Séries chronologiques - Partie 1

Généralités - Composantes d'une série chronologique

BUT Science des Données, deuxième année

Définition

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

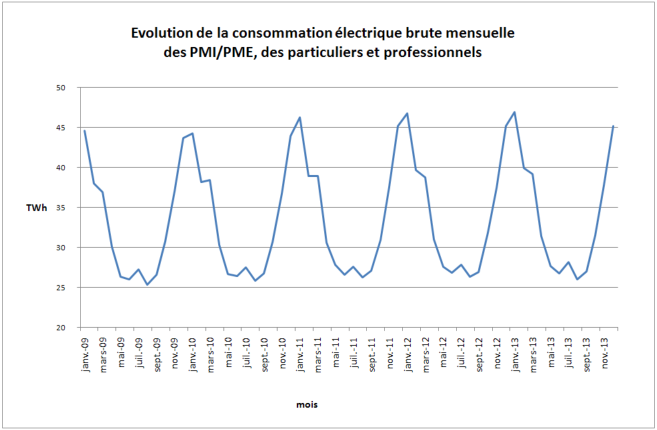
Le temps peut être selon les cas la seconde, la minute, l'heure, le jour, le mois, le trimestre, le semestre, l'année, ...

L'analyse des séries chronologiques a pour but de décrire, expliquer, et prévoir un phénomène évoluant au cours du temps.

Exemples

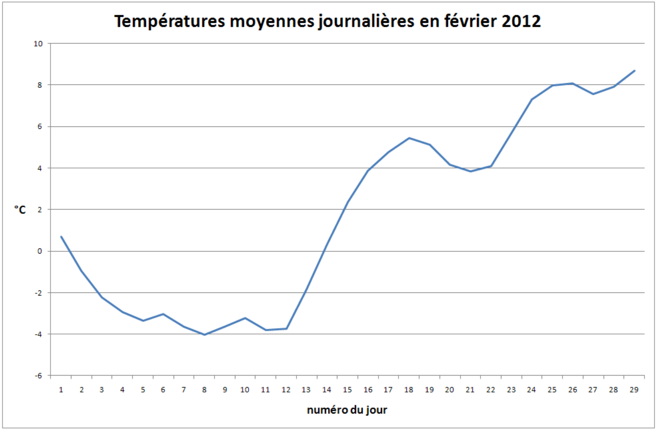
Elles se retrouvent dans de nombreux domaines.

- Economie : évolution d'indices (INSEE, bourse, ...), production-consommation d'un bien, ...



Elles se retrouvent dans de nombreux domaines.

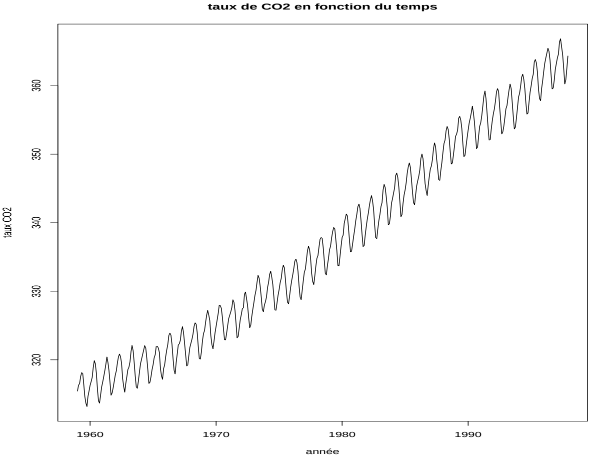
- Météorologie : pluies, températures, ...



Exemples

Elles se retrouvent dans de nombreux domaines.

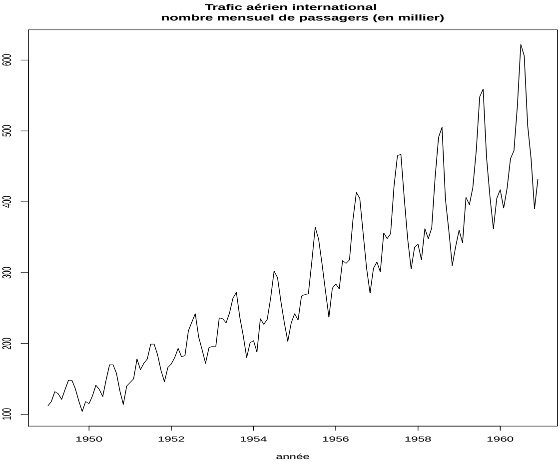
- Données climatiques: taux d'ozone, taux de CO2, . . .



