Ecole Nationale de la Statistique
et de l'Administration Économique
Paris

Projet de Séries Temporelles

Samy Mekkaoui et Simon Depaule

Sommaire

1 Les Données

1.1 Choix de la série

Dans ce projet de Séries Temporelles,nous allons nous intéresser à la fabrication mensuelle de produits amylacés de Janvier 1990 à Février 2002. Cette série est accessible sur le site de l'INSEE au lien suivant : `//https://www.insee.fr/fr/statistiques/serie/010537272`.

Nous traçons ci-contre l'allure de notre série :

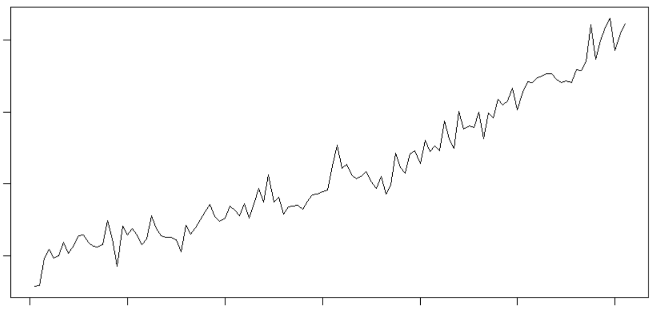

Figure 1 : Graphique représentant l'allure brute de notre série initiale

1.2 Mise en stationnarité de la série

Lorsque nous avons regardé graphiquement l'allure de notre série, il est assez clair qu'une tendance déterministe est présente. Elle semble par ailleurs être d'ordre 1 (d=1 dans le modèle ARIMA(p,d,q) puisqu'on a globalement l'allure d'une droite). Par ailleurs, il ne semble pas y avoir de saisonnalité donc il semblerait que seul un effet déterministe soit présent dans notre série. Ainsi, pour confirmer notre idée, posons $Y_t = X_t - X_{t-1}$ qui sera notre série corrigée. Ainsi, en différenciant notre série à l'ordre 1, notre série semble être stationnaire. Voici-ci dessous l'allure de cette dernière :

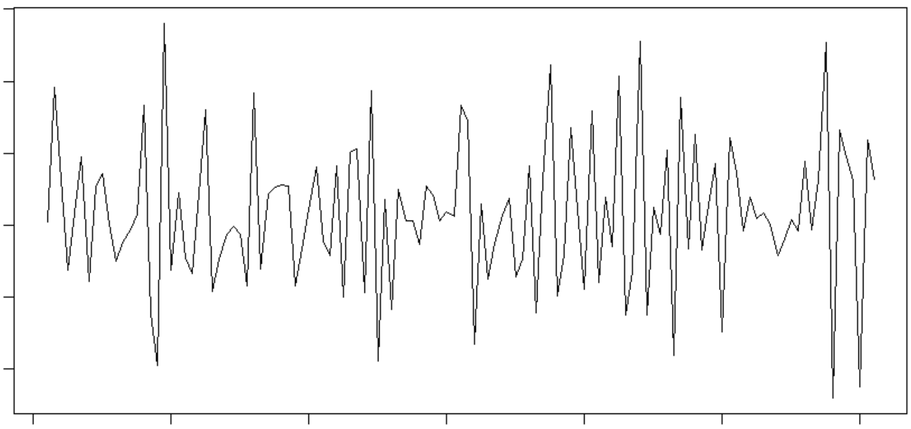

Figure 2 : Graphique obtenue après différenciation à l'ordre 1 de notre série initiale

Il est maintenant intéressant d'effectuer les tests de stationnarité pour confirmer notre hypothèse. Pour celà, nous allons nous intéresser aux tests classiques de racine unité : Dickey Fuller Augmenté,Phillips-Perron et le test KPSS. Par ailleurs, il est important de souligner que pour les tests de Dickey Fuller Augmenté et Phillips Perron, l'hypothèse nulle est la présence d'une racine unité et donc de la non stationnarité de la série alors que pour le test KPSS, l'hypothèse nulle est la stationnarité de la série. Regardons donc les résultats obtenus et résumés dans le tableau suivant :

