SCIENTIA FORESTALIS

ISSN 2318-1222 (online)

ORIGINAL ARTICLE

Variable transformation strategies in forest biomass modeling

Estratégias de transformação de variáveis na modelagem da biomassa florestal

Jobert Silva da Rocha[1], Allan Libanio Pelissari[1], Luan Demarco Fiorentin[1], Luciano Rodrigo Lanssanova[2], Vinicius Costa Cysneiros[3], Carla Krulikowski Rodrigues[1]

[1]Universidade Federal do Paraná – UFPR, Curitiba, PR, Brasil
[2]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso – IFMT, Juína, MT, Brasil
[3]Universidade Federal de Santa Catarina – UFSC, Curitibanos, SC, Brasil

How to cite: Rocha, J. S., Pelissari, A. L., Fiorentin, L. D., Lanssanova, L. R., Cysneiros, V. C., & Rodrigues, C. K. (2023). Variable transformation strategies in forest biomass modeling. *Scientia Forestalis*, *51*, e3976. https://doi.org/10.18671/scifor.v51.26

Abstract

In linear regression, residual normality and homoscedasticity assumptions must be satisfied for correct statistical inferences, which are commonly ignored in forest biomass modeling. In these cases, variable transformations can be applied as corrective measures, whose knowledge represents a scientific gap in forest science. This study aimed to evaluate data transformation strategies in linear regression models to predict total forest biomass and its components. Data for fitting linear regression models were sampled in native stands of *Mimosa scrabella* in Southern Brazil. Trees were sampled to determine the mass of stem, branch, leaf, bark, and root components. Data transformations were assessed to fit biomass with the diameter as a predictor. Models that simultaneously satisfied the residual normality and homoscedasticity assumptions were statistically evaluated. Modeling of total biomass and its components resulted in improved accuracy by applying variable transformations. The Root-Root transformation was indicated for predicting total and stem biomass. In addition, Log-Log transformation was recommended for estimating branch and bark biomass, as well as Manly for leaf biomass and Dual power for root biomass.

Keywords: Allometry; Native forest; Statistical assumptions; Homoscedasticity; Residual normality.

Resumo

Na regressão linear, as suposições de normalidade residual e homocedasticidade devem ser satisfeitas para corretas inferências estatísticas, as quais são comumente ignoradas na modelagem da biomassa florestal. Nesses casos, as transformações de variáveis podem ser aplicadas como medidas corretivas, cujo conhecimento representa uma lacuna científica na ciência florestal. Este estudo teve como objetivo avaliar estratégias de transformação de dados em modelos de regressão linear para predizer a biomassa florestal total e seus componentes. Os dados para ajuste de modelos de regressão linear foram amostrados em povoamentos nativos de *Mimosa scrabella* no Sul do Brasil. Foi realizada uma amostragem de árvores para determinar a massa dos componentes fuste, galhos, folhas, cascas e raízes. As transformações de dados foram avaliadas no ajuste da biomassa, tendo o diâmetro como preditor. Os modelos que atenderam simultaneamente as suposições de normalidade residual e homocedasticidade foram avaliados estatisticamente. A modelagem da biomassa total e seus componentes resultou em maior precisão por meio da aplicação de transformações de variáveis. A transformação Raiz-Raiz foi indicada para predizer as biomassas total e do fuste. Além disso, a transformação Log-Log foi recomendada para estimar as biomassas dos galhos e das cascas, bem como Manly para a biomassa das folhas e Dual power para a biomassa radicular.

Palavras-chave: Alometria; Floresta nativa; Suposições estatísticas; Homoscedasticidade; Normalidade residual.

Financial support: None.
Conflict of interest: Nothing to declare.
Corresponding author: allanpelissari@gmail.com
Received: 16 december 2022.
Accepted: 16 august 2023.
Editor: Mauro Valdir Schumacher.

INTRODUCTION

Linear regression is widely applied in forest mensuration for its practicality and easy interpretation, which is essential in allometric variable prediction, such as forest biomass and its components (Brown et al., 1989; Lima et al., 2012; Behling et al., 2018, 2019). Parametric analysis techniques require statistical assumptions about the necessary data. In this context, it is emphasized that the normal distribution of residues and the homoscedasticity are essential for regression modeling (Fox, 2016).

For its application in quantifying carbon in the climate change context (He et al., 2021; Huang et al., 2021; Li et al., 2022; Raiho et al., 2022), correct inferences in biomass predictions are essential in forest modeling. Despite its importance, statistical assumptions are commonly ignored, as noted by Cysneiros et al. (2020). The lack of attention to correct statistical inference is often observed in recent publications about the application of biomass equations (Mate et al., 2014; Anitha et al., 2015; Brahma et al., 2021; David et al., 2022).

Variables transformations can be applied when linear regression assumptions are not satisfied (Long & Ervin, 2000; Alexopoulos, 2010; Weisberg, 2014). These data transformations change the variable values by adding constants, raising to a power, converting to logarithmic scales, taking square root, as well as other methods (Box & Cox, 1964; Osborne, 2002; Rojas-Perilla, 2018; Medina et al., 2019). However, transformations are often applied without considering their theoretical and mathematical aspects, which result in a lack of information for allometric modeling (Eastaugh & Hasenauer, 2011; Monness, 2011; Fayolle et al., 2018).

By employing appropriate transformations on the variables, it becomes possible to fit statistically rigorous models that satisfy the assumptions of linear regression, providing accurate statistical inferences (Casella & Berger, 2002; Liu, 2011) on total biomass and its components. This property allows us to infer that the accuracy and efficiency of sample modeling are valid for biomass stock prediction in population inventories of native forest species for management and conservation efforts.

Therefore, this study aimed to evaluate data transformation strategies in linear regression models to predict total forest biomass and its components. Based on the results, we expect to recommend the most appropriate transformations that satisfy the residual normality and homoscedasticity assumptions for correct inferences in allometric modeling.

MATERIAL AND METHODS

DATABASE

Data were collected in four native stands of *Mimosa scrabella* Benth. located in the metropolitan region of Curitiba, Parana State, Brazil (Figure 1). This species is characterized as a pioneer, fast-growing, frost-resistant, and used as firewood and charcoal (Urbano et al., 2021), with economic and ecological importance in southern mixed forests. The climate of the region is Cfb-Köppen, with hot and rainy summers, cold winters with occasional dry periods, an average annual temperature of 17ºC, and average annual rainfall between 1,400 and 1,600 mm (Souza et al., 2014).

