

Rapport détaillé
Comité de suivi de thèse : 1ère année

Jeanne Clément

07 avril, 2023

Sommaire

1 Contexte scientifique et objectifs de la thèse	1
2 Fonctionnalités et contenu actuels du package jSDM	2
2.1 Définition des modèles joints de distribution des espèces envisagés	2
2.1.1 Modèle linéaire mixte généralisé multivarié (GLMM)	2
2.1.2 Modèle à variable latente (LVM)	3
2.2 Méthodes d'inférence bayésienne selon la fonction de lien choisie	4
2.2.1 Principe d'un échantillonneur de Gibbs	4
2.2.2 Priors utilisés	5
2.2.3 Modèle probit : échantillonneur de Gibbs et priors conjugués	5
2.2.4 Modèles logit et log : échantillonneur de Gibbs et algorithme de Metropolis adaptatif .	6
2.3 Comparaison des résultats obtenus avec ceux des packages boral et Hmsc	8
2.3.1 Description des jeux de données utilisés	8
2.3.2 Comparaison de la pertinence des résultats obtenus et des temps de calcul nécessaires avec chacun des packages	9
3 Obtention de cartes de communauté à l'échelle du territoire à partir de données d'inventaire forestier	10
3.1 Évolution de la biodiversité à Madagascar sous l'effet des changements climatiques	10
3.1.1 Description des données	10
3.1.2 Ajustement d'un JSDM à partir de ces données	11
3.1.3 Corrélation résiduelle entre les espèces estimée	15
3.1.4 Estimation de la richesse spécifique pour les placettes d'inventaire et comparaison à celle observée	15
3.1.5 Interpolation spatiale des paramètres associés aux sites d'inventaire	16
3.1.6 Evolution de la richesse spécifique à Madagascar	18
3.1.7 Evolution de la diversité β à Madagascar	18
3.2 Obtention de cartes de communauté en Guyane française	20
3.2.1 Données utilisées	20
3.2.2 Enjeux	20
4 Fonctionnalités en cours de développement du package jSDM	20
4.1 Estimation du nombre d'axes latents à prendre en compte et amélioration de leur convergence	20
4.2 Intégrer la phylogénie des espèces comme variable explicative des modèles	22
4.3 Ajuster des JSDMs spatialement explicites	23
4.4 Ajuster des JSDMs à partir de données de présence seule	23
Bibliographie	24

1 Contexte scientifique et objectifs de la thèse

Les changements climatiques risquent d'impacter fortement les forêts tropicales par des changements d'aire de distribution des espèces et de composition des communautés (Bunker et al. 2005 ; Vieilledent et al. 2016). Les modèles de distribution d'espèces (SDMs) sont couramment utilisés en écologie afin de prédire la niche écologique d'une espèce et sa vulnérabilité aux changements climatiques (Elith and Leathwick 2009). Les principales limitations des SDMs sont qu'ils ne prennent pas en compte les interactions entre espèces et qu'ils ne considèrent très souvent qu'un filtrage environnemental pour prédire l'occurrence des espèces (abondance ou probabilité de présence). Ce sont des modèles corrélatifs qui ne permettent pas toujours d'expliquer les différences de vulnérabilité entre espèces (via les traits fonctionnels par exemple).

Les modèles joints de distribution des espèces (JSDMs), qui sont apparus récemment en écologie (Warton et al. 2015), permettent de prendre en compte les interactions entre espèces pour prédire leur occurrence. Cette approche est particulièrement intéressante pour les espèces rares (nombreuses en forêt tropicale) qui peuvent ainsi emprunter de l'information aux autres espèces plus abondantes. De plus, ces modèles fournissent un cadre conceptuel permettant d'intégrer la phylogénie ou les traits fonctionnels pour expliquer les différences d'occurrence entre espèces (Warton et al. 2015 ; Ovaskainen et al. 2017) afin d'être en mesure d'interpréter les différences de vulnérabilité des espèces face au changement climatique en fonction de leurs traits spécifiques ou corrélations phylogénétiques. Ceci permettrait d'identifier des traits fonctionnels significatifs pour expliquer les caractéristiques de résistance à la sécheresse de certaines espèces par exemple. Cette approche d'écologie fonctionnelle peut également s'avérer particulièrement utile pour les espèces rares qui représentent une grande majorité des espèces en forêt tropicale et pour lesquelles on dispose de peu de données d'occurrence afin d'ajuster les modèles mais dont les traits spécifiques peuvent être mesurés même à partir de peu d'individus.

Les JSDMs ont connu une expansion rapide ces dernières années avec le développement de plusieurs librairies permettant d'ajuster ce type de modèles suivant différentes approches statistiques comme les packages R **Hmsc** (Ovaskainen et al. 2017), **gjam** (Clark et al. 2017), **BayesComm** (Golding, Nunn, and Purse 2015), **bora1** (Warton et al. 2015) ou **s-jSDM** (Pichler and Hartig 2020). Cependant, ces librairies peuvent présenter certaines limitations. Elles ne permettent pas toutes (i) le traitement de jeux de données conséquents en un temps raisonnable (ii) l'extrapolation entre les sites d'observation pour l'obtention de cartes prédictives, (iii) la gestion de données de présences seules (typique des données d'herbier par exemple) ou de données manquantes.

Mon projet de thèse s'appuie donc sur le développement du package **jSDM** (Vieilledent, Clément, and CIRAD 2019) afin de lever les limitations propres aux librairies et modèles de distribution d'espèces actuels. Pour ensuite utiliser ce package afin d'ajuster des modèles joints de distribution des espèces d'arbres pour la Guyane en combinant les données d'inventaires forestiers disponibles, les données sur les traits fonctionnels, la phylogénie et les données environnementales, dans l'objectif d'obtenir des cartes de communautés dans le présent et dans le futur sous l'effet des changements climatiques.

2 Fonctionnalités et contenu actuels du package jSDM

Le package jSDM comme les autres librairies mentionnées précédemment, permet l'ajustement de modèles joints de distribution des espèces prenant en compte les interactions entre espèces (<https://ecology.ghisla.inu.fr/jSDM>). J'ai développé ce package en grande partie, il fait appel à des routines en C++ qui utilisent des librairies C/C++ dédiées aux tirages aléatoires (GSL) et aux calculs matriciels (Armadillo). Le code est optimisé et permet l'estimation de paramètres pour de gros jeux de données en un temps limité. Pour chaque fonction du package j'ai rédigée une documentation détaillée accompagnée d'exemples et de vignettes afin de faciliter leur utilisation.

2.1 Définition des modèles joints de distribution des espèces envisagés

On s'est inspiré des articles Warton et al. (2015) et Ovaskainen et al. (2017) pour développer les approches hiérarchiques utilisées à la spécification des modèle joint de distribution des espèces dans le package jSDM.

Les données dont on dispose pour ajuster ce type de modèle sont les réalisations d'une variable réponse, $Y = (y_{ij})_{j=1,\dots,J}^{i=1,\dots,I}$ correspondant à des données de présence/absence ou d'abondance des espèces, ainsi que les variables explicatives $X = (X_i)_{i=1,\dots,I}$ avec $X_i = (x_{i0}, x_{i1}, \dots, x_{ip}) \in \mathbb{R}^{p+1}$ où p est le nombre de variables bioclimatiques considérées pour chaque site et $\forall i, x_{i0} = 1$.

On peut également prendre en compte les caractéristiques des espèces : $T = (T_j)_{j=1,\dots,J}$ avec $T_j = (t_{j0}, t_{j1}, \dots, t_{jn}) \in \mathbb{R}^n$ où n est le nombre de traits spécifiques considérés et $\forall j, t_{j0} = 1$.

2.1.1 Modèle linéaire mixte généralisé multivarié (GLMM)

D'une part on pourrait utiliser un modèle linéaire mixte généralisé multivarié (**GLMM**) de la forme :

$$\begin{aligned} g(\theta_{ij}) &= \alpha_i + X_i \beta_j + u_{ij}, \\ y_{ij} &\sim \text{Binomial}(n_i, \theta_{ij}), \text{ pour des données de présence/absence} \\ &\text{ou } y_{ij} \sim \text{Poisson}(\theta_{ij}), \text{ pour des données d'abondances} \\ u_i &\sim \mathcal{N}_J(0_{\mathbb{R}^J}, \Sigma) \text{ iid,} \\ \alpha_i &\sim \mathcal{N}(0, V_\alpha) \text{ iid et indépendant de } u_i. \end{aligned}$$

où

- n_i correspond au nombre de visites du site i et θ_{ij} à la probabilité de présence de l'espèce j sur le site i pour les données de présence/absence ou à l'abondance moyenne de l'espèce j sur le site i .
 - $g :]0, 1[\rightarrow]-\infty, +\infty[$ est une fonction de lien (probit, logit ou log).
 - α_i représente l'effet aléatoire ou fixe du site i ,
 - $\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})'$ sont l'intercept et les coefficients de régression correspondants aux variables bioclimatiques pour l'espèce j supposés être des effets fixes.
 - En l'absence de données sur les traits spécifiques, les effets espèces : β_j ; suivent la même distribution gaussienne *a priori* pour chaque espèce j telle que : $\beta_j \sim \mathcal{N}_{p+1}(\mu_\beta, V_\beta)$.
 - Si des données sur les traits spécifiques sont fournies, l'effet de l'espèce j : β_j ; suit une distribution gaussienne *a priori* telle que : $\beta_j \sim \mathcal{N}_{p+1}(\mu_{\beta_j}, V_\beta)$, où $\mu_{\beta_{jk}} = \sum_{r=0}^n t_{jr} \gamma_{rk}$ pour $k = 0, \dots, p$, prend différentes valeurs pour chaque espèce. Dans ce cas on suppose que $\gamma_{rk} \sim \mathcal{N}(\mu_{\gamma_{rk}}, V_{\gamma_{rk}})$ en tant que distribution *a priori*
 - $u_i = (u_{i1}, \dots, u_{iJ})$ est un effet aléatoire multivarié corrélé dont la matrice de variance covariance Σ contrôle la corrélation entre les espèces et est supposée être complètement non structurée.
- Cette dernière partie du modèle est problématique lorsque le nombre d'espèces J est important car le nombre de paramètres dans Σ augmente quadratiquement avec J .

