RMT analyse de données et modélisation, 2014

Introduction aux séries chronologiques

David Makowski
INRA

makowski@grignon.inra.fr

1. Définition et domaines d'application
2. Lissage
3. Processus stochastiques
4. Modèle linaire dynamique

1. Définition et domaines d’application

1. Définition et domaines d'application

SERIE CHRONOLOGIQUE

=

Suite d'observations indicées par le temps

$$y_t, \quad t = 1, \ldots, T$$

En général, dates d'observation régulières :

t* = *heure …

jour …

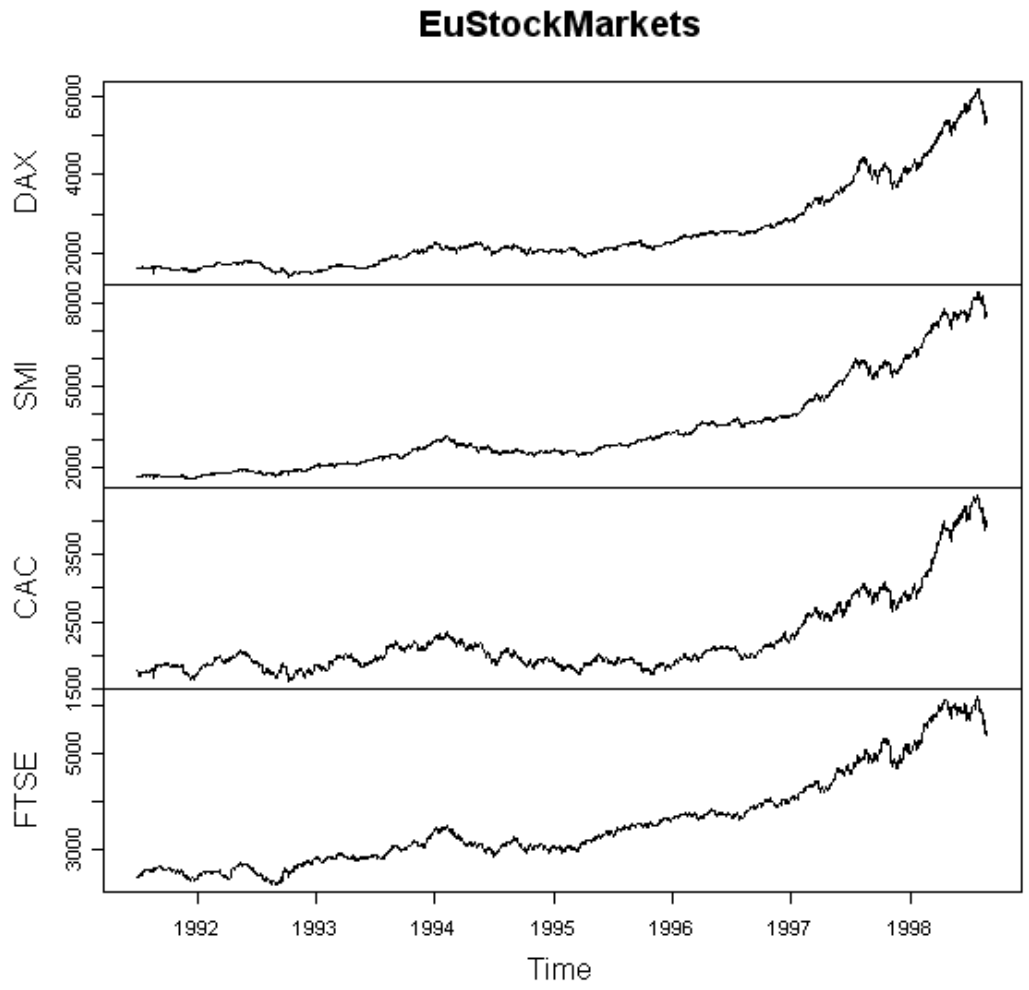
mois …

année …

Domaines d'application

Finance

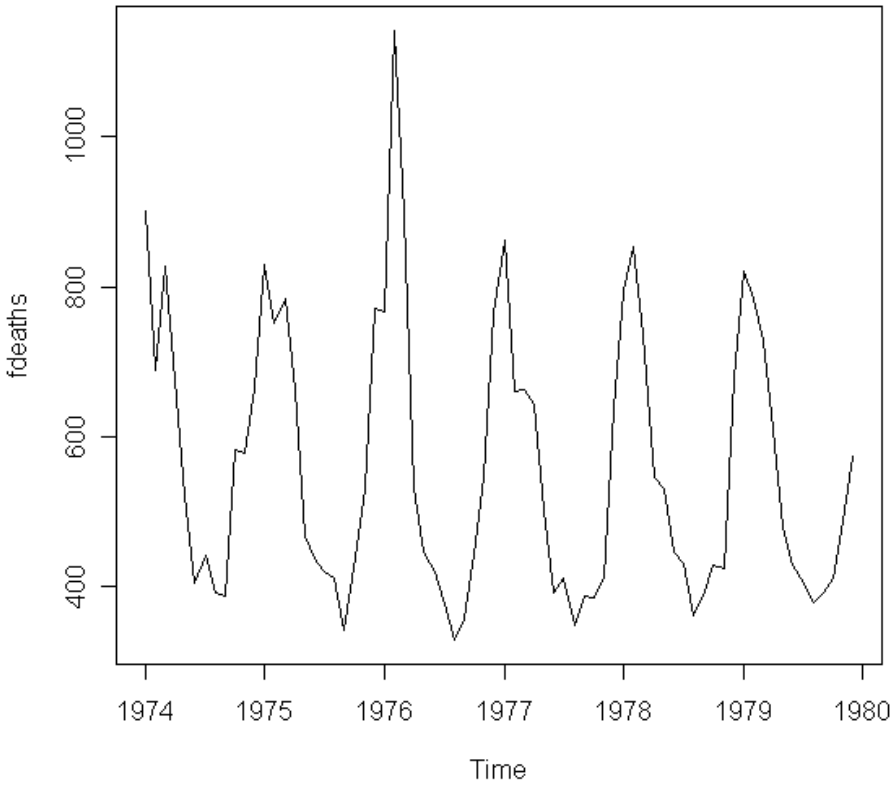
Cours de bourse quotidiens



1. Définition et domaines d'application

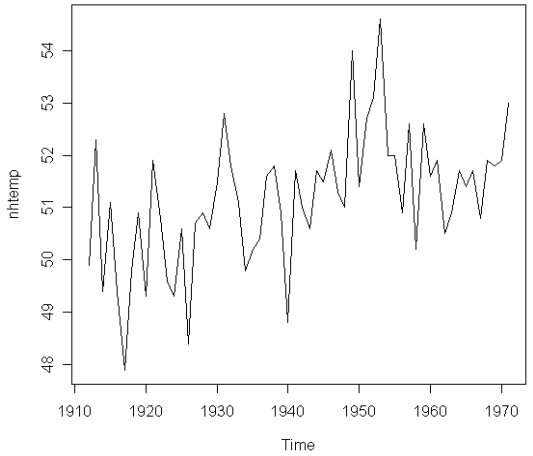
Epidémiologie

Décès mensuels
dus à des
maladies pulmonaires
(UK)



