A Preliminary Study of the Suitability of Deep Learning to Improve LiDAR-Derived Biomass Estimation

Jorge García-Gutiérrez[1](✉), Eduardo González-Ferreiro[2,3], Daniel Mateos-García[1], and José C. Riquelme-Santos[1]

[1] Department of Computer Languages and Systems,
University of Seville, Seville, Spain
{jorgarcia,mateosg,riquelme}@us.es

[2] Sustainable Forest Management Unit, Department of Agroforestry Engineering,
University of Santiago de Compostela, Santiago de Compostela, Spain
edu.g.ferreiro@gmail.com

[3] Department of Forest Ecosystems and Society,
Oregon State University, Corvallis, OR, USA

Abstract. Light Detection and Ranging (LiDAR) is a remote sensor able to extract three-dimensional information about forest structure. Biophysical models have taken advantage of the use of LiDAR-derived information to improve their accuracy. Multiple Linear Regression (MLR) is the most common method in the literature regarding biomass estimation to define the relation between the set of field measurements and the statistics extracted from a LiDAR flight. Unfortunately, there exist open issues regarding the generalization of models from one area to another due to the lack of knowledge about noise distribution, relationship between statistical features and risk of overfitting. Autoencoders (a type of deep neural network) has been applied to improve the results of machine learning techniques in recent times by undoing possible data corruption process and improving feature selection. This paper presents a preliminary comparison between the use of MLR with and without preprocessing by autoencoders on real LiDAR data from two areas in the province of Lugo (Galizia, Spain). The results show that autoencoders statistically increased the quality of MLR estimations by around 15–30%.

Keywords: Deep learning · LiDAR · Regression · Remote sensing · Soft computing

1 Introduction

Light Detection and Ranging (LiDAR) is a remote laser-based technology able to measure the distance from the source to an object or surface in addition to x-y position. LiDAR sensors have transformed the work previously done with expensive or not always-feasible fieldwork. One of the most important disciplines where LiDAR has become an important tool is forestry, specially for estimating forest biomass [1]. Biomass estimation is a key process in forestry management and also, is closely related to climate change since forests are important carbon deposits on Earth [2].

Regarding LiDAR and biomass estimation, researchers have focused on deriving variables related to the LiDAR's ability to extract vertical information and then, worked on establishing relations with field measurements [3]. Multiple linear regression (MLR) has usually been the selected technique to find those relations [4]. The main advantage of using MLR has been the simplicity and clarity of the resulting model, although other techniques have already proven to be more suitable for regression [5].

Regardless the technique selected, main concerns about LiDAR-derived models are related to their suitability to be applied from one area to another (overfitting) [6]. Scale-dependence is well-known in the literature [7] which makes researchers try to work at regional levels [8] in order to facilitate the carbon stocks calculation but with lower accuracy in the models. Regardless the selected scale, models have also problems related to noise in the LiDAR signal [9] which makes researchers avoid intensity (an important component in LiDAR data related to reflectance of the objects) because of the difficulties to well calibrate and sensitivity to multiple returns (which are often found in forested areas) [10]. All these problems are translated to the classical framework where MLR is applied to develop biomass estimation models. In this context, there is a need to explore new ways to reduce noise in LiDAR data whilst at the same time, clarify the relation between LiDAR-derived statistical features.

In recent times, deep learning has emerged as an important tool in machine learning. Deep learning is defined as a branch of machine learning that exploit layers of non-linear information (in the fashion of neural networks), supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification [11]. Autoencoders are a special type of deep neural network, whose output vectors are a (a presumably better) representation of the original input. They are often used for learning an effective or encoding representation of the original data as input vectors in the hidden layers. Vincent et al. [12,13] showed that the use of autoencoders to denoise training data could lead to an improvement of the performance of classification/regression tasks.

Our aim in this work is to compare the results of the classical MLR for biomass estimation with and without a preprocessing based on the use of autoencoders. Experiments carried out in two different areas of the province of Lugo (Galizia, Spain) are statistically validated and discussed.

