Projet de séries temporelles

BENOÎT CHOFFIN

Mai 2016

Chargé de TD : M. ZILIOTTO

TABLE DES MATIÈRES

TABLE DES FIGURES

1 LES DONNÉES

1) Que représente la série choisie ? (secteur, périmètre, traitements éventuels, transformation logarithmique...)

Les données utilisées pour la réalisation de ce mémoire de séries temporelles correspondent à l'indice brut de la production industrielle produit par l'INSEE (cf. [1]). Plus précisément, l'indice ici utilisé est celui concernant l'**extraction d'hydrocarbures** en France métropolitaine. Cet indice, qui retrace les variations des quantités produites du secteur étudié, répond à un double objectif d'aide à la décision macroéconomique et de comparaison entre les principaux pays développés.

Cet indice a pour base 100 l'année 2010 et est calculé par l'INSEE à partir des enquêtes de branche réalisées par l'INSEE, le SSP (Service de la Statistique et de la Prospective) du Ministère de l'Agriculture, le SOeS (Service de l'Observation et des Statistiques) du Ministère de l'écologie, du développement durable et de l'énergie, ainsi que les organismes professionnels.

La série, mensuelle, couvre les années 1990 à 2016 (février).

En observant une simple représentation graphique de la série, on constate assez rapidement une certaine volatilité de l'indice ; autrement dit, la variance des innovations ϵ_t est dépendante du temps. Elle diminue notamment beaucoup sur la fin de la période considérée (voir FIGURE 4.1). Pour résoudre ce problème d'hétéroscédasticité, j'ai donc choisi d'opérer une **transformation logarithmique** sur la série, afin d'écraser les écarts trop grands entre les variations. On peut effectuer une telle transformation car la série brute X_t est strictement positive pour tout t. On étudiera donc dorénavant la série $Y_t = \log X_t$. Pour une représentation graphique de la série temporelle après cette transformation logarithmique, voir la FIGURE 4.2.

2) Transformer si besoin la série pour la rendre stationnaire (désaisonnalisation, différentiation, suppression de la tendance déterministe...). Justifier soigneusement vos choix.

J'ai dans un premier temps **désaisonnalisé** la série transformée grâce à la fonction *decompose* de R, car elle présentait de manière évidente une saisonnalité au niveau de sa représentation graphique. Ce soupçon s'est trouvé confirmé par l'étude de la décomposition de la série en une tendance, une saisonnalité et un processus aléatoire (cf. FIGURE 4.3). J'ai également choisi un **modèle additif** pour la décomposition, car après la transformation logarithmique, l'amplitude des variations saisonnières est à peu près constante dans le temps. Pour une représentation graphique de la série désaisonnalisée, voir la FIGURE 4.4.

Une fois la série correctement désaisonnalisée, il s'agit maintenant de la rendre stationnaire par **différentiation**, car elle ne l'est pas encore. La stationnarité au second ordre s'applique en effet à une série temporelle dont $\mu_X(t) = \mathrm{E}(X_t)$ est indépendante de t et dont $\gamma_X(t, t+h) = \mathrm{Cov}(X_t, X_{t+h})$ est indépendante de t, pour tout h ; cela ne semble pas être le cas pour la série brute. On confirme cela par un test ADF (Augmented Dickey-Fuller test for unit roots) sur la série désaisonnalisée. Le résultat nous est donné dans l'annexe 4.3. La valeur de la statistique de test est de -1.6432 et le seuil à 5% pour la statistique de test est à -1.95. On en conclut qu'on ne peut pas rejeter l'hypothèse nulle (hypothèse de présence d'une racine unité pour la série désaisonnalisée, i.e. hypothèse de non stationnarité) au seuil de 5%. Pour la rendre stationnaire, on différencie donc une fois la série désaisonnalisée. La représentation graphique de cette série (cf. FIGURE 4.5) différenciée semble indiquer une stationnarité qu'il faut confirmer ou infirmer par un test de stationnarité.

On réalise donc un test ADF sur la série différenciée. Le résultat nous est donné dans l'annexe 4.3.
La p-value obtenue est inférieure à $2.2e^{-16}$, il y a donc un très fort soupçon contre l'hypothèse nulle. En outre, le seuil à 1% pour la statistique de test est de -2.58 et la valeur de la statistique de test est de -16.7577. On en conclut donc qu'on rejette l'hypothèse nulle au seuil de 1% (et même vraisemblablement à des seuils inférieurs).
On peut donc utiliser la série différenciée une fois pour la construction du modèle ARMA.

3) Représenter graphiquement la série avant et après transformation.

Voir les figures 4.1, 4.2, 4.4 et 4.5 en annexe.

2 MODÈLES ARMA

4) Choisir, en le justifiant, un modèle ARMA(p,q) (avec éventuellement une composante saisonnière) pour votre série corrigée X_t. Estimer les paramètres du modèle et vérifier sa validité.

Pour choisir un modèle ARMA convenable, nous allons appliquer la **méthodologie de Box-Jenkins**. Il ne sera pas nécessaire d'inclure une composante saisonnière car la série a déjà été désaisonnalisée dans la PARTIE 1.

