R et la prévision des séries temporelles Chapitre 4 - Modèles ARIMA

Michel CARBON

Université Laval de Québec

13 mai 2019

Le problème de la variance

Il faut, quand la variance n'est pas temporellement constante, effectuer une transformation de type Box et Cox.

Cela a été vu au chapitre 2. J'insiste sur le fait que c'est la première chose à faire dans toute étude de série univariée.

Stationnarité

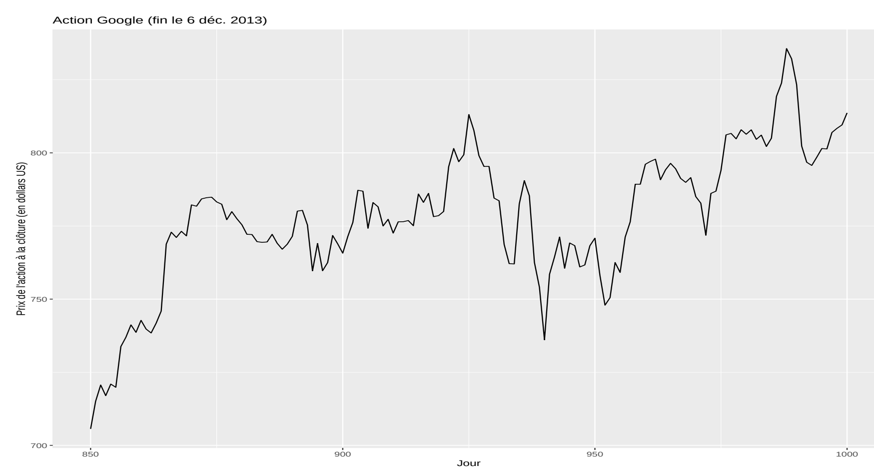
Une série stationnaire est une série dont les propriétés sont invariantes dans le temps.

Ainsi, les séries présentant une tendance ou une saisonnalité ne sont pas stationnaires.

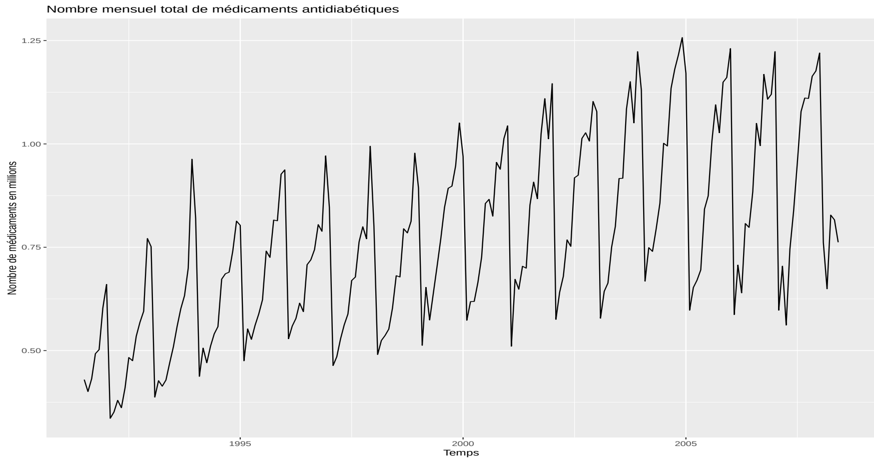
Un bruit blanc est stationnaire.

En général, une série stationnaire est globalement de moyenne constante, également de variance constante (même si un comportement cyclique est possible), sans composantes prédictibles.

