Analysis of multi-temporal classification techniques for forecasting image times series

R. Flamary, M. Fauvel, M. Dalla Mura, and S. Valero

***Abstract*—The classification of an annual times series by using data from past years is investigated in this paper. Several classification schemes based on data fusion, sparse learning and semi-supervised learning are proposed to address the problem. Numerical experiments are performed on a MODIS image time series and show that while several approaches have statistically equivalent performances, SVM with ℓ_1 regularization leads to a better interpretation of the results due to their inherent sparsity in the temporal domain.**

***Index Terms*—satellite image time series, classification, transfer learning**

I. Introduction

The study of the evolution of the Earth's environment has been boosted by the increasing availability of image time series. These images provide ways to detect and monitor changes over time and with a large spatial coverage, necessary characteristics for a broad range of applications [1]. Recently, the interest on multi-temporal image analysis has increased due to several facts such as: i) the growing amount of archive data (e.g., MODIS started acquiring in 2000 with a daily frequency); ii) the launch of new satellites with short revisit time, which allow the acquisition of longer time series; and iii) the spread of free and open policies for the distribution of satellite images (e.g. ESA Sentinel).

The land-cover classification of image time series for developing annual land-cover maps is one of the main challenges in the remote sensing community devoted to the monitoring of the landscape. In general, the classification of image time series for land cover/use determination is done at a pre-defined temporal scale such as years [2], seasons [3], composite periods (e.g., when using the 8/16 days or monthly products for MODIS [4], [5]) or even at daily acquisitions. In this scenario, a thematic class is assigned to each pixel in the scene for the whole period chosen in the analysis (assuming no change in the land cover during the period).

In the literature, many works have demonstrated that supervised learning approaches are suitable for the analysis of these data since they can produce classification results with high accuracy [6]. However, the supervised learning step requires the availability of training pixels to learn the parameters of the model. The number of referenced pixel as well as the quality of the labeling will strongly condition the training process. For instance, a limited number of training samples will result in poor results in terms of classification accuracy [7]. In remote sensing applications, it is often difficult to get enough samples for each class of interest. When considering the supervised land cover classification of image time series, the training phase relies in general on samples belonging the same temporal interval (i.e., a subset of the samples belonging to the the same time series with known label) [2], [3], [4], [5]. However, referenced data may become available a long time after the acquisition of the time series and users have to wait for classifying their images.

In the recent years some works have introduced methodologies based on transfer learning in order to address the above-mentioned issues in the context of updating land-cover maps [8]. Transfer learning aims at adapting to a new unlabeled image the classification model learned on another image (e.g., an image acquired previously with labeled data) [6, Chapter 9]. In these works, the adaptation of the classifier relies on the detection of the changes occurred between an image already known by the classifier and the new one. The aforementioned approaches perform the classification of a single new image and are typically adapted for high (metric) spatial resolution images, which in general are not available with a high temporal resolution.

The problem of multi-year classification is addressed in this work, that is the classification of an entire time series over one year without any prior information belonging to the same time series. In this context, the classification of the current year is still performed in a supervised scheme but by only exploiting the prior information available on the image times series coming from the previous years (i.e., the ground truth as well as the already learned classifiers). This problem is very challenging due to the non stationarity of the time series especially on natural landscapes. Indeed, the temporal signatures of vegetated areas show a significant variability among years [9], [10] due to i) the intrinsic variability given by the dependency of the phenology on the current season and ii) abrupt changes due to external factors such as the rotation of cultures, deforestation, fires and droughts. From the point of view of the application the scenario taken into account in this paper is of particular interest for the tasks requiring annual land cover maps [11] such as for the update of census maps of land use done yearly and the monitoring of the evolution of natural landscapes. In this specific context, Dynamic Time Warping (DTW) was proposed in [12] to classify image times

