*Brazilian Journal of Biometrics*, **43**, 1-16 (2025)
doi:10.28951/bjb.v43i3.743

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

# Regression analysis: a new *methodology* to compare equations[1]

José Antônio Aleixo da Silva*[1], Rinaldo Luiz Caraciolo Ferreira[1], Rute Berger[1], Daniel Alberto Avarez-Lazo[2] and Ouorou Ganni Mariel Guera [3]

[1] Department of Forest Science, Federal Rural University of Pernambuco, Recife-PE, Brazil
[2] Department of Forest Science, Pinar del Rio University, Pinar del Rio, Cuba
[3] Department of Forest Science, Polytechnic University of Morelos State, Morelos, Mexico
*Corresponding author. Email: jaaleixo@uol.com.br

(Received: June 7, 2024; Revised: November 25, 2024; Accepted: April 2, 2025; Published: July 30, 2025)

**Abstract**

When working on modeling with regression analysis, several models are generally tested in order to select a more accurate equation after fitting in relation to some statistical tests. The decision becomes more difficult when the results of the statistical tests are similar because they are usually tests by equation, meaning that they provide statistics for each equation separately. A methodology which classifies them according to the averages of estimates in relation to the average of actual data does not exist because the averages of estimates and actual data are the same. This work proposes a methodology where the average sum of squares of the fitted equations and the average sum of squares of the real data are used. Thus, it becomes possible to apply a mean comparison test to group similar equations and separate different equations. For this work, 105 trees of *Eucalyptus* spp. clones were used, rigorously cubed by the Smalian method at seven and a half years of age in their second rotation, in a forest experiment located in a semi-arid zone of Pernambuco, Brazil. In turn, five linear volumetric models and one non-linear model were fitted and comparisons among equations were performed using several statistical tests usually employed in forest modeling. The Tukey, Duncan, Dunnett and Scott-Knott tests were used to test the proposed methodology, which, in addition to presenting similar results to the traditional tests, enabled grouping similar equations, and having the real volumes of the trees as the control.

**Keywords**: *Eucalyptus*; Modeling; Average sum of squares; Mean comparison tests.

## 1. Introduction

The statistical technique of regression analysis models and investigates relationships among variables is considered one of the most used among all statistical methodologies. It is used in its breadth in the various fields of science, such as: Physics, Chemistry, Biology, Epidemiology, Econometrics, Administration, Social Sciences, and Forest Measurements, among others (Montgomery *et al.*, 2003; Figueiredo-Filho *et al.*, 2011; Fahrmeir *et al.*, 2013; Segel & Edelstein, 2013; Silva & Ferreira, 2021).

Regression is understood as the study of the relationship between two variables or groups of variables, in which one seeks to estimate the value of a dependent variable (Yi) from the knowledge of the values of one or more independent variable(s) (Xi). After fitting the models in modeling processes, there is a need to select the best equation that provides robust estimates of the modeled phenomenon.

There is a large number of statistics which are used to select equations, such as: simple and adjusted coefficient of determination, simple and corrected adjustment index, residual standard error, mean absolute error, standard error of the estimate, Furnival index, criterion of simple and corrected Akaike adjustment, Bayesian information criterion, weighted value of statistical scores, and graphical analysis of residuals, among others (Furnival, 1961; Akaike, 1974; Schwarz, 1978; Schlaegel, 1981; Schneider *et al*., 2009; Mehtätalo & Lappi, 2020).

These procedures always seek to select the best equation and it is often a difficult decision because other factors in addition to precision must be considered, such as: measurement techniques, costs to measure the independent variable(s), availability and accuracy of measuring instruments, equation selection tests, etc. However, selecting groups of equations that obtain similar results is not a common practice.

The simplest way to select groups of equations would be to apply a completely randomized experimental design with real values as comparators and equations as treatments and perform traditional analysis of variance (ANOVA), which is a fundamental statistical tool for testing randomness of data from a sample in order to select predictors and test hypotheses in experimental and modeling processes. If a significant difference was verified by the Snedecor F test (Snedecor, 1934), a mean separation test would be applied.

This procedure used in experimentation and other statistical techniques is unable to be used in regression analysis because the sum of observed values is equal to the sum of estimated values $\left(\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i\right)$, resulting in $\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i = 0$, as follows (expressions 1, 2, 3 and 4):

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0 \quad (1)$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} (b_0 + b_1 X_i) = 0 \quad (2)$$

But, $\hat{Y}_i = b_0 + b_1 X_i$ . So that:

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i = 0 \quad (3)$$

$$\sum_{i=1}^{n} \bar{Y}_i = \sum_{i=1}^{n} \hat{Y}_i \quad (4)$$

Therefore, given that the estimated means of all fitted equations are equal to the dependent variable mean or true mean $\left(\bar{\hat{Y}} = \bar{Y}\right)$, applying any mean comparison tests would always do not reject the null hypothesis because differences between any estimated mean and the real mean will always be zero.

In this work, what is proposed to solve this problem and to be able to form groups of similar equations is to compare the mean sum of squares of the real data, in this case the total sum of squares (TSS) with the means of the sum of squares of the regression (SSR) and carry out the F test normally as applied in the analysis of experiments. In this case, the null hypothesis ($H_0$) is not rejected when the means of SSR and TSS do not differ significantly at the α level of significance. At the same time, the means of the SSR of the different equations to be tested are also compared.

In the specific case of volumetric models in Forest Science and with possession of tree volumes (actual and estimated), variance analysis is conducted considering a completely randomized design and application of the F test. If $H_0$ is not rejected, the equations provide similar estimates to the real data, otherwise it is assumed that at least two equations differ from each other.

Applying a separation of means test, such as: Tukey (Tukey, 1949), Duncan (Duncan, 1955), Dunnett (Dunnett, 1964) among others, would allow grouping those equations that do not present significant difference among themselves, and an equation may belong to more than one group of equations in the Tukey and Duncan tests.

Although the Tukey and Duncan tests originate classifications of means from highest to lowest, the most important thing in this specific case of equation comparison is to consider the position of the mean of the total sum of squares (TSS) as being the main element of the comparisons, because it represents the actual volume.

It is recommended to apply the Scott-Knott test when the number of equations is high in order to avoid that an equation belongs to more than one group; it uses the likelihood ratio to test the significance that *i* treatments can be separated by groups that maximize the sum of squares between groups (Scott & Knott, 1974).

