Master 1 Mathématiques et Applications
U.E. ANAC++, année 2025–2026

FACULTÉ DES SCIENCES ET TECHNOLOGIES
Département Mathématiques

Université de Lille

Projet Statistique Computationnelle

Modèle hiérarchique bayésien pour l'écologie

RÉALISÉ PAR

Duc Hoang Viet NGUYEN
duc-hoang-viet.nguyen.etu@univ-lille.fr

Yvan Michel TCHONANG
yvan-michel.tchomang.etu@univ-lille.fr

Parcours Ingénierie Statistique et Numérique, Data Sciences (ISN)

21 Avril 2026

Table des matières

1 Extraction des données et prétraitement **2**
1.1 Préliminaires . . . 2
1.2 Résumé des occurrences . . . 2
1.3 Description des données . . . 2

2 Modèle hiérarchique bayésien pour l'écologie **3**
2.1 La fonction de log-vraisemblance . . . 3
2.2 Loi a priori pour μ et σ^2. . . . 4
2.2.1 Loi a priori pour μ . . . 4
2.2.2 Loi a priori pour σ^2 . . . 4
2.3 Loi a posteriori . . . 4

3 Estimation **4**
3.1 Algorithme Hamiltonian Monte-Carlo . . . 4
3.2 Application . . . 5
3.3 Résultats de l'exécution . . . 6
3.3.1 Analyse de la convergence et performances de l'algorithme . . . 6
3.3.2 Analyse de la variabilité inter-annuelle et intra-annuelle . . . 7
3.3.3 Analyser la qualité prédictive du modèle (PPC) . . . 8

4 Extension **9**
4.1 Loi a priori pour β . . . 9
4.2 Covariable environnementale X_{ij} . . . 9
4.3 Algorithme HMC - Extension . . . 9
4.4 Résultats de l'exécution . . . 9

5 Conclusion **10**

1 Extraction des données et prétraitement

1.1 Préliminaires

L'étape la plus simple était l'importation des données via l'API de GBIF. Une seule fonction était requise : la fonction `occ_search`. Elle prend en paramètres :

— le nom scientifique de l'animal (moustique tigre - Aedes albopictus).
— le pays d'intérêt (France).

Comme le dataset est trop grand et contient des variables superflues, nous allons nous concentrer sur les variables critiques telles que : `stateProvince`,`year`, `decimalLatitude` et `decimalLongitude`. La base de données extraite se présente comme illustré dans la Figure 1 (une table contenant 2209 lignes) :

stateProvince <chr>	year <int>	decimalLatitude <dbl>	decimalLongitude <dbl>
Provence-Alpes-Côte d'Azur	2025	43.64599	7.009733
Provence-Alpes-Côte d'Azur	2025	43.78476	7.511125
Languedoc-Roussillon	2025	43.85103	4.301107
Provence-Alpes-Côte d'Azur	2025	43.63114	7.126384
Languedoc-Roussillon	2025	43.52740	3.952216
Provence-Alpes-Côte d'Azur	2025	43.62485	7.031192

FIGURE 1 – La base de données extraite

Ce tableau se compose de quatre colonnes principales permettant une analyse spatiale et taxonomique :

1. `stateProvince` : pour identifier les subdivisions administratives (régions ou départements) concernées par la présence de l'espèce.
2. `year` : pour analyser l'évolution temporelle et la dynamique de propagation au fil des années.
3. `decimalLatitude` & `decimalLongitude` : les coordonnées géographiques en degrés décimaux. Par exemple les données analysées (approx. 43.6°N, 7.0°E) localisent ces observations principalement dans le sud de la France. Ils permettent une cartographie précise des points d'occurrence et effectuer des analyses spatiales.

1.2 Résumé des occurrences

Chaque ligne correspond à un enregistrement de présence du moustique tigre. Pour construire une table résumée, on a regroupé les observations par combinaison des variables `year` et `stateProvince`, en excluant toutes les lignes contenant des valeurs manquantes (`NA`). La table finale indique, pour chaque année et chaque site, le nombre d'occurrences relevées pour l'année et le site . Le résultat final contient 84 observations.

year <int>	stateProvince <chr>	n_occurrences <int>
2010	Provence-Alpes-Côte d'Azur	1
2014	Languedoc-Roussillon	2
2015	Provence-Alpes-Côte d'Azur	1
2016	Provence-Alpes-Côte d'Azur	1
2017	Languedoc-Roussillon	1
2017	Midi-Pyrénées	1

FIGURE 2 – Résumé des occurrences par année et site

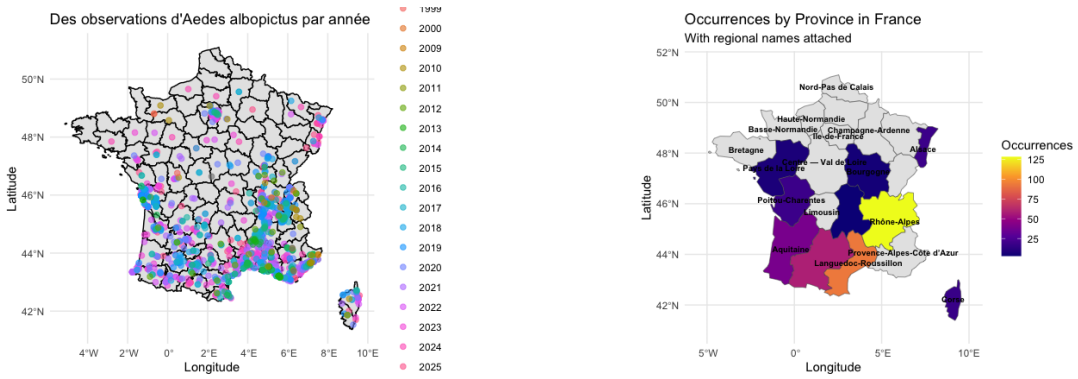
Remarque. À partir des données résumées dans la Figure2 on peut observer que, par exemple, en 2010, la région Provence-Alpes-Côte d'Azur présente 1 occurrence et en 2014, Languedoc-Roussillon enregistre 2 occurrences.

