

# Evolutionary feature selection to estimate forest stand variables using LiDAR

Jorge Garcia-Gutierrez, Eduardo Gonzalez-Ferreiro, Jose C. Riquelme-Santos, David Miranda, Ulises Dieguez-Aranda, Rafael M. Navarro-Cerrillo

*Keywords:*  
Evolutionary computation  
Forest-stand variables  
LiDAR  
Regression  
Stepwise selection

## A B S T R A C T

Light detection and ranging (LiDAR) has become an important tool in forestry. LiDAR-derived models are mostly developed by means of multiple linear regression (MLR) after stepwise selection of predictors. An increasing interest in machine learning and evolutionary computation has recently arisen to improve regression use in LiDAR data processing. Although evolutionary machine learning has already proven to be suitable for regression, evolutionary computation may also be applied to improve parametric models such as MLR. This paper provides a hybrid approach based on joint use of MLR and a novel genetic algorithm for the estimation of the main forest stand variables. We show a comparison between our genetic approach and other common methods of selecting predictors. The results obtained from several LiDAR datasets with different pulse densities in two areas of the Iberian Peninsula indicate that genetic algorithms perform better than the other methods statistically. Preliminary studies suggest that a lack of parametric conditions in field data and possible misuse of parametric tests may be the main reasons for the better performance of the genetic algorithm. This research confirms the findings of previous studies that outline the importance of evolutionary computation in the context of LiDAR analysis of forest data, especially when the size of fieldwork datasets is reduced.

## 1. Introduction

Light detection and ranging (LiDAR) has transformed the way in which forestry studies are carried out. The work previously done with costly fieldwork is partially replaced by the processing of LiDAR data. Thus, LiDAR is generally used for conducting forest inventories (Montaghi et al., 2013; Kronseder et al., 2012), developing fuel models (Garcia et al., 2011), and characterizing the main vertical forest information (Côté et al., 2011). Processing of LiDAR data for forestry applications usually starts by filtering the LiDAR point cloud to separate terrain and non-terrain returns (Gonçalves-Seco et al., 2011). Then, a set of statistics is extracted from the returns' normalized heights and (often) intensities. Later those statistics are used to develop models that regress a set of target variables (fieldwork measurements). Multiple linear regression (MLR) with previous stepwise variable selection is mostly responsible for the generation of the model (Tesfamichael et al., 2010; Dalponte et al., 2011; Sun et al., 2011; Tattoni et al., 2012).

Stepwise selection and MLR are both parametric techniques (i.e., they depend on the data distribution). This means that they require normality, homoscedasticity, and independence in the data to be used properly (Demsar, 2006), although these conditions are barely taken into account in most studies (Garcia et al., 2010b).

Machine learning has been paid increasing attention for both classification and regression on LiDAR in the last years (Chen et al., 2012; Finley et al., 2013). Moreover, some authors have recently proposed that the limitations of classical stepwise MLR on LiDAR data could be overcome by the use of non-parametric machine learning techniques. Thus, Zhao et al. (2011) showed that Gaussian processes could be a good tool to improve the results of classical MLR and more recently, Gleason and Im (2012) studied support vector machines and other modern regression techniques (e.g., random forests) for biomass estimation. The power of machine-learning non-parametric regression is based on the fact that it does not depend on any priori assumptions about the data. On the other hand, the resulting models may have a higher number of parameters and may lack the parametric models' simplicity and clarity (Millie et al., 2012). In addition, locally collected fieldwork datasets are usually reduced and can have a high level of error (Kitahara et al., 2010). Since machine-learning techniques have a higher intrinsic

dependence on training data (usually fieldwork), they might produce overconfident predictions (Zaffalon, 2005) and have a higher risk of overfitting (Hawkins, 2004), leading to subsequent problems in generalizing the resulting models to other areas.

Furthermore, the feature selection phase (also known as variable selection or screening) is an important step in the process of LiDAR knowledge generation. A good feature selection technique can make machine learning focus on the most important features and avoid redundant or irrelevant data which can affect machine learning final result or performance (Mallet et al., 2011). Some techniques such as random forests include their own feature selection (Guo et al., 2011) and for others such as the nearest neighbour family, it is well-known that their accuracies decrease in absence of feature selection (Hughes, 1968). In the case of regression, the variables selection filters the original set of potential predictors to keep models as simple as possible whilst their description power is maximized following the principle of parsimony (Peña, 2002).

Variable screening is classified as a hard problem in computer science since its complexity is exponential to the number of the possible parameters ( $2^p$  possible models for  $p$  parameters). This high level of complexity introduces the need for automatic or semi-automatic procedures to select the set of best predictors. As stated above, stepwise selection is the most usual variable-selection technique in LiDAR-related research. It selects the best subset of predictors by introducing variables to the candidate model and removing them from it in several iterations. A set of hypothesis tests is usually responsible for the selection of the variables that will be introduced or eliminated. The process ends when it is possible to neither introduce nor eliminate more variables. The fact that parametric tests ( $F$ -tests) are used to decide which variables are the best for the model can mean that stepwise selection works properly when the data meet parametric conditions (Garcia et al., 2010b). Normality is guaranteed when the data size is large enough according to the theorem of the central limit (Dudley, 1999). Unfortunately, reduced fieldwork datasets are a reality in many studies (Lovell et al., 2005) and the limit for normality is not known at the time of fieldwork planning. Besides, other reasons such as dependence of the results on the order of parameter entry (or deletion) or the absence of control for the inflation of type I errors in the sequence of statistical tests have been reported, which has discouraged the use of stepwise regression (Whittingham et al., 2006).

Some of the limitations of stepwise regression arise from the fact that it presumes that there is only one best subset with a number of variables. This affirmation is not always true and therefore some statisticians propose that such limitations should be overcome by the calculation of every possible regression with size  $X$  (Hill and Lewicki, 2006). This greedy type of technique is called best-subset selection and provides the best model but with a much greater computational effort.

In recent times, evolutionary computation (programming techniques inspired by the Darwinian evolution process) in the form of genetic algorithms (GAs) has appeared as a suitable tool to overcome the previous problems related to variable selection in environmental studies (Song et al., 2012). In the context of LiDAR regression, a GA has been defined as a variable-screening procedure, based on the principle of evolution by natural selection, which develops sets of variables that have evolved to fit a certain fitness function via cycles of a differential replication, recombination, and mutation procedure (Latifi et al., 2012). Thus, Latifi et al. (2010) showed that the use of this type of algorithm may improve the results of non-parametric prediction techniques for stand timber volume and biomass. In the same line, Latifi et al. (2012) showed that GAs may also play an important role in the selection of predictors not only in LiDAR but also fused with hyperspectral data. Li et al. (2013) also proved GAs power when they

are combined with a linear discriminant analysis to classify tree species. Although all of these proposals showed that GAs are suitable tools for improving LiDAR results, most authors used them to optimize machine-learning techniques such as nearest neighbour analysis or ensembles (combinations of several machine learning techniques) but not MLR.

The purpose of this article is to compare the results of two classical procedures (stepwise and best-subset) and a novel GA regression procedure for the selection of variables when classical MLR is applied on LiDAR data in two forest areas of the Iberian peninsula. Thus, we aimed to explore the generation of new regression models with advanced variable-screening techniques such as GAs, testing their performance and checking the significance of the differences from the other classical variable screening techniques.

## 2. Materials and method

### 2.1. Study sites

For this study, we used statistics extracted from aerial LiDAR data in two forest areas in the northwest part of the Iberian Peninsula (Fig. 1). The study area of Trabada (hereafter site A) was a  $1 \times 4$  km rectangle located in the municipality of Vilapena (Galicia, NW Spain; boundaries 644,800; 4806,600 and 645,800; 4810,600 UTM). A wide variety of landform types and an elevation range of 150 to 530 masl characterize this area. The forest type was dominated by *Eucalyptus globulus* Labill. stands, with low intensity silvicultural treatments and the presence of tall shrubs.

The study area of Guitiriz (hereafter site B) was also located in Galicia (NW Spain) and covered about 36 km<sup>2</sup> of *Pinus radiata* D. Don forests. In this case, the study site was defined as a  $4.130 \times 8.787$  km<sup>2</sup> rectangle (586,315; 4783,000 and 595,102; 4787,130 UTM). The forest type in this area was *P. radiata* stands, which were also characterized by low-intensity silvicultural treatments and by the presence of tall shrubs (Supplementary Information, Gonçalves-Seco et al., 2011 and Gonzalez-Ferreiro et al., 2012).

### 2.2. Field data

On site A, a total of 39 square plots of 225 m<sup>2</sup> were located and measured in the *E. globulus* plantations between February and March 2005. On site B, a similar process was carried out for a total of 54 square plots of 225 m<sup>2</sup> in *P. radiata* plantations between August and December 2007. The plots were selected to represent the existing range of ages, stand sizes, and densities of the studied forest. For every tree in each sample plot, two measurements of diameter at breast height (1.3 m above ground level) were made at right angles with a tree calliper. Measurements were made to the nearest millimetre, and the arithmetic mean of the two measurements was calculated. Total tree height was measured with a Vertex III hypsometer.

