Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Mathieu Fauvel, Vincent Thiérion, Millet Jerôme, Laurence Curtet

Graphical Abstract

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Hélène Touchais, Vincent Thierion, Jerome Millet, Laurence Curtet, Mathieu Fauvel

Highlights

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Hélène Touchais, Vincent Thierion, Jerome Millet, Laurence Curtet, Mathieu Fauvel

- Research highlight 1
- Research highlight 2

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas[a,*], Hélène Touchais[a], Vincent Thierion[a], Jerome Millet[b], Laurence Curtet[b], Mathieu Fauvel[a]

[a] *CESBIO Université de Toulouse, CNES/CNRS/INRAE/IRD/UT3-Paul Sabatier 31401 Toulouse France*
[b] *OFB ..., France*

Abstract

TODO

Keywords:

1. Introduction

Grasslands cover approximately 40% of the Earth's land area, encompassing nearly 70% of the global agricultural land area, and are distributed on all continents and all latitudes [1, 2]. Grassland dynamics influence global ecosystem functioning, and their impacts are widely modulated by management practices intensity [3].

Grasslands are subject to management practices such as mowing or grazing or a combination of both. These practices are primarily driven by grassland landscape maintenance, as well as by ecosystem service of provisioning offered by the grasslands. Therefore, monitoring grassland management practices is essential for assessing management intensity level, which in turn plays a critical role in studies related to biodiversity (XXXX), water (XXXXX) and carbon (XXXXX) cycling and others topics (XXXX).

In France, the National Observatory of Mowed Grassland Ecosystems (Il faudrait une référence, un lien ...) conducts birdlife monitoring in mowed grass-

*Corresponding author

lands and has related breeding failures to earlier mowing date. Indeed, Broyer et al. [4] has shown that early mowing intercepts birds' reproductive period and interrupts their breeding process. Traditionally, responsible agencies conduct occasional observation campaigns to support ecosystem-related public policies, but ground observations are not spatially exhaustive and are time-consuming, thus limited in terms of area covered and revisit frequency. Remote sensing Earth observation enables regular and global-scale monitoring, enabling tracking of vegetation dynamics. Currently, Sentinel-2 mission provides cost-free high resolution data at 10m spatial resolution with a 5-day temporal frequency (10 days before 2017), allowing intra-plot level observations and long-time analysis.

Grassland mowing events timing and intensity have been investigated using satellite image time series (SITS), mainly from features sensitive to vegetation status, such as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Leaf Area Index (LAI) and more. These methods usually exploit the temporal information contains ed? in SITS to detect moving events: a significant variation is usually associated to an events. The various methods differ in how the drop is computed. For instance, Estel et al. [5] assessed annual mowing frequency using temporal change analysis based on spline-adjusted MODIS NDVI time series was used?. Their approach involved identifying mowing events as instances where a local minima exhibited a change, relative to its preceding peak, exceeding 10% of growing season amplitude. The results showed an overall accuracy of 80%, which decreases as the frequency of events increases. In northern Switzerland, Kolecka et al. [6] also estimated mowing frequency employing similar temporal change analysis, but based on raw Sentinel-2 NDVI time series. Here, a drop in NDVI value greater than 0.2, between two consecutive cloud-free acquisition dates, was counted as a mowing event. Their method accurately identified 77% of observed events and highlighted that false detection can occur due to residual cloud presence, while sparse time series led to the omission of mowing events. Regarding Griffiths et al. [7], mowing events frequency and timing were mapped in Germany using 10-day composite Harmonized Landsat-Sentinel NDVI time series. Discrepancies between a hy-

pothetical bell-shaped curve and the current polynomial-fitted curve were evaluated. An event was counted when the difference exceeded 0.2 NDVI. Findings revealed consistent spatial patterns in mowing frequency (indicating extensive and intensive management). However, estimated dates exhibited significant discrepancies compared to observed dates (with a mean average error greater than days), which could be due to lower temporal resolution of Sentinel-2 before 2017 and the absence of reliable ground data for calibration and validation. Stumpf et al. [8] mapped grassland management (grazing or mowing) and its intensity based on biomass productivity and management frequency, respectively. The latter were extracted from n-day composite Landsat ETM + and Landsat OLI NDVI time series. As in previous cases, a management event was counted when NDVI loss is higher than a threshold, which was based on the probability density function of all NDVI changes across the time series and was specified for $p = 0.01$. Their approach yielded management patterns consistent with several management-related indicators (species richness, nutrient supply, slope, etc). Recently, Watzig et al. [9] estimated mowing events in Austria, using Sentinel-2 NDVI time series and implementing discrepancy analysis between a idealized unmowed trajectory and actual NDVI values. An event was recorded if the difference exceeded -0.061pourquoi?. Commission errors due to residual clouds were reduced via a subsequent binary classification of each estimated event using a gradient boosting algorithm trained over cloudy plots. Findings indicated an overall accuracy of 80% in correct event detection, with estimated dates closely aligning with observed dates (MAE $<$ 5 days).

Previous methods exploit only optical modality and are limited by clouds cover. To cope with optical SITS limitation, Vroey et al. [10] developed a algorithm for detecting mowing events across Europe using jointly raw Sentinel-2 NDVI and Sentinel-1 VH-coherence time series. A mowing event was deemed when temporal change exceeded -0.15 NDVI and 1.0 VH-coherence standard deviation, multiplied by a factor (3.0×10^{-7}), respectively. VH-coherence standard deviation was calculated from residuals of the six preceding observations. These residuals capture disparities between linear-fitted values and actual values. In

the final product, Sentinel-1 outputs were considered when Sentinel-2 omitted events due to cloud cover. Results demonstrated synergy between optical and radar data in detecting mowing events (F1-score of 79%). Using only Sentinel-2 data achieved maximum precision, but combining both sensors boosted recall significantly. Also relying on optical and radar data synergy, Reinermann et al. [11] mapped mowing frequency across Germany, from Sentinel-2 EVI and Sentinel-1 PolSAR entropy time series separately. A mowing event was counted when temporal change exceeded $-0.07\times$EVI, which was calculated between two consecutive critical points (local minima and its preceding local maxima). S1-based detection was used to find potentially missed mowing events in cloudy gaps ($>$ 25 days) in optical observations. Here, the change needed to exceed 0.05 entropy between a peak and a preceding trough. Findings showed that S2-based method correctly detected 60.3% of mowing events with an F1-Score of 0.64. However, combining S1 and S2 increased recall but also caused more false positives, lowering precision.

To reduce cloudy gap in optical time series, Schwieder et al. [12] combined Sentinel-2 and Landsat-8 EVI time series for mowing events detection in Germany. They analyzed the discrepancies between actual observations and an idealized temporal profile (unmowed regime). An event was recorded when difference exceeded the mean value of all absolute residuals. Also, the detected point needed a loss greater than 1.0 standard deviation (of actual time series) compared to the previous point. Overall, detected mowing dates exhibited an average absolute difference $<$ 12 days compared to observed dates. Mowing events were detected with an average F-score of 0.60, while the estimation of their frequency showed a mean error $<$ 40% of the actual number of mowing events. They highlighted that performance was lower in areas with less cloud-induced observations. In [13] Sentinel-1 and -2 SITS as well as climatic and topographic data were used to reconstruct continuous Sentinel-2 NDVI time series for mowing date estimation, based on NDVI drop analysis.

While threshold-based methods have been widely investigated, supervised learning based approaches have also been explored. Komisarenko et al. [14]

estimated mowing events timing at plot level in Estonia, using a 1-D Convolutional Neural Networks (CNN) on Sentinel-2- and Sentinel-1-based features time series. Although fourteen features were used, NDVI and the harmonic mean of VV and VH coherence were found to be the most relevant. Their approach yielded an accuracy of 73%. Most of the incorrectly estimated events were observed when optical time series were sparse or the size of the plot was small. Lobert et al. [15] also used a similar deep learning model (1-D CNN) on Sentinel-2/Landsat-8- feature and Sentinel-1-based features time series for mowing event frequency and timing detection. Among all tested feature combinations, the highest overall accuracy was reached when combined NDVI, backscatter cross-ratio and coherence with an F1-Score of 0.84. Estimated mowing dates showed a MAE of 3.79 days compared with the observed dates. In terms of management intensity, low-intensity grasslands were overestimated, while high-intensity grasslands were underestimated. Following a similar approach, Holtgrave et al. [16] tested four machine learning algorithms for mowing event detection in Germany. Sentinel-2/Landsat-8, Sentinel-1- and weather-based features time series were analyzed. Mowing events were detected by a binary classification approach (mown or unmown) for each observation in the time series, using the adjacent observations as predictors. As a preprocessing, the optical time series were gap-filled using machine learning regression and linear interpolation techniques. Here, 1D-CNN (F1-score 0.72 - 0.80) and Long Short-Term Memory (F1-score 0.89) algorithms performed better on unknown and known study site and years, respectively. Overall, optical data proves advantageous for known study sites and years, while the inclusion of both optical and SAR data yields favorable results for transferable models. Weather data were significant in classifying mowing events for known study sites and years, but caution is needed when including it in transferable models.

