



UNIVERSITÉ PARIS SUD

STAGE DE PREMIÈRE ANNÉE DE MASTER

**Modélisation de la forêt guyanaise par les  
processus ponctuels grâce aux méthodes  
INLA-SPDE**

*Léna KLAY*

supervisé par  
Eric MARCON, Stéphane TRAISSAC et Géraldine DERROIRE

Mai à juillet 2019

## Résumé

La forêt guyanaise couvre près de 96% du territoire et recèle une biodiversité exceptionnelle. Avec plus de 1500 espèces d'arbres qui y prospèrent, cet écosystème constitue un terrain d'étude aussi riche que complexe.

Les techniques récentes de modélisation de processus ponctuels et notamment les méthodes INLA-SPDE [6] ouvrent de nouvelles perspectives intéressantes dans ce domaine. Nous souhaitons adapter ces méthodes au site expérimental de Paracou, lieu d'étude de la forêt tropicale humide. Le but de ce stage est d'explorer les possibilités et les limites que nous offre ce nouvel outil.

Tout d'abord, nous avons pris le temps de comprendre les fondements des méthodes INLA-SPDE et les avantages inhérents. Nous avons ensuite cherché à tester ces approches sur des résultats connus. Enfin nous avons établi le modèle le plus satisfaisant possible afin de simuler la répartition des arbres au sein d'une espèce. Toutes les simulations ont été effectuées sous R version 3.6.0, et RStudio version 1.2.1335.



## Remerciements

Tout d'abord je voudrais remercier mes maîtres de stage, non pas parce que cela est d'usage mais bien parce qu'ils ont été des encadrants exceptionnels, tant d'un point de vue humain que pour me guider pendant mon stage. Merci à Eric Marcon qui a su prendre du temps pour discuter et se creuser la tête avec moi, malgré un emploi du temps très chargé ! Merci à Stéphane Traissac (alias le vengeur R-masqué), pour sa bonne humeur à toute épreuve et son coup de main pour m'y retrouver dans les mesures biscornues des coordonnées de Paracou. Merci à Géraldine Derroire pour ses bons conseils et l'analyse très pertinente de mon rapport.

Merci à Christine Keribin, qui même depuis Paris est toujours aussi efficace pour répondre à mes questions. Plus largement merci d'avoir été aussi disponible tout au long de cette année, d'avoir pris le temps de nous connaître et de nous avoir suivi comme vous l'avez fait.

Merci à Carole qui m'a chaleureusement accueillie dans sa bibliothèque lorsque mon ordinateur m'a lâché après deux semaines de stage, n'étant apparemment pas un adepte des climats tropicaux. Merci à Pascal notre informaticien que j'ai souvent embêté par la suite mais qui gardait toujours le sourire quand il me voyait débarquer dans son bureau.

Enfin je finirai par un petit mot à tous les stagiaires du campus, grâce à qui ce stage a aussi pu être un voyage. Merci à tous ces aventuriers qui ont délaissé le confort de la métropôle pour aller passer leurs week-end en hamac au milieu de la forêt ou dans un carbet au bord de l'eau. J'ai découvert des endroits magnifiques et appris un tas de choses en vrac sur l'écosystème tropical, sur les expériences et les voyages de chacun, sur moi même, merci pour tout ça.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>Dispositif expérimental de Paracou</b>	<b>7</b>
<b>3</b>	<b>Méthodes INLA-SPDE</b>	<b>9</b>
3.1	Méthode INLA . . . . .	9
3.2	Approche SPDE . . . . .	9
3.3	Package <code>inlabru</code> . . . . .	9
<b>4</b>	<b>Processus ponctuels</b>	<b>9</b>
4.1	Une rapide introduction . . . . .	9
4.1.1	propriété de premier ordre . . . . .	10
4.1.2	propriété de second ordre . . . . .	11
4.2	Processus ponctuels de notre étude . . . . .	11
4.2.1	Processus de Poisson . . . . .	11
4.2.2	Processus de Cox . . . . .	12
<b>5</b>	<b>Approche systématique de régression</b>	<b>13</b>
5.1	Régression de Poisson homogène . . . . .	13
5.2	Régression de Poisson non-homogène . . . . .	14
5.3	Régression de Cox homogène . . . . .	15
5.3.1	Problème d'implémentation dans <code>lgcp</code> . . . . .	16
5.4	Régression de Cox non-homogène . . . . .	17
5.4.1	Sur un processus de Cox non-homogène . . . . .	17
5.4.2	Sur un processus de Matérn . . . . .	19
<b>6</b>	<b>Ajout de la covariable altitude</b>	<b>21</b>
6.1	<i>Voucapoua americana</i> . . . . .	23
6.2	<i>Eperua falcata</i> . . . . .	25
6.3	<i>Oenocarpus bataua</i> . . . . .	28
<b>7</b>	<b>Limites du modèle</b>	<b>33</b>
<b>8</b>	<b>Conclusion</b>	<b>36</b>
<b>9</b>	<b>Accès au code</b>	<b>37</b>
 <b>Annexes</b>		
<b>A</b>	<b>Maillage</b>	<b>39</b>
<b>B</b>	<b>Covariance et corrélation de Matérn</b>	<b>41</b>
B.1	Fonction de covariance . . . . .	41
B.2	Fonction de corrélation . . . . .	42

<b>C Etudes complémentaires</b>	<b>43</b>
C.1 Test de la fonction <code>lgcp</code> dans le cas non-homogène . . . . .	43
<b>D Critères pour la comparaison des modèles</b>	<b>44</b>
<b>E Statistiques non-paramétriques</b>	<b>46</b>

## Abbréviations courantes

obs.	observée
moy.	moyenne
var.	variance

## Principales fonctions utilisées

Fonction	Package	A quoi sert-elle ?
<code>density</code>	<code>spatstat</code>	Calcule en tout point l'intensité d'un processus observé.
<code>lgcp</code>	<code>inlabru</code>	Effectue une régression de Cox.
<code>bru</code>	<code>inlabru</code>	Effectue plusieurs types de régression ( <i>family</i> à préciser).

## Table des figures

### Site de Paracou

1	Dispositif expérimental de Paracou . . . . .	8
---	--	---

### Approche systématique

#### Poisson non homogène

2	Fonction d'intensité $f : (x, y) \rightarrow x + y$ . . . . .	15
---	---	----

3	Intensité obs. ( <code>density</code> ) . . . . .	15
---	---	----

#### Cox homogène

4	Intensité moy. décrite par la réalisation du champ aléatoire $\mathcal{F}_1$ . . . . .	16
---	--	----

#### Cox non homogène

5	Valeurs moyennes du champ aléatoire $\mathcal{F}_2$ . . . . .	18
---	---	----

6	Intensité moy. décrite par la réalisation du champ aléatoire $\mathcal{F}_2$ . . . . .	18
---	--	----

7	Intensité obs. ( <code>density</code> ) . . . . .	19
---	---	----

8	Intensité moy. prédite ( <code>lgcp</code> ) . . . . .	19
---	--	----

9	Ecart-type de l'intensité prédite ( <code>lgcp</code> ) . . . . .	19
---	---	----

#### Matérn

10	Intensité obs. ( <code>density</code> ) . . . . .	20
----	---	----

11	Intensité moy. prédite ( <code>lgcp</code> ) . . . . .	20
----	--	----

12	Ecart-type de l'intensité prédite ( <code>lgcp</code> ) . . . . .	21
----	---	----

### Ajout de l'altitude en covariable

#### Altitude

13	Altitude sur la parcelle 16 de Paracou (en m au dessus de la mer)	22
----	---	----

#### *Voucapoua americana*

14	Répartition et altitude de <i>Voucapoua americana</i> . . . . .	23
----	---	----

15	Modèle altitude . . . . .	23
----	---------------------------	----

16	Modèle champ <i>field</i> . . . . .	23
----	-------------------------------------	----

17	Modèle altitude et champ <i>field</i> . . . . .	24
----	---	----

#### *Eperua falcata*

18	Répartition et altitude de <i>Eperua falcata</i> . . . . .	25
----	--	----

19	Modèle altitude . . . . .	26
----	---------------------------	----

20	Modèle champ <i>field</i> . . . . .	26
----	-------------------------------------	----

21	Modèle altitude et champ <i>field</i> . . . . .	26
----	---	----

#### *Oenocarpus bataua*

22	Répartition et altitude de <i>Oenocarpus bataua</i> . . . . .	28
----	---	----

23	Modèle altitude . . . . .	28
----	---------------------------	----

24	Modèle champ <i>field</i> . . . . .	28
----	-------------------------------------	----

25	Modèle altitude et champ <i>field</i> . . . . .	29
----	---	----

<b>Analyse des résultats des trois espèces</b>	
26 Tableau récapitulatif de l'intensité observée, moyenne et de l'écart-type pour le modèle altitude et champ <i>field</i> . . . . .	31
27 Tableau récapitulatif du paramètre <i>range</i> en centaines de m et de la fonction de corrélation pour chaque espèce (modèle altitude et champ <i>field</i> ) . . . . .	33

### **Etude de la répartition inter-spécifique**

28 Répartition de Vouacapoua americana et de Qualea rosea sur la parcelle 16 de Paracou . . . . .	35
29 Fonction M appliquée à l'espèce de référence Vouacapoua americana et l'espèce secondaire Qualea rosea (Mhat du package dbmss) . . . . .	35

### **Annexes**

#### **Maillage**

30 Mesh adapté à un exemple de points. . . . .	39
31 Interface du package INLA pour la construction d'un maillage . .	40
32 Indice de qualité : Approximate standard deviation . . . . .	40

#### **Covariance de Matérn**

33 Exemple de paramètres de la fonction de covariance Matérn. . .	42
34 Exemple de paramètres de sortie du champ aléatoire gaussien. .	42

## 1 Introduction

Ce stage s'est déroulé au sein de l'équipe de recherche Écologie des Forêts de Guyane (EcoFoG). Cette équipe est commune à différents organismes parmi lesquels on compte le CNRS<sup>a</sup>, AgroParisTech<sup>b</sup>, le Cirad<sup>c</sup>. Je suis officiellement rattachée au CNRS, Eric Marcon est directeur du centre de Kourou d'AgroParisTech et directeur de l'UMR EcoFoG, Stéphane Traissac est enseignant-chercheur à AgroParisTech et Géraldine Derroire est chercheuse au Cirad et responsable scientifique de la station de recherche forestière de Paracou.

Ce stage a eu lieu sur le campus agronomique de Kourou en Guyane Française et j'ai également eu l'occasion de me rendre sur le site forestier de Paracou d'où proviennent mes données.

---

a. CNRS : Centre National de la Recherche Scientifique

b. AgroParisTech : Institut des sciences et industries du vivant et de l'environnement

c. Cirad : Centre de Coopération Internationale en Recherche Agronomique pour le Développement

## 2 Dispositif expérimental de Paracou

Le dispositif expérimental de Paracou implanté sur la commune de Sinnamary est un site d'étude de l'écosystème forestier tropical humide. Géré par le Cirad, il comporte 16 parcelles d'étude qui couvrent au total 125 hectares sur lesquels 70000 arbres sont cartographiés, identifiés botaniquement et mesurés à des périodes régulières depuis 1984.

Parmi les 16 parcelles figurent 4 types de traitements expérimentaux qui reflètent le degré d'intervention de l'homme et son action. Les parcelles qui nous intéressent sont les parcelles dites de forêt non perturbée c'est à dire où il n'y a pas eu d'intervention de l'homme, et qui sont sur la figure ci-dessous en vert le plus foncé. Plus précisément, nous étudions la parcelle P16 (de 500m de côté).

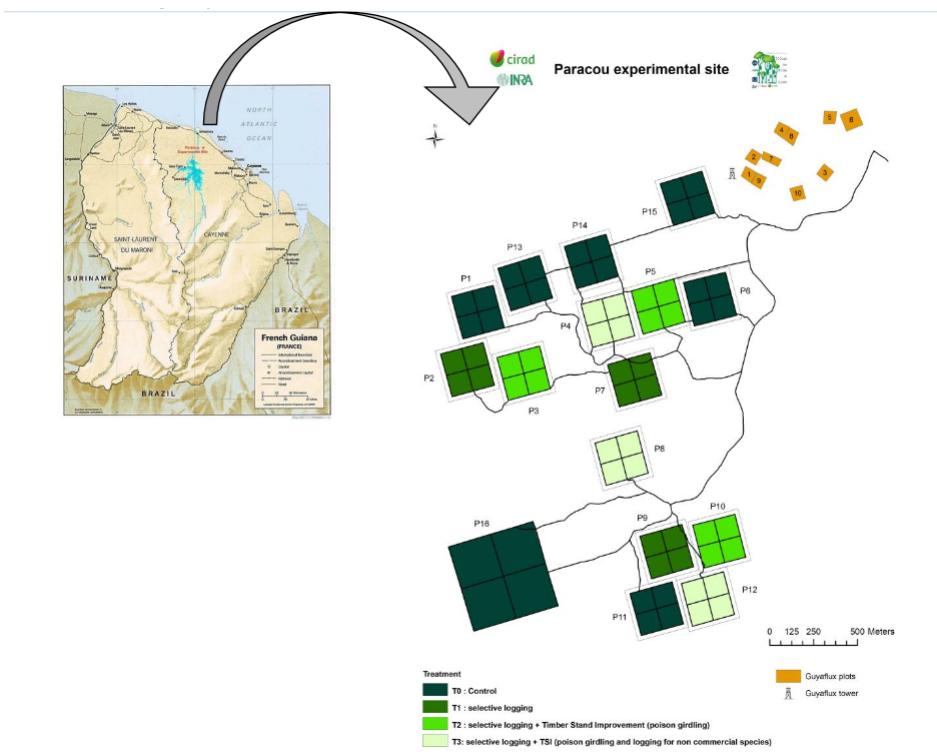


FIGURE 1 – Dispositif expérimental de Paracou

Nous nous intéressons à la structure spatiale des arbres à un temps  $t$  fixé, dans notre cas à la date des dernières mesures de 2015. Les coordonnées pour chaque arbre représentent la position de son centre et nous utilisons l'altitude comme variable supplémentaire du modèle. Le but de ce stage va être dans un premier temps de s'approprier les techniques de modélisation INLA-SPDE, puis de construire le modèle le plus performant possible afin de simuler la répartition des arbres au sein d'une espèce. Ces simulations pourraient notamment nous apporter des informations intéressantes d'un point de vue écologique, quant à la structuration des populations.

