# Modélisation de la dynamique spatio-temporelle de la rouille brune du blé à l'échelle de la France

Mehdi Boutaina

**MEHDI Boutaina**
Confidentiel

# RAPPORT DU STAGE D'APPLICATION EN STATISTIQUE DE 2E ANNEE

MEHDI Boutaina

**STRUCTURE D'ACCUEIL : Institut National de la Recherche Agronomique (INRA)-Unité de Biostatistique et Processus Spatiaux (BioSP)**

**THEME DU STAGE : Modélisation de la dynamique spatio-temporelle de la rouille brune du blé à l'échelle de la France**

LIEU DE STAGE : INRA Centre de Recherche PACA

Promotion : 2018

Maître de stage : Julien PAPAÏX & Emily WALKER

# TABLE DES MATIÈRES

# TABLE DES FIGURES

CHAPITRE

# 1

# REMERCIEMENTS

Ce stage n'aurait lieu d'être sans le déploiement des efforts de plusieurs personnes que je tiens à remercier par la présente.

Tout d'abord, j'adresse mes remerciements à mes deux encadrants Julien Papaïx et Emily Walker qui se sont toujours montrés présents pour moi et qui ont facilité mon intégration au sein de l'équipe. Je les remercie pour leur patience, gentillesse. Leurs conseils et leurs orientations m'ont permis d'avancer dans le sujet et apprendre énormément de choses en peu de temps. Leurs aide était bien plus que précieuse. Je présente mes remerciements à toute l'équipe BioSP, thésards et permanents, pour l'ambiance conviviale, et pour tous les échanges fructueux qui m'ont permis d'avoir un avant goût sur la recherche. Finalement, je remercie Meryem et Chaima pour leurs soutien indéfectible durant toute la période de stage.

# PRÉSENTATION DE L'ORGANISME DE STAGE

L'institut national de la recherche agronomique fût créé en 1946 dans le contexte d'après-guerre afin de moderniser l'agriculture française. Depuis il accompagne les changements du monde agricole, mais aussi des filières alimentaires afin de répondre aux attentes de la société, par exemple dans le domaine de la suffisance alimentaire. L'institut est placée sous la double tutelle du ministère de l'Enseignement supérieur et de la Recherche et du ministère de l'Agriculture et de la Pêche. Avec la mondialisation, les prérogatives de l'INRA ont changé. Parmi ses principales missions, on trouve l'amélioration de l'agriculture en termes de performance économique, sociale aussi bien qu'environnementale ;le développement de systèmes alimentaires sains et durables, la valorisation de la biomasse, ou encore l'atténuation et l'adaptation au changement climatique.

L'institut possède actuellement un dispositif de recherche décentralisé, et mutualisé comptant prés de 8000 salariés répartis sur dix-sept centres de recherche en France, s'organise en treize départements de recherche, et est dirigé depuis peu par Philippe Mauguin. L'INRA occupe le deuxième rang mondial et le premier en Europe pour le nombre de publications en sciences agricoles et en sciences de la plante et de l'animal (Voir Organigramme en annexe A)

Le centre de recherche Provence-Alpes-Côte d'Azur (PACA) rassemble 1000 agents, dont 700 agents permanents, répartis dans 26 unités, localisés sur 10 sites en PACA. La présidence du centre INRA PACA est assurée par Michel Bariteau et les thématiques qui y sont traitées sont souvent en relation avec l'environnement méditerranéen dans lequel il s'inscrit.

Mon stage se déroule au sein de l'unité biostatistique et Processus spatiaux qui dépend du département MIA (Mathématiques et informatique appliquées) dirigé par Etienne Klein (voir Annexe B). Elle compte également des scientifiques d'autres départements parmi ses effectifs : SPE (Santé des plantes et environnement) et EFPA (Ecologie des forêts, prairies et milieux aquatiques).L'unité BioSP développe des travaux en statistique, en systèmes dynamiques, en écologie-épidémiologie, et aux interfaces entre ces différentes disciplines avec un intérêt particulier pour les questions spatiales et spatio-temporelles. Les domaines d'application de ces travaux sont avant tout l'écologie, l'agriculture et l'environnement.

# INTRODUCTION

CHAPITRE

# 2

# PRÉSENTATION DES DONNÉES ET STATISTIQUES DESCRIPTIVES

## 2.1 Cartographie des parcelles et dispersion de la rouille brune

Arvalis- Institut du Végétal a mis a notre disposition la base Vigicultures sur la rouille brune du blé de la période 2009-2016. Sur cette période, nous disposons des données dans 64 départements pour chaque parcelle suivie (témoin non traité).

Le nombre de parcelles suivies par département change d'une année à une autre. Nous avons cartographié dans la figure 2.1 le nombre de parcelles par département. Toutefois, chaque année, la maladie n'est détectée que dans certains départements. On représente dans la carte de la figure 2.2 les départements où la maladie a été observée chaque année indépendemment du nombre de parcelles atteintes.

## 2.2 L'incidence comme indicateur du développement de la rouille brune

Dans chaque parcelle témoin, 20 plantes sont prélevées et on y observe les trois dernières feuilles développées. Ensuite l'incidence de la rouille est obtenue en notant la présence ou l'absence de la maladie sur chacune des feuilles et des plantes. Il est à noter que parmi toutes les dates observées, nous avons 90.79% des observations pour les trois feuilles, 2.43% pour deux feuilles, et 6.77% pour une seule feuille. Nous avons représenté l'histogramme des données d'incidence dans la figure 2.3. La distribution de celle-ci est unimodale autour de zéro. En revanche, parmi les incidences observées, 93.22% sont des zéros. Nous avons donc un excès de 0. Afin de voir la distribution des données sans les zéro, on représente les données relatives à des incidences strictement positives (figure 2.3).

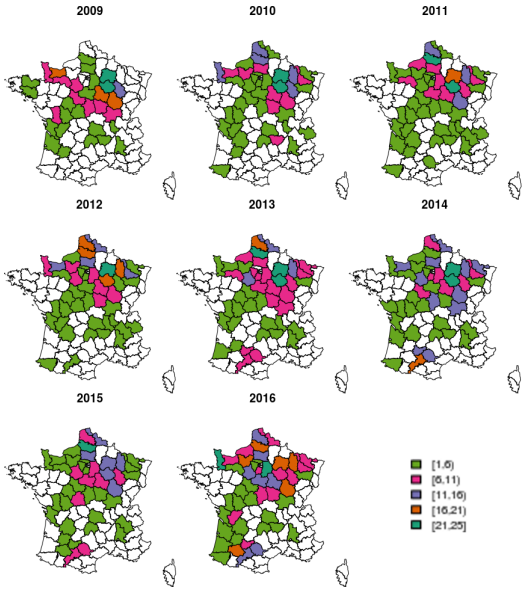



FIGURE 2.1 – Nombre de parcelles par departement

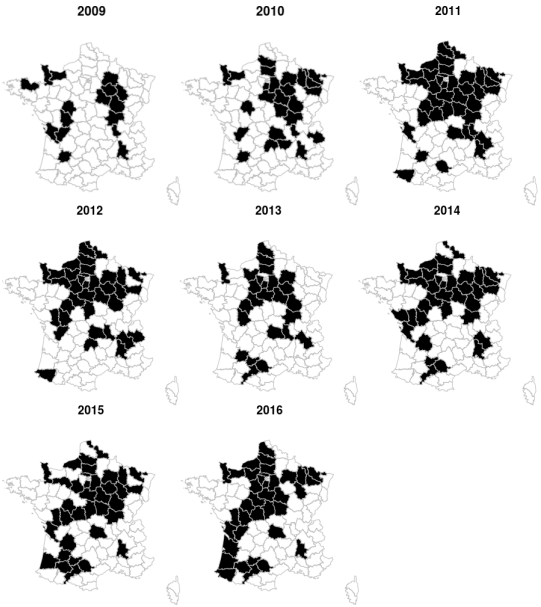



FIGURE 2.2 – Présence ou absence de la rouille brune par département

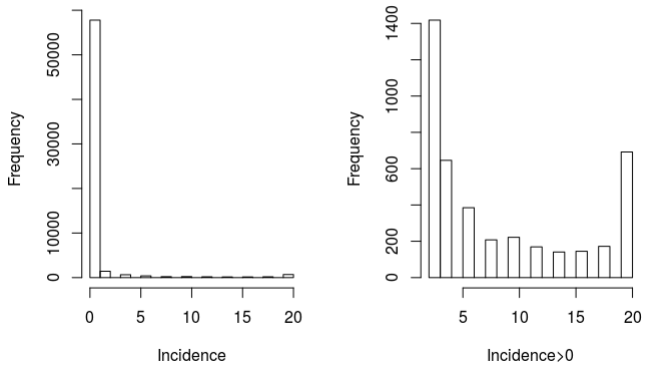



FIGURE 2.3 – Distribution de l'incidence sur toutes les données

## 2.3 Les variétés de blé

Dans les parcelles atteintes, la rouille brune se développe rapidement et s'adapte aux principales variétés en cultivées. De plus, les variétés ne sont pas toutes égales face au pathogène. Certaines témoignent d'une résistance contre la rouille, tandis que d'autres sont plus susceptibles d'être atteintes. C'est dans cette perspective qu'on inclut cette variable dans la modélisation de la propagation de la rouille.
Dans les 64 départements observés pendant la période 2009-2016, on trouve 172 variétés. L'effet de cette variable permettra de classer les variétés selon leur sensibilité à la rouille.
On représente dans la figure 2.4 les variétés les plus utilisées chaque année ( les variétés ayant été utilisées dans plus de10 parcelles).

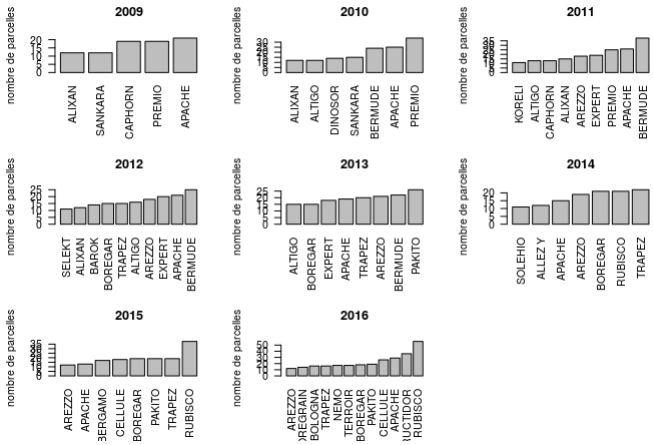



FIGURE 2.4 – Histogramme des variétés les plus utilisées 2009-2016

CHAPITRE

# 3

# MODÉLISATION DE LA DYNAMIQUE TEMPORELLE DE LA ROUILLE

## 3.1 Modélisation des observations

Nous disposons des données sur la propagation de la rouille brune dans des parcelles situées dans différents départements en France métropolitaine. Ces données représentent l'évolution dans le temps (une observation par semaine pendant la période épidémique) de l'incidence de la rouille sur 20 plantes et sur les 3 premières feuilles.

L'incidence représente le nombre de feuilles atteintes pour chaque étage foliaire (le nombre de plantes observées pour la feuille $j$ de la parcelle $i$ au temps $t$ est égal à 20), et peut donc être modélisée par une variable aléatoire $Y_{j,i,t}$ qui suit une loi binomiale $B(20, p_{j,i,t})$, avec :

- $j \in \{F1, F2, F3\}$ : l'étage foliaire
- $i \in \{1...N\}$ : la parcelle (avec $N$ le nombre de parcelles)
- $t \in \{1...T\}$ : le jour de l'observation (jour julien depuis le 01/04 de l'année étudiée)
- $p_{j,i,t}$ : la probabilité d'observer la maladie sur la feuille $j$ de la parcelle $i$ au temps $t$.

## 3.2 Modélisation hiéarchique du processus épidémique

Le processus épidémique est décrit par la variable latente $W_{i,t}$ qui représente l'intensité de la rouille dans la parcelle $i$ au temps $t$. Dans un premier temps nous supposons un modèle de croissance logistique tel que :

$$W_{i,t} = \alpha_0 + \phi_i * W_{i,t-1} * (1 - \frac{W_{i,t-1}}{K_i}) \tag{3.1}$$

où $\alpha_0$ représente l'inoculum, $\phi_i$ le taux de croissance de la rouille dans la parcelle $i$ et $K_i$ l'intensité maximale potentielle pour la parcelle $i$. A ce stade nous faisons les hypothèses suivantes :

- L'inoculum est constant en espace et en temps,

— L'intensité maximale est la même pour toutes les parcelles et fixée pour assurer l'identifiabilité (indétermination entre $\lambda$ et K, voir l'equation (3.2)).
— Le taux de croissance dépend de la variété cultivée, $log(\phi_i) = \phi_{0v_i}$.

L'incidence observée $Y_{j,i,t}$ est reliée à l'intensité de rouille $W_{i,t}$ via la probabilité $p_{j,i,t}$ d'observer une feuille malade via la fonction de lien suivante :

$$p_{j,i,t} = 1 - e^{-\lambda_j . W_{i,t}} \tag{3.2}$$

où $\lambda_j$ représente la partie des spores accessible à la feuille $j$.
La fonction de lien utilisée ici a été obtenue en supposant que le nombre de spores atteignant l'étage foliaire $j$ suit une loi de Poisson dont la moyenne est proportionnelle à l'intensité de rouille $\lambda_j.W_{i,t}$. En effet, sous cette hypothèse, la probabilité qu'il n'y ait aucune lésion sur une feuille donnée est $e^{-\lambda_j.W_{i,t}}$. Ainsi, la probabilité qu'il y ait au moins une lésion sur la feuille, c'est à dire la probabilité que la feuille soit dite malade, est égale à $1 - e^{-\lambda_j.W_{i,t}}$. On suppose ici que la détection de la maladie est parfaite.

En somme, nous nous plaçons dans le cadre des modèles hiérarchiques et menons l'inférence dans un cadre bayésien. Le graphe acyclique orienté représenté dans la figure 3.2 met en évidence la structure hiéarchique du modèle.