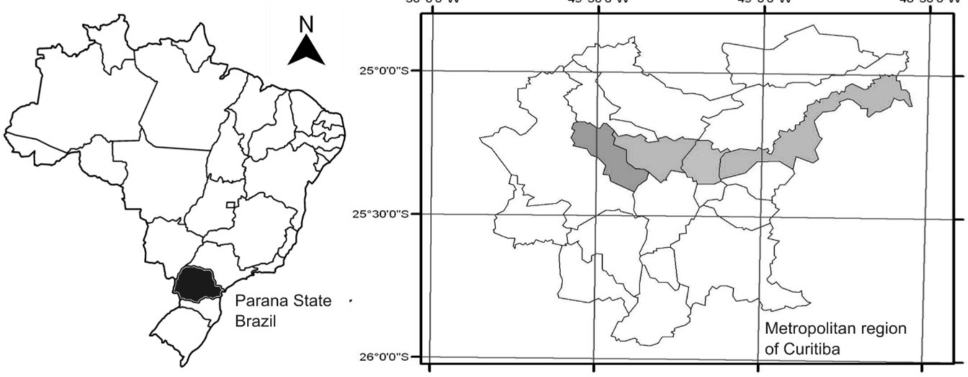

Figure 1. Location map of four Mimosa scrabella native stands sampled in the metropolitan region of Curitiba, Parana State, Brazil.

A sampling of 30 trees was carried out to measure the diameters at 1.3 m above ground (d) and determine the mass of stem, branch, leaf, bark, and root components (Figure 2). For stems, a cross-sectional wood sample was collected at mid-height. For branches, leaf, bark, and root, random samples of 1 kg were obtained for each component. Samples were dried in an oven at 65 ± 5°C to constant mass. The biomass of branch, leaf, bark, and root was calculated according to Equation 1.

$$w_c = \frac{mu_c \, . ms_a}{mu_a} \tag{1}$$

where w_c is the dry biomass of component (kg), mu_c is the total wet mass of component (kg), ms_a is the dry mass of component sample (kg), and mu_a is the wet mass of component sample (kg).

Basic wood density (ρ) was considered for stem biomass, which was determined by the ratio between dry mass and transversal section volume of stem samples. Based on the sampled average basic density ($\bar{\rho}$) and stem volumes (v), stem biomass was calculated using Equation 2. The total biomass (w_t) was obtained by the sum of component biomass stocks.

$$w_{stem} = \bar{\rho}.v \tag{2}$$

where w_{stem} is the dry stem biomass (kg), $\bar{\rho}$ is the average basic wood density (kg m^{-3}), and v is the stem volume (m^{-3}).

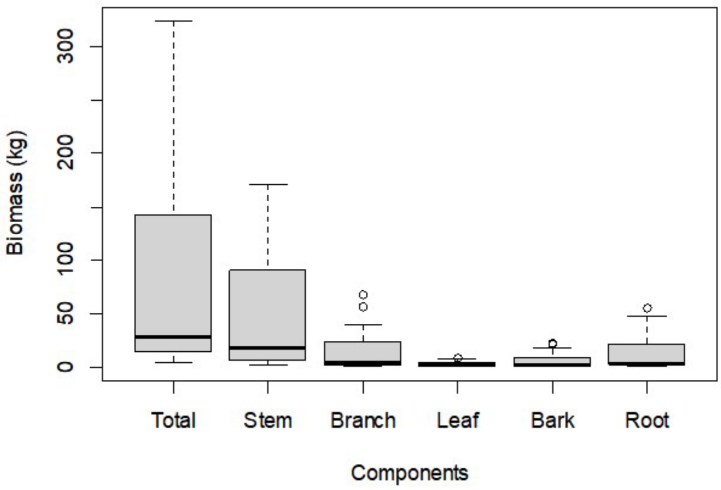

Figure 2. Summary of tree variables sampled for modeling total biomass and its components.

VARIABLE TRANSFORMATION METHODS

Transformation strategies were applied to fit the biomass as a function of diameter (Table 1). In the strategies in which a lambda parameter (λ) was estimated using maximum likelihood estimation, transformations were applied to the biomass variables. For strategies in which it was not required to estimate λ, the transformation procedure was performed for biomass variable or simultaneously for biomass and diameter variables. Therefore, in addition to the non-transformation, 15 strategies were fitted by R program, version 4.0.5 (R Core Team, 2021), using trafo package, version 1.0.1 (Medina et al., 2019).

$\mathcal{Y}$ is the response variable (total biomass and its components), $\mathcal{X}$ is the predictor variable (diameter at 1.3 m above ground), β_i is the regression coefficient, λ is the transformation parameter, and $\mathcal{Y}'$ is the transformed response variable.

Table 1. Summary of variables transformation strategies applied for modeling total biomass and its components.

