Etudiants : Steve LILEKE & Amadou Lindor FALL

Encadrant : Frederic PROÏA

M2 DATA SCIENCE -Université d'Angers

SÉRIES CHRONOLOGIQUES

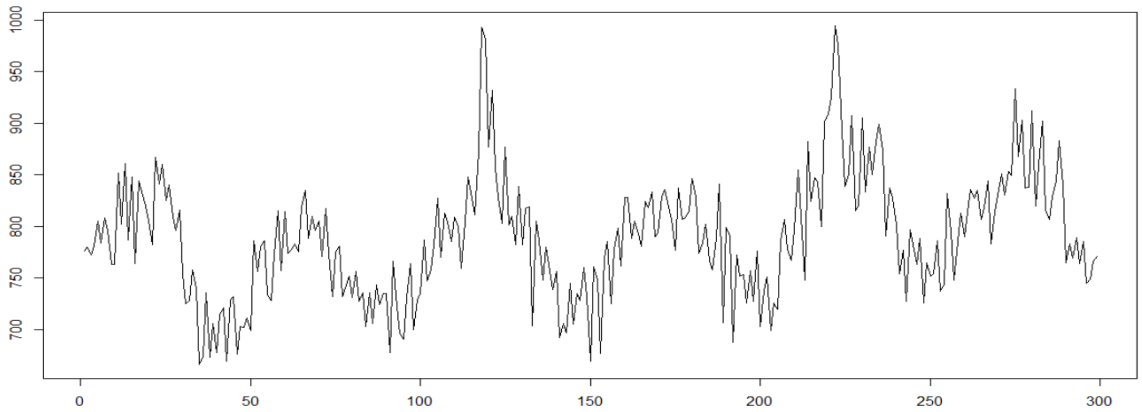
RAPPORT DU PROJET 2020-2021

Notre étude porte sur le jeu de données nécrologiques hebdomadaires des habitants du Minnesota. En effet, nous avons les observations allant de la semaine 40 de l'année 2010 à la semaine 26 de l'année 2016. Ce jeu de données connait une troncature de 78 semaines. Ces données tronquées sont censées être prédites avec précision grâce au meilleur modèles choisi. Ainsi, notre plan d'étude se fera en 3 principales étapes

1) *Exploration des modèles*
2) *Choix du meilleur model*
3) *Prédictions des valeurs futures*

1) EXPLORATION DES MODELES

Tout d'abord, nous aurions pu passer au log notre série, pour atténuer les fortes valeurs que nous avons. En effet, la moyenne des décès est de 789.1371 morts par semaine. Ce passage au log nous donnerait une moyenne de 6.670766 morts par semaine. La série suit un modèle additif. C'est d'ailleurs la raison pour laquelle il n'y a pas de différence remarquable entre la série originale et son logarithme. Nous trouvons que garder les valeurs de la série originale, sans passer au log, serait plus intéressant. Affichons graphiquement l'évolution de notre série :



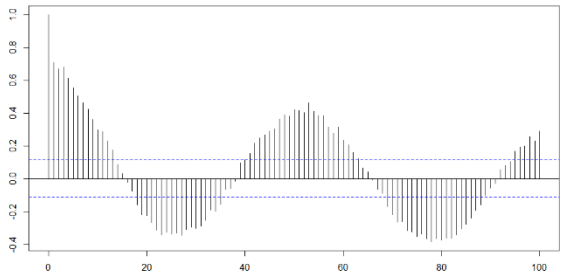
Evolution de la série

Nous remarquons une certaine périodicité avec une intensité constante de notre série. En revanche, cette série est loin d'être stationnaire. Notre premier reflexe fut de faire une régression linéaire et d'en tirer certaines informations comme la justification d'une relation croissante linéaire du nombre de décès par rapport au temps.

```
Call:
lm(formula = dataprojet$Deaths ~ Tps)

Residuals:
     Min       1Q   Median       3Q      Max
-120.137  -41.616    0.797   33.063  210.368

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 758.64515    6.48272 117.026  < 2e-16 ***
Tps           0.20328    0.03746   5.427 1.19e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.91 on 297 degrees of freedom
Multiple R-squared:  0.09021,   Adjusted R-squared:  0.08715
F-statistic: 29.45 on 1 and 297 DF,  p-value: 1.193e-07
```

