Titre

Votre nom

oday

Contents

Time Series Analysis

Introduction

Le but de ce rapport est de presenter les resultats du projet. Nous souhaitons déterminer le modèle étant le plus adapté pour décrire les 3
séries. L'analyse devra donc determiner un modèle et vérifier si celui-ci valide les hypothèses.

> **Note**
>
> Les séries choisies portent le nom de "Chuzeville.xls".

Serie 1

Importation de la série

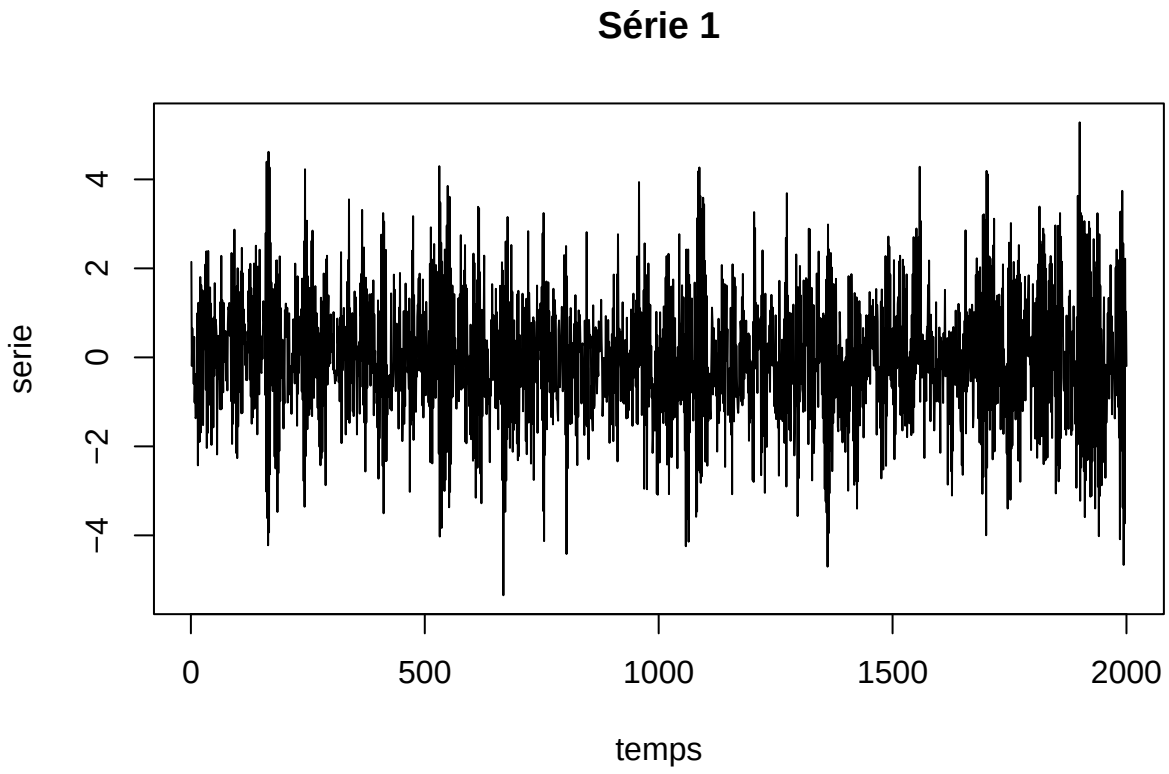
```
rm ( list = ls () )
cat (" \014 ")
```

```
library ( readxl )
serie1 = read_excel("C:/Users/PC/Documents/Time series-courses/Chuzeville.xls" , sheet = "serie1")
serie1
```

```
## # A tibble: 2,000 x 2
##    temps   serie
##    <chr>   <dbl>
##  1 1      2.15
##  2 2     -0.193
##  3 3      0.654
##  4 4     -0.0995
##  5 5      0.269
##  6 6     -0.596
##  7 7      0.460
##  8 8     -1.01
##  9 9     -0.375
## 10 10    -1.37
## # i 1,990 more rows
```

Note > > A chaque importation des données (3 fois), il faut modifier le chemin d'acccès correspondant à l'emplacement local du fichier.

```
plot(serie1, type='l', main ='Série 1')
```



Visualisation

La série semble stationnaire et oscille de manière assez chaotique (a priori entre -6 et 5 grossièrement), il est difficile après visualisation de définir une éventuelle périodicité.

Stationnarité Pour commencer il est bon de savoir si notre hypothèse de stationnarité a priori est bel et bien statistiquement acceptable. Pour se faire on réalise un test amélioré de Dickey-Fuller:

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
adf.test(serie1$serie)
```

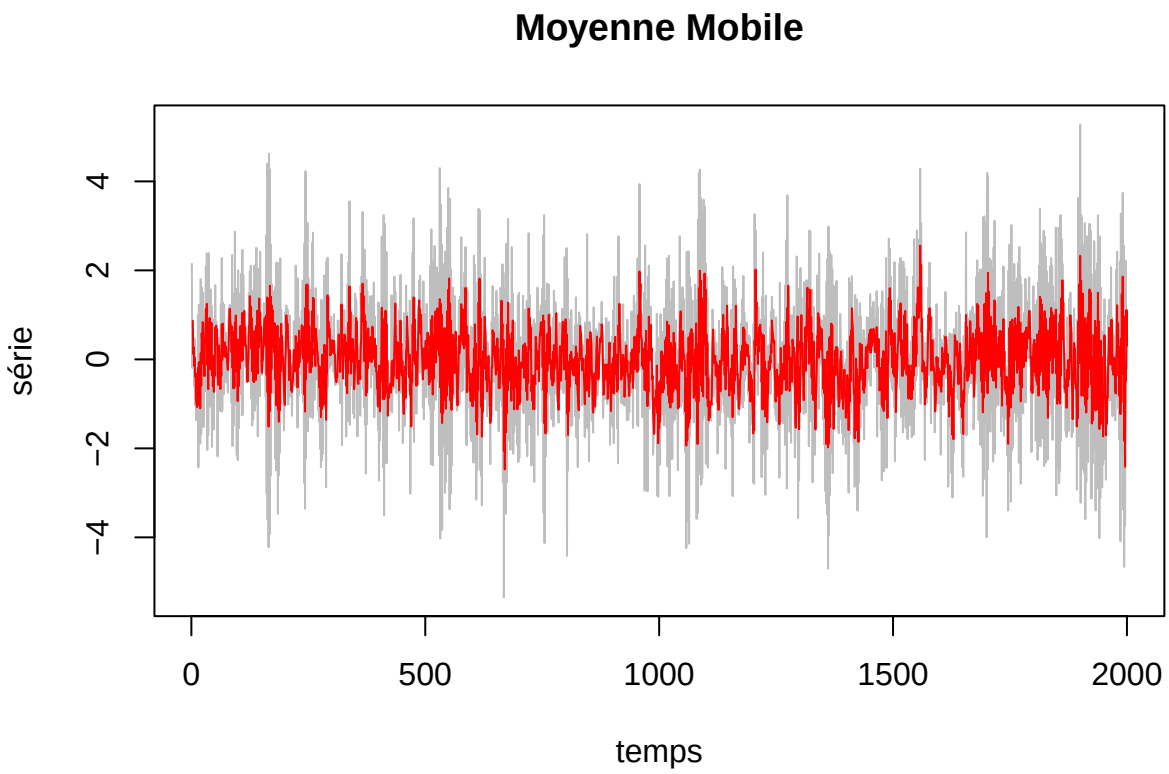
```
## Warning in adf.test(serie1$serie): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serie1$serie
## Dickey-Fuller = -11.759, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

Comme le test fournit un p-value de 0.01 qui est sous le seuil standard de rejet de 0.05 on peut rejeter l'hypothèse nulle qui consiste à affirmer la présence d'une racine unitaire à notre série, ceci entraine l'acceptation de notre hypothèse de stationnarité pour la Série1.

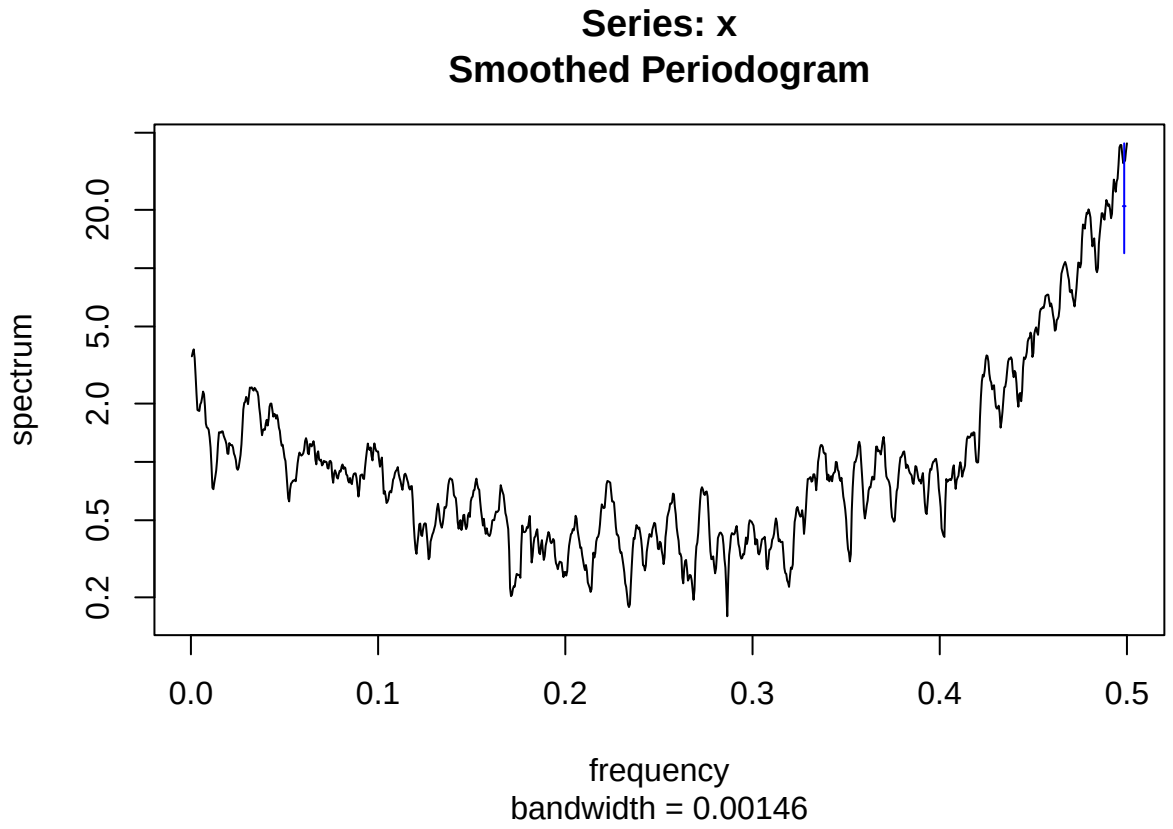
Moyenne mobile Pour avoir des informations plus précises au sujet de la tendance et la périodicité, traçons la moyenne mobile de la série.

```
moy_mob <- filter(serie1$serie,filter = rep(1/3, 3), method = 'convolution', sides = 1)
plot(serie1$serie, type='l',  col='grey', main='Moyenne Mobile', xlab='temps', ylab='série')
lines(moy_mob, col='red')
```

