ICASX 2026
8–10 July 2026 • Kraków, Poland

Crop yield inference from satellite imagery: a new dataset at the plot level

Antoine Lesauvage
SSP, Auzeville, France, antoine.lesauvage@agriculture.gouv.fr

Bertrand Ballet
SSP, Auzeville, France, bertrand.ballet@agriculture.gouv.fr

Damien Babet
SSP, Auzeville, France, damien.babet@agriculture.gouv.fr

Olivier Queyrut
CNES, Toulouse, France, Olivier.Queyrut@cnes.fr

Thomas Duthoit
Thalès Services Numériques, Toulouse, France, thomas.duthoit@thalesgroup.com

Thierry Chapuis
CNES, Toulouse, France, Thierry.Chapuis@cnes.fr

Antoine Barrau
CNES, Toulouse, France, Antoine.Barrau@cnes.fr

Abstract[1]

Estimating crop yields at field level from satellite imagery requires large, geolocated training datasets — a resource that is both scarce and rarely accessible to researchers. This paper presents a collaboration between the French ministry of Agriculture statistical service and the French space agency, combining data from the annual arable land survey, the French land parcel identification system, Sentinel-2 spectral indices, and ERA-5 meteorological data to train an XGBoost model that infers crop yields at parcel level across metropolitan France. The model is applied in an extrapolation setting — not for short-term forecasting, but to propagate yields observed in a representative statistical survey to all parcels for the same year. Parcel-level R^2 ranges from 0.42 (sunflower) to 0.71 (winter wheat), and NUTS3-level R^2 ranges from 0.81 (grain maize) to 0.98 (winter rapeseed), with no systematic bias. The main output is a multi-year (2017–2025) parcel-level yield dataset covering several hundred thousand parcels per crop and per year for four major crops — winter soft wheat, grain maize, winter rapeseed, and sunflower. To our knowledge, no equivalent dataset exists at this scale in Europe. Compliant with statistical confidentiality requirements, it is intended to be made available to researchers for model training, data linkage, and the production of fine-grained agricultural estimates.

[1] Text and materials are free of any copyright violation

1. Introduction

Estimating crop yields from satellite remote sensing is an important challenge for many stakeholders in the agricultural sector. Recent advances in satellite imagery — in particular the free availability of Sentinel-2 images at high spatial and temporal resolution — combined with progress in machine learning algorithms, open up unprecedented opportunities to estimate yields at a fine spatial scale across entire territories.

The main obstacle to the development of these approaches remains the lack of high-quality training data. Geolocated crop yield data are scarce, fragmented, and often confidential [i, ii]. Useful data do exist, however: national statistical offices collect large-scale representative yield data through agricultural surveys. But these data are subject to statistical confidentiality rules, and statistical offices do not always have the expertise needed to exploit them jointly with satellite imagery. Researchers who master these techniques, for their part, lack representative ground-truth data to calibrate and validate their models. The result is a silo effect that hampers the development of robust and generalisable methods.

This paper presents the results of a collaboration between the Statistical and Prospective Service (SSP, French Ministry of Agriculture) and the French National Centre for Space Studies (CNES), aimed precisely at removing this obstacle. Our approach consists in training a gradient boosting model (XGBoost) on yields collected through the arable land survey or Terres Labourables (TERLAB) — an annual survey covering approximately 17,000 farms distributed across the French territory — matched to parcels from the Land Parcel Identification System (LPIS, known in France as the RPG) and to Sentinel-2 spectral indices. This model is then used to infer a yield for each parcel in the RPG, producing a large-scale parcel-level yield dataset covering the entire metropolitan territory. The inference is not used for short-term forecasting, but for extrapolation over the same years as the training sample. In this context, the model predicts yields effectively: the parcel-level R^2 ranges from 0.42 (sunflower) to 0.71 (winter soft wheat), and ranges from 0.81 (grain maize) to 0.98 (winter rapeseed) at the departmental level (NUTS3), indicating the absence of systematic bias.

The main contribution of this work is the production of this dataset: several hundred thousand parcels per crop and per year, over nine years (2017–2025), for four major crops — winter soft wheat, grain maize, winter rapeseed, and sunflower. To our knowledge, no equivalent exists at this scale in Europe. Unlike agronomic simulation approaches, the yields made available are inferred from yields actually observed in a representative statistical survey, which gives them solid empirical validity. This dataset is intended to be made accessible to researchers and practitioners who wish to train models, perform linkages with other public or private data sources, or produce estimates for geographical areas or domains not covered by conventional surveys, while complying with statistical confidentiality requirements.

