Accounting for spatial regional variability in modelling and forecasting deforestation – Identifying forest refuge areas in Madagascar

Ghislain VIEILLEDENT[1,2,3,4,⋆] and Frédéric ACHARD[1]

[1] **European Commission** – JRC, Bio-economy Unit, I-21027 Ispra (VA), ITALY
[2] **CIRAD** – UPR Forêts et Sociétés, F-34398 Montpellier, FRANCE
[3] **CIRAD** – UPR AMAP, F-34398 Montpellier, FRANCE
[4] **Univ Montpellier** – AMAP, CIRAD, Montpellier, FRANCE

[⋆] **Corresponding author:** \E-mail: ghislain.vieilledent@cirad.fr \Phone: +33 4 67 61 49 09

Abstract

Deforestation models are useful tools in landscape ecology. They can be used to identify the main drivers of deforestation and estimate their relative contribution. When spatially explicit, models can also be used to predict the location of future deforestation. Deforestation forecasts can be used to estimate associated CO_2 emissions responsible of climate change, prioritize areas for conservation and identify refuge areas for biodiversity. Most of spatial deforestation models includes landscape variables such as the distance to forest edge, the distance to nearest road or the presence of protected areas. Such variables commonly explain a small part of the deforestation process and a large spatial variability remains unexplained by the model.

In the present study, we show how using an intrinsic conditional autoregressive (iCAR) model in a hierarchical Bayesian approach can help structure the residual spatial variability in the deforestation process and obtain more realistic predictions of the deforestation. We take Madagascar as a case study considering deforestation on the period 1990-2010 and forecasting deforestation on the period 2010-2050. We demonstrate that accounting for spatial autocorrelation increases the percentage of explained deviance of 21 points for the deforestation model in Madagascar. We also illustrate the use of the newly developed `forestatrisk` Python module to rapidly estimate the numerous parameters of a deforestation model including an iCAR process and efficiently forecast deforestation on large geographical areas at high spatial resolution.

We advocate the use of such models to obtain more accurate land-use change predictions. Such an approach could be used to estimate better the impact of future deforestation in the global carbon cycle and define more efficient strategies for biodiversity conservation in tropical countries.

Keywords: deforestation, forecasts, forest cover change, forest refuge areas, Madagascar, random effects, spatial autocorrelation, spatial modelling, variability

Target journals (selection): *Conservation Biology* (IF: 5.89), *Biological Conservation* (IF: 4.66), *Environmental Research Letters* (IF: 4.54), *Landscape Ecology* (IF: 3.83).

1 Introduction

Context

- Importance of modelling and forecasting deforestation.

Gap

- Difficult task as highly stochastic and mutlifactorial process.
- Deforestation process highly variable in space (Vieilledent *et al.*, 2018b; Yesuf *et al.*, 2019)
- Spatial explanatory variables explain only a small part of the deforestation process (Vieilledent *et al.*, 2013)
- Socio-economic factors are often obtained at large spatial scale (administrative boundaries) and are weak informative factors for small scale deforestation (Vieilledent *et al.*, 2013)
- Bad models can lead to completely wrong projections

Process-based vs. pattern-oriented models (Castella & Verburg, 2007).

Present study Spatial autocorrelated random effect to account for unmeasured hidden factors explaining the residual spatial variability (Clark, 2005).

2 Materials and Methods

2.1 Data

We used historical deforestation maps for Madagascar at 30m resolution for three time-periods: 1990–2000, 2000–2010, and 2010–2017 (Vieilledent *et al.*, 2018b). We tried to model the observed spatial deforestation process on the period 2000–2010 at the national level. Period 1990–2000 was used to compute the distance to past deforestation for each forest pixel in 2000. Period 2010–2017 was used to compare model forecasts with deforestation observations.

To explain the observed spatial deforestation on the period 2000–2010, we considered various spatial 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-tenure variables (protected area system). Characteristics of each variables are summarized in Tab. 1.

