nature portfolio

https://doi.org/10.1038/s41586-024-08106-4

Supplementary information

Global potential for natural regeneration in deforested tropical regions

In the format provided by the authors and unedited

Supplementary Information 1. Extended methodology and rationale for modelling approach

The primary purpose of this model is to facilitate the estimation of the probability of natural regeneration of tropical forests. The model is estimated/trained based on a sample of locations where natural regeneration did and did not occur. As such the dependent variable is binary. The suite of covariates selected represent a range of biophysical and socioeconomic variables (see Predictor Variables section in Methodology) that we anticipate are related to natural regeneration.

The statistical modelling and validation is based on a set of 6 million random points generated within the regeneration and non-regeneration spatial domains defined above, stratified so that each level of the dependent variable was equally represented but with no other form of spatial stratification. The regeneration points were randomly sampled as spatially random points generated within the forest regeneration polygons using the 'sf' library in R (function 'st sample')[1]. The non-regeneration points were sampled as spatially random points within latitude limits of +/- 25 degrees using the 'random points GCS' function in R, which accounts for the elliptical shape of Earth, and then discarding points falling outside of the target biomes (tropical and subtropical forest biomes), falling within excluded land cover categories (water, bare areas, urban areas, sparse vegetation; ESA CCI Landcover for year 2000), or falling within the regeneration or forestry polygons delineated in the Fagan and colleagues 2022[2] dataset. These random points were then intersected with the geospatial covariate datasets (see Predictor Variables section in Methodology). Any records having NoData values for any of the covariates were removed, resulting in a final set of 5.8 million records. NoData values can arise due to mismatches in the coverage of datasets, such as along coastlines.

Several modelling frameworks could be adopted. We choose Random Forests (RF), a classification tree machine-learning method, because of the high classification accuracies it commonly achieves while having lower risk of over-fitting compared to other potential algorithms (e.g. LightGBM), and because it has previously been applied to developing spatial predictions of the potential for natural regeneration[3]. Although RF models are fairly robust to correlations among independent variables, having many strongly correlated variables in the model can result in the model being unduly influenced by those factors. It is good practice, therefore, to evaluate correlations among variables explicitly (calculate Pearson's r correlation coefficients among all pairs of continuous variables) and either select a subset of any strongly correlated variables or use principal components analysis (PCA) to combine strongly correlated variables. We identified strong correlations among the 19 bioclimatic variables, which is expected as some of the variables are derived from others. Using PCA this dataset was reduced to 5 uncorrelated variables that capture 99.4% of the variation among the 19 variables (see Predictor Variables).

Although RF inherently includes 'out-of-bag' cross validation to reduce the potential for over-fitting, we also validated the model against data that was not used to train the model. Training and validation datasets were drawn without replacement, with equal numbers of records for each level of the binary dependent variable, from the pool of 5.8 M records described above. RF models can take considerable time to fit with large datasets. To identify an appropriate sample we fit ten models to different random samples of 500,000 records, fit RF models to each one, and examined variance in the accuracy and variable importance (decrease in accuracy and the Gini coefficient) among the models. There is a stochastic component to RF models so some level of variation is expected. However, there was little variation in either the accuracy (mean: 0.886, range:

0.885 - 0.887) or the variable importance among these models. We therefore used a sample of 500,000 records for the variable selection process, which involves fitting many RF models, and a sample of 1 million records to fit the final model.

Variable selection methods were used to identify a more parsimonious model and to reduce the computational burden associated with generating high-resolution predictions at a global scale. This was achieved using variable importance (mean decrease in accuracy metric) to rank the full set of variables, fitting a series of models of increasing complexity (adding one variable at a time in order of importance), and identifying the point at which the addition of further variables did not improve accuracy (Extended Data Fig. 1).

We contrasted models that included both socioeconomic and biophysical variables to models that included biophysical variables alone. The accuracy of both models was similar (0.886 and 0.880 respectively). We based the spatial predictions on the model containing biophysical variables only for two reasons. First, identifying areas having biophysical conditions amenable to natural regeneration regardless of whether socioeconomic conditions are favourable to regeneration is most closely aligned with the intended use of the predictions to inform forest restoration planning. The premise of this rationale is that socioeconomic factors can be altered or mitigated using a variety of policy and management interventions (e.g. payments for ecosystem services, legislation governing land management, etc), while biophysical conditions are more difficult to alter. Hence, the focus of this analysis is on identifying areas where there is the potential for natural regeneration if socioeconomic factors limiting regeneration are managed.

