

BERN01: Ecosystem modelling with random forests

Joseph A. Binns (jo4752bi-s@lu.se)

Group 15

Contents

1	Introduction	1
2	Methodology	1
2.1	Binary biome classification	2
2.1.1	Hyperparameters	2
2.1.2	Performance metrics	2
2.2	Multi-biome classification	2
2.2.1	Hyperparameters	2
2.2.2	Performance metrics	2
2.3	Regression	2
2.3.1	Hyperparameters	2
3	Results and discussion	3
3.1	Binary biome classification	3
3.2	Multi-biome classification	3
3.3	Regression	4
3.3.1	NPP	4
3.3.2	Vegetation carbon	4
4	Conclusion	4
5	References	5

1 Introduction

This project uses random forests to predict biome classes, net primary productivity (NPP) and vegetation carbon (VegC) from climate and soil variables [1]. We compare models trained on global data (excluding India), with a focus on their ability to predict these ecosystem properties within India.

2 Methodology

At discrete longitudes and latitudes of global landmass, we were provided with seasonal climate statistics (temperature, precipitation and shortwave radiation) and soil properties, each representing an input feature for our models. The corresponding LPJ-GUESS biome classification, NPP and vegetation carbon outputs served as the targets for classification and regression [2]. The modelling workflow builds on Group 15's Python codebase [3].

We used SHAP values to compare the importance of each feature to the model's prediction in an attempt to inform our feature selection. SHAP values are calculated by averaging a feature's effect across all subsets of the other features using Shapley weights [4].

We compare three feature sets applied to each task. Firstly, the exhaustive set is our baseline, including all climate and soil features. Secondly, the theoretical set is loosely motivated by Holdridge's 'Life Zones' model, and includes only the seasonal means of the temperature and precipitation [5]. Thirdly, the culled set is determined independently for each task after hyperparameter tuning by ranking

features by their mean absolute SHAP values and discarding the half deemed least important.¹

We performed limited hyperparameter tuning for each task and feature set whilst keeping the number of trees fixed at 100. We used scikit-learn [6] and compared its default random forest hyperparameters with two alternatives: increasing the minimum number of samples per leaf to four, or considering half the input features at each split. For rapid hyperparameter tuning, we used three-fold cross-validation on a restricted subset of 12,000 training cells outside India. We chose the hyperparameters with the highest average F1 for binary classification, the highest average macro F1 (continue reading) for multi-biome classification, and the lowest average mean squared error for regression.

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad \text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN}. \quad (1)$$

2.1 Binary biome classification

2.1.1 Hyperparameters

Set	Feature split	Min. leaf size
Exhaustive	6	1
Culled	4	1
Theoretical	4	1

2.1.2 Performance metrics

Split	Train / test	Majority accuracy	Accuracy	Macro F1	Weighted F1
Global random	47367 / 11824	0.935	0.977	0.898	0.976
India random	876 / 236	0.847	0.928	0.856	0.927
India holdout	58079 / 1112	0.855	0.875	0.586	0.832

baseline model

Majority accuracy: accuracy when predicting the most common training class. **Macro F1:** F1 uniformly averaged over all classes. **Weighted F1:** F1 averaged over all classes, weighted by each class's number of test samples.

2.2 Multi-biome classification

2.2.1 Hyperparameters

Set	Feature split	Min. leaf size
Exhaustive	6	1
Culled	11	1
Theoretical	4	1

2.2.2 Performance metrics

Split	Train / test	Majority accuracy	Accuracy	Macro F1	Weighted F1
Global random	47367 / 11824	0.186	0.877	0.785	0.875
India random	876 / 236	0.208	0.725	0.642	0.721
India holdout	58079 / 1112	0.018	0.441	0.342	0.378

baseline model

2.3 Regression

2.3.1 Hyperparameters

NPP

Set	Feature split	Min. leaf size
Exhaustive	23	1
Culled	11	1
Theoretical	4	1

Vegetation carbon

Set	Feature split	Min. leaf size
Exhaustive	23	1
Culled	11	1
Theoretical	4	1

We evaluate NPP and vegetation carbon predictions using mean squared error (MSE), the mean of the squared differences between predictions and LPJ-GUESS values. Squaring the errors gives large mistakes more weight, making MSE useful for highlighting the greatest prediction errors. As the units of NPP and vegetation carbon differ, we use normalized MSE (NMSE) to compare performance between the regression tasks.

