

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



Auteur : Jeanne Clément

Encadrant : Ghislain Vieilledent

Enseignante référente : Benoîte De Saporta



Introduction

- Les jeux de données disponibles sont de taille de plus en plus importante et répertorient de nombreuse espèces.
- On s'intéresse depuis quelques années aux modèles joints de distribution des espèces (JSDM) plutôt que de considérer les espèces comme indépendantes et ajuster un modèle pour chacune, ce qui présente les avantages que l'on verra par la suite.
- Les packages R déjà existants qui permettent d'ajuster ce types de modèles sont peu nombreux : boral, gjam, HMSC, BayesComm. Ils nécessitent des temps de calcul assez important et sont pour certains difficiles d'utilisation.
- L'objectif de mon stage était donc d'implémenter une fonction optimisée permettant d'ajuster ce type de modèle en choisissant un algorithme d'inférence performant atteignant assez rapidement la convergence, puis de construire un package R autour de cette fonction.

Plan

1 Construction du package R jSDM autour d'un Gibbs sampler en C++

- Echantillonneur de Gibbs
- Utilisation du package Rcpp
- RcppGSL pour les tirages aléatoires
- RcppArmadillo pour les calculs matriciels

2 Ajustement d'un modèle joint de distribution des espèces (JSDM)

- Utilité des JSDM
- Description des données utilisés
- Définition du modèle
- Ajustement de ce modèle avec boral et jSDM

3 Application aux données collectées à Madagascar

- Description des données
- Ajustement du JSDM sur ces données
- Représentation des résultats
- Obtention de cartes de biodiversité par interpolation spatiale

Plan

1 Construction du package R jSDM autour d'un Gibbs sampler en C++

- Echantillonneur de Gibbs
- Utilisation du package Rcpp
- RcppGSL pour les tirages aléatoires
- RcppArmadillo pour les calculs matriciels

2 Ajustement d'un modèle joint de distribution des espèces (JSDM)

- Utilité des JSDM
- Description des données utilisés
- Définition du modèle
- Ajustement de ce modèle avec boral et jSDM

3 Application aux données collectées à Madagascar

- Description des données
- Ajustement du JSDM sur ces données
- Représentation des résultats
- Obtention de cartes de biodiversité par interpolation spatiale

Principe d'un échantillonneur de Gibbs

L'échantillonnage de Gibbs est une approche d'inférence bayesienne permettant d'obtenir une réalisation de $\Theta = (\theta_1, \dots, \theta_n)$ selon la distribution jointe *a posteriori* $p(\theta_1, \dots, \theta_n | x)$.

- ➊ Initialisation de $\Theta^{(0)}$ par des valeurs arbitraires.
- ➋ Connaissant $\Theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_n^{(t)})$ l'échantillon obtenu à l'itération t , à l'itération $t + 1$ on génère pour $i = 0, \dots, n$:

$$\theta_i^{(t+1)} \sim p(\theta_i^{(t+1)} | \theta_0^{(t+1)}, \dots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \dots, \theta_n^{(t)}, x)$$

- ➌ On répète l'étape précédente N_{Gibbs} fois (Markov Chain Monte Carlo) et les échantillons ainsi obtenus permettent d'approcher la distribution jointe *a posteriori*.

On peut intégrer différents algorithmes (eg. Metropolis-Hastings adaptatif) pour implémenter une ou plusieurs étapes de l'échantillonneur de Gibbs.

Implémentation d'un échantillonneur de Gibbs

- Utilise des tirages aléatoires pour générer selon les lois conditionnelles.
- Nécessite des calculs matriciels dans le cas où on utilise une approche d'inférence bayesienne avec des priors conjugués.
- Implique boucles imbriquées sur les N_{Gibbs} itérations (MCMC) et le nombre de paramètres (Gibbs sampler).

Rcpp (C++), RcppGSL (tirages aléatoires) et RcppArmadillo (calcul matriciel), peuvent donc être utiles pour implémenter un échantillonneur de Gibbs performant.

