MCMC ou ABC ? Bonheurs et tourments de mise en oeuvre sur un cas de risque sanitaire

Clémence C. Rigaux, Sophie S. Ancelet, Frederic F. Carlin, Christophe C. Nguyen The, Isabelle I. Albert

Met@risk
Methods for Food Risk Analysis

MCMC ou ABC ?

Bonheurs et tourments de mise en œuvre sur un cas de risque sanitaire

Clémence RIGAUX

En collaboration avec Sophie Ancelet, Frédéric Carlin, Christophe Nguyen-Thé, Isabelle Albert

INRA

AppliBUGS , 17 juin 2011

Contexte : l'analyse du risque microbiologique dans les aliments

• Des modèles prédisant le comportement de bactéries pathogènes le long d'une chaîne de transformation des aliments (jusqu'à la maladie)

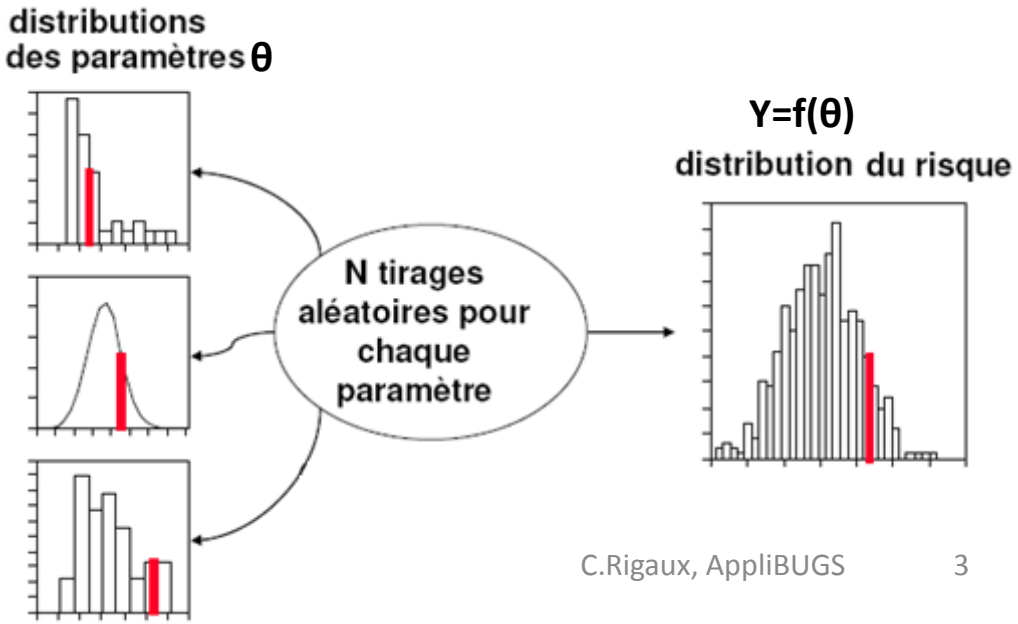
• Développement de méthodes quantitatives d'évaluation du risque microbiologique (QMRA : Quantitative Microbial Risk Assessment)

  • permettant de rendre compte des diverses situations par la prise en compte des sources de *variabilité* et d'*incertitudes*
  • permettant le calcul du risque de toxi-infection

INRA

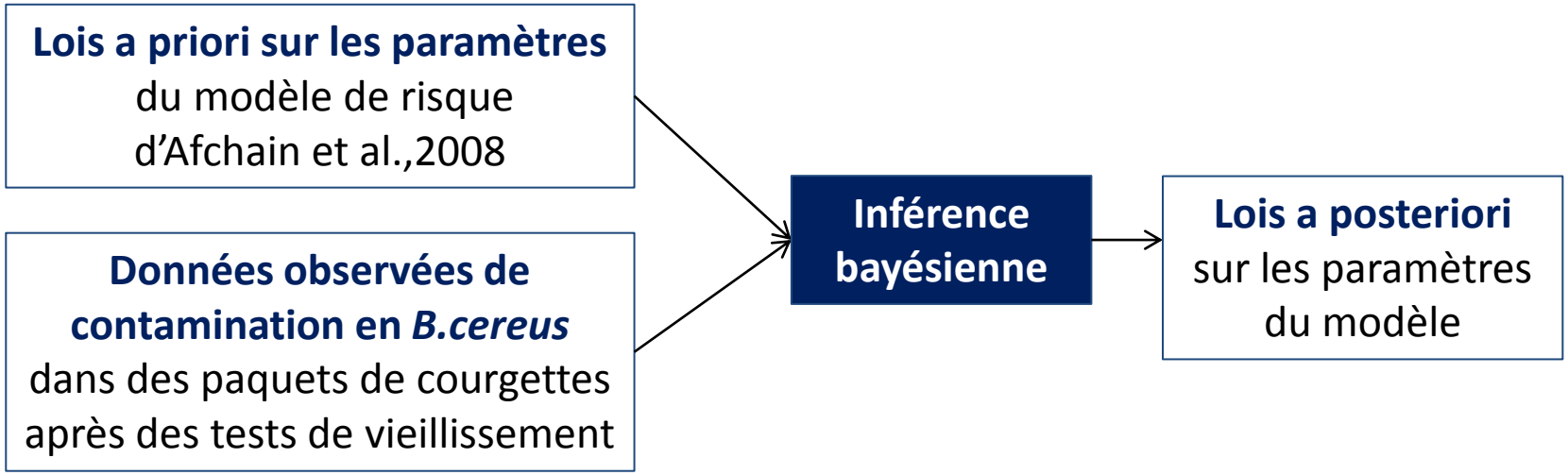
Cas d'étude : *B.cereus* dans la purée de courgettes

