Paris School of Business (PSB)

Étude de série temporelle: modèle ARIMA

MSc Data Management

Projet : **R**

par :

Arnaud Bruel YANKO

Sous la direction de :

M. Henri LAUDE

Enseignant

Année académique 2020-2022

♣ Table des matières ♣

0.1 Introduction

Les séries temporelles (ou chronologiques) sont des données associées à des indices temporels de tout ordre de grandeur : seconde, minute, heure, jour, mois, année, etc. En analyse de série temporelle, le temps est une variable explicative (ou dépendante) incontournable. L'émergence de cycles est une particularité des séries temporelles. Ceux-ci peuvent être analysés en vue d'en déterminer la tendance. Les séries temporelles peuvent également être modélisés en vue d'effectuer des prévisions.

Nous allons couvrir les concepts de base en analyse et modélisation de séries temporelles et plus particulièrement le modèle ARIMA qui fera l'objet de notre étude . Mais avant cela, voyons comment les données temporelles sont manipulées en R.

0.2 Séries temporelles

Définition 0.2.1. *Une série chronologique (ou temporelle) est une succession d'observations au cours du temps :* $U_t : t = 1, 2, \cdots, n, \cdots = (U_1, U_2, \cdots, U_n, \cdots)$

Par rapport aux autres types de données statistiques, la particularité des séries chronologiques tient à la présence d'une relation d'antériorité qui ordonne l'ensemble des informations. Les dates d'observations sont souvent équidistantes les unes des autres : on a des séries mensuelles, trimestrielles, etc, dans quel cas on peut les indexer par $t \in \mathbb{N}$

Exemple 0.2.1. — Nombre des moutons par année en Angleterre, entre 1867 et 2003 ;

- Nombre de voyageurs par mois (SNCF) entre 1990 et 2003 ;
- Nombre de voitures vendues par un garage, par trimestre entre 1995 et 1999 ;
- Taux de mortalité, per âge, entre 55 et 104 (c'est le premier exemple d'utilisation de splines, par Whittaker (1923))

Les séries temporelles sont le plus simple exemple d'une thématique plus large : l'estimation et prévision des processus stochastique, i.e. des familles des variables aléatoires U(x). Pour les séries temporelles/chrologiques, on s'intéresse en $x \in \mathbb{N}, \mathbb{Z}$.
On se propose d'estimer la valeur de la variable $U(x)$ en un point x quelconque connaissant les valeurs $U(x_i)$ aux points de mesure données x_i, pour $i = 1, ...N$. Le but principal est le choix d'un modèle *(estimation)* raisonnable, qui permettra à partir des valeurs connues la prédiction des valeurs inobservables (comme les valeurs futures des séries temporelles, ou moins accessibles physiquement, couteuses, etc).
On veut à la fois :

- (a) enlever du bruit d'observation eventuelle ;
- (b) *extrapoler* du connu au inconnu.

Domaines d'application :

- Prospection et exploitation pétrolières et minières
- Traitement du signal
- Imagerie Médicale
- Océanographie, météorologie, hydrogeologie, environnement,$\cdots$
- Séries temporelles, appliquées en économie, finances, météo, médecine

0.3 Premier abord aux séries temporelles/chroniques

Une règle générale en statistique descriptive consiste à commencer par regarder ses données, avant d'effectuer le moindre calcul. Ainsi, la figure suivante montre différentes séries chronologiques, qui méritent quelques commentaires.

