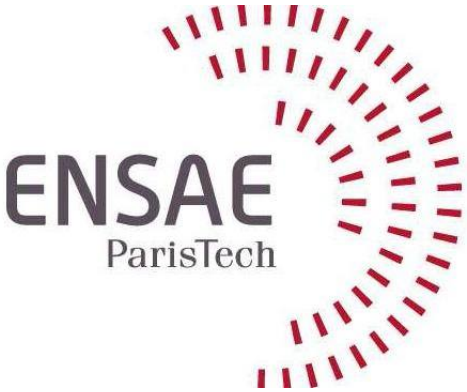


2018-2019

Projet de Séries Temporelles

Une étude menée sur l'indice de production industrielle dans le secteur de l'aéronautique

---

Auteurs :
Jeremy MARCK
Karim TIT

Encadrante :
Raphaël Lee

Table des matières

1 Les données

1.1 Série choisie

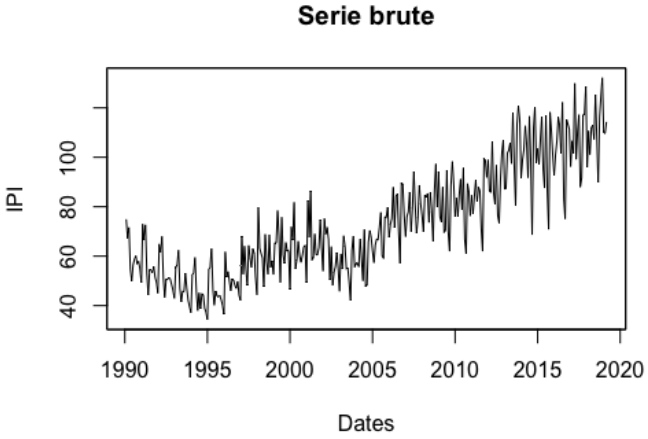
Le calcul des indices de la production industrielle répond à un impératif national et européen. Les indices de la production industrielle relèvent du règlement européen sur les statistiques de court-terme.

Les indices de la production industrielle permettent de suivre l'évolution mensuelle de l'activité industrielle de la France et de la construction. Ils représentent à ce titre une information primordiale pour le suivi du cycle conjoncturel en France et pour l'identification de points de retournement du cycle économique à un stade précoce, parallèlement ou en association avec d'autres grands indicateurs macro-économiques comme l'emploi, les indices de prix, les indices de la production dans les services, ou encore le commerce extérieur.

Par ailleurs, les indices de la production industrielle sont une des sources utilisées pour l'élaboration des comptes trimestriels français (PIB flash par exemple).

S'agissant de notre série, elle représente l'indice brut de la production industrielle (IPI) mensuelle (base 2015) dans le secteur de la construction aéronautique et spatiale en France métropolitaine, de janvier 1990 à février 2019. Aucune forme de correction n'a été apportée aux données. Il ne s'agit par exemple pas d'une série corrigée des variations saisonnières (CVS). Les données couvrent la production d'aeronefs et d'engins spatiaux, d'hélicoptères, de planeurs, de dirigeables, la construction d'avion et prennent en compte la TVA. Notre série brute est représentée ci-après.

FIGURE 1 – Serie IPI brute base 2015



La représentation de cette série suggère une non-stationnarité apparente des données pour deux raisons :

1. nous observons **une tendance déterministe** baissière puis haussière
2. **une régularité dans les variations autour de la tendance** est également observée.

1.2 Description et stationnarisation de la série

1.2.1 Description de la série choisie pour l'étude

Nous décidons de log-transformer la série brute afin d'attenuer l'effet de variations autour de la tendance. Cette série log-transformée est représentée en annexe.