Strategy	Model	Retransformation
Non-transformation	$\mathcal{Y} = \beta_0 + \beta_1\mathcal{X}$	not applied
Log	$\log(\mathcal{Y}) = \beta_0 + \beta_1\mathcal{X}$	$\mathcal{Y} = e^{\mathcal{Y}'}$
Reciprocal	$\frac{1}{\mathcal{Y}} = \beta_0 + \beta_1\mathcal{X}$	$\mathcal{Y} = \frac{1}{\mathcal{Y}'}$
Box-Cox	$\begin{cases} \frac{\mathcal{Y}^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda \neq 0 \\ \log(\mathcal{Y}) = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda = 0 \end{cases}$	$\begin{cases} \mathcal{Y} = (\lambda\mathcal{Y}' + 1)^{\frac{1}{\lambda}} & \text{if } \lambda \neq 0 \\ \mathcal{Y} = e^{\mathcal{Y}'} & \text{if } \lambda = 0 \end{cases}$
Manly	$\begin{cases} \frac{e^{\lambda\mathcal{Y}} - 1}{\lambda} = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda \neq 0 \\ \mathcal{Y} = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda = 0 \end{cases}$	$\begin{cases} \mathcal{Y} = \frac{log(\lambda\mathcal{Y}) + 1}{\lambda} & \text{if } \lambda \neq 0 \\ \mathcal{Y} = \mathcal{Y}' & \text{if } \lambda = 0 \end{cases}$
Bickel-Doksum	$\frac{\lvert\mathcal{Y}\rvert^{\lambda}\text{sign}(\mathcal{Y}) - 1}{\lambda} = \beta_0 + \beta_1\mathcal{X} \quad \text{if } \lambda > 0$	$\pm[sign(\mathcal{Y})(\lambda\mathcal{Y} + 1)]^{\frac{1}{\lambda}}$
Yeo-Johnson	$\begin{cases} \frac{(\mathcal{Y}+1)^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1\mathcal{X} & if\ \lambda \neq 0, \mathcal{Y} \geq 0 \\ log(\mathcal{Y}+1) = \beta_0 + \beta_1\mathcal{X} & if\ \lambda = 0, \mathcal{Y} \geq 0 \\ \frac{(1-\mathcal{Y})^{2-\lambda} - 1}{\lambda - 2} = \beta_0 + \beta_1\mathcal{X} & if\ \lambda \neq 2, \mathcal{Y} < 0 \\ -log(1-\mathcal{Y}) = \beta_0 + \beta_1\mathcal{X} & if\ \lambda = 2, \mathcal{Y} < 0 \end{cases}$	$\begin{cases} \mathcal{Y} = (\lambda.\mathcal{Y}' + 1)^{\frac{1}{\lambda}} - 1 & if\ \lambda \neq 0, \mathcal{Y} \geq 0 \\ \mathcal{Y} = e^{\mathcal{Y}'} & if\ \lambda = 0, \mathcal{Y} \geq 0 \\ 1 - [-\mathcal{Y}(2-\lambda) + 1]^{\frac{1}{2-\lambda}} & if\ \lambda \neq 2, \mathcal{Y} < 0 \\ -e^{-\mathcal{Y}} + 1 & if\ \lambda = 2, \mathcal{Y} < 0 \end{cases}$
Glog	$\log\left(\mathcal{Y} + \sqrt{\mathcal{Y}^2 + 1}\right) = \beta_0 + \beta_1\mathcal{X}$	$\mathcal{Y} = -\frac{[1 - (e^{\mathcal{Y}'})^2]}{2e^{\mathcal{Y}'}}$
Dual power	$\begin{cases} \frac{(\mathcal{Y}^{\lambda} - \mathcal{Y}^{-\lambda})}{\lambda} = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda \neq 0 \\ \log(\mathcal{Y}) = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda = 0 \end{cases}$	$\begin{cases} \mathcal{Y} = \left[\sqrt{1 + \lambda^2\mathcal{Y}'^2} + \lambda\mathcal{Y}'\right]^{\frac{1}{\lambda}} & \text{if } \lambda \neq 0 \\ \mathcal{Y} = e^{\mathcal{Y}'} & \text{if } \lambda = 0 \end{cases}$
G power	$\begin{cases} \frac{(\mathcal{Y} + \sqrt{\mathcal{Y}^2 + 1})^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda \neq 0 \\ \log(\mathcal{Y} + \sqrt{\mathcal{Y}^2 + 1}) = \beta_0 + \beta_1\mathcal{X} & \text{if } \lambda = 0 \end{cases}$	$\begin{cases} -\left[\frac{1 - (\lambda\mathcal{Y} + 1)^{\frac{2}{\lambda}}}{2(\lambda\mathcal{Y} + 1)^{\frac{1}{\lambda}}}\right] & \text{if } \lambda \neq 0 \\ -\frac{(1 - e^{\mathcal{Y}^2})}{2e^{\mathcal{Y}}} & \text{if } \lambda = 0 \end{cases}$
Log shift	$\log(\mathcal{Y} + \lambda) = \beta_0 + \beta_1\mathcal{X}$	$\mathcal{Y} = e^{\mathcal{Y}'} - \lambda$
Shifted square root	$\sqrt{\mathcal{Y} + \lambda} = \beta_0 + \beta_1\mathcal{X}$	$\mathcal{Y} = (\mathcal{Y}'^2 - \lambda)$
Log-Log	$\log(\mathcal{Y}) = \beta_0 + \beta_1\log(\mathcal{X})$	$\mathcal{Y} = e^{\mathcal{Y}'}$
Reciprocal-Reciprocal	$\frac{1}{\mathcal{Y}} = \beta_0 + \beta_1\frac{1}{\mathcal{X}}$	$\mathcal{Y} = \frac{1}{\mathcal{Y}'}$
Root-Root	$\sqrt{\mathcal{Y}} = \beta_0 + \beta_1\sqrt{\mathcal{X}}$	$\mathcal{Y} = \mathcal{Y}'^2$

The logarithmic scale is one of the most used transformations in models with non-normal residual distribution and heteroscedasticity (Feng et al., 2014). It was applied to the biomass variables (Log) and, simultaneously, to the biomass and diameter variables (Log-Log). The reciprocal transformation is useful for data with negative asymmetry (Medina et al., 2019) and was equally applied to the biomass variables and the biomass and diameter variables.

The Box-Cox strategy (Box & Cox, 1964) is one method of potential transformation, which applies a transformation parameter (λ) to the variable values. The parameter λ is iteratively estimated by maximum likelihood estimation (Fischer, 2016). The Manly transformation (Manly, 1976) was proposed for distributions with negative values as an alternative to the gaps of Box-Cox, becoming effective in converting unimodal distributions in almost symmetric ones (Chortirat et al., 2011).

The Bickel-Doksum power transformation (Bickel & Doksum, 1981) was designed for data with leptokurtic and platykurtic distributions (Medina et al., 2019). The Yeo-Johnson transformation was proposed to fit asymmetric distributions to normality, whether in positive or negative datasets (Yeo & Johnson, 2000). Durbin et al. (2002) proposed the Glog, also called generalized logarithm transformation, for normalizing and stabilizing variances of values close to zero.

Dual power was proposed by Yang (2006) due to the limitation of Box-Cox parameter λ, which does not transform the data to $\lambda = 1$, allowing to transform the data for any λ value. In addition, Kelmansky & Ricci (2017) proposed the G power as an extension of Glog for negative values and heavy-tailed distributions.

Log shift is an extension of Log strategy by adding a transformation parameter λ. This strategy is useful for correcting non-normality (Feng et al., 2016; Rojas-Perilla, 2018) by transforming asymmetric and log-normal distributions in symmetric ones. Also mentioned is its usefulness for data with excessive asymmetry, heteroscedasticity, and outlier observations (Feng et al., 2016).

The Shifted square root was proposed as an extension of the traditional square root, in which a parameter λ is estimated (Rojas-Perilla, 2018; Medina et al., 2019). The Root-Root transformation is a strategy used for stabilization of variances (Medina et al., 2019), since it is commonly applied for right-skewed distributions and recommended for data with extremely small values (Rojas-Perilla, 2018).

STATISTICAL EVALUATION OF REGRESSION MODELS

Exploratory analyses were carried out for the databases, using histograms, scatter plots, and Pearson linear correlation. In addition, linear regression coefficients (β_i) were estimated by ordinary least squares, in which the standard errors in percentage ($SE\beta_i\%$) and the significance of fitted coefficients were evaluated at the 5% probability level by the t-test. The statistical criteria of adjusted coefficient of determination ($R^2_{adj.}$), standard error of the estimate in percentage (SEE%), Bayesian Information Criterion (BIC), and the dispersion of observed by estimated values were reported for the best models with retransformed variables.

Quantile-quantile plots, in association with the Lilliefors test (Lilliefors, 1967) at 5% probability level, were used to examine the residual distribution to normality distribution by means of the transformation strategies. The dispersion of studentized residuals, in combination with the Breusch-Pagan test (Breusch & Pagan, 1979) at 5% probability level, were applied to verify the homoscedasticity.