Altitude (in m) and slope (in degree) at 90 m resolution were obtained from the SRTM Digital Elevation Database v4.1 (http://srtm.csi.cgiar.org/). Distances (in m) to nearest road, town and river at 150 m resolution were derived from the OpenStreetMap (OSM) project for Madagascar (http://www.geofabrik.de/). To obtain the road network in Madagascar, we considered the "motorway", "trunk", "primary", "secondary" and "tertiary" categories for the "highway" key in OSM. To obtain the network of populated places in Madagascar (that we simply call "towns" in the present study), we considered the "city", "town" and "village" categories for the "place" key in OSM. To obtain the river network in Madagascar, we considered the "river" and "canal" categories for the "waterway" key in OSM. For a more detailed description of each category, see the OSM wiki page (https://wiki.openstreetmap.org/wiki/Tags). Distance to forest edge was computed at 30 m resolution from the forest cover map in 2000. Distance to past deforestation was computed at 3 0m resolution from the 1990–2000 forest cover change map. For the protected area system, we used the 20/12/2010 version of the SAPM "Système des Aires Protégées à Madagascar" (http://rebioma.net/) and considered both Protected Areas (created before 2003) and New Protected Areas in the SAPM terminology. Polygons representing protected areas were rasterized at 30 m resolution. In total, we obtained 8 spatial explanatory variables to model deforestation location.

2.2 Models

We compared two deforestation models. The first model described in Eq. (1) is a simple logistic regression model, a special case of generalized linear model (GLM) for binary data. This model is denoted "glm" in subsequent sections and results. We considered the random variable y_i which takes value 1 if the forest pixel i is deforested on the period 2000–2010 and 0 if it is not. We assumed that y_i follows a Bernoulli distribution of parameter θ_i. In our model, θ_i represents the spatial probability of deforestation for pixel i. We assumed that θ_i 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 model parameters to be estimated.

$$\begin{aligned} y_i &\sim \mathcal{B}ernoulli(\theta_i) \\ \text{logit}(\theta_i) &= X_i\beta \end{aligned} \tag{1}$$

The second model described in Eq. (2) includes additional random effects ρ_j for each spatial cell j of a 10 × 10 km grid covering Madagascar. This grid resolution was choosen in order to have a reasonable balance between a good representation of the spatial variability of the deforestation process and the number of parameters. We assumed that random effects were spatially autocorrelated through an intrinsic conditional autoregressive (iCAR) model (Besag *et al.*, 1991; Banerjee *et al.*, 2014). This model is denoted "icar" in subsequent sections and results. In a 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'. In our case, the neighbouring cells are cells connected to the target cell j through a common border or corner (cells defined by the "king move" in chess). The number of neighbouring cells for cell j, which might vary, is denoted n_j.

$$\begin{aligned} y_i &\sim \mathcal{B}ernoulli(\theta_i) \\ \text{logit}(\theta_i) &= X_i\beta + \rho_j \\ \rho_j &\sim \mathcal{N}ormal(\sum_{j'} \rho_{j'}/n_j, V_\rho/n_j) \end{aligned} \tag{2}$$

The first model can be viewed as a "process-based" model for which variables are selected on an *a priori* knowledge of the deforestation process. For example, we assumed that the risk of deforestation decreases with the distance to road and forest edge, and is lower in protected areas. The second model can be viewed as a model combining a "process-based" part and a "pattern-oriented" part. Additional spatial random effects ρ_j account for unmeasured or unmeasurable variables (Clark, 2005) that explain a part of the residual spatial variation in the deforestation process (the residual spatial "pattern") that is not explained by the fixed environmental variables (X_i). While the first model has only 9 parameters to be estimated (one intercept parameter plus 8 slope parameters for the explanatory variables), the second model has 6,266 parameters to be estimated, including the 6,257 spatial random effects for the 10 × 10 km cells covering whole Madagascar (for which lands cover 587 000 km^2).

We used a random sample of 20,000 forest pixels in 2000 to fit the two models. The sample was stratified between 10,000 deforested pixels in 2000–2010 and 10,000 non-deforested pixels. A balanced sample between deforested and non-deforested pixels is preferable in our case (Dezécache *et al.*, 2017). First, deforestation events are rare (~1 %/yr) and a non-stratified sample would lead to very few observations of deforestation events, rendering difficult a good estimation of the slope parameters for the explanatory variables. Second, only the value of the linear model intercept is affected by this balanced sampling, which is not the case for the slope or random parameters. In our case, a biased estimate of the intercept is not an issue, as we are not interested in estimating the intensity of deforestation but the relative probability of deforestation between pixels. Function `sample()` from the `forestatrisk` Python package was used for fast stratified sampling and for extracting variable values at each point.

