Remote Sensing of Environment 173 (2016) 200–210

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Comparing Generalized Linear Models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile

J. Lopatin [a,b,*,1], K. Dolos [a,1], H.J. Hernández [b], M. Galleguillos [c,d], F.E. Fassnacht [a]

[a] Institute of Geography and Geoecology, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany
[b] Laboratory of Geomatics and Landscape Ecology, Faculty of Forest and Nature Conservation, University of Chile, 11315 Santa Rosa, Santiago, Chile
[c] Department of Environmental Sciences, School of Agronomic Sciences, University of Chile, 11315 Santiago, Chile
[d] Center for Climate and Resilience Research (CR)2, University of Chile, Santiago, Chile

ARTICLE INFO

Article history:
Received 6 May 2015
Received in revised form 13 November 2015
Accepted 23 November 2015
Available online xxxx

Keywords:
Species richness
LiDAR data
GLM
Random forest
Alpha-diversity
Bootstrap validation

ABSTRACT

Biodiversity is considered to be an essential element of the Earth system, driving important ecosystem services. However, the conservation of biodiversity in a quickly changing world is a challenging task which requires cost-efficient and precise monitoring systems. In the present study, the suitability of airborne discrete-return LiDAR data for the mapping of vascular plant species richness within a Sub-Mediterranean second growth native forest ecosystem was examined. The vascular plant richness of four different layers (total, tree, shrub and herb richness) was modeled using twelve LiDAR-derived variables. As species richness values are typically count data, the corresponding asymmetry and heteroscedasticity in the error distribution has to be considered. In this context, we compared the suitability of random forest (RF) and a Generalized Linear Model (GLM) with a negative binomial error distribution. Both models were coupled with a feature selection approach to identify the most relevant LiDAR predictors and keep the models parsimonious. The results of RF and GLM agreed that the three most important predictors for all four layers were altitude above sea level, standard deviation of slope and mean canopy height. This was consistent with the preconception of LiDAR's suitability for estimating species richness, which is its capacity to capture three types of information: micro-topographical, macro-topographical and canopy structural. Generalized Linear Models showed higher performances (r^2: 0.66, 0.50, 0.52, 0.50; nRMSE: 16.29%, 19.08%, 17.89%, 21.31% for total, tree, shrub and herb richness respectively) than RF (r^2: 0.55, 0.33, 0.45, 0.46; nRMSE: 18.30%, 21.90%, 18.95%, 21.00% for total, tree, shrub and herb richness, respectively). Furthermore, the results of the best GLM were more parsimonious (three predictors) and less biased than the best RF models (twelve predictors). We think that this is due to the mentioned non-symmetric error distribution of the species richness values, which RF is unable to properly capture.

From an ecological perspective, the predicted patterns agreed well with the known vegetation composition of the area. We found especially high species numbers at low elevations and along riversides. In these areas, overlapping distributions of thermopile sclerophyllos species, water demanding Valdivian evergreen species and species growing in *Nothofagus obliqua* forests occur.

The three main conclusions of the study are: 1) appropriate model selection is crucial when working with biodiversity count data; 2) the application of RF for data with non-symmetric error distributions is questionable; and 3) structural and topographic information derived from LiDAR data is useful for predicting local plant species richness.

1. Introduction

Today, biodiversity is considered to be an essential element of the Earth system from which all humans benefit directly or indirectly (Duffy, 2009). As a consequence of the dramatic impacts following human-induced changes to ecosystems worldwide, over the last few decades the current and future state of biodiversity has been receiving greater scientific and political interest. This interest is also motivated by an increased awareness of the adverse effects of reduced biodiversity on ecosystem services, on which human well-being depends (Balvanera et al., 2006; Carpenter, Bennett, & Peterson, 2006). To enable appropriate conservation and management strategies (with often limited resources), it is important to efficiently identify and monitor species rich sites (Turner et al., 2003). Theoretical and empirical studies have

* Corresponding author at: Institute of Geography and Geoecology, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany.
E-mail addresses: javier.lopatin@kit.edu (J. Lopatin), dolos@kit.edu (K. Dolos), jhernand@uchile.cl (H.J. Hernández), mgalleguillos@renare.uchile.cl (M. Galleguillos), fabian.fassnacht@kit.edu (F.E. Fassnacht).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.rse.2015.11.029
0034-4257/

suggested that local biodiversity is positively influenced by environmental heterogeneity (EH) (Stein, Gerstner, & Kreft, 2014). EH can be understood as the (co-) occurrence of a variety of environmental gradients and therefore habitat types (typically connected to high resource and structural complexity), offering a high diversity of niches over a comparably small area. A higher number of niches can in turn be colonized and inhabited by a greater number of species (e.g. Dufour, Gadallah, Wagner, Guisan, & Butler, 2006; Stein et al., 2014). In the special case of forests, topography for example can cause niche variability by separating the terrain into shaded and sunny slopes or by diversifying the local hydrology. Furthermore, vegetation structure can have a notable impact on niche diversity. For example, even aged forest stands provide fewer habitats than uneven aged multi-species forests (Gilbert & Lechowicz, 2004). As both passive (i.e. multi and hyperspectral) and active (i.e. LiDAR and Radar) sensors are able to deliver information on EH, they should also have a high potential for supporting the estimation and monitoring of species richness (Turner, 2014).

One important measure for biodiversity is the number (α-diversity) and variety of biotic species within a given geographic region (Kuenzer et al., 2014). A number of remote sensing studies in the last decade have attempted to map plant α-diversity, mostly using optical sensors (Rocchini et al., 2010). Within this context, Palmer, Earls, Hoagland, White, and Wohlgemuth (2002) formulated the spectral variation hypothesis (SVH), which states that spectral heterogeneity as measured by optical remote sensing systems relates to spatial (environmental) heterogeneity and thereby – as explained above – to species richness (Rocchini, 2007). A review on the state of the art of this research field is provided by Rocchini et al. (2010) who overview the differing aspects of remote sensing techniques that have been examined in the context of biodiversity assessment. These include the problem of scale (pixel size versus field sampling units), methods to measure spectral heterogeneity (crisp classification versus fuzzy methods or the direct application of non-classified reflectance values), as well as the question of how the derived spectral heterogeneity is connected to biodiversity. However, this last question requires a definition of biodiversity, which can be defined either taxonomically, functionally or genetically. According to Rocchini et al. (2010), most remote-sensing studies focus on taxonomic diversity. Finally, the success for estimating species richness from remote sensing data is also influenced by the structure of the field data (abundance data versus presence/absence) and the applied modeling techniques.

While the application of passive optical remote sensing sensors for estimating biodiversity has significantly advanced over the last two decades, the number of studies investigating the potential of active optical sensors such as Light Detection and Ranging (LiDAR) is still sparse. LiDAR has proven to be one of the most powerful data acquisition systems for obtaining topographical and vegetation-structural information (French, 2003; Lefsky, Cohen, Parker, & Harding, 2002). Both of these types of information were found in earlier studies to be able to estimate EH (Bergen et al., 2009; Dauber et al., 2003; Gaston, 2000). According to Bergen et al. (2009), this makes LiDAR information a good proxy for species richness, especially in forests with high vertical complexity. One focus of earlier studies was the application of LiDAR-derived forest structural and topographical information to predict forest fauna richness (e.g. Clawges, Vierling, Vierling, & Rowell, 2008; Goetz, Steinberg, Dubayah, & Blair, 2007; Vierling et al., 2011). A smaller number of studies also focused on forest flora richness with successful results (e.g. Hernández-Stefanoni et al., 2014; Lopatin, Galleguillos, Fassnacht, Ceballos, & Hernández, 2015; van Ewijk, Randin, Treitz, & Scott, 2014), confirming the suitability of LiDAR data for estimating plant species richness.

Generally, LiDAR data relate to three types of information which interact with plant species richness: micro-topographical, macro-topographical and canopy structural information. Macro-topography has been shown to be highly correlated with plant species distributions. Important factors include altitude above sea level, aspect and slope which relate to climate (e.g. irradiation, temperature, precipitation) and geomorphology (e.g. erosion intensity). These factors influence species composition by, for example, limiting the available light (limited irradiation on shaded slopes) or temperature (high altitudes) which may keep certain species from growing. Steep slopes may result in increased erosion risk, leading to areas with mechanical disturbances and poor soils which may only be suitable for stress-tolerating species.

Micro-topography (i.e. local slope or surface roughness conditions) as measured by LiDAR systems presumably acts as a proxy of small-scale habitat structures such as shaded humid sinks or areas with deeper soils (Moeslund et al., 2013; Silvertown, Dodd, Gowing, & Mountford, 1999). Depending on the number of LiDAR returns and the penetration ability of the applied scanning system, micro-topographic features might also be directly related to the presence of a dense herb or shrub layer, which cannot be penetrated by the LiDAR signal and therefore leading to increased surface roughness in the derived digital terrain model. The penetration capability of the LiDAR sensor is a general limitation which hampers the collection of information on micro-topographic conditions. For example, in the presence of a very dense overstory, only a limited number of returns may come from the ground.