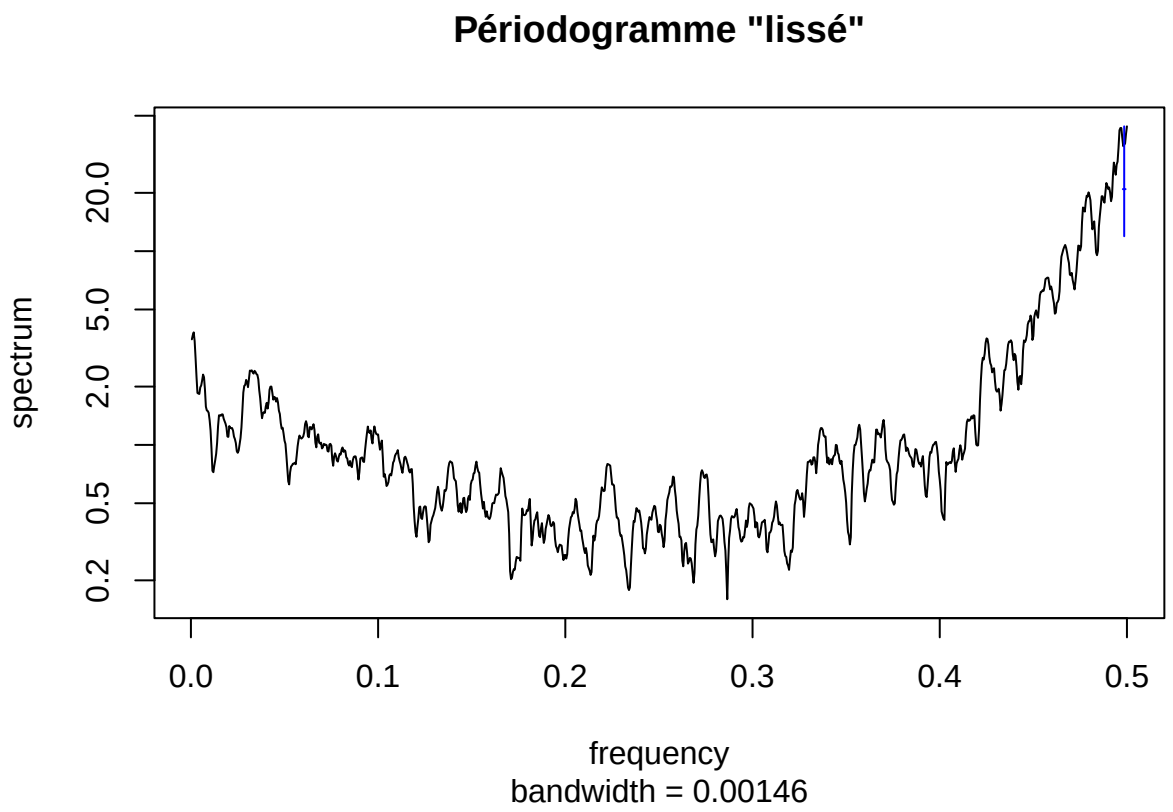


Il n'apparait rien de plus évident en considérantla moyenne mobile de cette série. Etudions alors le périodogramme de la Série 1 initiale:

```
spectre = spectrum ( serie1$serie, spans=10 )
```



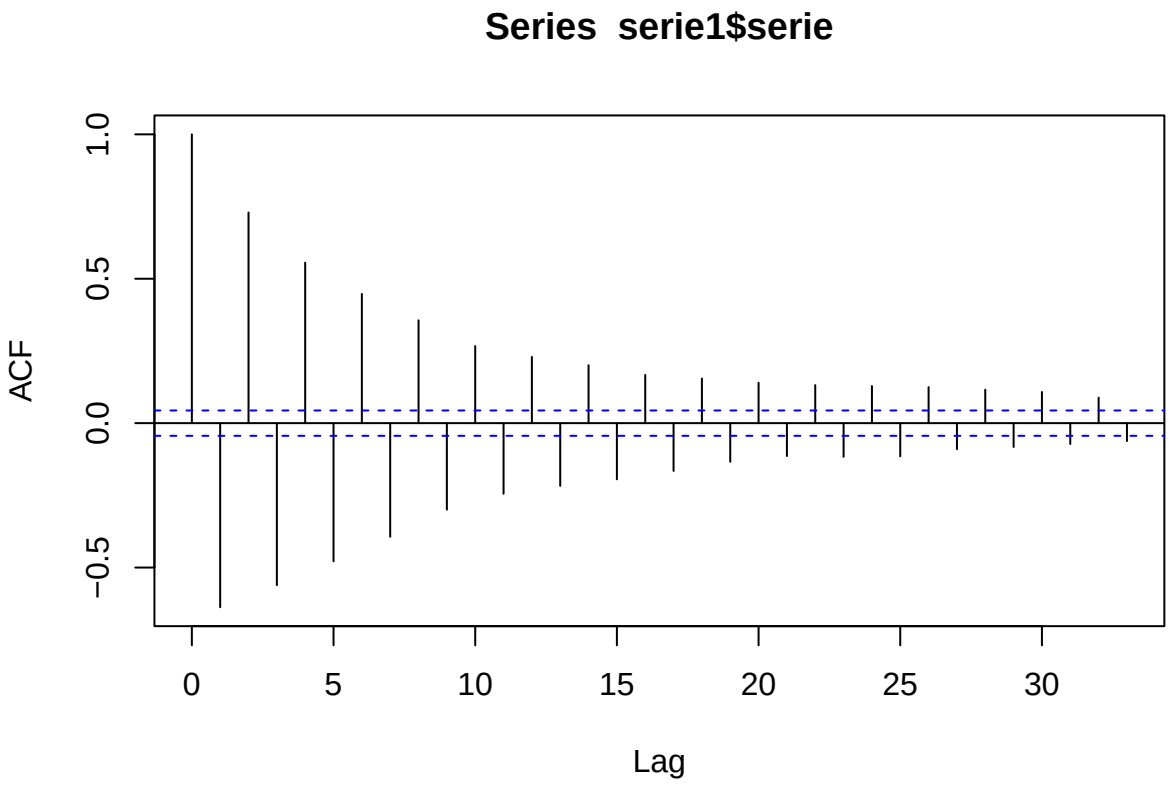

```
plot ( spectre , main = 'Périodogramme "lissé"')
```

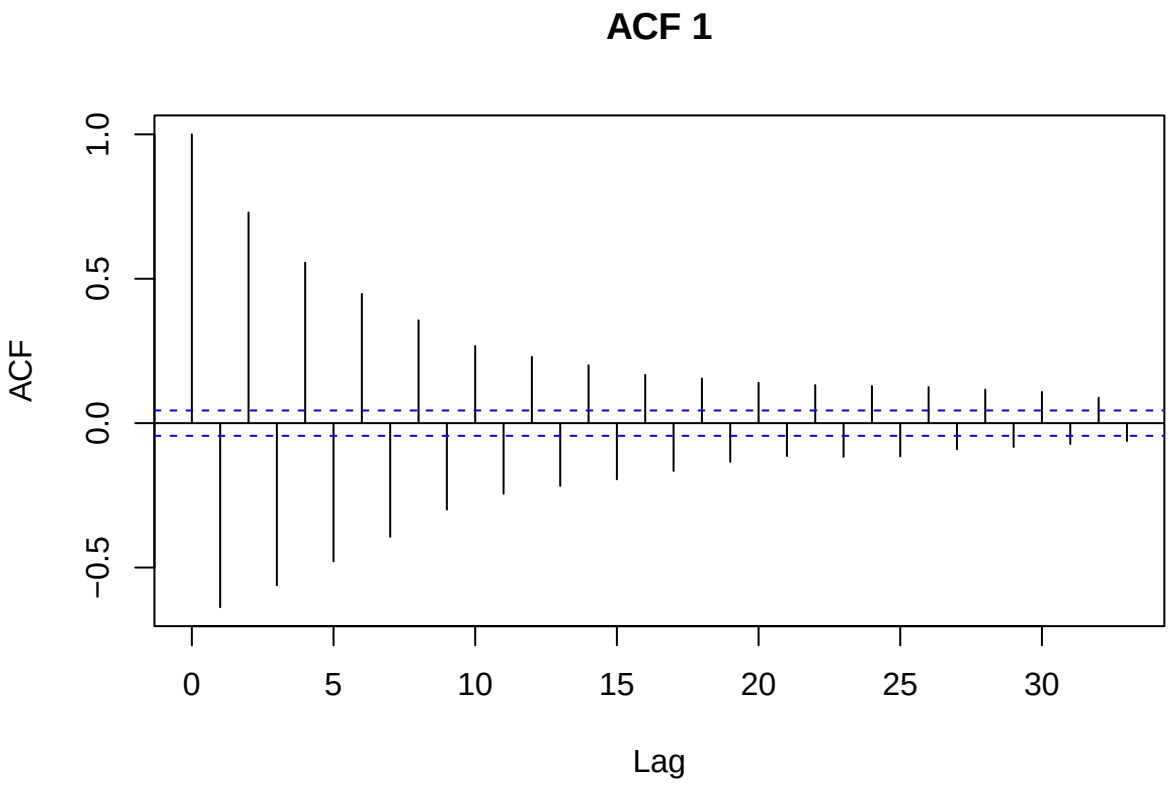


Nous n'obtenons rien de très convaincant par simple analyse visuelle. Passons donc à l'étude plus précise et au choix de notre modèle.

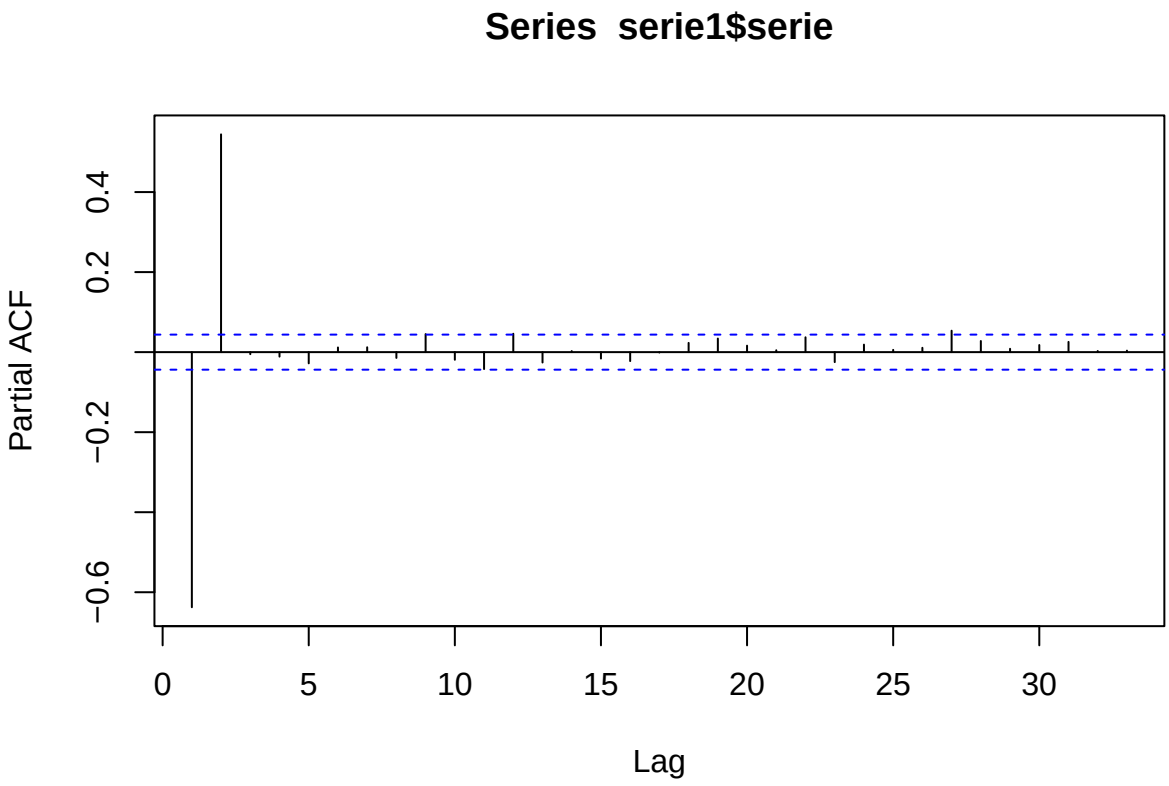
Choix de notre modèle Pour chosir le modèle qui conviendrait le mieux à cette série on s'intéresse tout d'abord aux ACF et PACF et de cette dernière:

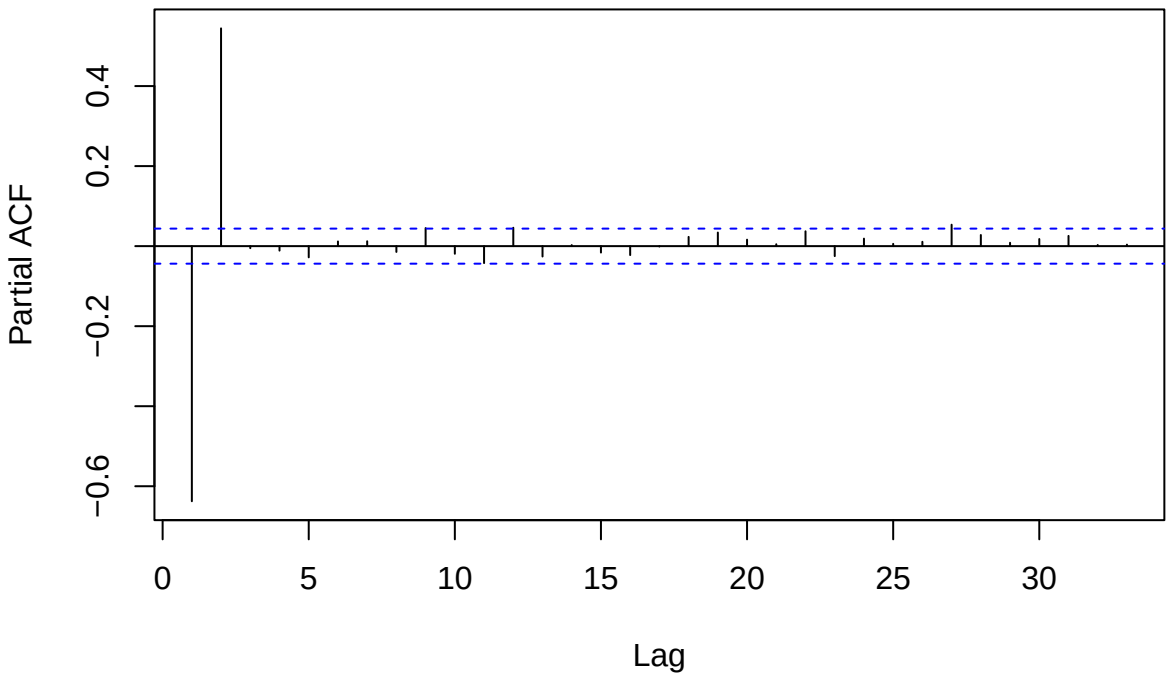
```
plot (acf(serie1$serie),main='ACF 1')
```


Series serie1$serie
ACF
1.0
0.5
0.0
−0.5
0
5
10
15
20
25
30
Lag




```
plot (pacf(serie1$serie), main='PACF 1')
```


Series serie1$serie
Partial ACF
0.4
0.2
−0.2
−0.6
0
5
10
15
20
25
30
Lag



Il est clair que l'autocorrélation est exponentiellement décroissante et alternée ceci suggère que le terme de moyenne mobile dans notre modèle est d'ordre 0 autrement dit qu'il n'y a pas de corrélation significative entre les retards. Concernant l'autocorrélation partielle, celle-ci présente 2 pics très significatifs ce qui suggèrerait que le terme auto regressif du modèle est d'ordre 2, c'est-à-dire que "les données du présent sont fortement reliées à celles antérieures de 2 rang".

