A Comparative Study of Machine Learning Regression Methods on LiDAR Data: A Case Study

Jorge Garcia-Gutierrez[1], Francisco Martínez-Álvarez[2],
Alicia Troncoso[2], and Jose C. Riquelme[1]

[1] Department of Computer Science, University of Seville, Spain
{jgarcia,riquelme}@lsi.us.es
[2] Department of Computer Science, Pablo de Olavide University, Spain
{fmaralv,ali}@upo.es

Abstract. Light Detection and Ranging (LiDAR) is a remote sensor able to extract vertical information from sensed objects. LiDAR-derived information is nowadays used to develop environmental models for describing fire behaviour or quantifying biomass stocks in forest areas. A multiple linear regression (MLR) with previous stepwise feature selection is the most common method in the literature to develop LiDAR-derived models. MLR defines the relation between the set of field measurements and the statistics extracted from a LiDAR flight. Machine learning has recently been paid an increasing attention to improve classic MLR results. 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 common regression techniques in machine learning (neural networks, regression trees, support vector machines, nearest neighbour, and ensembles such as random forests). The selected techniques are applied to real LiDAR data from two areas in the province of Lugo (Galizia, Spain). The results show that support vector regression statistically outperforms the rest of techniques when feature selection is applied. However, its performance cannot be said statistically different from that of Random Forests when previous feature selection is skipped.

Keywords: LiDAR, regression, remote sensing, soft computing.

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). In this context, research work focuses on the extraction of descriptive variables from LiDAR and their relation with field measurements. Following this philosophy, LiDAR is currently used to develop forest inventories [1] or fuel models [2] and to estimate biomass in forest areas [3], among other applications.

LiDAR-derived models are usually based on the estimation of parameters regressed from LiDAR statistics through multiple linear regression (MLR). The main advantage of using this type of methodology is the simplicity of the resulting model. In contrast, the selected method also has some drawbacks: this process results a set of highly correlated predictors with little physical justification [4] and, as a parametric technique, it is only recommended when assumptions such as normality, homoscedasticity, independence and linearity are met [5].

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 [6]. As a consequence, machine learning non-parametric regression techniques have recently started to be applied with success. For example, Hudak et al. [7] applied nearest neighbour to extract relations between LiDAR and fieldwork for several vegetation species at plot level. Chen and Hay [8] used support vector regression to estimate biophysical features of vegetation using data fusion (LiDAR + multiespectral). In the same line, Zhao et al. [9] provided a comparison between Gaussian processes and stepwise MLR where the first clearly improved the results after a set of composite features were extracted from a LiDAR point cloud. Decision trees in the form of random forests have also been applied with good results. Thus, Latifi et al. showed [10] how random forests could be used for biomass estimation and outperform classical stepwise regression after evolutionary feature selection.

Although machine learning seems to provide a suitable tool 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 [11] showed a partial comparison of methods where support vector regression outperformed random forests. Unfortunately, no statistical validation was performed which is necessary to generalize their conclusions.

Our aim in this work was to compare the most well-known regression techniques of machine learning in a common framework. We established a ranking when they were applied to forest variable estimation to help environmental researchers the selection of the most suitable technique for their needs. The different techniques were tested and statistically validated using their results 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 description of the LiDAR data used in this work as well as the methodology used. The results achieved, their statistical validation and the main findings in this work are shown in Section 3. Finally, Section 4 is devoted to summarize the conclusions and to discuss future lines of work.

2 Materials and Method

2.1 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. [12] and Gonzalez-Ferreiro et al. [13], 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). *E. globulus* stands, with low intensity silvicultural treatments and the presence of tall shrubs, dominated the forest type.

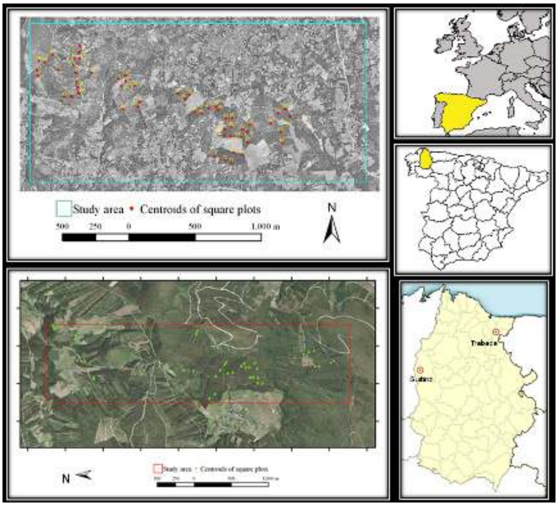

Fig. 1. Study sites located in the province of Lugo (NW of Spain). Top: study site of Guitiriz. Bottom: study site of Trabada.

