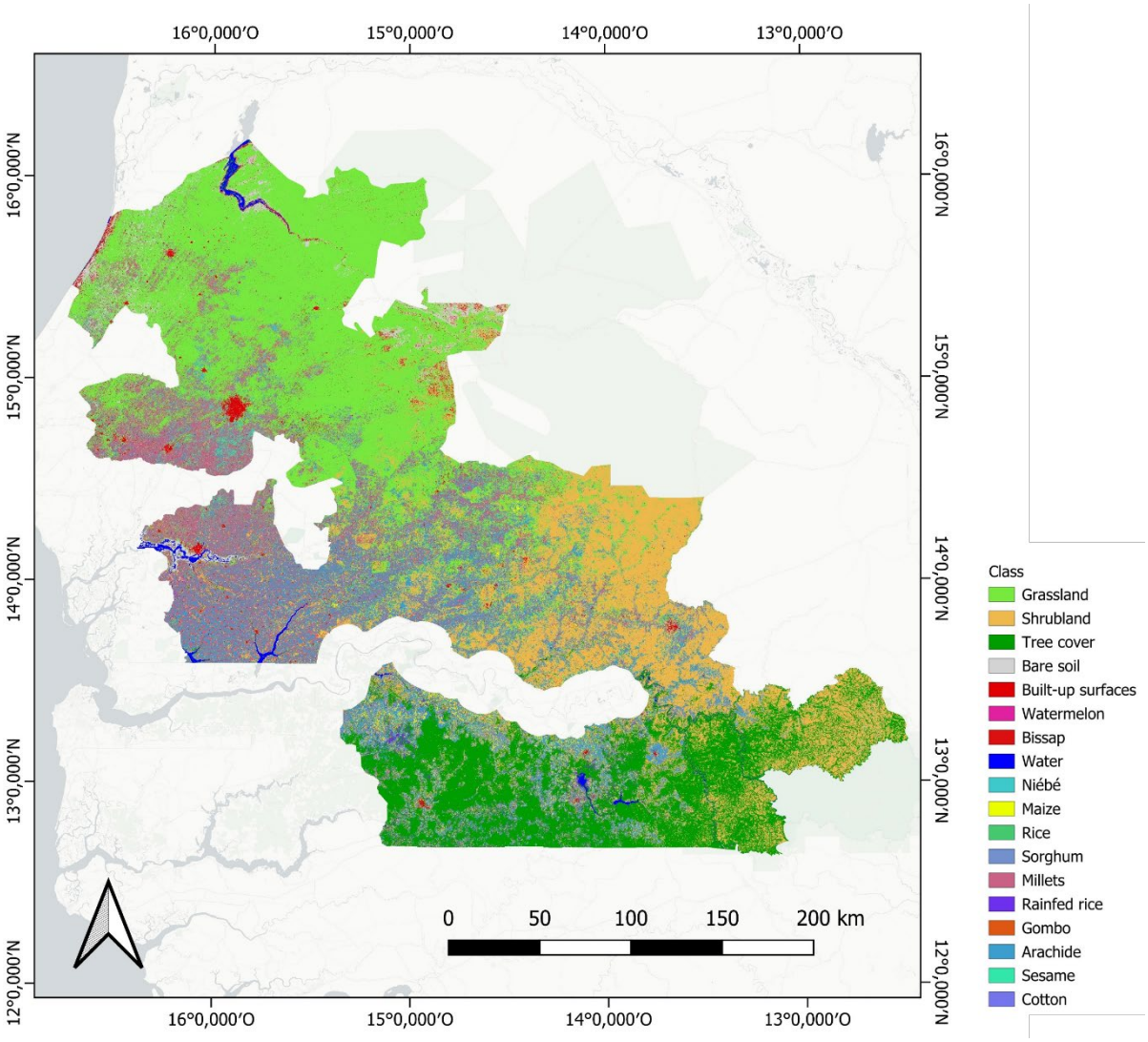


Figure 5-32. Crop type map 2023 over the 6 pilot departments

When looking at individual crop types (Figure 5-33), the ones that are best classified are groundnut (F-Score of 0,82), millets (F-Score of 0,72) and rice (F-Score of 0,972). The aggregation at the crop group levels allows significantly increasing the accuracy, showing a good accuracy for both the oilseed crops and cereals.