$$NMSE = \frac{MSE}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}. \quad (2)$$

¹My methodology strays slightly from the group's approach as I wanted to further explore feature culling based on SHAP values and the theoretical feature set in my individual report.

3 Results and discussion

3.1 Binary biome classification

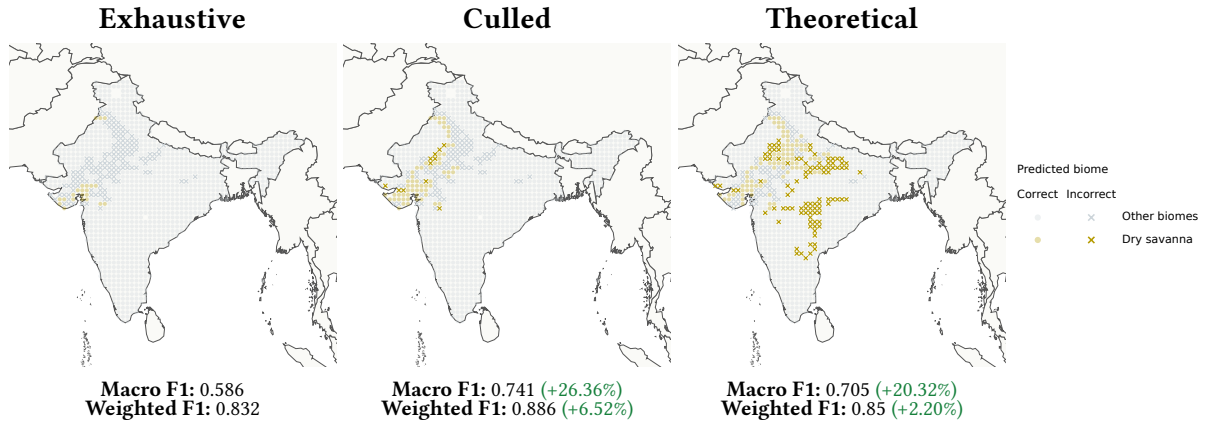


Figure 1: Binary biome classification in India, colored by prediction. Percentage changes are relative to the exhaustive model.

We observed that the exhaustive model performed poorly at predicting the dry savanna biome, favouring predictions of the other class, which occurs more frequently. This may be partially explained by dry savanna’s underrepresentation in the training data. On the other hand, the theoretical model actually overpredicted dry savanna. Savannas are characterized by seasonal rainfall and prolonged dry seasons [7], which led me to expect better performance from the theoretical feature set. However, whilst seasonal rainfall variables were ranked as the most important features, soil pH and organic carbon also ranked highly in the SHAP ranking used for culling. This suggests that climate data alone may be insufficient to distinguish between these more subtle biome categories, which are not explicitly named in Holdridge’s climatic ‘Life Zones’ model.

