SCIENTIA FORESTALIS
n. 69, p.38-50, dez. 2005

Melhoria da qualidade de ajuste de modelos biométricos florestais pelo emprego da teoria dos modelos não lineares generalizados

Improving fitting quality of forest biometric models by applying the generalized nonlinear model theory

Natalino Calegario
Cristina Lelis Leal Calegario
Romualdo Maestri
Richard Daniels

RESUMO: A presença de heterogeneidade da variância e da autocorrelação é comum em bases de dados biométricas longitudinais. Na modelagem de tais bases de dados, fundamentada nos modelos lineares e não lineares ordinários, as citadas características violam as pressuposições básicas da homogeneidade da variância e da independência entre as observações. No presente estudo foram utilizadas técnicas, baseadas nos modelos não lineares generalizados para a estimativa do crescimento em altura para árvores individuais clonais de *Eucalyptus*. Anteriormente à modelagem generalizada, foi feito um estudo para a seleção de um melhor modelo não linear para a descrição do crescimento. Entre 4 modelos, o Logístico foi o que apresentou melhor performance para a base de dados em questão. Após o ajuste dos parâmetros do modelo Logístico, foram incluídas variáveis relativas à árvore individual e ao povoamento, melhorando-se com isto a precisão do modelo. Com a modelagem da heterogeneidade da variância, a distribuição dos resíduos foi sensivelmente melhorada e o valor do logaritmo da máxima verossimilhança foi elevado de -4436 para -4404, sendo que tal diferença foi altamente significativa. Com a modelagem da autocorrelação, a distribuição dos resíduos foi melhorada e o logaritmo da máxima verossimilhança, que era de -4404, passou a -869. Tal diferença foi ainda mais significativa comparada àquela obtida com apenas a modelagem da heterocedasticidade. Além da distribuição dos resíduos, a modelagem da heterogeneidade da variância e da autocorrelação mudou os valores dos parâmetros ajustados para o modelo homocedástico. Tal mudança afeta de forma determinante o uso do modelo para predição e projeção com fins de planejamento florestal.

PALAVRAS-CHAVE: Modelos não lineares generalizados, Crescimento de *Eucalyptus*, Heterogeneidade de variância e autocorrelação

ABSTRACT: The heteroscedasticity and autocorrelation are common problems in longitudinal-biometric databases. In modeling such databases using ordinary nonlinear model, the assumptions of homocedasticity and independence are violated. In this study were used the generalized nonlinear model theory to model the individual tree height growth from *Eucalyptus* clonal stands. Firstly, we selected the Logistic model from 4 nonlinear models tested, based on the model performance to represent the database. In order to improve the model performance, we decomposed the model parameters and included some covariates associated to them. This procedure generated estimates that are more precise. The heteroscedasticity was modeled by power-variance modeling, increasing significantly the logLikelihood values from -4436 to -4404. After the autocorrelation modeling the residual distribution was more concentrated around the zero axis and the logLikelihood value changed from -4004 to -869. This increasing was also highly significant. Furthermore, the heteroscedasticity and autocorrelation modeling changed the fitted parameter values. This changed will affect the predicting and projecting process in a forest management planning and decision making.

KEYWORDS: Generalized nonlinear models, *Eucalyptus* growth, Heteroscedasticity and autocorrelation.

INTRODUÇÃO

A heterogeneidade da variância, também conhecida como heteroscedasticidade, e a autocorrelação são características presentes em muitas bases de dados representantes de diversas áreas do conhecimento. A primeira devido à variação na dispersão de uma variável com a variação de outra. A segunda devido a medições em um indivíduo ser repetida no tempo e, ou no espaço, ou seja, dados espaço-temporais. Quando a autocorrelação está presente, indivíduo com performance superior (inferior) tende a se manter superior (inferior).

O problema na modelagem de tais bases de dados é que tanto a heteroscedasticidade quanto a autocorrelação violam as pressuposições básicas da teoria dos modelos lineares e não lineares ordinários. Tais pressuposições consideram que a variável dependente, ou resposta, em um modelo tenha distribuição normal, com observações independentes e identicamente distribuídas. Quando a heterogeneidade de variância está presente, as observações não possuem distribuição idêntica. Quando existe autocorrelação, as observações não são independentes. Dependendo da característica da base de dados, ambas as violações podem estar presentes.

Alguns autores apresentam metodologias para se remediar a presença de tais problemas na base de dados (BOX e COX, 1987; COOK e WEISBERG, 1983; DAVIDIAN e CARROLL, 1987; ATKINSON, 1987; NETER *et al*., 1996; DRAPER e SMITH, 1998). Atualmente, a forma mais precisa de se tratar com o problema é o uso do método dos mínimos quadrados não lineares generalizados (CARROLL e RUPPERT, 1988; SEBER e WILD, 1989), o qual considera no ajuste tanto a heterogeneidade da variância quanto a autocorrelação. No Brasil, estudos com uso do método dos mínimos quadrados generalizados em biometria florestal não são muito freqüentes, com exceção de Díaz e Couto (1999) que utilizaram tal metodologia para a estimativa da mortalidade em povoamentos de *Eucalyptus*.

Como exemplo da aplicação do método, será selecionado um modelo não linear para a modelagem do crescimento da altura total (relação hipsométrica). Uma das dificuldades na modelagem de tal relação, conforme comentado por Batista *et al*.(2001), é o grande número de variáveis que a influenciam, dificultando a construção de modelos genéricos com base em métodos empíricos como a regressão linear e não linear. Diante disto, técnicas de decomposição dos parâmetros do modelo não linear, com a inclusão de covariantes relacionadas com árvores individuais e com o povoamento, podem melhorar a precisão do mesmo.

Portanto, o principal objetivo do presente estudo é a aplicação do método dos mínimos quadrados não lineares generalizados na modelagem do crescimento em altura de árvores individuais, visando a melhoria da qualidade do ajuste pela inclusão de covariantes e pela modelagem da heteroscedasticidade e da autocorrelação.

MATERIAL E MÉTODOS

Dados

A base de dados utilizada no estudo é proveniente de plantios comerciais de clones de *Eucalyptus*, cultivados na região costeira Brasileira, no estado do Espírito Santo. Os dados longitudinais foram registrados para árvores individuais com idades variando de 2,5 a 5 anos, sendo que a variável alvo foi a altura total (Figura 1). Como a figura está mostrando, o crescimento em altura tende a apresentar uma tendência não linear, com árvores atingindo diferentes assíntotas e com discreto aumento da variância à medida que a idade aumenta. Por se tratar de remedições no tempo dos mesmos indivíduos, espera-se também a presença de autocorrelação dentro das árvores individuais. Como estão faltando algumas observações para idades inferiores, a base de dados pode ser considerada como não balanceada.