****

""" In the literature, there is no consensus on the optimal satellite data for mowing event detection in grasslands. On the one hand, some studies demonstrated Sentinel-1 data potential, due to their sensitivity to changes in vegetation

cover structure, and their insensitivity to clouds [17]; but performance varies locally due to factors like soil moisture, vegetation water content, roughness, etc. On the other hand, some authors combined Sentinel-2 and Sentinel-1 data to reduce cloud effects in time series, and found enhanced performance in some cases [10, 18, 14, 15, 16]. Overall, most authors agree that optical data alone can effectively detect mowing events, provided that enough cloud-free observations are available [6, 7, 9, 8, 12]. In our study, we focused solely on optical data as previous studies have found no significant improvement by combining optical and radar data for grassland monitoring in France (e.g., [19]). """

Our approach:

In France, local remote sensing-based studies have already been conducted to discriminate grassland management practices in the northwest [20] and detect mowing events in the southeast [21].

Why first event only??? Considering environmental challenges in mowed grassland, as an indicator of management intensity, fulfillment of ecological policies for avifauna conservation, to support grassland management monitoring system at national level in France

Here, as a complement to previous efforts, we focused on mapping grassland first mowing event date using Sentinel-2-based features time series, primarily to support grassland management monitoring system at national level in France. Based on the state-of-the-art, we conducted a comprehensive evaluation of both machine learning- and threshold-based approaches, either already implemented in mowing event detection or chosen for their promising potential in this specific task. These methods were implemented via Iota[2] (`https://docs.iota2.net/`).

How is the document organized?

This paper is organized as follow : ...

2. Materials and Methods

2.1. Study area

Our study area covers permanent grasslands across the mainland France (except Corsica), which represent 68.5% of the total grassland area -including permanent, temporary and other grasslands-, declared in the Land Parcel Identification System - LPIS [22] in 2022 (Figure 1). According to Köppen–Geiger classification [23], the France climate is mainly oceanic across the country - with warm summers -, and is mediterranean in the south. Annual rainfall is around 800-1 000 (mm), with a contrast between the western (> 1 000 mm) and the southeastern (600-800 mm) regions. The average annual temperature is about 11-13 °C, with 20-25 °C in summer and 5-10 °C in winter (`https://meteofrance.com`).

Permanent grasslands are defined in the LPIS as surfaces with uninterrupted herbaceous cover for at least 6 years and are identified at the plot level with class code 18. These permanent grasslands alone account for approximately 27.5% (76 835 km^2) of the entire agricultural area reported for 2022. Grasslands cover regions that are less suitable for agricultural activities due to unfavorable climatic or site conditions (high altitudes, steep slopes, poor or wet soils). In mainland France, permanent grasslands are found in mountain chains in the center (Massif Central), western (Massif Armoricain), eastern (Jura and Vosges), Alps and Pyrenees, as well as in plains and wet regions (Figure 1). According to the LPIS, at least 75% of permanent grassland plots cover 2.80 hectares or less, and the largest plots -exceeding 20.0 hectares- are concentrated mainly in the center and eastern regions. Grasslands undergo various management practices such as mowing or grazing or a combination of both; and the intensity of these practices varies across plots, influenced by climate, site conditions and farmer decisions. Lower altitudes tend to offer more favorable conditions for mowing and more intensive management. In mainland France, grassland growing season spans from spring to autumn (March to October) and mowed grasslands are mainly managed extensively, with one or two mowing events per year (up

to six mowing events in intensive management).

From an avifauna diversity view point, timing of the first mowing event is more important than frequency of mowing events along growing season. In this sense, in our study area, intensive management refers to parcels for which the first mowing event occurs before June 15th, whereas extensive management, mowing happens after that date. Extensive management practices are beneficial for biodiversity [24] and birdlife [4], and are actively promoted by the European Common Agricultural Policy (CAP) through incentive payment mechanisms.

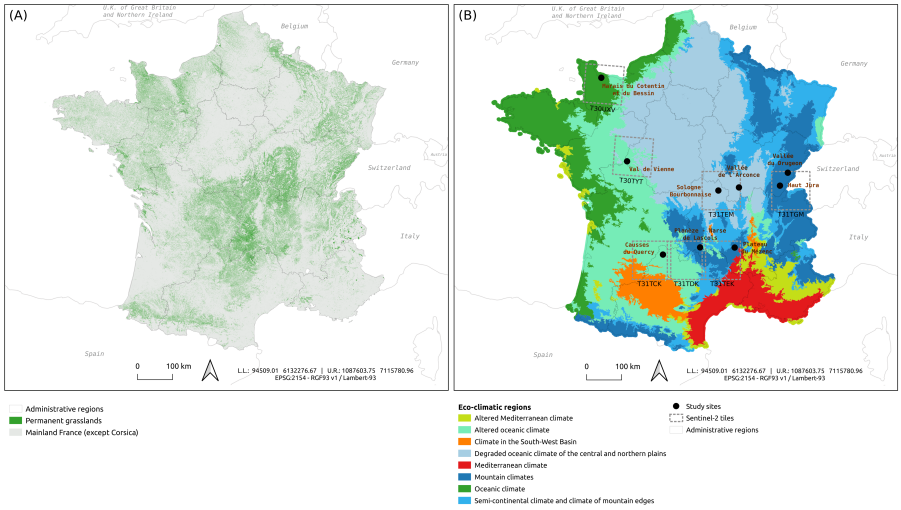

Figure 1: (A) Study area location. The gray color represents the delimitation of mainland France (except Corsica), while the green color represents the permanent grassland plots declared in the LPIS 2022. (B) Study sites location. The black dots represent the observation sites in 2022. The boxes in dashed gray lines represent the Sentinel-2 tiles that intercept each observation site. The color palette represents the eco-climatic regions in mainland France, as defined in [25].

2.2. Satellite data

All available Sentinel-2 (L2A) surface reflectance images, captured throughout the growing season (from January to September 2022) and intercepting mainland France, were used. This dataset comprised ninety tiles and seven of them, intercepting ground observation sites, were used for training and testing models (Figure 1). Each tile provided an average of sixty images. All spectral bands (except B1, B9 and B10) were used, after resampling 20m resolution

Table 1: Satellite data preprocessing and derived features according to implemented approach.

Approach	Spectral bands	Features	Preprocessing
Machine learning	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	1st derivative	linear interpolation
Threshold-based	-	NDVI	linear interpolation
Threshold-based	-	NDVI	raw data

bands to 10m resolution[1] to uniform pixel sizes to a common geographical grid.

These images had been preprocessed using MAJA algorithm [27] for atmospheric correction and cloud detection, and were downloaded from THEIA platform (`https://www.theia-land.fr`). All images were provided with a mask layer for clouds and shadows. For each tile, cloud- and shadow-free time series with a regular 10-day time interval were generated using a linear interpolator, as done in [28] or [29]. In addition to spectral bands, we also computed their temporal derivative -using finite differences-, as well as the Normalized Difference Vegetation Index - NDVI [30] (Table 1).

2.3. Reference data

In 2022, the French Biodiversity Agency (`https://www.ofb.gouv.fr`) coordinated an intensive campaign of ground observations throughout the mainland French territory, involving local government agencies participating in the National Observatory of Mowed Grasslands Ecosystem network. A total of eight sites (from north to south: Marais du Cotentin et du Bessin, Val de Vienne, Sologne Bourbonnaise, Vallée de l'Arconce, Vallée du Drugeon, Haut Jura, Plateau du Mézenc and Planèze - Narse de Lascols) were observed across four different eco-climatic regions (Figure 1), and covering a significant altitudinal gradient (Table 2). Observations were conducted once a week from May to August, covering a total of 2 227 plots with a balanced distribution among sites (Table 2). For each specific site, observed plots were chosen based on accessibility and the local observer's prior knowledge of the area.