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 *P. 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.

2.2 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. [14]. 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 stand variables in each plot: stand basal area (G), dominant height (H_d), mean height (H_m), and stand volume (V).

2.3 LiDAR Data

The LiDAR data from site A were acquired in November 2004. 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. A theoretical 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 returns 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 [13].

Intensity values in both study sites were normalized to eliminate the influence of path height variations [1]. Filtering, interpolation, and the development of Digital Terrain and Canopy Models (DTM/DCM) were performed by FUSION software [15]. 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 article.

After the LiDAR data processing, we obtained 60 databases with 48 independent variables ($cover_{FP}$ and $returns$ in Table 1 plus the rest calculated for intensity and heights). The first 20 datasets were composed of the previous statistics and each fieldwork variable as dependent variable (for each study site and resolution). The rest were obtained using two types of feature transformation (allometric and exponential, [13]), respectively.

2.4 Regression Techniques Comparison

The goal of this paper was to compare the results of several families of machine learning techniques when applied to LiDAR data for estimation of forest variables. For comparison, we selected the most extended machine learning algorithms in the literature from the software WEKA [16]: M5P (regression tree), SMOreg (support vector machine for regression), LinearRegression (classic multiple linear regression), MultilayerPerceptron (artificial neural network),

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		25th percentile	P25
returns over 2m	cover_FP	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

IBk (nearest neighbor). We also developed an ad-hoc Random Forest (ensemble of regression trees) based on the original implementation in Weka but 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. In any case, all algorithms were used with default parameters after applying a preprocessing phase of normalization, elimination of missing values, and feature selection (to avoid the Hughes phenomenon [17]) based on the Correlation Feature Selection (CFS) filter of Weka. The comparison was defined from the coefficients of determination (R^2) obtained in a process of 5-Fold Cross-Validation (5FCV).

A key factor for the performance of the techniques is the set of selected attributes in the preprocessing step. In certain cases, such as SVM and Random Forest, techniques perform their own selection of best attributes. A previous selection could therefore affect the quality of the predictions. To study feature selection's influence, we repeated the 5FCV in a second level of experimentation for the best two techniques without applying previous feature selection.

2.5 Statistical Analysis

After the generation of the quality results for the different models, a statistical analysis was used (using the open-source platform StatService [18]) to check the significance in the differences among multiple methods in terms of R^2. ANOVA is usually used for multiple comparison of results if parametric conditions (homoscedasticity, independence, normality) are met [19]. Parametric conditions were checked using the Shapiro-Wilk and Lilliefors tests for normality and the Levene test for homoscedasticity. If parametric conditions were not met, a non-parametric procedure would be selected. This procedure, firstly, would obtain 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 that the method would be the best for a test case, while a rank of n would mean it was the worst of the n compared methods. Finally the chosen procedure would use the Friedman test and the Holm post-hoc procedure (see [20] for a complete description of both non-parametric methods) to statistically validate the differences in the mean ranks.

In addition, for the second level comparison (between the two best methods) a similar procedure was done using a Student's T or a Wilcoxon test which are the corresponding parametric and non-parametric statistical test for pairwise comparisons, respectively [19].

3 Results

Due to the high number of datasets studied, we provide Table 2 which sums up the main statistics for every technique besides their mean ranking throughout the 60 datasets. The whole set of results are also depicted in Fig. 2 and 3. Figures show the results obtained by nearest neigbour (NN), support vector machines (SVM), artificial neural networks (ANN), multiple linear regression (MLR), regression trees (RT) and random forests (RF) for the 60 datasets separated in two subgroups for clarity. They both show the results of the globally best technique (SVM obtained the best mean ranking) when the complete data mining framework (including feature selection) was applied.

Table 2. Mean ranking and main statistics from the results obtained for every regression technique in terms of R^2 throughout the 60 datasets when preprocessing included feature selection

Technique	Mean ranking	R^2 Mean	R^2 Standard deviation
SVM	1.700	0.844	0.062
RF	2.783	0.827	0.071
RT	2.967	0.820	0.079
MLR	3.133	0.815	0.083
ANN	5.133	0.710	0.148
NN	5.283	0.723	0,098

Table 3. P-values and α values for each pairwise comparison in the Holm's procedure

DataSet	**p**	**Holm**
NN	0.000	0.010
ANN	0.000	0.013
MLR	0.000	0.017
RT	0.000	0.025
RF	0.002	0.050

Table 4. Mean ranking and main statistics from the results obtained for the two best regression techniques in terms of R^2 when preprocessing did not include feature selection

Technique	Ranking	R^2 Mean	R^2 Standard deviation
RandomForest	1.45	0.772	0.011
SVM	1.55	0.774	0.006

Rankings were used to assess the statistical significance of the study since Liliefors test rejected the normality hypothesis of the results with a p-value of 0.10 for an $\alpha = 0.05$. In this case, the p-value for the Friedman test was 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. The p-values for the several pairwise comparisons can be found in Table 3. As can be seen, every p-value was lower than the α required by Holm (Holm column in the table) so the procedure concluded that pairwise differences between SVM and the rest of the regression techniques were also statistically significant.