Parameter inference was done in a hierarchical Bayesian framework. Non-informative priors were used for all parameters: $\beta \sim \mathcal{N}ormal(\text{mean} = 0, \text{var} = 10^6)$ and $V_\rho \sim 1/\mathcal{G}amma(\text{shape} = 0.05, \text{rate} = 0.0005)$. We run a Markov Chain Monte Carlo (MCMC) of 7000 iterations. We discarded the first 2000 iterations (burn-in phase) and we thinned the chain each 5 iterations (to reduce autocorrelation between samples). We obtained 1000 estimates for each parameter. MCMC convergence was visually checked looking at MCMC traces and parameter posterior distributions. Function `model_binomial_iCAR()` from the `forestatrisk` Python package was used for parameter inference. This function calls an adaptive Metropolis-within-Gibbs algorithm (Rosenthal *et al.*, 2011) written in C for maximum computation speed.

2.3 Model comparison using percentage of deviance explained and cross-validation

We computed the deviance $\mathcal{D}$ of the two models with the formula $\mathcal{D} = -2 \log \mathcal{L}$, $\mathcal{L}$ being the likelihood of the model, i.e. the probability of observing the data given the model and estimated parameters. We compared the deviance of the two models with the deviances of both the "null" model and the "full" model. The "null" model assumes a constant probability of deforestation for all theobservations and has only one parameter, the intercept of the linear relationship. At the other extreme, the "full" model has as many parameters as there are observations. We then computed the percentage of deviance explained by each model, considering that the "null" model explains 0% of the deviance and the "full" model explains 100% of the deviance.

We also performed a cross-validation to compare models using an independent validation data-set of 20,000 forest pixels in 2000. Again, the sample was stratified between 10,000 deforested pixels in 2000–2010 and 10,000 non-deforested pixels. We used the fitted models to predict the deforestation probability of all the pixels of the validation data-set. To transform the deforestation probabilities into binary values, we identified the probability threshold respecting the percentage of deforested pixels (eg. the mode of the probabilities for a deforestation rate of 50%). Using model predictions and observations, we computed several accuracy indices: the Area Under the ROC Curve (AUC), the Figure of Merit (FOM), the Overall Accuracy (OA), the Expected Accuracy (EA), the Kappa of Cohen (K), the Specificity (Spe), the Sensitivity(Sen), and the True Skill Statistics (TSS). A detailed description of these indices can be found in Pontius *et al.* (2008) (for the FOM) and Liu *et al.* (2011) (for all the other indices). Formulas used to compute these indices are presented in Appendix 1.

Because the value of these indices depends on the deforestation rate (Pontius *et al.*, 2008), we computed the accuracy indices for various percentage of deforested pixels: 1, 5, 10, 25 and 50%. To do so, we selected subsamples of the deforested pixels in our validation data-set at random.

2.4 Comparing models' forecasting skill on the period 2000–2017

2.4.1 Computing the spatial probability of deforestation in 2010

We used the fitted models to predict the deforestation probability of all the forest pixels for the year 2010. Distances to forest edge in 2010 and to past deforestation in 2000–2010 were recomputed from forest cover maps in 2000 and 2010. All other explanatory variables were supposed constant. Spatial random effects were also supposed constant through time.

For the "icar" model, before computing the predictions of the deforestation probability, the spatial random effects at 10 km were interpolated at 1 km using a bicubic interpolation method. This was done in order to obtain spatial random effects at a resolution closer to the original forest raster resolution of 30 m, and to smooth the deforestation probability spatially.

Deforestation probabilities (float values in the interval $[0, 1]$) were transformed as integer values on the interval $[\![1, 65535]\!]$. This allows to save the large raster of probability as UInt16 type and save space on disk. We then obtained a map of the relative probability of deforestation for the year 2010 at 30 m resolution.

In 2010, Madagascar was covered by 9.3 Mha of natural forest corresponding to more than 104 M pixels at 30 m resolution. Predictions were computed using functions `predict_raster*()` from the `forestatrisk` Python package which make computation fast and efficient (with low memory usage) by treating raster data by blocks.

