Développement d'un modèle joint de distribution des espèces pour la réalisation d'une carte de biodiversité à Madagascar

Jeanne Clément

Rapport de stage, Février à Août 2019

Enseignant référent : Benoite De Saporta

Encadrant : Ghislain Vieilledent

Master Maths-Biostatistique

Université Montpellier 2

UMR AMAP - Montpellier

botAnique et Modélisation de l'Architecture des Plantes et des végétations

Remerciements

J'aimerai adresser mes plus sincères remerciements à G. Vieilledent qui m'a encadrée et conseillée durant ce stage riche en découvertes puisque le language C++, la construction de packages R ainsi que les modèles joints de distribution des espèces m'étaient inconnus. Il m'a beaucoup appris et encouragée à trouver des solutions par moi même. Je remercie également les chercheurs et autres stagiaires de l'UMR AMAP pour leur accueil chaleureux et leur bonne humeur communicative qui font du laboratoire un cadre de travail idéal et tout particulièrement G. Le Moguedec qui fut une référence précieuse en statistiques ainsi que l'instigateur de pic-niques au lac du Crès qui nous ont bien aidé à supporter la canicule.

Sommaire

1 Définition des modèles joints de distribution des espèces envisagés **2**
1.1 Modèle linéaire mixte généralisé (GLMM) . . . 2
1.2 Modèle à variable latente (LVM) . . . 2

2 Méthodes d'inférence bayesienne selon la fonction de lien choisie **3**
2.1 Principe d'un échantillonneur de Gibbs . . . 3
2.2 Modèle probit : échantillonneur de Gibbs et priors conjugués . . . 3
2.2.1 Définition du modèle . . . 3
2.2.2 Priors utilisés . . . 4
2.2.3 Propositions sur les priors conjugués . . . 4
2.2.4 Echantilloneur de Gibbs . . . 7
2.2.5 Modèle logit : échantillonneur de Gibbs et algorithme de Metropolis adaptatif . . . 7
2.2.6 Définition du modèle . . . 8
2.2.7 Priors utilisés . . . 8
2.2.8 Principe d'un algorithme de Metropolis adaptatif . . . 8
2.2.9 Echantilloneur de Gibbs . . . 8
2.3 Evaluation de la fiabilité de ces méthodes sur des données simulées . . . 10

3 Application aux données collectées à Madagascar **11**
3.1 Description des données . . . 11
3.2 Estimation des paramètres . . . 11
3.3 Prédictions par interpolation . . . 11
3.4 Prédictions avec auto-corrélation spatiale . . . 11
3.5 Analyse des résultats et mise en évidence de lieux refuges de la biodiversité . . . 11

Liste des figures

Liste des tableaux

Introduction

J'ai effectué mon stage au sein de l'UMR AMAP (botAnique et Modélisation de l'Architecture des Plantes et des végétations), qui se trouve à Montpellier. Il s'agit d'une unité interdisciplinaire hébergée par le Cirad ou « Centre de Coopération Internationale en Recherche Agronomique pour le Développement» et qui mène des recherches sur les plantes et les végétations, dans le but de prévoir la réponse des écosystèmes aux forçages environnementaux.

Ce stage s'inscrit dans le cadre du projet BioSceneMada qui vise à fournir des scénarios d'évolution de la biodiversité sous l'effet conjoint du changement climatique et de la déforestation à Madagascar. Pour ce faire, plusieurs jeux de données sur la biodiversité ont été collectés et regroupés pour différents groupes taxonomiques (mammifères, oiseaux, reptiles, amphibiens, arbres, plantes herbacées, invertébrés), parmi lesquels j'ai utilisé des inventaires forestiers répertoriant l'absence ou la présence d'espèces d'arbres sur différents sites de l'île ainsi que des variables bioclimatques afin d'ajuster un modèle joint de distribution des espèces permettent d'estimer la niche des espèces, de prédire leur distribution, tout en prenant en compte les intéractions entre espèces (Warton et al. (2015)).
Dans un premier temps j'ai implémenté différents échantillonneurs de Gibbs en C++ permettant d'estimer les paramètres de modèles joints de distribution des espèces (JSDM) comportant des variables latentes, à l'aide du package Rcpp. La construction du package R https://ecology.ghislainv.fr/jSDM/ autour de l'une de ces fonctions ainsi que sa présentation à la conférence useR 2019 ont constitué une partie importante de mon stage. L'ajustement d'un JSDM sur des données d'inventaires forestiers collectées à Madagascar ainsi que des variables climatiques et environnementales, m'a permis d'obtenir des carte reflétant la biodiversité sur l'île afin de par la suite identifier des zones refuges de la biodiversité sous l'effet du changement climatique en utilisant les variables bioclimatiques fournies par les scénarios du GIECC. Ces résultats seront utilisés pour des préconisations de gestion de la biodiversité dans le cadre du projet BioSceneMada.