1.3 Description des données

À l'aide des packages (`sf`, `maps`, `ggplot2`, `tidyverse`, `rgbif`, `dplyr` et `giscoR`) sous R, nous avons représenté la répartition géographique des points d'observation du moustique tigre, en utilisant les coordonnées de latitude et de longitude renseignées dans la base de données. Cela nous a permis d'obtenir la Figure 3, composée de la Carte 3a et de la Carte 3b.

Carte 1 : Évolution temporelle (Par temps)

D'après la Carte 3a, on a observé que le moustique tigre semble se déplacer progressivement vers le nord, ce qui pourrait être lié aux changements climatiques.



(a) Carte 1

(b) Carte 2

FIGURE 3 – Répartition géographique des points d'observation du moustique tigre.

— **Expansion géographique :** Les points colorés représentent les observations de 1999 à 2025. Au début, les observations étaient principalement concentrées le long des côtes du Sud (rouge, orange — premières années 2000). Cependant, les points bleus, violets et roses (années récentes, 2020–2025) apparaissent de plus en plus au nord, atteignant même la région parisienne et les provinces frontalières du Nord-Est.
— **Adaptation environnementale :** L'augmentation des températures et la modification des conditions climatiques dans les régions auparavant froides ont créé un environnement favorable à la survie hivernale et à la reproduction estivale du moustique tigre.

Carte 2 : Intensité par région (Par regions)

La Carte 3b permet de quantifier la gravité de la présence du moustique tigre dans chaque région à l'aide d'une échelle de couleurs allant du violet (faible) au jaune (élevé).

Points chauds (hotspots) : La région Rhône-Alpes apparaît en jaune intense, indiquant le nombre d'occurrences le plus élevé, dépassant 125 enregistrements. Viennent ensuite les régions du sud telles que Provence-Alpes-Côte d'Azur et Languedoc-Roussillon (orange/rouge).

Contraste régional :
— Les régions du Sud et Sud-Est sont les plus touchées en raison du climat chaud caractéristique.
— Les régions de l'Ouest et du centre comme Aquitaine ou Pays de la Loire commencent également à enregistrer une présence notable (violet foncé à rouge brique).
— Les régions du Nord extrême, telles que Nord-Pas de Calais ou Haute-Normandie, restent dans les zones grises (peu ou pas de données significatives), indiquant que l'invasion est encore en cours et que la densité n'a pas atteint celle du Sud.

2 Modèle hiérarchique bayésien pour l'écologie

Pour l'année i et le site j, on note $Y_{i,j}$ le nombre d'occurrences du moustique tigre. On propose le modèle hiérarchique suivant :

$$Y_{i,j} \mid \lambda_i \sim \text{Poisson}(\lambda_i),$$
$$\log(\lambda_i) \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2).$$

On cherche à estimer μ et σ^2, et à prédire les λ_i. Ces derniers, appelés effets aléatoires, représentent la moyenne (et également la variance) annuelle des occurrences.

2.1 La fonction de log-vraisemblance

On a $Y_{ij} \mid \lambda_i \sim \text{Poisson}(\lambda_i)$. Pour une observation y_{ij} donnée, la vraisemblance est :

$$V(y_{ij} \mid \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_{ij}}}{y_{ij}!}.$$

En supposant l'indépendance conditionnelle des observations, la vraisemblance s'écrit :

$$L(y_{ij} \mid \lambda_i) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \frac{e^{-\lambda_i} \lambda_i^{y_{ij}}}{y_{ij}!}.$$

On prend le logarithme : $\log(L(y_{ij} \mid \lambda_i)) = \ell(y_{ij} \mid \lambda_i) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (-\lambda_i + y_{ij} \log \lambda_i - \log(y_{ij}!))$. En regroupant les termes

selon i, on obtient l'expression de la log-vraisemblance :

$$\ell(y_{ij} \mid \lambda_i) = \sum_{i=1}^{I} \left[-J_i\lambda_i + \sum_{j=1}^{J_i} y_{ij} \log \lambda_i - \sum_{j=1}^{J_i} \log(y_{ij}!) \right].$$

2.2 Loi a priori pour μ et σ^2.

Le choix des paramètres des lois a priori est basé sur l'analyse des caractéristiques des données observées. Le nombre de moustiques observés dans les différents sites (J_i) varie entre 1 et 252 individus. Comme le modèle utilise la fonction de lien logarithmique ($Z_i = \log \lambda_i$), nous considérons les données sur l'échelle logarithmique : $\log(1) = 0$ et $\log(252) \approx 5{,}5$.

2.2.1 Loi a priori pour μ

Nous choisissons $\mu \sim \mathcal{N}(0, 4)$ (avec un écart-type $sd = 2$). Ce choix repose sur les arguments suivants :
Avec $sd = 2$, l'intervalle à 95% de μ est bien $[-4, 4]$. En passant à l'échelle des moyennes via $\exp(\mu)$:

— borne inférieure : $e^{-4} \approx 0{,}018$ moustique,
— borne supérieure : $e^{4} \approx 54{,}6$ moustiques.

Ce choix de prior couvre des niveaux moyens allant de presque 0 à environ 55 moustiques. Même si certaines observations atteignent des valeurs extrêmes (jusqu'à 252 dans certains sites), ces valeurs correspondent à des cas isolés. Le paramètre μ représentant une moyenne globale des sites, ce choix de variance joue un rôle de régularisation (regularization), en évitant que le modèle soit trop influencé par les valeurs extrêmes.

