Mixture of inhomogeneous matrix models for species-rich ecosystems

Frederic Mortier, Dakis-Yaoba Ouedraogo, Florian Claeys, Mahlet G. Tadesse, Guillaume Cornu, Fidele Baya, Fabrice Benedet, Vincent Freycon, Sylvie Gourlet-Fleury, Nicolas Picard

Research Article Environmetrics

Received: 22 May 2014, Revised: 12 October 2014, Accepted: 14 October 2014, Published online in Wiley Online Library: 11 November 2014

(wileyonlinelibrary.com) DOI: 10.1002/env.2320

Mixture of inhomogeneous matrix models for species-rich ecosystems

Frédéric Mortier[a]*, Dakis-Yaoba Ouédraogo[a], Florian Claeys[a,b,c], Mahlet G. Tadesse[d], Guillaume Cornu[a], Fidèle Baya[e], Fabrice Benedet[a], Vincent Freycon[a], Sylvie Gourlet-Fleury[a] and Nicolas Picard[a]

Understanding how environmental factors could impact population dynamics is of primary importance for species conservation. Matrix population models are widely used to predict population dynamics. However, in species-rich ecosystems with many rare species, the small population sizes hinder a good fit of species-specific models. In addition, classical matrix models do not take into account environmental variability. We propose a mixture of regression models with variable selection allowing the simultaneous clustering of species into groups according to vital rate information (recruitment, growth and mortality) and the identification of group-specific explicative environmental variables. We develop an inference method coupling the R packages `flexmix` and `glmnet`. We first highlight the effectiveness of the method on simulated datasets. Next, we apply it to data from a tropical rain forest in the Central African Republic. We demonstrate the accuracy of the inhomogeneous mixture matrix model in successfully reproducing stand dynamics and classifying tree species into well-differentiated groups with clear ecological interpretations. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: mixture models; lasso selection; species-rich ecosystems; usher models

1. INTRODUCTION

Understanding how environmental factors could impact population dynamics is of primary importance for animal and plant species conservation. Mathematical and statistical models are required to understand and predict these dynamics (Fieberg and Ellner, 2001; Demyanov *et al.*, 2006). Habitat models (Pearson *et al.*, 2002; Hargrove and Hoffman, 2004; García-López and Allué, 2011) use the spatial distribution of climate variables to predict the spatial range of species. These models are static in space and time and are conceptually unable to deal with situations where species are not in equilibrium with their environments (Stankowski and Parker, 2010). Ecophysiology-based dynamic global vegetation models (e.g., Scheiter and Higgins, 2009) precisely describe the biological processes that underlie growth, mortality and recruitment but require a huge amount of information. In species-rich ecosystems, limited information is available for each species. It is thus intractable to characterize different species with these models; instead, a plant functional type assumed to be representative of several species is modelled. As a consequence, these methods are more useful to predict biome changes at a continental scale than forest changes at a regional scale. Gap models (Solomon, 1986; Pastor and Post, 1988; Prentice *et al.*, 1993; Shao, 1996; Talkkari *et al.*, 1999), while using a simplified description of biological processes when compared with process-based models, still suffer from the same information limitation and are hardly used for species-rich forest ecosystems (Shugart and West, 1980).

Matrix population models, on the other hand, have been widely used to investigate the dynamics of age-, stage- or size-structured populations (Caswell, 2001; Stott *et al.*, 2010). They provide a simple way of integrating vital rate information such as birth, recruitment, growth or ageing and mortality (Crone *et al.*, 2011; Liang, 2010). In forest ecology and forest management, matrix models have been used to study natural successions, biodiversity dynamics and the impact of natural disturbances. They have also been used to evaluate economic outcomes and ecological impacts and to optimize management strategies (Buongiorno and Gilless, 2003).

Another challenge with species-rich ecosystems, such as tropical rain forests, tropical marine fish or coral reefs, is their high diversity, which implies that the sample size for most species is limited. The small sample size hinders development of species-specific models.

* *Correspondence to: F. Mortier, UPR Bsef, CIRAD, TA C-105/D, Campus International de Baillarguet, Montpellier, 34398, France. E-mail: fmortier@cirad.fr*

a *UPR Bsef, CIRAD, Montpellier, France*

b *AgroParisTech, Paris, France*

c *UMR Lef, AgroParisTech-INRA, Nancy, France*

d *Department of Mathematics and Statistics, Georgetown University, Washington, DC, U.S.A.*

e *Ministère des Eaux, Forêts, Chasse et Pêche, Bangui, Central African Republic*

To address this problem, modellers usually cluster species into groups using a variety of methods (Swaine and Whitmore, 1988; Steneck and Dethier, 1994; Favrichon, 1994; Bellwood and Wainwright, 2001; Gitay and Noble, 1997). Mixture models that cluster based on similar species responses rather than similar species traits have been proposed in the framework of generalized linear models (GLM) (Dunstan *et al.*, 2011; Dunstan *et al.*, 2013; Hui *et al.*, 2013; Ouédraogo *et al.*, 2013) and more recently in the context of homogeneous matrix population models (Mortier *et al.*, 2013).

In this paper, we propose a new class of mixture of inhomogeneous matrix population models that allows the simultaneous clustering of species based on vital rate processes (recruitment, growth and mortality) and selection of group-specific explicative environmental variables. The novelty of this method is that it provides the flexibility of selecting cluster-specific covariates in the context of multivariate GLM. It generalizes previous work for variable selection in multivariate Gaussian regression models (Brown *et al.*, 1998; Monni and Tadesse, 2009; Ouédraogo *et al.*, 2013) or in univariate GLM (Gupta and Ibrahim, 2007; Khalili and Chen, 2007; Städler *et al.*, 2010).

Section 2.2 is dedicated to the formulation of adaptive lasso regression mixture models and the associated expectation–maximization (EM) algorithm. Section 3 describes the simulation studies and a real dataset from the M'Baïki tropical rain forest in the Central African Republic, and Section 4 presents the corresponding results. The simulations demonstrate the effectiveness of the proposed method under various scenarios, while the real dataset highlights the performance of the mixture of inhomogeneous matrix models to predict stand characteristics of species-rich ecosystems in contrasted environmental conditions.

2. MODELS

2.1. Usher model

We first focus on a specific population labelled s and discuss the general setting that considers the whole stand in Section 2.2. The Usher matrix model applies to size-structured populations (Usher, 1966, 1969). It is based on the description of the change in the population size by a vector $N_s(t)$ containing the number of individuals in I ordered size classes at a discrete time t : $\mathbf{N}_s(t) = (N_{si}(t))_{i=1,\dots I}$, where $N_{si}(t)$ is the number of trees in the diameter class i at time t. The transitions between t and $t+1$ follow the Usher assumption that a tree can either stay in the same class, move up to the next class or die (moving backwards or moving up by more than one class are not allowed). The temporal change between times t and $t+1$ is defined by the recurrence relation

$$\mathbf{N}_s(t+1) = \mathbf{A}_s(t)\,\mathbf{N}_s(t) + \mathbf{R}_s(t) \tag{1}$$

where $\mathbf{A}_s(t)$ is the Usher $I \times I$ transition matrix for population s,

$$\mathbf{A}_s(t) = \begin{pmatrix} p_{s1}(t) & 0 & \dots & 0 \\ q_{s2}(t) & p_{s2}(t) & & 0 \\ & \ddots & \ddots & \\ 0 & & q_{sI}(t) & p_{sI}(t) \end{pmatrix} \tag{2}$$

and $\mathbf{R}_s(t)$ is the I-vector of recruitment for population s:

$$\mathbf{R}_s(t) = \begin{pmatrix} r_s(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3}$$

The transition parameters consist of: the stasis rate, $p_{si}(t)$, which corresponds to the probability of a tree in diameter class i at time t to stay alive and remain in the same diameter class at time $t+1$; the upgrowth rate, $q_{s,i+1}(t)$, which corresponds to the probability of a tree in diameter class i at time t to stay alive and to move up to diameter class $i+1$ at time $t+1$; and the recruitment flow, $r_s(t)$, which corresponds to the number of newly recruited trees in the first diameter class at time t. The transition parameters can be reparameterized as

$$\begin{aligned} q_{s,i+1}(t) &= q^{\bullet}_{s,i+1}(t) \times (1 - m_{si}(t)) \\ p_{si}(t) &= 1 - m_{si}(t) - q_{s,i+1}(t) \end{aligned} \tag{4}$$

where $q_{s,I+1}(t) = 0$ and $q^{\bullet}_{s,i+1}(t)$ is the conditional probability for a tree in diameter class i at time t to move up to diameter class $i+1$ given that it stays alive, and $m_{si}(t)$ is the probability for a tree in diameter class i to die between times t and $t+1$. Recruitment is assumed additive rather than proportional to the number of trees in each diameter class (Buongiorno and Michie, 1980). This means that the recruitment flow does not follow from the population alone but also involves an external inflow from the surrounding community. This additive recruitment is suited to the M'Baïki experimental case, where the observed plots are a sample of the whole forest (Caswell, 2001). A particular aspect of this matrix model is that the transition matrix $\mathbf{A}_s(t)$ and the recruitment vector $R_s(t)$ have explicit time dependence introduced through the linear associations of the demographic processes with time-varying environmental covariates. This contrasts with standard matrix models that are stationary.