1 Définition des modèles joints de distribution des espèces envisagés

Les données dont on dispose pour ajuster ce type de modèle sont les réalisations d'une variable réponse, $Y = (y_{ij})_{j=1,\dots,J}^{i=1,\dots,I}$ telle que :

$$y_{ij} = \begin{cases} 0 & \text{si l'espèce } j \text{ est absente du site } i \\ 1 & \text{si l'espèce } j \text{ est présente sur le site } i, \end{cases}$$

ainsi que de variables explicatives $X = (X_i)_{i=1,\dots,I}$ avec $X_i = (X_{i1}, \dots, X_{ip}) \in \mathbb{R}^p$ où p est le nombre de variables bioclimatiques considérées pour chaque site.
On note θ_{ij}, la probabilité de présence de l'espèce j sur le site i.

L'article Warton et al. (2015) développe deux approches hiérarchiques pouvant être utilisées à la spécification d'un modèle joint de distribution des espèces.

1.1 Modèle linéaire mixte généralisé (GLMM)

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

$$g(\theta_{ij}) = \alpha_i + \beta_{j0} + X_i\beta_j + u_{ij},$$

$$y_{ij} \mid u_{ij}, \alpha_i \sim \mathcal{B}ernoulli(\theta_{ij}),$$

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

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

où $g :]0,1[\to]-\infty, +\infty[$ est une fonction de lien, $\beta_j = (\beta_{j1}, \dots, \beta_{jp})'$ et β_{j0} sont les coefficients de régression correspondants aux variables bioclimatiques et l'intercept pour l'espèce j qui est supposé être un effet fixe, α_i représente l'effet aléatoire du site i, et $u_i = (u_{i1}, \dots, u_{iJ})$ est un effets aléatoires multivariés corrélés dont la matrice de variance covariance Σ controle la correlation 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.

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 prédicteurs non mesurés (ou "variables latentes") considérés et $\lambda_j = (\lambda_{j1}, \dots, \lambda_{jq})'$ les coefficients associés, on obtient le modèle à variables latentes **(LVM)** suivant :

$$g(\theta_{ij}) = \alpha_i + \beta_{j0} + X_i\beta_j + W_i\lambda_j$$

$$y_{ij} \mid W_i, \alpha_i \sim \mathcal{B}ernoulli(\theta_{ij}),$$

$$W_i \sim \mathcal{N}(0, I_q) \; iid$$

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

Ce qui revient à un cas particulier de GLMM multivarié auquel on impose la contrainte $\Sigma = \Lambda\Lambda'$ avec

$$\Lambda := \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1q} \\ \vdots & \vdots & \vdots \\ \lambda_{J1} & \dots & \lambda_{Jq} \end{pmatrix}$$

On préferera ce dernier modèle, 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).

On peut choisir de modéliser l'abondance absolue plutôt que l'abondance relative en supprimant l'effet site aléatoire α_i du modèle.

2 Méthodes d'inférence bayesienne selon la fonction de lien choisie

2.1 Principe d'un échantillonneur de Gibbs

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

L'**échantillonnage de Gibbs** consiste à :

— **Initialisation** : choix arbitraire de $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})$.

— **Itération t** : Générer $\theta^{(t)}$ de la manière suivante :

— $\theta_1^{(t)} \sim \pi\left(\theta_1 \mid \theta_2^{(t-1)}, \ldots, \theta_m^{(t-1)}, x\right)$

— $\theta_2^{(t)} \sim \pi\left((\theta_2 \mid (\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_m^{(t-1)}, x\right)$ ⋮

— $\theta_m^{(t)} \sim \pi\left(\theta_m \mid \theta_1^{(t)}, \ldots, \theta_{m-1}^{(t)}, x\right)$

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* $\pi(\theta \mid x)$.

Par conséquent l'implémentation d'un échantillonneur de Gibbs nécessite la connaissance des ditributions *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.

2.2 Modèle probit : échantillonneur de Gibbs et priors conjugués

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

2.2.1 Définition du modèle

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 observées en utilisant la proposition suivante :

Proposition 2.2.1.1 (Modèle probit par l'intermédiaire une variable latente)**.**

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

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

Alors on a $y_{ij} \mid Z_{ij} \sim \mathcal{B}ernoulli(\theta_{ij})$ *avec* $\text{probit}(\theta_{\text{ij}}) = \alpha_i + \beta_{j,0} + X_i\beta_j + W_i\lambda_j$.

Preuve 2.2.1.1.

$$\begin{aligned}\mathbb{P}(y_{ij}=1) &= \mathbb{P}(Z_{ij}>0)\\ &= \mathbb{P}(\alpha_i+\beta_{j0}+X_i\beta_j+W_i\lambda_j+\epsilon_{ij}>0)\\ &= \mathbb{P}(\epsilon_{ij}>-(\alpha_i+\beta_{j0}+X_i\beta_j+W_i\lambda_j)\ )\\ &= \mathbb{P}(\epsilon_{ij}\leq \alpha_i+\beta_{j0}+X_i\beta_j+W_i\lambda_j)\\ &= \Phi(\alpha_i+\beta_{j0}+X_i\beta_j+W_i\lambda_j)\end{aligned}$$

De la même façon on a :

$$\begin{aligned}\mathbb{P}(y_{ij}=0) &= \mathbb{P}(Z_{ij}\leq 0)\\ &= 1-\Phi(\alpha_i+\beta_{j,0}+X_i\beta_j+W_i\lambda_j)\end{aligned}$$

On définit le modèle probit à l'aide d'une variable latente afin d'être en mesure d'uiliser les propriétés des priors conjugués pour échantillonner les paramètres du modèle selon leur distributions conditionnelles *a posteriori*.