Remarque : Cet intervalle couvre la majorité des niveaux moyens observés. Le choix de $sd = 2$ agit ainsi comme un filtre de stabilité, rendant l'estimation plus robuste aux valeurs aberrantes.

2.2.2 Loi a priori pour σ^2

Nous choisissons σ^2 suivant la loi Exp(1). Ce choix repose sur les arguments suivants :

— La loi Exp(1) est un *prior* faiblement informatif qui privilégie les petites valeurs de σ^2. Cela permet de stabiliser le modèle en évitant une dispersion excessive entre les années. Elle favorise le regroupement des estimations vers la moyenne globale μ, ce qui est crucial pour les années où les données sont plus rares (J faible).

— Bien que les données brutes varient de 1 à 252, le modèle travaille sur l'échelle log ($Z = \log(\lambda)$). L'écart entre $\log(1) = 0$ et $\log(252) \approx 5,5$ est significatif, impliquant une variance de l'ordre de quelques unités. Une loi Exp(1) sur σ^2 laisse suffisamment de masse de probabilité pour capturer cette dynamique de croissance rapide sans pour autant laisser le modèle diverger vers des solutions instables.

— Mathématiquement, la variance σ^2 doit être strictement positive. La loi exponentielle définit naturellement σ^2 sur $[0, +\infty[$. Pour optimiser l'échantillonnage via l'algorithme HMC, nous utilisons la reparamétrisation $\gamma = \log(\sigma^2)$. Cette transformation permet au paramètre d'évoluer librement sur l'ensemble des réels $\mathbb{R}$, garantissant ainsi une meilleure exploration de l'espace des phases et une convergence plus rapide de la chaîne de Markov.

2.3 Loi a posteriori

La loi a posteriori $p(\log \lambda_1, \ldots, \log \lambda_I, \mu, \sigma^2 \mid y)$, où $y = (y_{ij})_{1 \le i \le I,\, 1 \le j \le J_i}$, est donnée par :

$$p(\log \lambda_1, \ldots, \log \lambda_I, \mu, \sigma^2 \mid y) \propto p(y_{ij} \mid \lambda_i) \cdot p(z_i \mid \mu, \sigma) \cdot \pi(\mu) \cdot \pi(\sigma^2)$$

$$\propto \prod_{i=1}^{I} \left(\prod_{j=1}^{J_i} \frac{(\lambda_i)^{y_{ij}} e^{-\lambda_i}}{y_{ij}!} \right) \cdot \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z_i-\mu)^2} \cdot \frac{1}{\sqrt{8\pi}} e^{-\frac{\mu^2}{8}} \cdot e^{-\sigma^2} \mathbb{1}_{\mathbb{R}_+}(\sigma^2)$$

$$\text{Comme } \lambda_i = e^{z_i} \Rightarrow \prod_{i=1}^{I} \left(\prod_{j=1}^{J_i} \frac{(e^{z_i})^{y_{ij}} e^{-e^{z_i}}}{y_{ij}!} \right) \cdot \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z_i-\mu)^2} \cdot \frac{1}{\sqrt{8\pi}} e^{-\frac{\mu^2}{8}} \cdot e^{-\sigma^2} \mathbb{1}_{\mathbb{R}_+}(\sigma^2)$$

$$\Rightarrow \log p(\log \lambda_1 \ldots \log \lambda_{12}, \mu, \sigma^2 \mid y) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} [y_{ij} z_i - e^{z_i} - \log(y_{ij}!)] + \sum_{i=1}^{I} \left[-\frac{(z_i-\mu)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \right] + \left(-\frac{\mu^2}{8} - \frac{1}{2}\log(8\pi) \right) - \sigma^2 (\sigma^2 > 0)$$

3 Estimation

3.1 Algorithme Hamiltonian Monte-Carlo

Algorithme HMC et méthode leapfrog L'idée générale de l'algorithme *Hamiltonian Monte Carlo (HMC)* introduite par Duane, Kennedy, Pendleton et Roweth en 1987, est de construire une dynamique auxiliaire permettant d'explorer la distribution cible $\pi(\theta)$. En partant d'une position initiale θ_0, on introduit une variable de moment r et l'Hamiltonien :

$$H(\theta, r) = -\log p(\theta) + K(r),$$

où $K(r) = \frac{1}{2} r^T M^{-1} r = \frac{1}{2} r^\top r = \frac{1}{2} \sum r_i^2$ est l'énergie cinétique associée au moment. ($M = I$)

Principe général. L'algorithme consiste à :

1. échantillonner un moment initial $r_0 \sim \mathcal{N}(0, M^{-1})$; $(M = I)$, ce qui simplifie l'expression de l'énergie cinétique.
2. simuler la dynamique hamiltonienne pendant un temps T ;
3. accepter ou rejeter le nouvel état selon une probabilité d'acceptation de type Metropolis.

La dynamique étant analytiquement insoluble, nous utilisons la méthode *Leapfrog* décrite par Neal (2011) et Chen (2019). On note ε le pas de discrétisation et L le nombre d'itérations, avec $\varepsilon L = T$.

