Package `jSDM` functionalities

Jeanne Clément

2022-06-27

Table des matières

1 Abstract

1. Species distribution models (SDMs) are commonly used in ecology to predict the ecological niche of species. However SDMs do not take into account interactions between species to estimate species distributions. Joint Species Distribution Models (JSDMs), which have recently emerged in ecology, allow to take into account co-occurrences between species to predict their distribution (Warton et al. 2015). Moreover these models provide a conceptual framework for integrating functional traits to explain differences in occurrence between species (Warton et al. 2015; Ovaskainen et al. 2017). Several libraries have been developed to fit such models following different statistical approaches (R packages `Hmsc` (Ovaskainen et al. 2017), `boral` (Hui 2016) …). However, these libraries may have some limitations. Not all of them allow the handling of large data-sets, the management of missing data and of data with multiple visits by site or the monitoring of latent variables convergence in JSDMs.

2. Here, we present the recently developed `jSDM` R package (https://ecology.ghislainv.fr/jSDM) for fitting JSDMs, to overcome the limitations of current libraries. Parameter inference is done in a hierarchical Bayesian framework and uses optimized C++ code to allow the analysis of large data-sets, with a high number of parameters in a limited amount of time.

3. We compared the calculation time, accuracy, and predictive performance of the `Hmsc`, `boral`, and `jSDM` libraries for fitting JSDMs on six real data-sets (Wilkinson et al. 2019; Choler 2005; Borcard and Legendre 1994) and one simulated. We also used the package `jSDM` for fitting a JSDM including species traits to analyze the response of alpine plants (Choler 2005) to snowmelt date, according to their specific leaf area (SLA). Furthermore, we developed a method to improve latent variables convergence, by selecting species with constrained factor loadings. Moreover, we used this package to fit a JSDM for tree species in Madagascar, combining tree occurrences from a large forest inventory and climate data. We used our model to derive distribution maps both at the species and community level (species richness and β-diversity maps).

4. We expect reduced computational time using `jSDM` rather than `Hmsc` or `boral` for fitting JSDMs with equivalent or better accuracy and predictive performance. JSDMs can be used to understand better the environmental filtering based on species functional traits. Also, community maps can be used for systematic conservation planning identifying priority areas for conservation based on species richness or uniqueness.

2 Introduction

2.1 State of the art

One of the key challenges in ecology is to understand the structure and assembly dynamics of ecological communities, shaped by both abiotic and biotic filtering (Vellend 2010). Indeed, the abundances and distributions of species depend not only on their individual responses to the abiotic environment, but also on their interactions (Wisz et al. 2013). During the last two decades, species distribution models (SDMs)

(Elith and Leathwick 2009)) have been commonly used in ecology to predict the ecological niches of species, independently of each other. However, SDMs do not take into account interactions between species and very often only consider environmental filtering, to predict the occurrence of species (abundance or probability of presence). Moreover, these correlative models do not allow to explain the differences in occurrence between species, via their functional traits for example. In addition, stacked SDMs (SSDMs) (Schmitt et al. 2017) are widely used to perform community-level predictions (e.g., community composition or species richness). This method combines multiple individual SDMs to produce a community-level model. It thus assumes that species respond individualistically to variation in environmental conditions (Guisan and Rahbek 2011), whereas interacting communities of species are more than the sum of their component species.

Methodological advances in multiple-species distribution modelling have recently experienced a rapid expansion, with the emergence of joint species distribution models (JSDMs) (Golding, Nunn, and Purse 2015; Hui 2016; Ovaskainen et al. 2017; Clark et al. 2017; Niku et al. 2019; Pichler and Hartig 2021), which allow to take into account interactions between species to predict their occurrence. Moreover these models provide a conceptual framework for integrating phylogeny relationships and functional traits to explain differences in occurrence between species (Warton et al. 2015; Ovaskainen et al. 2017), assuming, for example, that closely related species with like traits have more similar environmental niches, than expected by random. These models can thus be expected to have better predictive performance than SSDMs in terms of community composition or species richness, because biotic interactions represent an important driver of species local coexistence (Wisz et al. 2013). Indeed recent literature, as the article Norberg et al. (2019), supports this hypothesis. Even if it might be assumed that considering each species separately is more consistent with our limited current understanding of community assembly, because it can be difficult to predict *a priori* the joint structure of entire species assemblages responses to environmental variations (Norberg et al. 2019). Actually, treating each species individually, via SDMs, may involve a higher risk of overfitting, while borrowing information from other species, using JSDMs, may improve predictive performance, if the species respond similarly enough to abiotic variation (Ovaskainen and Soininen 2011; Hui et al. 2013). Indeed, recent literature on species distribution modelling shows that the JSDM approach, by assuming shared responses to the environment, can improve predictive performance especially for rare species, which can borrow information from other more common species (Ovaskainen and Soininen 2011; Hui et al. 2013; Tikhonov et al. 2017; Norberg et al. 2019). In fact, common species may present more reliable statistical relationships with environmental variables than rare species with large and dispersed distributions (Segurado and Araújo 2004), so considering species assemblages as a whole may actually improve the relevance of estimated environmental relationships for rarer species within communities (Ovaskainen and Soininen 2011; Hui et al. 2013). Thus, the use of JSDMs can be expected to be generally beneficial in community ecology studies, since most ecological communities consist of a few common and many rare species. Consequently, over the last few years, many libraries have been developed to fit JSDMs following different statistical approaches. Some of them, as well as their main features, are presented in the Table 1.

Table 1 – *Non-exhaustive list of softwares packages designed for fitting JSDMs with their main features. In the following table MC means Monte Carlo, Neg.Bin is for negative binomial distribution, LVM means latent variable model and LV latent variable.*

