HEC MONTRÉAL

Analyse comparative et de cohérence entre les modèles statistiques et par compartiments sur la prédiction des cas d'infections et décès de la COVID-19 au Québec

Par

Emmanuel Penka Fowe

Aurélie Labbe

HEC Montréal

Directrice de recherche

Sciences des données et analytique d'affaires

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Avril 2022

DÉDICACE

Ma femme Jacqueline, Tchouya Nana

REMERCIEMENTS

Ce mémoire n'aurait pu exister sans le soutien et la participation de plusieurs personnes que je ne puis nommer sur une seule page. Ainsi, je tiens à remercier les personnes suivantes qui ont joué des rôles importants dans le processus de ce travail:

— La professeure Aurélie Labbe d'avoir accepté de diriger ce travail, je lui suis très reconnaissant pour ses conseils et sa présence continue tout au long de ce mémoire. Elle m'a donné assez de libertés pour suivre mes propres intérêts de recherche, me permettant de commettre des erreurs et d'en apprendre de ses impacts. Je ne pouvais imaginer une si meilleure personne pour ma maîtrise, et sans son bon sens, sa perspicacité à toujours veiller aux moindres détails, je n'aurai jamais fini ce travail.

— Mes professeurs de la HEC pour leur disponibilité à me fournir tout le support dont j'ai eu besoin pendant mes études.

— Je suis redevable à mes nombreux camarades de classe pour avoir contribué à fournir un environnement stimulant et amusant dans lequel nous avons appris les uns des autres. Je remercie particulièrement Christelle Nangne.

— Toute ma famille Manuelle Penka Océane, Nathan Djamen Penka, Laétitia Penka, Noémie Penka, Emilienne Matchinda, Donald Pekan, Calode Penka, Micheline Penka, Elyse Penka, Chantal Penka, Jacques Penka pour leur soutien constant tout au long de ce travail.

RÉSUMÉ

Ce travail porte sur la comparaison de la performance de plusieurs modèles de série chronologiques pour la prédiction des cas d'infections et de décès liés à la COVID-19 au Québec. En particulier, trois modèles statistiques (ARIMA, SARIMA et Prophet) et deux modèles par compartiments (SIRD, SIRF) sont analysés. Les métriques utilisées pour la performance de ces modèles sont basées sur la racine carrée de la somme des carrés d'erreur (RMSE) et la moyenne absolue des pourcentages d'erreur (MAPE). Les modèles ont été construits en utilisant les données quotidiennes libre d'accès de la Santé Publique du Canada couvrant la période d'avril 2020 à juillet 2021. Les données de l'échantillon de test couvrent la période de novembre 2020 à juin 2021. Nos résultats montrent qu'en général, la performance du modèle SARIMA est supérieure à celle de tous les autres modèles pour les infections tandis que pour les décès, le modèle SIRD performe mieux. Nos analyses ont aussi montré que la performance du modèle Prophet s'améliore significativement (amélioration pouvant être supérieure à respectivement 200 et 40% pour les cas de décès et d'infections d'après les valeurs RMSE ou MAPE) quand on réduit de 7 à 5 mois la taille de l'échantillon d'entrainement en faveur des observations plus récentes.

La cohérence entre ces modèles a été mesurée en comptant le nombre de fois que la majorité des modèles a prédit avec succès la tendance mensuelle future des infections et décès. Nous avons trouvé en moyenne 5 cas sur un total de 8 possibilités (62.5%) pour les infections et 6 cas sur un total de 8 possibilités (75%) pour les décès laissant suggérer un potentiel bénéfice à agréger la prédiction de tendance de plusieurs modèles pour améliorer les prises de décision à priori face à la pandémie actuelle. Toutefois, ces résultats doivent être pris avec du recul compte tenu du faible nombre d'observations et de modèles utilisés.

ABSTRACT

In this work, we compare the performance of several time series models for the prediction of COVID-19 related infections and deaths in Quebec. Three statistical models (ARIMA, SARIMA, and Prophet) and two compartmental models (SIRD and SIRF) are used. To compare the prediction accuracy with respect to reported cases or deaths, accuracy metrics such as Mean Square Error (MSE) and Mean Absolute Percent Error (MAPE) metrics are then used as performance criteria. Data, publicly available are sourced from Canada's Public Health website and include daily COVID-19 cases and deaths cases from April 2020 to June 2021. Test sample covers the period from November 2020 to June 2021 whereas April 2020 to October is used for training. Regardless of which performance metrics is used, we find that SAMIRA performs overall better in predicting infections cases, whereas SIRD performs better for deaths. Our experiment also shows that the performance of the Prophet model increases significantly (could be greater than respectively 200% and 40% for deaths and infections cases with regards to RMSE or MAPE metrics) when shrinking the size of the training sample from 7 to 5 months toward more recent observations.

Consistency between these considered models is assessed by counting how often most models successfully predict the monthly trend of infections and deaths cases. Our results show 5 cases out of 8 possibilities (62.5%) for infections and 6 cases out of 8 possibilities (75%) for deaths. This result suggests a potential benefit of aggregating the predictions from several models to improve decision-making in the face of the current pandemic. However, these results should be taken with caution as few observations and models were used. Therefore, obtaining more data could empower the model for further validation.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ARIMA	Autoregressive Integrated moving average
AIC	Akaike information criterion
COVID-19	Maladie à coronavirus 2019
SARIMA	Seasonal Autoregressive Integrated Moving Average
SIR	Susceptible, Infectious, or Recovered
SIRD	Susceptible-Infectious-Recovered-Deceased
ARCH	Autoregressive conditional heteroscedastic
ADF	Augmented Dickey Fuller

INTRODUCTION

Le '*Monde d'avant*', celui où les évènements tels que le confinement, la distanciation sociale, le passeport vaccinal, le port de masques, les couvre-feux, la quarantaine, la fermeture des écoles, etc. étaient si rares pour ne pas dire presque qu'inexistants, continue de faire rêver bon nombre parmi nous. Ce monde dont certains pourraient s'en rappeler comme d'un lointain souvenir, c'était-il y'a à peine deux ans. Avant que ne soit identifié à Wuhan (Hubei) en Chine [1], une nouvelle souche de coronavirus, le SRAS-CoV-2, responsable de la COVID-19 déclarée par L'Organisation mondiale de la santé [2] comme une pandémie le 11 mars 2020. Un changement d'autant brusque pour ce '*Monde d'avant'* que le président Français dira dans un discours à sa nation au 16 Mars 2020 '*Nous sommes en guerre'*, un bouleversement dont les impacts ont donné lieu à de graves défis économiques, sociaux, voir éthiques pour de nombreux pays, dont les plus développés, comme l'Italie [3] [4], qui suite à une disproportion entre ressources hospitalières (personnels et lits) d'accueil en soins intensifs et malades de la COVID-19, a eu à faire un '*tri*'. Pour certains de ces malades, on a tenté de les sauver et d'autres la seule alternative a été la sédation et bien pire sans doute. Comme tous les pays, le Canada et la province du Québec en particulier n'ont pas été épargnés. Avec un nombre record de décès dans les CHLSD [5], le taux de décès le plus élevé[1] au Canada par 100 mille habitants[2], la province du Québec connait une campagne de vaccination avec un taux record[3] de plus de 84%. Toutefois, face à l'affut de nouveaux variants du virus, la belle province bat de nos jours (décembre 2021) de nouveaux records d'infections et s'est résolue à réimposer de nouvelles restrictions incluant le couvre-feu depuis le 30 décembre 2021.

Au milieu de cette énorme incertitude quant à l'avenir de la pandémie de la COVID-19, des modèles prédictifs entre autres connaissent un engouement particulier pour aider les services de santé et les gouvernements à planifier et à contenir la propagation de la maladie.

---

[1] Québec compte environ 137/100 000 au 03 Janvier 2022 d'après Santé-Infobase Canada

[2] Source : https://sante-infobase.canada.ca/covid-19/resume-epidemiologique-cas-covid-19.html?stat=rate&measure=deaths&map=pt#a2; consulté en décembre 2021

[3] Source : https://www.quebec.ca/sante/problemes-de-sante/a-z/coronavirus-2019/situation-coronavirus-quebec/donnees-sur-la-vaccination-covid-19. consulté en décembre 2021

Ces modèles prédictifs ont été utilisés par le passé pour les maladies comme la Polio [6], Ébola [7] [8] et ont connu une avancée significative par suite de l'association des percées dans plusieurs domaines allant de la médecine, de la biologie moléculaire à l'informatique et aux mathématiques appliquées etc. Cependant, ces progrès ont fait naître de nos jours plusieurs types de modèles de prédiction de séries temporelles, basés sur des hypothèses et concepts théoriques parfois très différents qui rendent compliqués le choix du plus approprié pour une modélisation d'un problème donné.

Ainsi, les objectifs principaux de ce travail sont en premier lieu de comparer la performance des modèles statistiques et par compartiments sur les infections et les décès de la COVID-19 au Québec et en second lieu de regarder s'il existe une cohérence entre ces modèles pour la prédiction des tendances des infections et décès mensuels au Québec.

La différence que nous faisons dans ce travail entre un modèle statistique et par compartiments est que le premier fait appel aux notions de probabilités sur les données (et parfois sur les hypothèses) pour la prédiction alors que le second divise nos données en blocs comme les guéris, les infectés, les vaccinés, etc. et fait des prédictions en solutionnant les équations différentielles issus des interactions entre ces blocs.

Les détails sur ces modèles sont donnés aux prochains chapitres et notons ici que les modèles statistiques comprennent les modèles de séries chronologiques (ARIMA, SARIMA) et bayésiennes ('*Phophet*') alors que les modèles par compartiments comprennent les SIRD et SIRF. Les bases de données utilisées sont celles de John Hopkins[4] et du ministère de la Santé Publique[5] « Health-infobase » qui tous deux sont libre d'accès au grand public. Pour ce travail, nos données couvrent la période allant d'avril 2020 à juillet 2021. La cohérence dans ce travail consiste à regarder si collectivement la majorité des modèles étudiés est capable de prédire les tendances mensuelles futures d'infections et de décès au Québec.

---

[4] Les bases de données de Johns Hopkins University (JHU) sont sans aucun doute les plus populaires utilisées dans les modèles associés au COVID. Elles possèdent l'information 181 pays et sont mise à jour quotidiennement.

Source : Source: https://github.com/CSSEGISandData/COVID-19. Consulté en décembre 2021

[5] Source: https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html. Consulté en décembre 2021

La motivation pour ce travail est que l'épidémie de la COVID-19 continue à impacter notre société sur le plan sanitaire, social, économique dans le monde entier et ce travail pourrait donc contribuer à fournir des outils d'aide à la décision pour contenir le virus.

Ce travail est divisé comme suit. Dans le chapitre 1, nous présentons la littérature relative aux deux approches de modélisation ci-dessus mentionnées. Le chapitre 2 fournit un aperçu (théories, hypothèses) sur ces différents modèles et nous y décrivons brièvement les métriques utilisées pour évaluer la performance des modèles. Dans le chapitre 3, nous décrivons les données et les détails sur l'implémentation et le test de nos différents modèles. Au chapitre 4, nous comparons la performance de nos modèles sur une période de 8 mois en en déduisons la cohérence entre ces modèles. Nous terminons ce travail par une conclusion où nous résumons les résultats essentiels de ce travail, ses limites et énonçons quelques perspectives d'amélioration.

CHAPITRE 1 REVUE DE LITÉRATURE

Les articles disponibles dans la littérature traitant de la modélisation associée à la COVID-19 sont assez nombreux et ne pouvaient tous être cités dans ce travail. Cependant, certains d'entre eux ont eu plus d'impact ou ont joué le rôle de catalyseur dans nos décisions de choisir ce sujet. Nous en citons quelques-uns dans ce chapitre et donnons un aperçu des résultats principaux. Des références sont fournies pour des lectures supplémentaires sur les méthodes utilisées. Ce chapitre est divisé en deux parties comprenant les approches statistiques et par compartiments.

Au regard de la littérature et au moment où nous rédigions ce mémoire, nous n'avons pas connaissance de l'existence d'un travail similaire où sont comparés les approches statistiques et par compartiments sur les données de la COVID au Québec.

1.1 Approches statistiques

Nous avons distingué dans ce travail les approches dites statistiques classiques et Bayésiennes ('*Prophet*') ou aussi noté simplement Prophet. Les premières font appel aux notions de probabilités sur les données alors que les secondes comparées aux premières incluent en plus des lois de probabilité à priori sur certains paramètres du modèle. Les statistiques classiques seront représentées par les modèles autorégressifs (ARIMA, SARIMA) alors que l'approche Bayésienne est basée sur Prophet une librairie développée par Facebook de prévision de séries temporelles. Les détails sur la théorie sont donnés au prochain chapitre.

De nombreuses recherches impliquant l'utilisation d'ARIMA pour la prédiction des cas d'infections à la COVID-19 ont été menées ces récentes années. On peut citer entre autres les travaux de Sahai et al. [9], qui ont porté sur la prédiction des cas d'infections de COVID-19 dans cinq pays (États-Unis, Brésil, Inde, Russie et Espagne) avec les données de février 2020 à juin 2020. Leurs résultats ont montré que le modèle ARIMA performait assez bien quant à la prédiction des infections sur un horizon de 18 jours. Lee et al. [10] ont aussi montré qu'en analysant la variabilité (première dérivée des données journalières de cas d'infections) des données de la COVID pendant la période allant de janvier 2020 à décembre 2021, le modèle ARIMA permettait de simuler adéquatement (comparer aux cas observés) les infections observées. Toujours dans le sens de montrer la performance du modèle ARIMA, Alabdulrazzaq et al. [11] ont montré que sur

les données des cas de la COVID du Koweit (de février 2020 à mai 2020), les valeurs observées étaient bien dans les limites de la précision du modèle ARIMA. Poleneni et al., ont aussi montré que sur les données de la COVID en Inde de janvier 2020 à juillet 2020, le modèle ARIMA était assez performant pour prédire le nombre cas actifs pour un horizon de 15 jours. D'autres recherches ont aussi regardé l'importance de tenir compte de la saisonnalité dans les données de la COVID. Ainsi, les résultats du modèle SARIMA ont été comparés à ceux du modèle ARIMA pour 16 pays par Arunkumar et al. [12] Leurs résultats ont suggéré que SARIMA donne des résultats plus réalistes des cas d'infections, des décès et des guérisons liés à la COVID. Des résultats similaires ont été trouvé par Malki et al. [13] dont les modélisations ont suggéré dès la mi-2020 un pic d'infections entre décembre 2020 et avril 2021 pour un ensemble de 20 pays. Leur étude a prévu également un boom de la pandémie si les mesures prises ou les précautions étaient complètement assouplies. L'approche bayésienne dans ce travail est faite avec la librairie Prophet [14]. Cette dernière, développée par l'équipe Data Science de Facebook, est relativement nouvelle dans la littérature. Ainsi, Suryaningsih Patandung and Ihsan Jatnika [15] ont récemment prédit avec Prophet qu'en ce contexte de pandemie, le taux de croissance des touristes étrangers en Inde devrait baisser jusqu'en février 2022 et remonter vers mars 2022. Satrio et al. [16] ont montré sur les données de COVID de l'Indonésie de janvier 2020 à mai 2020 qu'en général, ARIMA performait moins bien que Prophet pour un horizon de quatre semaines. De même, Khayyat et al. [17] ont noté que Prophet avait une bonne habileté à prévoir les décès sur les données de l'Arabie Saoudite, mais toutefois avait une faible capacité à prédire les cas de guérison. Tulshyan et al., ont analysé les infections et décès liés à la COVID-19 en Inde pendant et après la période de couvre-feu du 24 mars 2020 au 24 mai 2020. Les résultats ont montré que Prophet performait bien pendant le couvre-feu avec une précision autour de 87% et cette dernière chuterait autour de 60% après le couvre-feu.

1.2 Approches par compartiments

Les modèles épidémiologiques qui font appel à la division d'une population en compartiments sont souvent appelés modèles par compartiments [18] [19] [20] [21] [22]. Anastassopoulou et al. [23] ont montré qu'un modèle par compartiments à partir d'un échantillon de population subdivisée en susceptibles (S), infectés (I), guéris (R) et décédés (D), SIRD, était assez adéquat pour la prédiction de la fin de la pandémie à Hubei (Chine) en fin février 2020. Dans le même esprit, Cooper et al.

[24] ont montré que l'application des modèles SIR sur les données de certain pays (Chine, Corée du Sud, Inde, Australie, Italie, Texas au États-Unis) suggère que la pandémie aurait pu rapidement être sous contrôle dans tous ces pays si des restrictions appropriées et des politiques solides avaient été mises en œuvre pour contrôler les taux d'infection dès le début de la propagation de la maladie. Des variants du modèle SIR qui incluent un paramètre reflétant la distanciation sociale ont été utilisés par Bastos et Cajueiro [25], les résultats pour le Brésil ont confirmé l'intuition que la distanciation sociale aurait pu freiner le schéma d'infection de la COVID-19. Mieux encore, si cette contrainte n'était pas appliquée suffisamment longtemps, la conséquence aurait été un simple déplacement du pic de la pandémie. Feng et al., [26] ont montré que les modèles formés à partir des compartiments susceptibles (S), Exposés (E), infectés (I), guéris (R), SEIR étaient assez efficaces pour prédire les pics et tailles de l'épidémie de la COVID à Wuhan (en Chine). Mieux encore, les modélisations ont montré que les mesures de contrôle de l'épidémie prises par le gouvernement ont significativement réduit l'ampleur de la pandémie.

CHAPITRE 2 DESCRIPTION DES MODÈLES

Plusieurs modèles mathématiques sont de plus en plus utilisés pour comprendre la transmission des maladies infectieuses et évaluer l'impact potentiel des programmes (généralement gouvernementaux) de contrôle ou de réduction des effets néfastes liés à la propagation de la maladie. Ce chapitre vise à fournir, d'une part, un bref aperçu de certains de ces modèles sélectionnés pour ce mémoire, et d'autre part, à mettre en évidence leurs aspects intuitifs et conceptuels. Pour cela, le chapitre est divisé en deux grandes parties dont l'essence de la première porte sur la brève description théorique des modèles et leur utilisation alors que la seconde partie s'intéresse aux métriques utilisées pour la mesure de la performance des modèles.

La manipulation des équations mathématiques sous-jacentes à ces modèles est ignorée et des références sont fournies pour une lecture plus approfondie. De même, pour palier à toute erreur de traduction qui pourrait potentiellement changer le sens de certaines phrases, nous conservons certains mots dans la langue originale 'anglaise' de leur document de référence.

2.1 Détails sur les modèles

L'un des objectifs principaux de ce travail étant de comparer les modèles, nous présentons ici les détails sur les approches statistiques et par compartiments utilisées. Notons que dans le cadre de ce travail, notre choix s'est porté sur les modèles relativement simples d'application de chacune de ces approches (les détails sont donnés au Chapitre 4) afin de rendre plus interprétable des résultats y afférents.

2.1.1 Modèles statistique classique

Ici, nous présentons sommairement les modèles ARIMA et SARIMA où une série temporelle (ou chronologique) est définie par une suite de données $(Y_1, \dots, Y_T)$ représentant la valeur de la série au temps t et T étant la longueur de la série.

2.1.1.1 Modèle ARIMA

En statistique, les modèles '*AutoRegressive Integrated Moving Average'* ARIMA, ou aussi modèles de Box-Jenkins [27], sont parmi les plus utilisés en prévision sur les données de type séries temporelles [28] [29]. Conceptuellement, un modèle ARIMA est formé de trois éléments :

- Une première composante '*Autoregressive*' (AR) qui explique la corrélation entre la valeur actuelle de la série chronologique et certaines de ses valeurs passées. Habituellement, p représente l'ordre autorégressif (AR) ou le nombre de termes autorégressifs. Le modèle $AR(p)$ prend la forme de

$$Y_t = c + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + u_t = c + \sum_{i=1}^{p} \alpha_i Y_{t-i} + u_t. \quad (2.1)$$

  Les $\alpha_i (i = 1, .., p)$ sont les paramètres du modèle, p est un nombre entier positif, l'ordonnée à l'origine est représentée par c, et $u_t$est un bruit blanc de type Gaussien.

- La seconde composante '*Moving Average*' (MA) représente la durée de l'influence d'un choc aléatoire inexpliqué [30]. Habituellement, q représente l'ordre de la moyenne mobile où le nombre d'erreurs de prévision retardées dans l'équation de prévision. Le processus $MA(q)$ est défini comme suit :

$$Y_t = \mu + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \cdots + \beta_q u_{t-q} + u_t = \mu + \sum_{j=1}^{q} \beta_j u_{t-j} + u_t. \quad (2.2)$$

  Les termes d'erreur u_t sont supposés être du bruit blanc, les $\beta_i (i = 1, .., q)$ sont les paramètres du modèle, q est un nombre entier positif, et μ est une constante.

  Très souvent, les modèles AR (P) et MA(q) sont manipulés en utilisant les opérateurs de retard 'Lag'. Cet opérateur [31] [32] est défini par la lettre L et s'applique comme suit :

$LY_t = Y_{t-1}$. En appliquant deux fois cet opérateur, cela donne $L(LY_t) = L^2Y_t$. Les polynômes d'opérateurs de retard s'utilisent ainsi pour représenter les modèles MA(q) et AR(p) comme suit [32] :

- AR(p) : $u_t = \alpha(L)Y_t$
- MA(q) : $Y_t = \beta(L)u_t$.

Ici, α et β sont des fonctions d'opérateurs de retard et sont définis comme suit : $\alpha(L) = 1 - \sum_{i=1}^{p} \alpha_i L^i$ et $\beta(L) = 1 + \sum_{j=1}^{q} \beta_j L^j$

- La troisième composante '*Integrated*' (I) représente la quantité de différenciation à effectuer sur la série temporelle pour la rendre stationnaire. Habituellement, d représente l'ordre de différenciation ou le nombre de différences non saisonnières nécessaires à la stationnarité. La différence de second ordre est montrée dans l'équation suivante :

$$Y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2.} \quad (2.3)$$

- En rappel, la notation L^j veut dire que l'opérateur L est appliqué j fois.

En utilisant l'opérateur de retard, l'expression mathématique du modèle ARIMA pour une série non-stationnaire est donnée comme suit [32] :

$$\alpha(L)(1 - L)^d Y_t = \beta(L)u_t, \quad (2.4)$$

c.-à-d.

$(1 - \sum_{i=1}^{p} \alpha_i L^i)(1 - L)^d Y_t = (1 + \sum_{j=1}^{q} \beta_j L^j)u_t$.

En raison des trois entiers distincts (p, d, q) utilisés pour paramétrer les modèles ARIMA, ces derniers sont notés avec la notation $ARIMA\ (p, d, q)$.

2.1.1.2 Modèle SARIMA

Pour tenir compte des effets saisonniers, on utilise souvent le modèle $ARIMA$ saisonnier, qui est noté

$$SARIMA\ (p,d,q)(P,D,Q)_m \tag{2.5}$$

Ici, (p,d,q) correspondent aux paramètres non saisonniers décrits ci-dessus, tandis que (P,D,Q) suit la même définition mais appliquée à la composante saisonnière de la série chronologique. Le terme m est l'ordre de la saisonnalité ou la périodicité de la série temporelle (4 pour les périodes trimestrielles, 12 pour les périodes annuelles, etc.).

Dans l'équation ci-dessus, le terme de saisonnalité est représenté par trois nombres additionnels P, D, Q définis comme suit :

- P représente le terme autorégressif de la saisonnalité
- D exprime le terme de différence ou d'intégration de la saisonnalité
- Q représente le terme de la moyenne mobile de la saisonnalité

De plus amples détails sont fournis dans [33] et [34]. Ainsi, si par exemple pour illustrer le terme de différence que nous appelons ici y_t, nous avons

- D=0, D=1; alors $y_t = Y_t - Y_{t-m}$,
- D=1, D=1; alors $y_t = Y_t - Y_{t-1} - Y_{t-m} + Y_{t-m-1}$,

Très souvent, D n'est pas plus grand que 1, la somme d+D inférieure ou égale à 2, et lorsque la somme d+D=2, la constante (ordonnée à l'origine) est ignorée.

Par exemple, un modèle SARIMA (1,1,1)(1,1,1)4 aura l'expression ci-après (sans constante)

$(1-\alpha_1 L)(1-\alpha_1 L^4)(1-L)(1-L^4)Y_t = (1+\beta_1 L)(1+\beta_1 L^4)$.

Ici, dans le membre de gauche, le premier terme est un AR(1) sans saisonnalité, le second terme est AR(1) avec saisonnalité, le 3^e est le terme d'intégration sans saisonnalité, le 4^e est le terme d'intégration avec saisonnalité. Dans le membre de droite, le 1er terme est un MA(1) sans saisonnalité alors que le second terme est un MA(1) avec effet saisonnier. On en déduit que

$$Y_t = (1-\alpha_1)Y_{t-1} - \alpha_1 Y_{t-2} + (1-\alpha_1)Y_{t-4} - (1 + (1-\alpha_1 L^4)(1+\beta_1 L)(1+\beta_1 L^4).$$

2.1.2 Modèle statistique Bayésien: Facebook Prophet

Prophet est une librairie utilisée pour bâtir un modèle de prévision de séries chronologiques basé sur le modèle de régression additif [35] [36] et développé par l'équipe Data Science [14] de Facebook (par la suite, nous utiliserons Prophet pour désigner le modèle issu de la librairie de Prophet). La méthode Prophet utilise un cadre appelé « Analyst-in-the-Loop » comme illustré sur la Figure 2.1 (source Facebook Prophet [14]**)** ci-dessous.

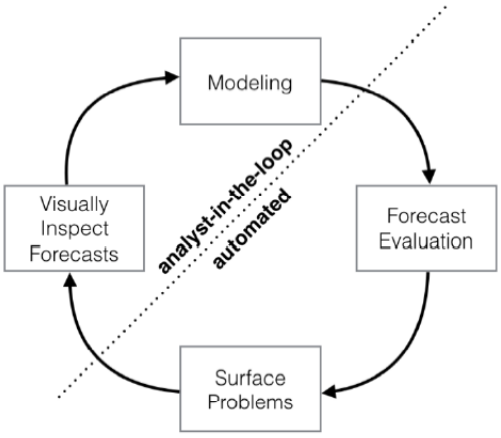


Figure 2.1 Schéma illustratif d'‘Analyst -in-the-Loop' de Prophet.

Le principe de modélisation de Prophet présenté dans la Figure 2.1, repose sur son approche à deux possibilités où, d'une part, l'ajustement du modèle est automatisé [37] et requiert aucune ou peu de connaissances statistiques, et d'autre part, son concept est assez flexible pour permettre à l'utilisateur d'y inclure les informations qu'il juge pertinentes en fonction de sa connaissance ou son expérience dans le domaine.

Le modèle Prophet peut être divisé en trois composantes principales : la tendance $g(t)$, la saisonnalité $s(t)$ et un effet de « vacances » ou des événements spéciaux $h(t)$ comme indiqué ci-dessous.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t. \quad (2.6)$$

Ici, $y(t)$ est la valeur de la série chronologique que nous observons au temps t, et ϵ_t est l'erreur du modèle.

Dans le cadre ce travail, nous avons choisi de garder uniquement les termes $g(t)$ et $s(t)$ dans notre modélisation[6]. De même, à cause de la nature Bayésienne associée au modèle, ses paramètres dépendent du degré initial de croyance ou de connaissance apriori qu'on se fait de leur distribution de probabilité qui correspond généralement à celle d'une une loi gaussienne. Nous utiliserons le terme 'Prior' ou loi a priori par la suite. La distribution a postériori utilisée par Prophet est de la même famille de distribution de probabilité que la probabilité a priori (les deux distributions sont dites conjuguées) et est donnée comme suit

$$\theta|y \sim N(g(t) + s(t), \epsilon_t). \quad (2.7)$$

θ est l'ensemble des paramètres qui décrivent $g(t)$et $s(t)$et pour lesquelles on définit une loi d'a priori pour chacune. L'Eq. 2.7 ayant une forme analytique (les distributions des probabilités d'a priori et d'a postériori étant de la même famille de loi de probabilité), Prophet offre la possibilité de la résoudre par une approche classique basée sur les techniques itératives simples (Newton-Raphson, gradient-conjugué, etc.) de maximisation de la fonction de posteriori (MAP) ou à travers une inférence bayésienne par les méthodes de Monte-Carlo à chaînes de Markov, ou les méthodes MCMC. Pour l'approche MCMC (non retenu dans ce travail), le temps de calcul est relativement long et l'usager devra spécifier le nombre d'échantillons à considérer.

[6] Le terme $h(t)$ reflète la modélisation des évènements spéciaux ('*Boxing Day*', '*Black Friday*', 'Thanksgiving', etc.) ou vacances. Bien que ces derniers puissent être des jours assez favorables pour le regroupement de la population, ils devraient être inclus comme étant des variables de type binaire dans le modèle. Toutefois, afin de garder la forme du modèle relativement simple, nous avons choisi de les exclure pour l'instant. Ce choix a été aussi fait par d'autres auteurs [58], [59].

2.1.2.1 Modèle de tendance g(t)

Pour modéliser le terme de tendance, la série chronologique est souvent divisée en plusieurs périodes ou fenêtres. Prophet offre le choix d'utiliser pour chacune de ces fenêtres soit une fonction logistique ou une fonction linéaire.

2.1.2.1.1 Modèle de croissance logistique

Prophet utilise un modèle de croissance logistique qui est un cas particulier de courbes de croissance logistique généralisées. Ces dernières sont très souvent utilisées pour appréhender l'évolution de la population dans bien des domaines liés au développement socio-économique et démographique [38] [39] [40] [41]. En bref, le modèle assume que le taux de croissance d'une variable d'intérêt devient presque nul au fur et à mesure que la fonction de la variable s'approche de sa valeur maximale. Ainsi, il est exprimé comme une courbe de type sigmoïde [37] dont l'expression est :

$$\mathrm{g(t)} = \frac{\mathrm{C(t)}}{1 + \exp\left[-(k + \boldsymbol{a}(t)^T \boldsymbol{\delta})\left(t - (m + \boldsymbol{a}(t)^T \boldsymbol{\gamma})\right)\right]}. \quad (2.8)$$

Les paramètres du modèle sont donnés ci-dessous:

- $\mathrm{C(t)}$: capacité de charge temporelle. Les valeurs maximales doivent être spécifiées.
- k: taux de croissance
- m: paramètre de décalage (*offset parameter*)
- $\boldsymbol{\delta}$: vecteur utilisé pour les ajustements de taux de croissance
- γ: vecteur dont chaque élément, γ_j, est défini comme $-s_j \delta_j$ ou s_j est le point de changement (appelé *changepoints* par la suite) ou de rupture délimitant chaque fenêtre de la série temporelle, et δ_j est le changement de taux qui se produit au temps s_j.
- $\boldsymbol{a}(t) \epsilon \{0,1\}^S$, tel que $a_j(t) \begin{cases} 1, & if\ t \geq s_j \\ 0, & sinon \end{cases}$

Notons que pour les valeurs de $\mathrm{C(t)}$, on peut se fier sur nos connaissances expertes sur le sujet ou utiliser toutes valeurs subjectives ou attendues. Ainsi, l'utilisateur fournit la valeur minimale et celle de la capacité maximale utilisée par Prophet pour toutes les fenêtres. Par exemple, si on

s'intéresse à prédire ce que pourrait être le nombre de personnes en soins intensifs, l'usager pourrait fournir le nombre maximal de lits en soins intensifs comme étant la capacité maximale alors que la capacité minimale serait le nombre de lits occupés en moyenne dans les jours normaux.

2.1.2.1.2 Modèle de Croissance Linéaire

Dans certains cas, il se pourrait que les problèmes de prédiction à l'étude ne présentent pas de plateau dans une fenêtre de notre analyse ou tout simplement, on souhaiterait éviter d'imposer à nos variables d'avoir une valeur maximale dans une fenêtre. Dans un tel cas, on peut utiliser une fonction linéaire avec un taux de croissance constant par fenêtre. Ce modèle est donné ci-dessous :

$$g(t) = (k + \boldsymbol{a}(t)^T \boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^T \boldsymbol{\gamma}). \quad (2.9)$$

Ici, $\boldsymbol{\gamma}$ est le paramètre d'ajustement utilisé pour maintenir la continuité de la fonction de croissance à travers deux fenêtres consécutives (voir la section 4.2 pour les applications). Les paramètres par défaut utilisés dans la librairie de Prophet sont donnés ci-dessous [37] :

- Le paramètre m suit une loi a priori normale. Par défaut, $m \sim N(0,5)$.
- Le paramètre k qui définit le taux d'accroissement suit une loi apriori normale. Par défaut, $k \sim N(0,5)$.
- Le paramètre δ suit une loi a priori de distribution de Laplace. Par défaut, elle est définie par : $\delta \sim Laplace\ (0, 0.05)$

Quant à la prédiction, Prophet suppose que la position des fenêtres futures ou des points de changement à venir se reproduisent à une fréquence (position des points) et à une intensité (taux d'augmentation de la croissance dans chaque fenêtre) semblable à celles observées dans les données de l'échantillon d'entrainement.