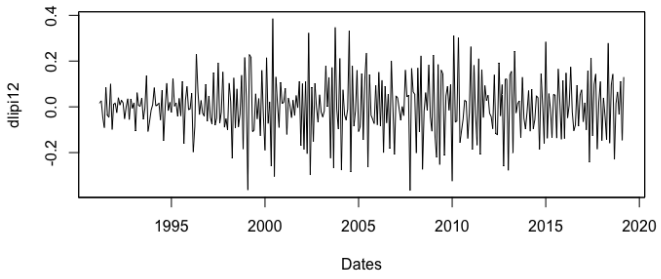
Identification de la saisonnalité : Nous analysons l'ACF de la série log-transformée, représentée en annexe. Cette dernière caractérise une saisonnalité d'ordre 12 (donc une saisonnalité annuelle car nous avons fait le choix d'une segmentation mensuelle de nos données).
Il est à noter que l'échelle en abscisses s'interpréte de la manière suivante : nous avons segmenté mensuellement nos données, et donc 1 correspond à 12/12. **Une saisonnalité d'ordre 12 est bien observée.**

Si notre série log-transformée et désaisonnalisée est stationnaire, alors nous pourrons l'approcher par un modèle de type SARIMA$(p,0,q)(P,1,Q)_{12}$. Cette stationnarité est étudiée par trois tests complémentaires présentés dans la section qui suit. S'agissant de la série log-transformée et désaisonnalisée, elle ne passe aucun des tests mis en place (ADF,PP et KPSS). Nous décidons alors de la différencier une fois et nous menons les tests sur cette série.

1.2.2 Tests de stationnarité

La série étudiée est donc la série log-transformée, désaisonnalisée (lag 12) et différenciée une fois.

FIGURE 2 – Serie log-transformée, désaisonnalisée et différenciée une fois



Test ADF : il s'agit du test Augmente Dickey-Fuller, ou test de racine unitaire. Dans le cadre d'une spécification auto-régressive pure, cette procédure teste la présence d'une racine unitaire, c"est à dire le fait que $\rho = 1$ dans le cas d'un modèle de type $X_t = \rho X_{t-1} + \varepsilon_t$ standard (AR(1)). Cette hypothèse est synonyme de non-stationnarité. On a donc $\mathcal{H}_0 : \rho = 1$. La p-valeur vaut 0.01 ce qui conduit à rejetter $\mathcal{H}_0$, soit l'hypothèse de non-stationnarité. Le test ADF conclut que notre série log-différenciée et désaisonnalisée est stationnaire.

Test PP (Philipps-Perron) : propose une hypothèse nulle de type $\mathcal{H}_0 : \rho = 1$ dans le cadre d'une modèle semi-paramétrique de la forme $X_t = a + bt + \rho X_{t-1} + u_t$ où u_t est un terme d'erreur très général. La mise en place de ce test conclut à la stationnarité de notre série log-différenciée et désaisonnalisée.

Test KPSS : cette fois-ci, , l'hypothèse nulle est la stationnarité. La spécification du modèle est :

$$X_t = \gamma t + r_t + \varepsilon_t \quad \text{où} \quad r_t = r_{t-1} + u_t$$

Pour $t \geq 1$ avec $= 0$ s'il n y'a pas de tendance déterministe et où r_0 sert de constante. Les u_t sont iid $(0,\sigma_u^2)$ et la nulle est $\mathcal{H}_0 : \sigma_u^2 = 0$ (stationnarité). Le test caractérise donc le processus comme la somme d'un trend déterministe, d'une marche aléatoire et d'une erreur stationnaire. Le test KPSS conduit à accepter la stationnarité de la série log-différenciée et désaisonnalisée.

2 Modèle SARIMA

Soit $(X_t)_{t\in 1,\ldots,350}$ la série log-différenciée. Nous supposons qu'elle suit un SARIMA$(p,1,q)(P,1,Q)_{12}$ et ce car nous avons vérifié la stationnarité de $Y_t = (1-B)(1-B^{12})X_t$ où B est l'opérateur Backward.

2.1 Caractérisation des retards p et q

Pour déterminer les ordres P et Q, nous retenons les derniers retards multiples de 12 significatifs des PACF et ACF respectivement. Il en va de même pour les ordres p et q mais en regardant cette fois les autocorrélogrammes pour les ordres inférieurs à 12. Les autocorrélogrammes sont disponibles en troisième graphique en annexe.

Nous retenons donc en premier lieu un SARIMA$(2,1,2)(2,1,2)_{12}$. La qualité de l'ajustement passe alors par deux étapes :

1. test sur la significativité des paramètres
2. puis test de la blancheur des résidus.

2.2 Qualité de l'ajustement

Nous menons ici les deux étapes sus-mentionnées.

