Detection of Cropland Abandonment through Multi-Temporal Landsat Data and Spatially Independent Machine Learning Validation

Luka Stojanović[1], Anka Lisec[2], Krištof Oštir[3], Bujar Fetai[4]

[1] Luka Stojanović, Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, lstojano@fgg.uni-lj.si
[2] Anka Lisec, Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, alisec@fgg.uni-lj.si
[3] Krištof Oštir, Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, kostir@fgg.uni-lj.si
[4] Bujar Fetai, Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, bfetai@fgg.uni-lj.si

Keywords: Cropland abandonment, Random Forest, XGBoost, Remote sensing, Landsat, Land use.

Abstract

Cropland abandonment (CA) is a major land-use change with important environmental and socio-economic implications. This study evaluates cropland abandonment using multi-temporal Landsat features and a spatially independent validation framework, comparing the performance and spatial behaviour of Random Forest and XGBoost classifiers. A set of temporally aggregated spectral indices (NDVI, BSI, NDBI, and MNDWI), including multi-year trends and variability measures, was integrated into a 56-band composite dataset. Training and validation samples were generated using 100 × 100-pixel windows centred on land-use parcels, with overlapping areas between different reference classes explicitly excluded to avoid label ambiguity. To reduce spatial autocorrelation, the data were split into separated training (1,582.6 km²) and testing (719.2 km²) areas within the Savinjska statistical region in Slovenia. Random Forest (RF) and XGBoost (XGB) classifiers were trained and evaluated using spatially separated validation data. Classification performance was assessed using standard accuracy metrics. Results indicate that XGB achieved a higher overall accuracy (0.705) compared to RF (0.670) and exhibited strong sensitivity in detecting cropland abandonment, while RF produced more conservative and spatially stable estimates of abandoned cropland area. Spatial error maps and area-based comparisons reveal systematic differences between the two classifiers, particularly in their tendency to overestimate abandonment extent. The findings highlight the importance of spatially explicit validation strategies, careful reference data preparation, and multi-temporal feature design for robust cropland abandonment mapping. The main contribution lies in the systematic assessment of model behaviour, spatial error patterns, and area estimates under strict spatial separation of training and testing data.

1. Introduction

Cropland abandonment (CA) has emerged as a widespread land-use change driven by socio-economic transformation, demographic shifts, agricultural intensification, and market restructuring. In many regions, CA has significant implications for food security, biodiversity, carbon cycling, and landscape management, making its timely and spatially explicit detection increasingly important for land-use planning and policy support (Lambin and Meyfroidt, 2011; Prishchepov et al., 2012).

Remote sensing has become a primary tool for CA mapping due to its ability to provide consistent, spatially continuous, and long-term observations of land surface dynamics (Wulder et al., 2019). Medium-resolution satellite missions, particularly the Landsat program, offer a unique temporal archive spanning several decades, enabling the analysis of vegetation dynamics associated with agricultural activity and abandonment processes (Alcántara et al., 2013; Kuemmerle et al., 2013). However, CA detection remains challenging because abandoned fields do not exhibit a universal spectral signature and may undergo heterogeneous post-abandonment trajectories, including grassland succession, shrub encroachment, or re-cultivation (Lambin and Geist, 2006; Estel et al., 2015).

To address these challenges, many studies have adopted multi-temporal approaches that exploit temporal trajectories of vegetation indices rather than relying on single-date imagery. Spectral indices such as the Normalised Difference Vegetation Index (NDVI) and Bare Soil Index (BSI), as well as their interannual trends and variability, have proven effective in distinguishing long-term CA from short-term fallow or crop rotation patterns. These approaches benefit from integrating multiple years of observations, thereby capturing persistent changes in land management rather than transient phenological effects (Prishchepov et al., 2012; Alcántara et al., 2013).

Recent advances in machine learning (ML) have further improved CA mapping by enabling the classification of high-dimensional, multi-temporal feature spaces. Non-parametric classifiers such as Random Forest (RF) and gradient-boosting methods, including XGBoost (XGB), are particularly well suited for remote sensing applications due to their robustness to multicollinearity, non-linear relationships, and noisy input data (Belgiu and Drăguţ, 2016; Maxwell et al., 2018). Several studies have demonstrated that ensemble-based ML models outperform traditional parametric classifiers in land-use and land-cover mapping, including CA detection (Estel et al., 2015; Stefanski et al., 2014).

Despite these methodological advances, a persistent limitation in CA mapping studies is the validation strategy used to assess classification performance. Many studies rely on random pixel-based or object-based train–test splits, which can lead to spatial dependence between training and validation samples. Such spatial autocorrelation often results in overly optimistic accuracy estimates and limited model generalisation to unseen areas. Addressing this issue is particularly critical in spatial classification problems, where neighbouring samples may share similar spectral and environmental characteristics (Roberts et al., 2017; Meyer et al., 2019).

Classification performance in cropland abandonment mapping is commonly evaluated using accuracy metrics that are robust to class imbalance and suitable for binary classification problems. Overall accuracy, user's and producer's accuracy, and the F1-score provide complementary information on model performance and error characteristics and are increasingly recommended for

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-679-2026

land-use classification tasks involving uneven class distributions (Pontius and Millones, 2011).

In this context, this study aims to detect CA using multi-temporal Landsat imagery and comparison of ML models, while explicitly addressing spatial validation and metric selection issues. A set of temporally aggregated spectral indices, including NDVI (Normalised Difference Vegetation Index), BSI (Bare Soil Index), NDBI (Normalised Difference Built-up Index), and MNDWI (Modified Normalised Difference Water Index), as well as multi-year trends and variability measures, is used to characterise long-term agricultural land-use dynamics. RF and XGB classifiers are evaluated using training and testing areas within the Savinjska statistical region.

The sub-objectives of this study are to:
(i) develop a multi-temporal feature representation suitable for CA detection using medium-resolution satellite data;
(ii) compare the performance of RF and XGB models; and
(iii) assess differences in spatial error patterns and abandoned cropland area estimates between the two models.

2. Study Area and Data

2.1 Study Area

The study was conducted in the Savinjska statistical region, located in north-eastern Slovenia. The region is characterised by a heterogeneous agricultural landscape, including intensively managed croplands, grasslands, and areas affected by long-term land-use change. Socio-economic restructuring and changes in agricultural practices over recent decades have contributed to the occurrence of CA, making the region suitable for evaluating CA detection using remote sensing data (Lambin and Meyfroidt, 2011; Lambin and Geist, 2006).

