



Applying data science and machine learning to create a new methodology of *Eucalyptus* productivity zoning

Clayton Alcarde Alvares^{1,2}

Ítalo Ramos Cegatta¹

Henrique Ferraço Scolforo¹

Reginaldo Gonçalves Mafía¹

¹Suzano SA, Limeira-SP (calcarde@suzano.com.br), ²UNESP-FCA, Botucatu-SP

ABSTRACT: *Foresters and resource managers often face difficulties in surveying reliable Eucalyptus productivity levels for the purpose of purchasing and prospecting lands. Data science and machine learning (ML) provide powerful approaches to making the best use of the large datasets available today. Based on more than 12,000 forest inventory plots of Eucalyptus plantations and climate, soil and altitude variable associated to them, a supervised decision tree (DT) approach was adjusted to model the stands productivity. The high quality of the model obtained made it possible to spatialize the DT rules providing a reliable cartographic visualization of the Eucalyptus productivity for a major region of forest-based industries in Brazil. Our study provided deeper analysis and reproducible tools to manage forest assets sustainably.*

Keywords: *forest inventory, decision tree algorithm, model tuning, yield gap*

Introduction

Currently, most forest-based companies hold huge datasets of forest inventory and environmental monitoring with meteorological series, and increasingly accurate soil and relief maps. However, in most cases those tree growth and environment datasets are collected and used for purely operational purposes, despite foresters knowing such data remain latent for deeper applications to generate insights and other more powerful predictions.

Given the massive amount of data, available computational power and balanced empirical and scientific knowledge, an opportunity arises to use data science and machine learning concepts to improve the *Eucalyptus* productivity zoning.

In the last 4 years, machine learning usage for farm risk management is tripled (Ghaffarian et al., 2022). Agriculture has made great use of these approaches (Almeida et al., 2021), and in this study, we explore how forestry can benefit as well.

Material and methods

The study area consists of approximately 11.4 million hectares located in southeastern Brazil, covering nearly 46% of the State of São Paulo (Figure 1A). Approximately 12,000 permanent forest



inventory plots of the Suzano's *Eucalyptus* plantations were available for this study. Each plot was populated with predictor environmental information as aridity index (weather data came from company and <https://mapas.inmet.gov.br>), altitude (<https://www.earthdata.nasa.gov/sensors/srtm>), and soil order and texture (Rossi, 2017) (Figure 1).