RESULTS

The diameter at 1.3 m above ground (d) and biomass components showed distributions with positive asymmetry and linear tendency (Figure 3). Additionally, Pearson correlation between variable d and total biomass and its components corroborated the linear association at 5% probability level. Although there was evidence of high variability of biomass values in the database, the correlations were greater than 0.9 for total biomass and its components.

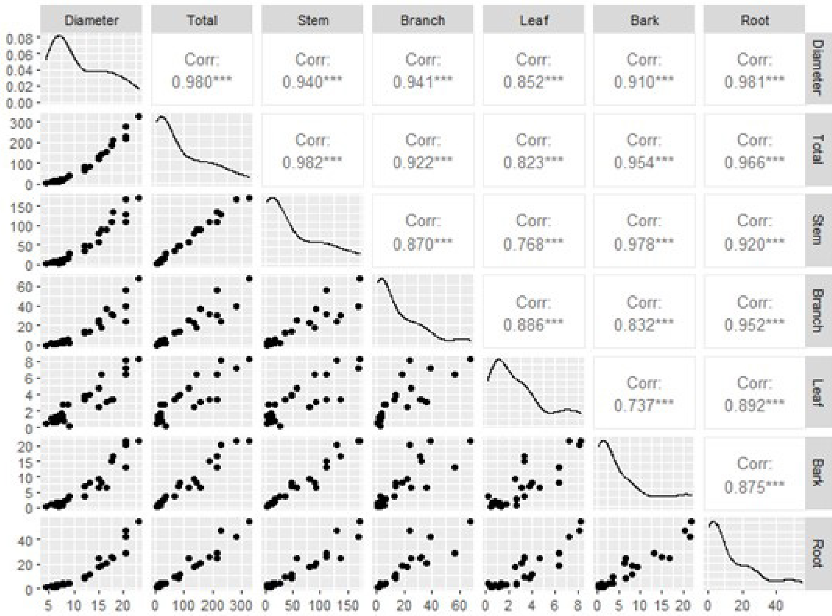

Figure 3. Exploratory analysis of diameter at 1.3 m above ground and total biomass and its components.

The estimated regression coefficients (β_i) were significant at 5% probability level for the main transformation strategies (Table 2). In addition, the fitted models were statistically more efficient with a reduction in residual standard error in percentage ($SE\beta_i\%$) of regression coefficients. The $SE\beta_0\%$ reduction for intercept coefficient (β_0) is highlighted for total and stem biomass with Root-Root transformations; with Log-Log for branch and bark biomass; with Shifted square root for leaf biomass; and with Box-Cox for root biomass (Table 2).

Table 2. Regression statistics of main transformation strategies for modeling total biomass and its components.

	Strategy	β_0	$SE\beta_0\%$	β_1	$SE\beta_1\%$	λ	$R^2_{adj.}$	SEE%	BIC	p-Li	p-BP
					Total biomass (w_{total})						
1	Non-transformation	-98.6074*	8.59	16.2546*	4.17		0.952	24.35	274.18	0.249	0.007
5	Manly	-43.2309*	5.63	8.6841*	2.27		0.969	19.59	261.13	0.188	0.142
15	Root-root	-11.6437*	4.45	5.9727*	2.59		0.986	16.59	251.17	0.101	0.615
					Stem biomass (w_{stem})						
1	Non-transformation	-56.1524*	10.18	9.2715*	4.93		0.934	28.75	250.59	0.291	0.003
4	Box-Cox	-3.7330*	12.62	1.0640*	3.54	0.439	0.965	26.85	246.48	0.995	0.063
15	Root-root	-9.0584*	6.27	4.5774*	3.71		0.962	24.12	240.05	0.958	0.374
					Branch biomass (w_{branch})						
1	Non-transformation	-18.1727*	15.15	2.9189*	7.55		0.857	45.50	206.78	0.012	0.001
2	Log	-0.7567*	23.74	0.2343*	6.15		0.901	73.06	235.19	0.171	0.961
13	Log-Log	-4.4179*	6.52	2.7340*	4.48		0.945	38.82	197.25	0.512	0.452
					Leaf biomass (w_{leaf})						
1	Non-transformation	-1.4353*	29.82	0.3806*	9.01		0.809	36.37	95.09	0.044	0.006
5	Manly	-0.3369*	68.36	0.2088*	8.86	-0.161	0.813	34.71	92.30	0.225	0.443
12	Shifted square root	0.7528*	14.00	0.0947*	8.87	0.790	0.812	35.13	93.01	0.121	0.260
					Bark biomass (w_{bark})						
1	Non-transformation	-6.9853*	13.95	1.1523*	6.77		0.882	39.51	144.48	0.308	0.001
7	Yeo-Johnson	-0.6177*	20.42	0.2105*	4.79	0.143	0.937	35.34	137.78	0.124	0.259
13	Log-Log	-4.8551*	6.99	2.5630*	5.63		0.916	33.06	133.79	0.324	0.159
					Root biomass (w_{root})						
1	Non-transformation	-15.8604*	13.04	2.5311*	6.54		0.889	39.74	189.61	0.478	0.001
4	Box-Cox	-1.8668*	7.79	0.4130*	2.81	0.290	0.978	23.14	157.17	0.591	0.236
9	Dual power	-1.4523*	8.63	0.3342*	2.99	0.427	0.975	22.99	156.78	0.885	0.425

β_i is the regression coefficient, * is significant at the 5% probability level, $SE\beta_i\%$ is the percentage standard error of regression coefficients, λ is the transformation parameter, $R^2_{adj.}$ is the adjusted coefficient of determination, SEE% is the standard error of the estimate in percentage, BIC is the Bayesian Information Criterion, p-Li is the p-value of Lilliefors test, and p-BP is the p-value of Breusch-Pagan test.

The Manly transformations for total and leaf biomass, as well as Box-Cox for stem and root biomass, were the most effective in reducing the residual standard error in percentage ($SE\beta_1\%$) of the angular coefficient (β_1). Additionally, Log-Log for branch biomass and Yeo-Johnson strategies for bark biomass resulted in a decrease in $SE\beta_1\%$ (Table 2).

The transformations of variables resulted in models with higher adjusted coefficients of determination ($R^2_{adj.}$) and lower values of standard error of estimate (SEE%) and Bayesian Information Criterion (BIC). In addition, the non-transformation strategy did not result in biomass models with homoscedasticity by Breusch-Pagan test (p-BP), as well as residual normality by Lilliefors test (p-Li) for branch and leaf biomass (Table 2), both at the 5% probability level. Therefore, the inferences from these models are not considered to be statistically reliable.

