FACULTÉ DES SCIENCES ET TECHNOLOGIES | Université de Lille

Projet Statistiques Computationnelles

Master 1 Mathématiques et applications parcours ISN

AUTEUR :
Moustapha SARR et Kobla Legbedje

PROFESSEUR :
Charlotte Baey

Année universitaire 2023-2024

Modélisation d'une épidémie par une approche bayésienne

Introduction

La maladie à virus Ébola doit son nom à une rivière nommée Ébola passant près de la ville de Yambuku dans le nord de la République Démocratique du Congo (RDC) [50]. C'est en effet dans l'hôpital de cette localité que faut observé pour la première fois Ébola en 1976 [65]. Cette découverte fut le début de la première épidémie d'Ébola, qui eu touché 318 personnes dont 280 en ont perdus la vie. Le virus Ébola appartient à la famille des filovirus, qui regroupe les virus à l'apparence filamenteuse caractéristique.
L'épidémie du virus Ébola (EBOV) de 2014 en Afrique de l'Ouest est la plus grande épidémie du genre Ebolavirus à ce jour. L'épidémie a débuté en Guinée en décembre 2013 et s'est ensuite propagée à la Sierra Leone, au Libéria et au Nigeria.

Figure 1 – Virus Ebola

L'objectif de ce projet est de modéliser la flambée épidémique de 2014, et d'estimer les paramètres clés de l'épidémie. On utilisera pour cela un modèle d'épidémiologie et une approche bayésienne.

On propose d'utiliser un modèle épidémiologique à compartiments de type SEIR (Susceptibles - Exposés - Infectés - Retirés, ou en anglais Susceptible - Exposed - Infected - Recovered). Ce type de modèle permet de décrire l'évolution d'une population en période d'épidémie, en considérant qu'un individu pris au hasard dans la population est dans l'un (et un seul) des états précédemment décrits. Le modèle SEIR peut être décrit par l'ensemble d'équations différentielles suivant :

$$\begin{cases} \frac{dS}{dt} & = -\beta \frac{S(t)I(t)}{N}, \\ \frac{dE}{dt} & = \beta \frac{S(t)I(t)}{N} - \sigma E(t), \\ \frac{dI}{dt} & = \sigma E(t) - \gamma I(t), \\ \frac{dR}{dt} & = (1-\mu)\gamma I(t). \end{cases}$$

Où :

S : Nombre de personnes susceptibles,
E : Nombre de personnes exposées,
I : Nombre de personnes infectées,
R : Nombre de personnes récupérées,
N : Population totale,
β : Taux de transmission,
σ : Taux d'incubation,
γ : Taux de contagiosité,
μ : Taux de létalité

Les individus susceptibles peuvent devenir exposés lorsqu'ils entrent en contact avec des individus infectieux.
Les individus exposés deviennent infectieux après une période d'incubation.
Les individus infectieux peuvent être retirés de la population infectieuse soit en récupérant et en devenant immunisés, soit en décédant.
Les individus retirés ne peuvent plus être infectés et ne participent pas à la transmission de la maladie.

Phase 1 :

Pour la première question, pour tester la sensibilité du modèle, nous avons fixé $\beta = 1.3/5.61$ car $R_0 = \beta/\gamma = 1.51$ pour la Guinée, R_0 est le nombre de personnes que risque de contaminer le patient I_0 avant que l'épidémie ne démarre , avec $\gamma = 1/5.61$, $\sigma = 1/5.3$ et $\mu = 0.74$. Les graphiques ci-dessous (Figure 2) montrent l'évolution des compartiments S, E, I et R du modèle SEIR.

