Majeure Data Science 2025-2026

(Analyse de) Séries Temporelles

(☺ « comment prévoir le futur en apprenant du passé ? »)

Contenu

1. Introduction
2. Approche probabiliste sur un exemple – série de chômage
3. Analyse exploratoire
4. Processus stationnaires
5. Modèles ARMA et extensions
6. Prévision des modèles ARMA
7. Analyse globale de séries temporelles

Objectifs du cours :

- Apporter les compétences de base minimales pour aborder l'étude ou analyse de séries chronologiques ainsi qu'un minimum de savoir faire…

- Faire comprendre les concepts fondamentaux de cette analyse, en particulier les concepts de stationnarité et d'auto-corrélation…

- Présenter le cadre probabiliste classique des modèles ARMA et leurs extensions (briques de base : AR, MA, ARMA, ARIMA, SARIMA)

Série temporelle = $x_1, \ldots, x_j, \ldots, x_n$ série de données (variable réelle) où le numéro d'observation $j \in \{1, 2, \ldots, n\}$ joue le rôle du « temps » t

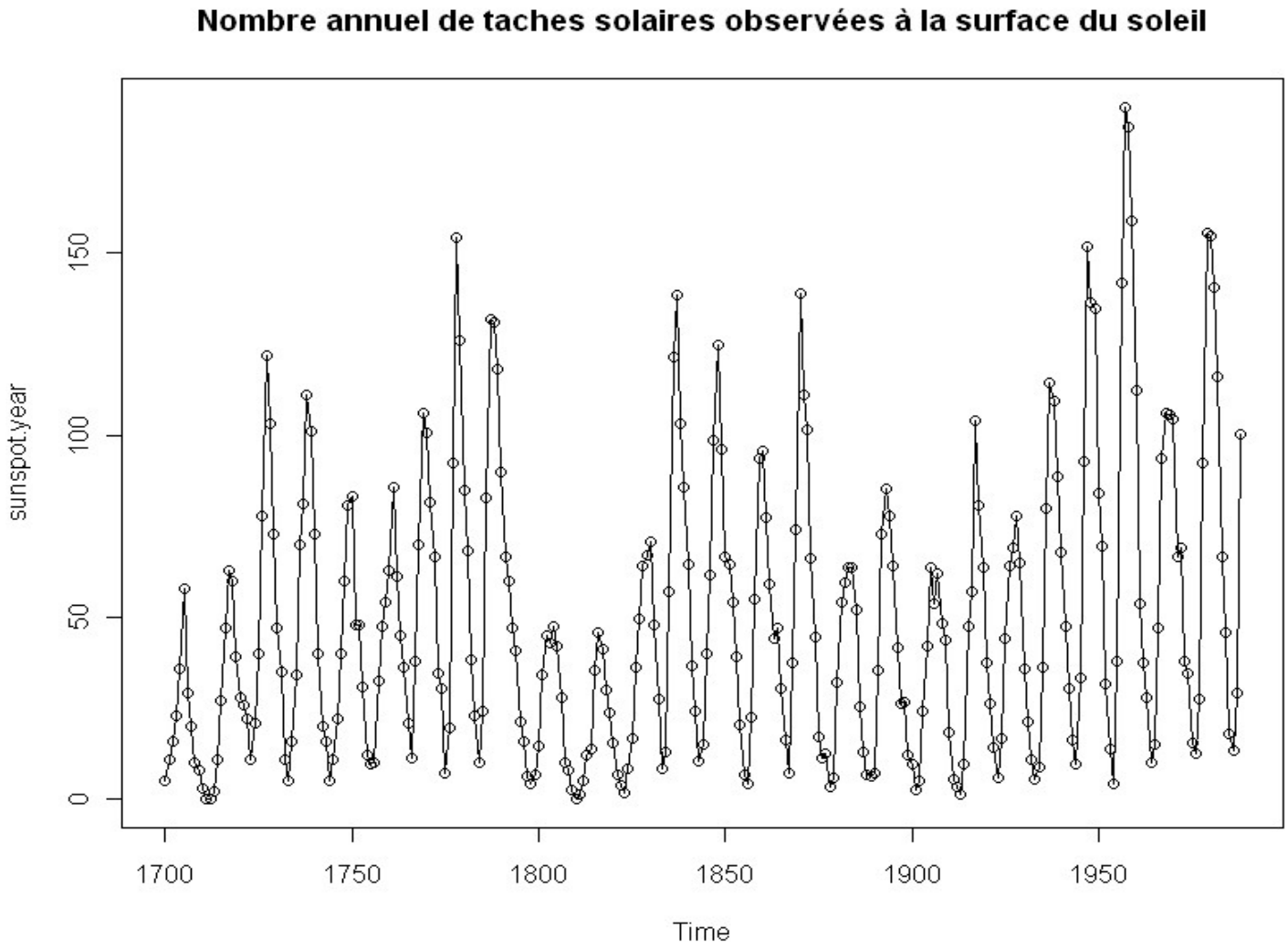
Chronologie dans le recueil des données :

série temporelle = série chronologique

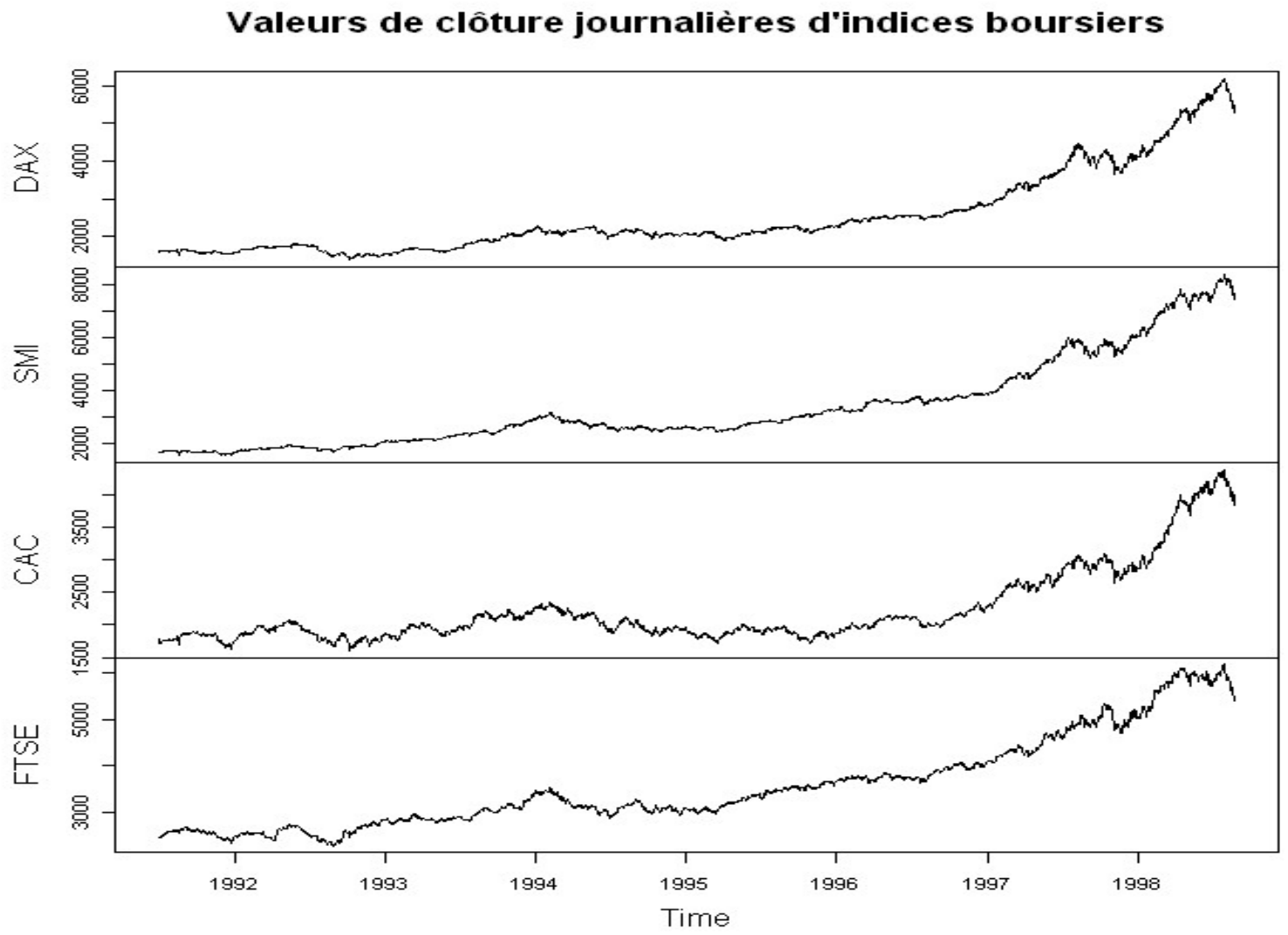
La nature de cette série est « complexe » :

- pas de relation ou mécanisme simple (déterministe) qui lie les observations dans le temps
- variabilité plus ou moins importante
- présence de composantes en apparence « imprévisibles »…

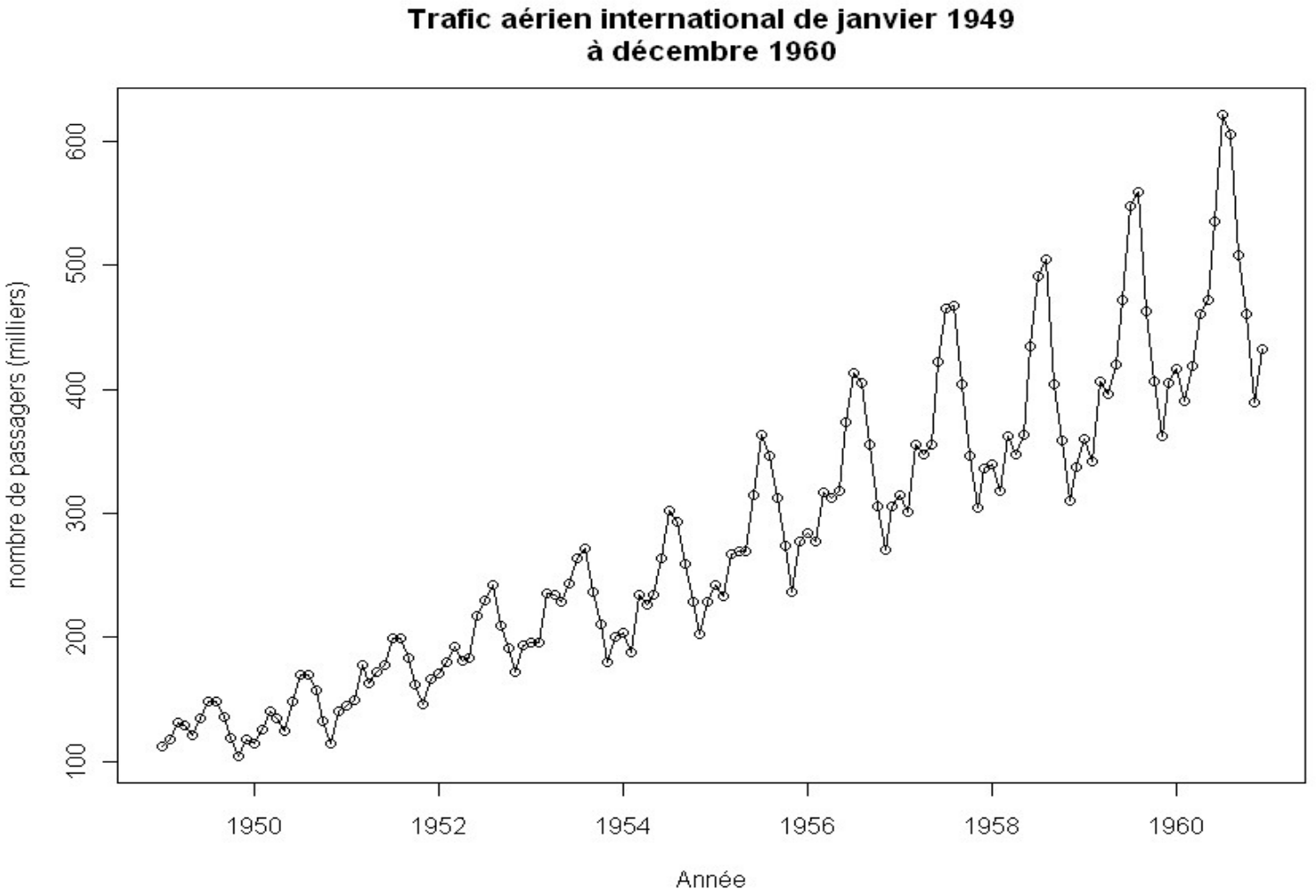
On donne maintenant des exemples :



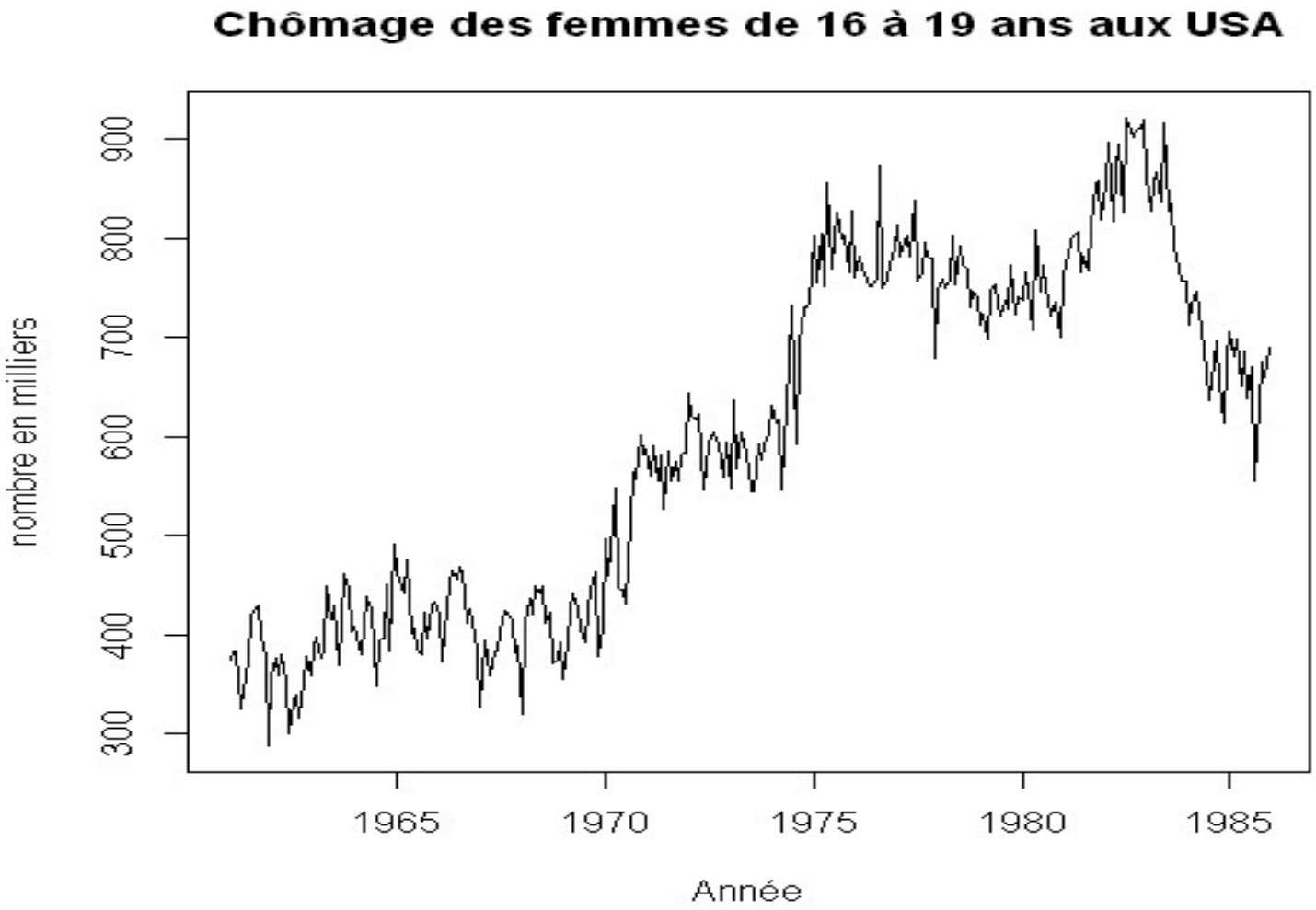
Exemple 1 (source : package R « datasets »)



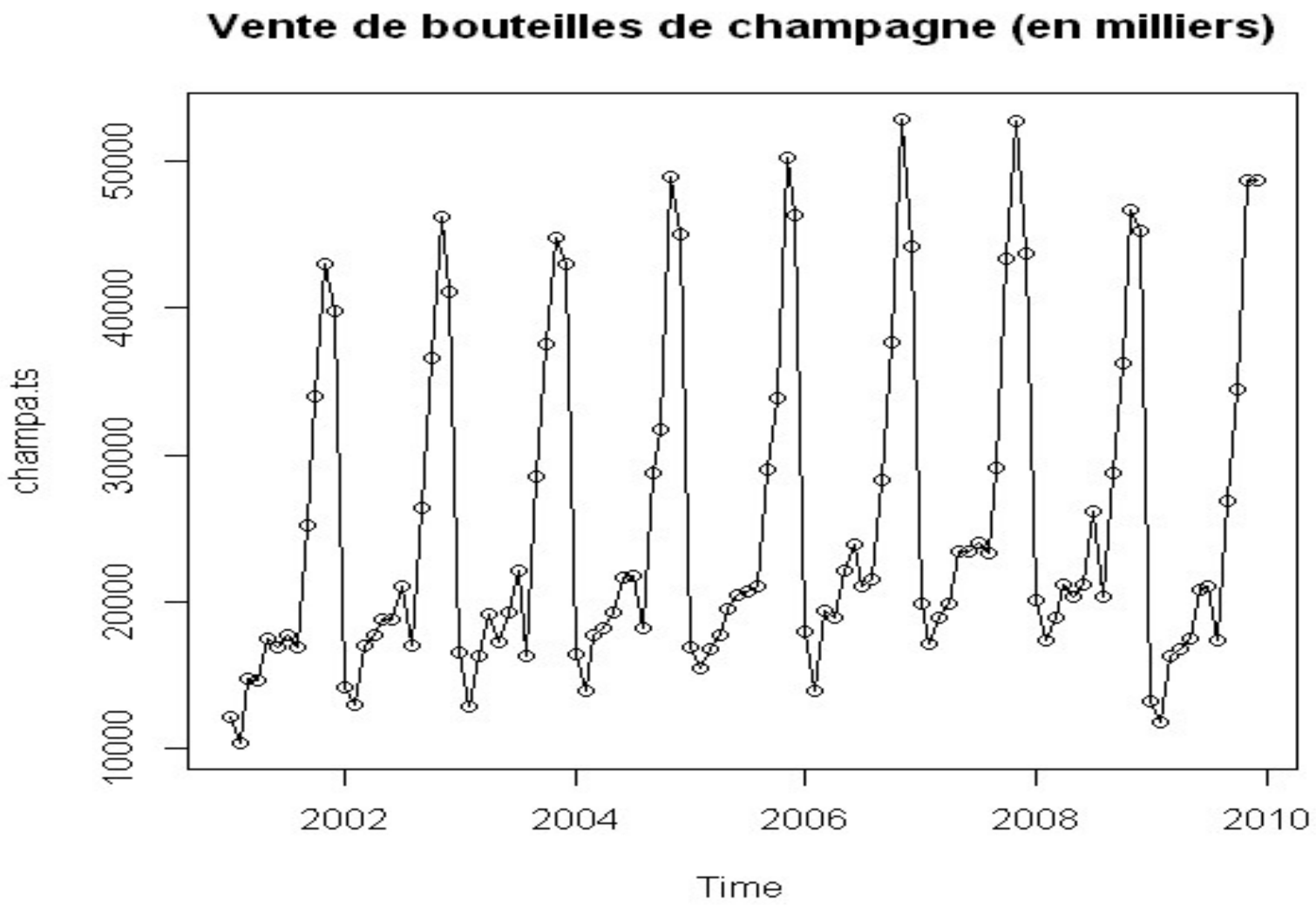
Exemple 2 (source : package R « datasets »)



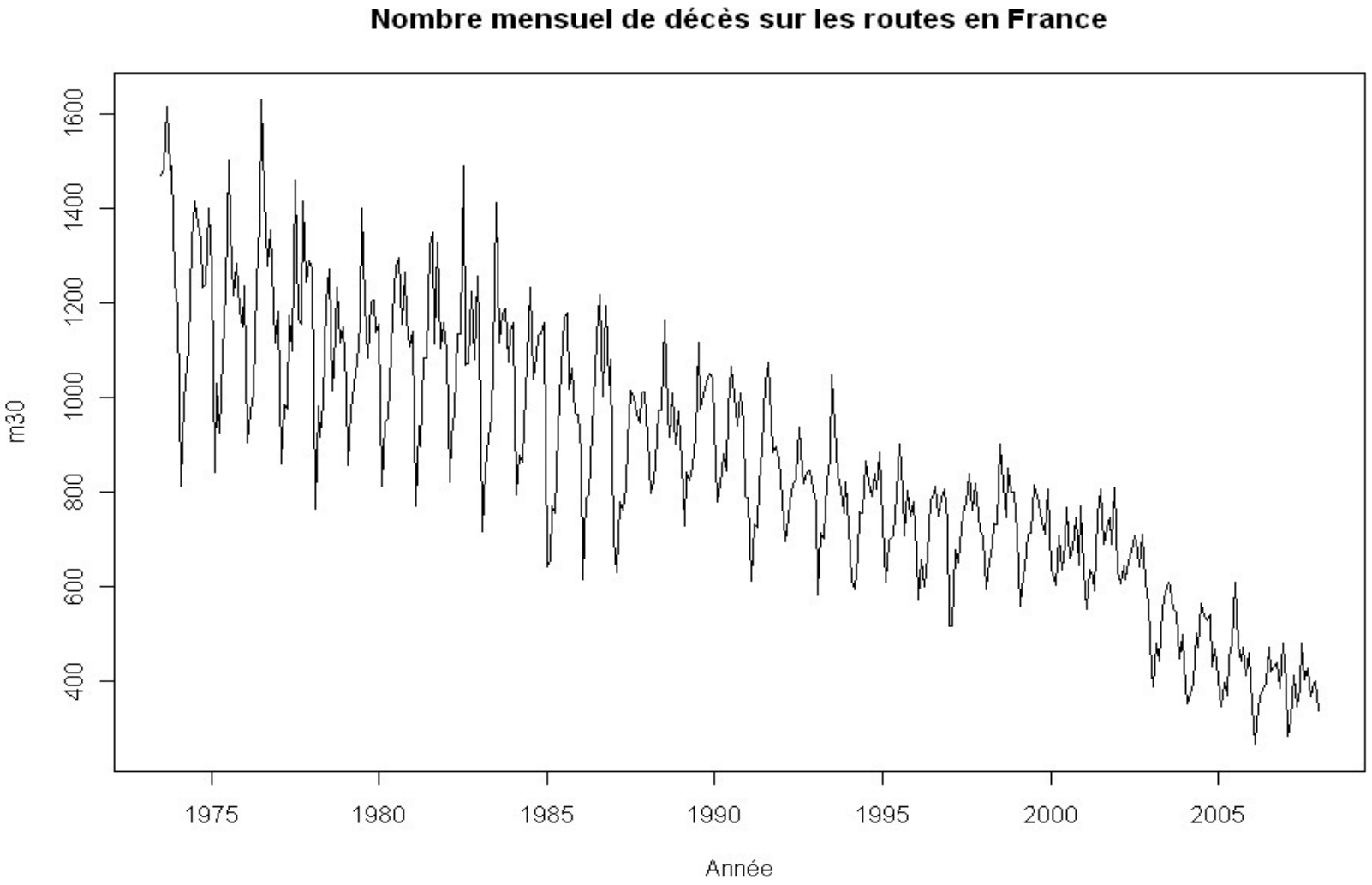
Exemple 3 = exemple 1 page 5 du support (Olivier Roustant, nov. 2008)
(source : package R « datasets », série AirPassengers)



