


Article

From Local to Global: A Transfer Learning-Based Approach for Mapping Poplar Plantations at Large Scale

Yousra Hamrouni ^{1,2} , Éric Paillassa ³ , Véronique Chéret ¹, Claude Monteil ¹  and David Sheeren ¹ 

¹ Université de Toulouse, INRAE, UMR DYNAFOR, Castanet-Tolosan, France

² Conseil National du Peuplier, Paris, France

³ Centre National de la Propriété Forestière, Institut pour le Développement Forestier, Bordeaux, France

* Correspondence: yousra.hamrouni@inrae.fr

Abstract: Reliable estimates of poplar plantations area are not available at the French national scale due to the unsuitability and low update rate of existing forest databases for this short-rotation species. While supervised classification methods have been shown to be highly accurate in mapping forest cover from remotely sensed images, their performance depends to a great extent on the labelled samples used to build the models. In addition to their high acquisition cost, such samples are often scarce and not fully representative of the variability in class distributions. Consequently, when classification models are applied to large areas with high intra-class variance, they generally yield poor accuracies. In this paper, we propose the use of active learning (AL) to efficiently adapt a classifier trained on a source image to spatially distinct target images with minimal labelling effort and without sacrificing classification performance. The adaptation consists in actively adding to the initial local model, new relevant training samples from other areas, in a cascade that iteratively improves the generalisation capabilities of the classifier, leading to a global model tailored to different areas. This active selection relies on uncertainty sampling to directly focus on the most informative pixels for which the algorithm is the least certain of their class labels. Experiments conducted on Sentinel-2 time series showed that when the same number of training samples was used, active learning outperformed passive learning (random sampling) by up to 5% of overall accuracy and up to 12% of class F-score. In addition, and depending on the class considered, the random sampling required up to 50% more samples to achieve the same performance of an active learning-based model. Moreover, the results demonstrate the suitability of the derived global model to accurately map poplar plantations among other tree species with overall accuracy values up to 14% higher than those obtained with local models. The proposed approach paves the way for national-scale mapping in an operational context.

Keywords: active learning; poplar plantations; spatial transfer; sentinel-2; large scale; image classification; random forest

1. Introduction

Poplar (*Populus spp.*) is one of the fast-growing and wood producing trees which are increasingly considered as an important resource to meet the global demand for natural forest products. Poplars are basic raw materials for industrial processing and provide valuable non-wood forest products. According to a report from the Food and Agriculture Organization of the United Nations (FAO) at the 25th session of the International Poplar Commission (IPC), France is the first European country in terms of planted poplar area with about 0.2 million hectares [1].

In France, poplar cultivation is a key local industry which involves silviculturists, nurserymen, forest managers and wood processors and is coordinated by the National Poplar Council (CNP). Despite the increased demand for poplar wood, planted areas continue to decrease [2]. Indeed, for the past 20 years, the poplar sector has undergone several economic, social and environmental upheavals and have consequently had an impact on the planting rate [3]. This deficit of several years has led the sector to an unavoidable wood shortage, which is expected to reach at least 500,000 m³/year in 2025, according to the CNP. Considering this risky situation, national strategies have been undertaken to encourage all the industry stakeholders, including providing financial incentives to replant poplar.

As the future of poplar in France relies on these replanted areas, obtaining spatially explicit information on newly-planted and cleared surfaces is crucial. This requires precise mapping and timely characterisation of their spatial and temporal dynamics. However, accurate and regularly updated maps of poplar plantations are still not available at the national scale and the update rate of existing forest databases is not suitable for this species because of its short rotation cycle (from 12 to 15 years on average).

Since the availability of high spatial, spectral and temporal resolution imagery from the Copernicus Sentinel sensors, new opportunities for monitoring poplar plantations over large areas have emerged. Several works have already demonstrated the potential of remote sensing for mapping plantations such as rubber [4–8] and palm oil [9–12] but only a few have focussed on poplar [13–17]. The few studies which did, have addressed the issue in various ways using different data sources and machine learning algorithms. Some used high or very high spatial resolution single-date images often acquired during the dry season [4,13,17]. These authors drew on the spectral information to separate monospecific plantations from the other species. However, the plantations were frequently confused with evergreen species or natural forest. For example, in [4], rubber trees were easily confused with eucalyptus trees despite the use of multi-spectral metrics from nine spectral bands of an ASTER image (15m). Other authors have focussed on temporal information and used multi-temporal approaches. Some considered seasonal variations in vegetation (i.e. *phenology*) using phenological features [8,12,18] while others extracted multi-seasonal texture measures [19,20] or time series derived features such as shapelets [6]. Despite the promising results reported, most of the studies were limited to local scale and the performances depend largely on the data used for training and validation. While different parametric and non-parametric classification approaches have been broadly developed, their application over large and heterogeneous areas remains challenging due to their limited generalisation ability [11,18].

Most classification methods assume that training and test sets share the same feature space and the same distribution [21]. The resulting classification models work well on other areas as long as the distribution has not changed. However, in remote sensing data, spectral reflectance signatures may vary according to acquisition conditions (incidence angle, illumination, relief, etc.). Moreover, in the particular case of vegetation and for a given species, the observed spectra may shift significantly due to several factors including the spatial variability of the phenology (seasonal changes), conditions at the site (type of soil, moisture conditions, etc.) and the forest health status [22]. Direct application of the original trained models in new areas (i.e. across-region generalisation [23]) will therefore result in poor classification performances [24].

Two options could be considered for spatial (and even temporal) generalisation. The first is to build local classification models adapted to each study area. This strategy can work if training data are available everywhere and are sufficiently representative of the spatial variability. At national scale, training data can easily be retrieved from forest databases but their quality may not be satisfactory as the data are not frequently updated. In this situation, photo interpretation is necessary but not practical, and is too expensive (in time and human effort) for large-scale production in an operational context. In addition, photo interpretation depends on the operator and may introduce some bias through information redundancy [24,25]. Indeed, several neighbouring pixels may contain the same

spectral information which can skew the learned model thereby reducing its generalisation ability (overfitting).

An alternative option would be to adapt a model trained locally using an initial feature space (the *source domain*) and make it able to accurately predict other geographical areas based on a different feature space (the *target domain*). From a machine learning point of view, this strategy is known as *Transfer Learning* (TL) [21]. TL has been developed since 1995 in response to these challenges, by leveraging the knowledge gained during initial learning tasks and adapting the subsequent ones.

In the literature, TL can be categorised in three branches, depending on the availability of labelled data and on the relationship between the source and target domains: *inductive*, *unsupervised* and *transductive* transfer learning [21]. The first category (*inductive*) refers to transfer learning with labelled data available in the target domain. In the *unsupervised* approach, no labelled data are available in either the source or target domains. When labelled data are only available in the source domain, the TL is named *transductive*. When the learning task is the same in the source and target domains, the *transductive* transfer learning branch is referred to as *domain adaptation* (DA). In the remote sensing literature, DA techniques can be grouped in four families [26]:

- **DA by selecting invariant features:** this consists in reducing the feature space by selecting only the image features (e.g. raw spectral bands or derived attributes) that are the most robust to change. The latter are described as landmarks in [27] and have the same distribution in both source and target domains. The new feature space is thus more stable and the gap between the two domains is reduced.
- **DA by adapting data distribution:** the aim of this second approach is to create a common feature space for both domains from the two distributions in order to fit the classifier that remains unchanged. Several strategies can be applied such as feature extraction [28] or similarity-based methods [29] for data alignment.
- **DA by adapting the classifier:** in this semi-supervised approach, the classification model already built with the source labelled data is modified to fit the target domain by only considering the distribution of the unlabelled data and never their labels. The classifier is therefore updated to accurately classify the target data.
- **DA by active learning:** it is a particular case of the previous approach in which the classifier is allowed to acquire some labelled data from the target domain. These data are selected according to their potential for improving the initial classification model to correctly predict the target domain [30].

Active learning (AL) has received a great deal of attention from the remote sensing community in the past decade [31]. It has mainly been applied to efficiently select a reduced set of optimal training samples needed by classifiers [25,32–35], sometimes taking the cost of the field data collection into consideration [36,37], but only a few works have been dedicated to its application to transfer learning between two distant domains [26,30,38] and for generalisation across space [39].

The objective of this paper is to evaluate the potential of AL techniques to map poplar plantations at large scale using Sentinel-2 (S2) time series acquired by the mission's two satellites (Sentinel-2A and Sentinel-2B on separate acquisition dates). The proposed approach is initiated by building independent local models derived from traditional supervised classifications and then, by adapting them using AL techniques for transferability, to finally obtain a global model optimised for mapping large areas. The main contributions of this paper are to (i) evaluate the performance of S2 time series in classifying a wide range of poplar plantations in France; (ii) define a robust active learning-based strategy for mapping poplar plantations at national scale, and (iii) enable a clear understanding of the functioning of AL techniques for large scale mapping. In particular, we investigate the impact of the AL direction (from one region to another), the impact of the AL ranking strategy (uncertainty and diversity) to select the most informative samples in the target domain, and the impact of missing classes. We also assess the contribution of AL to each class of the models.

2. Materials

2.1. Study areas

Three different study sites were chosen with forest partners. They are representative of the variability of French poplar plantations in terms of cultivars, silvicultural practices and climatic conditions. The poplar sites are located in northeastern, central and southwestern France and are covered by three S2 tiles with a surface area of 100 km² each (Figure 1).

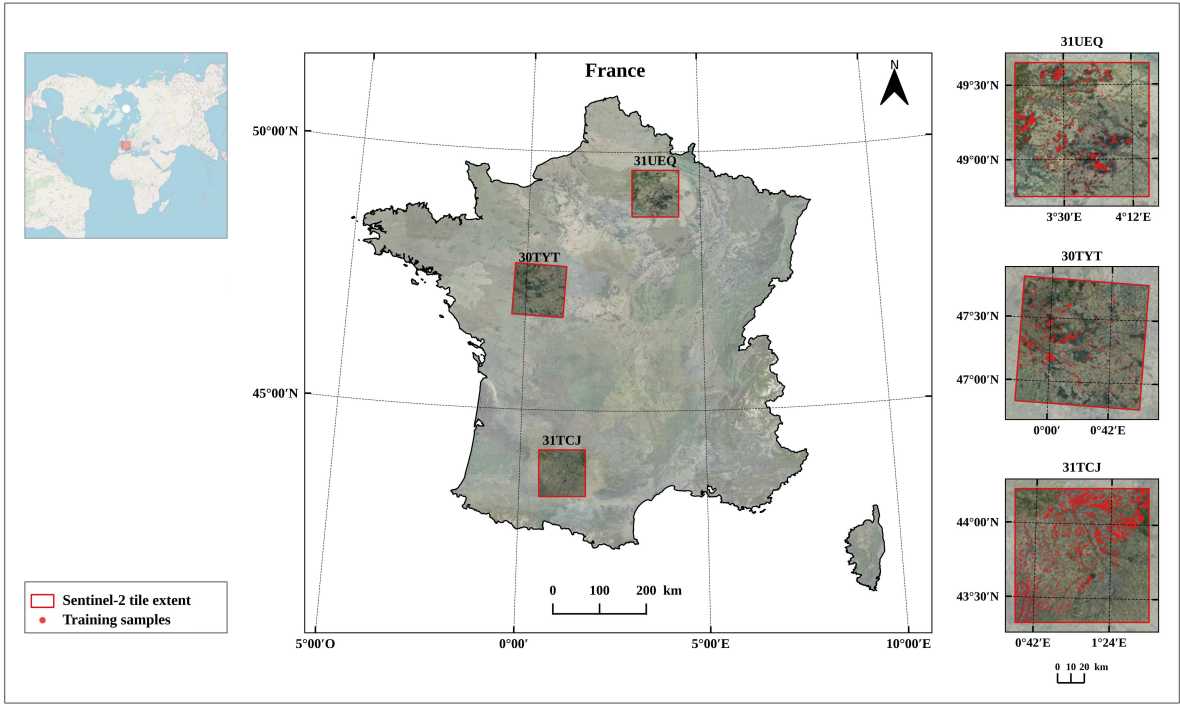


Figure 1. Outline of the study areas used in this work. 31UEQ, 30TYT and 31TCJ are the Sentinel-2 extents at the UTM tiling following the US-MGRS (US Military Grid Reference System).

2.2. Satellite imagery

In this study, Sentinel-2 optical Satellite Image Time Series (SITS) were used. The images were acquired during the Sentinel-2 mission as part of the Copernicus programme of the European Space Agency (ESA). The mission is based on a constellation of two satellites (Sentinel-2A and Sentinel-2B) equipped onboard with an optical MultiSpectral Imager (MSI) acquiring information in 13 different bands spanning from the visible through the short-wave infrared range: four bands at 10 m, six bands at 20 m and three bands at 60 m spatial resolution. The revisit frequency of the combined constellation is five days. For all three study areas (S2 tiles), all available acquisitions were downloaded from the French Theia Land Data Center. They are surface reflectance products (level 2A) processed by the MAJA¹ software, orthorectified, atmospherically corrected and provided with a cloud mask [40]. The number of downloaded dates per tile is reported in Table 1.

¹ Software developed in coordination between CNES/CESBIO and DLR using the multi-sensor atmospheric correction and cloud screening (MACCS) method.

Table 1. Properties of the Sentinel-2 tiles used in the study.

Tile code	Relative orbit number	No. of available dates in 2017
31UEQ	51	26
30TYT	94	34
31TCJ	51	36

From one tile to another, the acquired images are not synchronous, either because of sensor-specific constraints or the presence of clouds (Figure 2).

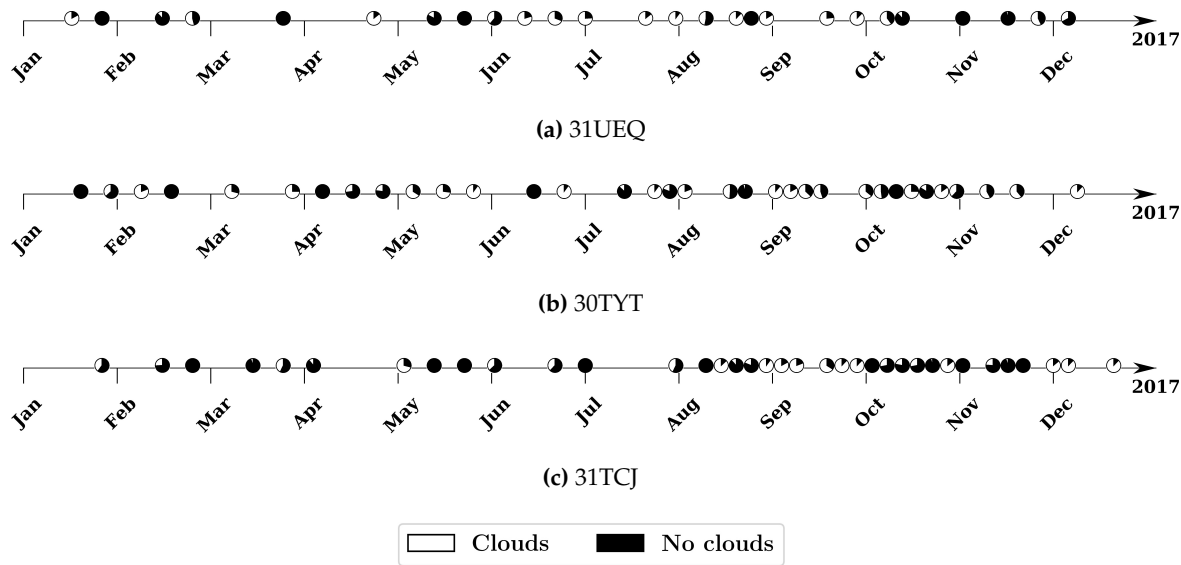


Figure 2. Sentinel-2 acquisitions available in 2017 and percent cloud cover/date in (a) the north-eastern tile 31UEQ, (b) the central tile 30TYT and (c) the south-western tile 31TCJ.

