Predicting Forest Stand Height and Canopy Cover from LANDSAT and LIDAR data using decision trees

Sašo Džeroski[1], Andrej Kobler[2],
Valentin Gjorgjioski[1], Panče Panov[1]
[1]Jožef Stefan Institute, Dept. of Knowledge Technologies
Jamova 39, 1000 Ljubljana, Slovenia
[2]Slovenian Forest Institute
Večna pot 2, 1000 Ljubljana, Slovenia

1 Introduction

With 57,4 % of the national territory covered by forests[7], Slovenia is one of the most forested countries of the European Union. Forests represent the most important CO_2 sink for Slovenia in the framework of the Kyoto protocol to the United Nations convention on climate change, predominantly due to accumulation of forest biomass. Only one quarter of the annually accumulated aboveground biomass is harvested in the form of timber, the rest represents a CO_2 sink according to the articles 3.3 and 3.4 of the Kyoto protocol. Furthermore, there is also a noticeable process of abandonment of arable land and pastures going on in Slovenia, due to the depopulation of rural areas. This leads to 0,4 % of annual forest cover increase, due to the spontaneous afforestation of abandoned agricultural areas ([2]). The forest biomass accumulation and the enlargement of forest areas are not only crucial in the global Slovenian CO_2 budget, but also also important items in trading of the CO_2 emission quotas. Furthermore, the accumulated forest biomass is also an important factor in potential risk of forest fire outbreaks and in forest fire behavior.

Airborne laser scanning (ALS), also termed airborne LIDAR (Light Detection And Ranging), is one of many laser remote sensing techniques [5]. By measuring the round trip time of an emitted laser pulse from the sensor to a reflecting surface and back again, the distance from the sensor to the surface

is determined. The 3D location of the reflecting surface is estimated taking into account the GPS-determined location of the platform. Through periodical deflection of the emitting direction across the flight path by an oscillating or rotating mirror and by the forward motion of the aircraft, a dense cloud of points is sampled from forest vegetation in the form of a swath. Compared to passive, optical remote sensing techniques, laser can penetrate the tree crowns, i.e., look through small gaps in the foliage, and reach the ground. Therefore, the distance to the ground below trees can also be measured. Because of its immediate generation of 3D data, high spatial resolution (in the order of a few centimeters) and accuracy, ALS data is becoming popular for detailed measurements of forest stand height and estimating other forest stand parameters [3].

2 Motivation and related work

Supporting information needs to be provided by a reliable forest monitoring system. The Slovenian Forestry Service already operates such monitoring system [6], which periodically provides a wide range of forestry related information using an extensive network of permanent field sample plots throughout Slovenia. Although this system is tested and reliable, it is also very labor-intensive and costly. Furthermore some of the forest stand attributes, such as canopy cover, are only roughly estimated by visual observation. Other items, such as forest stand height, are monitored only seldom due to technical difficulties of field measurements.

Our motivation for this study was to improve the consistency and accuracy, reduce the costs, and increase the spatial resolution for some of the information gathered by this monitoring system. Specifically we aimed to generate raster maps with 25 m horizontal resolution of forest stand height and canopy closure by using forecasting models based on multi-temporal Landsat ETM+ data. The calibration of the models was also to be done by remotely sensed data, acquired by very high resolution airborne laser scanning (ALS).

There were previous attempts to spatially extrapolate LIDAR-based forest stand metrics using multi spectral satellite data. The authors of [9] estimated relationships among ground measurements of Leaf Area Index (LAI), which is a measure of canopy cover fraction, high resolution IKONOS multi spectral satellite data, and LIDAR data at a ponderosa pine dominated site. They found a significant positive correlation (r=0.76, R^2=0.58) between the IKONOS-derived end member fraction for tree/shade, as well as between end member fraction and LIDAR tree canopy fraction (r=0.76). [4] used

waveform LIDAR (which is a variant of ALS) to predict forest structural attributes. For a dataset of 7700 field plots they found correlation $R^2 = 0.58$ between Landsat TM spectral data and LIDAR mean height, based on large footprint (5 - 15 m) SLICER waveform LIDAR data. The authors of [11] extended SLICER estimates of forest height from sample flight lines to a greater area using segmented Landsat TM data. They achieved correlation of $r^2 = 0.61$ between segment-level Landsat digital numbers and SLICER quantile-based estimates of mean canopy top height. All the mentioned studies used simple regression models to predict forest structure.