The second reason we favour the model with biophysical variables only is that it is difficult to make predictions using some socioeconomic variables. Specifically, we train our models based on recent historical natural regeneration using data from 2000, but make predictions of the potential for natural regeneration based on contemporary data from 2018. Although time-series of socioeconomic data are available there is often a trend in the data: in many areas there is a monotonic increase in factors such as population density and GDP. This is problematic as it can result in making model predictions based on data that fall outside of the range of values upon which the model was based. This can result in substantial errors.

The ten top-ranked covariates in the biophysical model included local forest density; distance to nearest forest; two soil variables (soil organic carbon density and soil pH); four bioclimatic variables (the first four axes of a principal components analysis based on all 19 bioclimatic variables); land use/cover; and biome (Extended Data Fig. 2).

Plots of each covariate against the predicted probability of natural regeneration (Extended Data Fig. 3) provide insight into the marginal effect of a variable on the class probability. The potential for natural regeneration as positively associated with forest density and soil organic carbon content, and negatively associated with distance to forest. There is a more complex relationship with soil pH and the relationship with the bioclimatic variables is more opaque because these variables are the axes of a principal components analysis representing combinations of all 19 original bioclimatic variables.

Predictions of the potential for natural regeneration were made using the spatial datasets representing the covariates in the final model. However, three of the variables were updated to reflect contemporary (2018) conditions: revised forest density and distance to forest covariates were derived from the Global Forest Watch 2018 tree cover dataset [4] at

a 30m resolution, and the revised land user/cover dataset was derived from the combination of the 2015 ESA CCI dataset with the aforementioned tree cover data. The result of this process is a raster dataset with cell values in the range [0, 1], representing the probability of natural regeneration within that cell. These data are made available in the following forms: (i) 30m resolution binary predictions (2.3 GB), (ii) 30m resolution continuous probabilities expressed as a percentage and rounded to integers to reduce file size (9.3 GB), and (iii) an approximately 1 km resolution representing the proportion of each cell that has the potential for natural regeneration. The 30 m resolution data were tiled into 10 degree lat/lon tiles.

A limitation of our analysis is that the combination of datasets with different resolutions may degrade the quality of the predictions. While some of the key variables were mapped at a 30 m resolution (tree cover, distance to forest, and forest density), the best available data on soils, climate and land cover variables were mapped at coarser resolutions (Appendix 2). Of these, we speculate that the lack of land cover and land use data at the native 30 m resolution may prevent identifying important associations between opportunity cost (especially in the context of agricultural and pastoral production) and the potential for natural regeneration.

The majority of geoprocessing in this analysis used open source libraries in R and Python (the ‘osgeo’, ‘numpy’ and ‘multiprocessing’ Python packages[5], and the ‘sp’, ‘raster’ and ‘sf’ libraries in R[1]), with some additional processing and map-making in ESRI ArcGIS[6].

Validation

The validation points were divided into 16 0.5-km distance intervals based on the distance from each validation point to the nearest training point and accuracy was estimated within each interval to examine evidence for autocorrelation issues. Accuracy

exceeded 95% for the points in the 0-0.5 km interval (Extended Data Fig. 4a), declined to 81.4% in the 2.0-2.5 km interval, and then increased again. This pattern was driven by differences in the classification accuracy between the binary outcomes (from Fagan and colleagues 2022 – an observed historic subset of data) and variation in the relative frequency of the outcomes as a function of distance to training points (Extended Data Fig. 4b). We therefore also quantified accuracy with balanced samples of outcomes in each distance interval using a bootstrapping approach (Extended Data Fig. 4c). The bootstrapping was necessary because there are relatively few occurrences of certain outcomes in some intervals. For example, in the 6-8 km intervals where the frequency of regeneration samples is low (Extended Data Fig. 4b), the bootstrapped accuracy estimates were based on only 5,000 samples. The estimates of model accuracy using this approach are lower (Extended Data Fig. 4c) but never worse than 75%. Regardless of how accuracy was quantified, the model performed sufficiently well to be deemed useful for our application.

