J.-M. Galharret,
A. Philippe,
N. Mercier

https://galharret.github.io/WEBSITE

# Modèle hiérarchique pour la détection de valeurs aberrantes

## Applications à la datation en archéologie

19 Décembre 2023

# Plan

# Problématique

En archéologie, quelle que soit la méthode de datation utilisée (C14,OSL,...), on est confronté au problème des outliers

- Erreur de mesure (laboratoire),
- Erreur de prélèvement (fouilles archéologiques).

Logiciels de modélisation chronologique

- OxCal : Modèle de Bronk Ramsey (2009),
- Chronomodel Lanos and Philippe (2017-18).

Notre approche :

- Identification des outliers via le modèle robuste,
- Ré-estimation du paramètre sur le sous-échantillon.

# Event Model

Modèle hiérarchique $\rightsquigarrow$ estimer l'âge d'un évènement à partir de la datation de $n$ objets le caractérisant.

- Pour le $i-$ème objet d'âge $A_i$, le laboratoire fournit la mesure $X_i$ avec une erreur $s_i$

$$X_i|A_i, s_i \sim \mathcal{N}(A_i, s_i^2).$$

- On suppose que les âges $A_1, ..., A_n$ sont contemporains de $A$

$$A_i|A, \sigma \sim \mathcal{N}(A, \sigma^2).$$


- Event model [1]

$$A_i|A, \sigma_i \sim \mathcal{N}(A, \sigma_i^2), \quad \sigma_i^2 \ i.i.d.$$

1. Lanos and Philippe (2017) Hierarchical Bayesian modeling for combining dates in archaeological context. Journal de la Société Française de Statistique, 158(2) :72(88)

# Loi de shrinkage



Loi de shrinkage uniforme pour les $\sigma_i^2$

$$\frac{s_0^2}{s_0^2 + \sigma_i^2} \underset{i.i.d}{\sim} \mathcal{U}[0,1],$$

où $s_0^{-2} = \sum_{i=1}^n s_i^{-2}$
Avantages de ce choix :

- La médiane de $\sigma_i^2$ est égale à $s_0^2$. $\rightsquigarrow$ poids identique entre les erreurs de mesure et de modèle.
- Cette loi a des moments infinis ($\rightsquigarrow$ outliers).

Inconvénient : Galharret & al. (2021) ont montré que la loi a posteriori a également des moments infinis.

# Plan

# Modèle hiérarchique

$X_1, ...X_n$ $n$ observations de vraisemblance

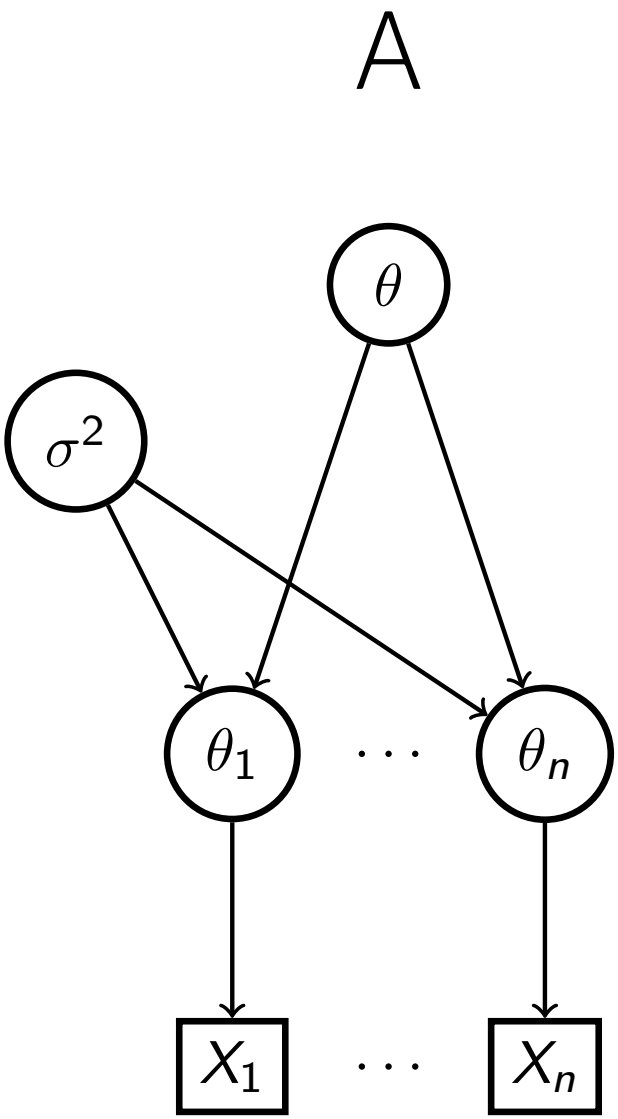
$$X_1, ..., X_n \sim p^{(n)}(\cdot \mid \theta) \tag{1}$$

Un modèle hierarchique classique (Spiegelhalter (2004), Congdon (2002) ,Gelman (2008)

$$\begin{aligned} p^{(n)}(X_1, ...X_n \mid \theta) &= \int f^{(n)}(X_1, ..., X_n, \theta_1, ..., \theta_n \mid \theta) \mathrm{d}\theta_1 ... \mathrm{d}\theta_n \\ &= \int f^{(n)}(X_1, ..., X_n \mid \theta_1, ..., \theta_n, \theta) \\ &\quad \times \pi_1(\theta_1, ..., \theta_n \mid \theta) \mathrm{d}\theta_1 ... \mathrm{d}\theta_n. \end{aligned}$$

$\rightsquigarrow$ Les variables aléatoires $|\theta_1 - \theta|, ..., |\theta_n - \theta|$ mesurent l'hétérogénéité entre $X_1, ..., X_n$.

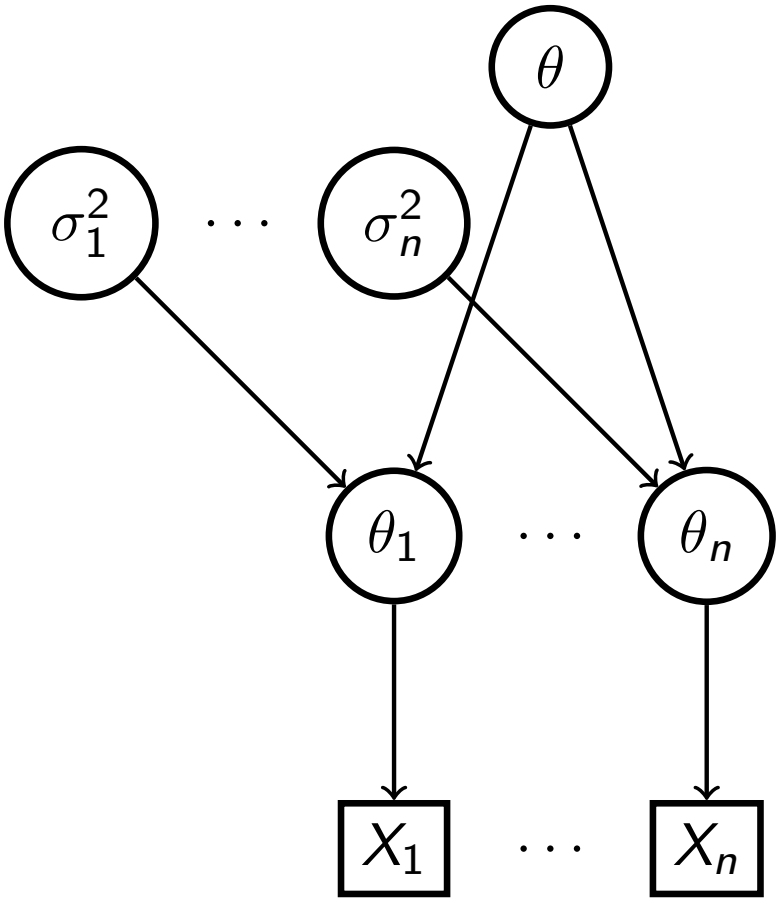
# $\theta_1, ..., \theta_n$ échangeables




$$\begin{aligned}
\pi_1(\theta_1, ..., \theta_n \mid \theta) &= \prod_{i=1}^{n} \pi_{\sigma^2}(\theta_i \mid \theta), \\
f^{(n)}(X_1, ..., X_n \mid \theta_1, ..., \theta_n, \theta) &= \prod_{i=1}^{n} f(X_i \mid \theta_i), \\
\sigma^2 &= \mathrm{Var}(\theta_i - \theta \mid \theta) \\
&= \mathrm{Var}(\theta_i \mid \theta).
\end{aligned}$$

