

Table des matières

Introduction au rapport	2
Taux de change US/Euro	3
Analyse primaire du jeu de données	3
Test de stationnarité	4
Transformation des données	4
Identification du modèle	6
Estimation des paramètres des différents modèles et tests de résidus	8
Prédictions	12
Conclusion	14
SAAQ	15
Analyse primaire du jeu de données	15
Test de stationnarité	16
Transformation des données	16
Identification du modèle	17
Annexe	20
Code	20

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 traitées 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.

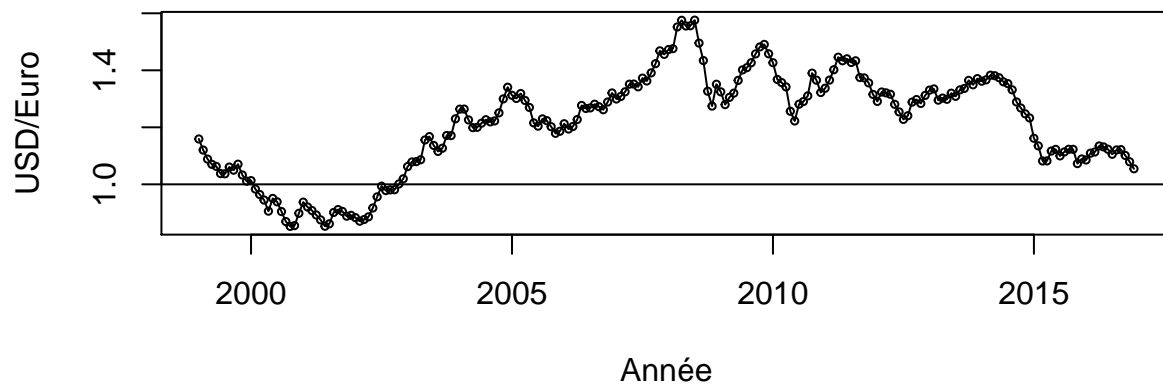


Figure 1. Taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

Grâce à une première analyse, on remarque sur le tableau de la fonction d'autocorrélation échantillonnale 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 débiter notre analyse.

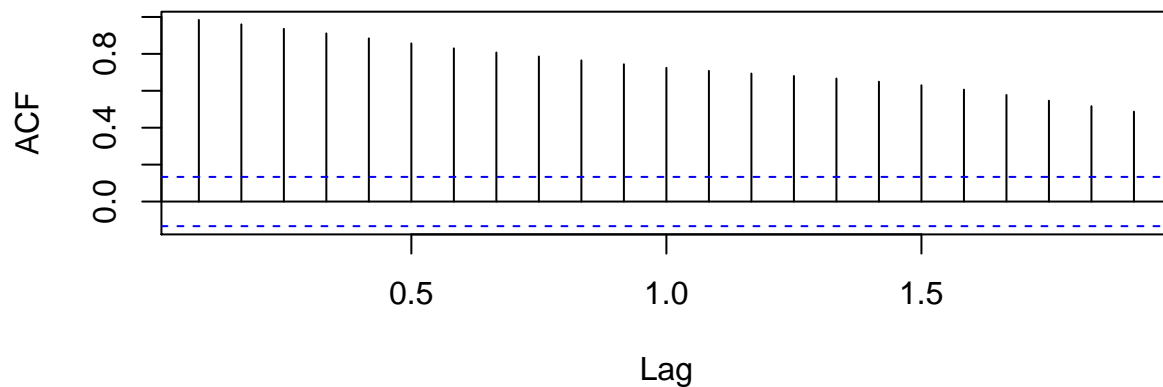


Figure 2. Fonction d'autocorrélation échantillonnales du taux de change US/Euro (lag étant exprimé en années).

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 appropriée 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*.

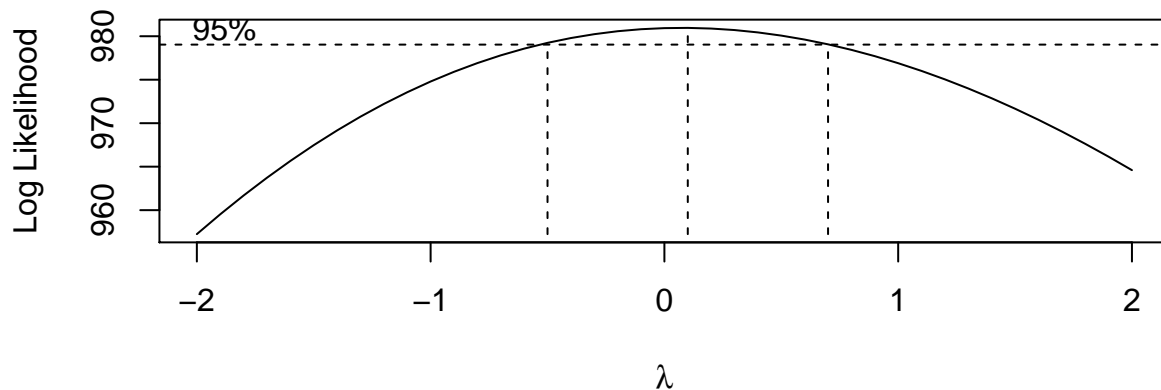


Figure 3. Fonction du logarithme de vraisemblance de la transformée de Box-Cox de la série chronologique du taux de change US/Euro.

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 intervalle 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 particulièrement intéressante puisqu'elle permet de modéliser non pas le «prix» du taux de change, mais son 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 suivante¹

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

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*.

