Gestão Sustentável da Produção de Manga no Vale do São Francisco por meio de Machine Learning

Wesley R. de Sousa[1], Diego S. Fonseca[1], Giovanna A. da Silva[1], Laécio A. Costa[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano Campus Petrolina, PE, Brasil

{wesley.rodrigues, diego.santos1, giovanna.alves}@aluno.ifsertao-pe.edu.br, laecio.costa@ifsertao-pe.edu.br

Abstract. *This paper evaluates five machine learning regression models — Random Forest, Gradient Boosting, MLP Neural Network, SVR, and Linear Regression, to predict mango fruit diameter from IoT-collected soil moisture data in the São Francisco Valley. Following the Design Science Research (DSR) methodology with real sensor data, Random Forest achieved the highest performance, confirming the non-linear relationship between soil moisture and fruit growth. The findings support precision irrigation in semi-arid regions and align with UN Sustainable Development Goals 2 and 12.*

Resumo. *Este artigo avalia cinco modelos de regressão por aprendizado de máquina, Random Forest, Gradient Boosting, MLP Neural Network, SVR e Regressão Linear, para prever o diâmetro dos frutos de manga a partir de dados de umidade do solo coletados via sensores IoT no Vale do São Francisco. Seguindo a metodologia Design Science Research (DSR) com dados reais de sensores, o Random Forest obteve o melhor desempenho, confirmando a natureza não linear da relação entre umidade do solo e crescimento do fruto. Os resultados apoiam a irrigação de precisão em regiões semiáridas e alinham-se aos ODS 2 e 12 da ONU.*

1. Introdução

O Vale do São Francisco, situado no semiárido nordestino do Brasil, consolidou-se ao longo das últimas décadas como um dos mais importantes pólos agrícolas do país. Graças a um conjunto de fatores favoráveis, elevada incidência solar, disponibilidade de água do rio São Francisco para irrigação e condições edafoclimáticas propícias, a região tornou-se referência nacional e internacional na produção de frutas tropicais, em especial a manga. Os estados da Bahia e Pernambuco respondem por cerca de 93% da manga nacional exportada, sendo 47,36% de origem baiana e 45,2% pernambucana, conforme dados do Anuário Brasileiro de Horti&Fruti (2024). A produção regional é destinada majoritariamente a mercados europeus e norte-americanos, onde o Brasil figura entre os principais fornecedores.

Apesar da reconhecida vocação agrícola, o Vale do São Francisco enfrenta um desafio estrutural: a escassez hídrica. O clima semiárido impõe precipitações irregulares e concentradas em poucos meses do ano, tornando a irrigação artificial indispensável para a manutenção da produção frutífera ao longo de todo o ciclo. O custo da irrigação para a cultura da manga varia entre R$ 6 mil e R$ 12 mil por hectare, a depender do nível tecnológico empregado [Revista Campo e Negócios 2022]. O manejo inadequado da lâmina d'água, seja por excesso ou por déficit, compromete diretamente o desenvolvimento dos frutos, reduz a produtividade, eleva os custos operacionais e pode provocar degradação permanente do solo, por salinização ou compactação.

Nesse contexto, o avanço das Tecnologias de Informação e Comunicação (TICs) aplicadas à agricultura tem aberto novas possibilidades de monitoramento e controle das variáveis produtivas. Segundo Elijah *et al.* (2018), a IoT oferece uma solução completa de monitoramento e tomada de decisão na agricultura, sendo uma ferramenta essencial para enfrentar desafios como as mudanças climáticas e a crescente demanda por alimentos. Esses dados, quando processados por algoritmos de Aprendizado de Máquina (Machine Learning — ML), podem alimentar modelos preditivos que antecipam o comportamento das culturas e orientam decisões de irrigação com maior precisão e menor desperdício.

A combinação de IoT e ML representa, portanto, uma oportunidade de tornar a agricultura semiárida mais eficiente, sustentável e competitiva. Modelos capazes de prever o crescimento dos frutos a partir de leituras de umidade do solo permitem ao produtor ajustar a lâmina irrigada de acordo com as necessidades reais de cada fase fenológica da mangueira, evitando tanto o estresse hídrico quanto o desperdício. Essa abordagem contribui diretamente para o ODS 12 da ONU (Consumo e Produção Responsáveis), bem como para o ODS 2 (Fome Zero e Agricultura Sustentável), ao viabilizar maior produtividade com menor impacto ambiental.