The best statistics are highlighted for models with Root-Root transformation for total and stem biomass, Log-Log for branch and bark biomass, Manly for leaf biomass, and Dual power for root biomass (Table 2). In addition, these models resulted in homoscedasticity (p-BP) and residual normality (p-Li) at the 5% of probability level.

The observed biomass data showed linearized dispersion by transformations, when compared to their predicted values (Figure 4). This feature had a positive influence on the statistical performance of fitted models. Additionally, lowest dispersions were noted with Root-Root transformations for total (w_{total}) and stem (w_{stem}) biomass, while higher dispersion resulted from leaf biomass (w_{leaf}), in which this non-woody biomass component has a lower correlation with diameter.

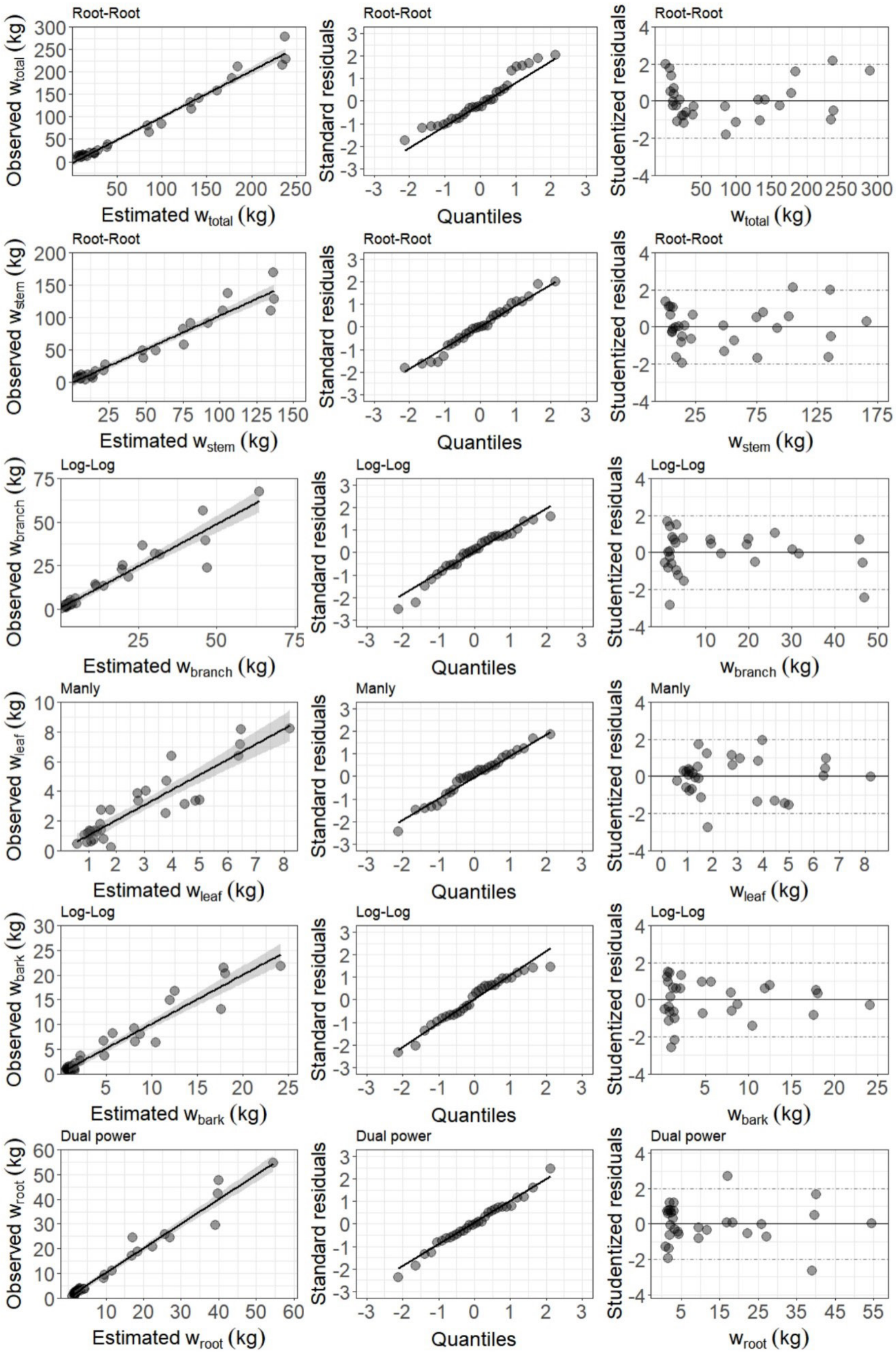

Figure 4. Scatter (left), quantile-quantile (middle), and studentized residual (right) plots of best transformation strategies for modeling total biomass and its components.

The quantile-quantile plots corroborated the results of Lilliefors test (Table 2), showing the presence of normally distributed residuals, evidenced by the lack of extreme quantiles in tail distributions (Figure 4). Otherwise, the studentized residuals showed balanced dispersions.

DISCUSSION

Modeling the relationship between variables is critical for predictions involving linear regression (Long & Ervin, 2000; Pretzsch, 2009; Weisberg, 2014; Dunn & Smyth, 2018). Therefore, it is highlighted that the variable transformations can be used to solve the statistical misspecification in allometric models (Porté & Bartelink, 2002; Pretzsch, 2009). Therefore, the transformation strategies in the present study provided accurate fits with correct statistical inferences for predicting total biomass and its components (Table 2).

Variable transformation strategies showed statistics improvements for fitting total biomass and its components (Table 2 and Figure 4), mainly from satisfying residual normality and homoscedasticity. It is emphasized that statistically incorrect allometric models were published, related to the lack of normally distributed residuals and the presence of heteroscedasticity, as highlighted by Cysneiros et al. (2020). Therefore, Sileshi (2014) related these models to common statistical problems, such as unreliable regression coefficient estimations. The data transformations proposed in this study provided a solution to satisfy these regression assumptions.

In the present study, modeling of total biomass and its components with the diameter at 1.3 m as a predictor resulted in improved accuracy applying variable transformations (Table 2). The presence of a strong linear association between the response variables and diameter may have contributed to the success of transformation strategies. These statistical considerations allow the conclusion that simple models based on a single predictor can be applied to estimate allometric variables (Picard et al., 2015; Forrester et al., 2017; Santos et al., 2020). However, in general, native forest inventories only provide access to the diameter variable, making it the most commonly available measure (Vibrans et al., 2015; Roik et al., 2018).

In addition, for biomass modeling with variable transformation, in which only the residual normality is satisfied, other strategies can be combined to ensure homoscedasticity. Since the estimates of regression coefficients are directly influenced by heteroscedasticity, this issue has been widely discussed in the literature, highlighting the heteroscedasticity-consistent covariance matrix (Long & Ervin, 2000), weighted least squares (Zea-Camaño et al., 2020; Coutinho et al., 2021), data mining (Sanquetta et al., 2015; Debastiani et al., 2019) and generalized linear modeling (Fox, 2016; Dunn & Smyth, 2018).

