Modélisation des flux de gènes: Quand les Bayésiens se mêlent de la question OGM.

Arnaud Bensadoun
En collaboration avec:
Antoine Messéan, Hervé Monod, Frédérique Angevin, David Makowski

Rencontre AppliBUGS
20 Juin 2013

Plan

Plan

Plan

Contexte de la coexistence

- Plusieurs types de cultures existent :
  Conventionnel, biologique ou biotechnologique (OGM).
- Chaque agriculteur doit pouvoir choisir le mode de production qu'il souhaite.
  $\Rightarrow$ Dispersion de pollen
  $\Rightarrow$ Risque de pollinisation croisée
- Le consommateur doit pouvoir choisir le produit qu'il souhaite consommer en connaissance de cause.
  $\Rightarrow$ Contrôle, Valeur seuil, étiquetage

Contexte de la coexistence

- Plusieurs types de cultures existent :
  Conventionnel, biologique ou biotechnologique (OGM).
- Chaque agriculteur doit pouvoir choisir le mode de production qu'il souhaite.
  $\Rightarrow$ Dispersion de pollen
  $\Rightarrow$ Risque de pollinisation croisée
- Le consommateur doit pouvoir choisir le produit qu'il souhaite consommer en connaissance de cause.
  $\Rightarrow$ Contrôle, Valeur seuil, étiquetage

Enjeux

$\Longrightarrow$ Assurer la séparation des filières compte tenu des risques de présence fortuite de transgène dans des lots considérés comme non OGM.

Règlementation et distance d'isolement

Commission Européenne formule des recommandations pour l'établissement de mesures de coexistence :

Règlementation et distance d'isolement

Commission Européenne formule des recommandations pour l'établissement de mesures de coexistence :

- Fonder ses décisions sur des résultats scientifiques
- Prendre des mesures spécifiques en fonction des espèces et variétés mais aussi du contexte régional

Règlementation et distance d'isolement

Commission Européenne formule des recommandations pour l'établissement de mesures de coexistence :

- Fonder ses décisions sur des résultats scientifiques
- Prendre des mesures spécifiques en fonction des espèces et variétés mais aussi du contexte régional

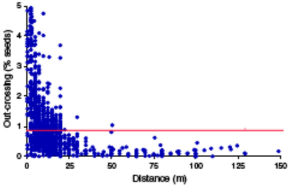


Figure: Taux de pollinisation croisée en fonction de la distance à la source OGM
Source : Riesgo et al., 2010

Règlementation et distance d'isolement

Commission Européenne formule des recommandations pour l'établissement de mesures de coexistence :

- Fonder ses décisions sur des résultats scientifiques
- Prendre des mesures spécifiques en fonction des espèces et variétés mais aussi du contexte régional

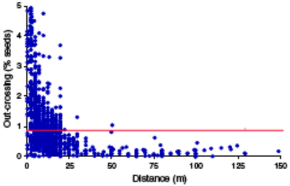


Figure: Taux de pollinisation croisée en fonction de la distance à la source OGM
Source : Riesgo et al., 2010

Enjeux

$\Longrightarrow$ Adapter les règles de coexistence à la diversité des situations possibles.
$\Longrightarrow$ Intégrer la variabilité et les incertitudes dans la prise de décision.

Règlementation et distance d'isolement

Etat membre	Maïs conventionnel	Maïs biologique
Allemagne	150	300
Autriche	∅	∅
Belgique	∅	∅
Danemark	150	150
Espagne	∅	∅
Finlande	∅	∅
France	∅	∅
Hongrie	400	400
Irlande	50	75
Italie	∅	∅
Lettonie	200	200
Lituanie	200	200
Luxembourg	600	600
Pays-Bas	25	250
Portugal	200	300
Rép. Tchèque	70	200
Roumanie	200	200
Slovaquie	200	300
Suède	50	50

Table 1 : Distance d'isolement (en m) entre OGM et nonOGM proposées dans différents Etats membre (hors production de semences)

