Baud Candice
Guybert Baptiste
Taleb Ahmed Raja

ENSAE

Séminaire de modélisation statistique

Année scolaire 2022-2023

Implémentation d'un modèle mathématique pour la surveillance du Covid-19

Cours encadré par Mme Kaleghi *Projet supervisé par* Mr Marc Lavielle

Mai 2023

Table des matières

1 Introduction

L'année 2020 a été marquée par la pandémie de COVID-19, sur le plan médical, social et économique. Afin de minimiser l'impact sanitaire de cette pandémie, les gouvernements et les organisations de santé publique ont pris des mesures pour surveiller et contrôler la propagation de la maladie. La compréhension des dynamiques de cette pandémie s'avère donc clé dans la mise en place de mesures adaptées.

Différents modèles mathématiques ont été développés pour décrire la dynamique de la pandémie et prédire la situation épidémiologique future, notamment les modèles épidémiologiques de type SIR ou SEIR. Ces modèles sont complexes et nécessitent une calibration pour obtenir un bon ajustement entre les calculs du modèle et les données observées.

L'approche présentée dans l'article diffère en ce qu'elle vise à développer un modèle simple et robuste qui s'ajuste bien aux données plutôt qu'à imiter avec précision la dynamique de l'épidémie. Le but d'un tel modèle n'est pas de prédire l'évolution future de l'épidémie, mais plutôt de décrire la dynamique passée et de détecter un changement dans cette dynamique le plus tôt possible si cela se produit. Les données utilisées pour cette surveillance sont les admissions quotidiennes à l'hôpital et les décès signalés par Santé Publique France.

Le modèle statistique proposé combine différents effets tels que la dynamique épidémique, un modèle hebdomadaire et des fluctuations irrégulières. La dynamique des admissions à l'hôpital est décrite en supposant une dynamique exponentielle avec une fonction de taux définie de manière linéaire par morceaux. L'ajustement de ce modèle aux données consiste à détecter des points de changement dans les données d'admission. Le modèle ajusté permet d'identifier les différentes vagues épidémiques observées en France depuis mars 2020.

Le but de ce projet n'est donc pas d'implémenter un modèle classique à compartiments mais plutôt de s'intéresser à la détection d'un changement de dynamique dès que possible. La prévision future se restreint alors à la prédiction de ce qui devrait se produire dans un proche avenir si la dynamique de l'épidémie ne change pas. Ainsi, avec l'aide de Marc Lavielle [1], nous établissons un modèle basé sur les points de rupture des différences logarithmiques de nos indicateurs et construisons des régressions sur les différents segments définis par ces points de rupture afin de capter les changements de dynamique de l'épidémie.

1. Marc Lavielle Inria, Saclay, France and CMAP, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris Route de Saclay 91128 Palaiseau Cedex, France
E-mail : Marc.Lavielle@inria.fr URL : http ://www.cmap.polytechnique.fr/ lavielle/

2 Modèles mathématiques

2.1 Première approche du problème par un modèle compartimental

Un première approche au problème de la surveillance du COVID-19 est une modélisation mathématique déterministe du problème. Il est en effet possible de modéliser l'évolution de l'épidémie par un modèle compartimental de type SIR. Ce genre de modèles consiste simplement à diviser la population en plusieurs groupes disjoints représentant l'entièreté de la population d'intérêt. Chaque groupe correspond à un état particulier de la maladie (par exemple un modèle SIR divise la population en un groupe sain, infecté et guéri). Le modèle proposé par M. Lavielle est le suivant :

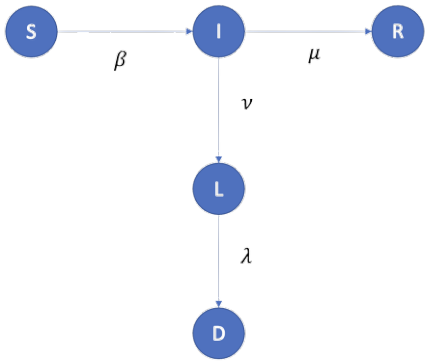

FIGURE 1 – Modèle compartimental

où $R(t)$ représente le nombre de personne saines, $I(t)$ le nombre d'infectés au temps t, $R(t)$ le nombre de personnes guéries, $L(t)$ le nombre de personnes hospitalisées et $D(t)$ le nombre de personne décédées. Dans ce cas, nous pouvons décrire les dynamiques de ce modèle comme suit :

$$\begin{aligned}
\dot{S}(t) &= -\beta S(t)I(t) \\
\dot{I}(t) &= \beta S(t)I(t) - \mu I(t) - \nu I(t) \\
\dot{R}(t) &= \mu I(t) \\
\dot{L}(t) &= \nu I(t) - \lambda L(t) \\
\dot{D}(t) &= \lambda L(t)
\end{aligned}$$

où β indique la proportion d'individus qui deviennent infectés, μ la proportion d'individus infectés qui guérissent, ν la proportion d'individus infectés qui sont hospitalisés, et enfin λ la proportion d'individus hospitalisés qui décèdent.

Dans le cas de l'épidémie de COVID-19 nos données ne concernent que les cas confirmés. Ainsi, il faudrait ajouter la variable $W_{\text{c}}(t)$ qui correspond au nombre de cas confirmés. Nous pouvons écrire sa dynamique ainsi :

$$\dot{W}_{\text{c}}(t) = \alpha \times \beta \times I(t) \times S(t)$$

où α est la proportion de personnes infectées confirmées.

Dans son modèle, Marc Lavielle compartimente la population entre les groupes $I_{\text{icu}}(t)$ [7] qui représente le nombre d'infectés en soin intensif, $I_{\text{Hosp}}(t)$, le nombre d'infectés en soin classique, $H(t)$ le nombre total de personnes hospitalisées, $D(t)$ le nombre de décès et enfin $O(t)$ le nombre de retour au domicile après avoir été guéri. La variation du nombre de patients hospitalisés suit donc la dynamique suivante :

$$H(t) = \dot{I}_{\text{Hosp}}(t) + \dot{I}_{\text{icu}}(t) - \dot{D}(t) - \dot{O}(t)$$

2.2 Approche du modèle via les données

Le modèle compartimental présente l'avantage d'être déterministe et d'être défini clairement par des équations mathématiques. Cependant, celui-ci est régi par un grand nombre de paramètres à estimer (ici α, β, μ, ν et λ), qui pour la plupart ne sont pas constants dans le temps. En effet, β dépend en réalité du temps : lorsque la population est confinée, celui-ci diminue, et inversement lorsque ce dernier est relâché. De même, α n'est pas une variable constante dans le temps, elle dépend du temps et du nombre de tests effectués.

L'approche effectuée ici est un modèle basé sur les données et non sur un modèle mathématique déterministe. Cela a l'avantage d'éviter d'estimer les paramètres pour fitter un modèle décrit mathématiquement. Le but est de pouvoir comprendre les tendances de l'épidémie à court terme en identifiant les points de rupture de la fonction $\dot{I}(t)$.

