Les processus ARIMA

MAP-STA2 : Séries chronologiques

Yannig Goude yannig.goude@edf.fr

2021-2022

Contents

Les processus ARIMA

Nous avons vu dans un chapitre antérieur comment corriger une série non-stationnaire de composantes déterministes telles qu'une tendance ou une saisonnalité. Entre autre, nous avons étudié l'opérateur de différenciation (rappel):

Notons Δ l'opérateur de différenciation: $\Delta X_t = X_t - X_{t-1}$. L'opérateur de différenciation d'ordre k correspondant est: $\Delta^k X_t = \Delta(\Delta^{k-1} X_t)$

Propriété soit un processus y admettant une tendance polynomiale d'ordre k:

$$y_t = \sum_{j=0}^{k} a_j t^j + \varepsilon_t$$

alors le processus Δy admet une tendance polyomiale d'ordre $k-1$.

Avec les notations des modèles ARMA, on peut remarquer que $\Delta X_t = (1-L)X_t$ et plus généralement $\Delta^d X_t = (1-L)^d X_t$. Ainsi un processus ARIMA est défini ainsi:

définition un processus stationnaire X_t admet une représentation ARIMA(p,d,q) minimale s'il satisfait

$$\Phi(L)(1-L)^d X_t = \Theta(L)\varepsilon_t, \ \ \forall t \in \mathbf{Z}$$

avec les conditions suivantes:

1. $\phi_p \neq 0$ et $\theta_q \neq 0$
2. Φ et Θ, polynômes de degrés resp. p et q, n'ont pas de racines communes et leurs racines sont de modules > 1
3. ε_t est un BB de variance σ^2

Un processus ARIMA(p,d,q) convient pour modéliser une série temporelle comprenant une tendance polynômiale de degrés d, l'opérateur $(1-L)^d$ permettant de transformer un polynôme de degré d en une constante.

Pour estimer les paramètres d'un modèle ARIMA, on procède de même que pour un ARMA sur le processus différencié $(1-L)^d X_t$.

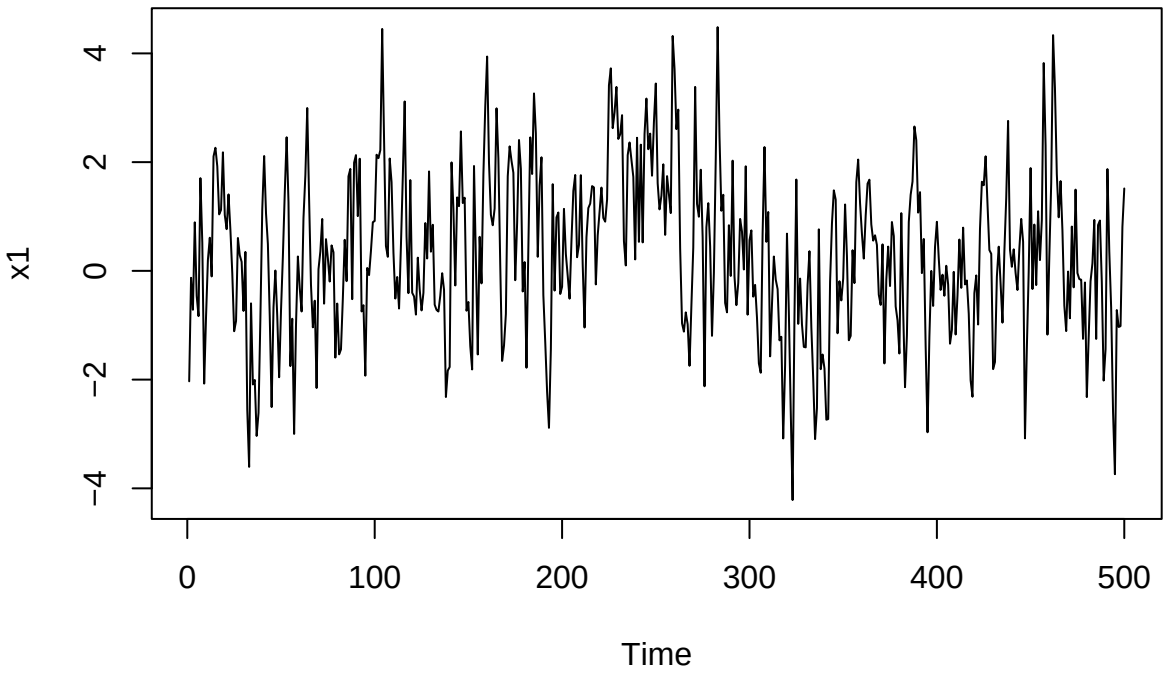
Les processus SARIMA

Certaines série chronologiques présentent des saisonnalités. Nous avons déjà abordé le cas de séries dont les composantes saisonnières sont déterministes et nous avons procédé par des méthodes de régressions (paramétriques, non-paramétriques) ou par différenciation pour les désaisonnaliser et ensuite leur appliquer une modélisation ARMA. Une autre démarche consiste à intégrer, dans un modèle ARIMA, des décalages multiples de la saisonnalité (par exemple retard de 12 pour des séries mensuelles possédant une saisonnalité annuelle). En théorie, si on choisit des ordres p et q suffisamment important, ces retards sont naturellement intégrés au modèle ARIMA(p,d,q), l'inconvénient étant qu'on rajoute un grand nombre de retards intermédiaires potentiellement non-significatifs.

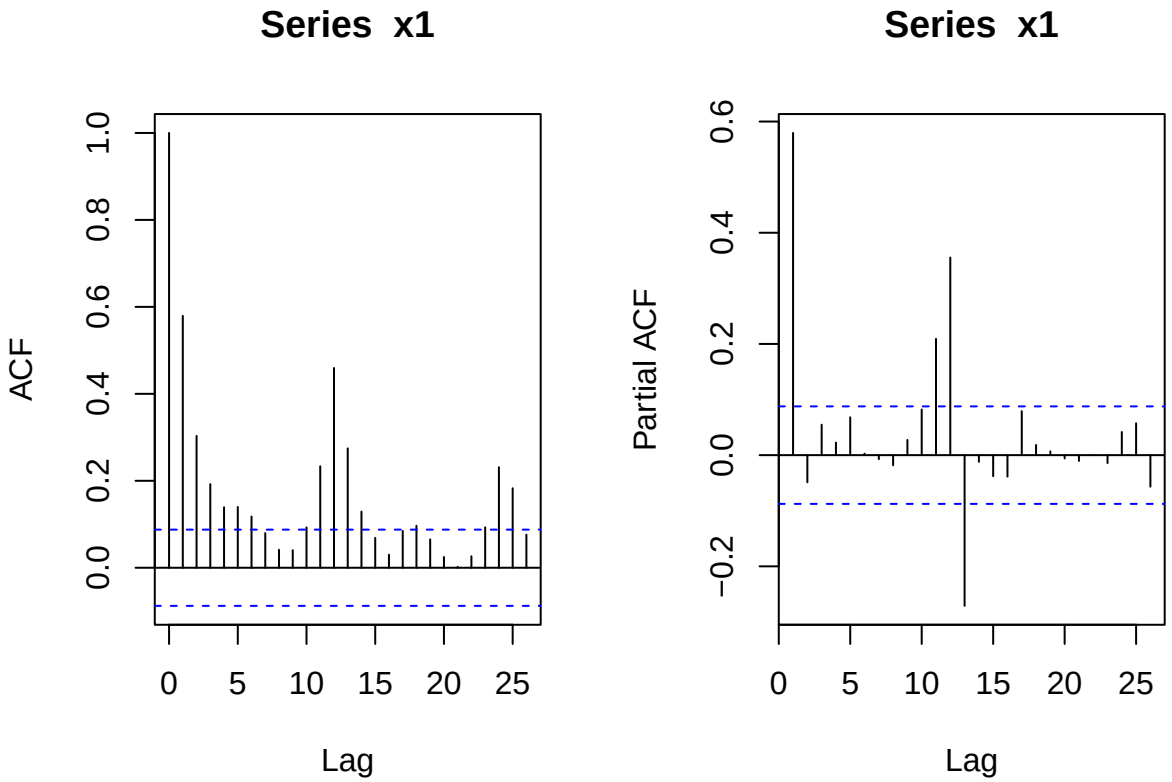
Par exemple, la série suivante:

```
set.seed(100)
n<-500
sigma<-1
eps<-rnorm(n,0,sd=sigma)
x1<-arima.sim(n = n, list(ar =c(.6,rep(0,10),.5,-.30),innov=eps))

plot(x1,type='l')
```



```
par(mfrow=c(1,2))
acf(x1)
pacf(x1)
```



