ISSN 0103-9954

AJUSTE E SELEÇÃO DE MODELOS TRADICIONAIS PARA SÉRIE TEMPORAL DE DADOS DE ALTURA DE ÁRVORES

FITTING AND SELECTING TRADITIONAL MODELS FOR TREE'S HEIGHT TIME SERIES DATA

Eduardo Pagel Floriano[1] Ivanor Müller[2] César Augusto Guimarães Finger[3] Paulo Renato Schneider[4]

RESUMO

A medição da altura das árvores é de extrema importância para o planejamento da produção florestal. Geralmente, é realizada por meio de amostragens por causa do tamanho das populações e das próprias árvores. Medições ao longo do tempo formam séries de dados temporais que implicam em certos problemas para o ajuste de equações que descrevam sua evolução. Muitos modelos de equações foram desenvolvidos com essa finalidade, sendo que neste trabalho são utilizados modelos lineares, logarítmicos, não-lineares linearizáveis e não-linearizáveis para descrever a altura ao longo do tempo. As estatísticas utilizadas para comparação entre modelos são o coeficiente de determinação (R^2), a estatística C_p de Mallows, o critério de informação de Akaike (Akaike*'s information criterion – AIC*), o quadrado médio dos resíduos (*QMres*) e a análise gráfica de resíduos. O objetivo deste trabalho foi desenvolver um exemplo de ajustamento de equações de crescimento para altura, verificar quais se adaptam melhor aos dados populacionais e determinar que critérios de seleção, entre os utilizados, têm mais relação com o verdadeiro melhor modelo. Para tanto, foi utilizada uma amostra de 64 árvores, provenientes de uma população de 531 árvores de *Pinus elliottii* Engelm. Nesse caso, as estatísticas da amostra são comparadas com as estatísticas da população, demonstrando qual modelo descreve melhor os dados da população. A qualidade do ajuste dos dados da população aos estimados por cada modelo foi avaliada pelo teste Qui-Quadrado e análise gráfica dos resíduos. O uso do critério de Akaike (*AIC*) mostrou-se adequado na seleção de modelos para os dados utilizados. As duas melhores equações foram a equação $h = b_0 + b_1.t + b_2.t^5$ e o modelo de Chapman-Richards, que não apresentaram diferenças significativas entre si para os critérios analizados. Nesse sentido, o critério de Akaike, calculado para os dados amostrais, mostrou-se eficiente como critério de seleção de equações para descrever a altura das árvores ao longo do tempo, para a população utilizada neste estudo. A generabilidade, calculada pelo teste Qui-Quadrado em relação à população, não mostrou diferença significativa entre os modelos 3 e 9. A seleção final, usando-se os critérios qualitativos de ligação do modelo com o processo estudado, sua interpretabilidade e compreensibilidade, determinou a escolha do modelo de Chapman-Richards como o melhor para descrever o crescimento em altura das árvores estudadas.

Palavras-chave: *Pinus*; modelos de crescimento em altura; ajuste; seleção.

ABSTRACT

Mesuring trees' height is very importance for planning forest production. Usually, it is accomplished through samplings due to the size of the populations and the size of the trees themselves. Measurements along time form a time series data with some problems for the adjustment of equations to describe its growth. Several models were developed with that purpose. The equations used in this paper were linear, logarithmic, and non-linear models. The statistics used for comparison of those models were the determination coefficient (R^2), Cp of Mallows, Akaike's information criterion (AIC), Schwarz's Bayesian criterion (SBC/BIC), squared mean of residues and the graphic analysis of residues. The objective of this work was to develop an example of adjustment of growth equations for height, to demonstrate which one adapts better to the population data and to determine which selection criteria have more relationship with the better true model. To do so, a sample of 64 trees was used, submitted to the trunk analysis, from a population of 531 trees of *Pinus elliottii*

1. Engenheiro Florestal, M.Sc., Doutorando do Programa de Pós-Graduação em Engenharia Florestal, Centro de Ciências Rurais, Universidade Federal de Santa Maria, CEP 97.105-900, Santa Maria (RS). Bolsista da CAPES. eduardofloriano@correios.com.br
2. Engenheiro Florestal, Dr., Professor Adjunto do Departamento de Estatística, Centro de Ciências Naturais e Exatas, Universidade Federal de Santa Maria, CEP 97.105-900, Santa Maria (RS). ivanormuller@smail.ufsm.br
3. Florestal, Dr., Professor Adjunto do Departamento de Ciências Florestais, Centro de Ciências Rurais, Universidade Federal de Santa Maria, CEP 97.105-900, Santa Maria (RS). finger@ccr.ufsm.br
4. Florestal, Dr., Professor Titular do Departamento de Ciências Florestais, Centro de Ciências Rurais, Universidade Federal de Santa Maria, CEP 97.105-900, Santa Maria (RS). paulors@ccr.ufsm.br

Recebido para publicação em 27/09/2005 e aceito em 17/04/2006.

Engelm. The statistics of the sample were compared to the statistics of the population, demonstrating which model describes better the data of the population. Quality of the adjustment to the population's data of each model was evaluated through the Chi-square test and graphic analysis of residues. The Akaike's information criterion (AIC) was appropriated to select models for the data. The two better equations were $h=b_0+b_1.t+b_2.t^5$ and Chapman-Richards' growth model, which showed no significant differences for the chosen criteria in this study. In this sense, the Akaike's information criterion (AIC) calculated to the sample data showed efficiency as an equations' selection criterion to describe the height of the trees along the time for the population used in this study. The generability, calculated by Qui-square test, in relation to the population, didn't show significant difference between models 3 and 9. Final selection, using the qualitative criteria of connection of the model to the studied process, its interpretability and comprehensibility, determined the choice of the Chapman-Richards' model as the best to describe the height growth for the studied trees.

Keywords: *Pinus*; heigth growth models; fitting; selecting.

INTRODUÇÃO

A medição da altura das árvores é de extrema importância para o planejamento da produção florestal, pois é fundamental na determinação do volume e para classificação de sítios (Finger, 1992).

Entretanto, medir a altura de todas as árvores de um povoamento ano após ano é inconcebível. Há muito que se utiliza a amostragem para estudar florestas por dois motivos principais. O primeiro é que, em conseqüência do tamanho das populações de árvores, torna-se quase impossível realizar um censo abrangendo todos os indivíduos de uma floresta. O segundo é que, por causa das grandes dimensões das árvores, torna-se difícil fazer medições. Quando é necessário estudar o crescimento das árvores ao longo do tempo, o problema se torna ainda mais complexo, pois envolve processos de aquisição de dados em séries temporais e análises especiais pelas características desse tipo de dados.

Conseqüentemente, tem-se usado a amostragem e a modelagem matemática dos dados para descrever a altura das árvores. Muitos modelos de equações foram desenvolvidos com essa finalidade e, dependendo da situação, algumas equações se adaptam melhor do que outras e algumas vezes há dificuldade em se decidir que modelo deve ser utilizado.

O objetivo deste trabalho é desenvolver um exemplo de ajustamento de equações de crescimento para altura, verificar quais se adaptam melhor aos dados populacionais e determinar que critérios de seleção, entre os utilizados, têm mais relação com o verdadeiro melhor modelo.

REVISÃO BIBLIOGRÁFICA

Séries de dados representando o crescimento das árvores ao longo do tempo podem ser obtidas com bom nível de precisão por parcelas permanentes, ou medição de anéis de crescimento (Schneider, 1993; Nemec, 1995). Assim, procura-se restringir o número de observações ao mínimo necessário para se obter uma amostra representativa da população em estudo. Amostras de altura em série temporal podem ser consideradas como amostra em pares de dados de altura e idade, que apresentam a vantagem de ter menor flutuação que amostras independentes (Wonnacott e Wonnacott, 1980), com isso, torna-se possível reduzir ainda mais seu tamanho, desde que não se perca a sua representatividade com relação à população estudada.

Modelos de crescimento são vitalmente importantes para o planejamento da produção florestal. Prever o crescimento e rendimento de sítios locais é uma condição prévia para planejar a administração de florestas em qualquer nível. Portanto, há necessidade de se conhecerem as técnicas de modelagem do crescimento e as suas limitações (Garcia, 1988).

Existem vários modelos de crescimento que se adaptam a diferentes situações. Há modelos lineares e não-lineares. Em muitas situações, mais de um modelo se adapta ao que se quer modelar; então, o problema passa a ser a escolha do melhor modelo, o que nem sempre é tarefa fácil. Muitos métodos já foram desenvolvidos com essa finalidade, geralmente, levando em consideração dois critérios gerais de avaliação: 1) o ajustamento da função aos dados; 2) e o afastamento dos dados com relação ao modelo. A forma mais simples de escolha de modelos baseia-se, portanto, no uso do coeficiente de determinação (R^2) da regressão, que estima o ajustamento, e no quadrado médio dos resíduos (*QMres*) que estima o afastamento, como recomendado por Sit (1994). Geralmente, a análise gráfica dos resíduos também é usada para decidir sobre o

melhor modelo conforme descrito por Bussab (1986). Outros critérios de seleção de modelos são usados em programas computacionais como o R^2 ajustado, a estatística C_p de Mallows, o critério de informação de Akaike (*Akaike's information criterion – AIC*), o critério bayesiano de Schwarz (Schwarz's Bayesian criterion – *SBC*) também chamado de critério de informação bayesiano de Schwarz (*Bayesian information criterion – BIC*) e o critério de predição de Amemiya (Amemiya's *prediction criterion – PC*) (SAS Institute, 1999b), ainda, o critério bootstrap de Efron (*Efron bootstrap criterion – EBC*), o critério de validação-cruzada (*cross-validation criterion – CVC*) (Zucchini, 2000) e o mínimo comprimento de descrição (minimum description length – *MDL*) (Myung *et al.*, 2003).

A seleção de modelos por causa de sistemas computacionais inclui a escolha do melhor número de nós escondidos em uma rede neural, a determinação do ponto de corte a ser executado em uma árvore de decisão e a escolha do grau de ajuste polinomial a uma série de pontos. Em cada um desses passos, a meta não é minimizar o erro nos dados de trabalho, mas minimizar o erro de generalização resultante (Kearns *et al.*, 1997), ou seja, o modelo deve descrever o melhor possível a população, não a amostra. Muitas comunidades de pesquisa propuseram diversos algoritmos para seleção de modelos usando vários tipos de análise para julgar o desempenho de cada algoritmo, incluindo consistência assintótica no sentido estatístico, otimização assintótica sob medidas de codificação teóricas e, mais raramente, taxas de convergência para o erro de generalização (Kearns *et al.*, 1997). Entretanto, pode-se dizer que, mesmo com diferentes formas de cálculo e diferentes critérios particulares, todas as formas de seleção de modelos usam, entre outros, um dos dois critérios gerais, ou ambos, separadamente ou combinados, quais sejam, o ajustamento e o erro, mas há outros critérios que são relevantes.

De acordo com Navarro e Myung (2004), ao avaliar um modelo, há vários fatores a considerar. Em termos gerais, podem ser usados métodos estatísticos para medir a suficiência descritiva de um modelo (ajustando-o aos dados e testando esses ajustes), como também sua generalização e simplicidade (usando ferramentas de seleção de modelos). Porém, a qualidade de um modelo também depende de sua interpretabilidade, de sua consistência com outros e de sua plausibilidade global. Isso implica em julgamentos inerentemente subjetivos, mas não menos importantes. Como sempre, não há nenhum substitutivo para avaliações pessoais e para o bom-senso, pois é crucial reconhecer que todos os modelos estão errados e uma meta realística de modelar é encontrar um modelo que represente uma "boa" aproximação à verdade em um senso estatisticamente definido. Testar a hipótese de nulidade é um método clássico de julgar o ajustamento de um modelo. A idéia é montar a hipótese nula de que o modelo está correto, então, obtém-se o valor de probabilidade "*p*" e toma-se uma decisão sobre rejeitar ou reter a hipótese, comparando o valor "*p*" resultante, com o nível *alfa* de probabilidade desejado.

