« Cours Statistique et logiciel R »

Rémy Drouilhet [1], Adeline Leclercq-Samson [1],
Frédérique Letué [1], Laurence Viry [2]

[1] Laboratoire Jean Kuntzmann, Dép. Probabilites et Statistique,

[2] Laboratoire Jean Kuntzmann, Dép. Modele et Algorithmes Déterministes

mars-avril 2016

Plan de la présentation

Qu'est ce qu'une série temporelle ?

Une **série chronologique**, ou **chronique** ou **série temporelle**, est une suite finie de données indexée par le temps.

Si $t_1, t_2, \ldots, t_n$ sont les n instants successifs d'observation et si y_{t_i} est la valeur mesurée à l'instant t_i, on notera la série chronologique $\{y_t\}_{t \in T}$ où T est l'ensemble ordonné $T = \{t_1, t_2, \ldots, t_n\}$.

Objectifs

- Décrire, expliquer un phénomène evoluant au cours du temps
- Prévoir des valeurs futures

Domaines d'application

- Le temps est indicé le plus souvent, en années, trimestres, mois, jours, heures,...
- Les données seront par exemple
  - Economie (prédiction d'indices économiques...)
  - Finance (évolution des cours de la bourse...)
  - Démographie (analyse de l'évolution d'une population...)
  - Météorologie (analyse de données climatiques...)
  - Médecine (analyse d'électrocardiogrammes...)
  - Géophysique (analyse de donnés sismiques...)
  - Traitement d'images (analyse d'images satellites, médicales...)
  - Energie (prévision de la consommation d'électricité...)

