



UNIVERSIDADE ESTADUAL DA PARAÍBA  
CENTRO DE CIÊNCIAS E TECNOLOGIA  
DEPARTAMENTO DE ESTATÍSTICA

Bruno Carneiro de Medeiros

**USO DE MODELOS ADITIVOS  
GENERALIZADOS PARA LOCAÇÃO, ESCALA  
E FORMA NA AVALIAÇÃO DO  
CRESCIMENTO DO EUCÁLIPTO.**

Campina Grande - PB

Agosto de 2019

Bruno Carneiro de Medeiros

**USO DE MODELOS ADITIVOS GENERALIZADOS  
PARA LOCAÇÃO, ESCALA E FORMA NA AVALIAÇÃO  
DO CRESCIMENTO DO EUCALIPTO.**

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

Coorientador: Profa. Dra. Ana Patricia Bastos Peixoto

Campina Grande - PB

Agosto de 2019

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

M488u Medeiros, Bruno Carneiro de.

Uso de modelos aditivos generalizados para locação, escala e forma na avaliação do crescimento do eucalipto [manuscrito] / Bruno Carneiro de Medeiros. - 2019.

27 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2019.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Coordenação do Curso de Estatística - CCT."

"Coorientação: Profa. Dra. Ana Patrícia Bastos Peixoto, Coordenação do Curso de Estatística - CCT."

1. Modelos Lineares Generalizados. 2. Análise de regressão. 3. Curvas de crescimento. I. Título

21. ed. CDD 519.5

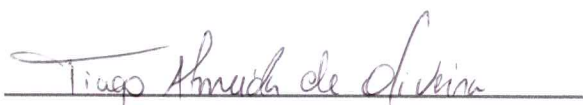
Bruno Carneiro de Medeiros

## USO DE MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA NA AVALIAÇÃO DO CRESCIMENTO DO EUCALIPTO.

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 16 de agosto de 2019.

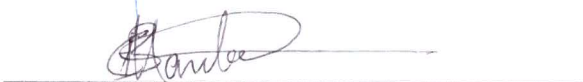
### BANCA EXAMINADORA



Prof. Dr. Tiago Almeida de Oliveira  
Universidade Estadual da Paraíba



Prof. Dr. Silvio Fernando Xavier Junior  
Universidade Estadual da Paraíba



Prof. Me. Ednário Mendonça Barbosa  
Universidade Estadual da Paraíba

# Agradecimentos

Em primeiro lugar a Deus, por me conceder a oportunidade de subir mais um degrau na escada da vida, pois sem ele nada é possível.

A minha mãe Maria do Socorro Carneiro, que me guiou no caminho certo desde os primeiros dias de minha vida, me dando não só a educação, mas todos os ensinamentos necessários para que eu pudesse me tornar a pessoa que sou hoje.

A meu pai Antônio Alves de Medeiros, por estar sempre ao meu lado, mesmo apesar da distância.

A meu irmão Weskley, em quem pude sempre me espelhar nessa vida acadêmica e que sei que posso contar para todos os momentos.

Aos meus filhos Anny Bheatriz e Brenno Henrick, razões para sempre seguir em frente mesmo diante das diversas adversidades as quais a vida nos remete.

A minha esposa Fabiana, pela compreensão e paciência frente a tantas ausências que a vida acadêmica nos submete.

A meu tio Amaro, por sempre me dar apoio.

A meus amigos José Reinaldo, Moniclaudia e muitos outros que aqui peço desculpas por não citar e que me ajudaram tanto durante todos esses anos.

A todos meus familiares, sempre presentes em minha vida.

A minha tia Rosana, que onde quer q esteja sempre estará torcendo por mim.

A todos os professores e profissionais que contribuíram para construir o conhecimento que tenho hoje, mas em especial aos professores Tiago Almeida e Ana Patricia, que além de orientadores são grandes amigos, pois sempre acreditaram em meu potencial, sendo peças mais que essenciais em minha formação acadêmica.

*“Uma resposta aproximada para o problema certo  
vale muito mais do que uma resposta exata para um problema aproximado.”  
(John W. Tukey)*

# Resumo

Os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS) são uma recente classe de modelos que flexibiliza ainda mais a distribuição da variável resposta. A análise de regressão tem sido utilizada para modelar fenômenos biológicos, e suas diversas modalidades têm atendido a necessidade do seu uso com precisão. Porém, existem situações nas quais se torna indispensável o ajuste de modelos com pressupostos mais flexíveis na especificação da distribuição da variável resposta, justificando assim o uso do GAMLSS. O estudo das curvas de crescimento de plantas tem ampla aplicação em pesquisas na área agropecuária, dessa forma, conhecer os hábitos de crescimento e desenvolvimento de espécies florestais é crucial para programas de reflorestamento e nas mais diversas pesquisas. O objetivo do estudo foi modelar o crescimento do eucalipto, por meio do ajuste de Modelos Aditivos Generalizados para Localização, Escala e Forma, afim de promover melhorias na produtividade da cultura. Considerando todos os parâmetros com efeito da variável independente (tempo) sob a modelagem da classe GAMLSS, o modelo de distribuição ST3 foi o que apresentou melhores resultados.

**Palavras-chaves:** Curvas de Crescimento; Eucalypto; GAMLSS.

# Abstract

Generalized Additive Models for Location, Scale and Shape (GAMLSS) are a recent class of model that further flexibilizes the distribution of the response variable. The regression analysis has been used to model biological phenomena, and its various modalities have met the need for its use with precision. However, there are situations in which the adjustment of models with more flexible assumptions in the specification of the distribution of the response variable becomes indispensable, thus justifying the use of GAMLSS. The study of plant growth curves has wide application in agricultural research, thus, know the habits of growth and development of forest species is crucial for reforestation programs and in the most diverse researches. The aim of the study was to model the growth of Eucalyptus, through of the adjusting of Generalized Additive Models for Location, Scale and Shape, in order to promote improvements on crop productivity. Considering all parameters of the independent variable (time) under GAMLSS class modeling, the distribution model ST3 that's what presented better results.

**Key-words:** Growth curves; Eucalytpus; GAMLSS.



# Lista de ilustrações

Figura 1 – Densidade da distribuição ST3, para valores diferentes de $\mu$ , fixando $\sigma = 1$ , $\tau = 2$ e $\nu = 6$ (Fonte: Eloy, 2016.) . . . . .	14
Figura 2 – Densidades Normal, t de Student, Cauchy e ST3, para diferentes valores de $\nu$ , fixando $\mu = 0$ , $\sigma = 1$ e $\tau = 1$ (Fonte: Eloy, 2016). . . . .	15
Figura 3 – Gráfico de dispersão de Tempo <i>versus</i> Altura de Eucaliptos. . . . .	19
Figura 4 – Gráficos de resíduos para a altura de Eucaliptos. . . . .	20
Figura 5 – Worm plot e gráficos residuais das distribuição Normal (média e dispersão). . . . .	21
Figura 6 – Worm plot e gráficos residuais das distribuição SHASHo. . . . .	21
Figura 7 – Worm plot e gráficos residuais das distribuição SHASH. . . . .	21
Figura 8 – Worm plot e gráficos residuais das distribuição ST3. . . . .	22

# Lista de tabelas

Tabela 1 – Critério de informação de Akaike Generalizado dos modelos probabilísticos ajustados via pacote <code>gamlss</code> . . . . .	20
Tabela 2 – Critério de informação de Akaike Generalizado dos modelos probabilísticos ajustados via pacote <code>gamlss</code> , nas diferentes configurações paramétricas.	22
Tabela 3 – Estatísticas dos Quantis Residuais dos modelos propostos . . . . .	23
Tabela 4 – Estatísticas do modelo <code>gamlss</code> ajustado sob a distribuição <i>Skew t</i> tipo 3.	23