## 3 Méthodes INLA-SPDE

### 3.1 Méthode INLA

La méthode INLA (*Integrated Nested Laplace Approximation*) décrite pour la première fois dans l'article de Rue, Martino et Chopin [6] est une méthode analytique d'inférence bayésienne. Elle se base sur les approximations de Laplace et offre ainsi une alternative aux méthodes de Monte-Carlo par Chaîne de Markov qui sont elles des méthodes d'approximation par simulation. Le principal avantage de l'approche INLA réside dans sa grande rapidité, mais nous pouvons également citer son caractère plus générique (elle couvre un plus large panel de problèmes). Cette méthode a été implémentée dans le package du même nom INLA.

### 3.2 Approche SPDE

Pour étudier la loi de processus ponctuels, nous cherchons à déterminer en particulier l'intensité de ce processus aux différents endroits du domaine. Nous représentons cette intensité sous la forme d'un champ aléatoire dont nous cherchons la loi à posteriori. La méthode SPDE détaillée dans l'article de Lindgren, Rue et Lindström [3] nous permet de faire cela efficacement en considérant une équation aux dérivées partielles stochastiques (SPDE) bien particulière dont la solution est un champ aléatoire gaussien avec une fonction de covariance Matérn (détails section B). L'équation se résoud par la méthode des éléments finis et fournit alors une approximation discrète de la solution sous la forme d'un champ aléatoire gaussien de Markov (sous la condition  $\nu > 0$ ). Ces champs ont la particularité d'avoir une matrice de précision sparse ce qui facilite considérablement les calculs. Cette méthode est également intégrée au package INLA.

Pour toutes questions plus avancées à ce sujet, le "SPDE book" [2] détaille de façon très claire la théorie mathématique et le fonctionnement du package.

### 3.3 Package `inlabru`

Le package `inlabru` est un package dérivant de celui d'INLA adapté à l'étude des processus ponctuels. Il est particulièrement bien adapté aux données écologiques et facile d'utilisation. Nous utilisons ces deux packages dans notre étude.

## 4 Processus ponctuels

### 4.1 Une rapide introduction

Un processus ponctuel est un processus aléatoire dont les réalisations sont des semis de points ou groupes de points (les arbres dans notre étude). Les

propriétés du processus définissent des contraintes sur ses réalisations (densité, voisinage, structure...). Un processus ponctuel peut donc générer une infinité de semis de points tous différents, mais qui partageront des propriétés communes puisque qu'étant les réalisations d'un seul et même processus. Lorsque l'on effectue une régression, on fixe la famille du processus aléatoire étudiée et on cherche les paramètres les plus probables au sein de cette famille pour que le processus correspondant ait engendré le semis. Cette étude se base bien évidemment sur les propriétés du semis.

On introduit les notations suivantes :

$\Omega$	le domaine d'étude (dans notre cas la parcelle 16)
$A$	une surface quelconque contenue dans $\Omega$
$n(A)$	le nombre de points du semis appartenant à la surface A
$dS$	une surface élémentaire quelconque contenue dans $\Omega$
$n(dS)$	le nombre de points du semis appartenant à $dS$ , qui vaut 0 ou 1 car c'est une surface élémentaire

Une des notions essentielles à comprendre pour l'étude des processus ponctuels est la suivante. L'observation de deux points proches l'un de l'autre dans le semis peut être attribuée à deux phénomènes :

- une intensité forte du domaine à cet endroit précis (biologiquement un environnement propice au développement des arbres),
- un phénomène d'agrégation, c'est à dire un phénomène d'attraction entre les deux arbres (biologiquement, deux espèces tirant bénéfice de leur proximité ou plus simplement une descendance)

ou potentiellement les deux en même temps.

Ces deux phénomènes sont quantifiés par ce que l'on appelle les propriétés de premier et second ordre.

#### 4.1.1 propriété de premier ordre

La propriété du premier ordre d'un processus ponctuel représente son intensité sur le domaine. Elle est notée  $\lambda(x)$  et caractérise la probabilité de présence d'un point du semis dans une surface élémentaire  $dS$  centrée en  $x$ .

$$\mathbb{P}(\text{présence d'un point dans } dS) = \mathbb{P}(n(dS) = 1) = \lambda(x)dS$$

Pour les processus dit homogènes, l'intensité  $\lambda$  est constante sur le domaine. Elle peut alors être estimée par :

$$\lambda \simeq \frac{\text{Nombre de points du semis}}{\text{Aire du domaine}}$$

#### 4.1.2 propriété de second ordre

La propriété de second ordre décrit la probabilité de présence conjointe de deux points dans les surfaces élémentaires  $dS_1$  et  $dS_2$  centrées en  $x_1$  et  $x_2$ . Elle est caractérisée par la fonction de densité des paires de points  $g(.,.)$ , définie par :

$$\begin{aligned} \mathbb{P}(\text{présence d'un point dans } dS_1 \text{ et d'un point dans } dS_2) = \\ \mathbb{P}(n(dS_1) = 1 \cap n(dS_2) = 1) = \lambda(x_1)\lambda(x_2)g(x_1, x_2)dS_1dS_2 \end{aligned}$$

La fonction  $g(.,.)$  exprime alors en quelque sorte les relations de voisinage entre deux points d'un semis, ce qui fait d'elle un bon outil (non-paramétrique) pour décrire les phénomènes d'agrégation.

#### Remarque

Il est normalement difficile de séparer les deux effets, l'intensité du domaine et les phénomènes d'agrégation avec des méthodes paramétriques. Cependant comme nous allons le voir ci-dessous le modèle de Cox nous offre cette possibilité.

## 4.2 Processus ponctuels de notre étude

### 4.2.1 Processus de Poisson

#### 4.2.1.1 Processus de Poisson homogène

Le processus de Poisson homogène résulte d'un tirage complètement aléatoire des points du semis, c'est à dire :

- la distribution des points est homogène, l'intensité est identique en tout point du domaine (*propriété du premier ordre*).
- la distribution des points est indépendante et il n'y a pas de phénomène d'agrégation (*propriété du second ordre*).

Grâce à ses propriétés, le processus de Poisson homogène est généralement choisi comme hypothèse nulle (hypothèse d'indépendance) dans les tests de répartition. Détaillons ses propriétés :

#### propriété du premier ordre

Le nombre de points du semis contenus dans une surface A quelconque appartenant à  $\Omega$  suit une loi de Poisson de paramètre  $\lambda||A||$  :

$$\mathbb{P}(n(A) = k) = e^{-\lambda||A||} \frac{(\lambda||A||)^k}{k!} \quad \forall k \in \mathbb{N}$$

où  $||A||$  est l'aire de A et  $\lambda$  est l'intensité, une constante. Ainsi l'espérance et la variance du nombre de points du semis contenus dans A valent :

$$\mathbb{E}[n(A)] = Var[n(A)] = \lambda||A||$$

#### propriété du second ordre

Les points étant distribués indépendamment les uns des autres, il en résulte que :

$$g(r) = 1 \quad \forall r \in \mathbb{R}^+$$

(Car dans le cas d'indépendance, la probabilité jointe de deux événements équivaut au produit des probabilités de réalisation de chacun d'entre eux)

##### 4.2.1.2 Processus de Poisson non-homogène

Le processus de Poisson non-homogène est une extension du processus de Poisson homogène pour laquelle l'intensité n'est pas constante sur le domaine. La propriété du premier ordre vaut pour une surface élémentaire quelconque dS de  $\Omega$  centrée sur x :

$$\mathbb{P}(n(dS) = k) = e^{-\lambda(x)} \frac{(\lambda(x))^k}{k!} \quad \forall k \in \{0, 1\}$$

On voit bien ici que l'intensité n'est plus une constante mais bien une fonction réelle définie pour chaque surface élémentaire dS par  $x \mapsto \lambda(x)$ . La propriété du second ordre vaut toujours  $g(r) = 1 \quad \forall r \in \mathbb{R}^+$  car la distribution des points est indépendante.

##### 4.2.2 Processus de Cox

Un processus de Cox est un processus de Poisson pour lequel le paramètre d'intensité  $\lambda$  est une variable aléatoire de loi log-normale. Ce processus peut être :

- homogène : dans ce cas, cette variable aléatoire  $\lambda$  est la même en tout point du domaine,
- non-homogène : le paramètre d'intensité  $\lambda$  est un champ aléatoire (défini par une variable aléatoire sur chaque surface élémentaire  $dS$ ).

On note  $\lambda(x)$  la variable aléatoire donnant l'intensité sur la surface élémentaire  $dS$  centrée en  $x$ . Dans le cas homogène :  $\lambda(x) = \lambda, \forall x$ . Pour générer un processus de Cox, il suffira donc de tirer une réalisation de ce champ aléatoire, puis de générer un processus de Poisson sur cette réalisation.

Les points étant des réalisations de variables aléatoires, on conserve la propriété d'indépendance :  

$$g(r) = 1 \quad \forall r \in \mathbb{R}^+$$

Cependant les variables aléatoires  $\lambda(x)$  elles, peuvent être corrélées et peuvent donc engendrer des agrégats de points.

En conclusion, la moyenne pour chaque  $\lambda(x)$  reflète l'intensité sur le domaine, tandis que la fonction de covariance des  $\lambda(x)$  reflète l'agrégation des points. Grâce au modèle de Cox, on peut donc distinguer les effets de la propriété du premier ordre de celle du second ordre.

## 5 Approche systématique de régression

La régression comme nous l'avons expliquée un peu plus haut, consiste à sélectionner les paramètres les plus probables au sein d'une famille de processus donnée, afin que le processus en résultant (processus appartenant à cette famille et avec ces paramètres précisément) ait engendré le semis observé.

Nous allons effectuer des régressions pour différentes familles de processus : celles sur lesquelles nous nous sommes attardés dans le paragraphe précédent. Le but final est de mettre en place une régression de Cox non-homogène car ce modèle pourrait nous fournir des propriétés intéressantes d'un point de vue biologique.

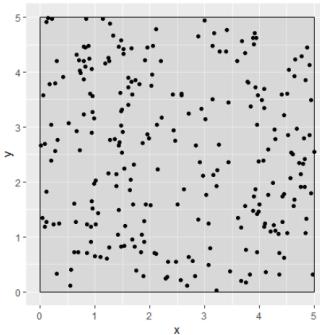
Avant de tester nos méthodes sur de vraies données, nous cherchons à les vérifier. Pour cela, nous générerons des processus dont nous connaissons les paramètres, puis nous effectuons une régression afin de contrôler la qualité de nos approximations.

### 5.1 Régression de Poisson homogène

Nous cherchons le paramètre d'intensité  $\lambda$ , constant. Par convention, la fonction `bru` du package `inlabru` effectue une régression log-linéaire :

$$\log(\lambda) = \text{Intercept}$$

Le paramètre  $\lambda$  nous donne alors l'intensité moyenne des points du semis sur  $\Omega$ , qui peut également se retrouver par le simple comptage des points divisé par l'aire de domaine. Dans notre démarche de vérification, nous obtenons les résultats suivants :



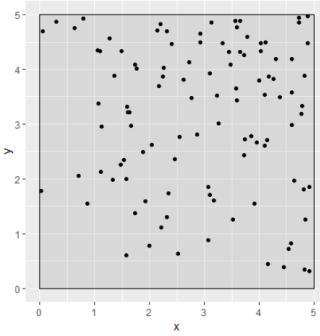
**Approche systématique**  
Processus de Poisson homogène généré

Intensité théorique	10
Intensité par comptage des points	10.2
Intensité par la régression brû	10.19996

Ces résultats sont très satisfaisants.

