Received: 31/07/2015
Accepted: 15/12/2015

**REVISTA DE TELEDETECCIÓN**
Spanish Association of Remote Sensing
(2016) 45, Special Issue, 27-40
ISSN 1133-0953
EISSN 1988-8740
http://dx.doi.org/10.4995/raet.2016.4066

# A comparative study of regression methods to predict forest structure and canopy fuel variables from LiDAR full-waveform data

Crespo-Peremarch, P.*[1,2], Ruiz, L.A.[1,2], Balaguer-Beser, A.[1,3]

[1] *Geo-Environmental Cartography and Remote Sensing Group (CGAT), Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain.*

[2] *Department of Cartographic Engineering, Geodesy and Photogrammetry, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain.*

[3] *Department of Applied Mathematics, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain.*

**Abstract:** Regression methods are widely employed in forestry to predict and map structure and canopy fuel variables. We present a study where several regression models (linear, non-linear, regression trees and ensemble) were assessed. Independent variables were calculated using metrics extracted from full-waveform LiDAR data, while the reference data used to generate the dependent variables for the prediction models were obtained from fieldwork in 78 plots of 16 m radius. Transformations of dependent and independent variables with feature selection were carried out to assess their influence in the prediction of response variables. In order to evaluate significant differences and rank regression models we used the non-parametric tests Wilcoxon and Friedman, and post-hoc analysis or post-hoc pairwise multiple comparison tests, such as Nemenyi, for Friedman test. Regressions using transformation of the dependent variable, like square-root or logarithmic, or the independent variable, increased $R^2$ up to 6% with respect to linear regression using unprocessed response variables. CART (Classification and Regression Tree) method provided poor results, but it may be interesting for categorisation purposes. Square-root transformation of the dependent variable is the method having the best overall results, except for stand volume. However, not always has a significant improvement with respect to other regression methods.

**Key words:** regression models, Random Forest, CART, M5, Wilcoxon, Friedman, forest structure, canopy fuel, LiDAR full-waveform.

## Estudio comparativo de métodos de regresión para la predicción de variables de estructura y combustibilidad a partir de datos LiDAR full-waveform

**Resumen:** Los métodos de regresión se utilizan ampliamente en el ámbito forestal para la predicción y el cartografiado de las variables de estructura y combustibilidad. En este artículo se evalúan diferentes modelos de regresión (lineal, no lineal, árboles de regresión y ensemble). Como variables independientes se utilizaron métricas extraídas de datos LiDAR full-waveform, mientras que los valores de las variables dependientes se generaron a partir de modelos basados en datos de campo obtenidos para 78 parcelas de 16 m de radio. Se llevaron a cabo transformaciones de las variables dependientes e independientes con selección de atributos para evaluar su influencia en la predicción de la variable respuesta. Con el fin de verificar diferencias significativas y ordenar los modelos de regresión se emplearon los tests no paramétricos de Wilcoxon y Friedman, y el análisis post-hoc o los tests de comparación post-hoc por pares, como el de Nemenyi, para el test de Friedman. Las regresiones basadas en la transformación de la variable dependiente,

* *Corresponding author: pabcrepe@cgf.upv.es*

como raíz cuadrada o logaritmo, o en la transformación de las variables independientes, obtuvieron un incremento de la $R^2$ de hasta un 6% con respecto a la regresión lineal. Mediante el método CART (*Classification and Regression Tree*) se obtuvieron resultados discretos, si bien su uso puede estar indicado para la categorización o estratificación. Con el método basado en la transformación de la variable dependiente mediante raíz cuadrada se consiguieron los mejores resultados comparativos en la predicción de variables forestales, excepto para el volumen. Sin embargo, su uso no siempre implica una mejora significativa con respecto a los otros métodos de regresión usados en este trabajo.

**Palabras clave:** modelos de regresión, *Random Forest*, CART, M5, Wilcoxon, Friedman, estructura forestal, combustibilidad, LiDAR full-waveform.

## 1. Introduction

Regression analysis is a statistical process that tries to estimate a dependent variable from one or more independent variables. Regressions are widely used to estimate forest variables. Standard forest biometric measurements from some trees were used to generate allometric equations to predict forest variables, such as aboveground biomass and other structure and canopy fuel variables across large areas (Skowronski *et al.,* 2011).

Regressions have also been employed for variable estimation using metrics extracted from LiDAR data (e.g. Means *et al.*, 2000; Hermosilla *et al.,* 2014). Previous studies have shown the very strong correlations between metrics extracted from LiDAR and data from forest biometric plots, what suggests that regression models can be calculated to estimate forest inventory parameters, and thus generate landscape-scale maps of these variables using a small number of field plots (Andersen *et al.,* 2005; Andersen and Breidenbach, 2007; Skowronski *et al.,* 2011). Airborne LiDAR collects a complete description of the forest vertical structure allowing laser pulses to penetrate through the canopy (Erdody and Moskal, 2010). Discrete LiDAR has however restrictions to get the different vegetation layers. In the last 20 years, several studies have shown that full-waveform LiDAR systems, that register the full wave that interacts with the canopy, let a better description of the physical and forest vertical structure properties, and therefore achieve good estimations of canopy fuel metrics (Lefsky *et al.,* 1999; Means *et al.,* 2000; Hermosilla *et al.,* 2014).

In forestry applications, a series of regression models have been implemented. The simplest but getting good fitting is the linear regression. Logarithmic and square-root transformations of the dependent variable are very widespread in studies related to canopy fuel parameters estimation (equations 1 and 2).

$$\ln(Y) = a_0 + a_1 \times X_1 + a_2 \times X_2 + \cdots + a_n \times X_n \quad (1)$$

$$\sqrt{Y} = a_0 + a_1 \times X_1 + a_2 \times X_2 + \cdots + a_n \times X_n \quad (2)$$

In Means *et al.* (2000) a logarithmic transformation of the dependent variable is used, whereas in Andersen *et al.* (2005) and in Erdody and Moskal (2010) both transformations are employed to increase accuracy.

These transformations of the dependent variable improve estimations because residuals meet regression models hypothesis: normally distributed, independent, homoscedastic (constant variance) and linear (Hannon and Knapp, 2003; Wang and Zhou, 2005). In these cases, since the dependent variable is modified and all the statistic results (coefficient of determination, root-mean-square error, etc.) are not calculated in the original, but in the transformed space, it is necessary to invert them to express the results in the units of the original space (Andersen *et al.,* 2005). There are some other non-linear methods used in ecological application, such as exponential models (Temesgen *et al.,* 2015), Random Forest (Baccini *et al.,* 2008) and regression trees (García-Gutiérrez *et al.,* 2011). These methods try to create groups or leaves of data keeping the homogeneity (De'Ath and Fabricius, 2013), and the regression trees are straightforward and easy to understand.

