Modelos Simétricos versus Regressão Kernel: Uma aplicação em Ciências Florestais

Luiz Medeiros de Araujo Lima Filho[1,2], Marcelo Rodrigo Portela Ferreira[2], Juliana Freitas Pires[2] e José Antônio Aleixo da Silva[3]

[1] Programa de Pós-Graduação em Biometria e Estatística Aplicada, Universidade Federal Rural de Pernambuco, CEP 52171-900 - Recife (PE) - Brasil
[2] Departamento de Estatística, Universidade Federal da Paraíba, CEP: 58051-900 - João Pessoa (PB)- Brasil
[3] Departamento de Ciência Florestal, Universidade Federal Rural de Pernambuco, CEP: 52171-900 - Recife (PE)- Brasil
luiz@de.ufpb.br

Resumo O Pólo Gesseiro do Araripe em Pernambuco é um grande consumidor de madeira para produção de gesso. Devido à grande necessidade de se buscar uma alternativa econômica e ambiental para a região é de interesse obter uma produção sustentável para o *Eucalyptus* ssp, uma vez que esta é um gênero de rápido crescimento e grande versatilidade. No planejamento do manejo florestal sustentado uma variável é de extrema importância: o crescimento. Sua modelagem é fundamental na prognose da produtividade, qualidade do local e dinâmica de populações. Geralmente, as curvas de crescimento são estudadas por meio de modelos não-lineares desenvolvidos empiricamente para relacionar, por exemplo, altura e idade. Um modelo não-linear bastante utilizado na prática para modelar curvas de crescimento é o modelo de Chapman-Richards. Em estudos deste tipo, em geral, assume-se que os erros seguem distribuição normal. Contudo, a modelagem sob a suposição de erros com distribuição normal é bastante sensível a valores atípicos que por ventura possam ocorrer, podendo distorcer as estimativas dos parâmetros. Uma alternativa para corrigir esse problema é adotar distribuições mais robustas que a distribuição normal. Desta forma, a classe de modelos simétricos se torna uma alternativa viável para corrigir tal problema. A classe dos modelos não paramétricos pode também ser considerada como uma alternativa viável, visto que muitas vezes as suposições feitas acerca da distribuição dos erros e da forma funcional do modelo podem ser muito restritivas ou até mesmo inadequadas. A idéia em regressão não paramétrica é estimar a relação entre a variável resposta e um conjunto de variáveis independentes diretamente dos dados, ao invés de estimar parâmetros. Neste trabalho, com a expectativa de obter melhores estimativas de crescimento em altura de *Eucalyptus* ssp, consideraremos os modelos simétricos, bem como o modelo de regressão não paramétrica via funções kernel. Para os modelos simétricos aplicaram-se ao modelo de Chapman-Richards as seguintes distribuições: normal, t de Student, Cauchy, exponencial potência e logística II. Dentre os modelos paramétricos, o modelo simétrico com distribuição Exponencial Potência e Logística II, de acordo com os critérios utilizados, foram os que apresentaram melhores estimativas de crescimento em altura de *Eucalyptus* ssp no Pólo Gesseiro de Pernambuco. Em contrapartida, o modelo de regressão kernel foi aplicado aos dados e apresentou melhores estimativas em comparação com os modelos paramétricos.

Palavras-chave: Log-Verossimilhança, Modelos Não-Lineares, Modelos Simétricos, Regressão Kernel.

1 Introdução

O Polo Gesseiro do Araripe, localizado na microregião de Araripina, semi-árido Pernambucano, é um grande consumidor de biomassa vegetal que é usada na calcinação da gipsita. Essa microregião abrange 10 municípios e é responsável por 95% do gesso industrializado no Brasil (ALBUQUERQUE, 2002).

O bioma caatinga, no qual está localizado o Pólo Gesseiro do Araripe-PE, vem sofrendo pressão visto que é explorado de forma desordenada. Esse fato se deve, principalmente, a crescente demanda por recursos naturais renováveis, aumentando gradativamente a sua degradação. Uma alternativa econômica e ambiental viável é a implementação e o manejo sustentado de povoamentos florestais nativos ou o reflorestamento com florestas de rápido crescimento, com destaque para o *Eucalyptus* ssp por sua elevada taxa de crescimento, a facilidade de reprodução, a rusticidade e o altíssimo nível de melhoramento genético em produtividade e qualidade da madeira.

Em face desse fato, torna-se de interesse quantificar o crescimento e a produção de florestas, promovendo um planejamento criterioso da produção através da prescrição de regimes de manejos adequados visando à qualidade do produto final. Sendo assim, pode-se dizer que a predição do crescimento e da produção é parte fundamental do processo de planejamento dos povoamentos florestais. Geralmente, as curvas de crescimento são estudadas por meio de modelos não-lineares. Um modelo não-linear bastante utilizado para descrever tais fenômenos em ciências florestais é o modelo de Chapman-Richards.

