REGRESSÃO PARAMÉTRICA E NÃO PARAMÉTRICA NA PREDIÇÃO VOLUMÉTRICA

Jackson Rosa Silva[1], Deivison Venicio Souza[1], Gabriela Cristina Costa Silva[1]

1 Universidade Federal do Pará, Altamira, PA, Brasil. E-mail: jackson.silva@altamira.ufpa.br; deivisonvs@ufpa.br; gbcristina.silva@gmail.com.
Autor correspondente: Jackson Rosa Silva. E-mail: jackson.silva@altamira.ufpa.br.

RESUMO

A inferência pode ser desafiadora, principalmente quando lidamos com conjuntos de variáveis que não seguem uma distribuição de probabilidade conhecida. O objetivo do trabalho foi comparar modelos paramétricos e não paramétricos para estimar o volume de 150 árvores da espécie *Eucalyptus* ssp. A seleção das 150 árvores-amostras de espécies comerciais foi realizada sob o processo de amostragem aleatória. Observou-se que o modelo paramétrico Schumacher-Hall foi o único que atendeu aos pressupostos lineares, o não paramétrico *Support Vector Machine* também apresentou boa acurácia (R^2 0,981 e RMSE% 1,7). Concluiu-se que a regressão não paramétrica é uma alternativa útil não apenas em situações em que a relação entre as variáveis é complexa ou não linear, mas também em amostras de dados com distribuições definidas.
Palavras-chave: Aprendizado de máquina; inferência; pressupostos

PARAMETRIC AND NON-PARAMETRIC REGRESSION IN VOLUMETRIC PREDICTION

ABSTRACT

Inference can be challenging, especially when dealing with sets of variables that do not follow a known probability distribution. The objective was to compare parametric and non-parametric models to estimate the volume of 150 Eucalyptus ssp trees. The selection of 150 sample trees of commercial species was carried out under the process of randomized. The parametric Schumacher-Hall model was the only one that met the linear budgets, the non-parametric Support Vector Machine also showed good accuracy (R^2 0.981 and RMSE% 1.7). Non-parametric regression is a useful alternative not only in situations where the relationship between variables is complex or non-linear, but also in exposing data with determined distributions.
***Key words**: Machine learning; inference; assumptions*

INTRODUÇÃO

A modelagem estatística impulsiona o avanço do conhecimento em várias áreas, incluindo a da mensuração florestal. Ela permite compreender as relações entre variáveis e fazer boas inferências do comportamento dos elementos de amostras e populações, sendo crucial, por exemplo, na estimativa do volume de árvores (Souza, 2020). Porém, tal tarefa pode ser desafiadora, principalmente quando lidamos com conjuntos de variáveis que não seguem uma distribuição de probabilidade conhecida. Nesses casos, é recomendável recorrer às técnicas de modelagem estatísticas avançadas que permitam uma análise mais precisa e abrangente dos dados (Souza, 2020).

Tem-se a modelagem matemática, que é uma ferramenta clássica e importante, pois ela permite a construção de fórmulas que descrevem o comportamento de vários fenômenos, o que incluí os do setor biológico (Silva & Santana, 2014). Por exemplo, a distribuição beta pode ser usada para modelar a proporção de área foliar necrosada por uma determinada doença em uma planta. Contudo, apesar deste tipo de modelagem produzir boas aproximações, é necessário considerar a natureza biunívoca entre as curvas que representam o padrão de distribuição de probabilidade específico e os elementos representativos dos dados observados, que podem variar em termos de intensidade.

Também por esta razão, frequentemente tem sido recomendado a aplicação de métodos de Aprendizado de Máquina para criar modelos que possam inferir sobre variáveis biométricas (Diamantopoulou, 2005). Embora existam inúmeras técnicas de regressão disponíveis na literatura, muitas delas ainda não foram amplamente experimentadas no campo da Mensuração Florestal (Araújo *et al.*, 2018).

Portanto, o objetivo desse trabalho é, comparar modelos paramétricos e não paramétricos para com a estimação volumétrica da espécie *Eucalyptus* ssp. L'Héritier. Mais precisamente, comparar os resultados dos modelos paramétricos: Schumacher&Hall, Naslund, Spurr e Hohenadl&Krenn com os dos métodos não paramétricos: Support Vector Machine, Artificial Neural Network, Gaussian Process Regression Polynomial e Random Forests na predição volumétrica da espécie *Eucalyptus* ssp.

MATERIAL E MÉTODOS

Caracterização da Área de Obtenção das Amostras

O estudo foi realizado em um povoamento plantado de híbrido *Eucalyptus* ssp. com oito anos de idade, sob espaçamento de 3mx3m e densidade de 1.089 árvores/ha-¹. Localizado no município de Pacajá, às margens da BR-230 (Rodovia Transamazônica), na mesorregião do Sudoeste paraense, sob coordenadas Latitude: N –03º38’59,20473” e Longitude: E –50º 57’ 19,45084″ (Datum WGS 84).