2.1.2.2 Saisonnalité, g(t)

Comme toute fonction continue, l'effet saisonnier peut être approximé par une série de fonctions harmoniques. Dans le cas de Prophet, les séries de Fourier [42] sont utilisées pour modéliser les effets périodiques et la formule est ci-dessous :

$$s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \cos\left(\frac{2\pi nt}{P}\right) \right] = X(t)\beta, \quad (2.10)$$

où X(t) et β sont exprimés sous forme vectoriel comme ci-dessous :

$$X(t) = \left[\cos\left(\frac{2\pi 1t}{P}\right), \sin\left(\frac{2\pi 1t}{P}\right), \ldots, \cos\left(\frac{2\pi Nt}{P}\right), \sin\left(\frac{2\pi Nt}{P}\right).\right.$$

$$\beta = [a_1, b_1, \ldots a_N, b_N]$$

Les paramètres N et P sont l'ordre et la période de la série de Fourrier. Les paramètres par défaut utilisés sont décrits ci-dessous [37] :

- Le paramètre β suit une loi a priori normale. Par défaut $\beta \sim N(0, \sigma^2)$, avec σ régularisant la force de saisonnalité.
- Pour une saisonnalité annuelle on a par défaut, N=10, P=365.25
- Pour une saisonnalité hebdomadaire par défaut, N=3, P=7

Les valeurs ci-dessus sont d'après les auteurs [14], celles qui semblent donner les meilleurs résultats. Toutefois nous avons estimé à travers une grille les paramètres les plus appropriés pour nos données.

2.1.3 Modèles par compartiments

Les modèles épidémiologiques qui divisent une population en compartiments sont souvent appelées modèles par compartiments [18] [19] [20] [21] [22]. Pour des raisons de commodité mathématique, ces compartiments sont généralement représentés par des lettres telles que S, E, I et R indiquant respectivement les populations Susceptible (*Susceptible)*, Exposées (*Exposed*), Infectées (*Infected*) et Retirées (*Removed*). Ces variables se définissent comme suit :

- *Susceptible* (*S*) fait référence aux individus qui sont vulnérables à l'infection. Cela signifie qu'ils n'ont jamais été infectés par l'agent pathogène et peuvent donc devenir infectés.
- *Exposed* (*E*) fait référence aux individus infectés par l'agent pathogène, mais pas encore infectieux, en raison de la période de latence de la maladie. Ils ne présentent pas de symptômes et sont incapables d'infecter les autres du même compartiment *E*.

- *Infectious/Infected* (I) sont des individus déjà infectés par l'agent pathogène et peuvent le transmettre à d'autres.
- *Removed* (R) sont les individus qui ne peuvent plus transmettre la maladie. Ce groupe comprend les individus guéris (également notés R) et les personnes décédées (notées D).

La liste des compartiments ci-dessus n'est pas exhaustive, dépendamment du contexte, elle peut être élargie afin d'inclure de nouveaux compartiments pour les personnes vaccinées, la capacité des soins intensifs, la capacité d'accueil mortuaire, etc.

Il convient de mentionner qu'un individu guéri (R) peut garder ce statut de façon permanente (on dit qu'il est immunisé) ou peut redevenir sensible au virus et bien évidemment retourner dans le compartiment S où il pourrait de nouveau s'infecter. Ceci est souvent observé lorsqu'il y'a une mutation de l'agent pathogène ou par suite d'une faible durée de la mémoire immunitaire. Les compartiments mentionnés ci-dessus peuvent être combinés pour former divers types de modèles reflétant les interactions possibles entre les différents compartiments qui la composent comme suit : SIS, SIR, SERD, SIRS et ainsi de suite. Ainsi, si l'on voulait utiliser un modèle qui prend en compte la réinfection d'une personne après une protection à la maladie (par guérison ou vaccination), un modèle de type SEIS ou SIS serait approprié pour la dynamique de la maladie. Des détails sur certains modèles sont fournis ci-après.

2.1.3.1 Modèles SIR

Le modèle mathématique de base SIR ou *Susceptible-Infected-Recovered* a été développé par [19]. Il consiste à diviser une population donnée en *Susceptible* (S), *Infected* (I) et *Recovered* (R). Etant donné qu'au fil du temps, le nombre de personnes dans chacun de ces compartiments change, on utilise la notation $S(t)$ pour caractériser le nombre d'individus susceptibles d'attraper la maladie au temps t, $I(t)$ pour définir le nombre d'individus infectés au temps t, et $R(t)$ pour le nombre d'individus décédés au temps t. Certaines hypothèses principales [43] fournies ci-dessous sont utilisées pour dériver les équations différentielles ordinaires (EDO) associées.

- La population initiale est supposée fixe dans le temps, ainsi, personne d'autre n'est ajouté au groupe susceptible ou sensible au fil du temps, puisque les naissances et l'immigration sont ignorées. Les seuls décès sont ceux imputés à la maladie.

- Tout individu susceptible peut se déplacer librement n'importe où dans l'espace occupé par la population.
- La seule façon pour un individu de quitter le groupe sensible est de s'infecter.
- En cas d'infection (nouvelle ou pas), l'individu infecté passe de la classe susceptible à la classe infectieuse.
- La loi de l'action de masse qui suppose que la transmission de la maladie dépend de taille de la population s'applique. Notons que ceci est aussi correct pour les infections comme la grippe, la malaria.

Le taux de changement dans le temps de $S(t)$ dépend (a) du nombre $S(t)$ déjà susceptibles et (b) de la quantité de contacts entre $S(t)$ et $I(t)$. En particulier, on admet que chaque individu infecté pourrait transmettre la maladie à un nombre fixe β de personnes susceptibles par unité de temps (en jours). On en déduit que

$$\frac{dS(t)}{dt} = -\beta S(t).$$

Le signe négatif indique que le taux de changement des susceptibles est toujours décroissant. Étant donné qu'il y'a au temps t une population de $I(t)$ personnes infectées et que chacune des personnes infectées pourrait transmettre la maladie aux personnes susceptible, l'équation ci-dessus peut se généraliser comme suit pour donner la 1ere équation différentielle ordinaire (EDO):

$$\frac{dS(t)}{dt} = -\beta S(t)I(t).$$

Une fraction fixe, γ, du groupe infecté se rétablira au cours d'une journée donnée. Le groupe retiré (*removed*) comprend les personnes qui se rétablissent et celles qui meurent, car les deux sont retirées du groupe infectieux [44].

Le nombre moyen de jours qu'il faut à un individu pour se remettre de la maladie, n, est inversement proportionnel à γ. Par exemple, si la durée moyenne de l'infection est de trois jours, alors, en moyenne, un tiers de la population actuellement infectée $I(t)$ se rétablit $R(t)$ chaque jour. Ceci permet d'avoir la seconde EDO :

$$\frac{dR(t)}{dt} = \gamma I(t)$$

Étant donné que (a) de nouvelles infections surviennent à la suite d'un contact entre les infectieux $I(t)$ et les susceptibles $S(t)$, et (b) étant donné que seuls les individus infectieux $I(t)$ peuvent entrer dans la classe retirée $R(t)$, la dernière EDO est exprimée sous la forme

$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t).$$

Par souci de simplicité, le modèle par compartiment est généralement représenté sous forme d'organigramme avec des flèches définissant les étapes clés du processus comme indiqué ci-après pour le modèle SIR.

$$\mathrm{S} \xrightarrow{\beta I} \mathrm{I} \xrightarrow{\gamma} \mathrm{R}$$

Cette notation signifie que le seul mouvement possible lorsqu'on est susceptible est d'aller dans le compartiment infecté. De même, une fois infecté, le seul déplacement possible est d'aller dans le compartiment R.

En combinant toutes les EDOs ci-dessus, on obtient pour le modèle SIR :

$$\mathrm{S} \xrightarrow{\beta I} \mathrm{I} \xrightarrow{\gamma} \mathrm{R}$$

$$\begin{cases} \dfrac{dS(t)}{dt} = -\beta S(t)I(t) \quad (a) \\ \dfrac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t) \; (b) \\ \dfrac{dR(t)}{dt} = \gamma I(t) \qquad (c) \end{cases} \tag{2.11}$$

Afin de simplifier la formulation du modèle, les variables peuvent subir quelques transformations [45] de leur unité pour des raisons purement numériques comme suit :

$$(S, I, R) = N \times (x, y, z) \text{ et } (t, \beta, \gamma) = (\tau t', \tau^{-1}\rho, \tau^{-1}\sigma),$$

où N est la taille de la population, et τ est un coefficient (généralement en [min]) utilisé pour la conversion du temps, t, en unité numérique, t', désirée. Ainsi, τ=1440, pour une conversion de jour en minutes ou τ=1, si on reste en jour). On obtient alors les EDOs transformées suivantes :

$$\frac{\mathrm{d}x}{\mathrm{d}t'} = -\rho xy$$
$$\frac{\mathrm{d}y}{\mathrm{d}t'} = \rho xy - \sigma y$$
$$\frac{\mathrm{d}z}{\mathrm{d}t'} = \sigma y \qquad (2.12)$$

- Initialement, $S(0) \approx N, I(0) > 0, et\ R(0) = 0.$
- De même, $x = x(t'), y = y(t'), z = \gamma(t').$

La force du modèle SIR réside dans sa simplicité d'interprétation, et peut jouer un rôle important en aidant à quantifier les stratégies de contrôle de la maladie. Par exemple, en permettant de se consacrer sur certains aspects particuliers de la maladie, en déterminant les quantités seuils qui définissent la survie de la maladie et en évaluant l'effet de stratégies de contrôle particulières mise en place. Pour cela, une des métriques les plus importantes est le nombre de reproduction de base [46], parfois appelé ratio de reproduction de base (*basic reproductive number,* R_0*, or basic reproductive ratio*). Ce dernier est brièvement commenté ci-dessous.

Note sur R_0 nombre de reproduction de base

R_0 *est défini comme le nombre moyen d'infections produits par un individu infecté introduit dans une population d'individus sensibles, où un individu infecté a contracté la maladie et où les individus sensibles sont en bonne santé mais peuvent contracter la maladie* [47].

En se concentrant sur l'équation 2.8 (b), le comportement initial de la pandémie est régi par le signe de $\beta S(t)I(t) - \gamma I(t)$, ou plus simplement $\beta SI - \gamma I$, où $S = \frac{S(t)}{N}$ et $I = \frac{I(t)}{N}$.

À l'équilibre (i.e. lorsque $\beta SI - \gamma I = 0$), le ratio des coefficients est noté par $R_0 = \frac{\beta}{\gamma}$ au stade précoce de la maladie, c'est à dire lorsque la fraction de personnes susceptibles d'être infectées est

S=1. Ainsi, le calcul de R_0 suppose au début une population où tous les individus sont en bonne santé, à l'exception de l'individu infectieux introduit (également appelé « patient zéro »). On en déduit que :

- Lorsque $\beta > \gamma$, on obtient $R_0 > 1$. La maladie se propage et devient une pandémie lorsque $R_0 > 1$.
- Lorsque $\beta < \gamma$, on obtient $R_0 < 1$. Le nombre d'individus infectieux décroît de façon monotone jusqu'à 0 et la maladie disparaît.

Ainsi, le calcul du taux de reproduction est crucial pour déterminer si l'épidémie est sous contrôle [48]. Contraindre $R_0 < 1$ nécessiterait de réduire la constante de vitesse de contact, β, ou d'augmenter la vitesse de guérison, γ. Les étapes vers la réalisation de cet objectif pourraient inclure des actions telles que :

- Diminution du taux de contact
  - Implémenter la distanciation sociale
  - Mettre en quarantaine des personnes exposées,
  - Imposer un couvre-feu et ou un blocus d'une région
  - Porter les masques, réduire l'existence des surfaces souillées, etc.
- Augmentation du taux de guérison
  - Utilisation des médicaments efficaces
  - Vaccination

Parce qu'une pandémie pourrait avoir différents stades (vagues d'infections, mutations de l'agent pathogène, etc.), R_0 est donc le nombre de reproduction de base au début de la pandémie ou au début d'une nouvelle vague [49]. Ceci s'explique par le fait que toute personne non infectée (ou dont la réponse immunitaire est faible) est considérée comme sensible dans la plupart des cas. Ceci induit donc un probable taux d'infection élevé, d'où R_0 élevé. Au cours du temps le nombre de reproduction de base varie (noté R_t) grâce aux mesures de contrôle ou de vaccination.

2.1.3.2 Modèles SIRD

Étant donné que l'on peut mesurer séparément le nombre de décès et de guérisons, on peut utiliser deux variables guéris (*Recovered*) et décès (*Deaths*) au lieu de *Removed* (R) dans le modèle SIR.

Ainsi, une personne infectée peut soit se remettre de la maladie, soit mourir[7] des suites de l'infection [50]. De plus, on suppose que toutes les personnes exposées au virus sont infectées immédiatement, c'est-à-dire qu'il n'y a pas de temps de latence entre l'exposition et l'infection. Le diagramme de flux est présenté ci-dessous ainsi que les EDOs y associées [51] [52].

$$\begin{array}{l} \mathrm{S} \xrightarrow{\beta I} \mathrm{I} \xrightarrow{\gamma} \mathrm{R} \\ \qquad\quad \mathrm{I} \xrightarrow{\alpha} \mathrm{D} \end{array}$$

Afin de simplifier la formulation du modèle, les variables peuvent subir quelques transformations [45] de leur unité comme suit.

$(S, I, R, D) = N \times (x, y, z, w)$ et $(t, \alpha, \beta, \gamma) = (\tau t', \tau^{-1}\kappa, \tau^{-1}\rho, \tau^{-1}\sigma)$

On obtient alors :

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t'} &= -\rho x y, \\ \frac{\mathrm{d}y}{\mathrm{d}t'} &= \rho x y - (\sigma + \kappa) y, \\ \frac{\mathrm{d}z}{\mathrm{d}t'} &= \sigma y, \\ \frac{\mathrm{d}w}{\mathrm{d}t'} &= \kappa y, \end{aligned} \tag{2.13}$$

où α est le taux de mortalité. Le nombre de reproduction de base peut être déduit facilement et on trouve $R_0 = \frac{\rho}{\sigma+\kappa} = \frac{\beta}{\gamma+\alpha}$

2.1.3.3 Modèles SIR-F

L'avantage des approches par compartiments réside à la flexibilité d'inclure de nouveaux compartiments à ceux existant. Ainsi, le modèle SIRD peut être légèrement modifié afin d'inclure nombreux cas qui ont été confirmés infectés de la pandémie après leur mort.

[7] Ceci exclut le cas de ceux qui bien que guéris de la maladie, peuvent mourir des séquelles y étant associées.

Ainsi, en considérant certaines personnes décédées avant d'être allées dans un centre hospitalier (Cas confirmé), le modèle SIRD peut prendre la forme élargie

$$\begin{aligned} S \xrightarrow{\beta I} S^* \xrightarrow{\alpha_1} D \\ S^* \xrightarrow{1-\alpha_1} I \xrightarrow{\gamma} R \\ I \xrightarrow{\alpha_2} D \end{aligned}$$

Ici, S^* décrit les cas qui sont réellement porteurs de la maladie qui peuvent soit avoir eu un statut confirmé uniquement après leur décès, soit avoir été transférés ou confirmé infectés après leur test. Tel que défini, S^* sert de compartiment auxiliaire pour séparer deux situations de décès en y introduisant un facteur de probabilité α_1 par defaut.

Les variables peuvent subir quelques transformations [53] de leur unité comme suit

$(S, I, R, F) = N \times (x, y, z, w)$ et $(t, \alpha_1, \alpha_2, \beta, \gamma) = (\tau t', \theta, \tau^{-1}\kappa, \tau^{-1}\rho, \tau^{-1}\sigma)$

Et les EDO sont comme suit

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t'} &= -\rho xy \\ \frac{\mathrm{d}y}{\mathrm{d}t'} &= \rho(1-\theta)xy - (\sigma+\kappa)y \\ \frac{\mathrm{d}z}{\mathrm{d}t'} &= \sigma y \\ \frac{\mathrm{d}w}{\mathrm{d}t'} &= \rho\theta xy + \kappa y \end{aligned} \tag{2.14}$$

Le nombre de reproduction de base peut être déduit facilement et on trouve [19]

$$R_0 = \rho(1-\theta)(\sigma+\kappa)^{-1} = \beta(1-\alpha_1)(\gamma+\alpha_2)^{-1}$$

2.2 Mesure de performance

La racine carrée de l'erreur quadratique moyenne [54] ou *Root Mean Square Error* (*RMSE*) a été choisie comme mesure d'évaluation en raison de sa capacité à pénaliser les erreurs importantes et, en tant que tel, elle est bien adaptée à la comparaison des performances de différents modèles. Une autre mesure d'erreur populaire serait le pourcentage d'erreur absolue moyenne ou *Mean Absolute Percentage Error* (MAPE) qui ne pénalise pas les erreurs importantes de la même manière que le RMSE et qui ne reflète pas les valeurs aberrantes possibles dans les données.

Pour simplifier, nous supposons que nous avons déjà N observations ou échantillons d'erreurs (entre les valeurs prédites et observées) de modèle ϵ notées $(e_i, i = 1,2, \dots, N)$. Le RMSE et le MAPE sont calculés pour les prédictions d'infections et de décès comme suit :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {e_i}^2} \tag{2.15}$$

$$MAPE = 100 \times \frac{1}{N}\sum_{i=1}^{N} \frac{|e_i|}{True_i} \tag{2.16}$$

CHAPITRE 3 PRESENTATION DES DONNÉES ET VISUALISATION

Dans ce chapitre, nous présentons et explorons l'ensemble de données utilisées dans le cadre de ce mémoire. Le chapitre est subdivisé en deux parties, la première parle de la source des données alors que la seconde partie porte sur l'analyse graphique de ces données

3.1 Source de données et extraction

Deux bases de données sont utilisées pour ce projet. Premièrement, les données chronologiques obtenues de la base de données John Hopkins[8] et qui se composent entre autres du nombre de cas actifs, du nombre de cas confirmés, du nombre de décès, et du nombre de cas guéris. A ce jour, c'est l'une des bases de données les plus utilisées en modélisation reliée à la COVID-19. Deuxièmement, nous avons téléchargé les données de COVID-19 au Canada à partir du site Web du ministère de la Santé Publique[9] « Health-infobase ». Cette base de données contient des informations journalières agrégées détaillées sur la propagation du virus au fil du temps et dans différentes régions du pays comme l'information relative au sexe, l'âge, les hospitalisations, soins intensifs, décès, nombre et type des tests effectués, nombre de cas associés aux variants préoccupants etc.

Étant donné que ce travail se concentre principalement sur le Canada et la province du Québec en particulier, la base de données de la Santé Publique du Canada « Heath-infobase » est utilisée comme source principale pour les données de ce travail. Cependant, nous avons comparé notre extraction avec celles disponible dans la base de données John Hopkins, et avons constaté qu'en juillet 2021, les deux bases de données contenaient des informations identiques.

---

[8] Les bases de données de Johns Hopkins University (JHU) sont sans aucun doute les plus populaires utilisées dans les modèles associés au COVID. Elles possèdent l'information 181 pays et sont mise à jour quotidiennement.

Source : Source: https://github.com/CSSEGISandData/COVID-19

[9] Source: https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html

3.1.1 Préparation des données

Les données disponibles dans les bases ci-dessus mentionnées vont de janvier 2020 jusqu'en juillet 2021. Nous avons noté quelques valeurs manquantes sur quelques lignes en début de l'année 2020 (soit le début de la pandémie). Par conséquent, pour éviter tout problème d'ordre numérique, les données avant le 1er avril 2020 ont été supprimées[10] et le 15 juillet 2021 (la dernière date d'extraction) a été utilisé comme date limite pour l'analyse. Ainsi, les données du 1er avril 2020 au 15 juillet 2021 ont été utilisées pour ce travail. Les données ont été extraites au format « csv » et nous avons effectué notre analyse à l'aide du langage de programmation « Python » à travers la plateforme « *Google Colab*[11] » de Google.

3.1.2 Faits saillants sur les données du Québec

Afin de mieux comprendre l'évolution de la contagion de la COVID-19 au Québec et leur possible impact sur notre stratégie de modélisation ou les sorties de nos modèles, nous devons prendre en compte les événements saillants (ils sont surlignés ci-dessous) et dates clés ayant marquées la santé publique au Québec. Ces informations sont libres d'accès et disponibles sur le site web de l'Institut National de La Santé Publique du Québec[12]. Ci-dessous dans le **Tableau 3.1**, nous donnons quelques faits marquants de la pandémie de la Covid-19 au Québec de janvier 2020 à juillet 2021.

---

[10] La raison de l'élimination des échantillons initiaux est qu'au début de la pandémie, les tests COVID ont eu de la peine à se mettre en place et les taux COVID observés pourraient ne pas refléter la réalité.

[11] Colaboratory, or "Colab" en bref, permet d'écrire et d'exécuter Python dans un navigateur, sans aucune configuration requise, un accès gratuit aux GPU et un partage facile.

Source: https://colab.research.google.com/?utm_source=scs-index.

[12] Source: https://inspq.qc.ca/covid-19/donnees/ligne-du-temps

Tableau 3.1 Quelques dates clés de la pandémie de la COVID-19 au Québec de janvier 2020 à juillet 2021.

Période	Faits saillants (les plus importants sont surlignés)
2020	o 27 février, Premier cas détecté o 29 février. Début de la semaine de relâche o 6 mars, détection du Variant Brésilien Gamma o 13 mars, Urgence sanitaire déclarée o 18 mars, premier décès rapporté o 21 mars, Interdiction de tout rassemblement intérieur et extérieur et Détection du variant Nigérian êta o 5 avril, Recrudescente dans les zones de la région de la Chaudière - Appalaches, des régions de la Capitale-Nationale, et de l'Outaouais. o 7 avril. On dénombre 10 000 cas au Québec o 26 avril, Émergence du variant Delta en Inde. o 4 mai, Réouverture des magasins o 11 juillet, Fin officiel de la 1ere vague o 10 août, plan de rentrée scolaire. Retour avec masque o 23 août début de la seconde vague o 11 novembre, on rapporte plus de 120 000 cas et 23% de la population a été testée. o 14 décembre, début de la vaccination
2021	o 6 février, Québec atteint 10 000 décès et plus de 250 000 cas o 20-21 mars, Fin de la seconde Vague et début de la 3eme Vague o 14 avril, le quart de la population est vaccinée o 26 avril. Détection du Variant préoccupant Delta

	○ 18 mai, La moitié de la population est vaccinée ○ 6 juin, 75% de la population de plus de 12 ans a été vaccinée ○ 17 juillet Fin de la 3eme Vague et début de la 4eme Vague. Fin des données utilisées pour la modélisation dans le cadre de ce travail

On note en général que la vie au Québec a été rythmée entre autres par des vagues successives de contamination (3 vagues) liées aux différents variants. La vaccination a aussi constitué un tournant majeur dans la lutte contre la pandémie alors que le retour en classes semble accélérer la pandémie.

3.2 Visualisation des données

Lors de l'extraction, nous avons conservé les données de toutes les provinces du Canada afin de comparer le statut sanitaire du Québec par rapport aux autres provinces. Trois variables principales sont présentes dans notre base de données finale. Elles sont cumulatives et correspondent respectivement au nombre de cas confirmés '*Confirmed*', de décès '*Deaths*' et de guérisons '*Recovered*'. De ces informations cumulées, on peut déduire le nombre de nouvelles variables comme les cas journaliers (utilisés pour la mesure de performance) et les cas actifs qui s'expriment comme suit:

$$Infections(t) = Confirmed\ (t) - Confirmed\ (t-1) \qquad (3.1)$$

$$Active(t) = Confirmed\ (t) - Recovered\ (t) - Deaths(t)$$

Pour notre travail, différentes bibliothèques Python telles que "*Numpy*", "*Pandas*", "*Matplotlib*", "Datetime", "*sklearn.metrics*" ont été utilisées pour la visualisation et manipulation des données. Ci-dessous, dans la Figure 3.1 nous présentons le nombre total cumulé de cas confirmés dans les différentes provinces du Canada.

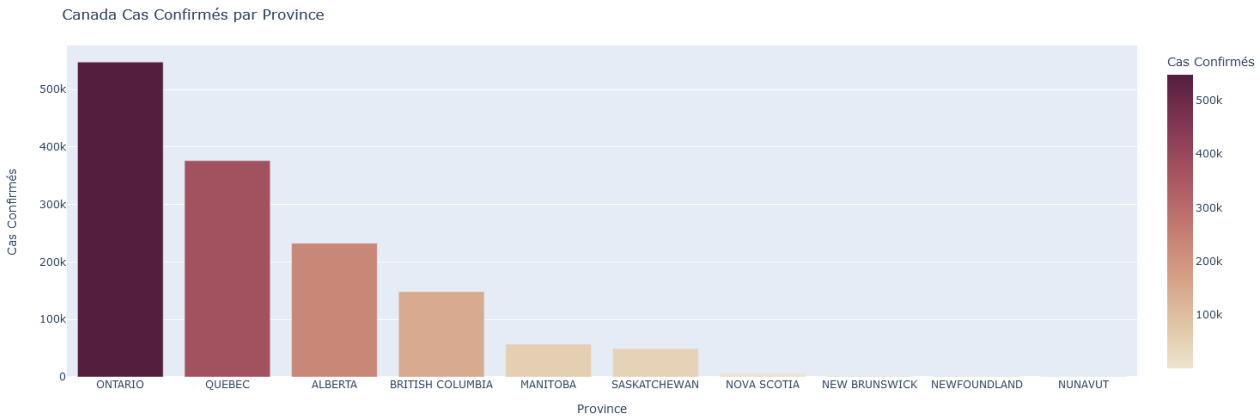


Figure 3.1. Cas COVID-19 confirmés cumulatifs par province jusqu'en juillet 2021

On note dans la **Figure 2.1** que jusqu'en juillet 2021, la province de l'Ontario avait le plus grand nombre de cas d'infections signalées. L'Ontario était suivi par Québec alors que Terre-Neuve et le Nouveau-Brunswick avaient signalé moins de cas. Le nombre total cumulé des cas de décès est discuté dans la **Figure 3.2** ci-dessous :

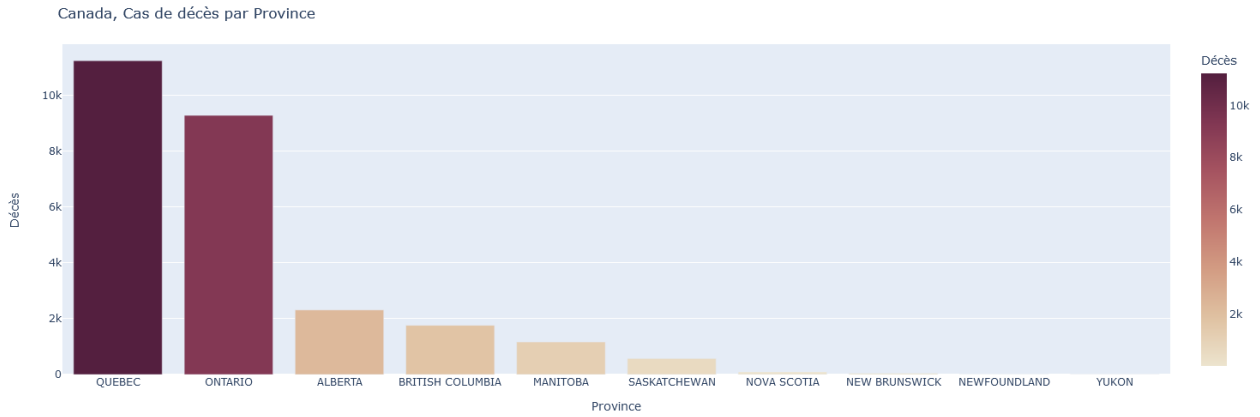


Figure 3.2. Nombre de décès cumulatifs dus à la COVID19 par province jusqu'en juillet 2021

Concernant les décès, la Figure 3.2 montre que jusqu'en juillet 2021, le Québec avait le plus grand nombre de cas cumulé de décès signalés. L'Ontario était le deuxième alors que Terre-Neuve et le Nouveau-Brunswick avaient signalé moins de cas parmi toutes les provinces. Ci-dessous, nous

montrons dans le **Tableau 3.3** la somme de tous les cas confirmés, actifs, de décès et guéris de toute notre base de données.

Tableau 3.2 Nombre de cas confirmés, décès, cas actives et guéris de la COVID-19 par province jusqu'en juillet 2021

	Province	Confirmés	Actifs	Décès	Guéris
0	ONTARIO	547705	1443	9275	536987
1	QUEBEC	376109	655	11232	364222
2	ALBERTA	232635	578	2312	229745
3	BRITISH COLUMBIA	148286	666	1761	145859
4	MANITOBA	57028	973	1164	54891
5	SASKATCHEWAN	49341	373	573	48395
6	NOVA SCOTIA	5870	22	92	5756
7	NEW BRUNSWICK	2343	9	46	2288
8	NEWFOUNDLAND	1433	50	7	1376
9	NUNAVUT	657	0	4	653
10	YUKON	497	77	6	414
11	PRINCE EDWARD ISLAND	208	0	0	208
12	NORTHWEST TERRITORIES	128	0	0	128
13	REPATRIATED CANADIANS	13	0	0	13
14	FEDERAL ALLOCATION	0	0	0	0

On note que l'Ontario a à la fois le plus grand nombre de cas actifs, confirmés et guéris alors que le Québec a le plus grand nombre de décès. On note également que le Manitoba présente le deuxième plus grand nombre de cas actifs de la série. Ci-dessous, dans le **Tableau 3.3** nous analysons le nombre de décès par 100 infections en moyenne. Ce taux de défaut est obtenu en divisant le nombre d'infections par le nombre de défauts par province

Tableau 3.3 Nombre de Décès vs. nombre de décès par 100 cas confirmés de la COVID-19 jusqu'en juillet 2021.

	Province	Décès	Décès /100 Cas Confirmés
0	QUEBEC	11232	2.990000
1	ONTARIO	9275	1.690000
2	ALBERTA	2312	0.990000
3	BRITISH COLUMBIA	1761	1.190000
4	MANITOBA	1164	2.040000
5	SASKATCHEWAN	573	1.160000
6	NOVA SCOTIA	92	1.570000
7	NEW BRUNSWICK	46	1.960000
8	NEWFOUNDLAND	7	0.490000
9	YUKON	6	1.210000
10	NUNAVUT	4	0.610000

On note que Québec a le taux de décès par cas confirmés le plus élevé. La raison du nombre plus élevé de décès liés à la COVID-19 au Québec pourrait être liée entre autres à la semaine de relâche en février (voir section 3.1.1) qui a favorisé les déplacements à l'étranger (infections suivies des décès), aux ports d'entrée ou d'expédition comme le grand port de Montréal et les aéroports dont celui de Montréal, et peut-être que la façon de compter les décès soit différente d'une province à l'autre. De plus, le Québec est une grande place touristique et compte une population plus âgée. La courbe des infections cumulatives dans la province du Québec est donnée ci-dessous dans la **Figure 3.3**.

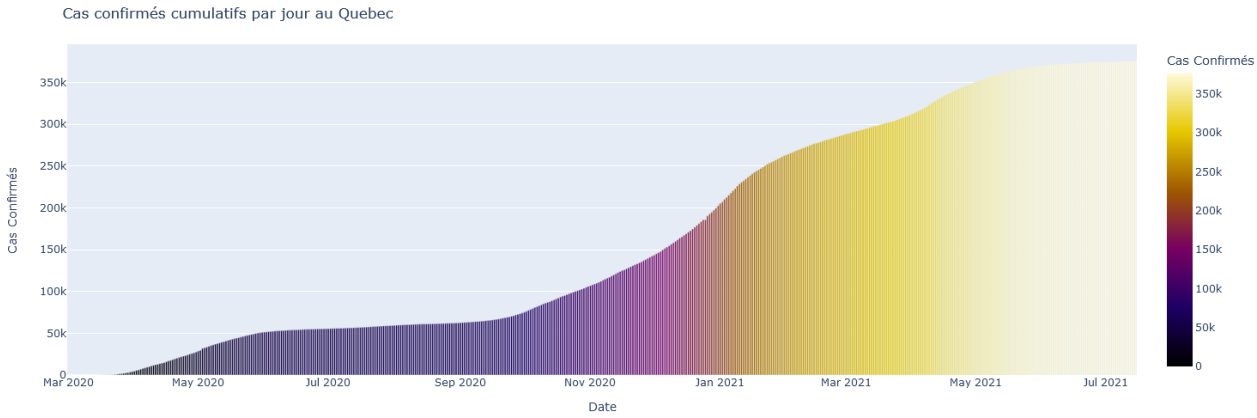


Figure 3.3. Cas confirmés cumulatifs de la COVID-19 par jour au Québec jusqu'en juillet 2021

La courbe montre une succession de sigmoïdes qui reflètent sans aucun doute les vagues successives ci-dessus mentionnées (section 3.1.2). La 1ère vague semble s'essouffler complètement en juillet 2020, alors que la seconde vague connait une vitesse fulgurante vers août-septembre 2020 (sans doute relié à la rentrée scolaire). Les choses semblent stagner un tout petit peu en février 2021 (sans doute lié au début de la vaccination) avant d'avoir une remontée autour du mois d'avril 2021 lié sans doute aux éclosions marquantes dans certaines régions du Québec comme la Chaudière-Appalaches, des régions de la Capitale-Nationale, et de l'Outaouais (voir section 3.1.2). Ci-dessous, dans la **Figure 3.4** nous présentons les sommes cumulatives des décès dans la province de Québec

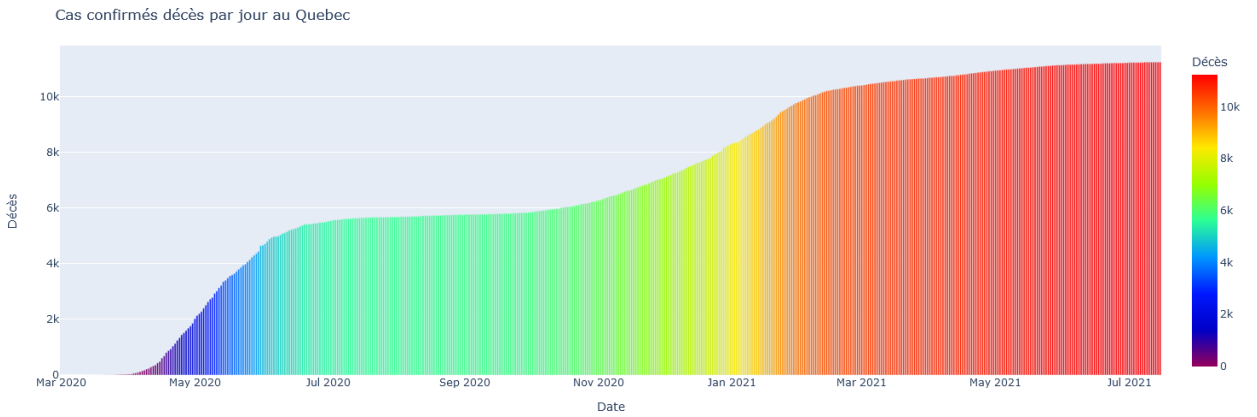


Figure 3.4. Cas de décès cumulatifs de la COVID-19 par jour au Québec jusqu'en juillet 2021

On note que comparé à la courbe cumulative d'infections, la courbe de décès semble laisser voir deux plateaux. L'un autour de juillet 2020 avec un cumulatif de 5000 cas environ et le second autour de juin 2021, avec environ 11000 décès. Ci-dessous dans la **Figure 3.5**, nous présentons les infections, les décès et les guérisons cumulatives en perspective à la même échelle.

