The Journal of Open Source Software

ForestAtRisk: A Python package for modelling and forecasting deforestation in the tropics

Ghislain Vieilledent[1, 2, 3, 4]

1 CIRAD, UMR AMAP, F–34398 Montpellier, France **2** CIRAD, Forêts et Sociétés, F–34398 Montpellier, France. **3** AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France. **4** European Commission, Joint Research Centre, I–21027 Ispra (VA), Italy.

DOI: DOIunavailable

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Reviewers:

- @Pending Reviewers

Submitted: N/A
Published: N/A

Summary

The ForestAtRisk Python package can be used to model the tropical deforestation spatially and forecast the future forest cover in the tropics. It includes functions to (i) sample the forest cover change observations and retrieve the information on potential spatial explanatory variables for each observation, (ii) estimate the parameters of various spatial deforestation models, (iii) predict the spatial probability of deforestation, (iv) forecast the likely forest cover in the future, (v) validate the models and the projected maps of forest cover change, and (vi) plot the results. Spatial information is provided through georeferenced raster files which can be very large (several gigabytes). Functions in the ForestAtRisk package treat large rasters by blocks of data, making computations fast and efficient. This allows analysis on large geographical extent (e.g. country scale), and at high resolution (e.g. 30 m). The ForestAtRisk package offers the possibility to use a logistic regression model with auto-correlated spatial random effects to model the deforestation process. Spatial random effects allow to structure the residual spatial variability (which is often large) within the model. They account for unmeasured or unmeasurable variables that explain a part of the residual spatial variation in the deforestation process that is not explained by the model's explanatory variables. In addition to bringing new features, the ForestAtRisk Python package is open source (license GPLv3), cross-platform and scriptable (through Python), user-friendly (functions provided with full documentation and examples), and easily extendable (with additional statistical models for example). The ForestAtRisk Python package has been recently used to model the deforestation and forecast the future forest cover by 2100 in the whole humid tropics.

Statement of Need

Commonly called the "Jewels of the Earth," tropical forests shelter 30 M species of plants and animals representing half of the Earth's wildlife and at least two-thirds of its plant species (Gibson et al., 2011). Through photosynthesis and carbon sequestration, they play an important role in the global carbon cycle, and in regulating the global climate (Baccini et al., 2017). Despite the many ecosystem services they provide, tropical forests are disappearing at an alarming rate (Keenan et al., 2015; Vancutsem et al., 2020), mostly because of human activities. Currently, around 8 Mha (twice the size of Switzerland) of tropical forest are disappearing each year (Keenan et al., 2015). Spatial modelling of the deforestation allow identifying the main factors determining the spatial risk of deforestation and quantifying their relative effects. Forecasting forest cover change is paramount as it allows to anticipate the consequences of deforestation (in terms of carbon emissions, biodiversity loss, or water supply) under various technological, political and socio-economic scenarios, and inform decision makers accordingly (J. S. Clark et al., 2001). Because both biodiversity and carbon vary greatly in space (Allnutt et al., 2008; Baccini et al., 2017), it is necessary to provide spatial forecasts of forest cover change to properly quantify biodiversity loss and carbon emissions associated to future deforestation.

The ForestAtRisk Python package we present here can be used to model the tropical deforestation spatially, predict the spatial risk of deforestation, and forecast the future forest cover in the tropics (Figure 1). Several other software allow modelling and forecasting forest cover change (J.-F. Mas et al., 2014). Most famous land cover change software include Dinamica-EGO (B. S. Soares-Filho et al., 2002), Land Change Modeller (Eastman & Toledano, 2017), and CLUE (Verburg & Overmars, 2009). Despite the many functionalities they provide, these software are not open source and might not all be cross-platform, scriptable, and completely user-friendly. Moreover, the statistical approaches they propose to model the land cover change do not allow to take into account the residual spatial variability in the deforestation process which is not explained by the model's variables, and which is often very large. The more recent sophisticated algorithms they use (genetic algorithms, artificial neural networks, or machine learning algorithms) might also have the tendency to overfit the data (J.-F. Mas et al., 2014). Finally, the possibility to use these software on large spatial scale (eg. national or continental scale) with high resolution data ($\leq$ 30 m) have not yet been demonstrated (but see Britaldo Silveira Soares-Filho et al. (2006)). The ForestAtRisk Python package aims at filling these gaps and enlarging the range of software available to model and forecast tropical deforestation.

