Méthodes Bayesiennes pour modèles de systèmes dynamiques

Arnaud Bensadoun, David Gouache, François Piraux, David Makowski, Daniel Wallach

INRA, Arvalis-Institut du végétal

RMT modélisation (INRA-ACTA-ITA)

Plan

- La problématique agronomique, modèle de l’espérance
- Approche, résultats Bayesiens

Contexte agronomique

La septoriose

- Maladie fongique la plus dommageable sur blé tendre en France
  - ✓ très fréquente dans le Nord
  - ✓ très dépendante du climat
- Formation de lésions entrainant la nécrose des feuilles
- Pertes de rendements jusqu'à 60 q/ha
- Très peu de lutte alternative aux fongicides

Dynamique de la maladie

- Démarrage
  - Spores survivent l'hiver
  - Infecte les feuilles
- Croissance
  - En conditions favorables, un spore donne une infection
  - L'infection grossit, produit des spores
- Transfert
  - Les spores infectent la même feuille ou d'autres feuilles

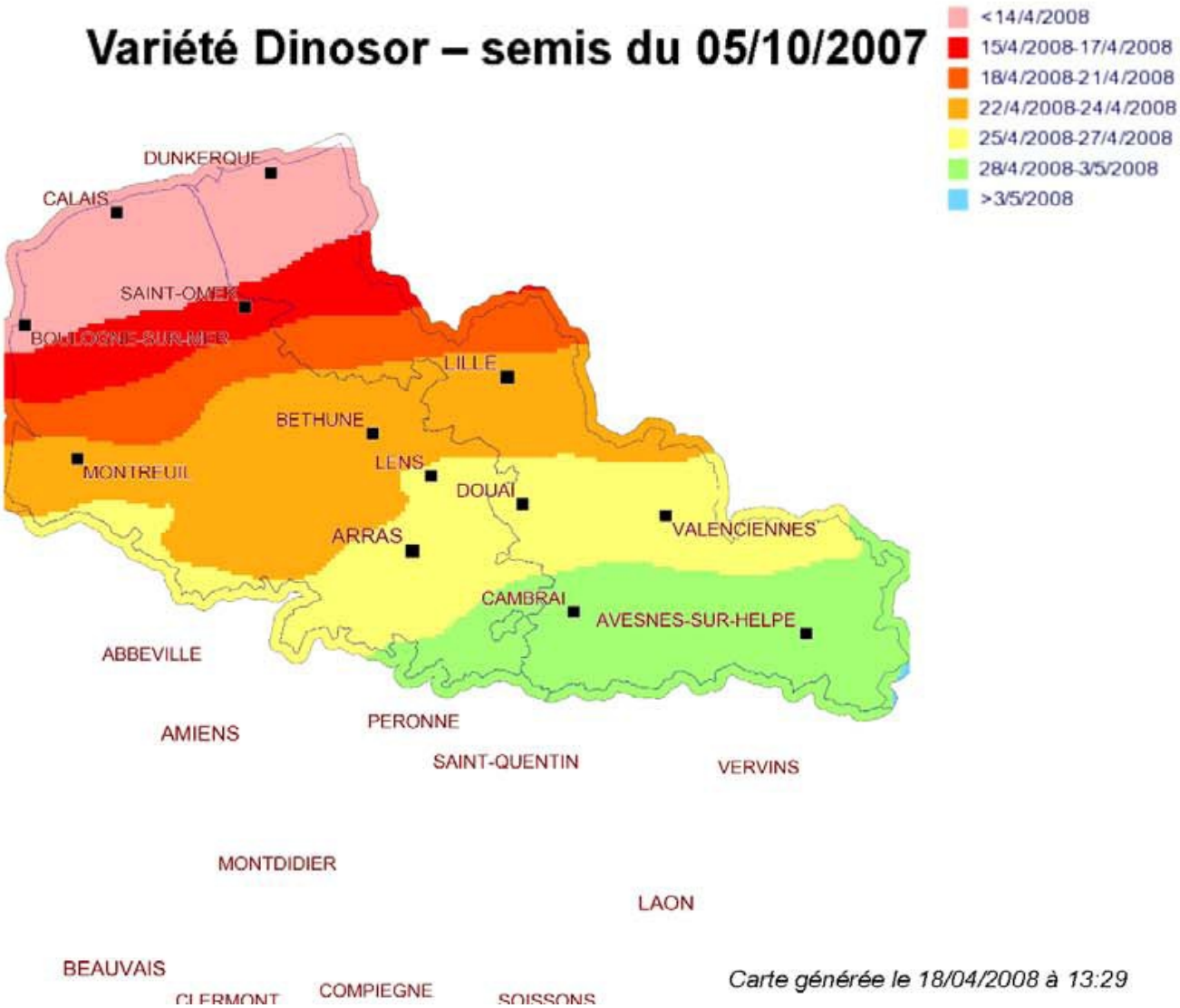
ersion

- Niveau d’attaque différent suivant lieu, année
- Important de ne pas commencer traitement avant que ce soit nécessaire
- Difficile de baser décision sur observations
- D’où utilité d’un modèle

SeptoLIS ®

- Modèle développé à Arvalis
  - Basé sur modèle littérature
- Conseillers et agriculteurs peuvent s'y abonner
  - Pour avoir résultats chaque jour sur internet

La septoriose et *Arvalis*



Le modèle de l'espérance SeptoLIS®

- $Y = f(X;\theta) + \varepsilon$
- Y = % nécrosé
- X = température, pluie chaque jour

Les équations, chaque jour

$$Y(t,l) = \sum_{i=1}^{t} I(i,l)\, y(T(t) - T(i))$$

$$U(t,l) = 1 - e^{-aY(t,l)}$$

$$F(t,l) = U_g e^{-kh_g(t)} + \sum_{j=l+1}^{M} U(t,j) e^{-kh_j(t)} + \rho U(t,l)$$

$$I(t,l) = L_V(t,l) L_C(t,l) R_{rain} F(t,l)$$

- Y=% surface atteinte
- U=nombre spores produits
- F=nombre de spores qui arrivent
- I= nombre d’infections réussies

C’est un modèle de système dynamique

- Ou modèle de processus
- Ou modèle mécaniste

Spécificité modèles de système dynamique (1)

- Complexe
  - On ne reprogramme pas le modèle
- Long (Chaque cas est une saison de croissance)
  - Temps de calcul limitant

