


Module MP(C)

Méthodes de prédiction des variables numériques

Simon Malinowski

M1 Miage - M1 Data Science, Univ. Rennes

Horaire de ce cours

- ① Partie 1 : Prédiction des séries chronologiques
 - ▶ 4h30 CM
 - ▶ 10h30 TD/TP (incluant un travail noté)
- ② Partie 2 : Prédiction à l'aide d'autres variables (modèles de régression)
 - ▶ 4h30 CM
 - ▶ 6h TD/TP
 - ▶ Projet de 4h30 (noté)
 - Concours Kaggle : 

Prédiction des séries chronologiques

Simon Malinowski

M1 Miage, Univ. Rennes 1

1 Introduction : contexte, définitions

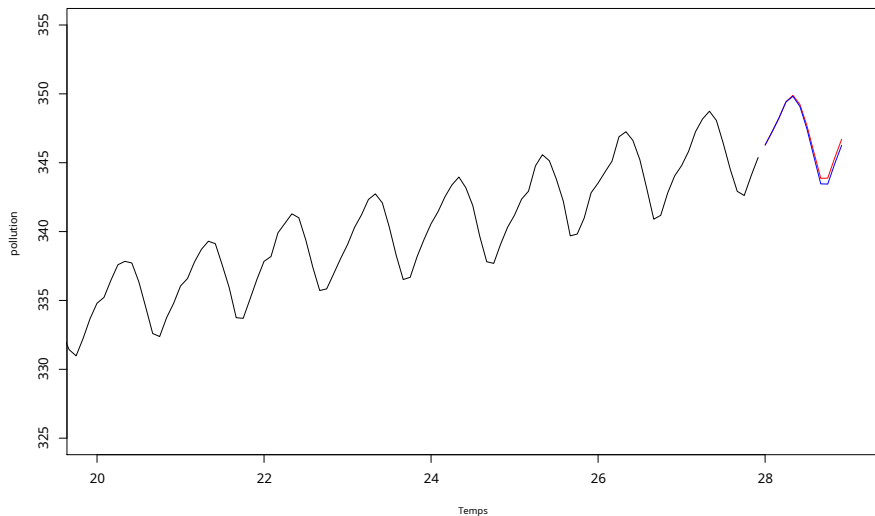
2 Décomposition des séries chronologiques

3 Prédiction des séries chronologiques

Exemples de séries chronologiques

- consommation d'électricité des utilisateurs chaque minute ou heure
- nombre de personnes dans les transports publics chaque jour
- température moyenne quotidienne dans une ville donnée

Notre objectif : prédire la ou les valeurs suivantes



Qu'est-ce qu'une série chronologique ?

Définition

Une série chronologique est une série finie d'entiers de points de données qui représentent l'évolution d'une certaine quantité sur des horodatages réguliers

Parfois aussi appelée série chronologique

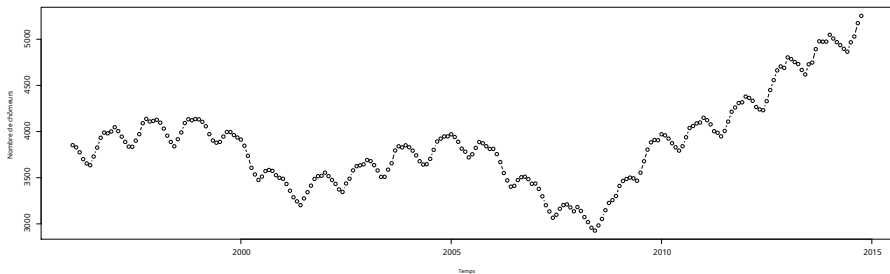
Désigné par $X(t) = x_1, x_2, \dots, x_n$, ou aussi $X(t) = x(1), x(2), \dots, x(n)$