1. Initialisation :
$$\theta^{(0)} = \theta_0, \quad r_0 \sim \mathcal{N}(0, M^{-1}).$$
2. Demi-pas sur le moment :
$$r^{(0)} = r_0 - \frac{\varepsilon}{2} \nabla \log p(\theta^{(0)}).$$
3. Pour $\ell = 1, \ldots, L-1$:
   (a) Mise à jour de la position :
   $$\theta^{(\ell)} = \theta^{(\ell-1)} + \varepsilon\, r^{(\ell-1)}.$$
   (b) Mise à jour du moment :
   $$r^{(\ell)} = r^{(\ell-1)} - \varepsilon \nabla \log p(\theta^{(\ell)}).$$
4. Mise à jour finale de la position :
$$\theta^{(L)} = \theta^{(L-1)} + \varepsilon\, r^{(L-1)}.$$
5. Demi-pas final sur le moment :
$$r^{(L)} = r^{(L-1)} - \frac{\varepsilon}{2} \nabla \log p(\theta^{(L)}).$$
6. Probabilité d'acceptation :
$$a = \min\left(1, \frac{\exp(-H(\theta^{(L)}, r^{(L)}))}{\exp(-H(\theta_0, r_0))}\right).$$
7. Acceptation :
$$\theta_{\text{new}} = \begin{cases} \theta^{(L)} & \text{avec probabilité } a \\ \theta_0 & \text{sinon} \end{cases}$$
8. Retourner θ_{new}.

3.2 Application

Dans l'algorithme HMC, la nouvelle écriture du vecteur de paramètres est :

$$\theta = (Z_1, \ldots, Z_I, \mu, \gamma) \quad \text{et} \quad r = (r_{Z_1}, \ldots, r_{Z_I}, r_\mu, r_\gamma).$$

où l'on pose $\gamma = \log(\sigma^2)$, ce qui équivaut à $\sigma^2 = e^\gamma$. Cette reparamétrisation garantit automatiquement la contrainte de positivité sur la variance ($\sigma^2 > 0$) tout en permettant à γ d'évoluer librement dans $\mathbb{R}$.

Par changement de variable, la densité a priori de γ, sachant que $\sigma^2 \sim \mathcal{E}(1)$, est obtenue via $\pi(\gamma) = \pi(\sigma^2) \left|\frac{d(\sigma^2)}{d\gamma}\right|$. Avec $\frac{d(\sigma^2)}{d\gamma} = e^\gamma$, on a :

$$\pi(\gamma) = e^{-e^\gamma} \cdot e^\gamma \implies \log \pi(\gamma) = -e^\gamma + \gamma$$

L'expression de la log-densité a posteriori devient alors (à une constante additive près) :

$$\log p(Z_1 \ldots Z_I, \mu, \gamma \mid y) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left[y_{ij} Z_i - e^{Z_i} - \log(y_{ij}!)\right] + \sum_{i=1}^{I} \left[-\frac{(Z_i - \mu)^2}{2e^\gamma} - \frac{1}{2}\gamma - \frac{1}{2}\log(2\pi)\right] - \frac{\mu^2}{8} - \frac{1}{2}\log(8\pi) - e^\gamma + \gamma$$

Ensuite, il faut calculer le gradient de la fonction d'évaluation $U(\theta) = -\log p(\theta \mid y)$, correspondant à l'opposé de la log-densité a posteriori. Ci-dessous, les dérivées partielles utilisées pour simuler les trajectoires de l'algorithme HMC :

1. Dérivée par rapport à z_i (Pour chaque i) :

$$\frac{\partial U}{\partial Z_i} = \frac{\partial}{\partial Z_i}\left[\sum_{j=1}^{J_i}(e^{Z_i} - y_{ij} Z_i + \log(y_{ij}!)) + \frac{(Z_i - \mu)^2}{2e^\gamma}\right] = J_i e^{Z_i} - \sum_{j=1}^{J_i} y_{ij} + \frac{Z_i - \mu}{e^\gamma}$$

2. Dérivée par rapport à μ :

$$\frac{\partial U}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{I} \frac{(Z_i - \mu)^2}{2e^\gamma} + \frac{\mu^2}{8}\right) = \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{I} \frac{Z_i^2 - 2Z_i\mu + \mu^2}{2e^\gamma} + \frac{\mu^2}{8}\right) = \sum_{i=1}^{I} \frac{-Z_i + \mu}{e^\gamma} + \frac{\mu}{4}$$

3. Dérivée par rapport à γ :

$$\frac{\partial U}{\partial \gamma} = \frac{\partial}{\partial \gamma}\left(\sum_{i=1}^{I}\left[\frac{(Z_i - \mu)^2}{2e^\gamma} + \frac{\gamma}{2}\right] + e^\gamma - \gamma\right) = -\sum_{i=1}^{I} \frac{(Z_i - \mu)^2}{2e^\gamma} + \frac{I}{2} + e^\gamma - 1$$

Expression finale du gradient Le gradient ∇U, évalué au point $\theta = (Z_1, \ldots, Z_I, \mu, \gamma)$, s'écrit formellement sous forme vectorielle à l'aide des dérivées partielles calculées précédemment :

$$\nabla U(\theta) = \left(\frac{\partial U}{\partial Z_1}, \ldots, \frac{\partial U}{\partial Z_I}, \frac{\partial U}{\partial \mu}, \frac{\partial U}{\partial \gamma} \right)^\top$$

Ce trio fondamental — le vecteur de paramètres θ, la fonction d'évaluation $U(\theta)$ (la log-densité a posteriori négative) et son gradient $\nabla U(\theta)$ — rassemble toute l'information géométrique locale de la distribution. Il va être utilisé pour paramétrer et exécuter l'algorithme HMC, nous permettant de diriger les propositions d'états vers les zones de forte densité et d'échantillonner efficacement selon la loi a posteriori de notre modèle.

3.3 Résultats de l'exécution

Dans cette section, nous présentons et analysons les échantillons obtenus suite à l'exécution de l'algorithme HMC. Nous évaluerons la convergence des chaînes de Markov et interpréterons les estimations a posteriori des paramètres μ, σ^2 (via γ) et des effets aléatoires spatiaux Z_i.