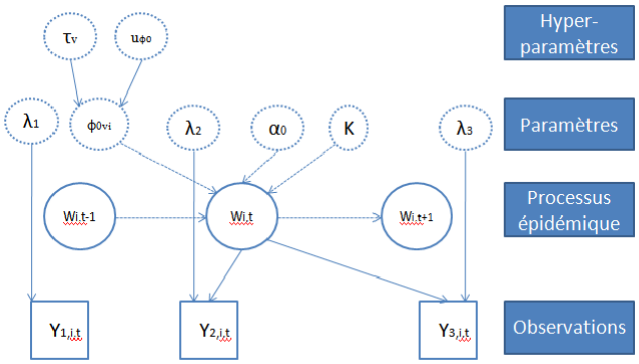



FIGURE 3.1 – Graphe acyclique orienté

## 3.3 Estimation des paramètres du modèle par une approche bayésienne

Dans le cadre bayésien, on donne des distributions *a priori* aux paramètres, et l'ajustement du modèle prend en compte à la fois ces *a priori* et les données (via la vraisemblance) pour donner une distribution *a posteriori* jointe des paramètres du modèle via la formule de Bayes : $\pi(\theta|y) = \frac{\pi(y,\theta)}{\pi(y)} = \frac{\mathscr{L}(y|\theta)\pi(\theta)}{\pi(y)}$
avec :
- $\pi(\theta)$ la loi *a priori* du modèle
- $\pi(y)$ la vraissemblance marginale
-$\mathscr{L}(y|\theta)$ la vraissemblance des données

Pour la suite on pose $\theta_y = (\lambda_1, \lambda_2, \lambda_3)$ et $\theta_w = (\phi_{0v_i}, \alpha_0)$.
De plus, les lois normales $\mathcal{N}(\mu, \tau^2)$ sont paramétrées par défaut par leur espérance $\mu$ et leur précision $\tau^2$ au lieu de leur variance ($\sigma^2$) pour assurer la cohérence avec le programme utilisé ensuite sous JAGS et on a $\tau^2 = \frac{1}{\sigma^2}$.

On choisit pour les lois *a priori* des paramètres, des lois peu informatives :

$\alpha_0 \sim \mathcal{U}[0, 1000]$,
$\phi_{0v_i} \sim \mathcal{N}[\mu_v, \tau_v]$,
$\sigma_v \sim \mathcal{U}[0, 10]$,

$\forall j \in \{1, 2, 3\}, \lambda_j \sim \mathcal{U}(0, 1)$,
$\mu_v \sim \mathcal{N}[0, 1]$

On peut alors écrire la vraisemblance du modèle :

$$\mathscr{L}(y|\theta_y, W) = f(y|\theta_y, W) = \prod_{i=1}^{N} \prod_{j=1}^{3} \prod_{t=1}^{100} f_{i,j,t}(y_{i,j,t}|\theta_y, W)$$

$$f_{i,j,t}(Y_{i,j,t} = y_{i,j,t}|\theta_y, W) = \binom{20}{y_{i,j,t}} * p_{j,i,t}^{y_{i,j,t}} * (1 - p_{j,i,t})^{20 - y_{i,j,t}}$$

Afin d'estimer les paramètres de notre modèle, on utilise la méthode de Monte-Carlo par chaîne de Markov qui permet de construire des chaînes de Markov qui ont pour lois stationnaires les distributions à échantillonner. Dans cette perspective, on utilise un programme JAGS (Just Another Gibbs Sampler) destiné à l'analyse de modèles bayésiens hiérarchiques, qui implémente l'algorithme de Gibbs (échantillonneur). JAGS va donc nous permettre d'échantillonner la loi *a posteriori* de nos paramètres.

Après avoir determiné les lois *a priori* des paramètres du modèle ainsi que la vraissemblance, nous avons lancé l'estimation du modèle de dynamique temporelle sur 3 chaînes de 30000 itérations avec un thinning interval de 150 afin d'éviter les autocorrélations (consulter annexe C).

## 3.4 Diagnostic de convergence et qualité d'ajustement

### 3.4.1 Diagnostic de convergence des chaînes

L'inférence sur les paramètres à partir des distributions *a posteriori* n'est valide que si les chaînes MCMC associées aux paramètres d'intérêt convergent. Pour s'assurer de cela, on utilisera la règle de convergence proposée par Gelman Rubin qui consiste à mesurer s'il existe une différence significative entre la variance intra et inter-chaînes.

Le package *"coda"* sous R fournit *Shrink factor* ( Facteur de réduction ) pour juger de la convergence. Une valeur égale à 1 signifie qu'il y'a égalité entre la variance inter et intra-chaînes tandis qu'une valeur différente de 1 signifie qu'il existe encore une différence significative entre les chaînes. Une règle de décision mise par Gelman est qu'une valeur inférieur ou égale 1.1 justifie une convergence adéquate.
Nous avons vérifié la convergence des chaînes pour tous les paramètres du modèle ( Voir annexe D ). Nous détaillons ci-dessous le diagnostic de convergence pour les paramètres du taux de croissance $\phi_i$. On remarque que les trois chaînes se mélangent bien entre elles et convergent vers la même dis-

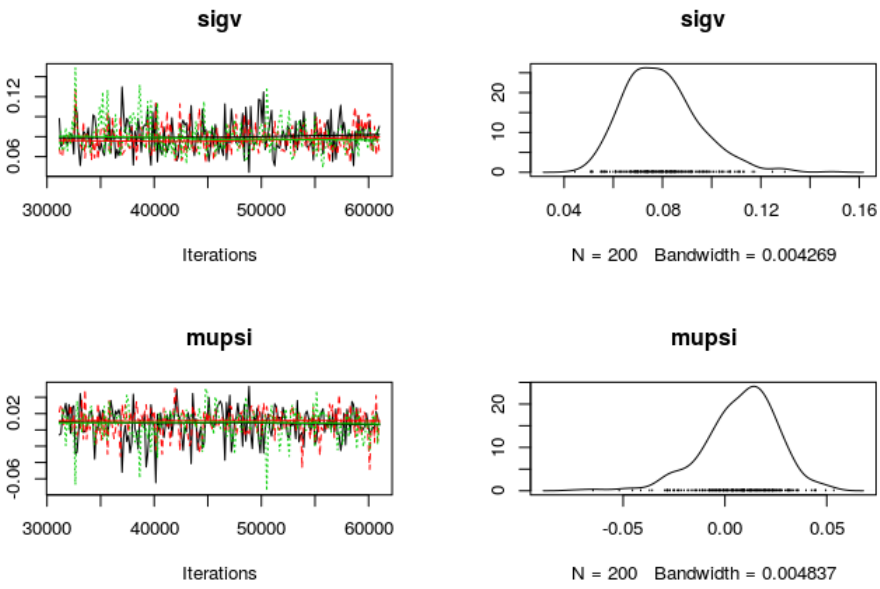



FIGURE 3.2 – Diagnostic de convergence pour le taux de croissance

tribution *a posteriori*. La règle de Gelman fournit des facteurs de réduction très proches de 1 ( voir figure 3.3), ce qui confirme la convergence des trois chaînes.

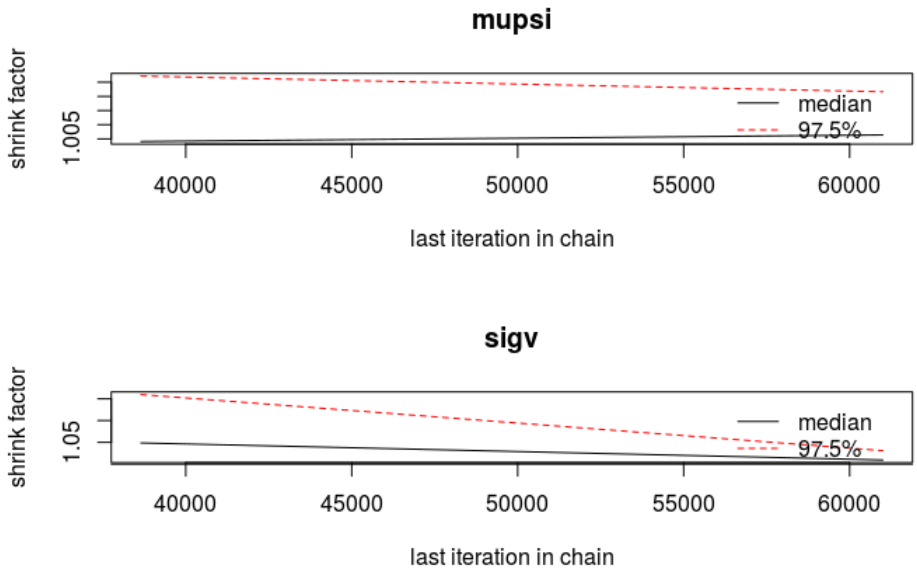



FIGURE 3.3 – Diagnostic de Gelman-Rubin

### 3.4.2 Qualité d'ajustement

Afin de juger de l'ajustement du modèle, nous avons tracé le graphe de l'incidence prédite en fonction de l'incidence observée pour les trois feuilles. Le nuage de points obtenu sera comparé à la première bissectrice. Dans la figure 3.4, on remarque à première vue que le modèle a tendance à sous-estimer les incidences élevées et à sur-estimer les incidences faibles. Toutefois, pour certains points l'incidence prédite est très proche ou égale à celle réalisée .On constate également que nous avons un excès de zéros par rapport à la loi qu'on stipule. Une option à envisager est de surdisperser la loi en 0 en multipliant l'incidence par la probabilité que la maladie soit observée. De plus, une saturation de l'incidence à vingt est remarquable, il faut probablement revoir le niveau de la capacité de charge K.

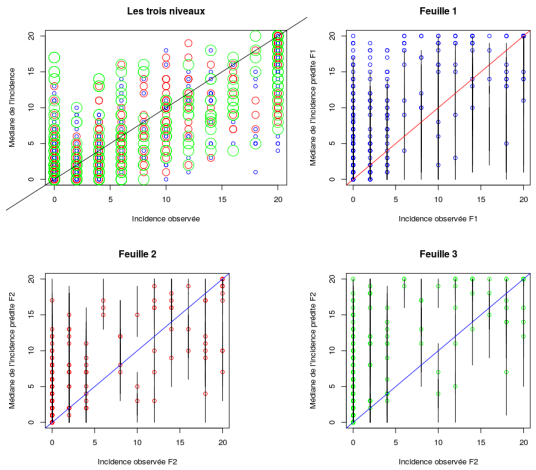



FIGURE 3.4 – L'incidence prédite en fonction de l'incidence observée pour F1,F2,F3

Par ailleurs, en se basant sur les observations prédites pour chaque point, nous avons calculé un intervalle de confiance au niveau de 95% dont le but est d'avoir un taux de couverture de nos observations dans l'intervalle de confiance. On obtient un taux de 95.85% en faisant un ratio entre le nombre d'observations d'incidence qui s'inscrivent dans leurs intervalle de confiance par rapport au nombre total d'observations non manquantes. Sans les zéros , on obtient un taux de couverture égale à 67,44%. Il est à noter que le modèle 3.1 ne différencie pas de façon directe entre les étages foliaires, c'est dans ce sens que nous avons calculé un taux de couverture des étages foliaires dans leur ensemble. On peut conclure que les prédictions obtenues sont assez bien dans ce premier modèle de croissance logistique qui tient compte de l'effet de la variété. Cependant certaines sous-estimations et sur-estimations restent apparentes et sont à améliorer(voir figure 3.5)

Nous avons regardé plus en détails les probalilités d'observer la maladie pour une parcelle malade (Voir figure 3.6 ). On remarque que la plupart probabilités $p_{j,i,t}$ d'observation de la maladie sont dans l'intervalle de confiance construit à un niveau de risque de 5%. Le graphe d'ajustement pour cette parcelle confirme la cohérence des probabilités retrouvées. Toutefois, pour les parcelles pour lesquelles il y'a de fortes sur-estimations ou sous-estimations de l'incidence, nous pouvons se tromper quant au jugement sur la probabilité d'observation de la maladie. D'où la nécessité d'améliorer notre premier modèle.

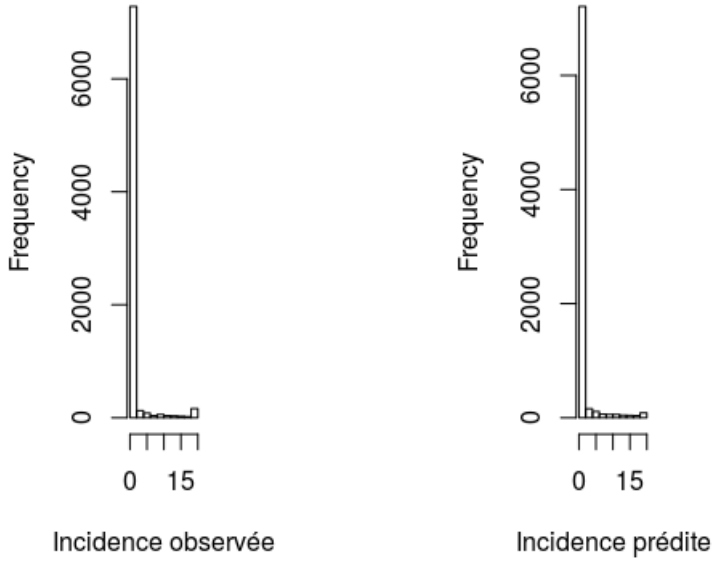



FIGURE 3.5 – Histogramme de l'incidence pour F1,F2,F3

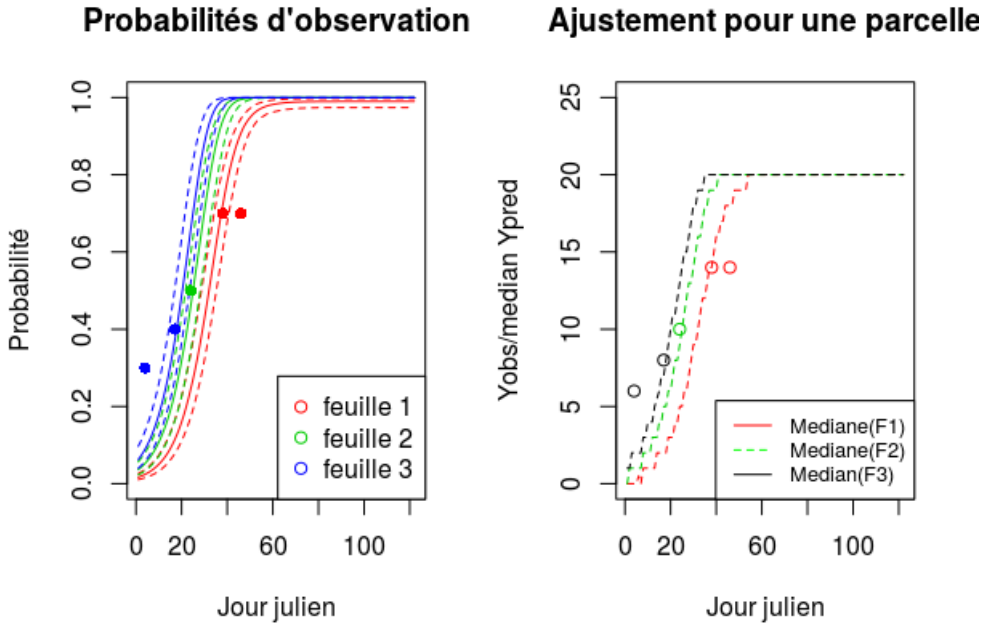



FIGURE 3.6 – Probabilité d'observation de la maladie et ajustement pour une parcelle dans le temps

## 3.5 Analyse des résultats

### 3.5.1 L'inoculum

Nous avons choisi une loi *a priori* peu informative pour l'inoculum $\alpha_0$.
D'après la densité et quantiles de la loi *a posteriori* de $\alpha_0$ de la figure 3.7, l'estimation de ce paramètre est très faible, ce qui peut être relié au fait que l'inoculum est considérée indépendemment pas de

la parcelle.En revanche, étant donnée la nature autorégressive du modèle, on peut dire qu'il y'a un effet de cumul sur l'inoculum.