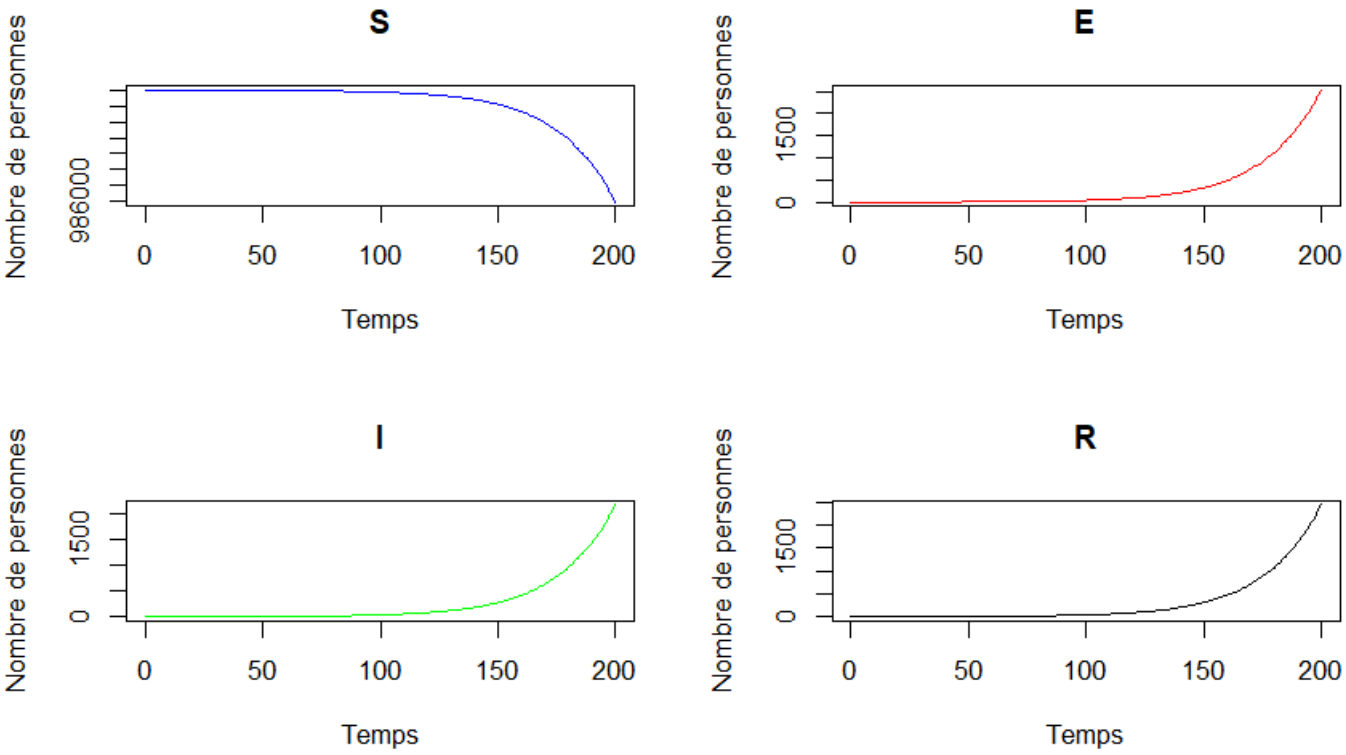

FIGURE 2 – La variation des compartiments du modèle SEIR

Nous observons une diminution du nombre d'individus susceptibles à partir du 100 ème jour et une augmentation du nombre d'individus exposés à la maladie, ainsi qu'une augmentation du nombre d'individus infectés et récupérés de la maladie pendant cette même période.

En maintenant le paramètre β constant, et en modifiant les valeurs des autres paramètres , on constate que le modèle reste inchangé, ce qui signifie que les mêmes graphiques sont obtenus.

Cependant, si nous varions la valeur de β, tout en maintenant les autres paramètres constants, nous obtenons les graphiques suivants :