To explicitly address spatial autocorrelation and avoid spatial dependence between training and testing samples, the study area was divided into two spatially independent zones (Roberts et al., 2017; Meyer et al., 2019). The separation was defined using a railway corridor that crosses the region and represents a clear, continuous, and geographically meaningful boundary (Figure 1). This infrastructure element was selected because it provides a stable and easily reproducible division that minimises spatial overlap and neighbourhood effects between the two zones.

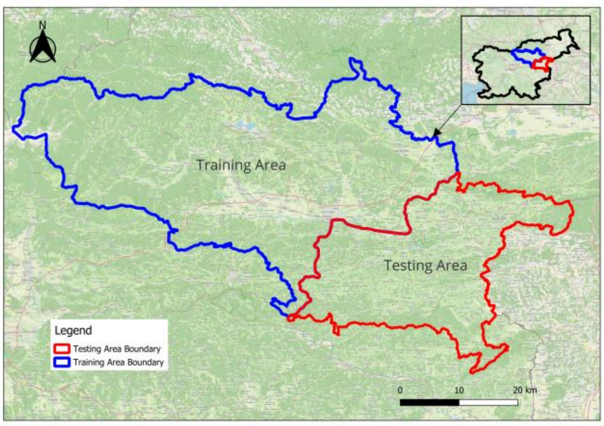

Figure 1: Location of the Savinjska statistical region and spatial division into training and testing areas divided by the railway corridor.

The area north-west of the railway line was used exclusively for model training, while the area south-east of the railway line was reserved for independent testing. This spatial split resulted in a training area of 1,582.6 km² and a testing area of 719.2 km². By enforcing a strict geographic separation between training and testing data, the adopted validation strategy reduces the influence of spatial autocorrelation and provides a more conservative and realistic assessment of model generalization performance (Roberts et al., 2017; Meyer et al., 2019).

2.2 Reference Data and Sampling Strategy

Reference land-use data were obtained from the official Slovenian land-use database RABA, provided through the national spatial data portal. Land-use data were interpreted based on land-use transitions observed between 2005 and 2025. Parcels that remained under continuous agricultural use during this period were classified as active cropland, whereas parcels exhibiting a cessation of agricultural activity were classified as cropland abandonment. The class definitions and corresponding land-use transitions used in this study are summarised in Table 1, which provides a transparent overview of the reference data interpretation applied for training and testing.

RABA ID 2005	Land use type 2005	RABA ID 2025	Land use type 2025	Abandoned/Active Class
1100	Arable Land	1160	Hop field	Active
1130	Temporary grassland	1300	Permanent grassland	Active
1160	Hop field	1160	Hop field	Active
1180	Permanent crops on arable land	1100	Arable Land	Active
1190	Greenhouse	1190	Greenhouse	Active
1211	Vineyard	1222	Extensive Orchard	Active
1221	Intensive Orchard	1221	Intensive Orchard	Active
1222	Extensive Orchard	1500	Trees and shrubs	Abandoned
1240	Other pemrnent crop	1300	Permanent grassland	Abandoned
1300	Permanent grassland	1800	Forest trees on agricultural land	Abandoned
1330	Mountain pasture	1300	Permanent grassland	Active
1410	Overgrown agricultural area	2000	Forest	Abandoned
1500	Trees and shrubs	1500	Trees and shrubs	Abandoned
1600	Uncultivated agricultural land	1300	Permanent grassland	Abandoned
1800	Forest trees on agricultural land	2000	Forest	Abandoned

Table 1: Definition of cropland classes based on land-use transitions (2005–2025)

All other land-use categories were excluded from the analysis to focus explicitly on the binary CA classification problem (Belgiu and Drăguț, 2016).

Instead of using full land-use polygons directly for training and validation, a centroid-based sampling strategy was adopted. For each reference polygon, a centroid point was extracted and used to generate a 100 × 100-pixel rectangular sampling window. Before sample extraction, areas where reference polygons of different classes spatially overlapped were explicitly excluded to avoid label ambiguity and reduce uncertainty in the training data.

This approach was selected to (i) reduce the influence of polygon boundary uncertainty, (ii) minimise mixed-pixel effects near parcel edges, and (iii) ensure consistent spatial units for feature extraction across all samples. The window size was chosen as a compromise between capturing within-field spectral variability and avoiding contamination from adjacent land-use types.

Each sample was overlaid on the multi-temporal 56-band image dataset, and only pixels fully contained within the window were used for model training and testing. Model training was performed at the pixel level, with all pixels inside each window treated as independent samples, while the sampling windows themselves served to guide spatially consistent reference data extraction.
Sampling was performed independently for the spatially separated training and testing areas defined in Section 2.1. In the training area, a total of 11,802 samples were generated for active

cropland and 26,367 samples for CA. These samples corresponded to 82,243 pixels for active cropland and 210,943 pixels for CA, indicating a moderate class imbalance. In the testing area, 8,425 active cropland samples and 14,864 CA samples were used for independent validation, corresponding to 60,520 and 117,489 pixels, respectively, shown in Table 2.

By combining centroid-based pixel sampling, explicit removal of inter-class overlaps, and strict spatial separation between training and testing zones, the adopted strategy reduces label ambiguity and spatial dependence between samples and supports a more realistic evaluation of model performance in cropland abandonment detection (Roberts et al., 2017; Meyer et al., 2019).

Area type	Class	Samples count	Pixel count
Training	Active	11802	82243
Training	Abandoned	26367	210943
Testing	Active	8425	60520
Testing	Abandoned	14864	117489

Table 2: Summary of reference data and sampling statistics for training and testing datasets

3. Methodology

3.1 Overall Workflow

The methodological workflow adopted in this study is illustrated in Figure 2 and was designed to enable robust detection of CA using multi-temporal remote sensing data while explicitly addressing spatial dependence in model validation. The workflow integrates satellite data pre-processing, reference data preparation, multi-temporal feature extraction, supervised classification, and accuracy assessment.

Multi-temporal Landsat Collection 2 Level-2 imagery (2000–2025) served as the primary data source (Wulder et al., 2019). Image pre-processing and feature generation were conducted within the QGIS environment. The workflow follows the spatially independent training and testing design described in Section 2.1.
Reference land-use data were processed separately for the training and testing zones. Multi-temporal spectral indices and temporal descriptors were then derived from the Landsat archive and combined into a single multi-band feature dataset.

Supervised classification was performed using Random Forest (RF) and XGBoost (XGB) models trained on the spatially defined training dataset (Belgiu and Drăguț, 2016; Chen and Guestrin, 2016). Model performance was evaluated on the independent testing dataset using standard accuracy metrics and spatial error analysis. The workflow provides a transparent and reproducible framework for CA mapping using medium-resolution satellite data.

