Neurocomputing 167 (2015) 24–31

Contents lists available at ScienceDirect

Neurocomputing

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

A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables

J. García-Gutiérrez [a,*], F. Martínez-Álvarez [b], A. Troncoso [b], J.C. Riquelme [a]

[a] Department of Computer Science, University of Seville, Spain
[b] Department of Computer Science, Pablo de Olavide University of Seville, Spain

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form
2 August 2014
Accepted 17 September 2014
Available online 14 May 2015

Keywords:
LiDAR
Machine learning
Regression
Remote sensing

ABSTRACT

Light Detection and Ranging (LiDAR) is a remote sensor able to extract three-dimensional information. Environmental models in forest areas have been benefited by the use of LiDAR-derived information in the last years. A multiple linear regression (MLR) with previous stepwise feature selection is the most common method in the literature to develop those models. MLR defines the relation between the set of field measurements and the statistics extracted from a LiDAR flight. Machine learning has emerged as a suitable tool to improve classic stepwise MLR results on LiDAR. Unfortunately, few studies have been proposed to compare the quality of the multiple machine learning approaches. This paper presents a comparison between the classic MLR-based methodology and regression techniques in machine learning (neural networks, support vector machines, nearest neighbour, ensembles such as random forests) with special emphasis on regression trees. The selected techniques are applied to real LiDAR data from two areas in the province of Lugo (Galizia, Spain). The results confirm that classic MLR is outperformed by machine learning techniques and concretely, our experiments suggest that Support Vector Regression with Gaussian kernels statistically outperforms the rest of the techniques.

1. Introduction

Light Detection and Ranging (LiDAR) is a remote laser-based technology which differs from optic sensors in its ability to determine heights of objects. LiDAR is able to measure the distance from the source to an object or surface providing not only x–y position, but also the coordinate z for every impact. The distance to the object is determined by measuring the time between the pulse emission and detection of the reflected signal taking into account the position of the emitter.

LiDAR sensors have transformed the way to perform many important tasks for the natural environment. The work previously done with expensive or not always-feasible fieldwork has partially been replaced by the processing of airborne LiDAR point cloud (initial product obtained from a LiDAR flight). Although the development of digital elevation models has traditionally been the main use of LiDAR [1,2], applications for other purposes can also be found in the literature. Thus, research work often aims to the extraction of descriptive variables from LiDAR and their use to develop products related to urban or environmental mapping and forest management. For those tasks, machine learning and, more precisely supervised learning, is usually the selected tool in the form of both classification (used in most urban or environmental mapping approaches) and regression (most common in estimation of biophysical variables).

Regarding classification, we can find techniques such as Support Vector Machines or Random Forests (RF) applied to LiDAR (isolated or fused with other information sources, e.g., multispectral images) for the development of forest inventories [3,4] or fuel models [5]. But even when classification has been also important for LiDAR, researchers have specially focused on deriving variables related to the LiDAR's ability to extract vertical information and then, worked on establishing relations with field measurements. Hence, regression techniques have been paid a greater attention to improve empirical models. Following this philosophy, LiDAR can currently be found for different tasks such as estimation of biomass in forest areas [6] or prediction of building ages [7].

In the case of forest LiDAR-derived models, we can observe multiple linear regression (MLR) has usually been the main tool for the estimation of parameters regressed from LiDAR statistics. The main advantage of using this type of methodology is the simplicity and clarity of the resulting model. In contrast, the selected method also has some drawbacks: this process provides a set of highly correlated predictors with little physical justification [8] and, as a

* Corresponding author.
E-mail addresses: jorgarcia@us.es (J. García-Gutiérrez), fmaralv@upo.es (F. Martínez-Álvarez), ali@upo.es (A. Troncoso), riquelme@us.es (J.C. Riquelme).

http://dx.doi.org/10.1016/j.neucom.2014.09.091
0925-2312/

parametric technique, it is only recommended when assumptions such as normality, homocedasticity, independence and linearity are met [9].

With the previous in mind, it is important to outline that methodologies to develop regression models between field-work data and LiDAR are being reviewed [10]. As a consequence, machine learning non-parametric regression techniques have started to be applied with success.

Our aim in this work is to compare the most well-known regression techniques of machine learning in a common framework. Thus, we can establish a ranking when they are applied to forest variable estimation to help environmental researchers for the selection of the most suitable technique for their needs. The different techniques have been tested and statistically validated on two LiDAR datasets from two different areas of the province of Lugo (Galizia, Spain).

The rest of the paper is organized as follows. Section 2 provides a general review of the state of the art and Section 3 shows a description of the LiDAR data used in this work. The methodology is presented in Section 4. The results achieved, their statistical validation and the main findings are shown in Section 5. Finally, Section 6 is devoted to summarize the conclusions and to discuss future lines of work.

2. Related work

Researchers have already explored advanced regression techniques for forest variables estimation and recent literature provides examples which show their suitability in comparison with MLR. Thus, Chen and Hay [11] used Support Vector Regression (SVR) to estimate biophysical features of vegetation from LiDAR data and multispectral images outperforming classical stepwise regression. Their results were confirmed by Jachowski et al. [12] although SVR was only applied to multispectral images in this case.

Decision trees in the form of ensembles such as RF have been applied with good results. Thus, Latifi et al. showed [13] how they could be used for biomass estimation and outperform classical stepwise regression after an evolutionary feature selection. Even without the evolutionary feature selection, similar results have been also reported by other researchers [14]. Moreover, they were not only used for onshore biomass estimation but also employed to model and predict seafloor standing stocks [15] and for large datasets where they showed a performance as good as that of smaller ones [16].

In addition to SVR and RF as the most extended machine learning regression techniques recently published in the literature, other techniques have also been explored. Thus, Zhao et al. [17] provided a comparison between Gaussian Processes (GP) and stepwise MLR where the first ones clearly improved the results after a set of composite features were extracted from a LiDAR point cloud. Hudak et al. [18] applied nearest neighbours to extract relations between LiDAR and fieldwork for several vegetation species at plot level.