2.2.1 Définition des variables

Le modèle s'établit à l'aide des variables suivantes :

1. $z_{1,j}$: Nombre d'admissions en unité de soins normale pour le jour j.
2. $z_{2,j}$: Nombre d'admissions en soins intensifs pour le jour j.

3. $z_{3,j}$: Nombre de décès pour le jour j.
4. $z_{4,j}$: Nombre de sorties d'hôpital pour le jour j.
5. t_j : Le jour j.
6. $f_l(t_j)$: La tendance pour la l-ème série.
7. s_{lj} : Un composant périodique hebdomadaire.
8. ε_{lj} : Une séquence d'erreurs résiduelles.
9. $I_{ntw}(t)$: Nombre total de patients admis en unité de soins normale entre le temps t_0 et le temps t.
10. $I_{icu}(t)$: Nombre total de patients admis en soins intensifs entre le temps t_0 et le temps t.
11. $D(t)$: Nombre total de décès à l'hôpital entre le temps t_0 et le temps t.
12. $O(t)$: Nombre total de sorties d'hôpital entre le temps t_0 et le temps t.
13. $H(t)$: Nombre de patients présents à l'hôpital (en unité de soins normale ou en soins intensifs) au temps t.
14. $k_{ntw}(t)$: La fonction de taux pour les admissions en unité de soins normale, qui peut varier avec le temps.
15. $k_{icu}(t)$: La fonction de taux pour les admissions en soins intensifs, qui peut varier avec le temps.
16. $\gamma_{deaths}(t)$: Le taux de mortalité à un instant donné.
17. $\gamma_{out}(t)$: Le taux de sortie d'hôpital à un instant donné.

2.2.2 Modèle statistique

Le modèle statistique combine une tendance générale, un composant périodique hebdomadaire et des fluctuations irrégulières pour chaque série. Il est défini par l'équation suivante :

$$z_{lj} = f_l(t_j) + f_l^{\alpha_l}(t_j)(s_{lj} + \varepsilon_{lj}), \quad l = 1, 2, 3, 4. \tag{1}$$

2.2.3 Modèle dynamique

Le modèle dynamique utilise des dynamiques de type exponentiel pour les admissions, où les fonctions de taux peuvent varier avec le temps. Les équations (3) et (4) sont les suivantes :

$$I''_{ntw}(t) = k_{ntw}(t)I'_{ntw}(t), \tag{2}$$

$$I''_{icu}(t) = k_{icu}(t)I'_{icu}(t). \tag{3}$$

Les fonctions de taux sont supposées linéaires par morceaux :

$$k_{ntw}(t) = b_{ntw} + 2c_{ntw}t + 2\sum_{k=1}^{K_{ntw}} h_{ntw,k}\max(t - \tau_{ntw,k}, 0), \tag{4}$$

$$k_{icu}(t) = b_{icu} + 2c_{icu}t + 2\sum_{k=1}^{K_{icu}} h_{icu,k}\max(t - \tau_{icu,k}, 0). \tag{5}$$

Pour les décès et les sorties d'hôpital, les taux de mortalité et de sortie dépendent du nombre de patients hospitalisés à un instant donné :

$$D'(t) = \gamma_{deaths}(t)H(t), \tag{6}$$

$$O'(t) = \gamma_{out}(t)H(t). \tag{7}$$

Les auteurs supposent que les logarithmes de ces fonctions sont également des fonctions quadratiques par morceaux :

$$\log(\gamma_{deaths}(t)) = a_{deaths} + b_{deaths}t + c_{deaths}t^2 + \sum_{k=1}^{K_{deaths}} h_{deaths,k}\max(t - \tau_{deaths,k}, 0)^2, \tag{8}$$

$$\log(\gamma_{out}(t)) = a_{out} + b_{out}t + c_{out}t^2 + \sum_{k=1}^{K_{out}} h_{out,k}\max(t - \tau_{out,k}, 0)^2. \tag{9}$$

3 Modélisation des données

3.1 Choix des données

Afin d'avoir une modélisation pertinente de la dynamique de l'épidémie, il est avant tout nécessaire d'avoir des données de qualité. Nous utilisons les données hospitalières de la base de données SI-VIC disponibles au lien suivant : https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-

lepidemie-de-covid-19/. Nous utilisons le jeu de données quotidien par département, pour lequel nous agrégeons les données pour les avoir au niveau national. Ces données sont publiées par Santé Publique France, l'agence nationale de santé publique en France, créée en mai 2016 et sous tutelle du ministère de la Santé.

Nous supprimons les données correspondant aux DOM-TOM qui pourraient fausser l'étude, la dynamique de l'épidémie n'étant pas la même qu'en métropole.

Nous disposons des taux d'incidence :

- des hospitalisations
- des réanimations
- des décès
- et des retours à domicile (notés RAD)

Ces données sont présentées en figure 1 [2]. Elles présentent une saisonnalité hebdomadaire qui correspond à un problème de transmission et de récupération des données les week-ends. Cela crée artificiellement des pics assez importants le lundi et assez faibles le week-end comparativement à la moyenne sur la semaine. Afin de supprimer cette tendance hebdomadaire, nous effectuons la moyenne mobile sur 7 jours de nos indicateurs que nous représentons en rouge sur la figure [2]. Il est important de noter que la courbe des hospitalisations est légèrement décalée dans le temps par rapport à la courbe des infections : en effet, si une prise en charge hospitalière survient, c'est qu'une infection a préalablement été effectuée. Ainsi, une augmentation des admissions reflète une augmentation des contaminations dans les jours précédents, et inversement pour une diminution. En se concentrant sur les admissions, il est alors possible de détecter les changements de dynamique des infections.

Les données utilisées par M. Lavielle dans le papier [5] s'étendent du début de la série à fin 2022. Nous décidons d'étudier le reste de la courbe pour établir un modèle, c'est à dire, toute l'année 2022. Les données de 2023 serviront à comparer les prédictions du modèle avec les 'vraies' valeurs. La figure [9] en annexe présente la série restreinte à ces dates.

