



RAPPORT DE COMITE DE SUIVI INDIVIDUEL (CSI)

En 2^{ème} année pour la réinscription en 3^{ème} année

POUR EN FACILITER LA LECTURE, LES TEXTES DES FORMULAIRES DE COMITE DE SUIVI SONT VALABLES AU MASCULIN COMME AU FEMININ

Document mis à jour 18 janvier 2022

La tenue d'un comité de suivi individuel (CSI) en absence de la direction de thèse (arrêté du 25 mai 2016, article 13) est obligatoire pour toute inscription en troisième année, et au delà.

Arrêté du 25 mai 2016, article 13

« [...] L'inscription est renouvelée au début de chaque année universitaire par le chef d'établissement, sur proposition du directeur de l'école doctorale, après avis du directeur de thèse et, à partir de la troisième inscription, du comité de suivi individuel du docteurant. En cas de non-renouvellement envisagé, après avis du directeur de thèse, l'avis motivé est notifié au docteurant par le directeur de l'école doctorale. Un deuxième avis peut être demandé par le docteurant auprès de la commission recherche du conseil académique ou de l'instance qui en tient lieu, dans l'établissement concerné. La décision de non-renouvellement est prise par le chef d'établissement, qui notifie celle-ci au docteurant. [...] »

Un comité de suivi individuel du docteurant veille au bon déroulement du cursus en s'appuyant sur la charte du doctorat et la convention de formation. Il évalue, dans un entretien avec le docteurant, les conditions de sa formation et les avancées de sa recherche. Il formule des recommandations et transmet un rapport de l'entretien au directeur de l'école doctorale, au docteurant et au directeur de thèse. Il veille notamment à prévenir toute forme de conflit, de discrimination ou de harcèlement. Les modalités de composition, d'organisation et de fonctionnement de ce comité sont fixées par le conseil de l'école doctorale. **Les membres de ce comité ne participent pas à la direction du travail du docteurant.**

AVANT DE COMMENCER A REMPLIR CE DOCUMENT : Merci de vérifier qu'il s'agit de la dernière version, disponible sur le site internet de l'ED GAIA (au bas de la page d'accueil, rubrique "Documents à télécharger"). Les versions précédentes ne seront pas prises en compte.

NB: Penser à consulter la convention de formation individuelle (CFI) (arrêté du 25 mai 2016, article 12) signée entre le docteurant et la direction de thèse comprenant des informations sur la situation contractuelle du docteurant, son projet doctoral, la planification et la valorisation des travaux, les modalités de l'encadrement, les projets et le parcours de formation.

Consignes pour l'organisation du CSI :

- **Le CSI sera organisé entre le 1^{er} mars et le 7 octobre AU PLUS TARD et devra dater de moins de 6 mois au moment de la réinscription**
- Le comité, présidé par le référent, se déroule sans les directeurs/encadrants de thèse.
- Le référent doit également prévoir un entretien avec l'équipe de direction en l'absence du docteurant.
- Le rapport du CSI sera déposé par le référent via son compte personnel ADUM au plus tard 1 semaine après la tenue du CSI.
- Le référent doit envoyer le rapport du CSI à l'ensemble des membres du comité, au docteurant et à l'équipe de direction de la thèse.

Première partie : A pré-remplir par le Doctorant (et à vérifier par le DT)

Nom et prénom du doctorant : Clément Jeanne

Date de première inscription en thèse : 11/01/2021

Date du comité de suivi de thèse : 27/09/2022

Date de soutenance envisagée : 15/12/2023

Titre de la thèse : Prédire la vulnérabilité des espèces d'arbres au changement climatique en forêt Guyanaise via l'utilisation de modèles joints de distribution des espèces

Filière doctorale : EERGP - Écologie, Evolution, Ressources Génétique, Paléobiologie

Discipline du doctorat : Ecologie et Biodiversité

Unité de recherche de rattachement : UMR AMAP

Mode de financement : Labex CEBA

Confidentialité : OUI / NON (rayer la mention inutile)

Cotutelle internationale : OUI / NON (rayer la mention inutile)

Noms du directeurs (HDR), du co-directeur (HDR) et des encadrants de la thèse :

- Directeur de thèse (avec HDR): Pierre Couteron (AMAP, IRD, Montpellier).
- Co-encadrant(s) de thèse : Ghislain Vieilledent (AMAP, Cirad, Montpellier).

Composition du comité de suivi individuel : Absence de la direction et de l'encadrement Pour chaque membre, préciser impérativement : nom, fonction, organisme et ville

- Membre/s extérieur/s à la thèse et à l'ED (**non impliqué dans le projet de thèse** et notamment **pas l'industriel financeur** dans le cas d'une CIFRE) : Frédéric Gosselin (Inrae, Nogent-sur-Vernisson).
- Représentant de la direction de l'UMR/UR d'accueil (**nommé par la direction de l'UMR, non impliqué dans la thèse**) : Raphaël Pélissier (AMAP, IRD, Montpellier).
- Référent de l'ED GAIA (Nom et Unité de recherche, doit être dans GAIA mais **extérieur à l'Unité de rattachement**) : Xavier Morin (CEFE, CNRS, Montpellier)

Seconde partie : A remplir par le docteurant

Cette partie est contrôlée par le référent lors du comité de thèse

Tous ces points sont à aborder absolument lors du comité de thèse, et doivent faire chacun l'objet d'un commentaire de longueur adaptée.

**A. Résumé des objectifs scientifiques et stratégie de recherche, (1/2 page)
(Un rapport plus détaillé est à mettre en annexe à la fin du document)**

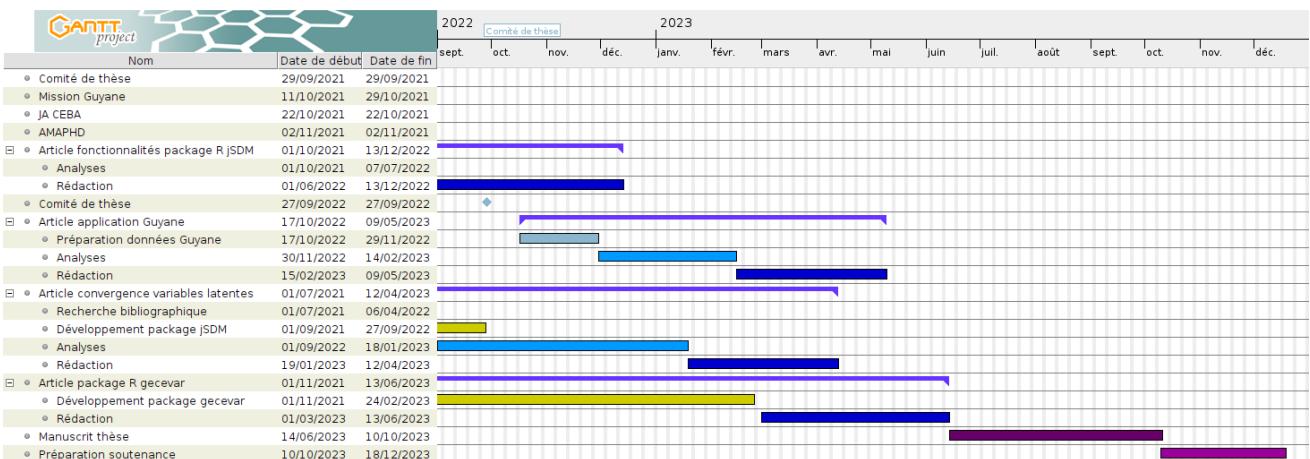
L'objectif de ce projet de thèse est (i) d'expliquer les différences de vulnérabilité des espèces aux changements climatiques d'un point de vue fonctionnel et phylogénétique et (ii) de prédire les changements d'aires de distribution des espèces et de composition des communautés d'arbres en Guyane sous l'effet du changement climatique.

Ce projet s'appuiera sur le développement du package R jSDM (Vieilledent et al. 2019) afin de lever les limitations propres aux SDMs et aux JSDMs actuels. Le package jSDM permet l'ajustement de modèles joints de distribution des espèces prenant en compte les interactions entre espèces (<https://ecology.ghislainv.fr/jSDM>). Le package inclut un échantillonneur de Gibbs avec un tirage en bloc des paramètres. Le code, que j'ai développé en grande partie, est écrit en C++ et fait appel à des librairies C/C++ dédiées aux tirages aléatoires et aux calculs matriciels (GSL et Armadillo). Le code est optimisé et permet l'estimation de paramètres pour de gros jeux de données en un temps limité.