Core functionalities

Processing of large raster data

Data for forest cover change and spatial explanatory variables are commonly available as georeferenced raster data. Raster data consists of rows and columns of cells (or pixels), with each cell storing a single value. The resolution of the raster data set is its pixel width in ground units. Depending on the number of pixels (which is function of the raster's geographical extent and resolution), raster files might occupy a space of several gigabytes on the hard drive. Processing such large rasters in memory might be prohibitively intensive. Functions in the ForestAtRisk package process large rasters by blocks of pixels representing subsets of the raster data. This makes computation efficient, with low memory usage. Reading and writting subsets of raster data is done by using two methods from the GDAL Python bindings (GDAL/OGR contributors, 2020): `gdal.Dataset.ReadAsArray()` and `gdal.Band.WriteArray()`. Numerical computations on arrays are performed with the NumPy (https://numpy.org) Python module whose core is mostly made of optimized and compiled C code which runs fast (Harris et al., 2020). This allows the ForestAtRisk Python package to model and forecast forest cover change on large spatial scale (eg. national or continental scale) using high resolution data ($\leq$ 30 m), even on personal computer with average performance hardware. For example, the ForestAtRisk Python package has been used on a personal computer to model and forecast the forest cover change at 30 m resolution for the Republic Democratic of the Congo (total area of 2,345 millions km^2), processing large raster files of 71,205 $\times$ 70,280 cells without issues.

Statistical model with autocorrelated spatial random effects

The ForestAtRisk Python package, which can be abbreviated `far`, includes a method called `far.model_binomial_iCAR()` to estimate the parameters of a logistic regression model including autocorrelated spatial random effects. The model considers the random variable y_i which takes value 1 if a forest pixel i is deforested in a given period of time and 0 if it is not. The model assumes that y_i follows a Bernoulli distribution of parameter θ_i (Equation 1). θ_i represents the spatial relative probability of deforestation for pixel i and is linked, through a logit function, to a linear combination of the explanatory variables $X_i\beta$, where X_i is the vector of explanatory variables for pixel i, and β is the vector of effects $[\beta_1, \dots, \beta_n]$ associated to the n variables. The model can include or not an intercept α. To account for the residual spatial variation in the deforestation process, the model accounts for additional random effects $\rho_{j(i)}$

for the cells of a spatial grid covering the study-area. The spatial grid resolution has to be chosen in order to have a reasonable balance between a good representation of the spatial variability and a limited number of parameters to estimate. Each observation i is associated to one spatial cell $j(i)$. Random effects ρ_j are assumed to be spatially autocorrelated through an intrinsic conditional autoregressive (iCAR) model (Besag et al., 1991). In an iCAR model, the random effect ρ_j associated to cell j depends on the values of the random effects $\rho_{j'}$ associated to neighbouring cells j'. The variance of the spatial random effects ρ_j is denoted V_ρ. The number of neighbouring cells for cell j, which might vary, is denoted n_j. Spatial random effects ρ_j account for unmeasured or unmeasurable variables (James S. Clark, 2005) that explain a part of the residual spatial variation in the deforestation process that is not explained by the fixed spatial explanatory variables (X_i). Parameter inference is done in a hierarchical Bayesian framework. The `far.model_binomial_iCAR()` method calls an adaptive Metropolis-within-Gibbs algorithm (Rosenthal & others, 2011) written in C for maximum computation speed.