Since heteroscedasticity is commonly observed in forest biomass modeling (Dutcă et al., 2022), the weighted linear regression would become a feasible alternative as a support to biomass variable transformations present in this study. Some authors have proposed weighting functions for heteroscedastic biomass models, mainly applying the inverse of the predictor variable by a power (Picard et al., 2012; McRoberts & Westfall, 2014; Kralicek et al., 2017). From this perspective, the knowledge of data transformations showed in this study represents an advance in forest biomass modeling and carbon assessments, considering the high costs and operational difficulties that limit sampling of tree biomass.

The data variability in native forest inventories can influence the accuracy of biomass estimations (Li et al., 2019; Liepiņš et al., 2022). In this context, variable transformation, as applied in the present study, can enhance the quality of predictive fit. In addition, variable transformation can also improve models for projecting biomass growth and yield in planted forests, such as *Eucalyptus* and *Pinus* plantations, in which high estimation errors are common (Schikowski et al., 2013; Hall et al., 2020). This assumption is essential, as biomass shows typical characteristics of a biological variable over time (Eloy et al., 2018; Repo et al., 2021). The temporal biomass dynamics may influence the correlation of biomass production with predictor variables (Dean & Cao, 2003; Sağlam & Sakici, 2023), for which different transformation methods can be applied to enhance predictions.

In this study, we observed the difficulty of developing more accurate models for biomass components, such as branch, leaf, and bark, primarily due to the absence of distinct predictor variables beyond the diameter above the ground. While variable transformations certainly enhance the accuracy of models, the incorporation of crown dimensions can significantly improve the quality of estimates (Goodman et al., 2014), particularly for non-woody components such as leaf biomass (Djomo & Chimi, 2017; Timilsina et al., 2017). Additional variables, like wood density, hold the potential to function as predictors, further enhancing the overall model quality for estimating total biomass and its constituent components (Guangyi et al., 2017; Romero et al., 2022).

CONCLUSIONS

Based on results of the present study, variable transformations are recommended when at least one of the assumptions of residual normality and homoscedasticity has not been satisfied in linear modeling of forest biomass. Root-Root transformation is indicated for predicting total and stem biomass. In addition, Log-Log transformation is recommended for estimating branch and bark biomass, as well as Manly for leaf biomass and Dual power for root biomass.

REFERENCES

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, *14*(1, Suppl.1), 23-28. PMid:21487487.

Anitha, K., Verchot, L. V., Joseph, S., Manuri, S., & Avitabile, V. (2015). A review of forest and tree plantation biomass equations in Indonesia. *Annals of Forest Science*, *72*(8), 981-997. http://dx.doi.org/10.1007/s13595-015-0507-4.

Behling, A., Péllico Netto, S., Sanquetta, C. R., Corte, A. P. D., Simon, A. A., Rodrigues, A. L., & Caron, B. O. (2019). Additive and non-additive biomass equations for black wattle. *Floresta e Ambiente*, *26*(4), 1-10. http://dx.doi.org/10.1590/2179-8087.043917.

Behling, A., Péllico Netto, S., Sanquetta, C. R., Corte, A. P. D., Affleck, D. L. R., Rodrigues, A. L., & Behling, M. (2018). Critical analyses when modeling tree biomass to ensure additivity of its components. *Anais da Academia Brasileira de Ciências*, *90*(2), 1759-1774. PMid:29791562. http://dx.doi.org/10.1590/0001-3765201820170684.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, *76*(374), 296-311. http://dx.doi.org/10.1080/01621459.1981.10477649.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B. Methodological*, *26*(2), 211-243. http://dx.doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brahma, B., Nath, A. J., Deb, C., Sileshi, G. W., Sahoo, U. K., & Kumar, A. (2021). A critical review of forest biomass estimation equations in India. *Trees. Forests and People*, *5*, 100098. http://dx.doi.org/10.1016/j.tfp.2021.100098.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287-1294. http://dx.doi.org/10.2307/1911963.

Brown, S., Gillespie, A., & Lugo, A. E. (1989). Biomass estimation methods for tropical forests with applications to forest inventory data. *Forest Science*, *35*(4), 881-902.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed). Pacific Grove: Thomson Learning.

Chortirat, T., Chomtee, B., & Sinsomboonthong, J. (2011). Comparison of four data transformation methods for Weibull distributed data. *Agriculture and Natural Resources (Bangkok)*, *45*(2), 366-383.

Coutinho, V. M., Oliveira, T. W. G., Fiorentin, L. D., Sanquetta, M. N. I., Sanquetta, C. R., & Corte, A. P. D. (2021). How to estimate black wattle aboveground biomass from heteroscedastic data? *Floresta*, *51*(1), 28-36. http://dx.doi.org/10.5380/rf.v51i1.65236.

Cysneiros, V. C., Gaui, T. D., Silveira Filho, T. B., Pelissari, A. L., Machado, S. D. A., Carvalho, D. C., & Amorim, H. B. (2020). Tree volume modeling for forest types in the Atlantic Forest: generic and specific models. *IForest (Viterbo)*, *13*(5), 417-425. http://dx.doi.org/10.3832/ifor3495-013.

David, H. C., Barbosa, R. I., Vibrans, A. C., Watzlawick, L. F., Trautenmuller, J. W., Rafaelo, B., Ribeiro, S. C., Jacovineh, L. A. G., Corte, A. P. D., Sanquetta, C. R., Silva, A. C., Freitas, J. V., & MacFarlane, D. W. (2022). The tropical biomass & carbon project–an application for forest biomass and carbon estimates. *Ecological Modelling*, *472*, 110067. http://dx.doi.org/10.1016/j.ecolmodel.2022.110067.

Dean, T. J., & Cao, Q. V. (2003). Inherent correlations between stand biomass variables calculated from tree measurements. *Forest Science*, *49*(2), 279-284.

Debastiani, A. B., Sanquetta, C. R., Dalla Corte, A. P., Pinto, N. S., & Rex, F. E. (2019). Evaluating SAR-optical sensor fusion for aboveground biomass estimation in a Brazilian tropical forest. *Annals of Forest Research*, *62*(2), 109-122. http://dx.doi.org/10.15287/afr.2018.1267.

Djomo, A. N., & Chimi, C. D. (2017). Tree allometric equations for estimation of above, below and total biomass in a tropical moist forest: case study with application to remote sensing. *Forest Ecology and Management*, *391*, 184-193. http://dx.doi.org/10.1016/j.foreco.2017.02.022.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R* (573 p.). New York: Springer. http://dx.doi.org/10.1007/978-1-4419-0118-7.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., & Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics (Oxford, England)*, *18*(1, Suppl.1), S105-S110. PMid:12169537. http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S105.