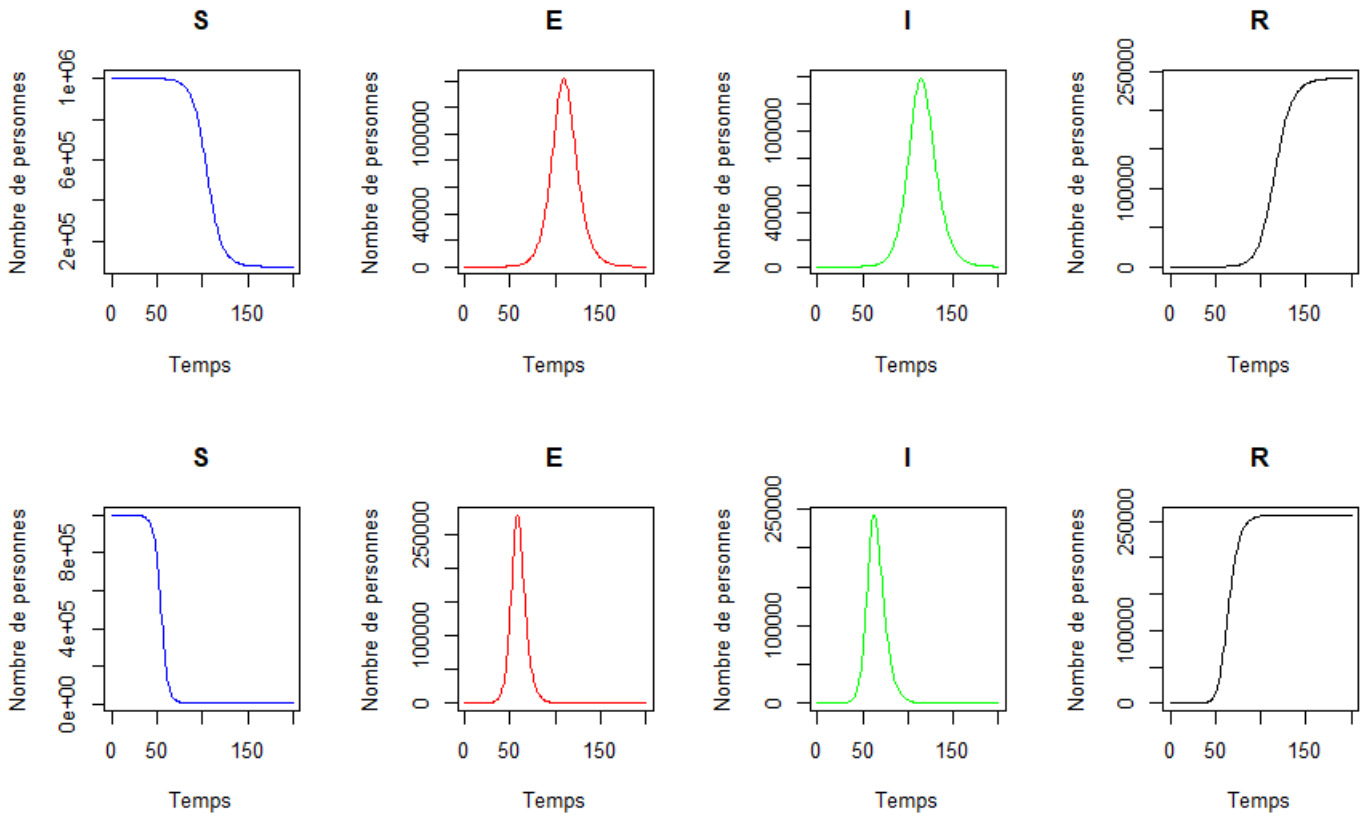

FIGURE 3 – La variation des compartiments du modèle SEIR pour différentes valeurs de β

Dans la figure 3, nous avons augmenté la valeur de β. Nous observons que les individus susceptibles diminuent au fil du temps on peut dire à partir du 90 ème jour jusqu'au 150 ème jour où leur nombre devient constant. En revanche, nous observons un pic important dans les graphiques des individus infectés et exposés à la maladie ainsi qu'une augmentation des individus retirés de la maladie à partir du 90 ème jour. Cela indique qu'une augmentation du taux de transmission entraîne une augmentation très rapide du nombre d'individus exposés et infectés. Cependant, à partir du 130 ème jour, leur nombre diminue jusqu'au 150ème jour où leur nombre devient constant , ce qui peut être dû à une immunité acquise ou à d'autres interventions de santé publique.
On peut conclure que la variation des compartiments du modèle dépend du paramètre β, s'il diminue, y aura moins d'individus dans les autres compartiments et plus qu'il augmente, on va avoir un pic d'épidémie qui devient de plus en plus important au niveau des individus exposés à la maladie et ceux infectés.

Question 2 :
Les données de la Guinée, du Libéria et de la Sierra Leone sont reportés aux mêmes dates dans un champ en format JJ/MM/AA.
Les données détaillées sur le nombre de cas et le nombre de morts de chaque pays sont fournies pour chaque date.
Deux autres champs sont également utilisés pour évaluer le nombre total de décès et le nombre total de cas dans les trois pays.

L'étude concerne l'épidémie de virus Ebola qui a eu lieu entre 2014 et 2016 en Afrique de l'Ouest.
Dans cette étude, nous allons décrire l'épidémie de virus Ebola à l'aide d'un modèle épidémiologie de type SEIR et une approche bayésienne.
Ce modèle nous permet d'estimer les paramètres clés de l'épidémie à savoir le taux de transmission β, le taux d'incubation σ, le taux de contagiosité γ et le taux de létalité μ.
Nous avons fait le choix de la Guinée pour effectuer l'étude.

