Parameter identifiability in bivariate spatiotemporal point processes for ecological presence-only data under incomplete sampling

Florian Lasgorceux [1] & Julien Papaïx [1] & Yoann Bunz [2] & Damien Combrisson [2] & Thomas Opitz [1]

[1] *INRAE - Unité BioSP, 228 route de l'Aérodrome, CS 40509, Domaine Saint-Paul - Site Agroparc, 84914 Avignon Cedex 9.*
florian.lasgorceux@inrae.fr , thomas.opitz@inrae.fr, julien.papaix@inrae.fr.
[2] *Parc national des Ecrins, Domaine de Charance, 05000 Gap.*
yoann.bunz@ecrins-parcnational.fr , damien.combrisson@ecrins-parcnational.fr

Résumé. Les gestionnaires d'espaces naturels ont de plus en plus recours à des bases de données de comptage d'espèce alimentées de manière opportuniste (*données de présence seule*). Ces données offrent une meilleure couverture spatio-temporelle mais sont affectées de multiples biais d'acquisition du fait de l'absence de protocole d'observation. Ceci se traduit en particulier par un effort d'échantillonnage hétérogène et inconnu. De plus, la mise en évidence des interactions entre espèces à partir de ce type de données n'est pas aisée puisque la co-occurrence d'observations ne présage pas forcément d'une influence d'une espèce sur l'autre. Dans ce contexte, nous nous sommes intéressés à la modélisation jointe de la distribution d'espèces et à l'inférence d'un tel modèle à partir de données opportunistes. Nous développons, dans un premier temps, un modèle de processus ponctuel bivarié (pour deux espèces), à savoir un processus de Cox log-gaussien, dont les intensités intègrent des covariables environnementales et des effets aléatoires spatiaux, dont un commun aux deux variables. L'identifiabilité des paramètres de ce modèle est étudiée par simulations. Dans un deuxième temps, un effort d'échantillonnage hétérogène en espace est ajouté au modèle et nous discutons des conséquences du choix de modélisation de ce dernier. Les simulations et l'estimation des paramètres du modèle sont réalisées dans un cadre bayésien via la méthode INLA-SPDE. L'estimation est explorée pour différentes configurations des paramètres liés à l'abondance et la variabilité spatiale de chaque espèce, à l'interaction entre les deux espèces et à la structure de l'effort d'échantillonnage. Le cadre de simulation est adapté au contexte du Parc National des Ecrins pour lequel des données opportunistes des communautés aviaires sont disponibles.

Mots-clés. Processus de Cox log-gaussien, INLA, données de présence seule, effort d'échantillonage, données opportunistes, écologie des communautés, modèle de distribution d'espèce.

Abstract. Managers of natural parcs and protected areas are increasingly using databases of species counts, which are supplied in an opportunistic way (presence-only data). These data guarantee a better spatio-temporal coverage than available protocoled data but suffer from strong biases due to the lack of a proper observation protocol, which leads

to a heterogeneous and unknown sampling effort. Moreover, the study of species interactions through this kind of data is arduous since empirical co-occurrence patterns do not necessarily imply an influence of one species on another. Our study deals with the joint modeling of species distributions and the inference of such models from opportunistic data. We firstly establish a bivariate (for two species) point process model constructed as a log-gaussian Cox process. The intensity functions of the two species take into account environmental covariates and spatial random effects, including one that is common to both variables to account for species interaction. Our main goal here is to study the identifiability of parameters using simulations. Moreover, a sampling effort that may be heterogeneous in space is added to the model, and we discuss the consequences of modeling the latter. The simulations and the estimation of the model parameters are performed through Bayesian inference using the INLA-SPDE approach. The simulation framework is adapted to the Ecrins National Park, for which opportunistic data of bird communities are available.

Keywords. Log-Gaussian Cox Process, INLA, presence-only data, sampling effort, opportunistic data, ecological community, species distribution models.

1 Contexte d'application

Les modèles de distribution d'espèces permettent de prédire, dans l'espace et dans le temps, la présence et l'abondance d'espèces animales ou végétales. Ils sont donc un pré-requis à la compréhension des dynamiques écologiques et de leurs réponses aux changements climatiques et d'usage des sols. De plus, ils servent de base à la mise en place de nouvelles mesures de gestion ainsi qu'à l'évaluation de mesures déjà en place.