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 :

$$\begin{array}{lcl} V_\alpha & \sim & \mathcal{IG}(\text{shape}=0.5,\text{rate}=0.005)\text{ avec rate}=\frac{1}{\text{scale}},\\ \beta_{jk} & \sim & \mathcal{N}(0,10^6)\text{ pour } j=1,\dots,J \text{ et } k=1,\dots,p,\\ \lambda_{jl} & \sim & \begin{cases}\mathcal{N}(0,10) & \text{si } l<j\\ \mathcal{N}(0,10)\text{ tronquée à gauche par 0} & \text{si } l=j\\ P \text{ tel que } \mathbb{P}(\lambda_{jl}=0)=1 & \text{si } l>j\end{cases}\\ & & \text{pour } j=1,\dots,J \text{ et } l=1,\dots,q.\end{array}$$

En effet pour assurer l'identifiabilité du modèle les valeurs de Λ sont contraintes à des valeurs strictements 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).
La fonction *boral()* du package du même nom, permettant d'ajuster toutes sortes de modèles utilise ces distributions *a priori* pour le modèle qui nous intéresse. Dans l'article Warton et al. (2015) l'ajustement de modèle joints de distributions des espèces est réalisé avec *boral()* qui fonctionne avec JAGS (Just Another Gibbs Sampler) un programme de simulation à partir de modèles hiérarchiques bayésiens utilisant des méthodes MCMC, implémenté en C++. Cependant la fonction *jSDM_probit_block()* du package jSDM que j'ai implémentée utilise une distribution *a priori* jointe pour les effets espèces fixes de la manière qui suit.

2.2.3 Propositions sur les priors conjugués

Effets espèces fixes :

On se ramène à un modèle de la forme $Z^*=X\beta+\epsilon$, en posant $Z^*_{i,j}=Z_{ij}-\alpha_i=\beta_{j0}+X_i\beta_j+W_i\lambda_j+\epsilon_{ij}$, afin d'estimer simultanément les β_j et λ_j pour chacune des espèces j, ce qui revient en écriture matricielle à :

$$Z^*_j=\begin{pmatrix}Z^*_{1j}\\ \vdots\\ Z^*_{Ij}\end{pmatrix}=\underbrace{\begin{pmatrix}1 & X_{11} & \dots & X_{1p} & W_{11} & \dots & W_{1q}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & X_{I1} & \dots & X_{Ip} & W_{I1} & \dots & W_{Iq}\end{pmatrix}}_{D}\underbrace{\begin{pmatrix}\beta_{j0}\\ \beta_{j1}\\ \vdots\\ \beta_{jp}\\ \lambda_{j1}\\ \vdots\\ \lambda_{jq}\end{pmatrix}}_{P_j}+\begin{pmatrix}\epsilon_{1j}\\ \vdots\\ \epsilon_{Ij}\end{pmatrix}$$

$$= DP_j+\epsilon_j \quad \text{avec } \epsilon_j\sim\mathcal{N}_I(0_{\mathbb{R}^I},I_I).$$

On suppose que $P_j\sim\mathcal{N}_{p+q+1}(m,V)$ avec $m=0_{\mathbb{R}^{p+q+1}}$ et $V=diag(\underbrace{10^6,\dots,10^6}_{\times p+1},\underbrace{10,\dots,10}_{\times q})$, par exemple.

Bien que cette distribution *a priori* ne prennent pas en compte les contraintes sur Λ, elle permet l'échantillonage selon une loi normale multivariée des effets espèce fixes. On imposera les contraintes aux λ_{jl} concernés après les avoir simulés.

On applique la proposition suivante :

Proposition 2.2.3.1.

$$\begin{cases} Y \mid \beta \sim \mathcal{N}_n(X\beta, I_n) \\ \beta \sim \mathcal{N}_p(m, V) \end{cases} \Rightarrow \begin{cases} \beta|Y \sim \mathcal{N}_p(m^*, V^*) \ avec \\ m^* = (V^{-1} + X'X)^{-1}(V^{-1}m + X'Y) \\ V^* = (V^{-1} + X'X)^{-1} \end{cases}$$

.

