Introduction au cours de séries temporelles

Angelina Roche

Executive Master Statistique et Big Data

2018–2019

Plan du cours d'aujourd'hui

Qu'est-ce qu'une série temporelle ? Exemples

Représentations graphiques

Tendance, saisonnalité, résidus

Remarques complémentaires et déroulement du cours

Plan

Définition d'une série temporelle

Définitions

- Une série temporelle est un ensemble d'observations $X_{t_1}, X_{t_2}, \ldots$ mesurées sur des temps $t_1, t_2, \ldots$ distincts.
- Un processus stochastique $\{X_t, t \in T\}$ est une famille de variables aléatoires définies sur le même espace de probabilité.

On modélisera généralement un série temporelle comme un processus stochastique.

Exemples

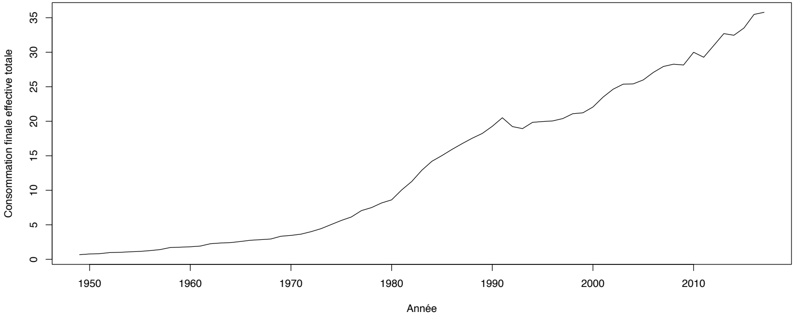


Figure – Évolution de la consommation effective totale annuelle (en milliards d'euros) en France de 1949 à 2017 (cf TP 2 d'analyse de données).

Exemples

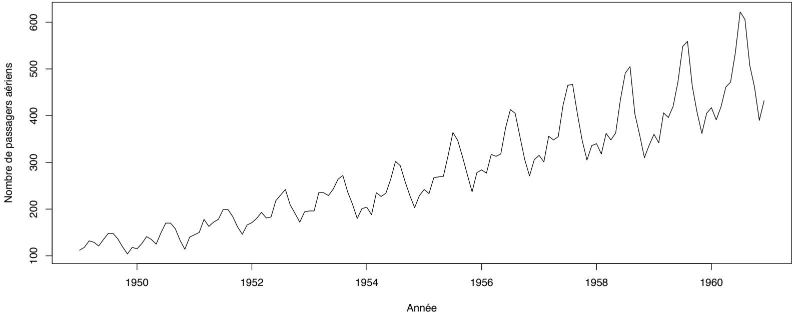


Figure – Nombre mensuel de passagers aériens sur des vols internationaux en milliers de 1949 à 1960.

Exemples

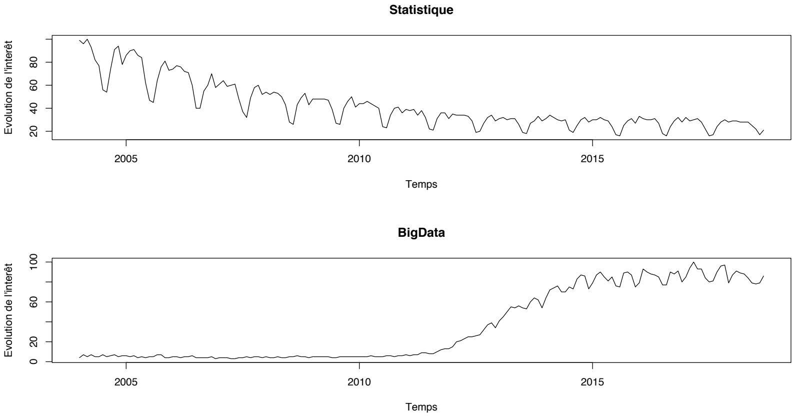


Figure – Évolution de l'intérêt pour les mot-clefs 'Statistique' et 'Big Data' (source : Google trends).