2.1.1. *Predicting growth*

The upgrowth transition rate $q^{\bullet}_{s,i+1}(t)$ is computed from $a_{si}(t)$, defined as the 'typical' diameter at breast height (dbh) growth rate of a tree in class i at time t. Let u_i and u_{i+1} be the boundaries of class i and let τ be the time step of the matrix model. All trees with dbh ranging from $u_{i+1} - a_{si}(t)\,\tau$ to u_{i+1} will grow up to the next class, whereas trees with a diameter ranging from u_i to $u_{i+1} - a_{si}(t)\,\tau$ will remain in the same class. The proportion of trees that grow up to the next diameter class can thus be computed as

$$q^{\bullet}_{s,i+1} = \frac{a_{si}(t)\,\tau}{d_i} \tag{5}$$

where $d_i = u_{i+1} - u_i$ is the width of diameter class i. The typical dbh growth rate $a_{si}(t)$ can be estimated using growth data from class i only or using a regression model that relates growth and size over the entire size range (Rogers-Bennett and Rogers, 2006). The advantages and limitations of each estimator have been discussed elsewhere (Picard *et al.*, 2008). Here, we use the regression approach and predict the typical dbh growth rate as

$$a_{si}(t) = X^{G}_{si}(t)\beta_s \tag{6}$$

where the β_s's are population-specific coefficients to be estimated from the data and $X^{G}_{si}(t)$ are a set of known time-varying environmental covariates associated to the growth process.

2.1.2. *Predicting mortality*

The probability $m_{si}(t)$ that a tree in diameter class i dies between times $t-1$ and t is computed as

$$m_{si}(t) = \text{logit}^{-1}\left[X^{M}_{si}(t)\gamma_s\right] \times (\tau/\Upsilon) \tag{7}$$

where $\text{logit}^{-1}(x) = (1 + \exp(-x))^{-1}$ is the inverse logit function, the γ_s's are population-specific coefficients to be estimated from the data, $X^{M}_{si}(t)$ are a set of known time-varying environmental covariates associated to the mortality process, and Υ is the time step for death observations. The ratio τ/Υ must ensure that $m_{si}(t) < 1$, which in practice is satisfied even when τ is 10-fold Υ because of the very small value of the inverse logit term.

2.1.3. *Predicting recruitment*

The number of recruits $r_s(t)$ at time t in the first diameter class is computed as

$$r_s(t) = \exp\left[X^{R}_{s}(t)\alpha_s\right] \times (\tau/\Upsilon) \tag{8}$$

where the α_s's are population-specific coefficients to be estimated from the data, $X^{R}_{s}(t)$ are a set of known time-varying environmental covariates associated to the recruitment process, and Υ is the time step for recruitment observations.

2.2. Mixture of regression models and variable selection

So far, we have considered a single population. We now consider the whole stand, with as many populations as there are species. Because there are a lot of species with very few individuals, the parameters α_s, β_s and γ_s cannot be estimated for all the species of the stand. Thus, we aim to group species based on their common behaviour (growth, mortality or recruitment) as well as their similar association patterns with environmental factors. Species in the same group will share the same estimated parameters.

Species clustering is defined separately for growth, recruitment and mortality processes, and the clustered responses are related to the predictors defined in Equations (6)–(8). We develop a unified method to simultaneously (i) classify species according to their response to the predictors, (ii) select the significant predictors and (iii) estimate the parameters α_s, β_s, and γ_s of Equations (6)-(8) for each species group. We use a finite mixture of GLM to classify species into groups and estimate the model parameters, and we incorporate an adaptive lasso penalty to select the predictors for each group (Städler *et al.*, 2010).

Let S be the number of species, T the number of measurement times, n_{st} the number of trees from species s measured at time t (where $s = 1, \ldots, S$ and $t = 1, \ldots, T$), and $n = \sum_{s=1}^{S}\sum_{t=1}^{T} n_{st}$ the total number of observations in the dataset. The time considered here is a chronological one used to model annual differences and does not correspond to tree age. Let $\mathbf{Y}$ be the random vector of observations associated with either growth increments or death events. We assume that the growth rate for a tree from species s in dbh class i (conditionally on the tree staying alive) follows a Gaussian distribution with expectation equal to $a_{si}(t)$ and variance σ^2_s and that the death event is distributed as a Bernoulli random variable with probability $m_{si}(t)$. Using mixture models to group species with similar characteristics, the log-likelihoods of the growth and mortality processes for the n observations are computed as

$$\ell_n(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} \log\left[\sum_{k=1}^{K} \pi_k\, f(Y_{stj}|\mathbf{X}, \boldsymbol{\psi}_k)\right] \tag{9}$$

where K is the number of species groups, π_k is the mixing proportion of group k, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K)$ with $\boldsymbol{\psi}_k$ the model parameters for group k, and $\mathbf{X}$ is the design matrix of explanatory variables. For the growth model, f is the Gaussian density function, $Y_{stj} = \Delta D_{stj}/\Upsilon$, where ΔD_{stj} is the diameter increment between times t and $t + \Upsilon$ for the j-th tree from species s and Υ is the time step between successive observations, and $\boldsymbol{\psi}_k = (\beta_k, \sigma_k)$. For the mortality model, f is the Bernoulli probability mass function, $Y_{stj} = M_{stj}$, where M_{stj} is a binary indicator of whether the j-th tree from species s died between times $t-1$ and t, and $\boldsymbol{\psi}_k = \gamma_k$.

The log-likelihood for the recruitment process is given by

$$\ell_n(\boldsymbol{\psi}|\mathbf{Y}) = \sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\sum_{k=1}^{K} \pi_k \, f(Y_{st}|\mathbf{X}, \alpha_k)\right] \tag{10}$$

where f is the probability mass function associated to the Poisson distribution with expected value $\exp(\mathbf{X}\alpha_k)$, $Y_{st} = R_{st}$ is the observed number of recruited trees for species s at time t, and $\boldsymbol{\psi}_k = \alpha_k$. It should be noted that the Poisson distribution is restrictive because of its assumption of equal expectation and variance, which is often not satisfied for ecological count data (Flores *et al.*, 2009). The negative binomial distribution can be a solution but may not be sufficient to accommodate the large number of zeros often recorded for recruitment processes. An alternative would be to use zero-inflated distributions (Poisson or negative binomial).

The relevant covariates associated to the different processes may vary from one group to another. We propose using the adaptive lasso approach to select the group-specific covariates (Zou, 2006; Städler *et al.*, 2010). The estimator $\hat{\boldsymbol{\psi}}$ for the model parameters $\boldsymbol{\psi}$ then corresponds to the maximum of a penalized log-likelihood:

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \{\ell_n(\boldsymbol{\psi}|\mathbf{Y}) - \mathcal{P}_n(\boldsymbol{\psi})\}$$

where $\mathcal{P}_n$ is the adaptive lasso penalty:

$$\mathcal{P}_n(\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \eta_{nk} \sum_{l=1}^{L} \frac{|\psi_{kl}|}{\left|\hat{\psi}_{kl}\right|} \tag{11}$$

with ψ_{kl} the lth element of $\boldsymbol{\psi}_k$, $\left|\hat{\psi}_{kl}\right|$ the maximum likelihood estimator of ψ_{kl}, and η_{nk} a parameter selected using cross-validation.