Ces analyses nous encouragent à opter pour un modèle ARMA(2,0).

Vérification des hypothèses du modèle On génère donc notre modèle à partir de notre série:

```
model = arima(serie1$serie, order=c(2,0,0))
model
```

```
##
## Call:
## arima(x = serie1$serie, order = c(2, 0, 0))
##
## Coefficients:
##           ar1     ar2  intercept
##       -0.2902  0.5445    -0.0216
## s.e.   0.0188  0.0187     0.0297
##
## sigma^2 estimated as 0.9827:  log likelihood = -2821.02,  aic = 5650.05
```

On calcule la variance et la moyenne des résidus de notre modèle:

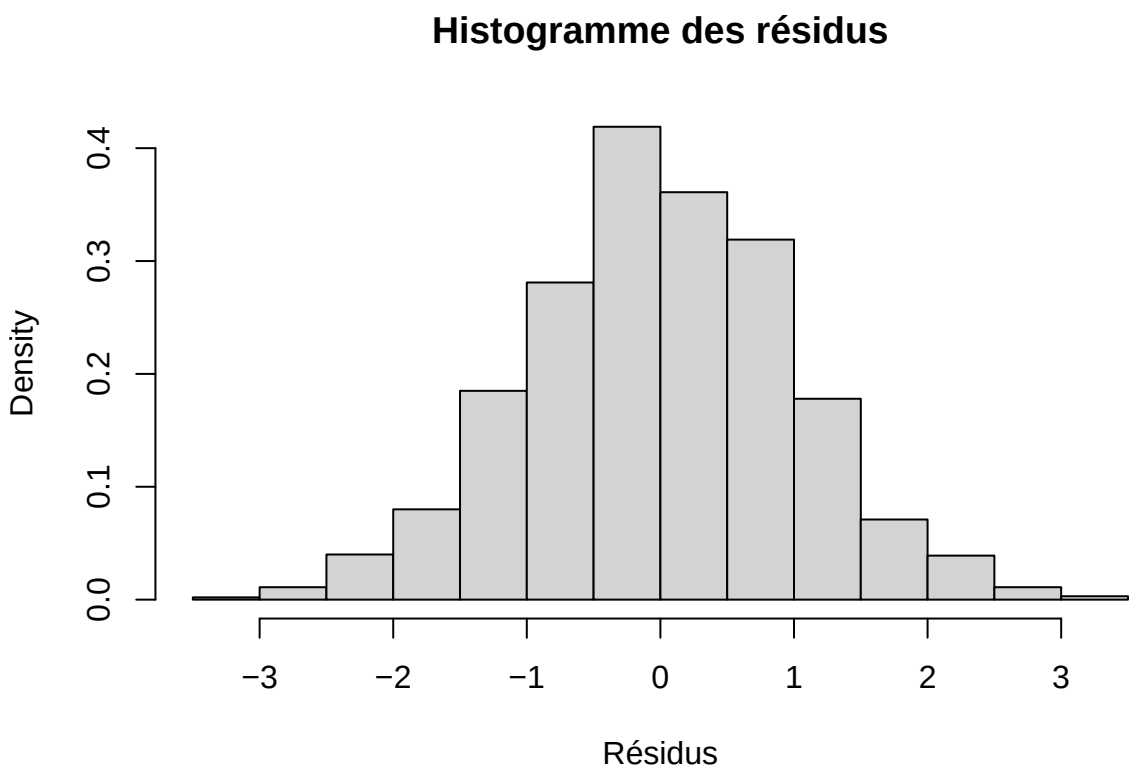
```r
res = model$residual
var(res)
```

```
## [1] 0.9831766
```

```r
mean(res)
```

```
## [1] -0.0003192402
```

```r
hist(res,main='Histogramme des résidus',xlab = 'Résidus',freq=F)
```



La variance des résidus est très proche de 1 et leur moyenne proche de 0. Ensuite on se demande s'il s'agit d'un bruit blanc c'est-à_dire que ces résidus suivent bien la loi normale N(0,1). On réalise alors le test de Shapiro-Wilk:

```r
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.99939, p-value = 0.7982
```

Comme le test fournit un p-value de 0.01 qui est sous le seuil standard de rejet de 0.05 on peut rejeter l'hypothèse nulle qui consiste à affirmer la présence d'une racine unitaire à notre série, ceci entraine l'acceptation de notre hypothèse de stationnarité pour la Série1.

La p-value ($0.7982 >> 0.05$) ne permet pas de rejeter l'hypothèse de non-normalité. Il est donc vraisemblable que ces résidus soient bel et bien des bruits blancs.

On teste la stationnarité de ses résidus via le test de Dickey-Fuller:

```
adf.test(res)
```

```
## Warning in adf.test(res): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res
## Dickey-Fuller = -12.547, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

Comme le test fournit un p-value de 0.01 qui est sous le seuil standard de rejet de 0.05 on peut rejeter l'hypothèse nulle qui consiste à affirmer la présence d'une racine unitaire à notre série, ceci entraine l'acceptation de notre hypothèse de stationnarité pour les résidus du modèle.

AIC/BIC du modèle, optimisation On se demande si notre modèle est celui qui s'approche le mieux de notre série en ayant un nombre de paramètres raisonnable. Pour se faire on calcule l'AIC et le BIC des différents modèles ARIMA:

```
l_aic <-c()
l_bic <- c()
for (i in 0:3){
  for (j in 0:3){
    for (k in 0:3) {
test = arima ( serie1$serie , order=c (i , j , k) )
l_aic <- c(l_aic, i,j ,k, AIC(test))
l_bic <- c(l_bic, i,j ,k, BIC(test))
}
}
}
```

```
## Warning in arima(serie1$serie, order = c(i, j, k)): problème de convergence
## possible : optim renvoie un code = 1
## Warning in arima(serie1$serie, order = c(i, j, k)): problème de convergence
## possible : optim renvoie un code = 1
```

> **Note**
>
> Ici on teste les modèles ARIMA(i,j,k) pour i,j et k allant de 0 à 4 (ce qui est suffisant)

On renvoie le meilleurs choix de paramètre obtenu via l'AIC puis via le BIC:

```
M_aic <- matrix(l_aic, ncol=4, byrow=T)
M_bic <- matrix(l_bic, ncol=4, byrow=T)
Maic = as.data.frame(M_aic)
Mbic = as.data.frame(M_bic)
Maic[Maic[4]==min(Maic[4])]
```

```
## [1]    2.000    1.000    1.000 5649.956
```

```
Mbic[Mbic[4]==min(Mbic[4])]
```

```
## [1]    2.000    1.000    1.000 5672.358
```

Après calcule de l'AIC et du BIC il semblerait que ARIMA(2,1,1) fournissent de meilleurs résultats, cependant en obsvervant les valeurs de l'AIC et du BIC pour ARIMA(2,0,0) et en les comparant à celle de ARIMA (2,1,1) on a:

```
print('AIC ARIMA(2,1,1)')
```

```
## [1] "AIC ARIMA(2,1,1)"
```

```
print(min(Maic[4]))
```

```
## [1] 5649.956
```

```
print('AIC ARIMA(2,0,0)')
```

```
## [1] "AIC ARIMA(2,0,0)"
```

```
print(Maic[4][Maic[1]==2 & Maic[2]==0 & Maic[3]==0])
```

```
## [1] 5650.045
```

```
print('BIC ARIMA(2,1,1)')
```

```
## [1] "BIC ARIMA(2,1,1)"
```

```
print(min(Mbic[4]))
```

```
## [1] 5672.358
```

```
print('BIC ARIMA(2,0,0)')
```

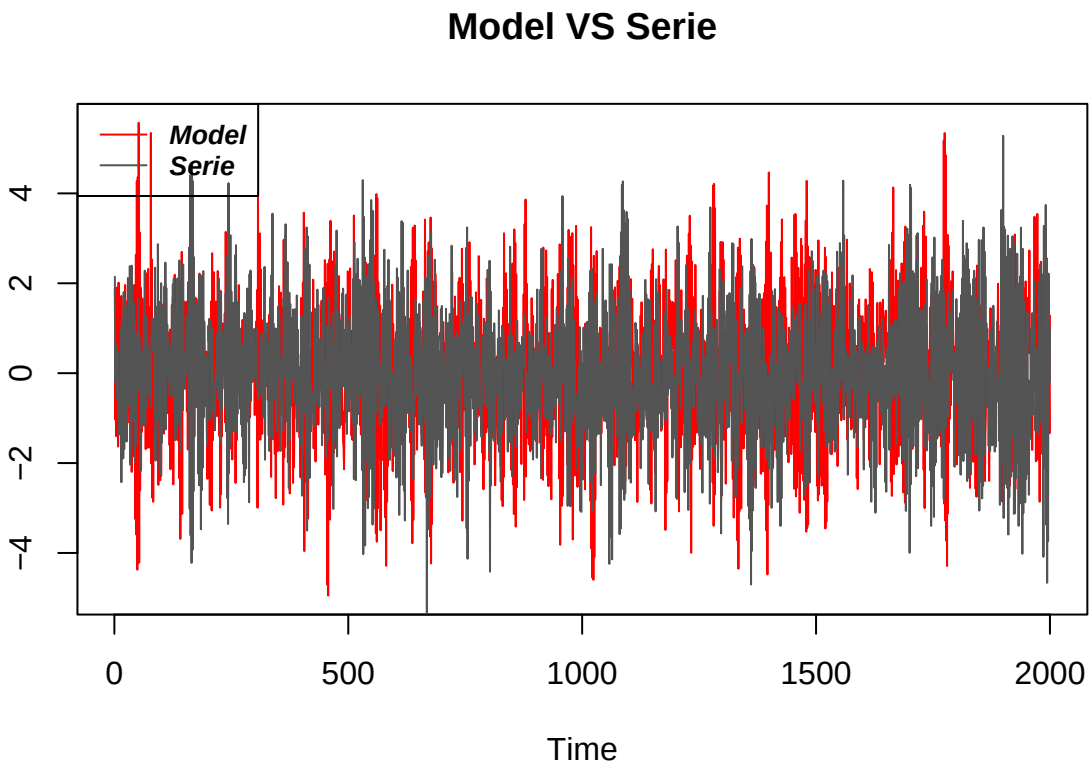
```
## [1] "BIC ARIMA(2,0,0)"
```

```r
print(Mbic[4][Mbic[1]==2 & Mbic[2]==0 & Mbic[3]==0])
```

```
## [1] 5672.449
```

On constate qu'à la fois l'AIC et le BIC sont extrèmement proches entre les modèles ARIMA(2,1,1) et ARIMA(2,0,0). Puisque que ce dernier modèle est plus simple que le précédent il fait sens de le valider comme modèle adéquat.