Pode-se classificar os critérios de seleção em qualitativos e quantitativos. Os critérios qualitativos consideram a ligação do modelo com o processo estudado, sua interpretabilidade e compreensibilidade. Os critérios quantitativos levam em conta a falseabilidade, a qualidade do ajustamento, a complexidade e a generabilidade do modelo (Myung *et al.*, 2003).

Os critérios qualitativos para seleção de um modelo dizem respeito à sua suficiência explicativa. Um modelo satisfaz o critério de suficiência explicativa se suas suposições são plausíveis, consistentes com os resultados encontrados, e se a relação teórica é razoável para o processo de interesse. Em outras palavras, o modelo deve fazer mais que redescrecer os dados observados. O modelo também deve ser interpretável, fazer sentido e ser compreensível. É importante que os componentes do modelo, especialmente seus parâmetros, estejam ligados aos processos estudados. Em outras palavras, não há razão em se escolher um modelo que não se pode explicar (Myung *et al.*, 2003).

Os principais critérios quantitativos desejados em um modelo, de acordo com Myung *et al.* (2003), são descritos a seguir:

a) Não deve apresentar **falseabilidade**, ou seja, deve descrever todos os indivíduos da população. Um modelo é tanto mais falseável quanto maior o número de indivíduos da população que não consegue descrever; isso pode ser avaliado pela distribuição dos resíduos em um gráfico;

b) Deve ajustar-se perfeitamente aos dados. A **qualidade do ajustamento** pode ser medida pelas estatísticas: soma de quadrados do erro (*SQres*), quadrado médio do erro (*QMres*), erro-padrão

de estimativa (*Syx*), coeficiente de determinação ou proporção da variância explicada em relação à variação total (R^2) e a máxima verossimilhança (maximum likelihood – ML); todos podem ser obtidos pelos procedimentos PROC AUTOREG e PROC ARIMA do SAS System (Souza, 1998); o PROC REG provê os quatro primeiros;

c) Não deve ser complexo. A **complexidade** de um modelo é definida por meio de, pelo menos, duas dimensões, o número de parâmetros e a forma funcional do modelo; quanto menor o número de parâmetros, quanto menos cálculos envolve e quanto menor sua complexidade geométrica, menor a sua complexidade geral e melhor o modelo é considerado, desde que se adapte aos dados e à população estudada;

d) Deve ser generalizável. A **generabilidade** de um modelo é entendida como a capacidade do modelo em descrever não só os dados da amostra, mas de toda a população ao longo do período de tempo considerado e este é considerado o principal critério para a seleção de modelos; os critérios *AIC* e *BIC* levam em consideração a generabilidade.

Quando dados em séries temporais são usados em análise de regressão, geralmente o termo de erro não é independente no tempo. Ao contrário, os erros são seriamente correlacionados ou autocorrelacionados. Se o termo de erro é autocorrelacionado, as estimativas de parâmetros pelos mínimos-quadrados ordinários (OLS) são adversamente afetadas e as estimativas de erro-padrão são tendenciosas. Portanto, não é aconselhável usar a análise de regressão ordinária para dados de séries de tempo porque as pré-suposições nas quais o modelo de regressão linear clássico é baseado geralmente são violadas. (SAS Institute, 1999a).

Séries temporais normalmente têm distribuição não-linear e o método ordinário de mínimos quadrados estima erroneamente o erro-padrão, além disso, deve-se considerar que estatísticas como o *F* de Fischer e o *t* de Student podem apresentar tendenciosidade na análise de regressão de séries temporais, (Souza, 1998); nesses casos, é preferível usar critérios como o *AIC* e o *BIC* para seleção de modelos e avaliá-los usando estatísticas como a de Durbin Watson, por causa da possibilidade de existência de correlação serial. Não é adequado usar os testes *F* ou *t* para verificar a significância de modelos não-lineares ou para comparação de modelos, mas é possível utilizar o *AIC* calculado de forma aproximada como um indicativo para seleção de modelos não-lineares (Collett, 2003). No SAS System, o teste de Durbin-Watson é realizado por meio da opção DW da declaração MODEL do procedimento PROC REG para equações lineares e da opção DW da declaração FIT do procedimento PROC NLIN para equações não-lineares, sendo válido somente para séries temporais de dados (SAS System, 1999b).

De acordo com Motulsky e Christopoulos (2003), o ajustamento na regressão não-linear é quantificado como a soma dos quadrados das distâncias verticais da curva até os pontos de dados (soma de quadrados do erro). Pode-se equivocadamente assumir que o melhor modelo é o que minimiza a soma de quadrados do erro, mas não é assim tão simples comparar um modelo com outro. O problema é que um modelo mais complexo (com maior número de parâmetros) geralmente produz uma curva mais flexível que uma curva definida por um modelo mais simples. Isso significa que um modelo mais complexo pode se "retorcer" mais e então se ajustar melhor aos dados. Para comparar modelos, portanto, nem sempre se pode escolher o que se ajusta melhor aos dados amostrais e gera a menor soma de quadrados do erro. Há casos em que é necessário utilizar ferramentas estatísticas de aproximação. Deve-se considerar que os modelos podem estar relacionados, quando então são ditos aninhados. Modelos aninhados ocorrem quando são partes de um modelo maior em seqüência. Como exemplo têm-se as funções de Richards (4 parâmetros), Chapman-Richards (3 parâmetros) e Weber (2 parâmetros) em que uma é parte da outra; ou equações lineares do tipo simples, quadrática, cúbica, etc, com mesma variável independente básica. Nesse caso, a comparação de modelos é possível por meio de dois métodos distintos: pela soma de quadrados do erro usando-se o teste da *Razão de F* que tem a mesma distribuição do *F* de Fischer, por exemplo, e de ferramentas de aproximação como o critério de informação de Akaike (*AIC*). Entretanto, se os modelos não forem aninhados, é incorreto utilizar a soma de quadrados do erro, e o teste *F* não pode ser usado para a comparação; neste caso, devem ser utilizadas ferramentas de comparação como o *AIC*.

O cálculo da *Razão de F* (e.1) é realizado com a seguinte equação (Motulsky e Christopoulos, 2003):

$$Razão\ de\ F = ((Sq_1 - Sq_2)/(p_1 - p_2))/(Sq_2/p_2) \quad \text{(e.1)}$$

Sendo: Sq_1 = soma de quadrados da regressão 1; Sq_2 = soma de quadrados da regressão 2; p_1 = número de parâmetros da equação 1; p_2 = número de parâmetros da equação2.

Os cálculos do *AIC* (e.2), da diferença Δ *AIC* (e.3) entre duas equações com diferente número de parâmetros, sendo o primeiro mais simples que o segundo, a probabilidade *p* (e.4) de se escolher a equação correta entre duas equações em teste e a correção necessária do AIC_c (e.5) quando o número de observações for menor do que 12 vezes o número de parâmetros da equação[1], são realizados pelas seguintes equações (Motulsky e Christopoulos, 2003):

$$AIC = n \cdot ln(SQ/n) + 2K \quad \text{(e.2)}$$

$$\Delta AIC = AIC_2 - AIC_1 = n \cdot ln(SQ_2 / SQ_1) + 2(K_2-K_1) \quad \text{(e.3)}$$

$$p = e^{(-0,5.\ \Delta AIC)} / (1+e^{(-0,5.\ \Delta AIC)}) \quad \text{(e.4)}$$

$$AIC_c = AIC + 2K(K-1)/(n-K-1) \quad \text{(e.5)}$$

Em que: *AIC* = critério de informação de Akaike; AIC_c = critério de informação de Akaike corrigido para quando o número de observações for menor do que dez vezes o número de parâmetros; Δ *AIC = diferença entre os AIC* de duas equações com número diferente de parâmetros; *n* = número de observações; *ln* = logaritmo de base natural; *e* = base do logaritmo natural; *SQ* = soma de quadrados do erro; *K* = número de parâmetros da equação, inclusive o intercepto; SQ_1 = soma de quadrados do erro da equação com menor número de parâmetros; SQ_2 = soma de quadrados do erro da equação com maior número de parâmetros; K_1 = número de parâmetros da equação 1, mais simples; K_2 = número de parâmetros da equação 2, mais complexa. Quanto menor o valor do *AIC*, melhor é a equação. Quando o resultado da diferença (Δ *AIC*) é positivo, entre duas equações com diferente número de parâmetros, o modelo 1 com menor número de parâmetros é melhor e se for negativo, a equação 2 é melhor. Observação: se a equação for não-linear e não possuir um intercepto, soma-se 1 ao valor de *K*.

A lógica do teste pelo *AIC* é que não há hipótese sendo testada como no teste *F*. Ao contrário, o teste permite que se determine qual modelo é o mais correto e quanto. O teste pode ser utilizado para comparar qualquer tipo de modelo: lineares, não-lineares, aninhados e não-aninhados. A base teórica matemática do método de Akaike é bastante complexa; combina a teoria da máxima verossimilhança, a teoria da informação e o conceito de entropia da informação e é descrita na obra: *Model selection and multimodel inference – A practical information-theoretic approach, 2ª ed, 2003. de K.P. Burnham e D.R. Anderson.*

A seqüência de procedimentos para escolher equações pelo *AIC* deve respeitar os seis passos a seguir (Motulsky e Christopoulos, 2003):

1. Ajustar as equações;
2. Anotar a soma de quadrados do erro de cada equação; no caso de fatores ponderados, usar a soma ponderada de quadrados do erro;
3. Determinar o número de observações (*n*); se foi usada uma variável de freqüência, ela deve ser considerada como multiplicador, e o valor total de *n* deve ser aquele determinado pela soma das freqüências de cada classe de valor;
4. Determinar o valor de K para cada equação; nas equações não-lineares sem um intercepto, deve-se adicionar 1 ao valor do número de parâmetros da equação, porque a regressão não-linear estima o valor da soma de quadrados;
5. Calcular o *AIC,* ou se necessário, o AIC_c (*corrigido*);

[1] No caso do cálculo do *AIC*, na amostragem para ajuste de equações para dados em série temporal, aconselha-se considerar a amostra como pequena, quando o número de observações é menor do que 200 vezes o número de parâmetros da equação ajustada; assim, quando a amostra for considerada grande (≥ 200 vezes o número de parâmetros), a correção será muito pequena, da ordem de 1% para equações com três parâmetros, e será pouco significativa, mas pode ser aplicada.

6. O modelo com menor *AIC* ou AIC_c é o mais próximo de ser o correto; somente se pode comparar um dos dois, ou o AIC simples ou o corrigido (AIC_c), nunca o *AIC* de uma com o AIC_c de outra equação.

Ainda, segundo Motulsky e Christopoulos (2003), somente devem ser comparados modelos que se ajustam bem aos dados, devendo-se antes eliminar todos os modelos que não apresentam bons resultados, ficando-se com os dois ou três melhores para a comparação final por meio de ferramentas estatísticas como o *AIC*.

O teste utilizado pelo SAS System, para comparação de equações lineares aninhadas, é o C_p de Mallows, permitindo que se reduza o número de variáveis independentes. O SAS System calcula o *AIC* para as funções lineares, entre outros índices, com o procedimento PROC REG e, das não-lineares, pelo procedimento PROC NLMIXED, possibilitando compará-las entre si dentro de cada grupo (SAS Institute, 1999b).

