DANA 4800 – Project 1A – Part I: Write-Up or Report

1. Background Information (5 pts)

The paper chose "The global distribution and drivers of wood density and their impact on forest carbon stocks," published in Nature Ecology & Evolution. It was picked to study, because it explores how ecological data can be analyzed using machine learning models like random forests. The authors use these techniques to understand and predict global patterns in wood density, which is important for estimating forest carbon stocks. The paper is found relevant since it deals with real-world data science challenges, such as working with a Global scale dataset, integrating different types of environmental data, and building predictive models. It also demonstrates good practices in reproducible research and data management, which are essential skills in data science. Overall, the paper provides useful insights into applying AI in ecology, especially for climate change research.

Publication Date: October 15, 2024
DOI: 10.1038/s41559-024-02564-9
Link : [The global distribution and drivers of wood density and their impact on forest carbon stocks | Nature Ecology & Evolution]

2. **Methodology**

The study combines wood density data from the Global Wood Density Database with environmental variables sourced from global climate and soil datasets, such as WorldClim and SoilGrids. The dataset covers thousands of tree species sampled globally, providing a rich, heterogeneous set of observations.

Data cleaning involved removing missing or inconsistent records and standardizing species taxonomy to ensure data quality and comparability. Geographic coordinates were used to match wood density observations with corresponding climate and soil variables.

The core analysis used **random forest models**, a machine learning algorithm capable of modeling nonlinear relationships and complex interactions without requiring explicit functional assumptions. The model was trained on a subset of the data and validated using cross-validation to prevent overfitting and ensure generalizability.

Key environmental predictors included mean annual temperature, precipitation seasonality, and soil texture variables. The model output was evaluated using standard performance metrics such as the coefficient of determination (R^2) and root mean square error (RMSE).

Spatial predictions were then generated by applying the trained model to global environmental layers, creating maps of predicted wood density worldwide. This spatially explicit modeling approach enables more accurate assessments of forest carbon stocks at regional and global scales.

3. Executive Summary

Wood density is a fundamental trait influencing forest biomass and carbon storage, critical components in global climate regulation. This study provides a comprehensive global assessment of wood density distribution and its environmental drivers, leveraging machine learning to analyze extensive ecological and environmental data.

The research team compiled a large, harmonized dataset consisting of wood density measurements from diverse forest ecosystems worldwide, spanning numerous tree species. To contextualize these data, environmental variables including temperature, precipitation, and soil properties were integrated from high-resolution global datasets.

Using **random forest models**, the authors identified key environmental factors influencing wood density variation. Climate variables such as mean annual temperature and precipitation seasonality emerged as significant drivers, along with soil texture characteristics. The results revealed that tropical forests typically exhibit lower wood density, whereas temperate and boreal forests tend to have higher values.

This spatial variability in wood density has substantial implications for estimating forest carbon stocks. Traditional models often assume constant wood density values across regions, which can introduce bias and under- or overestimate carbon storage. By generating spatially explicit wood density predictions, the study enhances the accuracy of biomass and carbon assessments at multiple scales.

The paper also highlights the utility of machine learning techniques in ecological research, demonstrating their ability to integrate heterogeneous datasets and capture complex ecological patterns. The publicly available R code repository ensures reproducibility, enabling researchers and practitioners to apply similar methods to other ecological datasets.

In conclusion, the study advances our understanding of global wood density patterns and provides a robust, scalable framework to improve forest carbon stock estimation. This contributes valuable insights for climate change mitigation efforts by refining key parameters used in carbon accounting models.

Appendix

The R code supporting these analyses is openly available on GitHub (https://github.com/LidongMo/GlobalWoodDensityProject), which includes scripts for data cleaning, modeling, and visualization, facilitating reproducibility and transparency.