Test	Valeur du t-test	Valeur de la p-value
Dickey Fuller	-6.1062	< 0.01
Phillips-Perron	-149.77	<0.01
KPSS	0.088	>0.1

Figure 3 : Résultats des tests ADF,PP et KPSS

Ainsi, nous obtenons pour les tests Dickey-Fuller et Phillips-Peron une p-value inférieur à 0,01 ce qui nous incite à rejeter l'Hypothèse nulle et donc à plutôt accepter que la série est stationnaire. Par ailleurs, pour le test KPSS, la p-value est plus grande que 0,1 donc on a tendance à accepter l'Hypothèse nulle qui est que notre série est stationnaire. Ainsi, il semble que notre série Y_t soit bien stationnaire

1.3 Représentation Graphique de la Série avant et après correction

Ainsi, après notre travail de mise en stationnarité de la série initiale brute, on peut comparer nos 2 séries celle initiale avec tendance linéaire déterministe X_t et celle corrigée Y_t

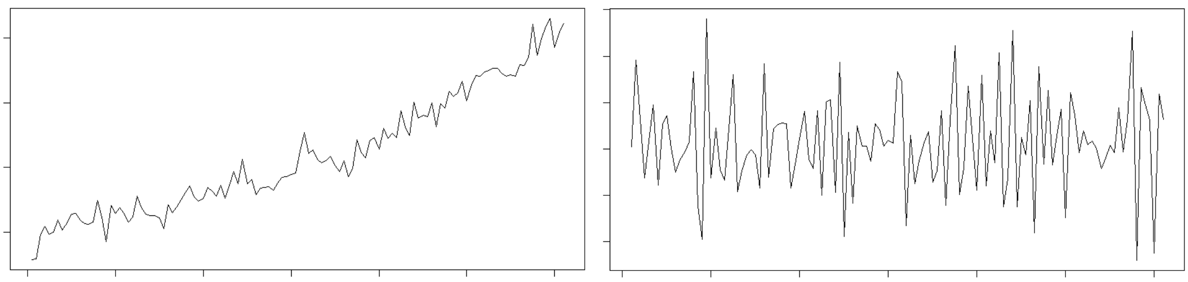

Figure 4 :Comparaison de la série initiale brute et de la série corrigée

2 Modèles ARMA

Nous allons maintenant nous intéresser ici au choix du bon modèle ARMA pour notre série corrigée Y_t

2.1 Choix du Modèle ARMA et estimations des paramètres associés au modèle

D'une part, d'après l'analyse faite précédemment, la série $Trend$ apparaît stationnaire donc on aura d=0 dans notre processus ARIMA(p,d,q). Analysons maintenant les autocorrélogrammes et les autocorrélogrammes partiels pour déterminer les coefficients potentiels de notre processus ARIMA. (la valeur de p étant reliée au PACF et celle de q à l'ACF). Tout d'abord, on regarde l'autocorrélogramme partiel de notre série $Y'_t = Y_t - Y_{mean}$ que voici ci-dessous :