Graphical representations

1) R global representation : les points (t, x_t) sont représentés sur un tracé

Exemple 1 : nombre de chômeurs inscrits à Pôle Emploi chaque mois

La veille de la France de 2001 à 2014

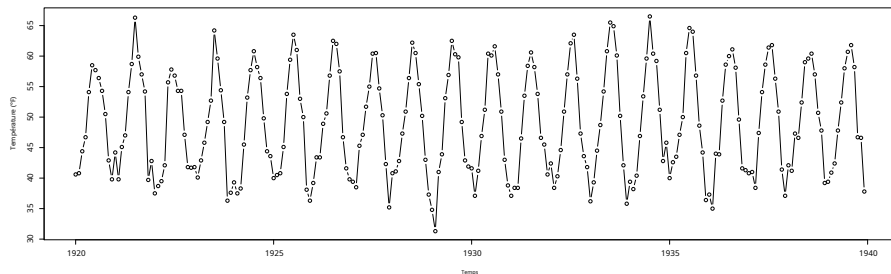


Graphical representation

Représentant mondial : les points (t, x_t) sont représentés sur un

Exemple 2 graphique : Température moyenne mensuelle à Nottingham

Période 1920-1940

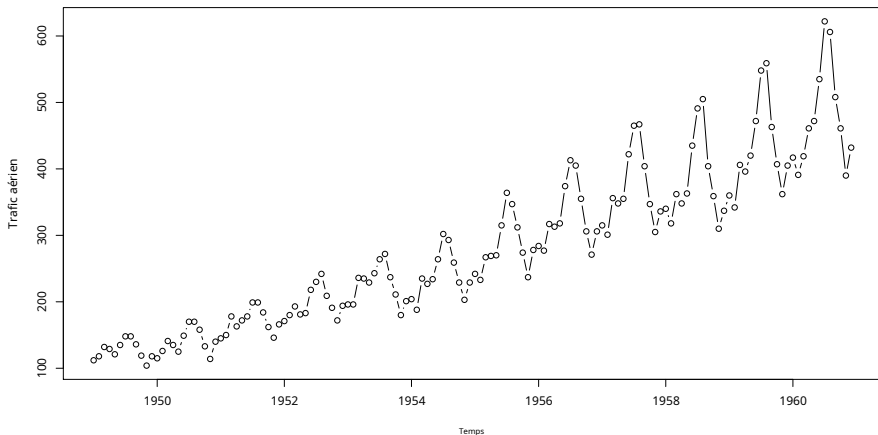


Graphical representations

Représentant mondial présentation : les points (t, x_t) sont représentés sur un graphique :

Exemple 3 Trafic aérien mensuel moyen

Période 1949-1961



1 Introduction : contexte, définitions

2 Décomposition des séries chronologiques

3 Prédiction des séries chronologiques

Analyser et comprendre une série chronologique

Les séries chronologiques représentent souvent un phénomène complexe, difficile à analyser de manière simple.

Les méthodes que nous utiliserons pour prédire une série temporelle sont basées sur la décomposition d'une série temporelle en éléments plus simples à

- manipuler avec
- contrôle

Les éléments pris en compte sont :

- 1 la tendance T_t
- 2 la composante saisonnière S_t
- 3 la composante résiduelle E_t

Modèles de décomposition

Modèle additif

Nous supposons que X_t peut être écrit comme

$$X_t = T_t + S_t + E_t$$

Modèle multiplicatif

Nous supposons que X_t peut être écrit comme

$$X_t = (T_t * S_t) + E_t$$

$$X_t = T_t * S_t * E_t$$

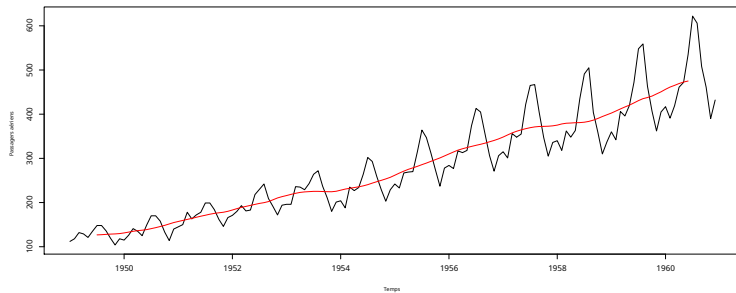
Définition

Ce composant représente l'évolution globale du phénomène considéré

Hypothèses sur ce composant

- variant lentement
- déterministe
- peut être estimée comme une fonction mathématique (simple)

Tendance ou pas tendance ?



