A Comparative Study between Two Regression Methods on LiDAR Data: A Case Study

Jorge García-Gutiérrez[1], Eduardo González-Ferreiro[2], Daniel Mateos-García[1], Jose C. Riquelme-Santos[1], and David Miranda[2]

[1] Department of Computer Languages and Systems, University of Seville, Reina Mercedes s/n, 41012 Seville, Spain
{jgarcia,mateos,riquelme}@lsi.us.es
[2] Land Laboratory - Department of Agroforestry Engineering, University of Santiago de Compostela, Campus Universitario s/n, 27002 Lugo, Spain
{eduardo.gonzalez,david.miranda}@usc.es

Abstract. Airborne LiDAR (Light Detection and Ranging) has become an excellent tool for accurately assessing vegetation characteristics in forest environments. Previous studies showed empirical relationships between LiDAR and field-measured biophysical variables. Multiple linear regression (MLR) with stepwise feature selection is the most common method for building estimation models. Although this technique has provided very interesting results, many other data mining techniques may be applied. The overall goal of this study is to compare different methodologies for assessing biomass fractions at stand level using airborne LiDAR data in forest settings. In order to choose the best methodology, a comparison between two different feature selection techniques (stepwise selection vs. genetic-based selection) is presented. In addition, classical MLR is also compared with regression trees (M5P). The results when each methodology is applied to estimate stand biomass fractions from an area of northern Spain show that genetically-selected M5P obtains the best results.

Keywords: Tasmanian blue gum, *Eucalyptus globulus*, remote sensing, regression trees, multiple linear regressions, stand biomass estimation.

1 Introduction

In order to guarantee forest sustainability, it is vital to consider both the economic and ecological functions of forests. Therefore, it is necessary to quantify existing resources for the strategic, tactical and operational planning of silvicultural treatments and forest operations. In the case of the forest biomass, it provides an indication of carbon sequestration in trees and an estimate of cellulosic material as a potential source of renewable energy [1].

Many forest management planning systems are based on the use of stand mean values of biophysical variables [2] often measuring in field. However, stand

variables characterization and quantification methods are very expensive, time-consuming [3] and limited by the cost of establishing sufficient sample plots to capture the existing variability [4]. Furthermore, biomass estimation often involves destructive sampling [1]. In this context, the use of Airborne Laser Scanning (ALS), also referred to as Light Detection and Ranging (LiDAR), has been explored to reduce costs transforming the way change detection and forest mensuration is performed. LiDAR is a remote laser-based technology that can determine the distance from the source placed on an aerial platform to an object or surface providing not only X-Y position, but also the returned energy (laser intensity) and 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.

A very important subset of forest applications like forest inventories [5], biomass estimation [6] or fuel models [7] are based on the estimation of variables in order to build models. If LiDAR is being used, those variables will usually be estimated by multiple linear regression (MLR) between field measurements and LiDAR metrics. The main advantage of using MLR is the simplicity of the resulting model. In contrast, the selected method also has some drawbacks: in most studies, the regression employs a suite of frequency-based metrics calculated from the previous LiDAR height and intensity data, which are systematically eliminated from a full model using a stepwise process which results in a set of predictors with little physical justification [8]. Thus, the methodologies to build regressions between some key variables for forest characterization and LiDAR data are being reviewed [9]. Moreover, new non-parametric techniques and genetic algorithms applied to the predictor selection [10] have been used [11] improving the results but also losing part of the linear regression model's simplicity and clarity.

To the best of our knowledge, the joint use of genetic algorithms and regression trees has not been accurately exploited in the context of biomass estimation since they can maintain the simplicity of stepwise-selected MLR improving its performance. Thus, two comparisons are presented in this work. First, the traditional stepwise selection is compared with a genetic feature selection. Then, a comparison between MLR and a M5P regression tree [12] both genetically-selected is also proposed. These comparisons aim to fulfill three objectives:

- Show the higher level of accuracy when genetic algorithms are applied in lieu of the classical stepwise feature selection.
- Show the improvement on the regression quality when more complex data mining techniques such as M5P replace MLR.
- Establish a solid study to back the exploration of new improvements in regression trees in order to enhance LiDAR products.