2. Objectives

We build a model for predicting crop yields at the field level by combining data from a statistical sample survey, the Land Parcel Identification System (LPIS/RPG), geographic characteristics and weather conditions on these fields, together with satellite imagery.

The objective is to estimate, for recent years with satellite imagery and training data, the crop yields of all fields under arable farming across the national territory. These exhaustive data, made available to researchers, will enable analyses at a very fine level of detail.

3. Method

Data

The Terres Labourables survey

The annual Terres labourables (TERLAB) survey aims to provide data on the yields of the most important field crops in the French cropping system (cereals, oilseeds, protein crops, potatoes, sugar beet, etc.) at the NUTS3, NUTS2 and national levels. It contributes to the early estimation of changes in cropped areas and production of major crops. It is a major source for the annual agricultural statistics (SAA), which compile agricultural production figures (areas, yields, production, livestock numbers) at the same geographical levels [iii].

The survey covers a sample of around 17,000 farms [iv] located throughout France and having declared at least 5 hectares of arable land in order to be eligible for common agricultural policy (CAP) payments. This sample is drawn with a sampling rate of approximately 10% from about 176,000 agricultural holding, which altogether use nearly 13 million hectares of cultivated land.

The TERLAB questionnaire covers the area and yield of all major crops grown on the farm, the volume of irrigation water for maize, and the sowing intentions for the following year.

Crop yields collected through the Terres labourables survey

Yields collected in the TERLAB survey are reported in quintals per hectare with two possible decimal places (1 quintal = 100 kg). They are calculated, estimated, or supplied by the collector of the harvest (cooperative or trader). These are average yields calculated at farm level for each crop.

For training the satellite-based yield estimation model, the survey yields are matched with parcels from the Land Parcel Identification System (LPIS/RPG).

The Land Parcel Identification System (RPG)

The RPG is a national geospatial database derived from the CAP information system.

With more than 9 million fields updated annually based on farmers' declarations, the RPG is today the only georeferenced database that is homogeneous across France and produced at a fine scale.

Thanks to its national coverage, annual update, and spatial accuracy, RPG data are used in many areas: management and control of agricultural subsidies, policy monitoring, analysis of crop evolution and farming practices, climate change studies, biodiversity protection, water quality management, etc.

Building the training dataset: matching survey yields with RPG parcels

The labeled datasets used for training the crop yield prediction model are constructed by matching:

- RPG agricultural parcels, which provide geographic location, geometry, and crop type
- crop-specific yields reported by farmers in the TERLAB survey.

The matching is based on a unique farm identifier ("pacage") present in both the RPG and the TERLAB survey. A correspondence table between the crop nomenclatures used in TERLAB and in the RPG ensures that the correct yields are assigned to the appropriate parcels.

These labeled datasets have been produced for the years 2017 to 2025.

Satellite indices

Radiometric indices used as predictors of crop yield are computed from time series of Sentinel-2 satellite images corrected for atmospheric effects (surface reflectance – level 2A), together with validity masks indicating whether the surface is visible on a given date [v].

For each RPG parcel, NDVI, NDWI, NDRE1 and NDRE2 spectral indices are computed from the mean surface reflectance over the parcel. These indices describe photosynthetic activity and surface moisture.

The minimum (Min) and maximum (Max) values of these indices are retained as predictors, calculated over the following four two-month periods:

P1: Jan–Feb / P2: Mar–Apr / P3: May–Jun /P4: Jul–Aug

Satellite index extraction is performed for each of the 82 tiles of 110×110 km covering France, with a 10-km overlap, consistent with the Sentinel-2 data production grid [vi].

Other predictors

Additional field characteristics used as predictors include:

- Geographic variables
    - Area
    - Latitude
    - Longitude
    - Altitude
    - Slope
    - Orientation

