

Modéliser la distribution spatio-temporelle d'espèces à partir de données opportunistes

F. Lasgorceux, J. Papaïx,
D. Combrisson, Y. Bunz, T. Opitz.



RÉPUBLIQUE
FRANÇAISE

*Liberté
Égalité
Fraternité*

INRAE

BIO/P

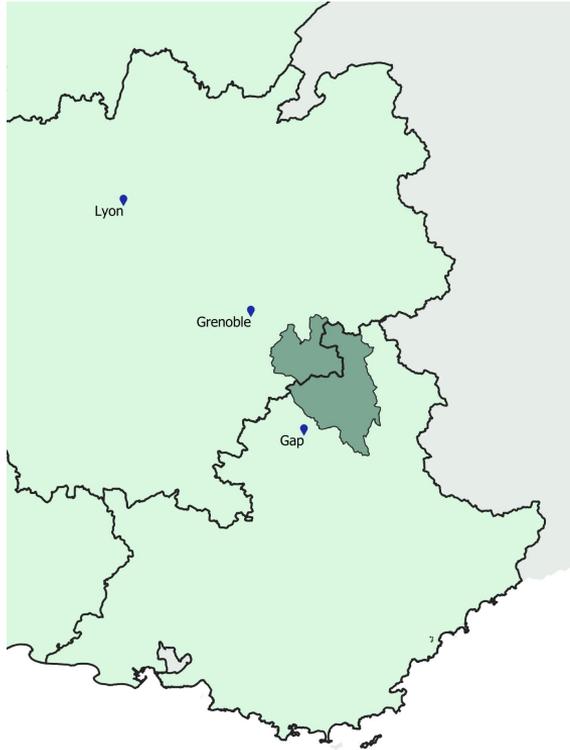


Parc national
des Ecrins

Sommaire

1. Contexte de la thèse
2. Etude de la distribution jointe de deux espèces
 - a. Modèles de Cox log-gaussien
 - b. Inférence bayésienne via l'approche INLA-SPDE
 - c. Plan d'expérience numérique
 - d. Application aux données des Ecrins
3. Perspectives et suite des travaux

Le parc national des Ecrins



- Aire d'adhésion : 159 600 ha.
- Altitude du parc : entre 700 et 4 102 m.
- 2 500 espèces végétales, plus de 350 espèces de vertébrés.

Prédire la distribution des espèces pour les mesures de gestion (Guisan et *al.* 2013)

Objectifs :

Cartographier les aires de répartition selon le climat et l'occupation / usage des sols



Etudier les réponses des espèces au changement climatique/utilisation des sols.



Créer des zones protégées pour les espèces menacées (ex: COP15).



Limiter les effets d'une espèce invasive.



Ce qu'apportent les Species Distribution Models (SDM) :

Permet d'apporter une réponse quantitative.

Quantifier des décalages de phénologie, des remontées en altitude.

Identifier les habitats les plus favorables et les zones les plus riches en terme de diversité.

Prédire les zones à risque.

Données opportunistes VS données protocolées

	Opportunistes	Protocolées
Professionnels	Données du PNE	Données taxonomiques
Citoyens	PI@ntNet, eBird	Breeding Bird Survey, STOC

Avantages

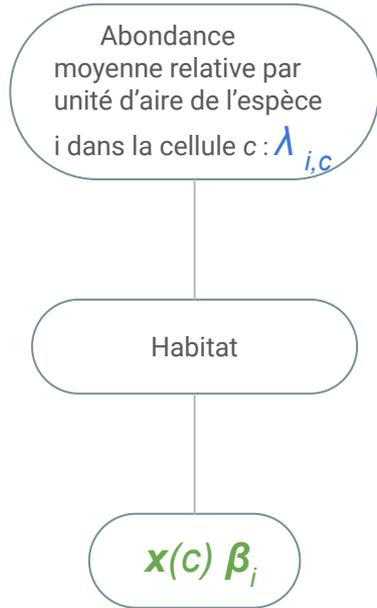
- Coût d'acquisition faible → grand volume de données.
- Couverture spatiale inégalable.
- Détection de comportements inhabituels (phénologie).
- Pas d'erreur d'identification (hypothèse).

Inconvénients

- Effort d'échantillonnage inconnu et hétérogène dans l'espace et dans le temps.
- Hétérogénéité d'acquisition chez les observateurs.
- Hétérogénéité d'acquisition selon les taxons.