Preuve 2.2.3.1.

$$\begin{aligned} p(\beta \mid Y) &\propto p(Y \mid \beta)\, p(\beta) \\ &\propto \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(Y - X\beta)'(Y - X\beta)\right) \frac{1}{(2\pi)^{\frac{p}{2}}|V|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\beta - m)'V^{-1}(\beta - m)\right) \\ &\propto \exp\left(-\frac{1}{2}\left((\beta - m)'V^{-1}(\beta - m) + (Y - X\beta)'(Y - X\beta)\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\beta'V^{-1}\beta + m'V^{-1}m - m'V^{-1}\beta - \beta'V^{-1}m + Y'Y + \beta'X'X\beta - Y'X\beta - \beta'X'Y\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\beta'(V^{-1} + X'X)\beta - \beta'(V^{-1}m + X'Y) - (Y'X + m'V^{-1})\beta + m'V^{-1}m + Y'Y\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\beta'(V^{-1} + X'X)\beta - \beta'(V^{-1}m + X'Y) - (X'Y + V^{-1}m)'\beta + m'V^{-1}m + Y'Y\right)\right) \\ &\propto \exp(-\frac{1}{2}\left(\beta - (V^{-1} + X'X)^{-1}(V^{-1}m + X'Y)\right)'(V^{-1} + X'X)\left(\beta - (V^{-1} + X'X)^{-1}(V^{-1}m + X'Y)\right) \\ &\quad - (V^{-1}m + X'Y)'(V^{-1} + X'X)^{-1}(V^{-1}m + X'Y) + m'V^{-1}m + Y'Y) \\ &\propto \exp\left(-\frac{1}{2}\left(\beta - \underbrace{(V^{-1} + X'X)^{-1}(V^{-1}m + X'Y)}_{m^*}\right)' \underbrace{(V^{-1} + X'X)}_{V^{*-1}}\left(\beta - (V^{-1} + X'X)^{-1}(V^{-1}m + X'Y)\right)\right) \end{aligned}$$

On obtient :

$$\begin{cases} Z_j^* \mid P_j \sim \mathcal{N}_I(DP_j, I_I) \\ P_j \sim \mathcal{N}_{p+q+1}(m, V) \end{cases} \Rightarrow \begin{cases} P_j \mid Z_j^* \sim \mathcal{N}_{p+q+1}(m^*, V^*) \text{ avec} \\ m^* = (V^{-1} + D'D)^{-1}(V^{-1}m + D'Z_j^*) \\ V^* = (V^{-1} + D'D)^{-1} \end{cases}$$

.

Prédicteurs non mesurés (ou "variables latentes") :

De la même façon, on pose : $Z_{ij}^\star = Z_{ij} - \alpha_i - \beta_{j0} - X_i\beta_j = W_i\lambda_j + \epsilon_{ij}$, afin d'estimer W_i pour chaque site i.
En appliquant la proposition précédente, on obtient :

$$\begin{cases} Z_i^* := (Z_{i1}^*, \ldots, Z_{iJ}^*)' \mid W_i \sim \mathcal{N}_J(\Lambda W_i', I_J) \\ W_i' \sim \mathcal{N}_q(0_{\mathbb{R}^q}, I_q) \end{cases} \Rightarrow \begin{cases} W_i' \mid Z_i^* \sim \mathcal{N}_q(m^*, V^*) \text{ avec} \\ m^* = (I_q + \Lambda'\Lambda)^{-1}(\Lambda' Z_i^*) \\ V^* = (I_q + \Lambda'\Lambda)^{-1} \end{cases}$$

.

Effets site aléatoires et variance associée :

En ce qui concerne l'effet site aléatoire $(\alpha_i)_{i=1,\ldots,I}$, on pose $Z_{i,j}^* = Z_{ij} - D_iP_j = \alpha_i + \epsilon_{i,j}$, avec $D_i = (1, X_{i1}, \ldots, X_{ip}, W_{i1}, \ldots, W_{iq})$. On a ainsi $Z_{ij}^* \mid \alpha_i \sim \mathcal{N}(\alpha_i, 1)$ *iid* pour $j = 1, \ldots, J$, puis on applique la proposition suivante :

Proposition 2.2.3.2.

$$\begin{cases} x_i \mid \theta \sim \mathcal{N}(\theta,\ \sigma^2) \ iid \ pour \ i = 1, \dots, n \\ \theta \sim \mathcal{N}(\mu_0, {\tau_0}^2) \\ \sigma^2 \ connu \end{cases} \Rightarrow \begin{cases} \theta |\ x_1, \dots, x_n \sim \mathcal{N}(\mu_1, {\tau_1}^2) \ avec \\ \mu_1 = \dfrac{{\tau_0}^{-2}\mu_0 + \sigma^{-2}\sum_{i=1}^n x_i}{{\tau_0}^{-2} + n\sigma^{-2}} \\ {\tau_1}^{-2} = {\tau_0}^{-2} + n\sigma^{-2} \end{cases}$$

.