2.2.1. *Expectation–maximization algorithm*

Because of the sum within the log in Equations (9) and (10), the penalized log-likelihood cannot be maximized analytically but can be numerically maximized using the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008). The EM algorithm is an iterative procedure that alternates between two steps, the E (or expectation) step and the M (or maximization) step. It starts with a random assignment of the species to the K groups. This gives the initial values $w_{stjk}^{(0)}$ of the posterior probability that the j-th tree from species s at time t belongs to species group k : $w_{stjk}^{(0)} = 1$ if species s is initially assigned to group k, and 0 otherwise.

In the E-step, the posterior probability that the j-th tree from species s at time t belongs to species group k is computed as

$$w_{stjk}^{(m)} = \frac{\pi_k^{(m)} \prod_{t'=1}^{T} \prod_{j'=1}^{n_{st'}} f\left(Y_{st'j'}|\mathbf{X}, \boldsymbol{\psi}_k^{(m)}\right)}{\sum_{l=1}^{K} \pi_l^{(m)} \prod_{t'=1}^{T} \prod_{j'=1}^{n_{st'}} f\left(Y_{st'j'}|\mathbf{X}, \boldsymbol{\psi}_l^{(m)}\right)} \tag{12}$$

where the superscript m is the iteration index of the algorithm. An important point to notice is that $w_{stjk}^{(m)}$ does not depend on t and j. This is peculiar to situations with replicate measurements for the clustered unit and ensures that when a species is assigned to a group, all its conspecifics are also assigned to the same group. In other words, posterior group probabilities are computed at the species level rather than at the individual tree level. We adopt the approximation used in Khalili and Chen (2007) to update the mixing proportions as

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} w_{stjk}^{(m)}$$

An improved update of the mixing proportions is provided in Städler *et al.* (2010).

In the M-step, the penalized log-likelihood is maximized for each component separately using the posterior probabilities of the observations as weights. This gives estimates for component k's parameters at the m-th iteration of the algorithm as

1. For the growth process

$$\hat{\beta}_k^{(m)} = \arg\max_{\beta_k} \left\{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{n_{st}} w_{stjk}^{(m-1)} \log f\left(\Delta D_{stj}/\Upsilon | X_{kj}^{G}\beta_k, \sigma_k^2\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\beta_k|}{|\hat{\beta}_k|}\right\} \tag{13}$$

where f is the density of the Gaussian distribution.

2. For the death process

$$\hat{\gamma}_k^{(m)} = \arg\max_{\gamma_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{j=1}^{n_{st}} w_{stjk}^{(m-1)} \log f\left(M_{stj} | X_{kj}^{M} \gamma_k\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\gamma_k|}{|\hat{\gamma}_k|} \right\} \tag{14}$$

where f is the probability mass function associated to the Bernoulli distribution.

3. For the recruitment process

$$\hat{\alpha}_k^{(m)} = \arg\max_{\alpha_k} \left\{ \sum_{s=1}^{S} \sum_{t=1}^{T} w_{stk}^{(m-1)} \log f\left(R_{st} | X_k^{R} \alpha_k\right) - \pi_k^{(m-1)} \eta_{nk} \frac{|\alpha_k|}{|\hat{\alpha}_k|} \right\} \tag{15}$$

where f is the probability mass function associated to the Poisson distribution.

2.2.2. *Number of components and species allocations*

The model fitting described in the previous paragraphs supposes that the number of groups K is known. In order to estimate it, we fit the finite mixture of GLM for $K = 1, 2, 3, \ldots$, and we select the value of K that minimizes an information criterion. Different criteria have been used, such as the Akaike information criterion (Akaike, 1974), the Bayesian information criterion (Schwarz, 1978), or the integrated completed likelihood criterion (ICL) (Biernacki *et al.*, 2000). We adopt the ICL, which has been specifically developed for mixture models and takes into account the quality of the classification. The ICL penalization is given by

$$\mathcal{P}_{\text{ICL}} = \nu_K log(n) + 2 \sum_{s=1}^{S} n_s \sum_{k=1}^{K} w_{sk} \log(w_{sk})$$

where the first term corresponds to the Bayesian information criterion penalization with ν_K equal to the number of free parameters in the model with K components, $n_s = \sum_{t=1}^{T} n_{st}$ is the number of tree observations for species s, and w_{sk} is the estimated posterior probability that species s belongs to group k (Equation 12). The maximum *a posteriori* estimate is then used to determine each species' allocation.

2.3. Mixture of inhomogeneous matrix models

The mixture of GLM gives K_g species groups for growth, K_r for recruitment and K_m for mortality. Crossing these classifications gives $K_g \times K_r \times K_m$ combinations of groups. These combinations are named $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$, with $1 \leqslant x \leqslant K_g, 1 \leqslant y \leqslant K_r$, and $1 \leqslant z \leqslant K_m$. Because of the additive recruitment, each of the K_r recruitment groups contributes to several combinations of groups. Therefore, the number of recruits $r_y(t)$ for recruitment group y must be distributed between the combinations $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$. The estimated number of recruits for the combination $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$ is computed as $\rho_{xyz}\, r_y(t)$, where $\rho_{xyz} = N_{xyz} / \sum_{x'} \sum_{z'} N_{x'yz'}$ is the ratio of the total number of alive trees in combination $\mathrm{g}_x\mathrm{r}_y\mathrm{m}_z$, N_{xyz}, over the total number of alive trees in recruitment group y, such that $\sum_x \sum_z \rho_{xyz} = 1$ for all y.

Each species exclusively belongs to one combination of groups. Because the parameters of the growth, mortality and recruitment models are estimated for each group, the look-up table assigning each species to growth group x, recruitment group y and mortality group z defines a matrix population model for it. Therefore, the combinations of groups define what we call *the mixture of inhomogeneous matrix models*.

3. APPLICATION

3.1. Simulations

We simulated mixture regression models with the true number of components set to three. We generated 30 species, and within each species, we sampled the number of trees from a Poisson(30). Within each tree in a given species, the number of repeated measures was sampled from a Poisson(15). To evaluate the effect of ignoring the time dependence in our model, we considered a first-order autoregressive correlation structure (AR1(ρ)) with varying correlation parameters $\rho = (0, 0.1, 0.3, 0.5, 0.7, 0.9)$; this autoregressive dependence was applied on the residuals for the Gaussian case and on the linear predictors for the Bernoulli and Poisson cases. The species were randomly assigned to the three groups with mixing proportions set to $\pi = (0.60, 0.25, 0.15)$. A total of five covariates were generated from a multivariate normal distribution with mean 0 and an AR1(0.7) covariance matrix. We also considered a scenario where the design matrix $\mathbf{X}$ has dependence structure across covariates with correlation of 0.5 in addition to the temporal AR1(0.7) correlation for repeated measures within the same covariate. The parameters associated to each covariate had a 0.5 probability of being zero, and the nonzero parameters were simulated as described in Table 1. We generated 50 datasets. For each simulation, a K-component mixture model was fit three times with different starting points for $K = 1, \ldots, 7$. We retained the fit and the K value that yield the lowest ICL. The computations were performed using the R software (R Core Team, 2014) by integrating functionalities of the `flexmix` (Leisch, 2004; Grün and Leisch 2007, 2008) and the `glmnet` (Friedman *et al.*, 2010) packages (see the Supporting information for the complete R code). For the algorithm to converge, it is necessary to use the same cross-validation partitioning across the EM iterations, that is, the subsamples for cross-validation must be defined at the beginning using the `foldid` option in the function `FLXMRglmnet` (see documentation in `glmnet`).

Table 1. Parameters used for simulations

Distribution	Intercept	Covariates coefficient	Variance
Gaussian	$\{-1.5, 0, 1.5\}$	$\mathcal{U}[-2, -1, 1, 2]$	1
Bernoulli	$\{-1.5, 0, 1.5\}$	$\mathcal{U}[-2, -1, 1, 2]$	—
Poisson	$\{-1, 0, 1\}$	$\mathcal{U}[-1, 1]$	—

Intercepts are fixed, one for each group, along a gradient of values set to -1.5, 0, 1.5 in the Gaussian and Bernoulli cases and equal to -1, 0, 1 in the Poisson case. The nonzero coefficients associated to the relevant covariates are randomly drawn from a discrete uniform distribution $\mathcal{U}$ in the set of values given between brackets.