2.3. Reference data

Reference data for training and testing were obtained from the most recent version of the French Forest database (BD Forêt® IGN, v.2) created by the national forest inventory and mapping agency (IGN). The vector data are provided by district and created through photo interpretation of aerial photographs. It took about 11 years to get a complete national coverage with orthophotos (Figure 3) and the forest maps generally became available from two to six years later.

In this study, reference data were extracted for all existing deciduous species, as polygons of pure and mixed classes of a minimum area of one hectare. The polygons include poplar plantations which are always defined as pure stands in the database. This phase was essential to better understand the behaviour of poplars and to detect possible confusion with other deciduous species.

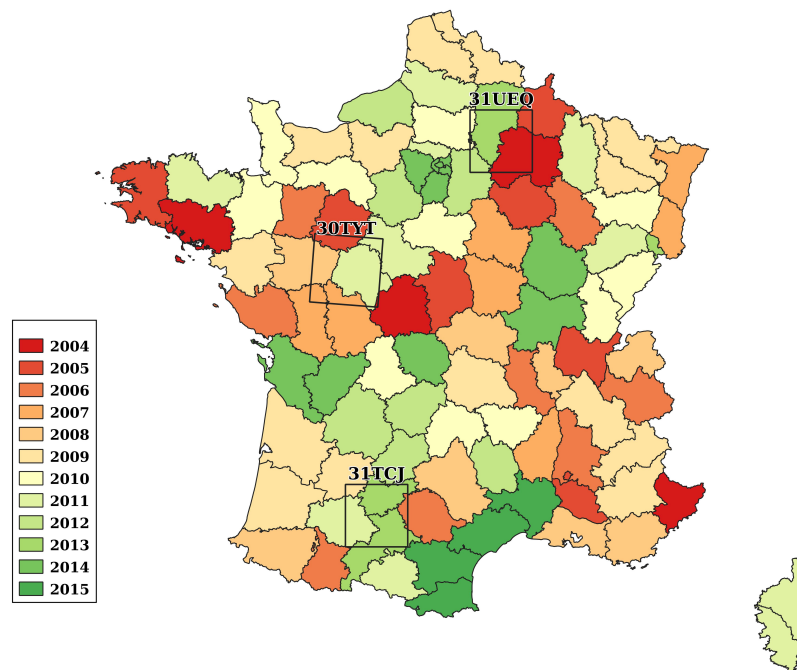


Figure 3. Acquisition years of aerial photographs per district in the forest database used as reference.

The delay between S2 images year (2017), the forest database production period (from 2007 to 2018) and the aerial photographs acquisitions (from 2004 to 2015) could be a potential source of error in the references. With a short production cycle (15 years on average), poplar plantations are the forest stands most likely to be affected by change. Consequently, additional work was required for this class. It consisted in visually inspecting all poplar polygons to check for possible changes and to ensure all references were up to date. Photo interpretation was performed using 50-cm spatial resolution orthophotos and S2 images acquired in 2017. All selected polygons were checked to make sure they had not changed in 2017. It should be noted that the photo-interpreted poplar samples correspond to relatively mature plantations (more than three years old) since there is more uncertainty when the trees are younger, poplar stands may be confused with other species and the canopy cover may be partial. Three experts were responsible for photo interpretation and validation and the error rate was deemed negligible (less than 1% misidentification).

The sample size varied between 2,500 and 7,700 pixels depending on the representativeness of the smallest class within each of the three S2 tiles to ensure class balance in the reference data. Table 2 provides an overview of the total number of samples used (for both training and testing) for each class and in each tile.

Table 2. Number of reference samples per class extracted from each S2 tile. The samples were derived from the French forest database (BD Forêt® IGN, v.2). For poplar plantations which only include stands more than three years old, all the samples were checked and validated by photo interpretation.

Tile code	Sample size in pixels per class						
	Poplar (photo-interpreted)	Locust	Chestnut	Oak	Beech	Closed deciduous forest (mixed)	Open deciduous forest (mixed)
31UEQ	2500	2500	NA ¹	2500	2500	2500	2500
30TYT	4000	4000	4000	4000	NA ¹	4000	4000
31TCJ	7700	7700	7700	7700	NA ¹	7700	7700

¹ Class not available in the study area or poorly represented.

3. Methods

The processing chain we proposed is shown in Figure 4. It consisted of three main steps: pre-processing, local processing and finally global processing. For large scale mapping, only one local model is required to initiate the learning process. The model is then adapted iteratively for spatial transferability, by introducing a limited number of reference data from other areas using active learning. The adaptation is intended to build a global model that can be used beyond the new area while maintaining good predictive performance on the initial one. In order to better assess the predictive power of this global model, independent local models were also learned in each area. This is why three individual models appear in the local processing step (Figure 4) but all the local models are not required in the global process.

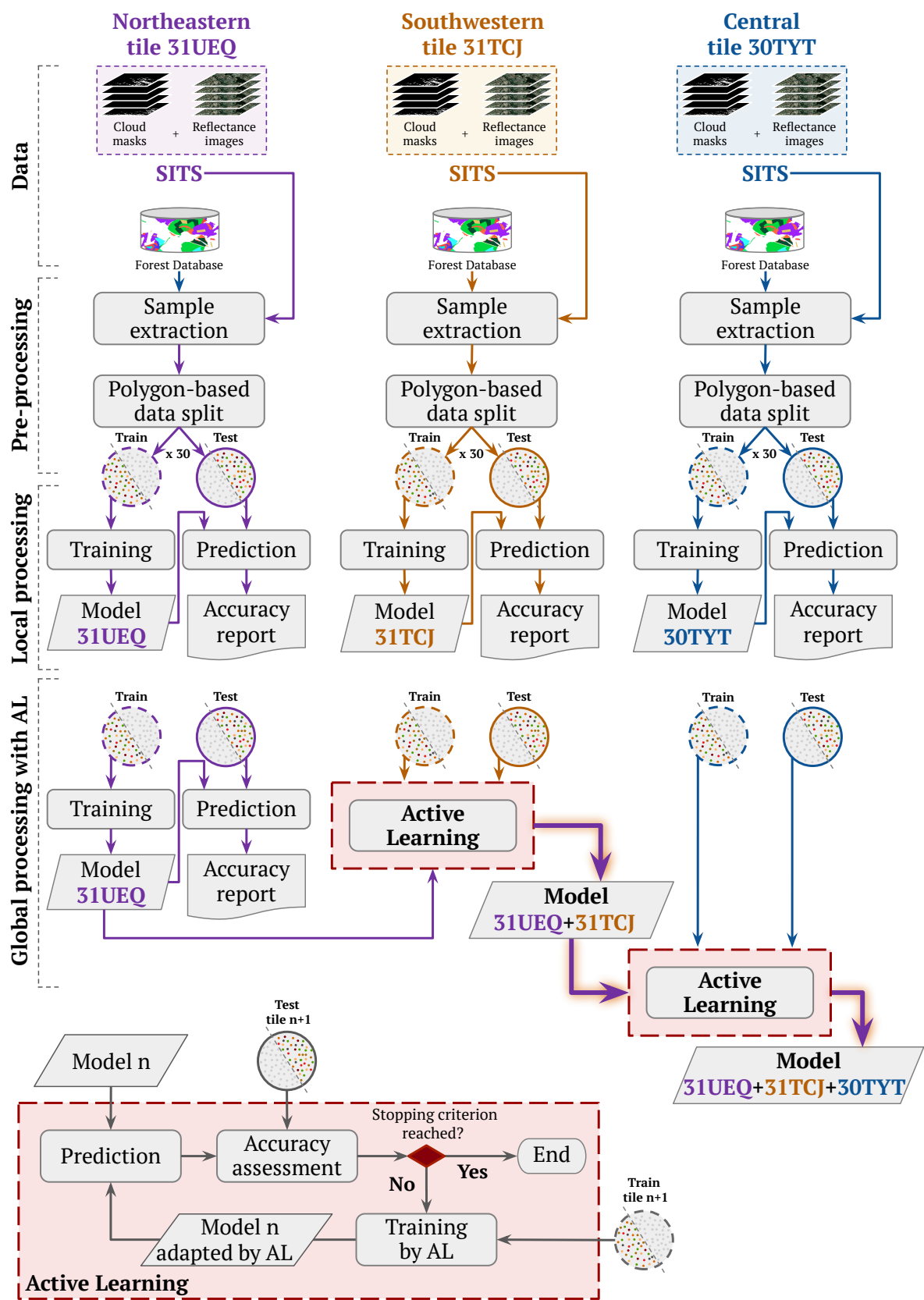


Figure 4. Flow chart of the proposed approach for large-scale classification. The chart illustrates the local classification process used for each of the three study tiles as well as an example of a global processing along a transfer direction from the north-eastern (31UEQ) to the south-western tile (31TCJ) and ending with the central one (30TYT).

3.1. Pre-processing

The pre-processing step was the same for all the study areas represented by the three S2 tiles. As illustrated in Figure 2, the acquisitions are asynchronous between the three tiles as well as the cloud coverage. For each Sentinel-2 SITS, the missing data were therefore filled using multitemporal linear interpolation. This involves replacing each invalid pixel (detected in the cloud mask) by an interpolated value with the closest and valid pixels in the SITS [41]. The time series is then resampled at 10-m spatial resolution with a 10-day time step common to all tiles. The resulting gap-filled SITS comprise 339 bands corresponding to 10 spectral bands * 39 resampled dates over the course of the year.

Reference polygons were randomly split, 50% for training and 50% for testing. To reduce spatial autocorrelation [42], we avoided splitting pixels from the same polygons into training and testing datasets. Sampling was repeated 30 times to quantify variability related to the random selection. Training and testing samples were then built using stratified random sampling of equal size for each class belonging to the same S2 tile.

3.2. Local classification approach

For each of the three S2 processed SITS, a random forest (RF) supervised classification was carried out independently. The main objective was to assess the ability of the algorithm to identify poplar plantations at local scale in the same tile.

RF is one of the most popular non-parametric classifiers and was developed by [43]. It has been used successfully by the remote sensing community thanks to its ability to manage high dimensional data and the easy tuning of its hyperparameters [44]. The RF classifications were performed using the *scikit-learn* Python library [45] with prior optimisation of the algorithm's hyperparameters: the number of trees, the maximum depth of the tree, and the number of features at each split. Optimisation was conducted using a grid-search approach in order to find the best combination of the hyperparameters based on 5-fold cross-validation of training data. The grid search value for the number of trees ranged from 10 to 150 with a step of 10 while the maximum depth varied from 5 to 100 with an interval of 10. The default value was also tested, it consisted in expanding the tree nodes until all the leaves became pure. Finally, the number of features to split a node was fixed as the square root of the total number of input variables [43,44,46]. This value was tested in addition to a range of values from 1 to 20 with a 2-value step. A total of 1,694 combinations were thus performed.

The classifiers were trained using the same number of samples for each class belonging to the same tile. The values of each sample (i.e. one S2 pixel of 10 m² area) were extracted from the 339 spectral bands of the gap-filled S2 SITS. Classifications were repeated 30 times corresponding to the 30 random splits of the reference data into 50% for training and 50% for testing. Confusion matrix and performance measures (Overall Accuracy, F-score, Precision and Recall) were then computed and averaged over the 30 repetitions. The Kappa (κ) metric was deliberately disregarded [47]. This process was conducted for the three S2 study tiles and the results were assessed for each tile separately.

In order to evaluate the importance of checking reference samples of poplar plantations by photo interpretation, the local classification models was similarly built with poplar samples directly derived from the forest database, with no update. The results are reported in Appendix A.

3.3. Global classification approach

The previous section detailed the traditional classification scheme where the learning process and the predictive performance are assessed in each tile independently, assuming that a relatively large quantity of reference samples is available for each region. In practice, this is rarely the case so this strategy is not suitable for large-scale mapping of poplar plantations. If accurate up-to-date reference samples are required (meaning a manual checking step of the forest database polygons is mandatory), to be operational, an alternative strategy will need to be defined. It could be based on the spatial

transfer of a local model. However, because of inter-specific variability across regions, the initial local model will have to be adapted to ensure robust predictions.

In this paper, active learning is proposed as a transfer strategy to build a global model from a local one with the addition of a limited number of reference samples from new geographical areas. The selection of extra labelled samples makes it possible to adapt the original model to other areas (i.e. to the target domain) while remaining valid for the first domain (i.e. the source domain) with a minimum effort required to collect new reference data.

3.3.1. Principles of active learning

Active learning (AL) is based on the assumption that a machine learning algorithm can perform better with fewer training samples than traditional methods if it is allowed to choose the data from which it learns [48]. AL makes iteratively effective queries to label the most informative samples rather than randomly selecting them. These labelled samples are often located in low confidence regions, which helps the initial learner to reduce its uncertainty [31]. The process is iterated until certain predefined stopping criteria are met [49] for instance, the maximum added sample or the maximum performance score.

In the AL literature, there are three common scenarios or ways in which the learner may query the labelled samples: query synthesis, stream-based active learning and pool-based active learning. The last one is the most popular scenario in the community [48].

A pool-based scenario assume that a large pool of unlabelled data \mathcal{U} is available and only the most useful samples \mathcal{L} are selected to be labelled [50,51]. The decision whether to query or reject a sample is made according to a ranking strategy that guides the model to choose the best samples (i.e. the most informative for the learning algorithm). The active learner calculates an information metric of all the samples in the entire pool, ranks them according to this value, selects the most informative ones and then queries their labels (i.e. the classes) to train a new model.