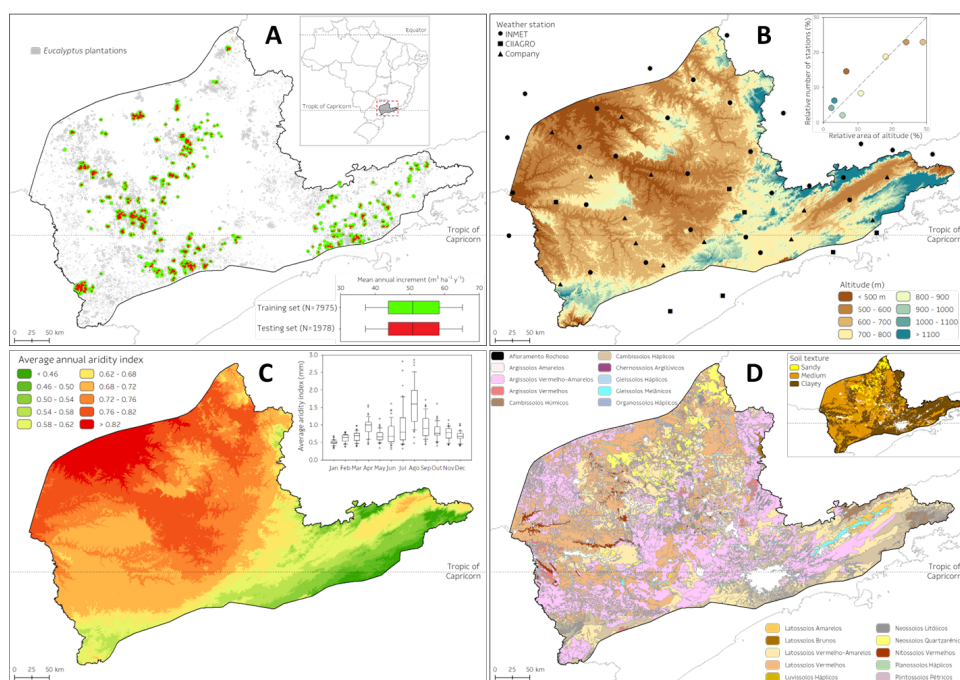


Figure 1. Distribution of *Eucalyptus* plantations (light gray), and the network of *Eucalyptus* permanent forest inventory plots used for training and testing decision tree modelling (A). Altitude and the network of weather station (B). Aridity index obtained is ratio of annual potential evapotranspiration and annual rainfall (C). Semi detailed soil map surveyed at scale 1:100,000 in the study area.

The plots were randomly divided into two datasets. The first, referred to as the training dataset, with 80% of all plots, was used in DT modelling exercise in order to build the productivity zones for planted *Eucalyptus* forests. Another group, called testing dataset, with 20% remaining plots, was used to perform the validation of the productivity zones produced by the DT.

We used the rpart package (Therneau et al., 2015) and its routines for R software (<https://www.r-project.org>) to predict the productivity zones. A broad multi-tuning of the DT hyperparameters was planned. Once the DT model was defined and its best-fitting hyperparameters were setup, a prediction was made transforming and spatializing each leaf node on productivity zones for the study area based on DT rules.



Results and discussion

Hyperparameters tuning approach yielded 450 DT models, with RMSE ranging from $60.1 \text{ m}^3 \text{ ha}^{-1}$ for worst model with three leaf nodes to $53.3 \text{ m}^3 \text{ ha}^{-1}$ for best model with 61 leaf nodes. A scatter plot with all tuning modelling outputs allowed to visualize a global exponential decay pattern between RMSE and number of leaf nodes each DT model (Figure 2A).

The validation of DT model showed that our algorithm had an optimum performance to predicting the existing testing dataset throughout study area (Figure 2B). By retro-applying the DT rules to the training data and classifying it at each productivity zone we were able to prospect the distribution and productivity probabilities in each of them. The observed productivities of each zone were adjusted to the normal distribution and the percentiles 15th, 50th, and 85th identified in the sophisticated ridgeline plot (Figure 2C).

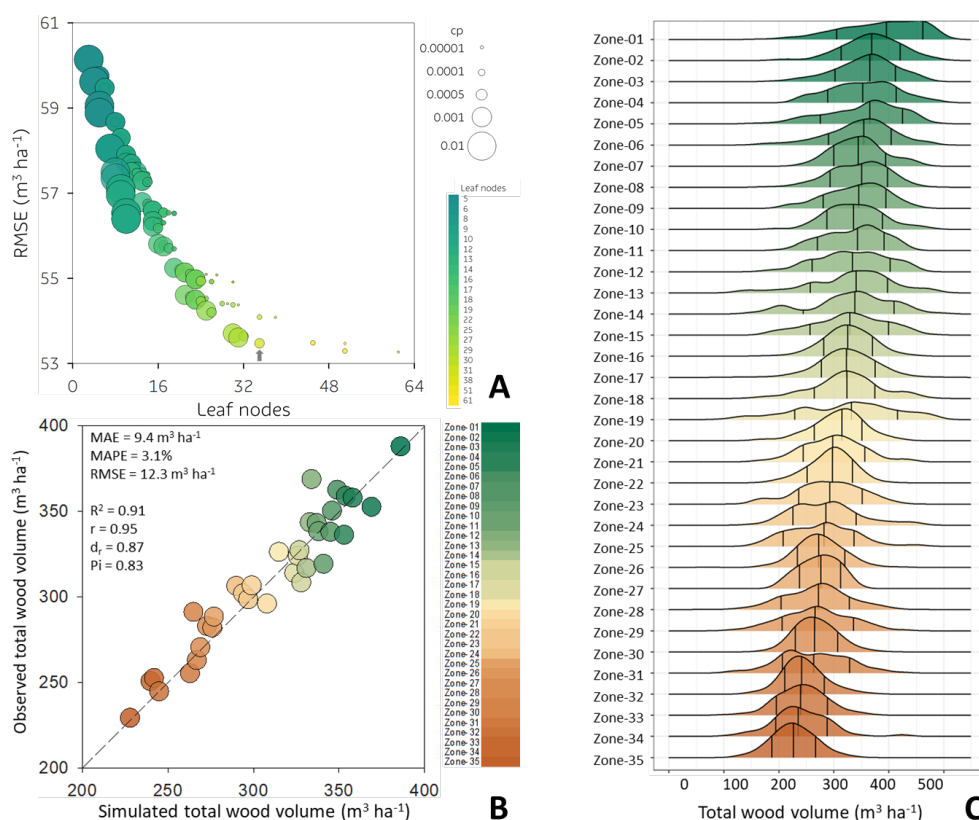
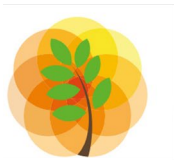