Types de tendances

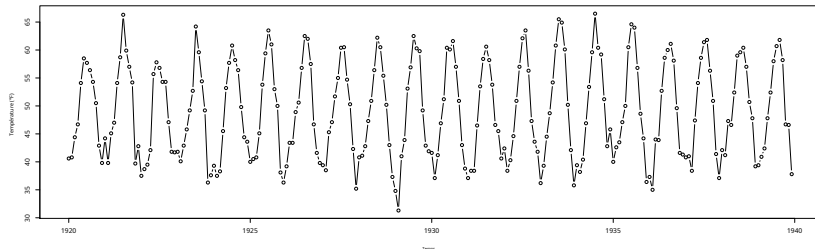
- 1 Linéaire: $T_t = u + b \times t$
- 2 Polynôme : $T_t = u_0 + u_1 \times t + \dots + u_n \times t^n$
- 3 Logarithmique : $T_t = u + b \times \ln t$
- 4 Exponentiel : $T_t = u \times \exp(bt)$

Les mers composant onal

Définition

La saisonnalité La composante al représente les fluctuations périodiques autour de la moyenne (de λ à l'intérieur d'un p longueur p) et qui se produisent presque de manière identique à partir de la période à la période

Exemple : Température à Nottingham

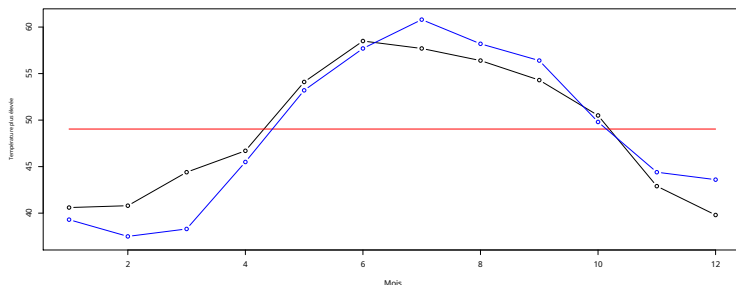


Les mers composant onal

Définition

La saisonnalité La composante al représente les fluctuations périodiques autour de la moyenne (de à l'intérieur d'un p longueur p) et qui se produisent presque de manière identique à partir de la période à la période

Exemple : Température à Nottingham



La composante saisonnière

Ce composant est entièrement défini par p coefficients saisonniers m_1, \dots, m_p qui représentent le profil saisonnier.

Nous avons $S_t = m_1, \dots, m_p, m_1, \dots, m_p, \dots$

D'où la composante saisonnière S_t est la répétition de ces p coefficients sur chaque période.

La composante résiduelle E_t

Cette composante représente des fluctuations irrégulières et « imprévisibles ».

Il s'agit d'une composante aléatoire, souvent considérée comme une variable aléatoire à moyenne nulle

Déterminer la composition d'une série chronologique

Toutes les séries chronologiques ne sont pas composées à la fois d'une tendance et d'une composante saisonnière

Les types typiques de séries chronologiques sont

- complètement aléatoire : $X_t = E_t$
- avec une composante de tendance uniquement : $X_t = T_t + E_t$
- avec une composante saisonnière uniquement : $X_t = S_t + E_t$
- avec à la fois une composante tendance et saisonnière : $X_t = T_t + S_t + E_t$

L'étape suivante est un outil pour vous aider à déterminer dans quel cas nous sommes