Exemples

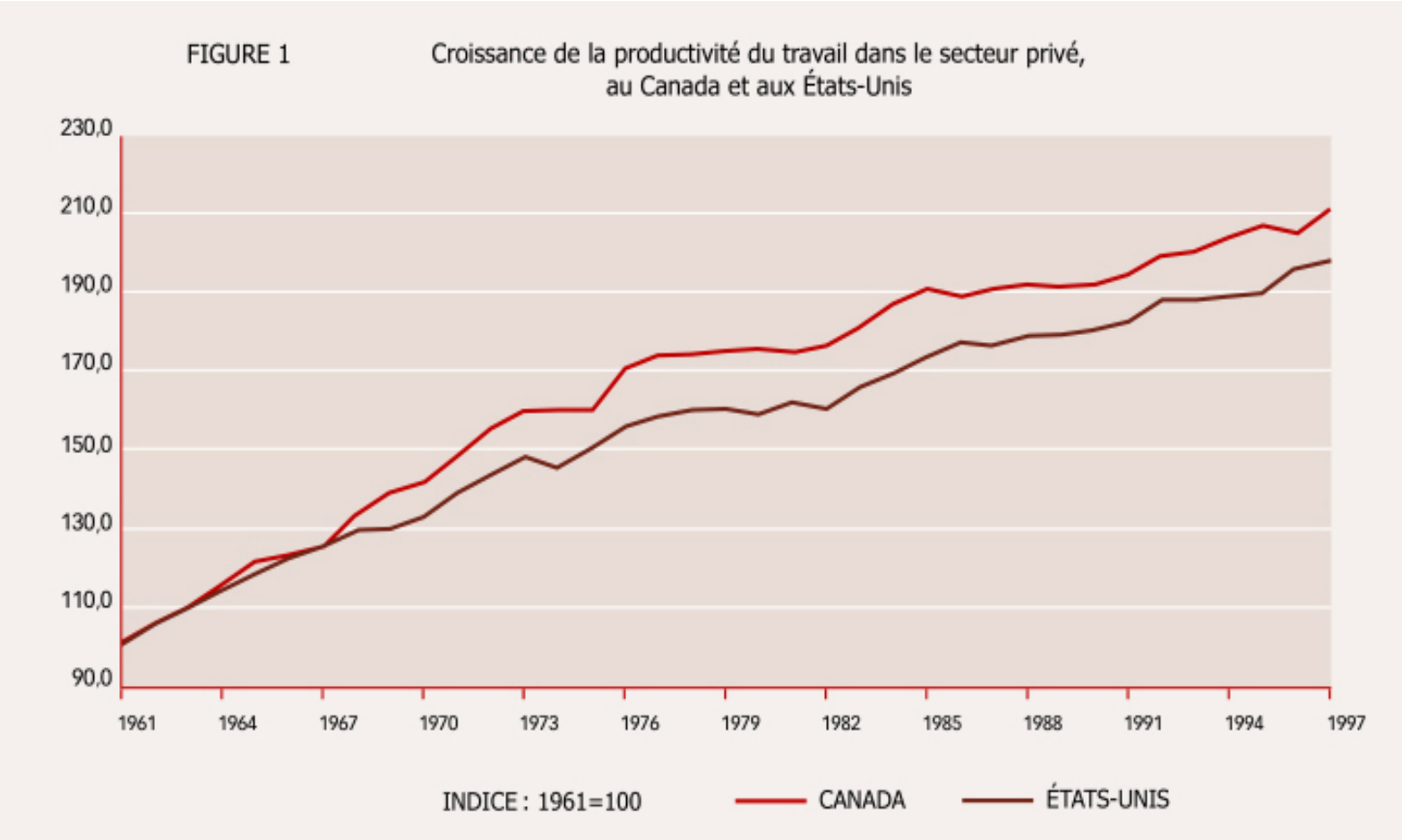
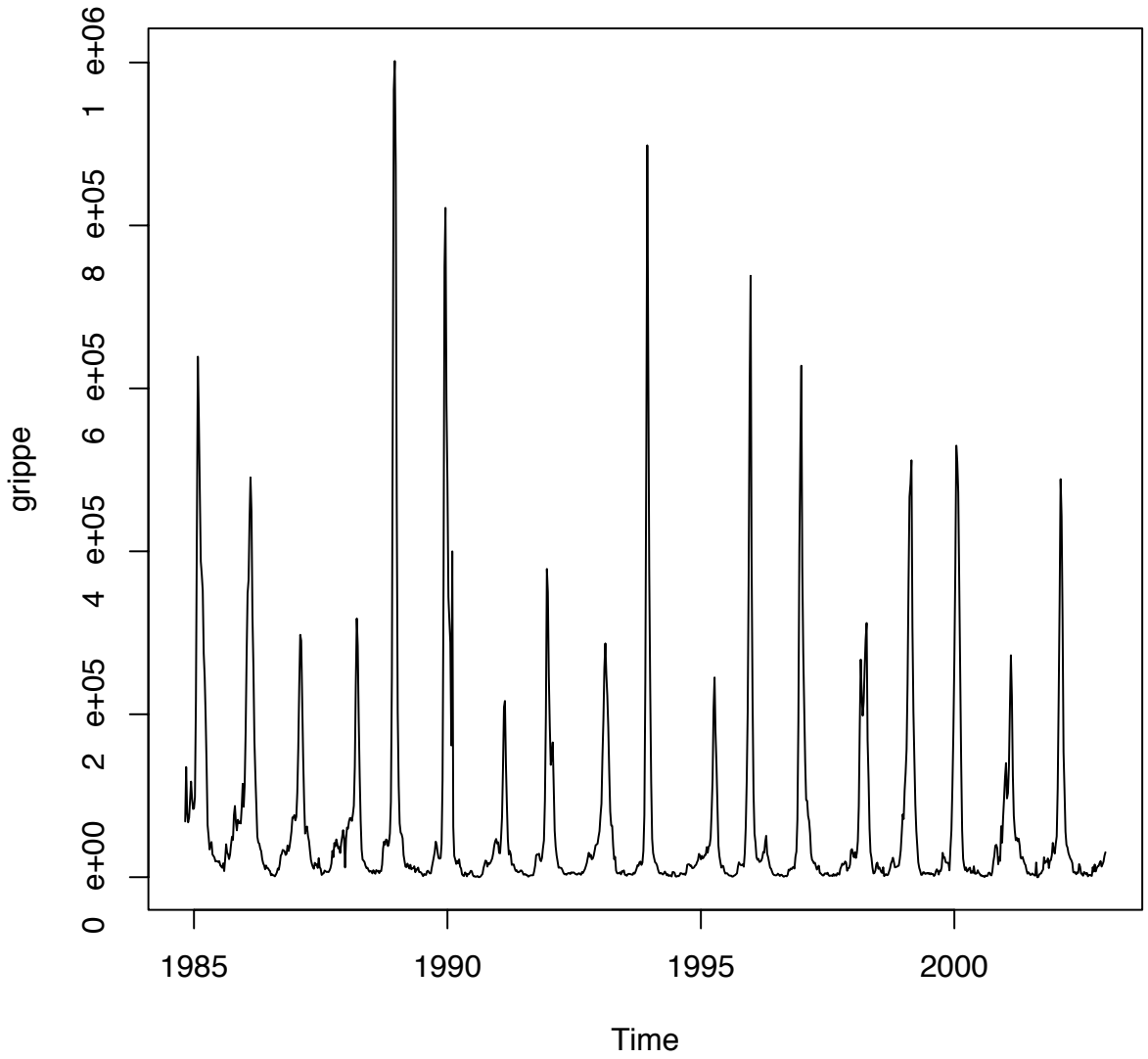
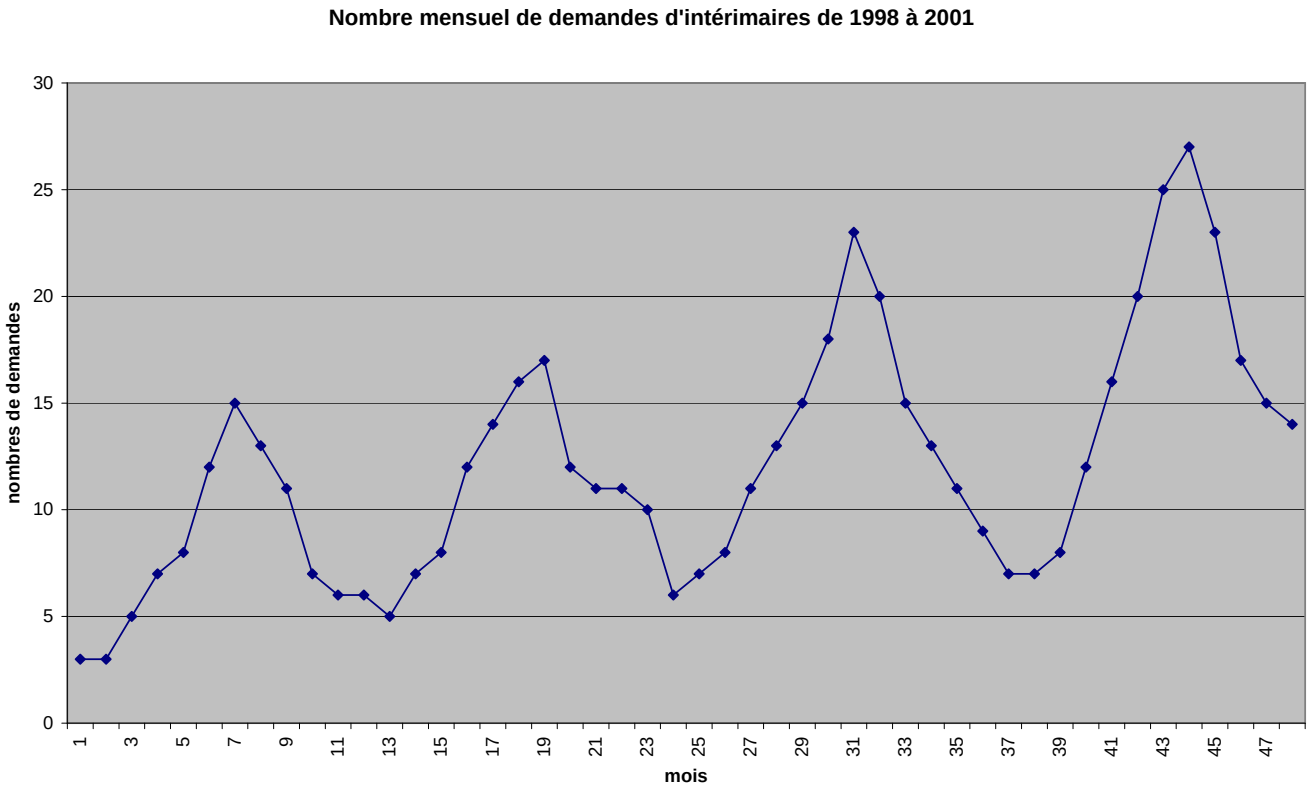


