Using Decision Trees to Predict Forest Stand Height and Canopy Cover from LANDSAT and LIDAR Data

Sašo Džeroski¹, Andrej Kobler, Valentin Gjorgjioski², Panče Panov³

Abstract
The motivation for this study was to improve the consistency and accuracy, and increase the spatial resolution of some of the supporting information to the forest monitoring system in Slovenia by using data mining techniques. Specifically we aim to generate raster maps with 25 m horizontal resolution of forest stand height and canopy cover, for the Kras region of Slovenia. We used predictive models based on multi-temporal LANDSAT data and calibrated it with high resolution airborne laser scanning (ALS) data. The visual inspection by a forestry expert of the resulting maps showed that the generated maps correspond to the actual forest cover in the Kras region, both in terms of forest stand height as well as canopy cover.

1. Introduction
Slovenia is one of the most forested countries of the European Union, with 57.4 % of the national territory covered by forest (Slovenian Forestry Service, 2006). Forests represent the most important CO₂ sink for Slovenia in the framework of the Kyoto protocol to the United Nations (http://lawref.org/KYOTO/index.html) convention on climate change, predominantly due to accumulation of forest biomass. One quarter of the annually accumulated aboveground biomass is harvested in the form of timber and the rest represents a CO₂ sink according to the articles 3.3 and 3.4 of the above mentioned protocol. In Slovenia there is also a noticeable process of abandonment of arable land due to the depopulation of rural areas. This leads to 0.4 % of annual forest cover increase, as a result of the spontaneous afforestation of abandoned agricultural areas (FAO, 2005). The forest biomass accumulation and the enlargement of forest areas are not only crucial in the global Slovenian CO₂ budget, but are also important items in the trading of CO₂ emission quotas. Accumulated forest biomass is an important factor in potential risk of forest fire outbreaks and in forest fire behavior.

The main idea of this study was to improve the consistency and accuracy (in order reduce the costs) and increase the spatial resolution of some of the information gathered by forest monitoring system. Specifically we are aimed to generate raster maps with 25 m horizontal resolution of forest stand height and canopy cover of Kras region of Slovenia by using predictive models based on multi-temporal LANDSAT ETM+ data. The calibration of the models was done by remotely sensed data, acquired by very high resolution airborne laser scanning (ALS).

Airborne laser scanning (ALS), also termed airborne LIDAR (Light Detection And Ranging), is one of many laser remote sensing techniques (Raymond et al., 1992) that is used in forestry for estimation of different parameters. 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 (Hyyppä et al., 2004).

¹Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Emails : Saso.Dzeroski@ijs.si, Valentin.Gjorgjioski@ijs.si, Pance.Panov@ijs.si
²Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, Slovenia
Email: Andrej.Kobler@gozdis.si
³Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, Slovenia
Email: Panče.Panov@ijs.si

EnviroInfo 2006 (Graz)
Managing Environmental Knowledge
Copyright © Shaker Verlag 2006. ISBN: 978-3-8322-5321-9
The rest of the paper is structured as follows. In section 2 we explain the motivation for this study and give some references to some related work in this field. Section 3 describes the dataset and section 4 describes data mining methodology used to build the predictive models. In section 5, we present and discuss the results. Section 6 covers the conclusions and plans for future work.

2. Motivation and related work

The motivation for this study is to provide supporting information to the forest monitoring system that would be useful in predicting the accumulated forest biomass. The Slovenian Forestry Service operates such a monitoring system (SFS, 1998), which periodically provides a wide range of forestry related information using an extensive network of permanent field sample plots throughout Slovenia. This system is tested and it is proven to be reliable, but it is also very labor-intensive and costly. Furthermore, some of the forest stand attributes, such as canopy cover, can only be roughly estimated by visual observation. Other items, such as forest stand height, can be monitored only seldom due to technical difficulties of field measurements. The work presented in this paper is focused on generating raster maps of forest stand height and canopy cover by using predictive models based on multi-temporal Landsat ETM+ data that are calibrated with high resolution LIDAR data.