Exemples

Elles se retrouvent dans de nombreux domaines.

- Trafic routier, trafic aérien, ...



Exemples

Elles se retrouvent dans de nombreux domaines.

- Démographie : comportement des familles (mariages, naissances, ...), évolution de la population rurale/urbaine/ d'un pays, ...
- Sociologie : crimes, mariages, ...
- Santé : épidémiologie (nombre quotidien de cas de grippes, de cas positifs covid, ...)
- Gestion : production, stocks, ventes, chiffres d'affaires ...
- Activités humaines : trafic routier, trafic téléphonique, ...
- ...

Formalisation 1

Soit Y notre variable d'intérêt qui évolue dans le temps, qu'on observe à n instants successifs.

Si $t_1, \ldots, t_n$ sont ces instants et 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 $T = \{t_1, t_2, \ldots, t_n\}$.

Formalisation 2

- La série chronologique $(y_t)_{t \in T}$ avec $T = \{t_1, t_2, \ldots, t_n\}$ peut aussi se définir comme une série statistique bidimensionnelle $(t_i, y_i)_{i=1,\ldots,n}$.
  - La première composante du couple définissant la série est le temps t et la deuxième composante est une variable numérique y prenant ses valeurs aux instants t.
  - Les valeurs de la première composante t sont rangées dans l'ordre chronologique, ce qui confère à la série (t, y_t) des propriétés particulières.

Considérons les données mensuelles de consommation d'électricité en France en 2013 :

Mois	Conso TWh
janvier 2013	52738
février 2013	45383
mars 2013	45177
avril 2013	37183
mai 2013	33557
juin 2013	32459

Mois	Conso TWh
juillet 2013	34094
aout 2013	31186
septembre 2013	32750
octobre 2013	37488
novembre 2013	43497
décembre 2013	50730

Rappel: 1 TW (térawatt)= 10^{12} watts.

Exemple

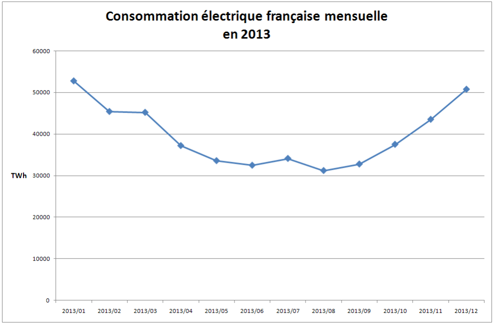
Considérons les données mensuelles de consommation d'électricité en France en 2013 :

i	t_i	y_{t_i}
1	janvier 2013	52738
2	février 2013	45383
3	mars 2013	45177
4	avril 2013	37183
5	mai 2013	33557
6	juin 2013	32459

i	t_i	y_{t_i}
7	juillet 2013	34094
8	aout 2013	31186
9	septembre 2013	32750
10	octobre 2013	37488
11	novembre 2013	43497
12	décembre 2013	50730

Une série chronologique se représente sous forme d'un nuage de points.

- Le temps t est porté en abscisses, dans l'ordre chronologique. La variable y_t est porté en ordonnées.
- Les points du nuage de points (t, y_t) sont reliés entre eux par des segments de droite : on obtient ainsi une courbe.



Exemple

Considérons les données relatives au nombre d'habitants aux Etats-Unis (en millions) :

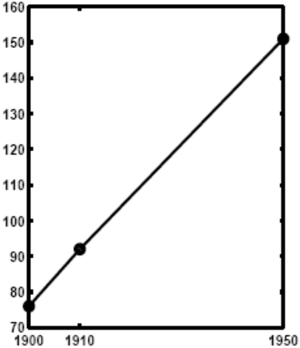
Année	Nb habitants
1900	76
1910	92
1950	151

Exemple

Considérons les données relatives au nombre d'habitants aux Etats-Unis (en millions) :

Indice	Date	Nb habitants
$i = 1$	$t_1 = 1900$	$y_1 = 76$
$i = 2$	$t_2 = 1910$	$y_2 = 92$
$i = 3$	$t_3 = 1950$	$y_3 = 151$

Vigilance lors de la représentation graphique



1. Représentation de $(t_i, y_i)_{i\in\{1,2,3\}}$ à gauche.
2. Représentation de $(i, y_i)_{i=1,2,3}$ à droite.

Convention

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

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

pour désigner la série chronologique

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

Cela revient à substituer la date par le numéro de l'observation.

