Using a Multistructural Object-Based LiDAR Approach to Estimate Vascular Plant Richness in Mediterranean Forests With Complex Structure

Javier Lopatin, Mauricio Galleguillos, Fabian E. Fassnacht, Andrés Ceballos, and Jaime Hernández

Abstract—A multistructural object-based LiDAR approach to predict plant richness in complex structure forests is presented. A normalized LiDAR point cloud was split into four height ranges: 1) high canopies (points above 16 m); 2) middle-high canopies (8–16 m); 3) middle-low canopies (2–8 m); and 4) low canopies (0–2 m). A digital canopy model (DCM) was obtained from the full normalized LiDAR point cloud, and four pseudo-DCMs (pDCMs) were obtained from the split point clouds. We applied a multiresolution segmentation algorithm to the DCM and the four pDCMs to obtain crown objects. A partial least squares path model (PLS-PM) algorithm was applied to predict total vascular plant richness using object-based image analysis (OBIA) variables, derived from the delineated crown objects, and topographic variables, derived from a digital terrain model. Results showed that the object-based model was able to predict the total richness with an $r^2$ of 0.64 and a root-mean-square error of four species. To predict richness using OBIA variables to predict richness. Furthermore, high-medium canopies (8–16 m) showed the biggest correlation with the total plant richness within the structural segments of the forest.

Index Terms—Bootstrapping, LiDAR, object-based analysis, partial least squares path model (PLS-PM), vascular plant richness.

I. INTRODUCTION

Biodiversity is an essential element of the Earth system from which all humans benefit directly or indirectly [1]. The current and future states of biodiversity have acquired increasingly more scientific and political relevance. This is based on the greater knowledge about the adverse effects that the decrease in biodiversity may produce on ecosystem services on which human well-being depends [2], [3].

Theoretical and empirical studies have suggested that local biodiversity is strongly influenced, in a positively correlated fashion, by the environmental heterogeneity [4]. More complex environments can host a greater number of ecological niches, which, in turn, can be colonized and inhabited by a greater number of species [5], [6]. Different remote sensors allow capturing spatial heterogeneity in two ways: spatial variability or heterogeneity (horizontal variability) and topographical—structural variability or heterogeneity (vertical variability) [6].

Some studies have used topographical information, i.e., microrelief and their spatial variability [7], [8], along with predictors describing the configuration and variability of the vertical structure of the aboveground vegetation [6] to estimate environmental heterogeneity. Topographical and structural information captured by LiDAR sensors may therefore be suitable to provide biodiversity-related variables in different ecosystems [9], particularly in forests with high vertical complexity [6]. Such data are expected to have large potential for biodiversity characterization [10]. For example, the suitability of a combination of LiDAR and hyperspectral data to estimate vascular plant richness in deciduous Mediterranean forests was demonstrated in [11]. A link between imaging spectroscopy and LiDAR with floristic composition and forest structure was presented by Higgins et al. [12], and the influences of terrain and vegetation structures on local avian species richness in two mixed-conifer forests were examined by Vogeler et al. [13].

Object-based image analysis (OBIA) is a segmentation/classification and feature extraction technique that has been used among other applications for environmental monitoring requirements, conservation goals, spatial planning enforcement, or ecosystem-oriented natural resources management [14]. The OBIA approach is able to group homogeneous pixels into meaningful objects based not only on their radiometric values but also on multiple attributes such as texture, context, and shape. This approach was found to produce more reliable and less noisy results when applied to high-resolution images [15], [16]. A large set of studies combined LiDAR data with OBIA technique for segmentation or feature extraction purposes. For example, a two-stage approach for characterizing the structure of Pinus sylvestris stands in forests of central Spain was presented in [17]. The urban landscape structure at the parcel level was also investigated by Zho and Troy [18] who applied an object-oriented approach on high-resolution digital aerial imagery and LiDAR data for the Baltimore area.
The objective of this study was to develop a method to predict local vascular plant richness in a secondary mixed deciduous Mediterranean forest. The underlying idea was to use only structural and topographical attributes derived from LiDAR in a multistructural object-oriented approach.

II. STUDY SITE AND DATA SET

A. Study Area

The study area, Monte Oscuro, is located in the Andes foothills of central Chile in the Maule Region (35°07'00" S, 79°35'30" W) [see Fig. 1(a)]. This area is associated with the Sub-Mediterranean Temperate bioclimatic zone, with a mean annual precipitation of 1000 mm, which is mainly concentrated between April and October. The site has an extent of 1295 ha, a mean altitude of 1075 m above sea level, and mostly southern aspect. Monte Oscuro is covered by secondary deciduous forest dominated by *Nothofagus obliqua* (Mirb.) Oerst., mixed with sclerophyll species.

