Assessment of Optimal Transport for Operational Land-Cover Mapping Using High-Resolution Satellite Images Time Series without Reference Data of the Mapping Period

Benjamin Tardy, Jordi Inglada, Julien Michel

remote sensing

MDPI

Article

Assessment of Optimal Transport for Operational Land-Cover Mapping Using High-Resolution Satellite Images Time Series without Reference Data of the Mapping Period

Benjamin Tardy [1,*], Jordi Inglada [1,2] and Julien Michel [2]

[1] Centre d'Etudes Spatiales de la BIOsphere (CESBIO), Université de Toulouse, CNES/CNRS/IRD/UPS/INRA, 31401 Toulouse, France; ingladaj@cesbio.cnes.fr
[2] Cente Nationnal d'Etudes Spatiales (CNES), 31401 Toulouse, France; Julien.Michel@cnes.fr
* Correspondence: benjamin.tardy@cesbio.cnes.fr; Tel.: +33-561-556-686; Fax: +33-561-558-500

Received: 20 March 2019; Accepted: 29 April 2019; Published: 3 May 2019

Abstract: Land-cover map production using remote-sensing imagery is governed by data availability. In our case, data sources are two-fold: on one hand, optical data provided regularly by satellites such as Sentinel-2, and on the other hand, reference data which allow calibrating mapping methods or validating the results. The lengthy delays due to reference data collection and cleansing are one of the main issues for applications. In this work, the use of Optimal Transport (OT) is proposed. OT is a Domain Adaptation method that uses past data, both images and reference data, to produce the land-cover map of the current period without updated reference data. Seven years of Formosat-2 image time series and the corresponding reference data are used to evaluate two OT algorithms: conventional EMD transport and regularized transport based on the Sinkhorn distance. The contribution of OT to a classification fusion strategy is also evaluated. The results show that with a 17-class nomenclature the problem is too complex for the Sinkhorn algorithm, which provides maps with an Overall Accuracy (OA) of 30%. In contrast, with the EMD algorithm, an OA close to 70% is obtained. One limitation of OT is the number of classes that can be considered at the same time. Simplification schemes are proposed to reduce the number of classes to be transported. Cases of improvement are shown when the problem is simplified, with an improvement in OA varying from 5% and 20%, producing maps with an OA near 79%. As several years are available, the OT approaches are compared to standard fusion schemes, like majority voting. The gain in voting strategies with OT use is lower than the gain obtained with standard majority voting (around 5%).

Keywords: land-cover; satellite image time series; Random Forests; Domain Adaptation; Optimal Transport; Classification Fusion

1. Introduction

A land-cover map is an image where each pixel value corresponds to a land-cover class label. Land-cover maps are essential in various projects for Earth monitoring, such as forest evolution [1], urban expansion [2], or climate modelling [3]. An efficient way to produce land-cover maps is to use satellite images, especially from the optical domain [4,5]. Recent decades have seen the launch of numerous satellites dedicated to Earth monitoring which provide high-quality data for land-cover mapping, such as the SPOT and Landsat families and more recently the Sentinel-2 constellation. The use of satellite image time series improves the classification accuracy in comparison to single date classification approaches [6]. Land-cover maps can be produced with two main procedures, human photo-interpretation and automatic mapping using machine learning classifiers. In practice,

image interpretation is a time-consuming hand task in contrast to automated processing. For an operational land-cover mapping over large areas, automated processing is preferred. Classification algorithms based on machine learning methodologies are often applied, which consist of a set of decision rules assigning a label to each pixel. Two categories of classification algorithms are considered for land-cover mapping: supervised and unsupervised ones. State of the art has shown that supervised classifiers such as Support Vector Machine (SVM) [7] or Random Forests (RF) [8] obtain better performance than unsupervised approaches [9]. Maximum Likelihood classifiers have also been used for land-cover mapping [10], but they do not behave well in high dimensional feature spaces as those encountered with multispectral image time series. SVM and RF can handle data sets with large numbers of features, such as the various spectral bands for the different dates of satellite image time series, and with many training samples covering a wide geographical extent. In addition to image time series, high-quality labelled reference data are an essential element for an efficient classification [11].

Reference data can come from various sources such as dedicated field surveys, or existing databases. Field surveys are expensive and tedious to perform since it is necessary to visit the fields several times a year to handle the seasonal land-cover classes such as crops. Therefore, reference data are generally available after the corresponding mapping period, whereas the remote-sensing images are available a few days after their acquisition. In essence, the delays in obtaining reference data are the main obstacles to the timely frequent production of land-cover maps. These delays are even longer when using publicly available databases since those have their own production and updating cycles. Naturally, the delays are proportional to the area covered and the kind of classes considered. Consequently, the large scale (i.e., over large areas) land-cover mapping [12] at the annual frequency cannot rely on field surveys nor existing data bases.

In this study, we propose and evaluate methods to overcome the absence of reference data for the period to be mapped. The purpose is to classify an image time series by using prior information as past image time series and past reference data.

However, the use of previous data is not straightforward due to changes between years. These changes can occur at two levels. At the image level, the behavior of the same vegetation class may differ between years due to climate differences, and the number and dates of the acquisitions may vary. At the class level, the same point of the map can belong to different classes according to the year, which requires the production of a new land-cover map each year. The land-cover changes or transitions can be natural or artificial. A natural transition occurs, for instance, when an edge river pixel becomes soil instead of water due to the water level. An artificial transition is often due to urban expansion when a new building was constructed on farmland. In our application case, the main artificial transitions correspond to crop classes, due to crop rotation, which consists mainly of alternating winter and summer crops for two consecutive years. For all these reasons, reference data have only a limited validity period and must be used only with the contemporaneous image time series.

In a previous work, we proposed simple strategies that use the history of the mapped area and fusion schemes [13]. First, a *naive* baseline was defined, consisting of training a classifier using past (image and reference) data, and classifying the image time series of the current period. The *naive* classification only considers one previous period at the same time. This approach yields to maps with an OA between 50% and 70%.

A second method, the *Single Classifier*, was trained using all the available previous data. In the case in which only one year is used, this is equivalent to the *naive* classifier approach. This approach provides an improvement of the OA of *naive* case between 5% and 10% depending on the number of previous data used. The Single Classifier has a major limitation as the method requires a large amount of training samples to be efficient. Then, it is necessary to keep all the data which may cause issues for an operational processing.

A third approach using voting methods was proposed. Majority voting using the *naive* case maps of several previous periods and a variant using the confidence of the *naive* classifiers yielded interesting

results approaching an OA of 82%. The main drawback of these approaches was the need for several past periods.

In the literature, the lack of reference data is handled with Domain Adaptation (DA) algorithms.

DA considers two domains, the source domain D_S which has reference data and the target one D_T for which few or no reference data are available.

Generally, for land-cover mapping two cases are considered:

- The reference data of the source domain is not representative of the area covered by the image [14].
- The source and target domains are two different images [15]. For this use of DA methods, both domains must be related in terms of land-cover classes and climate.

Tuia et al. [16] recently did a survey of DA methods used for remote-sensing data classification. The authors listed four categories of algorithms:

1. Adapting classifiers with unsupervised or semi-supervised approaches.
2. Adapting classifiers by Active Learning.
3. Selecting invariant features for classification.
4. Adapting data distributions.

The two first categories consist of adapting the decision rules of the classifier using the target domain data whereas the two last ones aim at adapting the data. The first step consists of finding common information in both domains. Then, the source domain is adapted to take into account shifts. Once this is done, the target domain can finally be classified.

Adapting the classifier is one of the most complex tasks. The main idea is to train a classifier on the source domain and then identify useful elements in the target domain and use them to modify the decision rules of the classifier. The first category of algorithms is automated using unsupervised or semi-supervised approaches. The work of Bruzzone et al. [14], considers a Maximum Likelihood classifier to propose an unsupervised update of the decision rules. This classifier, based on Bayes decision theory, uses the *a-priori* probability and the conditional probability density distribution of each class to determine a set of decision functions. Then, they use different methods to compute the same probabilities for the target domain and iteratively adapt the decision rules. In a similar context, Rajan et al. [17] use a knowledge transfer approach to compute weights based on target domain samples in an ensemble of Binary Hierarchical Classifiers (BHC) [18].

In our context, the land-cover maps are produced using RF. The decision trees in the RF are more elaborate than the ones constituting a BHC ensemble. Indeed, the BHC transforms a C class classification problem into a $C-1$ binary classification problem, instead of considering all the classes, as in the RF.

The second category of classifier adaptation algorithms is a set of methods based on Active Learning. Active Learning approaches achieve satisfying results as the user interacts several times during the process. Indeed, the term 'Active' refers to the fact that the user must select domain samples and label them by visual interpretation for instance. This type of method is prohibitive in our use because it is unusable in operational production at the national scale.

The feature extraction approach involves finding a set of features usable in both domains. The two domains are considered at the same time to determine a set of invariant features, i.e., the ones that are the least affected by shifts. Using the invariant features also allows the shifts to be estimated. A projection matrix is computed and used to transform both domains into a new one. Then, both the source and target data points can be processed indiscriminately in this single domain. In the literature, the authors mainly propose methods for the invariant feature detection and selection [15,19]. The authors propose different distance measures to determine the shifts between domains. These types of methods are widely used to correct two kinds of shifts, the illumination variations or effects due to the sensor viewing angle.

In our case, the used data are acquired with a constant viewing angle and the radiometric variations due to illumination changes are already corrected and expressed in surface reflectance, meaning the images are invariant to these shifts. The second case of use of feature extraction is to smooth the differences between two different parts of an image. The problem of classifying a large image with a small area covered by reference data is not addressed in this work, as in a current use the images are split according to eco-climatic areas before mapping [12] and the reference data are an excellent representation of the mapped area. In this work, the Formosat-2 images are included in a unique eco-climatic area. A split into eco-climatic areas seems to be more efficient than this use of DA in our case, as a complete satellite image can cover very different landscapes, including seaside mountains and plains for example, and an essential predicate for the DA is that the two domains must be similar. Moreover, removing features is not considered at the moment as the RF chooses relevant features during the training phase. Coupled with a scattered reference data on the image, the RF has a great generalization capability [20].