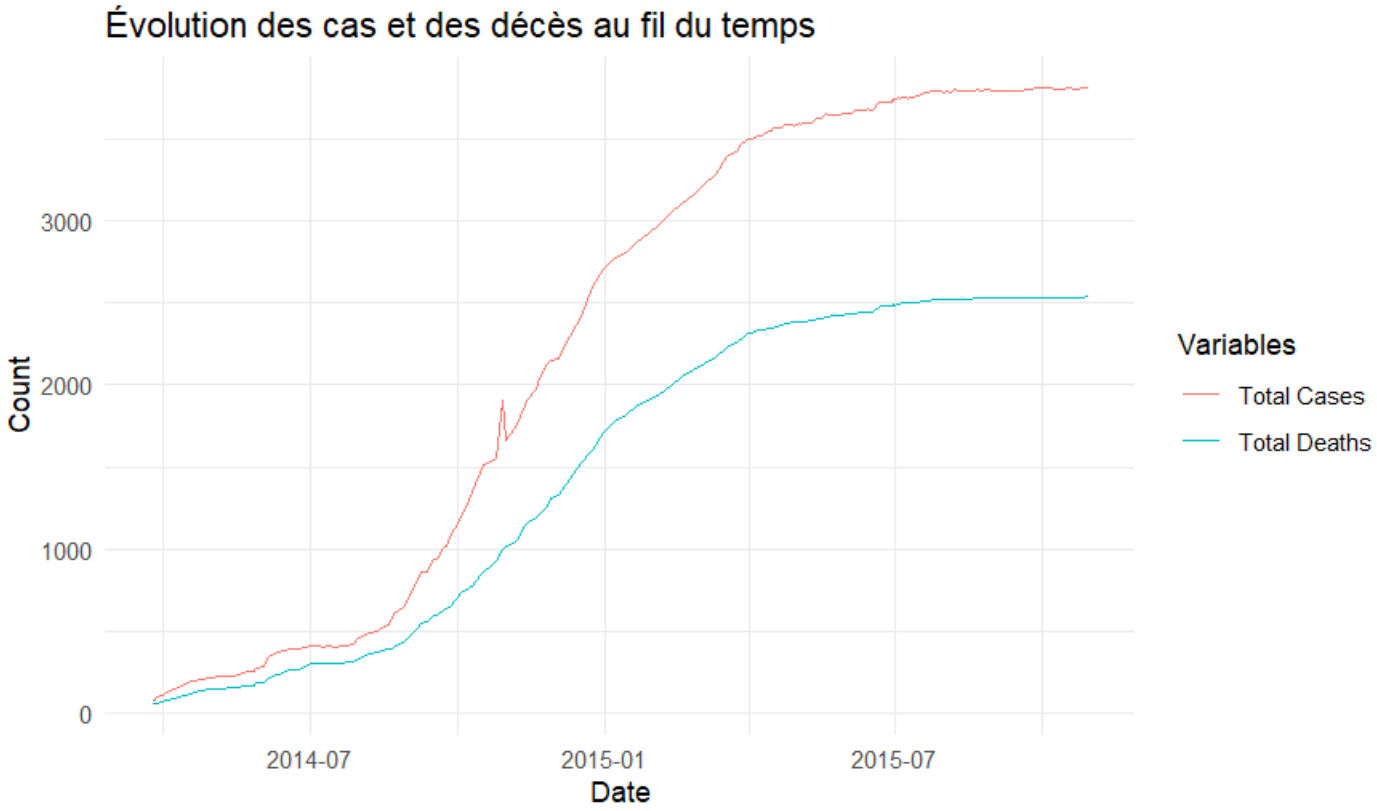

FIGURE 4 – Évolution des cas et des décès au fil du temps en Guinée

Question 3 :
On considère un modèle épidémiologique à compartiments de type SEIR dont on veut estimer ses paramètres, qui sont β, σ, γ et μ . Pour cela, on prend **N** comme étant la taille de la population de l'un des pays de la base de données, on note par **S** les individus qui sont susceptibles d'être infectés par le virus Ébola, par **E** les individus exposés au virus, par **I** les individus infectieux qui peuvent transmettre le virus, et par **R** les individus qui se sont rétablis de l'infection par le virus ou qui sont décédés de la maladie.
On note β la probabilité de transmission de la maladie d'un individu infectieux à un individu susceptible lors d'un contact. Étant donné que β dépend de divers facteurs tels que la densité de la population, les comportements individuels et les interventions de santé publique, on peut utiliser

une loi a priori non informatif comme par exemple **une distribution uniforme sur l'intervalle** $[0, 1]$. Étant donné que β est une quantité variant entre 0 et 1, il faut donc définir une mesure de probabilité sur $[0, 1]$.

Le taux d'incubation σ peut être influencé par la virulence spécifique du virus Ébola et d'autres facteurs biologiques. En disposant que de données sur le nombre de cas confirmés et le nombre de décès, on peut utiliser une loi a priori non informatif comme par exemple **une distribution uniforme sur l'intervalle** $[0, 1]$.

Étant donné que les données disponibles sont limitées aux cas confirmés et aux décès, il peut être difficile d'estimer directement le taux γ à partir de ces données.On peut utiliser une loi a priori non informatif comme par exemple **une distribution uniforme sur l'intervalle** $[0, 1]$.

Le taux de létalité μ est d'une importance capitale dans la distribution de l'épidémie d'Ebola. on peut utiliser une loi a priori non informatif comme par exemple **une distribution uniforme sur l'intervalle** $[0, 1]$.Étant donné que μ est une quantité variant entre 0 et 1, il faut donc définir une mesure de probabilité sur $[0, 1]$

On a fait ces hypothèses en se basant sur cette référence : "**C. L. Althaus. Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. PLoS currents, 6, 2014**".

Question 4 :
L'algorithme ABC (Approximate Bayesian Computation) est un ensemble de méthodes d'estimation de paramètres.
L'estimation des paramètres est faite dans une perspective bayésienne, c'est à dire on fournit des lois a posteriori.
Pour des modèles détaillés, composés de beaucoup de paramètres, il peut être difficile d'exprimer la fonction de vraisemblance $p(\mathcal{D}_{obs}|\theta)$. De plus, même s'il est possible de l'exprimer, la calculer numériquement et/ou calculer le maximum de vraisemblance peuvent être impossibles pour de grands jeux de données. Une solution est d'approcher la vraisemblance par simulation à l'aide de la méthode de calcul bayésien approché (ou ABC pour Approximate Bayesian Computation). La distribution a posteriori des paramètres θ est construite par méthode d'acceptation/de rejet basée sur la comparaison de données simulées et données observées.
Dans notre cas nous allons utiliser l'ABC qui repose sur le principe de l'acceptation-rejet pour des statistiques continues.
Cette méthode repose sur la génération d'échantillons de paramètres à partir de lois a priori spécifiées, puis sur la comparaison des données observées avec les données simulées à l'aide d'une fonction de distance appropriée.
Pour l'épidémie du virus Ebola, une distance appropriée pourrait être la distance euclidienne vue qu'on est dans un contexte où on veut comparer des données observées avec des donées simulées. .
On peut le définir de la façon suivante :

$$d(\mathcal{D}_{obs}, \mathcal{D}_i) = \sqrt{\sum (\mathcal{D}_{obs} - \mathcal{D}_i)^2}$$