Two ranking criteria can be considered in AL: *uncertainty* and *diversity*. Sampling using *uncertainty* consists in querying only the instances for which the initial learner is the least certain of their labels after prediction [50]. Following this strategy, the most informative samples are the most uncertain. For classification tasks, uncertainty is usually quantified by one of three measures: least confident (LC), margin sampling (MS) or entropy (H) [48]. These measures can be defined as follows (with x_M^* the best instance selected for query using an uncertainty measure M from the unlabelled data pool \mathcal{U}):

- **Least confident (LC):** this metric consists in querying the sample with the least confidence in its most likely label. Using this strategy, only the most probable label is considered and the other probabilities are disregarded:

$$x_{LC}^* = \underset{x}{\operatorname{argmin}} P_{\theta}(\hat{y} | x) \quad (1)$$

where y is the most probable label, \hat{y} is the prediction with the highest posterior probability under the model θ and P_{θ} denotes the probability value with the same model θ .

- **Margin sampling (MS):** this metric seeks to overcome the drawback of the least confident strategy by including the second most probable label. The closer the probabilities, the less confident the model (i.e. great doubt between the two labels). The best instance is the one that minimises this value:

$$x_{MS}^* = \underset{x}{\operatorname{argmin}} [P_{\theta}(\hat{y}_1 | x) - P_{\theta}(\hat{y}_2 | x)] \quad (2)$$

where \hat{y}_1 and \hat{y}_2 are respectively the first and second most probable labels.

- **Entropy (H):** for each sample, the entropy takes the probability of it belonging to all possible model classes into account [52]. When the degree of certainty is high, the entropy value is low (i.e. a high probability of belonging to a specific class). Great uncertainty arises when the probability

values are shared between many classes and none stands out (i.e. high entropy). The samples selected by the algorithm are therefore those with the highest entropy value:

$$\begin{aligned} x_H^* &= \operatorname{argmax}_x H_\theta(Y | x) \\ &= \operatorname{argmax}_x - \sum_y P_\theta(y | x) \log P_\theta(y | x) \end{aligned} \quad (3)$$

where H is Shannon's entropy and y ranges over all possible labels of x .

Regarding the diversity ranking criterion, diversity-based methods tend to exploit the data structure by selecting the samples that are the most representative or diverse of the input space distribution [48]. Density-weighted metrics or clustering-based approaches are generally used and they are frequently combined with an uncertainty metric [53].

Density metrics are based on similarity measures of distance or angle. Samples which minimise the similarity values (i.e. the most dissimilar samples) are favoured. Several similarity metrics are used including Euclidean distance [54] and Cosine similarity [55]. Euclidean distance measures the length of the segment of the straight line connecting two points in Euclidean space while cosine similarity is defined as the cosine of the angle between them.

Clustering-based approaches aim to exploit the data structure by seeking clusters or cluster centroids to obtain labels that are both representative and diverse. With this strategy, the number of clusters has to be determined in advance because it may not match the data labels [48].

3.3.2. Experimental setup for global classification with active learning

Our global classification scheme with AL is illustrated in Figure 4. It starts with an accurate and already trained local classifier on a first data source (the north-eastern S2 tile in the example shown). Then, the initial model (31UEQ) is used to predict the classes in a target domain (here, the south-western S2 tile). Without any adaptation, this first external prediction is assumed not to be very robust. If the predicted performance based on the initial model does not meet a predefined stopping criterion, new samples from the target domain are queried by AL and added to the initial model to be retrained. This procedure is repeated iteratively until the stopping condition (here, defined in terms of the number of samples added from the target domain) is reached. The resulting model (31UEQ+31TCJ) is expected to fit both source and target domains well. The same procedure is used in the following step to adapt the two-tile model (31UEQ+31TCJ) to the third one (30TYT) ultimately leading to a global model tailored to the three S2 tiles.

In order to check the symmetrical transferability of the models from one tile to another, we computed all six possible combinations to build the global model (starting from tile 31UEQ or the others). We assumed a pool-based active learning scenario in which new samples from the target training set are queried based on their informativeness. Two uncertainty metrics were tested to select new samples by AL: entropy (H) and margin sampling (MS). For the purpose of comparison, diversity metrics were also combined with the entropy and margin sampling uncertainty metrics. Euclidean distance and cosine angle were tested. Since AL with the combined uncertainty and diversity metrics performed worse, these results are listed in Appendix D.

In all the experiments, we considered only samples from the pure-species classes. Mixed classes were removed from the source and target sample sets, assuming that they could negatively affect the AL process (see discussion in Section 5.4). AL was run by adding 10 target samples at each iteration up to a maximum of 1,000. From an operational point of view, this maximum number is excessive but in these experiments, it enabled us to understand the impact of increasing the training set size in AL. In order to assess the contribution of AL, target samples were also queried using a random sampling (RS) strategy with the same sample size at each iteration. For both classification procedures (AL *vs.* RS), the performance metrics were calculated on testing sets of both source and target domains

and averaged over the 30 repetitions. The entire AL workflow was implemented in Python using the *modAL* framework [56].

4. Results

4.1. Tile by tile classification: local approach

Table 3 summarizes the results of the random forest classification for each of the three study tiles using the 2017 Sentinel-2 SITS. It will be recalled that our aim here was to assess the potential of S2 data to discriminate poplar plantations from the other deciduous species at the tile scale.

Overall, the results revealed a high capacity of S2 to identify poplar plantations with an average F-score ranging from 89.5% to 99.3%. Without checking reference samples by photo interpretation, the average F-score for poplars dropped from -10% to -20% (see Appendix A). Considering the other classes, accuracy was lower, leading to average values of global F-score (including all the classes) between 73.1% and 80.1%, depending on the S2 tiles.

Table 3. Results of local classification for each S2 tile averaged over 30 independent repetitions.

Tile code	Training size ¹ per class in pixels	No. classes	Overall Accuracy _(*30)	Global F-score _(*30)	Poplar F-score _(*30)	Poplar Precision _(*30)	Poplar Recall _(*30)
31UEQ	1250	6	73.7±1.7 %	73.1±2.0 %	89.5±3.3 %	90.2±6.7 %	89.2±3.2 %
30TYT	2000	6	74.9±1.8 %	75.0±1.9 %	99.3±0.2 %	99.4±0.5 %	99.3±0.6 %
31TCJ	3850	6	80.0±0.7 %	80.1±0.6 %	97.9±0.8 %	99.3±0.5 %	96.5±1.7 %

¹ Training samples represent 50% of the available reference data.

4.2. Global classification approach

4.2.1. Active learning *vs.* random sampling: assessment of overall accuracy

The average overall accuracy (OA) values of the two-tile models are reported in Table 4 according to the uncertainty measure used for AL, the transfer direction, and the number of added samples. In Table 5, we present the OA values achieved with three-tile-based global models for active learning (AL_{MS}).

Table 4. Average overall accuracy (in %) according to the transfer direction from one tile to another and the number of target samples added to the initial set for active learning with Entropy (AL_H) and margin sampling (AL_{MS}) uncertainty metrics. Comparisons were made with an adapted model including additional randomly selected (RS) target samples. The values are averaged over 30 independent repetitions.

		Target tiles														
		31UEQ					30TYT					31TCJ				
Source tiles	OA (%)	Number of additional samples														
		0	250	500	750	1000	0	250	500	750	1000	0	250	500	750	1000
31UEQ	RS	-	-	-	-	-	36	50	57	60	61	47	63	69	71	72
	AL _H	-	-	-	-	-	36	50	56	60	62	47	59	65	69	71
	AL _{MS}	-	-	-	-	-	36	54	61	63	64	47	68	73	74	75
30TYT	RS	31	45	52	55	57	-	-	-	-	-	53	69	70	71	72
	AL _H	31	44	50	53	56	-	-	-	-	-	53	67	70	71	72
	AL _{MS}	31	49	55	58	59	-	-	-	-	-	53	70	72	73	74
31TCJ	RS	40	50	58	61	63	52	61	64	64	65	-	-	-	-	-
	AL _H	40	50	59	63	63	52	62	63	64	65	-	-	-	-	-
	AL _{MS}	40	53	61	64	65	52	63	65	66	66	-	-	-	-	-

Generally, for AL, we observed a better performance with the margin sampling uncertainty metric (AL_{MS}) than with the entropy metric (AL_H). We also observed that classification performance was better with AL (especially using AL_{MS}) than with random sampling (RS). In addition, the results showed asymmetrical transferability particularly when only two tiles were used to build the global model. Performances varied according to the transfer direction. For instance, an average OA score of 61% was obtained with AL_{MS} from 31UEQ to 30TYT by adding extra 500 samples versus 55% in the opposite case (Table 4). However, when the global model included samples from the three S2 tiles (i.e. three-tile models), the classification performances were closer (see results in Table 5 for AL_{MS}). This is particularly true when the north-eastern tile (31UEQ) was included in the two-tile-based models to predict the third tile. From 500 extra samples, we found a difference in OA varying by only 1% or 2% between target tiles 30TYT and 31TCJ (e.g. OA=88% with 750 extra samples for 30TYT versus 89% with the same number of additional samples for 31TCJ). By contrast, when the north-eastern tile 31UEQ was the target, the three-tile-based models were less accurate (difference in OA of around 15% compared to the target tiles 30TYT and 31TCJ). Nevertheless, for all three tiles, the maximum achievable performance appeared to be reached with three-tile-based models and exceeded that of local models by 1% up to 14% depending on the tile considered (see Table 3 for comparison).

Table 5. Average overall accuracy (in %) according to the transfer direction from a two-tiles based model (i.e. an initial local model adapted by AL with 1000 extra samples from a second tile) to a third one. Target samples were added to the initial set by active learning based on margin sampling (AL_{MS}). The values are averaged over 30 independent repetitions.

Source tiles	OA (%)	Target tiles														
		31UEQ					30TYT					31TCJ				
		Number of additional samples														
		0	250	500	750	1000	0	250	500	750	1000	0	250	500	750	1000
30TYT+31TCJ	AL_{MS}	36	61	71	76	78	-	-	-	-	-	-	-	-	-	-
31TCJ+30TYT	AL_{MS}	33	51	65	71	75	-	-	-	-	-	-	-	-	-	-
31UEQ+31TCJ	AL_{MS}	-	-	-	-	-	50	79	85	88	89	-	-	-	-	-
31TCJ+31UEQ	AL_{MS}	-	-	-	-	-	57	79	85	87	88	-	-	-	-	-
31UEQ+30TYT	AL_{MS}	-	-	-	-	-	-	-	-	-	-	47	82	87	89	90
30TYT+31UEQ	AL_{MS}	-	-	-	-	-	-	-	-	-	-	52	84	87	89	90

An illustration of how OA scores fluctuated as a function of the additional target samples is given in Figure 5, in the particular case of transfer from the north-eastern tile (31UEQ) to the south-western tile (31TCJ). At first (i.e. with no additional samples from the target), the original 31UEQ model provided predictions on 31TCJ tile with an average OA of 47% (*vs.* 80% when a local model was used for the target tile, see Table 3, Section 4.1). In that case, the OA scores increased as the target samples were added randomly or by AL, with an advantage for AL_{MS} , especially with few extra samples. On average, AL_{MS} significantly outperformed RS model by 3.5% (Wilcoxon signed-rank test with a p -value < 0.01). A maximum difference of 5.5% between the two learning curves (AL_{MS} *vs.* RS) was also found based on Hausdorff distance. Concerning entropy-based active learning (AL_H), the approach yielded results that closely resembled those of RS.

As shown by the dotted lines, the adaptation of the models for the target domain (tile 31TCJ) does not reduce the classification accuracy in the source domain (tile 31UEQ). The OA scores remained

fairly constant when target samples were added (Figure 5). The same trend was observed regardless of the learning approach (RS or AL).

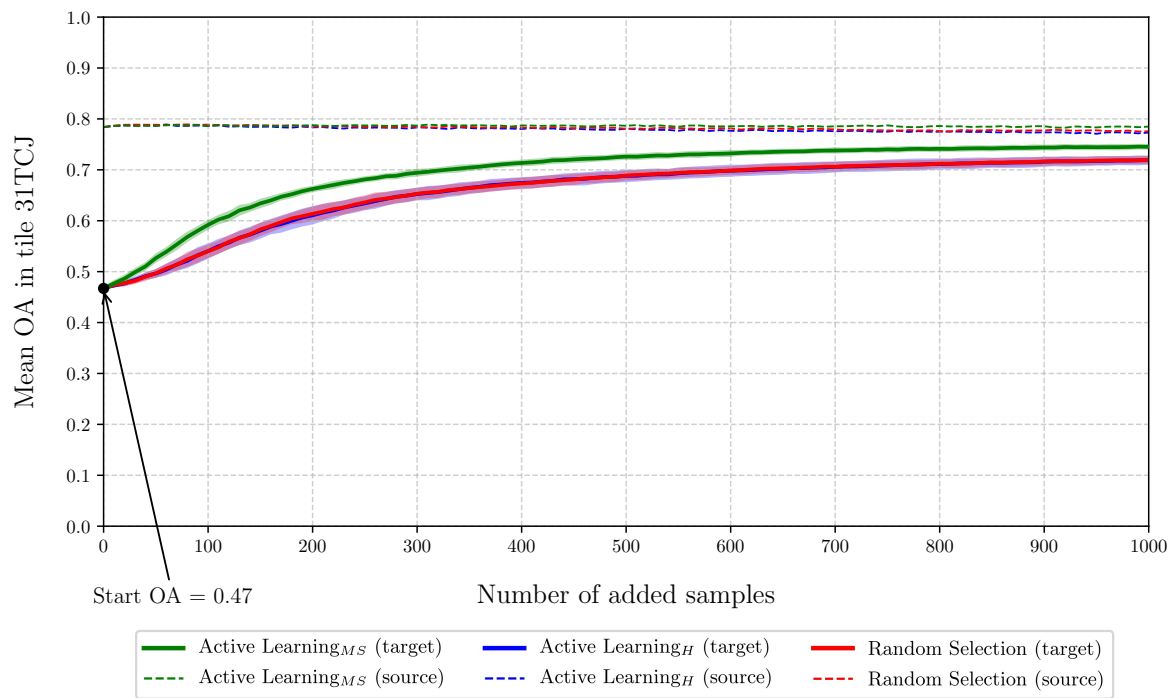
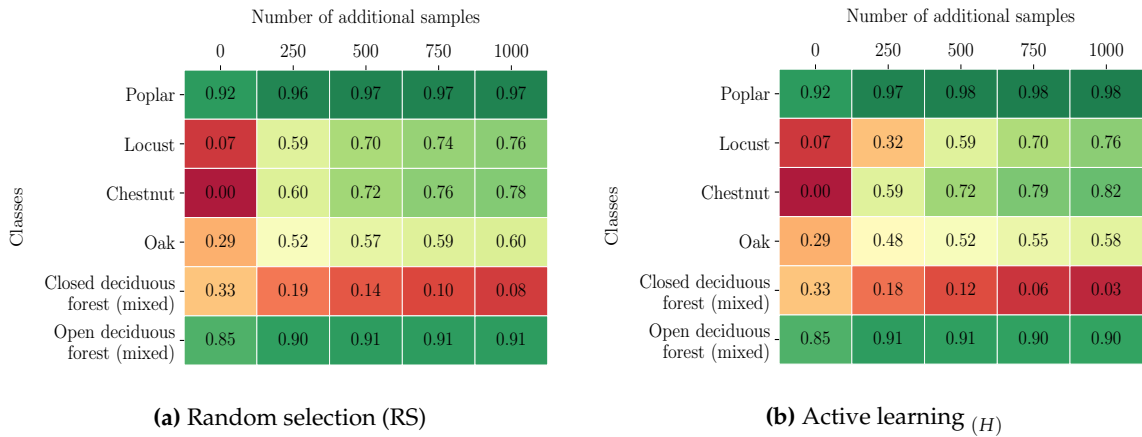