FIGURE 1
Croissance de la productivité du travail dans le secteur privé,
au Canada et aux États-Unis
230,0
210,0
190,0
170,0
150,0
130,0
110,0
90,0
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
INDICE : 1961=100
CANADA
ÉTATS-UNIS


grippe
0 e+00
2 e+05
4 e+05
6 e+05
8 e+05
1 e+06
1985
1990
1995
2000
Time


Nombre mensuel de demandes d'intérimaires de 1998 à 2001
nombres de demandes
mois
0
5
10
15
20
25
30
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47

Convention

On supposera dans toute la suite, que les dates sont équidistantes et donc nous adopterons la notation simplifiée

$$(y_i)_{i=1,\ldots,n}$$

pour désigner la série chronologique

$$\{y_t\}_{t\in T} \quad \text{avec} \quad T = \{t_1, t_2, \ldots, t_n\}.$$

Modélisation d'une série chronologique

Une règle générale en statistique descriptive consiste à commencer par regarder les données, avant d'effectuer le moindre calcul.

Ainsi, **l'examen du graphe** de la série peut mettre en évidence :

- une tendance : le phénomène étudiée a-t-il tendance à croître ou à décroître ?
- y a t-il un phénomène périodique ? Lié par exemple aux saisons ?
- y a t-il des variations exceptionnelles et peut-on les expliquer ?

En définitive, il s'agit de déterminer les éléments constitutifs de l'evolution globale d'une chronique : ils portent le nom de **composantes**.

Les composantes d'une chronique

On distingue :

- **la tendance** $(f_i, 1 \leq i \leq n)$:
  Elle représente l'évolution à long terme de la grandeur étudiée, et traduit l'aspect général de la série.

- **les variations saisonnières** $(s_i, 1 \leq i \leq n)$:
  - elles sont liées au rythme imposé par
    - les saisons météorologiques (production agricole, consommation de gaz, ...),
    - les activités économiques et sociales (fêtes, vacances, soldes, etc).
    - d'autres causes régulières...
  - elles sont de nature périodique, c'est-à-dire qu'il existe un entier p, appelé **période**, tel que

$$s_i = s_{i+p}, \quad \text{pour tout } i \geq 1.$$

Cette composante est donc entièrement déterminée par ses p premières valeurs $s_1, s_2, \ldots, s_p$.

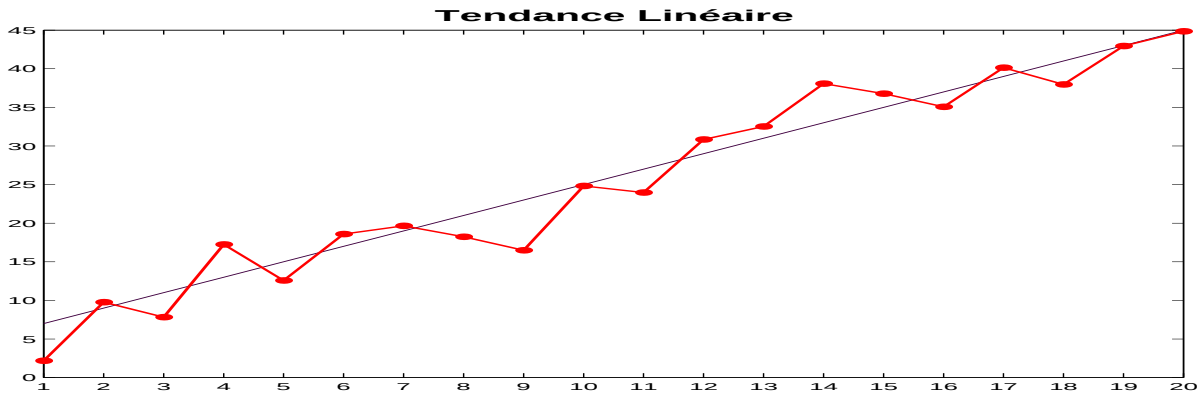
- **la composante résiduelle** ou **bruit** $(e_i, 1 \leq i \leq n)$:
  Les variations résiduelles (qui "restent" quand on a éliminé tous les autres mouvements) peuvent être de deux natures différentes :
  - La plupart du temps, elles proviennent d'un grand nombre de petites causes, sont inexpliquées
    - Elles sont de faible amplitude et de courte durée.
    - On parle alors de **fluctuations irrégulières**
  - Quelquefois, elles proviennent d'évenements accidentels de grandes ampleurs
    - dûs à des accidents importants et explicables : Mai 68, réunification de l'Allemagne, tempête, . . .).
    - On parle alors de **variations accidentelles**.
    - Graphiquement, elles correspondent à des valeurs isolées anormalement élevées ou faibles.

On considère qu'une série chronologique $(y_i, 1 \leq i \leq n)$ est la résultante de 3 composantes fondamentales :

- $(f_i, 1 \leq i \leq n)$, **la tendance** ou **trend** (intégrant éventuellement un cycle),
- $(s_j, 1 \leq j \leq p)$, **la composante saisonnière** ou **saisonnalité**,
- $(e_i, 1 \leq i \leq n)$, **la composante résiduelle** ou **bruit** (intégrant les fluctuations irrégulières et éventuellement des accidents).