# Outliers

B




$$\sigma_i^2 = \mathrm{Var}(\theta_i \mid \theta),$$

$$\pi_1(\theta_1, ..., \theta_n \mid \theta) = \prod_{i=1}^{n} \pi_{\sigma_i^2}(\theta_i \mid \theta).$$

# Règle de décision

Basée [2] sur la comparaison des lois a priori et a posteriori des $\sigma_i^2$
Soit $\alpha \in ]0,1[$ et soit $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi a priori de $\sigma_i^2$

$$\mathbb{P}(\sigma_i > q_{1-\alpha}) = \alpha,$$

l'observation $X_i$ est un outlier si

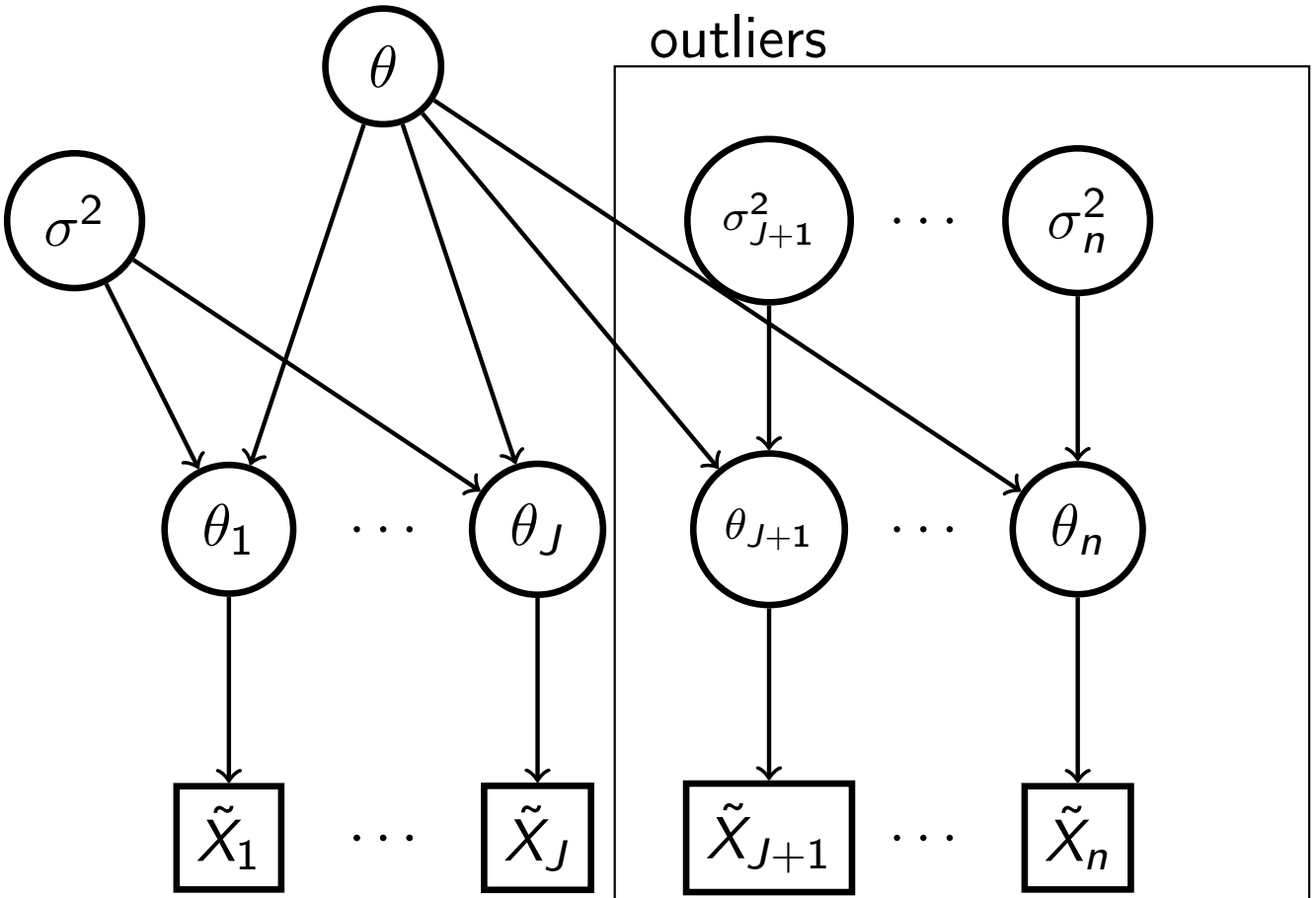
$$\mathbb{P}(\sigma_i > q_{1-\alpha} | X_1, ..., X_n) \geq \alpha,$$

Dans la suite on note $(\widetilde{X}_i)_{i\in\{1,...,J\}}$ le sous-échantillon de $(X_i)_{i\in\{1,...,n\}}$ d'où ont été exclus les outliers détectés $(\widetilde{X}_i)_{i\in\{J+1,...,n\}}$.

2. Galharret, J-M, Philippe, A., & Mercier, N. (2021). Detection of outliers with a Bayesian hierarchical model : application to the single-grain luminescence dating method. Electronic Journal Of Applied Statistical Analysis.

# Stratégies de ré-estimation

C



# Plan

# Simulations

Soit $\tau \in \{5\%, 10\%, 20\%\}$ le taux de contamination et $\mu \in \{5, 10, 20\}$ le décentrage. On simule $n$ observations $X_1, ..., X_n$ dont

- $J = [n\tau]$ vraies observations $X_1, ..., X_J$

$$X_i \sim \mathcal{N}(\theta, s_i^2),$$

- $n - J$ outliers $X_{J+1}, ..., X_n$

$$X_i \sim \mathcal{N}(\theta + \mu, s_i^2).$$

On effectue $B = 1000$ réplications.