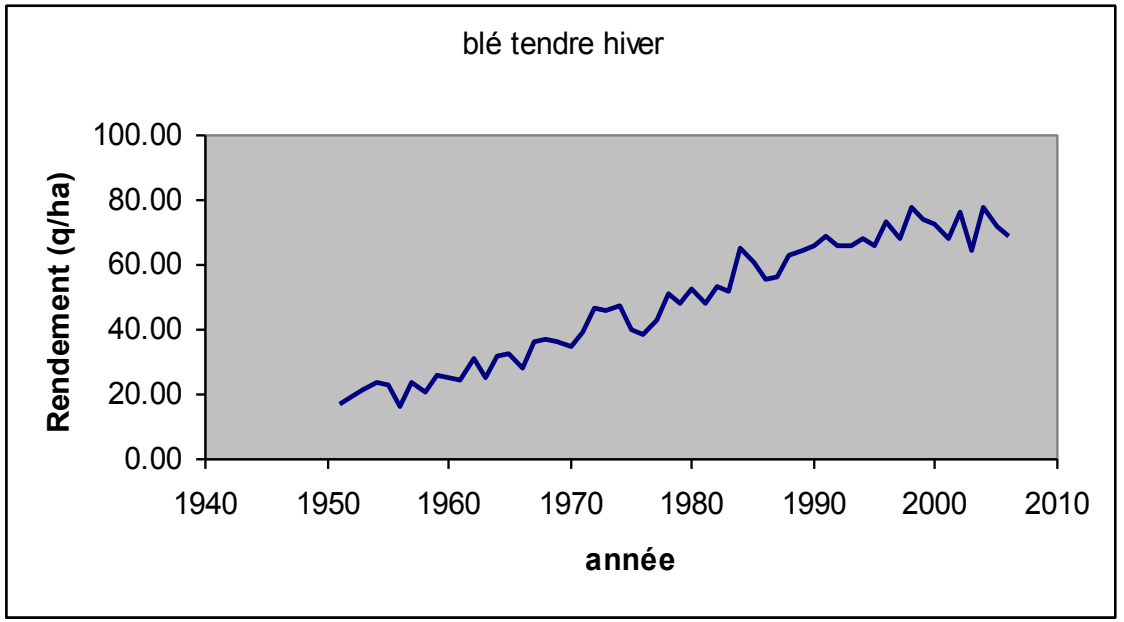
1. Définition et domaines d'application

Température annuelle
Moyenne à New Heaven



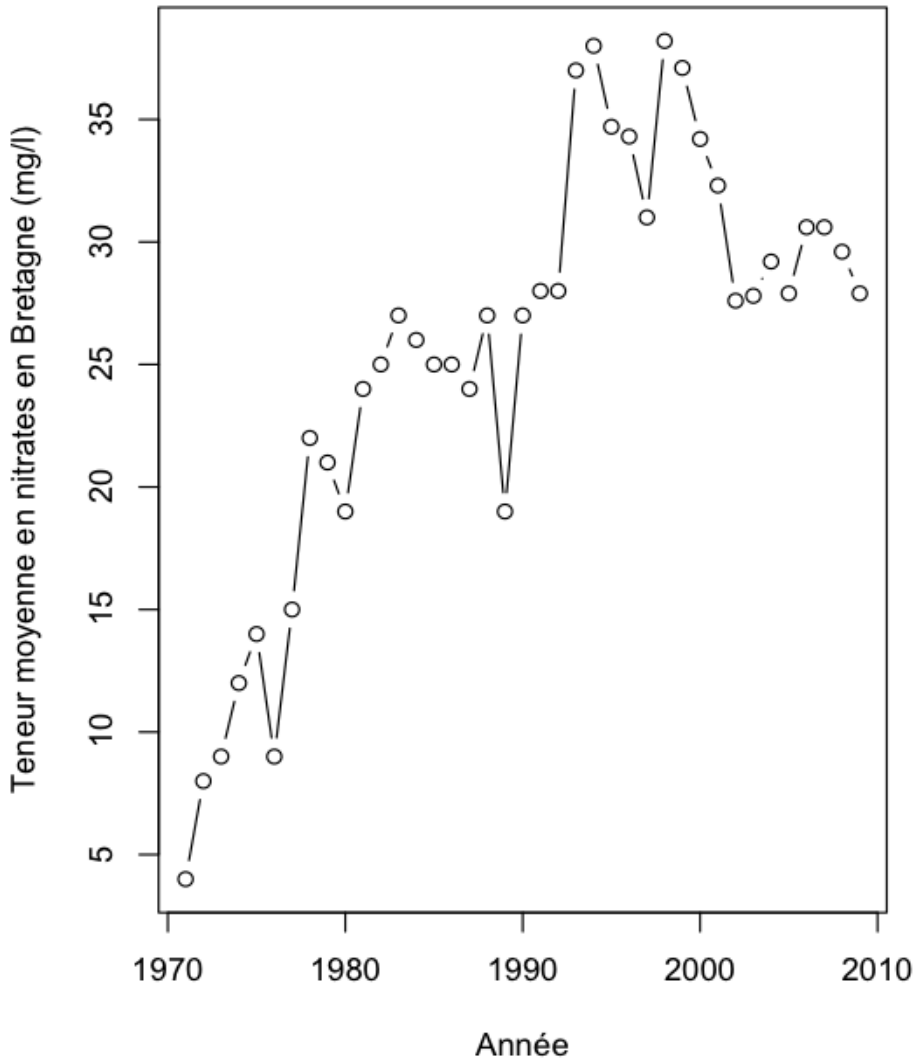
Écologie
Environnement
Météorologie
Hydrologie
etc.

Évolution du rendement de blé tendre d'hiver en France (source: SCEES)



1. Définition et domaines d'application

Moyenne annuelle de la concentration en nitrates dans les eaux superficielles en Bretagne 1971 - 2009



d'après DIREN Bretagne

Objectifs de l'étude d'une série chronologique

- **Analyser un phénomène**

  « *Décrire/comprendre/juger l'évolution de la série* »