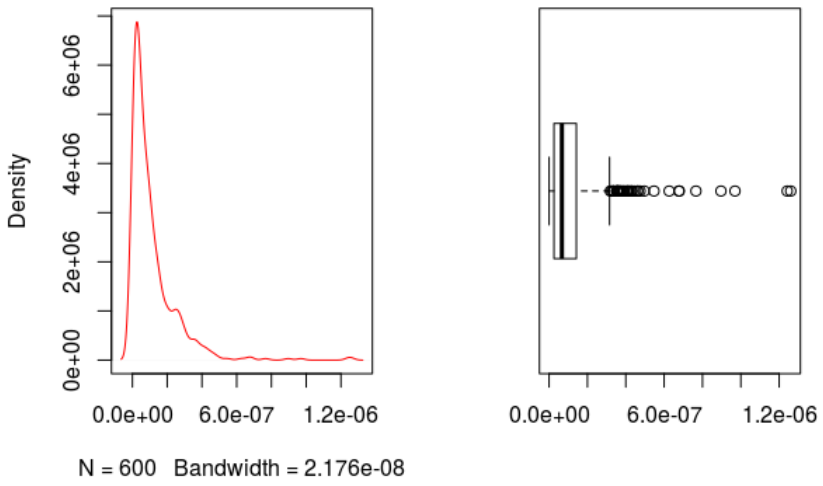



FIGURE 3.7 – Densité de la loi *a posteriori* de $\alpha_0$ et ses quantiles

### 3.5.2 Les étages foliaires

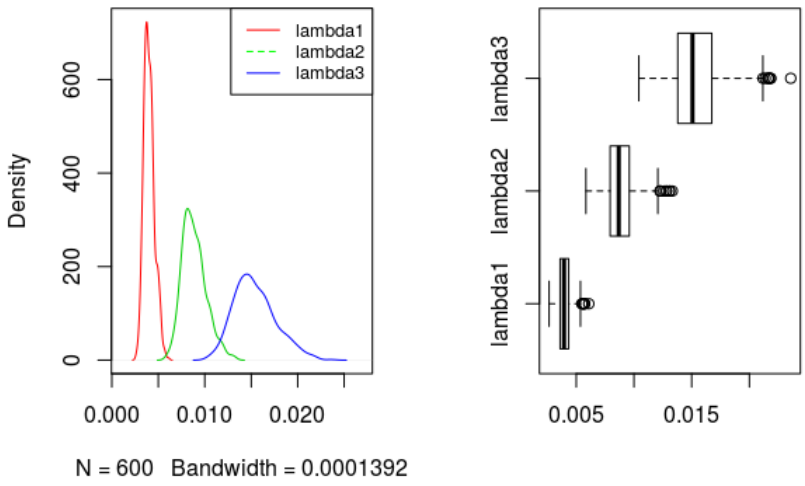



FIGURE 3.8 – Densité de la loi *a posteriori* de $\lambda_1$, $\lambda_2$, $\lambda_3$ et leurs quantiles

En se référant aux densités de la loi *a posteriori* pour chacun des paramètres de la figure 4.7, on remarque que $\lambda_3$>$\lambda_2$>$\lambda_1$, par conséquent les spores de la maladie sont accessibles aux feuilles dans cet ordre : F3, F2, et F1, ce qui est certainement lié à un effet d'ancienté des feuilles. Une possibilité serait d'introduire l'âge de la feuille au lieu de sa position.

### 3.5.3 Les variétés

On examine d'abord les paramètres de la loi que nous avons supposé*a priori* sur $\phi_i$( $\phi_{0v_i} \sim \mathcal{N}[\mu_v, \tau_v]$) : On constate sur la figure 3.11 que $\mu_v$ est approximativement centré en 0.01 et ne prend

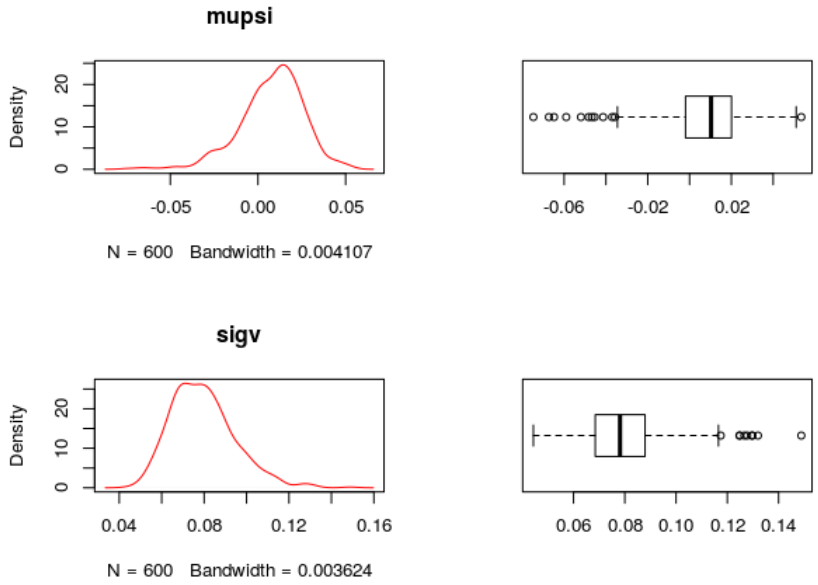



FIGURE 3.9 – Densité de la loi *a posteriori* de $\mu_v$ et $\sigma_v$ et leurs quantiles

que des valeurs proches de 0. Quant à la variance de $\phi_i$, elle est non nulle, avec un écart type médian de 0.08 et des valeurs dispersées (Intervalle de confiance plus large). On pourrait s'attendre alors à un effet significatif de la variété dans le modèle.

A travers les quantiles des lois *a posteriori* de $\phi_i$ de la figure 3.10, on constate que certaines variétés sont atteintes de la rouille plus que d'autres( plus malades/ moins malades). Afin de ressortir les

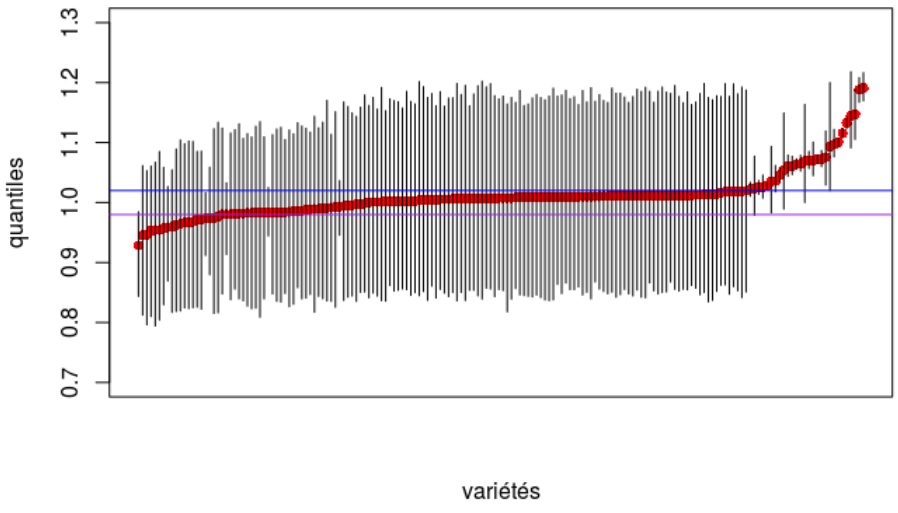



FIGURE 3.10 – Quantiles des lois *a posteriori* de $\phi_i$ (effet variété)

varités les plus/moins malades qui se démarquent, nous avons représenté dans la figure 3.11 les variétés ayant un $\phi i$>1.02 et $\phi i$< 0.98. On note que certaines variétés ont un intervalle de confiance assez large, il s'agit des variétés pour lesquelles on n'a pas beaucoup de données ( on retrouve ces variétés dans 17 parcelles seulement dans toute la france).En effet, la médiane de $\phi_i$de ces variétés est centré en 1, elles reproduisent donc la loi *a priori* posée sur $\phi_i$.

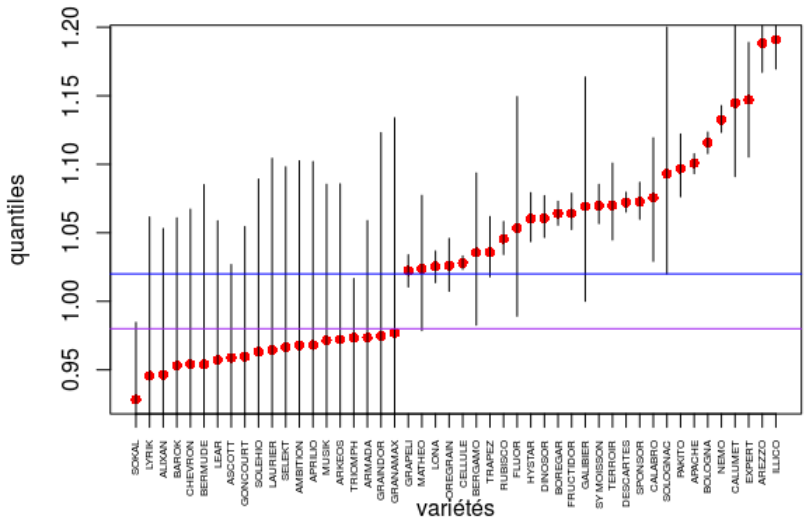



FIGURE 3.11 – Quantiles des lois *a posteriori* pour les variétés les plus/moins malade

### 3.5.4 Bilan du premier modèle

D'un point de vue **statistique**, la convergence de notre modèle a été bien validée en lançant l'algorithme MCMC sur 30 000 itérations. Les paramètres du modèle ont été bien estimés et l'ajustement a été jugé plutôt bien. Toutefois, certaines sous-estimations et sur-estimations de l'incidence prédite ressortent, d'où la nécessité d'améliorer notre premier modèle. Par ailleurs, étant donné que la dynamique de propagation de la rouille est étroitement liée à des conditions climatiques qui la favorise ainsi qu'à la dimension spatiale, nous allons donc introduire dans la partie suivante une dimension spatiale dans la modélisation.

CHAPITRE

# 4

# MODÉLISATION HIÉARCHIQUE SPATIALE

Dans une perspective d'amélioration de nos prédictions, nous avons choisi d'introduire une structure spatiale qui pourrait porter soit sur l'inoculum ou sur le taux de croissance qui dépend de la variété cultivée. Le premier cas peut être relié au fait que l'inoculum n'est pas constant en espace et peut varier d'une zone géographique à une autre, tandis que le deuxième peut être justifié par le fait que la croissance de la maladie dépend à la fois de la variété et de l'espace, dimension qui inclut plusieurs paramètres tels que les conditions climatiques et la proximité entre les parcelles. Dans cette partie, nous allons explorer la première possibilité. Par ailleurs, l'inoculum a été supposé constant en espace et en temps dans le premier modèle, hypothèse forte que nous allons relâcher partiellement, en supposant une structure spatiale sur $\alpha_0$.

## 4.1 Modélisation spatiale hiéarchique du processus épidémique

En utilisant les même notations que dans le chapitre 3, on écrit un modèle de croissance logistique spatial tel que :

$$W_{i,t} = \beta_i + \phi_i * W_{i,t-1} * (1 - \frac{W_{i,t-1}}{K_i}) \tag{4.1}$$

Dans ce cas, l'inoculum est constant dans le temps seulement et dépend de la parcelle i tel que : $log(\beta_i) = \alpha_0 + S_i$ . Avec :

- $\alpha_0$ est l'inoculum moyen
- $S_i$ représente un effet spatial tel que : $cov(S_i, S_{i'}) = C(S_i, S_{i'})$ avec $C$ une fonction de covariance qui doit vérifier certaines conditions :

  - $C$ est une fonction paire $C(-h) = C(h)$
  - $C(0)$ est constante
  - Peut être négatif $\mid C(h) \mid <= C(0)$
  - Il faut que C(h) soit définie postive

Soit :$S = \begin{pmatrix} S_1 \\ \vdots \\ S_{N_i} \end{pmatrix}$

On suppose que $S \sim \mathcal{NMV}\,[0, \sum]$ avec $\sum_{ii'} = C(Si, S_{i'})$ tel que :

$C(Si, S_{i'}) = \begin{cases} \tau^2 + \sigma^2 & \text{si } dist(i, i') = 0 \\ \sigma^2 exp(-\psi * dist(i, i')) & \text{sinon.} \end{cases}$

Avec dist(i,i') représente la distance entre deux parcelles i et i' , $\sigma$ un paramètre d'erreur, et $\psi$ la portée.La matrice de covariance, peut prendre plusieurs formes, nous avons utilisé une matrice de type exponentiel.

L'estimation des paramètres du modèle requiert l'inversion de la matrice de covariance, ce qui peut être coûteux en temps de calcul. Afin d'accélérer les calculs, on va utiliser une technique pour réduire la dimension de $\sum$ qui consiste à introduire un processus latent prédictif sur un nombre restreint de points.L'idée principale derrière le processus prédictif, est que l'ensemble représentatif des points du processus prédictif dans le domaine spatial devrait contenir suffisamment d'informations pour estimer le processus latent. En effet, pour choisir les points du processus prédictif, nous avons découpé la carte de France en une grille régulière. Les points du processus prédictif sont ensuite choisis en prenant le centroïde de chaque carré de la grille ( Voir la figure 4.1).