The rest of the paper is organized as follows. Section 2 provides a description of the LiDAR data used in this work and the methodology followed to carry out the experimentation in Sect. 2. The results achieved, their statistical validation and the main findings are shown in Sect. 3. Finally, Sect. 4 is devoted to summarize the conclusions and to discuss future lines of work.

2 Materials and Method

2.1 Experimental Datasets

Aerial LiDAR data used for this study was flown in two forest areas in the northwest of the Iberian Peninsula (Fig. 1, more details about both areas can be found in Gonzalez-Ferreiro et al. [14]).

Field data from the two study sites were collected to obtain the dependent variables. 39 and 54 instances (one per training plot in a study site) were located and measured on site A and B, respectively. From every study site, two different datasets were generated. One with the maximum resolution provided for the flight and another with a lower level of 0.5 pulses × m^{-1}.

Biomass fractions of each tree were estimated according to the field measurements (heights and diameters) and the equations for Eucalyptus globulus (site A) and Pinus radiata (site B) reported by [15]. For every plot, biomass fractions were used to calculate the following stand variables in per hectarea basis: stand

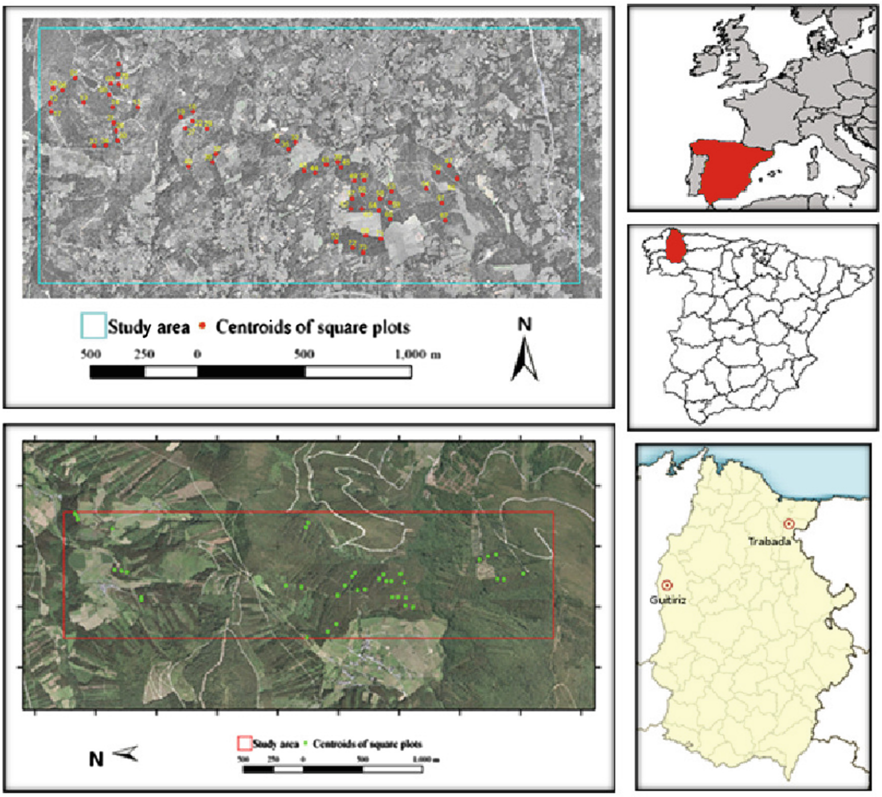

Fig. 1. Study sites located in NW of Spain. Bottom: study site of Trabada (site A). Top: study site of Guitiriz (site B).

crown biomass (W_{cr}), stand stem biomass (W_{st}) and stand aboveground biomass (W_{abg})for both sites A and B. Additionally, stand basal area (G), dominant height (H_d), mean height (H_m) and stand volume (V) were computed for site B.

