Ecole Nationale de la Statistique et de l'Administration Economique

Projet de Séries Temporelles

Indice CVS-CJO de la production industrielle (base 100 en 2015) - Extraction d'hydrocarbures : Identifiant 010537206

Anthony Ivanier, Victor Michel

9 mai 2022

Christian Francq

Sommaire

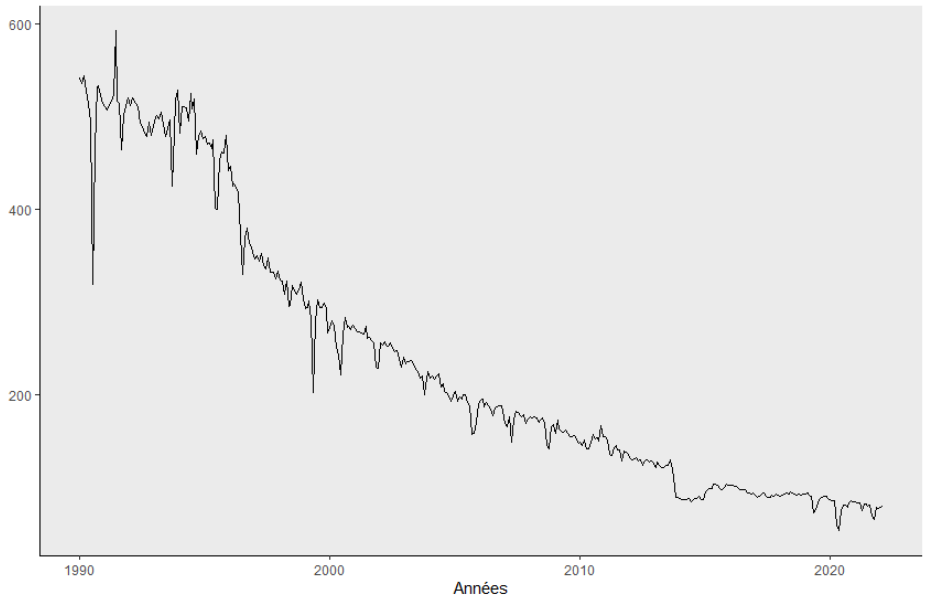
I Préparation des données

1 Description de la série

Notre série représente l'indice de la production industrielle en ce qui concerne l'extraction d'hydrocarbures.Disponible sur le site de l'INSEE, elle est corrigée des variations saisonnières et du nombre de jours ouvrables. Soixante-quatre gisements pétroliers et gaziers sont aujourd'hui en exploitation. Leur superficie totale représente environ 4 000 km2, principalement dans le Bassin aquitain et dans le Bassin parisien. La plupart de ces gisements ont été mis en production depuis 1980. La production est de 0,8 million de tonnes de pétrole et de 0,16 milliard de m3 de gaz en 2015.

Nous disposons d'observations mensuelles depuis janvier 1990 jusqu'en février 2022 soit 386 observations.

