École Nationale de la Statistique et de l'Administration économique

Compte rendu de Séries Temporelles

Sorba Marianne & Vincent-Cuaz Cédric

Indice de Production et distribution d'eau, assainissement, gestion des déchets et dépollution

Encadré par Benjamin Schannes

Partie 1 : Les données

Question 1 Notre série temporelle est extraite des données nationales de l'INSEE sur la production industrielle. C'est une série brute agrégée qui représente l'indice de production et distribution d'eau, assainissement et gestion des déchets et dépollution en France métropolitaine de Janvier 1990 à Mars 2017. La série temporelle brute initiale est représentée ci-dessous sur la Figure 1.

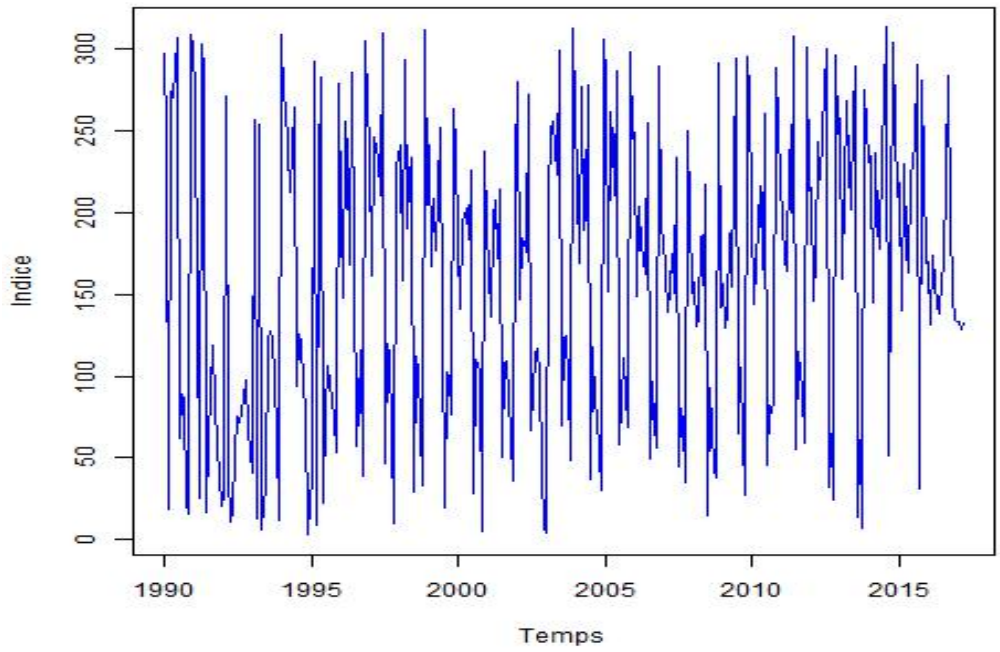


FIGURE 1 – Série temporelle brute initiale : Indice de production et distribution d'eau, assainissement, gestion des déchets et dépollution

Nous n'observons pas une décroissance/croissance visible de la variance. Il est donc inutile d'effectuer une transformation exponentielle ou logarithmique qui aurait pu supprimer une variance clairement croissante ou décroissante. La variance semble pourtant non constante dans le temps, avec des périodes marquées par une plus grande variance. C'est pour cela que nous avons choisi d'utiliser la transformation Box-Cox pour corriger la variance. Elle permet de rendre les données homoscédastiques et s'écrit sous cette forme :

$$B(x) = \frac{x^\lambda - 1}{\lambda} \quad \text{si} \quad \lambda \neq 0$$

$$B(x) = log(x) \quad \text{sinon}$$

La détermination du paramètre λ s'effectue à l'aide de la méthode du maximum de vraisemblance. Le paramètre λ correspond au pic de la courbe des maximums de vraisemblance, représentée en Figure 2.