Although machine learning seems to be suitable to extract meaningful information from LiDAR, few studies have been provided to compare the quality of the regressions obtained by different sets of techniques. For instance, Gleason and Im [19] showed a partial comparison of methods where SVR outperformed RF and Li et al. [20] established a deep comparison among machine learning techniques where SVMs and boosted decision trees obtained the best results though simple ordinary least squares approach performed just as well as any advanced machine learning method. Recently Gagliasso et al. [21] examined the predictive performance of linear regression, geographic weighted regression, gradient nearest neighbour, most similar neighbour, RF imputation, and k-Nearest Neighbour (kNN) to estimate biomass and basal area. A combination of ground inventory plots, LiDAR data, satellite imagery, and climate data was analyzed, and the root mean square error (RMSE) and bias were calculated to test the different methods. In this case, results showed that for biomass prediction, the kNN ($k=5$) had the lowest RMSE and the least amount of bias. Although a statistical validation of results is commonly required for a comparison of machine learning techniques [22], no statistical study was provided for either of the previous works which would have been desirable to generalize their conclusions.

Stojanova et al. [23] performed a comparison of the results of different types of regression trees, in particular, isolated trees and ensembles such as RF. The results confirmed the use of ensembles improved the performance when the estimation of forest variables was carried out. Although a deep statistical validation was applied to the results, no regression technique from other families such as kNN or SVR was compared with the regression trees.

After studying the possible improvements of the recent bibliography, we carried out a comparison of the most well-known regression techniques of machine learning in a common framework. Then we established a ranking when they were applied to the estimation of forest variables after being tested and statistically validated on two LiDAR datasets from two different areas of the province of Lugo (Galizia, Spain).

3. Materials

3.1. Study sites

Aerial LiDAR data in two forest areas in the northwest part of the Iberian Peninsula (Fig. 1) were used for this study (more details about both areas can be found in Goncalves-Seco et al. [24] and Gonzalez-Ferreiro et al. [25], respectively).

The first study area (hereafter site A) was located in Trabada, concretely in the municipality of Vilapena (Galicia, NW Spain; boundaries 644800; 4806600 and 645800; 4810600 UTM). *Eucalyptus globulus* stands, with low intensity silvicultural treatments and the presence of tall shrubs, dominated the forest type.

The second study area (hereafter site B) was also located in Galicia (NW Spain), in the municipality of Guitiriz, and covered about 36 km^2 of *Pinus radiata* forests (boundaries 586315; 4783000 and 595102; 4787130 UTM). *P. radiata* was the main forest type in this area and its stands were also characterized by low-intensity silvicultural treatments and by the presence of tall shrubs.

3.2. Field data

Field data from the two study sites were collected to obtain the dependent variables for the regressions in this work. Thus, 39 instances (one per training plot in the study site) were located and measured on site A. On site B, a similar process was carried out for a total of 54 plots. The plots were selected to represent the existing range of ages, stand sizes, and densities in the studied forests.

For site A and B, the dry weight of the biomass fractions of each tree was estimated using the equations for *E. globulus* in Galicia reported by Dieguez-Aranda et al. [26]. 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 field measurements (heights and diameters) and the estimated volumes and dry weight of the biomass fractions helped to estimate the following additional

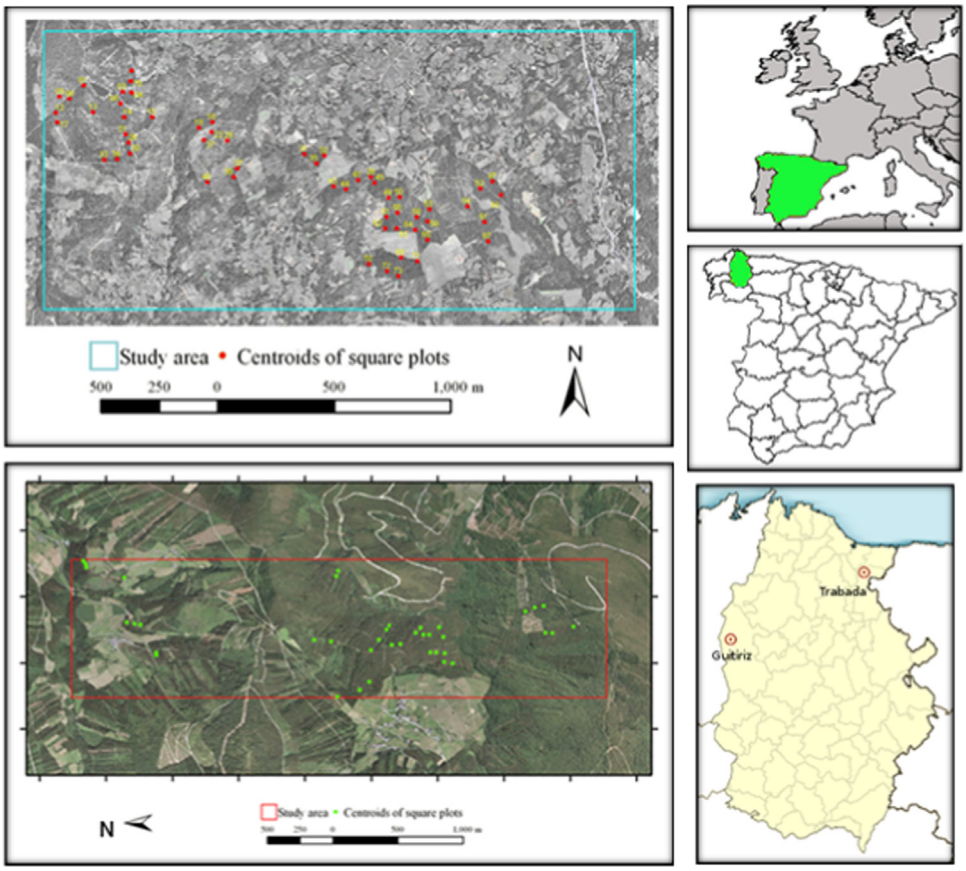


Fig. 1. Study sites located in the province of Lugo (NW of Spain). Top: study site of Guitiriz. Bottom: study site of Trabada.

stand variables in each plot: stand basal area (G), dominant height (H_d), mean height (H_m), and stand volume (V).

