



Article

Multi-Year Cropland Mapping Based on Remote Sensing Data: A Case Study for the Khabarovsk Territory, Russia

Konstantin Dubrovin ^{1,*}, Andrey Verkhoturov ¹, Alexey Stepanov ² and Tatiana Aseeva ²

¹ Computing Center Far Eastern Branch of the Russian Academy of Sciences, 680000 Khabarovsk, Russia; andrey@ccfebras.ru

² Far-Eastern Agriculture Research Institute, 680521 Vostochnoe, Russia; a.stepanov@ccfebras.ru (A.S.); aseeva59@mail.ru (T.A.)

* Correspondence: nob_keeper_93@mail.ru

Abstract: Cropland mapping using remote sensing data is the basis for effective crop monitoring, crop rotation control, and the detection of irrational land use. Classification using Normalized Difference Vegetation Index (NDVI) time series from multi-year data requires additional time costs, especially when sentinel data are sparse. Approximation by nonlinear functions was proposed to solve this problem. Time series of weekly NDVI composites were plotted using multispectral Sentinel-2 (Level-2A) images at a resolution of 10 m for sites in Khabarovsk District from April to October in the years 2021 and 2022. Missing values due to the lack of suitable images for analysis were recovered using cubic polynomial, Fourier series, and double sinusoidal function approximation. The classes that were considered included crops, namely, soybean, buckwheat, oat, and perennial grasses, and fallow. The mean absolute percentage error (MAPE) of each class fitting was calculated. It was found that Fourier series fitting showed the highest accuracy, with a mean error of 8.2%. Different classifiers, such as the support vector machine (SVM), random forest (RF), and gradient boosting (GB), were comparatively evaluated. The overall accuracy (OA) for the site pixels during the cross-validation (Fourier series restored) was 67.3%, 87.2%, and 85.9% for the SVM, RF, and GB classifiers, respectively. Thus, it was established that the best result in terms of combined accuracy, performance, and limitations in cropland mapping was achieved by composite construction using Fourier series and machine learning using GB. Similar results should be expected in regions with similar cropland structures and crop phenological cycles, including other regions of the Far East.

Keywords: remote sensing; crop mapping; Sentinel-2; NDVI



Citation: Dubrovin, K.; Verkhoturov, A.; Stepanov, A.; Aseeva, T. Multi-Year Cropland Mapping Based on Remote Sensing Data: A Case Study for the Khabarovsk Territory, Russia. *Remote Sens.* **2024**, *16*, 1633. <https://doi.org/10.3390/rs16091633>

Academic Editors: Jadu Dash, Wenjiang Huang, Kun Wang and Tiecheng Huang

Received: 4 March 2024

Revised: 15 April 2024

Accepted: 24 April 2024

Published: 3 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An effective approach for the information support of precision agriculture involves the quantitative assessment of cropland area and the analysis of cropland structure. Currently, the use of machine learning methods for the development of global and regional cropland masks, as well as masks of individual crops, has received much attention [1–4]. Remote sensing has emerged as a promising method for generating such masks. In particular, Sentinel imagery has made it possible to obtain information with the simultaneous coverage of significant areas with high periodicity, which is practically impossible to realize when using ground-based data.

The issue of cropland classification can be addressed by aggregating the values of visible and infrared spectrum bands to serve as input data. This process yields vegetation indices, such as the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), normalized difference water index (NDWI), and others. A popular method for constructing crop masks is the use of a single multispectral image [5–7]. However, this approach is more suitable for classifying land cover than for crops that are characterized by significant variations in vegetation index values during the vegetation season. Therefore,

time series of vegetation indices are extensively used. For instance, Luo et al. used Sentinel-2 *NDVI* time series data (with various phenological characteristics) to map crops in northern China [8], computing 20 informative features (time series and phenological) to achieve a classification accuracy of 93% using the random forest algorithm (RF). Ouzemou et al. used Landsat *NDVI* time series data for cropland mapping after resharpening them to a resolution of 15 m (10 images per season) [9].

A major disadvantage of optical data is the significant influence of weather phenomena, such as different types of clouds, haze, aerosols, and cloud shadows, on the vegetation indices' values. The expert or automated removal of unsuitable images is not a simple task and may often lead to a significant sparsity of time series. The process of the computation of vegetation indices' temporal composites can help address this issue by reducing the negative effects of atmospheric indices. Computing composites also reduces the amount of input data. For example, Erdanaev et al. identified agricultural crops in Uzbekistan, including cotton, wheat, rice, and others, using Landsat-8 and Sentinel-2 monthly *NDVI*, *EVI*, and *NDWI* composites [10]. Hao et al. also used Landsat-8 and Sentinel-2 imagery to construct 15-day *NDVI* and *EVI* composites [11]. In both cases, the authors were able to achieve a high classification accuracy. However, when working with data from different sensors, it is essential to perform data projection and resampling procedures. Low-resolution data are also used for land cover and crop mapping. For example, 16-day Moderate Resolution Imaging Spectroradiometer (MODIS) *NDVI* composites were used for classification [12,13]. However, using low spatial products, such as imagery with a spatial resolution of 250 m, does not allow for accurate crop mapping for relatively small sites or sites with multiple crops growing simultaneously.

To create an accurate classifier, it is necessary to account for the inter-annual variability associated with climatic factors, hydrological conditions, and anthropogenic factors. Consequently, studies using multi-year satellite data are more promising. For instance, Feyisa et al. classified croplands in Ethiopia using 8-day *NDVI* composites and the RF method and achieved the highest accuracy when using 3-year MODIS data [14]. Crop mapping using satellite data for previous (or, in the case of retrospective prediction, subsequent) years is also of great interest. For example, Konduri et al. used MODIS *NDVI* data for 7 previous years to generate cropland maps in the US in 2015–2018 [15]. Later, a classifier trained on monthly composites of Sentinel-2 bands and indices in 2020 performed a retrospective prediction of crops grown in 2019 using the RF algorithm to recover crop rotation information [16].

Often, there may be long periods between the dates of optical images suitable for analysis in regions that are characterized by distinct rainy seasons, long cloudy periods, and the presence of aerosols or fog during the vegetation season. If such a pause lasts for 2 weeks or longer, the calculation of composites (even over 15–16 days) is insufficient for the generation of continuous time series. In such cases, the missing values are recovered through interpolation, using neighboring values [17] and splines [18] or filters [19,20]. However, a more effective approach to filling the gaps in the data is time series function fitting. In this case, a function with a shape similar to that of the seasonal course of the vegetation index is selected. Extensive research has been conducted on the selection of functions for the function-fitting of the most popular vegetation indices. For example, polynomials [11], the Gauss function [21], logistic functions [22,23], and the Fourier series [24] have been used to fit *NDVI* time series.

A characteristic feature of the Russian Far East is the lack of reliable cropland maps and databases. Various machine learning (ML) models have been developed to validate ground-based observations, often using existing cropland maps for accuracy assessment [25–27]. Subsequently, the accuracy of global and federal cropland maps developed for different climatic conditions, crops, varieties, and agricultural practices is comparatively low. The lack of a regional database that takes into account crop phenology in the Russian Far East poses a serious challenge, hindering rational land use and the development of the agro-industrial complex as a whole. In addition, the inadequate control of crop rotation and

agricultural practices in the region results in non-compliance with crop rotations and the dominance of soybean. In some municipalities of the Far East, the share of soybean in the total amount of cropland exceeds 90% (<https://rosstat.gov.ru>, accessed on 15 April 2024). Repeating the cultivation of soybean as the most profitable crop, in cases of non-compliance with the crop rotation, leads to a decrease in soil fertility. To maintain fertility, the inclusion of oat and perennial grasses in crop rotation is recommended [28]. Gajduchenko et al. conducted 20-year research and demonstrated that crop rotation increases the soybean yield [29]. Moreover, a high-yielding crop such as buckwheat can be used in crop rotation with soybean for additional benefits in terms of the overall yield.

Various factors, including waterlogging, the decline in soil fertility, and the overgrowth of weeds and trees, coupled with primitive methods of fallow land identification, have resulted in a substantial amount of unused agricultural land in the region. Therefore, the introduction of forested and waterlogged fallow land into crop rotation is one of the most important tasks of agriculture in the Russian Far East. Remote sensing methods have not been previously used to solve these important problems in this region.

The main objective of this research is to develop an automated method for cropland classification using Sentinel-2 weekly *NDVI* composites and data reconstruction algorithms. An important problem that this study intends to address is to establish the feasibility of classification using time series over two years of observation. An approach that has not been used before for cropland mapping in the Russian Far East is relevant due to the large number of cloudy days during the growing season, as well as the different dates of images in different years, which create difficulties for the researcher. We have proposed to consider the possibility of using function-fitting to restore composite values. The study will evaluate the function-fitting accuracy using different algorithms and the classification accuracy using different ML methods.

2. Materials and Methods

2.1. Study Area

The study area is located in Khabarovsk Municipal District in the south of the Middle Amur Lowland. This region is on the right bank of the Amur River east of Khabarovsk between 48.31° and 48.64° N latitude and 134.81° and 135.57° E longitude (Figure 1). These lands lie in the temperate belt and experience a monsoon type of climate, i.e., winters are quite cold while summers are warm and sometimes hot, with high precipitation. The average annual precipitation is about 700 mm. The duration of the frost-free period is 130–150 days. The short duration of the vegetation season does not allow for the growth of winter crops.

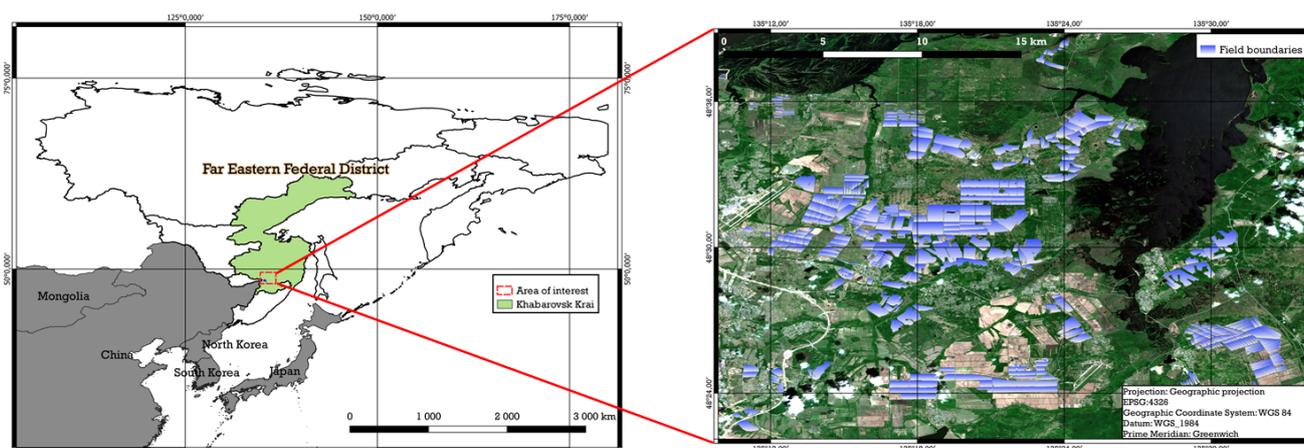


Figure 1. Study area.