Convention

En pratique, la série chronologique $(y_i)_{i=1,\ldots,n}$ est donnée sous la forme d'un tableau bidimensionnel où la date est remplacée par le numéro d'observation t.

Sur l'exemple de la consommation électrique :

t_i	y_i
1	52738
2	45383
3	45177
4	37183
5	33557
6	32459

t_i	y_i
7	34094
8	31186
9	32750
10	37488
11	43497
12	50730

Exercice

On considère la série $(y_i)_{i=1,\dots,n}$ du nombre d'iPhone vendus dans le monde par trimestre de début 2009 à fin 2013.

1. Que vaut n?
2. Que représentent les quantités suivantes :
   - $\frac{1}{5}\sum_{k=1}^{5} y_{4(k-1)+j}$ pour $1 \leq j \leq 4$?
   - $\sum_{j=1}^{4} y_{4(k-1)+j}$ pour $1 \leq k \leq 5$?

Rôle fondamental du graphe

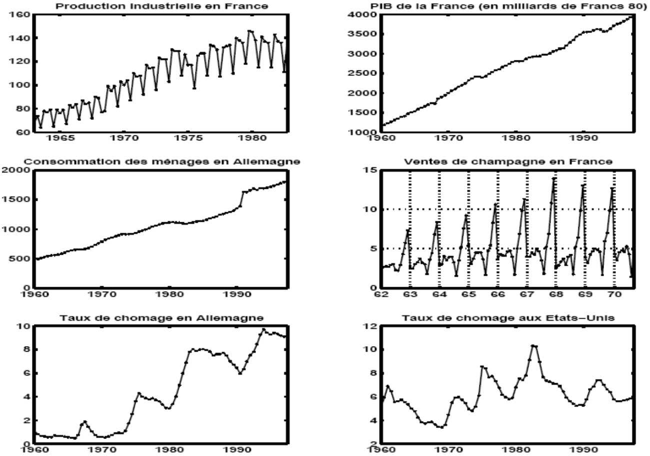
Une règle générale en statistique descriptive consiste à commencer par observer les données, avant d'effectuer le moindre calcul !

Ainsi, l'examen du graphe d'une série peut mettre en évidence :

- Une tendance: le phénomène étudié a-t-il tendance à croître ou à décroître ?
- Un comportement périodique : lié par exemple aux saisons.
- Des variations exceptionnelles : si oui peut-on les expliquer ?

En définitive, il s'agit de déterminer les éléments constitutifs de l'évolution globale d'une chronique, qu'on nomme composantes.

Que traduisent ces graphes ?



Nous pouvons décomposer la série brute en plusieurs éléments de base.

- La tendance générale $(f_i)_{i=1,\ldots,n}$.
  Elle représente l'évolution à long terme de la grandeur étudiée et traduit l'aspect général de la série.
- Le cycle (ou cycle conjoncturel) $(c_i)_{i=1,\ldots,n}$.
  Il regroupe les variations autour de la tendance avec des alternances de phases d'expansion et de récession.
  Ces phases durent généralement plusieurs années mais n'ont pas de durée fixe.

Formalisation

- Les variations saisonnières $(s_i)_{i=1,\dots,n}$.
  Elles sont liées au rythme imposé par les saisons météorologiques, les activités économiques et sociales,...

  Deux principes à la base de la notion et donc du calcul des $(s_i)_{i=1,\dots,n}$:

  1. Elles sont de nature périodique : il existe un entier p, appelé période, tel que la variation saisonnière s_i se répète à l'identique tous les p temps : $s_i = s_{i+p}$ pour $i \geq 1$.
     La composante saisonnière est entièrement déterminée par ses p premières valeurs $s_1, s_2, \dots, s_p$.
  2. L'influence de ces variations est neutre : $\sum_{j=1}^{p} s_j = 0$.