Nous trouvons un paramètre $\lambda = 0.844$. Après avoir transformé notre série, pour y voir plus clair, nous avons effectué une décomposition saisonnière de notre série temporelle présente en Annexe 1. La série transformée est représentée en Annexe 2. Il y apparaît très clairement une non-stationnarité

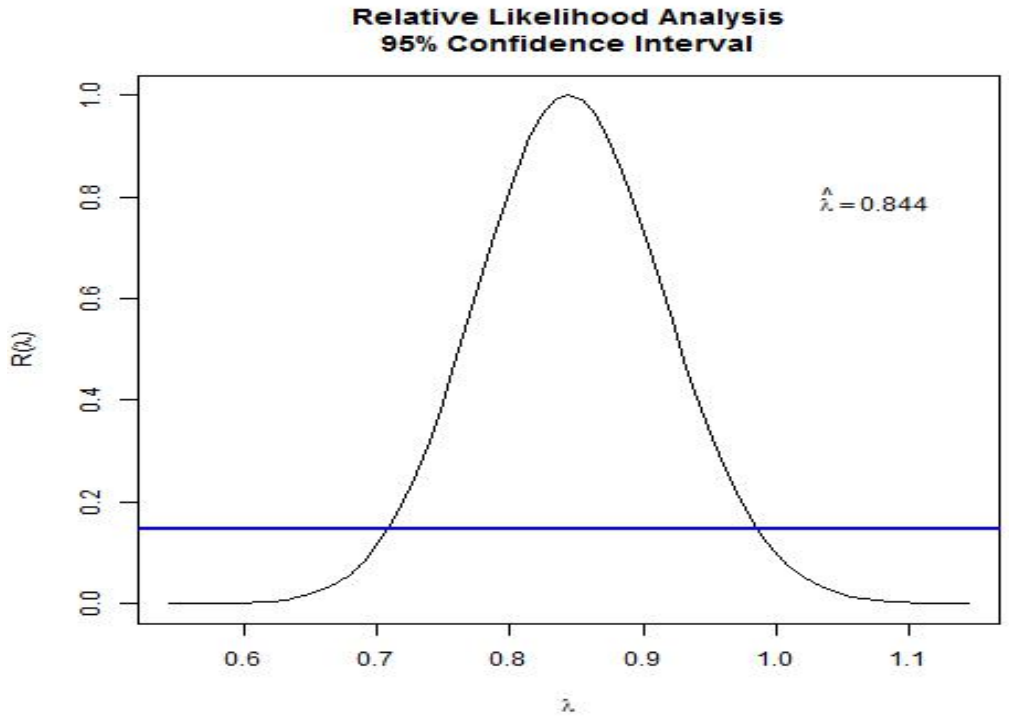


FIGURE 2 – Détermination du paramètre lambda optimal

déterministe : on observe une saisonnalité de 12 mois (seasonal). Cette première observation met en évidence la nécessité d'appliquer une différenciation saisonnière à notre série.

Question 2 Afin de confirmer les hypothèses faites sur la saisonnalité et la stationnarité de notre série, il convient de tracer la fonction d'autocorrélation et d'autocorrélation partielle de la série (Figure 3). La série n'est pas stationnaire : les pics significatifs répétés tous les 12 lags mettent en évidence la présence d'une saisonnalité annuelle (à l'ordre 12). Nous nous sommes débarrassés de cette saisonnalité en appliquant l'opérateur de différenciation saisonnière à notre série :
$\Delta_{12} X_t = X_t - X_{t-12}$