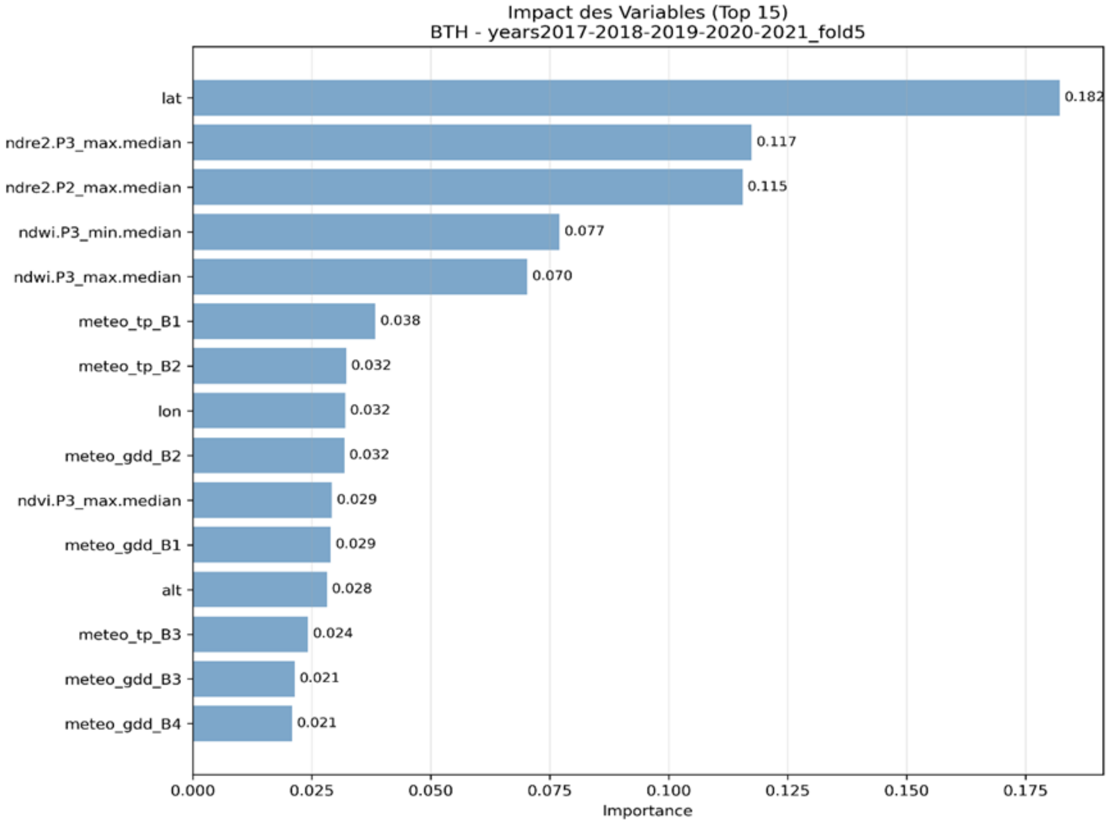
  derived from a digital elevation model.

- Weather variables related to vegetation phenology:
    - number of days with temperature > 30°C
    - number of days with temperature < 0°C
    - bimonthly precipitation totals
    - bimonthly growing degree days (GDD)

These data come from the Copernicus ERA-5 service [vii], calculated on 9×9 km grid cells. These meteorological variables were drawn from agronomic models [viii] and performed as well or better than more detailed meteorological features.

Post-hoc analysis shows that these predictors play a significant role in how the model estimates crop yield.

Figure 1 : Impact of the variables on the estimation of winter wheat yield



Developing the yield prediction model

Yield estimation at field level is performed using a boosted regression tree model based on the XGBoost algorithm [ix].

The method has been tested on the following major crops: winter wheat, winter oilseed rape, sunflower, grain maize. A separate model is trained for each crop.

<u>Cross-validation (k-fold)</u>

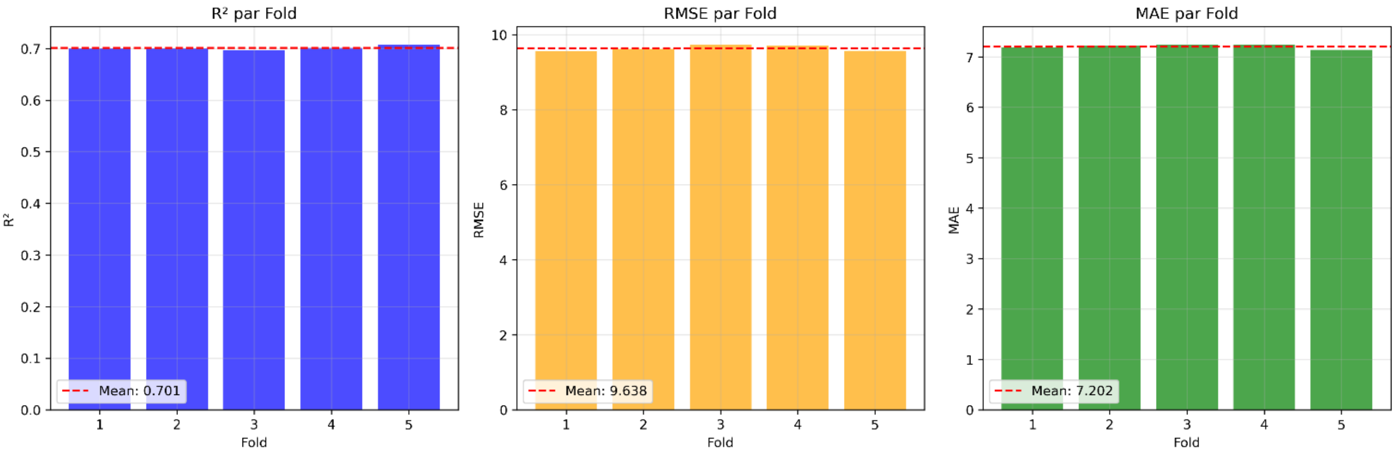
For each crop, the training dataset is split into 5 folds. Five models are trained, each time using 4 folds for training and 1 for testing. This yields 5 models per crop. Since TERLAB yields are defined at the farm level, a single yield value is associated with potentially several parcels. Therefore, fold separation must be done at the farm level to avoid contamination between train and test sets.