Plan

Connaissances et Besoins

- Connaissances sur les risques de mélange et en particulier sur les flux de gènes.
  $\Rightarrow$ Bases de données de résultats d'expérimentations au champ
- Modèles génériques de la dispersion du transgène à l'échelle des paysages : modèles MAPOD sur maïs (Angevin et al., 2008) et GeneSys sur colza (Colbach et al., 2001)

Connaissances et Besoins

- Connaissances sur les risques de mélange et en particulier sur les flux de gènes.
  $\Rightarrow$ Bases de données de résultats d'expérimentations au champ
- Modèles génériques de la dispersion du transgène à l'échelle des paysages : modèles MAPOD sur maïs (Angevin et al., 2008) et GeneSys sur colza (Colbach et al., 2001)

Objectifs de l'étude

- Elaboration de modèles simplifiés
- Quantification de l'incertitude des prédictions
- Intégration de la variabilité et des incertitudes dans la prise de décision
- Valorisation de données hétérogènes à l'échelle du paysage

$\Downarrow$

Intérêt de l'approche bayésienne

Plan

1. Introduction
   - Contexte
   - Besoins
2. Matériels et Méthodes
   - Données
   - Modèles de l'espérance
   - Modèles d'observations
3. Résultats
   - Estimation
   - Prédictions
   - Choix de modèles
4. Discussions

Plan

Dispositifs Expérimentaux

Les flux de gènes sur maïs ont été étudiés selon deux types de dispositif :

- Une seule parcelle émettrice (une ou plusieurs parcelles réceptrices). Dispositif le plus courant
- Dispositifs multisources (plusieurs émetteurs) à l'échelle du paysage dans des situations réelles de coexistence.

Dispositifs Expérimentaux

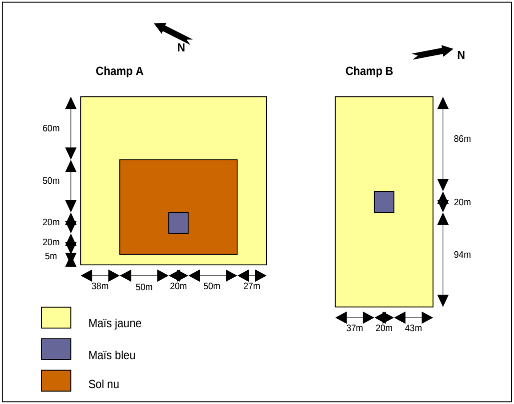
Les flux de gènes sur maïs ont été étudiés selon deux types de dispositif :

- Une seule parcelle émettrice (une ou plusieurs parcelles réceptrices). Dispositif le plus courant
- Dispositifs multisources (plusieurs émetteurs) à l'échelle du paysage dans des situations réelles de coexistence.

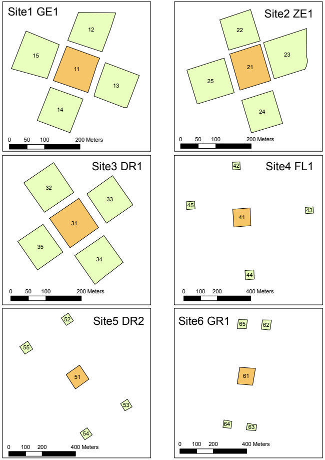
Observations :

- $y(s)$: nombre de grains contenant le transgène en un point s
- $\bar{N}$: Nombre moyen de grains par épi

Dispositif “Mono-source” (France)

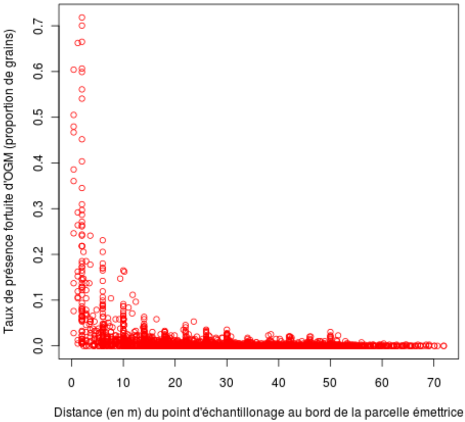


Dispositif "Mono-source" (Pays-Bas)



Dispositif “Multi-sources” (Espagne)

Exemple - Données : Essai Montargis 1998



Plan

Fonction de dispersion individuelle

Fonction de dispersion individuelle :
$\gamma_\theta(s, s')$ représente la probabilité qu'un grain de pollen émis en s' tombe et fertilise une plante en s.

