Predicting Vascular Plant Richness in a Heterogeneous Wetland Using Spectral and Textural Features and a Random Forest Algorithm

Julián Cabezas, Mauricio Galleguillos, and Jorge F. Perez-Quezada

***Abstract*—A method to predict vascular plant richness using spectral and textural variables in a heterogeneous wetland is presented. Plant richness was measured at 44 sampling plots in a 16-ha anthropogenic peatland. Several spectral indices, first-order statistics (median and standard deviation), and second-order statistics [metrics of a gray-level co-occurrence matrix (GLCM)] were extracted from a Landsat 8 Operational Land Imager image and a Pleiades 1B image. We selected the most important variables for predicting richness using recursive feature elimination and then built a model using random forest regression. The final model was based on only two textural variables obtained from the GLCM and derived from the Landsat 8 image. An accurate predictive capability was reported ($R^2 = 0.6$; RMSE $= 1.99$ species), highlighting the possibility of obtaining parsimonious models using textural variables. In addition, the results showed that the mid-resolution Landsat 8 image provided better predictors of richness than the high-resolution Pleiades image. This is the first study to generate a model for plant richness in a wetland ecosystem.**

***Index Terms*—Gray-level co-occurrence matrix (GLCM), Landsat, peatland, Pleiades, remote sensing, textural variables.**

I. Introduction

THE relationship between plant richness in ecosystems and the benefits they provide human beings has become the focus of attention in recent years [1]. One of the main challenges is the assessment of local variations using plant species richness as an indicator of biodiversity [2]. Environmental managers require continuous and detailed information of these indicators. In this sense, Earth observation systems are an optimal option for measuring the status and trends of biodiversity at different spatial scales [3].

Remote sensing has been widely used for plant richness modeling. Several studies have developed predictive models of species richness considering reflectance values or spectral indices of a single pixel, such as NDVI or DVI [4]. However, other studies have shown that spectral and spatial heterogeneity can be better predictors of plant richness [5]. This idea is based on the spectral variation hypothesis (SVH) [6], which states that spectral heterogeneity is positively correlated with species richness, because a more heterogeneous area can host a wider variety of ecological niches and habitats, increasing the possibility of finding a greater number of different species. Applying this hypothesis, Viedma *et al.* [7] tested several reflectance values and textural metrics obtained from a QuickBird image (2.4-m spatial resolution, 4 bands) as predictors of plant richness in a burned area of central Spain. Other authors, such as Nagendra *et al.* [8], have used the standard deviation of the NDVI or the metrics obtained from a gray-level co-occurrence matrix (GLCM) [9] as spatial heterogeneity metrics [7]. More examples of the use of spatial heterogeneity as a proxy of biodiversity can be found in Rocchini *et al.* [5].

Most of the work concerning richness predictions based on using remote sensing has been done for forested ecosystems [4], [7]–[9]. In wetland ecosystems, spatial predictions have not yet been documented, despite the fact that, in these ecosystems, there is a considerable number of flora species with strong singularity [10]. In addition, wetlands around the world are often under threatened scenarios (almost half of all wetland areas have been lost due to human activities) [10]. The closest study related to this topic was presented by Rocchini [11] about Montepulciano Lake, where the author explored the relationships between textural variables and species richness with no predictive model presented.

When working with biodiversity and remote sensing, spectral and spatial resolutions are important factors to consider [5]. Rocchini [11] showed that spectral heterogeneity derived from the reflectance values of a high-spatial-resolution QuickBird image and a Landsat ETM+ image (30-m spatial resolution, 6 bands) performed similarly when predicting plant richness, inferring that the higher spectral resolution of the Landsat compensates for its lower spatial resolution. Following a different approach, Stickler *et al.* [12] showed that the best model for habitat prediction was the one that combined Landsat and QuickBird data.

The random forest regression algorithm (RF) is a method created by Breiman [13] that builds large amounts of bootstrapped trees to obtain an average result [14]. The use of this methodology is frequent in land cover classification studies [15] but somewhat less frequent in regression analyses, mainly used

Manuscript received August 10, 2015; revised December 14, 2015; accepted January 13, 2016. Date of publication March 22, 2016; date of current version April 20, 2016. This work was supported by CONICYT through the FONDECYT project 1130935.