The climate and soil conditions favor the cultivation of leguminous crops, primarily soybean. In 2022, soybean grew on 59% of cropland in Khabarovsk District (10,722 ha), while oat and buckwheat occupied 8% (1446 ha) and 2% (327 ha) of total cropland, respectively.

2.2. Ground Truth Data

Marked sites of the Far-Eastern Agriculture Research Institute and agricultural enterprises of the region with crop information were used as ground observation data. Information on crop rotation in 2021 and 2022 was provided by the heads of organizations and verified during ground observations. Overall, in 2021, information on 50 sites with a total area of 1130 ha was collected. In 2022, information was collected on 80 sites, covering an area of 1992 ha. In 2021, the share of fallow land and soybean exceeded 30% of the total area, while buckwheat occupied 26% of these lands, and oat and perennial grasses together occupied just over 10%. In 2022, the number of sites with oat increased from 4 to 17 (with a corresponding increase in area from 67 to 511 ha), while those with perennial grasses increased from 2 to 29 (with a corresponding increase in area from 60 to 528 ha). Table 1 shows the number and total area of sites for each class.

Table 1. Number of sites and the area for each class included in this study.

| Class | Season | Sites | Area, ha |
|-------------------|--------|-------|----------|
| Fallow | 2021 | 15 | 356 |
| | 2022 | 12 | 282 |
| Soybean | 2021 | 15 | 347 |
| | 2022 | 14 | 355 |
| Oat | 2021 | 4 | 67 |
| | 2022 | 17 | 511 |
| Perennial grasses | 2021 | 2 | 60 |
| | 2022 | 29 | 528 |
| Buckwheat | 2021 | 14 | 300 |
| | 2022 | 8 | 316 |
| Overall | 2021 | 50 | 1130 |
| | 2022 | 80 | 1992 |

2.3. Crop Phenology

Active vegetation of soybean in Khabarovsk District begins in calendar weeks 26–28 (late June to mid-July), the vegetation maximum is reached by weeks 32–33 (mid-August), and the biomass then decreases until the end of the growing season (soybean harvesting is carried out not earlier than mid-October). Oat is sown in the first half of May (weeks 18–19), with the vegetation maximum occurring in early July (weeks 26–27); it is then harvested and the *NDVI* declines sharply. Perennial grasses are sown under the cover of grain crops in the previous year, allowing vegetation to begin immediately with the onset of the frost-free period. Biomass growth occurs at a rapid pace, reaching its peak by calendar weeks 25–27 (second half of June to early July). Following the summer mowing, the grass regrows rapidly.

In 2021, the experimental sowing of buckwheat was resumed in Khabarovsk District after a long break. Its vegetation period is approximately only 70 days, which allows for variations in the sowing and harvesting dates. Thus, in 2021, sowing was carried out in the first half of July (calendar weeks 27–28), and harvesting was performed in the second half of September (calendar weeks 38–39). Such dates are considered optimal for the local buckwheat variety. In 2022, sowing was carried out a month earlier, commencing in early June (weeks 22–23). Buckwheat maturation occurred by mid-August, with the end of the vegetative phase occurring in week 34, and harvesting took place in late August and early September.

The fallow land represents unused cropland with grass, shrubs, and woody vegetation. The insufficient melioration of cropland is an urgent problem in the Far East, resulting in waterlogging. Vegetation on fallow land experiences growth from the beginning of spring, maintains a consistently high level during the summer, and starts to wilt with a gradual decrease in biomass during the fall. Crop phenology for soybean, oat, buckwheat, and perennial grasses (with average temperatures) is presented in Figure 2.

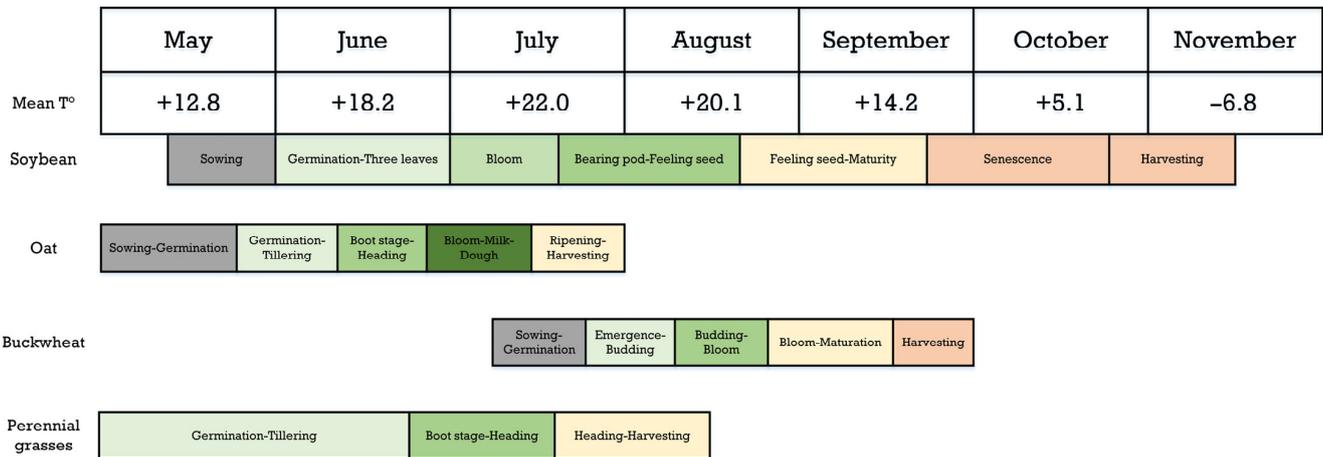


Figure 2. Crop phenology in Khabarovsk District.

2.4. Imagery Processing and NDVI Calculation

Multispectral Sentinel-2 data (Level-2A product) were accessed from the Copernicus Open Access Hub service. The Graph Processing Framework (GPF), available on the European Space Agency (ESA) Sentinel Application Platform (SNAP) v.9.0 (<http://step.esa.int/main>, accessed on 21 December 2023; ESA, Paris, France), was used to create graphs and set up chains for batch processing, as shown in Figure 3.

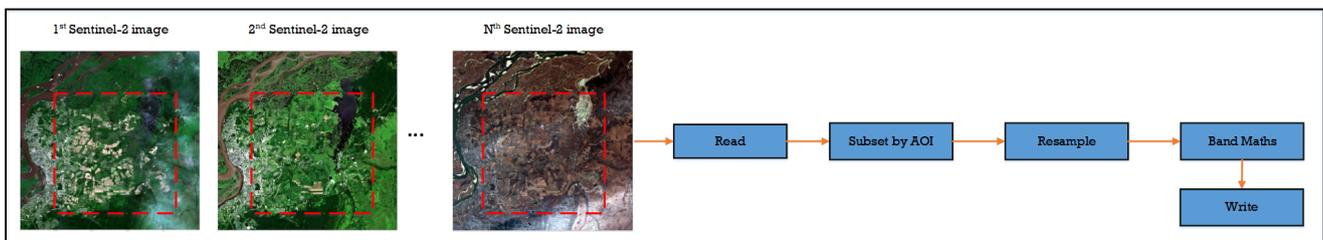


Figure 3. Flowchart for NDVI time series generation.

NDVI time series were calculated for images that fulfilled the condition of the cloud cover being less than 20% for the scene (Figure 4).

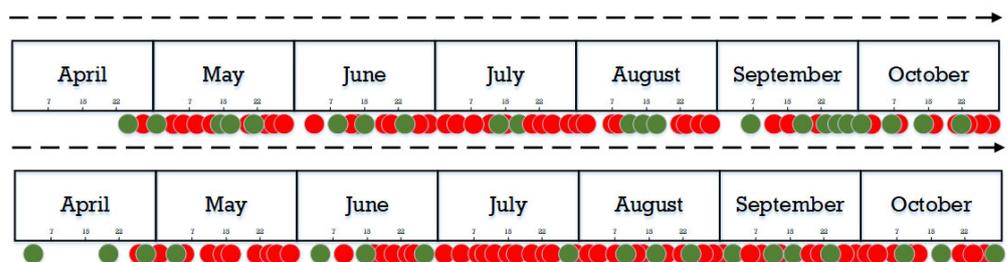


Figure 4. Selected Sentinel-2 scenes (marked in green) for the 2021 and 2022 seasons for the study area. The green color marked scenes with no more than 20% cloudiness and the red color marked scenes with over 20% cloudiness.

Additionally, for the selected images, Scene Classification (SCL) products were used together with Quality Indicators (QI) for clouds. Pixels were masked if they were contained in at least one of the following bands: “scl_cloud_high_proba”, “scl_cloud_shadow”, “opaque_clouds”, “scl_cloud_medium_proba”, “cirrus_clouds”, and “scl_thin_cirrus”. Initially, SCL products were resampled to 10 m. Further, using the near-infrared ($B8$) and red ($B4$) bands, the $NDVI$ for each scene was calculated. The $NDVI$ was calculated as follows:

$$NDVI = \frac{B8 - B4}{B8 + B4} \quad (1)$$

Before the dataset construction, we additionally performed “ 2σ -filtering”—for each field, the mean value and standard deviation of the $NDVI$ by site were calculated for each date. Pixels when over half of the time series values did not fall within the 2σ -interval were considered anomalous and were not included in the dataset. A total of 0.6% of the series in 2021 and 0.8% of the series in 2022 were filtered out, which shows insignificant $NDVI$ variability across the site. A flowchart of the current research (starting from time-series filtration, including time series restoration and classification) is presented in Figure 5.