2.1.2 Modèle à variable latente (LVM)

D'autre part en posant $u_{ij} = W_i \lambda_j$, avec $W_i = (W_{i1}, \dots, W_{iq})$ les q variables latentes (ou “prédicteurs non mesurés”) considérés et $\lambda_j = (\lambda_{j1}, \dots, \lambda_{jq})'$ les coefficients associés, on obtient le modèle à variables latentes (**LVM**) suivant :

$$\begin{aligned} g(\theta_{ij}) &= \alpha_i + X_i \beta_j + W_i \lambda_j \\ W_i &\sim \mathcal{N}(0, I_q) \text{ iid} \\ \alpha_i &\sim \mathcal{N}(0, V_\alpha) \text{ iid et indépendant de } W_i \end{aligned}$$

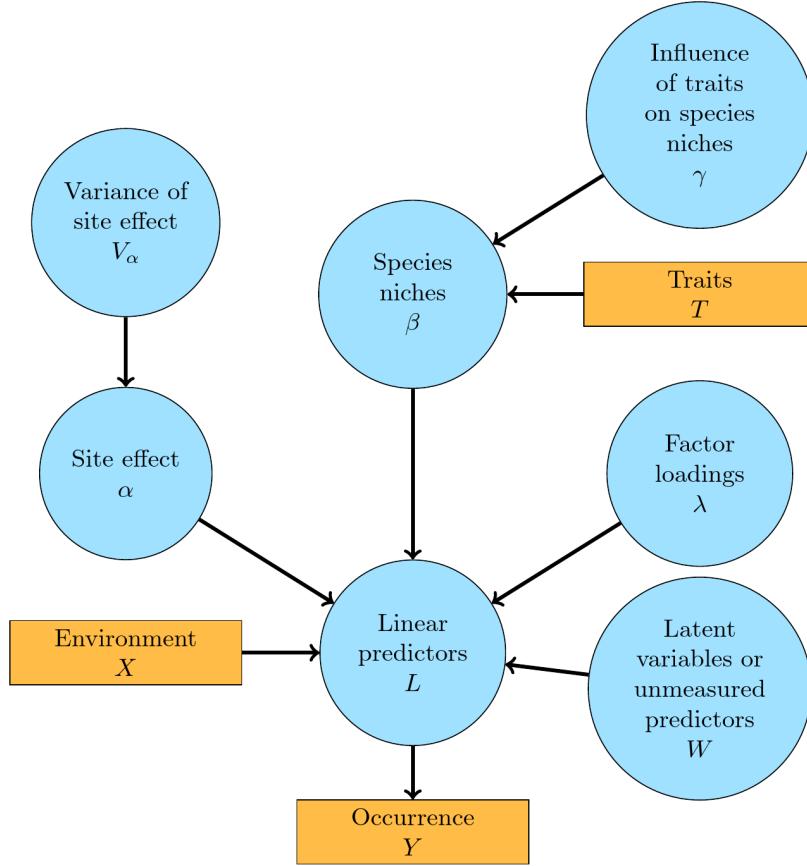
Ce qui revient à un cas particulier de GLMM auquel on impose la contrainte $\Sigma = \Lambda \Lambda'$ avec $\Lambda := (\lambda_{jl})_{j=1, \dots, J}^{l=1, \dots, q}$.

On préférera ce dernier modèle pour estimer la matrice de corrélation entre les espèces, en effet il comporte potentiellement beaucoup moins de paramètres que le GLMM précédent car Λ a autant de colonne qu'il y a de variables latentes (q) tandis que Σ présente autant de colonnes de paramètres qu'il y a d'espèces (J).

De plus, pour assurer l'identifiabilité du modèle les valeurs de Λ sont contraintes à des valeurs strictement positives sur la diagonale et nulles au dessus de celle-ci d'après l'article Warton et al. (2015), Λ est ainsi supposée être triangulaire inférieure.

Dans la suite on fixera $q = 2$, en effet comme Warton et al. (2015) on considérera des modèles à deux variables latentes par analogie avec une analyse par composante principale (ACP) sur les résidus pour lesquels on utilise souvent les deux ou trois premiers axes car l'intégration de variables latentes aux modèles s'apparente à une forme d'ordination.

Figure 1 – Un résumé graphique du modèle hiérarchique bayésien. Dans ce graphe acyclique dirigé (DAG), les cases orange font référence aux données, les cercles bleus aux paramètres à estimer, et les flèches aux relations fonctionnelles décrites à l'aide de distributions statistiques



2.2 Méthodes d'inférence bayésienne selon la fonction de lien choisie

Les méthodes d'inférence bayésiennes utilisées par le package pour estimer les paramètres des JSDMs sont résumées dans cette partie pour en savoir plus vous trouverez une présentation détaillée accompagnée de démonstrations dans la vignette Bayesian inference methods.

2.2.1 Principe d'un échantillonneur de Gibbs

Dans le cadre bayésien, l'algorithme de Gibbs permet d'obtenir une réalisation du paramètre $\theta = (\theta_1, \dots, \theta_m)$ suivant la loi *a posteriori* $\pi(\theta | x)$ dès que l'on est capable d'exprimer les lois conditionnelles : $\pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_m, x)$ pour $i = 1, \dots, m$.

L'échantillonage de Gibbs consiste à :

- **Initialisation** : choix arbitraire de $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_m^{(0)})$.
- **Itération t** : Générer $\theta^{(t)}$ de la manière suivante :
 - $\theta_1^{(t)} \sim \pi(\theta_1 | \theta_2^{(t-1)}, \dots, \theta_m^{(t-1)}, x)$
 - $\theta_2^{(t)} \sim \pi(\theta_2 | (\theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_m^{(t-1)}, x))$

- $\theta_m^{(t)} \sim \pi(\theta_m | \theta_1^{(t)}, \dots, \theta_{m-1}^{(t)}, x)$

Les itérations successives de cet algorithme génèrent les états d'une chaîne de Markov $\{\theta^{(t)}, t > 0\}$ à valeurs dans \mathbb{R}^m , on montre que cette chaîne admet une mesure invariante qui est la *loi a posteriori*.

Pour un nombre d'itérations suffisamment grand, le vecteur θ obtenu peut donc être considéré comme étant une réalisation de la loi *a posteriori* jointe $\pi(\theta | x)$.

Par conséquent l'implémentation d'un échantillonneur de Gibbs nécessite la connaissance des distributions *a posteriori* de chacun des paramètres conditionnellement aux autres paramètres du modèle, qui se déduisent des formules de priors conjugués dans le cas du modèle probit mais ne sont pas explicitement exprimables dans le cas où on utilise une fonction de lien logit ou log.

2.2.2 Priors utilisés

Afin d'utiliser une méthode d'inférence bayesienne on détermine une distribution *a priori* pour chacun des paramètres du modèle par défaut dans le package :

$$\begin{aligned} V_\alpha &\sim \mathcal{IG}(\text{shape} = 0.5, \text{rate} = 0.005) \text{ avec rate} = \frac{1}{\text{scale}}, \\ \beta_j &\sim \mathcal{N}_{p+1}(\mu_{\beta_j}, V_\beta) \text{ pour } j = 1, \dots, J \text{ où } V_\beta = \text{diag}(10) \\ &\quad \text{et } \mu_{\beta_{jk}} = \sum_{r=0}^n t_{jr} \cdot \gamma_{rk} \text{ pour } k = 0, \dots, p, \\ \gamma_{rk} &\sim \mathcal{N}(0, 10) \text{ pour } k = 0, \dots, p \text{ et } r = 0, \dots, n, \\ \lambda_{jl} &\sim \begin{cases} \mathcal{N}(0, 10) & \text{si } l < j \\ \mathcal{N}(0, 10) \text{ tronquée à gauche par 0} & \text{si } l = j \\ P \text{ tel que } \mathbb{P}(\lambda_{jl} = 0) = 1 & \text{si } l > j \end{cases} \\ &\quad \text{pour } j = 1, \dots, J \text{ et } l = 1, \dots, q. \end{aligned}$$

En effet pour assurer l'identifiabilité du modèle les valeurs de Λ sont contraintes à des valeurs strictement positives sur la diagonale et nulles au dessus de celle-ci, Λ est ainsi supposée être triangulaire inférieure d'après l'article Warton et al. (2015).

2.2.3 Modèle probit : échantillonneur de Gibbs et priors conjugués

D'une part, on peut utiliser un fonction de lien probit : $p \rightarrow \Phi^{-1}(p)$ où Φ correspond à la fonction de répartition d'une loi normale centrée réduite.

D'après l'article Albert and Siddhartha (1993), une modélisation possible est de supposer l'existence d'une variable latente sous-jacente liée à notre variable binaire observée en utilisant la proposition suivante :

Proposition 2.2.3.1 (Modèle probit par l'intermédiaire d'une variable latente).

Si $Z_{ij} = \alpha_i + \beta_{j0} + X_i \beta_j + W_i \lambda_j + \epsilon_{ij}$, $\forall i, j$ avec $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ iid et tel que :

$$y_{ij} = \begin{cases} 1 & \text{si } Z_{ij} > 0 \\ 0 & \text{sinon.} \end{cases}$$

Alors on a $y_{ij} | Z_{ij} \sim \text{Bernoulli}(\theta_{ij})$ avec $\text{probit}(\theta_{ij}) = \alpha_i + X_i \beta_j + W_i \lambda_j$.

On définit le modèle probit à l'aide d'une variable latente afin d'être en mesure d'utiliser les propositions sur les priors conjugués explicitées et démontrées dans la vignette Bayesian inference methods pour échantillonner les paramètres du modèle selon leurs distributions conditionnelles *a posteriori*.

2.2.4 Modèles logit et log : échantillonneur de Gibbs et algorithme de Metropolis adaptatif

De la même façon que pour le modèle probit, on peut définir les modèles logit et log par l'intermédiaire d'une variable latente mais dans ce cas les distributions *a priori* de la variable latente et des paramètres n'étant pas conjuguées, on n'est pas en mesure d'utiliser les propriétés des priors conjugués donc la modélisation à l'aide d'une variable latente ne présente pas d'intérêt. Par conséquent on échantillonnera les paramètres de ces modèles selon une estimation de leurs distributions conditionnelles *a posteriori* à l'aide d'un algorithme de Metropolis adaptatif.