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>10</b>
<b>2</b>	<b>REVISÃO DE LITERATURA</b>	<b>11</b>
<b>2.1</b>	<b>Modelos Lineares Generalizados</b>	<b>11</b>
<b>2.2</b>	<b>Modelos Aditivos Generalizados para Localização, Escala e Forma</b>	<b>12</b>
2.2.1	Tipos de distribuições dentro da família GAMLSS	13
2.2.1.1	Distribuição <i>skew-t</i> tipo 3	13
2.2.2	Seleção dos modelos	16
2.2.3	Análise de resíduos	17
<b>3</b>	<b>MATERIAIS E MÉTODOS</b>	<b>18</b>
<b>4</b>	<b>RESULTADOS E DISCUSSÕES</b>	<b>19</b>
<b>5</b>	<b>CONSIDERAÇÕES FINAIS</b>	<b>25</b>
	<b>REFERÊNCIAS</b>	<b>26</b>

# 1 Introdução

Conhecer os hábitos de crescimento e desenvolvimento de espécies florestais é crucial para programas de reflorestamento e nas mais diversas pesquisas. No caso do eucalipto, estudar as curvas de crescimento é de extrema importância, pois é possível definir o tempo ideal de corte, ou de rotação, que corresponde a época em que se obtém a máxima produtividade volumétrica e econômica de um povoamento florestal (GONÇALVES et al., 2017).

O estudo do crescimento de plantas tem ampla aplicação em pesquisas na área agropecuária, porquanto permite ao pesquisador a comparação do comportamento da planta ou de seus componentes relevantes, face às várias particularidades de cada experimento (MAIA et al., 2009). Ao modelar os dados de crescimento, tem-se por objetivo a obtenção de informações acerca dos parâmetros em que a curva é composta, para que assim se possa obter interpretações mais concretas desses parâmetros e de tal modo que seja possível construir um modelo padrão para as observações estudadas (LOPES, 2019).

No que se refere as técnicas de modelagem, a análise de regressão tem sido utilizada para modelar fenômenos biológicos, e suas diversas modalidades têm atendido a necessidade do seu uso com precisão. Porém, quando surgem situações em que há necessidade de ajustar modelos com pressupostos mais flexíveis na especificação da distribuição da variável resposta, (RIGBY; STASINOPOULOS, 2005) introduziram os Modelos Aditivos Generalizados para Localização, Escala e Forma (*Generalized Additive Models for Location, Scale and Shape – GAMLSS*), técnica essa que consegue reunir um grupo de modelos lineares e não-lineares, modelos lineares generalizados (*Generalized Linear Model – GLM*) (NELDER; WEDDERBURN, 1972), modelos aditivos generalizados (GAM) (HASTIE T. ; TIBSHIRANI, 1990) e modelos mistos.

Acredita-se que a classe de modelos GAMLSS é uma forte ferramenta quando se trata de modelagem, já que a mesma torna-se uma poderosa ferramenta que permite modelar não só a média, como nos modelos lineares generalizados (GLM) e modelos aditivos generalizados (GAM), mas também a dispersão, assimetria e a curtose, gerando mais possibilidades para o pesquisador e, em consequência, podendo alcançar melhores ajustes no processo de modelagem.

Diante dos fatos mencionados, o objetivo do estudo é modelar o crescimento do Eucalipto, por meio do ajuste de Modelos Aditivos Generalizados para Localização, Escala e Forma, afim de promover melhorias na produtividade da cultura.

## 2 Revisão de Literatura

Nesta seção serão abordados os principais métodos estatísticos que servirão de base para o entendimento sobre a aplicação dos Modelos Aditivos Generalizados para Localização, Escala e Forma.

### 2.1 Modelos Lineares Generalizados

Os modelos de regressão linear utilizados na análise de dados em diversas áreas da pesquisa levam em consideração que a variável resposta em estudo tenha distribuição normal, sendo este um pressuposto rigoroso e necessário para o seu uso correto, mas elimina uma série de situações práticas, como variáveis respostas binárias, de proporção, de contagem ou ainda categóricas, o que os tornam inadequados em algumas situações relacionadas a pesquisa acadêmica.

De posse dessas informações, Nelder e Wedderburn (1972) sugeriram os Modelos Lineares Generalizados (GLM), cuja ideia básica era expandir as opções de distribuições de probabilidades dos erros associados ao modelo de regressão, e conseqüentemente da variável resposta, de forma a abranger aquelas que pertençam a família exponencial, além de flexibilizar a relação funcional entre a variável resposta e as variáveis explicativas do modelo (PAULA, 2004).

A família exponencial de distribuições garante aos parâmetros uma estatística suficiente. Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias independentes, cada uma com função densidade (ou de probabilidade) na família exponencial da forma

$$f(y_i; \theta_i, \phi) = \exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \psi)], \quad i = 1, \dots, n \quad (2.1)$$

em que,  $b(\cdot)$  e  $c(\cdot)$  são funções conhecidas para cada observação,  $b(\theta_i)$  é o parâmetro natural ou canônico e  $\phi - 1 > 0$  é o parâmetro de dispersão.

Os modelos lineares generalizados envolvem os três componentes, sendo que no componente aleatório as variáveis respostas  $Y_1, Y_2, \dots, Y_n$  são independentes e seguem uma distribuição pertencente à família exponencial, com médias  $\mu_i$  e parâmetro de escala constante  $\phi$  e independentes das  $p$  variáveis explicativas. O componente sistemático é representado por  $g(\mu_i) = \eta_i$ , sendo que,  $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$  é o preditor linear, portanto, a parte sistemática que pode ser utilizada para realizar possíveis previsões para a equação da reta estimada. Já a função de ligação é uma função que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear.

Embora flexibilizem o pressuposto de normalidade dos Modelos de Regressão Linear (MRL), os GLM ainda revelam limitações, dado que julgam que a relação da média é linear às variáveis explicativas, mas nem sempre tal fato ocorre. Outra alternativa corresponde ao

uso de regressão não-paramétrica ou semi-paramétrica, abordagens muitas vezes superiores por sua flexibilidade, em detrimento às estruturas estritamente paramétricas (FLORENCIO, 2010). Dentre tais modelos, destacam-se os Modelos Aditivos Generalizados (*Generalized Additive Models* - GAM) introduzido por HASTIE T. ; TIBSHIRANI (1990) que configuram uma extensão dos GLM com um preditor linear incluindo a soma de funções de suavização não-paramétrica das covariáveis, configurando a possibilidade de os próprios dados guiarem sua relação com o preditor ( $\eta$ ), que geralmente, decorre de maneira não-linear. Os GAM permitem casualmente, solucionar questões dos resíduos assimétricos e/ou heterocedásticos.

Ao aplicar o GAM é necessário ter atenção ao realizar a análise do modelo, pois os erros padrão são inválidos na medida que só é possível serem calculados com base na própria covariável, pressupondo uma relação linear, não se importando com o efeito das funções de suavização como um todo. O ajuste de modelos semi-paramétricos tem como um dos seus aspectos não ser facilmente descrito em uma forma matemática, mas podem ser expostos de forma gráfica, que ao mesmo tempo pode nos remeter várias informações sobre a adequação do ajuste e resíduos.