3.3.1 Analyse de la convergence et performances de l'algorithme

La figure 4 présente les *traceplots* des paramètres clés (μ, γ et certains Z_i). On observe que les chaînes se mélangent de manière homogène et gravitent autour d'une valeur stationnaire après la période de *burn-in*. Cette allure en traceplot est un indicateur fort de la convergence de l'algorithme HMC vers la distribution cible.

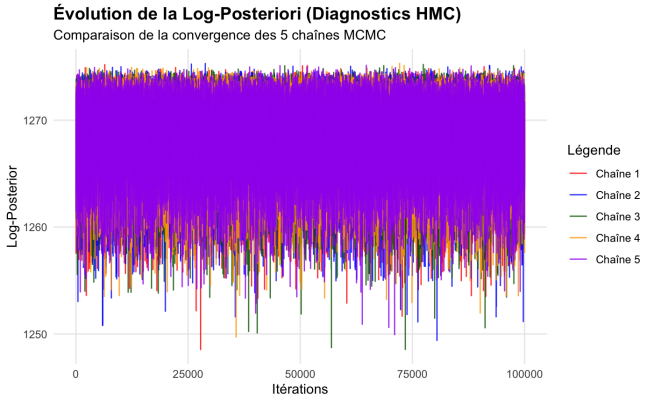


FIGURE 4 – Traceplots des paramètres du modèle montrant une bonne convergence des chaînes.

La figure ci-dessus présente l'évolution de la log-postériore pour les 5 chaînes simulées sur 100000 itérations.

- **Convergence rapide :** On observe que toutes les chaînes convergent quasi-immédiatement vers la même zone de haute densité (entre 1260 et 1275).
- **Mélange des chaînes :** La superposition parfaite des différentes couleurs indique un excellent mélange. L'algorithme explore l'espace des paramètres de manière consistante sans rester bloqué dans des optima locaux.
- **Stabilité :** La stationnarité du signal confirme que la période de chauffe (*burn-in*) de 1000 itérations est largement suffisante pour garantir la fiabilité des estimations postérieures présentées dans la suite du rapport.

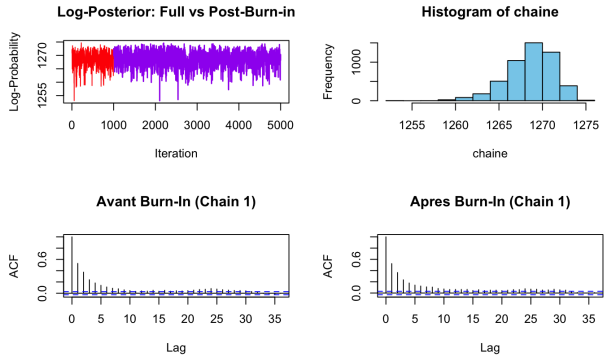


FIGURE 5 – Diagnostics de convergence de la chaîne 1.

Pour évaluer l'indépendance des échantillons générés, nous avons analysé la fonction d'autocorrélation (ACF) de la log-postériore. La Figure 5 illustre la stabilisation de la log-postérieure và la décroissance rapide de l'ACF pour une chaîne dans 5000 itérations.

— **Indépendance temporelle :** Les graphiques ACF montrent une décroissance rapide de la corrélation entre les itérations successives. Après un *lag* d'environ 15, l'autocorrélation devient négligeable (proche de zéro), ce qui indique que les échantillons conservés pour l'inférence sont quasi-indépendants.
— **Efficacité du HMC :** Cette chute rapide de l'ACF est une caractéristique propre à l'algorithme Hamiltonian Monte Carlo, qui surmonte les problèmes de "marche aléatoire" rencontrés par d'autres méthodes comme Metropolis-Hastings.
— **Validation du Burn-in :** La comparaison entre les phases *Avant* et *Après Burn-In* confirme que la chaîne s'est stabilisée. Les échantillons utilisés pour les résultats finaux proviennent d'une distribution stationnaire où l'information est maximisée.

3.3.2 Analyse de la variabilité inter-annuelle et intra-annuelle

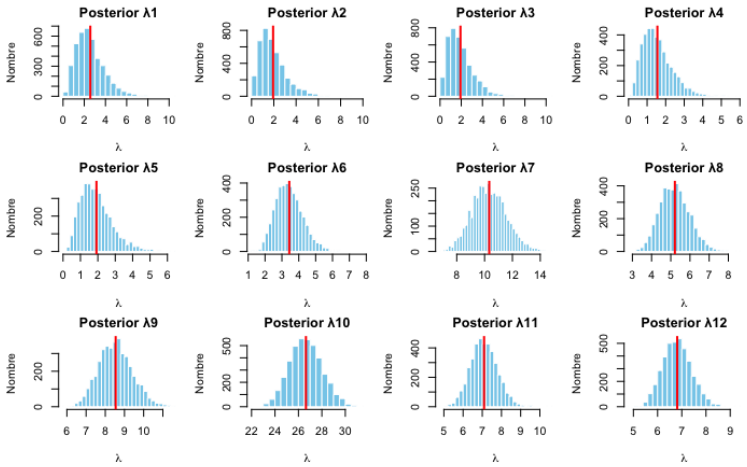


FIGURE 6 – Distributions postérieures des intensités de présence λ_1 à λ_{12}.

Analyse des résultats : Les distributions postérieures des intensités de présence λ_1 à λ_{12} sont représentées sur la Figure 6. On observe une progression nette des valeurs de λ au fil des années.