Où les lag 12 et 13 sont significatifs alors que les lag intermédaires ne sont pas significativement différents de 0. L'autocorrélogramme indiquant clairement une saisonnalité (pas immédiat en se basant sur la trajectoire).

Box et Jenkins propose des modèles multiplicatifs SARIMA définis ainsi.

définition un processus stationnaire X_t admet une représentation SARIMA[(p,d,q);(P,D,Q);s] minimale s'il satisfait

$$(1-L)^d\Phi_p(L)(1-L^s)^D\Phi_P(L^s)X_t = \Theta_q(L)\Theta_Q(L^s)\varepsilon_t, \quad \forall t \in \mathbf{Z}$$

avec des conditions similaires que pour les ARIMA.

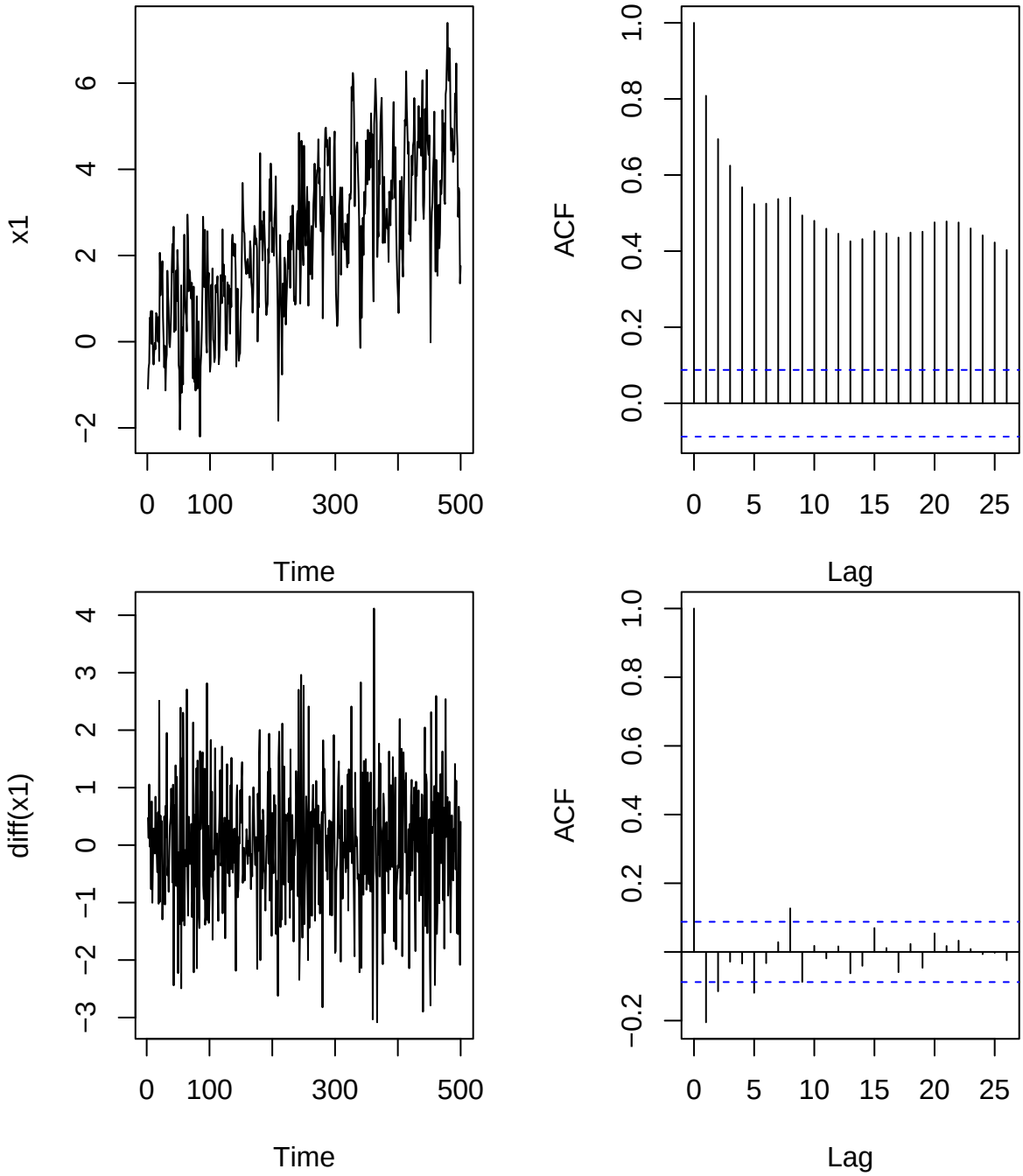
s correspond à la période du processus SARIMA qu'on peut identifier en regardant l'autocorrélogramme (cf exemple précédant) ou la densité spectrale. Les entiers d et D sont choisis de sorte que la série différenciée: $(1-L)^d(1-L^s)^D$ soit stationnaire (en pratique regarder la trajectoire, les autocorrélations). Les ordres p et q s'obtiennent comme pour les modèles ARMA(p,q) classiques (autocorrélation partielle et simple), les ordres P et Q en regardant les ordre multiples de s de l'autocorrélogramme. En pratique, on commencera par différencier en $1-B^s$ (choisir D) avant de choisir d car $1-B^s = (1-B)(1+B+B^2+...+B^{k-1})$.

L'approche de Box-Jenkins

L'approche de box jenkins, du nom des statisticiens George Box et Gwilym Jenkins, est une méthode empirique de choix et construction de modèle SARIMA mise au point dans les années 1970. Pour simplifier les choses nous nous intéressons ci-dessous aux processus ARIMA.