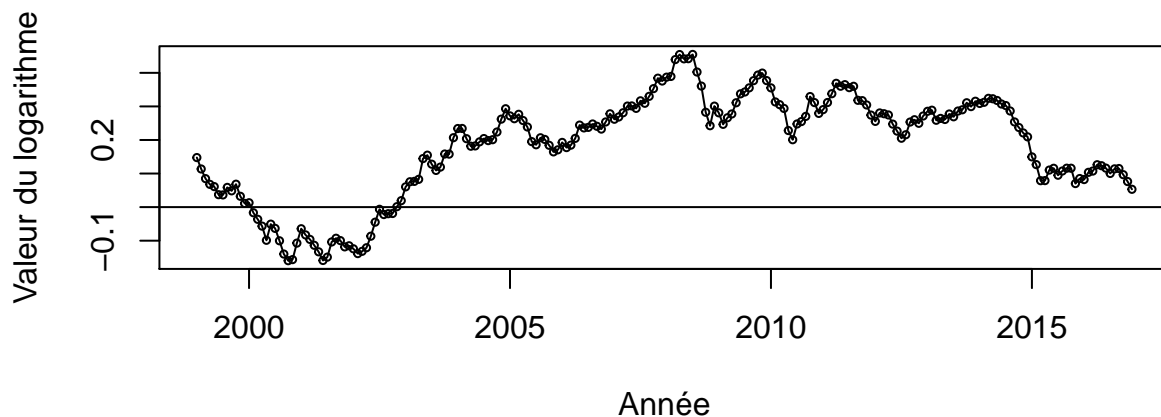


Figure 4. Logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2016.

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 rejeter pas l'hypothèse de stationnarité avec un p-value inférieur à 1%. Ainsi, tel que mentionné précédemment, 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*.

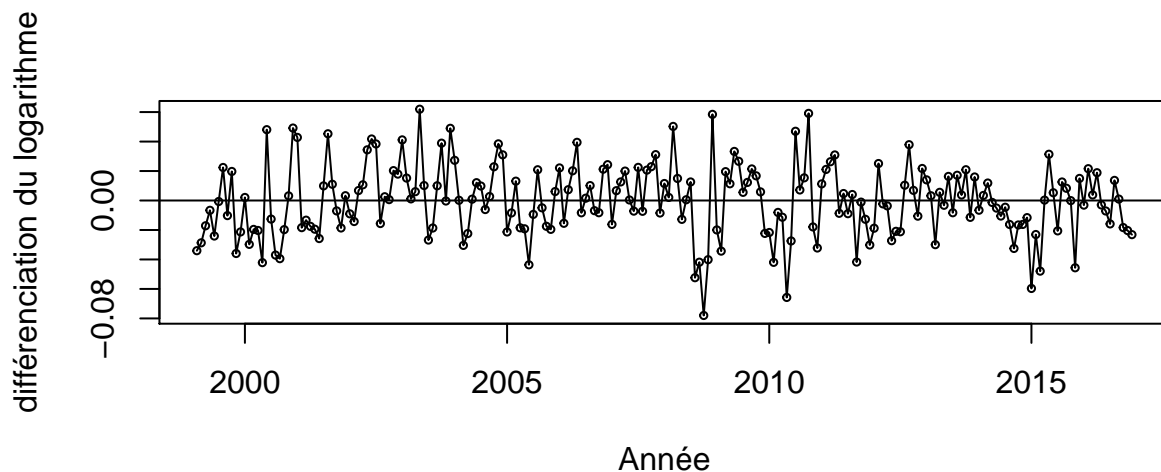


Figure 5. Première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

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 dorénavant 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 transformé 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.

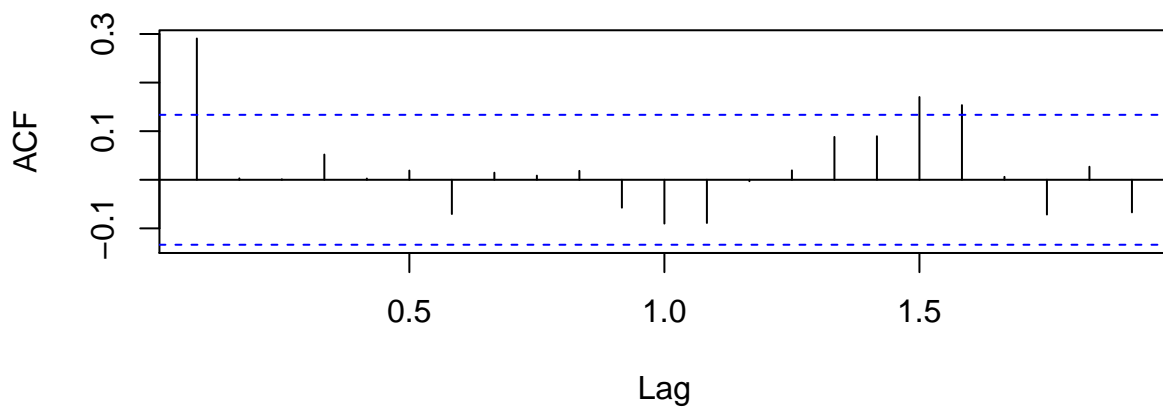


Figure 6. Fonction d'autocorrélation de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

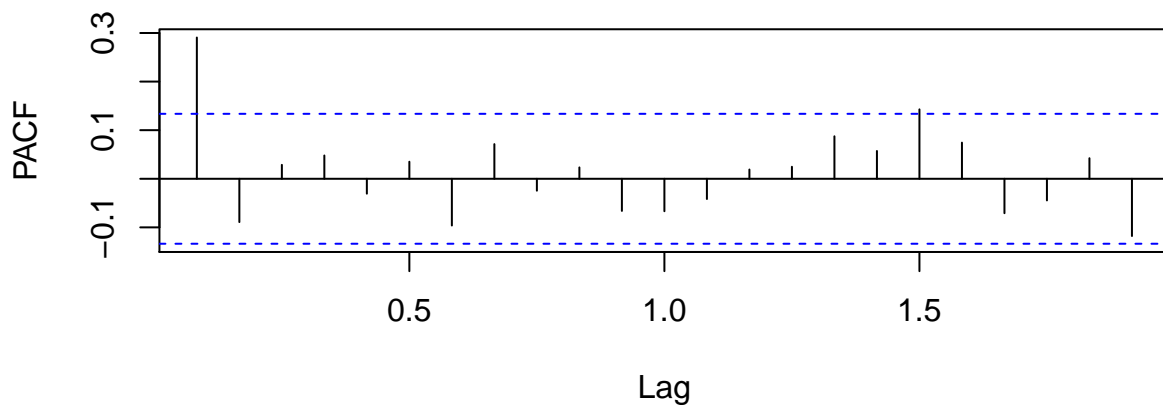


Figure 7. Fonction d'autocorrélation partielle de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

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 suggéré par le tableau EACF. Cependant, comme il fut déjà décidé d'évaluer la pertinence 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 pertinence de ce modèle selon le EACF.

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.