The rest of the paper is organized as follows. Section 2 provides a description of the real data used in this work, highlighting the final selected features. Section 3 describes the methodology used. The results achieved are shown in Section 4 and, finally, Section 5 is devoted to summarising the conclusions and to discussing future lines of work.

2 Data Description

The study area is located in the north of Galicia in the northwest of Spain (see Figure 1). The LiDAR data covered 4 km^2 of high density *Eucalyptus globulus* plantations and were acquired in November 2004.

Fig. 1. Image of the study area located in Trabada in the northern region of Galicia (Spain). In blue, the areas flown. In green, the centroids of the inventory plots.

A forest inventory of 39 square plots of 15 m^2 was conducted in mature *Eucalyptus globulus* plantations in February and March 2005. From that fieldwork, crown biomass (W_{cr}), stem biomass (W_{st}) and aboveground biomass (W_{abg}) were calculated.

A set of common metrics in literature [13,14,15] were calculated from the normalized intensity and height values of LiDAR data collected within the limits of the 39 field plots. These metrics are used as independent variables in the regression models whilst W_{cr}, W_{st} and W_{abg} are selected as dependent variables.

3 Method

In order to select the best predictors, a genetic feature selection from LiDAR metrics is carried out. A deeper description of the genetic algorithm (GA) and its characteristics is provided in the following paragraphs along with a brief description of the types of regression used in this study.

3.1 Initial Population

To execute the genetic algorithm, an individual representation is required. In this case, an individual of the population is an array whose cells each represent a weight for each feature in the training set. Each weight is initialized with a value of 1 or 0. Thus, if the corresponding feature is selected, the weight will be 1, otherwise 0.

The size of the population and the number of generations are genetic algorithm parameters which are set up with values of 200 and 100 respectively in this work.

These parameters were empirically selected and proved to reach the best results. In addition, every simple linear regression is part of the initial population which involves to start from the best possible minimum model.

3.2 Fitness Function

The fitness function for discriminating individuals who best fit each generation is based on the coefficient of determination R^2 which measures the adjustment with the training data. This value fluctuates between 0 and 1. The higher R^2, the better the individual.

A related problem with the simplicity of regression models is multicollinearity. The control of this detrimental effect is performed using the condition number as a threshold. The condition number is associated with the eigenvalues of the matrix built by the features selected in the individual. Moreover, it is well-known that a condition number that exceeds a value of 30 involves a high degree of multicollinearity. In this way, every individual with a condition number of 30 or higher is assigned a fitness value of 0.

3.3 Crossover and Mutation

In the design of a GA, it is always important to establish a coherent search criterion in the space of possible solutions. This can only be achieved with a proper selection of crossover and mutation operators.

A random crossover operation for two individuals (parents) selected by the roulette-wheel method is applied. The crossover selects a gen (weight) for each feature from two possible values (the parents values associated to the corresponding feature) randomly. In the end, the final set of genes is assigned to the new individual.

The mutation operator has been defined to change the value of a weight according to a probability.In our case a value of 0.1 was empirically selected. A mutation involves changing a gen value for its complementary (1 into 0 and vice versa).

3.4 Regression Models

Linear and allometric models were used to establish empirical relationships between field measurements and LiDAR variables. Their general expressions can be seen in Equation 1 and 2 respectively.

$$Y = \beta_0 + \beta_1 \, X_1 + \beta_2 \, X_2 + \ldots + \beta_n \, X_n \tag{1}$$

$$Y = \beta_0 \, X_1^{\beta_1} \, X_2^{\beta_2} \, \ldots \, X_n^{\beta_n} \tag{2}$$

where Y are field values of W_{cr} (kg ha^{-1}), W_{st} (kg ha^{-1}), and W_{abg} (kg ha^{-1}), and $X_1, X_2, \ldots X_n$ may be variables related to the metrics of heights and pulse intensities distributions or measurements related to canopy closure [13].

4 Results

In this work, two well-defined comparisons are proposed. In the first, we compare classical stepwise and genetic-based feature selections. In the second, after genetically selecting the best predictors, a comparison between the classical MLR and M5P regression tree is established. For both methods, we use the WEKA framework [16] to generate the results.

