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

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 un échantillonneur de Gibbs en C permettant d'estimer les paramètres d'un modèle joint de distribution des espèces comportant des variables latentes, puis j'ai appliqué ce modèle afin d'obtenir une carte de biodiversité β à Madagascar à partir de données d'inventaires forestiers et de variables climatiques et environnementales, ce qui m'a permis par la suite d'identifier les zones refuges de la biodiversité sous l'effet du changement climatique en considérant 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

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

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.

1.2.1 Modèle probit

Variables latentes : En effet pour assurer l'identifiabilité du modèle les valeurs des lambdas sont contraintes à des valeurs strictements positives sur la diagonale et à 0 au dessus de la diagonale. On considerera donc une distribution a priori normale tronquée à gauche par 0 pour les lambdas diagonaux. et et on suppose que $V_\alpha \sim \mathcal{IG}(\text{shape} = 0.5, \text{rate} = 0.005)$. On utilise une distribution a priori $\mathcal{N}(0, 10^6)$ pour tous les betas. D'une part on utilise un fonction de lien probit : probit : $p \to \Phi^{-1}(p)$ où Φ correspond à la fonction de répartition d'une loi normale centrée réduite. D'après l'article Albert and Siddhartha (1993), on a la proposition suivante :

Proposition 1.2.1.1 (Modèle de régression probit en utilisant une variable latente)**.** *Si* $z_{i,j} = \alpha_i + \beta_{j,0} + X_i\beta_j + W_i\lambda_j + \epsilon_{i,j},\ \forall i, j$ *avec* $\epsilon_{i,j} \sim \mathcal{N}(0,1)$ iid *et tel que :*

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

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

Preuve 1.2.1.1.

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

De la même façon on a :

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

On définit le modèle porbit à l'aide d'une variable latente afin d'utiliser les propriétés des priors conjugués pour estimer les distributions conditionnelles a posteriori de chacun des paramètres du modèle.

1.2.2 Modèle de regression logistique

1.3 Méthodes d'estimation utilisées

1.3.1 Echantillonneur de Gibbs

Echantilloneur de Gibbs : Cet algorithme est utilisé pour obtenir un échantillon d'une variable aléatoire dont la distribution de probabilité est connue, par exemple $Z = (Z_1, Z_2, Z_3)$ distribué selon $\pi_i(z_i)$ connue pour $i = 1, 2, 3$.

— **Initialisation** : $z^{(0)} = 0_{\mathbb{R}^3}$.

— **Itération t** : Générer $z^{(t)}$ de la manière suivante :
— $z_1^{(t)} \sim \pi_1\left(z_1\ |z_2^{(t-1)}, z_3^{(t-1)}\right)$
— $z_2^{(t)} \sim \pi_2\left(z_2 \mid z_1^{(t)}, z_3^{(t-1)}\right)$
— $z_3^{(t)} \sim \pi_3\left(z_3 \mid z_1^{(t)}, z_2^{(t)}\right)$

Par conséquent l'implémentation d'un Gibbs sampler 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.

1.3.2 Echantilloneur de Gibbs et priors conjugués pour le modèle probit

On se ramène à un modèle de la forme $Z^* = X\beta + \epsilon$ pour estimer les distributions conditionnelles a posteriori des effet espèces β_j et λ_j pour chacune des espèces j ainsi que celles des variables latentes W_i et des effets sites aléatoires α_i pour chaque site i.

D'une part afin d'estimer simultanément les β_j et λ_j pour chacune des espèces j, on pose $Z^*_{i,j} = Z_{i,j} - \alpha_i = \beta_{j,0} + X_i\beta_j + W_i\lambda_j + \epsilon_{i,j}$, ce qui revient en écriture matricielle à :

$$
\begin{aligned}
Z^*_j := \begin{pmatrix} Z^*_{1,j} \\ \vdots \\ Z^*_{I,j} \end{pmatrix} &= \underbrace{\begin{pmatrix} 1 & X_{i,1} & \dots & X_{i,p} & W_{i,1} & \dots & W_{i,q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{I,1} & \dots & X_{I,p} & W_{I,1} & \dots & W_{I,q} \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} \beta_{j,0} \\ \beta_{j,1} \\ \vdots \\ \beta_{j,p} \\ \lambda_{j,1} \\ \vdots \\ \lambda_{j,q} \end{pmatrix}}_{P} + \begin{pmatrix} \epsilon_{1,j} \\ \vdots \\ \epsilon_{I,j} \end{pmatrix} \\
&= DP + \epsilon_j
\end{aligned}
$$

Puis on applique la proposition suivante :

Proposition 1.3.2.1.

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

.

Preuve 1.3.2.1.

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

On obtient donc :

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

.

En ce qui concerne l'effet site aléatoire $(\alpha_i)_{i=1,\dots,I}$, on pose $Z_{i,j}^* = Z_{i,j} - DP = \alpha_i + \epsilon_{i,j}$, on a ainsi $Z_{i,j}^* \mid \alpha_i \sim \mathcal{N}(\alpha_i, 1)$ puis on applique la proposition suivante :

Proposition 1.3.2.2.

$$\begin{cases} x \mid \theta & \sim \mathcal{N}(\theta,\ \sigma^2) \\ \theta & \sim \mathcal{N}(\mu_0, {\tau_0}^2) \\ \sigma^2 & \textit{connu} \end{cases} \Rightarrow \begin{cases} \theta|\ x & \sim \mathcal{N}(\mu_1, {\tau_1}^2)\ \textit{avec} \\ \mu_1 & = \dfrac{{\tau_0}^2\mu_0 + x\sigma^2}{{\tau_0}^{-2} + \sigma^{-2}} \\ {\tau_1}^{-2} & = {\tau_0}^{-2} + \sigma^{-2} \end{cases}$$

.