2. Etude de la distribution jointe de deux espèces
 - a. Modèles de Cox log-gaussien

Utilisation des processus de Poisson inhomogènes



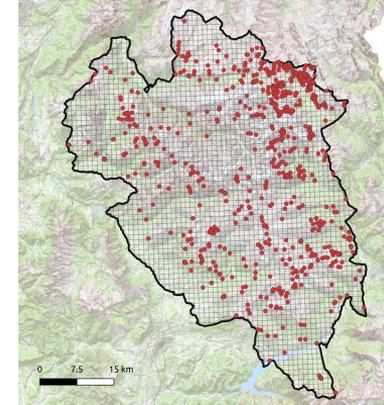
Processus de Poisson inhomogène

$$Y_i(s) \sim \text{Pois}(\lambda_i(s))$$

Hypothèse 1 : $\forall s \in c, \lambda_i(s) = \text{cste} = \lambda_{i,c}$

$$Y_{i,c} \sim \text{Pois}(|c| \lambda_{i,c})$$

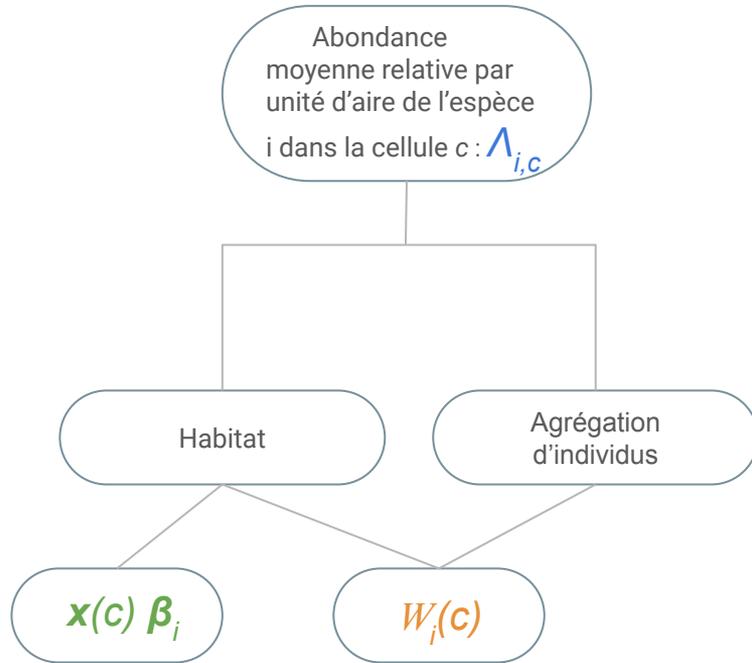
$$\log(\lambda_{i,c}) = \mathbf{x}(c) \beta_i$$



Hypothèse 2 : les covariables disponibles expliquent totalement $\lambda_{i,c}$ et les points d'observations sont indépendants les uns des autres conditionnellement à $\lambda_{i,c}$ (Renner et al. 2015).

Limites : si covariables significatives non mesurées, dispersion/agrégation d'individus, échantillonnage préférentiel, interactions entre espèces.

Utilisation des processus de Cox log-gaussien



Ajout d'une composante stochastique spatiale dans le terme de régression

$$Y_{i,c} | \Lambda_{i,c} \sim \text{Pois}(|c| \Lambda_{i,c})$$

$$\log(\Lambda_{i,c}) = x(c) \beta_i + W_i(c)$$

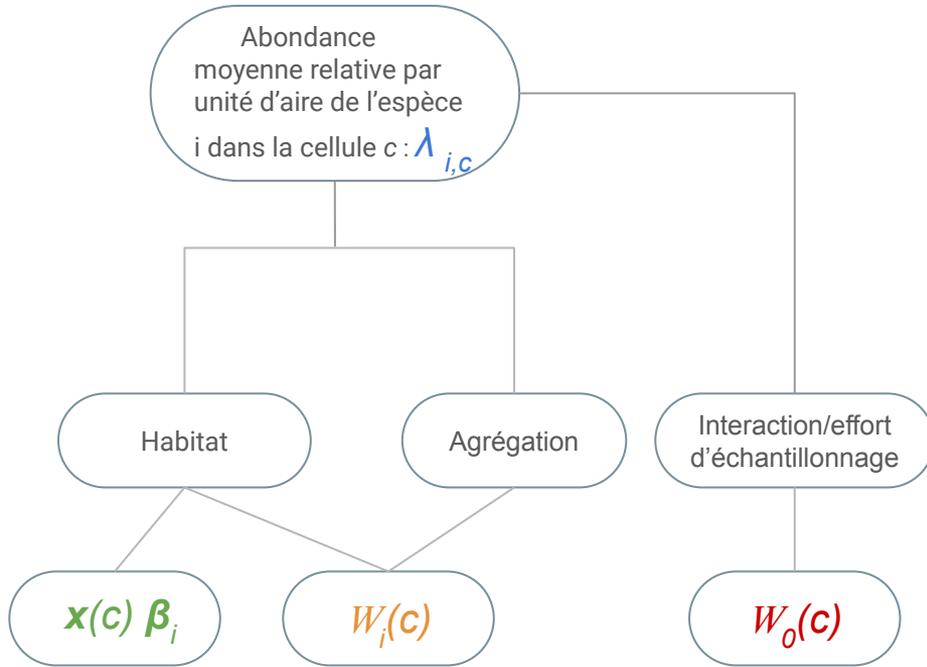
W_i est un champ gaussien latent de moyenne nulle, de fonction de covariance Matérn.

W_i dépend des hyperparamètres :

- σ^2 la variance
- ρ le range (paramètre d'échelle)
- ν paramètre de lissage (fixé à 1)

Limites : interaction entre espèces, échantillonnage préférentiel.

Utilisation des processus de Cox log-gaussien bivariés



Ajout d'une composante stochastique spatiale partagée dans les termes de régression :

Modèle W_0, W_1, W_2

$$Y_{i,c} | \Lambda_{i,c} \sim \text{Pois}(|c| \Lambda_{i,c}) \text{ pour } i = 1, 2.$$

$$\log(\Lambda_{1,c}) = \mathbf{x}(c) \beta_1 + W_1(c) + W_0(c)$$

$$\log(\Lambda_{2,c}) = \mathbf{x}(c) \beta_2 + W_2(c) + \gamma W_0(c)$$

γ est un paramètre de corrélation croisée entre les répartitions des deux espèces.

Limites : identifiabilité des différents champs gaussiens et de γ .

2. Etude de la distribution jointe de deux
espèces
b. Inférence bayésienne via l'approche
INLA-SPDE

Inférence Bayésienne à l'aide d'INLA (Lindgren et *al.* 2015)

INLA permet de calculer directement les densités *a posteriori* (\neq méthodes MCMC) pour les modèles gaussiens latents :

$$(\boldsymbol{\theta}) \sim p(\boldsymbol{\theta})$$

$$(\mathbf{x}|\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1})$$

$$\eta_{i,j} = \sum_j c_{i,j} x_j$$

$$(y_i|\mathbf{x}, \boldsymbol{\theta}) \sim p(y_i|\eta_i, \boldsymbol{\theta})$$

$\boldsymbol{\theta}$ sont les (hyper)paramètres, \mathbf{x} le champ gaussien latent, \mathbf{Q} une matrice creuse, $\boldsymbol{\eta}$ le prédicteur linéaire et \mathbf{y} est le vecteur des données.

Le prédicteur linéaire du processus de Cox-log gaussien $Y_{i,c}$:

$$\log(\Lambda_{i,c}) = \mathbf{x}(c) \boldsymbol{\beta}_i + W_i(c) + W_0(c)$$

doit suivre une loi gaussienne multivariée avec une matrice de précision creuse.