FIGURE 1 – Evolution de l'indice de production d'hydrocarbures au cours du temps

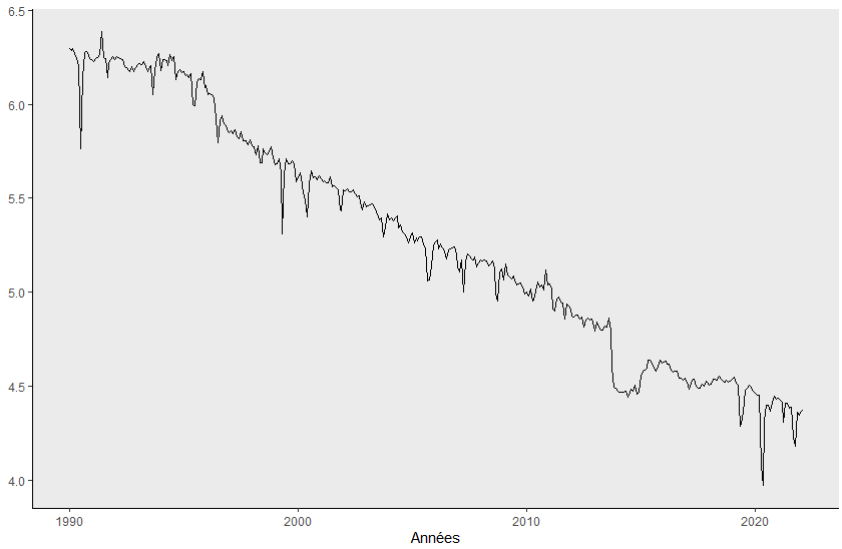


On voit clairement une décroissance linéaire sur cette série, ce qui nous laisse à penser que la série n'est pas stationnaire. La première étape consiste à essayer de corriger le caractère hétéroscédastique de la série en lui appliquant une transformation logarithmique ; cela a pour effet de « lisser » les pics. La présence d'une tendance est confirmé par une simple régression linéaire où on voit qu'il existe un coefficient directeur statistiquement non nul (régression disponible en annexe).

2 Stationnarisation et représentation graphique

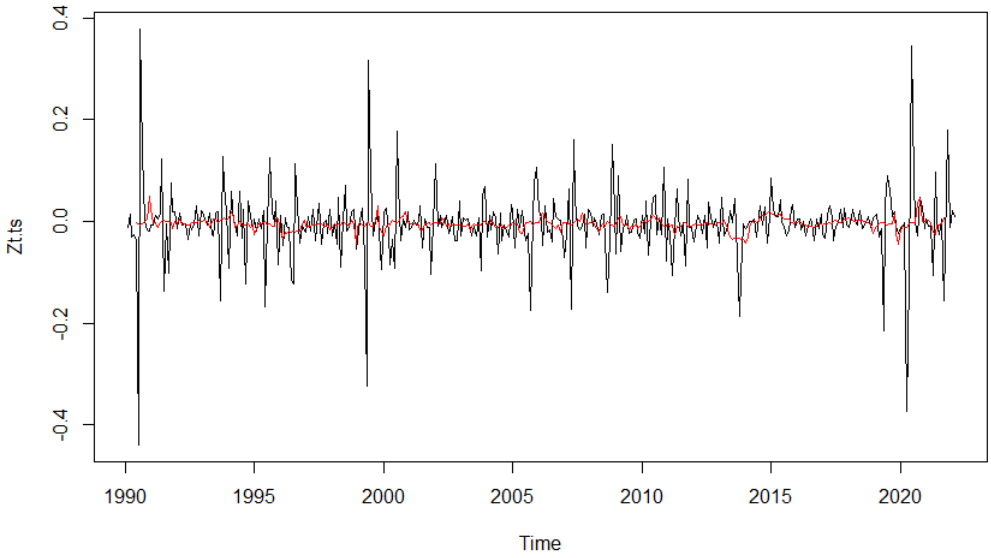
Nous effectuons donc les transformations évoqués : le passage au log transforme la série *Xt.ts* en *Gt.ts*.

FIGURE 2 – Transformé logarithmique de la série initiale



En affichant l'ACF de la log-série, nous nous aperçevons d'une décroissance linéaire des autocorrélations, ce qui nous indique qu'une différenciation est à effectuer.Nous effectuons donc une différenciation à l'ordre 1, et nous obtenons une série qui semble stationnaire à vue d'oeil. Nous appellerons cette série *Zt.ts*. Nous calculons également la moyenne mobile, donné par l'équation : $\bar{x}_n = \frac{1}{\mathrm{N}} \sum_{k=0}^{N-1} x_{n-k}$. On obtient donc notre série de travail :

FIGURE 3 – Série log-différenciée



3 Tests de stationnarité

Notre série de travail est donc exprimée de la façon suivante $Z_t = (G_t) - (G_{t-1})$, avec $G_t = \log(X_t)$.

Il convient maintenant de faire les tests de stationnarité de la série afin de confirmer notre hypothèse. Nous allons effectuer les tests classiques de racine unité : Augmented Dickey-Fuller, Phillips-Perron ainsi que le test de stationnarité KPSS. Nous rappelons que pour les deux premiers tests, l'hypothèse nulle est la présence de racine unité et donc la non stationnarité de la série alors que pour le test KPSS, l'hypothèse nulle est l'hypothèse de stationnarité de la série.

Les deux tests de racine unitaire (ADF et PP) nous ont donné une p-value inférieur à 0.01, ce qui nous permet de rejeter H0 au seuil de 1% en faveur de l'hypothèse alternative de stationnarité de la série. Concernant le test KPSS, nous ne parvenons pas à rejeter l'hypothèse nulle au niveau 0.1. Il n'est donc pas nécessaire d'effectuer une différentiation supplémentaire.