FUSION software [16] also provided the variables related to the height and return intensity distributions within the limits of the field plots. 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 the experimental datasets. Each dataset was compose of the 48 independent variables ($coverFP$ and $returns$ in Table 1 plus the rest calculated for LiDAR intensity and heights distributions) extracted from LiDAR data and a dependent variable

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

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

(fieldwork-derived forest variable). This procedure gave a total of 20 datasets (obtained for the ten biophysical variables and the two different resolutions of LiDAR-derived feature extraction).

2.2 Autoencoders

We used a traditional autoencoder model [17] to improve MLR performance. An autoencoder is a type of neural network which tries to learn an identity function defining a code/decode transformation. Thus, it firstly takes an input vector x and maps it to a hidden representation y through a deterministic mapping $y = f(x) = s(Wx+b)$, parametrized by W, b. W is a weight matrix and b is a bias vector. The resulting latent representation y is then mapped back to a *reconstructed* vector z in input space $z = g(y) = s(W'y + b')$. Every instance $x(i)$ is thus mapped to a corresponding $y(i)$ and a reconstruction $z(i)$ (see Fig. 2). The parameters of both transformations are calculated to minimize an error function between input and output usually based on the traditional squared error function.

In this work, the selected implementation of the autoencoder was obtained from Weka [18] source repository. The autoencoder was implemented as an unsupervised filter with two optional extra steps (normalization and standardization). We worked with just one hidden layer (since single-hidden-layer neural networks are universal approximators [19]) and the number of units in the hidden layer was set up to the number of features plus one (non-linear autoencoders with more hidden units than inputs have experimentally yielded useful representations [20]).

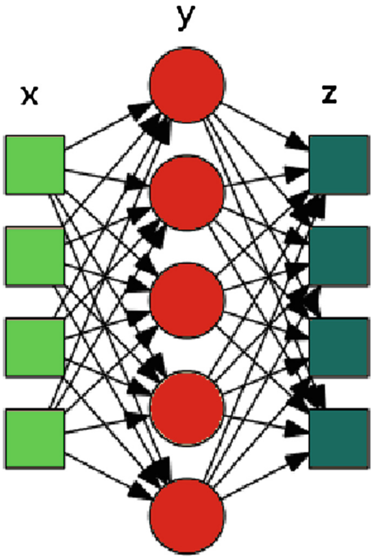

Fig. 2. Autoencoder structure.

2.3 Experimental Framework

In this work, we tested MLR with and without autoencoder preprocessing, applied to the 20 datasets with the 48 independent variables.

Coefficient of correlation (R) was selected to establish the comparison as was done in recent bibliography [21] although we also included root mean square error ($RMSE$) in the results section for information purpose. The coefficients were obtained in a process of 5-fold cross-validation on each dataset repeated five times. The mean value of the five repetitions was recorded for each technique and dataset in order to obtain robust results (independent from the random selection of folds).

Feature selection based on the well-known M5' filter from Weka was applied to avoid the Hughes phenomenon [22] in all the cases. Optionally, normalization and standardization were applied. Thus, experiments were repeated for every dataset and each compared technique (simple MLR and MLR with autoencoder) with no extra filtering, normalization, and standardization. Then, we selected the best R obtained by each technique and every dataset regardless the optimal filtering setup.

Finally, a T-student or Wilcoxon test (depending on whether or not parametric conditions are met) was applied to statistically validate the results [23].

3 Results

The results obtained by each option can be found in Table 2. If R is used as the reference measure of quality, we can observe that in 14 out of 20 cases, the use of an autoencoder improved the results obtained by MLR. If $RMSE$ is the quality measure, the results are even better (18 out of 20).

After the generation of the quality results for the models, a statistical analysis was applied by using the open-source platform StatService [24] to check the significance in the differences in terms of R. T-student test is usually used for pairwise comparison of results if parametric conditions (homoscedasticity, independence, normality) are met [25]. However, Shapiro-Wilk test rejected the normality hypothesis of the results with a p-value under 0.026 for an $\alpha = 0.05$, and therefore, a non-parametric test such as Wilcoxon's was selected.