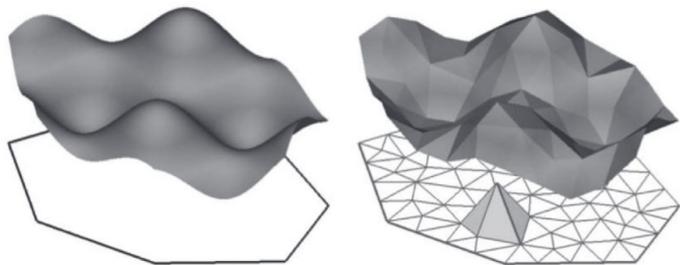
On met une **loi *a priori* gaussienne de matrice de précision diagonale** (indépendance) sur les paramètres des effets fixes $\boldsymbol{\beta}_i$.

Quid de $W_i(c)$ et $W_0(c)$?

Représentation des W_j via l'approche SPDE (Lindgren et al. 2011)

W_1 , W_2 et W_0 sont des champs gaussiens de fonction de covariance de Matérn.

Idée : travailler avec des Gaussian Markov Random Fields (GMRF) qui approchent les W_j



Blangiardo et al. (2013)

$$W_j(s) = \sum_{k=1}^{L_j} \psi_{j,k}(s) w_{j,k}$$

Les $\psi_{j,k}$ sont les éléments finis fixés grâce à la triangulation de l'espace.

Or, W_j est solution d'une équation aux dérivées partielles stochastiques (SPDE).

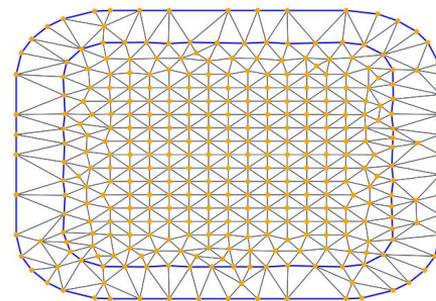
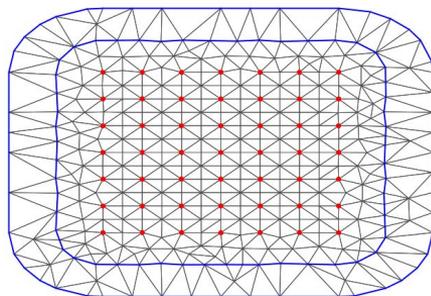
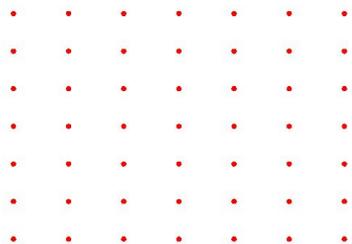
Il existe $(w_{j,1}, \dots, w_{j,L_j}) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_j^{-1})$ approchant la solution de cette SPDE, avec \mathbf{Q}_j une matrice de précision creuse.

2. Etude de la distribution jointe de deux espèces

c. Plan d'expérience numérique

Simulation des observations

Pts d'observations \longrightarrow Triangulation de l'espace \longrightarrow Simulation des champs W_j



Estimation des
(hyper)paramètres :

$\beta_1, \beta_2, \sigma^2_1, \rho_1, \sigma^2_2,$
 $\rho_2, \sigma^2_0, \rho_0, \gamma$

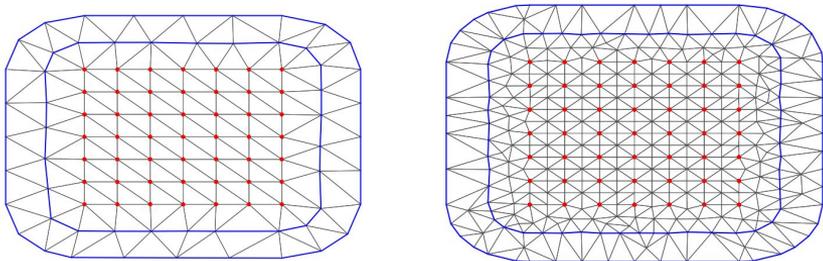
Observations =
Tirages des lois de
Pois($|c| \Lambda_{i,c}$)

Calculs de $\log(\Lambda_{1,c})$
et $\log(\Lambda_{2,c})$

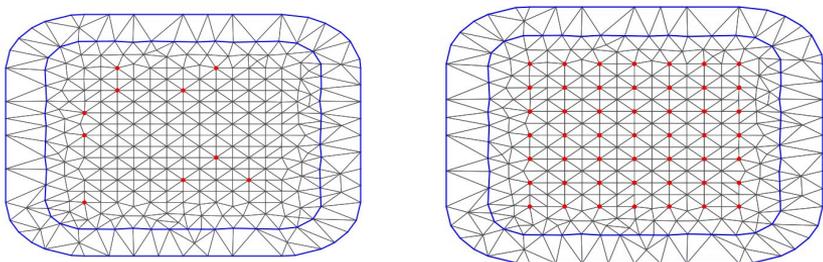
Scénarios de simulation

Sensibilité des estimations vis à vis des paramètres suivants :

- 3 paramètres de précision de mesh



- 3 nombres d'observations



- $\gamma \in \{-1, 0, 1\}$.
- 2 modèles :

Modèle W_0, W_1, W_2

$$Y_{i,c} | \Lambda_{i,c} \sim \text{Pois}(|c| \Lambda_{i,c}) \text{ pour } i = 1, 2.$$