Utilisation du package Rcpp

Rcpp

- **Rcpp** est un package R permettant d'implémenter et compiler facilement du code en C++ avec RStudio.
- Principal avantage : fonctions en C++ plus rapides que celles en R.
- Implémenté par **Dirk EDDELBUETTEL** et **Romain FRANCOIS**
- <http://www.rcpp.org/>

Utilisation du package Rcpp

Fonction `Rcpp::sourceCpp()`

- Compile le code en C++
- Exporte la fonction vers la session R
- Interchange automatiquement les types d'objets entre R et C++
- ... (et bien d'autres, voir `vignette("Rcpp-package")`)

Fonctions pour construire un package R

- `Rcpp.package.skeleton()` pour générer un nouveau package Rcpp (`DESCRIPTION` et `NAMESPACE`)
- `Rcpp::compileAttributes()` examine les fonctions en C++ afin que `Rcpp::exportAttribute` génère le code requis pour les rendre disponibles depuis R.

Exemple simple d'utilisation de Rcpp

Code C++ (in file Code/addition.cpp)

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
int addition(int a, int b) {
  return a + b;
}
```

Code R

```
Rcpp::sourceCpp("~/Code/Report/Code/addition.cpp")
addition(2, 2)

## [1] 4
```

RcppGSL pour les tirages aléatoires

GNU Scientific Library

- Librairie numérique pour programmeurs C++ et C
- Algorithmes de génération de nombres aléatoires fiables
- Echantillonnages selon diverses distributions aléatoires.
- Algèbre linéaire (matrices et vecteurs)
- <https://www.gnu.org/software/gsl/>



RcppGSL

- Interface entre R et GSL
- Utilisant Rcpp pour intégrer du C dans R
- <http://dirk.eddelbuettel.com/code/rcpp.gsl.html>

Avantages

- Diversité : GSL v2.6 inclus **38 distributions aléatoires** alors que R API en inclus “seulement” 24.
- Facilité à échantillonner selon des distributions construites à partir des distributions de base de GSL (e.g inverse gamma).
- Rapidité : certains tirages aléatoires plus performants avec GSL qu'avec R (eg. `gsl_ran_gamma()` vs. `R::rgamma()`).

RcppArmadillo pour les calculs matriciels

Armadillo

- Librairie C++ pour l'algèbre linéaire et le calcul scientifique
- Classes pour vecteurs, matrices et cubes
- Opérations matricielles, décomposition de matrices, etc.
- <http://arma.sourceforge.net/>



RcppArmadillo

- Interface entre R et Armadillo : objets définis avec Armadillo convertis automatiquement en objets pris en charge par R
- Utilisant Rcpp pour intégrer du C++ dans R
- <http://dirk.eddelbuettel.com/code/rcpp.armadillo.html>

Avantages

- Facilité d'utilisation : syntaxe et fonctionnalités de haut niveau pour la manipulation de vecteurs, matrices et cubes.
- Rapidité : calculs matriciels plus performants avec Armadillo qu'avec R.

Package R jSDM

jSDM 0.1.0  Get started Reference Articles ▾ Change log  

jSDM R Package

Package for fitting joint species distribution models (JSDM) in a hierarchical Bayesian framework (Warton *et al.* 2015). The Gibbs sampler is written in C++. It uses Rcpp, Armadillo and GSL to maximize computation efficiency.



Links

Browse source code at
<https://github.com/ghislainv/jSDM>

Report a bug at
<https://github.com/ghislainv/jSDM/issues>

License

GPL-3 | file [LICENSE](#)

Developers

Ghislain Vieilledent
Author, maintainer 

Jeanne Clément
Author 

 cirad
Copyright holder, funder

Dev status

build 

CRAN 

DOI [10.5281/zenodo.3253460](https://doi.org/10.5281/zenodo.3253460)

downloads 

System requirements

Make sure the GNU Scientific Library ([GSL](#)) is installed on your system.

Installation

Install the latest stable version of **jSDM** from [CRAN](#) with:

```
install.packages("jSDM")
```



Or install the development version of **jSDM** from [GitHub](#) with:

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

References

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K. (2015) So many variables: Joint

- <https://ecology.ghislainv.fr/jSDM>
- Construit à l'aide des packages Rcpp, RcppGSL et RcppArmadillo.

Plan

1 Construction du package R jSDM autour d'un Gibbs sampler en C++

- Echantillonneur de Gibbs
- Utilisation du package Rcpp
- RcppGSL pour les tirages aléatoires
- RcppArmadillo pour les calculs matriciels

2 Ajustement d'un modèle joint de distribution des espèces (JSDM)

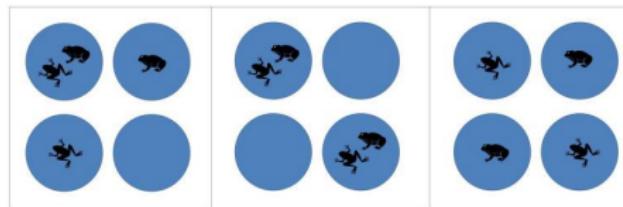
- Utilité des JSDM
- Description des données utilisés
- Définition du modèle
- Ajustement de ce modèle avec
boral et jSDM

3 Application aux données collectées à Madagascar

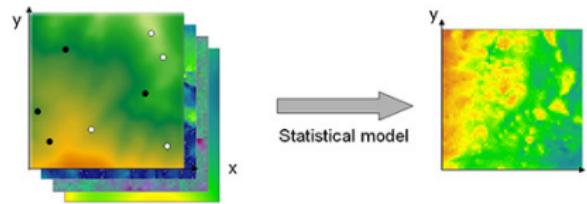
- Description des données
- Ajustement du JSDM sur ces
données
- Représentation des résultats
- Obtention de cartes de
biodiversité par interpolation
spatiale

Utilité des JSDM

- Utiliser l'information apportée par l'ensemble des espèces pour l'estimation des paramètres liés aux sites
- Prendre en compte les interactions entre les espèces



- Peuvent être utilisés pour expliquer/prédire la richesse spécifique des sites et produire différentes cartes de biodiversité



Field records and maps of environment

Map of probability species is present

Description des données utilisées

- Présence/absence de chaque espèce sur les sites
- Variables environnementales pour chaque site

Sites	Sp1	Sp2	...	Sp_nsp	X1	X2	...	X_nvar
Site1	0	0	...	1	-0.21	-1	...	-1.24
Site2	0	1	...	1	0.25	0	...	-0.53
...	-
Site_nsite	1	0	...	1	0.82	1	...	0.34

Définition du modèle

On définit le modèle suivant à partir de celui utilisé dans l'article Warton et al. (2015) :

$$Y = \{y_{ij}\}_{j=1, \dots, nsp}^{i=1, \dots, nsite}, \text{ avec :}$$