Wilcoxon test firstly obtains 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 model was the best for a test case, while a rank of 2 would mean it was the worst. Later, the test statistically validates the differences in the mean ranks. In our case, MLR without autoencoding obtained a mean ranking of 1.7 and with the use of an autoencoder reached a mean ranking of 1.1. Taking into account that the Wilcoxon statistic was 164.5 with 1 and 18 degrees of freedom and its corresponding p-value was 0.024, so we could state that the use of an autoencoder significantly improved MLR performance under an $\alpha = 0.05$.

Regarding the results obtained, we can observe that when an autoencoder is applied the improvement of $RMSE$ is around 15 % in average whilst it is 30 %

Table 2. Results obtained by MLR with (column 'Auto') and without (column 'Simple') autoencoding. The best in bold.

Site	Resolution	Biophysical variable	R		RMSE	
			Simple	Auto	Simple	Auto
Trabada	0.5	W_{cr}	0.41	**0.64**	7663.42	**5570.10**
		W_{st}	0.4	**0.74**	50424.57	**36029.75**
		W_{abg}	0.44	**0.69**	61899.45	**47395.80**
	4	W_{cr}	0.36	**0.69**	6200.81	**4553.23**
		W_{st}	0.30	**0.81**	51109.90	**31582.79**
		W_{abg}	0.27	**0.77**	56869.20	**36324.06**
Guitiriz	0.5	W_{cr}	**0.65**	0.55	10236.60	**9675.70**
		W_{st}	0.61	**0.72**	65573.57	**46515.49**
		W_{abg}	0.61	**0.69**	62827.15	**59505.88**
		G	0.50	**0.52**	12.58	**11.63**
		H_d	0.71	**0.79**	2.93	**2.84**
		H_m	0.76	**0.78**	2.48	**2.41**
		V	**0.68**	0.67	133.71	**125.00**
	8	W_{cr}	**0.76**	0.53	**9127.03**	9709.39
		W_{st}	**0.71**	0.69	58881.87	**45201.21**
		W_{abg}	0.69	**0.72**	63302.99	**54142.08**
		G	0.60	**0.64**	11.07	**10.63**
		H_d	0.76	**0.81**	3.44	**2.79**
		H_m	**0.78**	0.71	2.51	**2.18**
		V	**0.74**	0.70	**112.57**	121.22

if we compare averaged R. Similar results have reported [12] that demonstrated that the use of autoencoders bring benefits for other machine learning techniques such as Support Vector Machines. Our results show that parametric techniques can also be boosted by this type of preprocessing.

Among the possible reasons for such a good performance of autoencoders, we should outline that LiDAR-derived data is generated in several steps which may involve noise generation. Vincent et al. [13] showed how to develop denoising autoencoders which focus in noise reduction. Although the one applied in this work cannot be seen as a complete denoising autoencoder, the benefit could still be present due to the own nature of autoencoders which establish a mechanism to code/decode data avoiding spurious influence.

Also, notice that LiDAR-derived statistical features are usually limited and hand-crafted. These features might not have to completely describe the complexity in a forest area. In that case, relations between statistics could be explored by an autoencoder in order to find a better data representation by features combination. Taking into account that autoencoders take almost-negligible training time for small datasets (a common situation in forestry due to fieldwork high costs), their use is strongly advised.

Finally but not less important, autoencoders could have also decreased the risk of overfitting introducing some degree of distortion in training data although this possibility needs a deeper study to be confirmed.

4 Conclusions

This paper presented a preliminary study of the use of autoencoders to improve LiDAR-derived biomass estimation by MLR. The experimentation was carried out on real data from two areas in the province of Lugo (Galizia, Spain). The results showed that autoencoders statistically improved the use of classical MLR for biomass estimation. Nevertheless, results confirmed that autoencoders are a valuable tool to preprocess LiDAR-derived features by getting noise reduction and feature discovering.