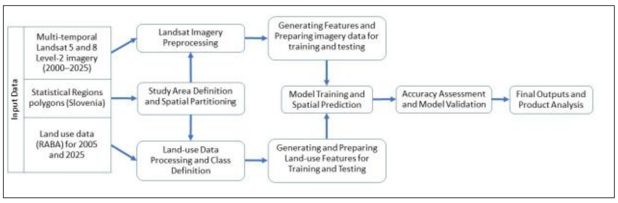

Figure 2: Methodological workflow for cropland abandonment detection using multi-temporal Landsat data and machine learning.

3.2 Remote Sensing and Feature Extraction

Multi-temporal Landsat imagery was used to derive a feature set suitable for cropland abandonment (CA) detection. Landsat Collection 2 Level-2 surface reflectance products were selected to ensure radiometric consistency across sensors and years and to support long-term change analysis (Wulder et al., 2019). The processing workflow was designed to capture persistent temporal patterns related to CA rather than single-date spectral states (Prishchepov et al., 2012; Alcántara et al., 2013).

For each year in the study period (2000–2025), Landsat scenes covering the study area were prepared by applying quality masking and spatial setting. Pixels affected by clouds and cloud shadows were excluded using the QA information provided with the Landsat Level-2 products. All raster images were clipped to the Savinjska statistical region extent to ensure consistent spatial coverage. The spatial split into independent training and testing zones was applied prior to sampling and modelling to prevent information leakage between zones (Roberts et al., 2017; Meyer et al., 2019). These pre-processing steps were carried out in QGIS 3.40.7 with the Semi-automatic Classification Plugin (SCP) 8.5.0 and raster processing tools.

The temporal aggregation into six five-year epochs was adopted as a compromise between temporal detail and model robustness. Using annual observations across the full 25-year period would increase feature dimensionality and sensitivity to noise caused by cloud contamination, sensor differences, and interannual phenological variability. Aggregating observations into multi-year periods allows the classifiers to focus on long-term land-use trends associated with cropland abandonment, while reducing redundancy and the risk of overfitting (Prishchepov et al., 2012; Alcántara et al., 2013).

From the pre-processed annual surface reflectance data, four spectral indices were calculated for each year: NDVI, BSI, NDBI, and MNDWI. These indices were selected because they capture complementary land surface characteristics relevant to CA detection, including vegetation activity, bare soil exposure, built-up sensitivity, and surface moisture conditions (Prishchepov et al., 2012; Alcántara et al., 2013). The indices were computed using standard formulations:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

$$MNDWI = \frac{GREEN - SWIR}{GREEN + SWIR}$$

$$BSI = \frac{(SWIR + RED) - (NIR + BLUE)}{(SWIR + RED) + (NIR + BLUE)}$$

Annual index layers represent year-specific land surface conditions and serve as the basis for subsequent temporal aggregation.
To characterise long-term land-use dynamics, temporal descriptors were derived from the annual NDVI and BSI time series. Two complementary temporal metrics were computed: (i) a linear trend (slope) to represent gradual directional change over time, and (ii) standard deviation to quantify interannual variability. Trend-based metrics have been widely used to distinguish persistent cropland abandonment from short-term

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-679-2026 |

fallow or crop rotation patterns (Prishchepov et al., 2012; Alcántara et al., 2013).

Beyond simple differences, linear temporal trends (slope) were calculated for NDVI and BSI across all six epochs (2000–2025). Time was coded as 0, 5, 10, 15, 20, and 25 to represent five-year intervals. Slopes were computed using the least squares regression formula:

$$slope = \frac{N\sum(t * y) - \sum t \sum y}{N\sum t^2 - (\sum t)^2}$$

where y is NDVI or BSI, t is the year index, and $N = 6$. A negative NDVI slope typically indicates declining vegetation activity (consistent with early abandonment or land degradation), whereas a positive slope suggests regrowth and secondary succession. Temporal trajectory analysis of vegetation indices is a well-established method for detecting gradual abandonment processes and is widely applied in LandTrendr-style approaches which rely on linear segments to approximate long-term NDVI dynamics (Prishchepov et al., 2012; Alcántara et al., 2013).

Interannual variability was quantified using the standard deviation σ:

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_1 - \overline{x})^2}$$

Together, these temporal descriptors support the separation of stable CA trajectories from actively managed cropland exhibiting strong interannual variability (Prishchepov et al., 2012; Alcántara et al., 2013).

All annual spectral index layers and temporal descriptors were merged into a single multi-band raster feature stack containing 56 bands, which was used consistently for model training and testing. The integration of multi-year indices and temporal metrics into a single feature space allows machine learning classifiers to exploit both spectral and temporal information when discriminating CA from active cropland (Belgiu and Drăguţ, 2016). Feature computation and raster stacking were performed using QGIS raster calculator and layer management tools, with additional temporal processing supported by Python-based workflows.

3.3 Reference Data Generation and Sample Design

Reference data for cropland abandonment (CA) mapping were derived from polygon-based land-use (RABA) datasets representing two classes: active cropland and cropland abandonment. All other land-use categories were excluded to focus explicitly on the binary CA classification problem. The preparation of reference data and sampling strategy was designed to ensure spatial consistency, reduce boundary-related uncertainties, and support spatially independent model validation (Roberts et al., 2017; Meyer et al., 2019).

Polygons located entirely within the training zone were used exclusively for model training, while polygons located within the testing zone were reserved for independent validation. This spatial separation was applied before any sampling or feature extraction to minimize spatial autocorrelation between training and validation samples and to avoid information leakage (Roberts et al., 2017; Meyer et al., 2019).

This approach was selected to reduce the influence of polygon boundary inaccuracies, which are common in agricultural land-use datasets, and to minimise mixed-pixel effects near parcel edges. By focusing on the interior of each polygon, the sampling strategy provides a more reliable representation of class-specific spectral and temporal characteristics. The use of fixed-size rectangular windows also ensures uniform sampling units across the study area and allows the extraction of multiple pixels per reference pixels, thereby capturing within-field variability while maintaining consistency between samples (Belgiu and Drăguţ, 2016).

For each 100 × 100 pixel rectangular sampling window, pixels located within the window were extracted and assigned the class label of the corresponding reference polygon. Only pixels within the rectangular sampling windows were used for training and testing, thereby limiting potential contamination from neighbouring land-use classes. This resulted in a set of spatially explicit, class-labelled pixel samples for both training and testing datasets. Sampling was performed independently for the training and testing zones, ensuring that no pixels from the testing area were used during model training or hyperparameter selection.