Fonction de dispersion individuelle

Fonction de dispersion individuelle :
$\gamma_\theta(s, s')$ représente la probabilité qu'un grain de pollen émis en s' tombe et fertilise une plante en s.

Plusieurs fonction de dispersion individuelle existent dans la littérature :

Normal Inverse Gaussian *(Klein et al., 2003)*

$$\gamma_{NIG}(x, y) = \frac{\delta_x \delta_y e^{\lambda_z}}{2\pi} \frac{q(x, y)^{-1/2} + p^{1/2}}{q(x, y)} e^{-\sqrt{pq(x,y)}} e^{x\delta_x \lambda_x + y\delta_y \lambda_y}$$

Avec récepteur en (x, y) et émetteur en $(0, 0)$

Fonction de dispersion individuelle

Fonction de dispersion individuelle :
$\gamma_\theta(s, s')$ représente la probabilité qu'un grain de pollen émis en s' tombe et fertilise une plante en s.

Plusieurs fonction de dispersion individuelle existent dans la littérature :

Bivariate Student *(Clark, 1998)*

$$\gamma_{2Dt}(r, \omega) = \frac{b-1}{\pi a^2}\left(1 + \frac{r^2}{a^2}\right)^{-b} e^{\kappa \cos(\omega - \omega_0)}$$

Avec récepteur en (x, y) et émetteur en $(0, 0)$
$r = \sqrt{x^2 + y^2}$ et ω l'angle entre $(0, 0)$ et (x, y)

Fonction de dispersion individuelle

Fonction de dispersion individuelle :
$\gamma_\theta(s, s')$ représente la probabilité qu'un grain de pollen émis en s' tombe et fertilise une plante en s.

Plusieurs fonction de dispersion individuelle existent dans la littérature :

Exponential *(Damgaard et al., 2005)*

$$\gamma_{Expo}(r^*) = \begin{cases} e^{-a_1 r^*} & r^* \leq D \\ e^{-a_1 D - a_2 r^*} & r^* \geq D \end{cases}$$

Avec récepteur en (x, y) et émetteur en $(0, 0)$
$r^* = r \times (1 - \theta_v \cos(\omega - \omega_0))$ et $r = \sqrt{x^2 + y^2}$

Modèle de l'espérance

2 Approches possibles pour une même fonction $\gamma_\theta(s, s')$

Modèle de l'espérance

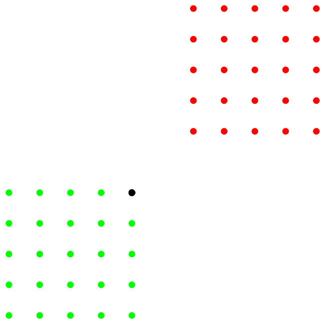
Approche Dispersion Globale :
L'espérance du nombre de grains marqués μ_θ en un point s dépend de la distance à l'émetteur OGM le plus proche.

$$\mu_\theta = \bar{N} \times \gamma_\theta(s, s')$$

Modèle de l'espérance

Approche Dispersion Globale :
L'espérance du nombre de grains marqués μ_θ en un point s dépend de la distance à l'émetteur OGM le plus proche.