- Un modèle de risque alimentaire dû au pathogène *Bacillus cereus* dans une chaîne de fabrication de purée de courgettes
  → Construit par Afchain et al., 2008
  → Tenant compte de la grande diversité génétique de *B.cereus* : 6 groupes
  → Analyse du risque par la méthode de simulation de Monte Carlo



Mise en place d'une méthode d'inférence bayésienne

But : utiliser l'information issue de données de contamination dans les paquets de courgettes pour améliorer le modèle de Afchain et al.,2008, qui peut être vu comme un réseau bayésien :



→ remontée de l'information de long de la chaîne

Mise en place d'une méthode d'inférence bayésienne

Le théorème de Bayes donne la **loi a posteriori**, ie la loi des paramètres θ sachant les données X :

$$[\theta|X] = \frac{[\theta][X|\theta]}{[X]}$$

→ Ici, deux méthodes possibles de calcul :

Méthode de Monte Carlo par Chaîne de Markov : MCMC

Chaînes de Markov convergeant vers la loi a posteriori

Méthode de calcul bayésien approché : ABC

Méthode approchée n'utilisant pas la vraisemblance des observations, utilisant la simulation des données

Plan

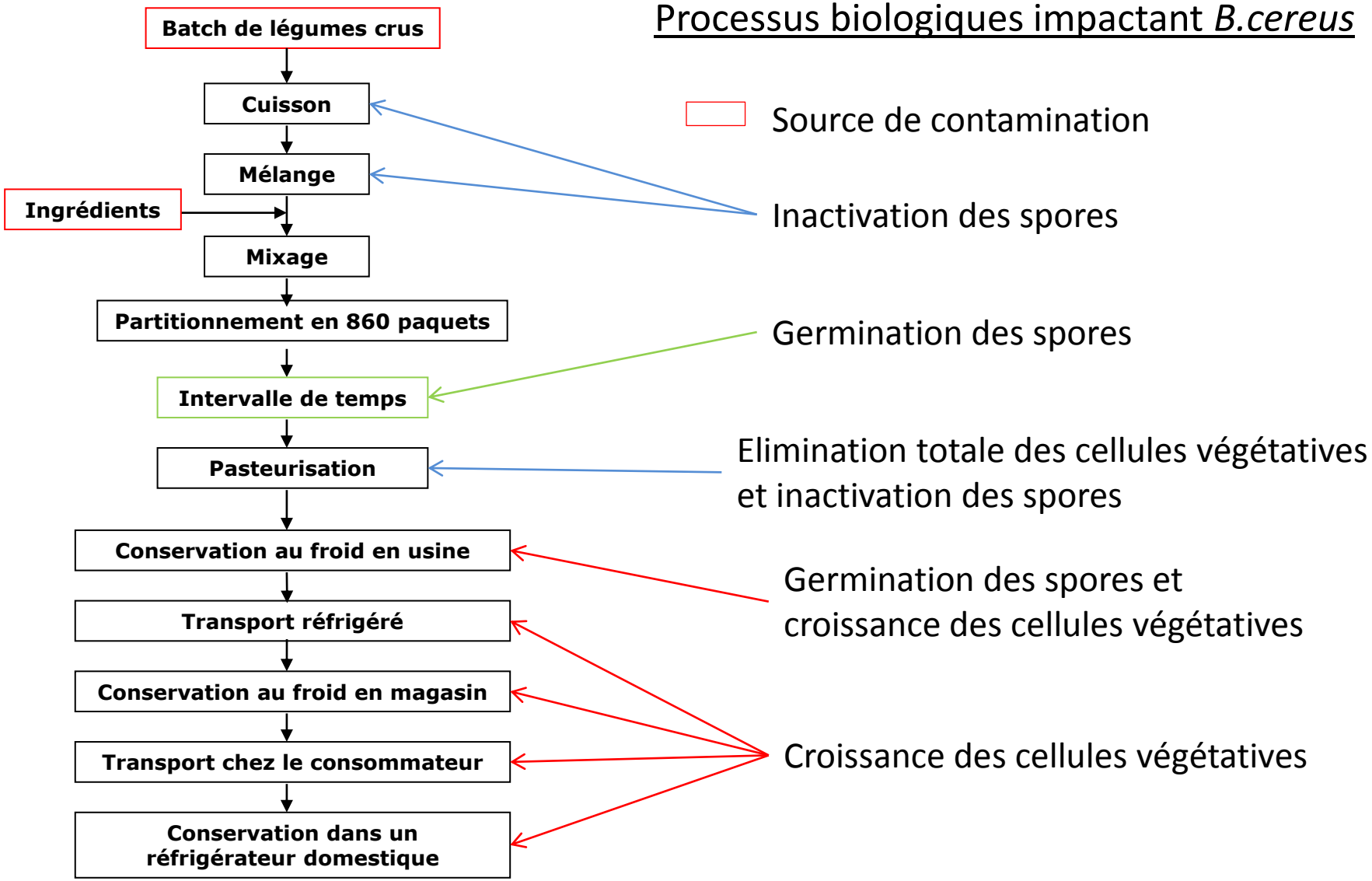
- ✓ Introduction

1. Présentation du modèle et des données
2. Essai d'inférence bayésienne par méthode ABC
3. Mise en place de l'inférence bayésienne par MCMC

