

O VIZINHO MAIS PRÓXIMO PONDERADO NA PREDIÇÃO DA BIOMASSA DA PARTE AÉREA DE ÁRVORES EM FLORESTAS TROPICAIS

Deivison Venicio Souza

Universidade Federal Pará - UFPA
Engenheiro Florestal, Me. Ciências Florestais
Programa de Pós-graduação em Engenharia Florestal - UFPR
(deivisonvs@ufpa.br)

**63^a Reunião Anual da Região Brasileira da Sociedade Internacional de
Biometria (RBBras)**



1 BIOMASSA

2 OBJETIVO GERAL

3 METODOLOGIA

4 RESULTADOS PARCIAIS

“Esta apresentação constitui uma pequena parte da pesquisa inédita de doutorado (em andamento) do palestrante que está sendo conduzida no Programa de Pós-graduação em Engenharia Florestal da UFPR.”

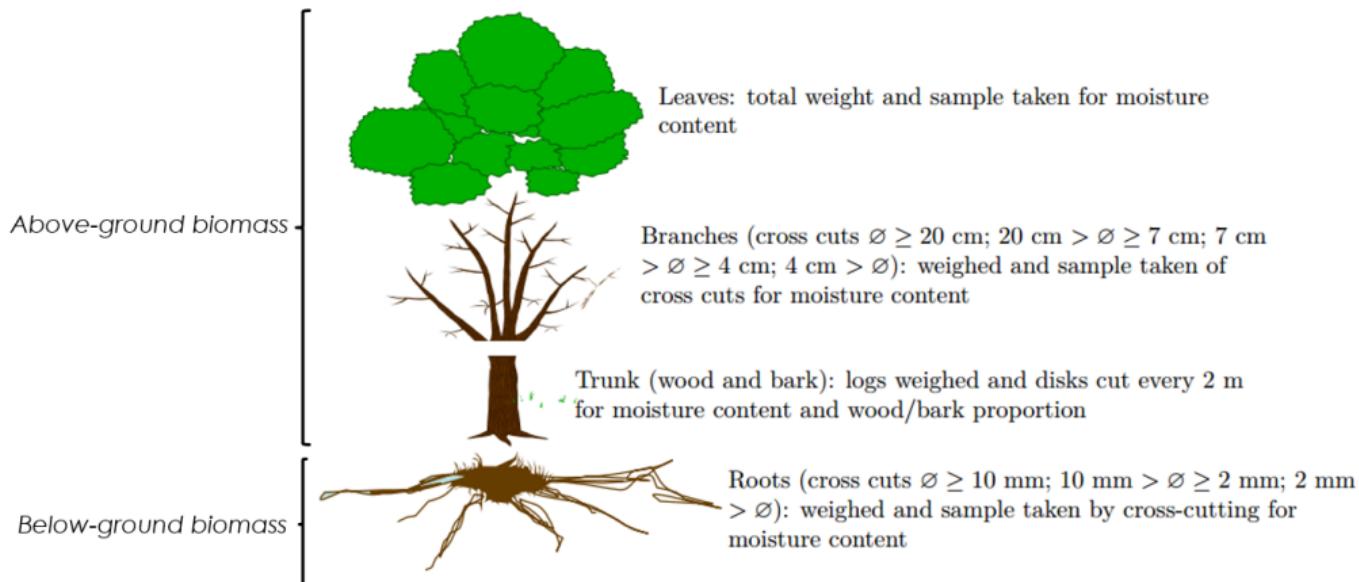
O que é?

O termo “biomassa” se refere à massa dos componentes de uma árvore excluindo-se a água, isto é, a **“massa seca”** da árvore (BATISTA et al., 2014).

Por que?

A biomassa precisa ser determinada e estimada de forma fidedigna. Do contrário não haverá consistência na **quantificação do carbono fixado nos ecossistemas florestais** (SANQUETTA, 2002).

Compartimentalização da biomassa de árvores

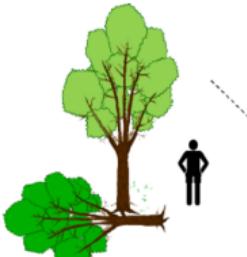


Fonte: Picard et al., 2012

Como quantificar a biomassa de árvores?