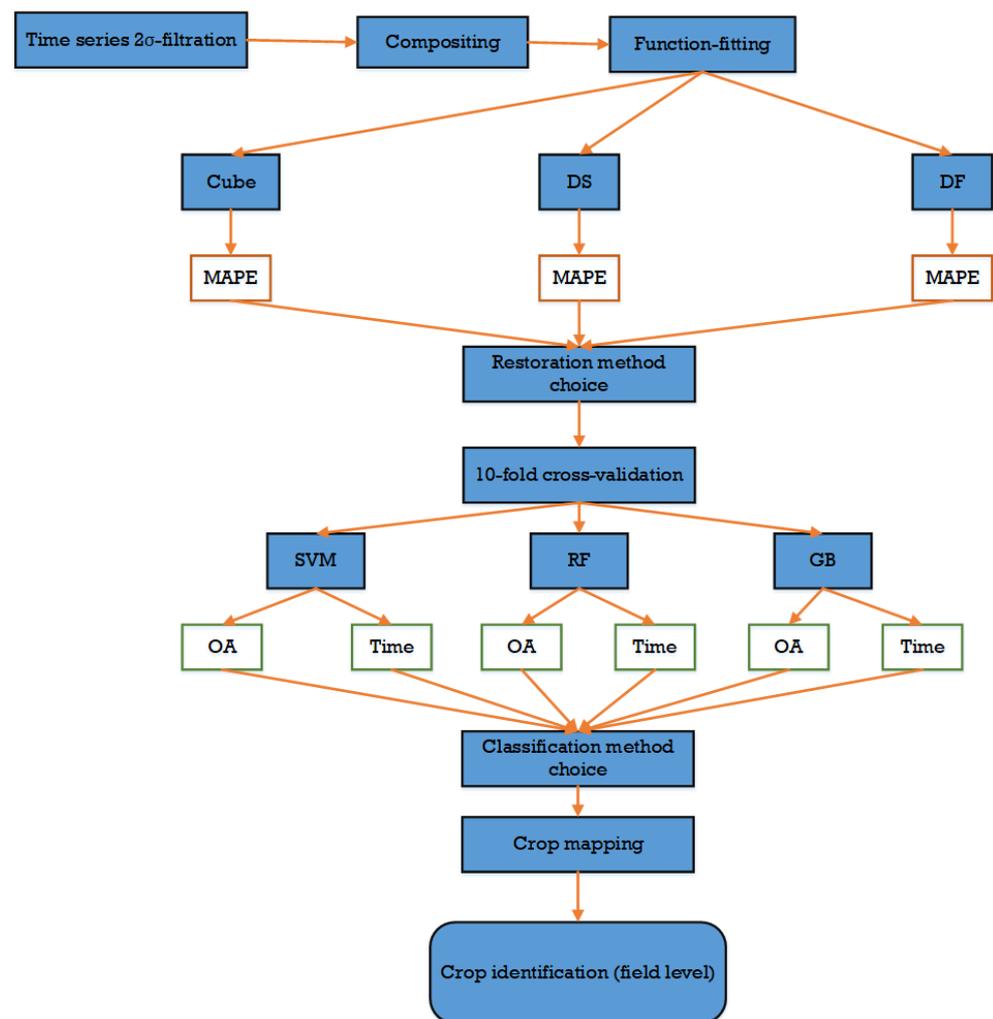


Figure 5. Flowchart for crop mapping.

2.5. Composites Construction and Data Restoration

To equalize the number of observations in the time series, we calculated weekly $NDVI$ composites. Within each time series, the values were divided by calendar weeks; in the case of several values within one calendar week, the average value was calculated. Thus,

all the *NDVI* composites time series included 27 observations (from 17 to 43 calendar weeks—from the end of April to the end of October).

The composites reduced the sparsity in the dataset, this was observed only for weeks with images. The dataset remained highly sparse, missing 36% of the values for 2021 and 47% of the values for 2022. An example of a time series with a large number of missing observations is shown in Figure 6.

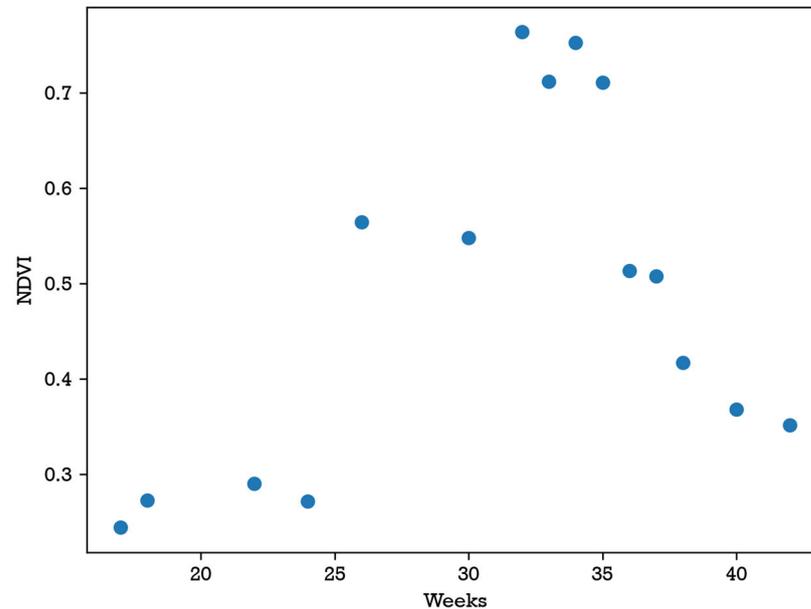


Figure 6. *NDVI* time series example for 2022.

To recover the missing values, we fitted each time series. The independent variable was the number of calendar weeks, while the dependent variable was the *NDVI* value. The parameters of the function were obtained for each time series. Then, using the number of the week as an argument, we obtained the missing values of the time series. The *lmfit* library (<https://lmfit.github.io/lmfit-py>; accessed on 15 November 2023) of the Python programming language was used for fitting function parameters by the least squares' method—specifically, the Levenberg–Marquardt algorithm [30]. The fitting functions were the cubic polynomial, double sinusoidal function, and Fourier series.

2.5.1. Cubic Polynomial

Time series function fitting using polynomials is among the simplest and most commonly used methods for time series reconstruction [11,23,31]. A cubic polynomial (Cube) has only four parameters and has the following form:

$$y = ax^3 + bx^2 + cx + d, \quad (2)$$

where a , b , c , and d represent the parameters of the model.

2.5.2. Double Sinusoidal Function

Trigonometric polynomials have demonstrated their effectiveness as fitting functions for periodic dependencies, which may include time series of vegetation indices [32]. In particular, the double sinusoidal function (DS) is suitable for dependencies that have two characteristic peaks. The DS function has six parameters and has the following general form:

$$y = a_1 \sin(b_1 x_1 + c_1) + a_2 \sin(b_2 x_2 + c_2) \quad (3)$$

where a_1 , b_1 , c_1 , a_2 , b_2 , and c_2 represent the parameters of the function.

2.5.3. Fourier Series

Fourier series are the best-known trigonometric series. Fourier series decomposition is often used in signal processing and serves to extract the amplitude component, thus allowing for the elimination of noise. There are several examples of the successful application of such decomposition for time series of vegetation indices [33,34]. Fourier series enables the representation of a complex relationship as the sum of simple trigonometric functions. In our study, we use the first two terms of the Fourier series expansion (DF):

$$y = a_0 + a_1 \cos(xw) + b_1 \sin(xw) + a_2 \cos(2xw) + b_2 \sin(2xw) \quad (4)$$

As in the case of the DS, the approximation is reduced to the search for six parameters represented by w , a_0 , a_1 , a_2 , b_1 , and b_2 . We calculated the MAPE for the averaged time series of all classes for 2021 and 2022, fitted by different numbers of terms in the Fourier series (Table S2). It was found that the mean error for one term and two terms differed significantly ($p < 0.05$), while the mean error for two terms and for three terms did not differ significantly ($p > 0.05$). An increase in the number of terms of Fourier series leads to an increase in the number of model parameters, which, first, requires more observations in time series (which can cause limitations in case of a lack of uncloudy data), and second, increases the complexity of the calculations and interpretation of the model. Therefore, using the first two terms of the Fourier series in the model is optimal.

2.5.4. Function Fitting Accuracy Evaluation

The mean absolute percentage error (MAPE) was calculated for each time series to compare fitting functions and to select the function that best fit the original data:

$$MAPE_i = \frac{1}{n-1} \sum_n \left| \frac{v_{ij}^{pr} - v_{ij}^{obs}}{v_{ij}^{obs}} \right| 100 \quad (5)$$

where i is the series index, v_{ij}^{pr} is the predicted NDVI in week j , v_{ij}^{obs} is the real NDVI in week j , and n is the week number with the known NDVI for series i . The MAPE values obtained for each series were averaged across classes and the entire dataset, as follows:

$$MAPE_{class} = \frac{1}{m} \sum_{i \in class} MAPE_i \quad (6)$$

$$MAPE_{overall} = \frac{1}{N} \sum_i MAPE_i, \quad (7)$$

where $class$ is the name of the class, m represents the number of series in the class, and N is the number of series in the dataset. The function with the lowest $MAPE_{overall}$ is used to reconstruct the time series when generating datasets for ML.

2.6. Machine Learning

The dataset for 2021 and 2022 was obtained after filtering. The number of series in the datasets is presented in Table 2. Partitioning into validation for 10-fold cross-validation sets was performed based on site locations to prevent the simultaneous inclusion of series from the same site in sets, as this could result in overfitting. A total of 295,673 labeled NDVI series belonging to 130 sites, with a total area of 3122 ha, were used in cross-validation. In general, the dataset can be considered balanced, thus avoiding distortions in the classification result.

To create classifiers, we used three ML algorithms, namely, support vector machines (SVM), RF, and gradient boosting (GB). These algorithms, along with other popular ML algorithms, were implemented using the scikit-learn package (<https://scikit-learn.org/stable/>; accessed on 15 November 2023) in the Python language.

Table 2. Number of time series for each class included in the dataset.

| Class | Series |
|-------------------|---------|
| Fallow | 62,156 |
| Soybean | 66,259 |
| Oat | 58,550 |
| Perennial grasses | 51,315 |
| Buckwheat | 57,393 |
| Overall | 295,673 |

2.6.1. Support Vector Machines

The SVM algorithm is widely applied in regression and classification tasks, including cropland mapping [9,35,36]. The essence of SVM is to determine the optimal hyperplane within the feature space that effectively separates points belonging to different classes. The competitive advantages of the SVM include efficiency with a large number of features, memory economy, and flexibility in kernel feature selection. In this study, a high-performance variation of the LinearSVC algorithm using a linear kernel was used. The parameters of SVM are the penalization norm, loss function, and multiclass strategy. The main parameter of the algorithm is the regularization coefficient C .

2.6.2. Random Forest

The RF algorithm is based on averaging the results of the ensembles of trees through bagging, with each tree being trained on a subsample of the dataset [37]. The RF method is used in both regression and classification problems. It is currently the most popular algorithm in cropland mapping [25,27,37]. The advantages of RF include high performance (especially for large datasets), high accuracy, and robustness regarding the presence of noise in the data [9,38]. Its parameters are the quality criterion, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features for each split. However, the main parameter affecting the performance of RF is the number of trees.