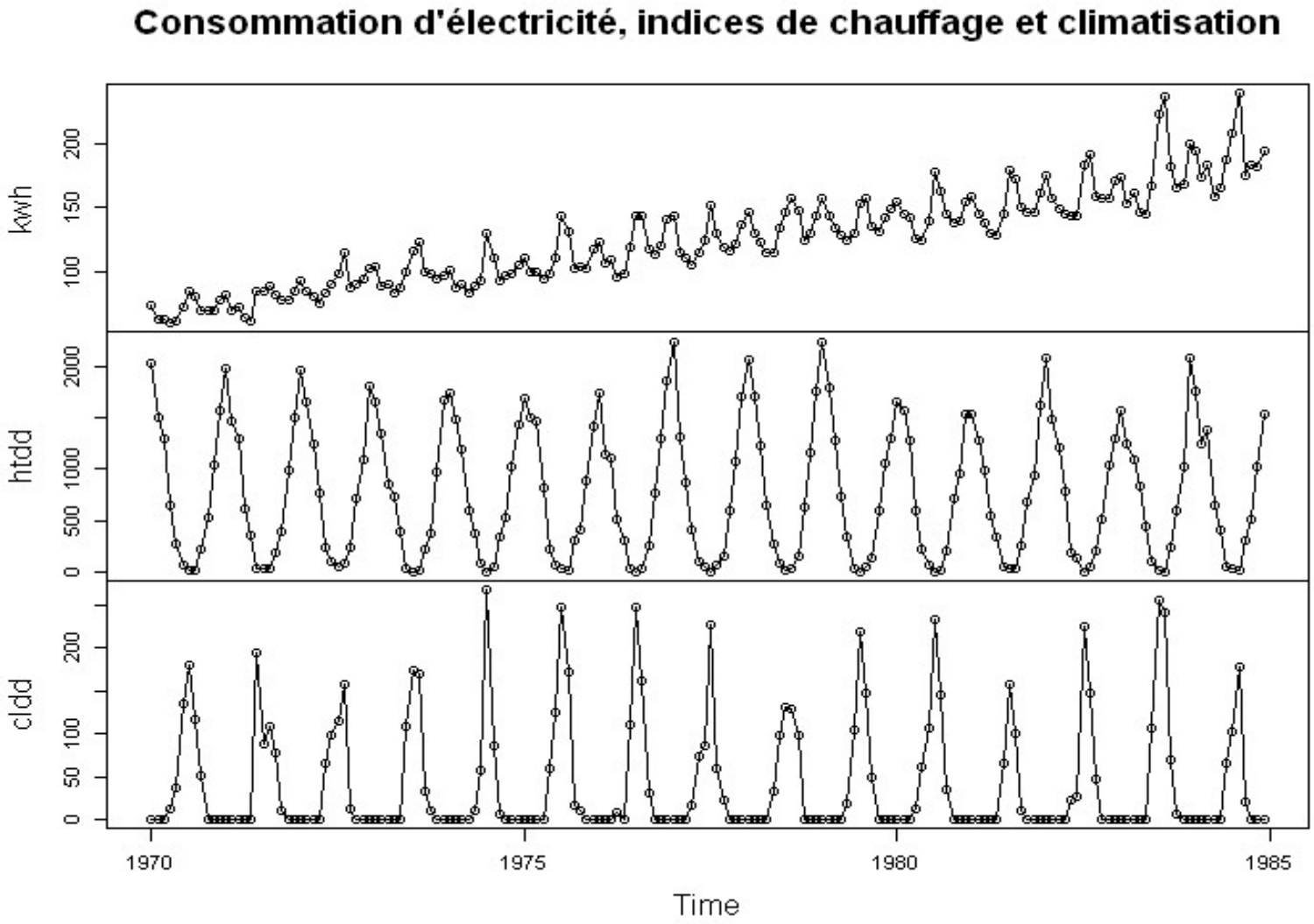
Exemple 4 = exemple 2 du support page 15



Exemple 5 (Source : pkg caschrono de R)



Exemple 6 (Source : pkg caschrono)



Exemple 7 (Source : pkg caschrono)

Objectif central de l'analyse statistique d'une série chronologique = faire de **l'inférence**, par exemple prévoir la suite de la séquence

(objectif propre à toute la Statistique)

☞ Nécessite donc un modèle probabiliste pour représenter les données.

Dans le cas d'un modèle paramétrique, il faudra estimer les paramètres et vérifier la validité du modèle (« goodness of fit »).

On pourra en retour utiliser le modèle estimé pour apporter une réponse à différentes problématiques :

- Décrire/analyser les données : tendance et/ou saisonnalité et fluctuations autour de cette composante

- Filtrage et prévision (objectif central dans ce cours)

- Tests d'hypothèses (par ex., dérive dans un « process » de fabrication, hausse ou non du chômage sur le dernier trimestre, réchauffement climatique dû à l'activité humaine, …)

- Simulations (gestion industrielle, gestion de flux dans l'entreprise, gestion de risque comme en finance ou assurance, …)

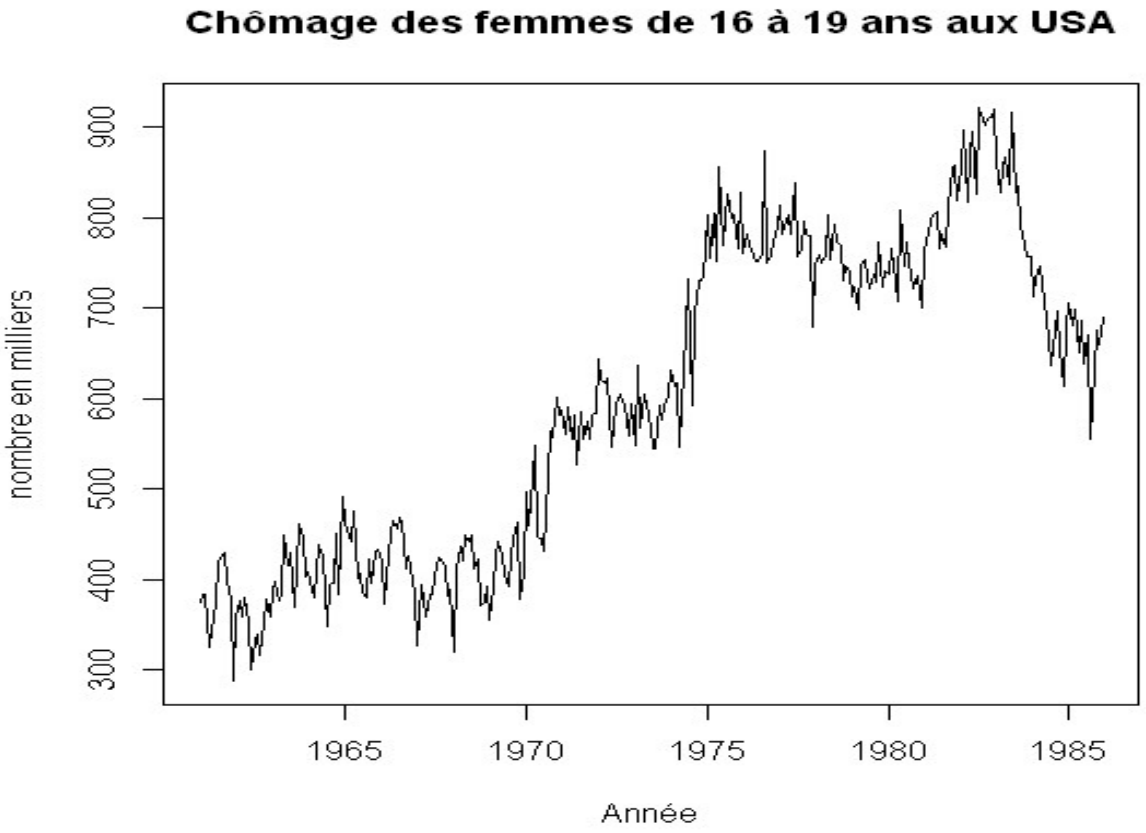
L'idée principale pour apporter une réponse aux problématiques précédentes est de se ramener à une série « stationnaire » avec deux approches classiques :

- On cherche à décrire la série par sa tendance et/ou saisonnalité, ce qui permet d'isoler une composante résiduelle (comme en régression)

- Approche de Box-Jenkins qui consiste à transformer la série initiale à l'aide notamment des opérateurs de différenciation $\nabla = I - B$, ∇^d mais aussi de différenciation saisonnière $\nabla_s = I - B^s$ où B désigne l'opérateur retard : $\{x_t\}_t \rightarrow \{x_{t-1}\}_t$

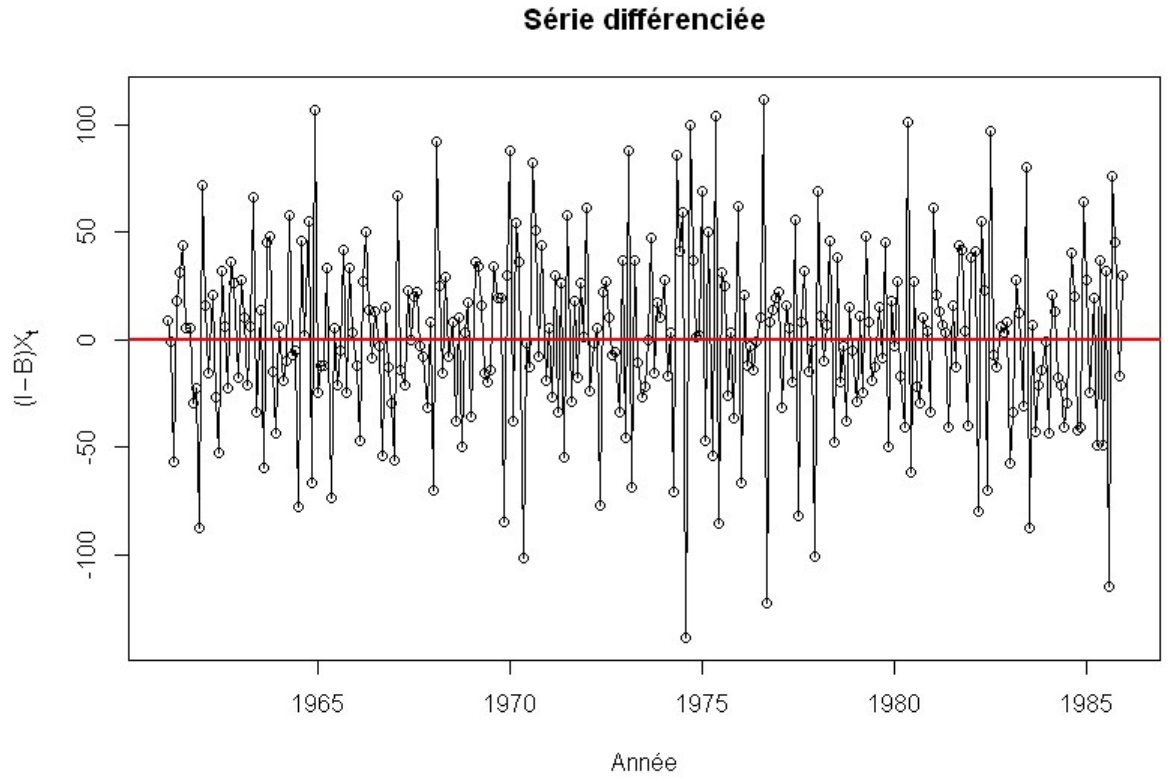
On rend compte ensuite de la partie stationnaire par un modèle probabiliste adéquat…

Exemple 2 du support de cours (page 14) :



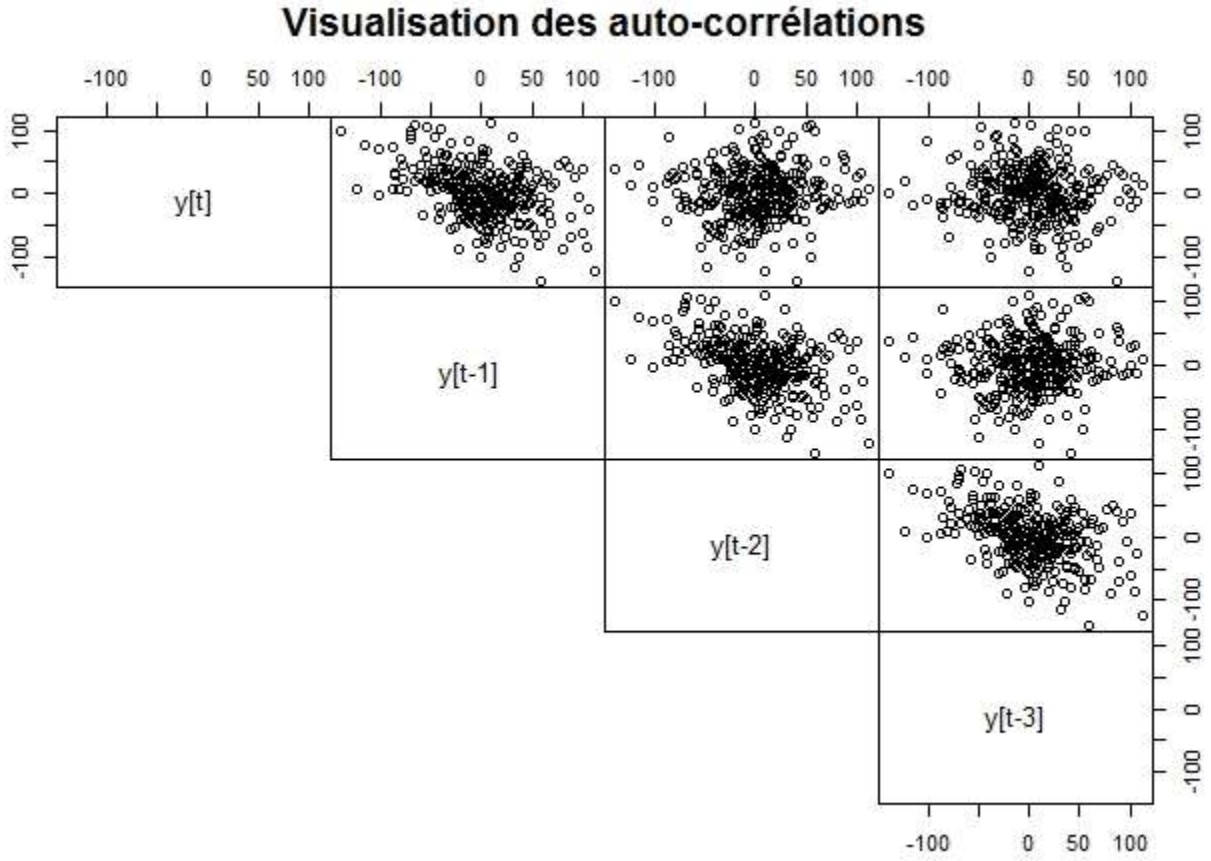
Chronogramme d'une série mensuelle X_t
pour t variant de janvier 1961 à décembre 1985, soit n = 25×12 valeurs

On considère la **série différenciée** : $Y_t = X_t - X_{t-1}$ ou $Y = (I - B)X$ avec **B opérateur retard**



☛ On considère que la série différenciée $y(t)$ est **stationnaire**

Se pose la question de savoir s'il s'agit d'un « **bruit** » (variables dé-corrélées), on réalise pour cela des **diagrammes retardés** (lag-plot) :



On observe clairement une dépendance linéaire pour le retard ou **décalage** h = 1 **(h = lag)**, on parle d'**autocorrélation d'ordre 1**.

Hypothèse : la série d'observations $(y_t)_t$ est la réalisation (partielle) d'un processus aléatoire $(Y_t)_t$ **stationnaire** au sens suivant :

- **$E(Y_t) = \mu$ constante**
- **$Var(Y_t) = \sigma^2$ constante > 0**
- **$Cov(Y_t, Y_{t+h}) = \gamma(h)$ pour h = 1, 2, …**

On notera **$\rho(h) = Cor(Y_t, Y_{t+h})$ coefficient de corrélation linéaire d'ordre h** entre les variables Y_t et Y_{t+h} :

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

La fonction γ est appelée fonction d'auto-covariance (ACvF) et ρ **fonction d'autocorrélation** (**ACF**) du processus Y.

☞ **temps t = …, -2, -1, 0, 1, 2, … $\in \mathbb{Z}$ (pas d'origine des temps)**

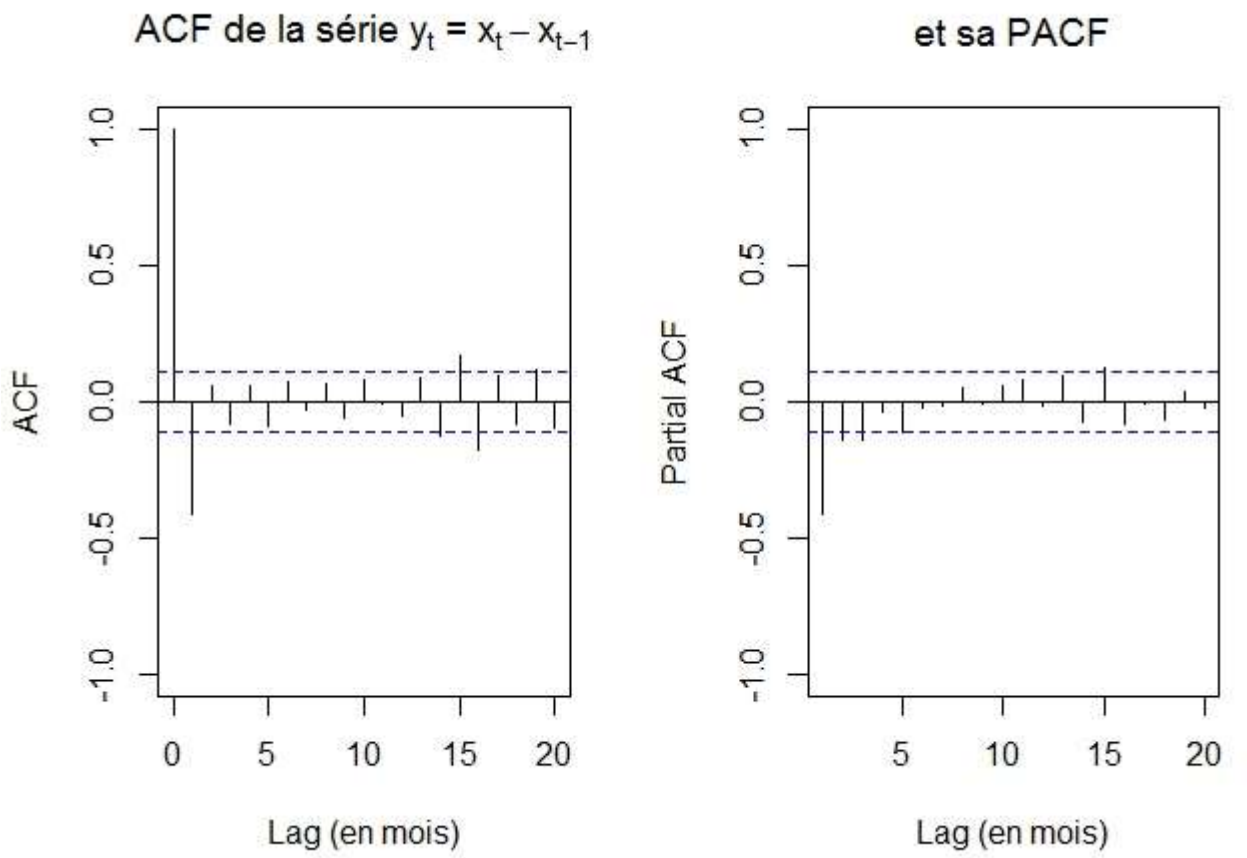
Estimations « naturelles » (moyennes temporelles par opposition à des moyennes spatiales)

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} y_t$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{\mu})^2$$

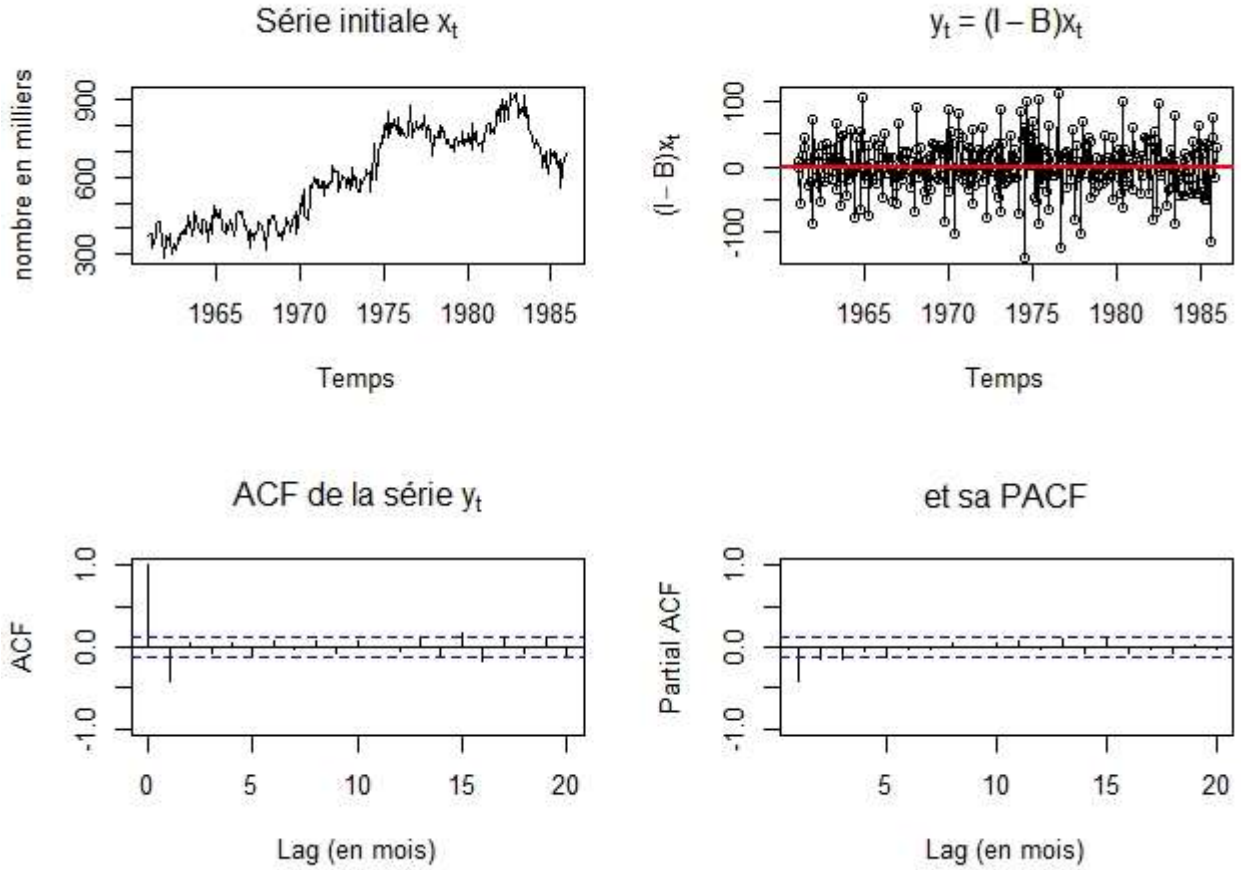
$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (y_{t+h} - \hat{\mu})(y_t - \hat{\mu})$$

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \quad (n \geq 50 \text{ et } h \text{ « petit » } \leq \frac{n}{5})$$



Fonction d'autocorrélation estimée (ACF) en fonction du retard h (lag) à gauche
Fonction d'autocorrélation partielle estimée à droite (cf. infra)
(corrélogramme et corrélogramme partiel)

Bilan provisoire (phase exploration de la série de chômage)



Modélisation probabiliste : le **corrélogramme** suggère que le processus $(Y_t)_t$ serait d'ACF théorique vérifiant

$$\rho_Y(h) = Cor(Y_{t+h}, Y_t) = 0 \text{ pour } h \geq 2$$

et

$$\rho_Y(1) \approx -0.5$$

- ☞ Un tel processus existe-t-il ?
- ☞ Est-il unique ?