3.2 Modèle statistique implémenté

Afin d'analyser les données hospitalières, nous utilisons un modèle pour estimer les taux d'incidence des hospitalisations ($q_{\text{ntw},j}$), des réanimations ($q_{\text{icu},j}$), des décès ($\frac{d_j}{h_j}$) et des retours à domicile ($\frac{o_j}{h_j}$). Les variables d_j, o_j et h_j représentent respectivement le nombre de décès, de sorties et de patients hospitalisés à l'instant t_j. Le modèle comprend une détection de points de changement pour chaque

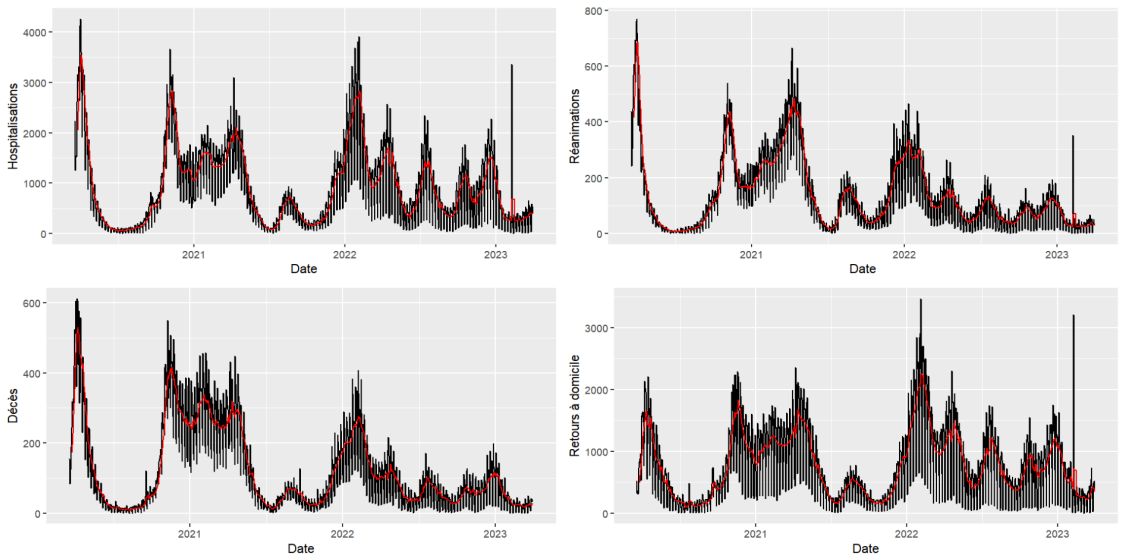

FIGURE 2 – Taux d'incidence bruts et avec moyenne mobile de 7 jours

série de données, en utilisant un critère des moindres carrés pénalisés pour estimer les paramètres du modèle et le nombre de points de changement.

Les équations suivantes décrivent les quatre séries de données :

$$\begin{aligned}
\log(q_{\text{ntw},j}) &= \log(I'_{\text{ntw}}(t)) + e_{\text{ntw},j} \\
\log(q_{\text{icu},j}) &= \log(I'_{\text{icu}}(t)) + e_{\text{icu},j} \\
\log\left(\frac{d_j}{h_j}\right) &= \log(\gamma_{\text{deaths}}(t_j)) + e_{\text{deaths},j} \\
\log\left(\frac{o_j}{h_j}\right) &= \log(\gamma_{\text{out}}(t_j)) + e_{\text{out},j}
\end{aligned}$$

Pour chaque série l, les variables $y_{l,j}$ sont définies comme suit :

$$y_{lj} = a_l + b_l t + c_l t^2 + \sum_{k=1}^{K_l} h_{l,k} \max(t - \tau_{l,k}, 0)^2 + e_{lj} \tag{10}$$

où a_l, b_l, c_l, $h_{l,k}$ sont des paramètres à estimer, et K_l et $\tau_{l,k}$ représentent respectivement le nombre et les instants de points de changement pour la série l. e_{lj} est le résidu pour la série l à l'instant t_j.

La fonction $f(t; \theta_K, T_K)$ est définie comme suit :

$$f(t; \theta_K, T_K) = a + bt + ct^2 + \sum_{k=1}^{K} h_k \max(t - \tau_k, 0)^2 \tag{11}$$

La minimisation du critère pénalisé $U(\theta_K, T_K, K)$ est effectuée pour estimer les paramètres et les

points de changement :

$$U(\theta_K, T_K, K) = \sum_{j=1}^{n} (y_j - f(t_j; \theta_K, T_K))^2 + \lambda K \tag{12}$$

Une valeur élevée du paramètre de pénalité λ favorise les configurations avec peu de points de changement, tandis qu'une valeur plus faible de λ autorise un nombre plus élevé de changements. Les paramètres et les points de changement sont estimés en minimisant le critère $U(\theta_K, T_K, K)$.

Pour une série de points de changement donnée T_K, la minimisation de U par rapport à θ_K est immédiate, car elle implique simplement le calcul de l'estimation des moindres carrés dans un modèle linéaire. Pour un nombre de changements donné K, cela se fait en fixant

$$\theta_K = \arg\min_{\theta} \sum_{j=1}^{n} (y_j - f(t_j; \theta, T_K))^2 \tag{13}$$

L'estimateur de T_K est alors défini comme

$$\hat{T}_K = \arg\min_{T_K} \sum_{j=1}^{n} (y_j - f(t_j; \hat{\theta}(T_K), T_K))^2 \tag{14}$$

Le nombre de changements K est choisi comme suit :

$$\hat{K} = \arg\min_{K} \sum_{j=1}^{n} (y_j - f(t_j; \hat{\theta}(\hat{T}_K), \hat{T}_K))^2 + \lambda K \tag{15}$$

L'estimation des points de changement comme défini en (10) est délicate. En effet, nous ne pouvons pas utiliser un algorithme de programmation dynamique car le critère à minimiser ne peut pas être décomposé en une somme de critères indépendants pour chaque segment en raison des contraintes de continuité sur f et sa dérivée.

Puisque les séries de données sont mises à jour quotidiennement, l'algorithme proposé est une procédure séquentielle qui nécessite peu de calcul. En fait, la configuration au jour $j+1$ est obtenue à partir des changements locaux dans la configuration obtenue au jour j. Supposons que $T^{(j)}$ soit la configuration optimale K obtenue au jour j. Nous calculons alors $T_K^{(j+1)}$ et $T_{K+1}^{(j+1)}$ comme les meilleures configurations avec K et $K+1$ ruptures, respectivement, lorsqu'il y a une nouvelle observation au temps t_{j+1}. Ces configurations sont obtenues par optimisation itérative, en changeant la position d'un seul point de changement à chaque itération. La meilleure de ces deux configurations est ensuite sélectionnée sur la base du critère pénalisé (11).

La valeur du paramètre de pénalité λ est ici ajustée manuellement afin que le résultat soit une segmentation qui "ressemble" visuellement à la segmentation que l'on créerait soi-même en regardant les données. Autrement dit, nous nous assurons que tous les changements que nous considérons comme significatifs sont bien détectés, tandis que les variations plus petites et plus irrégulières ne sont pas associées au signal, mais sont considérées comme des fluctuations aléatoires. Les résultats proposés ci-dessous ont tous été obtenus en choisissant $\lambda = 10^{-4}$.

3.3 Ajustement du modèle sur les données

Le modèle d'ajustement présenté dans l'article s'ajuste bien sur les séries des hospitalisations et réanimations, et nous nous concentrons ainsi dans la suite uniquement sur ces dernières.

3.3.1 Points de rupture

Dans un premier temps, afin de modéliser nos données, nous effectuons la différencielle logarithmique sur chacune de nos séries. Comme expliqué précédemment, celle-ci devrait être linéaire par morceaux.

