Table des matières

Introduction au rapport

Le présent rapport se veut être l'analyse de différentes séries chronologiques. Plus précisément, il a pour but de trouver le modèle qui s'ajuste le plus exactement à nos échantillons de données. En bref, le travail consiste à séparer nos bases de données en un échantillon *entraînement* qui sert à trouver un modèle et en un échantillon *test* qui sert à valider la précision du modèle retenu. Le travail est fait à partir de deux bases de données pour un total de cinq variables à modéliser. La première base de données, traitée au numéro 1, contient les données mensuelles du taux de change du dollar américain par rapport à l'euro de janvier 1999 à décembre 2016. La seconde base de données contient une série de statistiques de la *SAAQ*, soit le nombre d'accidents automobiles avec dommages corporels, le nombre de personnes accidentées, le nombre de demandes d'indemnités et le coût total de l'indemnisation (en dollar constant 2015) pour les années 1978 à 2015 inclusivement. Ces données seront traîtées au numéro 2. Afin d'alléger la lecture du présent rapport, le code R est placé en annexe.

Taux de change US/Euro

Analyse primaire du jeu de données

Tout d'abord, nous observons, à la *Figure 15*, la série chronologique du taux de change américain/européen depuis janvier 1999 jusqu'à décembre 2016. Les données ont été collectées de façon quotidienne pour ensuite être transformées en taux mensuelles.

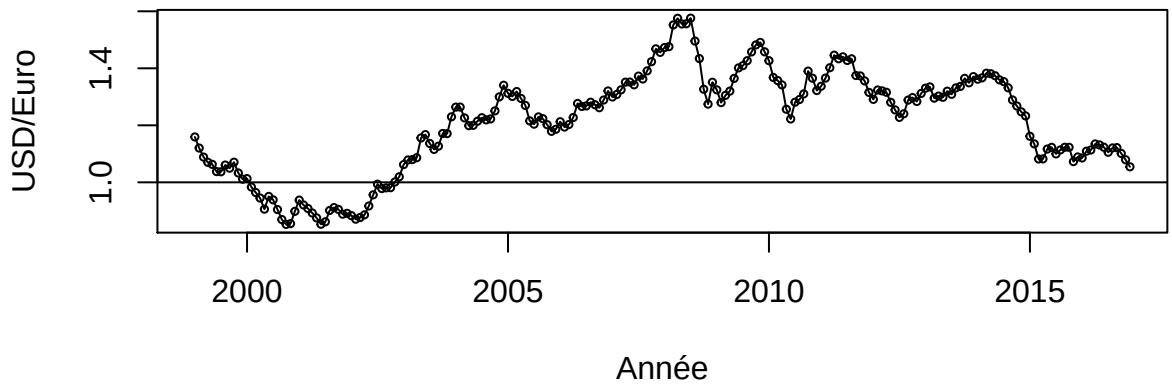


Figure 1. Taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

Grâce à une première analyse, on remarque sur le tableau de la fonction d'autocorrélation échantillonale qu'il y a une forte autocorrélation dans notre jeu de données et que celle-ci diminue lentement plus le lag augmente (voir *Figure 2*). On soupçonne donc fortement ce processus d'être non-stationnaire. Il n'est pas nécessaire de tracer le graphique de la fonction d'autocorrélation partielle puisqu'on doit d'abord stationnariser notre processus avant de réellement débuter notre analyse.

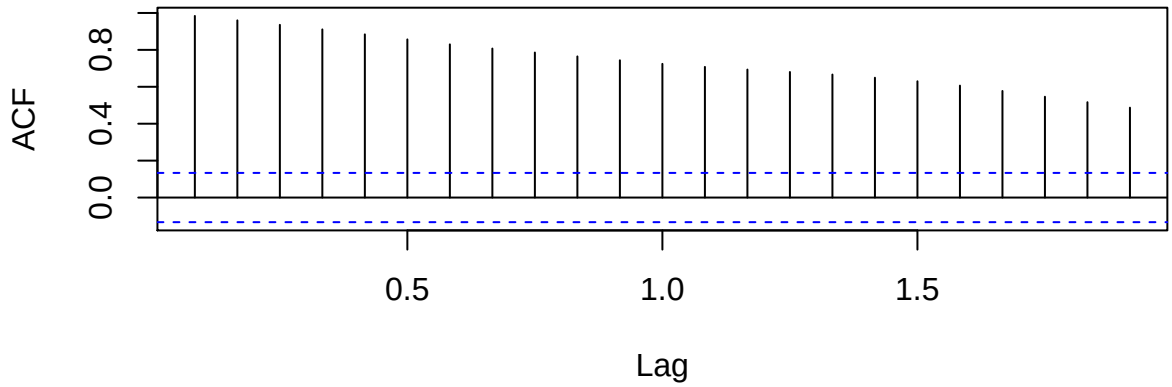


Figure 2. Fonction d'autocorrélation échantillonales du taux de change US/Euro (lag étant exprimé en années).

Test de stationnarité

Le test de stationnarité de Dickey-Fuller est alors effectué sur notre base de données. On remarque que, pour un ordre $k = 2$ de processus auto-autoregressif tel que proposé par la fonction *ar* de *R*, notre processus est non stationnaire à un niveau de confiance de 5% avec une *p-value* très forte de 91.33%. On vérifie alors si une transformée Box-Cox est apppropriée à notre modèle avant de tester la stationnarité d'une première différenciation.

Transformation des données

Étant donné que les données sont positives, il est possible d'utiliser la transformée de Box-Cox afin de stabiliser notre processus. On rappelle que la famille des fonctions de puissance est définie de la façon suivante:

$$g(x) = \frac{x^\lambda - 1}{\lambda} \times 1_{\{\lambda \neq 0\}} + \ln(x) \times 1_{\{\lambda = 0\}}$$

λ est donc déterminé en maximisant la fonction de log-vraisemblance de nos données fournie à la *Figure 3*.

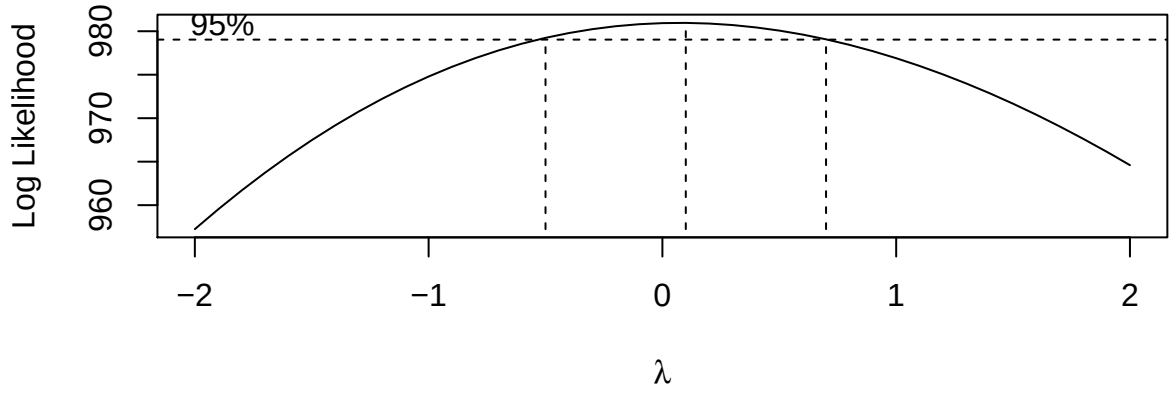


Figure 3. Fonction du logarithme de vraissemblance de la transformée de Box-Cox de la série chronologique du taux de change US/Euro.

On constate que $\lambda = 0.1$ semble être l'estimé MLE situé au centre de l'intervalle de confiance 95%, soit $]-0.5, 0.7[$. Puisque $\lambda = 0$ est dans notre intervale de confiance, cette valeur du paramètre est également à considérer. Dans le cas où le processus serait à différencier une seule fois, la transformation logarithmique est particuliairement intéressante puisqu'elle permet de modéliser non pas le «prix» du taux de change, mais sont rendement comme il est usage de le faire dans le cas d'outils financiers. En effet, soit Y_t le prix d'un outil financier au temps t, le modèle logarithmique modélise le rendement de la façon suivante [1]

$$Y_t = Y_{t-1} e^X$$

1. WEISHAUS, Abraham, SOA Exam MFE Models for Financial Economics 10th Edition.

où X, le rendement continu mensuel, est la variable aléatoire à modéliser. Ainsi, suite à une première différenciation de notre modèle logarithmique, on obtient direction ce rendement. Soit,

$$log\left(\frac{Y_t}{Y_{t-1}}\right) = X$$

On conserve alors cette transformation à condition que la première différenciation de notre processus soit stationnaire. Le graphique de cette transformation est affiché à la *Figure 4*.

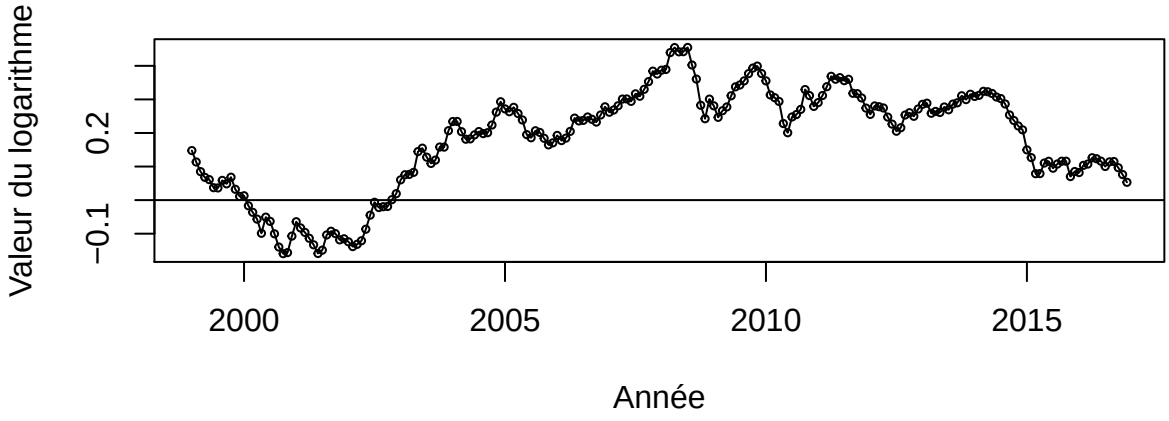


Figure 4. Logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2016.

Suite à une première différenciation, on effectue à nouveau le test augmenté de Dickey-Fuller. Avec un processus autorégressif d'ordre 1 tel que suggéré par la fonction *ar*, on remarque cette fois qu'on ne peut rejetter pas l'hypothèse de stationnarité avec un p-value inférieur à 1%. Ainsi, tel que mentionné précédement, la transformation logarithmique est conservée. Le graphique de la première différenciation du logarithme de la série chronologique à l'étude se trouve ci-dessous à la *Figure 5*.

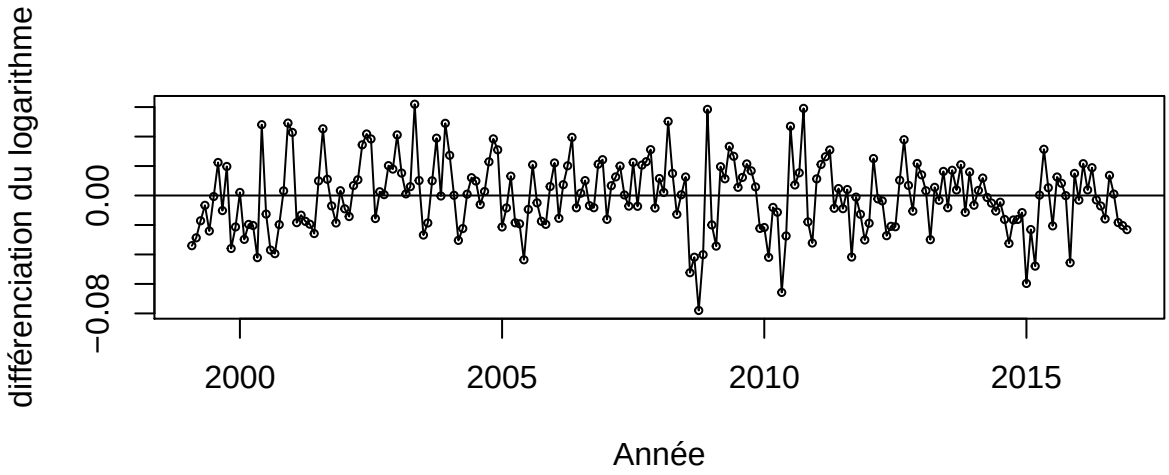


Figure 5. Première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

Identification du modèle

Maintenant que la série est stationnaire, on peut en identifier le modèle. La première différenciation du logarithme du taux de change US/Euro est dorénavent notre modèle de base. On fera ainsi référence à ce modèle par défaut.

On observe les fonctions d'autocorrélation et d'autocorrélation partielle affichées aux *Figure 6* et *Figure 7* de cette série. On remarque que la fonction d'autocorrélation suggère fortement un modèle à moyenne mobile d'ordre 1, soit IMA(1,1) pour notre modèle tansformé alors que la fonction d'autocorrélation partielle suggère un modèle autorégressif d'ordre 1, soit ARI(1,1). On testera alors également le modèle ARIMA(1,1,1) qui est suggéré par la combinaison de ces graphiques.

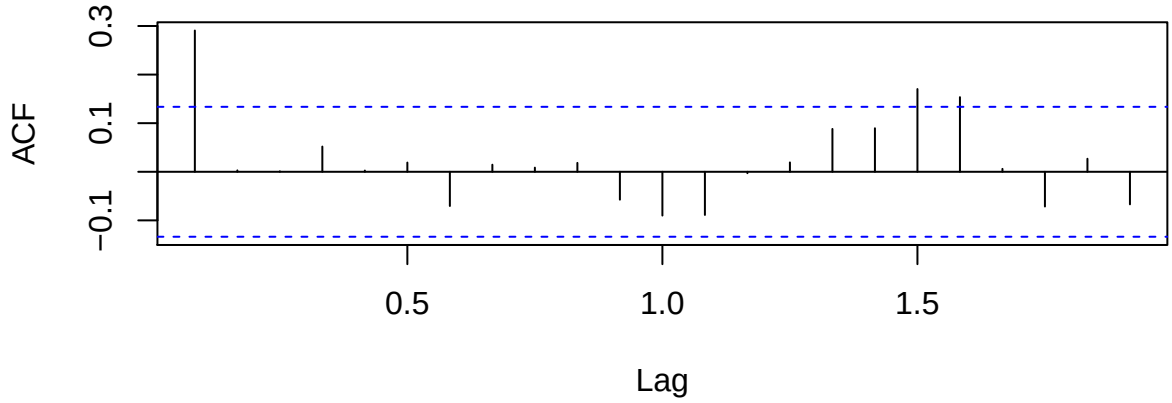


Figure 6. Fonction d'autocorrélation de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

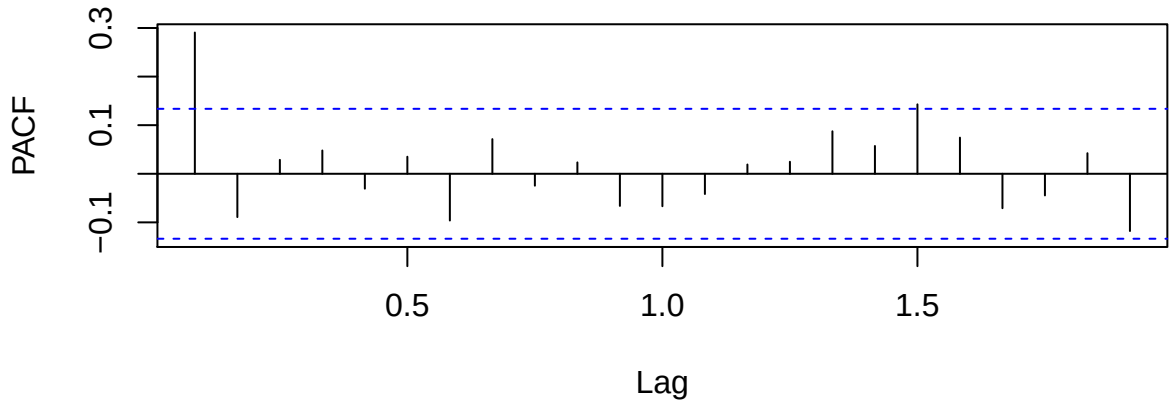


Figure 7. Fonction d'autocorrélation partielle de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

On observe ensuite la fonction d'autocorrélation étendue afin de voir si ce test à un autre modèle à proposer. Le tableau de la fonction EACF est à la *Table I*. On observe d'abord de ce tableau que le modèle IMA(1,1) est suggéré pour la différenciation de notre modèle transformé initial. On peut également chercher à savoir si la valeur du o en ARMA(0,1) et du x en ARMA(1,1) sont significativement différent de leur valeurs inverses (x pour le o et vice versa), sans quoi le modèle ARIMA(1,1,1) serait également suggeré par le tableau EACF. Cependant, comme il fut déjà décidé d'évaluer la pertinance du modèle ARIMA(1,1,1) suite à l'analyse des fonction d'autocorrélation et d'autocorrélation partielle, il n'est pas nécessaire de tester la pertinance de ce modèle selon le EACF.