J. Cabezas is with the Department of Environmental Science and Renewable Natural Resources, University of Chile, Santiago 1004, Chile.

M. Galleguillos is with the Department of Environmental Science and Renewable Natural Resources, University of Chile, Santiago 1004, Chile, and also with the Center for Climate and Resilience Research (CR2), University of Chile, Santiago 1004, Chile (e-mail: mgalleguillos@renare.uchile.cl).

J. F. Perez-Quezada is with the Department of Environmental Science and Renewable Natural Resources, University of Chile, Santiago 1004, Chile, and also with the Institute of Ecology and Biodiversity, Santiago 653, Chile.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2016.2532743

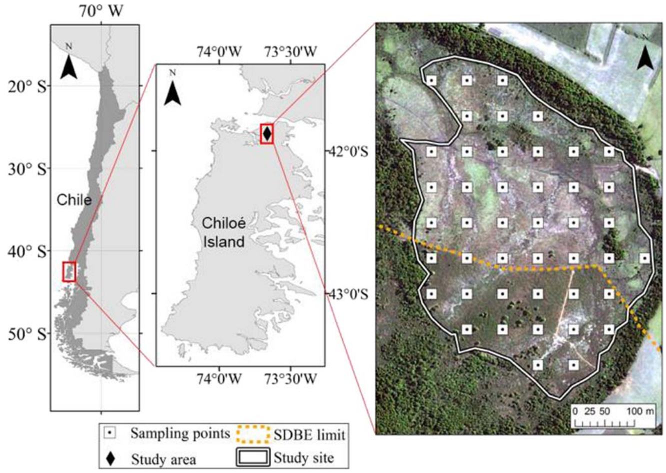

Fig. 1. Location of the study site in the north of Chiloé Island in southern Chile. The white squares represent the location of the sampling plots. SDBS property is on the southern side of the wetland.

for the prediction of aboveground biomass [16]. The ability of this method to handle large data sets [17] makes it suitable for predicting richness using a large number of different textural variables and indices.

This evidence suggests that a combination of high spatial resolution and the inclusion of multispectral data (the latter used for building vegetation indices) and the use of the spectral heterogeneity concept (through the extraction of textural data) can produce an acceptable prediction of plant richness in heterogeneous wetland ecosystems. The objective of this study was to develop a method for predicting plant richness at a local scale, using a recursive feature elimination procedure (RFE) and the capability of an RF algorithm that can handle large amounts of textural data to generate a parsimonious model, i.e., selecting only the most relevant predictors.

II. Methods

A. *Study Area*

The study area is a 16-ha wetland, located in the north of Chiloé Island in Chile (41°52′ S, 73°40′ W). This ecosystem has its origins in the cutting and burning of a *Tepualia stipularis* forest, leaving a poorly drained soil colonized by different species. The study area is located in a temperate climate with a strong ocean influence, with a two-month dry period during the summer; the average annual temperature is 10 °C, and the mean annual precipitation ranges from 2000 to 2500 mm.

The wetland is divided into two types of management: a conservation management area with 5.5 ha located inside the Senda Darwin Biological Station (SDBS; Fig. 1) and a productive management area that covers the rest of the peatland (10.5 ha). In the latter area, Sphagnum moss is harvested on a nonindustrial scale for commercial purposes, while this area is also used for grazing by four oxen. The particular characteristics of this ecosystem along with the different management activities have produced the formation of different microsites dominated by shrubs, competitive grasses, or weeds [18], causing a strong heterogeneity in this wetland.

TABLE I
Indices Calculated From the Landsat 8 OLI and Pleiades 1B Image

Index	Ref.	OLI	Pleiades
Normalized Difference Vegetation Index (NDVI)	[21]	✓	✓
Enhanced Vegetation Index (EVI)	[21]	✓	✓
Soil Adjusted Vegetation Index (SAVI)	[21]	✓	✓
Green Normalized Difference Vegetation Index (GNDVI)	[22]	✓	✓
All Normalized Difference Vegetation Index	[16]	✓	✓
Tasseled Cap Transformations: Greenness, Brightness and Wetness	[23]	✓	
Difference Vegetation Index (DVI)	[24]	✓	✓
Simple Ratio (SR)	[24]	✓	✓
Modified Simple Ratio (MSR)	[25]	✓	✓
Renormalized Difference Vegetation Index (RDVI)	[25]	✓	✓
Modified Soil Adjusted Vegetation Index 2 (MSAVI2)	[25]	✓	✓
Triangular Vegetation Index (TVI)	[25]	✓	

B. *Field and Remote Sensing Variables*

A systematic sampling of vascular plant richness was performed, creating a 60-m grid, which resulted in 44 sampling points where 2 × 2 m quadrants were sampled (Fig. 1). The sampling was done during the summer (January 2014).