3.3. LiDAR data

The LiDAR data from site A were acquired in November 2004. The first and last return pulses were registered. The whole study area was flown over 18 strips and each strip was flown over three times, which gave an average measurement density of about 4 pulses m^{-2}. The LiDAR data for site B were acquired in September 2007. An averaged laser pulse density of 8 pulses m^{-2} was obtained. In order to obtain two additional different resolutions, an artificial reduction based on a random selection of LiDAR pulses in a grid cell of 1 m^2 was carried out for each flight. They resulted in two new LiDAR datasets with a pulse density of 0.5 pulses [25].

Intensity values in both study sites were normalized to eliminate the influence of path height variations [3]. Filtering, interpolation, and the development of Digital Terrain and Canopy Models (DTM/DCM) were performed by FUSION software [27]. 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). Table 1 shows the complete set of metrics and the corresponding abbreviations used in this paper.

From field data and the statistics obtained from LiDAR we built 60 datasets. Each database was composed of the 48 independent variables ($COVER_{FP}$ and *RETURNS* in Table 1 together with the rest of the variables calculated for intensity and heights) extracted from LiDAR data and a dependent variable (fieldwork-derived forest variable), given a study site and a resolution. This procedure gave a total of 20 datasets (4 and 8 forest variables for site A and B, respectively, multiplied by 2 different resolutions). The remaining 40 datasets were obtained using two types of feature transformation (power and exponential) commonly used in the literature [25] on the previous 20.

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

Description	Abbreviation	Description	Abbreviation
Percentage of first returns over 2 m	COVER_FP	25th percentile	P25
		50th percentile	P50
Number of returns above 2 m	RETURNS	75th percentile	P75
Minimum	MIN	5th percentile	P05
Maximum	MAX	10th percentile	P10
Mean	MEAN	20th percentile	P20
Mode	MODE	30th percentile	P30
Standard deviation	SD	40th percentile	P40
Variance	V	60th percentile	P60
Interquartile distance	ID	70th percentile	P70
Skewness	SKW	80th percentile	P80
Kurtosis	KURT	90th percentile	P90
Average absolute deviation	AAD	95th percentile	P95

4. Regression techniques

4.1. Regression trees

Decision trees based on If–Then rules are one of the most popular methods used in machine learning for classification since they offer results that can be easily interpreted. Thus, this approach obtains ordered sets of If–Then rules for prediction that produces understandable models [28]. Regression trees thus offer

to LiDAR scientist the power of a non-parametric regression technique, which is frequently higher than MLR's when parametric conditions are not met, with an open-box (understandable) model which is often required by environmental users with no expertise in machine learning.

In this work, eight kinds of novel regression tree structures have been applied to forest variables prediction from LIDAR data. What makes them different from each other are the regression models considered in the leaf nodes. Specifically, we consider the following models:

Model MEAN-LM: This model is a constant linear model. This constant value is the mean of the target variable for all instances of the training set that reach the leaf.

Model MED-LM: This model is also a constant linear model, but the constant value is the median instead of the mean.

Model LM: Linear model in which the coefficients are computed by minimizing the mean square error.

Model MDL-LM: Linear model in which the coefficients are computed based on a minimum description length principle.

Model RLM: This model is a reduced linear model. Once the linear model is obtained by the least square method, it is simplified as in M5′ algorithm [29]. In particular, a greedy search is used to determine which variables of the linear function can be removed by minimizing the estimated error.

Model kNN-LM: This model is a local non-linear model, based on the *k-NN* algorithm. In this case, the forecasting is divided into two steps clearly differentiated. In a first step, the k nearest neighbours are searched for between the instances of the training set that reach the leaf, and secondly, the mean of the target variable of the k nearest neighbours is computed. Thus, the previously computed mean is the prediction.

Model WkNN-LM: This model is also a local non-linear model based on the weighted k nearest neighbours. The prediction is the weighted mean of the target variable of the k nearest neighbours and the weights w_i are given by a Gaussian kernel:

$$w_i = \exp\left(\frac{d(x_i, n(x_i))}{\sigma}\right) \quad (1)$$

where d is the Euclidean distance between a point x_i and its i-th nearest neighbour $n(x_i)$, σ is the width of the Gaussian kernel and $\exp(\cdot)$ is the exponential function.

Model LWLR: This model is a local linear model based on a locally weighted linear regression, in which the coefficients are computed by minimizing the weighted mean square error and the weights are given by a distance function.

Summarizing, six linear models (five global and one local) and two non-linear local models are considered to construct regression trees. In general, the linear models present some limitations since when models in leaves are too simple, some functions cannot be approximated. On the other hand, global models have a lower computational cost than local models, since the latter have to be rebuilt for each point of the test set. In this work, the package CORElearn [30] (available in software R) has been used to generate the different regression trees described above.