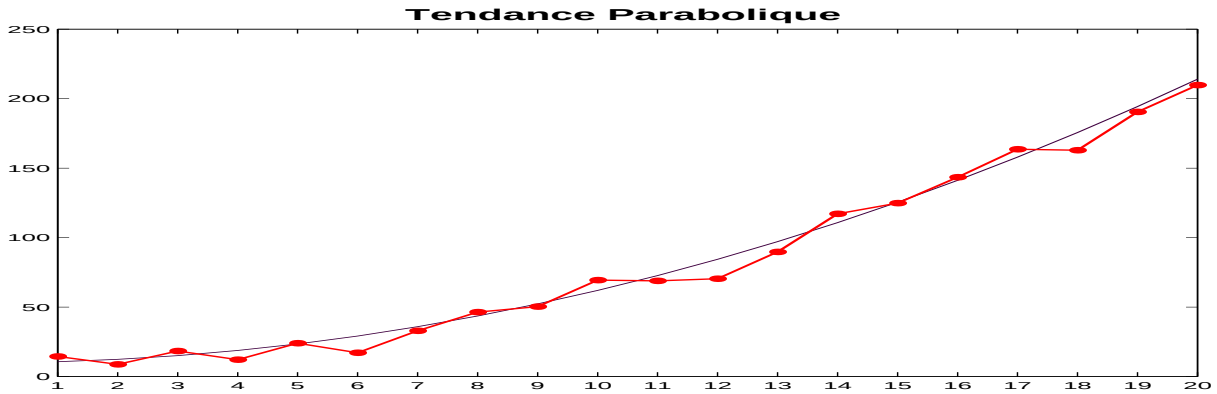
Modélisation de la tendance par une droite

$$f_\theta(t) = a\,t + b \quad \text{avec} \quad \theta = (a, b)'$$



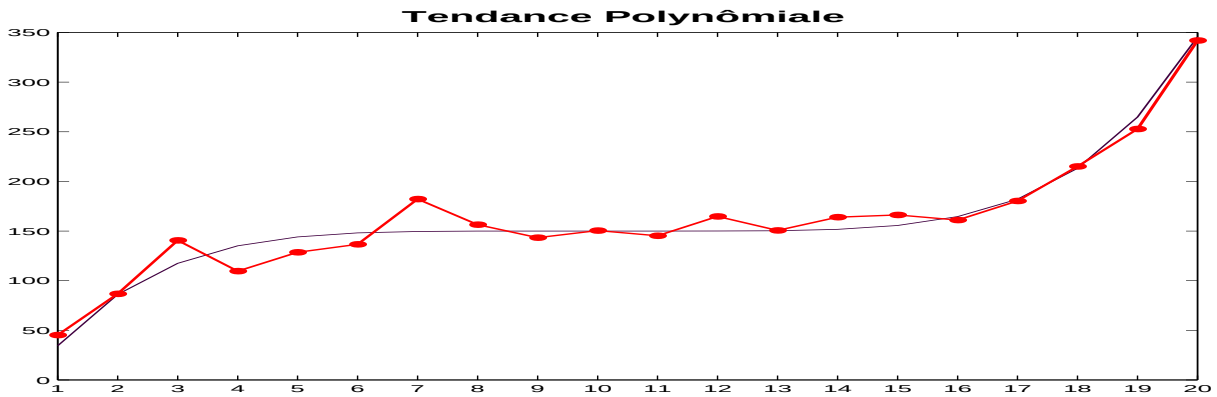
La parabole

$$f_\theta(t) = a\,t^2 + b\,t + c \quad \text{avec} \quad \theta = (a, b, c)'$$



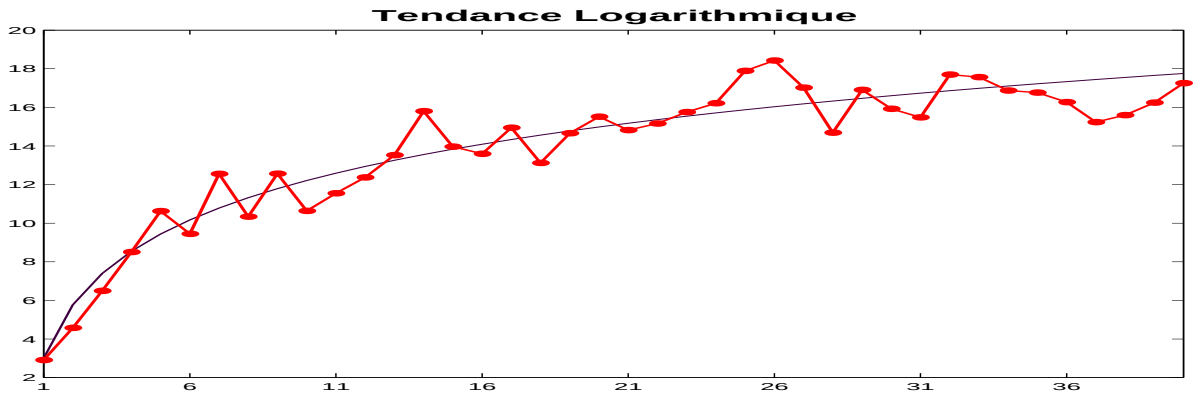
La courbe polynômiale

$$f_\theta(t) = a_p\, t^p + a_{p-1}\, t^{p-1} + \ldots + a_1\, t + a_0 \text{ avec } \theta = (a_p, \ldots, a_1, a_0)'$$