R. Flamary is with Laboratoire Lagrange, UMR 7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, Nice, France

M. Fauvel is with Université de Toulouse, INP-ENSAT, UMR 1201 DYNAFOR, France and INRA, UMR 1201 DYNAFOR, France

M. Dalla Mura is with GIPSA-lab, Grenoble Institute of Technology, France

S. Valero is with CESBIO, UMR CNRS 5126, ISIS, Toulouse, France

This work was supported by GdR ISIS/GRETSI; AMOR project. The authors gratefully acknowledge J.-P. Denux from Laboratoire de Teledetection et de Gestion des Territoires - DYNAFOR, for providing the MODIS time serie and the corresponding ground truth.

series by using satellite data captured at previous years. In it, the DTW approach is evaluated on high spatial resolution image times series composed by a short number of images with a different temporal sampling. Similarly, a kernel using DTW was proposed in [13].

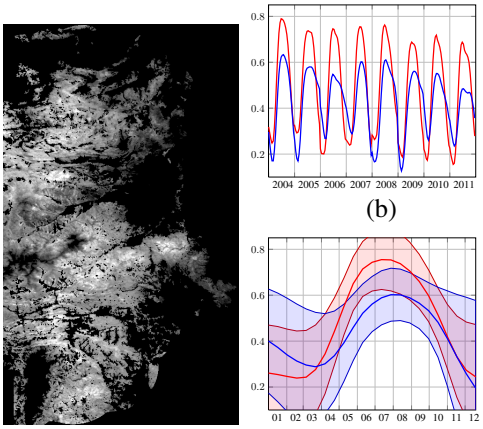
In this paper, we focus particularly on the classification step restricting the problem to data that are regularly sampled in the temporal domain and to long time series (more than 10 acquisitions per year) having a moderate spatial resolution. The contribution of this letter is the investigation of several classification strategies in multiyear classification. Roughly speaking, the past ground truth is used to learn one or several models (one per year or one for all the years) and several transfer strategies are investigated. The obtained results in terms of classification accuracy are compared to those obtained using the actual ground truth.

The remainder of the paper is organized as follows. The multitemporal data are introduced in Section II. The different classification schemes studied here are discussed in Section III. Experimental results are presented in Section IV where the performances of the different classification strategies are discussed. Finally, conclusions are drawn in Section V.

II. Multitemporal data set

Image time series of normalized difference vegetation index (NDVI) are considered. The NDVI is an index computed by a ratio between two spectral bands, which is used as an important descriptor for the characterization of the vegetated areas on a scene. The NDVI image sequences used here were produced by the USGS center of NASA from images acquired by the sensor MODIS [14]. The exploitation of vegetation indexes such as the NDVI has already largely proved their importance in crop classification allowing one to distinguish different crop types based on their phenology [11]. Each used times series is composed by 23 images per year having a spatial resolution of 250m. The entire data set spans the 2004-2011 period, totaling 184 images over 8 years. Fig.1(a) shows an example of NDVI image captured on 2007-05-25.

In the experimental analysis we considered the problem of discrimination between the class "Broad-leaved tree" and "Pine" that are of particular interest for forestal studies. 4481 and 3410 pixels per year have been labeled by experts agronomist for the two classes, respectively. Their mean NDVI values for the 2004-2012 period is reported in Fig. 1(b), whereas Fig. 1(c) shows the mean value of each considered class at one specific year. The shaded region of Fig. 1(c) represents the standard deviation around the mean value. This last figure illustrates the high degree of overlap that exists on most of the year for these two classes, making this a very challenging classification problem. From the Fig. 1(b), the periodic pattern of Fig. 1(c) can be observed over the years. Besides, it can be corroborated that small variations are present from one year to the other, both in terms of amplitude and phase.



(a) (b) (c)

Fig. 1. (a) NDVI value for the date 2007-05-25. White value indicate high value of NDVI (i.e., presence of vegetated areas). Black pixels correspond to pixels that are not included in the study site. (b) Mean NDVI value for the "Broad-leaved tree" (in red) and "Pine" classes (in blue) over the period 2004-2011. (c) Mean ($\pm$ standard deviation) of the NDVI values of the "Broad-leaved tree" (in red) and "Pine" classes (in blue) for the 23 images in the year 2007.