- La composante résiduelle ou bruit $(e_i)_{i=1,\ldots,n}$.
  Elle correspond aux variations résiduelles qui subsistent quand on a éliminé toutes les autres composantes.
  Elle peut être de différentes natures.
  - Les **fluctuations irrégulières**
    - Elles ont souvent un effet de faible intensité et de courte durée.
    - Elles proviennent d'un grand nombre de petites causes, de nature aléatoire, ce qui veut dire que, dans le cadre purement descriptif qui est le nôtre, elles sont inexpliquées.

- La composante résiduelle ou bruit $(e_i)_{i=1,\dots,n}$.
  Elle correspond aux variations résiduelles qui subsistent quand on a éliminé toutes les autres composantes.
  Elle peut être de différentes natures.
  - Les **fluctuations irrégulières**
  - Les **variations accidentelles**
    - Ce sont des valeurs isolées anormalement élevées ou faibles.
    - Ces variations brusques de la série proviennent d'évènements accidentels de grande ampleur, dus à des accidents importants et généralement explicables (grève, inondation, tempête, crise financière, crise politique, ...).

Composantes principales de la série

On considèrera qu'une série chronologique $(y_i)_{i=1,\ldots,n}$ est la résultante de 3 composantes fondamentales :

- $(f_i)_{i=1,\ldots,n}$ la tendance ou trend (intégrant éventuellement un cycle),
- $(s_i)_{i=1,\ldots,n}$ la composante saisonnière ou saisonnalité de période p, telle que $s_i = s_{i+p}$ pour tout $i \geq 1$,
- $(e_i)_{i=1,\ldots,n}$ la composante résiduelle (intégrant éventuellement des accidents).

Comment combiner ces composantes dans le but de modéliser la série ?

Deux modèles de décomposition

On étudiera 2 modèles de décomposition :

- Le modèle additif.
- Le modèle multiplicatif.

Il s'écrit :

$$y_i = f_i + s_i + e_i \text{ pour } i = 1, \ldots, n$$

avec :

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

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

Le modèle multiplicatif

Seulement pour des variables positives. Il s'écrit :

$$y_i = F_i \times S_i \times E_i \text{ pour } i = 1, \ldots, n$$

avec :

$$\textstyle\prod_{j=1}^{p} S_j = 1 \quad \text{et} \quad \prod_{j=1}^{n} E_j = 1.$$

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

Si l'on prend le logarithme des observations dans le modèle multiplicatif, on obtient

$$z_i = \log(y_i) = \log(F_i) + \log(S_i) + \log(E_i)\,,$$

c'est à dire un modèle additif avec $f_i = \log(F_i)$, $s_i = \log(S_i)$ et $e_i = \log(E_i)$.

On peut noter que les contraintes sont vérifiées: si $\prod_{i=1}^{p} S_i = 1$ alors

$$\sum_{i=1}^{p} s_i = \sum_{i=1}^{p} \log(S_i) = \log\left(\prod_{i=1}^{p} S_i\right) = 0\,.$$

Remarque: Il existe d'autres transformations possibles, autres que logarithmique, pour se ramener à un modèle additif, par exemple $g(x) = x^\gamma$ pour $\gamma \in]0,1[$. Ce sont les transformations de Box-Cox.

Remarque sur le modèle multiplicatif

Le modèle multiplicatif est souvent présenté avec $S_i = (1 + \tilde{s}_i)$ et $E_i = (1 + \tilde{e}_i)$, soit

$$y_i = F_i \times (1 + \tilde{s}_i) \times (1 + \tilde{e}_i) \text{ pour } i = 1, \ldots, n$$

avec les contraintes $\quad \sum_{j=1}^{p} \tilde{s}_j = 0 \quad$ et $\quad \sum_{j=1}^{n} \tilde{e}_j = 0$.
Ces contraintes ne correspondent pas au contraintes naturelles du modèle multiplicatif. Cependant si $\tilde{s}_i \sim 0$, alors $\log(1 + \tilde{s}_i) \sim \tilde{s}_i$, et

$$\sum_{i=1}^{p} \log(S_i) \sim \sum_{i=1}^{p} \tilde{s}_i = 0\,. \qquad (\star)$$