— La consommation des ménages en Allemagne et le Produit Intérieur Brut en France semblent avoir augmenté régulièrement ;

— Le taux de chomage en Allemagne semble avoir globalement augmenté depuis 1960, mais avec une alternance de baisses et de hausses soudaines. Le taux de chomage des Etats-Unis ne semble pas évoluer globalement, mais présente également cette alternance de baisses et de hausses ;

— Les ventes de champagnes, tout comme la production industrielle semblent exhiber un caractère périodique (ventes importantes de champagne en fin d'année, baisse de la production industrielle en éte, $\cdots$) ;

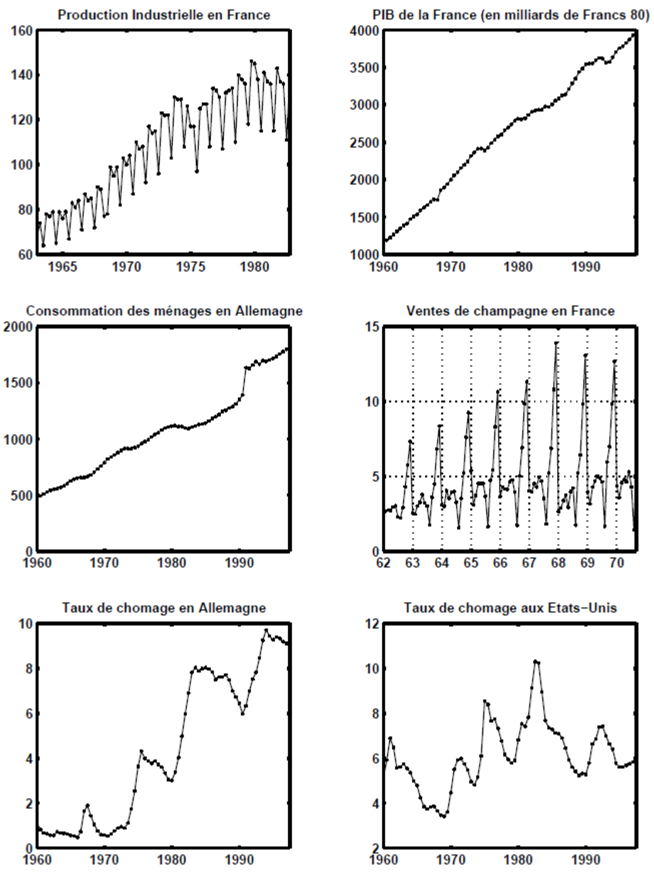
— D'autre part, les variations de ces 2 séries (indice de production industrielle et ventes de champagne) ont une amplitude qui semble augmenter au cours du temps.

— Toutes ces séries ont un aspect irrégulier. Ces fluctuations irrégulières ont parfois une amplitude anormalement élevée (PIB et production industrielle en France au second trimestre 1968, consommation en Allemagne en 1991).

Cette liste de remarques n'est bien sûre pas exhaustive. Elles traduisent simplement quelques comportements que l'on retrouve sur la plupart des séries chronologiques. Puisque notre ambition est de décrire et d'analyser ce genre de chroniques, il nous faut donc proposer des modèles qui intègrent les différentes caractéristiques que nous venons de relever.

0.3.1 Les composantes d'une série temporelle

Dans un premier temps, l'examen graphique de la série étudiée $(y_i, 1 \leq i \leq n)$ permet de dégager, lorsqu'on envisage une période de temps suffisamment longue, un certain nombre de composantes fondamentales de l'évolution de la grandeur étudiée.



Il faut alors analyser ces composantes, en les dissociant les unes des autres, c'est-à-dire en considérant une série comme résultant de la combinaison de différentes composantes, tel que chacune d'elles ait une évolution simple.

1. La tendance $(f_i, 1 \leq i \leq n)$ représente l'évolution à long terme de la grandeur étudiée, et traduit l'aspect général de la série. C'est une fonction monotone, souvent polynomiale.
2. Les variations saisonnières $(s_i, 1 \leq i \leq n)$ sont liées au rythme imposé par les saisons météorologiques (production agricole, consommation de gaz, $\cdots$), ou encore par des activités économiques et sociales (fêtes, vacances, solde, etc). Mathématiquement, ce sont des fonctions périodiques, c ?est-à-dire qu'il existe un entier p, appelé période, tel que :

$$s_i = s_{i+p}, \forall i \geq 1 \tag{1}$$

Cette composante est entièrement déterminée par ses p premières valeurs $s_1, s_2, \cdots, s_p$. On rencontre souvent aussi des phénomènes pour les quelles la période peut elle même varier. On parle alors de

