ARTIFICIAL NEURAL NETWORKS FOR MODELING HYPSOMETRIC RELATIONSHIPS OF *Pinus caribaea* Morelet var. *caribaea* Barr. & Golf.

Ouorou Ganni Mariel GUERA[1]
José Antônio Aleixo da SILVA[1]
Rinaldo L. Caraciolo FERREIRA[1]
Héctor Barrero MEDEL[2]
Daniel Álvarez LAZO[2]

- *ABSTRACT:* The present study was carried out to compare the performances of regression models and Artificial Neural Networks (ANNs) in hypsometric relationships modeling and to analyze the influence of ANN type and sample size on ANN performance. The database was consisted by 65 circular plots of 500 m² in which Diameter at Breast Height - DBH (cm) and Total Height - Ht (m) of 2538 trees were measured in plantations of *Pinus caribaea* var. *caribaea* in Macurije forest company, Cuba. The study was carried out in three stages: i) Adjustment of traditional hypsometric models and sigmoidal growth models; ii) ANNs training and comparison of the selected ANN with the regression model selected; iii) Analysis of sample size and ANN type influences on the estimates precision by means of a completely random experimental design with 5x2 factorial arrangement, with the factors sample size (N) and ANN type (R). The results indicated that the best equation to estimate trees heights was that of Gompertz. The ANNs MLP 1-4-1 and MLP 8-4-1 were superior to the selected equation (Gompertz). Multi-Layer Perceptron ANNs generated more accurate estimates and their performances were less influenced by the sample size.

- *KEYWORDS:* Artificial intelligence; regression analysis; diameter at breast height; total height.

1 Introduction

Different trees heights and diameters at breast height (DBH) are the most widely used dendrometric variables in forest measurement (CAMPOS and LEITE, 2013). Total height is an essential variable for forest surveys and serves as input variables for volume equations, site index (dominant height), taper model (MARTINS *et al*., 2016) and in forest biomass estimation (SANQUETTA *et al*., 2014a).

Tree height can be obtained by direct measurement in the felled tree, or indirectly, using hypsometers or clinometers (MOTTA *et al*., 2016). Such indirect measurement or height estimation at the naked eye as is most commonly done in numerous time-saving

[1]Universidade Federal Rural de Pernambuco - UFRPE, Laboratório de Biometria e Manejo Florestal, Departamento de Ciência Florestal, CEP: 52.171-900, Recife, PE, Brasil. E-mail: *gueraforest@gmail.com*; *jaaleixo@uol.com.br; rinaldo.ferreira@ufrpe.br*
[2]Universidad de Pinar del Río "Hermanos Saiz Montes de Oca", Departamento Forestal, Ciudad de Pinar del Río, Pinar del Río, Cuba. E-mail: *hbarrero@upr.edu.cu; daniel@upr.edu.cu*

forest inventories leads to inaccurate estimates. These indirect measurements are also hampered by visual obstructions in the field (LEI *et al*., 2009; BINOTI *et al*., 2013), which increases errors in measurements. These errors can lead to underestimation or overestimation of this variable and all that depend on it, such as volume, forest biomass, site index, etc.

An alternative of minimizing these errors and reducing significantly forest inventories costs is trees heights estimation by means of hypsometric models that express the *relationship* between tree *height* (*h*) and tree *diameter at breast height* (*dbh*) (SANQUETTA *et al*., 2009) and are based on regression analysis technique that many years of theory and application consecrate as a powerful tool for prediction (MARKHAM and RAKES, 1998). These hypsometric relationships, proposed by Ker and Smith (1957), are defined by Aldana (2010) as being the regression of height over diameter in a stand at a certain age. The equations obtained from these models adjustment are of great practical utility (ALDANA, 2010; AZEVEDO *et al*., 2011) for increasing surveys productivity and economic (MACHADO *et al*., 1993) for reducing forest inventories costs. Traditional hypsometric models, also known as local models, use *dbh* as a predictor variable, and the inclusion of any other variable such as age or site index generates models known mostly as generalist models (CAMPOS and LEITE, 2013).

In recent decades, techniques such as Artificial Neural Networks (ANNs) have been successfully used in the field of forest modeling (MARKHAM and RAKES, 1998, OZÇELIK *et al*., 2013), including hypsometric relationships (BINOTI *et al*. 2013). ANNs are mathematical models that have as metaphor, human brain functioning with its biological neural networks (VALENÇA, 2010). Among the studies that applied ANNs techniques for hypsometric relationships modeling, the researches of Binoti et al. (2013), Vendruscolo *et al*. (2015) and Martins et al. (2016) in *Eucalyptus sp*. plantations; that of Ozçelik *et al*. (2013) in *Crimean Juniper* plantations; that of Almeida (2015) in *Araucaria angustifolia* plantations and that of Campos *et al*. (2016) in plantations of *Eucalyptus grandis*, *Eucalyptus urophylla*, *Pinus caribaea* var. *hondurensis* and *Pinus oocarpa* are examples of successful applications of ANNs in hypsometric relationships modeling in forest plantations. In the results of these studies, ANNs presented performances superior or equal to those of traditional regression models. ANNs being statistical models, the sensitivity of their performances to sample size is an axiom. However, issues related to sample size influence on the performance of each type of neural network, mainly in forest modeling, should be studied in greater depth.

Based on the above, the present study was carried out with the objective of performing a comparative analysis of the capacities of regression models and ANNs in height estimation in plantations of *Pinus caribaea* Morelet var. *caribaea* Barr. & Golf. and to analyze the influence of neural network type and the sample size on the precision of the estimates made.

2 Materials and methods

2.1 Characterization of the study area

The present study was carried out in Macurije Forest Company located in the westernmost region of the province of Pinar del Rio, Cuba (Figure 1).

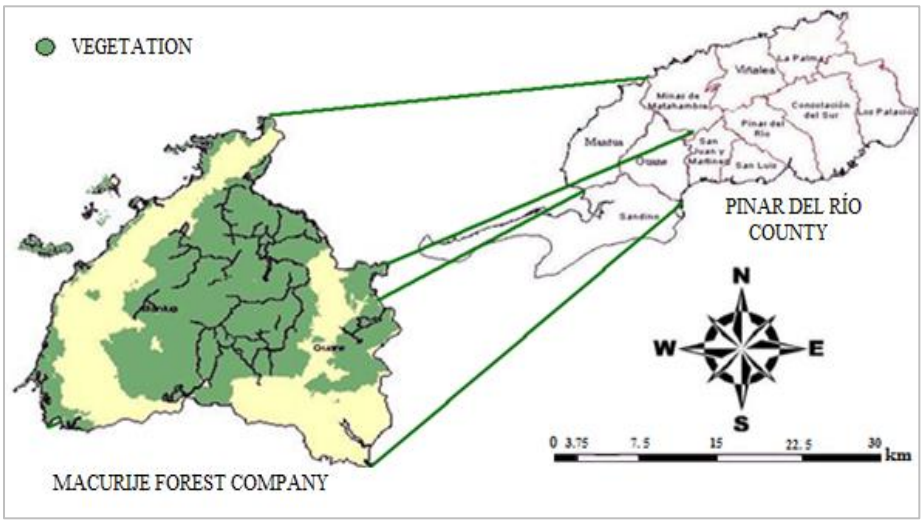

Figure 1 - Geographical location of Macurije Forest Company, Pinar del Río, Cuba.

2.2 Data collection and analysis of sample sufficiency

The data used in the modeling were constituted by 2538 trees (Table 1) measured in 65 circular plots of 500 m² raised in forest inventories made by means of a simple random sampling in *Pinus caribaea* var. *caribaea* plantations. The variables measured in each tree were: total height (*Ht*) and Diameter at Breast Height (*dbh*). Sample sufficiency was analyzed for an allowable error of 5% and 95% of probability, with expressions 01 and 02 (SANQUETTA *et al*., 2014b) given by

$$E_M(\%) = \frac{t_\alpha * S_{\bar{x}}}{\bar{X}} * 100 \quad (1)$$

and

$$n = \frac{t_\alpha^2 * S_x^2}{E^2} * 100, \quad (2)$$

where: E_M= Sampling error; n = Sample size; t_α= value of t student; $S_{\bar{x}}$= Standard Error of the Mean; S_x^2 = variance; $E = LE * \bar{x}$, LE = Limit of error allowed; $\bar{X}$= average of observed heights.