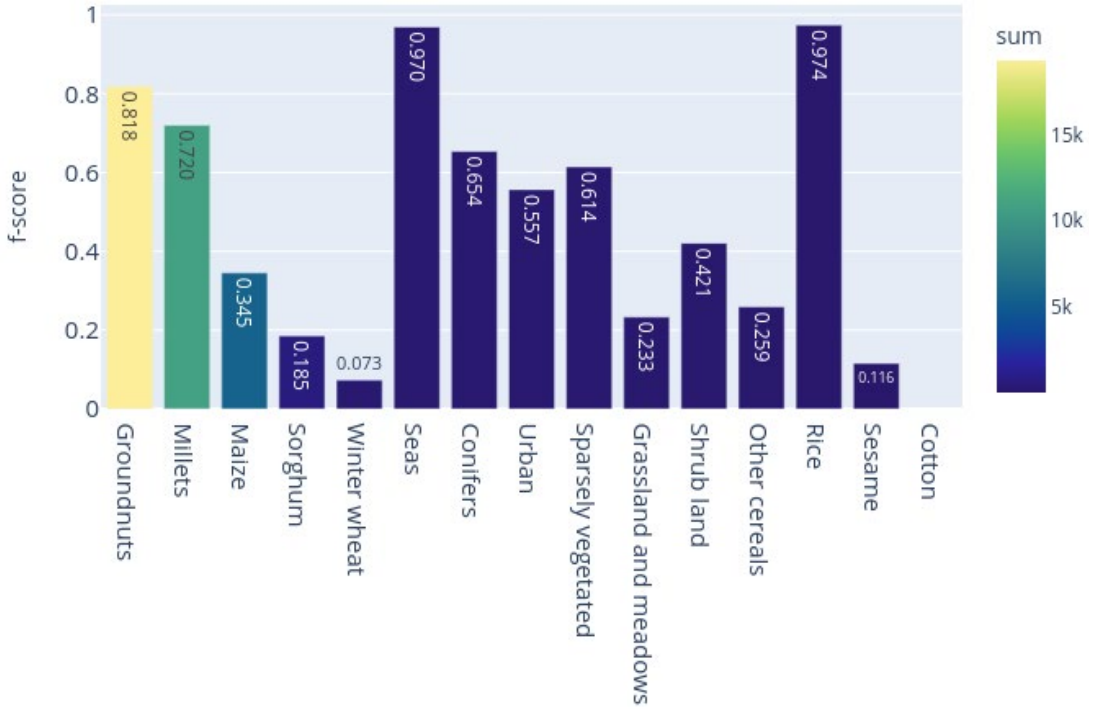


Figure 5-33. Accuracy metrics of individual crop and land cover types

5.2.2 Use cases

5.2.2.1 Survey protocol and uncertainty reduction in survey data

During the first field campaign, a comparative analysis was conducted between the parcel boundaries recorded by the tablet and by the GPS device.

The tablet dataset seems to be less accurate than the one recorded with the GPS and has a bias that underestimates the baseline value (Figure 5-34 and Figure 5-35). The trendline of the automatic ODK measurements (Figure 5-35 top, line orange) shows a small deviation from the line of the reference measurements (green) and thus a good overall accuracy, while for the ODK measurements in manual mode, the trendline deviates strongly from the reference (Figure 5-35 bottom, line orange). It is clear that the automatic point measurement provides a result that is much more accurate and closer to the reference result than the manual measurement. It avoids recurrent errors due to bad encoding of points, time to fix the GPS position of the tablet or problems in recording the points by the tool. Using the tablet in the automated mode is accurate enough to replace the GPS, and might allow to conduct the survey using one unique device.