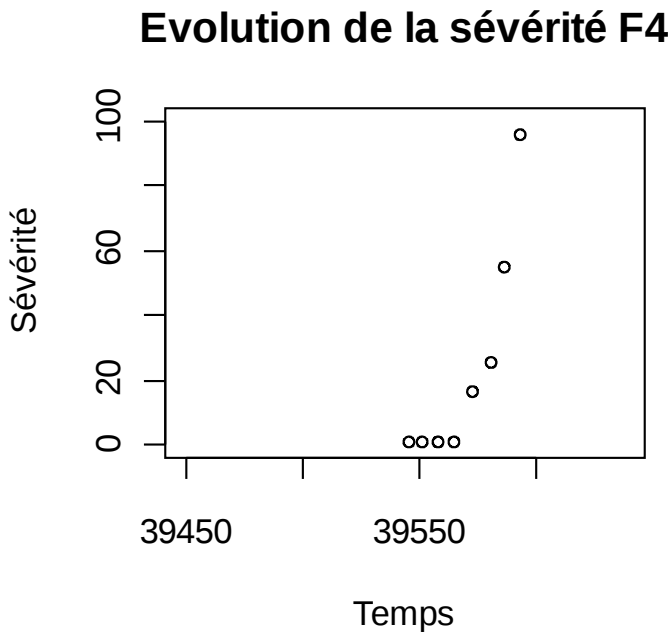
Paramètres

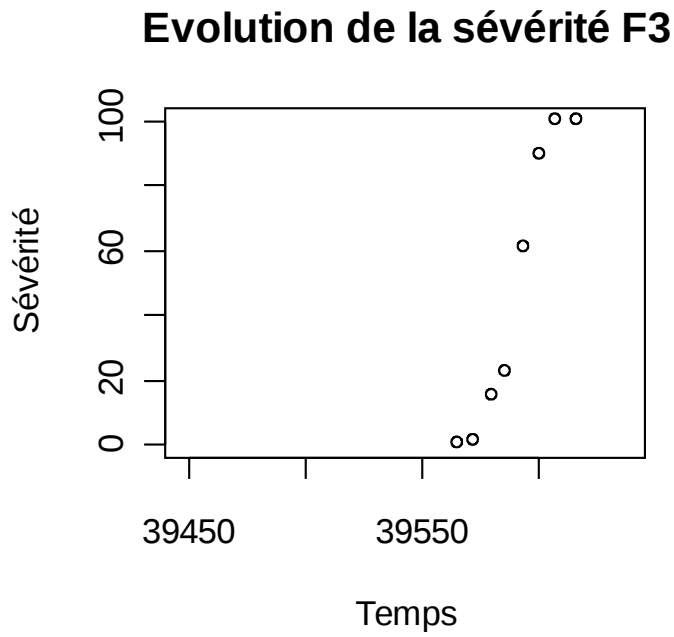
Parameters	*Values*		*Sources*
$inoc_{threshold}$	0.698	multiplication of ground inoculum	
$transf_{rainfall}$	-0.28	Spore transfert	from Audsley et al. (2005)
$transf_{height}$	0.16	Spore transfert	from Audsley et al. (2005)
vit_{nec}	0.015	Lesion expansion rate	adapted from Robert et al. (2008)
$appar_{nec}$	0.1	Initial lesion size	adapted from Robert et al. (2008)
$Tmin_{t}$	0.043	Temperature threshold at *t* for infection	
$Tmin_{t-1}$	6.756	Temperature threshold at *t-1* for infection	
$Rain_{threshold}$	0.119	Threshold rainfall for infection	
vit_{expan}	1.6	Leaf growth rate	from ARVALIS trials
limit	9	Rainfall threshold for maximal infection efficiency	from Audsley et al 2005
min_{lorin}	3.510	Minimal limit for ground inoculum at the end of the cycle	
$Tlim_{lorin}$	8.020	Temperature threshold for ground inoculum accumulation	
$inoc_{depletion}$	7.685	Depletion rate of ground inoculum	
$limit_{L1}$	9.351	Rainfall threshold for infection of extanding leaf 1	
Scale		Empirique	
$mult_{coeff}$		Empirique	
$Inoc_{prod}$		Empirique	

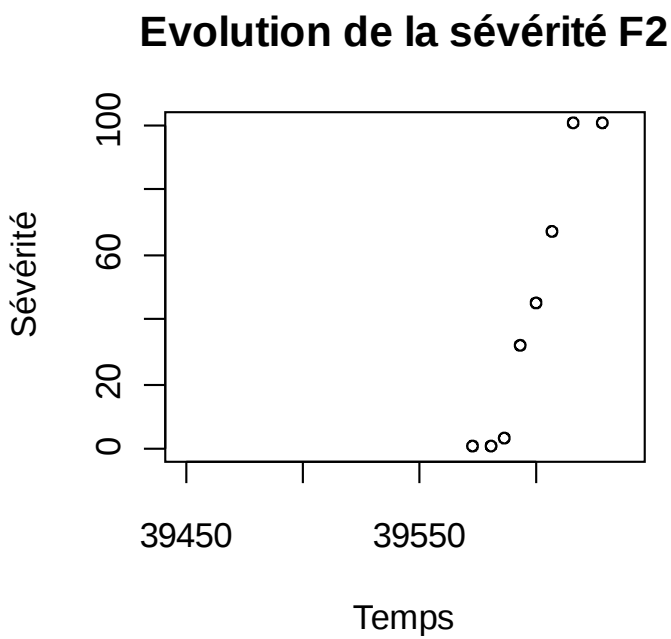
Spécificité modèles de système dynamique (2)

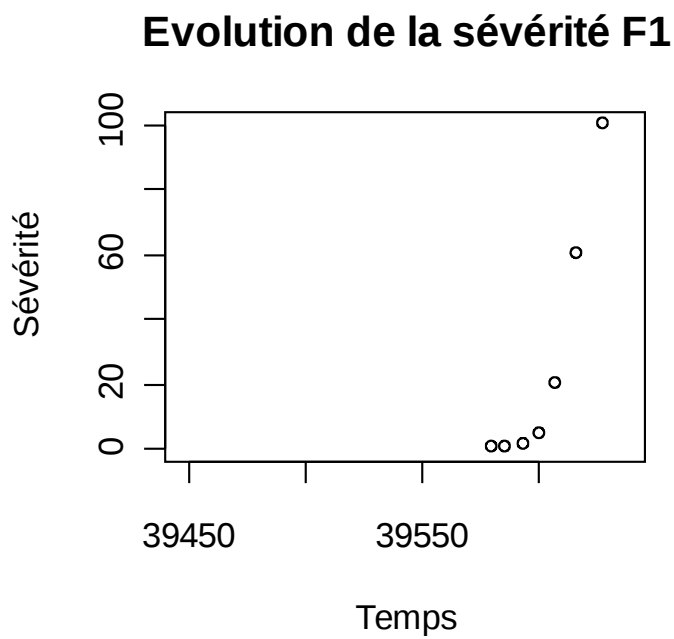
- Beaucoup de paramètres
  - On ne peut pas estimer tous par maximum de vraisemblance
- Une vraie information priori sur les paramètres
  - Mais difficile à quantifier exactement