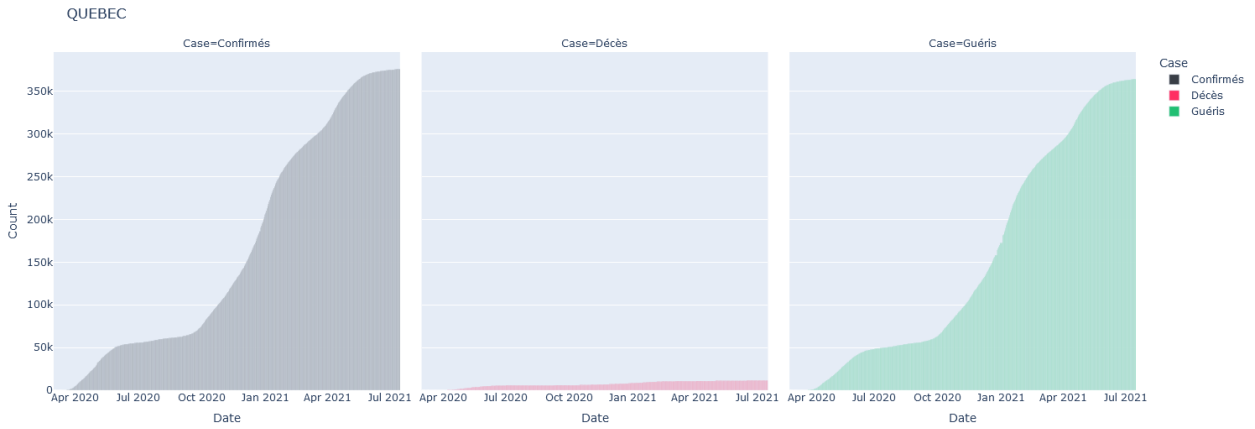


Figure 3.5. Cas confirmés, de décès et de guérison de la COVID-19 par jour et à la même échelle au Québec jusqu'en 2021

On note que Québec compte cumulativement plus de 300 000 cas confirmés de COVID-19 et plus de 10 000 décès en juillet 2021. Après avoir connu une hausse fulgurante au début de la pandémie, le nombre de décès semble se stabiliser sur un plateau à environ 11 000 cas. Ceci semble être le résultat de multiples confinements, port de masque et distanciation sociale. La campagne de vaccination battant son plein en juillet 2021, cela devrait avoir un impact sur la propagation du virus. Enfin, nous notons que la somme cumulative des cas confirmés et guéris est ponctuée de sauts. Ceci est lié au fait que le nombre de cas actifs ou guéris est régulièrement mis à jour. Ci-dessous dans la Figure 3.6 nous comparons les ratios de décès et guérisons pour 100 cas confirmés.

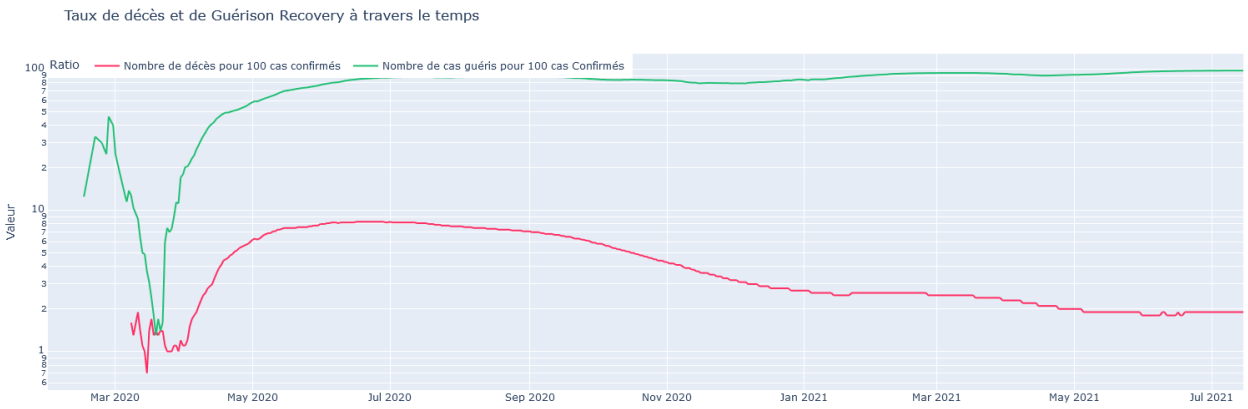


Figure 3.6. Nombre cas de décès et guéris par 100 cas confirmés au Québec jusqu'en juillet 2021.

On observe que le nombre de décès par 100 personnes infectées a connu un pic autour de juillet 2020 avec une valeur autour de 7% et a décru dès lors pour atteindre une valeur proche de 3% autour de juillet 2021. Inversement, depuis juillet 2020, le taux de guérison par 100 personnes infectées est stable autour de 97-98%.

Notez enfin que d'autres analyses relatives aux autres provinces du Canada sont disponibles sur le notebook dédié à ce mémoire[13].

3.3 Partition des données et analyse des variables d'intérêt

Nous rappelons qu'un des objectifs de ce mémoire est de comparer la performance de quelques modèles de prédiction des infections et décès liés à la COVID. Dans cette optique, les données sont divisées en un échantillon d'apprentissage, utilisé pour estimer les paramètres du modèle et d'un échantillon de validation ou test qui est utilisé pour évaluer la performance des modèles. La taille de la période de test a été maintenue fixée à 30 jours et correspond à la période qui suit directement l'échantillon d'entrainement (*Training*).

[13] Lien : https://colab.research.google.com/drive/1Q-oL8Th5BbOw0LaqI8baMTFtUQK3bvMZ?usp=sharing

Pour la suite de ce chapitre et sauf indication contraire, la population d'entrainement couvre la période d'avril 2020 à octobre 2020 et l'échantillon de test couvre les 30 prochains jours soit du 1er novembre 2020 au 30 novembre 2020. Notre intérêt pour ces périodes se justifie par le fait qu'elles excluent l'influence de la vaccination qui a démarré en décembre 2020 au Québec, ce qui aurait potentiellement biaisé nos analyses. Les résultats de nos modélisations basées sur les données qui couvrent une partie de la période de la vaccination sont discutés dans le Chapitre 4. De même, une analyse de sensibilité y est effectuée pour évaluer le potentiel impact de la segmentation (choix de la période de données utilisées pour l'entrainement) sur nos résultats.

Deux variables d'intérêt sont analysées à savoir les infections et les décès. La représentation graphique (journalière) de ces variables est donnée à la **Figure 3.7** ci-dessous

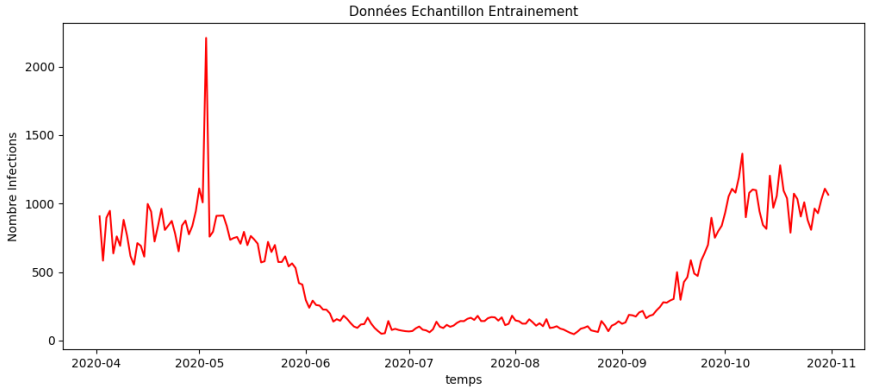


Figure 3.7. Cas COVID-19 confirmés journalier de l'échantillon d'entrainement

On note un pic des infections autour du début du mois de mai 2020. Ceci se justifierait par la combinaison d'une forte poussée d'infections dans les CHLSD à Montréal d'après les informations[14] du journal '*La Presse*' en ces jours, et d'une correction de la comptabilité des infections à la suite d'un

[14] Source : https://www.lapresse.ca/covid-19/2020-05-02/le-bilan-de-la-covid-19-s-alourdit-encore-au-quebec, Consulté en novembre 2021.

bug informatique au 03 Mai 2020 comme le rapporte le journal[15] '*Le Devoir*' ou aussi le journal[16] '*Le Soleil*' qui écrit « *…Selon les autorités, cette forte hausse est attribuable à un problème informatique. Selon elles, pas moins de 1317 cas qui ont été détectés entre le 2 et le 30 avril n'avaient pas encore été comptabilisés dans les données officielles…* » [Version web du journal du *Le Soleil* en date du 03 Mai 2020]. Étant donné que ces cas sont réels et bien que cela affecte les paramètres de notre modélisation, nous les gardons tels quels et une analyse de sensibilité sera effectuée. On note aussi que les cas d'infection semblent remonter au mois de septembre 2020 et ceci pourrait s'expliquer par la combinaison de la reprise des cours en présentielle (secondaire, primaire, Cegep), du variant Delta plus contagieux et de l'hiver qui favorise le regroupement intérieur. Le nombre cumulatif des cas confirmés a été montrée dans Figure 3.3. Ci-dessous, dans la **Figure 3.8** nous analysons les données journalières des décès de notre échantillon de *Training*.

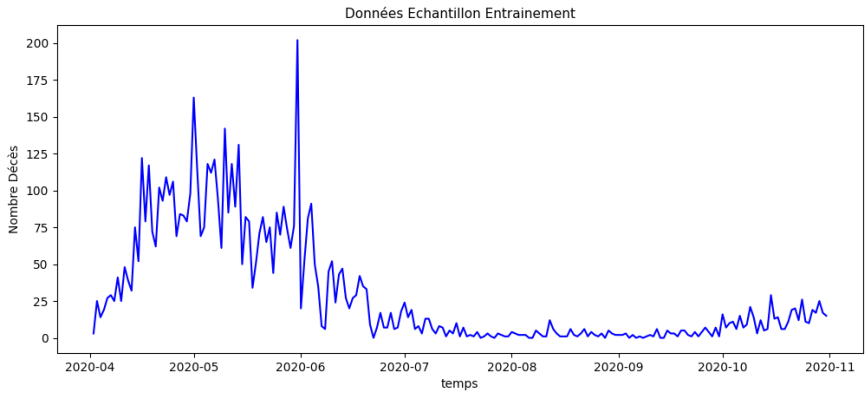


Figure 3.8. Cas décès de la COVID-19 de L'échantillon d'entrainement

On note que le nombre de décès a été assez élevé tout au début de la pandémie avec des pics de plus de 100 décès journaliers environ entre la fin avril 2020 et mi-mai 2020. Ceci se justifierait par

[15] Source : https://www.ledevoir.com/societe/sante/578200/coronavirus-bilan-quebec-3-mai. Consulté en novembre 2021.

[16] Source : https://www.ledevoir.com/societe/sante/578200/coronavirus-bilan-quebec-3-mai. Consulté en novembre 2021.

le nombre relativement élevé de décès dans les CHLSD de la province à cette période[17] (voir aussi 'Journal de Québec[18]'). On note aussi un pic de décès au Québec autour de la période du 31 mai 2020. Comme précédemment dans le cas des infections, cela correspond à des corrections des données qui n'avaient pas été pris en compte au cours des mois passés tel que rapporté dans le journal de Québec[19] en cette date : '…*De plus, la Belle Province a rapporté 37 décès dans les dernières 24 heures, mais l'ajout de 165 défunts qui n'avaient pas été comptabilisés à ce jour a fait gonfler les statistiques de **202** décès d'un seul coup*.' [Version web du journal du *Journal de Québec* en date du 31 mai 2020]. La somme cumulative des décès a été montrée dans la Figure 3.4.

Comme précédemment, nous avons gardé ces corrections dans nos données aux dates où elles ont eu lieu tout en étant conscient de leur potentiel impact sur les estimés de nos modèles. A cause de la forte variabilité des données journalières (voir **Figure 3.7** et **Figure 3.8**), nous avons utilisé sauf mention contraire les valeurs des sommes cumulatives car nos tests préliminaires ont laissé suggérer une performance médiocre avec les valeurs journalières. Les sorties du modèle ont été par la suite transformées en données journalières comme illustré en Eq. 3.1 pour la mesure de la performance.

---

[17] Source : https://www.journaldequebec.com/2020/05/02/covid-19-1008-nouveaux-cas-et-114-nouveaux-deces-au-quebec. Consulté en novembre 2021.

[18] Source : https://www.journaldequebec.com/2020/04/07/covid-19-devoilement-des-projections-sur-la-pandemie-au-quebec-aujourdhui. Consulté en novembre 2021.

[19] Source : https://www.journaldequebec.com/2020/05/31/covid-19-le-quebec-compte-37-deces-additionnels-1. Consulté en novembre 2021.

CHAPITRE 4 IMPLÉMENTATION DES MODÈLES ET RÉSULTATS

Nous décrivons dans ce chapitre les étapes suivies pour l'implémentation des modèles. Pour les approches statistiques le logarithme de la cumulative des observations journalières est utilisé. Ceci permet d'avoir des données lissées qu'on pourrait facilement rendre stationnaires (ceci est analogue à ce qu'ont fait certains auteurs [55], [56], [57]). Pour rappel (section 3.3), les étapes sont les suivantes: (a) nos données brutes sont des cumulatives[20] des observations journalières (voir **Figure 3.4**); b) nous appliquons le logarithme aux données des sommes cumulatives; c) le modèle est bâti et utilisé pour prédire (après conversion de l'expression logarithmique en nombre) les observations de l'échantillon test (1 mois); une différence d'ordre 1 est appliquée sur les données prédites et observées pour obtenir les valeurs journalières. Pour les modèles par compartiments, la transformation logarithmique n'a pas été nécessaire d'une part à cause de l'absence de contrainte de stationnarité sur des données, de l'absence des références littéraires qui appuieraient ce processus, et d'autre part à cause des tests préliminaires peu concluants.

4.1 Les Modèles statistiques: ARIMA et SARIMA

À des fins de modélisation, nous devons choisir des valeurs (p, d, q) pour les modèles ARIMA ainsi que les valeurs (P, D, Q) pour la composante saisonnière du SARIMA. Étant donné qu'il existe de nombreuses façons de faire ces choix, telles que l'examen des tracés d'autocorrélation, l'expérience du domaine, etc., nous avons choisi d'adopter l'approche ci-dessous :

1. Rendre la série stationnaire c'est-à-dire de trouver la valeur de d
2. Identifier les valeurs plausibles de la composante moyenne mobile (q) et du terme autorégressifs (p) à travers l'analyse visuelle des courbes des fonctions [58] d'autocorrélation (ACF) et d'autocorrélation partielle (PACF).

[20] Une option aurait été d'utiliser une moyenne mobile, cependant elle a été abandonnée à cause de sa dépendance à la fenêtre à considérer.

3. Identifier la fréquence de la saisonnalité dans le cas du modèle SARIMA à travers l'analyse de la décomposition de la série chronologique en ses composantes de tendance, saisonnalité et résidu.

4. Effectuer une recherche par grille sur plusieurs valeurs de (p, q, P, D, Q) ou (p, q) en utilisant le critère d'information d'Akaike (AIC). Ce dernier a été largement utilisé dans la littérature [59] [60] pour sélectionner les meilleurs modèles parmi d'autres potentiels candidats. L'approche consiste à estimer la qualité relative de plusieurs modèles ARIMA ou SARIMA en comparant leurs valeurs AIC. Ainsi, le meilleur modèle est celui avec une valeur AIC faible. La bibliothèque *pyramid-arima* de Python nous permet d'effectuer rapidement cette recherche de grille et aussi de créer un modèle (paramètres du modèle et analyse des résidu) pouvant s'adapter à nos données d'apprentissage. Cette bibliothèque a une fonction *auto_arima* qui aide à définir une plage de valeurs (p,d,q,P,D,Q) ou à fixer certaines de ces valeurs. Cette fonction conserve la combinaison de (p, d, q, P, D, Q) qui rapporte la meilleure valeur AIC. Les modèles ARIMA sont étudiés ci-dessous par variables d'intérêts (infections et décès).

4.1.1 Modèle ARIMA

4.1.1.1 Cas des infections

Il est apparu évident en observant visuellement la courbe de l'échantillon d'apprentissage (Figure 3.7 et Figure 3.8) que nous avons une tendance (moyenne variable) sur nos données. En annexe 1, nous montrons les différentes transformations (différence, racine carrée, logarithmiques, cubiques) qui ont été testées dans le but de rendre la série stationnaire. Nous avons trouvé qu'une transformation logarithmique suivie d'une différentiation d'ordre 1 rendait la série stationnaire puisque la valeur-p du test de Dickey-Fuller[21] obtenue est plus faible que 0.05. Ceci nous a conduit

[21] L'hypothèse nulle (H0) pour ce test est qu'il existe une racine unitaire. L'hypothèse alternative diffère légèrement selon l'équation que l'on utilise. L'alternative de base est que la série chronologique est stationnaire (ou tendance-stationnaire). Ainsi, si l'on recherche la stationnarité (ce qui est généralement le cas), on veut rejeter H0.

à rejeter l'hypothèse nulle (H0) i.e., que les données (avec un lag) n'ont pas de racine unitaire et sont donc stationnaires. Ainsi, nous avons fixé le terme d'intégration $d = 1$.

Pour les valeurs non négatives $p, q,$ qui pourraient être les plus probables, nous avons analysé les graphes de fonction d'autocorrélation (ACF) et de la fonction partielle d'autocorrélation (PACF). Pour rappel, le graphique de l'ACF permet d'avoir une idée de la force de corrélation entre nos données chronologiques et leurs valeurs décalées (ou 'lag'). Sur le graphique, l'axe des ordonnées indique le coefficient de corrélation tandis que l'axe des abscisses donne l'ordre sur le nombre de 'lags' ou décalages. Par exemple, une corrélation entre les valeurs d'une série temporelle au temps t et les valeurs de la même série au temps t-1 traduirait une forte corrélation d'ordre 1 (représentant l'ordre de décalage).

Une autocorrélation partielle (PACF) est assez similaire à l'ACF dans le sens qu'il donne un aperçu du lien entre les points dans une série chronologique et des observations décalées ou antérieures dont on a pris soin de supprimer la contribution d'observations intermédiaires. Ainsi, au premier lag ou décalage (corrélation entre les valeurs de la série temporelle au temps t avec les points de la même série au temps t-1), l'ACF et le PACF sont identiques. Toutefois, au second décalage, le PACF mesure la corrélation entre les valeurs de la série chronologique au temps t et les valeurs des données au temps t-2 après avoir contrôlé pour la corrélation entre les points de données au temps t avec ceux du temps t-1.

Plus généralement [58], le nombre de grands pics (ou de décalages) avant une chute brutale vers zéro (coefficient de corrélation non significatif) dans le PACF suggère l'ordre du processus autorégressif alors pour l'ACF, cela suggère que le processus est plutôt considéré comme ayant une moyenne mobile, indiquant ainsi l'ordre pour le terme de la moyenne mobile. Nous reportons les graphes de l'ACF et de PACF à la **Figure 4.1**.

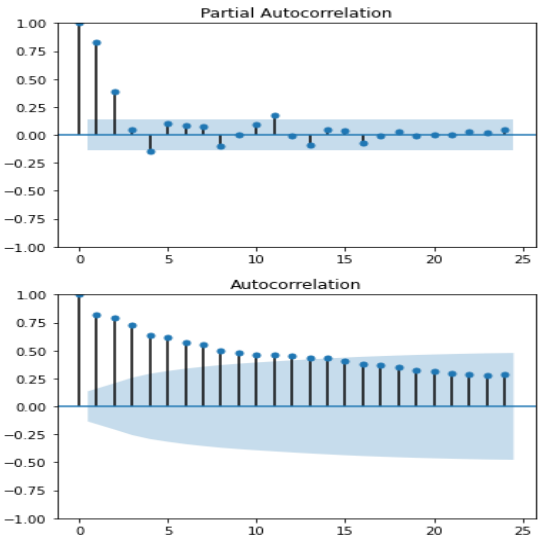


Figure 4.1. ACF et PACF issu du jeu de données stationnaires des infections

L'analyse visuelle du tracé laisse suggérer un point de coupure[22] autour de 3 (p=2 ou 4) pour le PACF alors qu'il est au-delà de 6 sur la courbe de l'ACF. L'absence évidente d'une chute brutale ou point de rupture sur la courbe de l'ACF rend difficile le choix des valeurs p et q. Pour pallier cette difficulté, à partir de ces points (p=2 ou 4), nous avons établi une grille servant à identifier le modèle ayant la plus faible valeur de AIC. Nous avons construit une grille sur des valeurs de p, d variant de 0 à 25. Pour rappel, certains articles [61], [62] proposent de fixer le nombre maximum de retards à 10 × log10(N) où N est la longueur de la série temporelle et log10 est la fonction logarithme base10. Dans notre cas, la valeur maximale serait autour de 30. Le résultat de notre balayage (fonction *auto_arima*) de paramètres est montré ci-dessous dans le **Tableau 4.1**.

[22] Pour estimer combien de termes sont associés à la composante AR ou MA, nous avons dénombré le nombre de segments en forme de boules au-dessus ou en dessous de l'intervalle de confiance définie comme la zone bleue. Bien évidemment, nous avons ignoré le premier terme situé au lag 0, car celle-ci exprime la corrélation parfaite entre l'observation de ce jour et soi-même.

Tableau 4.1 Paramètres du modèle ARIMA et analyse de la régression – infections

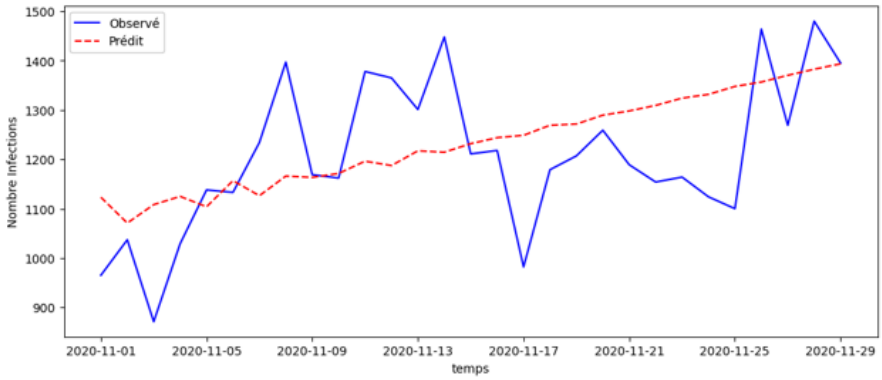
	Meilleur Modèle:	ARIMA (3,1,1)(0,0,0)[0]	
#	Coefficients	Ecart type	Valeur-p
ar.L1	-0.5198	0.032	0.000
ar.L2	0.7503	0.02	0.000
ar.L3	0.7661	0.015	0.000
ma.L1	0.8095	0.042	0.000
Sigma2	5.88E-05	1.83E-06	0.000
Analyse des résidu			
Ljung - Box (L1) (Q) :		2.43	
Prob(Q):		0.12	
Hétéroscédasticité (H) :		0.01	
Prob(H) (Deux-côtés) :		0.00	

Rappelons que Python utilise l'estimation du maximum de vraisemblance pour estimer les paramètres du modèle ARIMA. Lors du processus, le logiciel cherche les valeurs optimales (faible valeur de AIC) des hyperparamètres p, d, q et des paramètres du modèle qui offrent le plus de chance de répliquer les données que nous observons. La sortie du logiciel nous indique que les hyperparamètres du modèle qui correspondent le mieux à nos modèles vu nos données d'entrainement est ARIMA (3,1,1). La première valeur, 3, veut dire que pour le terme autorégressif est composé de trois décalages ou 'lags' qui sont *ar.L1*, ar.L2 et *ar.L3*. La seconde valeur, 1, indique que la série temporelle n'est pas stationnaire, nous devons donc prendre une différence de premier ordre et enfin la dernière valeur 1 nous indique que le modèle prend en compte le terme d'erreur d'une valeur précédente ou décalée, soit *ma.L1*. Pour déterminer si les paramètres du modèle ci-dessus définis sont statistiquement significatifs, il est usuel de devoir analyser la valeur-p de chaque coefficient du modèle. Nous rappelons que l'hypothèse nulle est que le coefficient n'est pas significativement différent de 0, ce qui indique qu'il est moins judicieux de le conserver dans le modèle. Habituellement, un niveau de signification (noté α) de 0,05 est admis. Nos résultats affichés dans le

Tableau 4.1 laissent croire d'une manière plus générale que :

- Les coefficients de la régression sont statistiquement significatifs[23] au seuil de 5%.
- Le modèle remplit la condition d'indépendance dans les résidus car la valeur-p du test de Ljung-Box[24] [63] (Prob(Q)) est de 0.12, ce qui est supérieur à 0,05. Ainsi, nous ne pouvons donc pas rejeter l'hypothèse nulle d'indépendance.
- On ne peut pas dire que la distribution résiduelle est homoscédastique (variance constante) car la valeur p du test d'hétéroscédasticité (Prob(H)) est inférieure à 0,05.

Bien que le modèle ne passe pas le test d'hétéroscédasticité[25], nous pensons que parfois il est difficile de trouver un modèle qui satisfasse à tous les tests. Ainsi, nous présentons ci-dessous la performance des modèles sur la population de test (soit novembre 2020).



Performance sur donnée test - Nov20

# Model	MAPE	RMSE	Moyenne (Observé)	Moyenne (prédit)
(4,1,2)	37	11	27.1	21.1

Figure 4.2. Performance du modèle ARIMA sur l'échantillon de test - infections.

[23] On rejette H0 : $\beta 1 = 0$ au niveau 5%; La valeur de p-value pour ces coefficients est inférieure à 0,05, ces coefficients sont donc statistiquement significatifs au niveau de 0,05. Toutefois, les valeurs-p pourraient être faussées à cause de la sélection de modèle basée sur le AIC.

[24] La statistique Ljung-Box Q (LBQ) teste l'hypothèse nulle selon laquelle les autocorrélations jusqu'au décalage k (lag) sont égales à zéro. En substance, les hypothèses sont :

Ho : Les données sont indépendamment distribuées. **H1** : les données ne sont pas indépendamment distribuées; il y a corrélation en série.

[25]En rappel, H0 : Homoscédasticité vs. H1 : Hétéroscédasticité

Dans la figure **Figure 4.2** ci-dessus, les valeurs prédites sont appelées ‘Forecast’ alors que celles observées sont notées ‘Observed’. Les résultats montrent qu’en moyenne, le modèle sous-estime le nombre d’infections journaliers pour le mois de novembre avec un taux d’erreur moyen estimé par le MAPE de 10% moyen. De même, le modèle ne semble pas capter les fluctuations journalières observées. Ceci justifie la valeur de RMSE assez loin de zéro.

4.1.1.2 Cas des décès

Pour rappel, les observations utilisées dans l’échantillon d’entrainement pour la variable décès sont représentées sur la **Figure 3.8** pour les cas journaliers et **Figure 3.4** pour les valeurs cumulatives journalières. Nous avons adopté les mêmes étapes décrites ci-dessus à la section 4.1.1.1 pour le cas des infections, i.e. nous avons utilisé la transformation logarithmique suivie d’un lag d’une unité pour rendre la série chronologique stationnaire, soit $d = 1$. Les tracés ACF et PACF affichés sur la **Figure 4.3**

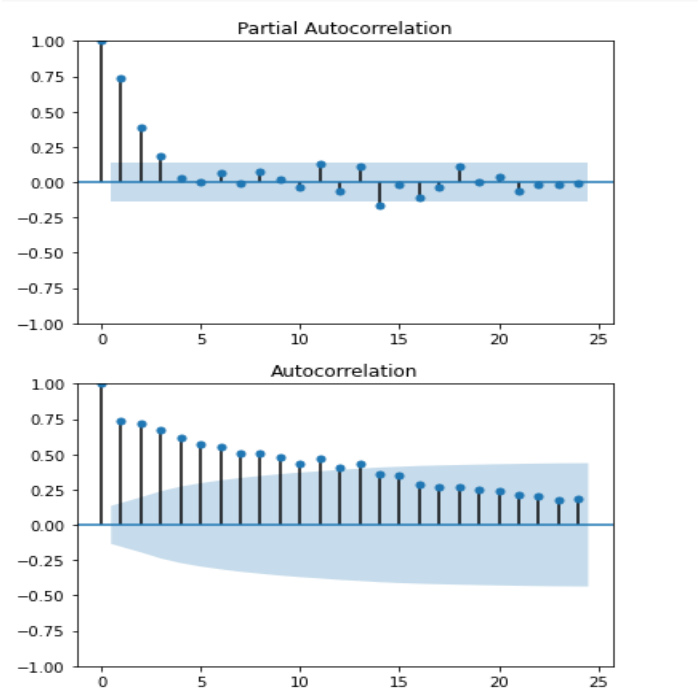


Figure 4.3. ACF et PACF issu du jeu de données stationnaires des décès

On observe pour le PACF ce qui ressemble à une décroissante des corrélations partielles vers zéro (définie comme la zone bleue) après 3 décalages ou lags alors que sur le graphique de l'ACF, on observe une série de corrélations qui demeure relativement large et significative (au-dessus de la zone bleue) au-delà de 13 décalages. Ceci laisse suggérer qu'on ne peut pas diagnostiquer avec précision à partir des graphes d'ACF et PACF les valeurs des paramètres p et q pour un modèle ARIMA. Toutefois, étant donné que notre série a été rendue stationnaire[26], l'analyse de l'ACF et PACF a été complétée par l'utilisation d'une grille (les valeurs de p, d variant de 0 à 25) servant à identifier le modèle ayant la plus faible valeur de AIC. Le résultat[27] de notre balayage (fonction *auto_arima*) de notre modélisation est montré ci-dessous dans le **Tableau 4.2**.

Tableau 4.2 Paramètres du modèle ARIMA et analyse de la régression – décès

	Meilleur Modèle:	ARIMA (4,1,2)(0,0,0)[0]	
#	Coefficients	Ecart type	Valeur-p
ar.L1	-1.39	0.017	0.000
ar.L2	0.0539	0.006	0.000
ar.L3	1.4258	0.016	0.000
ar.L4	0.9100	0.012	0.000
ma.L1	1.2045	0.036	0.000
ma.L2	0.6025	0.057	0.000
Sigma2	6.00E-03	1.77E-05	0.000
Analyse des résidu			
Ljung - Box (L1) (Q) :		2.52	
Prob(Q):		0.11	
Hétéroscédasticité (H) :		0.00	
Prob(H) (Deux-côtés) :		0.00	

[26] Pour rappel, nous avons fait une transformation logarithmique suivie du calcul de la différence première de la série (également appelée différenciation de la série) pour rendre la série stationnaire.

[27] Notons que le modèle ARIMA (5,1,1) est celui ayant la valeur la plus faible de l'AIC, toutefois, étant donné que ce modèle rejette l'hypothèse nulle de Ljung-Box relative à relative à l'indépendance des résidus, nous avons opté pour plus proche modèle ARIMA (4,1,1) n'ayant pas ce problème.