Diante desse cenário, o presente trabalho tem por objetivo apresentar os resultados de uma pesquisa que propôs, desenvolveu e avaliou modelos de ML para a predição do diâmetro de frutos de manga a partir de dados de umidade do solo coletados por sensores IoT em um pomar monitorado no Vale do São Francisco. O estudo foi conduzido entre abril e outubro de 2025 no Instituto Federal do Sertão Pernambucano (IF Sertão-PE), Campus Petrolina, em parceria com a Universidade Federal do Vale do São Francisco (Univasf). As seções seguintes apresentam os trabalhos relacionados, a metodologia adotada, os resultados obtidos e as perspectivas futuras da pesquisa.

2. Trabalhos Relacionados

A crescente pressão sobre os sistemas alimentares globais tem impulsionado a adoção de tecnologias digitais na agricultura. Projeções indicam que a produção agrícola mundial precisará aumentar em pelo menos 70% até 2050 para atender à demanda de uma população estimada em nove bilhões de pessoas [Massruhá 2020]. A IoT é definida como uma rede de objetos físicos equipados com componentes eletrônicos, sensores e conectividade de rede, capaz de coletar e trocar dados de forma autônoma [Gokhale 2018]. Na agricultura, a IoT viabiliza o monitoramento remoto de variáveis críticas como umidade do solo, temperatura e presença de pragas, sendo ferramenta essencial para enfrentar os desafios das mudanças climáticas e da crescente demanda por alimentos [Elijah *et al.* 2018].

No campo do aprendizado de máquina aplicado à agricultura, Andrade *et al.* (2023) avaliaram modelos preditivos de produtividade para culturas de uva, comparando algoritmos como PLSR, Cubist e Random Forest. Os autores constataram que modelos com variáveis meteorológicas alcançaram os melhores resultados, com o Random Forest destacando-se como o algoritmo mais robusto. O uso de dendrômetros, instrumentos que medem variações no diâmetro do caule ou dos frutos, tem sido progressivamente incorporado às pesquisas de fisiologia vegetal [Van der Maaten *et al.* 2016].

Do ponto de vista metodológico, o pré-processamento de dados constitui etapa crítica no desenvolvimento de modelos preditivos. Malley *et al.* (2016) descrevem técnicas como imputação de dados ausentes e normalização, enquanto a seleção de variáveis relevantes melhora a interpretabilidade dos modelos [Li *et al.* 2017]. A Análise Exploratória de Dados (EDA), com histogramas e boxplots, é recomendada para identificar padrões e anomalias [Larson 2014], e a validação cruzada com métricas como R^2 e MSE são ferramentas consagradas para avaliação de generalização [Raschka 2019]. No contexto da mangueira, Silva (2005) descreveu os padrões de evapotranspiração ao longo das fases fenológicas da cultura, demonstrando que a necessidade hídrica varia substancialmente com pico durante o crescimento dos frutos. Fraisse *et al.* (2022) reforçam que variáveis de umidade e temperatura viabilizam o treinamento de modelos de ML para suporte à irrigação.

3. Metodologia

3.1 Método de Pesquisa

A pesquisa foi conduzida no Instituto Federal do Sertão Pernambucano (IF Sertão-PE), no período de abril a outubro de 2025. A metodologia adotada baseou-se nos princípios do Design Science Research (DSR) [Peffers *et al.* 2018], estruturado em etapas iterativas: identificação do problema, definição dos objetivos, projeto e desenvolvimento do artefato, demonstração, avaliação e comunicação dos resultados. Os artefatos produzidos foram a base de dados tratada e os modelos preditivos de crescimento do fruto. A etapa de identificação do problema levantou os desafios do manejo hídrico na produção de manga por meio de revisão bibliográfica e interação com especialistas da área.

3.2 Infraestrutura de Coleta de Dados

Os dados utilizados neste estudo foram provenientes de sensores IoT instalados em um pomar de mangueiras em parceria com a Universidade Federal do Vale do São Francisco (Univasf). Os sensores registraram, em intervalos regulares, umidade volumétrica do solo em três profundidades (0–10 cm, 10–20 cm e 20–30 cm), utilizadas como variáveis preditoras, e o diâmetro do fruto (medido por dendrômetro), utilizado como variável-alvo. O conjunto de dados foi armazenado em formato CSV para posterior processamento.