Il est alors possible de définir les points de rupture de la courbe différenciée logarithmiquement. Nous l'effectuons dans un premier temps 'à la main' [8], c'est-à-dire en regardant 'selon nous' où ces derniers se trouvent. Les résultats sont déjà assez satisfaisants [10, 6.1], mais nous décidons d'effectuer une recherche plus précise des points exacts à l'aide d'une fonction recherchant les minimums et maximums locaux. Nous renseignons un intervalle dans lequel nous pensons que le point de rupture local se trouve et laissons l'algorithme nous donner le point exact. Une fois les points de rupture trouvés, nous pouvons *fitter* des régressions linéaires sur le logarithme de nos taux afin de retrouver l'expression de l'équation [11]. Les points de rupture et les tendances linéaires obtenus avec l'algorithme d'optimisation sont représentés dans la figure [3].

3.3.2 Régression linéaire

Une fois que les points de rupture ont été détectés, nous stockons alors ces derniers et ajoutons des variables à notre data set afin de pouvoir exécuter notre régression linéaire. Notre *design matrix* X s'écrit de la manière suivante pour chaque série que nous considérons [3.3.2] :

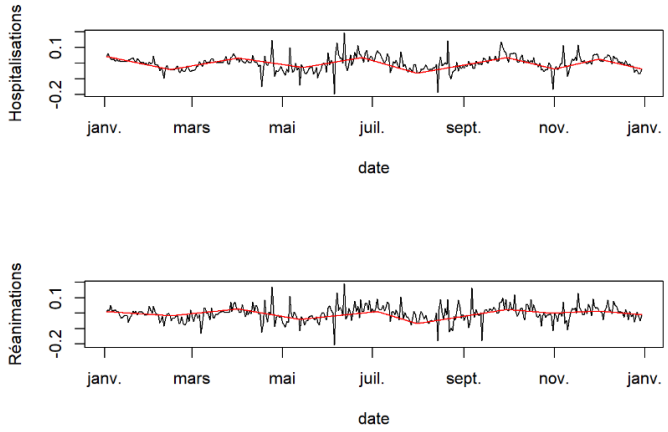

FIGURE 3 – Points de rupture optimisés

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & . & . & . & . & 0 \\ 1 & 2 & 0 & 0 & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & 1 & . & . & . & . & . & . \\ . & . & 2 & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & 1 & . & . & . & . & . \\ . & . & . & 2 & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 1 & n & n-\tau_1 & n-\tau_2 & . & . & . & . & n-\tau_K \end{bmatrix}$$

La première colonne correspond à la constante dans notre formule [11]. La seconde colonne représente une variable temps où la première date de la série est associée à 1, la deuxième à 2 etc... Les colonnes restantes correspondent aux coefficients associés aux points de rupture. C'est à dire, pour chaque τ_k (colonne k+2), les valeurs sont définies par $max(t - \tau_k, 0)$ où t est défini par la colonne 2.

Nous effectuons ensuite la régression linéaire pour chacun des indicateurs de $log(indicateur)$ sur une constante (colonne 1), la variable $temps$ (colonne 2), la variable $temps^2$ et les fonctions $max(t - \tau_k, 0)^2$ pour les τ_k points de rupture. Une fois que les coefficients sont estimés, nous avons donc la formule [11].

La formule s'écrit donc : $log(indicateur) = \hat{a} + \hat{b}t + \hat{c}t^2 + \sum_{k=1}^{K} \hat{h_k} \max(t - \tau_k, 0)^2 + \hat{\epsilon}$ où $\hat{\epsilon}$ correspond au résidu.

Les résultats des coefficients estimés sont disponibles en annexe [3, 4]. Ils apparaissent statistique-

ment significatifs à tous les seuils usuels. Les valeurs V_k apparaissant dans les tables correspondent aux $\hat{h_k}$. En passant à l'exponentielle, nous pouvons alors visualiser les courbes correspondantes.

3.3.3 Résultats obtenus sur les données

Nous choisissons K = 8 pour *fitter* le modèle sur les hopitalisations [4] et les réanimations [11]. La figure pour les hospitalisations est la suivante [4] :

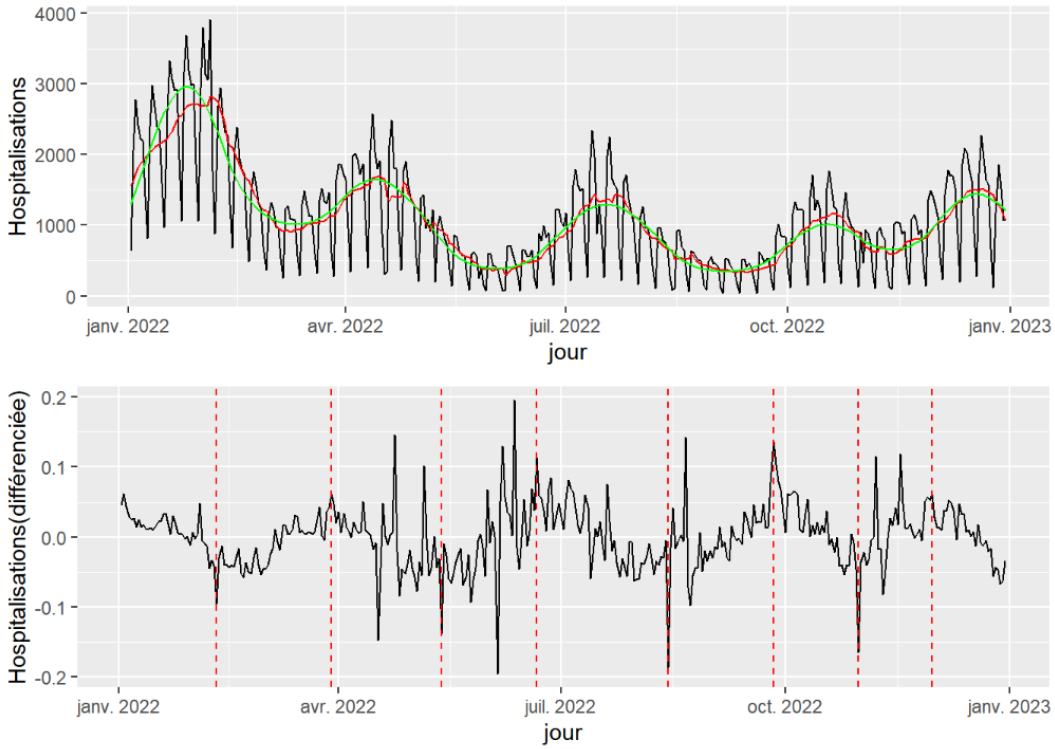

FIGURE 4 – Modèle ajusté sur les données et points de rupture - Hospitalisations

En noir sont représentées les séries initiales non corrigées des aberrations hebdomadaires. En rouge sont représentées les moyennes mobiles associées à chaque série, et en vert sont représentées les données ajustées au modèle. Il est intéressant de voir le lien entre ces points de rupture et l'évolution dynamique de l'épidémie [4, 11]. En effet, les pics de l'épidémie ne surviennent pas toujours simultanément aux points de rupture. Cette observation suggère l'impact d'autres facteurs, tels que les comportements individuels et les variations saisonnières, sur la propagation du virus. Les points de rupture peuvent également refléter les effets des interventions de santé publique, comme les restrictions de voyage, les campagnes de vaccination ou les mesures de distanciation sociale. L'examen de ces différentes représentations permet une meilleure compréhension des tendances et des effets des interventions de santé publique sur la dynamique de l'épidémie, mettant en évidence l'importance de la surveillance et de l'analyse de ces données pour orienter les politiques et les stratégies de

prévention. Les graphes plus détaillés pour chaque série sont représentés plus en détail en annexe [6.2].

