2026 International Nuclear Atlantic Conference – INAC 2026
Escola de Guerra Naval, Rio de Janeiro, RJ, August 24 - 28, 2026

ISBN: 978-65-272-2398-6

Random Forest Applied to the Classification of Wood Species Using XRF, FTIR, and Raman Spectroscopy

T. Zanin[1], M. M, Redígolo[1], and C. B. Zamboni[1]

[1]*IPEN/CNEN-SP, Brazil*
thaleszanin@usp.br
marcelo.redigolo@alumni.usp.br
czamboni@usp.br

1. Introduction

According to data from the National Forest Information System (SNIF), Brazil exports large volumes of timber products annually, including sawn timber, plywood, and logs [1]. However, illegal logging remains a major environmental and economic challenge, especially in the Brazilian Amazon. Recent reports by the Environmental Investigation Agency (EIA) describe schemes involving fraud and the international trade of illegally sourced timber destined for European and North American markets [2]. In this context, the development of rapid and reliable methods for identifying wood species is essential to support enforcement actions and the traceability of forest products. Conventional identification methods generally involve anatomical or chemical analyses that require sample preparation, the use of reagents, and specialized labor [3]. In contrast, spectroscopic techniques offer advantages such as rapid analysis, minimal sample destruction, and the potential for portable field applications. Energy-Dispersive X-Ray Fluorescence (EDXRF) provides information on the elemental composition of samples, while Raman and FTIR spectroscopy allow for the evaluation of structural and chemical differences associated primarily with lignin, cellulose, and hemicellulose [4-6]. These characteristics make these techniques promising for differentiating plant species through their spectral signatures. In recent years, machine learning methods have been widely applied in spectral data analysis due to their ability to process large sets of variables and identify complex patterns. Among these methods, the Random Forest (RF) algorithm stands out for its robustness and ability to handle high-dimensional data [7]. Recent studies demonstrate promising results using machine learning combined with spectroscopic techniques for the identification of wood species [5,8]. Thus, the objective of this study was to evaluate the application of the Random Forest algorithm in the classification of wood species from the Amazon biome using spectral data obtained by EDXRF, FTIR, and Raman, while also investigating the potential of combining spectral data to improve the performance of species discrimination.

2. Methodology

Twenty-two wood species from Rolim de Moura, Rondônia, within the Amazon biome, were analyzed. The samples were first dried and ground into a homogeneous powder for spectroscopic analysis.

XRF analyses were performed using an X-123SDD spectrometer (Amptek), consisting of an X-ray tube with silver (Ag) and gold (Au) targets, as well as a silicon drift detector (25 mm^2 × 500 μm) with a 12.5 μm beryllium window. Measurements were performed in triplicate using an acquisition time of 300 s, a current of 15 μA, and a voltage of 30 kV to ensure adequate resolution and sensitivity [9,10]. Instrument calibration was performed using the NIST 1573a certified reference material (Tomato Leaves).

The FTIR spectra were obtained using a Bruker Invenio S spectrometer coupled with a Specac Quest ATR module. The instrumental conditions used were: a resolution of 4 cm^{-1}, 64 scans for sample and background, a spectral range of 4000 to 400 cm^{-1}, and a DLaTGS detector. The instrument was calibrated using a certified polystyrene standard film. The acquired spectra were subsequently preprocessed for use in statistical models.

Raman analyses were performed using a Horiba XploRA Plus confocal Raman microscope. The instrumental parameters used were: spectral range from 40 to 4000 cm^{-1}, acquisition time of 10 s, 5 accumulations, 532 nm laser, 25% filter, 1200 lines/mm diffraction grating, 100 µm slit, and 300 µm aperture. Instrument calibration was performed using acetylsalicylic acid reference material.

The spectra obtained by XRF, FTIR, and Raman were used as input variables for supervised classification using the Random Forest (RF) algorithm. Each spectrum was converted into a feature vector containing intensity values as a function of energy (XRF) or wavenumber (FTIR and Raman). The data underwent preprocessing steps, including normalization, baseline correction, and removal of noisy spectral regions. The datasets from each technique were evaluated individually, as well as a data fusion approach combining XRF, FTIR, and Raman. The dataset was divided into training and test subsets using cross-validation. Model performance was evaluated using classification accuracy, confusion matrix, and spectral variable importance analysis.

3. Results and Discussion

An exploratory chemometric approach was used to investigate the multivariate chemical differentiation among samples. A fused dataset composed of elemental concentrations, spectroscopic variables (peak maxima) and tree species families was created to evaluate the ability of multimodal analytical information to discriminate wood samples based on their chemical fingerprints. Taxonomic family information was retained only for interpretation of the score plots and clustering patterns, and was not included as an active variable in the PCA/HCA models.