4.2. Other techniques

With the goal of comparing the results of the regression trees above with several families of machine learning techniques applied to LiDAR data for estimation of forest variables, we selected the most extended machine learning algorithms in the literature such as support vector machines for regression, artificial neural networks, nearest neighbour rule and ensembles of regression trees among others. Namely, we included in the study the algorithms implemented in the open source software for data mining WEKA [31]: M5P (another implementation of the M5' regression tree), SMOreg (SVR with polynomial, SMO-p, and Gaussian, SMO-g, kernels), LinearRegression (classical MLR), MultilayerPerceptron (MLP, a type of neural network), Gaussian processes (with Gaussian, GP-g, and polynomial, GP-p, kernels) and IBk (an implementation of kNN). On the other hand, we developed an ad hoc RF which consists in replacing its random trees by M5P trees. This change was necessary because the original implementation in WEKA only allows its use for classification and not for regression.

4.3. Comparison framework

The regression trees and the M5P, SMO-p, SMO-g, MLP, IBk, GP-g, GP-p, MLR and RF techniques described in Sections 4.1 and 4.2, respectively, were applied to the 60 testing datasets.

The comparison was defined from the coefficients of correlation, R, as was done in recent bibliography [17]. Also we included an analysis of the results in terms of root mean square error (RMSE). In our case, the coefficients obtained in a process of Leave-One-Out Cross-Validation (LOOCV) on each dataset since the number of instances was too small for a more adequate n-fold cross-validation. The best and mean values in 100 executions were recorded for each technique and dataset in order to obtain robust results (independent from parameterization).

All algorithms were used after applying a preprocessing phase of normalization and elimination of missing values. Moreover, feature selection to avoid the Hughes phenomenon [32] was also applied. The feature selection based on the well-known Correlation Feature Selection (CFS) filter from Weka was used. Since most techniques implement a feature selection phase in their own building (e.g., regression trees, support vector machines and RF) we decided to carry out each experiment with and without the CFS feature selection. Thus, we could obtain the best result regardless of whether the algorithms could obtain better results with their own feature selection.

As we said, all the considered algorithms were tested on every dataset 100 times using different configuration setups. Moreover, each tuple ⟨⟨configuration setup, algorithm, dataset⟩⟩ was carried out twice: one with feature selection (CFS) and one without. Then, we selected the best of the two to be part of the 100 partial results.

Regarding each algorithm setups, they were defined as random values for each parameter needed to execute the algorithm. Thus, a random value between 1 and 20 was selected for the number of neighbours in IBk, kNN-LM and WkNN-LM models and a σ parameter in the interval [1, number of instances] was also used for WkNN-LM. In the case of the support vector machines, the parameter C was defined in the interval [1,20]. In addition, SVM-g uses a Gaussian kernel which depends on a parameter g defined in the interval [0.01,1]. For SVM-p, which uses a polynomial kernel, the exponent for the polynomial was set up with a value between 1 and 3. Our RF only has two parameters: the number of trees and the number of predictors used per tree. They were in the intervals [1,20] and [1,100], respectively. Finally, the MLP performance is controlled by three parameters: number of hidden layers, momentum and learning rate. They were in the intervals [1,20], [0.01,1] and [0.01,1], respectively.

5. Results

This section provides the results obtained by the application of all the examined algorithms described in Section 4. Results for all the algorithms in terms of maximum and mean R and improvement of *RMSE* regarding the worst technique for each dataset are reported in Section 5. Tests to determine whether the results can be considered statistically different or not have been carried out in Section 5.2. Discussion of the results obtained can be found in Section 5.3.

5.1. Performance of the algorithms

Fig. 2 shows the influence of each statistic on the automatic selections made by the CFS algorithm. Refer to Table 2 to check how variables are mapped into the columns, where Column #1 is associated with the leftmost column in the figure, and Column #48 with the rightmost one. In this sense, a zero value (black) means that the variable was not selected, whereas a one value (white) means just the opposite, that is, the variable was selected.

Figs. 3 and 4 illustrate the maximum and mean of *R*. For visual purposes, Figs. 5 and 6 do not provide *RMSE* but the percentage of improvement regarding the *RMSE* of the worst regressor for each dataset when best and mean *RMSE* are studied, respectively. This representation has been chosen since RMSE is not a dimensionless measure which involves very different values if we are dealing with forest variables such as canopy biomass (kg/ha) or basal area (m^2). Each row of the figures identifies one dataset whereas the columns identify each of the seventeen algorithms. Refer to Table 3 to check how algorithms are mapped into the columns where Column #1 is associated with the leftmost column in the four figures, and Column #17 with the rightmost one. In addition, due to the high number of datasets studied, we provide Tables 4 and 5 which summarize the main statistics for every technique regarding *R* and the percentage of improvement of *RMSE*, respectively.

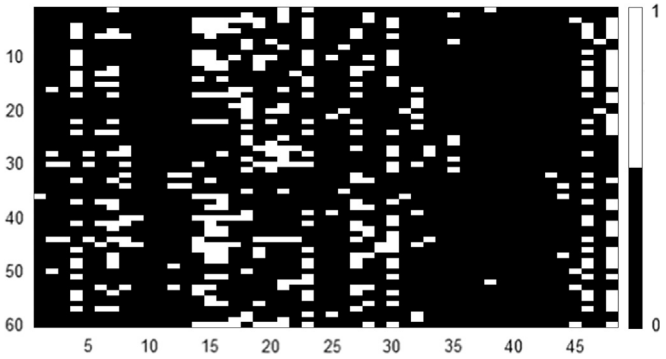


Fig. 2. Selection of LiDAR-derived features by the CFS method through the 60 datasets. A value of 1 means a feature was selected; otherwise, the value is 0.

Table 2
Correspondence between features and columns for Fig. 2. Variable names starting by H and I were obtained from height and intensity distributions, respectively.