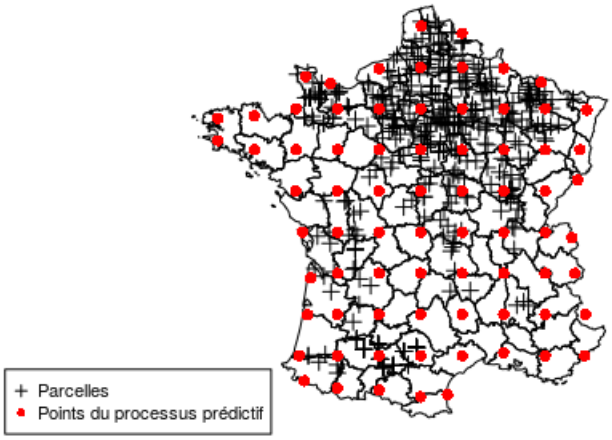



FIGURE 4.1 – Carte des parcelles et points du processus prédictif

On définit un processus spatial S* sur les points du processus prédictif tel que : $S^* \sim \mathcal{NMV}\,[0, \sum^*]$

Soit :

— $C.OBS.PP_{(N,N.PP)}$ la fonction de covariance entre les parcelles et le processus prédictifqui est de type exponentiel. Un élément de cette matrice s'écrit de la façon suivante :

$$C.OBS.PP[i,i'] = \sigma_s^2 * exp[-\psi * \frac{dist\_OBSPP[i,i']}{1000}] \tag{4.2}$$

tel que :

- $N$ : le nombre de parcelles
- $N.PP$ :le nombre de points du processus prédictifs( voir figure 4.1)
- $dist\_OBSPP_{(Ni,N.PP)}$ : la matrice de distance dans l'espace entre les points du processus prédictif et les parcelles.

- $C.PP_{(N.PP,N.PP)}$ la fonction de covariance entre les points du processus prédictif qui est de type exponentiel. Un élément de cette matrice s'écrit de la façon suivante :

$$C.PP[i,i'] = \sigma_s^2 * exp[-\psi * \frac{dist\_PP[i,i']}{1000}] \tag{4.3}$$

Avec :

- $dist\_PP_{(N.PP,N.PP)}$ : la matrice de distance dans l'espace entre les points du processus prédictif.

- Le vecteur des effets spatiaux du processus prédictif $S^* \sim \mathcal{MVN}\,[0, \sum^*]$ avec $\sum^*$=$[C.PP]$

Avec ces notations précédentes, nous pouvons écrire le vecteur des effets spatiaux :

$$S_{(Ni,1)} = C.OBS.PP_{(Ni,N.PP)} * [C.PP]^{-1}_{(N.PP,N.PP)} * S^*_{(N.PP,1)} \tag{4.4}$$

En utilisant les propriétés sur les matrices, on en déduit que S suit une loi normale multivariée d'espérance 0 et de matrice variance-covariance qui dépend des distances dans l'espace entre les points du processus prédictif et entre ces derniers et les parcelles.
Cependant, la réduction de la dimension de la matrice de covariance peut mener à une sur-estimation des variations à faible échelle de $W_{i,t}$, il faut donc envisager une variabilité aditionnelle qui peut être vue comme un bruit associé à la réplication de mesure dans la parcelle $i$. On rajoute une erreur additionnelle $S'$ tel que : $log(\beta_i) = \alpha_0 + S_i + S'_i$

On suppose que l'erreur additionnel $S'_i \sim \mathcal{N}\left[0, \tau_{S'_i}\right]$ tel que :

$$\tau_{S'_i} = \frac{1}{{\sigma_{S'_i}}^2} = {\sigma_s}^2 + {\tau_s}^2 - {}^t\,({}^t e_i * C.OBS) * [C.PP]^{-1} * {}^t\,({}^t e_i * C.OBS.PP) \tag{4.5}$$

où :

- $\tau_s$ est supposé null
- $e_i$ est le vecteur colonne où tous les éléments sont nulls sauf la i-ème ligne.

## 4.2 Estimation des paramètres du modèle spatial par une approche bayésienne

En suivant la même démarche que dans le chapitre précédent, on suppose des lois *a priori* sur les paramètres du modèle. On garde les même lois supposées pour le modèle de base sauf pour $\alpha_0$ :

$\alpha_0 \sim \mathcal{U}[0, 0.001]$, $\sigma_s \sim \mathcal{U}[0, 1]$
$\psi \sim \mathcal{U}[0.0001, 1]$,

On estime les paramètres du modèle par la méthode MCMC en le lançant sur trois chaînes en 30000 itérations avec un thinning de 300 pour éviter les autocorrélations.

## 4.3 Diagnostic de convergence et qualité d'ajustement

### 4.3.1 Diagnostic de convergence

Nous avons vérifié la convergence pour les paramètres du modèle en suivant la même démarche détaillée dans le chapitre précédent (Voir annexe F). Par ailleurs, tous les paramètres ont convergé sauf pour $\alpha_0$ et $\psi$ pour lesquels seulement deux chaînes se mélangent bien entre elles. Il serait donc judicieux de lancer le modèle sur un nombre d'itérations plus grand (voir figure 4.2 et 4.3)

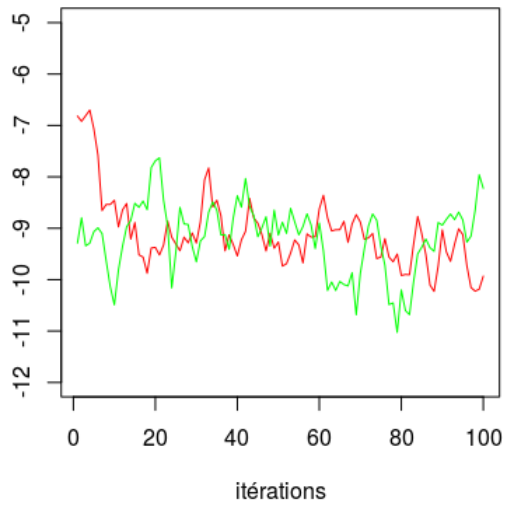



FIGURE 4.2 – convergence des deux chaînes de $\alpha_0$

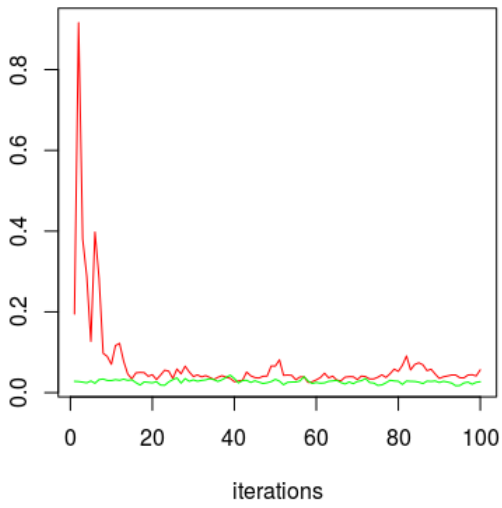



FIGURE 4.3 – convergence des deux chaînes de $\psi$

### 4.3.2 Qualité de l'ajustement

En procédent de la même manière que dans le modèle de base, on constate sur les figures 4.4 et 4.5 qu'on n'a pas un gain remarquable en terme d'ajustement. Les sur-estimations et sous-estimations persistent encore et le taux de couverture des observations dans l'intervalle de confiance s'élève à 68,8% (sans les incidences égales à zéro) soit un gain de 1.36% par rapport au premier modèle.

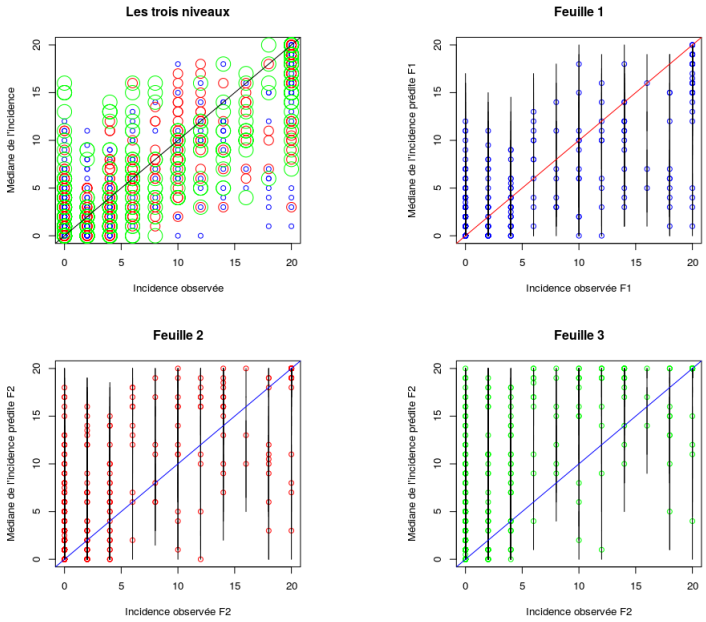



FIGURE 4.4 – L'incidence prédite en fonction de l'incidence observation pour F1,F2,F3

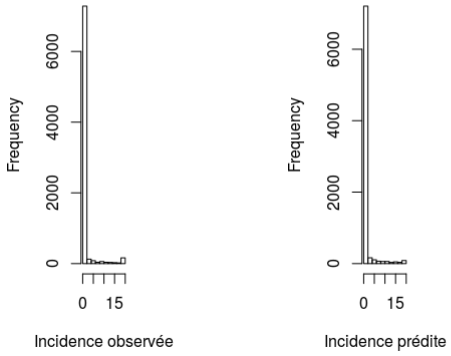



FIGURE 4.5 – Histogramme de l'incidence pour F1,F2,F3

## 4.4 Analyse des résultats

### 4.4.1 L'inoculum moyen

On représente la densité de la loi a posétriori pour les deux chaînes qui ont convergé dans la figure 4.6. L'estimation du paramètre est significativement non nulle et est négative avec une médiane de -8. L'inoculum est donc faible comme précédemment.

### 4.4.2 Les étages foliaires

Identiquement au chapitre précédent, on obtient les estimations des paramètres telles que : $\lambda_1$, $\lambda_2$, $\lambda_3$ et les spores sont plus accessibles à la Feuille 3 ( puis la feuille 2, et la feuille 1).

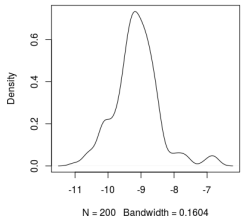



FIGURE 4.6 – Densité de la loi *a posteriori* de $\alpha_0$ sur les deux chaînes

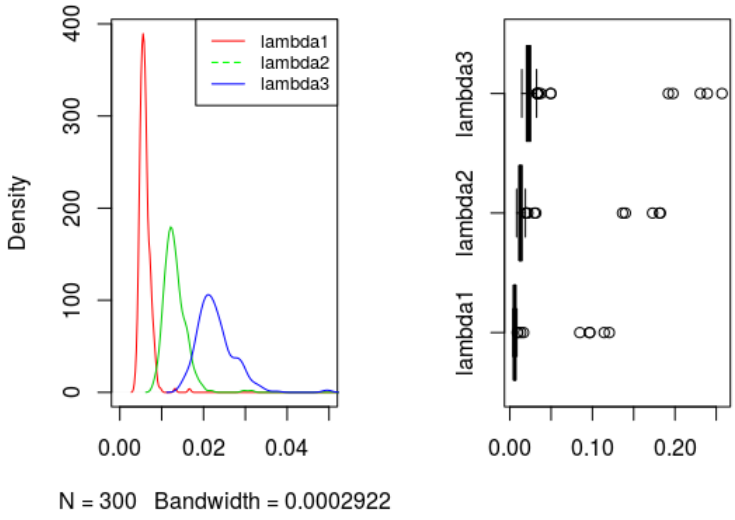



FIGURE 4.7 – Densité de la loi *a posteriori* de $\lambda_1$, $\lambda_2$, $\lambda_3$ et leurs quantiles

### 4.4.3 Les variétés

On remarque que $\mu_v$ est centrée sur -2, la variance est également non nulle avec un écart type médian de 2. On peut dire que l'effet variété serait plus important dans ce modèle. Par ailleurs, nous avons retrouvé la significativé de l'effet de la variété dans ce modèle également (Voir figure 4.9 ). Toutefois, on remarque que le gradient des variétés plus ou moins malades est plus apparent, avec moins de variétés très malades comparé au modèle précédent . De surcroît, certaines variétés dont les valeurs médianes de $\phi_i$ entre les deux seuils se démarquent. IL s'agit de six variétés, utilisées dans 35 parcelles seulement en 2016.

On regarde de plus près les variétés les plus malades ayant un $\phi_i$>0.97 ( voir figure 4.10). Ce modèle reste relativement en adéquation avec le précédent car ces variétés présentées ci-dessous ont été classifiées "plus malade"dans le modèle de base.

### 4.4.4 La portée

L'estimation du paramètre $\frac{1}{\psi}$ est centrée à 30km . Usuellement en géostatistique, on peut dire que les corrélations sont significatives à une échelle de $3.\frac{1}{\psi}$, qui s'élève dans notre cas à 90km. (Voir figure 4.11). On peut dire qu'il y'a un effet de l'espace.