[1] using a bicubic interpolation, as implemented in the Orfeo ToolBox and its SuperImpose application [26]

Plot boundaries were obtained from the 2020 LPIS. This database provides spatialized information on agricultural plot boundaries and crop types, but does not provides information about management practices. For permanent grasslands, a declared plot can be managed with two practices simultaneously (i.e., spatially separated mowing and grazing within the same plot). Therefore, prior to ground observation campaign, we visually assessed each chosen plot using a national database of aerial imagery (BD ORTHO, `https://geoservices.ign.fr/bdortho`) and Google Earth, to identify and separate sub-plots with homogeneous spatial structure. For each actual plot, a total of eleven observations were conducted throughout growing season. At each weekly visit, current management practice (mowing or grazing) and the corresponding date were recorded.

Based on these records, each plot was labeled as *mowed*, *grazed* or *mixed* (mowing + grazing). Then, these labeled plots were grouped into two management practice categories: *mowed* -including *mowed* and *mixed* plots (70.5% of plots)- and *unmowed* -including grazed plots-. A management practice could be ongoing during the visit or have occurred between the current visit and the previous visit. Consequently, in these *mowed* plots, observed date for a mowing event may have an average uncertainty of seven days. Here, 87% of *mowed* plots had one mowing event, while remaining plots had two mowing events.

Additionally, Causses du Quercy site was included (in the south, Figure 1), where 38 plots were observed with a lower temporal resolution. Here, observations were provided by the local observatory of the *Parc Naturel Régional des Causses du Quercy*, independently of the main observation campaign at the above-mentioned sites. At this site, 87% of observed plots were *mowed*, and all had one mowing event.

An unique site value was added to plots located within the same Sentinel-2 tile, corresponding to tile name. This grouping serves in the experimental validation to separate training and testing samples based on the tile membership: plots from a same site share the same satellite acquisition conditions and the same pedo-climatic conditions and should be used either for training or for test-

ing. In the following, the term "*site*" is used to denote plots belonging to the same Sentinel-2 tile.

Here, all pixels in observed *mowed* plots were selected and spectro-temporal profile and corresponding first mowing event day of the year (DOY) were extracted to serve as predictor and target values, respectively. Mowing event dates span from May 10th (DOY 130) to August 2nd (DOY 214), comprising a total of 328 451 pixels derived from the 1 605 observed *mowed* plots (Figure 2). Notably, 80% of these occurrences were concentrated between May 30th (DOY 150) and July 7th (DOY 188). The average observed date was June 16th (DOY 167), while the median was June 15th (DOY 166).

Finally, these observed mowing dates revealed a temporal pattern in this agricultural practice throughout the growing season, characterized by its occurrence during specific time periods. These time periods can be categorized sequentially throughout the year as follows: *early period* (before 150 DOY), *intermediate period* (between 150 and 188 DOY) and *late period* (after 188 DOY). Hence, observed plots were mainly mowed during *intermediate period* (Figure 2). The categorization of mowing periods relied solely on our observations and not on environmental criteria.

Table 2: Statistical description of the observed sites. The values represent the number of plots (# plots), average plot area (Av. area), number of mowed plots (# mowed plots) and approximate altitude. Tile column represents Sentinel-2 tile intercepting an observed site.

Site	**Tile**	**# plots**	**Av. area** (Ha)	**# mowed plots**	**Altitude** (m)
Marais du Cotentin et du Bessin	T30UXV	212	1.39	136	2-50
Val de Vienne	T30TYT	325	1.06	239	30
Sologne Bourbonnaise	T31TEM	267	2.47	119	230-280
Vallée de l'Arconce	T31TEM	288	2.23	174	280-390
Vallée du Drugeon	T31TGM	267	3.87	219	800-850
Haut Jura	T31TGM	272	2.55	213	800-950
Plateau du Mézenc	T31TEK	300	1.66	255	1100-1300
Planèze - Narse de Lascols	T31TDK	296	1.40	217	1000-1050
Causses du Quercy	T31TCK	38	0.50	33	309-775
Total		2 265		1 605	

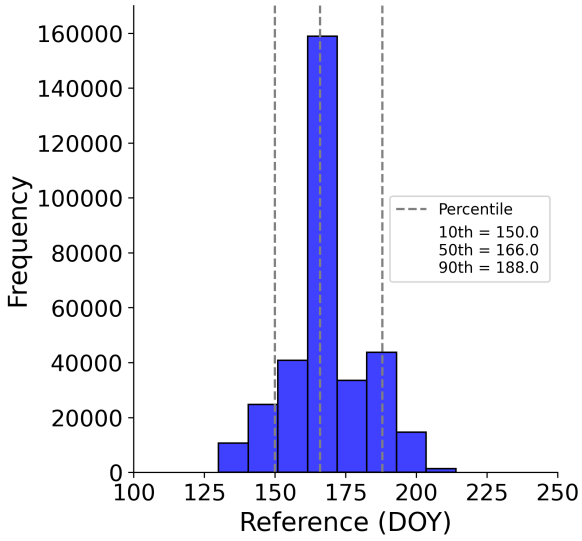

Figure 2: Distribution of pixel-level first mowing event dates observed in *mowed* plots across all sites in 2022. Reprendre la figure: mettre un peu plus de bins (20 ?), ne metttre que DOY sur l'axe horyzontal

2.4. Mowing events prediction

We predicted first mowing event date at pixel-level. For this purpose, we investigated several supervised regression algorithms, from conventional machine learning to recent deep learning ones. We then compared their performances against those obtained from unsupervised mowing event detection algorithms

found in recent literature, specifically those based on thresholding approaches.

Following Fauvel et al. [19], we set-up a spatial cross-validation to estimate the prediction accuracy (Figure 3). All observations from a site were excluded from reference data before training/calibrating models. Then, models were tested and assessed using the excluded site-specific observations. In other words, all observations from excluded site were used as testing data, and all observations from non-excluded sites were used as training data. This exercise was repeated seven times, so that each site was excluded once and considered as testing data. To estimate the test error, average prediction errors rate computed on 50 bootstrap set from each testing data were used.

In next sections, we provide a comprehensive overview of algorithms used to predict the first mowing event date.

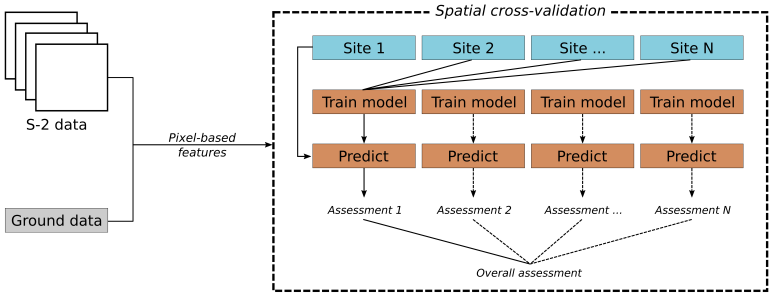

Figure 3: Workflow for mowing events prediction. The box in dashed lines represents the spatial cross-validation approach implemented in this study. Here, the different steps (model training, prediction and assessment) implemented for a given site are illustrated. The solid lines represent implementation for site 1, while dashed lines represent implementation for all remaining sites. In this example, all observations from site 1 were used as testing data, while all observations from the remaining sites were used as training data.

2.4.1. Machine learning approach

Machine learning algorithms are now widely used to estimate agro-ecological variables from remote sensing data. Yet, in recent years, few studies have applied machine learning-based methods to detect mowing event date, and traditional methods continue to dominate the field [31]. We implemented five generic models from the literature: conventional algorithms such as Random Forest (RF) and Ridge Regression, as well as cutting-edge deep learning architectures

such as Multilayer Perceptron (MLP), Fully Convolutional Network (FCN 1-D CNN, [32]) and Lightweight Temporal Attention Encoder (LTAE, [33]).

Linear regressor and MLP were used as a baseline, while RF was chosen due to its demonstrated high accuracy in large-scale prediction [28, 19]. FCN and LTAE were selected for their capacity to model temporal information, leveraging convolutional techniques and attention mechanisms, respectively. A brief review of these models is given in section 5.

The deep learning models were trained on 200 epochs with a batch size of 4096, using the optimization algorithm Adam[34]. 10% of the train dataset was used to form a validation set, which was used to perform early stopping and to reduce the learning rate by a factor of 10 when learning stagnated.

Oversampling techniques for minority ranges of mowing dates were also implemented using the *imbalanced-learn* toolbox [35]. More specifically, three oversampling techniques for classification problems were tested. The first is the naive Ramdom Over Sampling (ROS) method, which involves copying samples from minority classes and adding a small amount of noise. The two others are SMOTE [36] and ADASYN [37] algorithms, relying on a convex combination of existing samples. These oversampling techniques were defined for classification problem. We created 10 *fake* classes by dividing the interval of mowing date values into 10 sub-intervals of equal width and assigned each pixel to a class corresponding to the number of the interval in which its label fell[2]. For each method, we used the implementation provided by the imbalanced-learn library [35] and set the sampling strategy to 'not majority'. All classes were oversampled, except for the majority class, to obtained an equal number of samples in each class.