3 Description of the data

The study area encompassed 72226 hectares of the Kras region in western Slovenia. It is covered by a highly fragmented landscape of forests, shrubs and pastures. The forests contain mostly oak (Quercus pubescens) and pine (Pinus nigra) of various ages and stand compositions. Multi spectral Landsat ETM+ data were acquired on August 3^{rd}, 2001, May 18^{th}, 2002, November 10^{th}, 2002, and March 18^{th}, 2003, thus capturing the main phenological stages of forest vegetation in the area. The Landsat imagery was first geometrically corrected by orthorectification. Each of the 4 Landsat images was then segmented at two levels of spatial detail. The average image segment sizes were 4 ha for the fine segmentation and 20 ha for the coarse segmentation. Based on the data within each image segment 4 statistics (minimum reflectance, maximum reflectance, average reflectance, standard deviation of reflectance) were computed for each date, for each segmentation level, and for each of the Landsat image channels (2, 3, 4, 5, 7). This way 160 explanatory variables were derived for forest modeling. As the borders of individual segments were not identical between dates and segmentation levels, all 160 variables values were attributed back to individual image pixels.

An east-west transect measuring 2 km by 20 km across the typical part of the region was flown by ALS. The target variables were computed at the level of 25 m by 25 m squares from the LIDAR data set, corresponding to Landsat pixels. The average point cloud density of the LIDAR dataset was 7.5 points/m^2, thus 4687,5 discrete 3D LIDAR returns were contained on average in each square.

The forest stand height for each square (or Landsat pixel) was computed by averaging the heights of the LIDAR-based normalized digital surface model (nDSM) within the 25 m square. A nDSM is a high resolution raster map showing the relative height of vegetation above the bare ground. Our nDSM had a horizontal resolution of $1m^2$. A field validation of the nDSM

on a sample of 120 trees confirmed a vertical RMS error of 0.36 m and a vertical bias of -0.71 m. The canopy closure within this study is defined as the percentage of bare ground within a 25 m square (or a Landsat pixel), covered by a vertical projection of the overlying vegetation, higher than 1 m. The canopy closure CC for each 25 m square was computed from the following ratio between the number of first and the only LIDAR returns: $CC = (Nfirst + N_{only\_1})/(Nfirst + N_{only})$. The first returns are 3D locations of the first encounter of the emitted laser pulse with a reflecting surface in vegetation (several returns on a single pulse are possible). First returns mainly reside in the upper layers of forest. The only returns are reflected from solid objects, such as bare ground. Some of the only returns (denoted as $N_{only\_}1$ in the above formula) are reflected also from lower layers of forest. For the purposes of this study all the N_{only_1} returns were defined as those only returns, that exceeded 1 m relative height above bare ground digital terrain model (also derived from LIDAR dataset).

As it was described before we can make predictions about the forest stand height by using LANDsat images. The prediction task consist of building predictive models by using data mining algorithms and validating the models by using standard validation techniques. The problem of predicting FSH and canopy cover itself implies the use of techniques for multiple prediction because we have several different target variables that we want to predict in the same time.

4 Data analysis methodology

The analysis of data was done using several different ML algorithms for building decision trees. We used regression and model trees implemented in the WEKA[10] environment and predictive clustering trees implemented in the CLUS[8] system.

4.1 Regression/model trees

Decision trees are tree-shaped symbolic models that are frequently used in machine learning and data mining, in most cases for prediction tasks. Classification trees are the most common type and are used to predict a symbolic attribute, which is called the class. A second type of decision trees are regression trees. The latter can be used to predict the value of a numeric attribute. If the leaf contains a linear regression model that predicts the target value of examples that reach the leaf, it is called a model tree. Model trees have advantages over regression trees in both compactness and prediction accu-

racy, attributable to the ability of model trees to exploit local linearity in the data. Other difference is that regression trees will never give a predicted value lying outside a range observed in the training cases, whereas model trees can extrapolate. We used M5' model trees implemented in the WEKA DM Suite[10].

4.2 Predictive Clustering Trees

Data mining task of predicting forest stand height and canopy cover requires several variables to be predicted. We can do this by building separate models for each attribute or by using a model that can predict several target variables at once. Predictive Clustering Trees(PCT)[1] are such a methodology. They are implemented in a CLUS system, which can also built ordinary regression trees.