- Dans le cas du modèle logit on suppose que $y_{ij} | \theta_{ij} \sim \mathcal{B}(n_i, \theta_{ij})$, avec $\text{logit}(\theta_{ij}) = \alpha_i + X_i \beta_j + W_i \lambda_j$ et n_i le nombre de visites du site i .
- Dans le cas du modèle log, utilisé pour ajuster des JSDM à partir de données d'abondance des espèces, on suppose que $y_{ij} | \theta_{ij} \sim \mathcal{P}(\theta_{ij})$, avec $\log(\theta_{ij}) = \alpha_i + X_i \beta_j + W_i \lambda_j$.

Principe d'un algorithme de Metropolis adaptatif :

Cet algorithme appartient aux méthodes MCMC (Markov Chain Monte Carlo) et permet d'obtenir une réalisation du paramètre $\theta = (\theta_1, \dots, \theta_m)$ selon leurs distributions conditionnelles *a posteriori* $\pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_m, x)$, pour $i = 1, \dots, m$ connues à une constante multiplicative près.

On le qualifie d'adaptatif car la variance de la densité instrumentale conditionnelle utilisée est adaptée en fonction du nombre d'acceptation lors des dernières itérations.

- **Initialisation :** $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_m^{(0)})$ fixés arbitrairement, les nombres d'acceptation $(n_i^A)_{i=1, \dots, m}$ sont initialisés à 0 et les variances $(\sigma_i^2)_{i=1, \dots, m}$ sont initialisées à 1.
- **Itération t :** pour $i = 1, \dots, m$
 - Générer $\theta_i^* \sim q(\theta_i^{(t-1)}, \cdot)$, avec une densité instrumentale conditionnelle $q(\theta_i^{(t-1)}, \theta_i^*)$ symétrique, on choisira une loi $\mathcal{N}(\theta_i^{(t-1)}, \sigma_i^2)$ par exemple.
 - Calculer la probabilité d'acceptation :

$$\gamma = \min \left(1, \frac{\pi(\theta_i^* | \theta_1^{(t-1)}, \dots, \theta_{i-1}^{(t-1)}, \theta_{i+1}^{(t-1)}, \dots, \theta_m^{(t-1)}, x)}{\pi(\theta_i^{(t-1)} | \theta_1^{(t-1)}, \dots, \theta_{i-1}^{(t-1)}, \theta_{i+1}^{(t-1)}, \dots, \theta_m^{(t-1)}, x)} \right)$$

- Retenir

$$\theta_i^{(t)} = \begin{cases} \theta_i^* & \text{avec probabilité } \gamma \\ \theta_i^{(t-1)} & \text{si on est dans ce cas le nombre d'acceptation devient : } n_i^A \leftarrow n_i^A + 1 \\ & \text{avec probabilité } 1 - \gamma. \end{cases}$$

- **Durant le burn-in**, toutes les DIV itérations, avec

$$\text{DIV} = \begin{cases} 100 & \text{si } N_{Gibbs} \geq 1000 \\ \frac{N_{Gibbs}}{10} & \text{sinon} \end{cases}$$

, où N_{Gibbs} est le nombre total d'itérations effectuées. On modifie les variances en fonction des nombres d'acceptation de la manière suivante pour $i = 1, \dots, m$:

- On calcule le taux d'acceptation : $r_i^A = \frac{n_i^A}{\text{DIV}}$.
- On adapte les variances selon le taux d'acceptation et une constante fixée R_{opt} :

$$\sigma_i \leftarrow \begin{cases} \sigma_i \left(2 - \frac{1 - r_i^A}{1 - R_{opt}} \right) & \text{si } r_i^A \geq R_{opt} \\ \frac{\sigma_i}{2 - \frac{1 - r_i^A}{1 - R_{opt}}} & \text{sinon} \end{cases}$$

- On réinitialise les nombres d'acceptation : $n_i^A \leftarrow 0$.
- Toutes les $\frac{N_{Gibbs}}{10}$ itérations, on calcule et affiche les taux d'acceptation moyen $m^A = \frac{1}{m} \sum_{i=1, \dots, m} r_i^A$.

2.3 Comparaison des résultats obtenus avec ceux des packages `boral` et `Hmsc`

On a choisi de comparer nos résultats avec ceux de ces deux packages car dans l'article Warton et al. (2015) l'ajustement de modèles joints de distributions des espèces est réalisé à l'aide de `boral` qui fonctionne avec JAGS (Just Another Gibbs Sampler) un programme de simulation à partir de modèles hiérarchiques bayésiens utilisant des méthodes MCMC implémentées en C++. De plus, on s'est inspiré de l'article Wilkinson et al. (2019) qui compare l'ajustement sur différents jeux de données de modèle joints de distribution des espèces et en particulier ceux implémentés par les packages `Hmsc` et `boral`. Les résultats présentés dans la suite sont reproductibles en suivant les démarches détaillées dans les vignettes Comparison jSDM-boral et Comparison jSDM-Hmsc.

2.3.1 Description des jeux de données utilisés

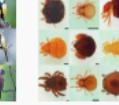
Parmi les jeux de données réels suivants, les données de présence-absence sont issues de l'article Wilkinson et al. (2019), le jeu de données qui recense la présence ou l'absence des oiseaux lors de plusieurs visites sur chaque site provient de l'article Kéry and Schmid (2006) et le jeu de données sur l'abondance des mites est décrit dans l'article Borcard and Legendre (1994). Chacun des ces jeux de données est disponible dans le package `jSDM` accompagné d'une description. De plus, on simule deux jeux de données selon le modèle probit décrit précédemment, d'une part on considère un modèle comprenant un effet site aléatoire α_i dans le cadre de la comparaison avec `boral` et d'autre part on génère un jeu de données qui ne comporte pas d'effet site pour la comparaison avec `Hmsc` dont le cadre bayésien ne définit pas les effets site de la même façon que dans `jSDM`.

Table 1 – Dimensions des jeux de données utilisés (*n.site* et *n.species*), nombre de covariables et axes latents considérés (*n.X.coefs* et *n.latent*) et nombre de paramètres à estimer (*n.param*) en effectuant *n.mcmc* itérations.

	Simulation	Moustiques	Eucalyptus	Grenouilles	Champignons	Oiseaux	Mites
type de données	presence-absence	presence-absence	presence-absence	presence-absence	presence-absence	presence-absence	abondance
distribution	bernoulli	bernoulli	bernoulli	bernoulli	bernoulli	binomiale	poisson
<i>n.site</i>	300	167	455	104	438	266	70
<i>n.species</i>	100	16	12	9	11	110	30
<i>n.latent</i>	2	2	2	2	2	2	2
<i>n.X.coefs</i>	3	14	8	4	13	4	12
<i>n.param</i>	1400	757	1485	366	1479	1458	630
<i>n.mcmc</i>	15000	15000	15000	15000	15000	15000	15000

2.3.2 Comparaison de la pertinence des résultats obtenus et des temps de calcul nécessaires avec chacun des packages

Table 2 – Temps de calcul en secondes nécessaire à l'ajustement du modèle pour chacun des jeux de données et déviances calculées à partir des paramètres estimés avec chacun des packages de la façon suivante : $D = -2 \sum_{i=1}^I \sum_{j=1}^J \log(\mathbb{P}(y_{ij}|\beta_j, \lambda_j, \alpha_i, W_i))$.

	Simulation	Moustiques	Eucalyptus	Grenouilles	Champignons	Oiseaux	Mites
Modèle probit avec effet site aléatoire							
Temps de compilation boral (secondes)	24970	3745	1406	297	1903	4561	734
Temps de compilation jSDM (secondes)	140	19	34	7	33	1737	162
Déviance boral	23327	1014	1811	157	1210	30158	6127
Déviance jSDM	23261	974	1779	107	1211	30122	6081
Modèle probit sans effet site							
Temps de compilation Hmsc (secondes)	641	283	171	125	197	1314	612
Temps de compilation jSDM (secondes)	177	20	39	7	38	178	123
Déviance Hmsc	15421	1768	2656	353	1926	14641	7047
Déviance jSDM	15191	1266	1830	203	1357	14121	6812
							

On constate que **jSDM** est **3 à 268** fois plus rapide que **boral** (JAGS) et **7 à 28** fois plus rapide que **Hmsc**.

Les temps de calcul de **jSDM** sont largement inférieurs à ceux nécessaires à **boral** et **Hmsc** ce qui est dû à l'utilisation du package **Rcpp** pour la construction de **jSDM**, qui permet l'intégration de routines en C++ au sein du package alors que **Hmsc** utilise uniquement du code R ainsi qu'aux méthodes d'inférence utilisées qui diffèrent entre les packages. En effet le tirage selon des lois normales multivariées de certains paramètres par **jSDM** présente un gain de temps considérable par rapport à la méthode MCMC estimant chaque paramètre séparément utilisée par **boral**.

De plus, les déviances obtenues avec **jSDM** sont inférieures à celles calculées avec les résultats de **boral** et **Hmsc** ce qui suggère que les modèles ajustés par **jSDM** correspondent mieux aux données.

Table 3 – Comparaison de la moyenne des carrés des erreurs ou (RMSE) pour Root-Mean-Square Error calculée da la façon suivante $RMSE = \sqrt{\frac{1}{IJ} \sum_{1 \leq i \leq I, 1 \leq j \leq J} (\theta_{ij} - \widehat{\theta}_{ij})^2}$ pour chacun des modèles ajustés sur les jeux de données simulés.

Modèle probit		Modèle probit	
avec effet site		sans effet site	
boral	jSDM	Hmsc	jSDM
RMSE	0.089	0.087	0.073
			0.072

Les RMSE associés à jSDM sont légèrement inférieures à ceux de boral et Hmsc ce qui indique que les résultats obtenus avec jSDM sur les jeux de données simulés sont un peu plus proches de ceux attendus que les paramètres estimés avec boral et Hmsc.

3 Obtention de cartes de communauté à l'échelle du territoire à partir de données d'inventaire forestier

On veut utiliser le package jSDM pour ajuster des JSDMs à partir des inventaires forestiers et des données bioclimatiques présentes et futures dont on dispose sur Madagascar et la Guyane française afin d'obtenir de cartes de communauté à l'échelle du territoire et d'être en mesure d'estimer l'évolution de la biodiversité sur ces territoires sous l'effet des changements climatiques.

3.1 Évolution de la biodiversité à Madagascar sous l'effet des changements climatiques

3.1.1 Description des données

On dispose des inventaires forestiers nationaux réalisés sur 751 sites de l'île de Madagascar et répertoriant la présence ou l'absence de 483 espèces végétales sur chacun de ces sites entre 1994 et 1996.