Softwares	jSDM	Hmsc	boral	sjSDM	gjam	BayesComm	gllvm	PyMC3, greta, JAGS, Stan
Model name	Joint species Distribution Models	Hierarchical Modeling of Species Communities	Bayesian Ordination and Regression Analysis	Scalable Joint species Distribution Models	Generalized Joint Attribute modeling	Bayesian community ecology Analysis	Generalized Linear Latent Variable Models	Probabilistic programming
Statistical inference framework	Gibbs sampler	Gibbs sampler	Gibbs sampler (JAGS)	Maximum Likelihood Estimate (MLE) MC approximation	Gibbs sampler	Gibbs sampler	Laplace or variational approximation	Hamiltonian MC Except for JAGS Gibbs sampler
Probabilistic programming	0	0	0	0	0	0	0	1
Speed								
CPU	1	1	1	1	1	1	1	1
GPU	0	0	0	1	0	0	0	1 (PyMC3, greta)
Language	C++/R	R	R/C++	Python/R	R/C++	R/C++	R	Python, R/C++, C++, C++
Data format	wide, long	wide	wide	wide	wide	wide	wide, long	wide, long
Data type								
Presence-absence	1	1	1	1	1	1	1	1
Count	1	1	1	1	1	0	1	1
Missing data	1	0	0	0	1	0	0	0
Presence only	0	0	0	0	1	0	0	0
Distribution (link)								
Bernoulli (probit)	1	1	1	1	0	1	1	1
Bernoulli (logit)	1	0	0	1	0	0	1	1
Binomial(probit)	0	0	1	0	0	0	0	1
Binomial (logit)	1	0	0	0	0	0	0	1
Poisson (log)	1	1	1	1	0	0	1	1
Neg.Bin (log)	0	0	1	0	0	0	1	1
Normal (identity)	0	1	1	1	1	0	1	1
Site effect								
None	1	1	1	1	0	1	1	1
Fixed	1	0	1	0	0	0	1	1
Random	1	1	1	0	1	0	1	1
Species effect								
Fixed	1	0	1	1	1	1	1	1
Random	0	1	0	0	1	0	1	1
Species traits	1	1	1	0	1	0	1	1
Spatial variation	0	1	1	1	0	0	1	1
LVM	1	1	1	0	1 f large data-set, 0 otherwise	0	1	1
Factor loadings								
Fixed	1	0	1	0	0	0	0	1
Random	0	1	0	0	1	0	1	1
Constraint choice	1	0	0	0	0	0	0	1
Number of LV								
Fixed	1	1	1	0	1	0	1	1
Estimated	0	1	0	0	0	0	0	1
User friendly								
Examples	1	1	1	1	1	1	1	1
Vignettes	1	1	0	1	1	0	1	1
Additional functions								
Predict	1	1	1	1	1	1	1	0
Residual correlation	1	1	1	1	1	1	1	0
Reference		Ovaskainen et al. (2017)	Hui (2016)	Pichler and Hartig (2021)	Clark et al. (2017)	Golding, Nunn, and Purse (2015)	Niku et al. (2019)	Salvatier, Wiecki, and Fonnesbeck (2016), Golding (2019), Plummer (2003), Stan Development Team (2021)
URL	jSDM	Hmsc	boral	sjSDM	gjam	BayesComm	gllvm	PyMC3, greta, JAGS, Stan

Note: The "Species effect > Random" row has an additional value "1" appearing in the PyMC3 column area.

2.2 Knowledge gap

However, these libraries may have some limitations. Indeed, only one of them (Clark et al. 2017), allows the fitting of JSDM from presence only data, only if a model for sampling effort is specified, while occurrence data

that do not record species absences, such as herbarium data or the huge GBIF database, gather thousands of data-sets concerning very numerous species. Furthermore, only two of these libraries (`jSDM` and `gjam`) are able to consider data-sets with missing data, e.g. inventories in which some species were not inventoried on certain sites. Probabilistic programming libraries (`PyMC3`, `greta`, `JAGS`, and `Stan`), allow a wide variety of JSDMs to be specified and fitted from different data types and formats, but it can be challenging for ecologists to develop Bayesian hierarchical models with constraints, latent variables and mixed effects or including spatial variation and species traits. Excluding these libraries, the only R packages capable of fitting JSDMs from long format data are `jSDM` and `gllvm`. In addition, the `jSDM` and `boral` packages are the only ones that can consider presence-absence data with multiple visits at sites, i.e., non-binary responses, following a binomial distribution with more than one trial, however `jSDM` allows specifying a number of trials for each site and `boral` can only consider different number of trials for each species.

— LVM

— monitoring of latent variables convergence and factor loadings in JSDMs.
— spatially explicit models
— taking into account species traits as explanatory variables of the model

Not all of them allow the handling of large data-sets,

— Speed and data-set size (Hmsc, boral), computation time on large data-sets.
— Heterogeneous model performance (Hmsc, boral, gjam (Norberg et al. 2019)). Wilkinson et al. (2019), for a comparative study of fitting JSDMs, using `boral`, `Hmsc`, `BayesComm` and `gjam` (see Table 1),

2.3 Response

— Objectives of the implementation of `jSDM` R package with optimized C++ code (Armadillo and GSL librairies) :

  — Open source, new functionalities easy to implement, evolving package
  — Fill some of this gaps.
  — New functionalities : fourth corner model, long format
  — Relatively fast inference (lower compilation time) The code is optimized and allows the estimation of parameters for large data-sets in a limited amount of time.
  — Easy to use functions –> model comparison and in depth study of the advantages of JSDMs.

— Application on Madagascar forest inventories, current climate conditions
— Application on alpine plants data set including species traits
— Selection of species with constrained factor loadings to improve convergence of latent variables models

3 Major functionalities of the `jSDM` R package

3.1 The structure of the JSDM problem

3.1.1 Data used

The data needed to fit JSDMs are the realizations of a response variable, $Y = (y_{ij})_{j=1,...,J}^{i=1,...,I}$ corresponding to the presence-absence or abundance data of the species, as well as the explanatory variables $X = (X_i)_{i=1,...,I}$, with $X_i = (x_{i0}, x_{i1}, ..., x_{ip}) \in \mathbb{R}^{p+1}$, where p is the number of bioclimatic or environmental variables considered for each site and $\forall i, x_{i0} = 1$. Species traits can also be included in the model as $T = (T_j)_{j=1,...,J}$ with $T_j = (t_{j0}, t_{j1}, ..., t_{jn}) \in \mathbb{R}^n$ where n is the number of species traits considered and $\forall j, t_{j0} = 1$.

These data can be provided in a wide format as above, or in a long format that allows to handle missing data, in the package `jSDM`.

— Figure : Ovaskainen fourth corner without Phylogenie

3.1.2 Model definition

We took inspiration from the articles Warton et al. (2015) and Ovaskainen et al. (2017) to develop the hierarchical approaches used to specify the joint species distribution model in the `jSDM` package.

We could use a multivariate generalized linear model **(GLMM)** of the form :

$$g(\theta_{ij}) = \alpha_i + X_i\beta_j + u_{ij},$$

$$y_{ij} \sim \mathcal{B}inomial(n_i, \theta_{ij}), \text{for presence/absence data}$$

$$\text{or } y_{ij} \sim \mathcal{P}oisson(\theta_{ij}), \text{for abundance data}$$

$$u_i \sim \mathcal{N}_J(0_{\mathbb{R}^J}, \Sigma) \; iid,$$

where

— n_i is the number of visits at site i, and θ_{ij} is the probability of presence of species j at site i for presence/absence data or the average abundance of species j at site i.

— $g : [0, 1[\rightarrow [-\infty, +\infty[$ is a link function (probit, logit or log). The inference method is able to handle only one visit by site with a probit link function so $\forall i, \; n_i = 1$ and $y_{ij} \sim \mathcal{B}ernoulli(\theta_{ij})$ in this case.

— α_i represents the random or fixed effect of site i. In the case of it is random $\alpha_i \sim \mathcal{N}(0, V_\alpha)$ iid and independent from u_i, we also assume that $V_\alpha \sim \mathcal{IG}(\text{shape}, \text{rate})$ as prior distribution.