III. Methods for multitemporal image classification

The objective is to classify each pixel given its observations along the current year $T+1$ and assuming that every pixel can belong to a unique class for the whole year (legitimate assumption for the classes considered). The training data set consists in n samples $\{\boldsymbol{x}_i, y_i\}_{i=1\ldots n}^{t}$ for each previous year $t = 1 \ldots T$, where $\boldsymbol{x}_i$ represents the temporal signature of a pixel and y_i their corresponding class. The features in $\boldsymbol{x}_i \in \mathbb{R}^d$ correspond to d regular sampling along the year of a given observed characteristic (i.e., here the NDVI). Note that while the number n of training samples per year is fixed and equal along the years in the available data, the analysis proposed in the following holds with n varying along the years.

This work focuses on the linear Support Vector Machines (SVM) classifier. In the last years, SVM became a state of the art method in remote sensing thanks to its ability to handle complex and possibly non linear discrimination. Nevertheless, recent works have highlighted the computational advantage of learning a linear classifier while estimating a non linear feature extraction [15] which do not require the computationally intensive use of kernels. The same classifier (SVM) was used for all the classification approaches for a fair comparison.

All the SVM approaches presented here are focused on linear classification of the form $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ and they can be divided in three classification strategies described in the following and illustrated in Fig. 2. Note that the linear model treats the time serie as a vector but still takes into account the temporal correlation between the features through the global optimization problem.

A. 1st Strategy: SVM learned on multiple years

A straightforward way to deal with the non-stationarity along the years is to learn a single SVM classifier for all the referenced pixels over the years. This is done by concatenating

all the training samples of all years creating a large data set of nT training samples. The parameter estimation is performed by minimizing the following problem

$$\min_{f} \quad R(f) + C \sum_{i=1}^{nT} H(y_i, f(\boldsymbol{x}_i)) \tag{1}$$

where $H(y_i, f(\boldsymbol{x}_i)) = \max(0, 1 - y_i f(\boldsymbol{x}_i))^2$ is the squared hinge loss and C is a regularization parameter. Both ℓ_2 and ℓ_1 regularization have been used in this work, i.e., $R(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ or $R(\cdot) = \|\cdot\|_1$ that both penalize large values of w_j, $j \in \{1, \ldots, d\}$. The latter has been widely used for sparse learning [16]. It has the advantage to provide parsimonious solutions in terms of w_j, i.e., some of them are exactly zero. In the context of this study, having sparse coefficients w_j means that some dates of the time series can be considered as not relevant, therefore, they are not used for the classification process.

By concatenating all the years, it is expected that if the number of years available for training is large enough, the non-stationarity will be correctly modeled as an additional variability of the data. Accordingly, the full set of available samples obtained from the past years is used here with the two SVM approaches. In the following, this will be referred to as SVM_{full} for the quadratic variant and $\text{SVM}_{\text{full}_{\ell_1}}$ for the ℓ_1 regularization variant. It must be remarked that any classifier could be used in this scheme (i.e., by considering all the data set available). For this reason, the use of Gaussian Mixture Models, denoted as GMM_{full}, has also been studied in our numerical experiments [7, Chapter 4].

B. 2nd Strategy: Semi supervised learning

The second approach considered in order to deal with the non-stationarity along years is the semi-supervised learning (SSL) scheme. SSL has been proposed in the literature to deal with data affected by spatial non-stationarity, which is defined by the dependence of a statistical model over its definition space. The base concept of SSL is that unlabeled pixels are used in the training step process [17]. Accordingly, in the context of time series classification, the unlabeled pixels are taken from the year to be classified, while the labeled pixels comes from the past observations.

Two standard semi-supervised schemes are investigated in this work that both rely on the following optimization problem:

$$\min_{f} \quad \frac{1}{2}\|f\|_2^2 + C \sum_{i=1}^{nT} H(y_i, f(\boldsymbol{x}_i)) + C_t \Omega(f) \tag{2}$$

where $\Omega(f)$ is a semi-supervised regularization term that takes into account both the labeled and unlabeled pixels and C_t is a parameter that weight the impact of the regularization term.