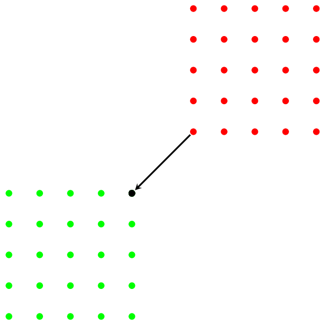
$$\mu_\theta = \bar{N} \times \gamma_\theta(s, s')$$



Modèle de l'espérance

Approche Dispersion Globale :
L'espérance du nombre de grains marqués μ_θ en un point s dépend de la distance à l'émetteur OGM le plus proche.

$$\mu_\theta = \bar{N} \times \gamma_\theta(s, s')$$



Modèle de l'espérance

Approche Dispersion Individuelle :
L'espérance du nombre de grains marqués μ_θ en un point s dépend du rapport entre les contributions des plantes OGM et les contributions de toute les plantes du paysage considéré.

$$\mu_\theta(s) = \bar{N} \times \frac{\int_{s' \in A} \gamma_\theta(s, s')\mathrm{ds}'}{\int_{s' \in A} \gamma_\theta(s, s')\mathrm{ds}' + \int_{s' \in B} \gamma_\theta(s, s')\mathrm{ds}'}$$

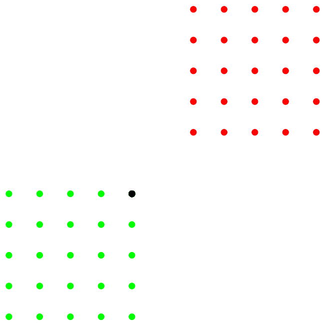
Avec A : Champ OGM, B : Champ non OGM

Modèle de l'espérance

Approche Dispersion Individuelle :
L'espérance du nombre de grains marqués μ_θ en un point s dépend du rapport entre les contributions des plantes OGM et les contributions de toute les plantes du paysage considéré.

$$\mu_\theta(s) = \bar{N} \times \frac{\int_{s' \in A} \gamma_\theta(s, s')\mathrm{d}s'}{\int_{s' \in A} \gamma_\theta(s, s')\mathrm{d}s' + \int_{s' \in B} \gamma_\theta(s, s')\mathrm{d}s'}$$

Avec A : Champ OGM, B : Champ non OGM

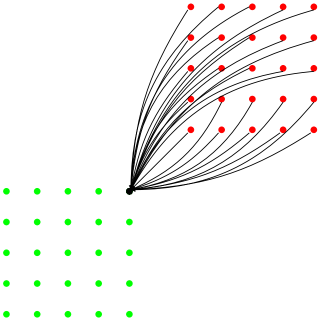


Modèle de l'espérance

Approche Dispersion Individuelle :
L'espérance du nombre de grains marqués μ_θ en un point s dépend du rapport entre les contributions des plantes OGM et les contributions de toute les plantes du paysage considéré.

$$\mu_\theta(s) = \bar{N} \times \frac{\int_{s' \in A} \gamma_\theta(s, s') \mathrm{d}s'}{\int_{s' \in A} \gamma_\theta(s, s') \mathrm{d}s' + \int_{s' \in B} \gamma_\theta(s, s') \mathrm{d}s'}$$

Avec A : Champ OGM, B : Champ non OGM



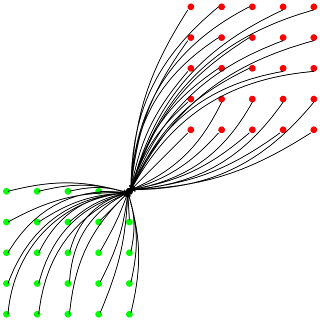
Modèle de l'espérance

Approche Dispersion Individuelle :

L'espérance du nombre de grains marqués μ_θ en un point s dépend du rapport entre les contributions des plantes OGM et les contributions de toute les plantes du paysage considéré.

$$\mu_\theta(s) = \bar{N} \times \frac{\int_{s' \in A} \gamma_\theta(s, s') \mathrm{d}s'}{\int_{s' \in A} \gamma_\theta(s, s') \mathrm{d}s' + \int_{s' \in B} \gamma_\theta(s, s') \mathrm{d}s'}$$

Avec A : Champ OGM, B : Champ non OGM



Plan

Paradigme Bayésien

$$P(\theta|Y) \propto \pi(\theta) \times \mathcal{L}(Y|\theta)$$

Nécessite :

- Des données Y
- Des connaissances *a priori* $\pi(\cdot)$ sur les paramètres θ
- Une fonction de vraisemblance $\mathcal{L}(Y|\theta)$

$\Rightarrow$ Requiert la définition d'un modèle probabiliste reliant les observations aux paramètres du modèle.