Outra forma de estimar alguns dos critérios de comparação e determinar os melhores modelos entre os testados é com o uso de procedimentos auto-regressivos de seleção de equações. O modelo de erro auto-regressivo corrige a correlação serial. O procedimento AUTOREG do SAS System pode ajustar o erro auto-regressivo de modelos lineares de qualquer ordem e pode ajustar subconjuntos de modelos auto-regressivos. É possível, também, especificar a auto-regressão stepwise para selecionar o modelo de erro auto-regressivo automaticamente. É comum que séries de dados apresentem regressores dependentes encapsulados (*lag regressors*); no procedimento AUTOREG, pode-se especificar até 12 regressores encapsulados; nesse caso, o procedimento PROC AUTOREG executa o t-teste de Durbin e o h-teste de Durbin para auto-correlação de primeira ordem e relata seus níveis de significância marginais; outro procedimento útil para o caso de regressões não-lineares é o PROC NLMIXED. (SAS Institute, 1999[a])

Os critérios *AIC* (e.6) e *BIC* (e.7), para um dado modelo, são definidos como (Myung *et al.*, 2003):

$$AIC = -2 \cdot ln\,[f(y|w^*)] + 2\,k \qquad \text{(e.6)}$$

$$BIC = -2 \cdot ln\,[f(y|w^*)] + k \cdot ln\,(n) \qquad \text{(e.7)}$$

Em que: w^* = estimativa de máxima verossimilhança (maximum likelihood estimation – *MLE*); ln = logaritmo natural de base *e*; k = número de parâmetros; n = número de elementos da amostra.

Para erros com distribuição normal e variância constante, o primeiro termo de ambos os critérios, $-2 \cdot ln\, f(y|w^*)$, é reduzido para $(n \cdot ln(SQE(w^*)) + c_o)$; em que c_o é uma constante que não depende do modelo. Em cada critério, o primeiro termo representa a medida da falta de ajustamento, o segundo representa a medição da complexidade e juntos eles representam a medida da falta de **generabilidade**. Quanto menor o valor encontrado para o critério, melhor a generabilidade; o modelo que minimiza um dos dois critérios tem preferência na escolha.

Os critérios *AIC* (e.8) e *SBC* (ou *BIC)* (e.9) são calculados pelo SAS System pelas seguintes equações (SAS Institute, 1999b):

$$AIC = -2 \cdot ln\,(L) + 2\,k \qquad \text{(e.8)}$$

$$SBC = -2 \cdot ln\,(L) + ln\,(n) \cdot k \qquad \text{(e.9)}$$

Em que: L = função de verossimilhança; n = número de resíduos que podem ser computados para a série de dados temporais; k = número de parâmetros livres; *ln* = logaritmo de base neperiana.

O coeficiente (e.10) de determinação é calculado como (Bussab, 1986):

$$R^2 = SQ_{reg} / SQ_{total} \qquad \text{(e.10)}$$

Em que: R^2 = coeficiente de determinação; SQ_{reg} = soma de quadrados da regressão; SQ_{total} = soma de quadrados totais.

O coeficiente de determinação ajustado (e.11) é calculado como (SAS Institute, 1999b):

$$R^2\,aj. = 1 - [(n-i)\,(1-R^2) / (n-p)] \qquad \text{(e.11)}$$

Em que: $R^2\ aj.$ = coeficiente de determinação ajustado; n = número de observações da amostra; i = indicador que assume o valor 1 (um) se o modelo possui intercepto e, se não possui, assume valor 0 (zero); p = número de parâmetros do modelo; R^2 = coeficiente de determinação.

A estatística C_p de Mallows é delineada contra o número de parâmetros (p); quanto mais próximo for C_p de p, menos tendenciosas são as estimativas dos parâmetros e melhor é o modelo. A estatística C_p de Mallows é dada por (e.12) (SAS Institute, 1999b):

$$C_p = [(SQE_p) / (s^2)] - (N - 2p) \quad \text{(e.12)}$$

Em que: p = número de parâmetros do modelo; s^2 = quadrado médio do erro para o modelo completo; SQE_p = soma de quadrados do erro para o modelo com p parâmetros, incluindo o intercepto se houver.

O erro-padrão de estimativa (S_{yx}) (e.13) pode ser expresso em percentagem em relação à média das observações da variável dependente ($\overline{y}$), podendo então ser chamado de coeficiente de variação do modelo ($CV\%$) (e.14), sendo o erro-padrão de estimativa dado pela raiz quadrada do quadrado médio dos resíduos (QM_{res}):

$$Syx = \sqrt{QMres} \quad \text{(e.13)}$$

$$CV\% = 100 \,.\, S_{yx} / \overline{y} \quad \text{(e.14)}$$

O teste Qui-Quadrado (χ^2) é a razão entre duas variâncias, sendo usado para verificar a aderência entre duas séries de dados; quanto mais próximo de 1, maior a aderência (Wonnacott e Wonnacott, 1980). O χ^2 é dado por (e.15):

$$\chi^2 = \frac{S^2}{\delta^2} \quad \text{(e.15)}$$

Em que: χ^2 = Qui-Quadrado; S^2 = variância da amostra; δ^2 = variância da população.

Sit (1994) recomenda que a comparação das equações logarítmicas com funções lineares normais e com funções não-lineares seja realizada pela variável dendrométrica estimada e não por meio da variável dependente transformada. O motivo é que as variáveis transformadas resultam em proporções diferentes quando se calculam as estatísticas, sendo válidas para uso dos testes F e t e para comparação entre modelos de mesmo tipo, mas não para comparação entre modelos de tipos diferentes.

A média é igual à razão da soma dos valores de um conjunto de dados e a quantidade de elementos do conjunto; quando os dados são transformados, como no caso de linearização por logaritmos, altera-se a estrutura do modelo matemático que expressa a média das observações, ou seja, a estrutura da média. Para comparação de modelos matemáticos, é necessário que as variáveis dependentes sejam de mesmo tipo e dimensão, com idêntica estrutura de médias (Zimmermann e Núñez-Antón, 2001).

Portanto, o QM_{res} das equações logarítmicas (e.16) deve ser calculado partindo da extração do antilogaritmo da variável dependente estimada e da observada, da seguinte forma (Sit, 1994):

$$QM_{res} = \sum (\ h_i - \hat{h}_i\ )^2 / GL_{res} \quad \text{(e.16)}$$

Em que: QM_{res} = quadrado médio dos resíduos; $h_i = e^{y_i}$ = altura observada; $\hat{h}_i = e^{\hat{y}_i}$ = altura estimada; em que: y_i = valor observado transformado; $\hat{y}_i$ = valor estimado pela equação logarítmica; e = base do logaritmo neperiano.

Entre os modelos utilizados para descrever a altura de árvores estão as equações polinomiais, as polinomiais inversas, a função logística, as exponenciais e potenciais, a equação de Chapman-Richards (e.17), a de Gompertz, Backman, Prodan, Weibull, entre outras citadas na literatura especializada. Dentre estas, o modelo mais compreensível e explicável em termos biológicos é o de Chapman-Richards que é descrito, de acordo com Prodan *et al*. (1997), como segue:

$$y = A\left\{\left[1 - e^{(-k.t)}\right]^{r}\right\} \qquad \text{(e.17)}$$

Em que: A = Assíntota superior, ou valor máximo que a variável dependente pode assumir; k = velocidade de crescimento, variando de *0* até *1*; r = ponto de inflexão da curva, entre o mínimo valor de y e a assíntota superior; y = variável dependente; t = variável independente (tempo).

MATERIAL E MÉTODOS

Neste trabalho, utilizou-se, como exemplo, os dados reais, de altura de árvores, provenientes de um experimento. Foram ajustados 11 modelos, sendo três lineares, três logarítmicos, dois não-lineares linearizados e três não-lineares. As estatísticas de cada modelo foram calculadas, sendo selecionado o melhor modelo de cada grupo pelas estatísticas normais (R^2, *QMres*, $C_{(p)}$, AIC, BIC) e, então, selecionado o melhor modelo geral por meio das estatísticas das estimativas de altura em metros, obtidas por meio de cada modelo para cada indivíduo da amostra. Como último passo, foi selecionado o verdadeiro melhor modelo por meio do teste Qui-Quadrado e da análise gráfica dos resíduos das estimativas de altura em metros, obtidas por cada modelo pré-selecionado, para cada indivíduo da população.

Dados

Os dados são provenientes de 16 parcelas com 70 árvores, originadas de um povoamento de *Pinus elliottii*, plantado em 1985, em Piratini, RS, com espaçamento de 3 m x 2 m (6 m² de área por árvore). Em 2000, foi realizado um desbaste sistemático da 6ª linha de plantio na eqüidistância de 3 metros, complementado por um desbaste por baixo nas linhas intermediárias, numa intensidade total aproximada de 40% do número de árvores, restando 531 árvores aos 18 anos. Foram medidas e separadas 64 árvores, quatro por parcela, para o presente trabalho.

Modelos

As equações utilizadas para modelagem dos dados da altura das 64 árvores de *Pinus elliottii*, considerando-se o período de 6 até os 18 anos idade, são as relacionadas na Tabela 1.

TABELA 1: Modelos de equações testados para descrição do crescimento em altura de 48 árvores de *Pinus elliottii*.

TABLE 1: Tested equations models for height growth description of 64 *Pinus elliottii* trees.

N. da equação	Equação	Autor	Tipo/Grupo
1	$h = b_0+b_1.ln\ t+b_2.ln^2\ t$	Backman [a]	
2	$h = b_0+b_1.1/t+b_2.t+b_3.t^2+b_4.ln\ t+b_5.ln^2\ t$	- [a]	Lineares
3	$h = b_0 + b_1.t + b_2.t^2 + b_3.t^3 + b_4.t^4 + b_5.t^5$	- [a]	
4	$ln\ h = b_0+b_1.ln\ t+b_2.ln^2\ t$	-	
5	$ln\ h = b_0+b_1.1/t+b_2.t+b_3.t^2+b_4.ln\ t+b_5.ln^2\ t$	-	Logarítmicas
6	$ln\ h = ln\ b_0 + b_1\ ln\ t + b_2\ t$	- [a]	
7	$1/(h-1,3) = b_0+b_1.1/t+b_2.1/t^2$	- [d]	Não-lineares
8	$t^2/h = (b_0+b_1.t+b_2.t^2)$	Prodan [c]	linearizadas
9	$h = A\left\{\left[1-e^{(-k.t)}\right]^{r}\right\}$	Chapman-Richards [b]	
10	$h = b_0\ .\ e^{e^{(b_1\ -\ b_2\ .t)}}$	Gompertz [a]	Não-lineares
11	$h = b_0\ t^{b1}\ e^{b2\ t}$	- [a]	

Em que: h = variável dependente (altura em metros); t = tempo em anos; ln = logaritmo neperiano; e = base do logaritmo neperiano; b_0, b_1, b_2, b_3, b_4, b_5 = parâmetros das equações; A, k, r = parâmetros da equação de Chapman-Richards. Fontes: (a) Sit (1994); (b) Prodan (1997) ; (c) Scheeren (2003); (d) Finger (1992).

Ajustamento dos modelos

O ajustamento das equações foi realizado com o SAS System, versão 8.

Inicialmente, cada grupo de modelos foi ajustado separadamente para a primeira seleção. Os modelos lineares foram ajustados com o procedimento PROC REG com a opção de seleção FORWARD, pelo Programa 1, em anexo, usado para selecionar as variáveis significativas, mantendo-se somente as recomendadas pelo programa, que usa o F a 10% de probabilidade para determinar que variáveis

permanecem no modelo e as estatísticas R^2 e C_p para determinar a melhor composição de variáveis. O mesmo procedimento foi utilizado para seleção das variáveis significativas dos modelos logarítmicos e dos não-lineares linearizados. Os modelos não-lineares foram ajustados pelo procedimento PROC NLIN e as suas estatísticas complementares, R^2 e QM_{res}, foram obtidas com o procedimento PROC MODEL.

A escolha do melhor modelo por grupo foi realizada usando as estatísticas *AIC* e *BIC*, obtidas com o Programa 2, com os procedimentos PROC REG para modelos lineares, logaritmicos e linearizados e PROC NLMIXED para os modelos não-lineares.