- le premier paramètre à choisir est le degré de différentiation d. Plusieurs moyens sont possibles pour détecter une non-stationnarité. La représentation temporelle de la série peut faire apparaitre des tendances polynomiale. On peut aussi calculer les autocorrélation empiriques et analyser la vitesse de décroissance vers 0 de cette fonction, si cette décroissance est lente (plus lente qu'exponentiel) on peut suspecter une non-stationnarité. Notons qu'en pratique le cas $d > 2$ est rarement rencontré. En effet, il est dangereux de sur-différencier un processus, cela peut conduire à une non-inversibilité des ARMA associés.

Par exemple dans le cas du processus X_t suivant:



- les ordres p et q des modèles AR et MA sont ensuites obtenus en regardant les autocorrélations et autocorrélations partielles. Bien souvent, ces ordre n'apparaissent pas de manière évidente. On peut

dans ce cas obtenir des bornes supérieures pour p et q puis sélectionner un modèle en minimisant un critère pénalisé de type AIC ou BIC.

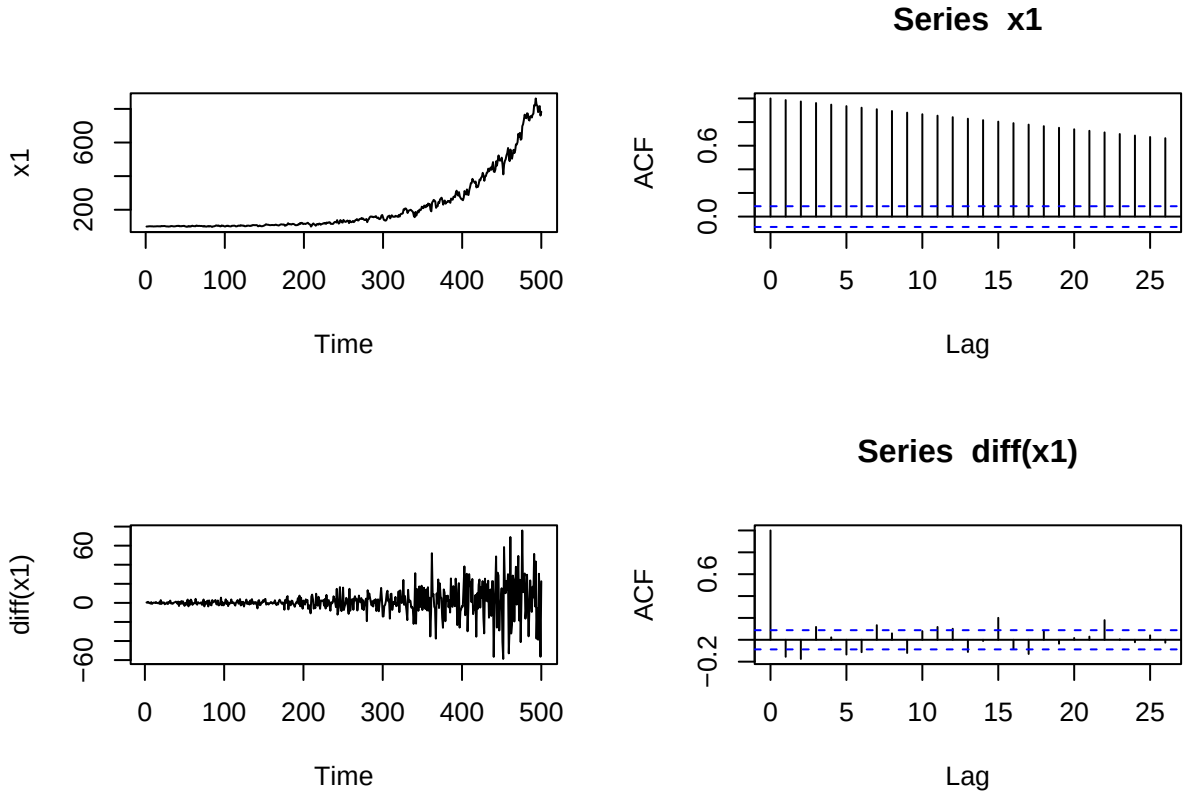
- l'estimation des paramètres se fait ensuite par les méthodes précédemment citées (maximum de vraisemblance, moindres carrés conditionnels).
- d'autres approches que la différenciation sont possibles pour "stationnariser" un processus. Par exemple, il est clair que si un processus possède une tendance non-polynomiale, par exemple exponentielle ou logarithmique, l'opérateur $(1-L)^d$ ne suffira pas à le rendre stationnaire. On peut alors soit modéliser par régression la partie déterministe du processus (cf théorême de Wold), ou appliquer des transformations à la série. Une transformation courante est le passage au log de la série (pour gérer des phénomènes de variance multiplicative par exemple). On peut généraliser ce type de transformation sous le terme de transformation de Box-Cox:

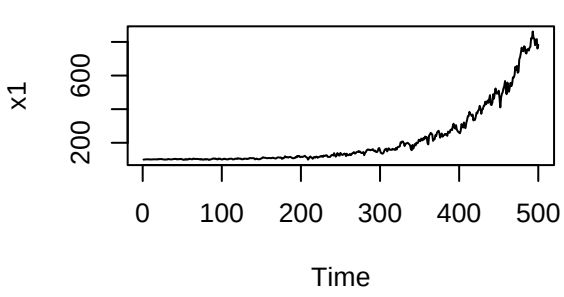
$$\text{BoxCox}_\lambda(x) = \frac{x^\lambda - 1}{\lambda}, \quad \lambda > 0$$

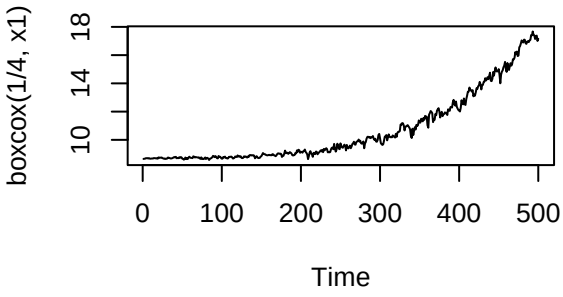
On peut remarquer que:

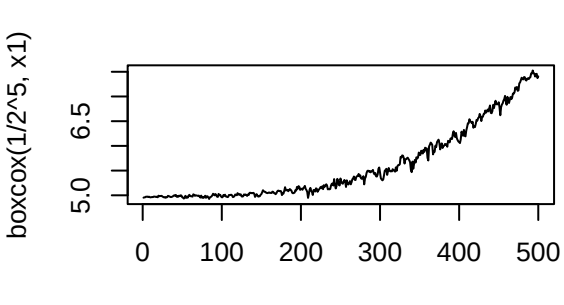
$$\frac{x^\lambda - 1}{\lambda} = \frac{1}{\lambda}(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + ... - 1) = \log(x) + \frac{1}{2}\lambda \log(x)^2 + ...$$

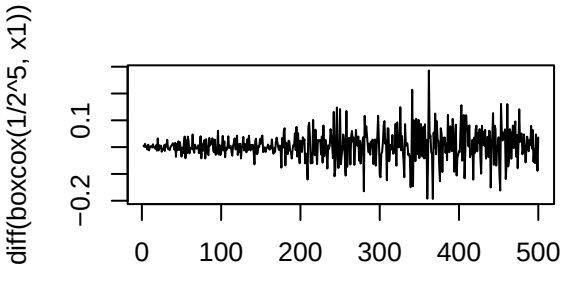
et donc que $\lambda = 0$ correspond à une transformation logarithmique.