B. Ground Data

We used the vascular plant richness as a response variable due to the clear dominance of this taxonomic group in Mediterranean forests and also because it is essential for the trophic network and ecosystem functioning [19]. A botanic survey was performed between January 2013 and January 2014. A 200 m × 200 m regular grid was applied to locate the square nested plots [20]. Steep zones (< 45%) and sites too close to trails (20 m) were not considered in the sampling [see Fig. 1(b)] to facilitate operational endeavor and avoid the effect of borders and bare soil in the plots. Finally, 80 square nested plots were located in the study site.

In each sampling unit, species within three height layers were registered and determined: trees (height of 2 m or more), shrubs (less than 2 m of height), and herbs (non-woody plants). Each plot was composed of six nested subplots. The largest one, in which only trees were registered, occupied 225 m² (15 m × 15 m), and it had its vertices aligned to the four cardinal points (N, E, S, and W). Shrubs were identified in two subplots of 25 m² (5 m × 5 m). Herbs were registered in three sub-subplots of 1 m² [see Fig. 1(c)].

Following the suggestions in [21], we considered the nested subplots to be a representative sample of the total plot area. Therefore, for each 225-m² plot, the total richness of vascular species was obtained by simply summing up all species encountered in the plots and subplots of the three height layers. Species occurring in more than one height layer (e.g., tree and shrub layers) were only counted once. The histogram of the richness values shows a bimodal distribution [see Fig. 2(a)]. This data distribution can be attributed to the fact that the area is considered as a second growth natural forest under some areas with silvicultural intervention, resulting in a high number of relatively young stands that were established after harvesting activities [22]. In Fig. 2(b), a distribution of the sample richness is illustrated.

C. LiDAR Data

A discrete pulse LiDAR from the Harrier 54/G4 Dual System (manufactured by Trimble Industries and provided by Digimapas Chile Ltd., Santiago, Chile) was used to survey the study site on March 2011. Characteristics of the Harrier 54/G4 Dual System include scanning frequency of 100 Hz, pulse rate of 100 kHz, scan angle [field of view (FOV)] of ±22.5°, laser beam divergence (instantaneous FOV) of 0.5 mrad, and an intensity of returns from a 1550-nm laser, obtaining an average point cloud density of 4.64 points/m² and a footprint of 29 cm. This instrument was also used in other studies, for example, in [22].

III. METHODS

A. Multistructural Height-Based Threshold

The LiDAR point cloud was classified into ground and nonground points for bare-earth extraction according to Briese [23]. Normalization of the LiDAR point cloud was performed to transform the altitude data units from meters above sea level to meters above the ground. Then, the normalized point cloud was classified into four classes of canopy heights (see Table I). These classes are based on the methods proposed by Etienne and Prado [24], who showed that they provide a representation of the vegetation structure of dominant species in the South-center of Chile.
TABLE I
HEIGHT-BASED THRESHOLD CLASSIFICATION

<table>
<thead>
<tr>
<th>Height-based threshold</th>
<th>Altitudinal ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Canopies (HC)</td>
<td>Over 16 m.</td>
</tr>
<tr>
<td>Middle-High Canopies (MHC)</td>
<td>8 – 16 m.</td>
</tr>
<tr>
<td>Middle-Low Canopies (MLC)</td>
<td>2 – 8 m.</td>
</tr>
<tr>
<td>Low Canopies (LC)</td>
<td>0 – 2 m.</td>
</tr>
</tbody>
</table>

B. DEMs
The point cloud classified into ground and nonground points was interpolated to obtain a digital terrain model (DTM) and a digital canopy model (DCM). Furthermore, interpolation of the four segmented normalized point clouds was performed to obtain four pseudo-DCMs (pDCMs). All the obtained digital elevation models (DEMs) had 1 m² spatial resolution.