There exist a large number of regression methods in the literature, and no one is the most appropriate for all the cases, depending this choice on the particular application and data, this is why

several comparative studies can be found in the literature. Naesset *et al.* (2005) compared three estimation techniques (ordinary least-squares, seemingly unrelated regression and partial least-squares regression), revealing that none of these techniques were superior to the others in predicting biophysical properties of forest stands over all combinations of strata and variables. In García-Gutiérrez *et al.* (2011) a comparative study between multiple linear regression (MLR) and regression trees (M5P) was done to predict crown, stem and aboveground biomass using LiDAR data, concluding that M5P outperforms MLR. In Marabel-García and Álvarez-Taboada (2014) a comparison between two different methods (Partial Least Square Regression –PLSR– and linear regression) were applied in order to find the best estimate of the aboveground biomass, recommending the use of linear regression due to its simplicity with respect to PLSR. In Li *et al.* (2015) several regression models (linear, exponential growth, support vector regression and neural network) were compared for aboveground biomass estimation, concluding that linear regression achieved the best performance. In other cases (Hyyppä *et al.,* 2000; Erdody and Moskal, 2010), a comparison of regression methods using different data sources (imagery, LiDAR or imagery + LiDAR) is performed.

In most of the abovementioned studies, the comparison is done only by analysing results attending to the coefficient of determination, the RMSE, etc. but in García-Gutiérrez *et al.* (2011) Wilcoxon test is employed. However, usually a significance test is needed in order to know if those performance results make a real difference in practical results or not. Therefore, having *n* data sets we need to test which model would perform the best, and if it is statistically different from others (Luengo *et al.,* 2012). Some parametric and non-parametric tests are available for this purpose. Among non-parametric tests, that are preferred over parametric ones (Demšar, 2006), we can find Wilcoxon and Friedman tests. The Wilcoxon signed rank test compares two methods, whereas the Friedman test compares all the methods between them, being needed a post-hoc test after rejecting null-hypothesis to make a pairwise comparison (Demšar, 2006).

Firstly, this paper aims to study several regression models to estimate structure and canopy fuel variables, and looking for improving the prediction of these variables applying transformations of dependent and independent variables, and carrying out feature selections. Secondly, this study compares and analyse the suitability of the regression methods tested to obtain prediction models, using non-parametric tests such as Wilcoxon and Friedman. Finally, some conclusions about methods are mentioned, such as the importance of using an evaluation set especially for Random Forest algorithm, and the usefulness of regression trees for data stratification.

## 2. Study area and data

The study area is located in Panther Creek, in the state of Oregon (USA) (see Figure 1). Elevation ranges from 100 to 700 m (see shaded relief in Figure 1). The dominant species is Douglas-fir (*Pseudotsuga menziesii*), being present in more than half of the total forested area. Occasionally this species is mixed with other conifers. The height of the trees of the study area is sometimes higher than 60 m, although it is very variable due to timber production in the zone.

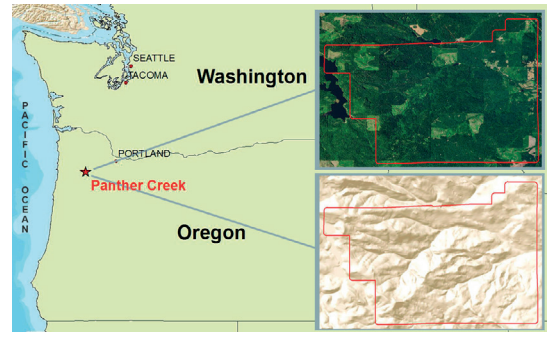


**Figure 1.** Location (left side), orthoimage (top right) and shaded relief (bottom right) of the study area.

Full-waveform data were collected in July 15th, 2010, by Watershed Sciences, Inc. using a Leica ALS60 sensor incorporated in a Cessna Caravan 208B. The system acquired data at a 105 kHz pulse rate, flown at an average altitude of 900 m above ground level, and a scanning angle of ±15° from nadir. The waveforms were recorded in 256

bins with a temporal sample spacing of 2 ns and beam footprint size of ~0.25 m, what yielded a pulse density of ≥8 points/m$^2$. The study area was surveyed with opposing flight line side-lap ≥50% (≥100% total overlap). Aircraft position was recorded with a frequency of 2 Hz by on-board differential GPS unit, altitude was acquired with a frequency of 200 Hz as pitch, roll and yaw from on-board IMU. LiDAR data were provided in LAS 1.3 format. Moreover, the company provided a digital terrain model (DTM) generated with the last return pulses. After DTM evaluation using 33 ground control points measured using RTK-GPS, the root-mean-square error was 0.19 m.

Field data were acquired in 78 circular plots (47 Douglas-fir and 31 mixed species) with 16 m radius. Plot positions were located with accuracy lower than 0.3 m in horizontal and vertical locations using Trimble R-8 GNSS and Leica TPS 800 total stations. All trees within plots having a diameter at breast height (DBH) greater or equal than 2.5 cm were tagged.

Using field measurements (tree height, DBH, species, and standard measurements) species-specific allometric equations derived by Standish *et al.* (1985) were calculated so as to estimate structure and canopy fuel variables for each plot (Table 1). These values were used as dependent variables to be predicted for regression models from LiDAR full-waveform metrics.

LiDAR full-waveform metrics were extracted following those proposed by Duong (2010) and further described by Cao *et al.* (2014), and used as independent variables in regression models (Table 2).

## 3. Methods

### 3.1. Feature selection

Three different feature selection processes were followed (Figure 2) in order to reduce the number of variables: one for regression trees, a second one for Random Forest regression and a third one for multiple regression methods.

In the case of regression trees we chose the three variables that were closer to the root, because when a variable is closer to the root it has more relevance in the overall estimation of the data. Regarding Random Forest algorithm, a relevance ranking is created based on the number of times that a variable appears in the nodes of all the trees generated with the algorithm. Variables with higher values are more used and therefore perform better. In all these cases a maximum of three variables was included in the models in order to avoid overfitting and to create more robust models (Hermosilla *et al.,* 2014).

A large variety of variable selection methods can be used, each of them motivated by various

**Table 1.** Structure and canopy fuel variables (dependent variables) summary of study area plots.

| Code (unit) | Variable | Description | Mean | s.d. |
|---|---|---|---|---|
| AGB (t·ha$^{-1}$) | Aboveground biomass | Weight of all the living biomass aboveground per unit of area | 309.57 | 202.23 |
| BA (m$^2$·ha$^{-1}$) | Basal area | Area occupied by tree trunks per unit of area | 46.14 | 22.62 |
| V (m$^3$) | Volume | Volume of canopy | 362.89 | 174.39 |
| CBD (kg·m$^{-3}$) | Canopy bulk density | Ratio between canopy fuel load and canopy depth | 0.136 | 0.086 |
| CFL (t·ha$^{-1}$) | Canopy fuel load | Total amount of biomass in the canopy fuel layer per unit of area | 48.49 | 23.51 |