2.2.1 Tests sur les paramètres

Nous avons testé plusieurs modèles et deux semblent éligibles : un SARIMA$(0,1,1)(0,1,2)_{12}$ et SARIMA$(0,1,2)(0,1,2)_{12}$. Pour obtenir une telle conclusion, nous avons mené des tests de sgignifiativité des paramètres (voir le script R en annexe). Parmi les deux modèles mentionnées ci-avant, nous retenons le deuxième de ces modèles car il présente un AIC plus faible. Le critère AIC, ou Akaike information criterion permet d'arbitrer entre ces deux modèles. L'idée est de se prémunir contre la surparamétrisation en introduisant un coût à l'introduction de chaque paramètre supplémentaire. La statistique mobilisée est la suivante :

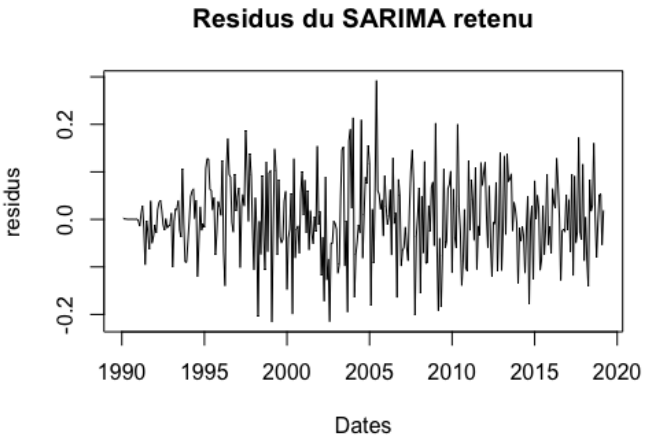
$$\text{AIC}(p,q) = \log\left(\widehat{\sigma}^2\right) + \frac{2(p+q)}{n} \quad \text{où} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} \widehat{\varepsilon}_t{}^2$$

Le modèle à retenir est celui qui présente l'AIC le plus faible (cas du modèle que nous avons retenu).

2.2.2 Tests de blancheur des résidus

L'analyse du graphique des résidus laisse à supposer que ces derniers sont assimilables à un bruit blanc. On rappelle que les résidus suivent un bruit blanc si **(1)** ils sont de moyenne nulle,**(2)** ils sont de même variance, **(3)** ils sont non autocorrélés.

FIGURE 3 – Graphe des résidus du SARIMA retenu



Pour s'en convaincre, nous menons un **test de Portmanteau (mobilisant la statistique de Ljung-Box)** qui teste l'hypothèse nulle d'absence d'autocorrélation des résidus. La statistique de test est la suivante :

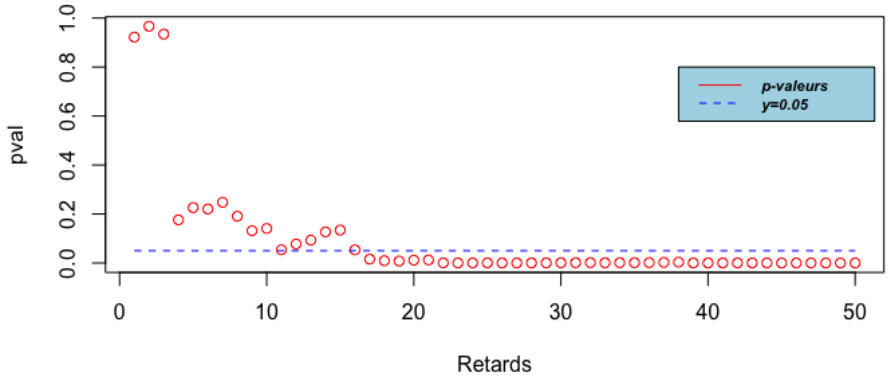
$$Q = n(n+2)\sum_{h=1}^{H} \frac{1}{n-h}\widehat{\rho}_{\widehat{\varepsilon}}{}^2(h) \quad \text{avec} \quad \widehat{\varepsilon}_t = \frac{\widehat{\Phi}(B)}{\widehat{\Psi}(B)}X_t \quad \text{et} \quad \widehat{\rho}(h) = \frac{\widehat{\gamma(h)}}{\widehat{\gamma(0)}}$$

Etant entendu que :

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{|h|+1}^{n} (X_t - X_n)\left(X_{t-|h|} - X_n\right)$$