Fonction d'autocorrélation

Laisser $X_t = x_1, \dots, x_n$ être une série chronologique, et k un entier $\in [0, n-1]$. La fonction d'autocorrélation de X_t est défini comme :

$$\rho(k) = \frac{\text{cove}(x_t, x_{t-k})}{\text{var}(x_t)} = \frac{\sum_{t=k+1}^n (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

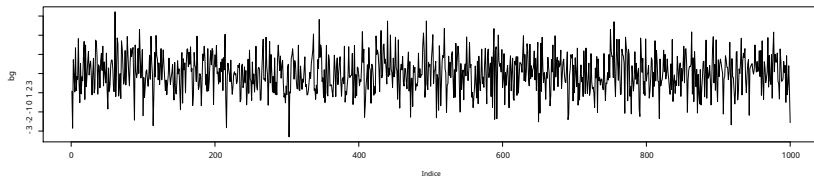
Propriétés :

- $\rho(0) = 1$
- $\rho(k) \leq 1, \forall k > 0$

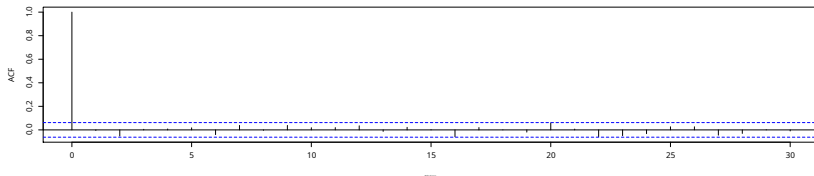
Autocorr fonction d'élacion

Pour un comp séries temporelles aléatoires ($X_t = E_t$),

$$|\rho(l)| < 1.96 * \frac{1}{\sqrt{n}}, \forall l > 0$$



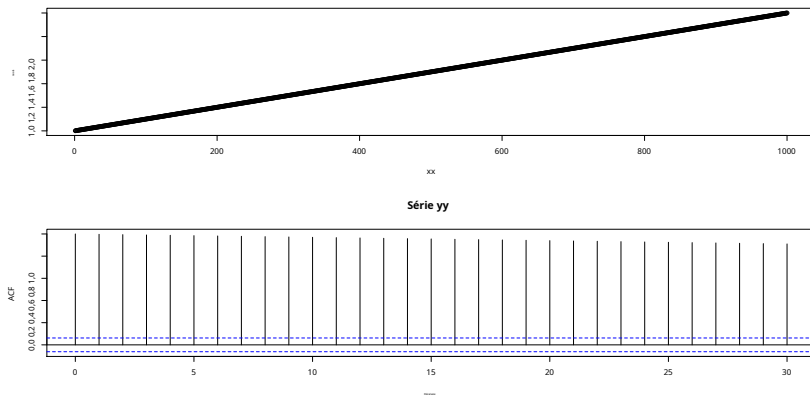
Série bbg



Fonction d'autocorrélation

Pendant un temps série avec une tendance ($X_t = T_t + E_t$), nous avons:

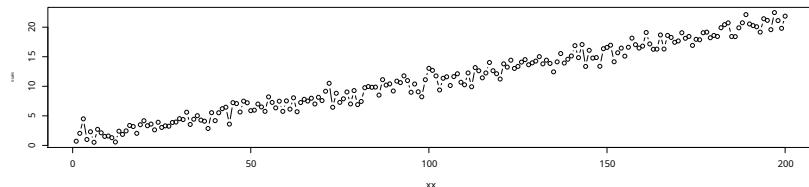
$$\rho(l) \rightarrow 1, \forall l > 0$$