Soient β, σ, γ et μ les paramètres du modèle SEIR, et soit le nombre total de cas confirmés et le nombre total de décès en Guinée du début de l'épidémie jusqu'à fin octobre 2015 l'ensemble des données observées, que l'on note $\mathcal{D}_{obs}$.

L'algorithme ABC consiste à :
pour i allant de 1 à n (taille de l'échantillon souhaitée) faire :

- Simuler β selon la loi uniforme sur l'intervalle $[0, 1]$
-Simuler σ selon la loi uniforme sur l'intervalle $[0, 1]$
-Simuler γ selon la loi uniforme sur l'intervalle $[0, 1]$
-Simuler μ selon la loi uniforme sur l'intervalle $[0, 1]$
-Simuler le nombre de cas et le nombre de décès selon le modèle SEIR, que l'on note $\mathcal{D}_i$
-Calculer la distance $d(\mathcal{D}_{obs}, \mathcal{D}_i)$ entre les données simulées et les données observées que l'on a déjà définit.
— Si cette distance est inférieure à un certain seuil ϵ, on garde la valeur des paramètres simulés, sinon on les rejette, avec $\epsilon > 0$.

Question 5 :
Le paramètre β contrôle le taux de transmission de l'infection. Les valeurs plus élevées de β entraînent une propagation plus rapide de l'infection dans la population. Si la loi a priori est uniforme sur [0, 1], la loi a posteriori peut être concentrée autour de valeurs plus élevées de β si les données observées indiquent une transmission rapide de l'infection. Cela peut correspondre à des distributions biaisées vers des valeurs plus élevées de β.

Le paramètre σ représente le taux de transition des individus exposés (E) à la phase infectieuse (I). Des valeurs plus élevées de σ signifient que les individus passent plus rapidement de l'état exposé à l'état infectieux. La loi a posteriori pour σ peut être influencée par les données observées, avec des valeurs plus élevées de σ si les données indiquent une période d'incubation plus courte.

Le paramètre γ contrôle le taux de transition des individus infectieux (I) à la récupération (R) ou à la mort. Si la loi a priori est uniforme sur [0, 1], la loi a posteriori pour γ peut être influencée par les données observées sur la durée moyenne de l'infection dans la population.

Le paramètre μ représente le taux de mortalité des individus infectieux. La loi a posteriori pour μ peut être influencée par les données sur le taux de mortalité observé dans la population.

Voici ci-dessous les diagrammes obtenus pour les paramètres acceptés.