Reference data preparation, centroid extraction, rectangular window generation, and pixel selection were implemented in QGIS using vector geometry tools and raster extraction workflows. The resulting training and testing samples were subsequently linked with the multi-temporal feature stack described in the previous subsection and used as input for supervised classification.

Figure 3 provides a spatial overview of the centroid-based rectangular sampling units generated within the training area. The map illustrates the relationship between reference land-use polygons, extracted centroids, and the resulting 100 × 100 m rectangular windows used for sample generation. For visualisation clarity, only a representative subset of sampling units is shown.

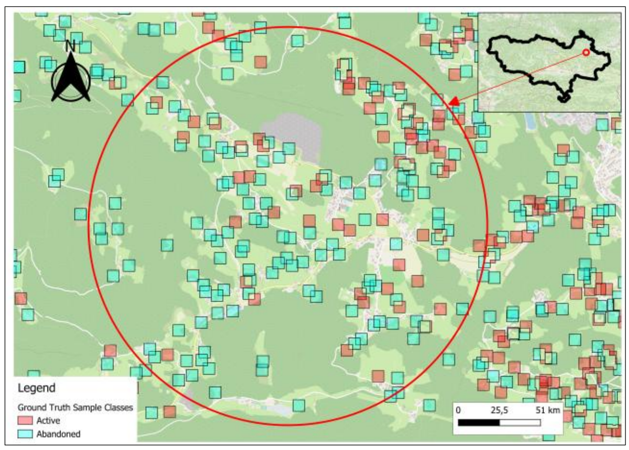

Figure 3: Overview of rectangular sampling units within the training area

This reference data preparation strategy ensures spatial independence between training and validation samples, reduces noise associated with polygon boundaries, and supports modest evaluation of CA detection performance using machine learning models.
Although reference information was derived from object-based land-use polygons, classification was performed at the pixel level. The rectangular windows were used exclusively as spatial

sampling units to assign class labels to pixels contained within each window. During model training, each pixel within a window was treated as an independent sample and associated with the class label of the corresponding reference polygon. Consequently, the adopted approach represents a pixel-based classification strategy informed by object-based sampling, which combines the advantages of polygon-level reference data with pixel-level spatial detail (Belgiu and Drăguţ, 2016).

3.4 Hyperparameter Selection

Hyperparameter values were selected based on a combination of commonly recommended settings in the literature and iterative empirical testing on the training dataset. Multiple parameter combinations were evaluated, and the final configuration was chosen to achieve stable performance while limiting overfitting under spatially independent validation. No reference data from the testing area were used during hyperparameter selection, ensuring an unbiased assessment of model generalisation (Belgiu and Drăguţ, 2016; Chen and Guestrin, 2016).

For Random Forest (RF), hyperparameters controlling the number of trees, tree depth, minimum sample sizes, and feature selection strategy were configured to ensure robust ensemble behaviour while accounting for moderate class imbalance. Such parameter choices have been shown to improve robustness and generalisation in land-cover and land-use classification tasks using Random Forest model (Belgiu and Drăguţ, 2016). For XGBoost (XGB), parameters related to tree depth, learning rate, subsampling, column sampling, and regularisation were selected to control model complexity and enhance generalisation to spatially independent data, following established practices for gradient boosting models (Chen and Guestrin, 2016).

The final hyperparameter configurations used for model training and prediction are summarised in Table 3. These settings were kept fixed for all experiments to ensure a fair and consistent comparison between the two classifiers.

RF hyperparamter	Values	XGB hyperparameter	Values
n_estimators	600	n_estimators	300
max_depth	12	max_depth	6
min_samples_split	50	learning_rate	0.1
min_samples_leaf	20	subsample	0.8
max_features	sqrt	colorsample_bytree	0.8
bootstrap	TRUE	min_child_weight	1
class-weight	balanced	gamma	0
/	/	reg_lamda	1

Table 3: Hyperparameter settings for RF and XGB models.

3.5 Classification and Training

Supervised classification was performed to distinguish active cropland from cropland abandonment (CA) using the multi-temporal feature stack described in the previous subsection. Two ensemble-based machine learning algorithms were employed: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). These models were selected due to their proven effectiveness in remote sensing applications, particularly for high-dimensional datasets and non-linear class boundaries (Belgiu and Drăguţ, 2016; Chen and Guestrin, 2016).

Both classifiers were trained exclusively on samples extracted from the spatially independent training zone, ensuring that no information from the testing area influenced model learning. Training samples consisted of pixel-level observations extracted from the sampling windows and linked to the corresponding multi-temporal feature vectors. The moderate class imbalance between active cropland and CA was addressed during model configuration, including the use of class weighting where applicable, to reduce bias toward the majority class.

For the RF classifier, model training relied on an ensemble of decision trees constructed using bootstrap sampling and random feature selection at each split. This design enhances robustness to noise and reduces overfitting, particularly when predictor variables are correlated, as is common in multi-temporal remote sensing features (Belgiu and Drăguţ, 2016).

XGB training followed a gradient boosting framework, where decision trees are built sequentially to minimise classification error through gradient-based optimisation. Regularisation and subsampling strategies were applied to improve generalisation performance and reduce sensitivity to noisy predictors (Chen and Guestrin, 2016).

After training, both models were applied to the independent testing zone using the same 56-band feature stack. This resulted in spatially explicit classification maps for RF and XGB, representing predicted CA and active cropland classes. Applying the models to a geographically separate testing area provides a conservative and realistic assessment of model transferability and spatial generalisation (Roberts et al., 2017; Meyer et al., 2019).

Model training and prediction were implemented using a QGIS machine learning plugin, which enables integration of raster-based feature stacks, reference samples, and classifier hyperparameters within a geospatial processing environment. The resulting classification outputs were exported as raster maps for subsequent accuracy assessment and spatial error analysis.
This classification setup allows a direct comparison between RF and XGB under identical data conditions and spatial validation constraints, supporting a robust evaluation of their performance in CA detection.

3.6 Accuracy Assessment and Evaluation Metrics

The performance of the supervised classification models was evaluated using an independent testing dataset and a set of standard accuracy metrics suitable for binary classification with class imbalance (Pontius and Millones, 2011). Accuracy assessment was conducted at the pixel level by comparing predicted class labels with reference data extracted from the testing zone, which was spatially independent from the training data.