Table 1 – Modeling data sample distribution by diameter (*dbh)* and height (H) class

DBH (cm) Classes	Ht (m) classes					Total
	2.86 -7.85	7.86-12.85	12.86-17.85	17.86-22.85	22.86-27.85	
2.41-7.40	223	15	1			239
7.41-12.40	221	217	17	3		458
12.41-17.40	16	316	323	20	1	676
17.41-22.40	1	149	594	96	12	852
22.41-27.40	1	19	173	60	17	270
27.41-32.40		3	14	11	7	35
32.41-37.40				3	5	8
Total	462	719	1122	193	42	2538

Height-DBH relationship described by Figure 2 indicates, as would be expected in the case of data consistency, a direct proportionality between *total height* and *dbh* growth.

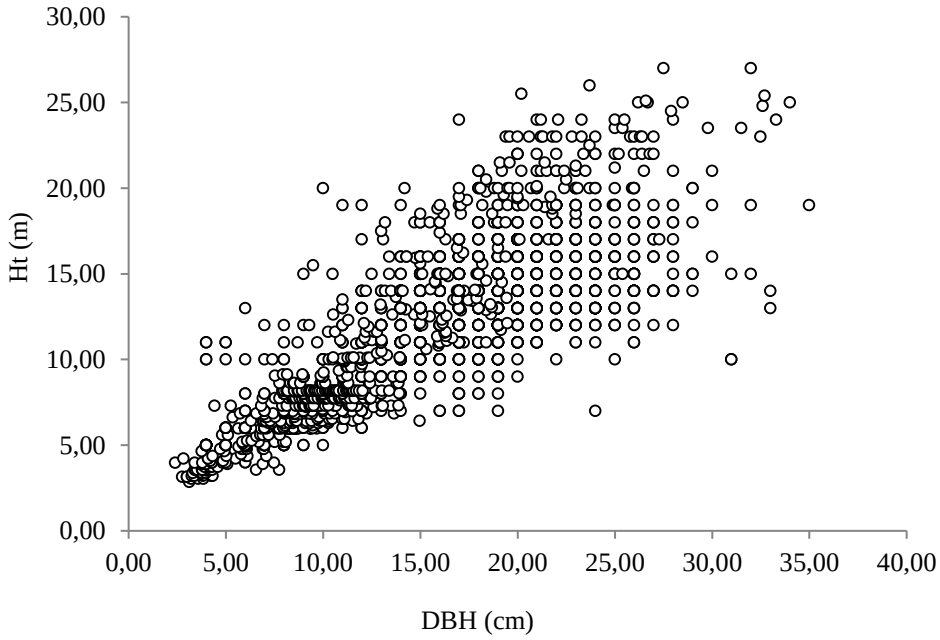

Figure 2 - Total height (*Ht*) - diameter at breast height (*dbh*) relationship in *Pinus caribaea* Morelet var. *caribaea* Barr. & Golf. plantations, Macurije Forest Company, Pinar del Río, Cuba.

2.3 Traditional regression models for hypsometric relationships modeling

The hypsometric equations for *Pinus caribaea* var. *caribaea* were obtained by fitting the models in Table 2.

Table 2 – Hypsometric models adjusted for height estimation

Models	Authors	Mathematical Expressions
01	Logistic/Verhulst (1838)*	$H_i = \frac{\beta_0}{1 + exp[(\beta_1 - DBH_i)/\beta_2]} + \varepsilon_i$
02	Gompertz (1832)*	$H_i = \beta_0\left[exp[-\beta_1 exp(-\beta_2 DBH_i)]\right] + \varepsilon_i$
03	Chapman-Richards/ Pienaar-Turnbull (1973)*	$H_i = \beta_0(1 - e^{-\beta_1 DBH_i})^{\beta_2} + \varepsilon_i$
04	Henricksen (1950)	$H_i = \beta_0 + \beta_1 LnDBH_i + \varepsilon_i$
05	Curtis (1967)	$Ln\ (H_i) = \beta_0 + \beta_1(\frac{1}{DBH_i}) + \varepsilon_i$
06	Stoffels – Soest (1953)	$Ln\ (H_i) = \beta_0 + \beta_1 Ln(DBH_i) + \varepsilon_i$

H_i = Total height; DBH_i= Diameter at breast height;β_0; β_1; β_2 = parameters to be estimated; ε_i = random error. * In those models, β_0 is the asymptotic parameter that corresponds to tree height at maturity; β_1 corresponds to the location parameter, without biological interpretation and β_2 determines the rate of expansion of height growth.

The linear models were adjusted using ordinary least squares method and the nonlinear ones using Levenberg-Marquardt or Gauss-Newton iterative methods.

2.4 Artificial Neural Network training for height estimation

The database were divided according to the following proportions recommended by Valença (2010): 50% for trainings, 25% for test and the remaining 25% for cross validation. To increase the efficiency of the backpropagation algorithm, the data were normalized by linear transformation to the intervals [0, 1] or [-1, 1] according to the activation function used. The functions tested were: Sine, Identity, Logistic, Exponential and Hyperbolic Tangent. The software SPSS version 20.0, Assistat version 7.7, Statistica version 8.0, DataFit 9 and Minitab 17.0. were used for statistical analyzes, regression models fits and ANNs trainings.

2.5 Criteria for evaluation and selection of regression models and ANNs

Every model has to undergo an evaluation or validation process to define the limits of its predictive capacity (ALDER, 1995). Models evaluation criteria used in the present study were: Adjusted coefficient of determination - R_{aj}^2 (%) (Equation 3), Root Mean Square Error - $RMSE$ (%) (Expression 04), Visual analysis of residuals distribution graphic (Expression 05), Bias (Expression 06), Standard error of the estimate - $S_{yx}(\%)$ (equation 7) as follow:

$$R_{aj}^2 = R^2 - \left[\frac{K-1}{N-K}\right].(1-R^2), \tag{3}$$

where: R^2= coefficient of determination; K= number of parameters; N= number of observations.

$$RMSE(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}\Big/\bar{Y}, \tag{4}$$

where: Y_i= observed height; $\hat{Y}_i$= estimated height; $\bar{Y}$= arithmetic mean of observed heights; n = number of observations.

$$E_i = \left(\frac{\hat{Y}_i - Y_i}{Y_i}\right) * 100, \tag{5}$$

where: E_i= Residue of i-th observation; Y_i= Observed height (m); $\hat{Y}$= Height estimated by the equation or ANN.

$$Bias = n^{-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right), \tag{6}$$

where: Y_i= Observed height (m); $\hat{Y}$= Height estimated by the equation or ANN; n= number of observations.

$$S_{yx}(\%) = \left(\frac{\sqrt{QMR}}{\bar{Y}}\right) * 100, \tag{7}$$

where: QMR = Mean Square Residual; $\bar{Y}$ = mean of observed values.

In the models in which the variables were subjected to logarithmic transformation, it was necessary to correct the logarithmic discrepancy of the estimated values (equation 9) and recalculate the standard error of the estimate (equation 8) with Meyer (1938) correction factor (equation 10).

$$S_{yx}corrected = \sqrt{\frac{\sum_{i=1}^{n}(H_i - \widehat{H}_i^*)^2}{n-p}}, \quad (8)$$

being:

$$\widehat{H}_i^* = \widehat{H}_i * F_M, \quad (9)$$

$$F_M = e^{0,5*QMR}, \quad (10)$$

where: H_i= observed height (m); $\widehat{H}_i$= height estimated by the equation (m); $\widehat{H}_i^*$= estimated height corrected (m); n = number of observations; p= number of parameters of the model; F_M= Meyer's correction factor; QMR= Mean Square Residual; $S_{yx}corrected$= corrected standard error of estimation.

The regression equation or the ANN selected was one that, in addition to having a high R_{aj}^2, a minor $RMSE$ and a non-tendentious and surroundings to zero residuals distribution, satisfied the assumptions of normality verified by Kolmogorov-smirnov test, homoscedasticity verified by White test and absence of residuals autocorrelation verified by Durbin-Watson test. The assumptions, although desired, was not required in ANNs since they can be considered non-parametric statistical techniques.

2.6 Regression equations performances versus ANNs performances in heights estimation in plantations of *Pinus caribaea* var. *caribaea*

The comparison between observed values and those estimated by the ANNs or selected hypsometric equations was performed using the statistical procedure L & O proposed by Leite and Oliveira (2002), testing the hypothesis H_0: observed values are equal to values estimated by selected regression equation and those estimated by trained ANNs. This procedure combines Graybill's F (H_0) test, t test for mean error ($t_{\bar{e}}$) and the linear correlation (r) between observed and estimated values.