**Table 2.** Summary of metrics extracted from LiDAR full-waveform (independent variables).

| Code | Variable | Description |
|---|---|---|
| HOME | Height of median energy | Height from ground to the waveform centroid |
| WD | Waveform distance | Height from ground to waveform beginning |
| NP | Number of peaks | Number of waveform peaks |
| ROUGH | Roughness of outermost canopy | Distance from beginning to the first peak |
| HTMR | Height/median ratio | Ratio between HOME and WD |
| VDR | Vertical distribution ratio | Difference between WD and HOME divided by WD |
| RWE | Return waveform energy | Area below waveform from the beginning to the ground |
| FS | Front slope angle | Vertical angle from beginning to the first peak |

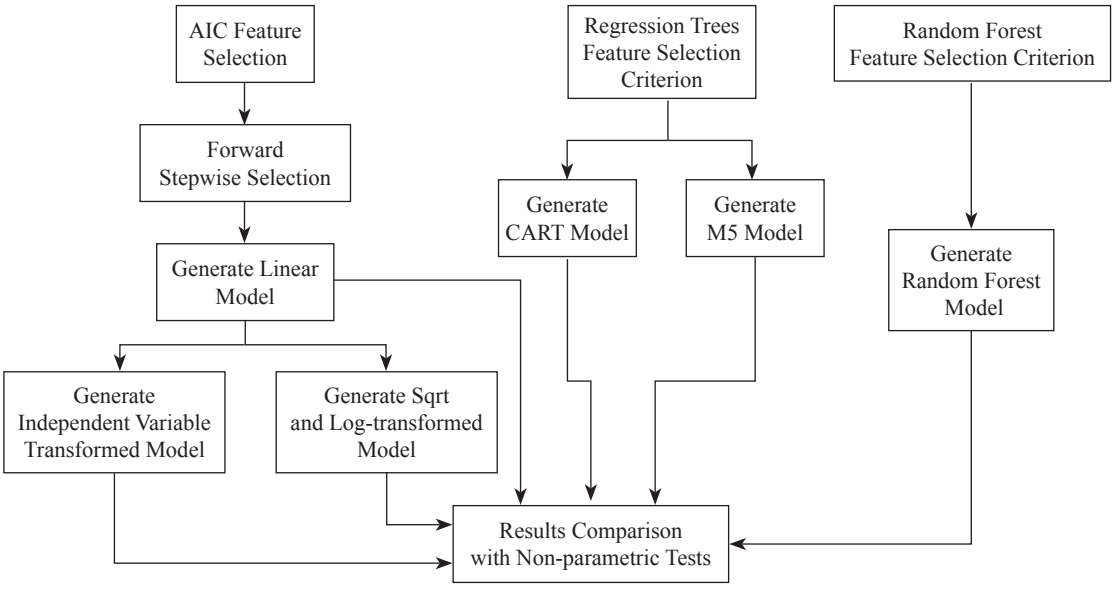


**Figure 2.** Diagram with the overall process followed for the comparison of regression methods.

theoretical arguments, but a unifying theoretical framework is lacking (Guyon and Elisseeff, 2003). Genetic approaches have made, for instance, a good variable selection in studies related to LiDAR for estimation of forest stand variables (García-Gutiérrez *et al.,* 2013). In our case, for multiple regression methods two criteria, the Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) were initially tested. These methods are robust in selecting variables, and very close results were reached with our data set and variables in the preliminary tests. Furthermore, using BIC method, a maximum of three variables was always obtained. Based on these results and following a methodological coherence with the Random Forest and regression trees algorithms, the models were restricted to a maximum of three variables. Since AIC minimise the risk function in finite sample sizes and when the true model is not among the candidate models or it is extremely complex, in the case of multiple regression methods we used the Akaike Information Criterion.

AIC is a relative estimator that represents the loss of information when a model is used to approximate a second one. It does not provide information about the quality of the model in absolute sense, but it is used for selecting the optimal number of variables. AIC value is calculated as follows:

$$AIC = -2l + 2K$$

where l is the maximised log-likelihood and K the number of parameters. This value is calculated for each model, selecting the one with the minimum AIC value, since it loses less information (Posada and Buckley, 2004).

In those cases that after applying AIC more than three variables still remained, we used forward stepwise selection using leave-one-out cross-validation from the linear regression generated by the AIC until having a maximum of three variables.

## 3.2. Regression models

Once three or less variables were selected, we generated and compared seven different regression models: Linear, Ind-trans, Log-trans, Sqrt-trans, CART, M5 and Random Forest. In this section these models are briefly described.

Linear model (Linear): Model obtained after linear fitting using the variables extracted with the forward stepwise selection.

Independent variable transformation (Ind-trans): Based on the initial linear regression, each independent variable is transformed (logarithmic, exponential, quadratic, cubic, inverse and S

curve) with the goal to increase the adjusted $R^2$. In order to know which transformation of each independent variable to employ, the linear correlation coefficient between the transformed and the dependent variable were compared, then a backward stepwise selection was carried out for these new transformed variables, verifying that the hypothesis of the multiple regression model was met.

Dependent variable transformation (Log-trans and Sqrt-trans): Based on the transformation of the dependent variable using logarithmic and square-root transformations, and following similar process as for the independent variable transformation. This transformation aims to obtain normal distribution, independence, homoscedasticity and linearity of residuals.

Regression trees: Classification and Regression Tree (CART) (Breiman *et al.,* 1984): This method can generate classification or regression trees, depending on the type of dependent variable. Rules are selected in order to better differentiate the data based on the variable to be estimated. This process can be repeated for each branch of the tree, generating new nodes, until data from the same leaf are homogeneous, and therefore they cannot be further split. For regressions, a value is set to each leaf (see Figure 3).

M5 (Quinlan, 1992) is also a tree-based method, but instead of having estimation values at their leaves, they have linear regressions, so that all the instances belonging to one leaf can have different values. M5 constructs a piecewise constant tree and then generates a linear regression to the data of each leaf (see Figure 4).

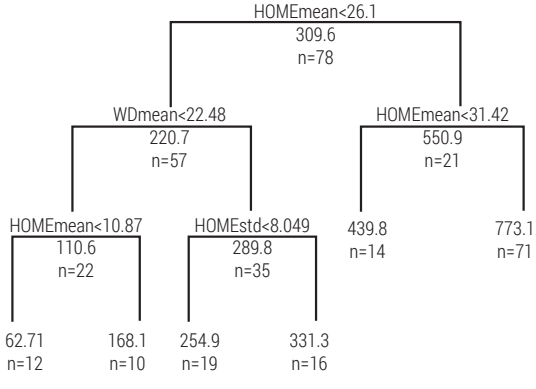


**Figure 3.** CART regression tree built for AGB variable, where in each node is showed the condition of the left branch, the estimated value and the number of samples of that node, respectively.

```
Model:

Rule 1: [69 cases, mean 261.46231, range 30.76877
to 597.8537, est err 50.96885]

if

WDmean <= 38.50956

then

outcome = 191.56349 + 24.7 HOMEmean - 568 HTMRmean

Rule 2: [9 cases, mean 678.41150, range 321.5639
to 924.2084, est err 95.07365]

if

WDmean > 38.50956

then

outcome = -1905.33629 + 58.9 WDmean + 2.9 HOMEmean
- 75 HTMRmean
```

**Figure 4.** Example of M5 tree-based rules for AGB variable in mixed forest.

Ensemble: Random Forest (Breiman, 2001) is an ensemble learning method for classification or regression. This algorithm is based on creating different subset of instances and variables randomly in order to generate a number of trees. For regressions, the estimated value is the average of estimated values for the different trees. This leads to have different models despite having the same data.