## 2.2 Modelos Aditivos Generalizados para Localização, Escala e Forma

Admitindo-se o ajuste de modelos cuja distribuição da variável resposta não necessariamente segue uma distribuição pertencente à família exponencial, propôs-se a substituição por uma família de distribuição geral. Neste sentido, Rigby e Stasinopoulos (2005) propuseram a classe modelos de regressão denominada GAMLSS - Modelos Aditivos Generalizados para Localização, Escala e Forma com a vantagem de modelar os parâmetros de assimetria e curtose, possibilitando uma maior flexibilidade e uma expansão ao número de distribuições consideradas. No GAMLSS, a parte sistemática do modelo é expandida para permitir modelar não apenas a média, mas todos os parâmetros de localização, escala e forma da distribuição da variável resposta.

Um modelo GAMLSS é descrito por,

$$Y \stackrel{ind}{\sim} D(\mu, \sigma, \nu, \tau)$$

$$\begin{cases} \eta_1 = g_1(\mu) = X_1\beta_1 + s_{11}(x_{11}) + \dots + s_{1J_1}(x_{1J_1}) \\ \eta_2 = g_2(\sigma) = X_2\beta_2 + s_{21}(x_{21}) + \dots + s_{2J_2}(x_{2J_2}) \\ \eta_3 = g_3(\nu) = X_3\beta_3 + s_{31}(x_{31}) + \dots + s_{3J_3}(x_{3J_3}) \\ \eta_4 = g_4(\tau) = X_4\beta_4 + s_{41}(x_{41}) + \dots + s_{4J_4}(x_{4J_4}) \end{cases} \quad (2.2)$$

em que,  $D(\mu, \sigma, \nu, \tau)$  é uma distribuição de quatro parâmetros,  $\mu$  normalmente é um parâmetro de localização,  $\sigma$  é constantemente, um parâmetro de escala,  $\nu$  e  $\tau$  são os parâmetros de forma da distribuição, comumente associados à assimetria e curtose, respectivamente,  $\eta_i$  representa os preditores lineares de cada parâmetro.  $X_1, X_2, X_3$  e  $X_4$  são as matrizes

que podem ou não coincidir, ou melhor, o preditor de cada parâmetro da distribuição que recebe diferentes variáveis explicativas (RIGBY; STASINOPOULOS, 2005).

As variáveis explicativas conseguem atingir os parâmetros da distribuição de diversas maneiras. Os GAMLSS permitem ajustes por meio de funções paramétricas lineares ou não lineares, e funções não-paramétricas. O pacote `gamlss` permite suavizar termos aditivos e, unido a outros pacotes, possibilita de modo simultâneo, o ajuste de árvores de decisão, efeitos aleatórios, redes neurais e suavizações multidimensionais (STASINOPOULOS et al., 2017).

A configuração da distribuição  $D(\mu, \sigma, \nu, \tau)$  (2.2) é geral e somente implica que a distribuição deve estar na configuração paramétrica. O pacote `gamlss` (STASINOPOULOS; RIGBY et al., 2007) implementado exclusivamente no *software* R, apresenta cerca de 200 distribuições discretas, contínuas e mistas implementadas na família GAMLSS, integrando algumas distribuições exclusivamente assimétricas, platicúrticas ou leptocúrticas.

As distribuições implementadas neste pacote permitem ser truncadas à direita, à esquerda ou em ambos os lados; censuradas em qualquer intervalo de resposta; misturadas criando distribuições de mistura; inflacionadas em zero e/ou um. Em relação às distribuições contínuas, é possível serem discretizadas, possibilitando serem criadas para modelar variáveis de resposta discreta. Também, as distribuições contínuas no intervalo  $(-\infty, \infty)$  podem ser transformadas em distribuições com intervalos de  $(0, \infty)$  e  $(0, 1)$  (RIGBY et al., 2017).

## 2.2.1 Tipos de distribuições dentro da família GAMLSS

Dentre as mais de 200 distribuições existentes na classe GAMLSS e implementadas no pacote `gamlss` do R. O tipo de distribuição a ser utilizada depende da variável resposta que está sendo modelada. Dentro da família GAMLSS há três tipos distintos de família de distribuições: as contínuas, as discretas e a mistura de distribuições. A tabela com os nomes das distribuições, espaço paramétrico e domínio das funções e respectivas funções de ligação na implementação do R, podem ser vistos em Stasinopoulos et al. (2017), sendo apresentada aqui neste texto apenas a distribuição *skew t* tipo 3 - ST3.

### 2.2.1.1 Distribuição *skew-t* tipo 3

A distribuição *skew-t* tipo 3 (ST3) foi obtida a partir da distribuição *t-student*, pelo método de (FERNANDEZ; STEEL, 1998), que consiste em uma transformação de uma distribuição simétrica em assimétrica, deste modo, a função de distribuição (ST3) é

dada por:

$$f_Z(z; \mu, \sigma, \tau, \nu) = \begin{cases} \kappa \sigma^{-1} \times \left[ \left( 1 + \frac{z^2}{\nu} \right) \frac{1}{\tau^2} \right]^{-\frac{(\nu+1)}{2}}, & z \geq 0 \\ \kappa \sigma^{-1} \times \left[ \left( 1 + \frac{z^2}{\nu} \right) \tau^2 \right]^{-\frac{(\nu+1)}{2}}, & z < 0 \end{cases}$$

em que,

$$z = \frac{x-\mu}{\sigma}, \quad \kappa = \frac{2}{\tau + \frac{1}{\tau}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) (\pi\nu)^{\frac{1}{2}}},$$

sendo que  $\mu \in R$  é o parâmetro de localização,  $\sigma > 0$  é o parâmetro de escala,  $\tau > 0$  é o parâmetro de forma e  $\nu > 0$  são os graus de liberdade.

A distribuição ST3 é uma das cinco versões assimétricas para a distribuição t, definida com quatro parâmetros, tendo suporte no conjunto dos reais. Para a notação, adotar-se-á  $Z \sim ST3(\mu, \sigma^2, \tau, \nu)$ . Na Figura (1), tem-se algumas formas de assimetria para a distribuição. É possível observar o comportamento das três distribuições citadas e a distribuição ST3 na Figura 2.

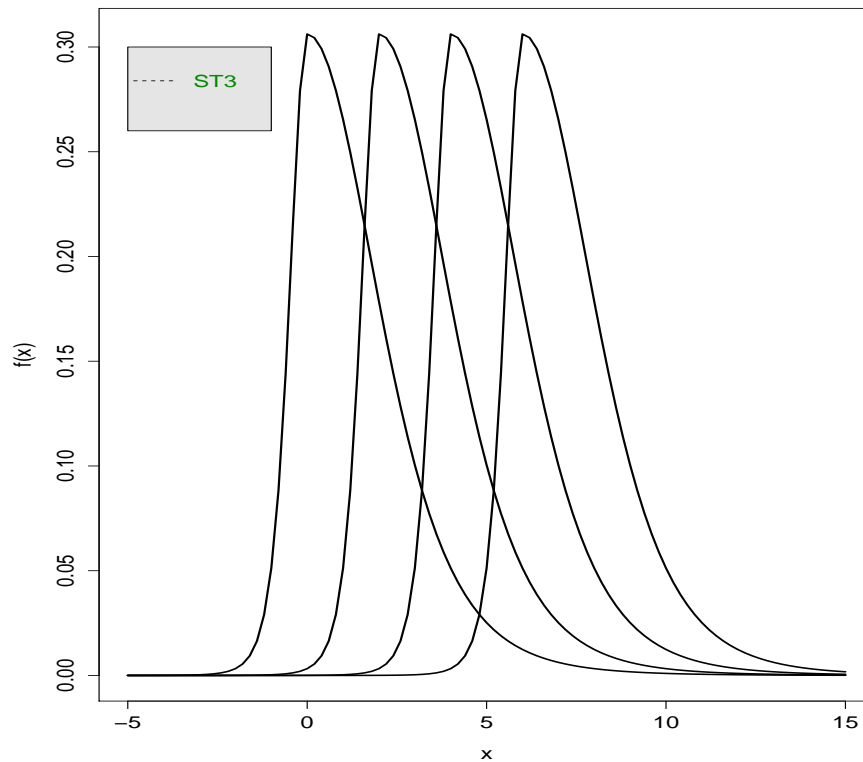


Figura 1 – Densidade da distribuição ST3, para valores diferentes de  $\mu$ , fixando  $\sigma = 1$ ,  $\tau = 2$  e  $\nu = 6$  (Fonte: Eloy, 2016.)