For each classifier (RF and XGB), a confusion matrix was constructed to summarise the agreement between predicted and reference class labels. The confusion matrix provides the basis for deriving class-specific and overall accuracy metrics and allows explicit identification of omission and commission errors for both active cropland and cropland abandonment (CA) (Pontius and Millones, 2011).
To comprehensively assess model performance, the following metrics were computed as follows:

Overall Accuracy (OA) representing the proportion of correctly classified pixels:

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

User's Accuracy (UA) or Precision, measures the reliability of the mapped classes:

$$UA = \frac{TP}{TP + FP}$$

Producer's Accuracy (PA) or Recall, measuring the ability to correctly identify reference samples:

$$PA = \frac{TP}{TP + FN}$$

F1-score, representing the mean of Precision and Recall:

$$F1 = 2 * \frac{UA * PA}{UA + PA}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively, considering CA as the positive class. These metrics were selected because they provide complementary information on classification performance and are robust to moderate class imbalance, which is common in CA mapping (Pontius and Millones, 2011).
In addition to numerical accuracy metrics, spatial error analysis was performed to examine the geographic distribution of classification errors. Error maps were generated by comparing predicted and reference class labels in the testing zone, allowing visualisation of correctly classified pixels as well as omission and commission errors. Such spatial analysis supports interpretation of model behaviour and helps identify systematic error patterns that may not be evident from summary statistics alone (Roberts et al., 2017; Meyer et al., 2019).

Finally, the spatial extent of CA estimated by each model was calculated from the classification maps and compared with the reference data. Area-based comparison provides an additional perspective on model bias, particularly in terms of overestimation or underestimation of CA, and complements pixel-based accuracy metrics (Pontius and Millones, 2011). All accuracy metrics, confusion matrices, and spatial analyses were computed using Python-based workflows in Jupyter Lab, ensuring transparent and reproducible evaluation of model performance.

4. Results

4.1 Classification Accuracy

Model performance was evaluated on the independent testing zone described in Section 2.1. Confusion matrices and class-specific accuracy metrics were derived at the pixel level and form the basis for the quantitative comparison between the two classifiers (Tables 4 & 5).

Confusion matrix RF	Predicted Active	Predicted Abandoned	Total
Reference Active	23175	27016	50191
Reference Abandoned	22509	77544	100053
Total	45684	104560	150244

Table 4: Confusion matrix for RF classification results on the independent testing dataset.

Confusion matrix XGB	Predicted Active	Predicted Abandoned	Total
Reference Active	10383	39808	50191
Reference Abandoned	4526	95527	100053
Total	14909	135335	150244

Table 5: Confusion matrix for XGB classification results on the independent testing dataset.

The confusion matrices reveal distinct error distributions for the two models. The RF model exhibits a relatively balanced distribution of omission and commission errors across the two classes. Both active cropland and cropland abandonment (CA) are affected by misclassification, indicating a more conservative classification behaviour. In contrast, the XGB model shows a pronounced tendency toward classifying pixels as CA, resulting in a substantially higher number of commission errors for CA and a corresponding reduction in correctly classified active cropland pixels.

These differences are reflected in the derived accuracy metrics (Table 6). The RF model achieved an overall accuracy (OA) of 0.670, while XGB reached a higher OA of 0.705. For CA, RF produced a user's accuracy (UA) of 0.742 and a producer's accuracy (PA) of 0.775, resulting in an F1-score of 0.757.

XGB, by contrast, achieved a very high PA for CA (0.954), indicating strong sensitivity in detecting abandoned cropland, but a lower UA (0.705), reflecting an increased rate of false positives. Consequently, XGB attained a higher F1-score for CA (0.811), despite its reduced discrimination for active cropland.

Model	Class	Precision (UA)	Recall (PA)	F1-score	Overall Accuracy (OA)
RF	Active	0.507	0.461	0.483	0.670
RF	Abandoned	0.741	0.775	0.757	0.670
XGB	Active	0.696	0.206	0.318	0.704
XGB	Abandoned	0.705	0.954	0.811	0.704

Table 6: Overall accuracy and class-specific accuracy metrics (UA, PA, F1) for RF and XGB models.

To facilitate a more intuitive comparison of class-specific performance, the accuracy metrics for both models were additionally visualized using bar charts (Figure 4). These graphs illustrate differences in UA, PA, and F1-score between RF and XGB for both classes and highlight the trade-off between detection sensitivity and classification reliability.

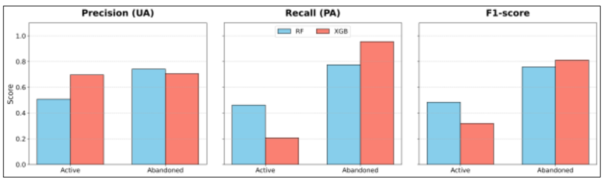

Figure 4: Comparison of user's accuracy, producer's accuracy, and F1-score for RF and XGB models per class.

The graphical comparison confirms that XGB prioritises cropland abandonment detection by achieving consistently higher producer's accuracy (recall), while RF exhibits a more balanced trade-off between recall and precision across classes. From an application perspective, these results suggest that XGB may be preferable in scenarios where minimising missed CA pixels is critical, whereas RF may be more suitable when limiting false positives and maintaining conservative CA estimates is a priority.

Overall, the quantitative accuracy assessment demonstrates that both models are capable of detecting CA using multi-temporal Landsat-derived features, but they differ substantially in class-wise error characteristics.

4.2 Spatial Distribution of Classification Results and Errors

In addition to quantitative accuracy metrics, the spatial characteristics of the classification results were analysed to better understand how the RF and XGB models behave across the landscape. Spatial analysis is particularly important for cropland abandonment (CA) mapping, as classification errors are often spatially structured and may reflect underlying landscape heterogeneity rather than random noise (Roberts et al., 2017; Meyer et al., 2019).

First, spatially explicit classification maps were produced for the independent testing zone using both RF and XGB models in QGIS. These maps are shown in Figure 5 & 6. The maps illustrate the geographic distribution of predicted active cropland and CA and provide an initial visual assessment of model behaviour.

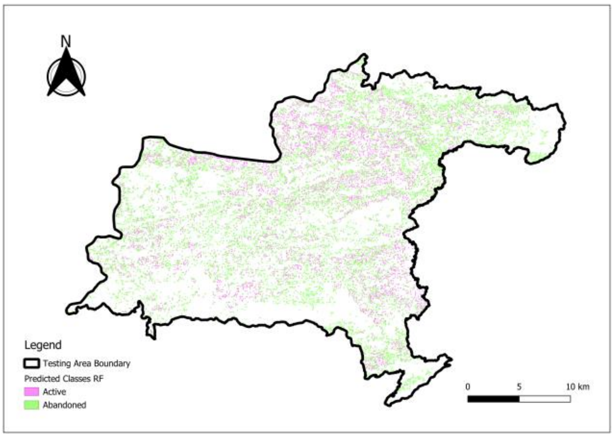

Figure 5: Predicted cropland abandonment maps for the testing area produced by the RF model.