For site A, the dry weight of the biomass fractions of each tree was estimated using the equations for *E. globulus* in Galicia reported by Diéguez-Aranda et al. (2009). In order to define the dependent variables, the field measurements (heights and diameters) and the estimated dry weight of the biomass fractions were used to calculate the following stand variables in each plot: stand crown biomass ( $W_{cr}$ ), stand stem biomass ( $W_{st}$ ), and stand aboveground biomass ( $W_{abg}$ ). In the case of site B, the equations developed by Diéguez-Aranda et al. (2009) were also used to estimate biomass variables for each tree (volume and dry weight of biomass fractions). Then the field measurements (heights and diameters) and the estimated volumes and dry weight of the biomass fractions

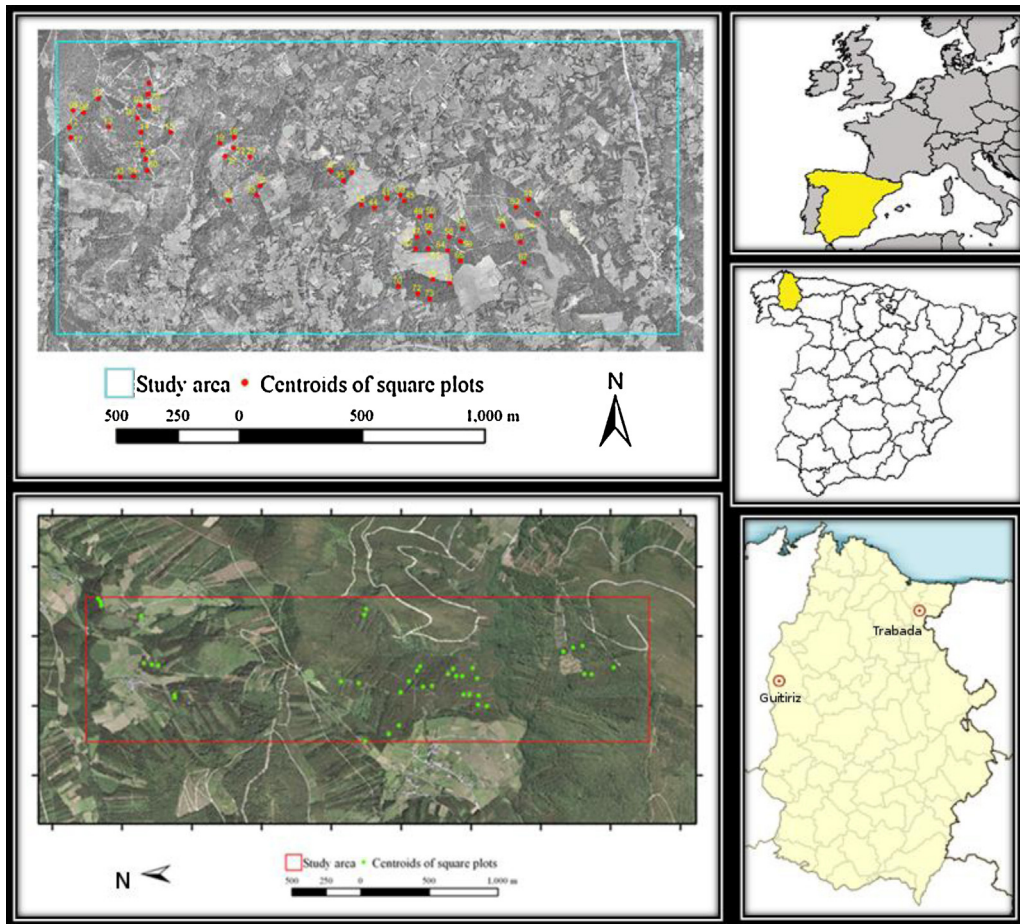


Fig. 1. Study sites located in the province of Lugo (NW of Spain). Top: study site of Guitiriz. Bottom: area of Trabada.

helped to estimate the following stand variables of each plot: stand crown biomass ( $W_{cr}$ ), stand stem biomass ( $W_{st}$ ), stand aboveground biomass ( $W_{abg}$ ), stand basal area ( $G$ ), dominant height ( $H_d$ ), mean height ( $H_m$ ), and stand volume ( $V$ ).

### 2.3. LiDAR data

The LiDAR data from site A were acquired in November 2004 with an Optech Airborne Laser Terrain Mapper (ALTM) 2033 sensor operated at a laser wavelength of 1064 nm from a flight altitude of 1500 m asl. The beam divergence was 0.3 mrad, the pulsing frequency 33 kHz, the scan frequency 50 Hz, and the maximum scan angle  $\pm 10^\circ$ . The first and last return pulses were registered. The whole study area was flown over in 18 strips and each strip was flown over three times, which gave an average measurement density of about 4 pulses  $m^{-2}$ . To obtain two different resolutions, an artificial reduction based on a random selection of LiDAR returns in a grid cell of 1  $m^2$  was carried out and resulted in LiDAR data with a return density of 0.5 returns  $m^{-2}$  (Gonzalez-Ferreiro et al., 2012).

The LiDAR data for site B were acquired in September 2007 using an Optech ALTM 3025 system, operated at 1064 nm, with a laser repetition rate of 25 kHz, a scan frequency of 200 Hz, a maximum scan angle of  $\pm 17^\circ$ , and a flying height of 1300 m asl. A theoretical laser pulse density of 8 pulses  $m^{-2}$  was obtained. As in the case of study site A, an artificial dataset with a pulse density of 0.5 pulses  $m^{-2}$  was generated to obtain a new dataset with lower resolution.

Furthermore, intensity values from LiDAR pulses in both areas were also used in this work. Intensities in both study sites were

normalized according to the equation described in Höfle and Pfeifer (2007) to eliminate the influence of path height variations (Garcia et al., 2010a). Unfortunately, no calibration was planned before the flights so intensity could not be corrected and therefore models that included intensity statistics could not be directly used in other areas. Moreover, problems related to multiple returns (Höfle and Pfeifer, 2007), which are very common in forest areas, were not taken into account either. Although the problems associated with the use of intensity might have suggested not to include it into the models, for the sake of comparison, we decided to follow the same methodology as in the previously published works for the study sites which did use intensity statistics (Gonçalves-Seco et al., 2011; Gonzalez-Ferreiro et al., 2012).

Filtering, interpolation, and the development of Digital Terrain and Canopy Models (DTM/DCM) were performed by FUSION software (McGaughey, 2009). For filtering, FUSION used an adapted algorithm from Kraus and Pfeifer (1998). This software also provided the variables related to the height and return intensity distributions within the limits of the field plots in the four datasets (original and reduced data from study sites A and B). As stated before, we used the same complete set of variables following the methodology described in the previous works on both sites.

Table 1 shows the complete set of metrics and the corresponding abbreviations used in this article. From now on, we will refer to them with a prefix (except for the cases of cover, FP and returns). Thus, for the same statistic, the prefix “h” involves a calculation from the height distribution while “i” refers to a calculation from the intensities distribution.

**Table 1**  
Statistics extracted from the LiDAR flights' heights and intensities used as independent variables for the regression models.

Description	Abbreviation
Percentage of first returns over 2 m	cover_FP
Number of returns above 2 m	returns
Minimum	min
Maximum	max
Mean	mean
Mode	mode
Standard deviation	SD
Variance	V
Interquartile distance	ID
Skewness	Skw
Kurtosis	Kurt
Average absolute deviation	AAD
25th percentile	P25
50th percentile	P50
75th percentile	P75
5th percentile	P05
10th percentile	P10
20th percentile	P20
30th percentile	P30
40th percentile	P40
60th percentile	P60
70th percentile	P70
80th percentile	P80
90th percentile	P90
95th percentile	P95

#### 2.4. Stepwise selection

Stepwise regression was primarily used to define the empirical relations between LiDAR and fieldwork incorporating and removing features from the model in several steps. It tries to preserve the independent variables that contain relevant information and to eliminate the redundant features following the so-called principle of parsimony (Hosmer and Lemeshow, 2000). An important fact is that the introduction or deletion of a candidate variable depends on the result of a parametric test ( $F$ -test) which requires normality, homoscedasticity, and independence, among other conditions (Watson, 1994). For this study, data were processed using stepwise selection implemented in SPSS 15.0 software ([www.spss.com](http://www.spss.com)).

#### 2.5. Best-subset selection

Once the stepwise regression had been applied, the Mallows'  $C_p$  selection method was used as the best-subset selection technique. Mallows'  $C_p$  statistic follows Eq. (1), where  $\hat{x}$  is the predicted value for  $x_i$ ,  $\sigma^2$  is approximately the variance of the predictions in the model generated by the parameters selected, and  $p$  is the number of parameters in the model.

$$C_p = \frac{\sum_{i=0}^n (x_i - \hat{x})^2}{\sigma^2} + 2p \quad (1)$$

We applied the Mallows'  $C_p$  technique implemented in the REG procedure of SAS/STAT (Clark et al., 2004), which performs all possible subset regressions and lists the models in ascending order of  $C_p$  (best models are those with lowest  $C_p$ ). Only models whose parameters were significant at a given level (in our case, 5%) were taken into account (Gonzalez-Ferreiro et al., 2012).

#### 2.6. Genetic selection

A third level of data analysis was done using genetic selection techniques (Goldberg, 1989; Holland, 1992). GAs are programming techniques that maintain a population of solutions (individuals) on

a computer along several generations, allow the best individuals to reproduce, and let the worst individuals die without descendants (Renner and Ekárt, 2003). The evolutionary search (which tries to solve an optimization problem) is conducted towards better regions of the search space on the basis of a fitness measure. Thus, each individual in a population is evaluated based on how well it solves the given problem (fitness function). The new individuals inherit the properties of their parents, and the fitter ones (with respect to the predefined fitness function) survive for the next generation. The final individuals (solutions) will be better than the initial population (usually randomly selected).

The genetic selection used in this paper was implemented using the framework Watchmaker (Dyer, 2006), and its complete process is described in Fig. 2. Our GA started with a random population of possible individuals. In our case, an individual was a set of binary (0, 1) values associated with the possible predictors in the final model. Thus, the information was organized as an array where a value of 0 in a cell indicated that the corresponding predictor would not be selected for the final model. Otherwise, it would be included.