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o o o o o o o o  o  o  o
## 1 x x o o o o o o o o o  o  o  o
## 2 x o o o o o o o o o o  o  o  o
## 3 x x x x o o o o o o o  o  o  o
## 4 x x o x o o o o o o o  o  o  o
## 5 x o x x o o o o o o o  o  o  o
## 6 x x x x o o o o o o o  o  o  o
## 7 x x x x x x o o o o o  o  o  o
```

Table I. Tableau de la fonction EACF de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

Le dernier test effectué pour ajuster un modèle à notre série chronologique est le test du BIC. Le tableau du BIC (tracé à la *Table II*) propose un modèle ARI(1,1). L'ordre maximal de moyenne mobile pour ce test a été établi à 1, sans quoi la fonction *armasubsets* permettant de tracer le tableau du BIC retourne des avertissements de forte dépendance linéaire entre les paramètres.

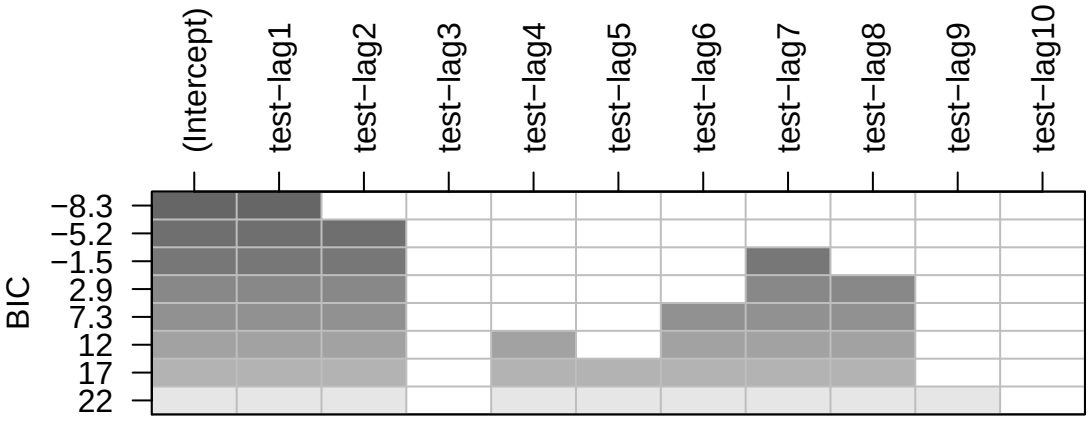


Table II. Tableau de la fonction BIC de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

On étudiera ainsi les quatre modèles proposés par les tests précédents qui peuvent être résumé de la façon suivante;

Résumé	
Tests et tableaux	Modèles suggérés
ACF	IMA(1,1) ARIMA(1,1,1)
PACF	ARI(1,1) ARIMA(1,1,1)
EACF	IMA(1,1)
BIC	ARI(1,1)

On teste alors respectivement les modèles ARI(1,1), IMA(1,1) et ARIMA(1,1,1) suivants

$$W_t - \mu = \phi(W_{t-1} - \mu) + e_t \quad (1)$$
$$W_t - \mu = e_t - \theta e_{t-1} \quad (2)$$
$$W_t - \mu = \phi(W_{t-1} - \mu) + e_t - \theta e_{t-1} \quad (3)$$

Où

$$W_t = \nabla \ln(Y_t)$$

et e_t est un bruit blanc et μ la moyenne de la série chronologique non-centrée. Il est ici important de noter que les paramètres estimés $\hat{\theta}$ et $\hat{\phi}$ ne seront pas les mêmes dans les trois modèles et devront donc être estimés pour chacun d'entre eux.

Estimation des paramètres des différents modèles et tests de résidus

On estime alors les valeurs des différents paramètres selon la méthode du maximum de vraisemblance à l'aide de la fonction *arima*. Les résultats se trouvent à la *Table III*.

Modèles	Paramètres	Valeurs estimés	AIC	Logarithme du max de vraisemblance
IMA(1,1)	μ	0	-1010	506
	θ	0.3118		
IMA(1,2)	μ	0	-1008	506
	θ_1	0.3207		
	θ_2	0.0207		
ARI(1,1)	μ	0	-1009	505
	ϕ	0.2933		
ARIMA(1,1,1)	μ	0	-1008	506
	θ	0.2635		
	ϕ	0.0557		

Table III. Tableau des valeurs estimées des différents coefficients obtenus de la fonction *arima* pour les trois modèles à l'étude du logarithme du taux de change US/Euro.

On peut alors faire l'analyse de nos résidus pour faire notre choix parmi nos trois modèles. Les résidus standardisés de ces trois modèles sont tracés à la *Figure 8*. On remarque d'abord que ces trois courbes se superposent et ne peuvent donc pas être utilisées pour sélectionner un modèle. De plus, on remarque de ce graphique que les valeurs de résidus supérieures à 2 en valeur absolue sont relativement fréquentes ce qui nous poussent à questionner la théorie selon laquelle les résidus suivent une loi normale. En effet, dans le cas

des résidus normaux, les valeurs supérieurs à 2 en valeur absolue ne devrait apparaître qu'environ 2.28% du temps. La fréquence observée est nettement supérieur.

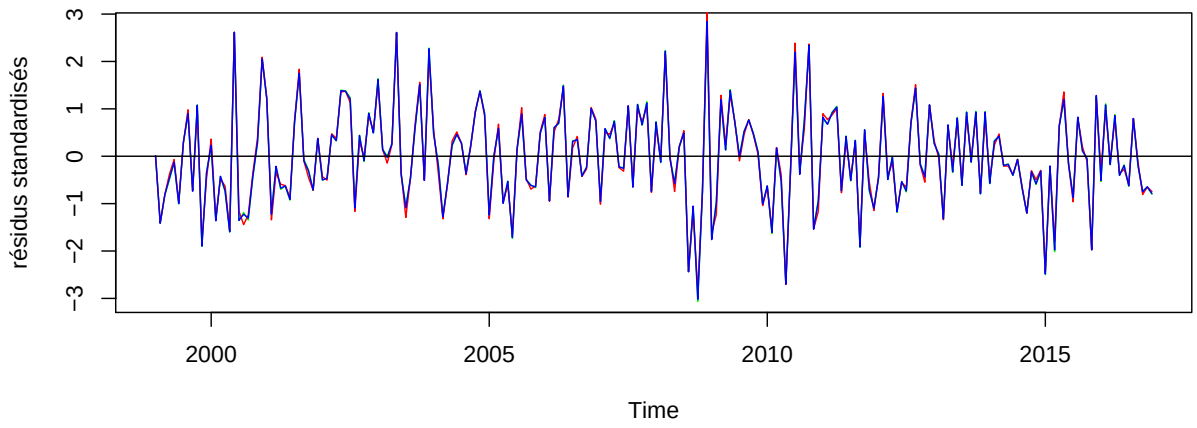


Figure 8. Superposition des graphiques des résidus standardisés des modèles ARI(1,1), IMA(1,1) et ARIMA(1,1,1) du logarithme du taux de change US/Euro.

On regarde ensuite les histogrammes et les graphiques QQ aux *Table IV* et *Figure 9* des résidus. On remarque que les trois histogrammes semblent tracer avec efficacité la cloche de la loi normal. De plus, on remarque des trois graphiques QQ, que les résidus observés sont en tout point simiaire à leur valeur théorique en cas de normalité. Ces deux test semblent donc justifier la distribution normale des résidus de nos trois modèles.

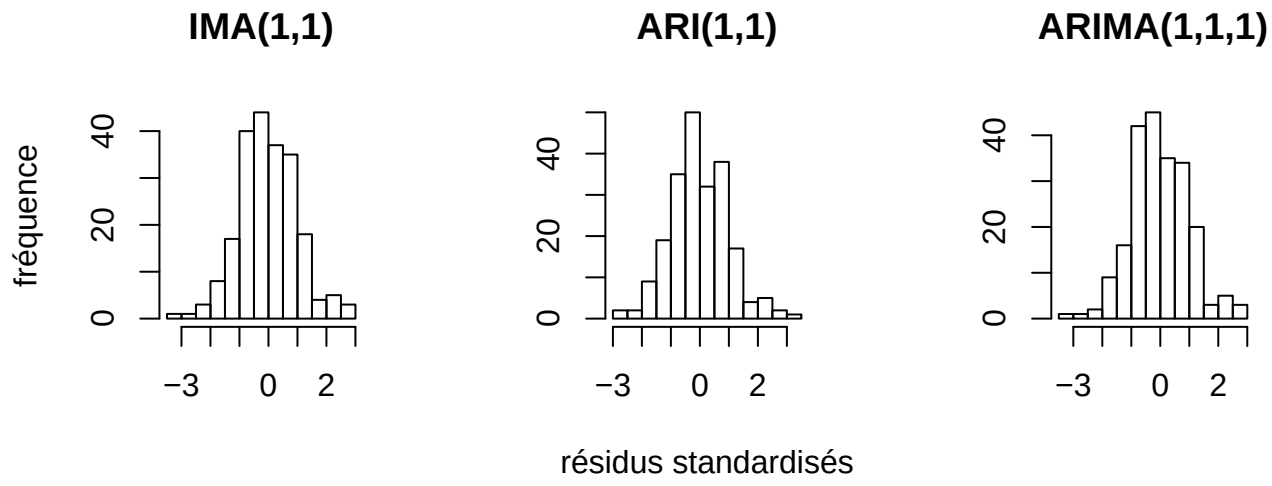


Table IV. Histogramme des résidus standardisés des modèles étuidés du logarithme du taux de change US/Euro.

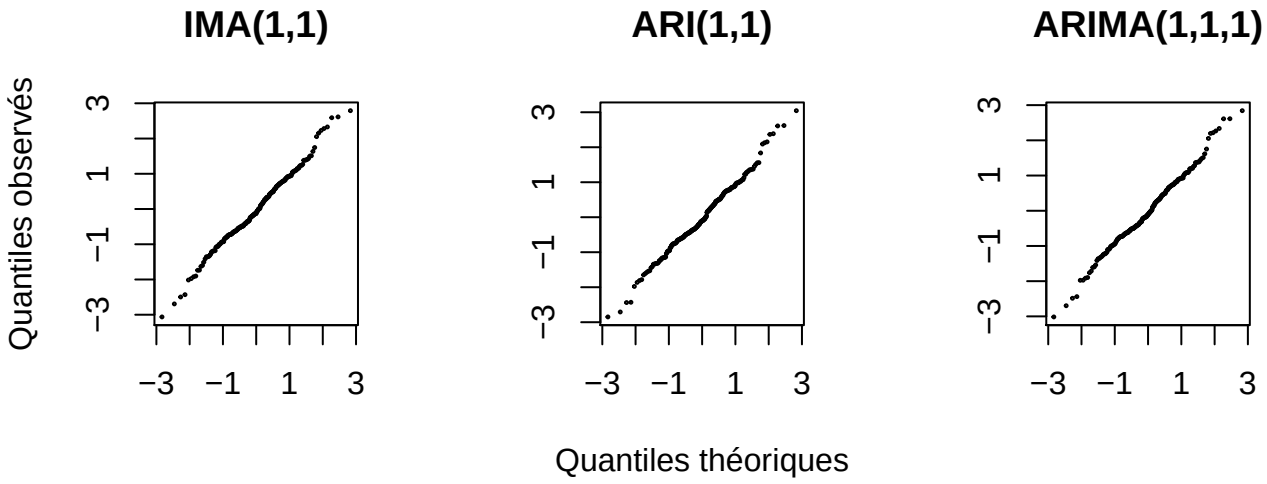


Figure 9. Graphiques QQ des modèles étudiés du logarithme du taux de change US/Euro.

On effectue alors, pour tenter de départager nos modèles, le test de Shapiro-Wilk évaluant le dégré de corrélation entre les quantiles des résidus standardisés et la loi normale standard. L'hypothèse nulle étant que la distribution des résidus suit une loi normale On confirme, avec des p-values respectives de 0.3247, 0.2846 et 0.3051, que les résidus engendrés par les modèles IMA(1,1), ARI(1,1) et ARIMA(1,1,1) sont tous normalement distribués pour un test à un niveau de confiance α de 5%.

On cherche également un modèle dont les résidus sont indépendants, sans quoi on ne retrouve pas un bruit blanc. On effectue alors le run-test sur nos modèles pour tester l'indépendance entre les résidus. Comme l'hypothèse nulle est que les résidus sont indépendants, on conclut à un niveau de confiane de $\alpha = 5\%$ qu'on ne peut rejetter l'hypothèse nulle si la p-value est supérieur à 5%. On trouve alors, pour les modèles IMA(1,1), ARI(1,1) et ARIMA(1,1,1) des valeurs respectives de 0.399, 0.81 et 0.793 pour les p-value de leur run-tests. Ainsi, aucun de ces modèles rejettent l'hypothèse de normalité.

On effectue un dernier test sur nos modèles afin de les départager. Le test de Ljung-Box à pour hypothèse nul que le modèle testé sur notre série chronologique est approprié. On remarque des *Figure 10* à *Figure 12* que les p-values des trois modèle ne permettent pas de rejetter l'hypothèse nulle. Les trois modèles sont donc, selon ce test, appropriés pour modéliser la série chronologique à l'étude.

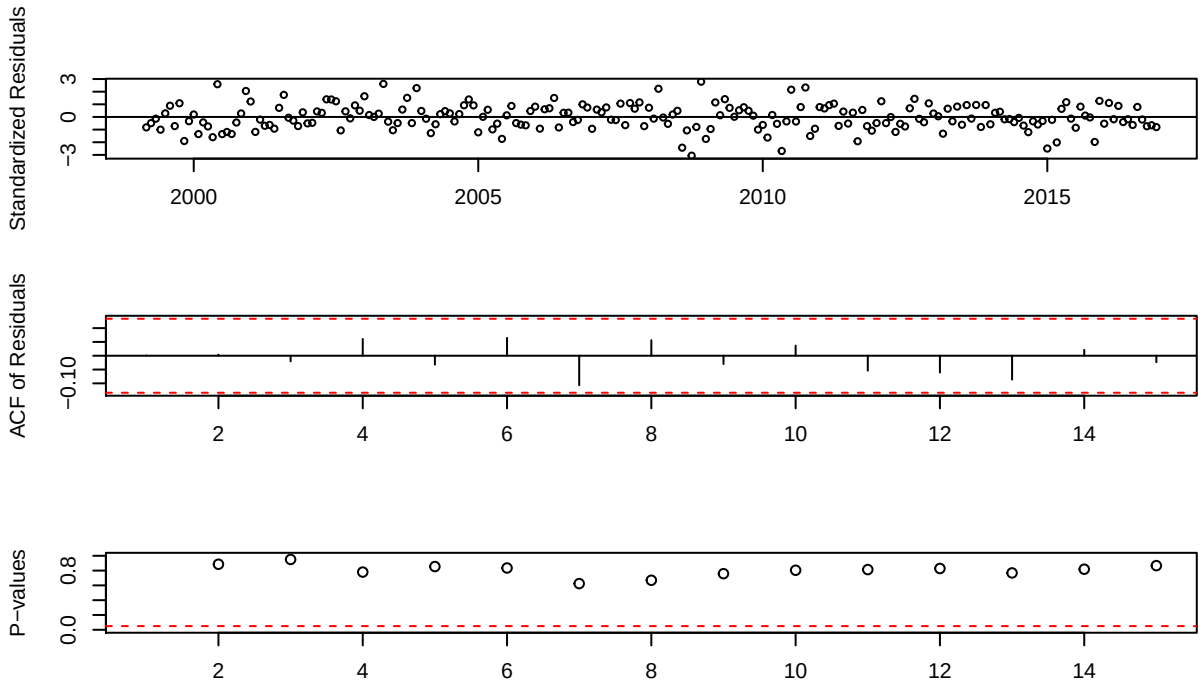


Figure 10. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle IMA(1,1) pour le logarithme du taux de change US/Euro.

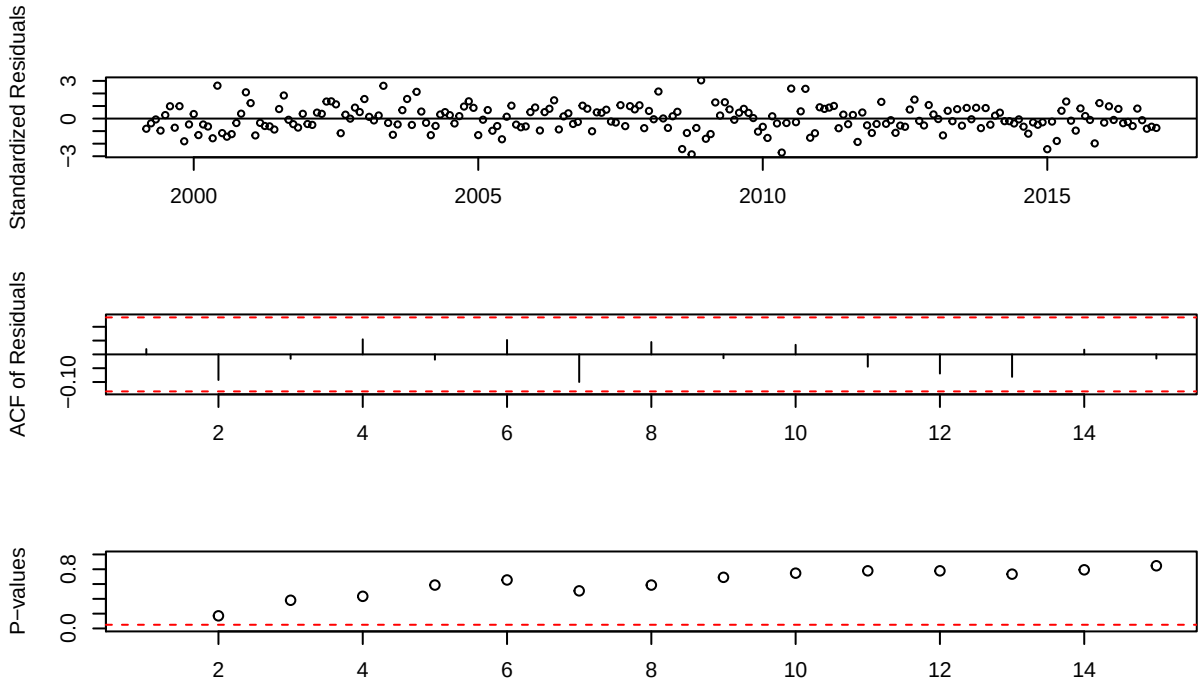


Figure 11. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle ARI(1,1) pour le logarithme du taux de change US/Euro.

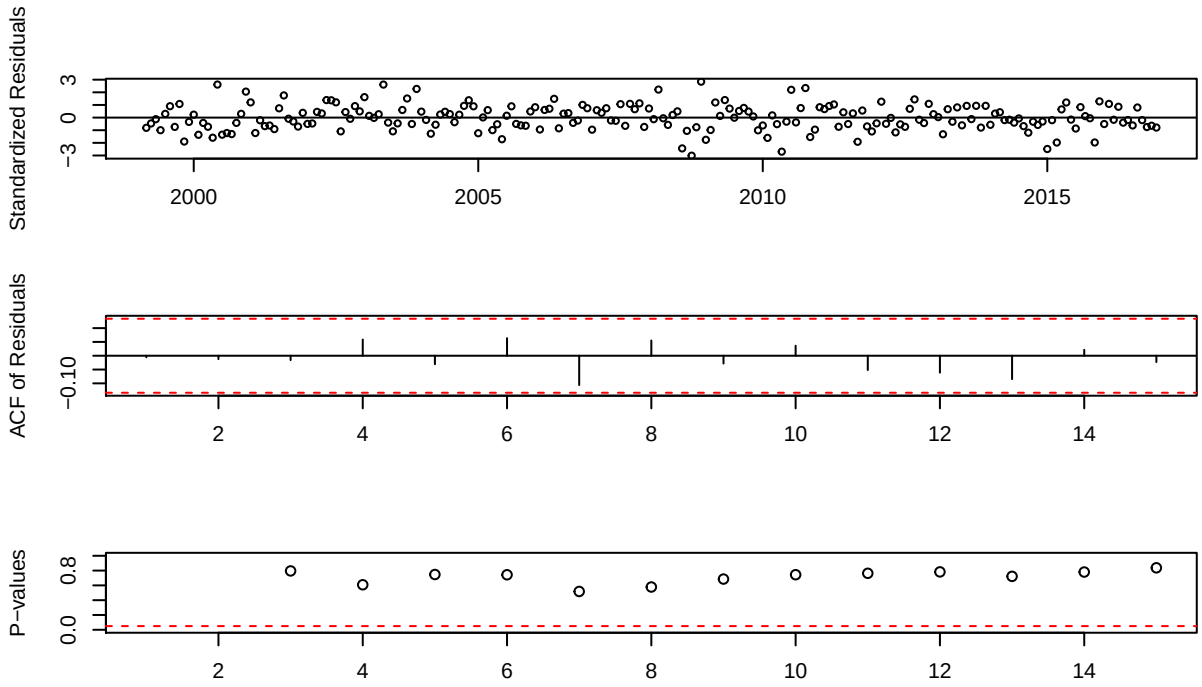


Figure 12. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle ARIMA(1,1,1) pour le logarithme du taux de change US/Euro.

On se trouve ainsi dans l'impossibilité de départager nos modèles à l'aide de l'étude de leurs résidus en plus de voir tous ces modèles être jugés appropriés selon le test de Ljung-Box. On choisit alors de conserver le modèle IMA(1,1) parce qu'il est simple, parce que la valeur maximale de sa fonction de vraisemblance est la plus élevée et parce que la valeur du test AIC est la plus faible selon la *Table III*. De plus, ce modèle nous permet de retrouver un modèle de loi lognormale pour modéliser le taux US/Euro qui est un modèle fortement utilisé dans le monde de la mathématique financière.

On regarde, pour finaliser le choix de modèle, la question de l'*overfiting*. Comme le modèle ARIMA(1,1,1) a déjà été testé, il suffit ici d'observer le comportement du modèle IMA(1,2). On remarque de la *Table III* que le paramètre θ_1 du modèle IMA(1,2) est très près du paramètre θ du modèle IMA(1,1) en plus d'avoir un paramètre θ_2 avoisinant 0. Sachant que $\hat{\theta}_2 \approx 0.0207$ et que son écart-type est de 0.0645, l'intervalle de confiance de $\hat{\theta}_2$ est $IC = \hat{\theta}_2 \pm Z_{1-\alpha/2}\sqrt{Var(\hat{\theta}_2)}$ = [-0.1057, 0.1471]. Puisque 0 est dans l'intervalle de confiance, on peut négliger ce paramètre et on peut alors écarter ce modèle. Le modèle IMA(1,1) est donc celui avec lequel les prédictions seront effectuées.