3. *Cycles* $(c_i, 1 \leq i \leq n)$, qui regroupent des variations à période moins précise autour de la tendance, par exemple les phases économiques d'expansion et de récession. Ces phases

durent généralement plusieurs années, mais n'ont pas de durée fixe. Sans informations spécifiques, il est généralement très difficile de dissocier la tendance du cycle. Dans le cadre de ce cours, la composante appelée tendance regroupera pour la plupart du temps aussi les cycles.

4. Les fluctuations *irrégulières/résidues/bruit* $(e_i, 1 \leq i \leq n)$ sont des variations de faible intensité et de courte durée, et de nature aléatoire (ce qui signifie ici, dans un cadre purement descriptif, qu ?elles ne sont pas complètement expliquables). En effet, elles ne sont pas clairement apercevables dans les graphiques, à cause de leur faible intensité par rapport aux autres composantes. Elles apparaissent clairement seulement après "l'enlèvement du signal" ; la question qui se posera alors sera : est-ce qu'ils contiennent encore du signal, ou est-ce que c'est vraiment du "bruit" ?
5. *Les variations accidentelles/observations aberrantes* sont des valeurs isolées anormalement élevées ou faibles de courte durée. Ces variations brusques de la série sont généralement explicables (Mai 68, réunification de l'Allemagne, tempête, $\cdots$). La plupart du temps, ces accidents sont intégrés dans la série des bruits (les fluctuations irrégulières).
6. Points de changement Ce sont des points où la série change complètement d'allure, par exemple de tendance. Ils sont normalement explicables, et imposent une analyse séparée de la série, par morceaux.

0.4 Modélisation stochastique des séries temporelles

Rappels 0.4.1. *On distingue 3 catégories de séries temporelles :*

— *Le modèle additif*

— *Le modèle multiplicatif*

— *modèle mixte*

Rappelons le modèle additif sans saisonnalité, qui cherche une décomposition de la forme :

$$Y_t = m_t + \epsilon_t \tag{2}$$

où :

— m_t représente la "tendance" (intuitivement un "mouvement lisse à long terme"), qui sera la composante la plus importante dans la prévision.

— $\epsilon_t = Y_t - m_t$ sont les "résidus" qui restent après qu'on enlève la partie structurée m_t. Elles représentent des "irrégularités/fluctuations imprévisibles", qui au début semblent inutilisables (à ignorer) pour la prévision (c'est correct du point de vue de la prévision ponctuelle, mais elles nous servirons quand-même dans la calcul des intervalles de confiance).

On s'arrangera toujours tel que les résidus ont la moyenne 0, mais ça n'est pas suffisant pour qu'ils soient un bruit totalement sans structure="bruit blanc" (et s'il y a encore une partie structurée, elle devrait être inclue en m_t).

Le "bruit blan" est notre premier exemple d'un processus stochastique : une formalisation du concept de séries temporelles, ayant des propriétés bien définies (voir prochaine chapitre). Inspirés par les propriétés de ce processus, on proposera des tests statistiques correspondant à ce modèle, qui nous permettrons de décider si ϵ_t ont les propriétés de manque de structure desirées. Pour tendance, plusieurs modèles se sont averés utiles :

1. regression sur des predicteurs exogènes "covariates"), implementé en logiciels R par "formules"

$$m_t \equiv X_t^{(1)} + X_t^{(2)} +, \cdots \tag{3}$$

2. modèles de superposition des chocs exterieurs/moyennes mobiles ϵ_t :

$$m_t = \sum_{i=1}^{q} \delta\epsilon_{t-i} \tag{4}$$

3. modèles autoregressifs :