The first semi-supervised approach is the graph regularization SSL ($\text{SVM}_{\text{graph}}$) [17]. It consists in constructing a graph $G_{i,j}$ of the pixels and promotes similar prediction score for similar pixels, i.e., close in terms of distance in the feature space. The regularization term is defined as $\Omega_{graph}(f) = \sum_{i,j}^{n',n'} G_{i,j}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2$ where $n' = nT + u$ is the total number of samples of the past years augmented by the u unlabeled samples from the actual year and $G_{i,j}$ is computed on all (labeled and unlabeled) examples. Typically a Gaussian kernel is used for $G_{i,j}$ in order to promote a manifold regularization on the classifier f. In order to limit the computational burden, the graph is constructed considering a k-means clustering performed on a subset of the data.

The second SSL approach investigated is the Transductive SVM ($\text{SVM}_{\text{trans}}$), which promotes the decision boundaries to be in areas of the feature space with low density of unlabeled samples [18]. Theoretically, the $\text{SVM}_{\text{trans}}$ can be used for adapting a classifier to a (slight) temporal change in the density of the data set. The corresponding regularization term is defined as $\Omega_{trans}(f) = \sum_{i=nT+1}^{n'} \exp\left(-3f(\boldsymbol{x}_i)^2\right)$ associated to the constraint $\frac{1}{n}\sum_{i=nT+1}^{n'} f(\boldsymbol{x}_i) = \frac{1}{nT}\sum_{i=1}^{nT} y_i$ that avoids all the unlabeled samples to be classified to the same class. It is clear from the previous expression that the classifier will induce large values for the prediction function on unlabeled pixels hence promoting a margin with few samples.

C. 3rd Strategy: Ensemble of classifiers

The third strategy presented here investigates the use of an ensemble of classifiers. Each year is considered as a separate source of information, therefore, a dedicated SVM is learned for each year. Hence, for T annual image times series coming from the past, T different classifiers $f_1, f_2, ..., f_T$ are learned. The classification of the actual year is performed by fusing the decisions provided by each single classifier f_t.

First, the absolute decision rule used in [19] is investigated to fuse the SVMs outputs. Taking into account the T classifier results, the decision rule is performed by extracting the prediction score for each class as $f_{T+1}(\boldsymbol{x}) = f_t(\boldsymbol{x})$ such as $|f_t(\boldsymbol{x})| > |f_1(\boldsymbol{x})|, \ldots, |f_T(\boldsymbol{x})|$.

However, the maximum operator leads to a disjunctive fusion, i.e., an optimistic fusion that is sensible to noise. For this reason, a second fusion approach is proposed by using an adaptive weighting scheme SVM_{weight}. The weight ν_t reflects the similarity between two years. It is computed as the power of the correlation between two time series $r(t,t')^q$, q being a positive parameter that allows to control the influence of each year in function of the correlation. The fusion rule is

$$f_{T+1}(\boldsymbol{x}) = \frac{\sum_{t=1}^{T} r(T+1,t)^q f_t(\boldsymbol{x})}{\sum_{t=1}^{T} r(T+1,t)^q}. \tag{3}$$

When $q = 0$, eq. (3) is the mean value over t of $f_t(\boldsymbol{x})$ and for $q \to +\infty$, eq. (3) tends to select only the most correlated year.

An illustration of the different approaches on a 2D toy example can be seen in Figure 2.

IV. Numerical experiments

The performances in terms of classification accuracy of the different strategies are evaluated in this section. In order to perform experiments with the same number of past years for predicting the actual year, a Leave-One-Year-Out (LOYO) validation scheme has been set up: a given year is withheld from the set, and the prediction is done using all the remaining years. Then the procedure is iterated for all the available