Tout d'abord, il faut utiliser l'autocorrélogramme et l'autocorrélogramme partiel de la série différenciée $\Delta \log X_t$ afin de déterminer les ordres maximum p_{max} et q_{max} de notre modèle ARMA. C'est la partie d'identification de la méthodologie de Box-Jenkins.
D'après l'autocorrélogramme (FIGURE 4.6) et l'autocorrélogramme partiel (FIGURE 4.7) de la série différenciée, on obtient $p_{max} = 6$ et $q_{max} = 2$. En effet, la dernière autocorrélation à dépasser le seuil de significativité est clairement la 2e. Quant à la PACF, le résultat semble un peu plus difficile à interpréter du fait de la 6e autocorrélation partielle qui dépasse le seuil de significativité ; au vu des autocorrélations partielles précédentes, on peut supposer qu'il s'agisse du fruit du hasard et que de ce fait, $p_{max} = 2$. Néanmoins, dans un souci de rigueur, nous allons prendre $p_{max} = 6$.

Dès lors que p_{max} et q_{max} ont été déterminés, on applique maintenant une **démarche ascendante** pour choisir le meilleur modèle ARMA. On commence donc par estimer les modèles ARMA(1,1), ARMA(2,2), ARMA(3,3), ARMA(4,4), ARMA(5,5) et ARMA(6,6) et on sélectionne le modèle qui minimise le critère de l'AIC.
On obtient, après estimation des six modèles, une valeur de -827.55 pour l'AIC de l'ARMA(1,1), -828.84 pour l'AIC de l'ARMA(2,2), -824.87 pour l'ARMA(3,3), -826.5 pour l'ARMA(4,4), -821.15 pour l'ARMA(5,5) et -821.6 pour l'ARMA(6,6). On sélectionne donc dans un premier temps le modèle ARMA(2,2).

Ensuite, on cherche à éliminer un ou plusieurs coefficients du modèle pré-sélectionné au-dessus — le critère de l'AIC a en général tendance à privilégier les gros modèles. Pour ce faire, on effectue des tests de nullité des paramètres ϕ_j et ψ_j. Plus précisément, on cherche à tester si, à partir du modèle ARMA(2,2), on ne peut pas éliminer un ou deux coefficients et privilégier un ARMA(1,2) ou un ARMA(2,1) — on ne peut en effet pas tester la nécessité de diminuer simultanément les degrés des polynômes autorégressif et moyenne mobile.
On utilise un test de Student pour tester la nécessité (au seuil 0.05) d'utiliser un modèle ARMA(1,2) ou ARMA(2,1). Le tableau suivant recense les valeurs des statistiques des deux tests respectifs :

	AR2	MA2
Statistique de test	0.783	1.462

Ces deux statistiques de test sont strictement inférieures à 1.96, on peut donc préférer (i.e. ne pas rejeter au seuil de 5% l'hypothèse nulle que l'un des deux modèles suivants soit le bon), soit un modèle ARMA(1,2), soit un modèle ARMA(2,1).
On estime ces deux modèles et on obtient un critère AIC de -830.12 pour l'ARMA(1,2) et de -829.02 pour l'ARMA(2,1). Comme le **modèle ARMA(1,2)** minimise le critère de l'AIC, on le sélectionne. Il ne reste plus qu'à lui faire passer des tests de blancheur des résidus.

La représentation graphique des résidus standardisés, l'ACF des résidus et les p-valeurs pour la statistique de Ljung-Box sont représentées dans la FIGURE 4.8
On constate dans un premier temps que la trajectoire des résidus standardisés ressemble bien à celle d'un bruit blanc, i.e. d'une série de variables aléatoires i.i.d., dont l'espérance est nulle (ce qui semble bien être le cas ici) et la variance, finie.
Enfin, pour tous les retards considérés entre 1 et 30, la p-value pour la statistique de Ljung-Box est supérieure à 0.9, on ne rejette donc pas l'hypothèse nulle d'indépendance du bruit, i.e. d'adéquation du modèle ARMA(1,2).

Il ne reste plus qu'à estimer les paramètres du modèle ARMA(1,2) :

	AR1	MA1	MA2	Intercept
Coefficients	0.3535	-0.5783	-0.2243	-0.0054
S.e.	0.1515	0.1545	0.0950	0.0011

On peut donc écrire le modèle de la manière suivante :

$$\boxed{X_t - 0.3535\, X_{t-1} = -0.0054 + \epsilon_t + 0.5783\, \epsilon_{t-1} + 0.2243\, \epsilon_{t-2}}$$

3 PRÉVISION

On note T la longueur de la série. Ici, $T = 312$ (12 mois sur 26 ans). On suppose que les résidus de la série sont gaussiens.

5) Écrire l'équation vérifiée par la région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2}).

Nous avons travaillé jusqu'ici sur des données couvrant la période allant de janvier 1990 à février 2016 ; il est donc important de noter que nous n'avons pas $T = 312$, mais $T = 314$.