2.4.2 Projecting the forest cover change on the period 2010–2017

We computed the observed deforestation D (in ha) on the period 2010–2017 from the forest cover maps at these two dates. To forecast the forest cover change in 2010–2017 with our models, we used the previously derived maps of relative probability of deforestation in 2010. The resolution of these maps is $r = 30$ m, equivalent to $r_{\text{ha}} = 0.09$ ha. We computed a probability threshold θ_T in the interval $[\![1, 65535]\!]$ identifying the n forest pixels in 2010 with the highest probability of deforestation so that $nr_{\text{ha}} = D + \epsilon$. Because deforestation probabilities have finite values in $[\![1, 65535]\!]$, some forest pixels might have the same deforestation probability and it might not be possible to identify θ_T such that $\epsilon = 0$. We thus selected the threshold θ_T minimizing ϵ, which was negligible ($<$ 20,000 ha) compared to D (874,211 ha) for both models. We considered those n forest pixels in 2010 as deforested on the period 2010–2017 and derived the forest cover change map on that period.

2.4.3 Validating the forest cover change on the period 2010–2017

We compared the forest cover change maps obtained from the modelling and forecasting procedure with the observed forest cover change map on the period 2010–2017. Forest cover change maps have a high resolution of 30 m. At high resolution, a pixel-to-pixel comparison provides limited information on the spatial accuracy of the predictions because it ignores neighborhood relationships. To overcome this problem, a multiple resolution accuracy assessment can be performed (Pontius *et al.*, 2004). Following the methodology proposed by Pontius *et al.* (2004), we computed two accuracy indices, the Figure of Merit (FOM) and the Overall Accuracy (OA), at multiple resolutions from 30 m to 15 km. In our case, the "quantity disagreement" census Pontius *et al.* (2004) between observations and predictions is negligeable. The accuracy indices we computed allow us to estimate a "location disagreement" census Pontius *et al.* (2004) between observations and predictions. Accuracy indices at multiple resolutions were compute with functions `resample_sum()`, `confmat()`, and `accuracy()` from the `forestatrisk` Python package.

2.5 Forecasting deforestation on the period 2017–2050

Using the same method as previously, we computed the deforestation probability of all the forest pixels for the year 2017. Distances to forest edge in 2017 and to past deforestation in 2000–2017 were recomputed from forest cover maps in 2000 and 2017. All other explanatory variables were supposed constant. Spatial random effects were also supposed constant through time. We then considered two scenarios for the deforestation intensity to forecast forest cover change on the period 2017–2050.

The first scenario is a "business-as-usual" scenario assuming that annual deforestation on the period 2017–2050 will be the same as the mean annual deforestion on the period 2000–2017 (85,000 ha/yr, Vieilledent *et al.* (2018a)). For this first scenario, a total of 2.8 Mha should be deforested on the period 2017–2050 and 5.6 Mha of natural forest should remain in 2050.

The second scenario takes into account the demographic growth in Madagascar, which is close to 3%/yr (Raftery *et al.*, 2012; Vieilledent *et al.*, 2013), and the link between population and deforestation. We considered the following model adapted from Barnes (1990): $\log(D_{t,t+1}) = \beta_0 + 0.607\log(F_t) + 0.493\log(P_t) + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}ormal(0, \sigma^2)$, and where $D_{t,t+1}$ is the mean annual deforestation (in ha/yr) between times t and $t+1$, F_t is the forest cover (in thousand hectares) at time t, and P_t is the population size (in thousand people) at time t. This model was fitted to data from 22 African countries in 1980, but not including Madagascar. Nonetheless, we assumed that the relationship between people and deforestation in continental Africa holds true for Madagascar as well, as deforestation is mainly due to small scale family farming in both regions (Curtis *et al.*, 2018). Using historical deforestation data from Vieilledent *et al.* (2018a) on the interval 1990–2017 (5 different time periods) and population data from the United Nations World Population Prospect (United Nations, 2017), we estimated parameters β_0 and σ^2 for Madagascar. We obtained $\beta_0 = -5.942$ and $\sigma^2 = 0.219$. Following Barnes (1990), we used this model to iteratively predict deforestation and forest cover from 2017 to 2050 using a time step of 3 years for 2017-2020 and 5 years for 2020-2050. Following this demographic scenario, we found that 3.6 Mha should be deforested on the period 2017–2050 and that only 4.8 Mha of natural forest should remain in 2050 (Appendix 2).

3 Results

4 Discussion

5 Ackowledgements

6 Tables

Table 1: **Set of explicative variables used to model the spatial probability of deforestation.** A total of height variables were tested. They described topography, forest accessibility, forest landscape, land tenure and deforestation history.