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

On suppose que $y_{ij} | W_i, \alpha_i \sim \text{Bernoulli}(\theta_{ij})$, avec :

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

où probit : $p \rightarrow \Phi^{-1}(p)$ avec Φ la fonction de répartition d'une loi normale centrée réduite,

α_i : effet site aléatoire tel que $\alpha_i \sim \mathcal{N}(0, V_\alpha)$ iid,

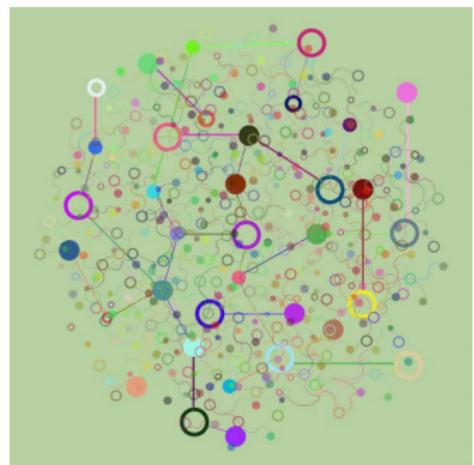
X_i : variables environnementales connues pour le site i ,

W_i : variables latentes pour le site i telles que $W_i \sim \mathcal{N}_{nlat}(0, I_{nlat})$ iid,

β_j, λ_j : effets espèce fixes.

Complexité du modèle

- Modèle linéaire généralisé de fonction de lien probit : variable réponse binaire.
- Multivarié : paramètres α_i pour les sites et β_j, λ_j pour les espèces.
- Modèle mixte avec effets sites aléatoires et effets espèces fixes.
- Variables latentes : W_i prédicteurs non mesurés ou non mesurables + principaux axes de covariation entre les taxons.



Contraintes d'identifiabilité

La complexité du modèle nous oblige à introduire les contraintes suivantes (cf Warton et al. (2015)) afin de le rendre identifiable :

$$\begin{cases} \lambda_{jl} = 0 & \text{si } l > j \\ \lambda_{jl} > 0 & \text{si } l = j, \end{cases}$$

pour $j = 1, \dots, nsp$ et $l = 1, \dots, nlat$.

Package R boral

- boral associe R et JAGS un programme de simulation à partir de modèles hiérarchiques bayésiens, implémenté en C++ permettant d'ajuster des JSDM.
- JAGS pour Just Another Gibbs Sampler :
<http://mcmc-jags.sourceforge.net/>
- Approche utilisée dans l'article Warton et al. (2015)
- boral par Francis K.C. Hui et JAGS de Martyn Plummer

Jeux de données utilisés

données	nsite	nsp	nobs	nX	nlat	npar	NA
Simulation	300	100	30000	2	2	1400	35000
Moustiques	167	16	2672	13	2	757	35000
Eucalyptus	458	12	5496	7	2	1494	35000
Grenouilles	104	9	936	3	2	366	35000
Champignons	800	11	8800	12	2	2565	35000