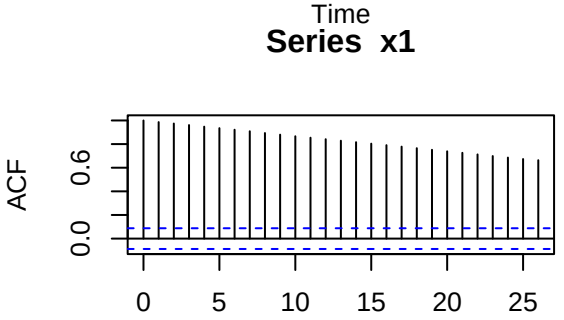


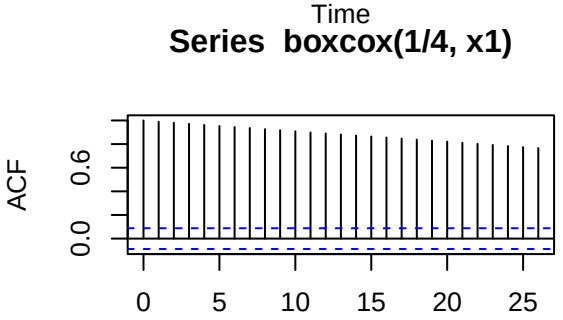


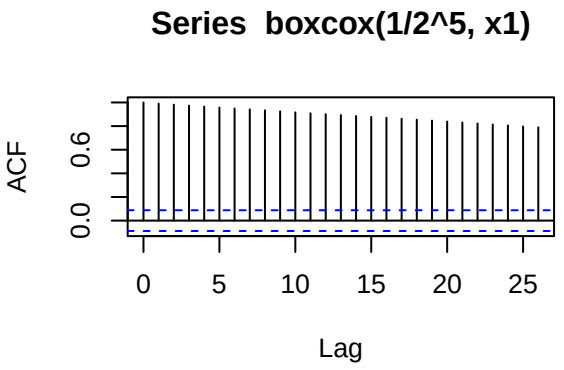


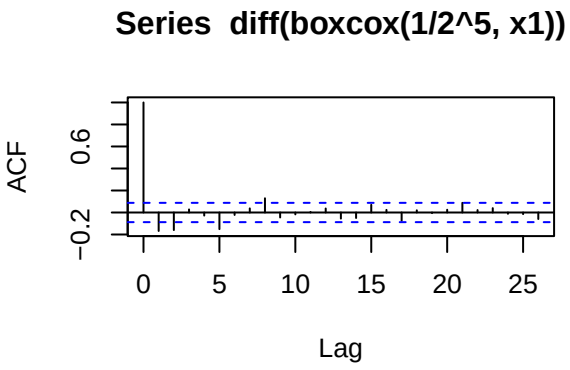












Diagnostics et tests

Une fois le modèle estimé, il est important de valider ou non nos choix de modélisation. Pour cela différents tests et diagnostics existent.

Analyse des résidus

Une fois un modèle ARIMA estimé, on peut obtenir une estimation de ε_t par:

$$\widehat{\varepsilon_t} = \widehat{\Theta}(L)^{-1}\widehat{\Phi}(L)(1-B)^d X_t$$

On commence par estimer la fonction d'autocorrélation des résidus (ACF empirique):

$$\widehat{\rho}_\varepsilon(h) = \frac{\sum_{t=1}^{n-h} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t+h}}{\sum_{t=1}^{n} \widehat{\varepsilon}_t^2}$$

idéalement, on aimerait obtenir la fonction d'autocorrélation d'un bruit blanc: $\widehat{\rho}_\varepsilon(h>0) = 0$. Si ce n'est pas le cas, il faut remettre en cause le choix des ordres ou la correction de la tendance (dans ce cas visualiser la trajectoire de ε peut s'avérer instructif).

Tests du portemanteau

Il existe 2 variantes du test du portemanteau (signifie fourre tout en anglais), test de blancheur d'une série (ici les résidus d'une modélisation SARIMA).

Test de Box-Pierce il permet de tester l'hypothèse que les résidus d'une série X_t suivant une modélisation ARMA(p,q) sont un bruit blanc ie, pour une série X_t et ses résidus associés $\widehat{\varepsilon_t} = \widehat{\Theta}(L)^{-1}\widehat{\Phi}(L)(1-B)^d X_t$ de fonction d'autocorrélation $\rho_\varepsilon(h)$ et son estimateur empirique asssocié:

$$H_0(h) : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(h) = 0$$

$$H_1(h) : \exists k \in (1, ..., h) \text{ t.q } \rho_\varepsilon(k) \neq 0$$

Il se base sur la statistique de Box-Pierce:

$$Q_{BP}(h) = n\sum_{j=1}^{h} \widehat{\rho}_\varepsilon(j)^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(h-k)$$

où k est le nombre de paramètre du modèle qu'on considère et vaut $p+q$ ou $p+q+1$ pour un ARMA(p,q) (avec constante ou non).

Test de Ljung-Box Il s'agit d'une variante du test précédant préconisée dans le cas ou la taille n de la trajectoire est petite.

$$Q_{LB}(h) = n(n+2)\sum_{j=1}^{h} \frac{\widehat{\rho}_\varepsilon(j)^2}{n-j} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(h-k)$$

Tests de significativité des paramètres

L'idée est ici de comparer des modèles ARMA(p,q) emboités pour savoir s'il est judicieux de reduire ou augmenter les ordres p et q. Il en résulte un algorithme itératif pour corriger ou affiner les ordre préalablement déterminés.

1. comparaison ARMA(p-1,q) ou ARMA(p,q-1) avec ARMA(p,q)

diminuer d'une unité l'ordre de l'AR(p) ou du MA(q) revient à tester la significativité du coefficient ϕ_p (resp. θ_q) ce qui peut être fait par un test de student, les estimateurs $\widehat{\phi}_p$ et $\widehat{\theta}_q$ obtenus par maximum de vraisemblance ayant les mêmes propriétés qu'en régression linéaire.

Soit $\widehat{\phi}_p$ le coefficient estimé et sa variance estimée $\sigma^2_{\widehat{\phi}_p}$, en faisant l'hypothèse que $n-p$ est grand (approximation d'un student $n-p$ par une gaussienne), on accepte le modèle ARMA(p-1,q) avec une erreure de première espèce de 5% si:

$$\frac{|\widehat{\phi}_p|}{\sigma_{\widehat{\phi}_p}} < 1.96$$

2. comparaison ARMA(p+1,q) ou ARMA(p,q+1) avec ARMA(p,q)

on peut estimer un modèle ARMA(p+1,q) (resp. ARMA(p,q+1)) puis tester la significativité de $\widehat{\phi}_{p+1}$ (resp. θ_{q+1}) et se ramener au point 1).

3. comparaison ARMA(p+1,q+1) avec ARMA(p,q)

on ne peut pas tester les deux ordres simultanément. En effet, un processus admettant une représentation ARMA(p,q) admet aussi une représentation ARMA(p+1,q+1) (il suffit de multiplier les deux polynômes Φ et Θ par $1+aL$ par exemple).

