Inter-annual land cover mapping

An approach for producing inter-annual maps by implementing scene based and pixel-based composites using Sentinel 2 imagery and an automatic labeling procedure for classification based on old maps.

Dissertation supervised by
Dr. Mario Caetano
Professor, Nova Information Management School
University of Nova Lisbon, Portugal

Dr. Edzer Pebezma
Professor, Institute for Geoinformatics
University of Münster, Germany

Dr. Jorge Mateu
Professor, Department of Mathematics
University of Jaume I Castellón, Spain

January 30, 2019

Abstract

Land Use Land Cover (LCLU) is very well established as one of the most efficient approaches for monitoring land cover and its changes (Gómez et al., 2016). Estimating the dynamic of the land cover system under different temporal scales may contribute to improve our understanding of its interaction with different biophysical process according with their temporal variation (Yang et al., 2013). In this thesis, we investigated the viability of producing inter-annual maps by implementing scene-based and pixel-based composites using sentinel 2 imagery, and an automatic labeling procedure for the classification based on old maps. Our results provide initial insights into the benefits and specific issues in this context. We found out that the use of pixel-base composites based on maximization of NDVI can lead to combat cloud contamination and reproduce artificial imagery with good quality for classification. Besides that, the reference old maps resulted in an extensive source of training data that also contain an important fraction of noise. To address this issue, we undertook the viability of using two procedures for refining mislabel data, one based on a iterative learning procedure using Random forest and entropies, and another based in the variability of NDVI signals. The proposed methodologies are tested using Sentinel 2 imagery for 2017 and the reference COS map developed by DGT. From the results, we discussed extensively the robustness of classifiers in the presence of different levels od noise data as well as parametrization in an environment of automatic mapping.

Keywords

Composites

Informativeness

Automatic feature selection

Support vector machine and random forest

Geographical information systems

Acronyms

DGT - Direção-Geral de Ordenamento do Território, Portugal

L1C Level-1C

L2A Level-2A

MMU Minimum Mapping Unit

SVM Support vector machine

RF Random forest

LCLU Land Cover Land Use

BAP Best available pixel

List of Figures

List of Tables

Contents

1 Introduction

1.1 Problem statement and Motivation

Land Use Land Cover (LCLU) is very well established as one of the most efficient approaches for monitoring land cover and its changes (Gómez et al., 2016). Estimating the dynamic of the land cover system under different temporal scales may contribute to improve our understanding of its interaction with different biophysical process according with their temporal variation (Yang et al., 2013). In this context, increasing the observation of LULC by new technology as it is improved the systematic production of LULC maps can lead to address with more consistency the analysis of process with inter-annual variability, such as models for fire propagation (Navarro et al., 2017), crop production (Vuolo et al., 2018) and climate (Bontemps et al., 2012).

For the regular characterization of the dynamic of the land system by using time series of Sentinel 2, the increase of dimensionality by the fact of an increase in the number of images to study, the possible differences in accuracy of the products and the clouds persistence in certain periods may determine special constraints in its application (Lu et al., 2004). With the increase of the revisit of observations of Sentinel 2, the approach of the best pixel available (BPA) offer new opportunities to produce imagery with regular frequency as well as better spectral features than can lead to perform better results in the task of classification (Gómez et al., 2016). Pixel- based composites have been developed by using different kind of protocols, which mainly depend on the use NDVI values and distances to the masked clouds to define the BPA (Hermosilla et al., 2015). In this context, based on experiments of Holben (1986) with VHRR time series, that is, making composites of NDVI by maximizing the NDVI in an arrange of several scenes in order to capture the state of the of vegetation when is more photosynthetically active, we propose to make seasonal composites as case of study. Besides that, this thesis aims at evaluating the viability of implementing a maximum value composites for classification, by retaining not only maximum NDVI per scene but also the rest of the spectral information associated to the index of the pixel with the highest NDVI.

Besides of giving special interest to the imagery in the dynamic of land cover at inter-annual time frame, the technique of classification may be also fundamental. Support Vector Machine (SVM) and Random Forest (RF) represent state of the art algorithms for its application in the production of LULC (Thanh Noi and Kappas, 2018); important for their ability to handle high dimensionality, being superior to unsupervised approaches and being insensitive to overfitting (Bishop, 2006). However, the performance of the supervise algorithms essentially depends on the quality of the labeled data for training (Tuia et al., 2009). Generally, the strategy for collecting labeled datasets consists of using interpreted data from very high spatial resolution satellite images or aerophotographies. Nonetheless, in most cases this selection turns out an expensive task not operationally efficient in the production of maps under the inter-annual reference (Inglada et al., 2017).

The availability of previous maps in the study area represent also an important reference (Colditz et al., 2011), and therefore an effective method for automation of collecting training data. However, even though they can represent a rich source of information, this data may content errors (Pelletier et al., 2017b). In this context, source of errors can be explained by 1: maps are a generalization of the land cover system, so random samples can fall over complexities of a wide diversity of classes that were simplified in one class in the map, 2: expected changes in the land cover due to the gap between the image acquisition and the date of production of the reference map, and 3: to the natural changes of certain communities of vegetation during the year that lead to consider temporarily in the classes. Although support vector machines and random forest are also known for being resistant to anomaly data, a classifier trained in a set of large amount of wrong labels can turn out in a wrong model (Pelletier et al., 2017a). Therefore, this thesis aim at refining the sampling by exploring the viability of implementing two cleaning procedures; one based on NDVI signal variability and other based on analysis of informativeness of samples per image.

Therefore, to continue advancing in the state of the art of inter-annual land cover mapping, is therefore important to know how the aforementioned proposals are going to be explored . In this context, this thesis aims at performing and comparing classification by using scene-based and pixel-based composites per season of Sentinel 2 images from January 2017 to December 2017 at central of Portugal. Besides that, it aims at implementing an automatic labeling procedure for supervised classification based on old maps and filtering preprocessing for mislabeled data.

1.2 Objectives

This thesis aims to research four questions:

- How do the classification approach based on multi-spectral data works over the year 2017 using random forest and support vector machine classifiers?
- Do a refining procedure of mislabeled data can lead to improve the performance of a scene classification.
- How usable is COS dataset in the classification task using multi-temporal and multi-spectral data from sentinel 2 imagery?
- Do a pixel-based image composites analysis achieve better results in classification than a scene based analysis?