3.3.4 Prédiction des prochaines valeurs

Le modèle établi précédemment est assez satisfaisant et nous pouvons alors tenter de prédire les prochaines valeurs de la série. La figure [5] montre les prédictions effectuées sur les quinze premiers jours du mois de janvier 2023 à l'aide du modèle établi sur l'année 2022. Le but de la prédiction étant de prédire les prochaines valeurs sans information additionnelle, nous ne pouvons donc pas prendre en compte les futurs points de rupture. Nous appliquons donc le modèle établi sur l'année 2022 avec le risque d'avoir un point de rupture survenant dans les prochains jours.

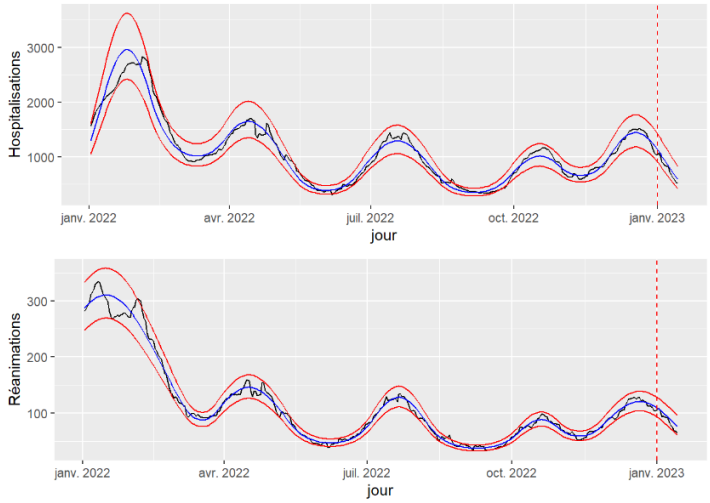

FIGURE 5 – Prédiction et intervalles de confiance

Il apparaît que les séries des hospitalisations et des réanimations sont prédites de manière plutôt correctes car les valeurs réelles sont toujours contenues dans l'intervalle de confiance à 95 %.

3.3.5 Résidus du modèle

Afin de regarder la qualité du *fit*, nous nous intéressons aux diagnostics de notre modèle. Pour rappel, le modèle que nous considérons [2.2.2] est exprimé ainsi :

$$z_{lj} = f_l(t_j) + f_l^{\alpha_l}(t_j)(s_{lj} + \varepsilon_{lj}), \quad l = 1, 2, 3, 4. \tag{16}$$

Ce qui signifie que $z_{l,j}$ contient une tendance globale f_l, une composante périodique s_{lj} et des résidus ε_{lj}. Notre but étant d'étudier les résidus sans la composante périodique, on définit :

$$w_{lj} = s_{lj} + \epsilon_{lj} = \frac{a_{lj} - \hat{f}_l(t_j)}{\hat{f}_l(t_j)^\alpha} \quad j = 1...n,$$
$$\hat{s}_{l,m} = \frac{1}{h} \sum_{k=0}^{h-1} w_{l,m+7k} \quad m = 1...7,$$
$$\hat{e}_{l,j} = \frac{z_{l,j} - \hat{f}_l(t_j) - \hat{f}_l^{\alpha_l}(t_j)\hat{s}_{lj}}{\hat{f}_l^{\alpha_l}(t_j)} \quad j = 1...n$$

Le paramètre α est choisi de manière à minimiser l'autocorrélation des séries $\hat{e}_{l,j}$ et $\hat{e}_{l,j+7}$. Nous choisissons $\alpha = 0.8$ comme dans l'article [5]. Nous pouvons alors représenter les séries $\hat{e}_{l,j}$, qui représentent nos résidus du modèle sans la composante périodique. Les résidus modifiés [6] ne semblent pas contenir de saisonnalité, ce qui était le but de nos opérations car nous voulions retrouver les résidus sans la composante périodique. Ils apparaissent gaussiens dans leur distribution, centrés et de variance constante, donc s'apparentant à un bruit blanc [12, 13].

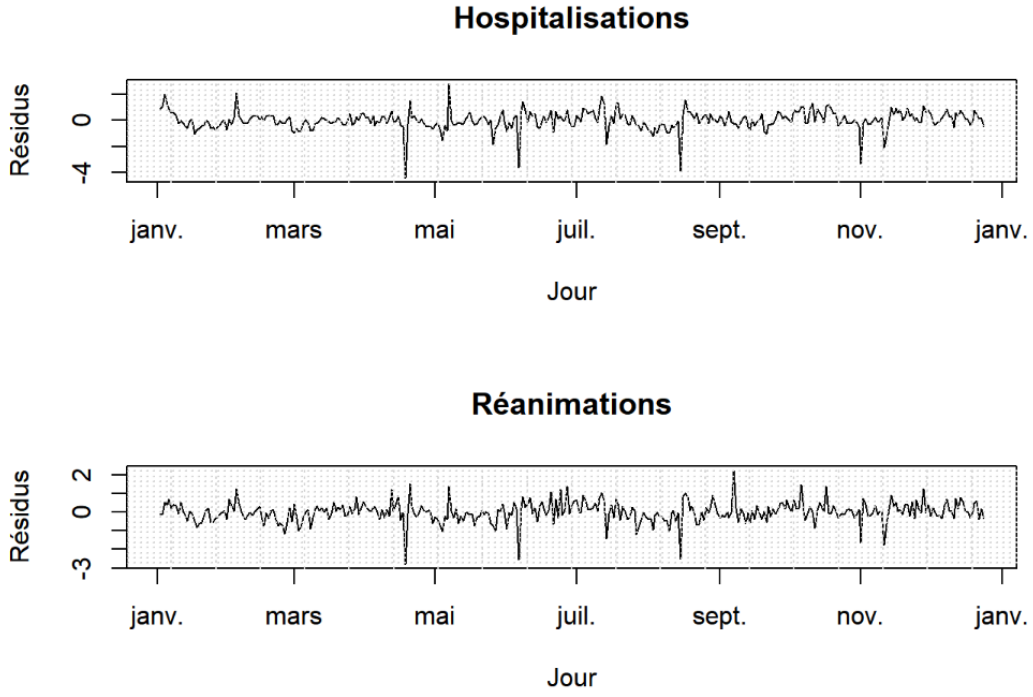

FIGURE 6 – Résidus

Au delà de la justification du modèle, l'analyse des résidus met en évidence l'effet qu'ont pu avoir les vacances ou certains longs week-end sur l'épidémie. En effet, nous observons des pics négatifs (par exemple en avril, en juin et en août) qui semblent indiquer une surestimation de notre modèle et donc un nombre de nouvelles entrées à l'hopital particulièrement faible par rapport à notre prédiction. Nous pouvons supposer que de nombreuses personnes ont fait plus attention aux gestes covid afin de s'assurer de partir en vacances ou dans l'idée de protéger leur famille. Cela pourrait expliquer ces valeurs plus faibles. Nous remarquons par ailleurs que la plupart de ces pics négatifs sont suivis par des valeurs plus élevées de nos résidus qui pourraient être dues à un plus grand laxisme pendant les vacances.