$$
\begin{aligned}
y_i &\sim \mathcal{B}ernoulli(\theta_i) \\
\text{logit}(\theta_i) &= \alpha + X_i\beta + \rho_{j(i)} \\
\rho_{j(i)} &\sim \mathcal{N}ormal(\sum_{j'} \rho_{j'}/n_j, V_\rho/n_j)
\end{aligned}
\tag{1}
$$

Using simple "glm" model is a commonly proposed approach for spatial modelling of deforestation (Eastman & Toledano, 2017; Ludeke et al., 1990; J.-F. Mas et al., 2007; Mertens & Lambin, 1997; Rosa et al., 2014; B. S. Soares-Filho et al., 2002; Britaldo S. Soares-Filho et al., 2001; Verburg et al., 2002-09-21). The random forest model (Breiman, 2001) is a machine learning approach using an ensemble of random classification trees (where both observations and features are chosen at random to build the classification trees) to predict the deforestation probability for a forest pixel. Random forest has been intensively used for species distribution modelling (Thuiller et al., 2009) and is now also commonly used for spatial modelling of deforestation (Grinand et al., 2020; Santos, 2019; Zanella et al., 2017). The "glm" and "rf" models were fitted using functions `LinearRegression` and `RandomForestClassifier` respectively, both available in the `scikit-learn` Python package (Pedregosa et al., 2011).

Figures

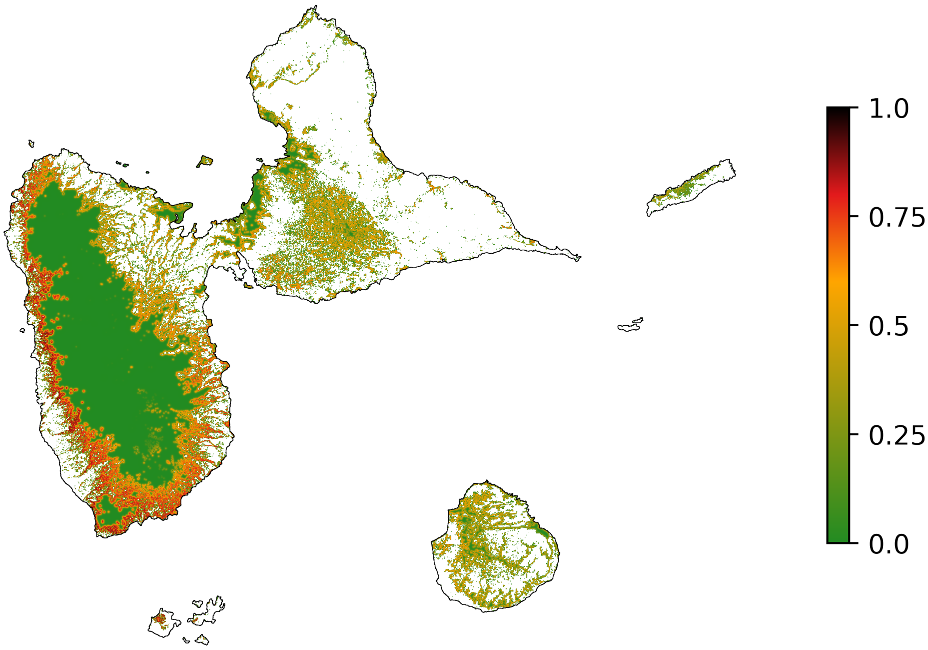

Figure 1: Map of the spatial probability of deforestation in the Guadeloupe archipelago. Colored pixels represent the extent of the natural old-growth tropical moist forest in 2020. A relative spatial probability of deforestation is computed for each forest pixel. The original map has a 30 m resolution. Probability of deforestation is a function of several explanatory variables describing: topography (altitude and slope), accessibility (distances to nearest road, town, and river), forest landscape (distance to forest edge), deforestation history (distance to past deforestation), and land conservation status (presence of a protected area). This map can be reproduced following the Get Started tutorial at https://ecology.ghislainv.fr/forestatrisk.