Método direto

Etapa 1
Preparação e derruba de árvores;



Operation 7

Etapa 2
Medição de árvores cortadas: perfil do tronco, marcação para corte transversal;



Operation 1

Etapa 3
Desfolha e desrama;



Etapa 4
Segmentação e obtenção de amostras de discos de madeira.



Etapa 5
Pesagem de folhas, galhos e toras de madeira.



Operation 2

Etapa 6
Retirada de amostras de folhas, galhos e discos do tronco



Operation 3

Etapa 7
Amostras, incluindo os discos, são levados para pesagem



Operation 4



Operation 6



Operation 5



Fonte: Picard et al., 2012



PROBLEMA?

Floresta Multiânea:

- Simplesmente impossível, impraticável!
- Portanto, é necessário aperfeiçoar os métodos indiretos de estimativa de biomassa florestal (HIGUCHI et al., 2004).

Potencial frente à Regressão Tradicional: Aprendizagem de Máquina.

Problemas com uso da regressão...

1 - Nem sempre se consegue obter modelos alométricos satisfatórios, em termos de precisão das estimativas que, muitas vezes, resultam em níveis de erro para além limiares de tolerância nas medidas florestais (SANQUETTA et al., 2013);

2 - Grande variabilidade natural dos dados de biomassa:
> variabilidade de biomassa > diâmetro \Rightarrow Espécies nativas de regiões tropicais e subtropicais (SANQUETTA et al., 2015);

Problemas com uso da regressão...

Problemática?

- 3) **Formulação matemática única** \Rightarrow Pode não reproduzir a elevada variabilidade natural dos dados de biomassa. Afetar a qualidade do modelo e fornecer estimativas viesadas (SANQUETTA et al., 2013); e
- 4) **Atendimento a pressuposições da regressão** \Rightarrow aditividade e linearidade, independência dos resíduos, homocedasticidade e normalidade de resíduos, inexistência de multicolinearidade.

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

Potencial: Aprendizagem de Máquina

Modelagem não-paramétrica

Estudos recentes têm despertado para o potencial da técnica de mineração de dados (**Data Mining**) na predição de variáveis biométricas. (ex.: (SANQUETTA et al., 2013; SANQUETTA et al., 2015))

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

OBJETIVO GERAL

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

Objetivo geral

O principal objetivo deste estudo foi aplicar uma abordagem não-paramétrica na estimativa da biomassa da parte aérea de árvores em florestas tropicais usando do algoritmo *Weighted k-Nearest-Neighbor (wkNN)* implementado por Schliep e Hechenbichler (2016) na library "kknn" do ambiente estatístico R; e

Prover uma comparação do desempenho preditivo do modelo *wkNN* final (ajustado com todos dados e usando da configuração ótima de hiperparâmetros) frente ao *Modelo Pantropical* proposto por Chave et al. (2014).

Chave et al. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees.
Global change biology, 20(10), 3177-3190.