Training is performed on TERLAB data from 2017 to 2024[2]. Hyperparameters are optimised on the first fold using a random grid search, selecting those that maximise R^2 on the test set while limiting overfitting. Model performance is evaluated using R^2, RMSE, and MAE.

[2] At the time of writing this draft

As illustrated in the following chart for winter wheat, the cross-validation results are consistent across the different folds:

Figure 2: Model performance metrics across the folds



4. Results

Performance is generally satisfactory, especially when train/test datasets are large. Winter wheat (nearly 88,000 parcels on average per year in the train/test dataset) yields good results; performance is acceptable for maize (nearly 35,000 parcels on average per year), moderate for oilseed rape (nearly 22,000 parcels on average per year), and relatively poor for sunflower (nearly 18,000 parcels on average per year). Overfitting remains noticeable.

Figures 3 to 6 show the comparisons between the predicted yields and the yields collected in TERLAB, for both the training and test samples, for winter wheat, maize, winter rapeseed and sunflower. For wheat, the training R^2 is intentionally lower: the model was tuned to limit overfitting, which naturally reduces its performance on the training data.

Figure 3 : Winter soft wheat

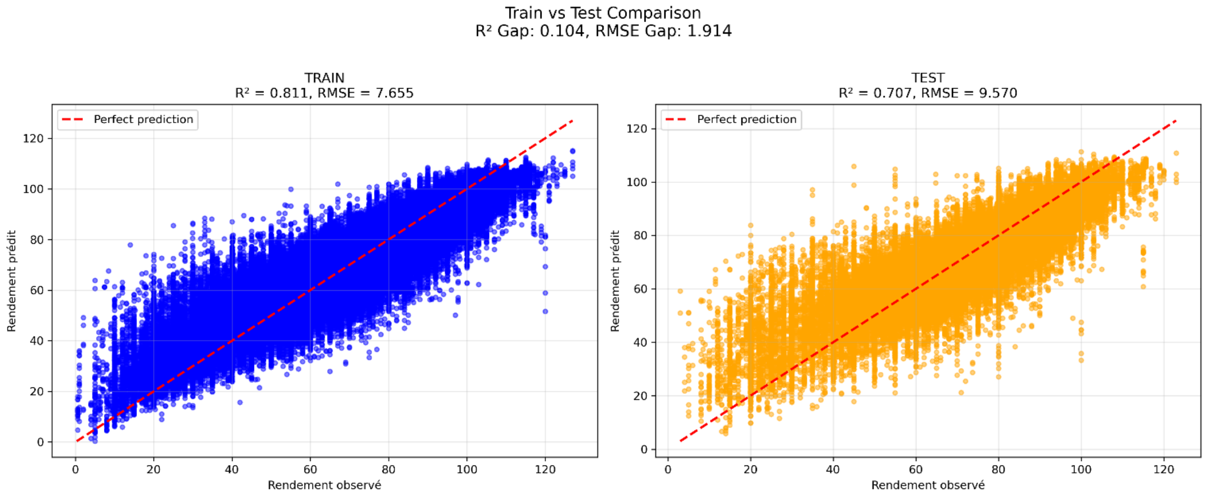


Figure 4 : Maize (MIS)