3.2 Multi-biome classification

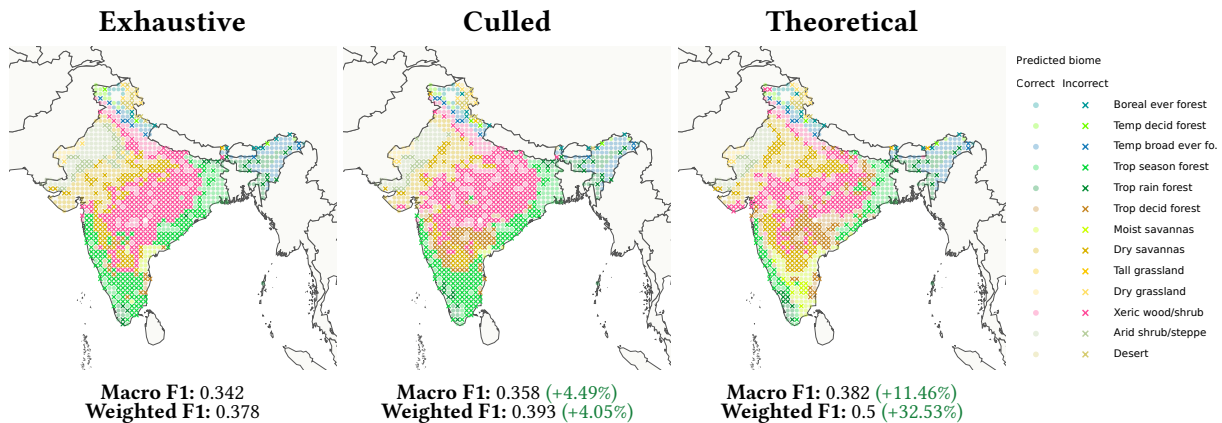


Figure 2: Multi-biome classification in India, colored by prediction. Percentage changes are relative to the exhaustive model.

For multi-biome classification, the theoretical model performed best on both reported F1 measures. The theoretical model did a particularly good job at predicting the moist savannas, something which the exhaustive and culled models struggled to do. Some of the theoretical model’s relative success may be attributed to the curse of dimensionality: India’s biome combinations are unevenly represented in the rest of the world, resulting in insufficient training data being diluted across many features. This may explain why simpler feature sets transferred well to this task.

3.3 Regression

3.3.1 NPP

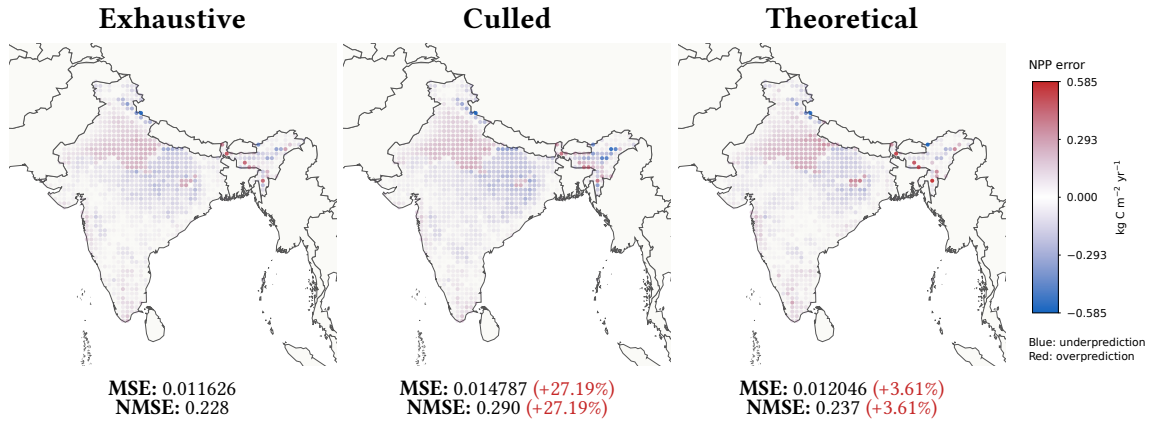


Figure 3: NPP prediction errors in India. Percentage changes are relative to the exhaustive model.

Interestingly, we found that the exhaustive model outperformed both the culled and theoretical models when predicting NPP. Sunlight and nutrient availability constrain plant growth and therefore NPP [8]. As such, removing these features appears to have outweighed the benefits of reduced dimensionality in this task. Finally, prediction errors appear concentrated in mountainous northern India. It seems plausible that such mountainous regions are poorly represented in the training data.