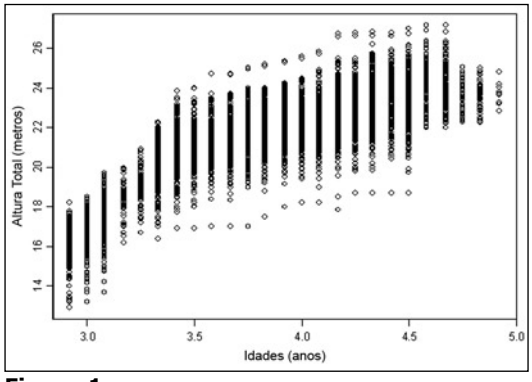

Figura 1.
Relação entre a idade e altura total para diferentes árvores de clone de *Eucalyptus.*
(Relationship between age and total height for different *Eucalyptus* clonal trees).

Modelos

Os modelos biométricos testados na estimativa da altura para a citada base são os mais comumente utilizados em tal situação, sendo que os mesmos possuem em comum a característica de representarem um crescimento em sigmóide (Tabela 1). Ou seja, as curvas geradas apresentarão crescimento monotônico até atingir um ponto de inflexão, onde a taxa máxima de crescimento é alcançada e, posterior a este ponto, a mesma declina e tende a zero na assíntota horizontal superior.

Obtenção dos valores iniciais para iteração

Além dos modelos utilizados no estudo, a Tabela 1 apresenta as formulações para a estimativa dos valores iniciais das iterações para os parâmetros dos mesmos. Para o modelo logístico, as formulações foram obtidas pelo isolamento dos parâmetros. Para os três outros modelos utilizados, as formulações foram baseadas em Ratkowsky (1983), com algumas modificações.

Uma das limitações do uso de modelos não lineares é a escolha correta dos valores iniciais dos parâmetros para o processo de iteração, os quais irão gerar as estimativas pela convergência do algoritmo baseado no método de Gauss-Newton, na maioria das vezes. Quando os valores iniciais dos parâmetros são distantes das estimativas para uma base de dados em questão, o processo de convergência do algoritmo falha e não ocorre a estimativa, ou o processo gera estimativas não confiáveis. Uma das formas de contornar o problema é o uso de expressões geradas com base na interpretação dos parâmetros do modelo e, ou, em expressões geradas pelo isolamento do parâmetro em função de outros parâmetros conhecidos e pontos específicos na base de dados.

Dos modelos citados, o logístico se apresenta como o de maior facilidade de interpretação dos parâmetros. O primeiro parâmetro representa a assíntota horizontal superior (AHS), a qual poderia ser visualizada na base de dados (Tabela 1), como o valor médio máximo da resposta (altura) no tempo (aproximadamente 24 m). Portanto, este seria um valor inicial com grande chance de convergência para o primeiro parâmetro neste modelo. Para o segundo parâmetro, ponto de inflexão, um valor razoável seria menor do que 3 anos, onde ocorre o ponto de inflexão da curva. Apesar do ponto estar fora da base de dados, qualquer valor entre 1 e 3 seria razoável. Para o terceiro parâmetro, escala, o qual é a diferença entre as idades a aproximadamente 70% da resposta média máxima e a idade no ponto de inflexão, um valor entre 0,1 e 1,0 seria uma escolha razoável.

Tabela 1.
Modelos não lineares utilizados na estimativa do crescimento em altura, com seus respectivos parâmetros iniciais para iterações, onde: AHS = Assíntota horizontal superior; H_{inf} = Valor aproximado da altura no ponto de inflexão; $Idade_{inf}$ = Valor aproximado da idade no ponto de inflexão; H_{INT}=Valor aproximado da altura no intercepto (Idade=0)
(Nonlinear models used to estimate the height growth, including the iteration start parameters where: AHS = Superior Horizontal Asymptote; H_{inf} = Approximated height value on the inflection point: H_{INT}= Approximated height value on the intercept (age=0))

		Valores Iniciais para Iteração			
Modelo	**Expressão**	Φ_1^0	Φ_2^0	Φ_3^0	Φ_4^0
Logístico	$H_i = \frac{\Phi_1}{1+\exp[(\Phi_2 - Idade_i)/\Phi_3]} + \varepsilon_i$	AHS	Ponto de Inflexão (Idade onde H=$\Phi_1^0/2$)	Idade a 0.7 Φ_1^0 subtraido de Φ_2^0	----
Richards	$H_i = \frac{\Phi_1}{1+\exp(\Phi_2 - \Phi_3 Idade)^{1/\Phi_4}} + \varepsilon_i$	AHS	$\ln[(\frac{\Phi_1^0}{H_{INT}})^{\Phi_4^0} - 1]$	$\frac{\Phi_2^0 - \ln\Phi_4^0}{Idade_{Inf}}$	$H_{Inf} = \Phi_1^0(1+\Phi_4^0)^{-1/\Phi_4^0}$
Gompertz	$H_i = \Phi_1[\exp(-\exp(\Phi_2 - \Phi_3 Idade_i)] + \varepsilon_i$	AHS	$\ln[-\ln(\frac{H}{\Phi_1^0})] = \Phi_2^0 - \Phi_3^0 Idade$	Mesmo que Φ_2^0	----
Weibull	$H_i = \Phi_1 - \Phi_2\exp[(-\Phi_3 Idade_i^{\Phi_4})] + \varepsilon_i$	AHS	$\Phi_2^0 = \Phi_1^0 - H_{INT}$	$\frac{\Phi_4^0 Idade_{Inf}^{\Phi_4^0}}{\Phi_4^0 - 1}$	$H_{Inf} = \Phi_1^0 - \Phi_2^0\exp[-\frac{\Phi_4^0 - 1}{\Phi_4^0}]$

No modelo proposto por Richards (1959), que pode ser considerado uma extensão do modelo logístico, a escolha dos valores iniciais para o início do processo iterativo não é tão direto quanto ao modelo logístico. O primeiro parâmetro também representa a assíntota horizontal superior. Após a escolha do primeiro parâmetro, encontra-se o valor aproximado da altura no ponto de inflexão (H_{Inf}) e, utilizando a expressão apresentada na Tabela 1, estima-se o valor inicial do quarto parâmetro. Em seguida, tendo-se o valor do quarto parâmetro, estima-se o valor segundo, e, na seqüência, o valor do terceiro, que é função do segundo, do quarto e da idade no ponto de inflexão (Tabela 1).

No terceiro modelo, Gompertz, o valor inicial do primeiro parâmetro, como nos dois anteriores, é obtido pela visualização do ponto de assíntota horizontal superior. Os outros dois são estimados por um modelo de regressão linear simples, onde a variável dependente, ou resposta, é função do primeiro parâmetro, dos valores da altura observada e a variável independente, ou regressor, é a idade (Tabela 1). Com este procedimento, são obtidas as estimativas dos valores iniciais dos outros dois parâmetros.