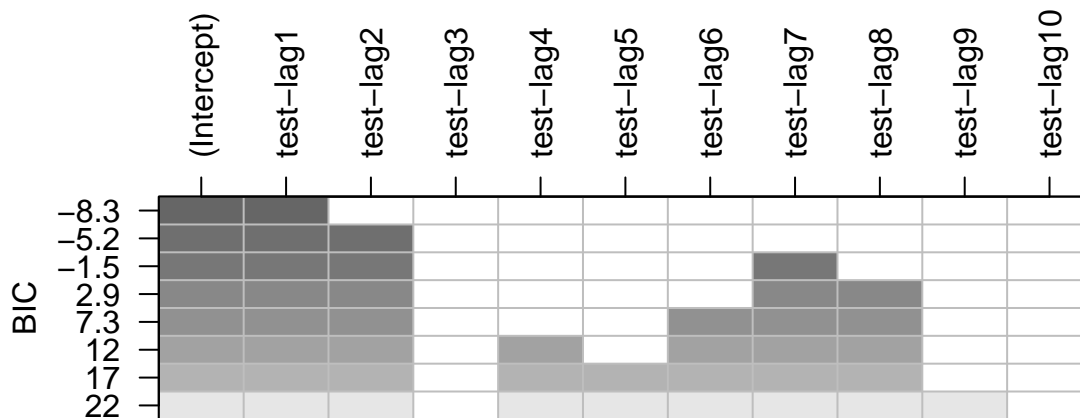


Table II. Tableau de la fonction BIC de la première différenciation du logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2015.

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.

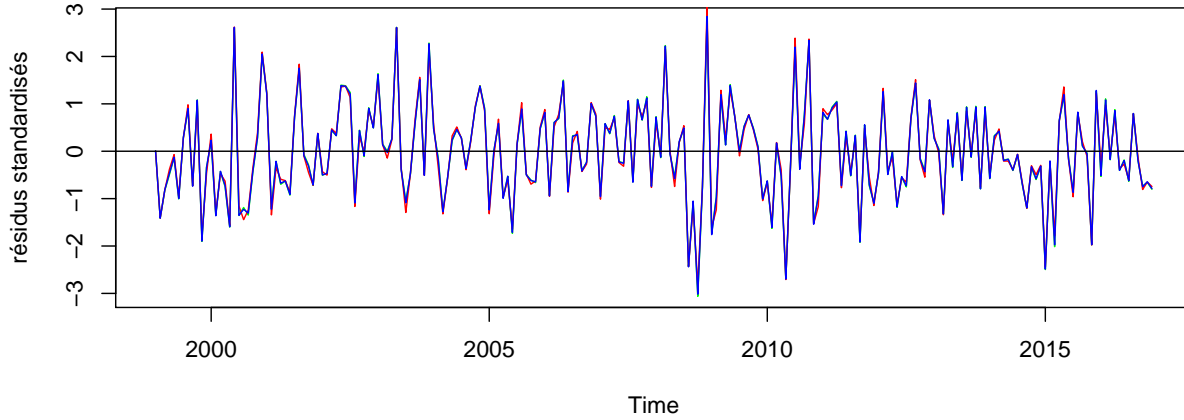


Figure 8. Superposition des graphiques des résidus standardisés des modèles $ARI(1,1)$, $IMA(1,1)$ et $ARIMA(1,1,1)$ du logarithme du taux de change US/Euro.

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.

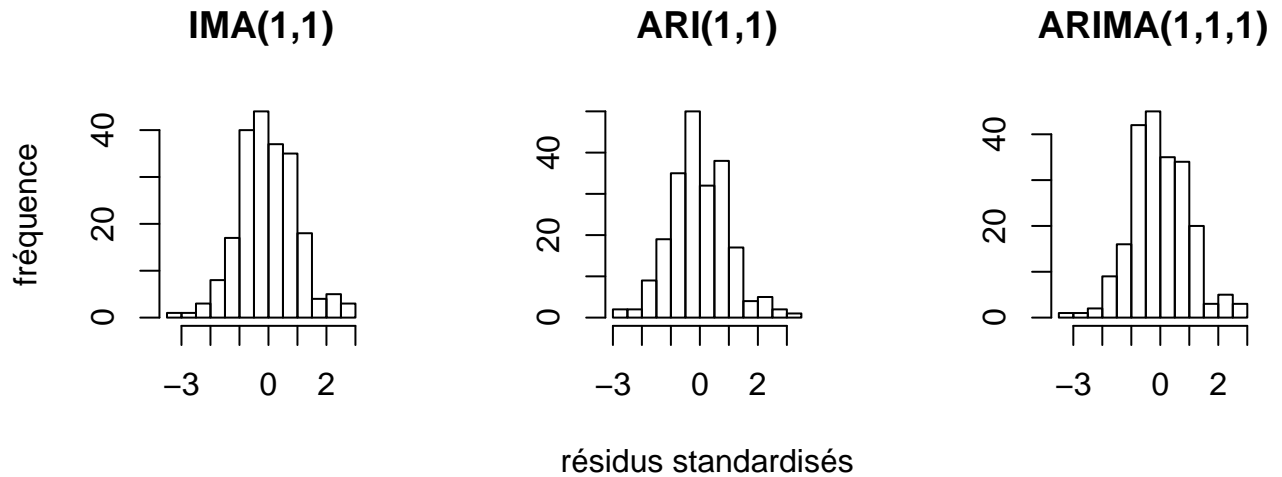


Table IV. Histogramme des résidus standardisés des modèles étudiés du logarithme du taux de change US/Euro.

