Ecological Informatics 5 (2010) 256–266

Contents lists available at ScienceDirect

Ecological Informatics

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

Estimating vegetation height and canopy cover from remotely sensed data with machine learning☆

Daniela Stojanova [a], Pančе Panov [b,*], Valentin Gjorgjioski [b], Andrej Kobler [a], Sašo Džeroski [b]

[a] Slovenian Forestry Institute, Večna pot 2, SI-1000 Ljubljana, Slovenia
[b] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 14 August 2009
Received in revised form 10 March 2010
Accepted 15 March 2010

Keywords:
Remote sensing
LiDAR
Landsat
Vegetation height
Canopy cover machine
Learning

ABSTRACT

High quality information on forest resources is important to forest ecosystem management. Traditional ground measurements are labor and resource intensive and at the same time expensive and time consuming. For most of the Slovenian forests, there is extensive ground-based information on forest properties of selected sample locations. However there is no continuous information of objectively measured vegetation height and canopy cover at appropriate resolution.

Currently, Light Detection And Ranging (LiDAR) technology provides detailed measurements of different forest properties because of its immediate generation of 3D data, its accuracy and acquisition flexibility. However, existing LiDAR sensors have limited spatial coverage and relatively high cost of acquisition. Satellite data, on the other hand, are low-cost and offer broader spatial coverage of generalized forest structure, but are not expected to provide accurate information about vegetation height.

Integration of LiDAR and satellite data promises to improve the measurement, mapping, and monitoring of forest properties. The primary objective of this study is to model the vegetation height and canopy cover in Slovenia by integrating LiDAR data, Landsat satellite data, and the use of machine learning techniques. This kind of integration uses the accuracy and precision of LiDAR data and the wide coverage of satellite data in order to generate cost-effective realistic estimates of the vegetation height and canopy cover, and consequently generate continuous forest vegetation map products to be used in forest management and monitoring.

Several machine learning techniques are applied to this task: they are evaluated and their performance is compared by using statistical significance tests. Ensemble methods perform significantly better than single- and multi-target regression trees and are further used for the generation of forest maps. Such maps are used for land-cover and land-use classification, as well as for monitoring and managing ongoing forest processes (like spontaneous afforestation, forest reduction and forest fires) that affect the stability of forest ecosystems.

1. Introduction

In forest management and forestry decision-making there is a continuous need for high quality information on forest resources. The state of forest resources can be monitored by using visualizations of forest properties for a specific spatial region in the form of a map. Forest maps are an effective tool for detecting the state of forest resources and monitoring ongoing spatial processes in forested landscapes. Examples of such processes include the enlargement of forest area by spontaneous afforestation of abandoned agricultural land, and the vertical growth of trees and transitions between developmental stages of existing forest stands. These processes affect the stability of forest ecosystems, an ever more important property due to extreme weather conditions, hydrological stress and the appearance of new diseases and pests.

One of the most important forest properties are: vegetation height and canopy cover. Vegetation height is the height of the vegetation in a stand, relative to the ground. It is a function of the species composition, climate and site quality, and can be used for land-cover classification or in conjunction with vegetation indices. If coupled with species composition and site quality information, vegetation height serves as an estimate of the stand age or the successional stages. Vegetation height is also a useful indicator of forest age and habitat quality. It is an important input variable for ecosystem and forest fire models, and is highly correlated with vegetation biomass and productivity. Biomass is the key component of the carbon circle (Skole and Tucker, 1993) and a surrogate for fuel loading estimation (Finney, 2004).

☆ The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

* Corresponding author. Tel.: +386 1 477 3307; fax: +386 1 477 3315.
E-mail addresses: daniela.stojanova@gozdis.si (D. Stojanova), pance.panov@ijs.si (P. Panov), valentin.gjorgjioski@ijs.si (V. Gjorgjioski), andrej.kobler@gozdis.si (A. Kobler), saso.dzeroski@ijs.si (S. Džeroski).

1574-9541/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2010.03.004

Forest canopy cover is defined as the percent cover of the tree canopy in a stand. It includes the cover from both trees and shrubs, but not herbal vegetation. Canopy cover describes the vertical projection of the tree canopy onto an imaginary horizontal surface representing the ground surface. Forest canopy cover is an ecologically very important forest property because it determines the occurrence and speed of forest regeneration. It is useful for distinguishing different plant and animal habitats, assessing forest floor microclimate, light conditions and estimating other forest variables (e.g., Leaf Area Index). Measurements of canopy cover are essential for silvicultural activities (Jennings et al., 1999).

Traditional ground-based field measurements of forest properties are made by using hand-held equipment. These measurements are expensive, subjective, time consuming and labor intensive, as well as difficult to perform, especially in dense forests (Buckley et al., 1999). Due to these reasons, other methods of estimating forest properties for larger areas are often used, such as remote sensing.

Over the course of the past few decades, remote sensing[1] (RS) has been a valuable source of information in mapping and monitoring forest activities. Remote sensing involves collecting of spatially organized data and information about an area of interest by detecting and measuring signals composed of radiation, particles and fields emanating from objects located beyond the immediate neighborhood of the sensor devices (Franklin, 2001). In this way, it offers a potential for more efficient resource assessment.