The objective of this work is to compare equations and form groups of equations through the means of total sums of squares of real data (TSS) and fitted equations (SSR).

## 2. Materials and Methods

For this work, 105 *Eucalyptus* spp. trees in second rotation, rigorously cubed by the Smalian method, were selected in a completely random sampling considering an adopted error (SE%=10%). The trees were planted in March 2011, were seven and a half years old when they were cubed and are part of an experiment installed at the Experimental Station of the Agronomic Institute of Pernambuco (IPA), in the municipality of Araripina, Pernambuco located in Chapada do Araripe, in the municipality of Araripina, Brazil, with geographic coordinates 07°27’37” S and 40°24”36” W, altitude of 831 meters, and with yellow latosol + red-yellow latosol soils. The climate is Bshw according to the Koppen classification, being semi-arid, hot and having summer-autumn rains. The average precipitation during the second rotation was 450 mm/year, concentrated between the months of November to May, which promotes water deficits from May to October (Gouveia *et al.*, 2015).

Before fitting with rigorous cubing data, the sampling error (actual) was calculated in percentage (SE%) (expression 5) in order to compare with the stipulated or adopted error (AE).

$$SE\% = \left[\frac{t_{(\alpha/2;n-2)}\, s_{\bar{V}}}{\bar{V}}\right] 100 \qquad (5)$$

where: $t_{(\alpha/2;n-2\ df)}$ = tabulated t-test value at the significance level $(\alpha/2)$ with n–2 degrees of freedom (df); $s_{\bar{V}}$ = standard error of the mean volume; $\bar{V}$ = mean volume.

The decision rule is that if SE% ≤ AE% the sample is sufficient, meaning that it represents the forest population. In the event of SE% > AE%, there is a need to calculate the sample size (n) using expression 6.

$$n = \frac{t^2_{(\alpha/2;n-2)} CV\%^2}{AE\%^2} \qquad (6)$$

where: n = number of samples needed to represent the forest population; $t_{(\alpha/2;n-2\ df)}$ = tabulated t-test value at the significance level $(\alpha/2)$ with n–2 degrees of freedom (df); CV% = coefficient of variation.

The variability among volume (V), diameter at height of 1.3 m from the ground (D) and the height of the tree (H) that will define sufficiency will be that which presents the greatest variability, as it corresponds to the need for a greater number of trees to represent the forest population considered. The calculation of n will indicate the need for cubing new trees so that SE% ≤ AE%. Five simple linear models and one non-linear model (Chart 1) were fitted using the least squares method for fitting the linear models and the iterative Marquardt method for fitting the non-linear model (Marquardt & Snee, 1974).

**Chart 1.** Adjusted linear and non-linear volumetric models of Eucalyptus spp. at 7.5 years old in Araripina, PE, Brazil

| Name | Model |
|---|---|
| Spurr or combined variable (Spurr, 1952) | $V_i = \beta_0 + \beta_1 (D^2H)_i + \varepsilon_i$ |
| Spurr without $\beta_0$ | $V_i = \beta_1 (D^2H)_i + \varepsilon_i$ |
| Non-linear Spurr | $V_i = \beta_0 (D^2H)_i^{\beta_1} . \varepsilon_i$ |
| Simple linear as a function of DBH | $V_i = \beta_0 + \beta_1 D_i + \varepsilon_i$ |
| Simple linear as a function of H | $V_i = \beta_0 + \beta_1 H_i + \varepsilon_i$ |
| Simple linear as a function of DH | $V_i = \beta_0 + \beta_1 DH_i + \varepsilon_i$ |

Where: Vi = Tree volume in m3; D = diameter at 1.30m from the ground in cm; H = total height of the tree in meters; β0 and β1= model parameters; εi = random error, uncontrollable part of the model.

The following statistical tests and graphical analysis of residuals were initially applied:

a) Adjusted coefficient of determination ($r^2_{adj}$) (expression 7)

$$r^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)(1 - r^2) \tag{7}$$

where: p = number of parameters in the model; $r^2$ = Coefficient of determination; $r^2 = \frac{SSR}{TSS}$,

where: SSR = Sum of squares of the regression; TSS = Total sum of squares.

b) Corrected Adjustment Index % (CAI%) (Schlaegel, 1981) (expression 8)

$$CAI\% = \left[1 - (1 - IA)\left(\frac{n-1}{n-p}\right)\right] 100 \tag{8}$$

where: $IA = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$, n = number of trees sampled; p = number of parameters in the model; $\hat{Y}_i$ = volume of the i-th tree estimated by regression; $\bar{Y}$ = mean volume observed; $Y_i$ = volume of the i-th tree observed.

c) Mean standard error of the residual (MRSE) (expression 9)

$$MRSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{9}$$

d) Absolute mean error (AME) (expression 10)

$$AME = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \quad (10)$$

e) Standard error of the estimate in percentage (SEE%) (expression 11)

$$SEE\% = \left(\frac{100}{\bar{Y}}\right)\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}} \quad (11)$$

f) Mean absolute standard error (MASE) (expression 12)

$$MASE = \frac{100}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i} \quad (12)$$

g) Mean square of the residual (MSR) (expression 13)

$$MSR = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}} \quad (13)$$

h) Coefficient of variation in percentage (CV%) (expression 14)

$$CV\% = \left[\frac{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}}{(\sqrt{n-p})\bar{Y}}\right]100 \quad (14)$$

i) Furnival Index (FI) (Furnival, 1961) (expression 15)

$$FI = \sqrt{MSR}\, exp\left[-\frac{(n-2)}{2n}\right] \quad (15)$$

j) Akaike Information Criterion (AIC) (Akaike, 1974) (expression 16)

$$AIC = n\, ln\left[\frac{SQRes}{n}\right] + 2p \quad (16)$$

l) Corrected Akaike Information Criterion ($AIC_c$) (Bozdogan, 1987) (expression 17)

$$AIC_c = AIC + \frac{2p(p+1)}{n-p-1} \quad (17)$$

m) Bayesian Information Criterion (BIC) (Schwarz, 1978) (expression 18)

$$BIC = AIC - p[2 - ln\,(n)] \quad (18)$$

n) Weighted value of statistical scores (WV) (expression 19)

The WV assigns values to each equation per test, with the best equation in a test given weight one, the second given weight two, and so on. The WV is expressed (19) by:

$$WV = \sum_{i=1}^{n}(Nr_i)(P_i) \quad (19)$$

where: $Nr_i$ = occurrence number in the i-th classification; $P_i$ = weight of the i-th classification.

o) Graphical analysis of the residuals

After fitting the models, the resulting equations were used to estimate the volumes of individual trees, and the results were subsequently submitted to ANOVA (α=5%; 1%) in a completely random design, considering each equation as a treatment and the real data as a comparator. The Tukey, Duncan, Dunnett and Scott-Knott mean comparison tests were applied at 5% and 1% significance levels.