Parmi les données climatiques et environnementales disponibles sur le site <https://madaclim.cirad.fr> concernant l'ensemble de l'île de Madagascar à l'heure actuelle (interpolations de données observées représentatives des années 1950-2000), on choisit d'utiliser les variables suivantes car elles ont un sens écologique qui les rend facilement interprétables et sont peu corrélées entre elles d'après l'article Vieilledent et al. (2013).

- Les températures (`temp`) moyennes annuelles qui sont exprimées en $^{\circ}C \times 10$.
- Les précipitations (`prec`) moyennes annuelles exprimées en millimètres.
- La saisonnalité des températures (`sais_temp`) qui correspond à l'écart type des températures mensuelles multiplié par 100.
- La saisonnalité des précipitations (`sais_prec`) sous la forme d'un coefficient de variation.
- Le déficit hydrique climatique (`cwd`) annuel exprimé en millimètres qui est calculé en fonction des précipitations et des évapotranspirations potentielles mensuelles (`pet`), définies comme la quantité d'évaporation qui se produirait en un mois si une source d'eau suffisante était disponible : $cwd = \sum_{m=1}^{12} \min(0, prec_m - pet_m)$.

On extrait les valeurs de ces variables climatiques correspondant aux coordonnées des placettes d'inventaires et on considère également les carrés de ces variables climatiques afin d'effectuer un régression quadratique, plus adaptée pour ajuster un modèle de niche qu'une régression linéaire.

On centre et on réduit ces variables afin de former une matrice de design X telle que pour $i = 1, \dots, 751$:

$$X_i = (1, \text{temp}_i, \text{prec}_i, \text{sais_temp}_i, \text{sais_prec}_i, \text{cwd}_i, \text{temp}_i^2, \text{prec}_i^2, \text{sais_temp}_i^2, \text{sais_prec}_i^2, \text{cwd}_i^2)$$

Les coordonnées des sites seront utilisées par la suite dans le cadre de l'interpolation spatiale et pour représenter spatialement les résultats.

De plus, on utilisera les données climatiques future obtenues à partir du portail de données climatiques du CGIAR CCAFS et qui sont disponibles sur <https://madaclim.cirad.fr/future-climate/>. Les conditions futures sont des données de modèle climatique global à échelle réduite provenant de CMIP5 (cinquième évaluation de la IPPC), à une résolution de 30 secondes d'arc (~1km). Nous considérons les conditions futures pour la période des années 2080, avec les scénarios d'émission de CO₂ RCP 8.5, caractérisés par une augmentation des émissions de gaz à effet de serre dans le temps, représentatifs des scénarios conduisant à des niveaux élevés de concentration de gaz à effet de serre dans la littérature. Les paramètres sous-jacents du scénario et la trajectoire de développement qui en résulte sont basés sur le scénario A2r détaillé dans l'article Riahi, Grubler, and Nakicenovic (2007). Nous considérerons ensuite la moyenne des probabilités de présence prédictes à partir des données climatiques future obtenues avec trois modèles climatiques globaux (GISS-E2-R, HadGEM2-ES et NorESM1-M) afin d'estimer l'évolution de la biodiversité à Madagascar sous l'effet du changement climatique et d'identifier des zones refuges de la biodiversité.

3.1.2 Ajustement d'un JSDM à partir de ces données

On ajuste un modèle joint de distribution des espèces de fonction de lien probit en considérant deux variables latentes et un effet site aléatoire à partir des données décrites précédemment, à l'aide de la fonction jSDM_binomial_probit du package `jSDM`, en effectuant 80000 itérations dont 40000 de burn-in et on retient $N_{\text{samp}} = 1000$ valeurs pour chaque paramètre du modèle que l'on va représenter en fonction du nombre d'itérations effectuées afin d'évaluer la convergence de l'algorithme de Gibbs.

De plus, on affiche une estimation de la densité des échantillons obtenus qui devrait correspondre aux distributions *a posteriori* définies et donc être de forme gaussienne.

On met en évidence les moyennes des N_{samp} -échantillons en bleu, que l'on utilisera comme estimateur pour les paramètres.

Table 4 – Dimensions des jeux de données utilisés (*n.site* et *n.species*), nombre de covariables et axes latents considérés (*n.X.coefs* et *n.latent*) et nombre de paramètres à estimer (*n.param*) en effectuant *n.mcmc* itérations ainsi que le temps de calcul nécessaire à l'ajustement du modèle sur les données de Madagascar et la déviance obtenue.

<i>n.sites</i>	<i>n.species</i>	<i>n.latent</i>	<i>n.X.coefs</i>	<i>n.param</i>	<i>n.mcmc</i>	Temps de calcul (heures)	Déviance
751	483	2	483	236991	80000	1.8	32969.6

Figure 2 – Traces et densités de la déviance du modèle

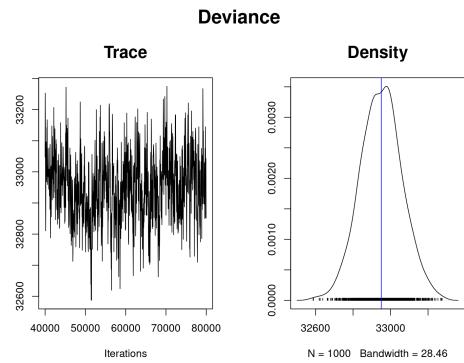


Figure 3 – Trace et densité d'un effet site et de la variance associée aux effets sites estimés

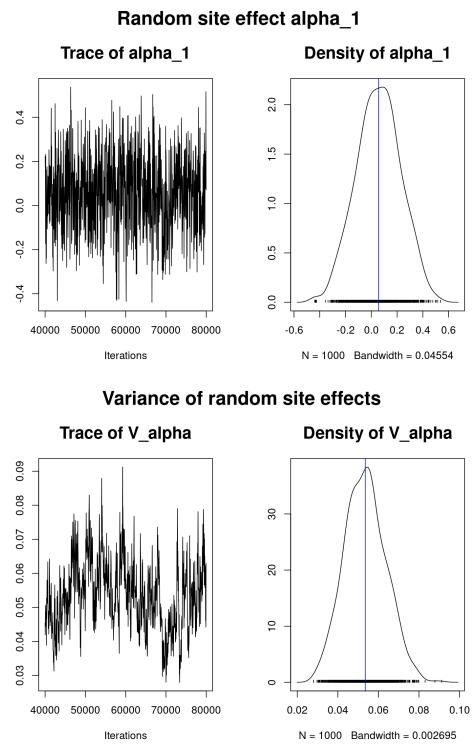


Figure 4 – Trace et densité estimées des variables latentes W_1 et W_2 estimées pour un site

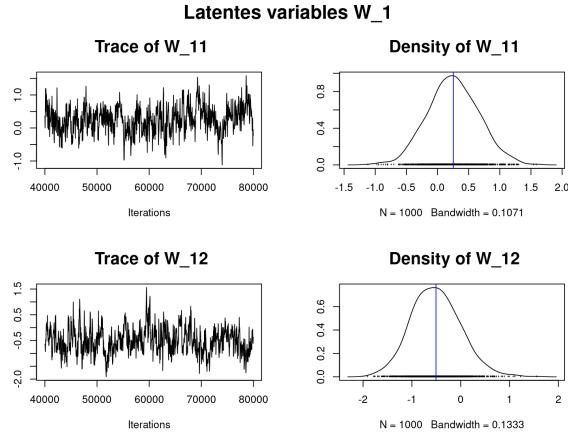


Figure 5 – Traces et densités des factor loadings $(\lambda_{jq})_{j=1,2}^{q=1,2}$ estimés pour les deux premières espèces

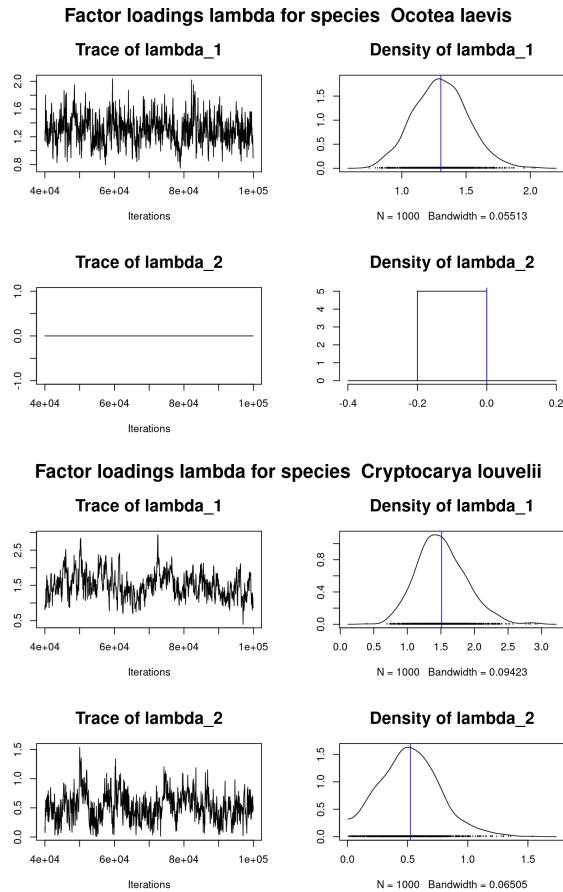
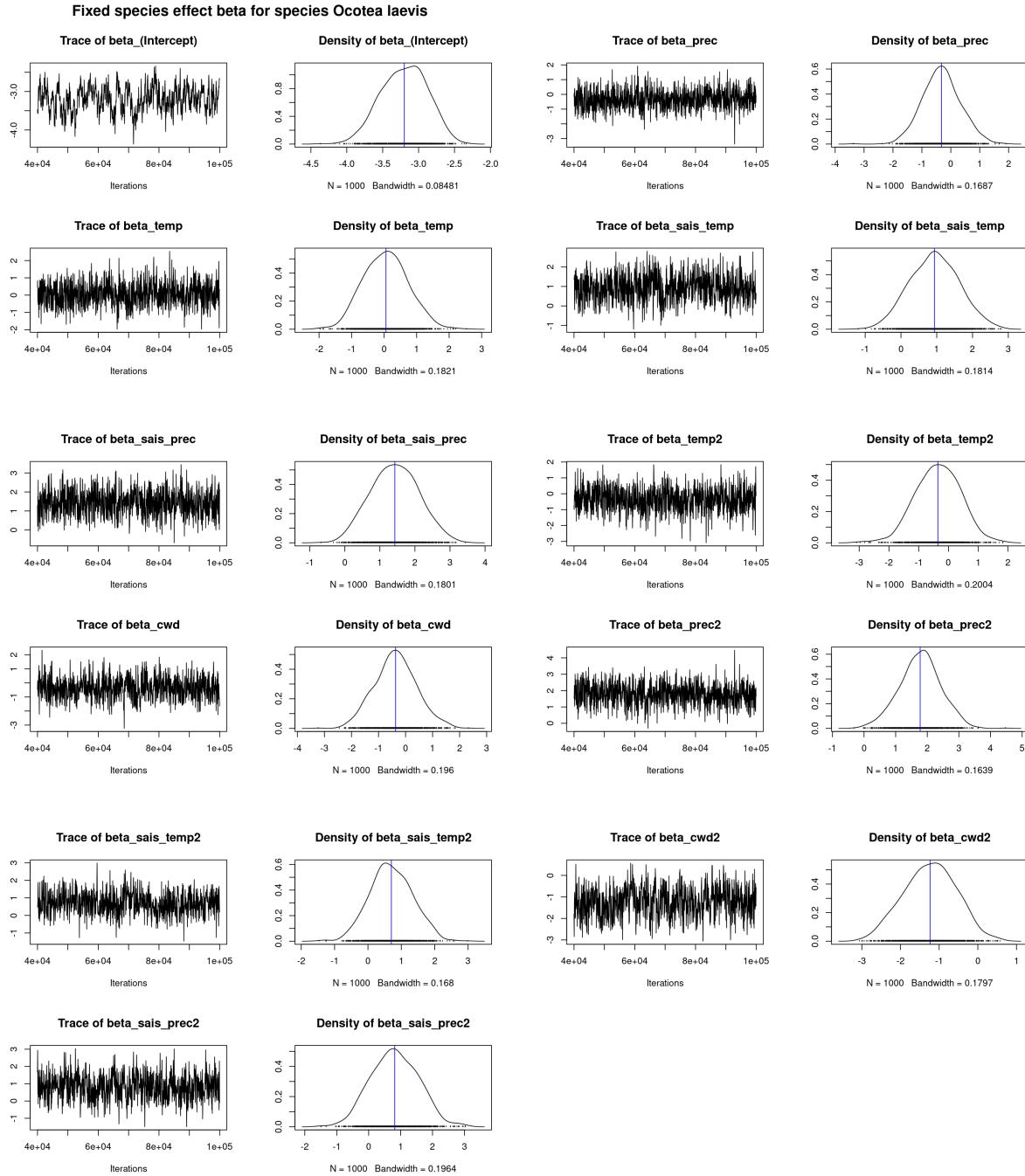