For each generation, the best individuals were selected and combined to produce new individuals (offspring). For the generation of each new offspring we used a classical roulette selection (Holland, 1992) to choose two parents for each new individual. The roulette selection technique assigned a probability to every individual (potential parent) according to Eq. (2) where  $f_i$  is the fitness reached by the  $i$ th individual. Then, it sorted the individuals by their probability (decreasing order) and for each parent selection, and it selected the first individual that made the accumulated probability be greater than a random value from 0 to 1.

$$P_i = \frac{f_i}{\sum_{i=1}^N f_i} \quad (2)$$

The crossing of each pair of parents resulted in two new individuals. This process was repeated until the production of a new offspring for each generation was completed, taking into account that the two best individuals of the previous generation always survived (elitism). New individuals were the result of a uniform crossover operator (Holland, 1992) which randomly selected one of the parent's binary values for each candidate variable to create a new individual. After the crossover, a mutation operator was applied with a given probability. Thus, a mutated individual included or removed a randomly selected predictor (changing the value associated with the predictor from 0 to 1 or vice versa, respectively).

Each individual was evaluated at the beginning of each generation using a fitness function which helped the genetic algorithm to select the best possible parents each generation. The fitness function used in this work was specially defined for the estimation of stand variables using MLR. Thus, it assigned the goodness of an individual according to the quality of a MLR model developed with the individual's selected variables. The goodness of the fitness was measured according to the Bayesian information criterion (BIC, Schwarz, 1978). BIC is expressed in Eq. (3), where  $L_p$  is the maximized log-likelihood of the model ( $\sum_{i=0}^n (x_i - \hat{x})^2 / \sigma^2$ ) as in the case of Mallows'  $C_p$ ,  $p$  is the number of predictors selected, and  $n$  is the number of plots in the dataset. Thus, BIC could provide a quality measurement penalizing the use of too many parameters (and therefore following the principle of parsimony).

It is also important to outline that our fitness function took into account multicollinearity. The condition index (CI) and the variance inflation factor (VIF) were checked among the explanatory variables selected by each individual. Individuals' predictor subsets with a condition index above 30 or a VIF above 10 were disregarded,

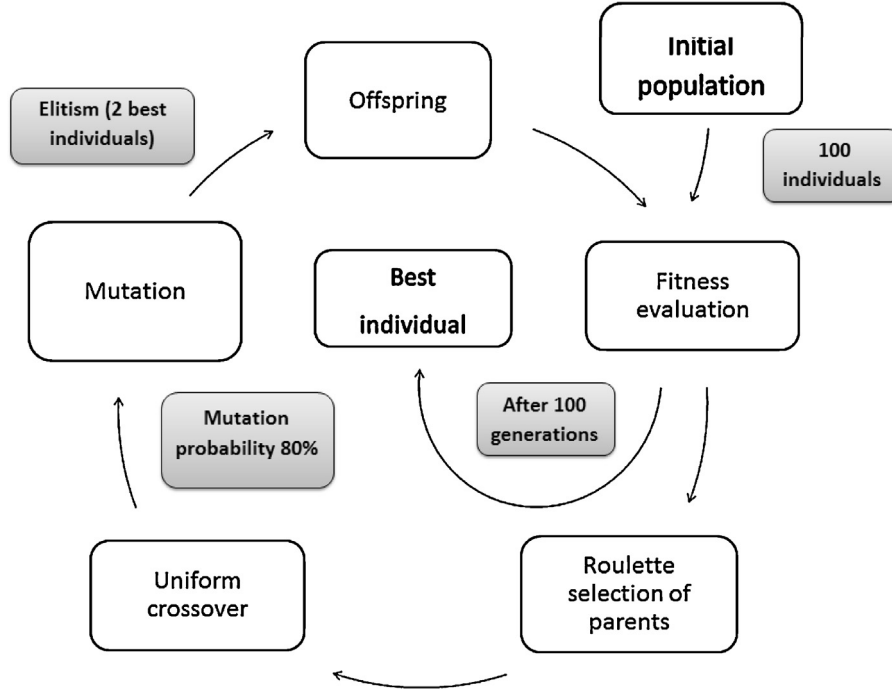


Fig. 2. Visual description of the general GA scheme followed in this work with the main parameters (numbers of generations, population size, probability of mutation, etc.).

as recommended by Belsley (1991) and Stevens (2002). For individuals with values over these thresholds, the worst possible fitness was assigned. This guaranteed the elimination of multicollinear solutions from the population in the genetic process.

$$\text{BIC} = -2L_p + p \ln(n) \quad (3)$$

A GA always depends on a set of parameters to control the trade-off between searching capacity and affordable computational time. Fig. 2 shows the selected parameters used in this work which were empirically adjusted. Finally, GAs also suffered from a high random influence. In this case, we reduced this undesirable effect by running the method 10 times for each dataset and variable. Then, the median of the genetic algorithm’s quality (the fifth best result) was selected for each case to be used in the comparison with the results of the rest of the methods.

### 2.7. Regression models

Linear, power function (multiplicative), and exponential MLR models were generated to define the relationships between the field measurements and the LiDAR variables selected by each method. The expressions of each type of model are as presented in Eqs. (4)–(6)

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon \quad (4)$$

$$Y = \beta_0 X_1^{\beta_1} + \dots + \beta_n X_n^{\beta_n} + \varepsilon \quad (5)$$

$$Y = \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n) + \varepsilon \quad (6)$$

where  $Y$  are the dependent variables obtained from fieldwork and  $X_1, \dots, X_n$  can be the metrics of the distributions of return heights and intensities or measurements related to canopy closure from the LiDAR data in a plot.

Before the model definition, and for selecting the best subset of independent variables obtained by each selection technique, the potential predictors and the dependent variables in multiplicative and exponential models were linearized by taking natural logarithms from both sides of Eqs. (5) and (6). Once the selection was

performed, the linearization was undone to obtain the exact quality measurements for each model.

The comparison of the estimates for the selected models was based on the following three statistics: the coefficient of determination ( $R^2$ ), the root mean square error (RMSE), and Bayesian information criterion (BIC). Although BIC was used as the final criterion for best model selection because of its better suitability for following the parsimony principle (Gonzalez-Ferreiro et al., 2012), it does not provide an intuitive idea of model precision.  $R^2$  and RMSE were therefore calculated to provide additional information. Thus,  $R^2$  indicates the proportion of the total variance of the dependent variable explained by the model and RMSE provides an idea of the precision of the estimates in the same units as the dependent variable.

We were not just interested in comparing candidate models in terms of their predictive capabilities but were interested in comparing variable-selection methods to obtain the best subset of predictors according to the parsimony principle. As stated before, a parsimonious model (Peña, 2002) maximizes its description power whilst it is as simple as possible. For the comparison of the models developed after the application of each selection technique, it must be taken into account that ordinary residuals are measures of quality of fit and not measures of the quality of future prediction (Myers, 1990). For prediction, models must be validated and to this end only a newly collected dataset is recommended (Kozak and Kozak, 2003). Therefore, model validation should be deferred until new datasets are available.

### 2.8. Statistical analysis

After the generation of the quality results for the MLR models with the variables selected by each method, a statistical analysis was used to check the significance in the differences among selection methods in terms of BIC. An analysis of variance (ANOVA) is usually used for multiple comparison of results if parametric conditions (homoscedasticity, independence, normality) are met (Demsar, 2006). Parametric conditions were checked using the Shapiro–Wilk (Shapiro and Wilk, 1965) and Lilliefors (Lilliefors,

**Table 2**  
Sets of predictors selected by the evolutionary procedure for the different forest variables (depending on the different pulses densities, sites and types of modelling).

Model	Var.	Site A		Site B	
		0.5 pulses m <sup>-2</sup>	4 pulses m <sup>-2</sup>	0.5 pulses m <sup>-2</sup>	8 pulses m <sup>-2</sup>
Mult.	$W_{cr}$	hmin, hV, hP25, imode	hP60	hKurt,hP90	hID, hKurt, iSkw
	$W_{st}$	hP60	hP60	hP50	hID, hKurt, iSkw
	$W_{abg}$	hP60	hP60	hP50	hID, hKurt, imin, iSkw
	$G$	–	–	hKurt,hP60	hID, hKurt, iSkw
	$H_d$	–	–	hP95	hP95
	$H_m$	–	–	hP50	hP40
	$V$	–	–	hP50	hID, hKurt, iSkw
	Exp.	$W_{cr}$	hP75	hP90	returns, hP50
$W_{st}$		hP70	hP75	returns, hP50	hKurt, hP75
$W_{abg}$		hP70	hP75	returns, hP50	returns, hKurt, hP75
$G$		–	–	returns, hSkw, hP50	returns,iSkw, hKurt, hAAD
$H_d$		–	–	hP95	hP95, iP10
$H_m$		–	–	returns, hP60, iAAD	hP40, iID
$V$		–	–	returns, hP50	returns, hKurt, hP75
Linear		$W_{cr}$	hP90	hP75	returns, hP40
	$W_{st}$	hmin, hmode, hV, hP10	hP75	hP50	returns, hP50, imax
	$W_{abg}$	hP50,hP75	hP75	hP50	returns, hP50, imax
	$G$	–	–	returns, hP40	hID, hKurt, imin, iKurt
	$H_d$	–	–	hP95	hP95
	$H_m$	–	–	hP50	hP30
	$V$	–	–	hP50	returns, hP50, imax

1967) tests for normality, and the Levene test (Levene, 1960) for homoscedasticity which are well-known techniques to check parametric conditions quantitatively. If normality or homoscedasticity conditions could not be met, a non-parametric procedure described in Garcia et al. (2010b) was applied instead the application of the ANOVA test. The non-parametric procedure firstly obtained the average ranks taking into account the position of the compared results with respect to each other. Thus, a value of 1 for a rank meant that the method was the best for a test case, while a rank of  $n$  meant that it was the worst of the  $n$  compared methods. Finally this non-parametric approach used the Friedman test and the Holm post-hoc procedure (see Luengo et al., 2009 for a complete description of both non-parametric methods) to statistically validate the differences in the mean ranks.