2.6.3. Gradient Boosting

Similar to RF, the GB [39] algorithm is based on an ensemble of trees. The main difference in this algorithm is the fact that training is performed sequentially. During each iteration, the deviations in predictions made by the already trained ensemble on the training sample are computed. The next model to be added to the ensemble will predict these deviations. New trees are added to the ensemble as long as the error decreases or until one of the “early stopping” rules is satisfied. GB is considered a flexible, computationally fast algorithm that works even with sparse data. It has been used in prior research for land cover mapping tasks [40–42]. Histogram-based GB is a simplified and fast implementation of GB in the scikit-learn library, developed to classify large datasets. These fast estimators first group the input samples X into integer-valued bins (typically, 256 bins). This significant reduction in the number of splitting points to consider enables the algorithm to utilize integer-based data structures (histograms) instead of relying on sorted continuous values when constructing the trees. The set of algorithm parameters is the minimum number of samples in a leaf node, the maximum number of leaves in a node, and the number of iterations.

2.6.4. Mapping Accuracy Evaluation

The GridSearchCV tool was used to optimize the key parameters of the algorithms. The parameters selection and algorithms comparison were performed using 10-fold cross-validation (10 sets by 13 sites). The random state control was used to ensure the reproducibility of the results. In this study, the metrics used to compare the algorithms were overall accuracy (OA), user accuracy (UA), performance accuracy (PA), and $F1$. These metrics were calculated from the confusion matrix.

The confusion matrix is a square matrix of the dimension $n \times n$, where n represents the number of classes. In the confusion matrix, each column represents the actual class and each row represents the predicted class. Each element of the matrix X_{ik} denotes the number of pixels of class k that fall into class i during classification. The diagonal elements of the matrix represent the number of correctly classified pixels. The ratio of the sum of the diagonal elements (true positive pixels TP) to the sum of all elements of the error matrix N (expressed as a percentage) is given by the OA metric:

$$OA = \frac{TP}{N} * 100\%. \quad (8)$$

The confusion matrix also shows the number of pixels of class k ($k \in [1, n]$) that are incorrectly assigned to other classes (false negative pixels FN_k). PA represents the map accuracy from the point of view of the map maker or producer. This metric represents the frequency at which real features on the ground are correctly shown on the classified map or the probability that a certain type of land cover of a given area on the ground is classified correctly. PA is the complement of the omission error and is given by the number of reference sites classified accurately divided by the total number of reference sites for that class:

$$PA_k = \frac{TP_k}{TP_k + FN_k} * 100\%. \quad (9)$$

In the confusion matrix, it is also possible to count the number of pixels erroneously assigned to class k (FP_k). UA represents the accuracy from the point of view of a map user rather than the map maker. UA represents the frequency at which a class on the map is present on the ground, which is referred to as reliability. It is the complement of the commission error and is given by the total number of correct classifications for a particular class divided by the sum of row elements:

$$UA_j = \frac{TP_k}{TP_k + FP_k} * 100\%. \quad (10)$$

The $F1$ score can be interpreted as the harmonic mean of precision and recall, which varies from 0 to 1. The relative contributions of precision and recall to the $F1$ score are equal:

$$F1_k = \frac{2 * UA_k * PA_k}{UA_k + PA_k} \quad (11)$$

Based on the results of comparing the metrics, the algorithm with the most suitable parameters for cropland mapping was selected. It is also important to determine the number of sites where the crop was correctly identified to evaluate the performance of the best classifier. The number of pixels assigned to a particular class by the classifier determined the crop in each site as the class label with the largest number of pixels assigned to the site. The number and proportion of sites with correctly identified crops were used to assess the classification accuracy at the parcel level.

3. Results

3.1. NDVI Time Series Restoration

Figure 7 shows the crop and year-averaged $NDVI$ composite series that were reconstructed using function fitting (Figure 7a–f). Figure 7a,b show that the Cube fit the $NDVI$ time series rather roughly, causing significant breaks in the seasonal curve. In particular, this pattern was observed at the beginning of the growing season—at week 21 in 2021 for perennial grasses and oats and at week 22 in 2022 for oat and buckwheat. In June 2021 (calendar week 25–26), an uncharacteristic drop in the $NDVI$ was observed for oat, buckwheat, and perennial grasses. In mid-August in 2021 (weeks 33 and 34), there was an underestimation of the $NDVI$ for soybean near the peak and an overestimation for soybean and perennial grasses (just after harvesting). In 2022, there was a sharp reversal in the

NDVI for soybean, buckwheat, and perennial grasses at week 23, with mid-season spikes for perennial grasses. The most serious anomalies using the Cube were observed at the end of the season—the NDVI of oat and perennial grasses increased at the end of October (week 43), but the NDVI of soybean was observed to exhibit a value less than 0. In general, the use of Fourier series and DS allows for the generation of time series without anomalous omissions, even in periods with long absences of satellite observations (such as weeks 29–31 in 2021 and weeks 19–21 and 27–29 in 2022). Overall, the restored seasonal NDVI pattern is consistent with the crop phenology.

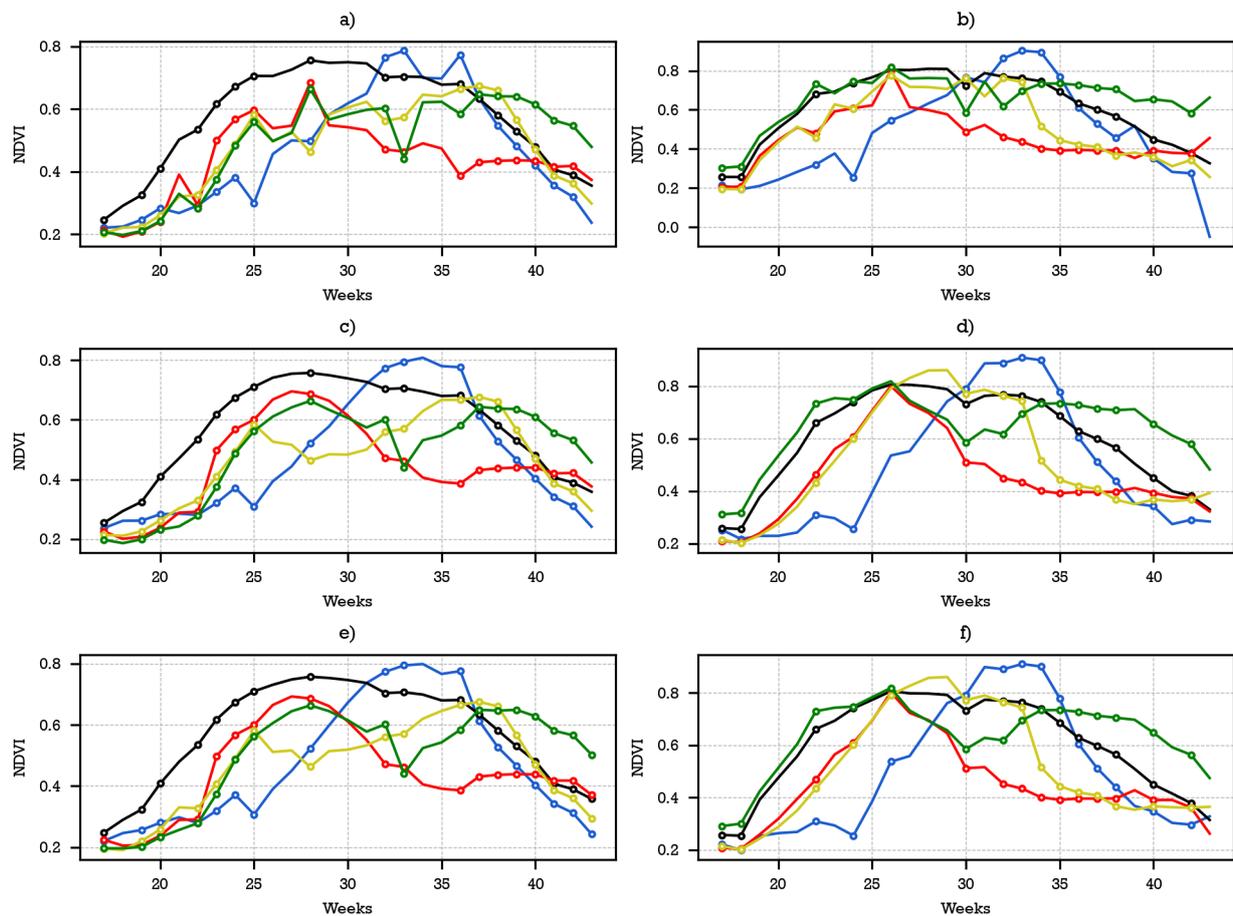


Figure 7. Restored NDVI composite time series: (a) Cube fitted in 2021, (b) Cube fitted in 2022, (c) DF fitted in 2021, (d) DF fitted in 2022, (e) DS fitted in 2021, (f) DS fitted in 2022. The points corresponding to the weeks when satellite observations were made are marked on the graphs.

An evaluation of the accuracy of the time series function fitting (in weeks with known NDVI) is presented in Table 3.

For the Cube, the largest error was observed for soybean ($MAPE_{soybean} = 26.1\%$). The $MAPE$ values for oats and buckwheat also exceeded 15% ($MAPE_{oat} = 19.8\%$, $MAPE_{buckwheat} = 17.4\%$), while that of perennial grasses was lower ($MAPE_{grasses} = 13\%$). A high approximation accuracy was achieved only for unused land NDVI fitting ($MAPE_{fallow} = 7.8\%$), characterized by the most stable pattern. Thus, $MAPE_{overall}$ for the Cube function was 16.8% (significantly higher than for other methods, Tukey test, $p < 0.01$), making it unsuitable for the recovery of the NDVI series.

Table 3. MAPE for different functions, crops, and seasons.

| Function | Crop | Season | MAPE, % | |
|-------------------|-------------------|---------|---------|------|
| Cube | Soybean | 2021 | 22.0 | |
| | | 2022 | 30.3 | |
| | Fallow | 2021 | 8.3 | |
| | | 2022 | 7.4 | |
| | Oat | 2021 | 24.2 | |
| | | 2022 | 15.4 | |
| | Buckwheat | 2021 | 15.2 | |
| | | 2022 | 19.6 | |
| | Perennial grasses | 2021 | 14.0 | |
| | | 2022 | 12.0 | |
| | DF | Soybean | 2021 | 11.6 |
| | | | 2022 | 12.0 |
| Fallow | | 2021 | 4.3 | |
| | | 2022 | 6.0 | |
| Oat | | 2021 | 6.9 | |
| | | 2022 | 8.3 | |
| Buckwheat | | 2021 | 10.1 | |
| | | 2022 | 7.1 | |
| Perennial grasses | | 2021 | 8.1 | |
| | | 2022 | 7.7 | |
| DS | | Soybean | 2021 | 12.9 |
| | | | 2022 | 11.5 |
| | Fallow | 2021 | 5.2 | |
| | | 2022 | 6.2 | |
| | Oat | 2021 | 7.1 | |
| | | 2022 | 8.9 | |
| | Buckwheat | 2021 | 14.3 | |
| | | 2022 | 7.2 | |
| | Perennial grasses | 2021 | 6.4 | |
| | | 2022 | 7.8 | |

The use of DS allowed for a nearly 50% reduction in error in comparison with the Cube ($MAPE_{overall} = 8.7\%$). The most significant error was observed for soybean ($MAPE_{soybean} = 12.2\%$) and buckwheat ($MAPE_{buckwheat} = 10.7\%$) due to the unsuccessful fitting of the composites for 2021. For the other classes, the approximation accuracy exceeded 90% ($MAPE_{oat} = 7.9\%$, $MAPE_{grasses} = 7.1\%$, $MAPE_{fallow} = 5.7\%$).