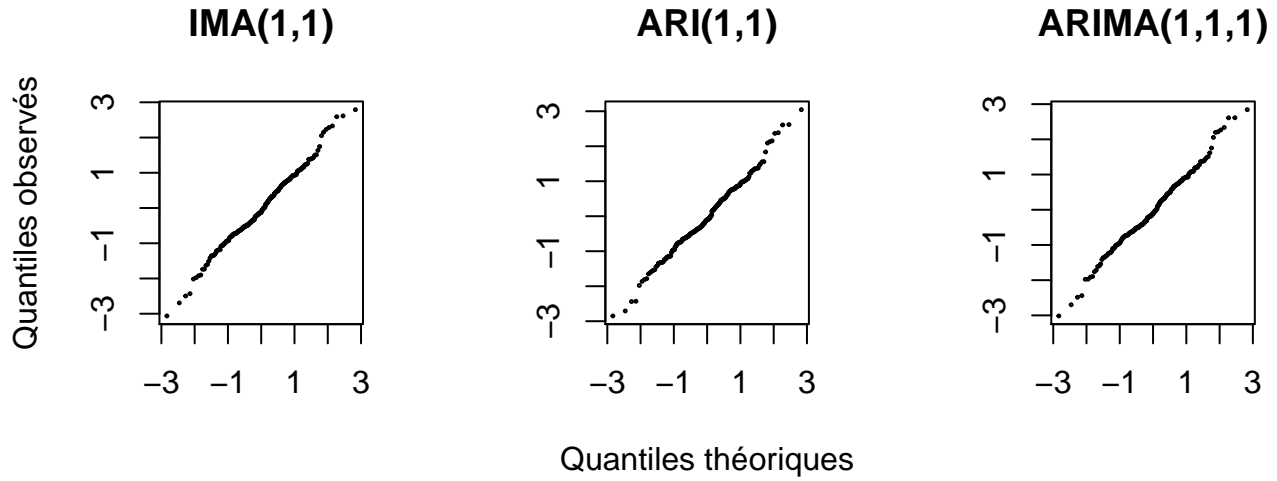


Figure 9. Graphiques QQ des modèles étudiés du logarithme du taux de change US/Euro.

On effectue alors, pour tenter de départager nos modèles, le test de Shapiro-Wilk évaluant le degré 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 confiance de $\alpha = 5\%$ qu'on ne peut rejeter 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 rejeter l'hypothèse nulle. Les trois modèles sont donc, selon ce test, appropriés pour modéliser la série chronologique à l'étude.

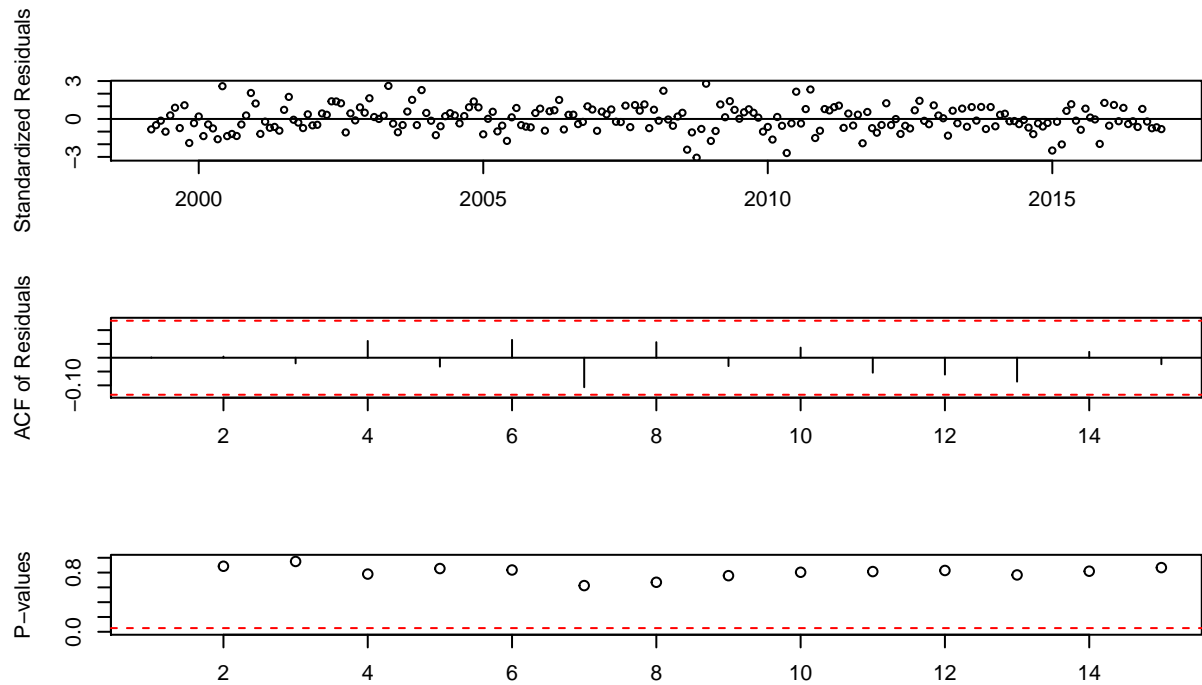


Figure 10. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle IMA(1,1) pour le logarithme du taux de change US/Euro.

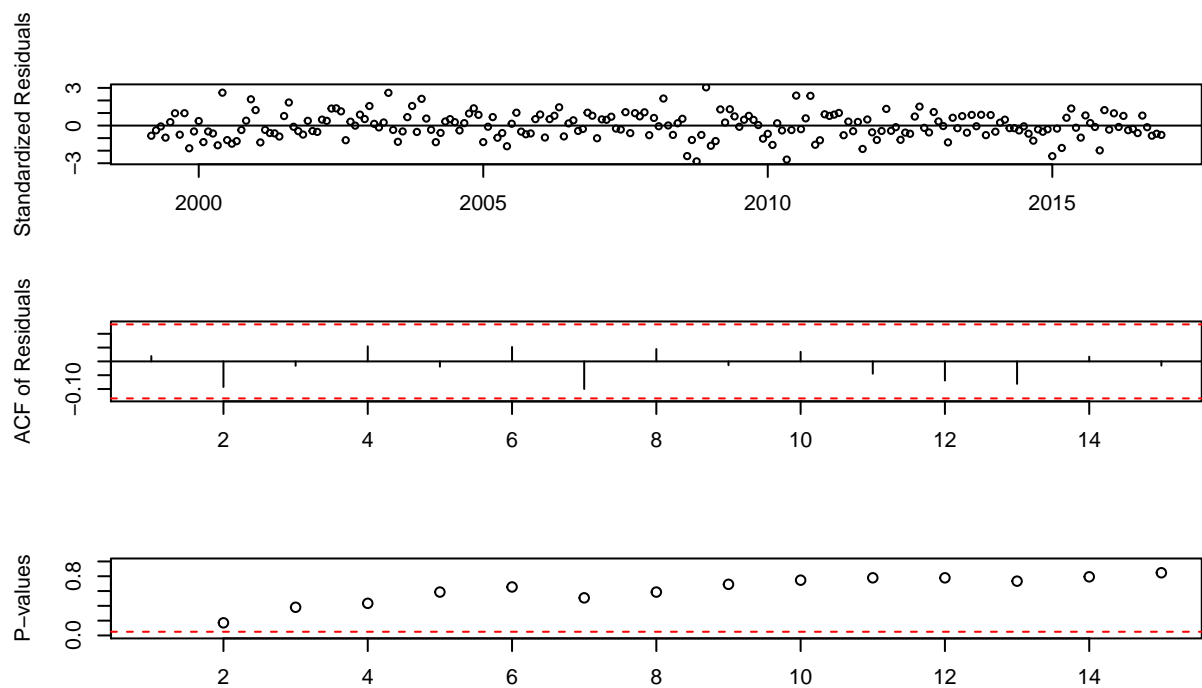


Figure 11. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle ARI(1,1) pour le logarithme du taux de change US/Euro.