— $\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})'$ are the intercept and regression coefficients corresponding to the bioclimatic or environmental variables for species j assumed to be fixed effects.

— In the absence of trait-specific data, the species effects : β_j; follow the same *a priori* Gaussian distribution for each species j such as : $\beta_j \sim \mathcal{N}_{p+1}(\mu_\beta, \Sigma_\beta)$.

— If trait-specific data are provided, we consider a fourth corner model to estimate traits-environment interactions, with the effect of species j : β_j; following a Gaussian distribution *a priori* such as : $\beta_j \sim \mathcal{N}_{p+1}(\mu_{\beta_j}, \Sigma_\beta)$, where $\mu_{\beta_{jk}} = \sum_{r=0}^n t_{jr}.\gamma_{rk}$ for $k = 0, \dots, p$, takes different values for each species. In this case we assume that $\gamma_{rk} \sim \mathcal{N}(\mu_{\gamma_{rk}}, V_{\gamma_{rk}})$ as an *a priori* distribution.

— $u_i = (u_{i1}, \dots, u_{iJ})$ is a correlated multivariate random effect whose variance-covariance matrix Σ controls the correlation between species and is assumed to be completely unstructured. This last part of the model is problematic when the number of species J is large, because the number of parameters in Σ increases quadratically with J.

However, in the package `jSDM`, we consider the following latent variable model **(LVM)** :

$$g(\theta_{ij}) = \alpha_i + X_i\beta_j + W_i\lambda_j$$

This LVM is equivalent to a particular case of GLMM with the constraint $\Sigma = \Lambda\Lambda'$ where $\Lambda := (\lambda_{jl})_{j=1,\dots,J}^{l=1,\dots,q}$, by assuming $u_{ij} = W_i\lambda_j$, with $W_i = (W_{i1}, \dots, W_{iq})$ the q latent variables (or "unmeasured predictors") considered, such as $W_i \sim \mathcal{N}(0, I_q)$ iid and independent from α_i, and $\lambda_j = (\lambda_{j1}, \dots, \lambda_{jq})'$ the corresponding factor loadings.

Moreover, to ensure the identifiability of the model, the Λ matrix is constrained to be lower triangular and to take strictly positive values on the diagonal, according to the article Warton et al. (2015). In fact, we use the following prior distribution, for $j = 1, \dots, J$ and $l = 1, \dots, q$:

$$\lambda_{jl} \sim \begin{cases} \mathcal{N}(\mu_\lambda, V_\lambda) & \text{if } l < j \\ \mathcal{N}(\mu_\lambda, V_\lambda) \text{ left truncated by } 0 & \text{if } l = j \\ \mathbb{P}(\lambda_{jl} = 0) = 1 & \text{if } l > j \end{cases}$$

.

We have chosen to consider this type of model, because it facilitates the estimation of the residual correlation matrix between species :

$$R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Indeed, a LVM includes potentially much less parameters to estimate than the corresponding GLMM, because Λ counts as many columns as there are latent variables in the model (q) while Σ contains as much columns as considered species (J).

In following applications we will fix $q = 2$. Indeed, we will consider models with two latent variables, like in the article Warton et al. (2015), by analogy with a principal component analysis (PCA) on the residuals, for which we often use the first two or three axes, because the integration of latent variables to the models is similar to a form of ordination.

Default values of hyper parameters, for each *a priori* distribution, are specified in the package, and can be easily modified by the user.

Figure 1 − *A graphical summary of the Bayesian hierarchical model considered by `jSDM`. In this directed acyclic graph (DAG), the orange boxes refer to the data, the blue circles to the parameters to be estimated, and the arrows to the functional relationships described using statistical distributions*
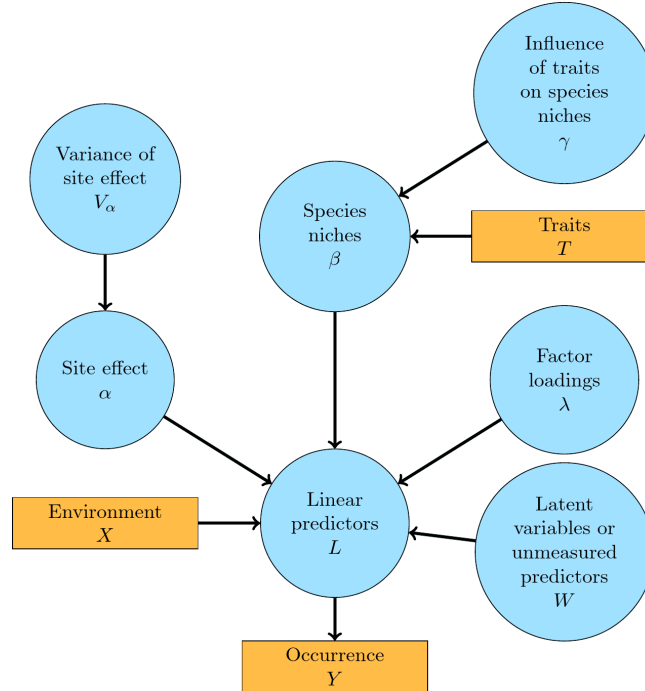
3.2   Bayesian inference methods

The Bayesian inference methods used by the package to estimate the parameters of the JSDMs are summarized in this section. For more information, you will find a detailed presentation with demonstrations in the vignette Bayesian inference methods.

Fitting of hierarchical Bayesian models defined above is performed by a Gibbs sampler written in C++ to maximize computation efficiency. It uses Rcpp to call routines in C++ within R functions, C/C++ libraries (GSL) for random draws and (Armadillo) for matrix calculations. However, the implementation of a Gibbs sampler requires the knowledge of the *a posteriori* distributions of each of the parameters, conditionally to the other parameters of the model.

3.2.1 Gibbs sampler with conjugate priors for probit link function

According to the article Albert and Siddhartha (1993), a possible formulation for models with probit link function, is to assume the existence of an underlying latent variable linked to our observed binary variable such as $z_{ij} = \alpha_i + X_i\beta_j + W_i\lambda_j + \epsilon_{ij} \ \forall i, j$ with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ *iid* and :

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{else.} \end{cases}$$

In this case, parameter's conditional *a posteriori* distributions can be deduced from the conjugate priors formulas.

3.2.2 Adaptive Metropolis-within-Gibbs algorithm for logit or log link functions

In the same way as for the probit model, we can define the models with logit or log link function, using an underlying latent variable, but in this case the *a priori* distributions of the latent variable and the parameters are not conjugated. The properties of conjugate priors cannot therefore be used to deduce an explicit expression for parameter's conditional *a posteriori* distributions. Consequently, the parameters of these models will be sampled according to an estimation of their conditional distributions *a posteriori*, using an adaptive Metropolis algorithm.