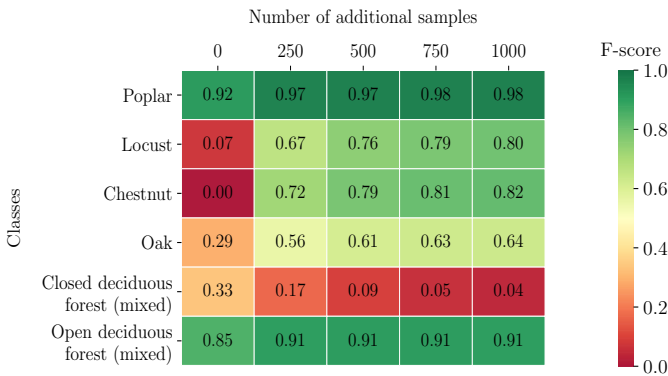
Figure 5. Changes in the average OA score (\pm standard deviation) on the south-western target tile (31TCJ) according to additional number of training samples from the initial model (north-eastern source tile 31UEQ). The average values were based on 30 independent runs for both random sampling (in red) and active learning using entropy (AL_H in blue) or margin sampling (AL_{MS} in green). Dashed lines show changes in the OA on the north-eastern source tile (31UEQ) with RS (red dashed line), AL_H (blue dashed line) and AL_{MS} (green dashed line).

4.2.2. Active learning *vs.* random sampling: class-specific assessment

In order to better understand what underlies the OA measurements, the F-score values by class were analysed in relation to the total number of additional samples. For the sake of simplicity and clarity, results are reported here for only one transfer direction.

Considering the transfer from the north-eastern (31UEQ) tile to the south-western (31TCJ) tile with no extra training samples, the F-score obtained for poplar plantations was 92% (Figure 6a-c). This score increased with the addition of new samples to reach a maximum at 97% and 98% respectively, for random and both active learning models (AL_{MS} and AL_H).





(c) Active learning (AL_{MS})

Figure 6. Average F-scores per class obtained over 30 independent repetitions as a function of the number of added samples with (a) random sampling, (b) entropy-based active learning (AL_H) and (c) margin sampling-based active learning (AL_{MS}). The transfer was performed from the north-eastern tile (31UEQ) to the southwestern tile (31TCJ).

Regarding the class of locust tree species, a low F-score of 7% was noted prior to the addition of target training samples. However, this value rose sharply to achieve 59%, 32% and 67% with 250 extra samples for respectively RS, AL_H and AL_{MS} models (Figure 6a-c). In the same way, the initial F-score for chestnut tree species was null in the target domain because of the absence of this class in the source domain. However, after few iterations, the accuracy jumped to around 60% for both random and AL_H models (250 extra samples). Like in the previous case, the rate of increase with the AL_{MS} model was faster and reached an F-score value of 72% at the 25th iteration. In these last two examples (locust and chestnut), RS required up to 50% more samples to achieve the same class F-scores as AL_{MS} . The same behaviour was observed for oak and open deciduous forest classes with an upward trend in the F-score when training samples were added. The improvement was particularly pronounced when the initial value was low. This was true except for the closed deciduous forest class (with mixed species) for which the F-score dropped with the addition of new samples whatever the learning approach (RS vs. AL) used.

4.3. Maps of poplar plantations

The three S2 tiles were predicted according to the different approaches: local models, non-adapted models (before AL) and finally with a global model offering the best performances (31UEQ+31TCJ+30TYT). The results are shown in Figure 7 with extracts from the three study tiles. Although visual inspection of the images is not obvious, a more accurate result was obtained using a global model with fewer noisy classification patterns and less over-detection of poplars.

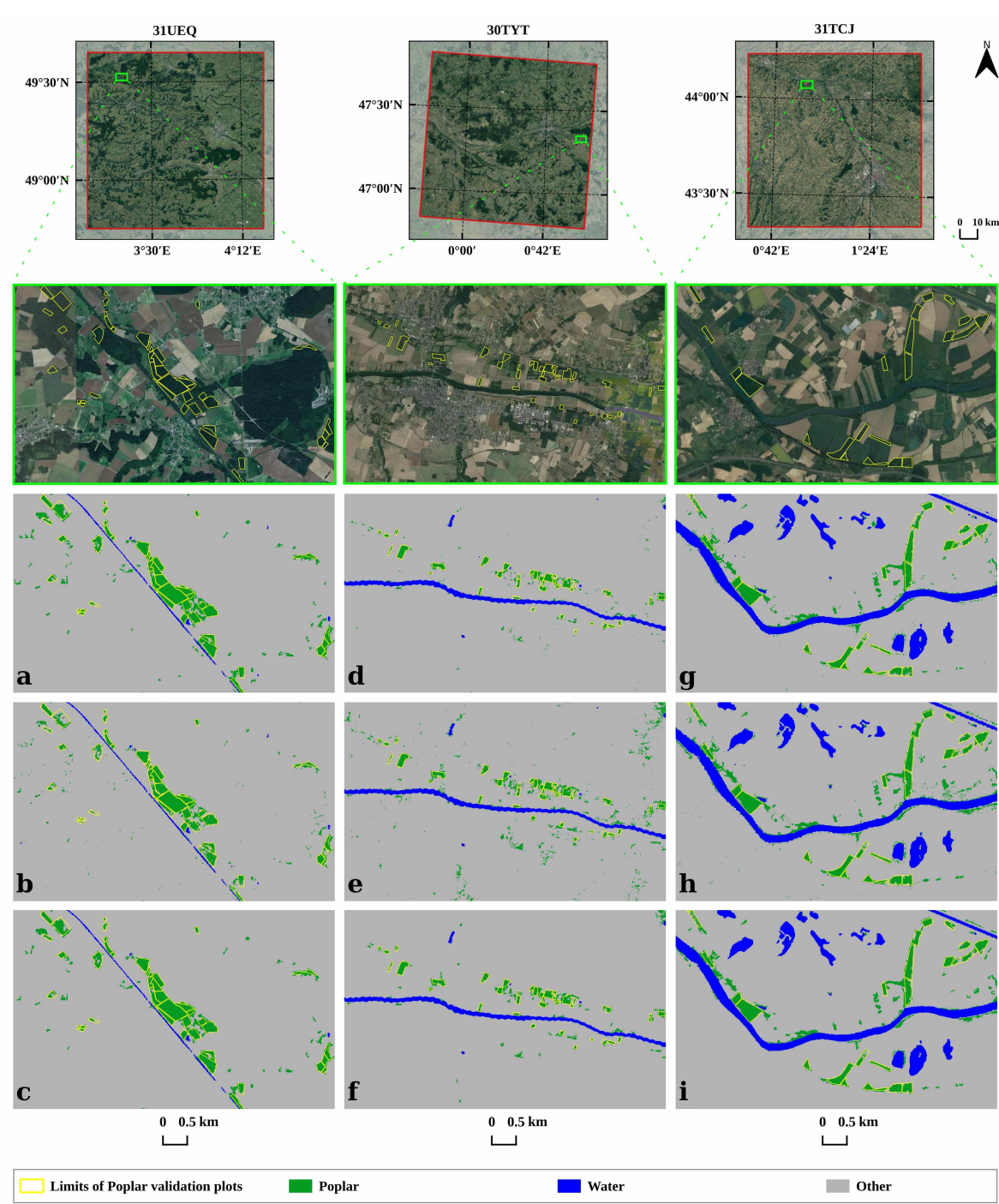


Figure 7. Zoomed prediction maps of the north-eastern, central and south-western tiles according to local models (respectively **a**, **d** and **g**), non-adapted (before AL) models (respectively **b**, **e** and **h**) and the global model (respectively **c**, **f** and **i**).

5. Discussion

5.1. Active learning: an efficient way to generalise across space with limited extra target samples

Our results revealed low classification performances with no adaptation of the initial model due to the non-stationarity of class distributions between the different study tiles. After querying samples from the target domain, the model improved its generalisation capabilities across space. This was true both for random sampling and active learning. However, compared to RS, the number of added target

samples was much lower with AL (and consequently, the cost of sample acquisition) while reaching the same level of accuracy. In some cases, the number of extra samples was twofold higher with RS (500 vs. 250 by AL_{MS} to obtain an average Chesnut F-score of 72% or 1,000 vs. 500 by AL_{MS} to obtain an average Locust F-score of 76%). For a same number of extra samples, AL_{MS} outperformed RS from 1% to 12% of F-score with respect to the transfer direction and the class considered. Furthermore, when the global models were based on all three tiles, overall accuracy was up to 14% higher than that achieved with the local models.

These results which are in line with those of [39], show that AL is a good way to minimise the number of samples for domain adaptation while maintaining high discrimination capabilities. The sample selection bias is corrected and the global model which is adapted for large scale mapping, is able to match the distributions of the local tiles in both the training and test set. However, to benefit from AL, the uncertainty measure must not be arbitrarily chosen. AL_H produced limited performances, with results that were very close to those obtained with RS. The entropy metric is highly dependent on unlikely classes making it less robust, as reported previously [57,58]. Although entropy is an intuitive indicator of uncertainty, the margin sampling (MS) metric is a more reliable and direct way to estimate uncertainty among the most confused classes [58].

5.2. Effect of transfer direction: case of the poplar class

Our results showed asymmetrical transferability of the models from one tile to another. This means that the classification accuracy may vary according to the tile used to initiate the learning process. When we considered the transfer from the north-eastern (31UEQ) to the south-western (31TCJ) tile, a high F-score value for poplar plantations was achieved without any additional target samples. The original model was already able to correctly identify (at 92% F-score) south-western poplars with the knowledge acquired from the north-eastern poplars (i.e. no adaptation was required). Moreover, the improvement in the F-score was almost the same between AL_{MS} and the random-based sampling model after the addition of new samples. By contrast, the number of poplar samples selected over the total number of additional samples revealed a marked difference. For an equivalent F-score value, the number of samples randomly selected (see red bars in Figure 8a) was about eight times higher than those selected by AL_{MS} (green bars in Figure 8a) indicating that extra training samples are highly redundant with random sampling. Here, AL minimises the need to hand-label target poplar samples without sacrificing classification performance. As highlighted in [59], when the class accuracy is high, the active learner avoids querying irrelevant samples.

When the transfer across space was performed in the opposite direction (i.e. from the south-western 31TCJ to the north-eastern 31UEQ tile), the result was slightly different. With no additional samples from the target, the poplar F-score was 76% (Figure 8b). It increased to achieve a stable value around 90% with 50 and 100 extra poplar samples (out of 400 additional samples) for AL_{MS} and RS models, respectively. In the first iterations, AL_{MS} queried more poplar samples than in the opposite direction. However, the number was still lower than with random sampling.

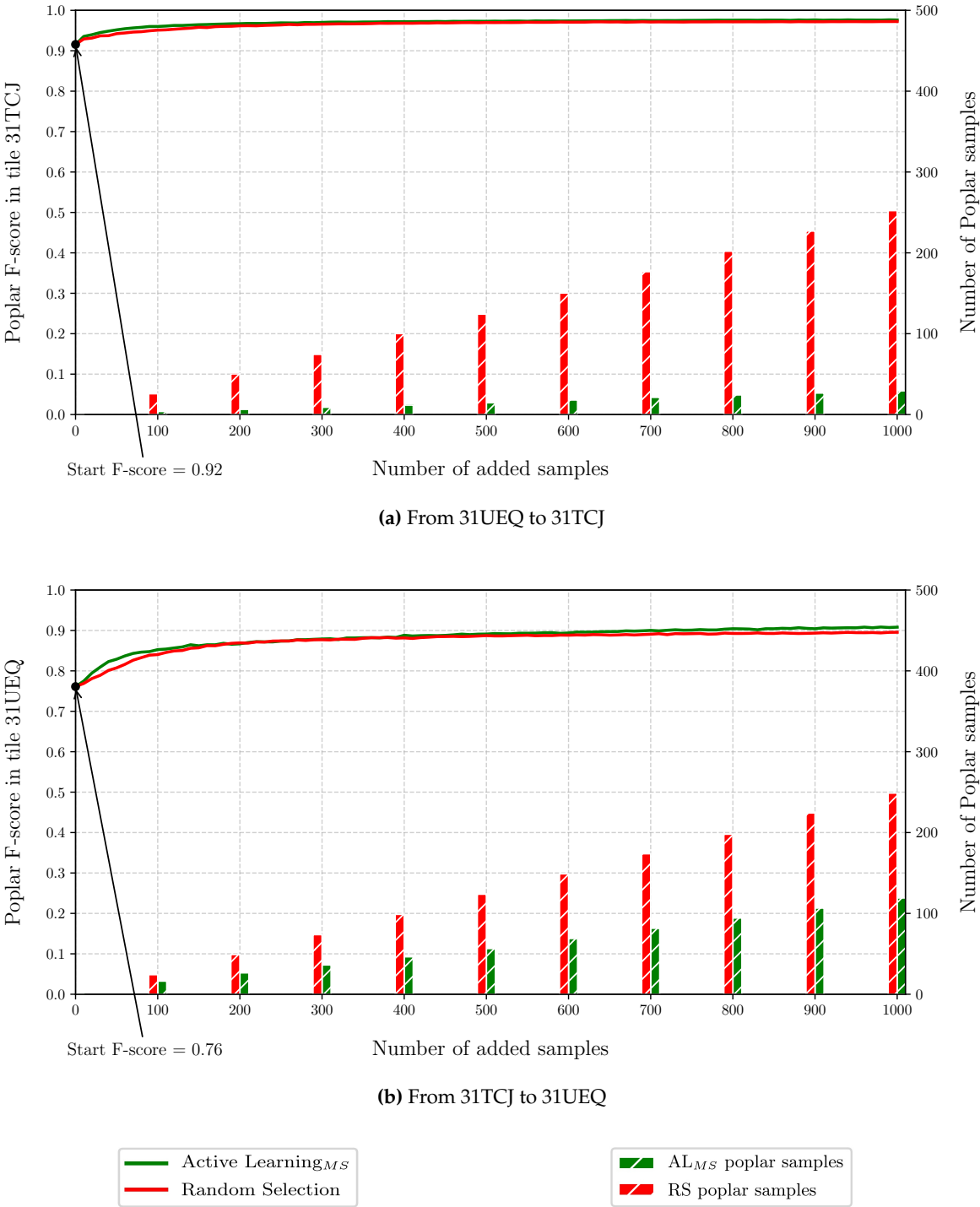


Figure 8. Changes in the average poplar F-score between the north-eastern (31UEQ) and south-western (31TCJ) tiles according to the number of added samples. In (a), the change in the average poplar F-score is given on the south-western target tile (31TCJ) with an initial classifier trained on the north-eastern source tile (31UEQ). In (b), the change in the average poplar F-score is given on the north-eastern tile based on the south-western model. Red (random sampling) and green (active learning-based sampling) bars show the number of target poplar samples selected (right y-axis) from the total number of additional training samples out of a total of 1,000 queried samples.