Finally, canopy characteristics such as differences in canopy height, leaf size and leaf orientation, lead to different canopy closure percentages or leaf area index values (Morsdorf, Kötz, Meier, Itten, & Allgöwer, 2006; Popescu, Wynne, & Nelson, 2003; Pope & Treitz, 2013; Woods, Lim, & Treitz, 2008). According to Lemenih, Gidyelew, and Teketay (2004) this influences the light conditions on the ground which in turn affects the species composition and richness. Thus, LiDAR information should be able to both provide a good description of the (upper) canopy structure as well as deliver valuable information concerning the understory conditions (Eskelson, Madsen, Hagar, & Temesgen, 2011; Su & Bork, 2007; Wing et al., 2012), which has been confirmed by a few earlier studies (e.g. Leutner et al., 2012; Wolf et al., 2012). Therefore, considering both the theoretical suitability of LiDAR data as well as the promising results of past studies, we think that it is valuable to further examine and refine the application of LiDAR data for estimating phyto-diversity.

One potential field for refinements is in the model building process. According to Rocchini et al. (2010), earlier studies focusing on the estimation of biodiversity from remote sensing data often followed simple univariate regression approaches (Oldeland, Wesuls, Rocchini, Schmidt, & Jurgens, 2010; Palmer et al., 2002; Rocchini, Chiarucci, & Loiselle, 2004) while others integrated weighting procedures into the univariate model set-up (Foody, 2005; Nagendra, Rocchini, Ghate, Sharma, & Pareeth, 2010). Furthermore, there are a few recent examples of advanced modeling techniques from the field of statistics, such as partial least square (PLS)-based models (Feilhauer & Schmidtlein, 2009) or Generalized Additive Models (GAM) (Fava et al., 2010), as well as from the field of machine learning, such as neural networks (Foody & Cutler, 2003). Studies following such approaches often used feature extraction approaches to address multi-collinearity originating from the multi- or hyperspectral bands (Fava et al., 2010; Higgins et al., 2014; Rocchini, 2007). Other studies applied feature selection approaches to reduce the feature space (Camathias, Bergamini, Küchler, Stofer, & Baltensweiler, 2013; Hernández-Stefanoni et al., 2014). Some earlier studies (e.g. Foody & Cutler, 2006) have claimed that simple methodological approaches such as the application of vegetation indices and standard regression techniques are not able to fully use the information content of remotely sensed data. In spite of this drawback, parametric statistical models are still useful because they provide an opportunity to account for the distribution of the response variable and the model residuals (Nelder & Wedderburn, 1972). As species richness is measured as count data (i.e. number of species), which are discrete and limited to non-negative values (Zeileis, Kleiber & Jackman, 2008), typical appropriate statistical families for the error distribution are the Poisson or negative binomial distribution. Applying techniques which assume symmetry or homoscedasticity – or even a Gaussian distribution of

the residuals – will often lead to a sub-optimal model fit in terms of precision and bias, which in the worst case can lead to a misinterpretation of the results (Hayes & Cai, 2007; Manning & Mullahy, 2001).

To resolve these issues, there are two tendencies in remote sensing. The first way is an evasion of these issues using the field of machine learning (e.g., Foody & Cutler, 2003; Foody & Cutler, 2006; Leutner et al., 2012). As many machine learning methods are described as being non-parametric, it is frequently assumed that there are no requirements concerning the error distribution. However, this is not true for regression trees and random forest methods, which either fit standard linear (Gaussian) regressions for tree nodes or are based on measures for node impurity, such as the sum of squared deviations to the mean (Loh, 2011), and thereby do not account for asymmetry and heteroscedasticity (Chaudhuri, Lo, Loh & Yang, 1995; Ciampi, 1991). The second way is to apply transformations of the dependent variable (e.g., Camathias et al., 2013; Hernández-Stefanoni et al., 2014). However, a well-known problem with data transformations is the trade-off between homoscedasticity and linearity (O'Hara & Kotze, 2010). The family of transformations used may not be able to correct one or both of these problems. An additional problem with the regression of transformed variables is that it can lead to impossible predictions, such as negative species numbers due to back-transformation of the response. Motivated by these challenges, Generalized Linear Models (GLMs) were developed (Nelder & Wedderburn, 1972). Among other options, GLMs allow for the specification of an error distribution and link function appropriate for count data such as species richness (e.g., Poisson or negative binomial). The option to choose an appropriate model family for the particular modeling task is an additional advantage of GLMs and similar approaches (e.g. GAMs, GLMMs) over standard machine learning methods, ordinary least squares and some PLS-based models.

In summary, the application of active LiDAR for predicting biodiversity, and plant diversity in particular, is still under-examined, although the suitability of the data has been demonstrated. Furthermore, random forest – one of the most frequently applied machine learning methods in remote sensing – is suspected to ignore the nature of count data. Therefore, in this study, we applied discrete return LiDAR data to model vascular plant species richness (α-diversity) in a highly complex second growth forest in Central Chile. We compared random forest with a GLM which we optimized for the response variables by assuming a negative binomial data distribution.

2. Material and methods

2.1. Study area

The study area, Monte Oscuro, is located in central Chile in the Maule region (35°07′00″ S, 70°55′30″ W) (Fig. 1 A). This area is associated with the Sub-Mediterranean Temperate bioclimatic zone. The total annual precipitation (1000 mm) is mainly concentrated between April and October, and monthly mean temperatures range from 8 °C in the coldest months (June to August) to 18 °C in the warmest months (December to February). The site covers an extent of 1295 ha, and features a mean altitude of 1075 m above sea level with most slopes facing south. Monte Oscuro is located in a transitional vegetation zone where species of the Mediterranean sclerophyll forest (e.g. *Quillaja saponaria* Mol., *Cryptocarya alba* (Molina) Looser, *Lithraea caustica* (Molina) Hook et Arn.) coexist with the Valdivian forest evergreen species in a matrix dominated by *Nothofagus obliqua* (Mirb.) Oerst. (Gajardo, 1994).

2.2. Ground data

The field survey focused on vascular plants because of the clear dominance of this taxon in Chilean Sub-Mediterranean forests. Furthermore, earlier studies showed the particular importance of vascular plants for the trophic network and other ecological processes (Palmer et al., 2002).

A vegetation assessment was conducted between January 2013 and January 2014. The 80 surveyed square nested plots were allocated in a 200 × 200 m regular grid (Fig. 1 B). To avoid the effect of borders and bare soil in the plots and also to facilitate operational effort, steep

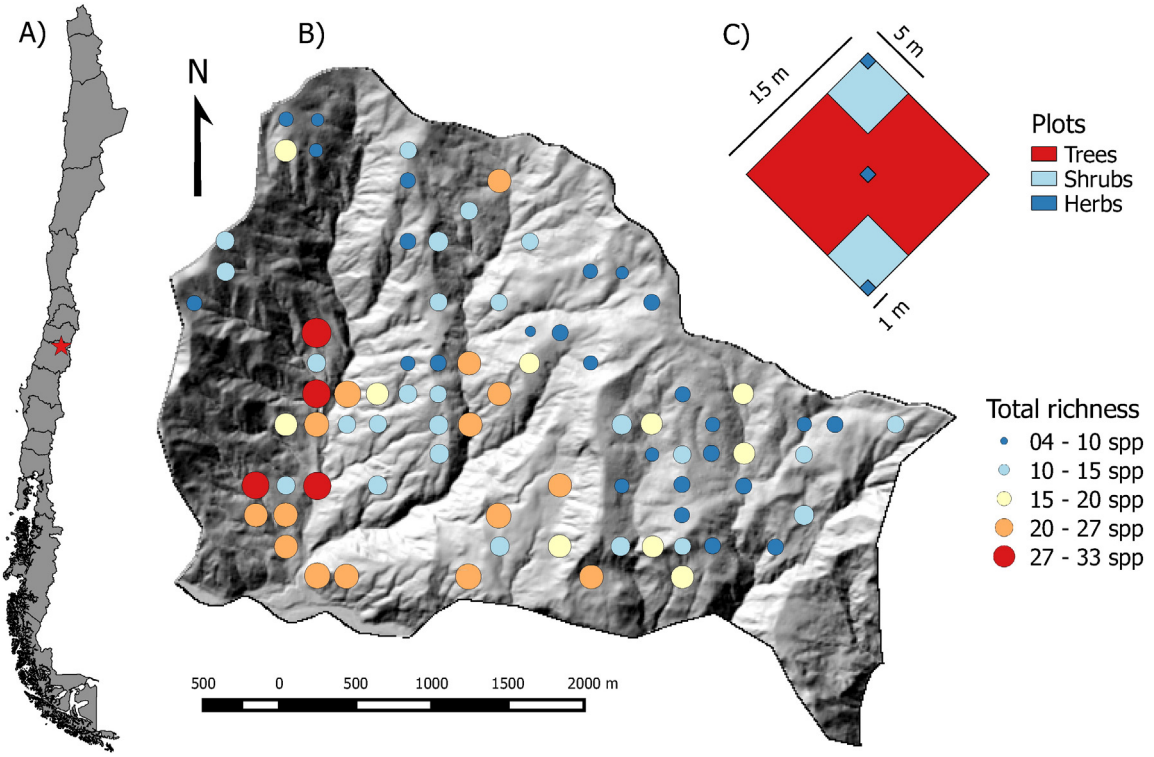

Fig. 1. A) Location of the study area; B) the distribution of field measured richness values are illustrated, and C) detail of a sampling unit with nested plots.

zones (>45%) and sites which were too close to trails (20 m) were excluded from the survey.