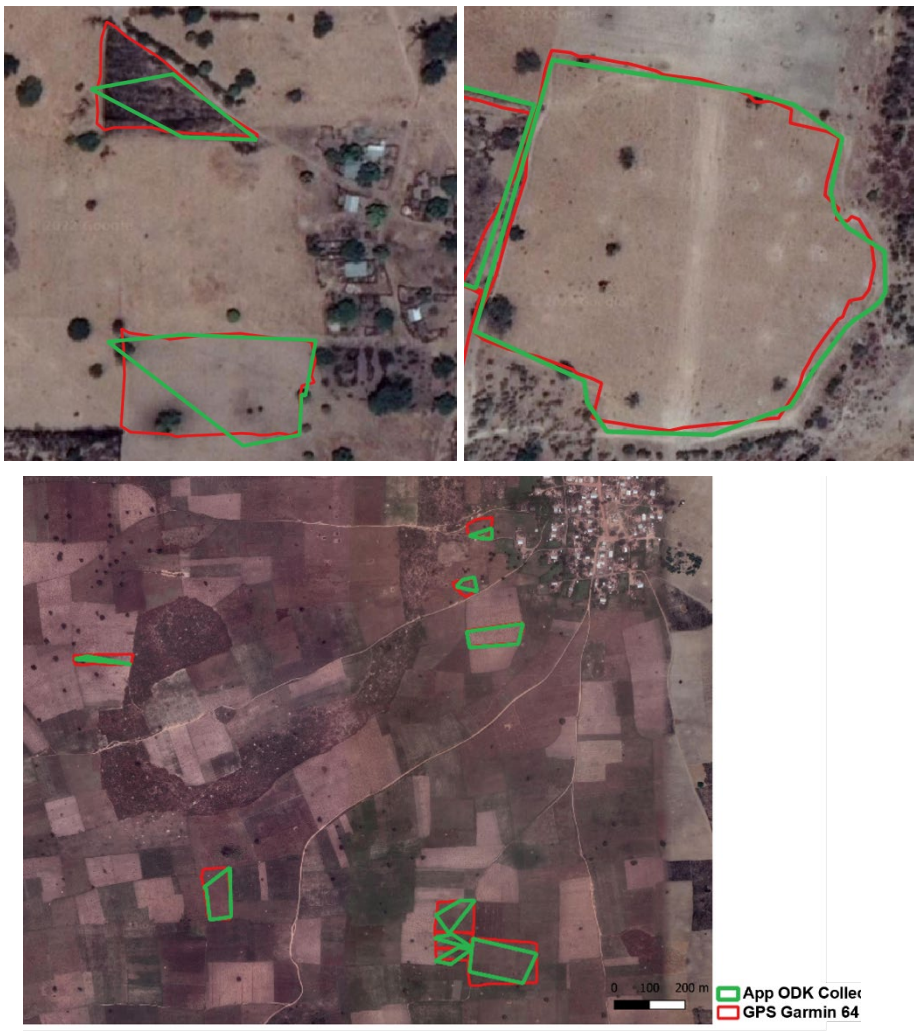
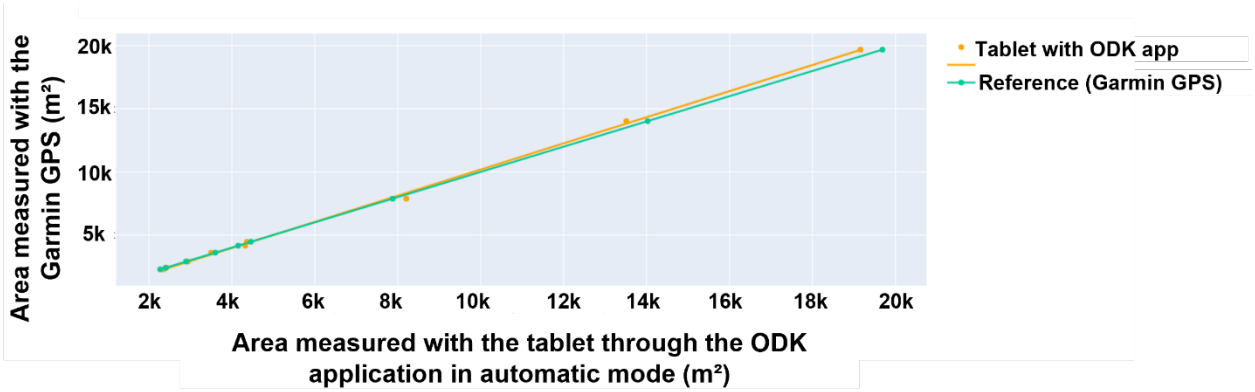


Figure 5-34. Parcel's polygons from Garmin GPS in red and from the tablet's GPS in green.