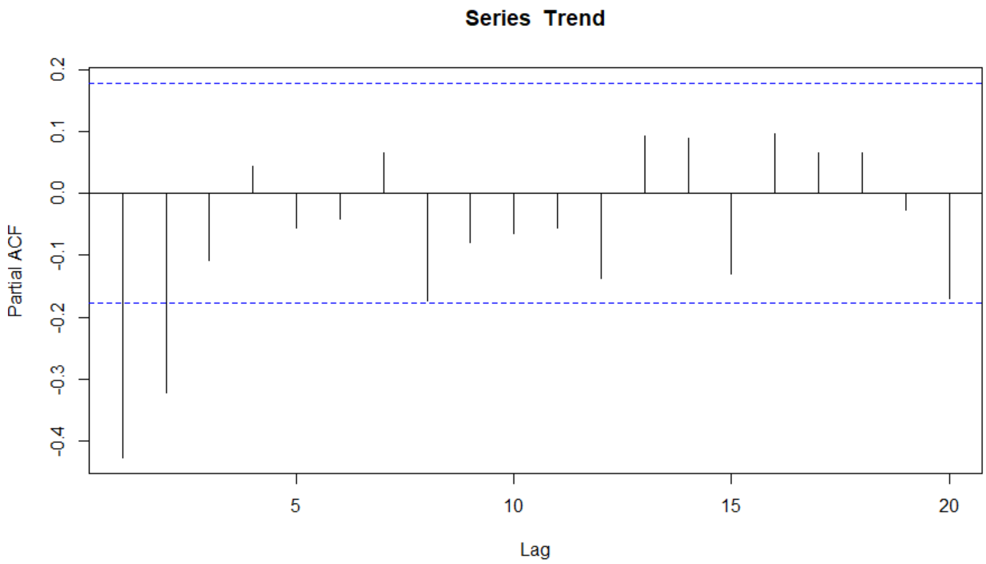

Figure 5 : PACF de notre série corrigée $Y' t$

Par ailleurs , on regarde le second autocorrélogramme de notre série :

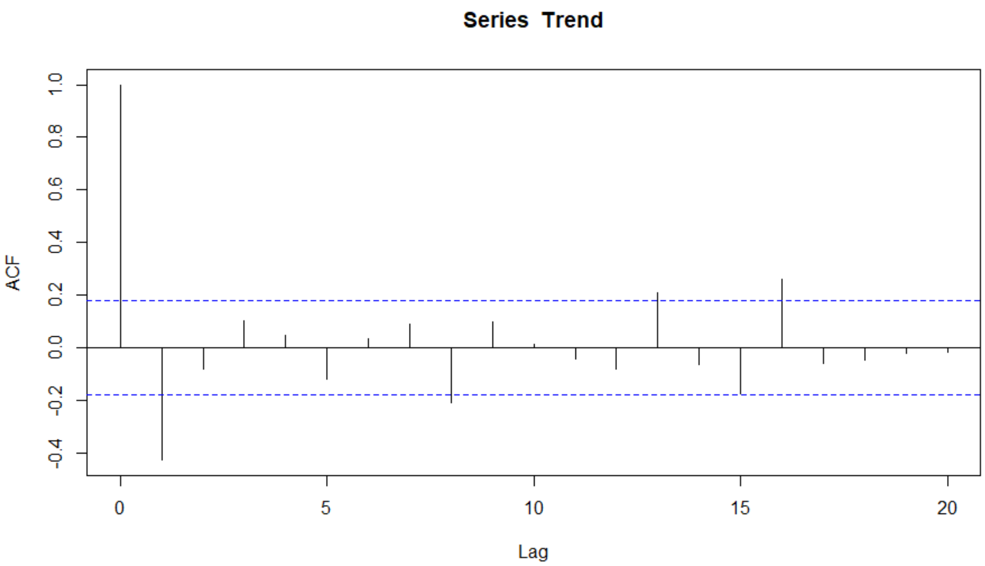

Figure 6 : ACF de notre série corrigée $Y' t$

Après visualisation des 2 figures précédentes, on voit que l'ACF et la PACF ne présentent des piques significatifs respectivement que pour 0≤p≤1 et 0≤p≤1. Par ailleurs, il est naturel à ce stade étant donné que les valeurs des autocorrélogrammes sont quasiment toutes nulles pour q≥2 qu'un modèle MA(1) soit cohérent pour nous série Y'_t

Ainsi, il nous faut déterminer le meilleur modèle ARMA. Pour se faire nous allons travailler sur le critère d'Akaike (AIC) et on choisira le modèle qui minimise ce critère. Pour se faire , nous allons ajuster un modèle ARMA(p,q) pour les valeurs de p et q retenues précédemment. Voici les résultats obtenus :

p \ q	0	1
0	516.36	482.37
1	494.89	484.34

Figure 7 : Comparaison des différentes valeurs du critère AIC

On choisit donc le modèle MA(1) pour notre série corrigée Y'_t. Par ailleurs, en utilisant une fonction particulière de R qui nous renseigne directement sur le meilleur modèle ARMA pour notre série (fonction auto.arima de R), on trouve également que le modèle MA(1) est le plus adapté pour notre série $Trend$.