Species were sampled in three height layers: trees (height of 2 m or more), shrubs (less than 2 m of height) and herbs (non-woody plants). Each of the 80 plots was composed of six nested subplots. In the largest plot with an area of 225 m^2 (15 m × 15 m) only trees were registered. Shrubs were sampled in two subplots of 25 m^2 (5 m × 5 m) while herbs were registered in three sub-subplots of 1 m^2 (see Fig. 1 C).

Finally, the total richness of vascular plant species within each of the 80 plots was calculated by summing up all species in all height layers, and as encountered in the plots and subplots. Species occurring in several height layers (e.g. tree and shrub layers) were only counted once.

2.3. LiDAR data

A discrete return LiDAR system (Harrier 54/G4 Dual System, manufactured by Trimble Industries and provided by Digimapas Chile Ltda., Santiago, Chile) was applied to gather airborne LiDAR data over the study site in March, 2011. The Harrier 54/G4 Dual System features a 1550 nm laser with a scanning frequency of 100 Hz, a pulse rate of 100 kHz, a scan angle (FOV) of ±22.5°, and a laser beam divergence (IFOV) of 0.5 mrad. The obtained point cloud had an average point density of 4.64 points per m^2 and a footprint diameter of 29 cm.

The LiDAR point cloud was classified into ground and non-ground points for bare-earth extraction according to Briese (2010). A digital terrain model (DTM) and a digital canopy model (DCM) of 1 m pixel size were calculated by interpolating the classified point cloud. The optimal pixel size was empirically selected based on further experiments that can be found in the Supplementary data. Consequently, each of the sampling plots contains ~225 pixels of DTM and DCM derived variables.

2.4. Predictor variables

At the location of each sample plot the mean values (and in two cases the standard deviation) of several topographical and vegetation-related predictor variables were derived from the LiDAR DTM and DCM and used in the models for species richness prediction (Table 1). The choice of the topographical variables was influenced by the results of earlier studies also focusing on the estimation of plant species richness (Bässler et al., 2010; Camathias et al., 2013; Ceballos, Hernández, Corvalán, & Galleguillos, 2015). We selected topographical variables carrying information on macro-topography (e.g. altitude above sea level (DTM), normalized heights) and micro-topography (e.g. standard deviation of the slope (sd slope)). The canopy related variables were derived from the DCM. Canopy structure, particularly its density, influences light conditions in all underlying forest layers (Lemenih et al., 2004). In the present case, we used textural information (co-occurrence textural matrix), the height (mean canopy height) and the homogeneity of the canopy (sd canopy height) as proxies for the canopy structure and its density. We assume that comparably smooth canopy surfaces indicate a closed canopy while lighter canopies show higher canopy roughness due to gaps. Both the co-occurrence textural matrix values as well as the variable sd canopy height should serve as descriptors of this canopy roughness. Furthermore, the mean canopy height serves as a proxy for the successional or developmental stage of forests. Differing forest succession and development stages have also been found to be related to special understory conditions in earlier studies, as further discussed below (Emborg, 1998; Guariguata & Ostertag, 2001).

2.5. Statistical models

Species richness was modeled using two approaches: Random forest (RF) and a Generalized Linear Model (GLM).

The ensemble regression tree method RF (Breiman, 2001) has been reported to be an efficient prediction approach, especially when – as in the present case – the numbers of observations are comparably low compared to the number of predictors (Svetnik et al., 2003). We applied the RF routine in R (package random Forest, Liaw & Wiener, 2002). Technical details on the applied RF algorithm can be found in Latifi, Fassnacht, and Koch (2012) and Ghosh and Joshi (2014). RF requires two parameters to be set: 1) mtry, the number of predictor variables performing the data partitioning at each node and 2) ntree, the total number of trees to be grown in the model run. Based on earlier experiences and recommendations from literature we set the number of ntree to 500, whereas mtry was fixed to 7 after some initial tuning experiments. The importance of predictor variables was measured by the Gini decrease in node impurity measure, which is computed by permuting the predictor variables with the out-of-bag data in the RF validation approach (details in Liaw & Wiener, 2002).

Processing of the data with GLMs can be subdivided into three steps:

(1) The first step was to identify an appropriate model family able to deal with the statistical properties of count data. We therefore compared the normalized quantile-plots of the residuals of several GLMs, calculated with model families which are generally recommended to be used with count data. We ran GLMs with log link functions and Poisson, Quasi-Poisson as well as negative binomial distributed residuals for each response variable. In addition, models assuming a Gaussian error distribution (which theoretically are not suitable for count data) with and without log link functions were calculated. In this first step the models were calculated based on a single predictor variable (mean canopy height) which was selected in a preceding analysis. Judging from the normalized quantile-plots of the residuals, the negative binomial error distribution with a log-link function appeared to perform slightly better than Poisson and Quasi-Poisson regression. As expected, the models assuming a Gaussian error distribution (with and without log-link) resulted in non-acceptable distributions of the residuals.

Table 1
Predictor variables used in both model approaches (RF and GLM) to estimate vascular plant species richness. Predictors were derived from the LiDAR DTM and DCM.

Variable	Description	Type		
Aspect	Mean of the aspect pixel values (°).	DTM	Macro-	Topography
DTM	Mean value of the altitude above sea level pixel values (m).			
sd Slope	Standard deviation of the slope pixel values (%).		Micro-	
TWI	Mean of the topographic wetness index pixel values (Beven & Kirkby, 1979). This index describes humidity patterns based on micro-topography.			
Normalized heights	Mean of the normalized height index (Böhner, Böhner, Blaschke, & Montanarella, 2008). This index describes the altitude difference of the altitude in a given pixel and the bottom of the next corresponding valley. It is therefore a measure for the distance from a riverside.			
Co-occurrence textural matrix	Mean pixel values of the co-occurrence textural matrix. These include the homogeneity, entropy, contrast, dissimilarity and second-moment of the canopy height model.	DCM		Canopy-structure
Mean canopy height	Mean of the canopy height model pixel values.			
Sd canopy height	Standard deviation of the canopy height model pixel values.			

(2) As GLMs cannot cope with multi-collinearity among independent variables, the search for an appropriate model family was followed by a variable selection procedure to find the most important and at the same time uncorrelated predictors. Scatterplots between the individual predictors and the response variables showed in most cases linear relations between the variables. This allowed us to use hierarchical partitioning in which we selected a negative binomial error distribution (compare step 1) and log link function (R-package "hier.part", Chevan & Sutherland, 1991) to measure the variable importance. This variable importance, calculated as percentage of total explained variance, was also compared with the rank of variable importance given by the Gini impurity index of RF.
(3) The resulting variable importances were used to select the predictor variables for the final models. The final number of predictor variables was determined by calculating several models and adding a single variable in each run (starting with the most important variable) and comparing the akaike information criterion (AIC) and deviance explained for the results. The best results were obtained using three predictors.

For validation purposes, the best GLM as well as the RF model were embedded in a bootstrap with 500 iterations. In each bootstrap iteration, we drew 80 times with replacement from the 80 available samples. In this procedure, on average 36.8% of the total number of samples (~28 samples) are not drawn. These samples were subsequently used as holdout samples for an independent validation (Fassnacht et al., 2014b). The model performances of RF and GLM were compared based on differences in the coefficients of determination (r^2 — calculated as the squared Pearson's correlation coefficient) and the normalized root mean square error (nRMSE) between predicted and observed richness values of the hold-out samples in the bootstrap. To enable sound comparisons between the four response variables, the normalized RMSE (nRMSE) calculated as [RMSE/[max(number of species) − min(number of species)]] × 100 was used, where RMSE is calculated as $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j-\hat{y}_j)^2}$ with y = reference species richness value, $\hat{y}$ = estimated species richness value and n = number of samples. High values of r^2 and low values of nRMSE indicate high model quality. The bias of prediction was measured as one minus the slope of a regression without intercept of the predicted versus observed values.

To test if model quality was statistically better for GLM than for RF, a one-sided bootstrap test was performed using the function "boot" available in the R package "boot" (Canty & Ripley, 2014). Test variables were calculated by subtracting the nRMSE and bias values of GLM from the ones produced by RF, and the other way around for r^2. For these distributions a one sided test was performed to test if the differences between GLM and RF values were larger than zero based on 500 bootstrap samples. The level for significance was set to $\alpha = 0.05$.