The top two regressors in the previous test were SVM and RF. Both were selected for a subsequent pairwise comparison where the experiment was replicated without performing previous feature selection. Their results are visually presented in Fig. 4 and summarized regarding the ranking and main statistics in Table 4. In this case, the use of the Wilcoxon test with a p-value of 0.9325 could not reject the null hypothesis (i.e., there were no significant differences between them) with a level of significance of $\alpha = 0.05$.

Through the analysis of the results of experimentation, it is possible to draw some important findings. First, our experiments confirmed that both SVM and RF are suitable tools to improve the performance of classical predictions (e.g., MLR or NN) for estimation of forest variables from LiDAR statistics although there is still room to improve the predictions.

Regarding the application of SVM or RF, our experiments showed that although feature selection outperformed every technique, SVM is more sensitive to the feature selection method since its performance decreased more in the second part of the experimentation. Moreover, if the feature selection is not adequate or does not exist, there would be little difference between RF and SVM (in our case, RF even globally outperformed SVM in most cases) as was shown in Table 4. This finding could justify the results of Gleason and Im [11] (where SVM outperformed RF) since the authors made the feature selection manually.

The fact that SVM outperformed RF can be also attributed to CFS feature selection provided a better set of attributes for SVM than for RF. More experimentation is needed to check if the same results can be obtain with other automatic feature selection techniques. In addition, it was also possible that the random nature of RF did not optimally combine the selected features. In any case, this study confirmed the well-known importance of feature selection for the performance of machine learning also in the LiDAR-regression context.

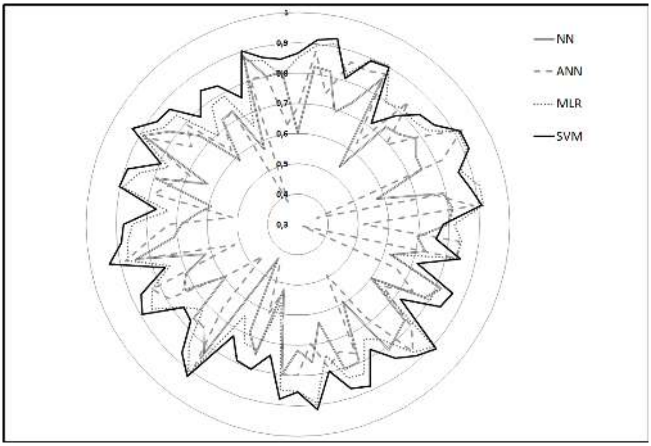

Fig. 2. Results of MLR, NN, ANN, and the averaged best technique (SVM) in terms of R^2 for the 60 datasets

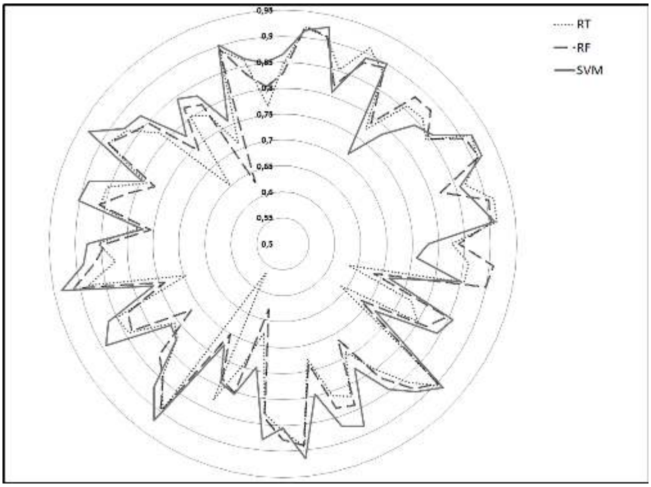

Fig. 3. Results of RT, RF, and the averaged best technique (SVM) in terms of R^2 for the 60 datasets

Finally, an issue not covered in this study and that should be considered in future studies is the influence of the parameters on the results. This point as feature selection will be addressed in future work.