Dans le **Tableau 4.2** est indiqué le modèle qui correspond le mieux à nos données d'entrainement, c-à-d ARIMA (4,1,2). La première valeur, 4, veut dire que pour le terme autorégressif est composé de trois décalages ou 'lags' qui sont *ar.L1*, *ar.L2*, *ar.L3* et *ar.L4*. La seconde valeur, 1, indique que la série temporelle n'est pas stationnaire, nous devons donc prendre une différence de premier ordre et enfin la dernière valeur 2 nous indique que le modèle prend en compte le terme de la moyenne mobile qui est caractérisée par les valeurs *ma.L1* et *ma.L2*.

D'une manière plus générale :

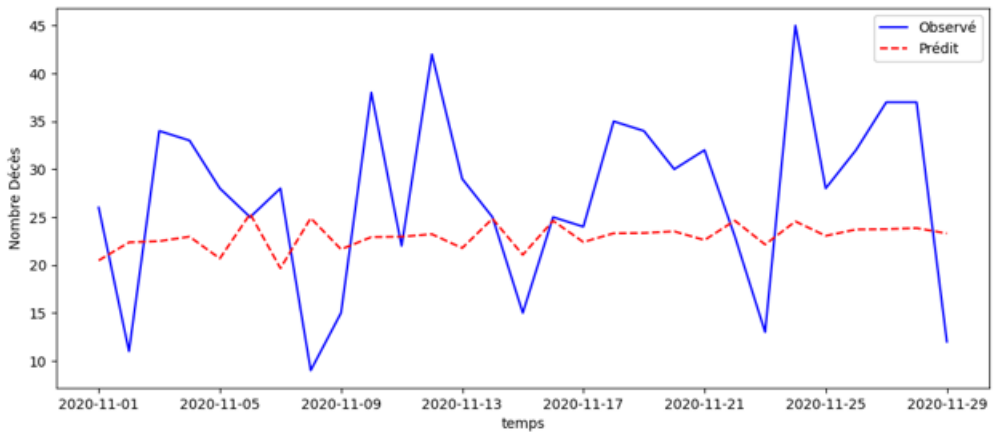
- Les coefficients de la régression sont statistiquement significatifs[28] au seuil de 5%.
- Le modèle remplit la condition d'indépendance des résidus car la valeur-p du test de Ljung-Box[29] [63] (Prob(Q)) est de 0.11, ce qui est supérieur à 0,05. Ainsi, nous ne pouvons donc pas rejeter l'hypothèse nulle d'indépendance.
- On ne peut pas dire que la distribution résiduelle est homoscédastique (variance constante) car la valeur p du test d'hétéroscédasticité (Prob(H)) est inférieure à 0,05.

Bien que le modèle ne passe pas le test d'hétéroscédasticité, nous pensons que parfois il est difficile de trouver un modèle qui satisfasse à tous les tests. Ainsi, nous présentons ci-dessous dans la **Figure 4.4** la performance des modèles sur la population de test (soit novembre 2020).

[28] On rejette H0 : $\beta 1 = 0$ au niveau 5%; La valeur-p pour ces coefficients est inférieure à 0,05, ces coefficients sont donc statistiquement significatifs au niveau de 0,05

[29] La statistique Ljung-Box Q (LBQ) teste l'hypothèse nulle selon laquelle les autocorrélations jusqu'au décalage k (lag) sont égales à zéro. En substance, les hypothèses sont :

Ho : Les données sont indépendamment distribuées. **H1** : les données ne sont pas indépendamment distribuées; il y a corrélation en série.



Performance sur donnée test - Nov20

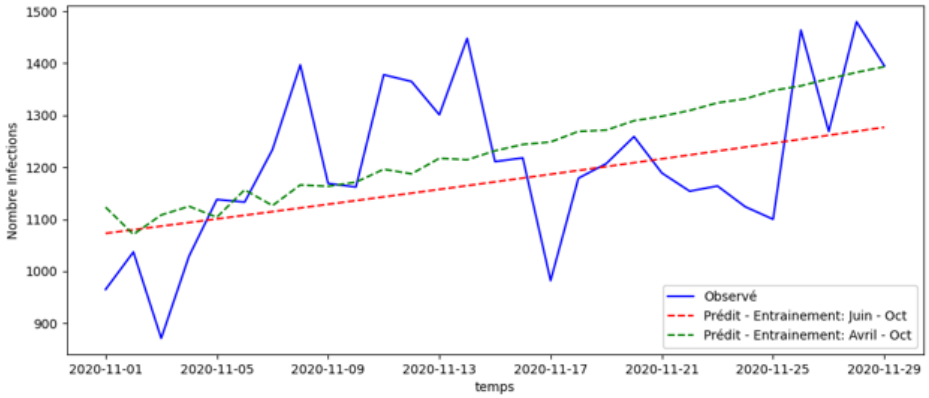
# Model	MAPE	RMSE	Moyenne (Observé)	Moyenne (prédit)
(4,1,2)	37	11	27.1	21.1

Figure 4.4. Performance du modèle ARIMA sur l'échantillon test - décès.

Les résultats montrent qu'en moyenne, le modèle sous-estime le nombre de décès journaliers pour le mois de novembre avec un taux d'erreur moyen estimé par le MAPE autour de 37%. Pire encore, le modèle ne semble pas capter les fluctuations journalières.

4.1.1.3 Analyse de sensibilité – modèles ARIMA

Nous avons testé la sensibilité du modèle ARIMA à la segmentation utilisée pour définir la période d'entrainement. Pour cela, nous avons bâti nos modèles sur des données d'entrainement allant d'avril 2020 à octobre 2020 et de juillet 2020 à octobre 2020. Ci-dessous dans la **Figure 4.5** nous comparons la performance de ces échantillons d'entrainement sur la période de performance de novembre 2020.

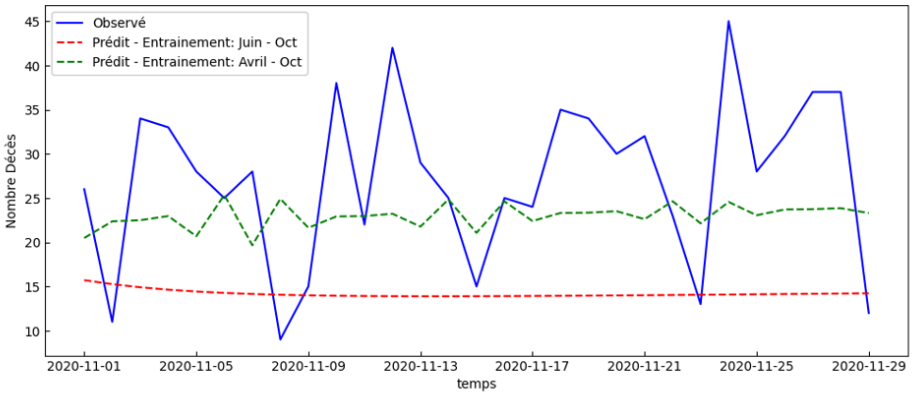


Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Avril20 -Oct20	9.8	140.6	1207	1234
Juin20-Oct20	8.9	138.4		1173

Figure 4.5. Impact de la taille de l'échantillon d'entrainement sur la performance du modèle ARIMA - infections.

On note que se limiter uniquement aux données les plus récentes (juin 2020 à octobre 2020) ne semble pas apporter une amélioration à la performance du modèle. Les valeurs MAPE et RMSE sont comparables. Ci-dessous dans la **Figure 4.6**, nous présentons la même analyse dans le cas des décès.



Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Avril20-Oct20	37	11	27.1	21.1
Juin20-Oct20	46.3	16		14.2

Figure 4.6. Impact de la taille de l'échantillon d'entrainement sur la performance du modèle ARIMA - Décès

On note que le nouveau modèle obtenu sur les données d'avril 2020 à octobre 2020 a une performance moins bonne se traduisant par une différence de MAPE et RMSE d'au moins 20% en comparaison du premier modèle bâti sur les données d'avril 2020 à octobre 2020. Ceci laisse envisager que la base de données d'entrainement d'avril à octobre serait plus appropriée pour la modélisation.

4.1.2 Modèle SARIMA

Étant donné que le modèle SARIMA peut être vu comme une extension du modèle ARIMA par l'inclusion de la composante de saisonnalité, il est important de vérifier l'ordre de la saisonnalité noté m dans l'Eq.2.4. Une façon de procéder consiste à utiliser une décomposition de la série chronologique en faisant une hypothèse sur la nature additive ou multiplicative entre les trois composantes (tendance, saisonnière, résiduelle) du modèle SARIMA. Ce processus peut être facilement opéré avec la fonction *seasonal_decompose* incluse dans la librairie statistique de Python *statsmodels.tsa.seasonal*.

4.1.2.1 Cas des infections

Le résultat de la décomposition de la série chronologique par l'approche additive en ses trois composantes (tendance, saisonnalité et résidu) est donné ci-après dans la Figure 4.7.

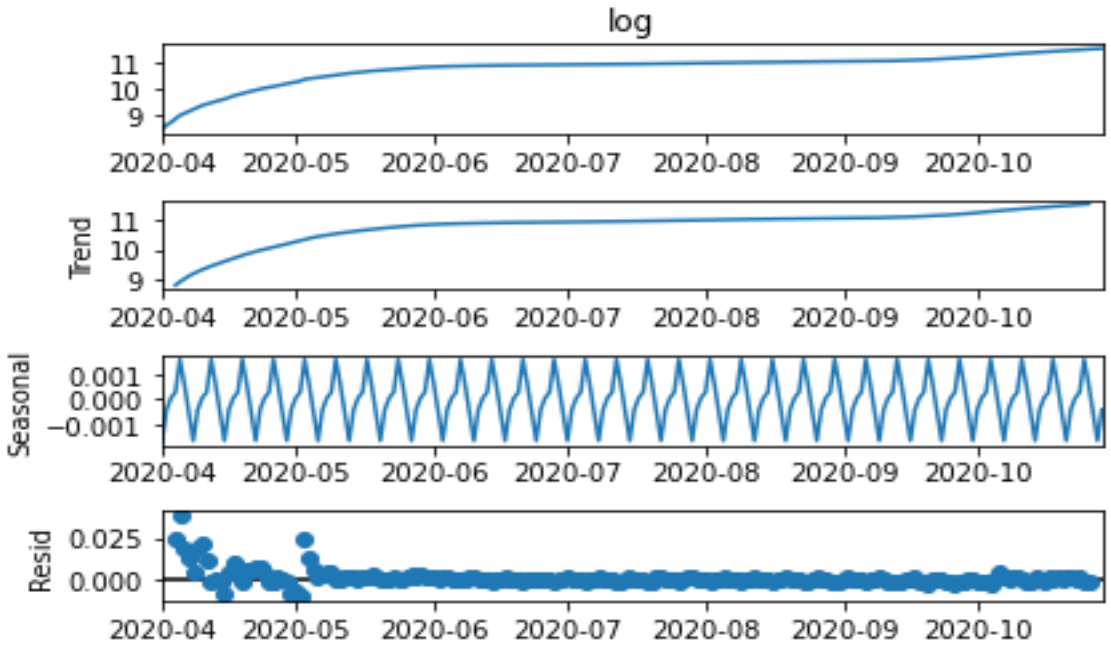


Figure 4.7. **Infection journalière cumulative (en haut) et ses trois composantes additives**

Les trois composantes additives (tendance, saisonnalité et résidus) sont présentées dans les trois derniers panneaux inférieurs de la **Figure 4.7**. Évidemment, en additionnant ces trois composantes, on peut générer la série chronologique affichée tout en haut de la même figure. D'après le graphique, il apparait qu'il y a environ 4 ou 5 pics identiques par mois ou un pic par semaine. Ce qui laisse suggérer que chaque semaine, il pourrait avoir un pattern sur le nombre cumulatif (logarithme) d'infections. Ainsi m=7 pourrait constituer donc un exemple de saisonnalité. Notons que si on ajoute les nouvelles données, l'ordre de saisonnalité pourrait bien évidemment être différent, car les observations récentes (santé publique du Québec) sur les données de la COVID-19 laissent suggérer que les cas tendent à augmenter suite au retour des élèves à l'école (hiver) et à baisser pendant les congés ou relâches scolaires.

Ainsi, en admettant que l'ordre de périodicité est autour de 7, nous avons refait de nouveau la recherche des paramètres optimaux du modèle avec les valeurs de la grille choisies telles que p, d varient de 0 à 25 alors que celles de P et Q varient de 0 à 15. Les résultats sont montrés ci-dessous dans le **Tableau 4.3**

Tableau 4.3 Paramètres du modèle SARIMA (ordre périodicité = 7) et analyse de la régression - infections

	Meilleur Modèle:	SARIMA (2,1,0)(0,1,2)[7]	
#	Coefficients	Ecart type	Valeur-p
ar.L1	0.312	0.022	0.000
ar.L2	0.540	0.02	0.000
ma.S.L7	-0.269	-0.269	0.000
ma.S.L14	0.021	0.053	0.695
Sigma2	7.28E-05	2.56E-06	0.000
Analyse des résidu			
Ljung - Box (L1) (Q) :		0.13	
Prob(Q):		0.71	
Hétéroscédasticité (H) :		0.02	
Prob(H) (Deux-côtés) :		0.00	

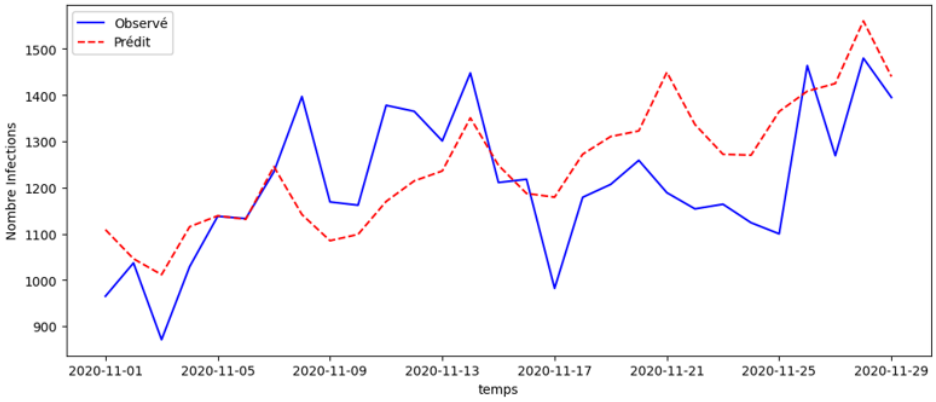
La sortie du logiciel nous indique que le modèle correspondant le mieux à nos données d'entrainement est SARIMA (2,1,0)(0,1,2)[7]. Comme expliqué précédemment, pour (2,1,0), la 1ère valeur, 2, veut dire que le terme autorégressif est composé de deux termes de décalage ou 'lags' qui sont *ar.L1* et *ar.L2*. La seconde valeur, 1, indique que nous avons fait une différence de premier ordre sur la série chronologique pour la rendre stationnaire. De façon analogue, (0,1,2) représente le terme de saisonnalité constitué du terme moyen mobile *ma.S.L7* et *ma.S.L14*. Ce terme saisonnier est stationnaire si on applique une différence de premier ordre. D'une manière plus générale :

- Les coefficients de la régression sont statistiquement significatifs[30] au seuil de 5%, à l'exception du terme de la moyenne mobile *ma.S.L14.*
- Le modèle remplit la condition d'indépendance des résidus car la valeur-p du test de Ljung-Box [63] (Prob(Q)) est de 0.71, ce qui est supérieur à 0,05. Ainsi, nous ne pouvons donc pas rejeter l'hypothèse nulle d'indépendance.

[30] On rejette H0 : $\beta 1 = 0$ au niveau 5%; La valeur-p pour ces coefficients est inférieure à 0,05, ces coefficients sont donc statistiquement significatifs au niveau de 0,05

- On ne peut pas dire que la distribution résiduelle est homoscédastique (variance constante) car la valeur-p du test d'hétéroscédasticité (Prob(H)) est inférieure à 0,05.

L'analyse de la performance du modèle sur l'échantillon test est donnée ci-dessous dans **la Figure 4.8.**



Performance sur donnée test - Nov20

# Model	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
(2,1,0)x(0,1,2)[7]	8.9	130.2	1207	1216

Figure 4.8. Performance du modèle SARIMA sur l'échantillon de test - infections.

Les résultats montrent qu'en moyenne, le modèle prédit assez bien le nombre d'infections journalières pour le mois de novembre avec un taux d'erreur moyen estimé par le MAPE moins de 10% en moyenne. Le modèle semble aussi refléter les fluctuations journalières des infections.

4.1.2.2 Cas des décès

Comme ci-dessus, le résultat de la décomposition de la série chronologique par l'approche additive en ses trois composantes (tendance, saisonnalité et résidu) est donné ci-après dans la **Figure 4.9**

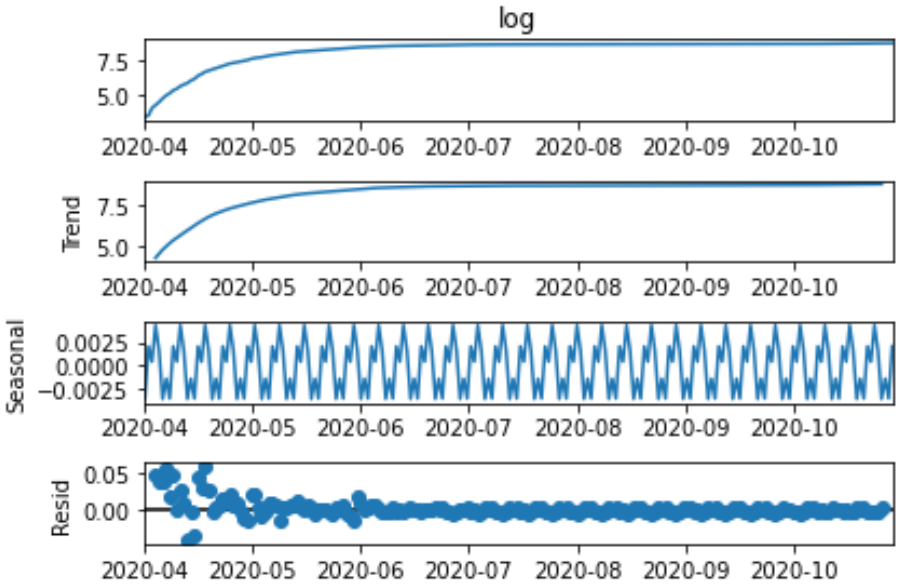


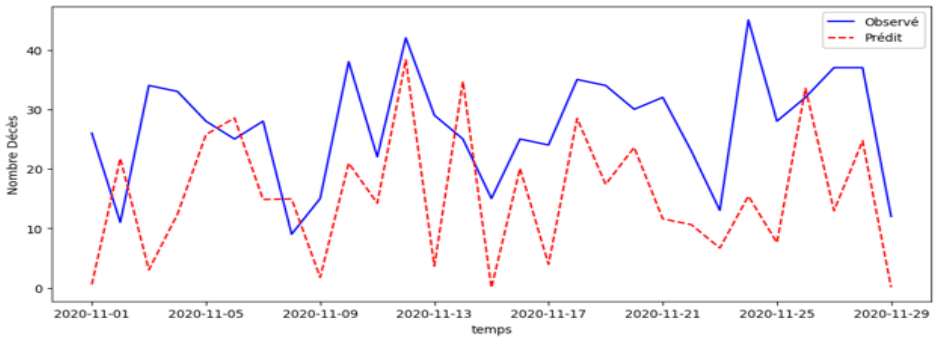
Figure 4.9. **Infection journalière décès (en haut) et ses trois composantes additives**

Il apparait qu'il y aurait environ 4 ou 5 pics identiques par mois ou un pic par semaine. Ceci laisse envisager un ordre de périodicité proche de 7. Pour déterminer les paramètres du modèle, nous avons utilisé la même grille définie comme dans le cas des infections. Les résultats obtenus sont affichés dans le **Tableau 4.4** ci-dessous.

Tableau 4.4 Paramètres du modèle SARIMA (ordre périodicité = 7) et analyse de la régression - décès

	Meilleur Modèle:	SARIMA (3,1,1)(1,1,0)[7]	
#	Coefficients	Ecart type	Valeur-p
ar.L1	-0.61	0.017	0.000
ar.L2	0.718	0.014	0.000
ar.L3	0.873	0.013	0.000
ma.L1	0.059	0.037	0.000
ar.S.L7	-0.743	0.019	0.000
Sigma2	6.00E-03	2.87E-05	0.000
Analyse des résidu			
Ljung - Box (L1) (Q) :		1.32	
Prob(Q):		0.25	
Hétéroscédasticité (H) :		0.00	
Prob(H) (Deux-côtés) :		0.00	

Comme dans les cas précédents, le modèle $SARIMA(3,1,1)(1,1,0)[7]$ remplit la condition d'indépendance dans les résidus car la valeur p du test de Ljung-Box est de 0.25, ce qui est supérieur à 0,05. Notons tout de même qu'un des coefficients de la régression n'est pas significatif au seuil de 5%. L'analyse de la performance du modèle est montrée ci-dessous à la Figure 4.10.



Performance sur donnée test - Nov20

# Model	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
(3,1,1)x(1,1,0)[7]	54	16.2	27.1	16

Figure 4.10. Performance du modèle SARIMA sur l'échantillon de test - décès.

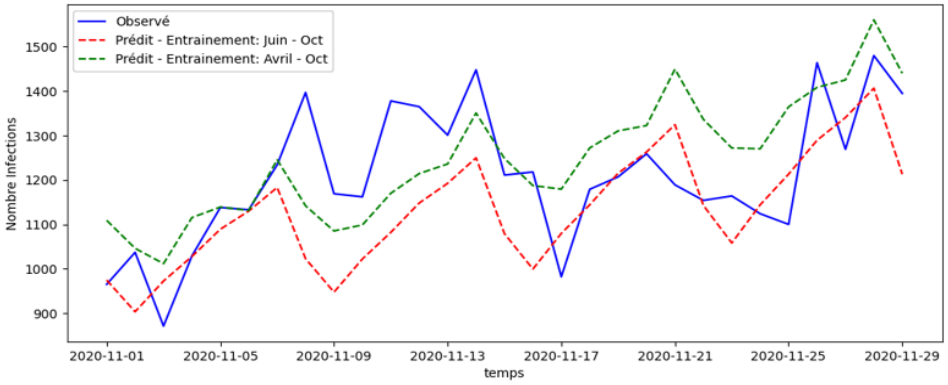
Les résultats montrent qu'en moyenne, le modèle tend à sous-estimer le nombre de décès journaliers pour le mois de novembre avec un taux d'erreur moyen estimé par le MAPE à environ 54%. Toutefois, le modèle semble aussi répliquer les fluctuations journalières des décès. La sensibilité du modèle est discutée ci-dessous.

4.1.2.3 Analyse de sensibilité – modèle SARIMA

Nous analysons ici la sensibilité de la performance du modèle à la segmentation et de l'ordre de la périodicité.

4.1.2.3.1 Impact de la segmentation de l'échantillon de 'Training'

Nous avons comparé la performance de deux modèles bâtis respectivement sur un échantillon d'entrainement couvrant la période d'avril à octobre 2020 et de juin à octobre 2020. Les données de l'échantillon test sont celles du mois de novembre 2020. Ci-dessous dans la **Figure 4.11** nous présentons l'analyse de la segmentation pour les Infections.



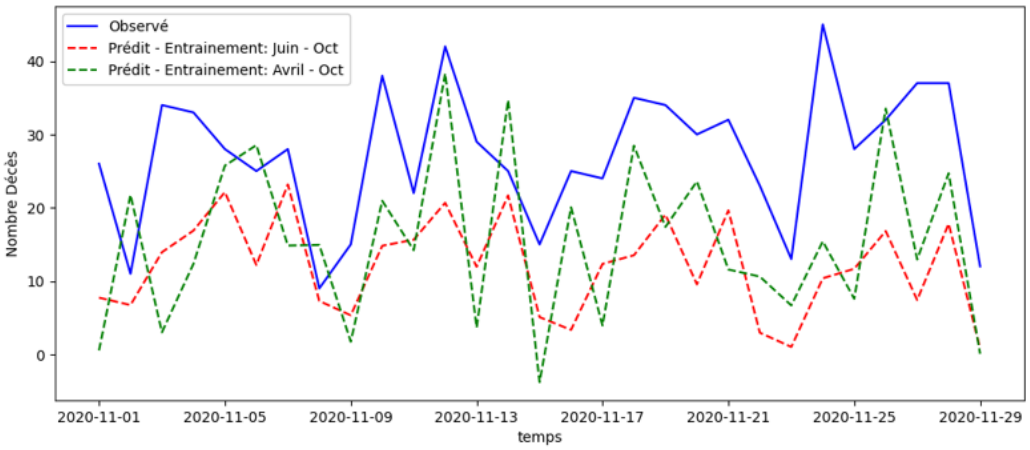
Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Avril20-Oct20	8.9	130.2	1207	1216
Juin20-Oct20	9.1	146		1133

Figure 4.11. Impact de la taille de l'échantillon d'entrainement sur la performance du modèle SARIMA - infections.

On note qu'en général, les deux modèles semblent répliquer les tendances des nouvelles infections pour le mois de novembre. De même, nos analyses montrent que le modèle obtenu de l'échantillon

d'entrainement de juin à octobre sous-estime en général le nombre d'infections alors que son homologue basé sur les données d'avril à octobre tend à surestimer les infections. Notons enfin que les RMSE et MAPE des deux modèles sont assez comparables (la différence est moins de 10%). Ci-dessous dans la **Figure 4.12,** les résultats reliés au cas des décès sont montrés.



Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Avril20-Oct20	54	16.2	27.1	16
Juin20-Oct20	32	11.5		18

Figure 4.12. Impact de la taille de l'échantillon d'entrainement sur la performance du modèle SARIMA - Décès.

On note que changer la taille de l'échantillon de 'Training' à un impact peu considérable sur la performance des modèles. La performance des deux modèles étant très semblable.

4.1.2.3.2 Impact de l'ordre de la périodicité

Pour analyser la sensibilité de notre modèle par rapport à l'ordre de la périodicité, nous avons pour chaque valeur de l'ordre de la périodicité m, cherché le modèle optimal (faible AIC) sur une grille dont les paramètre p, d, q, P, D, Q du modèle varient entre 0 et 20. Ci-dessous dans le **Tableau 4.5** nous présentons les résultats obtenus dans le cas des infections et décès.

Tableau 4.5 Sensibilité du modèle SARIMA à l'ordre de périodicité m.

	Echantillon Décès				
	Performance sur donnée test, Nov20				
périodiocité	**# Model**	**MAPE**	**RMSE**	**Moyenne(observed)**	**Moyenne(Prédit)**
m=5	(2,1,2)x(3,1,0)[5]	43	17	27.1	16.65
m=7	(3,1,1)x(1,1,0)[7]	54	16.2		16
m=10	(3,1,1)x(1,1,0)[10]	35	11.7		20.1
m=30	(3,1,1)x(0,1,1)[30]	39	11		21.33

	Echantillon Infections				
	Performance sur donnée test, Nov20				
périodiocité	**# Model**	**MAPE**	**RMSE**	**Moyenne(observé)**	**Moyenne(Predit)**
m=5	(2,1,0)x(2,1,0)[5]	13.9	191	1207	1317
m=7	(2,1,0)x(0,1,2)[7]	8.9	130.2		1216
m=10	(2,1,1)x(0,1,1)[10]	14	200.4		1331
m=30	(1,1,2)x(0,1,1)[30]	14	214		1296

Rappelons qu'augmenter la périodicité augmente significative le temps de calcul. Pour cela, nous nous sommes limités à *m=30*. Les résultats de notre modélisation montrent qu'en général, dans le cas des décès, l'augmentation de la multiplicité semble améliorer légèrement (moins de 20%) la performance des modèles selon nos valeurs de MAPE et RMSE. Toutefois, cette tendance n'est pas observée dans le cas des infections où on note qu'il n'y a pas un lien assez clair entre l'augmentation de l'ordre de la périodicité et la performance du modèle.

4.2 Les Modèles statistiques Bayésiennes: Prophet

4.2.1 Brève comparaison entre Prophet et ARIMA ou SARIMA

Pour mieux comprendre Prophet, il convient de rappeler ici sa différence fondamentale avec les modèles ARIMA ou SARIMA. Le principe de *Prophet* repose comme SARIMA ou ARIMA sur la décomposition de la série chronologique en ses termes de tendance, saisonnalité et de résidu. Dans le cas de *Prophet*, une composante de type binaire est ajoutée pour refléter les jours et événements particuliers de l'année. La différence fondamentale réside sur la façon de modéliser ces composantes. Par exemple pour le terme de la tendance, l'approche ARIMA ou SARIMA utilise

une approche stochastique [31], dans le cas de *Prophet* nous avons indiqué au chapitre précédent que l'approche est déterministe avec une pente fixe dans l'équation de la tendance. Bien évidemment, cette dernière hypothèse ne serait pas adéquate si *Prophet* ne faisait pas appel à des points de rupture qui permettent à la fonction de tendance de changer autour d'un point de rupture ou entre deux fenêtres consécutives. Cette aptitude à changer de pente rend ainsi le modèle comparable à l'approche ARIMA ou SARIMA. Notons enfin que la flexibilité de Prophet de changer de pente entre deux fenêtres devrait la rendre moins sensible aux données aberrantes ou manquantes (des observations différentes étant utilisées pour ajuster la pente de la tendance entre deux fenêtres). Ainsi, la qualité d'un modèle Prophet dépendrait du choix judicieux de ces points de rupture.

4.2.2 Modèle Prophet – implémentation

Nous avons utilisé la bibliothèque *Prophet* lancée par Facebook sous la forme d'API[32]. Pour utiliser la librairie, les éléments à fournir sont la colonne date et la colonne de la variable cible (notons que Prophet offre la possibilité d'inclure des variable explicatives, ce cas n'est pas discuté ici). En outre, il faut convertir la colonne de date au format *Date Heure*, puis renommer les deux colonnes en « ds » pour la date et « y » pour la cible.

- Afin d'ajuster le modèle défini par l'équation 2.7, *Prophet* utilise un ensemble d'équations linéaires par morceaux avec des pentes différentes entre les points de changement (point de rupture). Par défaut, *Prophet* spécifie 25 points de changement potentiels qui sont uniformément placés dans les premiers 80% des points de la série temporelle, mais seuls les points de changement pertinents sont conservés. Notons qu'il est aussi possible à l'utilisateur de spécifier ses points de rupture potentiels dont certains pourraient être rejetés par le modèle.

---

[31] Nous avons fait des tests de racine unitaire pour savoir si les données sont stationnaires ou tendancielles. Dans ce dernier cas, nous avons fait un mélange de transformation logarithmique et de différentiation pour rendre la série stationnaire. Ceci bien évidemment définit la composante stochastique de l'approche.

[32] API *Application Programming Interface* permet l'utilisateur de manipuler les données avec moins de programmations, car les fonctions sont pré-optimisées, adaptables et prêtes à l'emploi

- Pour implémenter l'équation 2.8, une série de Fourier (la somme de nombreux sinus et cosinus successifs) est utilisée et l'ordre de saisonnalité choisie est hebdomadaire comme dans les modèles SARIMA.

Pour rappel, nous avons aussi choisi de travailler avec les valeurs transformées par la fonction logarithmique car nos tests préliminaires sur nos données ont montré leur supériorité en termes de RMSE par rapport aux données non transformées.

4.2.2.1 Modèle Prophet – choix des paramètres.

- **Terme de tendance**

Nous avons choisi l'approche linéaire pour définir la tendance dans chaque fenêtre. Comme décrit dans la section 2.1.2.1, l'un des paramètres est le paramètre d'échelle de la distribution de Laplace[33] qui ajuste la flexibilité de la tendance $\delta_j \sim Laplace\ (\mu, \tau)$, (voir.Eq.2.7) où μ est le paramètre de position et τ étant le facteur multiplicatif ou d'échelle. Sans aller loin dans les détails mathématiques, une illustration de la fonction de masse de la distribution de Laplace est montrée dans la Figure 4.13 (source: Wikipedia[34]) avec $\mu = 0, \tau = b$.À

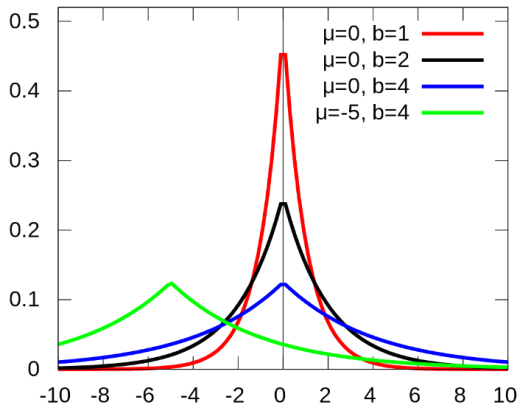


Figure 4.13. Illustration de la fonction de masse de la Loi de Laplace (source: Wikipedia)

[33] La loi de Laplace est aussi largement connue sous le nom de distribution doublement exponentielle.