Ainsi, nous pouvons supposer pour la suite de l'étude que notre série est stationnaire. On pourra se référer en annexe pour obtenir les résultats des tests.

II Modélisation ARMA et ARIMA

4 Modèle ARMA

Dans cette partie, nous allons tenter de trouver le modèle ARMA(p,q) qui correspond le mieux aux données de notre série différenciée, que l'on suppose maintenant stationnaire. Afin de sélectionner le modèle approprié, nous allons tout d'abord regarder la fonction d'autocorrélation (ACF) et la fonction d'autocorrélation partiel (PACF) de la série différenciée. Cela va nous permettre de trouver les ordres p et q maximales du modèle ARMA(p,q). Pour p, nous regarderons la PACF (partie AR) tandis que nous regarderons l'ACF pour q (partie MA).

D'après les graphiques ACF , nous décidons de choisir $p_{max} = 3$ et $q_{max} = 2$, puisque une propriété des MA(q) est que ρ (h) = 0 si h > q, avec ρ(h) l'autocorrélation de la série à l'ordre h.De la même façon, dans un modèle AR(p), on a que r(h) = 0 si h > p avec r(h) l'autocorrélation partielle de la série. On retrouve les ACF et PACF correspondants en annexe. Nous pouvons voir dans l'annexeque l'ACF et la PACF ne présentent plus de "piques" significativement différents de zéro respectivement au delà de 2 et 3 retards. Ainsi, nous estimons tous les modèles ARMA(p,q) avec $0 \leq p \leq 3$ et $0 \leq q \leq 2$.

Afin d'obtenir le meilleur modèle, nous allons utiliser les critères d'informations d'Akaike (AIC) ainsi que le critère critère d'information bayésien (BIC). Nous choisirons le modèle qui minimisent ces deux critères. Nous avons donc le choix entre un ARMA(1,2) et un ARMA(0,2) en fonction du critère choisie.

Table 1 – Matrices de minimisation

Table 2 – AIC

	q=0	q=1	q=2
p=0	FALSE	FALSE	FALSE
p=1	FALSE	FALSE	TRUE
p=2	FALSE	FALSE	FALSE
p=3	FALSE	FALSE	FALSE

Table 3 – BIC

	q=0	q=1	q=2
p=0	FALSE	FALSE	TRUE
p=1	FALSE	FALSE	FALSE
p=2	FALSE	FALSE	FALSE
p=3	FALSE	FALSE	FALSE

Il est important de noter que le critère BIC pénalise davantage l'inclusion de paramètre supplémentaire que le critère AIC. C'est la raison pour laquelle le critère BIC conduit à des modèles plus parcimonieux, comme c'est également le cas ici.Nous choisissons donc le modèle ARMA(0,2).

5 Validité et paramétrage du modèle

Nous effectuons un test de significativité des coefficient de notre modèle ARMA(0,2) par un test de Student. Ce test a pour hypothèse nulle la nullité des coefficients. Nous avons renseigné dans le tableau

de la figure 4 ci-après les résultats de ce test. Au vue des p-value des coefficients, nous pouvons rejeter l'hypothèse nulle à tous les seuils. Nos coefficients sont donc bien significatifs.

TABLE 4

ma1	ma2	intercept
0	0	0.00003

Enfin, nous devons vérifier la blancheur des résidus de notre ARMA(0,2) afin d'être certain qu'ils correspondent bien à un bruit blanc gaussien. Pour cela, nous pouvons réaliser le test du Portemanteau/Ljung-box. L'hypothèse nulle H0 de ce test est la présence d'un bruit blanc fort. On ne rejette pas H0 à 10%. L'hypothèse n'est pas rejetée à un niveau acceptable ce qui semble bien confirmer la blancheur des résidus du modèle.

TABLE 5 – Test de Portmanteau pour un ARMA(0,2)

lags	statistic	p-value
5	3.861	0.312
10	6.007	0.662
15	11.708	0.551
20	13.823	0.756
25	16.383	0.845
30	17.276	0.950

De plus, afin d'obtenir la prédiction la plus fiable possible, il est intéressant d'avoir des résidus normaux (donc gaussien).On effectue également un test de Shapiro dont l'hypothèse nulle H0 est la normalité des résidus, hypothèse qu'on rejette à 1%.