Moustiques



Eucalyptus



Grenouilles



Champignons

Comparaison des résultats obtenus avec boral et jSDM

Après 35000 itérations on obtient les résultats suivant avec chacun des packages :

Temps de calcul (en minutes)

Simulation	Moustiques	Eucalyptus	Grenouilles	Champignons
boral	96.9	5.8	17.2	1.2
jSDM	7.0	1.3	1.8	0.3

jSDM est **4 à 14** fois plus rapide que boral.

Root-mean-square error

Calculé pour $\text{probit}(\theta_{ij})$ sur les données simulées :

	boral	jSDM
RMSE	1.8	0.6

Comparaison des résultats obtenus avec boral et jSDM

Déviance

Simulation	Moustiques	Eucalyptus	Grenouilles	Champignons
boral	40486	6936	8779	884
jSDM	15651	1231	1922	150

Conséquences

- Petits jeux de données **et** modèles simples : boral, R + JAGS.
- Larges jeux de données **ou** modèles hiérarchiques complexes : jSDM, R + Rcpp + RcppGSL + RcppArmadillo.

Plan

1 Construction du package R jSDM autour d'un Gibbs sampler en C++

- Echantillonneur de Gibbs
- Utilisation du package Rcpp
- RcppGSL pour les tirages aléatoires
- RcppArmadillo pour les calculs matriciels

2 Ajustement d'un modèle joint de distribution des espèces (JSDM)

- Utilité des JSDM
- Description des données utilisés
- Définition du modèle
- Ajustement de ce modèle avec boral et jSDM

3 Application aux données collectées à Madagascar

- Description des données
- Ajustement du JSDM sur ces données
- Représentation des résultats
- Obtention de cartes de biodiversité par interpolation spatiale

Description des données

- Inventaires forestiers nationaux réalisés entre 1994 et 1996 et répertoriant la présence ou l'absence de 555 espèces végétales sur 751 placettes d'inventaire.
- Données climatiques et environnementales disponibles sur le site <https://madaclim.cirad.fr>, on choisit d'utiliser 5 variables pour lesquelles on extrait les valeurs correspondant aux coordonnées des placettes d'inventaire.

Jeu de données utilisé