Tukey's test, which compares contrasts between two means and is based on studentized amplitude, is expressed (20 and 21)by (Tukey, 1949):

$$\Delta = MSD = q_{(\alpha;n,n'df)}\sqrt{\frac{MSR}{J}}, \text{ when J (number of repetitions) by equation is equal.} \quad (20)$$

$$\Delta = MSD = q_{(\alpha;n,n'df)}\sqrt{\frac{1}{2}\left(\frac{1}{J_1}+\frac{1}{J_2}\right)MSR}, \text{ when J is different by equation.} \quad (21)$$

where: Δ= MSD = minimal significant difference in contrast involving two means; $q_\alpha$ = tabulated value of the studentized amplitude calculated according to the number of treatments to be compared (n) and the number of degrees of freedom of the residual (n') for the significance level α; MSR = mean square of the residual; J = number of repetitions per treatment (equations).

Duncan's multiple range test (1955) is more flexible than Tukey's test in terms of rejecting $H_0$, and it allows comparing contrasts between two means; however, it takes into account the number of means involved in the contract, meaning the mean(s) which is/are between the mean values to be compared. The test is expressed (22 and 23) by:

$$D = MSD = z_{(\alpha;n,n'df)}\sqrt{\frac{MSR}{J}}, \text{ when J (number of repetitions) by equation is equal.} \quad (22)$$

$$D = MSD = z_{(\alpha;n,n'df)}\sqrt{\frac{1}{2}\left(\frac{1}{J_1}+\frac{1}{J_2}\right)MSR}, \text{ when J is different by equation.} \quad (23)$$

where: D = MSD = Minimum significant difference between two treatments by Duncan's test; $z_\alpha$ = tabulated value as a function of the number of means involved in the contrast and the number of degrees of freedom of the residual for the significance level α.

Dunnett's test uniquely compares each equation with the control. In this test, the null and alternative hypotheses are respectively $H_0$: $\mu_c = \mu_i$ and the alternative hypothesis $\mu_c \neq \mu_i$, in which $\mu_c$ corresponds to the mean of the comparator or control, and $\mu_i$ the mean of treatment i to be compared with the control (Dunnett, 1964). The test is expressed (24 and 25) by:

$$\Delta = MSD = d_{(\alpha;n,n'df)}\sqrt{\frac{(2)MSR}{J}}$$ , when J (number of repetitions) by equation is equal. (24)

$$\Delta = MSD = d_{(\alpha;n,n'df)}\sqrt{\left(\frac{1}{J_1}+\frac{1}{J_2}\right)MSR}$$, when J is different by equation. (25)

The $d_{(\alpha;n,n'df)}$ corresponds to the tabulated value at the α significance level with n degrees of freedom (df) for treatments and n' degrees of freedom for the residual.

The Scott-Knott test was also applied to avoid letter coincidences among treatments (Scott & Knott, 1974).

The calculation procedure for this test is as follows:

1) Sort means in ascending order;
2) Calculate the partition between two groups that maximize the sum of squares between groups. Let $T_1$ and $T_2$ be the totals of the two groups with $k_1$ and $k_2$ treatments; The sum of squares of $B_0$ ($SSB_0$) is expressed (26) by:

$$SSB_0 = \frac{T_1^2}{k_1} + \frac{T_2^2}{k_2} - \frac{(T_1+T_2)^2}{k_1+k_2} \quad (26)$$

3) Determine the value of λ (expression 27)

$$\lambda = \left[\frac{\pi}{2(\pi-2)}\right]\left(\frac{SSB_0}{\hat{\sigma}_0^2}\right) = 1.37597\left(\frac{SSB_0}{\hat{\sigma}_0^2}\right) \quad (27)$$

$\hat{\sigma}_0^2$ is the estimator of $\sigma_{\bar{Y}}^2$. It is expressed (28) by:

$$\hat{\sigma}_0^2 = \left(\frac{1}{g+v}\right)\left[\sum_{i=1}^{n}(\bar{Y}_i - \bar{Y})^2 + v.s_{\bar{Y}}^2\right] \quad (28)$$

where: g = number of treatments involved in the mean group; v = degrees of freedom of the residual; $\bar{Y}_i$ = mean of treatment i; $\bar{Y}$ = mean of the mean of the treatments; $s_{\bar{Y}}^2 = \frac{MSR}{j}$

4) If $\lambda > \chi^2_{\left[\alpha;\frac{g}{\pi-2}\right]}$ , $H_0$ is rejected

where: $\chi^2_{\left[\alpha;\frac{g}{\pi-2}\right]}$ = tabulated value of $\chi_i^2$

Thus, when $H_0$ is rejected, it means that the groups of compared means are different. In this test, each equation will belong to a single group, which does not happen in the Tukey, Duncan and other means separation tests, thus preventing one mean from belonging to two or more groups.

The application of the Scott-Knott test using the volume in $m^3$ would generate a statistical analysis with extremely small values. Therefore, it was decided to use the volume in $cm^3$ for better understanding of the procedure because it will not change the ANOVA results, since the nature of the dependent variable in regression analysis is not changed depending on the nature of the independent variable(s) used in the modeling. For example, using the volume in $cm^3$, the diameter in cm and the height in m, the volume estimates will also be in $cm^3$.