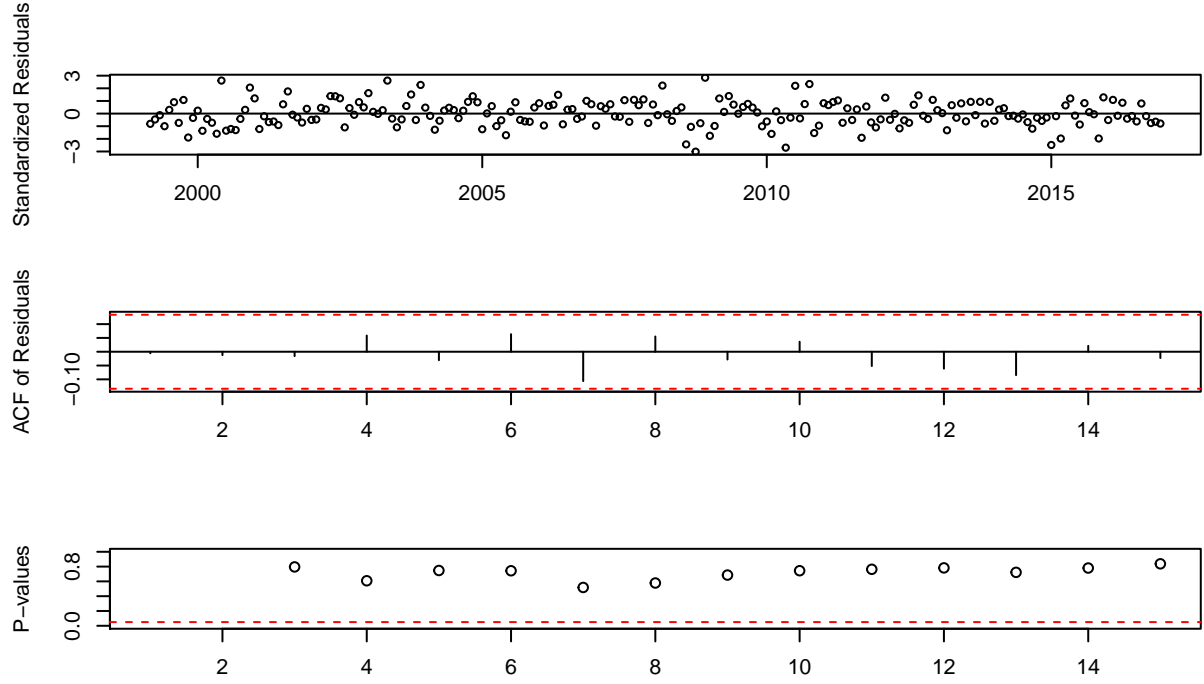


Figure 12. Graphiques des résidus standardisés, de l'autocorrélation des résidus et des p-values du test de Ljung-Box du modèle ARIMA(1,1,1) pour le logarithme du taux de change US/Euro.

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'*overfitting*. 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

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

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%.

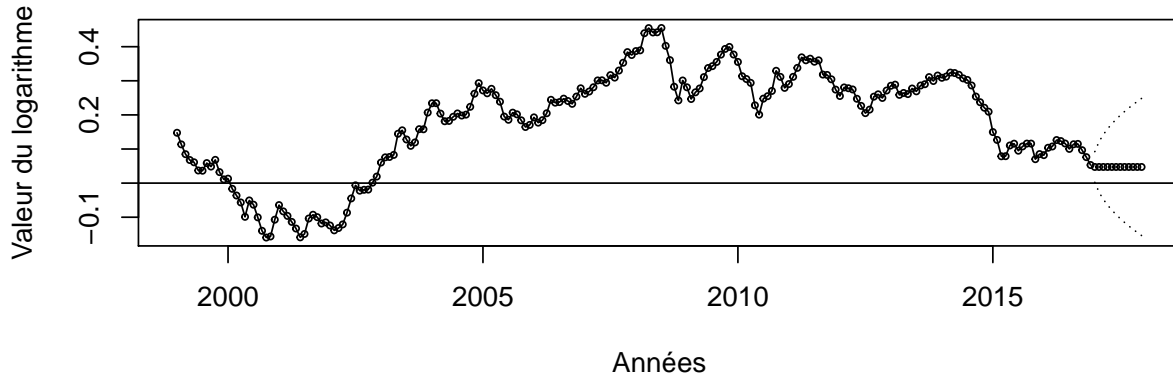


Figure 13. Logarithme du taux de change US/Euro mensuelle de janvier 1999 à décembre 2016 avec prédictions pour l'année 2017 et intervalle de confiance à 95%.