$$\log(\Lambda_{1,c}) = \mathbf{x}(c) \beta_1 + W_1(c) + W_0(c)$$

$$\log(\Lambda_{2,c}) = \mathbf{x}(c) \beta_2 + W_2(c) + \gamma W_0(c)$$

Modèle W_0 -only

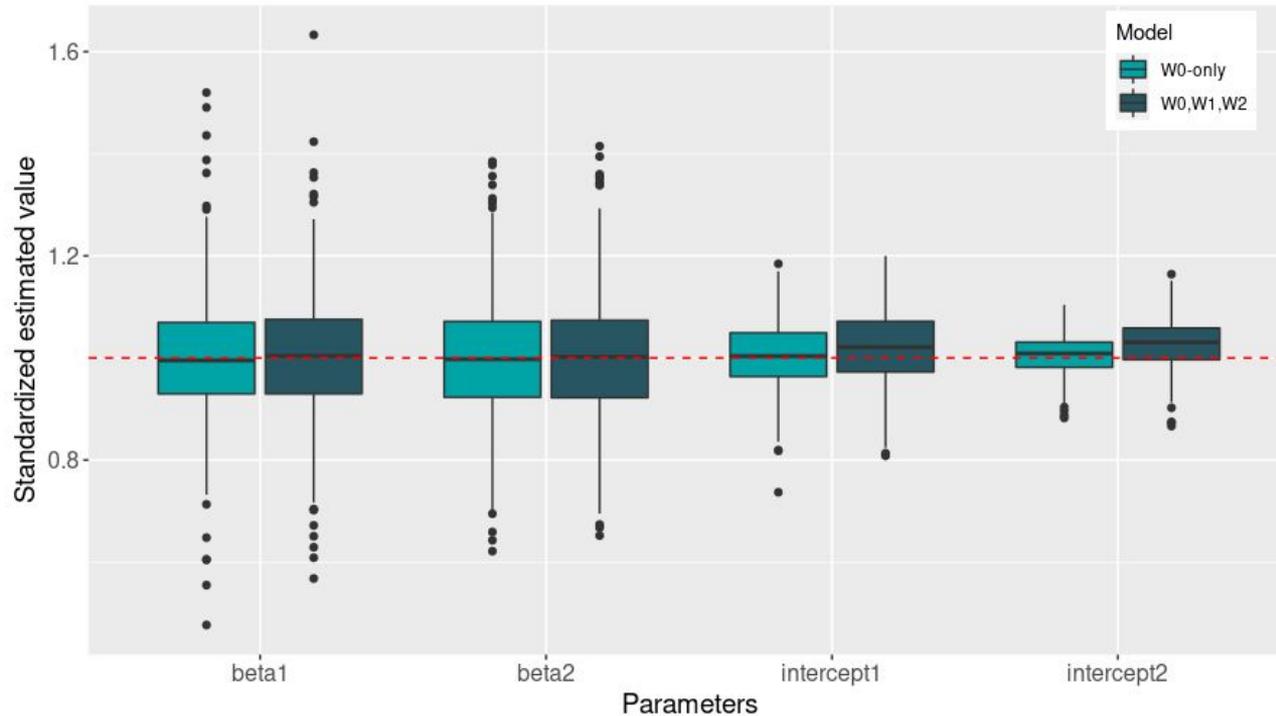
$$Y_{i,c} | \Lambda_{i,c} \sim \text{Pois}(|c| \Lambda_{i,c}) \text{ pour } i = 1, 2.$$

$$\log(\Lambda_{1,c}) = \mathbf{x}(c) \beta_1 + W_0(c)$$

$$\log(\Lambda_{2,c}) = \mathbf{x}(c) \beta_2 + \gamma W_0(c)$$

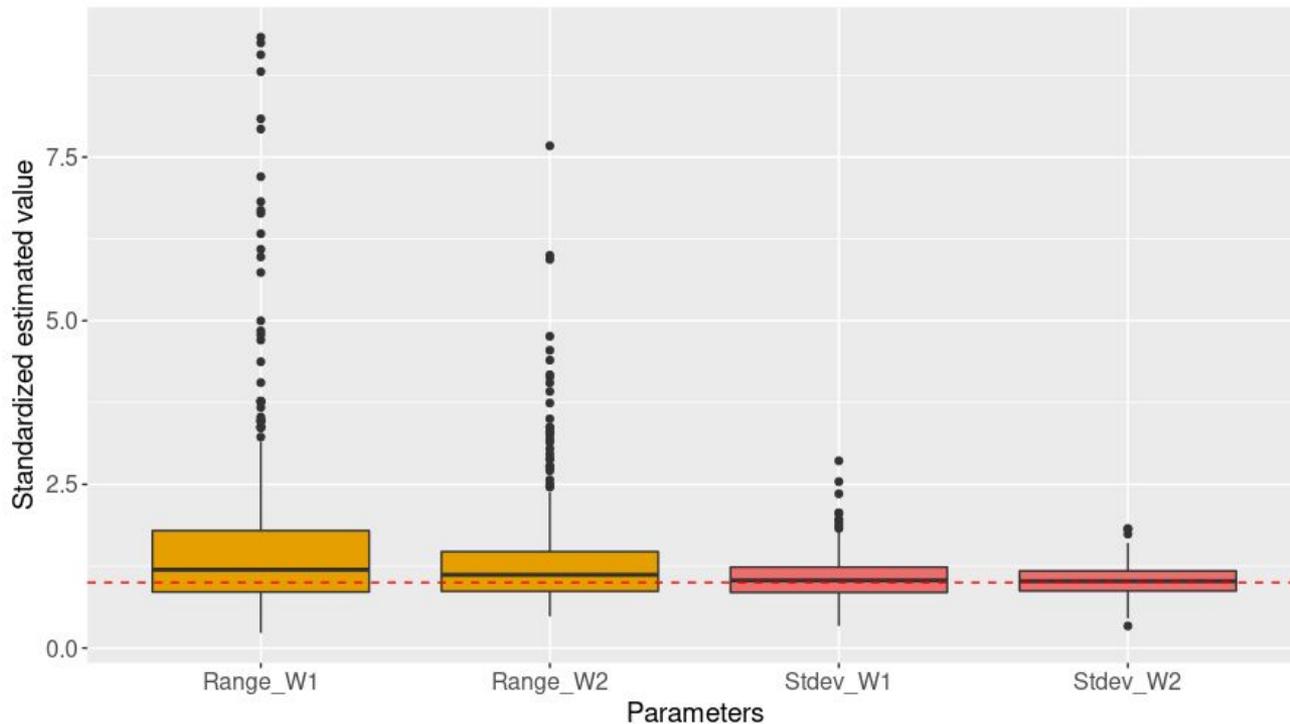
Estimation des effets fixes

Boxplots $\hat{\theta}/\theta$ obtenus à partir de $3 \times 3 \times 50 = 450$ simulations



