International Journal of Applied Earth Observation and Geoinformation 49 (2016) 99–106

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Mapping cultivable land from satellite imagery with clustering algorithms

R.B. Arango[a,*], A.M. Campos[a], E.F. Combarro[a], E.R. Canas[b], I. Díaz[a]

[a] *Computer Science Department, University of Oviedo, Spain*
[b] *Technical Direction, Bodegas Terras Gauda, Spain*

ARTICLE INFO

Article history:
Received 11 October 2015
Received in revised form 19 January 2016
Accepted 27 January 2016
Available online 16 February 2016

Keywords:
Land delimitation
Agriculture
Clustering
Cultivable land
Satellite
Smart agro-services

ABSTRACT

Open data satellite imagery provides valuable data for the planning and decision-making processes related with environmental domains. Specifically, agriculture uses remote sensing in a wide range of services, ranging from monitoring the health of the crops to forecasting the spread of crop diseases. In particular, this paper focuses on a methodology for the automatic delimitation of cultivable land by means of machine learning algorithms and satellite data. The method uses a partition clustering algorithm called Partitioning Around Medoids and considers the quality of the clusters obtained for each satellite band in order to evaluate which one better identifies cultivable land. The proposed method was tested with vineyards using as input the spectral and thermal bands of the Landsat 8 satellite. The experimental results show the great potential of this method for cultivable land monitoring from remote-sensed multispectral imagery.

1. Introduction

Satellite imagery and data products collected from satellite instruments integrate the datasets used by experts and researchers for modeling decision support systems related with the environmental and agricultural domains. These systems are intended to provide specific and high-value applications and services such as forecasting crop diseases or predicting the yield production. In this sense, the combination of open data satellite imagery with machine learning algorithms and other artificial intelligence techniques may improve such models.

One of the challenges addressed by these kinds of model is automatic land delimitation. In fact, the identification of homogeneous zones of crop land areas is a key factor (Schepers et al., 2004). From an agronomist standpoint, automatic land delimitation has been largely studied and researchers have proposed various techniques and algorithms such as expert systems (Le Ber, 1995), segmentation algorithms (Pedroso et al., 2010), clustering (Schuster et al., 2011; Kumar et al., 2011) and fuzzy algorithms (Fu et al., 2010; Liu and Samal, 2002).

Regarding the kind of features considered in the automatic delimitation of land, such as the classification of apparent soil electrical conductivity (Johnson et al., 2003; Peralta and Costa, 2013) or the analysis of yield maps (Blackmore et al., 2003), the following classification may be considered:

- **Soil properties**. Ortega and Santibá nez (2007) compared the results of the use of chemical properties of the soil with several techniques such as PCA and cluster analysis. Simbahan and Dobermann (2006) tested supervised classification algorithms with different datasets including soil maps, digital elevation models and apparent soil electrical conductivity. In the same line, other authors considered soil properties (Moral et al., 2011; Fu et al., 2010) and also yield and crop quality (Aggelopooulou et al., 2013).
- **Agricultural treatments and yield**. Schuster et al. (2011) identified homogeneous zones of a cotton field considering two datasets: the first one with two estimators of the yield and the second one considering geo-referenced field properties such as topographical characteristics and treatments applied to the field.
- **Biophysical features**. The use of biophysical features such as annual moisture deficit/surplus and mean annual precipitation

* Corresponding author.
E-mail addresses: rbarango@gmail.com (R.B. Arango), campos@uniovi.es (A.M. Campos), elias@aic.uniovi.es (E.F. Combarro), ecanas@terrasgauda.com (E.R. Canas), sirene@uniovi.es (I. Díaz).

http://dx.doi.org/10.1016/j.jag.2016.01.009
0303-2434/

was proposed by Liu and Samal (2002). The authors tested the same dataset with k-means and fuzzy algorithms concluding that a fuzzy approach generates more accurate delineations.