# Exemple

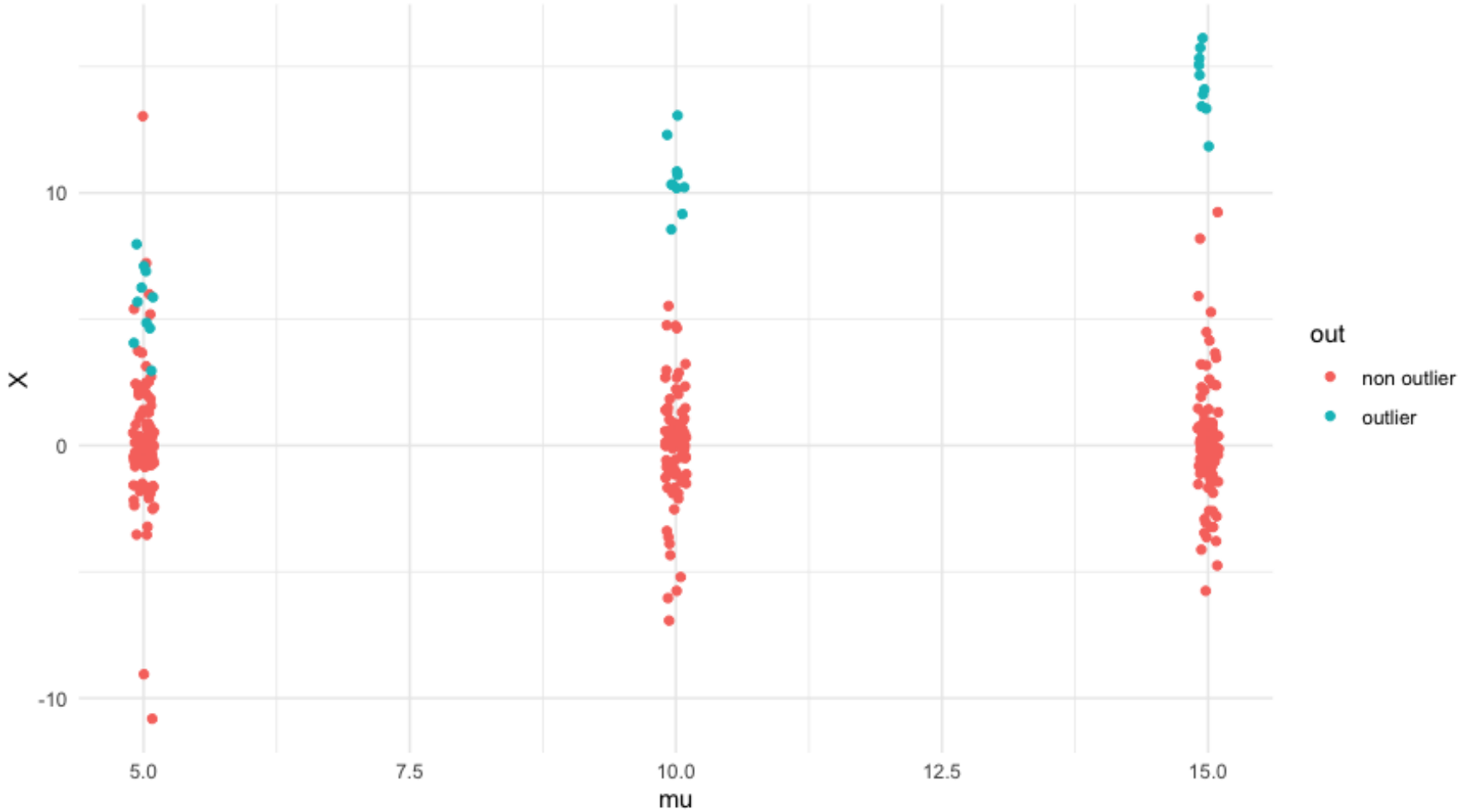



**Figure –** Exemple d'une simulation avec $\tau = 0.10$

# Calibrage de $\alpha$

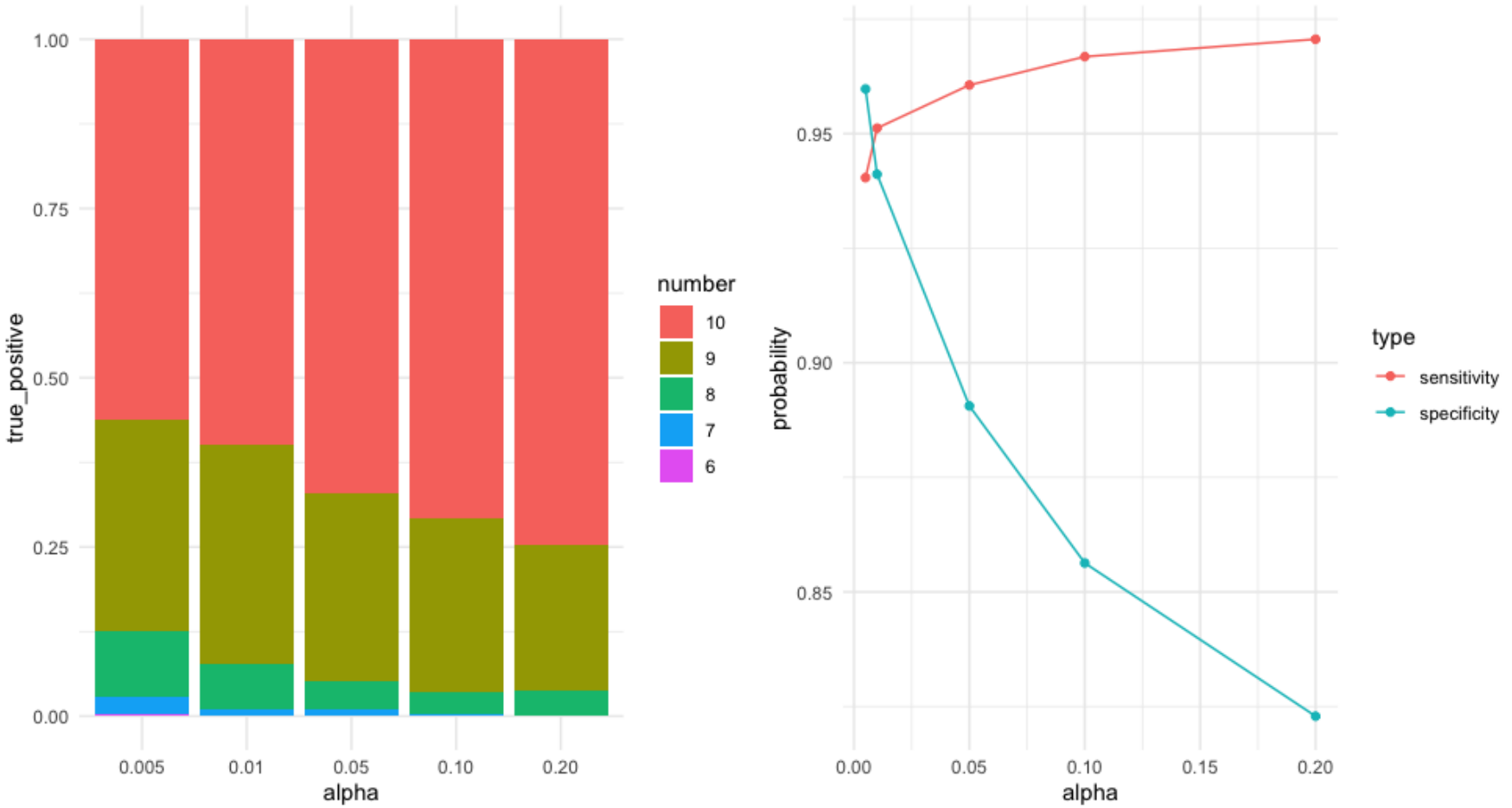



**Figure –** Variation du nombre de vrais positifs (gauche) et de la sensibilité/spécificité (droite) en fonction de $\alpha$ pour $\tau = 10\%$ et $\mu = 10$.

# Résultats des simulations

| | $\tau \backslash \mu$ | 15 | 10 | 5 |
|---|---|---|---|---|
| Sensitivity | 5% | 0.99 | 0.96 | 0.81 |
| | 10% | 0.99 | 0.96 | 0.80 |
| | 20% | 0.99 | 0.97 | 0.80 |
| Specificity | 5% | 0.89 | 0.89 | 0.89 |
| | 10% | 0.89 | 0.89 | 0.89 |
| | 20% | 0.89 | 0.89 | 0.89 |

**Table –** Estimation of the sensitivity and the specificity as function of contamination rate $\tau$ and the mean value $\mu$ of the distribution of the outliers. The cut-off is fixed to $\alpha = 0.05$, and the number of replications is $N = 1000$.