Du fait de l'étendue spatiale et temporelle nécessaire, les données permettant d'ajuster de tels modèles sont généralement fortement hétérogènes, les protocoles changent suivant les groupes taxonomiques, les lieux ou les périodes, et le nombre d'observateurs est grand. Parmi ces données, les gestionnaires d'espaces naturels ont de plus en plus recours aux données de contact opportuniste. Ces données, non protocolées, renseignent la présence d'un individu d'une espèce donnée en un lieu et un temps donné. L'avantage majeur de telles bases de données alimentées de manière opportuniste provient de leur importante couverture spatiale, de leur quasi-continuité temporelle et de l'information sur les comportements inhabituels (marge d'aire de répartition, décalage phénologique). Cependant, les données opportunistes sont difficiles à exploiter car fortement affectées par de multiples biais d'acquisition, en particulier l'échantillonnage inhomogène et préférentiel dans les dimensions de l'espace, du temps et des différentes espèces.

Dans ce travail, nous cherchons à caractériser la dynamique spatio-temporelle jointe de deux espèces à partir de données opportunistes. Nous développons un modèle de Cox log-Gaussien bivarié dont les intensités intègrent des covariables envionnementales et des

effets aléatoires spatiaux. Un des effets spatiaux est commun aux deux espèces et permet d'étudier les possibles interactions entre celles-ci. L'identifiabilité des paramètres est étudiée par simulations. Nous considérons ensuite un effort d'échantillonnage hétérogène en espace et nous discutons des conséquences du choix de modélisation de ce dernier.

2 Modélisation de la répartition spatiale de deux espèces

2.1 Du processus de Poisson au processus de Cox log-gaussien

L'ensemble des positions des individus d'une espèce observée dans une fenêtre d'observation spatiale peut être vu comme un semis de points aléatoires, et ce nuage de points peut être modélisé par un processus ponctuel. Le nombre d'individus se situant dans une unité de l'espace est un nombre entier aléatoire à valeurs dans $\mathbb{N} = \{0, 1, 2, \ldots\}$, et les nombres d'individus obtenus pour une partition de la fenêtre d'observation peuvent être modélisés via un processus de comptage. Les processus de comptage sont caractérisés par leur fonction d'intensité (ou simplement intensité) qui est définie comme le nombre moyen d'observation par unité de surface. L'intensité sera notée $\lambda(s)$ où $s \in \Omega \subset \mathbb{R}^2$ est point de l'espace. Comme le rappelle Renner et *al* (2015), le processus de comptage le plus naturel est le processus de Poisson inhomogène qui repose sur les hypothèses suivantes :

(i) Les points d'observations sont indépendants les uns des autres.

(ii) L'intensité $\lambda(s)$ varie dans l'espace.

L'hypothèse (ii) est souvent réduite à la modélisation de l'intensité par une fonction log-linéaire des covariables environnementales :

$$\ln(\lambda(s)) = \boldsymbol{x}(s)\boldsymbol{\beta},$$

où $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ est le vecteur des paramètres correspondant aux p covariables environnementales $\boldsymbol{x}(s) = (x_1(s), \ldots, x_p(s))$. On notera que la fonction de lien logarithme assure la positivité de l'intensité.
L'hypothèse (i), trop restrictive pour nos données de biodiversité, peut être contournée en modélisant les tendances d'agrégation ou de fragmentation des observations au delà de structures spatiales induites par les covariables à l'aide d'un champ aléatoire. On appelle un tel processus un processus de Cox. Ainsi, conditionnellement à la réalisation de ce champ, les observations forment un processus de Poisson inhomogène. Le processus de Cox qui nous intéresse a été proposé par Møller et *al.* (1998) et est appelé processus de Cox log-Gaussien (*Log-Gaussian Cox process*, LGCP en abrégé). Concrètement, l'intensité que l'on notera Λ puisqu'elle est aléatoire, est toujours fonction des covariables environnementales mais aussi désormais d'un champ aléatoire gaussien W :

$$\ln(\Lambda(s)) = \boldsymbol{x}(s)\boldsymbol{\beta} + W(s).$$

Ici, W est un champ gaussien de moyenne nulle et de fonction de covariance fixée arbitrairement, évalué en les points $s \in \Omega$. Le champ W décrit la variabilité, structurée dans l'espace, non expliquée par les covariables environnementales.