Modèles d'observations

Observation de comptages d'évènements relativement rares
$\Rightarrow$ Modèle de Poisson

$$Y \sim \mathcal{P}(\mu') \tag{1}$$

Modèles d'observations

Observation de comptages d'évènements relativement rares
$\Rightarrow$ Modèle de Poisson

$$Y \sim \mathcal{P}(\mu') \tag{1}$$

Mais observations très variables ET sur-représentation de zéros
$\Rightarrow$ Modèle de Poisson Zéro Inflaté

$$Y \sim ZI\mathcal{P}(1-q, \mu') \tag{2}$$

$$\Leftrightarrow$$

$$\begin{array}{l} Y|Z = 0 \sim \delta(\{0\}) \\ Y|Z = 1 \sim \mathcal{P}(\mu') \end{array}$$

Modèles d'observations

Observation de comptages d'évènements relativement rares
$\Rightarrow$ Modèle de Poisson

$$Y \sim \mathcal{P}(\mu') \tag{1}$$

Mais observations très variables ET sur-représentation de zéros
$\Rightarrow$ Modèle de Poisson Zéro Inflaté

$$Y \sim ZI\mathcal{P}(1-q,\mu') \tag{2}$$

$$\Leftrightarrow$$

$$Y|Z=0 \sim \delta(\{0\})$$
$$Y|Z=1 \sim \mathcal{P}(\mu')$$

$$Z \sim \mathcal{B}ern(q)$$

$$logit(q) = \beta_1(\beta_2 - d)$$

Modèles d'observations

Gestion de la variabilité supplémentaire
⇒ Modèle normal pour l'espérance

$$\mu' = \mu_\theta \tag{3}$$

$$\mu' \sim \mathcal{N}(\mu_\theta, \sigma^2) \tag{4}$$

$$\mu' \sim \mathcal{N}(\mu_\theta, \mu_\theta \times \sigma^2) \tag{5}$$

Protocole de simulation

- Trois fonctions de dispersion individuelle γ_θ : NIG, 2Dt, Exponentielle
- Deux approches pour γ_θ : Globale et Individuelle
- Deux modèles d'observations pour Y : Poisson, Poisson Zéro Inflaté
- Trois modèles pour l'espérance μ' : Fixe, Normal, Normal à variance proportionnelle

Protocole de simulation

- Trois fonctions de dispersion individuelle γ_θ : NIG, 2Dt, Exponentielle
- Deux approches pour γ_θ : Globale et Individuelle
- Deux modèles d'observations pour Y : Poisson, Poisson Zéro Inflaté
- Trois modèles pour l'espérance μ' : Fixe, Normal, Normal à variance proportionnelle

$$\Downarrow$$

Plan factoriel complet

$$3_{\gamma_\theta} \times 2_{Approches} \times 2_{observations} \times 3_{esperances} = 36 \text{ Modèles}$$

Plan d'expérience

Distribution Y	IDF γ_θ	Distribution μ'
$\mathcal{P}(\mu')$	Exponential	μ
$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \mu\sigma^2)$
$\mathcal{P}(\mu')$	2Dt	μ
$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \mu\sigma^2)$
$\mathcal{P}(\mu')$	NIG	μ
$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	Exponential	μ
$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	2Dt	μ
$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	NIG	μ
$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \mu\sigma^2)$

Plan

Plan

Estimation des paramètres

Méthode d'estimation : Monte Carlo Markov Chain (MCMC)

Software : JAGS

Convergence : OK

Temps d'estimation : Selon modèle entre 3 et 12h
(Avant optimisation/simplification du calcul de l'intégrale entre 2 et 8 jours !)