The last category of DA methods aims to adapt the data. These approaches seem to be the most interesting in our case. The main idea is to find a transformation to ensure that the data from the source domain matches the distribution of the data of the target domain. Once the source data are projected in the target domain, the supervised classifier can be used. The main difference with the feature extraction methods is that all the features are kept during the projection. In the literature, two major approaches are presented: a projection matrix estimation or an alignment of the data distributions. The proposed methods to align the distributions aim to estimate the distance between two different time series: for instance, the Dynamic Time Warping (DTW) [21]. In another example, the alignment is done by using histogram matching methods [22]. The most commonly used methods to estimate the projection matrix, are based on a kernel matrix. Mainly existing works propose different ways to estimate this matrix [23,24]. This category is widely referenced in the literature, with many algorithms, the interested reader can refer to the surveys [16,25].

Recently Optimal Transport (OT) was used for DA [26]. The OT problem is a well-known problem in statistics and mathematics. This problem has seen significant progress during the Second World War. Kantorovich [27] describes the transportation problem as follows: "Either $S_1, S_2, \cdots, S_m$ stations attached to a railway network deliver goods at a rate of $s_1, s_2, \cdots, s_m$ wagons per day, respectively. On this same railway network, there are $T_1, T_2, \cdots, T_n$ stations which consume these goods at a rate of $t_1, t_2, \cdots, t_n$ wagons per day, respectively. At the end of the day, the following equality must be respected: $\sum s_i = \sum t_k$ that means all wagons for the S stations have been received by the T stations. The cost $c_{i,k}$ representing the energy required by one wagon to be carried from the station S_i to the station T_k is known. The problem is, therefore, to assign to each consumption station, a path taking into account the production stations so that the total costs due to transport are minimal".

By considering the quantities of goods transported by a wagon as probability densities, OT has become a DA method, used to find an optimal transformation from Source to Target. Once an optimal coupling between the source and target domains is found, the coupling matrix is used to project the source data on the target domain. OT is presented in detail in Section 2.2.

Most of the methods listed above have limited use cases, making a transition to an operational use complex. Indeed, most of the time, the set of classes is restricted to classes with very different characteristics, such as water, building, forest, pasture, vineyards but rarely crops. The study area is generally reduced to an area of about 500×500 pixels. Domains are generally composed of a single satellite image, which raises the question of dimensions when using a complete time series.

In this paper, we propose to use the OT to produce the current land-cover map without using the corresponding reference data. The OT algorithms are compared to the *naive* classifier which can be seen as a very simple DA approach. The use of OT in fusion schemes is also shown.

The remainder of the paper is organized as follows. Section 2 introduces the data and the methods used. Section 3 presents the results obtained. Section 4 gives discussions about the different methods. Then conclusion and perspectives for future works are drawn in Section 5.

2. Materials and Methods

This section presents the used data, the classification process, and the OT algorithms.

2.1. Experimental Setup

2.1.1. Image Time Series

A large data set composed of 7 years of Formosat-2 images was used for this study. Formosat-2 has a one-day revisit cycle and a spatial resolution of 8m. The images were first atmospherically corrected using MACCS [28]. Masks for clouds, cloud shadows, and saturated pixels were provided as a by-product of the correction. Using the four spectral bands: blue, red, green, near infrared (NIR), two spectral indices were computed: NDVI (Normalized Difference Vegetation Index) and brightness (the norm of the spectral vector). As shown in Table 1, the number of available images and their repartition are different for each year, and therefore, the time series used as features for the classification have different lengths. To have the same number of features for each year, each time series is resampled onto a regular time grid. Linear interpolation together with the masks as mentioned earlier was used. Only the reflectance values of the four spectral bands are interpolated, the indices are computed after interpolation.

The period begins in October of the previous year and ends in December of the current year, so the entire phenological cycle of agricultural classes are covered. Finally, the satellite image time series are composed of the concatenation of the spectral bands, and the spectral indices of each date [29], yielding seven image time series.

Table 1. Number of available images per month for all the used periods. **0** represents a lack of images, either due to clouds or technical problems.

Month Year	2007	2008	2009	2010	2011	2012	2013
October (N-1)	3	3	3	1	2	2	**0**
November (N-1)	6	**0**	**0**	**0**	1	**0**	1
December (N-1)	**0**	1	**0**	1	**0**	1	2
January	**0**	**0**	**0**	**0**	1	1	**0**
February	2	1	1	**0**	**0**	1	1
March	**0**	**0**	2	1	**0**	2	1
April	1	**0**	**0**	2	1	**0**	**0**
May	1	**0**	1	1	3	1	2
June	1	2	3	2	1	2	3
July	1	2	3	2	1	2	3
August	2	2	3	3	**0**	2	2
September	3	1	3	1	2	**0**	2
October	3	3	1	2	2	**0**	2
November	**0**	**0**	**0**	1	**0**	1	1
December	1	**0**	1	**0**	1	2	1

2.1.2. Reference Data

Reference data for the seven satellite image time series were obtained by field surveys. These field surveys are designed to visit a high number of diverse crop plots while minimizing the travel distance. Each annual crop plot is visited each year. This may not be the case for perennial crops, which can be assessed by photo-interpretation. The resulting reference data is a stratified opportunistic sampling, which is not exactly proportional to the distribution of classes in the entire study area, although ranks are preserved. Figure 1 is an illustration of a true-color Formosat-2 image and the polygons resulting of the corresponding field campaign for 2010. The reference data covers large and diverse parts of study area.

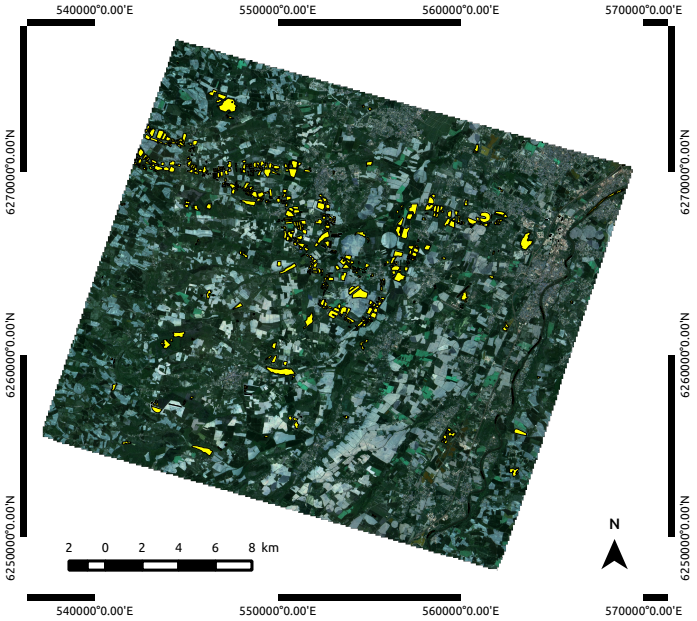

Figure 1. True-color Formosat-2 image and the corresponding reference data (yellow polygons) provided by the field campaign.

These data can be used for classifier training and validation purposes. With this objective, the data for each year are split into two independent data sets. To simulate a DA problem, we consider that the reference data are only available for the source domain. For the target domain, the reference data is only used for validation. The split between training and validation set and the use of reference data are detailed in the Section 2.1.4.

The reference data set contains 17 classes, representing either artificial classes such as buildings, natural classes such as forests or water bodies and finally, agricultural classes. The details of these classes are shown in Table 2. Zeros in bold represent a class which has not been sampled during the survey of that particular year. For instance, hemp was not grown in the region until 2009, and fallows were not included in the field work in recent years.

Table 2. Nomenclature and number of samples (pixels) for each period provided by the whole reference data set.

	Number of Samples of Each Period						
Classes	2007	2008	2009	2010	2011	2012	2013
Broad-leaved tree	33,659	39,060	40,905	28,702	39,743	39,743	39,989
Pine	10,160	13,112	6486	3703	3611	3611	3611
Wheat	66,116	49,848	23,854	66,047	340,803	58,476	97,825
Rapeseed	27,651	12,933	25,937	13,869	67,104	9885	40,508
Barley	1937	5908	3564	1203	35,799	12,055	20,270
Maize	58,438	39,185	49,570	54,858	142,214	29,063	105,107
Sunflower	5851	19,952	19,489	24,215	237,662	23,107	29,544
Sorghum	2040	1746	10,696	9829	8806	**0**	362
Soya	754	7921	8816	6497	12,482	**0**	2308
Artificial Surface	1550	1047	1047	1339	2089	1426	1496
Fallow land	16,148	5145	3396	**0**	35,110	**0**	**0**
Wasteland	1089	1299	9954	4142	10,357	10,357	14,208
River	5806	9092	6825	6736	13,298	8850	10,071
Lake	14,294	9997	10,090	20,070	4615	4440	4508
Gravel Pit	14,659	12,919	12,919	11,496	12,894	12,894	12,894
Hemp	**0**	**0**	960	1806	5881	670	279
Grass	42,656	11,900	13,571	18,379	120,299	21,182	25,858
Total	302,808	241,064	267,568	272,891	1,092,767	235,759	408,840

2.1.3. Land-Cover Map Production

The starting point of our procedure is the framework presented in [29], which consists of a RF supervised classification [8]. The forest is composed of 100 trees which are not pruned in the used implementation [30].

The standard supervised classification will be used here as an upper bound for the accuracy of the DA approaches presented in this paper. On the other hand, the straight application to the target domain images of a classifier trained in the source domain will constitute the lower bound for accuracy. This case is called *naive* in this paper.

An improvement of the *naive* case is proposed: the **Perennial Annual Split** (*PAS*) approach.

The goal of this method is to separate annual and perennial classes and to use the samples from the source domain and for the target domain, respectively. The perennial classes are broad-leaved trees, pine, artificial surfaces, river, and lake. The PAS approach is based on the fact that the land-cover class for perennial areas has a low probability of change from one year to another. The samples for perennial classes can be selected in the target domain according to the sample positions provided by the source domain reference data. For the annual classes (mainly crops), the sample position and their image features are provided by the source domain reference data. Then the classifier is trained using this mix of samples from both domains and used to classify the target domain time series.

The DA approaches will be applied as follows: the data of the source domain will be transformed to the target domain, the transformed data will be used for training the RF classifier and, finally, the classifier will be applied to the target domain data for which no reference data is available.

To be successful, DA must yield higher accuracies than the *naive* classification. In the ideal case, the accuracy would be equal to or higher than the supervised classification. In a previous work [13], voting methods based on the combination of the results of several *naive* classifications, were proposed. Majority voting of *naive* classifications was shown to improve the accuracy of each voter. These results are compared to the DA techniques proposed in this paper.