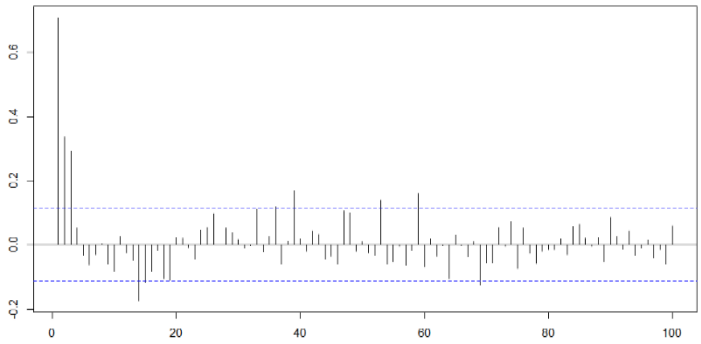
La relation linéaire n'est pas justifiée malgré la significativité des coefficients. De plus, le R2 ajusté est très petit.

Faisons une ACF et un PACF de notre jeu de données.



La sortie de notre ACF ne nous permet pas d'en tirer un modèle. En effet, à aucun moment nos pics se stabilisent et restent dans l'intervalle de confiance

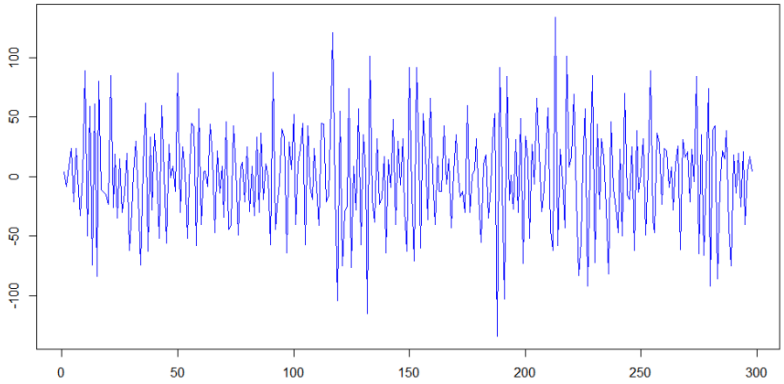
ACF de la série normale



Cependant, avec notre PACF, nous pouvons y sortir un AR (3). La PACF de la série originale est quasi nulle après le rang 3. Nous considérons que les pics qui suivent le 3 -ème sont des artefacts.
Donc **Modèle 1 : AR (3) [BIC=3043]**

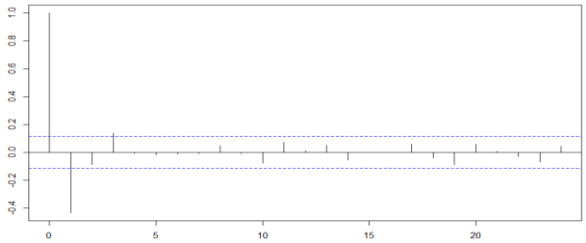
PACF de la série normale

Visuellement, nous avions affirmé que notre série est loin d'être stationnaire. Confirmons ceci par les tests kpss et adf . La p-value du test kpss est de 0.01 alors que celle du test adf est de 0.0448. Ces 2 tests se contredisent sur la stationnarité de la série. Nous pouvons essayer de travailler avec la série différenciée pour voir si le rendu est meilleur.



Evolution de la série différenciée

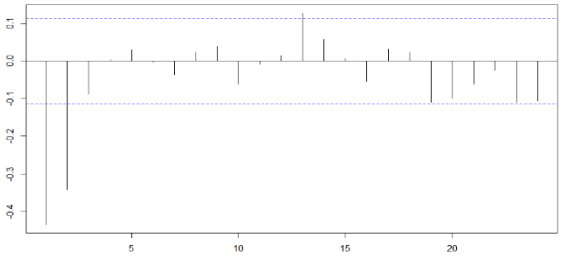
Contrairement à la série originale, cette série semble clairement stationnaire. La pvalue du test kpss est de 0.1 et celle du test adf est de 0.01. Ces 2 tests ne se contredisent pas ,contrairement au cas précédent. La stationnarité semble être justifiée