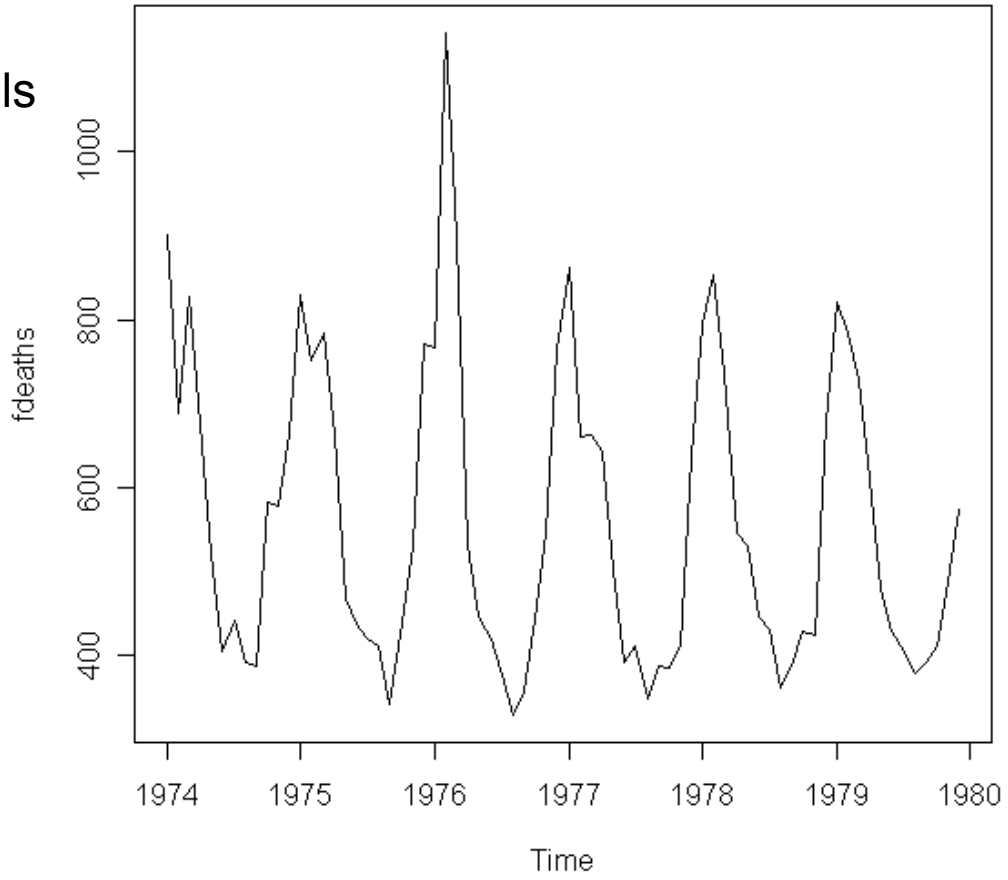
- **Prévision**

  « *Prévoir les valeurs futures de* Y_t »

1. Définition et domaines d'application

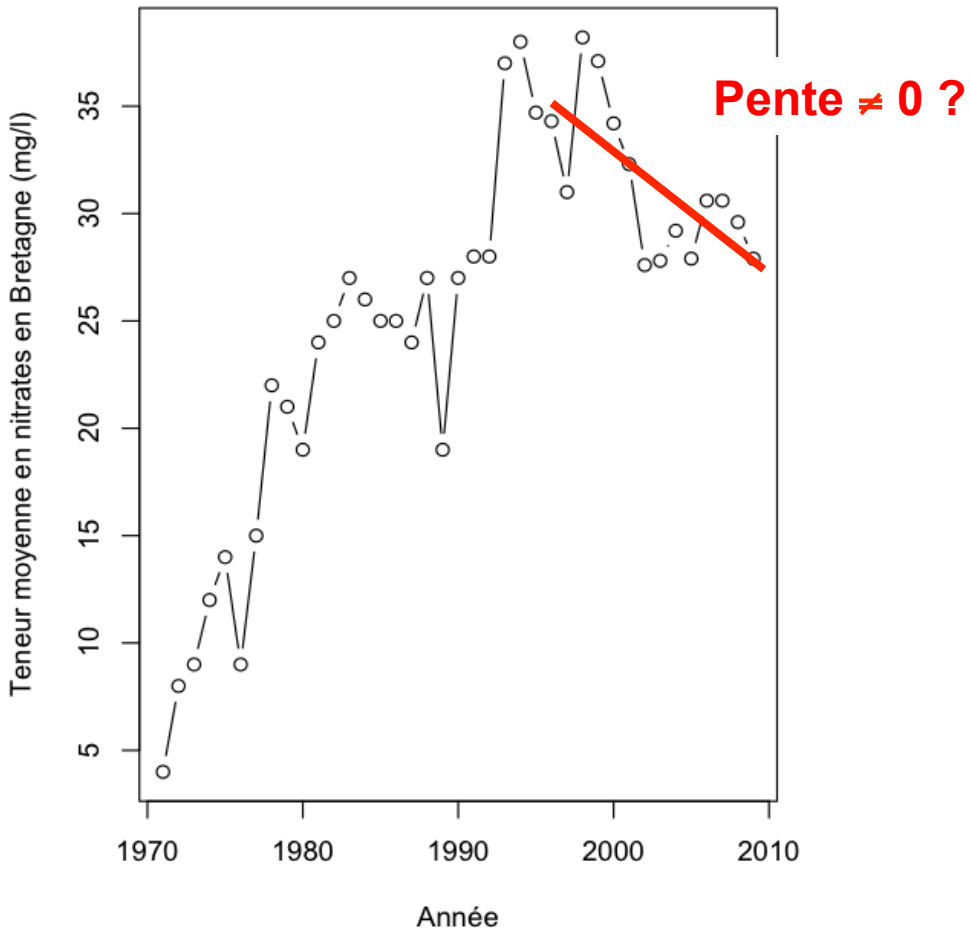
Quels sont les effets des saisons sur le nombre de décès ?

Décès mensuels (UK)