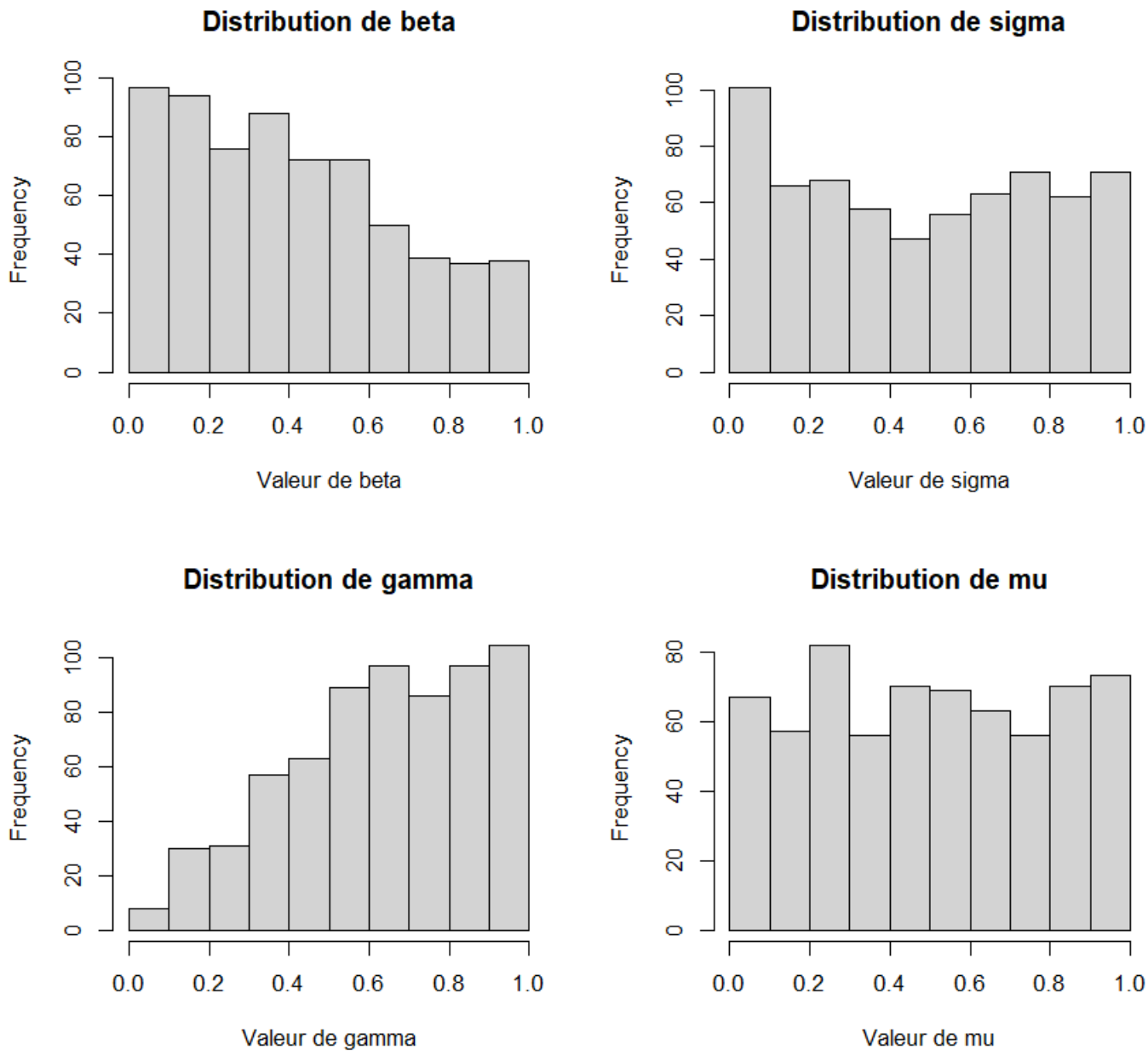

FIGURE 5 – Diagramme des paramètres acceptés par l'algorithme de ABC

Phase 2 :

Question 1 :
On peut utiliser un modèle stochastique à temps discret SEIR pour les maladies infectieuses en considérant que le taux de transmission β n'est plus une constante mais il va dépendre du temps. On prend $\beta(t) = \beta e^{-k(\tau - t)}$ en supposant que le taux de transmission décroît de façon exponentielle au taux k, avec τ qui représente le temps où des mesures de contrôle ont été introduites.
Donc le modèle (1) devient :

$$\begin{cases} \frac{dS}{dt} & = -\beta(t)\frac{S(t)I(t)}{N}, \\ \frac{dE}{dt} & = \beta(t)\frac{S(t)I(t)}{N} - \sigma E(t), \\ \frac{dI}{dt} & = \sigma E(t) - \gamma I(t), \\ \frac{dR}{dt} & = (1-\mu)\gamma I(t). \end{cases}$$

Question 2 :
(a) La dernière étape de la procédure ABC mis en place dans la première phase, peut ne pas être efficace quand les $\mathcal{D}_i$ et les $\mathcal{D}_{obs}$ sont de grande dimension et dans des espaces continues, dans ce cas, dans l'ABC, on se sert de statistiques résumées pour comparer les jeux de données $\mathcal{D}_i$ et $\mathcal{D}_{obs}$.
On note par :
$\mathcal{S} = s(\mathcal{D}_{obs})$: statistiques résumées observées
$\mathcal{S}_i = s(\mathcal{D}_i)$: statistiques résumées pour les paramètres simulées θ_i.
Algorithme :
pour i allant de 1 à n (taille de l'échantillon souhaitée) faire :
-Tirer indépendamment $\theta_i \quad (i = 1, \ldots, I)$ dans la loi a priori des paramètres du modèle (1) amélioré
–Simuler le nombre de cas et le nombre de décès selon le modèle SEIR, que l'on note $\mathcal{D}_i$
et calculer $\mathcal{S}_i$
-Accepter θ_i si $d(\mathcal{S}_i, \mathcal{S}) \leq \epsilon$ avec $\epsilon > 0$.
L'ensemble des points acceptés pour construire la loi a posteriori est définie par :

$$\{\theta_i : i = 1, \ldots, I \text{ et } d(S_i, S) \leq \epsilon\}$$

où I est l'ensemble des indices pour lesquels la valeur du paramètre a été acceptée.
On obtient alors des réalisations selon la loi a posteriori approchée $p(\theta | d(\mathcal{S}_i, \mathcal{S})) \leq \epsilon$

