Université de Strasbourg
UFR de Mathématiques et Informatique

Projet SIR

Etudiant : *Roussel Desmond Nzoyem*

UE : *Incertitudes* – Enseignant : *Pr. Frédéric Bertrand*
Date : *15 janvier 2021*

Introduction

L'objectif de ce travail est la simulation de l'épidémie de COVID-19 par un modèle de type SIR, dont les paramètres seront estimés principalement par calage déterministe. Nous avons choisi le modèle **SIRD** (Susceptibles-Infectés-Retirés(ou Rétablis)-Décédés) avec introduction du taux de natalité et de mortalité général de la population. Le modèle a été implémenté en langage Python avant d'être intégré à la librairie de calcul d'incertitudes **OpenTurns**. Ce modèle comporte neuf paramètres (y compris les conditions initiales) dont les lois de probabilités ont été estimées en utilisant les observations faites au **Cameroun** entre le 06 mars et le 14 juin 2020. Après une comparaison des résultats avec ceux obtenus dans une étude faite par **Nguemdjo et al.**, nous avons réussi à prédire l'évolution de l'épidémie. Les analyses de sensibilité par développement de Taylor et par calcul des indices de Sobol ont permis d'extraire des indications sur la limitation des ravages causés par l'épidémie au Cameroun.

I. Modélisation mathématique

D'entrée, le modèle choisi est limité par les données auxquelles on aura accès pour la partie 2 de cette étude (il s'agira de l'estimation des paramètres). Nous avons opté pour le modèle SIRD (Lisphilar, 2020) avec ajout du taux de natalité μ et du taux de mortalité ν de la population générale (Bayette et Monticelli, 2020)[1].

I.1 Le modèle SIRD

Commençons par décrire le modèle SIRD fréquemment rencontré dans la littérature. Nous sommes déjà familier avec le modèle classique SIR qui décrit respectivement les populations Susceptibles de contacter les virus, celles Infectées par le virus, et celles Retirées de l'étude. Dans le modèle SIR, les personnes Retirées (R) sont supposées soit guéries (et immunisées), soit décédés. Cela dit, les organismes d'agrégation de données comptent généralement les populations guéries et décédés séparément. On obtient alors le modèle SIRD dans lequel le R est mis pour Retirés[2] (ou Rétablis), et le D mis pour Décédés. Ce modèle est décrit à la figure 1.

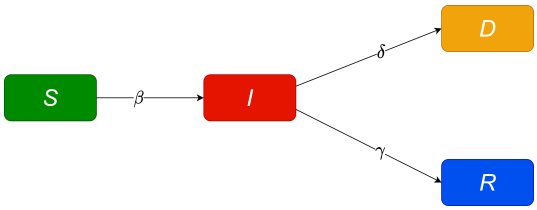


Figure 1 – *Illustration du modèle SIRD. β représente la proportion de personnes saines (susceptibles) qui deviennent infectées après un contact avec une personne infectée. Autrement dit, il s'agit du taux de transmission du virus. γ désigne le taux de guérison, et δ le taux de mortalité lié au virus.*

1. Attention, dans l'article du CNRS qui a inspirée ce modèle, les définitions de μ et ν sont inversées. Précisons aussi que dans cet article, l'introduction de la natalité et de la mortalité se fait à travers l'étude du modèle SEIR.
2. Bien que ce terme désigne à présent les population Rétablies, nous utiliserons le terme "Retirés" pour raison de concordance avec le modèle SIR classique.

I.2 Le modèle SIRD avec natalité et mortalité

Le modèle SIRD suppose que la population totale reste constante durant l'épidémie, ce qui est rarement le cas. Nous décidons donc de complexifié le modèle SIRD en y ajoutant les taux de natalité μ et du taux de mortalité ν liés à la population [3]. Dans la suite de ce rapport, nous désignerons ce modèle par SIRD*. Le schéma suivant illustre la dynamique dudit système.

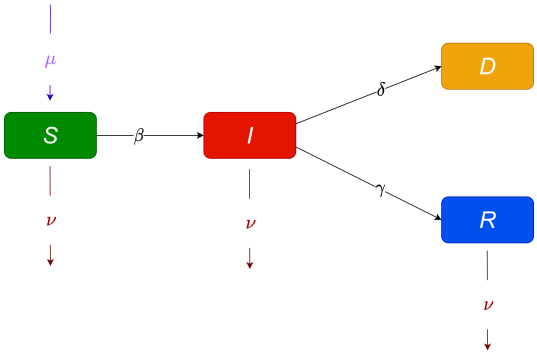


Figure 2 – *Illustration du modèle SIRD avec ajout du taux de natalité μ et du taux de mortalité ν de la population. Notons que contrairement à l'article du CNRS (*Bayette *et* Monticelli, *2020) qui a inspirée ce modèle, les definitions de μ et ν sont inversées.*