2.2 Expression du Modèle ARIMA

Intéressons nous maintenant aux coefficients de ce modèle ARMA(0,1) que nous avons choisi. Pour se faire, nous allons donner les résultats obtenus à partir de ce modèle dans le tableau ci-dessous :

	Estimate	Standard Error	Valeur du test	p-value
MA(1)	-0.59	0.075	-7.89	$3.3 * 10^{-15}$
Intercept	-0.015	0.065	-0.228	0.82

Figure 8 : Coefficients du modèle MA(1) choisi

Remarquons que la valeur du coefficient MA(1) est très significatif car la p-value est très faible et la valeur de l'intercept est très peu significative. Or, on sait par construction de notre série Y'_t qu'elle est de moyenne nulle donc l'intercept doit être nulle ce qui est cohérent avec notre modèle. Ainsi, comme la valeur du coefficient MA(1) est très significatif, on valide notre modèle MA(1) et c'est celui qu'on choisira pour la suite du projet.

Ainsi, nous avons choisi pour notre série $Trend$ un modèle MA(1) et voici l'expression de notre série : $Y'_t = \epsilon_t - 0.59\epsilon_{t-1}$

3 Prévision

Nous allons nous intéresser à la prévision de notre série Y'_t. Nous posons T la longueur de notre série et nous supposons les résidus de la série gaussiens donc :

$$\epsilon_t \sim \mathcal{N}(, \sigma^2)$$

Par ailleurs, on sait que notre modèle s'écrit : $Y'_t = \epsilon_t + \phi_1 \epsilon_{t-1}$

3.1 Equation de région de confiance de niveau α

On sait par le cours que les prévisions optimales en T sont données par :
$Y'_{T+1|T} = \phi_1 \epsilon_t$
$Y'_{T+2|T} = Y'_{mean} = 0$
Nous allons maintenant nous intéresser aux erreurs de prédiction : $Y'_{T+1} - Y'_{T+1|T}$ et $Y'_{T+2} - Y'_{T+2|T}$

Posons Y_{pred}= $\begin{pmatrix} Y'_{T+1|T} \\ Y'_{T+2|T} \end{pmatrix}$ et Y'= $\begin{pmatrix} Y'_{T+1} \\ Y'_{T+2} \end{pmatrix}$

Mais alors, on a :
$Y' - Y_{pred}$= $\begin{pmatrix} Y'_{T+1|T} - Y'_{T+1} \\ Y'_{T+2|T} - Y'_{T+2} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + \phi_1 \epsilon_{T+1} \end{pmatrix}$
Ainsi, on peut calculer la variance de chaque élement de ce vecteur aléatoire. On a :

$Var(Y'_{T+1|T} - Y'_{T+1}) = Var(\epsilon_{T+1}) = \sigma^2$
$Var(Y'_{T+2|T} - Y'_{T+2}) = Var(\epsilon_{T+2} + \phi_1 \epsilon_{T+1}) = \sigma^2(1+\phi_1^2)$ par non corrélation entre les résidus.