There were previous studies to spatially extrapolate LIDAR-based forest stand metrics using multispectral satellite data. Vierling et al. (2002) estimated relationships among ground measurements of leaf area index (LAI), which is a measure of canopy cover fraction, high resolution IKONOS multispectral satellite data, and ALS data at a ponderosa pine dominated site. They found a significant positive correlation ($r=0.76$) between the IKONOS-derived end member fraction for tree/shade, as well as between end member fraction and LIDAR tree canopy fraction ($r=0.76$). Lefsky et al. (1999) used waveform LIDAR (which is a variant of ALS) to predict forest structural attributes. For a dataset of 7700 field plots they found $R^2=0.581$ between Landsat TM spectral data and LIDAR mean height, based on large footprint (5 – 15 m) SLICER waveform lidar data. Wulder & Seeman (2003) extended SLICER estimates of forest height from sample flight lines to a greater area using segmented LIDAR TM data. They achieved $r^2=0.61$ between segment-level LIDAR digital numbers and SLICER (Scanning Ladar Imager of Canopies by Echo Recovery) quantile-based estimates of mean canopy top height. All the mentioned studies used simple regression models to predict forest structure. Hudak et al. (2002) also used kriging and cokriging in combination with linear regression.

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 3rd 2001, May 18th 2002, November 10th 2002, and March 18th 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) and 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 2km by 20km across the typical part of the Kras region was flown by ALS. The distance from the sensor to the surface is determined by measuring the round trip time of an emitted laser pulse from the sensor to a reflecting surface and back. The 3D location of the reflecting surface is estimated taking into account the GPS-determined location of the platform. A dense cloud of points is sampled from forest vegetation in the form of a swath, through a periodical deflection of the emitting direction across the flight path by an oscillating or rotating mirror and by the forward motion of the aircraft. Compared to passive, optical remote sensing techniques, laser beams can penetrate through the tree crowns, i.e., look through small gaps in the foliage, and reach the ground. Therefore, the distance to the ground below the trees can also be measured. The average point cloud density of the LIDAR dataset was 7.5 points/m², thus 4687.5 discrete 3D LIDAR returns were contained on average in each square. The target variables were computed at the level of 25 m by 25 m squares from the LIDAR data set, corresponding to Landsat pixels. We have 11 target variables describing the forest area and we will mention and describe them in this section.

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 1 m. 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 (Kobler et al., 2006).

The canopy cover 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 cover (CC) for each 25 m square was computed from the following ratio between the number of first and the only LIDAR returns: \( CC = \frac{N_{\text{first}}}{N_{\text{first}} + N_{\text{only}}} \). 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 \( N_{\text{only}} \) in the above formula) are reflected also from lower layers of forest. For the purposes of this study all the \( N_{\text{only}} \) 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 described before, we learn to make predictions about the forest stand height (FSH) and canopy cover (CC) by using Landsat images. The prediction task consists of building predictive models by using data mining algorithms and validating the models by using standard validation techniques. The problem of predicting FSH and CC itself implies the use of techniques for multiple prediction because we have several target variables that we want to predict at the same time.

4. Data mining methodology

The analysis of datasets was done using several different data mining techniques. We used regression and model trees (Quinlan, 1992) implemented in the WEKA (Witten et al., 2005) environment, model rules implemented CUBIST, commercial product from Rulequest Research (http://www.rulequest.com/), multi-objective regression trees (Struyf et al., 2006) implemented in the CLUS system and Microsoft Decision Trees algorithm implemented in SQL Server Analysis Services 2005.
4.1 Regression/model trees

Decision trees (Quinlan, 1986) 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 is regression tree (Quinlan, 1992). 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 (Quinlan, 1992) have advantages over regression trees in both compactness and prediction accuracy, attributable to the ability of model trees to exploit local linearity in the data. Another difference is that regression trees will never give a predicted value lying outside the range observed in the training cases, whereas model trees can extrapolate. We used the M5’ algorithm for building model trees, implemented in the WEKA DM Suite (Witten et al., 2005). An example of decision/regression tree built by WEKA is shown on the Figure 1(a).