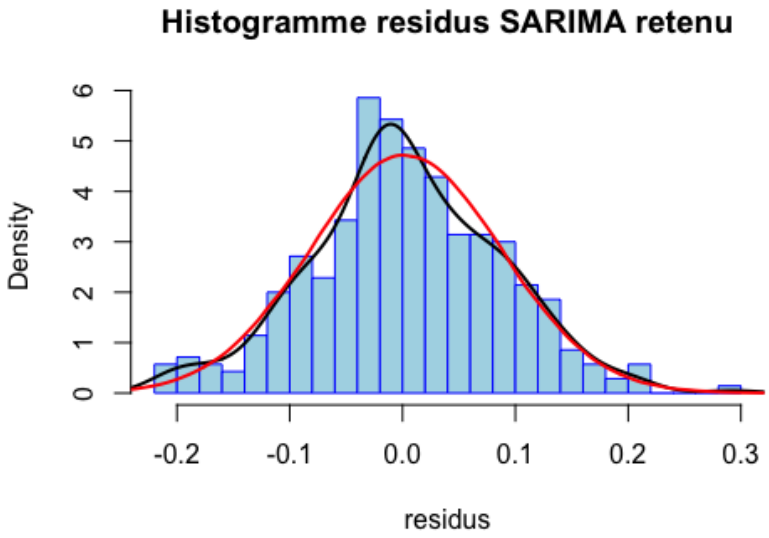
Le graphique qui suit représente la p-valeur pour le test sus-mentionné en fonction du nombre de retards.

FIGURE 4 – P-valeurs du test de Portmanteau en fonction des retards



La blancheur des résidus (hypothèse nulle) n'est pas rejetée pour des ordres inférieurs à 17 et à 5%. Elle l'est en revanche pour des ordres supérieurs. Notons que la onzième p-valeur peut paraitre ambigue : elle vaut précisément 0.054 et donc l'hypothèse nulle est acceptée à 5%.
La figure qui suit caractérise la normalité des résidus (en rouge est représentée la densité gausienne adéquate) : cela améliore la précision du test de Ljung-Box à distance finie.

FIGURE 5 – Histogramme des résidus



3 Prévision

3.1 Equation de la région de confiance

On a sélectionné le modèle $X_t \sim \text{SARIMA}(0,1,2)(0,1,2)_{12}$ ce qu'on peut écrire $(\mathbb{I}-B)\left(\mathbb{I}-B^{12}\right)X_t = \Psi(B)\Psi(B^{12})\varepsilon_t$ où $(X_t)_{t\in\{1,\dots,350\}}$ désigne la série log-transformée.
On a donc que $Y_t = (\mathbb{I}-B)\left(\mathbb{I}-B^{12}\right)X_t$ est un processus ARMA causal, donné par :

$$\begin{aligned} Y_t &= \psi(B)\Psi\left(B^{12}\right)\varepsilon_t = \left(1+\psi_1 B+\psi_2 B^2\right)\left(1+\Psi_1 B^{12}+\Psi_2 B^{24}\right) \\ &= \varepsilon_t+\psi_1\varepsilon_{t-1}+\psi_2\varepsilon_{t-2}+\Psi_1\varepsilon_{t-12}+\psi_1\Psi_1\varepsilon_{t-13}+\psi_2\Psi_1\varepsilon_{t-14}+\Psi_2\varepsilon_{t-24}+\psi_1\Psi_2\varepsilon_{t-25}+\psi_2\Psi_2\varepsilon_{t-26} \end{aligned}$$

On constate qu'en fait $Y_t \sim \text{MA}(26)$. En particulier :

$$\widehat{Y}_{T+1/T} = \mathbb{E}L\left[Y_{T+1}/\mathcal{F}_T\right] = \mathbb{E}L\left[\varepsilon_{T+1}\sum_{k=1}^{26}\psi_k\varepsilon_{T+1-k}/\mathcal{F}_T\right] = Y_{T+1} - \varepsilon_{T+1}$$

Soit encore :

$$\varepsilon_{T+1} = Y_{T+1} - \widehat{Y}_{T+1} \sim \mathcal{N}(0,\sigma^2)$$

où l'on suppose que $(\varepsilon_t)_{t\geq 0}$ est un bruit blanc fort gaussien.