Vascular plant richness (expressed as the number of species) was modeled using predictors derived from remote sensing. These predictive variables were obtained using an image from the Operational Land Imager (OLI) sensor onboard the Landsat 8 satellite, taken on December 24, 2013, and a second image from Pleiades 1B taken on January 28, 2014. The two images were radiometrically corrected using the gain and offset values given in the metadata of each of the images, and an atmospheric correction was applied using the Fast Line-of-sight Atmospheric Adjustment of Spectral Hypercubes (FLAASH) algorithm with the values for the atmospheric variables given by the MODTRAN4 algorithm for this zone [19]. All of the image preprocessing was performed using the software ENVI 5.1 (Exelis Visual Information Solutions, Boulder, CO, USA).

The predictive variables were the reflectance values for all bands from both images and several indices and transformations obtained from them (Table I); for each one, textural variables were calculated using a window of 3 × 3 pixels for the Pleiades and Landsat images. By using this window size, the Pleiades textural variables could account for microvariations around the sampling plot, whereas the Landsat variables could account for macrovariations produced, for example, by different management regimes in the wetland [18]. As per Mairota *et al.* [20], we separated the textural variables into first- and second-order statistical parameters. For the former, we used the median and standard deviation of the pixels in the window, calculated using the focal statistics tool in ArcGis 10 (ESRI, Redlands, CA, USA). As second-order statistical parameters, we calculated the GLCM metrics mean, variance, homogeneity, contrast, correlation, second moment, and dissimilarity. This resulted in a total of 533 predictor variables. The value for each sample point was extracted using bilinear interpolation.

C. Selection of Variables

A feature selection procedure was performed to select only the most relevant variables using an RFE algorithm. The RFE is a recursive method that ranks variables according to a measure of importance (in this case the measure of importance of RF regression [13]); then, the least important variable is deleted, and a measure of accuracy is calculated (in this case the root-mean-square error or RMSE). This algorithm was implemented using a tenfold cross-validation to stabilize the selection [26], obtaining the relation between the RMSE and the number of predictors and also obtaining the most important variables at each step. This procedure is implemented in the caret package of the software R. The RFE-RF has been used for classification, especially in cases with a large number of predictor variables, and is related to other disciplines, as, for example, in Granitto *et al.* [27], where the authors proved that the combination of these methods produced models with low multicollinearity.

We sought a model with low RMSE and low number of variables; thus, we followed the guidelines suggested by Hair *et al.* [28], which are the following: the number of observations must not be $<$ 5 per predictor, and ideally, the number of observations should be between 15 and 20 per explanatory variable. Once we obtained the first subset of variables that fulfilled these indications, we tested the correlation between variables using Pearson's correlation coefficient (R). Then, we selected the pairs of predictors with absolute R value $>$ 0.6 and eliminated the least important one by one (according to the importance ranking given by RFE-RF).

D. Prediction of Richness and Validation Procedure

Once the final subset of variables was obtained, an RF regression was performed. With the predicted richness values, we calculated the coefficient of determination (R^2), RMSE, and relative RMSE (RRMSE) (estimated by dividing the RMSE by the mean of the observations). Also, the bias of the model was calculated as one minus the slope of a regression of the predicted and the observed values, without the intercept.

To evaluate the quality of the model, we used a bootstrapping technique with 1000 iterations. On each iteration, the data were divided into train and test subsets, with an average of 36.2% of the data to test the model, calculating the RMSE and R^2 of the test subset in every iteration to plot their distribution [29]. All of the statistical analyses were carried out with the R 3.1.2 software (R Core Team, 2014).