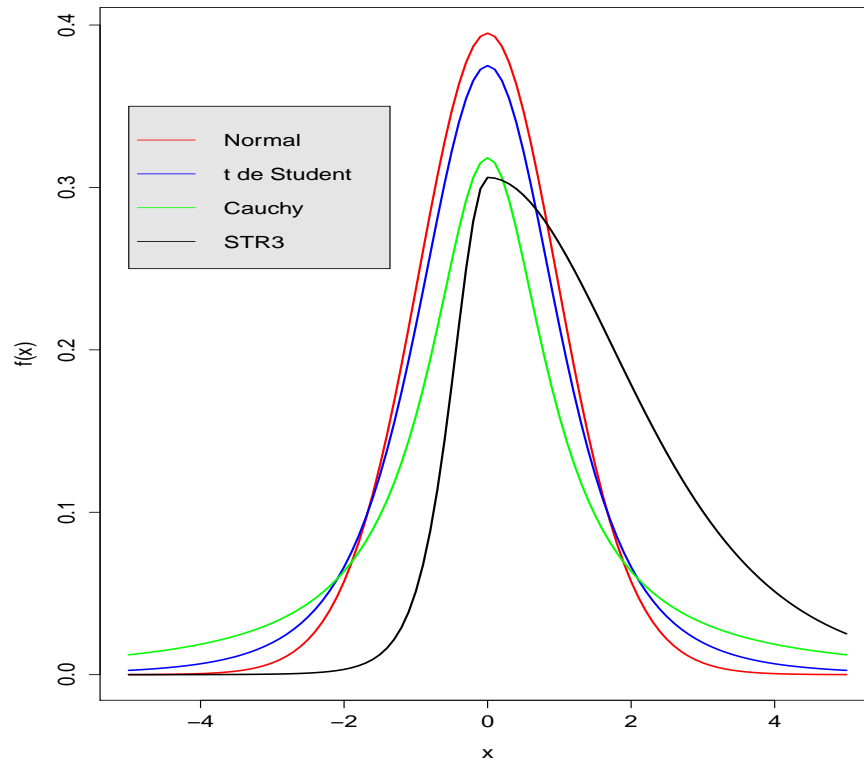


Figura 2 – Densidades Normal, t de Student, Cauchy e ST3, para diferentes valores de  $\nu$ , fixando  $\mu = 0$ ,  $\sigma = 1$  e  $\tau = 1$  (Fonte: Eloy, 2016).

Seja  $y_{11}, \dots, y_{IJ}$  uma amostra de tamanho  $n$  da distribuição *ST3*. Então, o logaritmo da função verossimilhança para o vetor de parâmetros  $\theta = (\mu, \sigma, \tau, \nu)^T$ , sob um modelo de regressão linear, em que:

$$l(\theta) = \sum_{i=1}^I \sum_{j=1}^J \log \left\{ k\sigma^{-1} \times \left[ \left( 1 + \frac{z_{ij}^2}{\nu} \right) \frac{1}{\tau^2} \right]^{-\frac{(\nu+1)}{2}} \right\} \quad (2.3)$$

onde,  $z_{ij} = (y_{ij} - (\beta_0 + \beta_1 x)) / \sigma$ .

As estimativas de máxima verossimilhança  $\hat{\theta}$  do vetor de parâmetros podem ser obtidas maximizando a log-verossimilhança. Para obter as componentes do vetor escore  $U(\theta)$  do modelo *ST3* derivou-se a função log-verossimilhança (2.3) em relação a  $\nu$ ,  $\mu$ ,  $\tau$  e

$\sigma$ . Dessa forma, as componentes do vetor escore  $U(\theta)$  são dadas por (ELOY, 2016):

$$\begin{aligned}
U_\nu(\theta) &= \sum_{i=1}^I \sum_{j=1}^J \left\{ \frac{\psi\left(\frac{\nu+1}{2}\right) \sqrt{\pi\nu}\tau^2 + \Gamma\left(\frac{\nu}{2}\right) \left[ \psi\left(\frac{\nu+1}{2}\right) \pi\nu - \psi\left(\frac{\nu}{2}\right) \pi\nu - \pi \right]}{2\sqrt{\pi\nu}} \right\} \\
&\quad + \sum_{i=1}^I \sum_{j=1}^J \left\{ -\frac{1}{2} \ln \left[ \frac{\nu\sigma^2 + (y_{ij} - \mu)^2}{\nu\sigma^2\tau^2} \right] + \frac{(\nu+1) + (y_{ij} - \mu)^2}{2\nu[\nu\sigma^2 + (y_{ij} - \mu)^2]} \right\}, \\
U_\mu(\theta) &= \sum_{i=1}^I \sum_{j=1}^J \left\{ \frac{\left(\frac{\nu+1}{2}\right) (-2y_{ij} + 2\mu)}{\nu\sigma^2 + (y_{ij} - \mu)^2} \right\}, \\
U_\tau(\theta) &= \sum_{j=1}^J \left\{ \frac{-\tau^2 + \Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu} - (\nu+1) \left[ \tau^2 + \Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu} \right]}{\tau \left( \tau^2 + \Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu} \right)} \right\}, \\
U_\sigma(\theta) &= \sum_{i=1}^I \sum_{j=1}^J \left\{ \frac{1}{\sigma} - \frac{(\nu+1)(y_{ij} - \mu)^2}{\sigma[\nu\sigma^2 + (y_{ij} - \mu)^2]} \right\}.
\end{aligned}$$

Igualando essas equações a zero e resolvendo-as, simultaneamente, obtêm-se as estimativas de máxima verossimilhança dos parâmetros utilizando métodos numéricos.

## 2.2.2 Seleção dos modelos

A seleção de modelos em GAMLSS inclui a seleção da melhor distribuição para a variável resposta, dos preditores adequados para os parâmetros da distribuição selecionada, das funções de ligação e dos hiperparâmetros.

O critério de Akaike generalizado (GAIC) leva em consideração o número de parâmetros e de graus de liberdade utilizados no modelo para penalizar os modelos mais complexos e evitar sobreajustes aos dados em amostras de grandes tamanhos (PAIVA; FREIRE; CECATTI, 2008). Ele é utilizado no mesmo contexto do critério de informação de Akaike (AIC) (AKAIKE, 1998). O GAIC é definido por (VOUDOURIS et al., 2012), como:

$$GAIC(k) = -2l(\hat{\theta}) + (k \times gl),$$

em que  $l$  é o logaritmo da função verossimilhança e  $gl$  são os graus efetivos de liberdade do modelo ajustado,  $k$  é constante e é a penalidade para cada grau de liberdade utilizado,  $\theta$  refere-se aos parâmetros. Refere-se a  $-2l(\hat{\theta})$ , como desvio global, pois o  $GAIC(k)$  é a estatística obtida pela adição do desvio global.