Dans le cadre de la thèse, des fonctions supplémentaires sont en cours de développement afin (i) d'intégrer la phylogénie et les traits fonctionnels comme facteur explicatifs des différences d'occurrence entre espèces et (ii) de pouvoir extrapoler spatialement les occurrences entre sites afin d'obtenir des cartes prédictives. En utilisant des jeux de données réels existant dans la littérature (Wilkinson et al. 2018; Pichler & Hartig 2020) et des jeux de données virtuels, on comparera les temps de calcul et les performances prédictives des différentes librairies et méthodes statistiques.

En combinant les données d'inventaires forestiers disponibles en Guyane, les données sur les traits fonctionnels, la phylogénie et les données environnementales, on cherchera à ajuster des modèles joints de distribution des espèces d'arbres pour la Guyane et à obtenir des cartes de communautés dans le présent et dans le futur sous l'effet des changements climatiques.

B. Points restant à développer, calendrier, moyens pour réaliser la thèse (1/2 page).



C. Environnement de la thèse/ Autres activités du doctorant

- Insertion dans l'équipe d'accueil :

Durant ma première année de thèse j'ai été amenée à collaborer avec d'autres doctorants comme Benjamin Deneu qui travaille sur l'interprétabilité des modèles de distribution des communautés d'espèces végétales appris par deep learning (CNN) avec qui nous comptons confronter nos approches ou Camille Girard-Tercieux, également encadrée par Ghislain Vieilledent, qui travaille sur le rôle de la variabilité intraspécifique sur la coexistence des espèces dans les forêts tropicales, avec qui nous travaillons à inférer une matrice de communauté. J'ai aussi apporté mon aide à Lily Cheng, une stagiaire en M1 BioGET, pour préparer les données nécessaires et implémenter une méthode d'estimation de la répartition écologique d'une espèce rare de poireau sauvage de la Séranne.

De plus j'ai été membre du comité d'organisation d'AMAPhD 2021, la journée des doctorants et non permanents de l'UMR AMAP. J'y ai déjà rapidement présenté mon travail en 2020 et 2021 et je vais faire une présentation plus détaillée de mon projet de thèse cette année.

Pour finir, durant les derniers mois, j'ai participé activement à l'encadrement de Pierre Guillaumont qui a effectué son stage de M2 MIND au sein de l'UMR AMAP, sur le sujet : "Classification des communautés d'arbres en Nouvelle-Calédonie à l'aide de modèle joint de distribution des espèces". Dans le cadre de son stage, Pierre a participé au développement du package R guyaclim, que j'avais commencé à implémenté avec Ghislain Vieilledent. Ce package doit permettre, à terme, d'extraire un ensemble de données climatiques et environnementales, sur n'importe quelle zone géographique, ainsi que de les re-projeter dans le système de coordonnées souhaité, et de les ré-échantillonner à la résolution 1km. Il a pour vocation de faciliter la récupération des données utilisées comme variables explicatives pour l'ajustement de modèles de distribution des espèces. De plus Pierre a utilisé le package R jSDM, que j'ai développé, afin d'ajuster un modèle joint de distribution des espèces en Nouvelle Calédonie, à partir des données climatiques et environnementales obtenues à l'aide du package guyaclim, ainsi que de données d'inventaire forestier répertoriant un très grand nombre d'espèces sur de nombreux sites d'inventaires.

- Modules de formation complémentaire réalisés et à prévoir le cas échéant (préciser le nombre d'heures ; séparer les formations d'ouverture des formations scientifiques) :

Intitulé formation	Nombre d'heures	Validée
Formations scientifique		
L'Intelligence Artificielle... avec intelligence !	15	OUI
Recherche reproductible : principes méthodologiques pour une science transparente	24	OUI
Intégrité scientifique dans les métiers de la recherche	15	OUI
Rédiger et publier un article scientifique	20	OUI
Formations d'ouverture		
Les bases de la prise de parole en public	15	OUI

J'ai suivi 89h de formation sur les 90h requises.

- Participations à des congrès nationaux/internationaux réalisées et à prévoir le cas échéant (préciser si elles ont donné lieu à une présentation orale ou à un poster) :

J'ai participé en octobre 2021 aux journées annuelles CEBA en distanciel, lors desquelles je ferai un présentation orale sous le format ma thèse en 180s.

J'ai également présenté un poster lors de la conférence internationale, European conference of tropical ecology, en juin 2022.

- Enseignements réalisés par le doctorant (cadre (MCE ou vacations) /nombre d'heures) (le cas échéant) :

- Publications (distinguer les ACL-Articles à Comité de Lecture) (état d'avancement, date de soumission effectuée ou prévue) :

Date prévue de soumission	Mots-Clés	Titre
Publications en tant que premier auteur		
Décembre 2022	Contenu et fonctionnalités du package R jSDM : modèles joints, modèles d'interaction traits-environnement, données de présence/absence ou d'abondance, modèles hiérarchiques bayésiens, échantillonneur de Gibbs, chaîne de Markov de Monte Carlo (MCMC) Comparaison : jSDM, boral, BayesComm, s-jSDM, HMSC, gjam, greta, gllvm, STAN	Main features and functionality of the R package jSDM and comparison with other approaches for the estimation of joint species distribution models.
Mai 2023	Carte des communautés d'arbres en Guyane se basant sur un modèle joint de distribution des espèces intégrant les interactions traits-environnement. Identification des espèces vulnérables aux changements climatiques et des zones de refuge de la biodiversité.	Identification of tree species vulnerable to climate change and of biodiversity refuge areas in French Guiana through the use of a joint species distribution model integrating trait-environment interactions.
Juillet 2023	R package gecevar ("GEtting Climatic and Environmental VARiables from open and free online datasets for a specific region") pour extraire un ensemble de données climatiques et environnementales, sur n'importe quelle zone géographique ainsi que de les re-projeter dans le système de coordonnées souhaité, et de les ré-échantillonner à la résolution 1km.	New R package to facilitate and automate the extraction and formatting of environmental and climatic data required for the adjustment of JSMDs.
Publication en tant que co-auteur		
Avril 2023	Ajout d'une variable dans le modèle Bayésien qui va estimer le nombre d'axes latents à considérer, évaluation de la convergence des paramètres associés aux axes latents, assurer leur identifiabilité en imposant les contraintes appropriées, réordonnancement des espèces	Diagnosis of latent axes and estimation of their number in generalized linear latent variable models for joint species distribution modeling

(Rappel : 1 publication acceptée est requise pour être autorisé à soutenir)

Troisième partie : A remplir par le référent

A. Entretiens menés par le référent de l'ED (à réaliser absolument) : si souci majeur ou mineur, faire impérativement remonter le souci au responsable de la filière concernée.

- Entretien avec la direction de thèse (en l'absence du doctorant) :
Bref résumé (obligatoire) :

- Entretien avec le doctorant (sans l'équipe de direction) :
Bref résumé (obligatoire) :

B. Le doctorant a été formé à l'éthique et à l'intégrité scientifique : Formation obligatoire pour être autorisé à soutenir ; arrêté de mai 2016.

OUI / NON (rayer la mention inutile)

C . Convention individuelle de formation. Contrôler et valider la convention avec la direction.
Si nécessaire la faire compléter et la déposer de nouveau sous ADUM

Quatrième partie : A remplir par le référent avec l'approbation de l'ensemble des membres du comité.

Dépôt du rapport définitif par le référent *via ADUM*.

1 semaine après la tenue du comité maximum

CONCLUSION et AVIS DU COMITE POUR UNE REINSCRIPTION EN TROISIEME ANNEE

Bilan du comité de suivi, préconisations (1/2 page maximum)

Conclusion :

**Avis : Favorable – Défavorable – Réserve
(rayer impérativement les mentions inutiles)**

Signature du référent*:

* signature électronique possible

Visa du directeur de thèse* montrant qu'il a pris connaissance du rapport de CSI et commentaire si nécessaire :

* signature électronique possible

Si changement de référent (si changement d'ED, d'UMR, départ à la retraite, mutations, ...), ne pas oublier de prévenir l'administration GAIA pour créer un lien informatique entre le doctorant et le référent pour que ce dernier puisse déposer le compte rendu via ADUM.