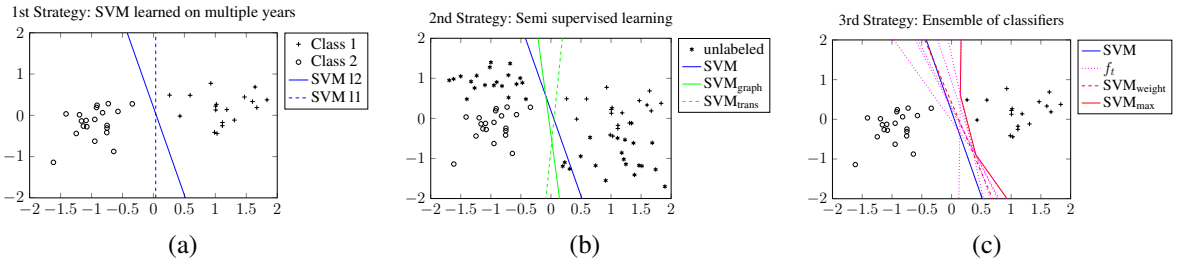


Fig. 2. Illustration of the different classification strategies on a simple 2D example. (a) Conventional SVM and its ℓ_1 counterpart: the second variable not selected with the ℓ_1 regularization since the variable is not discriminative, therefore the hyperplane is vertical. (b) Semi-supervised strategy: some unlabeled samples are added to the optimization problem and that results is a linear transformation of the hyperplane. (c) Ensemble of classifiers: several hyperplanes f_t are available and a fusion scheme (weighted mean or max rule) is applied. Note that with the max's rule the separating surface is now piecewise linear.

years and the final classification accuracy is obtained by averaging all the results. This validation strategy was already employed in [10] for examining the variability in year-to-year classification performance of MODIS time series and it is the typical scheme when assessing the performances of crop yield forecasting [20], [21].

Furthermore, a SVM, denoted as SVM_{curr}, is trained also on 500 labeled pixels from the year under investigation in order to get baseline results.

For each method, a cross validation is performed on the training data in order to automatically select the best regularization parameter. The best values for the parameter C and C_t for the semi-supervised approaches were selected over a grid ranging form 10^5 to 10^{10} and from 10^0 to 10^4, respectively. The parameter q has been set to 20 in all the experiments after some empirical runs. In order to limit the computational overload of the graph regularization method, the regularization graph was computed on a subset of 1000 samples (≈ 500 per class) picked randomly on the full set (past and current year) and using a k-means clustering with $k = 10$ clusters considering Euclidean distances for the construction the regularization graph. For the GMM, mean vectors, covariance matrices and priors are estimated by maximizing the likelihood.

In order to assess the statistical significance of the observed differences in terms of classification accuracy, a Wilcoxon rank test has been used on the difference of the classification rate over the 50 repetitions. If the median value of differences is zero, the results are said to be non significanlty different.

A. Comparison of classification strategies

The performances of the different strategies are evaluated by computing the recognition rate across the different years. The obtained results can be compared by looking at Figure 3.

A second evaluation of the obtained results is shown in Table I, where the recognition rate averaged along years and the p-value evaluation measure can be observed.

It can be seen that the best performing methods in the LOYO validation scheme are the SVM_{full} with both quadratic or ℓ_1 regularization, $\text{SVM}_{\text{graph}}$, SVM_{max} and $\text{SVM}_{\text{weight}}$. Despite the fact that those methods perform equally well, we want to emphasize that SVM_{full} is less computationally expensive than $\text{SVM}_{\text{graph}}$ suggesting the use of the former. Comparing the results obtained by the SSL approaches leads to interesting results. The graph regularization approach $\text{SVM}_{\text{graph}}$ does not

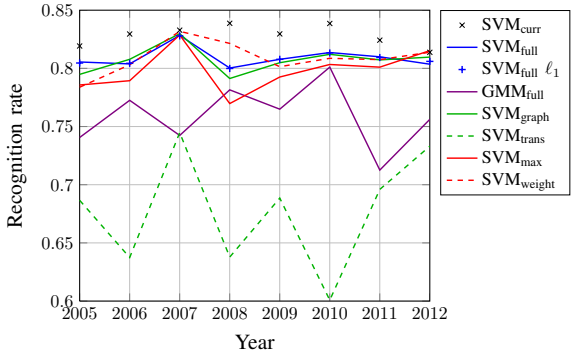


Fig. 3. Performance of the different methods along the years.