## 5.2 Régression de Poisson non-homogène

Nous cherchons le paramètre  $\lambda$  sous la forme d'un champ d'intensité : en d'autres termes, à chaque surface élémentaire du domaine est attribuée une valeur d'intensité constante. Il nous suffit donc de calculer l'intensité du semis sur le domaine : cela nous fournira alors le champ d'intensité le plus probable à ce qu'un processus de Poisson non-homogène ait généré ce pattern de points.



**Approche systématique**  
Processus de Poisson non-homogène  
généré à partir de la fonction d'intensité  
 $f : (x, y) \rightarrow x + y$

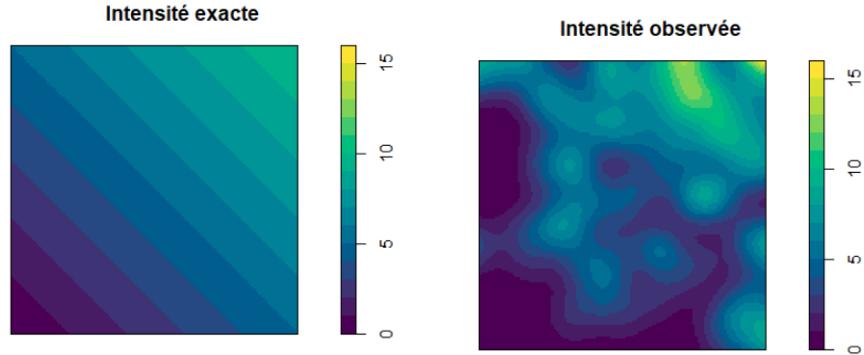


FIGURE 2 – Fonction d'intensité  
 $f : (x, y) \rightarrow x + y$

FIGURE 3 – Intensité obs. (`density`)

La fonction `density` de `spatstat` calcule l'intensité du semis sur le domaine. Pour cela, elle compte en chaque point du domaine le nombre d'observations présentes dans un disque de rayon  $r$  centré sur ce point. Chaque observation est pondérée par un poids gaussien (dont la moyenne est le centre du disque), cela implique donc que plus l'observation est proche du point, plus elle "compte". `density` effectue ensuite une somme pondérée de ces observations et divise le résultat par l'aire du disque, ce qui résultera en une approximation de l'intensité au point considéré. Par défaut,  $r$  (la bande passante) vaut un huitième du plus petit côté de la fenêtre d'observation. Cependant ce choix n'est souvent pas très judicieux, c'est pourquoi nous utilisons la fonction `bw.diggle` de `spatstat` qui fournit la bande passante minimisant l'erreur quadratique moyenne de l'estimateur (par cross-validation).

### 5.3 Régression de Cox homogène

Nous cherchons le paramètre d'intensité  $\lambda$  sous la forme d'une variable aléatoire de loi log-normale :

$$\log(\lambda) = \text{Intercept} + Y \quad \text{avec } Y \hookrightarrow \mathcal{N}(0, \sigma^2)$$

Le processus de Cox homogène est généré via la fonction `rLGCP` de `spatstat`. Pour cela nous définissons le champ aléatoire  $\mathcal{F}_1$  qui engendrera le processus : les variables de ce champ ont toutes la même loi log-normale puisque le processus est homogène. Nous fixons la moyenne de cette loi à 10 et sa variance à  $\exp(0.2)$ . Le champ est de plus muni d'une fonction de covariance de Matérn, dont nous choisissons les paramètres  $\alpha = 1/2$  et  $\nu = 1$ . (plus de détails sur la fonction de covariance de Matérn dans l'annexe B). Nous représentons la valeur moyenne de l'intensité ci-dessous :

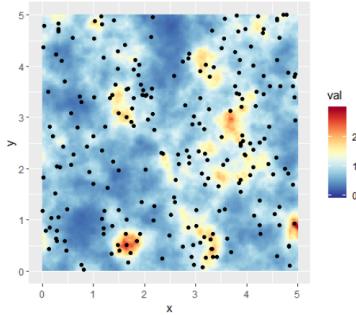
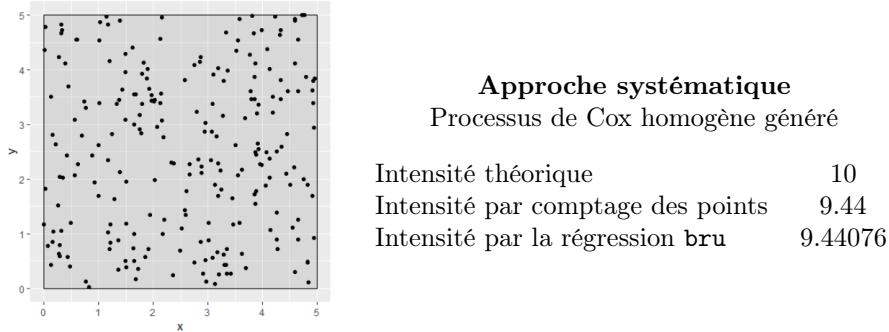


FIGURE 4 – Intensité moy. décrite par la réalisation du champ aléatoire  $\mathcal{F}_1$

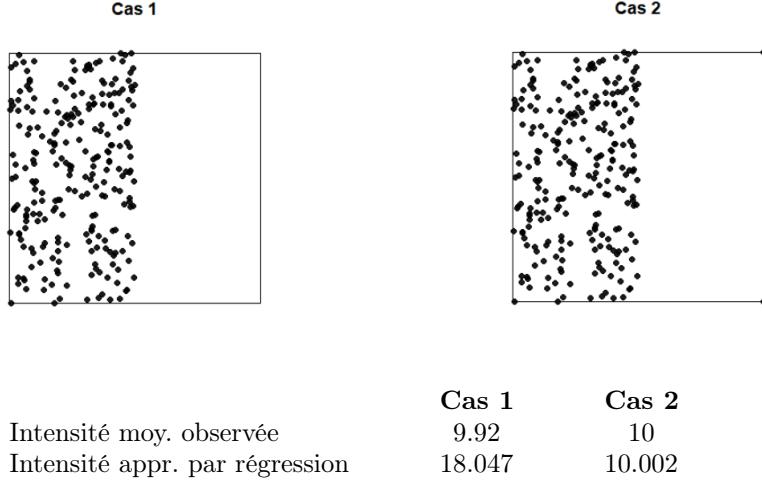
Remarque : L'intensité n'est pas constante sur le domaine bien que le processus soit qualifié d'homogène. Ce résultat n'est pas contradictoire ; les valeurs que nous observons sont les réalisations d'une variable aléatoire dont la moyenne est constante égale à 10.

La régression est effectuée grâce à la fonction `lgcp` du package `inlabru`. (`lgcp` : Log-Gaussian Cox process)



### 5.3.1 Problème d'implémentation dans `lgcp`

Les résultats ci-dessus semblent probants. Cependant en effectuant quelques analyses supplémentaires, nous remarquons que pour un processus trop éloigné d'un processus de Cox homogène, les approximations deviennent rapidement incorrectes. Afin de mettre cette imprécision en évidence, nous testons la régression sur deux patterns de points bien précis : un processus de Poisson homogène ( $\lambda = 20$ ) couvrant uniquement la moitié du domaine, et ce même exact processus auquel sont rajoutés deux points placés sur les coins opposés du domaine.



Tout d'abord nous remarquons qu'en ajoutant des points au semis (en passant du cas 1 au cas 2), nous faisons baisser la densité obtenue par régression ce qui témoigne d'une incohérence. Nous notons également que cette densité dans le cas 1 reste proche de 20, ce qui est l'intensité du processus de Poisson homogène généré sur la moitié du domaine. Tout laisse à penser que la fonction de régression agit "sans tenir compte" des aires qui ne contiennent pas de points, c'est à dire qu'elle cantonne son domaine d'étude à un voisinage du semis et non au domaine spécifié... Cette mauvaise implémentation doit être signalée, c'est pourquoi nous enverrons dès que possible un rapport de bug à son auteur.

## 5.4 Régression de Cox non-homogène

### 5.4.1 Sur un processus de Cox non-homogène

Nous cherchons cette fois à expliquer le paramètre  $\lambda$  sous la forme d'un champ aléatoire log-normal. À chaque surface élémentaire  $dS$  centrée en  $x$ , le logarithme de l'intensité est donnée par :

$$\log(\lambda(x)) = \text{Intercept} + \text{field}(x)$$

où  $\text{field}$  est un champ aléatoire gaussien centré (c'est à dire que chacune des variables aléatoires  $\text{field}(x)$  est gausienne centrée). Afin d'intégrer cette co-variable à notre modèle, nous devons d'abord discréteriser le domaine sous la forme d'un maillage. Cette étape est détaillée dans l'annexe A.

Afin de simuler le processus de Cox non-homogène, nous décidons de fixer la moyenne du champ aléatoire théorique  $\mathcal{F}_2$  à un damier de densité 5 ou 20 selon la case. Les autres paramètres du processus de la fonction de covariance Matérn

sont  $\text{var.} = 0.2$ ,  $\alpha = 1/2$ ,  $\nu = 1$  (voir détail annexe B).

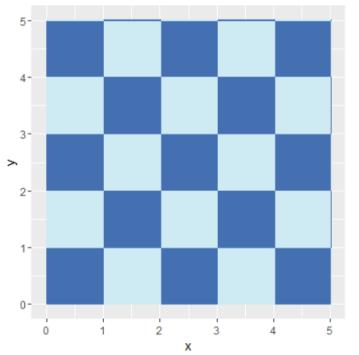


FIGURE 5 – Valeurs moyennes du champ aléatoire  $\mathcal{F}_2$

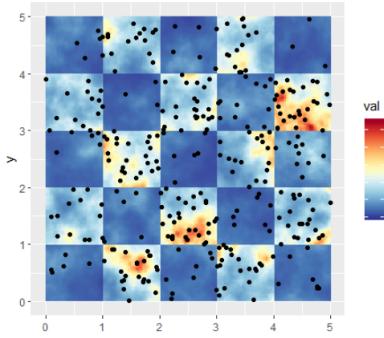
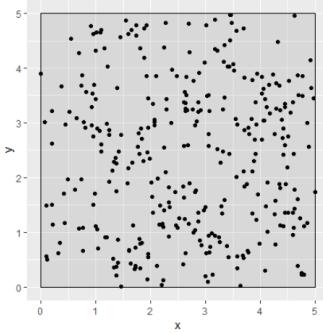


FIGURE 6 – Intensité moy. décrite par la réalisation du champ aléatoire  $\mathcal{F}_2$

Nous effectuons une régression de Cox non-homogène grâce à la fonction `lgcp` du package `inlabru`. Cette régression calcule l'intensité en chaque point du domaine, le problème d'implémentation détecté dans la section 5.3.1 ne devrait donc pas nous affecter ici. Afin d'en être totalement certains, nous refaisons l'expérience en annexe C.1 dans le cas non-homogène.



**Approche systématique**  
 Processus de Cox non-homogène  
 ayant pour moyenne la fonction damier  
 (détailée plus haut)  
 et pour paramètres de covariance  
 $\text{var.} = 0.2$ ,  $\alpha = 1/2$ ,  $\nu = 1$

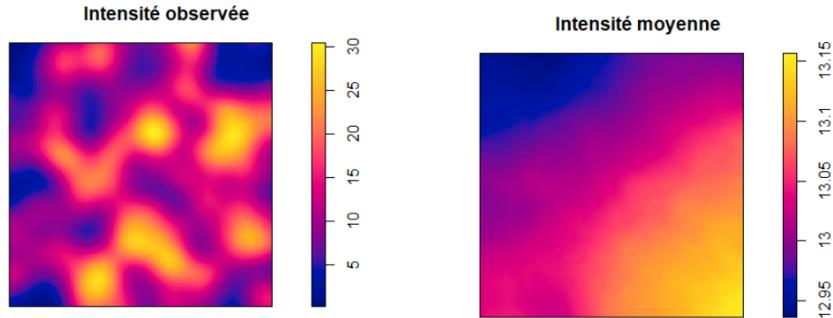


FIGURE 7 – Intensité obs. (`density`)

FIGURE 8 – Intensité moy. prédite (`lgcp`)

(Attention les échelles sont très différentes.) On remarque ici que la régression de Cox non-homogène a tendance à homogénéiser l'intensité ; en effet elle varie entre 12.9 et 13.2 alors que la fonction théorique de l'intensité moyenne (la fonction damier) varie entre 5 et 20. Une première explication pourrait se trouver dans le fait que la prédiction admet un écart-type plus important que la réalité afin de "combler" cette erreur :

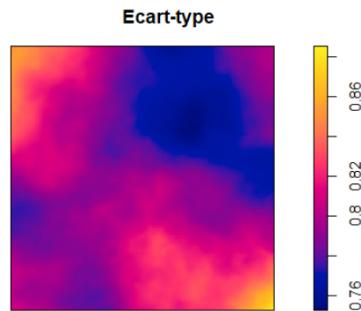


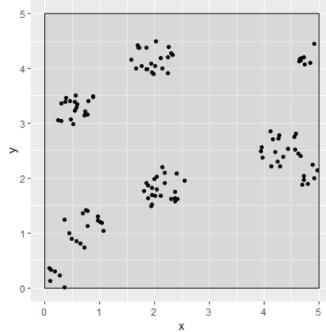
FIGURE 9 – Ecart-type de l'intensité prédite (`lgcp`)

Cependant cette variation (un écart-type maximum de 0.87) est trop faible pour même atteindre les valeurs de la fonction damier (moyenne théorique)... Il semble donc que la régression `lgcp` se basant uniquement sur le champ aléatoire `field` n'est pas très performante lorsque le champ d'intensité théorique varie trop abruptement.