Considering the Spurr equation or the combined variable $\widehat{V}_i = b_0 + b_1(D^2H)_i$ with the volume expressed in $cm^3$, the coefficients $b_0$ and $b_1$ are respectively expressed (29 and 30) by:

$$b_0 = \bar{V} - b_1 \left(\overline{D^2H}\right)_i \tag{29}$$

$$b_1 = \frac{\sum_{i=1}^{n}(D^2H)_i V_i - \frac{\left[\sum_{i=1}^{n}(D^2H)_i\right]\left(\sum_{i=1}^{n}V_i\right)}{n}}{\sum_{i=1}^{n}(D^2H)_i^2 - \frac{\left[\sum_{i=1}^{n}(D^2H)_i\right]^2}{n}} \tag{30}$$

Making the substitutions of the units, we have (expressions 31 and 32):

$$b_1 = \frac{(cm^2.m)(cm^3) - \frac{(cm^2.m)(cm^3)}{n}}{(cm^2.m)^2 - \frac{(cm^2.m)^2}{n}} \tag{31}$$

$$b_1 = \frac{(cm^6.m)-(cm^6.m)}{(cm^4.m^2)-(cm^4.m^2)} = \frac{(cm^6.m)}{(cm^4.m^2)} = \frac{cm^2}{m} \tag{32}$$

Substituting b1 for b0, we have (expressions 33):

$$b_0 = cm^3 - \left(\frac{cm^2}{m}\right)(cm^2.m) = cm^3 - cm^4 \tag{33}$$

So the equation $\widehat{V}_i = b_0 + b_1(D^2H)_i$ will provide estimates in $cm^3$ (expressions 34).

$$\hat{V}_i = b_0 + b_1(D^2H)_i = (cm^3 - cm^4) - \left(\frac{cm^2}{m}\right)(cm^2.m) = cm^3 - cm^4 + cm^4 = cm^3 \tag{34}$$

The tests used were carried out with the SYSTAT12 software program (DEMO).

## 3. Results and Discussion

Table 1 presents some statistics of the dependent and independent variables used in fitting the models.

**Table 1.** Descriptive statistics and dispersion measures for the measured variables for *Eucalyptus* spp. at 7.5 years old in Araripina, PE, Brazil

| Statistics | Variables | | |
|---|---|---|---|
| | Volume ($m^3$) | D (cm) | H (m) |
| Minimum value | 0.0099 | 4.7 | 4.6 |
| Maximum value | 0.0496 | 11.1 | 13.7 |
| Arithmetic mean | 0.0241 | 7.4 | 10.2 |
| Standard deviation | 0.0092 | 1.2 | 1.8 |
| Confidence interval | 0.0241 ± 0.0018 | 7.4 ± 0.2 | 10.2 ± 0.4 |
| Sampling error (SE%) | 7.47 | 2.7 | 3.9 |

As the variable that presented the highest sampling error was the dependent variable volume (SE% = 7.47), sample sufficiency was achieved. After making the adjustments, the resulting equations are shown in Table 2.

**Table 2.** Equations resulting from model fitting for *Eucalyptus* spp. at 7.5 years old in Araripina, PE, Brazil

| Name | Equations |
|---|---|
| Spurr or combined variable (Spurr, 1952) | $\widehat{V}_i = 0.004656 + 0.000033(D^2H)_i$ |
| Spurr without β_0 | $\widehat{V}_i = 0.000040(D^2H)_i$ |
| Non-linear Spurr | $\widehat{V}_i = 0.000142(D^2H)_i^{0.807587}$ |
| Simple linear as a function of DBH | $\widehat{V}_i = -0.026467 + 0.006813\ D_i$ |
| Simple linear as a function of H | $\widehat{V}_i = -0.009845 + 0.003347\ H_i$ |
| Simple linear as a function of DH | $\widehat{V}_i = -0.004691 + 0.000377\ DH_i$ |

The weighted values of the statistical scores for the different fitted equations are shown in Table 03. The equations were classified with the most accurate equation receiving a weight of 1.0 and the others receiving weights in ascending order.

**Table 3.** Results of the different statistical criteria used in the fitted equations for *Eucalyptus* spp. at 7.5 years old in Araripina, PE, Brazil

| Statistical | Equations | | | | | |
|---|---|---|---|---|---|---|
| | 01 | 02 | 03 | 04 | 05 | 06 |
| $r^2_{adj}$ | 0.9078 (2) | 0.8708 (3) | 0.9121 (1) | 0.8185 (5) | 0.4371 (6) | 0.8513 (4) |
| CAI | 0.9077 (2) | 0.8702 (3) | 0.9113 (1) | 0.8180 (5) | 0.4316 (6) | 0.8511 (4) |
| MRSE | 0.0028 (2) | – | 0.0027 (1) | 0.0039 (4) | 0.0069 (5) | 0.0035 (3) |
| AME | 0.0023 (1) | 0.0027 (2) | 0.0023 (1) | 0.0030 (3) | 0.0053 (4) | 0.0027 (2) |
| SEE% | 11.60 (2) | – | 11.41 (1) | 16.25 (4) | 28.75 (5) | 14.94 (3) |
| MASE | 10.01 (1) | 11.90 (2) | 10.01 (1) | 14.08 (4) | 24.96 (5) | 12.80 (3) |
| MSR | 0.0028 (1) | – | 0.0028 (1) | 0.0039 (3) | 0.0069 (5) | 0.0036 (2) |
| CV% | 11.60 (2) | – | 11.41 (1) | 16.43 (4) | 28.70 (5) | 14.95 (3) |
| FI | 0.0017 (1) | – | 0.0017 (1) | 0.0024 (3) | 0.0042 (4) | 0.0022 (2) |
| AIC | –11.74 (2) | – | –11.78 (1) | –11.06 (4) | –9.92 (5) | –11.26 (3) |
| AICc | –11.62 (2) | – | –11.66 (1) | –10.94 (4) | –9.45 (5) | –11.14 (3) |
| BIC | – 6.43 (2) | – | – 6.45 (1) | –5.75 (4) | – 4.61 (5) | –5.95 (3) |
| TOTALS | (20) | (10) | (13) | (47) | (60) | (35) |
| MEANS | 1.54 | 2.50 | 1.00 | 3.62 | 4.62 | 2.69 |

Where: The numbers in parentheses represent the classification assigned to the equation in the considered test. 01 (Spurr or combined variable); 02 (Spurr or combined variable without b0); 03 (Logarithmized and retransformed non-linear Spurr); 04 (Simple linear as a function of D); 05 (Simple linear as a function of H); 06 (Simple linear as a function of DH).

