Int J Appl Earth Obs Geoinformation 73 (2018) 39–51

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

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

# A data mining approach for global burned area mapping

Rubén Ramo[a,b,*], Mariano García[a,b], Daniel Rodríguez[c], Emilio Chuvieco[a,b]

[a] Department of Geology, Geography and Environment, University of Alcalá, Colegios 2, 28801 Alcalá de Henares, Spain
[b] Enviromental Remote Sensing Research Group, Department of Geology, Geography and the Environment, University of Alcalá, Colegios 2, 28801 Alcalá de Henares, Spain
[c] Department of Computer Science, University of Alcalá, Ctra. de Barcelona Km 33.6, 28871 Alcalá de Henares, Spain

ARTICLE INFO

*Keywords:*
Data mining
Burned area
MODIS
Remote sensing
Random forest
SVM
Neural Net
C5.0

ABSTRACT

Global burned are algorithms provide valuable information for climate modellers since fire disturbance is responsible of a significant part of the emissions and their related impact on humans. The aim of this work is to explore how four different classification algorithms, widely used in remote sensing, such as Random Forest (RF), Support Vector Machine (SVM), Neural Networks (NN) and a well-known decision tree algorithm (C5.0), for classifying burned areas at global scale through a data mining methodology using 2008 MODIS data. A training database consisting of burned and unburned pixels was created from 130 Landsat scenes. The resulting database was highly unbalanced with the burned class representing less than one percent of the total. Therefore, the ability of the algorithms to cope with this problem was evaluated.

Attribute selection was performed using three filters to remove potential noise and to reduce the dimensionality of the data: Random Forest, entropy-based filter, and logistic regression. Eight out of fifty-two attributes were selected, most of them related to the temporal difference of the reflectance of the bands. Models were trained using an 80% of the database following a ten-fold approach to reduce possible overfitting and to select the optimum parameters.

Finally, the performance of the algorithms was evaluated over six different regions using official statistics where they were available and benchmark burned area products, namely MCD45 (V5.1) and MCD64 (V6). Compared to official statistics, the best agreement was obtained by MCD64 (OE = 0.15, CE = 0.29) followed by RF (OE = 0.27, CE = 0.21). For the remaining three areas (Angola, Sudan and South Africa), RF (OE = 0.47, CE = 0.45) yielded the best results when compared to the reference data. NN and SVM showed the worst performance with omission and commission error reaching 0.81 and 0.17 respectively. SVM and NN showed higher sensitivity to unbalanced datasets, as in the case of burned area, with a clear bias towards the majority class. On the other hand, tree based algorithms are more robust to this issue given their own mechanisms to deal with big and unbalanced databases.

## 1. Introduction

Wildland fires are one of the most important disturbances in the Earth system, affecting the balance of greenhouse gases (van der Werf et al., 2010), vegetation distribution and society (Goldammer et al., 2008; Kloster et al., 2012; Schoennagel et al., 2009). Wildland fires are considered an Essential Climate Variable (ECV) by the Global Climate Observing System (GCOS) (2004); Hollmann et al., 2013) and has, therefore, been selected by the European Spatial Agency (ESA) as one of the ECV included in the Climate Change Initiative (CCI) program (Hollmann et al., 2013).

Burned area (BA) detection is an active research topic which has been studied over a variety of ecosystems. Many studies have shown the ability of high resolution sensors to map burned areas at local scale using high and medium resolution images (Dragozi et al., 2014; Mitri and Gitas, 2013). Nevertheless, to analyze global vegetation dynamics (Mouillot et al., 2014) or greenhouse gas emissions estimation (Leeuwen et al., 2013), global coverage is needed. In this framework, the most used products are those that use MODIS (Moderate-Resolution Imaging Spectroradiometer) images, such as MCD45 (Roy et al., 2005) or MCD64 (Giglio et al., 2013) products. In addition to these data, there are others BA products developed by different European projects in the last decade such as L3JRC (Tansey et al., 2008), Globcarbon (Plummer et al., 2005) based on SPOT-VEGETATION, or the Fire_cci product (Alonso-Canas and Chuvieco, 2015; Chuvieco et al., 2016) based on MERIS (Medium-Spectral Resolution Imaging Spectrometer). Given the high variety of the burning conditions (i.e. vegetation type, biomass consumption, time prevalence), most of the global BA products relies in

* Corresponding author at: Department of Geology, Geography and Environment, University of Alcalá, Colegios 2, 28801 Alcalá de Henares, Spain.
*E-mail address:* ruben.ramo@uah.es (R. Ramo).

https://doi.org/10.1016/j.jag.2018.05.027
Received 16 April 2018; Received in revised form 23 May 2018; Accepted 29 May 2018

the use of regional thresholds to discern between burned and unburned areas (Alonso-Canas and Chuvieco, 2015; Giglio et al., 2013; Plummer et al., 2005; Tansey et al., 2008), but none of them has been yet developed using machine learning algorithms, particularly using a global training dataset.

Data mining, defined as the computing process of discovering patterns and relationship from large dataset through the use of machine learning, statistics and database systems (Fayyad et al., 1996), has experienced an increase of popularity in the remote sensing field because of its capability to extract patterns from apparently unstructured data. For instance, it has been successfully applied to map natural disasters (Barnes et al., 2007; Goswami et al., 2016; Traore et al., 2017), land cover classification (DeFries and Chan, 2000; Zhou et al., 2013) or change detection (Boulila et al., 2011; Hussain et al., 2013). It has also been applied in fire applications such as forest fire prediction (Cheng and Wang, 2008) or to map burned areas (Özbayoğlu and Bozer, 2012; Quintano et al., 2011).