2.7 Analysis of influence of sample size, neural network type and their interaction on the precision of the estimates.

The analysis of the influence of sample size and ANN type in the precision of the estimates was made by means of a completely randomized experimental design with 5x2 factorial arrangement in which the factors were: 1) Sample Size (N) with 5 levels: N1 (n =

2538), N2 (n = 1269), N3 (n = 635), N4 (n = 318) and N5 (n = 159) obtained by random resampling carried out in the database and 2) ANN type with 2 levels: R1 = Multilayer Perceptron (MLP) and R2 = Radial Basis Function (RBF). The data of the treatments were obtained by performing five trainings (repetitions) of each type of ANNs with each one of the samples.

For each sample size, the optimal number of neurons in the intermediate layer was calculated using Baum-Haussler (1989) formula (equation 11) given by

$$N_{int} \leq \frac{N * \varepsilon_{train}}{N_{input} + N_{output}}, \tag{11}$$

where: N_{int} = number of neurons in the intermediate layer; N = number of observations used for training; ε_{train} = error tolerance in training; N_{input}= number of neurons in the input layer; N_{output} = number of neurons in the output layer.

The hypotheses tested in the factorial arrangements were as follows:

H_{01}: The sample size (N) used for ANNs trainings does not influence its performance in trees height estimation;

H_{02}: The ANN type (R) does not influence the precision of height estimates made by trained Artificial Neural Networks;

H_{03}: The interaction N×R does not have a significant effect on the performance of ANNs in tree heights estimation.

Equation 12 represents the structural model of the factorial design.

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}, \tag{12}$$

where: Y_{ijk}= value of i-th height observed under the combination of j-th level of factor N with k-th level of factor R; μ = overall mean of all experiment data; α_j = effect of j-th level of factor N; β_k = effect of k-th level of factor R; $(\alpha\beta)_{jk}$ = effect of interaction of j-th level of factor N and k-th level of factor R; ε_{ijk} = experimental error.

Multiple comparisons of means were performed using Tukey test at 5% significance level.

3 Results and discussion

3.1 Analysis of sample sufficiency for hypsometric relationship modeling

The pilot sample of 2538 trees with a sampling error of 2.33%, less than 5%, was considered definitive for tree height estimation. The minimum sample size to meet the maximum permissible error of 5% is 519 trees. The correlation coefficient of 0.80 between *dbh* and *Ht* indicates the existence of a strong positive relation between both. According to Aldana (2010), this result is indicative of well-formed and well-managed sites.

3.2 Hypsometric models for *Pinus caribaea* var. *caribaea* in Macuije forest company

The results of regression model's adjustments indicate that the candidate equations to be selected as suitable for heights estimation were those of Logistic, Gompertz, Chapman-

Richards and Stoffels because they presented the highest coefficients of determination and lower RMSE and S_{yx} (Table 3).

Table 3 – Estimation of parameters of the adjusted hypsometric models

Models	Coefficients			R^2_{aj}	S_{yx}	RMSE	Bias
	b_0	b_1	b_2	(%)	(%)	(%)	
Logistic	22.36*	13.01*	5.768*	84.7	6.045	5.034	1.310
Gompertz	25.23*	2.52*	0.087*	85.4	6.006	4.995	0.180
Chapman-Richards	24.25*	0.012*	4.229*	84.0	6.486	5.478	2.687
Henricksen (1950)	-8.31*	7.673*	-	65.49	5.291	6.357	1.003
Curtis (1967)	2.99*	-7.243*	-	69.68	6.126	6.512	-3.330
Stoffels (1953)	0.37*	0.771*	-	76.21	6.285	5.128	-0.220

*Parameter estimate significant to 99% of probability by t test.

Curtis and Henricksen equations presented low coefficients of determination and a marked tendency to underestimate small and large heights (Figure 3). Gompertz and Stoffels equations, due to their bias values below 1% (Table 3) and random residuals distributions (Figure 3), do not present any trend in the estimates. Because of its greater precision and satisfaction of all statistical assumptions (Table 4), Gompertz equation is considered adequate to generate accurate and non-biased estimates.

Table 4 – Verification of statistical assumptions in equations

Models	Normality		Serial Autocorrelation		Homoscedasticity	
	K-S[a]		Durbin-Watson		White	
	KS	Sig.	DW	Auto.	LM	Sig.
Logistic	0.083	0.000	1.347	+	22.554	1.42E-07*
Gompertz	0.077	0.162	1.834	A	1.598	0.463^{NS}
Chapman-Richards	0.079	0.100	1.089	+	8.648	0.57E-02*
Henricksen (1950)	0.088	0.000	1.428	+	13.451	0.24E-03*
Curtis (1967)	0.109	0.000	1.305	+	24.163	8.85E-07*
Stoffels (1953)	0.051	0.073	1.909	A	0.627	0.731^{NS}

* Significant test at 5% significance; a. Correction of significance of Lilliefors; A = No serial autocorrelation in the residues; + = Existence of positive autocorrelation.

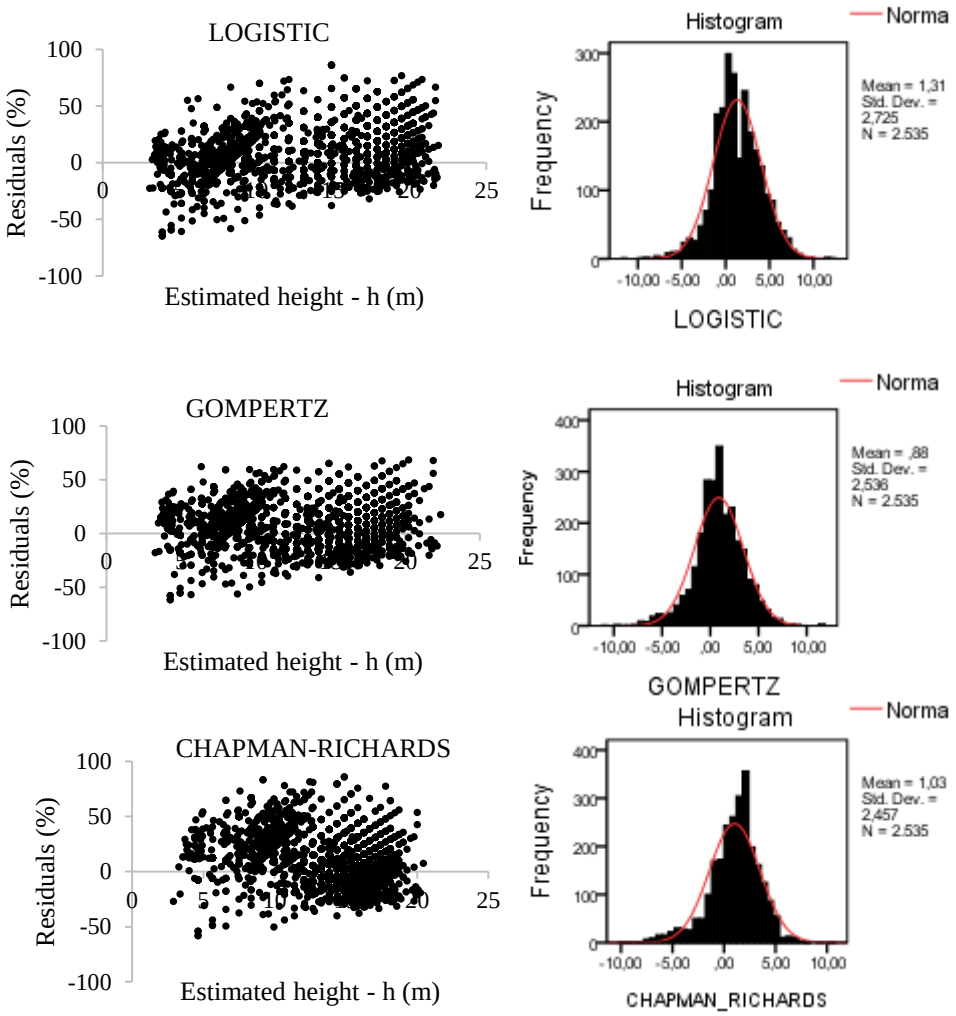

Figure 3 - Residuals distribution for hypsometric equations.