Preuve 2.2.3.2.

$$\begin{aligned} p(\theta \mid x_1, \dots, x_n) &\propto p(\theta) p(x_1, \dots, x_n \mid \theta) \\ &\propto \frac{1}{(2\pi{\tau_0}^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2{\tau_0}^2}(\theta - \mu_0)^2\right) \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2{\tau_0}^2}(\theta - \mu_0)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2{\tau_0}^2}(\theta^2 - 2\mu_0\theta) - \frac{1}{2\sigma^2}\sum_{i=1}^n (\theta^2 - 2\theta x_i)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\theta^2({\tau_0}^{-2} + n\sigma^{-2}) - 2\mu_0\theta{\tau_0}^{-2} - 2\theta\sigma^{-2}\sum_{i=1}^n x_i\right)\right) \\ &\propto \exp\left(-\frac{1}{2(\tau_0^{-2} + n\sigma^{-2})^{-1}}\left(\theta - \frac{\mu_0{\tau_0}^{-2} + \sigma^{-2}\sum_{i=1}^n x_i}{{\tau_0}^{-2} + n\sigma^{-2}}\right)^2\right) \end{aligned}$$

On obtient ainsi :

$$\begin{cases} Z_{ij}^* \mid \alpha_i \sim \mathcal{N}(\alpha_i,\ 1), \text{ iid } \forall j = 1, \dots, J \\ \alpha_i \sim \mathcal{N}(0, V_\alpha) \end{cases} \Rightarrow \begin{cases} \alpha_i |\ Z_{i1}^*, \dots, Z_{iJ}^* \sim \mathcal{N}(\mu_1, {\tau_1}^2) \text{ avec} \\ \mu_1 = \dfrac{\sum_{j=1}^J Z_{ij}^*}{V_\alpha^{-1} + J} \text{ et } {\tau_1}^{-2} = V_\alpha^{-1} + J. \end{cases}$$

Finalement pour estimer V_α, la variance des effets site aléatoires $(\alpha_i)_{i=1,\dots,I}$, on utilise la proposition suivante :

Proposition 2.2.3.3. *Si*

$$\begin{cases} x \mid \sigma^2 \sim \mathcal{N}_n(\theta,\ \sigma^2 I_n) \\ \sigma^2 \sim \mathcal{IG}(a, b) \\ \theta \ connu \end{cases} \Rightarrow \begin{cases} \sigma^2 | x \sim \mathcal{IG}(a', b') \ avec \\ a' = a + \frac{n}{2} \ et \ b' = \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2 + b. \end{cases}$$

Preuve 2.2.3.3.

$$\begin{aligned} p(\sigma^2 \mid x) &\propto p(x \mid \sigma^2)\ p(\sigma^2) \\ &\propto \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x - \theta)'(x - \theta)\right) \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-\left(\frac{n}{2} + a + 1\right)} \exp\left(-\frac{1}{\sigma^2}\left(b + \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right)\right) \end{aligned}$$

On a donc :

$$\begin{cases} (\alpha_1, \dots, \alpha_I)' \mid V_\alpha \sim \mathcal{N}_n(0_{\mathbb{R}^I}, V_\alpha I_n) \\ V_\alpha \sim \mathcal{IG}(a, b) \end{cases} \Rightarrow \begin{cases} V_\alpha \mid \alpha_1, \dots, \alpha_I \sim \mathcal{IG}(a', b') \text{ avec} \\ a' = a + \frac{I}{2} \text{ et } b' = b + \frac{1}{2}\sum_{i=1}^I \alpha_i^2. \end{cases}$$

2.2.4 Echantilloneur de Gibbs

L'algorithme utilisé pour estimer les paramètres de ce modèle est donc le suivant :

— Définir les constantes N_{Gibbs}, N_{burn}, N_{thin} telles que N_{Gibbs} correspond au nombre d'itérations effectuées par l'échantillonneur de Gibbs, N_{burn} au nombre d'itérations nécessaires pour le burn-in ou temps de chauffe et $N_{samp} = \frac{N_{Gibbs} - N_{burn}}{N_{thin}}$ au nombre de valeurs estimées retenues pour chaque paramètre. En effet on enregistre les paramètres estimés à certaines itérations, afin d'obtenir un échantillon de N_{samp} valeurs distribuées selon la distribution a posteriori pour chacun des paramètres.

— Initialiser tous les paramètres à 0 par exemple, excepté les valeurs diagonales de Λ initialisées à 1 et $V_\alpha^{(0)} = 1$.

— Gibbs sampler : à chaque itération t pour $t = 1, \ldots, N_{Gibbs}$ on répète chacune de ces étapes :

— Génerer la variable latente $Z^{(t)} = \left(Z_{ij}^{(t)}\right)_{i=1,\ldots,I}^{j=1,\ldots,J}$ telle que

$$Z_{ij}^{(t)} \sim \begin{cases} \mathcal{N}\left(\alpha_i^{(t1)} + \beta_{j0}^{(t1)} + X_i\beta_j(t1) + W_i^{(t1)}\lambda_j^{(t1)},\ 1\right) \text{ tronquée à droite par 0} & \text{si } y_{ij} = 0 \\ \mathcal{N}\left(\alpha_i^{(t1)} + \beta_{j0}^{(t1)} + X_i\beta_j(t1) + W_i^{(t1)}\lambda_j^{(t1)},\ 1\right) \text{ tronquée à gauche par 0} & \text{si } y_{ij} = 1 \end{cases}$$

,la variable latente est ainsi initialisée à la première itération en la générant selon ces lois normales centrées.