Clustering analysis

To check for potential spatial bias in the Fagan and colleagues 2022[2] regrowth patches, we conducted a clustering analysis. We first generated 10,000 bootstrapped samples from the full regrowth validation dataset from Fagan and colleagues 2022[2], using the same observed number of omission and commission error points in each bootstrap sample. Then, separately for the omission and commission bootstrap samples, we assessed potential spatial clustering of error by measuring nearest neighbour distances using the nndist tool from the spatstat package 7 in R8. These nearest-neighbour analyses indicated that the mean of nearest-neighbour distances for the original, observed omission error (6.38) is significantly more distant than nearest-neighbour distances for the full, bootstrapped regrowth validation dataset (4.26, CI 5/95% of means 3.09/5.75 (10,000 bootstrapped samples using the same number of omission error points)). In

other words, the omission errors are more spread out, on average, than by random chance. This indicates that the Fagan and colleagues 2022[2] model was on average more spatially accurate across different regions than would be expected by random chance, i.e. omission error was less clustered than by random chance. Further, we found that the mean nearest-neighbour distance for the original commission error (5.11) is similar to the means of the full, bootstrapped regrowth dataset (5.98, CI 5/95% of mean 0.51/9.30 (10,000 random samples using the same number of commission error points)) indicating that there is no clustering present (i.e. patterns are random).

To check for potential spatial bias in the Fagan and colleagues 2022[2] regrowth patches, we conducted a clustering analysis. We first generated 10,000 bootstrapped samples from the full regrowth validation dataset from Fagan and colleagues 2022[2], using the same observed number of omission and commission error points in each bootstrap sample. Then, separately for the omission and commission bootstrap samples, we assessed potential spatial clustering of error by measuring nearest neighbour distances using the nndist tool from the spatstat package[7] in R[8]. These nearest-neighbour analyses indicated that the mean of nearest-neighbour distances for the original, observed omission error (6.38) is significantly more distant than nearest-neighbour distances for the full, bootstrapped regrowth validation dataset (4.26, CI 5/95% of means 3.09/5.75 (10,000 bootstrapped samples using the same number of omission error points)). In other words, the omission errors are more spread out, on average, than by random chance. This indicates that the Fagan and colleagues 2022[2] model was on average more spatially accurate across different regions than would be expected by random chance, i.e. omission error was less clustered than by random chance. Further, we found that the mean nearest-neighbour distance for the original commission error (5.11) is similar to the means of the full, bootstrapped regrowth dataset (5.98, CI 5/95% of mean 0.51/9.30 (10,000 random samples using the same

number of commission error points)) indicating that there is no clustering present (i.e. patterns are random).

To check for potential bias the Fagan and colleagues 2022[2] regrowth patches, we conducted a clustering analysis. We used 10,000 bootstrapped samples from the full regrowth validation dataset from the Fagan et al. 2022[2] dataset using the same number of omission and commission error points to assess nearest neighbour distances using the nndist tool from the spatstat package[7] in R[8]. These nearest-neighbour analyses indicated that the mean of nearest-neighbour distances for the observed omission error (6.38) is significantly more distant than nearest-neighbour distances for the full regrowth validation dataset (4.26, CI 5/95% of means 3.09/5.75 (10,000 bootstrapped samples using the same number of omission error points)). In other words, the omission errors are more spread out, on average, than by random chance. This indicates that the Fagan et al. 2022 model was on average more spatially accurate across different regions than would be expected by random chance, ie. omission error was less clustered than by random chance. Further, we found that the mean nearest-neighbour distance for commission error (5.11) is similar to the means of the full regrowth dataset (5.98, CI 5/95% of mean 0.51/9.30 (10,000 random samples using the same number of commission error points)) indicating that there is no clustering present (ie. patterns are random).