# Comparaison sous $H_0$

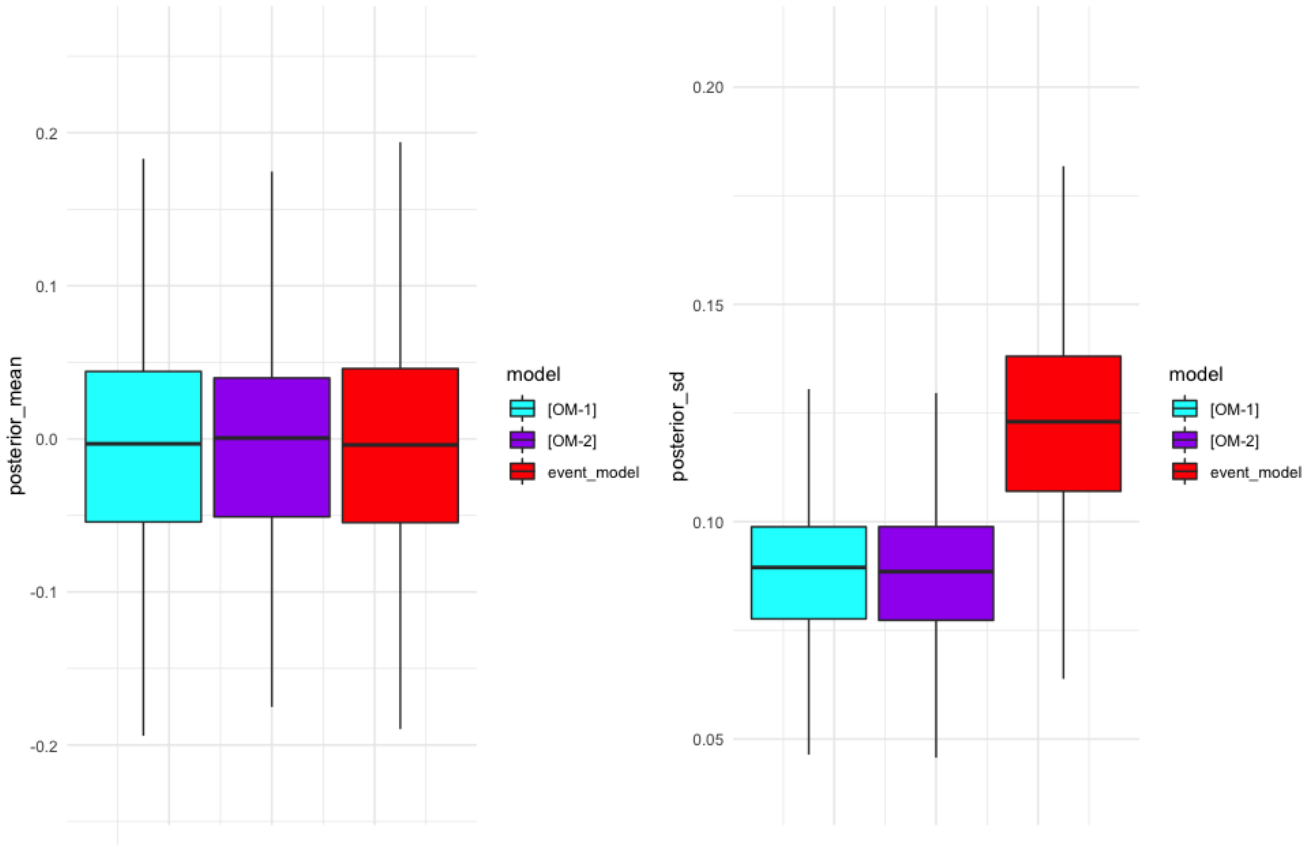
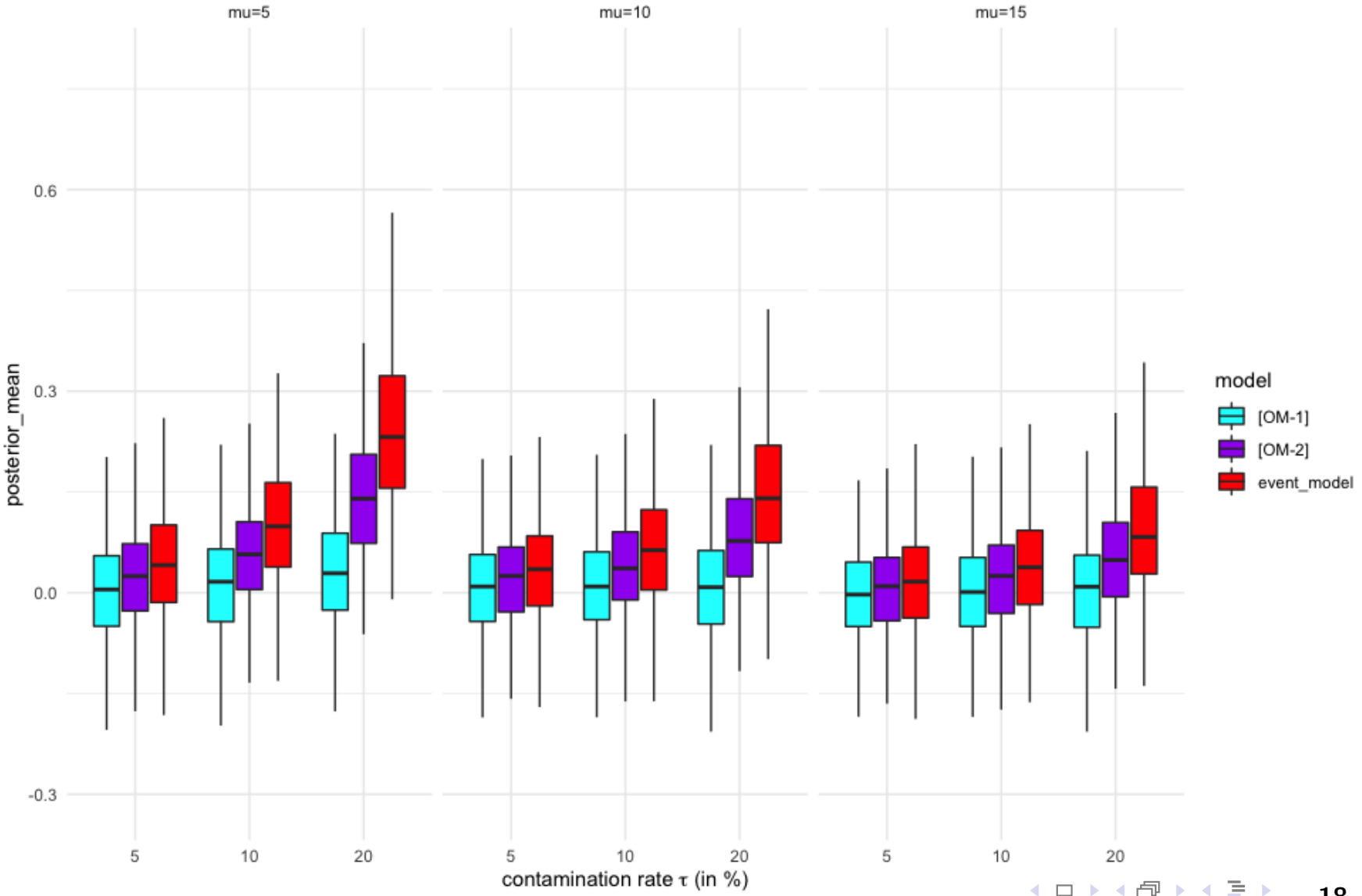



**Figure –** Comparison of the three following models : [OM-1], [OM-2] and the event model on simulated dataset without outlier ($\tau = 0$). We represent the boxplot of the mean (left) and standard deviation (Right) of the posterior distribution of $\theta$.