3.3 Pré-processamento e Análise Exploratória de Dados

A base de dados bruta passou por um rigoroso processo de preparação, composto pelas seguintes etapas:

- Análise de correlação: a correlação de Spearman mostrou-se mais adequada que a de Pearson, indicando relações monotônicas não lineares entre as variáveis ambientais e o crescimento do fruto;
- Detecção e tratamento de outliers pelo método IQR, substituindo observações fora do intervalo [Q1 − 1,5×IQR; Q3 + 1,5×IQR] pelos limites do intervalo;
- Preenchimento de lacunas temporais por interpolação linear, preservando a continuidade das séries.

A análise exploratória incluiu histogramas, boxplots e mapas de calor de correlação. A profundidade de 10–20 cm apresentou a maior correlação com o crescimento do fruto, sugerindo que a umidade na zona radicular ativa é o principal fator determinante do desenvolvimento da manga nas condições estudadas.

3.4 Algoritmos e Estratégia de Modelagem

Foram avaliados cinco algoritmos de regressão supervisionada: Regressão Linear, Random Forest, Gradient Boosting, MLP Neural Network e Support Vector Regressor (SVR). Os modelos foram treinados com 80% dos dados e avaliados nos 20% restantes. A validação cruzada foi realizada com 5 folds. As métricas utilizadas foram R^2, MSE e CVR^2 com desvio-padrão. O ambiente de desenvolvimento utilizou Python 3.10, com as bibliotecas Pandas, Scikit-learn, Matplotlib e Seaborn.

4. Resultados e Discussão

4.1 Desempenho Comparativo dos Modelos

Os experimentos realizados com a base de dados de sensores IoT resultaram no conjunto de métricas apresentado na Tabela 1. Os modelos baseados em ensemble (Random Forest e Gradient Boosting) superaram consistentemente os demais algoritmos em todos os indicadores avaliados, enquanto os modelos lineares apresentaram desempenho sensivelmente inferior.

Tabela 1. Comparativo de desempenho dos algoritmos de regressão avaliados.

Modelo	R^2 Score	MSE	Validação Cruzada (CVR^2)
Random Forest	0,8785	0,1368	0,8758 (±0,0284)
Gradient Boosting	0,8732	0,1428	0,8707 (±0,0195)
MLP Neural Network	0,8635	0,1538	0,8696 (±0,0262)
SVR (Support Vector)	0,8623	0,1551	0,8691 (±0,0177)
Regressão Linear	0,7413	0,2914	0,7544 (±0,0149)

O Random Forest Regressor obteve o melhor desempenho global, com R^2 = 0,8785 e CVR^2 = 0,8758 (±0,0284), indicando que o modelo explica aproximadamente 87,85% da variância do diâmetro do fruto. O MSE de 0,1368 reforça a acurácia preditiva do modelo. O Gradient Boosting ficou em segundo lugar (R^2 = 0,8732; CVR^2 = 0,8707 ±0,0195), apresentando o menor desvio-padrão na validação cruzada, o que indica alta estabilidade entre os folds.

Os modelos MLP Neural Network (R^2 = 0,8635) e SVR (R^2 = 0,8623) apresentaram desempenho intermediário, próximo entre si, com CVR^2 em torno de 0,869. Apesar de competitivos, esses modelos exigem maior custo computacional e ajuste de hiperparâmetros mais delicado. A Regressão Linear registrou R^2 = 0,7413 e MSE = 0,2914, aproximadamente o dobro do MSE do Random Forest, confirmando a inadequação de modelos lineares para capturar a dinâmica não linear entre variáveis

ambientais e crescimento do fruto, conforme já indicado pela correlação de Spearman na fase de EDA.

4.2 Análise Visual dos Resultados

As Figuras 1 e 2 apresentam a dispersão dos valores preditos em relação aos valores reais de diâmetro do fruto para o Random Forest Regressor e para a Regressão Linear, respectivamente. Essa comparação direta entre o melhor e o pior modelo avaliados permite visualizar com clareza o impacto da escolha do algoritmo sobre a qualidade das predições. As visualizações foram geradas com as bibliotecas Matplotlib e Seaborn a partir dos experimentos conduzidos sobre o conjunto de teste.