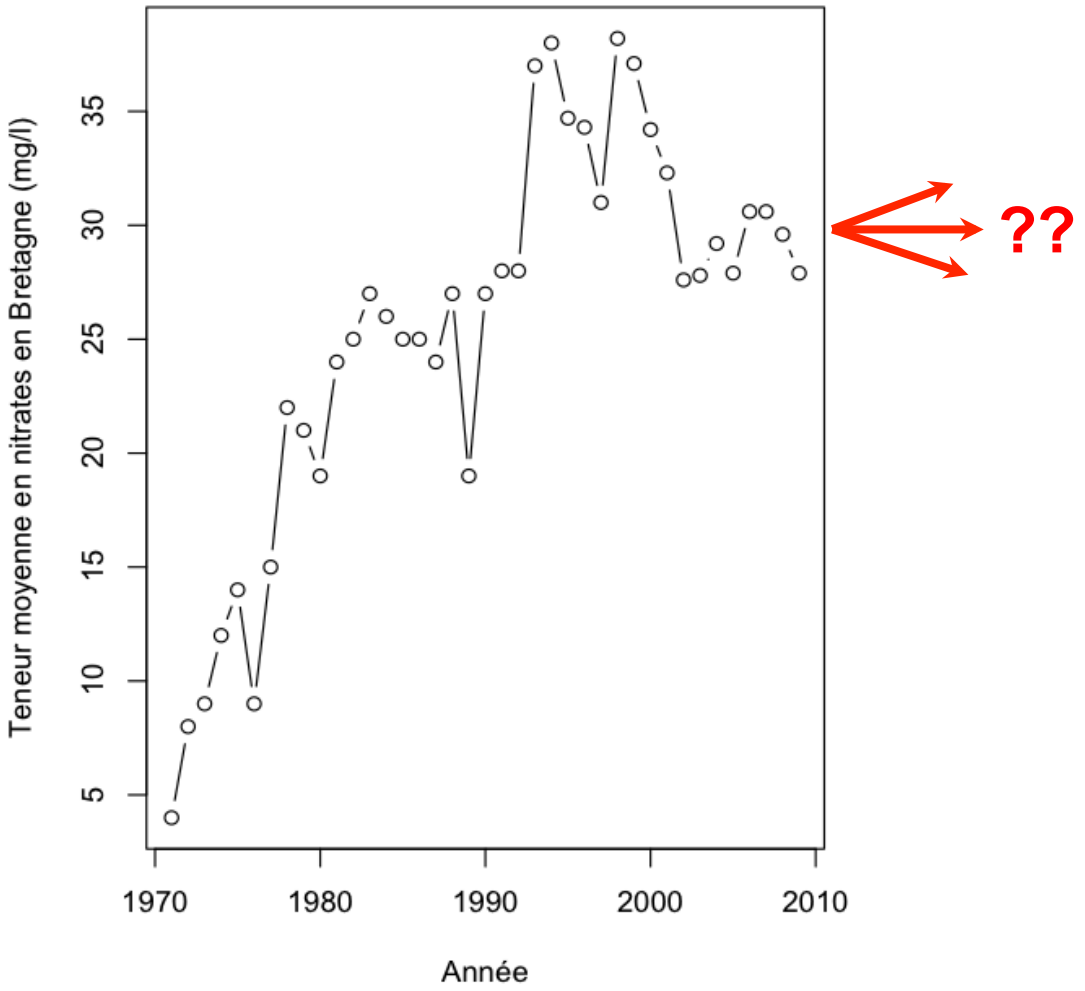
1. Définition et domaines d'application

Existe-t-il une tendance à la baisse après 1995 ?



1. Définition et domaines d'application

Quelle concentration future ?



Les étapes de l'analyse

i. Modélisation (définition des équations du modèle)

ii. Estimation des paramètres

iii. Prévision avec le modèle et évaluation des erreurs

1. Définition et domaines d’application
2. Lissage
3. Processus stochastiques
4. Modèle linéaire dynamique

2. Le lissage

2. Lissage

Principes

- « Lisser » les aléas afin d'observer la tendance
- Ajuster une fonction du temps à des données et extrapoler
- Donner plus de poids aux observations les plus récentes

De nombreuses méthodes

- Moyenne mobile
- Lissage exponentielle
- Méthode de Holt-Winters

…

Moyenne mobile

Permet de décomposer la série initiale en fonction d'une tendance et d'une composante saisonnière

Décomposition additive

Y_t = Tendance(t) + Saison(t) + Résidu(t)

Décomposition multiplicative

Y_t = Tendance(t) * Saison(t) * Résidu(t)

2. Lissage – moyenne mobile

Moyenne mobile *k* (*k* impair)

$$Z_t = \frac{y_{t-m} + \ldots + y_t + \ldots + y_{t+m}}{k}$$

avec $m = (k - 1)/2$

Pour une décomposition additive :

• La moyenne des résidus $y_t - Z_t$ obtenus pour une saison permet d'estimer l'effet de cette saison estim(s_t)

• y_t – estim(s_t) = **série corrigée des variations saisonnières**

2. Lissage – moyenne mobile

Moyenne mobile *k* (*k* pair)

$$Z_t^a = \frac{y_{t-m} + \ldots + y_t + \ldots + y_{t+m-1}}{k}$$

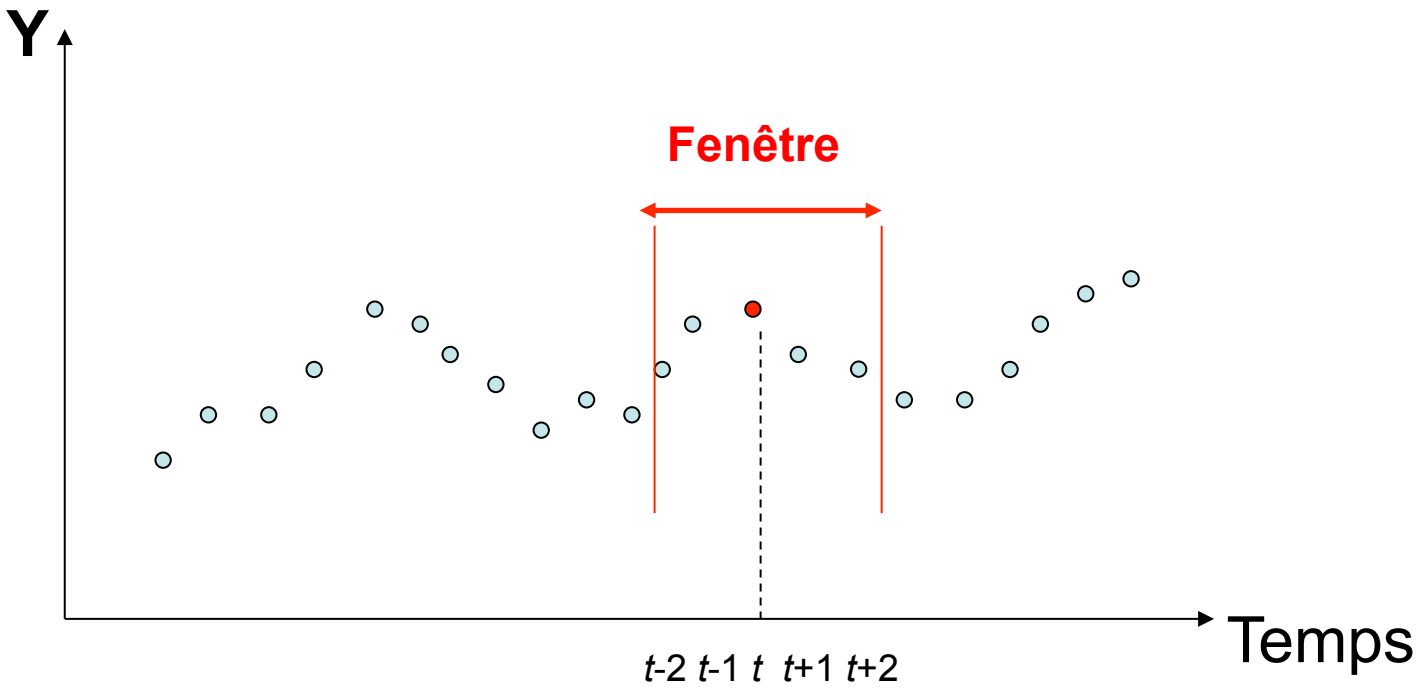
$$Z_t^b = \frac{y_{t-m+1} + \ldots + y_t + \ldots + y_{t+m}}{k}$$

$$Z_t = \frac{Z_t^a + Z_t^b}{2}$$

avec $m = k / 2$

2. Lissage – moyenne mobile

Moyenne mobile 5



Quel est l'effet du coefficient de lissage de la moyenne mobile ?