Comparação entre modelos

Seleção dentro dos grupos de equações

Numa primeira etapa de seleção, optou-se pelo melhor modelo de cada grupo de mesma estrutura de médias, por meio das estatísticas obtidas pelos relatórios emitidos pelo Programa 1: R², QM_{res}, *BIC*, *AIC* e gráfico de resíduos, calculados sobre a variável dependente de cada equação.

Na segunda etapa da seleção, foram estimados os erros dos melhores modelos dos tipos lineares e linearizados por meio das alturas observadas e estimadas com as equações. Primeiro, os resíduos das equações foram delineados em gráfico e avaliados e, então, calculadas as estatísticas R^2 e QM_{res}, *BIC* e *AIC*, sendo a decisão tomada com base nessas estatísticas, considerando-se o modelo que atendeu ao maior número de critérios. A seguir, foram processados os melhores modelos para obter a soma de quadrados dos resíduos em relação a diferença entre os valores observados e estimados para a altura em metros, usando-se o Programa 3, em anexo.

Quando as variáveis dependentes dentro de um mesmo grupo apresentam estruturas diferentes de médias, como é o caso dos modelos não-lineares linearizados utilizados, foi necessário executar o procedimento descrito por Sit (1994) para estimar o R^2 e o QM_{res} e calcular manualmente o *AIC* com as equações recomendadas por Motulsky e Christopoulos (2003).

A escolha do melhor modelo não-linear foi realizada com base no QM_{res} calculado pelo procedimento PROC NLIN e nos critérios *AIC* e *BIC* calculados pelo procedimento PROC NLMIXED do SAS.

Seleção entre grupos de equações

Quando a comparação direta pelas estatísticas emitidas pelos relatórios normais do SAS System não foi possível, procedeu-se às aproximações necessárias. No caso dos modelos logarítmicos e dos modelos não-lineares linearizados, as somas de quadrados foram calculadas pelas diferenças entre as alturas estimadas pelas equações e alturas observadas em metros, determinando-se o R^2, o *QMres* e o *AIC*.

O cálculo do *AIC* para comparação entre equações de grupos diferentes foi realizado de forma manual, usando-se a equação (e.5) recomendada por Motulsky e Christopoulos (2003):

$$AIC_c = AIC + 2K(K-1)/(n-K-1) \quad \text{(e.5)}$$

Na última etapa da seleção, foram realizadas as estimativas para a população com os melhores modelos e compararam-se os resultados obtidos. O melhor modelo, por meio da amostra, foi considerado o que atendeu ao maior número de critérios.

O melhor modelo foi encontrado considerando-se os critérios qualitativos de interpretabilidade e compreensibilidade e os critérios quantitativos de falseabilidade (verificada pela análise gráfica da distribuição de resíduos), qualidade do ajustamento (pelo R^2 e pelo QM_{res}), complexidade (número de parâmetros e tipo de modelo) e sua generabilidade; essa última característica sendo considerada a estatística mais importante e decisiva para a seleção de modelos, determinada por meio do *AIC,* em relação à amostra.

Seleção do verdadeiro melhor modelo

O verdadeiro melhor modelo foi considerado o que apresentou maior generalidade, determinada pelo Qui-Quadrado calculado entre as variâncias dos valores observados da população e dos valores estimados por meio de cada equação, sendo tanto melhor quanto mais o valor se aproxima de 1, ou seja, quanto mais a variância das estimativas se aproxima da variância das observações da população.

RESULTADOS E DISCUSSÃO

Ajuste dos modelos e cálculos estatísticos

Os modelos 1 a 8, ajustados pela opção de seleção FORWARD da declaração MODEL do procedimento PROC REG, reduzindo o número de variáveis independentes em decorrência do nível especificado de significância de 95% de probabilidade e os modelos 9 a 11 foram ajustados com os procedimentos PROC NLIN e PROC MODEL, e os resultados estão relacionados na Tabela 8.

O resultado do ajuste mostra que os grupos de equações com variável dependente h, dos tipos linear e não-linear apresentaram os maiores R^2, entretanto não se pode utilizar esses valores para comparar os modelos que apresentam médias com estruturas diferentes (Zimmermann e Núñez-Antón, 2001). Num processo de seleção tradicional, poderiam ser escolhidos os modelos logaritimicos, em detrimento dos demais em razão do menor erro-padrão de estimativa, mas isso poderia não estar correto.

TABELA 2: Equações para descrição do crescimento em altura de 48 árvores de *Pinus elliottii*, entre os 6 e 18 anos de idade.

TABLE 2: Equations for height growth description of 48 *Pinus elliottii* trees, from ages 6 to 18.

N. eq.	Variável dependente	Variáveis independentes retidas e parâmetros estimados					$R^2aj.$	$CV\%$	C_p	QM_{res}	Tipo/Grupo
		b_0	b_1		b_2						
		Parâmetro	Parâmetro	variável	Parâmetro	variável					
1	h	-22,78189	29,12430	ln t	-39,22331	$\ln^2 t$	0,8712	11,77	3,0	2,5545	
2	h	-2,41021	1,60455	t	-0,02066	t^2	0,8714	11,76	4,3	2,5521	Lineares
3	h	-0,96924	1,25766	t	-0,00000106	t^5	0,8715	11,76	2,7	2,5486	
4	ln h	0,48306	-0,22425	ln t	2,96536	$\ln^2 t$	0,8280	6,34	3,0	0,0259	
5	ln h	0,91282	-1,75963	1/t	2,04204	$\ln^2 t$	0,8280	6,34	2,8	0,0259	Logarítmicas
6	ln h	-0,74300	1,59857	ln t	-0,05011	t	0,8278	6,34	3,0	0,0260	
7	1/(h-1,3)	0,06554	-0,72490	1/t	10,47187	$1/t^2$	0,6100	41,07	3,0	0,0017	Não-lineares
8	t^2/h	3,66642	0,26735	t	0,02488	t^2	0,8539	12,44	3,0	1,8077	linearizadas
9	h	34,10967	0,04818	-	1,18267	-	0,8714	11,76	-	2,5520	
10	h	26,93389	0,95270	-	0,11504	-	0,8713	11,77	-	2,5531	Não-lineares
11	h	0,65857	1,39886	-	-0,03570	-	0,8714	11,76	-	2,5515	

Em que: h = variável dependente (altura em metros); t = tempo em anos; ln = logaritmo neperiano; e = base do logaritmo neperiano; b_0, b_1, b_2 = parâmetros das equações; nas equações não-lineares a variável independente (t) associa-se a mais de um parâmetro; $R^2aj.$= coeficiente de determinação ajustado; $CV\%$ = coeficiente de variação em percentagem; C_p = critério C_p de Mallows; QM_{res}= quadrado médio dos resíduos.

Nessa fase, foram selecionadas as variáveis significativas de cada modelo por meio do teste F da análise de variância da regressão, e determinadas às estatísticas para a equação final de cada modelo original.

O menor coeficiente de variação é o das equações logarítmicas, entretanto os coeficientes de determinação são menores que nas demais, exceto que o da equação 7. Por outro lado, a estrutura das suas médias são diferentes dos demais grupos, e não é possível a comparação direta.

As equações lineares podem ser comparadas entre si, equações 1 a 3, pois têm mesma estrutura de médias. Isso também é possível com as logarítmicas, equações 4 a 6, mas essas não podem ser comparadas diretamente com outras equações não-logarítmicas por que têm variável dependente com estrutura de médias própria. Já, as duas equações não-lineares linearizadas, equações 7 e 8, não têm a mesma variável dependente e, portanto, têm diferentes estruturas de médias e não podem ser comparadas diretamente entre si ou com as outras equações, sendo necessário antes estimar as alturas em metros e depois calcular a soma de quadrados dos resíduos de cada equação para depois estimar as demais estatísticas.

As equações não-lineares apresentam a mesma variável dependente e podem ser comparadas diretamente entre si pelo R^2 e QM_{res}; podem também ser comparadas diretamente com as lineares com a mesma variável dependente pelo AIC, pois a estrutura de médias para a variável dependente é a mesma (variável idêntica de mesma dimensão e mesma escala ou mesma proporção de variação), mas é necessário

determinar o *AIC* pelo procedimento recomendado por Motulsky e Christopoulos (2003) para a comparação final com os demais grupos, pois o *AIC* calculado pelos mínimos quadrados ordinários das equações lineares é diferente do *AIC* calculado pelos mínimos quadrados generalizados das não-lineares.

Entre os modelos lineares, equações 1 a 3, nenhum se destaca, e a decisão então é tomada pelo *AIC* apresentado na Tabela 3, em que a equação 3, com valor de 781,36 minimiza o *AIC* e é escolhida como a melhor das três.

Dos modelos logarítmicos, 4 a 6, o modelo 4 minimiza o *AIC* com valor de -3035,26 e AIC estimado sobre h de 784,56, sendo o escolhido.

Os modelos não-lineares linearizados são comparados com o *AIC* estimado pelo procedimento recomendado por Motulsky e Christopoulos (2003), sendo que o modelo 8 com *AIC* de 806,46 é considerado o melhor dos dois.

Verifica-se, pela Tabela 2, que as três equações não-lineares (9, 10 e 11) apresentam resultados muito semelhantes até mesmo quanto aos resíduos delineados nas Figuras 1-I, 1-J e 1-K. O *AIC, R^2 e CV* dos modelos 9 e 11 são idênticos e menores do que da equação 10. Embora a equação 11 apresente menor QM_{res} entre as duas, a decisão da seleção recai sobre o modelo 9, pois a diferença nos valores das estatísticas é mínima, e o modelo de Chapman-Richards (n. 9) apresenta maior interpretabilidade e compreensibilidade.

Os resíduos das equações 1 a 11 são delineados nas Figuras 1-A a 1-K respectivamente com as variáveis utilizadas na modelagem.