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

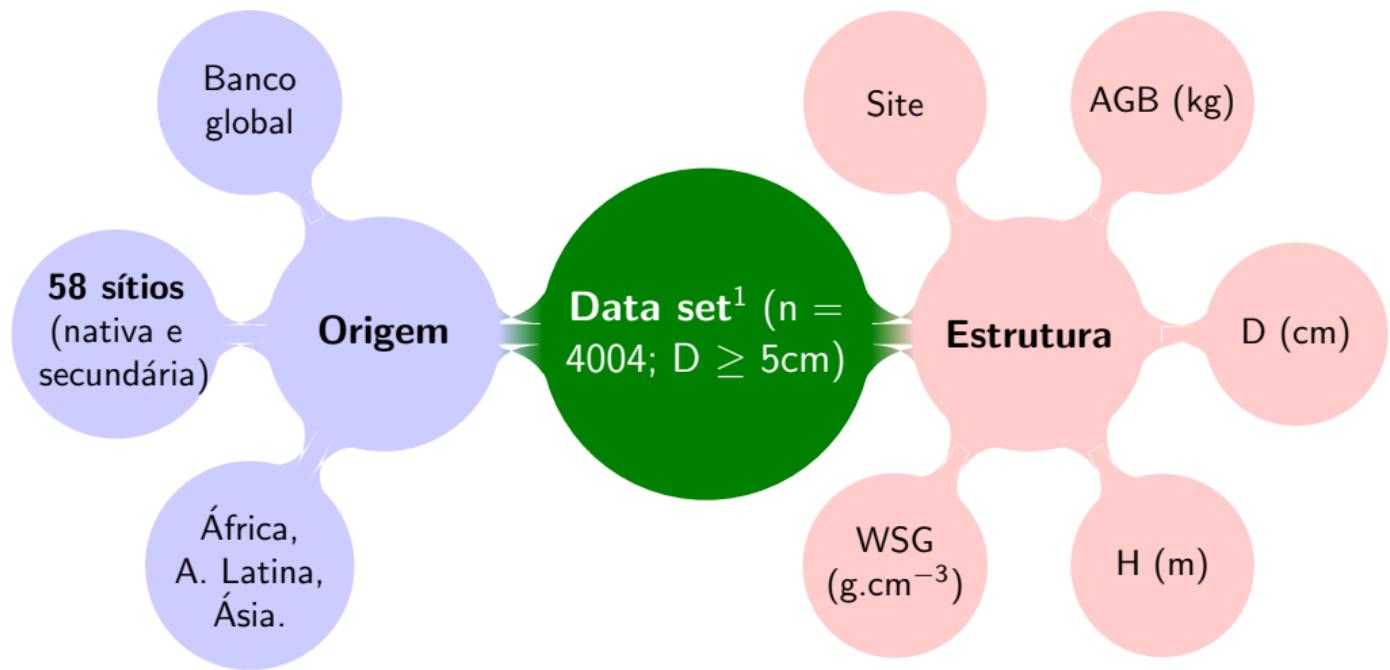
CONSTRUÇÃO DOS
MODELOS

PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

METODOLOGIA



¹Conjunto de dados compilado por Chave et al. (2014). Em que: AGB = Above-Ground Biomass; WSG = Wood Specific Gravity; D = Diameter at Breast Height; e H = Total Height.

Modelo Paramétrico: Regressão Linear (CHAVE et al., 2014)

$$\ln(AGB_{est.})_{(i)} = \alpha + \beta \underbrace{\ln(\rho D^2 H)_{(i)}}_{\text{Cov. Combinada}} + \epsilon_i \quad (1)$$

Modelo Pantropical: BIOMASS (REJOU-MECHAIN et al., 2018)

$$AGB_{est.(i)} = 0,0673 \times (\rho D^2 H)^{0,976} \quad (2)$$

$(S_{yx} = 0,357; AIC = 3130; df = 4002)$

Tabela: Bias e CV em nível de sítio j .

| Modelo | Bias _(j) (%) | CV _(j) (%) |
|--------|-------------------------|-----------------------|
| MP | 5,31% | 56,50% |

Modelo Paramétrico: Regressão Linear (CHAVE et al., 2014)

$$\ln(AGB_{est.})_{(i)} = \alpha + \beta \underbrace{\ln(\rho D^2 H)_{(i)}}_{\text{Cov. Combinada}} + \epsilon_i \quad (1)$$

Modelo Pantropical: BIOMASS (REJOU-MECHAIN et al., 2018)

$$AGB_{est.(i)} = 0,0673 \times (\rho D^2 H)^{0,976} \quad (2)$$

$(S_{yx} = 0,357; AIC = 3130; df = 4002)$

Tabela: Bias e CV em nível de sítio j .

| Modelo | Bias _(j) (%) | CV _(j) (%) |
|--------|-------------------------|-----------------------|
| MP | 5,31% | 56,50% |

Modelo Paramétrico: Regressão Linear (CHAVE et al., 2014)

$$\ln(AGB_{est.})_{(i)} = \alpha + \beta \underbrace{\ln(\rho D^2 H)_{(i)}}_{\text{Cov. Combinada}} + \epsilon_i \quad (1)$$

Modelo Pantropical: BIOMASS (REJOU-MECHAIN et al., 2018)

$$AGB_{est.(i)} = 0,0673 \times (\rho D^2 H)^{0,976} \quad (2)$$

$(S_{yx} = 0,357; AIC = 3130; df = 4002)$

Tabela: Bias e CV em nível de sítio j .

| Modelo | Bias $_{(j)}$ (%) | CV $_{(j)}$ (%) |
|--------|-------------------|-----------------|
| MP | 5,31% | 56,50% |

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA
ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO
CONSTRUÇÃO DOS
MODELOS
PREDITIVOS
ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

Indagação

Algoritmos de aprendizagem podem superar o modelo pantropical (PM) proposto por Chave et al. (2014)?