Finalement on obtient le modèle ARMA suivant : $Z_t = -0.005 + \varepsilon_t - 0.2906\varepsilon_{t-1} - 0.3392\varepsilon_{t-2}$

6 Modèle ARIMA

Si nous revenons à la log-série choisie G_t, le modèle correspondant est donc le modèle ARIMA (0,d,2). Or, nous avons différencié la série initiale une fois pour obtenir la série Z_t. Donc d = 1. Ainsi, le modèle correspondant à la log-série que nous avons choisie initialement est le modèle ARIMA(0,1,2).

III Prévision

7 Région de confiance sur les valeurs (X_{T+1}, X_{T+2})

Notre série différenciée est le ARMA(0,2) suivant : $Z_t = -0.005 + \varepsilon_t - 0.2906\varepsilon_{t-1} - 0.3392\varepsilon_{t-2}$. Par simplification, notons $Z_t = -0.005 + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2}$

Notons $\hat{Z}_{t+h}$ la prévision optimale de Z_{t+h}.

Calculons les erreurs de prédiction :

$$\begin{pmatrix} Z_{t+1|t} \\ Z_{t+2|t} \end{pmatrix} - \begin{pmatrix} \hat{Z}_{t+1|t} \\ \hat{Z}_{t+2|t} \end{pmatrix} = \begin{pmatrix} -0.005+\varepsilon_{t+1} + \phi_1\varepsilon_t + \phi_2\varepsilon_{t-1} - (-0.005 + \phi_1\varepsilon_t + \phi_2\varepsilon_{t-1}) \\ -0.005+\varepsilon_{t+2} + \phi_1\varepsilon_{t+1} + \phi_2\varepsilon_t - (-0.005+\phi_2\varepsilon_t) \end{pmatrix} = \begin{pmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+2} + \phi_1\varepsilon_{t+1} \end{pmatrix}$$

De plus, Z_t est stationnaire et les racines du polynôme $\phi(X)$ sont en dehors du cercle unité. Donc ε_t est l'innovation linéaire de Z_t.

$$\begin{pmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+2} + \phi_1\varepsilon_{t+1} \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$$

avec $\Sigma = \begin{pmatrix} \sigma^2 & \phi_1\sigma^2 \\ \phi_1\sigma^2 & \sigma^2 + \phi_1^2\sigma^2 \end{pmatrix}$ où σ^2 est la variance des ε_t

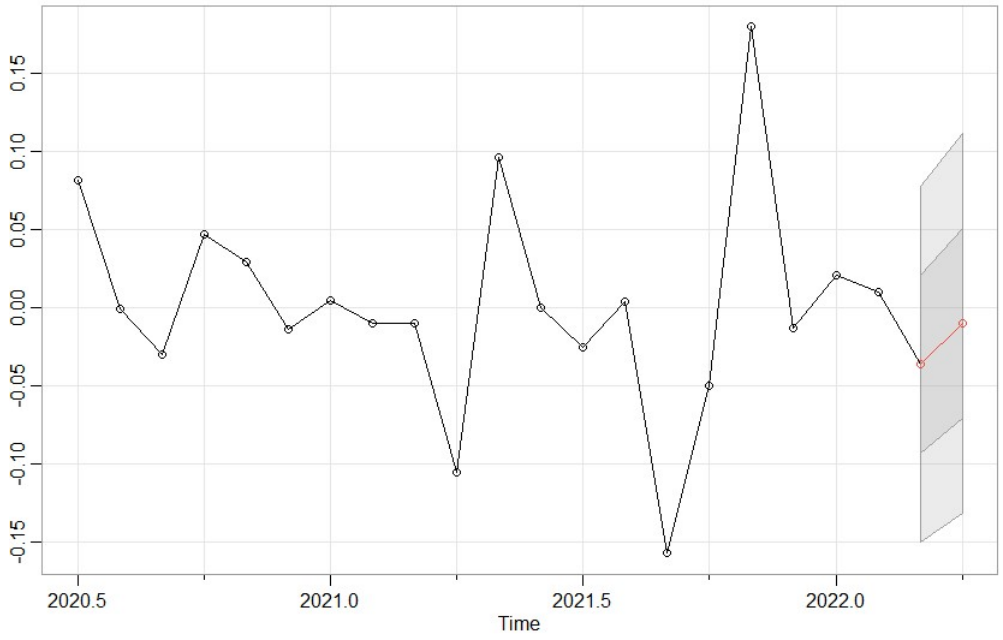
Or, $(Z_{t+1}\text{-}\hat{Z}_{t+1})'\ \Sigma^{-}1(\mathrm{Z}_{t+1}\text{-}\hat{Z}_{t+1}) \sim\chi^2_2$.
Donc notre intervalle de confiance est : $(\mathrm{X} \in R^2 \mid (Z_{t+1}\text{-}\hat{Z}_{t+1})'\ \Sigma^{-}1(\mathrm{Z}_{t+1}\text{-}\hat{Z}_{t+1}) \leq q^{1-\alpha}_{\chi^2_2}\ )$
Où $q^{1-\alpha}_{\chi^2_2}$ est le quantile d'ordre 1-α d'une loi χ^2_2.
C'est-à-dire $(\mathrm{X} \in R^2 \mid \begin{pmatrix}\varepsilon_{t+1} & \varepsilon_{t+2}+\phi_1\varepsilon_{t+1}\end{pmatrix}\ \Sigma^{-1}\begin{pmatrix}\varepsilon_{t+1}\\ \varepsilon_{t+2}+\phi_1\varepsilon_{t+1}\end{pmatrix} \leq q^{1-\alpha}_{\chi^2_2}\ )$