Acknowledgements

I am grateful to Clovis Grinand, Romuald Vaudry and Matthieu Tiberghien who gave me the opportunity to work on deforestation modeling when we were leading forest conservation projects in Madagascar. I also warmly thank Frédéric Achard and all the members of the IFORCE group for their invaluable support during the first phase of development of the package, while I was seconded to the JRC in Ispra. I would also like to thank Chris Garrad for writting the book untitled "Geoprocessing with Python" (Garrard, 2016) which has been of great help for the development of the `ForestAtRisk` package. This work benefited from funding by FRB-FFEM (BioSceneMada project, AAP-SCEN-2013 I), the European Commission (Roadless Forest project), and CNRT (RELIQUES project).

References

Allnutt, T. F., Ferrier, S., Manion, G., Powell, G. V. N., Ricketts, T. H., Fisher, B. L., Harper, G. J., Irwin, M. E., Kremen, C., Labat, J.-N., Lees, D. C., Pearce, T. A., & Rakotondrainibe, F. (2008). A method for quantifying biodiversity loss and its application to a 50-year record of deforestation across madagascar. *Conservation Letters*, *1*(4), 173–181. http://dx.doi.org/10.1111/j.1755-263X.2008.00027.x

Baccini, A., Walker, W., Carvalho, L., Farina, M., Sulla-Menashe, D., & Houghton, R. A. (2017). Tropical forests are a net carbon source based on aboveground measurements of gain and loss. *Science*, *358*(6360), 230–234. https://doi.org/10.1126/science.aam5962

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

Clark, James S. (2005). Why environmental scientists are becoming bayesians. *Ecology Letters*, *8*(1), 2–14. https://doi.org/10.1111/j.1461-0248.2004.00702.x

Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M., Pascual, M., Pielke, R., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W. H., Wall, D. H., & Wear, D. (2001). Ecological forecasts: An emerging imperative. *Science*, *293*(5530), 657–660.

Eastman, J. R., & Toledano, J. (2017). A short presentation of the land change modeler (LCM). In *Geomatic approaches for modeling land change scenarios* (pp. 499–505). Springer International Publishing. https://doi.org/10.1007/978-3-319-60801-3_36

Garrard, C. (2016). *Geoprocessing with python*. Manning Publications. ISBN: 978-1617292149

GDAL/OGR contributors. (2020). *GDAL/OGR geospatial data abstraction software library*. Open Source Geospatial Foundation. https://gdal.org

Gibson, L., Lee, T. M., Koh, L. P., Brook, B. W., Gardner, T. A., Barlow, J., Peres, C. A., Bradshaw, C. J. A., Laurance, W. F., Lovejoy, T. E., & Sodhi, N. S. (2011). Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, *478*(7369), 378–381. http://dx.doi.org/10.1038/nature10425

Grinand, C., Vieilledent, G., Razafimbelo, T., Rakotoarijaona, J.-R., Nourtier, M., & Bernoux, M. (2020). Landscape-scale spatial modelling of deforestation, land degradation, and regeneration using machine learning tools. *Land Degradation & Development*, *31*(13), 1699–1712. https://doi.org/10.1002/ldr.3526

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Keenan, R. J., Reams, G. A., Achard, F., Freitas, J. V. de, Grainger, A., & Lindquist, E. (2015). Dynamics of global forest area: Results from the {FAO} global forest resources assessment 2015. *Forest Ecology and Management*, *352*, 9–20. https://doi.org/https://doi.org/10.1016/j.foreco.2015.06.014

Ludeke, A. K., Maggio, R. C., & Reid, L. M. (1990). An analysis of anthropogenic deforestation using logistic regression and GIS. *Journal of Environmental Management*, *31*(3), 247–259. http://www.sciencedirect.com/science/article/pii/S0301479705800386