Notons que les personnes qui naissent sont bien évidemment non décédés, d'où l'ajout d'une somme (pondéré par le taux de natalité) des populations saines (S), infectées (I), et retirées (R) aux populations susceptibles (S). De manière similaire, les populations décédées (de cause non liée au virus) seront soustraites des populations S, I, et R. Cela nous donne le système d'EDO suivant [4] :

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} = -\beta\dfrac{SI}{N} + \mu(S+I+R) - \nu S \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = \beta\dfrac{SI}{N} - (\gamma+\delta)I - \nu I \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \nu R \\ \dfrac{\mathrm{d}D}{\mathrm{d}t} = \delta I \end{cases} \qquad (\mathcal{SIRD}^*)$$

Remarquons que contrairement aux modèles SIR et SIRD, la population totale N dans le modèle SIRD* n'est pas forcément constante. En effet, on a

$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}t} &= \frac{\mathrm{d}(S+I+R+D)}{\mathrm{d}t} \\ &= \mu(S+I+R) - \nu(S+I+R) \\ &= (\mu-\nu)(S+I+R) \end{aligned}$$

et donc

- Si $\mu > \nu$, la population totale augmente
- Si $\mu = \nu$, la population totale est constante
- Si $\mu < \nu$, la population totale diminue

Concernant le taux de reproduction du virus, on se focalise sur l'EDO sur le nombre d'infectés (I) dans le système ($\mathcal{SIRD}^*$) pour trouver :

$$\mathcal{R}_0 = \frac{\beta}{\gamma+\delta+\nu}$$

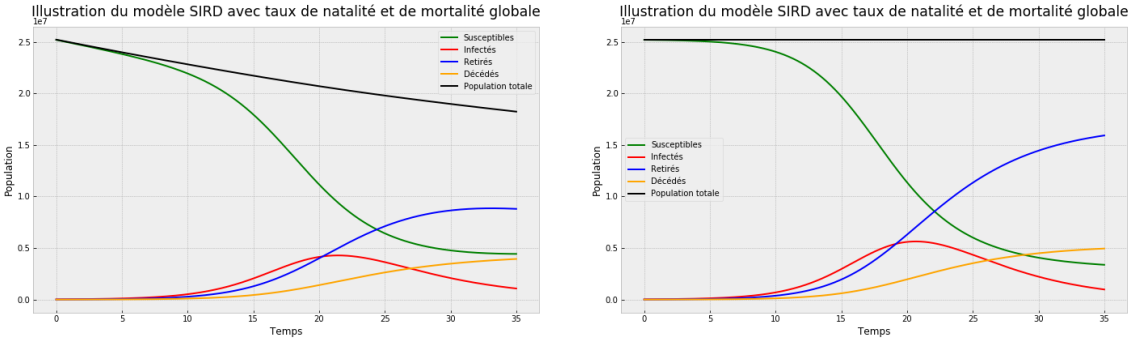
On sait que si :

3. Notons que ce taux de mortalité n'est pas lié au virus.
4. Pour plus de clarté, les dépendances des variables S, I, R, D ou N en fonction de t n'ont pas été précisées dans ces équations.

— $\mathcal{R}_0 > 1$, le virus se propage dans la population [5]

— $\mathcal{R}_0 \leq 1$, la propagation du virus va s'arrêter au bout d'un moment

C'est donc ce modèle que nous avons implémenté en Python par un schéma de Runge-Kutta d'ordre 4. Nous présentons ci-bas quelques résultats de simulation.



(a) $\mu = 0.02 \leq 0.03 = \nu$ (b) $\mu = \nu = 0$

Figure 3 – *Illustrations des résultats de l'implémentation du modèle SIR avec les coefficients* $\beta = 0.615, \gamma = 0.193, \delta = 0.06$ *fixés. Les conditions initiales sont* $S_0 = 25216237 - 2e4, I_0 = 2e4, R_0 = 0, D_0 = 0$*, et le temps de simulation est de* $T = 35$ *jours. Remarquons que la figure (b) permet de revenir à un modèle SIRD classique, où la population totale reste constante.*

II. Estimation des paramètres

Dans cette section, nous allons estimer les lois de probabilité suivies par cinq des paramètres du modèle. Il s'agit d'estimer les distributions sur $\beta, \gamma, \delta, \mu,$ et ν. Pour ce faire, nous devons obtenir des donnés observées sur une population. Nous choisissons naturellement le cas du Cameroun (CMR), où le premier cas d'infection a été constaté le 06 mars 2020.