Some statistical tests are not present for the equation without the $b_0$ coefficient, because they were performed in an ANOVA table with a different TSS that does not consider the correction.

It is observed that the results in the tests which use the mean square of the residual (MSR) were similar in terms of classifying the equations, only changing the value of the test. The same occurs for AIC, $AIC_c$ and BIC, with the exception of the case in which the dependent variable has many values less than zero (0.0). It is also observed that the use of many similar tests does not change the decision of choosing the best equation.

The equation that presented the best Weighted Value (WV) result was the non-linear Spurr equation (Eq. 03). However, doubts may arise about which equations are statistically similar or different, meaning that even with different values for the applied tests, the equations can be considered similar at a certain significance level $\alpha$.

The graphical analysis of residuals in percentage is an important tool in the modeling processes because it can help in the decision by showing how the data behave (linearly or non-linearly after the adjustment has been completed), especially when many equations present similar statistics.

It is observed that the distribution of residuals (Figure 1) corroborates the results of the weighted values, with linear (Eq. 01) and non-linear (Eq. 03) Spurr equations presenting the best distributions, while the last three (Eq. 04, 05 and 06) showed discrepant values (outliers) or bias in estimates (Eq. 02 and 06).

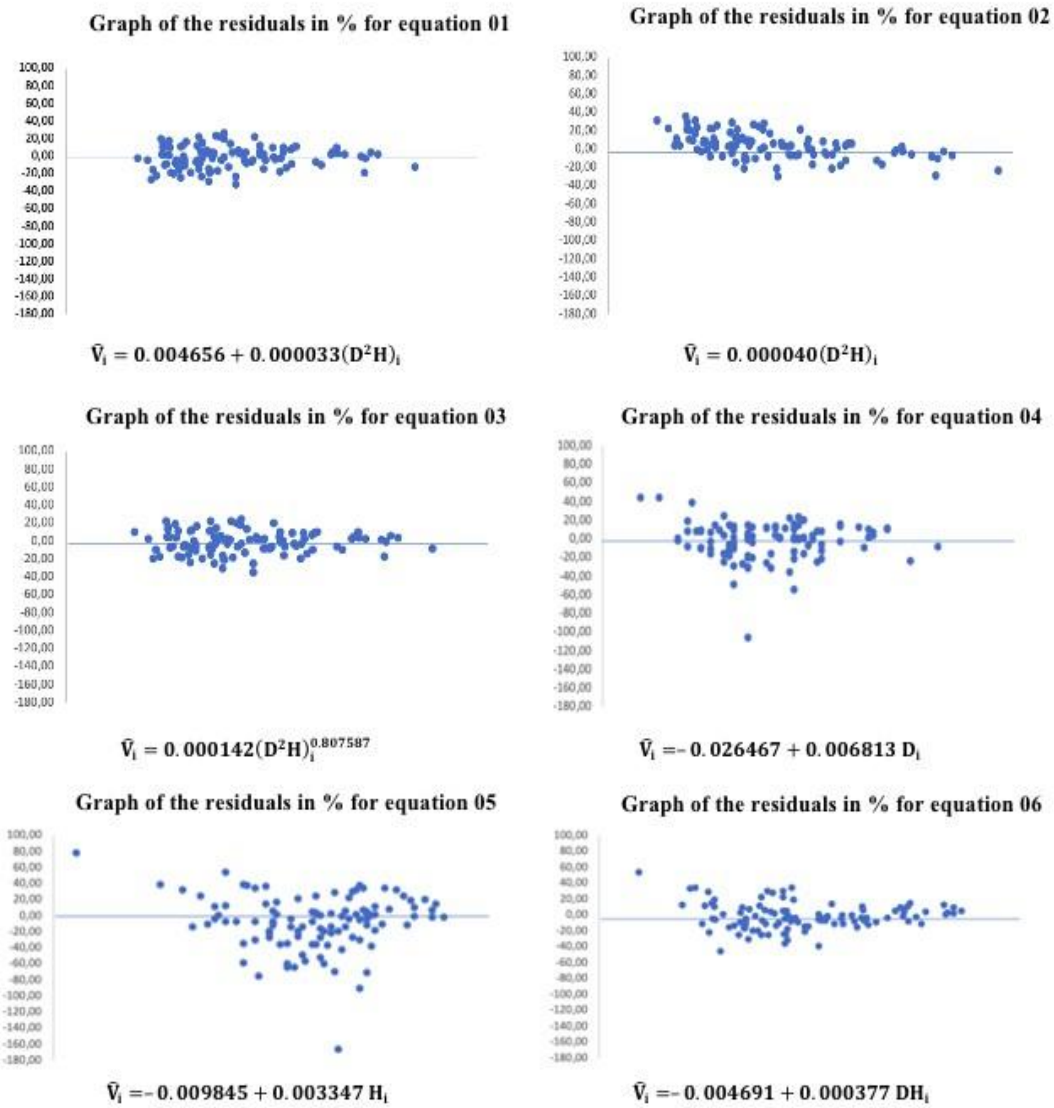


**Figure 1.** Graphic distribution of residuals (%) for the fitted equations for *Eucalyptus* spp. at 7.5 years old in Araripina, PE, Brazil.

The ANOVA (Table 4) shows that the F test was highly significant, indicating that at least one treatment differs statistically from the others ($\alpha$=1%). In applying Tukey's mean comparison test at 5% and 1% significance levels, we have that the values $q_{(\alpha;n,n'df)}$ with seven treatments (control and six equations) and 728 degrees of freedom for the residual are $q_{(0.05;7.728df)} = 4.18$ and $q_{(0.01;7.728df)c} = 4.90$. So, for comparisons between treatments by Tukey's test, the minimal significative difference (MSD) values at 5% and 1% significance levels will respectively be: 0.00004438 and 0.00005202 $(m^3)^2$.