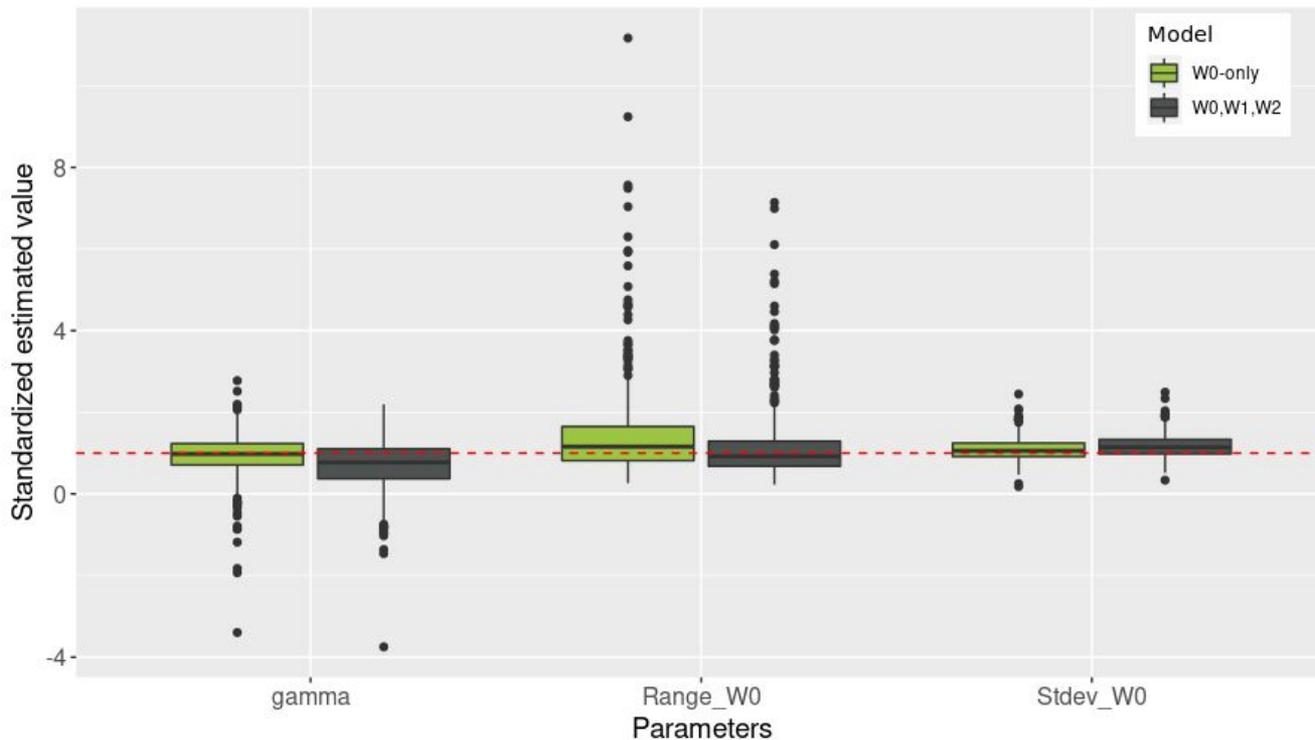
Estimation des hyperparamètres de W_1 et de W_2

Boxplots $\hat{\theta}/\theta$ obtenus à partir de $3 \times 3 \times 50 = 450$ simulations



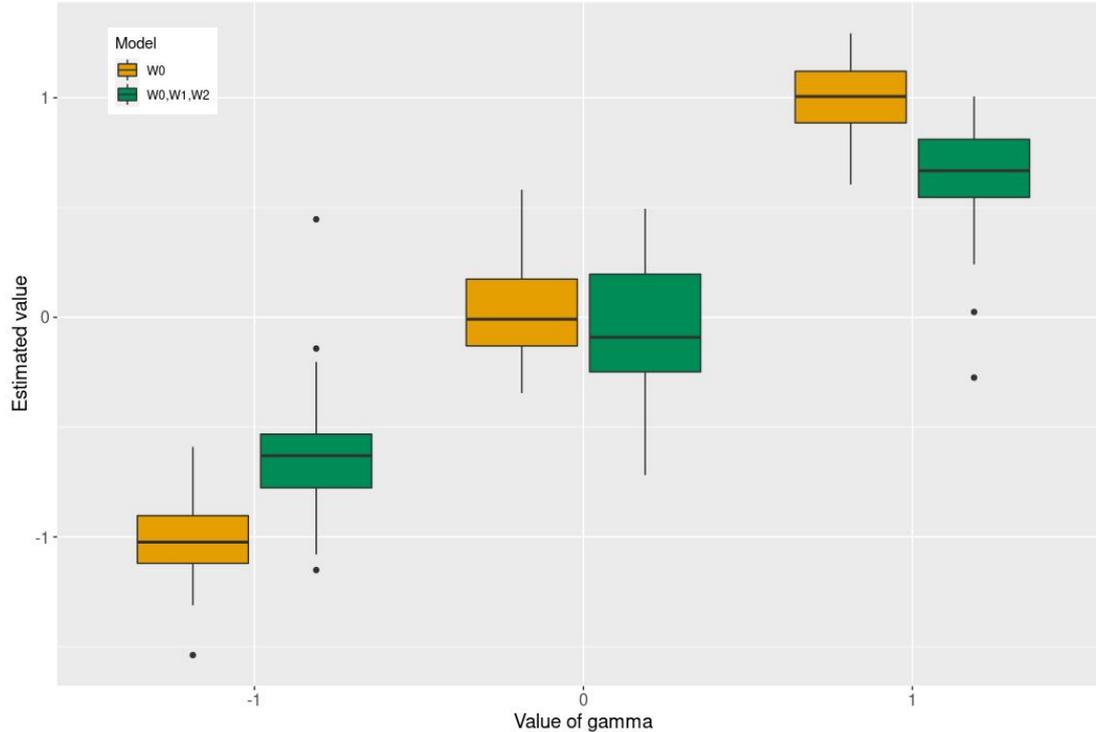
Estimation des hyperparamètres de W_0 et de γ

Boxplots $\hat{\theta}/\theta$ obtenus à partir de $3 \times 3 \times 50 = 450$ simulations