Estimation des paramètres

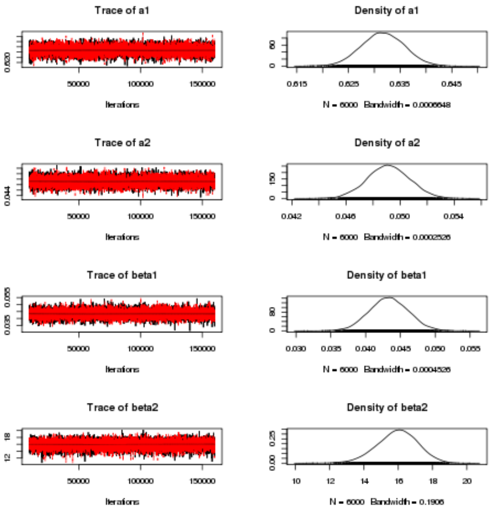


Figure: Trace et densité des paramètres

Plan

Prédictions

Distribution Y	IDF γ_θ	Distribution μ'
$\mathcal{P}(\mu')$	Exponential	μ
$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \mu\sigma^2)$
$\mathcal{P}(\mu')$	2Dt	μ
$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \mu\sigma^2)$
$\mathcal{P}(\mu')$	NIG	μ
$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$
$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	Exponential	μ
$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	2Dt	μ
$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \mu\sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	NIG	μ
$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$
$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \mu\sigma^2)$

Prédictions

Modèle de Poisson

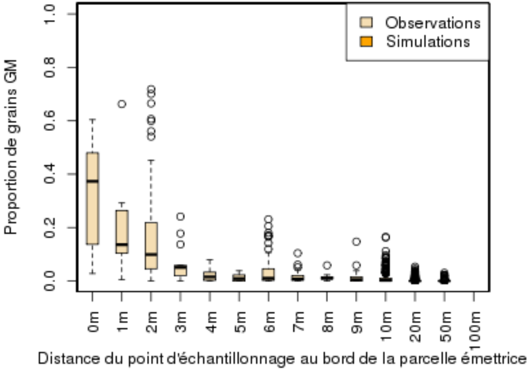


Figure: Observations en fonction de la distance

Prédictions

Modèle de Poisson

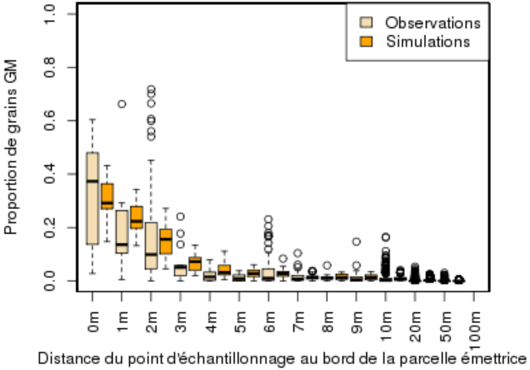


Figure: Observations en fonction de la distance vs Simulations avec le modèle Poisson 2Dt

Prédictions

Modèle de Poisson

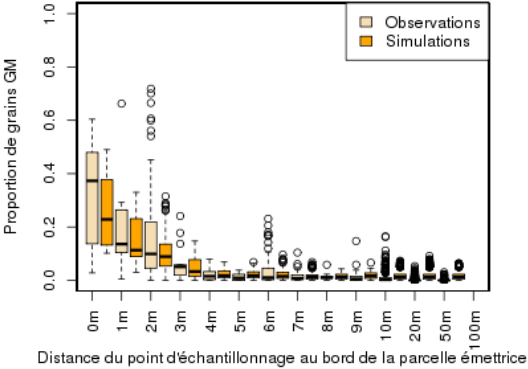


Figure: Observations en fonction de la distance vs Simulations avec le modèle Poisson 2Dt avec espérance normale

