Remote Sensing of Environment 187 (2016) 156–168

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas

Charlotte Pelletier[a,*], Silvia Valero[a], Jordi Inglada[a], Nicolas Champion[b], Gérard Dedieu[a]

[a] CESBIO-UMR 5126, 18 avenue Edouard Belin, 31401 Toulouse CEDEX 9, France
[b] IGN Espace-MATIS/Université Paris-Est, 6 avenue de l'Europe, 31521 Ramonville CEDEX, France

ARTICLE INFO

Article history:
Received 22 February 2016
Received in revised form 13 September 2016
Accepted 3 October 2016
Available online 15 October 2016

Keywords:
Land cover mapping
Satellite image time series
High resolution
Random forest
Temporal features
Sentinel-2

ABSTRACT

New remote sensing sensors will acquire High spectral, spatial and temporal Resolution Satellite Image Time Series (HR-SITS). These new data are of great interest to map land cover thanks to the combination of the three high resolutions that will allow a depiction of scene dynamics. However, their efficient exploitation involves new challenges, especially for adapting traditional classification schemes to data complexity. More specifically, it requires: (1) to determine which classifier algorithms can handle the amount and the variability of data; (2) to evaluate the stability of classifier parameters; (3) to select the best feature set used as input data in order to find the good trade-off between classification accuracy and computational time; and (4) to establish the classifier accuracy over large areas.
This work aims at studying these different issues, and more especially at demonstrating the ability of state-of-the-art classifiers, such as Random Forests (RF) or Support Vector Machines (SVM), to classify HR-SITS. For this purpose, several studies are carried out by using SPOT-4 and Landsat-8 HR-SITS in the south of France. Firstly, the choice of the classifier is discussed by comparing RF and SVM algorithms on HR-SITS. Both classifiers show their ability to tackle the classification problem with an Overall Accuracy (OA) of 83.3 % for RF and 77.1 % for SVM. But RF have some advantages such as a small training time, and an easy parameterization. Secondly, the stability of RF parameters is appraised. RF parameters appear to cause little influence on the classification accuracy, about 1% OA difference between the worst and the best parameter configuration. Thirdly, different input data – composed of spectral bands with or without spectral and/or temporal features – are proposed in order to enhance the characterization of land cover. The addition of features improves the classification accuracy, but the gain in OA is weak compared with the increase in the computational cost. Eventually, the classifier accuracy is assessed on a larger area where the landscape variabilities affect the classification performances.

1. Introduction

New satellite missions - such as Sentinel, Venμs, or Landsat Data Continuity Mission (LDCM) - will acquire High Resolution optical Satellite Image Time Series (HR-SITS). The large swath, the short revisit time, the high spatial resolution of about 10 m, and the spectral bands from visible to infra-red will become essential to monitor large territories. For instance, Sentinel-2 satellites will provide a global cover of continental surfaces every five days in 13 spectral bands from 10 to 60 m (Drusch et al., 2012).

The use of HR-SITS is of great interest for the development of high-level operational products, such as global land cover maps. For this specific purpose, the processing should be designed to operate with a robust classifier, and suitable input data. More precisely, standard classification processing chains need to be adapted in order to: (1) achieve a good trade-off between classification performances, the stability of the classifier and computational time; (2) provide the classifier with the best input data, which fully exploit the quantity of information given by HR-SITS; and (3) deal with the data variability arising from the landscape diversity over large areas. Note that HR-SITS will cover large areas where climate, human activities, and landscape soils and slopes may differ.

Multi-temporal classification issues have been tackled in different ways in remote sensing literature. Some approaches have proposed to select key dates representing discriminative phenological stages in order to classify multi-temporal data. The most common strategy consists in selecting images acquired on two different seasons (Rogan et al., 2002; Rodriguez-Galiano et al., 2012). These key dates generally correspond to a low cloud cover period,

* Corresponding author.

http://dx.doi.org/10.1016/j.rse.2016.10.010

and to higher differences in spectral signatures of vegetation categories. Carrão et al. (2008) and Masse et al. (2011) also proposed automated date selection methods using sorting criteria and genetic algorithms respectively. They sought to extract the most condensed and pertinent information from high temporal resolution SITS. However, the selection of key dates is a complex step for operational land cover mapping because the acquisition of (cloud free) images is not ensured at the key dates, and climatic change or human activities may change these key dates from one year to another.

More general methods (without date selection) classified SITS with a high temporal resolution, but at a low spatial resolution (Alcantara et al., 2012; Gong et al., 2006). On the contrary, the classification of high spatial resolution images has been introduced with single or few images (Lu et al., 2004). Generally, a unique image cannot distinguish all the land cover categories. Likewise, global land cover initiatives suffer from the lack of sensors combining high temporal and spatial resolutions (Wardlow and Egbert, 2008; Gong et al., 2013; Chen et al., 2015; Wang et al., 2015).

Therefore, an operational scheme for HR-SITS should work with high spatial, spectral, and temporal resolutions. In addition of large area processing, this scheme should also respect user requirements such as a frequent update, a high number of land cover categories, or an automated process.