Prenons un cas général et déterminons la région de confiance de niveau α sur une valeur future Z_{T+h}, avec $h > 0$. On utilise la forme $MA(\infty)$ de notre modèle ARMA(1,2) (adapté à la

série différenciée $Z_t = \Delta \log X_t$) :

$$Z_{n+h} = \epsilon_{n+h} + \sum_{j=1}^{\infty} c_j \epsilon_{n+h-j} \text{ avec } \epsilon_t \text{ l'innovation de } Z_t$$

Quand on projette cette relation sur le passé infini de Z_n, on obtient :

$$\hat{Z}_{n+h|Z_n\ldots} = \sum_{j=h}^{\infty} c_j \epsilon_{n+h-j}$$

L'équation précédente découle du fait que, d'après l'énoncé, ϵ_t est un **bruit blanc gaussien**, i.e. $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. On en déduit l'équation vérifiée par $e_t(h)$, l'erreur de prévision à horizon h :

$$e_t(h) = Z_{n+h} - \hat{Z}_{n+h|Z_n\ldots} = \epsilon_{n+h} + \sum_{j=1}^{h-1} c_j \epsilon_{n+h-j}$$

Par un calcul très simple (on sait notamment que $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$), on obtient que :

$$\text{Var}(e_t(h)) = \sigma^2 (1 + \sum_{j=1}^{h-1} c_j^2)$$

Et que donc :

$$e_t(h) \sim \mathcal{N}(0, \sigma^2 (1 + \sum_{j=1}^{h-1} c_j^2))$$

Ainsi, on a :

$$P(q_{\frac{\alpha}{2}} \leq \frac{e_t(h)}{\sigma \sqrt{(1 + \sum_{j=1}^{h-1} c_j^2)}} \leq q_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

avec q_α le quantile d'ordre α d'une loi normale centrée réduite.

D'où :

$$P(q_{\frac{\alpha}{2}} \sigma \sqrt{(1 + \sum_{j=1}^{h-1} c_j^2)} \leq Z_{n+h} - \hat{Z}_{n+h|Z_n\ldots} \leq q_{1-\frac{\alpha}{2}} \sigma \sqrt{(1 + \sum_{j=1}^{h-1} c_j^2)}) = 1 - \alpha$$

Ainsi :

$$P(\hat{Z}_{n+h|Z_n\ldots} + q_{\frac{\alpha}{2}} \sigma \sqrt{(1 + \sum_{j=1}^{h-1} c_j^2)} \leq Z_{n+h} \leq \hat{Z}_{n+h|Z_n\ldots} + q_{1-\frac{\alpha}{2}} \sigma \sqrt{(1 + \sum_{j=1}^{h-1} c_j^2)}) = 1 - \alpha$$

On en déduit l'intervalle de confiance de niveau α pour Z_{T+1} :

$$\boxed{[\hat{Z}_{T+1|Z_T\ldots} + q_{\frac{\alpha}{2}} \sigma \; ; \; \hat{Z}_{T+1|Z_T\ldots} + q_{1-\frac{\alpha}{2}} \sigma]}$$

Ainsi que l'intervalle de confiance de niveau α pour Z_{T+2} :

$$\boxed{[\hat{Z}_{T+2|Z_T\ldots} + q_{\frac{\alpha}{2}} \sigma \sqrt{1 + c_1^2} \; ; \; \hat{Z}_{T+2|Z_T\ldots} + q_{1-\frac{\alpha}{2}} \sigma \sqrt{1 + c_1^2}]}$$

6) Préciser les hypothèses utilisées pour obtenir cette région.

Les hypothèses utilisées pour obtenir cette région de confiance sont :
— (ϵ_t) est un bruit blanc gaussien ;
— (Z_t) est une série stationnaire au second ordre (pour la forme $MA(\infty)$, d'après Wold).
Ces deux conditions étant vérifiées pour notre série, on peut donc déterminer la région de confiance à 95% pour Z_{T+1} et Z_{T+2}.

7) Déterminer graphiquement cette région pour $\alpha = 95\%$. Commenter.

Voir la FIGURE 4.9.

Les valeurs prédites sont en rouge, tandis que les intervalles de confiance à 95% sont en orange.
On constate dans un premier temps que les intervalles de confiance sont centrés en l'espérance empirique de la série Z_t ; comme la série est elle-même stationnaire et d'espérance nulle, il s'agit de la prévision linéaire de Z_{T+1} et Z_{T+2}.
En outre, l'intervalle de confiance s'élargit pour la 2e prédiction. C'est normal car on a vu à la question précédente que l'intervalle de confiance pour Z_{T+2} différait de celui pour Z_{T+1} à cause du facteur multiplicatif $\sqrt{1+c_1^2}$ devant $q_\alpha\sigma$. Ce facteur étant nécessairement supérieur à 1, l'intervalle de confiance est donc plus grand pour les prédictions ultérieures.

8) Question ouverte : soit Y_t une série stationnaire disponible de $t = 1$ à T. On suppose que Y_{T+1} est disponible plus rapidement que X_{T+1}. À quelles conditions cette information permet-elle d'améliorer la prévision de X_{T+1} ?

Cette information permet d'améliorer la prévision de X_{T+1} à la condition que Y_t **cause instantanément** (au sens de Granger) X_t. Comme il est précisé que Y_t est stationnaire comme X_t, il faut donc que $\hat{X}_{T+1|\{X_u,Y_u,u\leq T\}\cup\{Y_{T+1}\}} \neq \hat{X}_{T+1|\{X_u,Y_u,u\leq T\}}$.

On peut tester l'hypothèse nulle de non-causalité au sens de Granger au moyen d'un test de Granger.