↓ **Bias** = 5,31% e ↓ **CV** = 56,50%

Iniciativa

Usar algoritmos implementados no ambiente R:

Weighted k-Nearest-Neighbor (wkNN)

Library "kknn" (SCHLIEP; HECHENBICHLER, 2016)

Modelo Não-Paramétrico: *Hiperparâmetros de ajuste - Pacote "kknn"* (SCH-LIEP; HECHENBICHLER, 2016)

O algoritmo *wkNN* possui três hiperparâmetros:

- 1 k = número de vizinhos mais próximos; ($k = 24$; start = 2; kmáx = 25)
- 2 d = métrica de distância; e (3 distâncias)
- 3 w = função de ponderação kernel. (10 funções)

Qual a melhor configuração de hiperparâmetros para predizer a AGB?

Modelo Não-Paramétrico: *Hiperparâmetros de ajuste - Pacote "kknn"* (SCH-LIEP; HECHENBICHLER, 2016)

O algoritmo *wkNN* possui três hiperparâmetros:

- 1 k = número de vizinhos mais próximos; ($k = 24$; start = 2; kmáx = 25)
- 2 d = métrica de distância; e (3 distâncias)
- 3 w = função de ponderação kernel. (10 funções)

Qual a melhor configuração de hiperparâmetros para predizer a AGB?

Modelo Não-Paramétrico: *Hiperparâmetros de ajuste - Pacote "kknn"* (SCH-LIEP; HECHENBICHLER, 2016)

O algoritmo *wkNN* possui três hiperparâmetros:

- 1 k = número de vizinhos mais próximos; ($k = 24$; start = 2; kmáx = 25)
- 2 d = métrica de distância; e (3 distâncias)
- 3 w = função de ponderação kernel. (10 funções)

Qual a melhor configuração de hiperparâmetros para predizer a AGB?

Modelo Não-Paramétrico: Hiperparâmetros de ajuste

Métricas de distâncias (ZHAO; CHEN, 2016):

Minkowski ($p=3$)

$$d = \left(\sum_{i=1}^n |q_i - x_i|^p \right)^{\frac{1}{p}} \quad (3)$$

Euclidean ($p=2$)

$$d = \sqrt{\sum_{i=1}^n (q_i - x_i)^2} \quad (4)$$

Manhattan ($p=1$)

$$d = \sum_{i=1}^n |q_i - x_i| \quad (5)$$

Modelo Não-Paramétrico: Hiperparâmetros de ajuste

Funções Kernel (HECHENBICHLER; SCHLIEP, 2004):