8 Hypothèses pour obtenir la région de confiance

Première hypothèse : Normalité des résidus. Comme évoqué précedemment, nous avons testé l'hypothèse de normalité des résidus avec le test de Shapiro avec l'hypothèse nulle de normalité des résidus que le test rejette à 1
Deuxième hypothèse :Le modèle est adapté et les paramètres sont correctement estimés.
Troisième hypothèse : Les erreurs sont des innovations et ne sont pas corrélées entre elles et aux valeurs passées de la série. Pour rappel, notre série est $Z_t = -0.005 + \varepsilon_t - 0.2906\varepsilon_{t-1} - 0.3392\varepsilon_{t-2}$. Ses racines sont -1,717 et 1,717 sont bien en dehors du cercle unité.

9 Représentation graphique de la région de confiance

Prévision et intervalle de confiance à 95% pour notre ARIMA(0,1,2) à 95%



10 Question ouverte

Pour que la connaissance de Y_{t+1} soit utile à la prédiction de X_{t+1} il faut que Yt cause instantanément Xt au sens de Granger. Autrement dit, $\hat{X}_{t+1|X_u,Y_u,u\leq t\cup Y_{t+1}} \neq \hat{X}_{t+1|X_u,Y_u,u\leq t}$
Pour vérifier que Yt cause instantanément Xt au sens de Granger, on doit trouver une corrélation entre leurs résidus dans un modèle VAR. Helmut Lütkepohl [1] montre également que l'on peut donner une définition de la causalité instantanée à l'aide des processus d'innovations respectifs des deux séries : On a $C_{X-Y<>}\ \mathrm{Corr}\left(X_{t+1} - X_{t+1} \mid (\widehat{X_u, Y_u}, u \leq t), Y_{t+1} - Y_{t+1} \mid (\widehat{X_u, Y_u}, u \leq t)\right) = 0$. On peut tester cela par un test de Wald sur la matrice de covariance (et celle de White si les innovations ne sont pas supposées gaussiennes). Il est également intéressant de voir les limites de cette analyse qui ne prend pas en compte les séries dans un système multivarié avec plus de deux séries qui peuvent s'inter-causer, on

1. New Introduction to Multiple Time Series Analysis, Helmut Lutkepohl p.45-47

parle de "liens indirects" qui la plupart du temps s'établissent sur plusieurs périodes et ne relèvent donc plus de la causalité instantanée [2] . De plus, il est à noter que le choix d'une décomposition mensuelle du temps peut avoir son importance car la causalité instantanée peut être trimestrielle (cela peut sembler paradoxal mais reste rigoureux mathématiquement). Enfin, il est possible que la série désaisonnalisée réponde à la causalité de Granger mais que la série initiale non, bien que cela dépasse la question posée ici [3]