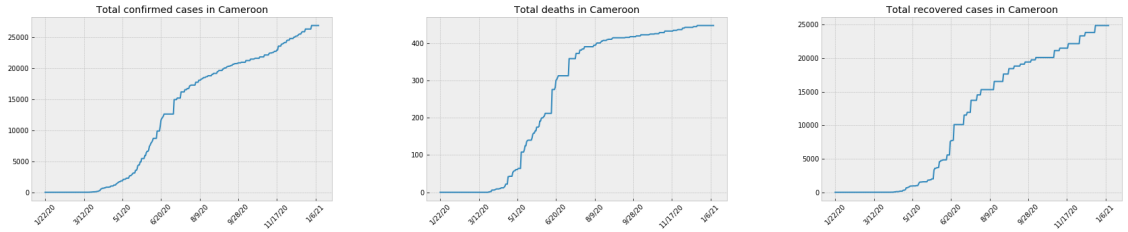
II.1 Création des entrées-sorties observées

En janvier 2020, au tout début de l'éruption du Coronavirus, le CSSE [6] de l'Université Johns Hopkins (CSSE, 2021a) s'est mise à regrouper des informations [7] visualisables sur ce site. Les données correspondantes sont conservées sur un dépôt GitHub (CSSE, 2021b) mis à jour quotidiennement. Nous y avons récupéré les informations liées au Cameroun sur la période du 22 janvier 2020 au 10 janvier 2021. Ces informations sont obtenues sous forme de trois séries temporelles (cf. figure 4) :

— le nombre de cas confirmés ("confirmed")

— le nombre de cas décédés ("deaths")

— le nombre de cas rétablis ("recovered")



(a) *Cas confirmés* (b) *Cas décédés* (c) *Cas rétablis*

Figure 4 – *Illustration des données brutes obtenues de la base de données de l'Université Johns Hopkins concernant l'évolution du COVID-19 au Cameroun du 22 janvier 2020 au 10 janvier 2021. Les effectifs présentés sur cette figure sont des effectifs cumulés.*

5. La probabilité d'avoir une épidémie n'est cependant pas garantie d'après Nguemdjo et al., 2020, p.4.
6. Center for Systems Science and Engineering
7. Ces données sont à l'échelle internationale et proviennent de plusieurs sources fiables.

Une fois les données brutes obtenues, il nous faut les classer en fonction des catégories qui sont intéressantes pour notre modèle SIRD*. Aucun recensement de population n'ayant eu lieu entre les dates d'observation, la population totale du Cameroun sera supposée constate $N = 25,216,237$. On a les relations suivantes (figure 5) :

— Décédés = Nombre de cas décédés

— Retirés = Nombre de cas rétablis

— Infectés = Nombre de cas confirmés - détectés - rétablis

— Susceptibles = Population totale - Nombre de cas confirmés

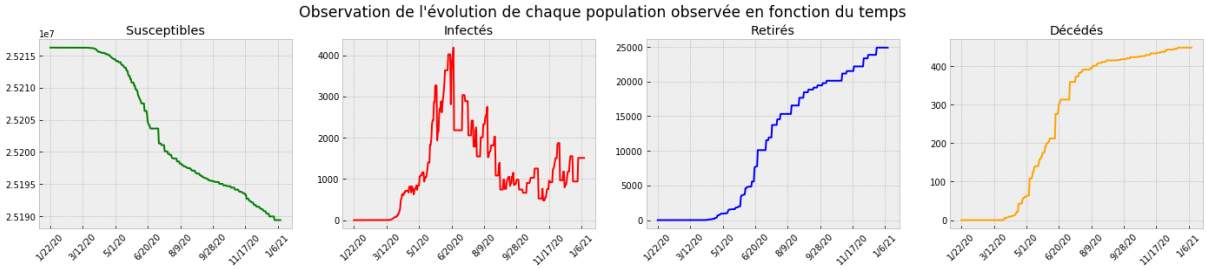


Figure 5 – *Illustration des données prétraitées sur la période du 22 janvier 2020 au 10 janvier 2021 au Cameroun.*

Une fois les données prétraitées et classées par catégories, on peut créer les entrées et les sorties observées. Mais avant, il faut choisir une fenêtre dans laquelle faire notre étude. Comme mentionné plus haut, le premier cas de COVID-19 a été détecté au Cameroun le 06 mars 2020 ; notre étude débutera donc à cette date, et s'achèvera 100 jours plus tard, soit le 14 juin 2020. Dans cette fenêtre, on prend des observations correspondant à des simulations sur 15 jours chacune. On peut en extraire 85 qui soient valides. En résumé, si $X(t)$ correspondant à l'état du système au temps t, alors l'entrée observée vaut $X(t)$ et la sortie observée correspondante vaut $X(t+15)$. Cela correspond au code ci-bas, dont le résultat est visualisé à la figure 6.

```
inputs = np.zeros((85, 4))
outputs = np.zeros((85, 4))
for k in range(85):
    inputs[k] = np.array([S[k], I[k], R[k], D[k]])
    outputs[k] = np.array([S[k+15], I[k+15], R[k+15], D[k+15]])
```