## 3.3. Evaluation and comparison methods

As mentioned above, leave-one-out cross-validation was used to select variables (using forward stepwise selection) and to compare the regression models. This split is required for the purpose that model accuracy validation must be an independent measure, and not to produce overfitting (Hernández-Orallo *et al.,* 2004). Regression models were evaluated using adjusted coefficient of determination ($R^2$), root-mean-square error (RMSE), normalised root-mean-square error (nRMSE) defined as the ratio of RMSE and the range of observed values, and coefficient of

variation (CV), that is the RMSE divided by the mean of observed values.

After generating the seven different regression models and in order to analyse statistical differences between models, we used two non-parametric tests: Wilcoxon and Friedman (see Figure 2). The aim of applying these tests was to establish a rank of the methods and to demonstrate if their performance differences are statistically significant. Since these tests need the algorithm results organised for each data set or fold, it is required to use cross-validation or different data sets. We used a leave-one-out cross-validation, so that each algorithm had 78 values for mixed species and 47 for Douglas-fir. The evaluation measure used was the residual error of each plot obtained by cross-validation.

The Wilcoxon signed-ranks test (Wilcoxon, 1945) compares two methods to show if the differences between them are significant or not. This test ranks absolute value of differences between results of methods for each data set, and compares the ranks for the positive and negative differences (Demšar, 2006). After performing a pairwise comparison and testing the differences, a ranking method was applied according to the total wins and no defeats (wins and ties) (Luengo *et al.,* 2012). The method with best results was the one that had more wins, and if two or more methods had the same number of wins, the one with a better rank was the method that had more wins and ties. If there was still a tie, an average rank was assigned to the tied methods.

In contrast to the Wilcoxon test, the Friedman test (Friedman, 1937; Friedman, 1940) is able to compare more than two methods. It consists on ranking methods for each data set separately, then considering the values of ranks by columns (Demšar, 2006), showing the general ranking and the statistical significance of the differences. In order to make a pairwise comparison about the significance of the differences between the methods, a post-hoc test like Nemenyi (Nemenyi, 1963) or *p-value* of Friedman test analysis was applied. These post-hoc tests can only be used when there is a significant difference considering all methods, i.e. after having rejected the null-hypothesis of the Friedman test.

Nemenyi test is equivalent to the Tukey test for ANOVA. According to this test, two methods are considered significantly different when their corresponding average ranks differ by at least the critical difference calculated with critical values based on the Studentized range statistic, number of data sets, and number of models (Demšar, 2006). There is also a post-hoc comparison for Friedman test (Zar, 1999) that calculates a z score using the rank difference between the two methods to be compared, number of data sets and number of total methods. Then, the *p-value* calculated for the z score must be greater than 0.05 (95% of confidence level) to consider that the performance of these two methods is significantly different.

## 4. Results and Discussion

For every regression model and dependent variables tested, a set of independent variables was selected. Table 3 shows the number of times each variable was selected and included in a prediction model. These variables represent the average (μ) and the standard deviation (σ) per plot of the different metrics displayed on Table 2. HOMEμ was used in almost all the regressions, whereas NPμ, NPσ, ROUGHσ, HTMRσ, RWEμ and FSμ were never selected.

**Table 3.** Percentage of use of the independent variables after feature selection.

| Variable | Use |
|---|---|
| HOMEμ | 97% |
| WDμ | 34% |
| HOMEσ | 22% |
| HTMRμ | 15% |
| VDR | 14% |
| WDσ | 8% |
| RWEσ | 6% |
| ROUGHμ | 2% |
| FSσ | 2% |
| NPμ | 0% |
| NPσ | 0% |
| ROUGHσ | 0% |
| HTMRσ | 0% |
| RWEμ | 0% |
| FSμ | 0% |

After the evaluation of the regression models using cross-validation for two different strata, mixed forest and only Douglas-fir, the results shown in Table 4 were obtained.

**Table 4.** Results of the different regression models for each estimated variable in stratum mixed forest.

| Variables | Statistical indicators | Linear | Ind-trans | Log-trans | Sqrt-trans | CART | M5 | Random Forest |
|---|---|---|---|---|---|---|---|---|
| AGB | $R^2$ | 0.84 | 0.88 | 0.88 | 0.88 | 0.73 | 0.85 | 0.86 |
| | RMSE ($t \cdot ha^{-1}$) | 78.80 | 67.54 | 70.88 | 67.07 | 101.76 | 78.30 | 74.36 |
| | nRMSE | 0.09 | 0.08 | 0.08 | 0.08 | 0.11 | 0.09 | 0.08 |
| | CV | 0.26 | 0.22 | 0.23 | 0.22 | 0.33 | 0.25 | 0.24 |
| BA | $R^2$ | 0.74 | 0.74 | 0.74 | 0.75 | 0.60 | 0.74 | 0.66 |
| | RMSE ($m^2 \cdot ha^{-1}$) | 11.36 | 11.36 | 11.36 | 11.14 | 14.02 | 11.35 | 12.94 |
| | nRMSE | 0.11 | 0.11 | 0.11 | 0.11 | 0.14 | 0.11 | 0.13 |
| | CV | 0.25 | 0.25 | 0.25 | 0.24 | 0.30 | 0.25 | 0.28 |
| V | $R^2$ | 0.72 | 0.72 | 0.72 | 0.71 | 0.55 | 0.69 | 0.61 |
| | RMSE ($m^3$) | 89.66 | 90.69 | 89.66 | 92.54 | 114.95 | 95.52 | 106.33 |
| | nRMSE | 0.11 | 0.11 | 0.11 | 0.11 | 0.14 | 0.12 | 0.13 |
| | CV | 0.25 | 0.25 | 0.25 | 0.26 | 0.32 | 0.26 | 0.29 |
| CBD | $R^2$ | 0.65 | 0.68 | 0.66 | 0.68 | 0.47 | 0.64 | 0.55 |
| | RMSE ($kg \cdot m^{-3}$) | 0.05 | 0.05 | 0.05 | 0.05 | 0.06 | 0.05 | 0.06 |
| | nRMSE | 0.13 | 0.12 | 0.13 | 0.12 | 0.16 | 0.13 | 0.14 |
| | CV | 0.36 | 0.35 | 0.37 | 0.35 | 0.46 | 0.38 | 0.41 |
| CFL | $R^2$ | 0.76 | 0.76 | 0.76 | 0.76 | 0.61 | 0.77 | 0.70 |
| | RMSE ($t \cdot ha^{-1}$) | 11.30 | 11.30 | 11.30 | 11.14 | 14.50 | 11.28 | 12.61 |
| | nRMSE | 0.11 | 0.11 | 0.11 | 0.11 | 0.14 | 0.11 | 0.12 |
| | CV | 0.23 | 0.23 | 0.23 | 0.23 | 0.30 | 0.23 | 0.26 |

As shown in Table 4, AGB has the best performance for mixed forest plots ($R^2$ = 0.88, nRMSE = 0.08, CV = 0.22), showing an important difference with respect to the rest of variables. Figure 5b and 5c show that AGB has the lowest values for nRMSE, while for other variables the values are similar. In contrast, CBD results are the poorest ($R^2$ = 0.68, nRMSE = 0.12, CV = 0.35) (see Table 4), showing dissimilarity with respect to other variables (see Figure 5a), and meaning that the accuracy of this variable cannot be improved as well in Douglas-fir stratum as in the case of other dependent variables.