ACF de la série différenciée

Cette sortie de l'acf, avec des pics restant dans notre intervalle de confiance à partir du rang 2, nous influence pour choisir un MA(2) pour une modélisation de cette série différenciée

Modèle 2 : MA (2) sur la série diff [BIC=3038.55]

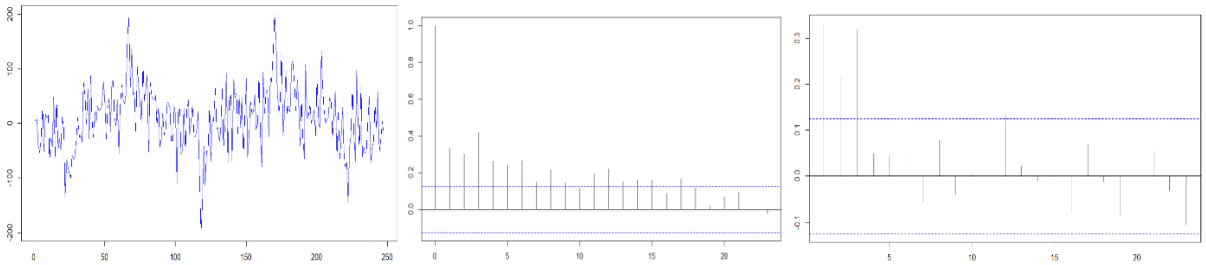


PACF de la série différenciée

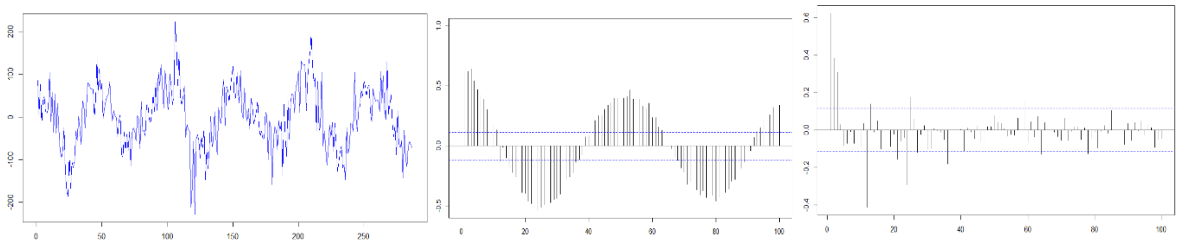
La sortie de notre Pacf , quant à elle, nous lance dans un AR(2) pour la modélisation de notre série différenciée

Modèle 3 : AR (2) sur la série diff [BIC=3034.21]

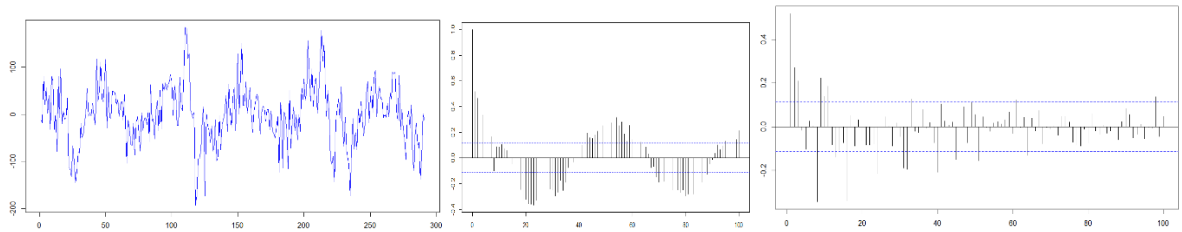
Toujours dans notre exploration des modèles, nous pouvons essayer de parcourir les différenciations saisonnières. Cependant, ce procédé pourrait supprimer une période entière de notre jeu de données. Ça ne serait pas idéal si la période est trop grande. Donc nous pouvons essayer de tester un certain nombre de périodes classiques, pour en tirer des modèles pertinents. De base, nous partons avec une période de 52 semaines, ce qui correspond à une période annuelle. Nous allons ensuite voir ce que nous pouvons avoir avec une période mensuelle (de 4 semaines), une période bimestrielle (8 semaines) et une période trimestrielle (12 semaines) . Nous avons respectivement les graphes faisant référence à l'évolution de la série suivant sa différenciation, son acf et son pacf ci-dessous :