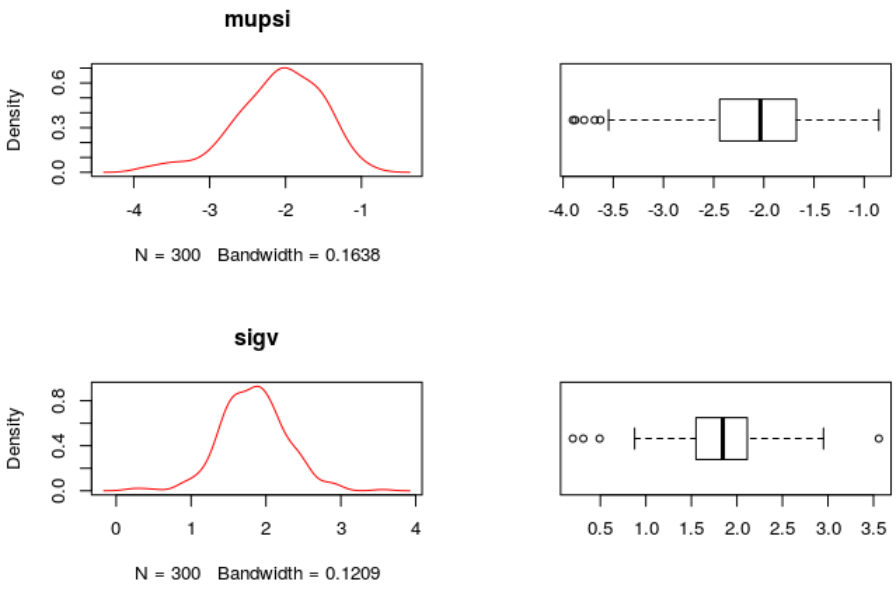



FIGURE 4.8 – Densité de la loi *a posteriori* de $\mu_v$ et $\sigma_v$ et leurs quantiles

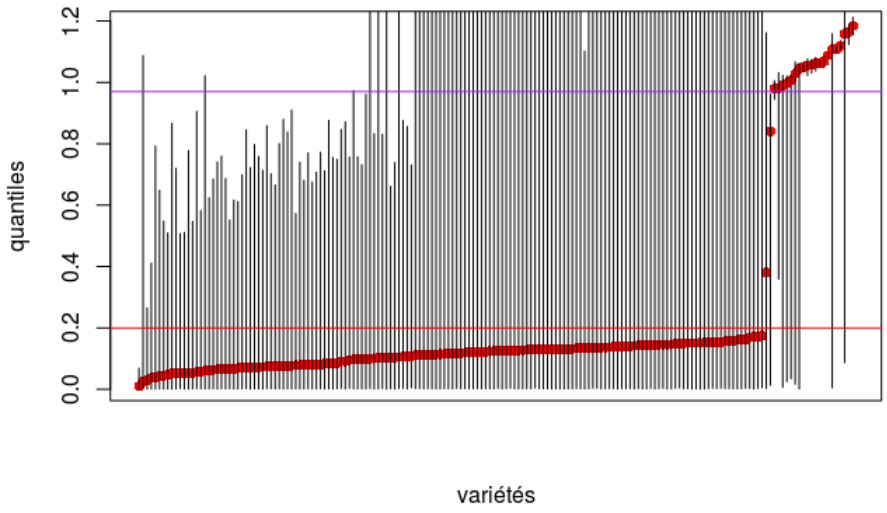



FIGURE 4.9 – Quantiles des lois *a posteriori* de $\phi_i$ (effet variété)

### 4.4.5 La structure spatiale

Afin d'examiner l'effet de l'introduction de l'effet spatiale dans le modèle, nous avons cartographié l'effet spatial dans le but de voir si il y'a une cohérance entre le modèle et les données.Nous avons représenté dans la carte de la figure 4.12 la médiane *a posteriori* des $S_i$ simulées par le modèle pour chaque parcelle. La taille du cercle est proportionnelle à la médiane de $S_i$.

De plus, nous avons calculé un degré d'incidence dans tous les départements de France pour les parcelles malades en se basant sur les notes d'incidence. En effet, nous avons sommé l'incidence des trois feuilles pour chaque parcelle, nous avons effectué ensuite une moyenne sur les parcelles du même département, ayant été observées dans la même semaine. Le but étant d'avoir une idée sur

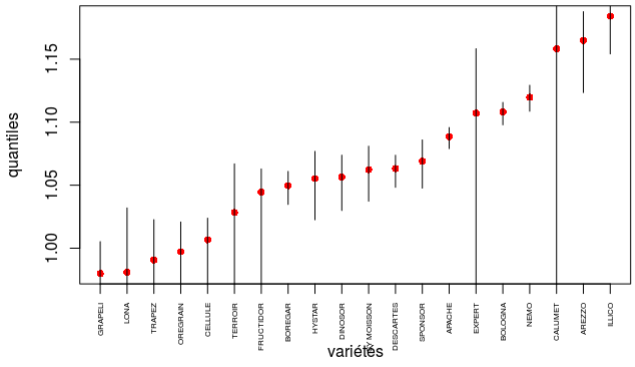



FIGURE 4.10 – Quantiles des lois *a posteriori* pour les variétés les plus malades

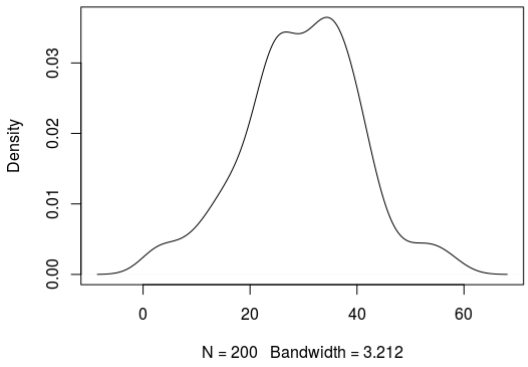



FIGURE 4.11 – Densité de la loi *a posteriori* de $\psi$

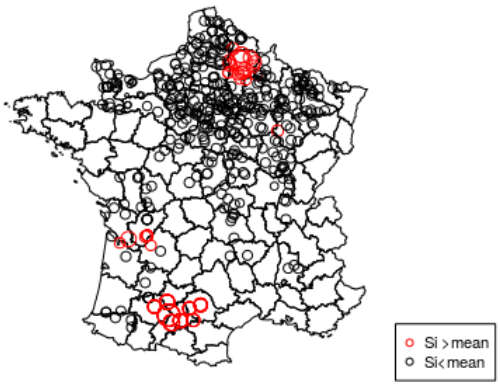



FIGURE 4.12 – Carte de l'effet spatiale $S_i$

l'odre de grandeur de l'incidence dans chaque département, nous avons pris le degré d'incidence

comme le maximum réalisé dans les semaines observées pour chaque département. On retrouve la cartographie du degré d'incidence dans la figure 4.13.

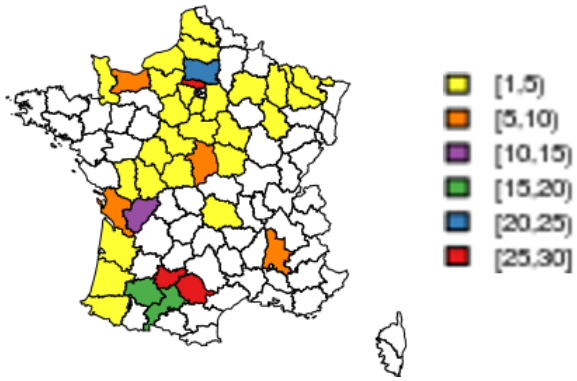



FIGURE 4.13 – Degré d'incidence dans les départements ayant des parcelles malades

On se sert des deux cartes ($S_i$ et degré d'incidence) pour juger de la cohérance entre le modèle spatial et les données. En effet, les parcelles qui se situent dans des départements où le degré d'incidence est important devraient avoir un effet spatial positif et plus important (un cercle de diamètre plus grand dans la figure 4.12) que les parcelles qui se trouvent dans des régions généralement moins atteintes de la rouille brune.

On remarque que l'effet spatial $S_i$ est plus important dans les parcelles situées aux départements 16,17, 31, 32, 81,82. Ce qui est cohérent avec la carte du degré d'incidence où on retrouve que ce sont les départements les plus atteints par la maladie en 2016. De plus, l'effet spatial n'est pas important dans les départements où le degré d'incidence est petit. En revanche, $S_i$ est sur-estimée au niveau des parcelles du département 02 où le degré d'incidence n'est pas important. En effet, il s'agit dans ce département de parcelles ayant de fortes variations de l'incidence.

En somme, la cartographie de l'effet spatial montre globalement une certaine cohérance entre le modèle et les données certainement en lien avec la dépendance de l'inoculum à d'autres paramètres, tels que les conditions météorologiques qui sont en relation avec la zone géographique. L'effet spatial permet donc d'avoir une vision plus large sur les paramètres participant à la propagation de la rouille.

### 4.4.6 Bilan du deuxième modèle

D'un point de vue **statistique**, certains paramètres n'ont pas convergé totalement. Il sera intéréssant de relancer le modèle sur un nombre d'itérations suffisamment grand pour assurer la convergence de tous les paramètres du modèle. De plus, l'amélioration de l'ajustement n'a pas été satisfaisante, il faut donc enrichir le modèle avec d'autres variables qui agissent sur l'évolution de l'épidémie.

# DISCUSSION ET PISTES D'AMÉLIOARTION

Les résultats obtenus à l'issue de ce travail de modélisation de la dynamique temporelle de la propagation de la rouille ont mené à la convergence de tous les paramètres du modèle, en plus de la significativé de la variable variété dans le modèle. Toutefois, l'effet de l'inoculum était faible comparé à celui de la variété et l'ajustement n'était pas satisfaisant en raison de de la sur-estimation de certaines incidences faibles et la sous-estimation de celles importantes. Etant donné que la répartition spatiale des parcelles n'est pas neutre, nous avons introduit une structure spatiale sur l'inoculum pour tenir compte de la variabilité dûe à l'espace. Le modèle qui en résulte n'a pas donné l'amélioration souhaitée au niveau de l'ajustement ( seulement 1.36% par rapport au premier modèle). Cependant la structure spatiale qui en découle est en adéquation avec les données d'incidence dont nous disposons.

De surcroît, certaines pistes sont envisageables pour améliorer le modèle. En effet, le cycle de germination des spores du champignon est favorisé par des conditions climatiques bien déterminées. Par conséquent, le taux de croissance de la maladie est lié aux conditions climatiques dans chaque département. Il serait intéréssant de spatialiser le taux de croissance pour tenir en compte l'effet des conditions climatiques. Une autre piste serait d'ajouter un effet de l'année, variable qui prendra en compte à la fois le risque climatique dans sa globalité dans l'année et le fait que les races de rouille évoluent chaque année et peuvent contourner la résistance des variétés. Par ailleurs, l'intensité de la rouille dans une parcelle est étroitement liée à celle des parcelles voisines, d'où l'intérêt de faire un modèle de dispersion qui tient compte du voisinage. Finalement, nous avons supposé dans les deux premiers modèles une intensité maximale potentielle constante pour toutes les parcelles, cependant, cette intensité peut dépendre de la surface agricole utile en blé de chaque département , variable fournie par l'INSEE et qui serait probablement influente dans le modèle.