— Générer les effets espèces fixes $P_j^{(t)} = (\beta_{j0}^{(t)}, \beta_{j1}^{(t)} \ldots, \beta_{jp}^{(t)}, \lambda_{j1}^{(t)}, \ldots, \lambda_{jq}^{(t)})'$ selon :

$$P_j^{(t)} \mid Z^{(t)}, W_1^{(t-1)}, \alpha_1^{(t-1)}, \ldots, W_I^{(t1)}, \alpha_I^{(t-1)} \sim \mathcal{N}_{p+q+1}(m^\star, V^\star), \text{ avec}$$

$$m^\star = (V^{-1} + {D^{(t)}}'D^{(t)})^{-1}(V^{-1}m + {D^{(t)}}'Z_j^\star) \text{ et } V^\star = \left(V^{-1} + {D^{(t)}}'D^{(t)}\right)^{-1},$$

$$\text{où } Z_j^\star = (Z_{1j}^\star, \ldots, Z_{Ij}^\star)' \text{ tel que } Z_{ij}^\star = Z_{ij}^{(t)} - \alpha_i^{(t-1)}.$$

Afin de contraindre les valeurs diagonales de $\Lambda = (\lambda_{jl})_{j=1,\ldots,J}^{l=1,\ldots,q}$ à des valeurs positives et de rendre la matrice triangulaire inférieure, on modifie les valeurs des $P^{(t)}$ simulées aléatoirement selon les conditions suivantes :

$$P_{jp+1+l}^{(t)} = \lambda_{jl}^{(t)} \leftarrow \begin{cases} 0 & \text{si } l > j \\ \lambda_{jl}^{(t-1)} & \text{si } l = j \text{ et } \lambda_{jl}^{(t)} < 0. \end{cases}$$

On pose $P^{(t)} = \left(P_1^{(t)} | \ldots | P_J^{(t)}\right)$.

— Générer les prédicteurs non mesurés (ou "variables latentes")

$$W_i^{(t)} \mid Z^{(t)}, P^{(t)}, \alpha_i^{(t-1)} \sim \mathcal{N}_q\left((I_q + {\Lambda^{(t)}}'\Lambda^{(t)})^{-1}({\Lambda^{(t)}}'Z_i^{\star\star}), (I_q + {\Lambda^{(t)}}'\Lambda^{(t)})^{-1}\right),$$

$$\text{où } Z_i^{\star\star} = (Z_{i1}^{\star\star}, \ldots, Z_{iJ}^{\star\star}) \text{ tel que } Z_{ij}^{\star\star} = Z_{ij}^{(t)} - \alpha_i^{(t-1)}\beta_{j0}^{(t)} - X_i\beta_j^{(t)}.$$

On pose $D_i^{(t)} = \left(1, X_{i1}, \ldots, X_{ip}, W_{i1}^{(t)}, \ldots, W_{iq}^{(t)}\right)$.

— Générer les effets sites aléatoires $\alpha_i^{(t)}$ pour $i = 1, \ldots, I$ selon :

$$\alpha_i | \ Z^{(t)}, P^{(t)}, W_i^{(t)} \sim \mathcal{N}\left(\frac{\sum_{j=1}^J Z_{ij}^{(t)} - D_i^{(t)}P_j^{(t)}}{{V_\alpha^{(t-1)}}^{-1} + J}, \left(\frac{1}{V_\alpha^{(t-1)}} + J\right)^{-1}\right)$$

— Générer la variance des effets site aléatoires $V_\alpha^{(t)}$ selon :

$$V_\alpha^{(t)} \mid \alpha_1^{(t)}, \ldots, \alpha_I^{(t)} \sim \mathcal{IG}\left(\text{shape} = 0.5 + \frac{I}{2}, \text{rate} = 0.005 + \frac{1}{2}\sum_{i=1}^I \left(\alpha_i^{(t)}\right)^2\right)$$

2.2.5 Modèle logit : échantillonneur de Gibbs et algorithme de Metropolis adaptatif

D'autre part on considère une fonction de lien logit : $p \to \ln\left(\frac{p}{1-p}\right)$.