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots) + \epsilon_t \tag{5}$$

Dans le manque des predicteurs exogènes, il est assez naturel d'adopter une modélisation autoregressive pour la tendance. Sous certaines conditions de régularité, ça ramènera à des prévisions autoregressives un pas en avant :

$$\tilde{Y_t} = f(Y_{t-1}, Y_{t-2}, \cdots) \tag{6}$$

Le modèle le plus simple est le processus $AR(1)$:

$$Y_t = \phi Y_{t-1} + b + \epsilon_t \tag{7}$$

Ce modèle est recommandable si on envisage une prévision

$$\tilde{Y_t} = \psi Y_{t-1} + b \longleftrightarrow (\tilde{Y_t} - a) = \psi(\tilde{Y_{t-1}} - a) \tag{8}$$

où $b = a(1 - \psi)$.
On vérifie que si la moyenne de Y_t est 0 on a $a = b = 0$; pour simplifier, on supposera normalement qu'on a déjà enlevé la moyenne de Y_t.
Pour utiliser ce modèle, on estimer le paramètre ϕ par une régression linéaire des points

$$(Y_{t-1}, Y_{t-1}), t = 2, \cdots, T \tag{9}$$

Le fait d'avoir enlevé la moyenne ramène à une droite passant par l'origine $y = \phi x$. En suite, on utilise la valeur trouvée pour résoudre l'équation. On trouve

$$Y_t = \sum_{i=0}^{t-1} \psi^i \epsilon_{t-i} + \psi^t Y_0 \tag{10}$$

Définition 0.4.1. *Processus stochastiques stationnaires*
Soit X un processus aléatoire indexé par $T = \mathbb{N}$ ou $\mathbb{Z}$. On dit que X est stationnaire (strict) si pour toute famille finie d'instants $t1...t_r \in T$ et tout entier s, les lois jointes de $(X_{t_1}, \cdots, X_{t_r})$ et de $(X_{t_{1+s}}, \cdots, X_{t_{r+s}})$ sont les mêmes.

Définition 0.4.2. *Soit X un processus aléatoire indexé par $T = \mathbb{N}$ ou $\mathbb{Z}$. On dit que X est stationnaire à l'ordre 2 si la moyenne $m(t)$ et la covariance $\tau(s, t)$ sont invariantes par translation dans le temps, i.e. si la moyenne est constante :*

$$\mathbb{E}X_t = m_t = m, \forall t \tag{11}$$

et si la covariance/corrélation dépend seulement de l'écart de temps $k = t - s$, i.e. il existe une fonction d'une variable $\gamma(k)$, paire, telle que :

$$45Cov(X_t, X_s) = C(t, s) = \gamma(t - s) = \gamma(k)\forall k = -2, -1, 0, 1, 2, 3, \cdots \tag{12}$$

Comme la plupart de séries n'est observable qu'une seule fois, l'utilité du concept de distributions et covariances théoriques n'est pas évidente pour les applications. Par contre, on peut toujours calculer des distributions et covariances empiriques, et sous l'hypothèse de stationnarité, les moyennes empiriques convergent vers les théoriques.

0.5 Les modèles $ARIMA(p, d, q)$

Avant de parler du modèle $ARIMA$, il est impératif de parler du modèle $ARMA$ modèle dont il est le dérivé. Raison pour laquelle on donnera juste la définition du dit modèle.

Définition 0.5.1. *Les modèles $ARMA(p, q)$*
On appelle processus $ARMA(p, q)$ un processus stationnaire Y_t, $t \in \mathbb{Z}$ vérifiant une relation de récurrence :

$$Y_t = \sum_{i=1}^{p} \psi^i Y_{t-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}, \forall t \in \mathbb{Z} \tag{13}$$

où les ψ, θ sont des réels et ϵ_t est un bruit blanc de variance σ^2.
La notation des polynômes de retard ramène (13) à la forme :

$$\psi(B)Y_t = \theta(B)\epsilon_t \tag{14}$$

Définition 0.5.2. *Les modèles $ARIMA(p, d, q)$*
On appelle processus $ARIMA(p, d, q)$ un processus non stationnaire X_t pour le quel le processus différencié d'ordre d, $Y_t = (1 - B)^d X_t, t \in \mathbb{Z}$ stationnaire, et vérifie une relation de