C. Object-Based Analysis
An object-oriented segmentation approach was used to delineate crown objects (polygons) of small size (less than 400 m²), based on the DCM and the four derive pDCMs. For the segmentation of the crown objects, we took into account the scale, the homogeneity, the shape, and the color of the DCM and the pDCMs. The scale is related to the minimum size required to identify a particular object, which depends on the resolution of the images [17]. Homogeneity is described by a mutually exclusive interaction between color and shape. Color refers, in this case, to the height of the objects, whereas shape is divided into two equally exclusive properties, i.e., smoothness and compactness, which define the boundaries of the polygons (crown objects) and their transition to others, respectively [14].

We applied a multiresolution segmentation algorithm, in eCognition software, to the DCM and the four pDCMs (scale parameter = 5, shape criterion = 0.3, and compactness criterion = 0.5). Results, which can be understood as crown objects, can be compound by a part of a crown, a single crown, or a group of crowns.

The crown objects derived from the DCM cover the full height ranges of the data set (LC to HC). These polygons were subsequently classified into the height classes given in Table I. No height-based classification was performed to the crown objects derived from the pDCMs (see Fig. 3).

D. Model Parameterization
Several topographical and OBIA predictor variables were obtained for the regression analysis. The topographic variables used are the generalized surface, the maximum heights, and the DEM, estimated in SAGA GIS(r) software. For each variable, the minima, mean, and maxima values within the plots were estimated. These variables were selected based on the findings of Ceballos et al. [11], who demonstrated that these topographic variables had high correlation with the total vascular richness in the same study area.

The object-based predictor variables were calculated by extracting: 1) the total number of crown-object polygons associated to each richness sampling plot; 2) their coverage (m²/ha); 3) the standard deviation; and 4) the average of their areas. In the case of the crown objects derived from the DCM, a total of 16 variables were obtained (i.e., 4 for each altitudinal range given in Table I; compare also Fig. 3), and in the case of each pDCM, 4 variables were obtained. Therefore, a total of 32 predictor variables were derived from OBIA (see Table II).

TABLE II
SUMMARY OF THE OBIA PREDICTOR VARIABLES

<table>
<thead>
<tr>
<th>Source of the variables</th>
<th>Number of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified DCM crown-objects</td>
<td>4+4+4+4</td>
</tr>
<tr>
<td>High Canopies (HC) crown-objects</td>
<td>4</td>
</tr>
<tr>
<td>Middle-High Canopies (MHC) crown-objects</td>
<td>4</td>
</tr>
<tr>
<td>Middle-Low Canopies (MLC) crown-objects</td>
<td>4</td>
</tr>
<tr>
<td>Low Canopies crown-objects</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
</tr>
</tbody>
</table>

E. Statistical Analysis
We used the partial least squares path model (PLS-PM) algorithm to predict total vascular plant richness. PLS-PM works as a fusion of regression analysis, principal components analysis, and path analysis [25]. This combination seems especially interesting in a remote sensing context, where often highly correlated predictors are applied and data reduction methods were found to improve model performances [26].

PLS-PM is a nonparametric method designed for prediction analysis and the development of theoretical models [27], which can be graphically represented by a path diagram that shows how the various elements (variables) are related to each other [25]. This algorithm was performed using the R-Project package “plspm.”

PLS-PM has a series of model validation techniques that work on different levels (i.e., loading, weights, and r²) [25], [27], with an overall validation of the model obtained with a pseudo-goodness-of-fit (GoF) proposed by Tenenhaus et al. [25]. In addition, a bootstrapping validation technique with 1000 boots was performed to assess the quality of the model. In each bootstrap iteration, we used the samples that were not drawn in the bootstrap (on average, 36.8% of the total
For the selection of the predictor variables in latent variables, the heuristic “10 cases per indicator” provided by Tenenhaus et al. [25] to maximize the efficiency of PLS-PM was used. This statement is a recommendation for robustness in PLS-PMs, where the authors made a series of empirical tests to provide this conclusion. The predictor variables with highest loading and communality were selected in each latent variable.

### IV. Results and Discussion

#### A. Variable Selection

Following the heuristic “10 cases per indicator” provided by Tenenhaus et al. [25] to maximize the efficiency of PLS-PM, a maximum of 8 predictor variables were selected for each latent variable (because 80 observations were obtained for the modeling). The OBIA predictors selected for each LV are shown in Table III. Finally, a total of 19 predictor variables were used for the modeling (8 are topographical, as proposed in [11], and 11 are OBIA based).

#### F. Richness Prediction

Species richness, defined as the number of species associated with a known area [21], has proven to be a relevant indicator to determine the biodiversity of a given area, as well as a valuable element to be used by decision makers at different levels [28] to establish spatially explicit conservation strategies for biodiversity [29].