Prédictions

Modèle de Poisson

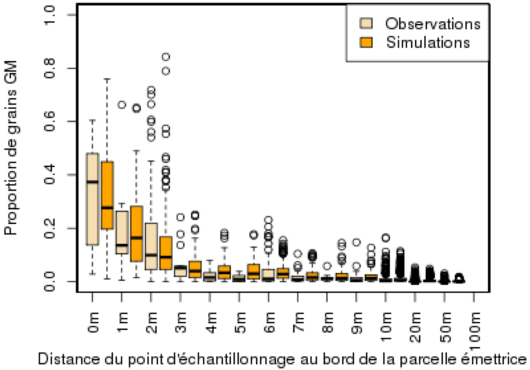


Figure: Observations en fonction de la distance vs Simulations avec le modèle Poisson 2Dt avec espérance normale à variance proportionelle

Plan

Choix de modèles

Pour chaque observation $y(s)$ on a une distribution empirique prédictive $F(s)$

Choix de modèles

Pour chaque observation $y(s)$ on a une distribution empirique prédictive $F(s)$

Critères "classiques"
$\Rightarrow$Considèrent uniquement la réponse moyenne $\hat{Y}_s = \mathbb{E}_{F(s)}$:

- Correlation : $r = \frac{\sigma_{Y\hat{Y}}}{\sigma_Y \sigma_{\hat{Y}}}$
- Root mean squared error : $RMSE = \sqrt{\frac{1}{N}\sum_1^N (y_s - \hat{Y}_s)^2}$
- Modelling efficiency : $EF = 1 - \frac{\sum_{s=1}^{N}(Y_s - \hat{Y}_s)^2}{\sum_{s=1}^{N}(Y_s - \mathbb{E}_Y)^2}$

Critères "Bayésiens"
$\Rightarrow$Considèrent toute la distribution :

- $DIC = p_D + \bar{D}$
- $CRPS(F(s), y(s)) = \frac{1}{2}\mathbb{E}_{F(s)}|Y - Y'| - \mathbb{E}_{F(s)}|Y - y(s)|$