site	Ocotea laevis	...	Bridelia pervilleana	temp	prec	sais.temp	sais.prec	cwd	long	lat
1	0	...	0	241	1302	1316	110	498	49.1	-12.4
2	0	...	0	243	1288	1320	111	521	49.1	-12.4
3	0	...	1	238	1321	1321	110	475	49.2	-12.4
4	0	...	0	239	1321	1323	110	482	49.1	-12.4
5	0	...	1	238	1325	1329	109	466	49.1	-12.5
6	0	...	0	196	1498	1366	97	216	49.2	-12.6

Ajustement du JSDM sur ces données

On ajuste un modèle joint de distribution des espèces de fonction de lien probit à l'aide de la fonction `jSDM_probit_block` du package `jSDM`.

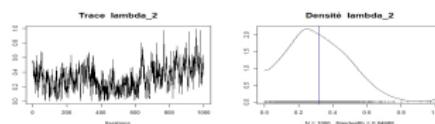
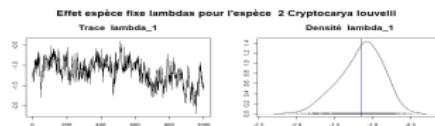
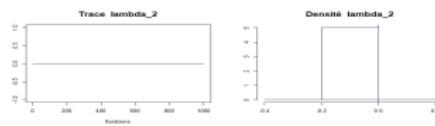
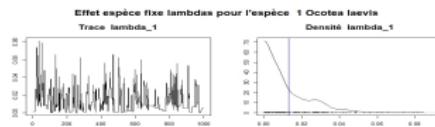
Ajustement du modèle et nombre de paramètres à estimer

nsite	nsp	nobs	nX	nlat	npar	ngibbs	temps
753	555	417915	10	2	9474	100000	12h

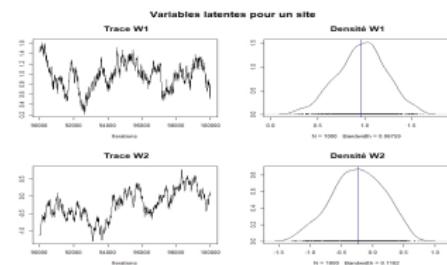
Traces et densité *a posteriori* des paramètres estimés

On met en évidence les **moyennes** des échantillons de $N_{\text{samp}} = 1000$ valeurs obtenus, que l'on utilisera comme estimateur pour les paramètres.

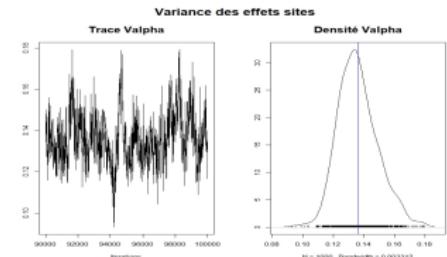
- Effets espèces fixes associés aux variables latentes



- Variables latentes W_1 et W_2

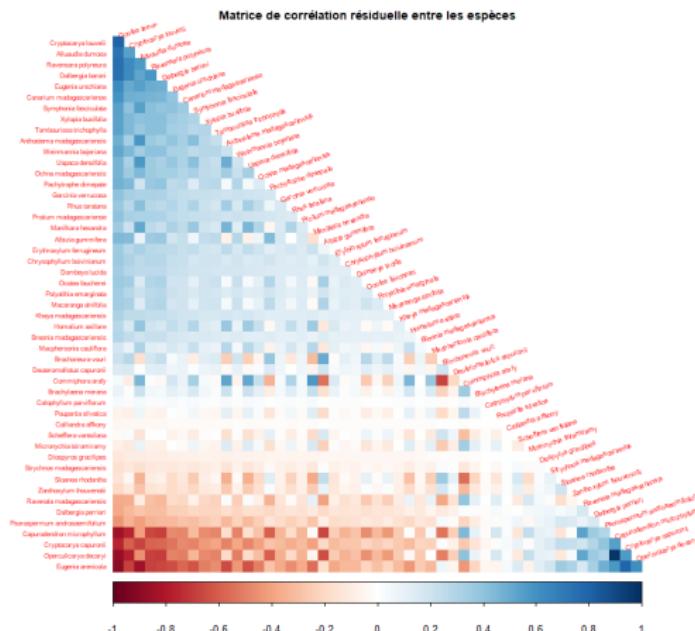


- Variance des effets sites aléatoires



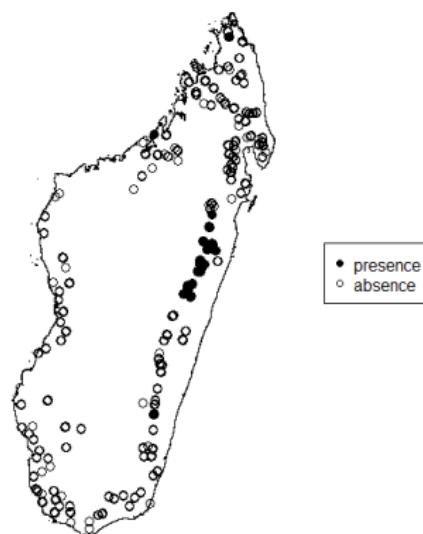
Matrice de corrélation résiduelle entre les espèces estimée

Matrice des corrélation résiduelle des probabilités d'occurrence des 50 espèces les plus présentes calculée comme dans l'article Warton et al. (2015).

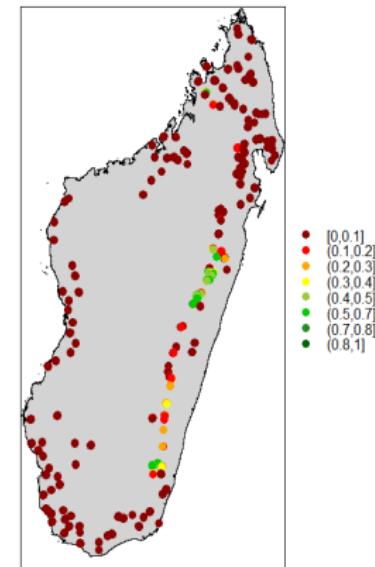


Probabilités de présence estimées comparées aux occurrences observées pour une espèce

Présences d'*Ocotea laevis* observées



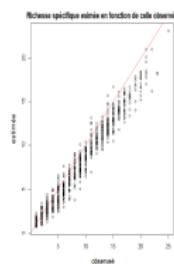
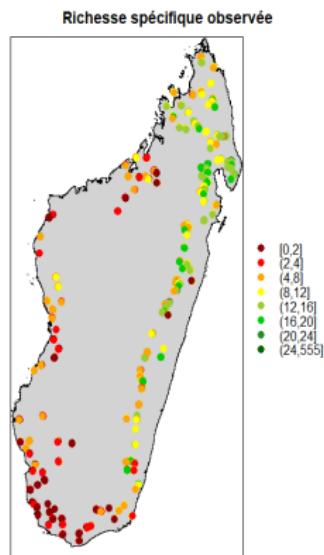
Probabilités de présence d'*Ocotea laevis* estimées



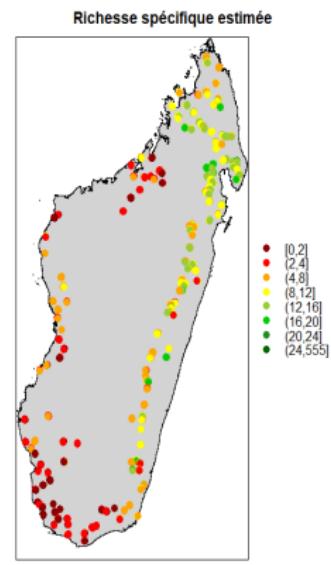
Richesse spécifique estimée comparée à celle observée