Listing 1 – *Code de génération des entrées-sorties afin d'effectuer l'estimation des paramètres. Les données S, I, R, et D correspondent à ce qui est observé à la figure 5.*

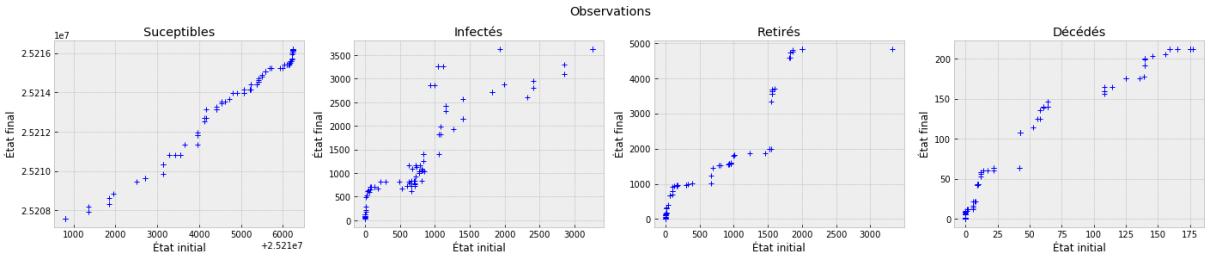


Figure 6 – *Illustration des données traitées observées sur la période du 06 mars au 14 juin 2020. Ces données seront utilisées pour l'estimation des paramètres. Observons que dû à la faible taille de la simulation (15 jours) à l'intérieur de la fenêtre d'étude, ces données ont une tendance linéaire que nous exploiterons dans la suite.*

II.2 Estimation des paramètres par calage

Maintenant que nous avons les données observées, nous pouvons estimer les paramètres qui s'accordent le mieux avec ces données. Pour cela, on se sert d'OpenTurns. Tout d'abord, on crée un objet de type `PythonFunction` pour introduire note modèle SIRD* (implémenté en Python) dans

OpenTurns. Ensuite, nous nous servons de la classe `LinearLeastSquaresCalibration` pour estimer les paramètres $\beta, \gamma, \delta, \mu$, et ν. La décision d'utiliser cet estimateur vient de la tendance linéaire des données observées à la figure 6. Cet estimateur linéarise le modèle et minimise la distance aux données observées [8], ceci autour d'un état (nommé "prior") que nous choisirons raisonnablement égale à `[0.467,0.4,0.001,0.0001,0.0001]` [9]. L'estimation des lois de probabilités de ces cinq paramètres produit les lois normales dont les moyennes, les écarts types, les valeurs supérieures et inférieurs des intervalles de confiance (CI) à 95 % sont présentées dans le tableau ci-après.

Paramètre	Valeur de référence	Moyenne	Écart type	Inf CI	Sup CI
β	0.4670	0.4503	0.0041	0.4192	0.4813
γ	0.4000	0.3670	0.0041	0.3353	0.3987
δ	0.0010	0.0057	0.0028	-0.0155	0.0269
μ	0.0001	0.0565	0.0019	0.0423	0.0707
ν	0.0001	0.0565	0.0019	0.0423	0.0707

Table 1 – *Résultats de l'estimation des paramètres. La valeur de référence indiquées est la valeur "prior", celle donnée à l'estimateur linéaire de moindres carrés d'OpenTurns. Les quatre colonnes d'après correspondent aux valeurs dites "posterior", celles qu'on obtient après l'estimation.*

	β	γ	δ	μ	ν
β	1	0.5705	0.2350	0.5248	0.5248
γ	0.5705	1	-0.2793	-0.0712	-0.0712
δ	0.2350	-0.2793	1	-0.2161	-0.2161
μ	0.5248	-0.0712	-0.2161	1	1
ν	0.5248	-0.0712	-0.2161	1	1

Table 2 – *Matrice de corrélation après l'estimation des différents paramètres. Cette matrice ne nous sert qu'à des fins d'analyse, et nous considérons ces cinq paramètres comme indépendants dans la suite.*

Il est intéressant de constater que l'estimation n'est pas aberrante, de par le fait que les paramètres μ et ν sont quasiment identiques en lois (cf. tableau 1), et parfaitement corrélés après calage (cf. tableau 2), ce qui découle de la constance de la population globale durant l'expérience. De façon visuelle, notre processus d'estimation des paramètres est présenté à la figure 7.