Dutcă, I., McRoberts, R. E., Næsset, E., & Blujdea, V. N. B. (2022). Accommodating heteroscedasticity in allometric biomass models. *Forest Ecology and Management*, *505*, 119865. http://dx.doi.org/10.1016/j.foreco.2021.119865.

Eastaugh, C. S., & Hasenauer, H. (2011). Incorporating management history into forest growth modelling. *IForest (Viterbo)*, *4*(5), 212-217. http://dx.doi.org/10.3832/ifor0597-004.

Eloy, E., Silva, D. A. D., Caron, B. O., Elli, E. F., & Schwerz, F. (2018). Effect of age and spacing on biomass production in forest plantations. *Revista Árvore*, *42*(2), http://dx.doi.org/10.1590/1806-90882018000200014.

Fayolle, A., Ngomanda, A., Mbasi, M., Barbier, N., Bocko, Y., Boyemba, F., Couteron, P., Fonton, N., Kamdem, N., Katembo, J., Kondaoule, H. J., Loumeto, J., Maidou, H. M., Mankou, G., Mengui, T., Mofack, G., Moundounga, C., Moundounga, Q., Nguimbous, L., Nchama, N. N., Obiang, D., Asue, F. O. M., Picard, N., Rossi, V., Senguela, Y., Sonké, B., Viard, L., Yongo, O. D., Zapfack, L., & Medjibe, V. P. (2018). A regional allometry for the Congo basin forests based on the largest ever-destructive sampling. *Forest Ecology and Management*, *430*, 228-240. http://dx.doi.org/10.1016/j.foreco.2018.07.030.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*(2), 105-109. PMid:25092958. http://dx.doi.org/10.3969/j.issn.1002-0829.2014.02.009.

Feng, Q., Hannig, J., & Marron, J. S. (2016). A note on automatic data transformation. *Stat*, *5*(1), 82-87. http://dx.doi.org/10.1002/sta4.104.

Fischer, C. (2016). Comparing the logarithmic transformation and the Box-Cox transformation for individual tree basal area increment models. *Forest Science*, *62*(3), 297-306. http://dx.doi.org/10.5849/forsci.15-135.

Forrester, D. I., Tachauer, I. H. H., Annighoefer, P., Barbeito, I., Pretzsch, H., Ruiz-Peinado, R., Stark, H., Vacchiano, G., Zlatanov, T., Chakraborty, T., Saha, S., & Sileshi, G. (2017). Generalized biomass and leaf area allometric equations for European tree species incorporating stand structure, tree age and climate. *Forest Ecology and Management*, *396*, 160-175. http://dx.doi.org/10.1016/j.foreco.2017.04.011.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (817 p.). Atlanta: Sage.

Goodman, R. C., Phillips, O. L., & Baker, T. R. (2014). The importance of crown dimensions to improve tropical tree biomass estimates. *Ecological Applications*, *24*(4), 680-698. PMid:24988768. http://dx.doi.org/10.1890/13-0070.1.

Guangyi, M., Yujun, S., & Saeed, S. (2017). Models for predicting the biomass of *Cunninghamia lanceolata* trees and stands in southeastern China. *PLoS One*, *12*(1), e0169747. PMid:28095512. http://dx.doi.org/10.1371/journal.pone.0169747.

Hall, K. B., Stape, J., Bullock, B. P., Frederick, D., Wright, J., Scolforo, H. F., & Cook, R. (2020). A growth and yield model for *Eucalyptus benthamii* in the Southeastern United States. *Forest Science*, *66*, 25-37. http://dx.doi.org/10.1093/forsci/fxz061.

He, X., Lei, X., & Dong, L. (2021). How large is the difference in large-scale forest biomass estimations based on new climate-modified stand biomass models? *Ecological Indicators*, *126*, 107569. http://dx.doi.org/10.1016/j.ecolind.2021.107569.

Huang, C., Liang, Y., He, H. S., Wu, M. M., Liu, B., & Ma, T. (2021). Sensitivity of aboveground biomass and species composition to climate change in boreal forests of Northeastern China. *Ecological Modelling*, *445*, 109472. http://dx.doi.org/10.1016/j.ecolmodel.2021.109472.

Kelmansky, D. M., & Ricci, L. (2017). A new distribution family for microarray data. *Microarrays (Basel, Switzerland)*, *6*(1), 5. PMid:28208652. http://dx.doi.org/10.3390/microarrays6010005.

Kralicek, K., Huy, B., Poudel, K. P., Temesgen, H., & Salas, C. (2017). Simultaneous estimation of above- and below-ground biomass in tropical forests of Viet Nam. *Forest Ecology and Management*, *390*, 147-156. http://dx.doi.org/10.1016/j.foreco.2017.01.030.

Li, Y., Li, C., Li, M., & Liu, Z. (2019). Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms. *Forests*, *10*(12), 1073. http://dx.doi.org/10.3390/f10121073.

Li, Y., Li, M., & Wang, Y. (2022). Forest aboveground biomass estimation and response to climate change based on remote sensing data. *Sustainability*, *14*(21), 14222. http://dx.doi.org/10.3390/su142114222.

Liepiņš, J., Lazdiņš, A., Kalēja, S., & Liepiņš, K. (2022). Species composition affects the accuracy of stand-level biomass models in hemiboreal forests. *Land (Basel)*, *11*(7), 1108. http://dx.doi.org/10.3390/land11071108.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399-402. http://dx.doi.org/10.1080/01621459.1967.10482916.

Lima, A. J. N., Suwa, R., Mello Ribeiro, G. H. P., Kajimoto, T., Santos, J., Silva, R. P., & Higuchi, N. (2012). Allometric models for estimating above-and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro, Brazil. *Forest Ecology and Management*, *277*, 163-172. http://dx.doi.org/10.1016/j.foreco.2012.04.028.

Liu, W. (2011). *Simultaneous inference in regression.* Boca Raton: CRC Press.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217-224.

Manly, B. F. (1976). Exponential data transformations. *The Statistician*, *25*(1), 37-42. http://dx.doi.org/10.2307/2988129.

Mate, R., Johansson, T., & Sitoe, A. (2014). Biomass equations for tropical forest tree species in Mozambique. *Forests*, *5*(3), 535-556. http://dx.doi.org/10.3390/f5030535.

McRoberts, R. E., & Westfall, J. A. (2014). Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. *Forest Science*, *60*(1), 34-42. http://dx.doi.org/10.5849/forsci.12-141.