4.2 Cubist’s model rules

Cubist (http://www.rulequest.com/) is a tool for generating rule-based predictive models from data. It builds models containing one or more rules, where each rule is a conjunction of conditions associated with a linear expression. The meaning of a rule is that, if a case satisfies all the conditions, then the linear expression is appropriate for predicting the target value. A Cubist model resembles a piecewise linear model, except that in this case models can overlap. A Cubist model consists of an unordered collection of rules, each of the form

\[ \text{If conditions then } \text{linear model} \]

A rule indicates that, whenever a case satisfies all the conditions, the linear model is appropriate for predicting the value of the target attribute. If two or more rules apply to a case, then the values are averaged to arrive at a final prediction. Each rule also carries some descriptive information: the number of training cases that satisfy the rule’s conditions, their target values’ mean and range, and a (rather erratic) estimate of the expected error magnitude of predictions made by the rule. Within the linear model, the attributes are ordered in decreasing relevance to the result. The final part of the model output describes the performance of the model on training data, and on the new cases in the test data (if present). An example of the generated Cubist rule is given on the Figure 1 (b).

A rule indicates that, whenever a case satisfies all the conditions, the linear model is appropriate for predicting the value of the target attribute. If two or more rules apply to a case, then the values are averaged to arrive at a final prediction. Each rule also carries some descriptive information: the number of training cases that satisfy the rule’s conditions, their target values’ mean and range, and a (rather erratic) estimate of the expected error magnitude of predictions made by the rule. Within the linear model, the attributes are ordered in decreasing relevance to the result. The final part of the model output describes the performance of the model on training data, and on the new cases in the test data (if present). An example of the generated Cubist rule is given on the Figure 1 (b).

Fig. 1: Models generated by WEKA and CUBIST
4.3 Multi-Objective Regression Trees

The data mining task of predicting forest stand height and canopy cover requires several variables to be predicted at the same time. We can do this by building separate models for each variable or by using a methodology that can predict several target variables at once. Multi-Objective Regression Trees (MORT) (Struyf, 2006) are such a methodology. MORTs are regression trees capable of predicting several numeric variables at once. This has two main advantages over building a separate regression tree for each target:

1. a single MORT is usually much smaller than the total size of the individual trees for all variables, and
2. a MORT explicates dependencies between the different target variables. MORTs are implemented in a CLUS system, which can also build ordinary regression trees. One example of MORT is shown on Figure 3.

![Fig. 2: Example of a MORT generated with the CLUS system](image)

4.4 Microsoft Decision Trees

The Microsoft Decision Trees algorithm is a classification and regression algorithm provided by Microsoft SQL Server 2005 Analysis Services (SSAS) for use in predictive modeling of both discrete and continuous variables. It also supports automatic feature selection. Feature selection automatically chooses the variables in a dataset that are most likely to be used in the model. This algorithm builds a data mining model by creating a series of splits, also called nodes, in the tree. The algorithm adds a node to the model every time an input column is found to be significantly correlated with the predictable column. If more than one column is set to predictable the algorithm builds a separate decision tree for each predictable column. In our case we will use this algorithm just to predict the main attributes, CC and FSH.

5. Experimental setup

As mentioned in section 2, the dataset that was analyzed consisted of 160 descriptive attributes and 11 target variables. There were 64000 examples of which 60607 described the vegetation outside a settlement and were used for building the predictive models. We used four different systems for learning predictive models: WEKA’s M5 algorithm was used to build regression and model trees using the default pruning parameters for every target variable. CUBIST was used to build model rules for every target variable with the default parameters for rule pruning. CLUS was used for building regression trees for every target variable separately and multi-objective regression trees (MORT) for several groups of target variables and for all target attributes at the same time. We used the default CLUS heuristics (intra cluster variance) for building trees and M5 pruning technique. The Microsoft decision tree algorithm which is included in SQL...
Server 2005 was also used to build a model tree. For CLUS, groups of target variables were formed by analysis of correlations (a matrix of correlations between the target variables was provided by the domain expert) between the target variables. In particular, we performed hierarchical (agglomerative) clustering of the target variables, where the similarity of two variables is determined from their correlation. The R (http://www.r-project.org/) system for statistical computing was used for this purpose. Based on the results of hierarchical clustering we decided to group the target variables in four groups. The idea of grouping of the target variables was to examine if simultaneous prediction of highly correlated target variables would improve the performance of MRTs vs. the simultaneous prediction of all target variables. The performance of the models was estimated with 10-fold cross-validation.