ANNEXE : rapport détaillé sur l'avancement des travaux de thèse (10 pages max)

Rapport détaillé

Comité de suivi de thèse : 2ème année

Jeanne Clément

09 novembre, 2022

Table des matières

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
2.2 Méthodes d'inférence bayésienne selon la fonction de lien choisie	4
2.3 Comparaison des résultats obtenus avec ceux des packages <i>boral</i> et <i>Hmsc</i>	5
2.4 Comparaison de la pertinence des résultats obtenus et des temps de calcul nécessaires avec chacun des packages	6
2.5 Comparaison de l'efficacité d'échantillonnage	7
3 Amélioration de la convergence pour l'estimation des axes latents	9
4 Développement du package <i>gecevar</i> pour faciliter la récupération des données explicatives nécessaires à l'ajustement des JSDMs	11
4.1 Objectifs	11
4.2 Fonctionnalités	11
5 Obtention de cartes de communauté à l'échelle du territoire à partir de données d'inventaire forestier	14
5.1 Éstimation de la biodiversité à Madagascar	15
5.2 Classification des communautés d'arbres en Nouvelle-Calédonie	22
5.3 Obtention de cartes de communauté en Guyane française	23
6 Perspectives de développement du package jSDM	24
6.1 Fonctionnalités en cours de développement	24
6.2 Perspectives de développement	25
6.3 Estimation du nombre d'axes latents à prendre en compte	25
6.4 Ajuster des JSDMs spatialement explicites	25
6.5 Ajuster des JSDMs à partir de données de présence seule	26
Bibliographie	27

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

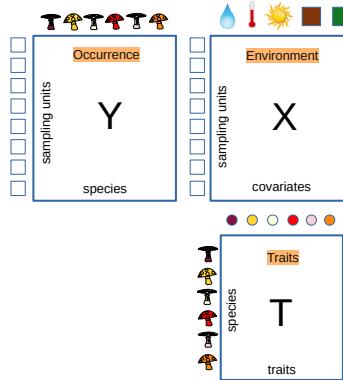
On s'est inspiré des articles Warton et al. (2015) et Ovaskainen et al. (2017) pour développer les approches hiérarchiques utilisées à la spécification des modèle joint de distribution des espèces dans le package jSDM.

2.1.1 Données utilisées

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 $\forall j, t_{j0} = 1$.

Figure 1 – Les données de présence/absence ou d'abondance des espèces (Y), les données bio-climatiques sur chaque site d'inventaire (X) et les mesures de traits pour pour chaque espèce considérées (T).



2.1.2 Modèle linéaire mixte généralisé multivarié (GLMM)

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

$$\begin{aligned} g(\theta_{ij}) &= \alpha_i + X_i \beta_j + u_{ij}, \\ y_{ij} &\sim \text{Binomial}(n_i, \theta_{ij}), \text{ pour des données de présence/absence} \\ &\text{ou } y_{ij} \sim \text{Poisson}(\theta_{ij}), \text{ pour des données d'abondances} \end{aligned}$$

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

$$\alpha_i \sim \mathcal{N}(0, V_\alpha) \text{ iid et indépendant de } u_i.$$

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 et $V_\alpha \sim \mathcal{IG}(\text{shape}, \text{rate})$ dans le cas où l'effet site est aléatoire.
 - $\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})'$ 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 : $\beta_j \sim \mathcal{N}_{p+1}(\mu_{\beta_jk}, V_\beta)$, 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 .

2.1.3 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 $\Lambda := (\lambda_{jl})_{j=1, \dots, J}^{l=1, \dots, q}$.

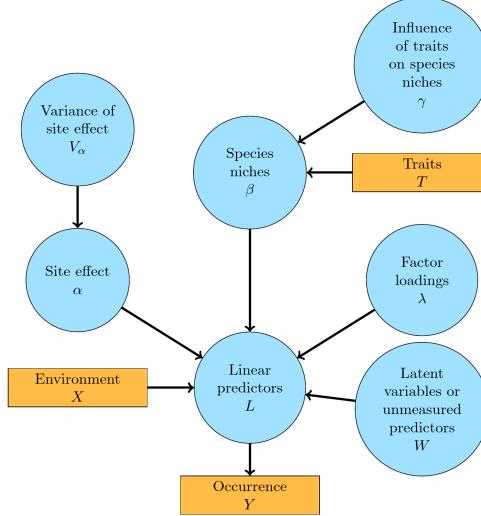
On préférera ce dernier modèle pour estimer la matrice de corrélation entre les espèces, en effet il comporte potentiellement beaucoup moins de paramètres que le GLMM précédent car Λ a autant de colonne qu'il y a de variables latentes (q) tandis que Σ présente autant de colonnes de paramètres qu'il y 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. En effet la distribution *a priori* des facteurs latents est définie comme suit :

$$\lambda_{jl} \sim \begin{cases} \mathcal{N}(\mu_\lambda, V_\lambda) & \text{si } l < j \\ \mathcal{N}(\mu_\lambda, V_\lambda) \text{ tronquée à gauche par 0} & \text{si } l = j \\ \delta(\lambda_{jl}) \text{ (Dirac distribution)} & \text{si } l > j \end{cases}$$

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 2 – 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](#).

Afin d'utiliser une méthode d'inférence bayesienne on a déterminé une distribution *a priori* pour chacun des paramètres du modèle, dont les hyperparamètres par défaut sont spécifiés dans le package.

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 $i = 1, \dots, m$.

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 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 2.2.2.1 (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*.

2.2.3 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 $y_{ij} | \theta_{ij} \sim \text{Binomial}(n_i, \theta_{ij})$, 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 $\log(\theta_{ij}) = \alpha_i + X_i\beta_j + W_i\lambda_j$.

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

D'une part, nous avons utilisé le package R **Hmsc** 3.0-11, pour "Hierarchical Modelling of Species Communities", qui est une approche basée sur un modèle pour analyser les données écologiques des communautés ((Ovaskainen et al. 2017)). D'autre part, nous avons utilisé le package R **boral** 2.0 (Hui 2016), pour "Bayesian Ordination and Regression Analysis", qui fonctionne avec **JAGS** ("Just Another Gibbs Sampler") un programme de simulation de modèles bayésiens hiérarchiques utilisant des méthodes MCMC, implémenté en C++. Ces packages et le package **jSDM** 0.2.1 permettent d'ajuster différents JSDMs, nous avons donc pu comparer la pertinence des résultats obtenus par chacun d'eux, leurs performances de calcul et l'efficacité des échantillonneurs sur différents jeux de données. Avec chacun de ces trois packages, nous avons effectué des régressions binomiales probit à partir d'un jeu de données simulé ainsi que quatre jeux de données de présence-absence, utilisés dans une étude comparative de JSDM récente par Wilkinson et al. (2019). Nous avons également effectué des régressions log-linéaires de Poisson à partir de deux jeux de données d'abondance (Choler 2005 ; Borcard and Legendre 1994). L'ajustement des modèles a été effectué sur le même ordinateur de bureau avec les spécifications suivantes : Ubuntu 20.04.4 OS, i5-1145G7 CPU @ 2.60GHz × 8, 15.4GB RAM) afin de comparer les performances de calcul.

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

Table 1 – Dimensions des ensembles de données utilisés (*n.sites* et *n.espèces*), nombre d’observations (*n.obs=n.site times n.espèces*), nombre de covariables, traits spécifiques et axes latents considérés (*n.col.X*, *n.traits* et *n.latent*), nombre de paramètres à estimer (*n.param*) en effectuant *n.mcmc* itérations, le temps de calcul en secondes nécessaire pour ajuster le modèle à partir de chacun des ensembles de données et la déviance calculée à partir des paramètres estimés avec chacun des packages comme suit : $D = -2 \sum_{i=1}^I \sum_{j=1}^J \log(\mathbb{P}(y_{ij}|\beta_j, \lambda_j, W_i))$. L’erreur quadratique moyenne (RMSE) est calculée pour les probabilités de présence θ_{ij} sur l’ensemble des données simulées. La sensibilité, la spécificité et la ”True Skill Statistic” (*TSS = sensibilit + spcificit - 1*) ont été calculées pour chaque ensemble de données, en tenant compte de la matrice de présence-absence prédite : \hat{Y} telle que $\hat{y}_{ij} = 1$ si la probabilité d’occurrence estimée correspondante $\hat{\theta}_{ij} \geq \tau_i$ et $\hat{y}_{ij} = 0$ sinon, où τ_i représente le seuil de présence pour le site *i*, défini par la valeur de probabilité d’occurrence estimée la plus faible parmi les n_i probabilités de présence d’espèces les plus élevées, où n_i correspond au nombre d’espèces observées sur le site *i*.