Regarding Douglas-fir stratum, in Table 5 we observe that AGB has also the best accuracy ($R^2$ = 0.89, nRMSE = 0.09), but CV value (CV = 0.23) is not better than CV for BA (CV = 0.22), V (CV = 0.21) and CFL (CV = 0.22). CBD has again the poorest values for Douglas-fir plots ($R^2$ = 0.68, nRMSE = 0.13, CV = 0.37).

The value of $R^2$ is sometimes higher in Douglas-fir stratum than in mixed forest (see Table 4 and 5), or it has similar value, as in CBD. However, this behaviour is not reflected in RMSE, nRMSE and

**Table 5.** Statistic results of the different regression models for each estimated variable for stratum Douglas-fir.

| Variables | Statistical indicators | Linear | Ind-trans | Log-trans | Sqrt-trans | CART | M5 | Random Forest |
|---|---|---|---|---|---|---|---|---|
| AGB | $R^2$ | 0.83 | 0.88 | 0.86 | 0.89 | 0.69 | 0.83 | 0.83 |
| | RMSE ($t \cdot ha^{-1}$) | 91.26 | 77.11 | 86.15 | 76.37 | 124.72 | 98.13 | 93.42 |
| | nRMSE | 0.10 | 0.09 | 0.10 | 0.09 | 0.14 | 0.11 | 0.10 |
| | CV | 0.27 | 0.23 | 0.26 | 0.23 | 0.37 | 0.29 | 0.28 |
| BA | $R^2$ | 0.78 | 0.78 | 0.78 | 0.79 | 0.62 | 0.77 | 0.74 |
| | RMSE ($m^2 \cdot ha^{-1}$) | 11.37 | 11.37 | 11.37 | 10.87 | 14.77 | 11.67 | 12.17 |
| | nRMSE | 0.11 | 0.11 | 0.11 | 0.11 | 0.15 | 0.12 | 0.12 |
| | CV | 0.23 | 0.23 | 0.23 | 0.22 | 0.30 | 0.24 | 0.25 |
| V | $R^2$ | 0.80 | 0.80 | 0.72 | 0.76 | 0.72 | 0.74 | 0.79 |
| | RMSE ($m^3$) | 82.82 | 82.82 | 99.07 | 91.19 | 98.84 | 97.09 | 85.50 |
| | nRMSE | 0.10 | 0.10 | 0.12 | 0.11 | 0.12 | 0.12 | 0.10 |
| | CV | 0.21 | 0.21 | 0.25 | 0.23 | 0.25 | 0.24 | 0.21 |
| CBD | $R^2$ | 0.66 | 0.66 | 0.68 | 0.68 | 0.42 | 0.57 | 0.51 |
| | RMSE ($kg \cdot m^{-3}$) | 0.05 | 0.05 | 0.05 | 0.05 | 0.07 | 0.06 | 0.06 |
| | nRMSE | 0.13 | 0.13 | 0.13 | 0.13 | 0.18 | 0.15 | 0.16 |
| | CV | 0.38 | 0.37 | 0.37 | 0.37 | 0.50 | 0.43 | 0.45 |
| CFL | $R^2$ | 0.78 | 0.78 | 0.78 | 0.79 | 0.62 | 0.78 | 0.75 |
| | RMSE ($t \cdot ha^{-1}$) | 11.67 | 11.67 | 11.67 | 11.14 | 15.38 | 11.99 | 12.29 |
| | nRMSE | 0.11 | 0.11 | 0.11 | 0.11 | 0.15 | 0.11 | 0.12 |
| | CV | 0.23 | 0.23 | 0.23 | 0.22 | 0.30 | 0.24 | 0.24 |

CV values and, in some cases, (e.g. AGB, CBD and CFL) results are not improved. For all the variables but V, using the square-root transformation the best results are achieved (see Table 4, 5 and Figure 5), whereas using regression tree CART the lowest. Taking linear regression as reference, square-root transformation can improve $R^2$ up to 6% (AGB in Douglas-fir stratum) (Figure 5a).

Tables 4, 5 and Figure 5a show that using regression methods such as independent variable transformation, logarithmic-transformed, square-root-transformed, M5 and Random Forest, the $R^2$ values for variables AGB, BA, CBD and CFL improve with respect to the linear regression method, whereas for estimation of V better results are achieved using linear regression. These differences