2.4.2. Threshold-based approach

Threshold-based methods are well-known in vegetation dynamics studies, and were widely used in mowing event frequency and timing detection [31].

[2]Other number of bins were investigated providing similar or worst results. For clarity we only report results for 10bins.

We implemented a recent specific mowing event detection algorithm introduced by Vroey et al. [10] as an integral monitoring tool within Sen4CAP program (`http://esa-sen4cap.org`). It was adapted to detect mowing event date, since it was primarily designed to detect mowing event time interval. The main differences compared to original algorithm are detailed in section 5.

The main idea developed in Vroey et al. [10] was to quantify temporal loss of NDVI, and to consider a mowing event when this loss is higher than a threshold value. In our study, the threshold value was set automatically using grid-search on training data (Table 3). Two types of thresholds methods were used:

1. A fixed threshold corresponding to fix loss of NDVI without taking into account the amplitude (specific pixel minimum and maximum value).
2. A relative threshold that determines a percentage of the pixel-wise NDVI seasonal amplitude.

Threshold-based algorithms were calibrated and tested using the same training and testing data used for machine learning-based algorithms. It should be noted that in assessing this method, our focus was solely on the configuration that yielded the optimal output among all possible configurations. This encompasses considerations such as raw or interpolated time series, relative or fixed thresholds, and determining the most effective threshold value.

Table 3: Algorithm-specific hyperparameters values. The **Value** column reports the selected values or the search range for the algorithm, with the following notation `start:step:end`. For Ridge Regression and threshold mode, cross-validation was used to select the optimal value.

Algorithm	**Hyperparameters**	**Value**	**Package**
Random Forest	Number of trees	100	Scikit-Learn
Rigde Regression	Regularization	1000:500:15500	Scikit-Learn
FCN (1-D CNN)	Learning rate	1e-3	Pytorch
LTAE	Learning rate	1e-3	Pytorch
MLP	Learning rate	1e-4	Pytorch
Fixed threshold	Minimum loss of NDVI	0.10:0.01:0.40	
Relative threshold	Minimum loss of NDVI	10:5:50 %	

2.5. Assessment of mowing events

The deviation between predicted and observed mowing dates in DOY were assessed using four metrics: Mean Absolute Error (MAE), Root Mean Square

Error (RMSE), Max error and the coefficient of determination (R^2), defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$
$$\text{Max error} = \max(|y_i - \hat{y}_i|),$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\hat{y}_i$ and y_i are predicted and observed moving dates at pixel i, respectively, n is the number of pixels in the testing data. In R^2 formula, $\bar{y}$ is the average of observed dates in all pixels in the testing data.

3. Results

All implemented algorithms were able to detect first mowing event date across the growing season in 2022. However, performances differed among models, with deep learning models performing better than conventional machine learning ones and threshold-based method (Figure 4). In the following sections, we present a comprehensive overview of the quantitative results obtained by all implemented algorithms. Subsequently, we focus on a more detailed analysis of the selected optimal model outputs. Finally, we visually interpret the mapping results of this selected model on observed *mowed* plots before extending its application to entire mainland France.

to remenber later: The article presents for each machine learning approach the best association between the three oversampling method, as well as no over sampling. The Appendix provides the results of all possible combinations machine learning methods/ oversampling techniques.

3.1. Evaluation of algorithms for mowing events prediction

Overall, machine learning models demonstrated better performances compared to threshold-based method (Figure 4 and Table 4). For machine learning-based approach, deep learning models (LTAE, LTAE_ADASYN, FCN_ADASYN, FCN and MLP_SMOTE) yielded higher performances compared to conventional ones (RF and Ridge). Notably, architectures that consider temporal dimension (LTAE and FCN) are placed as the best evaluated models (Figure 4).

FCN and MLP_SMOTE achieved similar scores in terms of R^2 and are at the inflection point between deep learning models ($R^2 > 0.48$) and conventional ones ($R^2 \leq 0.40$) (Table 4).

All machine learning-based models exhibited a temporal discrepancy (RMSE) about 10 days, an average MAE of 6.74 $\pm$ 1.07 days and Max_error of 57.55 $\pm$ 8.22 days (Table 4). For this approach, the linear Ridge model gave the worst performance in all evaluated metrics.

SMOTE oversampling technique only improved the R^2 of MLP model, increasing it from 0.40 to 0.47; while all other metrics remained similar (Table 4). In contrast, ADASYN algorithm did not contribute significantly in any metrics when implemented with LTAE and FCN.

Globally, deep learning models outputs (LTAE, LTAE_ADASYN, FCN_ADASYN, FCN, MLP_SMOTE) were statistically comparable with each other, while outputs of the other models did not show any similarity between them -except for RF and MLP- (Figure 7).

Regarding the threshold-based method, the implemented algorithm obtained an even lower overall performance (R^2= -1.36) compared to an algorithm (i.e., Mean) that always predicts the average of observed dates from training data (R^2= -0.10). Timing discrepancies were 14.02, 19.66 and 73.24 days for MAE, RMSE and Max_error, respectively (Table 4).

These average score values summarize the performance variability inherent to each model, either within a specific site or between studied sites (Figure 6). Here, all models performed differently across sites. Within each site, a model's performance varied across simulations -i.e., fifty folds per site, each fold containing 70% of the randomly selected observations- (Figure 6). Site-specific overall results are shown in Table 6 and the optimal model per site in Table 7. T31TDK exhibited the lowest variation (all models and folds) in R^2 (std.dev. of 0.05), while T31TEK had the highest (std.dev. of 0.10). Similarly, FCN exhibited the lowest variation across all sites and folds (std.dev. of 0.10), while LTAE had the highest (std.dev. of 0.18).

Finally, since our results will be used for biodiversity monitoring at plot level, we aggregated (by mean) pixel-level predictions to the corresponding plots. Subsequently, we re-evaluated these aggregated values against testing data. Here, overall inter-model performance trend remains basically unchanged. We can only observe that MLP_SMOTE and RF moved up a position to the detriment of FCN (Table 5), when compared to pixel-based evaluation (Table 4).

In conclusion, LTAE emerged as the optimal model due to its consistent performance across all sites. Consequently, in the following sections, all findings

are based on this model, and statistical interpretations centered around plot-level evaluations.

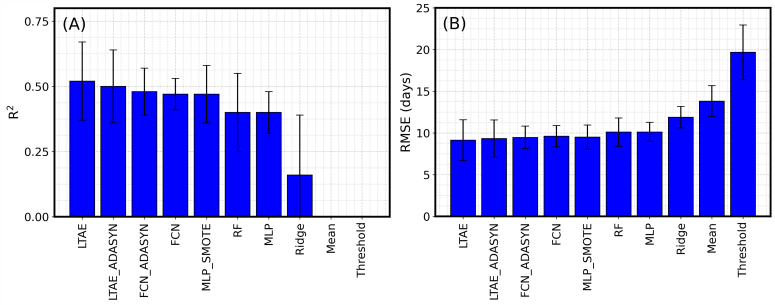

Figure 4: Algorithm-specific statistical summary in terms of (A) R^2 and (B) RMSE. The values represent weighted mean of all sites. A site-specific score was weighted using the number of pixels used for the evaluation (32 123 pixels in average). For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. Here, values less than zero are not shown. For comparison purposes, Mean corresponds to an algorithm that always predicts the average of observed dates from training data (poor model).

3.2. Mowing events prediction across sites

Although the optimal model differed among sites (Table 7), LTAE demonstrated consistent overall performance across all sites (Tables 4 and 5).

LTAE yielded a moderate overall performance, achieving a weighted average (# plots as weight) R^2 of 0.55 $\pm$ 0.17 across all sites (Figure 5 and Table 8). Temporal discrepancies between predicted and observed dates differed among sites; MAE ranged from 4.34 to 9.21 days, while RMSE from 6.76 to 14.61 days. The maximum error measured (all sites) was 61.04 days. Overall, across all sites (excluding T31TCK due to a limited # plots), this model exhibited its strongest performance at site T31TEK ($R^2 = 0.72$) and its lowest at site T30TYT ($R^2 = 0.18$).

At sites with larger prediction errors (e.g., RMSE $>$ 7.0 days), temporal discrepancies, between predicted and observed dates, were particularly accentuated at the extremes of the observed date values (Figure 5). These prediction errors at the extremes were mostly found in sites T30TYT, T30UXV and T31TEM. Notably, these prediction errors were positive during *early period*, while they

were negative during *late period* (Figure 8).