Données


Evolution de la sévérité F4
Sévérité
0
20
60
100
39450
39550
Temps


Evolution de la sévérité F3
Sévérité
0
20
60
100
39450
39550
Temps


Evolution de la sévérité F2
Sévérité
0
20
60
100
39450
39550
Temps


Evolution de la sévérité F1
Sévérité
0
20
60
100
39450
39550
Temps

- Pour l’estimation
  - 157 sites-années, 2001 données
  - (Pour étude Bayesienne on utilise 8)
- Données d’évaluation
  - 2008. 18 sites, 270 données

Spécificité modèles de système dynamique (3)

- Le modèle pour l'erreur est complexe
  - $Y = f(X;\theta) + \varepsilon$
  - Pas d'indépendance des erreurs (on peut avoir structure de corrélations complexe).
  - Donc, modéliser covariances
  - Sinon, au moins vérifier modèle statistique de l'erreur

Etude fréquentiste

- Approche
  - Estimer les 3 paramètres moindres carrés ordinaires
  - Fixer 14 paramètres à leur moyenne a priori
- Les désavantages
  - Information a priori dégradée (paramètre inconnu ou parfaitement connu)
  - Incertitude irréaliste pour les paramètres
    - Tout est mis sur 3 paramètres

Objectifs de cette étude

- Estimer distribution a posteriori des paramètres par approche Bayesienne
  - Mettre au point la méthode
  - Appliquer pour prédiction, pour évaluer les incertitudes
  - Evaluer intérêt approche Bayesienne

Approche, résultats Bayesiens

Metropolis-Hastings within Gibbs

- Recherche de distributions *a posteriori*
- Echantillonnage de distributions conditionnelles par étape

  - Etape 1 $P(\theta|Y,\sigma_\varepsilon^2)$

  - Etape 2 $P(\sigma_\varepsilon^2|Y,\theta)$

Etape 1 Metropolis-Hastings

$$P\left(\theta|Y,\sigma_\varepsilon^2\right) \propto P\left(Y|\theta,\sigma_\varepsilon^2\right) \times \pi(\theta)$$

- Vraisemblance $P\left(Y|\theta,\sigma_\varepsilon^2\right)$

  - Hypothèses sur les résidus:
    - ✓ Indépendance et normalité
    - ✓ Homogénéité de la variance: transformation arcsin($Y^{1/2}$)

$$P(Y|\theta,\sigma_\varepsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2/2\sigma_\varepsilon^2}$$

Etape 1 Metropolis-Hastings

Loi a priori $\pi(\theta)$

- Hypothèses sur les paramètres:
  - ✓Indépendance
  - ✓Normalité pour paramètres mesurées dans la littérature
  - ✓Uniforme pour paramètres empiriques

Etape 1 Metropolis-Hastings

$$\theta = (\theta_A, \theta_B)$$

$$\theta_A \sim U\left(\theta_{A_{min\,j}}, \theta_{A_{max\,j}}\right) \quad j=\{1,2,3\} \qquad \pi(\theta_A) = \prod_{j=1}^{3} \frac{1}{\left(\theta_{A_{max\,j}} - \theta_{A_{min\,j}}\right)}$$

$$\theta_B \sim N\left(\mu_{\theta_{B_i}}, \sigma^2_{\theta_{B_i}}\right) \quad i=\{1,2,\ldots,14\} \qquad \pi(\theta_B) = \prod_{i=1}^{14} \frac{1}{\sqrt{2\pi\sigma^2_{\theta_{B_i}}}} \times e^{-\frac{\left(\theta_{B_i} - \mu_{\theta_{B_i}}\right)^2}{2\sigma^2_{\theta_{B_i}}}}$$

$$\pi(\theta) = \pi(\theta_A) \times \pi(\theta_B) = \prod_{j=1}^{3} \frac{1}{\left(\theta_{A_{max\,j}} - \theta_{A_{min\,j}}\right)} \times \prod_{i=1}^{14} \frac{1}{\sqrt{2\pi\sigma^2_{\theta_{B_i}}}} \times e^{-\frac{\left(\theta_{B_i} - \mu_{\theta_{B_i}}\right)^2}{2\sigma^2_{\theta_{B_i}}}}$$

Etape 1 Metropolis-Hastings

Loi de proposition : $g\left(\theta^{*(n+1)}, \theta^{(n)}\right) = N\left(\theta^{(n)}, \Sigma\right)$

1er Run : *Σ*= tune x *Diag*(variance *a priori*)

2nd Run : *Σ*= tune x Variance *a posteriori(1er Run)*

Etape 1 Metropolis-Hastings

- On tire une valeur $\theta^{*(n+1)}$ dans $N(\theta^{(n)}, \Sigma)$ et on regarde le rapport

$$\alpha = \frac{P(Y|\theta^{*(n+1)}, \sigma_\varepsilon^{2(n)}) \times \pi(\theta^{*(n+1)})}{P(Y|\theta^{(n)}, \sigma_\varepsilon^{2(n)}) \times \pi(\theta^{(n)})}$$

- $Si\ \alpha \geq 1\ \ \theta^{(n+1)} = \theta^{*(n+1)}$

- $Sinon$ $\quad \theta^{(n+1)} = \theta^{*(n+1)}$ avec probabilité α

  $\theta^{(n+1)} = \theta^{(n)}$ avec probabilité $(1 - \alpha)$

Etape 2 Gibbs

$$P\left(\sigma_\varepsilon^2 \middle| Y, \theta\right) \propto \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2/2\sigma_\varepsilon^2} \times \frac{1}{\sigma_\varepsilon^2}$$

On pose $\tau = \frac{1}{\sigma_\varepsilon^2}$ pour obtenir

$$P(\tau|Y,\theta) \sim \Gamma\left(shape = \frac{N}{2} + 2, \qquad scale = 2/\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2\right)$$

➔Tirage direct dans la distribution *a posteriori*

Convergence

- Testée sur deux chaînes:

➢ Critère statistique: test de *Gelman et Rubin* (Package R *coda*)

➔ Comparaison variance intra/inter chaînes

➢ Critère graphique: Trace de la moyenne

Problème de temps de calcul

- 8 sites-années (73 mesures): 50 000 itérations ➔ 5 jours!