Estimation de γ

Boxplots $\hat{\theta}$ obtenus à partir de $3 \times 50 = 50$ simulations



2. Etude de la distribution jointe de deux espèces

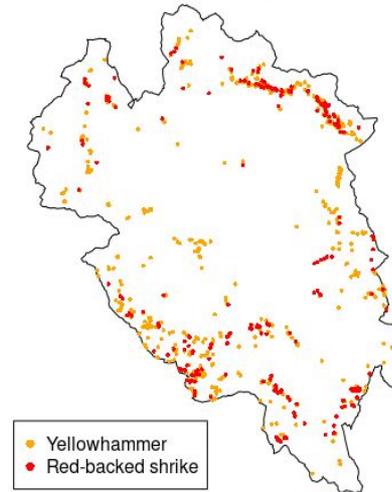
d. Application aux données des Ecrins

Application à des données de passereaux du PNE

- 722 observations de bruant jaune (*Emberizza citrinella*)
- 356 observations de pie-grièche écorcheur (*Lanius collurio*)
- Covariables = agrégation des surfaces de sols dans chaque pixel issues de la classification OSO (5mx5m) + altitude et pente normalisées entre 0 et 1.

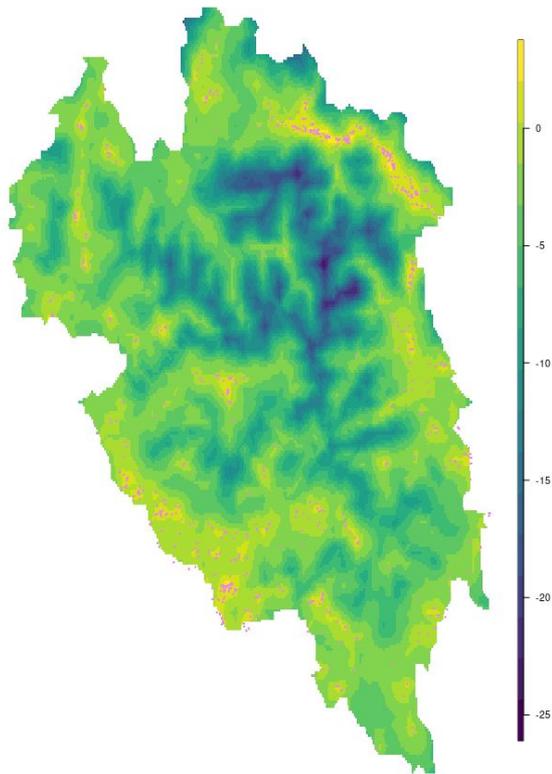


Observations between 2013 and 2020

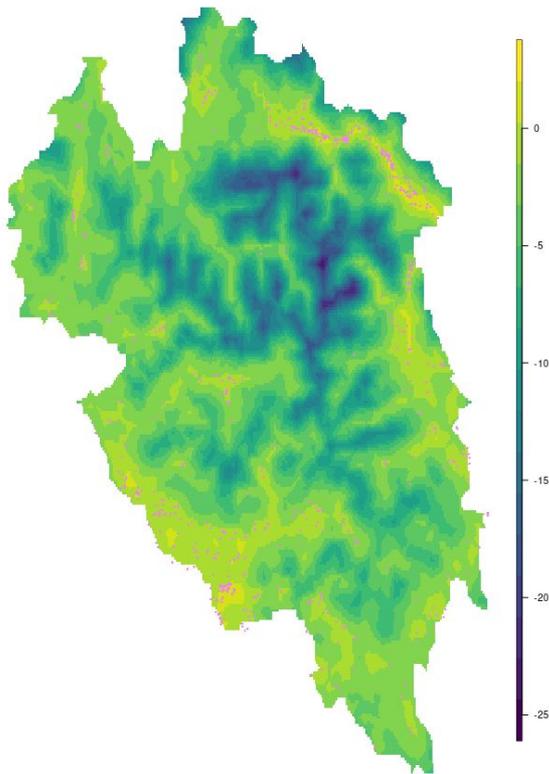


Champs prédits pour le bruant jaune

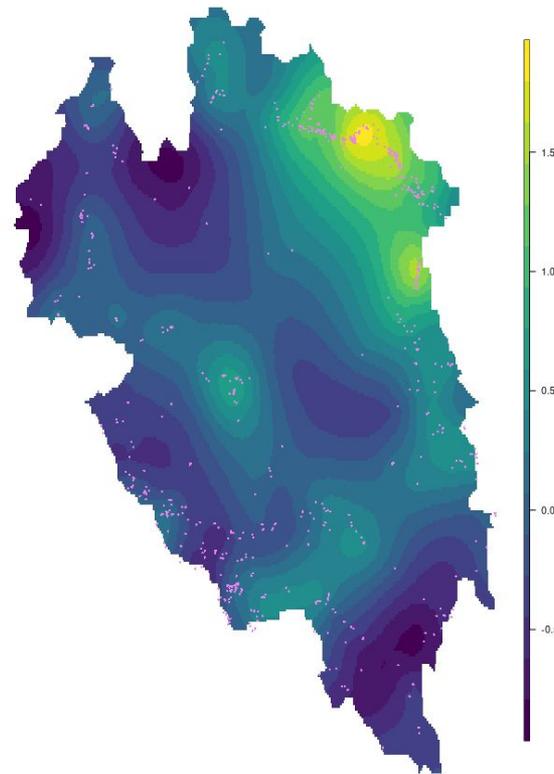
Predicted distribution ($X\beta_1 + W_1 + W_0$)



Predicted ecological niche ($X\beta_1 + W_1$)