The lowest error ($MAPE_{overall} = 8.2\%$) was observed when using DF. The highest accuracy was achieved for soybean ($MAPE_{soybean} = 11.8\%$), buckwheat ($MAPE_{buckwheat} = 8.6\%$), oat ($MAPE_{oat} = 7.6\%$), and fallow land ($MAPE_{fallow} = 5.1\%$). The error was slightly higher than that observed when using DS only for perennial grasses ($MAPE_{grasses} = 7.9\%$). Based on these results, DF was selected for data recovery to construct a common dataset.

3.2. Machine Learning Parameters Optimization

Table 4 shows the parameters of the Linear SVM algorithm that were obtained during optimization by evaluating the 10-fold cross-validation accuracy.

Table 4. Optimal SVM parameters.

| Parameter | Value |
|----------------------------|---------------|
| Penalization norm | L2 |
| Loss function | Squared hinge |
| Multi-class strategy | ovr |
| Regularization parameter C | 100 |

For the first three parameters, the default values were optimal. The main parameter affecting the performance of the algorithm was the regularization coefficient C . The dependence of the cross-validation accuracy on C is shown in Figure S1. The standard value of the regularization coefficient, i.e., 1, was not optimal. Increasing this value to 100 improved the cross-validation accuracy by 1.5% (from 65.8% to 67.3%). Any further increase in the value of the parameter led to a decrease in accuracy.

Table 5 summarizes the parameters of the RF algorithm, fitted using GridSearch.

Table 5. Optimal RF parameters.

| Parameter | Value |
|------------------------------|-----------|
| Criterion | Gini |
| Minimum samples to split | 2 |
| Minimum samples in leaf node | 1 |
| Features to best split | 5 |
| Maximum leaf in nodes | Unlimited |
| Trees | 150 |

The default value of 5 for the number of features used to determine the best split was found to be optimal. Limiting the number of leaves in nodes (by default, this number is not limited) also led to a decrease in cross-validation accuracy. The dependence of the cross-validation accuracy on the number of trees is presented in Figure S2. Increasing the number of trees led to higher training time costs, due to which capping this value at 150 trees was found to be feasible. Subsequently, the maximum accuracy achieved with RF during the 10-fold cross-validation was 87.2%.

Table 6 summarizes the GB algorithm parameters fitted using GridSearch during 10-fold cross-validation.

Table 6. Optimal GB parameters.

| Parameter | Value |
|-------------------------|-------|
| Minimum samples in leaf | 20 |
| Maximum leaf in nodes | 31 |
| Maximum iterations | 300 |

For the minimum number of samples in a leaf and the maximum number of leaves in a node, the default values (20 and 31, respectively) were optimal. The maximum number of iterations was increased from 100 to 300. The dependence of the cross-validation accuracy on the number of iterations is shown in Figure S3. The difference in accuracy after 100 and 300 iterations was 0.3% (85.6% and 85.9%, respectively). However, the higher computational performance of the histogram-based GB (an order higher than RF) did not lead to a significant increase in the computational cost.

3.3. Crop Mapping Results

When performing 10-fold cross-validation for DF-restored time series, the overall classification accuracy was 67.3% for the SVM method, 87.2% for the RF method, and 85.9% for the GB method.

Figure 8 shows the confusion matrix for the SVM method. The OA was 67.3%, which was not a satisfactory result. Overall, the classifier only managed to detect soybean ($PA_{soybean} = 0.89$, $UA_{soybean} = 0.95$, $F1_{soybean} = 0.92$). In particular, the SVM classifier faced challenges when attempting to separate fallow and perennial grasses. Consequently, 17,613 of 51,315 perennial grasses time series were recognized as fallow ($PA_{grasses} = 0.30$, $UA_{fallow} = 0.55$), resulting in relatively low values of $F1_{fallow}$ and $F1_{grasses}$ (0.58 and 0.33, respectively). For oat, the value of $F1_{oat}$ was 0.77, with parts of the series (14%) being recognized as either perennial grasses or buckwheat. The SVM classifier did not perform well

enough in recognizing buckwheat ($F1_{buckwheat} = 0.66$), and a substantial proportion of series (17%) were classified as fallow, while 12% of fallow series were classified as buckwheat.

| | | | | | | | | |
|-----------------|-----------|--------|--------|--------|--------|--------|-------|-------|
| Predicted class | Soybean | 58,791 | 855 | 386 | 684 | 1145 | 95.0% | 5.0% |
| | Fallow | 1217 | 36,187 | 1081 | 17,613 | 9797 | 54.9% | 45.1% |
| | Oat | 1723 | 65 | 41,787 | 5829 | 4396 | 77.7% | 22.3% |
| | Grasses | 1313 | 15,680 | 3669 | 13,033 | 2589 | 35.9% | 64.1% |
| | Buckwheat | 3215 | 7147 | 7233 | 6716 | 40,401 | 62.4% | 37.6% |

| | | | | |
|---------|--------|-------|---------|-----------|
| 88.7% | 60.4% | 77.2% | 29.7% | 69.3% |
| 11.3% | 39.6% | 22.8% | 70.3% | 30.7% |
| Soybean | Fallow | Oat | Grasses | Buckwheat |

True class

Figure 8. SVM confusion matrix (DF fitted time series).

Figure 9 shows the RF confusion matrix. The OA for the RF method was 87.2%, which was 19.9% higher than that of the SVM classifier. The values of $F1$ for soybean and oat exceeded 0.9 ($F1_{soybean} = 0.96$, $F1_{oat} = 0.91$). The $F1$ values for buckwheat and fallow exceeded 0.85 ($F1_{buckwheat} = 0.87$, $F1_{fallow} = 0.85$). The most difficult task for the RF classifier was the recognition of the perennial grasses' series ($F1_{grasses} = 0.69$): 21% of grasses time series were recognized as fallow. It could be related to the lack of plowing during the season in several fields with perennial grasses, making the time series resemble that of abandoned cropland. Nevertheless, the classification results with the RF method could be considered successful.

The GB classifier showed high accuracy ($OA = 85.9\%$). The GB confusion matrix is presented in Figure 10. In general, the results were similar to the results obtained from the RF classifier ($F1_{soybean} = 0.98$, $F1_{oat} = 0.88$, $F1_{grasses} = 0.62$, $F1_{fallow} = 0.84$, $F1_{buckwheat} = 0.88$). Moreover, similar to the other classifiers, the issue of grasses recognition (perennial or natural in unused cropland) is relevant ($PA_{grasses} = 0.54$ and $UA_{grasses} = 0.73$).

Table 7 summarizes the performance estimates for all three classifiers used in the study. The cross-validation OA for the three classes is present in Table S1. In terms of overall accuracy, the tree-based classifiers (RF and GB) significantly outperformed SVM (Tukey test, $p < 0.01$). Although the RF and GB classifiers achieved a similar classification accuracy (Tukey test, $p > 0.05$), the GB algorithm could significantly reduce the time cost (Tukey test, $p < 0.01$). Therefore, the use of a classifier based on GB is optimal in addressing the cropland mapping problem.

| | | | | | | | | |
|-----------------|-----------|------------|--------|--------|---------|-----------|-------|-------|
| Predicted class | Soybean | 64,318 | 671 | 176 | 223 | 1847 | 95.7% | 4.3% |
| | Fallow | 875 | 54,181 | 131 | 9444 | 2207 | 81.1% | 18.9% |
| | Oat | 172 | 123 | 50,890 | 4060 | 2848 | 87.6% | 12.4% |
| | Grasses | 111 | 4473 | 1320 | 26,949 | 1280 | 79.0% | 21.0% |
| | Buckwheat | 783 | 486 | 1639 | 3259 | 50,146 | 89.0% | 11.0% |
| | | 97.1% | 90.4% | 94.0% | 61.3% | 86.0% | | |
| | | 2.9% | 9.6% | 6.0% | 38.7% | 14.0% | | |
| | | Soybean | Fallow | Oat | Grasses | Buckwheat | | |
| | | True class | | | | | | |

Figure 9. RF confusion matrix (DF fitted time series).

| | | | | | | | | |
|-----------------|-----------|------------|--------|--------|---------|-----------|-------|-------|
| Predicted class | Soybean | 64,909 | 537 | 287 | 377 | 681 | 97.2% | 2.8% |
| | Fallow | 742 | 54,314 | 185 | 11,968 | 2394 | 78.0% | 22.0% |
| | Oat | 150 | 387 | 48,905 | 4473 | 2981 | 86.0% | 14.0% |
| | Grasses | 176 | 4362 | 2912 | 23,691 | 1427 | 72.7% | 27.3% |
| | Buckwheat | 282 | 334 | 1867 | 3426 | 50,845 | 89.6% | 10.4% |
| | | 98.0% | 90.6% | 90.3% | 53.9% | 87.2% | | |
| | | 2.0% | 9.4% | 9.7% | 46.1% | 12.8% | | |
| | | Soybean | Fallow | Oat | Grasses | Buckwheat | | |
| | | True class | | | | | | |

Figure 10. GB confusion matrix (DF fitted time series).

Table 7. Accuracy metrics for the function-fitted dataset.

| Metric | SVM | RF | GB |
|------------------|--------|--------|-------|
| OA, % | 67.3 | 87.2 | 85.9 |
| $F1_{soybean}$ | 0.92 | 0.96 | 0.98 |
| $F1_{fallow}$ | 0.58 | 0.85 | 0.84 |
| $F1_{oat}$ | 0.77 | 0.91 | 0.88 |
| $F1_{grasses}$ | 0.33 | 0.69 | 0.62 |
| $F1_{buckwheat}$ | 0.66 | 0.87 | 0.88 |
| Mean time, s | 275.86 | 221.17 | 20.54 |

3.4. Crop Identification at the Parcel Level

Table 8 shows the accuracy of the crop identification in the sites located in Khabarovsk District. A total of 9 errors were observed across 130 test sites, with a total accuracy of 93% at the parcel level. Buckwheat was identified on one of the unused sites. Two buckwheat sites were identified incorrectly (one as perennial grasses and one as fallow—in both cases, the percentage of correctly identified sites was near 40%). One oat site was identified as perennial grasses, which can be related to the oat-perennial grasses in-season crop rotation, when grains and perennial grasses grow together but farmers specify only one crop. Five perennial grass sites were recognized incorrectly: two sites as oat (grasses presence) and three sites as fallow (invisible plowing).