2. Lissage – moyenne mobile

Expression utile pour la prévision

$$Z_{T+1} = \frac{\sum_{j=0}^{k} y_{T-j}}{k+1}$$

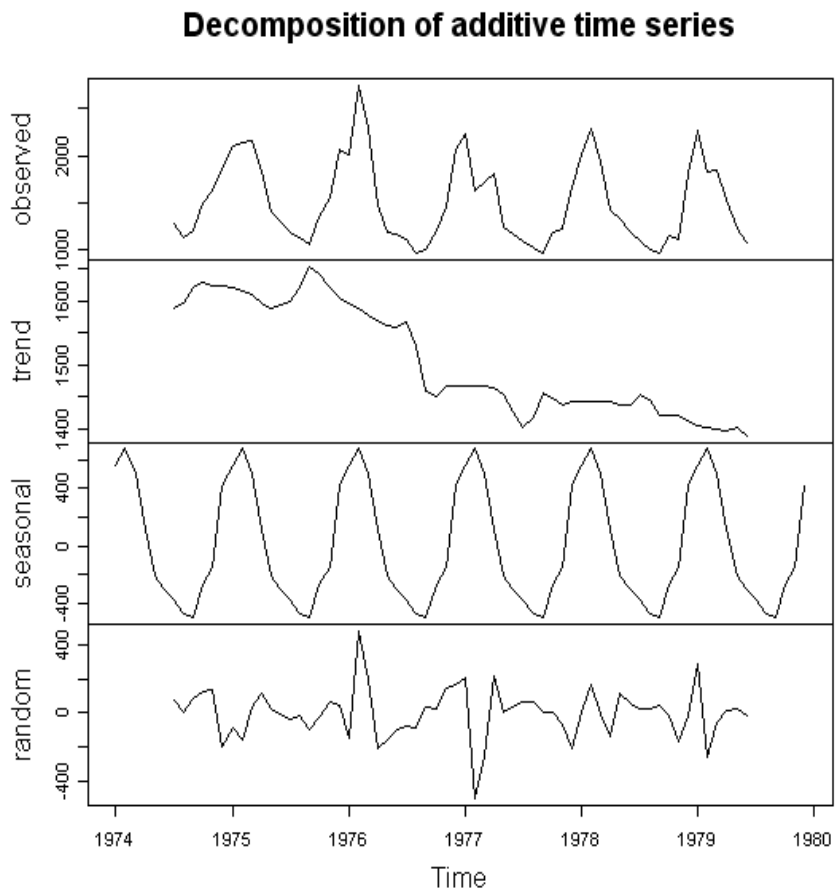
Moyenne mobile - estimation des paramètres

Un paramètre à estimer: *k*

- *k* faible: lissage faible.
- *k* fort: lissage fort.
- Estimation en fonction de la saisonnalité ou en minimisant les erreurs de prédiction.

2. Lissage – moyenne mobile

Exemple sur les données de décès UK

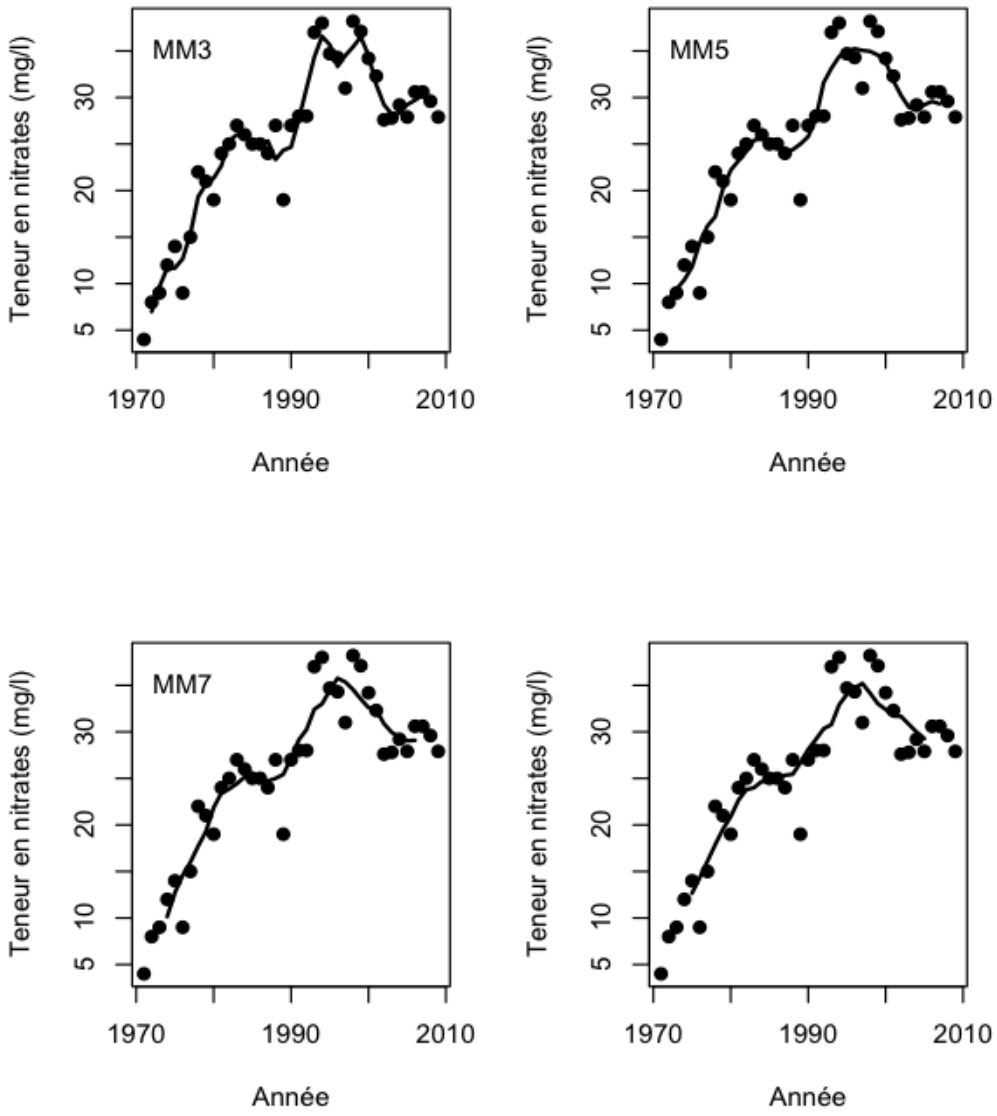


Approche moyennes mobiles

(fonction decompose de R)

2. Lissage – moyenne mobile

Moyenne annuelle de la concentration en nitrates dans les eaux superficielles en Bretagne



Lissage exponentiel simple

La prédiction au temps *t*+1 est une somme pondérée de la prédiction au temps *t* et de la mesure obtenue à cette date.

$$\hat{Y}_{t+1} = \lambda y_t + (1-\lambda)\hat{Y}_t = \hat{Y}_t + \lambda(y_t - \hat{Y}_t)$$

prédiction au temps *t*

erreur

$$0 < \lambda \leq 1$$

$$\hat{Y}_1 = y_1$$

Comme exprimer la prévision en fonction de y_1?

Comment évolue la prévision après la dernière date de mesure ?