RÉFÉRENCES

[1] Site de l'INSEE, http://www.insee.fr/fr/bases-de-donnees/bsweb/theme.asp?id=07

4 ANNEXES

4.1 CODE R

```
library(urca)
library(apt)
library(tseries)
library(foreign)
library(forecast)
library(fUnitRoots)

#PARTIE 1

#Extraction d'hydrocarbures
setwd("/Users/benoitchoffin/Desktop")

data <- read.csv("time_series.csv")

#transformation en objet "time series"
datats <- ts(data, frequency=12, start=c(1990,1))

#permutation de la serie, sinon elle est dans le mauvais sens
hycarb <- rev(datats)
hycarbts <- ts(hycarb, frequency=12, start=c(1990,1))

#Représentation graphique de la série brute
plot.ts(hycarbts, type="l",lwd=1, col="red", xlab="Time", ylab="Indice")

#on prend le logarithme de la série pour écraser
#les pics de saisonnalité
loghycarb <- log(hycarbts)
loghycarbts <- ts(loghycarb, frequency = 12, start = c(1990,1))

#Représentation graphique de la série log-transformée
plot.ts(loghycarbts, type="l",lwd=1, col="red", xlab="Time",
        ylab="Indice")

#désaisonnalisation de la série loghycarbts
df_desais <- decompose(loghycarbts)
```

```
plot(df_desais)

hycarb_desais <- loghycarbts - df_desais$seasonal
#Représentation graphique de la série désaisonnalisée
plot.ts(hycarb_desais,type="l",lwd=1, col="red", xlab="Time",
        ylab="Indice")
#Test ADF sur la série désaisonnalisée
urdfTest(hycarb_desais)

#création de la série différenciée
hycarbdiff1 <- diff(hycarb_desais, differences=1)
#Représentation graphique de la série différenciée
plot.ts(hycarbdiff1,type="l",lwd=1, col="red", xlab="Time",
        ylab="Indice")

#Test ADF pour déterminer la stationnarité de la série différenciée
urdfTest(hycarbdiff1)
#on obtient bien une stationnarité de la série différenciée
#(seuil de 1%)

#PARTIE 2
#Détermination des pmax et qmax à partir des ACF/PACF
acf(hycarbdiff1,lag.max=20) #a priori qmax = 2
pacf(hycarbdiff1,lag.max=20) #a priori pmax = 6

#Mise en place d'une stratégie ascendante :
#Estimation des modèles (1,1), (2,2)
#(3,3), (4,4), (5,5) et (6,6) : minimisation du critère AIC

mod1 <- Arima(hycarbdiff1, order=c(1,0,1)) #AIC = -827.55
mod2 <- Arima(hycarbdiff1, order = c(2,0,2)) #AIC = -828.84
mod3 <- Arima(hycarbdiff1, order = c(3,0,3)) #AIC = -824.87
mod4 <- Arima(hycarbdiff1, order = c(4,0,4)) #AIC = -826.5
mod5 <- Arima(hycarbdiff1, order = c(5,0,5)) #AIC = -821.15
mod6 <- Arima(hycarbdiff1, order = c(6,0,6)) #AIC = -821.6
#on choisit un ARMA(2,2) pour commencer, car il minimise l'AI
#entre les 6 modèles
#comme l'AIC privilégie les gros modèles, on va essayer d'enlever des
#paramètres grâce aux tests sur les paramètres

abs(mod2$coef)/sqrt(diag(mod2$var.coef))
#on peut supprimer deux coefficients (pas en même temps bien entendu)
#on teste donc les modèles ARMA(1,2) et ARMA(2,1)
```

```
mod12 <- Arima(hycarbdiff1, order=c(1,0,2)) #AIC = -830.12
mod21 <- Arima(hycarbdiff1, order=c(2,0,1)) #AIC = -829.02
#on sélectionne donc un ARMA(1,2) car il minimise l'AIC

#Testons maintenant la blancheur des résidus de ce modèle
tsdiag(mod12, gof.lag=30)

#PARTIE 3

#Prévision

hycarbforecasts <- forecast.Arima(mod12,h=2,level=c(95))
plot.forecast(hycarbforecasts,shadecols = 'orange',fcol='red',
              main = 'Intervalles de confiance à 95% pour T+1 et T+2')
```