One of the advantages of train global models is that after the training phase, the classification become fully automatic without the need of further calibrations or regional adaptations (Ramo and Chuvieco, 2017). However, the main difficulties of this approach are the necessity of generating a training database that includes the great variability of burned conditions, and the generation of balanced error rate models that classify burned area without overfitting or bias to the majority (or minority) class, obtaining similar error rates results among different regions.

The main objective of this study was to compare the capacity of four well-known machine learning algorithms, namely random forests (RF), support vector machine (SVM), artificial neural networks (ANN) and decision trees (C5.0), to map burned areas at global scale using a data mining approach. The algorithms were applied over six different regions (Australia, Angola, California, South Africa and Sudan) and the results validated in two ways. First, the performance was evaluated by leaving 20% of the training database for independent validation. Second, comparing the BA information yielded by the algorithms with existing official statistics (Australia, Canada and California), and two well-known BA products namely, MCD64 and MCD45.

## 2. Materials and methods

The proposed methodology consisted of several steps involving the training database compilation, attribute selection, algorithm training and evaluation, image classification and perimeter comparison. The flowchart of the applied methodology is presented in Fig. 1 to facilitate its interpretation.

### 2.1. Burned Area perimeters

To create the training dataset, the burned area perimeters from the Fire_cci project (http://www.esa-fire-cci.org/ last accessed April 2018) were used. This dataset has been previously used to validate global BA products (Padilla et al., 2015) such as MCD64 (Giglio et al., 2013), MCD45 (Roy et al., 2005) or the Fire_cci product (Alonso-Canas and Chuvieco, 2015). The Fire_cci validation dataset follows a global statistically designed sample (Padilla et al., 2014), thus the training sites were selected using a stratified random sampling were the strata were defined based on the proportion of burned area extracted from the Global Fire Emissions Database (GFED) (Giglio et al., 2013) and the Olson biomes reclassified in 7 categories based on their similarities and fire behavior (e.g. deserts, Tundra and Mangroves were merged in one class). Thus for each biome the proportion of burned area was computed and those with ≥80% of the area burned were grouped into the high burned area, and those with < 80% into the low burned area class, respectively. The Fire_cci validation dataset is composed of 130 Landsat pairs from 2008 (see Fig. 2) covering 1.58 million of km$^2$ from which 31,578 km$^2$ correspond to burned area. Burned areas include: Rainfed cropland (10.10%), mixed forest closed to open > 15% (10.63%), broad-leaved deciduous open 15–40% (5.45%), need-leaved evergreen closed to open > 15% (8.54%), shrubland (14.42%), grassland (16.16%), sparse vegetation (tree, shrub, herbaceous cover > 15%), and vegetation regularly flooded (5.13%).

### 2.2. MODIS data

The main source of information is the MCD43A4 (v6). This product was developed using Terra and Aqua observations to correct for the BRDF effect (Schaaf et al., 2002). The MCD43A4 has 500 m spatial resolution and includes the spectral information of seven different bands, Red (B1), Near-infrared (NIR, B2), Blue (B3), Green (B4) and three bands in the shortwave infrared region (SWIR, B5-B7). In addition to these bands, several spectral indices were computed to enhance the BA discrimination (Table 1).

### 2.3. Ancillary data

In addition to the information provided by the spectral bands and indices, information coming from hotspots (HS) was included. Thermal anomalies information has been extensively used for burned area detection because it provides higher contrast between burned and unburned pixels in comparison with other wavelength regions (Alonso-Canas and Chuvieco, 2015; Giglio et al., 2013). Hence, the MODIS MCD14ML (Version 5.1) product, which provides daily global coverage of hotspot with 1 km spatial resolution, was used. Using this data a distance matrix between each pixel to the closest HS was performed and included as an attribute.

Additionally, we included auxiliary data to adapt the model to regional environmental conditions of burned areas. In this case, we used the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) (Danielson and Gesch, 2011) from which the slope and aspect were computed. This information is useful for BA classification since it is related to the fire behaviour and the physical properties of the land. Land cover (LC) information also provides valuable data for BA mapping as it allows for characterizing the fire signal before and after the fire (Moreira et al., 2009) therefore, Land Cover CCI product was used.

Another important factor for BA mapping is related to the ecosystem variation. The condition of burned and the carbon footprint prevalence in the post-fire image is dependent on climate and vegetation type. In order to characterize this factor, we used the Olson biomes (Olson et al., 2001) which divide the world into 16 regions considering their geology, climate, and evolutionary history. Finally, we included the continental regions defined in the Global Fire Emission Database (GFED) that have been developed taking into account how the fire behaves (Giglio et al., 2013) and hence it can help to characterize the burned signal.

### 2.4. Training database

The database used for this study comprised the spectral and ancillary information previously described for two categories, namely burned and unburned pixels. Regarding the burned area, the database also included information of the burned proportion of the pixel and the date of the burned. The proportion of burned was extracted by overlapping the Landsat perimeters to the MODIS images. The HS was also used to assign the day of burned to each perimeter from the closest HS.

Our approach to map burned area was also based on a multitemporal analysis, therefore, we extracted the MODIS reflectance values for each band from an image acquired prior to the fire (t1) and another one after the fire (t2). For burned pixels, we constrained the search of post-fire images to a period between 2 and 12 days after the day of burned to avoid smoke plumes and clouds. Pixels with no valid observations in this period were rejected from the database. On the other hand, the search of pre-fire information was also constrained to a period of 1 to 10 days. For non-burned pixels, the t1 was set to the