Medina, L., Kreutzmann, A. K., Rojas-Perilla, N., & Castro, P. (2019). The R package *trafo* for transforming linear regression models. *The R Journal*, *9*(2), 99-123. http://dx.doi.org/10.32614/RJ-2019-054.

Monness, E. (2011). The power-normal distribution: application to forest stands. *Canadian Journal of Forest Research*, *41*(4), 707-714. http://dx.doi.org/10.1139/x10-246.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, *8*(6), 1-7.

Picard, N., Rutishauser, E., Ploton, P., Ngomanda, A., & Henry, M. (2015). Should tree biomass allometry be restricted to power models? *Forest Ecology and Management*, *353*, 156-163. http://dx.doi.org/10.1016/j.foreco.2015.05.035.

Picard, N., Saint-André, L., & Henry, M. (2012). *Manual for building tree volume and biomass allometric equations: from field measurement to prediction* (215 p.). Rome: FAO.

Porté, A., & Bartelink, H. H. (2002). Modelling mixed forest growth: a review of models for forest management. *Ecological Modelling*, *150*(1-2), 141-188. http://dx.doi.org/10.1016/S0304-3800(01)00476-8.

Pretzsch, H. (2009) *Forest dynamics, growth and yield: from measurement to model* (664 p.). Berlin: Springer. http://dx.doi.org/10.1007/978-3-540-88307-4.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Raiho, A. M., Paciorek, C. J., Dawson, A., Jackson, S. T., Mladenoff, D. J., Williams, J. W., & McLachlan, J. S. (2022). 8000-year doubling of Midwestern forest biomass driven by population- and biome-scale processes. *Science*, *376*(6600), 1491-1495. http://dx.doi.org/10.1126/science.abk3126.

Repo, A., Rajala, T., Henttonen, H. M., Lehtonen, A., Peltoniemi, M., & Heikkinen, J. (2021). Age-dependence of stand biomass in managed boreal forests based on the Finnish National Forest Inventory data. *Forest Ecology and Management*, *498*, 119507. http://dx.doi.org/10.1016/j.foreco.2021.119507.

Roik, M., Machado, S. D. A., Figueiredo Filho, A., Sanquetta, C. R., Roveda, M., & Stepka, T. F. (2018). Diameter increment modeling in an araucaria forest fragment using cluster analysis. *Floresta e Ambiente*, *25*(3), e20170625. http://dx.doi.org/10.1590/2179-8087.062517.

Rojas-Perilla, N. (2018). *The use of data-driven transformations and their applicability in small area estimation* (Doctoral thesis). Freie Universität Berlin, Berlin.

Romero, F. M. B., Jacovine, L. A. G., Torres, C. M. M. E., Ribeiro, S. C., Da Rocha, S. J. S. S., Novais, T. D. N. O., Gaspar, R. D. O., Da Silva, L. F., Vidal, E., Leite, H. G., Staudhammer, C. L., & Fearnside, P. M. (2022). Aboveground biomass allometric models for large trees in southwestern Amazonia. *Trees. Forests and People*, *9*, 100317. http://dx.doi.org/10.1016/j.tfp.2022.100317.

Sağlam, F., & Sakici, O. E. (2023). Allometric relationships between above-ground biomass increment and stand characteristics for Crimean pine in Taşköprü, Turkey. *Cerne*, *29*, e-103169. http://dx.doi.org/10.1590/01047760202329013169.

Sanquetta, C. R., Wojciechowski, J., Dalla Corte, A. P., Behling, A., Péllico Netto, S., Rodrigues, A. L., & Sanquetta, M. (2015). Comparison of data mining and allometric model in estimation of tree biomass. *BMC Bioinformatics*, *16*(1), 247. PMid:26250142. http://dx.doi.org/10.1186/s12859-015-0662-5.

Santos, M. F., Figueiredo Filho, A., Gama, J. R. V., Retslaff, F. A. S., & Costa, D. L. (2020). Species-specific equations: greater precision in commercial volume estimation in managed forests in the amazon. *Cerne*, *26*(3), 315-330. http://dx.doi.org/10.1590/01047760202026032741.

Schikowski, A. B., Corte, A. P. D., & Sanquetta, C. R. (2013). Modelagem do crescimento e de biomassa individual de *Pinus. Pesquisa Florestal Brasileira*, *33*(75), 269-278. http://dx.doi.org/10.4336/2013.pfb.33.75.503.

Sileshi, G. W. (2014). A critical review of forest biomass estimation models, common mistakes and corrective measures. *Forest Ecology and Management*, *329*, 237-254. http://dx.doi.org/10.1016/j.foreco.2014.06.026.

Souza, R. F., Machado, A. S., & Figueiredo Filho, A. (2014). Modelagem da biomassa total e da lenha por unidade de área para bracatingais nativos. *Revista Árvore*, *38*(1), 53-62. http://dx.doi.org/10.1590/S0100-67622014000100005.

Timilsina, N., Beck, J. L., Eames, M. S., Hauer, R., & Werner, L. (2017). A comparison of local and general models of leaf area and biomass of urban trees in USA. *Urban Forestry & Urban Greening*, *24*, 157-163. http://dx.doi.org/10.1016/j.ufug.2017.04.003.

Urbano, E., Machado, S. A., Figueiredo Filho, A., & Sanquetta, C. R. (2021). Modeling and dynamics of growth and yield of tree species in *Mimosa scabrella* stands. *Anais da Academia Brasileira de Ciências*, *93*(1), e20180891. PMid:33787682. http://dx.doi.org/10.1590/0001-3765202120180891.

Vibrans, A. C., Moser, P., Oliveira, L. Z., & Maçaneiro, J. P. D. (2015). Height-diameter models for three subtropical forest types in southern Brazil. *Ciência e Agrotecnologia*, *39*(3), 205-215. http://dx.doi.org/10.1590/S1413-70542015000300001.

Weisberg, S. (2014). *Applied linear regression* (370 p.). Minneapolis: John Wiley & Sons.

Yang, Z. (2006). A modified family of power transformations. *Economics Letters*, *92*(1), 14-19. http://dx.doi.org/10.1016/j.econlet.2006.01.011.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959. http://dx.doi.org/10.1093/biomet/87.4.954.

Zea-Camaño, J. D., Soto, J. R., Arce, J. E., Pelissari, A. L., Behling, A., Orso, G. A., & Eisfeld, R. D. L. (2020). Improving the modeling of the height–diameter relationship of tree species with high growth variability: robust regression analysis of *Ochroma pyramidale* (Balsa-Tree). *Forests*, *11*(3), 1-22. http://dx.doi.org/10.3390/f11030313.

Author contributions: JSR: investigation, writing – original draft; ALP: conceptualization, methodology, writing – review & editing; LDF: data curation, formal analysis; LRL: validation, writing – review & editing; VCC: validation, writing – review & editing; CKR: validation, writing – review & editing.