3.2. The M'Baïki forest case study

3.2.1. The experimental site

We applied the method to the M'Baïki species-rich tropical rainforest ecosystem. The M'Baïki experimental site (3°54'N, 17°56'E) was established in a lowland semi-deciduous tropical rain forest of the Central African Republic. The average annual rainfall for the period 1981 – 2008 is 1739 mm with a 4-month dry season and an annual average monthly temperature of 24.9°C (Ouédraogo *et al.*, 2013). The M'Baïki experimental site consists of 10 permanent sample plots, each of 4 ha (200 m × 200 m), established in two forests less than 10 km apart (Figure 1). Two blocks of three plots each were established in the Boukoko forest and one block of four plots in the La Lolé forest (Bedel *et al.*, 1998). These permanent sample plots have been inventoried every year since 1982 (except in 1997, 1999 and 2001): all trees $\geq$ 10 cm dbh have been individually marked and spatially located and have been measured yearly for dbh. All species present have been identified, and dead trees and newly recruited trees with dbh $\geq$ 10 cm have been surveyed. The type of soil in all plot, except one, has been mapped.

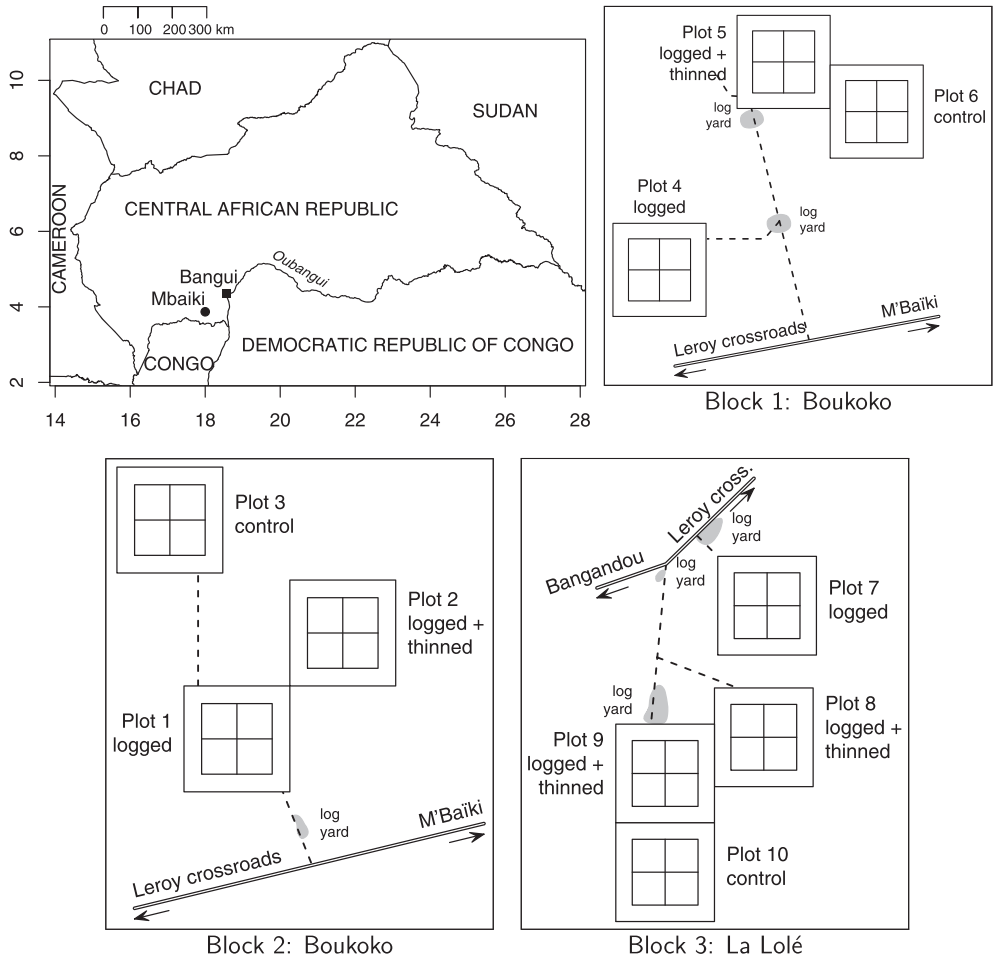

Figure 1. The M'Baïki forest experimental plots in the Central African Republic

Seven of the 10 plots across the three blocks were selectively logged between the 1984 and 1985 inventories. Three plots, one from each block, were left as controls. Logging consisted in harvesting trees with dbh $\geqslant$ 80 cm if belonging to one of 16 commercial species. Four of the seven plots logged (one from each of the Boukoko blocks and two from the La Lolé block) were thinned 2 years after logging to increase light penetration. Thinning consisted in poison girdling all nontimber trees with dbh $\geqslant$ 50 cm. This process was completed by cutting all lianas in the entire plot. The M'Baïki experimental site thus provides a perfect setting to observe the demographic processes across a wide range of disturbances, from undisturbed forests (unlogged plots) to highly disturbed forests (logged + thinned plots). Between 1982 and 2012, more than 37 000 trees from 230 genera have been monitored at this site. For this study, years for which complete data on the demographic processes and environmental variables are available were considered for analysis, resulting in $T = 18$.

3.2.2. *Growth, mortality and recruitment quantification*

The observations use an annual time step ($\Upsilon = 1$). To quantify the annual tree growth process, we calculated the annual tree diameter increments using only measurements from living trees that exhibit no trunk anomalies between two successive years. To further eliminate measurement errors, we only kept diameter increments between -0.4 cm (corresponding to stem shrinkage during dry seasons) (Baker *et al.*, 2002) and 4.456 cm, the 99th percentile of observed diameter increments of the fastest growing species *Musanga cecropioides* (Ouédraogo *et al.*, 2013).

The data were split into a training and a validation sets. The training dataset is taken to be Block 2 from the Boukoko forest and consists of three plots with the three different treatments (Figure 1). This block contains 197 species out of the 230 identified across all the M'Baïki plots and has data on 80 510 growth observations, 118 133 mortality observations and 42 816 recruitment observations. It is used to fit the growth, mortality and recruitment processes. The validation dataset consists of the other block in Boukoko and the block in La Lolé. It is used to evaluate the prediction quality of the mixture of inhomogeneous matrix models for plots sampled in contrasted environmental conditions.

To limit the discretization bias that may result from matrix modelling (Shimatani *et al.*, 2007; Picard *et al.*, 2010; Zuidema *et al.*, 2010), we use very thin dbh classes with a width of $d = 1$ cm. The time interval of the model has to be adjusted to the class width to meet the Usher assumption. This is achieved with a short time step of $\tau = 0.1$ year.

3.2.3. *Environmental covariates*

Five environmental variables and two variables describing the tree development stage were considered as potential covariates for the growth and mortality processes. The latter variables are the dbh and log-dbh (Di in cm and log-Di), which are commonly included in the model simultaneously to deal with the nonlinear association between dbh and growth (or mortality) (Zeide, 1993; Weiskittel *et al.*, 2011). The five environmental variables include two plot-level variables assessing competition for resources and three climate variables (see (Ouédraogo *et al.*, 2013) for details). The two competition indices are stand basal area (m^2 per hectares, BAst) and stand density (number of trees per hectares, Dst), which are computed on 1-ha subplots (100 m $\times$ 100 m) obtained as a subdivision of the initial 4-ha plots into four squares. This spatial unit was used because the environment is more homogeneous at this scale. The three climate variables are drought indices: the length of the dry season (number of months with rainfall $<$ 100 mm, LDS), the average rainfall during the dry season (RDS in millimetre) and the annual average soil water content (MSW in millimetre) (Ouédraogo *et al.*, 2013). For the recruitment process, potential predictors were restricted to BAst, Dst, LDS and RDS.