récurrence $ARMA(p, q)$ *:*

$$Y_t = \sum_{i=0}^{t-1} \psi^i \epsilon_{t-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}, \forall t \in \mathbb{Z} \tag{15}$$

où les ψ, θ *sont des réels et* ϵ_t *est un bruit blanc de variance* σ^2*.*
La notation des polynômes de retard ramène (14) à la forme :

$$\psi(B)(1-B)^d X_t = \psi(B) Y_t = \theta(B)\epsilon_t \tag{16}$$

$\psi(B), \theta_B$ *sont des polynômes relativement primes dans l'opérateur de retard* B *à ordres* p, q *avec coefficient libre 1, et avec racines dehors le cercle unitaire.*

Formellement, il s'agit des processus $ARMA$ ayant aussi la racine 1, et nous verrons qu'en effet, la prévision des processus $ARIMA(p, d, q)$ est donné par les mêmes formules que celle des processus stationnaires $ARMA(p, q)$.

0.5.1 Prévision linéaire des modèles autorégressifs $ARIMA(p, d, 0)$

Les deux méthodes principales pour l'estimation des paramètres sont la méthode des moments et la maximisation de la vraisemblance. La première méthode s'appui sur les formules théoriques des moments, en l'occurrence les corrélations.

Exemple 0.5.1. La prévision linéaire $X_t(k)$ pour le processus $ARIMA(0, 1, 0)$ à moyenne c satisfait la reccursion $YuleWalker$
$X_t(k) = X_t(k-1)$

et est donc constante
$X_t(k) = X_t$

Proposition 0.5.1. *La fonction de prévision "eventuelle" de Box-Jenkins pour les processus* $ARIMA(p, d, q)$ *est un élément de l'espace linéaire des solutions de la reccursion* $\psi(B)X_t(k), pour k > q$.

Par exemple, pour les processus $ARIMA(0, d, q)$ la fonction de prévision "eventuelle" est un polynome d'ordre $d?1$.

0.5.2 Prévision des processus $ARIMA(p,d,q)$

En conclusion, pour la pévision linéaire $\hat{X_t(k)}$ des processus $ARIMA(p,d,q)$, on aura toujours besoin d'une estimation de $\epsilon_{t-1}, \epsilon_{t-2}, \cdots$ ou au moins de $\epsilon_{-1}, \epsilon_{-2}, \cdots$ i.e. du "bruit inobservable passé" du modèle. On peut aussi recourir à la répresentation $AR(\infty)$, dans quel cas on aura besoin de $X_{-1}, X_{-2}, \cdots$, qui sont aussi inobservables. En plus, le resultat final demandra une approximation des valeurs precedant le debut d'observations 0 ; l'approximation la plus simple dans l'absence des moyennes est $\epsilon_k = Y_k$ pour $k < 0$.

Théorème 0.5.1. *Dans le cas d'un modèle $ARIMA(p,d,q)$, la meilleure prévision linéaire au temps t est :*

$$\hat{X_t(k)} = \mathbb{E}[X_{t+k}/F_t] = \sum_{i=1}^{p} \hat{\psi}_i X_t(\hat{k-i}) + \sum_{i=k}^{q} \theta_i \epsilon_{t+\hat{k}-i} \tag{17}$$