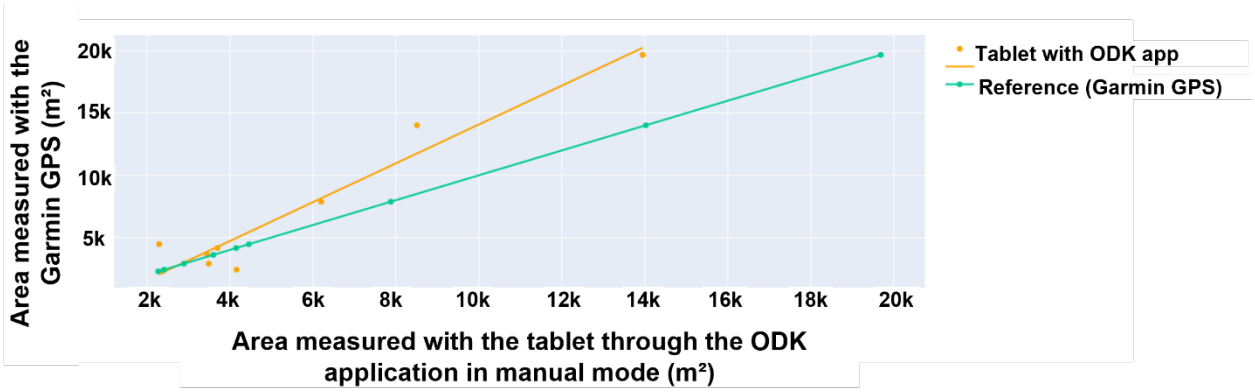


Figure 5-35. Comparison of the reference (Garmin 64 GPS) and the measurements made by the tablet via ODK Collect in automatic mode (top) manual mode (bottom)

5.2.2.2 Cost-efficiency

The implementation of the cost-efficiency use case in Senegal is more complex than in Spain, because the sampling design is not the same. While it is an area frame in Spain, this is a complex (4 stages) list frame of households in Senegal (Figure 5-36).

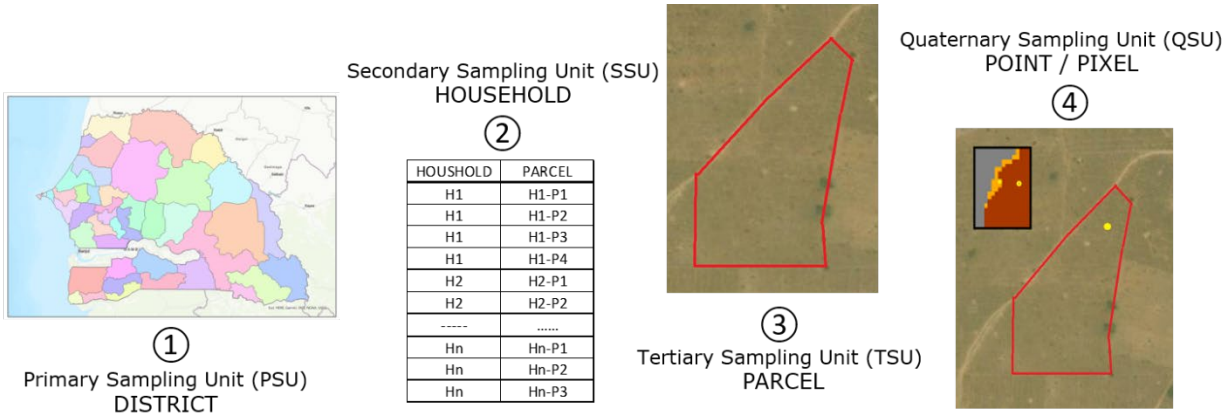


Figure 5-36. 4-stage list frame sample design in Senegal.

As a result, the integration of remote sensing and ground data cannot be done using linear models, but multinomial logit models are needed. These multinomial models deal with the uncertainties and generate probabilities that a pixel of a given class in the map is actually this given crop on the ground.

The cost-efficiency use case has been demonstrated in terms of crop acreage estimates using the crop type map generated over the department of Nioro. Figure 5-37 presents the crop acreage estimates in the department of Nioro, for the two main crops which are millet and groundnut while Figure 5-38 shows the efficiency of using the crop type map to support this estimation of crop acreages.

Crop type	Acreage (hectare)	Uncertainty				
		Standard error	Coefficient of variation (%)	Limits of 95% confidence interval		
				Lower	Upper	Amplitude
Millet	89215	3661.103	4.11	81978.88	96330.4	14351.52
Groundnut	78815	2923.94	3.71	73089.15	84550.98	11461.82

Figure 5-37. Crop acreage estimates using EO and ground data in Nioro (Senegal, 2021)

Crop type	Standard errors of proportion estimators		Relative efficiency of Classified RS data
	Using only ground data	Using ground & RS data	
Millet	3.37	1.90	3.13
Groundnut	3.34	1.52	4.80

Figure 5-38. Efficiency of using the crop type map for crop acreage estimation in Nioro (Senegal, 2021)

5.2.2.3 Domain and small area estimation use case

While it was not requested, the spatial disaggregation use case was also tested and successfully demonstrated: as shown in Figure 5-39, acreage estimates are available at the "arrondissement" levels with a reasonable error (expressed as the coefficient of variation).