```r
plot(arima.sim(list(order=c(2,0,0), ar= c(model$coef[1],model$coef[2])), n=length(serie1$serie)), col='
lines(serie1$serie, col='grey33')
legend("topleft", legend=c("Model", "Serie"),lty=1, col=c("red","grey33"), cex=0.8,text.font=4)
```



Conclusion: Après cette analyse nous optons pour le modèle ARIMA(2,0,0) pour cette Série 1.

Série 2

On reprend l'ensemble de la trame d'analyse de la première série:

```r
rm ( list = ls () )
cat (" \014 ")
```

```r
library ( readxl )
serie2 = read_excel("C:/Users/PC/Documents/Time series-courses/Chuzeville.xls" , sheet = "serie4")
```

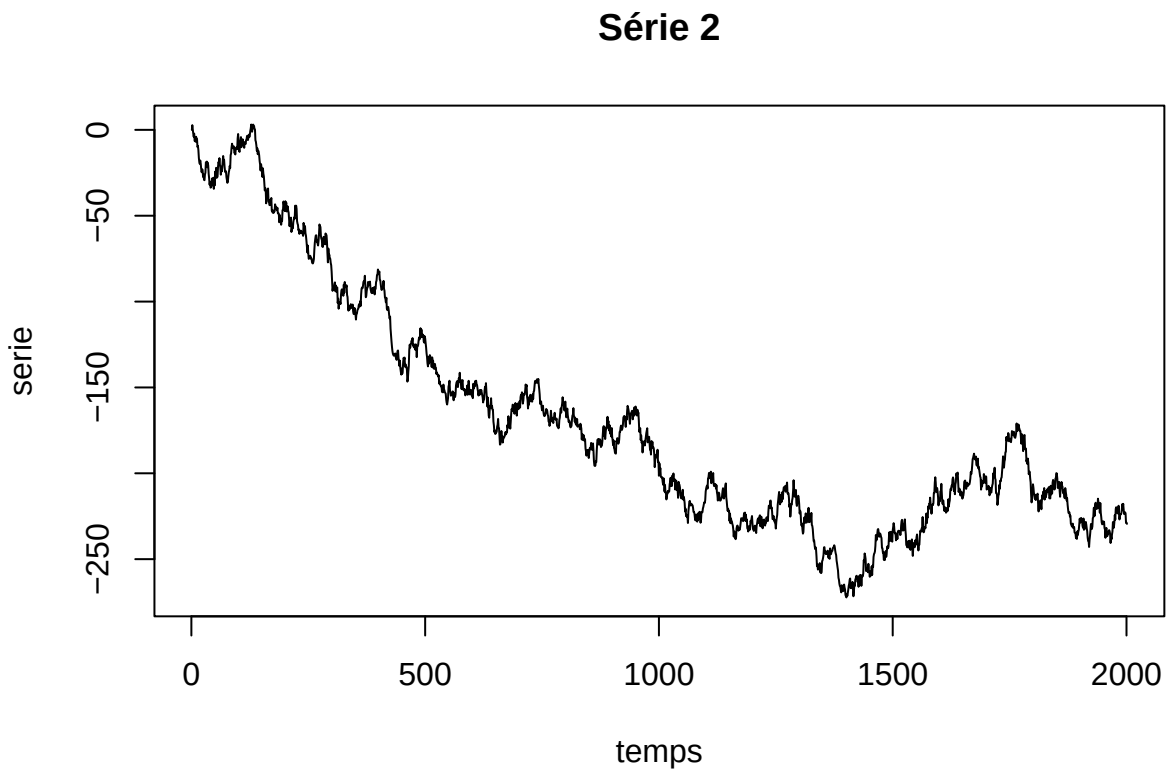
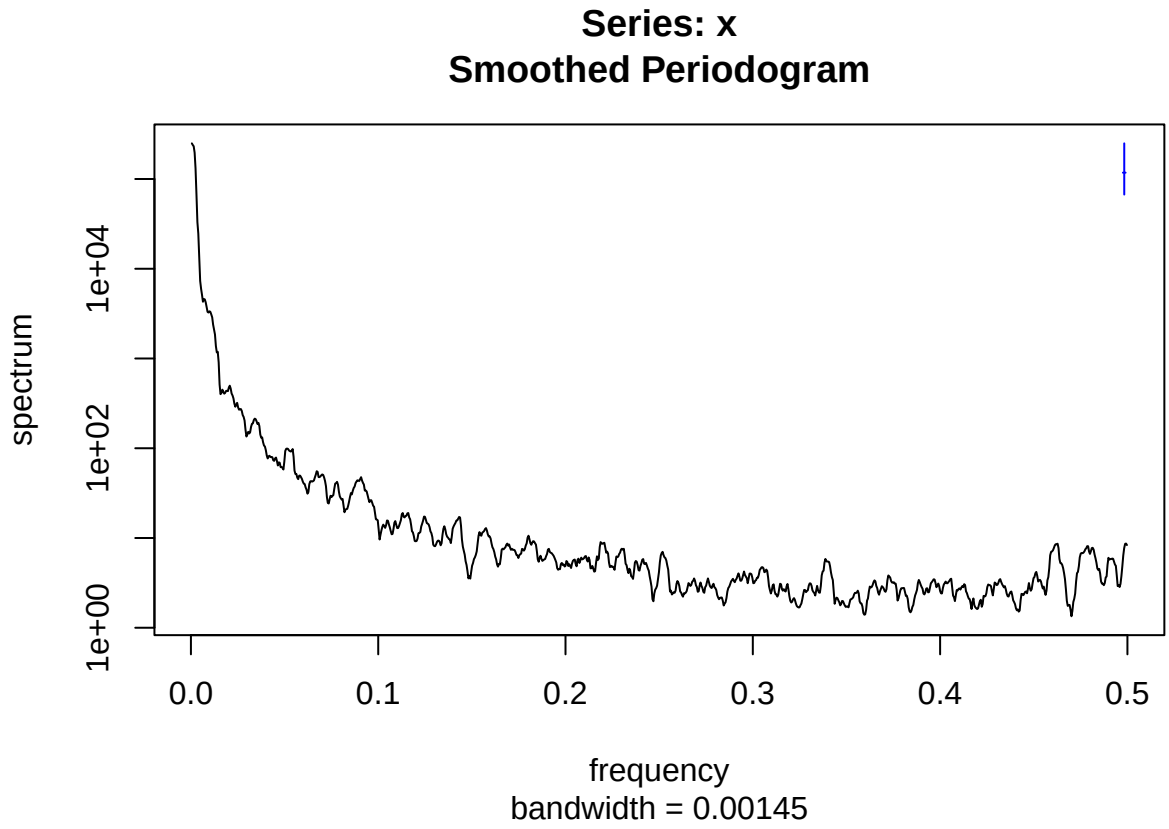
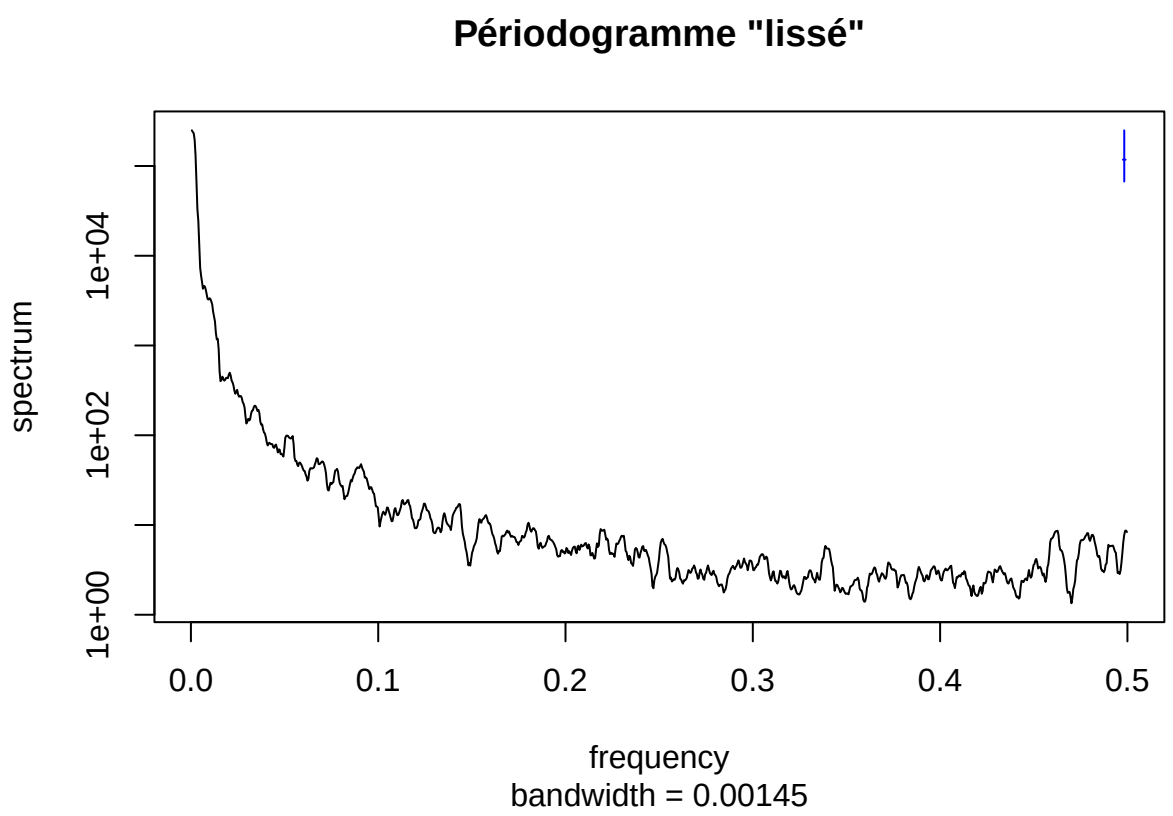


Figure 1: Série 2

Visualisation Cette nouvelle série est bien différente de la précédente, on constate directement qu'il semble y avoir une tendance décroissante, concernant les périodes il ne semble pas y en avoir.

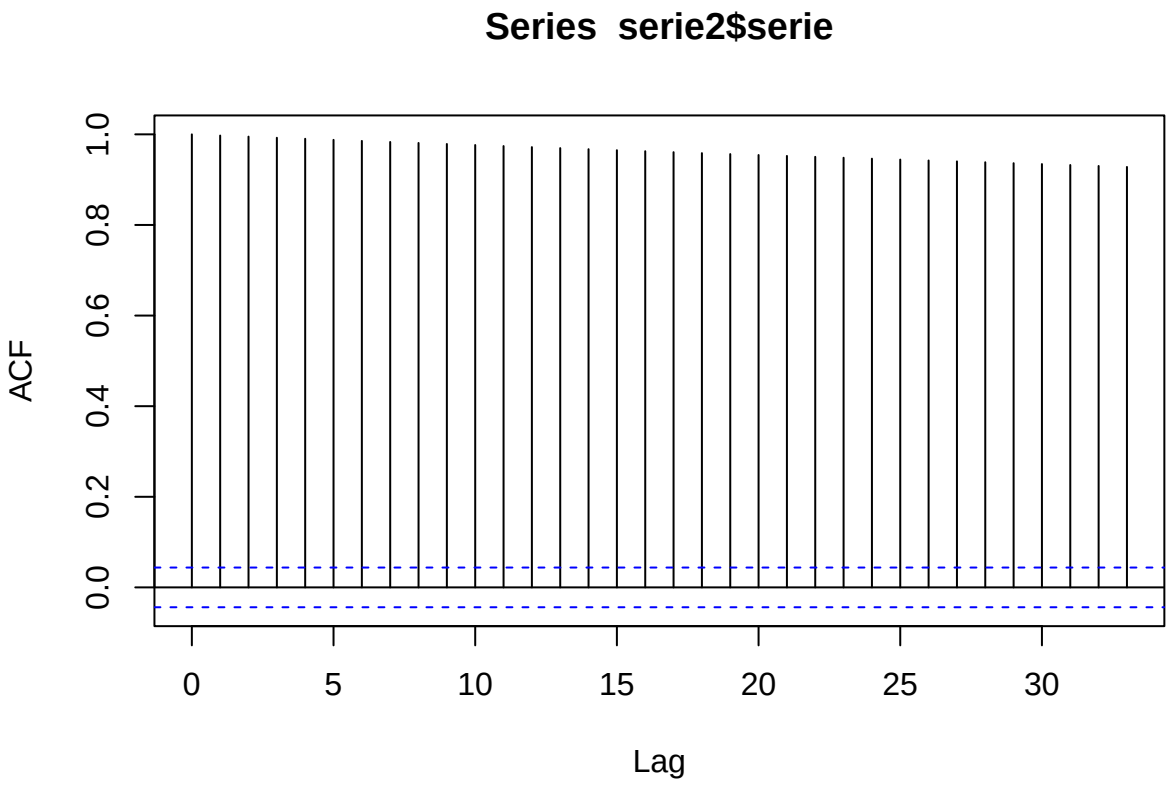
Etudions alors le périodogramme de la Série 2:

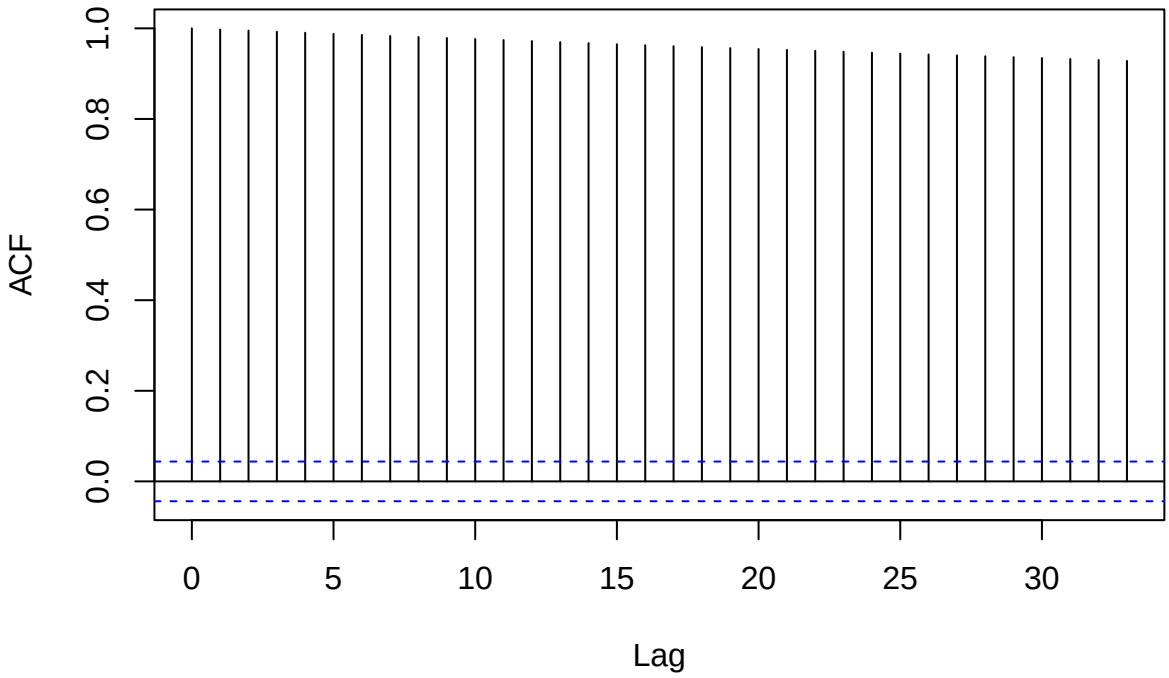

Series: x
Smoothed Periodogram
spectrum
1e+00
1e+02
1e+04
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00145

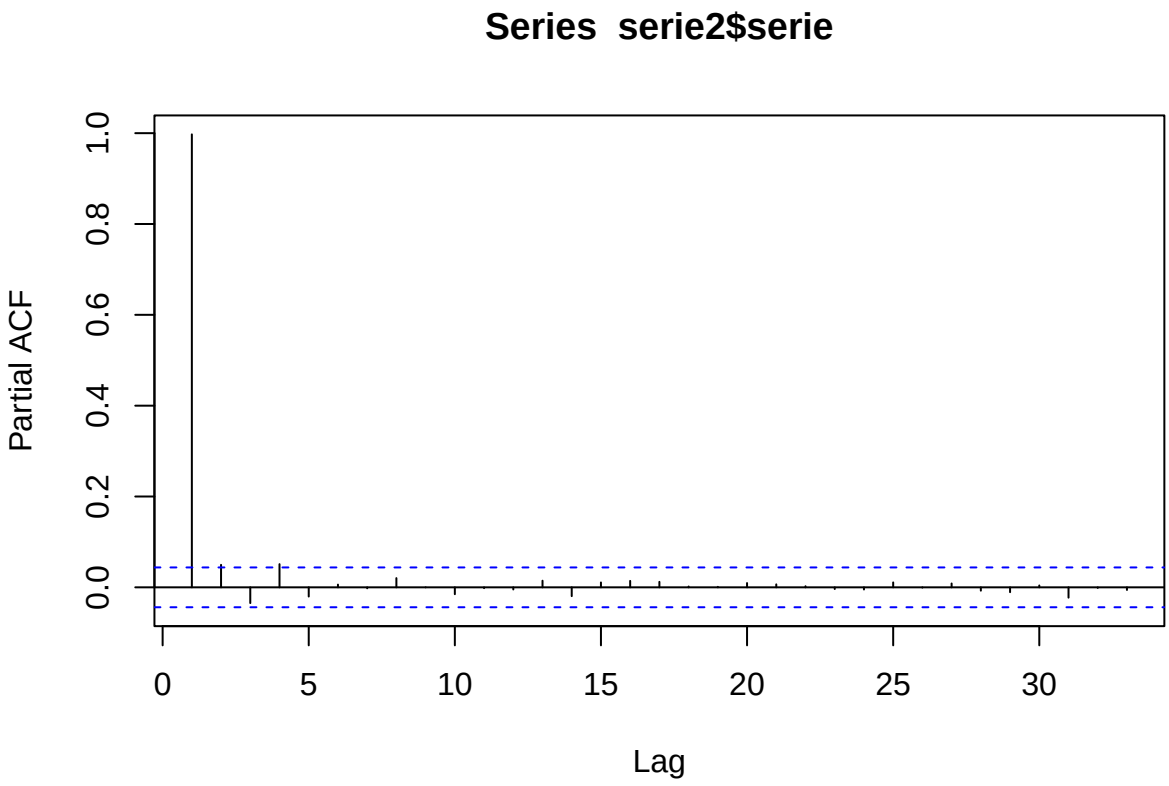


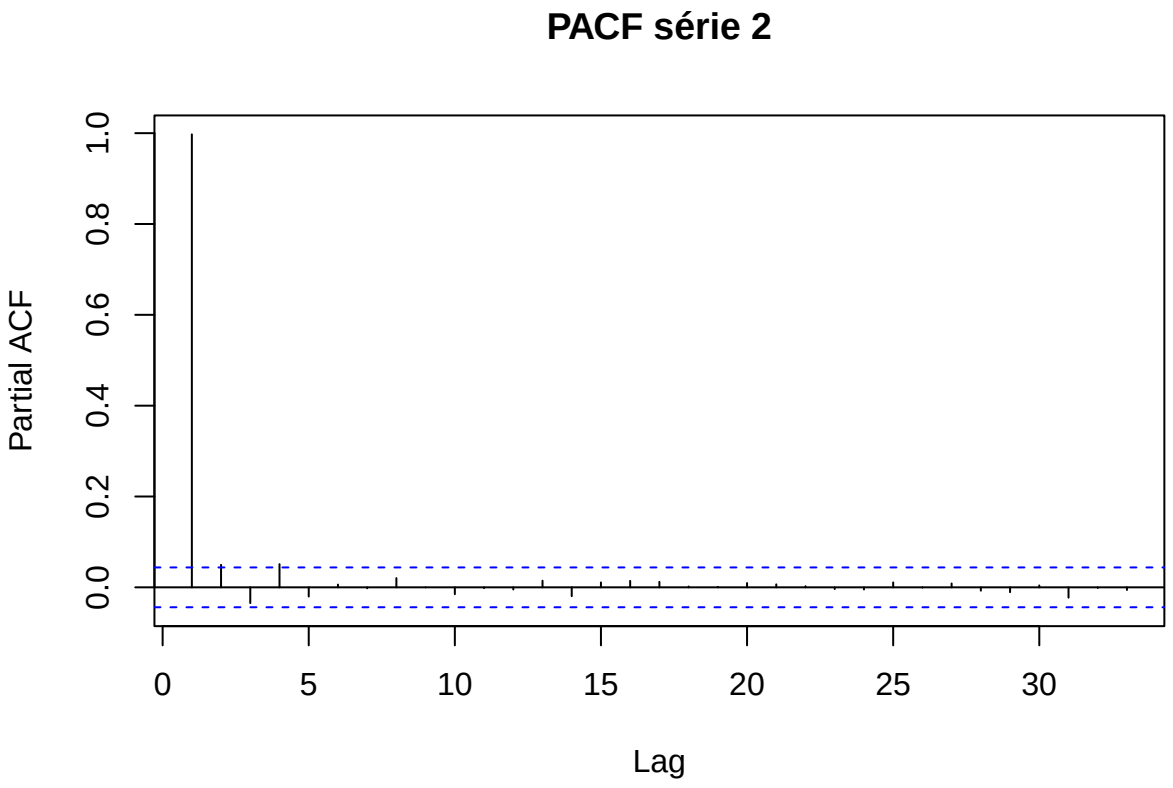
On observe aucun pic significatif, donc aucune période évidente a priori.

Choix du modèle Comme précédemment on trace l'ACF et la PACF:


Series serie2$serie
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag


ACF série 2
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag


Series serie2$serie
Partial ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag



L'ACF de cette série semble décroitre très lentement (elle est quasi constante dans le temps) il semble donc que le série ne soit pas stationnaire. Quant à l'autocorrélation partielle il n'y a qu'un seul pic significatif ceci indique que le terme de moyenne mobile est d'ordre 1.

Réalisons un test de stationnarité à cette série (test ADF):