Para o modelo do tipo Weibull (1951), baseado na distribuição do mesmo nome, a seqüência de obtenção dos valores iniciais dos parâmetros é basicamente a mesma. Primeiramente obtém-se a assíntota horizontal superior, por visualização na Figura 1. Em seguida estima-se o valor inicial do segundo parâmetro como uma função do primeiro e da altura no intercepto. Na seqüência, o valor inicial do quarto parâmetro é estimado como uma função da altura no ponto de inflexão e dos primeiro e segundo parâmetros. Finalizando, o valor inicial do terceiro parâmetro é estimado em função do quarto parâmetro e da idade no ponto de inflexão (Tabela 1).

Inclusão de covariantes no modelo

Anteriormente à modelagem da heterogeneidade da variância e da autocorrelação, e baseando-se no fato de que a variação da altura total de árvores individuais não é apenas explicada pela idade, os parâmetros do modelo selecionado foram decompostos associando aos mesmos variáveis relativas à árvore individual (DAP) e ao povoamento (Clone, G e HD). A grande flexibilidade deste método está no fato de que as variáveis podem estar associadas a um parâmetro e não a outro, dependendo da sua significância.

Heterogeneidade de variância

Como pode ser visto na Figura 1, ocorre uma discreta amplitude da variabilidade em torno da curva média de crescimento à medida que a idade dos indivíduos aumenta, classificando a base de dados como heterocedástica. Baseando-se nas pressuposições básicas dos modelos estatísticos ordinários, tal característica viola a homogeneidade da variância, sendo que a mesma deve ser corrigida.

A técnica apresentada aqui na modelagem de tal violação será baseada na proposta de Davidian e Giltinan (1995), os quais apresentaram as seguintes expressões para a definição geral da função de variância:

$$var(y_i) = \sigma^2 g^2(\mu_i, z_i, \Phi), \qquad \mu_i = f(x_i, \beta) \qquad (1)$$

A variância da resposta em (1) é função de g que, por sua vez, é função da média da resposta, de fatores fixos z, que podem ser representados por parte ou todos componentes de x, e do vetor de parâmetros Φ da função de variância. Como a resposta média é função dos parâmetros da regressão β, a variância também é função destes parâmetros. A função $g(.)$ pode possuir várias formas. As mais comuns são a forma exponencial, o parâmetro como uma potência da média e a forma com dois componentes, conforme expressão (2).

$$\begin{aligned} g(\mu_i, z_i, \Phi) &= \mu_i^{\Phi} \\ g(\mu_i, z_i, \Phi) &= exp(\mu_i^{\Phi}) \\ g(\mu_i, z_i, \Phi) &= \Phi_1 + \mu_i^{\Phi_2} \end{aligned} \qquad (2)$$

O processo de estimativa da função de variância é baseado nos mínimos quadrados generalizados. Após a estimativa do parâmetro Φ e da escolha dos valores iniciais para β, um processo iterativo gera valores definitivos para os parâmetros pela minimização da função de pseudo-verossimilhança, conforme expressão (3).

$$PV(\beta^{(0)}, \sigma, \Phi) = \sum_{i=1}^{n} \left(\frac{\{y_i - f(x_i, \hat{\beta}^{(0)})\}^2}{\sigma^2 g^2\{f(x_i, \beta^{(0)}), z_i, \Phi\}} + log[\sigma^2 g^2\{f(x_i, \beta^{(0)}), z_i, \Phi\}] \right) \qquad (3)$$

Tecnicamente, a minimização de (3) significa a maximização da verossimilhança em $\beta(0)$. Para a minimização da expressão (3), por iteração, é

necessário o conhecimento de Φ. Alguns métodos de estimativa do citado parâmetro podem ser encontrados em Carroll e Ruppert (1988). Independente da variância e da sua função $g(.)$, a minimização em (3) implica em minimização das somas de quadrados dos erros ($\{y_i - f(x_i,\hat{\beta}^{(0)})\}^2$). Porém, quanto mais apropriados os valores estimados da variância e da sua função $g(.)$, menor a soma do quadrado dos erros.

Autocorrelação

Além da violação da pressuposição de homogeneidade de variância, a base de dados da Figura 1, por ser proveniente de remedições de árvores individuais no tempo, ou seja, dados longitudinais, pode ser tratada como tendo também o problema da autocorrelação. Base de dados com medidas repetidas tem este problema devido ao fato das medidas serem feitas no mesmo indivíduo em tempo diferente, gerando o problema da autocorrelação, ou seja, os dados não são independentes. Isto viola uma das pressuposições básicas dos modelos lineares e não lineares clássicos.

Também para a modelagem de tal problema serão utilizados os conceitos associados aos mínimos quadrados generalizados. Na situação onde ocorre correlação entre as observações, como no caso da Figura 1, a qual representa uma base de dados longitudinais, esta correlação pode ser representada pela seguinte expressão:

$$Corr(e) = \Gamma(\alpha) \qquad (4.1)$$

Ou em forma matricial,

$$\Gamma(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \dots & \alpha^{n-1} \\ & 1 & \alpha & \alpha^2 & \dots & \alpha^{n-2} \\ & & 1 & . & . & . \\ & & & . & . & . \\ & & & & . & . \\ & & & & & \alpha \\ & & & & & 1 \end{bmatrix} \qquad (4.2)$$

A expressão (4.2), na realidade, representa a situação de correlação autorregressiva de ordem um (AR(1)). Caso, por exemplo, tenhamos duas medidas de altura do i-ésimo indivíduo nos tempos $t_{i,1}$ e $t_{i,2}$, os componentes da expressão (4.1) poderiam ser substituídos por $\alpha^{|t_{i,1}-t_{i,2}|}$. Existem vários outros padrões de autocorrelação disponíveis na literatura.

Um outro padrão de correlação seria a combinação da autorregressiva com a média móvel - ARMA (Box *et al.*, 1994). Tal padrão pode ser generalizado pela seguinte estrutura:

$$\varepsilon_t = \sum_{i=1}^{p} \varphi_i \varepsilon_{t-i} + \sum_{j=1}^{q} \eta_i a_{t-j} + a_t \qquad (6)$$

Onde ε_t refere-se a uma observação no tempo t e a_t é o erro termo. A primeira parte da expressão refere-se ao modelo autorregressivo (AR(p)) e a segunda parte ao de movimento médio (MA(q)). Se p=0, ter-se-á uma MA(q) situação e, ao contrário, se q=0, a situação seria de AR(p). Na AR(p) parte, φ representa os parâmetros de correlação com ordem p e (t-i) representa a distância entre duas observações (lag). A tendência é que os valores de φ decresçam com o tempo, indicando que observações próximas no tempo são mais correlacionadas do que observações distantes, o que é comum em estudos de dados longitudinais desta natureza. Na parte do movimento da média (MA(q)), o modelo assume que uma observação atual é uma função linear dos erros termos (a_t), identicamente e independentemente distribuídos.

Estrutura de covariância

Quando existe a presença de heterogeneidade de variância e da autocorrelação, como é o caso particular da base de dados representada na Figura 1, as duas pressuposições associadas a estas características dos modelos lineares e não lineares são violadas. Uma das formas de se remediar tais violações seria a modelagem da estrutura geral de covariância, a qual representa as variâncias e as correlações das medidas.