Met@risk
Methods for Food Risk Analysis

INRA

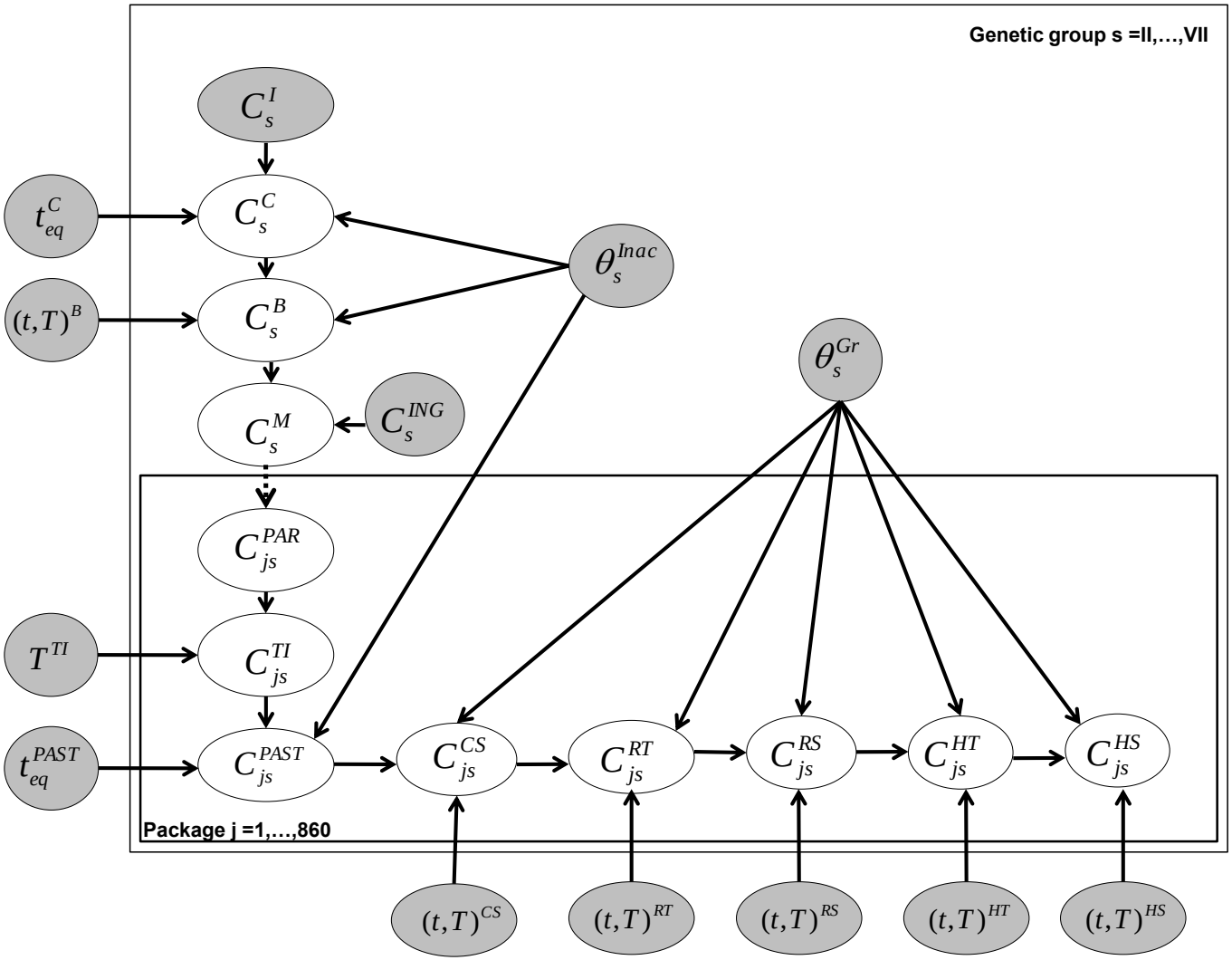
1. Présentation du modèle et des données



Modèle complexe : 132 variables, comprenant 66 inputs, 60 variables intermédiaires, 6 outputs

1. Présentation du modèle et des données

Ce modèle peut être vu comme un réseau bayésien :



1. Présentation du modèle et des données

Des données de contamination *après vieillissement*, souvent censurées à gauche :

75% des contaminations totales sous le seuil de détection

Pour les 25% des contaminations positives, censure des concentrations en souches non détectées