Figure 6 – Traces et densités des effets espèces fixes $(\beta_{jk})_{j=1}^{k=0, \dots, 10}$ estimés pour la première espèce



Dans l'ensemble, les traces et les densités des paramètres indiquent la convergence de l'algorithme. En effet on observe sur les traces que les valeurs oscillent autour de moyennes sans présenter de tendance croissante ou décroissante et on constate que les densités sont assez lisses et pour la plupart de forme gaussienne.

3.1.3 Corrélation résiduelle entre les espèces estimée

Après avoir ajusté le JSDM, d'après les articles Warton et al. (2015) et Tobler et al. (2019), la **matrice de corrélation résiduelle** $R = (R_{ij})_{j=1,\dots,J}^{i=1,\dots,J}$ peut être obtenue de la manière suivante :

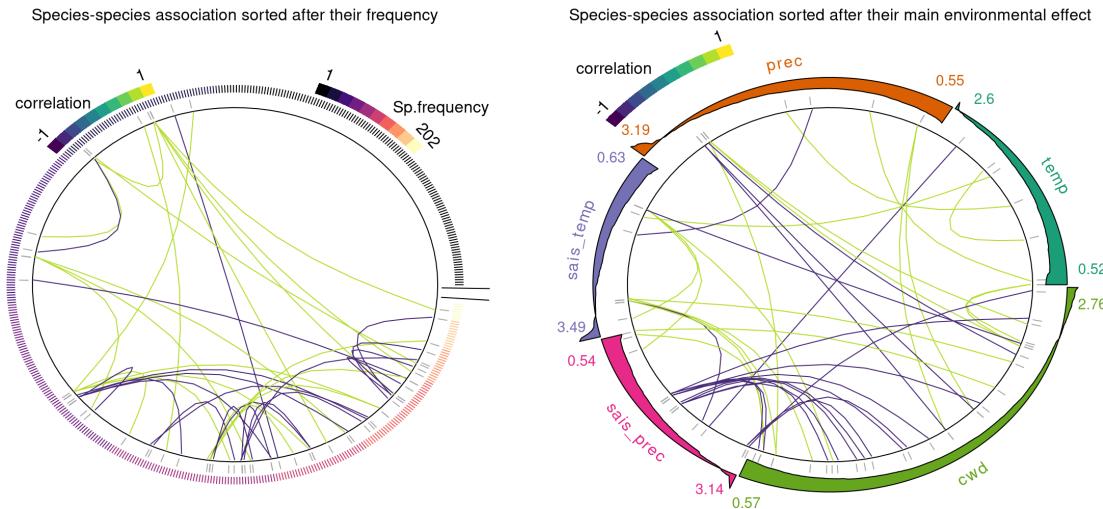
$$\Sigma_{ij} = \begin{cases} \lambda_i \cdot \lambda_j^T & \text{if } i \neq j \\ \lambda_i \cdot \lambda_j^T + 1 & \text{if } i = j \end{cases}$$

, on calcule ensuite les corrélations à partir des covariances :

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

Le nombre d'espèces considérées étant élevé, une représentation de la matrice de corrélation résiduelle à l'aide de la fonction `plot_residual_corr` serait illisible, on choisit donc de représenter les associations prépondérantes à l'aide de la fonction `plot_associations`.

Figure 7 – Représentation des corrélations résiduelles estimées des espèces et leur préférences environnementales. La figure de gauche montre les corrélations entre espèces, avec les 483 espèces triées selon le nombre de sites où elles sont présentes sur les 751 inventoriés et la figure de droite montre la même structure de covariance mais avec les espèces triées selon leurs coefficients environnementaux (β) les plus importants (l'anneau extérieur montre la distribution de l'effet environnemental pour les espèces au sein de l'échantillon).



Cette représentation des associations entre espèces permet d'observer les corrélations positives ou négatives entre les espèces qui sont interprétables en terme d'influence positive ou négative de la présence d'un espèce sur la probabilité d'occurrence d'une autre.

3.1.4 Estimation de la richesse spécifique pour les placettes d'inventaire et comparaison à celle observée

La richesse spécifique aussi appelée diversité α reflète le nombre d'espèces coexistant dans un milieu donné, on l'estime en additionnant les probabilités de présence estimées.

Figure 8 – Représentation spatiale de la richesse spécifique estimée pour chaque site par $\widehat{R}_i = \sum_{j=1}^{483} \widehat{\theta}_{ij}$.

Figure 9 – Représentation spatiale de la richesse spécifique observée pour chaque site calculée par $R_i = \sum_{j=1}^{483} y_{ij}$

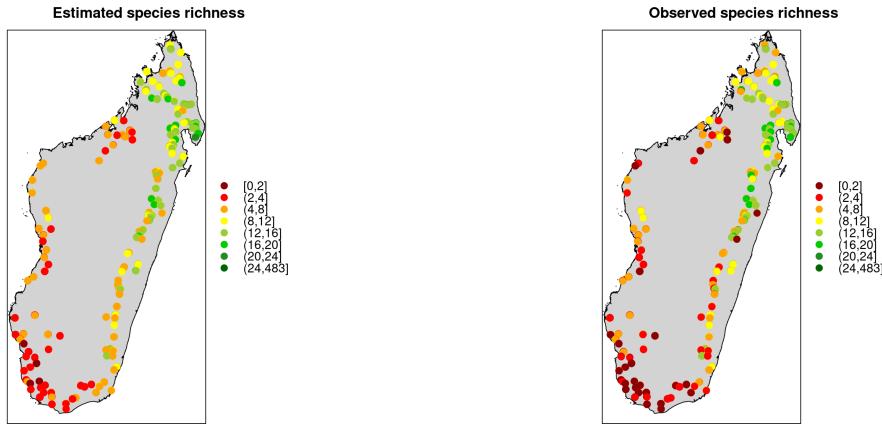
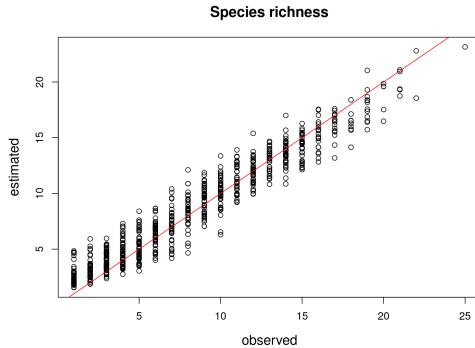


Figure 10 – Représentation de la richesse spécifique estimée en fonction de celle observée



3.1.5 Interpolation spatiale des paramètres associés aux sites d'inventaire

La présence d'une structure spatiale où les observations proches les unes des autres sont plus semblables que celles qui sont éloignées (auto-corrélation spatiale) est une condition préalable à l'application de la géostatistique, elle semble être remplie dans notre cas. Ainsi, il devrait être possible d'interpoler les paramètres des sites pour l'ensemble de l'île à partir de ceux estimés sur les placettes d'inventaire. On utilise maintenant la méthode d'interpolation spatiale appelée Regularized Spline with Tension (**RST**) du logiciel GRASS GIS via le package R **rgrass7**. Cette méthode est décrite dans l'article (Mitášová and Hofierka 1993).

Nous avons rencontré quelques difficultés pour effectuer cette étape d'interpolation. Avant d'utiliser la méthode **RST**, nous avons essayé trois méthodes d'interpolation spatiales disponibles dans le package **gstat** afin de choisir celle qui donne les meilleurs résultats en s'inspirant de l'article Robinson and Metternicht (2006).

D'une part, nous avons utilisé la méthode déterministe de pondération par distance inverse, nommée **IDW** pour l'interpolation multivariée à partir de l'ensemble connu de points dispersés. Les valeurs attribuées à des points inconnus sont calculées avec une moyenne pondérée des valeurs disponibles aux sites connus qui fait appel à l'inverse de la distance par rapport à chaque point connu lors de l'attribution des poids.