No componente representando as variâncias, uma matriz diagonal pode ser criada com a seguinte expressão:

$$G(\beta,\varnothing) = diag[g^2(\mu_1,z_1,\theta), g^2(\mu_1,z_1,\theta), \dots, g^2(\mu_n,z_n,\theta)] \qquad (7)$$

Como foi exposto anteriormente, a matriz de correlação é representada pela expressão $\Gamma(\alpha)$ que, combinada com a expressão (7), gera a seguinte expressão geral de variância-covariância para os resíduos:

$$Cov(e) = \sigma^2 G^{1/2}(\beta,\theta)\Gamma(\alpha)G^{1/2}(\beta,\theta) = R(\beta,\xi), \quad onde\ \xi = [\sigma,\theta^T,\alpha^T]^T \qquad (8)$$

Vale comentar que nas expressões acima o que está em negrito representa vetores ou matrizes e os caracteres sem negritos são escalares. Por exemplo, o componente de variância σ representa um escalar, o qual possui um valor único para a base de dados. Caso se tenha uma estrutura de variância-covariância combinando uma função de heterogeneidade da forma exponencial e uma estrutura de covariância autorregressiva heterogênea de ordem um, a expressão (8) teria a seguinte forma:

$$Cov(e) = \sigma^2 \begin{bmatrix} \exp(\mu_1\Phi)^{½} & 0 & 0 & \dots & 0 \\ 0 & \exp(\mu_1\Phi)^{½} & 0 & \dots & 0 \\ 0 & 0 & \exp(\mu_1\Phi)^{½} & \dots & 0 \\ 0 & 0 & 0 & \dots & \exp(\mu_1\Phi)^{½} \end{bmatrix} \begin{bmatrix} 1 & \rho_1 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \dots & \rho_{n-2} \\ . & . & \dots & . \\ \rho_{n-1} & \rho_{n-2} & \dots & 1 \end{bmatrix} \begin{bmatrix} \exp(\mu_1\Phi)^{½} & 0 & 0 & \dots & 0 \\ 0 & \exp(\mu_1\Phi)^{½} & 0 & \dots & 0 \\ 0 & 0 & \exp(\mu_1\Phi)^{½} & \dots & 0 \\ 0 & 0 & 0 & \dots & \exp(\mu_1\Phi)^{½} \end{bmatrix} \quad (9)$$

A expressão (9), após a multiplicação dos três componentes, resultará em uma matriz de dimensão (n x n), com a seguinte estrutura:

$$Cov(e) = \sigma^2 \begin{bmatrix} \exp(\mu_1\Phi) & \exp(\mu_2\Phi)\rho_1 & \dots & \exp(\mu_n\Phi)\rho_{n-1} \\ \exp(\mu_2\Phi)\rho_1 & \exp(\mu_3\Phi)\rho_2 & \dots & \exp(\mu_{n-1}\Phi)\rho_{n-2} \\ \exp(\mu_3\Phi)\rho_2 & \exp(\mu_3\Phi)\rho_3 & \dots & \exp(\mu_{n-2}\Phi)\rho_{n-3} \\ . & . & . & . \\ . & . & . & . \\ \exp(\mu_n\Phi)\rho_{n-1} & \exp(\mu_{n-1}\Phi)\rho_{n-2} & \dots & \exp(\mu_1\Phi)^{½} \end{bmatrix} \quad (10)$$

Em (10), a heterogeneidade da variância está sendo corrigida pelas expressões localizadas na diagonal principal da matriz de variância-covariância, as quais estão em função do escalar σ^2, da média de cada observação e de um parâmetro estimado Φ. A autocorrelação é corrigida pelos elementos fora da diagonal principal em (10), os quais estão em função de σ^2, da média de cada observação, do parâmetro estimado Φ e do coeficiente de correlação entre as observações ρ.

RESULTADOS E DISCUSSÃO

Seleção do melhor modelo

Como pode ser verificado na Tabela 2, os quatro modelos foram ajustados baseando-se em três parâmetros. Esta opção foi feita devido ao fato da inexistência de um ponto de inflexão claro na base de dados, gerando uma curva da forma sigmóide. Neste caso específico, como está representado na Figura 1, a base de dados representa uma tendência assintótica. Tal opção facilitou o processo de convergência na estimativa.

A Tabela 2 também mostra que todos os quatro modelos utilizados para explicar a variação altura em função da idade tiveram significância nos parâmetros (valor-p<0.0001). Portanto, considerando apenas este critério, qualquer dos quatro modelos poderia ser utilizado para a estimativa da altura em função da idade. Porém, baseando-se em outros critérios, alguns modelos são superiores a outros. Quando se considera o erro padrão de cada parâmetro, o modelo logístico apresenta menores valores, gerando intervalos de confiança de menor amplitude e maiores valores de t associados aos parâmetros. Outra característica positiva do modelo logístico está associada com as correlações entre parâmetros. Pode-se verificar na Tabela 2 que as correlações entre os pares de parâmetros foram menores, indicando que a presença dos parâmetros se faz necessária e que o modelo não possui um número excessivo de parâmetros. Nos outros três modelos houve uma alta correlação entre os parâmetros 2 e 3, indicando que a presença de um dos dois não se faz necessária. Outra característica importante do modelo logístico, como foi comentado anteriormente, é a sua interpretabilidade dos três parâmetros, o qual facilita o processo de convergência do algoritmo para a estimativa dos mesmos. Portanto, a partir daqui o modelo logístico vai ser utilizado para a complementação da modelagem.

Inclusão de covariantes no modelo

Conforme comentado, a variação da altura total de árvores individuais não é apenas explicada pela idade. Em vista disto, foram incluídas algumas variáveis relativas à árvore individual (DAP) e ao povoamento (Clone, G e HD), gerando um modelo com uma característica mais prática de utilização. A Figura 2 mostra a variação nos valores dos parâmetros em função das variáveis supracitadas. O parâmetro Assíntota foi o mais sensível, aumentando o seu valor com o aumento no valor das variáveis. Também pode ser visto diferentes valores dos três parâmetros para os diferentes clones, sugerindo a inclusão do clone como uma variável indicadora (dummy) no modelo.

Baseado nos resultados da Figura 2, os parâmetros do modelo logístico foram decompostos e estimados com a inclusão de variáveis DAP, Clone, G e HD, conforme expressão (11).

