GAMBAS – Tuesday, June 16th 2020 – Zoom meeting

jSDM R package for Joint Species Distribution Models

Ghislain VIEILLEDENT and Jeanne CLEMENT

Cirad, UMR AMAP, Montpellier, FRANCE
AMAP, **Univ Montpellier**, CIRAD, CNRS, INRAE, IRD, FRANCE

cirad

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

Plan

1. Introduction
   - State of the art
   - Obectives
2. The jSDM R package
   - Joint Species Distribution Models
   - Model specification
   - Rcpp* packages
3. Comparison with boral/JAGS
   - boral R package
   - Data-sets
   - Results
4. Perspectives
   - Additional functionalities
   - SDM vs. JSDM

Plan

1. Introduction
   - State of the art
   - Obectives
2. The jSDM R package
   - Joint Species Distribution Models
   - Model specification
   - Rcpp* packages
3. Comparison with boral/JAGS
   - boral R package
   - Data-sets
   - Results
4. Perspectives
   - Additional functionalities
   - SDM vs. JSDM

Available R packages for JSDMs

Community of coexisting R/Python packages

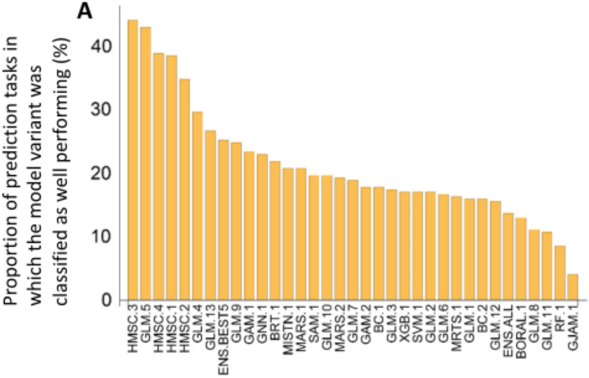
- **boral** (Warton and Hui)
- **HMSC** (Ovaskainen and Tikhonov)
- **gjam** (Clark and Gelfand)
- **BayesComm** (Golding)
- **s-jSDM** (Hartig and Pichler)
- . . .

Wilkinson, D. P. ; Golding, N. ; Guillera-Arroita, G. ; Tingley, R. ; McCarthy, M. A. ; Peres-Neto, P. 2018. A comparison of joint species distribution models for presence-absence data *Methods in Ecology and Evolution*, **10** :198-211. [doi : 10.1111/2041-210x.13106].

Pichler, M. ; Hartig F. 2020. A new method for faster and more accurate inference of species associations from novel community data. *arXiv* pre-print, https://arxiv.org/abs/2003.05331.

Limitations

- Computational speed (boral, HMSC)
- Model specifications (BayesComm, s-jSDM)
  - eg. site random effects, functional traits, phylogenetic data
- Heterogenous model performance (HMSC, boral, gjam)



Norberg et al. 2019

Obectives of the jSDM R package

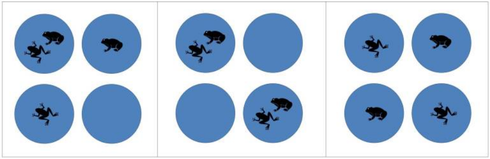
- Make our hands dirty to understand better the JSDM functioning
- Optimized code for fast MCMC computations
- User friendly : package, functions, articles, vignettes
- **A base for testing a large variety of models** :
  - occurrence and count data (Bernoulli/Binomial – Poisson/Neg-Binomial)
    - probit/logit link function for occurrences
    - functional traits and phylogenetic data
    - species and site random/fixed effects
    - presence-only data
- Companion for the `hSDM` R package, hierarchical **one**-species distribution models (mixed models, imperfect detection, spatial autocorrelation) https://ecology.ghislainv.fr/hSDM/

Plan

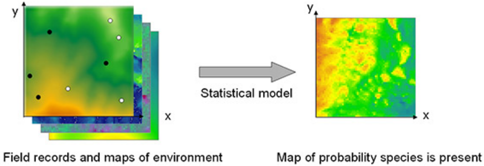
1. Introduction
   - State of the art
   - Obectives
2. The jSDM R package
   - Joint Species Distribution Models
   - Model specification
   - Rcpp* packages
3. Comparison with boral/JAGS
   - boral R package
   - Data-sets
   - Results
4. Perspectives
   - Additional functionalities
   - SDM vs. JSDM