Stationnarité - Exemple

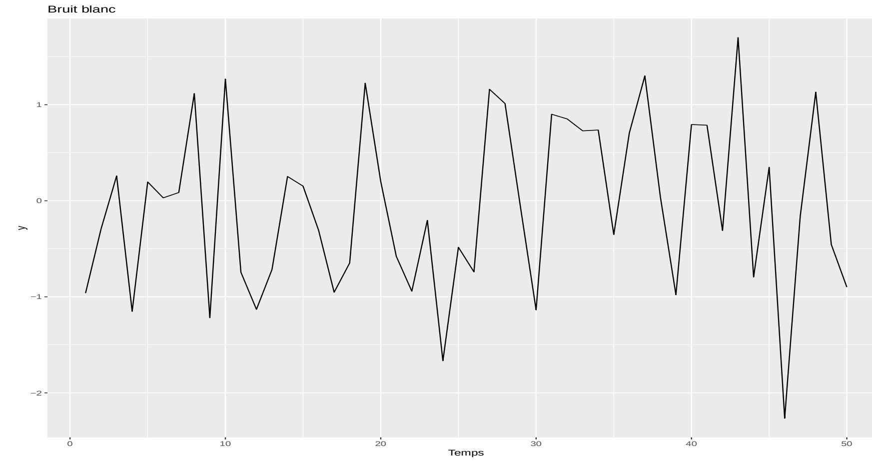


Stationnarité - Exemple



Stationnarité et différenciation Le modèle ARIMA - SARIM Le problème de la variance Stationnarité Différenciation Différe

Stationnarité - Exemple



Michel CARBON R et la prévision des séries temporelles Chapitre 4 - Modèl

Différenciation

La série "Google", notée (X_t), vue ci-dessus, présente une tendance croissante (la série est donc non stationnaire).

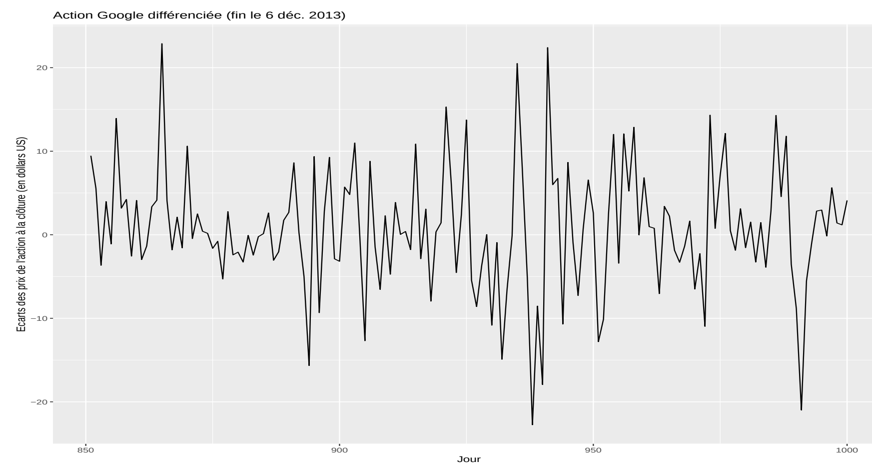
Mais par différenciation, on obtient une série stationnaire (Y_t) :

$$Y_t = (I - B)X_t = X_t - X_{t-1}$$

Michel CARBON R et la prévision des séries temporelles Chapitre 4 - Modèl

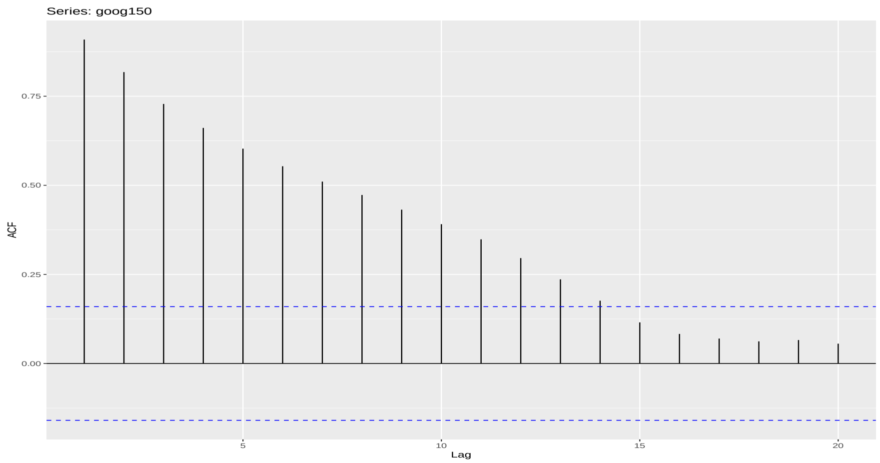
Stationnarité et différenciation Le modèle ARIMA - SARIM Le problème de la variance Stationnarité Différenciation Différe