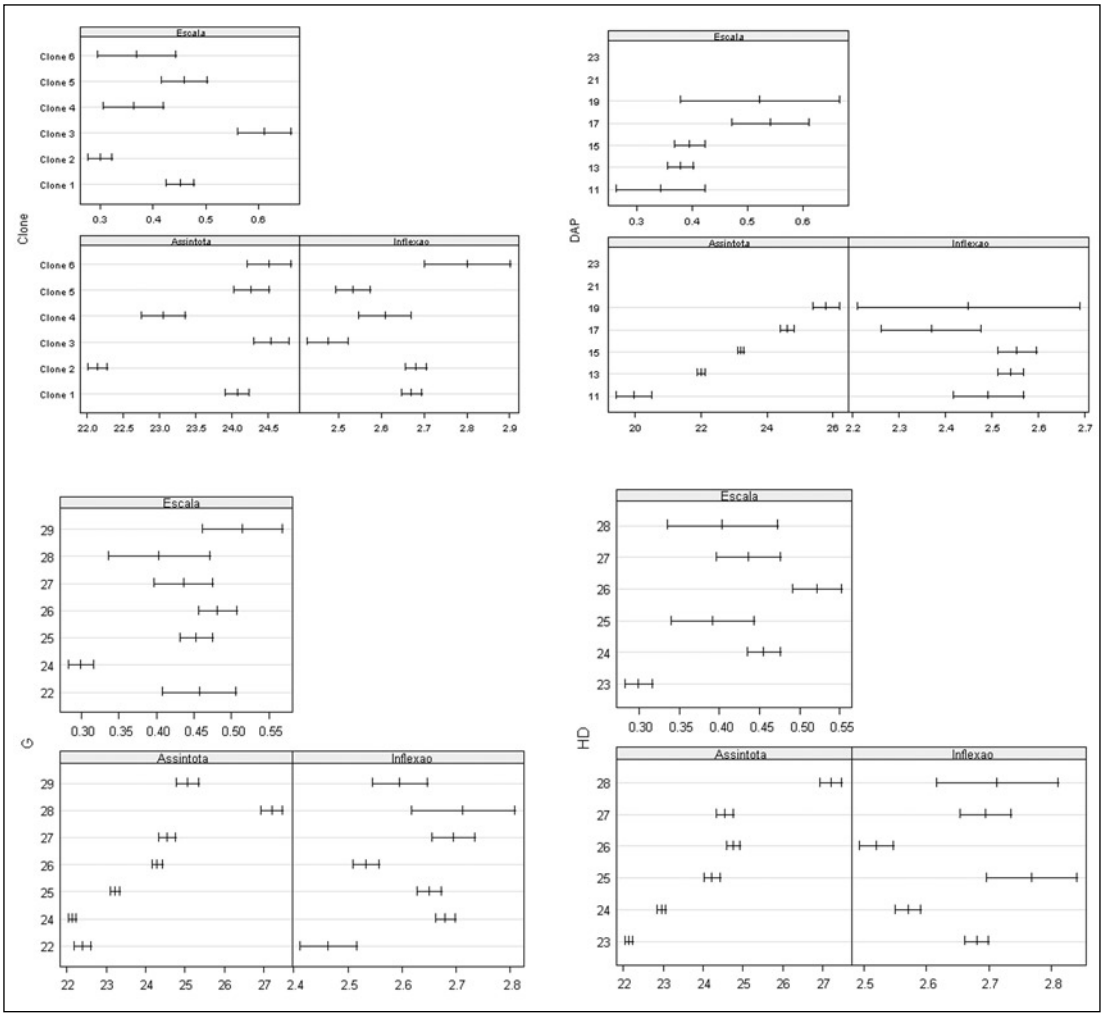

Figura 2.
Comportamento dos parâmetros do modelo logístico em função de algumas variáveis, onde: Clone = Código do clone relativo à parcela; DAP = Diâmetro à Altura do Peito por árvore, em cm; G = Área Basal das parcelas amostradas, em m²/parcela; e HD = Altura médias das árvores dominantes e co-dominantes de cada parcela, em metros.
(Parameter behavior of the logistic model as a function of the some variables, where: Clone = Clonal code of the sample unit; DAP = Diameter at Breast Height, in cm; G = Basal área by sample unit; HD = Average Height of the dominant and codominant trees, in meters).

$$\Phi_j = \phi_{j0} + (\sum_{i=1}^{5} \phi_{ji} + Clone(i)) + \phi_{j6} DAP + \phi_{j7} G + \phi_{j8} HD \quad (11)$$

em que:

Φ_j = Valor final do parâmetro (Assíntota para j=1; Ponto de inflexão para j=2; e Escala para j=3);

ϕ_{j0} = Valor do intercepto para o j-th parâmetro;

$\sum_{i=1}^{5} \phi_{ji} + Clone(i)) = \phi_{ji}$ = refere-se ao parâmetro associado ao i-th Clone e Clone(i) é uma variável indicadora com valor 1 para o i-th clone e 0 para os outros clones, para o j-th parâmetro;

$\phi_{j6}, \phi_{j7}, \phi_{j8}$ = Efeitos associados ao DAP, área basal e altura dominante, respectivamente, para o j-th parâmetro.

Tabela 2.
Estimativas e correlações para os quatro modelos utilizados na representação da variação da altura em função da idade.
(Estimates and correlations for the four models used to represent the height variation as a function of the age)

Parâmetro Estimado	Estimativas					Correlações	
	Valor	Erro Padrão	GL	Valor-t	Valor-p	Φ_2	Φ_3
Modelo Logístico (EPR=1,10736)							
Φ_1	23,6751	0,04264	1	555,2	<0.0001	-0,53	0,86
Φ_2	2,58177	0,00767	1	336,3	<0.0001	X	-0,85
Φ_3	0,46520	0,00787	1	59,07	<0.0001	X	X
Modelo de Richards (EPR=0.10735)							
Φ_1	23,6751	0,0426	1	555,2	<0.0001	-0,82	-0,85
Φ_2	5,54990	0,1083	1	51,2	<0.0001	X	0,99
Φ_3	2,14961	0,0364	1	59,0	<0.0001	X	x
Modelo de Gompertz (EPR=1,10467)							
Φ_1	23,8073	0,0477	1	499,10	<0.0001	-0,86	-0,88
Φ_2	4,5958	0,1016	1	45,23	<0.0001	X	0,99
Φ_3	1,8843	0,0344	1	54,77	<0.0001	X	X
Modelo de Weibull (EPR=1,10262)							
Φ_1	23,9768	0,0546	1	439,13	<0.0001	-0,89	-0,91
Φ_2	925,6	86,581	1	10,691	<0.0001	X	0,99
Φ_3	1,6208	0,0327	1	49,56	<0.0001	X	x

O modelo final, após a seleção das variáveis significativas associadas a cada parâmetro, teve a seguinte forma:

$$h_i = \frac{\phi_{10} + \sum_{i=1}^{5} \phi_{1i} + Clone(i) + \phi_{16} DAP + \phi_{17} G + \phi_{18} HD}{1 + \exp\left[\frac{(\phi_{20} + \sum_{i=1}^{5} \phi_{2i} + Clone(i) + \phi_{26} DAP + \phi_{28} HD\text{-}Idade_i)}{(\phi_{30} + \sum_{i=1}^{5} \phi_{3i} + Clone(i) + \phi_{37} G)}\right]} \quad (12)$$

As variáveis que contribuíram significativamente para explicar a variação da altura estão listadas na Tabela 3 e os seus valores encontram-se na Tabela 4. Pode ser verificado em (12) que apenas o parâmetro da assíntota foi influenciado significativamente por todas as variáveis, tanto em nível de árvore individual como de povoamento. O parâmetro relativo ao ponto de inflexão foi influenciado pelos clones, pelo DAP e pela altura dominante, enquanto o parâmetro relativo à escala teve influência apenas dos clones e da área basal. Pode-se perceber na Tabela 4 que, com exceção de alguns clones, todos os parâmetros associados às diferentes variáveis tiveram significância. Os clones que não foram significativos não foram agrupados por questões práticas, considerando que os mesmos podem ter diferentes usos finais em um processo produtivo. A melhoria da precisão da estimativa, com a inclusão das variáveis no modelo, pode ser verificado pela redução do erro padrão residual em 53% (de 1,28 para 0,60).