— **Variabilité inter-annuelle :** On observe une progression nette des valeurs de λ. La dynamique passe d'une faible intensité initiale ($\lambda_1 \approx 2,5$) à une explosion démographique vers l'année 10 ($\lambda_{10} \approx 26,5$). Cette variation d'une année à l'autre valide l'importance de la structure hiérarchique du modèle.
— **Précision des estimations :** La largeur des bases des histogrammes reflète l'incertitude. Plus l'année avance, plus les distributions deviennent "pincées" (notamment λ_{10}), témoignant d'une réduction de la variance postérieure grâce à des données plus abondantes.

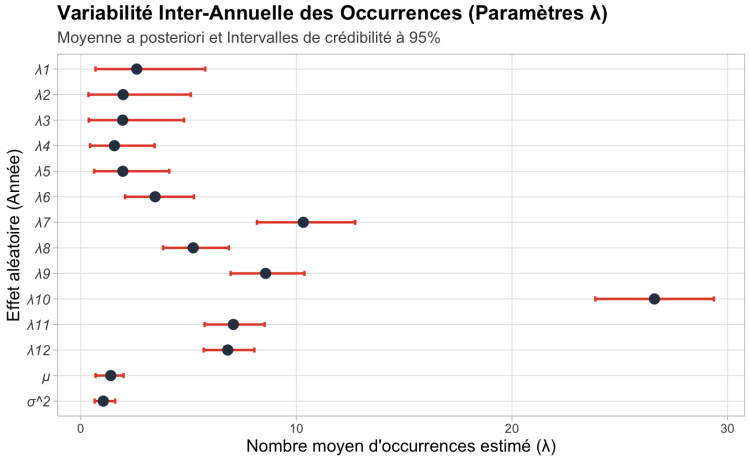


FIGURE 7 – Visualisation graphique (Forest Plot)

L'analyse des résultats numériques et graphiques (Figure 7) permet de tirer les conclusions suivantes sur la précision du modèle :

— **Estimation Ponctuelle (Moyenne a posteriori) :** L'estimation de chaque λ_i représente l'intensité moyenne d'occurrence pour une année donnée. On observe une transition majeure entre la première phase (années 1 à 6) où $\lambda < 5$ et la seconde phase (années 7 à 10) où l'intensité augmente drastiquement.

Remarque particulière : L'année 10 se distingue nettement avec une estimation moyenne très élevée ($\hat{\lambda}_{10} = 26,641$). Cette valeur est fortement influencée par la présence d'un site exceptionnel ayant enregistré un record de **252 moustiques**. Le modèle Bayesien hiérarchique parvient à intégrer cette valeur extrême tout en stabilisant l'estimation globale grâce à la structure de groupe.

— **Calcul de l'Intervalle de Crédibilité (IC à 95%) :** L'intervalle de crédibilité permet de quantifier l'incertitude :
  — *Incertitude élevée :* Pour λ_1, l'intervalle est large $[0,677\,;\,5,810]$, reflétant des données limitées.
  — *Précision accrue :* Pour l'année 10, malgré la présence de la valeur extrême (252), l'intervalle $[24,046\,;\,29,494]$ reste relativement étroit. Cela démontre que l'abondance des observations cette année-là permet de valider la réalité de cette explosion démographique avec une grande confiance statistique.

3.3.3 Analyser la qualité prédictive du modèle (PPC)

Afin d'évaluer la capacité du modèle à refléter la réalité des observations, nous avons effectué une vérification prédictive *a posteriori* (Posterior Predictive Check - PPC). La Figure 8 compare les statistiques issues des données observées (ligne rouge) aux distributions simulées par le modèle (histogrammes gris).

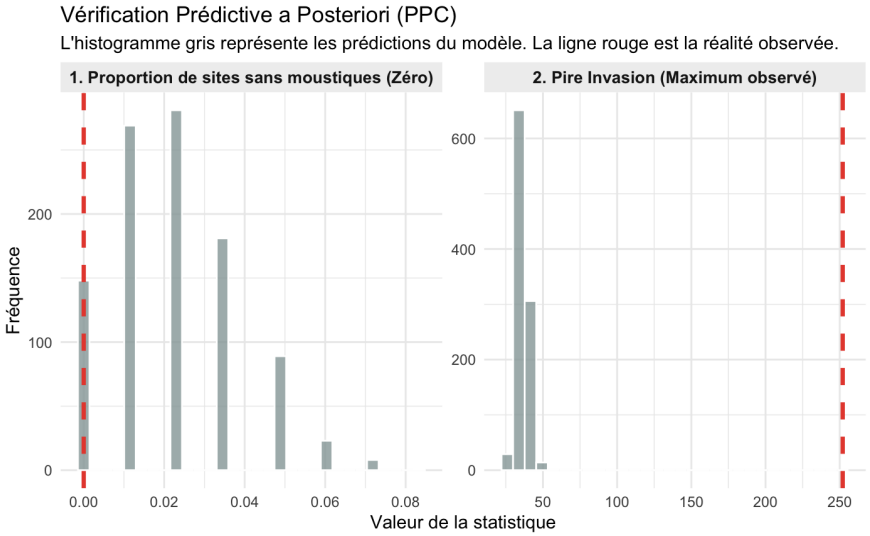


FIGURE 8 – Vérification Prédictive a Posteriori (PPC) pour la proportion de zéros et le maximum observé.

Interprétation des résultats :

— **Proportion de sites sans moustiques (Zéros) :** La ligne rouge représentant les données réelles se situe à l'extrémité gauche de la distribution simulée. Cela indique que notre modèle a tendance à *surestimer* légèrement la probabilité d'absence de moustiques. En d'autres termes, le modèle prédit un nombre de "zéros" un peu plus élevé que ce qui a été observé en réalité. Ce phénomène peut s'expliquer par les propriétés de la loi de Poisson qui, bien que robuste, ne capture pas parfaitement l'absence totale de moustiques sur certains sites lorsque les données sont très dispersées.