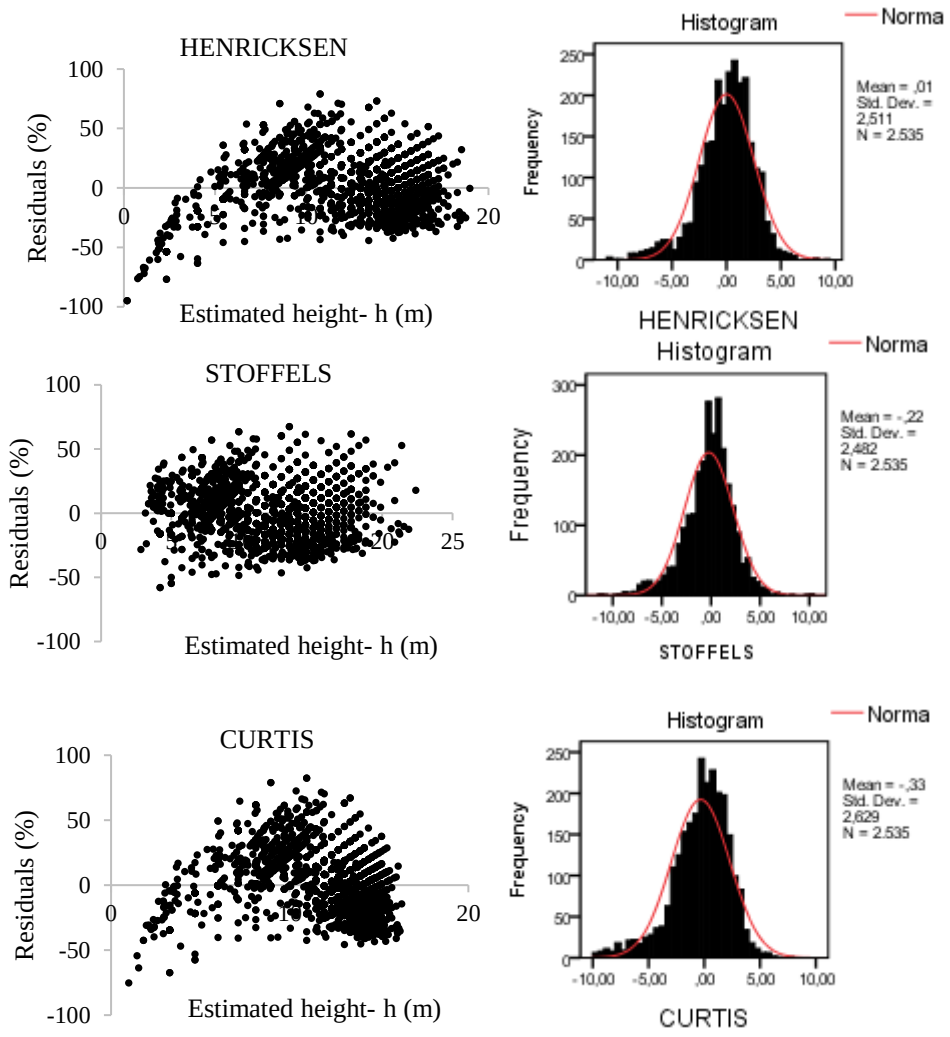

Figure 3 (Continuation) - Residuals distribution for hypsometric equations.

In Figure 4, it is possible to observe the behavior of each adjusted model and the good fit of Gompertz model to the database. It is also perceptible the tendency of underestimation of small and large heights identified in Curtis and Henricksen models.

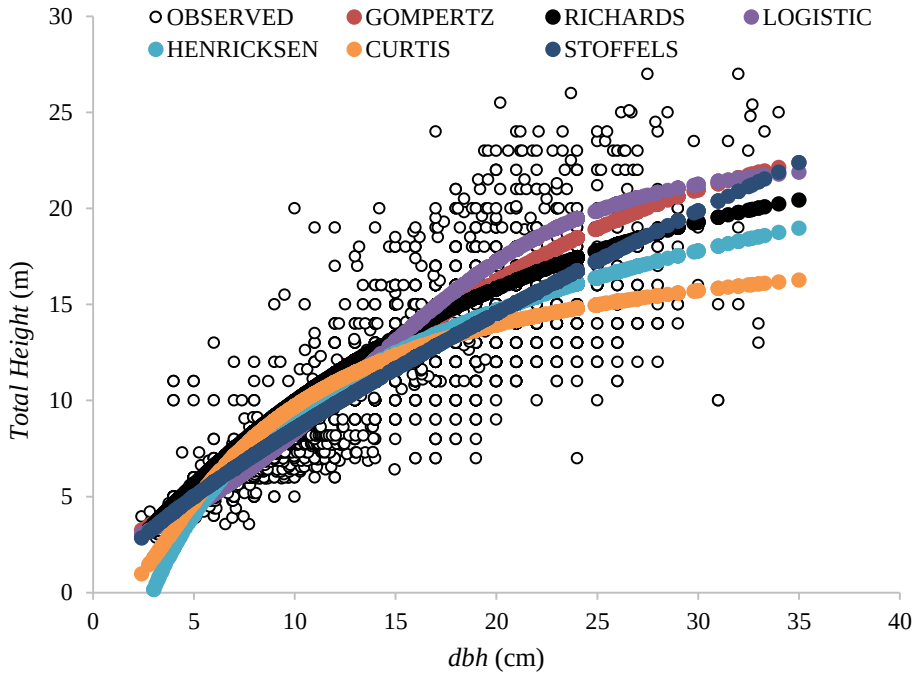

Figure 4 - Equations behaviors in height estimation in plantations of *Pinus caribaea* var. *caribaea*, Macurije Forest Company, Cuba.

3.3 Artificial Neural Networks for *Pinus caribaea* var. *caribaea* heights estimation

The trained Artificial Neural Networks presented satisfactory results with all the coefficients of determination above 85% and all RMSE less than 2% (Table 5). Based on these criteria, ANNs MLP 1-4-1 and MLP 1-6-1 are the best candidates.

Table 5 – Results of ANNs training for the estimation of the height

№	ANNs	Architecture	Var.	Activations Functions		R^2	RMSE (%)	Bias
				Intermediate	Output			
1	ANN1	MLP 1-2-1	DAP	Tanh	Tanh	85.71	1.865	0.025
2	ANN2	MLP 1-3-1	DAP	Identity	Tanh	85.70	1.867	-0.109
3	ANN3	MLP 1-6-1	DAP	Logistic	Tanh	85.71	1.863	-0.008
4	ANN4	MLP 1-4-1	DAP	Tanh	Identity	87.77	1.863	0.080

Residuals distribution indicated that all the heights estimated by the ANNs were accurate and without tendencies of underestimation or overestimation (Figure 5). All these results point to the selection of ANN MLP 1-4-1, whose parameter estimates are summarized in Table 6, as the most suitable for accurate height estimation in *Pinus caribaea* var. *caribaea* plantations at Macurije Forest Company.

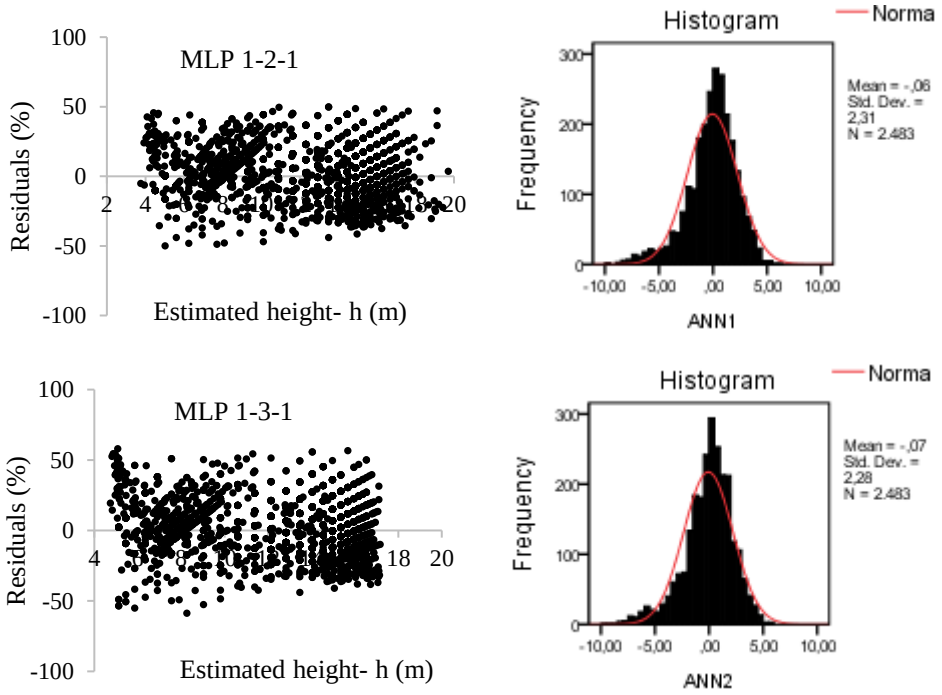

Figure 5 – Residuals distribution for ANNs in heights estimation.