Richesse spécifique observée

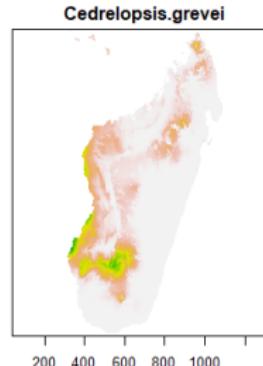
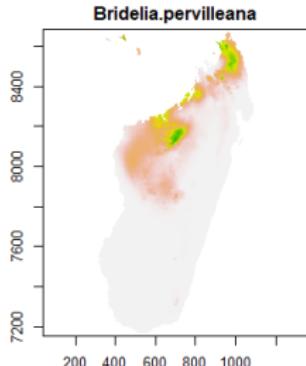
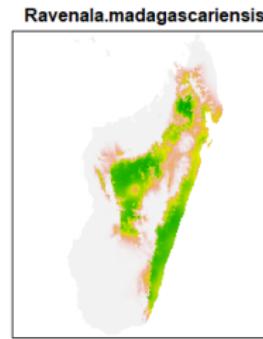
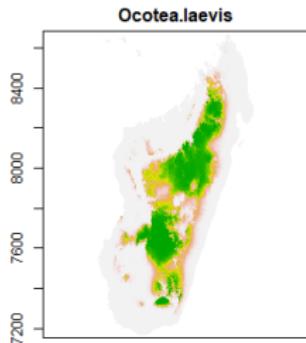
pour chaque site : $R_i = \sum_{j=1}^{555} y_{ij}$.



Richesse spécifique estimée pour
chaque site par $\widehat{R}_i = \sum_{j=1}^{555} \widehat{\theta}_{ij}$.

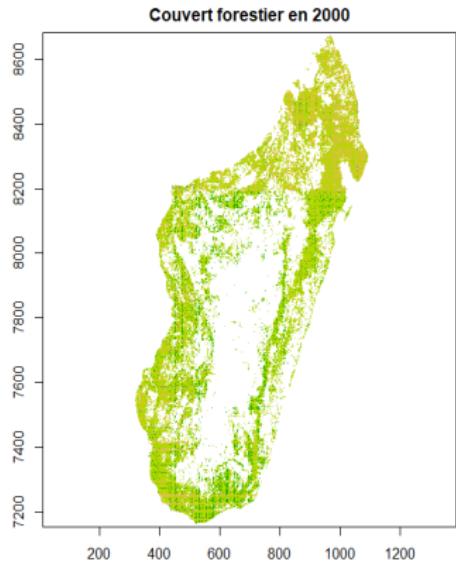


Probabilités de présence issues de l'interpolation par krigeage ordinaire

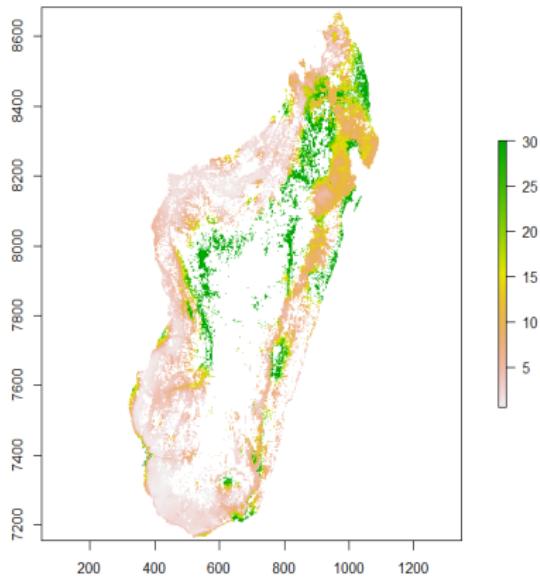


Estimation de la richesse spécifique à Madagascar

Couvert forestier issu de l'article
Vieilledent et al. (2018)



Richesse spécifique estimée restreinte au couvert forestier



Estimation de la diversité β à Madagascar

Diversité β estimée selon la méthode de l'article Allnutt et al. (2008) :

Diversité bêta estimée restreinte au couvert forestier



Différents types de forêt sur l'île d'après l'article Vieilledent et al. (2018) :



Les couleurs identiques représentent des communautés d'espèces similaires.

Conclusion

- **Limitations du package jSDM** : tirage en bloc des effets espèces fixes et prise en compte des contraintes, manière d'évaluer la convergence de l'algorithme, prédiction seulement sur les placettes d'inventaire
- **Perspectives de développement** : indice de Gelman-Rubin pour évaluer la convergence, ajustement de modèles avec traits spécifiques, prédiction sur l'ensemble du territoire considéré en intégrant une auto-corrélation spatiale (CAR ou 2D spline) au modèle.
- **Perspectives écologiques** : utiliser les paramètres estimés et les données climatiques prévisionnelles pour les années 2050 et 2080 afin de mettre en évidence des lieux refuges de la biodiversité.

Références

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>.

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>.



... Merci pour votre attention ...
<https://github.com/JeanneClement/Report>