Plan

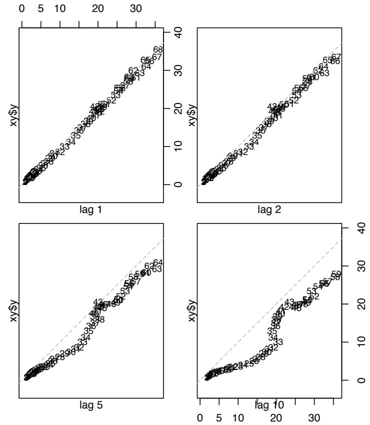
Représentations graphiques usuelles

- Le chronogramme : tracé du graphique $\{(t, X_t), t \in T\}$ où T est l'intervalle d'observation.
- Le lag-plot ou diagramme retardé : tracé des points $\{(X_{t-k}, X_t), t, t-k \in T\}$.

  Objectif : détecter des corrélations pour comprendre la dépendance de la série envers son passé.
- Le month-plot : représentation simultanée des chronogrammes associés à chaque saison ou mois.

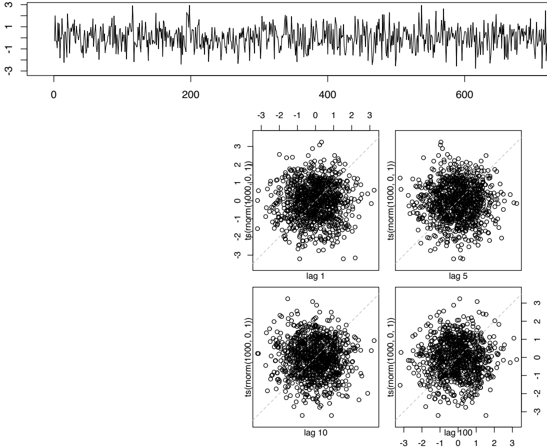
Example : lag-plot de la consommation effective annuelle

```
Total.ts <- ts(t(Total),start=1949)
lag.plot(Total.ts,set.lags=c(1,2,5,10),layout=c(2,2))
```



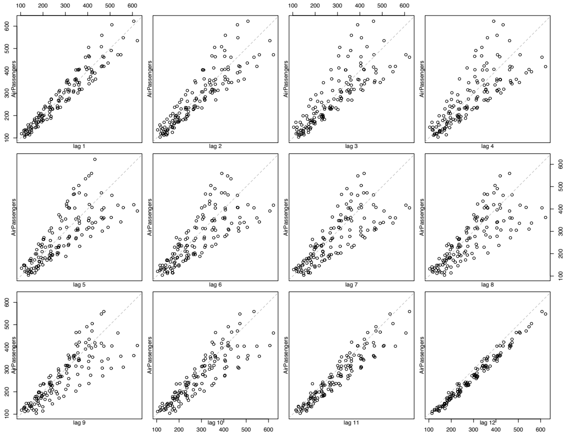
Example : lag-plot du bruit blanc (pas de corrélation)

```
lag.plot(ts(rnorm(1000,0,1)),set.lags=c(1,5,10,100),layout=c(2,2),do.lines=FALSE)
```



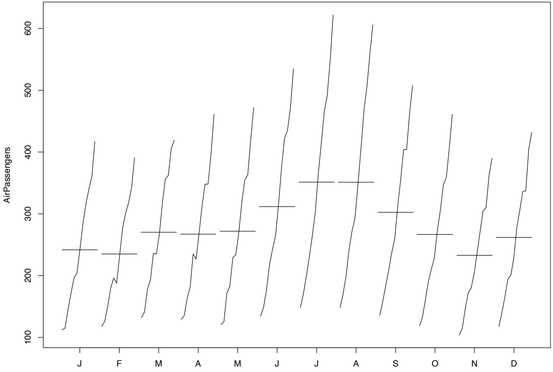
Example : lag-plot du nombre de passagers aériens

lag.plot(AirPassengers,lags=12,layout=c(3,4),do.lines=FALSE)



Example : month-plot du nombre de passagers aériens

monthplot(AirPassengers)



Plan

Tendance, saisonnalité, résidus

On décompose souvent une série temporelle $\{X_t, t \in T\}$ de manière additive

$$X_t = m_t + s_t + Z_t,$$

ou multiplicative

$$X_t = m_t s_t (1 + Z_t).$$

avec

- $t \mapsto m_t$: la tendance ou *trend* fonction qui varie peu (capte l'orientation à long terme),
- $t \mapsto s_t$: la composante saisonnière fonction périodique (comportement qui se répète),
- $\{Z_t, t \in T\}$: la composante d'erreur ou résidu, de moyenne nulle et idéalement de faible variabilité par rapport aux deux autres.