  - Parallélisation des calculs sur 4 cœurs d'un processeur multi-cœurs (Package R *SNOW*)

Problème de temps de calcul

- 8 sites-années (73 mesures): 50 000 itérations ➔ 5 jours!

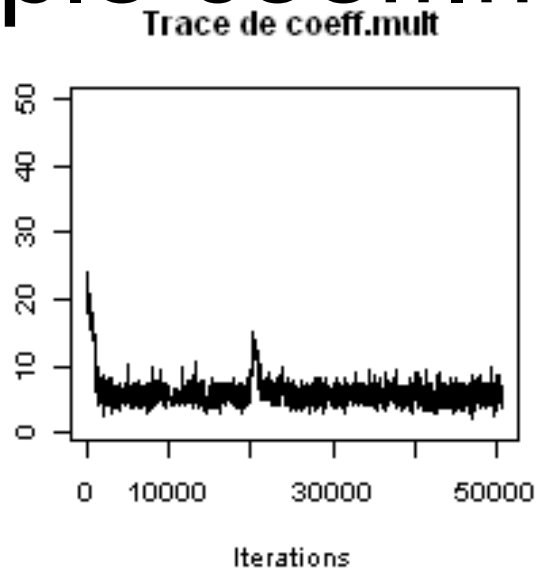
  - Parallélisation des calculs sur 4 cœurs d'un processeur multi-cœurs (Package R *SNOW*)

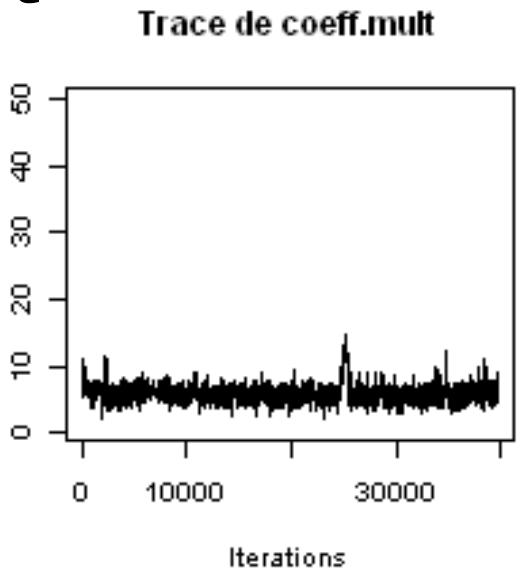
➔ 157 sites-années (2001 mesures) difficilement envisageable

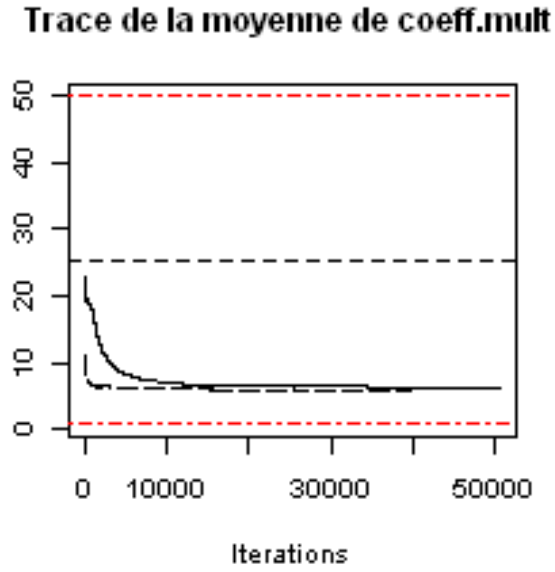
➔ Difficile de tester des idées (loi de proposition, point de départ, loi a priori…)