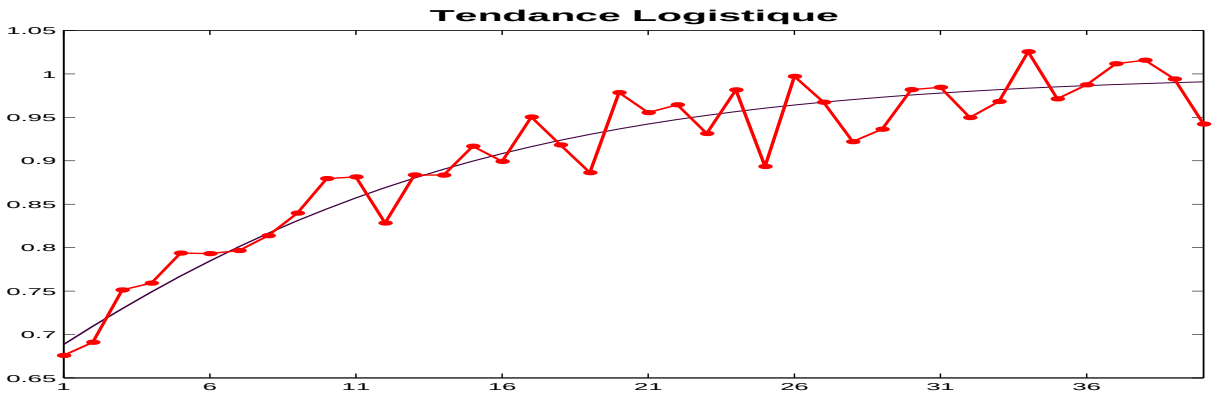
La courbe logarithmique

$$f_\theta(t) = a \log(t) + b \text{ avec } \theta = (a, b)'$$



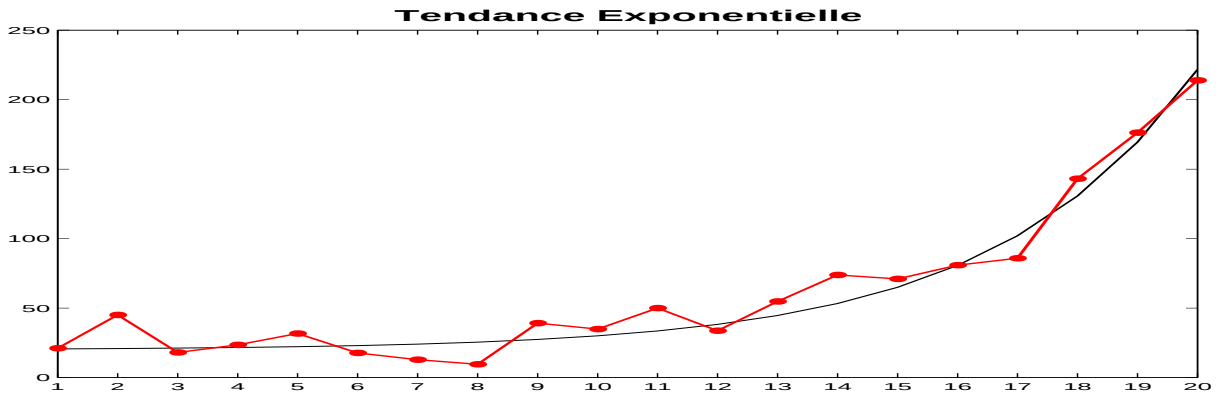
La courbe logistique

$$f_\theta(t) = \frac{1}{b\,\exp(-a\,t) + c} \quad \text{avec} \quad \theta = (a, b, c)'$$



La courbe exponentielle

$$f_\theta(t) = a \exp(b\,t) + c \quad \text{avec} \quad \theta = (a, b, c)'$$



Les modèles de décomposition

On étudiera deux modèles :

- le modèle additif
- le modèle multiplicatif

combinant chacun :

- une tendance (m_i) ;
- une composante saisonnière (s_i) ;
- une composante résiduelle (u_i).

Le modèle additif

Modèle additif

$$x_i = m_i + s_i + u_i \quad \text{pour} \quad i = 1, \ldots, n \tag{1}$$

avec

$$\sum_{j=1}^{p} s_j = 0 \quad \text{et} \quad \sum_{i=1}^{n} u_i = 0.$$

Graphiquement... Dans le modèle additif, l'amplitude de la composante saisonnière et du bruit reste constante au cours du temps. Ceci se traduit graphiquement par de fluctuations autour de la tendance d'amplitude **constante**.

Le modèle multiplicatif

Modèle multiplicatif

$$x_i = m_i\,(1+s_i)\,(1+u_i) \quad \text{pour} \quad i=1,\ldots,n \tag{2}$$

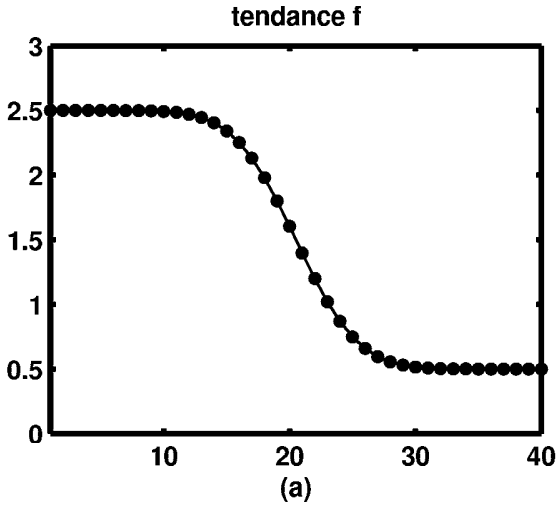
avec

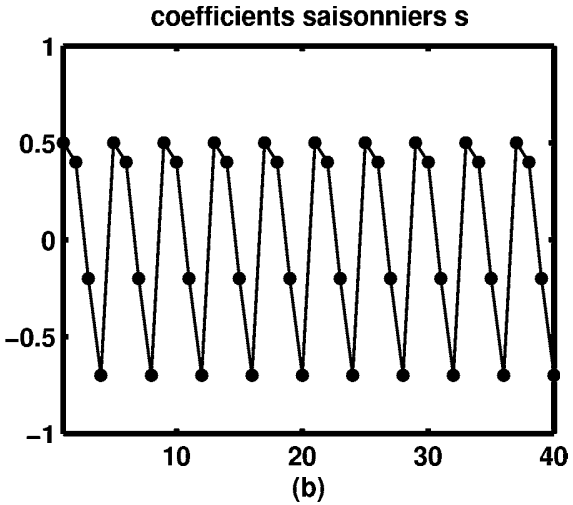
$$\sum_{j=1}^{p} s_j = 0 \quad \text{et} \quad \sum_{i=1}^{n} u_i = 0.$$