Données (source)	Produit	Conditions	Contamination en log CFU/g (groupes génétiques détectés)
14 paquets venant du même batch (INRA)	5 paquets de 800g	21 jours à 4°C	< 1,7
	4 paquets de 800g	21 jours à 10°C	5,5 (VI) 3 (VI) 3,8 (II) 3,1 (II)
	5 paquets de 800g	5 jours à 20-25°C	6 (IV) 7,2 (II et IV) 6,4 (IV) 6,3 (IV) 6,3 (IV)
50 paquets venant de batch différents (USINE)	27 paquets de 400g	20 jours à 4°C puis 10 jours à 8°C	3,7 (II) 2,7 (II) 3,7 (II) 5 (II) < 2 pour les 23 autres paquets
	23 paquets de 400g	10 jours à 4°C puis 20 jours à 8°C	5,2 (VI) 4,4 (VI) 4,7 (VI) < 2 pour les 20 autres paquets

Plan

- ✓ Introduction
1. Présentation du modèle et des données
2. Essai d'inférence bayésienne par méthode ABC
3. Mise en place de l'inférence bayésienne par MCMC

Met@risk
Methods for Food Risk Analysis

INRA

2. Essai d'inférence bayésienne par ABC

Motivations pour tenter l'ABC dans notre contexte :

- Complexité du modèle
- Problème initial de la loi multinomiale non programmable en Winbugs ou en Jags
- Difficulté de convergence de la méthode MCMC
- Comparaison de la méthode MCMC et de la méthode ABC

Objectif et principe de la méthode ABC :

Calcul des lois a posteriori sans utiliser de vraisemblance mais en comparant directement les données simulées (par le modèle de simulation de Monte Carlo) avec les données observées

2. Essai d'inférence bayésienne par ABC

Méthode ABC : algorithme de type acceptation/rejet :

1. Générer θ (vecteur de paramètres) à partir des lois a priori
2. Simuler des données avec θ (à l'aide du modèle)
3. Accepter θ si les données simulées sont *proches* des données observées
   Sinon, refuser θ.
   Retourner en 1.

Loi a posteriori = distribution des θ acceptés

Met@risk
Methods for Food Risk Analysis

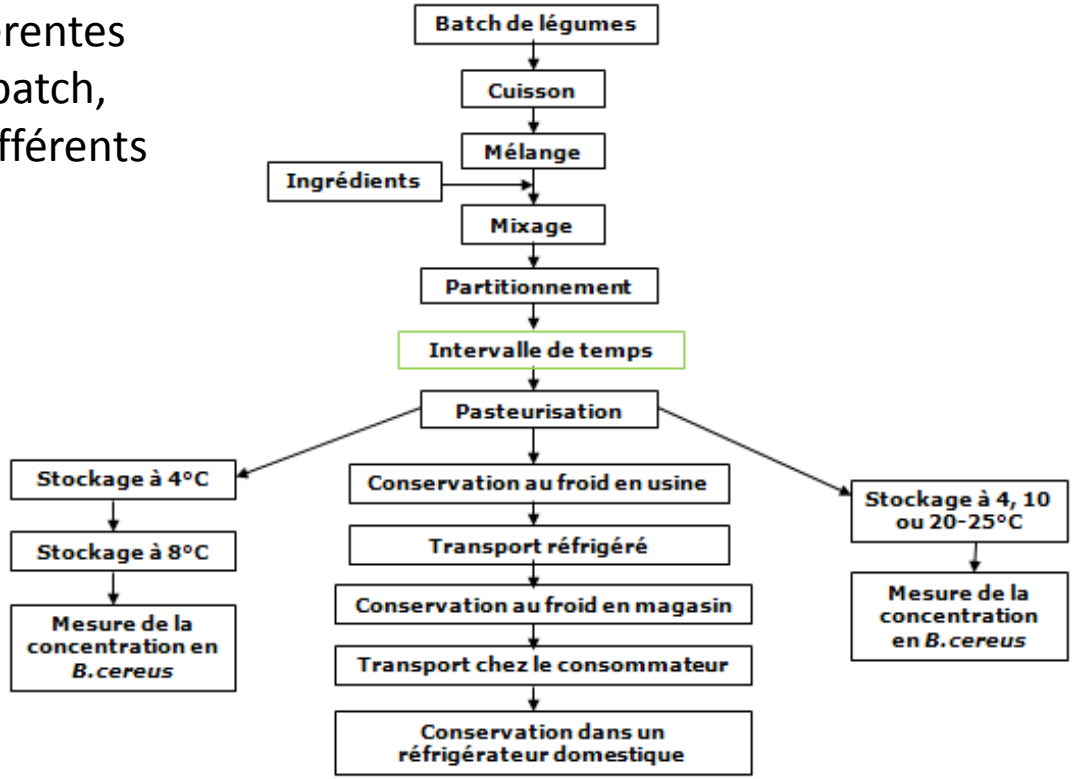
INRA

2. Essai d'inférence bayésienne par ABC

Mise en place de la méthode :

Simulation de 64 paquets dans les *mêmes conditions de vieillissement* que les données, par le modèle d'Afchain et al., 2008, *« augmenté »* :

- 5 conditions de vieillissement différentes
- 14 paquets Inra venant du même batch,
  50 paquets industriels de bachs différents



2. Essai d'inférence bayésienne par ABC

Mise en place de la méthode :

- Construction d'une ***distance*** entre données simulées et observées
- Ici en utilisant des ***statistiques résumées*** des concentrations simulées après vieillissement :

nécessaires pour obtenir une comparaison globale des concentrations des données avec les concentrations prédites (mais perte de précision)