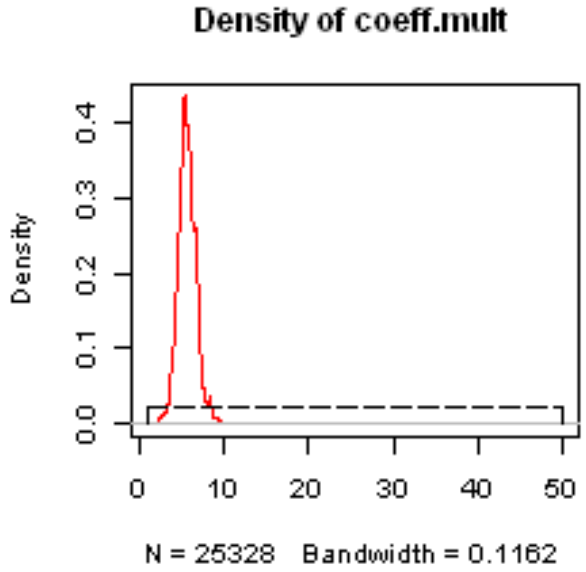
Distribution *a posteriori*

Exemple coeff.mult

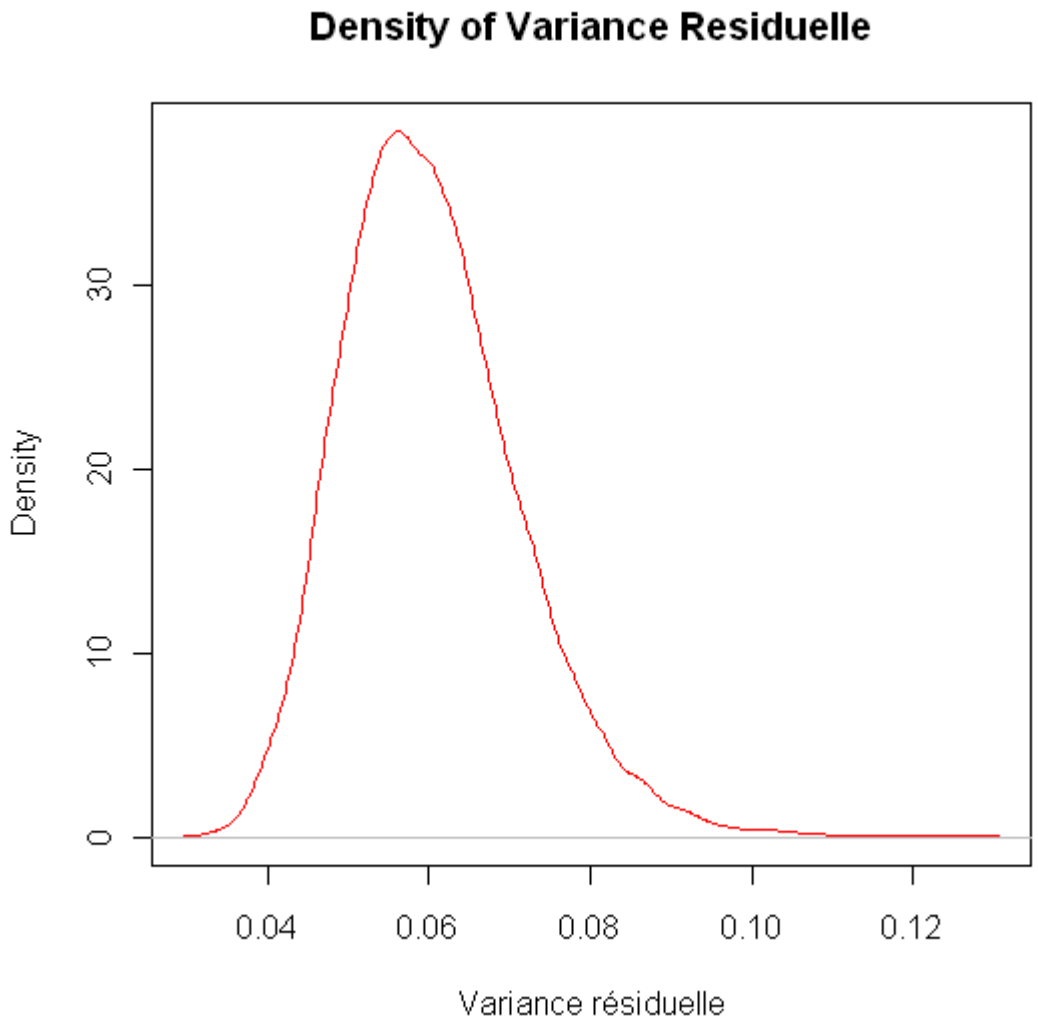








Variance résiduelle



PRIOR vs POSTERIOR

	Mean		Variance	
Parameter	PRIOR	POSTERIOR	PRIOR	POSTERIOR
coeff.mult	25.5	5.7373	200.083	1.0549
echelle	50.5	10.6837	816.75	42.638
transf.pluie	-0.28	-0.3344	0.0049	0.0031
transf.haut	0.160	0.1480	0.0016	0.0013
vit.nec	0.015	0.0187	1.00E-05	9.61E-06
prod.inoc	50.5	42.0501	816.75	795.091
vit.expan	1.60	0.8249	0.16	0.0673
Tlim.lorin	8.020	6.0706	4.02032	2.4955

Multivariate psrf= 1.07

PRIOR vs POSTERIOR

	Mean		Variance	
Parameter	PRIOR	POSTERIOR	PRIOR	POSTERIOR
coeff.mult	25.5	5.7373	200.083	1.0549
echelle	50.5	10.6837	816.75	42.638
transf.pluie	-0.28	-0.3344	0.0049	0.0031
transf.haut	0.160	0.1480	0.0016	0.0013
vit.nec	0.015	0.0187	1.00E-05	9.61E-06
prod.inoc	50.5	42.0501	816.75	795.091
vit.expan	1.60	0.8249	0.16	0.0673
Tlim.lorin	8.020	6.0706	4.02032	2.4955

Multivariate psrf= 1.07

PRIOR vs POSTERIOR

	Mean		Variance	
Parameter	PRIOR	POSTERIOR	PRIOR	POSTERIOR
coeff.mult	25.5	5.7373	200.083	1.0549
echelle	50.5	10.6837	816.75	42.638
transf.pluie	-0.28	-0.3344	0.0049	0.0031
transf.haut	0.160	0.1480	0.0016	0.0013
vit.nec	0.015	0.0187	1.00E-05	9.61E-06
prod.inoc	50.5	42.0501	816.75	795.091
vit.expan	1.60	0.8249	0.16	0.0673
Tlim.lorin	8.020	6.0706	4.02032	2.4955

Multivariate psrf= 1.07

PRIOR vs POSTERIOR

	Mean		Variance	
Parameter	PRIOR	POSTERIOR	PRIOR	POSTERIOR
coeff.mult	25.5	5.7373	200.083	1.0549
echelle	50.5	10.6837	816.75	42.638
transf.pluie	-0.28	-0.3344	0.0049	0.0031
transf.haut	0.160	0.1480	0.0016	0.0013
vit.nec	0.015	0.0187	1.00E-05	9.61E-06
prod.inoc	50.5	42.0501	816.75	795.091
vit.expan	1.60	0.8249	0.16	0.0673
Tlim.lorin	8.020	6.0706	4.02032	2.4955

Multivariate psrf= 1.07

PRIOR vs POSTERIOR

	Mean		Variance	
Parameter	PRIOR	POSTERIOR	PRIOR	POSTERIOR
coeff.mult	25.5	5.7373	200.083	1.0549
echelle	50.5	10.6837	816.75	42.638
transf.pluie	-0.28	-0.3344	0.0049	0.0031
transf.haut	0.160	0.1480	0.0016	0.0013
vit.nec	0.015	0.0187	1.00E-05	9.61E-06
prod.inoc	50.5	42.0501	816.75	795.091
vit.expan	1.60	0.8249	0.16	0.0673
Tlim.lorin	8.020	6.0706	4.02032	2.4955