Future work should address gaps not covered in this work. Thus, we must complete the framework with a deeper comparison with other types of autoencoders (denoising encoders, sparse encoders, etc.) and regression techniques (regression trees, Gaussian processes, etc.). We should also study the influence of the number of hidden units. Finally, we must assess autoencoders to avoid overfitting and thus overcome the problems to apply models from one area to another.

References

1. Wulder, M.A., Bater, C.W., Coops, N.C., Hilker, T., White, J.C.: The role of LiDAR in sustainable forest management. Forest. Chron. **84**(6), 807–826 (2008)
2. Hansen, E.H., Gobakken, T., Solberg, S., Kangas, A., Ene, L., Mauya, E., Nsset, E.: Relative efficiency of ALS and InSAR for biomass estimation in a tanzanian rainforest. Remote Sens. **7**(8), 9865 (2015)
3. Gonzlez-Ferreiro, E., Miranda, D., Barreiro-Fernandez, L., Bujan, S., Garcia-Gutierrez, J., Dieguez-Aranda, U.: Modelling stand biomass fractions in galician eucalyptus globulus plantations by use of different LiDAR pulse densities. For. Syst. **22**(3), 510–525 (2013)
4. Muss, J.D., Mladenoff, D.J., Townsend, P.A.: A pseudo-waveform technique to assess forest structure using discrete LiDAR data. Remote Sens. Environ. **115**(3), 824–835 (2011)
5. Garcia-Gutierrez, J., Martinez-Alvarez, F., Troncoso, A., Riquelme, J.: A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. Neurocomputing **167**, 24–31 (2015)
6. Bouvier, M., Durrieu, S., Fournier, R.A., Renaud, J.P.: Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. Remote Sens. Environ. **156**, 322–334 (2015)
7. Zhao, K., Popescu, S., Nelson, R.: LiDAR remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. Remote Sens. Environ. **113**(1), 182–196 (2009)

8. Hayashi, M., Saigusa, N., Yamagata, Y., Hirano, T.: Regional forest biomass estimation using ICESat/GLAS spaceborne LiDAR over borneo. Carbon Manage. **6**(1–2), 19–33 (2015)
9. Cao, N., Zhu, C., Kai, Y., Yan, P.: A method of background noise reduction in LiDAR data. Appl. Phys. B **113**(1), 115–123 (2013)
10. Hofle, B., Pfeifer, N.: Correction of laser scanning intensity data: Data and model-driven approaches. ISPRS J. Photogrammetry Remote Sens. **63**, 1415–1433 (2007)
11. Deng, L., Yu, D.: Deep learning: Methods and applications. Found. Trends Signal Process. **7**, 3–4 (2014)
12. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML 2008), pp. 1096–1103. ACM (2008)
13. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
14. 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)
15. Diéguez-Aranda, U., Rojo-Alboreca, A., Castedo-Dorado, F., González, J.A., Barrio-Anta, M., Crecente-Campo, F., González-González, J., Pérez-Cruzado, C., Rodríguez-Soalleiro, R., López-Sánchez, C., Balboa-Murias, M., Gorgoso-Varela, J.J., Sánchez-Rodríguez, F.: Herramientas selvícolas para la gestión forestal sostenible en Galicia, vol. 259. Consellería do Medio Rural, Xunta de Galicia (2009)
16. McGaughey, R.: FUSION/LDV: software for LIDAR data analysis and visualization. In: US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Seattle (2009)
17. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explor. **11**(1), 10–18 (2009)
19. Hassoun, M.: Fundamentals of Artificial Neural Networks. MIT Press, Cambridge (1995)
20. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Schölkopf, B., Platt, J., Hoffman, T., (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 153–160. MIT Press (2007)
21. Zhao, K., Popescu, S., Meng, X., Pang, Y., Agca, M.: Characterizing forest canopy structure with LiDAR composite metrics and machine learning. Remote Sens. Environ. **115**(8), 1978–1996 (2011)
22. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**, 55–63 (1968)
23. Fay, M., Proschan, M.: Wilcoxon-mann-whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat. Surv. **4**, 1–39 (2010)
24. 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)
25. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)