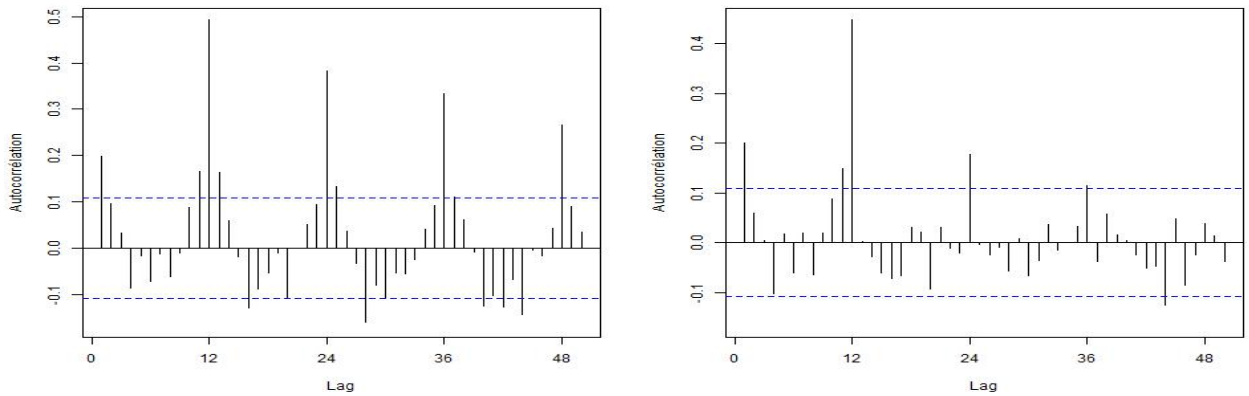


FIGURE 3 – Autocorrélogramme et autocorrélogramme partiel de la série transformée

Pour vérifier que la série obtenue est bien stationnaire, nous avons effectué un test de Dickey-Fuller augmenté ainsi qu'un test de Phillips–Perron afin de tester la présence d'une racine unitaire, synonyme de non-stationnarité de la série. Les résultats de ces tests sont présentés en Annexe 3. Les deux tests nous ont donné une p-value inférieur à 0.01 : nous avons donc rejeté au seuil de 1% l'hypothèse H_0 de non stationnarité de notre série désaisonnalisée en faveur de l'hypothèse alternative H_1 de stationnarité. Une différenciation supplémentaire n'est donc pas nécessaire. Afin de ne pas sur-différencier notre série, nous avons retenu la série transformée et désaisonnalisée pour la suite de notre projet.

Question 3 La série stationnaire que nous obtenons après transformation est tracée en Figure 4. Elle ne présente plus de saisonnalité ni de tendance visible. Elle présente néanmoins toujours une hétéroscédasticité visible non regulière, qui aurait pu être supprimée par un modèle GARCH. Ne ne l'avons cependant pas utilisé parce qu'il ne figure pas dans notre programme. Il s'agit donc maintenant d'entraîner un modèle de série temporelle $ARMA(p, q)$ à notre série.