2.6. Predictive species maps

Predictive maps of species richness were calculated for each layer based on the best obtained models. A convex hull mask was then applied to exclude all pixels outside of the value range of data used for fitting the models. This was necessary because predictions of statistical models are highly uncertain outside the data range used for fitting the model (e.g., there are no sampling plots in bare soil areas, so the model predictions for those areas are extrapolations and highly uncertain). The "alphahull" R-package was used (Pateiro-Lopez & Rodriguez-Casal, 2013) to calculate the mask. In the maps, white areas indicate areas out of the value range, which were excluded by applying the convex hull mask. Additionally, a map of the coefficient of variation (CV, given in %) values for the species richness predictions (as obtained from the 500 bootstrap runs) was produced for the entire area.

3. Results

3.1. Variable importance

Variable importance was determined for both modeling approaches (hierarchical partitioning and Gini impurity index for GLM and RF, respectively) and for each forest layer. The results of both approaches agreed that the three most important predictor variables for modeling species richness for all forest layers are mean canopy height, mean altitude above sea level (DTM) and sd slope. In almost all cases, mean canopy height was selected as the best predictor, except for the RF model for the tree layer where the DTM was found to be of highest importance (Fig. 2).

3.2. Model performances

For GLM the three most important variables were considered (mean canopy height, DTM and sd slope) in the model building process while for the RF models all variables were used. The model performance results were summarized in terms of r^2, nRMSE and bias for all 500 bootstrap values (see Fig. 3). GLMs showed systematically higher r^2, lower nRMSE values and less bias than RF (Table 2). The only exception was the herb layer which showed marginally higher errors for the GLM compared to RF. The best model fit was found when predicting total richness with GLM (median bootstrap r^2 of 0.65 and nRMSE of 16.60%), while the worst fit was observed for tree richness with RF (median bootstrap r^2 of 0.32 and nRMSE of 20.01%). Furthermore, both models showed a systematic tendency to overestimate small values and underestimate high values (Fig. 4). This effect was stronger in the RF models.

The bootstrap test for differences in the model quality measures was significant for six out of 12 tests. Regarding r^2 GLM was significantly better than RF for all layers but the shrub layer. For nRMSE GLM results were significantly better for total richness and tree richness estimates while the results for the herb and shrub layers were inconclusive. Differences in bias were only significant for the shrub layer (Table 3).

3.3. Prediction maps

The prediction map of vascular plant richness showed a general tendency towards high richness values in low altitude areas and near riversides (Fig. 5). This tendency was found in all height layers. In addition, increased herb richness close to bare soil areas (masked areas) was apparent. The coefficient of variation (CV) map obtained from the 500 bootstrap predictions of the GLMs (Fig. 6) showed that total, tree and shrub richness values were predicted with relatively low variation (0–5%), while the predictions for the herb layer showed higher variation (5–15%). These results agreed with the nRMSE errors which were also highest for the herb layer. Furthermore, there seemed to be a general tendency towards higher variation (10–20%) in the predictions along riversides.

4. Discussion

In the present study, the suitability of airborne discrete-return LiDAR data for the mapping of vascular plant species richness within a Sub-Mediterranean second growth forest ecosystem was examined. The findings of the study are discussed in three sections. First, we discuss the ecological meaning of the variables identified as the most important predictors (mean canopy height, DTM and sd slope). Second, we debate the spatial patterns of species richness distribution for all layers. Third, the results of the model comparison (GLM, RF) are discussed.

4.1. Selected predictor variables and their ecological implications

RF and GLM both identified mean canopy height as the most important variable for modeling the species richness of vascular plants in the

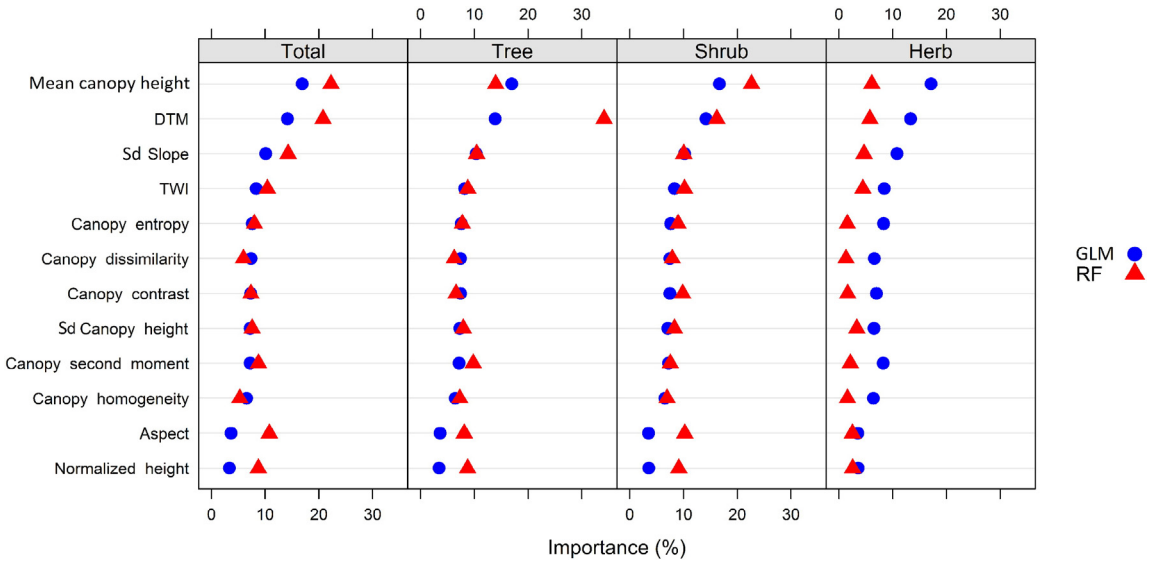

Fig. 2. Variable importance using the Gini impurity index (RF) and hierarchical partitioning with negative binomial model family and log-link function (GLM). The mean values of the 500 bootstraps are shown.

study region. This is in agreement with earlier studies focusing on species richness estimations from LiDAR data (Simonson, Allen, & Coomes, 2012; Wolf et al., 2012). In former studies, mean canopy height was proven to be a reliable proxy for forest properties such as above-ground biomass (Fassnacht et al., 2014a; Kattenborn et al., 2015; Latifi et al., 2012), or forest successional stages (Falkowski, Evans, Martinuzzi, Gessler, & Hudak, 2009). In differing successional or development stages of forests, the environmental conditions (e.g., light and water availability or micro-climatic conditions) in the stands differ and thereby influence the composition of the species' communities in all height layers (Emborg, 1998; Guariguata & Ostertag, 2001). Due to the use of the forests in the Monte Oscuro area by selective logging for construction wood and charcoal in the 1950s, a mosaic of several successional stages can be found in the study area. The harvested forest stands were abandoned and were subject to natural succession, while some other stands in less accessible areas were not intervened and were able to develop towards older development stages. It is likely that the importance of the mean canopy height relates to these processes.

Aside from the mean canopy height, the two topographic variables altitude above sea level (DTM) and sd slope (describing terrain heterogeneity on a micro-topographic scale) were identified as the second and third ranked variables in the feature selection. This again agrees well with earlier studies (e.g., Bässler et al., 2010; Bergen et al., 2009; Camathias et al., 2013; Ceballos et al., 2015; Rocchini et al., 2010).

In the present study, species richness was higher at lower altitudes. It is known that the species richness of the Chilean mountains generally declines with altitude (Gajardo, 1994). Within alpine communities, Henrik et al. (2006) described the relationship between altitude and species richness as a trade-off between a declining species pool (due to increasingly unfavorable environmental conditions) and the decreasing intensity of competition with altitude (due to reduced number of species). Therefore, in many cases unimodal relations between altitude and species richness could be observed. In the present case, the observed pattern of declining species numbers with increasing altitude must also be discussed in the context of the land use history of the study area. As mentioned before, significant interventions occurred in the 1950s for some forest stands in the study area. The extent and intensity of logging was stronger in easily accessible areas, which were usually located at low elevations. Therefore, the higher species richness at lower elevations, particularly in the tree layer, could be explained by the coexistence of species typically occurring in early (e.g., light demanding pioneer species such as *L. caustica*) and late (e.g., *N. obliqua*) successional stages of the forests.