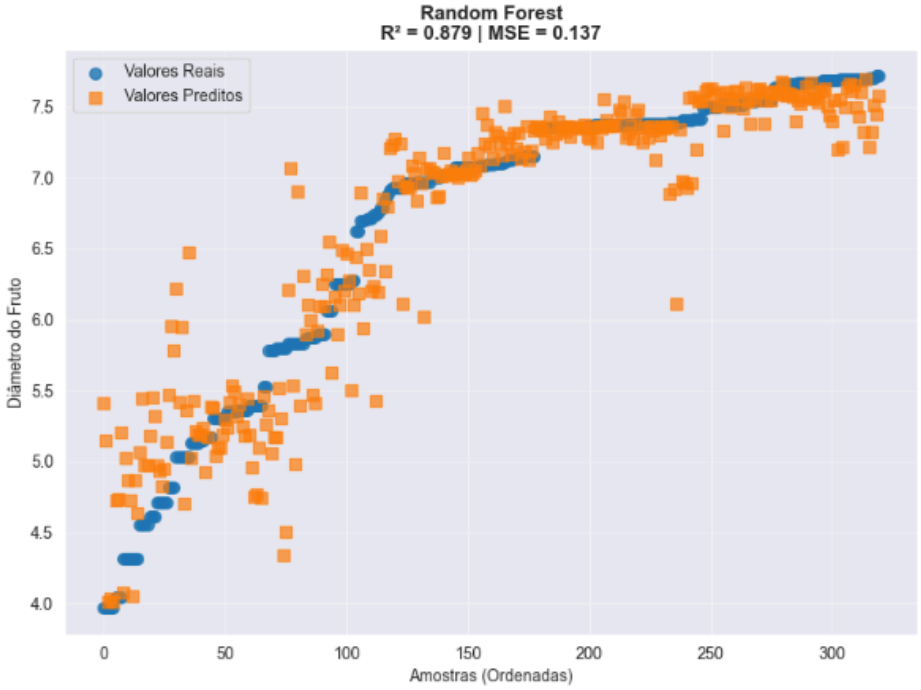

Figura 1. Dispersão dos valores preditos vs. reais — Random Forest Regressor. Fonte: elaborado pelos autores (2025).

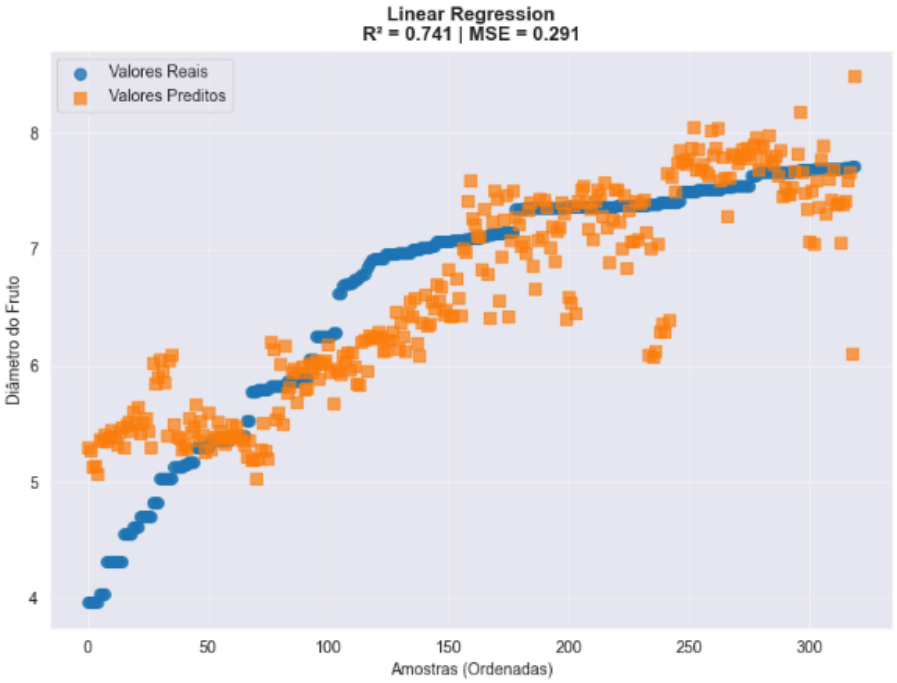

Figura 2. Dispersão dos valores preditos vs. reais — Regressão Linear.