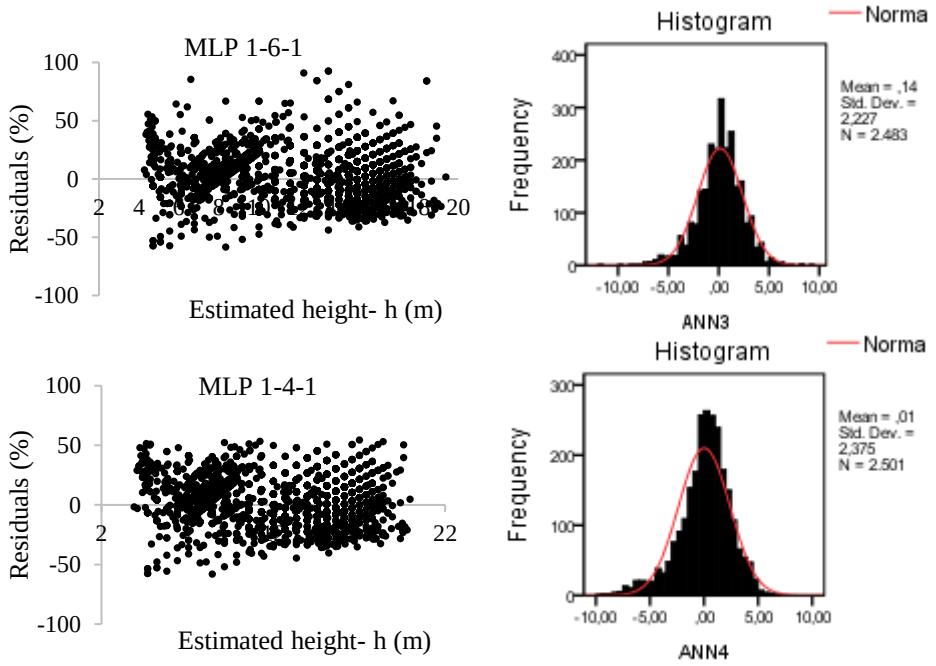

Figure 5 (continuation) – Residuals distribution for ANNs in heights estimation.

Table 6 – Estimates of ANN MLP 1-4-1 parameters

Predictors		Estimated				
		Hidden Layer				Output Layer
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	Height (m)
Input Layer	(Bias)	0.302	-1.236	-0.264	-0.243	
	DAP (cm)	0.460	-0.740	-0.432	0.400	
Hidden Layer	(Bias)					-1.284
	H(1:1)					0.032
	H(1:2)					-1.436
	H(1:3)					-1.148
	H(1:4)					0.023

Variables addition in statistical models tends to increase the dependent variable explained variance (increase of the coefficient of determination). It is not different in the

case of ANNs. The flexibility of ANN structures allowed categorical variables (Diametric class) inclusion in the training, whose results indicate a significant improvement in the accuracy of the estimates (Table 7 and Figure 6).

Table 7 – Results of ANNs training with inclusion of categorical variables

№	ANNs	Architectures	Var.	Activation functions		R^2 (%)	RMSE (%)	Bias
				Hidden	Output			
1	ANN1	MLP 8-4-1	DBH	Identity	Identity	90.747	1.321	-0.056
2	ANN2	MLP 8-7-1	DBH classes	Tanh	Identity	88.565	1.442	-1.644

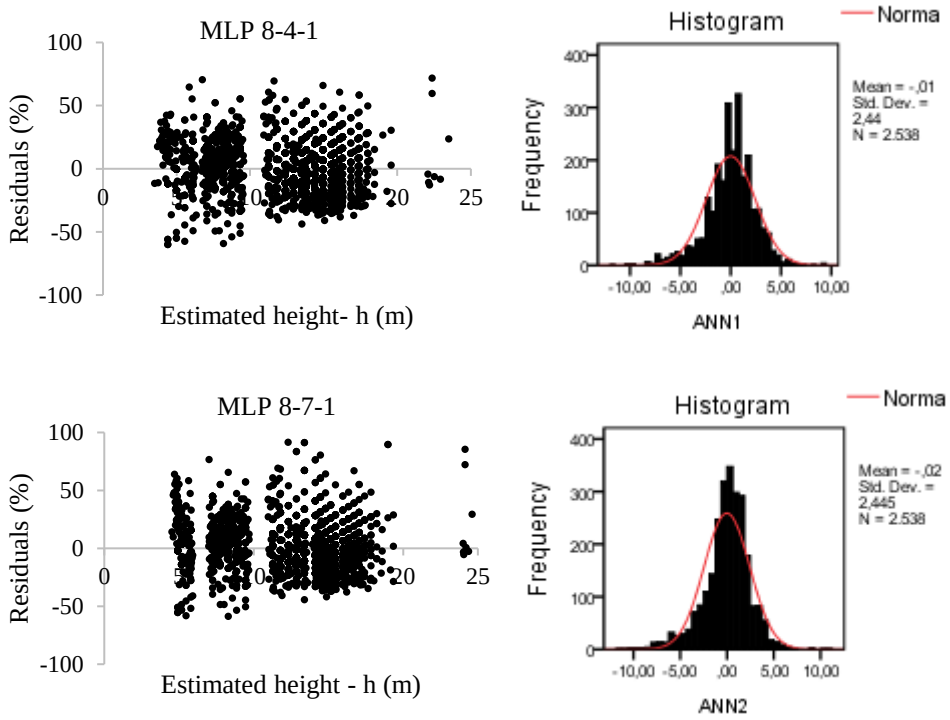

Figure 6 - Distribution of residues of ANNs with categorical variables.

Diametric class inclusion allowed to improve the predictive capacity of ANNs. The graph of Figure 7 indicates the variation of estimated heights by diameter class. A high

relationship between diametric class and heights was observed, with a lower variation in the last class (32.41 – 37.41) due to the trend of height stabilization in the largest diametric class mainly observed in mature stands in which the increases of these variables are lower.

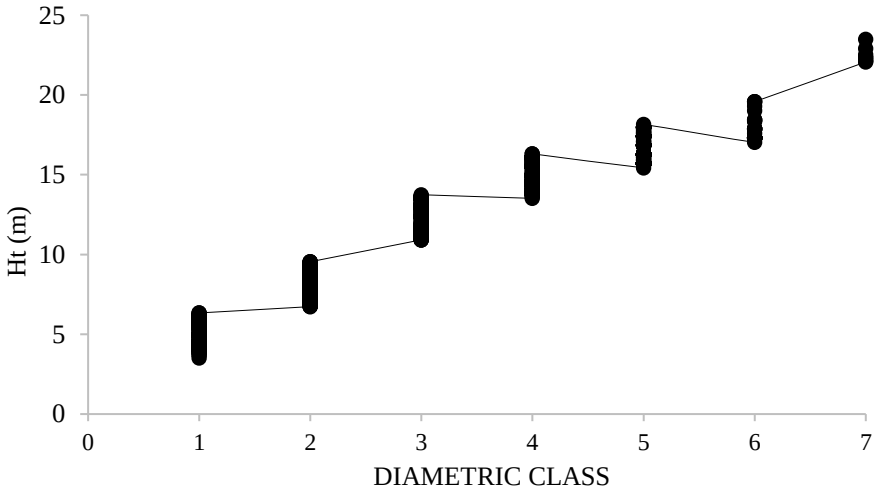

Figure 7 – Estimation of tree heights by diametric class with ANN MLP 8-4-1.

3.4 Regression equations vs ANNs in heights estimation

The results obtained by applying Leite and Oliveira (2002) statistical procedure for the best equation and selected ANNs are in Table 8. The non-significance of Graybill's test F (Ho) and t test for mean error ($t_{\bar{e}}$) and the satisfaction of the condition $r_{y_j y_i} \geq 1 - |\bar{e}|$ in all the models indicate that the observed values are statistically identical to those estimated by each of the selected models.