The third ranked variable sd slope describes terrain heterogeneity and was found to be positively correlated with species richness in

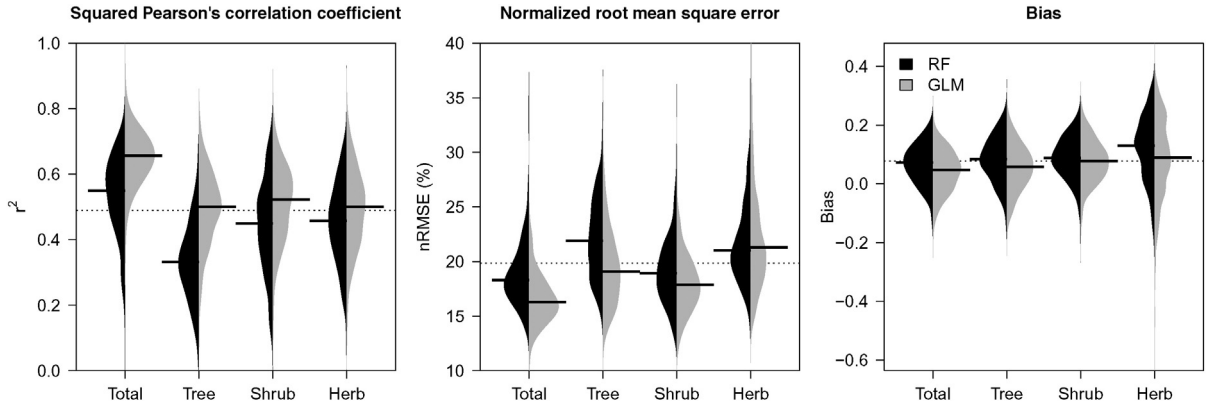

Fig. 3. Model accuracies of random forest (RF, black plots) and Generalized Linear Models (GLM, gray plots) in terms of r^2 (left panel), nRMSE (central panel) and bias (right panel). Beanplots display distribution of results for the 500 bootstrap runs. Black horizontal lines indicate the median values of the distribution.

Table 2
Model accuracies per forest layer. The median bootstrap values are displayed.

Forest layer	r^2 — GLM	r^2 — RF	nRMSE (%) — GLM	nRMSE (%) — RF	Bias — GLM	Bias — RF
Total	0.66	0.55	16.29	18.30	0.05	0.07
Tree	0.50	0.33	19.08	21.90	0.06	0.08
Shrub	0.52	0.45	17.89	18.95	0.08	0.09
Herb	0.50	0.46	21.31	21.00	0.09	0.13

Table 3
Results for bootstrap test to check for statistical differences in model quality measures r^2, nRMSE and bias as obtained by GLM and RF for tree, shrub and herb layer. H0: model quality measures do not differ, Ha: model quality is better for GLM than RF. Number of bootstrap samples = 500; $\alpha = 0.05$.

Layer/quality measure	r^2	nRMSE	Bias
Total	*	*	–
Tree	*	*	–
Shrub	–	–	*
Herb	*	–	–

* Indicates a significant test result.

Monte Oscuro. At the local scale, high terrain heterogeneity causes differences in soil conditions such as water availability (Silvertown et al., 1999; Moeslund et al., 2013) and vertical nutrient distribution (Everson & Boucher, 1998). This increases habitat heterogeneity and can foster increased species richness. Generally, micro-topography is difficult to consider in biodiversity studies due to the lack of reliable spatially continuous data over larger areas. In the present study the application of a LiDAR terrain model with 1 m pixel size allowed for the consideration of variables describing micro-topography. However, due to the functioning of LiDAR systems, uncertainty remains regarding the share of the signal that actually describes micro-topography, in contrast to the intermingling effect of dense understory vegetation. In the present case, where LiDAR data was collected over relatively dense canopies, the limited penetration ability of the LiDAR beams creates additional uncertainty concerning the information content of the variable sd slope. If the LiDAR beams predominantly penetrated to the ground, the sd slope variable would contain mainly information on micro-topography. If the beams did not reach the ground due to dense understory vegetation, the information would instead describe the surface roughness induced by understory vegetation. Since the presence of dense understory vegetation presumably also influences species richness, further examination would be necessary to determine the valuable information contained in the sd slope variable.

4.2. Spatial patterns of richness distribution

The prediction maps of species richness showed higher species richness for lower altitudes and areas close to riversides. In addition to the already discussed effects of the land-use history, the observed patterns agree well with the known natural vegetation composition of the study area. Monte Oscuro is located in the transition zone between Valdivian evergreen forests and Mediterranean sclerophyll forests in a matrix of *N. obliqua* (Ceballos et al., 2015; Luebert & Pliscoff, 1999). Within this transition zone, Sclerophyll species are more abundant in the lower altitudes due to favorable temperature conditions, while the Valdivian evergreen species are often located near the riversides due to their water needs (Corvalán, Galleguillos, & Hernández, 2014). Finally, *N. obliqua* is a competitive species typically dominant in the

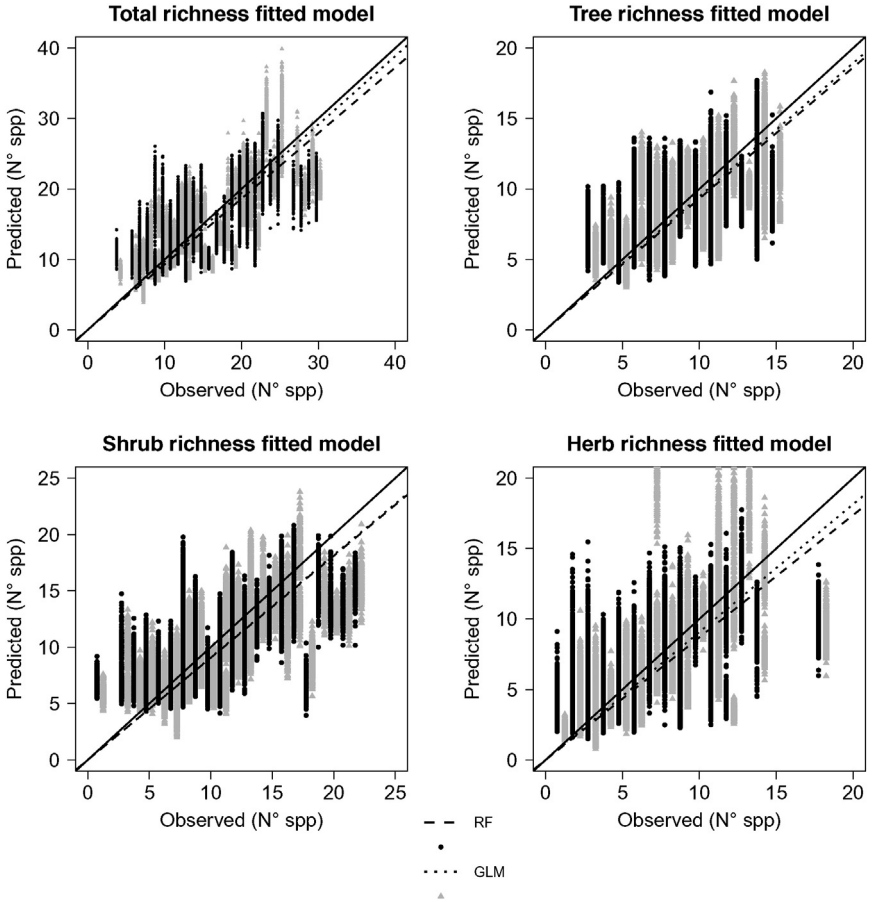

Fig. 4. Scatter plots of observed versus predicted values of total, tree, shrub and herb richness. Results from all 500 boots are displayed.