Modelagem da heterogeneidade da variância

Para a modelagem da heterogeneidade da variância, por questões práticas, foi utilizado o modelo original, sem a inclusão de covariantes no mesmo. A Tabela 5 mostra a comparação do modelo logístico entre sua forma homocedástica e heterocedástico. Como pode ser visto, tanto para os critérios de informação estatística (Akaike e Bayesiano – quanto menor melhor) quanto para o logaritmo da máxima verossimilhança (quanto maior melhor), o modelo heterocedástico teve melhor performance, gerando um valor de probabilidade altamente significativo. Baseando-se na expressão (2), a melhor forma encontrada para a base de dados em questão foi a exponencial, com a variável idade na base e expoente θ de 0,93. Uma outra característica positiva do modelo heterocedástico foi a redução do erro padrão residual em 88% (de 0,60 para 0,07), com a manutenção dos valores estimados dos parâmetros.

Tabela 3.
Teste F para os parâmetros significativos para o modelo da relação hipsométrica, onde: G.L.N. = o grau de liberdade do numerador; G.L.D. = o grau de liberdade do denominador.
(F test for the parameters in the hypsometric model, where: G.L.N. = Numerator Degrees of freedom; G.L.D. = Denominator Degrees of freedom)

Parâmetro	Variável Associada ao Parâmetro	G.L.N.	G.L.D.	Valor de F	Valor da Probabilidade > F
Assíntota	Intercepto	1	6600	2491	<.0001
	Clone	5	6600	6563	<.0001
	DAP	1	6600	4848	<.0001
	HD	1	6600	2044	<.0001
	Área Basal	1	6600	5781	<.0001
Ponto de Inflexão	Intercepto	1	6600	1419	<.0001
	Clone	5	6600	191	<.0001
	DAP	1	6600	1987	<.0001
	HD	1	6600	25	<.0001
Escala	Intercepto	1	6600	4832	<.0001
	Clone	5	6600	49	<.0001
	Área Basal	1	6600	48	<.0001

Tabela 4.
Parâmetros estimados para o modelo completo com suas respectivas estatísticas.
(Parameters estimated for the complete model and its statistics)

Parâmetro	Termo Adicionado ao Parâmetro	Valor	Erro Padrão	G.L.	Valor de t	Probabilidade
Assíntota	Intercepto	-1,340510	0,7948	6233	-1,687	0,0917
	Clone 1	0,550509	0,0635	6233	8,671	<,0001
	Clone 2	0,316457	0,0342	6233	9,257	<,0001
	Clone 3	-0,058607	0,0405	6233	-1,448	0,1476
	Clone 4	-0,024618	0,0243	6233	-1,012	0,3115
	Clone 5	0,064362	0,0266	6233	2,419	0,0156
	DAP	0,180806	0,0098	6233	18,464	<,0001
	G	0,299761	0,0363	6233	8,253	<,0001
	HD	0,582701	0,0468	6233	12,462	<,0001
Inflexão	Intercepto	2,486176	0,0738	6233	33,707	<,0001
	Clone 1	0,009945	0,0057	6233	1,747	0,0808
	Clone 2	-0,058013	0,0044	6233	-13,247	<,0001
	Clone 3	0,000394	0,0040	6233	0,099	0,9211
	Clone 4	-0,022186	0,0025	6233	-8,737	<,0001
	Clone 5	0,034589	0,0039	6233	8,794	<,0001
	DAP	-0,006963	0,0015	6233	-4,749	<,0001
	HD	0,008660	0,0030	6233	2,872	0,0041
Escala	Intercepto	-0,118836	0,0501	6233	-2,370	0,0178
	Clone 1	-0,026638	0,0049	6233	-5,384	<,0001
	Clone 2	0,054708	0,0036	6233	15,206	<,0001
	Clone 3	-0,017567	0,0036	6233	-4,881	<,0001
	Clone 4	0,008788	0,0022	6233	4,061	<,0001
	Clone 5	-0,008427	0,0029	6233	-2,884	0,0039
	G	0,020182	0,0020	6233	10,207	<,0001