Table 8 - Results obtained by applying the statistical procedure proposed by Leite and Oliveira (2002) for the best equation and trained ANNs

	GOMPERTZ	**MLP 1-4-1**	**MLP 8-4-1**
$\boldsymbol{F\ (H_o)}$	1.1354^{ns}	1.0533^{ns}	0.7947^{ns}
$\boldsymbol{t(e)}$	0.754^{ns}	1.3248^{ns}	1.0172^{ns}
$\boldsymbol{r}$	0.9193	0.9398	0.9521
$\boldsymbol{Mean\ Error\ (e)}$	0.1802	0.0800	-0.056
$\boldsymbol{r_{y_j y_i} \geq 1 - \lvert\bar{e}\rvert}$	Yes	Yes	Yes

ns = not significant at 95% probability.

Figure 8 presents a graphical view of the behavior of each selected model in heights estimation in plantations of *Pinus caribaea* var. *caribaea*. The proportionality observed, between growths in diameters and heights, is evidence of the biological realism of the selected models.

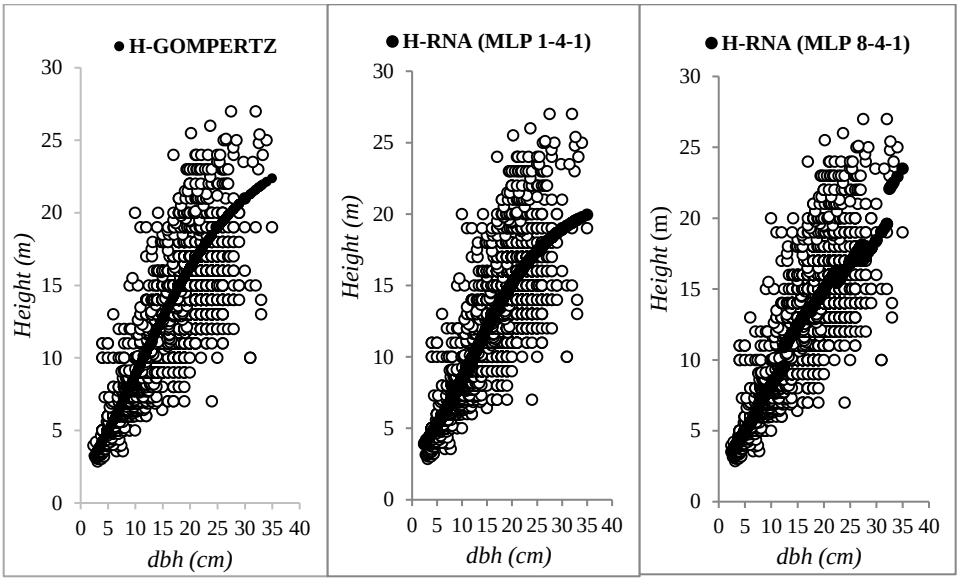

Figure 8 – Comparison between ANNs and equations in heights estimation.

The Neural Network MLP 1-4-1 and the Gompertz equation presented greater biological consistency since it was observed, in both models, a high growth rate in the beginning (small diameters) and this rate decreases and evolves towards a stabilization as the diameter stabilizes as indicated by Aldana (2010). In relation to ANN MLP 8-4-1, diametric class inclusion and ANN´s structure parallelism allowed to determine the best approximations by diametric class as indicated in Figure 8.

3.5 Influence of sample size and ANN type on the accuracy of estimates

ANNs trainings results are summarized in Table 9. They indicate that ANNs MLP-type showed better accuracy in all samples. All ANNs, except ANN RBF 1-4-1 in sample N3, were without bias. The bias value of this neural network, negative and greater than 5%, indicates that it has a general tendency to underestimate the heights.

The sample size is the most considered factor in studies of influence of data quality on the predictive capacity of ANNs. The results obtained in the present study (Table 9) indicate that the performances of ANNs are more influenced by the variance of the dependent variable measured by the coefficient of variation. Consideration of sample size

is due to the fact that its increase causes a reduction in variance but this does not always occur in practice. According to Table 9, the samples N3 and N4 presented lower coefficients of variation than the samples N1 and N2 that are superiors. It is then considered that a more objective analysis will be achieved based on the coefficient of variation (CV).

Table 9 - Results of MLP and RBF ANNs training in different sample sizes

SAMPLE SIZE	CV (%)	ANN TYPE	ARCHITECTURE	R^2 (%)	RMSE (%)	BIAS
N1	34,737	MLP	1-12-1	88.33	1.225	0.080
		RBF	1-12-1	85.98	1.293	0.117
N2	38,641	MLP	1-6-1	88.10	1.334	-1.233
		RBF	1-6-1	80.84	1.653	0.0825
N3	30,645	MLP	1-4-1	88.71	0.851	-0.048
		RBF	1-4-1	86.77	0.83	-5.863
N4	30,078	MLP	1-2-1	90.76	0.788	-0.534
		RBF	1-2-1	88.11	0.821	-0.590
N5	39,756	MLP	1-1-1	85.77	1.865	0.096
		RBF	1-1-1	76.28	2.713	0.078

As observed in Figure 9, the lower coefficients of variation coexist with the higher R^2 and the lower RMSE. The increase in CV had a negative influence on the accuracy of trained ANNs. ANNs, despite their high learning and generalization capacity, are affected by the variation of the data as any statistical model. Large samples are preferred because they generally have lower variances.

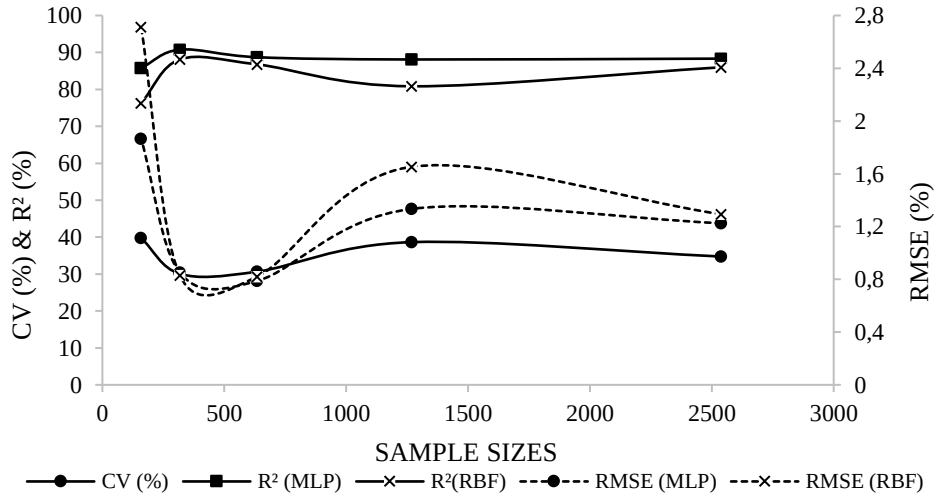

Figure 9 – Relationship between ANNs evaluation criteria and sample's coefficients of variation.

Either the Backpropagation algorithm used for ANN training or the least squares or iterative methods used to adjust the regression models are based on the convergence of the error expected value to zero (MARKHAM; RAKES, 1998). That explains the dependence of ANNs on the variance of the data. This variance depends on the nature of the variables but also on the errors of the forest inventories, one of the main data sources for forest modeling. These errors can be classified in sampling errors and non-sampling errors.

Non-sampling errors, which are measures of accuracy and not precision, are not objectives of the present study since their minimization does not depend on sample size and can occur even in censuses or total enumerations (SANQUETTA *et al*., 2009). These errors, difficult to detect and measure, are generally neglected in forest inventories total error calculation. According to Lesser and Kalsbeek (1997), these errors should be investigated and considered in forest inventories whose accuracy has a significant influence on quality of data used in forest modeling.

The sampling errors, attributed to the part of the population that was not sampled, decrease as the sampling intensity increases and is zero when the entire population is sampled (SANQUETTA *et al*., 2009). These sampling errors are functions of the variance that generally depends on the sample size (BORDERS, 2011). Sample size increase tends to decrease the variance in the dependent variable, which tends to cause an increase in the portion of the variance explained by the models (R^2) and a reduction in the standard error of the estimate (S_{yx}) by increasing residuals degree of freedom.