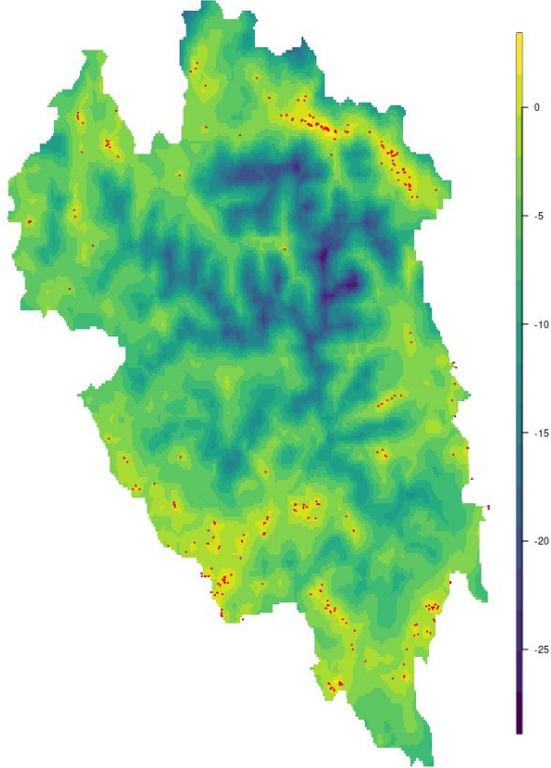


Specific field (W_1)

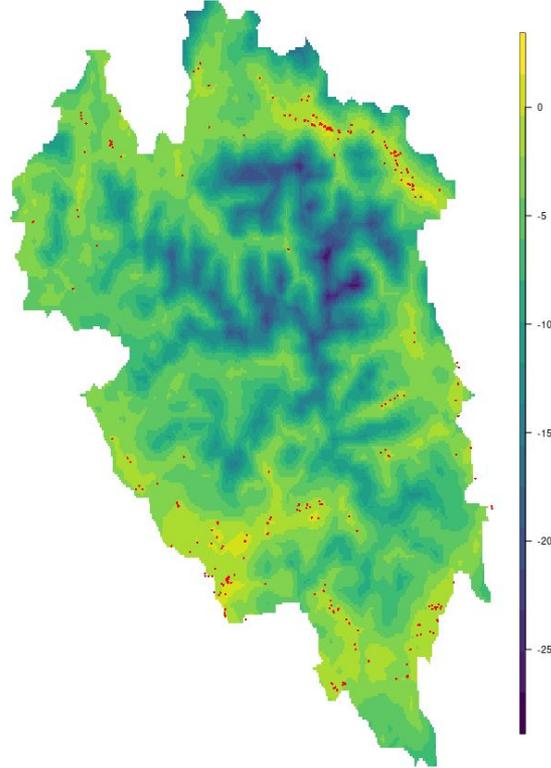


Champs prédits pour la pie-grièche écorcheur

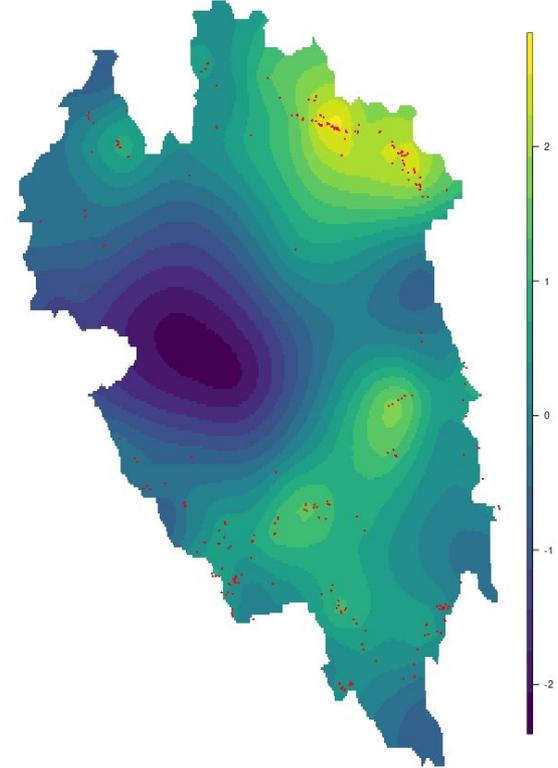
Predicted distribution ($X\beta_2 + W_2 + W_0$)



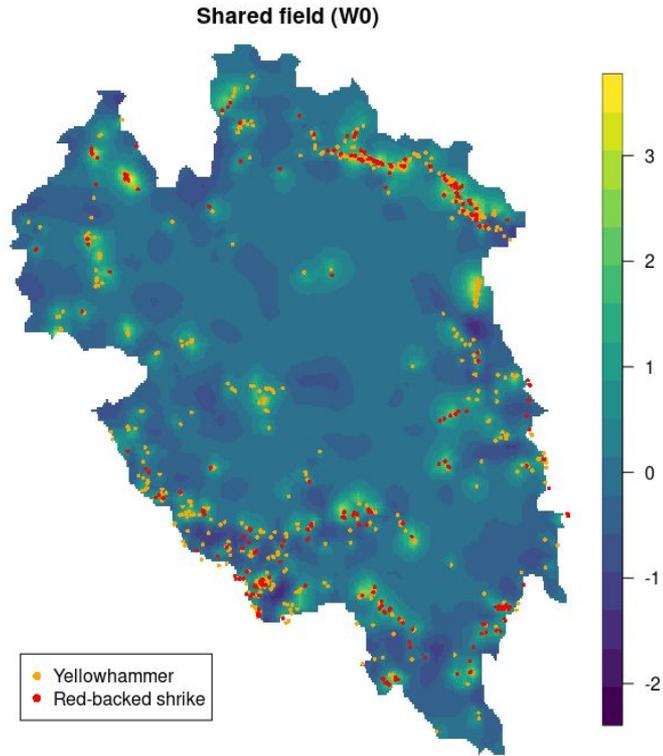
Predicted ecological niche ($X\beta_2 + W_2$)



Specific field (W_2)



Champ partagé prédit et hyperparamètres



Hyperparamètre	Mean	IC [0.025,0.975]
Range W_1	22.2	[6.1, 65.7]
Stdev W_1	0.8	[0.5, 1.2]
Range W_2	20.0	[5.7, 46.6]
Stdev W_2	1.5	[0.8, 2.5]
Range W_0	2.3	[1.7, 3.0]
Stdev W_0	1.3	[1.1, 1.5]
γ	1.2	[0.9, 1.4]

3. Perspectives et suites des travaux

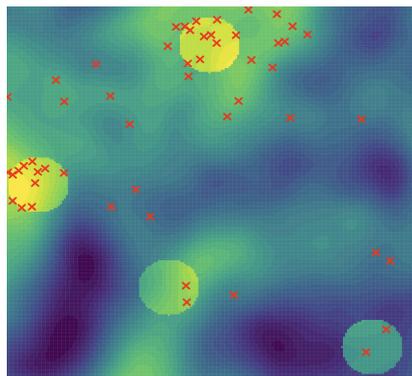
Perspectives sur les modèles multi-espèces

- Extraire plus d'informations du champ partagé : peut-on séparer l'effort d'échantillonnage des interactions entre espèces? Coupler les données opportunistes avec des données protocolées / utiliser des proxys de l'effort.
- Utiliser ce modèle pour d'autres taxons (mollusques).