3.3 Additionnal outputs

— Deviance
— Diagnostic functions (covariance, correlation, significance and plot functions)

4 Applications

4.1 Benchmarking our method against state-of-the-art JSDM implementation

4.1.1 Model fitting

On the one hand, we used the R package `Hmsc` 3.0-11, for Hierarchical Modelling of Species Communities, which is a model-based approach for analyzing community ecological data ((Ovaskainen et al. 2017)). On the other hand, we used the R package `boral` 2.0 (Hui 2016), for Bayesian Ordination and Regression Analysis, which runs with `JAGS` (Just Another Gibbs Sampler) a simulation program from hierarchical Bayesian models using MCMC methods, implemented in C++. These packages and the package `jSDM` 0.2.1 allow to fit various JSDMs, so we were able to compare the accuracy of results obtained by each of them, their computational performance and sampler efficiency on different data-sets. With each of these three packages, we performed binomial probit regressions from four presence-absence data-sets, used in a recent JSDM benchmark study by Wilkinson et al. (2019), and a simulated one. We also performed a Poisson log-linear regressions from two abundance data-sets (Choler 2005 ; Borcard and Legendre 1994). Model fitting was carried out on the same desktop computer with the specifications : Ubuntu 20.04.4 OS, i5-1145G7 CPU @ 2.60GHz × 8, 15.4GB RAM) to compare computational performance.

4.1.2 Statistical and computational comparison

Table 2 – *Dimensions of the data-sets used (n.sites and n.species), number of observations (n.obs=n.site times n.species), number of covariates, species traits and latent axes considered (n.col.X, n.traits and n.latent), number of parameters to estimate (n.param) by performing n.mcmc iterations, computation time in seconds required to fit the model from each of the data-sets and deviance computed from the parameters estimated with each of the packages as follows : $D = -2\sum_{i=1}^{I}\sum_{j=1}^{J}\log(\mathbb{P}(y_{ij}|\beta_j,\lambda_j,W_i))$. Root-Mean-Square Error (RMSE) is computed for the probabilities of presences θ_{ij} on the simulated data-set. Sensitivity, specificity and true skills statistic ($TSS = sensitivity + specificity - 1$) were computed for each data-set, considering the predicted presence-absences matrix : $\hat{Y}$ such as $\hat{y}_{ij} = 1$ if the corresponding estimated occurrence probability $\hat{\theta}_{ij} \geq \tau_i$ and $\hat{y}_{ij} = 0$ otherwise, where $\tau_i = \min\limits_{j,\ y_{ij}=1}(\hat{\theta}_{ij})$, represents the presence threshold for the site i, defined by the lowest estimated probability value for an occurrence point.*

	Simulated	Mosquitos	Eucalypts	Frogs	Fungi	Aravo	Mites
data type	presence-absence	presence-absence	presence-absence	presence-absence	presence-absence	abundance	abundance
distribution	bernoulli	bernoulli	bernoulli	bernoulli	bernoulli	poisson	poisson
n.sites	300	167	455	104	438	75	70
n.species	100	16	12	9	11	65	30
n.latent	2	2	2	2	2	2	2
n.col.X	3	14	8	4	13	3	12
n.traits	0	0	0	0	0	1	0
n.obs	30000	2672	5460	936	4818	4875	2100
n.param	1099	589	1029	261	1040	480	559
n.mcmc	20000	20000	20000	20000	20000	20000	20000
Computation time (secondes)							
boral	51405	3285	1806	93	2736	218	363
Hmsc	264	62	77	42	87	275	187
jSDM	117	15	28	6	28	131	126
Deviance							
boral	25210	1385	2134	318	1847	5034	7036
Hmsc	25392	1768	2632	376	1836	5653	7059
jSDM	25275	1423	2159	296	1603	5028	6857
RMSE							
boral	0.081						
Hmsc	0.079						
jSDM	0.08						
Sensitivity							
boral	0.8	0.73	0.8	0.61	0.76	0.72	0.84
Hmsc	0.8	0.69	0.71	0.61	0.79	0.66	0.84
jSDM	0.8	0.73	0.84	0.64	0.85	0.73	0.85
Specificity							
boral	0.8	0.93	0.96	0.94	0.96	0.9	0.85
Hmsc	0.8	0.9	0.94	0.95	0.96	0.88	0.84
jSDM	0.8	0.93	0.97	0.96	0.97	0.91	0.85
TSS							
boral	0.6	0.65	0.77	0.55	0.72	0.63	0.69
Hmsc	0.6	0.59	0.65	0.55	0.75	0.55	0.68
jSDM	0.6	0.66	0.81	0.61	0.82	0.63	0.69

4.1.3 Sampler efficiency comparison

Table 3 – *The number of effective values sampled per second, by each package. The effective sample size (ESS), is computed for each parameter, using the function `effectiveSize` from package `coda`. We assume that sampling time (T.sample) corresponds to half of the computation time (see Table 2) because the same number of sampling iterations (n.sample) and burn-in (n.burnin) iterations was performed to fit the models. To reduce the auto-correlation of the MCMCs, only one in n.thin of the sampling iterations was retained and a sample of size : sample.size, is returned for each parameter. To obtain de number of effective values sampled per second, in total or for the species intercept β_0, the species effect β, the factor loadings λ and the latent variables W, the sum of the ESSs computed for all parameters, or the average of ESSs for each kind of parameters, is divided by the sampling time in seconds.*

	Simulated	Mosquitos	Eucalypts	Frogs	Fungi	Aravo	Mites
n.burnin	10000	10000	10000	10000	10000	10000	10000
n.sample	10000	10000	10000	10000	10000	10000	10000
n.thin	10	10	10	10	10	10	10
sample.size	1000	1000	1000	1000	1000	1000	1000
T.sample (secondes)							
boral	25703	1643	903	47	1368	109	182
Hmsc	132	31	39	21	44	138	94
jSDM	58	7	14	3	14	65	63
Species intercept: β_0							
boral	0	1	1	21	1	9	5
Hmsc	1	5	4	7	3	1	2
jSDM	15	128	12	109	45	6	2
Species effects: β							
boral	0	1	1	22	1	9	6
Hmsc	7	25	13	16	10	0	1
jSDM	8	87	31	141	43	6	6
Factor loadings: λ							
boral	0	1	1	22	1	9	6
Hmsc	3	24	13	35	13	2	1
jSDM	5	7	2	39	3	3	4
Latent variables: W							
boral	0	1	1	22	1	9	6
Hmsc	5	30	19	38	14	2	1
jSDM	7	34	22	147	21	2	1
Total							
boral	0	1	1	22	1	9	6
Hmsc	5	27	19	35	14	1	1
jSDM	8	54	23	138	23	4	5

4.2 Fitting JSDM including species traits

Analyzing response of alpine plants (Choler 2005) to snowmelt Date according to their SLA

4.3 Selection of species with constrained factor loadings to improve LVM convergence

As a first step, we use the `jSDM_binomial_probit` function for fitting the JSDM in parallel from a simulated data-set, to obtain four MCMC chains of parameters. - Figure : trace and density lambda diags + Rhat to show convergence issues with arbitrarly chosen species

As a second step, we use the `jSDM_binomial_probit_sp_constrained` function for fitting the JSDM in parallel, to obtain four MCMC chains of parameters. At first, the function fit a JSDM with the constrained species arbitrarily chosen as the first ones in the presence-absence data-set. Then, the function evaluates the convergence of MCMC λ chains using the Gelman-Rubin convergence diagnostic ($\hat{R}$). It identifies the species ($\hat{j}_l$) having the higher $\hat{R}$ for $\lambda_{\hat{j}_l}$. These species drive the structure of the latent axis l. The λ corresponding to this species are constrained to be positive and placed on the diagonal of the Λ matrix for fitting a second model. This usually improves the convergence of the latent variables and factor loadings.