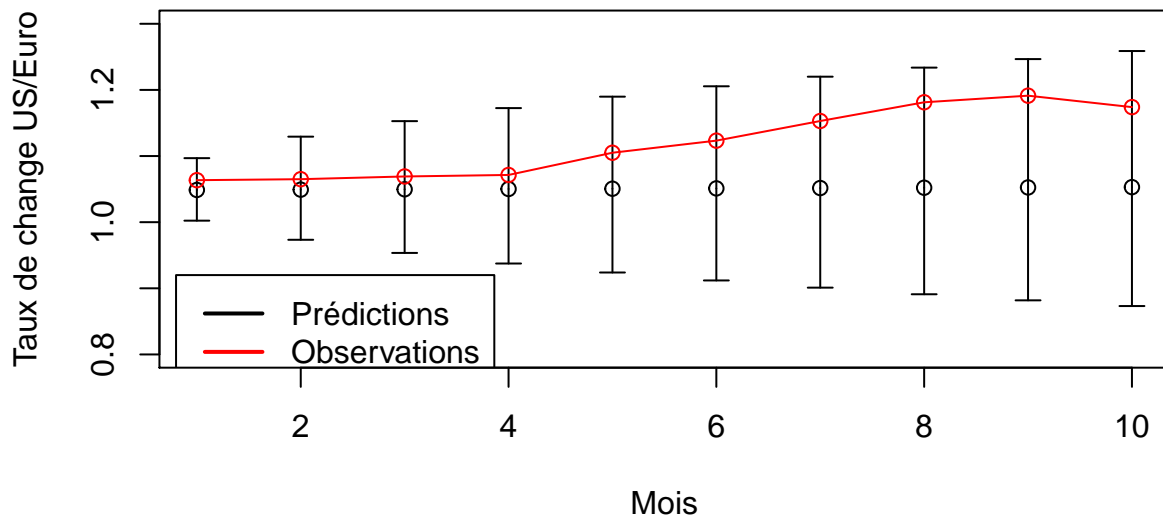


Figure 14. Comparaison entre les valeurs prédites du taux de change US/Euro grâce au modèle IMA(1,1) avec intervalle de confiance et les valeurs observées pour l'année 2017.

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 quatre séries chronologiques depuis 1978 jusqu'en 2015. Cela comprend la série du nombre d'accidents avec dommages corporels (NAADC), du nombre de personnes accidentées (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*

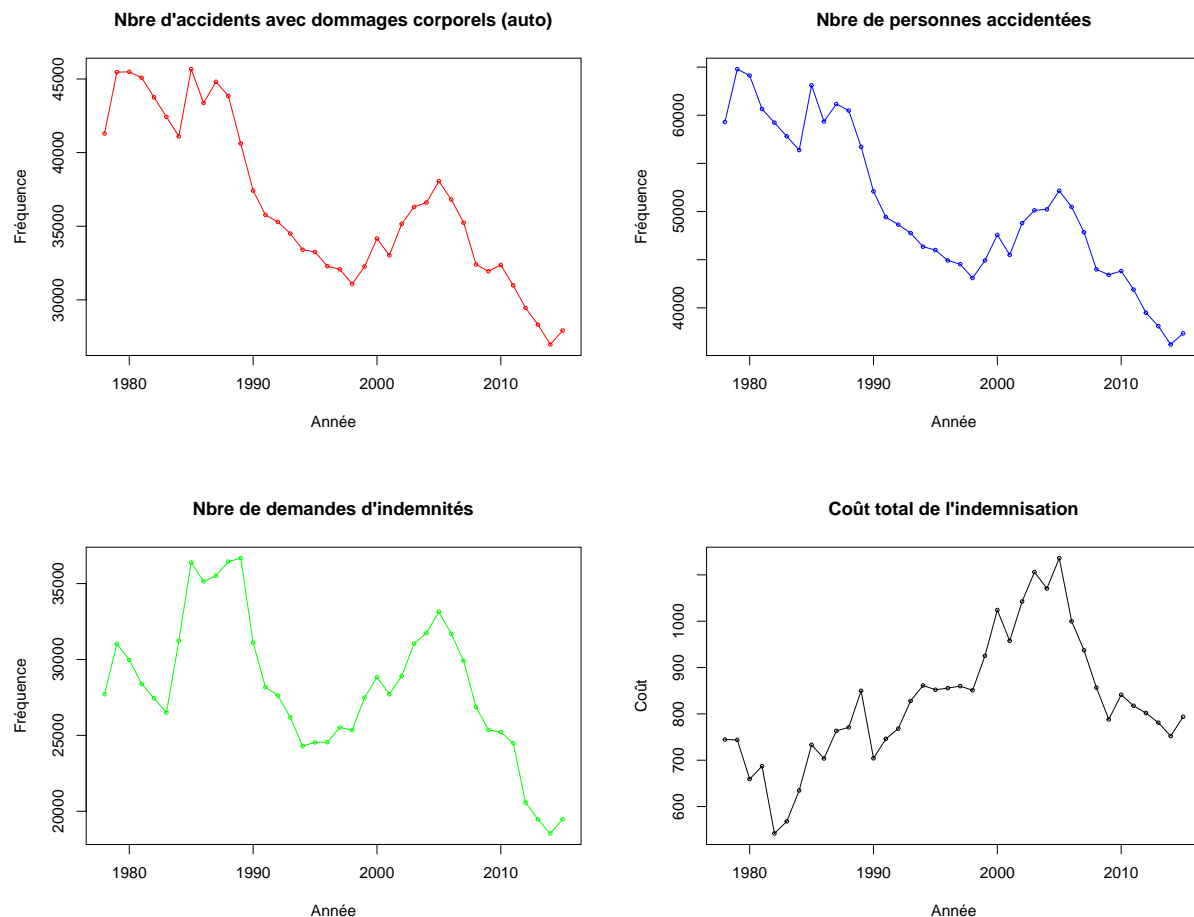


Figure 15. Graphiques des différentes séries chronologiques après modification des données aberrantes depuis 1978 jusqu'en 2015 (NAADC en rouge, NPA en bleu, NDI en vert et CTI en noir).

On remarque principalement 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.