Des combinaisons des modèles additif et multiplicatif sont aussi envisageables. Par exemple : $X_t = (m_t + s_t) Z_t$.

Estimation de la tendance par moindres carrés

- Objectif : on suppose que $X_t = m_t + Z_t$ avec $m_t = \beta_0^* + \beta_1^* t + ... + \beta_d^* t^d$ et l'on souhaite estimer $\beta^* = (\beta_0^*, ..., \beta_d^*)^t$.

- Idée : estimer β par moindres carrés à partir des observations $X_{t_1}, ..., X_{t_n}$.

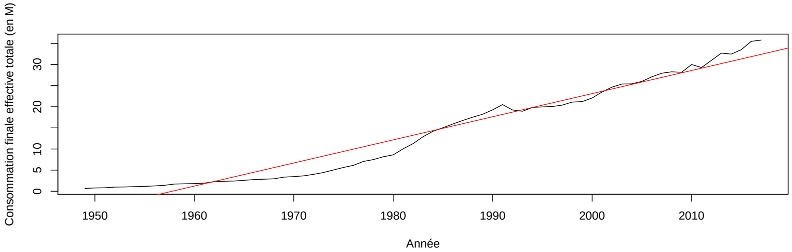
$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^{d+1}} \frac{1}{n} \sum_{i=1^n} (X_{t_i} - m_{t_i})^2 \Rightarrow \widehat{\beta} = (T^t T)^{-1} T^t X,$$

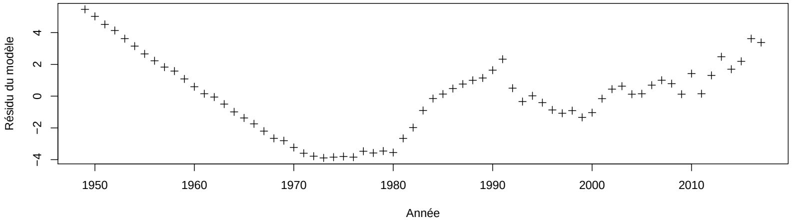
avec

$$T = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & t_n^2 & \dots & t_n^d \end{pmatrix} \text{ et } X = (X_{t_1}, ..., X_{t_n})^t.$$

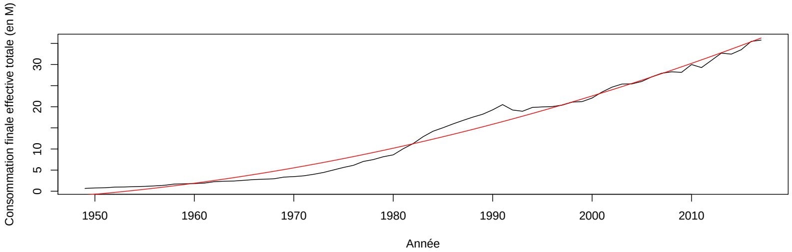
- On peut choisir d ensuite par des méthodes de choix de modèles (AIC, BIC,...) ou visuellement.

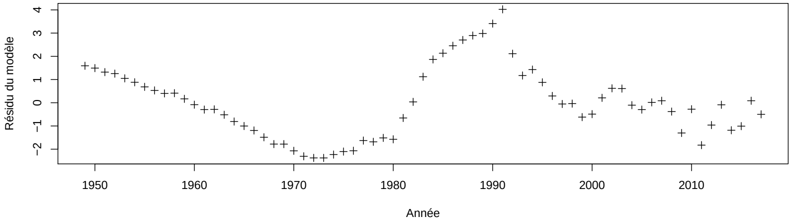
Exemple : consommation effective annuelle, d=1