Table 8. GB accuracy at the parcel level.

| Crop | Correctly Classified | Number of Fields | Accuracy, % |
|-------------------|----------------------|------------------|-------------|
| Soybean | 29 | 29 | 100 |
| Fallow | 26 | 27 | 96 |
| Oat | 20 | 21 | 95 |
| Perennial grasses | 26 | 31 | 84 |
| Buckwheat | 20 | 22 | 91 |
| Overall | 121 | 130 | 93 |

Figure 11 shows the results of crop mapping on individual sites. A correctly classified soybean site is presented in Figure 11a, with 2680 of 2713 pixels (98.8%) identified correctly. Figure 11b shows an example of unused land recognition. Fallow on the up-right site was recognized nearly perfectly, with only 9 out of 2239 pixels being recognized incorrectly. On the left-down site, 5% of time series were classified as soybean; however, the OA for this site was 93.1%. The site with oat in Figure 11c was identified correctly, with only 8 of 1102 series being classified incorrectly. The classification accuracy for both sites with perennial grasses in Figure 11d was higher than 97%; however, a limited number of the series were classified as fallow, which could be attributed to the presence of unused land at the site margins. Figure 11e shows an example of buckwheat cluster mapping. Four sites were identified with an accuracy higher than 96%. The case of misclassification is shown in Figure 11f (marked yellow). This error can be explained by the unsuccessful experience of growing buckwheat on individual sites—buckwheat did not grow, the site was overgrown with weeds, and on visual inspection, it was identified as unused. The resulting combined cropland map is shown in Figure S4.

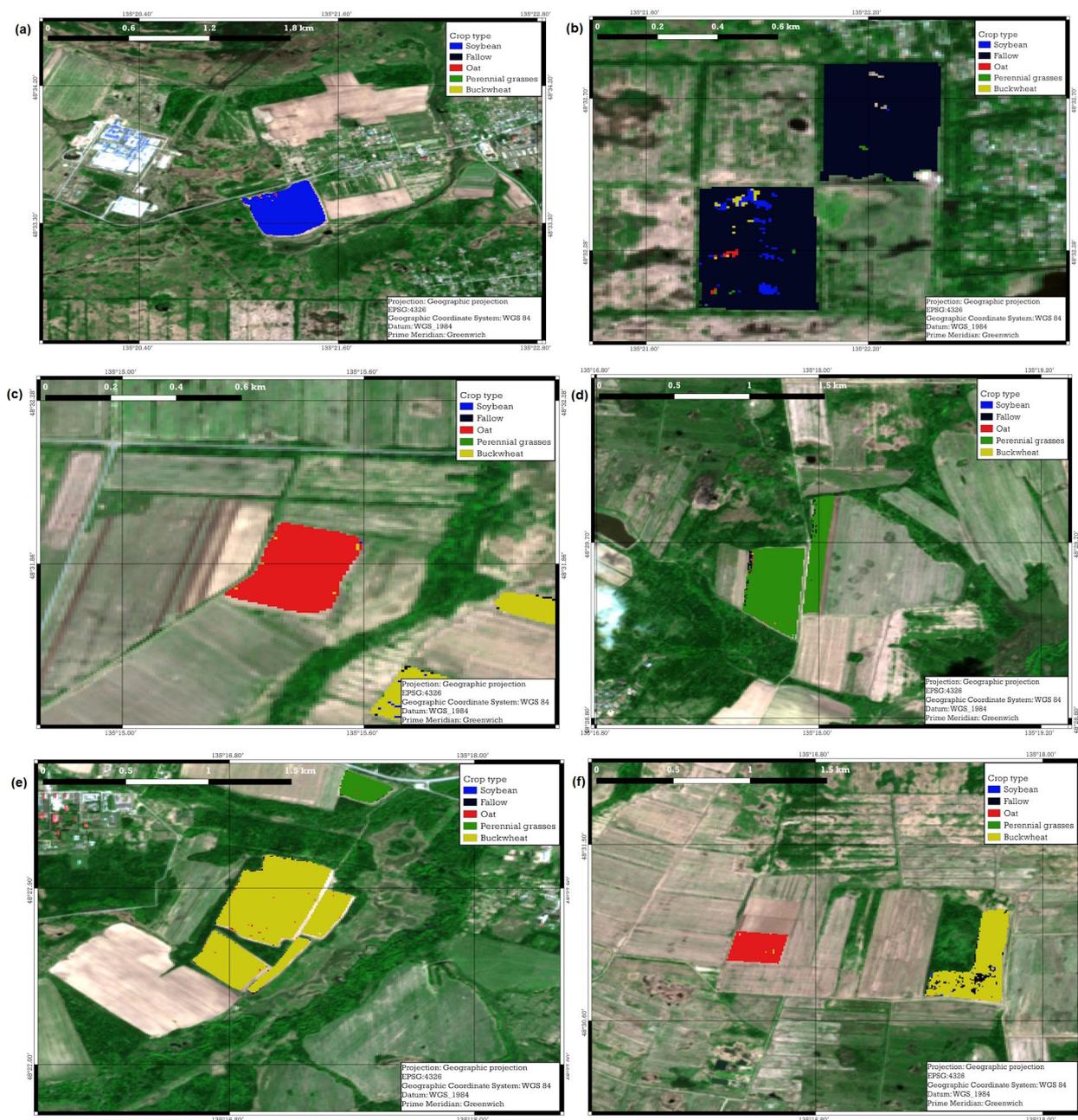


Figure 11. Crop mapping in test sites: (a) soybean site (marked blue), (b) fallow land recognition (marked black), (c) oat site (marked red), (d) grasses sites (marked green), (e) buckwheat sites (marked yellow), (f) misclassified site.

4. Discussion

4.1. NDVI Time Series Restoration

The use of compound trigonometric functions (sums of trigonometric functions) resulted in relatively high *MAPE* values when approximating *NDVI* series for soybean. It can be observed that the soybean *NDVI* plots showed a single peak and that employing simple functions with a single extremum may be better suited to fit such series (similar to the Gaussian function) [24,43]. However, single-peak functions were less effective for fitting the *NDVI* series for perennial grasses and buckwheat (when sown in July), which have two peaks [44]. In practice, unmarked data (where the crop was not known) were used when using the classifier. Therefore, a compromise solution had to be chosen when recovering the data.

The proposed data recovery methodology required the calculation of composites and obtaining the coefficients of functions for each series. Computationally, function fitting is much more complex than simple interpolation methods [17,18]. However, for a region with a large number of cloudy days during the vegetation season, the use of spatial and temporal interpolation is extremely difficult. Spatial interpolation requires the presence of *NDVI* values in the site, and the presence of cloud cover over the whole site hinders the use of this method. Temporal interpolation is most effective when there are only a few gaps in the time series; the lack of data for 3–4 weeks makes such interpolation impossible. It should be noted again that using this method is a relatively simple and efficient way to recover multi-year sparse time series for classification.

The obtained fitting accuracy is comparable to the results published in leading scientific journals. The average root mean square errors (RMSE) using DF for soybean, perennial grasses, buckwheat, oat, and fallow were 0.061, 0.053, 0.047, 0.043, and 0.031, respectively. For example, Vorobieva et al. fitted *NDVI* time series obtained from MODIS data for various sites across the Samara region in Russia [24]. When approximating fallow land (combined with perennial grasses) using the double logistic function, the RMSE was 0.038, while it was 0.043 for cereal crops (including oat). When the first four terms of the Fourier series were used, the RMSE was 0.043 for both fallow and cereal crops. The use of four terms of the Fourier series did not result in a substantial improvement in accuracy when applied with both single-peaked and double-peaked series, rendering it seemingly unnecessary. Berger et al. used a double logistic function to approximate 8-day MODIS *NDVI* soybean composites [23], resulting in an RMSE of 0.105, which is 0.044 more than that in our model. Sun et al. reconstructed the time series of *NDVI* composites [45] and reported an RMSE of 0.091 for soybean, which is 0.03 more than that in our model.

4.2. Crop Mapping

To validate the performance of these classification algorithms, it is important to compare the results of our study with the findings of prior research on *NDVI* time series classification for cropland mapping. For example, Erdanaev et al. used monthly Sentinel-2 *NDVI* composites and the SVM classifier for land cover and crop mapping, considering seven classes, including five crop types, water, and other [10]. They reported an overall accuracy of 87%, with SVM being more accurate than RF. Other vegetation indices and their combinations were also tested, but none of them achieved an accuracy exceeding 90%. Ouzemou et al. used pan-sharpened Landsat-8 *NDVI* data to classify cropland, including orchards, with the RF classifier achieving a maximum accuracy of 89% [9]. These studies used one-year sentinel data for crop mapping. Mapping cropland using multi-year data is substantially more challenging due to the inter-annual variation in vegetation indices and the need to calculate composites. Hu et al. used Sentinel-2 images for 2019 and 2020 to map four crops across croplands in North China (four crops) and achieved a classification accuracy of 92% [16]. However, the process involved a significant number of computations, as a range of bands and vegetation indices were required. Thus, the accuracy and computational performance of the GB-based classifier developed in this study can both be considered high.

Shapley values analysis [46,47] was performed to explain the contribution of each week to the GB prediction. Figure 12 shows the percentage contribution to the final prediction for each class (the five most valuable weeks were demonstrated in diagrams). All components of the time series are almost uniformly used for soybean series identification: among the most important parameters for prediction are *NDVI* values at the beginning of the soybean growing season, at its peak, and at the end of the growing season. For the identification of fallow, the most important are *NDVI* values in August (32–34 weeks), since only for fallow in this period is there a stability of values (for soybean, there is a peak, for oats and perennial grasses, the *NDVI* in this period is low, and for buckwheat in general, the *NDVI* growth is characteristic). For oat identification, the *NDVI* drop after its harvesting at the end of July–August (30, 32, 34 weeks) and the values on the approach to the vegetation

peak (24, 26 week) are of the greatest importance. For the recognition of perennial grasses, both the values at the beginning of the growing season (18, 22 weeks) and the values at the time of reaching the second peak (August–September) characteristic of this particular class are important. Due to the early sowing and harvesting of buckwheat in 2022, the most characteristic for buckwheat was the *NDVI* at week 34 (end of August)—its contribution was 20.8%. It was during this period that buckwheat in 2022 was characterized by a sharp drop in the *NDVI*. The analysis showed that although summer *NDVI* values have a greater impact in general, it is important to use the entire seasonal *NDVI* time series for multiclass classification.