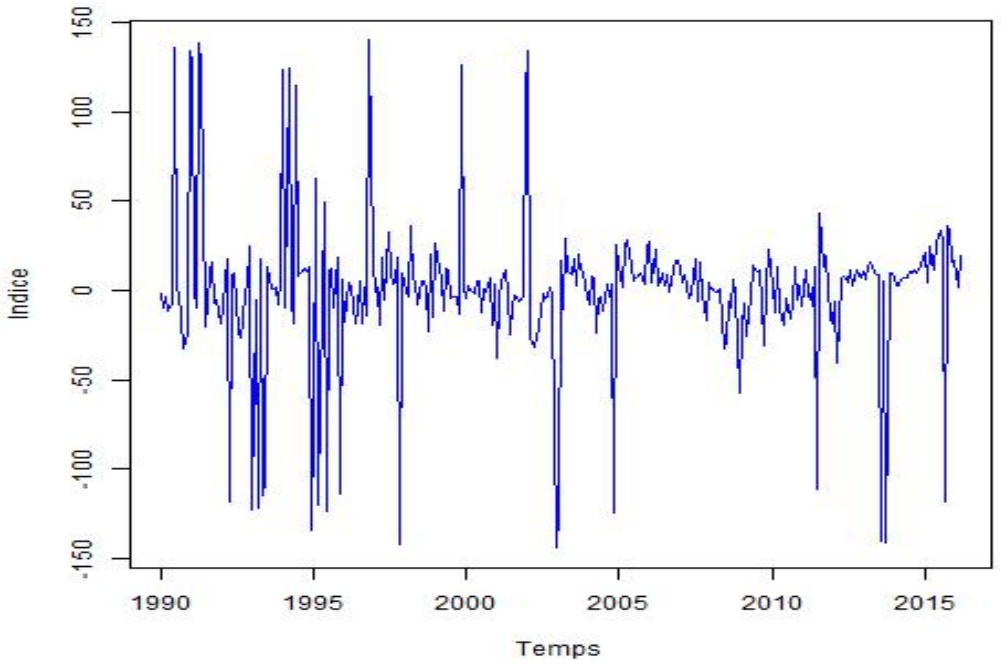


FIGURE 4 – Série transformée et désaisonnalisée

Partie 2 : Modèle ARMA

4. Pour identifier les paramètres optimaux p et q de notre modèle $ARMA(p, q)$, nous avons étudié les autocorrélations et les autocorrélations partielles de notre série corrigée.

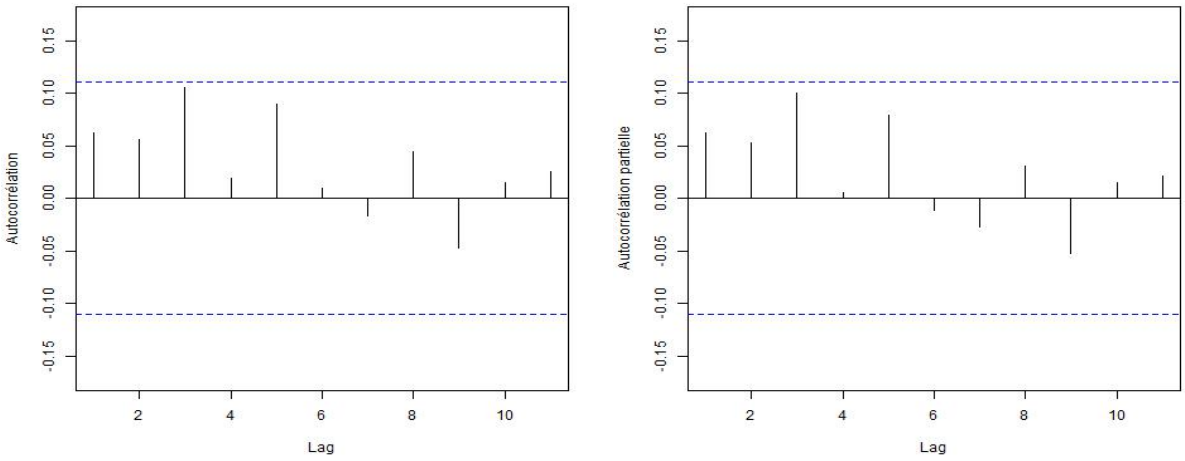


FIGURE 5 – Autocorrélogramme (gauche) et autocorrélogramme partiel (droite) de la série corrigée

Nous avons commencé par rechercher les paramètres maximaux p_{max} et q_{max}. Nous savons que dans un modèle $MA(q)$, l'autocorrélation de la série à l'ordre h est nulle pour tout $h \geq q+1$. Ainsi, q_{max} correspond au lag à partir duquel l'autocorrélation de la série n'est plus significativement différente de zéro.

De même dans un modèle $AR(p)$, l'autocorrélation partielle de la série à l'ordre h est nulle pour tout $h \geq p+1$. Ainsi, p_{max} correspond au lag à partir duquel l'autocorrélation partielle de la série n'est plus significativement différente de zéro.

Le dernier pic le plus proche de la significativité au seuil de 5% de l'autocorrélogramme correspond à un lag de 3, celui de l'autocorrélogramme partiel correspond à un lag de 3. Nous avons donc retenu $p_{max} = 3$, $q_{max} = 3$.