Multivariate psrf= 1.07

Utilisation de la loi *a posteriori*

Utilisation de la loi *a posteriori*

➔Calcul des distributions de prédictions de la sévérité

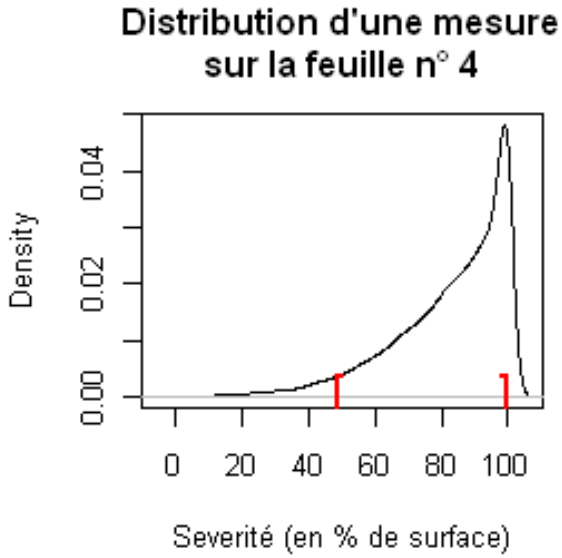
➔Calcul de la largeur moyenne des intervalles crédibles des prédictions

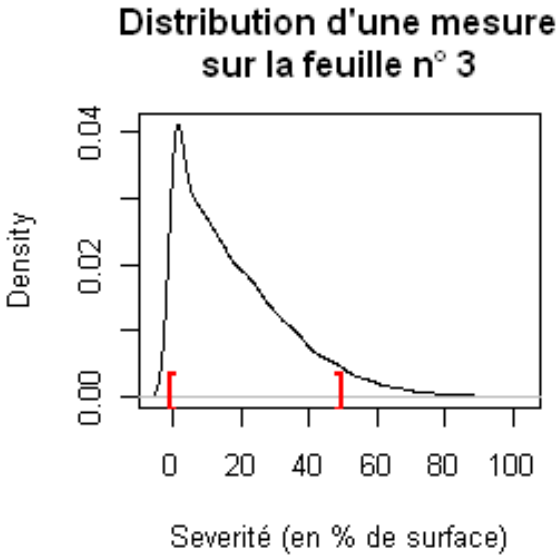
➔Séparation des contributions des sources d'erreur sur l'incertitude des prédictions

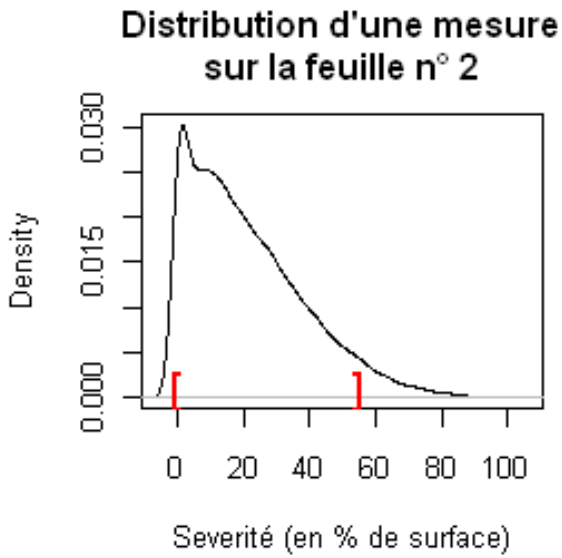
- Distribution des paramètres
- Distribution de la variance résiduelle

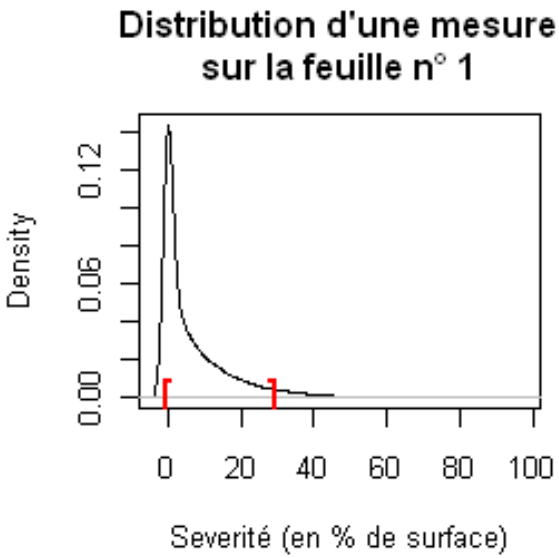
Distributions et intervalles crédibles des prédictions

Intervalles crédibles des prédictions









Intervalles crédibles des prédictions

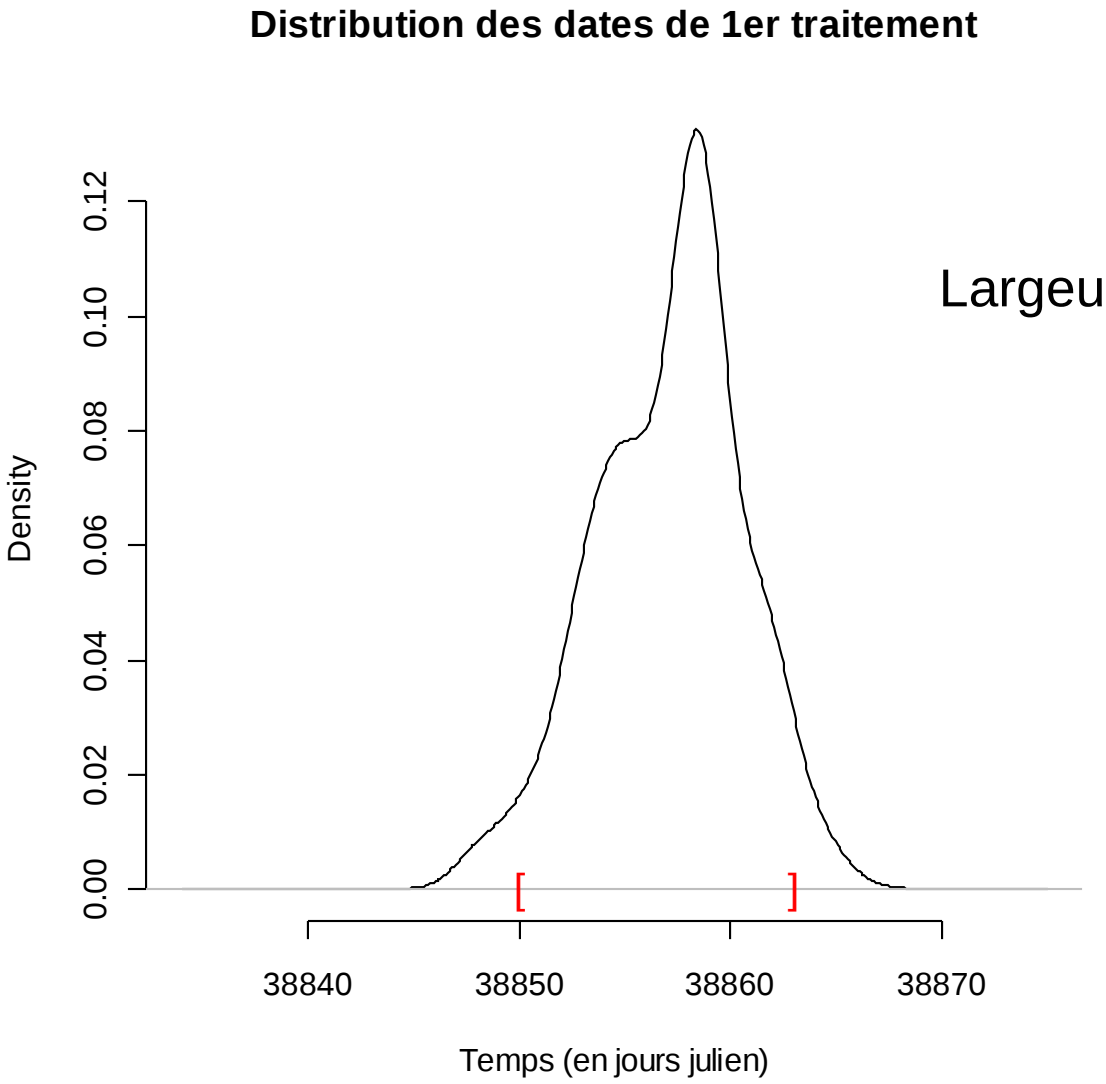
	Largeur moyenne de l'IC 90% (en % de sévérité)	
Source de variabilité	Jeu d'entraînement	Jeu de validation
θ	9,55	7,88
$\theta, \sigma_\varepsilon^2$	36,49	31,27