2.1.4. Validation

The validation of the DA techniques presented in this paper will be carried out classically since reference data are available for all image time series used in this work. The validation data will be used to compute confusion matrices and derived accuracy metrics. In the confusion matrix, the rows correspond to classes in the reference data, and the columns to the labels produced by the classifier. Each cell in the matrix indicates the number of samples belonging to the class given by the reference and classified as belonging to the class indicated by the column. The diagonal of the matrix represents the number of correctly classified pixels. The Overall Accuracy (OA), is the total number of correctly classified pixels divided by the total number of validation pixels. The FScore is the harmonic mean of precision and recall:

$$fscore = 2 \times \frac{recall \times precision}{recall + precision}$$

where:

- precision is the ratio between the correctly classified pixels and the sum of all pixels classified as this class;
- recall is the ratio between the correctly classified pixels and the total number of reference data pixels of that class.

Ten draws are made to separate each reference data set into training and validation sets. As the split uses plot polygon instead of pixels, the training and validation sets are independent and remain balanced. Once the reference set is split, all the pixels in the polygons are extracted. In the rest of the paper, the sample unit is the pixel. The ten draws are used to compute 90% confidence intervals and measure the robustness against the sample selection noise. In the case of voting methods, undecided pixels may appear if a tie occurs. These undecided pixels are removed for the computation of the

accuracy metrics [31]. The proportion of undecided pixels must be taken into account when interpreting the metrics.

2.2. Optimal Transport

Courty et al. [26] introduced the use of OT for DA in its discrete formulation.

The empirical distributions μ for the *source* and *target* domains are:

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{x_i^s}, \mu_t = \sum_{i=1}^{n_t} p_i^t \delta_{x_i^t} \tag{1}$$

where:

- δ_{x_i} is the Dirac at location x_i,
- p_i is the probability mass associated with the i^{th} sample.

These empirical distributions represent the masses that will be transported under the constraint of preservation of the conditional distributions between classes: $P_t(y|x^t) = P_s(y|T(x^s))$. To this end, the transformations are restricted to those for which the transformed source distribution matches the target domain distribution. These kinds of transformations are obtained considering a probabilistic coupling γ between $P(X_s)$ and $P(X_t)$. Then the set of probabilistic couplings β is defined as:

$$\beta = \{\gamma \in (R^+)^{n_s \times n_t} | \gamma \mathbf{1}_{n_t} = \mu_s, \gamma^{\mathbf{T}} \mathbf{1}_{n_s} = \mu_t\} \tag{2}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. All the existing couplings are represented in β. The Optimal Transport algorithm computes the one which represents the transportation with the minimal cost. The optimal coupling, according to the Kantorovich formulation, is given by:

$$\gamma_0 = \underset{\gamma \in \beta}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F \tag{3}$$

where C is the always positive cost function, and $\langle ., . \rangle_F$ the Frobenius dot product. The C cost function is a matrix representing the energy needed for moving a sample from x_i^s to x_j^t. In most cases, the cost is computed using a metric of the embedding space. In this work, the cost is defined by the squared Euclidean distance: $C(i,j) = \|x_i^s - x_j^t\|_2^2$. These equations are the so-called *Earth Mover's Distance* (EMD) formulation for the discrete OT problem. The EMD solution can be obtained by using linear solvers, but in the case of small samples size or outliers, the solution can over-fit the data.

To avoid over-fitting, a regularization based on the *Sinkhorn* distance [32] can be introduced. In this case, the minimization problem solution from Equation (3) becomes:

$$\gamma_0^\lambda = \underset{\gamma \in \beta}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega_s(\gamma) \tag{4}$$

where $\Omega_s(\gamma) = \sum_{i,j} \gamma(i,j) \log \gamma(i,j)$ is the negative entropy of γ.

Once the solution γ_0 is found, the source samples can be transported in the target domain by:

$$\hat{\mathbf{X}}_s = diag(\gamma_0 \mathbf{1}_{n_t})^{-1} \gamma_0 \mathbf{X}_t \tag{5}$$

Then the transported samples $\hat{\mathbf{X}}_s$ are used to train the model.

These OT formulations consider only the samples without class label information. This information can be introduced both in the sample selection or the cost computation.

2.2.1. Using Class Labels for Sample Selection

Since the OT is estimated using pairs of samples from the source and the target domains, the selection of the samples is crucial. Indeed, for computational complexity issues (the C cost function is a $n_s \times n_t$ matrix) not all available samples can be used.

Label information can be used to ensure that the prior probabilities of the source and target domains are preserved during this sample selection.

2.2.2. Using Class Labels for Cost Computation

In the Sinkhorn regularization, the use of label information allows samples from the same class to remain grouped during the transport. In the case where labels are provided to samples from both domains, the cost between two samples of different classes is set to a very high value (in practice, the maximum number in the numerical encoding of the platform), as illustrated in Figure 2.

This figure shows the cost and coupling matrices for a toy problem with three classes with Gaussian distributions. The samples are grouped by class, giving a block structure to the matrices. When the label information is not used (left column), the cost of coupling between samples of different classes is lower than when the label information is used (right column). Consequently, the optimal coupling matrices in the bottom row may present some couplings across classes when no label information is used (left column). These couplings cannot exist when the label information is used (right column).

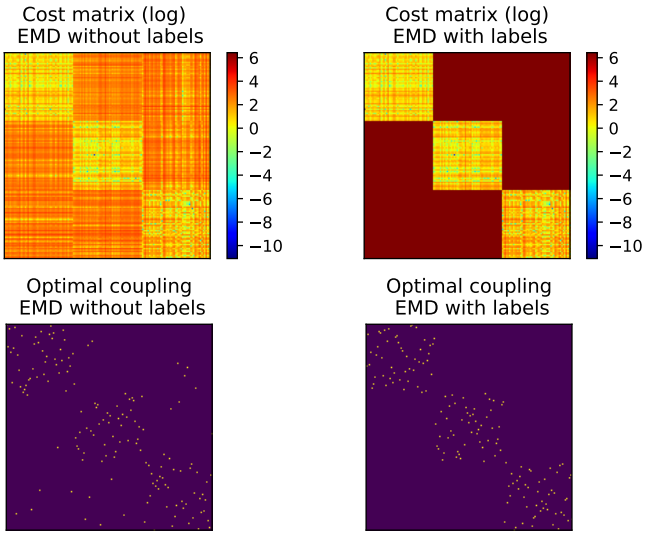

Figure 2. Cost and coupling matrices. The left column shows the EMD standard use and the right column, the use of labels from both domains. The used data are simulated using Gaussian distributions with only three classes. For a better visualization of the cost values, a logarithmic scale was used.

Depending on the algorithm, using the labels of the two domains does not have the same effect. Indeed, the EMD does not update the cost matrix, which has the effect of merely forbidding the coupling of samples of different classes. On the other hand, Sinkhorn will update this cost and therefore allow two close samples to be considered (even if they are from different classes) when estimating the transformation.

In this work, the *Python Optimal Transport* library [33] is used. It provides the implementation for the EMD solver and three regularization solvers based on the Sinkhorn distance.

2.2.3. Land-Cover Mapping with Optimal Transport

In this work, each time series constitutes a domain. To measure the method robustness, each possible pair of time series is used. This study considers the use of a single source domain to

classify a single target domain. The use of several source domains simultaneously could be considered in future works.

For land-cover mapping, OT is a pre-processing step as it aims to adapting the data distributions. Then the full procedure follows four major steps:

1. Select samples from both domains to estimate the transport.
2. Apply the transport to the training set given by the source domain reference data.
3. Train the classifier using these transported training samples.
4. Use the model learned to classify the target domain time series.

Figure 3 illustrates the OT processing for land-cover mapping. The first step is the most important, as the transformation must not degrade information. To this end, it is mandatory to ensure that the conditions to estimate the OT correctly are met. The second point is very important too, the transported training samples are always the same. This allows an efficient comparison of the methods. The training set is provided by the source domain reference data as described in the Section 2.1.2. This ensures that the methods can be compared as the number of samples used to train the model are always the same. The two last steps are similar to the *naive* use of RF. The main difference is that the training samples are adapted to the shifts in the target domain. This design reduces the bias associated with RF training and therefore allows a better comparison of different methods.

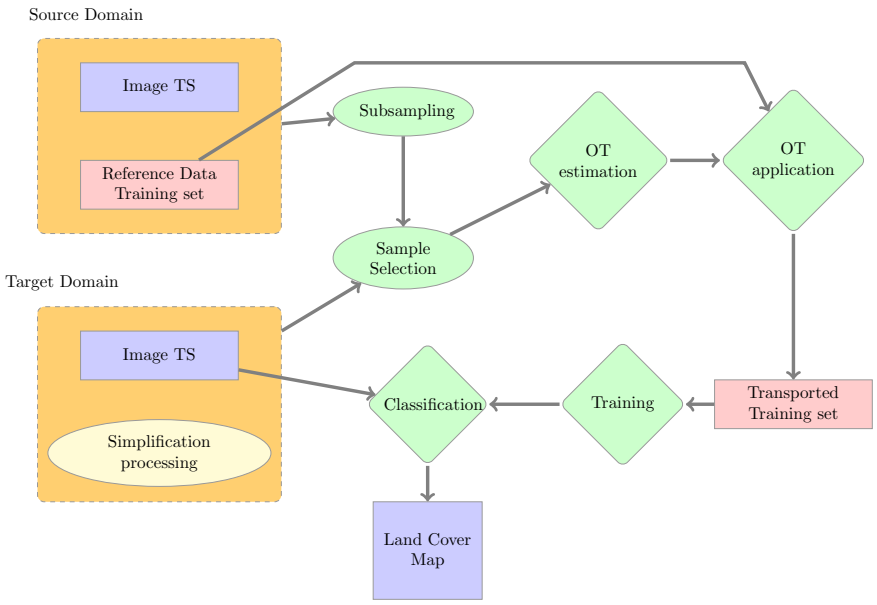

Figure 3. Land-cover mapping using OT for adapting the data. The algorithms used are represented by diamonds and the sampling strategies indicated by ellipses. The yellow ellipse represents the different simplification procedures proposed in this work.

2.3. Problem Simplification

With a data set composed of 17 classes, where each sample is characterized by 108 dimensions (four spectral bands, two spectral indices, for 18 dates), the standard data set sizes used in the literature on DA are exceeded by a significant amount [15,19]. Therefore, some adaptations are required for the use of these methods.

OT uses matrices whose sizes are proportional to the number of samples. These sizes are a numerical bottleneck in our case for both computational time and memory. As described in the Section 2.1.2, the number of samples for each period ranges from 200,000 to 500,000 samples. Reducing this number for transport estimation is therefore mandatory.