### 3. Results

Every selection method was tested using the stand variables and statistics extracted from the four datasets (site A with 0.5 and 4 pulses m<sup>-2</sup>, site B with 0.5 and 8 pulses m<sup>-2</sup>). The selected predictors obtained by the genetic approach for every stand variable can be seen in Table 2. Moreover Table 3 shows how often each technique obtained the best or worst result in terms of BIC,  $R^2$  and RMSE.

Figs. 3 to 8 show BIC,  $R^2$  and RMSE comparisons among the three types of variable selection methods for each site. In the figures, the results are presented in terms of the percentage of improvement in the corresponding statistic for each variable and type of model compared to that obtained by the worst variable screening method.

Once the results for stepwise, best-subset, and genetic regression had been obtained, a statistical analysis was performed to demonstrate whether the differences found among their BICs were

**Table 3**  
Positions reached by the results of each algorithm depending on the quality statistic after elimination of ties among all techniques (ties between two techniques were counted).

Result	Stepwise			Best-subset			Genetic		
	BIC	$R^2$	RMSE	BIC	$R^2$	RMSE	BIC	$R^2$	RMSE
Best	5	9	9	12	21	19	35	23	25
Worst	31	27	27	13	13	10	3	15	12

significant. It was possible to determine that the distribution of results did not meet the normality condition according to the Lilliefors test with a  $p$ -value of <0.01 and the Shapiro–Wilk test with a  $p$ -value of <0.0001, which were both significant at a level of  $\alpha = 0.05$ . Therefore, a non-parametric multiple comparison based on rankings was carried out. The mean rankings obtained were 1.55, 2.04, and 2.44 for the genetic, best-subset, and stepwise techniques respectively. Thus, the evolutionary approach ranked first on average and we could therefore apply the subsequent Friedman test and Holm’s post-hoc procedure to demonstrate that their differences were significant.

The null hypothesis for the Friedman test was that the ranks were not significantly different (they were not statistically different from the mean rank  $r = 2$ ). When the test was applied to assure the significance level of the results, its  $p$ -value was lower than  $1.47E - 5$ , so the null hypothesis was rejected. The next step was the use of the Holm procedure. The purpose of using Holm’s procedure was to check that the result of every pairwise comparison between an algorithm and our control method (genetic selection) was significant. The  $p$ -value obtained for every pairwise comparison was lower than the required value in every case (see the columns  $p$ -value and Holm’s threshold in Table 4), so all the null hypotheses (i.e., that there was no significant pairwise difference) were rejected.

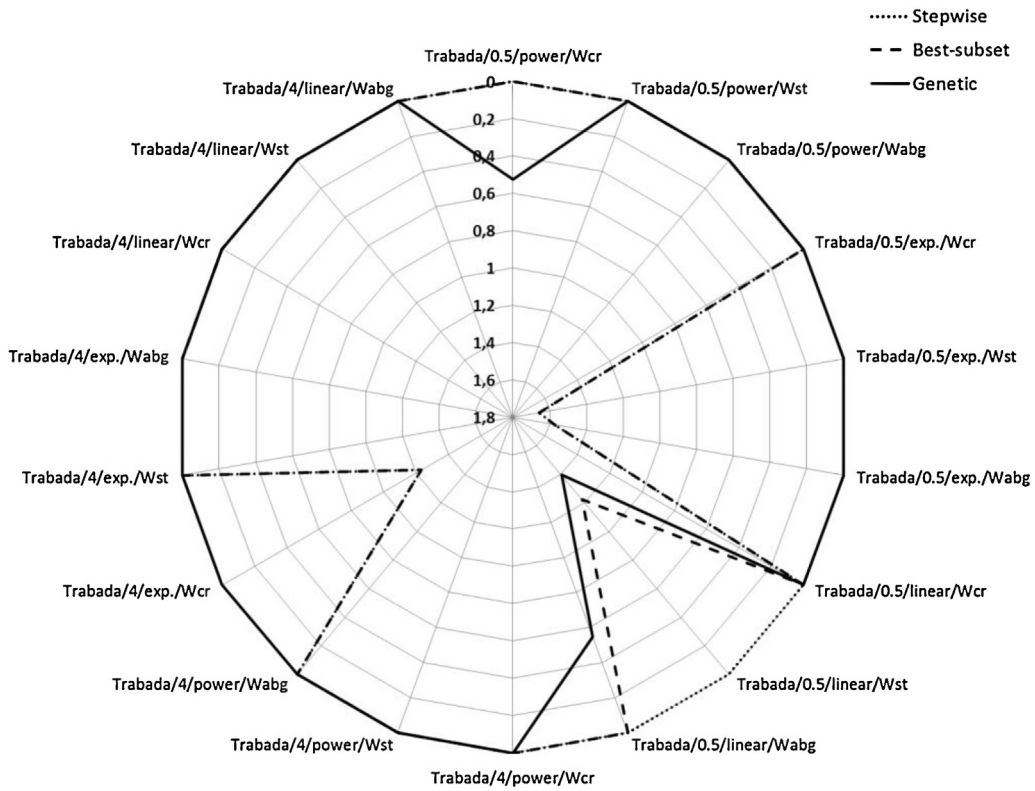
### 4. Discussion

Three variable screening techniques frequently used in LiDAR forest data analysis were tested in this work: the classical stepwise selection, the Mallows’  $C_p$  best-subset selection, and a genetic selection of predictors.

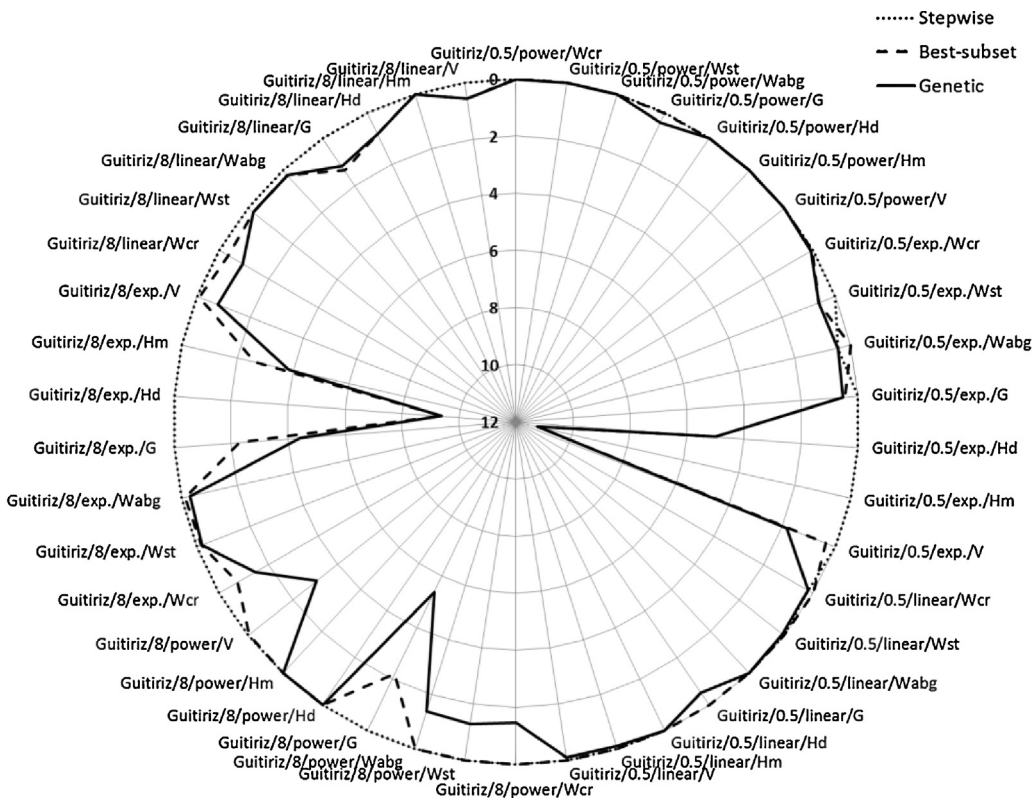
After analysing the BIC results obtained for every selection technique (see Table 3 and Appendix A), we observed some important facts. First, stepwise selection showed a limited power to detect the best subsets. It obtained the worst results in most cases (80% of the

**Table 4**  
Results of the Holm procedure comparison when our genetic selection was compared with every other algorithm for a level of significance of  $\alpha = 0.05$ .

Algorithm	$z = R_0 - R_i/SE$	$p$ -value	Holm’s threshold
Stepwise	4.70	2.59E–6	0.025
Best-subset	2.69	0.007	0.05



**Fig. 3.** Relative improvement with respect to the worst BIC for each variable and type of model in the area of Trabada with 0.5 and 4 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.



**Fig. 4.** Relative improvement with respect to the worst BIC for each variable and type of model in the area of Guitiriz with 0.5 and 8 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.