4.2 FIGURES

FIGURE 4.1 – Représentation graphique de la série brute

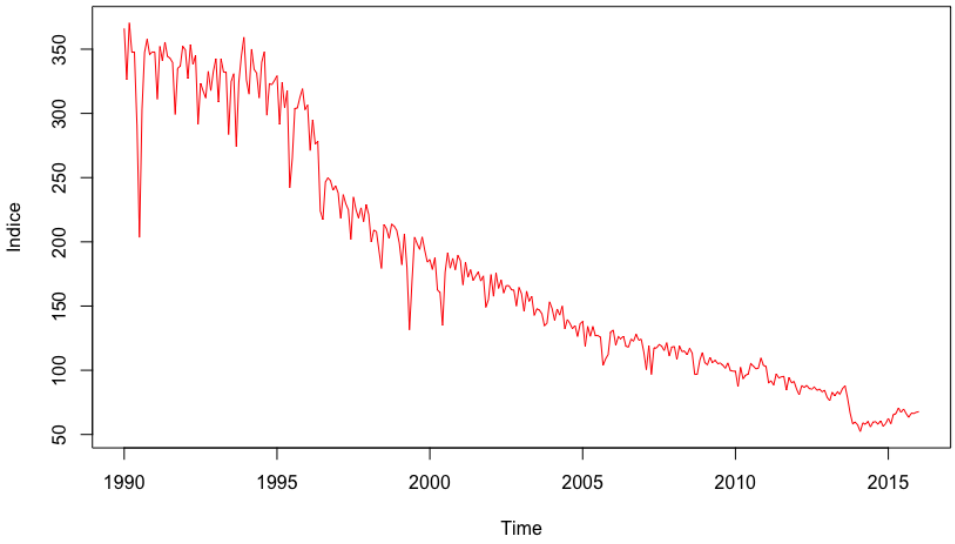


FIGURE 4.2 – Représentation graphique de la série log-transformée

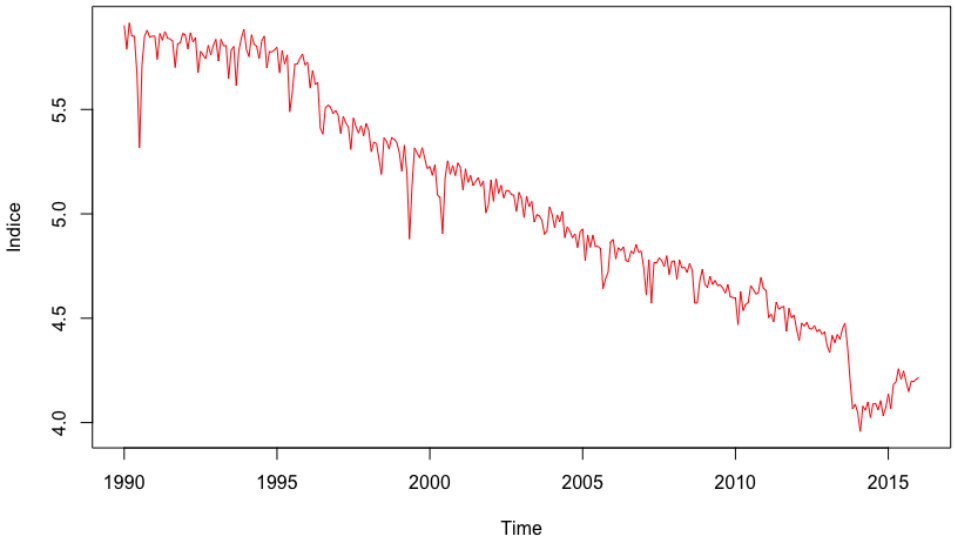


FIGURE 4.3 – Décomposition saisonnière additive de la série

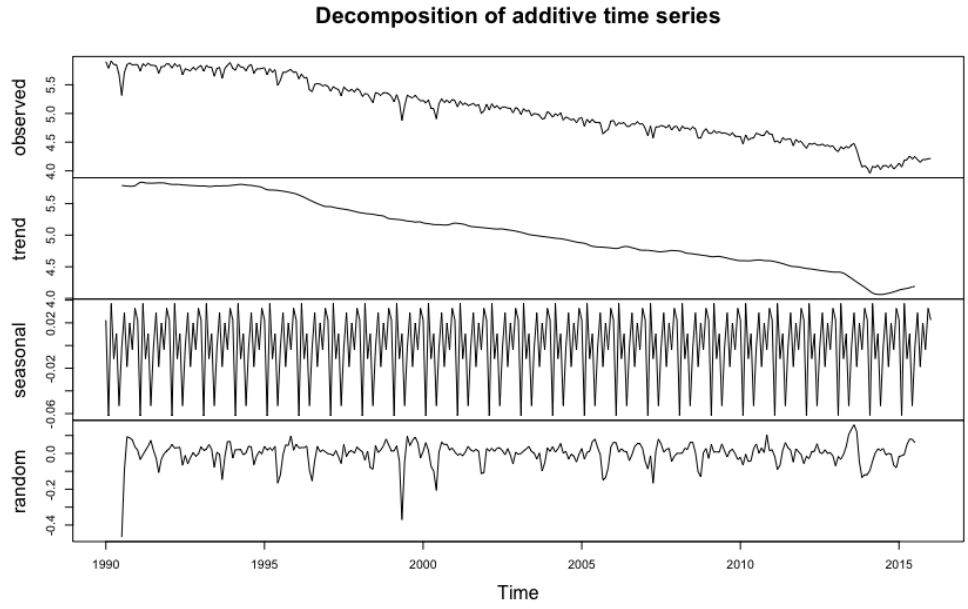