Prédictions

On peut alors effectuer les prédictions des taux de changes US/Euro des mois de l'année 2017 grâce au modèle suivant

$$\nabla \ln(Y_t) = e_t - 0.3118 e_{t-1}$$

Il est important de noter que, puisqu'une transformation logarithmique a été effectuée sur les données du taux de change US/Euro, le prédicteur de Y_t, $\hat{Y}_t(l)$, sera tel que

$$\hat{Y}_t(l) = \exp\left(\hat{Z}_t(l) + \frac{\text{Var}[e_t(l)]}{2}\right)$$

Où $\hat{Z}_t(l)$ est le prédicteur de Z_t défini tel que $Z_t = \ln Y_t$.

Le graphique des valeurs prédites se trouvent aux graphiques *Figure 13* et *Figure 14*. On remarque d'abord du second graphique que toutes les valeurs du taux de change US/Euro observées en 2017 se trouvent à l'intérieur de l'intervalle de confiance de nos prédictions. L'intervalle de confiance est fait à un niveau de confiance de 95%.

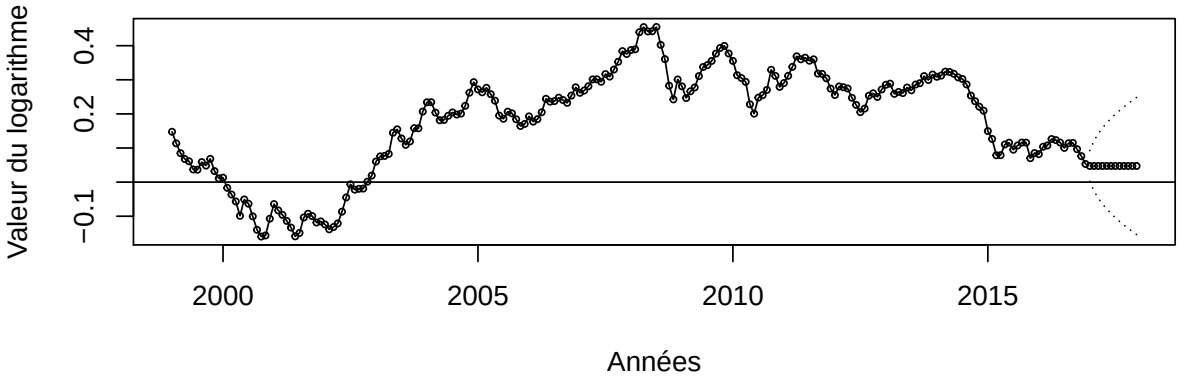


Figure 13. Logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2016 avec prédictions pour l'année 2017 et intervalle de confiance à 95%.

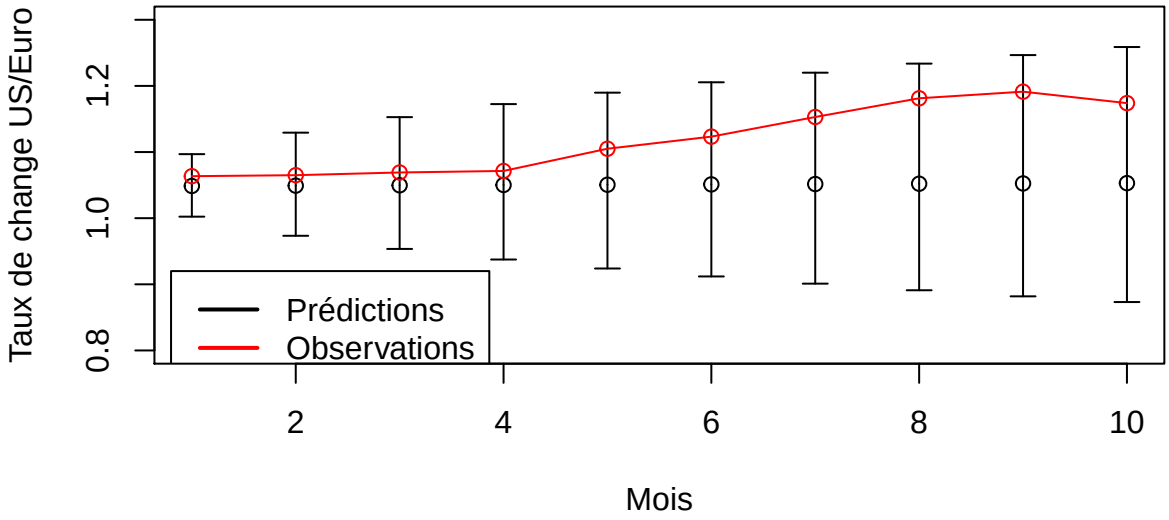


Figure 14. Comparaison entre les valeurs prédites du taux de change US/Euro grâce au modèle IMA(1,1) avec intervalle de confiance et les valeurs observerées pour l'année 2017.

On retrouve également les valeurs des prédictions du taux de change US/Euro pour les 10 premiers mois de l'année 2017 dans le tableau suivant

Mois	Y_{t+l}	$\hat{Y}_t(l)$	$\hat{e}_t(l)$	$\sqrt{Var(\hat{Y}_t(l))}$	Borne inf. de prédiction	Borne sup. de prédiction
Janvier	1.0635	1.0487	0.0148	0.023	1.0023	1.0968
Février	1.065	1.0492	0.0158	0.0379	0.9734	1.1293
Mars	1.0691	1.0497	0.0194	0.0484	0.9535	1.1528
Avril	1.0714	1.0502	0.0212	0.057	0.9375	1.1725
Mai	1.105	1.0506	0.0544	0.0645	0.9239	1.1898
Juin	1.1233	1.0511	0.0722	0.0712	0.9119	1.2055
Juillet	1.153	1.0516	0.1014	0.0773	0.901	1.22
Août	1.1813	1.0521	0.1292	0.083	0.891	1.2337
Septembre	1.1913	1.0525	0.1388	0.0883	0.8818	1.2466
Octobre	1.174	1.053	0.121	0.0933	0.8732	1.2588

où $\hat{e}_t(l) = Y_{t+l} - \hat{Y}_t(l)$ représente l'erreur de prédiction.

Conclusion

En observant le graphique et la valeur de nos prédictions, on observe que celles-ci semblent converger vers une valeur bien précise. Cependant, plus on tente de prédire pour des périodes éloignées, plus notre intervalle de confiance grandit et plus notre incertitude devient élevée.

SAAQ

Analyse primaire du jeu de données

Tout d'abord, on fait l'analyse du jeu de données en lien avec la SAAQ. On étudie les valeurs de nos quatres séries chronologiques depuis 1978 jusqu'en 2015. Cela comprend la série du nombre d'accidents avec dommages corporels (NAADC), du nombre de personnes accidentés (NPA), du nombre de demandes d'indemnités (NDI), et du coût total de l'indemnisation en millions de dollars, et en dollars constants 2015 (CTI), le tout sur une base annuelle. Puisque la banque de données semble comporter certaines valeurs aberrantes, on applique des changements à ces données grâce à la commande *tsoutliers*. Les nouvelles séries chronologiques sont affichées ci-dessous dans la *Figure 15*

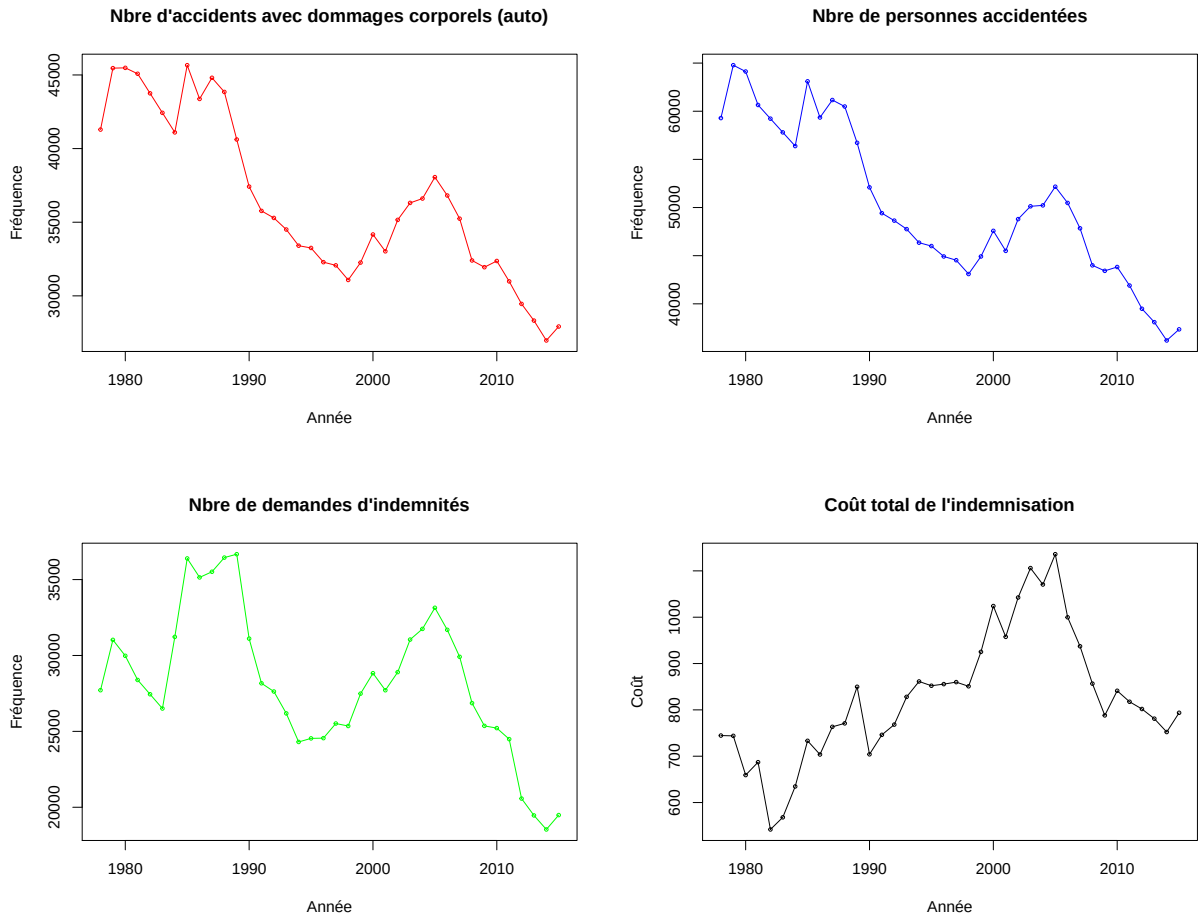


Figure 15. Graphiques des différentes séries chronologiques après modification des données aberrantes depuis 1978 jusqu'en 2015 (NAADC en rouge, NPA en bleu, NDI en vert et CTI en noir).

On remarque prinicipalement deux choses de ces graphiques. D'abord, les trois premières séries chronologiques semblent très fortement corrélées. Ce constat aurait également pu découler de la définition même de ces variables qui sont logiquement, très fortement corrélées.

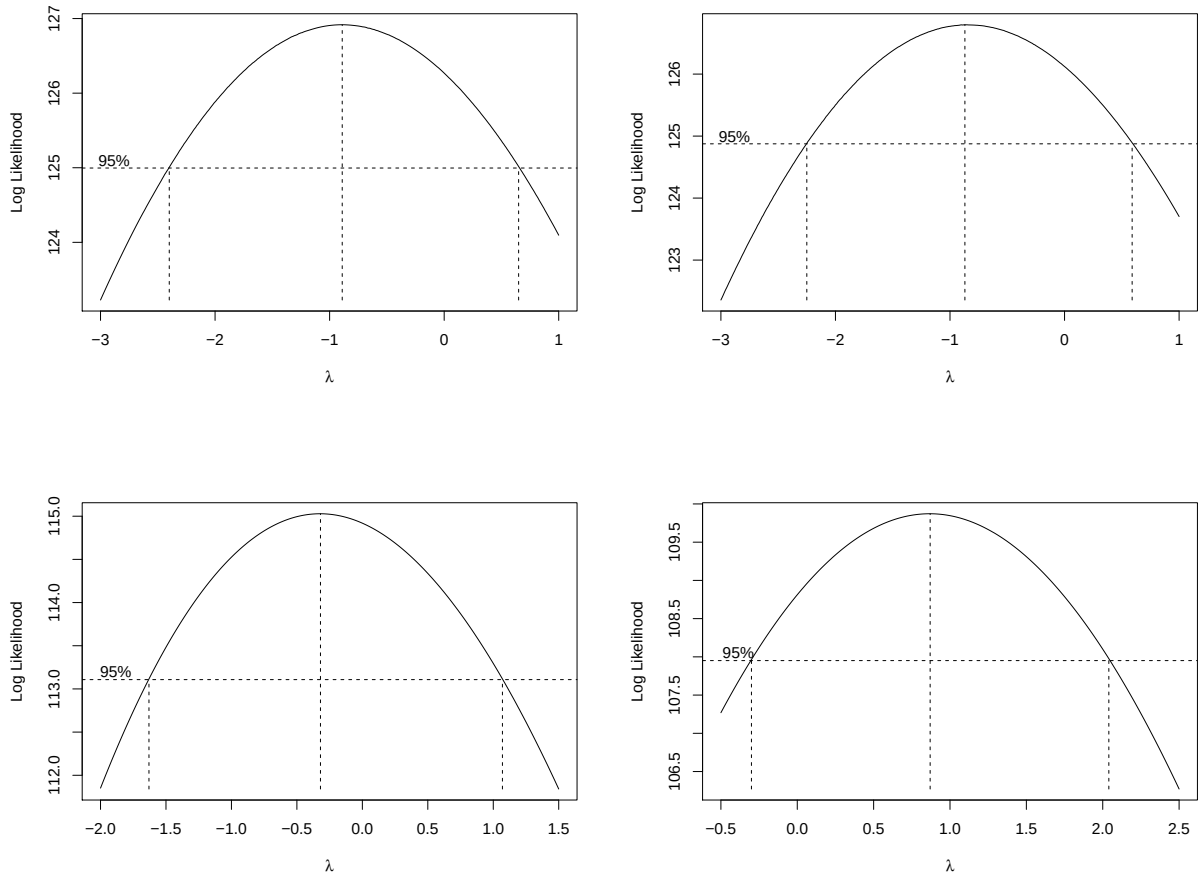
Test de stationnarité

L'analyse débute encore une fois par les tests de stationnarités. On commence ainsi par effectuer le test de Dickey-Fuller sur notre modèle pour tester ça stationnarité. Le test a été expliqué au numéro précédent. Nous obtenons les résultats suivants

Série chrono.	coeff. k dans AR(k)	p-value	Stationnarité
NAADC	2	0.4802	Non
NPA	2	0.499	Non
NDI	2	0.5238	Non
CTI	1	0.8737	Non

Étant donné que chacune des séries chronologiques a une p-value $> 5\%$, on ne peut pas rejeter l'hypothèse nulle du test qui stipule que la série est non-stationnaire. Une première différenciation s'impose. Cependant, il faudrait d'abord observer si une transformation est approprié pour stabiliser la variance. On effectue ainsi la transformation de Box-Cox.

Transformation des données



On peut donc observer grâce au maximum de vraisemblance que les valeurs de λ sont approximativement -1, -1, 0 et 1 pour les séries NAADC, NPA, NDI et CTI respectivement. On note qu'il n'y a donc pas de

transformation a appliquer à la série CTI.

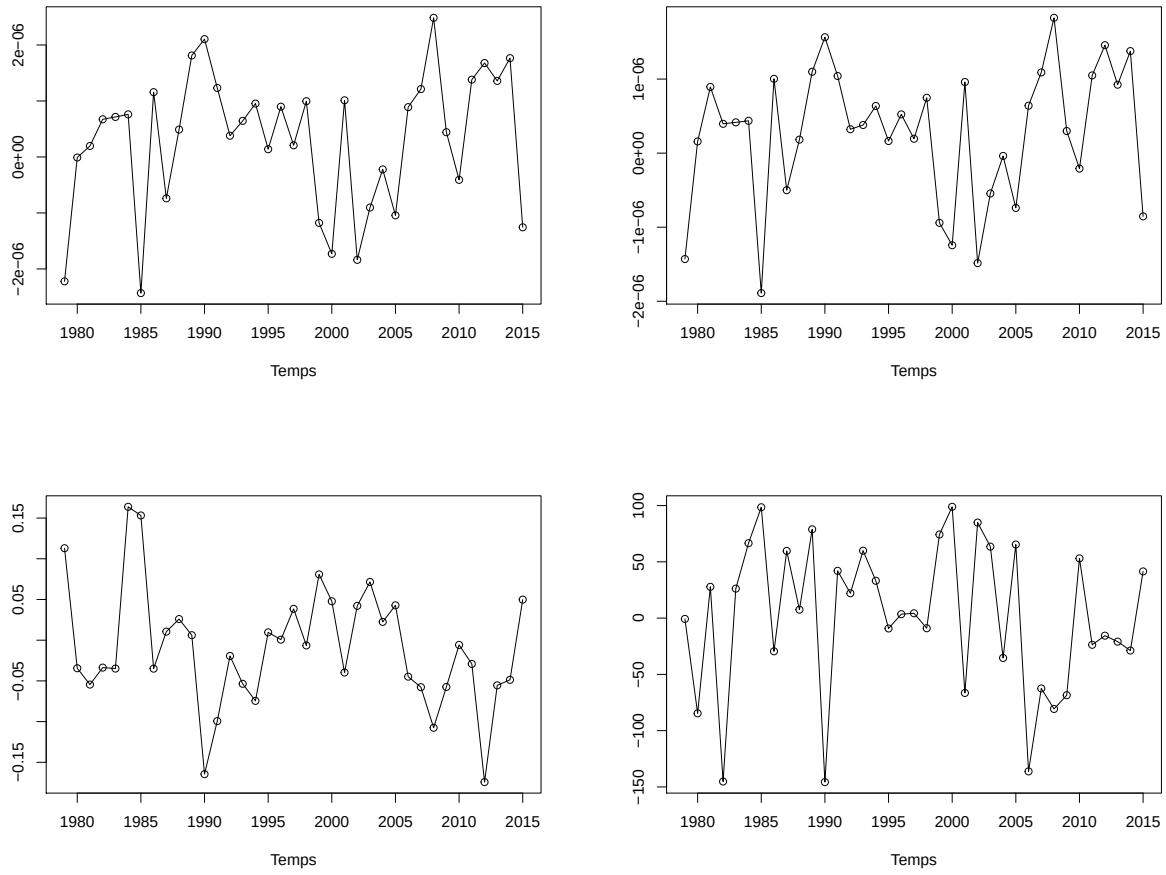
On peut alors, tester la stationnarité de la première différenciation des données transformées encore à l'aide du test de Dikey-Fuller. Ce test renvoie cette fois les valeurs suivantes

Série chrono.	coeff. k dans AR(k)	p-value	Stationnarité
NAADC	0	0.01	Oui
NPA	0	0.01	Oui
NDI	1	0.0323	Oui
CTI	0	0.01	Oui

On remarque cette fois que la p-value de chaque modèle est inférieure à 5%, ce qui nous permet d'affirmer la stationnarité de ceux-ci. On peut maintenant passer à l'étape du choix du modèle.

Identification du modèle

On affiche les graphiques de nos nouvelles séries chronologiques ci-dessous après avoir apporté la transformation et la différenciation à chacun



On poursuit en analysant les fonctions d'autocorrélation de chaque série transformée tel qui suit

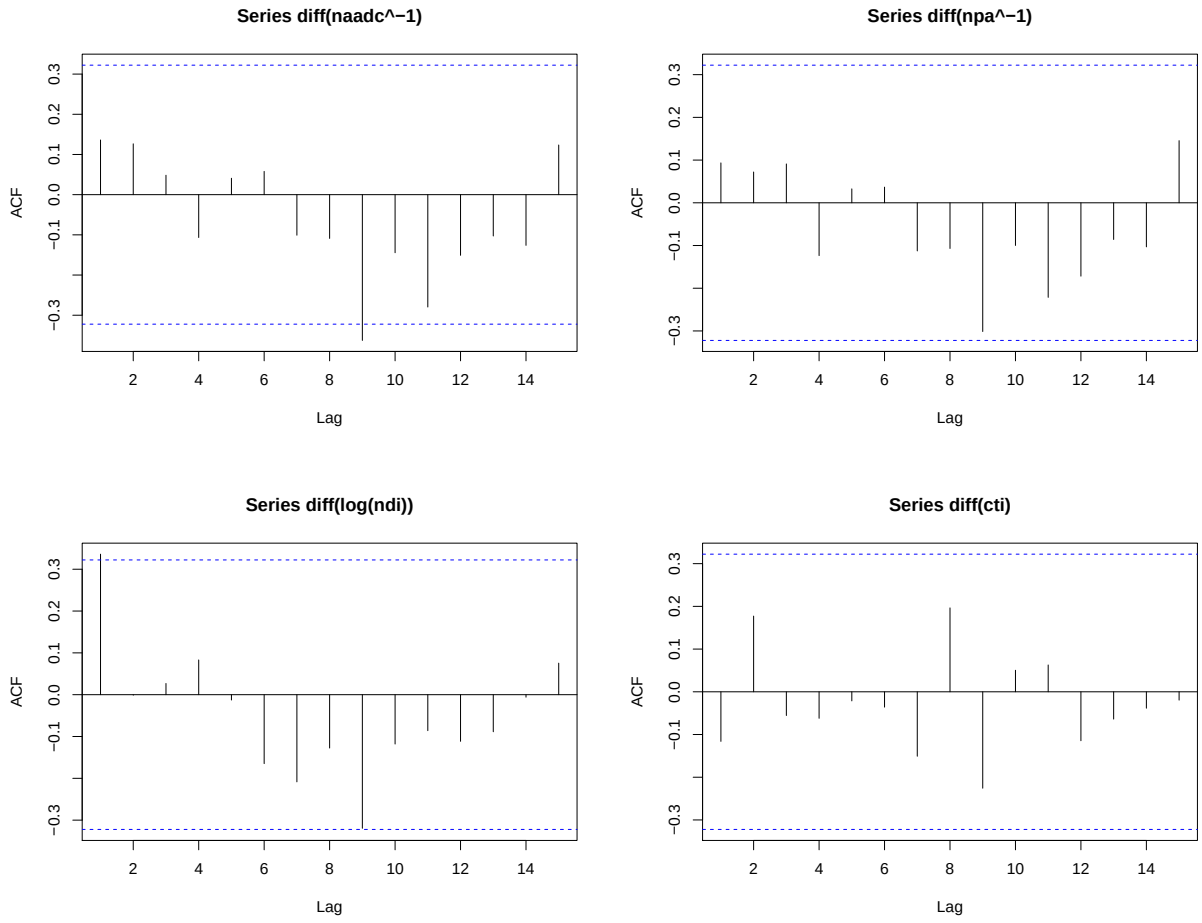
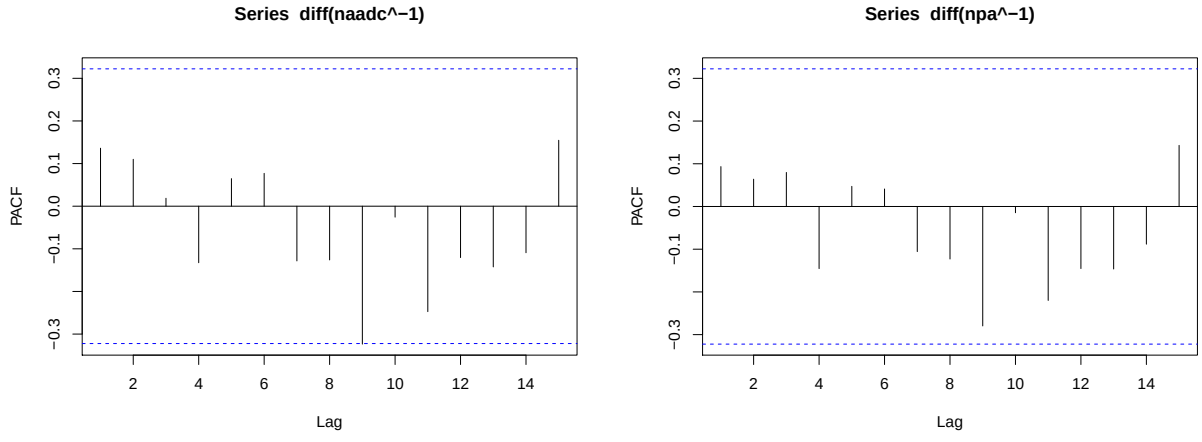
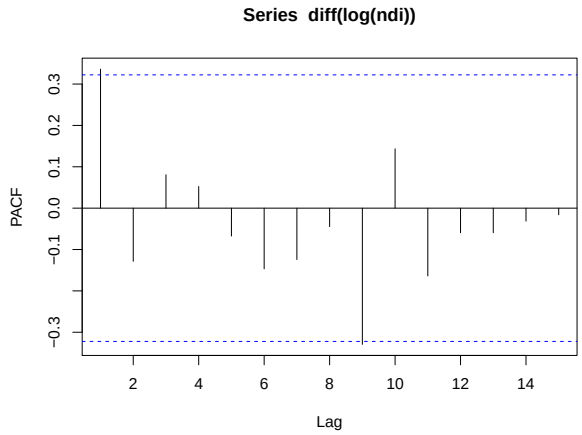


FIGURE WHATEVER ACF

Les graphiques ci-dessus suggère l'absence de coefficent de moyenne mobile dans notre série. Qu'en est-il de l'autocorrélation partielle?





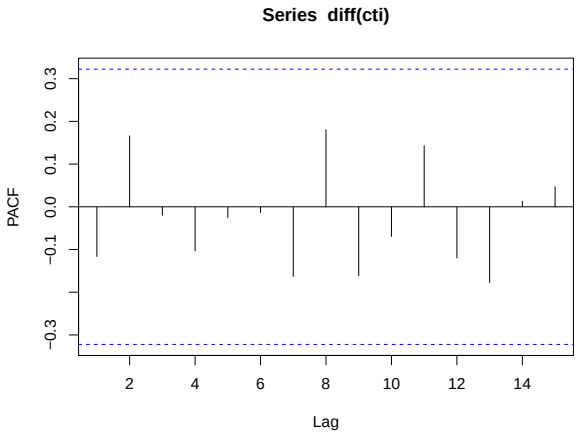


FIGURE WHATEVER PACF

Les graphiques de l'autocorrélation partielle semblent écarter la présence d'autorégression dans la différentiation de nos modèles. L'analyse des graphiques de l'autocorrélation et de l'autocorrélation partielle suggère donc un bruit blanc comme modèle à considérer pour chacune des séries.

On effectue un autre test pour ajuster un modèle aux séries chronologiques présentement à l'étude, le test de l'autocorrélation étendu. On obtient les tables qui suivent

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10
## 0 o o o o o o o o o o o
## 1 x o o o o o o o o o o
## 2 o o o o o o o o o o o
## 3 o x o o o o o o o o o
## 4 o x o o o o o o o o o
## 5 x o o o o o o o o o o
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10
## 0 o o o o o o o o o o o
## 1 x o o o o o o o o o o
## 2 x o o o o o o o o o o
## 3 x x o o o o o o o o o
## 4 o o o o o o o o o o o
## 5 x o o o o o o o o o o
```