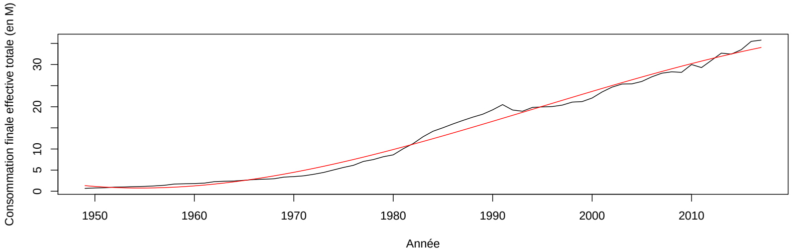


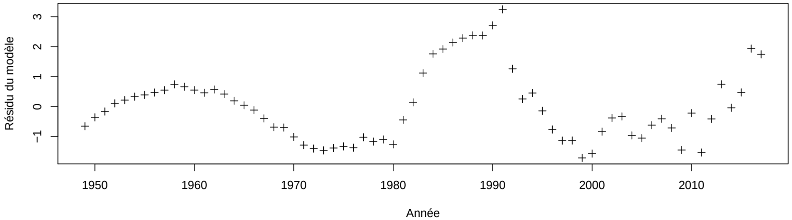
Exemple : consommation effective annuelle, d=2





Exemple : consommation effective annuelle, d=4





Autres méthodes

- Moyenne mobile :

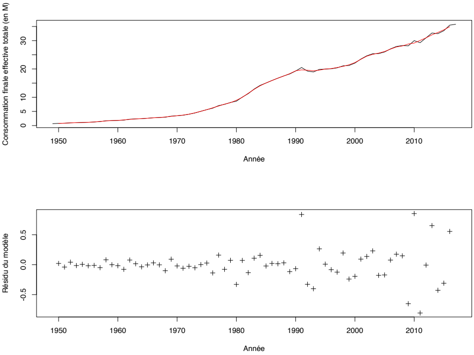
$$\widehat{m}_t = \frac{1}{2q+1} \sum_{k=-q}^{q} x_{t+k}.$$

- Autres méthodes de statistique non-paramétrique : noyau, polynômes locaux,...

Exemple : consommation effective annuelle

Estimation de la tendance par moyenne mobile ($q = 2$)

```
MA <- filter(Total.ts,filter=rep(1/3,3))
par(mfrow=c(2,1))
plot(Total.ts,xlab='Année',ylab='Consommation finale effective totale (en M)')
points(MA,col='red',type='l')
plot(Total.ts-MA,xlab='Année',ylab='Résidu du modèle',pch=3,type='p')
```



Estimation de la partie déterministe (tendance et saisonnalité)

- Il existe un grand nombre de méthodes.
- Fonctions implémentées dans `R` :
  - fonction `decompose()` : estime m_t par moyenne mobile ($\widehat{m}_t = \frac{1}{2q+1} \sum_{k=-q}^{q} x_{t+k}$) puis estime s_t à partir de $x_t - \widehat{m}_t$ (pour le modèle additif) par moyenne mobile également :

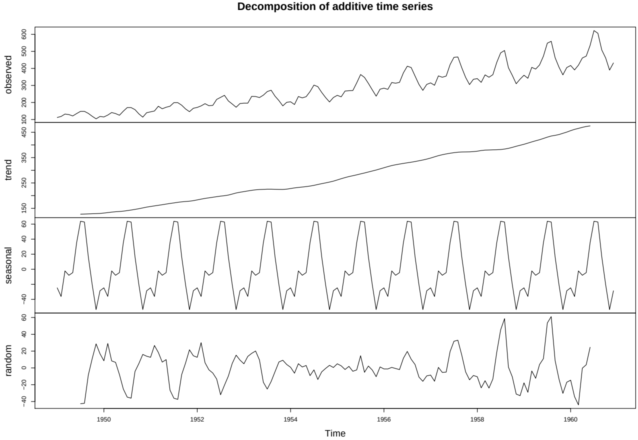
$$\widehat{s}_t = \frac{1}{k+1} \sum_{j=0}^{k} (x_{t+jT} - \widehat{m}_{t+jT}),$$

    où T est la période (estimée par la fonction) et k choisi le plus grand possible.
  - fonction `stl()` (plus élaboré) : estimation par maximum de vraisemblance dans un modèle additif.