3.3.2 Vegetation carbon

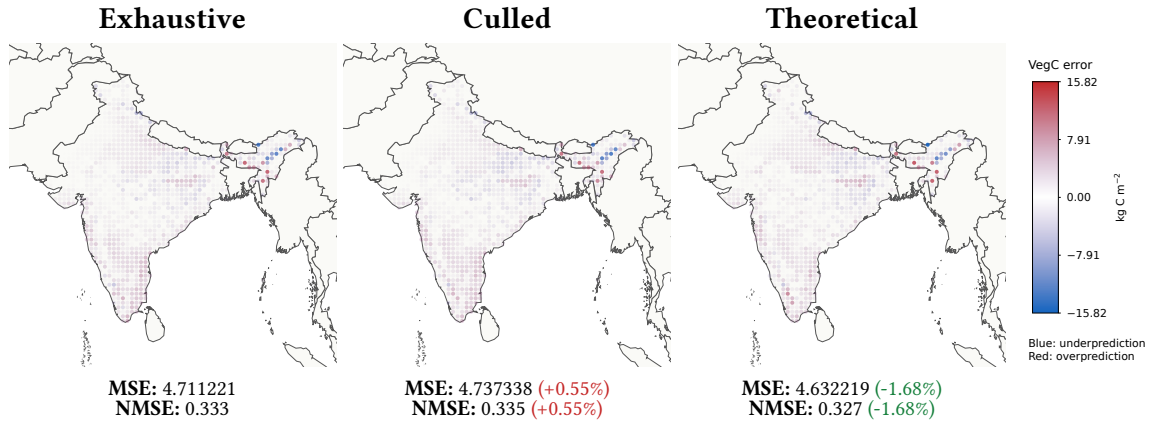


Figure 4: VegC prediction errors in India. Percentage changes are relative to the exhaustive model.

The culled and theoretical models showed much smaller relative changes in MSE when predicting vegetation carbon than when predicting NPP. As each model was only trained once with a single fixed random seed, it could be that the uncertainty in our results accounts for the small relative changes in performance we see between these models. Vegetation carbon predictions had a higher NMSE than NPP predictions, which may be explained by vegetation carbon accumulating over decades, whereas NPP varies on timescales better reflected by the seasonal data.

4 Conclusion

The theoretical model performed best among the tested models in multi-biome classification. However, precipitation and temperature alone were less effective than the culled feature set in distinguishing the subtly defined dry savanna biome. SHAP-based culling did not respect how similar features may dilute one another and ultimately offered no consistent improvement across tasks. These results suggest that reducing the curse of dimensionality often improves performance, but only when the simplification is tailored to the task based on a theoretical understanding.

5 References

- [1] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [2] B. Smith *et al.*, “Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model,” *Biogeosciences*, vol. 11, no. 7, pp. 2027–2054, 2014, doi: [10.5194/bg-11-2027-2014](https://doi.org/10.5194/bg-11-2027-2014).
- [3] J. Binns, M. Bleckmann, and P. Stover, “Group 15: BERN01 ecosystem modelling codebase.” 2026.
- [4] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, 2017. [Online]. Available: <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [5] L. R. Holdridge, “Determination of World Plant Formations From Simple Climatic Data,” *Science*, vol. 105, no. 2727, pp. 367–368, 1947, doi: [10.1126/science.105.2727.367](https://doi.org/10.1126/science.105.2727.367).
- [6] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>
- [7] I. N. Forseth, “Terrestrial Biomes,” *Nature Education Knowledge*, vol. 3, no. 10, p. 11, 2010, [Online]. Available: <https://www.nature.com/scitable/knowledge/library/terrestrial-biomes-13236757/>
- [8] H. Riebeek, “The Carbon Cycle.” [Online]. Available: <https://science.nasa.gov/earth/earth-observatory/the-carbon-cycle/>