Synthèse

En résumé la démarche empirique pour effectuer une modélisation ARIMA(p,d,q) est la suivante:

1. Stationnariser la série (régression, différenciation, transformation), vérifier avec l'autocorrélogramme.
2. Déterminer des ordres p et q plausibles à l'aide de respectivement l'autocorrélogramme partiel et l'autocorrélogramme.
3. Estimer les paramètres, si plusieurs modèles candidats les départager par l'AIC ou le BIC.
4. Valider ou non le modèle par un diagnostique des résidus (test, représentation graphique, autocorrélogramme).
5. Confirmer votre choix en simulant de la prévision (échantillon test).

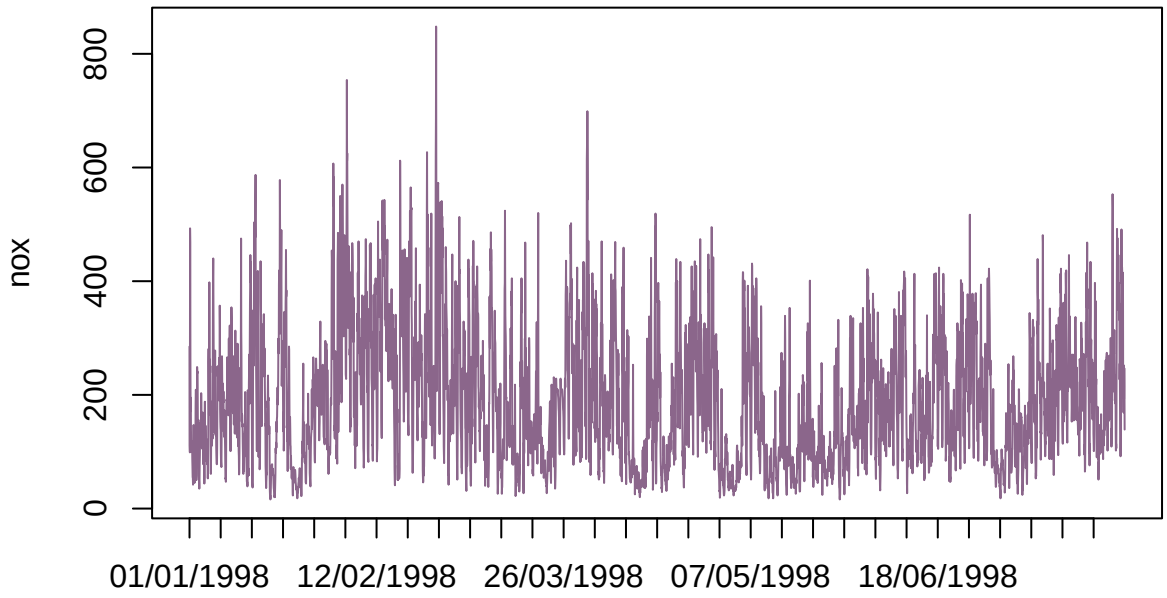
Dans le cas de modèles SARIMA[(p,d,q);(P,D,Q)]:

1. Identifier la saisonnalité s (autocorrélogramme, spectrogramme)
2. Stationnariser la série, en commençant par D puis d
3. Déterminer des ordres p et q plausibles à l'aide de l'autocorrélogramme partiel et l'autocorrélogramme. Les ordres P et Q en regardant les ordre mulitples de s de ces autocorrélogrammes.
4. Estimer les paramètres, si plusieurs modèles candidats les départager par l'AIC ou le BIC.
5. Valider ou non le modèle par un diagnostique des résidus (test, représentation graphique, autocorrélogramme).
6. Confirmer votre choix en simulant de la prévision (échantillon test).