	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
							

On constate que jSDM est **1.7 à 440** fois plus rapide que boral (JAGS) et **1.5 à 7.6** 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 le temps de calcul nécessaire à `jSDM` pour ajuster un modèle à partir des données d'abondance est plus important que pour les jeux de données de présence-absence, car la méthode d'inférence dépend de la fonction de lien choisie. En effet l'utilisation de l'algorithme de Metropolis Hastings au sein de l'échantillonneur de Gibbs dans le cas d'une fonction de lien log pour des données d'abondance, est moins efficace que les formules des priors conjugués utilisées avec une fonction de lien probit pour des données de présence absence.

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

Les spécificités, sensitivités et TSS obtenus avec `jSDM` sont aussi proches de 1 que ceux calculés avec `boral` et `Hmsc` ce qui suggère que la pertinence des résultats de `jSDM` est équivalente voire supérieure à celle des résultats obtenus avec les autres packages.

Enfin les RMSE associés à `jSDM` sont équivalents à ceux de `boral` et `Hmsc` ce qui indique que les résultats obtenus avec `jSDM` sur les jeux de données simulés sont aussi proches de ceux attendus que les paramètres estimés avec `boral` et `Hmsc`.

2.5 Comparaison de l'efficacité d'échantillonnage

D'autre part on compare l'efficacité des échantillonneurs utilisés dans les différents packages. Pour ce faire on calcule le nombre de valeurs efficaces échantillonées par seconde par chacun des packages et pour chacun des paramètres du modèle, afin d'évaluer leur efficacité d'échantillonnage pas seulement en fonction du temps de calcul nécessaire à l'ajustement des modèles mais aussi en prenant en compte l'auto-corrélation dans les chaînes MCMCs de paramètres renvoyées.

La taille d'échantillon effective (ESS) est calculée pour chaque paramètre à l'aide de la fonction '`effectiveSize`' du package `coda`.

Table 2 – Le nombre de valeurs efficaces échantillonnées par seconde, par chaque package. Nous supposons que le temps d'échantillonnage (`T.sample`) correspond à la moitié du temps de calcul (voir Tableau 1) car le même nombre d'itérations d'échantillonnage (`n.sample`) et de burn-in (`n.burnin`) ont été effectuées pour ajuster les modèles. Pour réduire l'auto-corrélation des MCMCs, seule une itération d'échantillonnage sur `n.thin` a été retenue et un échantillon de taille : `sample.size`, est retourné pour chaque paramètre. Pour obtenir le nombre de valeurs efficaces échantillonnées par seconde, au total ou pour l'intercept espèce β_0 , l'effet espèce β , les charges factorielles λ et les variables latentes W , la somme des ESS calculées pour tous les paramètres, ou la moyenne des ESS pour chaque type de paramètres, est divisée par le temps d'échantillonnage en secondes.

	Simulated	Mosquitos	Eucalypts	Frogs	Fungi	Aravo	Mites
<code>n.obs</code>	30000	2672	5460	936	4818	4875	2100
<code>n.param</code>	1099	589	1029	261	1040	480	559
<code>n.burnin</code>	10000	10000	10000	10000	10000	10000	10000
<code>n.sample</code>	10000	10000	10000	10000	10000	10000	10000
<code>n.thin</code>	10	10	10	10	10	10	10
<code>sample.size</code>	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

On constate que le nombre total de valeurs efficaces échantillonnées par seconde par **jSDM** est supérieur à celui de **Hmsc** pour chacun des jeux de données, l'échantillonneur de **jSDM** est donc plus efficace. Néanmoins si on s'intéresse uniquement aux facteurs latents λ , les nombres de valeurs efficaces échantillonnées par seconde par **Hmsc** et plus important que ceux de **jSDM**, ce qui peut s'expliquer par les méthodes d'inférences qui diffèrent entre les deux packages pour ces paramètres.

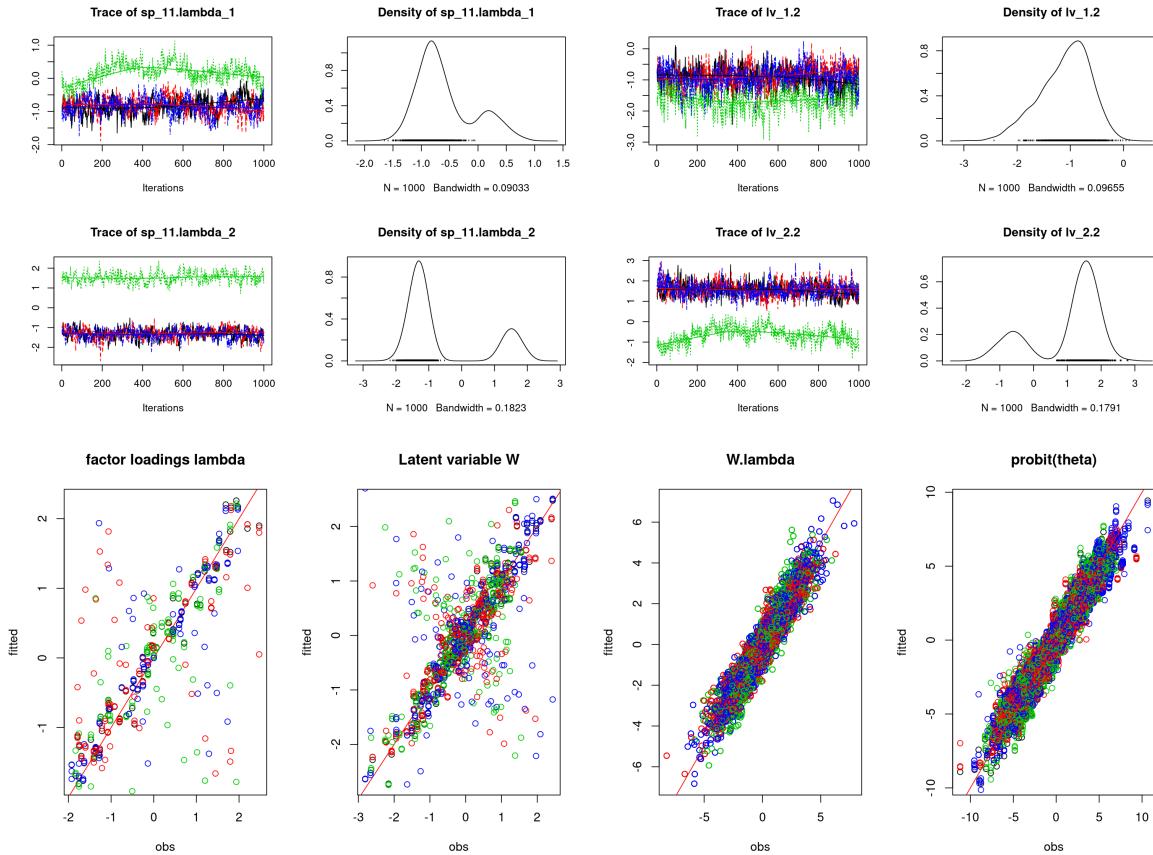
De plus **jSDM** présente un nombre total de valeurs efficaces échantillonnées par seconde plus élevé que **boral** concernant les jeux de données de présence-absence. En ce qui concerne les jeux de données d'abondance, l'échantillonneur de **boral** est légèrement plus efficace que celui de **jSDM**, mais c'est le cas car il s'agit de petits jeux de données, sur des jeux de données d'abondance plus conséquents, **jSDM** deviendrait vite plus performant que **boral**.

3 Amélioration de la convergence pour l'estimation des axes latents

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.

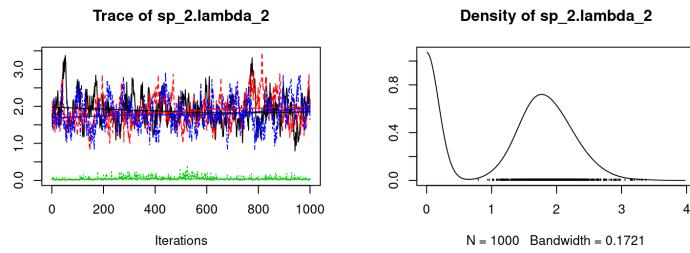
- Problème “symétrique” de convergence des facteurs latents et des axes latents

Figure 3 – Illustration des problèmes de convergence “symétrique” 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 4 – 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 contraint 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.