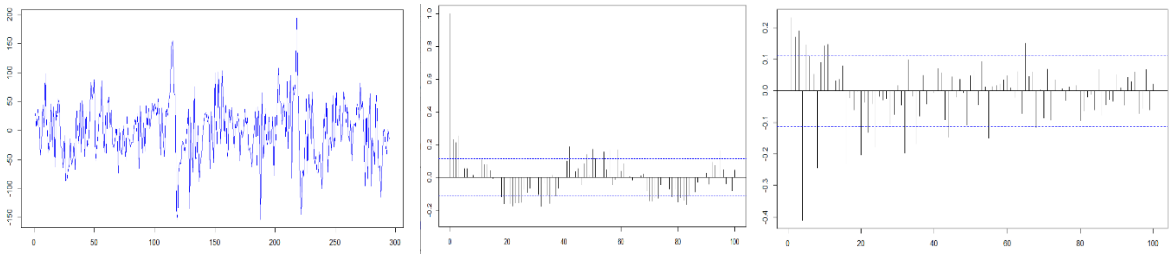
Période=52



Période=12



Période=8



Période=4

Par le même procédé que précédemment, nous retiendrons les modèles suivants :

Modèle 4 :AR (3), Période=52 **[BIC=2670.74]**

Modèle 5 : AR (3), Période=12 **[BIC=3084.55]**

Modèle 6 :AR (3), Période=8 **[BIC=3037.36]**

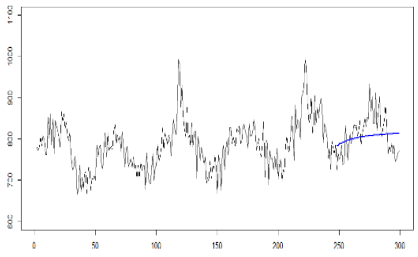
Modèle 7 :AR (3), Période=4 [**BIC=3148.96]**

Tous les modèles considérés ont pratiquement des résidus bruits blancs gaussiens.

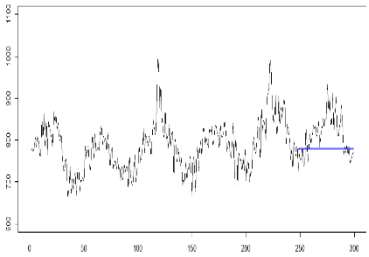
La commande de l'auto. arima sur la série originale, en autorisant la recherche d'une tendance et l'ajout d'une composante constante, nous donne un **ARMA (1,1)(Model 8 [BIC=2529])** . Une dernière alternative était de considérer le lissage de **Holt Winter (Model LHW)** , ce qui nous fait en un total de 9 modèles.

2) Choix du meilleur modèle

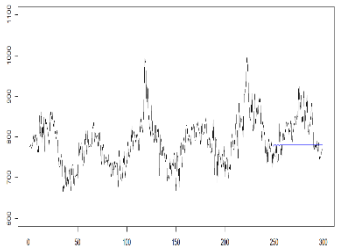
Pour éviter la saturation de ce présent rapport, nous accentuerons l'analyse sur les modèles les plus pertinents et les plus logiques.