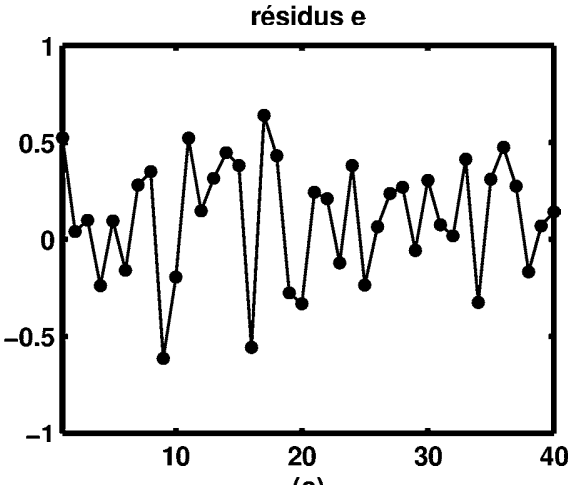
Graphiquement... Dans le modèle multiplicatif, l'amplitude de la composante saisonnière et du bruit n'est plus constante au cours du temps : elles varient au cours du temps proportionnellement à la tendance.

Apprendre à reconnaître le bon modèle

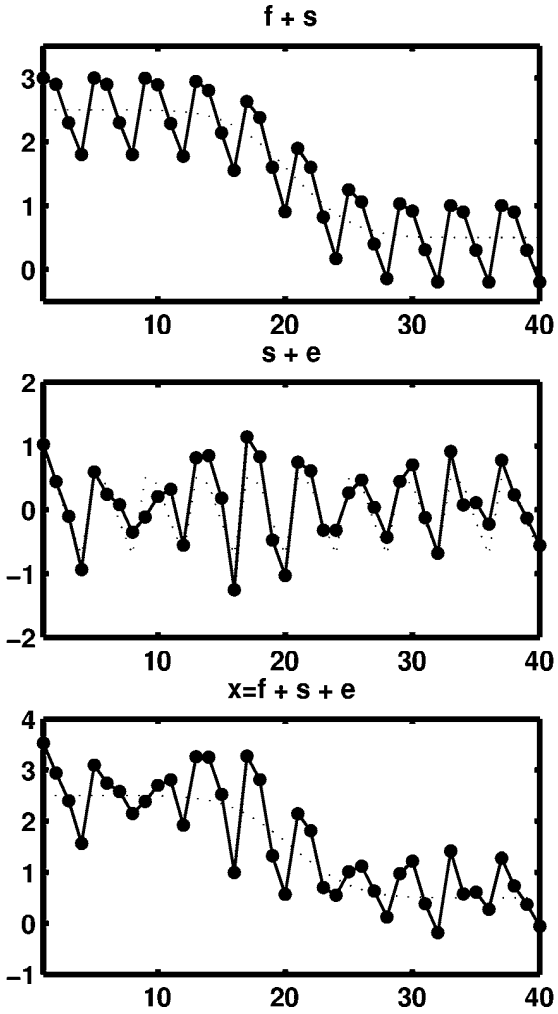
Considérons les trois composantes suivantes :



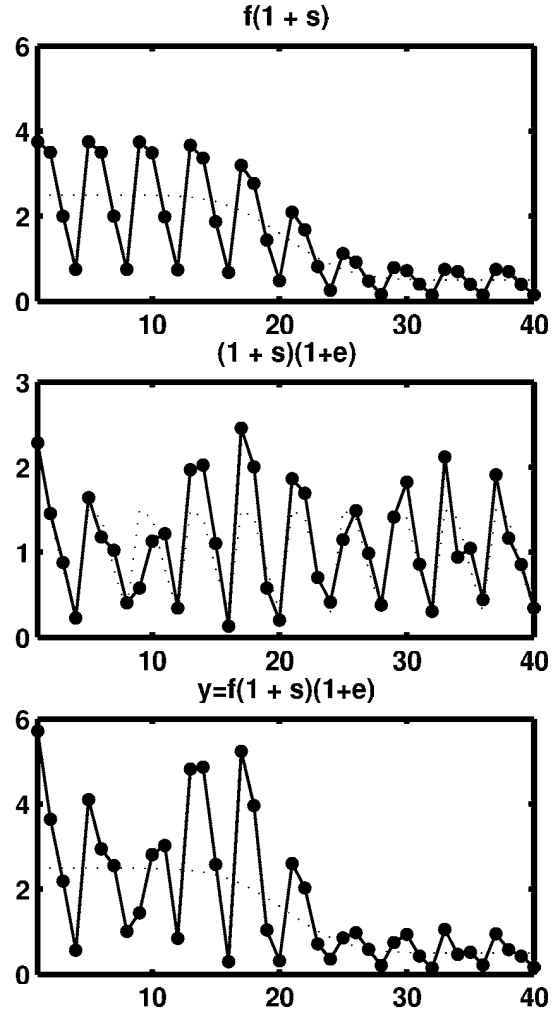




Modèle Additif



Modèle Multiplicatif



Objectifs

Buts de l'analyse des séries chonologiques :

- Description : décrire l'évolution du phénomène dans le temps. Phénomène périodique ? Survenue d'événements accidentels ?
- Déterminer les composantes d'une série chronologique. Supprimer l'effet saisonnier . En effet, les variations saisonnières peuvent masquer l'évolution principale du phénomène.
- Trouver une fonction simple du temps qui modèlise au mieux la tendance
- Faire de la prévision à court ou moyen terme : ayant observé $y_1, \ldots, y_n$, on veut prédire les valeurs futures $y_{n+1}, y_{n+2}, \ldots$

Ajustement de la tendance

On dispose d'une série chronologique $(y_i)_{i=1,\dots,n}$

Objectifs : trouver une fonction simple du temps qui modélise au mieux la tendance de la série $(y_i)_{i=1,\dots,n}$

Deux types de méthodes :

- par une méthode non paramétrique
- par une méthode paramétrique

Une méthode non paramétrique : lissage par moyenne mobiles

Les **moyennes mobiles** permettent de lisser directement la série sans hypothèse a priori sur la forme du modéle sous-jacent. La méthode est donc valable quel que soit le modéle de décomposition. Pour cette raison, on peut classer ce type de lissage dans les méthodes non paramétriques (par opposition aux méthodes paramétriques qui seront abordées dans la partie 3)

Objectif. Le but d'un lissage par moyenne mobile est

- de faire apparaître l'allure de la tendance
- en éliminant la composante saisonnière
- en atténuant les fluctuations irrégulières.

