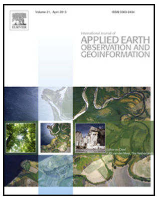
International Journal of Applied Earth Observation and Geoinformation 51 (2016) 1–10

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Urban land cover thematic disaggregation, employing datasets from multiple sources and RandomForests modeling

Dimitrios Gounaridis*, Sotirios Koukoulas

SAGISRS Lab, Department of Geography, University of the Aegean, Mytilene, Lesvos, Greece

ARTICLE INFO

Article history:
Received 26 September 2015
Received in revised form 5 April 2016
Accepted 18 April 2016
Available online 2 May 2016

Keywords:
Urban land cover
Thematic disaggregation
Urban atlas
Landsat
RandomForests

ABSTRACT

Urban land cover mapping has lately attracted a vast amount of attention as it closely relates to a broad scope of scientific and management applications. Late methodological and technological advancements facilitate the development of datasets with improved accuracy. However, thematic resolution of urban land cover has received much less attention so far, a fact that hampers the produced datasets utility. This paper seeks to provide insights towards the improvement of thematic resolution of urban land cover classification. We integrate existing, readily available and with acceptable accuracies datasets from multiple sources, with remote sensing techniques. The study site is Greece and the urban land cover is classified nationwide into five classes, using the RandomForests algorithm. Results allowed us to quantify, for the first time with a good accuracy, the proportion that is occupied by each different urban land cover class. The total area covered by urban land cover is 2280 km^2 (1.76% of total terrestrial area), the dominant class is discontinuous dense urban fabric (50.71% of urban land cover) and the least occurring class is discontinuous very low density urban fabric (2.06% of urban land cover).

1. Introduction

Urban areas determine, both positively and negatively, several functions of the Earth system, from local to global scales (DeFries et al., 2010; Folke et al., 1997). Accurate information about urban land cover (ULC) is critical to a wide range of social, economic, and environmental research questions not only for descriptive but also for analytical and predictive modeling purposes. Thus, reliable spatial information about ULC composition and configuration serves as a principal component in a variety of scientific activities, across several disciplines, while for studies related to global, environmental and/or climate change it is considered a pre-requisite (Grimm et al., 2008; Mills, 2007).

In pursuit of spatial information about land cover (LC), traditional field data approaches face certain drawbacks as they are limited to a local extent due to their prohibitively expensive nature in means of time, costs and personnel. Technological and methodological advances in remote sensing (RS) and geographic information systems (GIS) successfully provide spatially consistent LC information. Nowadays, an increased number of satellite sensors has been launched and facilitate the growing demand for multi-spectral and multi-temporal information of the Earth's surface over a wide range of scales and data types (Belward and Skøien, 2015).

A number of studies have generated several datasets regarding ULC, or LC in general. The majority of them consider studies at the scale of individual cities, analyzing changes and patterns over multiple years or exploiting spatial information and structure on a single date (Yu et al., 2014). On a global scale, more than ten datasets have been produced with spatial resolutions ranging from approximately 10 km to 30 m (Chen et al., 2015; Potere et al., 2009). The limitations and drawbacks of these global datasets have been discussed in detail by several researchers (Congalton et al., 2014; Giri et al., 2013; Potere et al., 2009). The predominant conclusion stressed by these studies is that the most prominent drawback is the variability in ULC definition.

On a regional scale, for Europe, the CORINE land cover (CLC) is the most frequently used dataset with a hierarchical classification scheme comprising of 44 classes (at level 3) and a minimum mapping unit of 25 ha. The urban category, denoted as 'urban fabric' is divided in two classes (continuous and discontinuous). Recently, the Copernicus land monitoring service has released the Urban Atlas (UA) database. It consists of LC maps for 305 European large urban zones with more than 100.000 inhabitants for

* Corresponding author.
E-mail address: gounaridis.d@geo.aegean.gr (D. Gounaridis).

http://dx.doi.org/10.1016/j.jag.2016.04.002
0303-2434/© 2016 Elsevier B.V. All rights reserved.

the reference year 2006 (European Commission, 2011). It has been derived by very high resolution satellite data (spatial resolution between 2.5 and 5 m) and has a significantly lower minimum mapping unit of 0.25 ha, compared to CLC. The thematic resolution of UA, regarding ULC is also much more detailed than CLC -although it has limited geographic coverage- dividing the urban class into five classes differentiated by their degree of imperviousness.