ANNEXE

# A

# ORGANIGRAMME DE L'INRA

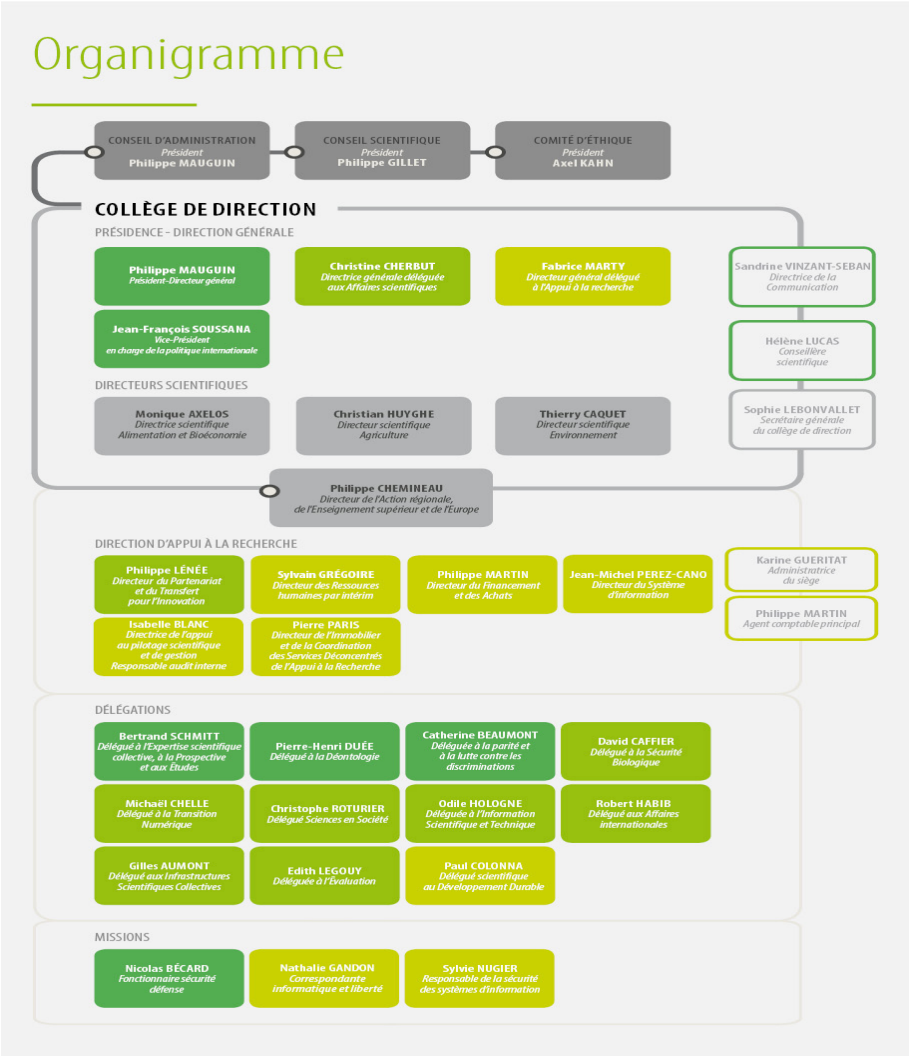



FIGURE A.1 – Organigramme de l'INRA

ANNEXE

# B

# ORGANIGRAMME DE BIOSP

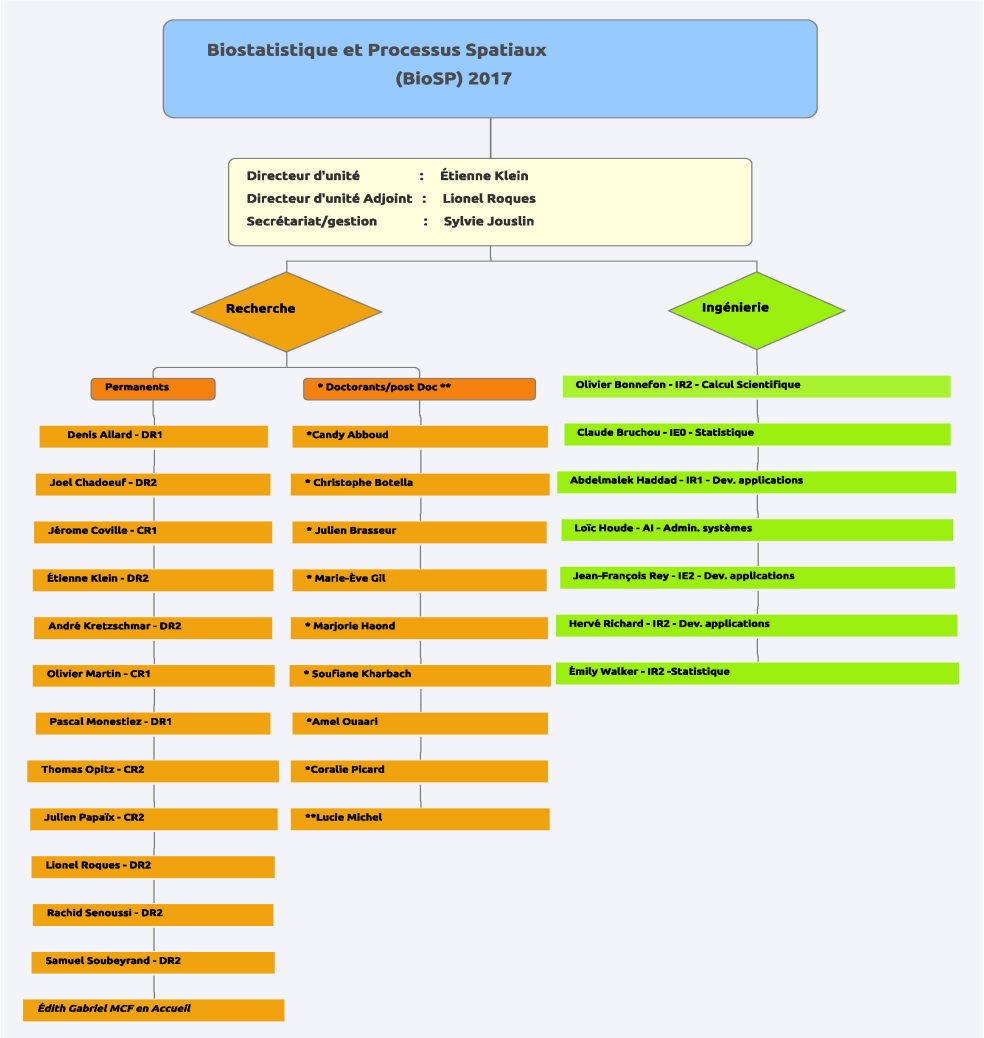



FIGURE B.1 – Organigramme de BioSP

ANNEXE

# C

# CODE JAGS DU MODÈLE TEMPOREL

```
#a priori
mupsi ~ dnorm(0,1)
alpha0 ~ dunif(0,1000)
sig1 ~ dunif(0,10)
sigv~ dunif(0,10)
tau1 <- 1/(sig1*sig1)
tauv<-1/(sigv*sigv)
K <-10000
for (j in 1:3){
#Etages foliaires
lambda[j] ~ dunif(0,1)
}
for(v in 1:Nv){
#variété
lpsi[v]~dnorm(mupsi,tauv)
log(psi[v])<-lpsi[v]
}
#variables latentes
for (i in 1:Ni){
lW[i] ~ dnorm(0,tau1)
log(W[i,1]) <- lW[i]
for (t in 2:Nt){
# Modèle
W[i,t] <- step(W[i,t-1]-K)*K+(1-step(W[i,t-1]-K))*(alpha0+ psi[variete[i]]*W[i,t-1]*(1-(W[i,t-1]/K)) )#+ eps[i,t]
}
}
#vraisemblance
for (j in 1:Nj){
for (i in 1:Ni){
for (t in 1:Nt){
p[j,i,t] <- 1-exp(-lambda[j]*W[i,t])
Y[j,i,t] ~ dbin(p[j,i,t],20)
Ypred[j,i,t] ~ dbin(p[j,i,t],20)
}
}
}
```

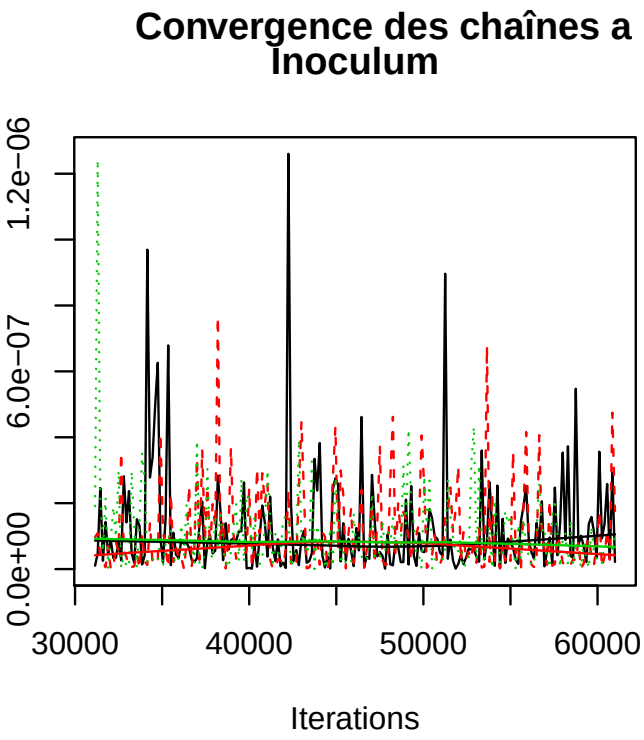
FIGURE C.1 – Code Jags du modèle temporel

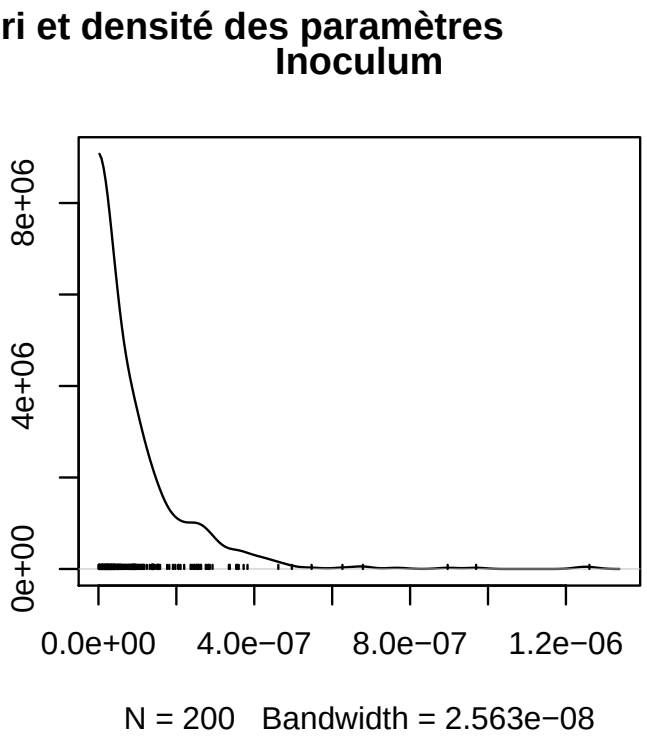
ANNEXE

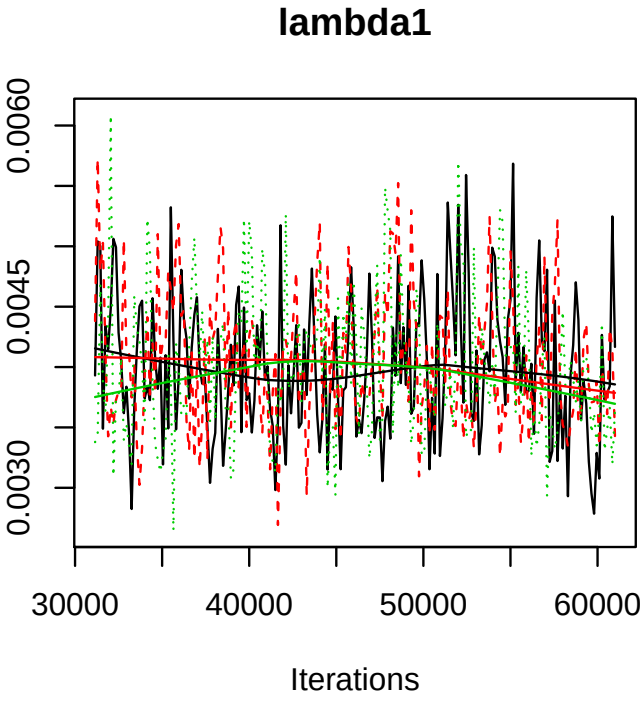
# D

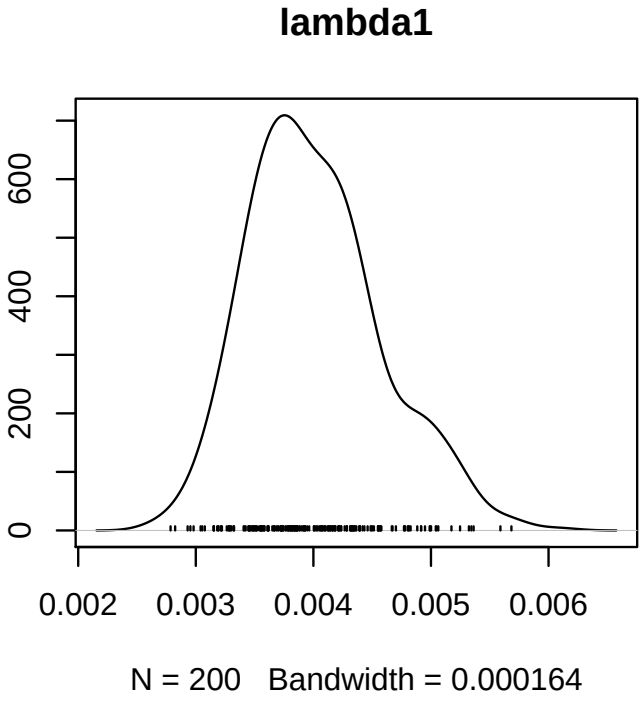
# CONVERGENCE DES PARAMÈTRES DU PREMIER MODÈLE