Principe du lissage par moyennes mobiles

On dispose d'une série chronologique $(y_i)_{i=1,\dots,n}$

Idée : on décide d'estimer la tendance en un point par une moyenne des observations qui l'entourent.

Moyennes mobiles simples

La **série des moyennes mobiles d'ordre** k, notée **MM(k)**, est la série des moyennes de k observations consécutives qui prend ses valeurs aux dates moyennes correspondantes :

- les moyennes de k termes consécutifs pour la variable y :

$$MM(k)_j = \frac{\overbrace{y_{j-m} + \cdots + y_{j-1}}^{m \text{ termes}} + y_j + \overbrace{y_{j+1} + \cdots + y_{j+m}}^{m \text{ termes}}}{2m+1}$$

Propriétés des moyennes mobiles
Elimination de la composante saisonnière

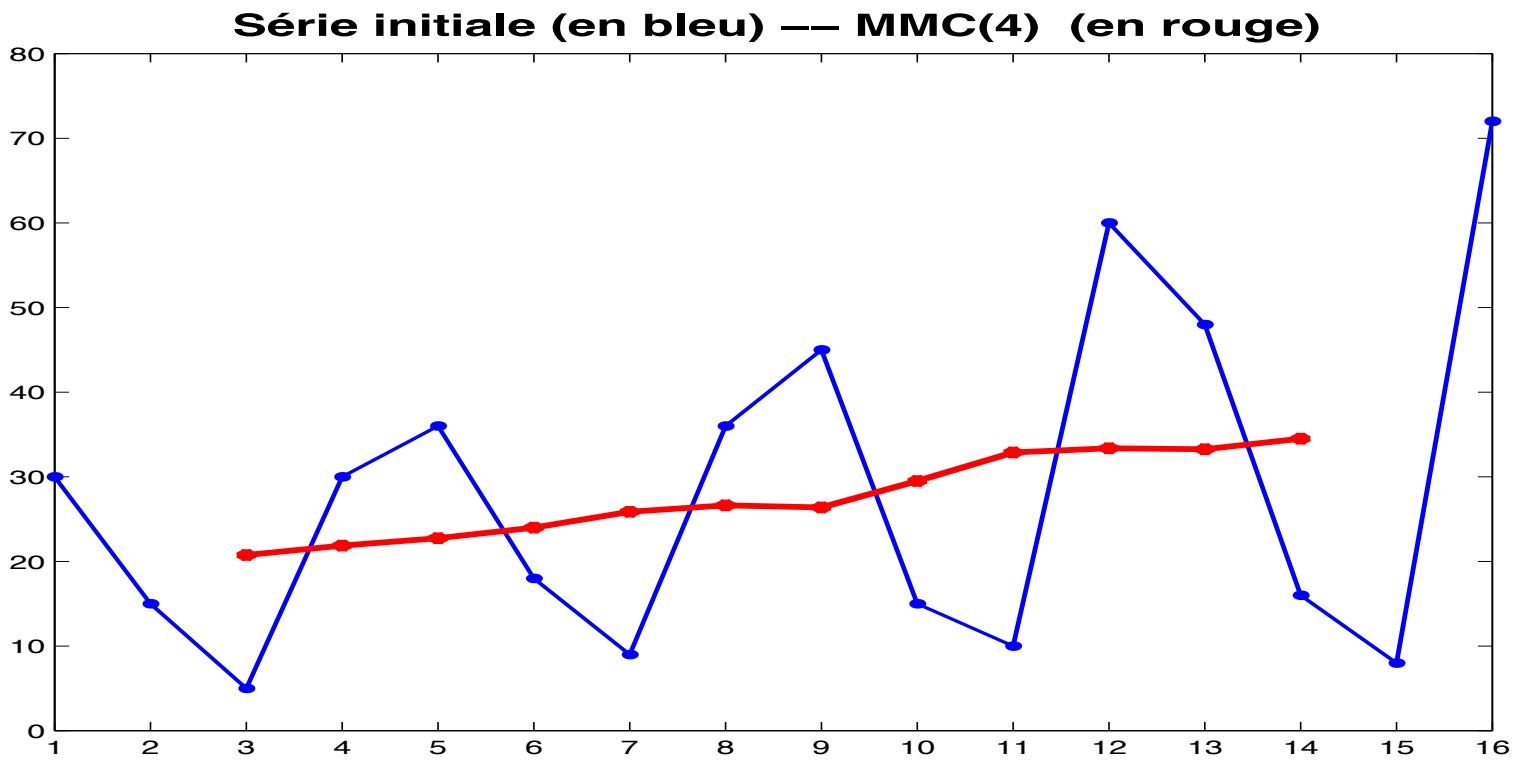
Elimination de la composante saisonnière

Si la série chronologique $(y_i)_{i=1,\dots,n}$ possède une composante saisonnière de période p, alors l'application d'une moyenne mobile d'ordre p supprime cette saisonnalité.

La série $MM(p)$ ou $MMC(p)$ ne possède plus de composante saisonnière de période p.

On se servira donc d'une moyenne mobile d'ordre p pour éliminer une composante saisonnière de période p.