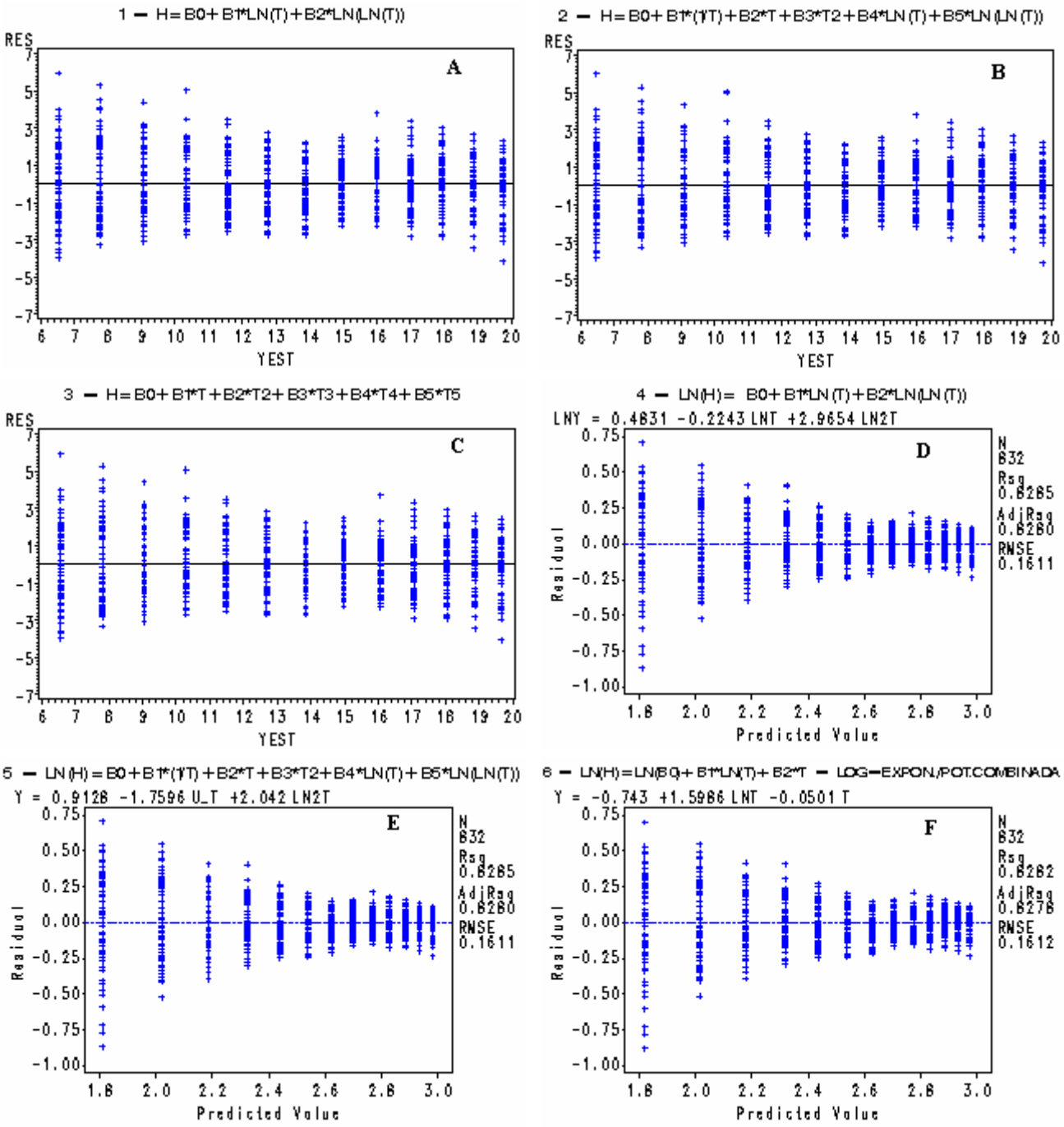
Para comparação dos resíduos entre grupos, é necessário utilizar a diferença entre altura observada e estimada ($h - \hat{h}$), como são apresentados nas Figuras 1-L a 1-P para os modelos logaritmicos, não sendo possível usar as Figuras 1-D a 1-H, pois são baseadas na diferença entre variáveis transformadas ($y - \hat{y}$). Como é possível ver pela própria forma que os resíduos assumem nos dois tipos de gráfico, baseados em variáveis de diferentes dimensões, reforçam as recomendações de autores como Sit (1994) e Collett (2003) sobre comparação de modelos com diferentes estruturas de médias.

As equações lineares (Figuras 1-A a 1-C) e as não-lineares (Figuras 1-I a 1-K) apresentam resíduos positivos e negativos razoavelmente equilibrados e com distribuição semelhante entre si, abrangendo todo o eixo das estimativas de forma semelhante.

Os modelos logarítmicos, Figuras 1-D a 1-F, apresentam distribuições e amplitude de resíduos semelhantes entre si, mas com uma distribuição pior do que as anteriores.

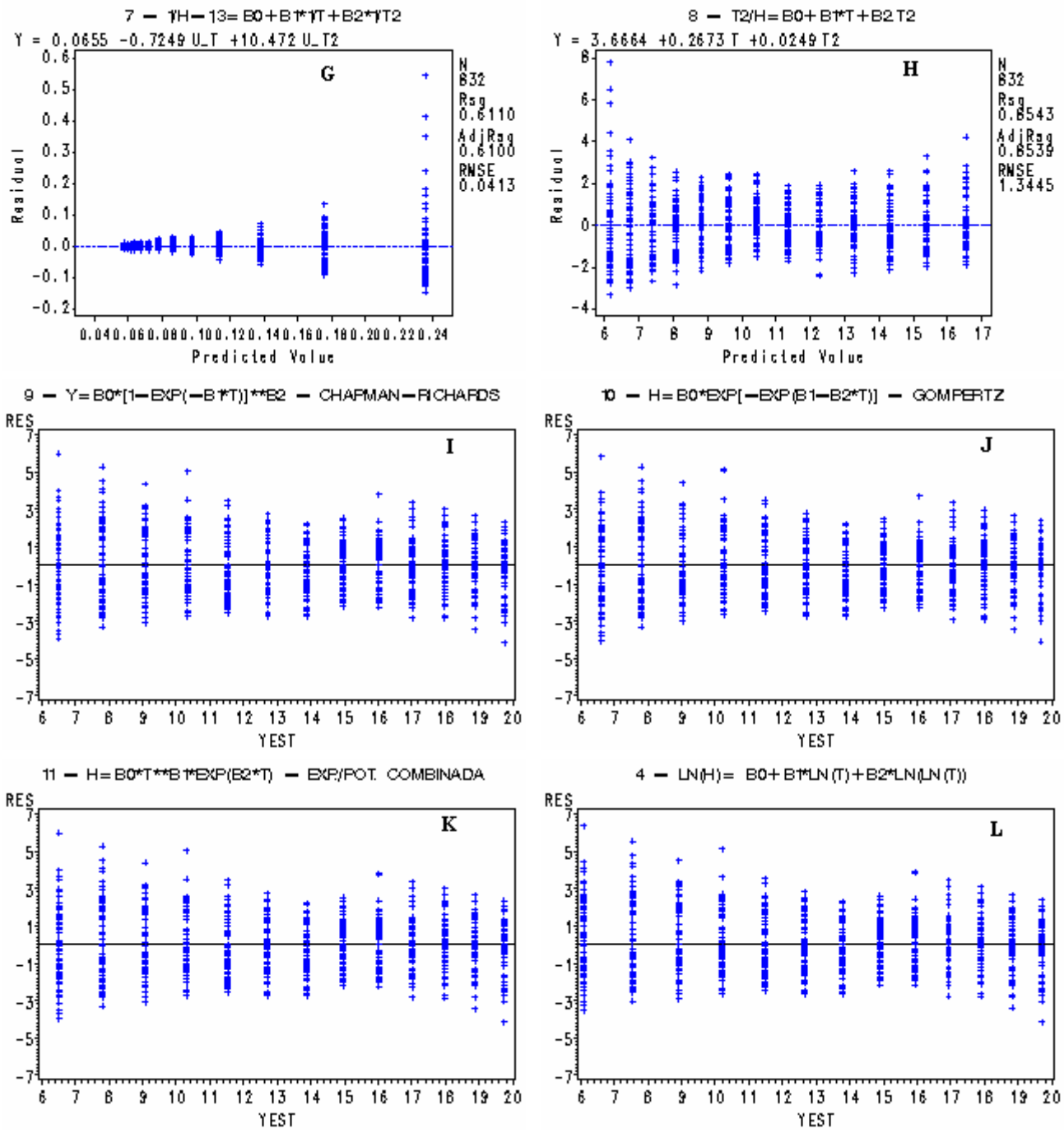
Já, os modelos não-lineares linearizados têm diferentes distribuições e amplitudes de resíduos entre si e entre os demais modelos, pelo fato de que cada um ter uma variável dependente com sua própria estrutura. A equação 7, Figuras 1-G e 1-O, e a equação 8, Figuras 1-H e 2-B, apresentam resíduos desequilibrados entre positivos e negativos, sendo que a equação 7 ainda tem uma amplitude mais restrita em relação ao eixo das estimativas.

Em todos os gráficos de resíduos, observa-se a formação de padrões, indicando tendências de heterocedasticidade da variância, característica inerente às séries temporais de dados, o que é determinante na utilização de critérios como o *AIC* e o *BIC* para comparação de modelos, em detrimento dos critérios tradicionais como o R^2 e QM_{res}.


1 — H=B0+B1*LN(T)+B2*LN(LN(T))
RES
A
YEST
2 — H=B0+B1*(1/T)+B2*T+B3*T2+B4*LN(T)+B5*LN(LN(T))
RES
B
YEST
3 — H=B0+B1*T+B2*T2+B3*T3+B4*T4+B5*T5
RES
C
YEST
4 — LN(H)= B0+B1*LN(T)+B2*LN(LN(T))
LNY = 0.4831 -0.2243 LNT +2.9654 LN2T
D
N 832
Rsq 0.8285
AdjRsq 0.8280
RMSE 0.1611
Residual
Predicted Value
5 — LN(H)=B0+B1*(1/T)+B2*T+B3*T2+B4*LN(T)+B5*LN(LN(T))
Y = 0.9128 -1.7596 U_T +2.042 LN2T
E
N 832
Rsq 0.8285
AdjRsq 0.8280
RMSE 0.1611
Residual
Predicted Value
6 — LN(H)=LN(B0)+B1*LN(T)+B2*T — LOG=EXPON./POT.COMBINADA
Y = -0.743 +1.5986 LNT -0.0501 T
F
N 832
Rsq 0.8282
AdjRsq 0.8278
RMSE 0.1612
Residual
Predicted Value

Continua …

Continuação …



Continua …

Continuação …

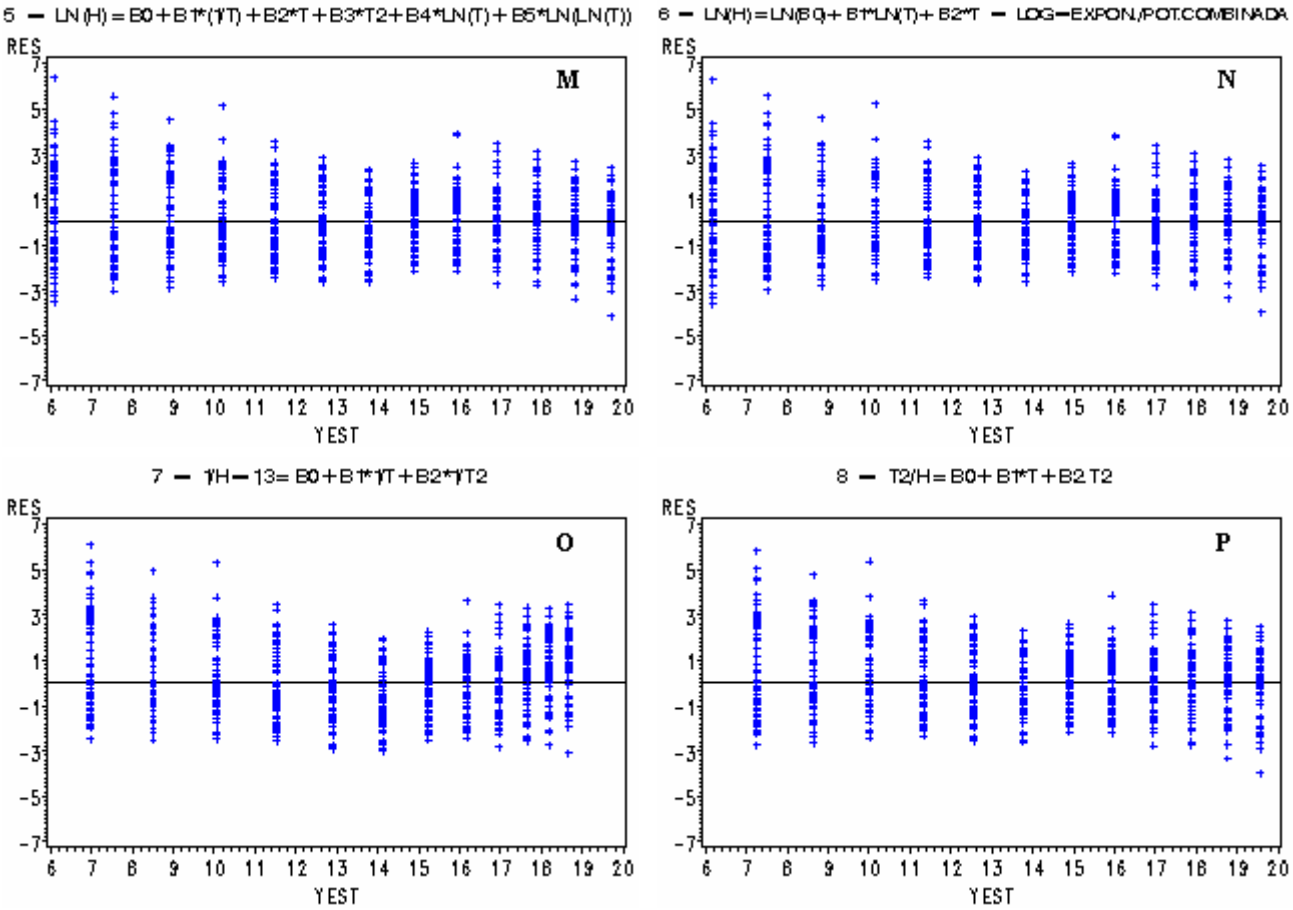


FIGURA 1: Resíduos das equações 1 a 16.
FIGURE 1: Residues of equation 1 and 16.