Ainsi, on voit que $Y_{pred} - Y'$ suit une loi normale de paramètres $\mu = 0$ et Σ avec Σ :a matrice de variance-covariance définie de la manière suivante : $\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\phi_1 \\ \sigma^2\phi_1 & \sigma^2(1+\phi_1^2) \end{pmatrix}$

Or, il est clair que det(Σ) =σ^2>0 d'après l'énoncé donc on a caractérisé notre erreur de prédiction.
On utilise maintenant le cours pour nous assurer que $(Y' - Y_{pred})^T \Sigma^{-1} (Y' - Y_{pred}) \rightsquigarrow \chi^2(2)$. Ainsi, on peut caractériser notre intervalle de confiance pour tout niveau $\alpha \in [0,1]$. En effet, la région de confiance de niveau α est alors $A_\alpha = \{Y' \in \mathbb{R}^2 \mid (Y' - Y_{pred})^T \Sigma^{-1} (Y' - Y_{pred}) \leq q^{1-\alpha}_{\chi^2(2)}\}$ où $q^{1-\alpha}_{\chi^2(2)}$ est le quantile d'ordre 1-α de la loi $\chi^2(2)$

3.2 Hypothèses utilisées pour la détermination de notre région de confiance

Afin de déterminer notre région de confiance, nous avons eu besoin d'effectuer plusieurs hypothèses que nous allons expliciter ci-dessous

1. On suppose que notre modèle est parfaitement connu et qu'il est bien un modèle ARMA(0,1) dans notre cas
2. Les coefficients estimés dans notre modèle sont les vrais coefficients de notre modèle
3. Les résidus sont bien gaussiens
4. $\sigma^2 > 0$ afin de pouvoir inverser la matrice de variance-covariance
5. La variance des résidus est connue

On a en effet utilisé notre série corrigée en faisant l'hypothèse qu'elle suivait un modèle ARMA dont nous avons montré que le meilleur était le modèle ARMA(0,1), puis nous avons eu besoin d'estimer les coefficients associés à ce modèle et on a eu besoin de la connaissance de la loi des résidus et notamment de la variance. Si cette dernière n'était pas connue, on aurait eu besoin de l'estimer et celà aurait rendu nos intervalles de confiance beaucoup moins précis.

3.3 Graphique obtenue pour un intervalle confiance de niveau 95%

Nous pouvons maintenant représenter graphiquement la région obtenue pour $\alpha = 95\%$.Notons que les points bleus correspondent aux valeurs prédites et les bandes grises aux intervalles de confiance à 95%. Voici-ci dessous ce que nous obtenons :

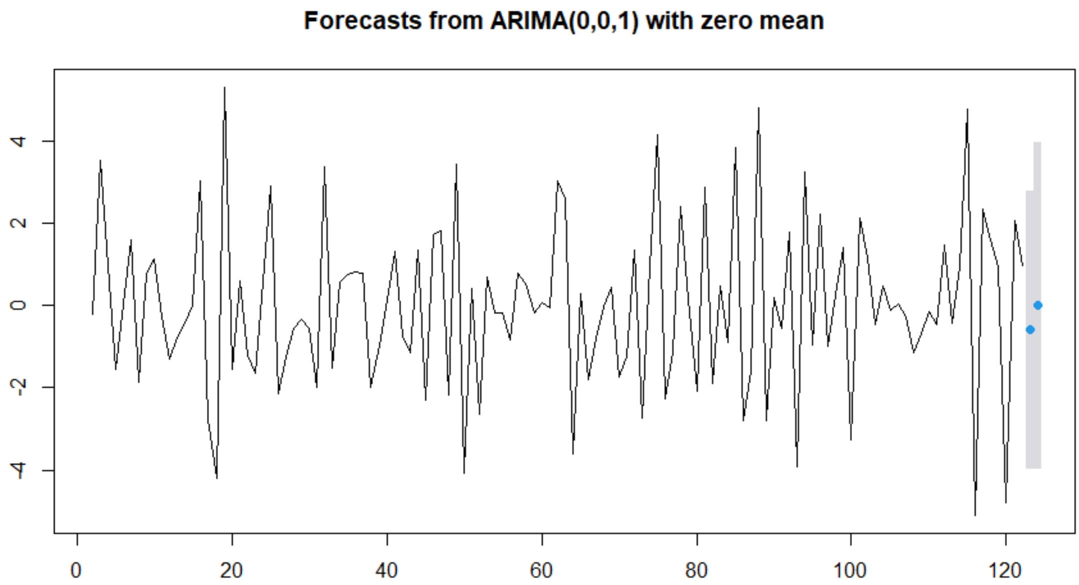

Figure 9 : Prédiction à horizon h=2 des valeurs de notre série corigée et des intervalles de confiance à 95%