```
adf.test(serie2$serie)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serie2$serie
## Dickey-Fuller = -1.8893, Lag order = 12, p-value = 0.6252
## alternative hypothesis: stationary
```

Le p-value du test est de 0.6252 ce qui signifie qu'on ne peut pas refuser l'hypothèse de non stationnarité de cette série, ce qui confirme ce qu'on avait établi a priori.

Ainsi on sait que l'ordre d'intégration de notre modèle ARIMA est strictement plus grand que 0.

Finalement cette analyse laisse à penser qu'il faut choisir un modèle de type ARIMA(p,q,1) avec p,q > 0.

Proposons donc le modèle préliminaire le plus simple qui correspond directement à notre analyse: le modèle ARIMA(1,1,1).

Vérification des hypothèses du modèle On génère donc notre modèle à partir de notre série:

```
model = arima(serie2$serie, order=c(1,1,1))
model
```

```
##
## Call:
## arima(x = serie2$serie, order = c(1, 1, 1))
##
## Coefficients:
##           ar1     ma1
##       -0.7908  0.6579
## s.e.   0.0450  0.0549
##
## sigma^2 estimated as 8.855:  log likelihood = -5018.94,  aic = 10043.88
```

On calcule la variance et la moyenne des résidus de notre modèle:

```
res = model$residual
print('Variance des résidus:')
```

```
## [1] "Variance des résidus:"
```

```
print(var(res))
```

```
## [1] 8.84005
```

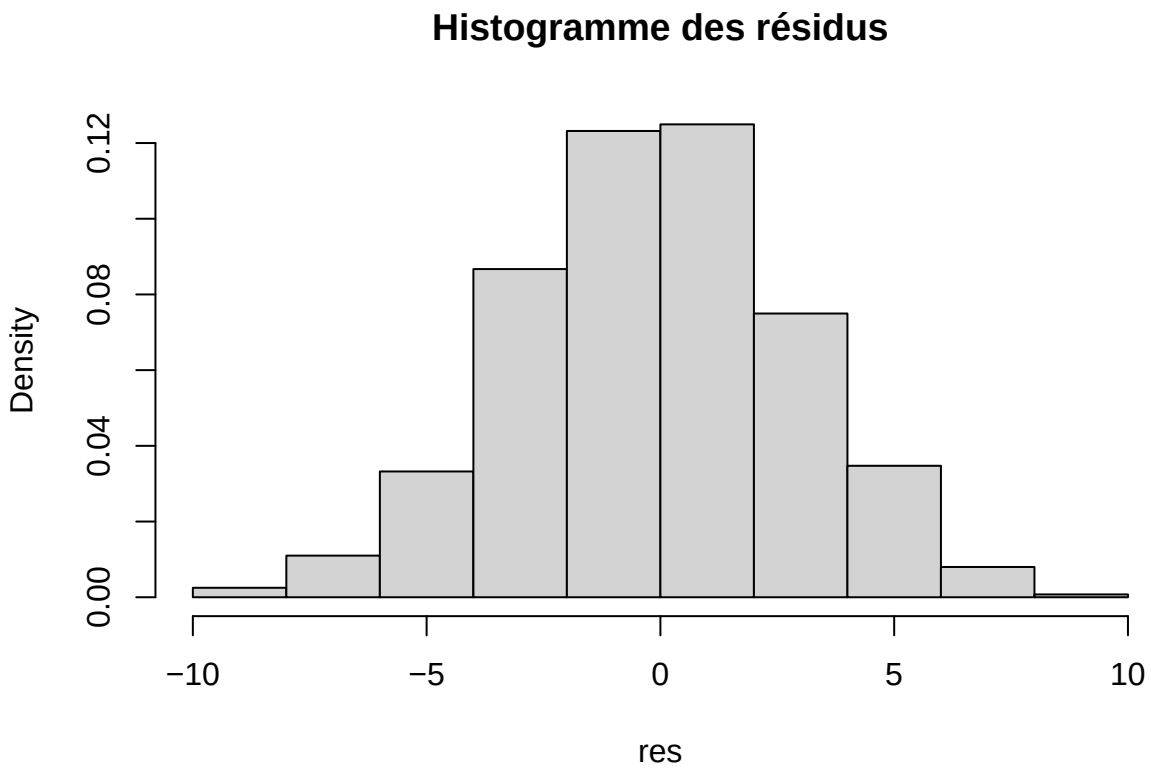
```
print('Moyenne des résidus:')
```

```
## [1] "Moyenne des résidus:"
```

```
print(mean(res))
```

```
## [1] -0.1237356
```

```
hist(res, freq=F, main='Histogramme des résidus')
```



On realise le test de normalité :

```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.99908, p-value = 0.4151
```

La p-value étant de 0.4151 on ne peut pas refuser l'hypothèse de non-normalité. ce qui indique que les résidus de notre modèle sont bien des bruits blancs.

On réalise le test de stationnarité:

```
adf.test(res)
```

```
## Warning in adf.test(res): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  res
## Dickey-Fuller = -11.813, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

La p-value est de 0.01 donc on peut rejeter l'hypothèse de non-stationnarité des résidus.