Comparação e seleção dos melhores modelos

As aproximações finais das estatísticas, para possibilitar a comparação entre os modelos com estruturas de médias de variável dependente diferentes entre si, ou que foram calculados por diferentes métodos de mínimos quadrados, são apresentados na Tabela 3. Verifica-se que os coeficientes de variação das equações logarítmicas (4, 5 e 6) aumentam de valor em relação aos valores da Tabela 2 e, quando estimados por meio das alturas calculadas em metros pela extração do antilogaritmo das variáveis estimadas, são maiores que os das equações lineares (1 a 3) e que das equações não-lineares (9 a 11).

Não é possível comparar diretamente as equações 7 e 8, pois apresentam variáveis dependentes de diferentes dimensões; nesse caso, deve-se adotar, antes, o procedimento de cálculo do QM_{res} recomendado por Sit (1994) e comparar as equações pelo *AIC* estimado sobre h pelo procedimento recomendado por Motulsky e Christopoulos (2003), resultando no modelo 8 como o melhor; o mesmo procedimento foi adotado para comparar as equações logarítmicas com as não-logarítmicas e lineares com não-lineares.

Os testes de diferença entre AIC_c foram significativos para os seguintes pares de equações: 4 e 1; 4 e 2; 4 e 3; 4 e 9; 4 e 10; 4 e 11; 5 e 1; 5 e 2; 5 e 3; 5 e 9; 5 e 10; 5 e 11; 6 e 1; 6 e 2; 6 e 3; 6 e 9; 6 e 10; 6 e 11; 7 e 1; 7 e 2; 7 e 3; 7 e 4; 7 e 5; 7 e 6; 7 e 8; 7 e 9; 7 e 10; 7 e 11; 8 e 1; 8 e 2; 8 e 3; 8 e 4; 8 e 5; 8 e 6; 8 e 9; 8 e 10 e 8 e 11. Resumidamente: as melhores equações foram as lineares típicas (1, 2 e 3) e as não-lineares (9, 10 e 11); as equações logaritmicas (4, 5 e 6) foram piores do que as lineares (1, 2 e 3) e do que as não-lineares (9, 10 e 11); as equações não-lineares linearizadas (7 e 8) foram piores do que todas as demais.

Observa-se que as equações 2, 3, 9 e 11, com menores coeficientes de variação, calculados sobre os dados de altura em metros (CV% = 11,76), são também as que apresentam os menores AIC (de 775,37 a 776,53).

Os melhores modelos de cada grupo (equações 3, 4, 8 e 9) são comparados entre si pelo *AIC* estimado sobre h, que é calculado pela mesma fórmula e mesma variável dependente, sendo que os modelos 3 e 9 apresentam semelhantes e menores *AIC* do que os demais; a decisão, então, deve ficar entre um dos dois, o que atende à orientação de reduzir a comparação final a dois, ou no máximo três modelos, de acordo com Motulsky e Christopoulos (2003).

O teste de Durbin-Watson confirma a análise visual da distribuição de resíduos, indicando a existência de correlação serial entre os erros das equações logarítmicas e das não-lineares linearizadas, inviabilizando o seu uso (Tabela 3), pois o teste mostrou que esses modelos podem gerar estimativas tendenciosas.

TABELA 3: Estatísticas para seleção de modelos de equações para descrição do crescimento em altura de 48 árvores de *Pinus elliottii*, dos 6 aos 18 anos de idade.

TABLE 3: Equations models estatistics for height growth description of 48 *Pinus elliottii* trees, from 6 to 18 years-old.

N. eq.	Variável dependente	*AIC*	*BIC*	QM_{res} sobre y	d (Durbin-Watson)	Variáveis independentes	SQ_{res} sobre *h*	AIC_c estimado sobre *h*	*CV%* sobre h	*k*
1	h	783,3	785,2	2,4832	1,92 n.s.	ln t, ln² t	2117,7	777,31	11,77	3
2	h	782,5	784,5	2,4766	1,92 n.s.	t, t²	2115,7	776,53	11,76	3
3	h	781,4	783,4	2,4792	1,92 n.s.	t, t^5	2112,8	775,37	11,76	3
4	ln h	-3035,3	-3033,2	0,0250	1,81 s.	ln t, ln² t	2136,2	784,56	11,82	3
5	ln h	-3035,3	-3033,2	0,0250	1,81 s.	1/t, ln² t	2136,3	784,59	11,82	3
6	ln h	-3034,0	-3031,9	0,0251	1,81 s.	ln t, t	2136,3	784,57	11,82	3
7	1/(h-1,3)	-5300,9	-5298,9	0,0018	1,70 s.	1/t, 1/t²	2373,1	872,05	12,46	3
8	t²/h	495,6	497,6	1,6356	1,82 s.	t, t²	2193,2	806,46	11,98	3
9	h	2292,7	2311,5	2,4796	1,92 n.s.	t	2115,6	776,50	11,76	3
10	h	2295,3	2314,2	2,4875	1,92 n.s.	t	2116,5	776,86	11,77	3
11	h	2292,7	2311,6	2,4798	1,92 n.s.	t	2115,2	776,35	11,76	3

Em que: N. eq. = número da equação; *AIC* = critério de informação de Akaike; AIC_c estimado sobre *h* = critério de informação de Akaike estimado pelo procedimento recomendado por Motulsky e Christopoulos (2003) calculado através das alturas observadas e estimadas em metros; QM_{res} sobre *y* = quadrado médio dos resíduos calculado pelos procedimentos tradicionais para a variável dependente (*y*) da equação; *d* = estatística *d* de Durbin-Watson, em que "s." indica significância e "n.s." indica não-significância ao nível de 5% de probabilidade; SQ_{res} sobre *h* = soma de quadrados dos resíduos calculados com as alturas estimadas e observadas; *CV%* sobre *h* = coeficiente de variação calculado com as alturas estimadas e observadas; *h* = altura em metros; *t* = idade em anos; *ln* = base do logaritmo neperiano; *k* = número de parâmetros da equação (*p*), nas não-lineares sem intercepto adiciona-se 1 ao valor de *p*.

Nas Figuras 2-A e 2-B, estão representadas as observações de altura das árvores da amostra e as estimativas realizadas com as equações n. 3 e 9 respectivamente em que se constata que as linhas geradas pelas equações são quase-idênticas, confirmando os resultados estatísticos semelhantes em relação aos dois modelos.

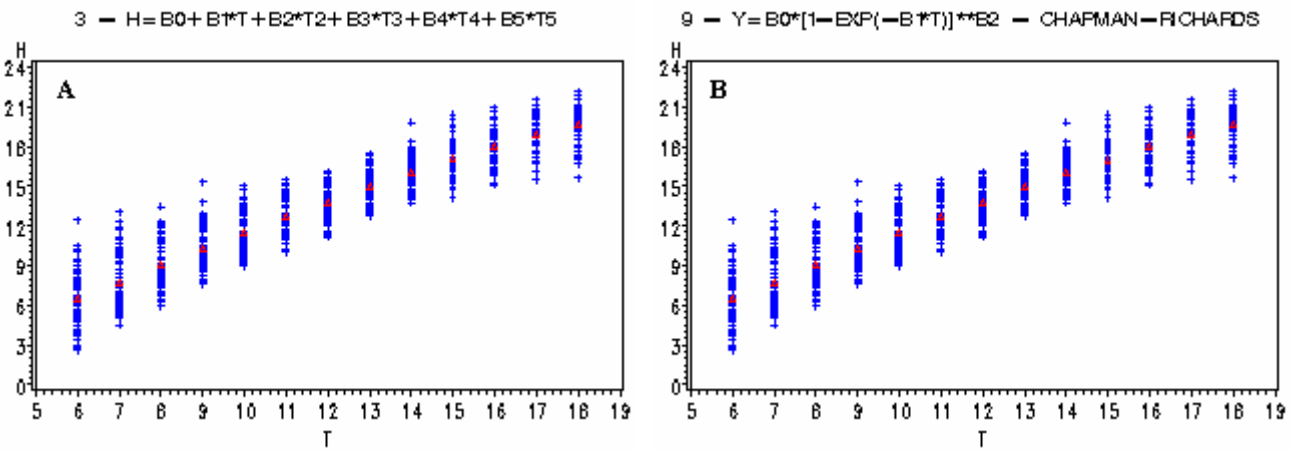


FIGURA 2: Observações (+) e estimativas (•) de altura (m) realizadas com as equações n. 3 e 9.
FIGURE 2: Observations (+) and estimates (•) of height (m) accomplished with equation n. 3 and 9.

Considerando-se a falseabilidade, verifica-se que a qualidade dos dois melhores modelos (equações 3 e 9) é muito próxima, pois os gráficos de resíduos são praticamente iguais. A complexidade pode ser considerada um pouco menor na equação 3, pois ambas têm três parâmetros, mas a equação 3 é linear, enquanto a 9 é não-linear e, portanto, mais complexa. O ajustamento é um pouco melhor no modelo 3, pois apresenta menor soma de quadrados dos resíduos que o modelo 9. As duas equações têm coeficientes de determinação (R^2) iguais e quadrados médios dos resíduos (QM_{res}) muito próximos. Há, portanto, uma leve tendência para a escolha do modelo n. 3. Entretanto, quando se aplicam os critérios qualitativos de interpretabilidade e compreensibilidade, o modelo de Chapman-Richards (equação 9) leva vantagem, até mesmo pelo amplo uso que tem tido na descrição do crescimento de árvores, possibilitando comparação com outros estudos. Até esse passo, os dois melhores modelos podem ser considerados empatados.

Seleção do verdadeiro melhor modelo

A definição do verdadeiro melhor modelo é realizada pela generalização verdadeira e não pelos demais critérios.

Nessa última etapa de seleção, foram estimadas as variâncias das estimativas feitas para cada indivíduo da população, pelas duas melhores equações e foi calculada a variância real da população. Então, foi realizado o cálculo do Qui-Quadrado e determinada qual é a equação que produz o melhor ajuste aos dados da população.

Os cálculos para computar as variâncias levaram em consideração apenas as 531 árvores remanescentes aos 18 anos, nos anos em que foram medidas em pé (aos 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 e 18 anos), resultando num total de 5.841 observações. Para tanto, utilizaram-se o procedimento PROC MEANS do SAS System e o Programa 4, em anexo. Os resultados são apresentados na Tabela 4, em que se observa que a equação n. 9 apresenta o maior Qui-Quadrado, aproximando-se mais dos dados populacionais, embora sem diferença significativa da equação 3, pois apresentam diferença proporcional de apenas 0,0031 em relação a esse critério.

Assim, verifica-se que os modelos 3 e 9 não apresentam diferenças estatisticamente significativas quanto aos critérios quantitativos. Mas, os critérios qualitativos, em especial de ligação do modelo com o processo estudado, sua interpretabilidade e compreensibilidade levam à escolha final do modelo de Chapman-Richards (equação 9).

TABELA 4: Variância da população e das estimativas e Qui-Quadrado das estimativas em relação a uma população de 531 árvores de *Pinus elliottii*.

TABLE 4: Population and predictions' variance, and predictions' Chi-square (χ^2) related to a population of 531 *Pinus elliottii* trees.