In addition to these technical constraints, other constraints related to the definition of the transport itself must be taken into account. The first constraint concerns the estimation of the empirical

distribution μ. While the distribution estimation for the source domain can be done using the supervised classification, the estimation for the target domain is not trivial. A satisfying compromise is to use a uniform distribution [34], ensuring that it is respected when selecting samples. Sample selection must also ensure that an acceptable representation of each class is maintained. Otherwise, the ability of the RF to generalize will be reduced.

A second constraint is the high number of classes, which impacts the regularization complexity. Moreover, some classes are very similar, as the three water classes (river, lake, gravel pit) or the pairs fallow land/grass area and barley/wheat.

The rest of this section presents the sample selection strategies considered for reducing the problem complexity.

2.3.1. Baseline Methods

Four baseline methods are proposed to evaluate the efficiency of different OA approaches. To clearly understand the behavior of the method, the reference data of both domains are used, so that no error due to sample selection in the target domain interferes with the transport estimation. As the labels are known, the reference data allows us to identify the classes of the samples, and therefore to balance the data. The number of samples required is defined by the class in the reference data with the fewest samples. This selection has a low error rate, as the reference data contains around 2% of errors. Once the samples of both domains are selected, there are four ways to estimate the transportation matrix. As these methods use the reference data of both domains their name starts by *SUP* as in *supervised*.

1. Only the samples are used with the EMD algorithm. This approach is called *SUP-EMD*.
2. The samples and the source labels are used with the Sinkhorn algorithm. This approach is called *SUP-Sinkhorn*.
3. The samples and both domains labels are used with the EMD algorithm. This approach is called *SUP-EMD-WL* (*WL* meaning With Labels).
4. The samples and both domains labels are used with the Sinkhorn algorithm. This approach is called *SUP-Sinkhorn-WL*.

With these baselines, the impact of the label use on the transport estimation was measured. As explained above, the use of labels in both domains has an impact on the cost and coupling matrices. These baselines give the upper limit expected by each method, but do not represent a real case of use, since the labels of the target domain are used. For real cases, where these labels are not available sampling strategies have to be used.

2.3.2. Sampling Selection Strategies

In a real case of use, several elements are available for each domain. The source domain is the most complete, with the time series, the reference data, the trained model and the supervised classification. For the target domain, the time series and *naive* classification are available. Four sample selection methods are proposed, using all these elements:

1. The source domain reference data provides the pixel positions. Then they are used to select samples in both domains:
    (a) The selected samples are used to estimate the EMD transport. This approach is called *SP-EMD*.
    (b) In the same way, the Sinkhorn transport is estimated, using the source domain labels. It is called *SP-Sinkhorn*.
2. A *naive* classification is produced for the target domain. Then the target samples are selected using the predicted labels.

(a) For each class produced in the *naive* map, an EMD transport is estimated. This method is named *EMD-PerClass*.
(b) Similar to the previous one, but only with annual classes, the *PAS* approach is considered with EMD transport. This approach is named *PAS-EMD*.
(c) On both *naive* and supervised maps, subclasses are extracted by clustering, and Sinkhorn transport is estimated for each class. This approach is called *SK-Cluster*.

The simplest selection method was not considered in our case. This approach would consider the whole image and randomly select the samples. This selection method is widely used in DA works, but in our case the area covered by an image is too large, resulting in a high imbalance of classes. Table 3 shows for each method the input data used and the use of label information for the OT estimation. The empty cells correspond to cases which have not been considered in this work.

The Same Position (SP) approach is the simplest selection method available in our work. At a given position the label can have one of the three following behaviors:

- It does not change, as for perennial classes such as artificial surfaces or lakes.
- It rarely may change, as is the case for stable classes such as maize or forests.
- It often changes, as is the case for crop classes according to the crop rotation (winter crop becomes a summer crop the next year).

The SP approach makes the hypothesis that stable classes do not change and that the crop rotation keeps class proportions stable between the two domains. At the end of sample selection with *SP*, it is expected that the sample distribution is uniform. The SP approach aims to reproduce the *SUP-EMD* and *SUP-Sinkhorn* baseline results.

Producing a *naive* classification allows using the provided labels to filter the sample selection. With this approach, the goal is to find relevant samples and then find an optimal transformation for a specific class. For the particular case of EMD, estimating the transport for each class gives similar results than providing labels for both domains. Estimating one transport for each class allows using more samples, and then providing a better representation of the class diversity. The EMD-PerClass approach aims to reproduce the *SUP-EMD-WL* baseline.

Based on the *naive* classification too, the *PAS-EMD* approach aims to combine the *PAS* and the *EMD-PerClass* approaches. With this approach, the number of classes to be transported is reduced, and correct samples are provided for perennial classes. The combined result should be the best as the classifier learns the perennial classes and corrects the shifts for samples from the source domain.

It is essential to maintain intra-class diversity for transport estimation, to ensure that the trained classifier is always as efficient. A K-Means clustering is applied to both domains, to maximize the intra-class diversity. For each class, only 5 clusters are used, expecting a subclass extraction. The time series are masked using the supervised map for the source domain, and the *naive* one for the target domain. Then, it is possible to estimate the Sinkhorn transport for each class. Instead of using class labels from the source domain, the cluster identifier was used to ensure subclasses to remain close during the transport. The goal of the *SK-Cluster* is to reproduce the *SUP-Sinkhorn-WL* baseline.

The four methods aim to reduce the problem complexity. The SP approaches remove noisy samples, such as plot boundaries which can correspond to mixed pixels. The *EMD-PerClass* aims to use more samples for each class, removing the constraint due to the distribution estimation. Finally, the *SK-Cluster* aims to decrease the regularization complexity, which is hard to resolve with 17 classes at the same time.

Table 3. The different proposed methods. The row indicates the used data for sample selection. The columns refer to how the labels are used in transport estimation. Labels come from source $Labels_S$ or target $Labels_T$ domains.

	Labels Use for OT Estimation		
Used data for Sample Selection	No labels	$Labels_S$	$Labels_S + Labels_T$
$RefData_{Source} + RefData_{Target}$	SUP-EMD	SUP-Sinkhorn	SUP-EMD-WL SUP-Sinkhorn-WL
$RefData_{Source}$	SP-EMD	SP-Sinkhorn	
$RefData_{Source} + Classif_{Target}$			EMD-PerClass PAS-EMD
$Classif_{Source} + Classif_{Target}$			SK-Cluster

3. Results

In this section, we present the evaluation of the performance of the proposed algorithms. First, a general overview is given by presenting results averaged for each year, followed by the results of all pairs of years. Secondly, the *EMD-PerClass* approach is analyzed in detail giving the performance for each class. Finally, the contribution of the majority voting using the maps produced with the *EMD-PerClass* approach is compared to majority voting using the *naive* maps.

3.1. Global Performance

To give an overall assessment of the proposed approaches and compare them to the *supervised* and *naive* baselines, the global averaged OA is presented. For the *supervised* case, the average is computed over ten random draws of the training data. For all other cases, the average includes the OA for all the combinations of a source domain year and all other years as target domain. For example, considering 2007 as source domain, the different pairs of years are $[(2007, 2008), (2007, 2009), \cdots, (2007, 2013)]$, then the OA obtained on the 10 draws for each pair of years are averaged and the 90% OA confidence interval is computed.

The OA obtained for each method are presented in Figure 4. Two types of plots are presented. The dotted curves represent methods that cannot be used in a real case because they use the labels in the target domain, which are not available in operational settings. The full curves indicate methods usable in an operational use case. In this figure, 12 curves are drawn, showing different source domain years on the x-axis and the OA on the y-axis. The *supervised* and *naive* cases are represented by the blue and dark curves, with OA around 90% and 70% respectively. These two baselines being upper and lower bounds of this work, only methods with an OA between them are interesting for land-cover map production. The *PAS* approach is represented with the yellow curve, but unexpectedly, the performances are similar or lower than the *naive* case. The Sinkhorn-based approaches, *SUP-Sinkhorn* (dashed green curve) and *SP-Sinkhorn* (cyan curve) are inefficient for all considered source domains, with OA between 25 and 45%. With more than 80% OA the *SUP-EMD-WL* (orange dashed curve) is the best method, although it is not a method usable in a real case. The five other methods are close to the *naive* OA. The *SK-Cluster* (grey curve) and *SP-EMD* (brown curve) have lower OA than the *naive* case, but with overlapping confidence intervals. The two baseline curves, *SUP-EMD* (light blue dashed curve) and *SUP-Sinkhorn-WL* (pink dashed curve) have slightly higher OA, but again, within the confidence intervals. These two methods cannot be considered in an operational production as they use the reference data for selected samples. The *EMD-PerClass* performances are represented with the red curve, showing a slight improvement with respect to the *naive* case. Finally, the last method is the *PAS-EMD* (purple curve). This approach produces maps with a statistically significative increase in OA compared to the *naive* case. However, as with all the proposed methods, the OA obtained remains very close, within the confidence interval of the *naive* case.

Despite this fact, three interesting results can be highlighted. First, considering the Sinkhorn algorithm, the OA obtained by the *SUP-Sinkhorn* and the *SUP-Sinkhorn-WL* show that the problem is too complex to be correctly solved with regularization. With an OA difference around 30%, this result is confirmed for all the source domains.

Second, the confidence intervals are narrow for all pairs of years, showing the stability of the methods for a given source domain. As five methods give similar results, the confidence intervals improve the methods comparison. The *SP-EMD* approach is the one with the widest confidence intervals, showing the importance of balancing the classes before the transport estimation. This point is shown by the OA obtained for 2011 as the source domain. For this year, the reference data cover almost entirely the image as it was produced by mixing two reference data sets. As the whole image is covered, the hypothesis for the SP approach that the crop rotations preserve the class distribution between two years (Section 2.3.2), is more likely. The confidence intervals confirm the inefficiency of the *SUP-Sinkhorn* and *SP-Sinkhorn*, as the difference between the maximum OA for Sinkhorn and the minimum for other method is more than 10%. At the opposite, the *SUP-EMD-WL* and *SUP-Sinkhorn-WL* have the narrowest confidence intervals, showing the importance of class balance during OT estimation again.

Thirdly, the link between the *naive* case and the *PAS-EMD* or *EMD-PerClass* OA is clearly shown, showing that improving the first should impact the second. This result opens the way for an iterative approach, where a new *PAS-EMD* can be estimated using the resulting map of the previous iteration.