JSDM utility

- Fit species distribution models
- Accounting for species co-occurrences



- Can be used to explain/predict species range and produce species range map



Statistical model

Field records and maps of environment

Map of probability species is present

Data to fit JSDM

- Species presence/absence on sites
- Environmental variables (climate, lancover) at each 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

Statistical model

$$y_{ij} = \begin{cases} 0 & \text{if species } j \text{ is absent on site } i \\ 1 & \text{if species } j \text{ is present on site } i. \end{cases}$$

We assume $y_{ij} \sim \mathcal{B}ernoulli(\theta_{ij})$, with :

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

α_i : site random effects, with $\alpha_i \sim \mathcal{N}(0, V_\alpha)$
X_i : known environmental variables on site i
W_i : latent variables for site i β_j, λ_j : species fixed effects

Latent variables W_i : missing predictors + main axes of covariation across taxa (see Warton et al. 2015 <doi : 10.1016/j.tree.2015.09.007>).

Statistical model

The previous latent variable model (LVM) :

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

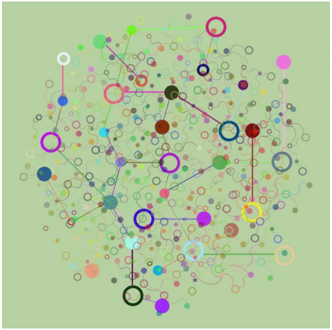
is equivalent to a multivariate probit regression (MPR) :

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

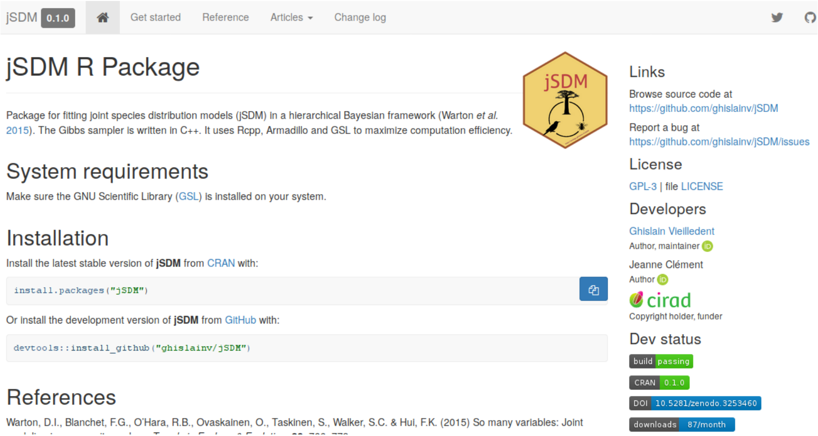
with $u_{ij} \sim \mathcal{N}(0, \Sigma)$ (where Σ is the variance-covariance matrix) and with the constraint that $\Sigma = \Lambda\Lambda'$ (where Λ is the full matrix of factor loadings, with the λ_j as its columns).

Complexity of the model

- Multi-dimensionality : parameters α_i for sites and β_j, λ_j for species
- Non Gaussian process
- Latent-variables W_i
- Mixed model with site random effects $\alpha_i \sim \mathcal{N}(0, V_\alpha)$



jSDM R package



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.

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

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

- https://ecology.ghislainv.fr/jSDM
- Made with Rcpp* packages

Rcpp R package

- **Rcpp** is an R package to extend R with C++ code
- Main advantage : C++ is fast, it accelerates R (see next sections)
- Written by **Dirk EDDELBUETTEL** and **Romain FRANCOIS**
- http://www.rcpp.org/

Simple Rcpp example

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

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

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

R code

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

```
## [1] 4
```

Rcpp advantages

Thanks to `Rcpp::sourceCpp()`

- Compile the C++ code
- Export the function to the R session
- Direct interchange of R objects (including S3, S4) between R and C++
- ... (many more, see `vignette("Rcpp-package")`)

In an R package

- `Rcpp.package.skeleton()` to generate a new Rcpp package (modifying `DESCRIPTION` and `NAMESPACE`)
- `Rcpp::compileAttributes()` scans the C++ files for `Rcpp::export` attributes and generates the code required to make the functions available in R.