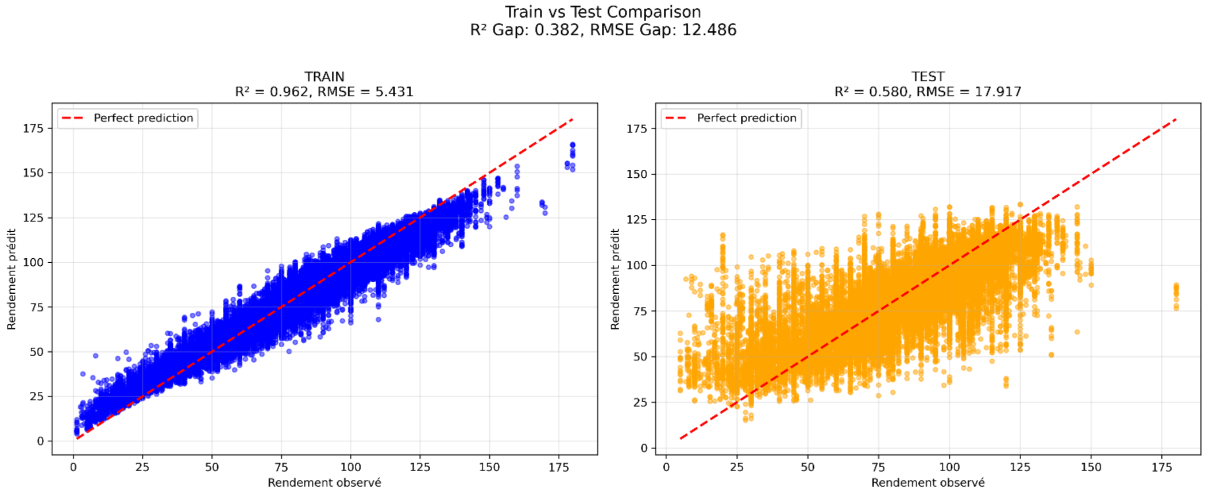


Figure 5 : Winter rapeseed

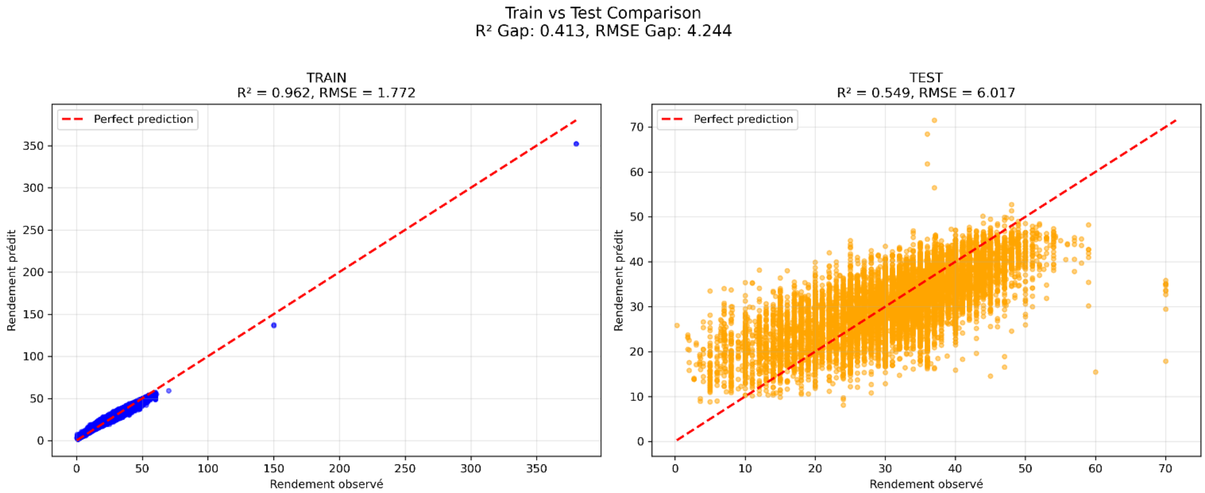
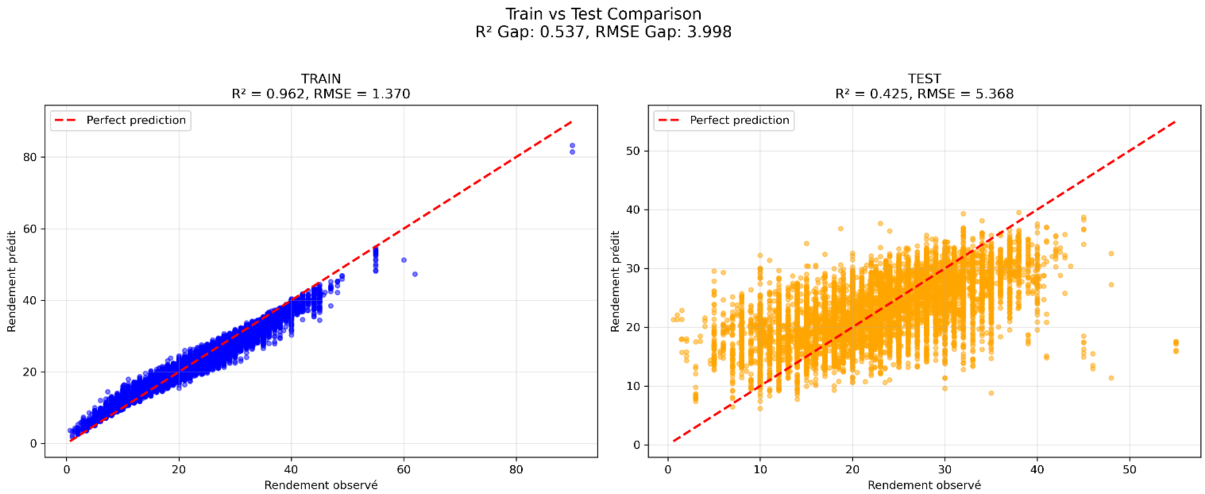