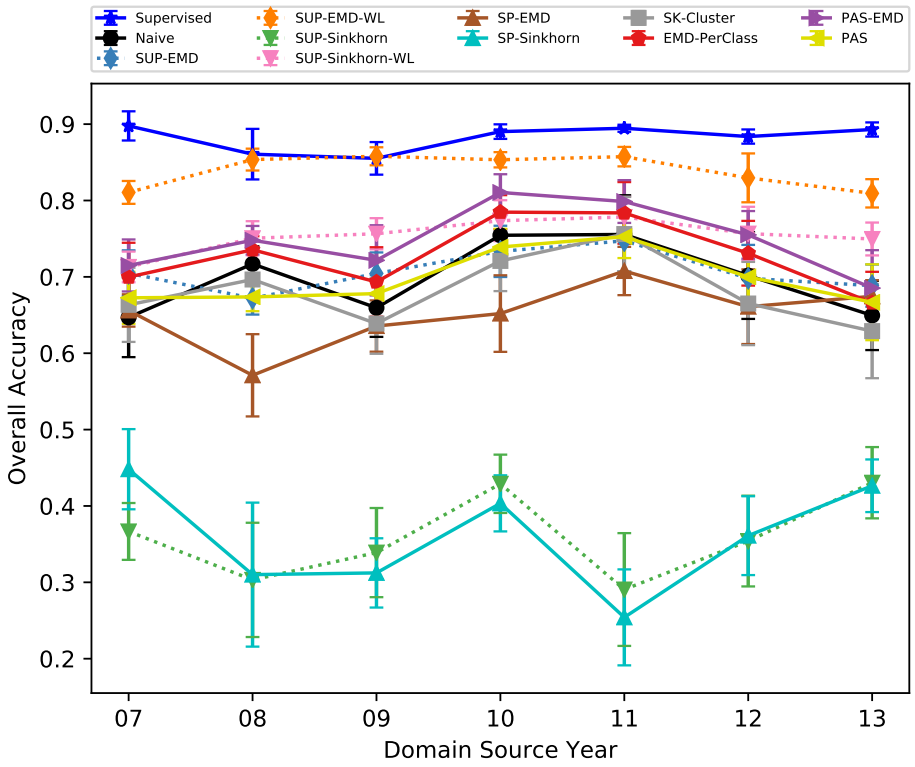

Figure 4. Mean Overall Accuracy for the different methods (y-axis) and the different source domains (x-axis). The nine methods using a unique source domain are represented in this figure, with the supervised, *naive* and *PAS* cases for reference. Full curves are methods usable in an operational case, whereas dashed curves represent methods for which the reference data of the target domain are used.

Impact of Domain Samples Selection

Figure 5 represents the matrices of differences between the OA of the *naive* case and all the other methods ($OA_{methods} - OA_{naive}$). Each matrix maps the source and target periods and, therefore, the diagonal represents the supervised case (set to zero here). Except for the first matrix, the values

represent the OA obtained by producing the map of the year on the y-axis from the data of the year on the x-axis according to the method indicated under the x-axis. The first matrix represents the difference between supervised and *naive* cases. For this particular case, the OA used for the supervised case is the one obtained for the target domain indicated on the y-axis. For the *naive* case, the value used is the one corresponding to the years on x-axis and y-axis. The color scale uses blue hues to show an OA increase in comparison to the supervised case and red hues for a decrease.

The color bar is not symmetrical, showing that the OA degradation is more significant than the improvement, but the color bar is centered on zero so the lighter the red, the less significant the difference is and the better the produced map is. The best cases are shown in dark blue.

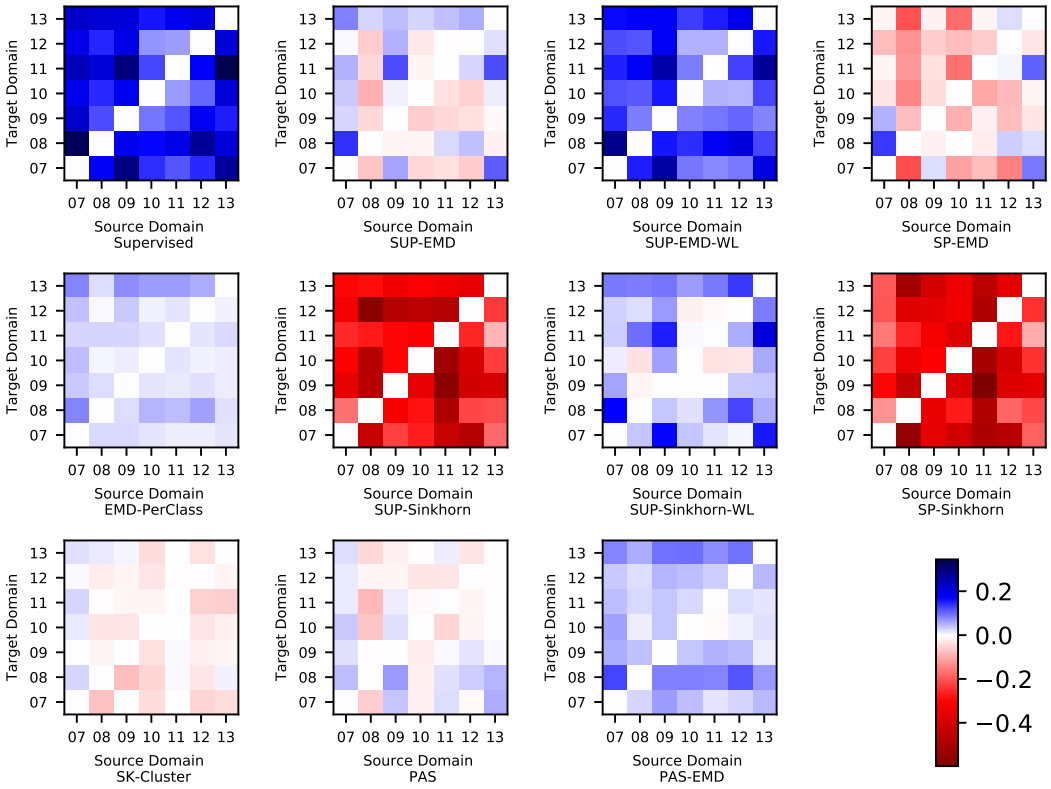

Figure 5. OA difference between the method on the x-label and the *naive* classification method. The color bar indicates the OA difference value. A red hue means the OA decreases using these methods, a blue hue indicates an increase. The x-axis represents the source domain. The y-axis represents the target domain. The diagonal values are set to zero as these cases only exist for supervised classification.

In the majority of cases, the proposed methods improve the *naive* case. The worst results are obtained by the Sinkhorn methods *SUP-Sinkhorn* and *SP-Sinkhorn* using labels on the source domain only, with an OA more than 40% lower than the *naive* case. For these two methods, the OA difference stays between -20% and -50% which is consistent with the overall view given by Figure 4. The two methods show relevant differences for some pairs of years as (2008, 2013) and (2010, 2012), which means that the selected samples have an impact on the regularization term. The impact of the sample selection is also shown with the *SUP-EMD* and *SP-EMD*, which provide different results.

The use of labels in both domains has a high impact, shown by the *SUP-Sinkhorn-WL* and *SUP-EMD-WL* graphs. For the Sinkhorn algorithm, the use of labels improves the OA in most of the cases because it allows being better than the *SUP-Sinkhorn* and *SP-Sinkhorn*. This result proves that simplifying the problem is required to use the Sinkhorn algorithm. The highest improvement is obtained with *SUP-EMD-WL* with an increase of OA near 10% for all cases. Again, the use of labels in the transport estimation allows for improving the performance.

The PAS approach provides similar results to the *naive* case, with very slight OA differences. Therefore, the *PAS-EMD* provides a significant improvement in comparison to the *PAS*. In most cases, it is better than the use of *EMD-PerClass* only. Also, the processing time is shorter for *PAS-EMD* than *EMD-PerClass* as the number of classes to be transported is lower.

The baseline cases provide three significant results:

1. The results are consistent since variations caused by the choices of the pairs of years used are limited to about 5 to 10%.
2. The used samples have an impact on the efficiency of the estimated transport.
3. Estimating the transport independently for each class seems more efficient because it is less complex.

The gain of per-class approach is shown by the *SK-Cluster*, the *EMD-PerClass* and the *PAS-EMD* methods. EMD seems to work better than the *SK-Cluster*. Introducing the class diversity in the regularization term is not as efficient as the *SUP-Sinkhorn-WL*, the main reason being the quality of *naive* classification which impacts the quality of used labels. The noise introduced by errors in the *naive* classification will impact the efficiency of the regularization term. In addition to the OA values obtained, the EMD per-class method is faster than the Sinkhorn processing, due in large part to the lack of the regularization step.

An interesting result is the absence of an optimal year pair for all methods. One can also observe from the structure of the matrices that each source domain does not have the same impact on a single target domain. This aspect opens the way to a fusion step of the maps produced after transport estimation if several source domains are available.

3.2. ParticulaR Case Analysis: PAS-EMD

We choose to focus the analysis on the one providing the best results, and usable in an operational context. In this section, a comparison between *naive* and *PAS-EMD* is provided, both for OA and per-class FScore.

Before considering the performance obtained by *PAS-EMD*, the FScore obtained by the *naive* classification is shown as it represents the lower baseline in this work. Figure 6 represents heat maps, where the color scale gives the FScore value. The source domain (the year of the trained model) is given on the x-axis, the target domain (the year of the produced land-cover map) is given on the y-axis. This figure shows that three categories of FScore are obtained, the first showing very high FScore values for all pairs of years (blue values), the second, at the opposite, very low FScore (red values) and the last, with average FScore depending on the considered years. Broad-Leaved tree and maize are majority classes and geographically stable in our data, which explains the excellent FScore obtained for all pairs of years. The worst FScore values are obtained with minority classes such as hemp or soybean. Several classes are known to be highly confused with majority classes such as barley being confused with wheat or sorghum with maize, which could explain the low FScores observed on these classes The heat maps are not symmetric. There is not a unique optimal pair of years for all classes shared by all the methods. These results are consistent with global analysis.

Figure 7 shows heat maps where the color represents the difference in FScore between *PAS-EMD* and *naive*. The classes are ordered as in Table 2, from left to right. The x-axis of each heat map represents the source domain year and the y-axis the target domain year. The values shown are computed as $FScore^{c}_{PAS-EMD} - FScore^{c}_{naive}$. The values are represented along a color bar where a blue hue means a higher difference in FScore and a red hue a lower difference. A white value equals zero, meaning equality between the two methods. The color bar extends from -0.2 to 0.8, but these extrema are not often reached. In most cases, the dominant color is blue, meaning that there is a global improvement. Each class has a different margin of improvement depending on the FScore value obtained by the *naive* case. The FScore has values between 0 and 1, so for classes where the FScore is higher than 0.8 an improvement of 0.1 is already satisfactory.