#### 5.4.2 Sur un processus de Matérn

Le processus de Matérn (qui n'a aucun lien avec la fonction de covariance de Matérn, si ce n'est son inventeur) est un autre type de processus ponctuel.

Des points, dits centres, sont générés selon un processus de Poisson homogène de paramètre  $\lambda_1$ . Dans un disque de rayon  $r$  autour de chaque centre, les points du semis sont ensuite générés selon un processus de Poisson homogène de paramètre  $\lambda_2$ . Les centres ne sont pas conservés dans le semis. Il en résulte un processus non-homogène et agrégé, dont voici un exemple :



### Approche systématique

Processus de Matérn  
de paramètres  $\lambda_1 = 0.3$ ,  
 $\lambda_2 = 40$  et  $r = 0.4$

Nous effectuons une régression de Cox non-homogène sur ce semis grâce à la fonction `lgcp` d'`inlabru`.

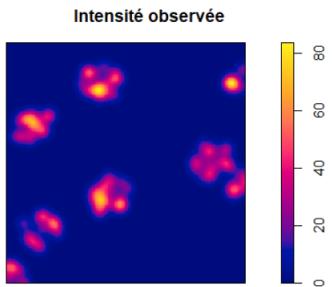


FIGURE 10 – Intensité obs. (`density`)

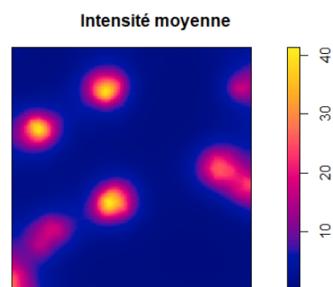


FIGURE 11 – Intensité moy. prédite (`lgcp`)

Cette régression nous fournit bien l'intensité initiale qui était de  $\lambda_2 = 40$ .

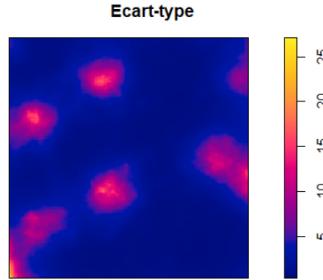


FIGURE 12 – Ecart-type de l'intensité prédictive (lgcp)

Les très grandes valeurs de l'écart-type aux emplacements des agrégats expliquent également les valeurs plus extrêmes que nous pouvons observer avec la fonction `density`. Les résultats semblent donc concluants.

A la vue de ces deux exemples, nous pouvons conclure à une meilleure performance dans le cas du processus de Matérn. Ce résultat peut s'expliquer par le fait que les agrégas de ce processus sont bien mieux délimités, il est donc plus facile pour la fonction de détecter une intensité non-homogène. Il semble ici que nous touchons du doigt une des limites de la régression par la fonction `lgcp`.

Le type d'agrégation observée dans l'exemple du processus de Cox non-homogène pourrait bien être similaire à certaine répartitions d'arbres, et cela serait problématique. Afin de limiter les erreurs de régression, nous décidons de fournir plus d'informations à la fonction en rajoutant une nouvelle covariable : l'altitude. Cette covariable devrait permettre d'expliquer la majeure partie des variations d'intensité, et donc d'aider à mieux cerner ces changements parfois abruptes d'intensité. Le modèle devrait alors être plus performant et la variable `field` plus à même de refléter ce pourquoi elle a été introduite : rendre compte des interactions spatiales des arbres au sein d'une espèce.

## 6 Ajout de la covariable altitude

L'altitude est une variable très significative dans la répartition des arbres sur la parcelle 16 de Paracou. Ce résultat a été démontré par Elodie Allié dans sa thèse [1].

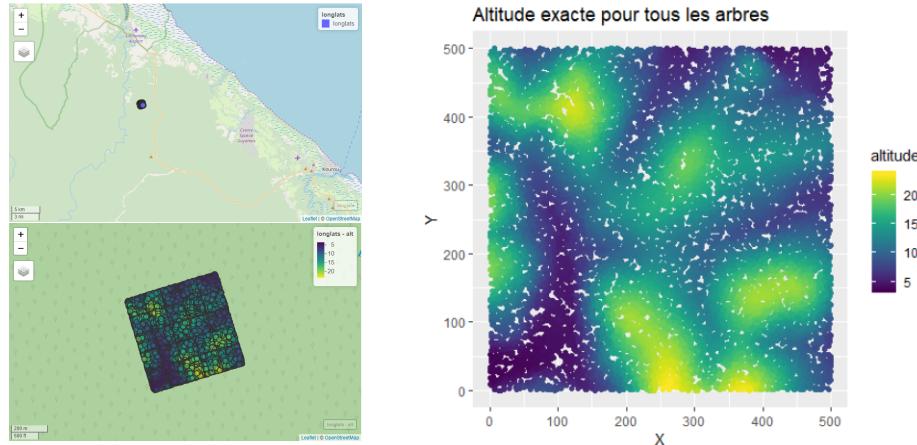


FIGURE 13 – Altitude sur la parcelle 16 de Paracou (en m au dessus de la mer)

Afin d'évaluer la qualité de notre modèle, nous allons le tester sur trois espèces de Paracou qui présentent des profils très différents : *Voucapoua americana*, *Eperua falcata* et *Oenocarpus bataua*.

*Voucapoua americana* semble s'accomoder de différentes altitudes et couvre environ la moitié du domaine, *Eperua falcata* a une structure très agrégative et se situe en général en basse altitude et *Oenocarpus bataua* se situe dans des altitudes intermédiaires et couvre l'ensemble du domaine.

Enfin, dans le but de distinguer l'effet de chaque variable, nous étudions chacun des modèles suivants pour chaque espèce :

$$\text{Modèle altitude : } \log(\lambda(x)) = \text{Intercept} + \beta \text{ alti}(x)$$

$$\text{Modèle champ } field : \log(\lambda(x)) = \text{Intercept} + field(x)$$

$$\text{Modèle altitude et champ } field : \log(\lambda(x)) = \text{Intercept} + \beta \text{ alti}(x) + field(x)$$

avec *field* un champ aléatoire gaussien centré,

*alti* un champ fixé (non-aléatoire) ayant pour valeur l'altitude en chaque point du domaine ( $\beta$  son coefficient dans le modèle linéaire).

Nous comparerons ensuite nos modèles sur la base des critères DIC et WAIC (annexe D). Ceux-ci sont fournis en sortie de la fonction `lgcp`. Attention cependant : dans le tutorial `inlabru`, il est très clairement expliqué que nous avons encore très peu de recul sur l'usage pratique de ces critères sur des modèles complexes spatialement. ([06a\_mod\_valid.pdf, tutorial `inlabru`] "INLA can compute DIC and WAIC. We have little experience with practical usage of

them for complex spatial models. It is not clear what they actually mean in the context of the models we look at here." ) Ils sont cependant les seuls outils de comparaison dont nous disposons dans le package, nous allons donc les utiliser mais avec précaution. Les régressions sont toutes faites grâce à la fonction `lgcp` et le mesh est le même pour tous les modèles au sein d'une espèce.

## 6.1 *Voucapoua americana*

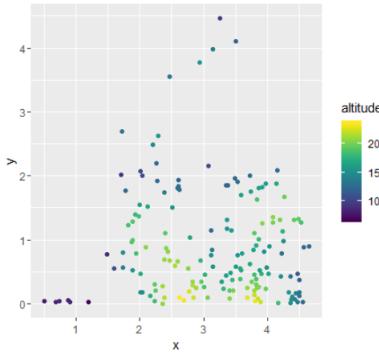


FIGURE 14 – Répartition et altitude de *Voucapoua americana*



Crédit photo : Laurent Asselin

La répartition des *Voucapoua americana* n'est pas homogène ; les arbres se trouvent plutôt au Sud de la parcelle et semblent concentrés sur les hautes altitudes (entre 15 et 20m), bien qu'on retrouve quelques arbres en basse altitude.

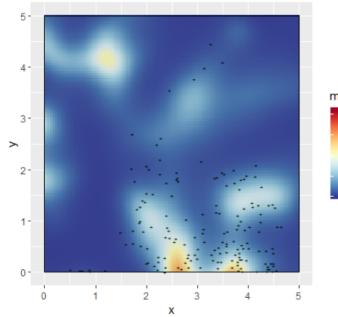


FIGURE 15 – Modèle altitude

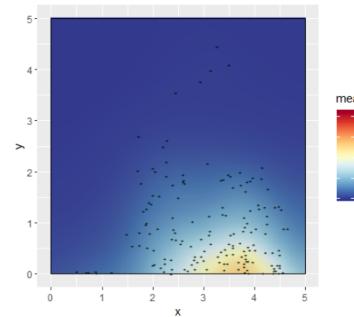


FIGURE 16 – Modèle champ *field*

Nous observons que les deux modèles apportent des informations différentes, l'ajout de la variable altitude pourrait donc s'avérer bénéfique.

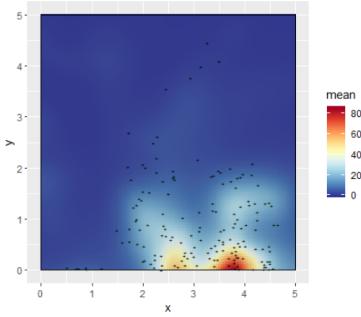


FIGURE 17 – Modèle altitude et champ *field*

Le modèle combiné semble visuellement être le plus adapté des trois modèles. Nous cherchons alors à vérifier cette première impression en comparant les modèles sur la base des critères DIC et WAIC (annexe D) :

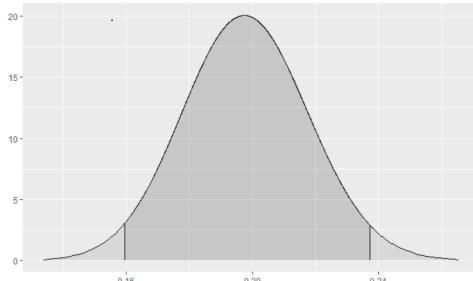
	DIC	WAIC
Modèle altitude	-511.2924	-513.7004
Modèle champ <i>field</i>	-566.1476	-566.5646
Modèle altitude + <i>field</i>	-569.8903	-570.2961

Plus les critères sont faibles, plus la prédiction du modèle est de bonne qualité : le modèle combiné est donc bien le meilleur des trois modèles. Ce résultat laisse à penser que l'altitude est une variable significative dans l'explication de la répartition des *Voucapoua americana*, ce que nous allons vérifier rigoureusement ci-dessous. Le champ de répartition intra-spécifique lui n'a pas de coefficient linéaire car ses coefficients sont internes (*range* et *sigma*) : c'est un sous-modèle de notre modèle (voir annexe B).

## Variable altitude

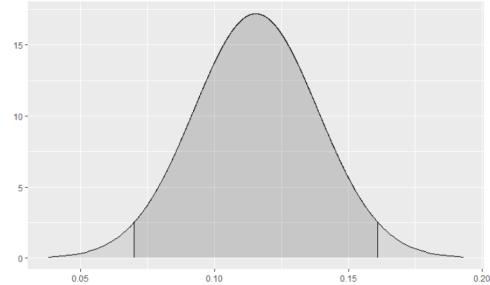
Modèle  
Coefficient  $\beta$   
Significativité

**Modèle altitude**  
0.1982447  
OUI



Loi à posteriori de  $\beta$  pour le modèle altitude

**Modèle altitude et field**  
0.1152663  
OUI



Loi à posteriori de  $\beta$  pour le modèle altitude et champ *field*

On observe que la variable altitude est significative dans les deux modèles où elle apparaît, et en particulier dans le modèle combiné qui semble être le plus adapté à notre étude. La répartition de *Voucapoua americana* dépend donc bien de l'altitude.

## 6.2 *Eperua falcata*

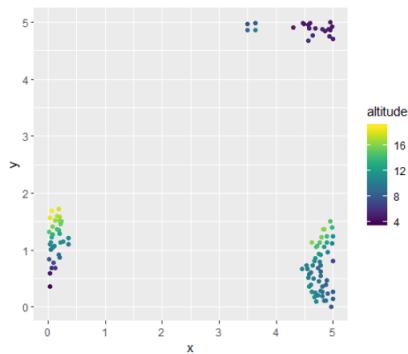


FIGURE 18 – Répartition et altitude de *Eperua falcata*



Crédit photo : Yves Caraglio

La répartition des *Eperua falcata* est très agrégative. Elle semble se concentrer majoritairement dans les zones de faible altitude (4 à 14m).