Product	Source	Variable derived	Unit	Resolution (m)
Forest maps (1990-2000-2010)	Vieilledent et al. 2018	distance to forest edge	m	30
		distance to past deforestation	m	30
Digital Elevation Model	SRTM v4.1 CSI-CGIAR	altitude	m	90
		slope	degree	90
Highways	OSM-Geofabrik	distance to roads	m	150
Places		distance to towns	m	150
Waterways		distance to river	m	150
Protected areas	Rebioma	presence of protected area	–	30

7 Figures

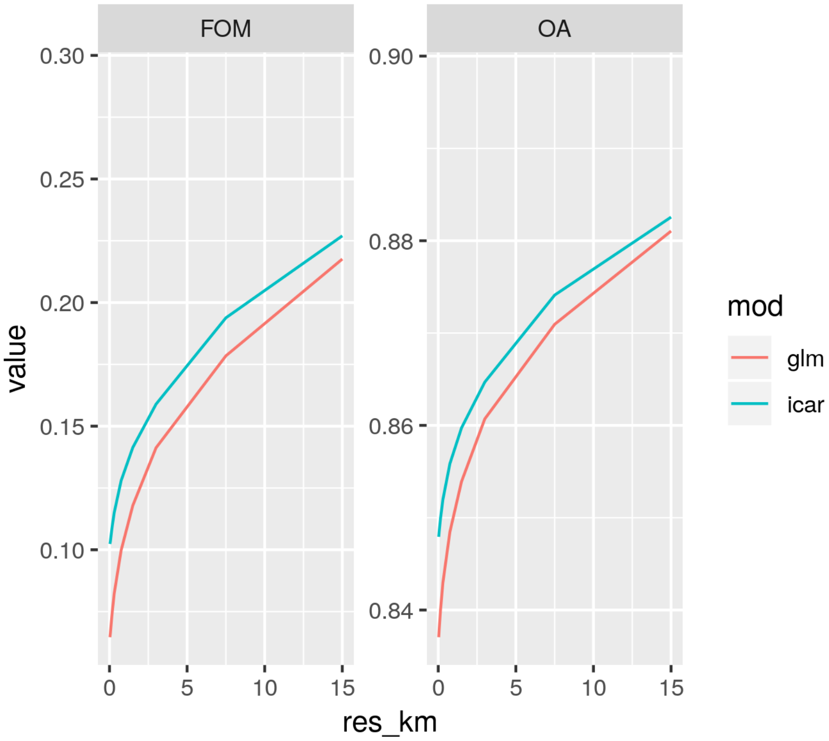

Figure 1: **Models' accuracy indices at multiple resolutions** We compared the projected deforestation maps of the two models (glm and icar) to the observed deforestation map on the period 2010-2017 (Vieilledent *et al.*, 2018a,b). We computed the Figure Of Merit (FOM) and the Overall Accuracy (OA) of the projections at multiple resolutions (from 30 m to 15 km).

8 Appendices

8.1 Appendix 1: Mathematical formulas for accuracy indices

Table 2: **Confusion matrix used to compute accuracy indices.** A confusion matrix can be computed to compare model predictions with observations.

		Observations		Total
		0 (non-deforested)	1 (deforested)	
Predictions	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
Total		n_{+0}	n_{+1}	n

Table 3: **Formulas used to compute accuracy indices.**. Several accuracy indices can be computed from the confusion matrix to estimate and compare models' predictive skill. We followed the definitions of Pontius *et al.* (2008) for the FOM and Liu *et al.* (2011) for the other indices. Note that the AUC relies on the predicted probabilities for observations 0 (non-deforested) and 1 (deforested), not on the confusion matrix per se.

Index	Formula
Overall Accuracy	$\text{OA} = (n_{11} + n_{00})/n$
Expected Accuracy	$\text{EA} = (n_{1+}n_{+1} + n_{0+}n_{+0})/n^2$
Figure Of Merit	$\text{FOM} = n_{11}/(n_{11} + n_{10} + n_{01})$
Sensitivity	$\text{Sen} = n_{11}/(n_{11} + n_{01})$
Specificity	$\text{Spe} = n_{00}/(n_{00} + n_{10})$
True Skill Statistics	$\text{TSS} = \text{Sen} + \text{Spe} - 1$
Cohen's Kappa	$\text{K} = (\text{OA} - \text{EA})/(1 - \text{EA})$
Area Under ROC Curve	$\text{AUC} = 1/(n_{+1}n_{+0}) \sum_{i=1}^{n_{+0}} \sum_{j=1}^{n_{+1}} \phi(\delta_i, \theta_j)$
	where $\phi(\delta_i, \theta_j)$ equals 1 if $\theta_j > \delta_i$, 1/2 if $\theta_j = \delta_i$, and 0 otherwise
	δ_i and θ_j are the predicted probabilities for $Y_i = 0$ and $Y_j = 1$