The south-western model struggled a little with the classification of north-eastern poplars. Plantations were mainly confused with closed deciduous forests (see confusion matrix in Appendix A3a). In this direction, the local environmental conditions of the north-eastern tile may have affected

the transferability of the original model and in particular, soil conditions together with related silvicultural practices. Indeed, among the essential requirements for the successful development of poplar plantations, is a well-drained soil with an easy access to water [60,61]. To satisfy this condition, tillage practices are applied in the driest regions (i.e. south tiles) to minimise competition for water from other plants (shrubs and herbaceous species) resulting in limited ground cover by the understory vegetation. In the north of France, water is not a limiting factor and most poplar plantations present a grassy or even shrubby cover. Thus, the predictors cover a wider range of variability in the north-eastern tile which makes the classification less accurate without a minimal adaptation (due to the *sample selection bias* and *covariate shift* as mentioned by [62]) leading to different but related source and target domains between tiles [21]. Stands with dense understory vegetation were almost invisible in the original south-western model and were therefore confused with the closed deciduous forest class in the north. This was confirmed by examining the nature of samples added in the first iterations of AL.

As illustrated in Figure 9, the NDVI distribution of the first 50 poplar samples added by AL_{MS} coincides with the uncertainty regions of the south-western model, when the transfer was carried out from the south-west to the north-east.

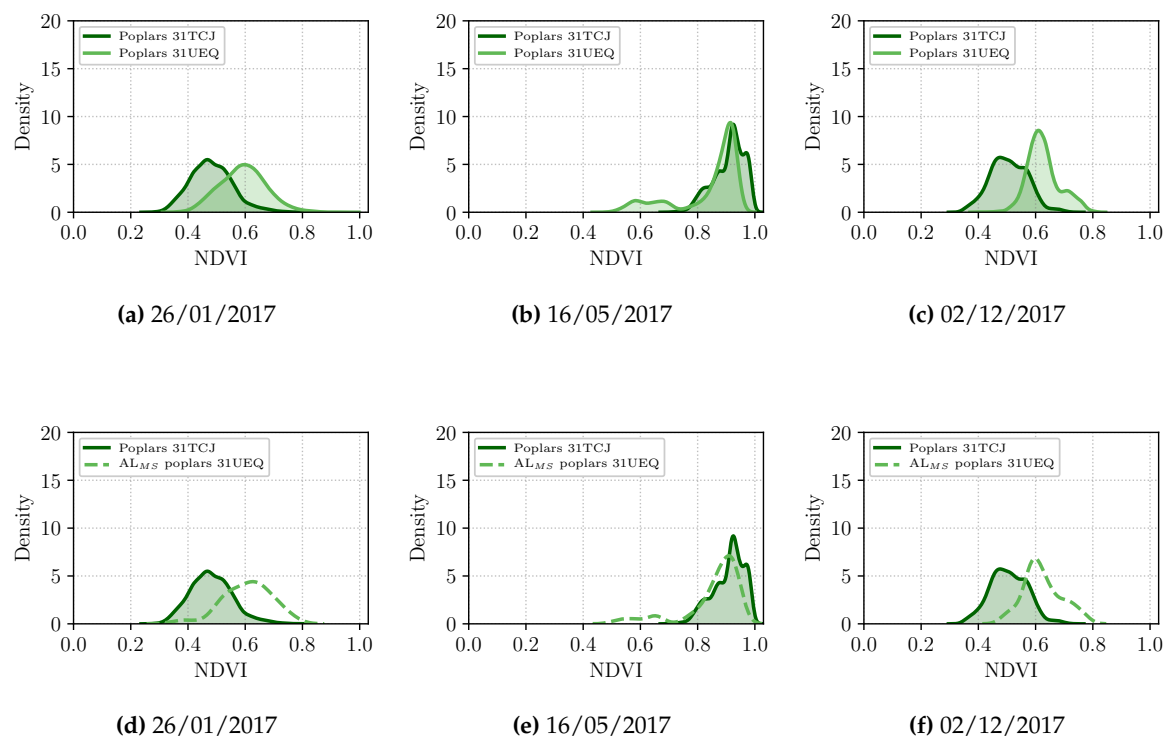


Figure 9. Kernel density of NDVI for poplars in the south-western tile (31TCJ, dark green solid line) and the north-eastern tile (31UEQ, light green solid line) for January 26th (a) and (d)), May 16th (b) and (e)) and December 2nd (c) and (f)), 2017. In d, e) and f) we added the NDVI distribution of the first 50 poplar samples selected by AL_{MS} (light green dashed line) from the north-eastern target tile when the transfer was conducted from the south-western tile.

The asymmetry of the transfer direction began to stabilise with three tiles especially when the north-eastern tile was no longer the target (Table 5). This could be attributed to the fact that all the classes are represented in the source tiles (i.e. with the north-east and south-west or the north-east and center models) unlike the case when the north-eastern tile is predicted (the beech class does not exist in the original models).

Hence, it might be better to start with a two-tile model including the north-east and then predict either the south-western or central tile in order to ensure the best classification performance in all three tiles. Following the three-tile model, all the class diversity and distribution should already be

accounted for (the covariate shift should disappear), the stability level should increase and equivalent performance would be expected regardless of the order of the tiles.

5.3. Impact of missing classes in the original model: case of the chestnut class

In the case of a transfer from the north-east to the south-west, the original model does not include the chestnut class which actually exists in the target tile (Table 2). When the source model was applied directly (before adding samples), the chestnut F-score was not surprisingly equal to zero since the chestnut class was initially unknown. All its samples were confused with the pre-existing classes but mainly with closed deciduous forest (51%) and oak (40%) (Appendix A1b). Both AL_{MS} and random models drew chestnut samples from the entire selected pool over the iterations but many more with the former. In fact, chestnut samples in AL_{MS} were considered to be the most uncertain and were therefore and primarily the most frequently chosen. As discussed in [24], the active learning followed an intuitive approach and immediately moved towards this unknown class supposedly difficult to classify. In either case, the F-score increased as chestnut samples were added from the target but considerably faster with AL_{MS} which promptly improved the class definition in the initial source model that successfully identified it (Figure 10). This behavior of AL was also observed for other missing classes with respect to the transfer direction. These results confirmed that the AL tends to select many samples from the hardest classes to discriminate among the others, as demonstrated in [58,59]. The good performance of random sampling can easily be explained by the high probability of the random process finding new classes only by chance [30]. From a learning point of view, both domains and learning tasks differ in this case: distribution values of features change because of spatial non-stationarity; tasks also vary because the label spaces between domains are not exactly the same [21].

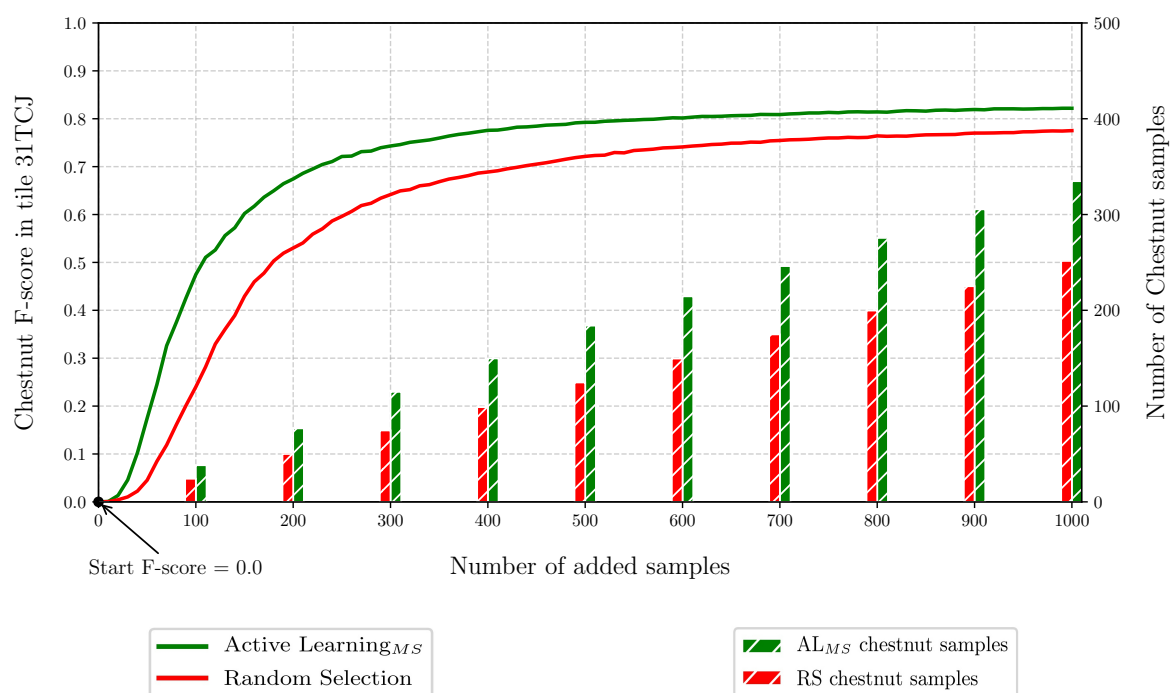


Figure 10. Changes in the average chestnut F-score on the south-western target tile (31TCJ) according to the number of added samples. The initial classifier was trained on the north-eastern source tile (31UEQ).

5.4. Effect of training with mixed classes: case of the locust class

In the experiments, all the models were trained using pure pixels of tree species only (Section 3.3.2). To check the functioning of AL, pixels of closed and open deciduous forests (mixed classes) were used for testing but not for training. In this way, we made AL more efficient, as explained below.

In the French forest reference database used as a source of samples, forest stands are defined as mixed species when the dominant species cover less than 75% of the total extent. Inside these stands, species can be spatially aggregated or scattered. Therefore, stands of mixed species can be composed of mixels (containing spectral information of several species in one pixel) or can be a mixture of pure pixels. If mixed-species classes are included in the training set, two issues may arise: (i) pure-species pixels may be confused with pixels labelled as mixed species since these pixels can be also pure at this level (the label making sense at the stand level); (ii) predictions of mixed-species classes on another S2 tile may be inefficient because of sample selection bias (the composition of mixed-species classes can vary across space from one tile to another) but this can be corrected by AL. We observed the first issue (confusion between pure and mixed-species classes) for several species. We illustrate it below with the locust species class.

At the beginning of the transfer from the north-eastern source tile (31UEQ) to the south-western target tile (31TCJ), the locust F-score was very low with no additional samples (7%, Table 6). This poor performance was surprising because pure pixels of locust species were present in both source and target tiles. Due to spectral and temporal overlaps (Appendix A4), there was considerable confusion with oak and closed deciduous forest classes (Appendix B.1).

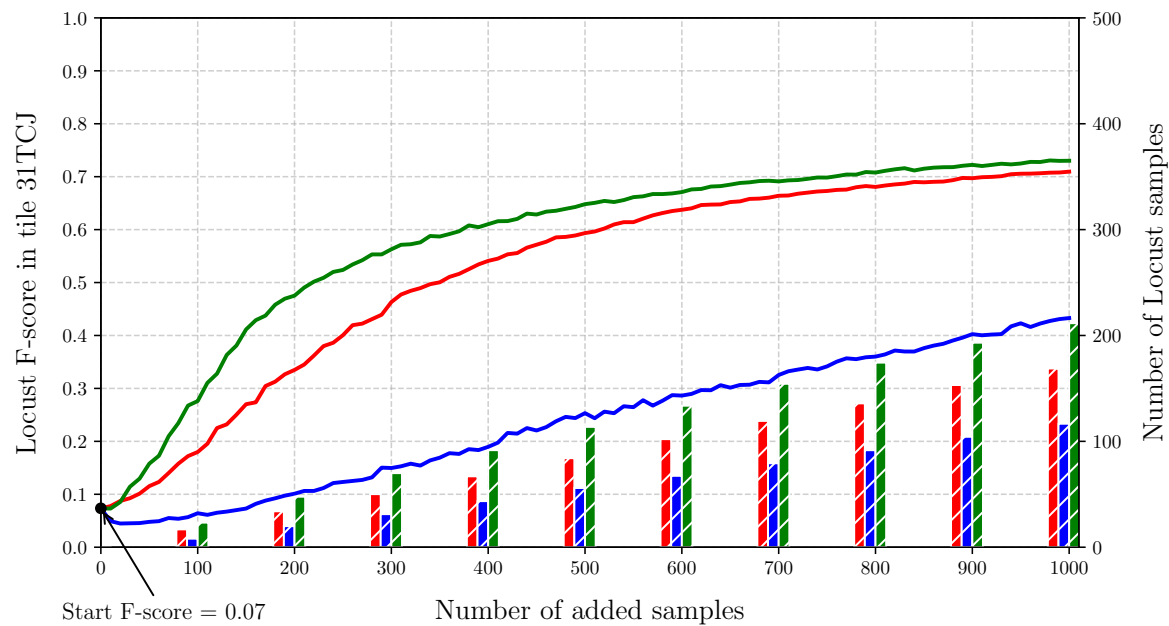
Considering mixed-species classes for training, the addition of target extra samples by AL led to a rapid increase in the locust F-score, especially by AL_{MS} (Table 6 and Figure 11a) but the performance boost was hampered by the closed deciduous forest class. Despite the low initial locust F-score, AL did not only select locust samples. Many uncertain samples belonging to the closed deciduous forest class were also queried (Figure 11b). After checking, some of the uncertain pixels also turned out to be locust pixels (the other ones being pixels of oak). Thus, AL seeks to improve the classification of locust in both cases but pixels can be found in two different classes. The existence of the mixed-species class makes AL less competitive (see Figures 11a and 11c for comparison). In addition, by selecting locust pixels labelled as closed deciduous forest in mixed stands, the model is progressively distorted. The negative effect of the mixed-species class is particularly pronounced with entropy-based AL. Because this metric takes the probabilities of belonging to all the classes into account (not only the two most probable classes, as in AL_{MS} which are locust and closed deciduous forest or oak), the model is thus influenced by low probabilities of unimportant classes which are rarely the right ones [58]. In other words, AL_H is less robust to class overlap. Regarding the model based on random sampling, more locust samples are selected than by AL_H which further improves the performance for this class. The same behaviour was encountered in [30], where random sampling outperformed active learning in a similar context with a highly mixed crop classes.