Fonte: elaborado pelos autores (2025).

A Figura 1 ilustra a dispersão das predições do Random Forest em torno da reta de identidade, demonstrando boa aderência do modelo aos dados observados, com pontos bem concentrados ao longo da diagonal e ausência de padrões sistemáticos de erro, o que confirma a inexistência de viés significativo nas previsões. A Figura 2, por sua vez, evidencia o comportamento da Regressão Linear: a maior dispersão dos pontos em relação à reta de identidade e a presença de erros mais acentuados nas extremidades da escala revelam a dificuldade do modelo linear em capturar a dinâmica não linear do crescimento do fruto. A comparação visual entre as duas figuras reforça quantitativamente a diferença de R^2 observada na Tabela 1 (0,8785 vs. 0,7413) e justifica a adoção de algoritmos de ensemble para este domínio de aplicação.

4.3 Discussão dos Resultados

Os resultados obtidos estão em consonância com a literatura existente sobre aplicação de ML na agricultura. Andrade *et al.* (2023) identificaram o Random Forest como o algoritmo mais eficaz para a predição de produtividade de uva, conclusão que se replica neste estudo para a predição de crescimento de manga. A robustez do Random Forest frente ao overfitting, característica estrutural do algoritmo, que combina múltiplas árvores com amostras e variáveis aleatórias, justifica seu desempenho superior mesmo diante de um conjunto de dados com variabilidade temporal.

A correlação de Spearman identificada na EDA revelou que a relação entre umidade do solo e diâmetro do fruto é essencialmente monotônica mas não linear, o que explica por que os modelos lineares, mesmo com regularização, não conseguiram capturar adequadamente a dinâmica do fenômeno. Em termos fisiológicos, esse resultado é coerente com o descrito por Silva (2005): o crescimento do fruto segue uma curva

sigmoidal ao longo do ciclo, com diferentes taxas de incremento em cada fase, e a resposta hídrica da planta varia de forma não proporcional à disponibilidade de água no solo.

A estabilidade dos modelos de ensemble na validação cruzada, com desvios-padrão inferiores a ±0,03 no CVR^2, indica boa capacidade de generalização, fator crítico para aplicações em campo com dados contínuos e condições ambientais variáveis ao longo das estações. Esse resultado sugere que os modelos desenvolvidos têm potencial para serem integrados a sistemas de apoio à decisão para irrigação de precisão, sem necessidade de retreinamento frequente.

5. Considerações Finais

Este trabalho demonstrou a viabilidade e a eficácia do uso de algoritmos de Machine Learning para a predição do crescimento de frutos de manga a partir de dados de umidade do solo coletados por sensores IoT no Vale do São Francisco. Dentre os cinco algoritmos avaliados, Regressão Linear, Random Forest, Gradient Boosting, MLP Neural Network e SVR, o Random Forest Regressor destacou-se como o mais eficaz, alcançando $R^2 = 0{,}8785$, MSE = 0,1368 e $CVR^2 = 0{,}8758$ (±0,0284). A confirmação da natureza não linear da relação entre umidade do solo e crescimento do fruto reforça a importância da escolha criteriosa do algoritmo para este domínio.

Do ponto de vista metodológico, a adoção do Design Science Research (DSR) mostrou-se adequada para estruturar as etapas de construção, avaliação e refinamento dos artefatos tecnológicos produzidos. Em termos de impacto prático, os modelos desenvolvidos têm potencial para subsidiar sistemas de apoio à decisão em irrigação de precisão, permitindo que produtores do Vale do São Francisco ajustem a lâmina irrigada com base em previsões confiáveis de crescimento dos frutos. Essa aplicação pode contribuir para reduzir o custo da irrigação que varia entre R$ 6 mil e R$ 12 mil/ha e minimizar o desperdício de água em uma região onde a escassez hídrica representa um desafio estrutural crescente, agravado pelas mudanças climáticas.