Ventes trimestrielles d'un produit



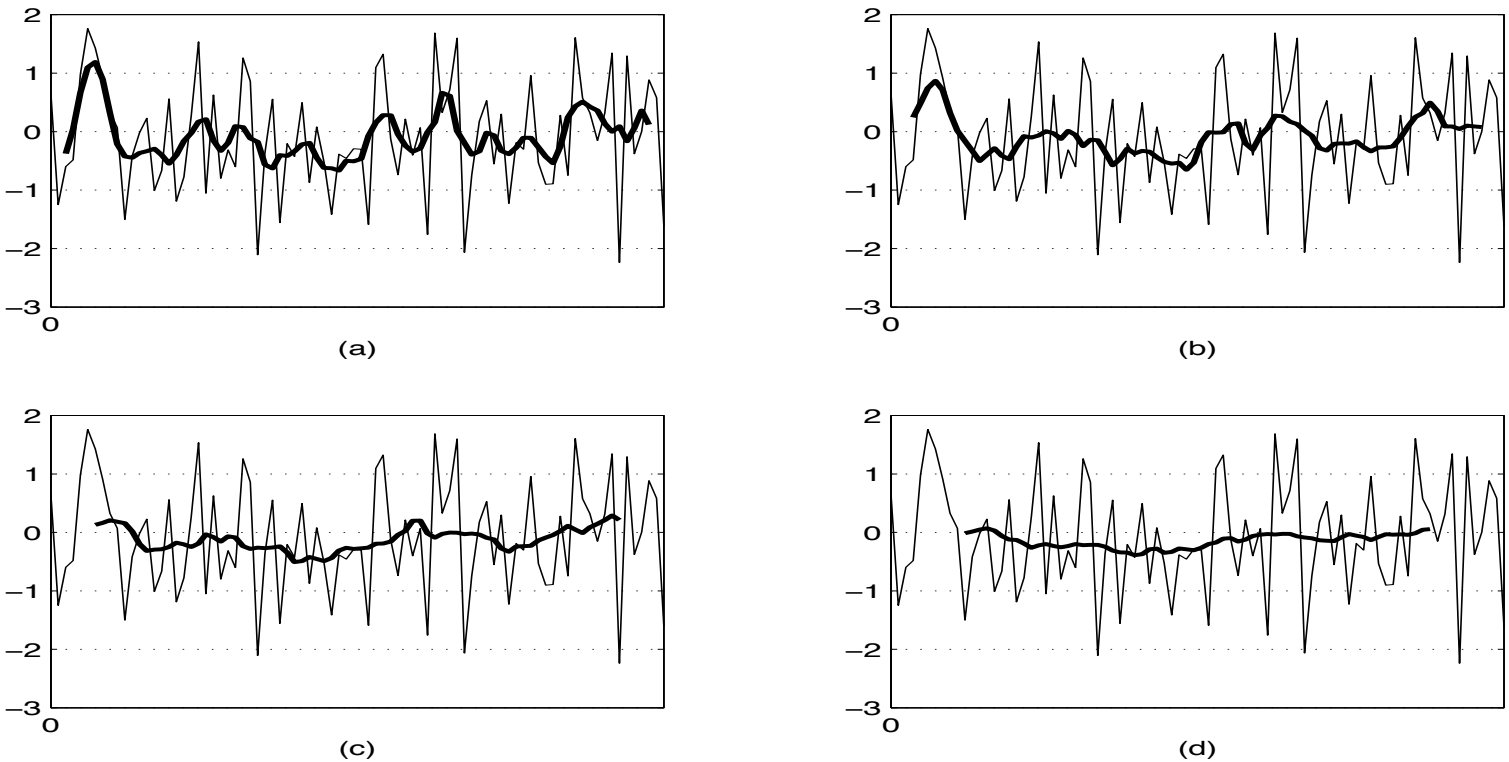
Atténuation des fluctuations irrégulières

Atténuation des fluctuations irrégulières

Par construction, une moyenne mobile consiste à faire des moyennes partielles de proche en proche. On obtient donc un "lissage" de la série.

1. Une moyenne mobile atténue l'amplitude des fluctuations irrégulières d'une chronique.
2. Plus l'ordre de la moyenne mobile est élevé, et plus cette atténuation est importante.

Moyennes mobiles sur une série de fluctuations irrégulières



La série des fluctuations irrégulières, avec en

(a), $MM(4)$ **(b)**, $MM(6)$
(c), $MM(12)$ **(d)**, $MM(20)$

Avec R

```
data = read.table('grippe.txt')
grippe = ts(data[,2],start=c(1984,44),frequency=52)
plot(grippe)
res=decompose(grippe,type="additive")
plot(res)
```

Ajustement paramétrique

Contexte : Nous supposerons dans cette partie que le modèle ne comporte plus de composante saisonnière : la série a été au préalable corrigée de ses variations saisonnières.

Ajustement paramétrique : On veut modéliser la tendance $(f_i)_{i=1,\dots,n}$ de la série chronologique $(y_i)_{i=1,\dots,n}$ par une fonction paramétrique du temps $f_\theta(\,.\,)$ où θ est un paramètre.

Il nous faut donc :

1. choisir une famille de fonctions $\{f_\theta\}$ dans une collection donnée de fonctions paramétriques
2. une fois la famille choisie, déterminer la valeur de θ qui conduit au *meilleur ajustement* (dans un sens à définir) de la série $(y_i)_{i=1,\dots,n}$.

Modélisation de la tendance

Choisir la famille de fonctions $\{f_\theta\}$

- Analyse graphique de la série chronologique détermine la famille de fonctions paramétriques à considérer
- Ajustement préliminaire de la tendance : méthode des moyennes mobiles

Choisir le meilleur ajustement

On veut ajuster la tendance de la série chronologique $(y_i)_{1 \leq i \leq n}$ par une fonction du temps f_θ, où θ est un paramètre de $\mathbb{R}^p$.

Une fois la famille de fonctions $\{f_\theta\}$ choisie, on veut déterminer la valeur de θ qui consuit au *meilleur ajustement.*

Le *meilleur ajustement* est déterminé à l'aide d'un critère, par exemple celui des moindres carrés :

On appellera estimation des moindres carrés, la valeur de θ qui minimise

$$\sum_{i=1}^{n} \Big(y_i - f_\theta(t_i) \Big)^2$$

On notera cette valeur $\hat{\theta}$.

Avec R

```
res = stl(grippe,s.window=52,s.degree=0)
plot(res)
tendance=res$time.series[,"trend"]
saison=res$time.series[,"seasonal"]
residus=res$time.series[,"remainder"]

#analyse de la tendance par un modele lineaire
lm.tendance=lm(tendance~time(tendance))
summary(lm.tendance)
plot(tendance)
abline(lm.tendance$coefficients[1],
lm.tendance$coefficients[2])
```