De la même manière, on a :

$$Y_t = (\mathbb{I} - B)\left(\mathbb{I} - B^{12}\right)X_t \iff X_t = Y_t + X_{t-1} + X_{t-2} + X_{t-12} - X_{t-13}$$

Et donc :

$$\begin{cases} X_{T+1} = Y_{T+1} + X_T + X_{T_1} + X_{T-11} - X_{T-12} \\ \widehat{X}_{T+1/T} = \widehat{Y}_{T+1/T} + X_T + X_{T-1} + X_{T-11} - X_{T_1 2} \end{cases}$$

Et donc :

$$X_{T+1} - \widehat{X}_{T+1} = Y_{T+1} - \widehat{Y}_{T+1/T} = \varepsilon_{T+1} \sim \mathcal{N}(0,\sigma^2)$$

En utilisant le théorème central limite de Lindeberg et le théorème de Slutsky, on obtient que :

$$\text{IC}_1 = \left[\widehat{X}_{T+1/T} - \widehat{\sigma}q_{1-\alpha/2}^{\mathcal{N}(0,1)}; \widehat{X}_{T+1/T} + \widehat{\sigma}q_{1-\alpha/2}^{\mathcal{N}(0,1)}\right] \quad \text{où} \quad \widehat{\sigma} = \frac{1}{T-1}\sum_{k=1}^{T}\widehat{\varepsilon_k}^2$$

Pareillement, on a :

$$\widehat{Y}_{T+2/T} = \mathbb{E}L\left[Y_{T+2}/\mathcal{F}_T\right] = \mathbb{E}L\left[\varepsilon_{T+2} + \sum_{k=2}^{26}\psi_k\varepsilon_{T+1-k} + \psi_1\varepsilon_{T+1}/\mathcal{F}_T\right] = Y_{T+2} - \varepsilon_{T+2} - \psi_1\varepsilon_{T+1}$$

Et donc :

$$Y_{T+2} - \widehat{Y}_{T+2/T} = \varepsilon_{T+2} + \psi_1\varepsilon_{T+1}$$

Mais

$$\begin{cases} X_{T+2} = Y_{T+2} + X_{T+1} + X_T + X_{T-10} + X_{T-11} \\ \widehat{X}_{T+2/T} = \widehat{Y}_{T+2/T} + \widehat{X}_{T+1/T} + X_T + X_{T-10} = +X_{T-11} \end{cases}$$

Donc :

$$X_{T+2} - \widehat{X}_{T+2/T} = \varepsilon_{T+2} + (1+\psi_1)\varepsilon_{T+1} \sim \mathcal{N}\left(0, \sigma^2(1 + (1+\psi_1)^2)\right)$$

Par le théorème central limite de Lindeberg et le théorème de Slutsky, on obtient l'intervalle de confiance :

$$\text{IC}_2 = \left[\widehat{X}_{T+2/T} - \widehat{\sigma}q_{1-\alpha/2}^{\mathcal{N}(0,1+(1+\widehat{\psi_1})^2)}; \widehat{X}_{T+2/T} + \widehat{\sigma}q_{1-\alpha/2}^{\mathcal{N}(0,1+(1+\widehat{\psi_1})^2)}\right]$$

Et pour déterminer la région de confiance pour (X_{T+1}, X_{T+2}), on calcule la covariance de $(X_{T+1} - \widehat{X}_{T+1/T})$ et $(X_{T+2} - \widehat{X}_{T+2/T})$. Soit :

$$\text{Cov}\left(X_{T+1} - \widehat{X}_{T+1/T}, X_{T+2} - \widehat{X}_{T+2/T}\right) = \text{Cov}\left(\varepsilon_{T+2} + (1+\psi_1)\varepsilon_{T+1}, \varepsilon_{T+1}\right) = (1+\psi_1)\sigma^2$$

Donc $(X_{T+1} - \widehat{X}_{T+1/T}, X_{T+2} - \widehat{X}_{T+2/T}) \sim \mathcal{N}(0,\Sigma)$ où :

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2(1+\psi_1) \\ \sigma^2(1+\psi_1) & \sigma^2\left(1 + (1+\psi_1)^2\right) \end{pmatrix}$$

Ainsi,

$$\begin{pmatrix} X_{T+1} - \widehat{X}_{T+1/T} \\ X_{T+2} - \widehat{X}_{T+2/T} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} X_{T+1} - \widehat{X}_{T+1/T} \\ X_{T+2} - \widehat{X}_{T+2/T} \end{pmatrix} \sim \chi^2(2)$$