**Table 4.** ANOVA for the seven treatments (control and six equations)

| FV | DF | SS | MS | F |
|---|---|---|---|---|
| Treatments | 6 | 3.2027325E-7 | 5.3378875E-8 | 4,51 ** |
| Residual | 728 | 8.6172098E-6 | 1.1836827E-8 | |
| Total | 734 | 8.9374830E-6 | | |

$F_{(0.01;6.728df)} = 2.80$

By applying Duncan's test at the 5% significance level, we obtain:

a) Involving two means, $z_{(0.05;2.105df)}$= 2.78. MSD = 0.00002951 $(m^3)^2$
b) Involving three means, $z_{(0.05;3.105df)}$= 2.9234. MSD = 0.00003104 $(m^3)^2$
c) Involving four means, $z_{(0.05;4.105df)}$= 3.0214. MSD = 0.00003208 $(m^3)^2$
d) Involving five means, $z_{(0.05;5.105df)}$= 3.0939. MSD = 0.00003285 $(m^3)^2$
e) Involving six means, $z_{(0.05;6.105df)}$= 3.1505. MSD = 0.00003345 $(m^3)^2$
f) Involving seven means, $z_{(0.05;7.105df)}$= 3.1967. MSD = 0.00003394 $(m^3)^2$

By applying Duncan's test at the 1% significance level, we obtain:

a) Involving two means, $z_{(0.01;2.105df)}$= 3.6525. MSD = 0.00003878 $(m^3)^2$
b) Involving three means, $z_{(0.01;3.105df)}$= 3.8060. MSD = 0.00004041 $(m^3)^2$
c) Involving four means, $z_{(0.01;4.105df)}$= 3.9105. MSD = 0.00004152 $(m^3)^2$
d) Involving five means, $z_{(0.01;5.105df)}$= 3.9887. MSD = 0.00004235 $(m^3)^2$
e) Involving six means, $z_{(0.01;6.105df)}$= 4.0509. MSD = 0.00004301 $(m^3)^2$
f) Involving seven means, $z_{(0.01;7.105df)}$= 4.1027. MSD = 0.00004356 $(m^3)^2$

Sorting the means in descending order, the Tukey's and Duncan's tests show the following results (Table 5):

**Table 5.** Tukey's and Duncan's tests at 5% and 1% significance levels

| **Equations** | **Means (SSR) $(m^3)^2$** | Tukey's test | | Duncan's test | |
|---|---|---|---|---|---|
| | | **5%** | **1%** | **5%** | **1%** |
| $\hat{V}_i = 0.000040(D^2H)_i$ | 0.00011304 | a | a | a | a |
| **Real volumes (control)** | **0.00008393** | a | a b | a b | a b |
| $\hat{V}_i = 0.004656 + 0.000033(D^2H)_i$ | 0.00007675 | a b | a b | b | a b c |
| $\hat{V}_i = 0.000142(D^2H)_i^{0.807587}$ | 0.00007656 | a b | a b | b | a b c |
| $\hat{V}_i = -0.004691 + 0.000377\ DH_i$ | 0.00007143 | a b | a b | b | a b c |
| $\hat{V}_i = -0.026467 + 0.006813\ D_i$ | 0.00006884 | b | a b | b | b c |
| $\hat{V}_i = -0.009845 + 0.003347\ H_i$ | 0.00003669 | b | b | c | c |

Means united by the same letters do not differ from each other by Tukey's and Duncan's tests at 5% and 1% significance levels.

For the Tukey's test at the 5% level of significance, it is observed that the linear Spurr equations with (Eq. 1) and without $b_0$ (Eq. 2), non-linear (Eq. 3) and with the independent variable DH (Eq. 6) provide estimates which do not statistically differ from the real volumes and are similar to each other, while the Spurr equation without $b_0$ (Eq. 2) and the real volumes differ from the equation which has the tree height as an independent variable.

Tukey's test at the 1% significance level does not show a significant difference from the equations for the real volume, since the area of acceptance of the null hypothesis $H_0$ is greater than the area of acceptance of $H_0$ at the 5% significance level.

Considering that Duncan's test is more flexible than Tukey's test in terms of rejecting $H_0$, its use results in an increase in the probability of committing a type I or α error, meaning do nor rejecting $H_1$ when $H_0$ is true. This type of error corresponds to the significance level, which is the maximum probability of a type I or α error occurring, which means rejecting the null hypothesis when it actually occurs (Silva & Silva, 1995).

Pimentel Gomes (1970) explains that the Duncan test is more flexible in finding significant differences because the probability level of do not reject $H_0$ decreases as the number of means involved in the contrast increases as a function of $(\alpha)^{n-1}$. For example, considering the probability level of 99% (α =0.99) for three means, the probability will be $(0.99)^{3-1}$= 0.9801, and involving four means it would be $(0.99)^{4-1}$= 0.9703.

In the example under consideration for the 1% level of probability in the contrast involving the seven means, the real probability of do not reject $H_0$ is not 0.99 because $(0.99)^{7-1}$= 0.9415. At the 95% probability level it is $(0.95)^{7-1}$= 0.7351.

The values $d_{(0.05;6.728df)} = 2.57$ and $d_{(0.01;6.728df)} = 3.00$ of the Dunnett bilateral test at the 5% and 1% significance levels, with 12 degrees of freedom obtain MSD equal to 0.00003859 and 0.00004685 $(m^3)^2$, respectively.

It is observed by the Dunnett's test (Table 6) for both the 5% and 1% significance levels that only the equation which has the tree height as an independent variable differs from the control (real data). This corroborates with the others, except Tukey's test at 1% (Table 6).

**Table 6.** Dunnett's test at 5% and 1% significance levels

| $\bar{V}$ | $\bar{\bar{V}}$ | $\bar{V} - \bar{\bar{V}}$ | 5% | 1% |
|---|---|---|---|---|
| $\bar{V}$ – 02 = 0.00008393 | – 0.00011304 = | – 0. 00002911 | n.s. | n.s. |
| $\bar{V}$ – 01 = 0.00008393 | – 0.00007675 = | 0.00000718 | n.s. | n.s. |
| $\bar{V}$ – 03 = 0.00008393 | – 0.00007656 = | 0.00000737 | n.s. | n.s. |
| $\bar{V}$ – 06 = 0.00008393 | – 0.00007143 = | 0.00001250 | n.s. | n.s. |
| $\bar{V}$ – 04 = 0.00008393 | – 0.00006884 = | 0.00001509 | n.s. | n.s. |
| $\bar{V}$ – 05 = 0.00008393 | – 0.00003669 = | 0.00004724 | * | ** |

Where: $\bar{V}$ (mean of the real volumes); $\bar{\bar{V}}$ (mean of the estimated volumes); 01 (Spurr); 02 (Spurr without $b_0$); 03 (non-linear Spurr); 04 (Simple linear as a function of DBHP); 05 (Simple linear as a function of H; 06 (Simple linear as a function of DH); n.s. = not significant; * = significant at a level of 5%; ** = significant at a level of 1%.