Choix de modèles

1 jeu de données d'entrainement

- Estimation
- Prédiction
- Calcul de critères

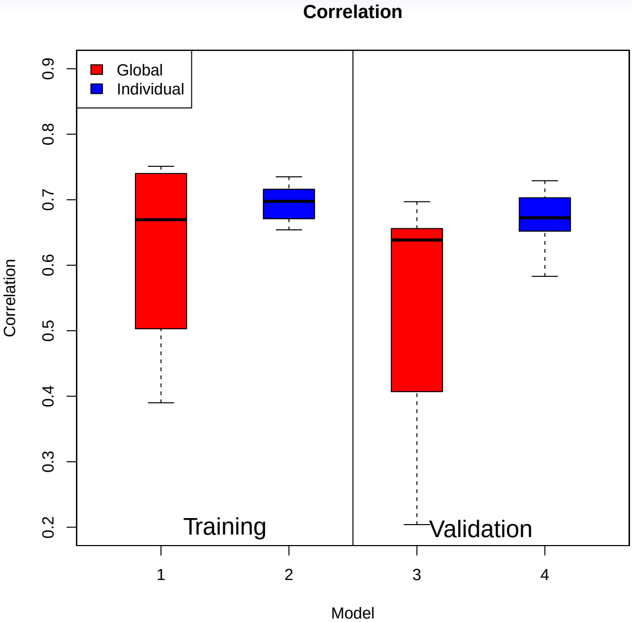
Choix de modèles

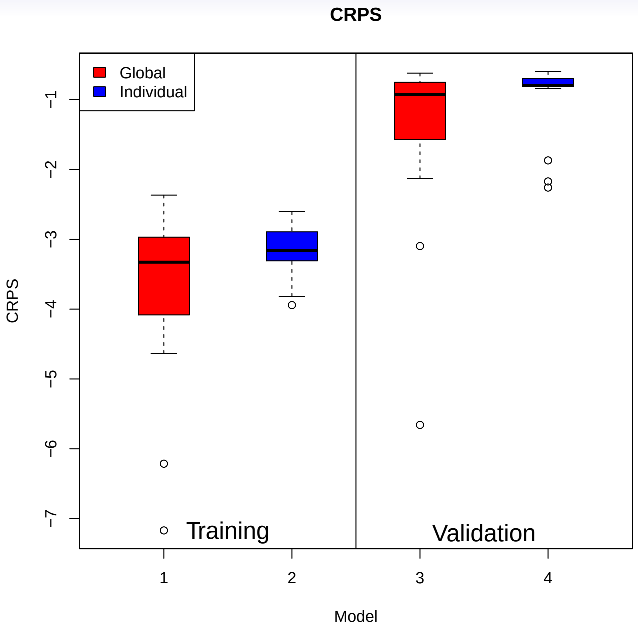
1 jeu de données d'entrainement

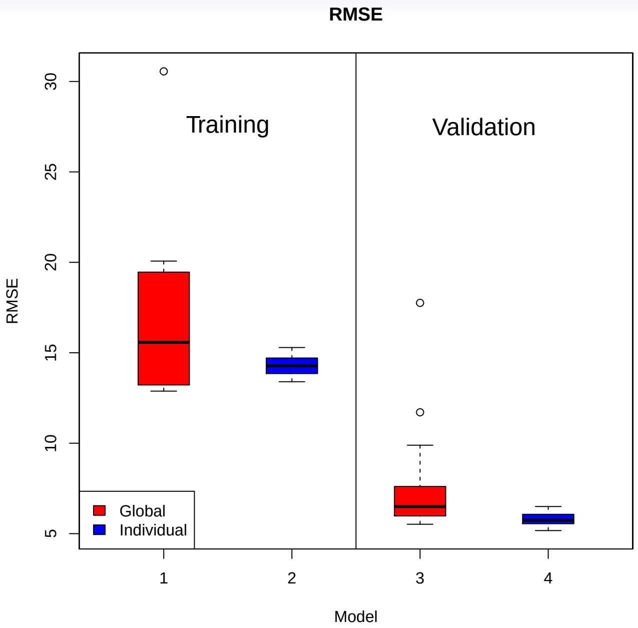
- Estimation
- Prédiction
- Calcul de critères

1 jeu de données de validation

- Prédiction avec θ estimés sur jeu d'entrainement
- Calcul de critères


Correlation
Global
Individual
Correlation
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Training
Validation
1
2
3
4
Model


CRPS
Global
Individual
CRPS
−1
−2
−3
−4
−5
−6
−7
Training
Validation
1
2
3
4
Model


RMSE
Training
Validation
RMSE
30
25
20
15
10
5
1
2
3
4
Model
Global
Individual

Plan

Discussions

- Méthodes bayésiennes
  - Application aux modèles de dispersion individuelle
  - Prise en compte de la variabilité

Discussions

- Méthodes bayésiennes
  - Application aux modèles de dispersion individuelle
  - Prise en compte de la variabilité
- Meilleure qualité d'ajustement pour l'approche globale
- Meilleure qualité prédictive pour l'approche individuelle

Discussions

- Méthodes bayésiennes
  - Application aux modèles de dispersion individuelle
  - Prise en compte de la variabilité
- Meilleure qualité d'ajustement pour l'approche globale
- Meilleure qualité prédictive pour l'approche individuelle
- Sélection de modèle
  - Meilleur modèle ?
  - Modèle moyen : *Bayesian Model Averaging* ?

Discussions

- Méthodes bayésiennes
  - Application aux modèles de dispersion individuelle
  - Prise en compte de la variabilité
- Meilleure qualité d'ajustement pour l'approche globale
- Meilleure qualité prédictive pour l'approche individuelle
- Sélection de modèle
  - Meilleur modèle ?
  - Modèle moyen : *Bayesian Model Averaging* ?
- Extension au multi-sources
  - Prise en compte du paysage
  - Obtention de covariables à cette échelle
  - Temps de calcul

Merci pour votre attention

Des questions ?

"La vie c'est comme une chaîne de Markov, on ne sait jamais sur quoi on va tomber !"