3.2.4. *Adjustment of the method to the M'Baïki forest*

The models were fit for each process using $K = 1, \ldots, 10$ groups. This was repeated 10 times with different initial random points for each K, and the fit with smallest ICL was chosen. The group structures for the growth and mortality processes were successfully identified. However, because of the large number of zeros in the recruitment, the mixture model did not work as well. We therefore made some adjustments to adapt the inference for this process. We assumed that the species groups identified for the growth process are nested within the groups of the recruitment process. This assumption is supported by the well-established positive correlation between species-specific recruitment rates and growth rates in disturbed forests, which is a direct consequence of the recruitment design that requires passing a 10 cm dbh threshold (Gourlet-Fleury *et al.*, 2005). Therefore, once we identified the growth groups, the recruitment groups were obtained by fitting a mixture of Poisson regression models to the number of recruits of the growth groups, instead of the number of recruits of the species.

A second adjustment to the general framework presented earlier was made to deal with species that could not be classified for various reasons, including situations in which the species were not available in the training data, environmental covariates were missing for the species, or the species had a single individual measurement. The strategy we adopted is presented in Section 4.

4. RESULTS

4.1. Simulations

The algorithm performs quite well even when the dependence across time is not taken into account. We are able to identify the correct number of underlying clusters for all the different processes with correlations as high as 0.9 between consecutive repeated measures (Figure 2). We use two matching indices, I_1 and I_2 (Mortier *et al.*, 2013), to assess the clustering performance and compare each species group allocation based on the maximum *a posteriori* estimate to the true group membership. These indices are based on the $K \times \hat{K}$ contingency table

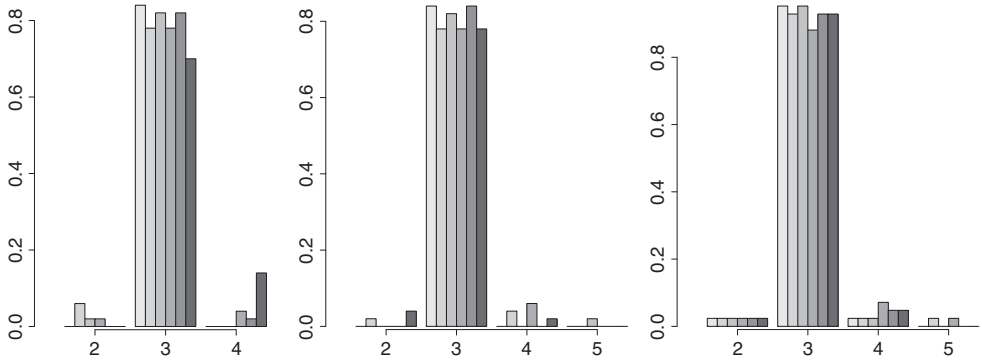

Figure 2. Distribution of the estimated number of groups based on 50 replications of a simulated dataset with three groups, when observations have either a (a) Gaussian, (b) Poisson or (c) Bernoulli distribution. We considered a first-order autoregressive correlation structure (AR1(ρ)) with varying correlation parameters from 0 (light grey) to 0.9 (dark grey)

$C = (C_{ij})$ with $i = 1, \ldots, K$ and $j = 1, \ldots, \hat{K}$ that cross-tabulates the species according to their true and estimated classifications:

$$I_1 = \frac{1}{S}\sum_{i=1}^{K} \max\left\{C_{i1}, \ldots, C_{i\hat{K}}\right\} \qquad I_2 = \frac{1}{S}\sum_{j=1}^{\hat{K}} \max\{C_{1j}, \ldots, C_{Kj}\}$$

These indices vary between $1/S$ and 1 with higher values corresponding to better classifications. For $\hat{K} = K$, we obtain 98% of the time $I_1 = I_2 = 1$. When considering $\hat{K} = K + 1$ (which occurred rarely), we obtain 93% of the time $I_1 = 1$ and the few instances where $I_1 < 1$ are due to a group being split into two subgroups (I_2 is always lower than one by construction).

The algorithm is also effective at selecting the component-specific relevant covariates for all the distribution types (Gaussian, Bernoulli or Poisson). For example, in the more complex scenario where the design matrix **X** has both temporal dependence and correlated covariates, we obtain the following results: in the Gaussian case, out of the 50 simulations, one false positive is included one time; in the Bernoulli case, one false positive is selected five times, and there is a single instance of two false negatives; in the Poisson case, one, three or four false positives are selected one time each, and there is a single instance of one false negative.

4.2. The M'Baïki forest case study

4.2.1. Species classification and ecological meaning

Six groups are identified for the growth process, labelled g_1 to g_6 in order of increasing maximum growth rate, which is used as a proxy for light requirement. These six groups are nested within four recruitment groups, $r_1, \ldots, r_4$: g_2 and g_6 correspond to r_2, g_5 and g_4 match with r_1, g_3 with r_4 and group g_1 constitutes r_3. We also identify three mortality groups, labelled m_1 to m_3. The growth ordering does not parallel the mortality ordering, and no obvious relationship can be found between growth and mortality groups. The ICL curves as well as parameter estimates are presented in the Supporting information.

Crossing these classifications gives $6 \times 4 \times 3 = 72$ possible combinations of groups, of which only 15 are nonempty. Accordingly, the mixture of matrix models is composed of 15 transition matrices. The nonempty combinations of groups contain between a single species up to 24 species (with known regeneration guild, (Bénédet *et al.*, 2014)) and correspond to groupings that are biologically meaningful, especially in terms of regeneration guild (Table 2). Moreover, the clusters uncovered by the mixture of Usher matrix models group species according to both their maximum growth rate and their maximum diameter (95th percentile). When plotting species along these two axes, the combinations of groups are well separated (Figure 3). Because these two axes can be used to order species along a continuum of ecological strategies (Turner, 2001; Alder *et al.*, 2002), this provides evidence that the mixture of inhomogeneous Usher matrix models is able to cluster species in a way that is consistent with their autecology (Picard *et al.*, 2012).

4.2.2. Prediction results, correction factors and asymptotic state

Among the 230 tree species at M'Baïki, 12 were not considered for analysis for various reasons (missing covariates and lack of replicate measurements) and remained unclassified. Out of the 218 tree species retained for analysis, 21 are not present in the training set but are present in the validation dataset and are classified *a posteriori*. It is still necessary to account for the 12 unclassified species when computing the stand basal area (Bast(t)) and the stand density (Dst(t)) to avoid underestimating these two competition indices. Hence, correction factors c_B and c_D are applied to Bast(t) and Dst(t), respectively. Factor c_B is computed as the ratio of the total stand basal area in 1992 over the cumulated basal area of classified species in 1992: $c_B = 1.00259\,(\pm 0.00027)$. Factor c_D is computed as the ratio of the total number of trees in 1992 over the cumulated number of trees from species that were classified in 1992: $c_D = 1.000351\,(\pm 0.00011)$.

Table 2. Floristic characteristics of the combinations of growth, recruitment and mortality groups identified at M'Baïki: number of species in each combination (size), regeneration guild (guild), phenology and dominant species.

	Classification characteristics			
Group	Size	Guild	Phenology	Dominant species
$g_1r_3m_3$	4	SB	Ever	Garcinia smeathmannii
$g_1r_3m_1$	20	NPLD-SB	Dec	Canarium schweinfurthii
$g_2r_2m_1$	24	NPLD-SB	Dec	Entandrophragma candollei
$g_2r_2m_2$	3	SB	Ever	Cola altissima
$g_2r_2m_3$	4	SB	Ever	Afrostyrax lepidophyllus
$g_3r_4m_2$	3	SB	Dec	Monodora myristica
$g_3r_4m_1$	24	NPLD-SB	Dec	Entandrophragma utile
$g_3r_4m_3$	1	P	Ind	Zanthoxylum lemairei
$g_4r_1m_2$	1	NPLD	Ever	Pycnanthus angolensis
$g_4r_1m_1$	22	NPLD	Dec	Entandrophragma angolense
$g_5r_1m_2$	1	P	Ind	Dictyandra arborescens
$g_5r_1m_1$	21	NPLD	Dec	Lovoa trichilioides
$g_5r_1m_3$	3	NPLD	Dec	Entandrophragma cylindricum
$g_6r_2m_3$	2	P	Ever	Cleistopholis glauca
$g_6r_2m_1$	11	P	Dec	Terminalia superba

SB, shade bearer; NPLD, nonpioneer light demander; P, pioneer; Ever, evergreen; Dec, deciduous; and Ind, unknown phenology.
Regeneration guild is determined for each group based on two aspects: the guild of the species with the largest number of trees in the group and the guild that contains the most species in the group. In most cases, the two agree, but when they are different, we provide both (e.g., NPLD-SB). Dominant species means that this species has the highest number of trees in the group.