6. Results and Discussion

In this section, we present the results of the experiments performed on the dataset described in Section 1. As mentioned before, we have 11 numerical target variables we want to predict denoted with the following names: CC, which is a percentage of vegetation cover within a pixel; FSH which is the highest reflection in a pixel (defined in previous section), delveg is the percentage of vegetation inside the pixel; vpv1_hmx vertical vegetation profile inside the pixel at maximum height; and the percentiles (99, 95, 75, 50, 25, 10, 5) of vegetation profiles. The results from the experiments are presented in the Tables 1 and 2.

<table>
<thead>
<tr>
<th>Target Variable</th>
<th>WEKA M5 RT</th>
<th>WEKA M5P MT</th>
<th>CUBIST</th>
<th>SSAS 2005 MDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.73</td>
</tr>
<tr>
<td>FSH</td>
<td>0.88</td>
<td>0.89</td>
<td>0.87</td>
<td>0.78</td>
</tr>
<tr>
<td>delveg</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>vpv1_hmx</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>vpv1_h99</td>
<td>0.82</td>
<td>0.83</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>vpv1_h50</td>
<td>0.83</td>
<td>0.84</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>vpv1_h75</td>
<td>0.83</td>
<td>0.84</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>vpv1_h50</td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>vpv1_h25</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>vpv1_h10</td>
<td>0.68</td>
<td>0.67</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>vpv1_h05</td>
<td>0.60</td>
<td>0.60</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Correlation coefficients of the obtained WEKA, CUBIST and MDT models

The tables give the correlation coefficients for all experiments we performed. In SQL Serve some of the experiments were not performed because of technical problems with SQL Server (out of memory), however experiments were performed for the most important variables for our study which are CC and FSH. In the Table 1, the first two columns contain results from the WEKA M5 regression and model trees. We can see that the correlation coefficient of model trees is slightly higher than that of regression trees. Last column of the table gives the correlations of the CUBIST rules. These are comparable to the ones obtained by WEKA’s model trees.
Results of CLUS and SQL Server are presented in Table 2. In the first column are the correlation coefficients obtained using SQL Server, while the next three columns contain results from regression trees induced by CLUS. The models built by WIRA M5 are similar to the ones built by CLUS. WIRA M5 results from Table 1 are given also in Table 2 for reference. Most interesting results were obtained by building one multi-objective regression tree for all target variables. The correlation coefficients are comparable with those of regression tree models built for every target variable separately. The usefulness of this model is that instead of having 11 separate models we have only one model that describes all variables with minimal loss of memory. The last column presents the correlation coefficients for the multi-target model built for correlated groups of target variables. We built four models by grouping the target variables into four clusters with hierarchical clustering. The results are comparable with those of ordinary regression trees and the accuracy is a little higher than that of the single MORT model predicting all targets.

In CLUS we are able to generate Python functions that represent the obtained decision tree. These functions were later used for predicting values for our target variables from the data and were plugged into the system which draws maps. Giving satellite image data to this system, it predicts target variables, and produces useful maps, some of which are presented in Figure 5.
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 models easily, the latter enabled us to implicitly include into our models the temporal dynamics, typical of individual forest stand types. We believe that this approach to modeling improved our model correlations. In fact, our correlations for very heterogeneous forest vegetation were better or comparable with previous studies done on much smaller and relatively homogeneous sites (Vierling et al., 2002), or done using large footprint SLICER-calibrated models.

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

7. Conclusions and further work

Forest stand height and canopy cover maps such as the ones generated within our study are a very effective tool for detecting ongoing spatial processes in forested landscapes. 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 for existing forest stands. These maps can be used not only in the process of monitoring the forest biomass accumulation and CO$_2$ in the Kyoto framework, but also in forest fire modeling. Due to their spatial continuity (vs. the 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 foreseeable 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 Landat ETM+ data for a 4-date multi temporal coverage was only about 0.1 US$ / km$^2$. Using Landat 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 cover, 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 and (5) Try to use different data mining techniques to build predictive models.

Bibliography


SFS Slovenian Forestry Service (1998): Slovenian forest and forestry. Zavod za gozdove RS, 24 pp


Copyright © Shaker Verlag 2006. ISBN: 978-3-8322-5321-9