2.2.6 Définition du modèle

Dans ce cas les distributions n'étant pas conjuguées, on ne peut appliquer les propositions précédentes par conséquent on approche les distributions a posteriori des paramètres à l'aide d'un algorithme de Metropolis adaptatif de la manière suivante :

2.2.7 Priors utilisés

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

$$\begin{array}{lcl} V_\alpha & \sim & \mathcal{IG}(\text{shape} = 0.5, \text{rate} = 0.005) \text{ avec rate} = \frac{1}{\text{scale}}, \\ \beta_{jk} & \sim & \mathcal{N}(0, 10^6) \text{ pour } j = 1, \ldots, J \text{ et } k = 1, \ldots, p, \\ \lambda_{jl} & \sim & \begin{cases} \mathcal{N}(0,10) & \text{si } l < j \\ \mathcal{U}(0,10) & \text{si } l = j \\ P \text{ tel que } \mathbb{P}(\lambda_{jl} = 0) = 1 & \text{si } l > j \end{cases} \\ & & \text{pour } j = 1, \ldots, J \text{ et } l = 1, \ldots, q. \end{array}$$

2.2.8 Principe d'un algorithme de Metropolis adaptatif

** Algorithme de Metropolis adaptatif ** :
Cet algorithme appartient aux méthodes MCMC et permet de générer une chaîne de Markov dont la distribution stationnaire est celle voulue. On l'utilise pour échantillonner les paramètres selon leurs distributions conditionnelles *a posterirori* connue à une constante multiplicative près.

- **Initialisation** : $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})$ fixés arbitrairement.
- **Iteration t** :
- Générer $\theta_i^* \sim q(\theta_i^{(t-1)}, .)$, comme densité instrumentale conditionnelle $q(\theta_i^{(t-1)}, .)$ symétrique, on utilise $\mathcal{N}(\theta_i^{(t-1)}, \sigma^{2\ (t)}_{\theta_i}$ par exemple.
- Calculer la probabilité d'acceptation :

$$\alpha = min\left(1, \frac{\pi(\theta_i^*)}{\pi(\theta_i^{(t-1)})}\right)$$

.

- Retenir

$$\theta_i^{(t)} = \begin{cases} \theta_i^* & \text{avec probabilité } \alpha \\ \theta_i^{(t-1)} & \text{avec probabilité } 1-\alpha. \end{cases}$$

2.2.9 Echantilloneur de Gibbs

- Initialisation
- Définition des constantes N_{Gibbs}, N_{burn}, N_{thin} et R_{opt} tels que N_{Gibbs} correspond au nombre d'itérations effectuées par l'algorithme, N_{burn} au nombre d'itérations nécessaires pour le burn-in ou temps de chauffe et R_{opt} au ratio d'acceptation optimal. On définit $N_{samp} = \dfrac{N_{Gibbs} - N_{burn}}{N_{thin}}$ correspondant au nombre de valeurs estimées retenues pour chaque paramètre. En effet on enregistre les paramètres estimés à certaines itérations afin d'obtenir N_{samp} valeurs, nous permettant de représenter une distribution a posteriori pour chacun des paramètres.
- Implémentation de fonctions approchant la log-vraisemblance du modèle à partir des paramètres estimés à l'itération t :
  On pose $\theta^{(t)} = (\theta_{i,j}^{(t)})_{j=1,\ldots,m}^{i=1,\ldots,n}$.

$$\log(L(\theta^{(t)})) = l(\theta^{(t)}) = \sum_{\substack{1\le i\le n \\ 1\le j\le m}} \log\left(\mathbb{P}(y_{i,j}|\theta_{i,j}^{(t)})\right) = \sum_{\substack{1\le i\le n \\ 1\le j\le m}} \log\left(\binom{n_i}{y_{i,j}} (\theta_{i,j}^{(t)})^{y_{i,j}} (1-\theta_{i,j}^{(t)})^{n_i - y_{i,j}}\right)$$

  et retournant une valeur approchée du log de la loi a posteriori pour chacun des paramètres :

$$\text{On utilise : } \log\left(\mathrm{p}(\theta^{(t)} \mid Y)\right) \propto l(\theta^{(t)}) + \log(\underbrace{\Pi(\theta^{(t)})}_{\text{loi a priori}})$$

— fonction **betadens** estime :

$$\log\left(\mathrm{p}(\beta_j^{k\,(t)} \mid y_{i,j},\ \beta_j^{-k\,(t)})\right) \propto l(\theta^{(t)}) + \log(\underbrace{\Pi(\beta_j^{k\,(t)})}_{\text{loi a priori}}$$

— fonction **zdens** estime :

$$\log\left(\mathrm{p}(z_{i,l}{}^{(t)} \mid y_{i,j},\ z_i^{-l\,(t)})\right) \propto l(\theta^{(t)}) + \log(\underbrace{\Pi(z_{i,l}{}^{(t)})}_{\text{loi a priori}}$$

— fonction **lambdadens** estime :

$$\log\left(\mathrm{p}(\lambda_j^{q\,(t)} \mid y_{i,j},\ \lambda_j^{-q\,(t)})\right) \propto l(\theta^{(t)}) + \log(\underbrace{\Pi(\lambda_j^{q\,(t)})}_{\text{loi a priori}}$$

— fonction **alphadens** estime

$$\log\left(\mathrm{p}(\alpha_i{}^{(t)} \mid y_{i,j},\ \alpha_1^{(t)}, \ldots, \alpha_{i-1}^{(t)}, \alpha_{i+1}^{(t)}, \ldots, \alpha_n^{(t)})\right) \propto l(\theta^{(t)}) + \log(\underbrace{\Pi(\alpha_i{}^{(t)})}_{\text{loi a priori}}$$