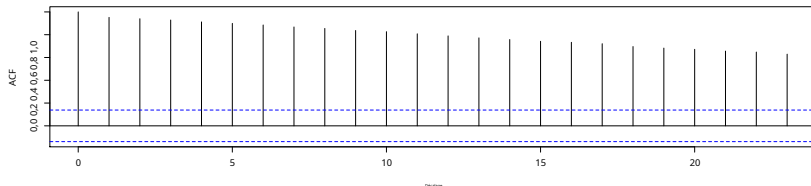
Autocorr fonction d'élation

Pendant un temps série avec une tendance ($X_t = T_t + E_t$), nous avons:

$$\rho(l) \rightarrow 1, \forall l > 0$$



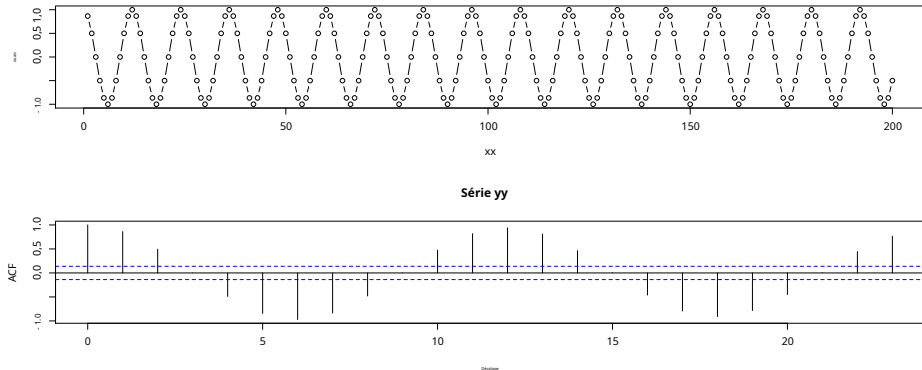
Series yy



Fonction d'autocorrélation

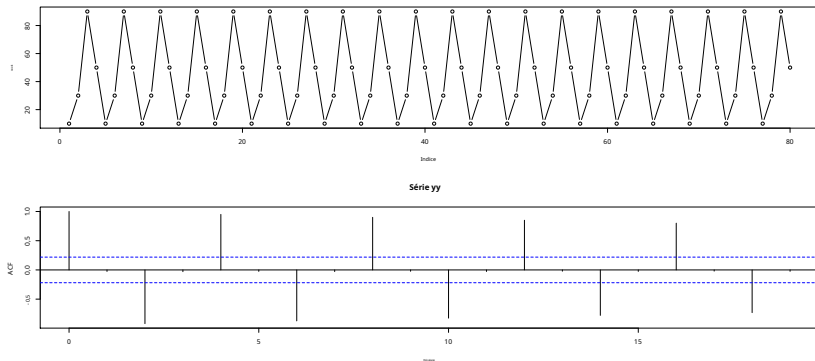
Pour une série chronologique du genre $x_t = \cos\left(\frac{2\pi}{T}t\right)$ (avec une composante saisonnière), nous avons :

$$\rho(h) = \cos\left(\frac{2\pi}{T}h\right), \forall h \geq 0$$



Autocorr fonction d'élation

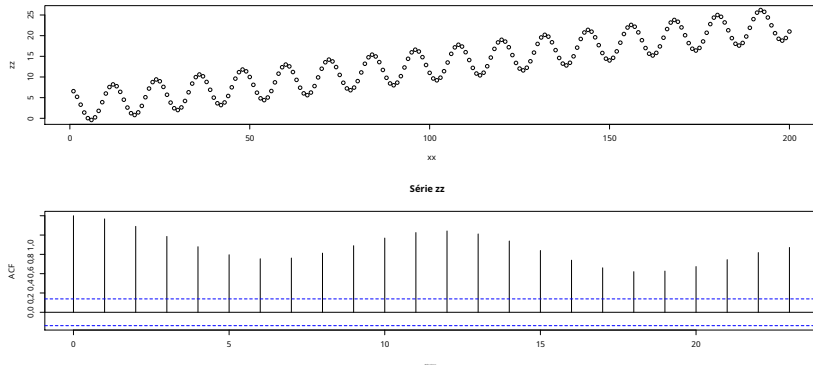
Exemple de série chronologique y_t avec une période de longueur 4