Most of the plots mowed during *early* or *late* periods exhibited prediction errors exceeding 15.0 days, while the plots mowed during *intermediate period* exhibited prediction errors lower than 15.0 days. In general, fewer than 10.0% of *mowed* plots (all sites and all mowing periods) experienced a prediction error exceeding 15.0 days (Figure 8).

”” Visual interpretation of observed plots???? ””

Regarding intra-plot spatial configuration of outputs, ...

Quel site? Quelle parcelle? 1 605 parcelles

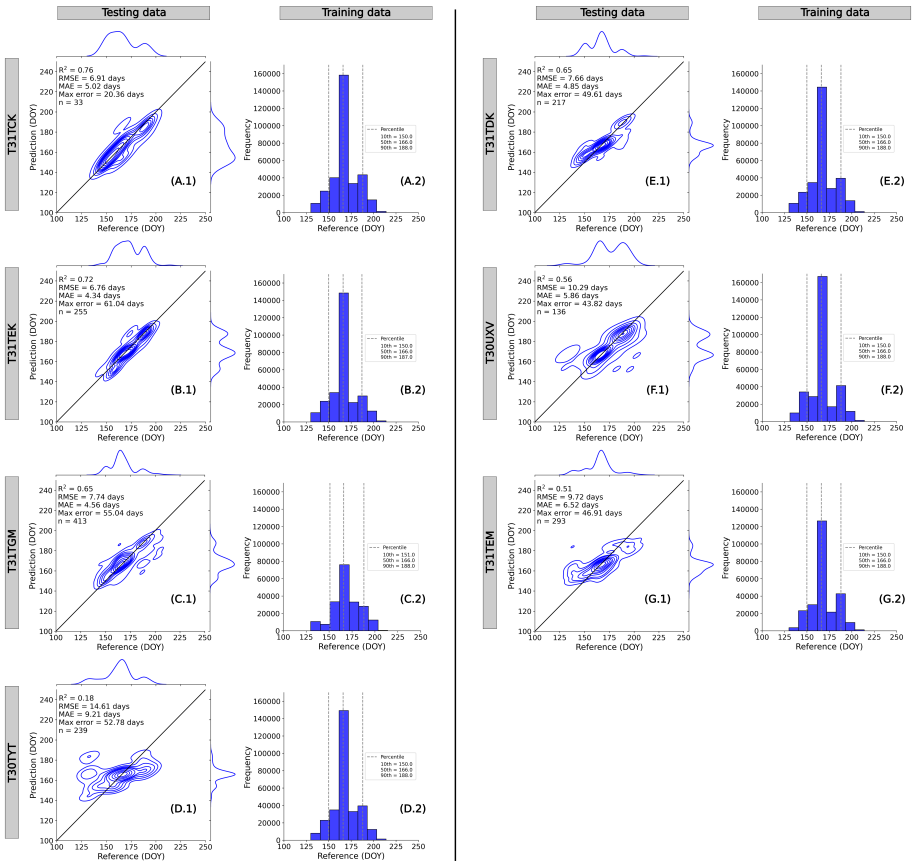

Figure 5: Site-specific LTAE outputs aggregated at plot-level. The columns represent datasets (testing or training), while the rows represent studied sites. For a given site, testing data included all samples from that specific site, while training data included all samples from all remaining sites. For a given site, the first column (e.g., A.1) shows predicted dates (y-axis) against observed dates (x-axis) at plot-level, while the second column (e.g., A.2) shows the distribution of pixel-level observed dates used to train the model. These predicted and observed dates are expressed in Days Of Year (DOY).

3.3. Mowing events prediction across mainland France

All plots of all sites as training data (ratio 1.0), not spatial CV!!!

Grassland mask description (two classes in a LPIS plot).

Spatial structure within some plots across sites??? pdf of plots with predictions

4. Discussion

4.1. Mowing event date

Time interval graph wiht delta t observed Expliquer que l'on estime la fin de l'intervale observé Faire une figure explicative Interpreter les dates estimées avec le biais. Si negative, delta t est diminué et donc moins d'incertitude. Si positive, delta t est augmenté et donc plus d'incertitude.

Laisser claire que les dates estimées sont en effet la fin de l'intervale de temps observée.

4.2. False positive mowing events

Remaining clouds and snow :

Kolecka et al. [6] found that the highest accuracy for detection of mowing events was achieved using additional clouds masking and size reduction of parcels, which allowed correct detection of 77% of mowing events. Additionally, we found that using only standard cloud masking leads to significant overestimation of mowing events (false positive).

4.3. Period of mowing event detection

Kolecka et al. [6] : We found that more than 40% of the study area was mown before 15 June, while the remaining part was either mown later, or was not mown at all.

4.4. Number of valid observations

Kolecka et al. [6] the detection based on sparse time series does not fully correspond to key events in the grass growth season.

4.5. Traditional and ML approach

Kolecka et al. [6] Our approach : First, it is not conditional upon the availability of reference data, which is often missing, contrary to widely used machine learning strategies, which require training data. Understanding seasonal and phenological aspects of management practices in the study area was sufficient

to allow discrimination between grass and non-grass clusters and, together with careful investigation of satellite imagery, allowed for development of a rule set for mowing event detection.

Komisarenko et al. [14] MLP showed lower performances than their CNN

Lobert et al. [15] NDVI time series alone mostly under performed in comparison to optical/SAR combinations but clearly outperformed input-sets that were solely based on SAR features.

Vroey et al. [10] : In this study, the Planet image interpretation approach allowed to rapidly gather a large reference dataset (n = 803) to validate the mowing detections in six countries along the whole season (April to October 2019). ??

here, our approach had more reference dataset (n = 1600)

Nationwide mapping

Even though remote sensing-based approaches have been shown valuable to gather such information, large- scale mapping approaches are still scarce (Reinermann et al., 2020).

When lower cloud-free obs is available, mowing event frequency led to a systematic underestimation of mowing events, when the general data availability was not as high as in the following years (compare Fig. 2). This becomes increasingly prob lematic towards the South and may thus hamper comparable analyses in the context of CAP in other European countries. The launch of Landsat 9 and the planned launches of Sentinel-2C, and –2D in 2021/24/25 would enable to increase the density of optical time series — but only if all sensors remain active.

While the usefulness of SAR data for detecting grassland management has already been tested in several studies with a regional focus (De Vroey et al., 2021; Tamm et al., 2016, Lobert et al. accepted), mowing detection algorithms that make use of SAR and op tical data together are still scarce. Even though mowing events can be identified in SAR time series, additional factors such as topography, parcel size and shape influence the results and there are still signal in teractions that need to be further explored (De Vroey et al., 2021).

In most common grasslands, exploitation activities (grazing or mowing) start from mid-April. In grasslands of high biological interest, supported by the EU CAP, mowing is only allowed after the 16th of June, for flowering purposes, and before the 31th of October.

Limitations : reference data with 7 days error We compared estimated date and observed date. However, observed date correspond to the end of observation time interval $\Delta t = 7days$.

5. Conclusion

References

[1] J. M. Suttie, S. G. Reynolds, C. Batello, Grasslands of the World, volume 34, Food & Agriculture Org., 2005.

[2] R. P. White, S. Murray, M. Rohweder, S. Prince, K. Thompson, et al., Grassland ecosystems, World Resources Institute Washington, DC, USA, 2000.

[3] Y. Zhao, Z. Liu, J. Wu, Grassland ecosystem services: a systematic review of research advances and future directions, Landscape Ecology 35 (2020) 793–814.

[4] J. Broyer, L. Curtet, M. Boissenin, Does breeding success lead meadow passerines to select late mown fields?, Journal of Ornithology 153 (2012) 817–823.

[5] S. Estel, S. Mader, C. Levers, P. H. Verburg, M. Baumann, T. Kuemmerle, Combining satellite data and agricultural statistics to map grassland management intensity in europe, Environmental Research Letters (2018). doi:10.1088/1748-9326/aacc7a.

[6] N. Kolecka, C. Ginzler, R. Pazúr, B. Price, P. H. Verburg, Regional scale mapping of grassland mowing frequency with sentinel-2 time series, Remote Sensing (2018). doi:10.3390/rs10081221.

[7] P. Griffiths, C. Nendel, J. Pickert, P. Hostert, Towards national-scale characterization of grassland use intensity from integrated sentinel-2 and landsat time series, Remote Sensing of Environment (2020). doi:10.1016/j.rse.2019.03.017.