Column	Variable	Column	Variable
#1	COVER_FP	#25	IAAD
#2	HAAD	#26	IID
#3	HID	#27	IKURT
#4	HKURT	#28	IMAX
#5	HMAX	#29	IMEAN
#6	HMEAN	#30	IMIN
#7	HMIN	#31	IMODE
#8	HMODE	#32	IP05
#9	HP05	#33	IP10
#10	HP10	#34	IP20
#11	HP20	#35	IP25
#12	HP25	#36	IP30
#13	HP30	#37	IP40
#14	HP40	#38	IP50
#15	HP50	#39	IP60
#16	HP60	#40	IP70
#17	HP70	#41	IP75
#18	HP75	#42	IP80
#19	HP80	#43	IP90
#20	HP90	#44	IP95
#21	HP95	#45	ISD
#22	HSD	#46	ISKW
#23	HSKW	#47	IV
#24	HV	#48	RETURNS

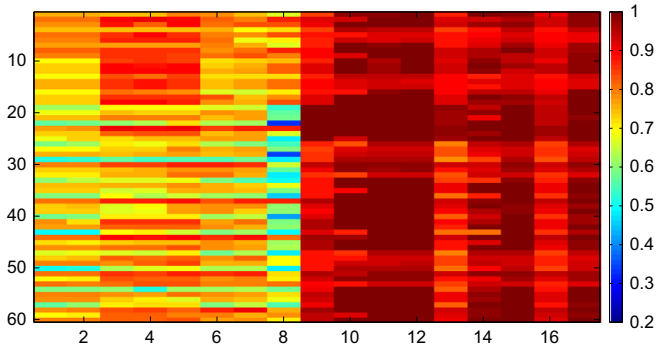


Fig. 3. Maximum *R* values after 100 executions for the 60 datasets each.

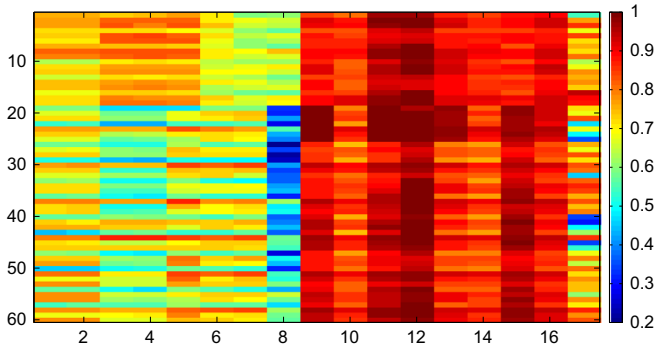


Fig. 4. Mean *R* values after 100 executions for the 60 datasets each.

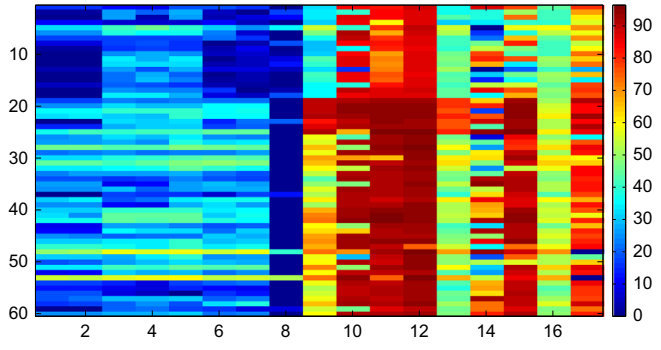


Fig. 5. Maximum percentage of improvement regarding the best *RMSE* reached by the worst competitor after 100 executions for the 60 datasets.

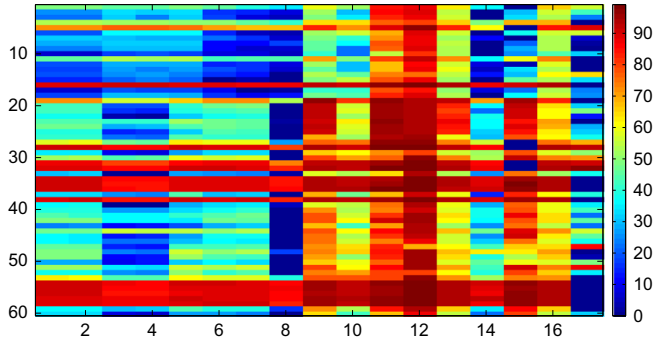


Fig. 6. Mean percentage of improvement regarding the mean *RMSE* reached by the worst competitor after 100 executions for the 60 datasets.

Table 3
Correspondence between algorithms and columns for Figs. 3–6.

Column	Algorithm
#1	MEAN-LM
#2	MED-LM
#3	LM
#4	MDL-LM
#5	RLM
#6	kNN-LM
#7	WkNN-LM
#8	LWLR
#9	MLR
#10	IBk
#11	SMO-p
#12	SMO-g
#13	M5P
#14	GP-g
#15	GP-p
#16	RF
#17	MLP

Table 4
Averaged best and mean *R* for the 60 datasets after 100 executions. The best values in bold.

Algorithm	Maximum *R*	Mean *R*
MEAN-LM	0.717	0.711
MED-LM	0.716	0.711
LM	0.785	0.780
MDL-LM	0.786	0.781
RLM	0.796	0.791
kNN-LM	0.765	0.699
WkNN-LM	0.768	0.696
LWLR	0.687	0.562
MLR	0.916	0.913
IBk	0.959	0.845
SMO-p	0.976	0.944
SMO-g	**0.979**	**0.974**
M5P	0.891	0.891
GP-g	0.946	0.851
GP-p	0.970	0.929
RF	0.906	0.898
MLP	0.969	0.711

5.2. Statistical analysis

After the generation of the quality results for the different models, a statistical analysis was applied by using the open-source platform StatService [33] to check the significance in the differences among multiple methods in terms of *R*. ANOVA is usually used for multiple comparison of results if parametric conditions (homoscedasticity, independence, normality) are met [22]. However, Levene test rejected the homoscedasticity hypothesis of the results with a *p*-value under 0.001 for an $\alpha = 0.05$, and therefore, a non-parametric procedure was selected. This 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 would mean a method was the best for a test case, while a rank of *n* would mean that it was the worst of the *n* compared methods. Finally, the chosen procedure used the Friedman test and the Holm post hoc procedure (see [34] for a complete description of both non-parametric methods) to statistically validate the differences in the mean ranks.