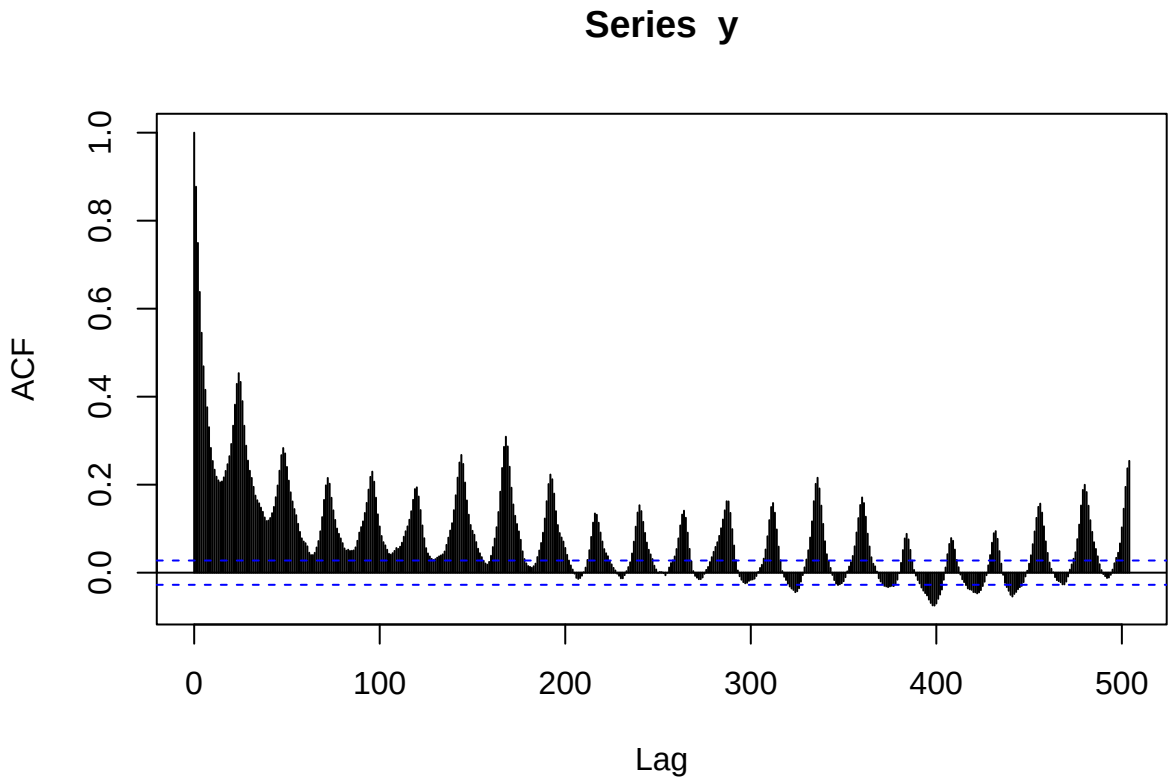
Etude de cas

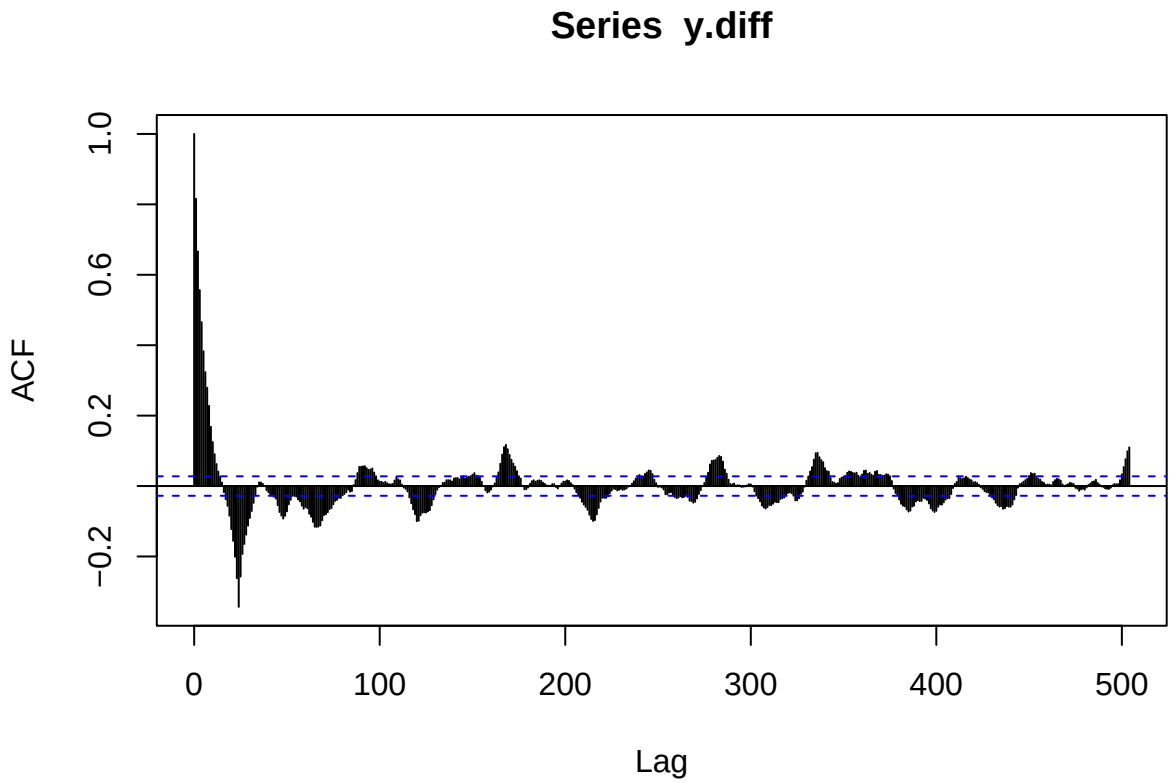
Données de pollution

Prenons l'exemple de données de polution de l'air mesurée à Londre à Marylebone Road (source: http://www.openair-project.org/). Rappelons qu'il s'agit de données au pas horaire. Nous considérons ici, les mesures de volume d'oxyde d'azote (nox: monoxyde d'azote plus dioxyde d'azote, voir: http://www.airparif.asso.fr/pollution/differents-polluants).



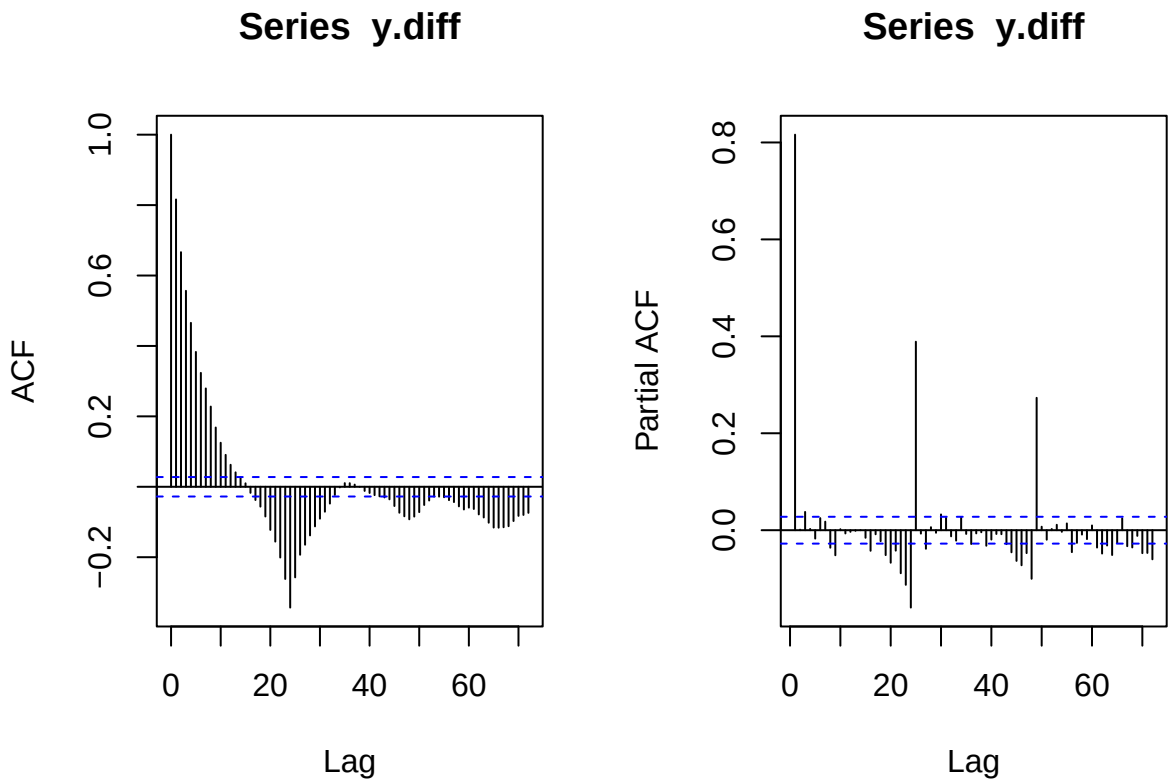
dont l'autocorrélogramme fait clairement apparaitre une saisonnalité de période 24 heures. On remarque que les autocorrélations non multiples de 24 décroissent rapidement vers 0 alors que ce n'est pas le cas pour celles d'ordre $k * 24$. On différencie une fois via l'opérateur $1 - L^{24}$ (fonction `diff` de r) et l'autocorrélogramme de la série différenciée est satisfaisant (décroissance rapide vers 0).