Après avoir établi p_{max} et q_{max}, il s'agit de tester la validité des différents modèles et d'appliquer le principe de parcimonie en choisissant parmi les modèles valides le modèle qui minimise le critère d'information AIC (Akaike Information Criterion) ou BIC (Bayesian Information Criteria). Pour ce faire, pour tout couple $(i, j) \in [\![0, p_{max}]\!] \times [\![0, q_{max}]\!]$ nous avons ajusté un $ARMA(i, j)$ à notre série corrigée. Nous avons en premier lieu testé la significativité de leurs coefficients à l'aide d'un test de Student : seul l'$ARMA(1, 1)$ a tous ses coefficients significatifs. Nous l'avons donc retenu pour la suite bien que l'$ARMA(1, 1)$ ne soit pas le modèle qui minimise le critère AIC comme le montre la Table 1.

Afin de vérifier si les résidus de notre $ARMA(1, 1)$ correspondent bien à un bruit blanc gaussien, nous avons réalisé un test de blancheur des résidus (test du Portemanteau/Ljung-Box test) : L'hypothèse H_0 d'indépendance des résidus ne peut pas être rejetée au seuil de 5%. Les résultats du test sont représentés en Annexe 4.

q \ p	0	1	2	3
0	3241.417	3242.311	3243.682	3242.702
1	3242.204	3241.431	3242.928	3243.852
2	3243.343	3243.009	3237.091	3236.987
3	3242.214	3243.958	3237.067	3239.227

TABLE 1 – AIC pour différents modèles $ARMA(p, q)$

	Estimate	Std.Error	z value	$\mathbb{P}(\geq \|z\|)$
ar1	0.748378	0.141939	5.2725	$1.346e-07$
ma1	-0.673973	0.154458	-4.3635	$1.280e-05$
intercept	-0.051576	2.984813	-0.0173	0.9862

TABLE 2 – Coefficients du modèle $ARMA(1, 1)$

Les coefficients de notre $ARMA(1, 1)$ sont présentés dans la Table 2. La constante de notre modèle n'est pas significative. En supposant que notre série corrigée (X_t) suit un $ARMA(1, 1)$, on a finalement :

$$X_t = \epsilon_t + 0.748X_{t-1} - 0.674\epsilon_{t-1}$$

Partie 3 : Prévision

Question 5 Nous considèrons que T est la longueur de notre série et que les résidus sont gaussiens tels que $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. Pour un modèle ARMA(1,1) :

$$X_t = \phi X_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

Sachant que $\mathbb{E}[e_{T+h}|T] = 0 \; \forall h > 0$, les prévisions optimales en T sont données par :

$$\begin{cases} \hat{X}_{T+1|T} = \phi X_T + \theta\epsilon_{T-1} \\ \\ \hat{X}_{T+2|T} = \phi\hat{X}_{T+1|T} = \phi^2 X_T + \phi\theta\epsilon_T \end{cases}$$

Les variances des erreurs des prévisions en T sont données par les équations :

$$\begin{cases} V(X_{T+1} - \hat{X}_{T+1|T}) = V(\epsilon_{T+1}) = \sigma_\epsilon^2 \\ \\ V(X_{T+2} - \hat{X}_{T+2|T}) = V(\epsilon_{T+2} + (\phi + \theta)\epsilon_{T+1}) = \sigma_\epsilon^2(1 + (\phi + \theta)^2) \end{cases}$$

Ainsi si on considère $(X - \hat{X}) = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix}$, on a $(X - \hat{X}) \sim N(0, \Sigma)$ où $\Sigma = \sigma_\epsilon^2 \begin{pmatrix} 1 & \phi + \theta \\ \phi + \theta & 1 + (\phi + \theta)^2 \end{pmatrix}$.
On voit facilement que $Det(\Sigma) = \sigma_\epsilon^2$, et est donc inversible si et seulement si $\sigma_\epsilon^2 > 0$, ce que nous supposons être vrai.