GSL and RcppGSL for fast random draws

GNU Scientific Library

- Numerical library for C and C++ programmers
- Reliable random number generator algorithms
- Thoroughly tested and fast random number distributions
- Linear algebra (matrices and vectors)
- https://www.gnu.org/software/gsl/

RcppGSL

- Interface between R and GSL
- Using Rcpp to interface R and C
- http://dirk.eddelbuettel.com/code/rcpp.gsl.html

GSL random number distributions

- GSL v2.6 includes **38 random number distributions** (see GNU GSL)
- It's easy to implement additional random number distributions from the GSL base distributions (e.g. truncated normal distribution)
- For comparison, R API includes "only" 24 random number distributions (see Writing R Extensions)
- Random draws are faster with GSL than with R (eg. `gsl_ran_gamma()` vs. `R::rgamma()`)

Armadillo and RcppArmadillo for high-performance linear algebra

Armadillo

- C++ library for linear algebra and scientific computing
- Provides high-level syntax and functionality : speed and ease of use
- Classes for vectors, matrices and cubes
- Matrix operations, matrix decomposition, linear model solver, etc.
- http://arma.sourceforge.net/

RcppArmadillo

- Interface between R and Armadillo
- Using Rcpp to interface R and C++
- http://dirk.eddelbuettel.com/code/rcpp.armadillo.html

GSL and Armadillo licenses

- Licenses : GNU General Public License, Apache License 2.0 for Armadillo
- Free software licenses : we can use, modify and redistribute these softwares

Plan

1. Introduction
   - State of the art
   - Obectives
2. The jSDM R package
   - Joint Species Distribution Models
   - Model specification
   - Rcpp* packages
3. Comparison with boral/JAGS
   - boral R package
   - Data-sets
   - Results
4. Perspectives
   - Additional functionalities
   - SDM vs. JSDM

boral R package

- R package interfacing R with JAGS for fitting Joint Species Distribution Models
- JAGS is Just Another Gibbs Sampler : http://mcmc-jags.sourceforge.net/
- Approach used in Warton et al. 2015 : <doi : 10.1016/j.tree.2015.09.007>
- boral by Francis K.C. Hui and JAGS by Martyn Plummer

Data-sets

dataset	nsite	nsp	nobs	nX	nlat	npar	nmcmc
Simulated	300	100	30000	2	2	1400	35000
Mosquitos	167	16	2672	13	2	757	35000
Eucalyptus	458	12	5496	7	2	1494	35000
Frogs	104	9	936	3	2	366	35000
Fungi	800	11	8800	12	2	2565	35000

Mosquitos

Eucalyptus

Frogs

Fungi

Comparison results

Compilation time (in minutes)

	Simulated	Mosquitos	Eucalyptus	Frogs	Fungi
boral	96.9	5.8	17.2	1.2	38.6
jSDM	7.0	1.3	1.8	0.3	4.1

jSDM is **4 to 14** times faster than boral/jags.

Root-mean-square error

Computed for probit(θ_{ij}) with the simulated data-set.

	boral	jSDM
RMSE	1.8	0.6

Deviance

	Simulated	Mosquitos	Eucalyptus	Frogs	Fungi
boral	40486	6936	8779	884	12871
jSDM	15651	1231	1922	150	1982

Conclusion

- Small data-sets **and** simple models : R, *BUGS, JAGS, Stan, INLA, MCMCglmm
- Large data-sets **or** complex hierarchical models : R + Rcpp + RcppGSL + RcppArmadillo

- With Rcpp* packages, the Gibbs sampler can typically be written in about half a day
- Code is reusable and easily packageable
- Tools with incomparable efficiency for statisticians

Plan

1. Introduction
   - State of the art
   - Obectives
2. The jSDM R package
   - Joint Species Distribution Models
   - Model specification
   - Rcpp* packages
3. Comparison with boral/JAGS
   - boral R package
   - Data-sets
   - Results
4. Perspectives
   - Additional functionalities
   - SDM vs. JSDM

Additional functionalities

- Count data (Poisson/Negative-Binomial)
- Logit link function for occurrences
- Functional traits and phylogenetic data
- Species and site random/fixed effects
- Presence-only data
- Spatial autocorrelation for α_i and W_i

SDM vs. JSDM

See notebook

. . . Thank you for attention . . .

🐦 @ghislainv

https://ecology.ghislainv.fr/presentations

ghislain.vieilledent@cirad.fr | jeanne.clement16@laposte.net