Met@risk
Methods for Food Risk Analysis

INRA

2. Essai d'inférence bayésienne par ABC

Définition de la statistique résumée : un ensemble de conditions résumant les données :

A l'intérieur de chaque condition de vieillissement :
- Le type de groupes génétiques dominants (ie au dessus du seuil de détection, avec les autres groupes génétiques en quantité 10 fois inférieure) (II,…,VI)
- Le nombre de paquets contaminés par ces groupes dominants
- La quantité de bactéries de ces groupes dominants (log_{10}ufc/g)
- Le nombre de paquets non contaminés

Distance :

On juge une série de 64 prédictions *similaire* aux 64 données si :

A l'intérieur de chaque condition de vieillissement :
- On retrouve le même type de groupes génétiques dominants
- Et un nombre similaire de paquets contaminés par cette souche (±1 ou ±2)
- Et en quantité similaire : marge d'erreur ± δ log_{10}ufc/g
- Et un nombre similaire de paquets non contaminés

2. Essai d'inférence bayésienne par ABC

Exemple sur une condition de vieillissement :

Condition de vieillissement INRA 2 : 5 paquets	5,5 (VI)
	3 (VI)
	7,5 (VI)
	3,8 (II)
	3,1 (II)

On accepte les 5 paquets simulés correspondants si ils comportent :

• De 1 à 3 paquets avec
- entre 3.1-δ et 3.8+δ log10 ufc/g de souche II
- des concentrations simulées en souches III à VII < à 1/10 de la concentration simulée en II dans ces paquets

• De 2 à 4 paquets avec
- entre 3 – δ et 7.5+δ log10 ufc/g de souche VI
- des concentrations simulées en souches III, III, IV, V et VII < à 1/10 de la concentration simulée en IV dans ces paquets

• Les 5 paquets simulés correspondent à un des cas ci-dessus, avec éventuellement un paquet non contaminé.

2. Essai d'inférence bayésienne par ABC

Résultat :

Avec 400 000 simulations (durée : 15h) et $\delta = 0.5$ $\log_{10}$ufc/g :

Conditions	Pourcentage de simulation acceptées
Inra 1	98,9%
Inra 2	0,00%
Inra 3	1,21%
Industriel 1	0,00%
Industriel 2	0,00%
Les 64 paquets	0,00%

→ Certaines conditions très difficiles à accepter

→ L'ensemble des conditions doit être accepté simultanément pour accepter une simulation. Donc 0 simulation acceptée.

Met@risk
Methods for Food Risk Analysis

INRA

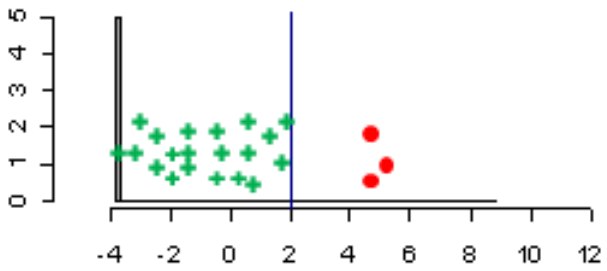
(*) résultats obtenus avec le 1er jeu de 70 données et programme initial A.L., plusieurs batchs par simulation

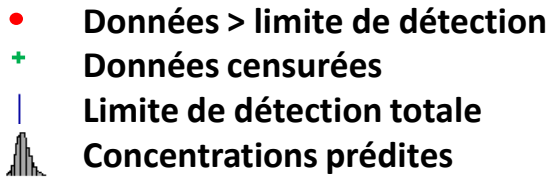
2. Essai d'inférence bayésienne par ABC

Comment expliquer qu'on n'accepte jamais ?

- Résultat du modèle multivarié, de dimension 64*6 = 384 → difficultés
- Données assez éloignées des prédictions du modèle
  - Souche VI : prévalence prédite par le modèle de 0.03%, mais dans les données prévalence supérieure à 7.8%
  - Condition Inra 2 : prédiction surtout de II, III et IV, mais dans les données: du II et du VI

Exemple : groupe génétique VI, condition Industriel 2





Remarque : <0.1 bactérie dans le paquet : assimilé à 0.1 bactérie -> - 4 log10ufc/g

2. Essai d'inférence bayésienne par ABC

Essais pour avoir des meilleurs taux d'acceptation :

- Assouplissements de la distance :
  augmentation de la marge d'erreur δ, autoriser quelques paquets non contaminés, diminution de la limite de censure pour les groupes génétiques non détectées, etc.

- Modification des priors (élargissements)
  → problème : les taux d'acceptation s'améliorent pour certaines conditions de vieillissement mais se dégradent pour d'autres

- Essais en supposant tous les paquets issus d'un même batch

  Tous les paramètres θ à ce niveau de modèle sont au niveau batch
  51 batchs → 51 jeux de θ par simulation, donc 66*51 = 3366 inputs !
  1 batch → 1 jeu de θ par simulation, donc 66 inputs seulement
  mais sous-estimation de la variabilité !