- Évaluer la convergence des paramètres associés aux axes latents et diagnostiquer l'espèce (j) qui se structure le plus clairement sur chaque axe (l) en essayant différents critères :
 - Maximise les moyennes de la valeur absolue des λ calculées sur chaque chaîne MCMC : $|\bar{\lambda}_{jl}|$.
 - Maximise les coefficients de variation calculées sur chaque chaîne MCMC : $\frac{|\bar{\lambda}_{jl}|}{\sqrt{\text{Var}(|\lambda_{jl}|)}}$.
 - Maximise les diagnostiques de Gelman–Rubin (\hat{R}_{jl}) calculé entre toutes les chaînes MCMC obtenues pour chaque λ , qui évalue la convergence en comparant les variances intra-chaînes et inter-chaînes estimées pour chaque paramètre.
- En déduire sur quelles espèces imposer les contraintes de positivité sur la diagonale et réordonner les espèces avant d'ajuster un nouveau JSDM.

On a intégré cette méthode dans package `jSDM` en implémentant la fonction `jSDM_binomial_probit_sp_constrained`, pour laquelle on a choisi de considérer les diagnostiques de Gelman–Rubin (\hat{R}) comme critère pour le choix de l'espèce de référence pour chaque axe et dont un exemple d'utilisation est présenté dans la vignette : https://ecology.ghislainv.fr/jSDM/articles/jSDM_binomial_probit_sp_constrained.html.

4 Développement du package `gecevar` pour faciliter la récupération des données explicatives nécessaires à l'ajustement des JSDMs

4.1 Objectifs

On a commencé le développement du package R, nommé `guyaclim` initialement, avec Ghislain Vieilledent, lors de notre mission en Guyane en octobre 2021 durant laquelle on a pu échanger avec les personnes travaillant sur place sur les bases de données disponibles sur la Guyane française et la manière d'accéder à ces données. On a décidé d'implémenter ce package, afin de faciliter et automatiser l'extraction et le formatage des données environnementales et climatiques nécessaires à l'ajustement de JSDMs en Guyane française. Il est amené à être renommé (`gecevar` pour “GEtting Climatic and Environmental VARiables from open and free online datasets for a specific region” par exemple), car ce package doit permettre, à terme, d'extraire un ensemble de données climatiques et environnementales, sur n'importe quelle zone géographique, pas seulement la Guyane, ainsi que de les re-projeter dans le système de coordonnées souhaité (EPSG), et de les ré-échantillonner à la résolution voulue (250m, 500m, 1km, 2km ou 5km). Il a pour vocation de faciliter la récupération des données utilisées comme variables explicatives pour l'ajustement de modèles de distribution des espèces.

De plus, j'ai participé activement à l'encadrement de Pierre Guillaumont qui a effectué son stage de M2 MIND au sein de l'UMR AMAP, sur le sujet : “Classification des communautés d'arbres en Nouvelle-Calédonie à l'aide de modèle joint de distribution des espèces”. Dans le cadre de son stage, Pierre a participé au développement du package R `gecevar`, en généralisant les codes qui avaient été implémentés pour la Guyane à l'extraction de données en Nouvelle Calédonie et en important des variables supplémentaires.

Pour finaliser le package, il me reste donc juste à organiser les différents scripts déjà implémentés pour la Guyane et la Nouvelle Calédonie sous forme de fonction et à structurer et à documenter le package.

4.2 Fonctionnalités

4.2.1 Spécification de la zone géographique d'intérêt

La zone géographique sur laquelle les données seront extraites peut être spécifiée par l'utilisateur au moyen d'un shapefile indiquant les frontières du territoire, de l'étendue de la zone définie selon le système de coordonnées spécifié pour reprojeter les données extraites, ou du code ISO 3166 correspondant s'il s'agit d'un pays.

4.2.2 Extractions de données climatiques actuelles et futures

Les données climatiques actuelles sont extraites du site chelsa-climate.org (moyennes sur la période de 1981 à 2010). Le téléchargement des données se fait uniquement à l'échelle mondiale, ce qui est un avantage pour généraliser l'acquisition à n'importe quelle zone géographique mais qui nécessite un temps de téléchargement et un espace de stockage relativement importants.

En ce qui concerne les données climatiques futures, on laisse à l'utilisateur du package la possibilité de spécifier pour quels modèles climatiques globaux (GCMs) et pour quels scénarios socio-économiques partagés (SSPs pour “Shared Socio-economic Pathways”), il souhaite importer ces données. Le choix de l'utilisateur est cependant limité par la disponibilité des données climatiques futures estimées selon le GCM et le SSP choisis, sur le site chelsa-climate.org. Les scénarios CMIP6 pour lesquels les données sont disponibles sur le sites source correspondent aux différentes combinaisons possibles de trois SSPs (SSP1-RCP2.6, SSP3-RCP7, SSP5-RCP8.5) et de cinq GCMs proposés par les laboratoires ci dessous :

- GFDL-ESM4, Princeton University Forrestal Campus, New Jersey, Etats Unis d'Amérique
- IPSL-CM6A-LR, Institut Pierre Simon Laplace, Paris, France
- MPI-ESM1-2-HR, Max Planck Institute, Göttingen, Allemagne
- MRI-ESM2-0, Meteorological Research Institute, Nagamine, Japon
- UKESM1-0-LL, Met Office Hadley Centre, Exeter, Royaume Uni

De plus les valeurs mensuelles pour les données climatiques futures, disponibles sur chelsa-climate.org, sont des moyennes sur des périodes de 30 ans (2011-2040, 241-2070, 2071-2100).

Table 3 – Données climatiques téléchargées par la fonction `get_chelsa_current` and `get_chelsa_future`.

Variable	Unité	Source
Températures mensuelles minimales	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Températures mensuelles maximales	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Températures mensuelles moyennes	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Précipitations mensuelles	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Couverture nuageuse	%	chelsa-climate.org
Déficit hydrique climatique (Thorntwaite)	kg.m^{-2}	calculé à partir des données de chelsa-climate.org
Evapotranspiration potentielle (Thorntwaite)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	calculée à partir des données de chelsa-climate.org
Nombre de mois secs (Thorntwaite)	mois	calculés à partir des données de chelsa-climate.org
Déficit hydrique climatique (Penman-Monteith)	kg.m^{-2}	chelsa-climate.org
Evapotranspiration potentielle (Penman-Monteith)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Nombre de mois secs (Penman-Monteith)	mois	calculés à partir des données de chelsa-climate.org
Température moyenne annuelle	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température moyenne annuelle (bio1)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Amplitude des températures diurnes (bio2)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Isothermalité (bio3=bio2/bio7)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Saisonnalité des températures (bio4)	$^{\circ}\text{C}/10$	chelsa-climate.org
Température maximale du mois le plus chaud (bio5)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température minimale du mois le plus froid (bio6)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Amplitude annuelle des températures (bio7=bio5-bio6)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température moyenne du trimestre le plus humide (bio8)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température moyenne du trimestre le plus sec (bio9)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température moyenne du trimestre le plus chaud (bio10)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Température moyenne du trimestre le plus froid (bio11)	$^{\circ}\text{C} \times 10$	chelsa-climate.org
Cumul annuel des précipitations (bio12)	$\text{kg.m}^{-2}.\text{an}^{-1}$	chelsa-climate.org
Cumul des précipitations du mois le plus humide (bio13)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Cumul des précipitations du mois le plus sec (bio14)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Saisonnalité des précipitations (bio15)	kg.m^{-2}	chelsa-climate.org
Précipitations du trimestre le plus humide (bio16)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Précipitations du trimestre le plus sec (bio17)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Précipitations du trimestre le plus chaud (bio18)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org
Précipitations du trimestre le plus froid (bio19)	$\text{kg.m}^{-2}.\text{mois}^{-1}$	chelsa-climate.org

4.2.3 Extraction de données environnementales

Les données environnementales proviennent de sources multiples. Il s'est avéré plus compliqué pour les données environnementales que pour les données climatiques de trouver des sites disposant de données précises et disponibles pour un maximum de localisations.