An SSR = 11836.83 $(cm^3)^2$ was obtained by applying the Scott-Knott test. The next step is to calculate $SSB_0$ for the possible partitions.

a) For the first partition (1) versus (2), (3), (4), (5), (6) and (7), $SSB_0$ = 1740.9897
b) For the second partition (1) and (2) versus (3), (4), (5), (6) and (7), $SSB_0$ = 1424.4384
c) For the third partition: (1), (2) and (3) versus (4), (5), (6) and (7), $SSB_0$ = 1400.5833
d) For the fourth partition: (1), (2), (3) e (4) versus (5), (6) and (7), $SSB_0$ = 1330.5836
e) For the fifth partition: (1), (2), (3), (4) and (5) versus (6) and (7), $SSB_0$ = 1502.5283

f) For the sixth partition: (1), (2), (3), (4), (5) and (6) versus (7), $SSB_0$ = 1632.3248

The first partition had the largest $SSB_0$, resulting in $\lambda = 20.6861^{**}$. The values of $\chi^2_{(0.05;6.13df)}$ are between 12.59 for 6 df and 14.07 for 7 df. The values for $\chi^2_{(0.01;6.13df)}$ are between 16.81 for 6 df and 18.48 for 7 df. Thus, it is admitted that (1) differs from (2), (3), (4), (5), (6) and (7), being isolated as a group. The partitions in (2), (3), (4), (5), (6) and (7) are verified in the next step.

a) First partition: (2) versus (3), (4), (5), (6) and (7). $SSB_0$ = 201.8132
b) Second partition: (2) and (3) versus (4), (5), (6) and (7). $SSB_0$ = 408.8002
c) Third partition: (2), (3) and (4) versus (5), (6) and (7). $SSB_0$ = 490.3420
d) Fourth partition: (2), (3), (4), (5) versus (6) and (7). $SSB_0$ = 834.3337
e) Fifth partition: (2), (3), (4), (5) and (6) versus (7). $SSB_0$ = 1170.5004

Therefore, the fifth partition was the one that presented the greatest SSB0, with $\lambda = 14.1782^{*}$. The values of $\chi^2_{(0.05;5.26\,df)}$ are between 11.07 for 5 df and 12.59 for 6 df. The values for $\chi^2_{(0.01;6.13df)}$are between 15.09 for 5 df and 16.81 for 6 df. Thus, it is admitted that (1) differs from (2), (3), (4), (5), (6) and (7) for the 5% significance level and it is accepted that it is the only group for the 1% significance level.

The next partition was analyzing the group (2), (3), (4), (5) and (6), and was performed at the 5% significance level because this group joins with (7) at the 1% significance level. So the next step is to check the partitions in (2), (3), (4), (5) and (6).

a) First partition: (2) versus (3), (4), (5) and (6). $SSB_0$ = 55.4778
b) Second partition: (2) e (3) versus (4), (5) and (6). $SSB_0$ = 96.0157
c) Third partition: (2), (3) e (4) versus (5) and (6). $SSB_0$ = 78.0208
d) Fourth partition: (2), (3), (4), (5) versus (6). $SSB_0$ = 88.7890

The second partition had the highest $SSB_0$ with $\lambda = 1.1781^{n.s.}$. The values of $\chi^2_{(0.05;4.38\,df)}$ are between 9.49 for 4 df and 11.07 for 5 df. Thus, $H_0$ is not rejected forming a group.

Next, there are three groups of equations by the Scott-Knott test (Table 7) at the 5% significance level; the first consists of the equation that has height as the independent variable, a second group with the Spurr equation without $b_0$, and the third (which is the most important because it contains the real volumes) is composed by the linear and non-linear Spurr equation, and the equations that have D and DH as independent variables. There is no participation of any equation in more than one group, which is an advantage of the Scott-Knott test, especially when the number of treatments is high. The result at 1% significance is similar to the other tests.

**Table 7.** Summary of the Scott-Knott test

| Equations | Means of SSR $(cm^3)^2$ | 5% | 1% |
|---|---|---|---|
| $\hat{V}_i = -0.009845 + 0.003347\,H_i$ (01) | 36.69 | a | a |
| $\hat{V}_i = -0.026467 + 0.006813\,D_i$ (02) | 68.84 | b | b |
| $\hat{V}_i = -0.004691 + 0.000377\,DH_i$ (03) | 71.43 | b | b |
| $\hat{V}_i = 0.000142(D^2H)_i^{0.807587}$ (04) | 76.56 | b | b |
| $\hat{V}_i = 0.004656 + 0.000033(D^2H)_i$ (05) | 76.75 | b | b |
| Real volumes (06) | 83.93 | b | b |
| $\hat{V}_i = 0.000040(D^2H)_i$ (07) | 113.04 | c | b |

Therefore, the separation of means tests are options for comparing equations and forming groups of equations. The results presented corroborate those found by the various tests considered, and it has the advantage of forming groups of similar equations.

Several studies on equation selection using regression analysis in forestry science have been published, including: Vanclay; Skovsgaard, 1997; Guendehou *et al*., 2012; Viana; Fernandes; Aranha, 2013; Zhang *et al*., 2014; Cysneiros *et al*., 2017; Santos *et al*., 2018. These studies used different statistical methods, but always seeking to select a better discovery, even if they present similar estimates, but the selected research most often presents a minimal difference (certainly not significant) from the others. Thus, the research studied may have independent variables that are more difficult to measure than other research studies that present similar results, but the tests used did not allow grouping.

The choice of the best equation within a group will be a decision made by the modeler, considering different statistical and applicability criteria, emphasizing that the best equation is the one that presents good precision with the smallest and easiest number of measurable variables. Also, the cost to measure independent variables can be considered when selecting the best equation of a group of similar equations.