TABLE I
PERFORMANCES OF THE DIFFERENT APPROACHES RECOGNITION RATE AVERAGED ALONG YEARS. THE P-VALUE IS COMPUTED *w.r.t.* THE BEST PERFORMING METHOD (SVM_{weight}). THE METHODS HIGHLIGHTED IN BOLD ARE ALL STATISTICALLY EQUIVALENT WITH A P-VALUE THRESHOLD $\alpha < 0.05$.

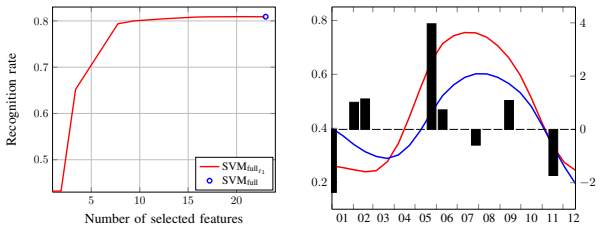
Method	Rec. rate	p-value
SVM_{curr}	0.8283	0.0078
SVM_{full}	**0.8091**	-
$\text{SVM}_{\text{full}_{\ell_1}}$	**0.8091**	0.5938
GMM_{full}	0.7590	0.0078
$\text{SVM}_{\text{graph}}$	**0.8072**	0.4844
$\text{SVM}_{\text{trans}}$	0.6781	0.0078
SVM_{max}	**0.7982**	0.0781
$\text{SVM}_{\text{weight}}$	**0.8089**	0.8438

yield any increase in performances, however, the Transductive $\text{SVM}_{\text{trans}}$ is clearly limited. This can be related to the equality constraint that might limit the adaptation of the variability among the years. However, besides the encouraging results, performances are significantly inferior to the SVM learned on the current year (baseline).

B. Feature selection via $\text{SVM}_{\text{full}_{\ell_1}}$

Table I shows that $\text{SVM}_{\text{full}_{\ell_1}}$ has the same performance than SVM_{full}. As the former is a sparsity inducing method, this suggests that not all 23 temporal features available each year are equally discriminant. For this reason, a short analysis of the different levels of sparsity is presented in this section.

The performances and sparsity of $\text{SVM}_{\text{full}_{\ell_1}}$ are investigated when selecting 0 features to all features. The average performances of $\text{SVM}_{\text{full}_{\ell_1}}$ in LOYO validation is plotted as a function of the number of selected features in Figure 4(a). The



(a) (b)

Fig. 4. (a) Prediction performances as a function of the number of selected features. (b) Coefficient value of the SVM$_{\text{full}_{\ell_1}}$ when 8 dates are retained (right axis) and mean spectra for the year 2007 for the two considered classes.

performance of SVM$_{\text{full}}$ using all 23 features is also reported as a blue circle. The red line shows how the performances are nearly constant when more than 15 features are selected. It suggests that only around 65% of the temporal samples are relevant to the classification task. It is also seen that a slight loss in performances of SVM$_{\text{full}_{\ell_1}}$ is obtained by selecting down to 8 features.

The Figure 4(b) shows the weights values of the SVM$_{\text{full}_{\ell_1}}$ model when 8 features are selected. The horizontal axis represents the 12 months of an annual NDVI times series. The weights of the 8 features are plotted on the mean NDVI values of Figure 1(c). The most important weight is given for the time "2007-05-25", which corresponds to an high growth of the trees: leafs are appearing in that period and the difference in terms of NDVI between Broad-leaved tree and Pine is high. Besides, some features are also selected in winter: the canopy of Broad-leaved is lost during this period whilst the conifers (e.g., the Pine trees) keeps it. This corroborates that there is an important correlation between the weights values of SVM$_{\text{full}_{\ell_1}}$ and the nature of the data. Note that in some cases (May and February), temporal neighbors have been both selected despite their strong temporal correlation, this highlights the ability of the ℓ_1 regularization to select correlated discriminant features in order to promote performances.