— **Pire invasion (Maximum Observé) :** Ce graphique illustre la limite structurelle du modèle actuel. La valeur réelle observée (indiquée par la ligne rouge à **252**) est très éloignée de la masse des simulations, lesquelles se concentrent autour de valeurs comprises entre 40 et 50.
  — *Analyse :* Le modèle sous-estime systématiquement les événements extrêmes. Une observation de 252 moustiques est considérée par le modèle comme extrêmement improbable sous l'hypothèse d'une loi de Poisson standard.
  — *Justification :* Cette divergence souligne que l'année 10 a été marquée par un événement exceptionnel et atypique qui dépasse la variabilité moyenne capturée par la structure hiérarchique.

Conclusion sur la qualité : Malgré une sous-estimation du maximum, le modèle s'avère très performant pour capturer la tendance globale et la variabilité inter-annuelle de la colonisation.

4 Extension

On pourra considérer l'extension suivante pour expliquer une partie de la variabilité entre années :

$$\begin{aligned} y_{ij} \mid \lambda_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \log \lambda_{ij} &= \alpha_j + \beta x_{ij} \\ \alpha_j \mid \mu, \sigma &\sim \mathcal{N}(\mu, \sigma^2), \end{aligned}$$

où x_{ij} est une covariable environnementale et β un paramètre supplémentaire à estimer.

4.1 Loi a priori pour β

Nous avons choisi une loi a priori pour β suivant une $\mathcal{N}(0, 10)$. Il s'agit d'un prior faiblement informatif. Ce choix se justifie par le fait que β est multiplié par la covariable X_{ij} (standardisée). Comme l'ampleur de l'effet de la température sur la dynamique des moustiques n'est pas connue a priori, cette variance de 10 offre suffisamment de souplesse au modèle pour que les données dictent l'estimation de β, tout en agissant comme une régularisation légère pour éviter des valeurs numériquement instables dans la fonction de lien log.

4.2 Covariable environnementale X_{ij}

Nous avons choisi la température à 2 mètres (T2M) comme covariable environnementale majeure, car elle constitue un facteur biologique déterminant pour le cycle de vie du moustique tigre (*Aedes albopictus*). Afin de garantir une précision spatiale optimale, nous avons exploité les coordonnées géographiques (latitude et longitude) présentes dans le jeu de données initial. En utilisant la librairie `nasapower` sous R, nous avons interfacé le modèle avec la base de données NASA POWER pour extraire les températures spécifiques à chaque province (*stateProvince*).

4.3 Algorithme HMC - Extension

La loi a posteriori $p(z_1, \ldots, z_J, \beta, \mu, \sigma^2 \mid y)$, où $y = (y_{ij})_{1 \le j \le J_i, 1 \le i \le I}$, est donnée par :

$$p(z_1, \ldots, z_J, \beta, \mu, \sigma^2 \mid y) \propto p(y_{ij} \mid \alpha_j, \beta) \cdot p(\alpha_j \mid \mu, \sigma) \cdot \pi(\beta) \cdot \pi(\mu) \cdot \pi(\sigma^2) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \frac{(e^{\alpha_j + \beta X_{ij}}) y_{ij} e^{-e^{\alpha_j + \beta X_{ij}}}}{y_{ij}!} \cdot \ldots \cdot \frac{1}{\sqrt{20\pi}} e^{-\frac{\beta^2}{20}} \cdot \ldots$$

Pour implémenter l'algorithme HMC, nous avons besoin du gradient du log-posterior par rapport au nouveau paramètre β :

$$\frac{\partial \log p(z, \beta, \mu, \sigma^2 \mid y)}{\partial \beta} = \sum_{j=1}^{J_i} \sum_{i=1}^{I} X_{ij} \left(y_{ij} - e^{\alpha_j + \beta X_{ij}} \right) - \frac{\beta}{10}$$

4.4 Résultats de l'exécution

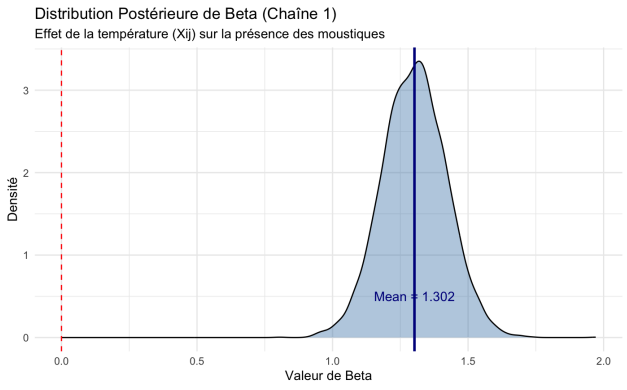


FIGURE 9 – Distribution postérieure du paramètre β (Chaîne 1)

La figure 9 présente la distribution postérieure du coefficient β associé à la température X_{ij}. On observe que la totalité de la masse de la distribution est située largement au-dessus de zéro, avec une valeur moyenne de $\hat{\beta} \approx 1.302$. L'absence de la valeur 0 dans l'intervalle de crédibilité confirme que la température a un effet positif significatif sur la présence des moustiques tigres. Plus précisément, une augmentation de la température normalisée entraîne une croissance exponentielle du taux de signalement λ_{ij}, validant ainsi l'apport crucial de cette covariable environnementale dans le modèle.

Le tableau suivant résume l'évolution des paramètres globaux entre le modèle de base et le modèle étendu :

Paramètre (Moyenne)	Modèle de base	Modèle extension
μ (Moyenne globale)	1.391	1.321
σ^2 (Variance)	1.042	1.336

TABLE 1 – Évolution des paramètres après l'introduction de la température.