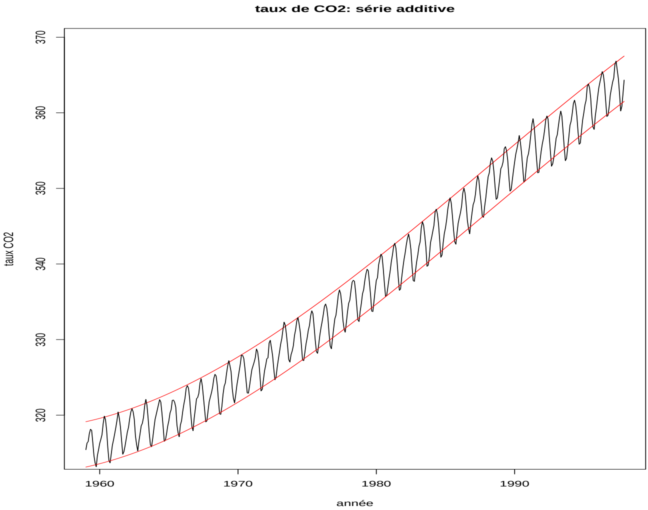
Les contraintes naturelles sont cependant préférables, car dans $(\star)$ on cumule les erreurs d'approximation.
Remarque : Lors de la décomposition du modèle multiplicatif sous R, ce sont les contraintes additives sur $\tilde{s}_i$ et $\tilde{e}_i$ qui sont utilisées.

Méthode de la bande

La méthode dite de la bande, est une méthode graphique qui consiste à tracer la courbe qui passe par les minima sur une période, ainsi que la courbe qui passe par les maxima sur une période.

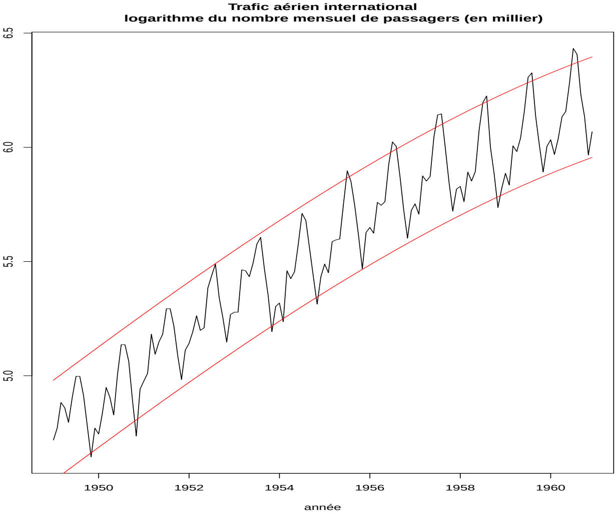
- Si les courbes sont à peu près parallèles, on choisit un modèle additif.
- Sinon, on peut envisager un modèle multiplicatif.

Un exemple de modèle additif



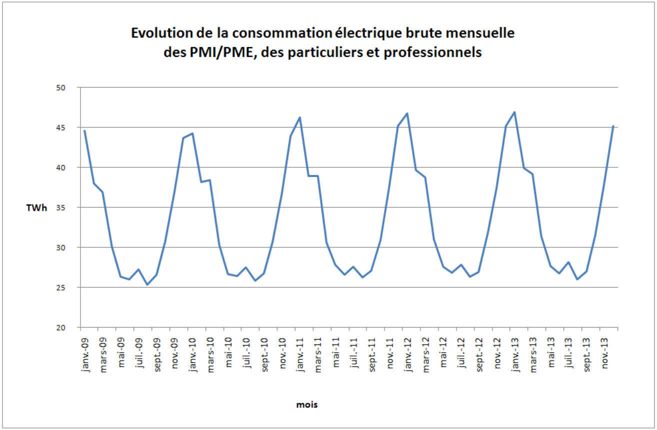
Un exemple de modèle multliplicatif

La série “Trafic aérien" log-transformée est additive:

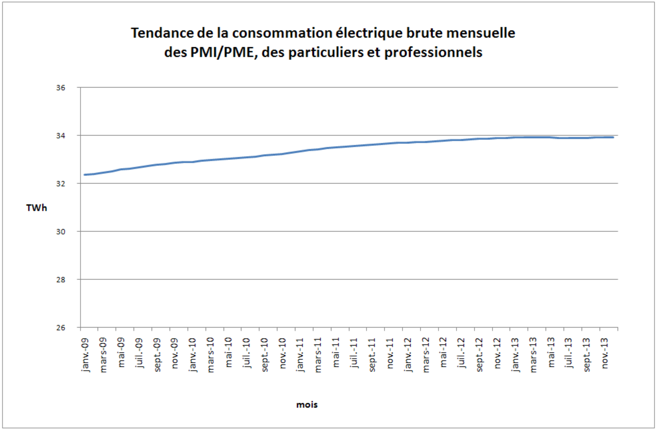


- Description : Identifier les différentes sources de variation de la série (évolution générale du phénomène dans le temps, phénomène périodique, évènements accidentels).