FIGURE 4.4 – Représentation graphique de la série désaisonnalisée

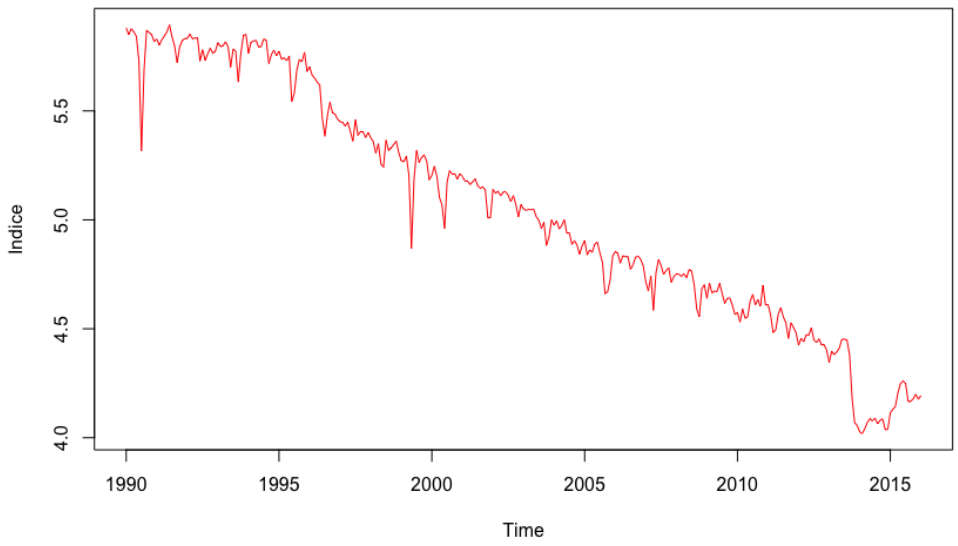


FIGURE 4.5 – Représentation graphique de la série différenciée

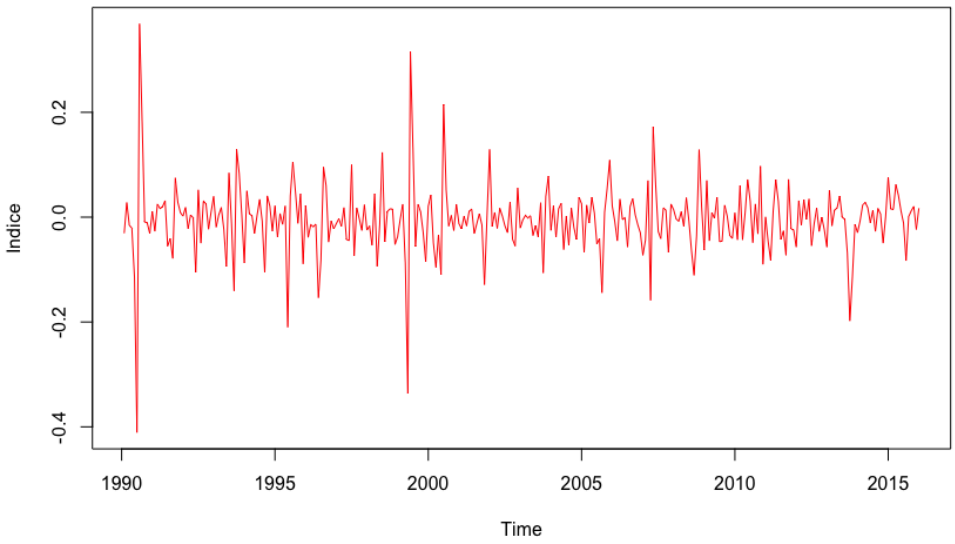


FIGURE 4.6 – Autocorrélogramme de la série différenciée

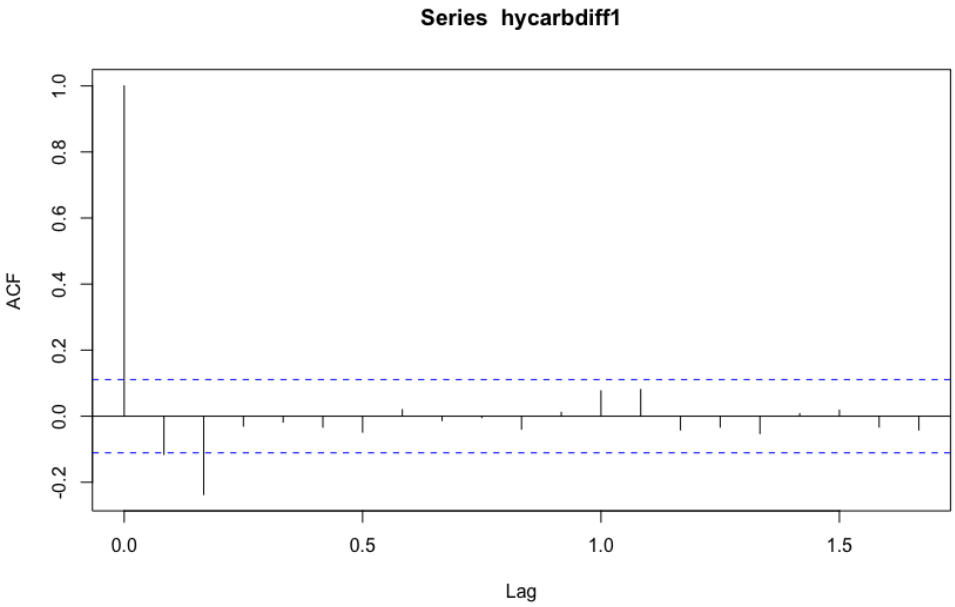