# Comparaison sous $H_1$



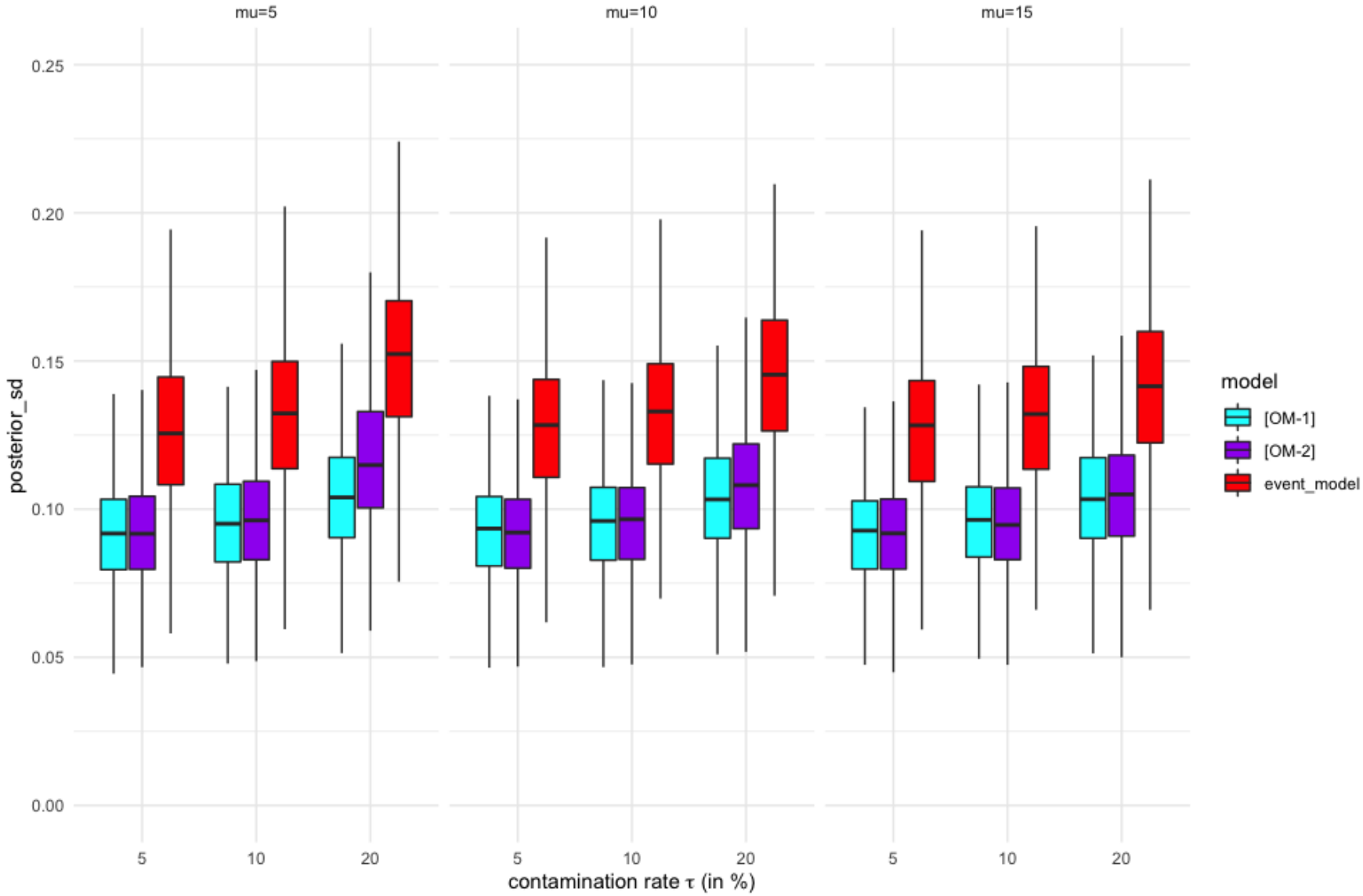
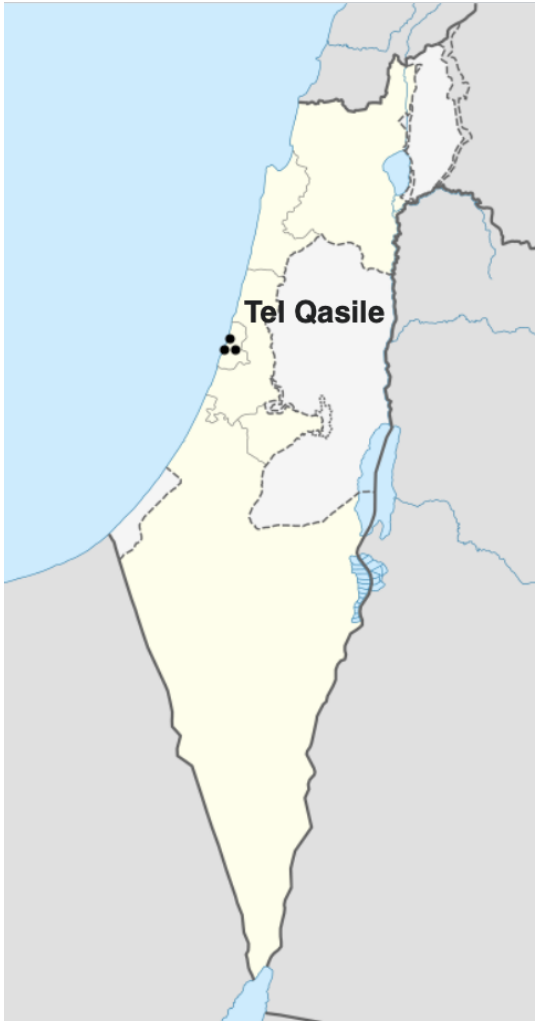