Above-mentioned classification techniques rely on supervised and unsupervised approaches. Many comparisons between both types have been performed in the literature, showing that supervised methods — Maximum Likelihood (ML), Neural Networks (NN) (multilayer perceptron of Atkinson and Tatnall (1997)), Support Vector Machines (SVM) (Vapnik, 1995, 1998), and Decision Trees (DT) (Breiman et al., 1984; Hansen et al., 1996; Friedl and Brodley, 1997) — outperform unsupervised methods (Szuster et al., 2011; Khatami et al., 2016).

More specifically, ensemble learning methods (bootstrap, boosting, *etc.*) have recently received a strong interest. They consist in learning several weak classifiers to generate a classifier with a strong decision rule. A well-known ensemble learning method is Random Forests (RF) of Breiman (2001), which has demonstrated its ability to yield accurate land cover maps (Belgiu and Drăguţ, 2016). It accomplished performances comparable to traditional classifiers such as DT or SVM, with a lower computational time (Inglada et al., 2015; Rodriguez-Galiano et al., 2012; Gislason et al., 2006).

To help the classifier to learn the decision rule, features (also named variables and attributes) are used as input data in the classification system. The number and quality of input features are related to resulting accuracies, but also to computational time. Hundreds of spectral features, such as the NDVI (Normalized Difference Vegetation Index) for vegetation depiction, NDWI (Normalized Difference Water Index) for water detection or NDBI (Normalized Difference Built-up Index) for building detection, have been proposed and compared in remote sensing domains (Mróz et al., 2004; Silleos et al., 2006; Yeom et al., 2013). In the same way, a great number of spatial features (geometric, texture, *etc.*) have been proposed (Haralick, 1979; Trias-Sanz, 2006; Lv et al., 2014). For land cover mapping, temporal features receive less interest because of the lack of high temporal SITS. However, they have proven their ability to improve the classification accuracies, especially on vegetation categories (Jia et al., 2014; Valero et al., 2016).

Features are mainly used to reduce the dimensionality of the data without discarding the main information. Before the introduction of accurate classifier methods that can handle complex and high quantity of data, only specific features were computed in line with the classification problem (Xiao et al., 2005). Currently, hundreds of features are computed (Dalla Mura et al., 2010; Huang and Zhang, 2013), and then the best subset is selected (Gressin et al., 2013; Paget et al., 2015). In addition, robust classification methods, such as SVM or RF, have shown that their performances are likely to remain unchanged even by adding insignificant features.

When working with SITS acquired at high temporal resolution, the contribution of several input features is uncertain (Fernández-Delgado et al., 2014). Indeed, the spectral signatures of temporal profiles can be enough to characterize land cover categories. However, features can help to deal with large variability of the landscape when working on large areas. Therefore, it becomes interesting to determine the best feature subset of smaller size among all available features in order to achieve equal accuracies and reduce computational cost.

This work aims at assessing the robustness of classification methods to provide accurate land cover maps over large areas with HR-SITS. Specifically, it addresses the evaluation of RF performances on large areas by using different feature sets as input data through several studies. Firstly, the choice of RF classifier is discussed by comparing it with the well-known SVM. Secondly, the RF parameter sensitivity is analyzed. Then, the use of different sets of features as input data in the classification system is studied. Finally, the classifier stability is tested on a larger area covering around 20,000 km.[2].

This paper is organized as follows: Section 2 describes the data; Section 3 details classification scheme, and more precisely the input features used, and RF classifier; Section 4 is devoted to results and discussions; and finally Section 5 draws the conclusion.

2. Data

2.1. Study area

Two study areas are selected in the south of France having a temperate climate and a mean annual precipitation of 650 mm (Fig. 1). The slope in the whole scene varies greatly with the presence of the Pyrénées mountains in the south, the Massif Central mountains in the northeast, and lowlands in the remaining area. Main land cover categories are agricultural fields (principally winter crops), roads, urban areas, and forests (broad-leaved and conifers).

2.2. Satellite images

Landsat-8 and SPOT-4 (Take-5 experiment, Hagolle et al. (2015)) images are used as a simulation of the upcoming ESA's Sentinel-2 data. The combination of both sensors provides images with a temporal average gap of 13 days, approaching the temporal resolution of Sentinel-2 of five days. Concerning the spatial resolution, Sentinel-2 has variable band resolutions from 10 to 60 m, but the main bands have a spatial resolution of 10 or 20 m close from the one of Landsat-8 and SPOT-4. The major difference with Sentinel-2 time series concerns the spectral resolution due to the absence of red-edge bands for Landsat-8 and SPOT-4 sensors. The satellite characteristics are summarized in Table 1.

USGS (United States Geological Survey) and THEIA Land Data Centre pre-processed Landsat-8 and SPOT-4 images respectively: they ortho-rectified images and converted digital number values to top-of-atmosphere reflectances. Then, top-of-atmosphere reflectances are converted to top-of-canopy reflectances by using MACCS processing chain (Multi-sensor Atmospheric Correction and Cloud Screening, Hagolle et al. (2015)) for both satellite images. Images with more than 80% of cloudy data, are omitted in the final time series.

As satellites need several orbital cycles to cover a large area, HR-SITS are composed of multiple footprint images acquired at different dates. In order to work with a regular temporal sampling in the whole scene, images are temporally resampled (linear interpolation) as done in Inglada (2016a). Their work showed that although sample information is modified, classification accuracy does not change significantly (Inglada, 2016a).