Table 4 – Données environnementales téléchargées par la fonction `get_env_variables`.

Variable	Unité	Source
Couvert forestier	%	calculé à partir des données de https://forestatrisk.cirad.fr
Altitude	m	https://srtm.cgiar.org/
Pente	degrés	calculée à partir des données de https://srtm.cgiar.org/
Orientation du sol (aspect)	degrés	calculé à partir des données de https://srtm.cgiar.org/
Rugosité	m	calculée à partir des données de https://srtm.cgiar.org/
Ensoleillement	$Wh.m^{-2}.jour^{-1}$	calculé à partir des données https://srtm.cgiar.org/
Distance à la forêt	m	calculée à partir des données de https://www.openstreetmap.org/
Distance à la mer	m	calculée à partir des données de https://www.openstreetmap.org/
Distance aux routes	m	calculée à partir des données de https://www.openstreetmap.org/
Distance aux villes/villages	m	calculée à partir des données de https://www.openstreetmap.org/
Distance aux points d'eau	m	calculée à partir des données de https://www.openstreetmap.org/
Zones protégées (WDPA)	catégories	https://www.protectedplanet.net/en
Densité de population	habitants/km ²	https://data.worldpop.org/GIS/Population
Données pédologiques	catégories	https://soilgrids.org
(type de sol et leur propriétés chimiques et physique)		

4.2.4 Efficacité de l'extraction et du formatage des données

Pour implémenter le package R `gecevar`, on a utilisé les packages `stars` et `terra` permettent de réaliser des opérations basiques sur les fichiers, comme redimensionner ou afficher le fichier dans R. Tandis que les package `rgdal` ou `rgrass7` permettent de réaliser des opérations plus complexes comme, re-projeter les données spatiales dans le système de coordonnées souhaité, les ré-échantillonner à la résolution voulue et calculer la pente ou les radiations solaires à partir de données d'altitude. Ces packages sont des interfaces entre R et C pour `rgrass7` et Python pour `rgdal`. Ils sont attractifs par leurs fonctionnalités variées mais aussi leur vitesse de calcul induite par les langages utilisés.

5 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 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 dans un second temps.

5.1 Éstimation de la biodiversité à Madagascar

5.1.1 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 $^{\circ}\text{C} \times 10$.
- 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 : $\text{cwd} = \sum_{m=1}^{12} \min(0, \text{prec}_m - \text{pet}_m)$.

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.

5.1.2 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`.

Table 5 – 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.

<code>n.sites</code>	<code>n.species</code>	<code>n.latent</code>	<code>n.X.coefs</code>	<code>n.param</code>	<code>n.mcmc</code>	Temps de calcul (heures)	Déviance
751	483	2	483	236991	80000	1.8	32916.8

5.1.3 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 :

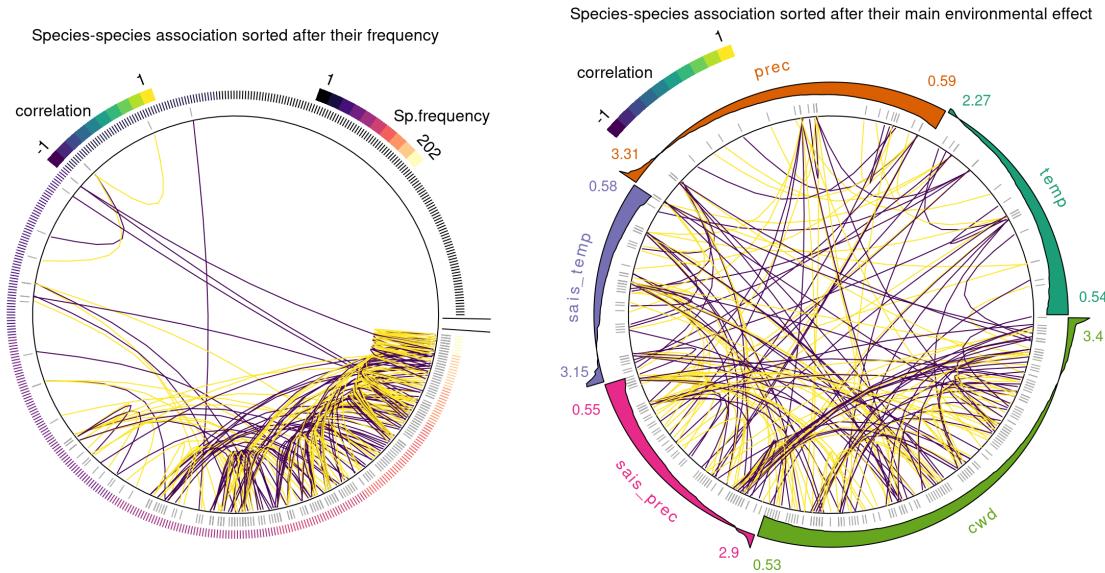
$$\Sigma_{gj} = \begin{cases} \lambda'_g \cdot \lambda_j & \text{if } g \neq j \\ \lambda'_g \cdot \lambda_j^T & \text{if } g = j \end{cases}$$

, on calcule ensuite les corrélation à partir des covariances :

$$R_{g,j} = \frac{\Sigma_{gj}}{\sqrt{\Sigma_{gg}\Sigma_{jj}}}$$

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 choisi donc de représenter les associations prépondérantes à l'aide de la fonction `plot_associations`.

Figure 5 – 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. On constate sur la figure de gauche que les corrélations résiduelles prépondérantes sont plus nombreuses entre les espèces les plus abondantes, c'est à dire celles observées sur un grand nombre de sites. De plus on peut voir sur la de droite que les espèces dont la distribution est influencée principalement par le déficit hydrique (cwb) et celles dont l'effet environnemental principal et la saisonnalité des précipitations (sais_prec) sont fortement corrélées.

5.1.4 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 estimée (Scherrer, Mod, and Guisan 2020).

Figure 6 – Représentation spatiale de la richesse spécifique estimée pour chaque site par $\widehat{R}_i = \sum_{j=1}^{483} \widehat{\theta}_{ij}$. **Figure 7 – Représentation spatiale de la richesse spécifique observée** pour chaque site calculée par $R_i = \sum_{j=1}^{483} y_{ij}$

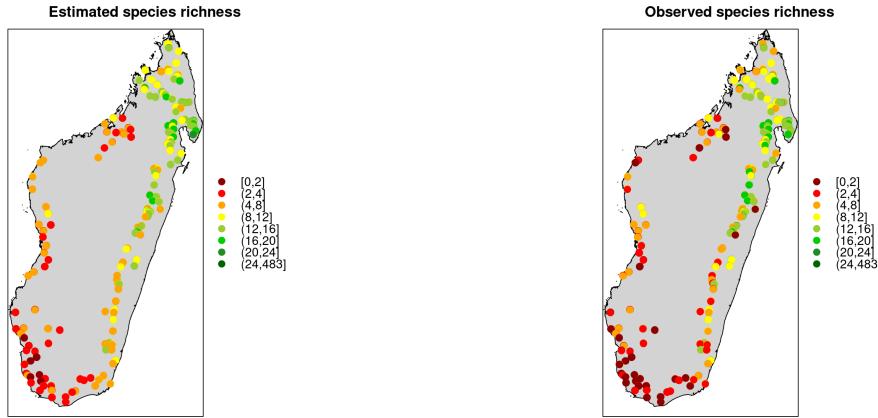
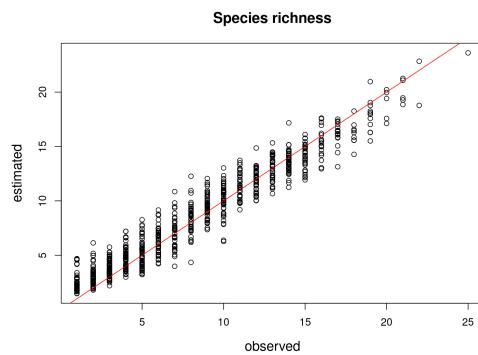


Figure 8 – Représentation de la richesse spécifique estimée en fonction de celle observée



On constate que la richesse spécifique observée sur les sites d'inventaire correspond assez bien à celle estimée. De plus on peut voir que la richesse spécifique observée est plus importante dans la forêt humide du nord-est de l'île, notamment dans la péninsule de Masoala.