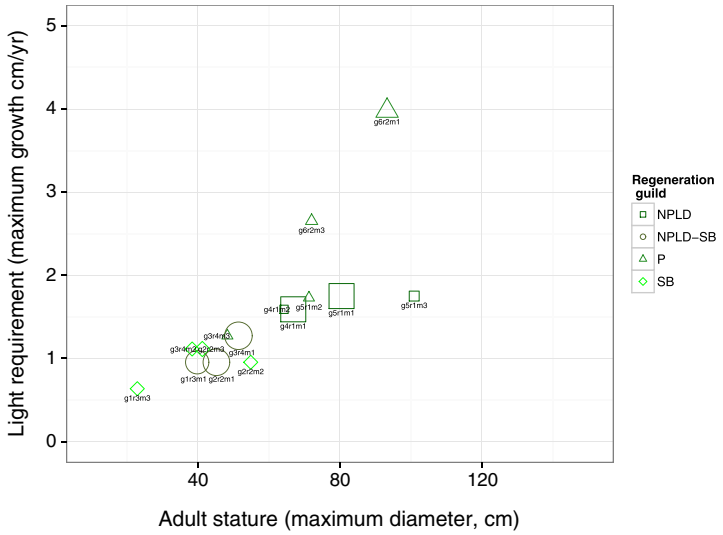

Figure 3. Projection of the species clustering obtained by the inhomogeneous mixture of Usher matrix models at M'Baïki on the two axes corresponding to the maximum diameter and the maximum growth rate. The labels $g_x r_y m_z$ correspond to the identified species groups. Each symbol corresponds to the dominant regeneration guild of each group. The size of the symbol is proportional to the number of species in the group

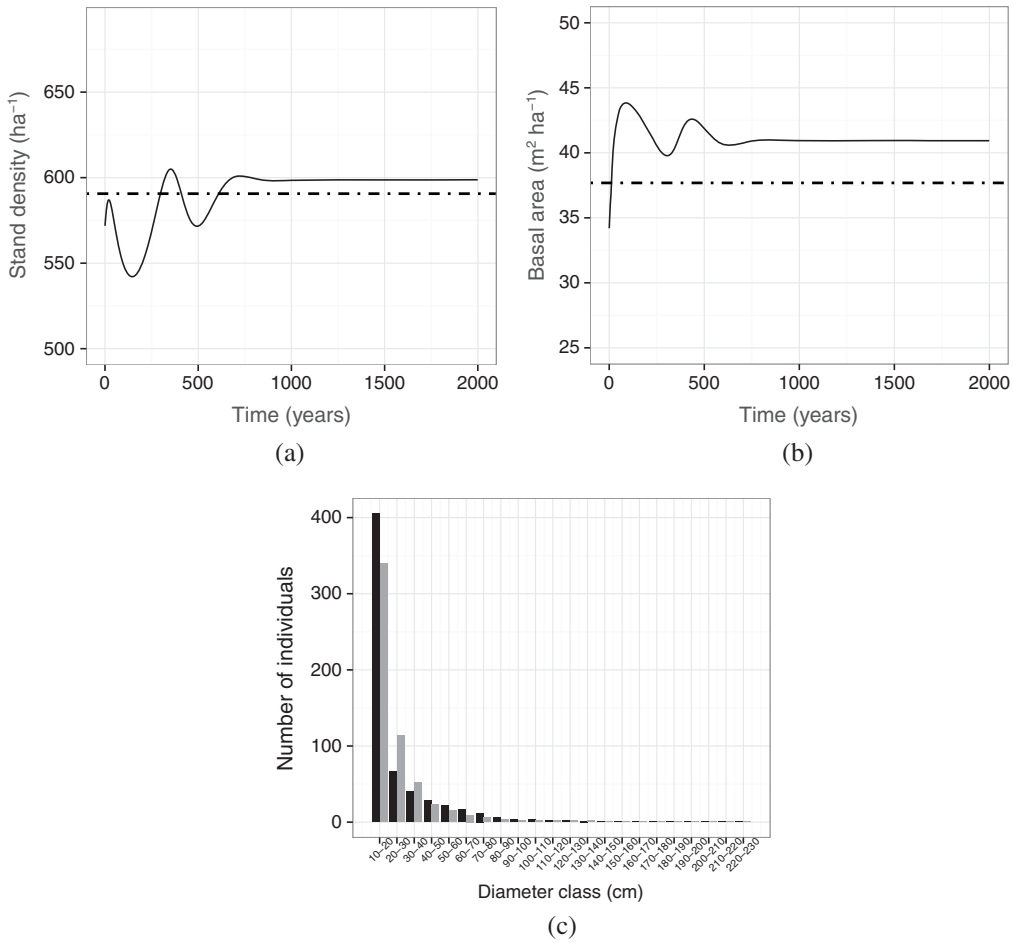

Figure 4. Density (number of tree per hectares), basal area (per hectares) and diametric structure. In (a) and (b), the solid lines correspond to the simulated forest and the dot-dashed lines to the observed stand in 2012 on the validation blocks. In (c), the black bars correspond to the simulated forest and the grey ones to the observed stand

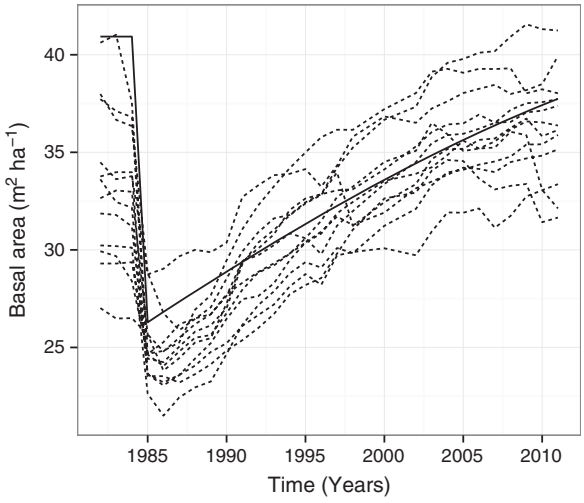

Figure 5. Dynamics of the basal area after logging (solid line: prediction; dashed lines: observations from 1982 to 2012 in the logged plots of the validation blocks)

Year 1992 is chosen because information for all processes and environmental variables is available from this time on. The two competition indices are then computed from the vector of the number of trees as $\mathrm{BAst}(t) = c_B \times \sum_s \mathbf{B}'\mathbf{N}_s(t)$ and $\mathrm{Dst}(t) = c_D \times \sum_s \mathbf{1}'\mathbf{N}_s(t)$, where $\mathrm{B} = \left(\frac{\pi}{4}\mathrm{D}_i^2\right)_{i=1\ldots I}$ is the vector of mean basal area for each diameter class, $\mathbf{1}$ is a vector of ones of length I and prime denotes the transpose operator.

The results of the simulated forest dynamics using the inhomogeneous matrix model over 2000 years starting with the observed forest stand in 1992 is shown in Figure 4 (see the Supporting information for the complete R code). The predicted asymptotic tree density, basal area and dbh structure match the observations of the validation data in 2012. In addition, the observed dbh distribution in 2012 at M'Baïki has an inverse-J shape that is typical of natural rain forests (Figure 4(c)). It could be fit by an exponential distribution with parameter 0.0724 (standard error 0.0047). In comparison, the predicted dbh distribution also presents an inverse-J shape and can be fit by an exponential distribution with parameter 0.0695.