Moyenne mobile d'ordre 1 (MA(1) - Moving Average)

Soit $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$ où $\{\varepsilon_t\}$ **est un bruit blanc centré de variance** σ^2.

Alors le processus Y est stationnaire et (réponse à la question de l'existence)

$$\rho_Y(0) = 1 \; ; \; \rho_Y(1) = \frac{\theta}{1+\theta^2}$$

et

$$\rho_Y(h) = 0 \text{ pour } h \geq 2.$$

On observe que $|\rho_Y(1)| \leq \frac{1}{2}$.

MA(1) de moyenne μ : $Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$

Unicité (problème d'identification) Soit Y un processus stationnaire d'ACF nulle à partir du rang 2. Alors :

☞ $|\rho_Y(1)| \leq \frac{1}{2}$

☞ Il existe θ unique dans [-1, 1] tel que $\rho_Y(1) = \frac{\theta}{1+\theta^2}$

☺ Y est un MA(1), i.e. de la forme $Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$ où $\{\varepsilon_t\}$ est un bruit blanc centré.

Bilan modélisation de la série de chômage $(x_t)_t$

La série différenciée $y_t = x_t - x_{t-1}$ (nombre de chômeurs en plus ou en moins d'un mois sur l'autre) est la réalisation d'un processus stationnaire $Y_t = X_t - X_{t-1}$ de type MA(1)

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \text{ avec } \varepsilon \text{ bruit blanc (centré) de variance } \sigma^2$$

☞ le modèle pour la série initiale $(x_t)_t$ est de la forme ARIMA(p, d, q) avec p = 0 (pas de partie autorégressive), d=1 (car la série est différenciée une fois) et q=1 pour l'ordre de la partie moyenne mobile : $x \sim$ ARIMA(0, 1, 1)

ARIMA = Autoregressive Integrated Moving Average
(moyenne mobile autorégressive intégrée)

On laisse provisoirement de côté le problème de l'estimation des paramètres θ et σ^2 (ajustement du modèle ou model fitting) ainsi que celui de la validation du modèle à partir des données (qualité de l'ajustement ou goodness of fit) pour s'intéresser au problème théorique de la **prévision** avec un tel modèle.

Prévision avec un modèle de type MA(1) : on suppose $\mu = 0$ sans perte de généralité et

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1} \text{ avec } |\theta| < 1.$$

On dispose des observations $Y_1, Y_2, \ldots, Y_n$. On note $\rho = \rho_Y(1) = \dfrac{\theta}{1+\theta^2}$

Prédicteur linéaire optimal à un pas de temps basé sur la dernière observation

$$\hat{Y}_{n+1,\,1} = E_L(\, Y_{n+1} \mid Y_n\,) = \rho Y_n$$

Variance de l'erreur de prévision correspondante

$$v_{n,\,1} = E(\, Y_{n+1} - \hat{Y}_{n+1,\,1}\,)^2 = (1 - \rho^2\,)\times\sigma_Y^2$$

Prédicteur linéaire optimal basé sur les deux dernières observations Y_n et Y_{n-1}

$$\hat{Y}_{n+1,2} = E_L(Y_{n+1} \mid Y_n, Y_{n-1}) = \frac{\rho}{1-\rho^2} Y_n - \frac{\rho^2}{1-\rho^2} Y_{n-1}$$

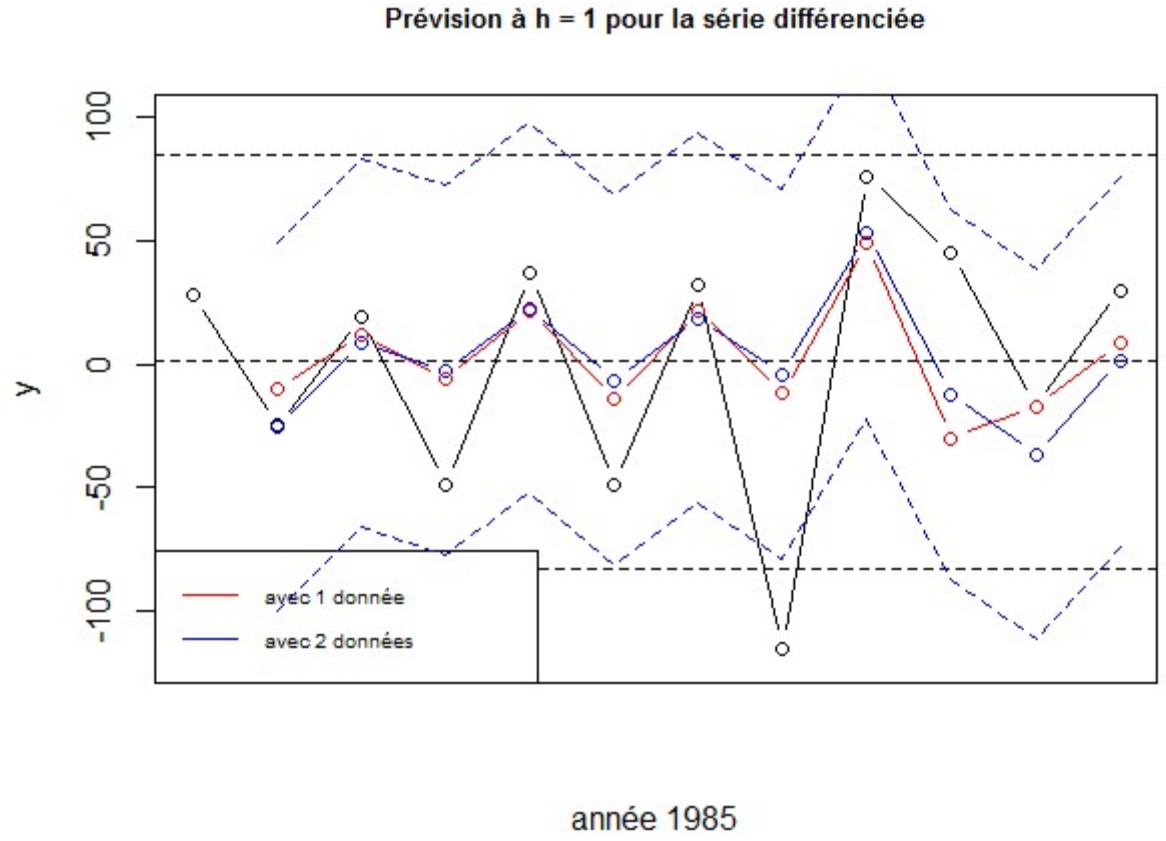
$$v_{n,2} = E(Y_{n+1} - \hat{Y}_{n+1,2})^2 = (1-\rho^2)\times(1-\pi_Y(2)^2)\times\sigma_Y^2$$

en notant

$$\pi_Y(2) = -\frac{\rho^2}{1-\rho^2} = Cor(Y_{n+1} - \rho Y_n, Y_{n-1} - \rho Y_n)$$

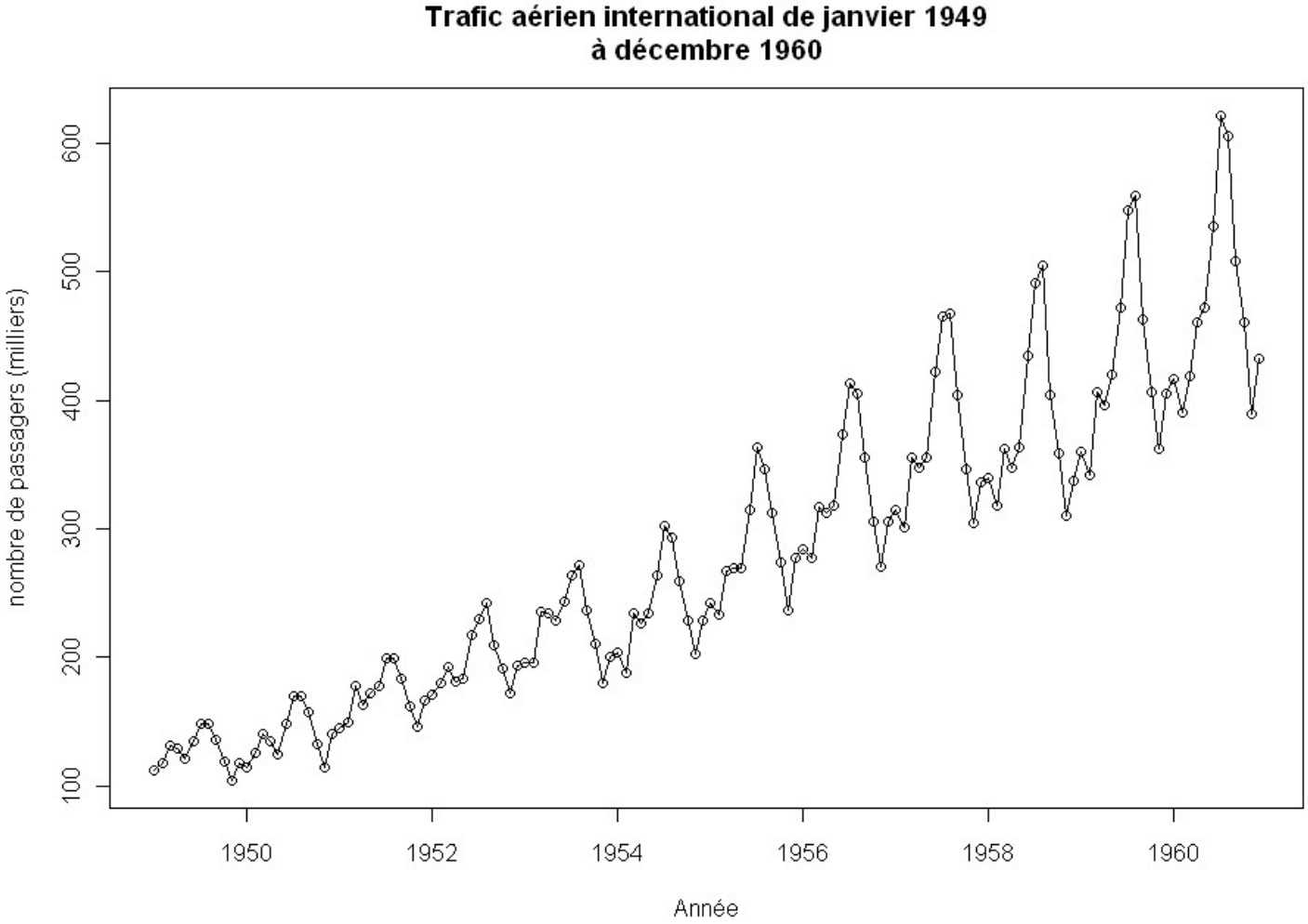
☞ **$\pi_Y(2)$ est appelé coefficient d'autocorrélation partielle d'ordre 2**

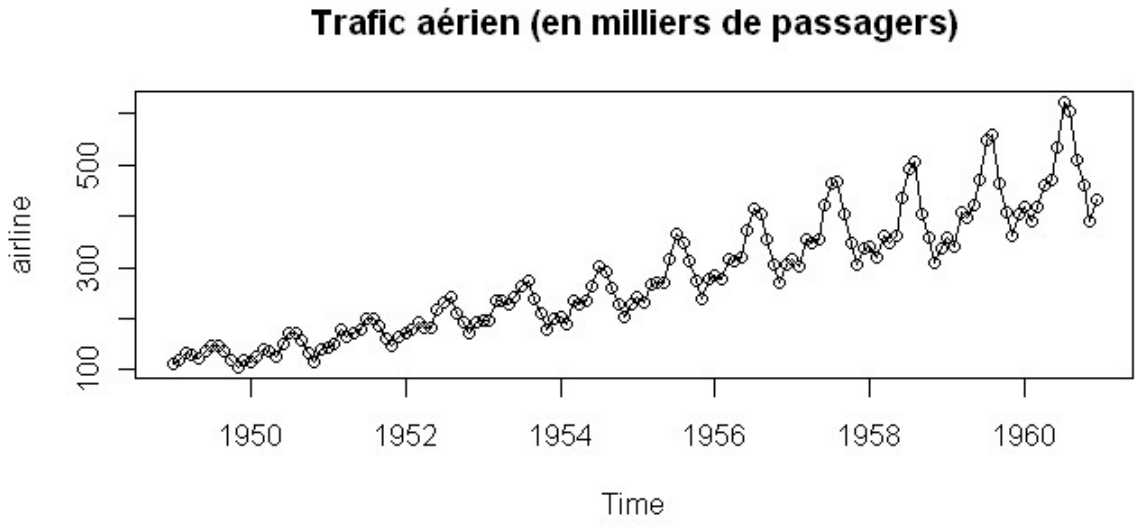
Bien que la variable Y_{n-1} soit non corrélée avec Y_{n+1}, elle améliore sa prévision car le coefficient de corrélation partielle est non nul !

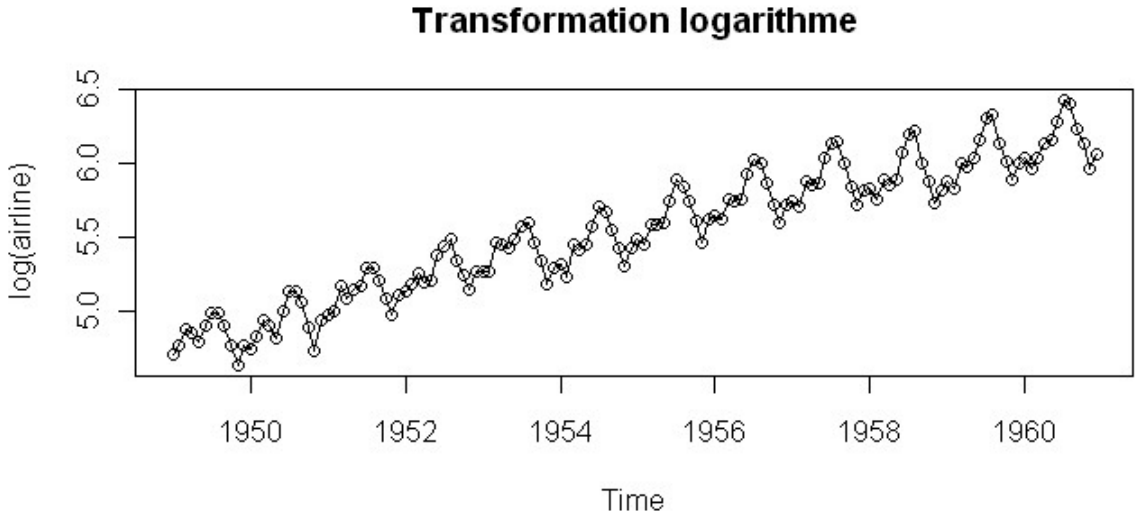


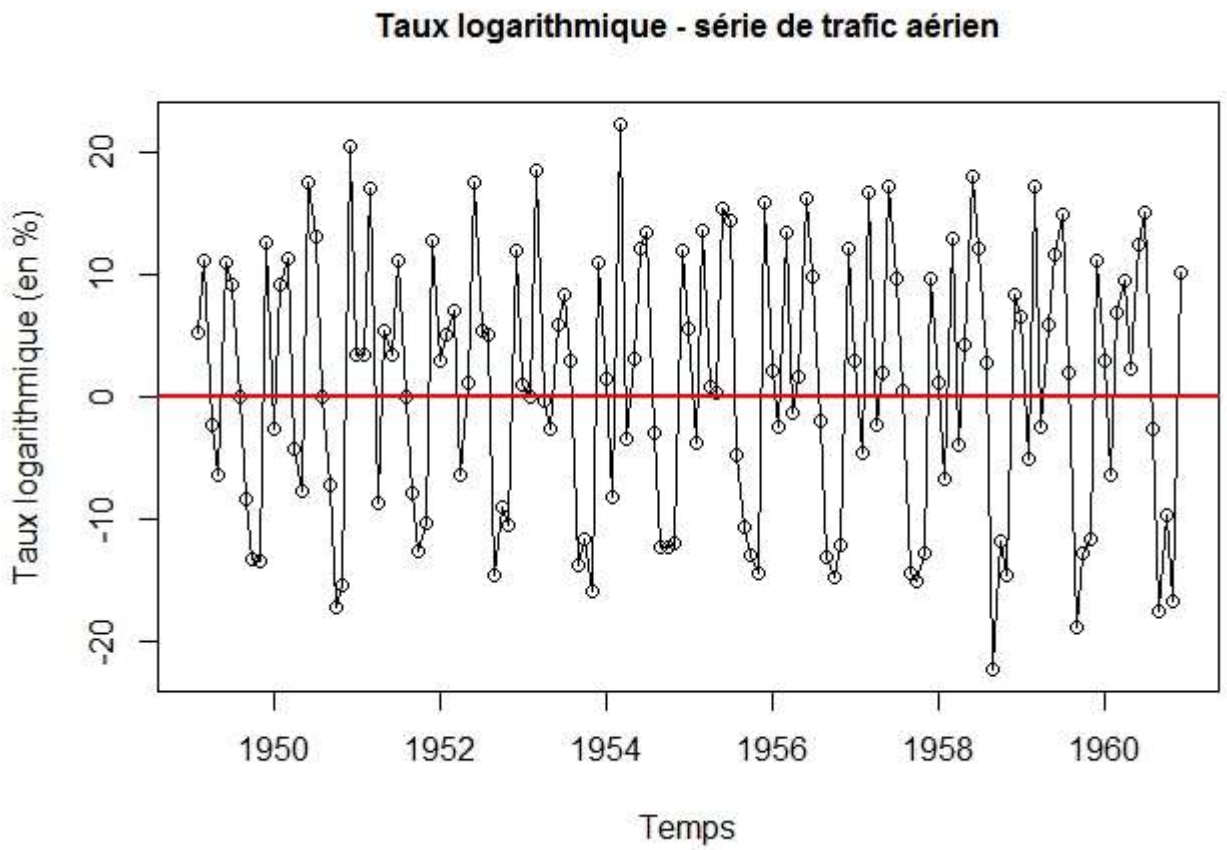
Test rétro-actif de prévision ou backtesting

Exemple 1 du support : chronogramme de la série de trafic aérien




Trafic aérien (en milliers de passagers)
airline
100
300
500
1950
1952
1954
1956
1958
1960
Time

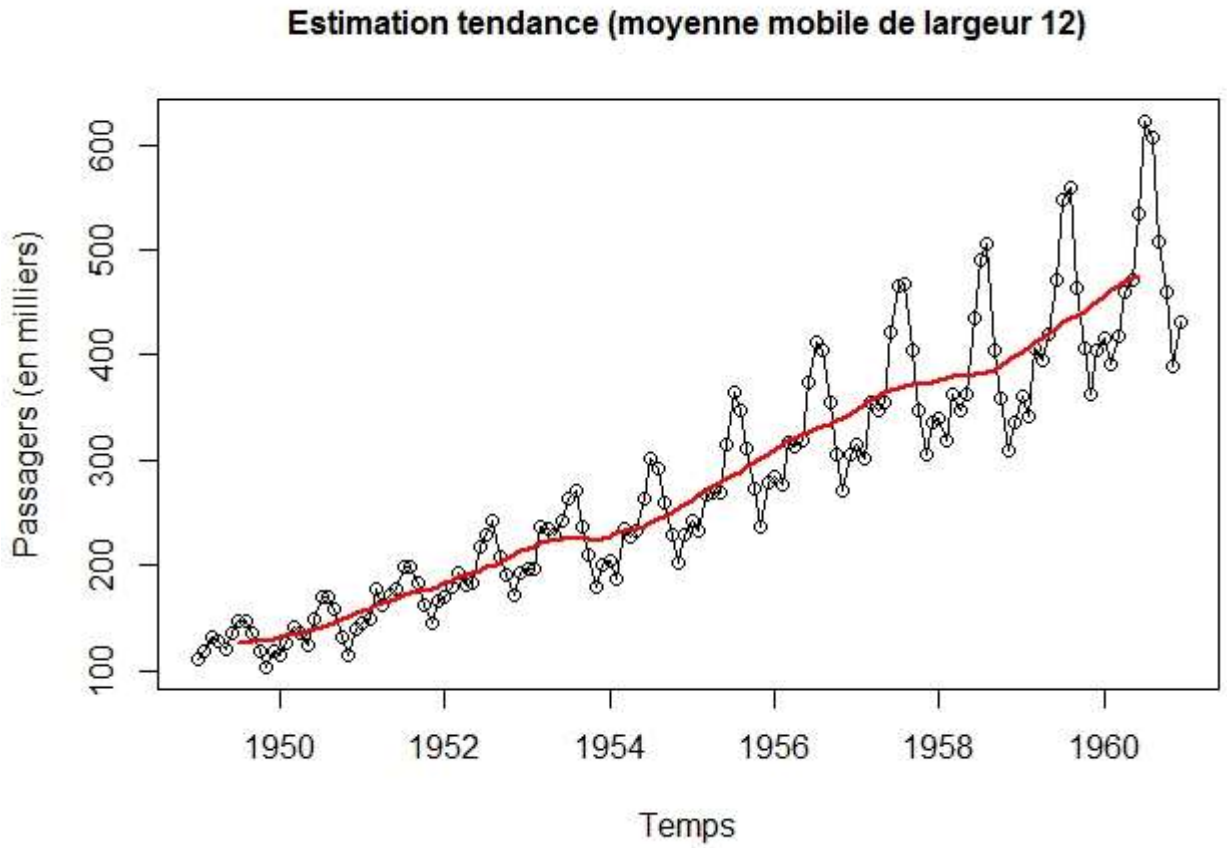

Transformation logarithme
log(airline)
5.0
5.5
6.0
6.5
1950
1952
1954
1956
1958
1960
Time