Table 6 summarizes the rankings obtained for maximum *R* and mean *R* when Friedman's test was applied. The statistics for Friedman with control method were 815.17 and 854.02, all of them distributed according to a chi-square distribution with 16 degrees of freedom. The *p*-value for the Friedman test were less than 0.0001 so it rejected the null hypothesis (all the techniques behave in a similar way) with a level of significance of $\alpha = 0.05$. Then, the Holm post hoc procedure was applied.

Table 7 shows the *p*-values, *z*-value and α, using SMO-g as control algorithm since it obtained the best ranking in terms of maximum *R*. Note that Holm procedure rejects all the algorithm except for SMO-p, whose α value was greater than the required by Holm. In other words, differences between kernels applied to SVM did not provide results statistically different, which was expected, given that both of them were SVMs and the similar results they had obtained on average and that had been previously reported in Table 4.

The same information can be found in Table 8 using again SMO-g as control algorithm since it also obtained the best ranking in terms of mean *R*. As can be seen, the null hypothesis is rejected for the rest of the 16 algorithms, thus concluding that SMO-g generated the best results and they were statistically different from those of the other algorithms.

5.3. Discussion

Through the analysis of the results of the experimentation, it was possible to draw some interesting findings.

Regarding the attribute selection technique used, we may observe that there is a pattern in the selected set. This pattern is consistent

Table 5
Averaged improvement regarding maximum and mean *RMSE* obtained by the worst regressor for each of the 60 datasets after 100 executions. The best values in bold.

Algorithm	% Maximum *RMSE*	% Mean *RMSE*
MEAN-LM	19.811	48.737
MED-LM	20.015	48.726
LM	28.163	40.740
MDL-LM	27.458	40.487
RLM	30.207	50.106
kNN-LM	23.443	47.409
WkNN-LM	23.373	46.670
LWLR	4.517	25.344
MLR	57.019	73.572
IBk	78.055	61.647
SMO-p	86.751	84.174
SMO-g	**91.224**	**91.981**
M5P	51.769	70.443
GP-g	55.472	42.376
GP-p	77.752	62.911
RF	51.703	69.255
MLP	70.682	29.988

Table 6
Rankings from maximum and mean *R* for the 17 algorithms after 100 executions. The best in bold.

Algorithm	Maximum *R*	Mean *R*
MEAN-LM	15.408	13.658
MED-LM	15.433	13.717
LM	12.175	10.325
MDL-LM	12.025	10.258
RLM	11.258	9.625
kNN-LM	12.900	14.167
WkNN-LM	12.967	14.167
LWLR	15.300	16.233
MLR	7.533	4.783
IBk	3.975	8.117
SMO-p	2.533	2.400
SMO-g	**1.450**	**1.200**
M5P	8.467	5.283
GP-p	3.883	4.133
GP-g	5.542	7.733
RF	7.450	4.683
MLP	4.700	12.517

Table 7
Post hoc analysis using Holm test and SMO-g as the control algorithm, in terms of maximum *R*.

Algorithm	*p*	*z*	Holm	H_0 rejected
MEAN-LM	0	15.1399	0.0033	√
MED-LM	0	15.1671	0.0031	√
LM	0	11.6329	0.0045	√
MDL-LM	0	11.4702	0.005	√
RLM	0	10.6386	0.0056	√
kNN-LM	0	12.4193	0.0042	√
WkNN-LM	0	12.4916	0.0038	√
LWLR	0	15.0224	0.0036	√
MLR	0	6.5983	0.0071	√
IBk	0.0062	2.7387	0.0167	√
SMO-p	0.24	1.175	0.05	
GP-p	0.0083	2.6393	0.025	√
GP-g	0	4.438	0.01	√
M5P	0	7.6106	0.0063	√
MLP	0.0004	3.5251	0.0125	√
RF	0	6.5079	0.0083	√

Table 8
Post hoc analysis using Holm test and SMO-g as control algorithm, in terms of mean *R*.

Algorithm	*p*	*z*	Holm	H_0 rejected
MEAN-LM	0	13.513	0.0042	√
MED-LM	0	13.5762	0.0038	√
LM	0	9.8975	0.005	√
MDL-LM	0	9.8251	0.0056	√
RLM	0	9.1382	0.0063	√
kNN-LM	0	14 .0643	0.0033	√
WkNN-LM	0	14.0643	0.0036	√
LWLR	0	16.3059	0.0031	√
MLR	0 .0001	3.8867	0.0125	√
IBk	0	7 .5022	0.0071	√
SMO-p	0.1931	1.3016	0.05	√
M5P	0	4.429	0.01	√
MLP	0	12.2746	0.0045	√
GP-p	0.0015	3.1816	0.025	√
GP-g	0	7.0864	0.0083	√
RF	0.0002	3.7782	0.0167	√

with that seen in other studies regarding the prevalence of variables related to heights (specially high percentiles) as key elements to extract knowledge from LiDAR [20]. The contribution of the intensity is mostly based on the use of minimum intensity, skewness and kurtosis. The number of returns was also a key factor to consider according to the output of the CFS method in certain databases. Selection of height percentiles is not novel but the fact that intensity variables appeared should be considered. In any case, the variables related to the intensity that were selected seem to be those most resistant to the effect of multiple returns (well-known issue in the literature [35]) since they work with the shape of their distributions (skewness, kurtosis) rather than specific values. In any case, most of the techniques used in this study had its own variable selection and the best performance (with or without previous CSF) was selected to establish the comparison with the rest of the competitors so that the importance of CFS selection should be relatively taken into account.