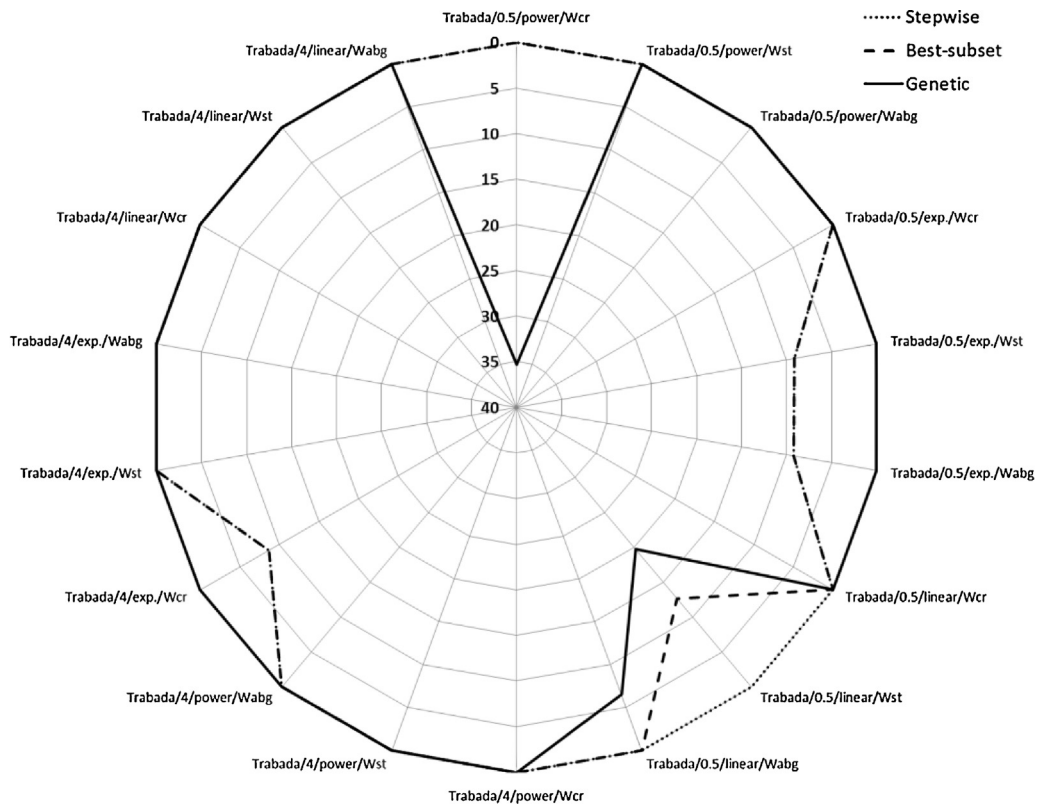


Fig. 5. Relative improvement with respect to the worst  $R^2$  for each variable and type of model in the area of Trabada with 0.5 and 4 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.

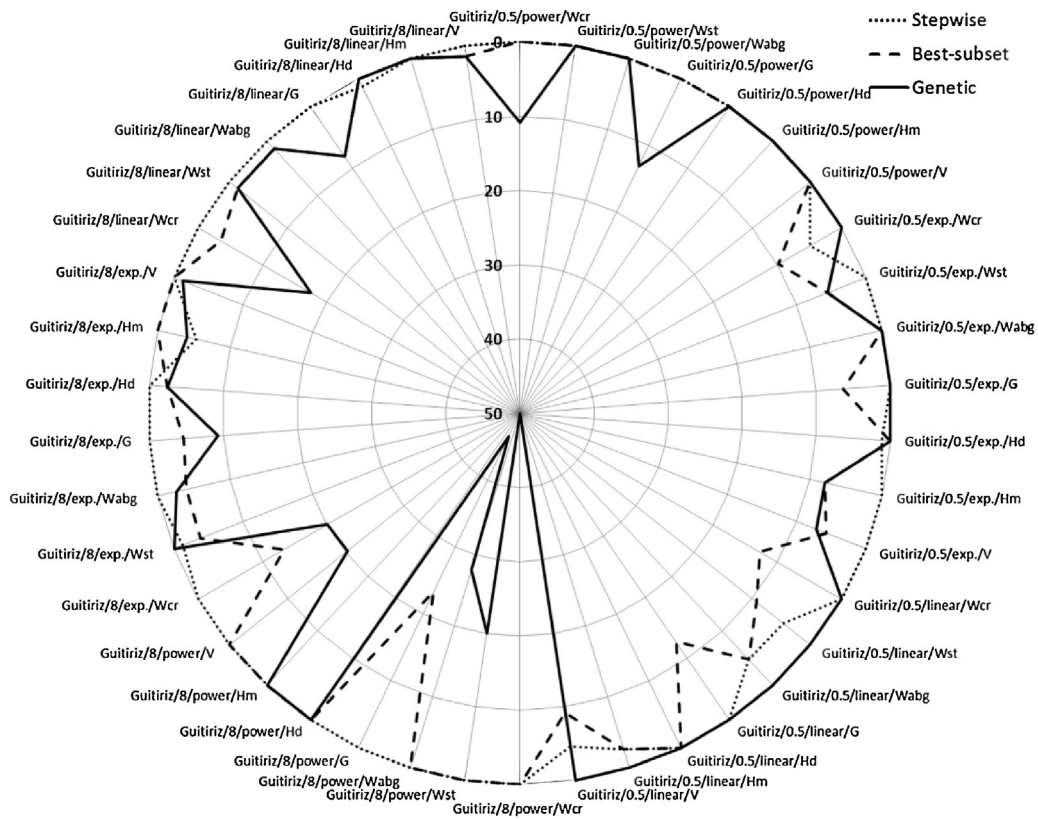


Fig. 6. Relative improvement with respect to the worst  $R^2$  for each variable and type of model in the area of Guitiriz with 0.5 and 8 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.



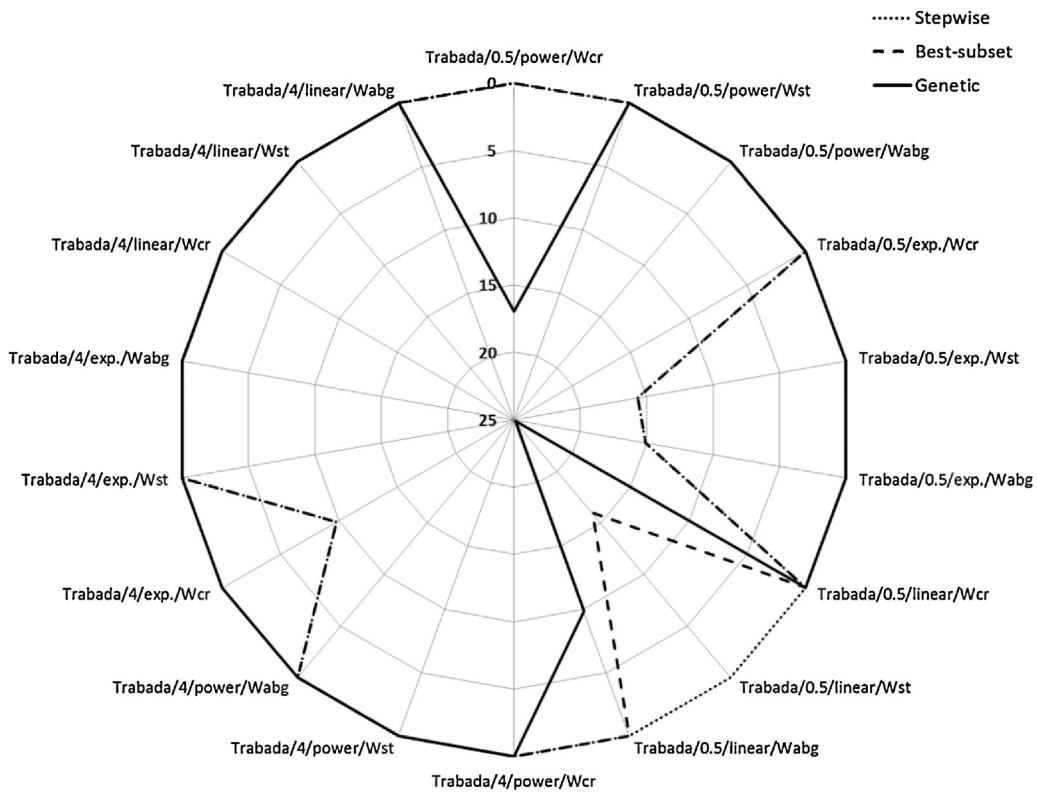


Fig. 7. Relative improvement with respect to the worst RMSE for each variable and type of model in the area of Trabada with 0.5 and 4 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.