O desvio global (GDEV), importante medida para a seleção de modelos em GAMLSS, define-se como,

$$GDEV = -2l(\hat{\theta}),$$

em que,  $l(\hat{\theta})$  é o logaritmo da função de verossimilhança ajustada.

O modelo com o menor valor do critério  $GAIC(k)$ , para algum  $k$  escolhido é então selecionado e considerado melhor ajuste, pois o  $GAIC(k)$  penaliza modelos com muitos parâmetros. O critério (AIC) (AKAIKE, 1998) e o critério de informação bayesiano (BIC) (SCHWARZ et al., 1978) são casos especiais do  $GAIC(k)$ , um caso se  $k = 2$  recai no (AIC) e um outro caso se  $k = \ln(n)$  recai no (BIC) (VOUDOURIS et al., 2012).

### 2.2.3 Análise de resíduos

Os resíduos quantílicos ( $\widehat{r}q_{ij}$ ) (DUNN; SMYTH, 1996) podem ser obtidos para qualquer distribuição da variável resposta, e fornecem uma maneira comum de verificar a adequação de um modelo (STASINOPOULOS et al., 2017).

Este método gera resíduos para uma distribuição contínua por meio dos desvios e dos resíduos quantis, respectivamente. O propósito de Dunn e Smyth (1996) era utilizar a  $i$ -ésima observação do resíduo quantílico como uma medida de discrepância entre  $y_{ij}$  e  $\hat{\mu}_{ij}$ . Dessa forma, o resíduo quantílico ( $\widehat{r}q_{ij}$ ) para o modelo (3.1) pode ser definido por:

$$\widehat{r}q_{ij} = \Phi^{-1} \{ \Phi_{ST3}(\hat{y}, \mu = \hat{\mu}, \sigma = \hat{\sigma}, \nu = \hat{\nu}, \tau = \hat{\tau}) \}, \quad (2.4)$$

em que  $\Phi(\cdot)^{-1}$  é a função quantílica da distribuição normal padrão e  $\Phi_{ST3}(\cdot) = Q_{ST3}(\cdot)$  é o quantil da distribuição t Student tipo 3.

O modelo *skew t ST3* estará bem ajustado, se os resíduos ( $\widehat{r}q_{ij}$ ) apresentarem um comportamento aleatório em torno de zero. Uma técnica importante aplicada para analisar os resíduos de um GAMLSS são os gráficos de minhoca (*worm plot*) o estágio de diagnóstico envolve o uso deste gráfico bem específico, que permite a detecção de inadequações no modelo globalmente ou dentro de um intervalo de uma variável explicativa (BUUREN; FREDRIKS, 2001; BUUREN, 2007). O *worm plot* funciona como o gráfico normal *quantil-quantil* (Q-Q) sem tendência a fim de evidenciar alterações locais no domínio de uma dada covariável.

Os gráficos *worm plots* permitem identificar regiões em que o modelo não é bem ajustado aos dados, seu eixo vertical descreve para cada observação a diferença entre a sua localização nas distribuições teórica e empírica. Os pontos observados em conjunto, representam uma curva que se assemelha a uma minhoca. Esta forma gráfica aponta como os dados se distanciam da distribuição assumida e, quando tomadas em conjunto, sugere modificações úteis no modelo, de tal forma que, se os pontos estão situados no interior da banda de confiança de 95% (entre as duas curvas elípticas), o ajuste do modelo é considerado satisfatório.

### 3 Materiais e Métodos

O experimento foi conduzido na UNESP/Campus de Ilha Solteira, em condições de cultivo protegido, com clones do híbrido VR3 *Eucalyptus grandis* × *Eucalyptus urophylla*, denominado *E. urograndis*, no período de fevereiro a junho de 2014. O solo utilizado foi coletado na camada de 0,0 a 0,20 m de profundidade, sob plantio de eucalipto, localizado no município de Três Lagoas-MS (Latitude: 20° 59'S e Longitude 51° 48' O). O solo foi peneirado (4mm) e acondicionado em sacos para mudas com 0,40 m de altura (8 kg de solo). Calcário e NPK (Nitrogênio, Fósforo e Potássio) foram incorporados, em acordo com a recomendação em uso na região para a cultura do eucalipto (1,5 mg/ha de calcário, 0,5 mg/ha NPK - 12-20-16). As mudas de eucalipto (*Eucalyptus*), utilizadas como indicadoras, foram doadas pela FIBRIA Ltda. Foram realizadas avaliações de altura 0, 2, 4, 6, 12, 24, 36, 48, 60 e 72 meses após o plantio.

Neste estudo, foi considerado um modelo da classe dos modelos de regressão, porém sob a suposição que as observações seguem uma distribuição da classe dos modelos GAMLSS, conforme está apresentada no pacote `gamlss` do *software* estatístico R. O modelo para descrever o crescimento tem a seguinte forma:

$$Y_{ij} = \beta_0 + \beta_1 x + \sigma \epsilon_{ij}, \quad (3.1)$$

em que  $Y_{ij}$  representa o valor observado em cada tempo  $i$ ,  $\beta_0$  é o efeito do intercepto,  $\beta_1$  é a taxa de crescimento  $i$  e  $\epsilon_{ij} \sim N(0, \sigma^2)$  representam os efeitos do fator não controlado do ensaio experimental, com  $i = 1, \dots, I$  e  $j = 1, \dots, J$ , sendo que  $I$  denota o número de observações no tempo e  $J$  o número de repetições e pode-se também atribuir  $\mu_i = \beta_0 + \beta_1 x_i$ .

Foi utilizado o *software* R para as análises (versão 3.5.1). O pacote `gamlss` foi utilizado para as análises e a função `fitDist` utilizada para escolher inicialmente as distribuições probabilísticas candidatas. O critério de GAIC foi aplicado para auxiliar na escolha do melhor modelo, com  $k = 1$ , além de gráficos de diagnósticos de resíduos e *worm plot*. Também foi considerado o teste de Shapiro-Wilks para os resíduos dos modelos GAMLSS bem como estatísticas dos quatro primeiros momentos da distribuição em estudo.

## 4 Resultados e Discussões

É possível observar por meio da Figura 3, a dispersão ao longo do tempo para os valores de altura (cm), no qual percebe-se que o crescimento tem uma fase de maior aceleração dos 10 à 40 meses, chegando a uma fase de estabilização para o corte comercial por volta dos 60 meses, que é o período no qual a colheita é geralmente determinada, devido este ser considerado o momento em que a floresta fornece o produto em qualidade e quantidade apropriada para haver um retorno econômico. Esta figura levanta também a possibilidade de se pensar em modelar o relacionamento de tempo e altura por meio de um modelo de regressão paramétrico ordinário. Os ajustes do modelo em relação aos dados quer mostrar a relação existente entre a altura das plantas de eucalipto ao longo dos meses e o gráfico de dispersão sugere uma curva de crescimento normal. Porém houve a necessidade de utilizar modelos mais precisos e que conseguisse explicar o desempenho e comportamento dos dados.