Figure 6 : Sunflower



Robustness of the methodological approach

To ensure the reliability and generalization of the results, several precautions were taken in the construction and evaluation of the model:

- **Multi-year evaluation**: The model is now tested on a dataset covering several years rather than a single year. This approach makes it possible to assess its ability to generalize to diverse conditions, independently of the specific climatic or agronomic features of any particular year.
- **Handling duplicated parcels**: Overlaps between Sentinel-2 tiles sometimes generated duplicated parcels, occasionally associated with differing radiometric values depending on the tile. These redundancies were corrected by removing duplicates and, where relevant, retaining the minimum and/or maximum radiometric index values to avoid biases linked to tiling artefacts.
- **Prevention of data leakage:** The separation between training and test datasets was strengthened by applying cross-validation at the farm level (rather than at the parcel level). Thus, for a given farm, crop and year, no parcel appears simultaneously in both training and test subsets. This prevents artificial inflation of performance metrics (e.g., overestimated R^2) caused by shared targets such as farm-level TERLAB yields.

Inference of plot-level yields

The five trained models (one per fold) each predict a yield for a given plot. The final predicted yield is the mean of the five values. An uncertainty indicator is computed from the standard deviation of the fold-specific predictions and the model's RMSE.

$$\sigma_{tot} = \sqrt{\sigma^2{}_{folds} + RMSE^2{}_{model}}$$

Where σ_{folds} is the standard deviation of the five predicted yield values for the plot, and $RMSE_{model}$ is the root mean squared error of the model, estimated on the training sample and averaged over the five folds.

Statistical validation of predicted yields at the departement (NUTS3) level

To validate the satellite-based prediction of crop yields at the plot level, we compute a NUTS3-level average from the predicted yields and compare it with NUTS3-level average yields obtained from the TERLAB survey.

Using the yields predicted by the XGBoost model for all RPG parcels, we calculate two types of predicted departemental averages:

- the predicted average yield computed over all RPG parcels (avg_predicted_yield_rpg)
- the predicted average yield computed only on RPG parcels belonging to the TERLAB sample (avg_predicted_yield_terlab)

We then compare these predicted departement-level yields with the observed averages calculated from TERLAB survey data (avg_observed_yield_terlab). These comparisons were carried out for the following crops and years:

- Winter wheat 2021

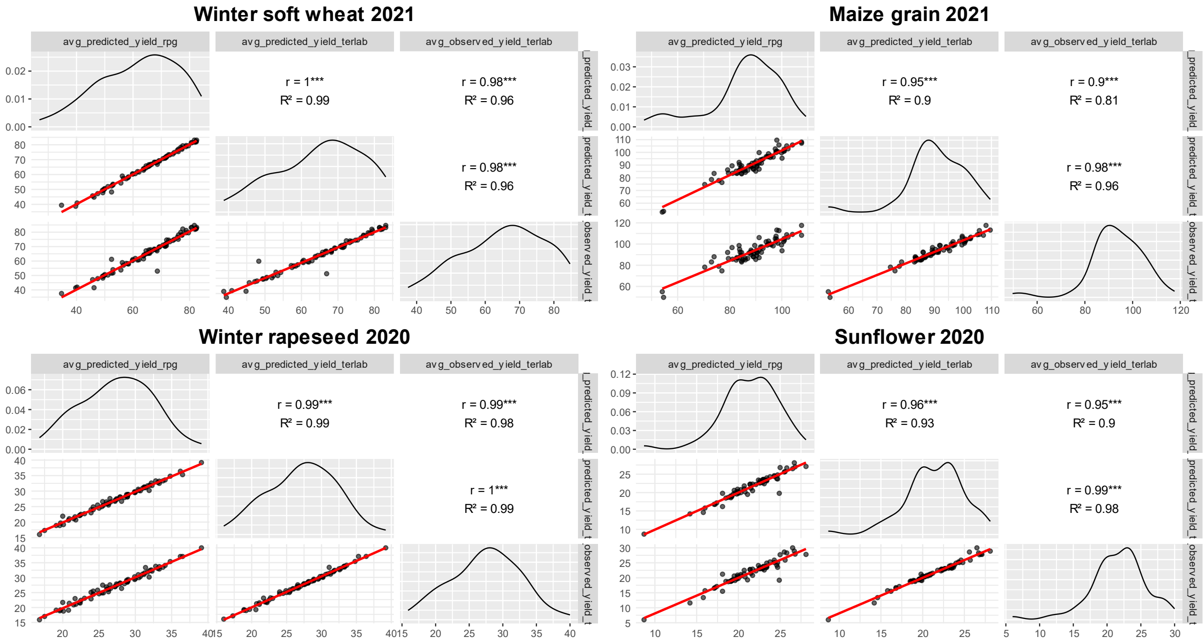
- Maize 2021
- Rapeseed 2020
- Sunflower 2020