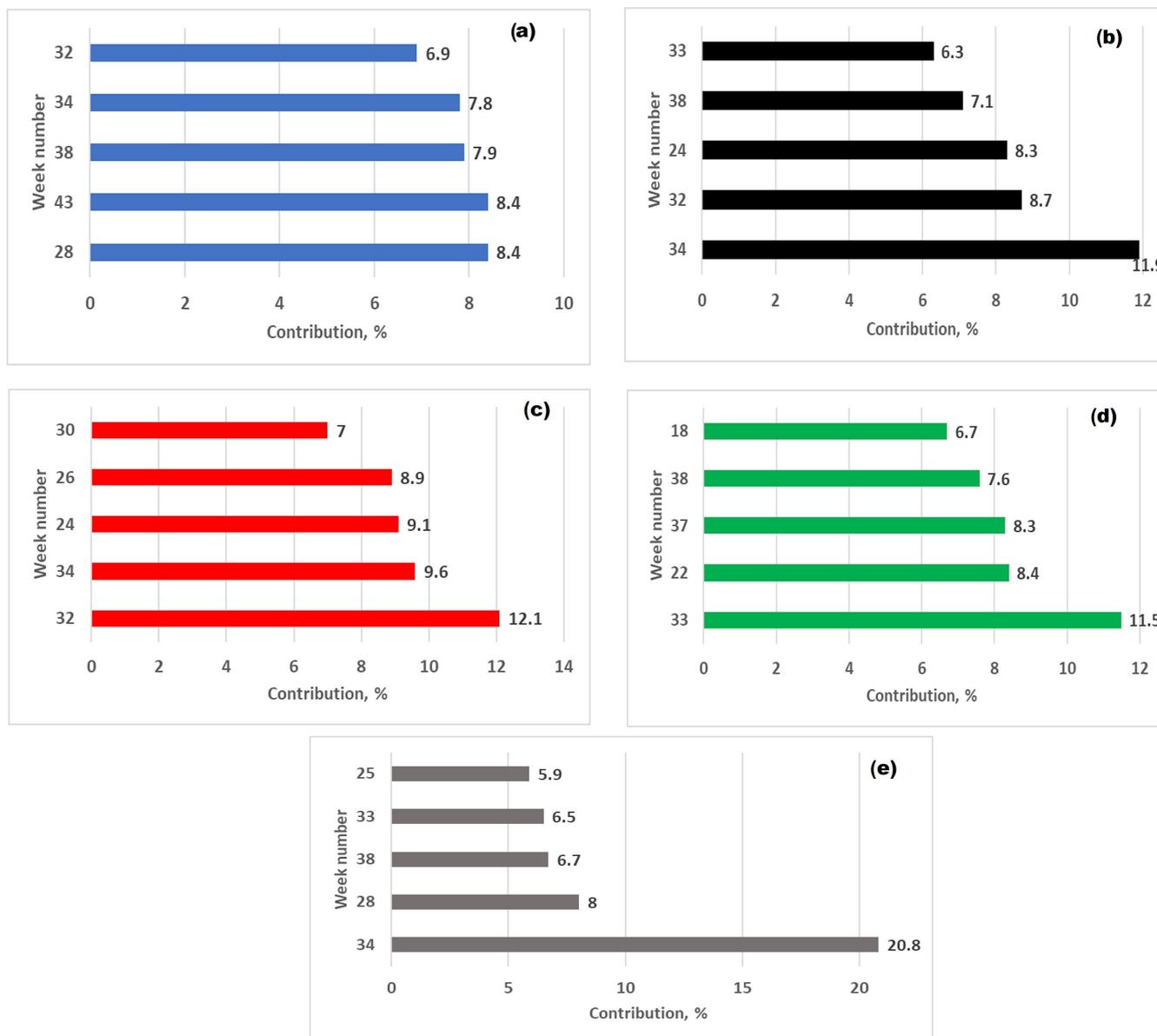


Figure 12. Contribution of *NDVI* values to GB prediction for: (a) soybean, (b) fallow, (c) oat, (d) grasses, and (e) buckwheat.

5. Conclusions

Time series of weekly *NDVI* composites for 2021 and 2022 were generated using multispectral Sentinel-2 images for crop sites in Khabarovsk District. Three types of nonlinear functions were fitted to recover observations for missing weeks. It was found that the Fourier series showed the highest accuracy in approximating the *NDVI* series, with a mean error of 8.2%.

Thus $MAPE_{soybean} = 11.8\%$, $MAPE_{buckwheat} = 8.6\%$, $MAPE_{oat} = 7.6\%$, $MAPE_{fallow} = 5.1\%$, and $MAPE_{grasses} = 7.9\%$. The use of a double sinusoidal function for fitting the *NDVI* series resulted in an increase in the error by up to 8.6%, while using a cubic polynomial resulted in an even higher increase in error by up to 16.8%.

Tenfold cross-validation was performed to optimize the hyperparameters of three classification methods (SVM with linear kernel, RF, and GB), with a histogram-based modification, and compare their accuracy. The overall accuracy when using the optimal hyperparameters was 67.3%, 87.2%, and 85.9% for SVM, RF, and GB, respectively. The *F1* values for different classes ranged from 0.33 to 0.92 for SVM, 0.69 to 0.96 for RF, and 0.62 to 0.98 for HGB. For the main crop of the Far East, soybean, the *F1* value when using GB was 0.98. An important factor supporting the use of this algorithm is that the time cost of using GB is an order of magnitude lower compared to the other methods. Thus, it has been shown that the use of fitting functions to reconstruct *NDVI* time series and subsequent GB machine learning make it possible to perform crop mapping based on multi-year data. The high accuracy and productivity of the proposed method determine the prospects for its use in regions with similar cropland structures and crop phenological cycles, including the developing agricultural regions of Northeast China that have large areas of cropland [3,48,49].

In future research, we plan to test the developed method in other regions using data for more years. It is also planned to use data from different satellites to eliminate sparsity in time series. The use of a combination of satellite bands and neural networks may be very promising.

The use of precision methods for cropland identification and control will make it possible to more effectively manage the agro-industrial complex. Accurate cropland mapping and vegetation analysis can also help assess the rational use of fertilizers, whose contribution to total nitrogen emissions is large enough that it can lead to significant changes in ecosystems [50]. Effective management in agriculture, coupled with well-managed urbanization, will allow for the faster adoption of modern agricultural methods, which, in turn, will reduce nitrogen emissions, potentially improving air and water quality [51]. Globally, climate change greatly impacts the production of major crops [52]. This impact is very uneven, especially for countries with large territories. The climate in Russia is warming about 2.5 times more intensely than the global average. The softening of winters in Russia could be an important factor in the future development of agriculture [53]. An increase in precipitation and climate warming leads to an increase in soybean yields. In general, the change in agro-climatic conditions in the Middle Amur River Region is relatively favorable for soybean (which is the major crop in this region) cultivation [54].

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs16091633/s1>, Figure S1: SVM accuracy with C values; Figure S2: RF accuracy with the number of trees; Figure S3: GB accuracy with the number of iterations; Figure S4: Crop map of research. Table S1: 10-fold cross-validation *OA*, % for different algorithms; Table S2: *MAPE* for different number of terms, crops, and seasons for Fourier fitting.