Figure 2. Relationship between RMSE (root mean squared error) and leaf (terminal) nodes for the 450 decision tree models obtained by tuning hyperparameters (A). Validation of decision tree model by comparison the simulated average Eucalyptus total wood volume at 6-years-old of each leaf node against testing dataset independent data (B). Distribution of Eucalyptus total wood volume at 6-years-old from the training dataset for each *Eucalyptus* forest zone (C).

The outputs from the DT were spatialized for the study area providing a reliable cartographic visualization of probability levels of the Eucalyptus productivity for the study area (Figure 3). The



ten most productivity zones summarized 2,380,000 ha and are in the wetter and higher altitude landscapes of the study area. The ten lowest productivity zones totaled 4,289,007 ha and are mostly spread in the center and northern region of the study area.

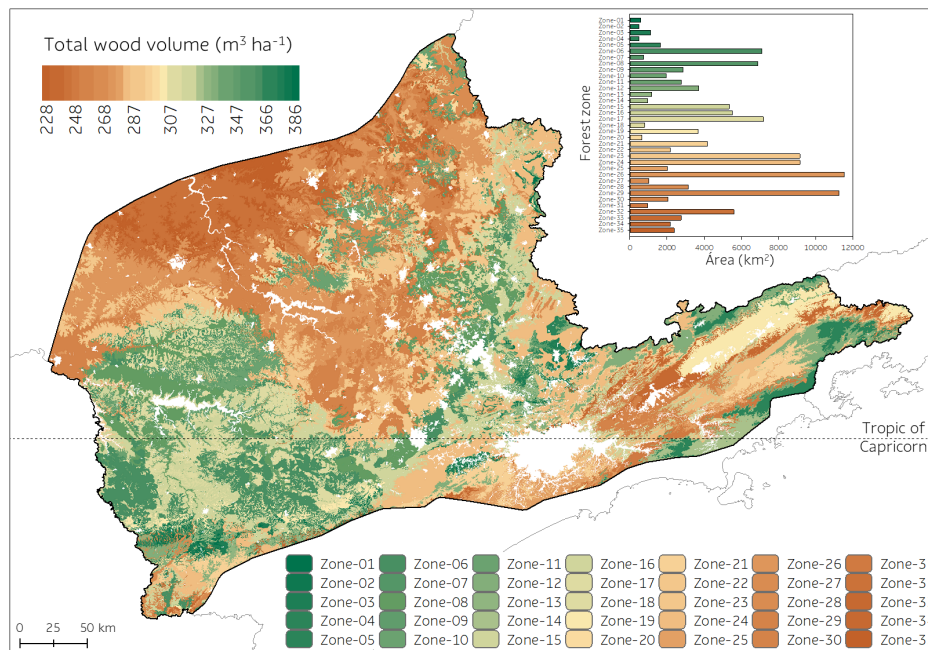


Figure 3. Decision tree' output spatialized for the study area. Leaf nodes was ranked in decreasing order (from highest to lowest productivity), named as Zone-01 till 35.