— Figure : trace and density lambda_jhat + Rhat, legend : The function returns the parameter posterior distributions for this second model. We represent the trace and density *a posteriori* of $\lambda_{\hat{j}_l}$ for each latent axis l.

4.4 Inference of biodiversity maps in Madagascar from large forest inventories

4.4.1 Datasets

4.4.1.1 Forest inventory National forest inventories carried out on 753 sites on the island of Madagascar are available, listing the presence or absence of 555 plant species on each of these sites between 1994 and 1996. We use these forest inventories to calculate a matrix indicating the presence by a 1 and the absence by a 0 of the species at each site by removing observations for which the species is not identified. This matrix therefore records the occurrences of 483 (n_{sp}) species at 751 (n_{site}) sites.

4.4.1.2 Current environmental variables Among the climatic and environmental data available on the site https://madaclim.cirad.fr, concerning the whole island of Madagascar at present (interpolations of representative observed data from the years 1960-1990), and with a resolution of 1km, we chose to use the following variables, because they have an ecological meaning which makes them easily interpretable and are little correlated between them, according to the article Vieilledent et al. (2013) :

Table 4 − *Climatic variables*

variable	description
temp	the average annual temperature ($°C \times 10$)
prec	the average annual precipitation (mm).
sais.temp	the seasonality of temperatures corresponds to the standard deviation of monthly temperatures multiplied by 100.
sais.prec	the seasonality of precipitation as a coefficient of variation.
cwd	the annual climatic water deficit (mm) is based on monthly precipitation (*prec*) and potential evapotranspiration (*pet*) which is defined as the amount of evaporation that would occur in a month if a sufficient water source were available : $\mathrm{cwd} = \sum_{m=1}^{12} \min(0, \mathrm{prec}_m - \mathrm{pet}_m)$.

We consider also the quadratic effects of these climate variables to perform a quadratic regression, which is more suitable for fitting a niche model than a linear regression. The values of these climate variables, on the coordinates of the inventory plots, are extracted, scaled and centered. The resulting data-set is used as explanatory variables for fitting the JSDM.

4.4.2 Fitting joint species distribution model (JSDM)

4.4.2.1 Model definition We use the latent variable model (LVM) defined previously to account for species co-occurrence on all sites :

$$\mathrm{probit}(\theta_{ij}) = \alpha_i + \beta_{0j} + X_i.\beta_j + W_i.\lambda_j$$

It can be easily shown that the response variable $y_{ij} \sim \mathcal{Bernoulli}(\theta_{ij})$ by defining the latent variable : $z_{ij} = \alpha_i + \beta_{0j} + X_i.\beta_j + W_i.\lambda_j + \epsilon_{i,j}$, with $\forall (i,j)\ \epsilon_{ij} \sim \mathcal{N}(0,1)$, such that :

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

— Latent variables : $W_i = (W_i^1, \dots, W_i^q)$ where q is the number of latent variables considered, which has to be fixed by the user (in this case $q = 2$). We assume that $W_i \sim \mathcal{N}(0, I_q)$ and we define the associated coefficients $\lambda_j = (\lambda_j^1, \dots, \lambda_j^q)'$, also known as "factor loadings" (Warton et al. 2015). We use a prior distribution $\mathcal{N}(0, 1)$ for all lambda not concerned by constraints to 0 on upper diagonal and to strictly positive values on diagonal.

— Explanatory variables : bioclimatic data about each site and their quadratic effects. $X = (X_i)_{i=1,\dots,n_{site}}$ with $X_i = (\text{temp}_i, \text{prec}_i, \text{sais.temp}_i, \text{sais.prec}_i, \text{cwd}_i, \text{temp}_i^2, \text{prec}_i^2, \text{sais.temp}_i^2, \text{sais.prec}_i^2, \text{cwd}_i^2) \in \mathbb{R}^p$ where $p = 10$ is the number of variables considered. The corresponding regression coefficients for each species j are noted : $\beta_j = (\beta_j^1, \dots, \beta_j^p)'$. We use a prior distribution $\mathcal{N}(0, 1)$ for all beta.

— β_{0j} corresponds to the intercept for species j which is assumed to be a fixed effect. We use a prior distribution $\mathcal{N}(0, 10)$ and define the prevalence of species j ($p_j = \sum_{i=1}^{n_{site}} \frac{y_{ij}}{n_{site}}$) as starting values for the species intercept to improve the estimation of this parameter.

— α_i represents the random effect of site i such as $\alpha_i \sim \mathcal{N}(0, V_\alpha)$ and we assume that $V_\alpha \sim \mathcal{IG}(\text{shape} = 0.1, \text{rate} = 0.1)$ as prior distribution by default.

4.4.2.2 Parameters inference We fit a binomial joint species distribution model, including two latent variables and a random site effect using the `jSDM_binomial_probit()` function to perform binomial probit regression considering all the species from the data described above, by performing 80000 iterations including 40000 of burn-in and we retain $N_{samp} = 1000$ values for each parameter of the model.

Table 5 – *Compuation time for fitting JSDM from Madagascar data-sets and deviance obtained*

Number of species	Number of sites	Number of observations	Number of estimated parameters	Iterations performed	Computation time (hours)	Deviance
483	751	362733	8532	80000	1.8	32916.8