Notons qu'une étude très similaire à la nôtre a été effectuée par Nguemdjo et al., et publiée en août 2020. Également sur la thématique de la prédiction de l'évolution du COVID-19 au Cameroun, cette étude s'est servie d'un modèle SIR et s'est focalisé sur la fenêtre du 06 mars 2020 au 10 avril 2020. Ses paramètres, obtenus à partir d'un estimateur par maximum de vraisemblance, sont présentés ci-bas.

	Original	Bias	Std. error	Bootstrap normal CI*	
				Inf	Sup
β	0.615	7.65e-06	0.003	0.610	0.619
γ	0.393	-3.69e-05	0.003	0.388	0.398
R_0	1.567	0.000	0.016	1.536	1.597
Maximum Infected	2,015,200	757.6864	76,463.73	1,864,576	2,164,309
Number of Days to reach the peak	81,06	-0.022	1.660	77.81	84.32

Figure 8 – *Résultats obtenus par Nguemdjo et al. sur l'estimation des paramètres d'un modèle SIR (du 06 mars au 10 avril 2020) pour modéliser l'évolution du COVID-19 au Cameroun (*Nguemdjo *et al., 2020, p.4).*

Remarquons que les résultats ci-hauts (cf. figure 8) diffèrent assez largement des nôtres (cf. tableau 1), pour les paramètres β et γ. En particulier le taux de propagation du virus β est considérablement plus bas chez nous. Cette différence s'explique principalement par le fait que la fenêtre

8. Au sens des moindres carrés.
9. En réalité, l'état "prior" a été cherché manuellement, de façon à correspondre à au moins une observation.