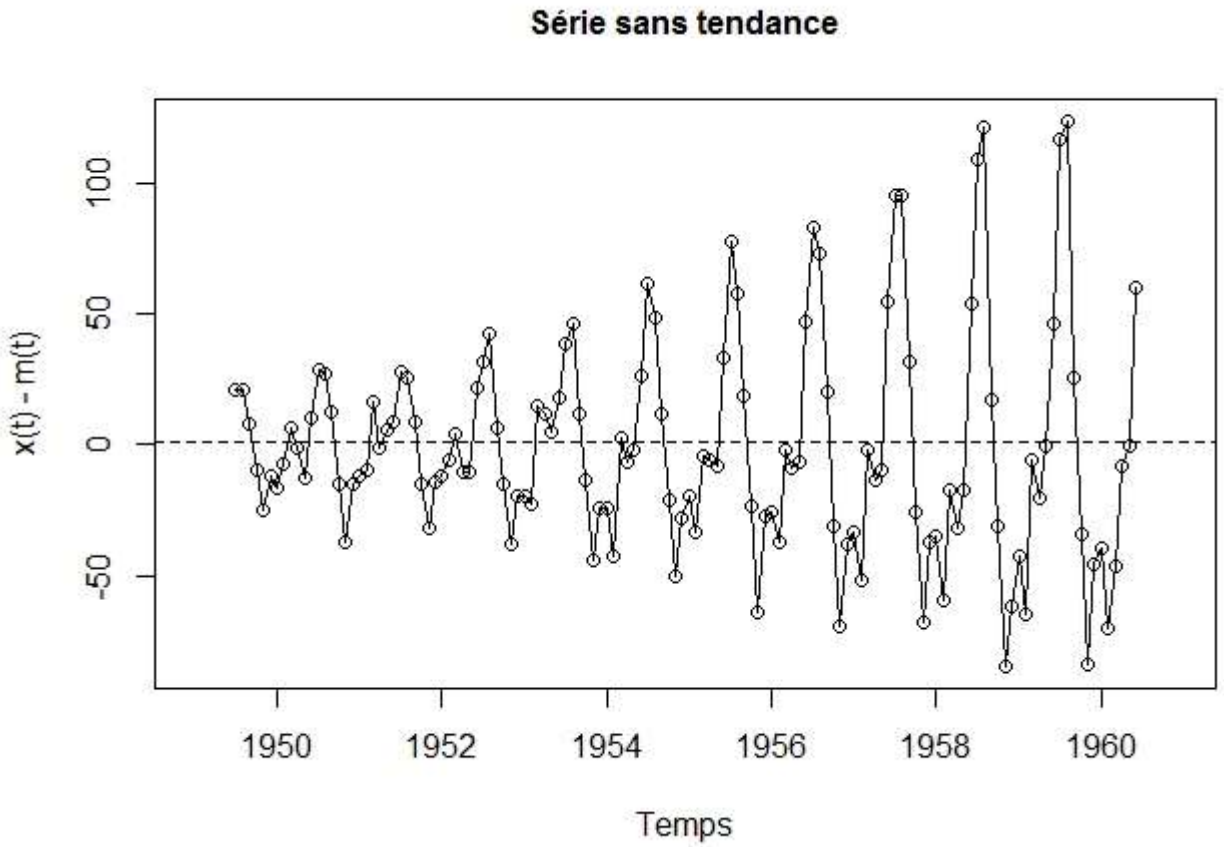


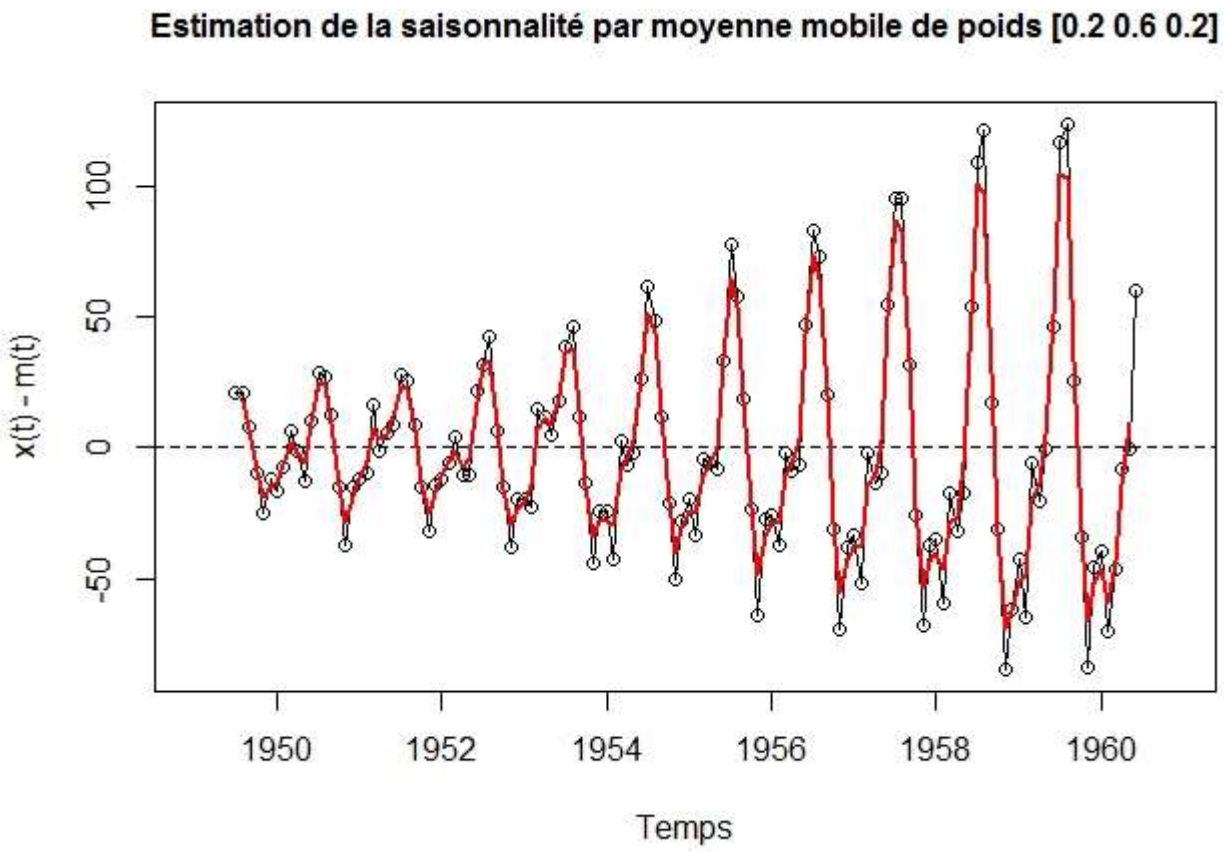
Interprétation (du passage au log). Avec $y_t = \log(x_t)$, on obtient que

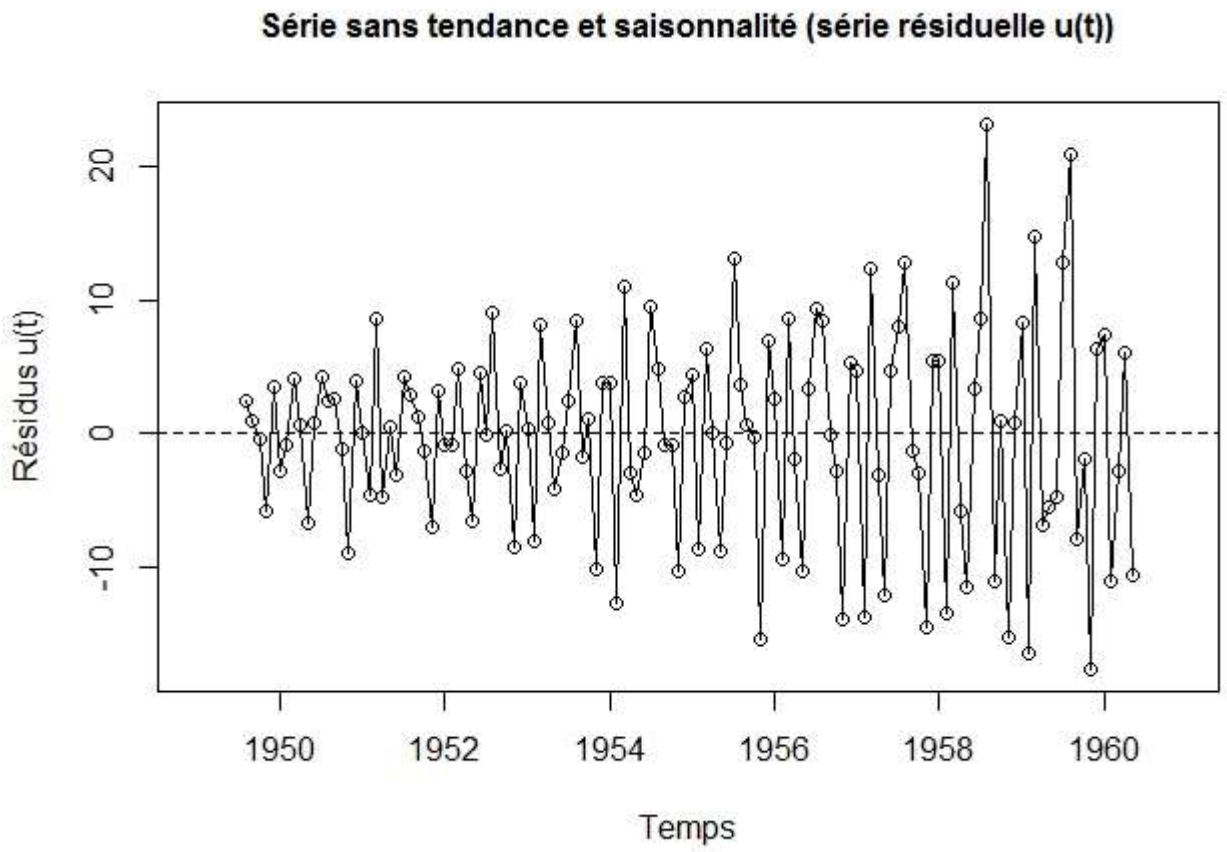
$$\nabla y_t = y_t - y_{t-1} \approx \frac{x_t - x_{t-1}}{x_{t-1}} \qquad \text{(si petites variations)}$$

Modèle additif : $x_t = m_t + s_t + u_t$ (cf. support page 6)

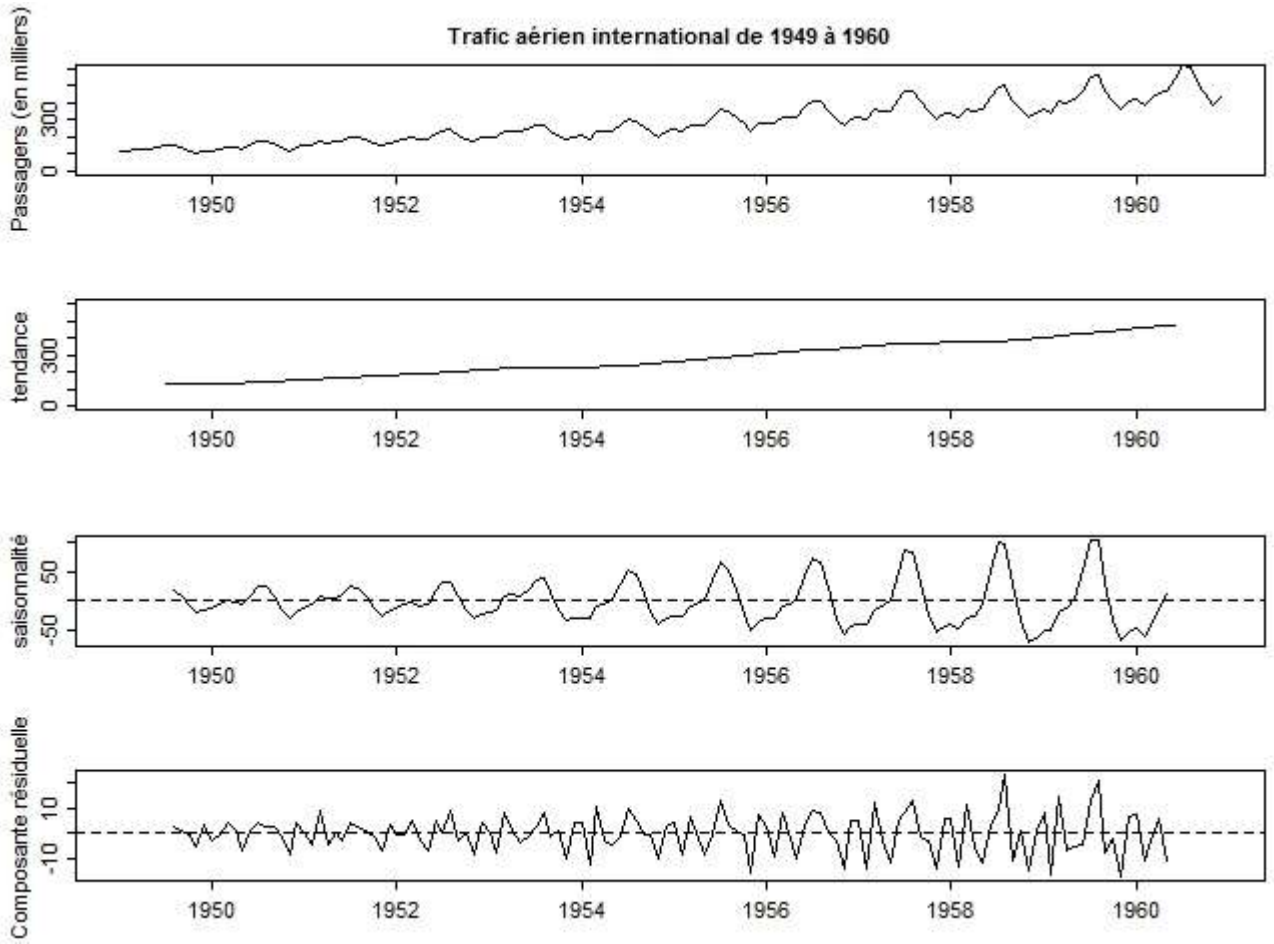



Série sans tendance
x(t) - m(t)
-50
0
50
100
1950
1952
1954
1956
1958
1960
Temps


Estimation de la saisonnalité par moyenne mobile de poids [0.2 0.6 0.2]
x(t) - m(t)
100
50
0
-50
1950
1952
1954
1956
1958
1960
Temps



Série stationnaire ?



Décomposition finale : $x_t = m_t + s_t + u_t$

Modèle multiplicatif :

$y_t = \log(x_t)$ transformation logarithmique

$y_t = m_t + s_t + u_t$ décomposition additive de y

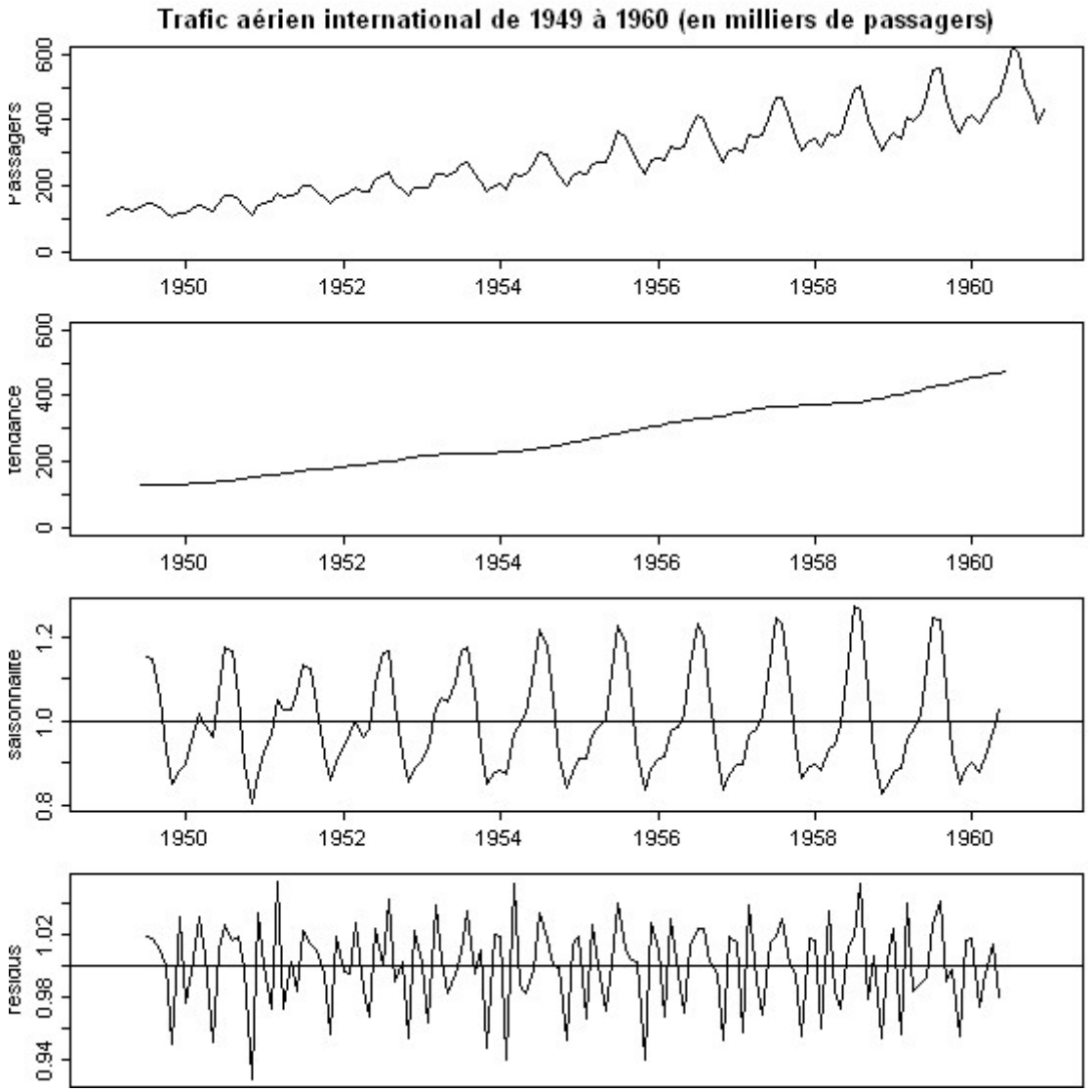
$\Rightarrow$ $\boxed{x_t = M_t \times S_t \times U_t}$ **décomposition multiplicative de x**