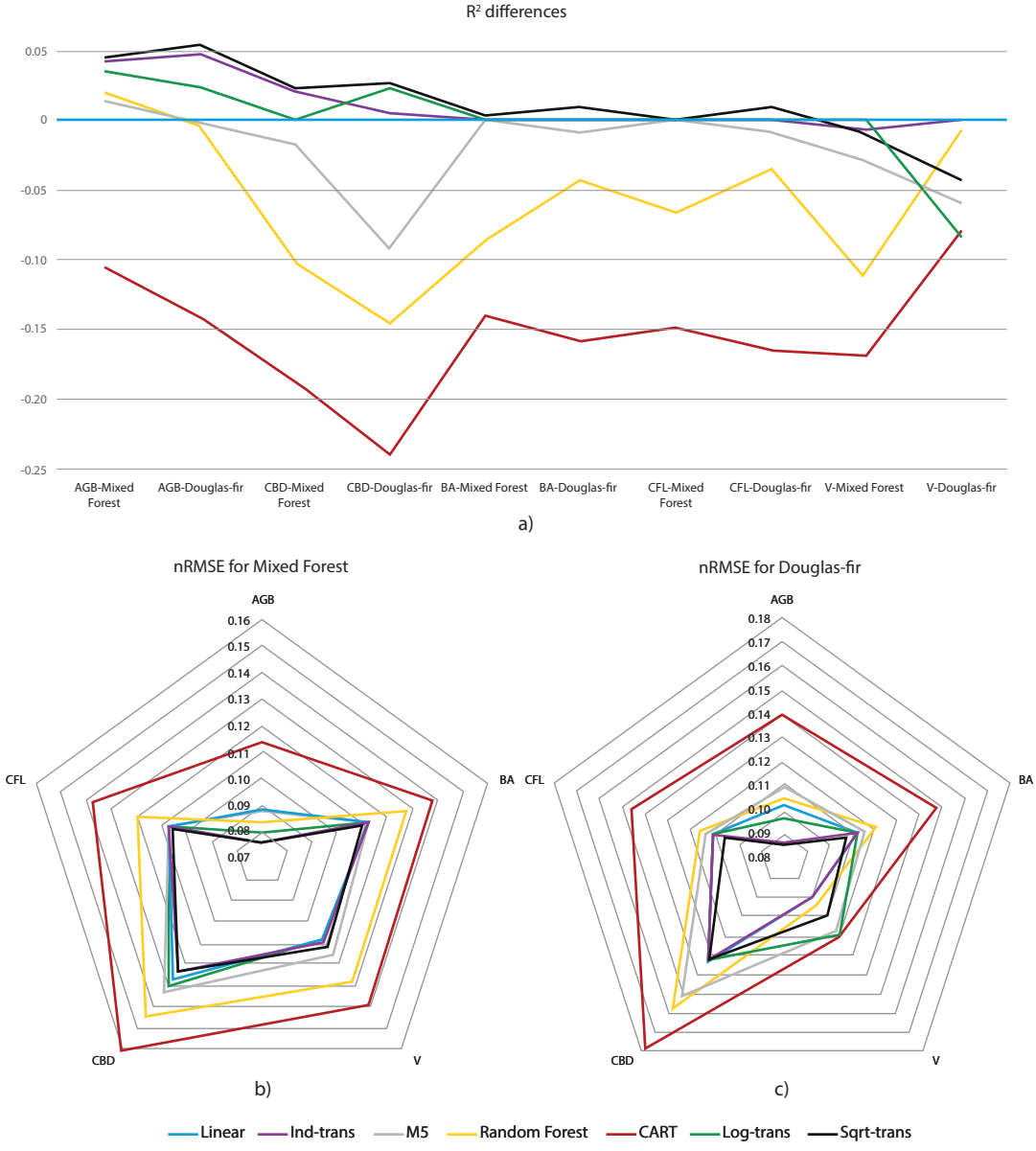


**Figure 5.** (a) Graph displaying differences in $R^2$ with respect to linear regression: positive values when the regression model outperforms the linear regression and negative values when the linear regression performs better; radial graph of different regression models for each variable estimation displaying nRMSE (b) for mixed forest plots and (c) for Douglas-fir stratum.

| Variable | Data | Test | Ranking | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| AGB | MF | F+p | Sqrt | I-trans | Ln | RF | Lin | M5 | CART |
| | | F+N | Sqrt | I-trans | Ln | RF | Lin | M5 | CART |
| | | W | Sqrt | I-trans | Ln | Lin | RF | M5 | CART |
| | DF | F+p | Sqrt | I-trans | Ln | RF | Lin | M5 | CART |
| | | F+N | Sqrt | I-trans | Ln | RF | Lin | M5 | CART |
| | | W | Sqrt | I-trans | Ln | RF | Lin | M5 | CART |
| BA | MF | F+p | Sqrt | Lin | M5 | RF | CART | | |
| | | F+N | Sqrt | Lin | M5 | RF | CART | | |
| | | W | Sqrt | Lin | RF | M5 | CART | | |
| | DF | F+p | Sqrt | Lin | M5 | RF | CART | | |
| | | F+N | Sqrt | Lin | M5 | RF | CART | | |
| | | W | Sqrt | Lin | M5 | RF | CART | | |
| V | MF | F+p | Lin | M5 | I-trans | Sqrt | RF | CART | |
| | | F+N | Lin | M5 | I-trans | Sqrt | RF | CART | |
| | | W | Lin | M5 | I-trans | Sqrt | RF | CART | |
| | DF | F+p | Ln | Sqrt | Lin | M5 | CART | RF | |
| | | F+N | Ln | Sqrt | Lin | M5 | CART | RF | |
| | | W | Ln | Sqrt | Lin | M5 | CART | RF | |
| CBD | MF | F+p | Sqrt | Lin | I-trans | Ln | RF | M5 | CART |
| | | F+N | Sqrt | Lin | I-trans | Ln | RF | M5 | CART |
| | | W | Sqrt | Lin | I-trans | Ln | RF | M5 | CART |
| | DF | F+p | Sqrt | Lin | Ln | I-trans | RF | CART | M5 |
| | | F+N | Sqrt | Lin | Ln | I-trans | RF | CART | M5 |
| | | W | Sqrt | Lin | Ln | I-trans | RF | CART | M5 |
| CFL | MF | F+p | Sqrt | M5 | Lin | RF | CART | | |
| | | F+N | Sqrt | M5 | Lin | RF | CART | | |
| | | W | Sqrt | Lin | M5 | RF | CART | | |
| | DF | F+p | Sqrt | RF | Lin | M5 | CART | | |
| | | F+N | Sqrt | RF | Lin | M5 | CART | | |
| | | W | Sqrt | Lin | RF | M5 | CART | | |

**Figure 6.** Regression models (Lin: linear, I-trans: independent variable transformation, Sqrt: square-root-transformed, Ln: logarithmic-transformed, CART, M5 and RF: Random Forest) ranking after applying Wilcoxon (W) and Friedman (F) tests and post-hoc *p-value* (p) and Nemenyi (N) tests using mixed forest (MF) and Douglas-fir (DF) for the different predicted variables. Regression no significantly different regarding the method placed in rank 1 is green-coloured.

are only in the range of 1% to 6% (Figure 5a) and, in some cases, they are not significant, as we can see in Figure 6, as described below.

In order to better interpret the differences between methods, regression model residuals of the predicted variables for the two strata types are represented using boxplots in Figure 7.

Analysing this boxplot graph (Figure 7), CART boxes are different from the other methods tested, being the residuals and the range higher. Independent variable transformation and square-root-transformed have smaller boxes (lower standard deviation range) than the rest. As observed in Tables 4 and 5, results for V variable were very similar, and the square-root-transformed was not the most accurate model.

In order to confirm that these differences between regression models were significant, non-parametric tests were used. Ranking and proof of significant differences between regression models using non-parametric and post-hoc tests, using the two plot types and for all the predicted variables tested, are displayed in Figure 6.

As shown in Figure 6, the ranking of regression methods for the different variables is very similar either using the Friedman or the Wilcoxon test. However, after applying Nemenyi post-hoc test, *p-values* of Friedman or Wilcoxon test to verify if the difference between methods is statistically

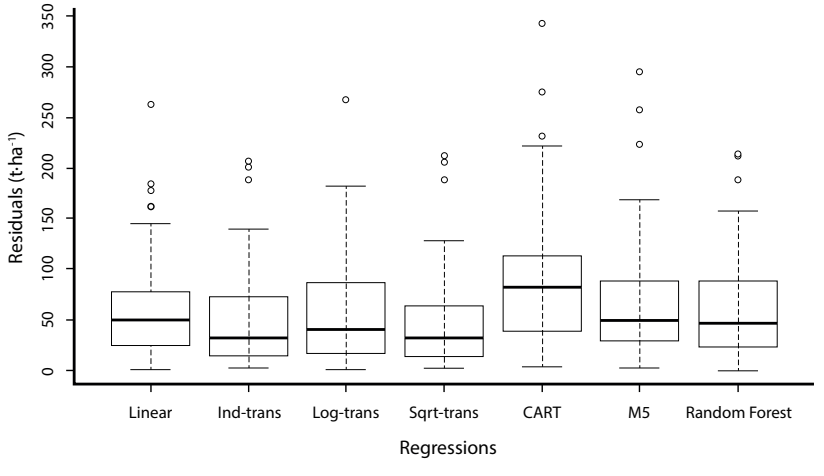


**Figure 7.** Boxplot of regression models residuals for AGB (t·ha$^{-1}$) in Douglas-fir stratum.

significant, then some differences are noticed due to the differences of test criteria.