On représente ensuite l'acf et la pacf de la série différenciée de manière à indentifier les ordres p,q,P et Q du modèle SARIMA que l'on veut ajuster aux données. On obtient les bornes max suivantes:

- qmax=12; Qmax=2
- pmax=2; Pmax=1



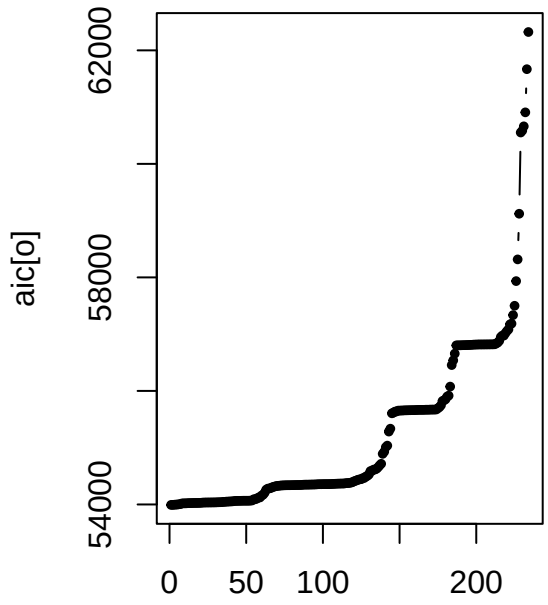
On procède ensuite à l'estimation de tous les modèles candidats:

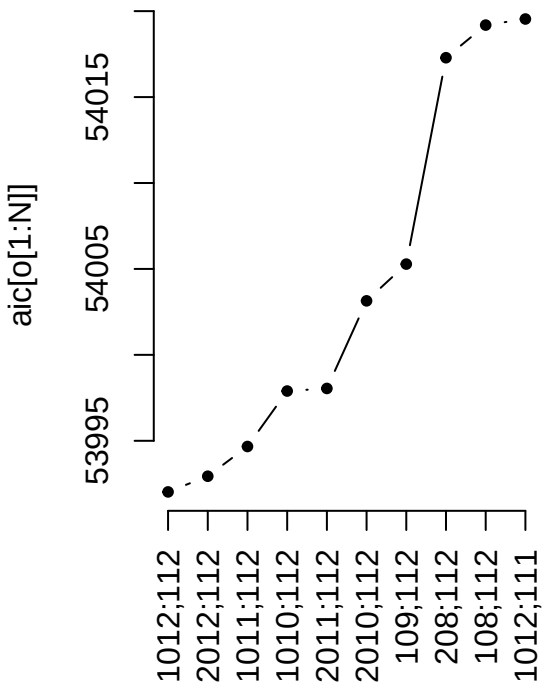
$$\text{SARIMA}[(0 \leq p \leq \text{pmax}, 0, 0 \leq q \leq \text{qmax}); (0 \leq P \leq \text{Pmax}, 1, 0 \leq Q0 \leq \text{Qmax})]$$

Ce qui fait au total 234 modèles. L'objectif final étant la prévision, nous calculons ensuite l'AIC et le représentons en l'ordonnant au préalable par ordre croissant. Notons que pour pouvoir estimer 234 modèles en un temps raisonable il est d'utiliser la méthode de moindre carrés conditionnel (option "CSS" dans la fonction `arima`, voir TP6).

Le meilleur modèle au sens de l'AIC est ici le modèle

$$\text{SARIMA}[(1, 0, 12); (1, 1, 2); 24]$$

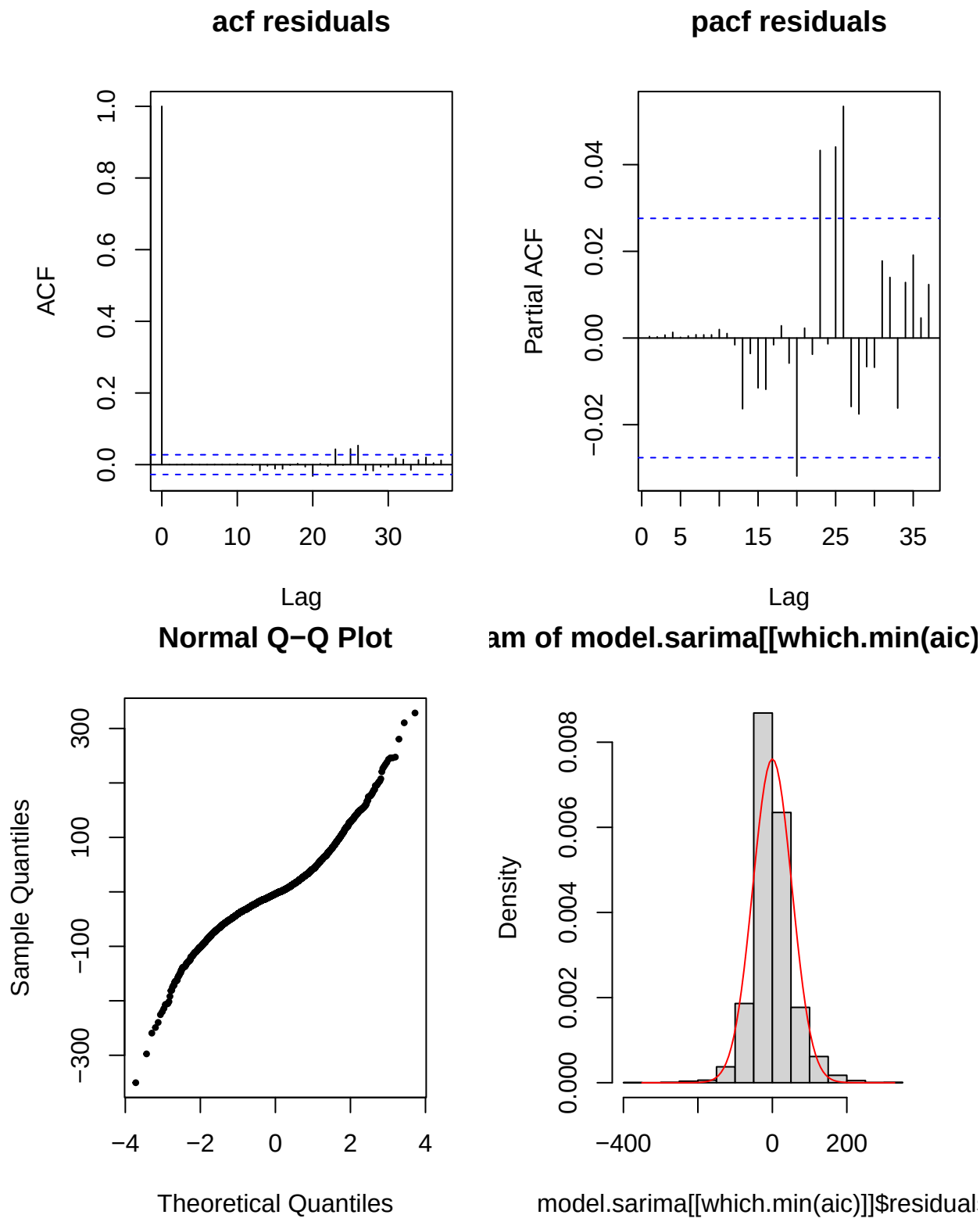




On constate que les tests de student de significativité des paramètres permettent de rejetter au seuil 5% l'hypothèse de nullité des paramètres sauf pour θ_7, θ_8 et θ_{11} ce qui n'est pas une remise en cause du choix de l'ordre car ne concerne pas les coefficients d'ordre maximal.

```
##          ar1          ma1          ma2          ma3          ma4          ma5
## 0.000000e+00 7.021050e-13 0.000000e+00 2.082379e-11 4.596166e-08 8.260351e-09
##          ma6          ma7          ma8          ma9         ma10         ma11
## 1.961620e-04 4.462638e-01 8.854397e-01 1.097078e-07 7.803898e-03 5.334191e-02
##         ma12         sar1         sma1         sma2
## 3.146292e-02 1.743045e-04 0.000000e+00 5.186587e-07
```

```
##   ar1   ma1   ma2   ma3   ma4   ma5   ma6   ma7   ma8   ma9  ma10  ma11  ma12
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE
##  sar1  sma1  sma2
## FALSE FALSE FALSE
```