Exemple : Consomme-t-on de plus en plus d'électricité en France ?



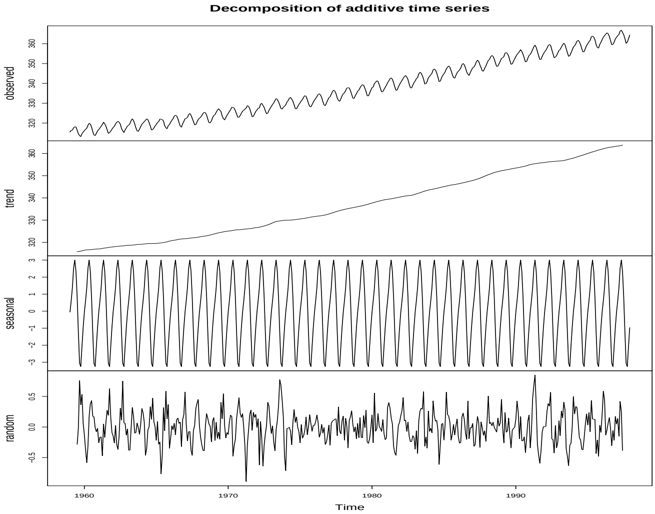
Exemple : Une réponse via la tendance ?



- Description : Identifier les différentes sources de variation de la série (évolution générale du phénomène dans le temps, phénomène périodique, évènements accidentels).
- Explication : Modéliser la série pour comprendre sa structure, la comparer à une autre série.

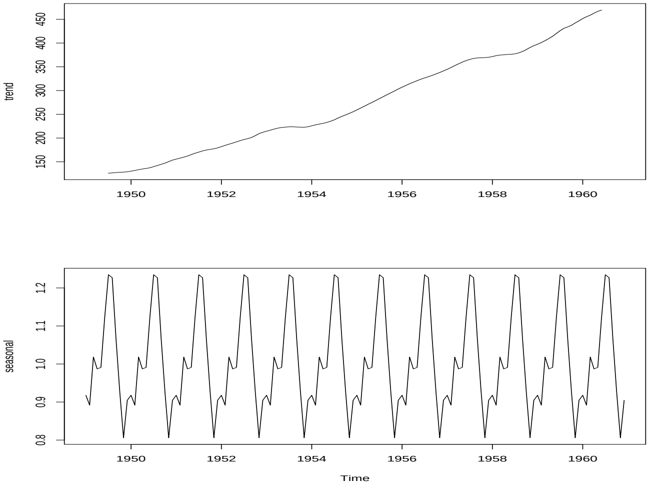
Exemple du taux de CO2

Les différentes composantes de la série



Exemple du trafic aérien

Tendance et composante saisonnière



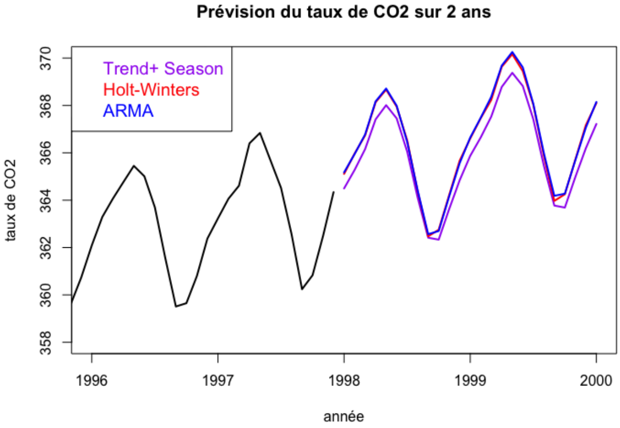
Exemple du trafic aérien

- Pour l'année 1952, y a- t-il eu une augmentation du nombre de passagers entre juin et juillet si l'on omet les variations saisonnières?
- Même question pour les années 1953, 1954, 1955 et 1956.
- Solution à venir . . .