Despite the unquestionable value of the datasets produced so far, their application to a range of research applications and management activities is inefficient. The reason is their resolution, both spatial and thematic, a constraint in cases when these data are to be used in studies that finer scale of analysis is mandatory (e.g., urban planning) or in studies that require sufficient thematic ULC detail (e.g., population density mapping). As far as spatial resolution is concerned, the previous efforts mainly employed coarse resolution satellite data for feasibility reasons (data availability, technical innovation, human and financial resources). However, ULC delineation employing coarse resolution primary data is not a simple task. On the one hand, ULC class has a limited areal extent in comparison with other classes, while on the other hand, it is a class with extreme variability in terms of spectral and textural characteristics. Thus, data derived by coarse spatial resolution are due to the mixed pixel effect, especially for the ULC class, where small area urban areas are often completely omitted, spatial details are lacking and the edges are erroneously presented (Potere et al., 2009; Schneider et al., 2010). Thematic resolution refers to the number of classes and the detail in their definition that determines the amount of geospatial information of hard classified categorical data. The more detail in a land user/cover map, the more meaningful and insightful the map is for a wide range of research questions. Several authors have explored the effects of thematic resolution in land use modeling (Conway, 2009; Pontius and Malizia, 2004), land-cover pattern analyses (Buyantuyev and Wu, 2007) and landscape indices behavior (Bailey et al., 2007), converging that the outcomes are significantly influenced. Whilst thematic resolution is important to a range of applications, available regional and global datasets in most cases represent ULC lumped into one or two broad classes, a fact that is far from reality on the ground, given the heterogeneity of urban areas across space (Potere et al., 2009).

In this paper we successfully disaggregate ULC patterns into five categories achieving nationwide coverage (for Greece). Additionally, we demonstrate a sequence of steps towards the improvement of existing shortcomings and scarcity of high quality data related to ULC. Our main focus was to achieve the highest possible thematic resolution without compromising accuracy. To this end, we employ the Random Forests (RF) machine learning algorithm (Breiman, 2001) that is proven to perform well in the face of heterogeneous classes. Our model is trained intensively by the polygon centroids of the UA dataset -available for nine cities- to finally 'predict' ULC for the rest of the geographic coverage of Greece. Road density, population, LC and spectral indices derived by Landsat satellite, serve as predictor variables.

The rest of the paper is structured as follows: We first present the study site, Greece, along with information about morphology, recent population dynamics and some causal factors that contributed to the existing ULC scenery. Next, we present an overview and the data used as both response and predictor variables to train our models, along with the data pre-processing steps. Then, the RF classifier application and the accuracy assessment process are described in detail. In the next section, we present the obtained results and we discuss the model performance. Finally, in the last section we discuss the conclusions drown and we highlight some key points.

2. Material and methods

2.1. Study site

Greece is a Mediterranean country of Southeast Europe situated between latitudes 34° and 42°N, and longitudes 19° and 30°E (Fig. 1) and is populated by approximately 11 million inhabitants. Two-thirds of the inhabitants live in urban areas, while the remaining one-third are rural inhabitants (Hellenic Statistical Authority, 2013). Almost two thirds of the Greek territory is mountainous, with Mount Olympus being the highest at 2.917 m (European Environment Agency, 2010). Extensive agricultural plains are primarily located in Thessaly, Central Macedonia and Thrace regions, constituting key economic sources.

Greece has a long history of land use, ranging from prehistoric to present times constituting a country of people with strong dependency on the land. The last decades of the 20th century the economic potential of urban centers motivated a constant societal demand to capture new economic opportunities, a fact that consequently triggered a shift of rural population (Kasimis et al., 2003). In turn, rural land abandonment progressively led to marginalization of remote areas especially in the uplands (MacDonald et al., 2000) while leading to agricultural and farming intensification at the lowlands (Beopoulos and Skuras, 1997). Significant expansion of the tourism sector as well as a trend in second homes gave a boost in growth dynamics of the built environment, especially in the coastal zones. At the same time, the aforementioned developments are perceived as both consequences and driving forces of public works and transport infrastructures expansion. Thus, the demographic dynamics and the major socio-economic changes that have taken place progressively brought radical changes in Greek landscapes (Zomeni et al., 2008).

Statistics clearly advocate all the aforementioned. According to the latest census of 2011, the Greek ULC scenery consists of 13220 settlements, 746 (5.6%) of which are uninhabited, 6897 (52.2%) have less than 100 residents and 8806 (66.6%) have less than 200 inhabitants. At the same time, according to the 2011 census, Attica prefecture is inhabited by 3.827.624 residents (35% of total population) and the prefecture of Thessaloniki by 1.880.058 residents (17% of total population).

2.2. Overview

The RF algorithm is a robust non-parametric machine learning algorithm (Breiman, 2001) that has been widely used for LC classification. Initially, the algorithm uses a randomly selected part of training observations (response variable) as well as a sample of predictor variables, resulting in a number of independent to each other classification trees. This process is repeated several hundreds of times, thus forming a 'forest' of classifiers. Each tree contributes with a single vote to the assignment of the most frequent class. The final outputs of classification are determined from the majority of votes for each class (Breiman, 2001). The main advantages of adopting RF in our task are: (i) the independency of each classification tree, on the one hand, and the randomness of variable selection, on the other, reduce the problem of overfitting and at the same time make the models insensitive to noise and outliers (Breiman, 2001; Chan and Paelinckx, 2008). (ii) The algorithm can efficiently handle predictor variables, with different nature (both continuous and categorical) and from multiple sources (Gounaridis et al., 2014; Gounaridis et al., 2015) which is the case for our approach. (iii) The first two advantages of RF contribute to good performance in the classification of heterogeneous landscapes (Rodriguez-Galiano et al., 2012a; Timm and McGarigal, 2012) such as ULC. (iv) RF can handle large datasets and thousands of input variables being computationally faster than other classifiers (Rodriguez-Galiano et al.,