➔La plus grande part de variabilité des résultats est due à la variance résiduelle!

➔L'amélioration du modèle ne passe donc pas par une meilleure approximation des paramètres mais par une meilleure prise en compte de la variabilité résiduelle dans le modèle

Utilisation des résultats

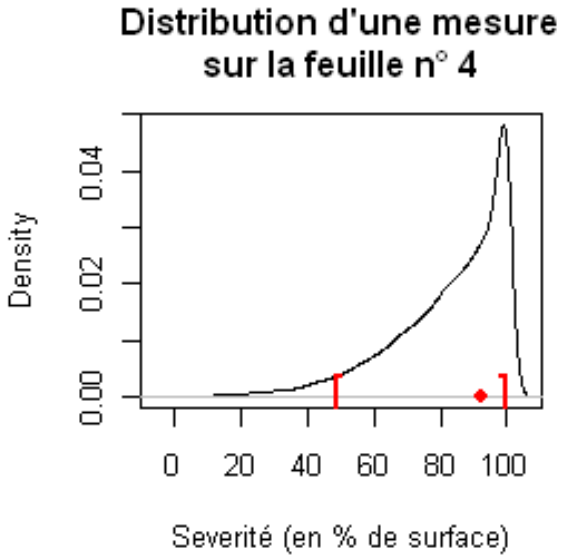
Utilisation des résultats

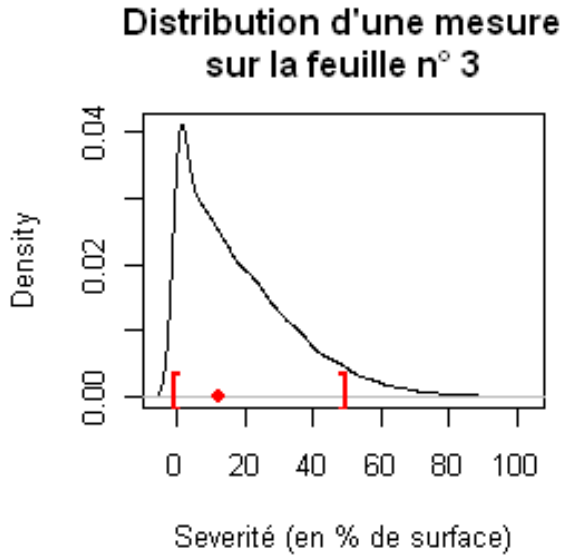


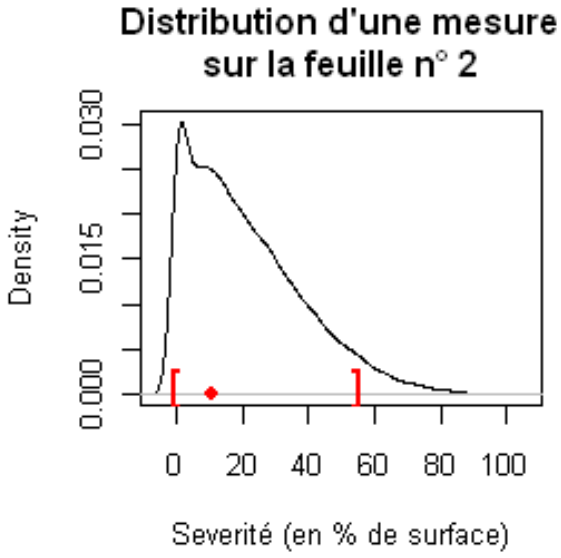
Largeur IC 90%: 13 jours

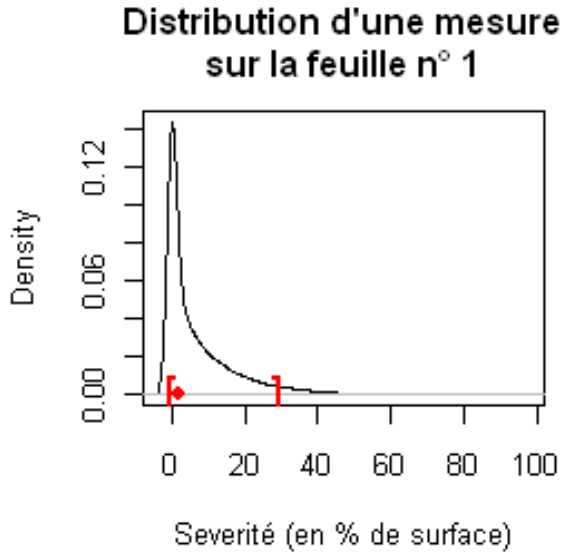
Vérification *a posteriori*

Vérification *a posteriori*









Vérification *a posteriori*

	% de Valeurs dans l'IC 90%	
Source de variabilité	Jeu d'entraînement	Jeu de validation
$\theta, \sigma_\varepsilon^2$	90	91,4