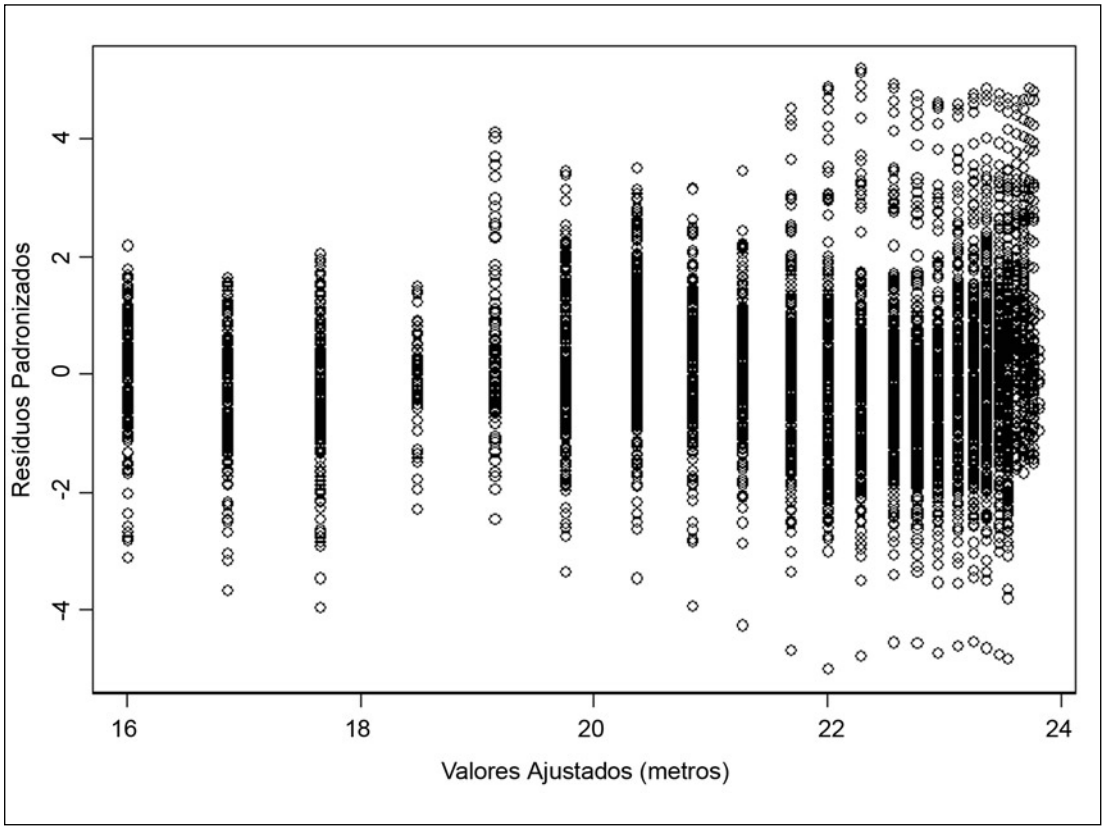

Figura 3.
Distribuição dos resíduos padronizados para o modelo logístico homocedástico.
(Standard residual distribution for the homoscedastic logistic model)

Tabela 5.
Comparação entre o modelo homocedástico e o heterocedástico, onde: GL=Graus de Liberdade; CIA=Critério de Informação de Akaike; CIB=Critério de Informação Bayesiano; LogMV=Logaritmo da Máxima Verossimilhança; TRMV=Teste da Razão da Máxima Verossimilhança; Valor-p=Valor da probabilidade acima do valor do qui-quadrado calculado.
(Comparing the homocedastic and the heterocedastic model, where: GL= Degrees of freedom; CIA= Akaike information criterion; CIB= Bayesiano information criterion; LogMV= Logaritm of the Likelihood; TRMV=Likelihood Ratio test; Valor-p= probability value above the critical qui-square value).

Modelo	GL	CIA	CIB	LogMV	TRMV	Valor-p
1- Homocedástico	4	8880	8903	-4436	1vs.2=64	<0,0001
2- Heterocedástico	5	8818	8848	-4404		
3-Heterocedástico/Autocorrelação	5	1748	1777	-869	1vs.3=7134	<0,0001

Antes da modelagem da heterogeneidade da variância (Figura 3), a distribuição dos resíduos apresentava uma forma de funil pelo modelo homocedástico. Esta foi corrigida pelo modelo heterocedástico (Figura 4), o qual apresentou uma distribuição praticamente homogênea para os diferentes valores ajustados, não violando as pressuposições básicas para os modelos não lineares clássicos.

Modelagem da autocorrelação

Também na modelagem da autocorrelação foi utilizado o modelo logístico na sua forma simples, sem a inclusão de covariantes. Por se tratar

de uma base de dados longitudinais, ou séries temporais, com medidas repetidas por indivíduo, espera-se autocorrelação entre as medidas no tempo e, conseqüentemente, uma melhora na qualidade do ajuste pela modelagem da mesma, já que também viola uma das pressuposições básicas dos modelos não lineares clássico, que é a independência entre as observações.

Também foram testadas combinações do modelo autorregressivo e de média móvel, conforme expressão (6). A melhor combinação foi o modelo ARMA(1,1), ou seja, modelo autorregressivo e de média móvel de grau um para ambos, conforme Tabela 5. Pode-se observar que tanto os critérios estatísticos quanto a máxima verossimilhança tiveram valores significativamente melhores quando comparados com o modelo apenas heterocedástico, sendo que o teste da razão da máxima verossimilhança gerou um valor de -869, com probabilidade menor que 0,0001, ou seja, altamente significativo.

A Tabela 6 mostra uma análise comparativa na estimativa dos parâmetros para os três modelos logísticos estudados. Os parâmetros estimados tiveram maior variações para o modelo combinando heterogeneidade de variância e autocorrelação, com redução do erro padrão residual. Tal resultado se torna importante na utilização da equação gerada para fins de predição. Se os parâmetros forem mudados, a curva muda de posição e as estimativas da altura em função da idade são mais precisas e exatas gerando, também, melhor confiabilidade nas predições e projeções futuras das estimativas. Com isto, o processo de planejamento do uso da produção, por exemplo, se torna mais confiável.

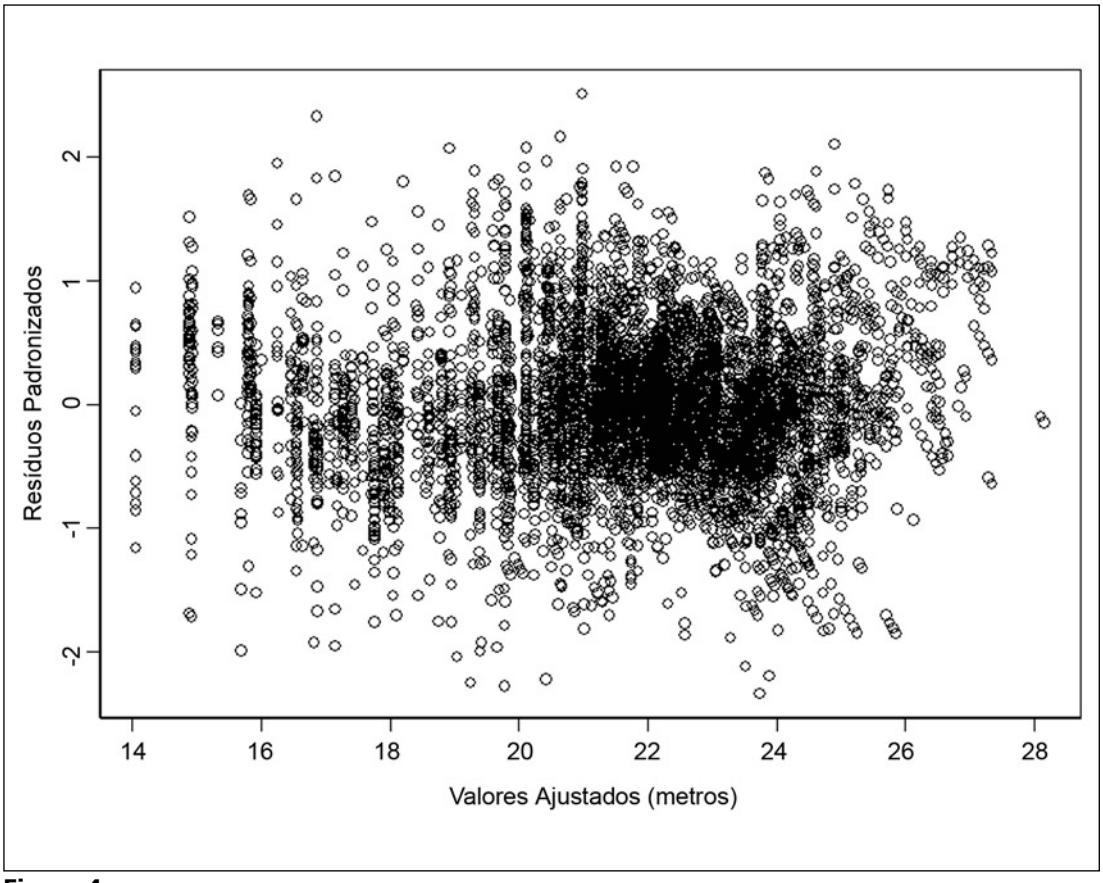

Figura 4.
Distribuição dos resíduos padronizados para o modelo logístico heterocedástico.
(Standard residual distribution for the heteroscedastic logistic model)