Différenciation - Exemple

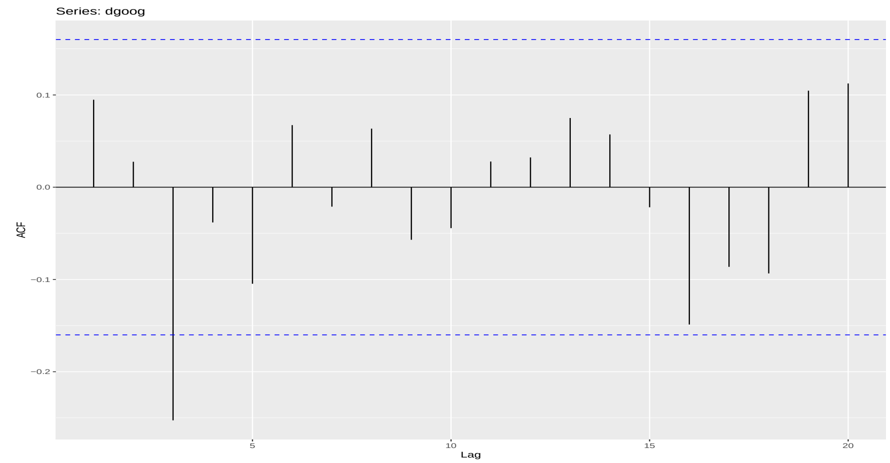


Michel CARBON R et la prévision des séries temporelles Chapitre 4 - Modèl

Différenciation - ACF de "Google" brute



Différenciation - ACF de "Google" différenciée

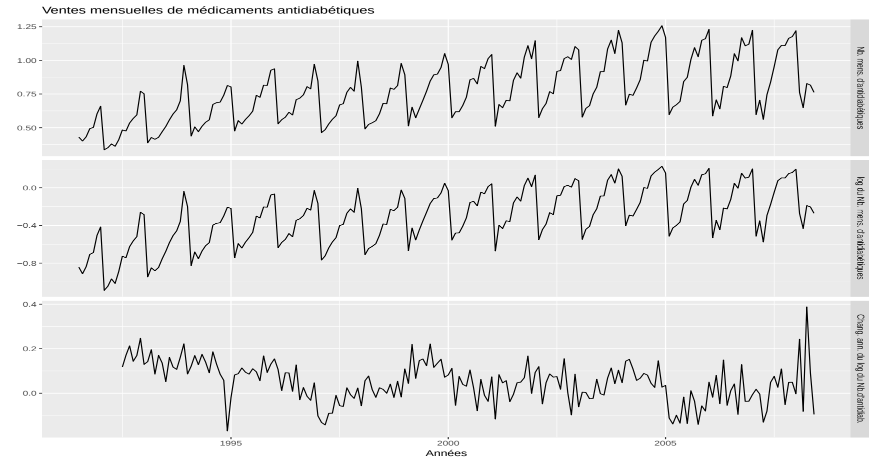


Différenciation saisonnière

On peut être amené à différencier saisonnièrement. Par exemple, sur la série (X_t) des "médicaments antidiabétiques", on est amené à différencier la saison ($s = 12$), donnant la série :

$$Y_t = (I - B^{12})X_t = X_t - X_{t-12}$$

Différenciation saisonnière - Exemple



Différenciation multiple

Il existe des tests statistiques pour savoir quel degré de différenciation appliquer :

1. Le test de Dickey-Fuller augmenté : l'hypothèse nulle est que la série est non stationnaire et non saisonnière.
2. Le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) : l'hypothèse nulle est que la série est stationnaire et non saisonnière.
3. D'autres tests sont applicables pour les séries saisonnières.