[8] F. Stumpf, M. K. Schneider, A. Keller, A. Mayr, T. Rentschler, R. Meuli, M. E. Schaepman, F. Liebisch, Spatial monitoring of grassland management using multi-temporal satellite imagery, Ecological Indicators (2020). doi:10.1016/j.ecolind.2020.106201.

[9] C. Watzig, A. Schaumberger, A. Klingler, A. Dujakovic, C. Atzberger, F. Vuolo, Grassland cut detection based on sentinel-2 time series to respond to the environmental and technical challenges of the austrian fodder production for livestock feeding, Remote Sensing of Environment (2023). doi:10.1016/j.rse.2023.113577.

[10] M. D. Vroey, L. D. Vendictis, M. Zavagli, S. Bontemps, D. Heymans, J. Radoux, B. Koetz, P. Defourny, Mowing detection using sentinel-1 and sentinel-2 time series for large scale grassland monitoring, Remote Sensing of Environment (2022). doi:10.1016/j.rse.2022.113145.

[11] S. Reinermann, U. Gessner, S. Asam, T. Ullmann, A. Schucknecht, C. Kuenzer, Detection of grassland mowing events for germany by combining sentinel-1 and sentinel-2 time series, Remote Sensing 14 (2022). doi:10.3390/rs14071647.

[12] M. Schwieder, M. Wesemeyer, D. Frantz, K. Pfoch, S. Erasmi, J. Pickert, C. Nendel, P. Hostert, Mapping grassland mowing events across germany based on combined sentinel-2 and landsat 8 time series, Remote Sensing of Environment (2022). doi:10.1016/j.rse.2021.112795.

[13] A. Garioud, S. Giordano, S. Valero, C. Mallet, Challenges in Grassland Mowing Event Detection with Multimodal Sentinel Images, in: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), IEEE, Shanghai, France, 2019, pp. 1–4. URL: https://hal.science/hal-02387167. doi:10.1109/Multi-Temp.2019.8866914.

[14] V. Komisarenko, K. Voormansik, R. Elshawi, S. Sakr, Exploiting time series of sentinel-1 and sentinel-2 to detect grassland mowing events using deep learning with reject region, Scientific Reports (2022). doi:10.1038/s41598-022-04932-6.

[15] F. Lobert, A.-K. Holtgrave, M. Schwieder, M. Pause, J. Vogt, A. Gocht, S. Erasmi, Mowing event detection in permanent grasslands: Systematic

evaluation of input features from sentinel-1, sentinel-2, and landsat 8 time series, Remote Sensing of Environment (2021). doi:10.1016/j.rse.2021.112751.

[16] A.-K. Holtgrave, F. Lobert, S. Erasmi, N. Röder, B. Kleinschmit, Grassland mowing event detection using combined optical, sar, and weather time series, Remote Sensing of Environment (2023). doi:10.1016/j.rse.2023.113680.

[17] M. D. Vroey, J. Radoux, P. Defourny, Grassland mowing detection using sentinel-1 time series: Potential and limitations, Remote Sensing (2021). doi:10.3390/rs13030348.

[18] S. Reinermann, S. Asam, C. Kuenzer, Remote sensing of grassland production and management—a review, Remote Sensing (2020). doi:10.3390/rs12121949.

[19] M. Fauvel, M. Lopes, T. Dubo, J. Rivers-Moore, P.-L. Frison, N. Gross, A. Ouin, Prediction of plant diversity in grasslands using sentinel-1 and -2 satellite image time series, Remote Sensing of Environment 237 (2020) 111536. URL: https://www.sciencedirect.com/science/article/pii/S0034425719305553. doi:https://doi.org/10.1016/j.rse.2019.111536.

[20] P. Dusseux, F. Vertès, T. Corpetti, S. Corgne, L. Hubert-Moy, Agricultural practices in grasslands detected by spatial remote sensing, Environmental monitoring and assessment 186 (2014) 8249–8265.

[21] D. Courault, R. Hadria, F. Ruget, A. Olioso, B. Duchemin, O. Hagolle, G. Dedieu, Combined use of formosat-2 images with a crop model for biomass and water monitoring of permanent grassland in mediterranean region, Hydrology and Earth System Sciences 14 (2010) 1731–1744.

[22] P. Cantelaube, M. Carles, Le registre parcellaire graphique: des données

géographiques pour décrire la couverture du sol agricole, Le Cahier des Techniques de l'INRA (2014) 58–64.

[23] M. C. Peel, B. L. Finlayson, T. A. McMahon, Updated world map of the köppen-geiger climate classification, Hydrology and Earth System Sciences 11 (2007) 1633–1644. URL: https://hess.copernicus.org/articles/11/1633/2007/. doi:10.5194/hess-11-1633-2007, publisher: Copernicus GmbH.

[24] J. S. Petermann, O. Y. Buzhdygan, Grassland biodiversity, Current Biology 31 (2021) R1195–R1201.

[25] D. Joly, T. Brossard, H. Cardot, J. Cavailhes, M. Hilal, P. Wavresky, Les types de climats en france, une construction spatiale, Cybergeo: European Journal of Geography (2010). URL: https://journals.openedition.org/cybergeo/23155#annexe. doi:10.4000/cybergeo.23155, publisher: CNRS-UMR Géographie-cités 8504.

[26] O. D. Team, Orfeo toolbox 8.1.2, 2023. URL: https://doi.org/10.5281/zenodo.8178641. doi:10.5281/zenodo.8178641.

[27] V. Lonjou, C. Desjardins, O. Hagolle, B. Petrucci, T. Tremas, M. Dejus, A. Makarau, S. Auer, MACCS-ATCOR joint algorithm (MAJA), in: A. Comerón, E. I. Kassianov, K. Schäfer (Eds.), Remote Sensing of Clouds and the Atmosphere XXI, volume 10001, International Society for Optics and Photonics, SPIE, 2016, p. 1000107. URL: https://doi.org/10.1117/12.2240935. doi:10.1117/12.2240935.

[28] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, I. Rodes, Operational high resolution land cover map production at the country scale using satellite image time series, Remote Sensing 9 (2017). URL: https://www.mdpi.com/2072-4292/9/1/95. doi:10.3390/rs9010095.

[29] V. Bellet, M. Fauvel, J. Inglada, Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features, IEEE Transactions

on Geoscience and Remote Sensing (2023). URL: https://hal.science/hal-03781332. doi:10.1109/TGRS.2023.3234527.

[30] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering, et al., Monitoring vegetation systems in the great plains with erts, NASA Spec. Publ 351 (1974) 309.

[31] Z. Wang, Y. Ma, Y. Zhang, J. Shang, Review of remote sensing applications in grassland monitoring, Remote Sensing (2022). doi:10.3390/rs14122903.

[32] C. Pelletier, G. I. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sensing 11 (2019) 523.

[33] V. S. F. Garnot, L. Landrieu, Lightweight temporal self-attention for classifying satellite images time series, in: V. Lemaire, S. Malinowski, A. Bagnall, T. Guyet, R. Tavenard, G. Ifrim (Eds.), Advanced Analytics and Learning on Temporal Data, Springer International Publishing, Cham, 2020, pp. 171–181.

[34] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), San Diega, CA, USA, 2015.

[35] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research 18 (2017) 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[36] L. O. H. N. V. Chawla, K. W. Bowyer, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research (2002). doi:10.1613/jair.953.

[37] E. A. G. Haibo He, Yang Bai, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, IEEE International Joint Conference on Neural Networks (2008). doi:10.1109/IJCNN.2008.4633969.

[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[39] A. Ivanda, L. Šerić, M. Bugarić, M. Braović, Mapping chlorophyll-a concentrations in the kaštela bay and brač channel using ridge regression and sentinel-2 satellite images, Electronics 10 (2021) 3004.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[41] L. Breiman, Random forests, Machine Learning (2001). doi:10.1023/a:1010933404324.

[42] M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS journal of photogrammetry and remote sensing 114 (2016) 24–31.

[43] A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, Dive into Deep Learning, Cambridge University Press, 2023. https://D2L.ai.

[44] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, Deep learning classification of land cover and crop types using remote sensing data, IEEE Geoscience and Remote Sensing Letters 14 (2017) 778–782.

[45] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, P. M. Atkinson, A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification, ISPRS Journal of Photogrammetry and Remote Sensing 140 (2018) 133–144.

[46] C. Zhang, I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, P. M. Atkinson, Joint deep learning for land cover and land use classification, Remote sensing of environment 221 (2019) 173–187.

[47] B. Vinayak, H. S. Lee, S. Gedem, Prediction of land use and land cover changes in mumbai city, india, using remote sensing data and a multilayer perceptron neural network-based markov chain model, Sustainability 13 (2021) 471.