[34] Source: https://en.wikipedia.org/wiki/Laplace_distribution

On note que lorsque $\mu = 0$, la densité de masse (probabilité) est très importante autour du point[35] 0. Ainsi,

- Si τ est faible, il est plus probable d'avoir plus de points tel que $\delta_j \to 0$, ce qui veut dire que les points de rupture en j seront ignorés ou rejetés par l'algorithme. Ceci pourrait engendrer un sous apprentissage du modèle qui va fonctionner avec moins de fenêtres ou de points de rupture qu'il n'en faut (bien qu'il n'ait aucun moyen de connaitre le nombre exact). Le modèle ne pourra probablement pas capter adéquatement la dynamique de variations des tendances dans nos données (hausse ou baisse des infections).
- Si τ élevé, il est plus probable d'avoir plus de points tel que $\delta_j \neq 0$, ce qui veut dire que nous pourrions avoir plus de points de ruptures dans les modèles qu'il n'en faut et ceci peut engendrer un surapprentissage du modèle.

Exemple d'illustration.

Afin de mieux illustrer ce processus clé de de la libraire Prophet, supposons une série chronologique avec 10 points sur laquelle nous voulons implémenter le modèle défini par l'Eq.2.7.

$$g(t) = (k + \boldsymbol{a}(t)^T \boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^T \boldsymbol{\gamma}).$$

a) Considérons k=1, m=5. Nous rappelons que k représente le taux de croissance de la pente dans une fenêtre, alors que m est utile pour relier deux pentes issues de deux fenêtres consécutives afin de rendre continue les droites des pentes. La matrice $\boldsymbol{a}(t)$ est simplement un identificateur (0,1) utile pour ajuster le taux de croissance, k, dans la fenêtre considérée (elle vaut 0 avant cette fenêtre). Enfin, $\gamma_j = -s_j \delta_j$.

b) Définissions 4 points de rupture aux points 1,3,5 et 8

35 Soit une variable aléatoire U qui suit dans l'intervalle [-1/2, 1/2] une loi uniforme continue, dans ce cas,

$\delta = -\tau * sgn(U) * ln(1 - 2|U|)$ est distribué selon la loi de Laplace de paramètres 0 et τ

c) Supposons que le résultat que du choix aléatoire de quatre points provenant de la distribution de Laplace est [-0.55, -0.78, 0.01, 0.65]

Ainsi, comme on peut le voir dans la Figure 4.14, le 3e point de rupture est rejeté par le modèle car $\delta_j \rightarrow 0$. Le script en Python qui permet d'implémenter notre exemple est donné en Appendix 3 de ce mémoire.

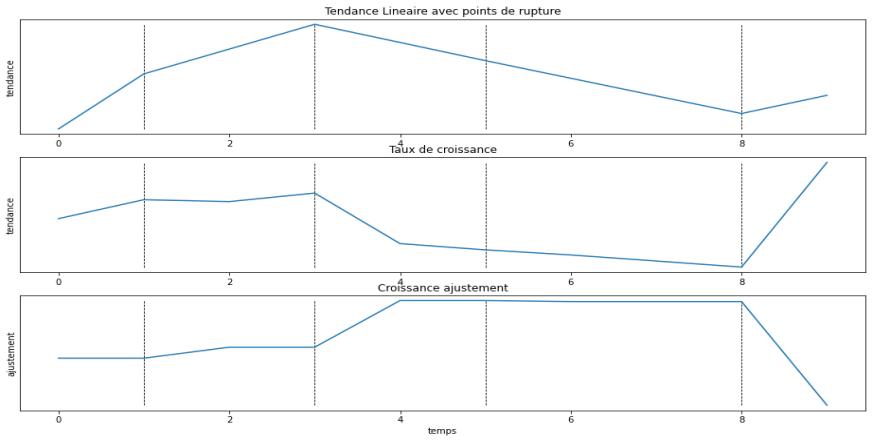


Figure 4.14. Illustration l'implémentation de la fonction de croissance utilisée dans la librairie de Prophet.

- Il est important de noter ici que la majeure partie de notre analyse ci-dessus a porté sur les points de rupture en lien avec la fonction d'apriori qui suit la loi de Laplace. Nous devons rappeler ici que l'Eq. 2.7 est en fait un produit entre la distribution de la fonction d'apriori et la fonction de vraisemblance. Ainsi, une fois la fonction de croissance déterminée (la pente), il est important que cette pente reflète les valeurs observées dans les données pour la fenêtre considérée, ou encore que la valeur de la vraisemblance basée sur ce paramètre soit non nulle.

- **Terme de saisonnalité**

L'ordre, *N*, de l'équation de Fourier (Eq.2.8) définit à quelle vitesse la saisonnalité peut changer. Étant donné que nous avons admis une saisonnalité hebdomadaire, nous avons choisi la valeur par défaut (l'ordre N=3, et la période est P=7). La dernière étape de l'implémentation du modèle est de

définir la distribution de $\beta \sim N(0, \sigma^2)$ qui sert à définir le poids de la contribution du terme de saisonnalité dans le modèle. Ainsi le terme σ est associé à la variance de la loi du paramètre β qui régularise la force de saisonnalité. Un σ proche de zéro, revient pratiquement à exclure la contribution de la saisonnalité, ce qui entrainera un sous-apprentissage s'il y a de la saisonnalité dans nos données. De même, on aura un sur-apprentissage pour des valeurs de σ élevées. *Prophet* propose de prendre une valeur entre [0.01, 10], la valeur par défaut étant 10.

Notons que la saisonnalité peut être additive ou multiplicative comme dans le modèle SARIMA. Nous avons gardé l'approche additive.

Les deux paramètres σ, τ ci-dessus sont les seuls qui à notre avis pourraient avoir un potentiel impact sur le modèle. La sensibilité de nos résultats par rapport aux autres facteurs tel que le nombre de points rupture initiaux, la contribution des jours fériés, le taux de croissance) n'a pas étudié dans ce travail et la valeur par défaut a été utilisée. Une revue de la littérature [64] récente montre que l'impact sur la performance du modèle de ces valeurs n'a pas été aussi analysé.

Notons pour terminer que pour prédire le nombre de décès et d'infections, Prophet suppose que les changements de tendance des infections et décès dans le futur sont semblables à celles observées historiquement. En particulier, Prophet suppose que la fréquence et l'ampleur moyennes des changements de tendance à l'avenir seront les mêmes que celles observées historiquement.

4.2.2.2 Modèle Prophet – infections

Dans la Figure 4.15, les points de rupture sont représentés par des lignes rouges verticales. Par défaut ils étaient au nombre de 25 et équidistants, mais on en dénombre environ 18 car les autres ont été rejetés. Les points noirs représentent les valeurs observées alors que les lignes rouges entre les points de ruptures sont estimées par le modèle. On observe qu'en général sur l'échantillon d'entrainement, l'ajustement est assez correct. Nous pouvons également voir les prédictions faites sur les données de test au mois de novembre ainsi que l'intervalle de crédibilité y associée en bleu sur la Figure 4.15. Nous rappelons que les données ont connu une transformation logarithmique.

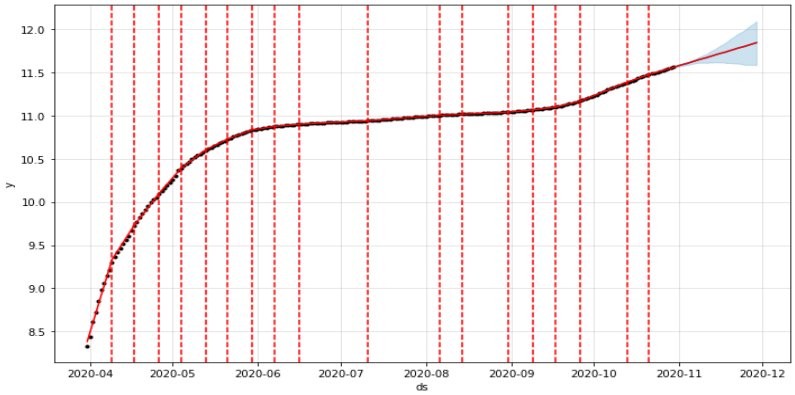
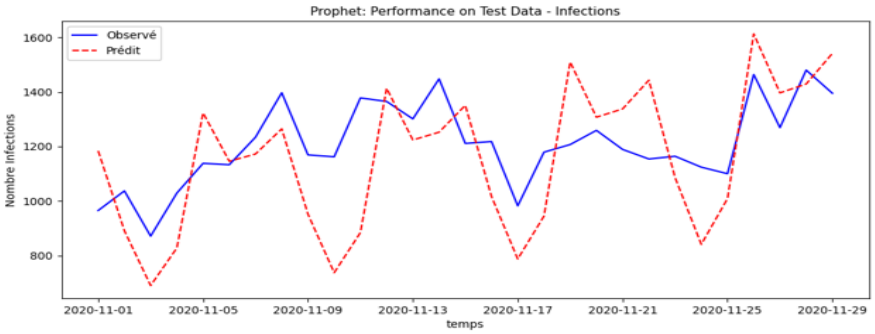


Figure 4.15. Performance du modèle Prophet sur l'échantillon cumulatif d'entrainement et test' – infections (échelle logarithmique)

Pour rappel, nous modélisons la cumulative (en réalité le logarithme de la cumulative) des cas d'infections et de décès. Ceci implique que nos résultats prédisent la cumulative de ces deux variables pour la période de test considérée et nous pouvons avoir les données sur les valeurs journalières. Ci-dessous sur la **Figure 4.16** nous pouvons voir en détail comment le modèle performe sur la population de test.



Performance sur donnée test - Nov20

MAPE	RMSE	Moyenne (Observé)	Moyenne (prédit)
14.9	205.7	1207	1157

Figure 4.16. Performance du modèle Prophet sur l'échantillon de test' - infections

Le modèle semble refléter les fluctuations journalières d'infection de l'échantillon de test. Ceci semble être corroboré par les valeurs relativement faibles de MAPE (moins de 15%). En moyenne, le modèle prédit 1157 infections journalier pour le mois de novembre contre 1207 observés

4.2.2.3 Modèle Prophet – décès

Ci-dessous dans la **Figure 4.17** nous affichons les résultats de nos modélisations dans le cas de la variable décès.

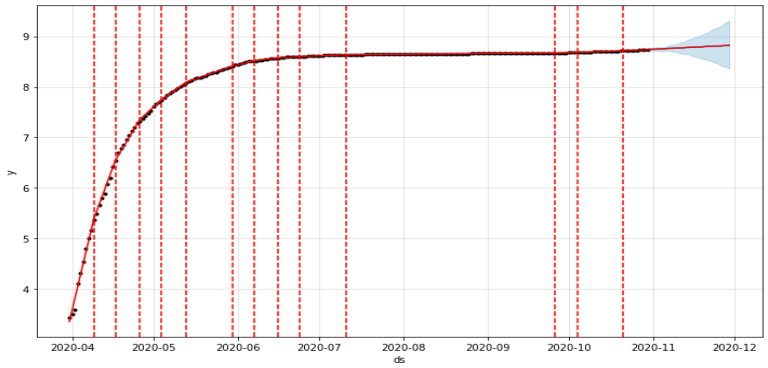
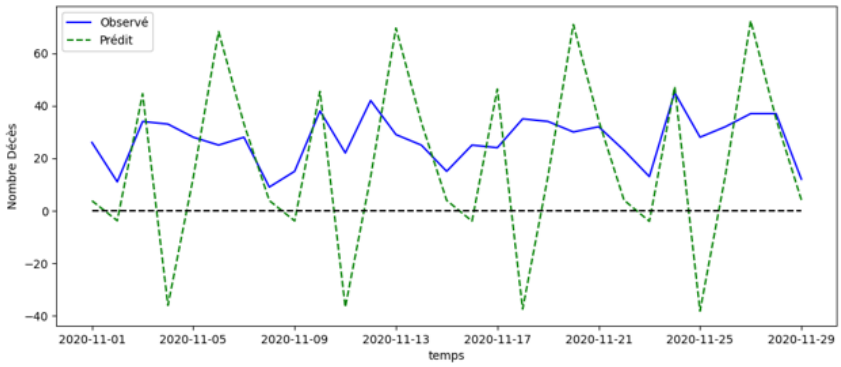


Figure 4.17. Performance du modèle Prophet sur l'échantillon cumulatif d'entrainement et test' – décès (échelle logarithmique)

On observe que sur les 25 potentiels points de rupture initiaux, environ la moitié a été rejetée. Une fois de plus, les valeurs estimées semblent bien s'ajuster aux valeurs observées sur l'échantillon d'entrainement. Ci-dessous dans la **Figure 4.18**, la performance sur l'échantillon de test est affichée.



Performance sur donnée test - Nov20

MAPE	RMSE	Moyenne (Observé)	Moyenne (prédit)
101	32	27.1	17.6

Figure 4.18. Performance du modèle Prophet sur l'échantillon de test' – décès

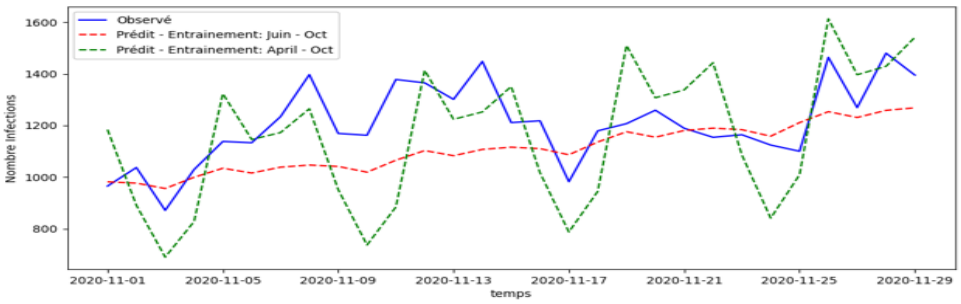
Il apparait que le modèle semble moins refléter les fluctuations journalières d'infection de l'échantillon de test. Ceci semble être corroboré par les valeurs de MAPE assez élevées comparé aux autres modèles étudiés. En moyenne, le modèle prédit 18 décès journalier pour le mois de novembre contre 27 observés. Notons que le modèle prédit des valeurs négatives d'infections. Ceci pourrait être attribué à la transformation logarithmique que nous avons effectuée sur les données. Nous avons choisi de garder les valeurs inchangées pour le calcul des mesures de performances MAPE et RMSE.

4.2.2.4 Analyse de sensibilité

Comme dans les modèles ARIMA et SARIMA, nous avons analysé la sensibilité du modèle Prophet par rapport à la segmentation de l'échantillon d'apprentissage et le choix des paramètres initiaux du modèles. Pour rappel, la segmentation consiste à bâtir deux modèles, le premier est basé sur les données d'entrainement d'avril 2020 à octobre 2020 et le second sur l'échantillon d'entrainement couvrant la période de juin à octobre 2020. La période de test est celle des données du mois de novembre 2020.

4.2.2.4.1 Sensibilité à la segmentation de l'échantillon d'entrainement

Les résultats de nos modélisations sont affichés dans la **Figure 4.19** ci-dessous dans le cas des infections.



Performance sur donnée test - Nov20

Training	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Apr-Oct20	14.9	205.7	1207	1157
Jun-Oct20	9.9	156		1109

Figure 4.19. Impact de la taille de l'échantillon d'entraiment sur la performance du modèle Prophet - infections.

Les résultats montrent une amélioration significative (plus de 40%) des valeurs RMSE et MAPE en faveur de l'utilisation d'un échantillon basé sur les observations plus récentes (juin-octobre 2020). Dans la **Figure 4.20**, les résultats de notre analyse sont montrés pour la variable décès.

Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Avril20-Oct20	196	32	27.1	17.6
Juin20-Oct20	31.6	11		19

Figure 4.20. Impact de la taille de l'échantillon d'entrainement sur la performance du modèle Prophet - décès.

Une fois de plus, on note une amélioration significative de plus de 200% des valeurs de MAPE et RMSE. Ceci laisserait suggérer que Prophet performe mieux si les données sont plus récentes. Cependant d'autres analyses devraient être faites pour confirmer cette observation.

4.2.2.4.2 Sensibilité aux valeurs initiales des paramètres du modèle

Nous avons cherché les valeurs optimales des paramètres σ, τ sur nos données. Le processus consiste à faire varier ces paramètres suivant les valeurs de σ entre 0.01 et 10; et τ entre 0.01 et 0.5 tel qu'illustré dans le **Tableau 4.6** ci-dessous ainsi que la valeur optimale trouvée.

Tableau 4.6 Paramètres optimaux des modèles de Prophet

Prophet - Infections				
Composantes	**#**	**Valeur Par défaut**	**Intervalle**	**Valeur Optimale**
Tendance	facteur à priori point de rupture (σ)	0.05	[0.001, 0.5]	0.01
Saisonnalité	facteur à priori saisonalité (τ)	10	[0.01, 10]	0.01

Prophet - Décès				
Composantes	**#**	**Valeur Par défaut**	**Intervalle**	**Valeur Optimale**
Tendance	facteur à priori point de rupture (σ)	0.05	[0.001, 0.5]	0.5
Saisonnalité	facteur à priori saisonalité (τ)	10	[0.01, 10]	0.01

Notons ici que dans le cas des décès, la valeur optimale du terme de tendance est celle de la borne supérieure de la grille que nous avons choisie alors que pour le terme de saisonnalité, la valeur optimale est celle de la borne inférieure de la grille.

- Pour le facteur apriori de la saisonnalité, τ , la valeur optimale obtenue est 0.01 et correspond à la borne inférieure de la grille pour les variables infections et décès. Cette valeur optimale est proche de zéro, ce qui veut dire que le terme de saisonnalité (Eq. 2.5) est négligeable. Il n'est donc pas nécessaire d'augmenter la borne inférieure de la grille tel qu'illustré à la **Figure 4.23**.
- Pour le facteur apriori du point de rupture, σ, la variable optimale obtenue pour la variable décès est celle de la borne supérieure de la grille. Pour cela, nous avons analysé dans la

Figure 4.21, la sensibilité de la performance du modèle par rapport à la valeur optimale obtenue à la suite de l'augmentation de la taille de la grille.

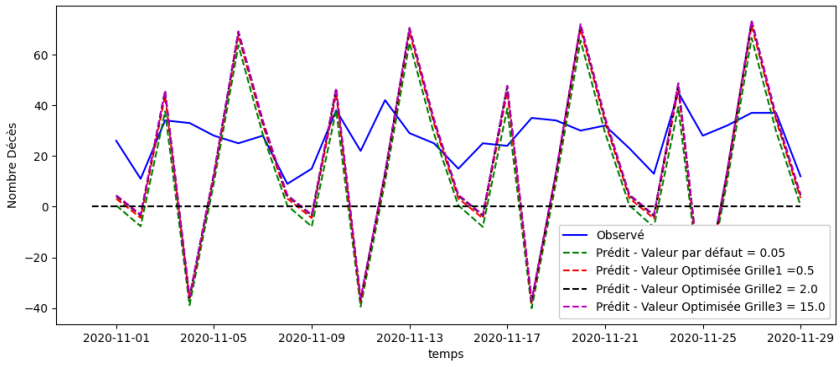
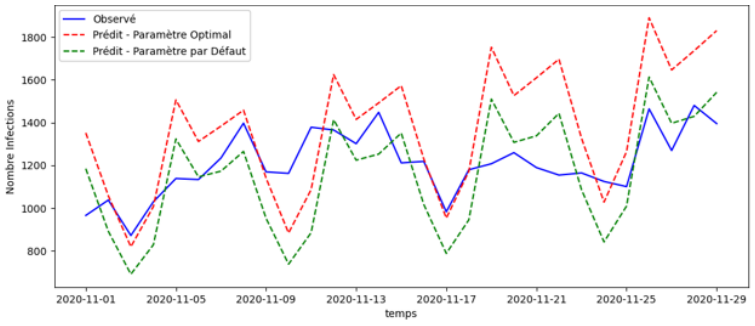


Figure 4.21. Sensibilité de la performance du modèle Prophet – Décès par rapport au paramètre de la fonction de distribution d'apriori

Nous résultats de la **Figure 4.21** montrent que l'impact pour la variable décès, faire varier le facteur à priori du point de rupture σ n'a pas d'effet sur la performance du modèle. Ci-dessous dans la **Figure 4.22** les résultats pour la variable Infection sont reportés.



Performance sur donnée test - Nov20

Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
Valeur par défaut	14.9	207.66	1207.65	1157.71
Valeur optimisée	17.61	275.82		1371.16

Figure 4.22. Paramètres par défaut vs. paramètres optimisés sur la performance du modèle Prophet - infections.

Les résultats montrent que la performance du modèle avec les paramètres optimaux est moins bonne que celle des paramètres par défaut. Ces derniers sont donc utilisés par la suite. On note également que la courbe issue des paramètres optimisés semble être le résultat d'un déplacement vers le haut des valeurs prédites issue du modèle utilisant les paramètres par défaut. Pour mieux analyser ce qui pourrait être à l'origine de ce déplacement en parallèle, nous analysons individuellement l'impact des paramètres (optimisés vs par défaut) de tendance et de saisonnalité sur la performance de ces modèles. L'impact du paramètre de saisonnalité est montré dans la figure **Figure 4.23** où on note que ce paramètre (valeur par défaut vs valeur optimisée) semble avoir peu d'impact sur la performance de ces modèles.

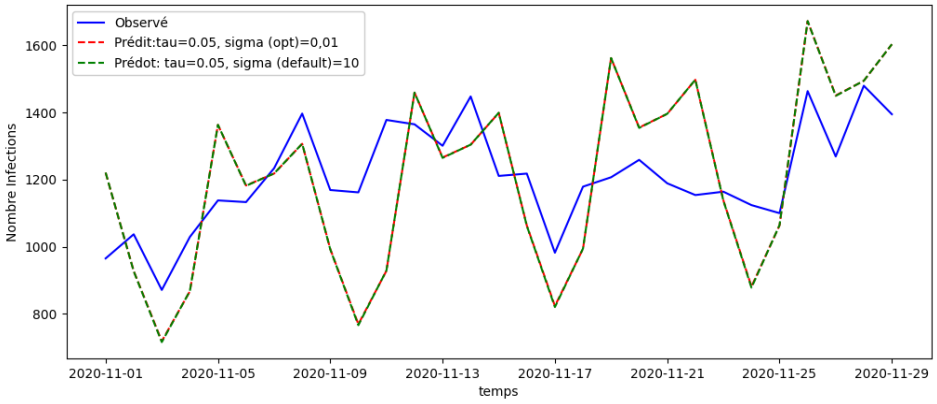


Figure 4.23. Paramètre par défaut vs. paramètres optimisés du facteur de saisonnalité sur la performance du modèle Prophet - infections.

L'analyse de l'influence du paramètre de tendance est montrée dans la **Figure 4.24** où on note que les valeurs prédites avec les paramètres par défaut sont décalées vers le haut comparé à celles utilisant les valeurs optimisées. Ceci laisse suggérer que l'utilisation de valeurs élevées du facteur de tendance (τ) de la fonction de Laplace tend à augmenter la valeur du taux de croissance dans l'équation de la pente dans chaque fenêtre définie par Prophet.

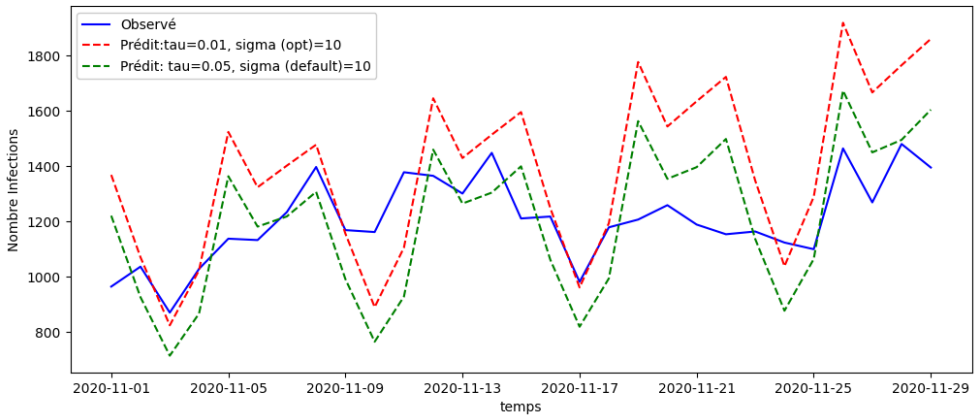


Figure 4.24. Paramètre par défaut vs. paramètres optimisés du facteur de tendance sur la performance du modèle Prophet - infections.

4.3 Modèles par compartiments

La modélisation par compartiments dans ce travail consiste à implémenter le modèle de l'Eq.2.9. Cependant, pour obtenir une prévision précise (décès ou incidence maximale, taux de reproduction, etc.), les paramètres de l'équation 2.9 doivent être déterminés par les observations les plus récentes de l'échantillon d'entrainement. Une façon de le faire est de subdiviser l'échantillon d'entraînement en phases délimitées par des points de changement ou des points de rupture analogue au modèle *Prophet* ci-dessus. L'idée [65] est de faire une analyse de tendance basée sur l'utilisation de la solution du modèle SIR (Eq.2.8) pour définir les points de rupture entre les phases. Pour se débarrasser de la variable liée à l'infection dans l'équation 2.8, l'hypothèse utilisée est qu'une épidémie est censée s'arrêter non seulement lorsque tout le monde a été infecté, mais aussi lorsque plus personne n'est nouvellement infectée. L'équation utilisée est donc

$$S(R) = N\, e^{-a\,R}$$

Ici N est la population totale; $a = \frac{\beta}{N\gamma}$; S(R)veut dire que S est une fonction de R. Ceci entraine

$$log\ S(R) = -a\ R + log\ N \qquad (4.1)$$

Dans l'équation ci-dessus appelée aussi modèle SR, la pente de la ligne peut changer lorsque les valeurs des paramètres β, γ sont modifiées. Pour identifier les points de rupture, il faut trouver des points pour lesquels la relation (coefficients de la régression) de changement de log S en fonction de R cesse de s'ajuster bien aux données de la phase ou la fenêtre correspondante. L'un des défis consiste à identifier leur nombre et leur emplacement optimaux. Différents algorithmes de détection de ces points de changement sont disponibles dans la littérature. Les plus connus sont ceux de PELT, Binseg et Bottom-up.

L'hypothèse de base de l'algorithme PELT est que le nombre de points de ruptures augmente linéairement avec l'augmentation de la taille des données. Ainsi, pour identifier les multiples points de changement ou rupture, l'algorithme est d'abord appliqué à l'ensemble des données de manière itérative et indépendante à chaque partition jusqu'à ce qu'aucun autre point de changement ne soit détecté. Pour des détails sur l'algorithme PELT, voir par exemple la référence [66].

Concernant l'algorithme Binseg, elle a pour base une approche séquentielle. D'abord, un point de changement est détecté dans la donnée initiale, puis la série est scindée autour de ce point de changement en deux parties. Ensuite l'opération de recherche de points de changement est répétée dans chacune des sous parties et ainsi de suite. Ainsi, Binseg est très apprécié pour sa flexibilité d'étendre toute méthode de détection de point de ruptures à la détection de plusieurs points.

Pour une analyse théorique et algorithmique de Binseg, voir par exemple les références [67] et [68]. Enfin, contrairement à l'algorithme de Binseg, l'algorithme BottomUp est une approche de détection dont l'essence est de considérer de nombreux points de changement initialement et de supprimer successivement les moins significatifs. Ainsi, la série chronologique est divisée en plusieurs sous périodes le long d'une grille régulière. Par la suite, les fenêtres ou périodes contiguës sont successivement fusionnées en fonction d'une mesure de leur similarité. Pour une analyse théorique et algorithmique de BottonUp, voir par exemple les références [69] et [70].

Les algorithmes PELT, Binseg et BottomUp sont disponibles dans les librairies Ruptures sous Python ou R. Pour la construction de ces algorithmes, on a l'option de choisir des fonctions de

bases normales (‘*Normal*’) ou radiales (‘*RBF*’). Pour l’utilisation des modèles par compartiments, nous avons utilisé *CovSirPhy* [71], [45], une bibliothèque Python pour la COVID-19. Les sorties de la modélisation relative à l’indentification des points de ruptures et la tendance entre ces points de rupture sont données dans la **Figure 4.25** ci-dessous.

	Debut	Fin
0	01Avril2020	16Avril2020
1er	17Avril2020	01Mai2020
2e	02Mai2020	19Mai2020
3e	20Mai2020	03Juin2020
4e	04Juin2020	15Juillet2020
5e	16Juillet2020	12Août2020
6e	13Août 2020	14Septembre2020
7e	15Septembre2020	29Septembre2020
8e	30Septembre2020	14Octobre2020
9e	15Octobre2020	30Octobre2020

Algorithme = 'PELT-RBF'

	Debut	Fin
0	01Avril2020	21Avril2020
1er	22Avril2020	13Mai2020
2e	14Mai2020	29Mai2020
3e	30Mai2020	19Juillet2020
4e	20Juillet2020	06Septembre2020
5e	07Septembre2020	25Septembre2020
6e	26Septembre2020	12Octobre2020
7e	13Octobre2020	30Octobre2020

Algorithme = 'Binseg-RBF'

	Debut	Fin
0	01Avril2020	25Avril2020
1er	26Avril2020	24Mai2020
2e	25Mai2020	19Juin2020
3e	20Juin2020	16Juillet2020
4e	17Juillet2020	11Août2020
5e	12Août2020	06Septembre2020
6e	07Septembre2020	02Octobre2020
7e	03Octobre2020	30Octobre2020

Algorithme = 'BottomUp-RBF'

	Debut	Fin
0	01Avril2020	25Avril2020
1er	26Avril2020	24Mai2020
2e	25Mai2020	16Juillet2020
3e	17Juillet2020	06Septembre2020
4e	07Septembre2020	02Octobre2020
5e	12Août2020	06Septembre2020

Algorithme = 'BottomUp-Normal'

Figure 4.25. Sélection des points de ruptures en fonction des algorithmes choisis

On trouve qu’avec l’algorithme ‘PELT – RBF, on identifie 10 points de rupture[36] alors que pour les algorithmes ‘Binseg-RBF’, ‘BottomUp-RBF’ et ‘BottomUp-Normal’ on identifie respectivement 8, 8 et 6 points de rupture. On note aussi que bien que les nombres et les localisations des points de ruptures diffèrent selon l’algorithme choisi, ceux de la dernière phase ou de la dernière fenêtre sont presque identiques entre les périodes allant du 03 au 30 Octobre pour les algorithmes “*BottomUp-rbf*”, “*BottomUp-normal*” et du 15 au 30 Octobre pour “*Pelt-rbf*”, “*Binseg-rbf*”. Étant donné que seule la dernière phase ou fenêtre est importante pour la prédiction, notre analyse se limitera aux algorithmes “*Pelt-rbf”* utilisé comme référence et l’Algorithme “*BottomUp-rbf*” utilisé comme ‘challenger’ dans l’analyse de sensibilité.

Les modèles relatifs aux Eqs. 2.11 et 2.12 associés aux approches SIRD et SIRF sont implémentés dans *CovSirPhy*. Cette librairie dispose des fonctions pour rechercher les paramètres optimaux du modèle pour chacune des phases ci-dessous. Les sorties des modèles SIRD et SIRF sont analysées ci-dessous pour les variables infections et décès.

[36] Nous rappelons que seuls les paramètres de la dernière phase ou fenêtre ou les paramètres de la dernière fonction de tendance de l’échantillon d’entrainement sont utilisés pour la prédiction, incluant les potentiels points de rupture futurs.

4.3.1 Modèles par compartiments – paramètres optimaux

Nous avons utilisé la librairie *CovSirPhy* pour estimer les paramètres optimaux des modèles pour chacune des fenêtres mentionnées ci-dessus. Les valeurs obtenues comme paramètres des deux modèles SIRD et SIRF considérés sont reportés dans le **Tableau 4.7**. L'algorithme 'PERL -rbf' a été utilisé.

Tableau 4.7 Paramètres optimaux des modèles SIRD et SIRF

Fenêtre		Population	Modèle	Rt	Paramètres optimaux					
Debut	**Fin**				**theta (Θ)**	**Kappa (κ)**	**Rho (ρ)**	**sigma (σ)**	**RMSE**	**MAPE**
15-Oct-20	30-Oct-20	8537674	SIRD	1.08		0.00194	0.110614	0.1009	1670.57	0.133
15-Oct-20	30-Oct-20	8537674	SIRF	1.07	0.000343	0.00116	0.118385	0.1097	1794.64	0.136

La fenêtre indique la dernière période utilisée pour estimer les paramètres des modèles SIRD et SIRF. Tel que reporté dans la **Figure 4.25**, ces fenêtres sont les mêmes. La population initiale qui est une donnée d'entrée du modèle est celle de la province du Québec d'environ 9 millions d'habitant. Nos résultats montrent aussi que les paramètres optimaux trouvés pour les deux modèles (Eqs 2.11 et 2.12) sont assez faibles et relativement semblables. Cette faible valeur obtenue serait liée à la différence entre les proportions des personnes décédées ou cas actifs[37] par rapport à celles guéries comme illustré ci-dessous sur la **Figure 4.26**. Notons enfin que les sauts observés sur la **Figure 4.26** sont dus à la mise à jour régulière du nombre de cas actifs ou guéris.