4 Conclusion et discussion

Ce projet apporte une modélisation plus originale du covid-19, basée sur les données et non sur un modèle mathématique dont on cherche à estimer les paramètres. Sa valeur ajoutée réside probablement dans sa simplicité d'estimation des paramètres et de fit du modèle, utilisant uniquement un modèle linéaire.

Nous avons uniquement déterminé localement les points de rupture à l'aide d'un algorithme basé sur notre propre observation des données. L'article [5] va bien plus loin en implémentant une détection automatique de ces points de rupture. Pour cela, il s'agit d'observer à quel moment les données observées quittent l'intervalle de confiance du modèle prédictif. Cela permet de déterminer et d'analyser à quels moments un changement de dynamique s'est opéré, ce qui est *in fine* un des objectifs de ce modèle.

En effet, le but de ce modèle n'est pas de prédire les dynamiques à venir mais d'expliquer *a posteriori* ce qu'il s'est passé. Dans ce cadre, ce modèle met parfaitement en évidence les points de rupture qui ont eu lieu : les moments où l'épidémie a explosé et ceux où elle a été contenue. En mettant ces informations en lien avec les politiques qui ont été menées, il est possible de mieux évaluer l'efficacité de certaines décisions gouvernementales et ainsi mieux réagir à l'avenir. Cependant, comme Marc Lavielle l'explique dans son rapport, il manquerait une analyse plus fine en fonction de certaines autres variables. L'âge, par exemple, a un rôle déterminant sur l'évolution du COVID-19 et ne pas inclure cette dimension au modèle peut manquer.

Ainsi, bien que des analyses pourraient être affinées, ce modèle apporte un éclairage certain sur l'évolution de l'épidémie et sur notre façon de la gérer.

5 Références et bibliographie

[1] *Using Hospital Data for Monitoring the Dynamics of COVID-19 in France*, Lavielle, M. (2022). Journal of Data science, Statistics and Visualisation

https://jdssv.org/index.php/jdssv/article/view/48

[2] *Github*, Baud Candice, Guybert Baptiste, Taleb-Ahmed Raja

https://github.com/candicebaud/modelisation_stat

Table des figures

6 Annexes

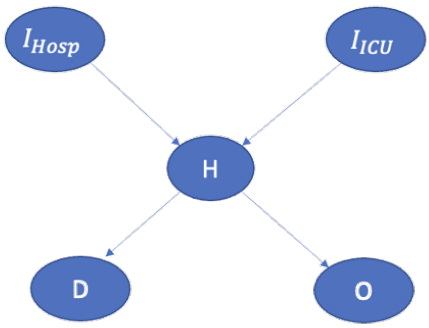

FIGURE 7 – Le modèle

6.1 Annexes de la sélection à la main

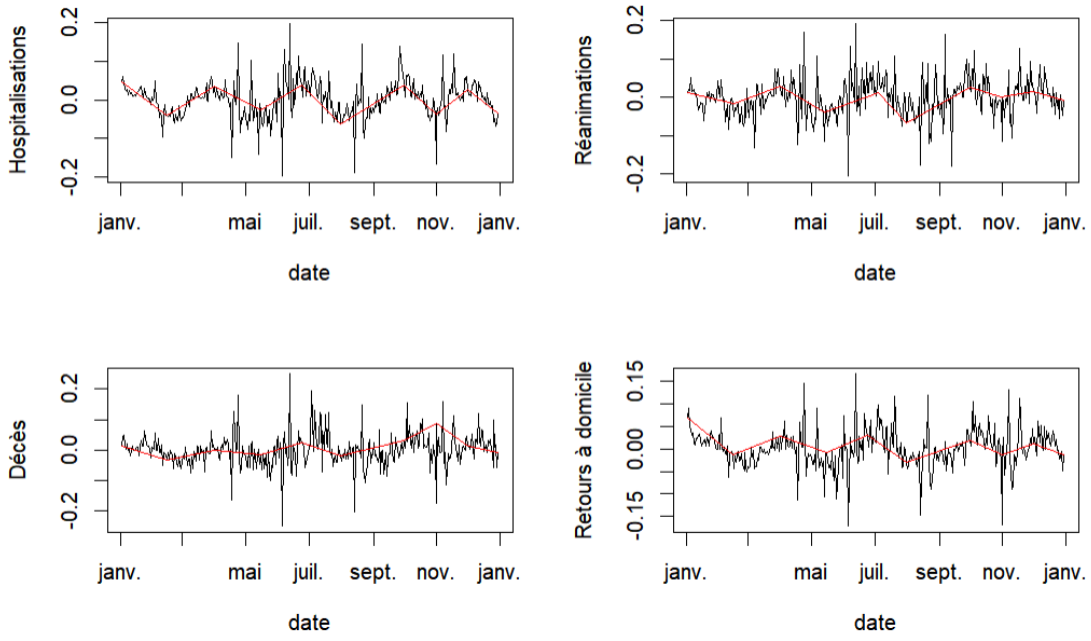

FIGURE 8 – Points de rupture des différentes séries

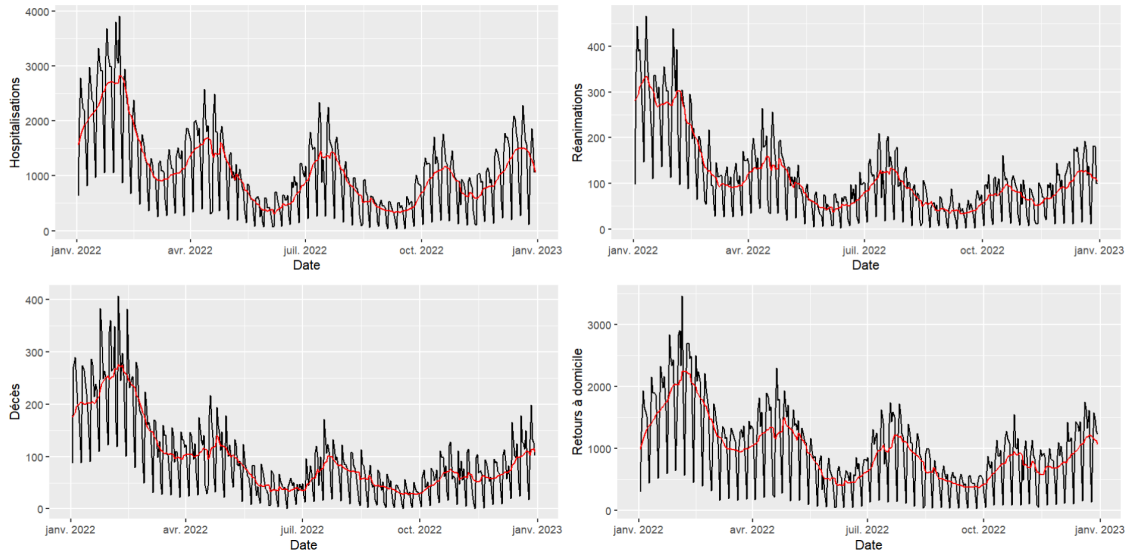

FIGURE 9 – Taux d'incidence bruts et avec moyenne mobile de 7 jours sur la période restreinte