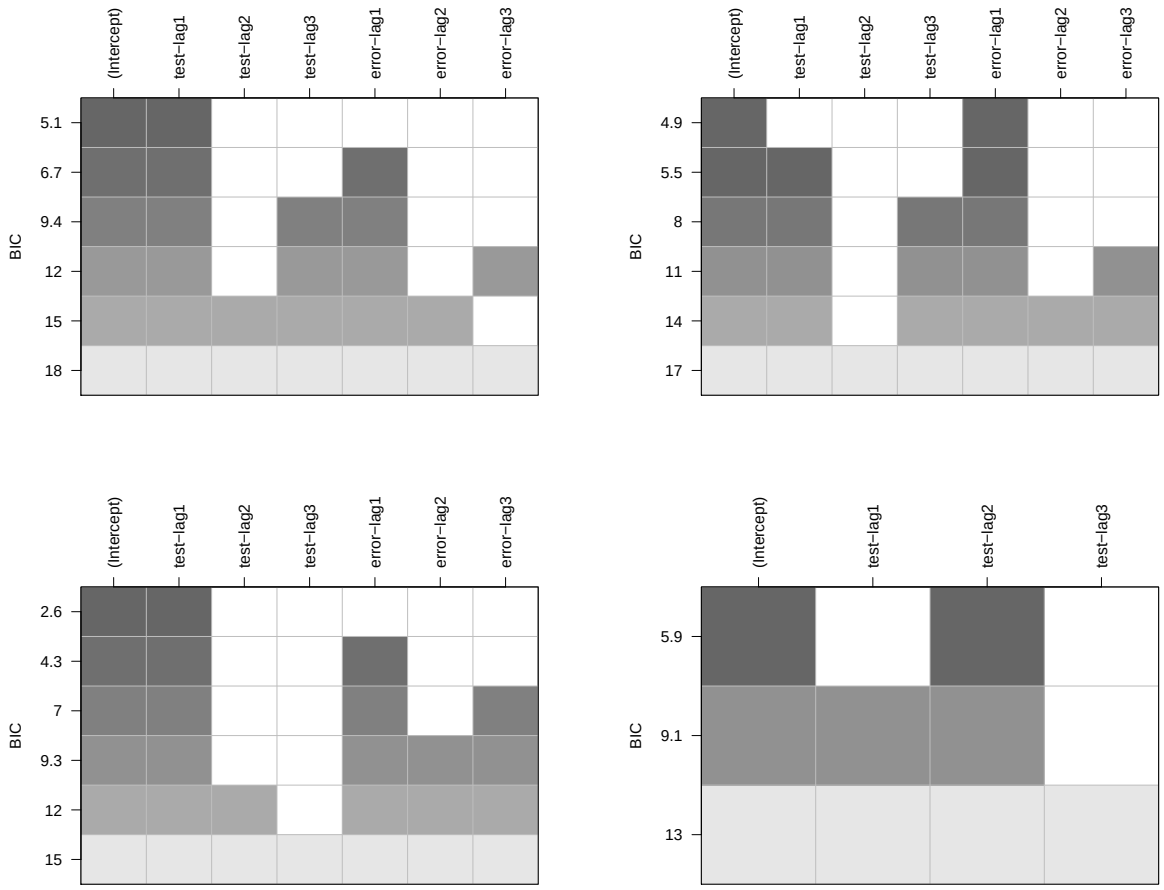
```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10
## 0 x o o o o o o o o o o
## 1 x o o o o o o o o o o
## 2 x o o o o o o o o o o
## 3 o x o o o o o o o o o
## 4 o o o o o o o o o o o
## 5 x o o o o o o o o o o
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10
## 0 o o o o o o o o o o o
## 1 x o o o o o o o o o o
## 2 o o o o o o o o o o o
## 3 o o o o o o o o o o o
## 4 o o o o o o o o o o o
## 5 x o o o o o o o o o o
```