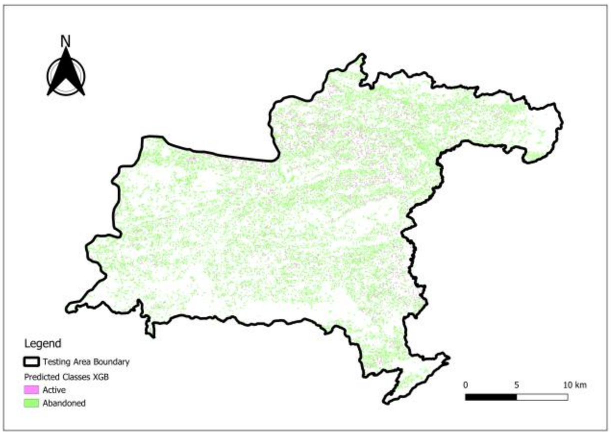

Figure 6: Predicted cropland abandonment map for the testing area produced by the XGB model.

Visual inspection of the predicted maps reveals notable differences between the two classifiers. The RF map shows a more fragmented spatial pattern of CA, with abandoned areas occurring in spatially discrete patches. In contrast, the XGB map exhibits more spatially extensive and contiguous areas classified as CA, particularly in heterogeneous agricultural regions. This pattern indicates a stronger tendency of XGB to assign the CA class, consistent with its higher producer's accuracy (recall) for CA reported in Section 4.1.

To further investigate classification reliability, spatial error maps were generated by comparing predicted class labels with reference data in the testing zone, considering CA as the positive class for both models in (Figure 7 & 8). Pixels were categorized as correctly classified, omission errors (reference CA classified as active cropland), or commission errors (reference active cropland classified as CA).

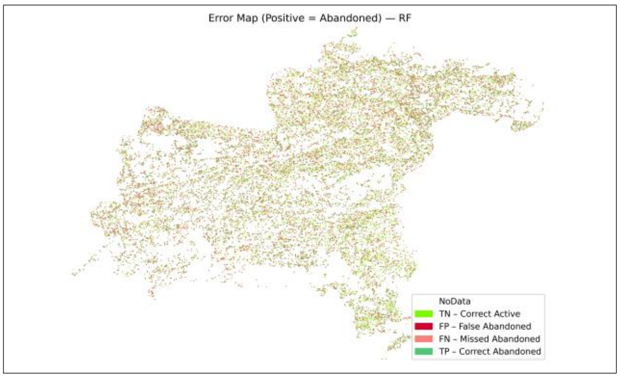

Figure 7: Spatial error maps for RF classification in the testing area (CA as the positive class).

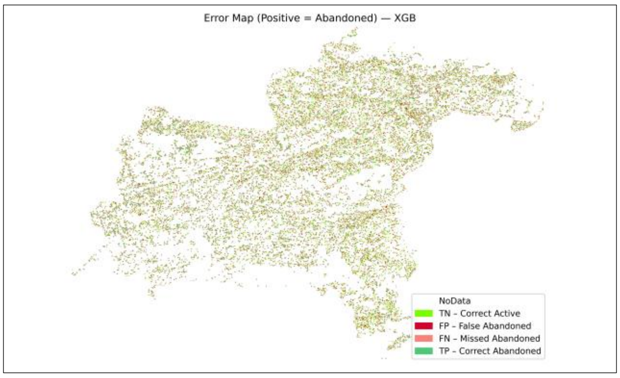

Figure 8: Spatial error maps for XGB classification in the testing area (CA as the positive class).

The RF error map indicates a relatively balanced distribution of omission and commission errors. Omission errors tend to occur in areas where abandonment signals are temporally ambiguous, while commission errors are generally spatially dispersed and limited in extent. This pattern reflects the more conservative classification behaviour of RF and its relatively balanced user's and producer's accuracy for CA.

In contrast, the XGB error map is dominated by commission errors forming large and spatially coherent clusters. These errors correspond to areas of active cropland exhibiting temporal signatures partially resembling abandonment, leading XGB to overclassify CA. The spatial coherence of these commission errors suggests that XGB captures broad temporal patterns but applies them more uniformly across space, increasing the risk of systematic overestimation in certain landscape contexts.

Comparing the predicted maps and error maps highlights that the differences between RF and XGB are not limited to overall accuracy values but are strongly expressed in the spatial structure of classification outputs. While RF produces more spatially conservative and fragmented CA patterns, XGB generates smoother and more extensive CA regions, which may be

advantageous for detecting widespread abandonment but problematic for precise spatial delineation.
These spatial differences have direct implications for estimating the total extent of CA and are examined quantitatively in the following subsection through an area-based comparison of classification results.

4.3 Comparison of Cropland Abandonment Area Estimates

In addition to pixel-level accuracy and spatial error patterns, the reliability of cropland abandonment (CA) mapping was further evaluated by comparing the total area of CA estimated by each classification model. Area-based assessment is particularly important for regional-scale land-use analyses, where classification outputs are often used to support monitoring, reporting, and policy-related applications (Pontius and Millones, 2011).
The total CA area was calculated from the predicted classification maps for the independent testing zone and compared with the reference data. The resulting area estimates are summarised in Table 7.

Source	Abandoned area (km²)
GT	105.97
RF	94.10
XGB	121.80

Table 7: Comparison of cropland abandonment area estimates derived from reference data, RF, and XGB.

According to the reference dataset, the total CA area within the testing zone amounts to 105.97 km²**.** The RF model estimated a CA area of 94.10 km²**,** corresponding to an underestimation relative to the reference data. In contrast, the XGB model produced a substantially larger estimate of 121.80 km², indicating a pronounced overestimation of abandoned land extent.

These differences are consistent with the class-specific accuracy metrics and spatial error patterns presented in the previous subsections. The RF model's relatively balanced user's and producer's accuracy for CA results in an area estimate that remains closer to the reference value, despite a tendency toward omission errors. This behaviour reflects the more conservative spatial patterns observed in the RF predicted and error maps, where commission errors are limited in extent.

The XGB model's area overestimation can be directly linked to its strong tendency to classify pixels as CA. As demonstrated in Sections 4.1 and 4.2, XGB achieves very high producer's accuracy for CA but at the expense of a high number of commission errors. These false positives accumulate spatially, leading to a substantial inflation of the estimated CA area. This effect is particularly evident in heterogeneous agricultural landscapes, where temporal signals may partially resemble abandonment without representing true CA. The area comparison demonstrates that higher classification accuracy does not necessarily translate into more reliable area estimates (Pontius and Millones, 2011).