[37] Rappel : Cas actifs = Cumulative confirmées – Cumulative Guéris – Cumulative Décès

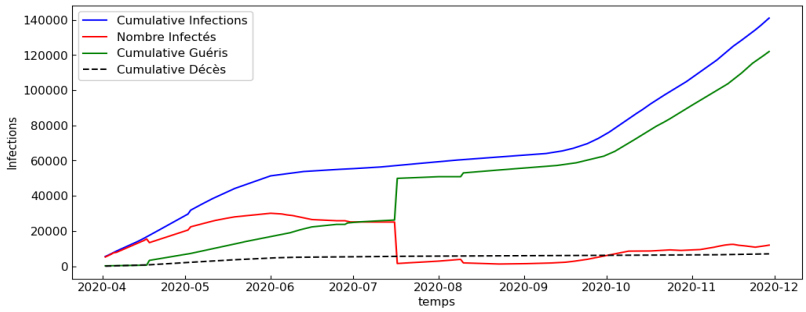
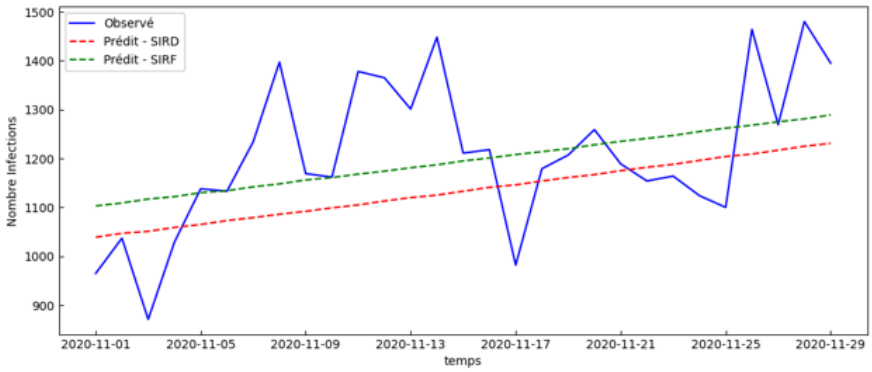


Figure 4.26. Dynamique historique des principales variables du modèle compartimental.

4.3.2 Modèles par compartiments – performance

Pour rappel, l'échantillon d'entrainement couvre la période de de mars 2020 à octobre 2020 et les observations du mois de novembre 2020 ont été utilisées pour le test du modèle. Nous présentons dans la **Figure 4.27** la performance du modèle sur l'échantillon de test.



	Performance sur donnée test - Nov20				
Modèle	Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
SIRD	Avril20-Oct20	8.8	153.2	1207	1134
SIRF		9.7	135.2		1194

Figure 4.27. Performance des modèles SIRD et SIRF sur l'échantillon de test - infections.

Les résultats de nos modélisations montrent qu'en moyenne, les valeurs prédites par nos modèles SIRD et SIRF sont assez proches des valeurs observées. Toutefois, au niveau granulaire, nos deux modèles s'avèrent moins aptes à répliquer la fluctuation quotidienne des cas d'infections. Ceci se traduit ainsi par des valeurs assez élevées de MAPE et RMSE.

On note que le modèle prédit pour le mois de novembre un taux de croissance constant sur l'échantillon de test. Ceci s'explique par le fait que pour cette période, notre modèle est purement linéaire à cause d'une faible volatilité au niveau des données observées (voir **Figure 4.28** et **Figure 4.29**). De même, on note que les droites que forment les cas d'infections prédites par nos deux modèles sont parallèles. Ceci s'explique par la faible différence observée entre les valeurs des paramètres optimaux des deux modèles SIRD et SIRD (voir **Tableau 4.7**)

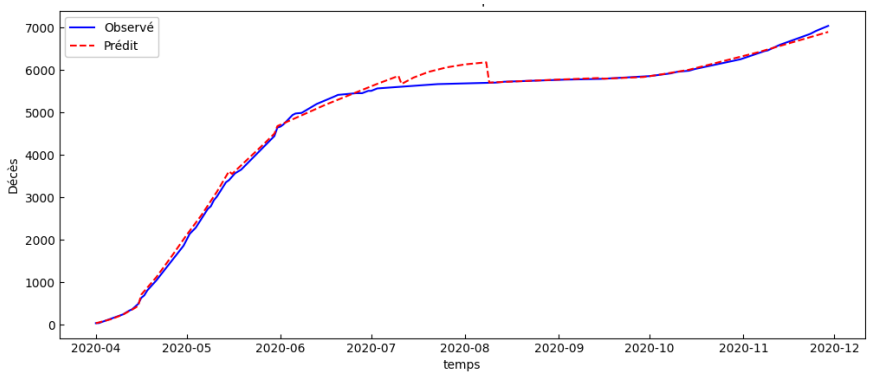


Figure 4.28. Performance historique du modèle SIRD pour la variable décès.

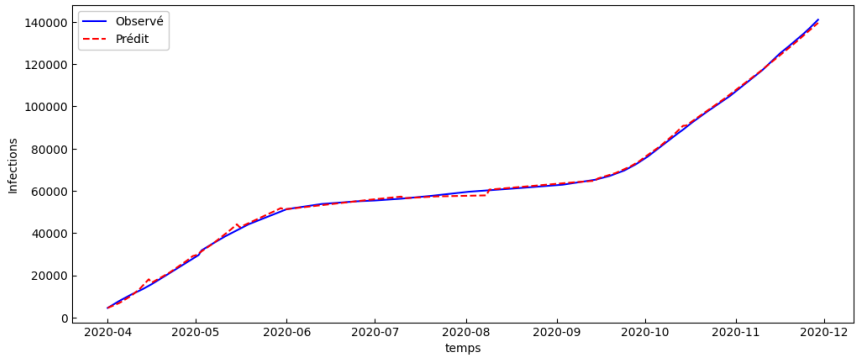
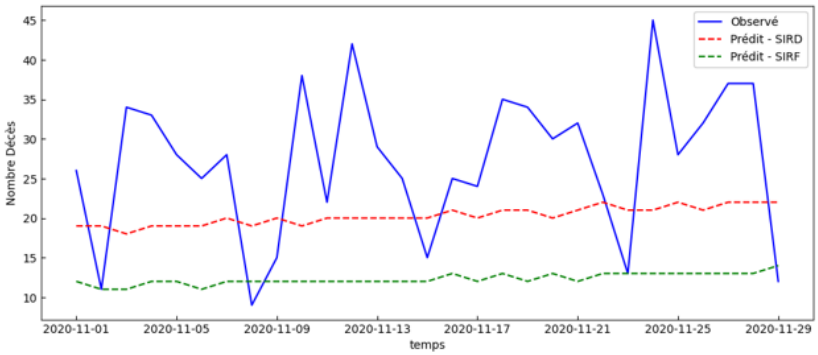


Figure 4.29. Performance historique du modèle SIRD pour la variable infection.

Dans la **Figure 4.30**, nous reportons les résultats de nos modélisations pour la variable décès. Nous notons que nos résultats sont similaires à ceux observés dans la **Figure 4.27** dans le cas des infections. Les valeurs prédites par nos deux modèles sont des droites qui en général sous-estiment les nombres de décès observés. Cette mauvaise performance de nos modèles semble être plus

importante pour l'approche SIRF comparé à l'approche SIRD si on se réfère aux valeurs des RMSE et MAPE.



	Performance sur donnée test - Nov20				
Modèle	Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
SIRD	Avril20-Oct20	39	11.47	27.1	20.3
SIRF		49.8	17.46		12.3

Figure 4.30. Performance des modèles SIRD et SIRF sur l'échantillon de test – décès

4.3.3 Modèles par compartiments – analyse de la sensibilité.

Nous avons analysé la sensibilité du modèle en fonction des algorithmes PELT-rbf et BottomUp-rbf utilisés pour déterminer les fenêtres utilisées dans l'analyse de la tendance. Les résultats sont montrés dans la **Figure 4.9**

Tableau 4.8 Sensibilité de l'algorithme sur la performance du modèle SIRF- infections

	Performance sur donnée test - Nov20: Infections				
Modèle	Echantillon	MAPE	RMSE	Moyenne (Observé)	Moyenne (Prédit)
PELT-RBF	Avril20-Oct20	8.8	135.2	1207	1194
BottomUp-RBF		8.7	135		1192

Les résultats montrent que le choix d'algorithme a peu d'impact sur la performance du modèle.

4.4 Résumé et comparaison de la performance de tous les modèles.

Ci-dessous dans le **Tableau 4.9** nous comparons la performance de tous nos modèles sur l'échantillon test que nous avons choisi (novembre 2020). Notons que nous avons obtenu différents modèles en prenant en compte : (a) la taille de l'échantillon d'entrainement d'avril à octobre 2020 ou de juin à octobre 2020; (b) d'autres considérations comme l'impact de l'ordre de la périodicité, les hyperparamètres par défaut vs ceux optimisés, l'algorithme utilisé pour détecter une tendance dans une série chronologique.

Tableau 4.9 Comparaison des modèles sur l'échantillon de test (novembre 2020)

	#	Training	Intrant1	MAPE	RMSE
ARIMA	**Infections**	Apr20 -Oct20		**9.8**	140.6
		Jun - Oct20		8.9	138.4
	Décès	Apr20 -Oct20		37	**11**
		Jun - Oct20		46.3	16
SARIMA	**Infections**	Jun - Oct20	m=7	9.1	**120**
		Apr20 -Oct20	m=5	13.9	191
			m=7	13.9	191
			m=10	14	200
			m=30	14	214
	Décès	Jun - Oct20	m=7	29.2	11.7
		Apr20 -Oct20	m=5	43	17
			m=7	29	**11**
			m=10	35	12
			m=20	39	11
Prophet	**Infections**	Jun - Oct20	Parametre par défaut	**9.9**	156
		Apr20 -Oct20		14.9	207
			Parametre Optim	17.6	276
	Décès	Jun - Oct20	Parametre par défaut	101	32
		Apr20 -Oct20		96	32
			Parametre Optim	97	33
Modèle par compartiments	**Infections**	Apr20 -Oct20	SIRD - PELT_RBF	**8.8**	153
			SIRF-PELT_RBF	9.7	135
			SIRF-BottomUp_RBF	8.7	135
	Décès		SIRD- PEL_RBF	39	**11**
			SIRF-PELT_RBF	50	17.5
			SIRF-BottomUp_RBF	49	18

Les résultats affichés dans le **Tableau 4.9** laissent suggérer que pour la variable infection, les modèles par compartiment offrent en général les meilleures performances au regard des valeurs de MAPE et RMSE obtenus. Dans le cas de la variable décès, le modèle SAMIRA s'avère avoir la meilleure performance au regard des valeurs de MAPE et RMSE. Notons enfin que le modèle SARIMA semble relativement sensible à l'ordre de la périodicité choisie.

CHAPITRE 5 BACKTESTING ET COHERENCE ENTRE MODELES

Dans les Chapitre 2, Chapitre 3 et Chapitre 4 précédents, nous avons développé, implémenté nos modèles et les avons testé sur une période d'un mois. Dans ce chapitre, nous étendons le même schéma sur un large ensemble de données historiques couvrant les périodes de décembre 2020 à juin 2021. De plus, nous mesurerons la cohérence entre nos modèles, c'est-à-dire leur aptitude à prédire collectivement l'augmentation ou la diminution des infections ou décès au cours du mois à venir. Ce chapitre commence par l'analyse de la performance de nos modèles pendant une période de huit mois et se termine par l'analyse de la cohérence ou similarité entre modèles.

Nous avons mesuré la cohérence à travers la notion de signal parce qu'à cause de nombreux soubresauts entourant la pandémie (nouveaux variants, multiplication des tests, nouveaux types de test rapide, vaccinations, mesures de control de la pandémie tel que le confinement, les masques, etc.), il est parfois essentiel et suffisant de savoir si les infections ou décès augmenteront ou pas les mois futurs afin de prendre les décisions à priori contre la pandémie.

5.1 Backtesting des modèles

La **Figure 5.1** ci-dessous montre la division que nous avons fait de notre base de données. Ainsi, la taille de notre échantillon d'apprentissage a augmenté chaque fois d'un mois tout au long du processus d'ajustement du modèle, celle de l'échantillon de test est restée fixe et égale au mois qui succède la date de fin de l'échantillon d'apprentissage.

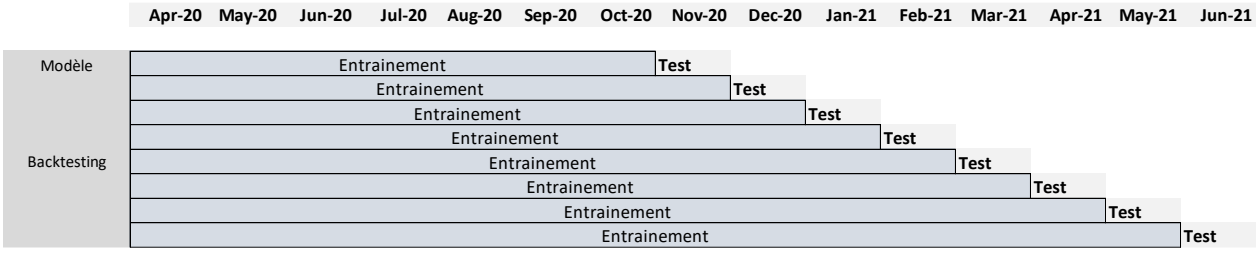


Figure 5.1. Division de la base de données en échantillons d'entrainement et de test.

Nous avons ajusté nos modèles afin qu'ils reflètent les nouvelles informations incluses dans les données. Ainsi, dans le cas des modèles SARIMA ou ARIMA, nous avons fait la transformation logarithmique et maintenu l'ordre du terme d'intégration $d = 1$, et vérifié que nos données restaient stationnaires. De même, nous avons maintenu l'ordre de la périodicité $m = 7$, et avons permis aux autres paramètres du modèle p, q, P, Q de fluctuer entre 0 et 20. Pour les modèles Prophet, nous avons conservé les valeurs par défaut indiquées dans le **Tableau 4.6**. Enfin, pour les modèles par compartiments, nous avons conservé le nombre minimal d'observations égal à 15 pour construire chaque fenêtre utilisée de l'échantillon d'entrainement et seul l'algorithme "Pelt-rbf" a été utilisé pour déterminer les fenêtres utilisées en analyse de la tendance.

Pour notre analyse, nous avons ajouté les indicateurs définis ci-dessous :

- Moyenne des sept dernières observations de l'échantillon d'entrainement noté '*Moyenne Entrainement*'. Ce dernier est simplement le nombre d'infections ou décès moyen observé sur la dernière semaine du dernier mois de l'échantillon d'apprentissage. Ce choix a été fait pour avoir une valeur lisse des dernières observations afin de limiter l'impact d'une possible fluctuation journalière.
- Le nombre de décès ou infections moyen observés dans l'échantillon de test est noté '*Observée*' alors que ceux prédits sont notés '*Predite*'.
- L'erreur relative notée '*Relative Error*' est la variation relative entre le nombre de décès ou infections prédits '*Prédite*' par le modèle et ceux observés *'Observée*' dans l'échantillon de test.
- Les indicateurs (signaux) de diminution ou d'augmentation moyen de décès ou infections se calculent comme le rapport entre le nombre moyen de décès ou infections sur la moyenne des sept dernières observations de l'échantillon d'entrainement, soit alors $signal = \frac{Prédite\ ou\ Observée}{\text{Moyenne Entrainement}}$. Selon le signe de ce ratio, le signal sera noté $'+'$ pour une valeur positive et $'-'$ dans le cas contraire.
- La cohérence, ou similarité entre signaux est une variable binaire qui vaut '1' pour des signaux de même signe et '0' dans le cas contraire.

Ci-dessous nous présentons et analysons nos résultats.

5.1.1 Cas des infections

Ci-dessous dans le **Tableau 5.1**, nous présentons le résultat de performance (Backtesting) pour le modèle SARIMA et pour le cas des infections.

Tableau 5.1 Backtesting du modèle SARIMA - infections

Sarima Infections									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	960.9	1,207.7	1,216.5	130.3	8.9	1%	+	+	1
Dec-20	1,309.3	1,920.1	1,648.0	668.0	15.9	-14%	+	+	1
Jan-21	2,328.4	1,978.0	3,095.0	1,694.0	84.0	56%	+	-	0
Feb-21	1,312.0	904.0	823.5	129.0	12.1	-9%	-	-	1
Mar-21	802.0	744.6	762.8	96.0	10.4	2%	-	-	1
Apr-21	908.4	1,298.0	1,067.5	334.3	20.1	-18%	+	+	1
May-21	1,012.1	673.4	1,024.1	418.2	70.1	52%	+	-	0
Jun-21	381.0	152.0	225.0	95.0	59.7	48%	-	-	1
	+ augmentation du nombre d'infections								
	- baisse du nombre d'infections								

Pour comprendre le **Tableau 5.1**, prenons la deuxième ligne par exemple, le mois de test est décembre 2020. Comme illustré dans la **Figure 5.1**, l'échantillon d'apprentissage couvre la période allant d'avril 2020 à novembre 2020. Les derniers sept jours du mois de novembre 2020 ont été utilisés pour calculer la valeur observée de '*Moyenne Entrainement*', soit la période du 24 au 30 novembre. La valeur trouvée est en moyenne 1309 infections. Le nombre moyen d'infections observé pendant le mois de test (décembre) est de 1920 et celui prédit par le modèle est de 1648 infections. Le RMSE et MAPE sont les mesures de performance du modèle dans l'échantillon de test tel qu'expliqué dans les chapitres précédents. L'erreur relative est calculée comme $\frac{1648}{1920.1} - 1 = -14\%$.

- Le signal prédit $\frac{Predite}{Moyenne\ Entrainement} = \frac{1648}{1309.28} - 1 = 25\% > 0,$ on utilise le signe +, pour dire qu'on prédit une augmentation au cours du mois prochain par rapport à la semaine dernière.

- On obtient $\frac{Observée}{Moyenne\ Entrainement} = \frac{1920}{1309.28} - 1 = 46\% > 0$, on utilise le signe +, pour le signal observé. Enfin pour la cohérence, elle vaut '1' car les deux signaux sont de même signe tel qu'affiché dans le **Tableau 5.1**.

On trouve qu'en valeur absolue, l'erreur relative de prédiction varie de 1 à 50% avec une assez forte déviation observée en décembre 2020 et janvier 2021 (Voir Appendice 2). D'après des informations rapportées par le journal[38] « La Presse » du 27 décembre 2020, une correction des données a été faite le 27 décembre pour tenir compte des infections non comptabilisées pour la période du 24 au 27 décembre. Bien que cela puisse influencer considérablement les paramètres de nos modèles, nous avons choisi de garder ces valeurs inchangées car ces données sont bien observées malgré l'incertitude sur les dates réelles associées. De même, nos modèles performent moins bien en avril 2021, sans doute dû à la forte recrudescente des infections dans les zones de la région de la Chaudière-Appalaches, des régions de la Capitale-Nationale, et de l'Outaouais (voir section 3.1.2).

On note aussi que le modèle SARIMA performe moins bien (surestime) en mai et juin (voir : Appendice 2 : Graphique du Backesting des Modèles), sans doute à cause de l'impact positif de l'inoculation du vaccin à environ 25% de la population Québécoise. Ci-dessous dans le **Tableau 5.2**, l'analyse pour le modèle ARIMA est présenté.

Tableau 5.2 Backtesting du modèle ARIMA - infections

ARIMA Infections									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	960.85	1207.65	1234.7	139.9	9.8	2%	+	+	1
Dec-20	1309.28	1920.1	1611.9	703.4	16.4	-16%	+	+	1
Jan-21	2328.42	1978	3390	1716.0	94.0	71%	+	-	0
Feb-21	1312	904	1855.0	1028.0	113.0	105%	+	-	0
Mar-21	802	744.6	1254.0	549.1	69.6	68%	+	-	0
Apr-21	908.42	1298	896.0	468.0	29.0	-31%	-	+	0
May-21	1012.14	673.35	1084.6	465.0	80.0	61%	+	-	0
Jun-21	381	152	821.0	734.0	598.0	440%	+	-	0
	+ augmentation du nombre d'infections								
	- baisse du nombre d'infections								

[38] Source: https://www.lapresse.ca/covid-19/2020-12-27/covid-19-au-quebec/6783-nouveaux-cas-et-110-deces-depuis-le-24-decembre.php

En général, les valeurs RMSE, MAPE et l'erreur relative sont élevées. On note aussi qu'en terme de cohérence, le modèle ARIMA prédit correctement pour 2 mois sur 8 mois possibles la tendance des infections observée entre la période de novembre 2020 à juin 2021. Ci-dessous au **Tableau 5.3** l'analyse pour le modèle Prophet est présentée.

Tableau 5.3 Backtesting du modèle Prophet - infections

Prophet - Infections									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	960.85	1207.65	1157.0	206.0	15.0	-4%	+	+	1
Dec-20	1309.28	1920.1	1451.0	804.2	17.1	-24%	+	+	1
Jan-21	2328.42	1978	2908	1278.0	66.2	47%	+	-	0
Feb-21	1312	904	1489.0	743.0	74.0	65%	+	-	0
Mar-21	802	744.6	807.0	451.0	47.0	8%	+	-	0
Apr-21	908.42	1298	838.0	795.0	48.0	-35%	-	+	0
May-21	1012.14	673.35	1356.0	955.6	152.0	101%	+	-	0
Jun-21	381	152	511.0	775.0	550.0	236%	+	-	0
	+ augmentation du nombre d'infections								
	- baisse du nombre d'infections								

En général, les valeurs RMSE, MAPE et l'erreur relative sont élevées. On note aussi qu'en terme de cohérence, le modèle Prophet prédit correctement pour 2 mois sur 8 mois possibles la tendance des infections observée entre la période de novembre 2020 à juin 2021. Ci-dessous au **Tableau 5.4** l'analyse pour le modèle SIRF est présenté.

Tableau 5.4 Backtesting du modèle par compartiment SIRF - Infections

Compartmental SIRF - Infections									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	960.85	1207.65	1195.0	135.2	8.7	-1%	+	+	1
Dec-20	1309.28	1920.1	1079.0	1093.0	21.0	-44%	-	+	0
Jan-21	2328.42	1978	2937	1303.0	68.6	48%	+	-	0
Feb-21	1312	904	862.0	147.0	14.0	-5%	-	-	1
Mar-21	802	744.6	563.3	261.8	25.3	-24%	-	-	1
Apr-21	908.42	1298	623.0	720.0	51.0	-52%	-	+	0
May-21	1012.14	673.35	1009.0	381.0	67.0	50%	-	-	1
Jun-21	381	152	156.0	267.0	49.0	3%	-	-	1
	+ augmentation du nombre d'infections								
	- baisse du nombre d'infections								

On note aussi qu'en terme de cohérence, le modèle SIRF prédit correctement pour 5 mois sur 8 possibles la tendance des infections observée entre la période de novembre 2020 à juin 2021. Ci-dessous au **Tableau 5.5** l'analyse pour le modèle SIRD est présenté.

Tableau 5.5 Backtesting du modèle par compartiment SIRD- infections

Compartmental SIRD - Infections									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	960.9	1,207.7	1,133.9	153.2	9.7	-6%	+	+	1
Dec-20	1,309.3	1,920.1	1,016.1	1,539.9	14.0	-47%	-	+	0
Jan-21	2,328.4	1,978.0	3,048.0	1,392.0	73.6	54%	+	-	0
Feb-21	1,312.0	904.0	789.2	177.0	17.0	-13%	-	-	1
Mar-21	802.0	744.6	493.9	305.0	31.0	-34%	-	-	1
Apr-21	908.4	1,298.0	694.5	639.0	44.0	-46%	-	+	0
May-21	1,012.1	673.4	1,104.5	467.0	83.0	64%	+	-	0
Jun-21	381.0	152.0	370.6	219.5	22.0	144%	-	-	1
	+ augmentation du nombre d'infections								
	- baisse du nombre d'infections								

Notre modèle prédit correctement pour 4 mois sur 8 possibles la tendance des infections observée entre la période de novembre 2020 à juin 2021

En conclusion, le modèle SARIMA réussit à prédire correctement 6 fois sur 8 la tendance (augmentation ou diminution) des décès du prochain mois. Cette entente tombe en moyenne à 5/8 et 2/8 respectivement pour les modèles par compartiments et de Prophet. Toutefois, à cause du faible nombre d'observations (huit observations pour huit mois), on ne saurait à l'état actuel tirer une conclusion définitive. Notons qu'en terme de performance (RMSE), les résultats des tables (**Tableau 5.1**, **Tableau 5.2**, **Tableau 5.3**, **Tableau 5.4**, **Tableau 5.5**) montrent qu'en moyenne la qualité des performances de nos modèles suit l'ordre SARIMA> Prophet > ARIMA> Compartiment (SIRF) > Compartiment (SIRD).

5.1.2 Cas des décès

Les résultats de nos modélisations sont présentés dans les tables (**Tableau 5.6**, **Tableau 5.7**, **Tableau 5.8**, **Tableau 5.9**) ci-dessous pour respectivement les modèles SARIMA, ARIMA, Prophet, SIRF et SIRD). Le résultat du modèle SARIMA est ci-dessous la table **Tableau 5.6**

Tableau 5.6 Backtesting du modèle SARIMA – décès

SARIMA Décès									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	17.86	27.13	18.9	11.3	29.2	-30%	+	+	1
Dec-20	30.57	39.86	30.52	27.797	36.0	-23%	-	+	0
Jan-21	53.14	49.44	107.6	85.5	154.7	118%	+	-	0
Feb-21	46.57	21.88	35.9	28.1	131.8	64%	-	-	1
Mar-21	12.28	9	-69.0	86.0	1271.2	-867%	-	-	1
Apr-21	5.71	8.48	7.0	7.1	89.6	-17%	+	+	1
May-21	9.42	6.857	15.4	11.1	190.5	124%	+	-	0
Jun-21	6.14	2.724	5.2	3.6	28.0	90%	-	-	1
	+ augmentation du nombre décès								
	- baisse du nombre de décès								

On note en général que la performance de SARIMA n'est pas bonne car valeurs du RMSE et MAPE sont assez élevées particulièrement pour la période de décembre 2020 à mars 2021. On peut aussi noter que le modèle prédit pour le mois de mars un nombre de décès négatif. Bien que cela ne soit pas intuitif, ceci s'explique par le fait que le modèle prédit une tendance fortement à la baisse pour le mois de mars 2021. En prenant la différence d'ordre 1 pour avoir les valeurs journalières on peut avoir des valeurs négatives (voir section 4.1.2.2). Nous notons aussi que l'erreur relative reste au-delà de 17%. Ceci traduit également la faiblesse de notre modèle pour la prédiction du nombre moyen de décès mensuels. Toutefois, notons qu'en terme de cohérence, notre modèle prédit correctement pour 5 mois sur 8 mois possibles, la tendance des décès observée entre la période de novembre 2020 à juin 2021. Ci-dessous dans la **Tableau 5.7**, nous analysons la performance du modèle ARIMA.

Tableau 5.7 Backtesting du modèle ARIMA – décès

ARIMA Décès									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	17.86	27.1	21.1	11.0	36.9	-22%	+	+	1
Dec-20	30.57	39.9	17.9	30.9	48.1	-55%	-	+	0
Jan-21	53.14	49.4	62.6	37.2	82.6	27%	+	-	0
Feb-21	46.57	21.9	50.5	31.1	198.0	131%	+	-	0
Mar-21	12.28	9.0	12.1	12.1	136.7	34%	-	-	1
Apr-21	5.71	8.5	4.4	5.9	58.6	-49%	-	+	0
May-21	9.42	6.9	9.1	4.0	78.1	33%	-	-	1
Jun-21	6.14	2.8	2.4	3.2	67.0	-13%	-	-	1
	+ augmentation du nombre décès								
	- baisse du nombre de décès								

En général, les valeurs RMSE, MAPE et l'erreur relative sont élevés de décembre 2020 à février 2021. On note aussi qu'en terme de cohérence, notre modèle prédit correctement pour 4 mois sur 8 mois possibles la tendance des décès observée entre la période de novembre 2020 à juin 2021. Ci-dessous au **Tableau 5.8**, l'analyse pour le modèle Prophet est présentée.

Tableau 5.8 Backtesting du modèle Prophet – décès

Prophet Décès									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	17.86	27.13	17.5	32.0	96.0	-35%	-	+	0
Dec-20	30.57	39.857	31.7	40.7	68.0	-21%	+	+	1
Jan-21	53.14	49.44	55.59	34.6	65.5	12%	+	-	0
Feb-21	46.57	21.88	66.3	57.1	304.8	203%	+	-	0
Mar-21	12.28	9	12.0	34.6	418.6	34%	-	-	1
Apr-21	5.71	8.48	8.2	36.1	448.7	-3%	+	+	1
May-21	9.42	6.857	10.2	30.7	423.8	49%	+	-	0
Jun-21	6.14	2.724	6.6	24.2	467.0	142%	+	-	0
	+ augmentation du nombre décès								
	- baisse du nombre de décès								

En général, les valeurs RMSE, MAPE et l'erreur relative sont élevés pour tous les mois considérés. On note aussi qu'en terme de cohérence, notre modèle prédit correctement pour 3 mois sur 8 mois possibles la tendance des décès observée entre la période de novembre 2020 à juin 2021. Ci-dessous dans le **Tableau 5.9** est présentée l'analyse pour le modèle SIRF.

Tableau 5.9 Backtesting du modèle SIRF – décès

Compartmental SIRF - Décès									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	Signal Prédit	Signal Observé	Cohérence
Nov-20	17.857	27.13	20.3	17.5	49.8	-25%	+	+	1
Dec-20	30.57	39.857	26.43	21.2	33.5	-34%	-	+	0
Jan-21	53.14	49.44	54.13	17.0	38.5	9%	+	-	0
Feb-21	46.57	21.88	30.068	10.4	62.1	37%	-	-	1
Mar-21	12.28	9	11.56	3.8	64.4	28%	-	-	1
Apr-21	5.71	8.48	12.46	5,34	88.7	47%	+	+	1
May-21	9.42	6.857	7.16	3.2	69.1	4%	-	-	1
Jun-21	6.14	2.724	2.57	1.4	12.0	-6%	-	-	1
	+ augmentation du nombre décès								
	- baisse du nombre de décès								

On note que les valeurs RMSE, MAPE et l'erreur relative faibles pour tous les mois considérés. On note aussi qu'en terme de cohérence, notre modèle prédit correctement pour 6 mois sur 8 mois possibles la tendance des décès observée entre la période de novembre 2020 à juin 2021. Ci-dessous dans le **Tableau 5.8**, nous présentons nos résultats pour le modèle SIRD.

Tableau 5.10 Backtesting du modèle SIRD – décès

Compartmental SIRD - Décès									
test	Moyenne Entrainement	Observée	prédite	RMSE	MAPE	Erreur Relative	ignal Prédi	gnal Obser	Cohérence
Nov-20	17.86	27.13	20.3	11.5	39.0	-25%	+	+	1
Dec-20	30.57	39.86	27.43	20.77	24.0	-31%	-	+	0
Jan-21	53.14	49.44	56.266	18.2	41.0	14%	+	-	0
Feb-21	46.57	21.88	31.2	11.2	66.6	42%	-	-	1
Mar-21	12.28	9	12.9	5.2	10.0	43%	+	-	0
Apr-21	5.71	8.48	6.9	3.9	40.6	-19%	+	+	1
May-21	9.42	6.857	7.7	3.2	74.2	12%	-	-	1
Jun-21	6.14	2.724	3.9	2.6	14.0	45%	-	-	1
	+ augmentation du nombre décès								
	- baisse du nombre de décès								

On note que les valeurs RMSE, MAPE et l'erreur relative faibles pour tous les mois considérés. On note aussi qu'en terme de cohérence, le modèle SIRD prédit correctement pour 5 mois sur 8 mois possibles la tendance des décès observée entre la période de novembre 2020 à juin 2021.

En général, on note que compte tenu de la baisse significative des décès au fil des mois après décembre 2020, la plupart des modèles (à l'exception des modèles SIRD et SIRF) prédisent des nombres négatifs de décès. Bien que ces résultats ne fassent aucun sens, nous avons gardé ces valeurs inchangées dans nos calculs et comme impact, nos modèles ont eu des performances médiocres se traduisant par les valeurs élevées MAPE et RMSE. Les graphiques de comparaison entre les cas de décès observés et ceux prédits pour chacun des modèles étudiés sont fournis en Appendice 2 : Graphique du Backesting des Modèles de ce travail. On trouve aussi qu'en général, les modèles par compartiments prédisent assez bien la tendance des décès au prochain mois. Les taux de bonne prédiction sont d'environ 6/8 et 5/8 pour les modèles SIRF (voir Appendice 2) et SIRD, suivi par le modèle SARIMA et ARIMA dont les chiffres sont de 5/8 et 4/8. La plus faible entente est observée avec le modèle Prophet.

5.2 Cohérence entre modèles

Nous comparons entre modèles leur capacité à pouvoir prédire conjointement la tendance des décès ou infections au cours du prochain mois. Notre mesure de cohérence est la loi de la majorité[39] qui est très utilisée en apprentissage machine. Le principe est simplement de prendre comme signal final des modèles la tendance prédite par la majorité des modèles pour un mois considéré. Ceci est illustré dans le **Tableau 5.11** dans le cas des infections.