Despite the evidence of preference for larger sample sizes for ANNs as for any model based on statistical principles, the issue of resource availability (money and time) is very common in practice and mainly in hypsometric relationships modeling in which a smaller portion of the population is measured to make heights estimates for the remaining unmeasured portion. Time and costs minimization being the reason for the existence of these models, incur in high costs for their elaboration would be unintelligent. Is it really worth making investments in high quality data to get more accurate? Does the accuracy achieved in heights estimation justify the financial investment made?

All samples that were superior to the minimum sample of 519 trees (indicated by sample sufficiency analysis) presented satisfactory results. The sample N4, although inferior to the minimum sample, presented one of the smallest coefficients of variation, which allowed the obtaining of ANNs with good generalization capacities. With this sample, a neural network was obtained with good accuracy but without guaranteeing estimates accuracy since a good part of the population variance was not considered. Artificial Neural Networks trained with small databases are often at risk of overfitting and when this problem occurs, ANNs generally exhibit optimal training parameters and lower generalization capacities.

The results of analysis of variance (Table 10) between estimates obtained with ANNs of Table 9 indicated that factors sample size (N), ANN type (R) and their interaction (NxR) have a significant influence on the generalization capacity of Artificial Neural Network.

The results of N×R interaction analysis by means of Tukey test (Table 11) showed that in the three smaller samples (N3, N4 and N5), there was no significant difference between the heights estimated by the two types of neural network. However, in samples N1 and N2, it is observed that RBF type of ANNs generated higher estimates than those generated by MLP ANNs. There was then no difference between the uses of the two types of ANNs in the smaller samples.

Table 10 – Variance analysis for factorial arrangement 5×2

SOURCES OF VARIATION	DF	SS	MS	F
Treatments	9	144.036	16.004	61.444 **
Sample Size (N)	4	126.914	31.728	121.816 **
Ann Type (R)	1	12.955	12.955	49.740 **
N*R	4	4.167	1.042	3.999 **
Errors	1190	309.951	0.260	
Total	1199	453.987		

** = p < 0,01

Table 11 – Tukey test for the interaction N×R

Sample Size (N)	ANN Type (R)	
	R1 (MLP)	*R2 (RBF)*
N1 (n=2538)	13.223 aB	13.517 aA
N2 (n=1269)	13.195 aB	13.437 aA
N3 (n=635)	12.949 aA	13.016 bA
N4 (n=318)	12.615 bA	12.760 cA
N5 (n=159)	12.454 cA	12.545 dA

Averages followed by the same letters (lowercase in the columns and uppercase in the lines) do not differ from each other for the Tukey test at 5% significance.

Regarding the analysis of sample size variation influence on each type of ANN, a significant difference was observed between the performances of ANN MLP in samples N4 and N5 and their stabilization, from sample N3 to sample N1 (Table 11). The RBF ANNs presented different performances in almost all the samples. These results indicate that RBF ANNs are more influenced by sample size than MLP ANNs. This result is in correspondence with those of Ferreira et al. (2014) according to which MLP architecture is the most suitable and accurate for heights estimation of *Tibouchina granulosa* seedlings. According to the same authors, MLP neural networks are able to capture the biological realism of the data, to learn and to generalize the assimilated knowledge.

The influence of sample size, coefficient of variation and ANN type (1 = MLP; 2 = RBF) on ANNs accuracy were analyzed. The regressions of R^2 and RMSE on the variables sample size, coefficient of variation and ANNs type (1 = MLP; 2 = RBF) in the software DataFit 9 allowed to obtain the equations of expressions 14 and 15. Sample size exclusion is due to the non-significance of its contribution in the estimation of R^2 or RMSE.

$$R^2 = -2695.559 + 283720.933/CV - 9642219.006/CV^2 + 108932126.975/CV^3 + 9.476/ANN \quad (R^2_{aj} = 74.50\%;\ S_{yx} = 2.523\%) \tag{13}$$

$$REMQ = -438{,}107 + 38{,}395 * CV - 1{,}115 * CV^2 + 0{,}01076 * CV^3 - 0{,}499/ANN \quad (R^2_{aj} = 82{,}16\%; S_{yx} = 19{,}26\%) \tag{14}$$

It is observed as discussed above, a tendency to increase the coefficient of determination with coefficient of variation decrease (Figure 10). MLP networks presented higher coefficients of determination than those of RBF networks.

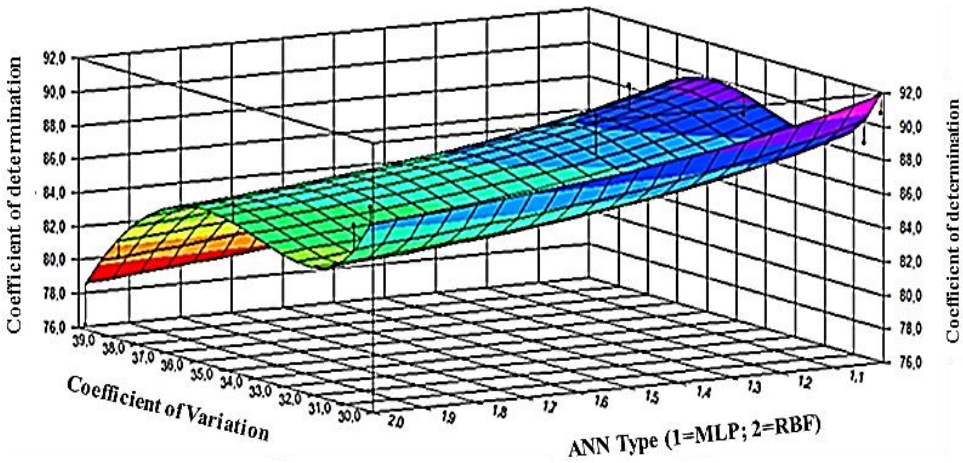

Figure 10 – Variation of ANNs coefficient of determination according to the coefficient of variation and ANN type.

Obviously, an inverse trend is observed for RMSE (Figure 11). Increase of height coefficient of variation caused an increase of RMSE as expected. It is also noticeable that MLP networks presented RMSE smaller than those of RBF networks.

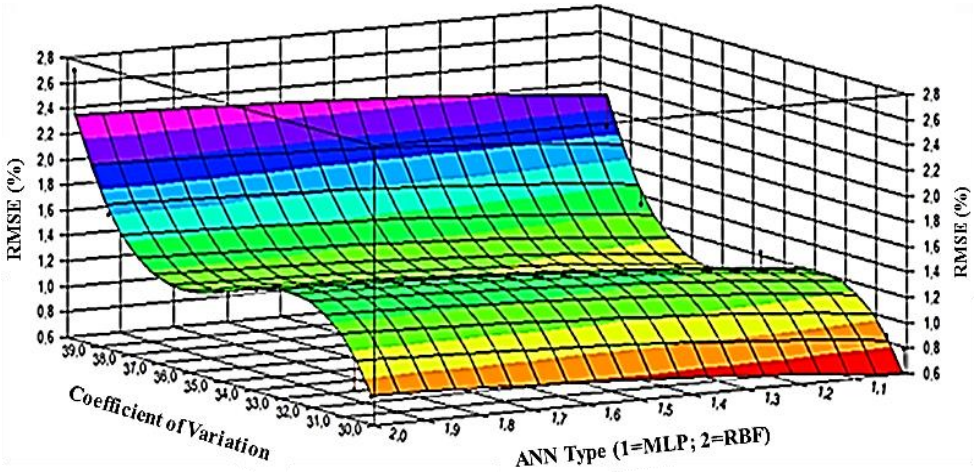

Figure 11 – Variation of ANNs RMSE according to the CV and ANN type.

Conclusions

The hypsometric equation obtained from Gompertz model was the most suitable for trees height estimation in *Pinus caribaea* var. *caribaea* plantations at Macurije Forest Company.

Both Gompertz equation and trained ANNs (MLP 1-4-1) generated accurate estimates that did not differ statistically from the observed values.

Diametric class inclusion as input variables improved the generalization capacity of Artificial Neural Networks (MLP 8-4-1) in local hypsometric relationships modeling.

Multilayer Perceptron Artificial Neural Networks allowed to obtain more accurate estimates than those obtained with Radial Basis Function networks and their performances were less influenced by sample size variation.

Acknowledgements

The authors are grateful to Macurije Forestry Company for providing their areas for carrying out the research and for the logistical support, to the forest department of University of Pinar del Río (UPR) and the biometric and forest management laboratory of the Federal Rural University of Pernambuco (LBMF / DCFL / UFRPE). Also, authors thank Reviewers and Editors for their comments.