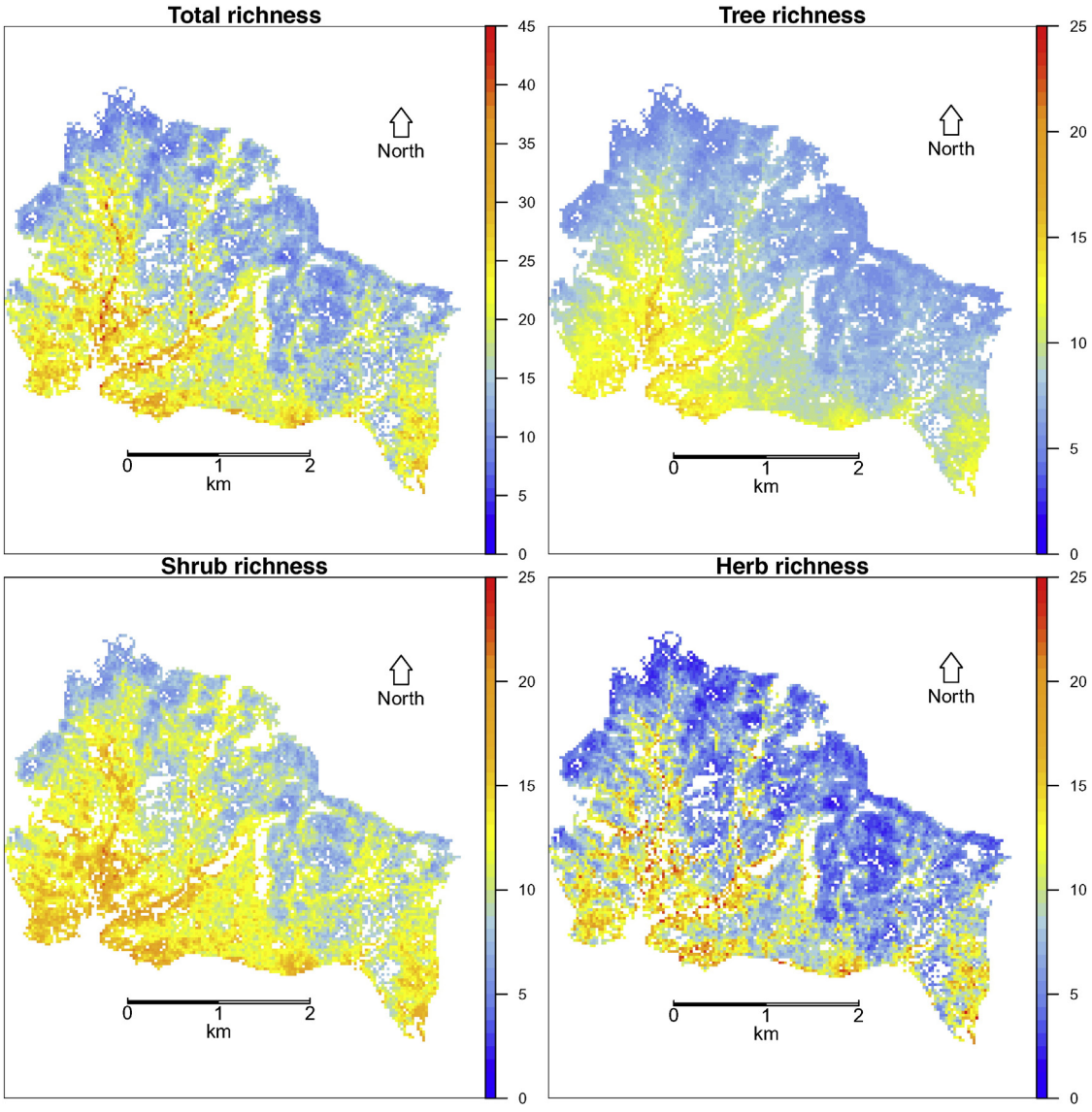

Fig. 5. Maps of predicted richness per layer using the GLM models. White areas indicate areas out of the value range which were excluded by applying the convex hull mask.

area (Gajardo, 1994) and occurs with associated species over the largest fraction of the environmental gradients in the study area. At low elevations and close to riversides, it is therefore expected that species from each forest type add to an overall increased species number.

For trees and herbs this observed tendency was stronger compared to the shrubs, which also reached higher richness values at higher altitudes. High tree richness values correlated strongly with low mean canopy height values of about 7 m (Fig. S3), which supports the earlier formulated hypothesizes that the land-use history of the site and the corresponding successional stages have had a strong influence on the observed species richness patterns.

As previously mentioned, high shrub richness values were distributed over a larger part of the study area. This observation could be related to the definition of the shrub layer, which was solely based on a height threshold. We assume that at intermediate elevation levels, several shrub and tree species which are unable to grow at the higher elevations due to unfavorable water and temperature situations are able to establish themselves and grow up to a certain height. However, they might have problems reaching the tree layer (as defined by height threshold in this study) due to the increasingly unfavorable conditions with increasing altitude.

4.3. Model comparisons

Here, the best GLM was superior to the best RF model. The observed superiority of the GLM is presumably connected to the special statistical properties of species richness count data, which typically shows non-symmetrical error distributions. In the field of machine learning, algorithms such as RF are often referred to as being non-parametric. In many cases this is misinterpreted as “no requirements concerning data and error distributions exist”. RF is indeed less sensitive than GLM when unsuitable error distributions are used. However, standard implementations of RF are not designed to deal with non-symmetrical error distributions, unlike GLM. In parametric statistics, GLMs allow for the definition of non-Gaussian distributed model residuals such as Poisson and negative binomial error distributions.

Earlier studies using LiDAR data to estimate plant richness have been conducted in various ecosystems including Mediterranean forest (Simonson et al., 2012), tropical forest (Hernández-Stefanoni et al., 2014; Higgins et al., 2014; Wolf et al., 2012), coniferous forest (van Ewijk et al., 2014; Vogeler et al., 2014) and continental forest (Camathias et al., 2013; Leutner et al., 2012). It can be observed that in many of these studies machine learning algorithms are preferred over

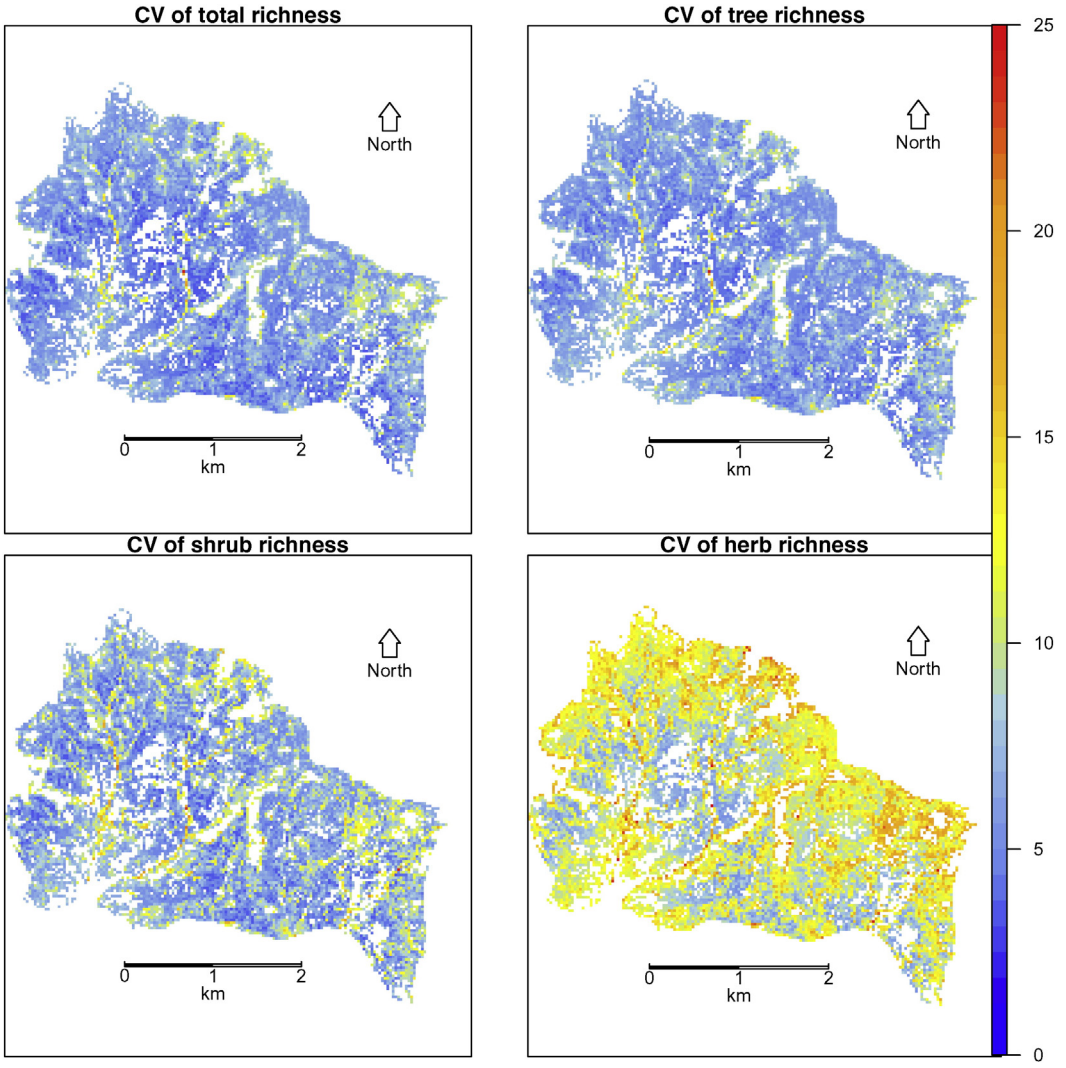

Fig. 6. Coefficient of variation (CV) maps in percentage (%) obtained from the 500 bootstrapped GLM model runs. White areas indicate areas out of the value range which were excluded by applying the convex hull mask.

parametric models. In some studies GLMs have been applied without further specification of the chosen model family; thus, it remains unclear if the standard linear regression with a Gaussian error distribution and identity-link function or a more suitable error distribution was used in those studies. Furthermore, bias was only rarely considered as a parameter for judging the quality of the models in earlier studies. We strongly recommend that this should be done, as both r^2 and nRMSE can suffer from notable offset errors when bias is not considered (Bennett et al., 2013).

In the present case, it became evident that the best GLM using only three predictors showed in all but one cases higher performances (i.e. higher r^2 and lower nRMSE) and less bias than the best RF models which used all 12 predictors. Differences among the models were most pronounced for the tree layer and the total richness values for which the bootstrap test found significantly better results for GLMs in terms of r^2 and nRMSE values. Although test results for the herb (significant only for r^2) and shrub layer (significant only for bias) were less clear, we still think a clear tendency in favor of GLMs is apparent.