  → modèle d'Afchain et al.,2008, de Monte Carlo de 1er degré : ne sépare pas la variabilité et l'incertitude
  →problème de prise en compte de la *variabilité entre les paquets* qui viennent d'un même batch et les paquets venant de batch différents

2. Essai d'inférence bayésienne par ABC

Résultats des essais pour avoir des meilleurs taux d'acceptation :
(en supposant tous les paquets issus d'un même batch)

	Pourcentage de simulation acceptées		
Conditions	Modèle normal 1000 simulations Erreur = 1 log10ufc/g	Priors modifiés (10) 300 000 simulations Erreur = **1** log10ufc/g	Priors modifiés (10) 1 000 000 simulations Erreur = **5** log10ufc/g
Inra 1	99,8%	98,2%	98,3%
Inra 2	0,00%	0,1%	0,8%
Inra 3	1,5%	12,6%	19,6%
Industriel 1	0,60%	1,5%	1,8%
Industriel 2	0,00%	0,55%	1,44%
Les 64 paquets	0,00%	0,00%	0,00%

Configuration 10 : élargissement des priors de δ_{VI}, z, T_{minII}, $Log_{10}N_{max}$, t_{Ceq}, t_{eqpas}, T_b, TTI

→Toujours aucune acceptation simultanée !
??

(*) résultats obtenus avec le jeu de 64 données et programme bis issu du programme MCMC, 1 batch par simulation

Plan

- ✓ Introduction
1. Présentation du modèle et des données
2. Essai d'inférence bayésienne par méthode ABC
3. Mise en place de l'inférence bayésienne par MCMC

Met@risk
Methods for Food Risk Analysis

INRA

3. Inférence bayésienne par MCMC

Mise en place :

- Utilisation de Jags
- «Augmentation» du modèle : rajout d'étapes de vieillissement des paquets dans les mêmes conditions que les données
- *Vraisemblance* $[X|\theta]$: raccordement des données de concentrations en *B.cereus* (en ufc/g) aux concentrations simulées via une loi lognormale censurée (i : batch, j : paquet, s : groupe génétique, α_j = LoD ou $\log(C_{jS}^{obs}/10)$):

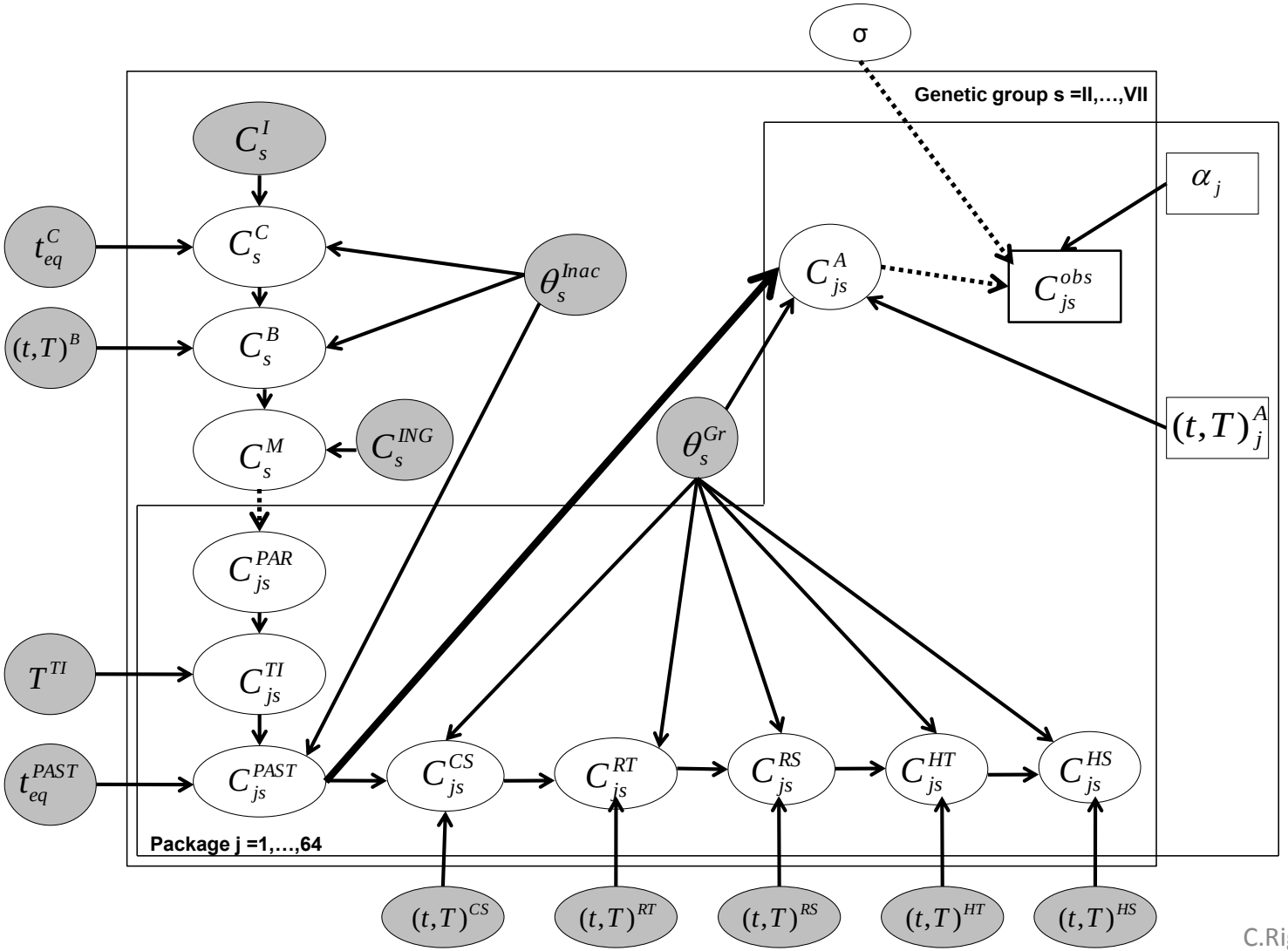
$$\log(C_{js}^{obs}) \sim N(\log(C_{js}^{A}), \sigma^2) C(;\alpha_j)$$