Ao longo dos anos, modelos supondo erros normais vêm sendo utilizados para descrever a maioria dos fenômenos aleatórios, dado que a suposição de normalidade sempre é muito atrativa para os erros dos modelos de regressão com resposta contínua. Contudo, observa-se que as estimativas obtidas para os coeficientes dos modelos normais se mostram sensíveis à presença de observações extremas. Desta forma, alternativas à suposição de erros normais têm sido propostas na literatura. Lange et al. (1989) propuseram o modelo baseado na suposição de erros t de Student. Taylor (1992) propôs o ajuste de um modelo de regressão linear supondo erro com distribuição exponencial potência com um parâmetro extra de forma. Galea et al. (2005) apresentaram alguns resultados sobre modelagem, em particular sobre o desenvolvimento da análise inferencial e de diagnóstico na classe não-lineares com erros simétricos independentes.

Os modelos de regressão não paramétricos têm recebido considerável atenção na última década de pesquisadores de diversas áreas e vêm se mostrando bastante eficazes em problemas de predição quando as suposições necessárias aos modelos paramétricos não se verificam (ver, por exemplo, Ruppert et al (2003)). Tais modelos trazem consigo a vantagem da flexibilidade por não estarem retritos a uma forma funcional específica, permitindo que "os dados falem por si próprios". Essa grande vantagem tem um preço: em muitas situações, estimadores não paramétricos são menos eficientes do que suas contrapartidas paramétricas quando o modelo paramétrico é válido (DIAS, 2001). Dentre os métodos de regressão não paramétrica podemos citar os estimadores por *splines*, a regressão via kernel (ou estimador de Nadaraya-Watson) e os modelos generalizados aditivos. O estimador de Nadaraya-Watson, considerado nesse trabalho, parte de uma idéia simples baseada na estimação da esperança condicional da variável resposta através de funções kernel.

O objetivo deste trabalho consiste em estimar a altura dos *Eucalyptus* ssp através de modelos simétricos não-lineares baseados em erros com distribuições mais robustas que a distribuição normal e através de regressão kernel.

2 Distribuições Simétricas

A família de distribuições simétricas gera uma classe geral de distribuições com a mesma simetria que a distribuição normal padrão. Entre essas distribuições podemos citar: t de Student, Cauchy, exponencial potência e logística II. Para maiores detalhes sobre a família de distribuições simétricas em modelos de regressão, podem ser encontradas em Cysneiros e Paula (2005a).

Diz-se que a variável aleatória Y tem distribuição simétrica, com suporte em $\Re$, com parâmetros de locação $\mu \in \Re$ e de escala $\phi > 0$, se sua função de densidade de probabilidade é dada por

$$f(y;\mu,\phi) = \frac{1}{\sqrt{\phi}} g\left\{u\right\}, \qquad y \in \Re, \tag{1}$$

para alguma função $g(\cdot)$ denominada função geradora de densidade, em que $u = \frac{(y-\mu)^2}{\phi}$, com $g(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} g(u) du = 1$. Essa condição é necessária para que $f(y;\mu,\phi)$

seja uma função densidade de probabilidade. Assim, denota-se por $Y \sim S(\mu, \phi)$ e denomina-se de variável aleatória simétrica.

A seguir, são apresentadas algumas distribuições simétricas com suporte na reta real para $Y \sim S(\mu, \phi)$ em que $u = (y - \mu)^2/\phi$.

2.1 Distribuição Normal

Diz-se que $Y \sim S(\mu, \phi)$ tem distribuição normal se sua função geradora de densidade $g(\cdot)$ é da forma

$$g(u) = \frac{1}{\sqrt{2\phi}} \exp\left(-u/2\right), \qquad u > 0, \tag{2}$$

então, Y tem distribuição normal denotada por $Y \sim N(\mu, \phi)$. O coeficiente de curtose desta distribuição é $\gamma_2 = 3$.

2.2 Distribuição t de Student

A variável aleatória $Y \sim S(\mu, \phi)$ tem distribuição t de Student se sua função geradora de densidade $g(\cdot)$ é da forma

$$g(u) = \frac{\nu^{\nu/2}}{B(1/2, \nu/2)} \left(\nu + u\right)^{-\frac{\nu+1}{2}}, \qquad \nu > 0, u > 0, \tag{3}$$

em que $B(\cdot, \cdot)$ é a função Beta. Assim, Y é denotada por $Y \sim t(\mu, \phi, \nu)$. O coeficiente de curtose é $\gamma_2 = 3 + \frac{6}{\nu - 4}$, para $\nu > 4$. Este coeficiente é maior que o coeficiente da distribuição normal.