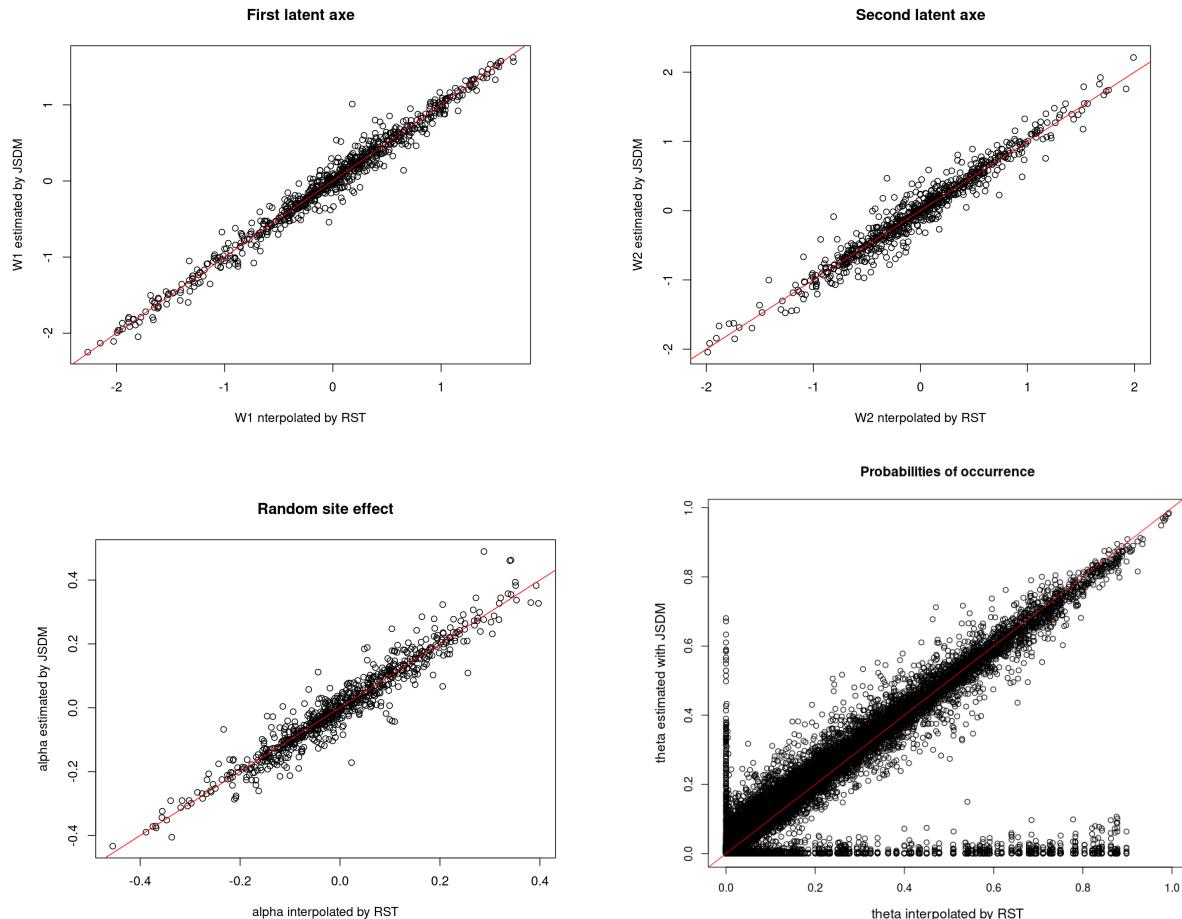
D'autre part, nous avons procédé par krigeage ordinaire (**OK**) ce qui consiste à considérer les valeurs des

sites inconnus comme une combinaison linéaire des valeurs connues dont les coefficients sont estimés en minimisant la variance de l'erreur d'estimation théorique qui dépend des coefficients ainsi que du variogramme expérimental tenant compte non seulement de la distance entre les données et les points d'estimation, mais également des distances entre les données deux-à-deux.

Enfin nous avons appliqué la méthode **TPS** pour thin plate spline pour laquelle une fonction thin plate spline en deux dimensions est ajustée sur les coordonnées et les valeurs des points connus afin d'interpoler les valeurs non observées en fonction de leurs positions.

Nous avions choisi d'interpoler les variables latentes et les effets sites estimés pour les placettes d'inventaires par krigeage ordinaire (OK) car cette méthode présentait un RMSE, calculé par validation croisée entre les effets sites estimés et ceux prédits par interpolation, inférieur à ceux obtenus avec les méthodes TPS et IDW. Cependant, nous avons constaté que l'interpolation ne conservait pas les valeurs des paramètres estimées par le JSDM sur les sites d'inventaire, ce qui biaisait largement les prédictions. La méthode RST devrait résoudre ce problème mais d'après les figures suivantes ce n'est pas le cas.

Figure 11 – Représentation des paramètres estimées par le JSDM sur les sites d'inventaire en fonction de ceux interpolés par RST.



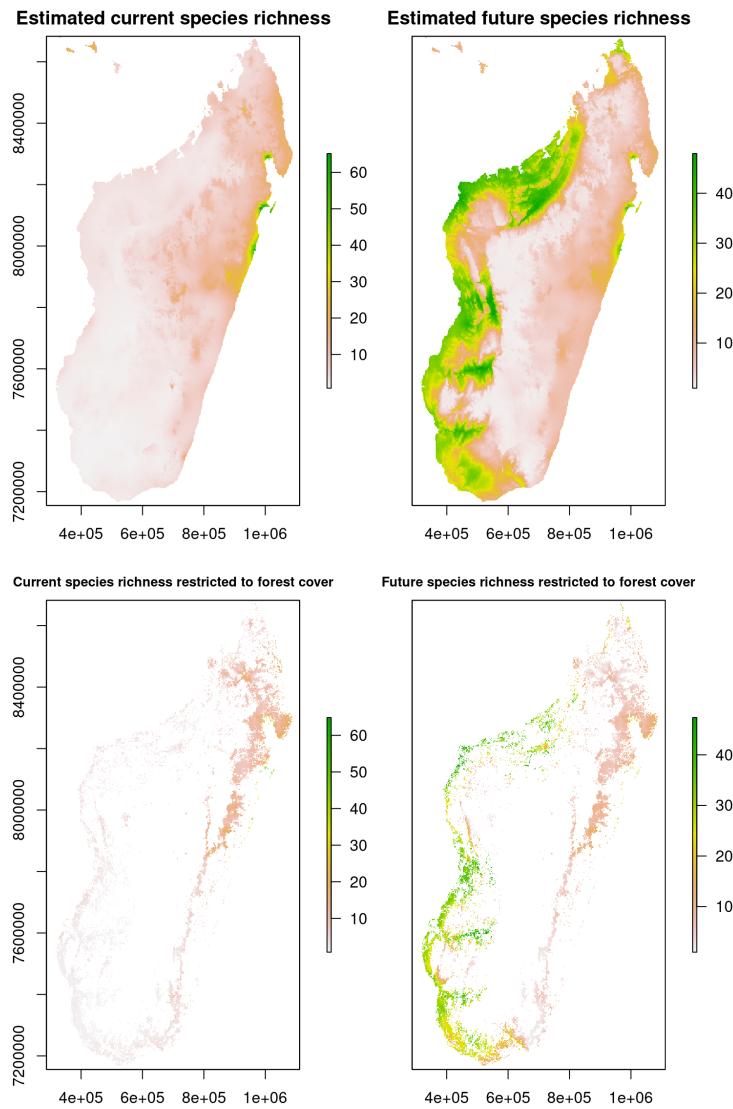
Nous avons tout de même poursuivi la démarche avec l'intention d'améliorer la méthode d'interpolation plus tard. Nous avons utilisé les paramètres interpolés par RST pour calculer les probabilités de présence sur l'île de chacune des espèces en fonction des variables climatiques présentes et futures définies précédemment dont les valeurs sont connues pour l'ensemble de l'île et sont centrées et réduites pour le calcul.

3.1.6 Evolution de la richesse spécifique à Madagascar

Nous avons additionné les probabilités de présence interpolées pour chacune des espèces afin d'obtenir la richesse spécifique actuelle et dans le futur.

Cependant, le modèle utilisé ne prend pas en compte la présence humaine qui se manifeste en particulier par la déforestation de l'île, on utilise donc les données sur le couvert forestier restant en 2000 provenant de l'article Vieilledent et al. (2018) afin de remplacer par des valeurs nulles les richesses spécifiques interpolées à des endroits où on sait qu'il n'y a pas de forêt.

Figure 12 – Richesse spécifique interpolée sur l'ensemble de l'île et restreinte au couvert forestier



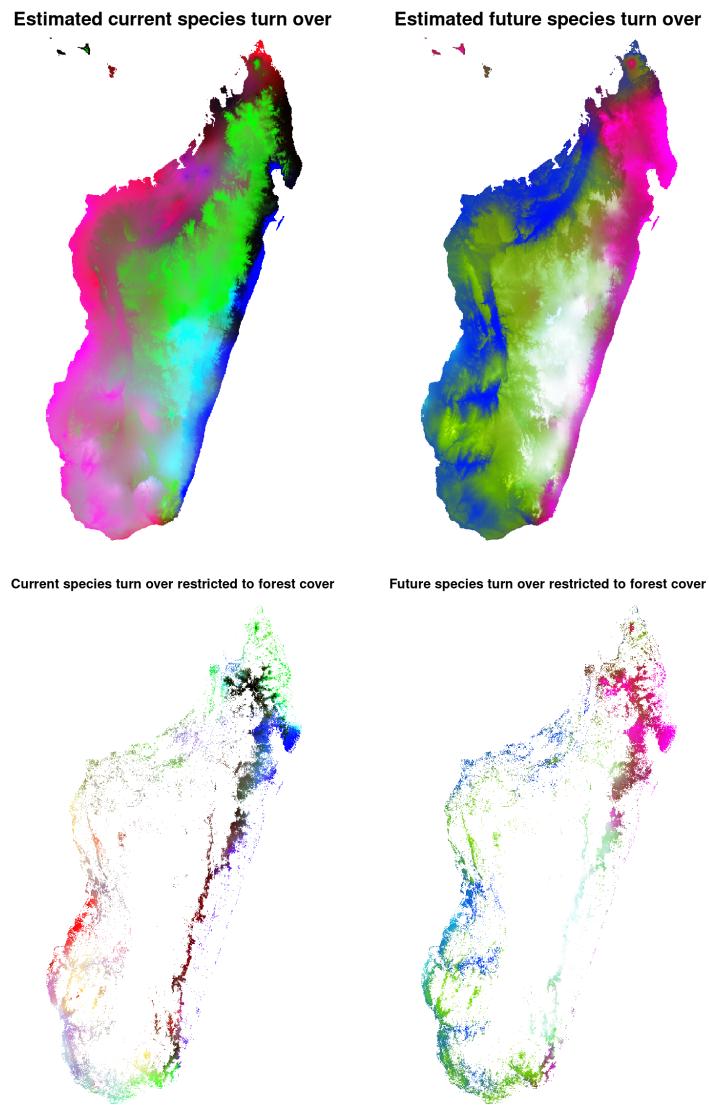
3.1.7 Evolution de la diversité β à Madagascar

La diversité β est une mesure de la biodiversité qui consiste à comparer la diversité des espèces entre écosystèmes ou le long de gradients environnementaux, en utilisant le nombre de taxons qui sont uniques à chacun des écosystèmes.