4.3 Experimental setup

Models were built using four different algorithms: M5 regression trees, M5 model trees, CLUS regression trees, CLUS predictive clustering trees. After the results of this analysis were obtained, we decided to go one step further to make hierarchical clustering of the target variables and to use CLUS to process four groups of variables. We used different pruning methods and best results were obtained using M5 pruning. To get smaller trees, we also considered using a depth constraint and made some experiments but the results were not as good as was expected. The models were validated by 10-fold cross-validation technique to obtain reliable estimates of their predictive accuracy. The results of the experiments are presented in the next section.

5 Results

We will present a table of the obtained results from the experiments using Pearson correlation for the target variables. As mentioned before we have 11 numerical target variables we want to predict: sklep which is a percentage of vegetation cover within a pixel, ndsm which is the highest reflection in a pixel(defined in prev.section), delveg is the percentage of vegetation inside the pixel and next variables are maximum of vertical vegetation profile inside the pixel, and there percentiles(99,95,75,50,25,10,5) respectively. The results from the experiments are presented in Table 1. As we can see from the table we have the correlation coefficients for all experiments. The first two columns contain results from M5 regression and model trees. We can see

		WEKA		CLUS			
N^o	Target Attr.	M5 RT	M5P MT	M5 RT	MORT	HC	HC group
1	sklep	0.857	0.862	0.847	0.843	0.848	Cluster1
2	delveg	0.857	0.861	0.847	0.842	0.849	Cluster1
3	ndsm	0.877	0.885	0.871	0.867	0.871	Cluster2
4	vpv1_hmx	0.815	0.820	0.811	0.796	0.811	Cluster2
5	vpv1_h99	0.823	0.834	0.818	0.807	0.821	Cluster2
6	vpv1_h95	0.830	0.841	0.823	0.814	0.826	Cluster2
7	vpv1_h75	0.828	0.836	0.819	0.815	0.815	Cluster3
8	vpv1_h50	0.802	0.813	0.796	0.794	0.795	Cluster3
9	vpv1_h25	0.753	0.761	0.744	0.744	0.744	Cluster3
10	vpv1_h10	0.676	0.672	0.659	0.655	0.662	Cluster4
11	vpv1_h05	0.602	0.606	0.579	0.571	0.585	Cluster4

Table 1: Pearson correlation coefficient of the obtained models

that the correlation coefficient of model trees is little higher than that of the regression trees. The reason for this was mentioned in previous section. The next three columns contain results from regression trees induced by CLUS system. The models built by WEKA are similar with the ones built by CLUS. Most interesting results were obtained by building one multi-objective regression tree containing all the target variables. The correlation coefficients are comparable with regression tree models for every target variable separately. The usefulness of this model is that instead of heaving 11 separate models you have only one model that describes all variables with accuracy differing from the original regression models very little. The last column we present the correlation coefficients of multi target model. We built four models by grouping the target variables into clusters with hierarchical clustering. The results are comparable with the ordinary regression trees and the accuracy is little higher than the original MORT model.

6 Discussion

With our approach to modeling forest structure from integrated LIDAR and multi spectral satellite data we used machine learned regression trees instead of simple regression models, and multi temporal satellite data instead of mono temporal data. While the former enabled us to make simple modeling, the latter enabled us to implicitly include into our models the temporal dynamics, typical of individual forest stand types. We believe (assume, suppose in any case, we did not try to prove this) that this improved our model correlations.

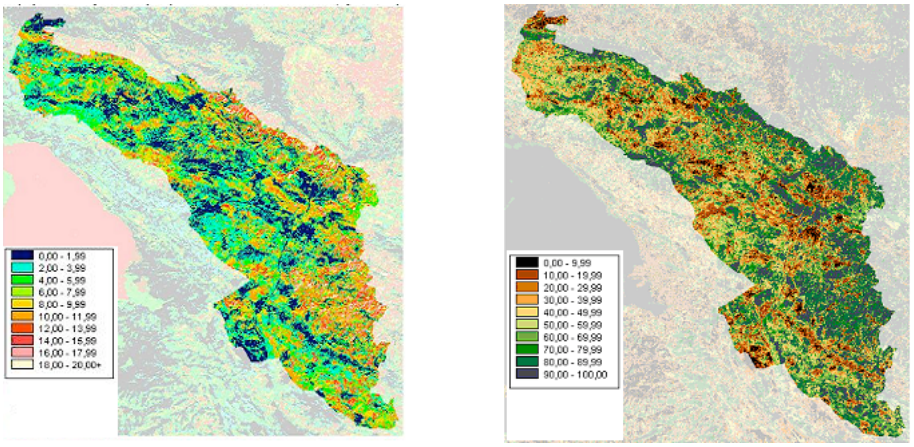

Figure 1: Maps of predicted forest stand height(left) and canopy cover(right)

In fact our correlations for very heterogeneous forest vegetation were better or comparable with previous studies done on much smaller and relatively homogeneous sites (e.g.[9]), or done using large footprint SLICER -calibrated models.