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.2367	0.2853	25.37	0.0000
temps	0.0510	0.0211	2.41	0.0163
I(temps^2)	-0.0011	0.0003	-3.33	0.0009
I(V1^2)	0.0018	0.0005	4.02	0.0001
I(V2^2)	-0.0018	0.0003	-5.89	0.0000
I(V3^2)	0.0026	0.0003	8.30	0.0000
I(V4^2)	-0.0034	0.0003	-9.79	0.0000
I(V5^2)	0.0026	0.0003	9.53	0.0000
I(V6^2)	-0.0022	0.0004	-6.27	0.0000
I(V7^2)	0.0029	0.0007	4.51	0.0000
I(V8^2)	-0.0035	0.0012	-2.92	0.0037

TABLE 1 – Hospitalisations

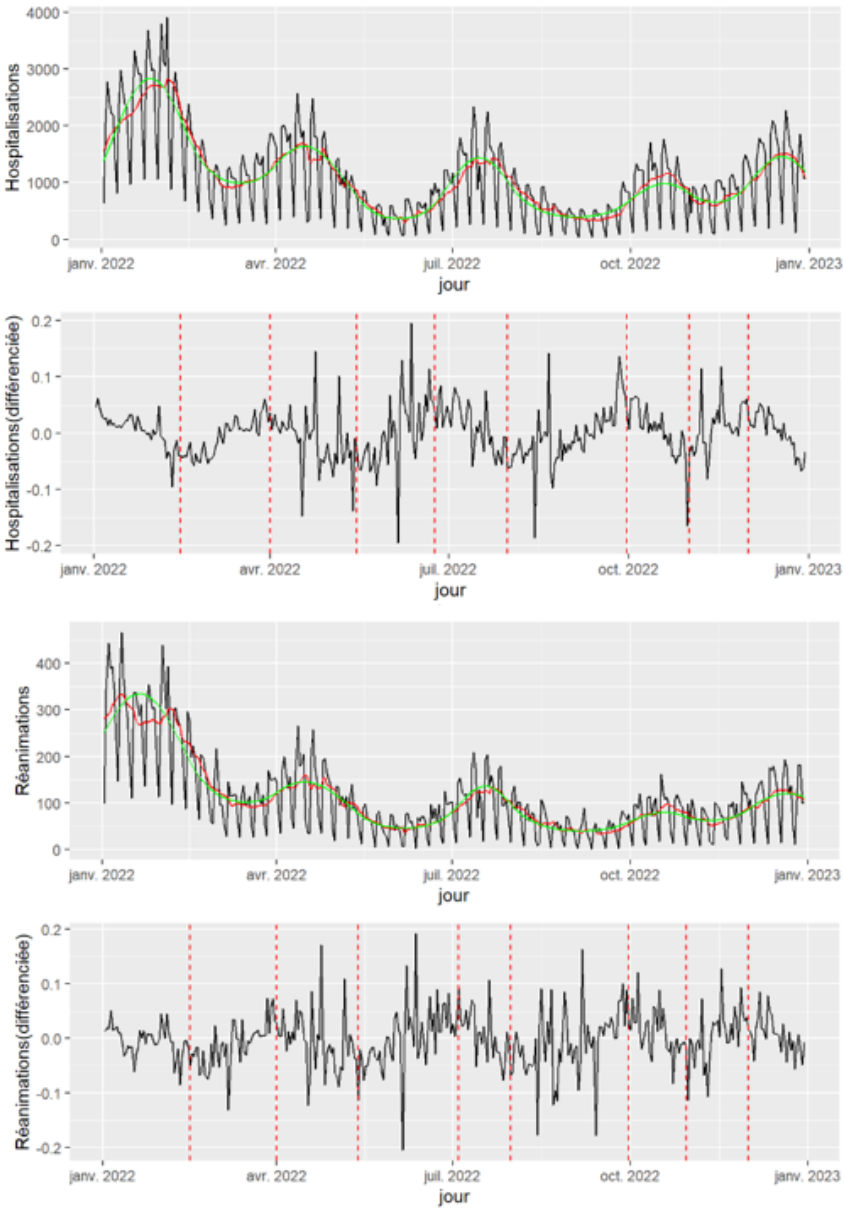

FIGURE 10 – Modèle ajusté sur les données et points de rupture pour toutes les séries

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.4594	0.3147	17.35	0.0000
temps	0.0285	0.0230	1.24	0.2150
I(temps^2)	-0.0008	0.0003	-2.32	0.0210
I(V1^2)	0.0015	0.0005	3.05	0.0024
I(V2^2)	-0.0016	0.0003	-4.59	0.0000
I(V3^2)	0.0018	0.0003	6.06	0.0000
I(V4^2)	-0.0032	0.0004	-7.67	0.0000
I(V5^2)	0.0030	0.0004	7.57	0.0000
I(V6^2)	-0.0020	0.0004	-4.59	0.0000
I(V7^2)	0.0022	0.0007	3.06	0.0024
I(V8^2)	-0.0022	0.0013	-1.70	0.0902

Table 2 – Reanimations

6.2 Annexes avec sélection optimisée des points de rupture

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0992	0.0425	166.90	0.0000
temps	0.0737	0.0034	21.73	0.0000
I(temps^2)	-0.0015	0.0001	-27.28	0.0000
I(V1^2)	0.0023	0.0001	31.31	0.0000
I(V2^2)	-0.0018	0.0000	-42.69	0.0000
I(V3^2)	0.0024	0.0000	56.17	0.0000
I(V4^2)	-0.0024	0.0000	-63.76	0.0000
I(V5^2)	0.0022	0.0000	59.05	0.0000
I(V6^2)	-0.0024	0.0001	-42.25	0.0000
I(V7^2)	0.0025	0.0001	27.46	0.0000
I(V8^2)	-0.0025	0.0002	-15.51	0.0000