GUERA, O. G. M., SILVA, J. A. A., FERREIRA, R. L. C., MEDEL, H. B., LAZO, D. A. Redes Neurais Artificiais para modelagem de relações hipsométricas de *Pinus caribaea* Morelet var. *caribaea* Barr. & Golf. *Rev. Bras. Biom.*, Lavras, v.36, n.4, p.891-915, 2018.

- *RESUMO: O presente estudo foi realizado com o objetivo de comparar os desempenhos de modelos de regressão e de Redes Neurais Artificiais (RNAs) na modelagem de relações*

hipsométricas e analisar a influência do tipo de RNA e o tamanho da amostra no desempenho das RNAs. A base de dados foi composta por 65 parcelas circulares de 500 m² nas quais foram medidas as variáveis DAP (cm) e Ht (m) de 2538 árvores em plantações de Pinus caribaea var. caribaea da empresa Macurije, Cuba. O estudo foi realizado em três etapas: i) Ajuste de modelos hipsométricos tradicionais e modelos sigmoidais de crescimento; ii) treinamento de RNAs e comparação da RNA selecionada com o modelo de regressão selecionado; iii) Análise da influência do tamanho da amostra e do tipo de RNA sobre a precisão das estimativas por meio de um delineamento experimental completamente aleatório com esquema fatorial 5x2, com os fatores tamanho da amostra (N) e tipo de RNA (R). Os resultados indicaram que a melhor equação para estimar a altura das árvores foi a de Gompertz. As RNAs MLP 1-4-1 e MLP 8-4-1 foram superiores à equação de Gompertz. As RNAs do tipo MLP geraram estimativas mais precisas e desempenhos menos influenciados pelo tamanho da amostra.

- *PALAVRAS-CHAVE: Inteligência artificial; análise de regressão; diâmetro à altura do peito; altura total.*

References

ALDANA, P. E. *Medición forestal.* La Habana: Editorial Félix Varela, 2010. 266p.

ALDER, D. *Growth Modelling for Mixed Tropical Forest*. Oxford: Tropical Forestry Papers. No. 30. OFI (FRP), 1995. 231 p.

ALMEIDA, T. G. *Comparação entre redes neurais artificiais e modelos tradicionais de regressão para estimar altura e volume em povoamentos de Araucaria angustifolia* (Bertol.) O. Kuntze, 2015. 56p. TCC - Universidade Federal do Paraná, Curitiba. 2015.

AZEVEDO, G. B.; SOUSA, G. T. O.; SILVA, H. F.; BARRETO, P. A. B.; NOVAES, A. B. Seleção de modelos hipsométricos para quatro espécies florestais nativas em plantio misto no planalto da conquista na Bahia. *Enciclopedia Biosfera*, v.7, n.12; p.1-13, 2011.

BAUM, E. B.; HAUSSLER, D. What net size gives valid generalization? *Neural Computation*, n.1, p. 151-160, 1989.

BINOTI, M. L. M. S.; BINOTI, D. H. B.; LEITE, H. G. Aplicação de redes neurais artificiais para estimação da altura de povoamentos equiâneos de eucalipto. *Revista Árvore*, v.37, n.4, p. 639-645, 2013.

BORDERS, B. E. *Timber inventory information use in timber management planning: what you don´t know, may cost you*. Athens: Statistical Issues in Forest Management. 2011. 65p.

CAMPOS, B. P. F.; SILVA, G .F.; BINOTI, D. H. B.; MENDONÇA, A. R.; LEITE, H. G. Predição da altura total de árvores em plantios de diferentes espécies por meio de redes neurais artificiais. *Pesquisa Florestal Brasileira,* v. 6, n.88, p.375-385, 2016.

CAMPOS, J. C. C.; LEITE, H. G. *Mensuração florestal: perguntas e respostas*. 4.ed. Viçosa: UFV, 2013. 605p.

CURTIS, R. O. Height-diameter and height-diameter-age equations for second-growth Douglas-fir. *Forest Science*, v.13, n.4, p.365-375, 1967.

GOMPERTZ, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transaction of the Royal Society*, v.115, p. 513-585, 1832.

HENRIKSEN, H. A. Height-diameter curve with logarithmic diameter: brief report on a more reliable method of height determination from height curves, introduced by the State Forest Research Branch. *Dansk Skovforeningens Tidsskrift*, v.35, n.4, p.193-202, 1950.

KER, J.; SMITH, J. Sampling for height-diameter relationships. *Journal of forestry*, v.55, n.3, p.205-207, 1957.

LEI, X. D.; PENG, C. H.; WANG, H. Y.; ZHOU, X. L. Individual height-diameter models for young black spruce (*Picea mariana*) and jack pine (*Pinus banksiana*) plantations in New Brunswick, Canada. *The Forestry Chronicle*, v.85: p.43–56, 2009.

LESSER, V. M.; KALSBEEK, W. D. Nonsampling error considerations in environmental surveys. In: AMERICAN STATISTICAL ASSOCIATION, 1997, *Proceedings of the Section on Survey Research Methods*, 1997. p.96-104.

MACHADO, S. A.; BASSO, S. F.; BEVILACQUA JUNIOR, V. G. Teste de modelos matemáticos para o ajuste da relação hipsométrica em diferentes sítios e idades para plantações de *Pinus elliottii* no Estado do Paraná. In: CONGRESSO FLORESTAL BRASILEIRO, 7., 1993, Curitiba. *Anais*. São Paulo: Sociedade Brasileira de Silvicultura, 1993. v.2, p.553-556.

MARKHAM, I. S.; RAKES, T. R. The effect of the sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computers Ops Research*, v.25, n.4, p.251-263, 1998.

MARTINS, E. R.; BINOTI, M. L. M. S.; LEITE, H. G.; BINOTI, D. H. B.; DUTRA, C. Configuração de redes neurais artificiais para estimação do afilamento do fuste de árvores de eucalipto. *Revista Brasileira de Ciência Agronômica*, v.11, n.1, p.33-38, 2016.

MOTTA, A. S.; ALMEIDA, E. J.; VENDRUSCOLO, D. G. S.; SOUZA, H. S.; MEDEIROS, R. A.; SILVA, R. S. Modelagem da altura de *Tectona grandis* L.f. clonal e seminal. *Revista Brasileira de Biometria*, v.34, n.3, p.395-406, 2016.

OZÇELIK, R.; DIAMANTOPOULOU, M. J.; CRECENTE-CAMPO, F.; ELER, U. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. *Forest Ecology and Management*, v.306, p.52-60, 2013.

PIENAAR, L. V.; TURNBULL, K. J. The Chapman-Richards generalization of Von Bertalanffy's growth model for basal area growth and yield in even-aged stands, *Forest Science*, v.19, n.1, p.2-22, 1973.

RICHARDS, F. J. A. Flexible growth function for empirical use. *Journal of Experimental Botany*, v.10, n.2, p.290-300, 1959.

SANQUETTA, C. R.; CORTE, A. P. D.; MOGNON, F.; RODRIGUES, A. L. Estimativa de carbono individual para Araucária angustifólia. *Pesquisa Agropecuária Tropical*, v.44, p.1-8, 2014a.

SANQUETTA, C. R.; WATZLAWICK, L. F.; CORTE, A. P. D.; RODRIGUES, L. A. *Inventários florestais: planejamento e execução*. 3.ed. Curitiba: Mult-Graf, 2014b. 406p.

SANQUETTA, C. R.; WATZLAWICK, L. F.; DALLA CORTE, A. P.; FERNANDES, L. A. V.; SIQUEIRA, J. D. P. *Inventários Florestais: Planejamento e Execução.* Curitiba: Mult-Graf, 2009. 316p.

VALENÇA, M. *Fundamentos das redes neurais: Exemplos em Java.* 2.ed. Olinda: Livro Rápido, 2010. 386p.

VENDRUSCOLO, D. G. S.; DRESCHER, R.; SOUZA, H. S.; MOURA, J. P. V. M.; MAMORÉ, F. M. D.; SIQUEIRA, T. A. S. Estimativa da altura de Eucalipto por meio de regressão não linear e redes neurais artificiais. *Revista Brasileira de Biometria*, v.33, n.4, p.556-569, 2015.

VERHULST, P. F. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathématique et physique*, v.10, p.113-121, 1838.

Received on 05.06.2017

Approved after revised on 14.03.2018