While XGB outperforms RF in terms of overall accuracy and CA F1-score, its tendency to overestimate CA extent may limit its suitability for applications requiring accurate quantification of abandoned land. Conversely, RF provides more conservative area estimates that are closer to the reference data, which may be preferable for regional monitoring and reporting purposes. Overall, the area-based assessment highlights an important trade-off between detection sensitivity and area reliability in CA mapping. These findings are further interpreted and placed in a broader methodological and application context in the following section.

5. Discussion

This study demonstrates that multi-temporal Landsat-derived features combined with machine learning classifiers can effectively detect cropland abandonment (CA) under a spatially independent validation framework. The results highlight not only differences in overall classification accuracy between Random Forest (RF) and XGBoost (XGB), but also substantial contrasts in class-specific behaviour, spatial error patterns, and area estimates, which are critical for interpreting CA mapping outcomes (Prishchepov et al., 2012; Alcántara et al., 2013).

5.1 Influence of Multi-Temporal Features on Cropland Abandonment Detection

The results demonstrate that the use of multi-temporal spectral features plays a central role in detecting cropland abandonment (CA) from medium-resolution satellite data. By integrating annual spectral indices with temporal descriptors, the proposed approach captures long-term land-use dynamics that are not observable in single-date imagery (Prishchepov et al., 2012; Alcántara et al., 2013). This is particularly important for CA mapping, as abandonment typically manifests as gradual changes in vegetation cover and management intensity rather than abrupt spectral shifts (Lambin and Geist, 2006; Prishchepov et al., 2012).

Temporal information derived from NDVI and BSI proved especially relevant for distinguishing CA from active cropland. Long-term trends reflect persistent changes in vegetation activity and soil exposure, while interannual variability provides insight into the stability of land management practices. Actively managed cropland often exhibits pronounced year-to-year variability associated with crop rotation and cultivation cycles, whereas abandoned land tends to show more stable or gradually evolving temporal signatures. This pattern is consistent with previous studies highlighting the importance of vegetation trajectories and temporal aggregation for CA detection (Prishchepov et al., 2012; Alcántara et al., 2013).

The integration of multiple spectral indices further enhances the representation of land surface conditions. While NDVI captures vegetation productivity, BSI provides complementary information on soil exposure, and indices such as NDBI and MNDWI help account for built-up influence and moisture-related effects (Alcántara et al., 2013). Combining these indices across multiple years into a single feature space allows machine learning models to exploit both spectral and temporal contrasts between abandoned and actively cultivated fields (Belgiu and Drăguţ, 2016).

However, the results also indicate that the effectiveness of multi-temporal features depends on how classification algorithms utilise them. Although both RF and XGB were trained on the same feature stack, their different learning strategies resulted in distinct interpretations of temporal information, leading to contrasting classification behaviours. This suggests that feature design alone does not determine mapping outcomes but interacts strongly with model characteristics, which is further discussed in the following subsection.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-679-2026

5.2 Comparison of RF and XGB Classification Behaviour

Although both classifiers were trained using the same multi-temporal feature set and identical training samples, the results reveal clear differences in how Random Forest (RF) and XGBoost (XGB) interpret and exploit the available information for cropland abandonment (CA) detection. These differences are evident in class-specific accuracy metrics, spatial error patterns, and area estimates.

RF exhibited a more conservative and balanced classification behaviour. Its performance is characterised by moderate producer’s and user’s accuracy for CA, reflecting a trade-off between detecting abandoned land and limiting false positives. This behaviour is reflected in the spatial error maps, where RF errors are more fragmented and less spatially coherent. Such patterns suggest that RF responds more locally to variations in the feature space, reducing the risk of large-scale overclassification but increasing the likelihood of missing subtle abandonment signals (Belgiu and Drăguţ, 2016).

In contrast, XGB showed a strong tendency toward identifying the CA class. The very high producer’s accuracy for CA indicates that XGB successfully captures most abandoned pixels in the testing zone. However, this sensitivity comes at the cost of reduced discrimination for active cropland, as reflected by the high number of commission errors and the systematic overestimation of CA area. Spatially, these commission errors form contiguous clusters, indicating that XGB applies learned temporal patterns more uniformly across the landscape (Chen and Guestrin, 2016).

These contrasting behaviours are consistent with the underlying learning mechanisms of the two algorithms. RF relies on ensemble averaging of independently grown decision trees, which tends to stabilise predictions and limit extreme responses to specific feature patterns (Belgiu and Drăguţ, 2016). XGB, by contrast, employs a boosting strategy that emphasises patterns contributing most to error reduction, which can amplify systematic tendencies toward a particular class when strong but ambiguous temporal signals are present (Chen and Guestrin, 2016).

From a practical perspective, these findings highlight an important trade-off in CA mapping. XGB may be more suitable for applications prioritising comprehensive detection of potential abandonment areas, where omission errors are critical. RF, on the other hand, may be preferable when conservative mapping and reliable area estimation are required. This comparison underscores that model selection should be guided by application-specific objectives rather than overall accuracy alone.

5.3 Importance of Spatially Independent Validation

The use of a spatially independent training and testing design had a significant influence on the observed classification performance and the interpretation of results. By dividing the study area into geographically distinct zones prior to sample generation, the validation strategy explicitly reduced spatial autocorrelation between training and testing data. This approach addresses a well-known limitation in spatial classification studies, where random pixel-based or object-based splits can lead to overly optimistic accuracy estimates due to spatial dependence among neighbouring samples (Roberts et al., 2017; Meyer et al., 2019).

The moderate overall accuracy values obtained for both RF and XGB should therefore be interpreted in the context of this strict validation framework. Unlike random sampling approaches, the spatial split forces the models to generalize to areas with different local conditions, management histories, and landscape structures. As a result, the reported accuracy metrics provide a more realistic estimate of model performance when applied to new geographic regions with similar characteristics. The spatial validation strategy also helps explain the differences observed between the two classifiers. The tendency of XGB to overclassify cropland abandonment becomes more pronounced when applied to a spatially independent testing zone, where temporal patterns learned during training may not fully correspond to local land-use dynamics. RF, while less sensitive to subtle abandonment signals, demonstrates greater stability under spatial transfer, producing more conservative and spatially heterogeneous classification outputs.

Overall, the adoption of spatially independent validation strengthens the methodological robustness of this study and enhances the credibility of the reported results. It also highlights the importance of considering spatial dependence explicitly when evaluating machine learning models for cropland abandonment mapping and other land-use classification tasks (Roberts et al., 2017; Meyer et al., 2019).