Lissage exponentiel simple

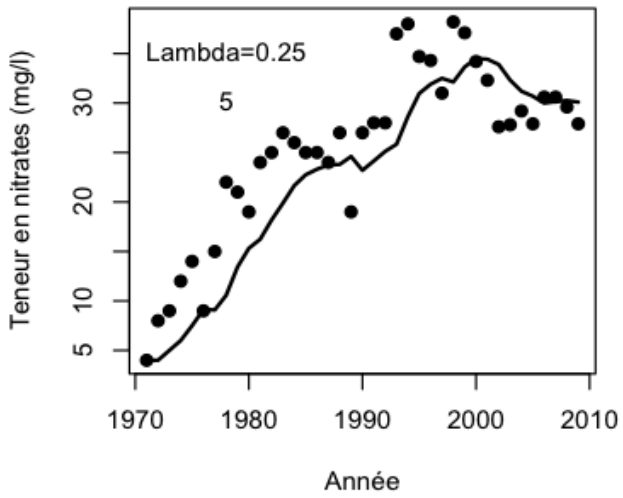
- **Le résultat dépend de la valeur choisie pour la constante de lissage λ**

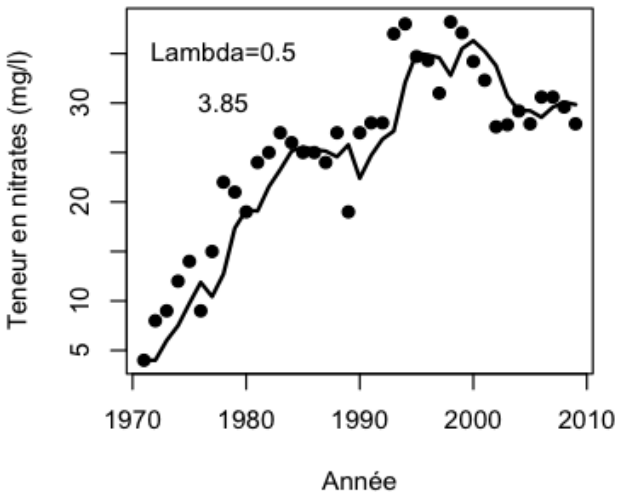
- **La valeur peut être déterminée en cherchant à minimiser les erreurs de prédiction**

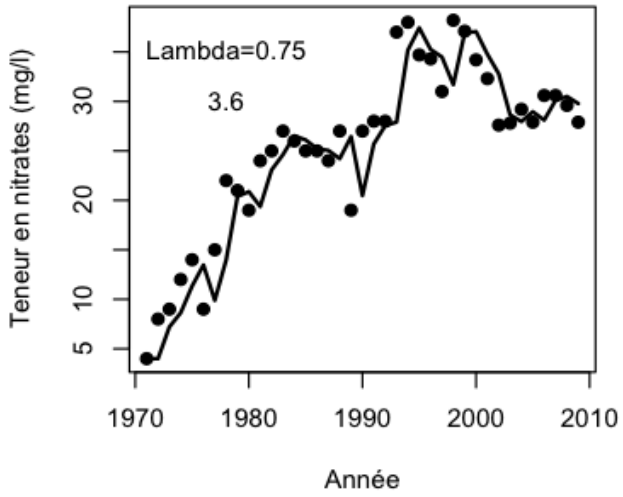
- **La prédiction est constante après la dernière observation**

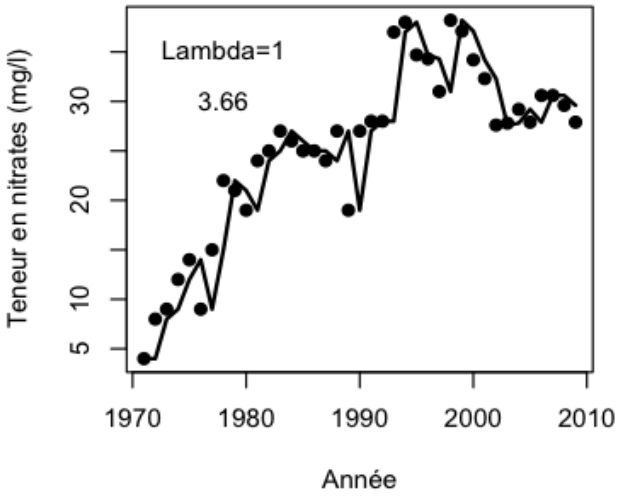
2. Lissage – Lissage exponentiel simple

Prédictions des moyennes annuelles de concentration en nitrates et RMSE



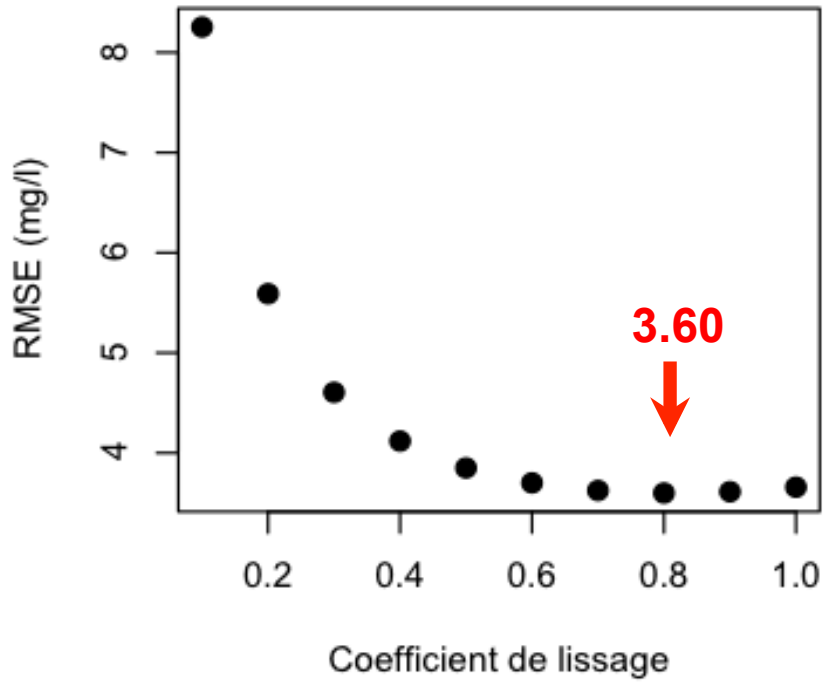


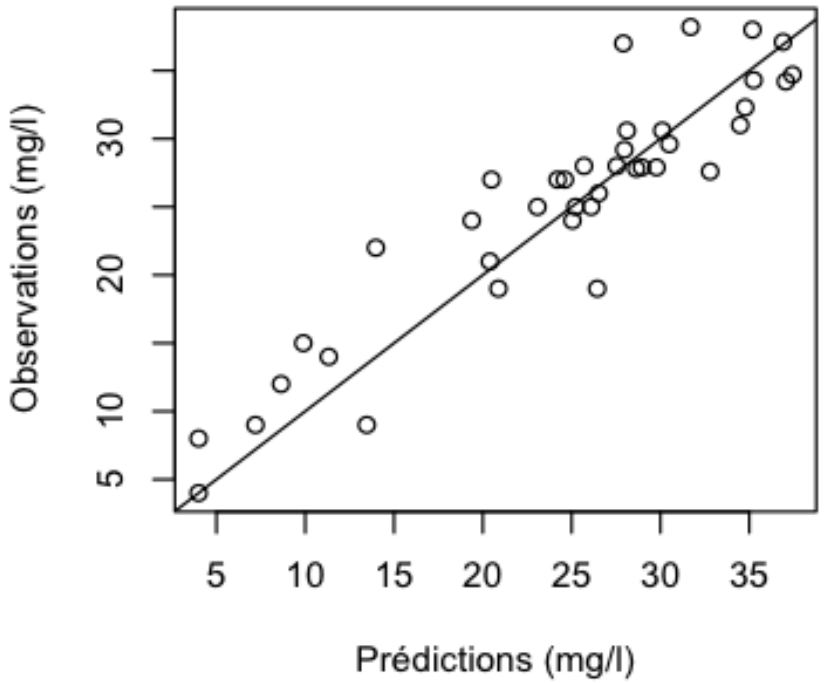




2. Lissage – Lissage exponentiel simple

RMSE en fonction du coefficient de lissage





2. Lissage – Lissage exponentiel simple

Prédiction post 2009 avec λ =0.8: 28.3 mg/l

Très proche de la mesure 2009 (27.9 mg/l)