Equação n.	Variável considerada	N	Altura Média (m)	Variância $(\delta_h^2, S_{\hat{h}}^2)$	Altura Mínima (m)	Altura Máxima (m)	$\chi^2 = S_{\hat{h}}^2 / \delta_h^2$
-	h	5841	11,1378	22,6594	0,7000	23,3000	-
3	$\hat{h}$	5841	12,6899	15,2417	6,5685	19,6630	0,6726
9	$\hat{h}$	5841	12,6961	15,3100	6,5097	19,7628	0,6757

Em que: N = número de observações; h = altura observada em metros; $\hat{h}$ = altura estimada em metros; δ_h^2 = variância das observações; $S_{\hat{h}}^2$ = variância das estimativas; χ^2 = Qui-quadrado.

CONCLUSÕES

As duas melhores equações foram a número 3 ($h=b_0+b_1.t+b_2.t^5$) e a número 9 (modelo de Chapman-Richards), de acordo com o AIC. Também não houve diferença significativa quanto à generalidade, de acordo com o teste Qui-Quadrado, entre os modelos 3 e 9.

O uso do critério de Akaike (*AIC*), calculado sobre os dados amostrais, mostrou-se adequado como critério de seleção de modelos para representar uma série temporal de dados de altura de árvores, permitindo determinar que modelos apresentaram maior generalidade para a população de *Pinus elliottii* utilizada neste estudo, selecionando as equações 3 e 9 como melhores.

Os critérios quantitativos de falseabilidade, qualidade do ajustamento, complexidade e generabilidade permitiram definir os modelos 3 e 9 como os dois melhores e sem diferenças estatisticamente significativas entre si. O auxílio dos critérios qualitativos de seleção, levando-se em conta a ligação do modelo com o processo estudado, sua interpretabilidade e compreensibilidade permitiram a escolha final do modelo de Chapman-Richards.

REFERÊNCIAS BIBLIOGRÁFICAS

BUSSAB, Wilton de O. **Análise de variância e de regressão.** São Paulo: Atual, 1986. 147p.

COLLETT, Dave. Ordered cathegorical data – Lecture 14, 14p. In: **MSc generalised linear models course**. [Reading]: The University of Reading, School of Applied Statistics, 2003. Disponível em: < http://www.personal.rdg.ac.uk/~snscolet/MScGLMs/Lecture1.pdf>. Acesso em: 17 out. 2004.

FINGER, César A. G. **Fundamentos de biometria florestal**. Santa Maria : UFSM, CEPEF : FATEC, 1992. 269p.

GARCIA, O. Growth modelling : a (Re)view. **New Zealand Forestry**, n. 33, p. 14-17, 1988.

KEARNS, Michael; MANSOUR, Yishay; NG, Andrew Y.; RON, Dana. An experimental and theoretical comparison. **Machine Learning**, n. 27, p. 7–50, 1997.

MOTULSKY, Harvey; CHRISTOPOULOS, Arthur. **Fitting models to biological data using linear e nonlinear regression : a pratical guide to curve fitting.** San Diego : GraphPad Software, 2003. 351p.

MYUNG, In J.; PITT, Mark A.; KIM, Woojae. **Model evaluation, testing and selection**. Columbus : Ohio State University, Department of Psychology, 2003. 45p.

NAVARRO, Daniel J.; MYUNG, In Jae. **Model evaluation and selection**. Columbus, USA: Ohio State University, Department of Psychology, 2004. 6p.

NEMEC, Amanda F. Linnell. **Analysis of repeated measures and time series: an introduction with forestry examples - Handbook 6**. Victoria, Canada: Ministry of Forests, Forest Science Research Branch, Biometrics information, 1995. 83p.

PRODAN, M; PETERS, R.; COX, F.; REAL, P. **Mensura forestal**. San José: Instituto Interamericano de Cooperación para la Agricultura, 1997. 562 p.

SAS Institute. **SAS/ETS® User's Guide**, Version 7-2. Cary : SAS Institute, 1999a. 1550p.

____. **SAS/STAT® User's Guide**, Version 8. Cary : SAS Institute, 1999b. 3365p.

SCHEEREN, Luciano W. **Estruturação da produção de povoamentos monoclonais de *Eucalyptus saligna* Smith manejados em alto fuste**. 2003. 181f. Tese (Doutorado em Engenharia Florestal) – Universidade Federal de Santa maria, Santa Maria, 2003.

SCHNEIDER, Paulo R. **Introdução ao manejo florestal.** Santa Maria: UFSM,CEPEF : FATEC, 1993. 348p.

SIT, Vera. **Catalog of curves for curve fitting - Handbook 4**. Victoria: Ministry of Forests, Forest Science Research Branch, Biometrics information, 1994. 110p.

SOUZA, Geraldo da S. **Introdução aos modelos de regressão linear e não linear.** Brasília: EMBRAPA, 1998. 489p.

WONNACOTT, Thomas H.; WONNACOTT, Ronald J. **Introdução à estatística**. Rio de Janeiro: Livros Técnicos e Científicos, 1980. 589p.

ZIMMERMAN, Dale L.; NÚÑEZ-ANTÓN, Vicente. Parametric modelling of growth curve data: An overview, p.1-41. In: Modelling curve data. **Test, Sociedad de Estadística e Investigación Operativa,** v. 10, n. 1, p. 111-999, 2001.

ZUCCHINI, Walter. An introduction to model selection. **Journal of Mathematical Psychology**, n. 44, p. 41-61, 2000.

ANEXOS
PROGRAMAS SAS

```
* PROGRAMA 1: Selecao de variaveis e ajuste das equacoes e calculo de D de Durbin-Watson das nao-lineares;
DATA VARIAVS;
INFILE 'C:\CRESCIMENTO\AMOSTRA.DAT';
INPUT BL TR ARV IDADE H;
T=IDADE; LNT=LOG(T); LN2T=LOG(LOG(T)); U_T=1/T; T2=T**2; T3=T**3; T4=T**4; T5=T**5;
  LNH=LOG(H); U_H_13=1/(H-1.3); U_T2=1/T2; T2_H=T2/H;
PROC SORT DATA=VARIAVS; BY BL TR ARV IDADE;
GOPTIONS VSIZE=6CM HSIZE=8.5CM HTEXT=4PCT HTITLE=4.5PCT;
PROC GPLOT DATA=VARIAVS;
PLOT H*T / VAXIS=0 TO 24 BY 3 HREF=5 HAXIS=5 TO 19;
%MACRO GRAFICO;
PROC SORT DATA=RES; BY BL TR ARV IDADE;
PROC GPLOT DATA=RES; SYMBOL1 VALUE=PLUS CV=BLUE; SYMBOL2 VALUE=TRIANGLE
  CV=RED; PLOT RES*YEST / VREF=0 VAXIS=-7 TO 7 BY 2 HAXIS=6 TO 20; PLOT H*T YEST*T /
  OVERLAY VAXIS=0 TO 24 BY 3 HREF=5 HAXIS=5 TO 19; RUN;
%MEND GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=H; LNY=LOG(Y);
PROC REG DATA=VARIAVS; TITLE '1 - H=B0+B1*LN(T)+B2*LN(LN(T))'; VAR T; MODEL Y=LNT
  LN2T / SELECTION=FORWARD; OUTPUT OUT=RES P=P R=R; DATA RES; SET RES; YEST=P;
  RES=R; %GRAFICO;
PROC REG DATA=VARIAVS; TITLE '2 - H=B0+B1*(1/T)+B2*T+B3*T²+B4*LN(T)+B5*LN(LN(T))';
  MODEL Y=U_T T T2 LNT LN2T / SELECTION=FORWARD; OUTPUT OUT=RES P=P R=R; DATA
  RES; SET RES; YEST=P; RES=R; %GRAFICO;
PROC REG DATA=VARIAVS; TITLE '3 - H=B0+B1*T+B2*T²+B3*T³+B4*T4+B5*T5'; MODEL Y=T
  T2 T3 T4 T5 / SELECTION=FORWARD; OUTPUT OUT=RES P=P R=R; DATA RES; SET RES;
  YEST=P; RES=R; %GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '4 - LN(H)= B0+B1*LN(T)+B2*LN(LN(T))'; VAR T Y; MODEL
  LNY=LNT LN2T / SELECTION=FORWARD; PLOT R.*P.; OUTPUT OUT=RES P=ESTIMAT; DATA
  RES; SET RES; YEST=EXP(ESTIMAT); RES=H-YEST; %GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '5 -
  LN(H)=B0+B1*(1/T)+B2*T+B3*T²+B4*LN(T)+B5*LN(LN(T))'; MODEL Y=U_T T T2 LNT LN2T /
  SELECTION=FORWARD; PLOT R.*P.; OUTPUT OUT=RES P=ESTIMAT; DATA RES; SET RES;
  YEST=EXP(ESTIMAT); RES=H-YEST; %GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '6 - LN(H)=LN(B0)+B1*LN(T)+B2*T - LOG-
  EXPON./POT.COMBINADA'; MODEL Y=LNT T / SELECTION=FORWARD; PLOT R.*P.; OUTPUT
  OUT=RES P=ESTIMAT; DATA RES; SET RES; YEST=EXP(ESTIMAT); RES=H-YEST; %GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=U_H_13;
PROC REG DATA=VARIAVS; TITLE '7 - 1/H-1,3=B0+B1*1/T+B2*1/T²'; VAR T; MODEL Y=U_T
  U_T2 / SELECTION=FORWARD; PLOT R.*P.; OUTPUT OUT=RES P=ESTIMAT; DATA RES; SET
  RES; YEST=(1/ESTIMAT)+1.3; RES=H-YEST; %GRAFICO;
DATA VARIAVS; SET VARIAVS; Y=T2_H;
PROC REG DATA=VARIAVS; TITLE '8 - T²/H=B0+B1*T+B2.T²'; MODEL Y= T T2 /
  SELECTION=FORWARD; PLOT R.*P.; OUTPUT OUT=RES P=ESTIMAT; DATA RES; SET RES;
  YEST=T2/ESTIMAT; RES=H-YEST; %GRAFICO;
PROC NLIN DATA=VARIAVS; TITLE '9 - Y=B0*[1-EXP(-B1*T)]**B2 - CHAPMAN-RICHARDS';
  PARMS B0=37.27296 B1=0.04818 B2=1.18267; BOUNDS 10<=B0<=50; BOUNDS 0.0001<=B1<=1;
  BOUNDS 0<=B2<=20; X=T; EBX=EXP(-B1*X); EBX1=1-EBX; EBXB2=(EBX1)**B2; MODEL
  H=B0*EBXB2; DER.B0=EBXB2; DER.B1=B0*X*B2*EBX*EBX1**(B2-1);
```

```
  DER.B2=B0*EBXB2*LOG(EBX1); OUTPUT OUT=RES P=YEST R=RES;
PROC MODEL DATA=VARIAVS; TITLE '9 - Y=B0*[1-EXP(-B1*T)]**B2 - CHAPMAN-RICHARDS';
  PARMS B0=37.27296 B1=0.04818 B2=1.18267; BOUNDS 10<=B0<=50; BOUNDS 0.0001<=B1<=1;
  BOUNDS 0<=B2<=20; X=T; EBX=EXP(-B1*X); EBX1=1-EBX; EBXB2=(EBX1)**B2; H=B0*EBXB2;
  DER.B0=EBXB2; DER.B1=B0*X*B2*EBX*EBX1**(B2-1); DER.B2=B0*EBXB2*LOG(EBX1); FIT H
  / DW; %GRAFICO;
PROC NLIN DATA=VARIAVS; TITLE '10 - H=B0*EXP[-EXP(B1-B2*T)] - GOMPERTZ'; PARMS
  B0=30.0 B1=1.5 B2=0.22; BDT = EXP(B1-B2*T); BCBDT = EXP(-BDT); BBB = BDT*BCBDT*B0;
  MODEL H = BCBDT*B0; DER.B0 = BCBDT; DER.B1 = -BBB; DER.B2 = T*BBB; OUTPUT
  OUT=RES P=YEST R=RES;
PROC MODEL DATA=VARIAVS; PARMS B0=30.0 B1=1.5 B2=0.22; BDT = EXP(B1-B2*T); BCBDT =
  EXP(-BDT); BBB = BDT*BCBDT*B0; H = BCBDT*B0; DER.B0 = BCBDT; DER.B1 = -BBB; DER.B2
  = T*BBB; FIT H / DW; %GRAFICO;
PROC NLIN DATA=VARIAVS; TITLE '11 - H=B0*T**B1*EXP(B2*T) - EXP./POT. COMBINADA';
  PARMS B0=2.0 B1=2.0 B2=0.1; EBCX = T**B1*EXP(B2*T); MODEL H = B0*EBCX; DER.B0 =
  EBCX; DER.B1 = B0*EBCX*LOG(T); DER.B2 = B0*T*EBCX; OUTPUT OUT=RES P=YEST R=RES;
PROC MODEL DATA=VARIAVS; PARMS B0=2.0 B1=2.0 B2=0.1; EBCX = T**B1*EXP(B2*T); H =
  B0*EBCX; DER.B0 = EBCX; DER.B1 = B0*EBCX*LOG(T); DER.B2 = B0*T*EBCX; FIT H / DW;
  %GRAFICO;
RUN;
```