The visual inspection by a forestry expert of the resulting maps should that they correspond to the actual forest cover in the Kras region, both in terms of forest stand height as well as canopy closure. No such continuous maps existed previously for this region.

7 Conclusions, further work

Forest stand height and canopy closure maps such as the once generated within our study are a very effective tool for detecting ongoing spatial processes in forested landscape. These processes involve both enlargement of forest areas by spontaneous afforestation of abandoned agricultural land, as well as vertical growth and gradual closing of canopy cover of existing forest stands. These maps can be used not only in the process of monitoring the forest biomass accumulation and CO_2 sink in the Kyoto framework, but also in the forest fire modeling. Due to their spatial continuity (vs. discrete sampling layout of current forest monitoring schemes) the potential applications also include the study of forest habitats and transitional agricultural-forest habitats, visual landscape assessments, land use suitability analysis, visibility analysis for cell phone networks etc.

Although such maps could be generated with exceeding precision and accuracy purely from LIDAR data, this seems impractical for the forseable

future due to the very high cost of high resolution ALS data (in our case 660 US$ / km^2). On the other hand, the price of Landsat ETM+ data for a 4-date multi temporal coverage was only about 0,1 US$ / km^2. Using Landsat data as the main data source therefore ensures a very acceptable cost benefit ratio. On the other hand ALS as used here for model calibration seems a very good replacement for sample plot field measurements of forest stand height and canopy closure, due to the even higher costs and difficulty or imprecision of the field measurements.

In further work the following issues should be investigated: (1) to lower the cost of the ALS data needed for model calibration, only ALS data within sampling plots could be used, (2) Analysis of the influence of the relative size of sampling plots on the quality of the resulting models, (3) Upgrading of LANDsat data by radiometric correction, (4) adding quantile-based estimators at the segment level into the models.

References

[1] H. Blockeel. *Top-down induction of first order logical decision trees*. PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, 1998. `http://www.cs.kuleuven.ac.be/~ml/PS/blockeel98:phd.ps.gz`.

[2] FAO-Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment Update 2005, Slovenia Country Report*. 2005.

[3] Hyyppa H. Litkey P. Yu X. Haggren H. Ronnholm P. Pyysalo U. Juho Pitkanen J. Maltamo M. Hyyppa, J. *Algorithms and Methods of Airborne Laser-Scanning for Forest Measurements*. International Archives of Photogrammetry and Remote Sensing, Vol XXXVI, 8/W2, Freiburg, Germany, 2004.

[4] A. Hudak S.A. Acker Lefsky M.A., W.B. Cohen and J.L. Ohmann. Integration of lidar, landsat etm+ and forest inventory data for regional forest mapping. In *Proceedings of the ISPRS Workshop Mapping surface structure and topography by airborne and spaceborne lasers,9-11 NOVEMBER 1999*, 1999.

[5] Raymond M. Measures. *Laser remote sensing: fundamentals and applications*. Malabar, Fla., Krieger Pub. Co., 1992. 510 p. G70.6.M4, 1992.

[6] SFS Slovenian Forestry Service. *Slovenian forest and forestry.* Zavod za gozdove RS, Ljubljana, Slovenia, 24 pp, 1998.

[7] SFS Slovenian Forestry Service. *Slovenian forest cover statistics 2004, Unpublished manuscript.* Zavod za gozdove RS, Ljubljana, Slovenia, 2006.

[8] J. Struyf and S. Džeroski. Constraint based induction of multi-objective regression trees. 2005. Submitted to the Workshop on Knowledge Discovery in Inductive Databases (KDID'05) at the 17th European Conference on Machine Learning (ECML'05).

[9] Rowell E. Chen X. Dykstra D. Vierling, L. and K. Vierling. *Relationships among airborne scanning LiDAR, high resolution multispectral imagery, and ground-based inventory data in a ponderosa pine forest.* 2002.

[10] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005. 2nd Edition.

[11] A.W. Wulder and D. Seeman. Forest inventory height update through the integration of lidar data with segmented landsat imagery. *Canadian Journal of Remote Sensing, 29, 536-543*, 2003.