(b) L'algorithme de Monte Carlo en chaîne de Markov ABC, ABC-MCMC part d'une valeur θ arbitrairement choisie et procède comme suit :

pour i allant de 1 à n (taille de l'échantillon souhaitée) faire :
-Tirer indépendamment $\theta_i \quad (i = 1, \ldots, I)$ dans la loi a priori des paramètres du modèle (1) amélioré
-Simuler le nombre de cas et le nombre de décès selon le modèle SEIR, que l'on note $\mathcal{D}_i$
et calculer $\mathcal{S}_i$ pour chaque donnée simulée $\mathcal{D}_i$
-Accepter θ_i si $d(\mathcal{S}_i, \mathcal{S}) \leq \epsilon$ avec $\epsilon > 0$.
A l'iteration i + 1 faire :
-Tirer un θ^* candidat selon la loi de proposition $q(\theta^* | \theta_i)$
-Simuler le nombre de cas et le nombre de décès selon le modèle SEIR, que l'on note $\mathcal{D}_i$
et calculer $\mathcal{S}^* = s(\mathcal{D}^*)$
-Si $d(\mathcal{S}^*, \mathcal{S})) \leq \epsilon$ et $U(0,1) \leq \min\left\{1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)}\right\}$ alors
$\theta_{i+1} = \theta^*$ où π est la loi a priori des paramètres.
-Sinon $\theta_{i+1} = \theta_i$
A l'état stationnaire on échantillonne dans la loi

$$p(\theta | d(S_i, S) \leq \epsilon) = \frac{p\{d(S_i, S) \leq \epsilon | \theta\}\pi(\theta)}{p\{d(S_i, S) \leq \epsilon\}}$$

(c) Pour le choix itératif du seuil ϵ on peut procéder comme suit :
1-Initialisation du seuil ϵ : définir arbitrairement un $\epsilon > 0$ pour notre cas
2-Itération de l'algorithme ABC : commencer l'algorithme ABC avec le seuil initial choisi
3-Évaluation de la performance : évaluer la performance de l'algorithme en fonction du seuil choisi,

cela peut se faire en analysant la convergence de l'algorithme
4-Mise à jour du seuil ϵ :
-Si l'algorithme ABC converge rapidement et que la couverture de la distribution postérieure est satisfaisante, on peut diminuer le seuil pour obtenir une meilleure approximation de la distribution postérieure.
-Si l'algorithme ABC converge lentement ou que la couverture de la distribution postérieure est insatisfaisante,on peut augmenter le seuil.
5-Répéter les étapes 2 à 4.

(d) L'algorithme ABC-regression :
-Calculer $\mathcal{S}$
-Tirer indépendamment n valeurs de paramètres $\theta_i \quad (i = 1, \ldots, I)$ dans la loi a priori des paramètres du modèle (1) amélioré
-Simuler le nombre de cas et le nombre de décès selon le modèle SEIR, que l'on note $\mathcal{D}_i$
et calculer $\mathcal{S}_i$ pour chaque donnée simulée $\mathcal{D}_i$
-Associer à chaque paire $s(\mathcal{D}_i, \theta_i)$ un poids $w_i = 0$ si la distance $d(\mathcal{S}_i, \mathcal{S})$ n'appartient pas au quantile P de la distribution empirique des distances $d(\mathcal{S}_i, \mathcal{S})$, et un poids défini par une fonction de noyau statistique $d(\mathcal{S}_i, \mathcal{S})$ sinon
-Construire un modèle de régression linéaire $m(\mathcal{S}_i) = \theta$ à partir des paires pondérées $s(\mathcal{D}_i, \theta_i)$
-Calculer la distribution a posteriori à partir du modèle de régression telle que
$\theta_i^* = \theta_i + m(\mathcal{S}) - m(\mathcal{S}_i)$.
Question 3 :