For all the dependent variables, except V, square-root transformed method is always the best in the ranking, although in most of the cases the difference with the next method is not significant. In these cases, the difference could be due to randomness (Luengo *et al.,* 2012).

Results of Figure 6 are graphically confirmed in Figure 7. In Figure 6, the results related to the existence of significant differences obtained from Friedman+Nemenyi with respect to those obtained from Friedman+*p-value* or Wilcoxon tests are different. While using the first there is not dissimilarity between square-root transformation, independent variable transformation, logarithmic transformation, linear and M5 regressions, using the latter tests there is no significant difference only between square-root and independent variable transformation methods. In Figure 6, Friedman+*p-value* and Wilcoxon test results show that square-root transformation of the dependent variable performs better than the rest of models. As previously mentioned, non-parametric tests confirmed that regression models results for variable V in mixed forest are very similar and there is not a significant difference between them, except for CART.

As mentioned above, CART, which is a regression tree method (see Figure 3), does not achieve a high performance. When a decision tree is used for regressions, an average value is assigned to each leaf of the tree, so all the instances meeting the conditions of that leaf will have the same value, transforming a continuous variable to a discrete one. However, the generation of a regression tree can be sometimes interesting, because it is very easy to understand by no experts and, since each leaf is represented by homogenous data, it is possible to differentiate several strata using the variable to estimate. This stratification could be done depending on the number of leaves, and each stratum would be represented by the value of the leaf.

M5 tree-based method improves CART algorithm, but it does not achieve high performance results. This method generates as many linear regressions as rules (see Figure 4), as opposite to CART that just gets a unique predicted value for each leaf, being able to have different estimated values for a same leaf. M5 does not have a better accuracy, this can be due to the fact that the node condition to split the data is not doing a good data discrimination.

It is known that Random Forest overfits for some noisy data (Segal, 2004). As we can see in Figure 8, there is an important difference between results obtained with Random Forest by using or not using cross-validation for evaluation. This difference can vary between 10% and 37% for Random Forest algorithm, while for the rest of the

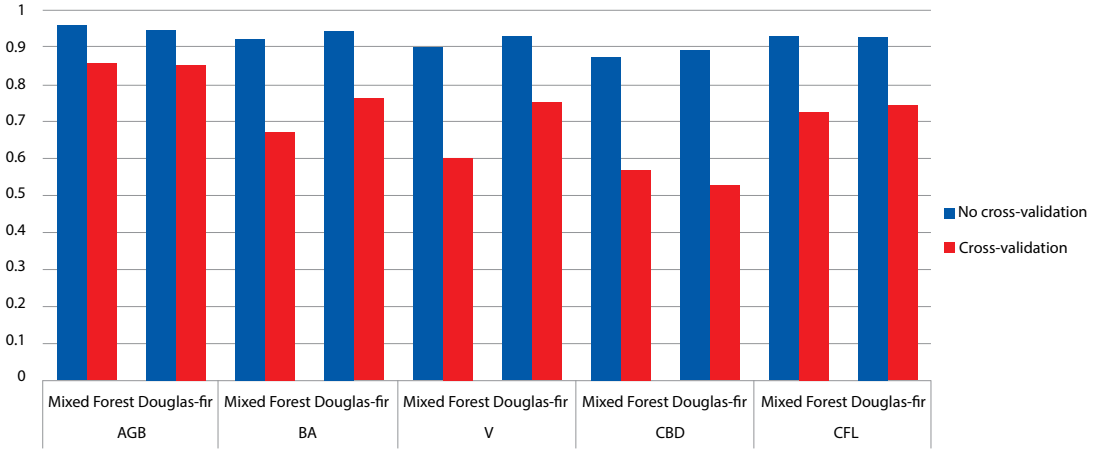


**Figure 8.** Differences between with cross-validation and without cross-validation in $R^2$ for Random Forest regression.

methods is only about 2%. Hence cross-validation or, in general, independent evaluation is crucial, and especially relevant when using Random Forest, due to overestimation.

## 5. Conclusions

In this paper we analysed seven multiple regression methods (linear, independent variable transformation, logarithmic-transformed, square-root-transformed, CART, M5 and Random Forest) to assess their potential to predict forest structure and canopy fuel variables.

Sometimes a linear regression can be improved by transforming the dependent variable with a logarithmic, a square-root transformation or transforming some of the independent variables. This happens when data transformation makes the residuals meet the regression models hypotheses: normal distribution, independence, homoscedasticity (constant variance) and linearity.

We observe (see Table 3) that HOMEμ (97%), WDμ (34%), HOMEσ (22%), HTMRμ (15%) and VDR (14%) are the most used variables in regression models, especially HOMEμ, that is present in almost all of them, showing a relevant performance of these features for forest structure and canopy variable prediction. However, NPμ, NPσ, ROUGHσ, HTMRσ, RWEμ and FSμ were never selected in our tests.

For all variables except for the volume, square-root-transformed is the method that achieved better results. However, in most of the cases the difference between this method and the second method ranked is not statistically different, so mainly all the regression models (green-coloured in Figure 6) could be used to obtain similar results. Analysing Tables 4 and 5 we observe that square-root, logarithmic and independent variable transformations have the highest results, but in some cases the difference between these regression models and the rest are not significant (see Figure 6). This improvement of results using square-root and logarithmic transformation is also mentioned in Means *et al.* (2000) and Andersen *et al.* (2005).

The CART method does not have good results for prediction, however, since it organises each leaf containing homogenous instances, it could be used for data stratification or categorization.

Finally, it is well known that is crucial to use cross-validation or independent evaluation data sets for training and evaluation, but it becomes indispensable when Random Forest algorithm is employed, so as not to overestimate results (see Figure 8).

## Acknowledgments

This research has been funded by the Spanish Ministerio de Economía y Competitividad and FEDER, in the framework of the project CGL2013-46387-C2-1-R.

The authors thank the Bureau of Land Management and the Panther Creek Remote Sensing and Research Cooperative Program for the data provided for this research.