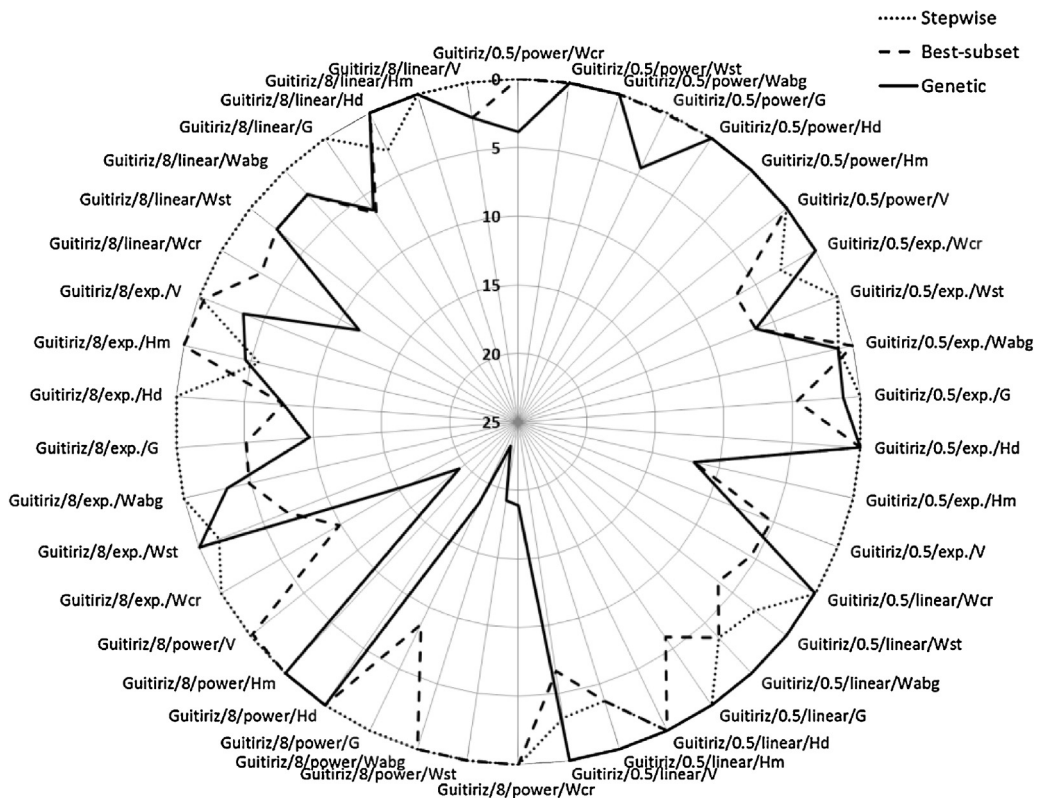


Fig. 8. Relative improvement with respect to the worst RMSE for each variable and type of model in the area of Guitiriz with 0.5 and 8 pulses  $m^{-2}$ . Every result is identified by its "site/resolution/model type/variable", where "variable" stands for each of the variables in Table 1.

cases with no tie among all the techniques) producing increases of up to 30% in RMSE and 10.78% on average. Best-subset selection performed better than stepwise selection but also obtained limited results if those are compared with the evolutionary solution. Mallows's  $C_p$  provided the worst results in the 33% of cases with no tie among all the techniques. It also resulted in increases of nearly 30% in the RMSE and 11.23% on average. Finally, GAs models mostly provided better results than the other variable selection methods since they obtained the same or better results (in terms of BIC) than the rest of the methods for every type of regression and variable except for three cases in site A and one case in site B. The last results might be attributed to the well-known random nature of GAs, which are techniques that obtain pseudo-optimal solutions, not necessarily the real best solution (Holland, 1992).

Although GAs obtained better results, the differences among the variable screening techniques were not high in general, as can be seen in Figs. 3 and 4. Our experiments showed that when the relationship between dependent and independent variables could be explained with just one variable (as in site A), the differences among the performance of the different methods became more reduced. Otherwise (site B) the differences increased. Nevertheless, even though site A provided fewer differences, GA results were statistically better than those provided by the other approaches.

From the results obtained, an important question emerges: Why do statistics-based selection methods (stepwise and best-subset) obtain worse results than a metaheuristic technique such as a GA? As previously mentioned, all of the methods provided nearly the same models with just one explicative variable but they differed when several variables were important for the model. Therefore, we can suppose that their differences were obtained in the process followed to select multiple variables.

The differences found between our genetic selection and the stepwise procedure could thus be based on the nature of their decision-making procedure. GAs techniques decide whether to introduce a variable or not with the aim of accomplishing an objective (fitness function) but without making any previous assumptions about the data (Holland, 1992). On the other hand, stepwise selection usually applies an  $F$ -test, which is a parametric test based on an  $F$  distribution, to decide whether a candidate variable is included in the final model or not (Ratner, 2012). The  $F$ -test's statistic can be seen in Eq. (7), where RSS is the residual sum of squares,  $n$  is the number of fieldwork measurements, and  $p$  is the number of predictors for the model tested.

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)} \quad (7)$$

The  $F$ -distribution is defined as the ratio of two  $\chi^2$  distributions. Furthermore, a  $\chi^2$  distribution is defined as the sum of normal distributions. Thus, the model's residual distribution in stepwise selection should be normally distributed to approximate the two  $\chi^2$  distributions. Otherwise the  $F$ -statistic would not have the claimed distribution and therefore the  $p$ -values yielded would not have the proper meaning (Ratner, 2012). When the fieldwork dataset is large enough, the theorem of the central limit (Dudley, 1999) guarantees normality, but for small datasets this condition should be explicitly checked to assure it is met (LeBlanc, 2004).

To check normality, we used statistical tests such as the Kolmogorov, Shapiro–Wilk, and Lilliefors tests (Luengo et al., 2009) applied to the residual distributions obtained by the stepwise models with more than one explicative variable. As was expected, the tests rejected the normality condition of the RSSs in most cases at a significance level of  $\alpha = 0.05$ , and concretely, the Lilliefors test rejected every single case.

Although non-normality in reduced datasets might be blamed for the problems of statistically dependent selection techniques, another concern related to stepwise selection may increase its undesirable effects. Thus, some researchers have reported problems regarding the use of multiple  $F$ -tests to build stepwise models when  $F$ -tests are actually designed for unique statistical comparisons (Whittingham et al., 2006). This inadequate use involves an increase in type I errors, which can reduce confidence in the final model. An example may clarify this point. Suppose we conduct tests on two parameters,  $X_1$  and  $X_2$ , with a significance level of  $\alpha = 0.05$  in each test. If we try to extract conclusions from both, the probability of making a correct decision on both parameters will be  $0.95 \times 0.95 = 0.90$ . In general, when testing  $n$  potential predictors, the overall error rate (also known as the family-wise error rate) in the corresponding  $n$  tests is  $(1 - \alpha)^n$  (Garcia et al., 2010b). In a context where we work with tens or hundreds of potential predictors, this fact seriously affects the statistical selection techniques and therefore decreases their reliability.

In the case of the best-subset regression, a previous manifold selection is usually needed since computers may not work in the proper manner on sets with many predictors; for example, a 1-GHz processor could need about 80 h to deal with the variations of the 48 predictors used in this study ( $2^{48}$  variations), but this number can easily be higher in data fusion (Latifi et al., 2012). Although human guidance could have caused the differences in the results, the problems related to the misuse of the  $F$ -test to assure models with statistically significant variables can again be the responsible for their limited performance since the  $F$ -test is used to generate models in which all of the variables are statistically different.

Taking into account the limitations of statistical selection methods, this work showed that genetic selection for LiDAR-derived regression can be a better option in studies related to LiDAR for estimation of forest stand variables as it has been observed in other areas of environmental knowledge (Song et al., 2012, Ramadana et al., 2001). Thus, genetic selection is especially interesting for those datasets with many variables from remote data fusion (e.g., LiDAR + hyperspectral, LiDAR + multispectral, etc.) and when it is suspected that parametric conditions are violated (e.g., reduced datasets). This finding agrees with other important results in the area (Latifi et al., 2012) and with previous work criticizing stepwise-related selection techniques (Whittingham et al., 2006).

## 5. Conclusions

This paper provided a hybrid approach based on the joint use of MLR and a novel genetic algorithm for the estimation of the main forest stand variables. We also showed a comparison between our genetic approach and other common methods for the selection of predictors. The results obtained for several LiDAR datasets with different pulse densities in two areas of the Iberian Peninsula indicated that GAs statistically outperformed the rest of the variable screening methods. Preliminary studies suggested that a lack of parametric conditions in field data and possible misuses of parametric tests could be the main reasons for the better performance of GAs.

Although our results seem to confirm that evolutionary variable screening outperforms classical approaches and that its use to develop MLR provided results that can compete with other machine-learning techniques, further research with independent validation data and a deeper comparison with other advanced techniques such as Gaussian processes (Zhao et al., 2011) or nearest-neighbour analysis (Latifi et al., 2010) is needed. Moreover transferability into industry is another issue not covered in this article. Concepts such as machine learning or evolutionary computation require a solid background which potential users

do not usually have. In the future, a specific software should be provided to make evolutionary feature selection as available as stepwise selection which is implemented in most commercial statistical packages. Finally, in the light of the results, we recommend that forestry researchers avoid the use of the classical stepwise procedure as predictors selection technique for LIDAR. Instead we recommend the use of evolutionary selection or if it is not available, a best-subset feature approach especially when more than one predictor is suspected to be included in the final models.

## Acknowledgements

This paper has been partially supported by the Spanish government under projects TIN2007-68084-C00, and

TIN2011-28956-C02; the Galician Government, Xunta de Galicia, Dirección Xeral de Montes (09MRU022291PR); Norvento (multinational energy company) (PGIDT09REM023E); Dirección Xeral de Ordenación e Calidade do Sistema Universitario de Galicia (Consellería de Educación e Ordenación Universitaria), the European Social Fund (Official Journal of Galicia—DOG n° 9, p. 2246, exp. 2011/14) and the DIVERBOS (CGL2011-30285-C02-02)/THERMOLIDAR (FP7-SME-2011-BSG). We would also like to thank the Regional Ministry of Andalusia for its help in the development of this work and Ph. D. Mariano Garcia and Prof. Luis Gonçalves-Seco for their invaluable advice during this research.

## Appendix A.

### Tables A.1–A.3

**Table A.1**

BICs,  $R^2$ , and RMSE ( $\text{kg ha}^{-1}$ ) obtained by the different models used to estimate forest biomass in site A with 0.5 and 4 pulses  $\text{m}^{-2}$ . The best BICs are shown in bold type.