Méthode de Holt-Winters

- **Permet de prendre en compte une tendance et/ou une saisonnalité**
- **Basée sur deux ou trois équations récursives**
- **Deux ou trois constantes de lissage à estimer**

1. Définition et domaines d’application
2. Lissage
3. Processus stochastiques
4. Modèle linéaire dynamique

3. Processus stochastiques

Définitions

- Décrit l'évolution dans le temps d'une variable aléatoire Y_t

- Un processus stochastique est un mécanisme qui permet de générer des observations y_t sur une période $t = 1, \ldots, T$

- Peut être utilisé pour générer une infinité de jeux de données. Chacun correspond à une réalisation du processus

- Les processus stationnaires constituent un cas particulier important

Processus stationnaire

Stationnarité faible

$$E(Y_t) = \mu$$

$$\text{var}(Y_t) = E\left[(Y_t - \mu)^2\right] = \gamma(0)$$

$$E\left[(Y_t - \mu)(Y_{t-\tau} - \mu)\right] = \gamma(\tau)$$

Les propriétés du processus ne changent pas dans le temps.

Intérêt de la stationnarité

On peut estimer l'espérance, la variance et les autocovariances à partir d'une seule réalisation.

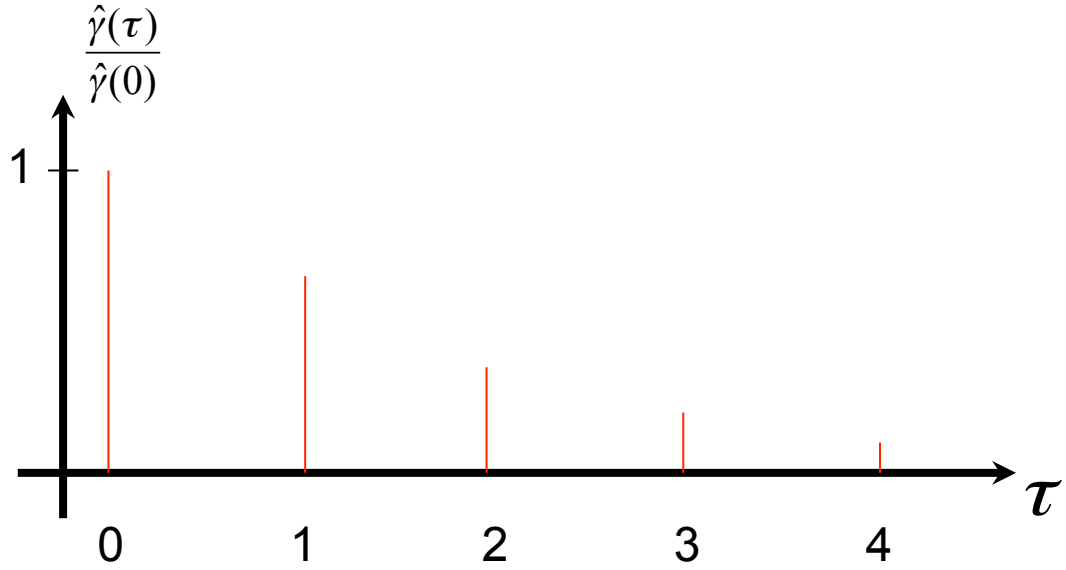
$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

$$\hat{\gamma}(0) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2$$

$$\hat{\gamma}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T}\left(y_t - \bar{y}\right)\left(y_{t-\tau} - \bar{y}\right)$$

Corrélogramme

Graphique des autocorrélations $\frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}$ en fonction de τ



- **Ce graphique est très utile pour identifier la nature du processus.**
- **Les autocorrélations partielles sont également utiles.**

Les différents types de processus stationnaires

- Les modèles « autorégressif » (AR)
- Les modèles « moyenne mobile » (MA)
- Les modèles « mixtes » (ARMA)

Les processus non stationnaires

- L'espérance et/ou la variance évoluent dans le temps

- On va chercher à éliminer la tendance et à stabiliser la variance

Estimer/Eliminer la tendance

- **Différenciation → Modèle ARIMA**
- **Ajustement d'une tendance déterministe → Régression**

Stabiliser la variance

Transformation des données, par exemple en prenant le logarithme des mesures ou une puissance (transformation de Box-Cox).

Démarche de Box-Jenkins (1979)

- Identifier un modèle
- Estimer les paramètres
- Evaluer
- Prévoir

Georges Box (né en 1919 en Angleterre)

George Edward Pelham Box (18 October 1919 –) is one of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.

He was born in Gravesend, Kent, England and originally trained as a chemist. During World War II, he worked on biochemical experiments on the effect of poison gases on small animals for the British Army. He needed statistical advice to analyze the results of his experiments but could not find a statistician who could give him guidance, so he taught himself statistics from available texts. After the war, he enrolled at University College London and obtained a bachelor's degree in mathematics and statistics. He received a Ph.D. from the University of London in 1953.

From 1948 to 1956, Box worked as a statistician for Imperial Chemical Industries (ICI). While at ICI, he took a leave of absence for a year and served as a visiting professor at the University of North Carolina at Chapel Hill. He later went to Princeton University where he served as Director of the Statistical Research Group.

In 1960, Box moved to the University of Wisconsin-Madison to create the Department of Statistics. He was appointed Vilas Research Professor of Statistics (the highest honor accorded to any faculty member at the University of Wisconsin-Madison) in 1980. George Box and Bill Hunter co-founded the Center for Quality and Productivity Improvement at the University of Wisconsin-Madison in 1984. Box officially retired in 1992, becoming an Emeritus Professor.

Throughout his career, George Box has written numerous research papers and published many books. His most important books include Statistics for Experimenters (1978), Time Series Analysis: Forecasting and Control (1979, with Gwilym Jenkins) and Bayesian Inference in Statistical Analysis. (1973, with George C. Tiao). Today, his name is associated with important results in statistics such as Box-Jenkins models, Box-Cox transformations, Box-Behnken designs and numerous others.

He served as President of the American Statistical Association in 1978 and of the Institute of Mathematical Statistics in 1979. He received the Shewhart Medal from the American Society for Quality Control in 1968, the Wilks Memorial Award from the American Statistical Association in 1972, the R. A. Fisher Lectureship in 1974, and the Guy Medal in Gold from the Royal Statistical Society in 1993. He was elected a member of the American Academy of Arts and Sciences in 1974 and a Fellow of the Royal Society in 1979.

The often quoted phrase, "Essentially, all models are wrong, but some are useful", is attributed to George Box.

Box married Joan Fisher, second of Ronald Fisher's five daughters. In 1978, Joan Fisher Box published a very well-received biography of her father. He supervized Lars Pallesen, who was appointed rector (president) of the Technical University of Denmark in 2007.