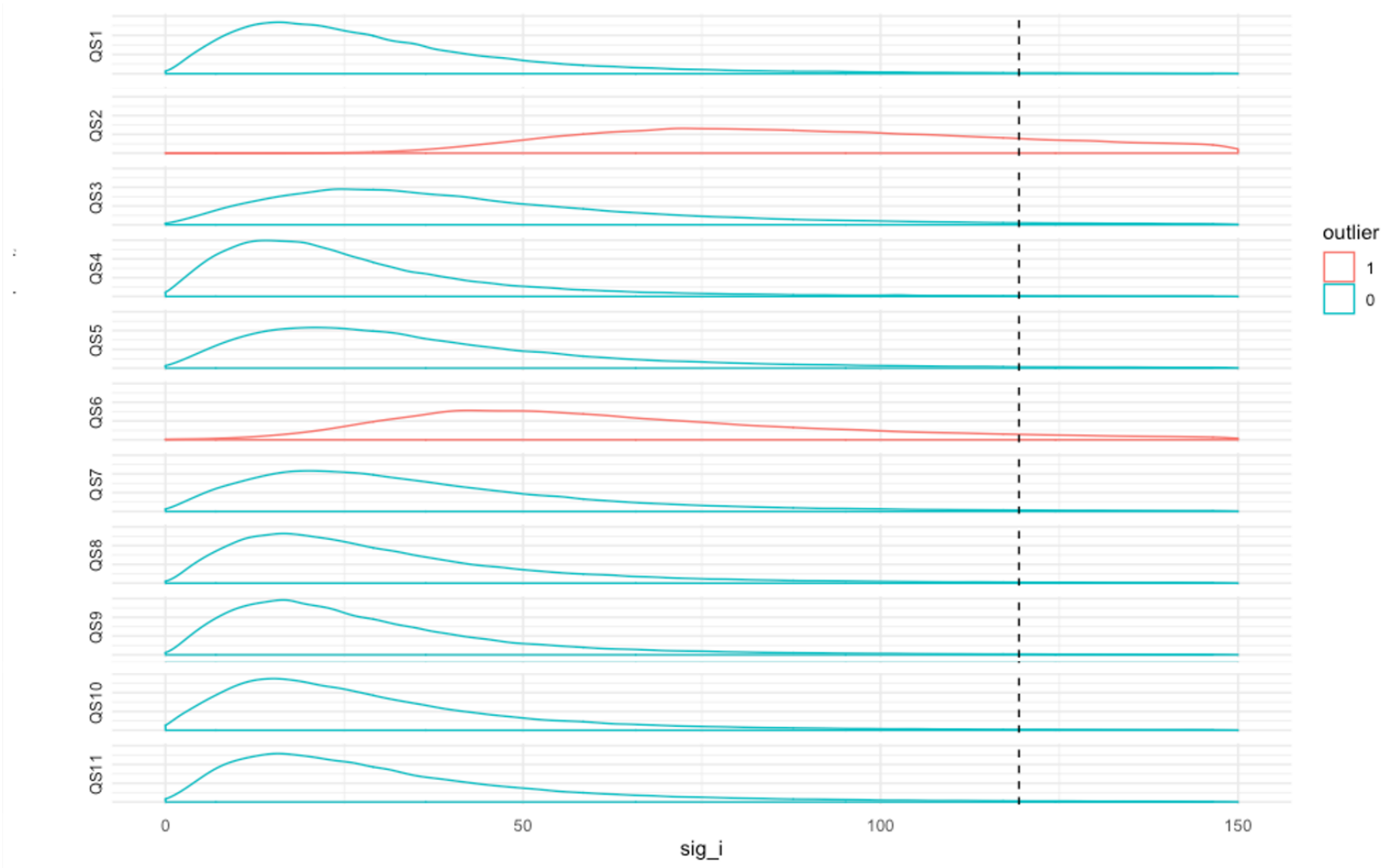

**Figure –** Comparison of the three models [OM-1], [OM-2], event model on simulated dataset with outliers for different values of contaminated rates $\tau$ and parameter $\mu$.

# Plan

# Tel Quasile

- Tel Quasile est un site archéologique situé en Israël.
- Célèbre pour avoir contribuer à la connaissance des Phillistins (installés entre le XII et X siècle avant JC)
- Bronk Ramsey a identifié les outliers dans les 12 échantillons datés par C14 en utilisant un mélange.




QS1
QS2
QS3
QS4
QS5
QS6
QS7
QS8
QS9
QS10
QS11
0
50
100
150
sig_i
outlier
1
0

# Comparaison

| method | ident | date $X_i$ | $s_i$ | $\mathbb{P}(\sigma_i > q_{.95} \mid X_1, ..., X_n)$ |
|---|---|---|---|---|
| $^{14}C$ | QS1 | 2818 | 26 | 0.017 |
| $^{14}C$ | QS2 | 2692 | 24 | **0.358** |
| $^{14}C$ | QS3 | 2911 | 26 | 0.046 |
| $^{14}C$ | QS4 | 2853 | 25 | 0.016 |
| $^{14}C$ | QS5 | 2895 | 25 | 0.030 |
| $^{14}C$ | QS6 | 2753 | 22 | **0.128** |
| $^{14}C$ | QS7 | 2800 | 25 | 0.030 |
| $^{14}C$ | QS8 | 2882 | 28 | 0.020 |
| $^{14}C$ | QS9 | 2864 | 40 | 0.015 |
| $^{14}C$ | QS10 | 2818 | 38 | 0.019 |
| $^{14}C$ | QS11 | 2897 | 44 | 0.023 |

**Table –** Dates from Tell Qasile X and outputs of the decision rule : bold values indicate radiocarbon dates detected as outliers.

# Plan

- La thermoluminescence permet de dater des matériaux archéologiques qui ont autrefois été chauffés.
- On date la dernière chauffe ou la dernière exposition à la lumière d'un échantillon contenant des minéraux de quartz.
- Méthode : [luminescence stimulée optiquement] l'énergie cumulée est alors libérée sous forme lumineuse et la quantité de lumière émise est proportionnelle au temps écoulé depuis la dernière chauffe.

$\rightsquigarrow$ Outliers proviennent de grains mal blanchis (c'est à dire qui n'ont pas rejeté toute l'énergie cumulée).

# Relation fondamentale

Dans le cadre de la datation de quartz par luminescence on estime son âge en utilisant la relation suivante

$$D \overset{\mathcal{L}}{=} A\dot{d}$$

où $\dot{d}$ est le débit de dose, $D$ est la dose équivalente absorbée et $A$ l'âge cible.

- $\dot{d}$ n'est pas observable mais on sait simuler des échantillons de $\dot{d}$ à une erreur systématique *int* prêt

$$\dot{d} = \widetilde{\dot{d}}(1 + int\dot{\varepsilon}), \quad \dot{\varepsilon} \sim \mathcal{N}(0,1).$$

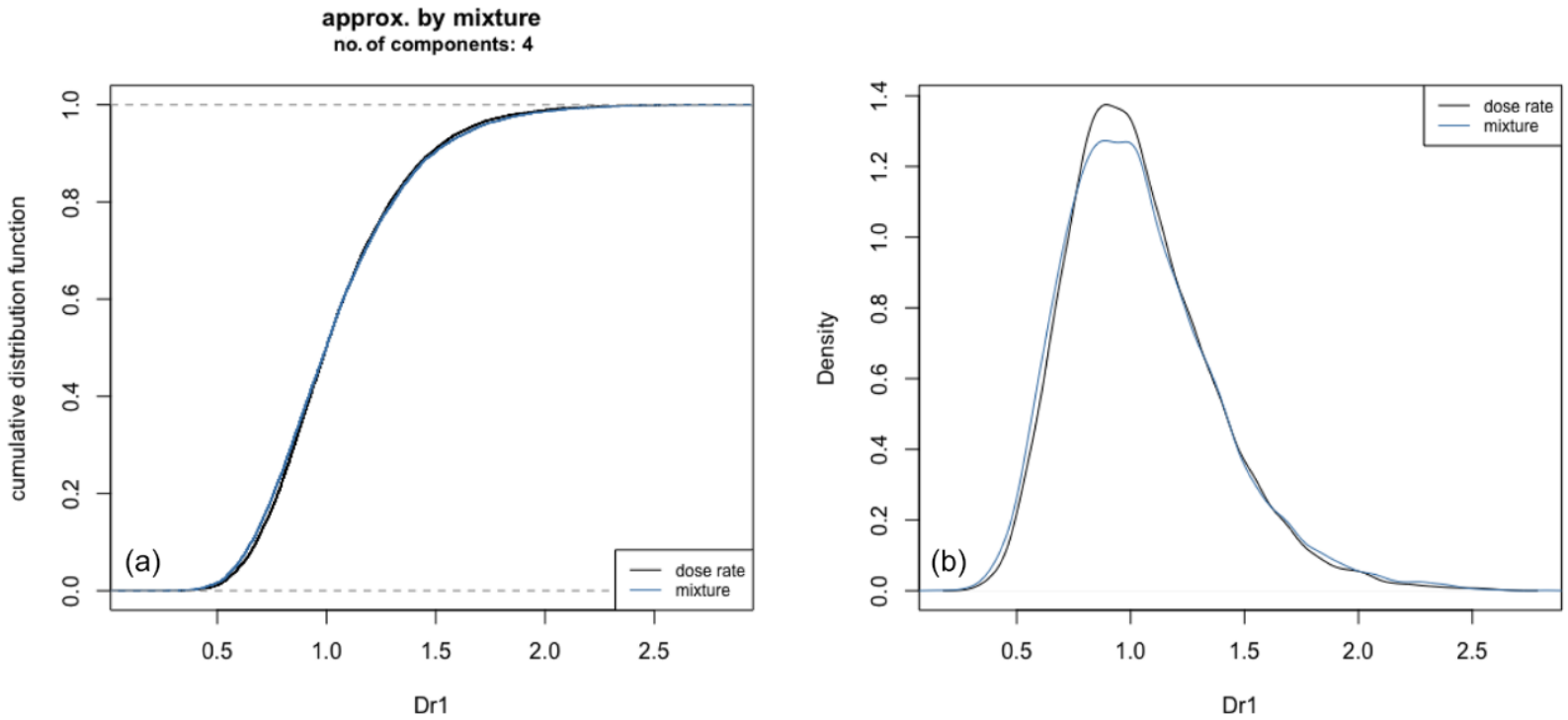
- Première étape : on ajuste $\dot{d}$ par un mélange gaussien de paramètres $(K, p_1, ..., pK, \dot{\mu}_1, ..., \dot{\mu}_K, \dot{\sigma}_1, ..., \dot{\sigma}_K)$.
- On a alors

$$D \simeq \sum_{k=1}^{K} \dot{p}_k \mathcal{N}(A\dot{\mu}_k, A^2\dot{\sigma}_k^2).$$

# Ajustement de $\dot{d}$

- Pour $K$ fixé,les paramètres du mélange sont estimés par l'algorithme EM.
- Pour le choix de $K$, on utilise $BIC$.