où

$M_t = \exp(m_t)$ tendance de la série x

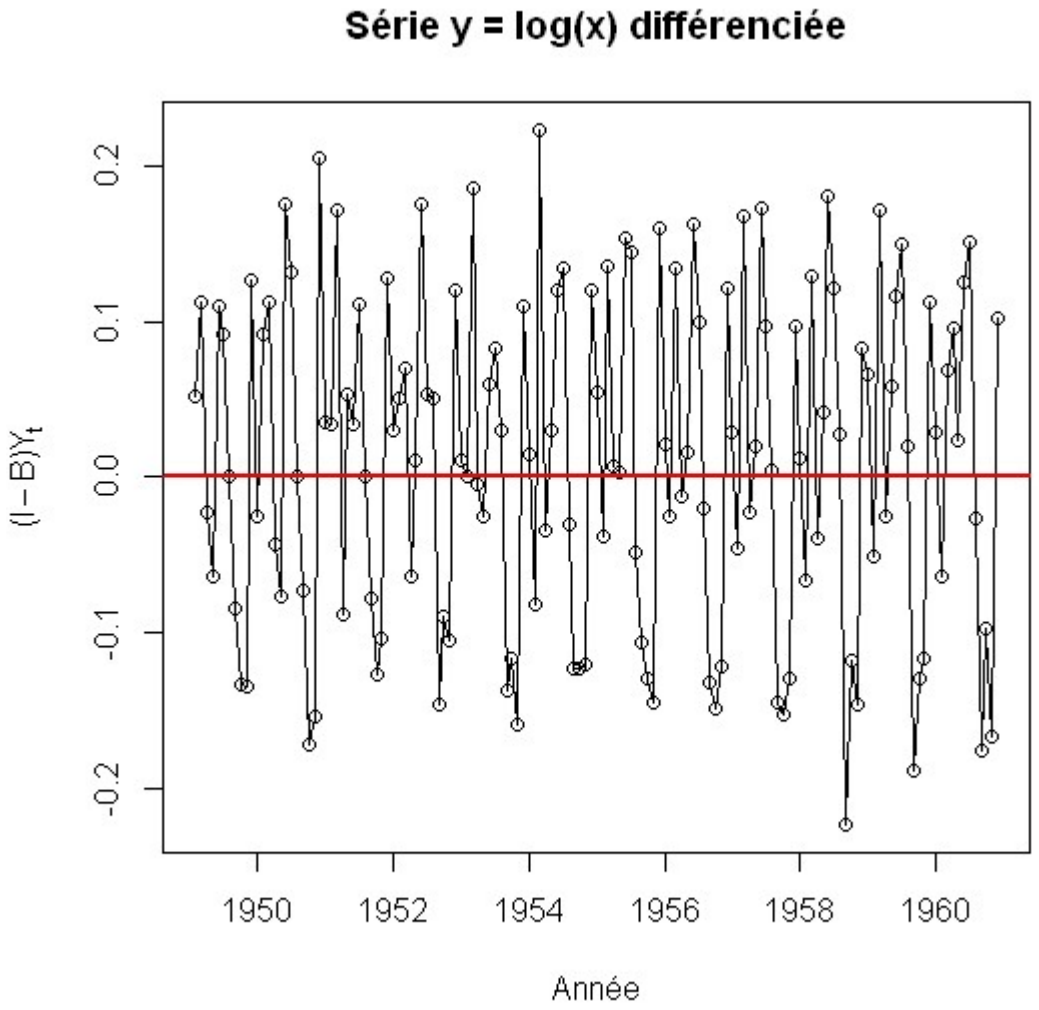
$S_t = \exp(s_t)$ indice saisonnier

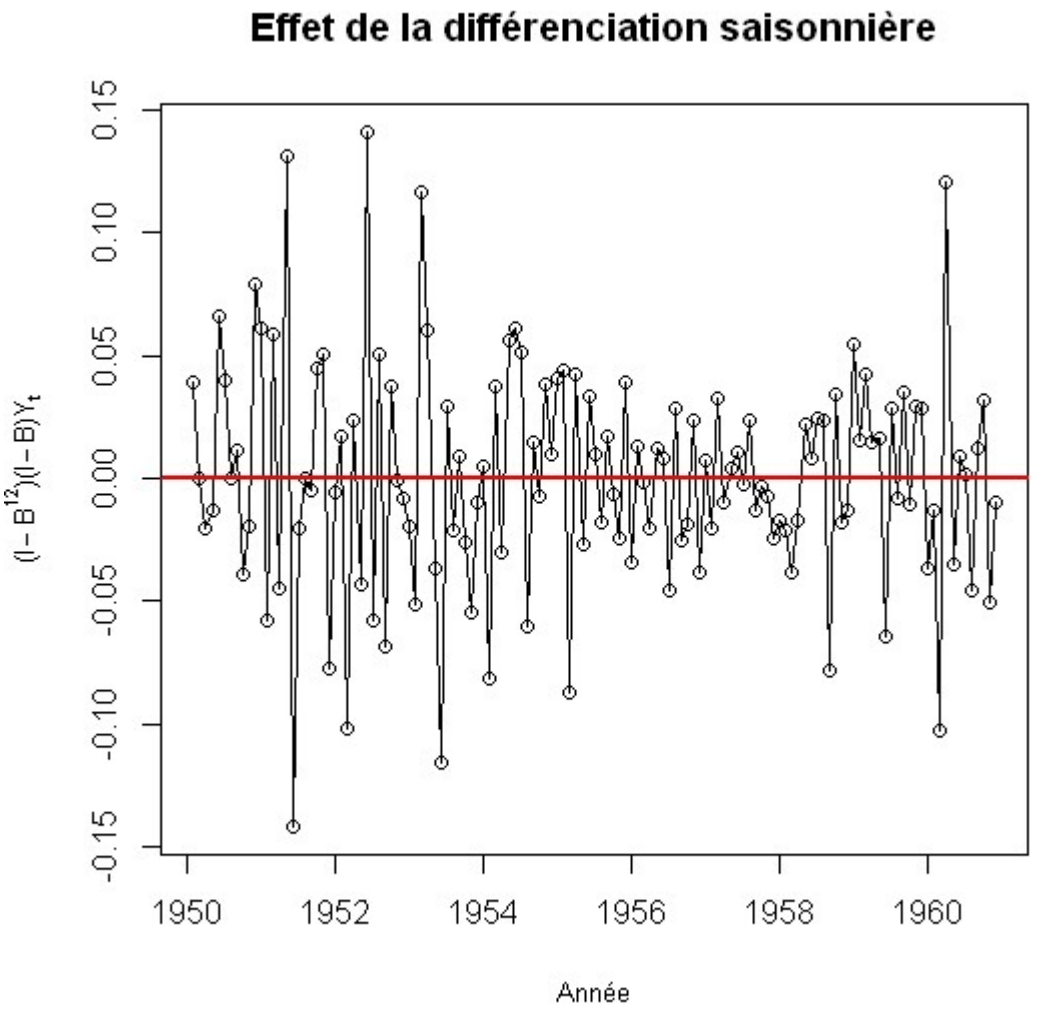
$U_t = \exp(u_t)$ indice aléatoire

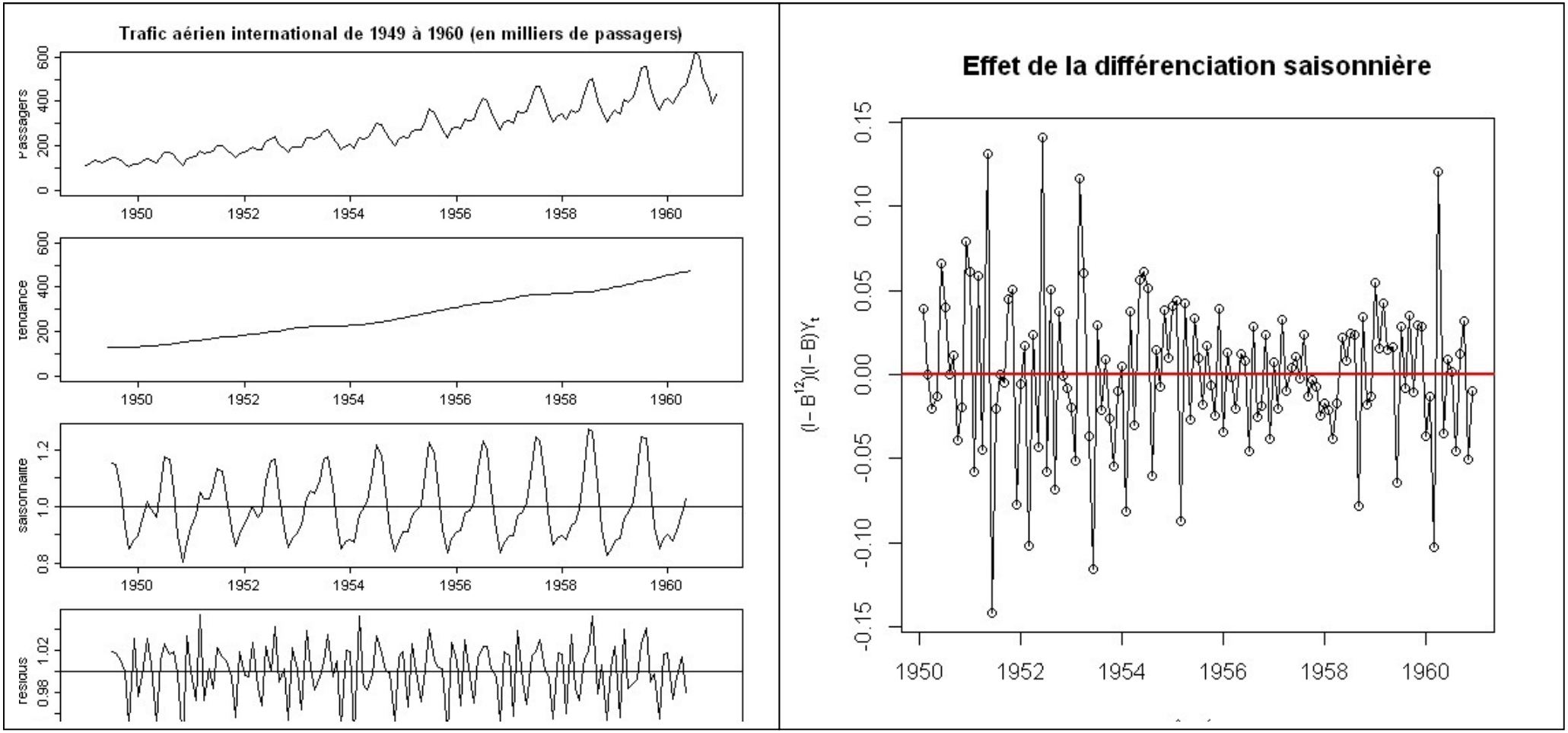


Décomposition multiplicative $x_t = M_t \times S_t \times U_t$

Seconde approche par différenciation (méthodologie de Box&Jenkins)

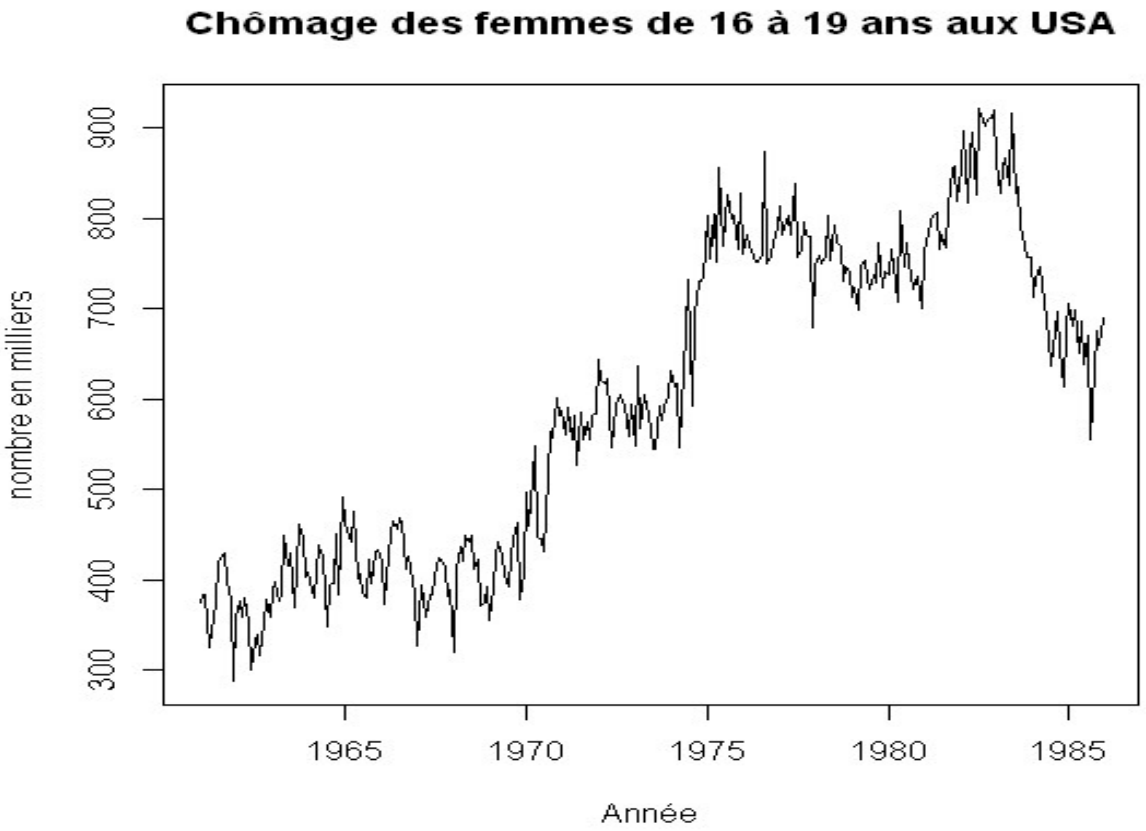



Effet de la différenciation saisonnière
$(I-B^{12})(I-B)Y_t$
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
1950
1952
1954
1956
1958
1960
Année

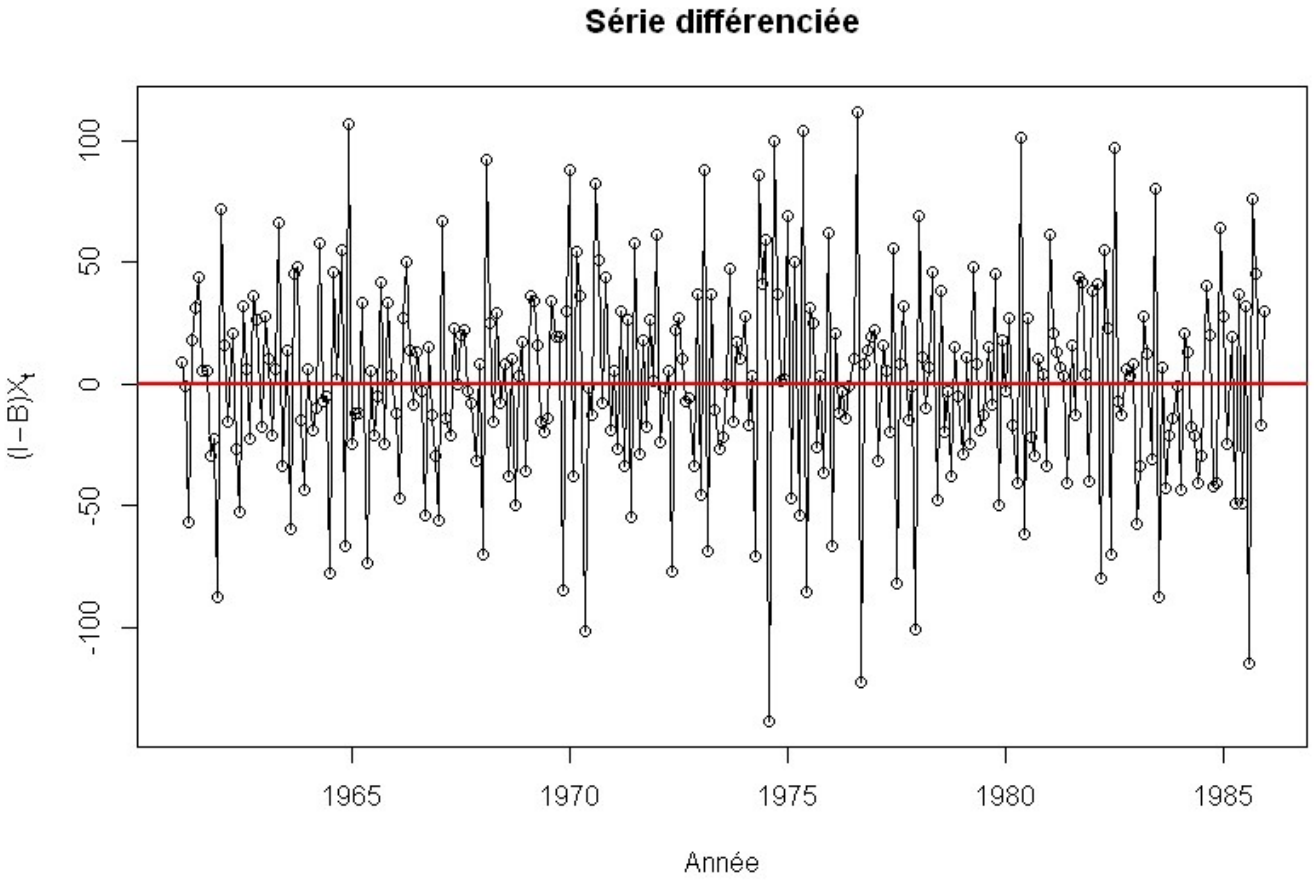


« Stationnarisation » de la série de trafic aérien (2 approches)

Exemple 2 du support (pour rappel, série modélisée par un ARMA(0,1,1)) :

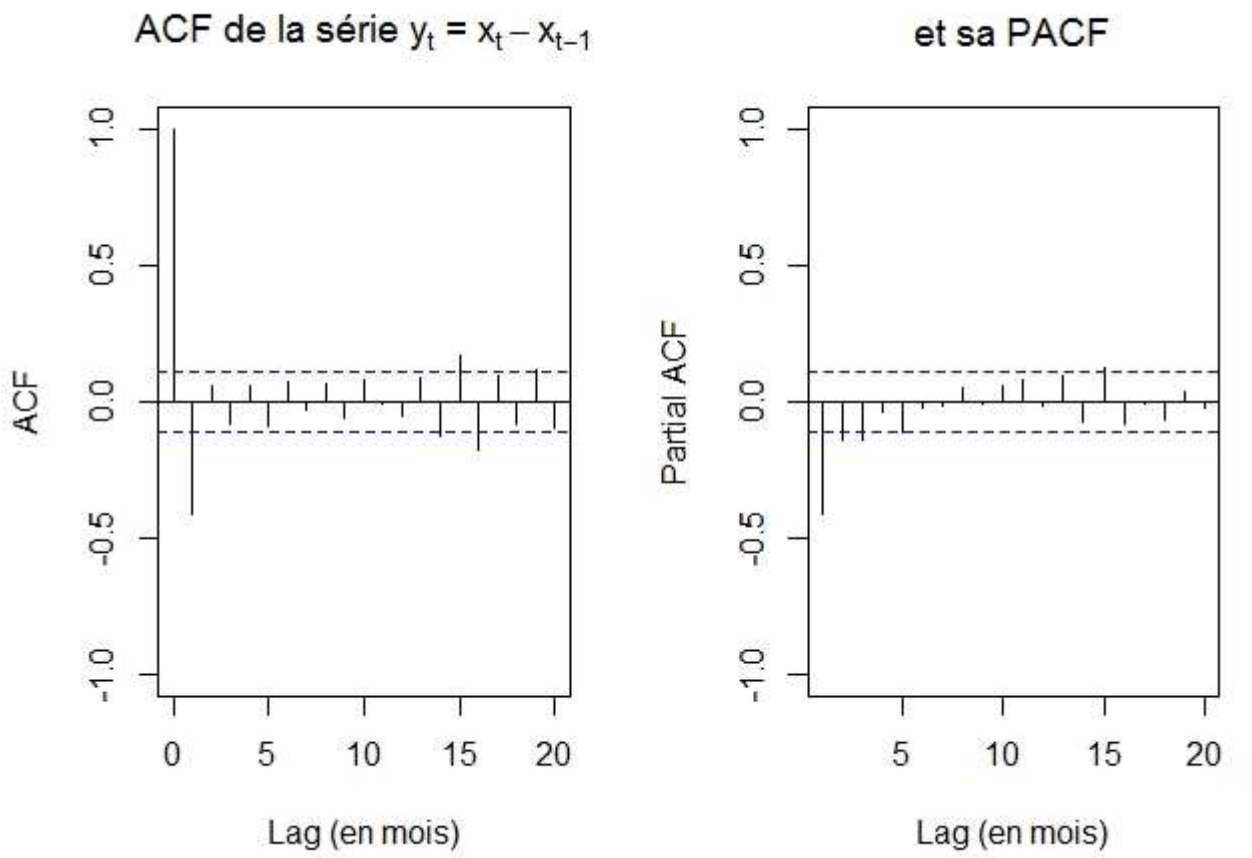


Chronogramme de la série de chômage



Variations d'un mois sur l'autre : série stationnaire

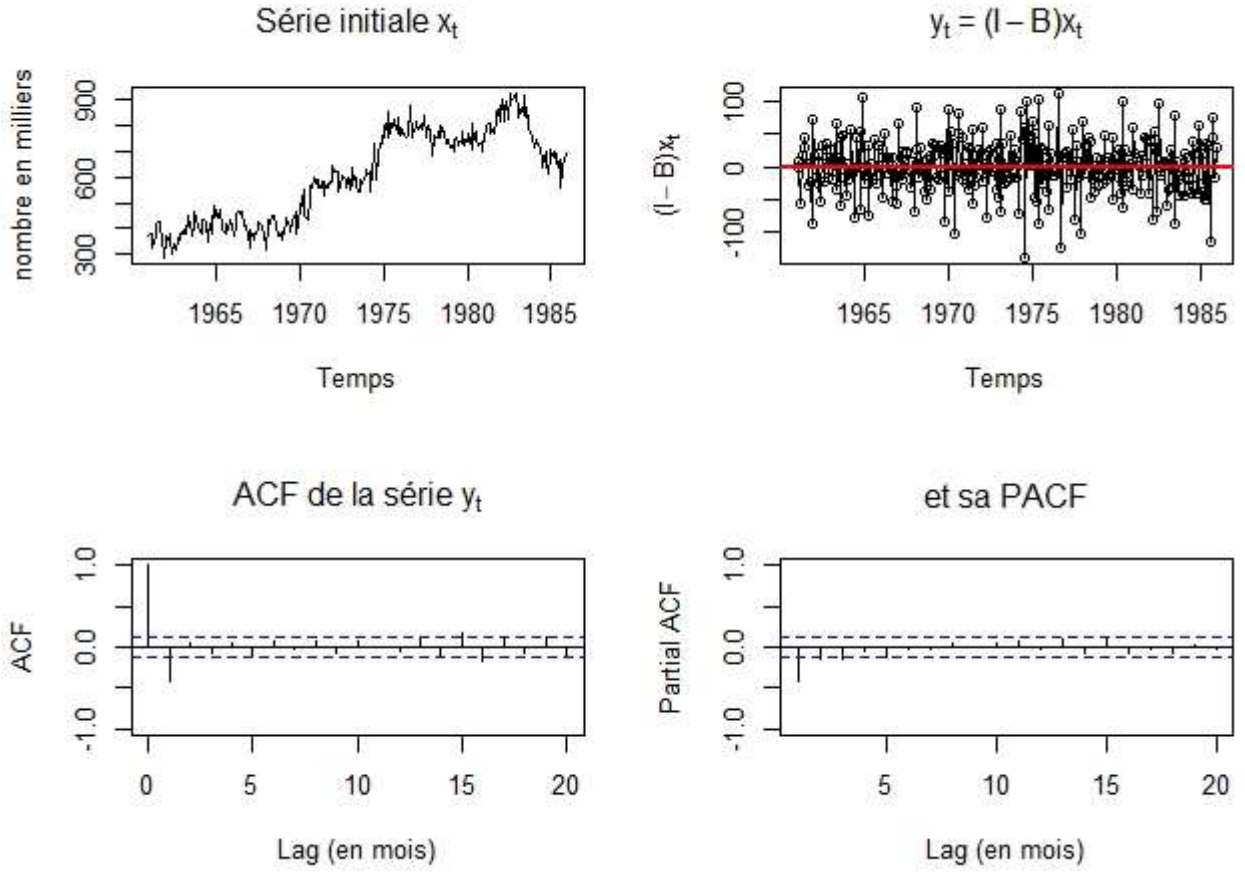
$$Y_t = (I - B)X_t = X_t - X_{t-1}$$



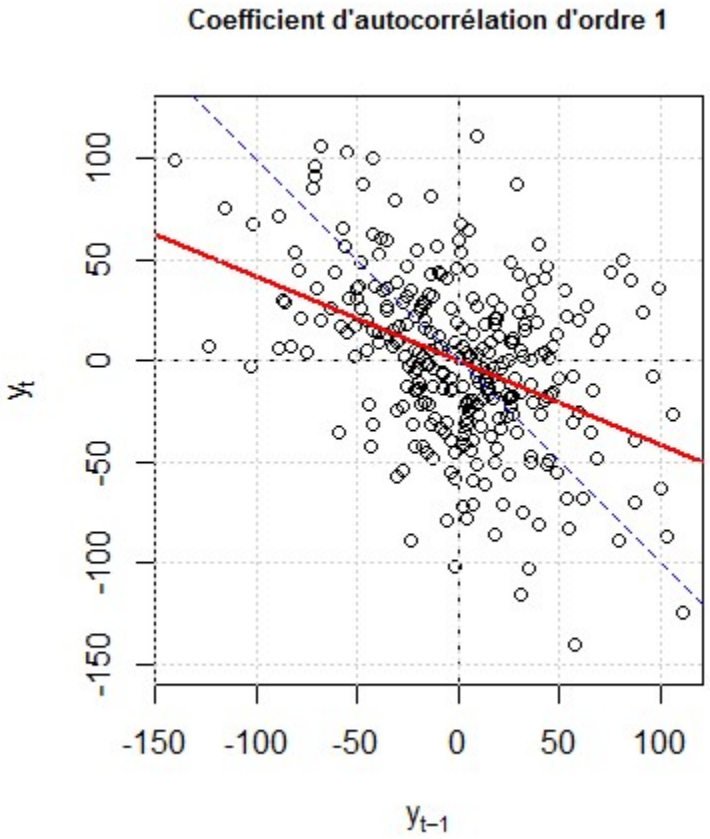
Fonction d'autocorrélation (ACF) en fonction du retard h (lag) à gauche
Fonction d'autocorrélation partielle à droite (cf. infra)

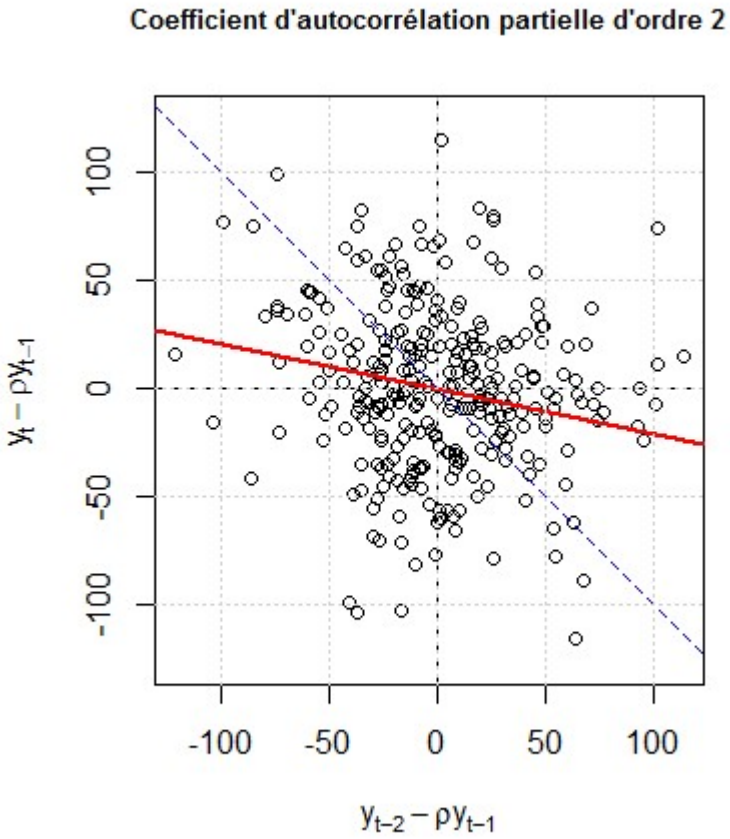
ACF nulle à partir de h = 2 ⇒ MA(q=1)

Bilan stationnarisation de la série de chômage :



☛ on considère que la série différenciée *y*(t) est stationnaire





$$\rho(1) = \pi(1) \approx -0.41 \text{ et } \pi(2) \approx -0.21$$

.