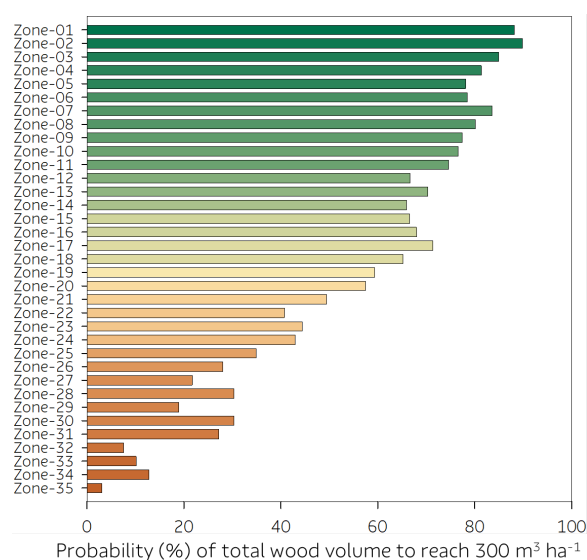


Figure 4. Probability of total wood volume to reach 300 m³ ha⁻¹ at 6-years-old for each zone.

A stakeholder prospecting lands to leasing for growing eucalypt forest whose baseline



production is planned to be reach $300 \text{ m}^3 \text{ ha}^{-1}$ at 6 years ($50 \text{ m}^3 \text{ ha}^{-1} \text{ y}^{-1}$) has a following question: What is the probability that threshold will be reached at each zone? The answers with the probabilities are in Figure 4 and this is one of the significant leap forwards provided by our study. The Zone-01, Zone-02, Zone-03, and Zone-04 have more than 80% probability of productivity reaching threshold $300 \text{ m}^3 \text{ ha}^{-1}$, while for the Zone-32, Zone-33, Zone-34, and Zone-35 the probability dropped to less than 15%.

We assuming that all current 781,654 ha (Figure 1A) of *Eucalyptus* plantations within the study area were 6-years-old, the trade secret here was applying the DT for every hectare to predict by classifying it as each one of the productivity zones. We estimated the *Eucalyptus* wood stock in the study region was approximately $2.44 \times 10^8 \text{ m}^3$ for a 50% probability. Based on 15th percentiles, there was an 85% probability the wood stock was greater than $1.99 \times 10^8 \text{ m}^3$. Considering the 85th percentile threshold as a proxy for attainable *Eucalyptus* productivity at each zone, the wood stock at this productivity level by amount to $2.86 \times 10^8 \text{ m}^3$. We made a yield gap analysis using a very pragmatic approach. The difference between the attainable and average productivities of wood stocks of current *Eucalyptus* plantations can be considered as the yield gap, and that huge amount sounds like $4.2 \times 10^7 \text{ m}^3$. The sum of each zone's yield gap ratios by its attainable productivity (85th percentile) is few more than 113,000 ha, or 15% of current forest base.

Conclusions

Our study appears to be a complex approach, but it is necessary to perceive that both data science and machine learning, supported by ecophysiological knowledges, will be from now on day-to-day tools to support make small and big decisions for cash crop business, as well as for Brazil's fast-growing *Eucalyptus* plantations for industrial supply.

References

- ALMEIDA, G.M.; PEREIRA, G.T.; SOUZA BAHIA, A.S.R.; FERNANDES, K; JÚNIOR, J.M. Machine learning in the prediction of sugarcane production environments. *Computers and Electronics in Agriculture*, v.190, 106452, 2021.
- GHAFFARIAN, S.; VAN DER VOORT, M.; VALENTE, J.; TEKINERDOGAN, B.; DE MEY; Y. Machine learning-based farm risk management: A systematic mapping review. *Computers and Electronics in Agriculture*, v.192, 106631, 2022.
- ROSSI, M. Mapa pedológico do Estado de São Paulo: revisado e ampliado. Instituto Florestal: São Paulo, SP, Brasil, 2017. 118p.
- THERNEAU, T.; ATKINSON, B.; RIPLEY, B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. 2015.