2.3 Distribuição Cauchy

A variável aleatória $Y \sim S(\mu, \phi)$ tem distribuição de Cauchy se sua função geradora de densidade $g(\cdot)$ é da forma

$$g(u) = \frac{1}{\pi(1 + u)}, \qquad u > 0. \tag{4}$$

Essa distribuição, denotada por $Y \sim C(\mu, \phi)$, é também conhecida como distribuição de Pearson Tipo VII.

Por questão de brevidade, outras distribuições simétricas podem ser encontradas em Cysneiros et al. (2005).

3 Modelos Simétricos

A família simétrica de densidades de locação-dispersão guarda a estrutura da distribuição normal, mas elimina a forma específica da densidade normal para incluir densidades simétricas com caudas mais leves ou mais pesadas do que as caudas da normal.

Para introduzir uma estrutura regressora na classe de distribuições (1), toma-se a componente sistemática do modelo linear generalizado para o vetor da média $\mu = E(Y)$ dado por

$$g(\mu) = \eta_i(\beta) = h(x_i, \beta), \tag{5}$$

em que $g(\cdot)$ é conhecida e duas vezes diferenciável, $\eta_i(\beta)$ é o preditor não-linear, X é uma matriz $n \times p$ de posto completo e $\beta = (\beta_1, ..., \beta_p)^T$ é um conjunto de parâmetros não-lineares desconhecidos a serem estimados.

Os modelos simétricos assumem que as variáveis aleatórias $Y_1, ..., Y_n$ podem ser tratadas como distribuidas independentemente seguindo a componente aleatória (1) e a componente sistemática (5). Desta forma, o modelo definido em (1) e (5) é dito modelo simétrico não-linear.

O principal objetivo na análise de modelos simétricos é fazer inferências no vetor de parâmetros β e no parâmetro de dispersão ϕ. A log-verossimilhança para os parâmetros do modelo pode ser expressa como:

$$l(\beta, \phi) = -\frac{n}{2}\log\phi + \sum_{i=1}^{n} \log g\left\{\phi^{-1}(y_i - \mu_i)^2\right\}. \tag{6}$$

A log-verossimilhança apresentada pode ser maximizada incondicionalmente usando alguns softwares como o SAS, Matlab, R ou a linguagem de programação Ox.

4 Regressão Kernel Univariada

Considere que observações são coletadas de uma variável aleatória Y em n valores de uma variável independente X, isto é, sejam os pares (x_i, y_i), $i = 1, \ldots, n$, tais que o seguinte modelo de regressão pode ser proposto

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{7}$$

onde $g(\cdot)$ é uma função suave e contínua e as variáveis aleatórias ε_i têm média zero, variância comum σ^2, são não correlacionadas e independentes de Y.

A esperança condicional de Y dado $X = x$ é dada por

$$E(Y|X = x_i) = \int y f(y|x_i)dy = \int y \frac{f(x_i, y)}{f_X(x_i)}dy = g(x_i). \tag{8}$$

Nadaraya (1964) e Watson (1964) propuseram o seguinte estimador para a curva de regressão $g(x_i)$

$$\widehat{g}_h(x_i) = \frac{\sum_{j=1}^{n} K_h(x_i - x_j)y_j}{\sum_{j=1}^{n} K_h(x_i - x_j)}, \tag{9}$$

onde a constante h é chamada constante de suavização e $K_h(t) = h^{-1}K(h^{-1}t)$, onde $K(\cdot)$ é a função kernel, uma função não-negativa tal que $\int_{-\infty}^{+\infty} K(t)dt = 1$. Usualmente, mais não sempre, $K(\cdot)$ será uma função densidade de probabilidade simétrica, como, por exemplo, a densidade normal padrão $K(t) = (2\pi)^{-1/2}\exp\{-t^2/2\}$. Pode-se provar que $\widehat{g}_h(x_i)$ é um estimador consistente para $g(x_i)$ quando $h \to 0$ e $nh \to \infty$ (ver, por exemplo, Härdle (1990)).

A constante h controla o grau de suavização aplicado aos dados: se $h \to 0$, teremos uma curva com muito ruído; por outro lado, se $h \to \infty$, teremos uma curva muito suave. Em outras palavras, com h muito pequeno a curva tenderá a interpolar perfeitamente os dados, implicando em viés pequeno e grande variância, enquanto que com h muito grande teremos perda de detalhes na curvatura dos dados, implicando em viés grande e pequena variância. Claramente, há a necessidade de que a escolha de h seja feita considerando um compromisso entre viés e variância. A constante de suavização ótima é encontrada através da minimização de alguma função perda, usualmente a raiz do erro quadrático médio (RMSE). Seletores automáticos da constante de suavização, obtidos através de validação cruzada, têm sido usados na prática e, em geral, fornecem bons resultados.