AIC/BIC du modèle, optimisation On se demande si notre modèle est celui qui s'approche le mieux de notre série en ayant un nombre de paramètres raisonnable. Pour se faire on calcule l'AIC et le BIC des différents modèles ARIMA:

```
l_aic <-c()
l_bic <- c()
for (i in 1:3){
  for (j in 1:3){
    for (k in 1:3) {
test = arima ( serie2$serie , order=c (i , j , k) )
l_aic <- c(l_aic, i,j ,k, AIC(test))
l_bic <- c(l_bic, i,j ,k, BIC(test))
}
}
}
```

```
## Warning in arima(serie2$serie, order = c(i, j, k)): problème de convergence
## possible : optim renvoie un code = 1
```

> **Note**
>
> Ici on teste les modèles ARIMA(i,j,k) pour i,j et k allant de 0 à 4 (ce qui est suffisant)

On renvoie le meilleur choix de paramètres obtenu via l'AIC puis via le BIC:

```
M_aic <- matrix(l_aic, ncol=4, byrow=T)
M_bic <- matrix(l_bic, ncol=4, byrow=T)
Maic = as.data.frame(M_aic)
Mbic = as.data.frame(M_bic)
print("Meilleur AIC")
```

```
## [1] "Meilleur AIC"
```

```
print(Maic[Maic[4]==min(Maic[4])])
```

```
## [1]     1.00     1.00     1.00 10043.88
```

```
print('Meilleur BIC')
```

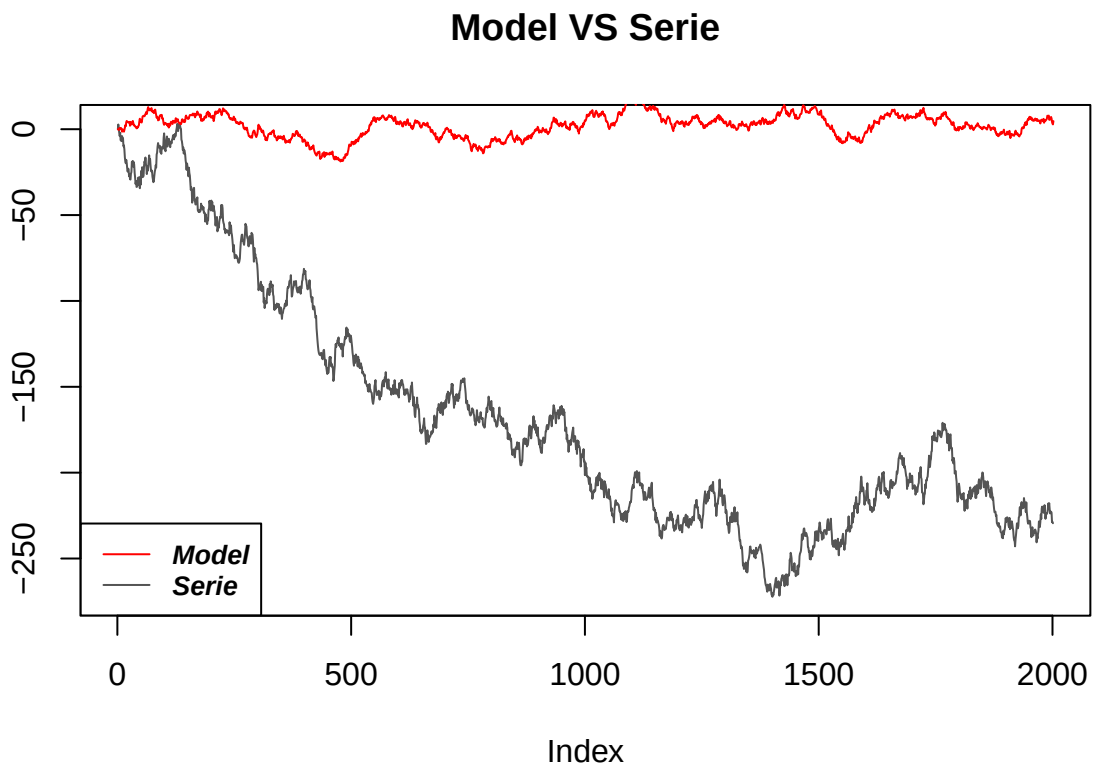
```
## [1] "Meilleur BIC"
```

```
print(Mbic[Mbic[4]==min(Mbic[4])])
```

```
## [1]     1.00     1.00     1.00 10060.68
```

Pour cette série 2, à la fois l'AIC et le BIC confirme le caractère "optimal" du modèle préliminaire choisi c'est-à-dire le modèle ARIMA(1,1,1).

```r
plot(serie2$serie, col='grey33',main = 'Model VS Serie',ylab='',type='l')
lines(arima.sim(list(order=c(1,1,1), ar= c(model$coef[1]), ma=c(model$coef[2])), n=length(serie2$serie))
legend("bottomleft", legend=c("Model", "Serie"),lty=1, col=c("red","grey33"), cex=0.8,text.font=4)
```



Note > > Parfois pour avoir un meilleur rendu graphique il faut relancer le code (pour que le modèle et la série trouve leur place dans la fenêtre)
Conclusion:

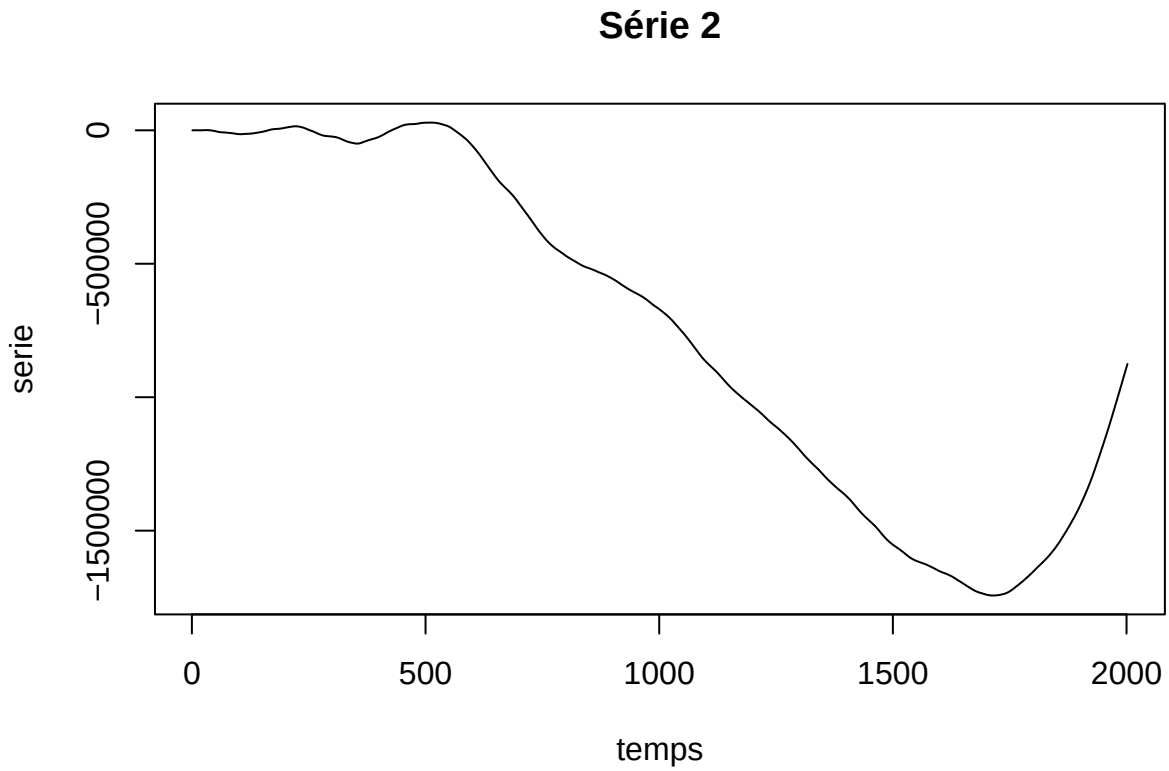
Après cette analyse nous optons pour le modèle ARIMA(1,1,1) pour cette Série 2.

Série 3

```r
rm ( list = ls () )
cat (" \014 ")
```

```
library ( readxl )
serie3 = read_excel("C:/Users/PC/Documents/Time series-courses/Chuzeville.xls" , sheet = "serie5")
```

```
plot(serie3, type='l', main ='Série 2')
```