A total of 22 wood samples representing 22 different species were analyzed. Because only one specimen was available for each species, the study was designed as an exploratory chemometric investigation rather than a supervised classification study. In this context, Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were employed to evaluate natural sample grouping patterns and identify the variables contributing most strongly to multivariate differentiation.

Prior to analysis, all variables were standardized using Z-score scaling to place the XRF, Raman, and FTIR variables on comparable scales.

Principal Component Analysis biplot was performed on the scaled fused dataset to reduce dimensionality and investigate the major sources of variance within the data matrix (Fig. 1). The first principal componente (PC1) explained 31.29% of the total variance, while the second principal component (PC2) explained 28.18%. Together, the first two components accounted for 59.47% of the total variance.

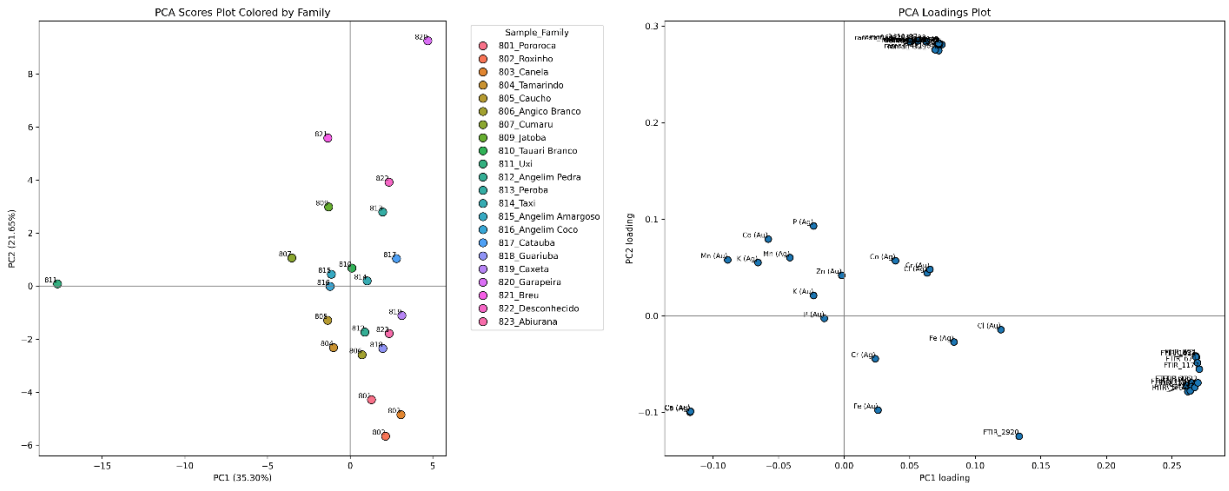

Figure 1: PCA biplot and PCA Loadings statistical analysis

The loadings plot indicates grouping patterns among variables originating from the different analytical techniques. The Raman variables formed a compact cluster with high positive loadings on PC2, indicating strong internal correlation among Raman bands. This behavior is consistent with the sensitivity of Raman spectroscopy to lignin, cellulose and aromatic structural vibrations. The FTIR variables also exhibited grouped behavior, but occupied a distinct region of the PCA space compared to the Raman variables. This pattern suggests that FTIR identified molecular information that was partly distinct from the Raman variables. Several FTIR peaks displayed strong positive contributions to PC1, suggesting that these variables were important drivers of overall sample differentiation. In contrast, XRF elemental variables were more dispersed throughout the loading space, indicating more heterogeneous behavior among elemental concentrations. This pattern suggests that different elements contributed independently.

Among the elemental variables, Ca showed one of the most distinct loading positions, suggesting that its concentration may represent an important discriminating factor. Other elemental variables, such as Fe, Cr and Cl, also contributed to the principal components.

Hierarchical Cluster Analysis (HCA) was performed using Ward linkage and Euclidean distance in order to investigate similarity relationships among the wood samples (Fig. 2).