Afin d'estimer cet indicateur, on procède de la même façon que dans l'article Allnutt et al. (2008) en effectuant une ACP normée sur les probabilités de présence des espèces interpolées pour chaque pixel de l'image affichée. On utilise les coordonnées obtenues pour les trois premiers axes de l'ACP qui reflètent la composition de la communauté d'espèces occupant probablement le pixel correspondant. Ces coordonnées sont mises à l'échelle [0, 255] afin d'être représentables par des niveaux de couleur rouge pour le premier axe, verte pour le deuxième et bleue pour le troisième, l'association de ces trois niveaux de couleur détermine la coloration de chaque pixel de la carte de diversité β affichée. Par conséquent une différence de couleur entre deux pixels indique que les espèces présentes ne sont pas les mêmes tandis que des pixels de couleur identiques hébergent des communautés d'espèces similaires.

De la même manière que précédemment, on restreint les valeurs obtenues pour la diversité β au couvert forestier restant en 2000.

Figure 13 – Diversité β interpolée sur l'ensemble de l'île et restreinte au couvert forestier connu en 2000



J'ai représenté ces cartes de biodiversité α et β afin d'illustrer la méthode mise en oeuvre pour les obtenir mais elle ne sont pas cohérentes car elles sont basées sur les résultats de l'interpolation par RST qui nécessite d'être améliorée.

3.2 Obtention de cartes de communauté en Guyane française

3.2.1 Données utilisées

Une fois qu'on aura réussi à obtenir des résultats satisfaisants pour Madagascar, on suivra la même méthode pour obtenir des cartes de communauté en Guyane française en utilisant les inventaires forestiers et données climatiques présentes et futures disponibles mais aussi les bases de données de traits fonctionnels uniques dont dispose la communauté scientifique du CEBA.

En effet pour la Guyane on ajustera un JSDM prenant en compte des traits spécifiques comme variables explicatives du modèles. On considérera des traits fonctionnels classiques comme la densité du bois ou la surface foliaire spécifique mais les études récentes ne montrent pas de lien évident entre ces traits et la résistance à la sécheresse (Maréchaux et al. 2020). Par conséquent on utilisera également des traits plus mécanistes et écophysiologique (soft traits) comme le point de perte de la turgescence des feuilles (ou leaf turgor loss point) qui présentent un lien plus direct avec la résistance à la sécheresse (Maréchaux et al. 2018).

Je vais d'ailleurs participer en octobre à une mission de mesure de traits hydrauliques en forêt guyanaise dans le cadre du projet stratégique du CEBA nommé METRADICA qui vise à prédire l'évolution de l'abondance et de la distribution des espèces d'arbres en Amazonie sous l'effet du changement climatique en fonction des interactions entre traits mécanistes et environnement. Il me paraît important de participer à l'effort de récolte de ces données sans lesquelles ma thèse ne serait pas faisable mais aussi d'apprendre la façon dont les données sont mesurées et organisées afin d'être en mesure de les analyser et de les interpréter au mieux.

3.2.2 Enjeux

Ce projet de thèse contribuera à apporter des éléments de réponse à une question fondamentale au centre de nombreux projets de recherche en écologie actuellement qui est de savoir si la forêt amazonienne sera capable de résister au changement climatiques à venir.

Dans un premier temps on utilisera le package jSDM afin d'ajuster des modèle joints de distribution des espèces à partir des données (bio-climatiques, d'inventaire forestier et de traits fonctionnels) disponibles sur la forêt Guyanaise pour ensuite être en mesure de prédire l'évolution des aires de distribution des espèces sous l'effet du changements climatique.

D'une part les résultats obtenus pourraient laisser supposer que la forêt amazonienne montera une certaine résilience face au changement climatique qui correspondrait à un changement des aires de répartition des espèces et donc de la composition de la forêt tropicale mais avec conservation du couvert forestier et de la capacité des forêts à absorber et stocker le dioxyde de carbone (CO₂) et à diffuser fraîcheur et humidité sur les continents.

D'autre part les prédictions pourraient au contraire mettre en évidence la vulnérabilité de la forêt amazonienne face au changement climatiques qui se manifesterait par une contraction généralisée des aires de répartition des espèces associée à un phénomène de mortalité en masse. Si ce scénario s'avère le plus probable il est important d'être en mesure d'anticiper un emballement soudain du réchauffement climatique engendré par les effets de rétroaction sur le climat de l'effondrement des écosystèmes forestiers qui amplifierait les bouleversements climatiques.

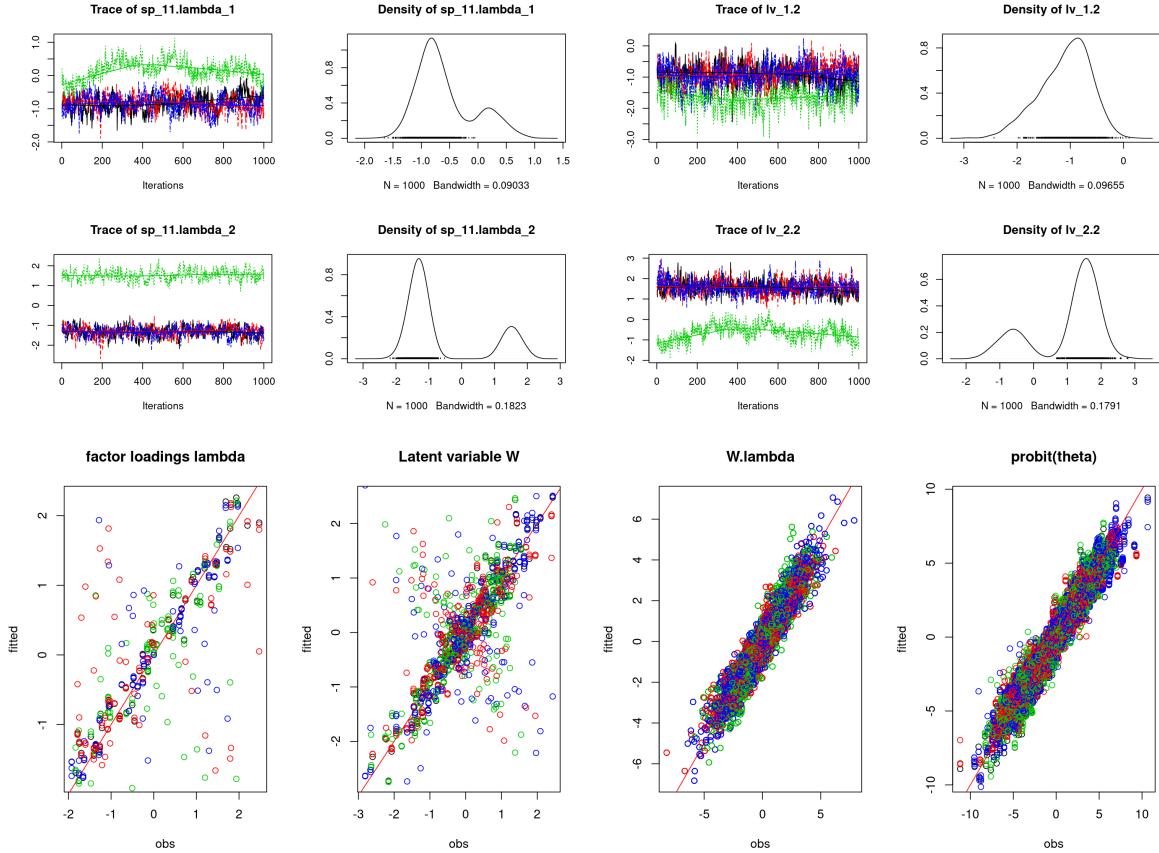
4 Fonctionnalités en cours de développement du package jSDM

4.1 Estimation du nombre d'axes latents à prendre en compte et amélioration de leur convergence

On rencontre des difficultés à estimer les axes latents (W) et les facteurs associés (Λ) liées aux contraintes imposées à la matrice Λ supposée être triangulaire inférieure et strictement positives sur la diagonale pour assurer l'identifiabilité du modèle.

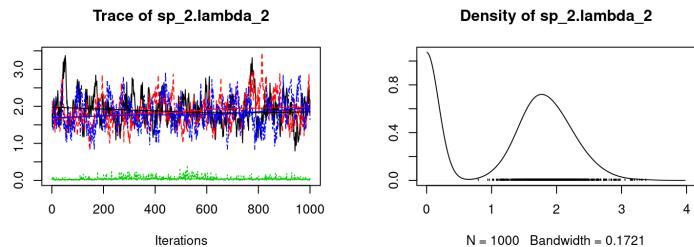
- Problème “symétrique” de convergence des facteurs latents et des axes latents

Figure 14 – Illustration des problèmes de convergence des facteurs latents et des axes latents en représentant les traces et les densités de 4 chaînes MCMC ajustées à partir de différentes valeurs initiales sur un jeu de données simulé ainsi que les paramètres estimés en fonction des valeurs attendues.



- Associés à des problèmes non symétriques pour les espèces “références” de l’axe latent, c’est à dire les espèces j correspondants aux facteurs latents diagonaux contraints à la positivité : λ_{jl} tels que $j = l$.