Ainsi $(X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$. On en déduit que la région de confiance de niveau α $\forall \alpha \in [0; 1]$ est donnée par :

$$\left\{ X \in R^2 | (X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \leq q^{1-\alpha}_{\chi^2(2)} \right\}$$

Où $q^{1-\alpha}_{\chi^2(2)}$ est le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(2)$.

Question 6 Afin d'obtenir cette région de confiance pour un niveau α donné, nous avons dû supposer que le modèle est parfaitement connu et que les coefficients obtenus à la partie précédente sont les vrais coefficients de notre modèle. De plus nous avons supposé que le bruit blanc suivait une loi normale $\epsilon_t \sim N(0, \sigma^2_\epsilon)$ où $\sigma^2_\epsilon > 0$.

Question 7 La région de confiance obtenue à la question 6 est présentée sur la figure 6. En abscisse se trouve la prévision pour X_{t+1} et en ordonnée celle de X_{t+2}.

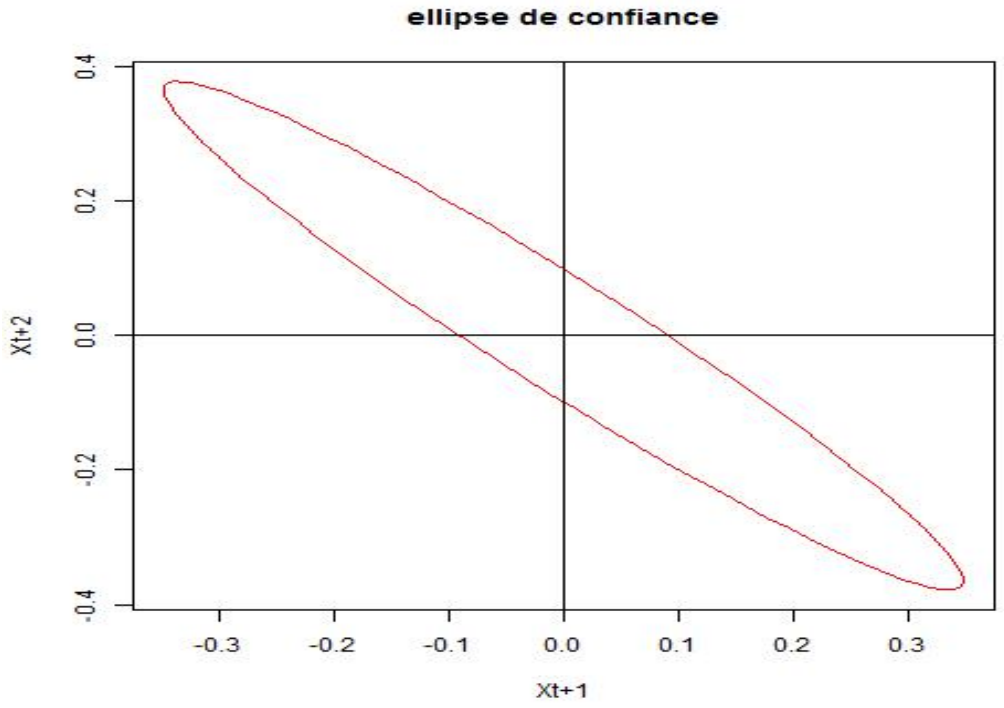
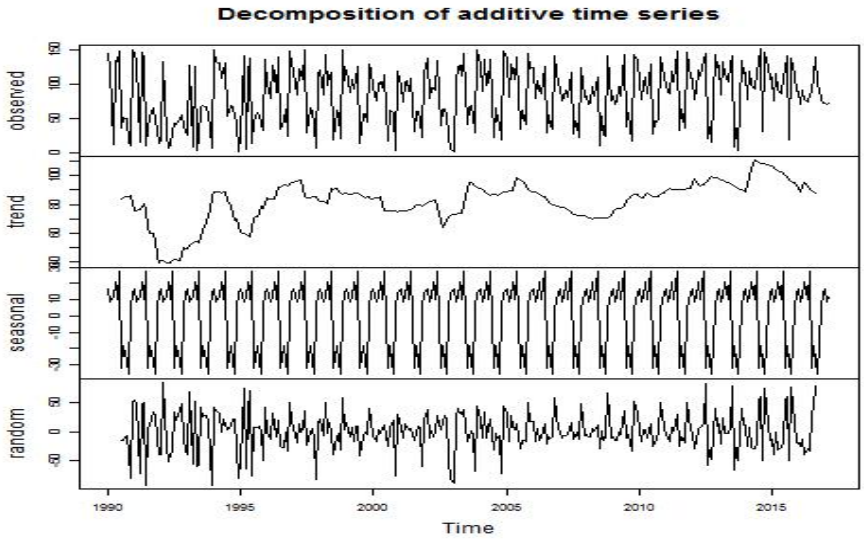