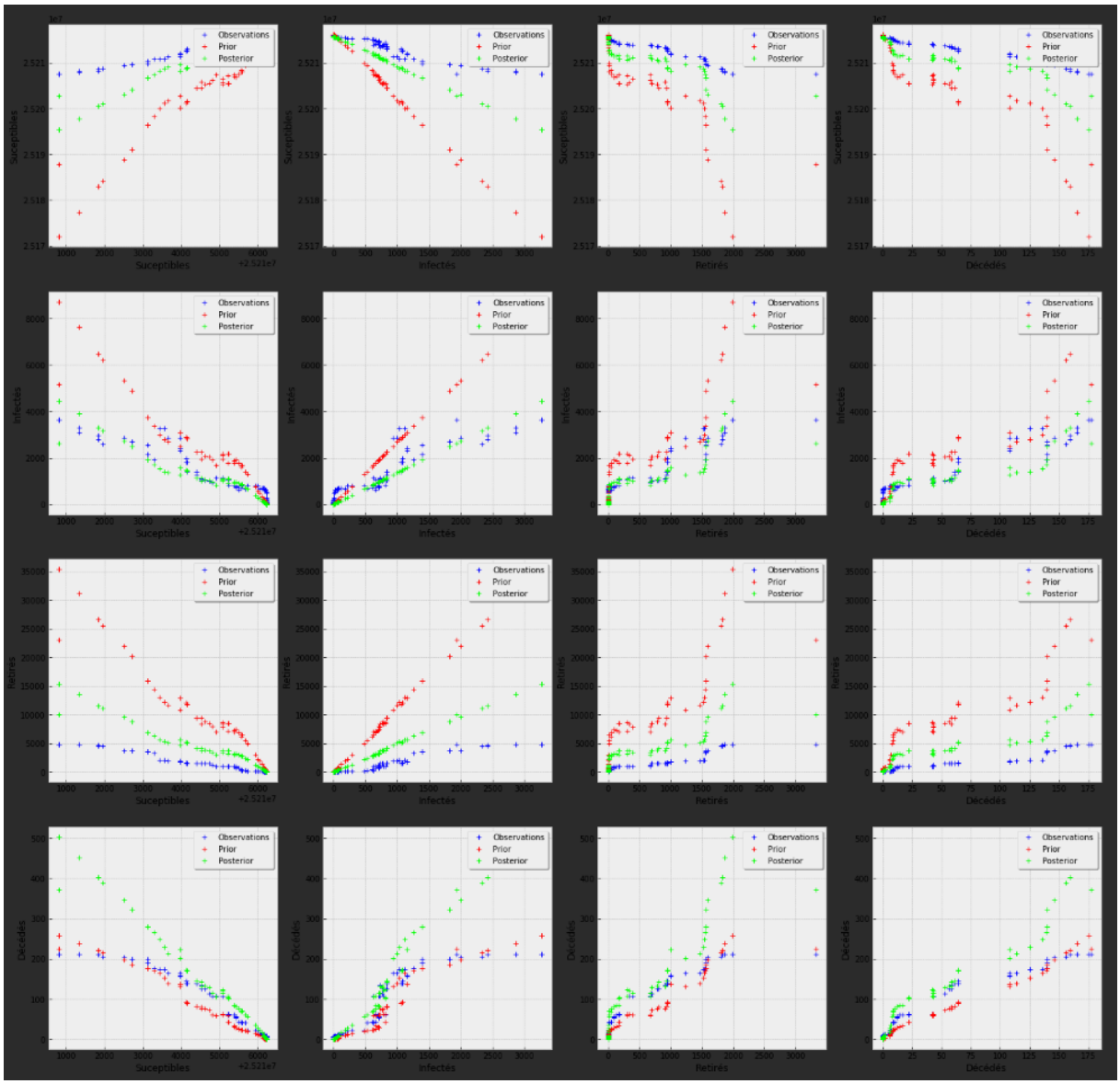


Figure 7 – *Visualisation des états "prior" et "posterior" pour l'estimation des paramètres. On observe bien que les états "posterior" sont globalement plus proche des observations que les états "prior". Notons que ceci n'est pas seulement le cas pour le figures sur la diagonale.*

d'étude que nous avons choisi (du 06 mars au 14 juin 2020) s'achève juste avant le pic de l'épidémie (cf. figure 5), qui correspond à un ralentissement des contaminations, d'où la diminution de β. En plus, le gouvernement Camerounais a eu le temps de renforcer ses mesures de confinement entre le 10 avril et le 14 juin 2020, ce qui a réduit le taux de propagation du virus [10].

II.3 Analyse de sensibilité

Nous allons effectuer une analyse de sensibilité de notre modèle SIRD*. En plus des paramètres étudiés dans la partie précédente, nous décidons ici d'analyser aussi l'effet des paramètres initiaux sur le modèle. En résumé, nous analyserons la sensibilité de S, I, R et D par rapport aux paramètres $\beta, \gamma, \delta, \mu, \nu, S_0, I_0, R_0$,et D_0. Nous devons au préalable obtenir les distributions de chacun de ces paramètres.

Les cinq premiers paramètres $\beta, \gamma, \delta, \mu, \nu$ ont été estimés dans la section précédente. Leurs loi

10. Dans le code fourni avec ce rapport, il est possible de choisir la même fenêtre d'étude que Nguemdjo et al., même si ce faisant, nous réduisons considérablement la taille des données observées, causant une perte de précision lors de l'estimation des paramètres.

de probabilités correspondrons à des lois normales dont les moyennes et les écarts-types sont présentées dans le tableau 1 [11]. En ce qui concerne les conditions initiales S_0, I_0, R_0,et D_0, elles suivent naturellement les même lois que S, I, R,et D. En observant la figure 5 sur la duré du 06 mars au 14 juin 2020, on constate qu'on peut approximer ces lois par des lois Beta. On implémente cela dans OpenTruns et on obtient le résultat du tableau 3. Le QQ-plot de la figure 9 confirme effectivement notre choix d'utiliser des lois Beta.

Paramètre	**alpha**	**beta**	**a**	**b**
S_0	0.9460	0.3721	2.52075e+07	2.52163e+07
I_0	0.4778	1.0105	-34.6078	3668.61
R_0	0.2528	0.7599	-47.4118	4883.41
D_0	0.2850	0.4967	-2.07843	214.078

Table 3 – *Estimation de lois Beta suivi par les paramètres initiaux par intuition au vu des observations de la figure 5.*

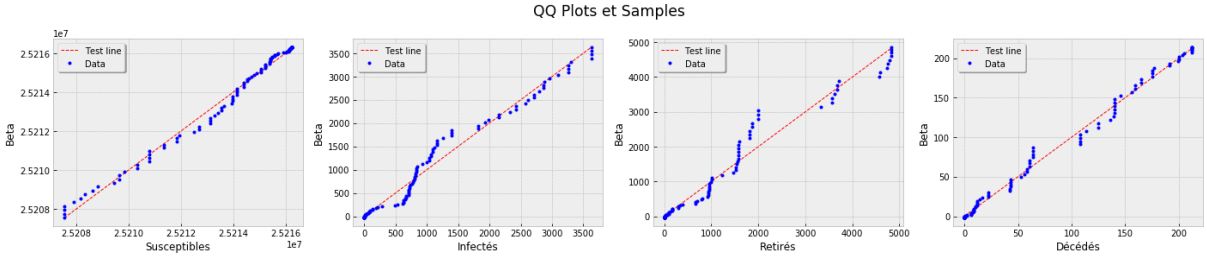


Figure 9 – *Confirmation du choix des lois Beta pour les conditions initiales à travers des QQ-plots.*

Connaissant les lois suivies par tout les paramètre en entré du nôtre modèle SIRD*, nous pouvons effectuer les analyses de sensibilité voulues. Nous commençons par une analyse de sensibilité par **développement de Taylor** au premier ordre. On obtient les facteurs d'importance ci-bas.