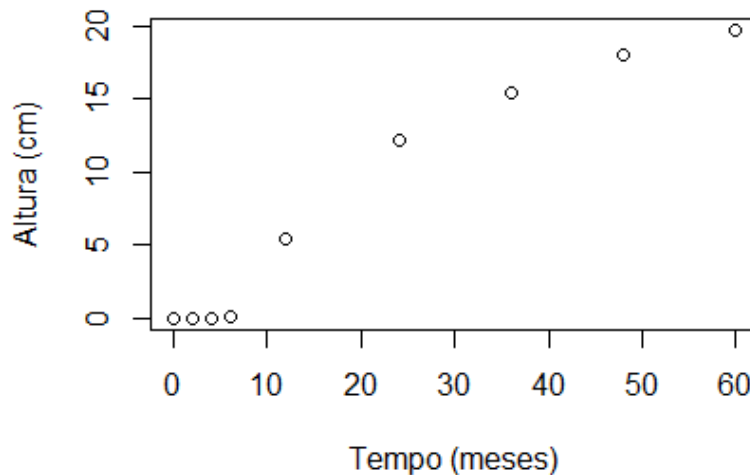


Figura 3 – Gráfico de dispersão de Tempo *versus* Altura de Eucaliptos.

O ajuste do modelo de regressão linear sob suposição de distribuição normal para os resíduos, apesar de não rejeitar aparentemente a suposição de normalidade, no gráfico dos resíduos *versus* os valores ajustados a uma aparente falta de algum termo a ser considerado no modelo (Figura 4). E mesmo ajustando um modelo polinomial este gráfico não apresenta melhoras bem como o comportamento normal dos resíduos não fica aparente, além da visualização de um possível ponto influente. Neste sentido, os modelos GAMLSS podem ser utilizados para situações em que se têm comportamentos assimétricos que a modelagem tradicional não consegue corrigir (STASINOPOULOS; RIGBY et al., 2007). Desta forma,

foi ajustado uma variedade de modelos probabilísticos aos dados afim de escolher o que melhor caracterizava o fenômeno (Tabela 1).

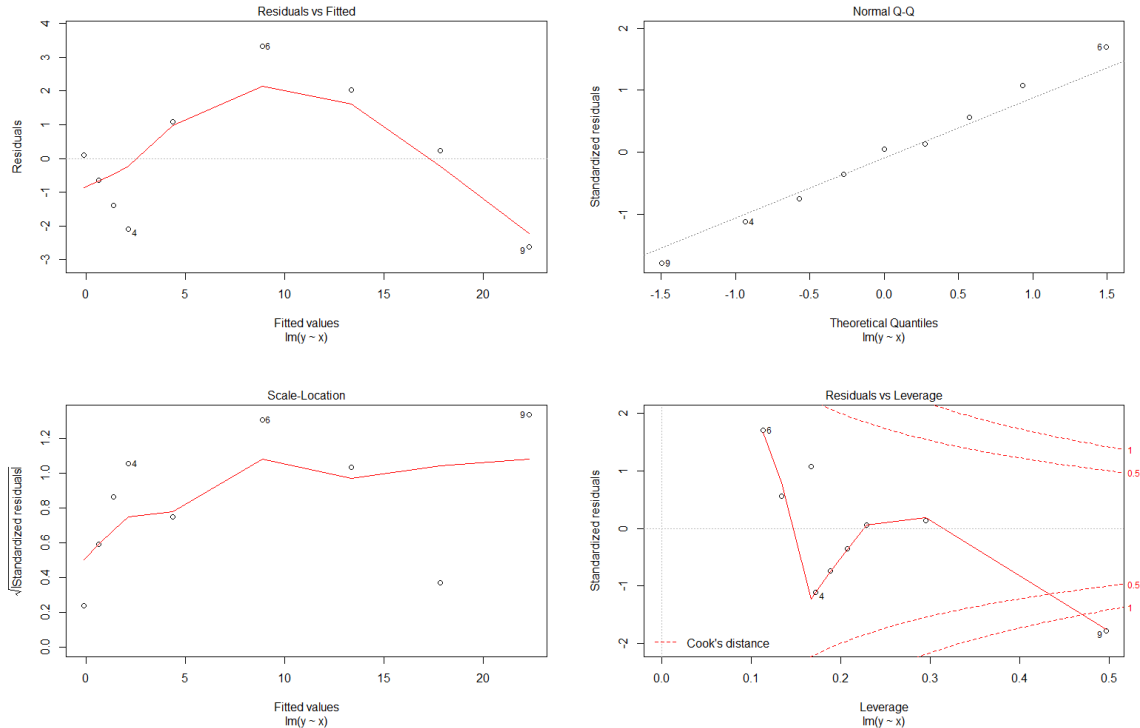


Figura 4 – Gráficos de resíduos para a altura de Eucaliptos.

Foram ajustadas 60 distribuições contínuas, das quais as 30 melhores são apresentadas na Tabela 1. As distribuições são classificadas para os dados via o menor valor critério de Informação de Akaike Generalizado - GAIC. As siglas relacionadas as distribuições citadas na tabela 1 podem ser melhor detalhadas quanto ao seu significado e as suas propriedade estatísticas no livro (STASINOPOULOS et al., 2017).

Tabela 1 – Critério de informação de Akaike Generalizado dos modelos probabilísticos ajustados via pacote gamlss

Distrib.	GAIC	Distrib.	GAIC	Distrib.	GAIC	Distrib.	GAIC
SHASH	29,79366	LOGNO	40,82479	PARETO2o	42,52187	SN2	62,56368
GG	35,34865	LOGNO2	40,82479	GP	42,52187	ST3	64,56368
GIG	38,01075	BCCGo	41,12412	BCTo	43,12412	ST2	64,56370
GA	38,02769	BCCG	41,12412	JSU	48,39924	ST1	64,56402
WEI	39,86231	IGAMMA	41,35272	IG	49,21181	SST	64,66178
WEI2	39,86231	GB2	42,27532	EXP	57,15536	RG	66,10766
WEI3	39,86231	PARETO2	42,52187	SEP1	61,48246	NO	66,91545

A distribuição *Sinh-Arcosinh* - SHASH foi a melhor distribuição via GAIC inicialmente ajustada, porém ao se incluir o efeito linear do tempo (Figura 7), tanto a distribuição SHASH como as demais da lista não apresentam ajuste satisfatório quando comparados a distribuição normal, conforme Tabela 2. Devido a proximidade entre as distribuições

também foi feito o ajuste com a distribuição *Sinh-Arcosinh original* - SHASHo, porém esta também não apresentou bons resultados (Figura 6). Neste sentido, a distribuição *Skew t* tipo 3 (ST3), apresentou melhores valores das estatísticas residuais dos quantis bem como melhor gráfico de resíduos e *worm plot* (Figura 8).

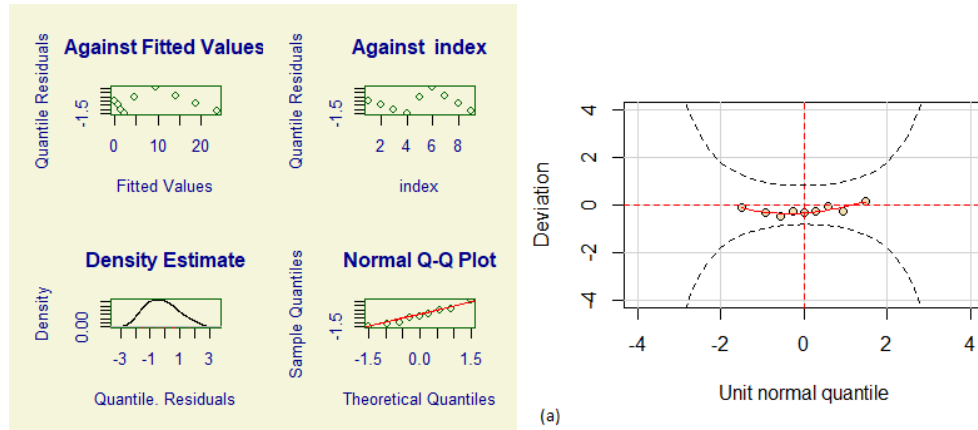


Figura 5 – Worm plot e gráficos residuais das distribuição Normal (média e dispersão).