- Généraliser le modèle à un grand nombre d'espèces.
- Réaliser des assemblages d'espèces à partir des paramètres écologiques estimés à l'aide des tirages *a posteriori* des lois estimées univariées.

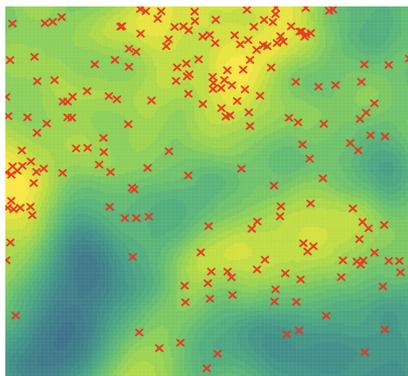
Perspectives sur l'effort d'échantillonnage

Observations



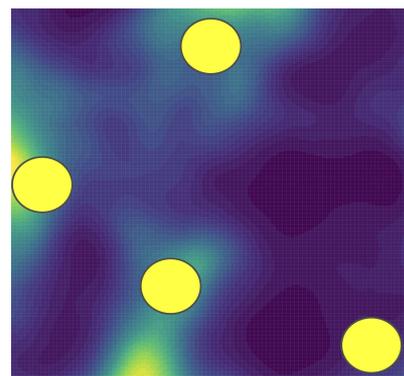
=

Abondance réelle



x

Effort d'échantillonnage



$$Y_i^{obs}(s,t) \mid \Lambda_i^{obs}(s,t) \sim \text{Poisson}(\Lambda_i^{obs}(s,t))$$

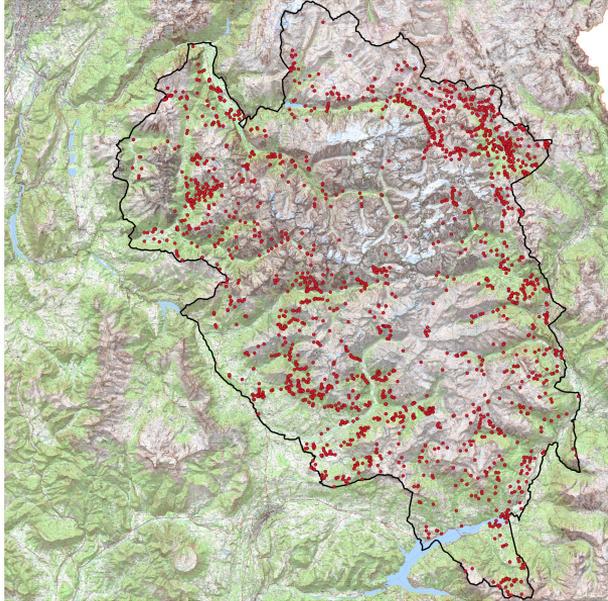
$$\text{Hypothèse 1 : } \Lambda_i^{obs}(s,t) = \Lambda_i(s,t) E_i(s,t)$$

$$\log(\Lambda_i^{obs}(s,t)) = \log(\Lambda_i(s,t)) + \log(E_i(s,t))$$

$$= \eta_i(s,t ; \text{prédicteurs}) + e_i(s,t ; \text{prédicteurs})$$

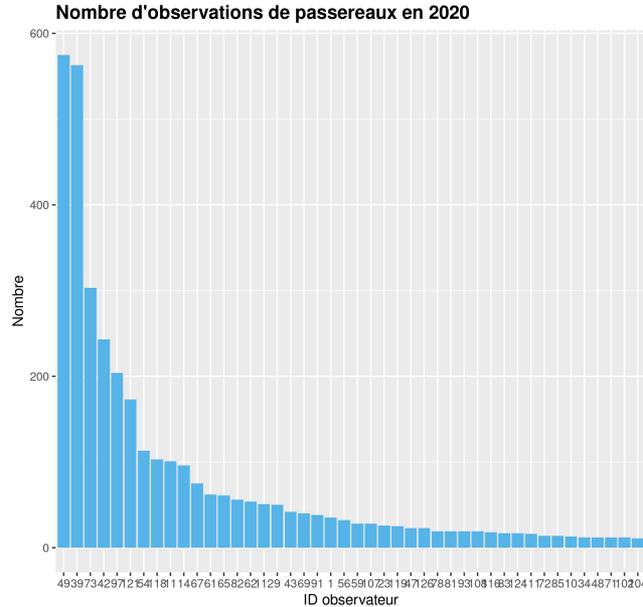
Identification des biais induits par l'effort d'échantillonnage

Biais de prospection spatio-temporel



→observateur

Biais de reporting



→observateur

Biais taxonomique



→taxonomique
et observateur

Biais de détectabilité



→taxonomique et milieu

Merci pour votre attention !



Références

- Guisan, A., et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424-1435. <https://doi.org/10.1111/ele.12189>
- Renner, I.W., et al. (2015). Point process models for presence-only analysis. *Methods Ecol Evol*, 6: 366-379. <https://doi.org/10.1111/2041-210X.12352>
- Blangiardo, M., et al. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*. 4C. 33-49. <https://doi.org/10.1016/j.sste.2013.07.003>
- Lindgren, F., et al. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73: 423-498. <https://doi.org/10.1111/j.1467-9868.2011.00777.x>
- Finn, Lindgren., Håvard, Rue. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63(1):1-25. <https://doi.org/10.18637/jss.v063.i19>