To test for potential correlations between the spatial environmental variables used in our potential for natural regeneration analysis and the validation dataset from the Fagan and colleagues 2022[2] analysis we used glm from the R base library[8] to conduct binomial logistic regressions. We found that the occurrence of omission errors in the validation data displays very little apparent correlation with a number of spatial environmental variables as most variables were non-significant (14/15) apart from slope ($p = 0.001^{**}$). This is to be expected because, as Fagan and colleagues 2022[2] highlight, their product was derived from both optical and microwave data and as such local variation in topography and forest structure affected classification accuracy - their

product tended to have lower accuracy in mountainous areas (where microwave returns vary). Similarly, we found that the occurrence of commission errors in the validation data displays very little apparent correlation with the same spatial environmental variables as most were non-significant (14/15) apart distance to urban areas (p = 0.001**). Synthetic aperture radar (a key input to the dataset of Fagan and colleagues 2022[2]) has elevated and variable values in urban areas[9–11], decreasing the likelihood of predicting forest in these regions. We note that urban areas are a very small percentage of the humid tropics, and therefore not an issue for the present study. We can therefore conclude that there is minimal detectable spatial bias in the Fagan and colleagues 2022[2] model's producer's accuracy. In this study, our spatial predictive model includes topography and distance to urban areas as predictors, so it is able to correct for this minimal bias in omission and commission error observed in the input data.

Confusion matrix for Fagan and colleagues 2022

Table A1.1. Confusion matrix depicting the independent model accuracy assessment for the humid tropics, and for each subregion (TP = tree plantations, REG = natural regrowth, NG = no-gain, NGP = no gain possible), using only reference polygons predicted with medium to high confidence. Percent accuracy estimates are presented for overall accuracy (mean +/- confidence interval), user's accuracy (100-commission error), and producer's error (100-omission error). Confusion matrix counts represent both the number of polygons, and the area-weighted estimates.

Map Accuracy Testing Dataset							
Humid tropical biome			**Reference data points (n=2981)**				
Predicted	*TP*	*REG*	*NG*	Total	Map Area (Mha)	*% User Acc.*	*Corrected Area (Mha)*
TP	160	12	1	173	9.9	*90.7 ±6.0*	*25.0 ±6.0*
REG	18	150	6	174	4.7	*85.1 ±5.6*	*21.6 ±6.1*
NG	25	28	1432	1485	875.4	*98.2 ±0.5*	*1711.0 ±8.5*
NGP	0	0	1149	1149	867.5		
Total	203	190	2588			***OA %***	

% Prod. Acc.	*78.8 ±5.6*	*78.9 ±5.8*	*99.7 ±0.20*	***97.0 ±0.5***
% Prod. Acc. (area-based)	*35.8 ±8.7*	*18.7 ±5.4*	*99.99 ±0.01*	***98.1 ±0.5***

Confusion matrices for potential for natural regeneration model

Table A1.2. Confusion matrix depicting model accuracy assessment for the potential for natural regeneration (NREG = No potential, REG = Potential) globally. Percent accuracy estimates are presented for overall accuracy (mean +/- confidence interval), user's accuracy (100-commission error), and producer's error (100-omission error). Confusion matrix counts represent both the number of points, and the area-weighted estimates.

Confidence intervals for area-based estimates - Global						
Area available for restoration			**Reference data points (n=1,000,000)**			
Predicted	*NREG*	*REG*	Total	Map Area (Mha)	*% User Acc.*	*Corrected Area (Mha)*
NREG	420,160	79,840	500000	1,974	*84 ±0.0001*	*1,677 ±2.0107*
REG	41,884	458,116	500000	215	*92 ±0.0001*	*512 ±2.0107*
Total	462,044	537,956			***OA %***	
% Prod. Acc.	*90.9 ±0.102*	*85.2 ± 0.077*			***87.8 ±0.064***	
% Prod. Acc. (area-based)	*99 ±0.0001*	*38 ±0.004*			***84.8 ±0.001***	

Table A1.3. Confusion matrix depicting model accuracy assessment for the potential for natural regeneration (NREG = No potential, REG = Potential) for the Neotropics. Percent accuracy estimates are presented for overall accuracy (mean +/- confidence interval), user's accuracy (100-commission error), and producer's error (100-omission error). Confusion matrix counts represent both the number of points, and the area-weighted estimates.