$$\text{Rectangular} = \frac{1}{2} \quad (6)$$

$$\text{Triangular} = (1 - |d|) \quad (7)$$

$$\text{Epanechnikov} = \frac{3}{4}(1 - d^2) \quad (8)$$

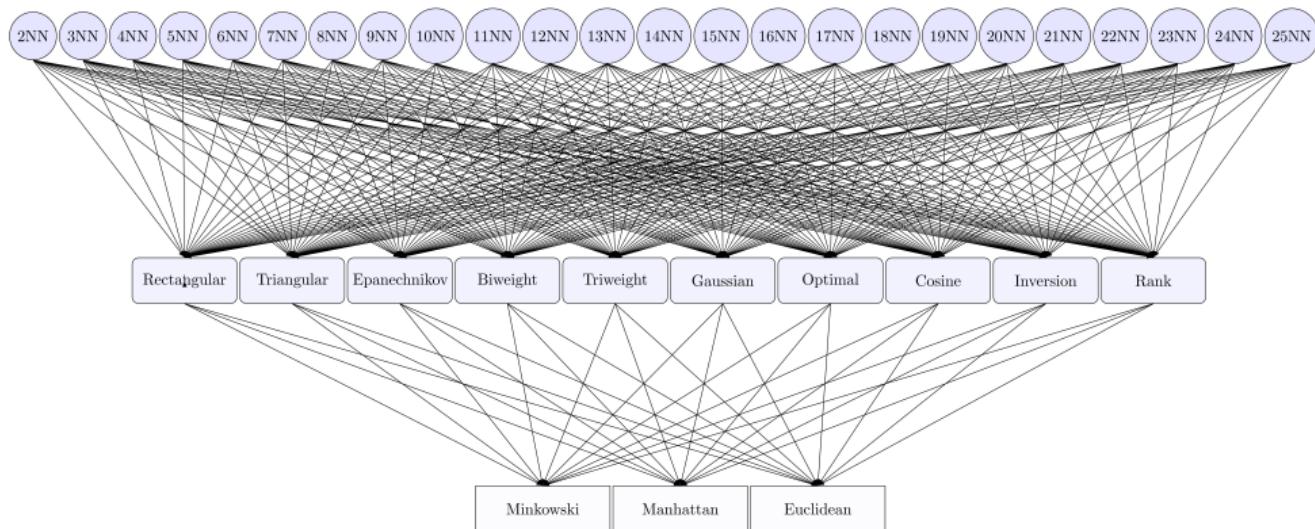
$$\text{Biweight} = \frac{15}{16}(1 - d^2)^2 \quad (9)$$

$$\text{Triweight} = \frac{35}{32}(1 - d^2)^3 \quad (10)$$

$$\text{Cosine} = \frac{\pi}{4} \cos\left(\frac{\pi}{2}d\right) \quad (11)$$

$$\text{Gauss} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{d^2}{2}\right) \quad (12)$$

Modelo Não-Paramétrico: candidatos à hiperparâmetro de ajuste ótimo



Total = $24 \times 10 \times 3 = 720$ variantes do algoritmo wkNN.

Modelo Não-Paramétrico: Variação de preditores (covariáveis)

Foram testadas três variações de preditores da AGB:

$$3 \text{ tipos} = \begin{cases} \mathbf{Vpred}_1 \implies AGB_{est(i, j)} = f(D, H, WSG) \\ \mathbf{Vpred}_2 \implies AGB_{est(i, j)} = f(D, H) \\ \mathbf{Vpred}_3 \implies AGB_{est(i, j)} = f(D) \end{cases}$$

Em que: AGB = Above-Ground Biomass; WSG = Wood Specific Gravity;
D = Diameter at Breast Height; e H = Total Height.

Modelo Não-Paramétrico: Pacote **caret** (Classification and Regression Training) (KUHN; JOHNSON, 2013; KUHN, 2018)

- 1 **Diferencial:** Interface uniforme para treinamento e previsão de diversos modelos.
- 2 **Emprego:** divisão dos dados, pré-processamento, método de reamostragem (repeats of k -folds cross-validation).

| Tarefa | Função |
|------------------------------------|---------------------|
| Divisão de dados | createDataPartition |
| Pré-processamento | center e scale |
| Modelagem usando resampling | repeatedcv |
| Configuração de recursos de treino | trainControl |
| Aprendizado dos modelos | train |

Modelo Não-Paramétrico: Estimativa da capacidade de generalização

1. Root Mean Square Error (RMSE)

$$RMSE_{(k)} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(AGB_{obs(i, j)} - AGB_{est(i, j)} \right)^2} \quad (13)$$

2. Relative Root Mean Square Error (rRMSE)

$$rRMSE_{(k)} = \frac{100}{MAGB_{obs}} \sqrt{\frac{1}{n} \sum_{i=1}^n \left(AGB_{obs(i, j)} - AGB_{est(i, j)} \right)^2} \quad (14)$$

Modelo Não-Paramétrico: Estimativa da capacidade de generalização

3. *R-squared* (R^2)

$$R_{(k)}^2 = \left(\frac{\sum_{i=1}^n (AGB_{est(i,j)} - MAGB_{est}) (AGB_{obs(i,j)} - MAGB_{obs})}{\sqrt{\left[\sum_{i=1}^n (AGB_{est(i,j)} - MAGB_{est})^2 \right] \left[\sum_{i=1}^n (AGB_{obs(i,j)} - MAGB_{obs})^2 \right]}} \right)^2 \quad (15)$$