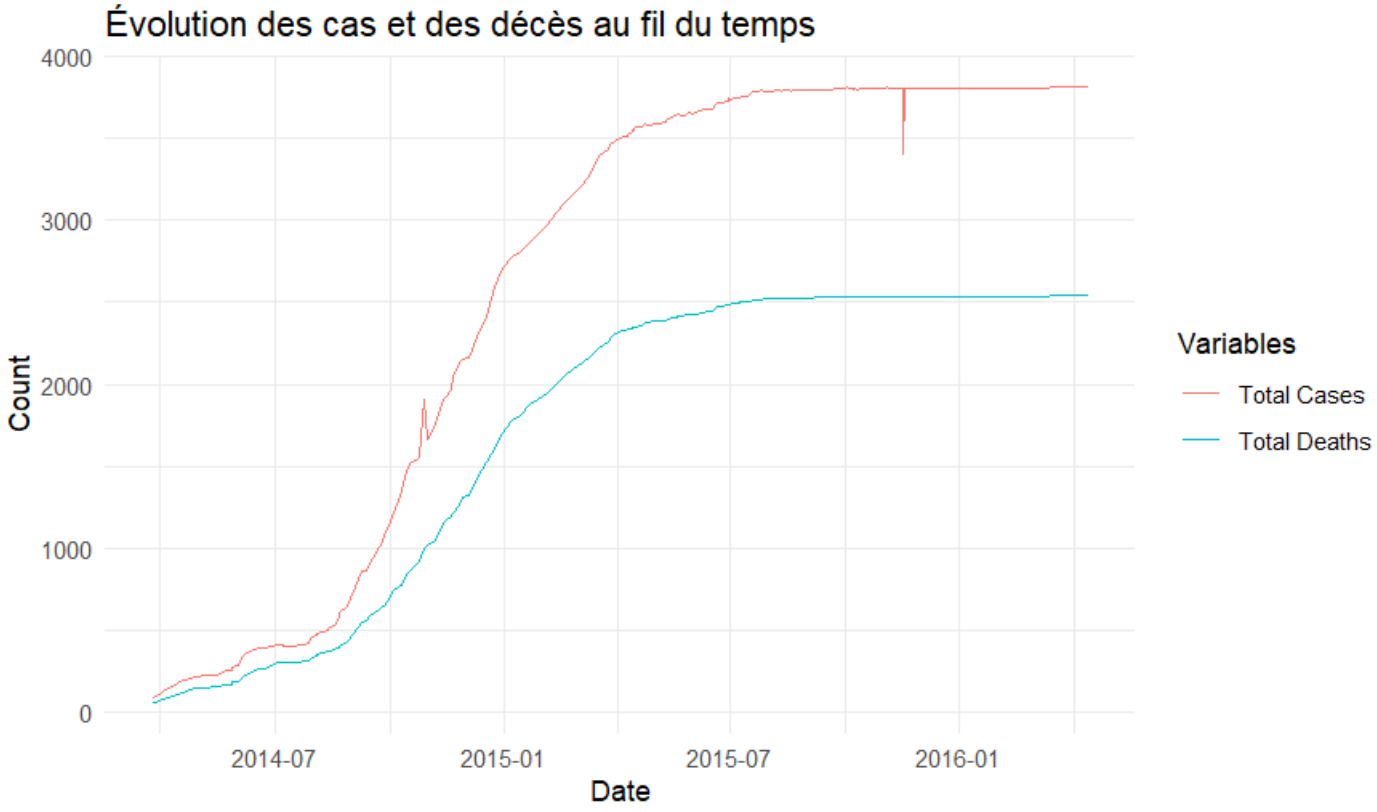

FIGURE 6 – Évolution des cas et des décès au fil du temps en Guinée du début de l'épidémie jusqu'à la fin

Conclusion :

Ce projet avait pour objectif de modéliser l'épidémie d'Ebola survenue en Afrique de l'Ouest entre 2014 et 2016, en utilisant un modèle épidémiologique SEIR et une approche bayésienne. L'épidémie d'Ebola constitue un défi majeur pour la santé publique en raison de sa gravité et de sa capacité à se propager rapidement dans les populations.

En appliquant le modèle SEIR aux données de l'épidémie, nous avons pu estimer les paramètres clés du modèle, tels que le taux de transmission β , le taux d'incubation σ , le taux de guérison ou de décès γ, et le taux de létalité μ. Nous avons utilisé une approche bayésienne, en particulier la méthode ABC, pour estimer ces paramètres en tenant compte de l'information a priori disponible.

Les résultats de notre analyse ont révélé des lois a posteriori approximées pour les paramètres du modèle, ce qui nous a permis de mieux comprendre les dynamiques de l'épidémie d'Ebola de 2014. Nous avons également identifié des intervalles de crédibilité pour ces paramètres, ce qui nous donne une indication de la fiabilité de nos estimations.

En examinant les résultats obtenus, nous constatons que notre approche a fourni des estimations cohérentes des paramètres du modèle, en accord avec les connaissances existantes sur l'épidémiologie de l'Ebola. Cependant, des améliorations sont possibles, notamment en explorant des méthodes d'optimisation de l'algorithme ABC et en affinant le modèle épidémiologique pour mieux capturer les caractéristiques de l'épidémie.

En conclusion, ce projet a contribué à la modélisation et à la compréhension de l'épidémie d'Ebola de 2014, en utilisant des méthodes statistiques avancées et des outils informatiques modernes. Ces résultats sont pertinents pour la santé publique et la gestion des épidémies futures, en fournissant des outils pour évaluer et prévoir la propagation des maladies infectieuses à l'échelle mondiale.

Nous sommes encouragés par les résultats de ce projet et nous restons engagés à poursuivre la recherche dans ce domaine crucial de la santé publique.

References :

C. L. Althaus. Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. PLoS currents, 6, 2014.

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate bayesian computation. Biometrika, 96(4) :983–990, 2009.

M. G. Blum and O. François. Non-linear regression models for Approximate Bayesian Computation. Statistics and computing, 20(1) :63–73, 2010.

P. E. Lekone and B. F. Finkenstädt. Statistical inference in a stochastic epidemic seir model with control intervention : Ebola as a case study. Biometrics, 62(4) :1170–1177, 2006.

P. Marjoram and S. Tavaré. Modern computational approaches for analysing molecular genetic variation data. Nature Reviews Genetics, 7(10) :759–770, 2006.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes : a study of Y chromosome microsatellites. Molecular biology and evolution, 16(12) :1791–1798, 1999.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. Genetics, 145(2) :505–518, 1997.