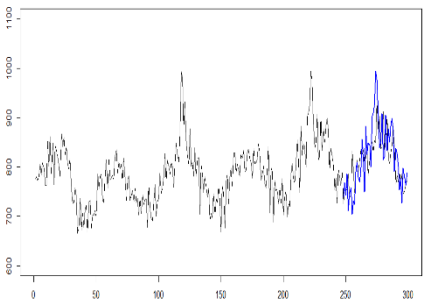
Modèle 1 -MSE=2089



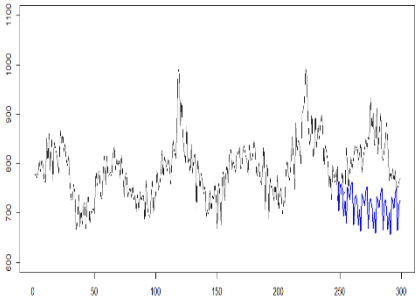
Modèle 2 -MSE=3325



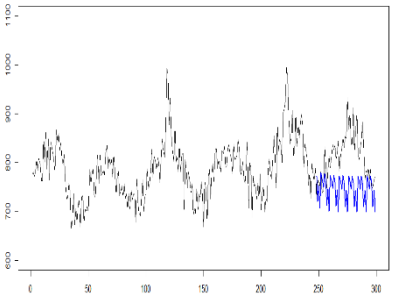
Modèle 3-MSE =3273



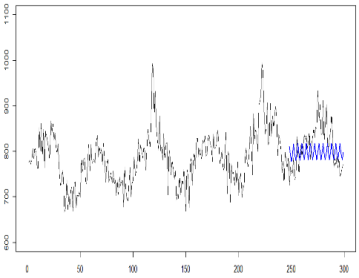
Modèle 4-MSE= 2709



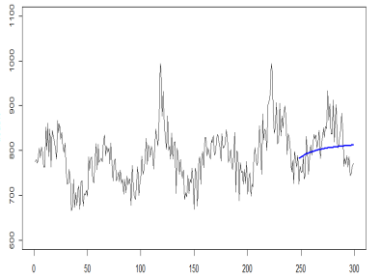
Modèle 5 -MSE=12768



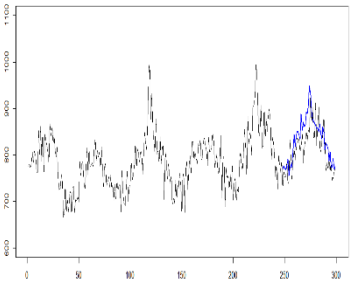
Modèle 6 - MSE=7436



Modèle 7 -MSE=2505



Modèle 8 – MSE=2134

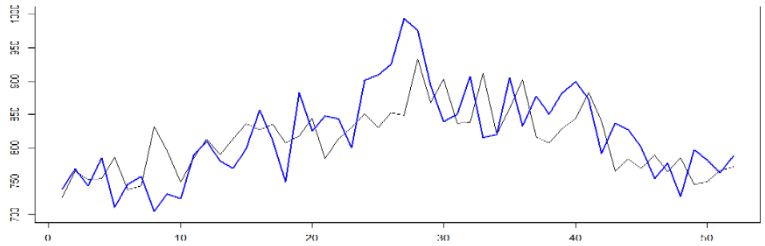


Modèle LHW-MSE=1470

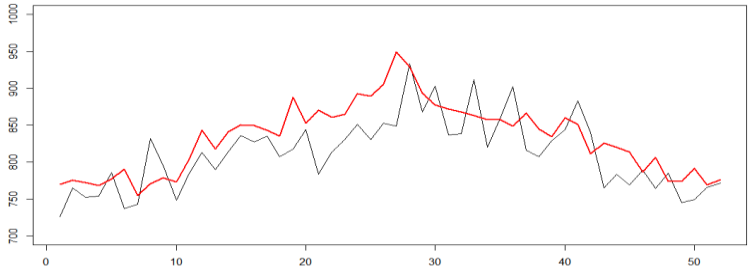
Pour chacun des modèles énoncés, nous avons son MSE et son BIC. Le MSE n'est en rien un critère suffisant pour témoigner de la pertinence d'un modèle. En effet, il y a des points où les prédictions sont exactement égales aux valeurs réelles. Ce qui réduit fortement le MSE sans que le modèle suive parfaitement l'évolution du signal. De ce fait, nous privilégions le modèle qui offre le meilleur compromis entre minimisation du BIC, du MSE et superposition du graphe de la prédiction sur la série. **Dans cette optique, nous choisissons le modèle 4 et le dernier modèle, Holt Winter**

Le Holt Winter ou plus précisément le lissage exponentiel de Holt est applicable dès que le nombre d'observations est au moins égal à 2. La taille de notre signal suffira pour lui permettre de faire de bonnes prédictions. Les paramètres alpha, beta et gamma sont estimés par le logiciel en minimisant

les erreurs comme dans les moindres carrés. En effet, ce modèle est celui qui présente le plus petit MSE parmi les modèles testés. Graphiquement, nous avons :