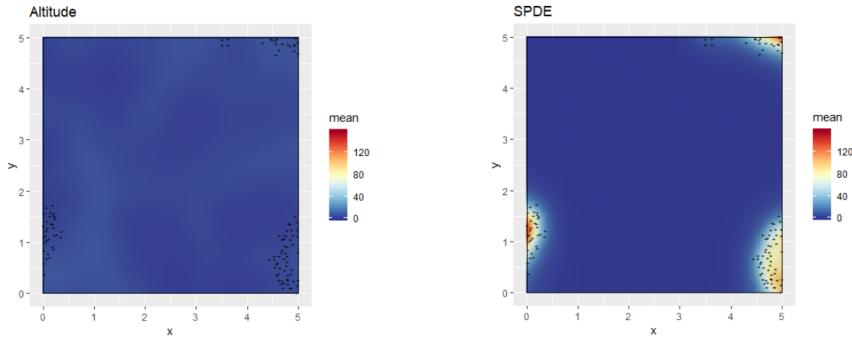


FIGURE 19 – Modèle altitude

FIGURE 20 – Modèle champ *field*

Bien que nous ayons remarqué que l'*Eperua falcata* se situe majoritairement dans les zones de basses altitudes, nous notons cependant qu'elle n'occupe de loin pas toutes ces zones. L'altitude ne semble donc pas la covariable la plus importante, ce qui est visible avec la prédiction du modèle altitude (la prédiction n'est pas nécessairement constante, mais à l'échelle du second modèle c'est l'impression que nous avons). La répartition intra-spécifique semble elle très significative grâce aux agrégats extrêmement bien délimités, que le second modèle arrive très bien à déceler.

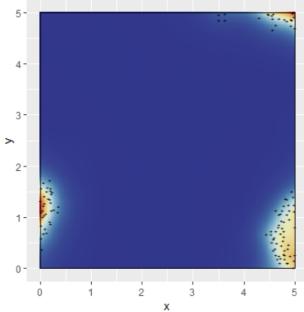


FIGURE 21 – Modèle altitude et champ *field*

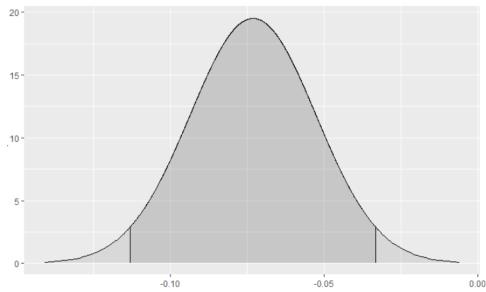
Le modèle combiné ressemble très fortement au modèle champ *field*, puisque l'altitude semble avoir peu d'influence. Comparons les critères DIC et WAIC (annexe D) :

	DIC	WAIC
Modèle altitude	-144.5606	-145.2060
Modèle champ <i>field</i>	-609.1620	-607.60515
Modèle altitude + <i>field</i>	-608.4624	-606.7850

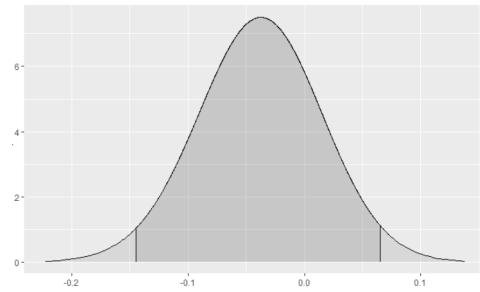
Nous remarquons que le modèle comprenant uniquement l'altitude n'est pas de bonne qualité comme nous nous y attendions, et que le modèle combiné et le modèle *field* sont de qualités similaires. Il semble donc que *Eperua falcata* soit très peu dépendante de l'altitude. Afin de confirmer cette hypothèse, nous étudions la significativité de la variable altitude plus en détail :

### Variable altitude

Modèle	Modèle altitude	Modèle altitude et field
Coefficient $\beta$	-0.06904678	-0.03269948
Significativité	OUI	NON



Loi à posteriori de  $\beta$  pour le modèle altitude



Loi à posteriori de  $\beta$  pour le modèle altitude et champ *field*

Nous remarquons que la variable altitude n'est plus significative dès lors nous rajoutons le champ *field* dans le modèle : ceci prouve que la répartition de *Eperua falcata* n'est presque pas influencée par l'altitude, ou du moins que cette influence est négligeable en comparaison de l'influence qu'a le champ *field* sur cette variable.

### 6.3 *Oenocarpus bataua*

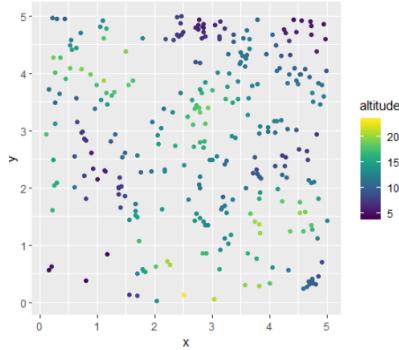


FIGURE 22 – Répartition et altitude de *Oenocarpus bataua*



Crédit photo : Michael Calonje

La répartition de *Oenocarpus bataua* est la répartition la plus homogène des trois exemples, même si nous observons une densité légèrement plus forte dans le partie Nord-Est et légèrement plus faible dans la partie Sud-Ouest. L'espèce semble préférer les zones d'altitude intermédiaires (de 5 à 20m).

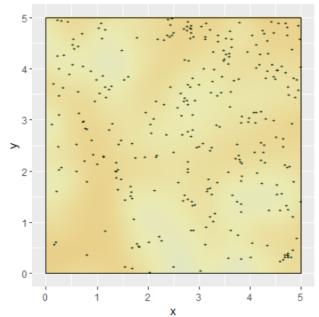


FIGURE 23 – Modèle altitude

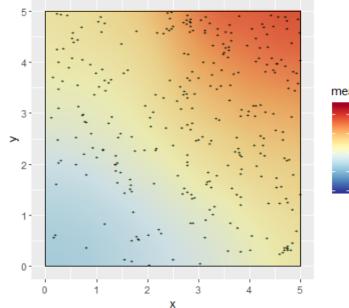


FIGURE 24 – Modèle champ *field*

Nous remarquons que la répartition intra-spécifique (champ *field*) semble avoir un influence plus forte que l'altitude sur la répartition des *Oenocarpus bataua* dans la parcelle.

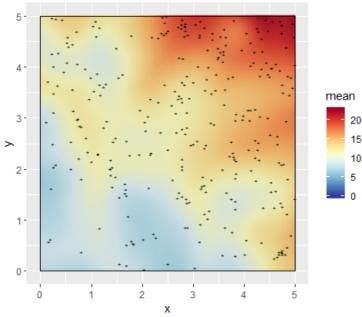


FIGURE 25 – Modèle altitude et champ *field*

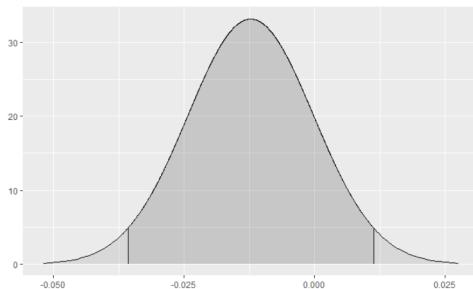
Comparons les critères DIC et WAIC (annexe D) de ces trois modèles :

	DIC	WAIC
Modèle altitude	-901.7093	-903.7281
Modèle champ <i>field</i>	-933.0700	-933.6495
Modèle altitude + <i>field</i>	-933.5876	-934.0818

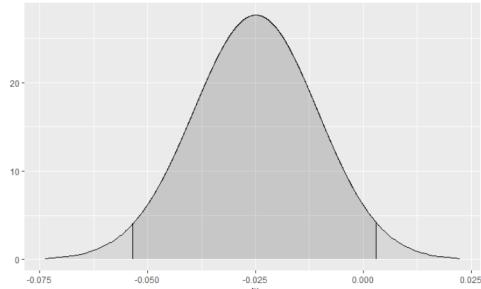
Nous observons une performance pratiquement équivalente du modèle champ *field* et du modèle altitude + *field*. Le modèle altitude est le moins bon. Détaillons l'influence de l'altitude sur la répartition de *Oenocarpus bataua* :

### Variable altitude

Modèle	Modèle altitude	Modèle altitude et field
Coefficient $\beta$	-0.01179511	-0.02434751
Significativité	NON	NON



Loi à posteriori de  $\beta$  pour le modèle altitude



Loi à posteriori de  $\beta$  pour le modèle altitude et champ *field*

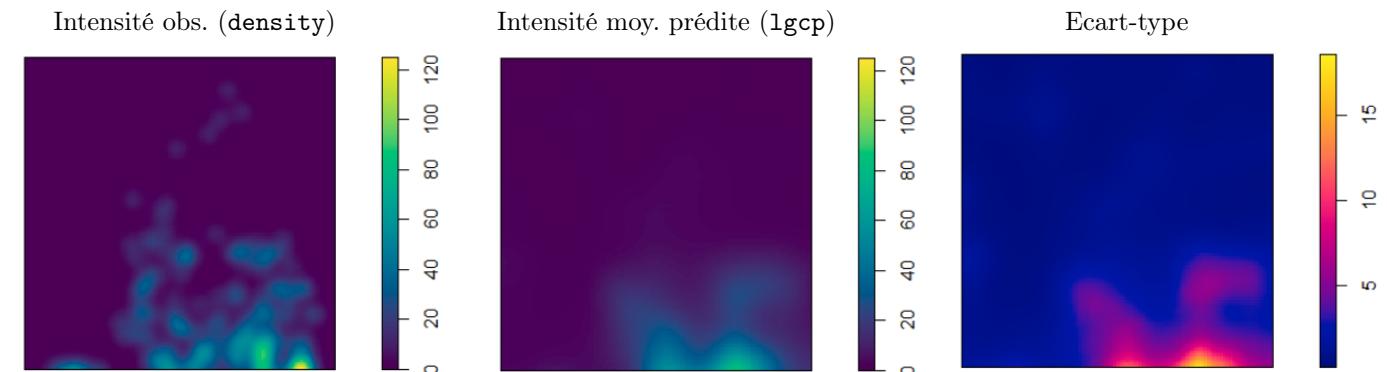
Cette fois, la variable altitude n'est significative dans aucun des deux modèles. *Oenocarpus bataua* est donc une espèce dont la répartition n'est pas significativement influencée par l'altitude.

**En conclusion**, si l'espèce tend à favoriser certaines altitudes pour son habitat, le modèle combiné est le meilleur des modèles testés. A l'inverse si l'habitat de l'espèce est indépendant de l'altitude, alors le modèle combiné et le modèle champ *field* sont de qualité similaire (précision de la prédition comparable). Il est difficile d'aller plus loin dans l'analyse sachant que nous n'avons encore que très peu de recul concernant l'application des critères DIC et WAIC à ce type de modèles. Cependant nous pouvons noter de façon certaine que la répartition intra-spécifique apporte une vraie amélioration dans la qualité de la prédition sur tous nos exemples. Cet aspect du modèle serait donc intéressant à développer dans de futures études.

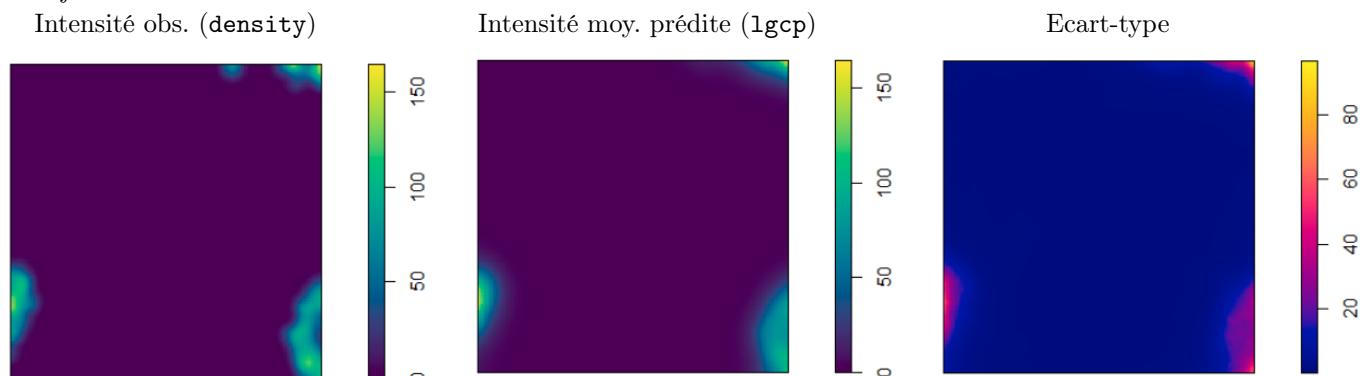
Il semble également que les problèmes rencontrés dans la section 5.4.1 ne nous concernent pas sur ces exemples. L'ajout de la covariable altitude malgré ce que nous pensions, n'est pas la raison de ce dénouement car les résultats sont déjà relativement bons dans le modèle champ *field*. Il est plus probable que les exemples biologiques dont nous disposons ont des champs d'intensité (qualité du sol, altitude, eau...) dont les changements sont beaucoup moins abruptes que dans l'exemple 5.4.1. Nous ne pouvons malheureusement pas tester notre modèle sur l'exemple problématique de Cox puisqu'il ne contient pas de variable altitude (approche systématique).