Une région de confiance asymptotique est donc donnée par l'équation :

$$\text{RC} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \quad \begin{pmatrix} X_{T+1} - \widehat{X}_{T+1/T} \\ X_{T+2} - \widehat{X}_{T+2/T} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} X_{T+1} - \widehat{X}_{T+1/T} \\ X_{T+2} - \widehat{X}_{T+2/T} \end{pmatrix} \leq q_{1-\alpha/2}^{\chi^2(2)} \right\}$$

On constate qu'il s'agit d'une ellipse dans le plan R^2

3.2 Hypothèses mobilisées

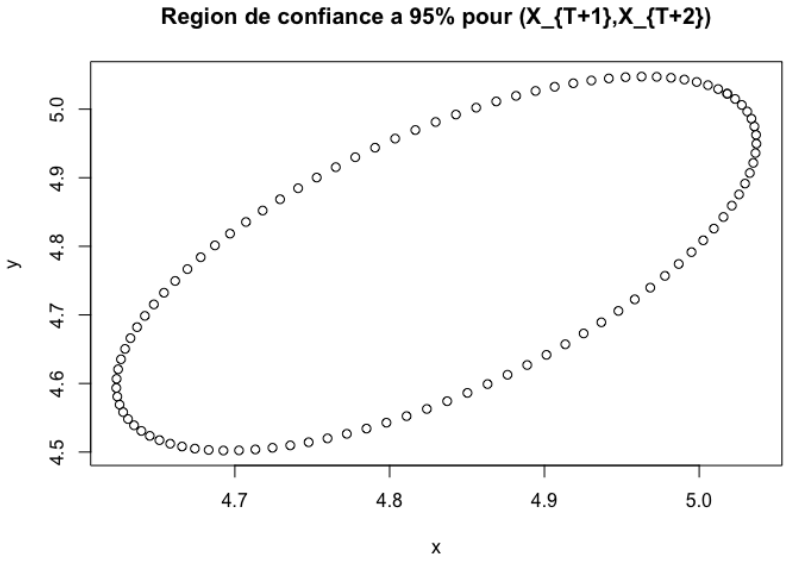
Ce sont les conditions suivantes :

1. (ε_t) est l'innovation de $X_t \sim \text{SARIMA}(0,1,2)(0,1,2)_{12}$
2. (ε_t) est bruit blanc fort gaussien
3. $\forall t \geq 1$, (Y_t) n'est pas corrélé avec valeurs initiales $(X_{-13}, X_{-12}, X_{-1}, X_{-2}, X_0)$

3.3 Représentation graphique pour $\alpha = 95\%$

On obtient la région de confiance caractérisée par l'ellipse suivante :