2.2 Modèle mis en oeuvre dans cette étude

Nous avons fait le choix de modéliser l'intensité du processus d'observation de l'espèce $i \in \{1, 2\}$, notée Λ_i^{obs}, comme le produit d'un effort d'échantillonage dépendant de l'espace et de l'espèce $e_i(s)$ et d'un champ aléatoire, $\Lambda_i(s)$ représentant l'abondance de l'espèce i. C'est-à-dire,

$$\Lambda_i^{obs}(s) = e_i(s)\Lambda_i(s). \tag{1}$$

Le champ aléatoire Λ_i est fonction des covariables environnementales et de deux champs gaussiens, afin d'exprimer la variabilité spatiale non expliquée en une partie propre à l'espèce, et une autre issue de l'interaction entre les espèces. Si l'on note Λ_1 (*resp.* Λ_2) l'intensité du processus associé à l'espèce 1 (*resp.* à l'espèce 2), W_1 (*resp.* W_2) le champ gaussien propre à l'espèce 1 (*resp.* à l'espèce 2), et W_0 le champ gaussien partagé par les deux espèces, on obtient :

$$\ln(\Lambda_1(s)) = \boldsymbol{x}(s)\boldsymbol{\beta}_1 + W_1(s) + W_0(s), \tag{2}$$
$$\ln(\Lambda_2(s)) = \boldsymbol{x}(s)\boldsymbol{\beta}_2 + W_2(s) + \gamma W_0(s), \tag{3}$$

où γ (équation (3)) est un paramètre de corrélation entre les répartitions des deux espèces au travers du champ W_0. Ainsi, conditionnellement aux covariables environnementales et aux réalisations des champs gaussiens, un paramètre γ positif aura tendance à agréger les observations des deux espèces tandis qu'un paramètre γ négatif aura tendance à fragmenter les observations entre les deux espèces. Les vecteurs de coefficients $\boldsymbol{\beta}_1$ (*resp.* $\boldsymbol{\beta}_2$) sont quant à eux les paramètres correspondant aux p variables environnementales $\boldsymbol{x}(s)$ pour l'espèce 1 (*resp.* pour l'espèce 2).

3 Estimation des paramètres

Nous nous concentrons dans cette partie sur l'identifiabilité des paramètres mis en jeu dans le modèle décrit par les équations (2) et (3). On y présente la méthode INLA (Integrated Nested Laplace Approximation) et les différents scénarios de simulation.

3.1 La méthode INLA

Supposons pour le moment l'effort d'échantillonnage constant égal à 1 (voir équation (1)) afin de se concentrer sur l'abondance des espèces $\Lambda_i(s)$ pour $i \in \{1, 2\}$. Supposons

que nous faisions intervenir p covariables environnementales dans notre modèle et que les fonctions de covariance de chaque processus latent (W_0, W_1 et W_2) sont des covariances de Matérn (ici avec le paramètre de régularité fixé à $\nu = 1$). Chaque fonction de covariance est alors contrôlée par deux paramètres, la variance et la portée spatiale. Au total, nous avons besoin d'estimer $2 \times (p + 1)$ coefficients linéaires et 7 hyperparamètres (trois variances, trois portées spatiales, et le paramètre d'interaction γ). Pour cela nous utilisons une méthode d'inférence bayésienne reposant sur les approximations de Laplace appelée INLA développée par Rue, Martino et Chopin (2009). L'avantage d'INLA par rapport aux autres méthodes d'estimation bayésiennes du type Markov Chain Monte Carlo (MCMC) provient du gain en temps de calcul pour les modèles pouvant être exprimés comme des champs gaussien latents, et ce gain est particulièrement important en utilisant des champs gaussiens markoviens (souvent abrégés *GMRF* pour Gauss–Markov Random Fields). Avec INLA, nous utilisons l'approche des équations aux dérivées partielles stochastiques (*Stochastic Partial Differential Equations*, SPDE) pour simuler et estimer les champs gaussiens W_0, W_1 et W_2 grâce à des fonctions de base du type des éléments finis. Cette approche décrite par Lindgren, Rue et Lindström (2011), consiste en une approximation numériquement avantageuse de la fonction de covariance de Matérn via des matrices de précision creuses $\boldsymbol{Q}_j$ (donc de type GMRF) pour le vecteur des coefficients et mène à la représentation suivante :