We also compared the predicted dynamics following a 28-year wait after disturbance of the asymptotic state to the observed dynamics between 1982 and 2012 in the logged plots of the validation dataset (Figure 5). The simulated disturbance for the asymptotic state consisted of removing with probability 1/2 trees with dbh greater than 80 cm from the asymptotic dbh distribution. This corresponds to a perturbation of the same magnitude as the one realized in 1984 at M'Baïki in terms of lost basal area but performed on a wider range of species. The model successfully predicts the reconstitution rate of the basal area after disturbance (slope of dynamics): the predicted rate is 0.4329, while the observed rates in the logged plots of the validation data have a mean of 0.4517 and standard error 0.0929.

5. DISCUSSION

The proposed mixture of inhomogeneous matrix models is an original method that simultaneously fits matrix population models for species-rich ecosystems, clusters species into ecologically meaningful groups and selects relevant environmental covariates. As such, it is an integrated alternative to classical methods for building matrix population models, for classifying species or for selecting variables in regression models. The coupling of modern covariate selection methods and mixture model approaches that we have put forward in the mixture of inhomogeneous matrix models can be straightforwardly incorporated into any model where individual growth is regressed against size and environmental covariates. In particular, it could also be implemented in individual-based models (Dunstan *et al.*, 2011) or in integral projection models (Zuidema *et al.*, 2010).

Compared with other modelling approaches, the mixture of inhomogeneous matrix models combines the power of modern and technically complex statistical methods with the simplicity of matrix modelling. In this paper, we considered a few potential covariates, but the proposed method has the flexibility to handle a large number of covariates and select the relevant ones to model the dynamics and refine the predictions. For example, species-specific functional traits, such as the 99th percentile of diameter or wood density, as proposed by Hérault *et al.* (2011) could be included as potential covariates. For the front-end user, the model is as simple to use as any other matrix model. We thus expect the mixture of inhomogeneous matrix models to be useful in all application areas where matrix population models have been found to be useful decision tools, such as population viability analysis (Morris and Doak, 2002) or the management of wildlife population with harvest (Jensen, 1996), in particular when operating in a variable environment.

Taking into account environmental variability in matrix models is crucial to better understand and predict consequences of environmental variations on population dynamics. In the particular case of the M'Baïki tropical rain forest, we demonstrated the model's ability to reproduce the stand structure at equilibrium and the dynamics after disturbance. We showed, using simple exploitation rules, that the model could successfully reproduce post-logging dynamics over a 25-year period. Climate variables were also included in the environmental variables, thus paving the way for predicting the impact of climate change (Liang *et al.*, 2011), including the change in species composition or the interaction between disturbance and climate change, caused by the species differentiated responses to climate. The role of climate in forest dynamics at M'Baïki will be investigated in a future study.

Further work should be pursued to address some issues that were not taken into account in this paper. In particular, (i) explicitly modelling the time dependence between observations within the same tree, (ii) addressing the zero inflation in the recruitment process and (iii) investigating the impact of imbalanced class distributions on the results of the mixture models. For the first, mixed models offer a flexible method to handle longitudinal dependence (Bondell *et al.*, 2010; Schelldorfer *et al.*, 2014). Our method can be extended to accommodate this by considering mixtures of generalized linear mixed models with variable selection. However, this is computationally challenging and requires the development of efficient algorithms. For the second, zero-inflated distributions provide a general framework to overcome the presence of a large number of zeros (Flores *et al.*, 2009). However, the challenge of using zero-inflated models in the context of model-based clustering is the complexity of nesting two levels of mixtures: one corresponding to the mixture of a point mass at zero and a Poisson (or negative binomial) distribution and the other corresponding to the mixture of distributions used to identify groups of species. For the imbalanced class distribution issue, which may compromise the performance of clustering, sampling methods, such as random undersampling (Tseng and Wong, 2005), are commonly used to achieve a more balanced distribution. The integration of such sampling strategies with ensemble learning methods, such as bagging (Breiman, 1996) and boosting (Friedman, 2000), has been shown to improve the performance of imbalanced data classification/clustering (He and Garcia, 2009). However, the problem is more complicated in our context, where the clustering is performed at the species level and the imbalanced distribution occurs both at the level of the species and the varying number of trees within species.

Finally, we have fit the growth, mortality and recruitment models separately. This ensures an optimal fit for each dynamic component. However, because growth, mortality and recruitment are nonlinearly combined into the matrix model, this does not ensure an optimal fit at the matrix model level. Combining equations estimated separately may induce a prediction bias at the population level. Although scarcely documented in the scientific literature, this prediction bias is a well-known issue among forest modellers and occurs in different types of forest dynamic models. The problem is usually addressed by tuning *a posteriori* some coefficients (Favrichon, 1998). An alternative to deal

with this problem and a possible extension of our proposed model would be to formulate a unified approach that allows the fit of the three demographic processes simultaneously using an integrated population model (Abadi *et al.*, 2010). This can be achieved within a Bayesian hierarchical framework (Cressie *et al.*, 2009) by defining a first level that models the number of trees in a diameter class $\mathbf{N}_s(t)$ conditionally on the growth, mortality and recruitment processes and a second level that models these demographic processes using mixture models with variable selection similarly to the method we have proposed here.

Acknowledgements

This research was supported by the CoForChange project (http://www.coforchange.eu/) funded by the ERA-Net BiodivERsA with the national funders ANR (France) and NERC (UK), part of the 2008 BiodivERsA call for research proposals involving 16 European, African and international partners including a number of timber companies (see the list on the website, http://www.coforchange.eu/partners), and by the CoForTips project funded by the ERA-Net BiodivERsA with the national funders FWF (Austria), BelSPO (Belgium) and ANR (France), part of the 2011–2012 BiodivERsA call for research proposals (http://www.biodiversa.org/519). We would also like to thank the two anonymous referees for their constructive comments.

REFERENCES

Abadi F, Gimenez O, Arlettaz R, Schaub M. 2010. An assessment of integrated population models: bias, accuracy, and violation of the assumption of independence. *Ecology* **91**(1):7–14.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6):716–723.

Alder D, Oavika F, Sanchez M, Silva JNM, Van der Hout P, Wright HL. 2002. A comparison of species growth rates from four moist tropical forest regions using increment-size ordination. *International Forestry Review* **4**(3):196–205.

Baker TR, Affum-Baffoe K, Burslem D, Swaine MD. 2002. Phenological differences in tree water use and the timing of tropical forest inventories: conclusions from patterns of dry season diameter change. *Forest Ecology and Management* **171**(3):261–274.

Bedel F, Durrieu de Madron L, Dupuy B, Favrichon V, Maître H, Bar-Hen A, Narboni P. 1998. Dynamique de croissance dans des peuplements exploités et éclaircis de forêt dense africaine. le dispositif de m'baïki en république centrafricaine (1982-1995). *CIRAD Forêt, Montpellier Série FORAFRI, document* **1**:1–72.

Bellwood D, Wainwright P. 2001. Locomotion in labrid fishes: implications for habitat use and cross-shelf biogeography on the great barrier reef. *Coral Reefs* **20**(2):139–150.

Bénédet F, Vincke D, Fayolle A, Doucet F, Gourlet-Fleury S: Cofortraits, African plant traits information database. version 1.0. http://coforchange.cirad.fr/african_plant_trait, access to database can be granted upon request. [accessed on 24 September, 2013]

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7):719–725.

Bondell H, Krishna A, Ghosh S. 2010. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**(4):1069–1077.

Breiman L. 1996. Bagging predictors. *Machine Learning* **24**(2):123–140.

Brown P, Vannucci M, Fearn T. 1998. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**(3): 627–641.

Buongiorno J, Gilless J. 2003. *Decision Methods for Forest Resource Management*. Academic Press: Elsevier Science (USA).

Buongiorno J, Michie B. 1980. A matrix model of uneven-aged forest management. *Forest Science* **26**(3):609–625.

Caswell H. 2001. *Matrix Population Models, Construction, Analysis, and Interpretation* deuxième edition. Sinauer Associates, Inc. Publishers: Sunderland, Massachusetts.

Cressie N, Calder CA, Clark JS, Wikle CK. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19**(3):553–570.

Crone E, Menges E, Ellis M, Bell T, Bierzychudek P, Ehrlén J, Kaye T, Knight T, Lesica P, Morris W, Oostermeijer G, Quintana-Ascencio P, Stanley A, Ticktin T, Valverde T, Williams J. 2011. How do plant ecologists use matrix population models? *Ecology Letters* **14**(1):1–8.