De plus le test de box-pierce ne rejette pas au seuil 5% et à l'ordre 20 l'hypothèse de non-corrélation des résidus:

```
pvalue_BP(model.sarima[[which.min(aic)]],K=20)
```

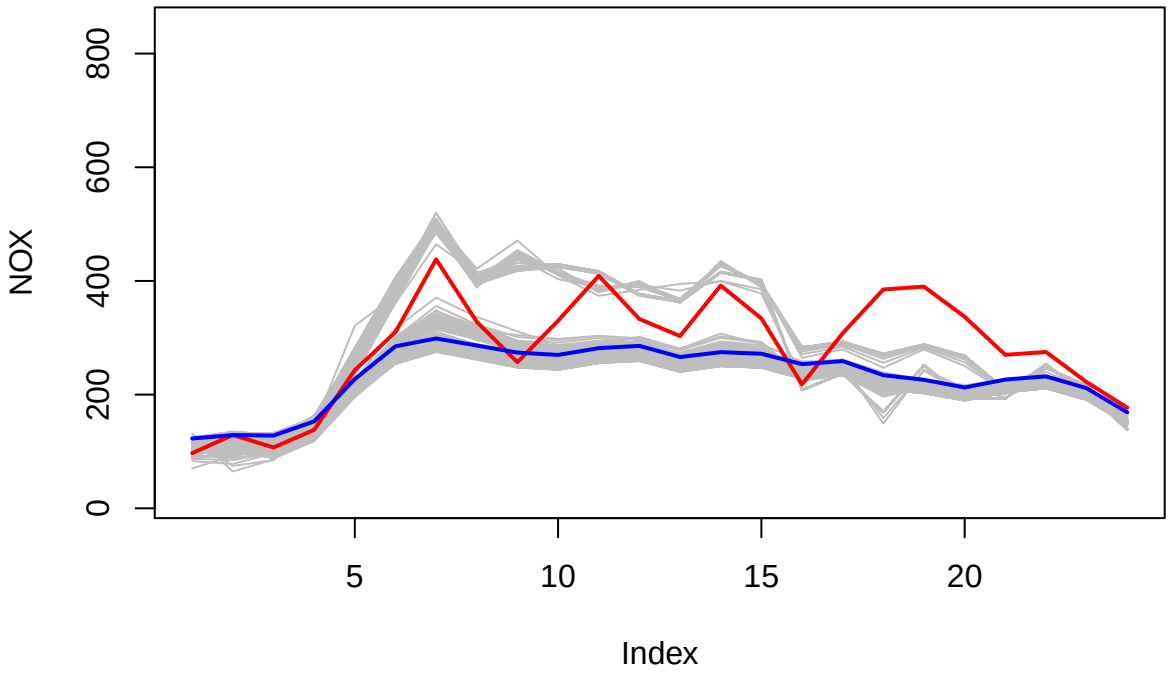
```
## [1] 0.08603576
```

Enfin, nous pouvons effectuer une prévision à 24h d'horizon du jour suivant la période d'estimation (en rouge), pour l'ensemble des modèles candidats (en gris), et le modèle retenu (en bleu):

```
f<-function(x){as.numeric(predict(x,n.ahead=24)$pred)}
forecast<-lapply(model.sarima,f)

erreur<-unlist(lapply(forecast,function(x){mean((ynew-x)^2)}))

par(mfrow=c(1,1))
plot(ynew[1:24],type='l',ylim=range(y,forecast,ynew),ylab='NOX')
for(i in c(1:length(aic)))
{
  lines(forecast[[i]],col='grey')
}
lines(ynew[1:24],col='red',lwd=2)
lines(forecast[[which.min(aic)]],col='blue',lwd=2)
```

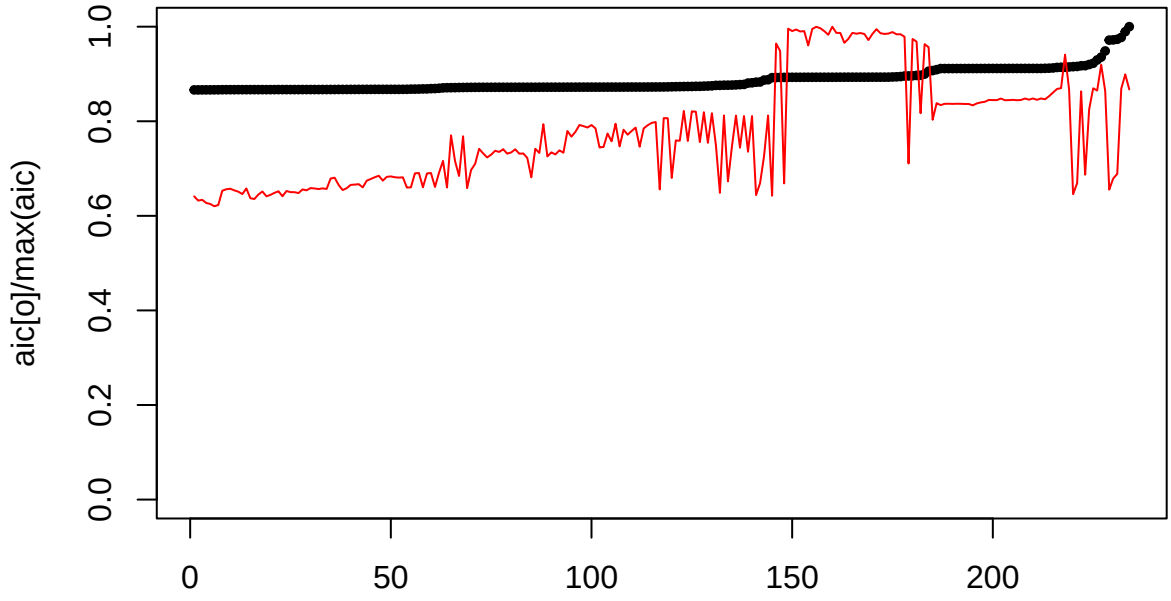


Enfin, comparons l'erreur de prévision (de 1 à 24h tous les jours pendant 3 semaines) avec l'aic:

```
setwd("/Users/yannig/Documents/Enseignement/2017-2018/Serie_temp/cours")
load("prev_nox.RData")

erreur_list<-unlist(lapply(prev_list,function(x){mean((ynew[-c(1:24)]-x[1:(length(x)-24)])^2)}))

o<-order(aic)
plot(aic[o]/max(aic),type='b',pch=20,axes=T,xlab='',cex=0.8,ylim=range(0,1))
lines(erreur_list[o]/max(erreur_list),col='red')
```



On constate que le choix de l'AIC s'est avéré plutôt bon pour la prévision à 24h.

```
par(mfrow=c(1,1))
plot(ynew[-c(1:24)],type='l',ylim=range(ynew),ylab='NOX')
for(i in c(1:length(aic)))
{
  lines(prev_list[[i]][1:(length(ynew)-24)],col='grey')
}
lines(ynew[-c(1:24)],col='red',lwd=1)
lines(prev_list[[which.min(aic)]][1:(length(ynew)-24)],col='blue',lwd=1)
```