Fonction d'autocorrélation

Et pour autant que Série présentant à la fois une composante tendance et une composante saisonnière :

$$x_t = a + b \times t + \text{parce que} \left(\frac{2\pi}{T} t \right)$$

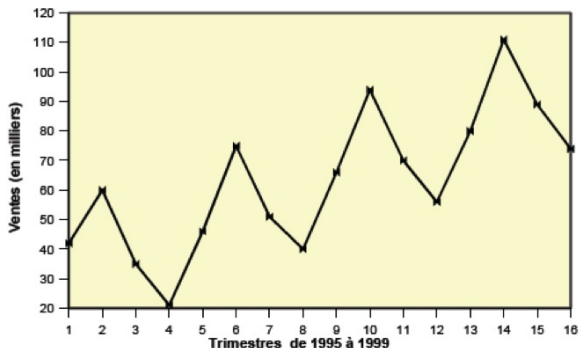


Résumé:

- ① Intuition sur le phénomène
- ② Représentations graphiques des séries temporelles
- ③ Analyse de la fonction d'autocorrélation
 - ▶ diminution lente des coefficients → s'orienter
 - ▶ périodicité des coefficients → saisonnalité
 - ▶ combinaison des deux : tendance + saisonnalité

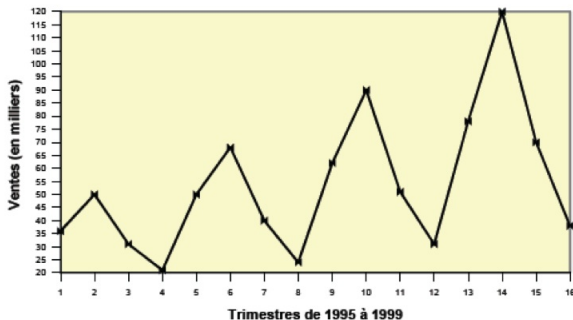
Composante saisonnière : additive ou multiplicative ?

- Pour un modèle additif, la série temporelle reste à l'intérieur d'une « bande » constante autour de la tendance



Composante saisonnière : additive ou multiplicative ?

- Pour un modèle multiplicatif, la « bande » autour de la moyenne n'est pas constante dans le temps, c'est-à-dire que les composantes saisonnières dépendent de la tendance



1 Introduction : contexte, définitions

2 Décomposition des séries chronologiques

3 **Prédiction des séries chronologiques**

- **Série sans tendance ni saisonnalité**

Exemple

Prédiction d'une série sans tendance ni saisonnalité

C'est le cas le plus simple, rarement utilisé en pratique. Mais il permet de comprendre et d'appréhender certaines choses importantes qui serviront plus tard.

Hypothèse : la série temporelle prend ses valeurs autour d'un niveau stable μ , il n'a aucune tendance ni aucune composante saisonnière.

$$X_t = \mu + \epsilon_t,$$

avec ϵ_t une composante résiduelle.

Problème : comment estimer μ afin de prédire les valeurs futures de X_t ?

Prédiction d'une série sans tendance ni saisonnalité

Laisser $X_t = x_1, \dots, x_n$ une série sans tendance ni saisonnalité. On cherche à prédire la valeur suivante : \hat{x}_{n+1} (et pourquoi pas les prochains).

1 Naïve : la dernière valeur x_n est utilisée comme prédiction au moment $n+1$

$$\hat{x}_{n+1} = x_n$$

- + la valeur suivante est susceptible d'être proche de la précédente
- seule la valeur précédente est utilisée pour prédire (les informations passées sont perdues)

Sur notre exemple :

Prédiction d'une série sans tendance ni saisonnalité

La prédiction est la moyenne de toutes les valeurs passées :