Exemple : nombre de passagers aériens

Estimation par la fonction `decompose()` (modèle additif par défaut)

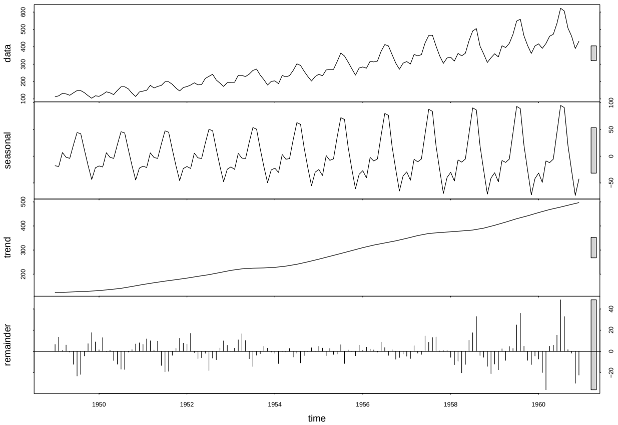
```
fit.decompose <- decompose(AirPassengers)
plot(fit.decompose)
```



Exemple : nombre de passagers aériens

Estimation par la fonction `stl()`

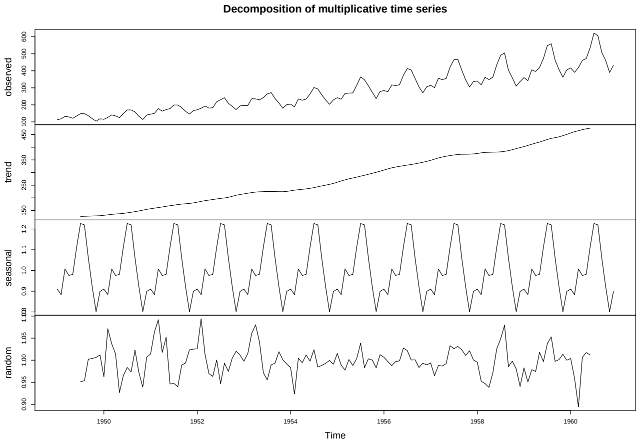
```
fit.stl <- stl(AirPassengers,s.window=12)
plot(fit.stl)
```



Exemple : nombre de passagers aériens

Estimation par la fonction decompose() (modèle multiplicatif)

```
fit.decompose2 <- decompose(AirPassengers,type='multiplicative')
plot(fit.decompose2)
```



Plan

Étapes et objectifs de l'analyse d'une série temporelle

1. Décrire, préparer :
   - Corriger la série des effets systématiques (prise en compte des jours ouvrables, grèves, pannes,...), imputation des données manquantes (fonction `na.approx()` du package `zoo` par exemple),...
   - Représentations graphiques : chronogramme, lag-plot, month-plot,... Permet de détecter les valeurs atypiques, anomalies,... que l'on peut traiter éventuellement comme des données manquantes.
2. Modéliser : expliquer la valeur en un instant par des modèles ayant peu de paramètres.
3. Prévoir : à partir d'un modèle ou construites à partir des données (lissage exponentiel par exemple).

Remarques

- Le choix d'un modèle, l'incorporation d'une composante peuvent s'apprécier d'après le graphique et tirer leur validation de l'analyse.

- Une même série peut-être analysée de différentes façons. Rien n'interdit de superposer plusieurs approches car
  - On étudie une seule réalisation d'un même processus sur un intervalle de temps fini (caractéristiques à long terme éventuellement non visible).
  - Une série n'est pas nécessairement modélisée par un modèle de notre catalogue.

    Box (1978) "All models are wrong but some are useful."

Déroulement du cours

- Cours 1 (20 septembre) : introduction, représentations graphiques, tendance, saisonnalité.
- Cours 2 (18 octobre) : processus stationnaires, modèles ARMA.
- Cours 3 (15 novembre) : lissage exponentiel.
- Cours 4 (13 décembre) : modèles non stationnaires (modèles ARIMA et SARIMA, tests de non-stationnarité).

Quelques références

- Aragon, Y. *Séries temporelles avec R*.
- Site web de Arthur Charpentier : `freakonometrics.hypotheses.org`.
- Brockwell, P. et Davis, R. *Time Series : Theory and Methods*.
- Brockwell, P. et Davis, R. *Introduction to time series and forecasting*.
- Box, G. et Jenkins, G. *Time Series analysis – forecasting and control*.