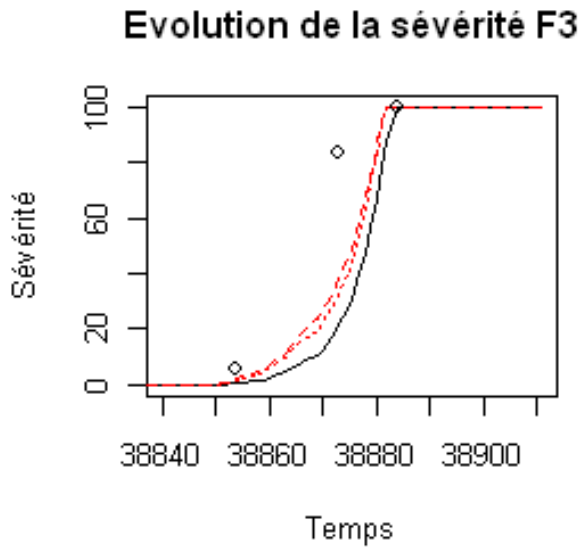
➔ Couverture réelle ≈ Couverture nominale

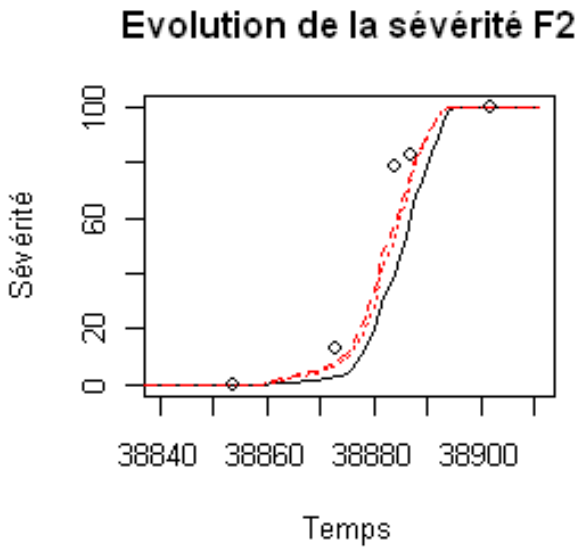
Comparaison avec d’autres approches

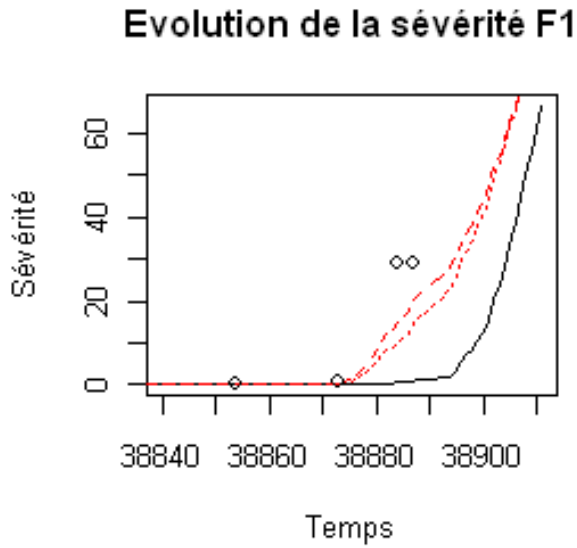
Comparaison avec d'autres approches

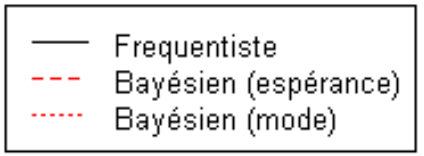
- Estimateurs Bayésien
  - Espérance des prédictions *a posteriori*
  - Prédictions avec l'espérance marginale *a posteriori*: $E(\theta|y)$
  - Prédictions avec le mode *a posteriori*
- Estimateur Fréquentiste
  - Maximum de vraisemblance (3 paramètres)
- 1 critère de comparaison ➔ MSE

Comparaison avec d'autres approches









Comparaison avec d’autres approches

	MSE (jeu d'entrainement)
Bayésien moyenne des prédictions	0.0576
Bayésien E(θ|y)	0.0606
Bayésien MAP	0.0645
Fréquentiste	0.0890

- ✓ Meilleure performance des estimateurs bayésiens

Comparaison avec d’autres approches

	MSE (jeu d'entrainement)	MSE (jeu de validation)
Bayésien moyenne des prédictions	0.0576	0.0527
Bayésien E(θ|y)	0.0606	0.0594
Bayésien MAP	0.0645	0.0567
Fréquentiste	0.0890	0.0634

- ✓ Meilleure performance des estimateurs bayésiens
- ✓ Pas d’inflation de MSE sur des données indépendantes

Conclusions

- Méthodologique

  - ✓ On peut faire les calculs (sous R)
  - ✓ Intervalles crédibles, malgré les approximations:

    couverture réelle ≈ couverture nominale
  - ✓ Meilleure performance des estimateurs Bayésiens par rapport au Fréquentiste

Conclusions

- Méthodologique
  - ✓ On peut faire les calculs (sous R)
  - ✓ Intervalles crédibles, malgré les approximations:

    couverture réelle ≈ couverture nominale
  - ✓ Meilleure performance des estimateurs Bayésiens par rapport au Fréquentiste
- Pratique
  - ✓ Largeur d'intervalle crédible de la sévérité ≈35%
  - ✓ Largeur d'intervalle pour 1ère date de traitement ≈ 13 jours
  - ✓ L'essentiel de l'erreur vient de l'erreur résiduelle
  - ✓ Pour améliorer, il faudrait travailler sur le modèle (pas sur les paramètres)

La suite…

- Amélioration de l'algorithme
  - ✓ Problème temps de calcul
  - ✓ Accélérer la convergence
- Revenir sur les premières approximations
  - ➢ Prendre en compte les corrélations entre les erreurs
    - ✓ Intra/inter site-année
    - ✓ Intra/inter-feuille
- Quelles (autres) possibilités de vérification *a posteriori* ?

To be continued…

« La vie c’est comme une chaîne de Markov… On ne sait jamais sur quoi on va tomber !»