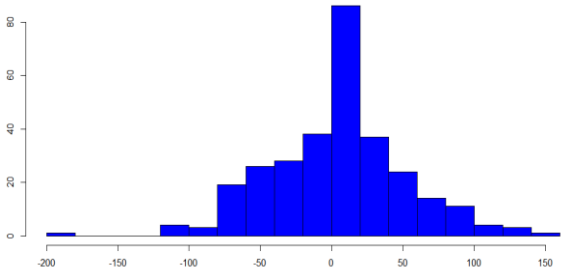


Superposition de la prédiction du modèle 4 sur la dernière période



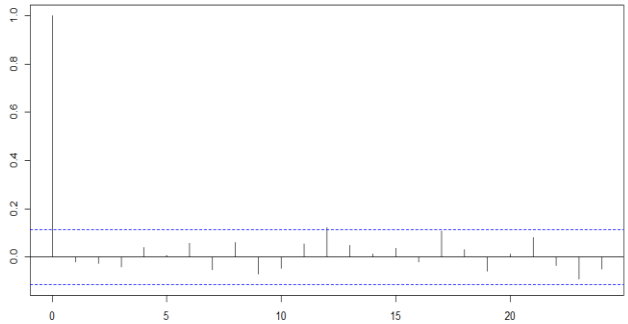
Superposition de la prédiction du modèle Holt Winter sur la dernière période

Ces 2 modèles suivent assez bien l'évolution de la série. Visuellement, il nous est impossible de dire quel modèle se rapproche le plus des vraies valeurs. Un des derniers critères sur lequel nous pouvons nous baser pour choisir notre model est le comportement de ses résidus

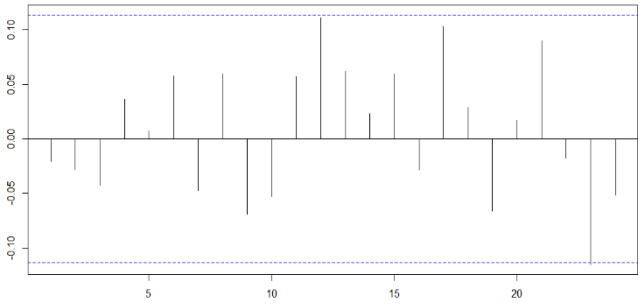


Histogramme des résidus du modèle 4

La forme de l'histogramme des résidus du modèle 4 ressemble à une courbe en cloche de la densité gaussienne. Cependant, le test de Shapiro affirme le contraire avec une p-value de 0.00077. On décide de privilégier l'aspect visuel au test de Shapiro qui a des résultats parfois douteux, voire peu fiables.