4.4.2.3 Evaluation of MCMC convergence multiple chains Rhat ?

4.4.3 Predictive performance of the model

compute AUC, Rprecision on inventory sites to assess the predictive performance of the model ?
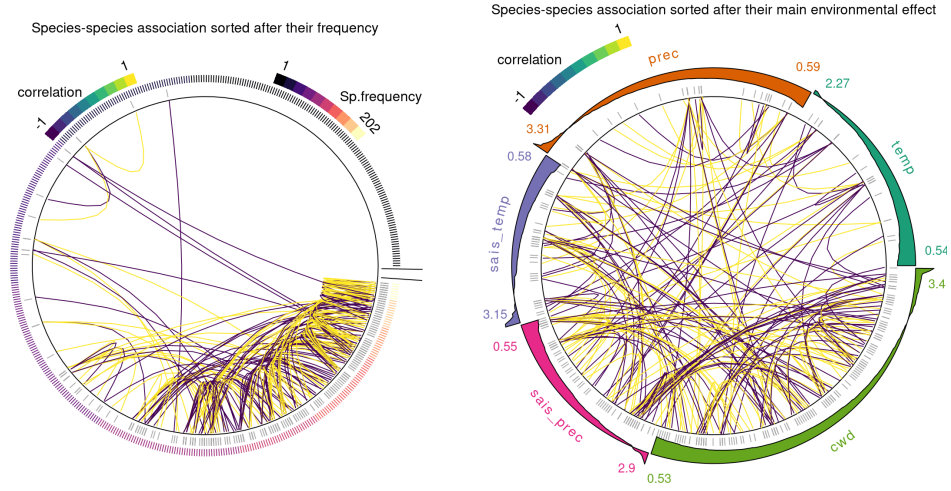
4.4.4 Estimated species-species associations

After fitting the JSDM with latent variables, the **full species residual correlation matrix** $R = (R_{ij})_{j=1,\dots,n_{species}}^{i=1,\dots,n_{site}}$ can be derived from the covariance in the latent variables such as :

$$\Sigma_{ij} = \lambda_i . \lambda_j^T$$

, then we compute correlations from covariances :

$$R_{i,j} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$$

.

(a) *The top 100 residual correlations between species, with the 483 species sorted according to the number of sites where they occur out of the 751 inventoried.*

(b) *The top 100 residual correlations between species, with the species sorted according to their largest environmental coefficients (β) (the outer ring shows the distribution of environmental effect for species within the sample.)*

Figure 2 – *Representation of the top 100 estimated residual correlations between species, based on their number of observed occurrences and environmental preferences.*

The figure 2, inspired by the article Pichler and Hartig (2021), allows us to observe the positive or negative residual correlations between species which can be interpreted in terms of positive or negative influence of the presence of a species on the probability of occurrence of another.

— main correlation more between abundant species
— correlation between species whith main env effect cwd and sais_prec

4.4.5 Interpolation of parameters between inventory sites

The presence of a spatial structure where observations close to each other are more similar than those far away (spatial auto-correlation) is a prerequisite for the application of geostatistics and according to the spatial representations in supplementary, it seems to be satisfied.

Thus, we can to interpolate sites' parameters (α and W) for the whole island from those estimated for the inventory plots, using the spatial interpolation method called regularized spline with tension (**RST**) from `GRASS GIS` software via `rgrass7` R package. This method is described in the article (Mitášová and Hofierka 1993). We then calculate species occurrence probabilities between inventory sites using interpolated site parameters.
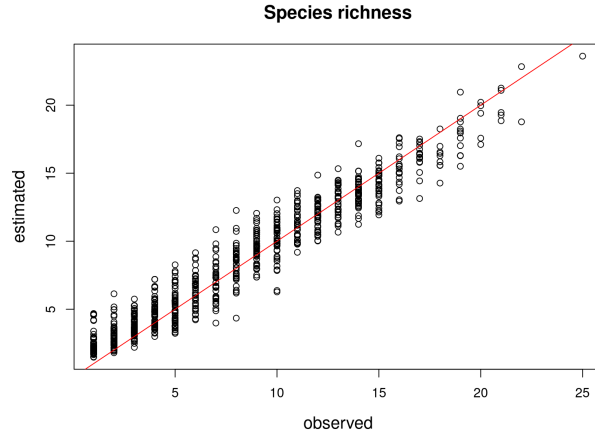
4.4.6 Estimated species richness

Species richness, also known as α-diversity, reflects the number of species coexisting in a given environment and is calculated by summing the number of species present at each site.
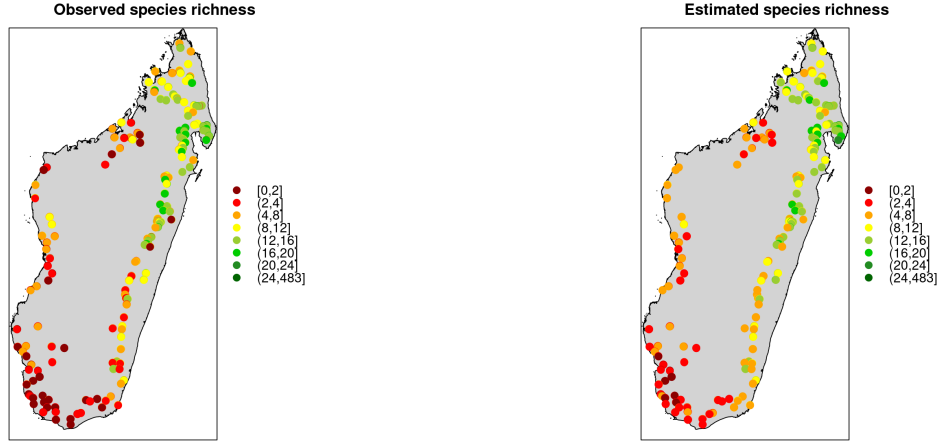
(a) *Spatial representation of the species richness observed at each inventory site i defined by $R_i = \sum\limits_{j=1}^{n_{sp}} y_{ij}$.*

(b) *Spatial representation of the species richness estimated, following Scherrer, Mod, and Guisan (2020), by summing the estimated occurrence probabilities of all species on each inventory site : $\widehat{R}_i = \sum\limits_{j=1}^{n_{sp}} \widehat{\theta}_{ij}$.*