Dempster A, Laird N, Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**(1):1–38.

Demyanov V, Wood S, Kedwards T. 2006. Improving ecological impact assessment by statistical data synthesis using process-based models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55**(1):41–62.

Dunstan P, Foster S, Hui F, Warton D. 2013. Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics* **18**(3):357–375.

Dunstan PK, Foster SD, Darnell R. 2011. Model based grouping of species across environmental gradient. *Ecological Modelling* **222**(4):955–963.

Favrichon V. 1994. Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d'un modèle de dynamique de peuplement en forêt guyanaise classification of guiana forest tree species into functional groups for a model of vegetation dynamic. *Revue d'écologie* **49**:379–403.

Favrichon V. 1998. Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest Science* **44**(1):113–124.

Fieberg J, Ellner S. 2001. Stochastic matrix models for conservation and management: a comparative review of methods. *Ecology Letters* **4**(3):244–266.

Flores O, Rossi V, Mortier M. 2009. Autocorrelation offsets zero-inflation in models of tropical saplings density. *Ecological Modeling* **220**(15):1797–1809.

Friedman J. 2000. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**(5):1189–1232.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.

García-López J, Allué C. 2011. Modelling phytoclimatic versatility as a large scale indicator of adaptive capacity to climate change in forest ecosystems. *Ecological Modelling* **222**(8):1436–1447.

Gitay H, Noble I. 1997. *What are Functional Types and How Should We Seek Them?* Cambridge University Press: Cambridge,3–19.

Gourlet-Fleury S, Blanc L, Picard N, Sist P, Dick J, Nasi R, Swaine MD, Forni E. 2005. Grouping species for predicting mixed tropical forest dynamics: looking for a strategy. *Annals of Forest Science* **62**(8):785–796.

Grün B, Leisch F. 2007. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis* **51**:5247–5252.

Grün B, Leisch F. 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**(4):1–35.

Gupta M, Ibrahim J. 2007. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102**(479):867–880.
Hargrove W, Hoffman F. 2004. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management* **34**(1):39–60.
He H, Garcia E. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9):1263–1284.
Hérault B, Bachelot B, Poorter L, Rossi V, Bongers F, Chave J, Paine C, Wagner F, Baraloto C. 2011. Functional traits shape ontogenetic growth trajectories among rain forest tree species. *Journal of Ecology* **99**(6):1431–1440.
Hui FC, Warton DI, Foster SD, Dunstan PK. 2013. To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* **94**(9):1913–1919.
Jensen AL. 1996. Density-dependent matrix yield equation for optimal harvest of age-structured wildlife populations. *Ecological Modelling* **88**(1–3): 125–132.
Khalili A, Chen J. 2007. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**(479):1025–1038.
Leisch F. 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **11**(8):1–18.
Liang J. 2010. Dynamics and management of Alaska boreal forest: an all-aged multi-species matrix stand growth model. *Forest Ecology and Management* **260**(4):491–501.
Liang J, Zhou M, Verbyla D, Zhang L, Springsteen AL, Malone T. 2011. Mapping forest dynamics under climate change: a matrix model. *Forest Ecology and Management* **262**(12):2250–2262.
McLachlan G, Krishnan T. 2008. *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics. Wiley Series in Probability and Statistics: New York.
Monni S, Tadesse M. 2009. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis* **4**(3):413–436.
Morris WF, Doak DF. 2002. *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis*. Sinauer Associates, Inc.: Sunderland, MA, 480 pp.
Mortier F, Rossi V, Guillot G, Gourlet-Fleury S, Picard N. 2013. Population dynamics of species-rich ecosystems: the mixture of matrix population models approach. *Methods in Ecology and Evolution* **4**(4):316–326.
Ouédraogo DY, Mortier F, Gourlet-Fleury S, Freycon V, Picard N. 2013. Slow-growing species cope best with drought: evidence from long-term measurements in a tropical semi-deciduous moist forest of Central Africa. *Journal of Ecology* **101**(6):1459–1470.
Pastor J, Post W. 1988. Response of northern forests to CO2-induced climate change. *Nature* **334**:55–58.
Pearson R, Dawson T, Berry P, Harrison P. 2002. Species: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling* **154**(3):289–300.
Picard N, Köhler P, Mortier F, Gourlet-Fleury S. 2012. A comparison of five classifications of species into functional groups in tropical forests of French Guiana. *Ecological Complexity* **11**:75–83.
Picard N, Mortier F, Chagneau P. 2008. Influence of estimators of the vital rates in the stock recovery rate when using matrix models for tropical rainforests. *Ecological Modelling* **214**(2–4):349–360.
Picard N, Ouédraogo DY, Bar-Hen A. 2010. Choosing classes for size projection matrix models. *Ecological Modelling* **221**(19):2270–2279.
Prentice IC, Sykes M, Cramer W. 1993. A simulation model for the transient effects of climate change on forest landscapes. *Ecological Modelling* **65**(1–2):51–70.
R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: Austria.
Rogers-Bennett L, Rogers D. 2006. A semi-empirical growth estimation method for matrix models of endangered species. *Ecological Modelling* **195**(3–4):237–246.
Scheiter S, Higgins SI. 2009. Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach. *Global Change Biology* **15**(9):2224–2246.
Schelldorfer J, Meier L, Bühlmann P, Winterthur AXA, Zürich ETH. 2014. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using $\ell 1$-penalization. *Journal of Computational and Graphical Statistics* **23**(2):460–477.
Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**(2):461–464.
Shao G. 1996. Potential impacts of climate change on a mixed broadleaved-Korean pine forest stand: a gap model approach. *Climatic Change* **34**(2):263–268.
Shimatani I, Kubota Y, Araki K, Aikawa SI, Manabe T. 2007. Matrix models using fine size classes and their application to the population dynamics of tree species: Bayesian non-parametric estimation. *Plant Species Biology* **22**(3):175–190.
Shugart H, West D. 1980. Forest succession models. *BioScience* **30**(5):308–313.
Solomon A. 1986. Transient response of forests to CO2-induced climate change: simulation modeling experiments in eastern North America. *Oecologia* **68**(4):567–579.
Städler N, Bühlmann P, Van De Geer S. 2010. $\ell 1$-penalization for mixture regression models. *Test* **19**(2):209–256.
Stankowski P, Parker WH. 2010. Species distribution modelling: does one size fit all? A phytogeographic analysis of Salix in Ontario. *Ecological Modelling* **221**(13–14):1655–1664.
Steneck R, Dethier M. 1994. A functional group approach to the structure of algal-dominated communities. *Oikos* **69**(3):476–498.
Stott I, Townley S, Carslake D, Hodgson D. 2010. On reducibility and ergodicity of population projection matrix models. *Methods in Ecology and Evolution* **1**(3):242–252.
Swaine M, Whitmore T. 1988. On the definition of ecological species groups in tropical rain forests. *Vegetation* **75**(1–2):81–86.
Talkkari A, Kellomäki S, Peltola H. 1999. Bridging a gap between a gap model and a physiological model for calculating the effect of temperature on forest growth under boreal conditions. *Forest Ecology and Management* **119**(1–3):137–150.
Tseng G, Wong W. 2005. Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**(1):10–16.
Turner IM. 2001. *The Ecology of Trees in the Tropical Rain Forest*. Cambridge University Press: Cambridge.
Usher M. 1966. A matrix approach to the management of renewable resources, with special reference to selection forests. *Journal of Applied Ecology* **3**(2):355–367.
Usher M. 1969. A matrix model for forest management. *Journal of Biometric Society* **25**(2):309–315.
Weiskittel A, Hann D, Kershaw J,Jr, Vanclay J. 2011. *Forest Growth and Yield Modeling*. Wiley: Chicester.
Zeide B. 1993. Analysis of growth equations. *Forest Science* **39**(3):594–616.
Zou H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476):1418–1429.
Zuidema P, Jongejans E, Chien P, During H, Schieving F. 2010. Integral projection models for trees: a new parameterization method and a validation of model output. *Journal of Ecology* **98**(2):345–355.

SUPPORTING INFORMATION

Additional information and supplementary material for this article, including R code, are available online at the journal's website.