Multi-spectral RS is often used to map structural metrics at moderate resolution and broader scale. Multi-spectral satellite imagery is well suited for capturing horizontally distributed (2D) conditions, strictures and changes (Wulder et al., 2008). However, it cannot capture the 3D forest structure directly and is easily influenced by topographical covers and weather conditions.

Light Detection And Ranging (LiDAR) technology, on the other hand, provides horizontal and vertical information (3D) at high spatial resolution and vertical accuracies. It is good for characterizing the vertical structure of vegetation, but has limited spatial coverage mostly due to pricing. By combining remotely sensed data, that describe the horizontal distribution of target phenomena, with LiDAR data, we can improve the measurement, mapping and monitoring of forest properties and provide means of characterizing forest canopy parameters and dynamics.

In this context, many papers have been recently published on the joint use of LiDAR and other active and passive sensors in forest properties estimation problems (Lefsky et al., 1999; Hyde et al., 2006; Maltamo et al., 2006). These studies perform estimation of the forest structure directly from LiDAR measurements and extend them, over limited areas, to spatially homogeneous spectral segments derived from the optical data sets. Medium resolution RS data, such as Landsat images, are relatively inexpensive to acquire over large areas (Franklin and Wulder, 2002), whereas LiDAR covers small areas, at a high cost per unit area (Lim et al., 2003). As a result, these two data types may be combined to generate estimates of vegetation heights and canopy cover over large areas at a reasonable cost (Hudak et al., 2002).

Latest studies (Wulder et al., 2008) of the integration of LiDAR and satellite data point out possible high correlations between different satellite images and forest properties (vegetation height and canopy cover). Hyde et al. (2006) compared the performance of step-wise linear regression models using waveform LiDAR, RaDAR, Landsat, Quickbird and InSAR in a statistical combination of structural information in an attempt to estimate the mean canopy height and biomass. The addition of Landsat ETM+ metrics significantly improved LiDAR estimates of large tree structure — the combination of all sensors is more accurate than using LiDAR alone, but only marginally better than the combination of LiDAR and Landsat ETM+.

[1] Remote sensing. See also:http://rst.gsfc.nasa.gov (accessed February 11, 2010).

Machine learning techniques, such as regression trees, artificial neural networks and support vector machines have been widely used in many remote sensing forestry applications (Lefsky et al., 1999; Moghaddam et al., 2002; Wulder and Seeman, 2003). The typical machine learning task in all these studies is to learn a predictive model that uses a set of remote sensing observations with the aim of predicting the value of forest conditions or properties for unseen cases. The data input to the machine learning system consists of information extracted from different RS data sources, while the output of the system is a predictive model (or a set of predictive models called an ensemble) that describe the forest property.

The main objective of this study is to estimate the vegetation height and canopy cover from an integration of LiDAR and Landsat data in a diverse and unevenly distributed forest. This kind of integration uses the accuracy and precision of LiDAR data and the wide coverage of satellite data in order to generate cost-effective realistic estimation of the forest properties over a geographically large area. The study area is located in the Kras region in western Slovenia, near the border with Italy. The input to the machine learning system are the independent explanatory variables generated from multi-temporal Landsat data and the target variables (representing forest properties that we want to model): The latter are estimated from the 3D LiDAR data and serve as a very good substitute for field-base sample plot measurements. The machine learning system outputs a predictive model of the forest property at hand, which is then used to generate forest vegetation maps that can be used in a variety of forest management applications.

Although forest vegetation maps can be generated with high precision and accuracy purely from LiDAR data, this seems impractical for the nearest future due to the very high cost of high resolution LiDAR data (in our case 4 EUR/ha). On the other hand, the price of Landsat ETM+ data for a multi-temporal coverage is significantly lower (in our case it is free of charge). Using Landsat data as the main data source therefore ensures a very acceptable cost benefit ratio. On the other hand, LiDAR as used here for model calibration seems a very good substitute for field-based sample plot measurements of vegetation height and canopy cover, due to the even higher costs of field measurements which can in some cases also be very difficult and imprecise.

In our preliminary work (Džeroski et al., 2006a,b; Taškova et al., 2006), we introduce the problem of prediction of forest parameters from Landsat and LiDAR data, and present preliminary results using a limited set of machine learning algorithms. The predictive models for estimating the vegetation height and canopy cover from LiDAR and Landsat data, using model and regression trees, pointed out a possible high correlation between satellite data and vegetation properties (Džeroski et al., 2006b). These results were enhanced by using additional machine learning techniques (bagging of model trees) in Taškova et al. (2006).

In this study, we significantly extend and upgrade the work presented in the preliminary work. Here we investigate the performance of a broader set of state-of-the-art machine learning techniques. We confirm the results from our preliminary work by systematically repeating the experiments using the same machine learning techniques. In addition, we apply other state-of-the-art machine learning techniques, i.e., ensemble methods that aim at improving the predictive performance of a given machine learning technique, using single (learning an ensemble for each target variable separately) as well as multi-target setting (learning an ensemble for all target variables together). We use a more carefully chosen experimental methodology that allows extensive comparisons of the predictive performances of all algorithms and perform statistical significance testing. Finally, we use the model with the best predictive power for generation of vegetation height and canopy cover maps of the Kras region of Slovenia and provide a more comprehensive discussion of the experimental results and the use of the map products.