[48] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (cnn) in vegetation remote sensing, ISPRS journal of photogrammetry and remote sensing 173 (2021) 24–49.

[49] D. Guidici, M. L. Clark, One-dimensional convolutional neural network land-cover classification of multi-seasonal hyperspectral imagery in the san francisco bay area, california, Remote Sensing 9 (2017) 629.

[50] L. Zhong, L. Hu, H. Zhou, Deep learning based multi-temporal crop classification, Remote sensing of environment 221 (2019) 430–443.

[51] C. Liao, J. Wang, Q. Xie, A. A. Baz, X. Huang, J. Shang, Y. He, Synergistic use of multi-temporal radarsat-2 and venμs data for crop classification based on 1d convolutional neural network, Remote Sensing 12 (2020) 832.

[52] S. Ofori-Ampofo, C. Pelletier, S. Lang, Crop type mapping from optical and radar time series using attention-based deep learning, Remote Sensing 13 (2021) 4668.

[53] Z. Li, G. Chen, T. Zhang, Temporal attention networks for multitemporal multisensor crop classification, Ieee Access 7 (2019) 134677–134690.

[54] S. K. McFeeters, The use of the normalized difference water index (ndwi) in the delineation of open water features, International Journal of Remote Sensing (1996). doi:10.1080/01431169608948714.

[55] R. Escadafal, Remote sensing of arid soil surface color with landsat thematic mapper, Advances in space research 9 (1989) 159–163.

[56] J. Inglada, A. Vincent, M. Arias, B. Tardy, iota2-a25386, 2016. URL: https://doi.org/10.5281/zenodo.58150. doi:10.5281/zenodo.58150.

Supplementary materials

Review of models

Conventional machine learning models

Ridge Regression is a regularized linear model that seeks a linear relationship between the predictors (here the Sentinel-2 spectro-temporal features) and the output (here the observed mowing date) [38]. A regularized version was used to cope with the high number of spectro-temporal features [38, Chapter 3]. This method serves as a baseline for supervised model: its learning capacity is limited w.r.t. other non-parametric regression methods but has provided accurate results for some case, such as *chlorophyll-a* concentration mapping [39]. The regularization parameter value was selected using 10-folds cross-validation on the training data, as implemented in Scikit-learn [40].

Random Forest is a non-parametric and non-linear regression model introduced by Breiman [41]. It is an ensemble-based model learning multiple independent decision trees, using bootstraps of training samples and features. It has been widely used in remote sensing time series applications, mainly for land cover/use mapping [28] and estimation of continuous variables [42]. Several hyperparameters can be selected for training. The most important one is the number of decision trees in the forest. As shown in Inglada et al. [28], Fauvel et al. [19], setting it to a large value is enough to provide accurate results. In this experiment we found that 100 trees was a good compromise: increasing the values did not lead to an improvement of the precision while the processing complexity (time and memory footprint) was much higher. Random Forest was implemented in [40].

Deep learning models

One conventional and two advanced DL models were implemented: a Multiple Layer Perceptron (MLP), a 1D-CNN and the Lightweight Temporal Attention Encoder (LTAE), respectively. The MLP was composed of three "linear layer + batchnormalization layer + rectified linear activation layer" modules

and last linear output layer [43]. Such architecture has been widely used in remote sensing for land cover/use mapping [44, 45, 46] or land cover/use changes analysis [47].

The 1D-CNN was defined to perform along the temporal dimension, as in Kattenborn et al. [48], Kussul et al. [44], Guidici and Clark [49], Zhong et al. [50], Liao et al. [51], Pelletier et al. [32], to take into account the temporal dependence between the acquisition dates. From the MLP configuration, we replace the linear layer by a 1D convolutional layer and add max-pooling operation, as usually done with CNN models [43].

LTAE used temporal attention mechanism to make use of the acquisition dates [33]. Attention mechanism has showed to perform really well for land-cover mapping [52, 33, 53, 29]. The same architecture proposed by Garnot and Landrieu [33] was used in this work, just adapting the last layer and loss function to perform regression rather than classification.

Threshold-based method

We implemented a the specific mowing event detection algorithm introduced by Vroey et al. [10] and integrated into the Sen4CAP toolbox (`http://esa-sen4cap.org`) toto facilitate the monitoring of grassland management activities across Europe, aligning with the European Common Agricultural Policy. In our study, this method was adapted to detect mowing event date, since it was primarily designed to detect mowing event time interval.

Vroey et al. [10] proposed two independent change detection algorithms, whereby raw Sentinel-2 NDVI and Sentinel-1 VH-coherence time series were evaluated separately. In the final product, Sentinel-1 outputs were considered only when Sentinel-2 omitted events due to cloud cover. Here, we reproduced and adapted their Sentinel-2-based algorithm for evaluating pixel-based time series, as opposed to the original method that used object-based approaches.

To account for a mowing event, the original algorithm performed the following steps:

1. Each observation NDVI(t) is compared to the last available cloud-free observation NDVI($t-1$).
2. If the loss of NDVI, between NDVI(t) and NDVI($t-1$), is greater than 0.15 NDVI (NDVI(t) $<$ NDVI($t-1$) -0.15), a mowing event is considered. As an additional condition, two consecutive mowing events must be separated by a minimum temporal distance of 28 days, and if a mowing event is detected within the time interval $[t-1, t]$, it is assumed that the actual event took place within 60 days before t. If $[t-1, t]$ spans more than 60 days, the detection interval is adjusted to $[t-60, t]$. For each detected mowing event, the confidence level was estimated through a normalization function as follows:

$$f(x; \min, \max) = \max - (\max - \min) \times \exp(-x), \tag{1}$$

where x is the difference NDVI($t-1$) $-0.15-$ NDVI(t), $[\min, \max]$ were set to fit the confidence limits from 0.5 to 1.

The first mowing event among the four most confident detections was retained, as opposed to the original method that retained all four most confident detections. In contrast to the original method, where the time interval $[t-1, t]$ was kept for each detected mowing event, we retained the specific date t. Therefore, in our study, additional checks in step 2 were ignored.

Grassland management map

A map of grassland management practices -*mowed* or *unmowed*- was generated to constrain mowing date prediction to areas of mowed grassland. We performed a pixel-based classification task within a nationwide grassland mask (Figure 1), derived from permanent grassland plots declared in the 2022 LPIS (section 2.1). This database provides spatialized information on agricultural plot boundaries and crop types, but does not provides information about management practices.

Here, we trained a Random Forest classifier using a grassland management

practices dataset, derived from ground observations in 2022 (section 2.3). In this reference dataset, *mowed* class included 1 605 plots and *unmowed* class 660 plots (Table 2). Reference data were split into a 70% training dataset and a 30% test dataset, ensuring classes and sites representation through stratified sampling. Sentinel-2-based time series were used as predictor. In this dataset, in addition to spectral bands, we also computed three spectral indices: Normalized Difference Vegetation Index - NDVI [30], Normalized Difference Water Index - NDWI [54] and Brightness Index - BI [55].

The classification was done using IOTA^2 software [56]. Grassland management map achieved an overall precision of 90%, with *mowed* class showing an F-score of 0.93 and *unmowed* class exhibiting an F-score of 0.81. Findings showed that *mowed* class was slightly overestimated. In addition, in each plot of the initial grassland mask, resulting classes exhibited a coherent spatial structure, showing unimodal or bimodal intra-plot management patterns.

It is important to note that all quantitative evaluation results presented in this paper are not based on the grasslands management map, as they were computed on the reference data (observed *mowed* plots). This map was used for visual evaluation only

Tables and figures

Table 4: Algorithm-specific statistical summary. Each score value represents weighted mean of all sites. A site-specific score was weighted using the number of pixels used for the evaluation (32 123 pixels in average). For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. MAE, RMSE and Max_error are represented in days. For comparison purposes, Mean corresponds to an algorithm that always predicts the average of observed dates from training data (poor model). The lines are sorted based on R^2 values (descending order).

Algorithm	MAE	RMSE	R^2	Max_error
LTAE	5.63	9.13	0.52	59.58
LTAE_ADASYN	5.51	9.32	0.50	64.31
FCN_ADASYN	6.67	9.46	0.48	48.57
FCN	6.84	9.60	0.47	49.30
MLP_SMOTE	6.50	9.50	0.47	54.23
RF	6.80	10.09	0.40	56.69
MLP	7.02	10.11	0.40	54.28
Ridge	9.02	11.88	0.16	73.46
Mean	10.28	13.81	-0.10	42.06
Threshold	14.02	19.66	-1.36	73.24

Table 5: Algorithm-specific statistical summary. Each score value represents weighted mean of all sites. A site-specific score was weighted using the number of plots used for the evaluation (158 plots in average). For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. MAE, RMSE and Max_error are represented in days. For comparison purposes, Mean corresponds to an algorithm that always predicts the average of observed dates from training data (poor model). The lines are sorted based on R^2 values (descending order).