Différenciation multiple - Exemple

```
> adf.test(goog150)

Augmented Dickey-Fuller Test

data:  goog150
Dickey-Fuller = -3.1913, Lag order = 5, p-value = 0.09201
alternative hypothesis: stationary
```


Différenciation multiple - Exemple

```
> summary(ur.kpss(goog150))

#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 1.5337

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```


Le modèle ARIMA - SARIMA

Le modèle complet est :

$$\Phi_p(B)\,\Phi_P\left(B^S\right)\nabla^d\nabla_S^D X_t = \Theta_q(B)\,\Theta_Q\left(B^S\right)\varepsilon_t \qquad (1)$$

Notations

Notations :

S est la période de la saisonnalité ($S = 12$ dans l'exemple précédent, $S = 4$ pour des séries trimestrielles,...) .

$\nabla = 1 - B$; $\nabla_S = 1 - B^S$; $\Phi_p, \Phi_P, \Theta_q, \Theta_Q$ sont des polynômes de degrés respectifs p, P, q, Q dont les racines sont de modules strictement supérieurs à 1, et $(\varepsilon_t)_{t \in \mathbb{Z}}$ est un bruit blanc.

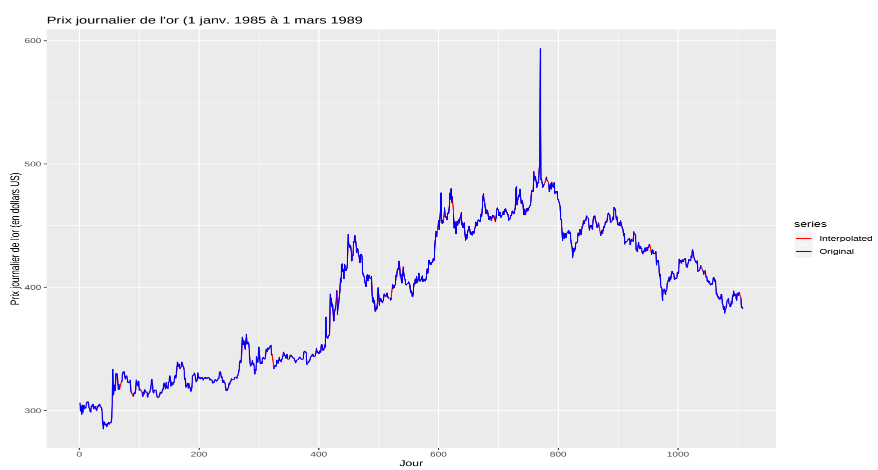
Un processus X_t satisfaisant (1) est appelé processus

$$SARIMA_S\ [(p, d, q), (P, D, Q)]$$

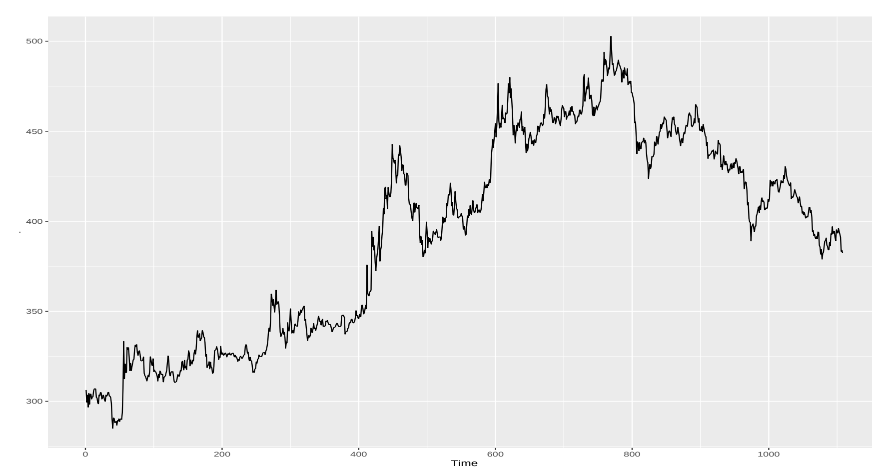
La commande "auto.arima"

1. Recherche éventuelle de points aberrants ou manquants.
2. Le cas échéant, stabilisez la variance via une transformation de Box et Cox.
3. Utilisez la commande "auto.arima" pour sélectionner un modèle.
4. Faire une analyse des résidus qui devraient être la réalisation d'un bruit blanc gaussien. Si cela n'a pas lieu, il faudra modifier le modèle.
5. Une fois cela réalisé, on peut utiliser le modèle pour faire des prévisions.

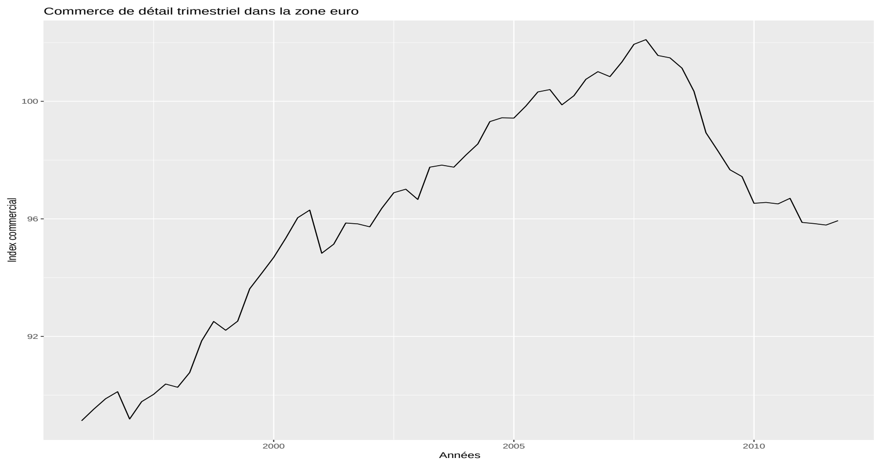
Valeurs manquantes



Valeurs manquantes et aberrantes - Commande "tsclean"



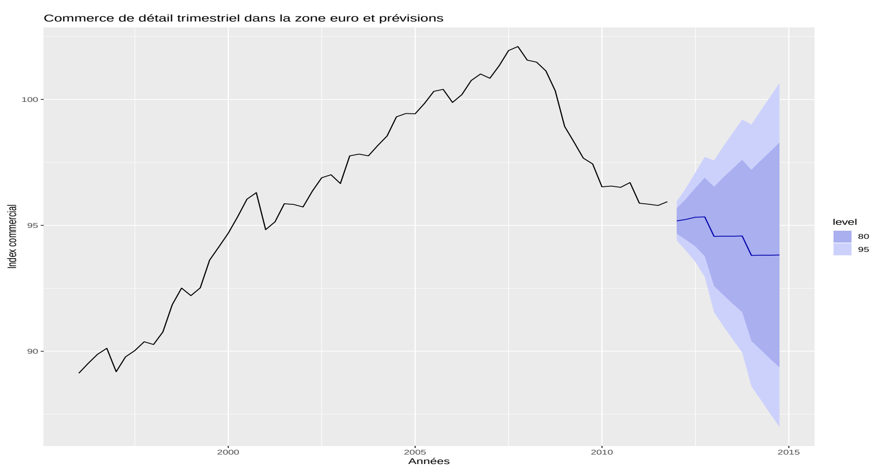
Commande "auto.arima" - Exemple



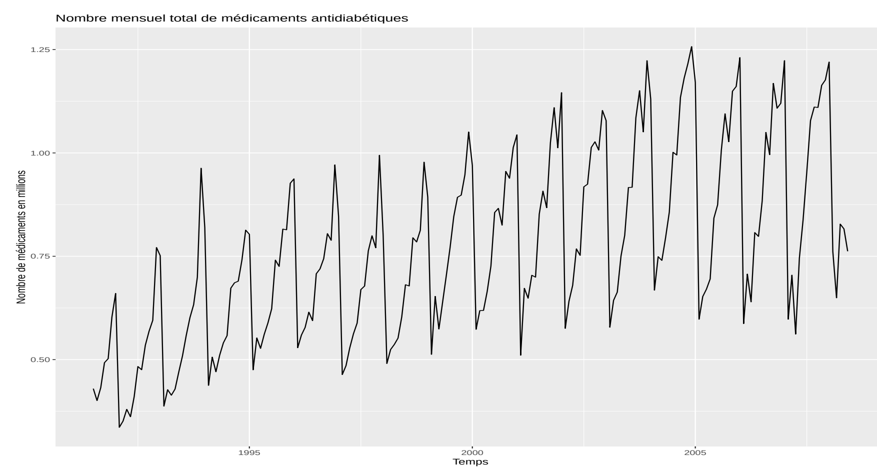
Commande "auto.arima" - Exemple