*** PROGRAMA 2: Calculo do AIC e D de Durbin-Watson das lineares e linearizadas;**

```
GOPTIONS VSIZE=5.5CM HSIZE=8.5CM HTEXT=5PCT HTITLE=5PCT;
DATA VARIAVS; INFILE 'C:\CRESCIMENTO\AMOSTRA.DAT'; INPUT BLOCO TRAT ARVORE
  IDADE H; T=IDADE; LNT=LOG(T); LN2T=LOG(LOG(T)); U_T=1/T; T2=T**2; T3=T**3; T4=T**4;
  T5=T**5; LNH=LOG(H); U_H_13=1/(H-1.3); U_T2=1/T2; T2_H=T2/H;
PROC SORT DATA=VARIAVS; BY IDADE TRAT ARVORE;
DATA VARIAVS; SET VARIAVS; Y=H;
PROC REG DATA=VARIAVS; TITLE '1 - H=B0+B1*LN(T)+B2*LN(LN(T))'; MODEL Y=LNT LN2T /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y=LNT LN2T / DW;
PROC REG DATA=VARIAVS; TITLE '2 - H=B0+B1*T+B2*T²'; MODEL Y= T T2 /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y= T T2 / DW;
PROC REG DATA=VARIAVS; TITLE '3 - H=B0+B1*T+B2*T5'; MODEL Y=T T5 /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y=T T5 / DW;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '4 - LN(H)= B0+B1*LN(LN(T))'; VAR T; MODEL Y=LNT LN2T
  / SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y=LNT LN2T / DW;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '5 - LN(H)=B0+B1*1/T+LN(LN(T))'; MODEL Y=U_T LN2T /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y=U_T LN2T / DW;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS; TITLE '6 - LN(H)=LN(B0)+B1*LN(T)+B2*T - LOG-
  EXPON./POT.COMBINADA'; MODEL Y=LNT T / SELECTION=RSQUARE MSE CP AIC BIC
  BEST=1; MODEL Y=LNT T / DW;
DATA VARIAVS; SET VARIAVS; Y=U_H_13;
PROC REG DATA=VARIAVS; TITLE '7 - 1/H-1,3=B0+B1*1/T+B2*1/T²'; MODEL Y=U_T U_T2 /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y=U_T U_T2 / DW;
DATA VARIAVS; SET VARIAVS; Y=T2_H;
PROC REG DATA=VARIAVS; TITLE '8 - T²/H=B0+B1*T+B2.T²'; MODEL Y= T T2 /
  SELECTION=RSQUARE MSE CP AIC BIC BEST=1; MODEL Y= T T2 / DW;
PROC NLMIXED DATA=VARIAVS; TITLE2 '9 - Y=B0*[1-EXP(-B1*T)]**B2 - CHAPMAN-
  RICHARDS'; PARMS B0=34.10967 B1=0.064784 B2=1.461857; EBT=EXP(-B1*T); EBT1=1-EBT;
```

```
  EBTB2=(EBT1)**B2; H = B0*EBTB2+ERRO; DER.B0=EBTB2;
  DER.B1=B0*T*B2*EBT*EBT1**(B2-1); DER.B2=B0*EBTB2*LOG(EBT1); MODEL H ~
  NORMAL(H,2.4796);
PROC NLMIXED DATA=VARIAVS; TITLE2 '10 - H=B0*EXP[-EXP(B1-B2*T)] - GOMPERTZ';
  PARMS B0=26.93389 B1=1.088928 B2=0.124987; BDT = EXP(B1-B2*T); BCBDT = EXP(-BDT); BBB
  = BDT*BCBDT*B0; H = BCBDT*B0+ERRO; DER.B0 = BCBDT; DER.B1 = -BBB; DER.B2 = T*BBB;
  MODEL H ~ NORMAL(H,2.4875);
PROC NLMIXED DATA=VARIAVS; TITLE2 '11 - H=B0*T**B1*EXP(B2*T) -
  EXPONENCIAL/POTENCIAL COMBINADA'; PARMS B0=0.658573 B1=1.398858 B2=-0.0357; EBCX
  = T**B1*EXP(B2*T); H = B0*EBCX+ERRO; DER.B0 = EBCX; DER.B1 = B0*EBCX*LOG(T);
  DER.B2 = B0*T*EBCX; MODEL H ~ NORMAL(H,2.4798);
RUN;
```

*** PROGRAMA 3: Calculo da SQRes (soma((h-hest)**2)) das equacoes com variaveis transformadas;**

```
OPTIONS LS=120 PS=58 NODATE NOSTIMER;
GOPTIONS VSIZE=5.5CM HSIZE=8.5CM HTEXT=5PCT HTITLE=5PCT;
DATA VARIAVS; INFILE 'C:\CRESCIMENTO\AMOSTRA.DAT'; INPUT BLOCO TRAT ARVORE
  IDADE H; T=IDADE; LNT=LOG(T); LN2T=LOG(LOG(T)); U_T=1/T; T2=T**2; T3=T**3; T4=T**4;
  T5=T**5; LNH=LOG(H); U_H_13=1/(H-1.3); U_T2=1/T2; T2_H=T2/H;
 PROC SORT DATA=VARIAVS; BY IDADE TRAT ARVORE;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS NOPRINT; TITLE2 '4 - LN(H)= B0+B1*LN(T)+B2*LN(LN(T))'; VAR T;
  MODEL Y=LNT LN2T; OUTPUT OUT=RES4 P=YEST;
DATA RES4;SET RES4; HEST=EXP(YEST);H=EXP(LNH); QRES=(H-HEST)**2;
PROC SUMMARY; VAR QRES; OUTPUT OUT=RES4 SUM(QRES)=SQRES; PROC PRINT;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS NOPRINT; TITLE2 '5 - LN(H)=B0+B1*1/T+B2*LN(LN(T))'; MODEL
  Y=U_T LN2T; OUTPUT OUT=RES5 P=YEST;
DATA RES5;SET RES5; HEST=EXP(YEST);H=EXP(LNH); QRES=(H-HEST)**2;
PROC SUMMARY; VAR QRES; OUTPUT OUT=RES5 SUM(QRES)=SQRES; PROC PRINT;
DATA VARIAVS; SET VARIAVS; Y=LNH;
PROC REG DATA=VARIAVS NOPRINT; TITLE2 '6 - LN(H)=LN(B0)+B1*LN(T)+B2*T - LOG-
  EXPON./POT.COMBINADA'; MODEL Y=LNT T ; OUTPUT OUT=RES6 P=YEST;
DATA RES6;SET RES6; HEST=EXP(YEST);H=EXP(LNH); QRES=(H-HEST)**2;
PROC SUMMARY; VAR QRES H; OUTPUT OUT=RES6 SUM(QRES)=SQRES MEAN(H)=HMED;
  PROC PRINT;
DATA VARIAVS; SET VARIAVS; Y=U_H_13;
PROC REG DATA=VARIAVS NOPRINT; TITLE2 '7 - 1/(H-1,3)=B0+B1*1/T+B2*1/T²'; VAR T; MODEL
  Y=U_T U_T2; OUTPUT OUT=RES7 P=YEST;
DATA RES7;SET RES7; HEST=(1/YEST)+1.3;H=(1/Y)+1.3; QRES=(H-HEST)**2;
PROC SUMMARY; VAR QRES; OUTPUT OUT=RES7 SUM(QRES)=SQRES; PROC PRINT;
DATA VARIAVS; SET VARIAVS; Y=T2_H;
PROC REG DATA=VARIAVS NOPRINT; TITLE2 '8 - T²/H=B0+B1*T+B2.T²'; MODEL Y= T T2;
  OUTPUT OUT=RES8 P=YEST;
DATA RES8;SET RES8; HEST=T2/YEST;H=T2/Y; QRES=(H-HEST)**2;
PROC SUMMARY; VAR QRES H; OUTPUT OUT=RES6 SUM(QRES)=SQRES MEAN(H)=HMED;
  PROC PRINT;
RUN;
```

*** PROGRAMA 4: Variancia da população e das estimativas pelas duas melhores equacoes;**

```
DATA MEDICOES;
```

```
INFILE 'C:\CRESCIMENTO\POPULACAO.DAT';
INPUT BL TR ARV T D H; H=H/10; MERGE_AUX=1;
PROC SORT DATA=MEDICOES; BY BL TR ARV T;
DATA VARIAVS; INFILE 'C:\CRESCIMENTO\AMOSTRA.DAT';
INPUT BL TR ARV T H; T2=T**2; T5=T**5; LNH=LOG(H); T2_H=T2/H; LNT=LOG(T);
  LN2T=LOG(LOG(T));
PROC SORT DATA=VARIAVS; BY BL TR ARV T;
* EQUACAO 3;
PROC REG DATA=VARIAVS OUTEST=PARAM3 NOPRINT;
MODEL H=T T5; DATA PARAM3 (KEEP=B30 B31 B32 MERGE_AUX); SET PARAM3;
  MERGE_AUX=1; B30=INTERCEPT; B31=T; B32=T5;
* EQUACAO 9;
PROC NLIN DATA=VARIAVS NOPRINT; PARMS A=37.27296 K=0.04818 R=1.18267; BOUNDS
  10<=A<=50; BOUNDS 0.0001<=K<=1; BOUNDS 0<=R<=20; X=T; EBX=EXP(-K*X); EBX1=1-EBX;
  EBXR=(EBX1)**R; MODEL H=A*EBXR; DER.A=EBXR; DER.K=A*X*R*EBX*EBX1**(R-1);
  DER.R=A*EBXR*LOG(EBX1); OUTPUT OUT=PARAM9 PARMS=A K R; DATA PARAM9
  (KEEP=A K R MERGE_AUX); SET PARAM9; MERGE_AUX=1;
DATA MEDICOES;
MERGE MEDICOES PARAM3 PARAM9; BY MERGE_AUX;
T2=T**2; T5=T**5; LNT=LOG(T); LN2T=LOG(LOG(T));
* EQUACAO 3; HEST3=B30+B31*T+B32*T**5; RES3=H-HEST3; QRES3=RES3**2;
* EQUACAO 9; HEST9=A*(1-EXP(-K*T))**R; RES9=H-HEST9; QRES9=RES9**2;
PROC SORT; BY BL TR ARV T;
PROC MEANS DATA=MEDICOES; VAR H HEST3 HEST9;
PROC MEANS DATA=MEDICOES; VAR QRES3 QRES9;
RUN;
```