FIGURE 6 – Région de confiance de la prévision selon $ARMA(1, 1)$

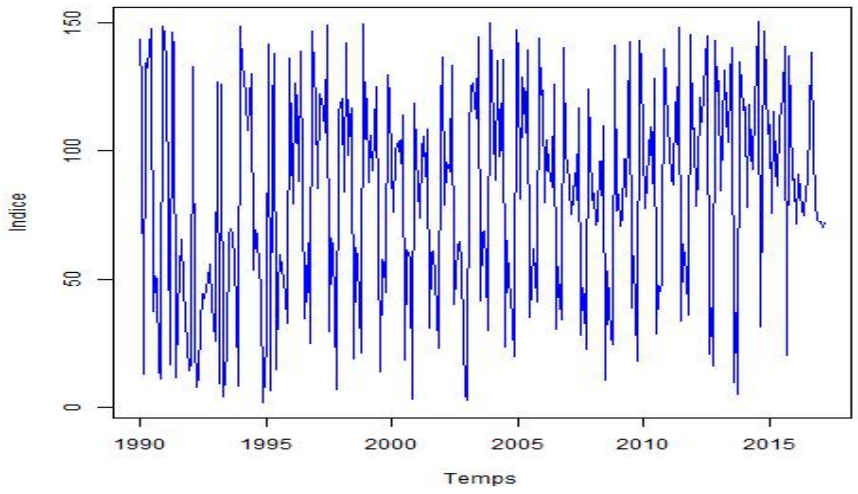
Question 8 Soit Y_t une série chronologique disponible de $t = 1$ à T . Si Y_{T+1} est disponible plus rapidement que X_{T+1}, cette information permet d'améliorer la prévision X_{T+1} si la variable (Y_t) cause instantanément la variable (X_t) au sens de Granger. C'est-à-dire que la série (Y_t) cause (X_t) si conditionnée aux valeurs passées de (Y_t) l'erreur quadratique moyenne de prédiction de X_{T+1} est inférieure à celle si (Y_t) serait omise.

Annexe

Annexe 1 – Décomposition saisonnière de la série brute



Annexe 2 – Série transformée



Annexe 3 – Tests ADF et PP sur la série corrigée

	lag order	t	$\mathbb{P}(\geq \lvert t \rvert)$
Augmented-Dickey Fuller	6	-5.9289	≤ 0.01
Phillips-Peron	5	-324.57	≤ 0.01