FIGURE 6 – Région de confiance à 95%



4 Annexe

4.1 Graphiques annexes

FIGURE 7 – Serie log-transformée

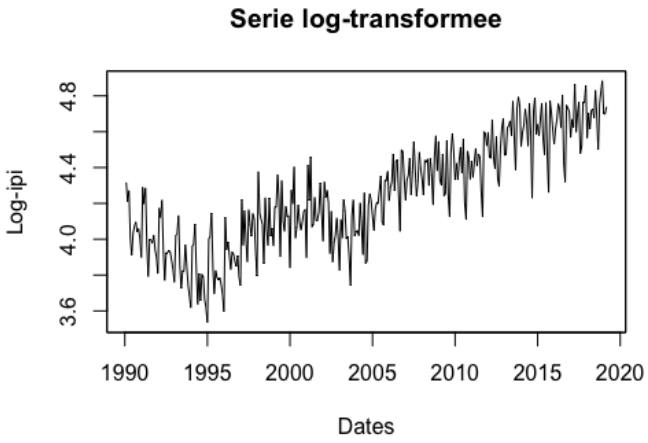


FIGURE 8 – ACF de la série log-transformée

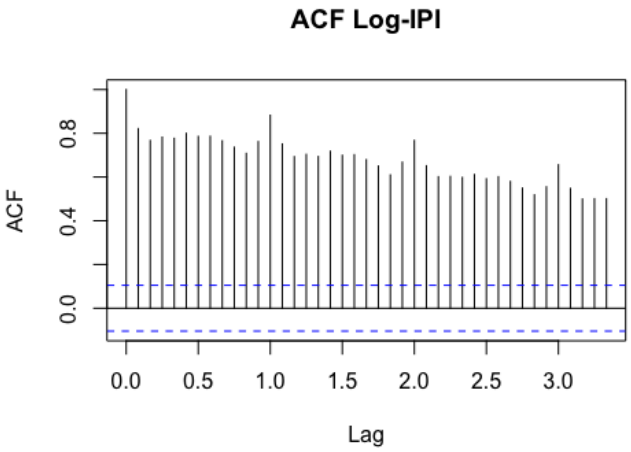
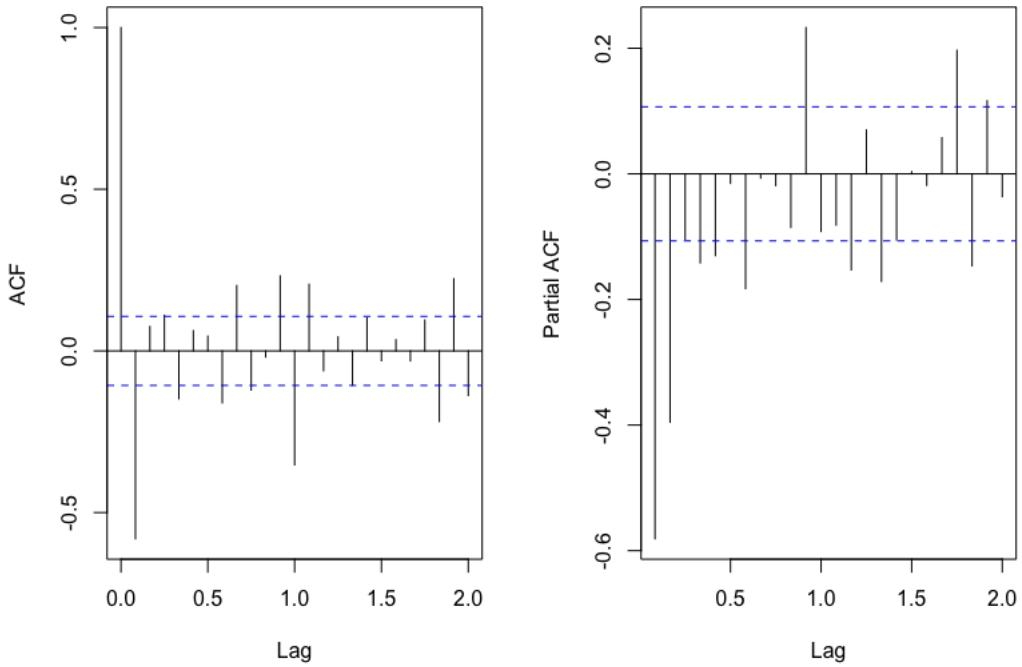


FIGURE 9 – ACF et PACF de la série log-transformée, désaisonnalisée et différenciée une fois