Tableau 5.11 Cohérence entre modèles – infections

Models Coherence - Infections										
test	Moyenne Entrainement	Observed	SARIMA - Signal	ARIMA - Signal	Prophet - Signal	SIRF - Signal	SIRD-Signal	Average Signal	Observed Signal	Agreement
Nov-20	961	1,208	+	+	+	+	+	+	+	1
Dec-20	1,309	1,920	+	+	+	-	-	+	+	1
Jan-21	2,328	1,978	+	+	+	+	+	+	-	0
Feb-21	1,312	904	-	+	+	-	-	-	-	1
Mar-21	802	745	-	+	+	-	-	-	-	1
Apr-21	908	1,298	+	-	-	-	-	-	+	0
May-21	1,012	673	+	+	+	-	+	+	-	0
Jun-21	381	152	-	+	+	-	-	-	-	1

Dans la première ligne du tableau ci-dessus, on note par exemple que tous les modèles prédisent que les cas augmenteront en décembre 2020. Ceci correspond au '*average signal'* qui est positif. Pour le mois d'Avril 2021, 4/5 signaux sont négatifs, ce qui donne un '*average signal*' négatif. En comparant ce signal résultant avec ceux observés (avant dernière colonne du tableau), on note que l'entente entre les signaux prédits et observés s'établit à environ 5/8 pour les infections et 6/8 pour les décès (voir **Tableau 5.12**). Ces chiffres sont en général meilleurs que ceux de la plupart des modèles pris individuellement (voir section précédente). Ceci laisse suggérer que prendre collectivement le signal de plusieurs modèles pourrait améliorer les prises de décision à priori face à la pandémie actuelle. Toutefois, ces résultats doivent pris avec du recul compte tenu du faible nombre d'observations, de modèles utilisés et des changements de règles sanitaires

[39] Pour qu'un modèle de vote soit sans biais, il devrait respecter certaines hypothèses comme la diversité minimale (absence de forte corrélation entre modèles), la nécessité pour chacun des modèles de faire mieux qu'un modèle aléatoire et enfin la pluralité des modèles. Bien que cette dernière ne soit pas respectée dans notre travail, il pourrait être pris comme base pour des travaux futurs.

Tableau 5.12 Cohérence entre modèles – décès

Cohérence des modèles - Décès										
test	Moyenne Entrainement	Observed	ARIMA - Signal	Prophet - Signal	SIRF - Signal	SIRD - Signal	SARIMA - Signal	Signal Majoritaire	Signal Observé	Cohérence
Nov-20	17.86	27.13	+	-	+	+	+	+	+	1
Dec-20	30.57	39.86	-	+	-	-	-	-	+	0
Jan-21	53.14	49.44	+	+	+	+	+	+	-	0
Feb-21	46.57	21.88	+	+	-	-	-	-	-	1
Mar-21	12.28	9	-	-	-	+	-	-	-	1
Apr-21	5.71	8.48	-	+	+	+	+	+	+	1
May-21	9.42	6.857	-	+	-	-	+	-	-	1
Jun-21	6.14	2.724	-	+	-	-	-	-	-	1
			+ augmentation du nombre décès							
			- baisse du nombre de décès							6/8

CHAPITRE 6 CONCLUSION

Ce travail a porté sur la comparaison de la performance des modèles statistiques (ARIMA, SARIMA et Prophet) et des modèles par compartiments (SIRD, SIRF) pour la prédiction des cas d'infections et de décès de la COVID au Québec. Nos résultats ont montré une performance en général supérieure pour le modèle SARIMA dans le cas des infections et du modèle SIRD dans le cas des décès. Toutefois, nous avons trouvé que la performance des modèles SARIMA semblait un peu sensible aux valeurs initiales utilisées pour définir l'ordre de la périodicité du modèle alors que la performance du modèle Prophet s'améliorait significativement quand on tronquait la taille de l'échantillon d'entrainement en faveur des observations plus récentes. Le fait que Prophet et SARIMA (tous deux incluent une composante saisonnière) donnent des résultats si différents laisserait suggérer que l'impact de la saisonnalité pourrait être moindre sur la performance des modèles. Pour confirmer cette hypothèse et comme piste d'amélioration pour ce travail, il serait judicieux de faire nos analyses en y incluant des données récentes (au-delà de juillet 2021). Enfin l'intervalle de confiance de nos prédictions n'a pas été analysée dans ce travail.

Concernant la cohérence, nous avons trouvé que collectivement nos modèles pouvaient prédire correctement les tendances (augmentation et diminution) pendant 5 mois sur une période de 8 pour les infections et pendant 6 mois sur une période de 8 pour les décès. Ces résultats laissent suggérer un possible bénéfice à agréger la prédiction de tendance de plusieurs modèles. Toutefois, face au faible nombre d'échantillons et de modèles utilisés, ces résultats méritaient d'être pris avec du recul. Comme piste de solution, ce travail pourrait être amélioré en y incluant d'autres types de modèles comme les réseaux de neurone récurrents, les supports vecteurs machines, les modèles de croissance de population et surtout les vecteurs autorégressifs qui permettraient d'analyser en une étape les variables d'états (infections et décès) de ce travail. Il est aussi important de noter que face à la difficulté en temps de pandémie de dénombrer adéquatement les personnes infectées, il serait judicieux d'inclure une variable relative aux hospitalisations.

Au moment (janvier 2022) où nous concluons ce travail, l'actualité de la pandémie est particulièrement dominée par la circulation active du variant 'omicron'. La capacité d'accueil de certains centre hospitaliers (Laval, Laurentides, Estrie) est presque arrivée à saturation. Les modèles que nous venons de bâtir pourraient être utilisés comme outils d'aide à la décision.

BIBLIOGRAPHIE

[1] C. G. Meyer et T. Velatan, «The COVID-19 epidemic,» *Trop. Med. Int. Health*, vol. 25, p. 278, 2020.

[2] WHO, «Coronavirus Disease 2019: Situation Report,» 2020.

[3] S. Camporesi et M. Mori, «Ethicists, doctors and triage decisions: who should decide? And on what basis?,» *J Med Ethics.*, p. 106499, 2020.

[4] L. Rosenbaum, «Facing Covid-19 in Italy - Ethics, Logistics, and Therapeutics on the Epidemic's Front Line,» *N Engl J Med*, p. 382, 2020.

[5] INSPQ, «Surmortalité et mortalité par COVID-19 au Québec en 2020,» 14 Juin 2021. [En ligne]. Available: https://www.inspq.qc.ca/sites/default/files/publications/3143-surmortalite-mortalite-covid-19-2020.pdf.

[6] Y. Berchenko, . Y. Manor, L. Freedman , E. Kaliner, I. Grotto, E. Mendelson et A. Huppert, «Estimation of polio infection prevalence from environmental surveillance data,» *Sci Transl Med*, vol. 9, p. 383, 2017.

[7] H. Boujakjian, «Modeling the spread of ebola with seir and optimal control, (2016),» *SIAM*, p. 026113, 2016.

[8] D. Attila, A. Gumel, B. Podolsky et N. Rosen, «Modeling the impact of quarantine during an outbreak of ebola virus disease,» *Infectious Disease Modelling*, p. 12, 2019.

[9] A. K. Sahai, N. Rath et P. M. Singh, «ARIMA modelling & forecasting of COVID-19 in top five affected countries,» *Diabetes Metab Syndr*, vol. 14, p. 1419, 2020.

[10] D. H. Lee, Y. S. Kim, Y. Y. Koh, K. Y. Song et I. H. Chang, «Forecasting COVID-19 Confirmed Cases Using Empirical Data,» *Healthcare*, vol. 9, p. 254, 2021.

[11] H. Alabdulrazzaq, . M. N. Alenezi, . Y. Rawajfih, B. Alghannam, . A. A. Al-Hassan et F. S. Al-Anzi, «On the accuracy of ARIMA based prediction of COVID-19 spread.,» *Results Phys* , vol. 27, p. 104509, 2021.

[12] K. E. Arunkumar, D. Kalaga, M. Kumar, G. Chilkoor, M. Kawaji et T. Brenza, «Forecasting the dynamics of cumulative COVID-19 cases (confirmed,recovered, and deaths) for top-16 countries using statistical machine learning models: ARIMA and SARIMA,» *Applied Soft Computing Journal*, vol. 103, p. 107161, 2021.

[13] Z. Malki, E.-S. Atlam, A. Ewis, G. Dagnew, A. R. Alzighaibi, G. Elmarhomy, M. Elhosseini, A. E. Hassanien et I. Gad, «ARIMA models for predicting the end of COVID-19 pandemic,» *Neural Computing and Applications*, vol. 33, p. 2929, 2021.

[14] B. Letham et S. J. Taylor, «Forecasting at Scale,» *PeerJ Preprints*, 2017.

[15] S. Patandung et I. Jatnika, «The FB Prophet Model Application to the Growth Prediction of International Tourists in,» *International Research Journal of Advanced Engineering and Science, Indonesia during the COVID-19 Pandemic*, vol. 6, p. 110, 2021.

[16] C. B. A. Satrio, W. Darmawan, B. Nadia et N. Hanafiah, «Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET,» *Procedia Computer Science*, vol. 179, p. 524, 2021.

[17] M. Khayyat, K. Laabidi, N. Almalki et M. Al-zahrani, «Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction,» *Computers, Materials & Continua*, vol. 67, p. 3782, 2021.

[18] M. Kretzschmar et J. Wallinga , Mathematical Models in Infectious Disease Epidemiology 2010, New York,: Springer, 2010.

[19] W. Kermak et A. McKendrik, «A contribution to the mathematical theory of epidemics,» *Proc Roy Soc Lond, 1927,* p. 700, 1927.

[20] F. Brauer et C. Castillo-Chavez , «Mathematical Models in Population Biology and Epidemiology,» *Springer-Verlag Inc,* p. 273, 2001.

[21] S. Patrak, A. Maiti et G. P. Samanata, «Rich Dynamics of an SIR epidemic model,Nonlinear Analysis,» *Modeling and Control ,* p. 71, 2010.

[22] A. Abou-Ismail, «Compartmental Models of the COVID-19 Pandemic for Physicians and Physician-Scientists,» *SN Compr Clin Med.,* p. 1, 2020.

[23] C. Anastassopoulou, L. Russo, L. Tsakris et A. Siettos, «Data-based analysis, modelling and forecasting of the Covid-19 outbreak,» *PLOS One,* vol. 15, p. 230505, 2020.

[24] I. Cooper, A. Mondal et C. Antonopoulos, «A SIR model assumption for the spred of COVID-19 in different communities,» *Chaos Solitons Fractals,* vol. 139, p. 110057, 2020.

[25] S. Bastos et D. Cajueiro, «Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil,» *Nature - Scientific Reports,* vol. 10, p. 19457, 2020.

[26] S. Feng, Z. Feng, C. Ling et C. Chang , «Prediction of the COVID-19 epidemic trends based on SEIR and AI models,» *PLOS ONE ,* vol. 16, p. e0245101, 2021.

[27] G. Box et G. M. Jenkins, Time Series Analysis: Forecasting and Control., San Francisco: Holden-Day, 1970.

[28] P. Newbold, «Some Recent Developments in Time Series Analysis,» *International Statistical Review,* p. 53, 1981.

[29] N. Gujarati et D. Porter, Basic Econometrics, New York: McGraw-Hill/Irwin, 2008.

[30] B. Choi, ARMA Model Identification, New York: Springer-Verlag, 1992.

[31] J. H. Cochrane , Time Series for Macroeconomics and Finance, Chicago: Spring, 1997.

[32] W. Greene, Econometric Analysis, 7th éd., Upper Saddle River: Prentice Hall,, 2012.

[33] J. Cryer et K. Chan, Time Series Analysis: with Application in R, New York: Springer, 2008.

[34] G. Box et G. Jenkins, Time Series Analysis, Forecasting and Control., New Jersey: John Wiley and Sons, 2008.

[35] J. H. Friedman et W. Stuetzle, «Projection Pursuit Regression,» *Journal of the American Statistical Association,* vol. 76, p. 813, 1981.

[36] A. Buja, T. Hastie et R. Tibshirani, «Linear Smoothers and Additive Models,» *The Annals of Statistics ,* vol. 17, p. 452, 1989.

[37] S. J. Taylor et B. Lethan, «Forecasting at Scale,» *American Statistician,* p. 37, 2018.

[38] M. Dawed, P. Koya et A. Goshu, «Mathematical Modelling of Population Growth: The Case of Logistic and Von Bertalanffy Models,» *Open Journal of Modelling and Simulation,* p. 113, 2014.

[39] J. Sunday, A. James, E. Ibijola, R. Ogunrinde et S. Ogunyebi, «A Computational Approach to Verhulst-Pearl Model,» *IOSR Journal of Mathematics,* p. 06, 2012.

[40] J. S. Cramer, «The Early Origin of the logit model,» *Biol Biomed Sci.,* p. 613, 2004.

[41] A. Ayiomamitis, «Logistic curve fitting and parameter estimation using nonlinear noniterative least-squares regression analysis,» *Computers and Biomedical Research*, p. 142, 1986.

[42] A. Harvey et S. Peters, «Estimation Procedures for Structural Time Series Models,» *Journal of Forecasting*, p. 89, 1990.

[43] D. Smith et L. Moore, «The SIR Model for Spread of Disease - The Differential Equation Model,» *Convergence*, 2004.

[44] K. Sasaki, «COVID-19 dynamics with SIR model.,» 2020. [En ligne]. Available: https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html.

[45] H. Takaya, «Kaggle Notebook, COVID-19 data with SIR model,» 2020. [En ligne]. Available: https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model.

[46] J. M. Heffernan, R. J. Smith et L. M. Wahl, «Perspectives on the basic reproductive ratio,» *Journal of the Royal Society Interface*, p. 281, 2005.

[47] P. V. D. Driessche, «Reproduction numbers of infectious disease models,» *Infect Dis Model.*, p. 288, 2017.

[48] R. Ross, «An application of the theory of probabilities to the study of a priori pathometry,» *Proceedings of the Royal Society of London*, p. 204, 1916.

[49] S. Gao, Z. Teng, J. Nieto et A. Torres, «Analysis of an SIR epidemic model with pulse vaccination and distributed time delay,» *Journal of Biomedicine & Biotechnology*, p. 64870, 2007.

[50] P. Wang et J. Jia, «Stationary distribution of a stochastic SIRD epidemic model of Ebola with double saturated incidence rates and vaccination,» *Adv. Differ. Equ.*, p. 433, 2019.

[51] O. Amenaghawon et D. Aboubakary , «Mathematical Modelling of the Transmission Dynamics of Ebola Virus,» *Applied and Computational Mathematics*, p. 313, 2015.

[52] S. Gounane, Y. Barkouch, A. Atlas et M. Bendahmane, «An adaptive social distancing SIR model for Covid-19 disease spreading and forecasting,» *Epidemiol. Methods*, p. 10, 2021.

[53] S. Alanazi, M. M. Kamruzzaman, . N. Alshammar, S. A. Alqahtani et A. Karime, «Measuring and Preventing COVID-19 Using the SIR Model and Machine Learning in Smart Health Care,» *Journal of Healthcare Engineering*, p. 12, 2020.

[54] T. Chai et R. Draxler, «Mean Square Error (RMSE) or Mean Absolute Error (MAE)? — Arguments against Avoiding RMSE in the Literature,» *Geoscientific Model Development*, vol. 7, p. 1247, 2014.

[55] N. Kumar et S. Susan, «COVID-19 Pandemic Prediction using Time Series Forecasting Models. arXiv:2009.12176v1 [physics.soc-ph],» *Physics.soc-ph*, p. 2009, 2020.

[56] M. Mohan, K. Swathi, H. Kumar et A. Sravani, «Forecasting the Spread of COVID-19 Pandemic with Prophet,» *Revue d'Intelligence Artificielle*, vol. 35, p. 115, 2021.

[57] T. Chafiq, M. Ouadoud et K. Elboukhari, «Covid-19 forecasting in Morocco using FBprophet Facebook's Framework in Python,» *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, p. 8654, 2020.

[58] P. S. Cowpertwait et A. Metcalfe, INTRODUCTORY TIME SERIES WITH R, New York: Springer, 2009.

[59] Y. Rahkmawati, M. Sumertajaya et M. Nur Aidi, «Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach,» *International Journal of Scientific and Research Publications,* vol. 9, p. 439, 2019.

[60] K. Aho, D. Derryberry et T. Peterson, «Model selection for ecologists: the worldviews of AIC and BIC,» *Ecology,* vol. 95, p. 631, 2014.

[61] Rdocumentation, «Acf: (Partial) Autocorrelation and Cross-Correlation Function Estimation,» 2022. [En ligne]. Available: https://www.rdocumentation.org/packages/forecast/versions/8.16/topics/Acf.

[62] R. Hyndman, «Discussion of High-dimensional autocovariance matrices and optimal linear prediction,» *Electronic Journal of Statistics,* vol. 9, p. 782, 2015.

[63] G. M. Ljung et G. E. Box, «On a Measure of a Lack of Fir in Time Series Models,» *Biometrika,* vol. 65, p. 297, 1978.

[64] E. Zunic, K. Korjenic et D. Donko, «Application of Facebook's Prophet Algorithm for Succesful Sales Forecasting Based on Real-World Data,» *International Journal of Computer Science & Information Technology,* vol. 12, p. 23, 2020.

[65] T. M. Balkew, The SIR Model When S(t) is a Multi-Exponential Function, Electronic Theses and Dissertations: Tennessee State University, 2010.

[66] R. Killick, P. Fearnhead et I. Eckley, «Optimal detection of change points with a linear computational costs,» *JASA,* vol. 107, p. 1590, 2012.

[67] J. Bai, «Estimating multiple breaks one at a time,» *Econometric Theory,* vol. 13, p. 315, 1997.

[68] P. Fryzlewicz, «Wild binary segmentation for multiple change-point detection.,» *The Annals of Statistics,* vol. 42, p. 2243, 2014.

[69] E. Keogh, S. Chu, D. Hart et M. Pazzani, «An online algorithm for segmenting time series. , 289–296.,» *Proceedings of the IEEE International Conference on Data Mining (ICDM),* p. 289, 2001.

[70] P. Fryzlewicz, «Unbalanced Haar technique for nonparametric function estimation,» *Journal of the American Statistical Association,* vol. 102, p. 1318, 2007.

[71] CovsirPhy, «CovsirPhy Development Team,» 2020-2021. [En ligne]. Available: https://github.com/lisphilar/covid19-sir.

[72] J. Taylor et B. Letham, «Automatic Forecasting Procedure,» 2019. [En ligne]. Available: https://facebook.github.io/prophet/.

[73] D. Kucharavy et R. De Guio, «Application of Logistic Growth Curve,» *Procedia Engineering,* p. 280, 2015.

[74] K. C. Tjorve et L. Underhill, «Growth sibling rivalry and their relationship tofledging success in African black oystercatchers Haematopus moquini,» *Zoology,* p. 27, 2009.

[75] T. Haryanti et B. Pamukti, «A Comparison of The Predictive Ability betweenLogistic and Gompertz Model on COVID-19 Outbreak,» *Int. Journal of Applied IT ,* p. 02, 2020.

[76] A. King, M. DeCelles, F. Magpantay et P. Rohani, « Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola.,» *Proc R Soc B. Biol. Sci.,* pp. 0-6, 2015.

[77] C. Truong, L. Oudre et N. Vayatis, «Selective review of offline change point detection methods,» *Signal Processing*, p. 167, 2020.

[78] B. Gompertz, «On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies,» *In a letter to Francis Baily, Esq. FRS &c. Philosophical transactions of the Royal Society of London*, vol. 115, p. 513, 1825.

[79] M. Zwietering, I. Jongenburger, F. Rombouts et K. Van't Riek, «Modeling of the bacterial growth curve,» *Appl Environ Microbiol.*, vol. 56, p. 1875, 1990.

[80] P. Gerlee, «The model muddle: in search of tumor growth laws,» *s. Cancer research.* , vol. 73, p. 2407, 2013.

[81] F. Khan, A. Saeed et S. Ali, «Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using Vector Autoregressive model in Pakistan,» *Chaos Solitons Fractals*, vol. 140, p. 110189, 2020.

[82] A. Marques, B. Gois, J. Xavier-Neto et S. J. Fong, Predictive Models For Decision Support In The Covid-19 Crisis, New York: Springer, 2021.

[83] J. Sun, «Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models,» *Comput Methods Programs Biomed Update*, vol. 1, p. 100029, 2021.

[84] S. Mahmud, «Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model,» *SSRN Electron. J.*, vol. 19, p. 2020, 2020.

[85] A. Sharma, D. Yadav, U. Chandra et H. Maheswari, «Outbreak Prediction of COVID-19 in India Using ARIMA and Prophet Model with Lockdown and Unlock,» *Advances in Science and Technology*, vol. 105, p. 318, 2021.

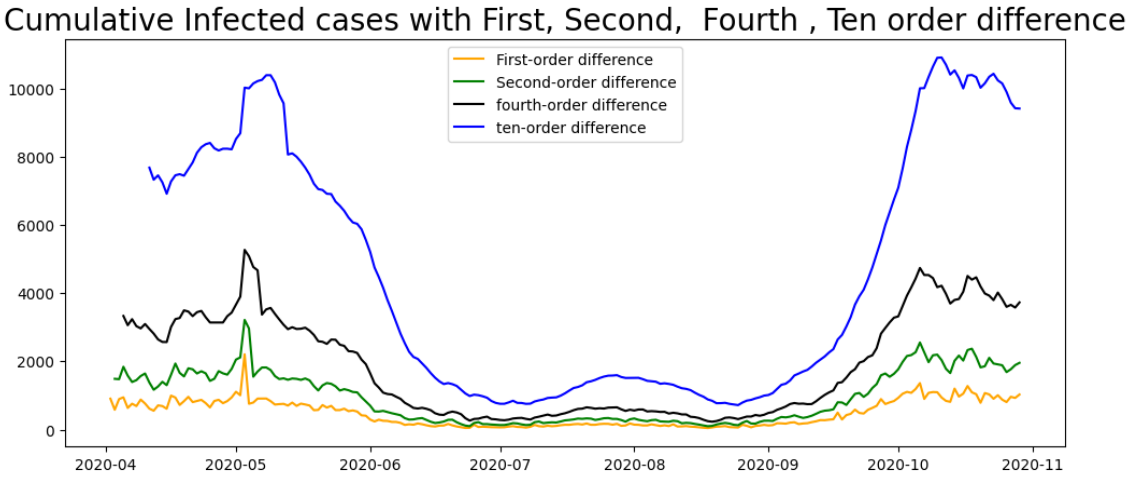
[86] V. S. D. &. M. M. Tulshyan, «On Eye on the Future of COVID-19: Prediction of Likely Positive Cases and Fatality in India over a 30-Day Horizon Using the Prophet Model.,» *Disaster medicine and public health preparedness*, p. 1, 2020.

APPENDICE

APPENDICE 1 : STATIONARITÉ

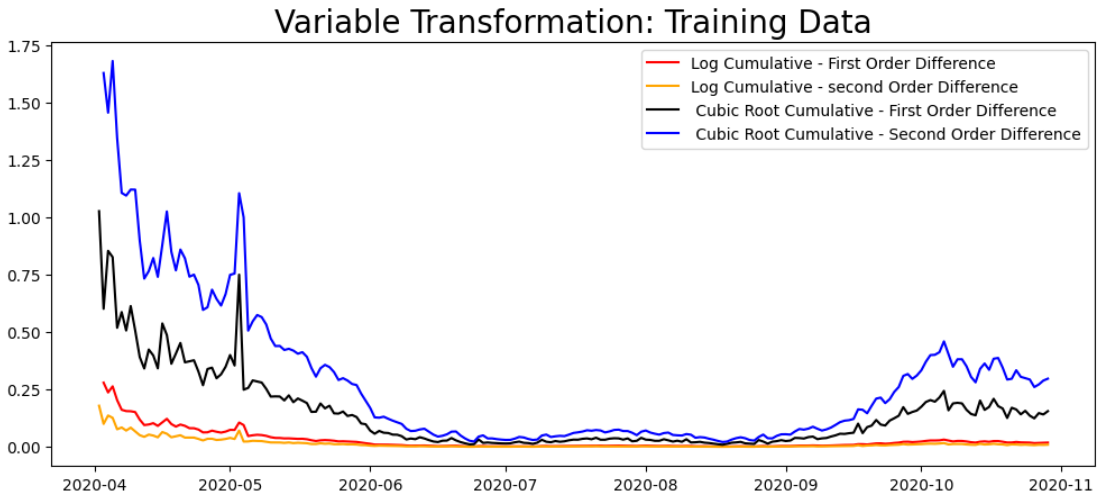
1) Échantillon d'entrainement.

Ci-dessous, une sérié de calcul de différence ont été appliquée à la population d'entrainement des infections pour la rendre stationnaire



Il s'est avéré qu'aucune de ces courbes n'était stationnaire (ADF test)

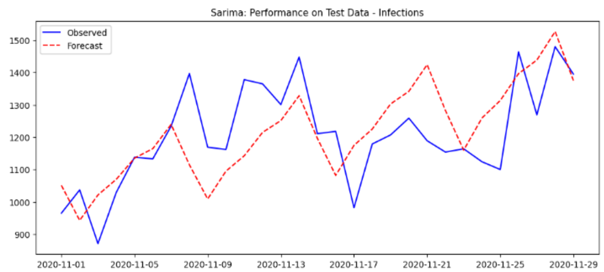
Ci-dessous, une combinaison de transformation (logarithmique, racine cubique) et de différence ont été appliquée à l'échantillon d'entrainement des infections pour la rendre stationnaire.

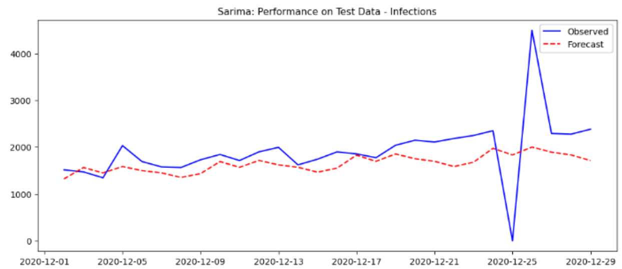


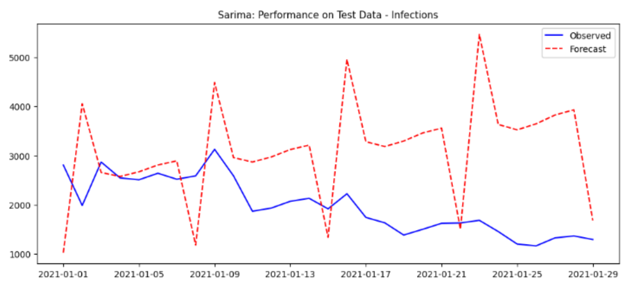
Il s'est avéré que la première différence de la fonction logarithme était stationnaire (ADF test)

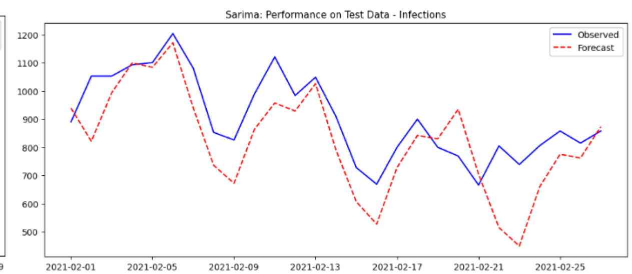
APPENDICE 2 : GRAPHIQUE DU BACKESTING DES MODÈLES

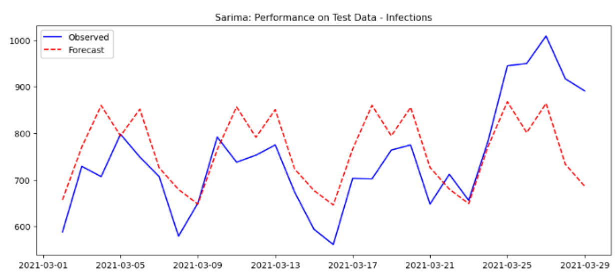
SARIMA - INFECTIONS

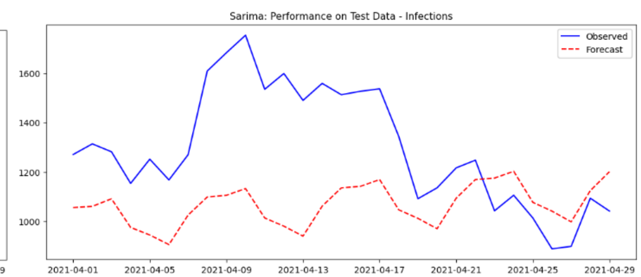

Sarima: Performance on Test Data - Infections
Observed
Forecast

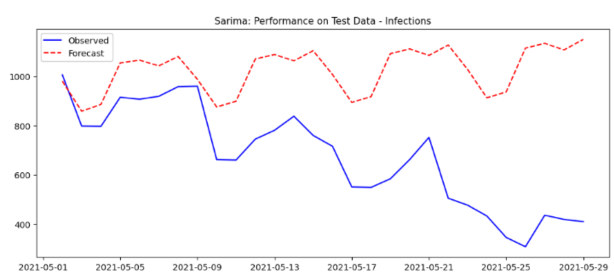

Sarima: Performance on Test Data - Infections
Observed
Forecast

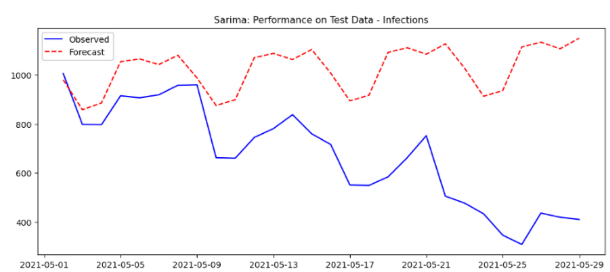

Sarima: Performance on Test Data - Infections
Observed
Forecast


Sarima: Performance on Test Data - Infections
Observed
Forecast

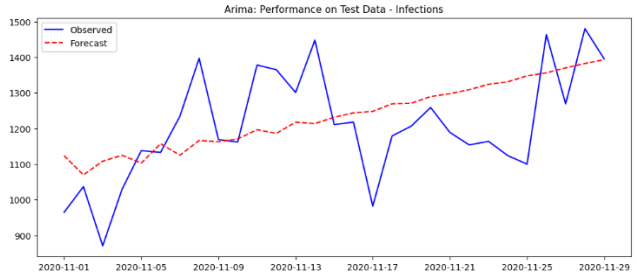

Sarima: Performance on Test Data - Infections
Observed
Forecast

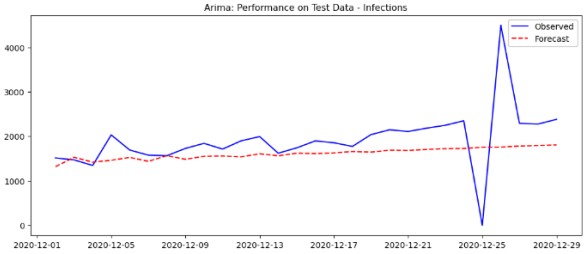

Sarima: Performance on Test Data - Infections
Observed
Forecast

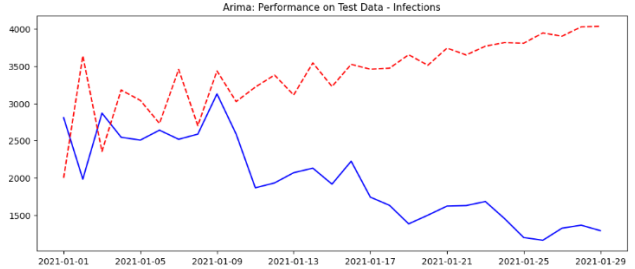

Sarima: Performance on Test Data - Infections
Observed
Forecast

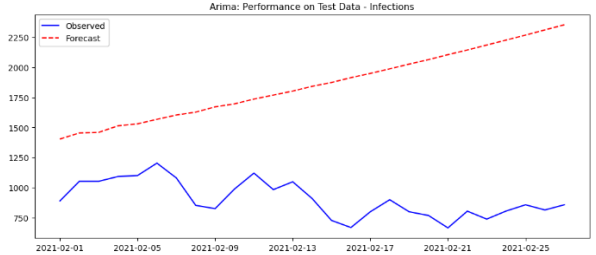

Sarima: Performance on Test Data - Infections
Observed
Forecast

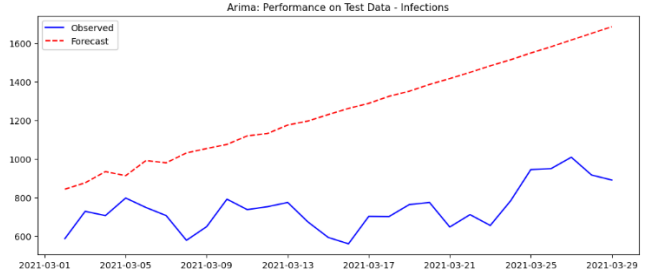
ARIMA - INFECTIONS

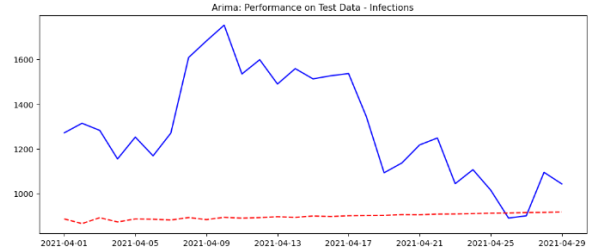


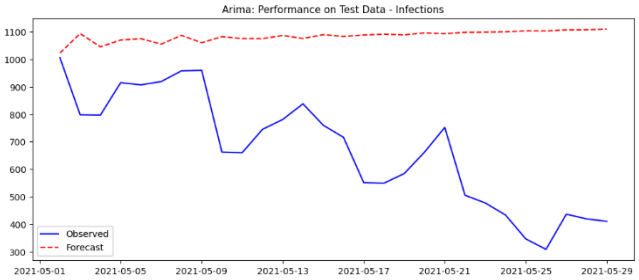


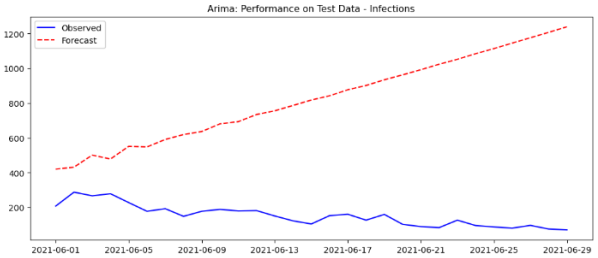




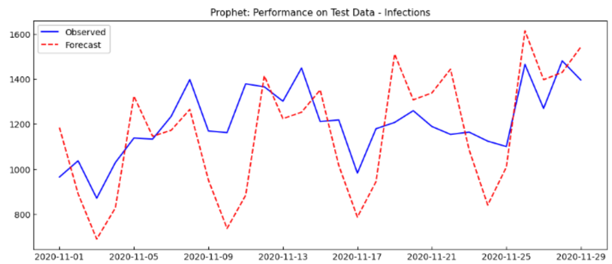


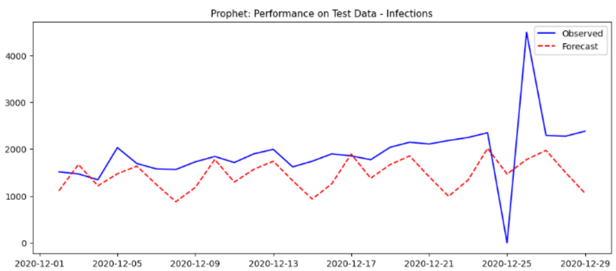


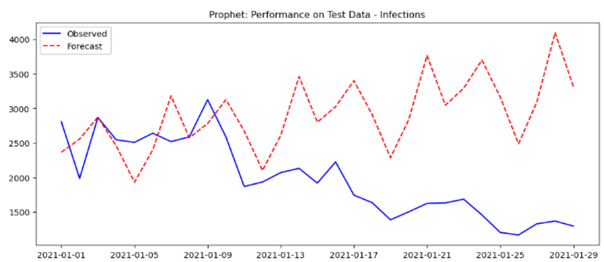

Arima: Performance on Test Data - Infections
Observed
Forecast
1100
1000
900
800
700
600
500
400
300
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17
2021-05-21
2021-05-25
2021-05-29