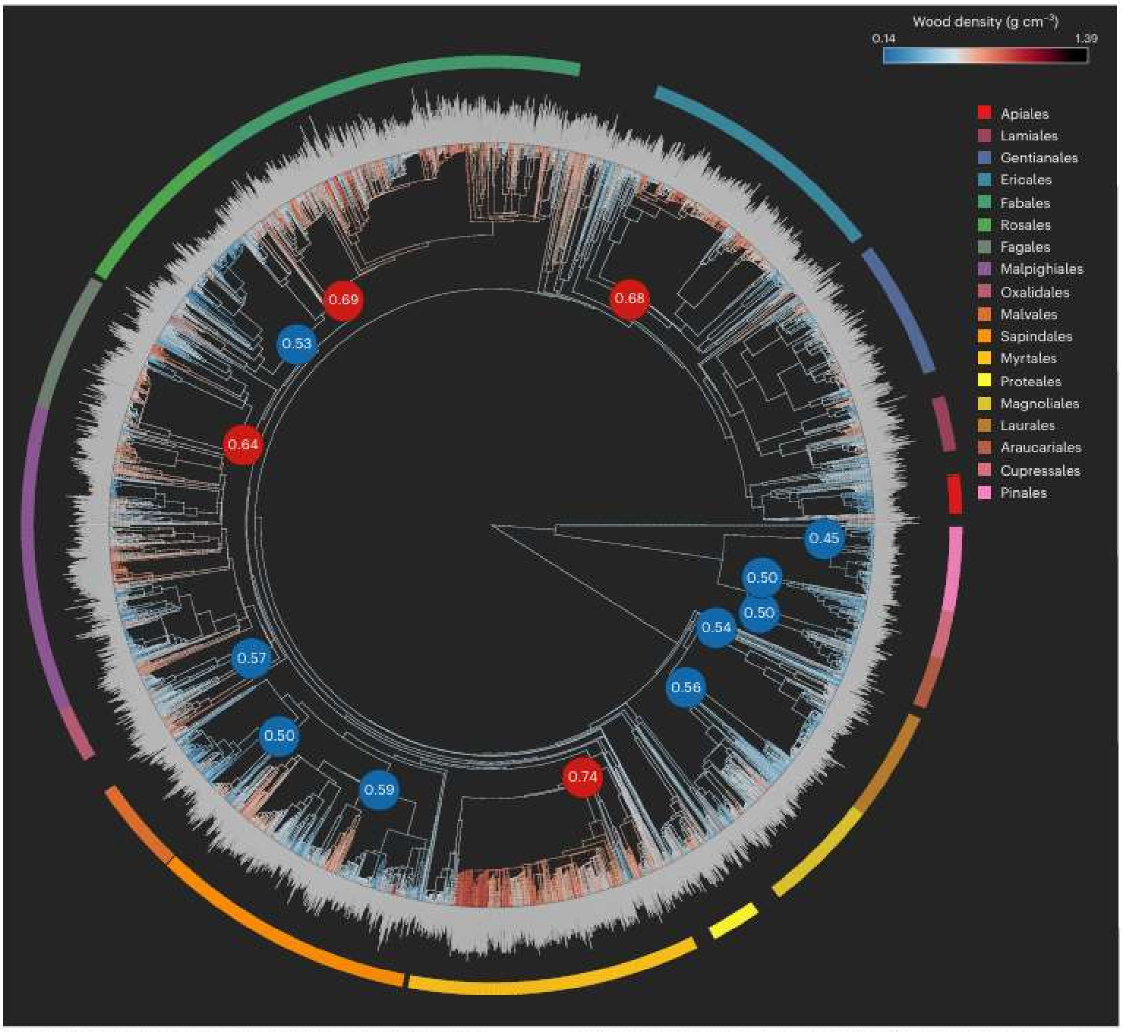

Fig. 2 | Phylogenetic tree and wood density information of global tree species. The phylogenetic tree was constructed using the R package

all other tree species, we conducted a two-tailed significance test by comparing the order-level wood density with 999 randomized wood density values from