2. Amélioration de la prévision et causalité entre deux séries d'une système multivarié autorégressif stationnaire, C. Bruneau et J-P. Nicola ´ ¨ı, Annales d'econométrie et de statistiques

3. . New Introduction to Multiple Time Series Analysis, Helmut Lutkepohl p.48-50

IV Annexe

11 Tests

Table 6 – Régression linéaire de Xt.ts par rapport au temps

	Dependent variable :
	Valeur
date	−0.041***
	(0.001)
Constant	776.043***
	(9.928)
Observations	386
R^2	0.891
Adjusted R^2	0.891
Residual Std. Error	48.667 (df = 384)
F Statistic	3,141.264*** (df = 1 ; 384)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Table 7 – Test de Dickey-Fuller pour la série log-différenciée

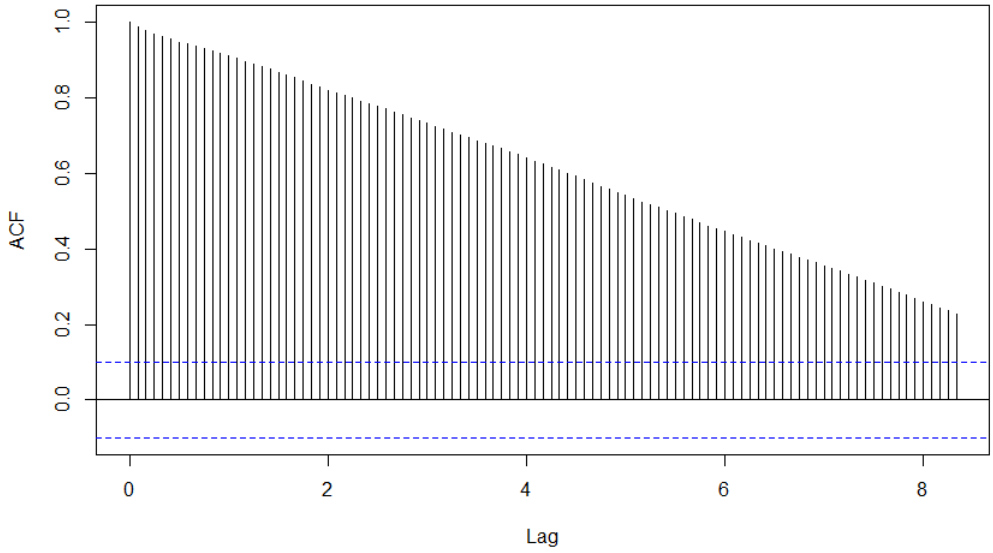
lag	ADF	p.value
0	-22.269	0.010
1	-19.008	0.010
2	-15.748	0.010
3	-13.453	0.010
4	-12.111	0.010
5	-12.316	0.010

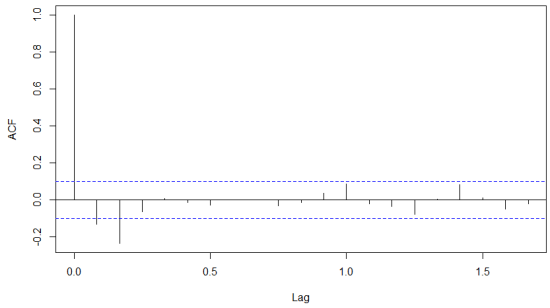
Table 8 – Test de Phillips-Perron pour la série log-différenciée

	lag	Z	p.value
type 1	5	-340.841	0.010
type 2	5	-336.754	0.010
type 3	5	-336.772	0.010

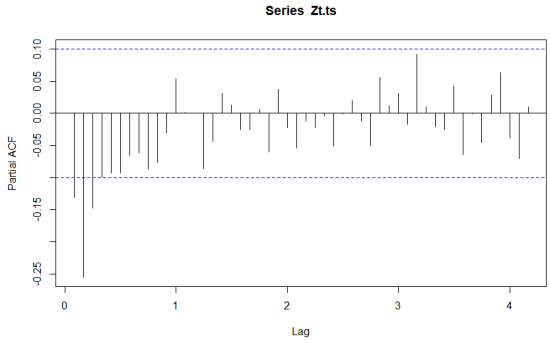
12 ACF et PACF

FIGURE 4 – ACF de la log-série





(a) ACF



(b) PACF

FIGURE 5 – ACF et PACF de la série log-différenciée