- **Remote-sensed data**. Other approaches (Kumar et al., 2011) studied the use of the k-means algorithm with the MODIS-based greenness index and the seasonal leaf area index, developing a parallel implementation able to delimit 1000 agroecozones in 700 s using 2048 processors. On the other hand, Duro et al. (2012) studied the classification of agricultural landscapes by means of image analysis techniques with SPOT-5 (https://goo.gl/LpIaT4/) satellite imagery.

The major drawback of the first two approaches is that they require some kind of in-field exploration in order to measure the values of the features. On the other hand, biophysical features can be acquired from public sources such as agro-meteorological stations but the results will depend on the number and the geographical distribution of the stations. However, remotely-sensed imagery from satellites does not require in-field exploration and there are satellite programs with global coverage and free distribution of their data products. Kumar et al. (2011) followed this approach, however the process of the delimitation of the land only considers two features with a low spatial resolution for PA applications (250 m). Even though Duro et al. (2012) have worked with high spatial resolution imagery (10 m), SPOT-5 data products are not publicly available.

The main purpose of this research was to develop a methodology for the automatic delimitation of land able to be used for farming, cultivable land (Fawcett, 1930), using clustering algorithms with publicly available satellite data (Arango et al., 2015). In addition, it was also intended to study the performance of spectral and thermal bands provided by Landsat 8 satellite in detecting cultivable land. The approach developed was tested and applied to the vineyards of Terras Gauda, a well-known wine producer from Galicia, Spain. The study considered three plots with topographical dissimilarities and compared the results of cultivable land delimitation from clusters obtained using different spectral and thermal bands. Results suggest that the use of a clustering approach together with data from Landsat 8 satellite is promising for mapping cultivable land.

The remainder of the paper is organized as follows: Section 2 explains the methodology for the automatic delimitation of cultivable land. Section 3 describes the application of the method in a study case and shows the results of the validation. Finally, Section 4 draws the conclusions and presents some ideas for future work.

2. Cultivable land delimitation methodology

The proposed method groups the pixels of multispectral images acquired by on-board satellite instruments for a desired land zone and period of time, in two main clusters: cultivable and non-cultivable land. Each element of the clustering represents one pixel of the image and is assigned to a group by means of a dissimilarity metric. The method tries to group as well the pixels in three, four and five clusters, with the aim of select the better clustering. In this regard, the scientific literature proposes the calculation of indices (Milligan and Cooper, 1985) that measure the goodness or quality of the clustering. Section 2.3 explains the index used by the method. If as a result the clustering has more than two groups then the method generates two metaclusters merging the clusters.

Note also that one of the objectives of this work is the study of the performance of Landsat 8 satellite in detecting cultivable land. Thus, the clustering algorithm is applied as many times as the considered bands and taking as input one satellite band each time. The main steps are the following:

1 Download and process the Landsat 8 data products corresponding to the region under study, getting the raw values of the spectral and thermal bands.
2 For each band B_i:
  - Perform the clustering with number of clusters k varying from 2 to 5.
  - Select the number of clusters k maximizing a quality clustering index (see Section 2.3).
  - If $k > 2$, merge clusters in 2 groups (associated to cultivable and not cultivable land) by computing distances between each pair of cluster representatives.

As output of the method, each pixel from the considered image is labeled as cultivable or non-cultivable land. In this point it is possible to generate, for instance, a new layer for the raster image with the aim of producing a map of cultivable/non-cultivable land (see Fig. 1).

All these steps are susceptible of being automated using, for instance, R o Python. In fact, we have developed R scripts and functions for the whole process with the exception of the download of satellite imagery. Part of this work was published as a R package in Github (https://github.com/rbarango/LST8). The package is intended to process Landsat 8 data products and extract the geolocated values of the satellite instruments for the desired geographical extension. The data obtained by this R package would be considered as input for the clustering algorithm. The representation of the clusters (cultivable and non-cultivable land) in a map are also automated with scripts.

In the following subsections, all the steps involved in the methodology are described in detail.

Fig. 1. Mapping cultivable land cluster for the land parcel 2 with the Band 5 (NIR). The zeroes represent the land correctly classified as non-cultivable. The ones represent the land incorrectly classified as non-cultivable. The zones without numbers belong to the cultivable-land cluster.