Table 6. Average locust F-score on the south-western tile according to the total number of additional samples with and without the presence of mixed classes in the training set.

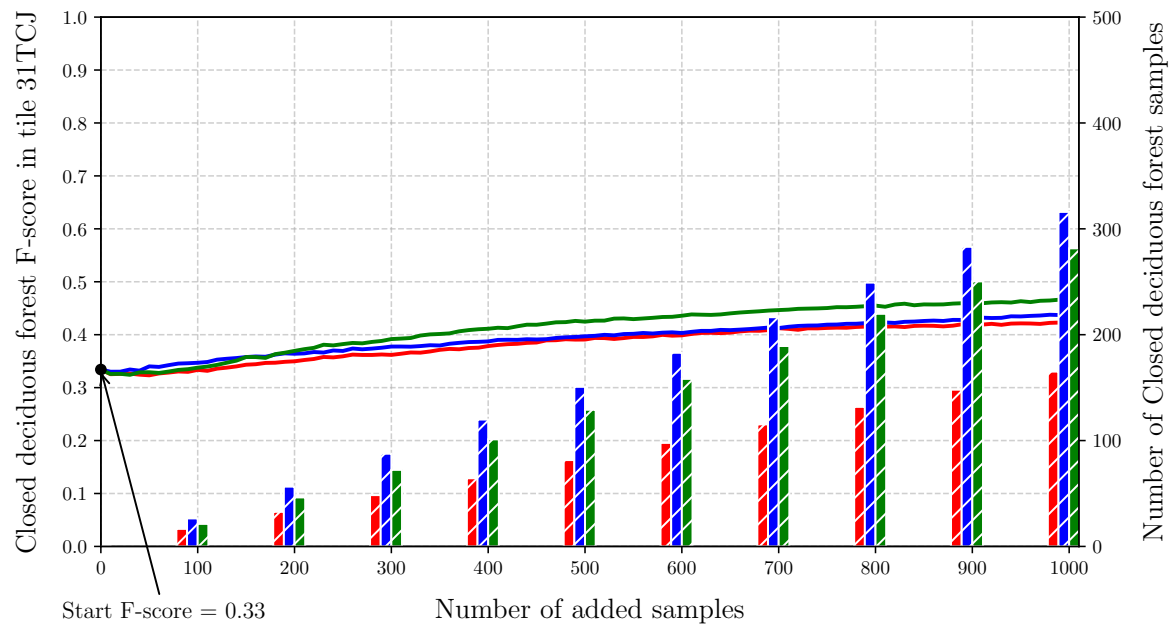
Average locust F-score (%) _(*30)	Number of additional samples (total)									
	With mixed classes					Without mixed classes				
	0	250	500	750	1000	0	250	500	750	1000
Random Selection	7	40	59	67	71	7	59	70	74	76
Active Learning _(H)	7	12	25	34	43	7	32	59	70	76
Active Learning _(MS)	7	52	65	70	73	7	67	76	79	80

With pixels of pure-species class only, model performance improved, especially with few extra samples (Table 6 and Figures 11c and 11d). The model not only improves the definition of pure classes

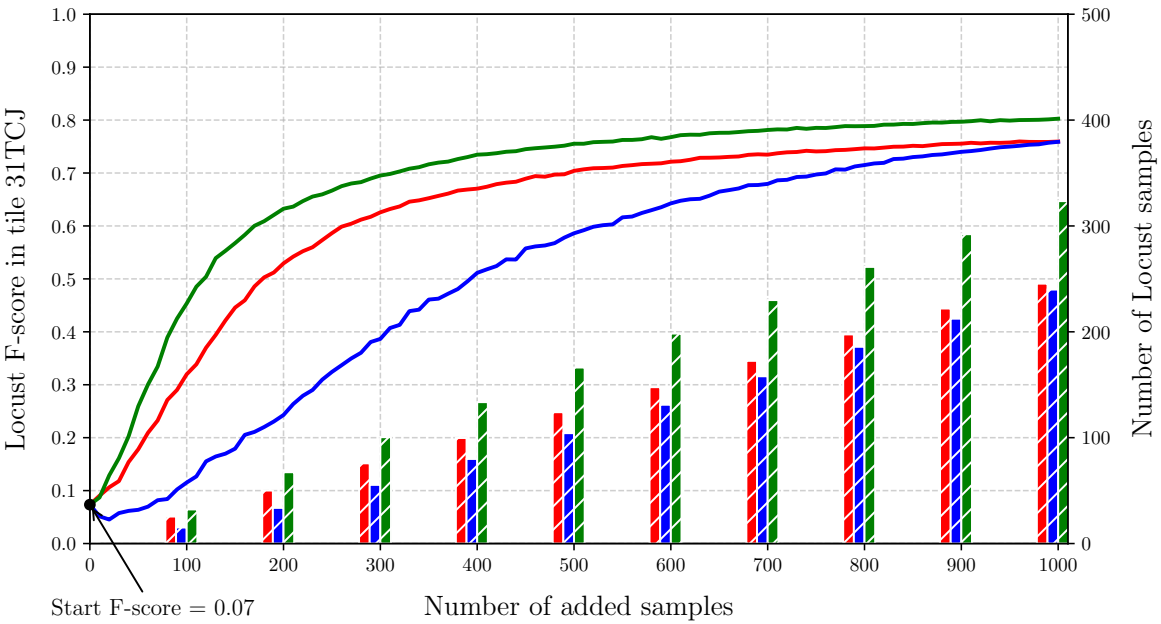
but also allows a better characterisation of the species in the mixed-species stands, opening interesting perspectives to enrich the forest reference map.



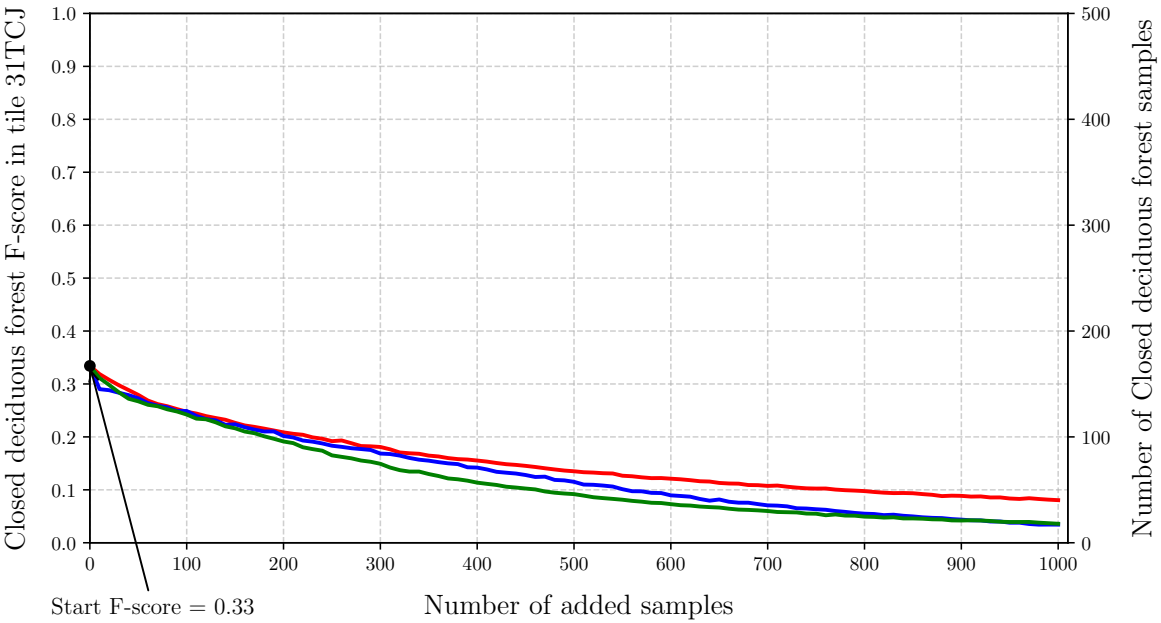
(a) Training with mixed classes; Locust class



(b) Training with mixed classes; Closed deciduous forest class



(c) Training with mixed classes; Locust class



(d) Training with mixed classes; Closed deciduous forest class



Figure 11. Changes in the average F-score on the south-western target tile (31TCJ) for locust (a and c) and closed deciduous forest (b and d) classes according to the number of additional samples. The initial classifier was trained on the north-eastern source tile (31UEQ). In (a) and (b), the training set contained pure and mixed species while in (c) and (d) only pure classes were considered (explaining why no extra pixels of closed deciduous forest were added in (d)).

6. Conclusions

In this paper, we propose the use of an active learning approach for the classification of poplar plantations, among other tree species, in a large-scale context. The results show the good capability of Sentinel-2 time series to map poplar plantations with all their diversity related to cultivars, management practices and climatic conditions. In particular, we demonstrate the potential of active learning to adapt a local model into a global model suitable for large-scale mapping. By adaptability, we have not only shown its efficiency to query relevant samples from the unexplored areas, but also its capacity to discover new classes. Following the analysis of the different transfer combinations by AL between the three study S2 tiles, we conclude that spatial transferability is strongly dependent on the transfer direction but the performance gap tends to decrease with three-tile-based models. The resulting global models achieved high classification performance with few training samples reflecting the potential of AL to considerably reduce the cost of labelling. With respect to untested uncertainty metrics, margin sampling provided the best classification results and proved to be more robust to class overlap than entropy. We therefore recommend the use of MS metric for multi-class tasks, especially when dealing with mixed classes.

It is important to note that in real-world applications, it would be interesting to use AL to guide field surveys or photo interpretation operators in collecting labels for the most useful pixels rather than providing a pool of samples from which to choose the most useful ones [36,37]. Another direction for future work is to consider other transfer learning methods for large scale image classification. It could be multitask learning, in which the target and source domains are learned simultaneously. However, such a framework assumes that a lot of training samples are widely available [21].

7. Patents

Author Contributions: Conceptualization, Y.H. and D.S.; Methodology, Y.H. and D.S.; Investigation, Y.H.; Software, Y.H.; Visualization, Y.H.; Validation, Y.H., E.P., V.C., C.M. and D.S.; Funding acquisition, E.P. and D.S.; Supervision, E.P., V.C., C.M. and D.S.; Writing—original draft preparation, Y.H. and D.S.; Writing—review and editing, E.P., V.C. and C.M.

Funding: Y.H. received a PhD scholarship from the French Ministry of Higher Education and Research (ANRT/CIFRE grant number 2017/0228). The project is spearheaded by the French National Poplar Council (CNP) and was supported by public funding from the French Ministry of Food and Agriculture (grant number BOP 149-26-12), the regions of Nouvelle Aquitaine and Grand Est and the County Council of Lot-et-Garonne. The project involved as well private funders, namely the Codifab (Professional Committee for the Development of French Furniture Industries), France Bois Forêt, Alliance Forêts Bois and the company Garnica Plywood. The study has also received financial support from the French Space Agency CNES, as part of TOSCA Parcelle project.

Acknowledgments: We thank Johann Hübelé and Nicolas Vanderheeren for their technical expertise in poplar cultivation and for their field assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Local classification results without prior photo interpretation of poplar samples

When all reference polygons were retrieved from the forest database, including poplar samples without being photo-interpreted, a loss of about 17%, 13% and 9% of the poplar F-score was observed for the north-eastern, central and south-western tiles respectively. The impact was more noticeable on the north-eastern tile and could be related to the acquisition year of the aerial photographs used to generate the forest database (Figure 3). They are dated from 2004 to 2013 for the north-eastern districts, from 2008 to 2011 for the central ones and from 2012 to 2013 for the south-western districts.

Table A1. Local classification results in two Sentinel-2 tiles without photo interpretation of poplars.

Tile code	Training size ¹ per class in pixels	No. classes	Overall Accuracy _(*30)	Global F-score _(*30)	Poplar F-score _(*30)	Poplar Precision _(*30)	Poplar Recall _(*30)
31UEQ	1250	6	65.6±6.5 %	65.2±6.9 %	72.6±5.7 %	69.5±8.2 %	77.3±3.9 %
30TYT	2000	6	65.8±2.2 %	70.6±2.2 %	86.7±1.7 %	85.6±2.8 %	87.9±2.4 %
31TCJ	3850	6	79.5±3.3 %	79.4±3.7 %	89.1±3.9 %	88.6±5.2 %	90.8±3.2 %

¹ Training samples represent 50% of the available reference data.

The poplars derived from the forest database presented various errors corresponding to logged or abandoned plantations as well as grasslands. The results were therefore statistically and cartographically unsatisfactory. This led us to choose systematic photo interpretation of all poplar samples in order to ensure the best classification results.

Appendix B. Results before Active Learning adaptation

Appendix B.1. Source tile: north-eastern (31UEQ)

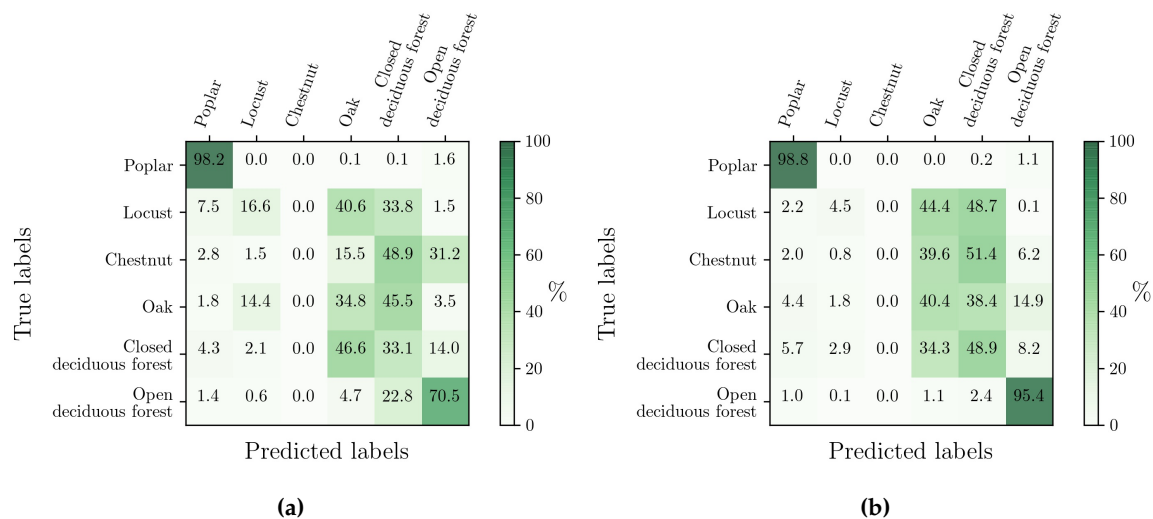


Figure A1. Normalised confusion matrices obtained before active learning adaptation (before adding new samples from target) when the transfer was performed from the north-eastern source tile (31UEQ) to (a) the central (30TYT) and (b) the south-western (31TCJ) target tiles.