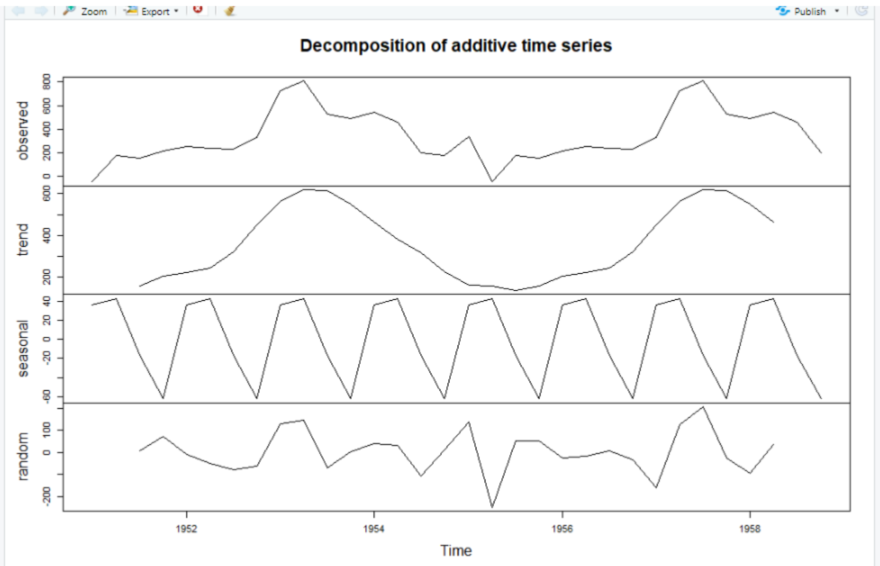
où les $\hat{\psi}_i$ sont les coefficients du polynôme $\psi(B)(1-B)^d$

0.6 Application sur R

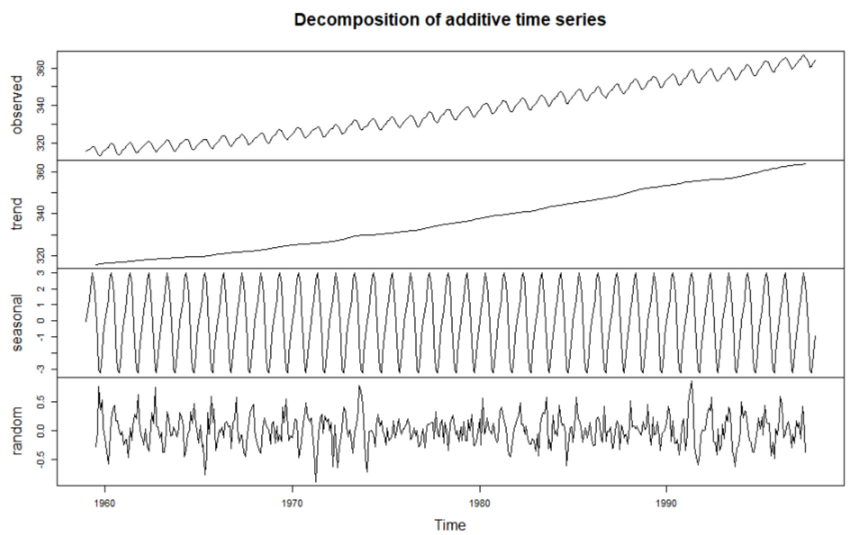
0.6.1 Modélisation d'une série temporelle