Analysons cependant un peu plus en détail nos trois exemples :

*Voucapoua americana*



*Eperua falcata*



*Oenocarpus bataua*

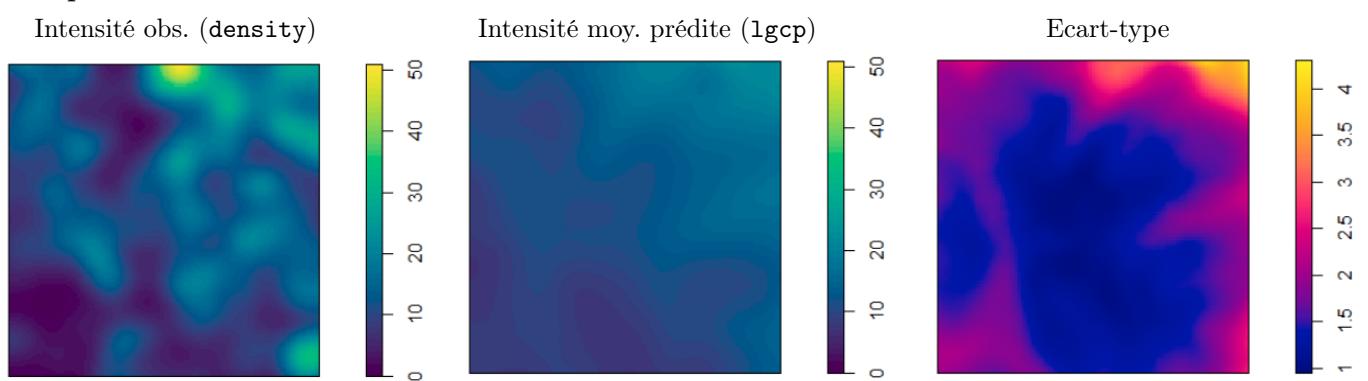
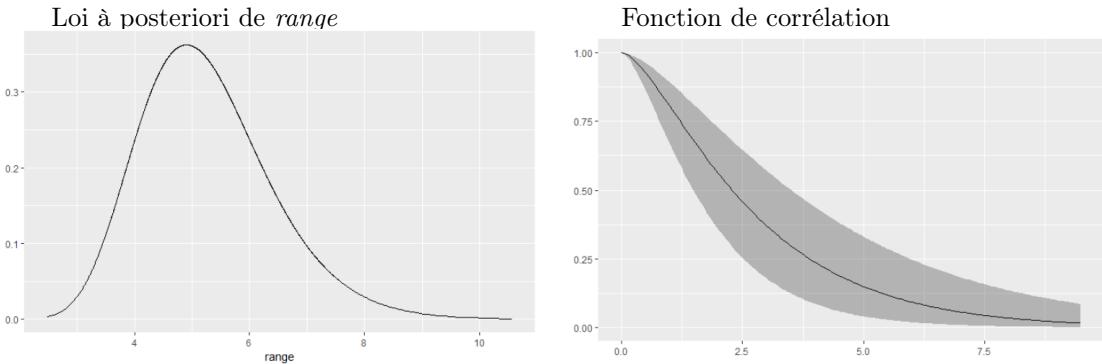


FIGURE 26 – Tableau récapitulatif de l'intensité observée, moyenne et de l'écart-type pour le modèle altitude et champ *field*.

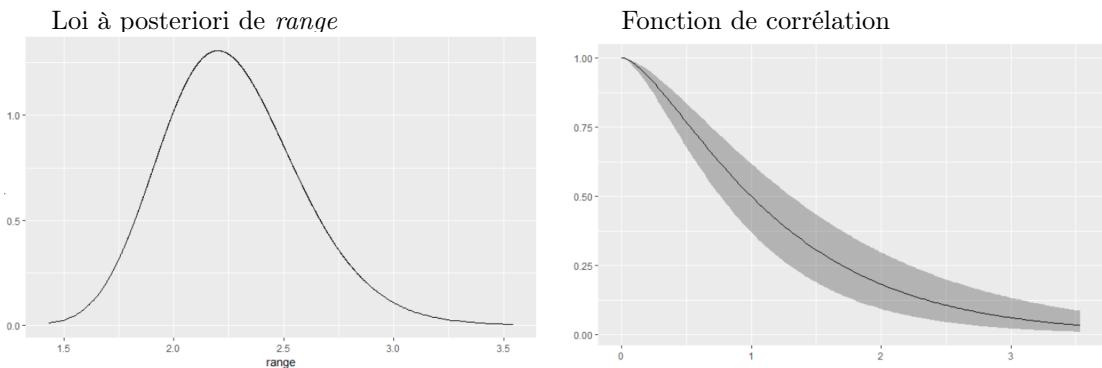
Nous remarquons que pour les deux premières espèces (*Voucapoua americana* et *Eperua falcata*), l'intensité observée est très proche de l'intensité moyenne prédictive, ce qui est satisfaisant. Dans le cas de la dernière espèce (*Oenocarpus bataua*), nous observons une légère homogénéisation de l'intensité moyenne prédictive. Celle-ci reste cependant raisonnable et peut s'expliquer au niveau de la variabilité introduite par l'écart-type du champ. La dernière observation que nous pouvons faire concerne également l'écart-type ; il semble que celui-ci augmente lorsque les données sont très agrégées, ce qui fait sens puisque les changements d'intensité sont alors beaucoup plus abruptes.

Pour finir nous nous intéressons au paramètre *range* qui devrait nous retourner la distance à partir de laquelle la fonction de corrélation est nulle, c'est à dire la distance minimale entre deux points telle que l'existence d'un des points n'a plus aucune influence sur la probabilité d'obtenir l'autre point.

### *Voucapoua americana*



### *Eperua falcata*



## *Oenocarpus bataua*

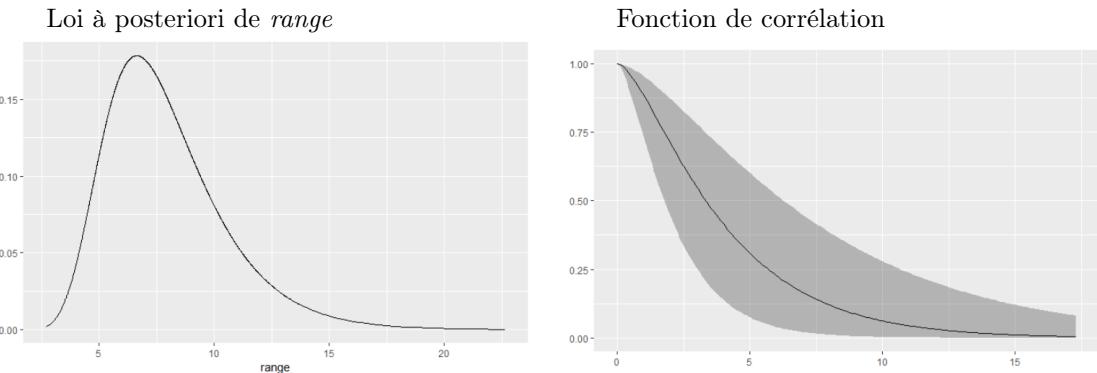


FIGURE 27 – Tableau récapitulatif du paramètre *range* en centaines de m et de la fonction de corrélation pour chaque espèce (modèle altitude et champ *field*)

L'évolution du paramètre *range* semble être cohérente avec le degré d'agrégation des trois espèces étudiées : plus les données sont agrégées, plus la valeur de *range* est faible.

En théorie, la corrélation est de presque zéro aux distances pour lesquelles la probabilité que *range* soit supérieur est très faible. Il est donc normal qu'à la valeur la plus probable de *range*, la corrélation soit encore assez éloignée de zéro car il y a encore une probabilité non négligeable pour que *range* soit en réalité plus grand.

## 7 Limites du modèle

Nous avons étudié précédemment la dépendance intra-spécifique de la répartition des arbres. Cependant il pourrait aussi être très intéressant d'un point de vue biologique d'obtenir des informations sur la dépendance inter-spécifique de la répartition des arbres. En d'autres termes, nous pourrions détecter des phénomènes d'attraction ou de répulsion entre deux espèces.

Dans un premier temps, nous avons considéré le cas où le semis ne contiendrait que deux espèces distinctes. Nous avons réfléchi à plusieurs approches afin de mettre en évidence ces interactions, seulement même dans un cas très simple comme celui-ci, aucune d'elles ne semble applicable en pratique. L'idée est d'étudier la fonction de covariance (dépendant toujours uniquement de la distance) construite grâce aux informations apportées par les couples mixtes et seulement ceux-là. Par mixte, j'entends les couples formés par un individu de chaque espèce. Cette fonction représente l'interaction de ces deux espèces en

fonction de la distance qui les sépare. Si nous notons A et B les deux espèces, notons  $\text{cor}(A,B)$  cette fonction. Enfin notons  $\text{cor}(A,A)$  (resp.  $\text{cor}(B,B)$ ) la fonction construite de façon similaire à partir des seuls individus de l'espèce A (resp. de l'espèce B).

Plusieurs possibilités s'offrent alors à nous :

- si les deux espèces sont présentes dans le processus étudié, alors la fonction de corrélation est calculée grâce aux résultats de  $\text{cor}(A,B)$ , mais aussi grâce aux résultats de  $\text{cor}(A,A)$  et de  $\text{cor}(B,B)$ , ce qui fausse totalement notre analyse,
- placer l'espèce B en covariable de l'espèce A n'est non plus pas une option car la fonction de corrélation est calculée grâce au champ aléatoire *field*, qui lui même se base sur le pattern observé ce qui reviendrait donc à ne considérer que  $\text{cor}(A,A)$
- nous avons aussi considéré l'idée de construire deux champs aléatoires *field* pour chacune des espèces, seulement il ne semble pas possible de spécifier deux pattern de points distincts à la fonction `lgcp` (l'argument `data` est unique),
- enfin nous avons tenté d'approfondir les options que nous offraient les processus ponctuels marqués. L'idée est de considérer les deux espèces dans le processus, en spécifiant sur chaque point l'espèce à laquelle il appartient par une marque. Le modèle retourne alors deux champs distincts : un champ pour l'intensité spatiale (*field*) et un champ pour les marques. Il aurait alors été intéressant de regarder la fonction de corrélation sur le champ des marques, seulement là encore cette fonction témoigne de l'influence que peut avoir une marque A sur l'apparition d'une marque B un peu plus loin, mais aussi d'une marque A un peu plus loin. L'information serait donc également faussée dans ce cas là.

Il semble qu'à ce jour, les méthodes paramétriques ne nous permettent pas encore de répondre à cette question. Nous sommes cependant en mesure de détecter ces interactions grâce à des méthodes non paramétriques, notamment grâce à la fonction M (voir annexe E). Cette fonction est implémentée sous le nom de `Mhat` du package `dbmss` [5].

Ci-dessous, un exemple avec *Vouacapoua americana* comme espèce de référence et *Qualea rosea* comme espèce secondaire :

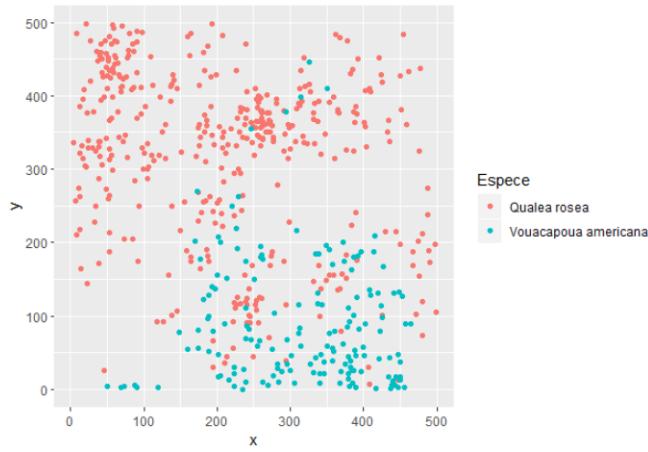


FIGURE 28 – Répartition de Vouacapoua americana et de Qualea rosea sur la parcelle 16 de Paracou

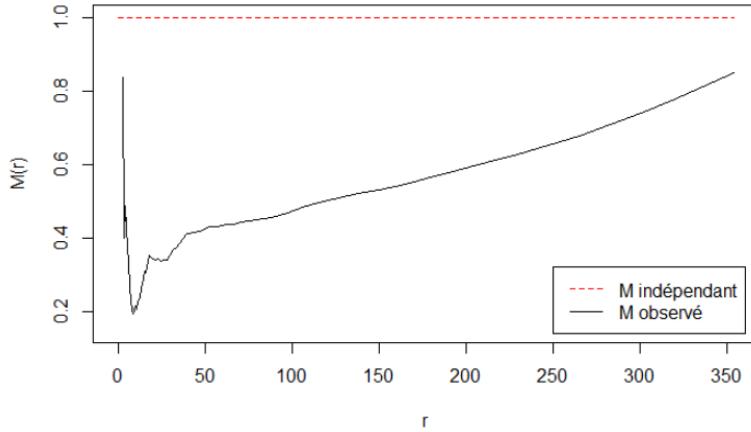


FIGURE 29 – Fonction M appliquée à l’espèce de référence Vouacapoua americana et l’espèce secondaire Qualea rosea ( $\text{Mhat}$  du package `dbmss`)

Il a été montré dans l’article de E. Marcon, F. Puech et S. Traissac [4] que ces deux espèces ont tendance à se repousser. Ici ce phénomène s’observe grâce à la fonction M : nous remarquons un rapport inférieur à 1 quelque soit la distance considérée. Il est également intéressant de noter que ce phénomène s’estompe avec la distance comme nous pouvions nous y attendre. Enfin la phase de décroissance que nous observons à très faible distance correspond en réalité à du bruit : à une distance si restreinte le nombre de voisins est très faible et le rapport extrêmement variable. Ensuite l’évolution de celui-ci se stabilise et le rapport croît vers 1.