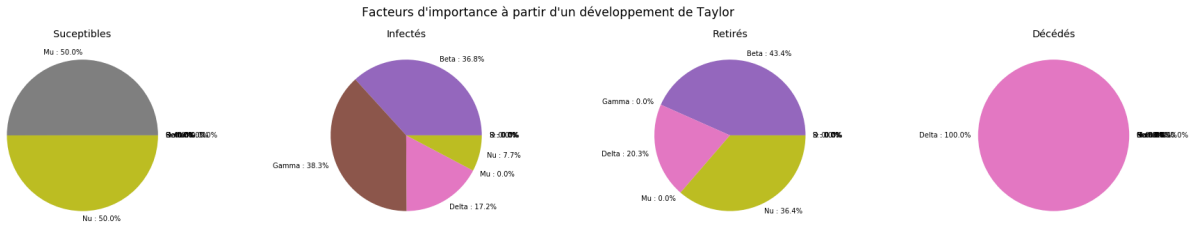


Figure 10 – *Analyse de sensibilité par développement de Taylor au premier ordre. On constate que les conditions initiales n'influent pas sur les sorties, ce qui est dû au caractère linéaire de ce développement limité.*

Ensuite, nous faisons une analyse de sensibilité par calcul des **indices de Sobol** au premier ordre. Les résultats obtenus sont présentés ci-bas [12].

11. Notons que dans le code fourni avec ce rapport, il s'agit des lois marginales de la distribution qui nous est retournée par l'estimateur linéaire décrit à la section précédente.
12. Le lecteur est redirigé vers le code accompagnant ce rapport pour les valeurs exactes de ces indices.

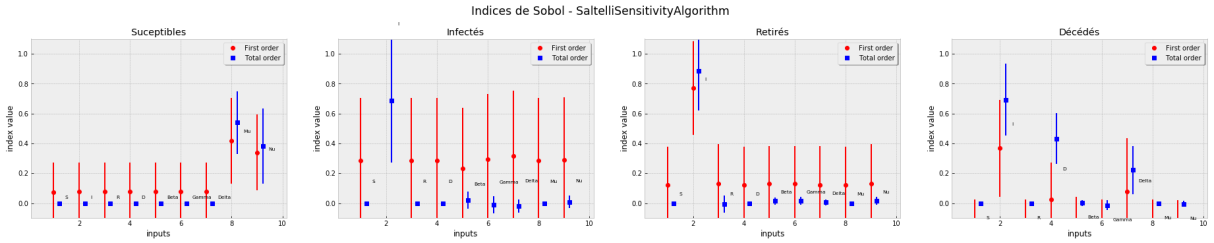


Figure 11 – *Analyse de sensibilité par calcul des indices de Sobol au premier ordre et totaux. Remarquons qu'en général, l'indice du premier ordre est différent de l'indice total correspondant, indiquant que les variables étudiés interagissent entre elles, comme nous l'avons montré au tableau 2.*

En analysant les figures 10 et 11 ci-haut, on constate que les paramètres les plus influents sur le nombre de cas susceptible dans la population Camerounaise sont ses taux de natalité μ et de mortalité ν. Le paramètre le plus influent sur le nombre d'infectés qu'on obtient à la fin d'une simulation est le nombre d'infectés initial. Ce même nombre est crucial pour la détermination des nombres finaux de guéris et de décédés. Remarquons aussi que le nombre de décédés est grandement influencé par plusieurs paramètres, parmi lesquels le taux de mortalité du virus δ.

III. Prédiction de l'évolution de l'épidémie

Pour prédire l'évolution du virus, nous calculons la distribution suivie par son taux de reproduction $\mathcal{R}_0 = \frac{\beta}{\gamma+\delta+\nu}$. Nous rappelons que les paramètres suivis par les lois β, γ, δ, et ν ont été estimés dans la partie 2 de ce rapport. La densité de probabilité de $\mathcal{R}_0$ est présentée ci-bas.

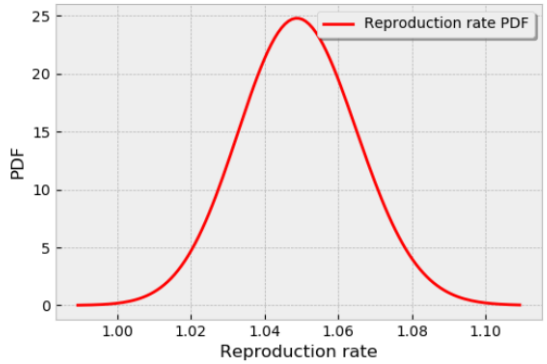


Figure 12 – *Densité de probabilité du taux de reproduction du virus. Cette loi visiblement normale a pour moyenne* 1.04928 *et pour écart type* 0.0160945*. Il y a de très fortes chances que le virus continue sa propagation, ce qui a effectivement été observé au Cameroun après le 14 juin 2020.*

Pour une prédiction plus ambitieuse, nous proposons de faire une simulation pour prédire l'état des différentes groupes 15 jours après la fin de notre fenêtre d'étude (du 06 mars au 14 juin 2020). En d'autres termes, nous prédisons l'état de la population Camerounaise le 29 juin 2020. On obtient la figure ci-bas.