The correlation analysis between these different departement-level yield estimates leads to the following conclusions:

- The TERLAB sample is highly representative of all plots in each departement: predicted averages computed on all parcels are strongly correlated ($R^2 \geq 0.93$) with the averages computed only on TERLAB parcels, except for maize ($R^2 \geq 0.9$), possibly due to differences between crop classifications in TERLAB and the RPG.
- Average yields are accurately predicted by the XGBoost model within the TERLAB sample: predicted NUTS3 sample averages are close to TERLAB observed value averages ($R^2 \geq 0.96$).

Correlation plots for winter wheat, maize, rapeseed and sunflower are shown below:

Figure 7: Correlation diagrams of predicted and observed yields by crop



Reading note: The correlograms above display **correlation statistics for each pair of yield estimation methods**. The diagonal shows the distribution of departement-level yields for each method (density). The lower triangle shows scatterplots of departement-level yields with the corresponding regression line, while the upper triangle displays numerical correlation indicators (r, significance test, R^2).

Pearson's r: checks whether differences between departements are consistent across the two methods. A value close to 1 means that both methods produce very similar variations between departements and rank them almost identically.

R^2: checks whether the two methods produce comparable yield levels. A high R^2 means that one method explains the values of the other well: their numerical results are very close.

5. **Perspectives**

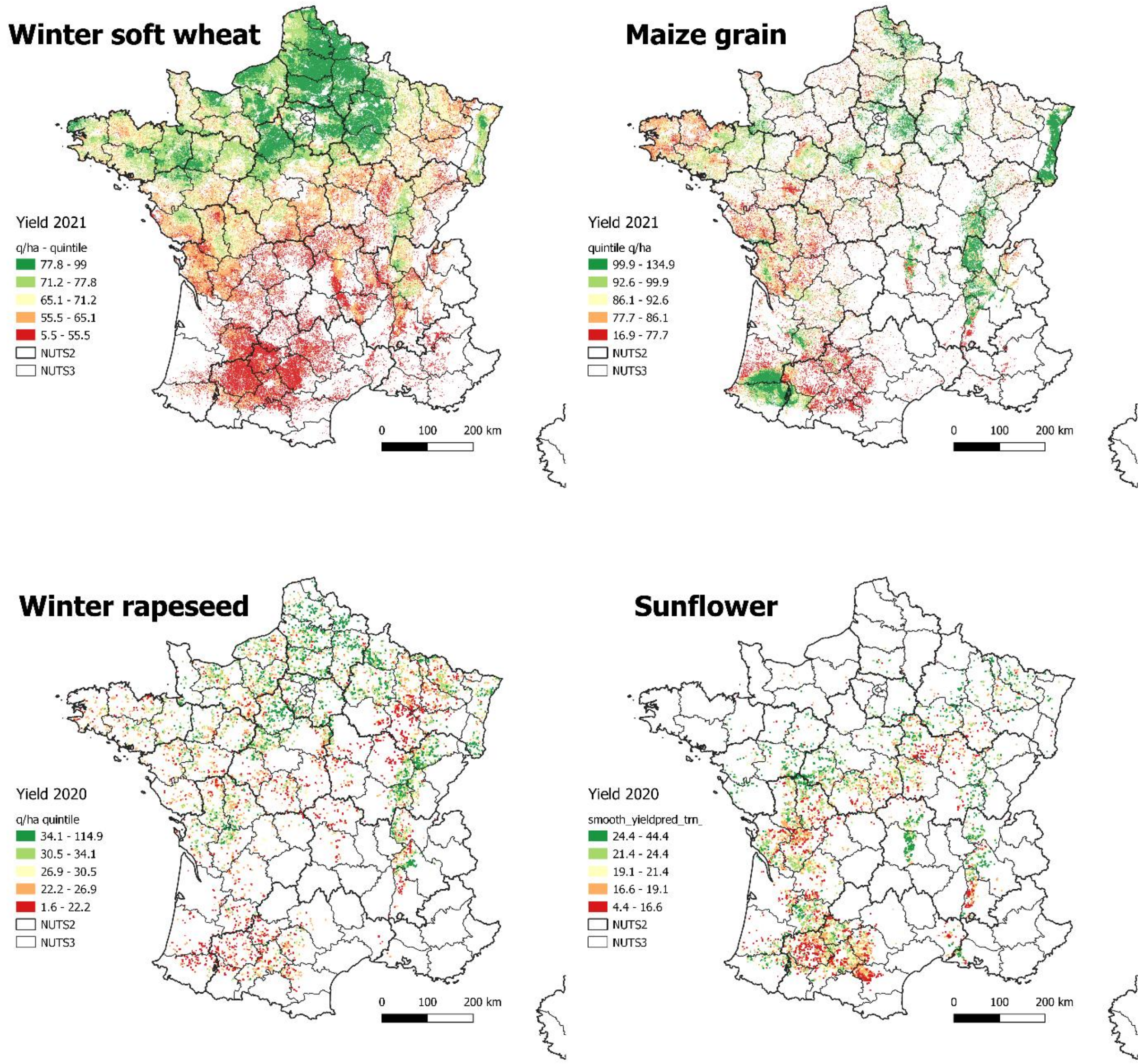
<u>Dissemination of predicted yield maps</u>