## 8 Conclusion

L'objectif de mon stage était d'évaluer le potentiel des méthodes INLA-SPDE dans le cadre du site de Paracou. Notamment il était utile pour l'équipe EcoFoG que j'éclaircisse les points sur lesquels la documentation était trop succincte, que j'explore les possibilités offertes par cette nouvelle approche et que je rende réutilisable les modèles développés en lien avec les données de Paracou.

Le modèle qui se base sur l'altitude et l'interaction intra-spécifique résume bien la répartition au sein d'une espèce et pourrait être une piste d'adaptation pour de futures études s'intéressant à l'agrégation spatiale. La composante "interaction intra-spécifique" n'est en effet pas encore nécessairement prise en compte dans les autres modèles développés par l'équipe à ce jour. Cependant l'interaction spatiale inter-spécifique qui constituait le réel espoir de cette méthode s'avère encore inenvisageable pour le moment. A ce jour, les seuls outils qui nous permettent de mettre en évidence ce type de relations sont des méthodes non paramétriques.

Mais pourquoi vouloir à ce point effectuer une transition vers les méthodes paramétriques ? Quelles sont leurs avantages ? Tout d'abord les statistiques non-paramétriques dont nous disposons (g, K, M... voir annexe E) ne sont pas très maniables. Elles s'opposent à un modèle nul qu'on ne sait généralement pas paramétriser avec les covariables qui nous concernent. Ensuite le grand avantage des méthodes paramétriques repose sur le fait qu'elles permettent de simuler une population partageant les mêmes propriétés que la population observée. Cette propriété peut être utile à de nombreuses études, par exemple pour des modèles individu-centrés<sup>1</sup> où il pourrait être souhaitable d'avoir différentes situations de départ "plausibles", afin de tester la puissance du modèle. Simuler une gamme de populations similaires peut également servir à calculer des intervalles de confiance pour des statistiques décrivant l'organisation forestière. Par exemple si nous souhaitons étudier la proportion d'arbres d'une certaine espèce dans une parcelle, il nous suffit de simuler de nombreuses fois le peuplement sur cette parcelle afin d'obtenir la valeur des quantiles de la statistique qui nous intéresse.

Pour conclure, il est intéressant de noter que de manière générale, la modélisation de l'organisation et de l'évolution de la forêt guyanaise a de nombreuses visées pratiques comme théoriques. Elle peut servir en sylviculture<sup>2</sup>, pour simuler différents scénarios d'exploitation et nous permettre de choisir la solution minimisant notre impact sur l'écosystème. Elle peut permettre d'anticiper les bouleversements liés au réchauffement climatique, de les quantifier et de prévenir

---

1. Les modèles individu-centrés disposent pour chaque arbre d'une loi pour grandir et mourir qui dépend des paramètres environnants de l'arbre. Un processus de régénération est également intégré au modèle afin de reproduire la dispersion des graines au sein de la forêt. On dispose alors d'un simulateur forestier pas à pas comprenant des règles bien précises pour passer du temps t au temps t+1.

2. Exploitation rationnelle des arbres forestiers (entretien, reboisement, etc.).

des conséquences sur la forêt. Enfin la modélisation permet également d'acquérir des connaissances théoriques sur le milieu forestier tropical humide, en étudiant par exemple la résilience de la biodiversité face à différentes perturbations. Ces études contribuent à une meilleure compréhension de l'organisation générale de l'écosystème, sans impacter de quelque façon que ce soit le milieu naturel.

## 9 Accès au code

Le script R permettant de refaire les analyses et les figures de ce rapport est disponible sur [Github](#) et sur le réseau EcoFoG à l'adresse \\roucou.ecofog.gf\Users\Lenaklay\Public\Stage.

## Références

- [1] E. ALLIÉ. « Assemblage des communautés d’arbres à une échelle locale en forêt tropicale : Apport d’une approche intégrative ». Thèse de doct. École doctorale Diversités, santé et développement en Amazonie (Cayenne), 2016. URL : [http://www.ecofog.gf/greybase/files/allie/2016/265\\_Allie2016.pdf](http://www.ecofog.gf/greybase/files/allie/2016/265_Allie2016.pdf).
- [2] E. KRAINSKI et al. *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Gitbook or CRC Press/Taylor et Francis Group, 2018. URL : <https://becarioprecario.bitbucket.io/spde-gitbook/index.html>.
- [3] F. LINDGREN, H. RUE et J. LINDSTRÖM. « An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields : The Stochastic Partial Differential Equation Approach ». In : *J. R. Statist. Soc.B Vol.73 No.4* (2011), p. 423–98. URL : <https://pdfs.semanticscholar.org/0a7c/66e70b84b983fa5e85e8af90b81d88b19856.pdf>.
- [4] E. MARCON, F. PUECH et S. TRAISSAC. « Characterizing the Relative Spatial Structure of Point Patterns ». In : *International Journal of Ecology (Article ID 619281)* (2012), p. 1–11. URL : <https://www.hindawi.com/journals/ijecol/2012/619281/>.
- [5] E. MARCON et al. « Tools to Characterize Point Patterns : dbmss for R ». In : *Journal of Statistical Software. Vol.67(3)* (2015), p. 1–15. URL : <https://www.jstatsoft.org/article/view/v067c03/v67c03.pdf>.
- [6] H. RUE, S. MARTINO et N. CHOPIN. « Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations ». In : *J. R. Statist. Soc.B Vol.71 Part.2* (2009), p. 319–39. URL : <https://inla.r-inla-download.org/r-inla.org/papers/inla-rss.pdf>.

# Annexes

## A Maillage

Afin d'effectuer la regression d'un processus non-homogène et d'obtenir la valeur de l'intensité en tout point du domaine, nous devons définir une approximation discrète de ce dernier sous la forme d'un maillage.

Dans le package `inlabru`, ce maillage (*mesh*) est un maillage triangulaire défini sur deux zones : une zone interne et une zone externe. Nous avons décidé de définir la zone interne comme étant notre domaine, et de laisser libre choix à la fonction pour la frontière externe. Les paramètres à définir dans ces deux zones sont par exemple : la taille minimale et maximale d'un côté, l'angle maximal et minimal autorisé dans un triangle, le nombre de noeuds à ne pas dépasser...

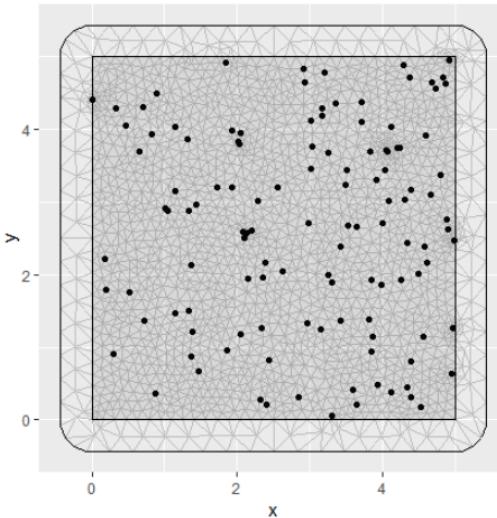


FIGURE 30 – Mesh adapté à un exemple de points.

Le maillage est une approximation discrète du domaine. En effet lors de la régression, le modèle est évalué uniquement sur les noeuds du maillage. On comprend alors qu'un trop grand nombre de noeuds résultera en un temps de calcul très long, mais à l'inverse si le nombre de noeuds est insuffisant notre approximation risque de ne pas être assez précise.

Une application shiny a été mise en place dans le package `INLA` (commande `meshbuilder()`) afin de rendre plus accessible la paramétrisation du maillage.

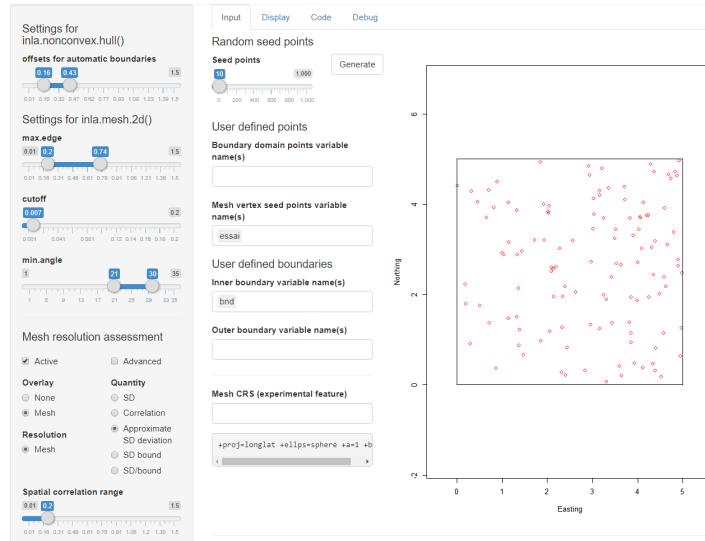


FIGURE 31 – Interface du package INLA pour la construction d'un maillage

Grâce à cette interface, nous pouvons notamment évaluer la qualité du maillage grâce à l'outil "Approximate standard deviation" (onglet : Display). Cet outil retourne en chaque point le rapport entre la variance du modèle discret et celle du modèle continu désiré. Un mesh de bonne qualité tendra donc vers 1 pour ce critère au moins dans la zone interne du maillage.

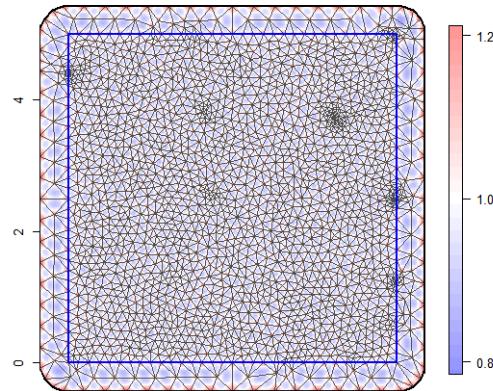


FIGURE 32 – Indice de qualité : Approximate standard deviation

## B Covariance et corrélation de Matérn

### B.1 Fonction de covariance

Dans le cas du modèle de Cox non-homogène, le logarithme de l'intensité est un champ aléatoire gaussien. Comme tout champ gaussien, celui-ci est défini par sa moyenne en chaque point et sa fonction de covariance pour chaque paire de points. (On rappelle que la variance en tout point peut-être déduite de la fonction de covariance par la formule  $\text{cor}(X, X) = \text{var}(X)$ .)

Pour une question de simplicité, il est très courant que la fonction de covariance dépende uniquement de la distance euclidienne  $r$  séparant les variables du champ : on dit alors que la fonction de covariance est isotrope. Cette simplification est justifiée car elle reste fidèle à ce que l'on peut observer d'un point de vue biologique. On remarque également empiriquement que la fonction de covariance est généralement nulle lorsque que la distance est plus grande qu'un certain seuil (car si deux arbres sont trop éloignés, il n'y a plus d'interactions entre eux). Cette distance limite va être également intégrée à notre modèle, sous le nom de *range*.

La fonction de covariance isotrope choisie dans le package **INLA** est celle de Matérn, car elle confère certains avantages détaillés dans la section 3.1. Elle est définie comme suit :

$$c(r) = \sigma^2 \frac{(\alpha r)^\nu}{2^{\nu-1} \Gamma(\nu)} \kappa_\nu(\alpha r)$$

avec

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt \text{ la fonction gamma}$$

$$\kappa_\nu(x) = \frac{\pi}{2} \frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\pi\nu)} \text{ la fonction de Bessel modifiée de deuxième espèce}$$

$$\text{et } I_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{+\infty} \frac{1}{k! \Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}$$

$$\nu > 0 \text{ (*smooth parameter*)}$$

$$\alpha > 0 \text{ (*scale parameter*) qui peut aussi être écrit } \alpha = \frac{2\sqrt{\nu}}{l} \\ \text{avec } l \text{ (*correlation length*)}.$$

On pose  $\rho = \frac{\sqrt{8\nu}}{\alpha}$  (*range parameter*). Pour  $r = 0$ ,  $c(r) = \sigma^2$  est la variance. Il est alors possible de définir la fonction de covariance Matérn seulement avec  $\rho$  (range) et  $\sigma$  (sigma) : c'est le choix qui a été fait dans le package **INLA** et ceux seront donc deux des paramètres que nous chercherons à déterminer dans notre

régression. Ils apparaissent sous cette forme en sortie du modèle :

param <fctr>	mean <dbl>	var <dbl>	sd <dbl>	lq <dbl>	median <dbl>	►
range	4.1242321	0.672239494	0.81990212	2.7563351	4.0390892	
log.range	1.3975471	0.039091464	0.19771561	1.0123327	1.3957847	
variance	0.3103198	0.007253303	0.08516632	0.1761721	0.2989736	
log.variance	-1.2065429	0.073659563	0.27140295	-1.7383159	-1.2076920	

FIGURE 33 – Exemple de paramètres de la fonction de covariance Matérn.