FIGURE WHATEVER EACF

Les graphiques provenant des tests d'autocorrélation étendue nous proposent de façon respective une ARMA(0,0), ARMA(0,0), ARMA(0,1) et ARMA(0,0) pour les modèles transformés et différenciés NAADC, NPA, NDI et CTI.

Il ne reste plus qu'à effectuer le critère du BIC pour conclure quels modèles seront retenus.



Les graphiques provenant du critère BIC nous proposent de façon respective une AR(1), MA(1), AR(1) et MA(3) (avec θ_1 et θ_2 égal à 0) pour les modèles transformés et différenciés NAADC, NPA, NDI et CTI. Il est pertinent de mentionner que le BIC tente de renvoyer le modèle ARMA(p,q) qui s'applique le mieux à la série sans jamais proposer le bruit blanc.

Les modèles testés seront donc les suivants:

— NDAAC

$$\nabla Y_t^{-1} - \mu = e_t \tag{1}$$
$$\nabla Y_t^{-1} - \mu = e_t + \phi\left(\nabla Y_{t-1}^{-1} - \mu\right) \tag{2}$$

— NPA

$$\nabla Y_t^{-1} - \mu = e_t \tag{3}$$
$$\nabla Y_t^{-1} - \mu = e_t - \theta e_{t-1} \tag{4}$$

— NDI

$$\nabla lnY_t - \mu = e_t \tag{5}$$
$$\nabla lnY_t - \mu = e_t - \theta e_{t-1} \tag{6}$$
$$\nabla lnY_t - \mu = e_t + \phi\left(\nabla lnY_{t-1} - \mu\right) \tag{7}$$

— CTI

$$\nabla Y_t - \mu = e_t \tag{8}$$
$$\nabla Y_t - \mu = e_t - \theta e_{t-3} \tag{9}$$

Estimation des paramètres des différents modèles et tests de résidus

Il faut maintenant estimer les paramètres de nos différents modèles. Les tableaux ci-dessous fournissent l'information nécessaire

Modèle pour NAADC	Paramètres	Valeurs estimés	AIC	Logarithme du max de vraisemblance
ARIMA(0,1,0)	μ	0	-901	451.6
ARIMA(1,1,0)	μ	0	-900	452
	ϕ	0.1592		

Modèle pour NPA	Paramètres	Valeurs estimés	AIC	Logarithme du max de vraisemblance
ARIMA(0,1,0)	μ	0	-923	462.5
ARIMA(0,0,1)	μ	0	-921	462.7
	θ	0.0954		

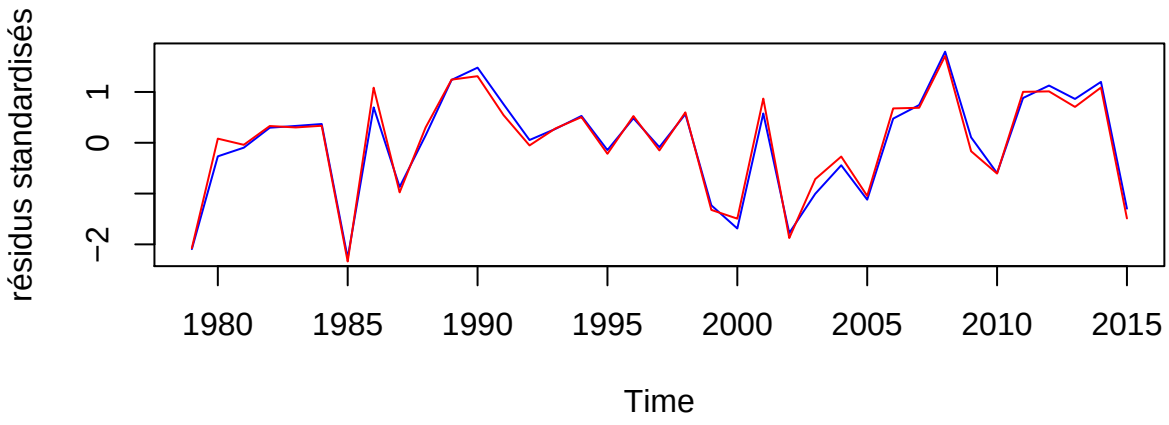
Modèle pour NDI	Paramètres	Valeurs estimés	AIC	Logarithme du max de vraisemblance
ARIMA(0,1,0)	μ	-0.0095	-87	44.6
ARIMA(0,1,1)	μ	-0.0071	-91	47.6
	θ	0.4391		
ARIMA(1,1,0)	μ	-0.0068	-90	47
	ϕ	0.3622		

Modèle pour CTI	Paramètres	Valeurs estimés	AIC	Logarithme du max de vraisemblance
ARIMA(0,1,0)	μ	0	417	-207.3
ARIMA(0,0,3)	μ	1.4595101	419	-207.3
	θ	-0.0624432		

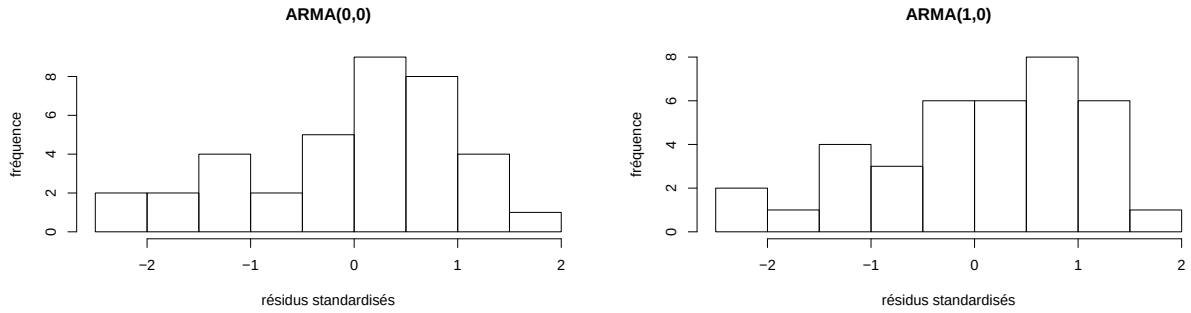
On peut maintenant faire l'analyse de nos résidus pour faire notre choix de modèle. Puisque nous avons plusieurs modèles par série chronologique, nous allons séparer les graphiques en 4 sections, i.e. pour les 4 modèles transformés différenciés.

NAADC

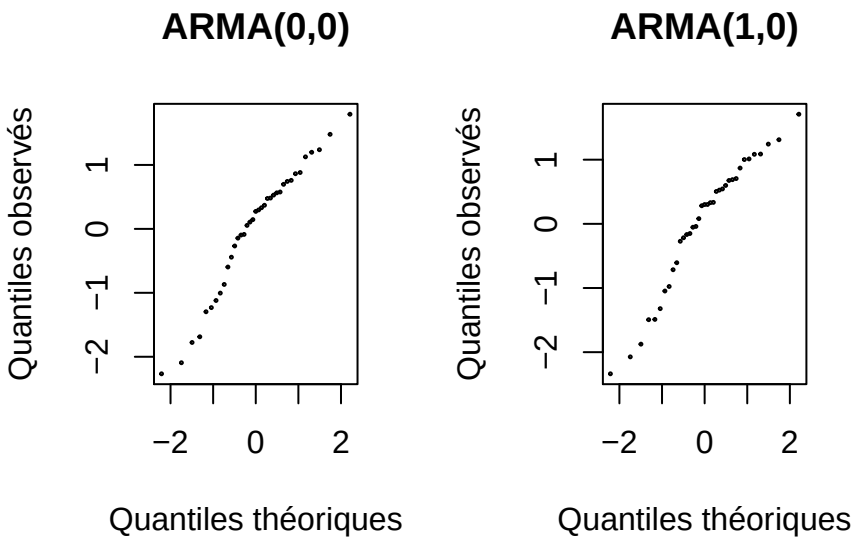
Les résidus standardisés des deux modèles que nous avons gardés apparaîssent dans le graphique ci-dessous. On remarque que les courbes sont plutôt superposées. Puisque les valeurs des résidus standardisés sont généralement comprises dans une fourchette de ± 2, l'hypothèse de normalité semble adéquate.



On regarde maintenant les histogrammes des résidus. Les deux histogrammes ne semblent pas tracer la cloche de la loi normale tel que nous la connaissons.