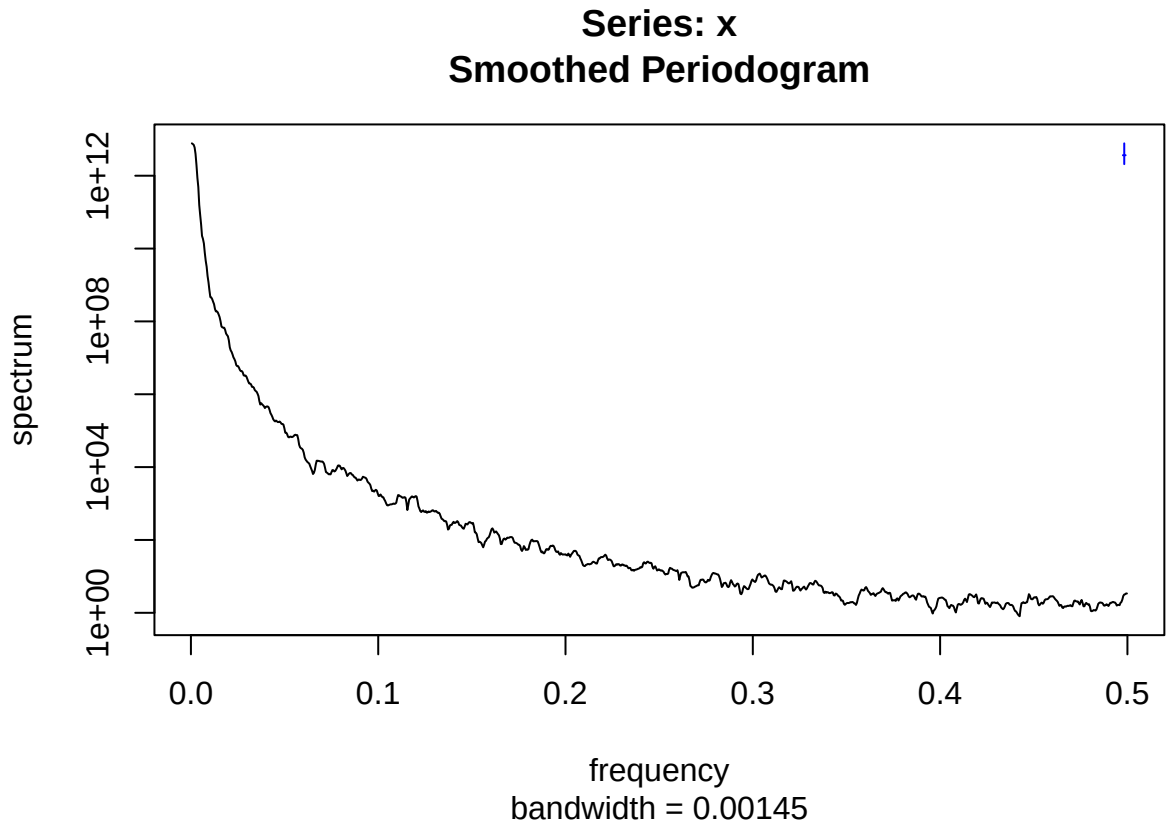


Visualisation

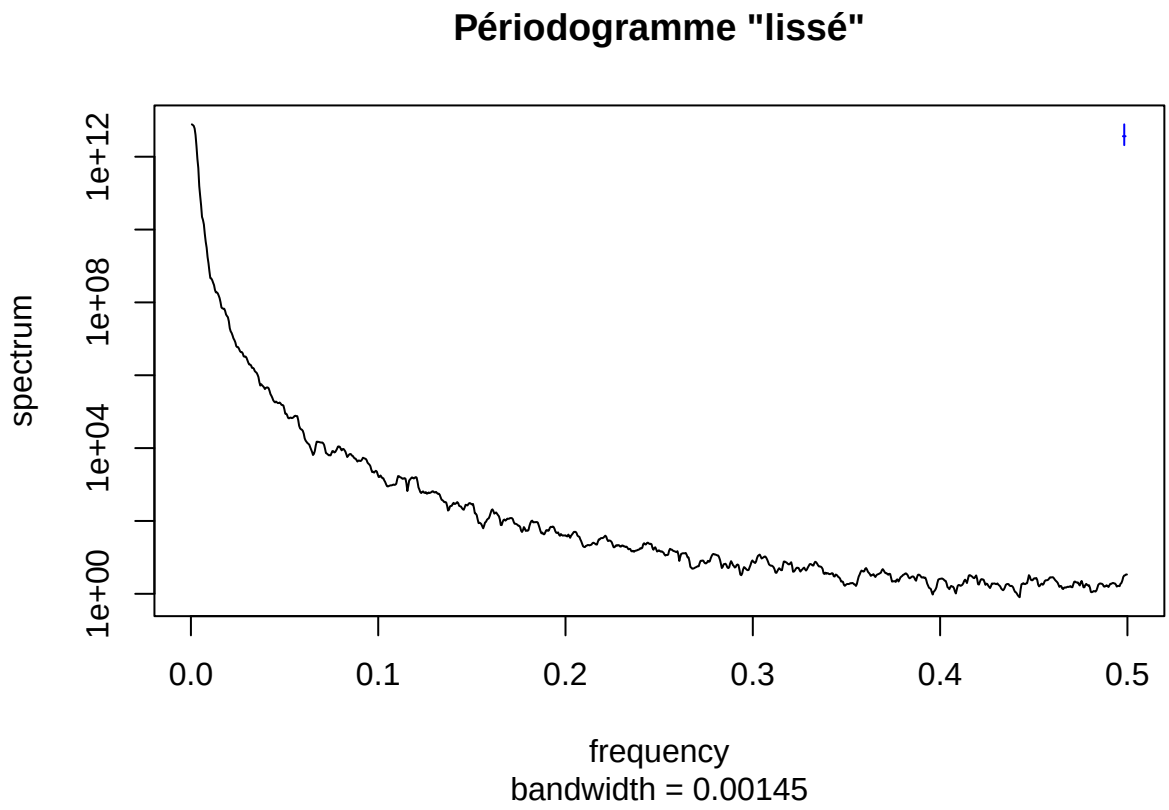
A première vue cette série semble composée de 3 morceaux: une étape initiale stationnaire puis une décroissance suivie d'une croissance. Cette série n'est pas stationnaire et ne présente visiblement pas de période. Le caractère lisse de cette série indique une absence de bruit significatif. (Lors d'un choix de modèle ARIMA il est vraisemblable que le terme de moyenne mobile soit d'ordre 0).

Etudions alors le périodogramme de la Série 2:

```
spectre = spectrum ( serie3$serie, spans=10 )
```



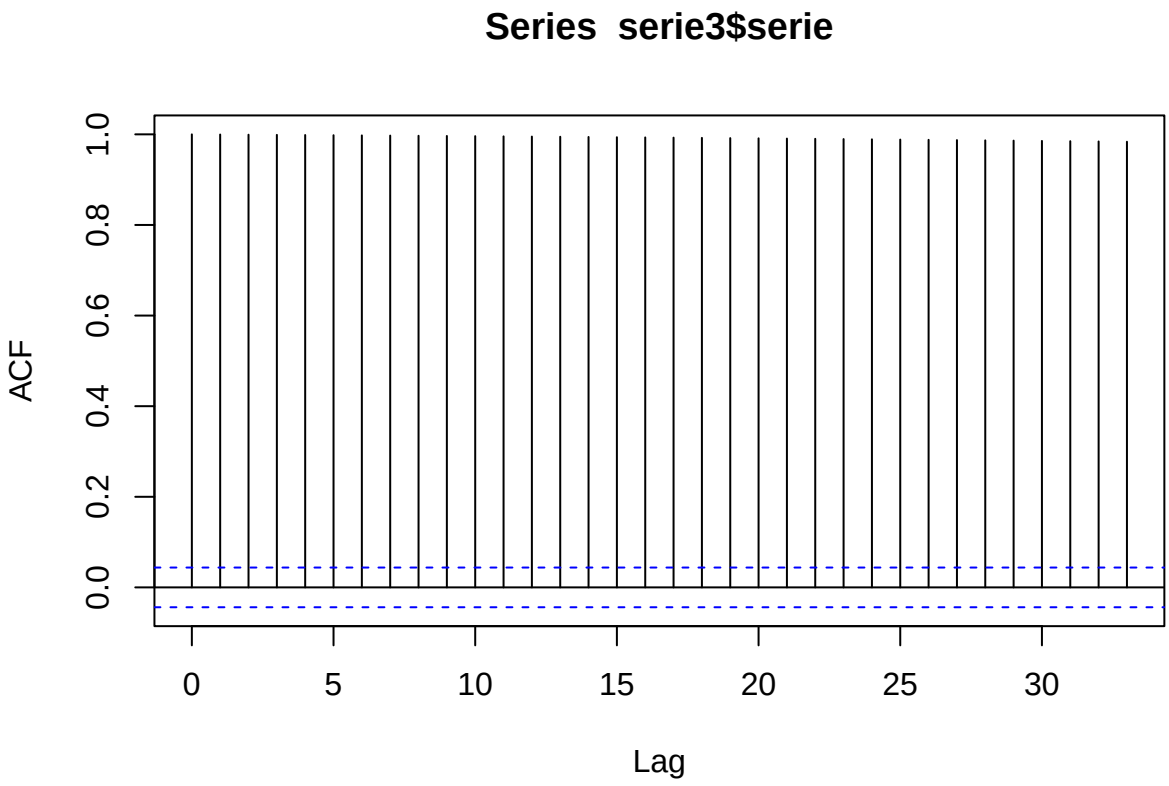

```
plot ( spectre , main = 'Périodogramme "lissé"')
```

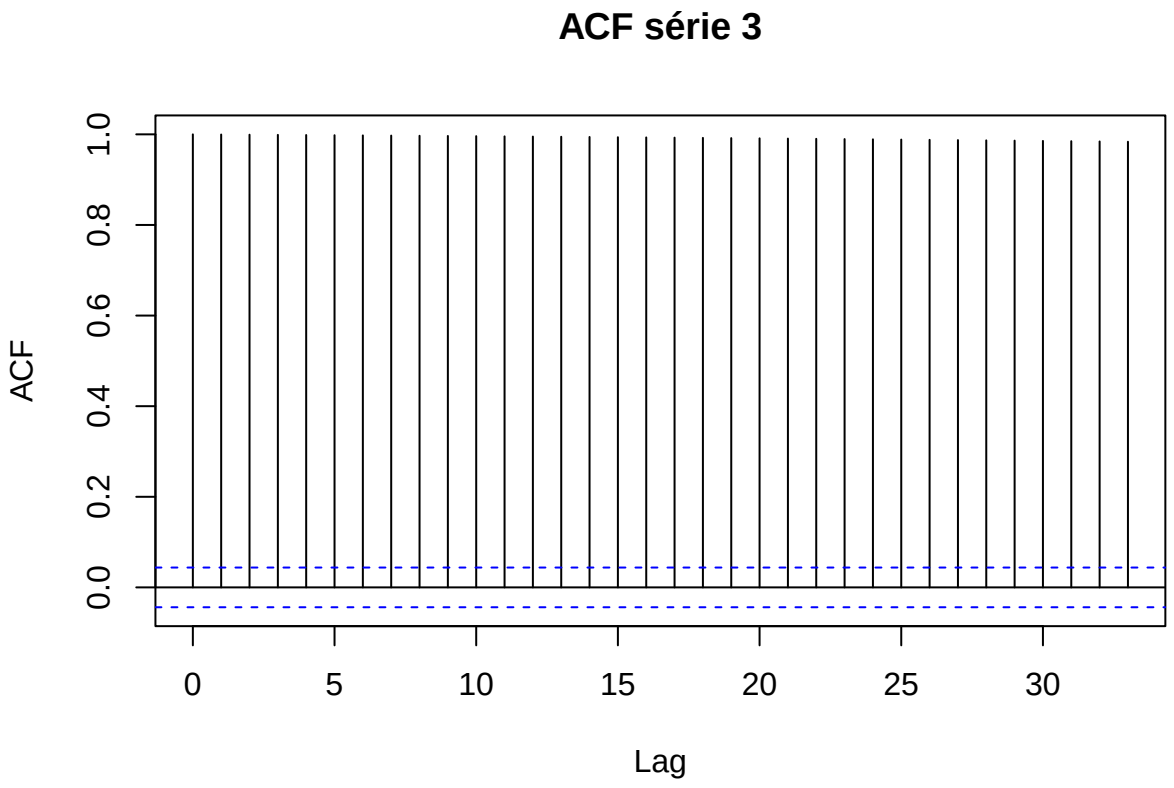


Le périodogramme vérifie bien cette absence de période (aucun pic n'y est présent).

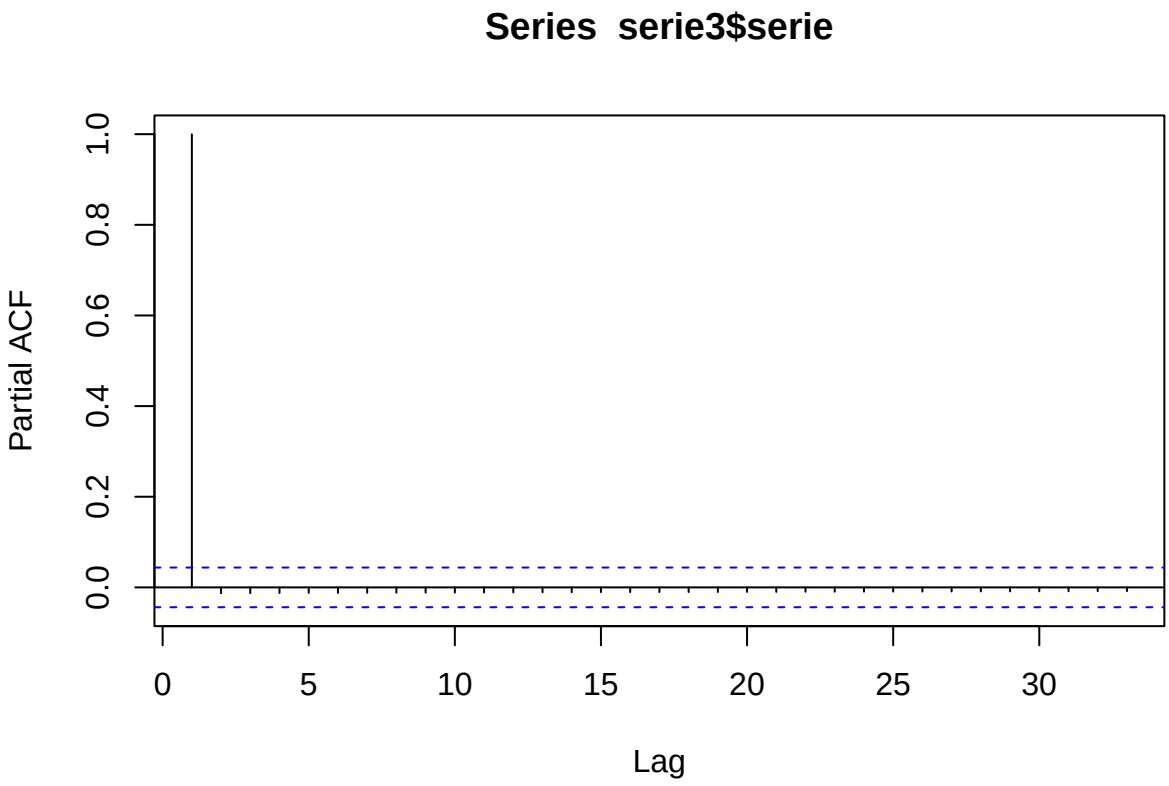
Choix du modèle Comme précédemment on trace l'ACF et la PACF:

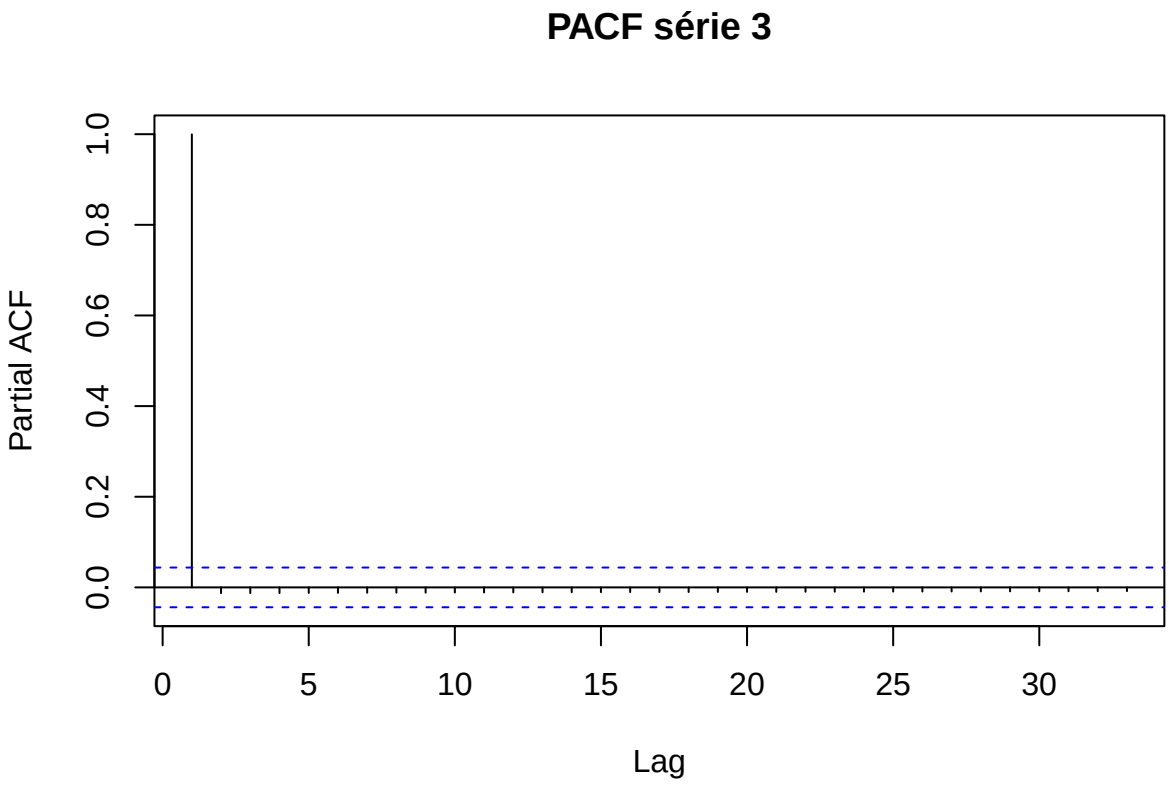
```
plot (acf(serie3$serie), main='ACF série 3')
```


Series serie3$serie
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag




```
plot (pacf(serie3$serie), main='PACF série 3')
```


Series serie3$serie
Partial ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag



L'ACF est constante tandis que la PACF ne présente qu'un seul pic significatif (en 1) ce qui indique "une corrélation uniquement du présent avec le passé immédiat".

Tout ceci couplé avec l'absence de périodicité ainsi que l'aspect lisse de cette série mènent à considérer un modèle ARIMA(1,j,0) avec j>1.

Nous proposons donc le modèle intermédiaire le plus simple: ARIMA(1,2,0) pour cette Série 3.

Vérification des hypothèses du modèle On génère donc notre modèle à partir de notre série:

```
##
## Call:
## arima(x = serie3$serie, order = c(1, 2, 0))
##
## Coefficients:
##          ar1
##       0.9052
## s.e.  0.0095
##
## sigma^2 estimated as 101.8:  log likelihood = -7461.59,  aic = 14927.18
```

AIC/BIC du modèle, optimisation On se demande si notre modèle est celui qui s'approche le mieux de notre série en ayant un nombre de paramètres raisonnable. Pour se faire on calcule l'AIC et le BIC des différents modèles ARIMA d'ordre d'intégration plus grand:

```
l_aic <-c()
l_bic <- c()
  for (j in 2:4){
    test = arima ( serie3$serie , order=c (1 , j , 0) )
l_aic <- c(l_aic, 1,j ,0, AIC(test))
l_bic <- c(l_bic, 1,j ,0, BIC(test))
}
```

On renvoie le meilleur choix de paramètre obtenu via l'AIC puis via le BIC:

```
M_aic <- matrix(l_aic, ncol=4, byrow=T)
M_bic <- matrix(l_bic, ncol=4, byrow=T)
Maic = as.data.frame(M_aic)
Mbic = as.data.frame(M_bic)
print('Meilleur AIC:')
```

```
## [1] "Meilleur AIC:"
```

```
print(Maic[Maic[4]==min(Maic[4])])
```

```
## [1]     1.00     2.00     0.00 14927.18
```

```
print('Meilleur BIC')
```

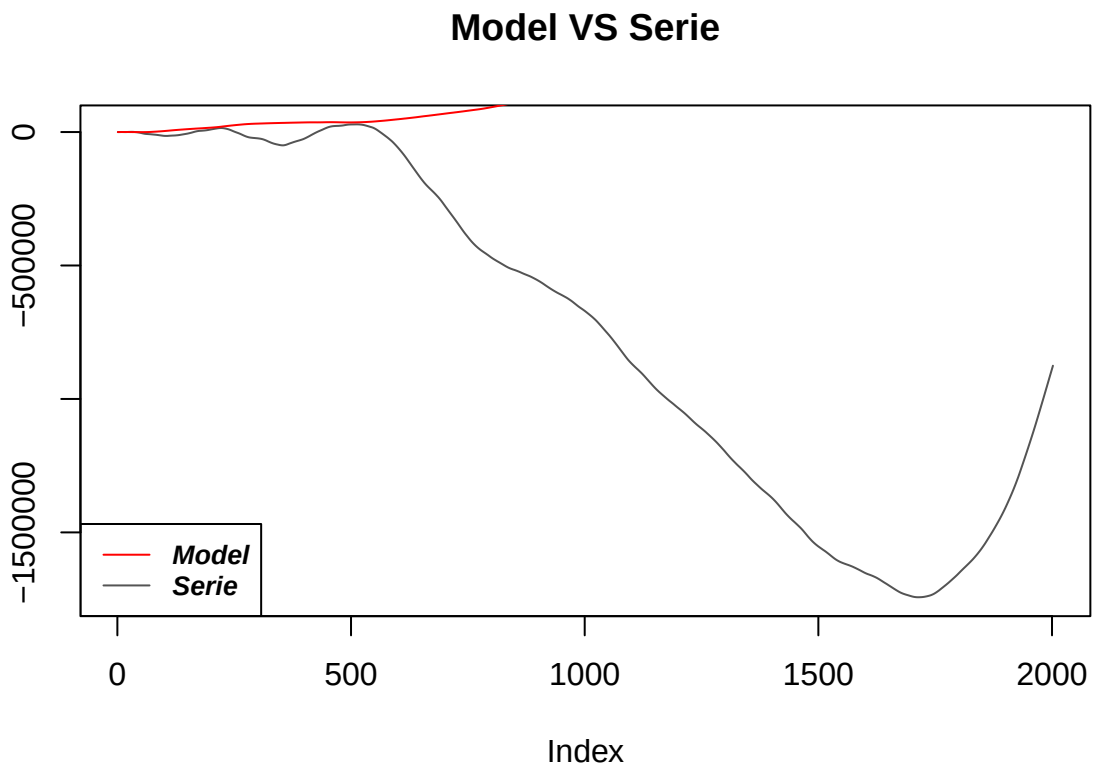
```
## [1] "Meilleur BIC"
```

```
print(Mbic[Mbic[4]==min(Mbic[4])])
```

```
## [1]     1.00     2.00     0.00 14938.39
```

Pour cette série 3, à la fois l'AIC et le BIC confirme le caractère "optimal" du modèle préliminaire choisi c'est-à-dire le modèle ARIMA(1,2,0).

```
plot(serie3$serie, col='grey33',main = 'Model VS Serie',ylab='',type='l')
lines(arima.sim(list(order=c(1,2,0), ar= c(model$coef[1])), n=length(serie3$serie)), col='red',type='l')
legend("bottomleft", legend=c("Model", "Serie"),lty=1, col=c("red","grey33"), cex=0.8,text.font=4)
```

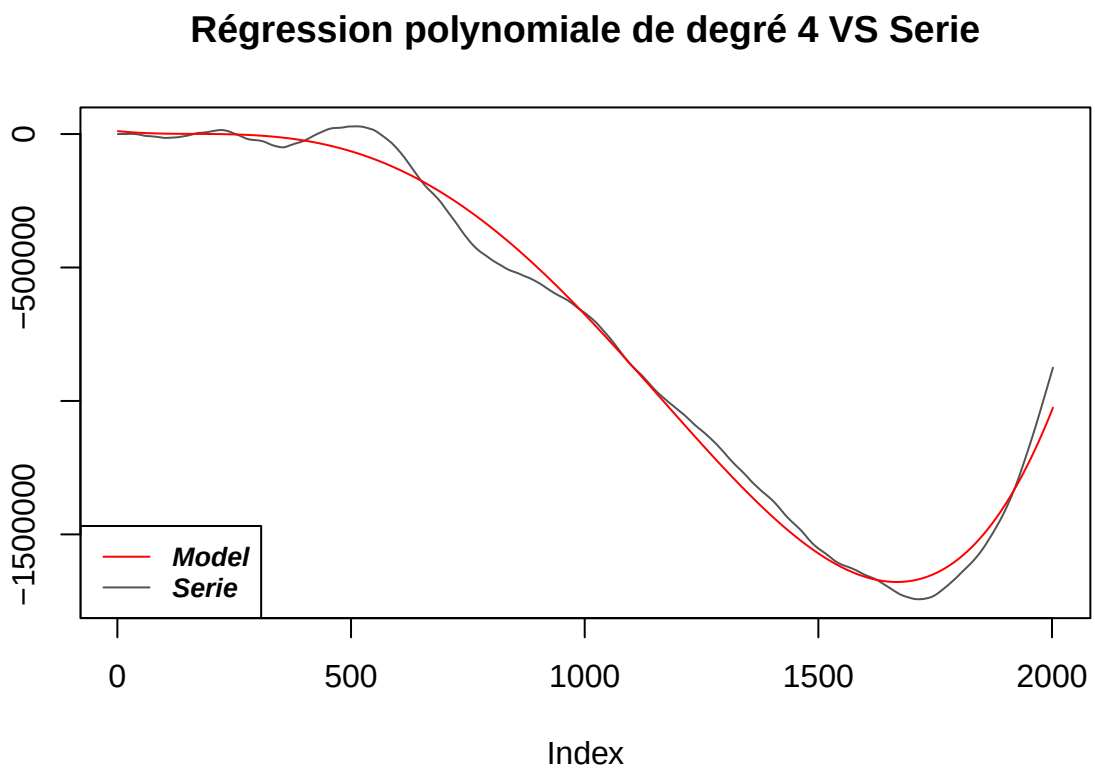


Conclusion: Après cette analyse nous optons pour le modèle ARIMA(1,2,0) pour cette Série 3.

Le choix d'un modèle ARIMA pour cette dernière série n'est peut-être pas le plus approprié, par exemple une régression polynomiale pourrait aussi être intéressante.

Par exemple avec une regression à l'aide d'un polynôme de degré 4 on obtient d'ores et déjà quelque chose de visiblement satisfaisant:

```
reg_model<-lm(serie3$serie~poly(c(1:length(serie3$serie)),4, raw=TRUE),data=as.data.frame(serie3$serie))
plot(serie3$serie, col='grey33',main = 'Régression polynomiale de degré 4 VS Serie',ylab='',type='l')
lines(reg_model$fitted.values, col='red',type='l')
legend("bottomleft", legend=c("Model", "Serie"),lty=1, col=c("red","grey33"), cex=0.8,text.font=4)
```



Derniers mots:

Pour chacune de ses séries nous avons établi les meilleurs modèles ARIMA permettant de les représenter, étant donnée l'absence d'information supplémentaire sur la nature des données, il est difficile de savoir si ces modèles sont cohérents et vraisemblables. Notamment pour la dernière série, nous avons évoqué l'idée qu'une régression polynomiale pourrait avoir plus de sens éventuellement.