A região possui clima tropical úmido do tipo “af”, conforme a classificação de Kõppen, com estações quentes e chuvosas, caracterizado por umidade relativa do ar de 85%, de temperatura anual variando de 21 ºC a 32 ºC e índice pluviométrico anual de 2300 mm. O solo é caracterizado como Argissolo Vermelho-Amarelo Distrófico de textura argilosa/muito argilosa e médio argilosa (IBGE, 2023).

Coleta da Amostra

A seleção das 150 árvores-amostras de espécies comerciais foi realizada sob o processo de amostragem aleatória, e baseada na lista de espécies autorizadas para exploração florestal e constantes no inventário florestal 100% (IF100%) da Unidade de Produção Anual (UPA). As árvores selecionadas foram cubadas usando a metodologia de Huber.

Modelagem Preditiva

Na modelagem preditiva foram estudados modelos paramétricos (Tabela 1) e não paramétricos (Tabela 2).

Tabela 1. Modelos paramétricos testados para a estimativa de volume em árvores da espécie *Eucalyptus* ssp

Modelo	Denominação
$\ln(V)_i = \beta_0 + \beta_1 \ln(D)_i + \beta_2 \ln(H)_i + \varepsilon_i$	Schumacher-Hall
$V_i = \beta_0 + \beta_1 D_i^2 + \beta_2 (D^2H)_i + \beta_3 (DH^2)_i + \beta_4 H_i^2 + \varepsilon_i$	Naslund
$V_i = \beta_0 + \beta 1 (D^2H)_i + \varepsilon_i$	Spurr
$V_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \varepsilon_i$	Hohenadl-Krenn

H = altura total da árvore (m); D = diâmetro medido a 1,30 m do solo (cm); Ln = logaritmo neperiano; β_0, β_1 e β_2 = parâmetros do modelo; e εi = erro aleatório.

Tabela 2. Modelos não paramétricos testados para a estimativa de volume em árvores da espécie *Eucalyptus ssp.*

Algoritmo	Autor
Support Vector Machine	Cortes & Vapnik (1995)
Artificial Neural Network	Venables & Ripley (2002)
Gaussian Process Regression Polynomial	Rasmussen & Williams (2003)
Random Forests	Breiman (2001)

Treino e Teste dos Modelos

Os dados foram divididos em 70% para treino e 30% para teste. Foi utilizado o método de validação cruzada *Repeated Cross-Validation*, que envolve a divisão aleatória dos dados de treinamento em k partes iguais, sendo k-1 usadas para o ajuste do modelo e a parte restante usada para a avaliação da performance. No caso, houve 15 vezes repetições e k igual a 4. Tal procedimento ajudou a avaliar a performance do modelo de forma mais robusta e reduzir o risco de overfitting, principalmente para os métodos não paramétricos. Todos os modelos foram treinados no software R, versão 4.2.2. Os pacotes utilizados foram: caret, mlbench, pROC, caretEnsemble, data.table, lmteste, nortest, parallel, Rweka e rJava.

Fórmulas e Conceitos

Teorema Gauss-Markov

Desde que sejam satisfeitas algumas condições: os erros ε com média zero e variância constante; os erros ε não sejam correlacionados entre si e que os erros ε tenham distribuição

normal, o método dos Mínimos Quadrados Ordinários (MQO) produzirá as melhores estimativas lineares não-enviesadas dos coeficientes de regressão (Greene, 2003).

Verificação dos Resíduos

A verificação dos resíduos foi feita usando o teste de correlação de Pearson (Pearson, 1896), o teste de Breusch-Pagan para homoscedasticidade (Breusch & Pagan, 1979) e o teste de Shapiro-Wilk para normalidade (Shapiro & Wilk, 1965).

Avaliação dos Modelos

Os modelos foram avaliados utilizando as estatísticas raiz do erro quadrático médio (RMSE), coeficiente de determinação (R^2) e erro absoluto médio (MAE) (Greene, 2003), além de uma análise gráfica dos resíduos do conjunto de teste.

RESULTADOS E DISCUSSÃO

Correlação dos Vetores

Pelo teste de Pearson, foi observada uma forte correlação positiva de 0,96 entre os vetores de diâmetro e volume. Além disto, foi encontrada uma correlação positiva moderada de 0,78 entre os vetores de altura e volume, bem como uma correlação positiva moderada de 0,73 entre os vetores de altura e diâmetro.