Como trabalhos futuros, propõe-se: (i) validação em tempo real do modelo com dados contínuos do dendrômetro ao longo de um ciclo produtivo completo; (ii) desenvolvimento de um dashboard interativo para visualização das predições e geração automática de recomendações de irrigação para os produtores; (iii) ajuste fino de hiperparâmetros dos modelos por meio de técnicas como Grid Search e Bayesian Optimization para redução adicional do MSE; (iv) expansão do conjunto de variáveis preditoras, incorporando dados de radiação solar, velocidade do vento e nutrientes do solo; e (v) avaliação da transferibilidade dos modelos para outras cultivares de manga e regiões semiáridas com características edafoclimáticas distintas.

6. Agradecimentos

Os autores agradecem ao Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano (IF Sertão-PE) pelo apoio institucional e financeiro referente ao processo PIBIC nº 23302.100324/2025-12, e à Universidade Federal do Vale do São Francisco (Univasf) pelo fornecimento dos dados experimentais de campo utilizados na fase de modelagem com dados reais.

7. Referências

Andrade, C. B.; Moura-Bueno, J. M.; Comin, J. J.; Brunetto, G. (2023). Grape yield prediction models: Approaching different machine learning algorithms. Horticulturae, v. 9, n. 12, p. 1294. DOI: 10.3390/horticulturae9121294.

Anuário Brasileiro de Horti&Fruti. (2024). Santa Cruz do Sul: Editora Gazeta Santa Cruz, 94 p. ISSN 2107-0897. Disponível em: https://editoragazeta.com.br/wp-content/uploads/2024/04/HF_2024_DUPLAS.pdf. Acesso em: 10 mar. 2025.

Elijah, O.; Rahman, T. A.; Orikumhi, I.; Leow, C. Y.; Hindia, M. N. (2018). An overview of internet of things (IoT) and data analytics in agriculture: Benefits and challenges. IEEE Internet of Things Journal, v. 5, n. 5, p. 3758–3773. DOI: 10.1109/JIOT.2018.2844296.

Fraisse, C.; Ampatzidis, Y.; Guzmán, S.; Lee, W.; Martinez, C.; Shukla, S.; Singh, A.; Yu, Z. (2022). Artificial intelligence (AI) for crop yield forecasting. EDIS/IFAS. Disponível em: https://edis.ifas.ufl.edu/publication/AE571. Acesso em: 6 nov. 2024.

Gokhale, P.; Bhat, O.; Bhat, S. (2018). Introduction to IoT. International Advanced Research Journal in Science, Engineering and Technology, v. 5, n. 1, p. 41–44. DOI: 10.17148/IARJSET.2018.517.

Larson, R.; Farber, B. (2014). Elementary Statistics. Pearson Education UK.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys, v. 50, n. 6, p. 1–45. DOI: 10.1145/3136625.

Malley, B.; Ramazzotti, D.; Wu, J. T. (2016). Data pre-processing. In: Secondary Analysis of Electronic Health Records, p. 115–141. DOI: 10.1007/978-3-319-43742-2_12.

Massruhá, S. M. F. S. (2020). Agricultura digital: pesquisa, desenvolvimento e inovação nas cadeias produtivas. Embrapa. Disponível em: http://www.alice.cnptia.embrapa.br/alice/handle/doc/1126213.

Peffers, K.; Tuunanen, T.; Niehaves, B. (2018). Design science research genres. European Journal of Information Systems, v. 27, n. 2, p. 129–139. DOI: 10.1080/0960085X.2018.1458066.

Raschka, S.; Mirjalili, V. (2019). Python Machine Learning. Packt Publishing Ltd.

Revista Campo e Negócios. (2022). Irrigação em mangas: fundamental para a produção. Disponível em: https://revistacampoenegocios.com.br/irrigacao-em-mangas-fundamental-para-a-producao.

Silva, G. J. N. e. (2005). Manejo da indução floral da mangueira. In: Anais do I Simpósio de Manga do Vale do São Francisco. Petrolina: Embrapa Semiárido.

Van der Maaten, E.; van der Maaten-Theunissen, M.; Smiljanić, M.; Rossi, S.; Simard, S.; Wilmking, M.; Deslauriers, A.; Fonti, P.; von Arx, G.; Bouriaud, O. (2016). dendrometeR: Analyzing the pulse of trees in R. Dendrochronologia, v. 40, p. 12–16. DOI: 10.1016/j.dendro.2016.06.001.