On voit que les valeurs prédites sont bien dans les intervalles de confiance calculés ce qui nous parait cohérent d'après notre modèle.

3.4 Réponse à la Question Ouverte

Considérons une série stationnaire Y_t dont nous disposons des valeurs de 1 à T et nous savons que Y_{T+1} est disponible avant X_{T+1}. Alors, on sait que l'on pourra utiliser Y_{T+1} si la série Y_t cause instantanément au sens de Granger la série X_t c'est à dire si l'erreur quadratique moyenne de prédiction de X_{T+1} conditionnellement aux valeurs passés de Y_t est inférieure à celle que nous aurions eu sans l'information apportée par la série Y_t

4 Annexes

4.1 Import des données

Voici le code associée à la partie Import des données :

```
#Import des données

serie=read.csv("C:\\Users\\samym\\Desktop\\Séries Temporelles Linéaires\\Série2.csv")
View(serie)

# On remet la série dans le bon ordre


serie1=serie$Valeurs[122:1] #Nous ne conservons que les données de Janvier 1990 à Février 2002 ce qui correspond
#aux 122 premières données

plot(serie1,type="l") # on la plot pour voir à quoi elle ressemble

# On voit une tendance linéaire dans notre série. On va tester une différenciation à l'ordre 1
```

4.2 Différenciation à l'ordre 1

Voici le code associée à la partie Différenciation à l'ordre 1 :

```
#Mise de la série au format Zoo
serie2=zoo(serie1)
plot(serie2)


#Différenciation à l'ordre 1 de notre série pour la rendre stationnaire

Trend=serie2-lag(serie2,-1)
plot(Trend)

# Notre série Trend semble être stationnaire après cette différenciation à l'ordre 1


# On rend notre série de moyenne nulle par soucis de simplification
Trend=Trend-mean(Trend)
```

4.3 Tests de Stationnarité

Voici le code associé à la partie Tests de Stationnarité :

```
# Tests de Stationnarité
adf.test(Trend) # Test de Dickey-Fuller Augmenté
pp.test(Trend) # Test de Phillips-Perron
kpss.test(Trend) # Test KPSS
```

4.4 Choix du meilleur modèle

Voici le code associée au choix du meilleur modèle ARMA(p,q) pour notre série corrigée Y'_t

```
#On va tester les modèles pour p<=1 et q<=1

model1=arma(Trend,order=c(0,0))
summary(model1)

model2=arma(Trend,order=c(0,1))
summary(model2)

model3=arma(Trend,order=c(1,0))
summary(model3)

model4=arma(Trend,order=c(1,1))
summary(model4)


# On voit que le modèle 2 est celui qui minimise le critère AIC


#On vérifie que le ARMA(0,1) est  le meilleur modèle avec la fonction auto.arima de R

auto.arima(Trend,seasonal=FALSE)

# Elle redonne bien le modèle ARMA(0,1) comme meilleur modèle ARMA de notre série
```

4.5 Prédiction de notre série à horizon T=2

Voici le code associée à la prédiction de notre série à horizon T=2

```
# On va prédire à horizon 2 ( h=2) les valeurs de notre série avec un intervalle de confiance à 95%

mod=auto.arima(Trend,seasonal=FALSE)

pred_forecast=forecast(mod,h = 2,level=95)
plot(pred_forecast) # On plot  à horizon h=2 les 2 valeurs prédites par le modèle ainsi que les intervalles
# de confiance à 95%
```