Mas, J.-F., Kolb, M., Paegelow, M., Olmedo, M. T. C., & Houet, T. (2014). Inductive pattern-based land use/cover change models: A comparison of four software packages. *Environmental Modelling & Software*, *51*, 94–111. https://doi.org/10.1016/j.envsoft.2013.09.010

Mas, J.-F., Paegelow, M., Jong, B. de, Guerrero, G., Follador, M., Olgien, M., Díaz, J. R., Castillo, M. A., & García, T. (2007). Modelling tropical deforestation: A comparison of approaches. *Symposium Remote Sensing of Environment*, 3.

Mertens, B., & Lambin, E. F. (1997). Spatial modelling of deforestation in southern cameroon: Spatial disaggregation of diverse deforestation processes. *Applied Geography*, *17*(2), 143–162. http://www.sciencedirect.com/science/article/pii/S0143622897000325

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rosa, I. M., Purves, D., Carreiras, J. M., & Ewers, R. M. (2014). Modelling land cover change in the brazilian amazon: Temporal changes in drivers and calibration issues. *Regional Environmental Change*, 1–15.

Rosenthal, J. S., & others. (2011). Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, *4*(10.1201).

Santos, V. A. B., Fabián AND Graw. (2019). A geographically weighted random forest approach for evaluate forest change drivers in the northern ecuadorian amazon. *PLOS ONE*, *14*(12), 1–37. https://doi.org/10.1371/journal.pone.0226224

Soares-Filho, Britaldo S., Assunção, R. M., & Pantuzzo, A. E. (2001). Modeling the Spatial Transition Probabilities of Landscape Dynamics in an Amazonian Colonization Frontier: Transition probability maps indicate where changes may occur in the landscape, thus enabling better evaluation of the ecological consequences of landscape evolution. *BioScience*, *51*(12), 1059–1067. https://doi.org/10.1641/0006-3568(2001)051%5B1059:MTSTPO%5D2.0.CO;2

Soares-Filho, B. S., Cerqueira, G. C., & Pennachin, C. L. (2002). DINAMICA - a stochastic cellular automata model designed to simulate the landscape dynamics in an amazonian colonization frontier. *Ecological Modelling*, *154*(3), 217–235. https://doi.org/10.1016/S0304-3800(02)00059-5

Soares-Filho, Britaldo Silveira, Nepstad, D. C., Curran, L. M., Cerqueira, G. C., Garcia, R. A., Ramos, C. A., Voll, E., McDonald, A., Lefebvre, P., & Schlesinger, P. (2006). Modelling conservation in the amazon basin. *Nature*, *440*(7083), 520–523. https://doi.org/10.1038/nature04389

Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373.

Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragao, L., & Nasi, R. s. (2020). Long-term (1990-2019) monitoring of tropical moist forests dynamics. *bioRxiv*. https://doi.org/10.1101/2020.09.17.295774

Verburg, P. H., & Overmars, K. P. (2009). Combining top-down and bottom-up dynamics in land use modeling: Exploring the future of abandoned farmlands in europe with the dyna-CLUE model. *Landscape Ecology*, *24*(9), 1167–1181. https://doi.org/10.1007/s10980-009-9355-7

Verburg, P. H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., & Mastura, S. S. A. (2002-09-21). Modeling the spatial dynamics of regional land use: The CLUE-s model. *Environmental Management*, *30*(3), 391–405. http://dx.doi.org/10.1007/s00267-002-2630-x

Zanella, L., Folkard, A. M., Blackburn, G. A., & Carvalho, L. M. T. (2017). How well does random forest analysis model deforestation and forest fragmentation in the brazilian atlantic forest? *Environmental and Ecological Statistics*, *24*(4), 529–549. https://doi.org/10.1007/s10651-017-0389-8