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

RESULTADOS PARCIAIS

Em termos gerais, do teste de variações de preditores observou-se que:

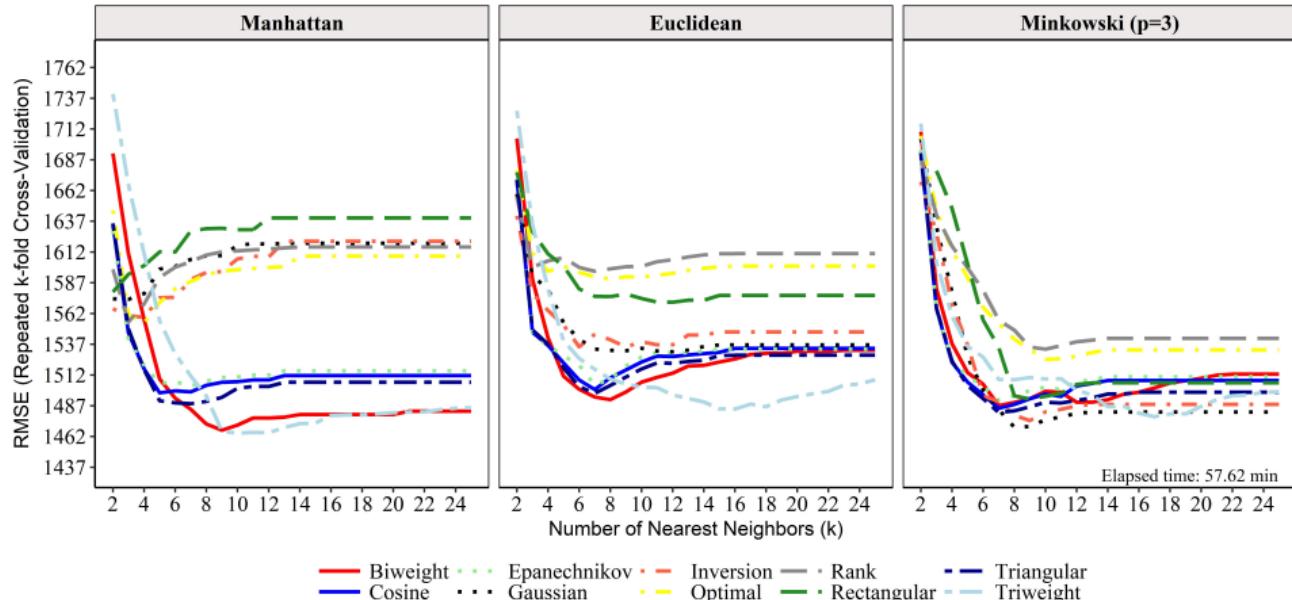
1. O uso de todas as variáveis disponíveis ($Vpred_1$), em suas formas naturais, possibilitou a construção de modelos *wkNN* **mais precisos** (menor RMSE), **menos complexos** (exigência de menor k) e com **menor variação do RMSE** no esquema 5x10-folds CV, ao mesmo tempo em que manteve um menor RMSE médio).

$$Vpred_1 \implies AGB_{est(i, j)} = f(D, H, WSG) \implies "Best\ models"$$

Variação: $V_{pred_1} \implies AGB_{est}(i, j) = f(D, H, WSG)$
Best model: 9trwg1