5.4 Implications for Cropland Abandonment Mapping

The results of this study have important implications for cropland abandonment (CA) mapping, particularly when classification outputs are used not only for spatial pattern analysis but also for quantifying abandoned land extent (Pontius and Millones, 2011). The comparison between RF and XGB demonstrates that differences in model behaviour directly affect CA area estimates, even when overall classification accuracy appears comparable.

The area-based assessment revealed that XGB substantially overestimates CA extent, whereas RF produces more conservative estimates that are closer to the reference data. This finding highlights that high sensitivity in detecting CA pixels does not necessarily translate into reliable estimates of abandoned land area (Pontius and Millones, 2011). In regional-scale applications, such overestimation may lead to inflated assessments of abandonment intensity and potentially misleading conclusions for land-use monitoring and policy support. Conversely, the more conservative behaviour of RF suggests that it may be better suited for applications where accurate area quantification is a primary objective. Although RF misses a higher proportion of CA pixels compared to XGB, its lower rate of commission errors limits the accumulation of false positives and results in more stable area estimates. This trade-off illustrates that model selection should be guided by the intended application, whether prioritising comprehensive detection of abandonment or reliable estimation of its spatial extent.

These findings emphasise the need to evaluate CA mapping approaches using both pixel-based accuracy metrics and area-based indicators (Pontius and Millones, 2011). Relying solely on overall accuracy or F1-score may obscure important differences in how classification errors propagate into area estimates. Integrating spatial error analysis and area comparison into the evaluation framework provides a more complete understanding of model performance and supports informed decision-making in CA mapping applications.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-679-2026 | © Author(s) 2026. CC BY 4.0 License.

6. Conclusion

This study demonstrated the potential of multi-temporal Landsat data combined with machine learning methods for detecting cropland abandonment (CA) under a spatially independent validation framework. By integrating annual spectral indices with temporal trend and variability descriptors, the proposed approach effectively captured long-term land-use dynamics associated with CA. The comparison between Random Forest (RF) and XGBoost (XGB) revealed substantial differences in classification behaviour. While XGB achieved higher overall accuracy and stronger sensitivity in detecting CA, it exhibited a pronounced tendency to overestimate abandoned land extent due to extensive commission errors. RF, in contrast, provided more conservative and spatially stable classification results, producing CA area estimates closer to the reference data. These findings highlight that higher classification accuracy does not necessarily translate into more reliable area estimates (Pontius and Millones, 2011).

A key methodological contribution of this study is the application of spatially independent training and testing zones, which reduced the influence of spatial autocorrelation and enabled a more realistic assessment of model generalization performance. Overall, the results underline the importance of combining multi-temporal feature design, spatially explicit validation strategies, and complementary evaluation metrics when mapping cropland abandonment. The proposed workflow provides a transferable framework for regional-scale CA mapping using medium-resolution satellite data and supports informed model selection depending on application-specific objectives.

Acknowledgements

This paper is part of the dissemination activities of the research project “Geospatial information technologies for resilient and sustainable society – GeoAI” (GC 0006) funded by the Slovenian Research and Innovation Agency (ARIS).

References

Alcántara, C., Kuemmerle, T., Prishchepov, A. V., & Radeloff, V. C., 2013. Mapping abandoned agriculture with multi-temporal MODIS satellite data. Remote Sensing of Environment, 124, 334–347.
https://doi.org/10.1016/j.rse.2012.05.019

Belgiu, M., & Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31.
https://doi.org/10.1016/j.isprsjprs.2016.01.011

Chen, T., & Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
https://doi.org/10.1145/2939672.2939785

Estel, S., Kuemmerle, T., Alcántara, C., Levers, C., Prishchepov, A. V., & Hostert, P., 2015. Mapping farmland abandonment and recultivation across Europe using MODIS NDVI time series. Remote Sensing of Environment, 163, 312–325.
https://doi.org/10.1016/j.rse.2015.03.028

Kuemmerle, T., Hostert, P., Radeloff, V. C., van der Linden, S., Perzanowski, K., & Kruhlov, I., 2013. Cross-border comparison of post-socialist farmland abandonment in the Carpathians. Ecosystems, 11(4), 614–628.
https://doi.org/10.1007/s10021-008-9146-z

Lambin, E. F., & Geist, H. J., 2006. Land-Use and Land-Cover Change: Local Processes and Global Impacts. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/3-540-32202-7

Lambin, E. F., & Meyfroidt, P., 2011. Global land use change, economic globalization, and the looming land scarcity. Proceedings of the National Academy of Sciences, 108(9), 3465–3472.
https://doi.org/10.1073/pnas.1100480108

Maxwell, A. E., Warner, T. A., & Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784–2817.
https://doi.org/10.1080/01431161.2018.1433343

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling, 411, 108815.
https://doi.org/10.1016/j.ecolmodel.2019.108815

Pontius, R. G., & Millones, M., 2011. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32(15), 4407–4429.
https://doi.org/10.1080/01431161.2011.552923

Prishchepov, A. V., Radeloff, V. C., Dubinin, M., & Alcántara, C., 2012. The effect of Landsat ETM/ETM+ image acquisition dates on the detection of agricultural land abandonment in Eastern Europe. Remote Sensing of Environment, 126, 195–209.
https://doi.org/10.1016/j.rse.2012.08.017

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.
https://doi.org/10.1111/ecog.02881

Stefanski, J., Kuemmerle, T., Chaskovskyy, O., Griffiths, P., Havryluk, V., Knorn, J., Korol, N., Sieber, A., Waske, B., 2014. Mapping land management regimes in western Ukraine using optical and SAR data. Remote Sensing of Environment, 6(6), 5279-5305.
https://doi.org/10.3390/rs6065279

Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., Allen R. G., Anderson M. C., Belward, A. S., Cohen, W. B., Dwyer, J., Erb, A., Gao, F., Griffiths, P., Helder, D., Hermosilla, T., Hipple, J. D., Hostert, P., Hughes, J. M., Huntington, J., Johnson, D. M., Kennedy, R., Kilic, A., Li, Z., Lymburner, L., McCorkel, J., Pahlevan, N., Scambos, T. A., Schaaf, C., Schott, J. R., Sheng, Y., Storey, J., Vermote, E., Vogelmann, J., White, J. C., Wynne, R. H., Zhu, Z., 2019. Current status of Landsat program, science, and applications. Remote Sensing of Environment, 225, 127–147.
https://doi.org/10.1016/j.rse.2019.02.015