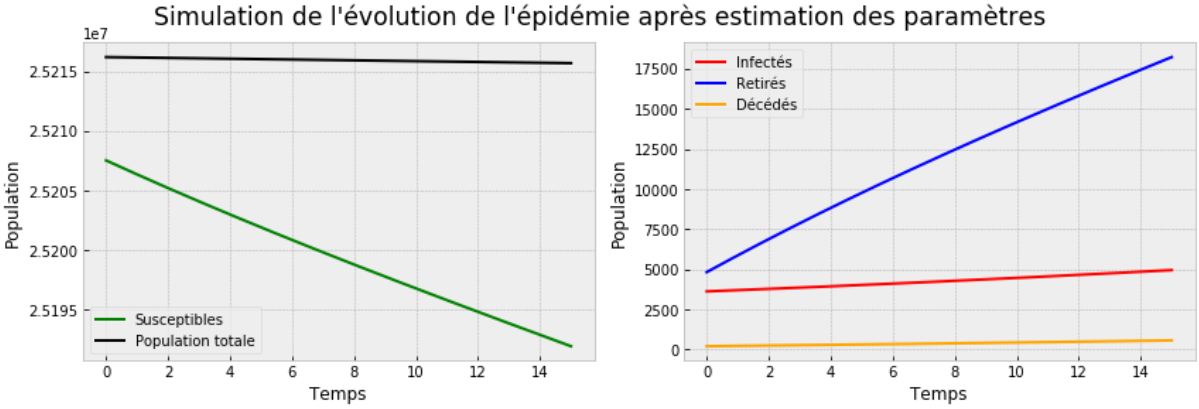


Figure 13 – *Prédiction de l'état de la population Camerounaise le 29 juin 2020. Les populations sont séparées en deux figures en raison des échelles de grandeurs très différentes. Cette simulation correspond effectivement à une très bonne approximation de ce qui a été observé au Cameroun à la date indiquée. En effet, partant de* $S_0, I_0, R_0, D_0 = 25207556, 3633, 4836, 212$ *le 14 juin 2020, nous avons prédit* $S_{pred}, I_{pred}, R_{pred}, D_{pred} = 25191938, 4962, 18251, 576$ *; et les quantités* $S_{true}, I_{true}, R_{true}, D_{true} = 25203645, 2179, 10100, 313$ *ont été observées. Naturellement, une prédiction plus loin dans le temps produit des résultats moins précis.*

Conclusion

En conclusion, nous avons étudié l'évolution du COVID-19 au Cameroun à travers un modèle SIRD incluant les taux de natalité et de mortalité de la population. L'analyse de sensibilité qui en découle a montré que le nombre de décès est influencé par plusieurs paramètres, ce qui indique que les autorités ont plusieurs options pour minimiser les dégâts liés à ce virus. Notre étude s'est focalisée sur la fenêtre du 06 mars au 14 juin 2020, durant laquelle l'épidémie sévissait à grande ampleur. Nous pourrions étendre cette étude avec des estimations et des prédictions de l'évolution de l'épidémie durant la période actuelle de calme relatif (septembre 2020-janvier 2021). On pourrait par exemple rechercher la date à laquelle l'épidémie va s'achever [13]. Idéalement, il faudrait rechercher la loi de probabilité des différents paramètres en fonction du temps et des mesures de limitations appliquées par les gouvernements.

13. Tout ceci est possible dans le code de calcul qui est fournie nommé "Code.ipynb" ; il suffit juste de changer la date de début de l'étude et la duré de l'étude à l'endroit indiqué.

Références

Bayette, C. et M. Monticelli (2020). « MODÉLISATION DUNE ÉPIDÉMIE, PARTIE 2 ». In : *Images des Mathématiques*. url : http://images.math.cnrs.fr/Modelisation-d-une-epidemie-partie-2.html?lang=fr#nh1.

CSSE (2021a). « COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) ». In : *Arcgis*. url : https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

— (2021b). « Time series summary ». In : *GitHub*. url : https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

Lisphilar (2020). « COVID-19 data with SIR model ». In : *Kaggle*. url : https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model#SIR-D-model.

Nguemdjo, Ulrich et al. (2020). « Simulating the progression of the COVID-19 disease in Cameroon using SIR models ». In : *PloS one* 15.8, e0237832. url : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237832.