Predicted yields are disseminated through smoothed maps showing large-scale spatial patterns rather than parcel-level variability. Plot-level predictions are projected onto a grid and spatially smoothed using distance-weighted averages.

- A 300×300 m grid and 1-km radius are used for wheat and maize.
- A 1-km² grid and 3-km radius are used for sunflower and rapeseed.

Figure 8 : Spatial Distribution of Predicted Yield

Spatial smoothing (BTB package) - Quintile classification



Sources : Terres labourables surveys (SSP), RPG-LPIS (ASP), Sentinel2 (Copernicus)
Processing : Cnes/Lab'OT, SSP/DéMéSIS/BMS

Dissemination of plot-level predicted yields

A multi-year (2017-2025) field-level dataset will be created, complementing the RPG, and made available to researchers in agronomy and agricultural statistics. The k-fold validation strategy ensures that the predicted yield for a plot is always generated by a model that did not use the survey response of that farm. This makes confidentiality-compliant dissemination possible.

Given the promising results, SSP plans to continue producing annual plot-level predicted yields using the updated sample survey and satellite data, for coming years and additional crops. Further work includes testing short-term yield forecast, and using predicted yields as calibration margin to correct for sampling noise in TERLAB and other surveys. The methodology can be adapted to any country with LPIS data and large yield surveys.

6. Références

[i] Corcoran, E., Bebber, D.P., Curceac, S. et al. A comprehensive UK crop yield dataset incorporating satellite, weather, and soil type information. Sci Data (2026). https://doi.org/10.1038/s41597-025-06528-x

[ii] Perich, G., Turkoglu, M. O., Graf, L. V., Wegner, J. D., Aasen, H., Walter, A., & Liebisch, F. (2023). Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. Field Crops Research, 292, 108824.

[iii] Statistique agricole annuelle 2024 – Chiffres définitifs : https://agreste.agriculture.gouv.fr/agreste-web/disaron/Chd2516/detail/

[iv] Terres labourables : https://agreste.agriculture.gouv.fr/agreste-web/methodon/S-TerLab2025/methodon/

[v] Inglada, J., Exploitation de l'enquête TERLAB pour l'estimation du rendement des cultures à la parcelle à partir de séries temporelles Sentinel-2. [Rapport de recherche] CESBIO. 2020. ⟨hal-02935469⟩

[vi] Tile viewer : https://www.cesbio.cnrs.fr/multitemp/sentinel-2/s2-tiles-tuiles-sentinel-2/

[vii] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 (Accessed on 19-11-2025)

[viii] Brisson N., Launay M., Mary B., Beaudoin N., Conceptual basis, formalisations and parameterization of the STICS crop model (https://hal.inrae.fr/hal-02824114)

[ix] Chen T., Guestrin C.,. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[x] Queyrut O., Duthoit T., Chapuis T., Barrau A., Babet D., Ballet B., Lesauvage A., Prediction of crop yields at field scale from earth observation data, NTTS 2025 poster, 2025