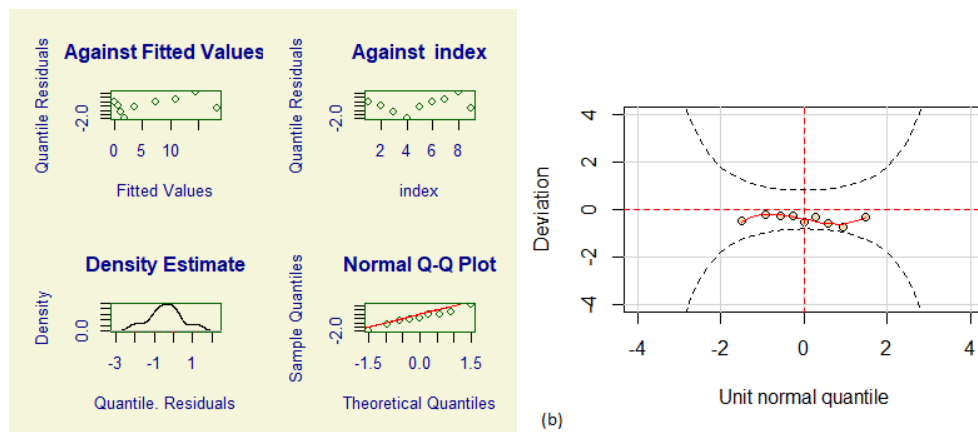


Figura 6 – Worm plot e gráficos residuais das distribuição SHASHo.

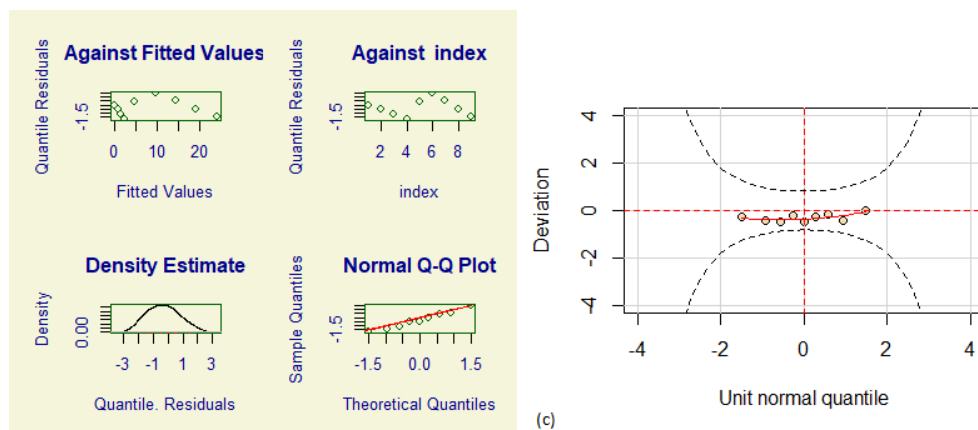


Figura 7 – Worm plot e gráficos residuais das distribuição SHASH.

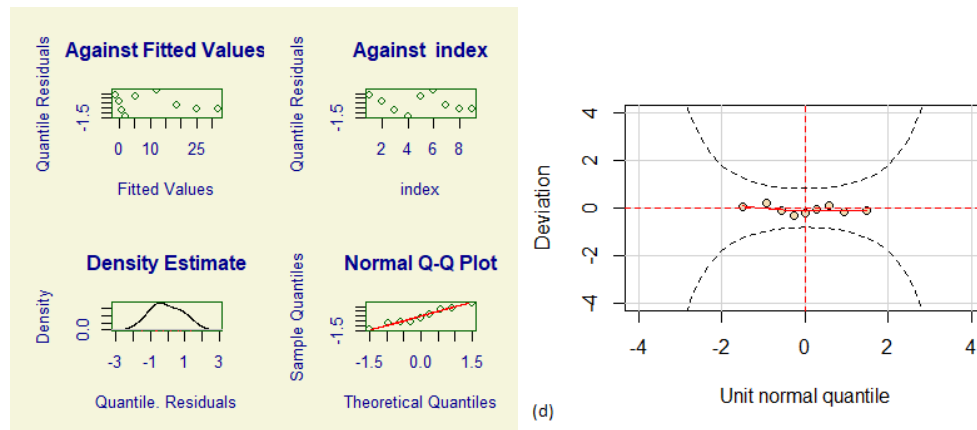


Figura 8 – Worm plot e gráficos residuais das distribuição ST3.

Após a inclusão dos efeitos fixos nos modelos probabilísticos, foram observadas as distribuições Normal acrescida da modelagem do parâmetro  $\sigma$  (Normal+ $\sigma$ ), a Normal, a SHASHo, a SHASH e a ST3, com diferentes configurações paramétricas (Tabela 2), onde obteve-se um melhor valor de GAIC para a distribuição SHASHo, porém os valores dos quantis residuais e da análise gráfica residual (Figura 6), não foram satisfatórios, levando a escolha da distribuição ST3, que obteve menor valor de GAIC que a distribuição Normal e melhores valores dos quantis residuais que as demais distribuições avaliadas.

Tabela 2 – Critério de informação de Akaike Generalizado dos modelos probabilísticos ajustados via pacote gamlss, nas diferentes configurações paramétricas.

Distribuição	GAIC	$\mu$	$\sigma$	$\tau$	$\nu$
SHASHo	38,01075	x	x	x	x
SHASH	38,02769	x	x	x	x
Normal+ $\sigma$	38,73810	x	x		
ST3	39,16977	x	x	x	x
Normal	39,37939	x	x		

× intercepto; x Efeitos de covariáveis (Linear ( $\beta_1$ ) ou Quadrática ( $\beta_2$ )).

Pela Tabela 3, os resultados referentes a média nas cinco distribuições consideradas (Normal, Normal +  $\sigma$ , SHASHo, SHASH e ST3) foram próximos de zero, com maior diferença para o modelo SHASHo e SHASH. No que se refere a variância a SHASHo e a ST3 foram as distribuições que tiveram pior desempenho com uma diferença de 0,20 unidades. Com relação ao Coeficiente de Assimetria, a Normal+ $\sigma$  tem maior assimetria seguida da Normal e SHASH, com a ST3 e SHASHo com melhor desempenho neste quesito. Para a curtose apenas a SHASHo obteve um desempenho satisfatório e nenhuma diferença pronunciada pode ser relacionada ao Coeficiente de Correlação de Filliben.

Percebe-se que as estatísticas da ST3 foi uma das melhores com um dos menores valores de assimetria. Ao se analisar conjuntamente com os gráficos da Figura 5, podemos



Tabela 3 – Estatísticas dos Quantis Residuais dos modelos propostos

Estatísticas	Modelos				
	Normal	Normal + $\sigma$	SHASHo	SHASH	ST3
Média	$-0,2337058 \times 10^{-18}$	-0,214825	-0,4105777	-0,3067954	-0,07357571
Variância	1,125	1,073074	0,8098819	1,058978	0,8001895
Coefficiente de Assimetria	0,240712	0,3458527	0,00167738	0,2057842	0,1143107
Coefficiente de Curtose	1,660571	1,786863	2,265013	1,745959	1,629309
Coefficiente de Correlação de Filliben	0,9918957	0,9856064	0,9819349	0,9897563	0,9849912

observar que a mesma foi a distribuição que melhor se adequou aos dados.