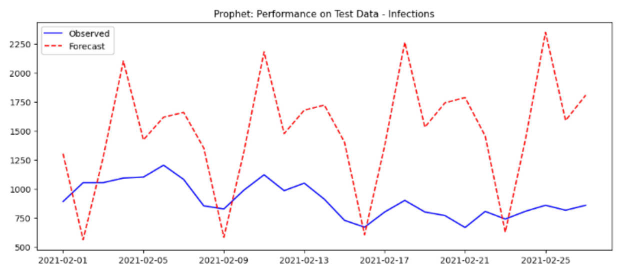

Arima: Performance on Test Data - Infections
Observed
Forecast
1200
1000
800
600
400
200
2021-06-01
2021-06-05
2021-06-09
2021-06-13
2021-06-17
2021-06-21
2021-06-25
2021-06-29

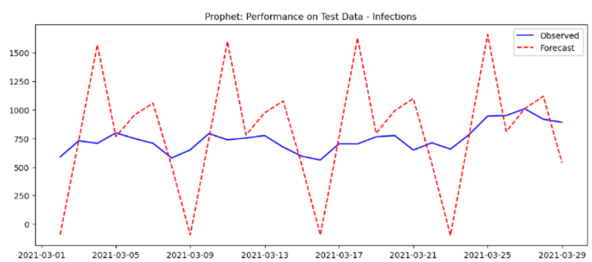
PROPHET INFECTIONS

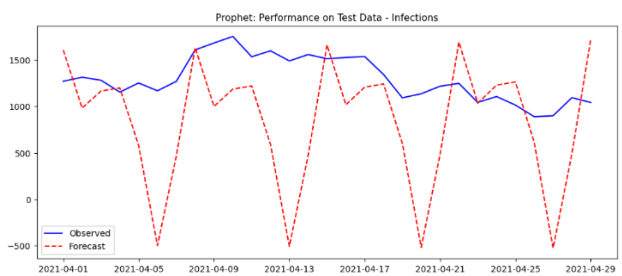


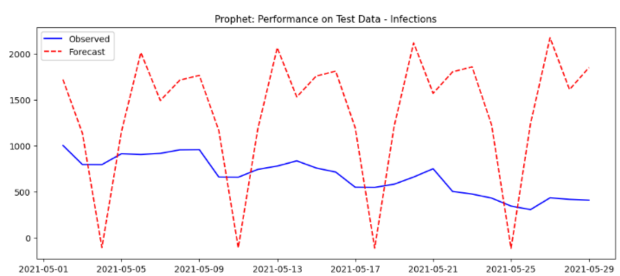


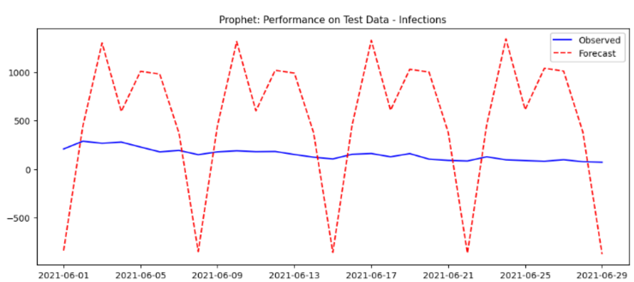




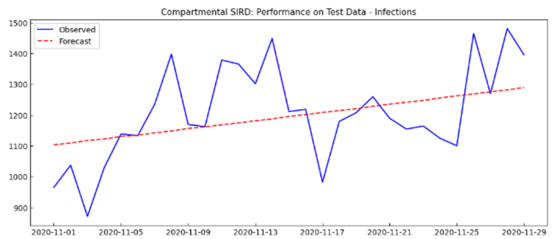


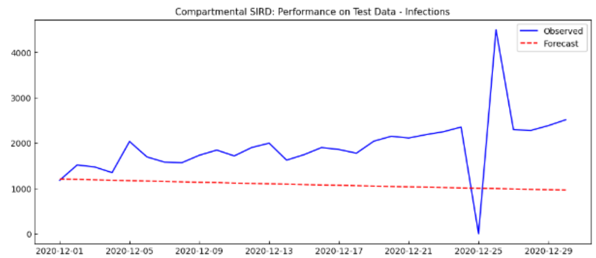


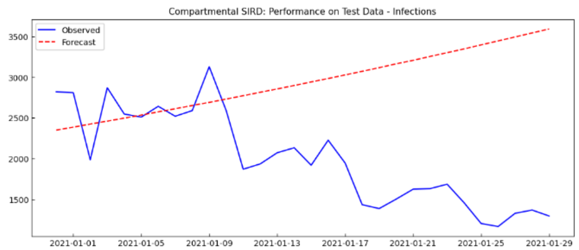


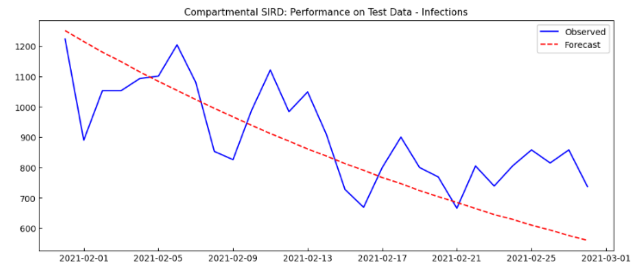


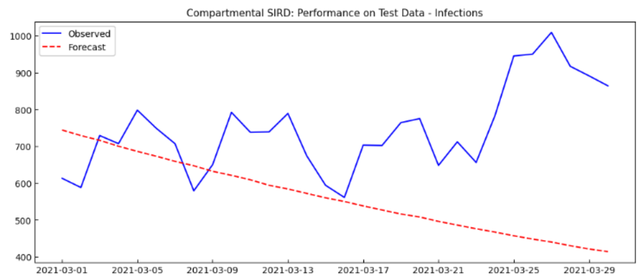
Compartimental Infections

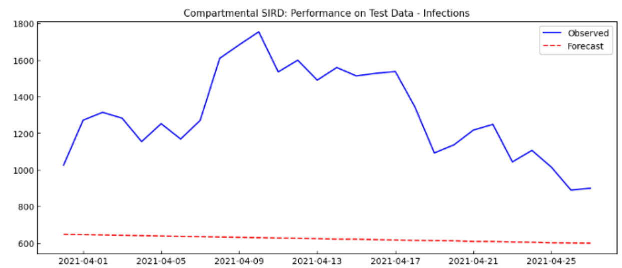


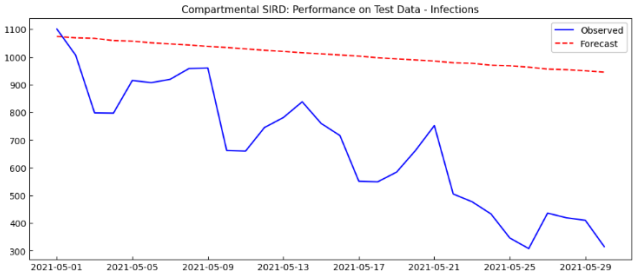


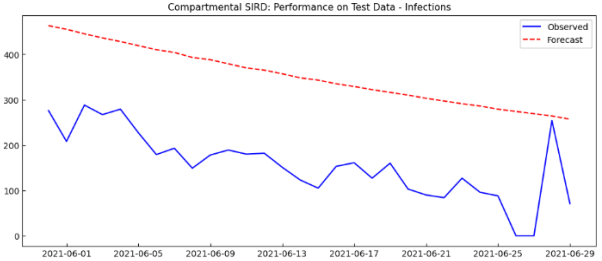




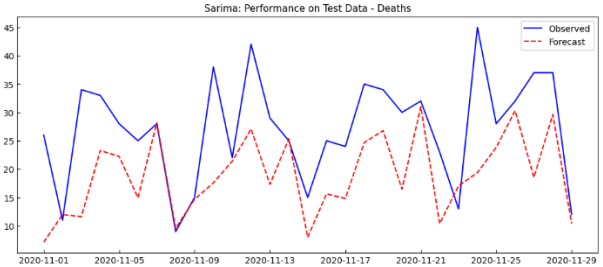


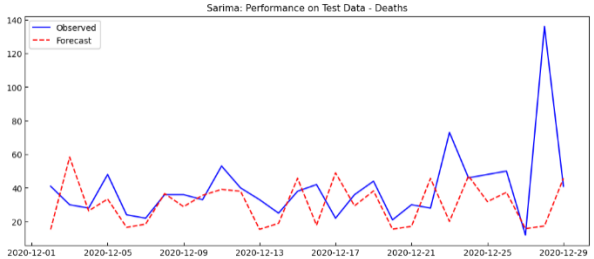


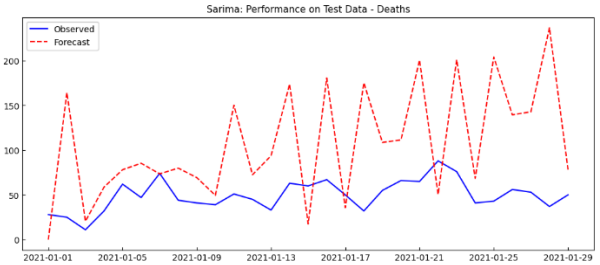

Compartmental SIRD: Performance on Test Data - Infections
Observed
Forecast
1100
1000
900
800
700
600
500
400
300
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17
2021-05-21
2021-05-25
2021-05-29

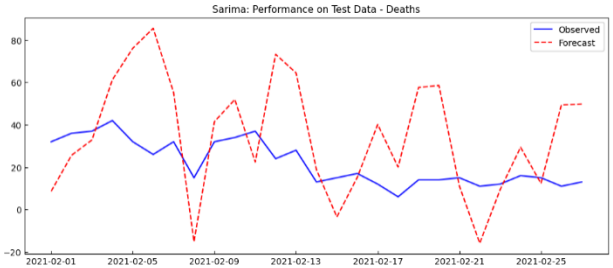

Compartmental SIRD: Performance on Test Data - Infections
Observed
Forecast
400
300
200
100
0
2021-06-01
2021-06-05
2021-06-09
2021-06-13
2021-06-17
2021-06-21
2021-06-25
2021-06-29

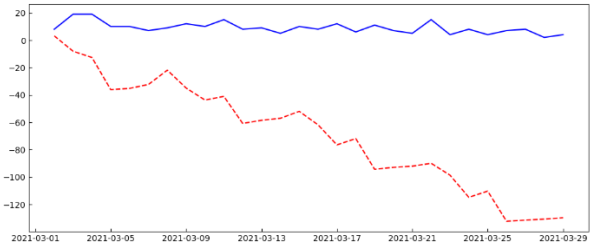
SARIMA – Décès

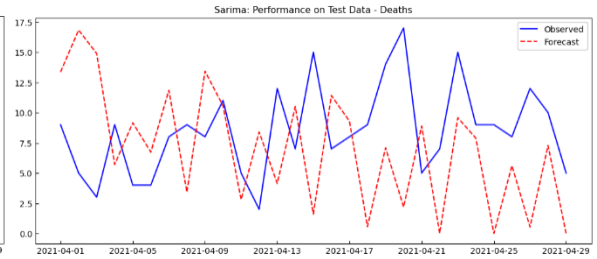


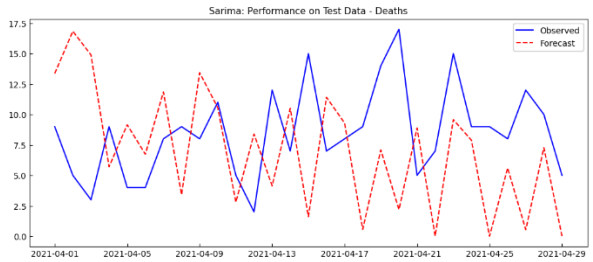


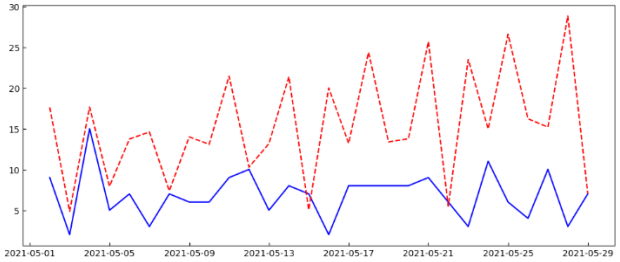


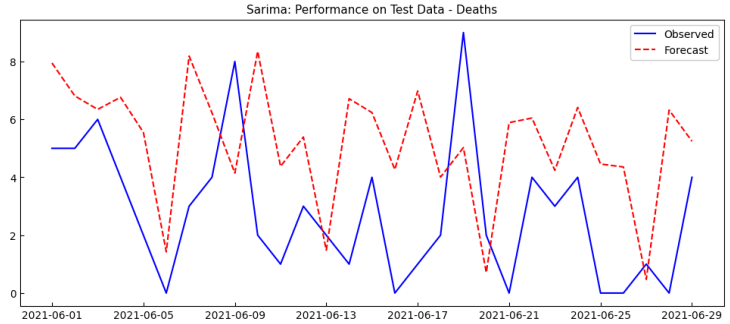




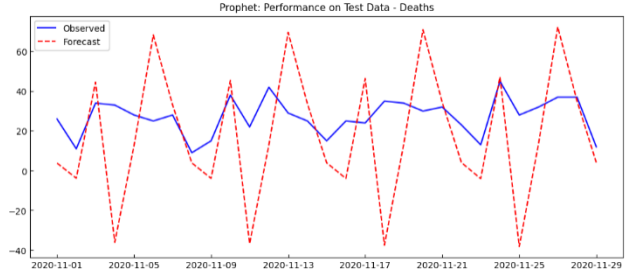


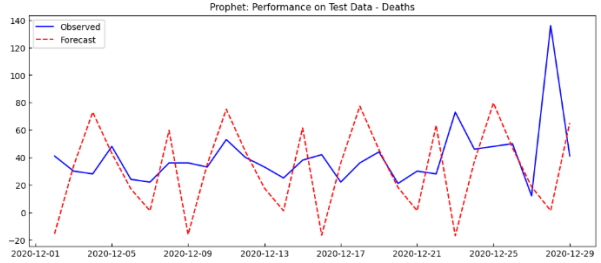


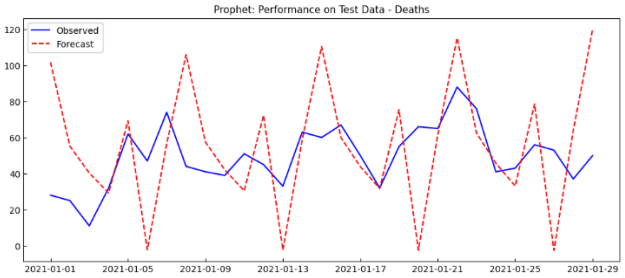


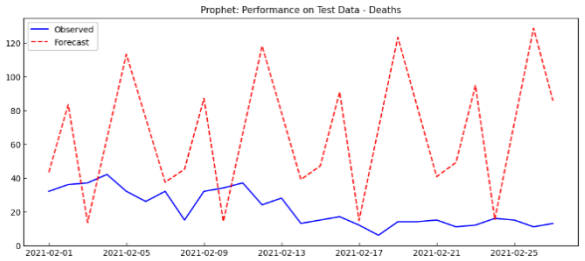


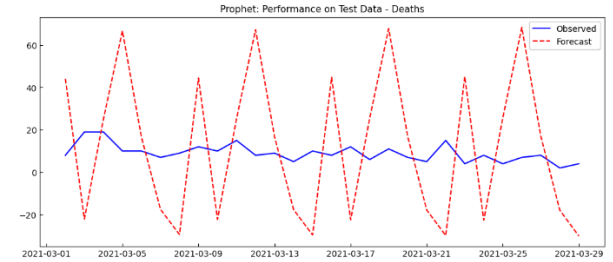
PROPHET Décès

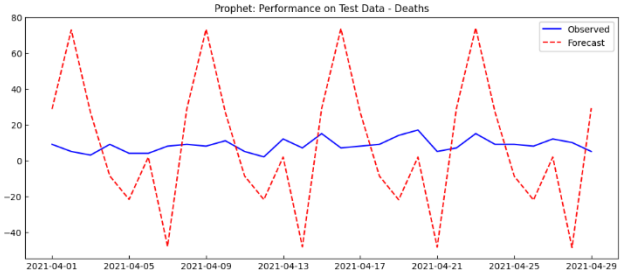


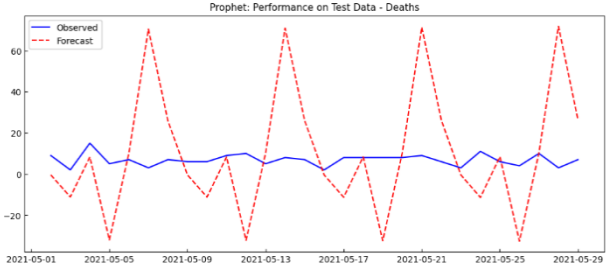


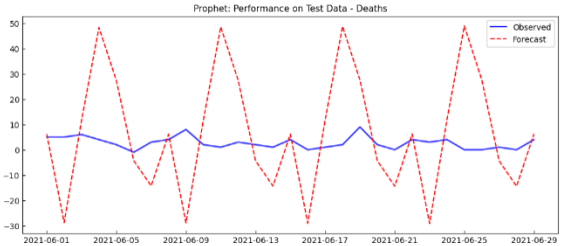

Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-01-01
2021-01-05
2021-01-09
2021-01-13
2021-01-17
2021-01-21
2021-01-25
2021-01-29


Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-02-01
2021-02-05
2021-02-09
2021-02-13
2021-02-17
2021-02-21
2021-02-25

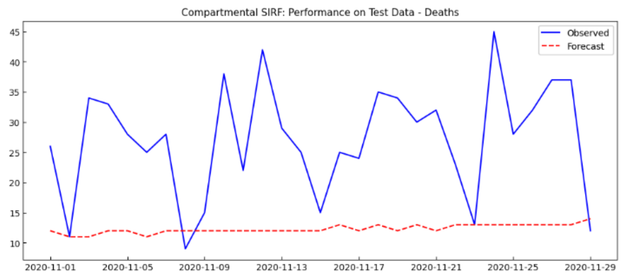

Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-03-01
2021-03-05
2021-03-09
2021-03-13
2021-03-17
2021-03-21
2021-03-25
2021-03-29

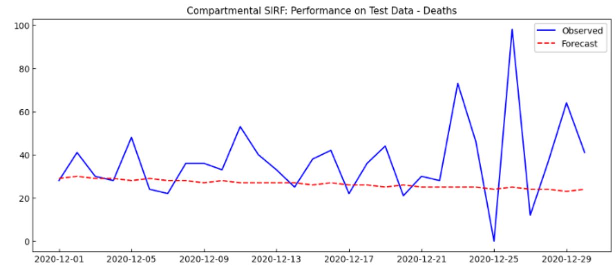

Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-04-01
2021-04-05
2021-04-09
2021-04-13
2021-04-17
2021-04-21
2021-04-25
2021-04-29

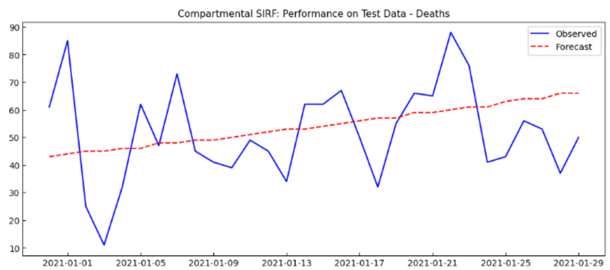

Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17
2021-05-21
2021-05-25
2021-05-29

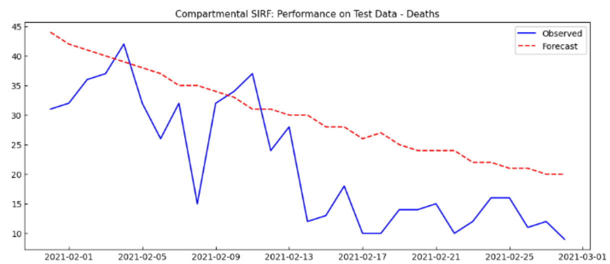

Prophet: Performance on Test Data - Deaths
Observed
Forecast
2021-06-01
2021-06-05
2021-06-09
2021-06-13
2021-06-17
2021-06-21
2021-06-25
2021-06-29

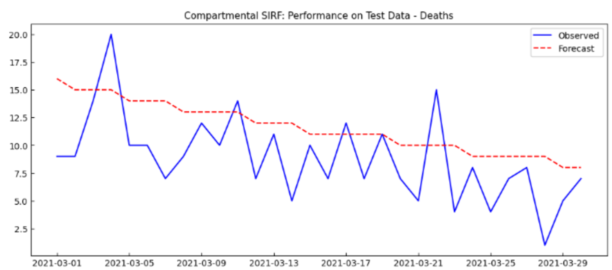
Compartimental

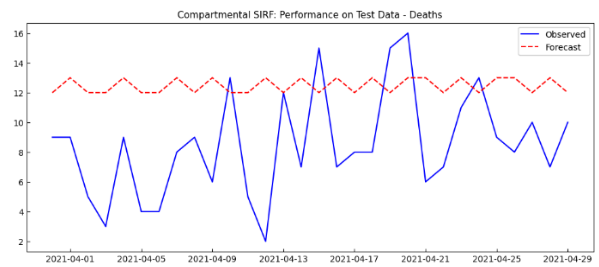


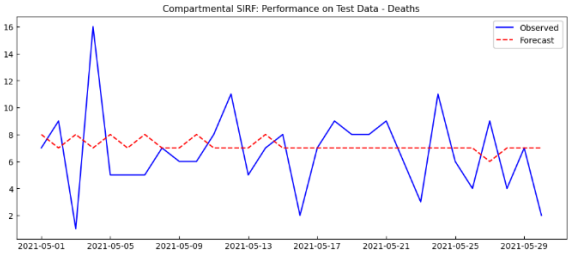


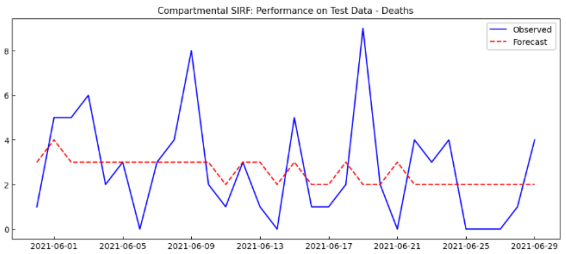






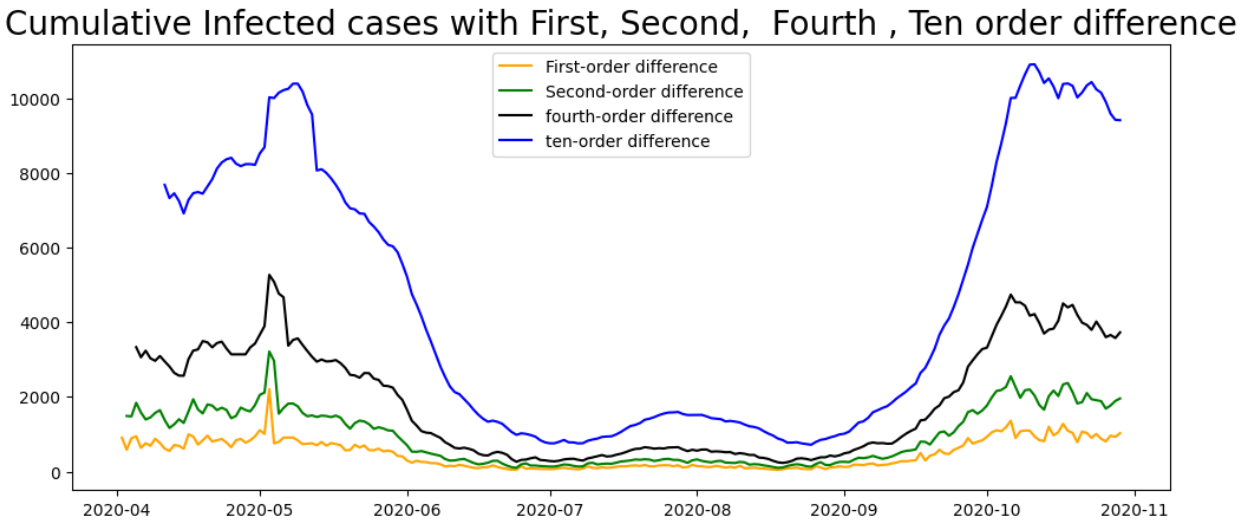



Compartmental SIRF: Performance on Test Data - Deaths
Observed
Forecast
16
14
12
10
8
6
4
2
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17
2021-05-21
2021-05-25
2021-05-29


Compartmental SIRF: Performance on Test Data - Deaths
Observed
Forecast
8
6
4
2
0
2021-06-01
2021-06-05
2021-06-09
2021-06-13
2021-06-17
2021-06-21
2021-06-25
2021-06-29

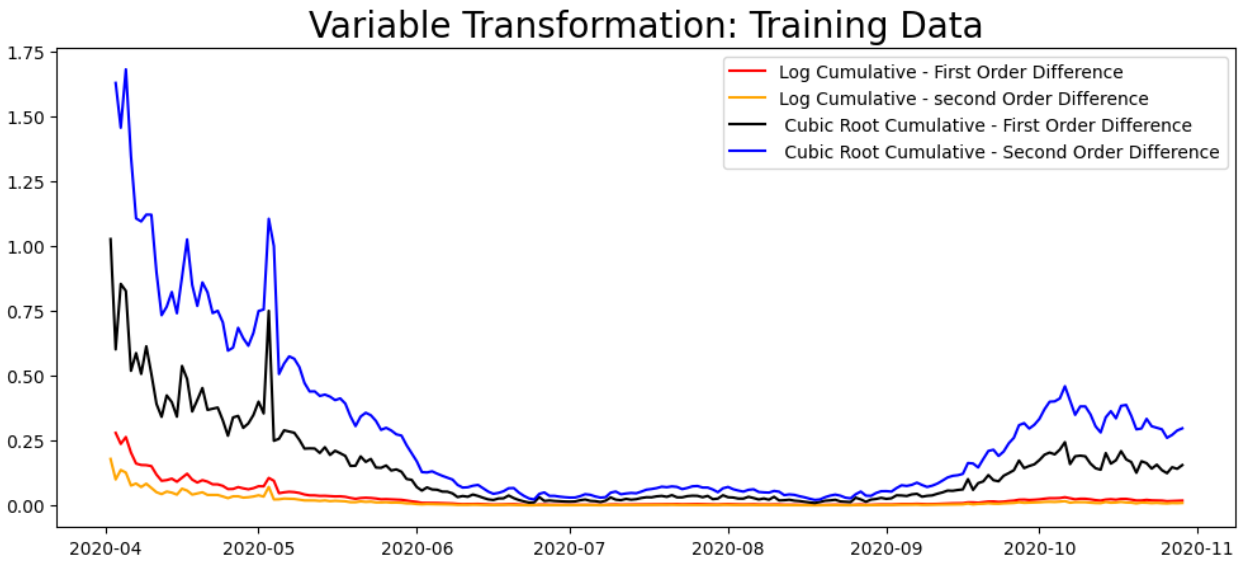
1) Échantillon d'entrainement.

Ci-dessous, une sérié de calcul de différence ont été appliquée à la échantillon d'entrainement des infections pour la rendre stationnaire



Il s'est avéré qu'aucune de ces courbes n'était stationnaire (ADF test)

Ci-dessous, une combinaison de transformation (logarithmique, racine cubique) et de différence ont été appliquée à l'échantillon d'entrainement des infections pour la rendre stationnaire.



Il s'est avéré que la première différence de la fonction logarithme était stationnaire (ADF test)

APPENDICE 3 : CODE ILLUSTRATION PROPHET

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma)$$

```python
np.random.seed(25)
n_changepoints = 4
t = np.arange(10)
s = np.sort(np.random.choice(t, n_changepoints, replace=False))
A = (t[:, None] > s) * 1
delta = np.random.normal(size=n_changepoints)#delta = pm.Laplace('delta', 0,
print(t)
print('----S----')
print(s)
print('----A---')
print(A)
print('----delta----')
print(delta)
print('---adjT=A @ delta-----')

adjT=A @ delta
print(adjT)
print('---growth = (k + A @ delta) * t-----')
growth = (k + A @ delta) * t
print(growth)
print('---gamma = -s * delta-----')
gamma = -s * delta
print(gamma)

print('---offset =m + A @ gamma-----')
offset = m + A @ gamma
print(offset)
print('trend = growth + offset-----')
```

```
-----time-----
[0 1 2 3 4 5 6 7 8 9]
----S----
[1 3 5 8]
----A---
[[0 0 0 0]
 [0 0 0 0]
 [1 0 0 0]
 [1 0 0 0]
 [1 1 0 0]
 [1 1 0 0]
 [1 1 1 0]
 [1 1 1 0]
 [1 1 1 0]
 [1 1 1 1]]
----delta----
[-0.55229723 -0.78527129  0.02512239 -0.3785463 ]
---adjT=A @ delta-----
[ 0.          0.         -0.55229723 -0.55229723 -1.33756852 -1.33756852
 -1.31244612 -1.31244612 -1.31244612 -1.69099243]
---growth = (k + A @ delta) * t-----
[ 0.          1.          0.89540555  1.34310832 -1.35027407 -1.68784259
 -1.87467674 -2.18712286 -2.49956898 -6.21893183]
---gamma = -s * delta-----
[ 0.55229723  2.35581387 -0.12561197  3.02837042]
---offset =m + A @ gamma-----
[ 5.          5.          5.55229723  5.55229723  7.9081111   7.9081111
  7.78249913  7.78249913  7.78249913 10.81086954]
trend = growth + offset-----
[5.         6.         6.44770277 6.89540555 6.55783703 6.22026851
 5.90782239 5.59537627 5.28293014 4.59193772]
```

```
np.random.seed(6)
def fourier_series(t, p=365.25, n=10):
    # 2 pi n / p
    x = 2 * np.pi * np.arange(1, n + 1) / p
    # 2 pi n / p * t
    x = x * t[:, None]
    x = np.concatenate((np.cos(x), np.sin(x)), axis=1)
    return x

n = 1
t = np.arange(1000)
beta = np.random.normal(size=2 * n)
plt.figure(figsize=(16, 6))
plt.plot(fourier_series(t, 365.25, n) @ beta)
print(beta)
```