Données — Prédictions — Censure : uniquement pour les données censurées

Prior mis sur l'erreur : $\sigma \sim Unif(0,100)$

3. Inférence bayésienne par MCMC

Le réseau bayésien augmenté :



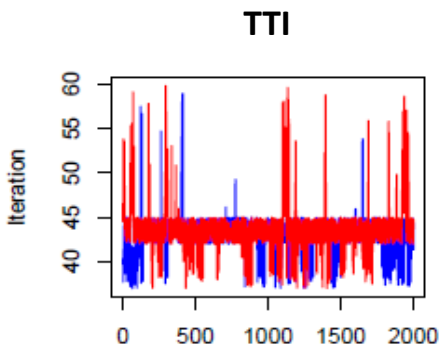
3. Inférence bayésienne par MCMC

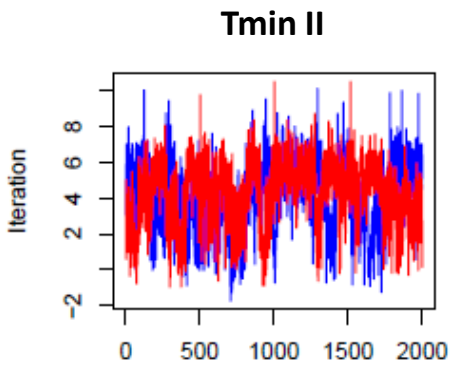
Méthodes de calcul :

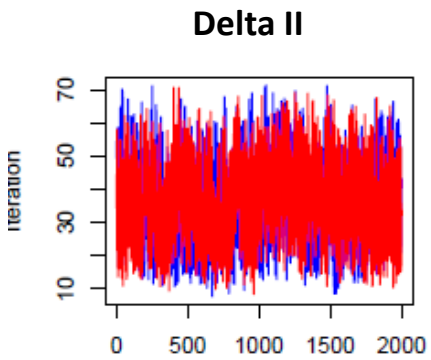
- Simulation de 1 batch et de 64 paquets par batch
  - → Hypothèse : les paquets sont tous issus d'un même batch
  - → Même problème de prise en compte de la variabilité/incertitude que pour l'ABC

- Avec le logiciel Jags, simulation de 2 chaînes de Markov indépendantes
- Période de chauffe (« Burn-in ») de 1 000 000 itérations
- 2 000 000 itérations avec un pas de 1000

➢ Convergence atteinte pour la majorité des inputs sauf difficultés pour 2 ou 3 : grosse autocorrélation