III. Results and Discussion

A. Selection of Variables

Following the recommendations of Hair *et al.* [28] and the number of observations in our study $(n = 44)$, the number of variables should have been lower than eight, and the ideal number is around three. When the RFE was applied, a preliminary number of seven variables were chosen because, in this step, the RMSE was minimized (Fig. 2). The analysis showed high correlation values $(R > 0.6)$ between five pairs of predictor variables, so five variables were eliminated, yielding a final number of two variables (Table II). Obtaining a model with only two variables is significant because, even though RF can deal with a large number of predictors, the performance of the model can decrease with a large number of noisy variables [14].

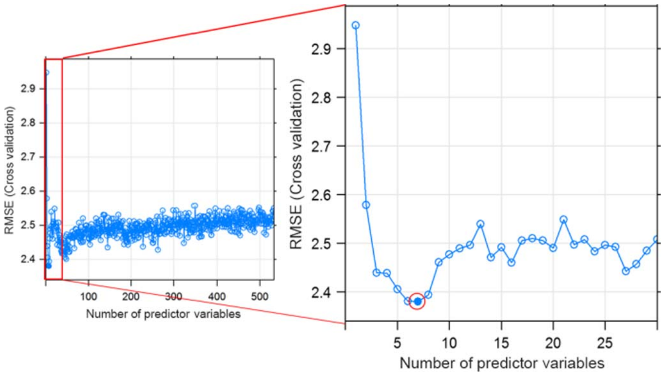

Fig. 2. RFE. The preliminary number of variables $(n = 7)$ is marked with a red circle.

TABLE II
Variables Selected for Predicting Vascular Plant Richness

Index or Variable	Texture metric	Source (pixels window)
Normalized difference between blue and red bands	Contrast (GLCM)	Landsat 8 (3×3)
Normalized difference between NIR and SWIR2 bands	Variance (GLCM)	Landsat 8 (3×3)

All of the final selected predictors were textural variables derived from the GLCM (contrast and variance; Table II). This confirms the power of this method to characterize the spatial heterogeneity of the image, providing good proxies for habitat spatial heterogeneity and, consequently, for plant richness [20]. These variables are textural metrics of normalized differences between bands from the Landsat image. This coincides with the results of Mohammadi and Shataee [4], who showed that the texture metrics of spectral indices perform better than the texture metrics of single bands. Also, this result confirms the idea behind the SVH, showing that spectral heterogeneity is related to richness. In this case, the GLCM metrics worked better than other characterizations of spatial biodiversity, such as the standard deviation of the reflectance, used by Palmer *et al.* [6] when they first proposed the SVH.

Our results are similar to Nagendra *et al.* [8], showing that, for predicting richness, the variables obtained from Landsat were chosen over the variables obtained from the higher spatial resolution images from Pleiades because it increases the pixel-to-pixel variation. This can be due to the fact that one of the selected predictors includes the SWIR bands (which are not available in the Pleiades image), where the water absorption is stronger, making this band suitable to differentiate wetland conditions [30]. The fact that all of the selected variables were extracted from Landsat can be explained because an overly

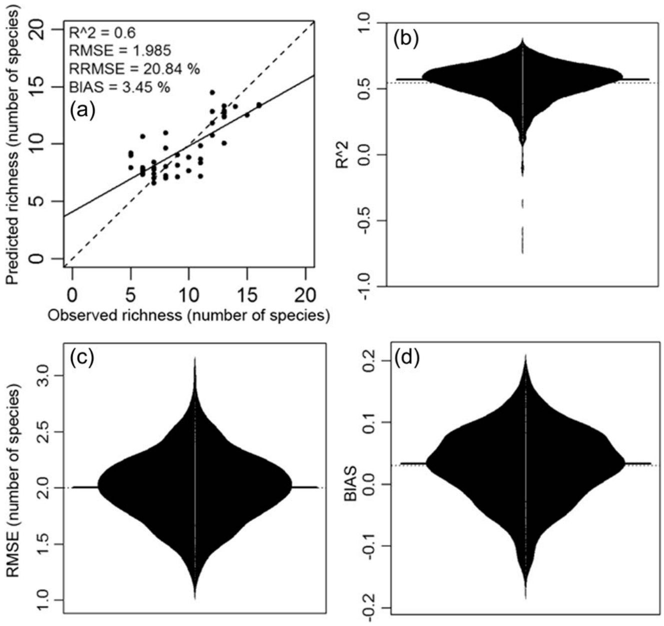

Fig. 3. Model adjustment and validation. Panel (a) shows the observed versus predicted values; panels (b)–(d) show the distribution after 1000 bootstrapping iterations for R^2 (b), RMSE (c), and bias (d).

sharp resolution can overestimate the spatial variation, being influenced by shadows produced by the objects on the terrain [5].