Algorithm	MAE	RMSE	R^2	Max_error
LTAE_ADASYN	5.43	8.94	0.58	48.03
LTAE	5.76	9.19	0.55	48.53
FCN_ADASYN	6.71	9.62	0.51	42.70
MLP_SMOTE	6.46	9.66	0.51	45.73
RF	6.61	9.73	0.50	45.76
FCN	6.95	9.94	0.48	44.05
MLP	6.86	10.04	0.47	44.05
Ridge	8.27	11.27	0.32	43.39
Mean	10.78	14.47	-0.10	41.14
Threshold	12.98	17.24	-0.64	56.38

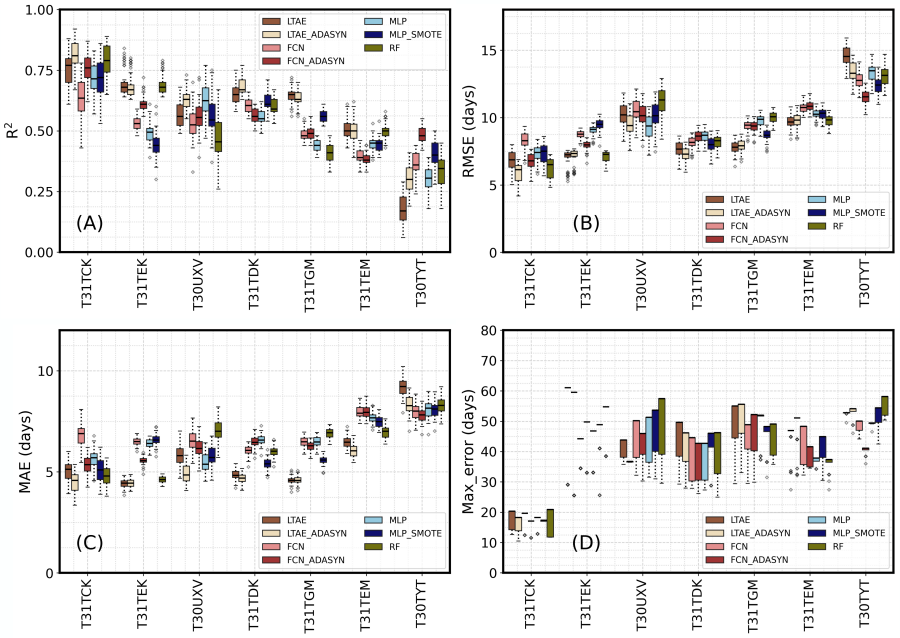

Figure 6: Algorithm-specific outputs in terms of (A) R^2, (B) RMSE, (C) MAE and (D) Max_error. The sites are represented on the x-axis. The color palette represents the algorithms. For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. The number of plots used for evaluation was: T31TCK (23.0), T31TEK (178.0), T30UXV (95.0), T31TDK (152.0), T31TGM (289.0), T31TEM (205.0) and T30TYT (167.0). Ridge, Mean and Threshold outputs are not shown in this figure.

Table 6: Site-specific statistical summary. Each score value represents mean of all models (except Ridge, Mean and Threshold outputs) and folds. For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. The number of plots used for the evaluation is represented by n. MAE, RMSE and Max_error are represented in days. The lines are sorted based on R^2 values (descending order).

Site	MAE	RMSE	R^2	Max_error	n
T31TCK	5.304771	6.948514	0.739886	17.523229	23.0
T31TDK	5.721743	8.059086	0.611000	41.570743	152.0
T31TEK	5.505600	8.044171	0.593800	48.275429	178.0
T30UXV	5.984714	10.191029	0.564000	44.685886	95.0
T31TGM	5.849600	8.970257	0.523257	48.181800	289.0
T31TEM	7.226886	10.177057	0.453571	41.949286	205.0
T30TYT	8.259800	13.000257	0.340571	50.503600	167.0

Table 7: Site-specific statistical summary. Each score value represents mean of all folds of the corresponding optimal model (in terms of R^2). For each site, fifty folds were synthetically generated for individual evaluation, each fold containing 70% of the randomly selected observations. The number of plots used for the evaluation is represented by n. MAE, RMSE and Max_error are represented in days. The lines are sorted based on R^2 values (descending order).

Site	Optimal model	MAE	RMSE	R^2	Max_error	n
T31TCK	LTAE_ADASYN	4.4704	5.8876	0.8144	16.3618	23.0
T31TEK	LTAE	4.4212	6.9634	0.6980	54.6360	178.0
T31TDK	LTAE_ADASYN	4.6706	7.2848	0.6830	42.6492	152.0
T31TGM	LTAE	4.5920	7.7784	0.6438	51.1618	289.0
T30UXV	MLP	5.4642	9.3572	0.6324	46.0896	95.0
T31TEM	LTAE	6.4670	9.6552	0.5084	44.9112	205.0
T30TYT	FCN_ADASYN	7.8062	11.5166	0.4850	40.7350	167.0

Table 8: Site-specific LTAE outputs aggregated at plot-level. Each score value was calculated from all plots. The number of plots used for the evaluation is represented by n. MAE, RMSE and Max_error are represented in days. The lines are sorted based on R^2 values (descending order).

Site	MAE	RMSE	R^2	Max_error	n
T31TCK	5.02	6.91	0.76	20.36	33.0
T31TEK	4.34	6.76	0.72	61.04	255.0
T31TDK	4.85	7.66	0.65	49.61	217.0
T31TGM	4.56	7.74	0.65	55.04	413.0
T30UXV	5.86	10.29	0.56	43.82	136.0
T31TEM	6.52	9.72	0.51	46.91	293.0
T30TYT	9.21	14.61	0.18	52.78	239.0

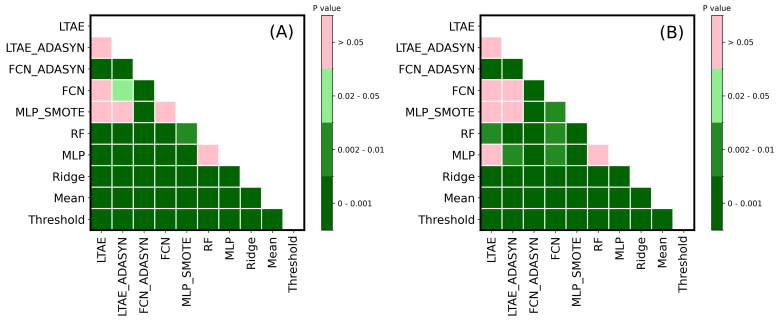

Figure 7: T-test between each pair of algorithms in terms of (A) R^2 and (B) RMSE. The color palette represents p-value scores from the statistical test. Here, the number of observations by algorithm was 350 (7 sites x 50 folds each).

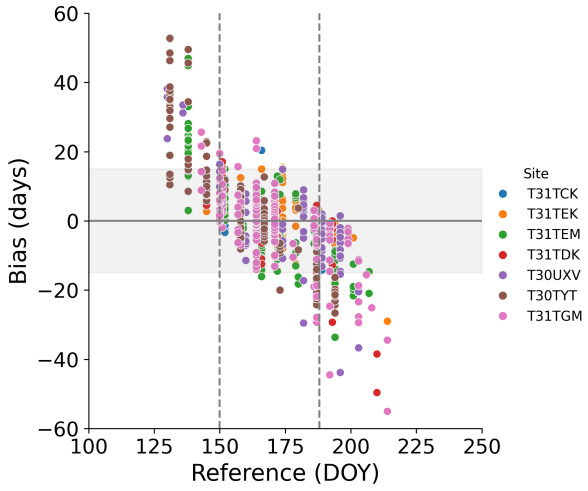

Figure 8: Site-specific LTAE outputs aggregated at plot-level. For each plot, the predicted date was compared to the observed date. The x-axis represents the observed date, while the y-axis represents the prediction error. Each plot is represented by a dot on the graph, and its color indicates the observation site. A positive bias means that predicted date is higher than observed date, while a negative bias means that predicted date is lower than observed date. The gray band shows a tolerance margin of ± 15 days. The dashed vertical lines represent the average value of the 10th (left, 150.0 DOY) and 90th (right, 188.0 DOY) percentiles. These percentiles were derived from each site's training dataset, before averaging.