5.1.5 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).

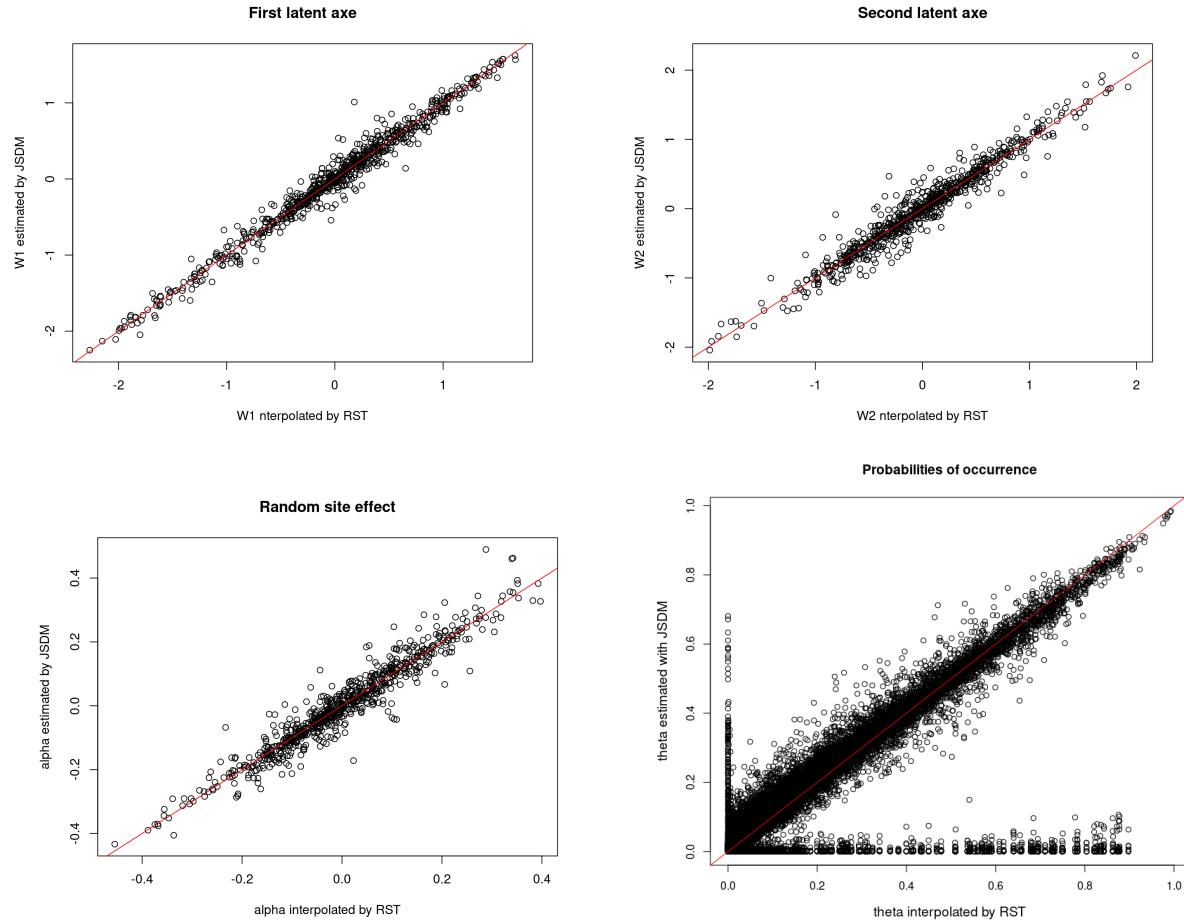
D'une part, nous avons utilisé la méthode déterministe de pondération par distance inverse, nommée **IDW** pour l'interpolation multivariée à partir de l'ensemble connu de points dispersés. Les valeurs attribuées à des points inconnus sont calculées avec une moyenne pondérée des valeurs disponibles aux 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 9 – Représentation des paramètres estimés par le JSDM sur les sites d'inventaire en fonction de ceux interpolés par RST.



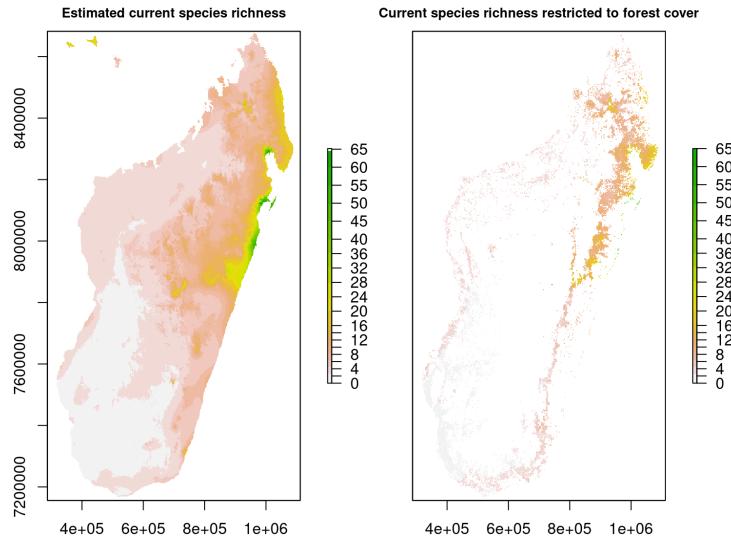
Nous avons tout de même poursuivi la démarche avec l'intention d'améliorer la méthode d'interpolation plus tard. Nous avons utilisé les paramètres interpolés par RST pour calculer les probabilités de présence sur l'île de chacune des espèces en fonction des variables climatiques présentes définies précédemment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

5.1.6 Estimation de la richesse spécifique à Madagascar

J'ai représenté les cartes suivantes de biodiversité α et β afin d'illustrer la méthode mise en oeuvre pour les obtenir mais elles peuvent présenter des incohérences car elles sont basées sur les résultats de l'interpolation par RST qui nécessite d'être améliorée.

Nous avons additionné les probabilités de présence interpolées pour chacune des espèces afin d'estimer la richesse spécifique. 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 10 – Richesse spécifique interpolée sur l’ensemble de l’île et restreinte au couvert forestier



Nous pouvons voir sur la figure 10 que la richesse spécifique, calculée à partir des probabilités d’occurrence interpolées, est plus importante dans la forêt humide du nord-est de l’île, en particulier dans la péninsule de Masoala, comme attendu étant donné la richesse spécifique observée sur les sites d’inventaire. Cela peut s’expliquer par le fait que cette zone est située près de l’équateur et reçoit le plus haut niveau de précipitations à Madagascar. En conséquence, cette zone bénéficie de niveaux élevés de ressources (lumière et eau) et est caractérisée par un faible stress environnemental et une longue saison de croissance (Vieilledent et al. 2013). Notez que la richesse spécifique estimée ne représente pas nécessairement le nombre “réel” d’espèces présentes, mais plutôt une richesse spécifique relative entre les sites (un site est plus riche en espèces qu’un autre).

5.1.7 Estimation 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.

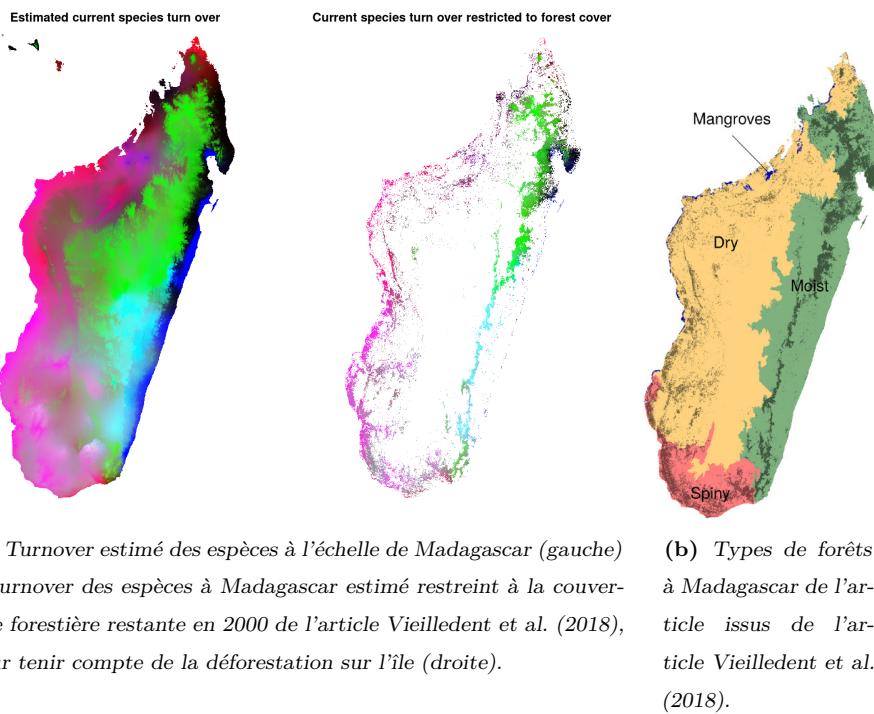


Figure 11 – Diversité β interpolée sur l'ensemble de l'île et restreinte au couvert forestier connu en 2000 et types de forêts à Madagascar.