B. Model Validation

The final model showed a good prediction capability, with an R^2 of 0.6, an RMSE of 1.99 species (20.84% of the mean), a bias of 3.45%, and a slight tendency to overestimate the total richness [Fig. 3(a)]. The distribution of the bootstrapped values of these metrics showed a low probability of obtaining a low adjustment, with most of the values close to the mean [Fig. 3(b)–(d)].

This is the first model for plant richness prediction in a wetland ecosystem. The prediction of this model surpasses the model constructed with the mean NDVI and the SD of the NDVI by Gillespie [31], with an R^2 of 0.44 in the tropical dry forest of Florida (USA). Our model has a similar R^2 compared with the model for predicting tree richness reported by Mohammadi and Shataee [4] in the Hyrcanian forest of Iran (adjusted $R^2 = 0.59$). In a similar way, Ceballos *et al.* [32] constructed a model for richness prediction using hyperspectral and LiDAR data from a Mediterranean forest in Chile, obtaining good results ($R^2 = 0.59$). Using the same data set, Lopatin *et al.* [29] achieved better results ($R^2 = 0.64$) by processing LiDAR data with an object-based image analysis. The present model was achieved only with the usage of remote sensing variables, and it has high prediction performance with 21% of inaccuracy. This model could be potentially improved by using additional variables (e.g., topography) to obtain even better results.

IV. Conclusion

A parsimonious model has been obtained using only two textural variables for predicting vascular plant richness in a heterogeneous wetland. The results showed the possibility of using RFE-RF and a correlation analysis for effective feature selection, achieving easily interpretable results. Furthermore, second-order textural variables (derived from GLCM), combined with spectral indices, showed a good prediction capability of the vascular plant richness, representing the first documented predictive model for wetland ecosystems. Finally, we have found that the mid-resolution image (Landsat 8) provided good predictors of plant richness in this kind of ecosystem, leaving open the opportunity of such research for larger spatial extensions.

References

[1] J. Duffy, “Why biodiversity is important to the functioning of real-world ecosystems,” *Front. Ecol. Environ.*, vol. 7, no. 8, pp. 437–444, Aug. 2009.

[2] K. J. Gaston, “Global patterns in biodiversity,” *Nature*, vol. 405, no. 6783, pp. 220–227, May 2000.

[3] W. Turner *et al.*, “Remote sensing for biodiversity science and conservation,” *Trends Ecol. Evol.*, vol. 18, no. 6, pp. 306–314, Jun. 2003.

[4] J. Mohammadi and S. Shataee, “Possibility investigation of tree diversity mapping using Landsat ETM+ data in the Hyrcanian forests of Iran,” *Remote Sens. Environ.*, vol. 114, no. 7, pp. 1504–1512, Jul. 2010.

[5] D. Rocchini *et al.*, “Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges,” *Ecol. Inform.*, vol. 5, no. 5, pp. 318–329, Sep. 2010.

[6] M. W. Palmer, P. G. Earls, B. W. Hoagland, P. S. White, and T. Wohlgemuth, “Quantitative tools for perfecting species lists,” *Environmetrics*, vol. 13, no. 2, pp. 121–137, Mar. 2002.

[7] O. Viedma, I. Torres, B. Pérez, and J. M. Moreno, “Modeling plant species richness using reflectance and texture data derived from QuickBird in a recently burned area of Central Spain,” *Remote Sens. Environ.*, vol. 119, pp. 208–221, Apr. 2012.

[8] H. Nagendra, D. Rocchini, R. Ghate, B. Sharma, and S. Pareeth, “Assessing plant diversity in a dry tropical forest: Comparing the utility of Landsat and Ikonos satellite images,” *Remote Sens.*, vol. 2, no. 2, pp. 478–496, Feb. 2010.