— Pour $t = 1, \ldots, N_{Gibbs}$ à l'itération t on fait une boucle sur $i = 1, \ldots, I$ et sur $j = 1, \ldots, J$:

1. Calculer $\mathrm{logit}(\theta_{i,j}^{(t-1)}) = \alpha_i^{(t-1)} + \beta_{j,0}^{(t-1)} + X_i'\beta_j^{(t-1)} + z_i^{(t-1)'}\lambda_j^{(t-1)}$,
   puis $\theta_{i,j}^{(t-1)} = logit^{-1}(\Phi_{i,j}^{(t-1)}) = \dfrac{\exp(\Phi_{i,j}^{(t-1)})}{1 + \exp(\Phi_{i,j}^{(t-1)})}$.
2. **Algorithme Metropolis Hastings** :
   Pour chacun des paramètres on a un algo pour z_i par exemple :
   On initialise le nombre d'acceptation $nA^i = (nA_1^i, \ldots, nA_q^i) = 0_{\mathbb{R}^q}$ et le taux d'acceptation $Ar^i = (Ar_1^i, \ldots, Ar_q^i) = 0_{\mathbb{R}^q}$.
   Boucle sur $l = 1, \ldots, q$:
   — On pose $z_{now\,i,l} = z_{i,l}{}^{(t-1)}$.
   — On génère $z_{prop_{i,l}} \sim \mathcal{N}(z_{now\,i,l}, \sigma_{z_{i,l}}^{(t)})$ avec $\sigma_{z_{i,l}}^{(t)}$ adapté en fonction du nombre d'acceptation et initialisé par la valeur 1.
   — On calcule $p_{now} = \mathrm{zdens}(z_{now\,i,l})$ et $p_{prop} = \mathrm{zdens}(z_{prop_{i,l}})$.
   — On calcule la probabilité d'acceptation :

$$\alpha = exp(p_{prop} - p_{now}) = \frac{exp(p_{prop})}{exp(p_{now})} = \frac{L(\theta^{(t)})\Pi(z_{prop_{i,l}})}{L(\theta^{(t)})\Pi(z_{now\,i,l})}$$

   — On pose

$$z_{i,l}{}^{(t)} = \begin{cases} z_{prop_{i,l}} & \text{avec probabilité } \alpha \text{ si on est dans ce cas on fait } nA_{i,l} = nA_{i,l} + 1 \\ z_{now\,i,l} & \text{avec probabilité } 1 - \alpha \end{cases}$$

   — On pose

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

   .
   — **Durant le burnin** et lors des itérations t telles que $t + 1$ est multiple de DIV ($t < N_{burn}$ et $t + 1 \equiv 0 \pmod{DIV}$) pour $l = 1, \ldots, q$:
   On calcule $Ar_{i,l} = \dfrac{nA_{i,l}}{\mathrm{DIV}}$ puis on définit

$$\sigma_{z_{i,l}}^{(t)} = \begin{cases} \sigma_{z_{i,l}}^{(t-\mathrm{DIV})}\left(2 - \frac{1-Ar_{i,l}}{1-R_{opt}}\right) & \text{si } Ar_{i,l} \geq R_{opt} \\ \dfrac{\sigma_{z_{i,l}}^{(t-\mathrm{DIV})}}{2 - \frac{1-Ar_{i,l}}{1-R_{opt}}} & \text{sinon} \end{cases}$$

   On réinitialise les nombres d'acceptation : $nA_{i,l} \leftarrow 0$.

— **Après le burnin** et lors des itérations t telles que $t+1$ est multiple de DIV ($t \geq N_{burn}$ et $t+1 \equiv 0 \pmod{DIV}$) pour $l = 1, \ldots, q$:
On calcule $Ar_{i,l} = \dfrac{nA_{i,l}}{\text{DIV}}$ puis on réinitialise les nombres d'acceptation : $nA_{i,l} \leftarrow 0$.

— On calcule et affiche le taux d'acceptation moyen $mA^i = \dfrac{1}{q} \sum\limits_{l=1,\ldots,q} Ar_{i,l}$.

2.3 Evaluation de la fiabilité de ces méthodes sur des données simulées

3 Application aux données collectées à Madagascar

3.1 Description des données

On dispose d'inventaires forestiers réalisés sur différents sites de l'île de Madagascar (plot sites).

3.2 Estimation des paramètres

3.3 Prédictions par interpolation

3.4 Prédictions avec auto-corrélation spatiale

3.5 Analyse des résultats et mise en évidence de lieux refuges de la biodiversité

Conclusion

References

Albert, James H., and Chib Siddhartha. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88 (422) : 669–79. doi :10.1080/01621459.1993.10476321.

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). Elsevier : 766–79. doi :10.1016/j.tree.2015.09.007.