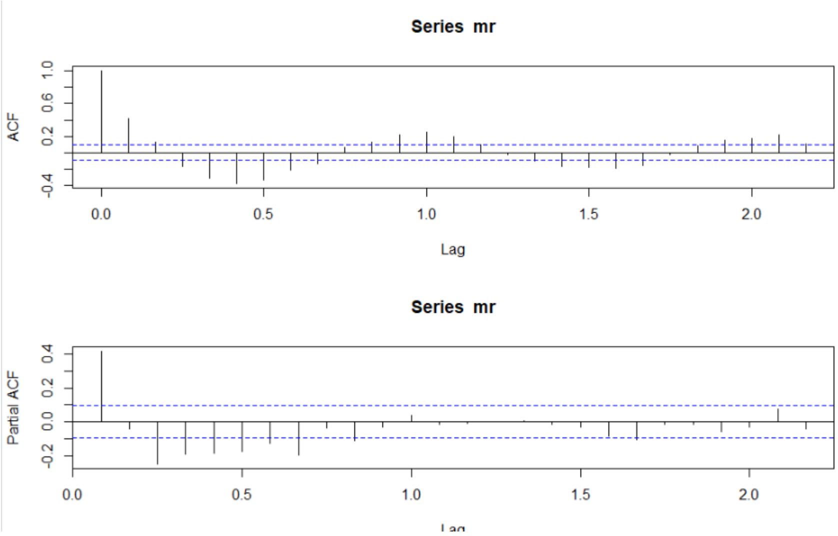
Ici on modélise une série temporelle dit "classique" qui est un exemple de série temporelle monter par moi même.

```
> x <- c(-50, 175, 149, 214, 247, 237, 225, 329, 729, 809, 530, 489, 540, 457, 195, 176, 337)
> x <- ts(x, start = c(1951, 1), end = c(1958, 4), frequency = 4)
> m <- decompose(x)
> plot(m)
>
```




```
require(graphics)
m <- decompose(co2)
names(m)
plot(m)
mr<-window(m$random,start=c(1960,1),end=c(1996,4))
summary(mr)
layout(1:2)
acf(mr)
pacf(mr)
mrs<-as.ts(mr)
mr.ar <- ar(mrs, method = "mle")
mr.ar
```






```
Coefficients:
      1        2        3        4        5        6        7        8        9       10
 0.2377  -0.0509  -0.2358  -0.1941  -0.1961  -0.1688  -0.1043  -0.1962  -0.0110  -0.1151

Order selected 10  sigma^2 estimated as  0.04462
```

0.6.2 Modélisation : Modèle $ARIMA(1,1,2)$

Considérons le modèle $ARMA(1,2)$ suivant :

$$y_t = -0.9 - 0.8y_{t-1} + z_t - 0.3z_{t-1} + 0.6z_{t-2}, t = 1, \cdots, 200 \tag{18}$$

Avec z_t qui suit $\mathbf{N}(0,4)$. On peut reécrire cette équation sous la forme :

$$y_t = -0.5 + \frac{1 - 0.3B + 0.6B^2}{1 + 0.8B} z_t, t = 1, \cdots, 200 \tag{19}$$

Le modèle donnée par l'équation (19) est un $ARIMA(1,1,2)$

Question : Comment simuler cette série et l'analyser ?

1. Simulation de $ARMA(1,2)$ avec la fonction *arima.sim*

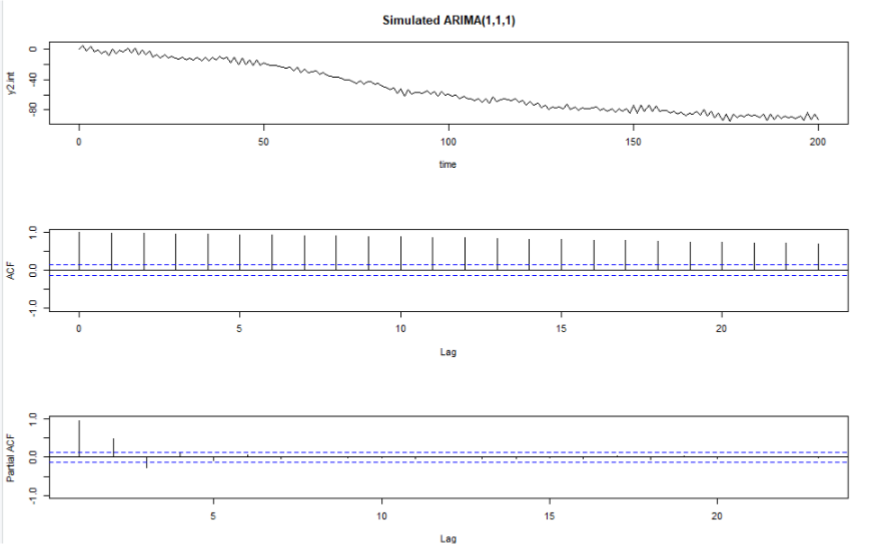
```
set.seed(121181)
yd.n = -0.5 + arima.sim(n = 200, list(ar = -0.8, ma = c(-0.3,
                                                  + 0.6)), sd = 2, n.start = 50)
```