Importância Específica das Variáveis Dependentes por meio da Permutação

A importância da variável baseada em permutação é definida como a mudança relativa nos desempenhos preditivos do modelo entre conjuntos de dados com e sem valores permutados para a variável associada (Fisher *et al.*, 2019). Na Figura 1 está a representação gráfica dos resultados dessa aplicação para o caso.

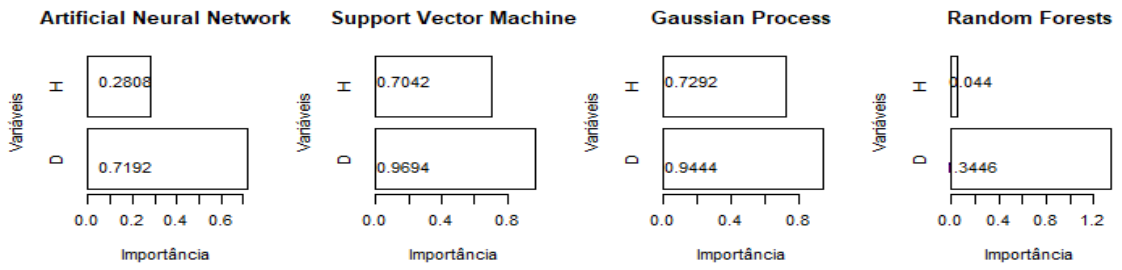

Figura 1. Importância específica das variáveis dependentes

Normalidade e Homocedasticidade dos Resíduos dos Modelos Paramétricos

Quando os resíduos de um modelo de regressão não são homocedásticos, significa que a variância dos erros não é constante em toda a faixa de valores das variáveis independentes. Isto pode levar a uma subestimação da variabilidade do modelo, o que pode resultar em uma superestimação da significância dos coeficientes de regressão e, consequentemente, do R^2 (Montgomery *et al.*, 2021). Fato observável no modelo Naslund, com coeficiente de determinação alto, mas com resíduos demasiadamente heterocedásticos e não gaussianos, conforme Tabela 3.

Tabela 3. Testes de normalidade e homoscedasticidade da variância dos modelos paramétricos avaliados para estimativa do volume de *Eucalyptus ssp*

Modelo	Teste de Normalidade de Shapiro-Wilk (p-valor)	Teste de Homocedasticidade de Breusch-Pagan (p-valor)
Naslund	0,0174*	0,0009*
Hohenadl-Krenn	0,6082^{ns}	0,0077*
Spurr	0,0351*	0,7911^{ns}
Schumacher-Hall	0,7103^{ns}	0,4633

Em que: * significativo ($\alpha < 0,05$); ns não significativo ($\alpha \geq 0,05$)

Resíduos Padronizados, Índices de RMSE, R^2 e MAE (Inferência Base Teste)

Alguns modelos paramétricos apresentaram bons coeficientes de determinação e erros médios quadráticos (Figura 2), contudo não atenderam aos pressupostos paramétricos. O único modelo paramétrico que pode ser útil para inferir sobre a população, por ainda possuir variância mínima e não possuir viés dos estimadores da variância dos β estimados, é o Schumacher-Hall sob o intercepto = -0,2927, coeficiente de log(D) = 1,9115 e coeficiente de log(H) = 0,7737. Já para o caso dos não paramétricos, o melhor foi o SVM com hiperparâmetro degree 2 (referente ao grau do polinômio usado para transformar os dados), hiperparâmetro scale em 0.1 (tamanho da janela de pesquisa em torno do ponto de consulta para encontrar os pontos de dados mais próximos) e o parâmetro "C" em 1, que controla o trade-off entre a precisão da regressão e a maximização da margem.