## References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory*. Akadémia Kiado, Budapest, Hungary. pp. 267-281.

Andersen, H.E., McGaughey, R.J., Reutebuch, S.E. 2005. Estimating forest canopy fuel parameters using LiDAR data. *Remote Sensing of Environment*, 94(4), 441-449. http://dx.doi.org/10.1016/j.rse.2004.10.013

Andersen, H.E., Breidenbach, J. 2007. Statistical properties of mean stand biomass estimators in a lidar-based double sampling forest survey design. In: *ISPRS Workshop on Laser Scanning 2007 and SilviLaser, 2007*. Espoo, Finland, September 12-14. pp. 8-13.

Baccini, A., Laporte, N., Goetz, S.J., Sun, M., Dong, H. 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. *Environmental Research Letters*, 3(4), 1-9. http://dx.doi.org/10.1088/1748-9326/3/4/045011

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. *Classification and regression trees*. New York: Chapman and Hall.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Cao, L., Coops, N.C., Hermosilla, T., Innes, J., Dai, J., She, G. 2014. Using small-footprint discrete and full-waveform airborne LiDAR metrics to estimate total biomass and biomass components in subtropical forests. *Remote Sensing*, 6, 7110-7135. http://dx.doi.org/10.3390/rs6087110

De'Ath, G., Fabricius, K.E. 2013. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), 3178-3192. http://dx.doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Reasearch*, 7, 1-30.

Duong, H.V. 2010. Processing and application of ICESat large footprint full waveform laser range data. *Ph.D. Thesis*, Delft University of Technology, Netherlands.

Erdody, T.L., Moskal, L.M. 2010. Fusion of LiDAR and imagery for estimating forest canopy fuels. *Remote Sensing of Environment*, 114(4), 725-737. http://dx.doi.org/10.1016/j.rse.2009.11.002

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675-701. http://dx.doi.org/10.1080/01621459.1937.10503522

Friedman, M. 1940. A comparison of alternative tests of significance for the problem of *m* rankings. *The Annals of Mathematical Statistics*, 11(1), 86-92. http://dx.doi.org/10.1214/aoms/1177731944

García-Gutiérrez, J., González-Ferreiro, E., Mateos-García, D., Riquelme-Santos, J.C., Mirando, D. 2011. A comparative study between two regression methods on LiDAR data: A case study. *Lecture Notes in Artificial Intelligence*, 6679, 311-318. http://dx.doi.org/10.1007/978-3-642-21222-2_38

García-Gutiérrez, J., González-Ferreiro, E., Riquelme-Santos, J.C., Miranda, D., Diéguez-Aranda, U., Navarro-Cerrillo, R.M. 2013. Evolutionary feature selection to estimate forest stand variables using LiDAR. *International Journal of Applied Earth Observation and Geoinformation*, 26, 119-131. http://dx.doi.org/10.1016/j.jag.2013.06.005

Guyon, I., Elisseeff, A. 2003. An introduction to variables and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.

Hannon, L., Knapp, P. 2003. Reassessing nonlinearity in the urban disadvantage/violent crime relationship: an example of methodological bias from log transformation. *Criminology*, 41(4), 1427-1448. http://dx.doi.org/10.1111/j.1745-9125.2003.tb01026.x

Hermosilla, T., Ruiz, L.A., Kazakova, A.N., Coops, N.C., Moskal, L.M. 2014. Estimation of forest structure and canopy fuel parameters from small-footprint full-waveform LiDAR data. *International Journal of Wildland Fire*, 23(2), 224-233. http://dx.doi.org/10.1071/WF13086

Hernández-Orallo, J., Ramírez, M.J., Ferri, C. 2004. *Introducción a la minería de datos*. Madrid: Pearson Educación S.A.

Hyyppä, J., Hyyppä, H., Inkinen, M., Engdahl, M., Linko, S., Zhu, Y-H. 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, 128(1-2), 109-120. http://dx.doi.org/10.1016/S0378-1127(99)00278-9

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., Harding, D. 1999. Lidar remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. *Remote Sensing of Environment,* 70(3), 339-361. http://dx.doi.org/10.1016/S0034-4257(99)00052-8

Li, L., Guo, Q., Tao, S., Kelly, M., Xu, G. 2015. Lidar with multi-temporal MODIS provide a means to upscale predictions of forest biomass. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102, 198-208. http://dx.doi.org/10.1016/j.isprsjprs.2015.02.007

Luengo, J., García, S., Herrera, F. 2012. On the Choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32(1), 77-108. http://dx.doi.org/10.1007/s10115-011-0424-2

Marabel-García, M., Álvarez-Taboada, F. 2014. Estimación de biomasa en herbáceas a partir de datos hiperespectrales, regresión PLS y la transformación continuum removal. *Revista de Teledetección*, 42, 49-59. http://dx.doi.org/10.4995/raet.2014.2286

Means, J.E., Acker, S.A., Fitt, B.J., Renslow, M., Emerson, L., Hendrix, C.J. 2000. Predicting forest stand characteristics with airborne scanning lidar. *Photogrammetric Engineering & Remote Sensing*, 66(11), 1367-1371.

Naesset, E., Bollandsas, O.M., Gobakken, T. 2005. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment*, 94(4), 541-553. http://dx.doi.org/10.1016/j.rse.2004.11.010

Nemenyi, P.B. 1963. Distribution-free multiple comparisons. *Ph.D. Thesis*, Princeton University, New Jersey, USA.

Posada, D., Buckley, T.R. 2004. Model selection and model averaging in Phylogenetics: advantages of Akaike Information Criterion and Bayesian Approaches over Likelihood Ratio tests. *Systematic biology*, 53(5), 793-808. http://dx.doi.org/10.1080/10635150490522304

Quinlan, J.R. 1992. Learning with continuous classes. *Machine Learning*, 92, 343-348.

Segal, M.R. 2004. Machine learning benchmarks and Random Forest regression. *Technical report, Center for Bioinformatics & Molecular Biostatistics*, University of California, San Francisco, USA.

Skowronski, N.S., Clark, K.L., Duveneck, M., Hom, J. 2011. Three-dimensional canopy fuel loading predicted using upward and downward sensing LiDAR systems. *Remote Sensing of Environment*, 115(2), 703-714. http://dx.doi.org/10.1016/j.rse.2010.10.012

Standish, J.T., Manning, G.H., Demaershalk, J.P. 1985. Development of biomass equations for British Columbia tree species. *Canadian Forestry Service, Pacific Forest Research Center*, Information Report BC-X-264, Victoria, BC, Canada.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464. http://dx.doi.org/10.1214/aos/1176344136

Temesgen, H., Strunk, J., Andersen, H., Flewelling, J. 2015. Evaluating different models to predict biomass increment from multi-temporal lidar sampling and remeasured field inventory data in south-central Alaska. *Mathematical and computational forestry and natural resource sciences*, 7(2), 66-80.

Wang, L., Zhou, X-H., 2005. A fully nonparametric diagnostic test for homogeneity of variances. *The Canadian Journal of Statistics*, 33(4), 545-558. http://dx.doi.org/10.1002/cjs.5550330406

Wilcoxon, F., 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83. http://dx.doi.org/10.2307/3001968

Zar, J.H., 1999. *Biostatistical analysis*. Upper Saddle River, New Jersey: Prentice Hall.