If we focus on the performance of the regression techniques and as can be inferred from the analysis of Fig. 3 and the subsequent statistical analysis, SMO-g (column #13) obtained the best results regarding *R*, with an averaged maximum value of 0.979. Moreover, from all the 100 executions with different configuration setups, it reached an *R* of 0.974 on average. In addition, most of the rest of the studied machine learning techniques obtained high *R* values. SMO-p even obtained non-statistically different results in terms of maximum. This involves that parameterization was adequate although, in some cases, this point can be difficult to deal with. Specially difficult was the parameterization of MLP (as can be seen comparing maximum and mean values) which suffered from a higher number of parameters than the rest which makes it harder to optimize. On the other hand, regression trees (including their ensembles, RF), which are easier to parameterize, did not get as good results showing a limited prediction power. Among them, our results show that M5P obtained the best results and its use in ensembles (RFs) increase the quality of its results even though they could not compete against SVR.

The results in terms of *RMSE* (Figs. 5 and 6, Table 5) also outline the previous finding. Light differences were found between the use of *R* and *RMSE* for validation. This may be explained by the fact that RMSE is very sensitive to extreme values. This issue is well-known in the literature [36] and in an environment in which databases are very small (in our case, less than 60 instances), the LOOCV procedure makes certain instances very difficult to deal with. This situation specially affects those algorithms prone to overfitting though potentially more powerful. By contrast, regression trees seem to be more resistant to overfitting although they globally do not fit as well as other techniques such as SVMs or GPs.

Finally, the most important point of this study was maybe that SVMs and GP-p outperformed MLR results according to the ranking obtained in Table 6. This fact concurs with that previously reported [11,12,17,19,20] and shows machine learning as an important tool which must be taken into account for LiDAR-derived estimation of forest variables such as the ones studied in this work.

6. Conclusions

This paper presented a comparison between common regression techniques in machine learning (neural networks, support vector machines, nearest neighbour, Gaussian processes, several types of regression trees, and ensembles such as Random Forest) and the classic MLR-based methodology for the estimation of forest variables from LiDAR data. The selected techniques were applied to real LiDAR data from two areas in the province of Lugo (Galizia, Spain). The results showed that SVMs statistically outperformed the rest of the techniques. Nevertheless, results confirmed recent bibliography since machine learning techniques (SVMs, GPs) obtained better results than those of classic MLR.

Future work should address gaps not covered in this work. Thus, we must complete the framework with an ad hoc feature selection for each specific method specially for those techniques that do not have it in their own building. For those techniques CFS could obtain a non-optimal subset of features when comparing with an ad hoc feature selection specially designed for them. In the same line, automatic parameterization could make the use of the most powerful techniques available for the general LiDAR users. Both problems can be solved at the same time with the application of evolutionary computation although a trade-off between optimization and run time should be reached for industrial use.

Acknowledgements

The authors would like to thank Spanish Ministry of Science and Technology, Junta de Andaluca and Pablo de Olavide University for the support under projects TIN2011-28956-C02, P12-TIC-1728 and APPB813097, respectively.

References

[1] L. Goncalves-Seco, D. Miranda, R. Crecente, J. Farto, Digital terrain model generation using airborne LIDAR in forested area of Galicia, Spain, Lisbon, Portugal, in: Proceedings of 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 2006, pp. 169–180.

[2] W.L. Lu, K.P. Murphy, J.J. Little, A. Sheffer, H. Fu, A hybrid conditional random field for estimating the underlying ground surface from airborne lidar data, IEEE Trans. Geosci. Remote Sens. 47 (8/2) (2009) 2913–2922.

[3] M. Garcia, D. Riano, E. Chuvieco, F.M. Danson, Estimating biomass carbon stocks for a Mediterranean forest in Central Spain using LiDAR height and intensity data, Remote Sens. Environ. 114 (4) (2010) 816–830.

[4] R. Pino-Mejias, M.D.C. de-la Vega, M. Anaya-Romero, A. Pascual-Acosta, A. Jordan-Lopez, N. Bellinfante-Crocci, Predicting the potential habitat of oaks with data mining models and the R system, Environ. Model. Softw. 25 (7) (2010) 826–836.

[5] M. Mutlu, S.C. Popescu, C. Stripling, T. Spencer, Mapping surface fuel models using LiDAR and multispectral data fusion for fire behavior, Remote Sens. Environ. 112 (1) (2008) 274–285.

[6] E. Gonzalez-Ferreiro, U. Dieguez-Aranda, L. Gonçalves-Seco, R. Crecente, D. Miranda, Estimation of biomass in *Eucalyptus globulus* labill. forests using different LiDAR sampling densities, in: Proceedings of ForestSat, 2010.

[7] T.R. Tooke, N.C. Coops, J. Webster, Predicting building ages from LiDAR data with random forests for building energy modeling, Energy Build. 68 (Part A) (2014) 603–610.

[8] J.D. Muss, D.J. Mladenoff, P.A. Townsend, A pseudo-waveform technique to assess forest structure using discrete LiDAR data, Remote Sens. Environ. 115 (3) (2010) 824–835.

[9] J. Osborne, E. Waters, Four assumptions of multiple regression that researchers should always test, Pract. Assess. Res. Eval. 8 (2) (2002).

[10] C. Salas, L. Ene, T.G. Gregoire, E. Næsset, T. Gobakken, Modelling tree diameter from airborne laser scanning derived variables: a comparison of spatial statistical models, Remote Sens. Environ. 114 (6) (2010) 1277–1285.

[11] G. Chen, G.J. Hay, A support vector regression approach to estimate forest biophysical parameters at the object level using airborne lidar transects and quickbird data, Photogramm. Eng. Remote Sens. 77 (7) (2011) 733–741.

[12] N.R. Jachowski, M.S. Quak, D.A. Friess, D. Duangnamon, E.L. Webb, A.D. Ziegler, Mangrove biomass estimation in Southwest Thailand using machine learning, Appl. Geogr. 45 (0) (2013) 311–321.