**Figure –** Exemple d'ajustement



# Modèlisation

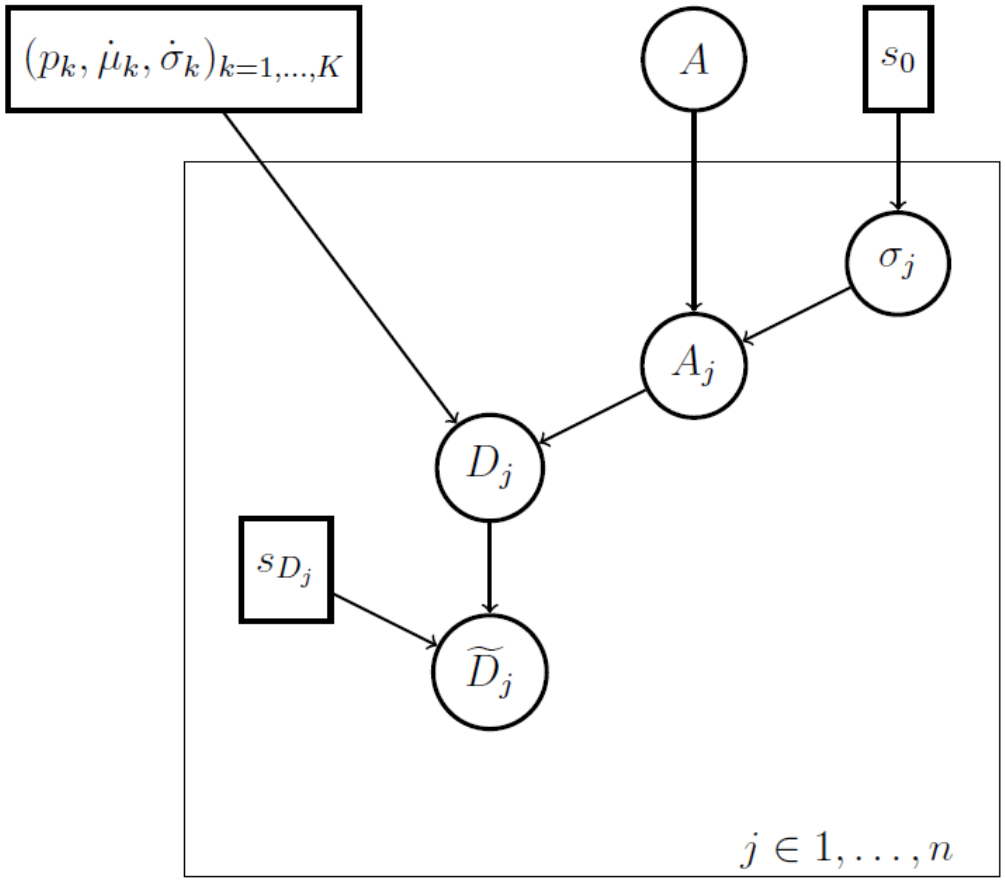
$$\tilde{D}_j \sim \mathcal{N}(D_j, s^2_{D_j})$$

$$D_j \sim \sum_{k=1}^{K} \dot{p}_k \mathcal{N}(A_j \dot{\mu}_k, A_j^2 \dot{\sigma}_k^2)$$

$$A_j \sim \mathcal{N}(A, \sigma_j^2)$$

$$\sigma_j^2 \sim \mathcal{S}(s_0^2)$$

$$A \sim \mathcal{U}niform[\underline{A}, \overline{A}]$$




Mercier N., Galharret J.-M., Tribolo C., Kreutzer S., and Philippe A. (2022) Luminescence age calculation through Bayesian convolution of equivalent dose and dose-rate distributions : the De_Dr model Geochronology, 4, 297–310, 202

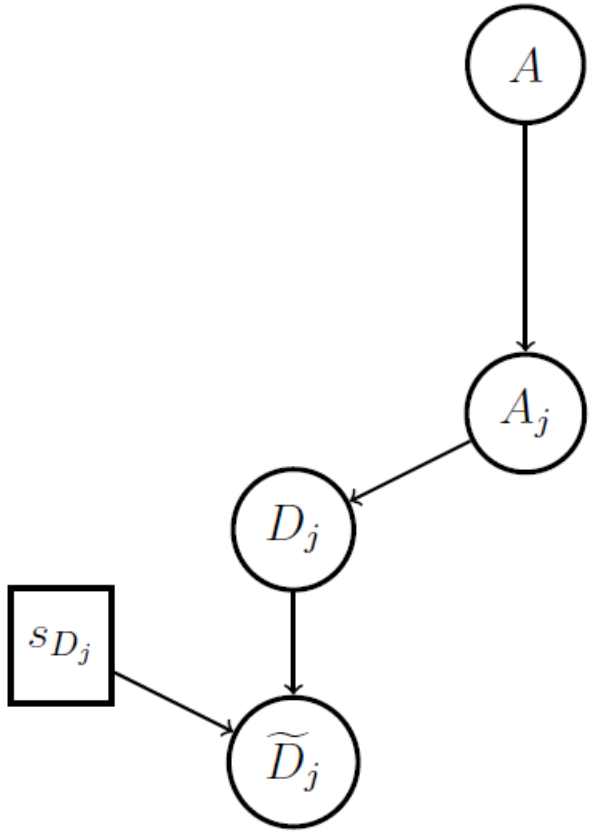
# Estimation de $s_0^2$

$$\widetilde{D}_j \sim \mathcal{N}(D_j, s^2_{D_j}),$$

$$D_j \sim \sum_{k=1}^{K} \dot{p}_k \mathcal{N}(A_j \dot{\mu}_k, A_j^2 \dot{\sigma}_k^2),$$

$$A_j \sim \mathcal{U}niform[\underline{A}, \overline{A}]$$

$s_0^2$ est estimée par la moyenne harmonique de $Var(A_j \mid \widetilde{D}_j)$



# Estimation sur données réelles

On considère un échantillon de $n = 53$ observations.