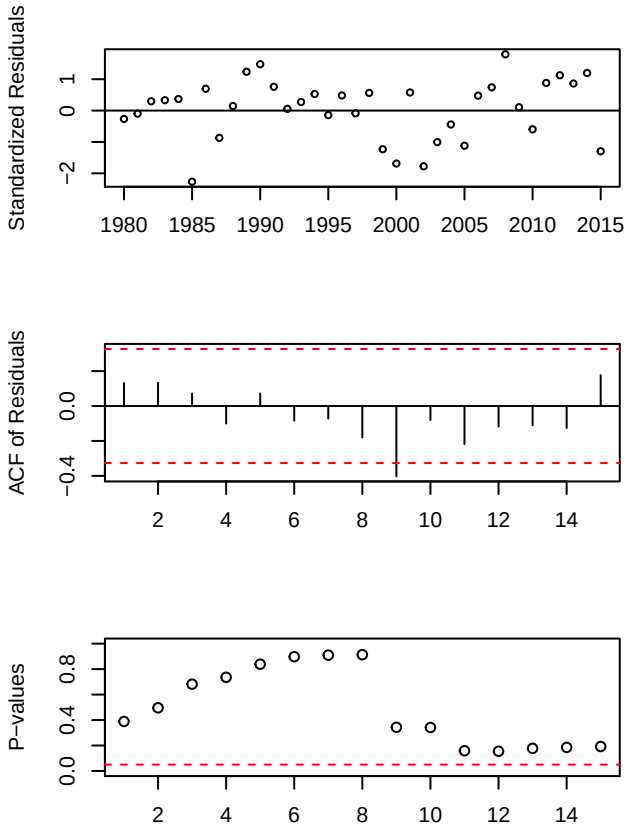


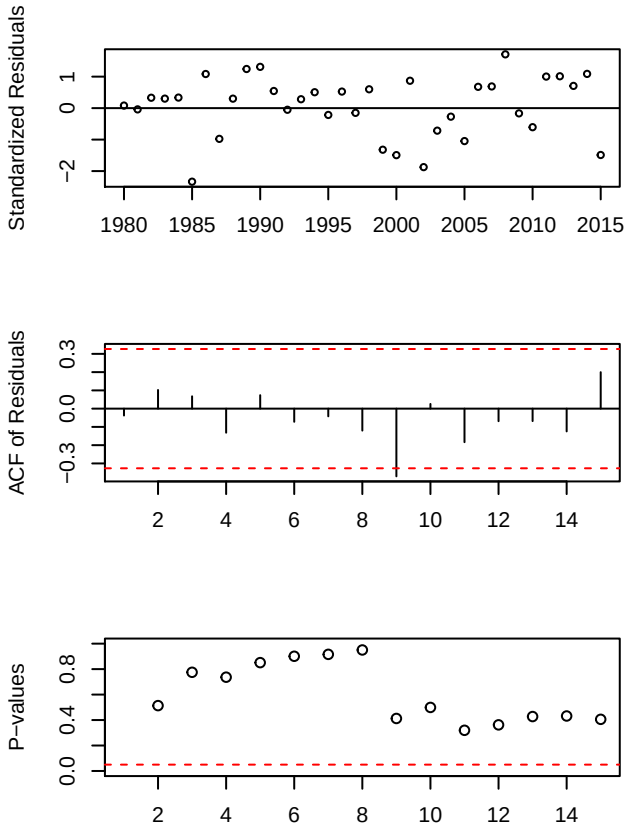
On poursuit notre étude des résidus avec le graphique QQ. On remarque que les résidus observés ne semblent pas être similaire à leur valeur théorique en cas de normalité.



On effectue maintenant le test de Shapiro-Wilks pour départager nos modèles. Les p-values respectives du modèle ARMA(0,0) et ARMA(1,0) sont 0.1381 et 0.0627. Puisque les p-values sont supérieures à 5%, nous gardons l'hypothèse nulle qui stipule que les résidus suivent une loi normale.

Par la suite, il faut effectuer le run-test pour déterminer si les résidus sont indépendants. Les p-values de nos modèles ARMA(0,0) et ARMA(1,0) sont 0.642 et 0.651. Puisqu'elles sont nettement supérieures à 5%, on accepte l'hypothèse nulle qui stipule l'indépendance entre les résidus.


Standardized Residuals
1
0
−2
1980
1985
1990
1995
2000
2005
2010
2015
ACF of Residuals
0.0
−0.4
2
4
6
8
10
12
14
P−values
0.8
0.4
0.0
2
4
6
8
10
12
14



Ljung-Box nous dit pour les modèles ARMA(0,0) et ARMA(1,0) que leur p-value sont respectivement de 0.2618 et 0.4991. Puisqu'elle sont supérieures à 5%, les modèles sont appropriés.

Puisque les résidus des 2 modèles semblent suivre une normale, que les log de max de vraisemblance sont identiques tout comme les AIC, on choisit de conserver le modèle ARMA(0,0), i.e. le bruit blanc parce qu'il est simple et que les graphiques ACF et PACF supportent cette hypothèse.

On regarde maintenant la question de la surparamétrisation. Dans notre cas, puisqu'on a déjà testé le modèle ARMA(1,0), il ne reste qu'à tester le modèle ARMA(0,1). Ce modèle semble adéquat avec un paramètre ϕ = 0.128972 et une moyenne μ = 2.988855e-07. Cependant, on reste avec notre hypothèse de départ et on suppose que le modèle final du NAADC est un bruit blanc, i.e. $\nabla Y_t^{-1} = e_t$.

NPA

On cherche alors à savoir si les résidus suivent des lois normales.

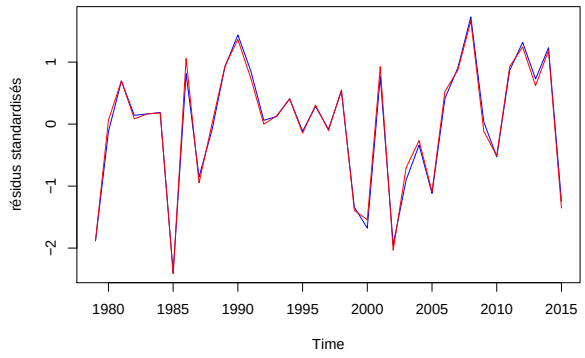


Figure ?? Superposition des graphiques des résidus standardisés des modèles ARIMA(0,1,0) et ARIMA(0,1,1) du data NPA.

On remarque que les résidus standardisés sont inférieurs à 2 en valeurs absolues ce qui est souhaitable pour des lois normales centrées réduites.

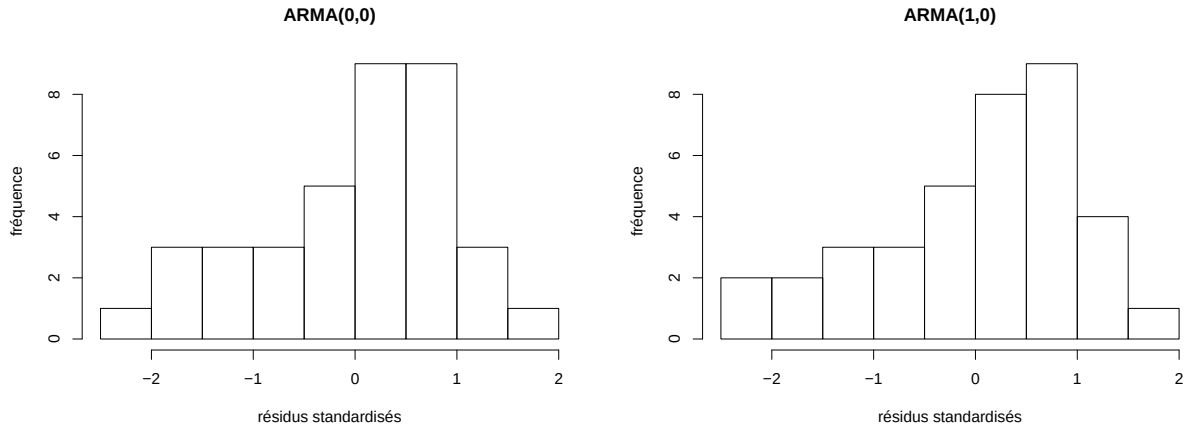


Figure ?? Histogramme des résudies standardisés des modèles ARIMA(0,1,0) et ARIMA(0,1,1) du data NPA.

Les histogrammes des deux modèles à l'étude semblent difficilement tracer de loi normal tel qu'on remarque du graphique ci-dessus.

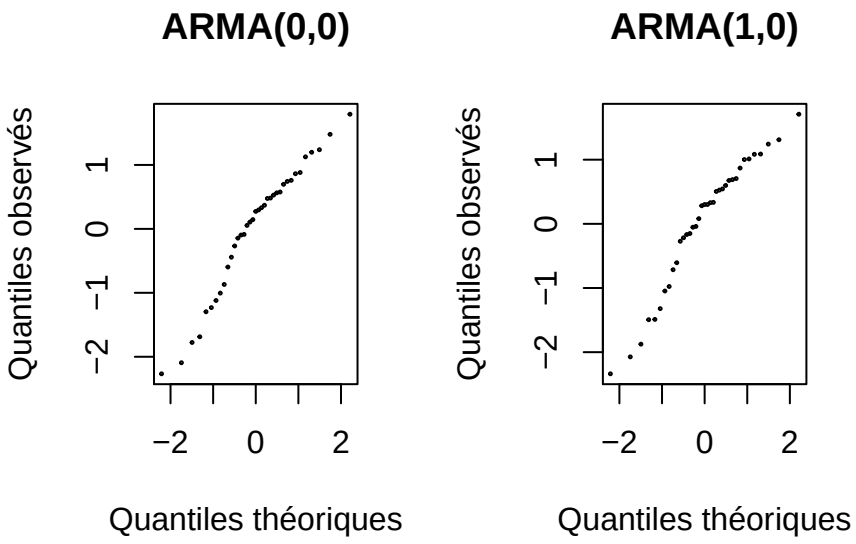
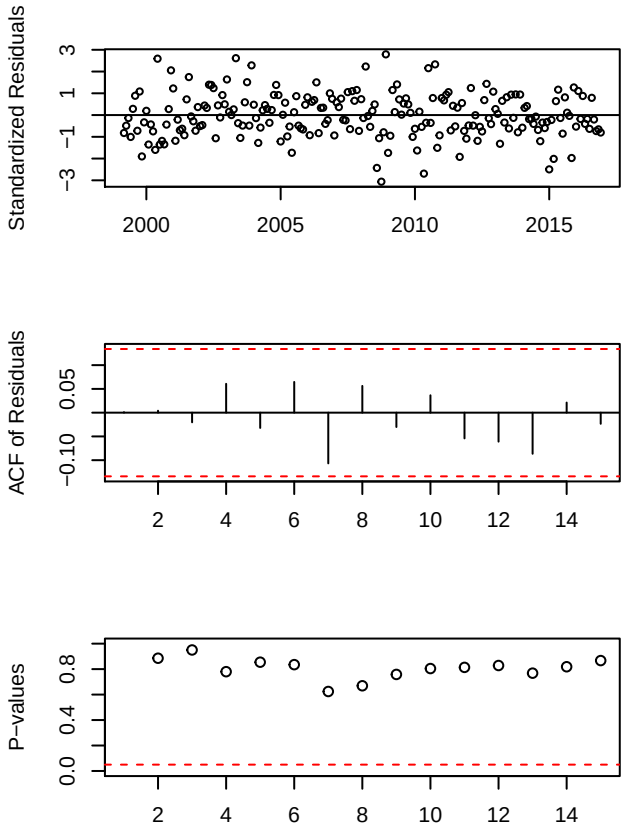


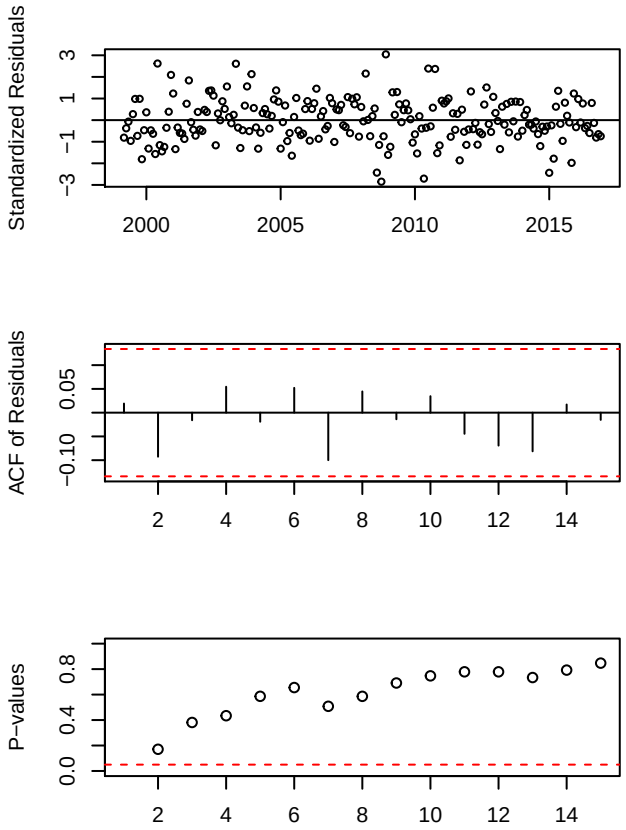
Figure ?? graphique QQ des modèles ARIMA(0,1,0) et ARIMA(0,1,1)

Les graphiques QQ ne semblent pas tracer de belle droite ce qui fait douter de la fiabilité de l'hypothèse de normalité.

Le test de Shapiro-Wilk ne permet toutefois pas de rejetter l'hyptohèse nulle de noramlité des résidus, ni dans le cas ARIMA(0,1,0), ni dans le cas du ARIMA(0,1,1) avec des p-values respectivent de 0.12 et 0.082.

On ne peut ici rejetter l'hypothèse nulle d'indépendance des résidus du run-test avec des p-values respectives de0.642 et 0.903 pour les mêmes modèles que ceux testés précédemment.


Standardized Residuals
3
1
−1
−3
2000
2005
2010
2015
ACF of Residuals
0.05
−0.10
2
4
6
8
10
12
14
P−values
0.8
0.4
0.0
2
4
6
8
10
12
14



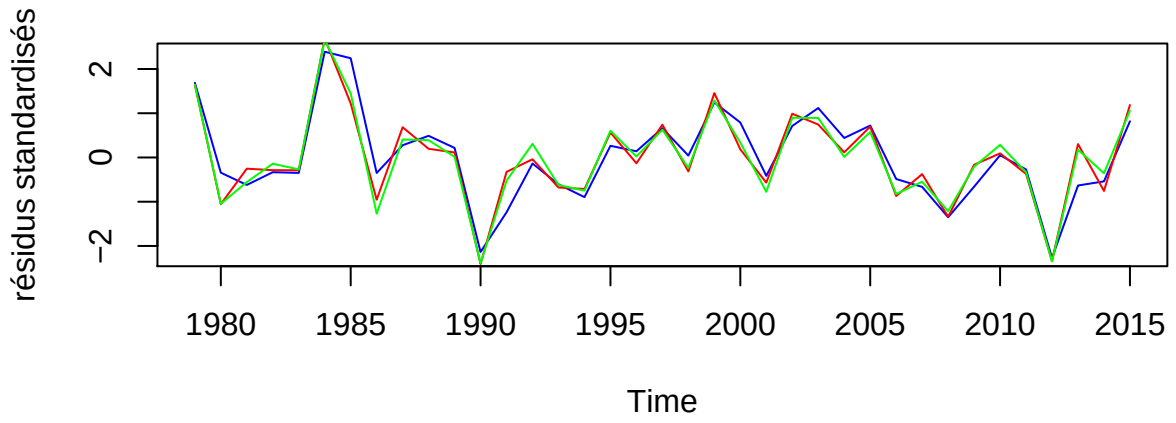
Vient finalement le test de Ljung–Box qui a pour hypothèse nul que le modèle testé est approprié. On obtient, pour les modèles ARIMA(0,1,0) et ARIMA(0,0,1), les p-values respectives 0.525 et 0.637 qui viennent toutes deux justifiées leur modèles respectifs à un niveau de confiance arbitraire de 5%.

Puisque les deux modèles engendrent les mêmes conclusion pour tous les tests, on se fit aux graphiques de l'ACF et du PACF pour retenir le bruit blanc comme modèle approprié. On se penche alors sur la question de l'overfiting. Comme le modèle ARIMA(0,1,1) a déjà été testé, on se penche uniquement sur le cas du modèeles ARIMA(1,1,0) pour l'overfiting.