Model/resolution	Var.	Stepwise			Best-subset			Genetic		
		BIC	$R^2$	RMSE	BIC	$R^2$	RMSE	BIC	$R^2$	RMSE
Mult./0.5	$W_{cr}$	670.33	0.51	4914.79	670.33	0.51	4914.79	<b>666.80</b>	0.69	4079.95
	$W_{st}$	<b>818.44</b>	0.64	32824.95	<b>818.44</b>	0.64	32824.95	<b>818.44</b>	0.64	32824.95
	$W_{abg}$	<b>829.02</b>	0.63	37592.31	<b>829.02</b>	0.63	37592.31	<b>829.02</b>	0.63	37592.31
Exp./0.5	$W_{cr}$	<b>649.67</b>	0.71	3771.46	<b>649.67</b>	0.71	3771.46	<b>649.67</b>	0.71	3771.46
	$W_{st}$	<b>787.69</b>	0.84	22128.16	<b>787.69</b>	0.84	22128.16	800.97	0.77	26237.74
	$W_{abg}$	<b>799.26</b>	0.83	25667.62	<b>799.26</b>	0.83	25667.62	812.03	0.76	30234.05
Linear/0.5	$W_{cr}$	<b>650.51</b>	0.71	3811.87	<b>650.51</b>	0.71	3811.87	<b>650.51</b>	0.71	3811.87
	$W_{st}$	809.93	0.71	29429.76	800.03	0.80	24731.75	<b>798.64</b>	0.85	22117.22
	$W_{abg}$	809.99	0.77	29452.93	809.99	0.77	29452.93	<b>805.55</b>	0.82	26546.40
Mult./4	$W_{cr}$	<b>659.79</b>	0.63	4293.74	<b>659.79</b>	0.63	4293.74	<b>659.79</b>	0.63	4293.74
	$W_{st}$	<b>801.34</b>	0.77	26359.82	<b>801.34</b>	0.77	26359.82	<b>801.34</b>	0.77	26359.82
	$W_{abg}$	<b>812.82</b>	0.75	30541.58	<b>812.82</b>	0.75	30541.58	<b>812.82</b>	0.75	30541.58
Exp./4	$W_{cr}$	<b>645.05</b>	0.75	3554.54	<b>645.05</b>	0.75	3554.54	653.10	0.69	3940.95
	$W_{st}$	<b>784.45</b>	0.85	21229.33	<b>784.45</b>	0.85	21229.33	<b>784.45</b>	0.85	21229.33
	$W_{abg}$	<b>795.36</b>	0.84	24415.85	<b>795.36</b>	0.84	24415.85	<b>795.36</b>	0.84	24415.85
Linear/4	$W_{cr}$	<b>648.94</b>	0.72	3735.92	<b>648.94</b>	0.72	3735.92	<b>648.94</b>	0.72	3735.92
	$W_{st}$	<b>794.86</b>	0.80	24260.93	<b>794.86</b>	0.80	24260.93	<b>794.86</b>	0.80	24260.93
	$W_{abg}$	<b>805.27</b>	0.80	27722.93	<b>805.27</b>	0.80	27722.93	<b>805.27</b>	0.80	27722.93

**Table A.2**

BICs,  $R^2$ , and RMSE ( $\text{kg ha}^{-1}$ ) obtained by the different models to estimate biomass in site B with 0.5 pulses  $\text{m}^{-2}$ . The best BICs are shown in bold type.

Model	Var.	Stepwise			Best-subset			Genetic		
		BIC	$R^2$	RMSE	BIC	$R^2$	RMSE	BIC	$R^2$	RMSE
Mult.	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	986.78	0.46	8629.51	986.78	0.46	8629.51	<b>986.53</b>	0.51	8297.40
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	<b>1149.22</b>	0.67	38834.54	<b>1149.22</b>	0.67	38834.54	<b>1149.22</b>	0.67	38834.54
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	<b>1170.97</b>	0.64	47497.25	<b>1170.97</b>	0.64	47497.25	<b>1170.97</b>	0.64	47497.25
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	258.03	0.46	10.13	257.96	0.46	10.12	<b>257.08</b>	0.52	9.68
	$H_d$ ( $\text{m ha}^{-1}$ )	<b>83.32</b>	0.84	2.01	<b>83.32</b>	0.84	2.01	<b>83.32</b>	0.84	2.01
	$H_m$ ( $\text{m ha}^{-1}$ )	<b>78.30</b>	0.76	1.92	<b>78.30</b>	0.76	1.92	<b>78.30</b>	0.76	1.92
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	<b>501.71</b>	0.66	96.70	<b>501.71</b>	0.66	96.70	<b>501.71</b>	0.66	96.70
Exp.	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	974.92	0.64	7181.70	974.64	0.67	6903.33	<b>974.15</b>	0.61	7398.72
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	1144.74	0.72	35903.71	<b>1137.60</b>	0.76	33608.97	<b>1137.60</b>	0.76	33608.97
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	<b>1157.68</b>	0.74	40476.63	1162.87	0.74	40926.34	<b>1157.68</b>	0.74	40476.63
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	248.54	0.63	8.62	247.41	0.67	8.22	<b>247.20</b>	0.63	8.51
	$H_d$ ( $\text{m ha}^{-1}$ )	87.58	0.85	2.01	<b>83.20</b>	0.84	2.01	<b>83.20</b>	0.84	2.01
	$H_m$ ( $\text{m ha}^{-1}$ )	83.44	0.76	1.94	<b>74.07</b>	0.82	1.71	<b>74.07</b>	0.82	1.71
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	499.48	0.70	91.29	497.72	0.74	86.55	<b>490.36</b>	0.75	83.90
Linear	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	981.44	0.55	7915.85	983.40	0.62	7486.98	<b>981.40</b>	0.55	7912.76
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	1152.49	0.68	38577.01	1152.60	0.71	37214.38	<b>1151.71</b>	0.65	39739.89
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	1171.62	0.66	46052.73	1171.62	0.66	46052.73	<b>1171.49</b>	0.63	47725.83
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	<b>253.00</b>	0.55	9.32	254.29	0.62	8.76	<b>253.00</b>	0.55	9.32
	$H_d$ ( $\text{m ha}^{-1}$ )	<b>82.19</b>	0.85	1.99	<b>82.19</b>	0.85	1.99	<b>82.19</b>	0.85	1.99
	$H_m$ ( $\text{m ha}^{-1}$ )	77.11	0.79	1.83	77.11	0.79	1.83	<b>77.03</b>	0.77	1.90
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	504.22	0.68	95.38	504.29	0.71	91.98	<b>503.68</b>	0.65	98.48

**Table A.3**

BICs,  $R^2$ , and RMSE ( $\text{kg ha}^{-1}$ ) obtained by the different models to estimate forest biomass in site B with 8 pulses  $\text{m}^{-2}$ . The best BICs are shown in bold type.

Model	Var.	Stepwise			Best-subset			Genetic		
		BIC	$R^2$	RMSE	BIC	$R^2$	RMSE	BIC	$R^2$	RMSE
Mult.	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	990.32	0.42	8917.63	990.32	0.42	8917.63	<b>975.71</b>	0.63	7234.44
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	1152.57	0.65	40056.37	1152.57	0.65	40056.37	<b>1137.52</b>	0.78	32365.70
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	1174.45	0.61	49053.00	1174.45	0.61	49053.00	<b>1157.99</b>	0.78	37701.36
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	261.39	0.43	10.45	255.64	0.53	9.55	<b>247.30</b>	0.63	8.52
	$H_d$ ( $\text{m ha}^{-1}$ )	<b>81.93</b>	0.85	1.98	<b>81.93</b>	0.85	1.98	<b>81.93</b>	0.85	1.98
	$H_m$ ( $\text{m ha}^{-1}$ )	<b>81.68</b>	0.74	1.98	<b>81.68</b>	0.74	1.98	<b>81.68</b>	0.74	1.98
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	505.50	0.64	100.15	505.27	0.64	99.94	<b>490.04</b>	0.77	80.62
Exp.	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	981.15	0.60	7608.05	973.82	0.68	6851.45	<b>966.72</b>	0.72	6415.11
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	1132.85	0.80	30997.20	1130.83	0.82	29318.35	<b>1130.60</b>	0.79	31499.03
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	1155.44	0.77	38208.33	1154.03	0.80	36345.64	<b>1151.93</b>	0.79	36985.59
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	250.09	0.65	8.42	244.32	0.68	7.99	<b>239.01</b>	0.71	7.60
	$H_d$ ( $\text{m ha}^{-1}$ )	89.38	0.84	2.05	<b>80.99</b>	0.86	1.89	<b>80.99</b>	0.86	1.89
	$H_m$ ( $\text{m ha}^{-1}$ )	82.02	0.79	1.84	79.87	0.75	1.95	<b>78.87</b>	0.78	1.86
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	487.27	0.79	78.57	486.82	0.79	78.25	<b>483.48</b>	0.80	75.86
Linear	$W_{cr}$ ( $\text{kg ha}^{-1}$ )	985.11	0.57	7892.12	981.43	0.59	7628.16	<b>975.80</b>	0.67	6977.88
	$W_{st}$ ( $\text{kg ha}^{-1}$ )	1151.09	0.72	36698.23	<b>1148.39</b>	0.73	35793.09	<b>1148.39</b>	0.73	35793.09
	$W_{abg}$ ( $\text{kg ha}^{-1}$ )	1170.80	0.70	44047.91	<b>1168.19</b>	0.71	42996.57	<b>1168.19</b>	0.71	42996.57
	$G$ ( $\text{m}^2 \text{ha}^{-1}$ )	249.92	0.62	8.73	<b>246.53</b>	0.67	8.15	246.98	0.67	8.18
	$H_d$ ( $\text{m ha}^{-1}$ )	81.73	0.86	1.91	<b>81.06</b>	0.85	1.97	<b>81.06</b>	0.85	1.97
	$H_m$ ( $\text{m ha}^{-1}$ )	<b>78.34</b>	0.76	1.92	<b>78.34</b>	0.76	1.92	<b>78.34</b>	0.76	1.92
	$V$ ( $\text{m}^3 \text{ha}^{-1}$ )	503.74	0.71	91.52	<b>500.95</b>	0.72	89.18	<b>500.95</b>	0.72	89.18