Author Contributions: Conceptualization, A.S. and T.A.; methodology, A.S.; software, K.D. and A.V.; validation, A.S. and T.A.; formal analysis, T.A.; investigation, K.D. and A.V.; resources, T.A.; data curation, T.A.; writing—original draft preparation, K.D.; writing—review and editing, A.S. and A.V.; visualization, A.V.; supervision, A.S.; project administration, T.A.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Russian Science Foundation, project number 23-76-00007.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Feng, F.; Gao, M.; Liu, R.; Yao, S.; Yang, G. A deep learning framework for crop mapping with reconstructed Sentinel-2 time series images. *Comput. Electron. Agric.* **2023**, *213*, 108227. [[CrossRef](#)]
- Pott, L.P.; Amado, T.J.C.; Schwalbert, R.A.; Corassa, G.M.; Ciampitti, I.A. Mapping crop rotation by satellite-based data fusion in Southern Brazil. *Comput. Electron. Agric.* **2023**, *211*, 107958. [[CrossRef](#)]
- Yan, S.; Yao, X.; Zhu, D.; Liu, D.; Zhang, L.; Yu, G.; Gao, B.; Yang, J.; Yun, W. Large-scale crop mapping from multi-source optical satellite imageries using machine learning with discrete grids. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102485. [[CrossRef](#)]
- You, L.; Sun, Z. Mapping global cropping system: Challenges, opportunities, and future perspectives. *Crop Environ.* **2022**, *1*, 68–73. [[CrossRef](#)]
- Wang, Z.; Zhang, H.; He, W.; Zhang, L. Cross-phenological-region crop mapping framework using Sentinel-2 time series Imagery: A new perspective for winter crops in China. *ISPRS J. Photogramm. Remote Sens.* **2022**, *193*, 200–215. [[CrossRef](#)]
- Xia, T.; He, Z.; Cai, Z.; Wang, C.; Wang, W.; Wang, J.; Hu, Q.; Song, Q. Exploring the potential of Chinese GF-6 images for crop mapping in regions with complex agricultural landscapes. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102702. [[CrossRef](#)]
- Meng, S.; Wang, X.; Hu, X.; Luo, C.; Zhong, Y. Deep learning-based crop mapping in the cloudy season using one-shot hyperspectral satellite imagery. *Comput. Electron. Agric.* **2021**, *186*, 106188. [[CrossRef](#)]
- Luo, K.; Lu, L.; Xie, Y.; Chen, F.; Yin, F.; Li, Q. Crop type mapping in the central part of the North China Plain using Sentinel-2 time series and machine learning. *Comput. Electron. Agric.* **2023**, *205*, 107577. [[CrossRef](#)]
- Ouzemou, J.-E.; El Harti, A.; Lhissou, R.; El Moujahid, A.; Bouch, N.; El Ouazzani, R.; El Bachaoui, M.; El Ghmari, A. Crop type mapping from pansharpened Landsat 8 NDVI data: A case of a highly fragmented and intensive agricultural system. *Remote Sens. Appl. Soc. Environ.* **2018**, *11*, 94–103. [[CrossRef](#)]
- Erdanaev, E.; Kappas, M.; Wyss, D. Irrigated Crop Types Mapping in Tashkent Province of Uzbekistan with Remote Sensing-Based Classification Methods. *Sensors* **2022**, *22*, 5683. [[CrossRef](#)]
- Hao, P.; Tang, H.; Chen, Z.; Yu, L.; Wu, M. High resolution crop intensity mapping using harmonized Landsat-8 and Sentinel-2 data. *J. Integr. Agric.* **2019**, *18*, 2883–2897. [[CrossRef](#)]
- Chen, Y.; Lu, D.; Moran, E.; Batistella, M.; Dutra, L.V.; da Silva, I.; Huang, J.; Luiz, A.; de Oliveira, M. Mapping croplands, cropping patterns, and crop types using MODIS time-series data. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *69*, 133–147. [[CrossRef](#)]
- Bellón, B.; Bégue, A.; Lo Seen, D.; Lebourgeois, V.; Evangelista, B.A.; Simões, M.; Ferraz, R.P.D. Improved regional-scale Brazilian cropping systems' mapping based on a semi-automatic object-based clustering approach. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *68*, 127–138. [[CrossRef](#)]
- Feyisa, G.L.; Palao, L.K.; Nelson, A.; Gumma, M.K.; Paliwal, A.; Win, K.T.; Nge, K.H.; Johnson, D.E. Characterizing and mapping cropping patterns in a complex agro-ecosystem: An iterative participatory mapping procedure using machine learning algorithms and MODIS vegetation indices. *Comput. Electron. Agric.* **2020**, *17*, 105595. [[CrossRef](#)]
- Konduri, V.S.; Kumar, J.; Hargrove, W.W.; Hoffman, F.M.; Ganguly, A.R. Mapping crops within the growing season across the United States. *Remote Sens. Environ.* **2020**, *251*, 112048. [[CrossRef](#)]
- Hu, Y.; Zeng, H.; Tian, F.; Zhang, M.; Wu, B.; Gilliams, S.; Li, S.; Li, Y.; Lu, Y.; Yang, H. An Interannual Transfer Learning Approach for Crop Classification in the Hetao Irrigation District, China. *Remote Sens.* **2022**, *14*, 1208. [[CrossRef](#)]
- Julien, Y.; Sobrino, J.A. Comparison of cloud-reconstruction methods for time series of composite NDVI data. *Remote Sens. Environ.* **2010**, *114*, 618–625. [[CrossRef](#)]
- Bradley, B.A.; Jacob, R.W.; Hermance, J.F.; Mustard, J.F. A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data. *Remote Sens. Environ.* **2007**, *106*, 137–145. [[CrossRef](#)]
- Tang, L.; Zhao, Z.; Tang, P.; Yang, H. SURE-based optimum-length S-G filter to reconstruct NDVI time series iteratively with outliers' removal. *Int. J. Wavel. Multiresolut. Inf. Process.* **2020**, *18*, 2050001. [[CrossRef](#)]
- Chen, Y.; Cao, R.; Chen, J.; Liu, L.; Matsushita, B. A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 174–190. [[CrossRef](#)]
- Jonsson, P.; Eklundh, L. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1824–1832. [[CrossRef](#)]
- Yang, Y.P.; Luo, J.C.; Huang, Q.T.; Wu, W.; Sun, Y.W. Weighted double-logistic function fitting method for reconstructing the high-quality Sentinel-2 NDVI time series data set. *Remote Sens.* **2019**, *11*, 2342. [[CrossRef](#)]
- Berger, A.; Ettl, G.; Quincke, C.; Rodríguez-Bocca, P. Predicting the Normalized Difference Vegetation Index (NDVI) by training a crop growth model with historical data. *Comput. Electron. Agric.* **2019**, *161*, 305–311. [[CrossRef](#)]
- Vorobiova, N.; Chernov, A. Curve fitting of MODIS NDVI time series in the task of early crops identification by satellite images. *Procedia Eng.* **2017**, *201*, 184–195. [[CrossRef](#)]
- Tran, K.H.; Zhang, H.K.; McMaine, J.T.; Zhang, X.; Luo, D. 10 m crop type mapping using Sentinel-2 reflectance and 30 m cropland data layer product. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102692. [[CrossRef](#)]
- Blickensdorfer, L.; Schwider, M.; Pflugmacher, D.; Nendel, C.; Erasmi, S.; Hostert, P. Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. *Remote Sens. Environ.* **2022**, *269*, 112831. [[CrossRef](#)]
- Yaramasu, R.; Bandaru, V.; Pnvr, K. Pre-season crop type mapping using deep neural networks. *Comput. Electron. Agric.* **2020**, *176*, 105664. [[CrossRef](#)]

28. Aseeva, T.A. Methods of expanding reproduction of heavy loamy soils fertility of the Priamurye region. *Soil Sci. Agrochem.* **2015**, *2*, 107–116. (In Russian)
29. Gajduchenko, A.N.; Oborskij, S.L.; Toporova, L.I. Scientifically proved crop rotation and optimization of technological ways of cultivation—A basis of increasing of efficiency of soya. *Far East. Agrar. Her.* **2009**, *2*, 30–33. (In Russian)
30. Gavin, H.P. The Levenberg–Marquardt Method for Nonlinear Least Squares Curve-Fitting Problems. Available online: <http://people.duke.edu/~hpgavin/ce281/lm.pdf> (accessed on 23 March 2023).
31. Michishita, R.; Chen, J.; Xu, B. Empirical comparison of noise reduction techniques for NDVI time-series based on a new measure. *ISPRS J. Photogramm. Remote Sens.* **2014**, *91*, 17–28. [[CrossRef](#)]
32. Zhou, X.; Wang, J.; He, Y.; Shan, B. Crop Classification and Representative Crop Rotation Identifying Using Statistical Features of Time-Series Sentinel-1 GRD Data. *Remote Sens.* **2022**, *14*, 5116. [[CrossRef](#)]
33. Atkinson, P.M.; Jeganathan, C.; Dash, J.; Atzberger, C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sens. Environ.* **2012**, *123*, 400–417. [[CrossRef](#)]
34. Son, N.T.; Chen, C.F.; Chen, C.R.; Duc, H.N.; Chang, L.Y. A Phenology-Based Classification of Time-Series MODIS Data for Rice Crop Monitoring in Mekong Delta, Vietnam. *Remote Sens.* **2014**, *6*, 135–156. [[CrossRef](#)]
35. Asgarian, A.; Soffianian, A.; Pourmanafi, S. Crop type mapping in a highly fragmented and heterogeneous agricultural landscape: A case of central Iran using multi-temporal Landsat 8 imagery. *Comput. Electron. Agric.* **2016**, *127*, 531–540. [[CrossRef](#)]
36. Zhang, H.; Kang, J.; Xu, X.; Zhang, L. Accessing the temporal and spectral features in crop type mapping using multi-temporal Sentinel-2 imagery: A case study of Yi'an County, Heilongjiang province, China. *Comput. Electron. Agric.* **2020**, *176*, 105618. [[CrossRef](#)]
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
38. Van Tricht, K.; Gobin, A.; Gilliams, S.; Piccard, I. Synergistic Use of Radar Sentinel-1 and Optical Sentinel-2 Imagery for Crop Mapping: A Case Study for Belgium. *Remote Sens.* **2018**, *10*, 1642. [[CrossRef](#)]
39. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, USA, 13–17 August 2016.
40. Chirici, G.; Scotti, R.; Montaghi, A.; Barbati, A.; Cartisano, R.; Lopez, G.; Marchetti, M.; McRoberts, R.E.; Olsson, H.; Corona, P. Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *25*, 87–97. [[CrossRef](#)]
41. Candido, C.; Blanco, A.C.; Medina, J.; Gubatanga, E.; Santos, A.; Sta Ana, R.; Reyes, R.B. Improving the consistency of multi-temporal land cover mapping of Laguna Lake watershed using light gradient boosting machine (LightGBM) approach, change detection analysis, and Markov chain. *Remote Sens. Appl. Soc. Environ.* **2021**, *23*, 100565. [[CrossRef](#)]
42. Lu, B.; Meng, X.; Dong, S.; Zhang, Z.; Liu, C.; Jiang, J.; Herrmann, H.; Li, X. High-resolution mapping of regional VOCs using the enhanced space-time extreme gradient boosting machine (XGBoost) in Shanghai. *Sci. Total Environ.* **2023**, *905*, 167054. [[CrossRef](#)] [[PubMed](#)]
43. Hird, J.N.; McDermid, G.J. Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sens. Environ.* **2009**, *113*, 248–258. [[CrossRef](#)]
44. Stepanov, A.; Dubrovin, K.; Sorokin, A. Function fitting for modeling seasonal normalized difference vegetation index time series and early forecasting of soybean yield. *Crop J.* **2022**, *10*, 1452–1459. [[CrossRef](#)]
45. Sun, L.; Gao, F.; Xie, D.; Anderson, M.; Chen, R.; Yang, Y.; Yang, Y.; Chen, Z. Reconstructing daily 30 m NDVI over complex agricultural landscapes using a crop reference curve approach. *Remote Sens. Environ.* **2021**, *253*, 112156. [[CrossRef](#)]
46. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.-W.; Newman, S.-F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [[CrossRef](#)] [[PubMed](#)]
47. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [[CrossRef](#)] [[PubMed](#)]
48. Xuan, F.; Dong, Y.; Li, J.; Li, X.; Su, W.; Huang, X.; Huang, J.; Xie, Z.; Li, Z.; Liu, H.; et al. Mapping crop type in Northeast China during 2013–2021 using automatic sampling and tile-based image classification. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103178. [[CrossRef](#)]
49. Zhen, Z.; Chen, S.; Yin, T.; Gastellu-Etchegorry, J.-P. Improving Crop Mapping by Using Bidirectional Reflectance Distribution Function (BRDF) Signatures with Google Earth Engine. *Remote Sens.* **2023**, *15*, 2761. [[CrossRef](#)]
50. Liu, L.; Xu, W.; Wen, Z.; Liu, P.; Xu, H.; Liu, S.; Lu, X.; Zhong, B.; Guo, Y.; Lu, X.; et al. Modeling global oceanic nitrogen deposition from food systems and its mitigation potential by reducing overuse of fertilizers. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2221459120. [[CrossRef](#)] [[PubMed](#)]
51. Deng, O.; Wang, S.; Ran, J.; Huang, S.; Zhang, X.; Duan, J.; Zhang, L.; Xia, Y.; Reis, S.; Xu, J.; et al. Managing urban development could halve nitrogen pollution in China. *Nat. Commun.* **2024**, *15*, 401. [[CrossRef](#)] [[PubMed](#)]
52. Mohammadi, S.; Rydgren, K.; Bakkestuen, V.; Gillespie, M.A.K. Impacts of recent climate change on crop yield can depend on local conditions in climatically diverse regions of Norway. *Sci. Rep.* **2023**, *13*, 3633. [[CrossRef](#)] [[PubMed](#)]

-
53. Gordeev, R.V.; Pyzhev, A.I.; Zander, E.V. Does Climate Change Influence Russian Agriculture? Evidence from Panel Data Analysis. *Sustainability* **2022**, *14*, 718. [[CrossRef](#)]
 54. Fedorova, T.N.; Aseeva, T.A. Changes in regional climatic characteristics of the Middle Amur Region and their impact on soybean yield. *Vestn. FEB RAS* **2022**, *3*, 138–148. (In Russian) [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.