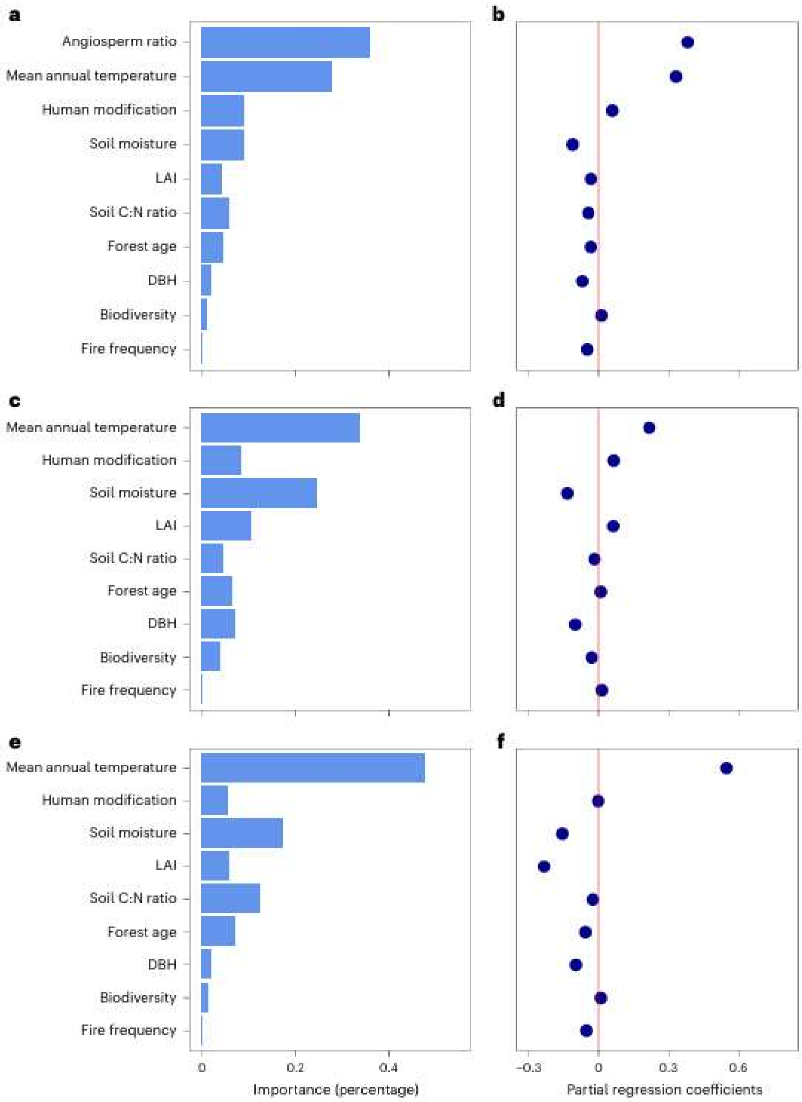

Fig. 4 | Variable importance of the selected environmental metrics. a–f, The environmental metrics are based on random-forest models (**a**,**c**,**e**) and linear partial regression models (**b**,**d**,**f**). **a**,**b**, Variable importance of the selected covariates across global forests, including angiosperm ratio to control for wood density differences between angiosperms and gymnosperms. **c**,**d**, Variable importance within angiosperm-only communities. **e**,**f**, Variable importance within gymnosperm-only communities. Mean decrease in accuracy values in **a**, **c** and **e** represents the relative contribution of each variable to CWD variation, whereby we averaged the values of 100 bootstrapped random-forest models. Bootstrapped partial regression coefficients for each variable (**b**,**d**,**f**) were calculated by averaging the partial regression coefficients from 100 multivariate linear models. All variables were standardized to allow for direct effect size comparison. In addition, we quantified the absolute effects of these covariates using partial regression analysis, as detailed in Supplementary Table 5.

Data Set

the next links get the data used for the paper

- ✓ Code repository https://github.com/LidongMo/GlobalWoodDensityProject
- ✓ Data pictures https://zenodo.org/records/13331493
- ✓ Data csv https://zenodo.org/records/14054077

References:

- Mo, L., Crowther, T. W., Maynard, D. S., van den Hoogen, J., Ma, H., Bialic-Murphy, L., Liang, J., de-Miguel, S., Nabuurs, G. J., Reich, P. B., Phillips, O. L., Abegg, M., Adou Yao, Y. C., Alberti, G., Almeyda Zambrano, A. M., Alvarado, B. V., Alvarez-Dávila, E., Alvarez-Loayza, P., Alves, L. F., ... Zohner, C. M. (2024). The global distribution and drivers of wood density and their impact on forest carbon stocks. *Nature Ecology & Evolution*, *8*(12), 2195–2212. https://doi.org/10.1038/s41559-024-02564-9Wageningen University & Research