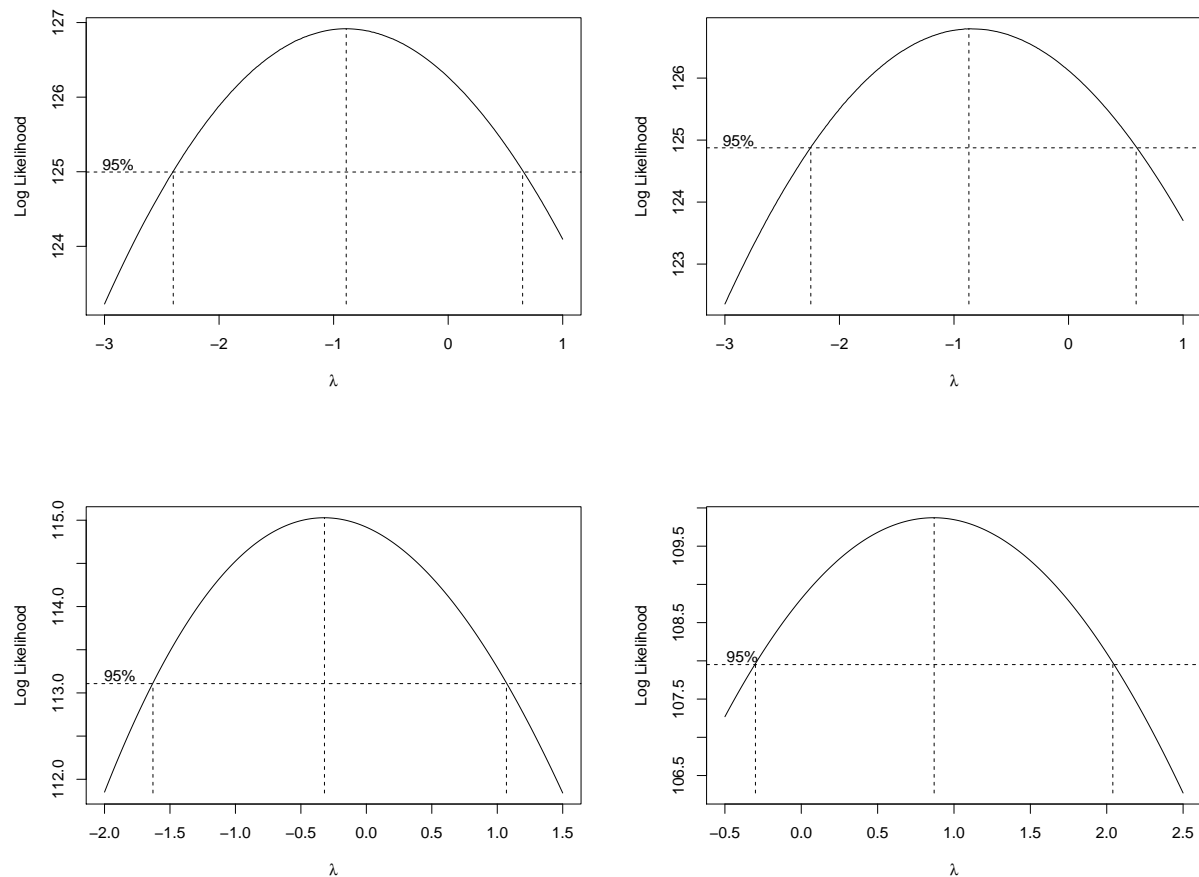
Test de stationnarité

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.

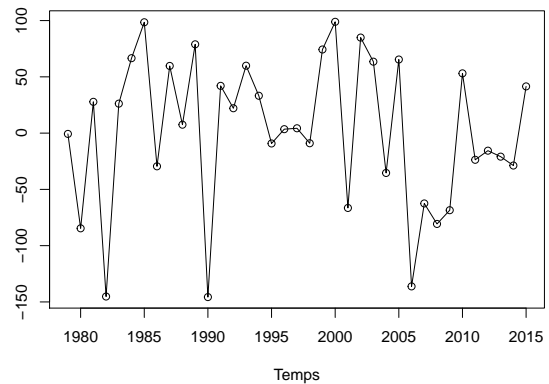
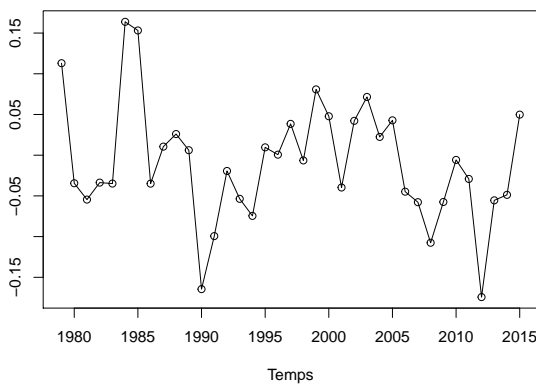
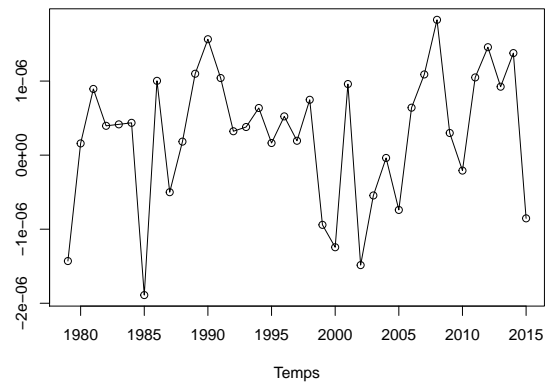
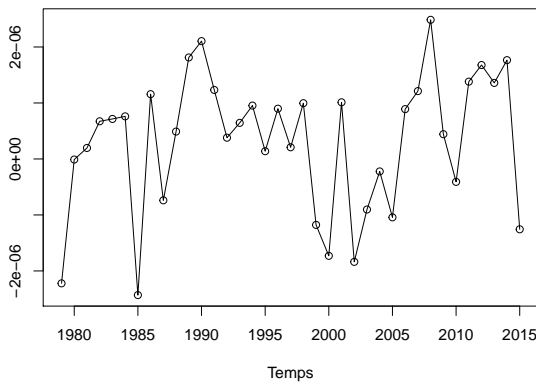
On peut alors, tester la stationnarité de la première différenciation des données transformées encore à l'aide du test de Dikey-Fuller. Ce test renvoie cette fois les valeurs suivantes