## 4. Conclusions

Compared with the classic criteria expressed by the WV, the results of the means tests showed similar results. Another possible advantage of this methodology is that one equation is compared with another based on estimates and not on their coefficients. This methodology which has been applied to simple linear models and a simple non-linear model can be applied to multiple linear and non-linear models using the same procedures.

### Conflicts of Interest

The authors declare no conflict of interest.

### Author Contributions

**Conceptualization:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Data curation:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Formal analysis:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Funding acquisition:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Investigation:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Methodology:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Project administration:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Software:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Resources:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Supervision:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Validation:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Visualization:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Writing - original draft:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M. **Writing - review and editing:** SILVA, J. A. A., FERREIRA, R. L. C., BERGER, R., AVAREZ-LAZO, D. A., GUERA, O. G. M.

## References


1. Akaike, H. A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19** (6), 716-723 (1974).
2. Bozdogan, H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987). DOI: https://doi.org/10.1007/BF02294361
3. Cysneiros, V. C.; Pelissari, A. L.; Machado, S. A.; Filho, A. F. & Souza, L. Modelos genéricos e específicos para a estimativa do volume comercial em uma floresta sob concessão na Amazônia, *Sci. For.* **45** (114), 295-304 (2017). DOI: https://doi.org/10.18671/scifor.v45n114.06
4. Duncan, D. B. Multiple range and multiple F tests, *Biometrics* **11**, 1-42 (1955). DOI: 10.2307/3001478
5. Dunnet, C. W. New tables for multiple comparisons with control, *Biometrics* **40**, 482-491 (1964). DOI:
https://doi.org/10.2307/2528490
6. Fahrmeir, L.; Kneib, T.; Lang, S. & Marx, B. *Regression, models, methods and applications* (Springer-Verlag, 2013).
7. Figueiredo-Filho, D.; Nunes, F.; Rocha, E. C.; Santos, M. L.; Batista, M. & Silva-Júnior, J. A. O que fazer e não fazer com a regressão: pressupostos e aplicações do modelo linear de mínimos quadrados ordinários (MQO). *Revista Política Hoje* **20** (1), 44-99 (2011).
8. Furnival, G. M. An index for comparing equation used in construction of volume table. *Forest Science* **7** (4), 337-341 (1961). DOI: https://doi.org/10.1093/forestscience/7.4.337
9. Gouveia, J. F.; Silva, J. A.A.; Ferreira, R. L. C.; Gadelha, F. H. L. & Lima-Filho, L. M. A. Modelos volumétricos mistos em clones de Eucalyptus no polo gesseiro do Araripe, Pernambuco. *Floresta* **45** (3), 587-598 (2015). DOI: https://doi.org/10.5380/rf.v45i3.36844
10. Guendehou, G. H. S.; Lehtonen, A.; Moudachirou, M.; Mäkipää, R. & Sinsin, B. Volume models of selected tropical tree species in West Africa. *Southern Forests: Journal of Forest Science* **74** (2), 77-88 (2012). DOI: https://doi.org/10.2989/20702620.2012.701432
11. Marquardt, D. W. & Snee, R. D. Test statistics for mixture models. *Technometrics* **16** (4), 533-537 (1974). DOI: https://doi.org/10.2307/1267604
12. Mehtätalo, L. & Lappi, J. *Biometry for forestry and environmental data with examples in R* (CRC Press, 2020).
13. Montgomery, D. C.; Peck, E. A. & Vining, G. G. *Introduction to linear regression analysis* (John Wiley & Sons Pte. Ltd., 2013).
14. Pimentel-Gomes, F. *Curso de Estatística Experimental* (Livraria Nobel, 4a Ed., 1970).
15. Santos, K. N. F.; Rode, R.; Andrade, D. F. C.; Correa, K. S. & Lopes, L. S. S. Ajuste de equações volumétricas e redes neurais artificiais na estimativa do volume de tauari na floresta nacional do Tapajós. *Agroecossistemas* **10** (1), 1 – 17 (2018). DOI: https://doi.org/10.18542/ragros.v10i1.5130
16. Schlaegel, B. D. Testing, reporting, and using biomass estimation models. *In: Proceedings of 1981*, Southern Forest Biomass Workshop, Clemson University, Clemson, South Carolina. 95-112 (1981).
17. Schneider, P. R.; Schneider, P. S. P. & Souza, C. A. M. *Análise de regressão aplicada à Engenharia Florestal* (FACOS-UFSM, 2 ed., 2009).
18. Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* **6** (2), 461-464 (1978). http://www.jstor.org/stable/2958889
19. Scott, A. J. & Knott, M. A. A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30**, 507-5042 (1974). DOI: https://doi.org/10.2307/2529204
20. Segel, L. A. & Edelstein, L. *A premier on Mathematical models in Biology* (SIAM-Society for Industrial and Applied Mathematics, 2013).

21. Silva, J. A. A. & Ferreira, R. L. C. Influences of determined and estimated dendrometric variables on the precision of volumetric modelling. *Journal of Forest Science* **67** (12), 553–561 (2021). DOI: https://doi.org/10.17221/105/2021-JFS
22. Silva, J. A. A. & Silva, A, I. P. *Estatística experimental aplicada à ciência florestal*. (Imprensa Universitária da UFRPE, 2 ed., 1995).
23. Snedecor, G. W. Calculation and interpretation of analysis of variance and covariance. *Journal of Farm Economics* **16** (4), 728-730 (1934). DOI: https://doi.org/10.1037/13308-000
24. Spurr, S. H. *Forest inventory* (New York: Ronald Press, 1952).
25. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* **5**, 99-114 (1949). DOI: https://doi.org/10.2307/3001913
26. Viana, H.; Fernandes, P. & Aranha, J. Equações para estimar a biomassa aérea das principais lenhosas arbustivas no Norte e Centro do País. *Silva Lusitana*, 99-109 (2013).
27. Vanclay, J.K. & Skovsgaard, J. P. Evaluating forest growth models. *Ecological Modelling* **98**, 1-12 (1997). DOI: https://doi.org/10.1016/S0304-3800(96)01932-1
28. Zhang, X.; Duan, A.; Zhang, J. & Xiang, C. Estimating tree height-diameter models with the Bayesian method. *Hindawi Publishing Corporation, Scientific World Journal* (Article ID 683691, 9 p., 2014). DOI: https://doi.org/10.1155/2014/683691