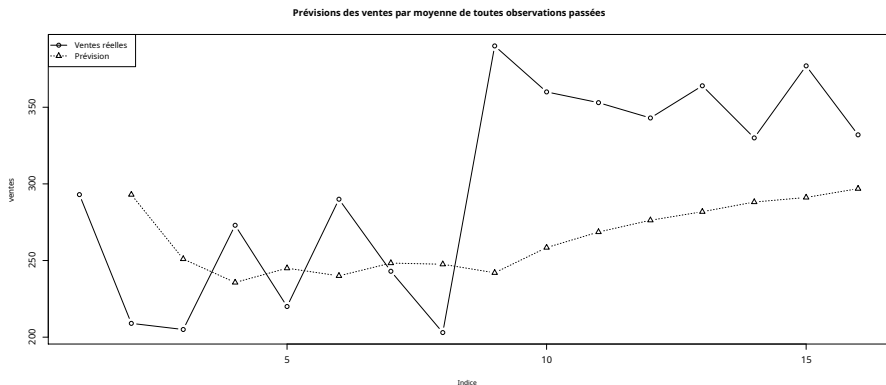
$$\hat{x}_{n+1} = \frac{1}{n} \sum_{t=1}^n x_t$$

+ toutes les valeurs passées sont utilisées pour la prédiction

—

Sur notre exemple :

Prédiction n en utilisant la moyenne des valeurs passées

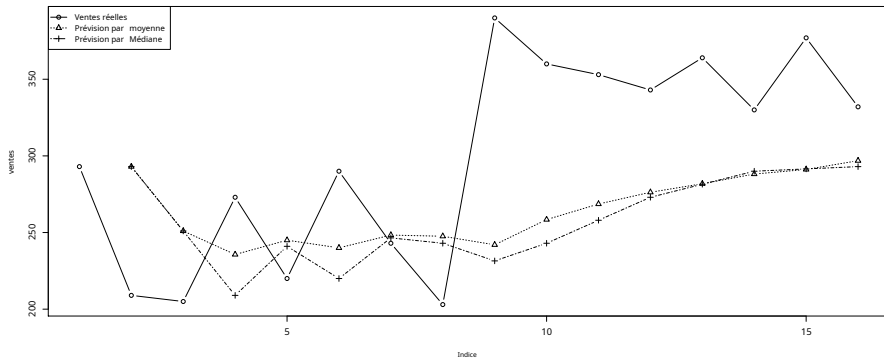


Prédiction d'une série sans tendance ni saisonnalité

- 3 la prédiction est la médiane de toutes les valeurs passées
- + toutes les données disponibles sont prises en compte
- + les valeurs extrêmes ont moins d'influence
- le passé proche est aussi important que le passé lointain

Dans notre exemple :

Prédiction n en utilisant la médiane des valeurs passées



Prédiction d'une série sans tendance ni saisonnalité

4 Somme pondérée de toutes les valeurs passées

$$\hat{x}_{n+1} = \frac{\sum_{je=1}^n \omega_{je} x_{je}}{\sum_{je=1}^n \omega_{je}}$$

- + nous pouvons décider de donner plus d'influence à certaines valeurs passées
- + toutes les données disponibles sont utilisées
- choix de ω_{je} est fastidieux

Dans notre exemple :

Prédiction d'une série sans tendance ni saisonnalité

5 Lissage exponentiel simple

- + plus d'influence sur les valeurs récentes
- + toutes les données disponibles sont utilisées
- + flexible et facile à appliquer

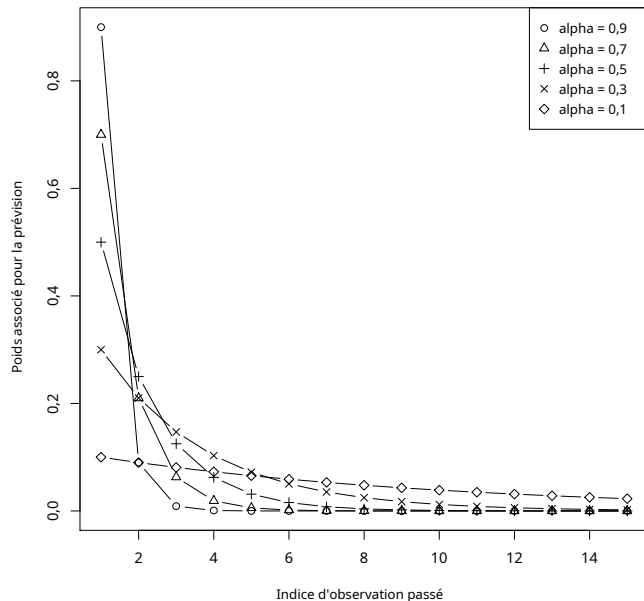
Principe:

Lissage exponentiel simple : analyse

$$\begin{aligned}\hat{x}_{n+1} &= \alpha x_n + (1 - \alpha) \hat{x}_n \\ &= \alpha x_n + (1 - \alpha) (\alpha x_{n-1} + (1 - \alpha) \hat{x}_{n-1}) \\ &= \alpha x_n + \alpha(1 - \alpha) x_{n-1} + (1 - \alpha)^2 (\alpha x_{n-2} + (1 - \alpha) \hat{x}_{n-2}) = \dots \\ &= \alpha x_n + \alpha(1 - \alpha) x_{n-1} + \alpha(1 - \alpha)^2 x_{n-2} + \dots + \alpha(1 - \alpha)^k x_n \\ &\quad - \alpha(1 - \alpha)^{n-1} x_1 + (1 - \alpha)^n \hat{x}_1\end{aligned}$$

Lissage exponentiel simple : exemple

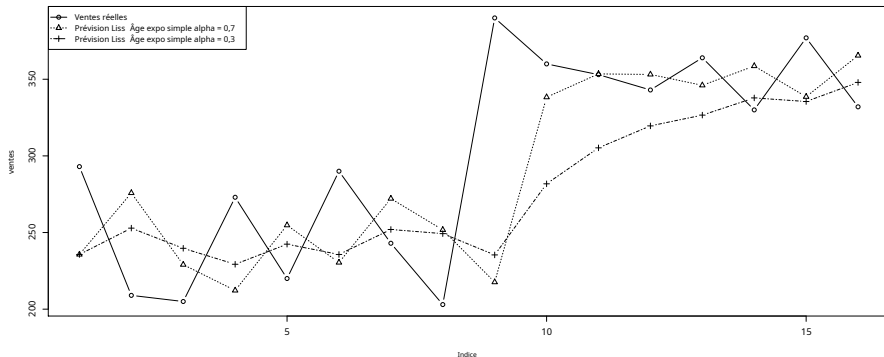
Lissage exponentiel simple : analyse



Influence du paramètre α

- si α est grand (proche de 1), on donne beaucoup d'influence aux valeurs récentes. A la limite ($\alpha=1$), je prédis la dernière valeur observée
 - les prévisions fluctuent ; réaction rapide aux changements dans les données
- si α est petit (proche de 0), plus d'influence est accordée aux valeurs passées (tend vers la moyenne de toutes les valeurs passées)
 - les prévisions sont stables mais réagissent lentement aux changements dans les données

E simple Lissage exponentiel



Choix de α et x_0

Modèles auto-régressifs (AR)

Principe :

$$\hat{x}_n = \beta_0 + \sum_{k=1}^p \beta_k x_{n-k}$$

- p est l'ordre du modèle : AR(p)
- les coefficients β sont estimés sur la base de la série originale

Modèles auto-régressifs : exemple

Comment choisir la commande p ?

Pour une donnée p , lorsqu'un modèle est ajusté (les coefficients sont estimés), nous pouvons calculer le score BIC (Bayesian Information Criterion).

Il s'agit d'un compromis entre la précision des prédictions intermédiaires et la complexité du modèle (nombre de coefficients)

Le meilleur rapport qualité/prix p pour une série temporelle donnée, on peut choisir en fonction de ce score (plus le BIC est bas, mieux c'est)

Résumé