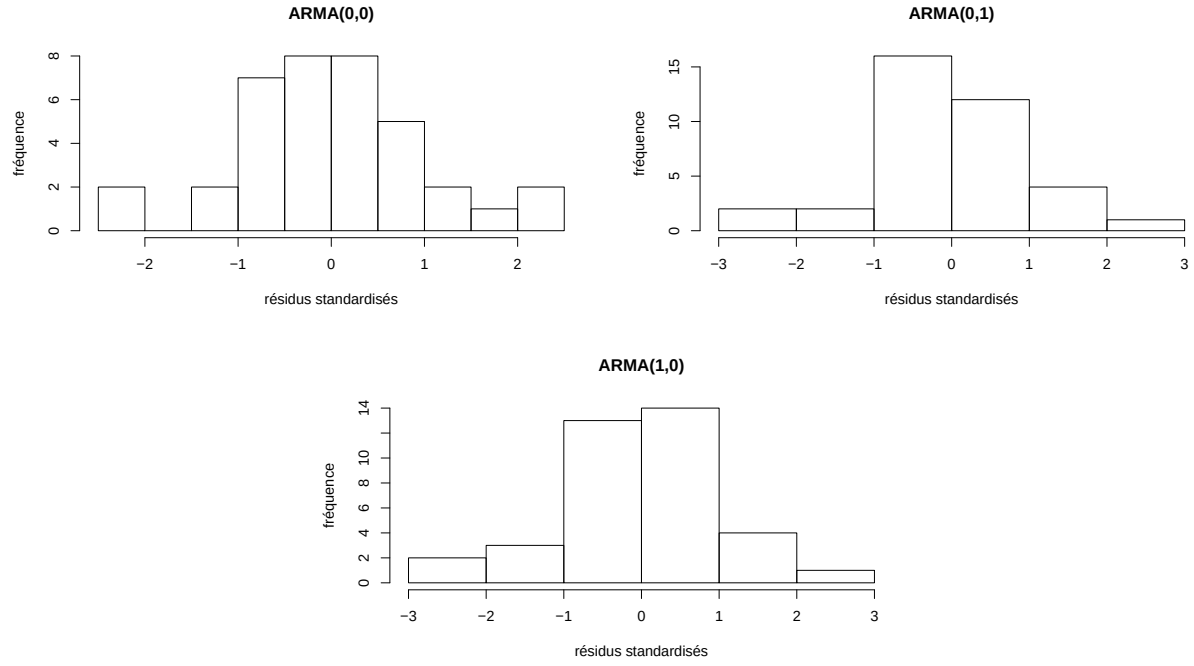
La sortie R de la fonction arima pour le modèle ARIMA(1,1,0) renvoit un coefficient autorégressif de 0.1592 ce qui ne tend pas significativement vers 0. Ce modèle est donc à considérer. Toutefois, le test de Shapiro-Wilks pour la normalité des résidus de ce modèle renvoit un valeur de p-value de 6.003%, qui est concluant pour un test à niveau de confiance de 5%, mais pas très fortement. Comme la normalité des résidus du modèle est plus concluante pour le modèele ARIMA(0,1,0) et que les graphiques de l'ACF et du PACF suggèrent très fortement ce modèles, c'est celui qui sera retenu.

NDI

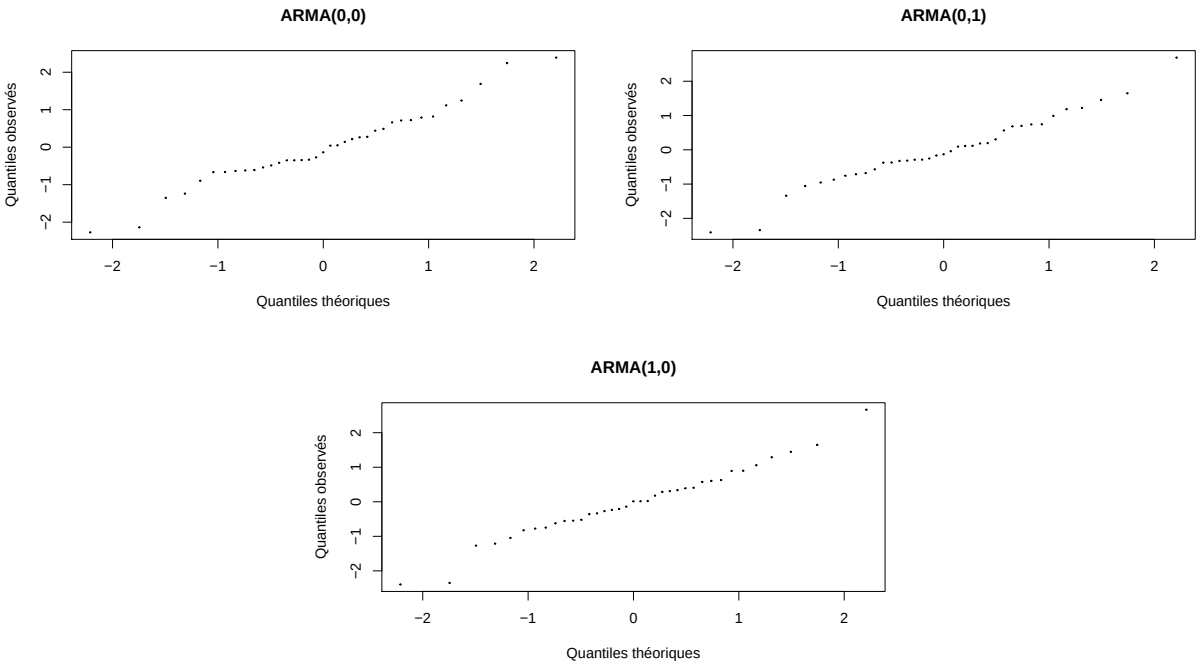
Les résidus standardisés des trois modèles que nous avons gardés apparaîssent dans le graphique ci-dessous. On remarque que les courbes sont plutôt superposées. Les valeurs des résidus standardisés sont à quelques reprises à l'extérieur d'une fourchette de ± 2, l'hypothèse de normalité semble plus ou moins convaincante.



On regarde maintenant les histogrammes des résidus. Les trois histogrammes semblent tracer la cloche de la loi normale assez bien.

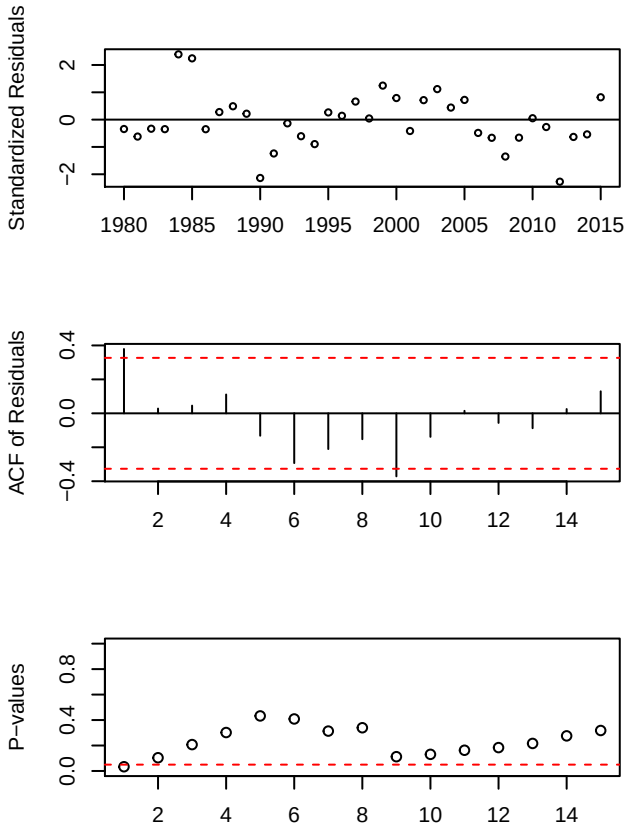


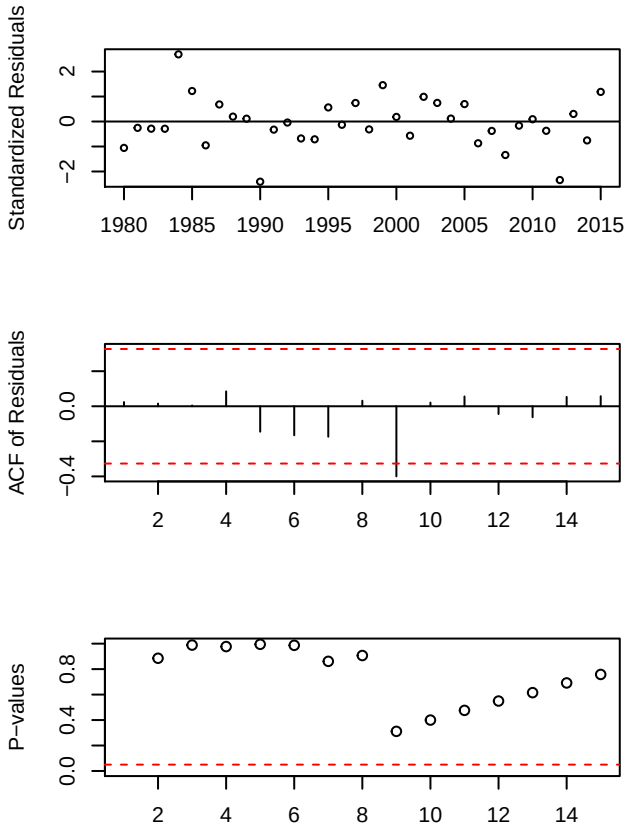
On poursuit notre étude des résidus avec le graphique QQ. On remarque que les résidus observés sont sensiblement similaire à leur valeur théorique en cas de normalité.

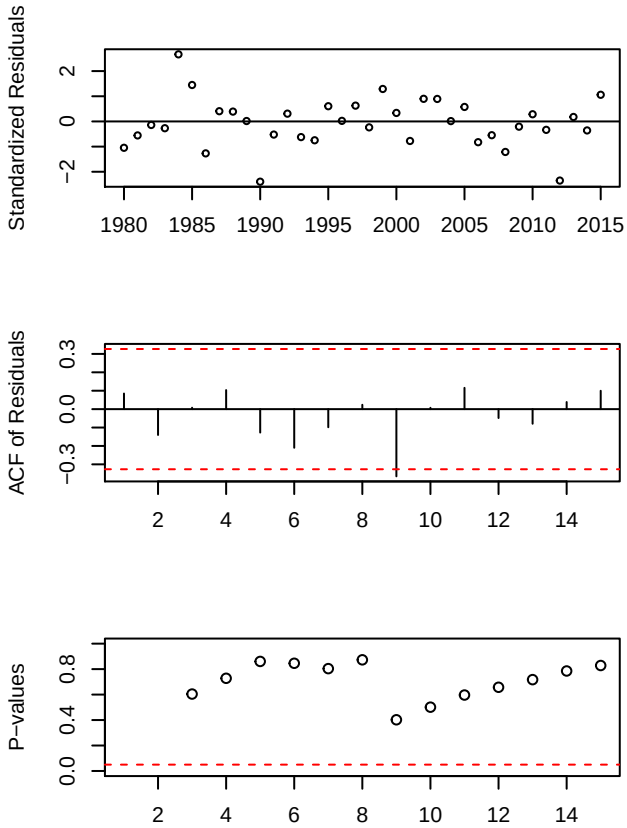


On effectue maintenant le test de Shapiro-Wilks pour départager nos modèles. Les p-values respectives du modèle ARMA(0,0), ARMA(0,1) et ARMA(1,0) sont 0.4395, 0.5029 et 0.0627. Puisque les p-values sont supérieures à 5%, nous gardons l'hypothèse nulle qui stipule que les résidus suivent une loi normale. Cependant, on remarque que le modèle ARMA(1,0) est tout prêt de refuser cette hypothèse. On est en droit de mettre en doute la force du test.

Par la suite, il faut effectuer le run-test pour déterminer si les résidus sont indépendants. Les p-values de nos modèles ARMA(0,0), ARMA(0,1) et ARMA(1,0) sont 0.0437, 0.971 et 1. On remarque pour le modèle du bruit blanc (ARMA(0,0)) que la p-value est inférieure à 5%. Dans ce cas, on rejette l'hypothèse que les résidus sont indépendants. Cependant, pour le modèle autorégressif et moyenne mobile, puisque les p-values sont nettement supérieures à 5%, on accepte l'hypothèse nulle qui stipule l'indépendance entre les résidus.


Standardized Residuals
2
0
−2
1980
1985
1990
1995
2000
2005
2010
2015
ACF of Residuals
0.4
0.0
−0.4
2
4
6
8
10
12
14
P−values
0.8
0.4
0.0
2
4
6
8
10
12
14


Standardized Residuals
−2
0
2
1980
1985
1990
1995
2000
2005
2010
2015
ACF of Residuals
−0.4
0.0
2
4
6
8
10
12
14
P−values
0.0
0.4
0.8
2
4
6
8
10
12
14



Ljung-Box nous dit pour les modèles ARMA(0,0), ARMA(0,1) et ARMA(1,0) que leur p-value sont respectivement de 0.0899, 0.3999 et 0.5012. Puisqu'elle sont supérieures à 5%, les modèles sont appropriés.

Nous avons donc écarté l'hypothèse de modèle basé uniquement sur le bruit blanc. De plus, puisque le graphique ACF semble montrer une forme autorégressive, et que le BIC supporte cette hypothèse, et que l'histogramme des résidus du ARMA(1,0) est plus représentatif d'une loi normale que l'ARMA(0,1), nous décidons de garder le modèle ARMA(1,0).

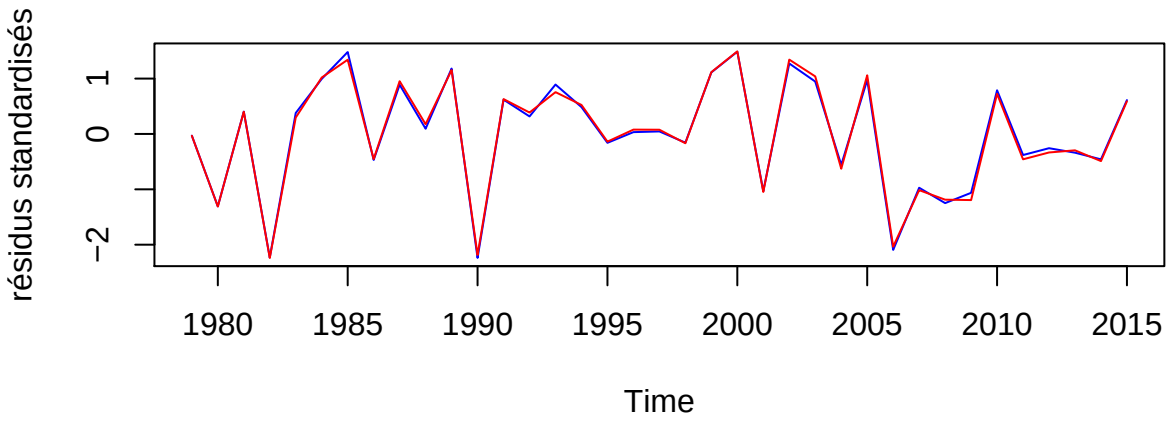
On regarde maintenant la question de la surparamétrisation. On teste le modèle ARMA(2,0) et ARMA(1,1).

Pour le modèle ARMA(2,0), les coefficients ϕ_1 et ϕ_2 sont respectivement de 0.4441 et -0.1627. Puisque le deuxième paramètre du modèle ARMA(2,0) n'est pas significativement différent de 0 (selon les intervalles de confiance), il est jugé de refuser le modèle ARMA(2,0) et garder notre modèle initial.

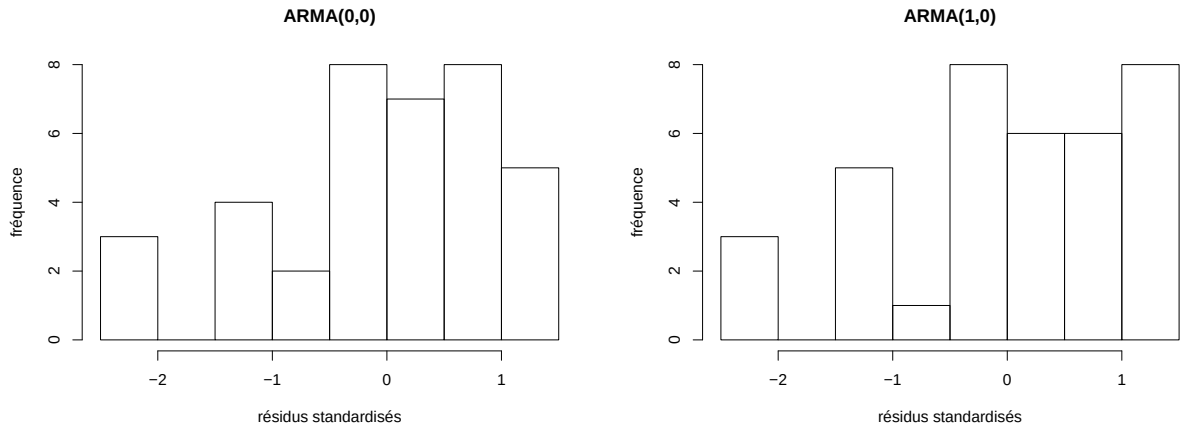
Pour ce qui est du modèle ARMA(1,1), la partie autorégressive ne semble pas convaincante puisque le coefficient ϕ tend vers 0. Il penche donc plus vers une moyenne mobile MA(1) qui est un des modèles que nous avions rejeté initialement. Donc, nous préférerons garder une fois de plus le modèle ARMA(1,0).