## References

- Belsley, D., 1991. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.
- Chen, G., Hay, G.J., St-Onge, B., 2012. A GEOBIA framework to estimate forest parameters from lidar transects, Quickbird imagery and machine learning: a case study in Quebec, Canada. *International Journal of Applied Earth Observation and Geoinformation* 15, 28–37.
- Clark, V., Arnold, T., Agnelli, R., DeVault, C., Edwards, E.S., Gibbs, P., Hayes, D., Kiernan, K., Routten, E., Sanders, A., Savarese, P.T., Schlotzhauer, D., Stockstill, M., 2004. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc., Cary, NC.
- Côté, J.-F., Fournier, R.A., Egli, R., 2011. An architectural model of trees to estimate forest structural attributes using terrestrial LiDAR. *Environmental Modelling & Software* 26 (6), 761–777.
- Dalponete, M., Martinez, C., Rodeghiero, M., Gianelle, D., 2011. The role of ground reference data collection in the prediction of stem volume with lidar data in mountain areas. *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (6), 787–797.
- Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, 1–30.
- Diéguez-Aranda, U., Rojo-Alboreca, A., Castedo-Dorado, F., Álvarez-González, J.G., Barrio-Anta, M., Crecente-Campo, F., González-González, J.M., Pérez-Cruzado, C., Rodríguez-Soalleiro, R., López-Sánchez, C.A., Balboa-Murias, M.A., Gorgoso-Varela, J.J., Sánchez-Rodríguez, F., 2009. Herramientas selvícolas para la gestión forestal sostenible en Galicia. Xunta de Galicia.
- Dudley, R.M., 1999. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.
- Dyer, D.W., 2006. Watchmaker Framework for Evolutionary Computation. Online.
- Finley, A.O., Banerjee, S., Cook, B.D., Bradford, J.B., 2013. Hierarchical Bayesian spatial models for predicting multiple forest variables using waveform LiDAR, hyperspectral imagery, and large inventory datasets. *International Journal of Applied Earth Observation and Geoinformation* 22, 147–160.
- García, M., Riano, D., Chuvieco, E., Danson, F.M., 2010a. Estimating biomass carbon stocks for a mediterranean forest in central Spain using lidar height and intensity data. *Remote Sensing of Environment* 114 (4), 816–830.
- García, M., Riano, D., Chuvieco, E., Salas, J., Danson, F.M., 2011. Multispectral and LiDAR data fusion for fuel type mapping using support vector machine and decision rules. *Remote Sensing of Environment* 115 (6), 1369–1379.
- García, S., Fernandez, A., Luengo, J., Herrera, F., 2010b. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Information Science* 180 (10), 2044–2064.
- Gleason, C.J., Im, J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment* 125, 80–91.
- Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, Boston, MA.
- Gonçalves-Seco, L., Gonzalez-Ferreiro, E., Dieguez-Aranda, U., Fraga-Bugallo, B., Crecente, R., Miranda, D., 2011. Assessing attributes of high density *Eucalyptus globulus* stands using airborne laser scanner data. *International Journal of Remote Sensing* 32 (24), 9821–9841.
- Gonzalez-Ferreiro, E., Dieguez-Aranda, U., Miranda, D., 2012. Estimation of stand variables in *Pinus radiata* D. Don plantations using different lidar pulse densities. *Forestry* 85 (2), 281–292.
- Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (1), 56–66.
- Hawkins, D.M., 2004. The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44, 1–12.
- Hill, T., Lewicki, P., 2006. *Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining*. StatSoft.
- Höfle, B., Pfeifer, N., 2007. Correction of laser scanning intensity data: data and model-driven approaches. *ISPRS Journal of Photogrammetry and Remote Sensing* 63, 1415–1433.
- Holland, J.H., 1992. *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.
- Hosmer, D.W., Lemeshow, S., 2000. *Applied Logistic Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience Publication, Hoboken, NJ.
- Hughes, G.F., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14 (1), 55–63.
- Kitahara, F., Mizoue, N., Yoshida, S., 2010. Effects of training for inexperienced surveyors on data quality of tree diameter and height measurements. *Silva Fennica* 44 (4), 657–667.
- Kozak, A., Kozak, R., 2003. Does cross validation provide additional information in the evaluation of regression models? *Canadian Journal of Forest Research* 33, 976–987.
- Kraus, K., Pfeifer, N., 1998. Determination of terrain models in wooded areas with airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing* 53 (4), 193–203.
- Kronseider, K., Ballhorn, U., Böhm, V., Siegert, F., 2012. Above ground biomass estimation across forest types at different degradation levels in Central Kalimantan using LiDAR data. *International Journal of Applied Earth Observation and Geoinformation* 18, 37–48.
- Latifi, H., Fassnacht, F., Koch, B., 2012. Forest structure modeling with combined airborne hyperspectral and lidar data. *Remote Sensing of Environment* 121, 10–25.
- Latifi, H., Nothdurft, A., Koch, B., 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/lidar-derived predictors. *Forestry* 83 (4), 395–407.
- LeBlanc, D., 2004. *Statistics: Concepts and Applications for Science*. Jones and Bartlett.
- Levene, H., 1960. Robust tests for equality of variances. In: Olkin, I (Ed.), *Contributions to Probability and Statistics*, pp. 278–292.
- Li, J., Hu, B., Noland, T.L., 2013. Classification of tree species based on structural features derived from high density LiDAR data. *Agricultural and Forest Meteorology* 171–172, 104–114.
- Lilliefors, H., 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62, 399–402.
- Lovell, J.L., Jupp, D.L.B., Newnham, G.J., Coops, N.C., Culvenor, D.S., 2005. Simulation study for finding optimal lidar acquisition parameters for forest height retrieval. *Forest Ecology Management* 214 (1–3), 398–412.
- Luengo, J., García, S., Herrera, F., 2009. A study on the use of statistical tests for experimentation with neural networks: analysis of parametric test conditions and non-parametric tests. *Expert Systems with Applications* 36 (4), 7798–7808.
- Mallet, C., Bretar, F., Roux, M., Soergel, U., Heipke, C., 2011. Relevance assessment of full-waveform lidar data for urban area classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (6), S71–S84.

- McGaughey, R., 2009. FUSION/LDV: Software for LIDAR Data Analysis and Visualization. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Seattle.
- Millie, D.F., Weckman, G.R., Young II, W.A., Ivey, J.E., Carrick, H.J., Fahnenstiel, G.L., 2012. Modeling microalgal abundance with artificial neural networks: demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences. *Environmental Modelling & Software* 38, 27–39.
- Montaghi, A., Corona, P., Dalponte, M., Gianelle, D., Chirici, G., Olsson, H., 2013. Airborne laser scanning of forest resources: an overview of research in Italy as a commentary case study. *International Journal of Applied Earth Observation and Geoinformation* 23, 288–300.
- Myers, R.H., 1990. *Classical and Modern Regression with Applications*, second ed. Duxbury Press, Belmont.
- Peña, D., 2002. *Regresión y Diseño de experimentos*. Alianza Editoria, Madrid.
- Ramadana, Z., Songa, X.-H., Hopke, P.K., Johnson, M.J., Scow, K.M., 2001. Variable selection in classification of environmental soil samples for partial least square and neural network models. *Analytica Chimica Acta* 446, 233–244.
- Ratner, D., 2012. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, second ed. CRC Press, Boca Raton, FL.
- Renner, G., Ekárt, A., 2003. Genetic algorithms in computer aided design. *Computer-Aided Design* 35, 709–726.
- Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52 (3–4), 591–611.
- Stevens, J., 2002. *Applied Multivariate Statistics for the Social Sciences*. L. Erlbaum Associates Inc., Hillsdale, NJ.
- Song, K., Li, L., Li, S., Tedesco, L., Hall, B., Li, Z., 2012. Hyperspectral retrieval of phycocyanin in potable water sources using genetic algorithm–partial least squares (GA–PLS) modeling. *International Journal of Applied Earth Observation and Geoinformation* 18, 368–385.
- Sun, G., Ranson, K., Guo, Z., Zhang, Z., Montesano, P., Kimes, D., 2011. Forest biomass mapping from lidar and radar synergies. *Remote Sensing of Environment* 115 (11), 2906–2916.
- Tattoni, C., Rizzolli, F., Pedrini, P., 2012. Can lidar data improve bird habitat suitability models? *Ecological Modelling* 245, 103–110.
- Tesfamichael, S., Ahmed, F., van Aardt, J., 2010. Investigating the impact of discrete-return lidar point density on estimations of mean and dominant plot-level tree height in *Eucalyptus grandis* plantations. *International Journal of Remote Sensing* 31 (11), 2925–2940.
- Watson, B., 1994. *Statistics for Management and Economics: Solutions Manual*. Allyn and Bacon, Boston, MA.
- Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* 75, 1182–1189.
- Zaffalon, M., 2005. Credible classification for environmental problems. *Environmental Modelling and Software* 20 (8), 1003–1012.
- Zhao, K., Popescu, S., Meng, X., Pang, Y., Agca, M., 2011. Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment* 115 (8), 1978–1996.