ACF des résidus du modèle 4



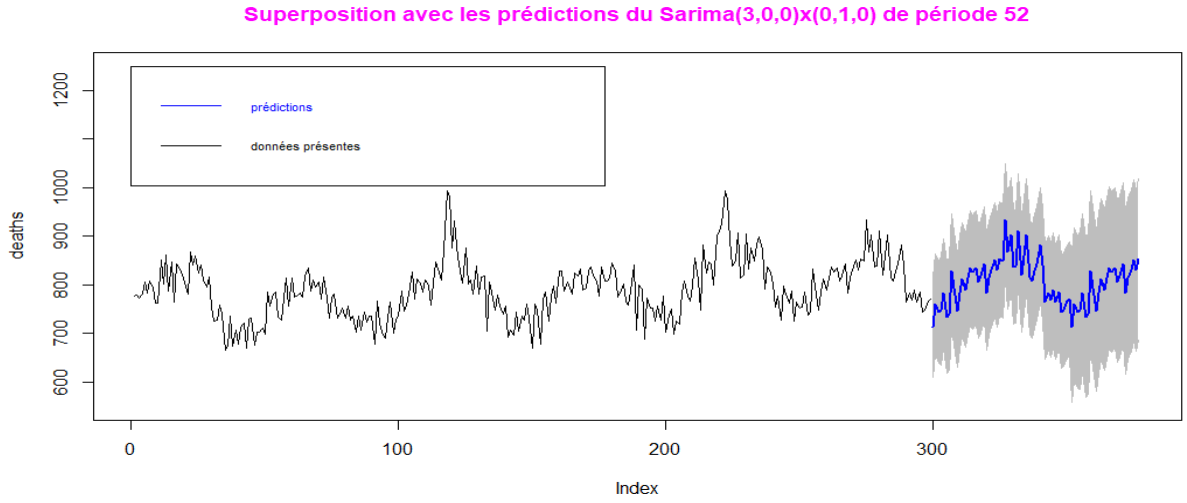
PACF des résidus du modèle 4

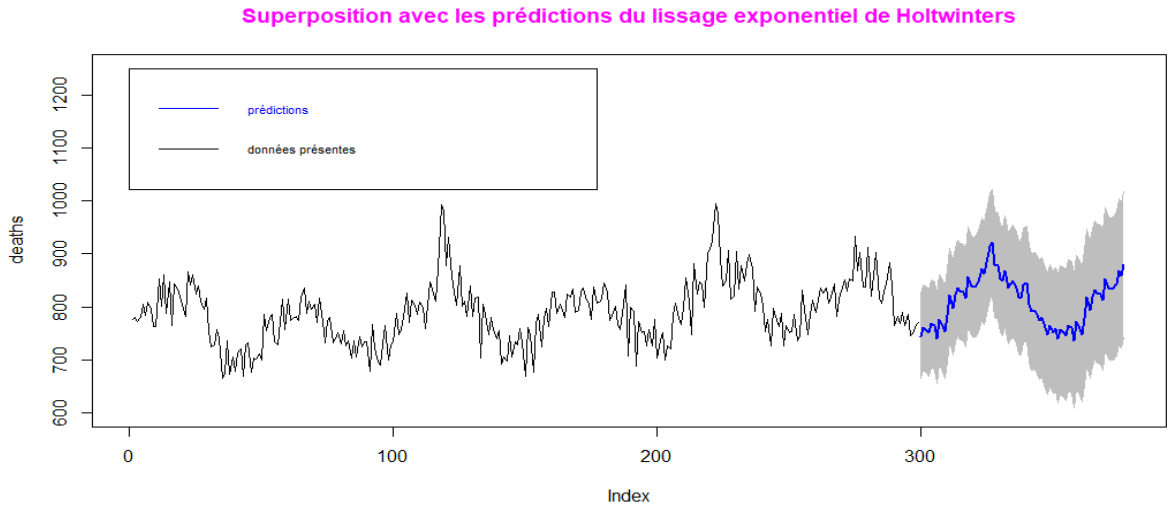
Les sorties de ces autocorrélations montrent que ces résidus sont bruits blancs. Ce qui est confirmé avec le test de Ljung-Box avec une p-value 0.7236, nous permettant de ne pas rejeter l'hypothèse nulle d'absence de corrélations.

Nous portons notre choix final de modèle sur le SARIMA(3,0,0)(0,1,0)[52] compte tenu de la robustesse et de la rigueur de ses fondements théoriques.

3)Prédictions des valeurs futures

Il est clair que notre choix portera sur ces deux modèles pour prédire le reste de notre jeu de données, à savoir les 78 prochaines semaines. De ce fait, en représentant graphiquement ces futures données ainsi que leurs intervalles de prédiction, nous avons :





De là, notre choix de préférer le module 4 est encore plus appuyé par le fait que, visuellement, ses prédictions semblent être plus compatibles avec le tracé des périodes précédentes que celui du modèle s'appuyant sur le lissage exponentiel de Holt. Nous vous invitons à consulter le fichier R joint à ce rapport pour avoir les valeurs de ces prédictions.

Il est important de garder à l'esprit que ces prévisions sont purement hypothétiques et que la prédiction de la mort relève d'un nombre conséquent de facteurs. Si on avait utilisé ce modèle pour prédire le nombre de décès enregistré durant l'année 2020, ces prédictions seraient fortement biaisées par l'impact qu'a eu cette pandémie que nous vivons actuellement .