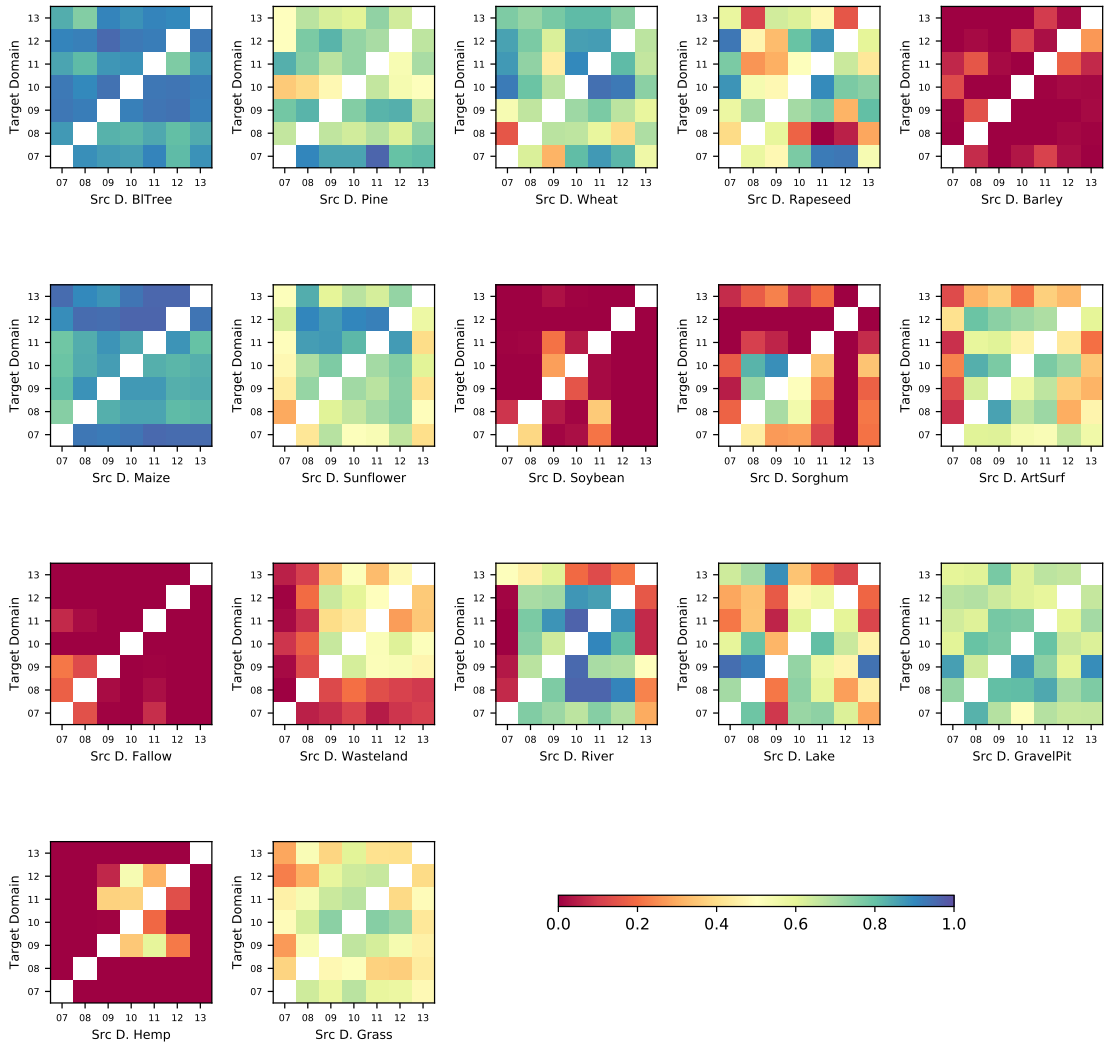

Figure 6. FScore of each class obtained in *naive* baseline. Each heat map corresponds to a class showing the results combining the various possible pairs of years. The diagonal represents the supervised case, so the values are set to 0. Source domain year and class are indicated on the x-axis, the target domain year on the y-axis. The color scale shows the difference range, with a low FScore value in red and very high FScore in blue.

Therefore, Figures 6 and 7 should be compared to understand the improvement brought by *PAS-EMD* use for each class. A significant improvement is shown for the river class, where a difference of about 0.8 is achieved. In general, perennial classes benefit from the *PAS-EMD* approach with an overall improvement of the FScore. The only exception is the artificial surface class for which degradation is observed when 2007 is the target domain. The FScore degradation cases are mainly grouped in the lower diagonal part of the matrix showing a sensitivity to the source domain used.

For the annual classes, the results are quite robust with an overall improvement between 0.1 and 0.3. Maize is the annual class that benefits the least from the contribution of OT, but the FScore degradation is limited to 0.05 which remains negligible.

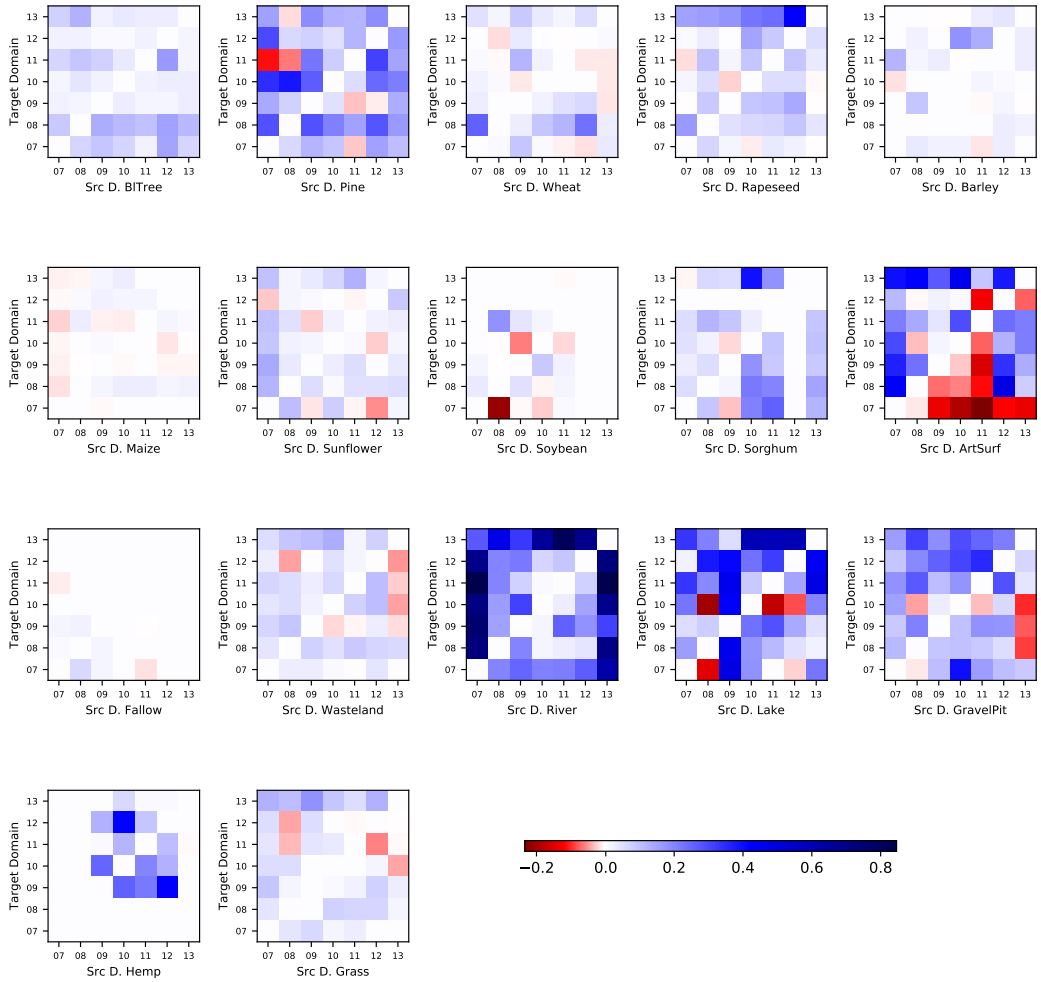

Figure 7. FScore difference between the *naive* case and *PAS-EMD* transport estimation. Each plot corresponds to a class for each pair of years available. The diagonals are empty. The source domain year and class is indicated on the x-axis, the target domain year on the y-axis. The color scale shows the difference range. Red hues represent a decrease in FScore and blues hues an increase.

3.3. Majority Voting after OT

The previous results are obtained using only one source domain at the same time. We investigate here the use of majority voting (MV) to use several source domains to predict one target domain. Figure 8 represents the OA obtained (on the y-axis) depending on the number of source domains used for voting (x-axis). Two kinds of curves are presented, the horizontal lines representing methods using only one domain source as input, and the full curves representing the MV performance. The horizontal lines correspond to the mean values over all source domains from Figure 4. The reference cases are the *supervised* (blue dashed line), the *naive* (black dashed line) and the *PAS* (yellow dashed line) cases which are mostly confounded. The two methods providing the best results are considered to be input for voting: *EMD-PerClass* (light blue dashed line) and *PAS-EMD* (orange dashed line) and are also shown for reference. Then, the MV is applied using the maps produced with these methods. The *naive* case maps are used to produce the *MV-Naive* maps, and this is also the case for the other methods. Then the OA shown are the *MV-Naive* (green curve), the *MV-PAS* (pink curve), the *MV-EMD-PerClass* (brown curve) and finally the *MV-PAS-EMD* (purple curve).

By averaging the values over all source domains, the difference between the *naive* and the *PAS* is negligible. The use of transport alone improves performance by 2.5% and the combined use of transport and *PAS* increases performance by 5% compared to the *naive* case. This gain is found when using the MV with an average gap between 2 and 5% between the *MV-naive* and the *MV-PAS-EMD*

voting. The use of MV with the *PAS-EMD* increases the OA from 0.75 to 0.85, and is better than the other voting approaches. The confidence intervals are very narrow, which shows the robustness of the voting methods concerning the isolated classification error. All voting methods show the same behavior, a very high OA when two years are used then the OA decrease between 2 and 5% when 3 or more years are used.