Weighted k-Nearest-Neighbor

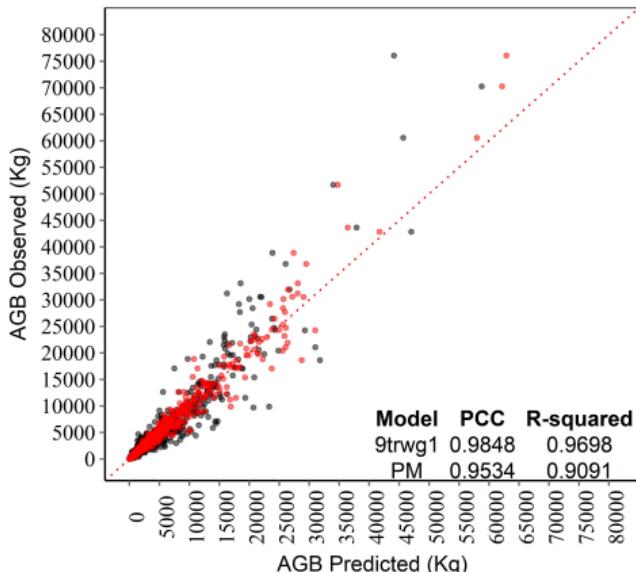
$$AGBest. = f(D, H, WSG)$$



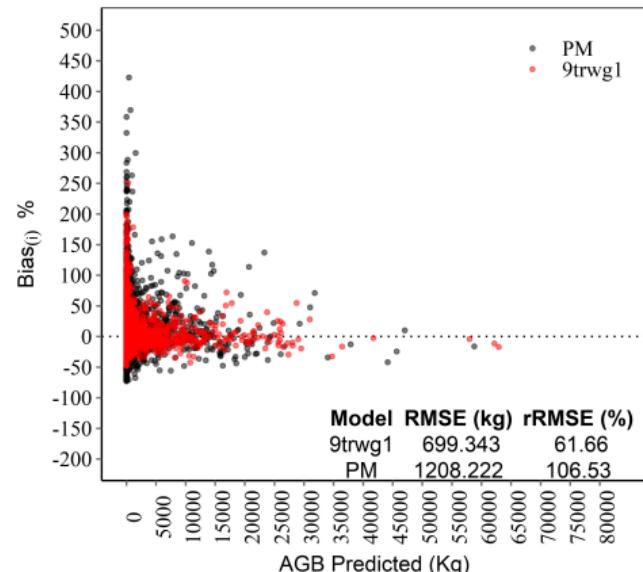
Source: The author

Pantropical Model versus Modelo 9trwg

Predicted versus Observed



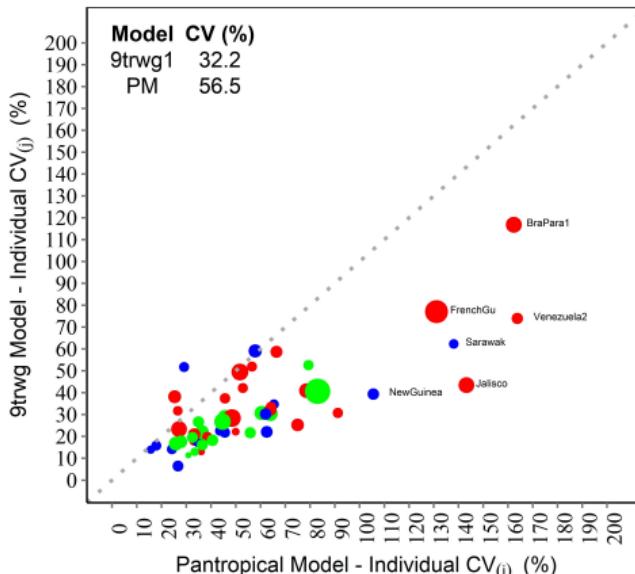
Single-tree bias



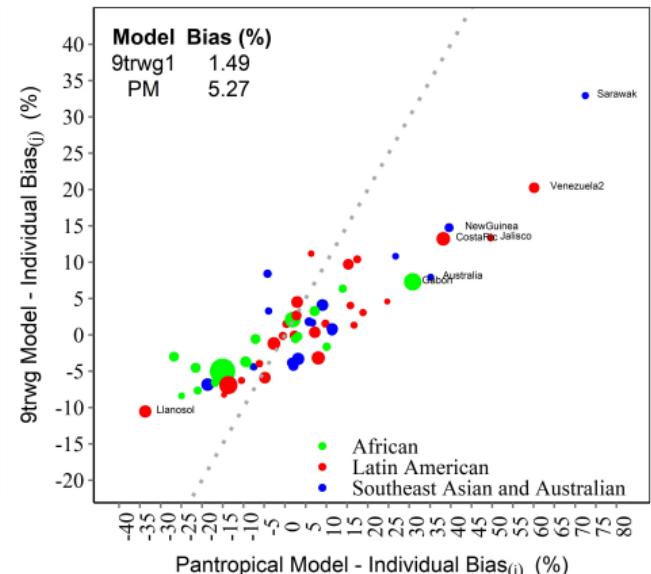
Source: The author

Pantropical Model versus Modelo 9trwg

Individual CV at each site



Individual bias at each site



Source: The author

CONSIDERAÇÃO FINAL

Estamos Entusiasmados!