$$W_j(s) = \sum_{\ell=1}^{L_j} w_{j,\ell}\psi_{j,\ell}(s), \quad \boldsymbol{w}_j = (w_{j,1}, \ldots, w_{j,L_j}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_j^{-1}),$$

où $\psi_{j,\ell}$ sont les éléments finis fixés grâce à une triangulation de l'espace.

3.2 Scénarios de simulation

Afin d'étudier les capacités d'inférence de nos paramètres à l'aide d'INLA, différents scénarios de simulations suivant le modèle bivarié (équations (2) et (3)) seront présentés. Le premier scénario portera sur la simulation de données gaussiennes dans un carré unité. Un intérêt particulier sera porté à la qualité d'estimation des paramètres lié au champ partagé W_0 et au paramètre γ de corrélation entre les répartitions des deux espèces. Nous nous attacherons en particulier à l'estimation du signe de ce paramètre d'interaction ainsi qu'à sa significativité, où nous considérons un effet comme significatif si son intervalle de crédibilité ne contient pas la valeur zéro. En effet, au-delà de sa valeur quantitative, du signe du paramètre γ dépendent les interprétations écologiques.
Le deuxième scénario de simulation présentera l'adaptation de ces simulations de données gaussiennes au domaine du Parc National des Ecrins. Ce jeu de simulation permettra d'étudier l'importance de la résolution spatiale du maillage choisi pour la simulation des champs gaussiens à l'échelle des Ecrins en abordant le compromis entre le temps de calcul et la précision des estimations.

Enfin, le troisième scénario de simulation portera sur les simulations de données de comptage dans le parc national des Ecrins. En outre, la qualité des estimations des différents paramètres dans le cas de données de comptage (relativement peu informatives pour les champs gaussiens latents) sera comparée à celle obtenue pour des données gaussiennes avec une erreur de mesure (amenant plus d'information utile pour l'estimation des champs gaussiens latents).

4 Modélisation spatio-temporelle et effort d'échantillonnage

Une fois les résultats des simulations obtenus dans le cadre spatial, nous présenterons l'extension du modèle dans le cadre spatio-temporel à l'aide d'une structure autorégressive. Pour tout $i \in \{1, 2, 3\}$, le champ W_i s'écrit en le point $(s, t) \in \Omega \times \mathbb{R}_+$ comme

$$W_i(s,t) = W_i(s) + \varepsilon(t) \qquad \text{avec} \quad \varepsilon(t+1) = \rho\varepsilon(t) + \sqrt{1-\rho^2}\tilde{\varepsilon}(t)$$

où les $\tilde{\varepsilon}(t)$ sont $i.i.d$ de loi $\mathcal{N}(0, \sigma^2)$ et $\varepsilon(1) = \tilde{\varepsilon}(1)$.
Enfin on abordera la question de la modélisation de l'effort d'échantillonnage et de son estimation. Une première étude consistera à intégrer un effort d'échantillonnage dépendant de l'espace et du temps comme un offset de l'abondance en considérant les observations de l'ensemble des espèces. Dans une seconde étude, à l'instar de Botella et al. (2019), nous modéliserons et estimerons de façon explicite un champs spatial décrivant cet effort d'échantillonnage.

Bibliographie

Botella, C. and Joly, A. and Monestiez, P. and Munoz, F. (2019). Bias in presence-only niche models related to sampling effort and species niches : Lessons for background point selection. *PLoS ONE*, 15(5) : e023207, pp. 1-18.

Lindgren, F. and Rue, H. and Lindström, J. (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(4), pp.423-498.

Møller, J. and Syversveen, A. R. and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25, pp. 451-482.

Renner, I.W and Elith, J. and Baddeley, A. and Fithian, W. and Hastie, T. and Phillips, S. and Popovic, G. and Warton, D.I. (2015). Point process models for presence-only analysis - a review, *Methods in Ecology and Evolution*, 6, pp. 366-379.

Rue, H. and Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(2), pp. 319-392.