On retrouve ces mêmes paramètres dans une autre fenêtre du summary :

- Range for field indiquant la distance à partir de laquelle la fonction de covariance est nulle (si les deux variables sont trop éloignées, elles n'ont plus d'influence l'une sur l'autre),
- Stdev for field qui indique l'écart-type du champ aléatoire ( $\sigma$ ).

	mean <dbl>	sd <dbl>	0.025quant <dbl>	0.5quant <dbl>	0.975quant <dbl>	►
Range for field	4.1252283	0.82613209	2.7513545	4.0403608	5.982607	
Stdev for field	0.5520801	0.07530927	0.4191794	0.5469125	0.714752	

FIGURE 34 – Exemple de paramètres de sortie du champ aléatoire gaussien.

## B.2 Fonction de corrélation

Pour deux variables aléatoires X et Y, la définition générale de la corrélation est la suivante :

$$cor(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \quad \text{avec } \sigma_{\bullet} \text{ l'écart type de } \bullet$$

Dans notre cas, la fonction de corrélation dépend donc uniquement de la distance r et si on note  $\gamma$  cette fonction de corrélation, on a naturellement pour toute variable aléatoire X :

$$cor(X, X) = \gamma(0) = 1$$

On rappelle que pour toute variable aléatoire X :

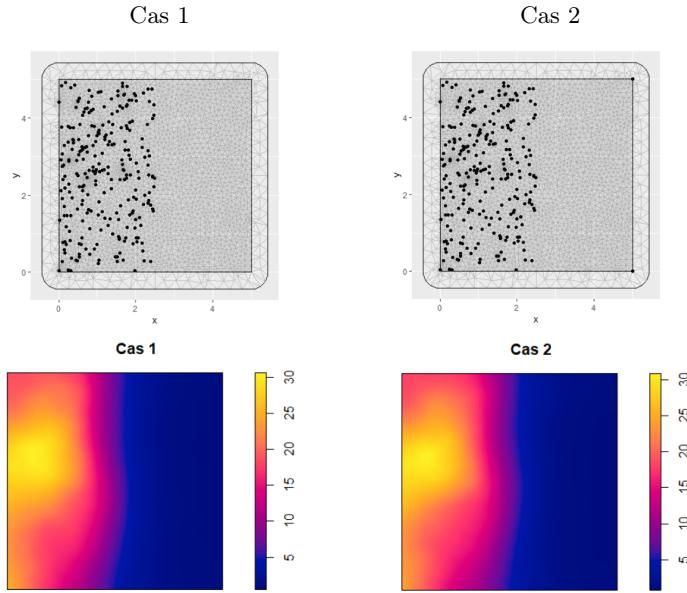
$$cov(X, X) = c(0) = \sigma_X^2$$

ce qui confirme ce que l'on vient de dire plus tôt (avec c la fonction de covariance).

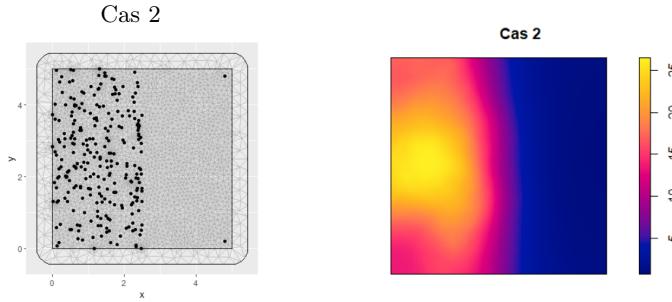
## C Etudes complémentaires

### C.1 Test de la fonction `lgcp` dans le cas non-homogène

Nous reprenons ici l'expérience menée dans la section 5.3.1 en effectuant cette fois-ci une régression de Cox non-homogène.



Les résultats sont cohérents et correspondent à nos attentes. Nous pouvons cependant noter que la régression de Cox non-homogène néglige les points situés dans les coins du carré; cela pourrait laisser penser que le modèle est moins bon aux frontières... Cependant si nous refaisons l'expérience en rapprochant les points du centre, nous nous apercevons que la régression les néglige toujours :



Cette expérience nous montre que l'existence d'un point seul n'a pas un impact fondamental dans le calcul de l'intensité par la fonction `lgcp`.

## D Critères pour la comparaison des modèles

Il existe de nombreux critères pour comparer l'efficacité de deux modèles, dont bons nombres se basent sur la fonction de déviance. Soit la vraisemblance définie par :

$$L(\theta|x) = f(x|\theta)$$

avec  $f(x|\theta)$  la distribution à priori,  $x$  les observations et  $\theta$  le paramètre du modèle. La déviance est alors définie comme :

$$D(\theta) = -2 \log L(\theta|x)$$

Plus la déviance est grande, moins le modèle est ajusté aux données. Cependant utilisée seule, la déviance n'est pas une bonne mesure discriminante puisqu'elle est biaisée en faveur des modèles de grandes dimensions (phénomène de sur-apprentissage). Il est donc intéressant de rajouter un deuxième terme qui pénalise la complexité du modèle, comme dans les critères AIC ou BIC :

$$AIC(\mathcal{M}) = D(\theta) + 2\dim(\theta)$$

$$BIC(\mathcal{M}) = D(\theta) + \log(n)\dim(\theta)$$

avec  $\mathcal{M}$  le modèle considéré et  $n$  le nombre d'observations.

### Le critère DIC

Un des critères que nous allons considérer pour la comparaison de nos modèles est le critère DIC (*Deviance information criteria*) :

$$\begin{aligned} DIC(\mathcal{M}) &= E(D(\theta)|x) + p_{DIC} \\ &= E(D(\theta)|x) + \{E(D(\theta)|x) - D(E(\theta)|x)\} \\ &= D(E(\theta)|x) + 2p_{DIC} \end{aligned}$$

Le facteur  $E(D(\theta)|x)$  peut être vu comme une mesure d'ajustement aux données, tandis que  $p_{DIC}$  évalue la complexité du modèle (c'est le *nombre effectif de paramètres*). Dans le cas d'un modèle bayésien, ce critère est facilement calculable grâce à la méthode de Monte-Carlo par chaînes de Markov, puisque nous cherchons une approximation de  $E(D(\theta)|x)$  et non plus  $D(\theta)$ .

### Le critère WAIC

Le deuxième critère que nous utilisons dans notre étude est le WAIC (*Watanabe-Akaike information criterion*). Il est également composé de deux termes : lppd (*log-pointwise-predictive-density*) :

$$lppd = -2 \sum_{i=1}^n \log L(\theta|x_i)$$

qui est considéré comme l'analogue de la déviance point par point, et  $p_{\text{WAIC}}$  :

$$p_{\text{WAIC}} = 2 \sum_{i=1}^n \{E(\log L(\theta|x_i)^2|x) - E(\log L(\theta|x_i)|x)^2\}$$

Le critère WAIC est alors donné par :

$$\text{WAIC}(\mathcal{M}) = lppd + p_{\text{WAIC}}$$

La particularité de ce critère réside dans le fait qu'il évalue la flexibilité d'un modèle en fonction de l'ajustement à chacune des observations. Cela peut s'avérer très utile dans le cas où certaines observations comportent des incertitudes différentes.

#### **Brève bibliographie des critères de sélection :**

AIC	H. Akaike (1973)
BIC	G.E. Schwarz (1978)
DIC	D.J. Spiegelhalter et al. (2002)
WAIC	S. Watanabe (2010)

Pour plus de précisions, l'article [Understanding predictive information criteria for Bayesian models](#) (A. Gelman et al., 2013) met en parallèle ces différents critères de façon très intéressante.

## E Statistiques non-paramétriques

Les statistiques non-paramétriques sont des outils permettant de décrire la structure d'un semis de points sans se restreindre à une famille de modèle (avec un nombre de paramètres fini) au préalable. Pour ce faire, elles opposent la distribution observée à un modèle nul. Nous détaillons ici quelques uns des outils non-paramétriques utilisés dans l'étude des phénomènes d'agrégation spatiale. L'hypothèse nulle à laquelle nous nous confrontons est la distribution aléatoire des points (pas d'agrégation ni de répulsion), c'est à dire un processus de poisson homogène d'intensité égale à l'intensité moyenne du processus observé.

### Fonction g de densité des paires de points

La fonction  $g(.,.)$  décrit la probabilité de présence conjointe de deux points dans les surfaces élémentaires  $dS_1, dS_2$  centrées en  $x_1, x_2$ .

$$\begin{aligned}\mathbb{P}(\text{présence d'un point dans } dS_1 \text{ et d'un point dans } dS_2) &= \\ \mathbb{P}(n(dS_1) = 1 \cap n(dS_2) = 1) &= \lambda(x_1)\lambda(x_2)g(x_1, x_2)dS_1dS_2\end{aligned}$$

Cette fonction reflète la propriété du second ordre, c'est à dire les phénomènes d'agrégation et de répulsion. Dans le cas de l'hypothèse nulle, c'est à dire une distribution indépendante des points, g est constante égale à 1. Si le processus étudié est stationnaire (invariable par translation) et isotrope (invariable par rotation), on dit alors la distribution est homogène au second ordre. La fonction g ne dépend alors que de la distance r qui sépare  $x_1$  et  $x_2$  et est proportionnelle au nombre de couples de points à distance r l'un de l'autre. Cette hypothèse concorde bien avec nos observations des processus forestiers, ce qui simplifie beaucoup les calculs. Dans la suite, nous nous plaçons dans ce cadre.

### Fonction K de Ripley

La fonction K de Ripley est un indicateur cumulatif qui se base sur la fonction g :

$$K(r) = \int_0^r g(u)2\pi u \, du$$

Cette fonction est égale au rapport du nombre de voisins attendu dans un disque de rayon r sur l'intensité moyenne du processus (le disque peut-être centré n'importe où sur le domaine, tant qu'il y est entièrement inclu). Cette valeur est à comparer avec une distribution indépendante des points entre eux, pour laquelle  $g(r)=1 \forall r$  et donc  $K(r) = \pi r^2 \forall r$ .

- si  $K(r) > \pi r^2$ , la densité à l'intérieur du disque de rayon r est plus forte que sur l'ensemble du domaine : les points sont agrégés.
- si au contraire  $K(r) < \pi r^2$ , la densité dans le disque est plus faible : les points ont tendance à se repousser.

Un estimateur naturel de K(r) autour d'un point x du domaine est :

$$\hat{K}(r) = \frac{n(V_{x,r})}{\hat{\lambda}}$$

avec  $n(V_{x,r})$  le nombre de voisins de  $x$  dans un rayon  $r$  et  $\hat{\lambda}$  l'intensité observée moyenne qui est le rapport du nombre de points du semis sur l'aire du domaine. En tenant en compte des effets de bord (si un point du semis est proche du bord, son nombre de voisins observés est bien moindre...), il est possible de calculer une bonne approximation de la fonction  $K$  sur le domaine.

### Fonction M

La fonction  $M$  compare la proportion de points d'intérêt dans un certain voisinage à celle que l'on observe sur l'ensemble du domaine. Elle se présente donc sous la forme d'une comparaison de ratio qui peut témoigner :

— de la répartition intra-spécifique (ici l'espèce A),

$$M(r) = \sum_{a \in A} \frac{n_A(V_{a,r})}{n(V_{a,r})} / \frac{n_A}{n}$$

— de la répartition inter-spécifique (ici l'espèce B autour de l'espèce A).

$$M(r) = \sum_{a \in A} \frac{n_B(V_{a,r})}{n(V_{a,r})} / \frac{n_B}{n}$$

où  $n$  désigne le nombre de points du semis,  $n_A$  le nombre de points du semis de l'espèce A,  $n(V_{a,r})$  le nombre de voisins du point  $a$  à une distance inférieure à  $r$ .

Lorsque la distribution des points est indépendante, on retrouve une proportion voisine égale à la proportion sur le domaine ; donc  $M(r)=1 \forall r$ . Une valeur supérieure de  $M(r)$  traduit un phénomène d'agrégation, tandis qu'une valeur inférieure témoigne d'une tendance à la répulsion. Les valeurs de  $M$  peuvent s'interpréter comme une comparaison de ratios : si  $M(r) = 3$ , cela signifie qu'il y a en moyenne 3 fois plus de points d'intérêts dans un rayon  $r$  autour des points considérés que sur l'ensemble du domaine.