As mentioned previously, checking whether the classical stepwise process on the possible predictors is enhanced by a genetically-based feature selection is one of the main objectives of this work. Due to the random nature of genetic algorithms and in order to establish the comparison, the execution of the genetic algorithms was repeated thirty times and the averaged values were taken. In Table 1, the selected predictors collected by each method can be seen beside the coefficient of determination R^2 achieved by the MLR technique for each case.

For every biomass dependent variable (W_{cr}, W_{st}, W_{abg}) an improvement is reached for both allometric and linear models when genetic selection is applied.

Table 1. Prediction capacity (coefficient of determination, R^2) for MLR when a stepwise and a genetic selection are respectively applied

Variable	Stepwise R^2	Stepwise predictors	Genetic R^2	Genetic predictors
allometric W_{cr}	0.619	h_{60}	0.759	h_{75}
allometric W_{st}	0.740	h_{60}	0.863	h_{75}
allometric W_{abg}	0.727	h_{60}	0.853	h_{75}
linear W_{cr}	0.708	h_{90}	0.753	h_{90},i_{mode},i_{ID}
linear W_{st}	0.801	h_{SKw},h_{75}	0.814	h_{SKw},h_{75}, i_{mode}, i_{70}
linear W_{abg}	0.771	h_{95}	0.809	h_{min},h_V,h_{75}

The next step consists in comparing the classical MLR and M5P regression tree generator when a GA is applied to make the feature selection. It is important to outline that the fitness function chosen for the GA optimizes the use of MLR so M5P regression tree starts with some disadvantage. Anyway, as seen in Table 2, M5P gets the same value of R^2 as MLR in the worst case, overcoming MLR in two out of six tests.

To statistically validate the differences between MLR and M5P, a test of statistical significance is needed. Since the real data is too small (just 39 instances in only one dataset), the study has to be built from other sources. Thus, the 10-fold cross-validation results on 27 well-known datasets [17] are collected. Once the coefficients of determination for every dataset are obtained for both methods (see Table 3), it is possible to establish a statistical analysis of their prediction capacity. Traditionally, parametric statistical tests such as ANOVA are applied for this type of analysis. However, for a comparison of these types of tests to be correct, the data must meet the criteria of independence, normality, and homoscedasticity [18]. Through a D'Agostino-Pearson test [19], it could thus be

Table 2. Prediction capacity (coefficient of determination, R^2) of genetically-selected MLR and M5P respectively

Variable	MLR averaged R^2	M5P averaged R^2
allometric W_{cr}	0.759	0.759
allometric W_{st}	0.863	0.863
allometric W_{abg}	0.853	0.853
linear W_{cr}	0.753	0.753
linear W_{st}	0.814	0.820
linear W_{abg}	0.809	0.826

confirmed that the data obtained for this study did not meet the criteria of normality. For this reason, a non-parametric approximation (Wilcoxon test) was selected [20]. The p-value results in a value less than 0.0001 so it can be said that differences between the methods are statistically significative (at $\alpha = 0.05$).

Having found that the number of wins is higher for M5P (5 for MLR and 12 for M5P) and knowing their differences are statistically significative, we can conclude that M5P outperforms MLR.

Table 3. Prediction capacity (coefficient of determination, R^2) of genetically-selected MLR and M5P respectively when both methods are applied to 27 datasets

Dataset	MLR R^2	M5P R^2	Dataset	MLR R^2	M5P R^2
auto93.arff	0.631	0.631	autoHorse.arff	0.801	0.801
autoMpg.arff	0.698	0.736	autoPrice.arff	0.808	0.823
bodyfat.arff	0.976	0.978	breastTumor.arff	0.000	0.000
cholesterol.arff	0.049	0.044	echoMonths.arff	0.124	0.124
housing.arff	0.636	0.830	hungarian.arff	0.302	0.298
kdd_coil_train1.arff	0.298	0.470	kdd_coil_train2.arff	0.164	0.164
kdd_coil_train3.arff	0.115	0.115	kdd_coil_train5.arff	0.114	0.114
kdd_coil_train6.arff	0.120	0.120	kdd_coil_train7.arff	0.066	0.066
kdd_el_ninosmall.arff	0.793	0.811	machine_cpu.arff	0.865	0.946
meta.arff	0.110	0.075	pbc.arff	0.266	0.305
pharynx.arff	0.000	0.000	pyrim.arff	0.752	0.718
quake.arff	0.006	0.040	stock.arff	0.532	0.746
strike.arff	0.098	0.234	triazines.arff	0.318	0.487
wisconsin.arff	0.219	0.209			