Appendix B.2. Source tile: central (30TYT)

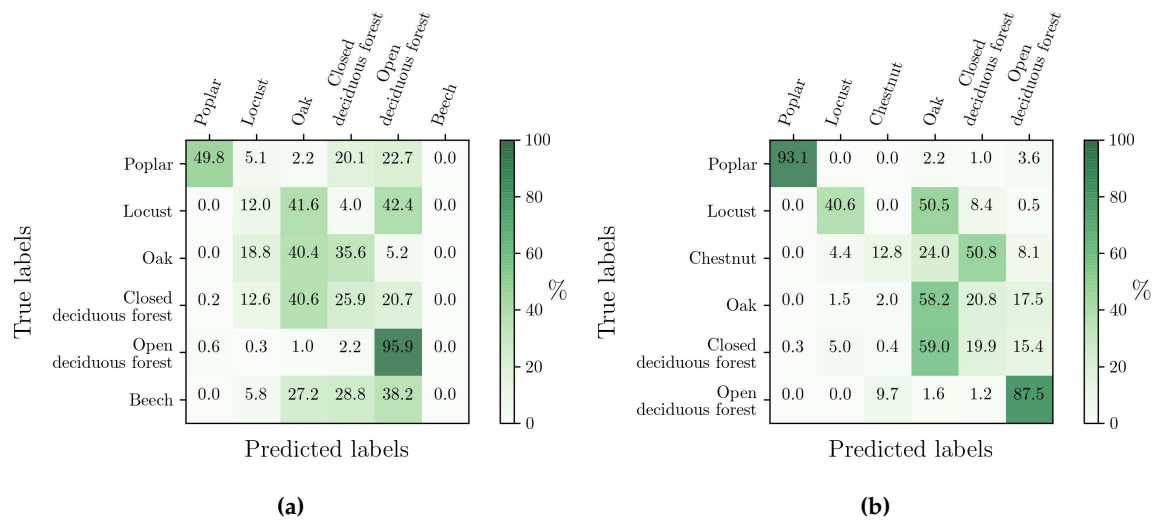


Figure A2. Normalised confusion matrices obtained before active learning adaptation (before adding new samples from target) when the transfer was performed from the central source tile (30TYT) to (a) the north-eastern (31UEQ) and (b) the south-western (31TCJ) target tiles.

Appendix B.3. Source tile: south-western (31TCJ)

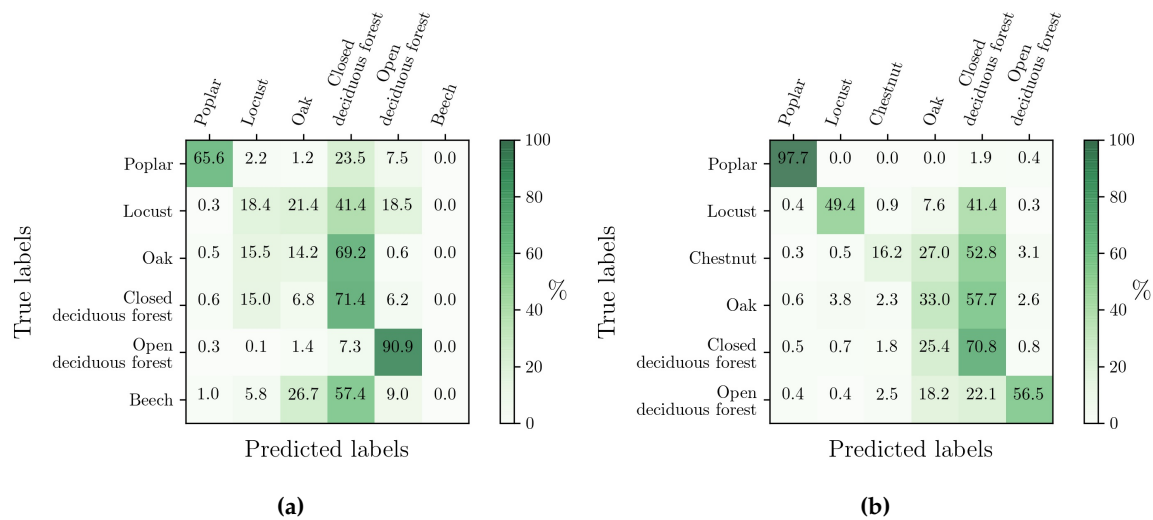


Figure A3. Normalised confusion matrices obtained before active learning adaptation (before adding new samples from target) when the transfer was performed from the south-western source tile (31TCJ) to (a) the north-eastern (31UEQ) and (b) the central (30TYT) target tiles.

Appendix C. Distribution of NDVI values within the deciduous classes of each tile

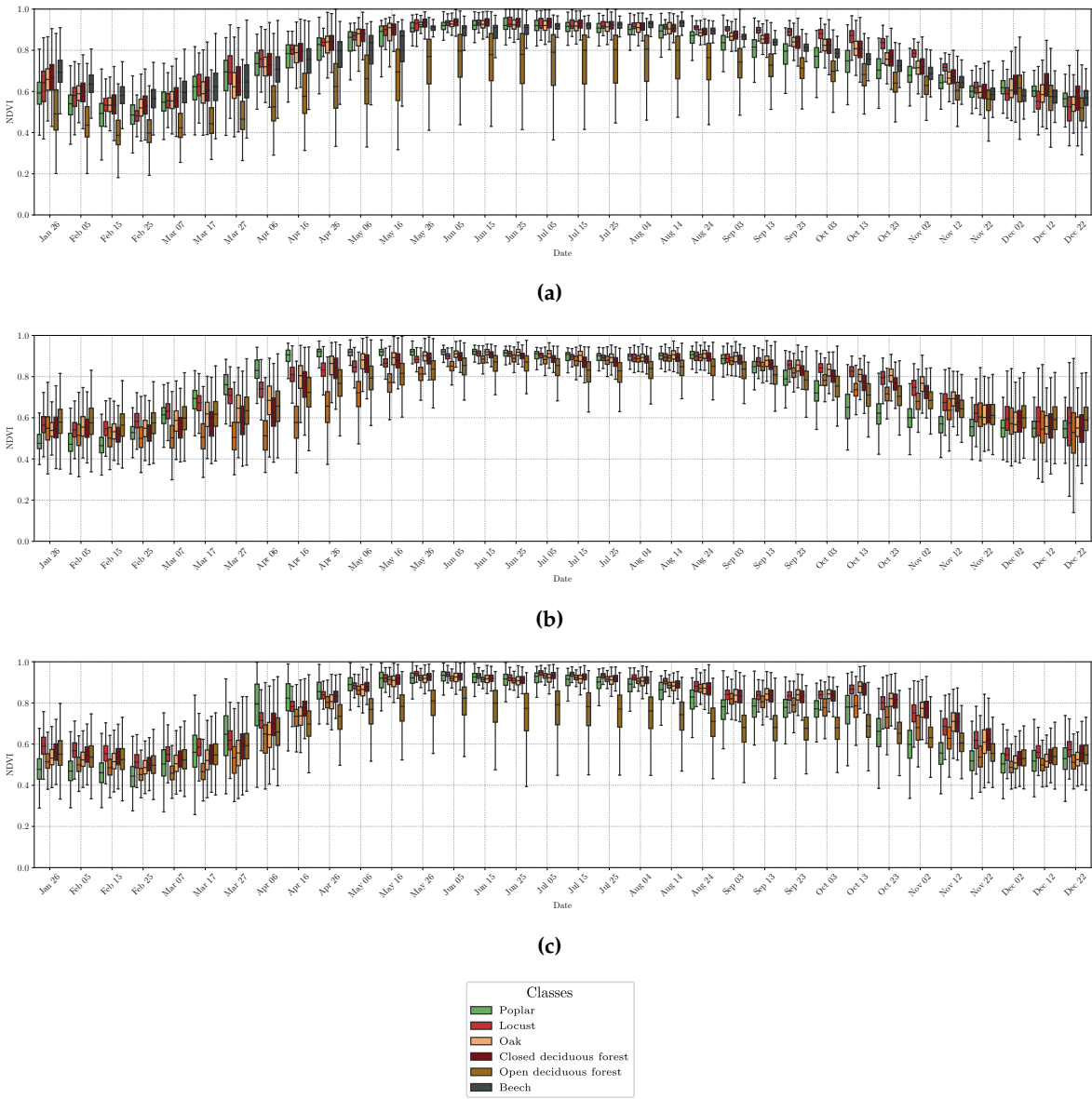


Figure A4. Spectral variability of Sentinel-2 NDVI over the deciduous classes in (a) the north-eastern (31UEQ), (b) the central (30TYT) and (c) the south-western (31TCJ) tiles.

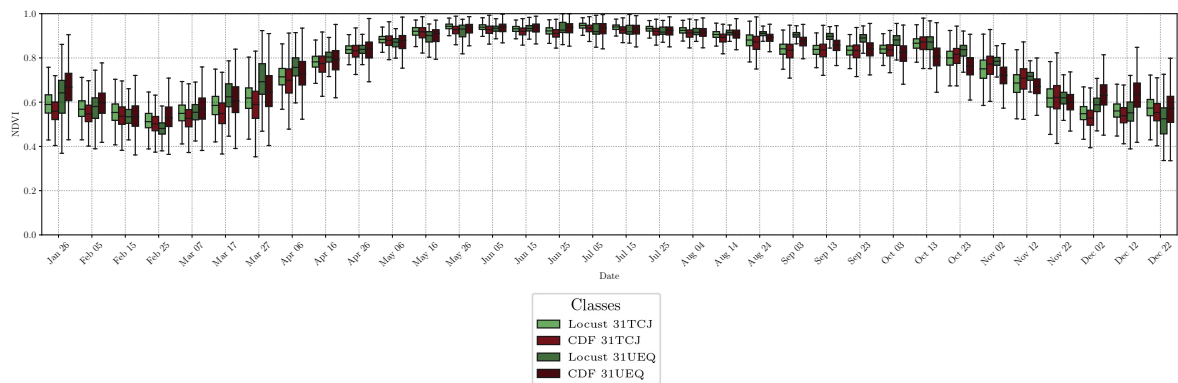


Figure A5. Spectral variability of Sentinel-2 NDVI between locusts and closed deciduous forests (CDF) on the north-eastern (N.E) and south-western (S.W) tiles.

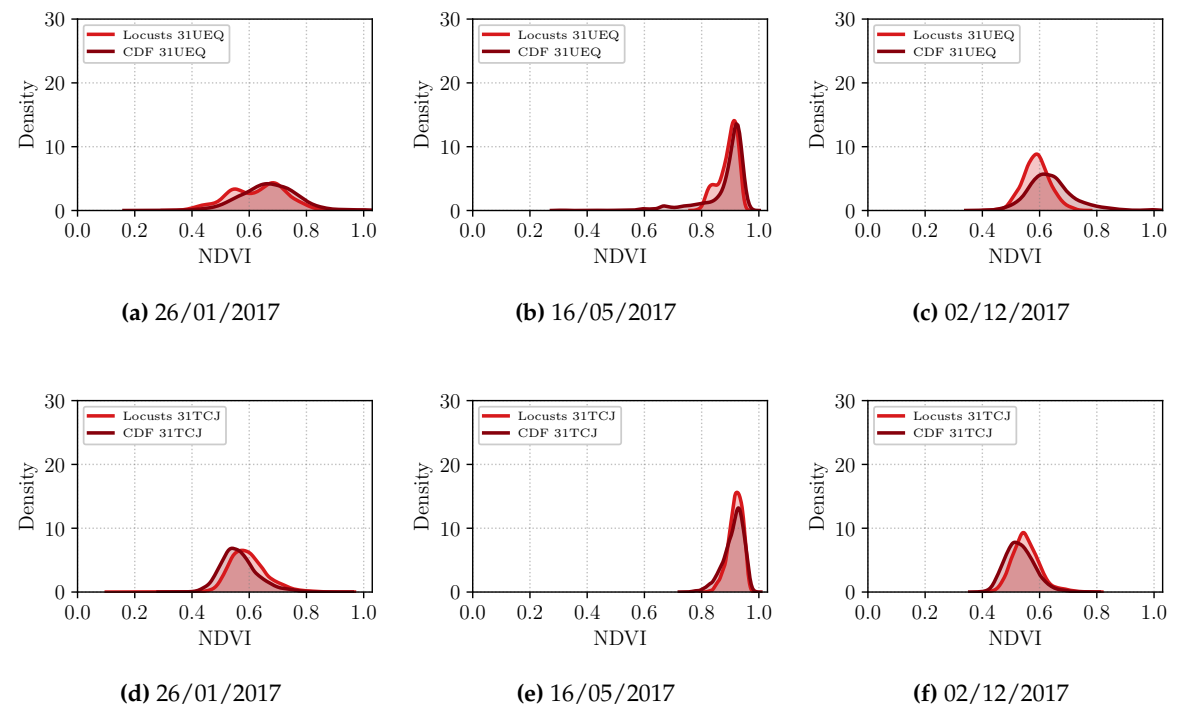


Figure A6. kernel density of NDVI for locusts (red line) and closed deciduous forest (CDF) (burgundy line) in the north-eastern (31UEQ) and south-western (31TCJ) tiles for January 26th (a) and d)), May 16th (b) and e)) and December 2nd (c) and f)).

Appendix D. Active learning results with a combination of uncertainty and diversity measures

Table A2. Overall accuracies (in %) according to the direction of the transfer and the number of target samples added to the initial set for active learning with margin sampling (MS) uncertainty metric combined with Euclidean distance (AL_{MS-Eu}) and Cosine (AL_{MS-C}) diversity metrics. The values are averaged over 30 independent repetitions.

		Target tiles														
		31UEQ					30TYT					31TCJ				
Source tiles	OA (%)	No. of additional samples														
		0	250	500	750	1000	0	250	500	750	1000	0	250	500	750	1000
31UEQ	AL_{MS-Eu}	-	-	-	-	-	36	53	59	62	63	47	62	69	71	73
	AL_{MS-C}	-	-	-	-	-	36	50	57	60	62	47	55	64	68	70
30TYT	AL_{MS-Eu}	31	47	53	56	57	-	-	-	-	-	53	66	69	70	71
	AL_{MS-C}	31	46	52	55	57	-	-	-	-	-	53	64	67	69	70
31TCJ	AL_{MS-Eu}	40	53	61	63	64	52	62	64	65	65	-	-	-	-	-
	AL_{MS-C}	40	54	60	63	64	52	61	64	65	65	-	-	-	-	-

References

1. FAO. Poplars and Other Fast-Growing Trees - Renewable Resources for Future Green Economies. Synthesis of Country Progress Reports, Berlin, Germany, 2016.