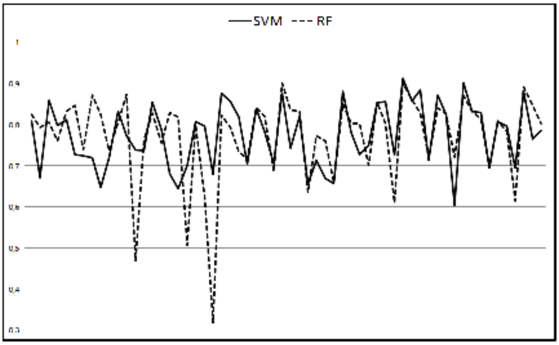

Fig. 4. Results for every dataset in terms of R^2 of Random Forests and SVM when no feature selection was applied

4 Conclusions

This paper presented a comparison between common regression techniques in machine learning (ANN, RT, SVM, NN, and ensembles such as RF) and the classic MLR-based methodology. The selected techniques were applied to real LiDAR data from two areas in the province of Lugo (Galizia, Spain). The results showed that support vector regression statistically outperformed the rest of techniques when feature selection is applied but its performance could not be said statistically different from that of Random Forests when feature selection was skipped. Nevertheless, results confirmed recent bibliography since SVM and RF behaved the best for the 60 experimental datasets.

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. In the same line, parametrization will have to be addressed as another important issue which can change the results of the predictors. 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 uses.

References

1. Garcia, M., Riano, D., Chuvieco, E., Danson, F.M.: Estimating biomass carbon stocks for a mediterranean forest in central spain using LiDAR height and intensity data. Remote Sensing of Environment 114(4), 816–830 (2010)
2. Mutlu, M., Popescu, S.C., Stripling, C., Spencer, T.: Mapping surface fuel models using LiDAR and multispectral data fusion for fire behavior. Remote Sensing of Environment 112(1), 274–285 (2008)
3. Gonzalez-Ferreiro, E., Dieguez-Aranda, U., Gonçalves-Seco, L., Crecente, R., Miranda, D.: Estimation of biomass in eucalyptus globulus labill. forests using different LiDAR sampling densities. In: Proceedings of ForestSat (2010)

4. Muss, J.D., Mladenoff, D.J., Townsend, P.A.: A pseudo-waveform technique to assess forest structure using discrete LiDAR data. Remote Sensing of Environment 115(3), 824–835 (2010)
5. Osborne, J., Waters, E.: Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research and Evaluation 8(2) (2002)
6. Salas, C., Ene, L., Gregoire, T.G., Næsset, E., Gobakken, T.: Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models. Remote Sensing of Environment 114(6), 1277–1285 (2010)
7. Hudak, A.T., Crookston, N.L., Evans, J.S., Halls, D.E., Falkowski, M.J.: Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LIDAR data. Remote Sensing of Environment 112, 2232–2245 (2008)
8. Chen, G., Hay, G.J.: A support vector regression approach to estimate forest biophysical parameters at the object level using airborne lidar transects and quickbird data. Photogrammetric Engineering and Remote Sensing 77(7), 733–741 (2011)
9. Zhao, K., Popescu, S., Meng, X., Pang, Y., Agca, M.: Characterizing forest canopy structure with lidar composite metrics and machine learning. Remote Sensing of Environment 115(8), 1978–1996 (2011)
10. Latifi, H., Nothdurft, A., Koch, B.: Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: Application of multiple optical/LiDAR-derived predictors. Forestry 83(4), 395–407 (2010)
11. Gleason, C.J., Im, J.: Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sensing of Environment 125, 80–91 (2012)
12. Goncalves-Seco, L., Gonzalez-Ferreiro, E., Dieguez-Aranda, U., Fraga-Bugallo, B., Crecente, R., Miranda, D.: Assessing attributes of high density eucalyptus globulus stands using airborne laser scanner data. International Journal of Remote Sensing 32(24), 9821–9841 (2011)
13. Gonzalez-Ferreiro, E., Dieguez-Aranda, U., Miranda, D.: Estimation of stand variables in pinus radiata d. don plantations using different lidar pulse densities. Forestry 85(2), 281–292 (2012)
14. Dieguez-Aranda, U., et al.: Herramientas selvicolas para la gestion forestal sostenible en Galicia. Xunta de Galicia (2009)
15. McGaughey, R.: FUSION/LDV: Software for LIDAR Data Analysis and Visualization. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Seattle (2009)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1) (2009)
17. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory 14, 55–63 (1968)
18. Parejo, J.A., García, J., Ruiz-Cortés, A., Riquelme, J.C.: Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas. In: Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados (2012)
19. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
20. Luengo, J., Garcia, S., Herrera, F.: A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests. Expert Systems with Applications 36, 7798–7808 (2009)