[9] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE T. Syst. Man. Cybern.*, vol. 3, no. 6, Nov. 1973.

[10] J. B. Zedler and S. Kercher, “Wetland resources: Status, trends, ecosystem services, and restorability,” *Annu. Rev. Envon. Resources*, vol. 30, no. 1, pp. 39–74, Nov. 2005.

[11] D. Rocchini, “Effects of spatial and spectral resolution in estimating ecosystem α-diversity by satellite imagery,” *Remote Sens. Environ.*, vol. 111, no. 4, pp. 423–434, Dec. 2007.

[12] C. M. Stickler and J. Southworth, “Application of multi-scale spatial and spectral analysis for predicting primate occurrence and habitat associations in Kibale National Park, Uganda,” *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2170–2186, May 2008.

[13] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer-Verlag, 2009.

[15] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random forests for land cover classification,” *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, Mar. 2006.

[16] O. Mutanga, E. Adam, and M. A. Cho, “High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm,” *Int. J. Appl. Earth Observ.*, vol. 18, pp. 399–406, Aug. 2012.

[17] C. Crisci, B. Ghattas, and G. Perera, “A review of supervised machine learning algorithms and their applications to ecological data,” *Ecol. Model.*, vol. 240, pp. 113–122, Aug. 2012.

[18] J. Cabezas *et al.*, “Evaluating the impacts of management in an anthropogenic peatland using field and remote sensing data,” *Ecosphere*, vol. 6, no. 12, pp. 1–24, Dec. 2015.

[19] M. W. Matthew *et al.*, “Status of atmospheric correction using a MODTRAN4-based algorithm,” in *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI*, vol. 01803, S. S. Shen and M. S. Descour, Eds. Orlando, FL, USA: SPIE Digital Library, 2000, pp. 199–207.

[20] P. Mairota *et al.*, "Very high resolution Earth observation features for monitoring plant and animal community structure across multiple spatial scales in protected areas," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 37, pp. 100–105, May 2015.

[21] A. R. Huete, H. Q. Liu, K. Batchily, and W. Van Leeuwen, "A comparison of vegetation indices over a global set of TM images for EOS-MODIS," *Remote Sens. Environ.*, vol. 59, no. 3, pp. 440–451, Mar. 1997.

[22] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzljak, "Use of a green channel in remote sensing of global vegetation from EOS-MODIS," *Remote Sens. Environ.*, vol. 55, no. 3, pp. 289–298, Dec. 1996.

[23] M. H. A. Baig, L. Zhang, T. Shuai, and Q. Tong, "Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance," *Remote Sens. Lett.*, vol. 5, no. 5, pp. 423–431, Jun. 2014.

[24] C. Ling *et al.*, "Study on above-ground biomass estimation of East Dong Ting lake wetland based on WorldView-2 data," in *Proc. 3rd Int. Workshop Earth Observ. Remote Sens. Appl.*, Jun. 2014, pp. 428–432.

[25] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, Apr. 2004.

[26] L. Breiman, "Heuristics of instability and stabilization in model selection," *Ann. Statist.*, vol. 24, no. 6, pp. 2350–2383, 1996.

[27] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, 2006.

[28] J. F. J. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

[29] J. Lopatin, M. Galleguillos, F. E. Fassnacht, A. Ceballos, and J. Hernández, "Using a multistructural object-based LiDAR approach to estimate vascular plant richness in Mediterranean forests with complex structure," *IEEE Geosci. Remote Sens.*, vol. 12, no. 5, pp. 1008–1012, May 2015.

[30] E. Adam, O. Mutanga, and D. Rugege, "Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review," *Wetlands Ecol. Manage.*, vol. 18, no. 3, pp. 281–296, Dec. 2009.

[31] T. W. Gillespie, "Predicting woody-plant species richness in tropical dry forests: A case study from south Florida, USA," *Ecol. Appl.*, vol. 15, no. 1, pp. 27–37, 2005.

[32] A. Ceballos, J. Hernández, P. Corvalán, and M. Galleguillos, "Comparison of airborne LiDAR and satellite hyperspectral remote sensing to estimate vascular plant richness in deciduous Mediterranean forests of central Chile," *Remote Sens.*, vol. 7, no. 3, pp. 2692–2714, Mar. 2015.