Figure 15 – Illustration des problèmes de convergence des facteurs latents sur la diagonale en représentant les traces et les densités de 4 chaînes MCMC ajustées à partir de différentes valeurs initiales sur un jeu de données simulé.



On observe sur la figure 14 une inversions des signes d’une partie des facteurs latents et des variables latentes

par rapport au valeurs attendues, qui ne faussent pas les prédictions mais rendent plus difficile la convergence de l'algorithme.

On constate sur la figure 15 que l'un des facteurs latents constraint à la positivité oscille autour de valeurs très proches de 0 sur l'une des chaîne MCMC considérée ce qui peut signifier que l'espèce associée n'est pas la plus indiquée pour structurer les axes latents.

On travaille sur ce sujet en collaboration avec Frédéric Gosselin, chercheur INRAE qui s'intéresse à cette problématique. Il a développé les méthodes suivantes pour améliorer la convergence des axes latents et des facteurs associés mais aussi pour estimer le nombre d'axes latents (n_{latent}) que le JSDM devrait considérer plutôt que de fixer ce nombre arbitrairement comme on le faisait jusqu'à maintenant.

- Évaluer la convergence des paramètres associés aux axes latents et diagnostiquer l'espèce qui se structure le plus clairement sur chaque axe, en déduire sur quelles espèces imposer les contraintes de positivité sur la diagonale et réordonner les espèces avant d'ajuster un nouveau JSDM.
- Ajouter une variable dans le modèle hiérarchique Bayésien qui va estimer le nombre d'axes latents à considérer.

On va essayer d'intégrer ces méthodes dans le cadre du package `jSDM`, on envisage aussi de modifier les contraintes imposées aux facteurs latents en s'inspirant de l'article Peeters (2012).

4.2 Intégrer la phylogénie des espèces comme variable explicative des modèles

D'après l'article Ovaskainen et al. (2017), pour tenir compte des relations phylogénétiques (résumées par la matrice C), on peut modéliser la structure de la covariance de la distribution normale multivariée des effets espèces β de la façon suivante :

$$\beta \sim \mathcal{N}(\mu, V \otimes [\rho C + (1 - \rho)I])$$

où le symbole \otimes représente le produit de Kronecker et $0 \leq \rho \leq 1$ mesure la force du signal phylogénétique. Par conséquent pour $\rho = 0$, la variance résiduelle est indépendante entre les espèces (comme décrit par la matrice d'identité I), ce qui implique que les espèces étroitement liées n'ont pas des niches environnementales plus similaires que les espèces éloignées. Lorsque ρ s'approche de $\rho = 1$, les niches environnementales des espèces sont entièrement structurées par leur phylogénie, ce qui implique que les espèces apparentées auront des niches plus similaires.

On va reproduire cette approche dans le package `jSDM`.

4.3 Ajuster des JSDMs spatialement explicites

Il est essentiel de développer le package pour lui permettre d'ajuster des JSDMs spatialement explicites afin d'être en mesure de prédire les probabilités de présence des espèces au-delà des sites d'inventaires, sans passer par une interpolation des paramètres dont les résultats ne sont pas satisfaisants comme expliqué précédemment.

Pour ce faire j'ai envisagé deux méthodes explicitées dans les articles Guélat and Kéry (2018) Latimer et al. (2006) afin d'intégrer une auto-corrélation spatiale dans le modèle :

D'une part j'ai développé une fonction ajustant un modèle gaussien auto-régressif conditionnel (CAR) intrinsèque en utilisant une grille sur l'ensemble du territoire considéré dont chacune des cellules possède au plus huit voisines. On estimera ainsi les valeurs des effets sites et des variables latentes pour chacune des cellules en fonction de ceux estimés pour les cellules voisines.

En effet dans le contexte des modèles de répartition des espèces, on suppose que la présence ou l'absence d'une espèce à un endroit est associée à sa présence ou son absence dans le voisinage. Afin de prendre en compte les voisinages, les distributions *a priori* des paramètres liés aux sites sont centrées sur la moyenne des valeurs prises par ces paramètres dans les cellules voisines et leurs variances dépendent du nombre de cellules partageant des frontières avec la cellule considérée. Cependant cette méthode induit un temps de calcul important et présente des difficultés à converger car les paramètres à estimer sont très nombreux en raison du nombre conséquent de cellules constituant la grille utilisée.

D'autre part j'ai essayé d'intégrer un 2D splines dans le modèle en redéfinissant les effets sites par le produit d'une matrice calculée en fonction de la distance du site par rapport à des noeuds répartis uniformément sur le territoire étudié, et de paramètres à estimer qui sont aussi nombreux que les noeuds choisis. On ajoute également les coordonnées des sites parmi les variables explicatives du modèle ce qui induit deux paramètres supplémentaires à estimer.

Cependant ces fonctions intégrant une auto-corrélation spatiale dans le modèle, qui rendraient possible la prédiction sur des sites non observés sont implémentées mais ne sont pas abouties, en effet les résultats obtenus ne sont pas satisfaisants pour l'instant.

4.4 Ajuster des JSDMs à partir de données de présence seule

On voudrait être en mesure d'ajuster des JSDMs à partir de jeux de données d'occurrence qui ne répertorient pas les absences des espèces, comme les données d'herbier ou l'immense base de donnée GBIF qui rassemble des milliers de jeux de données concernant de nombreuses espèces. Pour ce faire, le plus simple serait de générer des pseudo-absences en suivant une méthode adaptée comme celle développée dans l'article Barbet-Massin et al. (2012), afin de prendre en charge les données de présence seule en utilisant les mêmes algorithmes que pour les données de présence-absence. Cependant cette idée n'a pas encore été développée et s'avérera peut être difficile à mettre en oeuvre.

Bibliographie

- 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/10.1080/01621459.1993.10476321>.
- 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>.
- Barbet-Massin, Morgane, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. 2012. "Selecting Pseudo-Absences for Species Distribution Models : How, Where and How Many?" *Methods in Ecology and Evolution* 3 (2) : 327–38. <https://doi.org/10.1111/j.2041-210X.2011.00172.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>.
- Bunker, Daniel, Fabrice Declerck, Jason Bradford, Robert Colwell, Ivette Perfecto, Oliver Phillips, Mahesh Sankaran, and Shahid Naeem. 2005. "Ecology : Species Loss and Aboveground Carbon Storage in a Tropical." *Science (New York, N.Y.)* 310 (December) : 1029–31. <https://doi.org/10.1126/science.1117682>.
- Clark, James S., Diana Nemergut, Bijan Seyednasrollah, Phillip J. Turner, and Stacy Zhang. 2017. "Generalized Joint Attribute Modeling for Biodiversity Analysis : Median-Zero, Multivariate, Multifarious Data." *Ecological Monographs* 87 (1) : 34–56. <https://doi.org/10.1002/ecm.1241>.
- 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, 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>.
- Guélat, Jérôme, and Marc Kéry. 2018. "Effects of Spatial Autocorrelation and Imperfect Detection on Species Distribution Models." *Methods in Ecology and Evolution* 9 (6) : 1614–25. <https://doi.org/10.1111/2041-210X.12983>.
- Kéry, Marc, and Hans Schmid. 2006. "Estimating Species Richness : Calibrating a Large Avian Monitoring Programme." *Journal of Applied Ecology* 43 (1) : 101–10. <https://doi.org/10.1111/j.1365-2664.2005.01111.x>.
- Latimer, Andrew M., Shanshan Wu, Alan E. Gelfand, and John A. Silander. 2006. "Building Statistical Models To Analyze Species Distributions." *Ecological Applications* 16 (1) : 33–50. <https://doi.org/10.1890/04-0609>.
- Maréchaux, Isabelle, Damien Bonal, Megan Bartlett, Benoit Burban, Sabrina Coste, Elodie Courtois, Maguy Dulormne, et al. 2018. "Dry-Season Decline in Tree Sapflux Is Correlated with Leaf Turgor Loss Point in a Tropical Rainforest." *Functional Ecology* 32 (July). <https://doi.org/10.1111/1365-2435.13188>.
- Maréchaux, Isabelle, Laurent Saint-André, Megan K. Bartlett, Lawren Sack, and Jérôme Chave. 2020. "Leaf Drought Tolerance Cannot Be Inferred from Classic Leaf Traits in a Tropical Rainforest." *Journal of Ecology* 108 (3) : 1030–45. <https://doi.org/10.1111/1365-2745.13321>.
- 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>.
- 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>.
- Peeters, Carel F. W. 2012. "Rotational Uniqueness Conditions Under Oblique Factor Correlation Metric." *Psychometrika* 77 (2) : 288–92. <https://doi.org/10.1007/s11336-012-9259-3>.
- Pichler, Maximilian, and Florian Hartig. 2020. *A New Method for Faster and More Accurate Inference of Species Associations from Novel Community Data*.
- Riahi, Keywan, Arnulf Grubler, and Nebojsa Nakicenovic. 2007. "Scenarios of Long-Term Socio-Economic and Environmental Development Under Climate Stabilization." *Technological Forecasting and Social Change* 74 (1) : 1–16.

- Change* 74 (September) : 887–935. <https://doi.org/10.1016/j.techfore.2006.05.026>.
- Robinson, T. P., and G. Metternicht. 2006. “Testing the Performance of Spatial Interpolation Techniques for Mapping Soil Properties.” *Computers and Electronics in Agriculture* 50 (2) : 97–108. <https://doi.org/10.1016/j.compag.2005.07.003>.
- Tobler, Mathias W., Marc Kéry, Francis K. C. Hui, Gurutzeta Guillera-Arroita, Peter Knaus, and Thomas Sattler. 2019. “Joint Species Distribution Models with Species Correlations and Imperfect Detection.” *Ecology* 100 (8) : e02754. <https://doi.org/10.1002/ecy.2754>.
- Vieilledent, Ghislain, Jeanne Clément, and CIRAD. 2019. “jSDM : Joint Species Distribution Models.” <https://CRAN.R-project.org/package=jSDM>.
- 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, Oliver Gardi, Clovis Grinand, Christian Burren, Mamitiana Andriamanjato, Christian Camara, Charlie J. Gardner, et al. 2016. “Bioclimatic Envelope Models Predict a Decrease in Tropical Forest Carbon Stocks with Climate Change in Madagascar.” *Journal of Ecology* 104 (3) : 703–15. <https://doi.org/10.1111/1365-2745.12548>.
- 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>.
- 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>.