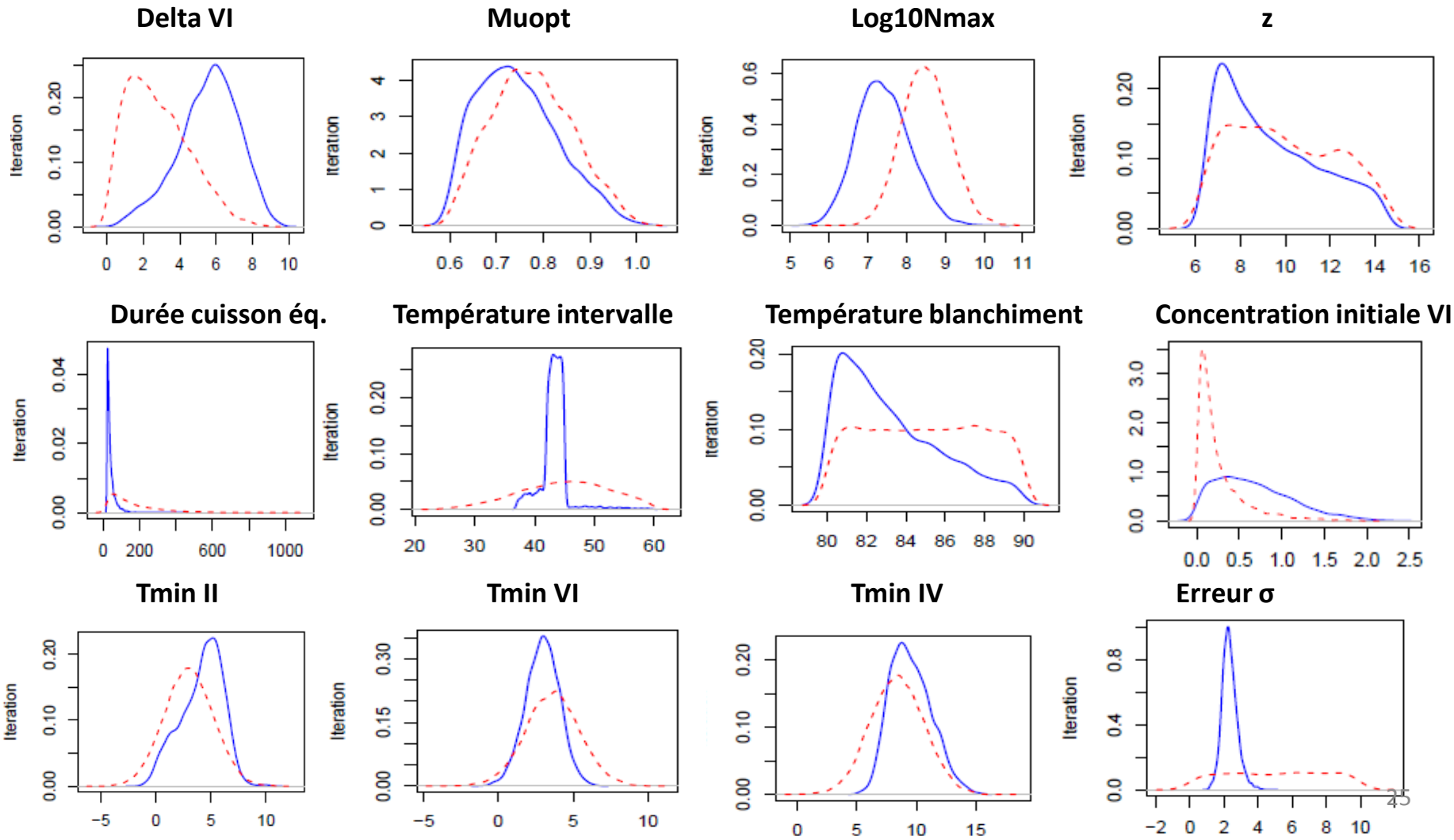






3. Inférence bayésienne par MCMC

Résultats : prior (--) versus posterior (—) : les + importantes modifications



3. Inférence bayésienne par MCMC

Résultats : prévalence

Augmentation de la prévalence finale pour les souches II et VI, diminution pour les autres

MC prevalence (%)							
Genetic Group	Initial	After cooking	Milk (ingredient)	Starch (ingredient)	After mixing with ingredients and partitioning	After pasteurization	After home cold storage
II	100	60	100	100	28	13	13
III	100	48	100	100	80	40	40
IV	100	45	100	100	100	84	84
V	100	24	100	100	5	0.1	0.1
VI	100	45	100	100	6	0.03	0.03
VII	100	64	100	100	10	1	1
Total	100	83	100	100	100	88	88
MCMC prevalence (%)							
II	100	100	100	100	80	59	59
III	100	87	100	100	65	1	1
IV	100	86	100	100	100	72	72
V	100	82	100	100	6	0.00	0.00
VI	100	93	100	100	50	15	15
VII	100	99	100	100	19	0.05	0.05
Total	100	100	100	100	100	90	90

3. Inférence bayésienne par MCMC

Retour sur résultats

- Modifications des contaminations : cohérentes avec les données (II, VI)
- Modifications des priors :

  - soit en accord avec les avis d'experts (ex : Tmin)
    → réduction d'incertitude sur les lois de certains paramètres

  - soit assez surprenantes aux yeux des experts
    → questionnement du modèle, des connaissances ou de la méthode

    Ex : Grosse réduction des durées équivalentes de cuisson
    → problème avec le modèle de temps équivalents?

Met@risk
Methods for Food Risk Analysis

INRA

Conclusion

Inférence bayésienne sur un réseau AQR :

Technique puissante permettant de faire remonter l'information apportée par les données le long de la chaîne de production

- permet de réduire l'incertitude sur les paramètres
- et d'éventuellement s'interroger sur certains aspects du modèle

Méthode ABC : n'a pas abouti – modèle trop éloigné des données ??

Méthode MCMC : fonctionne, avec des résultats cohérents

- mais convergence lente à atteindre, car forte autocorrélation
- prise en compte difficile de la variabilité entre les paquets de batchs différents, due à la structure du modèle

Met@risk
Methods for Food Risk Analysis

INRA

Merci de votre attention !

Quelques références :

Afchain, A.L., Carlin, F., Nguyen-the, C., Albert, I., 2008. Improving quantitative exposure assessment by considering genetic diversity of *B. cereus* in cooked, pasteurised and chilled foods. International Journal of Food Microbiology, 128, 165-173.

Albert, I., Grenier, E., Denis, J.B., Rousseau, J., 2008. Quantitative Risk Assessment from Farm to Fork and Beyond: a global Bayesian Approach Concerning Food-Borne Diseases. Risk Analysis, 28, 557-571.

Delignette-Muller, M.L., Cornu, M., Pouillot, R., Denis, J.B., 2006. Use of Bayesian modelling in risk assessment: application to growth of *Listeria monocytogenes* and food flora in cold-smoked salmon. International Journal of Food Microbiology,106, 195–208.

Met@risk
Methods for Food Risk Analysis

INRA