```
euretail %>% auto.arima() %>% forecast(h=12) %>% autoplot() +
  ylab("Index commercial") + xlab("Années")+
  ggtitle("Commerce de détail trimestriel dans la zone euro et

Series: .
ARIMA(0,1,3)(0,1,1)[4]

Coefficients:
         ma1     ma2     ma3     sma1
      0.2630  0.3694  0.4200  -0.6636
s.e.  0.1237  0.1255  0.1294   0.1545

sigma^2 estimated as 0.156:  log likelihood=-28.63
AIC=67.26   AICc=68.39   BIC=77.65
```


Commande "auto.arima" - Exemple



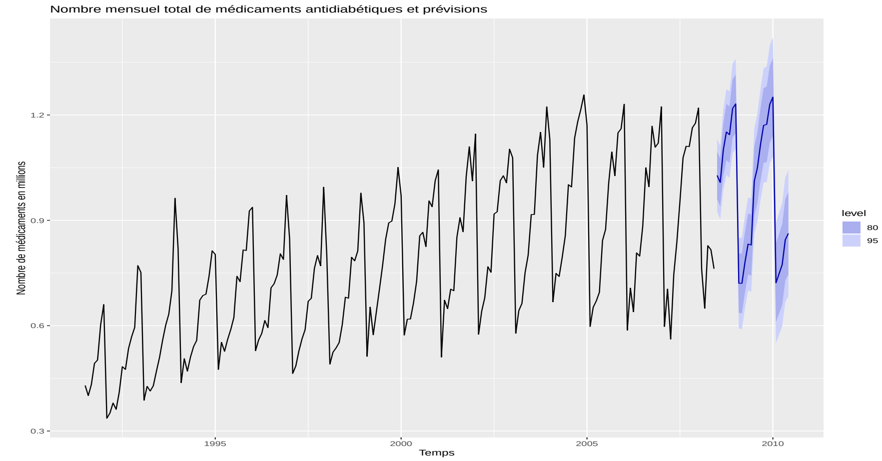
Commande "auto.arima" - Exemple



Commande "auto.arima" - Exemple

```
h02 %>% auto.arima()%>% forecast(h=24) %>% autoplot()+
  ggtitle("Nombre mensuel total de médicaments antidiabétiques et prév
  xlab("Temps")+
  ylab("Nombre de médicaments en millions")

 Series: .
ARIMA(4,1,1)(0,1,2)[12]

Coefficients:
         ar1     ar2     ar3      ar4      ma1     sma1     sma2
      0.0888  0.3386  0.2302  -0.2233  -0.9068  -0.4798  -0.1624
s.e.  0.1063  0.0976  0.0894   0.0850   0.0853   0.0913   0.0930

sigma^2 estimated as 0.00276:  log likelihood=291.7
AIC=-567.4   AICc=-566.6   BIC=-541.38
```


Commande "auto.arima" - Exemple

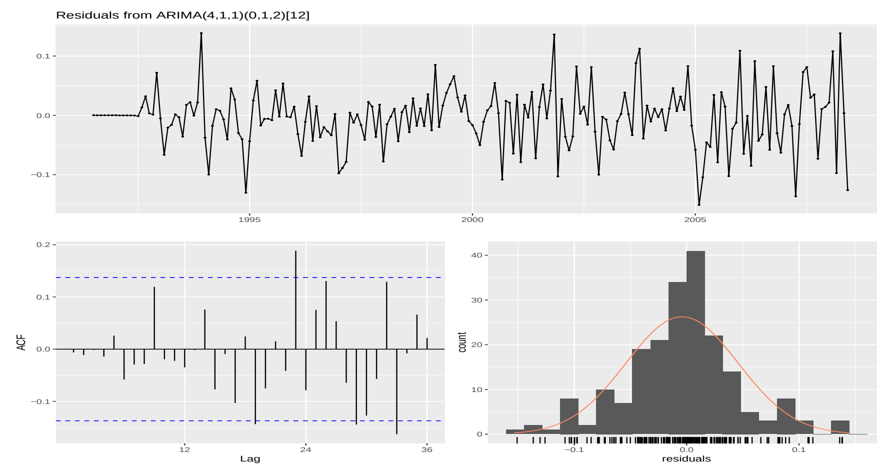


Commande "auto.arima" - Exemple

On doit analyser les résidus :