Arrondissement	Millet		Groundnut	
	Acreage (has.)	Error (CV%)	Acreage (has.)	Error (CV%)
Medina Sabakh	20067.21	8.6	19765.36	7.3
Paoskoto	38316.02	5.3	35018.93	4.0
Wack Ngouna	30831.77	11.9	24030.71	10.7
Total Nioro	89215,00	4.11	78815,00	3.71

Figure 5-39. Crop acreage estimates at the district (arrondissement) level in Nioro (Senegal, 2021)

6 System uptake

6.1 Pilot countries

During the project, the main effort in terms of system uptake has been targeted towards the pilot NSOs, with dedicated trainings. However, it shall be noted that the NSOs usually don't have staff with the required expertise to master the Sen4Stat system and to be able to run it. Their request was just to understand how it works but their plan was to outsource it through other departments (IT or Geomatics for instance) or through existing computing center.

6.2 System release and forum

The first version of the Sen4Stat system was made available on the website on January 2023, with a beta version delivered to beta users from October 2022.

Minor updates were made on this version and regularly published on the website between January 2023 and the end of the project.

In order to support the system release, a number of tools have been put in place during the last month of the project, which will be better exploited during the extension: forum and online support.

A website was also set-up at the start of the project, with a major update at the end to better reflect the outcomes and be in agreement with the ESA branding.

6.3 System dissemination in partnership with international donors

Promoting the use of the EO data by the NSO and building capacity to ensure a proper uptake of the system is a process that takes time, more time than the project lifetime. As a result, we collaborated closely with the FAO and the World Bank to support the system dissemination world-wide. Figure 6-1 shows the different countries were Sen4Stat is being implemented (i.e. demonstrated with statistical use cases and with capacity building activities for the uptake) and is being demonstrated (i.e. feasibility study to possibly move for an implementation).

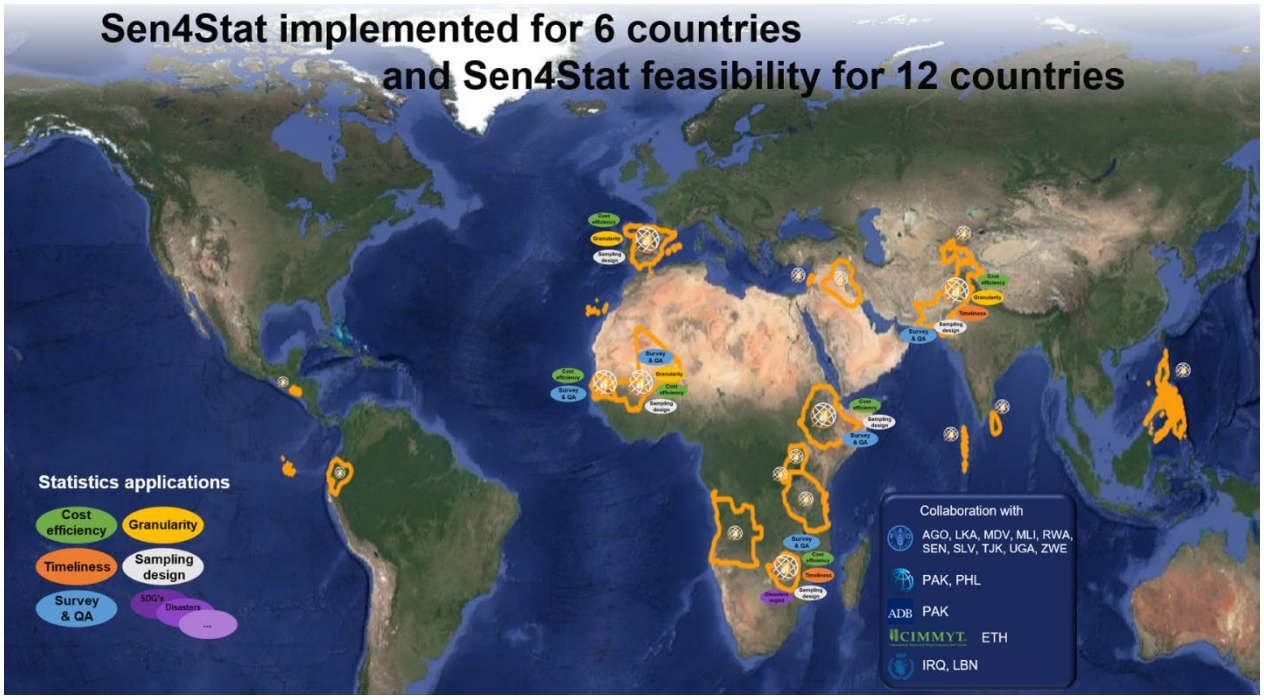


Figure 6-1. Sen4Stat dissemination, partnering with international institutions