GLMs and RFs each have different advantages (e.g. the option to choose the residual distribution family and the adaptability to evolve with data, for GLM and RF respectively), and the link between these two approaches presents an interesting development opportunity for modeling count data. The first methods heading towards such linked algorithms are Generalized regression trees (Ciampi, 1991), or Poisson regression trees which can nowadays be fitted in R (R Core Team, 2014) with the package "rpart" (Therneau, Atkinson, & Ripley, 2015). Also, methods to deal with more specific issues of modeling count data such as over-dispersion have already been developed (Choi, Ahnb, & Chen, 2005; Mathlouthi, Fredette, & Larocque, 2015). With the current dataset, however, the Poisson recursive partitioning regression trees and the Poisson boosted regression trees performed worse than RF (results not presented here). Nevertheless, we still think that an efficient integration of generalized regression trees into the framework of random forests has a high potential for improving our ability to model ecological count data with remote sensing data.

An extension of the conducted comparison between RF and GLMs or any of the abovementioned methods for comparable data sets would be desirable to finally assess their potentials and weaknesses. One approach could also be to conduct these comparisons based on artificial datasets of species richness. A similar approach has been presented earlier for species distribution models (Meynard & Quinn, 2007).

5. Conclusion

We applied LiDAR derived variables to estimate vascular plant species richness in a Mediterranean forest ecosystem in central Chile.

A model comparison between GLMs and RF showed that RF seem to be unable to fully exploit the potential of statistics to model species

richness count data from remote sensing data. GLMs, which are able to account for asymmetric error distributions, were found to deliver better results in terms of precision and bias in the present study. Therefore, the application of RF to model count data such as species richness may be questionable.

Altitude above sea level, terrain heterogeneity expressed as the standard deviation of the slope and mean canopy height were found to be the most important predictor variables during the variable selection. These variables agree well with our original hypothesis on the three information types (vegetation structure, macro-topography, micro-topography) contained in LiDAR data.

The findings of the study support the claimed potential of LiDAR data for mapping environmental information related to plant species richness. In combination with the choice of an appropriate statistical model accounting for the special statistical properties of count data, LiDAR data may therefore enhance the capabilities for mapping species richness as one major aspect of biodiversity and foster the integration of remote sensing data into monitoring for nature conservation.

Acknowledgments

This work was partially funded by CONICYT project, Integration of Advanced Human Capital into the Academy, code 791100013 and by the U-INICIA VID 2012, code 1/0612, University of Chile. The authors would furthermore like to thank two anonymous reviewers for their valuable comments that helped to improve an earlier version of the manuscript. Kyle Pipkins is acknowledged for proof-reading the manuscript. Finally, we would like to thank Dr. Florian Hartig for his advice concerning the selection of the statistical test.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2015.11.029.

The R-codes used in this study can be accessed at: https://github.com/JavierLopatin/SpeciesRichness-GLMvsRF-LiDAR.

References

Balvanera, P., Pfisterer, A., Buchmann, N., He, J., Nakashizuka, T., Raffaelli, D., & Schmid, B. (2006). Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology Letters*, *9*, 1146–1156.

Bässler, C., Stadler, J., Müller, J., Förster, B., Göttlein, A., & Brandl, R. (2010). LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. *Biodiversity and Conservation*, *20*, 465–481.

Bennett, N., Croke, B., Guariso, G., Guillaume, J., Hamilton, S., Jakeman, A., ... Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1–20.

Bergen, K., Goetz, S., Dubayah, R., Henebry, G., Hunsaker, C., Imhoff, M., ... Radeloff, V. (2009). Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for LiDAR and radar spaceborne missions. *Journal of Geophysical Research*, *114*, 1–13.

Beven, K.J., & Kirkby, M. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, *24*, 3–69.

Böhner, J., Böhner, J., Blaschke, T., & Montanarella, L. (2008). *SAGA — Seconds out*. Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie (113 pp.).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Briese, C. (2010). Extraction of digital terrain models. In G. Vosselman, & H. Maas (Eds.), *Airborn and terrestrial laser scanning*. Taylor & Francis (318 pp.).

Camathias, L., Bergamini, A., Küchler, M., Stofer, S., & Baltensweiler, A. (2013). High-resolution remote sensing data improves models of species richness. *Applied Vegetation Science*, *16*, 539–551.

Canty, A., & Ripley, B. (2014). *Boot: Bootstrap R (S-Plus) functions. R package version 1, 3–13*.

Carpenter, S., Bennett, E., & Peterson, G. (2006). Scenarios for ecosystem services: An overview. *Ecology and Society*, *11*(1), 1–29.

Ceballos, A., Hernández, J., Corvalán, P., & Galleguillos, M. (2015). Comparison of airborne LiDAR and satellite hyperspectral remote sensing to estimate vascular plant richness in deciduous Mediterranean forests of Central Chile. *Remote Sensing*, *7*(3), 2692–2714.

Chaudhuri, P., Lo, W. D., Loh, W. Y., & Yang, C. C. (1995). Generalized regression trees. *Statistica Sinica*, *5*, 641–666.

Chevan, A., & Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician*, *45*, 90–96.

Choi, Y., Ahnb, H., & Chen, J.J. (2005). Regression trees for analysis of count data with extra Poisson variation. *Computational Statistics and Data Analysis*, *49*, 893–915.

Ciampi, A. (1991). Generalized regression trees. *Computational Statlstics and Data Analysis*, *12*, 57–78.

Clawges, R., Vierling, K., Vierling, L., & Rowell, E. (2008). The use of airborne LiDAR to assess avian species diversity, density, and occurrence in a pine/aspen forest. *Remote Sensing of the Environment*, *112*, 2064–2073.

Corvalán, P., Galleguillos, M., & Hernández, J. (2014). Presencia, abundancia y asociatividad de *Citronella mucronata* en bosques secundarios dominados por *Nothofagus obliqua* de la precordillera de Curicó, región del Maule, Chile. *Bosque*, *35*, 269–278.

Dauber, J., Hirsch, M., Simmering, D., Waldhardt, R., Otte, A., & Wolters, V. (2003). Landscape structure as an indicator of biodiversity: Matrix effects on species richness. *Agriculture, Ecosystems and Environment*, *98*, 321–329.

Duffy, E. (2009). Why biodiversity is important to the functioning of real-world ecosystems. *Frontiers in Ecology and the Environment*, *7*, 437–444.

Dufour, A., Gadallah, F., Wagner, H., Guisan, A., & Butler, A. (2006). Plant species richness and environmental heterogeneity in a mountain landscape: Effects of variability and spatial configuration. *Ecography*, *29*, 573–584.

Emborg, J. (1998). Understorey light conditions and regeneration with respect to the structural dynamics of a near-natural temperate deciduous forest in Denmark. *Forest Ecology and Management*, *106*, 83–95.

Eskelson, B.N.I., Madsen, L., Hagar, J.C., & Temesgen, H. (2011). Estimating riparian understory vegetation cover with beta regression and copula models. *Forest Science*, *57*, 212–221.

Everson, D., & Boucher, H. (1998). Tree species-richness and topographic complexity along the riparian edge of the Potomac River. *Forest Ecology and Management*, *109*, 305–314.

van Ewijk, K.Y., Randin, C.F., Treitz, P.M., & Scott, N. (2014). Predicting fine-scale tree species abundance patterns using biotic variables derived from LiDAR and high spatial resolution imagery. *Remote Sensing of the Environment*, *150*, 120–131.

Falkowski, M.J., Evans, J.S., Martinuzzi, S., Gessler, P.E., & Hudak, A.T. (2009). Characterizing forest succession with LiDAR data: An evaluation for the Inland Northwest, USA. *Remote Sensing of Environment*, *113*, 946–956.

Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014a). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, *154*, 102–114.

Fassnacht, F., Neumann, C., Förster, M., Buddenbaum, H., Ghosh, A., Clasen, A., Joshi, P. K., & Koch, B. (2014b). Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites. *IEEE Journal of Selected Topics in Applied Earth Observation and remote Sensing*, *7*, 2547–2561.

Fava, F., Parolo, G., Colombo, R., Gusmeroli, F., Della Marianna, G., Monteiro, A., & Bocchi, S. (2010). Fine-scale assessment of hay meadow productivity and plant diversity in the European Alps using field spectrometric data. *Agriculture, Ecosystems & Environment*, *137*, 151–157.

Feilhauer, H., & Schmidtlein, S. (2009). Mapping continuous fields of forest alpha and beta diversity. *Applied Vegetation Science*, *12*, 429–439.

Foody, G. (2005). Mapping the richness and composition of British breeding birds from coarse spatial resolution satellite sensor imagery. *International Journal of Remote Sensing*, *26*, 3943–3956.

Foody, G., & Cutler, M. (2003). Tree biodiversity in protected and logged Bornean tropical rain forests and its measurement by satellite remote sensing. *Journal of Biogeography*, *30*, 1053–1066.