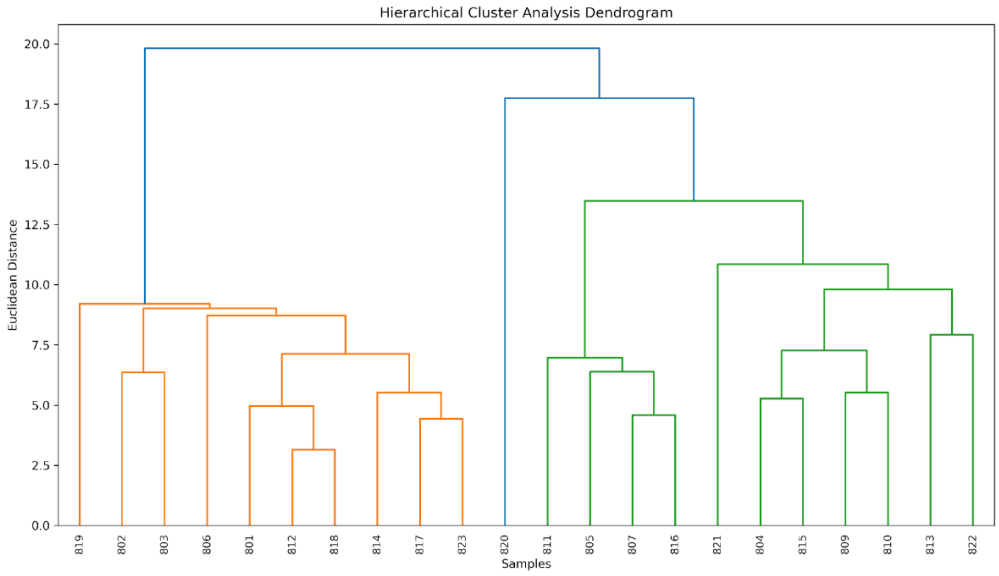

Figure 2: Hierarchical Cluster Analysis Dendrogram

Two major chemometric groups can be observed within the dataset, indicating substantial multivariate differentiation among the analyzed woods. The large linkage distance separating these two clusters suggests significant chemical differences between the two major groups. Several pairs of samples presented high similarity, forming compact subclusters with relatively low linkage distances. Examples include samples 812–818, 817–823 and 805–807. These relationships indicate similar fused spectroscopic and elemental profiles among these woods.

Sample 820 exhibited a relatively isolated position within the dendrogram, corroborating its separation observed in the PCA scores plot. The similar behavior of sample 820 in both PCA and HCA supports the interpretation that this specimen differs chemically from most of the dataset.

4. Conclusion

The fused XRF, Raman, and FTIR dataset revealed an exploratory structure among the 22 wood specimens. The first two PCA components explained 59.47% of the total variance, while HCA visualization showed consistent grouping patterns, including the distinct behavior of sample 820. The spectroscopic variables tended to behave as correlated molecular blocks, whereas the elemental variables

varied more independently across samples. These results demonstrate the relevance of multimodal data fusion as an exploratory approach and support future studies using larger, replicated datasets suitable for supervised classification.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2022/12732-5. As well as the National Nuclear Energy Commission (CNEN).

References

[1] Serviço Florestal Brasileiro (SFB). *Florestas do Brasil 2024: Dados e Estatísticas*. Sistema Nacional de Informações Florestais (SNIF), 2024.

[2] Environmental Investigation Agency (EIA). *Tricks, Traders and Trees: How Illegal Logging Drives Forest Crime in the Brazilian Amazon and Feeds EU and U.S. Markets*. 2025.

[3] Krasznai, D. J.; Champagne Hartley, R.; Roy, H. M.; Champagne, P.; Cunningham, M. F. Compositional analysis of lignocellulosic biomass: conventional methodologies and future outlook. *Critical Reviews in Biotechnology*, v. 38, n. 2, p. 199–217, 2018.

[4] Agarwal, U. P. Analysis of Cellulose and Lignocellulose Materials by Raman Spectroscopy: A Review of the Current Status. *Molecules*, v. 24, n. 9, 2019.

[5] Jesus, E.; França, T.; Calvani, C.; et al. Making wood inspection easier: FTIR spectroscopy and machine learning for Brazilian native commercial wood species identification. *RSC Advances*, v. 14, p. 7283–7289, 2024.

[6] Makarem, M.; Lee, C. M.; Kafle, K.; et al. Probing cellulose structures with vibrational spectroscopy. *Cellulose*, v. 26, n. 1, p. 35–79, 2019.

[7] Breiman, L. Random Forests. *Machine Learning*, v. 45, p. 5–32, 2001.

[8] Shugar, A. N.; Drake, B. L.; Kelley, G. Rapid identification of wood species using XRF and neural network machine learning. *Scientific Reports*, v. 11, 17533, 2021.

[9] Potts, P. J. et al., "Atomic spectrometry update: X-ray fluorescence spectrometry," *Journal of Analytical Atomic Spectrometry*, vol. 19, pp. 1397–1419 (2004).

[10] Amptek Inc., "DPPMCA Display & Acquisition Software" (2019). Available at: https://www.amptek.com/software/dp5-digital-pulse-processor-software/dppmca-display-acquisition-software