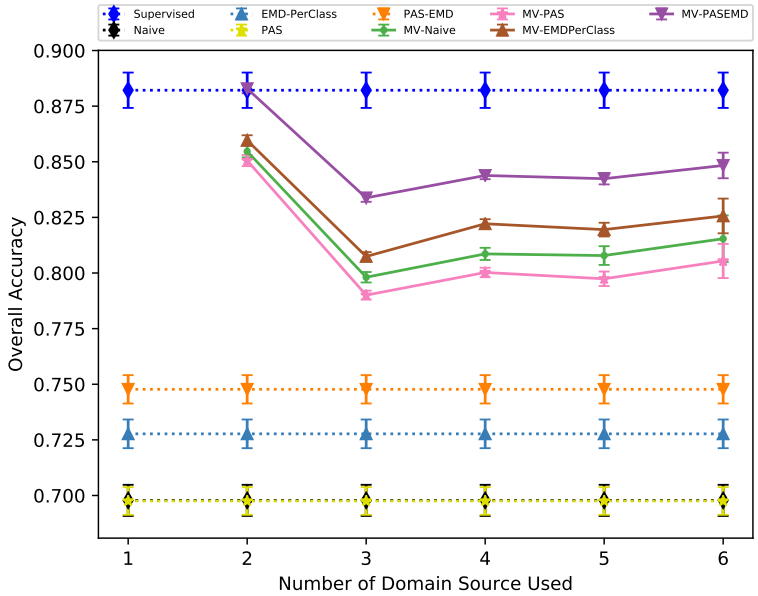

Figure 8. Overall Accuracy on the y-axis as a function of the number of source domain years used on the x-axis. Horizontal lines represent reference cases: *supervised* in blue, *naive* in black, *PAS* in yellow, *EMD-PerClass* in light blue and *PAS-EMD* in orange. Other curves represent the majority voting: *MV-Naive* in green, *MV-PAS* in pink, *MV-EMD-PerClass* in brown, *MV-PAS-EMD* in purple.

As introduced earlier, standard measures are insufficient for the evaluation of voting methods. Therefore, the OA obtained must be tempered by looking at the undecided pixels ratio. Figure 9 is a ratio graph, in which the x-axis represents the number of source domains used for voting, and the y-axis represents the pixel ratio between correctly classified (top plot), incorrectly classified (bottom plot), and undecided pixels (middle plot). The three plots in this figure do not have the same range, as the top one varies between 0.6 and 0.9%. The two bottom plots vary between 0 and 0.33%. The presented methods are the same as in Figure 8, with the same color code and legend. Each method has three curves associated, one for each category in a different subplot, except for the reference methods for which undecided pixels do not exist.

An interesting result highlighted by this figure is the percentage of incorrectly classified pixels by each method. Voting methods reduce the ratio of incorrectly classified pixels by about 10 to 20% compared to reference methods. When only two years are used, the MV allows a map to be produced with fewer errors than the supervised case, but the ratio of undecided pixels is very high. This explains the behavior of voting curves, as the ratio of undecided pixels is around 35% for two years and falls under 10% when more than three source domains are used.

This figure shows that the results of the *MV-PAS-EMD* are better than the others because it has the highest number of correctly classified pixels, the fewest undecided, and especially the fewest incorrectly classified pixels.

Figure 10 presents the FScore value for each class (y-axis) as a function of the number of input source domains (x-axis). The color bar starts with red hues for values near 0 and goes to blue hues for high FScore values near 1. There are four heat maps in this figure, one for each voting method. As expected, the FScore is always higher in the first column, due to the high number of undecided pixels obtained when two inputs are used for voting. This phenomenon is also observed for an even

number of inputs because it is a favorable situation for cases of indecision. This figure highlights the improvement of FScore compared to voting based on the *naive* case. The *MV-PAS-EMD* is better than the other methods with respect to OA, pixel ratio, and FScore.

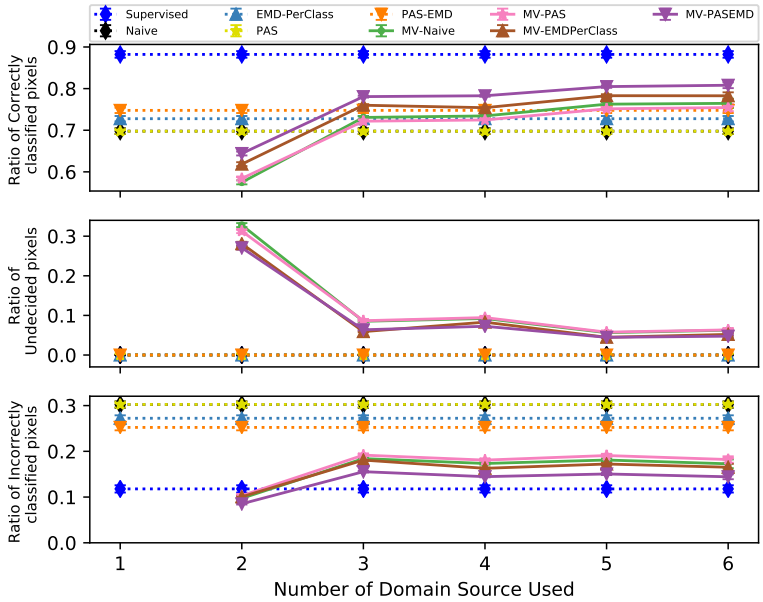

Figure 9. Pixel ratio between correctly classified (**top**), incorrectly classified (**bottom**) and undecided pixels (**middle**). The pixel ratio is displayed on the y-axis with different range and the number of source domains used on the x-axis. Horizontal lines represent reference cases: *supervised* in blue, *naive* in black, *PAS* in yellow, *EMD-PerClass* in light blue and *PAS-EMD* in orange. Other curves represent the majority voting: *MV-Naive* in green, *MV-PAS* in pink, *MV-EMD-PerClass* in brown, *MV-PAS-EMD* in purple.

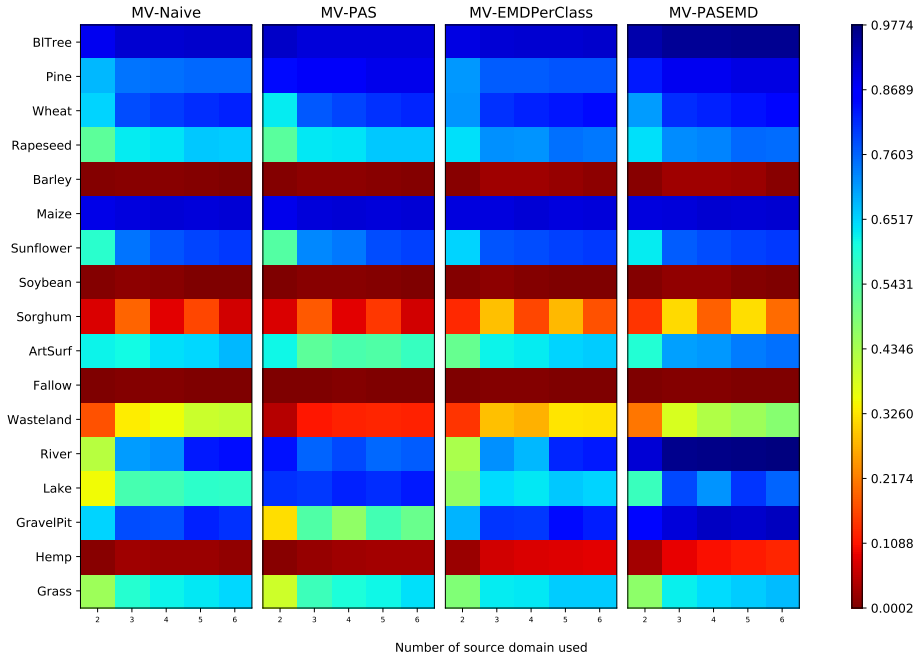

Figure 10. FScore per class (on the y-axis) as a function of the number of input source domains used (x-axis). The four graphs are *MV-Naive*, *MV-PAS*, *MV-EMD-PerClass* and *MV-PAS-EMD*, the name of method indicates the input maps used. The color bar indicates the average FScore value. The red hues indicate low FScore and blue hues indicate high FScore value.

4. Discussions

Figures 4 and 5 allow a quick comparison of the OA obtained using the different methods. Two types of analyses are carried out: first, baselines using reference data of both domains and, second, using only the reference data of the source domain. Baseline methods where samples are controlled allow defining the trends that can be expected from OT in the best cases. The *SUP-Sinkhorn* case thus highlights the inefficiency of the algorithm, which is explained by the complexity of the used data: there are too many classes for optimal minimization.

The management of samples is double-edged. On one hand, the number of samples must be high enough to represent each class correctly. On the other hand, this number must be small so that the cost matrices can be computed. Otherwise, the algorithm does not find the optimal solution because of the high number and the diversity of classes considered.

This interpretation is confirmed by the results obtained by *SUP-Sinkhorn-WL*, which shows that with sufficient simplifications the algorithm can provide robust transport with an improvement of about 5% of OA compared to the *naive* case, but more than 35% compared to the *SUP-Sinkhorn*. All these results confirm that our problem is too complex to use the Sinkhorn algorithm directly.

The EMD algorithm achieves satisfying results, with shorter processing time but it requires some help to be efficient, as shown using labels in both domains. The main impact of the use of labels is on the cost matrix, where for two different classes, the cost between samples is set to a very high value. As the EMD is a linear operation, the problem can be easily separated in C transport estimations, where C is the number of classes. This approach gives significant results and some cases of improvements, whereas the OA obtained by standard transport estimation, considering all classes at the same time, oscillates and remains close to the *naive* case. The main drawback of this approach is that a first classification is needed, and its quality significantly impacts transport efficiency.

If the time series of both domains are very different, the use of OT provides a correct gain regardless of the *naive* classification. For example, the results obtained for 2008 for which many dates are missing in winter, or for 2011 for which there is no image for August.

The best case is achieved when the initial classification is done in a supervised way, but this case does not need transport and is just studied here for comparison purposes.

The maximum OA of each method corresponds to a different year pair. That indicates that the essential element of transport is the selection of relevant samples.

The major drawback of the *naive* case is the loss of the minority classes which are drowned out by the majority classes in the training data sets. This is slightly corrected thanks to transport. However, if two classes are too similar (for instance wheat and barley or water classes), the main difficulty remains to separate them. Indeed, the analysis of the confusion matrices shows that most of the errors made by the classifiers remain consistent. For instance, vegetation classes are confused with grass, or several pixels in lake boundaries are classified as gravel pit. The FScore analysis gives good insight of the advantage of using transport. This result is confirmed by the analysis of the impact of the per-class EMD on the MV FScore (Figure 10), showing that despite the slight increase in FScore per class for minority classes, the voting predicts some samples as members of these classes. That means that using OT allows keeping the minority and hard to classify classes during the processing and paves the way for an iterative method, for instance, to improve the prediction of these classes.