Les différents objectifs de l'analyse des séries chronologiques

- Description : Identifier les différentes sources de variation de la série (évolution générale du phénomène dans le temps, phénomène périodique, évènements accidentels).
- Explication : Modéliser la série pour comprendre sa structure, la comparer à une autre série.
- Prévision : Prévoir les valeurs futures connaissant le passé.

Exemple : prévoir le taux de CO2



Nous verrons en TP ces trois façons de faire des prévisions.

Série corrigées des variations saisonnières

La série corrigée des variations saisonnières $(CVS_i)_{i=1,\dots,n}$ est obtenue en supprimant la composante saisonnière $(s_i)_{i=1,\dots,n}$ du modèle.

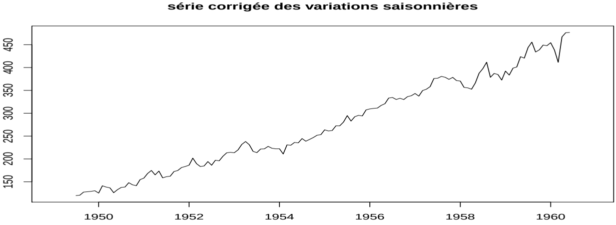
- Pour un modèle additif : $CVS_i = f_i + e_i$.
- Pour un modèle multiplicatif : $CVS_i = F_i \times E_i$.

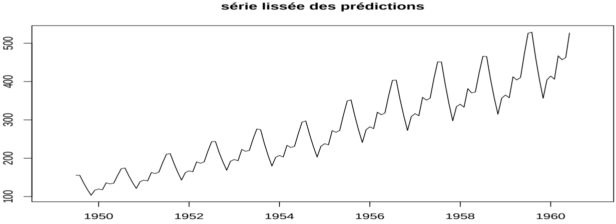
Série lissée des prédictions

La série lissée des prédictions ou série des valeurs ajustées $(\widehat{y}_i)_{i=1,\ldots,n}$ est obtenue en supprimant la composante résiduelle $(e_i)_{i=1,\ldots,n}$ du modèle.

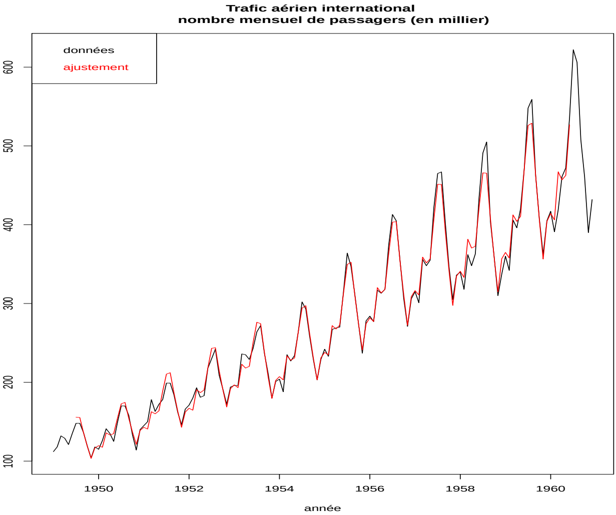
- Pour un modèle additif : $\widehat{y}_i = f_i + s_i$.
- Pour un modèle multiplicatif : $\widehat{y}_i = F_i \times S_i$.

Retour à l'exemple du trafic aérien

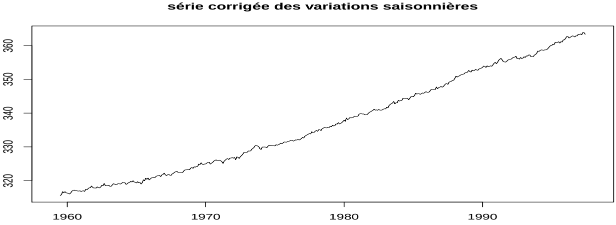


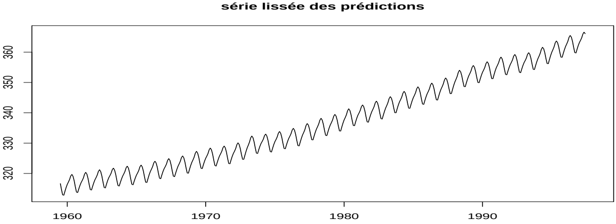


Retour à l'exemple du trafic aérien



Retour à l'exemple du taux de CO2





Retour à l'exemple du taux de CO2