Avanço: Precisamos compreender melhor e avançar na implementação de algoritmos de **Aprendizagem de Máquina** para predição de variáveis nas Ciências Florestais!



Referencial teórico

-  BATISTA, J. L.; COUTO, H. d.; FILHO, D. d. S. Quantificação de recursos florestais: árvores, arvoredos e florestas. **São Paulo: Oficina de Textos**, 2014.
-  BICHARA, J.-P.; LIMA, R. A. Uma análise da política nacional sobre mudança do clima de 2009. **Cadernos de Direito**, v. 12, n. 23, p. 165–192, 2012.
-  CERON, L. F.; PORTO, L. P. Convênio-quadro das nações unidas: Protocolo de kyoto e a política nacional sobre mudança do clima. **Revista Eletrônica do Curso de Direito da UFSM**, v. 8, p. 529–540, 2013.
-  CHAVE, J. et al. Improved allometric models to estimate the aboveground biomass of tropical trees. **Global change biology**, Wiley Online Library, v. 20, n. 10, p. 3177–3190, 2014.
-  FAO. Climate change guidelines for forest managers. **FAO Forestry Paper**, n. 172, 2013.

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

- HECHENBICHLER, K.; SCHLIEP, K. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.
- HIGUCHI, N. et al. Dinâmica e balanço do carbono da vegetação primária da amazônia central. **Floresta**, v. 34, n. 3, 2004.
- KUHN, M. **caret: Classification and Regression Training**. [S.I.], 2018. R package version 6.0-79. Disponível em: <<https://CRAN.R-project.org/package=caret>>.
- KUHN, M.; JOHNSON, K. **Applied predictive modeling**. [S.I.]: Springer, 2013. v. 810.
- REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: **Encyclopedia of database systems**. [S.I.]: Springer, 2009. p. 532–538.
- REJOU-MECHAIN, M. et al. **BIOMASS: Estimating Aboveground Biomass and Its Uncertainty in Tropical Forests**. [S.I.], 2018. R package version 1.2. Disponível em: <<https://CRAN.R-project.org/package=BIOMASS>>.

- SANQUETTA, C. R. Métodos de determinação de biomassa florestal. **As florestas e o carbono. Curitiba**, p. 119–140, 2002.
- SANQUETTA, C. R. et al. On the use of data mining for estimating carbon storage in the trees. **Carbon balance and management**, Springer, v. 8, n. 1, p. 6, 2013.
- SANQUETTA, C. R. et al. Comparison of data mining and allometric model in estimation of tree biomass. **BMC bioinformatics**, BioMed Central, v. 16, n. 1, p. 247, 2015.
- SCHLIEP, K.; HECHENBICHLER, K. **kknn: Weighted k-Nearest Neighbors**. [S.I.], 2016. R package version 1.3.1. Disponível em: <https://CRAN.R-project.org/package=kknn>.
- ZHAO, M.; CHEN, J. Improvement and comparison of weighted k nearest neighbors classifiers for model selection. **Journal of Software Engineering**, v. 10, n. 1, p. 109–118, 2016.

MACHINE
LEARNING E
BIOMETRIA
FLORESTAL

DEIVISON V.
SOUZA

BIOMASSA

OBJETIVO GERAL

METODOLOGIA

ORIGEM E
ESTRUTURA DO DATA
SET

MODELO
PARAMÉTRICO

MODELO
NÃO-PARAMÉTRICO

CONSTRUÇÃO DOS
MODELOS
PREDITIVOS

ESTIMATIVA DE
DESEMPENHO

RESULTADOS
PARCIAIS

OBRIGADO!

Deivison Venicio Souza (UFPA)

Programa de Pós-graduação em Engenharia Florestal - UFPR

Email: deivisonvs@ufpa.br