2. Appliquer la fonction *diffinv* sur le $ARMA(1,2)$:

```
y2.int <- diffinv(yd.n)
```

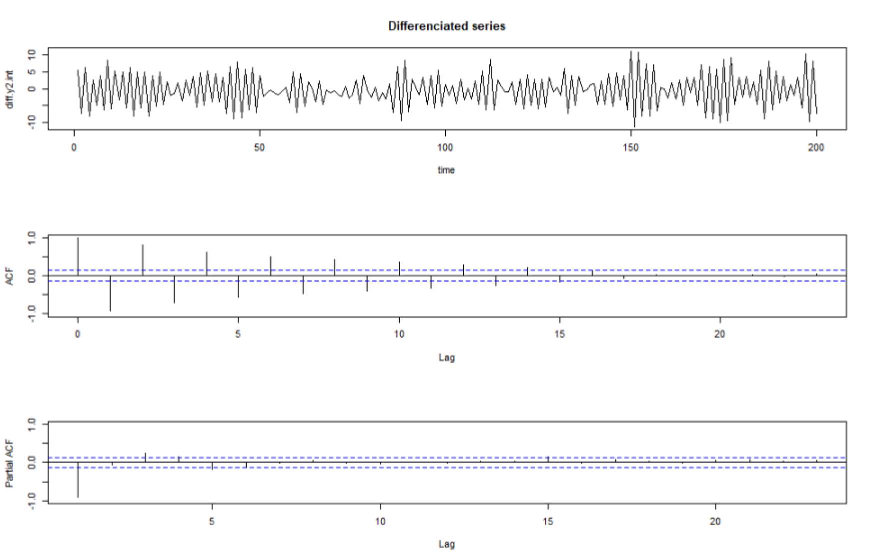
$(-)$ L'analyse de l'ACF (la décroissante de manière non exponentielle) confirme la non stationnarité de la série.

```
op <- par(mfrow = c(3, 1))
plot(y2.int, type = "l", xlab = "time", main = "Simulated ARIMA(1,1,1)")
acf(y2.int, main = "", ylim = c(-1, 1))
pacf(y2.int, main = "", ylim = c(-1, 1))
par(op)
```



$(-)$ ACF et $PACF$ de la série différenciée

```r
diff.y2.int <- diff(y2.int)
op <- par(mfrow = c(3, 1))
plot(diff.y2.int, type = "l", xlab = "time", main = "Differenciated series")
acf(diff.y2.int, main = "", ylim = c(-1, 1))
pacf(diff.y2.int, main = "", ylim = c(-1, 1))
par(op)
```



3. Modélisation $ARIMA(1, 1, 2)$

   $(-)$ utilisation la fonction $Arima$ incluse dans le package $Forecast$ pour modéliser le $ARIMA(1, 1, 2)$ de la serie initiale.

```r
require(forecast)
```

```r
install.packages("forecast")
require(forecast)
y2.fit = Arima(y2.int, order = c(1, 1, 2), include.drift = TRUE)
y2.fit
```

```
Series: y2.int
ARIMA(1,1,2) with drift

Coefficients:
          ar1      ma1     ma2    drift
      -0.8175  -0.2345  0.4635  -0.4608
s.e.   0.0475   0.0719  0.0743   0.0891

sigma^2 estimated as 3.544:  log likelihood=-409.5
AIC=828.99   AICc=829.3   BIC=845.48
>
```

♣ Bibliographie ♣

[1] Site le CRAN *https ://cran.r-project.org/web/packages/available packages by name.html available-packages-E*

[2] *https ://cran.r-project.org/web/packages/*

[3] *Modèles ARIMA et SARIMA, prédiction et choix - Ceremadewww.ceremade.dauphine.fr*

[4] *Séries temporelles : régression, et modélisation ARIMA(p,d,q)web.univ-pau.fr*