These conclusions are confirmed by the combined use of the *PAS* and the *EMD-PerClass* approaches. The OA performances are improved in most of the cases, and the FScore values are also improved in comparison to the *naive* case. The improvement of FScore is relative to the initial value, but in most cases, this approach provides excellent results.

A limitation of the PAS approach is shown by the FScore analysis as the lower diagonal of the artificial surface class FScore matrix presents a degradation of FScore values. This result is specific to this class and can be explained by its nature.

Indeed, unlike other classes, artificial areas are more likely to appear at a point in time or remain stable, rather than transition to another class. In particular, this means that using a position indicating

a building from a previous year will nearly always be valid for a future year, but the opposite is not systematically exact. Not all samples used for transport and training phases are valid, which explains the observed performance.

The results obtained by the OT only are not very satisfactory and provide, in the best cases, a relative gain of 5% of OA. The use of voting methods allows several source domains to be jointly used and therefore benefits from the gain due to transport to improve the performance of voting methods. Figures 8 and 9 show that this gain remains minimal given the workload involved in the effective use of OT. Again, the performances are boosted by the combined use of *PAS* and *EMD-PerClass*, allowing an improvement in OA near to 10% in comparison to the *naive*-based voting. This approach provides a better OA with mostly correctly classified pixels and fewer undecided and incorrectly classified pixels. The results of this study provide a negative overview of the use of OT, which is mainly due to the quality of the dataset, and its strong heterogeneity. The 2008 case is significant because the absence of images for a long duration (see Table 1) makes the use of this year as a source domain difficult. It is, therefore, evident that in similar cases, the use of OT makes the improvement of the classification accuracy possible. Because of all the difficulties encountered for using OT for land-cover mapping, the relative improvement of 5% OA is a therefore positive result.

In the particular case where only one source domain is available, the use of transport is recommended because it allows a more accurate classification to be produced. On the other hand, if two or more years are available, producing naive classifications and fusing them with a simple MV is faster and produces a map of similar quality and therefore, the trade-off between complexity and accuracy must be considered.

5. Conclusions

In this work, the problem of using DA and more particularly OT in the case of operational land-cover mapping when no reference data for the mapping period exists is addressed. Most of the algorithms in the literature are applied to small data sets (low number of samples, low data dimensionality, or low number of classes). In this paper, we proposed baseline methods to show the limits of the standard OT methods and their impact on the OA of the produced land-cover maps. These baselines show that the standard use of such methods produces lower accuracies than the use of the *naive* classification, which uses past data for training and applies it on a different period time series. The Sinkhorn algorithm is for example not able to handle a problem with 17 classes and a satellite image time series with 108 features. We also showed that the sample selection for OT estimation is an essential element of the transport efficiency. Therefore, several sampling schemes are proposed, coupled with strategies to reduce the problem complexity and help the OT to find an optimal coupling. PAS EMD transport estimation gives significant results, allowing improvement of the *naive* case up to 10% of OA. To this end, first, a *naive* classification is produced to identify the samples in the target domain. The results obtained after OT estimation depend on the quality of this *naive* classification. The measured improvements for the other methods are not significant enough for the operational use of OT, as OA obtained remain close in the *naive* case confidence intervals. The thematic analysis with the FScore shows a slight improvement in the prediction of the minority classes. A favorable way of improving the use of OT is to separate annual and perennial classes and consider only the annual classes for transport. This simple strategy reduces the number of classes considered during the transport estimation and leverages the stability of perennial classes to improve the classification. In this work, the importance of the selected samples for OT estimation is shown, and a particular work should be done to find efficient sample selection methods. All the obtained results are encouraging because the recognition of minority classes is a very complicated task, especially when only past reference data are used. In future works, the use of OT could be considered to be the first step in an iterative process. However, for operational contexts, the computational costs of OT still seem too high for the small gain in accuracy that they bring. Therefore, it would be interesting to investigate the possibility of performing (approximate) OT with lower computational burden.

Author Contributions: Conceptualization, B.T., J.I., and J.M.; Formal analysis, B.T. and J.I.; Investigation, B.T.; Methodology, B.T., J.I., and J.M.; Project administration, J.I. and J.M.; Supervision, J.I. and J.M.; Validation, B.T., and J.I.; Visualization, B.T.; Writing—original draft, B.T. and J.I.; Writing—review & editing, B.T., J.I., and J.M.

Funding: This research was funded by CNES and the region Occitanie grant number 2442 of 22th June 2015. The APC was funded by CNRS.

Acknowledgments: This work is funded by the French space agency (CNES) and the Région Occitanie, and it has been carried out at CESBIO laboratory. The authors would like to thank CESBIO colleagues for their help in data collection and reviewing of this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

OT	Optimal Transport
EMD	Earth's Movers Distance
PAS	Perennial Annual Split
SUP	SUPervised i.e., the use of reference data of both domains for samples selection
SUP-EMD-WL	SUPervised EMD With use of Labels
SUP-Sinkhorn-WL	SUPervised Sinkhorn With use of Labels
SUP-EMD	SUPervised EMD without use of labels
SUP-Sinkhorn	SUPervised Sinkhorn without use of labels
SP-EMD	Same Position for target samples than source samples used with EMD algorithm
SP-Sinkhorn	Same Position for target samples than source samples used with Sinkhorn algorithm
MV	Majority Voting. All method with MV in prefix is a voting method using the result of this method as inputs

References

1. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **2013**, *342*, 850–853. [CrossRef]
2. Dewan, A.M.; Yamaguchi, Y. Land Use and Land Cover Change in Greater Dhaka, Bangladesh: Using Remote Sensing To Promote Sustainable Urbanization. *Appl. Geogr.* **2009**, *29*, 390–401. [CrossRef]
3. Jung, M.; Henkel, K.; Herold, M.; Churkina, G. Exploiting Synergies of Global Land Cover Products for Carbon Cycle Modeling. *Remote Sens. Environ.* **2006**, *101*, 534–553. [CrossRef]
4. Srivastava, P.K.; Han, D.; Rico-Ramirez, M.A.; Bray, M.; Islam, T. Selection of Classification Techniques for Land Use/land Cover Change Investigation. *Adv. Space Res.* **2012**, *50*, 1250–1265. [CrossRef]
5. Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience With Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
6. Waske, B.; Braun, M. Classifier Ensembles for Land Cover Mapping Using Multitemporal Sar Imagery. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 450–457. [CrossRef]
7. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
8. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
9. Congalton, R.G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [CrossRef]
10. Wang, L.; Sousa, W.P.; Gong, P. Integration of Object-Based and Pixel-Based Classification for Mapping Mangroves With Ikonos Imagery. *Int. J. Remote Sens.* **2004**, *25*, 5655–5668. [CrossRef]
11. Pelletier, C.; Valero, S.; Inglada, J.; Champion, N.; Sicre, C.M.; Dedieu, G. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping With Satellite Image Time Series. *Remote Sens.* **2017**, *9*, 173. [CrossRef]
12. Inglada, J.; Vincent, A.; Arias, M.; Tardy, B.; Morin, D.; Rodes, I. Operational High Resolution Land Cover Map Production At the Country Scale Using Satellite Image Time Series. *Remote Sens.* **2017**, *9*, 95. [CrossRef]

13. Tardy, B.; Inglada, J.; Michel, J. Fusion Approaches for Land Cover Map Production Using High Resolution Image Time Series Without Reference Data of the Corresponding Period. *Remote Sens.* **2017**, *9*, 1151. [CrossRef]
14. Bruzzone, L.; Prieto, D. Unsupervised Retraining of a Maximum Likelihood Classifier for the Analysis of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 456–460. [CrossRef]
15. Persello, C.; Bruzzone, L. Kernel-Based Domain-Invariant Feature Selection in Hyperspectral Images for Transfer Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2615–2626. [CrossRef]
16. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
17. Rajan, S.; Ghosh, J.; Crawford, M. Exploiting Class Hierarchies for Knowledge Transfer in Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 3408–3417. [CrossRef]
18. Kumar, S.; Ghosh, J.; Crawford, M.M. Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis. *Pattern Anal. Appl.* **2002**, *5*, 210–220. [CrossRef]
19. Bruzzone, L.; Persello, C. A Novel Approach To the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3180–3191. [CrossRef]
20. Ham, J.; Chen, Y.; Crawford, M.; Ghosh, J. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [CrossRef]
21. Petitjean, F.; Weber, J. Efficient Satellite Image Time Series Analysis Under Time Warping. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1143–1147. [CrossRef]
22. Inamdar, S.; Bovolo, F.; Bruzzone, L.; Chaudhuri, S. Multidimensional Probability Density Function Matching for Preprocessing of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1243–1252. [CrossRef]
23. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3550–3564. [CrossRef]
24. Bailly, A.; Chapel, L.; Tavenard, R.; Camps-Valls, G. Nonlinear Time-Series Adaptation for Land Cover Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 896–900. [CrossRef]
25. Patel, V.M.; Gopalan, R.; Li, R.; Chellappa, R. Visual Domain Adaptation: A Survey of Recent Advances. *IEEE Signal Process. Mag.* **2015**, *32*, 53–69. [CrossRef]
26. Courty, N.; Flamary, R.; Tuia, D. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 274–289.
27. Kantorovitch, L. On the translocation of masses. *Manag. Sci.* **1958**, *5*, 1–4. [CrossRef]
28. Hagolle, O.; Huc, M.; Pascual, D.V.; Dedieu, G. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sens. Environ.* **2010**, *114*, 1747–1755. [CrossRef]
29. Inglada, J.; Arias, M.; Tardy, B.; Hagolle, O.; Valero, S.; Morin, D.; Dedieu, G.; Sepulcre, G.; Bontemps, S.; Defourny, P.; et al. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sens.* **2015**, *7*, 12356. [CrossRef]
30. Igel, C.; Heidrich-Meisner, V.; Glasmachers, T. Shark. *J. Mach. Learn. Res.* **2008**, *9*, 993–996.
31. Lam, L.; Suen, S. Application of Majority Voting To Pattern Recognition: an Analysis of Its Behavior and Performance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **1997**, *27*, 553–568. [CrossRef]
32. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300, arXiv:1306.0895.

33. Flamary, R.; Courty, N. POT Python Optimal Transport Library. 2017. Available online: https://github.com/rflamary/POT (accessed on 1 May 2019).
34. Tuia, D.; Flamary, R.; Rakotomamonjy, A.; Courty, N. Multitemporal classification without new labels: A solution with optimal transport. In Proceedings of the 2015 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp), Annecy, France, 22–24 July 2015.