5 Conclusions

LiDAR technology has become an important tool for carrying out several important tasks for the natural environment and, in particular, for biomass estimation. Lately, biomass estimation models have been built by means of LiDAR data processing. In this work, two different comparisons were established when regression techniques were applied to LiDAR data. First, a comparison between a genetically-based and a classical stepwise feature selection was presented. The

study concluded that the GA outperformed the stepwise process when MLR was built using each set of selected predictors. Then, from a genetically-based feature set, two regression methods were tested: classical MLR and M5P regression trees. In this case, the results showed that M5P obtained better results when both methods were applied to real data from Galicia (Spain).

According to the results, new intelligent techniques applied to regression trees can be explored to improve the results when applied to biomass estimation. With this purpose, evolutionary computation could be used to overcome some M5P limits, optimizing the predictor selection and controlling the thresholds of the regression tree branches. Furthermore, a more in-depth comparison of regression trees with other non-parametric methodologies (support vector machines, neural networks) is required. Finally, an important aspect not explored in this work is the ability of regression trees for detecting the most important predictors (regression tree roots) which should be developed in future research.

References

1. Popescu, S.C.: Estimating biomass of individual pine trees using airborne lidar. Biomass and Bioenergy 31, 646–655 (2007)
2. Naesset, E., Gobakken, T., Holmgren, J., Hyyppä, H., Hyyppä, J.: Laser scanning of forest resources: the nordic experience. Scand. J. Forest. Res. 19, 482–499 (2004)
3. Hall, S., Burke, I., Box, D., Kaufmann, M., Stoker, J.: Estimating stand structure using discrete-return lidar: an example from low density, fire prone ponderosa pine forests. Forest. Ecol. Manag. 208, 189–209 (2005)
4. Lovell, J., Jupp, D., Newnham, G., Coops, N., Culvenor, D.: Simulation study for finding optimal lidar acquisition parameters for forest height retrieval. Forest. Ecol. Manag. 214, 398–412 (2005)
5. Anderson, J.E., Plourde, L.C., Martin, M.E., Braswell, B.H., Smith, M.L., Dubayah, R.O., Hofton, M.A., Blair, J.B.: Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. Remote Sensing of Environment 112(4), 1856–1870 (2008)
6. 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)
7. 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)
8. Muss, J.D., Mladeno, D.J., Townsend, P.A.: A pseudo-waveform technique to assess forest structure using discrete lidar data. Remote Sensing of Environment (2010) (in Press)
9. 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)
10. Gong, B., Im, J., Mountrakis, G.: An artificial immune network approach to multisensor land use/land cover classification. Remote Sensing of Environment 115(2), 600–614 (2011)
11. 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)

12. Quinlan, R.J.: Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348 (1992)
13. González-Ferreiro, E., Diéguez-Aranda, U., Gonçalves-Seco, L., Crecente, R., Miranda, D.: Assessing biomass in Eucalyptus globulus plantations in Galicia using different LiDAR sampling densities. In: Miranda, D., Suárez, J., Crecente, R. (eds.) Proceedings of ForestSat 2010: 4th international conference on Operational tools in forestry using remote sensing techniques, Lugo, Spain, September 6-10, pp. 37–41 (2010)
14. Antonarakis, A., Richards, K., Brasington, J.: Object-based land cover classification using airborne LIDAR. Remote Sensing of Environment (112), 2988–2998 (2008)
15. 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)
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. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
18. Zar, J.: Biostatistical Analysis. Prentice-Hall, Englewood Cliffs (1999)
19. Trujillo-Ortiz, A., Hernandez-Walls, R.: DagosPtest: D'Agostino-Pearson's K2 test for assessing normality of data using skewness and kurtosis. A MATLAB file (2003)
20. Cardillo, G.: Wilcoxon test: non parametric wilcoxon test for paired samples (2006)