Tabela 6.
Resumo das estimativas para os três modelos propostos, onde: GL=Graus de Liberdade; Φ´s são parâmetros estimados para o modelo logístico; EPR=Erro Padrão Residual; θ = Potência estimada na modelagem da heterogeneidade da variância; φ e η = Parâmetros estimados na modelagem da autocorrelação, referentes aos componentes autoregressivo e de média móvel, respectivamente.
(Summary of the estimates for the three models, where: GL= Degrees of freedom; Φ´s are the estimated parameters; EPR= Residual standard error; θ = Estimated power in modeling the variance heterogeneity; φ e η = Estimated parameters in modeling the autocorrelation of the autoregressive and moving mean componentes, respectively).

Modelo	GL	Parâmetros Estimados			EPR	Het.	Aut.	
		Φ_1	Φ_2	Φ_3		θ	φ	η
Homocedástico	4	24,574	2,6137	0,4817	0,0305	-	-	-
Heterocedástico	5	24,575	2,6144	0,4815	0,0020	1,3532	-	-
Heterocedástico/Autorregressivo	7	24,792	2,7028	0,4475	0,0051	1,1050	0,9782	-0,014

CONCLUSÕES

• A heterogeneidade da variância e, ou a autocorrelação estão presentes com bastante freqüência nas bases de dados florestais;

• Tais características violam as pressuposições básicas dos modelos lineares e não lineares clássicos, sendo necessário o uso da metodologia dos modelos lineares e não lineares generalizados para a modelagem de tais bases de dados;

• A combinação software/hardware disponíveis na atualidade possui determinante importância no sucesso deste tipo de modelagem, devido a sua complexidade e as dimensões matriciais geradas no processo de estimativa;

• Dos quatro modelos não-lineares testados (Logístico, Richards, Gompertz e Weibull), o de melhor performance para a base de dados do estudo foi o Logístico, o qual apresentou menores correlações entre os parâmetros estimados, indicando a não presença de excesso de parâmetros no modelo. Adicionalmente, o citado modelo possui a importante característica da interpretabilidade biológica dos parâmetros, a qual facilita significativamente na escolha dos valores iniciais para o processo iterativo do algoritmo utilizado na estimativa dos parâmetros;

• A inclusão de covariantes no modelo selecionado, associados individualmente aos parâmetros, melhorou significativamente a precisão do modelo;

• A distribuição dos resíduos foi sensivelmente melhorada com o modelo heterocesdástico, eliminando a forma de funil do mesmo, o qual gerou diferentes precisões nas estimativas da variável resposta;

• Na modelagem da autocorrelação, a melhor combinação foi ARMA(1,1), a qual elevou o valor do logaritmo da máxima verossimilhança de -4404 para -869, gerando um valor para o teste da razão da máxima verossimilhança altamente significativo (pr<0,0001).

• A modelagem da autocorrelação provocou mudanças nos parâmetros originais estimados para o modelo Logístico, influenciando, também, no uso do mesmo para o processo de predição e projeção.

AUTORES

Natalino Calegario é Professor Doutor do Departamento de Ciências Florestais da Universidade Federal de Lavras – UFLA - Caixa Postal 3037 - Lavras, MG – 37200- 000 – E-mail: calegari@ufla.br

Cristina Lellis Leal Calegario é Doutora em Economia Agrícola pela Universidade da Geórgia, USA, e técnica do Departamento de Administração e Economia da Universidade Federal de Lavras - UFLA - Caixa Postal 3037 - Lavras, MG – 37200- 000 – E-mail: cristinaleal@ufla.br

Romualdo Maestri é Doutor e pesquisador da Aracruz Celulose S/A. Aracruz, ES - E-mail: rmaestri@aracruz.com.br

Richard F. Daniels é Professor Doutor da Escola de Recursos Florestais Daniel B. Warnell – Geórgia University – UGA - Athens, GA - 30602-2152. E-mail: ddaniels@smokey.forestry.uga.edu

REFERÊNCIAS BIBLIOGRAFIAS

ATKINSON, A.C. **Plots, transformations, and regression**. Oxford: Clarendon Press, 1987. 296p.

BATISTA, J.L.F; COUTO, H.T.Z.; MARQUESINI, M. Desempenho de modelos de relações hipsométricas: estudo em três tipos de floresta. **Scientia Forestalis**, Piracicaba, n.60, p.149-163, 2001.

BOX, G.E.P.; COX, D.R. An analysis of transformations. J**ournal of the Royal Statistical Society**, Series B, London, v.26, p.211-246, 1987.

BOX, G.; JENKINS, G.; REINSEL, G. **Time series analysis: forecasting and control**. New Jersey: Prentice Hall, 1994.

CARROLL, R.J.; RUPPERT, D. **Transformations and weighting in regression**. New York: Chapman & Hall, 1988. 264p.

COOK, R.D.; WEISBERG, S. Diagnostics for heteroscedasticity in regression. **Biometrika**, Oxford, v.70, p.1-10, 1983.

DAVIDIAN, M.; CARROLL, R.J. Variance function estimation. **Journal of the American Statistical Society**, Alexandria, v.82, p.1079-1091, 1987.

DAVIDIAN, M.; GILTINAN, D.M. **Nonlinear models for repeated measurement data**. London: Chapman and Hall, 1995. 359p.

DÍAZ, M.P.; COUTO, H.T.Z. Modelos generalizados para a mortalidade de árvores de *Eucalyptus grandis* no Estado de São Paulo, Brasil. **Scientia Forestalis**, Piracicaba, n.56,p.101-111, 1999.

DRAPER, N.R.; SMITH, H. **Applied regression analysis**. New York: John Wiley, 1998. 736p.

NETER J.; KUTNER, M.H.; NACHTSHEIM, C.J.; WASSERMAN, W. **Applied linear statistical models**. New York: McGraw-Hill, 1996. 1408p.

RATKOWSKY, D.A. **Nonlinear regression modeling: a unified practical approach**. New York: Marcel Dekker, 1983. 276 p.

RICHARDS, F.J. A flexible growth function for empirical use. **Journal of Experimental Biology**, v.10, p.290-300, 1959.

SEBER, G.A.F.; WILD, C.J. **Nonlinear regression**. New York: John Wiley, 1989. 792p.

WEIBULL, W. A statistical distribution function of wide applicability. J**ournal of Applied Mechanics**, v.18, p.293-296, 1951.