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

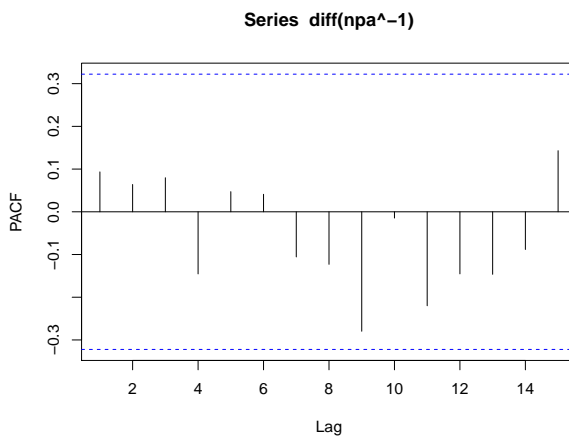
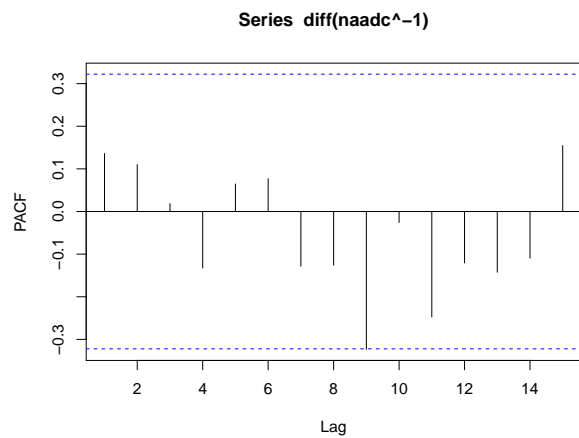
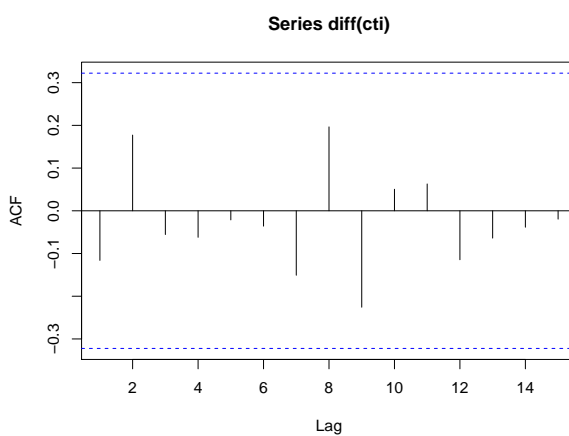
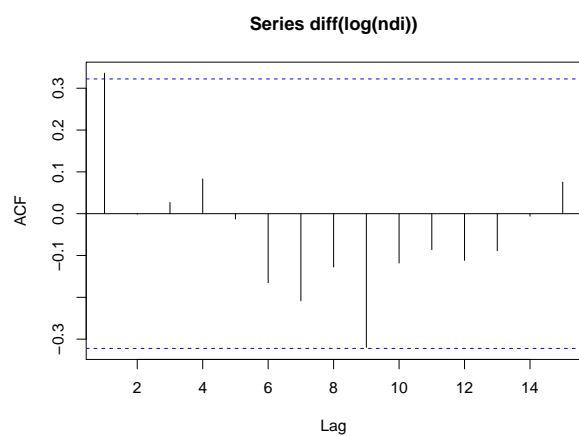
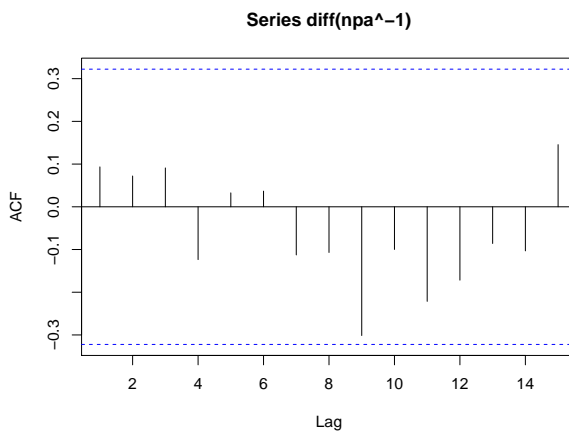
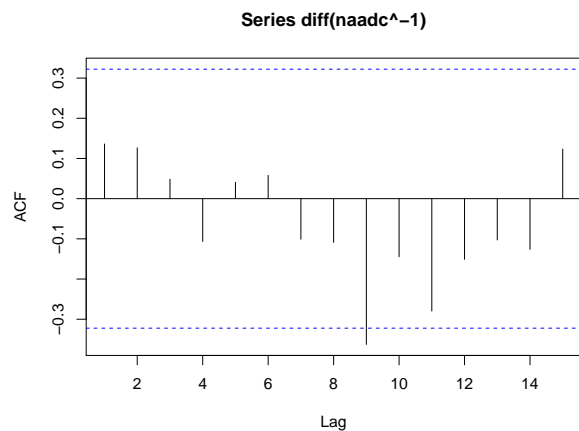
On remarque cette fois que la p-value de chaque modèle est inférieure à 5%, ce qui nous permet d'affirmer la stationnarité de ceux-ci. On peut maintenant passer à l'étape du choix du modèle.

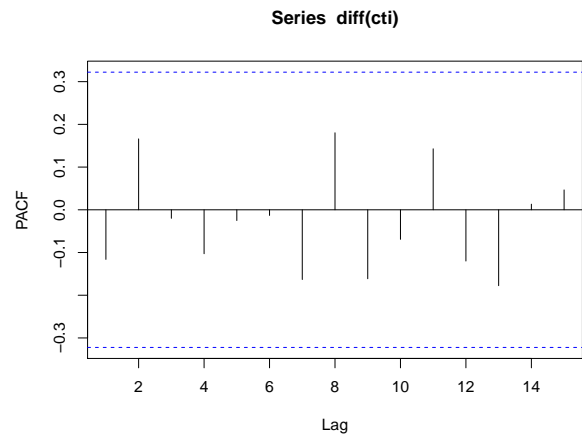
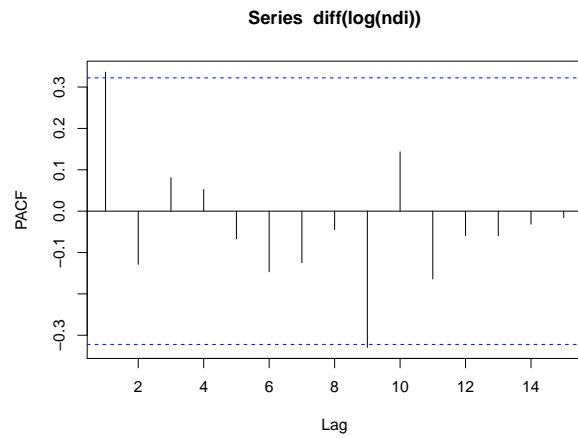
Identification du modèle

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 et d'autocorrélation partielle de chaque série transformée tel qui suit





Annexe

Code