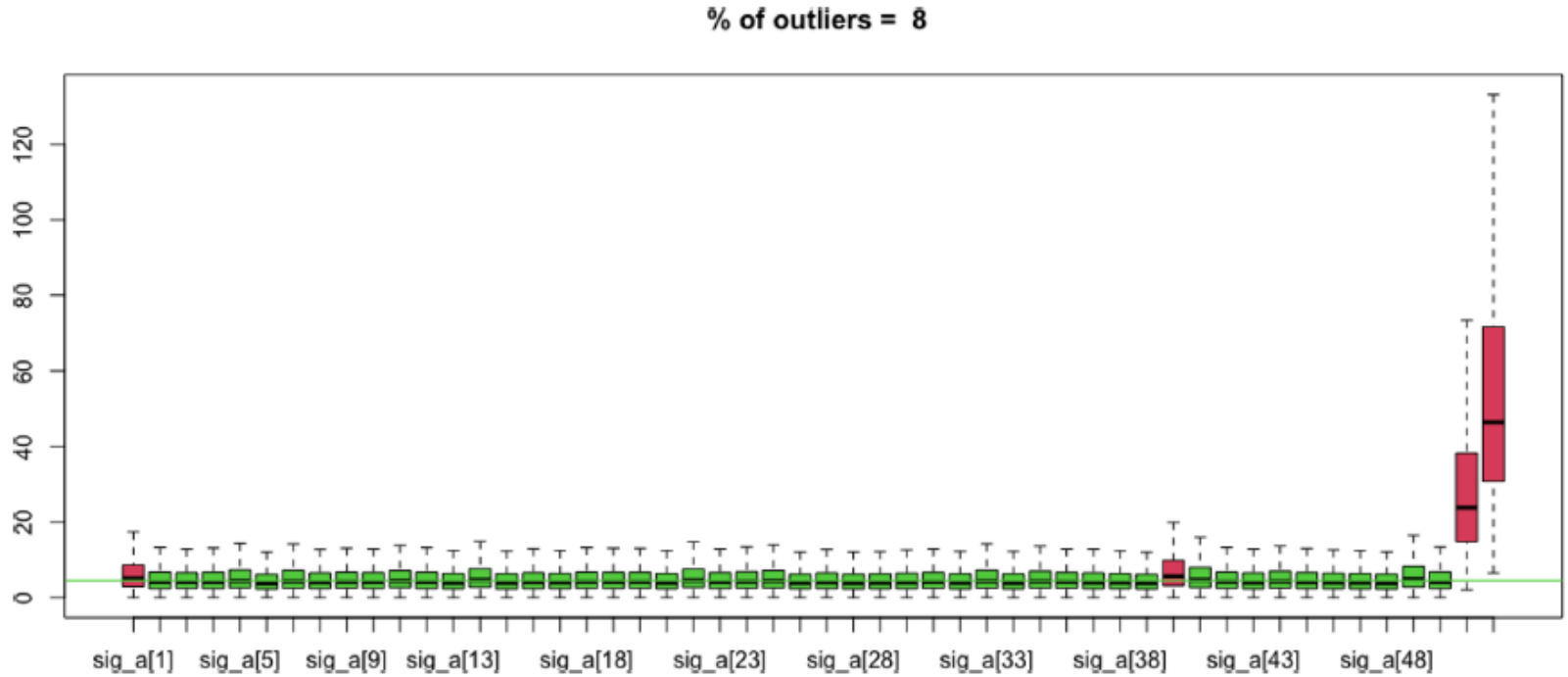



**Figure –** Boxplot des lois a posteriori des $\sigma_j$ ordonnées selon les médianes a posteriori des âges.

# Validation finale du modèle



On revient à la relation fondamentale $D \overset{\mathcal{L}}{=} A\dot{d}$ et on va comparer

- La fonction empirique des $(D_j)_{j\in J}$ (notée $F_D$)
- La fonction de répartition de $A.\dot{d}$ (notée $F_{A.\dot{d}}$)

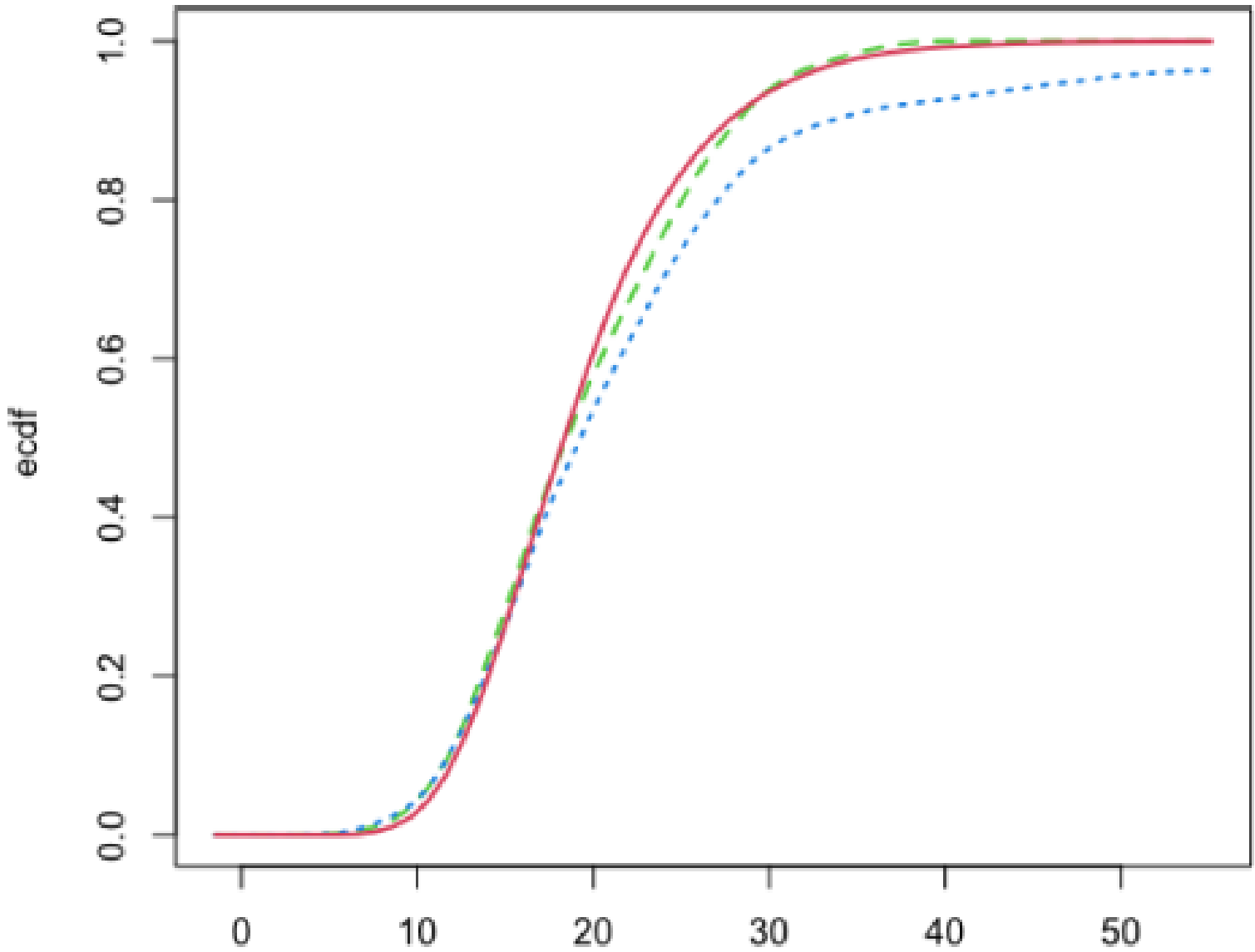
Ces deux fonctions de répartition dépendent respectivement des paramètres inconnus $(D_j)_{j\in J}$ et $A, \dot{\varepsilon}$. On calcule leurs estimateurs de Bayes

$$\mathbb{E}\left(F_D|(\widetilde{D}_j)_{j\in J}\right) = \frac{1}{J}\sum_{j\in J} F_{D_j|\widetilde{D}_j}(t)$$

$$\mathbb{E}\left(F_{A.\dot{d}}|(\widetilde{D}_j)_{j\in J}\right) = \mathbb{E}\left(\sum_{k=1}^{K} \dot{p}_k \dot{F}_k(\frac{t}{A} - \varepsilon)|(\widetilde{D}_j)_{j\in J}\right)$$

où $\dot{F}_k$ est la fonction de répartition de $\mathcal{N}(\dot{\mu}_k, \dot{\sigma}_k)$

**Figure –** Comparaison des fonctions de répartitions empiriques de $A\dot{d}$ (rouge), $D$ (bleue) et $D$ après avoir retiré les outliers détectés (vert)



# Conclusion

- La modélisation proposée dans le cadre de la datation OSL est implémentée dans le package *Luminescence*.
- La procédure pour le contexte normal-normal a été ajouté dans le logiciel Chronomodel.
- On a aussi proposé une détection des outliers dans les modèles de comptage.
- On s'intéresse actuellement à l'adaption du modèle lorsque l'on a des contrainte stratigraphiques (contraintes d'ordre).

# Merci de votre attention