2. Robert, A. Poplar Plantations in France, at the Heart of a Conflict between Provisioning Services and Cultural (Dis)Services. Ecosystem Services in a Changing World: Moving from Theory to Practice, San Sebastián, Spain, 2018.

3. Paillassa, É. Les Peupleraies : Quels Enjeux Pour l’avenir de La Populiculture Française ? *Revue Forestière Française* **2014**, 66, 301–311. doi:10.4267/2042/56060.

4. Li, Z.; Fox, J.M. Integrating Mahalanobis Typicalities with a Neural Network for Rubber Distribution Mapping. *Remote Sensing Letters* **2011**, 2, 157–166. doi:10.1080/01431161.2010.505589.

5. Dong, J.; Xiao, X.; Sheldon, S.; Biradar, C.; Xie, G. Mapping Tropical Forests and Rubber Plantations in Complex Landscapes by Integrating PALSAR and MODIS Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **2012**, 74, 20–33. doi:10.1016/j.isprsjprs.2012.07.004.

6. Ye, S.; Rogan, J.; Sangermano, F. Monitoring Rubber Plantation Expansion Using Landsat Data Time Series and a Shapelet-Based Approach. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, 136, 134–143. doi:10.1016/j.isprsjprs.2018.01.002.

7. Han, P.; Chen, J.; Han, Y.; Yi, L.; Zhang, Y.; Jiang, X. Monitoring Rubber Plantation Distribution on Hainan Island Using Landsat OLI Imagery. *International Journal of Remote Sensing* **2018**, 39, 2189–2206. doi:10.1080/01431161.2017.1420933.

8. Xiao, C.; Li, P.; Feng, Z. Monitoring Annual Dynamics of Mature Rubber Plantations in Xishuangbanna during 1987-2018 Using Landsat Time Series Data: A Multiple Normalization Approach. *International Journal of Applied Earth Observation and Geoinformation* **2019**, 77, 30–41. doi:10.1016/j.jag.2018.12.006.

9. Rosenqvist, A. Evaluation of JERS-1, ERS-1 and Almaz SAR Backscatter for Rubber and Oil Palm Stands in West Malaysia. *International Journal of Remote Sensing - INT J REMOTE SENS* **1996**, 17, 3219–3231. doi:10.1080/01431169608949140.

10. Lazecky, M.; Lhota, S.; Penaz, T.; Klushina, D. Application of Sentinel-1 Satellite to Identify Oil Palm Plantations in Balikpapan Bay. *IOP Conference Series: Earth and Environmental Science* **2018**, 169, 012064. doi:10.1088/1755-1315/169/1/012064.

11. Cheng, Y.; Yu, L.; Xu, Y.; Liu, X.; Lu, H.; Cracknell, A.P.; Kanniah, K.; Gong, P. Towards Global Oil Palm Plantation Mapping Using Remote-Sensing Data. *International Journal of Remote Sensing* **2018**, 39, 5891–5906. doi:10.1080/01431161.2018.1492182.

12. Poortinga, A.; Tenneson, K.; Shapiro, A.; Nquyen, Q.; San Aung, K.; Chishtie, F.; Saah, D. Mapping Plantations in Myanmar by Fusing Landsat-8, Sentinel-2 and Sentinel-1 Data along with Systematic Error Quantification. *Remote Sensing* **2019**, 11, 831. doi:10.3390/rs11070831.

13. Chardenon, J.; Flouzat, G. The application of remote sensing to poplar growing: identification and inventory of poplar groves, prediction of timber production; France, Italy. *Revue Forestiere Francaise* **1981**, *33*, 478–493. doi:10.4267/2042/21534.
14. Borry, F.C.; de Roover, B.P.; Leysen, M.M.; de Wulf, R.R.; Goossens, R.E. Evaluation of SPOT and TM Data for Forest Stratification: A Case Study for Small-Size Poplar Stands. *IEEE Transactions on Geoscience and Remote Sensing* **1993**, *31*, 483–490. doi:10.1109/36.214924.
15. Heyman, O.; Gaston, G.; Kimerling, A.; Campbell, J. A Per-Segment Approach to Improving Aspen Mapping from High-Resolution Remote Sensing Imagery. *Journal of Forestry* **2003**, *101*, 29–33. doi:10.1093/jof/101.4.29.
16. Grignetti, A.; Coaloa, D.; Niccolini, G. Classification of poplar stand areas by high-resolution satellite images. *Forest@ - Rivista di Selvicoltura ed Ecologia Forestale* **2009**, *6*, 299–311. doi:10.3832/efor0590-006.
17. Eslami, A.; Zahedi, S.S. Providing Poplar Plantation Map by Indian Remote Sensing (IRS) Satellite Imagery in Northern Iran. *African Journal of Agricultural Research* **2011**, Vol. 6, 4769–4774. doi:10.5897/AJAR11.590.
18. Li, Z.; Fox, J.M. Mapping Rubber Tree Growth in Mainland Southeast Asia Using Time-Series MODIS 250 m NDVI and Statistical Data. *Applied Geography* **2012**, *32*, 420–432. doi:10.1016/j.apgeog.2011.06.018.
19. Gao, T.; Zhu, J.; Zheng, X.; Shang, G.; Huang, L.; Wu, S. Mapping Spatial Distribution of Larch Plantations from Multi-Seasonal Landsat-8 OLI Imagery and Multi-Scale Textures Using Random Forests. *Remote Sensing* **2015**, *7*, 1702–1720. doi:10.3390/rs70201702.
20. Descals, A.; Szantoi, Z.; Meijaard, E.; Sutikno, H.; Rindanata, G.; Wich, S. Oil Palm (*Elaeis Guineensis*) Mapping with Details: Smallholder versus Industrial Plantations and Their Extent in Riau, Sumatra. *Remote Sensing* **2019**, *11*, 2590. doi:10.3390/rs11212590.
21. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345–1359. doi:10.1109/TKDE.2009.191.
22. Lambert, J.; Drenou, C.; Denux, J.P.; Balent, G.; Cheret, V. Monitoring Forest Decline through Remote Sensing Time Series Analysis. *GIScience & Remote Sensing* **2013**, *50*, 437–457. doi:10.1080/15481603.2013.820070.
23. Woodcock, C.E.; Macomber, S.A.; Pax-Lenney, M.; Cohen, W.B. Monitoring Large Areas for Forest Change Using Landsat: Generalization across Space, Time and Landsat Sensors. *Remote Sensing of Environment* **2001**, *78*, 194–203. doi:10.1016/S0034-4257(01)00259-0.
24. Crawford, M.M.; Tuia, D.; Yang, H.L. Active Learning: Any Value for Classification of Remotely Sensed Data? *Proceedings of the IEEE* **2013**, *101*, 593–608. doi:10.1109/JPROC.2012.2231951.
25. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.; Emery, W. Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2009**, *47*, 2218–2232. doi:10.1109/TGRS.2008.2010404.
26. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geoscience and Remote Sensing Magazine* **2016**, *4*, 41–57. doi:10.1109/MGRS.2016.2548504.
27. Gong, B.; Grauman, K.; Sha, F. Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation. *International Conference on Machine Learning*, 2013, pp. 222–230.
28. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2015**, *53*, 3550–3564. doi:10.1109/TGRS.2014.2377785.
29. Petitjean, F.; Ketterlin, A.; Gançarski, P. A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. *Pattern Recogn.* **2011**, *44*, 678–693. doi:10.1016/j.patcog.2010.09.013.
30. Tuia, D.; Pasolli, E.; Emery, W.J. Using Active Learning to Adapt Remote Sensing Image Classifiers. *Remote Sensing of Environment* **2011**, *115*, 2232–2242. doi:10.1016/j.rse.2011.04.022.
31. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing* **2011**, *5*, 606–617. doi:10.1109/JSTSP.2011.2139193.
32. Stumpf, A.; Lachiche, N.; Malet, J.; Kerle, N.; Puissant, A. Active Learning in the Spatial Domain for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2014**, *52*, 2492–2507. doi:10.1109/TGRS.2013.2262052.

33. Rougier, S. *Apport Des Images Satellites à Très Haute Résolution Spatiale Couplées à Des Données Géographiques Multi-Sources Pour l'analyse Des Espaces Urbains*; Strasbourg, 2016.
34. Ma, L.; Fu, T.; Li, M. Active Learning for Object-Based Image Classification Using Predefined Training Objects. *International Journal of Remote Sensing* **2018**, *39*, 2746–2765. doi:10.1080/01431161.2018.1430398.
35. Amor, I.B.S.B.; Chehata, N.; Bailly, J.; Farah, I.R.; Lagacherie, P. Parcel-Based Active Learning for Large Extent Cultivated Area Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2018**, *11*, 79–88. doi:10.1109/JSTARS.2017.2751148.
36. Demir, B.; Minello, L.; Bruzzone, L. Definition of Effective Training Sets for Supervised Classification of Remote Sensing Images by a Novel Cost-Sensitive Active Learning Method. *IEEE Transactions on Geoscience and Remote Sensing* **2014**, *52*, 1272–1284. doi:10.1109/TGRS.2013.2249522.
37. Malek, S.; Miglietta, F.; Gobakken, T.; Næsset, E.; Gianelle, D.; Dalponte, M. Optimizing Field Data Collection for Individual Tree Attribute Predictions Using Active Learning Methods. *Remote Sensing* **2019**, *11*, 949. doi:10.3390/rs11080949.
38. Matasci, G.; Tuia, D.; Kanevski, M. SVM-Based Boosting of Active Learning Strategies for Efficient Domain Adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2012**, *5*, 1335–1343. doi:10.1109/JSTARS.2012.2202881.
39. Alajlan, N.; Pasolli, E.; Melgani, F.; Franzoso, A. Large-Scale Image Classification Using Active Learning. *IEEE Geoscience and Remote Sensing Letters* **2014**, *11*, 259–263. doi:10.1109/LGRS.2013.2255258.
40. Hagolle, O.; Huc, M.; Villa Pascual, D.; Dedieu, G. A Multi-Temporal and Multi-Spectral Method to Estimate Aerosol Optical Thickness over Land, for the Atmospheric Correction of FormoSat-2, LandSat, VEN μ S and Sentinel-2 Images. *Remote Sensing* **2015**, *7*, 2668–2691. doi:10.3390/rs70302668.
41. Inglada, J. Otb Gapfilling, A Temporal Gapfilling For Image Time Series Library, 2016. doi:10.5281/zenodo.45572.
42. Karasiak, N.; Dejoux, J.F.; Fauvel, M.; Willm, J.; Monteil, C.; Sheeren, D. Statistical Stability and Spatial Instability in Mapping Forest Tree Species by Comparing 9 Years of Satellite Image Time Series. *Remote Sensing* **2019**, *11*, 2512. doi:10.3390/rs11212512.
43. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
44. Belgiu, M.; Drăguț, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *114*, 24–31. doi:10.1016/j.isprsjprs.2016.01.011.
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Ghosh, A.; Joshi, P.K. A Comparison of Selected Classification Algorithms for Mapping Bamboo Patches in Lower Gangetic Plains Using Very High Resolution WorldView 2 Imagery. *International Journal of Applied Earth Observation and Geoinformation* **2014**, *26*, 298–311. doi:10.1016/j.jag.2013.08.011.
47. Foody, G.M. Explaining the Unsuitability of the Kappa Coefficient in the Assessment and Comparison of the Accuracy of Thematic Maps Obtained by Image Classification. *Remote Sensing of Environment* **2020**, *239*, 111630. doi:10.1016/j.rse.2019.111630.
48. Settles, B. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2012**, *6*, 1–114. doi:10.2200/S00429ED1V01Y201207AIM018.
49. Vlachos, A. A Stopping Criterion for Active Learning. *Computer Speech & Language* **2008**, *22*, 295–312. doi:10.1016/j.csl.2007.12.001.
50. Lewis, D.D.; Gale, W.A. A Sequential Algorithm for Training Text Classifiers. In *SIGIR '94*; Croft, B.W.; van Rijsbergen, C.J., Eds.; Springer London: London, 1994; pp. 3–12. doi:10.1007/978-1-4471-2099-5_1.
51. McCallum, A.; Nigam, K. *Employing EM and Pool-Based Active Learning for Text Classification*; ICML '98, Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1998; p. 350–358.
52. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
53. Brinker, K. Incorporating Diversity in Active Learning with Support Vector Machines. *International Conference on Machine Learning*; AAAI Press: Washington, DC, USA, 2003; pp. 59–66.
54. Gong, Z.; Zhong, P.; Hu, W. Diversity in Machine Learning. *IEEE Access* **2019**, *7*, 64323–64350. doi:10.1109/ACCESS.2019.2917620.

55. Zhong, P.; Gong, Z.; Li, S.; Schönlieb, C.B. Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3516–3530. doi:10.1109/TGRS.2017.2675902.
56. Danka, T.; Horvath, P. modAL: A Modular Active Learning Framework for Python **2018**. [<https://arxiv.org/abs/1805.00979>].
57. Rajan, S.; Ghosh, J.; Crawford, M.M. An Active Learning Approach to Hyperspectral Data Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2008**, *46*, 1231–1242. doi:10.1109/TGRS.2007.910220.
58. Joshi, A.J.; Porikli, F.; Papanikolopoulos, N. Multi-Class Active Learning for Image Classification. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2372–2379. doi:10.1109/CVPR.2009.5206627.
59. Di, W.; Crawford, M.M. Critical Class Oriented Active Learning for Hyperspectral Image Classification. 2011 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Vancouver, BC, Canada, 2011; pp. 3899–3902. doi:10.1109/IGARSS.2011.6050083.
60. Duchaufour, P. Les Sols à Peupliers. *Revue Forestière Française* **1955**, p. 539. doi:10.4267/2042/27114.
61. Fischer, M.; Zenone, T.; Trnka, M.; Orság, M.; Montagnani, L.; Ward, E.J.; Tripathi, A.M.; Hlavinka, P.; Seufert, G.; Žalud, Z.; King, J.S.; Ceulemans, R. Water Requirements of Short Rotation Poplar Coppice: Experimental and Modelling Analyses across Europe. *Agricultural and Forest Meteorology* **2018**, *250–251*, 343–360. doi:10.1016/j.agrformet.2017.12.079.
62. Leiva-Murillo, J.M.; Gomez-Chova, L.; Camps-Valls, G. Multitask Remote Sensing Data Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2013**, *51*, 151–161. doi:10.1109/TGRS.2012.2200043.