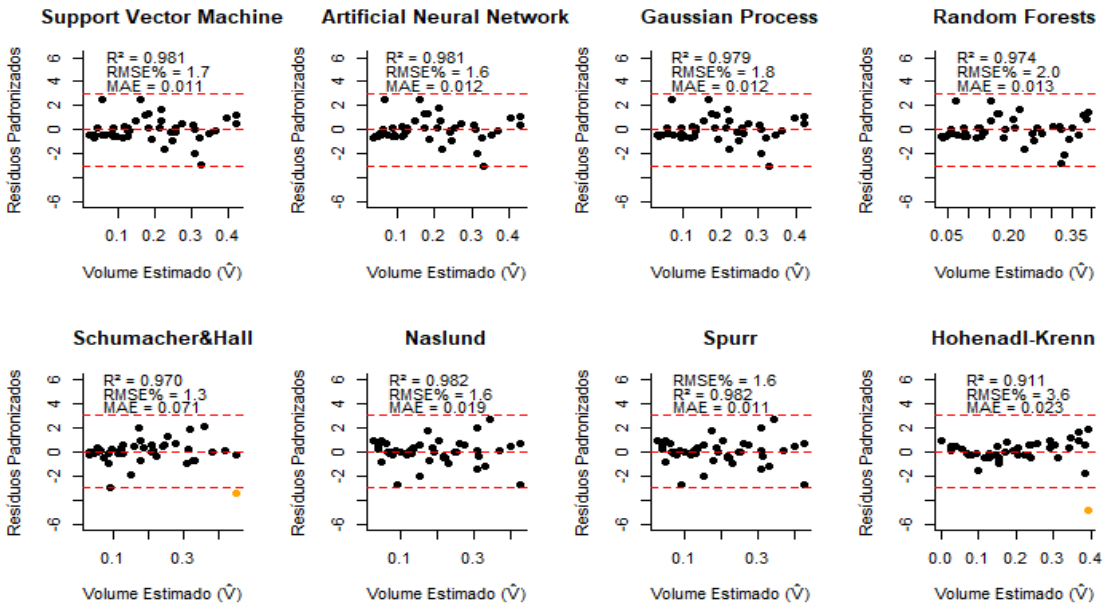

Figura 2. Resíduos padronizados, raiz do erro quadrático médio (RMSE), coeficiente de determinação (R^2) e erro absoluto médio (MAE)

CONCLUSÃO

A regressão não paramétrica é uma alternativa útil não apenas em situações em que a relação entre as variáveis é complexa ou não linear, mas também em amostras de dados com distribuições definidas. Uma das vantagens é que os modelos não paramétricos, geralmente, não dependem das suposições de que os resíduos sejam homocedásticos e normalmente distribuídos.

REFERÊNCIAS BIBLIOGRÁFICAS

Araújo, E. J. G.; Loureiro, G. H.; Sanquetta, C. R.; Sanquetta, M N. I.; Dalla Corte, A. P.; Péllico Netto, S.; Behling, A Allometric models to biomass in restoration areas in the atlantic rainforest. **Floresta e Ambiente**, v. 25, n. 1, e201601932018, 2018. https://doi.org/10.1590/2179-8087.019316.

Breiman, L. Random forests. **Machine Learning**, v. 45, p.5-32, 2001. https://doi.org/10.1023/A:1010933404324.

Breusch, T. S.; Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. **Econometrica**, v. 47, n. 5, p.1287–1294, 1979. https://doi.org/10.2307/1911963.

Cortes, C.; Vapnik, V. Support-vector networks. **Machine Learning,** v. 20, p.273–297, 1995. https://doi.org/10.1023/A:1022627411411.

Diamantopoulou, M. J. Artificial neural networks as an alternative tool in pine bark volume estimation. **Computers and Electronics in Agriculture**, v. 48, n. 3, p.235–244, 2005. https://doi.org/10.1016/j.compag.2005.04.002.

Greene, W. H. **Econometric analysis**. 5.ed. Upper Saddle River: Prentice Hall, 2003. 1024p.

Instituto Brasileiro de Geografia e Estatística – IBGE. **Pedologia**. Amazônia Legal. Disponível em: https://www.ibge.gov.br/geociencias/informacoes-ambientais/pedologia/15819-amazonia-legal.html. Acesso em: 15 Jul. 2023.

Montgomery D. C.; Peck E. A.; Vining, G. G. **Introduction to linear regression analysis**. 6.ed. New Jersey: Wiley, 2021. 704p.

Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. **Philosophical Transactions of the Royal Society A**, v. 187, p.253–318, 1896. https://doi.org/10.1098/rsta.1896.0007.

Rasmussen, C. E.; Williams, C. K. I. **Gaussian processes for machine learning**. Cambridge: MIT Press, 2006. 248p. Disponível em: https://gaussianprocess.org/gpml/chapters/RW.pdf. Acesso em: 10 Jun. 2023.

Shapiro, S.S.; Wilk, M. B. An analysis of variance test for normality. **Biometrika**, v. 52, n. 3/4, p.591-611, 1965. https://doi.org/10.1093/biomet/52.3-4.591.

Silva E. N.; Santana A. C. Modelos de regressão para estimação do volume de árvores comerciais, em florestas de Paragominas. **Revista Ceres**, v. 61, n. 5, p.631-636, 2014. https://doi.org/10.1590/0034-737X201461050005.

Souza, V. D. **Aprendizado de máquina para a predição de biomassa e volume comercial de árvores em florestas tropicais**. 2020. 171f. Tese (Doutorado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2020. Disponível em: https://hdl.handle.net/1884/74596. Acesso em: 10 Jun. 2023.

Venables, W. N.; Ripley, B. D. **Modern applied statistics with S-Plus**. New York: Springer, 2002. 493p. https://doi.org/10.1007/978-0-387-21706-2.