Preuve 1.3.2.2.

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta)\ p(\theta) \\ &\propto \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \frac{1}{(2\pi{\tau_0}^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2{\tau_0}^2}(\theta-\mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2{\tau_0}^2}(\theta-\mu_0)^2 - \frac{1}{2\sigma^2}(x-\theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2{\tau_0}^2}(\theta^2 - 2\mu_0\theta) - \frac{1}{2\sigma^2}(\theta^2 - 2x\theta)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\theta^2({\tau_0}^{-2} + \sigma^{-2}) - 2\mu_0\theta{\tau_0}^{-2} - 2x\theta\sigma^{-2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2(\tau_0^{-2} + \sigma^{-2})^{-1}}\left(\theta^2 - 2\theta\frac{\mu_0{\tau_0}^{-2} + x\sigma^{-2}}{{\tau_0}^{-2} + \sigma^{-2}}\right)\right) \end{aligned}$$

On obtient ainsi :

$$\begin{cases} Z_{i,j}^* \mid \alpha_i \sim \mathcal{N}(\alpha_i,\ 1),\ \text{iid}\ \forall j = 1, \dots, J \\ \alpha_i \sim \mathcal{N}(0, V_\alpha) \end{cases} \Rightarrow \begin{cases} \alpha_i|\ x & \sim \mathcal{N}(\mu_1, {\tau_1}^2)\ \text{avec} \\ \mu_1 & = \dfrac{{\tau_0}^2\mu_0 + x\sigma^2}{{\tau_0}^{-2} + \sigma^{-2}} \\ {\tau_1}^{-2} & = {\tau_0}^{-2} + \sigma^{-2} \end{cases}$$

.

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

Proposition 1.3.2.3. *Si*

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

Preuve 1.3.2.3.

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

1.3.3 Echantilloneur de Gibbs et algorithme de Metropolis adaptatif pour le modèle logit

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

** Algorithme de Metropolis Hastings ** :
On l'utilise pour générer une variable aléatoire selon une estimation de sa distibution conditionnelle a posteriori.

- **Initialisation** : $z^{(0)} = 0_{\mathbb{R}^3}$.
- **Iteration t** :
- Générer $z^* \sim q(z^{(t-1)}, .)$, comme densité instrumentale conditionnelle $q(z^{(t-1)}, .)$ on utilise $\mathcal{N}(z^{(t-1)}, 1)$.
- Calculer la probabilité d'acceptation :
$$\alpha = min\left(1, \frac{\pi(z^*)}{\pi(z^{(t-1)})}\right)$$
.
- Retenir
$$z^{(t)} = \begin{cases} z^* & \text{avec probabilité } \alpha \\ z^{(t-1)} & \text{avec probabilité } 1-\alpha. \end{cases}$$
- Initialisation en utilisant les lois a priori définies pour chacun des paramètres :
- Générer $\beta_j^{(0)}$ pour $j = 1, \ldots, J$ avec $\beta_{j,k} \sim \mathcal{N}(\mu_{\beta_{j,k}}, \sigma^2_{\beta_{j,k}})$, où $\mu_{\beta_{j,k}} = 0$ et $\sigma^2_{\beta_{j,k}} = 1e+06$ pour $k = 1, \ldots, p$.
- Générer $\beta_{j,0}^{(0)}$ pour $j = 1, \ldots, J$ selon une $\mathcal{N}(0, 1e+06)$.
- Générer $\lambda_j^{(0)}$ pour $j = 1, \ldots, J$ avec $\lambda_{j,l} \sim \mathcal{N}(0, \sigma^2_{\lambda_{j,l}})$ où $\sigma^2_{\lambda_{j,l}} = 20$ pour $l = 1, \ldots, q$.
- Générer $W_i^{(0)}$ pour $i = 1, \ldots, n$ selon une $\mathcal{N}(0, I_q)$.
- Générer $\alpha_i^{(0)}$ pour $i = 1, \ldots, I$ selon une $\mathcal{N}(0, 100)$ si effet fixe.
- 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 $N_{samp} = \frac{N_{Gibbs} - N_{burn}}{N_{thin}}$ 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 :

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

— fonction **betadens** estime :

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

— fonction **zdens** estime :

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

— fonction **lambdadens** estime :

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

— fonction **alphadens** estime

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

— 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 $Ar_{i,l} = \dfrac{nA_{i,l}}{\text{DIV}}$ puis on définit

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

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 $mA^i = \dfrac{1}{q} \sum\limits_{l=1,\ldots,q} Ar_{i,l}$.

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

2 Application aux données collectées à Madagascar

2.1 Description des données

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

2.2 Estimation des paramètres

2.3 Prédictions par interpolation

2.4 Prédictions avec auto-corrélation spatiale

2.5 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.