```
fit.arima<-auto.arima(h02)
checkresiduals(fit.arima)

Ljung-Box test

data:  Residuals from ARIMA(4,1,1)(0,1,2)[12]
Q* = 26.081, df = 17, p-value = 0.07299

Model df: 7.   Total lags used: 24
```


Commande "auto.arima" - Exemple

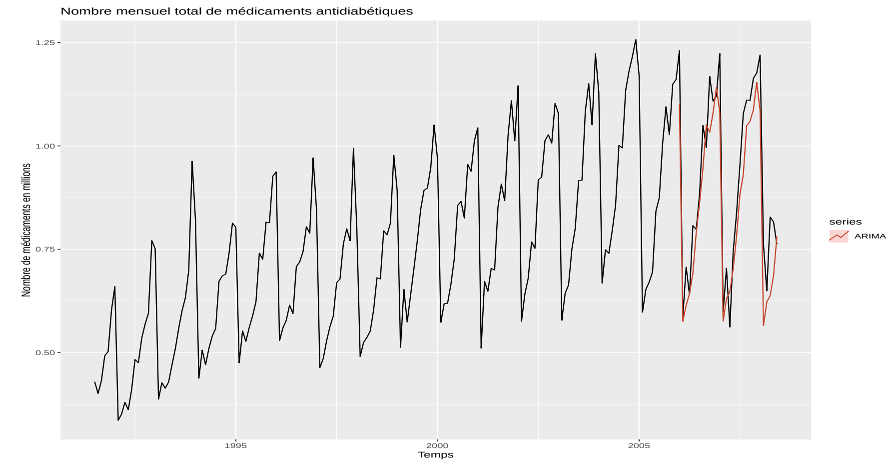


Commande "auto.arima" - Exemple

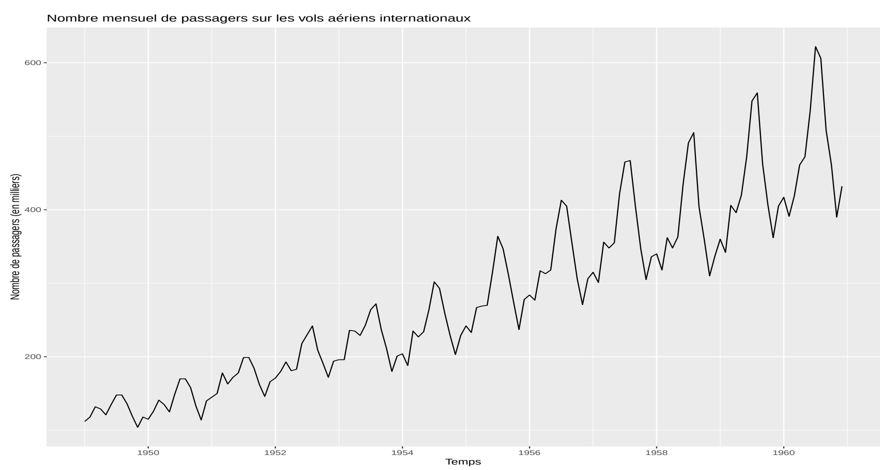
On peut faire un test d'évaluation de la pertinence des prévisions.
Ici, on va retrancher 2 ans.