$(X_t)_t$ processus aléatoire du second ordre de **moyenne μ constante** est dit **stationnaire** (à l'ordre 2) si

$\gamma(h) = Cov(X_{t+h}, X_t)$ ne dépend pas de t pour tout $h = 0, \pm 1, \pm 2, \ldots$

ce qui permet de définir l'objet fondamental (sous ses deux formes classiques):

(**ACvF**) Fonction d'auto-covariance : $\gamma(h) = Cov(X_{t+h}, X_t)$

(**ACF**) Fonction d'auto-corrélation : $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = Cor(X_{t+h}, X_t)$

☞ La propriété de stationnarité est difficile à tester directement (en général) à partir d'une seule série de données. Cette seule propriété ne suffit pas en pratique…

On définit maintenant une classe importante de processus stationnaires ayant de très bonnes propriétés et au cœur de l'analyse de séries chronologiques.

Un processus **stationnaire** (à l'ordre 2) est de type **ARMA(p, q)** (p et q entiers) s'il vérifie une relation de la forme

$$\Phi(B)X_t = \Theta(B)Z_t \qquad \text{avec } \{Z_t\} \sim WN(0, \sigma^2)$$

et où $\Phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ est un polynôme de degré p

$\Theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ est un polynôme de degré q

Notations : **WN = « White Noise »** = « bruit blanc » ; $\{Z_t\}$ noté parfois $\{\varepsilon_t\}$

σ^2 = variance du bruit (si ambiguïté, on la notera σ_Z^2 ou σ_ε^2)

Cas particuliers : p = 0, ARMA(0, q) = MA(q)

q = 0, ARMA(p, 0) = AR(p)

On appelle **processus linéaire** tout processus $\{X_t\}$ tel que

$$X_t = \sum_{j=-\infty}^{+\infty} \psi_j Z_{t-j} \quad \text{avec} \quad \sum_{j=-\infty}^{+\infty} |\psi_j| < +\infty$$

Propriétés. Soit Y stationnaire (centré) et $X_t = \sum_{j=-\infty}^{+\infty} \psi_j Y_{t-j}$ ($\sum_{j=-\infty}^{+\infty} |\psi_j| < +\infty$).

Alors, X est stationnaire d'ACvF :

$$\gamma_X(h) = \sum_{j,k=-\infty}^{+\infty} \psi_j \psi_k \gamma_Y(h+k-j)$$

En particulier, si $Y = Z \sim WN(0, \sigma^2)$, alors : $\gamma_X(h) = \sum_{j=-\infty}^{+\infty} \psi_j \psi_{j+h} \sigma^2$

Processus linéaire = sortie d'un filtre (linéaire) stationnaire où l'entrée est un bruit blanc

Propriétés de l'ACF (**A**uto**C**orrelation **F**unction / fonction d'autocorrélation) :

$\rho(0) = 1 \quad ; \quad |\rho(h)| \leq 1 \quad ; \quad \rho(h) = \rho(-h)$ (ρ paire)

$\rho(h) = \cos(\alpha(h))$ où $\alpha(h)$ = angle entre X_t et X_{t+h} (si X centré)

$\gamma(h) = \gamma(0) \times \rho(h) = \sigma_X^2 \times \rho(h)$ (ACvF = **A**uto**C**o**v**ariance **F**unction)

Moyenne μ, variance $\sigma_X^2 = \gamma(0)$ et ACF ρ du processus $\Rightarrow$ on dispose de toute la connaissance nécessaire pour calculer le prédicteur linéaire optimal à un pas de temps (mais aussi à plusieurs pas de temps)

$$\hat{X}_{n+1} = E_L(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1) = \phi_{n,1} X_n + \ldots + \phi_{n,n} X_1$$

ainsi que la variance de l'erreur de prévision : $v_n = E(X_{n+1} - \hat{X}_{n+1})^2$

Réversibilité (dans le temps) : si X est un processus stationnaire de moyenne μ, de variance σ_X^2 et de structure de corrélation linéaire ou ACF $\rho(h)$, alors il en est de même du processus

$$\tilde{X} : t \to X_{-t}$$

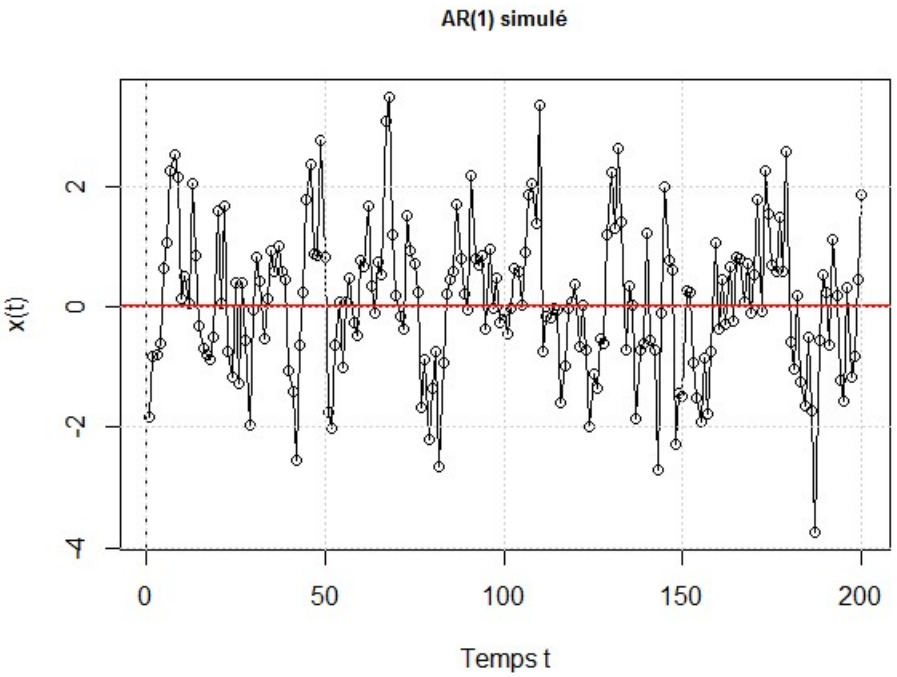
En particulier (pour remonter dans le passé, d'où l'on vient ? ☺), on a :

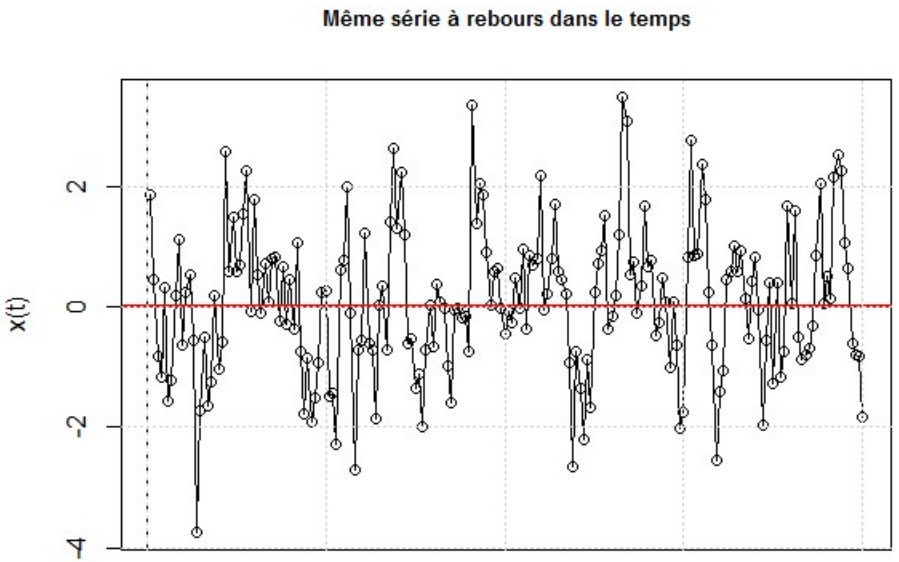
$$\hat{X}_0 = E_L(X_0 \mid X_1, X_2, \ldots, X_n) = \phi_{n,1} X_1 + \ldots + \phi_{n,n} X_n$$

ainsi que la même erreur de prévision

$$E(X_0 - \hat{X}_0)^2 = v_n = E(X_{n+1} - \hat{X}_{n+1})^2$$

Illustration sur un AR(1)





Pour aider à identifier un processus stationnaire, on définit encore la

PACF (Partial AutoCorrelation Function / fonction d'autocorrélation partielle)

$$\pi(h) = \mathrm{Cor}\left(X_{t+h} - E_L(X_{t+h} \mid X_{t+h-1}, \ldots, X_{t+1}) \, , \, X_t - E_L(X_t \mid X_{t+1}, \ldots, X_{t+h-1}) \right)$$

où l'on convient $\pi(1) = \rho(1)$ ($\pi(0)$ non défini).

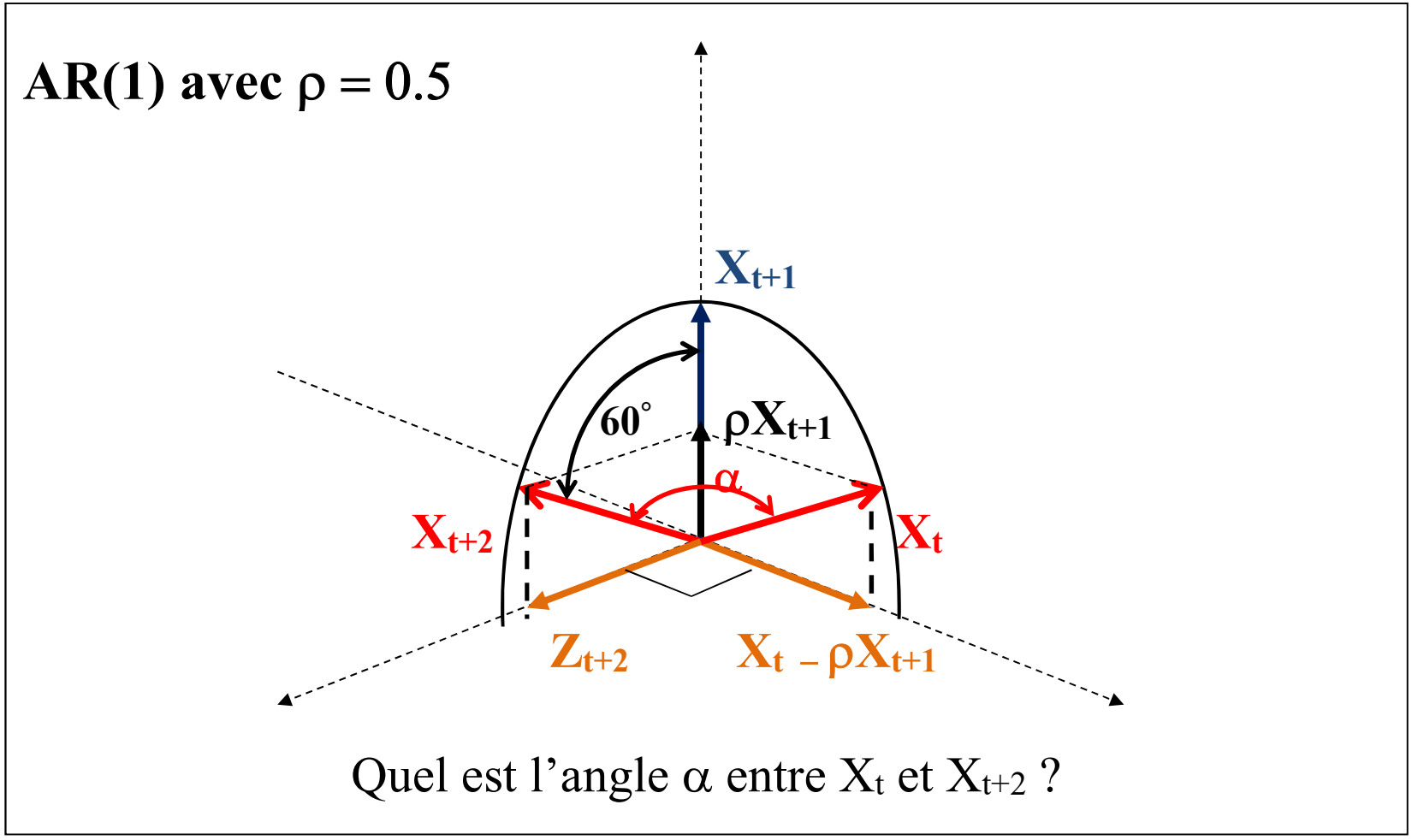
Cas d'un AR(1) : $X_t - \rho X_{t-1} = Z_t$

$$\pi(1) = \rho \text{ et } \pi(h) = 0 \text{ pour } h \geq 2$$

Cas d'un MA(1) : $X_t = Z_t + \theta Z_{t-1}$

$$\pi(h) = (-1)^{h-1} \frac{\theta^h (\theta^2 - 1)}{\theta^{2(h+1)} - 1} \qquad (h \geq 1)$$

Illustration pour un AR(1) (voir aussi support page 23)



Proposition : de la projection orthogonale,

$$E_L(X_{t+h} \mid X_{t+h-1}, \ldots, X_{t+1}, \mathbf{X_t}) = \phi_{h,1} X_{t+h-1} + \ldots + \mathbf{\phi_{h,h} X_t}$$

on déduit simplement le coefficient d'autocorrélation partielle d'ordre h :

$$\pi(h) = \mathbf{\phi_{h,h}}$$

Interprétation : soit $\hat{X}_{n+1} = E_L(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1)$ le prédicteur optimal. Alors

$$\mathbf{X_{n+1} - E_L(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_2) = \phi_{n,n} \times (X_1 - E_L(X_1 \mid X_2, X_3, \ldots, X_n)) + X_{n+1} - \hat{X}_{n+1}}$$

et l'erreur de prévision $v_n = E(X_{n+1} - \hat{X}_{n+1})^2$ vérifie

$$\mathbf{v_n = (1 - \phi_{n,n}^2) \times v_{n-1} \; ; \; v_0 = \gamma(0)}$$

Le résultat fondamental :

$$\mathbf{v_n = E(X_{n+1} - \hat{X}_{n+1})^2 = \sigma_X^2 \times (1 - \rho^2) \times (1 - \pi(2)^2) \times \ldots \times (1 - \pi(n)^2)}$$

où $\rho = \rho(1)$ coefficient d'autocorrélation d'ordre 1 et π = PACF

On a défini un modèle **ARMA(p, q)** (p et q entiers) comme un processus stationnaire (à l'ordre 2) qui vérifie une relation du type

$$\Phi(B)X_t = \Theta(B)Z_t \; ; \; \{Z_t\} \text{ WN}(0, \sigma^2), \; \Phi \text{ et } \Theta \text{ polynômes de degrés resp. p et q}$$

☞ On suppose que les racines de Φ sont toutes de module > 1 et pour Θ de module ≥ 1 et enfin que ces polynômes n'ont pas de racine commune : X est dit **causal** et **inversible**.

Un modèle ARMA(p, q) non centré sera de la forme

$$X = \mu + Y \text{ avec } Y \text{ ARMA}(p, q) \Leftrightarrow X - \mu \sim \text{ARMA}(p, q)$$

Pour ajuster un tel modèle, il faut commencer par estimer sa moyenne μ.

Estimateur usuel de la moyenne : $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$

C'est un estimateur sans biais et :

$$\mathbf{Var}(\bar{X}) = \frac{1}{n} \sum_{h=-(n-1)}^{n-1} (1 - \frac{|h|}{n}) \gamma(h) \ .$$

De plus, $\sqrt{n}\,(\bar{X} - \mu)$ **est asymptotiquement de loi** $N(0 ; \sigma_X^2 (1 + 2 \sum_{h \geq 1} \rho(h)))$

On estime maintenant l'ACvF du processus X ainsi que l'ACF (coefficients d'auto-corrélation $\rho(h)$).

Estimation ACvF et ACF par leurs analogues empiriques :

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} x_t$$

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \hat{\mu})(x_t - \hat{\mu}) \; ; \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \; (n \geq 50 \text{ et } h \text{ « petit » } \leq \frac{n}{5})$$

Estimation de la PACF : $\hat{\pi}(h) = \hat{\phi}_{h,h}$ sachant que

$$E_L(X_{t+h} \mid X_{t+h-1}, \ldots, X_t) = \phi_{h,1} X_{t+h-1} + \ldots + \phi_{h,h} X_t$$

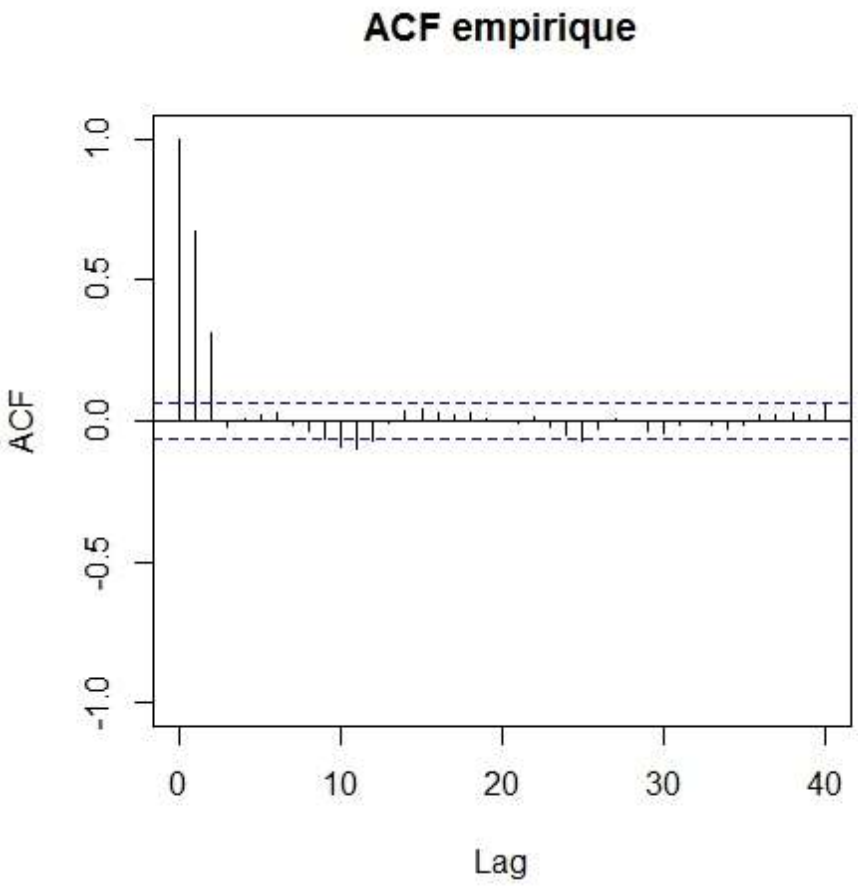
Cas particulier h = 0 :

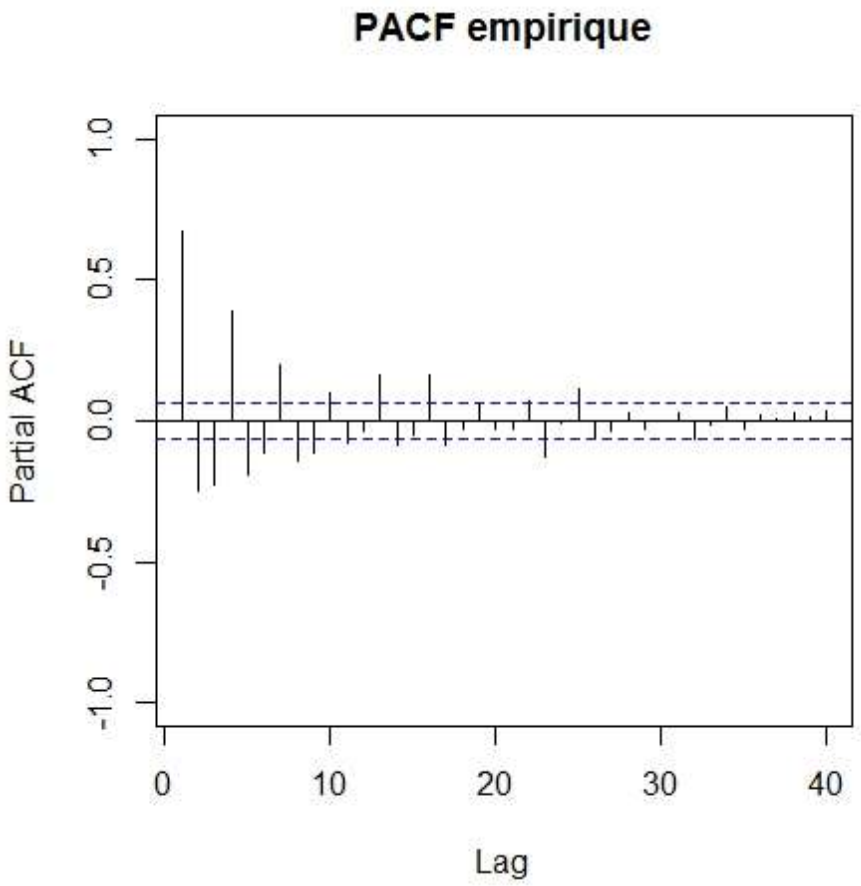
$$\hat{\sigma}_X^2 = \hat{\gamma}(0) = \frac{1}{n} \sum_{t=1}^{n} (X_t - \bar{X})^2$$

Considérations pratiques

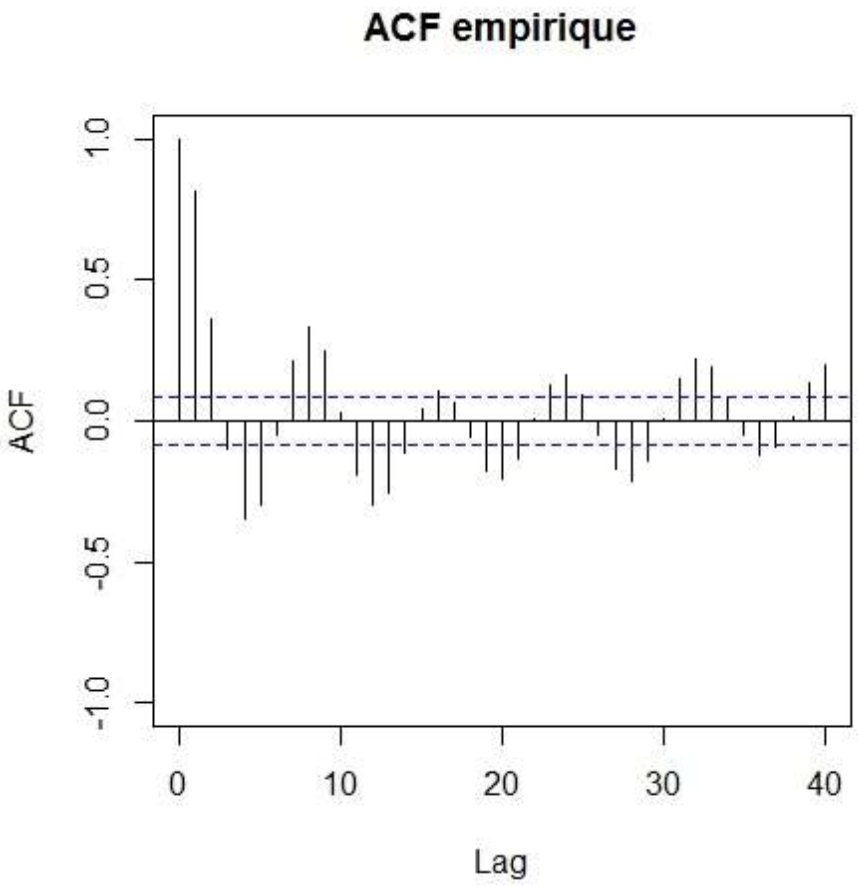
- Examen de l'ACF empirique : on considère qu'une valeur à l'intérieur des bornes $\pm \frac{1.96}{\sqrt{\mathbf{n}}}$ n'est pas significative (hypothèse nulle = WN(0, σ^2)). C'est l'examen global qui est important et peut conduire à proposer un modèle de type MA(q). Dans ce cas, les coefficients $\hat{\rho}$ (h) pour h > q sont asymptotiquement i.i.d. N(0, $\frac{1}{n}$ [1 + 2($\rho(1)^2$ + … + $\rho(q)^2$)]). Une décroissance d'aspect « exponentiel » signale la présence d'une partie autorégressive…