Table 3 – Hospitalisations avec optimisation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.6476	0.0253	223.65	0.0000
temps	0.0122	0.0015	8.16	0.0000
I(temps^2)	-0.0004	0.0000	-22.57	0.0000
I(V1^2)	0.0023	0.0001	38.54	0.0000
I(V2^2)	-0.0027	0.0001	-46.84	0.0000
I(V3^2)	0.0017	0.0000	62.75	0.0000
I(V4^2)	-0.0022	0.0000	-75.91	0.0000
I(V5^2)	0.0022	0.0000	73.67	0.0000
I(V6^2)	-0.0023	0.0000	-45.93	0.0000
I(V7^2)	0.0028	0.0001	32.25	0.0000
I(V8^2)	-0.0021	0.0001	-18.31	0.0000

Table 4 – Réanimations avec optimisation

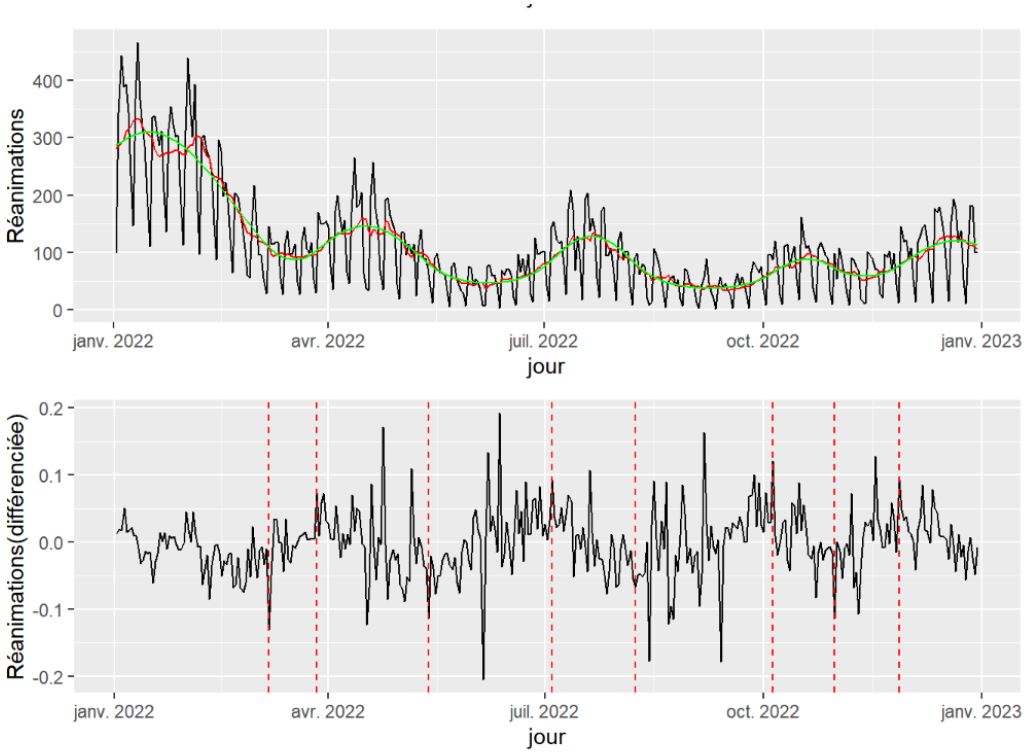

FIGURE 11 – Modèle ajusté sur les données et points de rupture - Réanimations

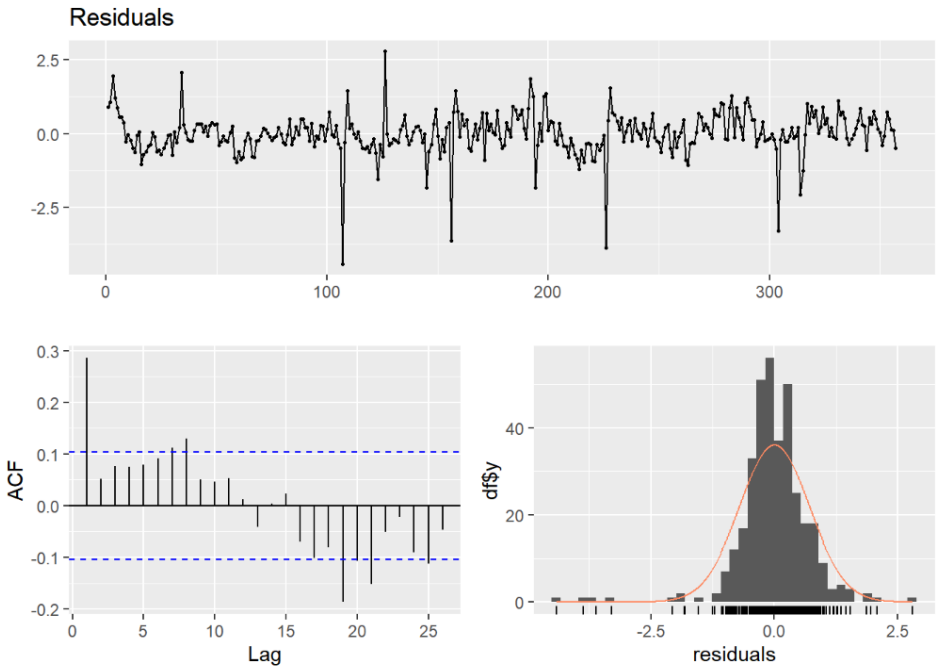

FIGURE 12 – Résidus modifiés hospitalisations

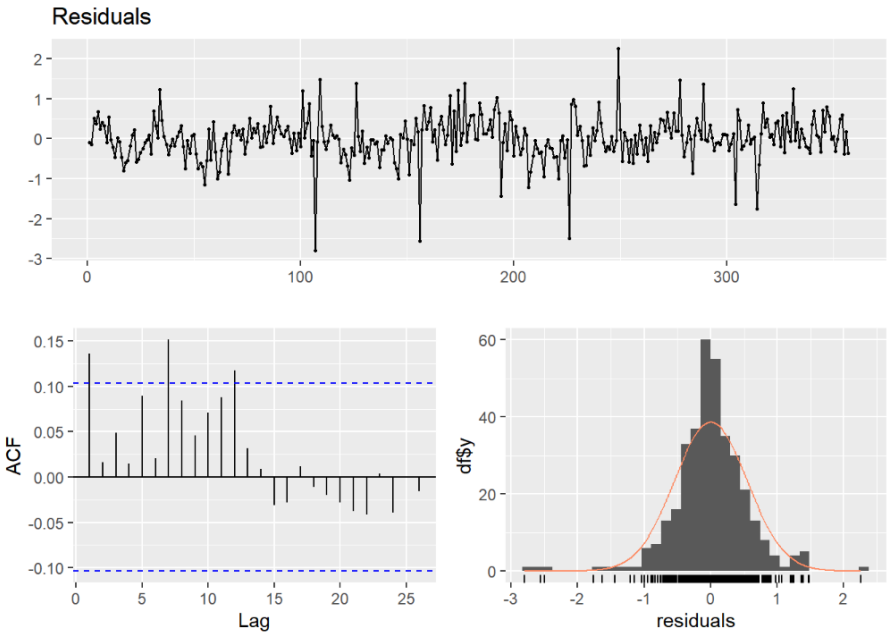

FIGURE 13 – Résidus modifiés réanimations