(c) *Estimated species richness according to the observed one.*

Figure 3 − *Representation of species richness at inventory sites.*

We can see on figures 3.a and 3.b that the observed species richness is greater in the moist forest of the north-east of the island, particularly in the Masoala peninsula. Moreover, it can be seen on figure 3.c that the observed species richness on the inventory sites corresponds fairly well to that estimated.
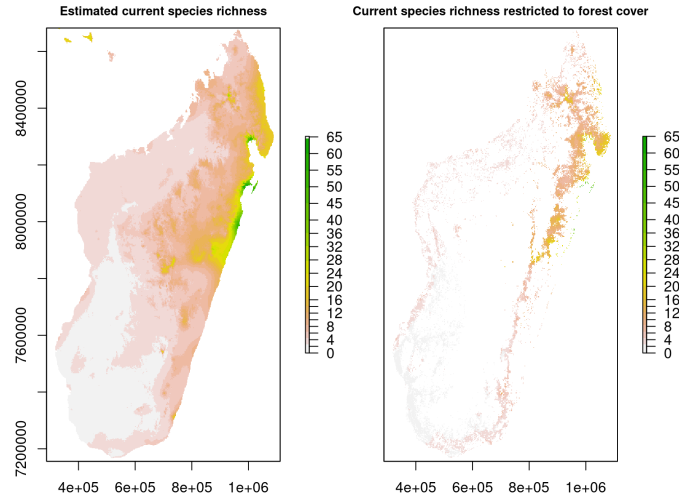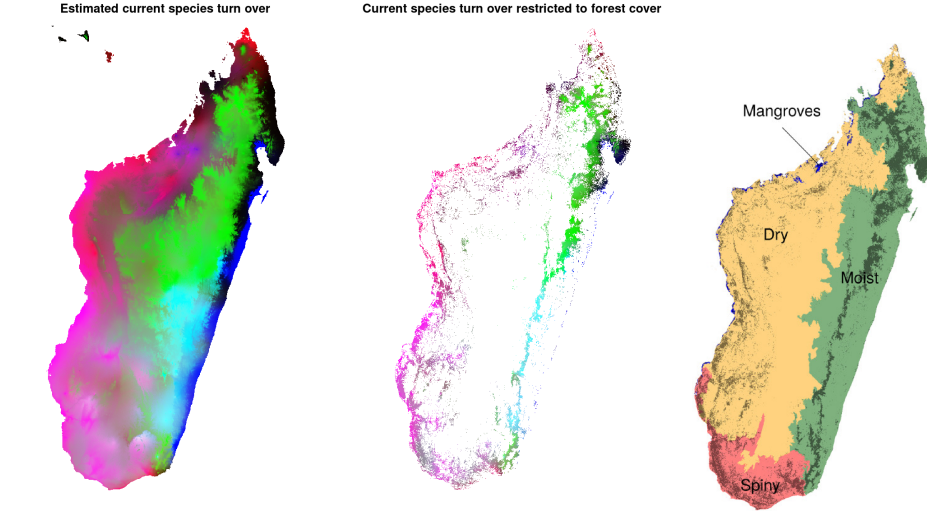
Figure 4 − *Estimated species richness at the Madagascar scale by summing interpolated occurrence probabilities (left) and estimated species richness in Madagascar restricted to remaining forest cover in 2000 from the article Vieilledent et al. (2018), to account for deforestation on the island (right).*

We can see on figure 4 that the species richness, computed from interpolated occurrence probabilities, is greater in the moist forest of the north-east of the island, particularly in the Masoala peninsula as expected given the observed species richness on inventory sites. It can be explained by the fact that this area is located close to the Equator and is receiving the highest level of precipitation in Madagascar. As a consequence, this area benefits from high levels of resources (light and water) and is characterized by a low environmental stress and a long growing season (Vieilledent et al. 2013). Note that the estimated species richness does not necessarily stand for a 'true' number of present species but represent a relative species richness between sites (one site is richer in species than another site).

4.4.7 Estimated species turnover at Madagascar scale

Species turnover, also known as β-diversity, is a measure of biodiversity that compares the diversity of species between ecosystems or along environmental gradients, using the number of taxa that are unique to each ecosystem.

(a) *Estimated species turnover at the Madagascar scale (left) and estimated species turnover in Madagascar restricted to remaining forest cover in 2000 from the article Vieilledent et al. (2018), to account for deforestation on the island (right). To obtain these maps, we perform a standard PCA on the species occurrence probabilities, interpolated for each pixel, referring to the article Allnutt et al. (2008). The coordinates obtained for the first three PCA's axes, that reflect the composition of the species community likely to occupy the corresponding pixel, are scaled $ [0,255] $ to be represented by red, green and blue color levels. The combination of these three color levels determines the coloring of each pixel in the displayed $ $-diversity map. Therefore, a color difference between two pixels indicates that the species present are not the same, while pixels of identical color host communities of similar species.*

(b) *Forest types in Madagascar from the article Vieilledent et al. (2018).*

Figure 5 − *Estimated species turnover and forest types in Madagascar.*

We can see that the estimated map of β-diversity on the figure @ref(fig : diversity-beta).a agrees with the map of Madagascar's forest types from the article Vieilledent et al. (2018) on figure @ref(fig : diversity-beta).b. Indeed, communities of species represented by shades of blue and green correspond to the eastern moist forest, communities represented by shades of red match the western dry forest and communities in magenta correspond to the southern spiny forest. Moreover, the β-diversity map we obtain allows to identify substantially different species communities within these three forest types. For example, we can clearly identify three different tree communities (in blue, green, and black) within the eastern moist forest.

5 Discussion

5.1 Other software comparison

Computational performance :

Inferential performance :

jSDM advantages : - lower compilation time than most other softwares - Open source - new features easy to implement (evolving package) - complete documentation

5.2 Limitation of jSDM and perspective of improvement

— very large data set still too long fitting, sjSDM useful in this case
— phylogeny
— Spatial autocorrelation for fitting spatially explicit models

— Random species effect and factor loadings
— Negative-binomial distribution for abundance data
— presence only data

5.3 Implications and outlook for ecological data analysis

— to predict species distributions area
— to estimate species diversity (α), species turnover (β) or other diversity indices
— to predict the evolution of biodiversity under climate change
— to identify refuge zone of biodiversity
— to explain difference of vulnerability to climate change between species in relation to their functional traits

6 Installation

A stable version of the package (jSDM version 0.2.1 at the time of writing) can be downloaded from Comprehensive R Archive Network (CRAN, https://cran.r-project.org/package=jSDM). This version run with the current version of R (R 3.5.0 at the time of writing) and will provide the features introduced in this publication. At the same time, a developmental version is available on Github (https://github.com/ghislainv/jSDM), where it will be tested before being released on CRAN. The current stable Github version can be downloaded manually or via `install_github` from the `devtools` suite of functions (Wickham et al. 2021) :

```
devtools::install_github("ghislainv/jSDM")
```

Authors' contribution

Frederic Gosselin has developed the method of selecting the species with constrained factor loadings to improve latent variable models convergence.

Conflict of interest

The authors declare that there are no conflict of interest.

Acknowledgments

Data Accessibility

Supplementary information

References

Albert, James H., and Chib Siddhartha. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88 (422) : 669–79. https://doi.org/10.1080/0162 1459.1993.10476321.

Allnutt, Thomas F., Simon Ferrier, Glenn Manion, George V. N. Powell, Taylor H. Ricketts, Brian L. Fisher, Grady J. Harper, et al. 2008. "A Method for Quantifying Biodiversity Loss and Its Application to a 50-Year Record of Deforestation Across Madagascar." *Conservation Letters* 1 (4) : 173–81. https://doi. org/10.1111/j.1755-263X.2008.00027.x.

Borcard, Daniel, and Pierre Legendre. 1994. "Environmental Control and Spatial Structure in Ecological Communities : An Example Using Oribatid Mites (Acari, Oribatei)." *Environmental and Ecological Statistics* 1 (1) : 37–61. https://doi.org/10.1007/BF00714196.

Choler, Philippe. 2005. "Consistent Shifts in Alpine Plant Traits Along a Mesotopographical Gradient." *Arctic, Antarctic, and Alpine Research* 37 (4) : 444–53. https://doi.org/10.1657/1523-0430(2005)037%5B0444:csiapt%5D2.0.co;2.