Annexe 4 – Tests de blancheurs des résidus sur $ARMA(1,1)$

Modèle	df	χ^2	p-value
ARMA(1,1)	1	0.17248	0.6779

Code R

```
library(foreign)
library(tseries)
library(forecast)
library(FitAR)
library(lmtest)
library(ggplot2)
df<- read.csv("C:\\Users\\maria\\Documents\\ENSAE\\Projet\\Serie temp\\Valeurs.csv",header=TRUE,sep=";")
colnames(df)<-c("Annee","Mois","Indice de Production et distribution d'eau, assainissement, gestion des dechets et depollution")
df<-df[c(-1,-2),]
jpeg('ST_brute.jpg')
time_serie<-ts(df[,"Indice de Production et distribution d'eau, assainissement, gestion des dechets et depollution"],frequency=12,start=c(1990,1))
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps")
dev.off()
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps", main="Serie temporelle brute: Indice de Production et distribution d'eau, assainissement, gestion des dechets et depollution")
#Tendance a la hausse, une composante saisonniere
#modele: X_t=a0 + a1t+st+Yt
jpeg('Box_Cox.jpg')
BoxCox(time_serie)
dev.off()
lambda=0.844
BoxCox(time_serie)
time_serie=(time_serie^0.844 - 1)/0.844
jpeg('ST_tf_Box_Cox.jpg')
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps")
dev.off()
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps", main="Serie temporelle transforme: Indice de Production et distribution d'eau, assainissement, gestion des dechets et depollution")
jpeg("Decomposition.jpg")
plot(decompose(time_serie))
dev.off()
plot(decompose(time_serie))
#saisonnalite
jpeg('ACF_ST_Brute.jpg')
Acf(time_serie,lag.max=50,ylab="Autocorrelation",xlab="Lag",main="")
```

```
dev.off()
Acf(time_serie,lag.max=50,main="ACF Serie brute",ylab="Autocorrelation",xlab="
    Lag")
#saisonnalite
jpeg('PACF_ST_Brute.jpg')
Pacf(time_serie,lag.max=50,ylab="Autocorrelation",xlab="Lag",main="")
dev.off()
Pacf(time_serie,lag.max=50,main="ACF Serie brute",ylab="Autocorrelation",xlab="
    Lag")
#operateur de differenciation saisonniere
jpeg("ST_saison.jpg")
time_serie<-time_serie- lag(time_serie,12)
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps")
dev.off()
plot.ts(time_serie,col="blue",lwd=1,xlab="Temps", main="Serie desaisonnalisee et
     tranformee")
#test de stationarite
adf.test(time_serie)#dicker fuller stationnarite
pp.test(time_serie)#phillip
kpss.test(time_serie)
#acf: On s'arrete au 4 eme pic, MA max=qmax=4
jpeg("ACF_ST_saison_BC.jpg")
Acf(time_serie,lag.max=11,ylab="Autocorrelation",xlab="Lag",main="")
dev.off()
Acf(time_serie,lag.max=11,main="ACF Serie transformee et desaisonnalisee ",ylab=
    "Autocorrelation",xlab="Lag")
#pacf: On s'arrete au 4 eme pic, MA max=qmax=4
jpeg("PACF_ST_saison_BC.jpg")
Pacf(time_serie,lag.max=11,ylab="Autocorrelation partielle",xlab="Lag",main="")
dev.off()

Pacf(time_serie,lag.max=11,main="PACF Serie transformee et desaisonnalisee ",
    ylab="Autocorrelation partielle",xlab="Lag")
for (p in 0:4){
  for (q in 0:4){
    fit<-arima(time_serie, order=c(p,0,q))
    print(Box.test(fit$residuals))#p value enorme: on ne rejette pas
    print(coeftest(fit))}#on utilise bic car beaucoup de parametre
}
mat<-matrix(nrow=4,ncol=4)
fit<-arima(time_serie,c(3,0,2))
Box.test(fit$residuals)#p value enorme: on ne rejette pas
coeftest(fit)
for (p in 0:3){
  for (q in 0:3){
    mat[p+1,q+1]<-AIC(arima(time_serie, order=c(p,0,q)))}
}

mat
min(mat)#(p=3,q=3 1278.128 ou 1279.553 (1,1) : maximise le critere
fit<-arima(time_serie,c(3,0,2))
Box.test(fit$residuals)
coeftest(fit)
fit<-arima(time_serie,c(2,0,3))
Box.test(fit$residuals)
coeftest(fit)
fit<-arima(time_serie,c(2,0,2))
Box.test(fit$residuals)
```

```
87 coeftest(fit)
```