4.2 Script R

```
####### PROJET DE SERIES TEMPORELLES ########
################################################

rm(list=ls())

## Importation packages :
require(zoo)
require(tseries)
require(fUnitRoots)
require(stargazer)
require(urca)
require(FitAR)
require(caschrono)
require(tidyverse)
require(pracma)

# Importation des donnees :
############################
data=read.csv("/Users/jeremymarck//Desktop/Projetst/data1.csv",sep=";",header=T)
colnames(data)=c('Dates','IPI','Codes')

# Mise en place des dates :
############################
data_sources=as.character(data$Date)
ipi_source=data$IPI
start=1990+1/12
end=2019+2/12
dates_sequence=seq(from=start,to=end,by=1/12) # définit un séquencage des dates
```

```
dates=as.yearmon(dates_sequence)
ipi=zoo(ipi_source,order.by = dates)
dev.off()

# Plot de la serie brute :
############################
dev.off()
plot(ipi,main="Serie brute",ylab="IPI",xlab="Dates")
## Deux remarques :
# - presence d'un trend baissier puis haussier
# - presence d'une saisonnalite
acf(ipi)
pacf(ipi)

## Recuperation du log-ipi :
############################
lipi=log(ipi)
plot(lipi,main="Serie log-transformee",ylab="Log-ipi",xlab="Dates")
acf(lipi,lag.max=40,main="ACF Log-IPI")
pacf(lipi,lag.max=40)
# Saisonnalite d'ordre 12

## Differenciation saisonniere :
################################
lipi12=diff(lipi,12)
plot(lipi12)
acf(lipi12)
pacf(lipi12) # Saisonnalite donc on differencie


## Differenciation d'ordre 1: car sinon on n'avait pas quelque chose de stationnaire
##############################
dlipi12=diff(lipi12,1)
plot(dlipi12,xlab='Dates')
acf(dlipi12) # lagmax de 13 pour le test adf


## Tests de stationnarisation :
################################
test=ur.df(y=dlipi12,type="trend",lags=13)
summary(test)
kpss.test(x=dlipi12,null='Trend')
stargazer(a)# Apparement c'est bon
# Ici l'hypothese nulle est la stationnarite et on accepte l'hypothse nulle
help(kpss.test)

pp.test(dlipi12)
stargazer(a)# Stationnaire
plot(dlipi12)
adf.test(dlipi12)
stargazer(a)# Stationnaire
# Pour les deux derniers tests, on rejette H0 qui est l'hypothèse de racine unité
# donc on rejette la non-stationnarité.
# Ok, on a un modèle stationnaire

## Choix d'un modèle :
######################
statio=dlipi12
```

```
par(mfrow=c(1,2))
acf(statio,lag.max=24,main='')
pacf(statio,lag.max=24,main='')

# On retient un SARIMA (2,1,2)(2,1,2)
#####################################
hessian=FALSE
modele1=arima(x=lipi,order=c(2,1,1),seasonal=list(order=c(1,1,2),period=12),method='ML')
t_stat(modele1)
modele1$aic
# AR(1) pas signif et sar(1) non plus --> a enlever, on commence par sar1
modele2=arima(x=lipi,order=c(1,1,1),seasonal=list(order=c(1,1,2),period=12),method='ML')
t_stat(modele2)
modele2$aic
#
modele3=arima(x=lipi,order=c(1,1,1),seasonal=list(order=c(0,1,2),period=12),method='ML')
t_stat(modele3)
modele3$aic
# AR(1) a enlever
modele4=arima(x=lipi,order=c(0,1,1),seasonal=list(order=c(0,1,2),period=12),method='ML')
t_stat(modele4)
modele4$aic
#
modele5=arima(x=lipi,order=c(0,1,2),seasonal=list(order=c(0,1,2),period=12),method='ML')
t_stat(modele5)
modele5$aic
stargazer((a))

## Modèle 5 retenu soit un SARIMA(0,1,2)(0,1,2)_{12}
#####################################################

#### Blancheur des résidus : ###
#################################

## Graphique et histogramme :
modele=modele5
residus=modele$residuals
dev.off()
plot(residus,main="Residus du SARIMA retenu",xlab="Dates")
hist(residus,breaks=20,freq=F,main="Histogramme residus SARIMA retenu",col='lightblue',border='blue')
lines(density(residus),lwd=2)
lines(density(rnorm(1000000,mean(residus),sd(residus))),col='red')

## Graphique des p-valeurs :
pval = c()
for(i in 1:50){
  pval[i]=Box.test(x=residus,lag=i,type="Ljung-Box")$p.value
}
vec = rep(0.05,50)
c=(rep(2,10))
plot(pval,xlab="Retards",col="red")
lines(vec,col="blue",lty=2)
legend(38,0.8,legend=c("p-valeurs", "y=0.05"),col=c("red", "blue"),lty=1:2,cex=0.7,text.font=4, bg='lightblu
help(plot)
pval[11] # Point ambigu --> elle vaut 0.054, on accepte la blancheur des résidus :)


#### PREDICTION ####
```

```
####################

### Prediction de X_{T+1} et X_{T+2} :
x1=predict(modele,n.ahead = 2)$pred[1]
x2=predict(modele,n.ahead = 2)$pred[2]

### Variance estimee des residus du modele :
sigma2=var(residuals(modele))

### Parametre psi1 intervenant dans la variance de la prevision de X_{T+2} :
psi1=modele$coef[4]
psi1

### Matrice de variance-covariance :
v2=sigma2*(1+(1+psi1)^2)
cov=sigma2*(1+psi1)
sigma=matrix(c(sigma2,cov,cov,v2),nrow=2,ncol=2)

# Région de confiance :
library(ellipse)
plot(ellipse(sigma,centre=c(x1,x2),type="l"),main="Region de confiance a 95% pour (X_{T+1},X_{T+2})",)
```