Nous pouvons constater que la carte estimée de la diversité β sur la figure 11.a concorde avec la carte des types de forêts de Madagascar de l'article Vieilledent et al. (2018) sur la figure 11.b. En effet, les communautés d'espèces représentées par des nuances de bleu et de vert correspondent à la forêt humide de l'est, les communautés représentées par des nuances de rouge correspondent à la forêt sèche de l'ouest et les communautés en magenta correspondent à la forêt épineuse du sud. De plus, la carte de la diversité β que nous obtenons permet d'identifier des communautés d'espèces sensiblement différentes au sein de ces trois types de forêts. Par exemple, nous pouvons clairement identifier trois communautés d'arbres différentes (en bleu, vert et noir) au sein de la forêt humide orientale.

5.2 Classification des communautés d'arbres en Nouvelle-Calédonie

J'ai participé activement à l'encadrement de Pierre Guillaumont qui a effectué son stage de M2 MIND au sein de l'UMR AMAP, sur le sujet : "Classification des communautés d'arbres en Nouvelle-Calédonie à l'aide de modèle joint de distribution des espèces".

5.2.1 Ajustement d'un JSDM en Nouvelle Calédonie

Dans le cadre de son stage, Pierre a participé au développement du package R `gecevar`, Pierre a utilisé le package R `jSDM`, que j'ai développé, afin d'ajuster un modèle joint de distribution des espèces en Nouvelle Calédonie, à partir des données climatiques et environnementales obtenues à l'aide du package `gecevar`, ainsi que de données d'inventaire forestier répertoriant un très grand nombre d'espèces sur de nombreux sites d'inventaires.

Figure 12 – Variables explicatives du modèle et les ordres considérés.

Variables explicatives du modèle	
Statut ultramafique	
Moyenne annuelle des températures	
Saisonnalité des températures	Ordre 1 uniquement
Précipitation annuelle	Ordre 1 et Ordre 2
Variance des précipitations mensuelles	
Déficit hydrique	
Logarithme de l'aire du site d'inventaire	

Figure 13 – Ajustement d'un JSDM à partir des données disponibles en Nouvelle Calédonie.

Nombre d'espèces	877
Nombre de sites	554
Nombre d'observations	45258
Nombre d'itérations	15000
Temps de calcul (en heure)	1
Déviance	50480.55

On peut voir que malgré les jeux de données très conséquents considérés, l'ajustement du JSDM n'a pris qu'une heure en utilisant le package `jSDM` alors qu'avec d'autres packages l'ajustement du modèle aurait pris bien plus longtemps voire n'aurait pas été possible.

5.2.2 Obtention de cartes de biodiveristé en Nouvelle Calédonie

En utilisant le JSDM ajusté, il a pu obtenir des cartes de biodiversité en Nouvelle Calédonie en suivant les mêmes méthode que celle présentée pour Madagascar.

Figure 14 – Richesse spécifique estimée en Nouvelle Calédonie, restreinte au couvert forestier restant en 2000

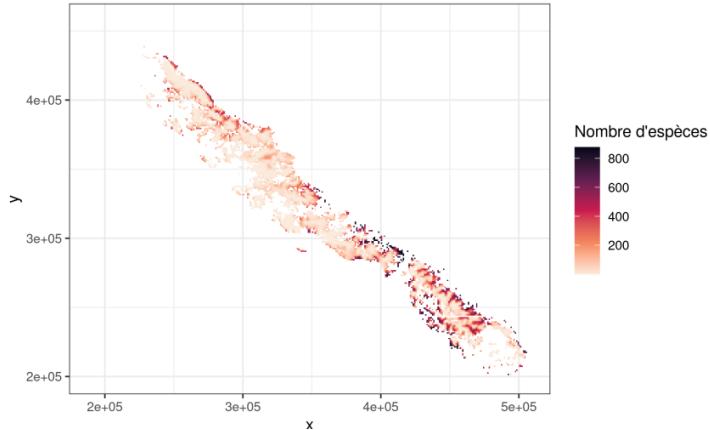
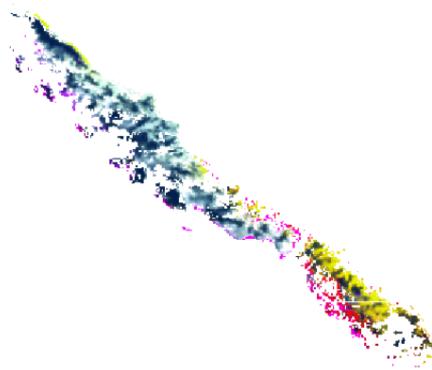


Figure 15 – Diversité β estimée en Nouvelle Calédonie, restreinte au couvert forestier restant en 2000



5.3 Obtention de cartes de communauté en Guyane française

5.3.1 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 les données climatiques et environnementales rassemblées à l'aide du package `gecevar` 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).

J'ai d'ailleurs participé en octobre 2021 à une mission en forêt guyanaise, de mesure de traits foliaires liés à la gestion de l'eau, 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.

5.3.2 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 (CO_2) 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.

6 Perspectives de développement du package `jSDM`

6.1 Fonctionnalités en cours de développement

6.1.1 Ajuster un modèle avec des effets espèces aléatoires

Dans le cas où on considère que les effets espèces β sont aléatoire, leur distribution *a priori* est de la forme : $\beta_j \sim \mathcal{N}_{p+1}(\mu_\beta, V_\beta)$ et on suppose que $V_\beta \sim \text{Wishart}^{-1}(M, \nu)$.

Pour ajuster un modèle avec effet espèce aléatoire, on utilise les mêmes méthodes d'inférence que précédemment. De plus le choix de cette distribution *a priori* inverse Wishart pour la matrice de variance-covariance des effets espèces, nous permet d'utiliser les formules de priors conjugués pour l'estimation de ces paramètres.

6.1.2 Ajuster un JSDM à partir d'une variable réponse continue

Afin de permettre au package d'ajuster un JSDM à partir d'une variable réponse continue, on a développé la fonction `jSDM_gaussian`.

Dans ce cas on suppose que $y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$, pour des données continues et on considère le modèle linéaire à variables latentes suivant :

$$\mu_{ij} = \alpha_i + X_i \beta_j + W_i \lambda_j + \epsilon_{ij}$$

où $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ et σ est un paramètre de dispersion à estimer.

Pour ajuster ce modèle, on utilise un échantillonneur de Gibbs associé aux formules des priors conjugués, de la même façon que pour le modèle probit.

6.1.3 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`.

6.2 Perspectives de développement

6.3 Estimation du nombre d'axes latents à prendre en compte

Plutôt que de fixer arbitrairement un nombre d'axe latents à considérer (q) dans le modèle, il serait préférable d'estimer q en ajoutant une variable dans le modèle hiérarchique Bayésien qui va estimer ce nombre, en s'inspirant du package `Hmsc` qui propose déjà cette fonctionnalité.

6.4 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 implantées mais ne sont pas abouties, en effet les résultats obtenus ne sont pas satisfaisants pour l'instant.

6.5 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>.
- 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:cspapt%5D2.0.co;2](https://doi.org/10.1657/1523-0430(2005)037%5B0444:cspapt%5D2.0.co;2).
- 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>.
- 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>.
- 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>.
- Pichler, Maximilian, and Florian Hartig. 2020. *A New Method for Faster and More Accurate Inference of Species Associations from Novel Community Data*.
- 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](https://doi.org/10.1016/j.compag.2005.07.003).

- 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>.
- 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>.