**Convergence des chaînes a postériori et densité des paramètres**

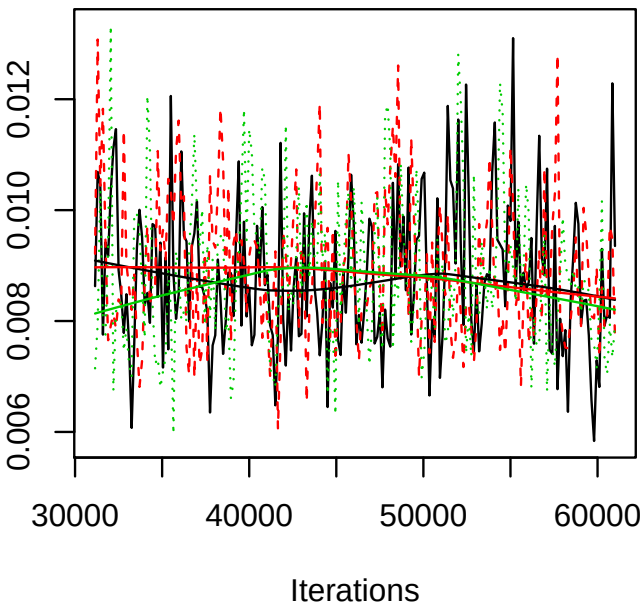


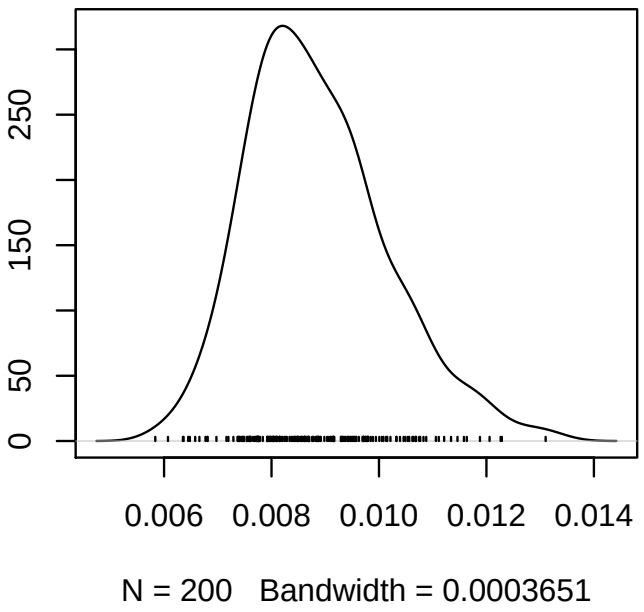



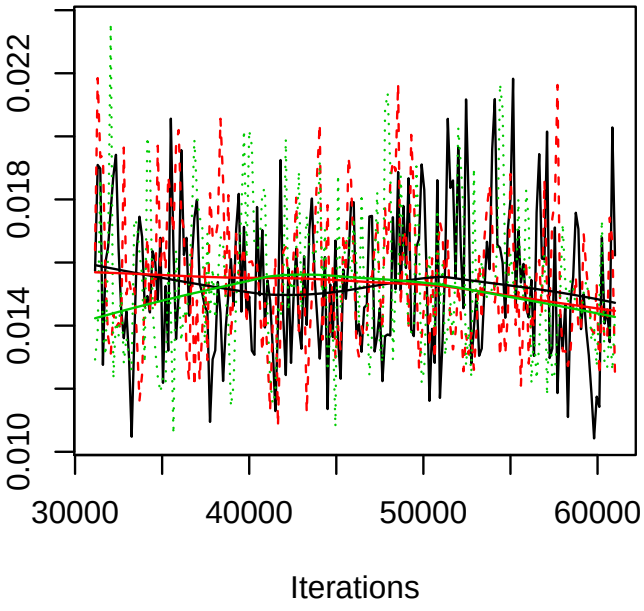



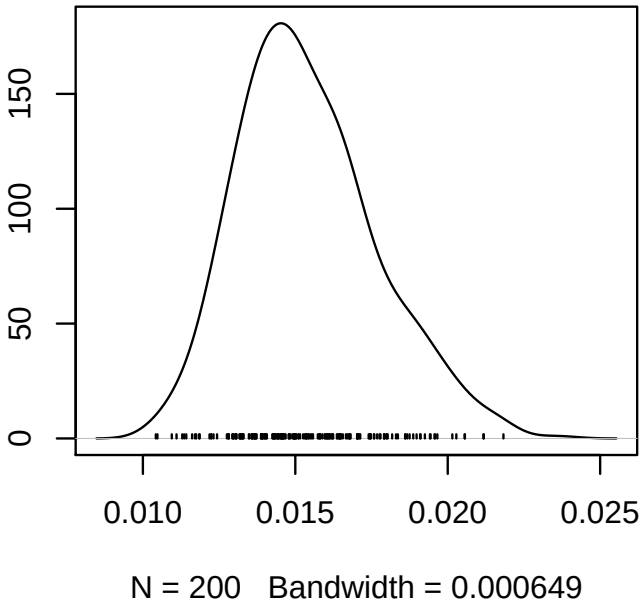



**Convergence des chaînes a postériori et densité des paramètres**

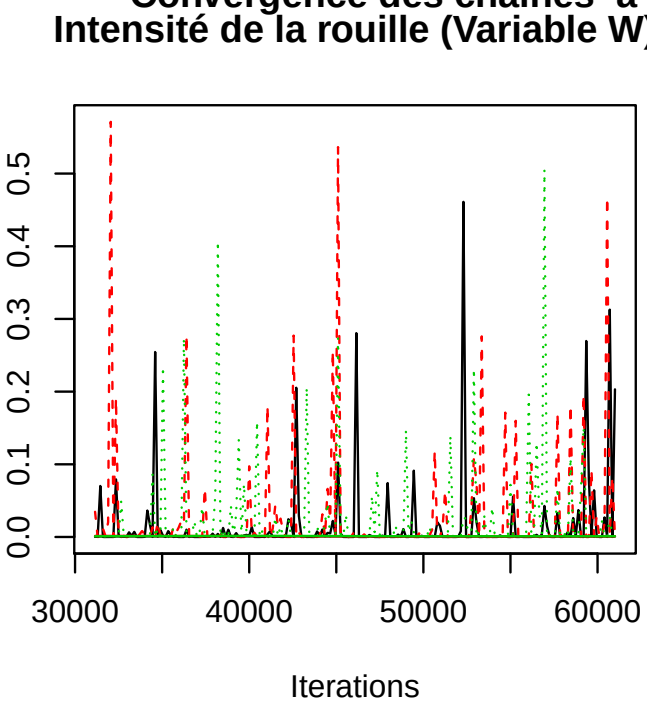


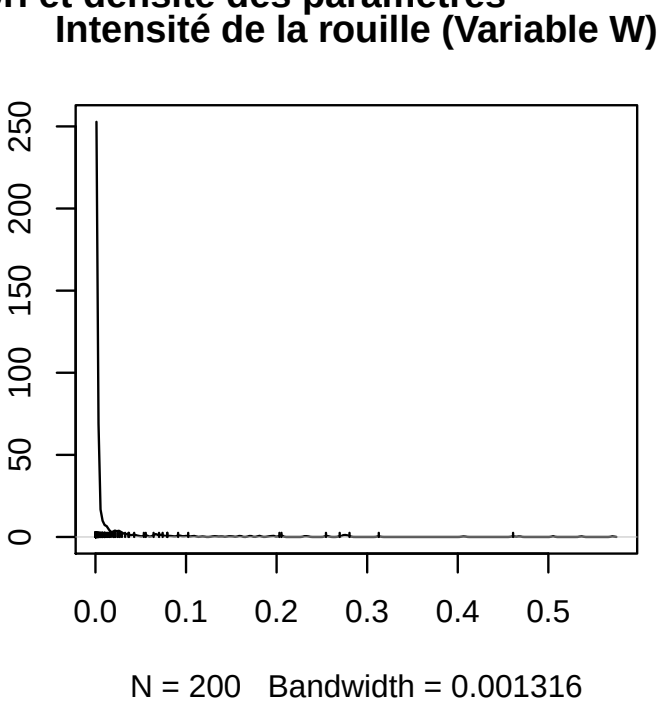



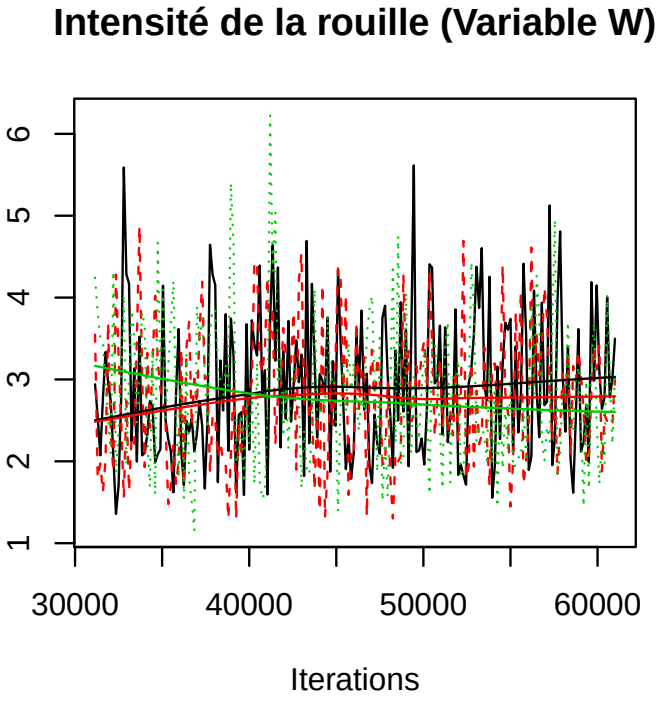



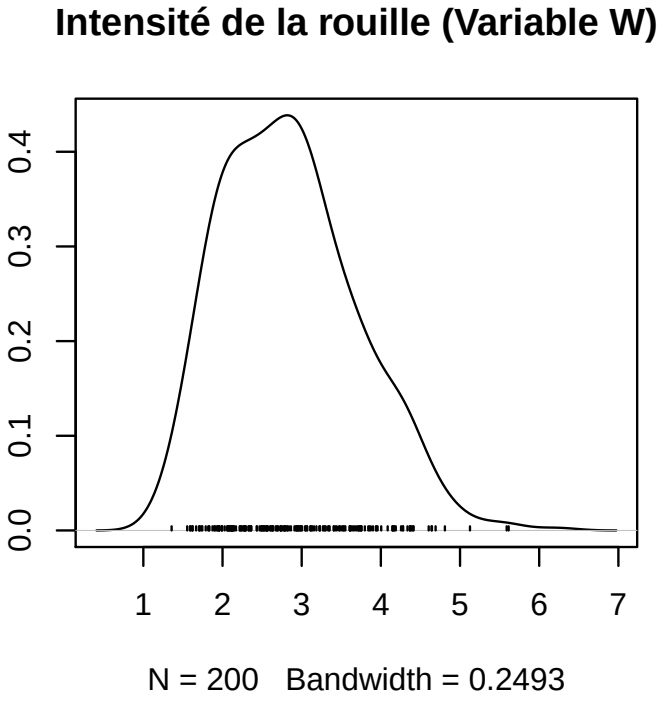



**Convergence des chaînes a postériori et densité des paramètres**



ANNEXE

# E

# CODE JAGS DU MODÈLE SPATIAL

```
#a priori
mupsi ~ dnorm(0,1)
alpha0 ~ dnorm(0,0.001)
lambdaP~dunif(0.0001,1)
 sig1 ~ dunif(0,10)
sigv~ dunif(0,10)
sigs ~ dunif(0,10)
taus <-0 #~ dunif(0,10)
tau1 <- 1/(sig1*sig1)
tauv<-1/(sigv*sigv)
K <-10000
for (j in 1:3){
lambda[j] ~ dunif(0,1)
}
for(v in 1:Nv){
lpsi[v]~dnorm(mupsi,tauv)
log(psi[v])<-lpsi[v]

}
#process predictif
C.PP.inv <- pow(sigs,-2)*inverse(C.PP)
for(i in 1:N.PP){
mu.S.PP[i] <- 0
for(j in 1:N.PP){
C.PP[i,j] <- exp(-lambdaP*dist_PP[i,j]/1000)
}
}
#variables latentes
for (i in 1:Ni){
lW[i] ~ dnorm(0,tau1)
log(W[i,1]) <- lW[i]
#interpolation aux observation
for(j in 1:N.PP){
C.obs.PP[i,j] <- sigs*sigs*exp(-lambdaP*dist_obsPP[i,j]/1000)
}
}
#définition des erreurs w.PP et interpolation aux w
S.PP[1:N.PP] ~ dmnorm(mu.S.PP, C.PP.inv)
S[1:Ni] <- C.obs.PP%*%C.PP.inv%*%S.PP[1:N.PP]
```

FIGURE E.1 – Code Jags du modèle spatial

```
for(i in 1:Ni){
#spatial
correction[i] <- t(C.obs.PP[i,])%*%C.PP.inv%*%C.obs.PP[i,]
.tot[i] <- pow(sigs,2) + pow(taus,2) - correction[i]
prec[i] <- 1/.tot[i]
Sprim[i] ~ dnorm(0, prec[i])
log(alpha[i]) <- alpha0 + S[i] + Sprim[i]
for (t in 2:Nt){
# Modèle
W[i,t] <- step(W[i,t-1]-K)*K+(1-step(W[i,t-1]-K))*(alpha[i]+ psi[iete[i]]*W[i,t-1]*(1-(W[i,t-1]/K)) )#+ eps[i,t]
}
}


#vraisemblance

for (j in 1:Nj){
for (i in 1:Ni){
for (t in 1:Nt){
p[j,i,t] <- 1-exp(-lambda[j]*W[i,t])
Y[j,i,t] ~ dbin(p[j,i,t],20)
Ypred[j,i,t] ~ dbin(p[j,i,t],20)
}
}
}
```

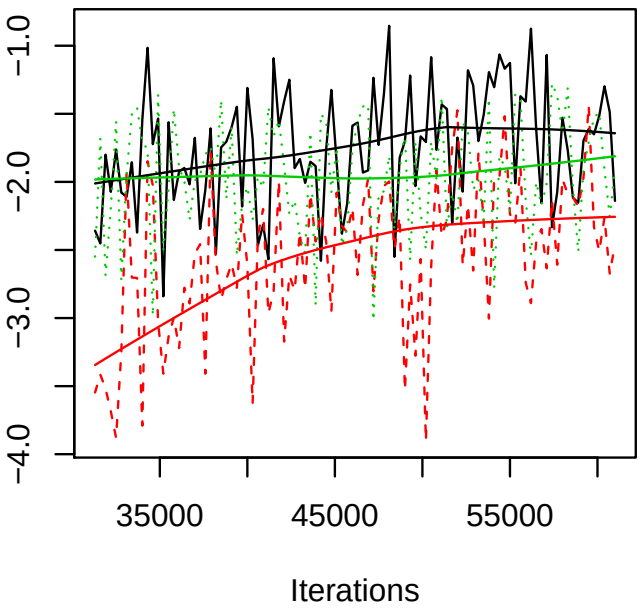
FIGURE E.2 – Code Jags du modèle spatial

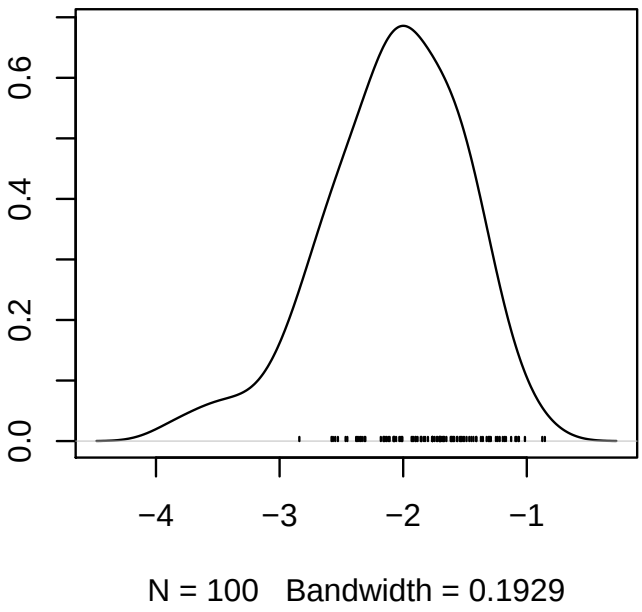
ANNEXE

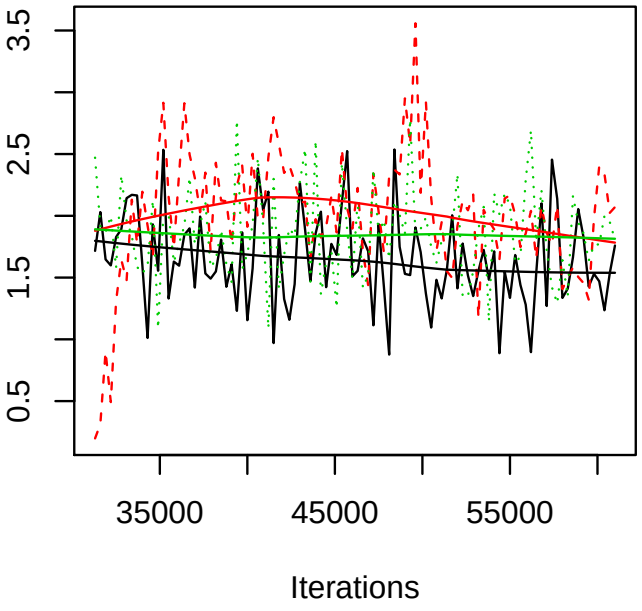
# F

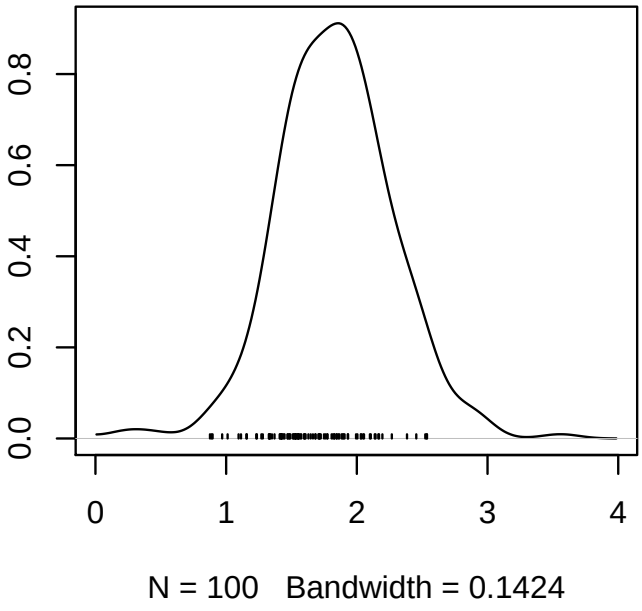
# CONVERGENCE DES PARAMÈTRES DU MODÈLE SPATIAL