V. Conclusion

In this paper, the problem of optical remote sensing image time series classification has been considered. Different strategies aiming at classifying multi-temporal NDVI satellite image times series by using past prior information have been proposed. The classification of non-stationary data among years has been investigated by several approaches. In this context, three different classification schemes exploiting the capabilities of linear SVM classifier have been studied. A comparative analysis has been performed by several experiments addressing a two-class problem.

Experiments have shown that SVM$_{\text{full}}$ obtains the best results. However, from a statistical point of view, it has been seen that other strategies such as SVM$_{\text{graph}}$, SVM$_{\text{max}}$ and SVM$_{\text{weight}}$ obtain equivalent results. Furthermore, it has been analyzed how the coefficients obtained by SVM$_{\text{full}_{\ell_1}}$ allows the interpretation of the nature of the data.

References

[1] F. Bovolo, L. Bruzzone, and R. King, "Introduction to the special issue on analysis of multitemporal remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, pp. 1867–1869, April 2013.

[2] D. Muchoney, J. Borak, H. Chi, M. Friedl, S. Gopal, J. Hodges, N. Morrow, and A. Strahler, "Application of the MODIS global supervised classification model to vegetation and land cover mapping of central america," *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1115–1138, 2000.

[3] J. O. Sexton, D. L. Urban, M. J. Donohue, and C. Song, "Long-term land cover dynamics by multi-temporal classification across the Landsat-5 record," *Remote Sensing of Environment*, vol. 128, no. 0, pp. 246 – 258, 2013.

[4] M. Friedl, D. McIver, J. Hodges, X. Zhang, D. Muchoney, A. Strahler, C. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao, and C. Schaaf, "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of Environment*, vol. 83, no. 12, pp. 287 – 302, 2002.

[5] H. Carrao, P. Gonalves, and M. Caetano, "Contribution of multispectral and multitemporal information from MODIS images to land cover classification," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 986 – 997, 2008.

[6] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*. Wiley Online Library, 2009.

[7] D. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New Jersey: John Wiley and Sons, 2003.

[8] B. Demir, F. Bovolo, and L. Bruzzone, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 300–312, 2013.

[9] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang, "MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 168 – 182, 2010.

[10] J. C. Brown, J. H. Kastens, A. C. Coutinho, D. de Castro Victoria, and C. R. Bishop, "Classifying multiyear agricultural land use data from mato grosso using time-series {MODIS} vegetation index data," *Remote Sensing of Environment*, vol. 130, no. 0, pp. 39 – 50, 2013.

[11] B. D. Wardlow, S. L. Egbert, and J. H. Kastens, "Analysis of time-series MODIS 250m vegetation index data for crop classification in the u.s. central great plains," *Remote Sensing of Environment*, vol. 108, no. 3, pp. 290 – 310, 2007.

[12] F. Petitjean, J. Inglada, and P. Gancarski, "Satellite image time series analysis under time warping," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, p. 3081, 2012.

[13] P. Dusseux, T. Corpetti, and L. Hubert-Moy, "Temporal kernels for the identification of grassland management using time series of high spatial resolution satellite images," in *Geoscience and Remote Sensing Symposium, IGARSS 2013*, pp. 3258–3260, July 2013.

[14] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," *Remote Sensing of Environment*, vol. 83, pp. 195–213, Nov. 2002.

[15] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse SVM," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 52, no. 10, pp. 6062–6074, 2014.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[17] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, 2007.

[18] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*, vol. 2. MIT press Cambridge, 2006.

[19] M. Fauvel, J. Chanussot, and J. Benediktsson, "A combined support vector machines classification based on decision fusion," in *Geoscience and Remote Sensing Symposium, IGARSS 2006*, pp. 2494–2497, July 2006.

[20] D. K. Bolton and M. A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agricultural and Forest Meteorology*, vol. 173, no. 0, pp. 74 – 84, 2013.

[21] D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the united states," *Remote Sensing of Environment*, vol. 141, no. 0, pp. 116 – 128, 2014.