Confidence intervals for area-based estimates - Neotropics

Area available for restoration	**Reference data points (n=1,000,000)**					
Predicted	*NREG*	*REG*	Total	Map Area (Mha)	*% User Acc.*	*Corrected Area (Mha)*
NREG	151177	33753	184930	1020.52	82 *±0.002*	*843 ±1.801*
REG	19225	187061	206286	98.78	91 *±0.001*	*276 ±1.801*
Total	170402	220814	391216		***OA %***	
% Prod. Acc.	*88.7 ±0.176*	*84.7 ±0.125*			***86.5 ±0.107***	
% Prod. Acc. (area-based)	*99 ±0.0002*	*32 ±0.006*			**82.5 *±0.002***	

Table A1.4. Confusion matrix depicting model accuracy assessment for the potential for natural regeneration (NREG = No potential, REG = Potential) for the Afrotropics. Percent accuracy estimates are presented for overall accuracy (mean +/- confidence interval), user's accuracy (100-commission error), and producer's error (100-omission error). Confusion matrix counts represent both the number of points, and the area-weighted estimates.

Confidence intervals for area-based estimates - Afrotropics						
Area available for restoration	**Reference data points (n=1,000,000)**					
Predicted	*NREG*	*REG*	Total	Map Area (Mha)	*% User Acc.*	*Corrected Area (Mha)*
NREG	192,266	13,006	205,272	323.93	*94 ±0.001*	*306 ±0.345*
REG	11,023	91,434	102,457	25.5	*89.2 ±0.002*	*43 ±0.345*
Total	203,289	104,440	307,729		***OA %***	
% Prod. Acc.	*94.6 ±0.105*	*87.5 ±0.1897*			***92.2 ±0.0948***	
% Prod. Acc. (area-based)	*99 ±0.0002*	*53 ±0.008*			***93.3 ±0.001***	

Table A1.5. Confusion matrix depicting model accuracy assessment for the potential for natural regeneration (NREG = No potential, REG = Potential) for Indomalaysia. Percent accuracy estimates are presented for overall accuracy (mean +/- confidence interval), user's accuracy (100-commission error), and producer's error (100-omission error). Confusion matrix counts represent both the number of points, and the area-weighted estimates.

Confidence intervals for area-based estimates - Indomalaysia

Area available for restoration	**Reference data points (n=1,000,000)**					
Predicted	*NREG*	*REG*	Total	Map Area (Mha)	*% User Acc.*	*Corrected Area (Mha)*
NREG	76,717	33,081	109,798	629.19	*70 ±0.003*	*445 ±1.7104*
REG	11,636	179,621	191,257	90.86	*93.9 ±0.001*	*275 ±1.7104*
Total	88,353	212,702	301,055		***OA %***	
% Prod. Acc.	*86.8 ±0.271*	*84.4 ±0.107*			***85.1 ±0.127***	
% Prod. Acc. (area-based)	*99 ±0.0002*	*31 ±0.006*			***72.9 ±0.002***	

References

1. Pebesma, E. & Bivand, R. Simple features for R: Standardized support for spatial vector data. *Chapman and Hall/CRC* (2023) doi:10.32614/rj-2018-009.
2. Fagan, M. E. *et al.* The expansion of tree plantations across tropical biomes. *Nat. Sustain.* **5**, 681–688 (2022).
3. Crouzeilles, R. *et al.* Achieving cost-effective landscape-scale forest restoration through targeted natural regeneration. *Conserv. Lett.* **13**, (2020).
4. Hansen, M. C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).
5. GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction software Library. *Open Source Geospatial Foundation* (2023) doi:10.5281/zenodo.5884351.
6. ESRI. ArcGIS Desktop: Release 10.8. *Redlands, CA: Environmental Systems Research Institute* (2011).
7. Baddeley, A. & Turner, R. spatstat: An*R*Package for Analyzing Spatial Point Patterns. *J. Stat. Softw.* **12**, (2005).
8. R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing. *Vienna, Austria* (2021) doi:ISBN3-900051-07-0.
9. Ferro, A., Brunner, D., Bruzzone, L. & Lemoine, G. On the relationship between double bounce and the orientation of buildings in VHR SAR images. *IEEE Geoscience and Remote Sensing Letters* **8**, 612–616 (2011).
10. Henderson, F. M. & Xia, Z.-G. SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report. *IEEE Trans. Geosci. Remote Sens.* **35**, 79–85 (1997).
11. Lee, W. K. Analytical investigation of urban SAR features having a group of corner reflectors. *IGARSS 2001. Scanning the Present and Resolving the Future.*

Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217) **3**, 1282–1284 (2001).