FIGURE 4.7 – Autocorrélogramme partiel de la série différenciée

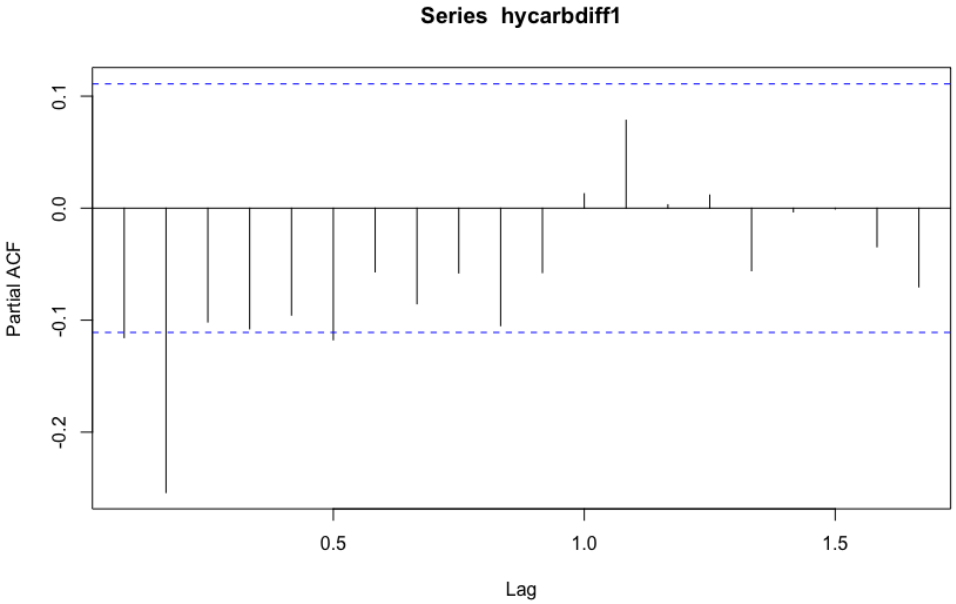


FIGURE 4.8 – ARMA(1,2) : Résidus standardisés, ACF des résidus et p-valeurs pour la statistique de Ljung-Box

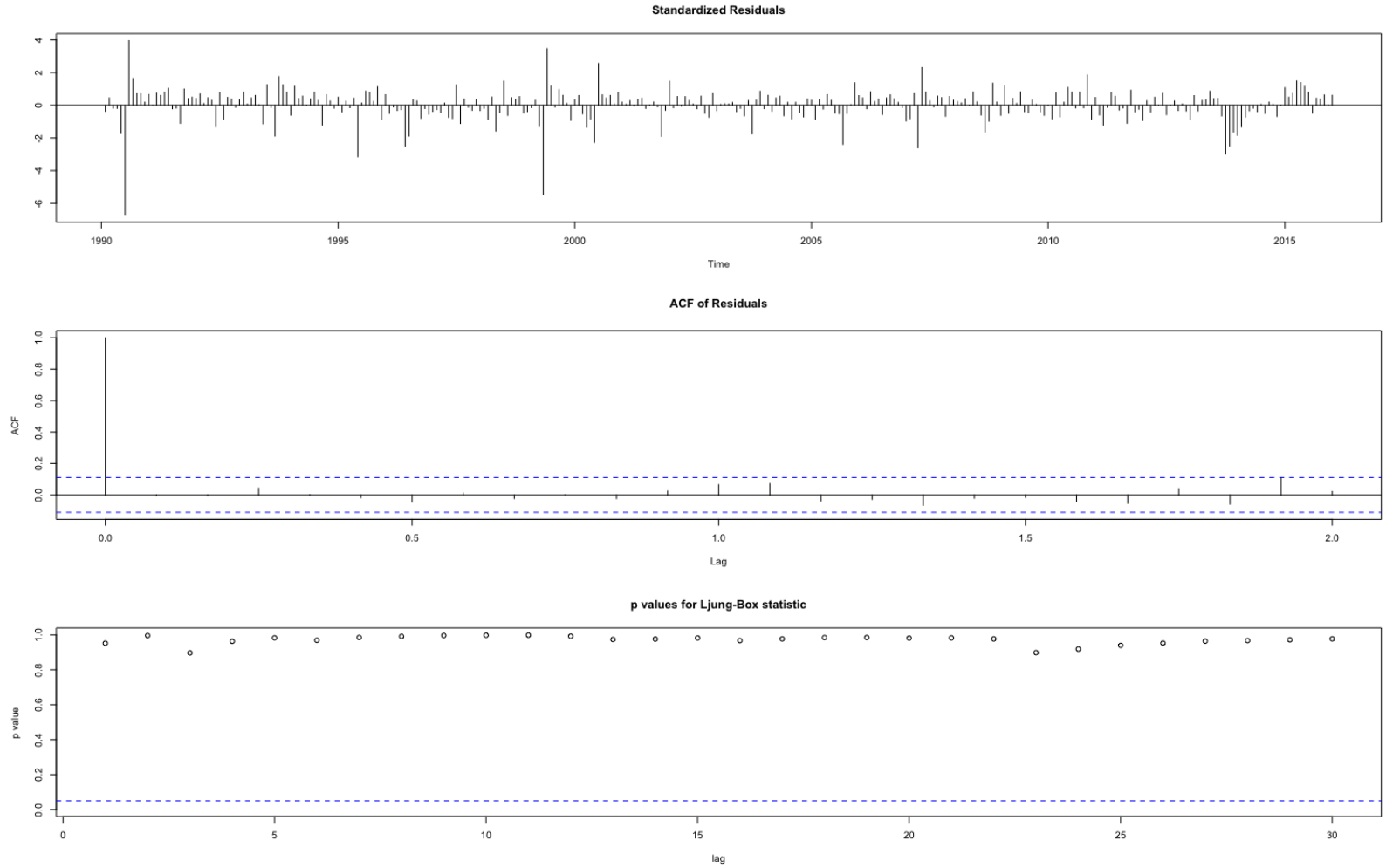
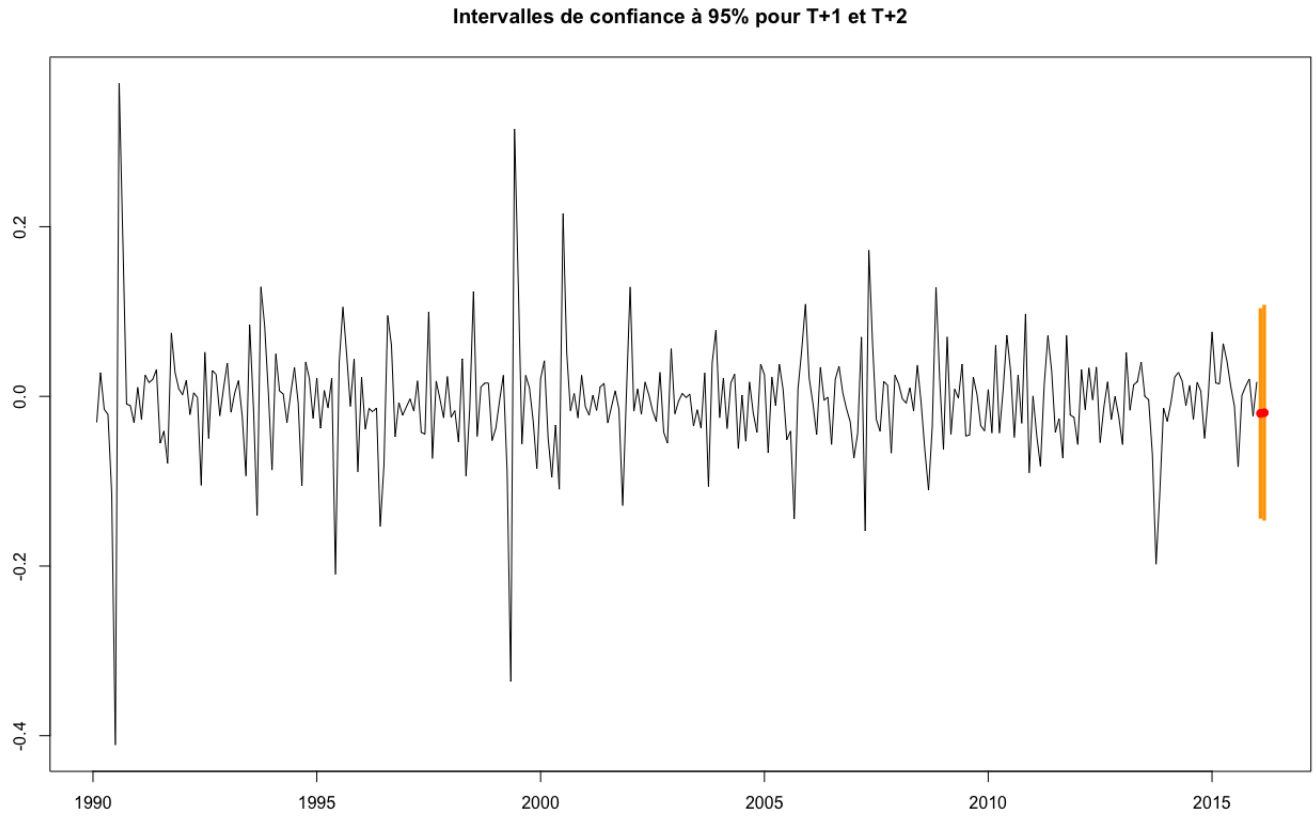