CTI

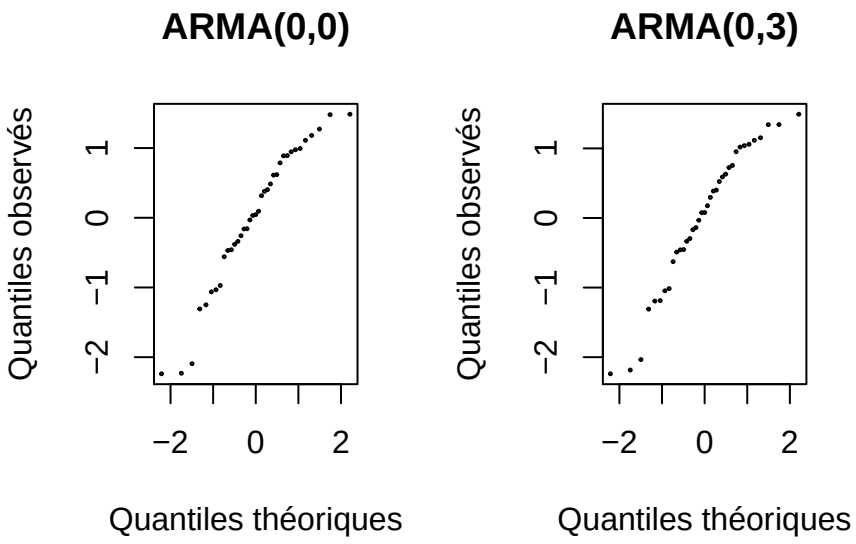
Les résidus standardisés des deux modèles que nous avons gardés apparaîssent dans le graphique ci-dessous. On remarque que les courbes sont plutôt superposées. Puisque les valeurs des résidus standardisés sont généralement comprises dans une fourchette de ± 2, l'hypothèse de normalité semble adéquate.



On regarde maintenant les histogrammes des résidus. Les deux histogrammes ne semblent pas tracer la cloche de la loi normale tel que nous la connaissons.

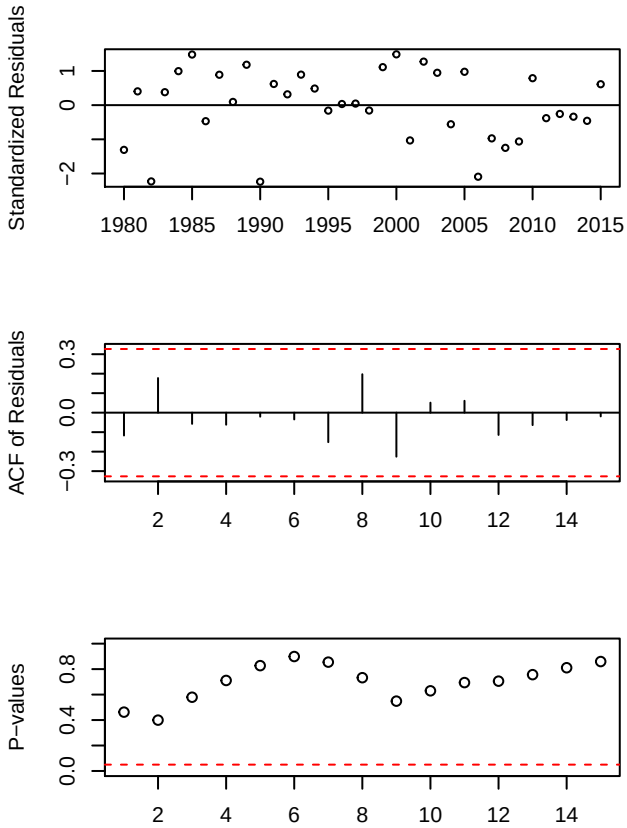


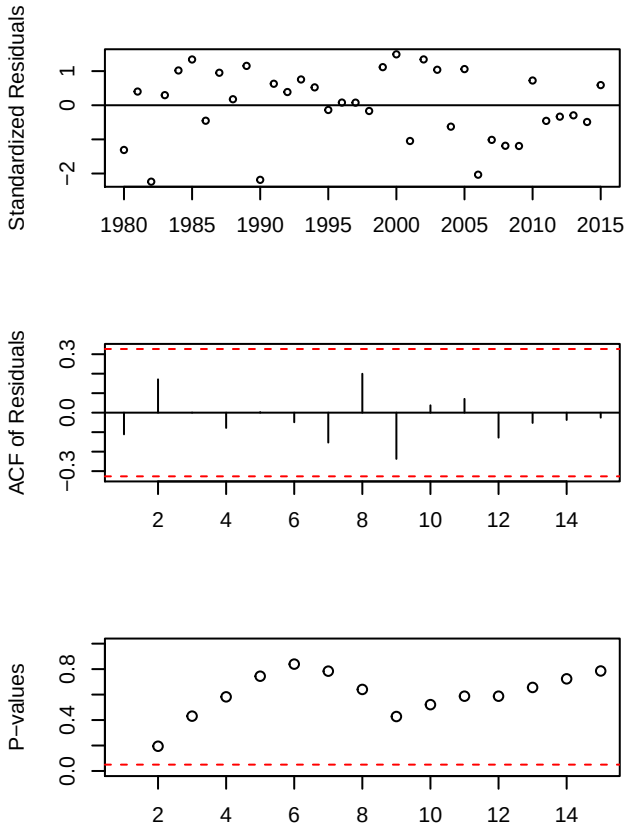
On poursuit notre étude des résidus avec le graphique QQ. On remarque que les résidus observés ne semblent pas être similaire à leur valeur théorique en cas de normalité.



On effectue maintenant le test de Shapiro-Wilks pour départager nos modèles. Les p-values respectives du modèle ARMA(0,0) et ARMA(0,3) sont 0.0739 et 0.0808. Puisque les p-values sont supérieures à 5%, nous gardons l'hypothèse nulle qui stipule que les résidus suivent une loi normale.

Par la suite, il faut effectuer le run-test pour déterminer si les résidus sont indépendants. Les p-values de nos modèles ARMA(0,0) et ARMA(0,3) sont 0.971 et 0.971. Puisqu'elles sont nettement supérieures à 5%, on accepte l'hypothèse nulle qui stipule l'indépendance entre les résidus.


Standardized Residuals
1
0
−2
1980
1985
1990
1995
2000
2005
2010
2015
ACF of Residuals
0.3
0.0
−0.3
2
4
6
8
10
12
14
P−values
0.8
0.4
0.0
2
4
6
8
10
12
14



Ljung-Box nous dit pour les modèles ARMA(0,0) et ARMA(0,3) que leur p-value sont respectivement de 0.5349 et 0.521. Puisqu'elle sont supérieures à 5%, les modèles sont appropriés.

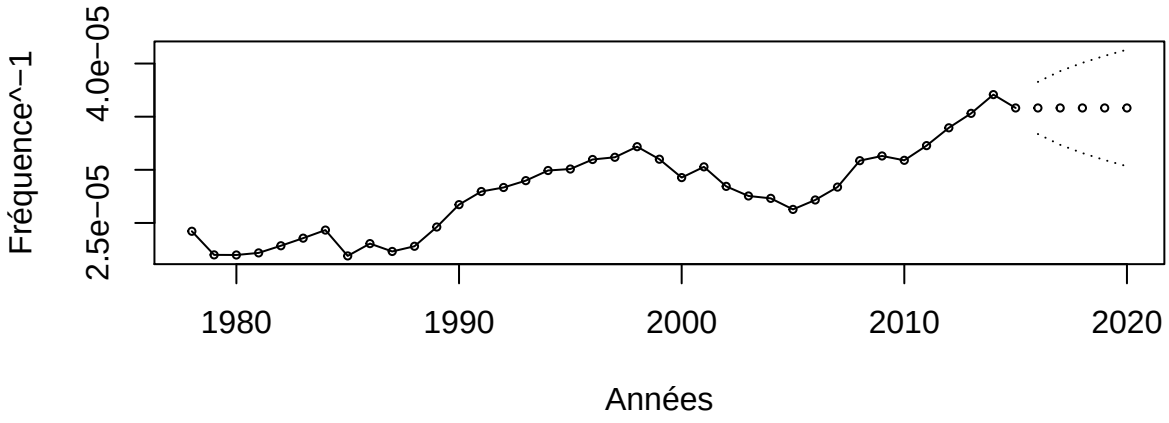
Puisque les résidus des 2 modèles semblent suivre une normale, que les log de max de vraisemblance sont identiques tout comme les AIC, on choisit de conserver le modèle ARMA(0,0), i.e. le bruit blanc parce qu'il est simple et que les graphiques ACF et PACF supportent cette hypothèse.

On regarde maintenant la question de la surparamétrisation. Dans notre cas, on doit tester le modèle ARMA(1,0) et ARMA(0,1). Les p-values respectives du modèle ARMA(1,0) et ARMA(0,1) sont 0.1537 et 0.1361. Les résidus suivent bien une loi normale, mais ça reste que les graphiques ACF et PACF appuient le fait qu'il serait préférable de garder le modèle du bruit blanc, i.e. $\nabla Y_t = e_t$.

Prédictions

NAADC

On rappelle que notre modèle final est $\nabla Y_t^{-1} = e_t$. Les prédictions ainsi que les bornes de nos prédictions du modèle NAADC modifiés et différés sont affichés ci-dessous



Donc, les valeurs sont dans le tableau suivant

Année	Y_{t+l}	Borne inf.	Borne sup.
2016	27917	2.6128322×10^4	2.9968572×10^4
2017	27917	2.5452824×10^4	3.0909451×10^4
2018	27917	2.4957719×10^4	3.1672461×10^4
2019	27917	2.4555048×10^4	3.2345594×10^4
2020	27917	2.4210903×10^4	3.2962798×10^4

NPA

On rappelle que notre modèle final est $\nabla lnY_t = e_t + \phi \nabla lnY_{t-1}$. Les prédictions ainsi que les bornes de nos prédictions du modèle NPA modifiés et différés sont affichés ci-dessous

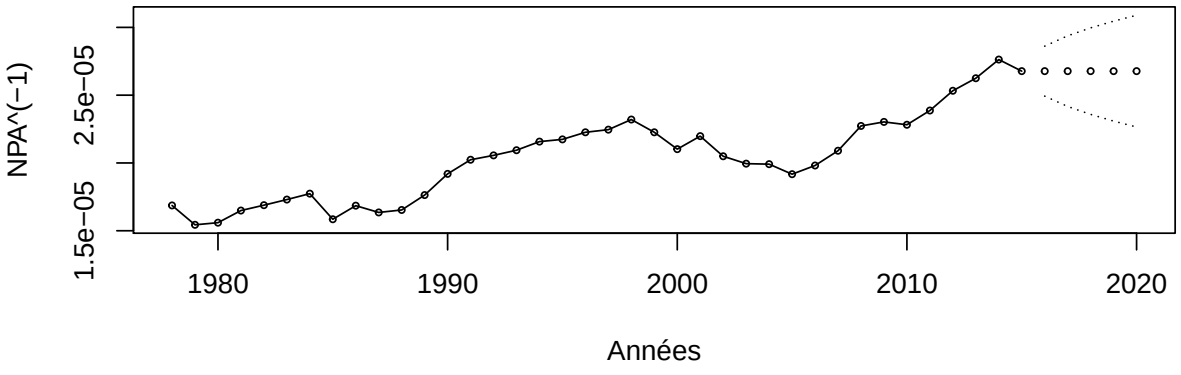


Figure ?? Projection sur 5 ans du nombre d'accidents avec personnes accidentés (NPA).

On remarque que l'intervalle de confiance augmentes avec le temps ce qui est dû à la variance croissante. On obtient ainsi les prédictions suivantes.

Donc, les valeurs sont dans le tableau suivant

Année	Y_{t+l}	Borne inf.	Borne sup.
2016	37351	3.4945419×10^4	4.0112257×10^4
2017	37351	3.4037393×10^4	4.1379362×10^4
2018	37351	3.3372011×10^4	4.2407275×10^4
2019	37351	3.2830951×10^4	4.3314369×10^4
2020	37351	3.23686×10^4	4.4146307×10^4

NDI

On rappelle que notre modèle final est $\nabla Y_t^{-1} = e_t$. Les prédictions ainsi que les bornes de nos prédictions du modèle NPA modifiés et différés sont affichés ci-dessous

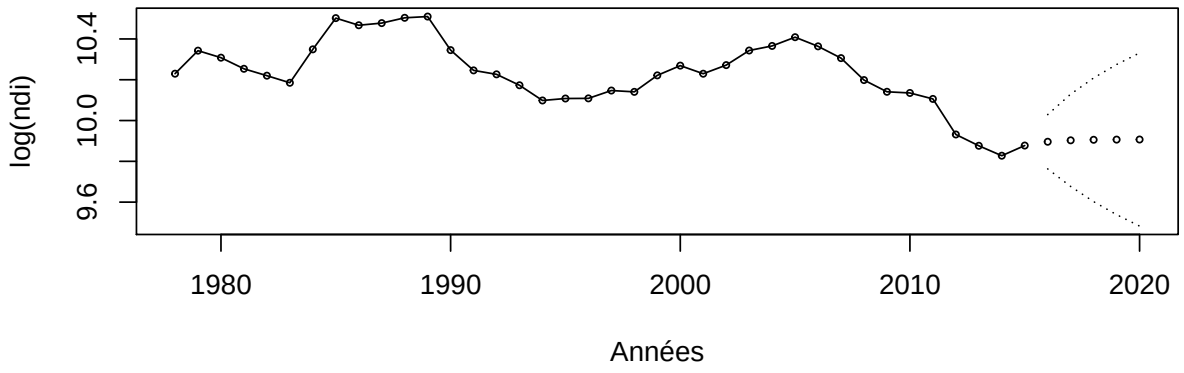


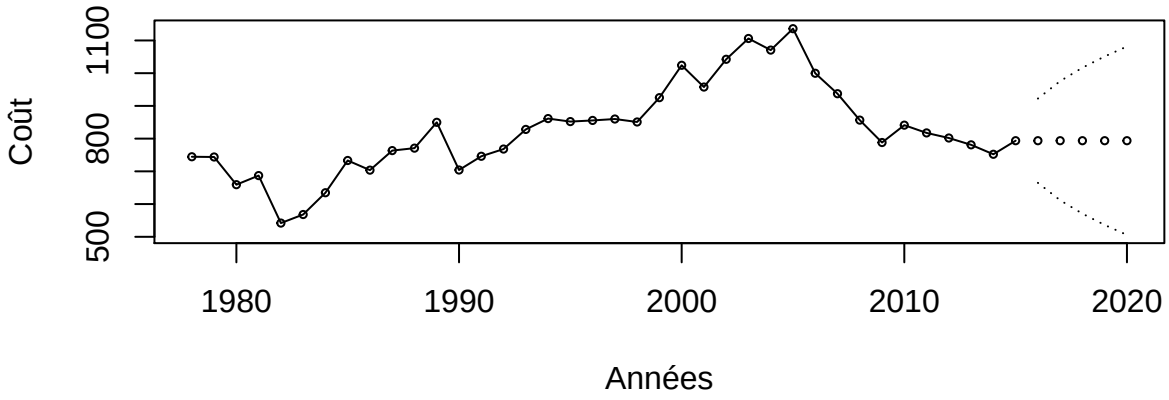
Figure ?? Projection sur 5 ans du nombre d'accidents avec personnes accidentés (NPA).

On remarque que l'intervalle de confiance augmentes avec le temps ce qui est dû à la variance croissante. On obtient ainsi les prédictions suivantes.

Année	Y_{t+l}	Borne inf.	Borne sup.
2016	19896.61	1.7375012×10^4	2.2679161×10^4
2017	20123.51	1.5939885×10^4	2.5068477×10^4
2018	20283.27	1.4799742×10^4	2.7141086×10^4
2019	20418.52	1.3883616×10^4	2.8988626×10^4
2020	20545.05	1.312785×10^4	3.0679926×10^4

CTI

On rappelle que notre modèle final est $\nabla Y_t = e_t$. Les prédictions ainsi que les bornes de nos prédictions du modèle CTI sont affichés ci-dessous


Coût
500
800
1100
1980
1990
2000
2010
2020
Années

Annexe

Code