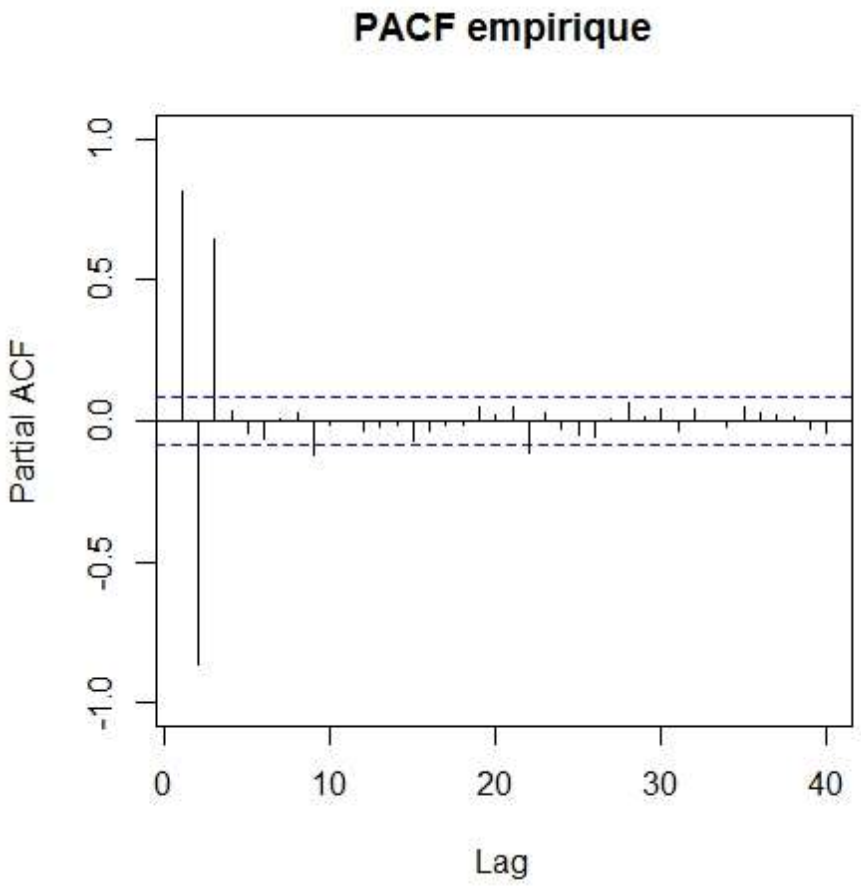
- Examen de la PACF empirique : on considère encore qu'une valeur à l'intérieur des bornes $\pm \frac{1.96}{\sqrt{\mathbf{n}}}$ n'est pas significative. Dans le cas d'un modèle AR(p), les coefficients $\hat{\pi}$ (h) pour h > p sont à peu près i.i.d. N(0, $\frac{1}{n}$). L'examen global de la PACF peut conduire cette fois à la considération d'un modèle de type AR(p)…


ACF empirique
ACF
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
Lag

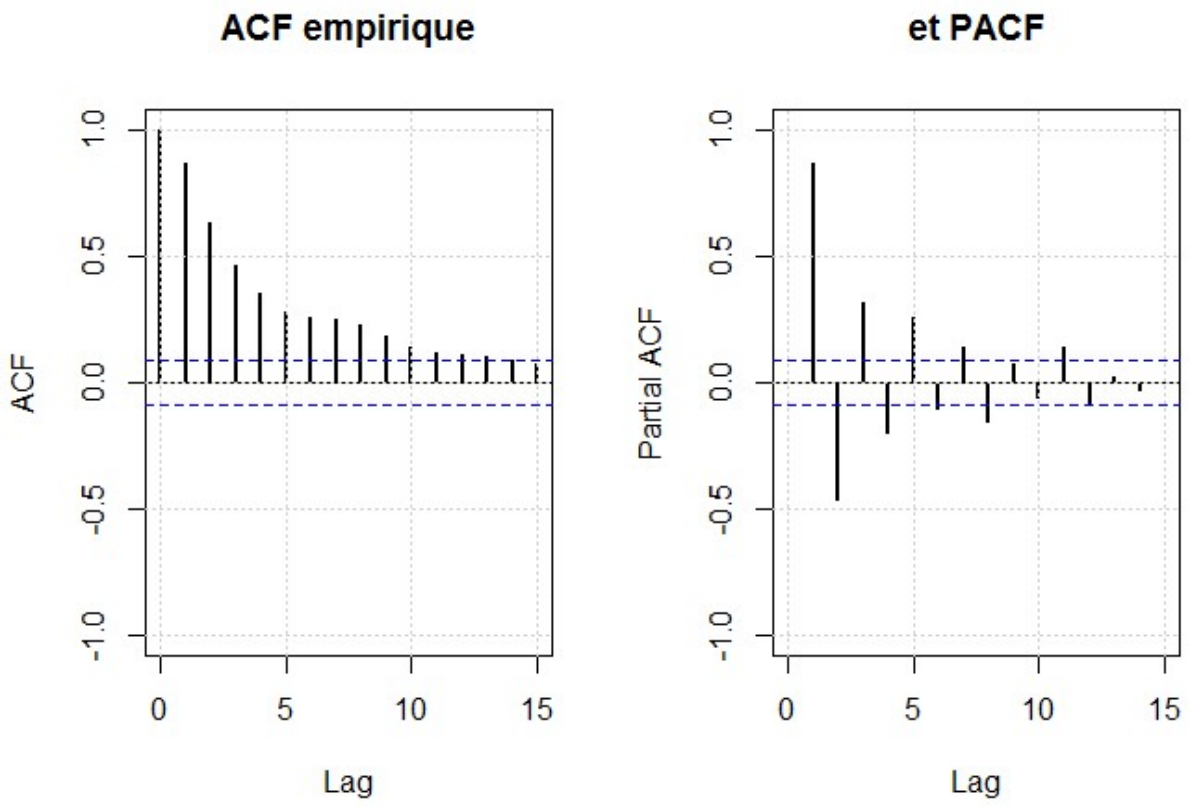

PACF empirique
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
Lag

ARMA ?


ACF empirique
ACF
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
Lag


PACF empirique
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
Lag

ARMA ?


ACF empirique
et PACF
ACF
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
Lag
Lag

ARMA ?

Extensions des modèles ARMA :

- $X \sim ARIMA(p, \mathbf{d}, q)$ si $Y = (I - B)^{\mathbf{d}} X_t \sim ARMA(p, q)$

  X ARIMA(p, d, q) de moyenne μ si $X - \mu \sim ARIMA(p, d, q)$

- Modèle SARIMA est un modèle ARIMA saisonnier (voir support page 28)

  Exemple 1 du support : série de trafic aérien

  $\mathbf{Y_t = (I - B^{12})(I - B)X_t}$ série différenciée et désaisonnalisée considérée stationnaire

  Examen ACF et PACF suggère un modèle ARMA saisonnier pour Y_t :

$$Y_t = (I + \theta_{12}B^{12})(I + \theta_1 B)Z_t \text{ avec } Z_t \sim WN(0, \sigma^2)$$

  Notation : $SARIMA(p=0, d=1, q=1)(P=0, D=1, Q=1)_{s=12}$

Prévision optimal à un pas de temps : on dispose d'un historique $x_1, \ldots, x_n$ de taille n d'une série temporelle modélisée par un processus X ~ ARMA(p, q)

$$\hat{X}_{n+1} = E_L(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1) \qquad \textbf{prédicteur} \text{ (v.a.)}$$

$$\hat{x}_{n+1} = E_L(X_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \qquad \textbf{prévision} \text{ (ponctuelle)}$$

Erreur de prévision à un pas de temps : $v_n = E(X_{n+1} - \hat{X}_{n+1})^2$

Le calcul de $\hat{X}_{n+1}$ peut se faire simplement par résolution d'un système linéaire mais on préfère des algorithmes de mise à jour beaucoup plus efficaces.

Pour un AR(1), c'est très simple puisque $\hat{X}_{n+1} = \phi_1 X_n$ et $v_n = \sigma^2$ (variance de Z_{n+1}).

Prévision à plusieurs pas de temps : on note h l'horizon de prévision

$$\hat{X}_{n,\,h} = E_L(X_{n+h} \mid X_n, X_{n-1}, \ldots, X_1) \qquad \text{prédicteur optimal}$$

$$v(n, h) = E(X_{n+h} - \hat{X}_{n,\,h})^2 \qquad \text{erreur de prévision}$$

Attention au support (chapitre 5) qui considère la situation « asymptotique » (valable si l'historique est de taille suffisamment grande)

$$\hat{X}(n, h) = E_L(X_{n+h} \mid X_n, X_{n-1}, \ldots, X_1, \mathbf{X_0, X_{-1}, \ldots})$$

$$e(n, h) = E(X_{n+h} - \hat{X}(n, h))^2$$

Cas d'un AR(p) « pur » : $e(n, h) = \sigma^2(1 + \psi_1^2 + \ldots + \psi_{h-1}^2) \rightarrow \sigma_X^2$ lorsque $h \uparrow +\infty$

Cas d'un MA(q) « pur » : $e(n, h) = \sigma^2(1 + \theta_1^2 + \ldots + \theta_q^2) = \sigma_X^2$ si $h \geq q+1$

(cf. exercices 1 et 2 du support, chapitre 5, page 40)

Attention : on dispose ici de données uniquement pour la série temporelle d'intérêt (série univariée).

Un **chronogramme** (simple tracé de la série de valeurs x_t en fonction du « temps » t) permet de visualiser la série, de détecter **tendance** et **saisonnalité** éventuelles.

On cherche ensuite à se ramener à une **série stationnaire** (transformations, estimation des composantes déterministes, différenciation simple ou saisonnière).

On trace les **ACF** et **PACF empiriques** de la série « stationnarisée » pour voir si un modèle de type ARMA peut être raisonnablement envisagé.

La *méthodologie de Box et Jenkins* s'inscrit dans cette démarche (support, chapitre 4, page 29)!

En pratique, il est souvent difficile d'identifier directement un « bon » modèle ARMA à partir des seules ACF et PACF empiriques, ce qui pose le problème du choix des ordres p et q.

On utilise alors un critère, par exemple :

- AIC pour Aikaike's Information Criterion
- SBC Schwartz Bayesian Criterion

L'idée générale est comme en régression de **pénaliser** la (fonction de) vraisemblance des données en fonction du nombre de paramètres (p+q+1) du modèle sachant que l'on préfère toujours un modèle avec le moins de paramètres possibles…

☞ Les propriétés de l'estimateur du maximum de vraisemblance sont données support de cours pages 32 et 33. L'analyse se fait exactement comme en régression.

Phase de validation : il s'agit d'analyser les résidus « estimés » du modèle (le processus de bruit $\{Z_t\}$).

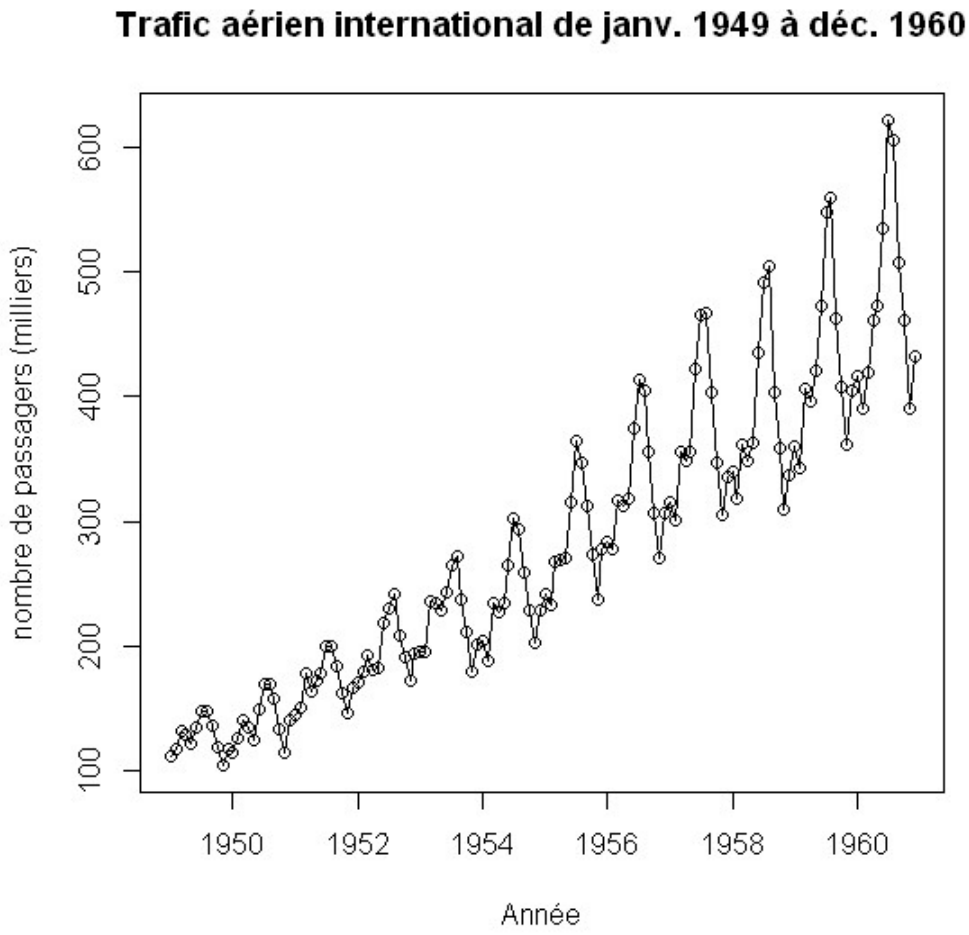
Vérifications graphiques : ACF et PACF des résidus mais aussi du carré des résidus pour tester une éventuelle dépendance synonyme de modèle non gaussien. On peut aussi tracer la droite de Henry pour tester la normalité des résidus (qqplot).

Test statistique de Ljung-Box : la statistique est la suivante (support page 35)

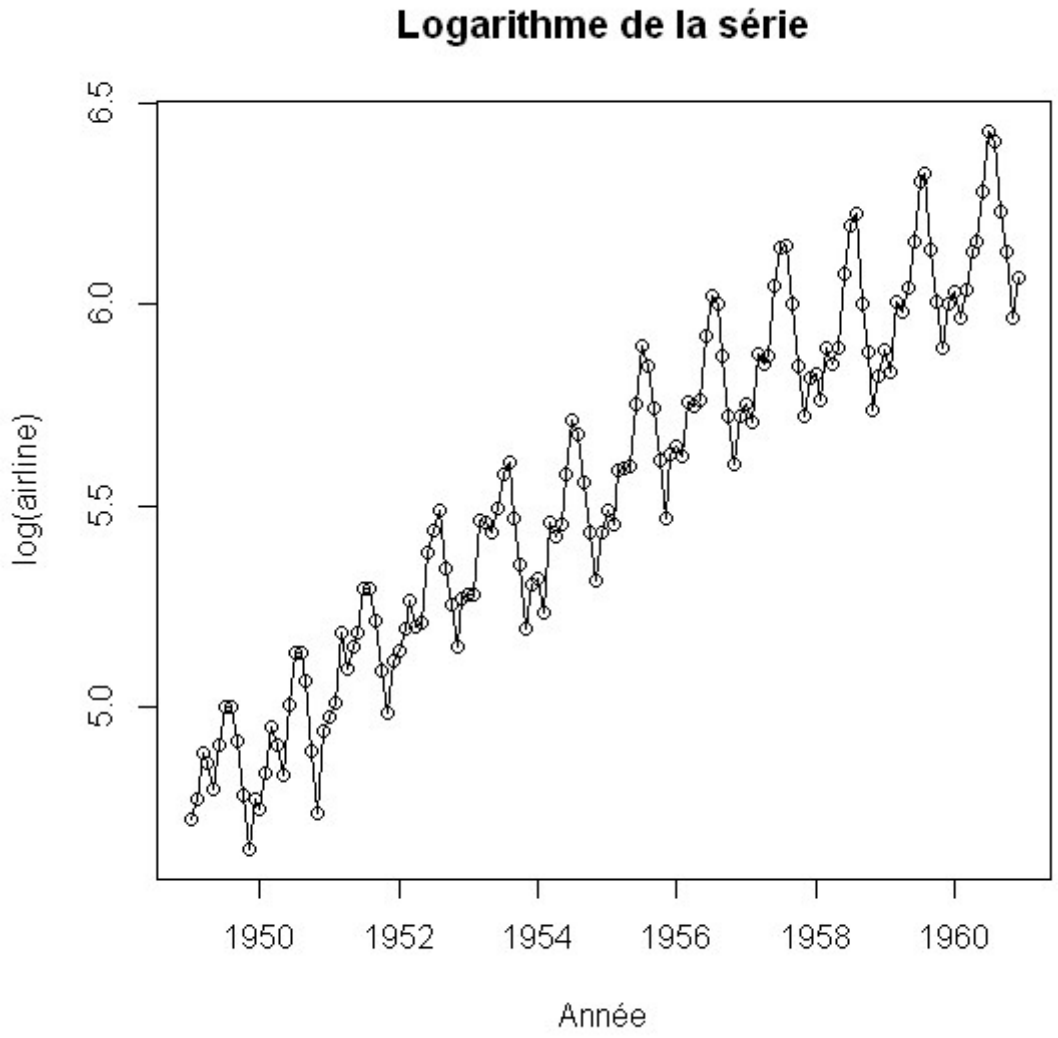
$$Q_Z = n(n+2) \sum_{j=1}^{h} \frac{\hat{\rho}_Z^2(j)}{n-j}$$

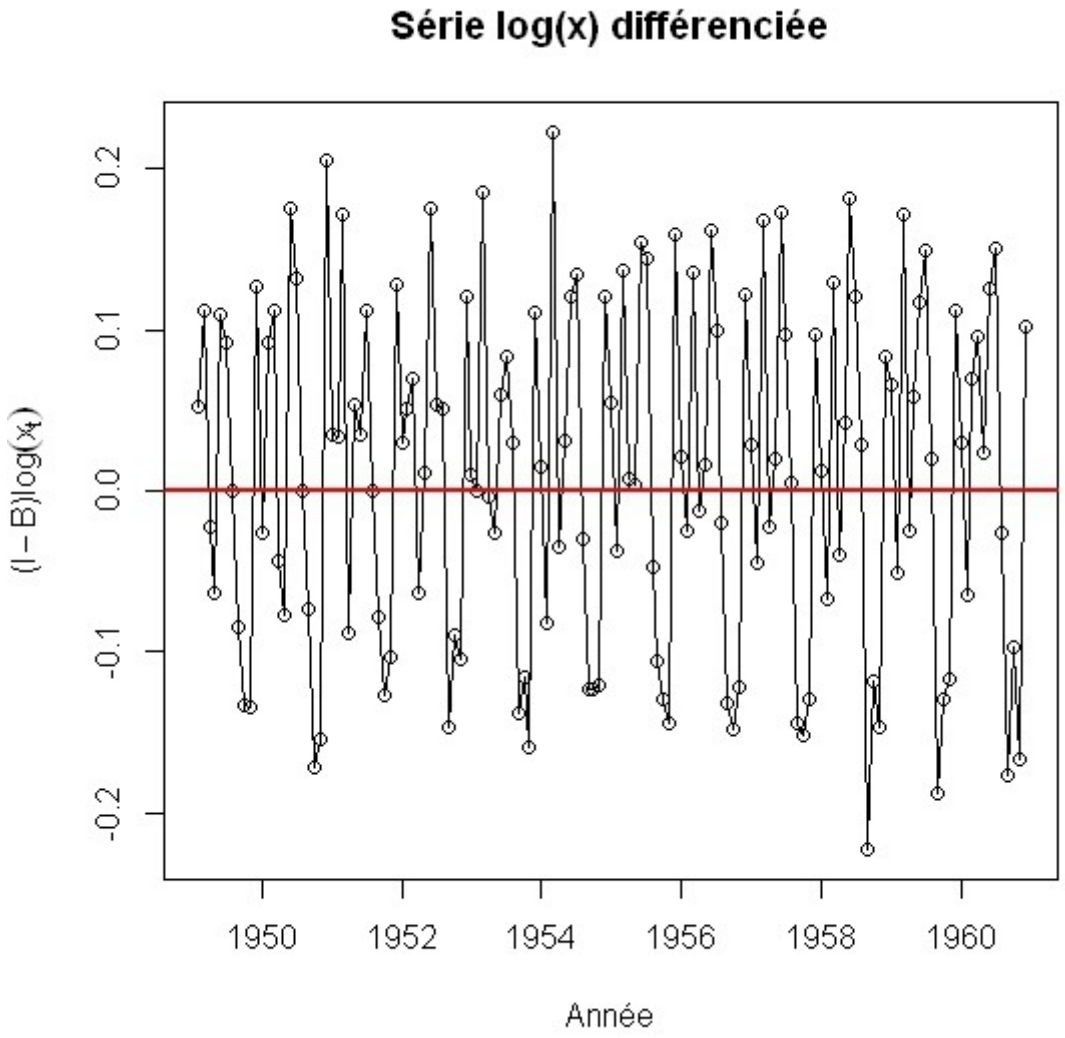
Sous l'hypothèse H_0 d'un modèle ARMA(p, q), Q_Z est approximativement de loi $\chi^2(h-(p+q))$…