Clark, James S., Diana Nemergut, Bijan Seyednasrollah, Phillip Turner, and Stacy Zhang. 2017. "Generalized Joint Attribute Modeling for Biodiversity Analysis : Median-Zero, Multivariate, Multifarious Data." *Ecological Monographs.* http://sites.nicholas.duke.edu/clarklab/code/.

Elith, Jane, and John Leathwick. 2009. "Species Distribution Models : Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution and Systematics* 40 (December) : 677–97. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Golding, Nick. 2019. "greta : Simple and Scalable Statistical Modelling in r." *Journal of Open Source Software* 4 (40) : 1601. https://doi.org/10.21105/joss.01601.

Golding, Nick, Miles A. Nunn, and Bethan V. Purse. 2015. "Identifying Biotic Interactions Which Drive the Spatial Distribution of a Mosquito Community." *Parasites & Vectors* 8 (1) : 367. https://doi.org/10.1186/s13071-015-0915-1.

Guisan, Antoine, and Carsten Rahbek. 2011. "SESAM – a New Framework Integrating Macroecological and Species Distribution Models for Predicting Spatio-Temporal Patterns of Species Assemblages." *Journal of Biogeography* 38 (8) : 1433–44. https://doi.org/10.1111/j.1365-2699.2011.02550.x.

Hui, Francis K. C. 2016. "Boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r." *Methods in Ecology and Evolution* 7 (6) : 744–50. https://doi.org/10.1111/2041-210X.12514.

Hui, Francis K. C., David I. Warton, Scott D. Foster, and Piers K. Dunstan. 2013. "To Mix or Not to Mix : Comparing the Predictive Performance of Mixture Models Vs. Separate Species Distribution Models." *Ecology* 94 (9) : 1913–19. https://doi.org/10.1890/12-1322.1.

Mitášová, Helena, and Jaroslav Hofierka. 1993. "Interpolation by Regularized Spline with Tension : II. Application to Terrain Modeling and Surface Geometry Analysis." *Mathematical Geology* 25 (6) : 657–69. https://doi.org/10.1007/BF00893172.

Niku, Jenni, Francis K. C. Hui, Sara Taskinen, and David I. Warton. 2019. "Gllvm - Fast Analysis of Multivariate Abundance Data with Generalized Linear Latent Variable Models in r." *Methods in Ecology and Evolution* 10 : 2173–82.

Norberg, Anna, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, et al. 2019. "A Comprehensive Evaluation of Predictive Performance of 33 Species Distribution Models at Species and Community Levels." *Ecological Monographs* 89 (3) : e01370. https://doi.org/10.1002/ecm.1370.

Ovaskainen, Otso, and Janne Soininen. 2011. "Making More Out of Sparse Data : Hierarchical Modeling of Species Communities." *Ecology* 92 (February) : 289–95. https://doi.org/10.2307/41151138.

Ovaskainen, Otso, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. 2017. "How to Make More Out of Community Data ? A Conceptual Framework and Its Implementation as Models and Software." *Ecology Letters* 20 (5) : 561–76. https://doi.org/10.1111/ele.12757.

Pichler, Maximilian, and Florian Hartig. 2021. "A New Joint Species Distribution Model for Faster and More Accurate Inference of Species Associations from Big Community Data." *Methods in Ecology and Evolution* 12 (11) : 2159–73. https://doi.org/10.1111/2041-210X.13687.

Plummer, Martyn. 2003. "JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling."

Salvatier, John, Thomas V Wiecki, and Christopher Fonnesbeck. 2016. "Probabilistic Programming in Python Using PyMC3." *PeerJ Computer Science* 2 : e55.

Scherrer, Daniel, Heidi K. Mod, and Antoine Guisan. 2020. "How to Evaluate Community Predictions Without Thresholding ?" *Methods in Ecology and Evolution* 11 (1) : 51–63. https://doi.org/10.1111/2041-210X.13312.

Schmitt, Sylvain, Robin Pouteau, Dimitri Justeau, Florian de Boissieu, and Philippe Birnbaum. 2017. "Ssdm : An r Package to Predict Distribution of Species Richness and Composition Based on Stacked Species Distribution Models." *Methods in Ecology and Evolution* 8 (12) : 1795–1803. https://doi.org/10.1111/2041-210X.12841.

Segurado, Pedro, and Miguel B. Araújo. 2004. "An Evaluation of Methods for Modelling Species Distributions." *Journal of Biogeography* 31 (10) : 1555–68. https://doi.org/10.1111/j.1365-2699.2004.01076.x.

Stan Development Team. 2021. "Stan Modeling Language Users Guide and Reference Manual, Version 2.29." https://mc-stan.org.

Tikhonov, Gleb, Nerea Abrego, David Dunson, and Otso Ovaskainen. 2017. "Using Joint Species Distribution Models for Evaluating How Species-to-Species Associations Depend on the Environmental Context." *Methods in Ecology and Evolution* 8 (4) : 443–52. https://doi.org/10.1111/2041-210X.12723.

Vellend, Mark. 2010. "Conceptual Synthesis in Community Ecology." *The Quarterly Review of Biology* 85 (2) : 183–206. https://doi.org/10.1086/652373.

Vieilledent, Ghislain, Cyrille Cornu, Aida Cuní Sanchez, Jean-Michel Leong Pock-Tsy, and Pascal Danthu. 2013. "Vulnerability of Baobab Species to Climate Change and Effectiveness of the Protected Area Network in Madagascar : Towards New Conservation Priorities." *Biological Conservation* 166 (October) : 11–22. https://doi.org/10.1016/j.biocon.2013.06.007.

Vieilledent, Ghislain, Clovis Grinand, Fety A. Rakotomalala, Rija Ranaivosoa, Jean-Roger Rakotoarijaona, Thomas F. Allnutt, and Frédéric Achard. 2018. "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 (June) : 189–97. https://doi.org/10.1016/j.biocon.2018.04.008.

Warton, David I., F. Guillaume Blanchet, Robert B. O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C. Walker, and Francis K. C. Hui. 2015. "So Many Variables : Joint Modeling in Community Ecology." *Trends in Ecology & Evolution* 30 (12) : 766–79. https://doi.org/10.1016/j.tree.2015.09.007.

Wickham, Hadley, Jim Hester, Winston Chang, Jennifer Bryan, and RStudio. 2021. "Devtools : Tools to Make Developing R Packages Easier." https://CRAN.R-project.org/package=devtools.

Wilkinson, David P., Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A. McCarthy. 2019. "A Comparison of Joint Species Distribution Models for Presence-Absence Data." *Methods in Ecology and Evolution* 10 (2) : 198–211. https://doi.org/10.1111/2041-210X.13106.

Wisz, Mary Susanne, Julien Pottier, W. Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F. Damgaard, Carsten F. Dormann, et al. 2013. "The Role of Biotic Interactions in Shaping Distributions and Realised Assemblages of Species : Implications for Species Distribution Modelling." *Biological Reviews* 88 (1) : 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x.