Foody, G., & Cutler, M. (2006). Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. *Ecological Modelling*, *195*, 37–42.

French, J.R. (2003). Airborne LiDAR in support of geomorphological and hydraulic modelling. *Earth Surface Processes and Landforms*, *28*, 321–335.

Gajardo, R. (1994). *La vegetación natural de Chile*. Editorial Universiaria (165 pp.).

Gaston, K. (2000). Global patterns in biodiversity. *Nature*, *405*, 220–227.

Ghosh, A., & Joshi, P.K. (2014). A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 298–311.

Gilbert, B., & Lechowicz, M. (2004). Neutrality, niches, and dispersal in a temperate forest understory. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 7651–7656.

Goetz, S., Steinberg, D., Dubayah, R., & Blair, B. (2007). Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. *Remote Sensing of the Environment*, *108*, 254–263.

Guariguata, M.R., & Ostertag, R. (2001). Neotropical secondary forest succession: changes in structural and functional characteristics. *Forest Ecology and Management*, *148*, 185–206.

Hayes, A., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Henrik, H., Moen, J., Virtanen, R., Grytnes, J., Oksanen, L., & Angerbjörn, A. (2006). Effects of altitude and topography on species richness of vascular plants, bryophytes and lichens in alpine communities. *Journal of Vegetation Science*, *17*, 37–46.

Hernández-Stefanoni, J., Dupuy, J., Johnson, K., Birdsey, R., Tun-Dzul, F., Peduzzi, A., ... López-Merlín, D. (2014). Improving species diversity and biomass estimates of tropical dry forests using airborne LiDAR. *Remote Sensing*, *6*, 4741–4763.

Higgins, M., Asner, G., Martin, R., Knapp, D., Anderson, C., Kennedy-Bowdoin, T., ... Wright, S. (2014). Linking imaging spectroscopy and LiDAR with floristic composition and forest structure in Panama. *Remote Sensing of the Environment*, *154*, 358–367.

Kattenborn, T., Maack, J., Fassnacht, F.E., Enßle, F., Ermert, J., & Koch, B. (2015). Mapping forest biomass from space — Fusion of hyperspectral EO1-hyperion data and

tandem-X and WorldView-2 canopy height models. *International Journal of Applied Earth Observation and Geoinformation*, *35*, 359–367.
Kuenzer, C., Ottinger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., ... Wikelski, M. (2014). Earth observation satellite sensors for biodiversity monitoring: Potentials and bottlenecks. *International Journal of Remote Sensing*, *35*, 37–41.
Latifi, H., Fassnacht, F.E., & Koch, B. (2012). Forest structure modeling with combined airborne hyperspectral and LiDAR data. *Remote Sensing of Environment*, *121*, 10–25.
Lefsky, M.A., Cohen, W.B., Parker, G.G., & Harding, D.J. (2002). LiDAR remote sensing for ecosystem studies. *Biosciences*, *52*(1), 19–30.
Lemenih, M., Gidyelew, T., & Teketay, D. (2004). Effects of canopy cover and understory environment of tree plantations on richness, density and size of colonizing woody species in southern Ethiopia. *Forest Ecology and Management*, *194*, 1–10.
Leutner, B., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., & Wegmann, M. (2012). Modelling forest α-diversity and floristic composition — On the added value of LiDAR plus hyperspectral remote sensing. *Remote Sensing*, *4*, 2818–2845.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.
Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 14–23.
Lopatin, J., Galleguillos, M., Fassnacht, F.E., Ceballos, A., & Hernández, J. (2015). Using a multistructural object-based LiDAR approach to estimate vascular plant richness in Mediterranean forests with complex structure. *IEEE Geoscience and Remote Sensing Letters*, *12*, 1008–1012.
Luebert, F., & Pliscoff, P. (1999). *Sinopsis bioclimática y vegetacional de Chile*. Santiago, Chile: Editorial Universitaria (1999, In Spanish).
Manning, W., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, *20*, 461–494.
Mathlouthi, W., Fredette, M., & Larocque, D. (2015). Regression trees and forests for non-homogeneous Poisson processes. *Statistics and Probability Letters*, *96*, 204–211.
Meynard, C.N., & Quinn, J.F. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, *34*, 1455–1469.
Moeslund, J., Arge, L., Bocher, P., Dalgaard, T., Odgaard, M., Nygaard, B., & Svenning, J. (2013). Topographically controlled soil moisture is the primary driver of local vegetation patterns across a lowland region. *Ecosphere*, *4*, 1–26.
Morsdorf, F., Kötz, B., Meier, E., Itten, K.I., & Allgöwer, B. (2006). Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction. *Remote Sensing of Environment*, *104*, 50–61.
Nagendra, H., Rocchini, D., Ghate, R., Sharma, B., & Pareeth, S. (2010). Assessing plant diversity in a dry tropical forest: Comparing the utility of Landsat and IKONOS satellite images. *Remote Sensing*, *2*, 478–496.
Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*, 370–384.
O'Hara, R., & Kotze, D. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*, 118–122.
Oldeland, J., Wesuls, D., Rocchini, D., Schmidt, M., & Jurgens, N. (2010). Does using species abundance data improve estimates of species diversity from remotely sensed spectral heterogeneity? *Ecological Indicators*, *10*, 390–396.
Palmer, M., Earls, P., Hoagland, B., White, P., & Wohlgemuth, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics*, *13*, 121–137.
Pateiro-Lopez, B., & Rodriguez-Casal, A. (2013). Alphahull: Generalization of the convex hull of a sample of points in the plane. R package version 1.0. http://CRAN.R-project.org/package=alphahull
Pope, G., & Treitz, P. (2013). Leaf area index (LAI) estimation in borealmixedwood forest of Ontario, Canada using light detection and ranging (LiDAR) and Worldview-2 imagery. *Remote Sensing*, *5*, 5040–5063.
Popescu, S., Wynne, R.H., & Nelson, R.F. (2003). Measuring individual tree crown diameter with LiDAR and assessing its influence on estimating forest volume and biomass. *Remote Sensing of Environment*, *29*, 564–577.
R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (URL http://www.R-project.org/).
Rocchini, D. (2007). Effects of spatial and spectral resolution in estimating ecosystem ?—Diversity by satellite imagery. *Remote Sensing of Environment*, *111*, 423–434.
Rocchini, D., Balkenhol, N., Carter, G.A., Foody, G.M., Gillespie, T.W., He, K.S., ... Neteler, M. (2010). Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges. *Ecological Informatics*, *5*, 318–329.
Rocchini, D., Chiarucci, A., & Loiselle, S.A. (2004). Testing the spectral variation hypothesis by using satellite multispectral images. *Acta Oecologica*, *26*, 117–120.
Silvertown, J., Dodd, M., Gowing, D., & Mountford, J. (1999). Hydrologically defined niches reveal a basis for species richness in plant communities. *Nature*, *400*, 61–63.
Simonson, W., Allen, H., & Coomes, D. (2012). Use of an airborne LiDAR system to model plant species composition and diversity of Mediterranean oak forests. *Conservation Biology*, *26*, 840–850.
Stein, A., Gerstner, K., & Kreft, H. (2014). Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecology Letters*, *17*, 866–880.
Su, J.G., & Bork, E.W. (2007). Characterization of diverse plant communities in aspen parkland rangeland using LiDAR data. *Applied Vegetation Science*, *10*, 407–416.
Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*, 1947–1958.
Therneau, T., Atkinson, B., & Ripley, B. (2015). *Rpart: Recursive partitioning and regression trees. R package version 4.1–9 http://CRAN.R-project.org/package=rpart*
Turner, W. (2014). Sensing biodiversity. *Science*, *346*, 301–302.
Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., & Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology and Evolution*, *18*, 306–314.
Vierling, K., Bässler, C., Brandl, R., Vierling, L., Weiss, I., & Müller, J. (2011). Spinning a laser web: predicting spider distributions using LiDAR. *Ecological Applications*, *21*, 577–588.
Vogeler, J.C., Hudak, A.T., Vierling, L., Evans, J., Green, P., & Vierling, K.T. (2014). Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. *Remote Sensing of Environment*, *147*, 13–22.
Wing, B.M., Ritchie, M.W., Boston, K., Cohen, W.B., Gitelman, A., & Olsen, M.J. (2012). Prediction of understory vegetation cover with airborne LiDAR in an interior ponderosa pine forest. *Remote Sensing of Environment*, *124*, 730–741.
Wolf, J.A., Fricker, G.A., Meyer, V., Hubbell, S.P., Gillespie, T.W., & Saatchi, S.S. (2012). Plant species richness is associated with canopy height and topography in a neotropical forest. *Remote Sensing*, *44*, 4010–4021.
Woods, M., Lim, K., & Treitz, P. (2008). Predicting forest stand variables from LiDAR data in the Great Lakes—St Lawrence forest of Ontario. *The Forestry Chronicle*, *84*, 827–839.
Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, *27*, 1–25.