FIGURE 4.9 – Représentation graphique des intervalles de confiance à 95% pour $T+1$ et $T+2$



```
Title:
 Augmented Dickey-Fuller Unit Root Test

Test Results:

  Test regression none

  Call:
  lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

  Residuals:
       Min       1Q   Median       3Q      Max
  -0.41680 -0.02804  0.00376  0.02658  0.32792

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  z.lag.1    -0.0012386  0.0007538  -1.643   0.1014
  z.diff.lag -0.1153122  0.0564456  -2.043   0.0419 *
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Residual standard error: 0.06735 on 309 degrees of freedom
  Multiple R-squared:  0.02034,        Adjusted R-squared:  0.014
  F-statistic: 3.208 on 2 and 309 DF,  p-value: 0.04179


  Value of test-statistic is: -1.6432

  Critical values for test statistics:
        1pct  5pct 10pct
  tau1 -2.58 -1.95 -1.62
```

Test ADF sur la série désaisonnalisée

```
Title:
 Augmented Dickey-Fuller Unit Root Test

Test Results:

  Test regression none

  Call:
  lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

  Residuals:
       Min       1Q   Median       3Q      Max
  -0.43123 -0.03360 -0.00228  0.02370  0.28590

  Coefficients:
            Estimate Std. Error t value Pr(>|t|)
  z.lag.1    -1.37836    0.08225 -16.758  < 2e-16 ***
  z.diff.lag  0.24394    0.05524   4.416 1.39e-05 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Residual standard error: 0.06569 on 308 degrees of freedom
  Multiple R-squared:  0.5806,        Adjusted R-squared:  0.5779
  F-statistic: 213.2 on 2 and 308 DF,  p-value: < 2.2e-16


  Value of test-statistic is: -16.7577

  Critical values for test statistics:
        1pct  5pct 10pct
  tau1 -2.58 -1.95 -1.62
```

Test ADF sur la série différenciée