```
train <- window(h02,end=c(2005,12))
(fit.arima<-auto.arima(train))
diab1<-fit.arima %>% forecast(h=24)%>%
  accuracy(h02)
diab1[,c("RMSE","MAPE","MASE")]

                   RMSE     MAPE      MASE
Training set 0.04588769 4.568735 0.5600418
Test set     0.07927912 7.013943 1.0942005
```


Commande "auto.arima" - Exemple



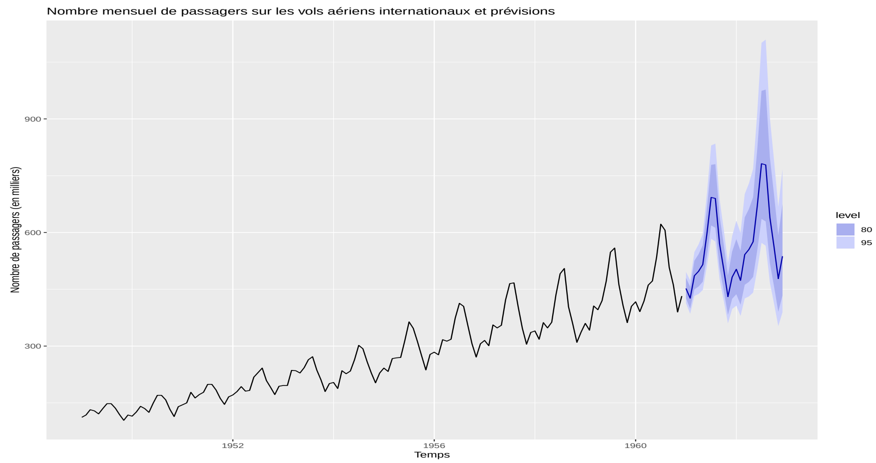
Commande "auto.arima" - Exemple



Commande "auto.arima" - Exemple

```
airline %>% auto.arima(lambda="auto") %>% forecast(h=24)%>%
 autoplot()+
  ggtitle("Nombre mensuel de passagers sur les vols aériens interna
  xlab("Temps")+ ylab("Nombre de passagers (en milliers)")

  Series: .
ARIMA(0,1,1)(0,1,1)[12]
Box Cox transformation: lambda= -0.2947156

Coefficients:
          ma1     sma1
      -0.4355  -0.5847
s.e.   0.0908   0.0725

sigma^2 estimated as 5.855e-05:  log likelihood=451.6
AIC=-897.19   AICc=-897.01   BIC=-888.57
```

Commande "auto.arima" - Exemple

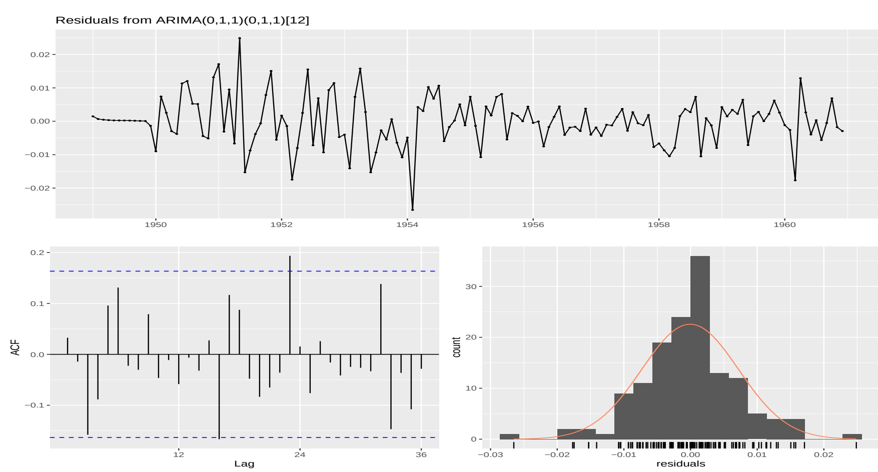


Commande "auto.arima" - Exemple

On doit analyser les résidus :

```
fit1.arima<-auto.arima(airline,lambda="auto")
checkresiduals(fit1.arima)

Ljung-Box test

data:  Residuals from ARIMA(0,1,1)(0,1,1)[12]
Q* = 28.635, df = 22, p-value = 0.1556

Model df: 2.   Total lags used: 24
```


Commande "auto.arima" - Exemple



Exercice

À vous !

1. Importez la série "electricity.dat", série mensuelle débutant en janvier 1986. Faites-en un graphique.
2. En utilisant un modèle SARIMA, faites des prévisions sur 2 ans.
3. Tronquez fictivement les 60 dernières observations et faites des prévisions via un modèle SARIMA et une approche "ets". Ì
4. Comparez l'efficacité prévisionelle des deux approches.