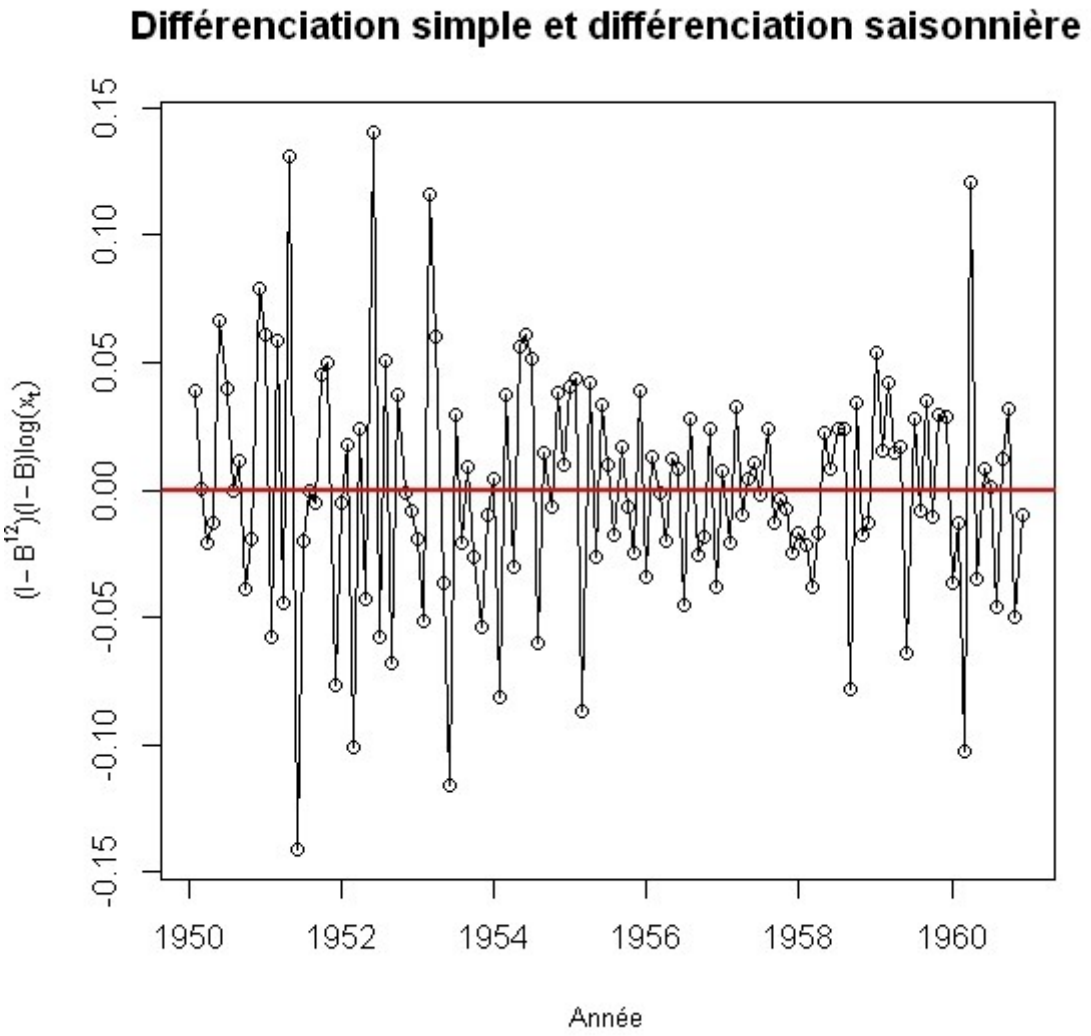
Un exemple : série de trafic aérien internationale (airline, ex.1 du support)



Méthodologie de Box&Jenkins : transformation « logarithme »




Série log(x) différenciée
(I − B)log(x_t)
0.2
0.1
0.0
-0.1
-0.2
1950
1952
1954
1956
1958
1960
Année


Différenciation simple et différenciation saisonnière
$(1-B^{12})(1-B)\log(x_t)$
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
1950
1952
1954
1956
1958
1960
Année

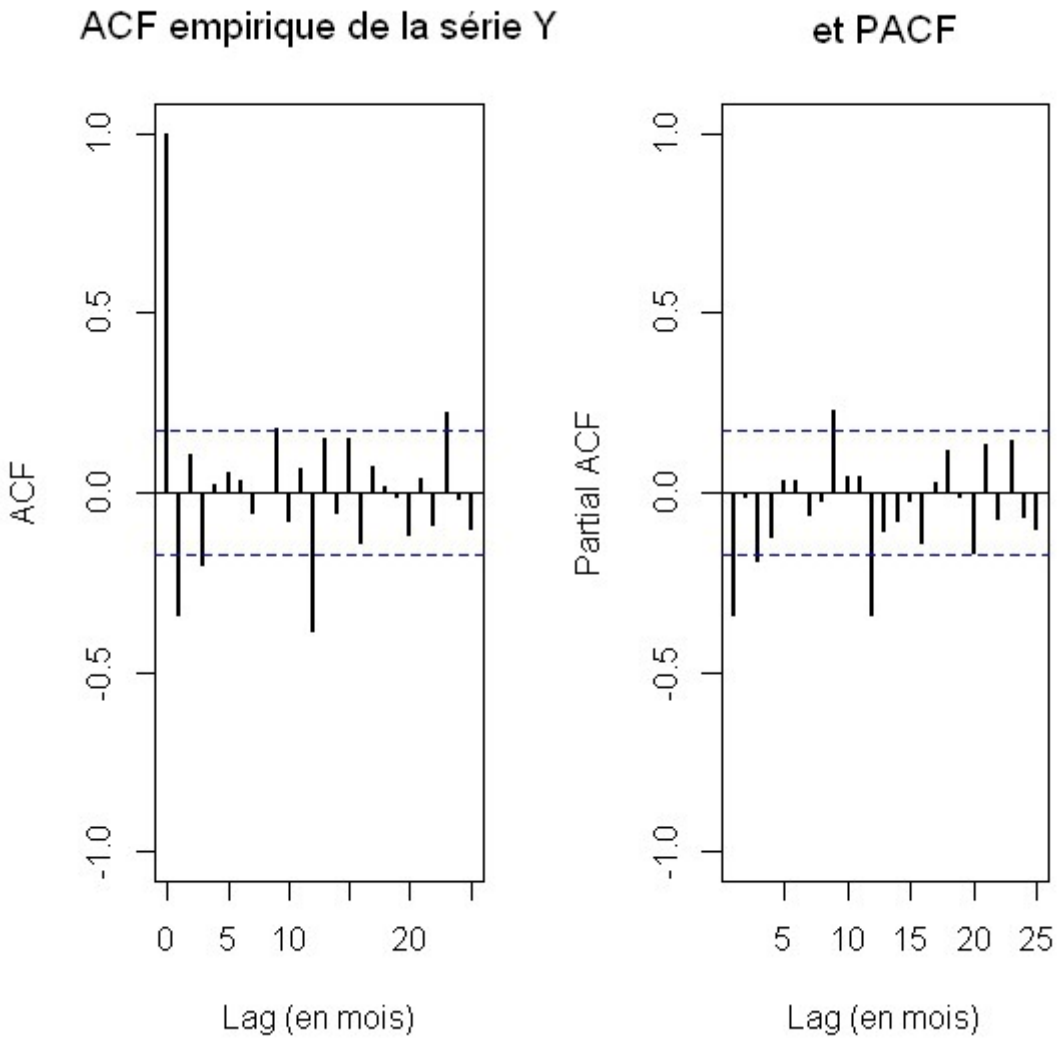
Fin étape de stationnarisation :

on considère que la série

$$Y_t = (I - B^{12})(I - B)\log(X_t)$$

est stationnaire (hypothèse de travail).

On examine maintenant l'ACF et PACF empiriques de Y avec l'idée d'en rendre compte par un processus ARMA saisonnier.


ACF empirique de la série Y
et PACF
ACF
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
20
5
10
15
20
25
Lag (en mois)
Lag (en mois)

Si l'on ignore le coefficient d'autocorrélation d'ordre h = 12 correspondant à la saisonnalité, on est amené à proposer un modèle de type MA(q=1)

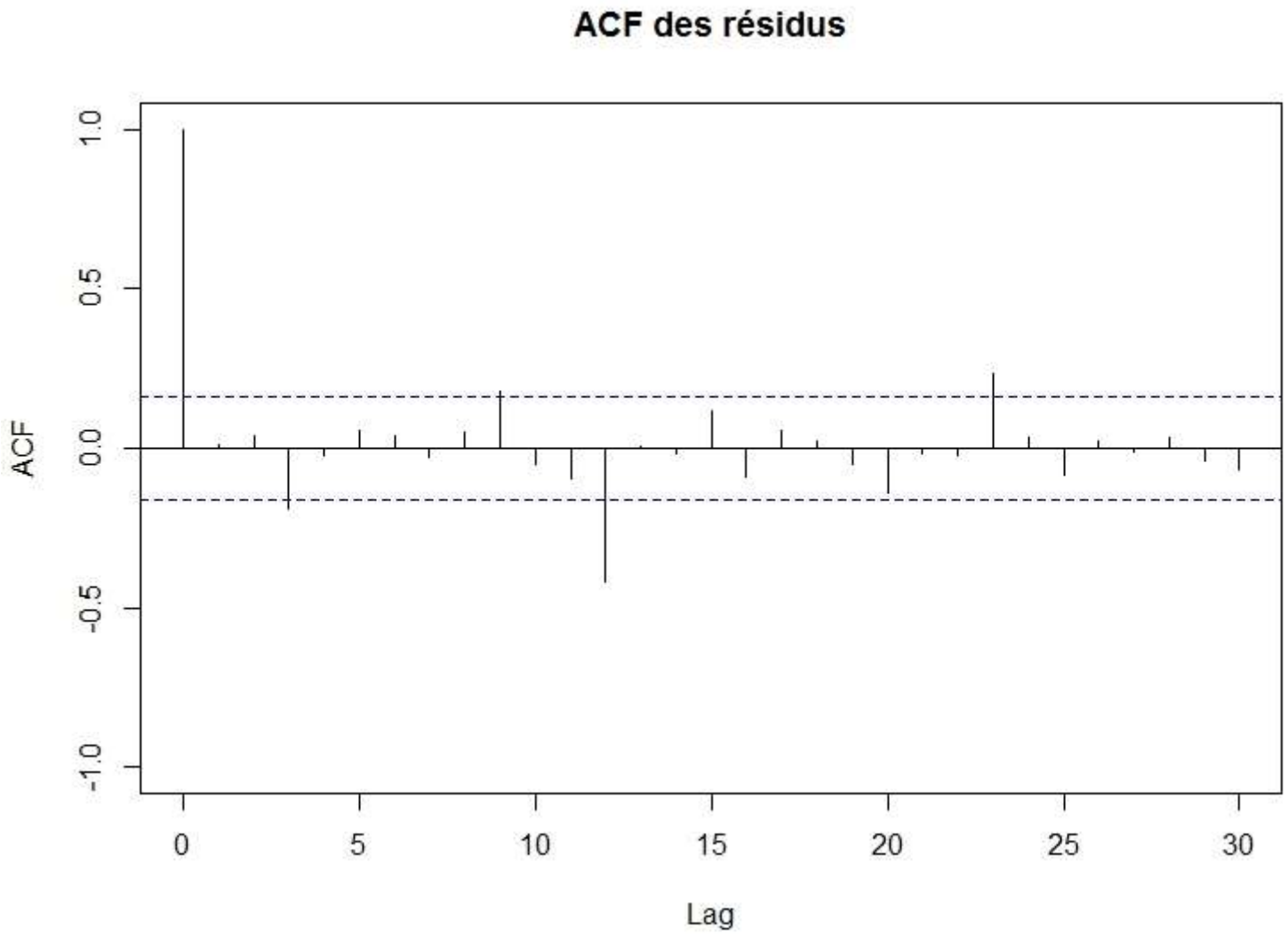
$$Y_t = (I + \theta_1 B)W_t$$

de manière à capter la dépendance sur les premiers mois.

Comme la série présente une forte saisonnalité, il est logique de s'attendre à ce que le « résidu » $\{W_t\}$ ait une structure de corrélation saisonnière.

A l'aide de la fonction **arima()**, on cale ce premier modèle de la forme MA(1) puis on trace l'ACF des « résidus » obtenus :

```
modele <- arima(logair, order = c(0,1,1), seasonal = list(order=c(0,1,0), period = 12))
residus <- modele$residuals
acf(as.vector(residus),lag.max=30,ylim=c(-1,1),main="ACF des résidus")
```


ACF des résidus
ACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
25
30
Lag

On observe bien la dépendance saisonnière, soit la « non blancheur » du processus $\{W_t\}$. Il est naturel d'essayer d'en tenir compte en considérant que l'unité de temps est la période s = 12, ce qui conduit à considérer le modèle

$$W_t = (I + \theta_{12}B^{12})Z_t \qquad \text{avec } Z \text{ bruit blanc}$$

En revenant au processus $Y_t = (I + \theta_1 B)W_t$, on obtient le modèle

$$Y_t = (I + \theta_1 B)(I + \theta_{12}B^{12})Z_t \text{ avec } \{Z_t\} \sim WN(0,\sigma^2)$$

Il s'agit d'un modèle de type **$SARMA(p=0, q=1)(P=0, Q=1)_{s=12}$**

En revenant à la série initiale $\{X_t\}$

$$(I - B^{12})(I - B)\log(X_t) = (I + \theta_1 B)(I + \theta_{12}B^{12})Z_t$$

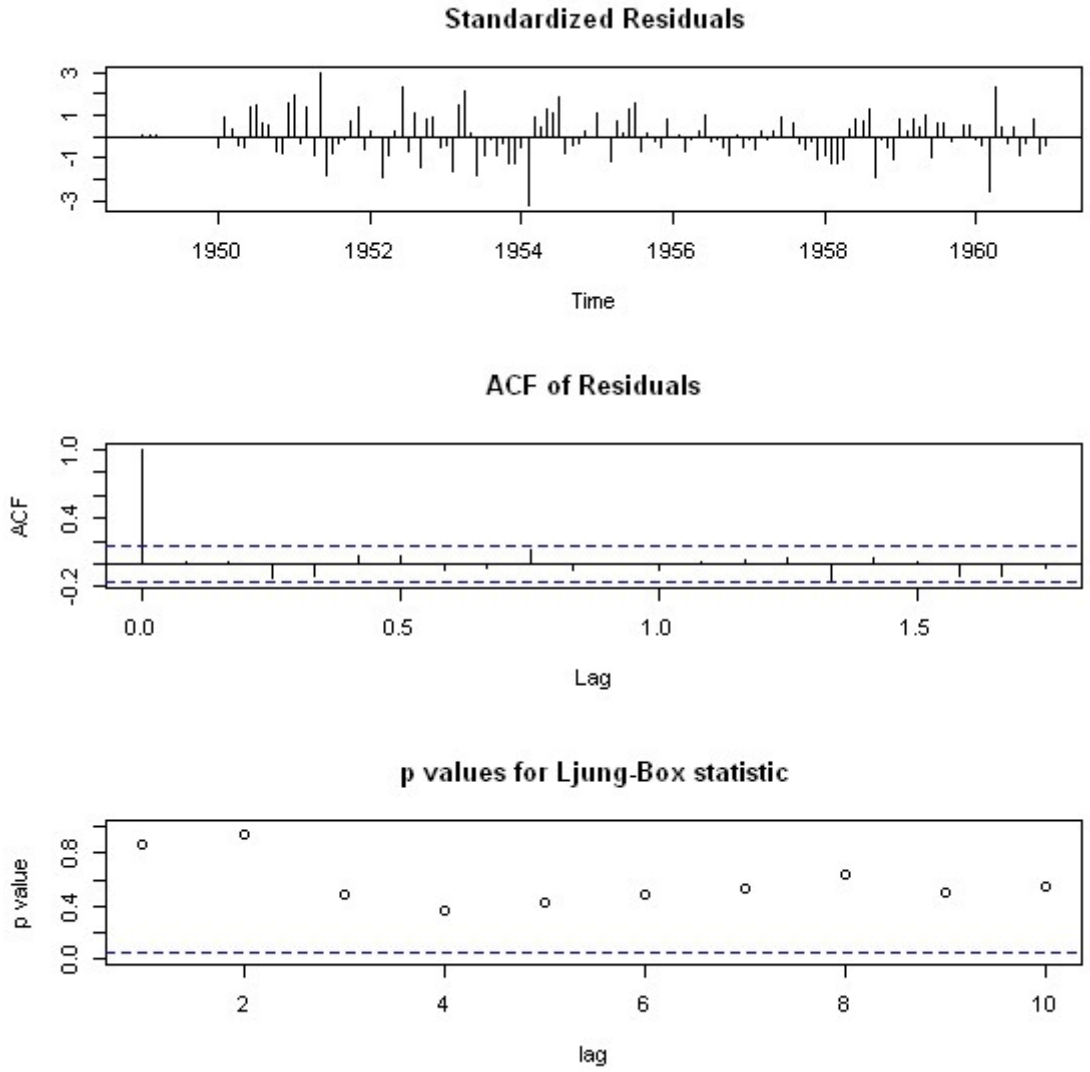
En abrégé : **$\log(X_t) \sim SARIMA(p,d,q)(P,D,Q)_s = SARIMA(0,1,1)(0,1,1)_{s=12}$**

Ajustement du modèle avec la fonction R « arima » :

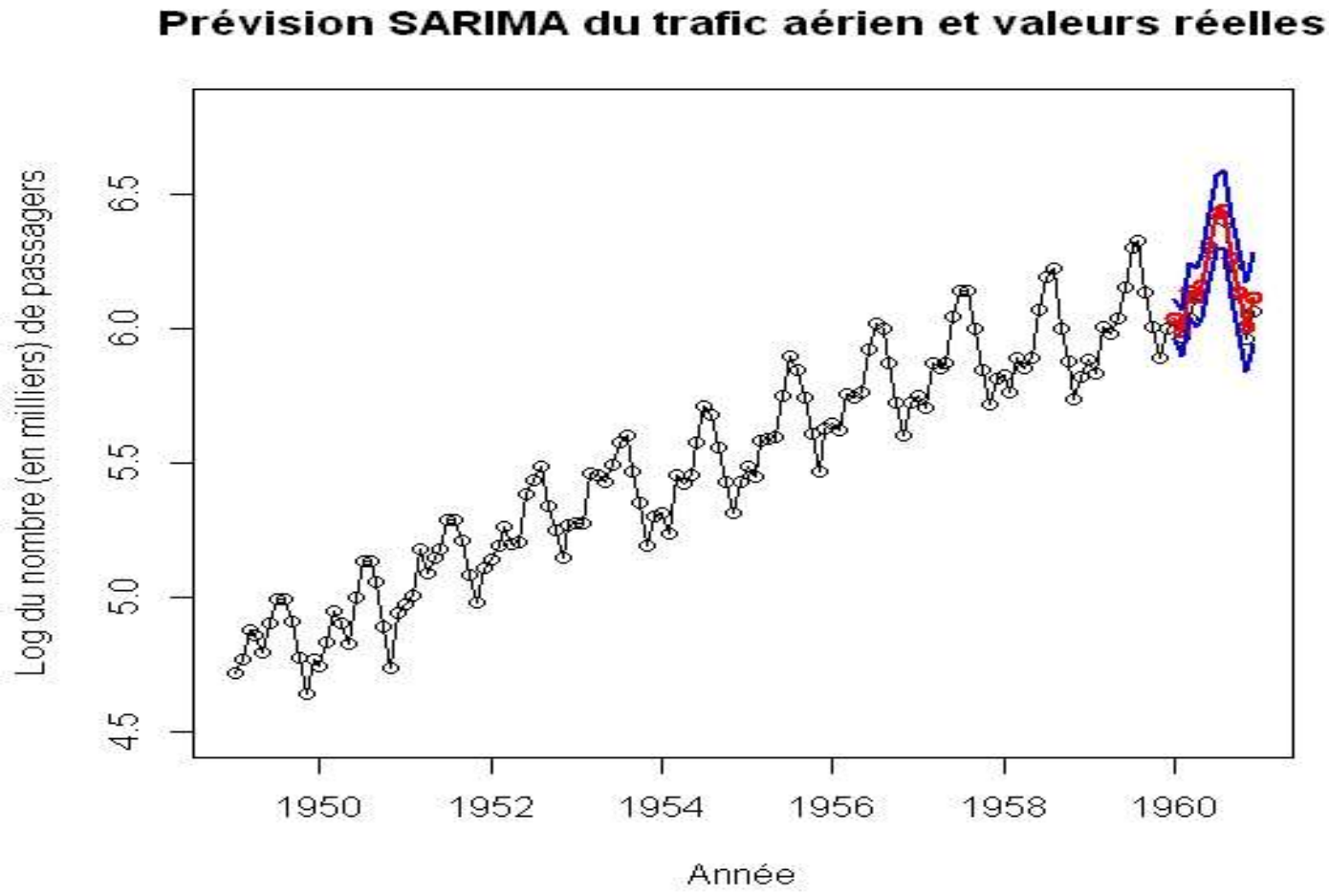
```
> modele <- arima(logair, order = c(0,1,1), seasonal = list(order=c(0,1,1), period = 12))
> print(modele)
```

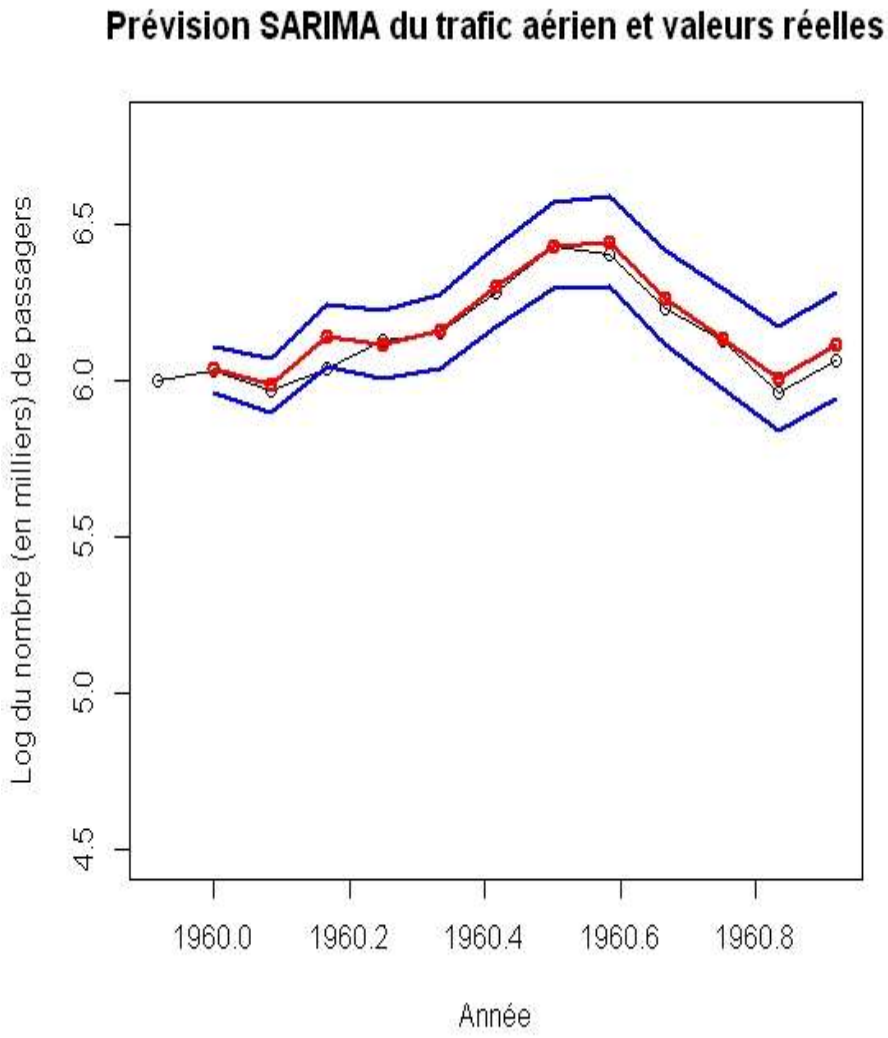
```
Call:
arima(x = logair, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = -483.4
```

On valide ensuite avec la fonction « tsdiag » pour *time series diagnostics*.


Standardized Residuals
-3
-1
1
3
1950
1952
1954
1956
1958
1960
Time
ACF of Residuals
ACF
-0.2
0.4
1.0
0.0
0.5
1.0
1.5
Lag
p values for Ljung-Box statistic
p value
0.0
0.4
0.8
2
4
6
8
10
lag

Back-testing du modèle SARIMA (on enlève les 12 dernières valeurs que l'on cherche à prédire et on compare aux valeurs réelles)





Prévision et valeurs réelles de janvier 60 à décembre 60