5 Material e método

Em março de 2002, foi implantado na Estação Experimental do Instituto Agronômico de Pernambuco na Chapada do Araripe - PE o Módulo de Experimentação Florestal para o Pólo Gesseiro do Araripe. Foram utilizadas 83 árvores, sobreviventes das 100 árvores plantadas no início do experimento, sendo 25 árvores em cada uma das quatro repetições. A variável altura foi medida em todas as árvores ao longo do tempo, durante 6 anos e meio, totalizando 14 medições. O tempo inicial considerado foi de 2 meses que corresponde à idade em que as mudas foram plantadas no campo.

Para estimar a altura dos *Eucalyptus* ssp através dos modelos paramétricos, foi utilizado o modelo de Chapman-Richards na estrutura de modelos lineares generalizados para a média $\mu = E(Y)$, definido por

$$\mu = U(1 - \exp(-kt))^{\theta}, \tag{10}$$

com t correspondendo à idade da árvore em meses. Os modelos foram ajustados supondo diferentes distribuições simétricas para os erros. Este procedimento foi realizado por meio da Proc NLP do SAS.

A previsão através de regressão não paramétrica foi feita através da estimação da curva de regressão pelo estimador de Nadaraya-Watson, implementado na função `ksmooth` do `R`.

Para comparar os modelos ajustados aos dados foram utilizados o erro percentual absoluto médio (MAPE) e o erro quadrático médio (EQM).

6 Aplicação a dados reais

Para a seleção dos modelos ajustados, apresentamos na Tabela 1 os critérios MAPE e EQM. Assim, para os dados de altura o modelo supondo erro com distribuição Exponencial Potência obteve, dentre os modelos paramétricos, menor valor para o critério MAPE, enquanto que o modelo Logística II apresentou menor EQM. O modelo de regressão kernel com constante de suavização $h = 6,71$ selecionado por minimização do RMSE através de validação cruzada obteve os melhores resultados dentre todos os modelos considerados segundo ambos os critérios.

Tabela 1. Estatísticas para seleção dos modelos.

Modelo	**MAPE** (%)	**EQM**
Normal	9,52	0,65
Student t_2	9,18	0,67
Exp. Potência ($l = 0,1$)	9,11	0,65
Cauchy	9,16	0,70
Logística II	9,13	0,64
Regressão Kernel	7,82	0,47

7 Conclusão

Diante dos resultados, é possível concluir que, na classe dos modelos paramétricos, os modelos simétricos são bastantes relevantes para os estudos de modelos de crescimento de forma prática e bastante útil para análise de dados reais, contribuindo assim de forma efetiva, no sentido de ampliar as possibilidades de análise para os modelos de crescimento adotados em Ciências Florestais. Além disso, verifica-se que a classe de modelos não paramétricos pode se constituir numa alternativa viável quando as suposições acerca da distribuição dos erros e/ou da forma funcional dos modelos paramétricos não se verificarem.

Referências

1. ALBUQUERQUE, J.L. **Diagnóstico ambiental e questões estratégicas: Uma análise considerando o Pólo Gesseiro do Sertão do Araripe - Estado de Pernambuco**. 2002. Tese (Doutorado em Ciências Florestais) - Universidade Federal do Paraná, Brazil.
2. CYSNEIROS, F.J.A.; PAULA, G.A. Restricted Methods in Symmetrical Linear Regression Models. **Computational Statistics and Data Analysis**. V. 49, n. 3, p. 689-708, 2005a.

3. CYSNEIROS, F.J.A.; PAULA, G.A.; GALEA, M. **Modelos Simétricos Aplicados**. São Pedro: 9ª Escola de Modelos de Regressão, 2005b.
4. DIAS, R. Regressão Não Paramétrica. Relatório Técnico, UNICAMP, 2001.
5. GALEA, M.; PAULA, G.A.; CYSNEIROS, F.J.A. On Diagnostic in Symmetrical Nonlinear Models. **Statistics and Probability Letters**. v. 73, n. 4, p. 459-467, 2005.
6. HÄRDLE, W. **Smoothing Techniques With Implementation in S**. New York: Springer-Verlag, 1990.
7. LANGE, K.L.; LITTLE, R.J.A.; TAYLOR, J.M.G. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, v. 84, p. 881-896, 1989.
8. NADARAYA, E.A. On Estimating Regression. **Theory of Probability and its Applications**, v. 10, p. 186-190, 1964.
9. RUPPERT, D; WAND, M.P.; CARROL, R.J. **Semiparametric Regression**. New York: Cambridge University Press, 2003.
10. TAYLOR, J.M.G. Properties of modelling the error distribution with an extra shape parameter. **Computational statistics and data analysis**, v. 13, p. 33-46, 1992.
11. WATSON, G.S. Smooth Regression Analysis. **Sankya A**, v. 26, p. 359-372, 1964.