Quanto aos efeitos das estimativas dos parâmetros  $\mu$ ,  $\sigma$ ,  $\tau$  e  $\nu$ , percebe-se que o efeito linear de  $\beta_1$  sobre o crescimento foi de 0,54 centímetros a cada unidade de tempo, resultado diferente do encontrado se fosse utilizado o modelo de regressão sob distribuição residual normal, em que, a cada unidade de tempo haveria um acréscimo médio de 0,37 centímetros. Os demais parâmetros não são interpretados neste trabalho, porém tem-se que, a exceção de  $\tau$ , os parâmetros foram significativos, indicando a necessidade de se considerar estes parâmetros para modelar a dispersão, assimetria e curtose.

Tabela 4 – Estatísticas do modelo gamlss ajustado sob a distribuição *Skew t* tipo 3.

Função de Ligação de $\mu$ : Coeficientes de $\mu$				
	Estimativa	Erro Padrão	valor t	Pr(> t )
$\beta_0$	-1,13	0,86	-1,31	0,26
$\beta_1$	0,54	0,05	11,24	0,00
Função de Ligação de $\sigma$ : log. Coeficientes de $\sigma$				
	Estimativa	Erro Padrão	valor t	Pr(> t )
$\beta_2$	$-5,28 \times 10^{-6}$	$-1,31 \times 10^{-6}$	-401,30	0,00
Função de Ligação de $\nu$ : log. Coeficientes de $\nu$				
	Estimativa	Erro Padrão	valor t	Pr(> t )
$\beta_1$	-0,04	0,01	-4,97	0,01
Função de Ligação de $\tau$ : log. Coeficientes de $\tau$				
	Estimativa	Erro Padrão	valor t	Pr(> t )
$\beta_1$	0,56	0,98	0,57	0,60

Para curvas de crescimento e modelos de crescimento de formal geral, o GAMLSS vem sendo pesquisado e utilizado com frequência. Curvas de referência ligadas ao comportamento do crescimento de fetos, para valores de parâmetros antropométricos na população, foram utilizadas por Paiva, Freire e Cecatti (2008), os resultados encontrados indicam que a classe de modelos GAMLSS é o estado da arte para pesquisa de curvas de referência. Fumes et al. (2017) estudaram uma classe de modelos para ajustar curvas de crescimento para diâmetro e altura de *Eucalyptus grandis*  $\times$  *Eucalyptus urophylla* em um estudo realizado em casa de vegetação, no Departamento de Recursos Naturais da Escola de Agricultura, Botucatu, São Paulo, Brasil. As variáveis de crescimento referentes à altura e ao diâmetro foram avaliadas, sendo medido cinco e quatro vezes, respectivamente. A metodologia foi realizada em um modelo longitudinal misto usando uma nova abordagem baseada na distribuição Box-Cox Normal (BCN) e as comparações com este modelo foram

feitas assumindo normalidade dos dados. Os resultados revelaram que o modelo misto do BCN forneceu resultados semelhantes ao modelo padrão para estimar as curvas de crescimento; no entanto, o modelo BCN foi o melhor resultado segundo o critério de Akaike, considerando a pequena assimetria no conjunto de dados, tendo sido o escolhido, de forma que os autores usaram critérios semelhantes aos utilizados neste trabalho. Neste sentido, Thomas et al. (2018) afirmam que o GAMLSS permite que diferentes tipos de efeitos aleatórios sejam incluídos na formulação do modelo, os autores realizaram uma análise de um experimento com dados de crescimento de plantas usando a metodologia GAMLSS com efeito aleatório e compararam a uma abordagem de modelo linear misto, obtendo melhores resultados para o modelo GAMLSS sob a distribuição Gama Generalizada (GG).

## 5 Considerações Finais

O uso dos modelos GAMLSS foi adequado para modelar dados de crescimento de eucaliptos ao longo do tempo, demonstrando a versatilidade desta metodologia nas mais diversas áreas de aplicação da estatística. Os modelos da classe GAMLSS obtiveram melhores resultados que a distribuição normal, tendo pelo menos três modelos competitivos para modelar o crescimento (SHASHo, SHASH, ST3). O modelo sob distribuição ST3 foi o melhor, considerando todos os parâmetros modelados com efeitos da variável independente (tempo).

## Referências

- AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: *Selected papers of hirotugu akaike*. [S.l.]: Springer, 1998. p. 199–213. Citado 2 vezes nas páginas 16 e 17.
- BUUREN, S. V. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 219–242, 2007. Citado na página 17.
- BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001. Citado na página 17.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 17.
- ELOY, R. A. O. Distribuição assimétrica t-student tipo 3: uma aplicação a delineamentos inteiramente casualizados. 2016. Citado na página 16.
- FERNANDEZ, C.; STEEL, M. F. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 93, n. 441, p. 359–371, 1998. Citado na página 13.
- FLORENCIO, L. Engenharia de avaliações com base em modelos gamlss (dissertação de mestrado). *Universidade Federal de Pernambuco, Recife*, 2010. Citado na página 12.
- FUMES, G. et al. Growth curves for diameter and height using mixed models: An application in eucalyptus seedling. *Open Journal of Forestry*, Scientific Research Publishing, v. 7, n. 04, p. 403, 2017. Citado na página 23.
- GONÇALVES, J. C. et al. Análise económica da rotação florestal de povoamentos de eucalipto utilizando a simulação de monte carlo. *Ciência Florestal*, Universidade Federal de Santa Maria, v. 27, n. 4, p. 1339–1347, 2017. Citado na página 10.
- HASTIE T. ; TIBSHIRANI, R. Generalized additive models. [s.l.]. *Wiley Online Library*, 1990. Citado 2 vezes nas páginas 10 e 12.
- LOPES, B. G. *Classificação da variação, tamanho ótimo de parcela e curva de crescimento para experimentos com eucalipto*. Dissertação (Mestrado) — Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira, 2019. Citado na página 10.
- MAIA, M. C. C. et al. Seleção simultânea para produção, adaptabilidade e estabilidade genotípicas em clones de cajueiro, via modelos mistos. *Pesquisa Agropecuária Tropical*, Escola de Agronomia e Engenharia de Alimentos, v. 39, n. 1, p. 43–50, 2009. Citado na página 10.
- NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado 2 vezes nas páginas 10 e 11.

- PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, v. 12, n. 3, p. 289–310, 2008. Citado 2 vezes nas páginas 16 e 23.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado na página 11.
- RIGBY, R. A. et al. Distributions for modelling location, scale, and shape: Using gamlss in r. *URL [www.gamlss.org](http://www.gamlss.org)*. (last accessed 5 March 2018), 2017. Citado na página 13.
- RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado 3 vezes nas páginas 10, 12 e 13.
- SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 17.
- STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007. Citado 2 vezes nas páginas 13 e 19.
- STASINOPOULOS, M. D. et al. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2017. Citado 3 vezes nas páginas 13, 17 e 20.
- THOMAS, G. et al. Analysis of a longitudinal multilevel experiment using gamlss. *arXiv preprint arXiv:1810.03085*, 2018. Citado na página 24.
- VOUDOURIS, V. et al. Modelling skewness and kurtosis with the bcpe density in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012. Citado 2 vezes nas páginas 16 e 17.