PLS-PM uses linear approaches to predict the total forest vascular plant richness, as described in (1). The prediction formula was standardized to obtain positive weights and to eliminate the intercept terms [27] as follows:

$$R = \left\{ \sum_{k=1}^{n} \beta_{jk} \left( \sum_{j=1}^{n} (w_{ij}y_{j}) \cdot \frac{\sigma_{j} - \overline{\sigma}_{j}}{\sigma_{jk}} \right) \cdot \frac{\sigma_{jR}}{\sigma_{jk}} + \varepsilon_{jk} \right\} + \varepsilon_{jk}$$

where $R$ is the richness in the number of species, $k$ is the flux of $R$ in each predictor latent variable ($LV_i$), $i$ corresponds to every predictor variable of $LV_j$, $\sigma_{j}$ is the standard deviation of $LV_j$, $\overline{\sigma}_{j}$ is the mean value of $LV_j$, $\sigma_{jk}$ is the standard deviation of the ground data richness, $\varepsilon_{jk}$ is the mean of the ground data richness, and $\varepsilon_{jk}$ is the associated error. The error of the model was estimated using root-mean-square error (RMSE).

#### B. Model Validation

The loadings $r$ and communalities were highly significant ($< 0.9$) in all cases, demonstrating that the predictor variables were strongly represented by the latent variables ($LV$’s).

Furthermore, the redundancy of the richness was high (about 0.7), indicating that the $LV$’s had a highly significant predictive ability.

The path diagrams of the prediction models of richness, with the path coefficients ($\beta$) between the predictor variables and the $LV$’s, are shown in Fig. 4(b). The LV with the highest coefficient was topography, indicating that, at local scale, the (micro-) topography and the edaphic conditions provide more accurate predictive information for the estimation of plant richness than the structural vegetation information. It is possible that, with a smaller scale study, this relationship could change.

Middle-high canopies and the low canopies were the best forest structure segments to predict total richness, accounting for 84% and 20% of the effects of the vegetation in the total forest richness, respectively. On the contrary, the worst forest structure segments were the middle-low canopies and the high canopies, accounting for only 7% and 12% of the effects of the vegetation on the total forest richness, respectively.

The model showed a significant predictive capacity, with a GoF of 0.82, an $r^2$ of 0.64, and an RMSE of 4.13 species.

TABLE III

<table>
<thead>
<tr>
<th>Latent Variable</th>
<th>Selected variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCM</td>
<td>Middle crown-objects, the Middle cover, and the Low cover.</td>
</tr>
<tr>
<td>HC</td>
<td>Number, Cover, and Standard Deviation of the crown-objects size.</td>
</tr>
<tr>
<td>MHC</td>
<td>Number and Cover of the crown-objects.</td>
</tr>
<tr>
<td>MLC</td>
<td>Number of crown-objects.</td>
</tr>
<tr>
<td>LC</td>
<td>Number and Cover of the crown-objects.</td>
</tr>
</tbody>
</table>

### FIGURES

#### Fig. 4. Model adjustment of the data set and path diagram. (a) Scatterplot between observed and predicted richness values. (b) Path diagram. (c) and (d) $r^2$ and RMSE distribution obtained when applying the bootstrapping validation. ∗ Indicates that the path coefficient is significant with 95% confidence.
for predicting species richness in tropical forest with an $r^2$ of 0.64 was applied by [30].

The model showed a slight tendency to overestimate species richness in areas with low species diversity (under 15 species) and to produce a slight underestimation of species richness in zones of high field-measured species diversity (over 15 species). This is consistent with the bimodal distribution of the richness data set [see Fig. 2(a)], where the mean of the two modes (the valley between the two local maxima) is in the 15-species threshold value. The high number of young stands makes the local maximum of the distribution higher than the right local maxima, producing the mitigation effect of the model.

**V. CONCLUSION**

An empirical PLS-PM based on object-based LiDAR data has been developed to predict plant vascular richness in a highly complex secondary Mediterranean forest in central Chile. Results have demonstrated that the inclusion of object-based information provided accurate results. Furthermore, we found that, in our study site, the effects of the (micro-) topography and edaphic conditions are more important than the vegetation structural features at a local scale. Moreover, middle-high canopies are the most useful variables to predict the total forest richness. Finally, LiDAR data showed to be a reliable source of information for the use of object-based approaches in forest ecosystems.

**REFERENCES**