[13] H. Latifi, A. Nothdurft, B. Koch, 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) (2010) 395–407.

[14] J. Zhang, S. Huang, E.H. Hogg, V. Lieffers, Y. Qin, F. He, Estimating spatial variation in Alberta forest biomass from a combination of forest inventory and remote sensing data, Biogeosci. Discuss. 10 (12) (2013) 19005–19044.

[15] C.L. Wei, et al., Global patterns and predictions of seafloor biomass using random forests, PLoS ONE 5 (December (12)) (2010) e15323.

[16] J. Mascaro, et al., A tale of two forests: random forest machine learning aids tropical forest carbon mapping, PLoS ONE 9 (January (1)) (2014) e85993.

[17] K. Zhao, S. Popescu, X. Meng, Y. Pang, M. Agca, Characterizing forest canopy structure with lidar composite metrics and machine learning, Remote Sens. Environ. 115 (8) (2011) 1978–1996.

[18] A.T. Hudak, N.L. Crookston, J.S. Evans, D.E. Halls, M.J. Falkowski, Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LIDAR data, Remote Sens. Environ. 112 (2008) 2232–2245.

[19] C.J. Gleason, J. Im, Forest biomass estimation from airborne LiDAR data using machine learning approaches, Remote Sens. Environ. 125 (2012) 80–91.

[20] M. Li, J. Im, L. Quackenbush, T. Liu, Forest biomass and carbon stock quantification using airborne lidar data: a case study over Huntington wildlife forest in the Adirondack park, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., in press, 2015.

[21] D. Gagliasso, S. Hummel, H. Temesgen, A comparison of selected parametric and non-parametric imputation methods for estimating forest biomass and basal area, Open J. For. 4 (2014) 42–48.

[22] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[23] D. Stojanova, P. Panov, V. Gjorgjioski, A. Kobler, S. Dzeroski, Estimating vegetation height and canopy cover from remotely sensed data with machine learning, Ecol. Inform. 5 (4) (2010) 256–266.

[24] L. Goncalves-Seco, E. Gonzalez-Ferreiro, U. Dieguez-Aranda, B. Fraga-Bugallo, R. Crecente, D. Miranda, Assessing attributes of high density eucalyptus globulus stands using airborne laser scanner data, Int. J. Remote Sens. 32 (24) (2011) 9821–9841.

[25] E. Gonzalez-Ferreiro, U. Dieguez-Aranda, D. Miranda, Estimation of stand variables in *Pinus radiata* D. Don plantations using different lidar pulse densities, Forestry 85 (2) (2012) 281–292.

[26] U. Dieguez-Aranda, et al., Herramientas selvicolas para la gestion forestal sostenible en Galicia. Xunta de Galicia, 2009.

[27] R. McGaughey, FUSION/LDV: software for LIDAR data analysis and visualization, US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Seattle, 2009.

[28] A. Fakhari, A.M. Moghadam, Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval, Appl. Soft Comput. 13 (2) (2013) 1292–1302.

[29] J.R. Quinlan, Learning with continuous classes, in: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, 1992, pp. 343–348.

[30] M. Robnik-Sikonja, r-cran-corelearn, 2013 ⟨http://mloss.org/software/view/310/⟩.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. 11 (1) (2009).

[32] G.F. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Inf. Theory 14 (1968) 55–63.

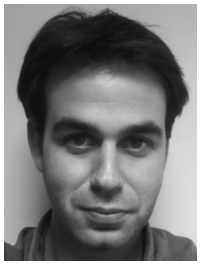
[33] J.A. Parejo, J. García, A. Ruiz-Cortés, J.C. Riquelme, Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheur'isticas. In: Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados, 2012.

[34] J. Luengo, S. Garcia, F. Herrera, A study on the use of statistical tests for experimentation with neural networks: analysis of parametric test conditions and non-parametric tests, Expert Syst. Appl. 36 (2009) 7798–7808.

[35] B. Hofle, N. Pfeifer, Correction of laser scanning intensity data: data and model-driven approaches, ISPRS J. Photogramm. Remote Sens. 63 (2007) 1415–1433.

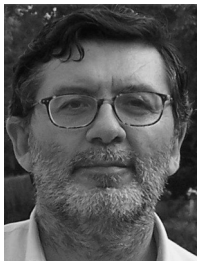
[36] J. Tinoco, A.G. Correia, P. Cortez, Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time, Constr. Build. Mater. 25 (3) (2011) 1257–1262.

Jorge García-Gutiérrez received the Ph.D. degree in Computer Engineering from the University of Seville, Spain, in 2012. He has been working for the Department of Computer Science of the University of Seville since 2008 where he is currently Lecturer Professor. His primary areas of interest are machine learning techniques, big data, remote sensing, data fusion and evolutionary computation.



Francisco Martínez-Álvarez received the M.Sc. degree in Telecommunications Engineering from the University of Seville, and the Ph.D. degree in Computer Engineering from the Pablo de Olavide University. He has been with the Department of Computer Science at the Pablo de Olavide University since 2007, where he is currently an Associate Professor. His primary areas of interest are time series analysis, data mining, and evolutionary computation.

Alicia Troncoso was born in Carmona, Spain, in 1974. She received the Ph.D. degree in Computer Science from the University of Seville, Spain, in 2005. From 2002 to 2005, she was with the Department of Computer Science, University of Seville. Presently, she is an Associate Professor at the Pablo de Olavide University of Seville. Her primary areas of interest are time series analysis, control and forecasting, and optimization techniques.



Jose C. Riquelme received the M.Sc. degree in Mathematics and the Ph.D. degree in Computer Science from the University of Seville, Spain. Since 1987 he has been with the Department of Computer Science, University of Seville, where he is currently Full Professor. His primary areas of interest are data mining, machine learning techniques, and evolutionary computation.