Interprétation des résultats :

L'apport de la covariable température se manifeste par deux changements clés :

1. **Baisse de μ (1.391 $\rightarrow$ 1.321) :** Une partie de l'abondance moyenne est désormais expliquée par la température. Le modèle est plus précis car il ne repose plus uniquement sur une constante globale, mais sur un facteur climatique endogène.
2. **Hausse de σ^2 (1.042 $\rightarrow$ 1.336) :** Cependant, elle s'explique par le passage d'une structure de données temporelle (12 années) à une structure plus fine (23 sites sur 12 ans). L'introduction de la température capte l'effet climatique global, mais révèle parallèlement une forte hétérogénéité spatiale entre les départements français qui n'était pas visible dans le modèle simplifié..

Ces résultats confirment que si la température est un moteur de la dynamique, la variabilité géographique reste un facteur majeur de la présence des moustiques.

5 Conclusion

Ce projet a permis de mettre en place un modèle Bayésien hiérarchique pour analyser la dynamique d'invasion du moustique tigre sur une période de 12 ans. À travers les résultats obtenus, nous pouvons tirer les conclusions majeures suivantes :

— **Efficacité de l'Inférence :** L'utilisation de l'algorithme HMC (*Hamiltonian Monte Carlo*) a permis d'obtenir une convergence robuste, confirmée par les tracés de traceplots et une décroissance rapide de l'autocorrélation (ACF). Le modèle a su transformer des données de terrain hétérogènes en estimations statistiques fiables et biologiquement cohérentes.
— **Dynamique de Colonisation :** Les résultats mettent en évidence une progression nette de l'intensité de présence (λ), passant de valeurs initiales faibles à une explosion démographique marquée, particulièrement lors de l'année 10. La structure hiérarchique du modèle a été déterminante pour capturer cette variabilité inter-annuelle tout en gérant l'incertitude liée aux variations de l'effort d'échantillonnage (J).
— **Robustesse et Limites :** Bien que le modèle décrive avec précision la tendance globale, l'analyse prédictive *a posteriori* (PPC) a révélé les limites intrinsèques de la loi de Poisson face à des événements extrêmes, tel que le record observé de **252 moustiques** sur un site unique. Cela souligne la complexité des phénomènes biologiques où des micro-facteurs locaux peuvent provoquer des pics de densité imprévisibles, dépassant la moyenne du groupe.

De plus, le modèle dans la partie **Extension** nous révèle des enseignements cruciaux sur les facteurs environnementaux :

— $\beta \approx 1.302$: Ce résultat confirme que la température a un impact positif très fort sur l'invasion. La chaleur agit comme un moteur direct favorisant la prolifération des moustiques.
— **Diminution de μ (**$1.391 \rightarrow 1.321$**)** : Cette baisse démontre que la température explique désormais une partie de l'abondance moyenne. Une fraction de la dynamique qui était auparavant "inexpliquée" est maintenant captée par cette variable climatique.
— **Augmentation de σ^2 (**$1.042 \rightarrow 1.336$**)** : En analysant les données à l'échelle des sites (départements), le modèle détecte une hétérogénéité spatiale beaucoup plus importante. Cela montre que le modèle étendu parvient à identifier des disparités locales que le modèle initial, limité à une vision annuelle, ignorait.

Par conséquent, ce travail démontre que si la température favorise globalement l'invasion, la dynamique reste complexe et géographiquement contrastée, faisant de ce modèle étendu un outil plus robuste pour la surveillance écologique.

L'analyse des résultats a mis en évidence une forte hétérogénéité de la présence du moustique. D'une part, on observe une forte variabilité inter-annuelle avec des années présentant des pics d'occurrences marqués. D'autre part, la variabilité spatiale (intra-annuelle) démontre que l'invasion n'est pas uniforme, l'espèce se concentrant massivement dans certains « hotspots » régionaux tandis que d'autres territoires restent moins touchés.

C'est précisément dans cette perspective que nous avons enrichi ce modèle statistique en intégrant la température comme covariable environnementale. Cette approche a permis de mieux expliquer les variations observées et d'améliorer les capacités prédictives du modèle face à cette problématique majeure de santé publique et d'écologie.

Références

[1] R. M. Neal, *MCMC using Hamiltonian dynamics*, University of Toronto, Canada.
Disponible sur : `https://arxiv.org/pdf/1206.1901`

[2] E. Walker, S. Soubeyrand, *Hamiltonian Monte Carlo in Practice*, Institut National de la Recherche Agronomique (INRA), Avignon, France.
Disponible sur : `https://biosp.mathnum.inrae.fr/sites/default/files/2021-03/RR2016_49.pdf`

[3] Y. C. Chen, *Lecture 9 : Hamiltonian Monte Carlo (leapfrog)*, Faculty of Washington.
Disponible sur : `https://faculty.washington.edu/yenchic/19A_stat535/Lec9_HMC.pdf`

[4] IBM Documentation, *Statistiques bayésiennes (SPSS Statistics)*, IBM Docs.
Disponible sur : `https://www.ibm.com/docs/fr/spss-statistics/cd?topic=ctas-bayesian-statistics`

[5] IRD (Institut de Recherche pour le Développement), *Changement climatique : le moustique tigre gagne du terrain en Europe*. Disponible sur :
`https://lemag.ird.fr/fr/changement-climatique-le-moustique-tigre-gagne-du-terrain-en-europe`

[6] ANSES (Agence nationale de sécurité sanitaire), *Le moustique tigre (Aedes albopictus)*.
Disponible sur : `https://www.anses.fr/fr/content/le-moustique-tigre`