References

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2014) *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman and Hall/CRC.
URL https://doi.org/10.1201%2Fb17115

Barnes, R.F.W. (1990) Deforestation trends in tropical Africa. *African Journal of Ecology*, **28**, 161–173.
URL https://doi.org/10.1111/j.1365-2028.1990.tb01150.x

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

Castella, J.C. & Verburg, P.H. (2007) Combination of process-oriented and pattern-oriented models of land-use change in a mountain area of Vietnam. *Ecological Modelling*, **202**, 410–420.
URL https://doi.org/10.1016/j.ecolmodel.2006.11.011

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

Curtis, P.G., Slay, C.M., Harris, N.L., Tyukavina, A. & Hansen, M.C. (2018) Classifying drivers of global forest loss. *Science*, **361**, 1108–1111.
URL https://doi.org/10.1126/science.aau3445

Dezécache, C., Salles, J.M., Vieilledent, G. & Hérault, B. (2017) Moving forward socio-economically focused models of deforestation. *Global Change Biology*, **23**, 3484–3500.
URL https://doi.org/10.1111/gcb.13611

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, **34**, 232–243. ISSN 1600-0587.
URL http://dx.doi.org/10.1111/j.1600-0587.2010.06354.x

Pontius, R., Boersma, W., Castella, J.C., Clarke, K., de Nijs, T., Dietzel, C., Duan, Z., Fotsing, E., Goldstein, N., Kok, K., Koomen, E., Lippitt, C., McConnell, W., Mohd Sood, A., Pijanowski, B., Pithadia, S., Sweeney, S., Trung, T., Veldkamp, A. & Verburg, P. (2008) Comparing the input, output, and validation maps for several models of land change. *The Annals of Regional Science*, **42**, 11–37. ISSN 0570-1864.
URL http://dx.doi.org/10.1007/s00168-007-0138-2

Pontius, R.G., Huffaker, D. & Denman, K. (2004) Useful techniques of validation for spatially explicit land-change models. *Ecological Modelling*, **179**, 445–461.
URL https://doi.org/10.1016/j.ecolmodel.2004.05.010

Raftery, A.E., Li, N., Sevcikova, H., Gerland, P. & Heilig, G.K. (2012) Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*,

109, 13915–13921.
URL https://doi.org/10.1073/pnas.1211452109

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

United Nations (2017) *World Population Prospects: The 2017 Revision. Revision, custom data acquired via website. United Nations, Department of Economic and Social Affairs, Population Division.*
URL https://population.un.org/wpp/

Vieilledent, G., Grinand, C., Rakotomalala, F.A., Ranaivosoa, R., Rakotoarijaona, J.R., Allnutt, T.F. & Achard, F. (2018a) Output data from: Combining global tree cover loss data with historical national forest-cover maps to look at six decades of deforestation and forest fragmentation in Madagascar.
URL http://dx.doi.org/10.18167/DVN1/AUBRRC

Vieilledent, G., Grinand, C., Rakotomalala, F.A., Ranaivosoa, R., Rakotoarijaona, J.R., Allnutt, T.F. & Achard, F. (2018b) Combining global tree cover loss data with historical national forest cover maps to look at six decades of deforestation and forest fragmentation in Madagascar. *Biological Conservation*, **222**, 189–197.
URL https://doi.org/10.1016/j.biocon.2018.04.008

Vieilledent, G., Grinand, C. & Vaudry, R. (2013) Forecasting deforestation and carbon emissions in tropical developing countries facing demographic expansion: a case study in Madagascar. *Ecology and Evolution*, **3**, 1702–1716.
URL https://doi.org/10.1002/ece3.550

Yesuf, G., Brown, K.A. & Walford, N. (2019) Assessing regional-scale variability in deforestation and forest degradation rates in a tropical biodiversity hotspot. *Remote Sensing in Ecology and Conservation*.
URL https://doi.org/10.1002/rse2.110