**Convergence des chaînes a postériori et densité des paramètres**

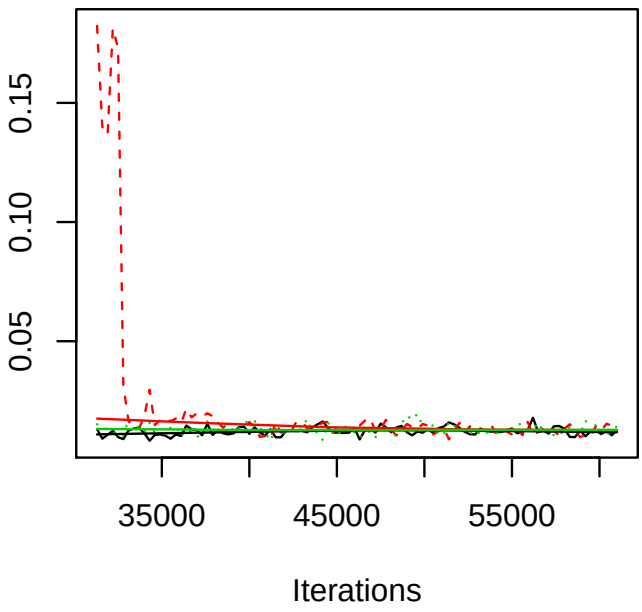


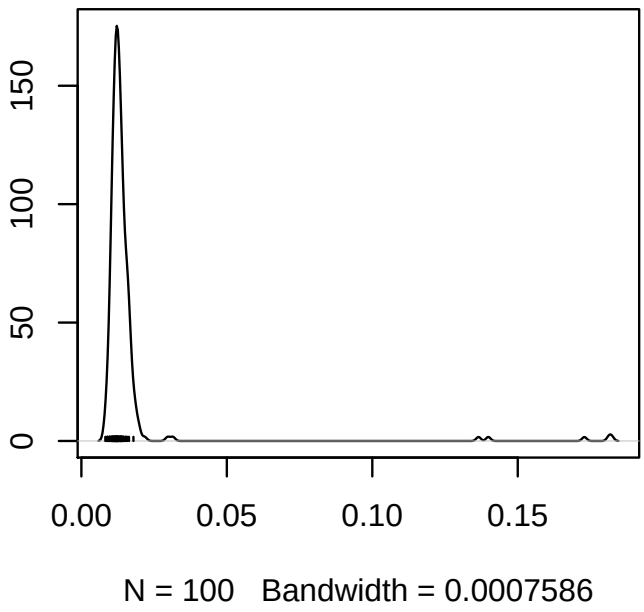



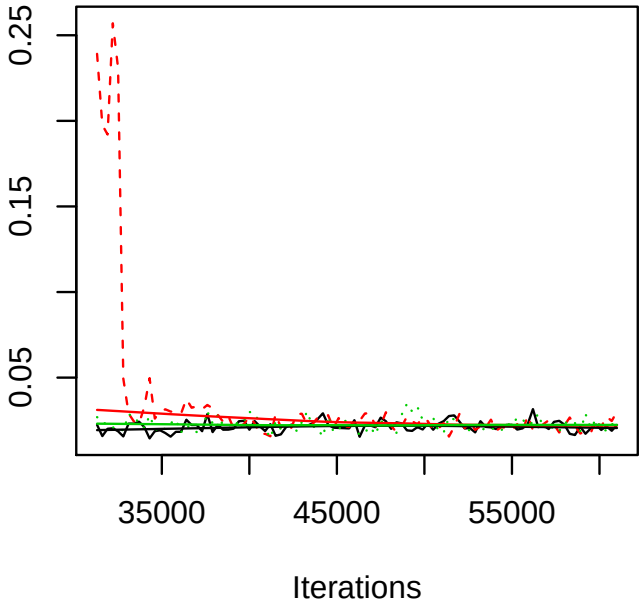



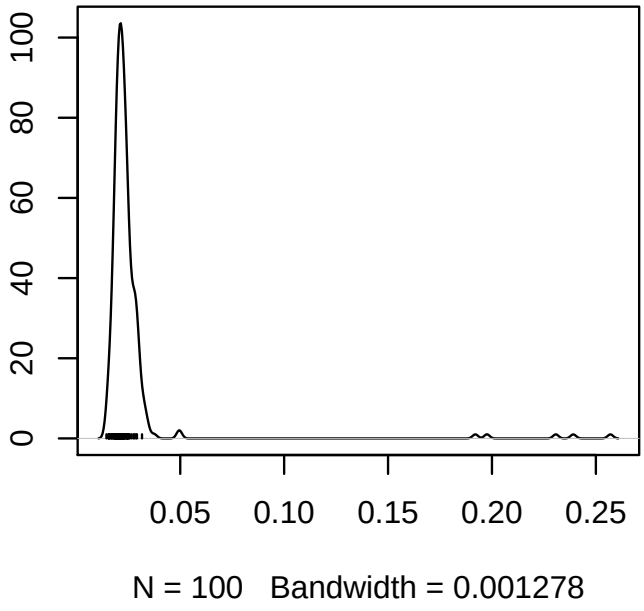



**Convergence des chaînes a postériori et densité des paramètres**

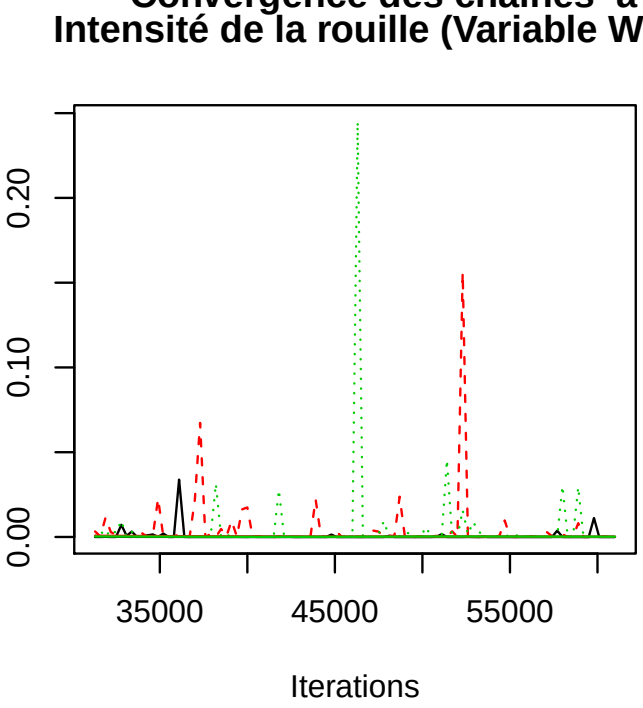


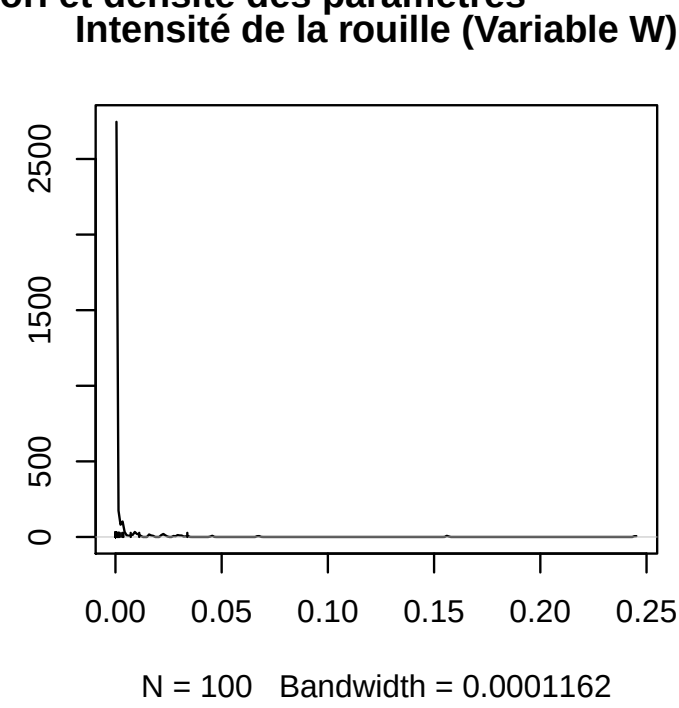



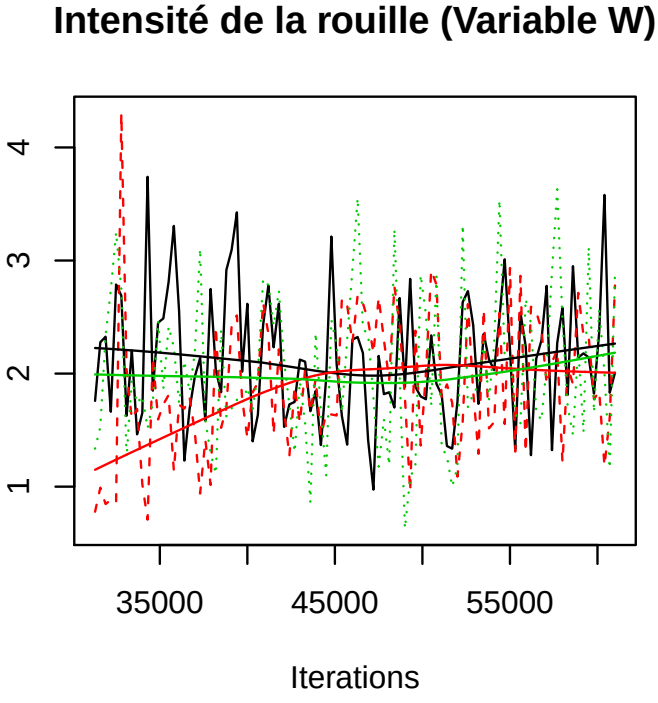



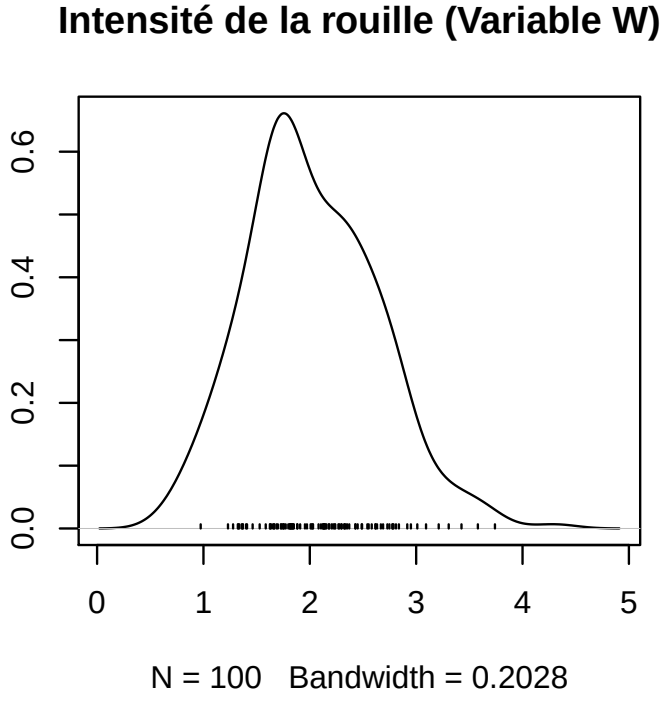



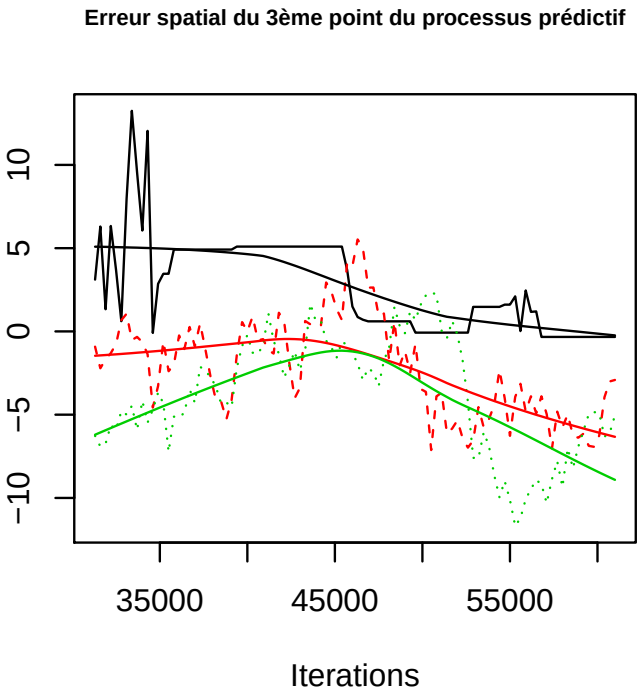
**Convergence des chaînes a postériori et densité des paramètres**

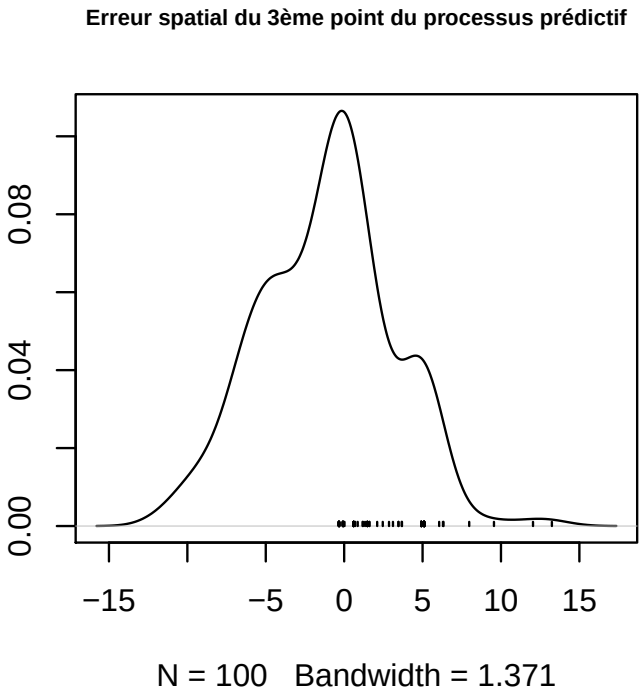



Erreur spatial du 3ème point du processus prédictif
10
5
0
−5
−10
35000
45000
55000
Iterations


Erreur spatial du 3ème point du processus prédictif
0.08
0.04
0.00
−15
−5
0
5
10
15
N = 100 Bandwidth = 1.371