Wikipedia

Démarche de Box-Jenkins (1979)

i. Identifier un modèle

ii. Estimer les paramètres

iii. Evaluer

iv. Prévoir

i. Identifier un modèle

• Déterminer si la stationnarité est plausible (analyse graphique, test)

• Si ce n'est pas le cas, différencier les données ou ajouter une tendance.

• Déterminer la nature du processus (AR, MA, ARMA) (analyse des corrélogrammes)

• Etape délicate, plus ou moins subjective

Modèle AR(1) (autorégressif d'ordre 1)

$$Y_t = \varphi Y_{t-1} + \xi_t$$

$$E(\xi_t) = 0$$

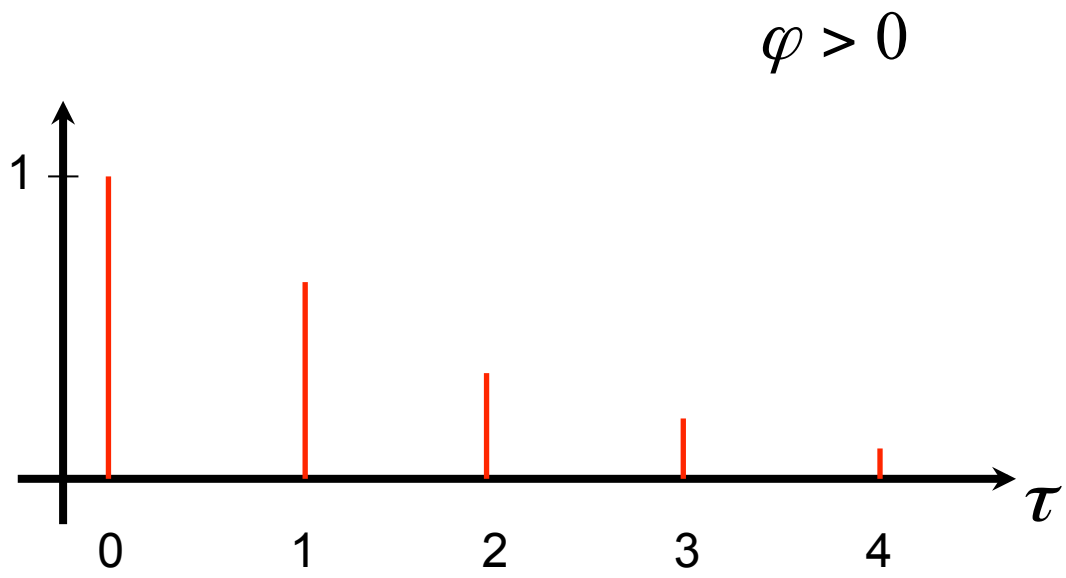
$$\mathrm{var}(\xi_t) = \sigma^2$$

Le processus est stationnaire si $|\varphi| < 1$

$$\gamma(0) = \frac{\sigma^2}{(1-\varphi^2)}$$

$$\rho(\tau) = \varphi^\tau$$

Quelle est l'allure du corrélogramme ?


$\varphi > 0$
1
0
1
2
3
4
τ

ii. Estimer les paramètres

Méthode du maximum de vraisemblance

$$L = prob\left(y_1,...,y_T \left| \varphi_1,...,\varphi_p,\theta_1,...,\theta_q \right. \right)$$

iii. Evaluer le modèle

- **Analyse des résidus**

- **Critères de comparaison de modèles**

Le critère d'Akaïke

$$AIC = -2 \times \log(L) + 2 \times (p + q)$$

Estimation de la distance entre le vrai processus ayant généré les données et le modèle.

iv. Prévoir

Espérance de *Y* conditionnelle aux dernières observations

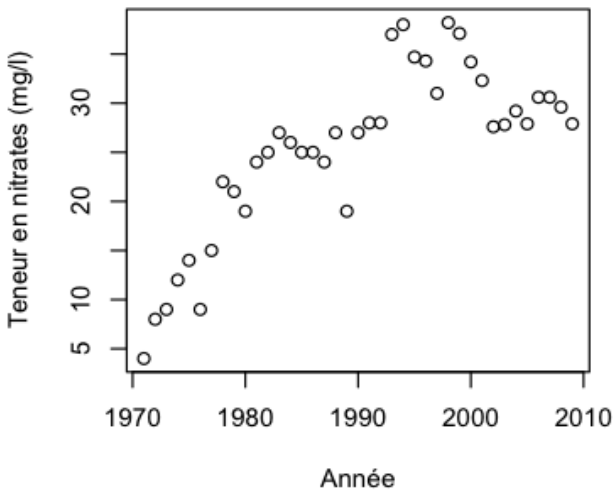
$$\hat{Y}_{T+k} = \hat{E}\left(Y_{T+k} \left| y_T, y_{T-1}, \ldots, y_{T-p} \right.\right)$$

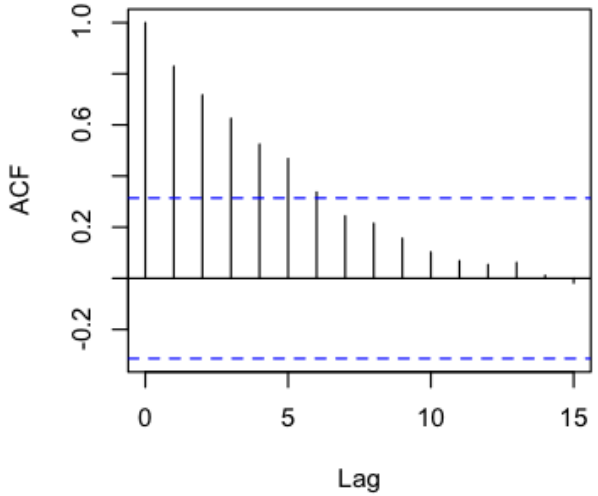
Exemple avec un AR(1)

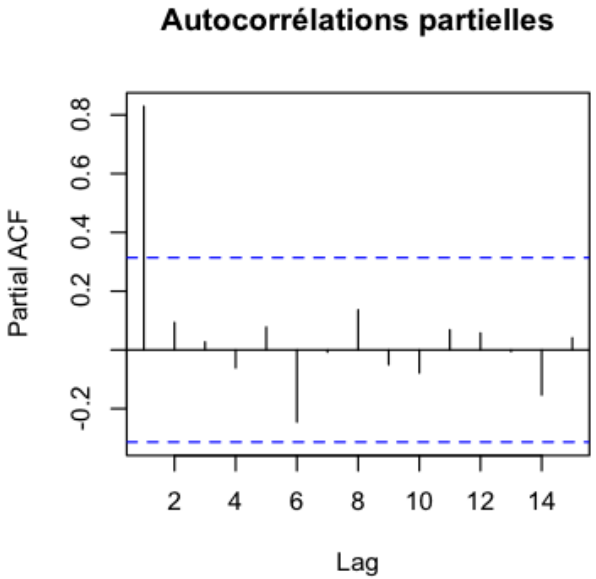
$$\hat{Y}_{T+k} = \hat{\varphi}^k y_T$$

Exemple: prévision de la teneur en nitrate de l’eau

Série initiale

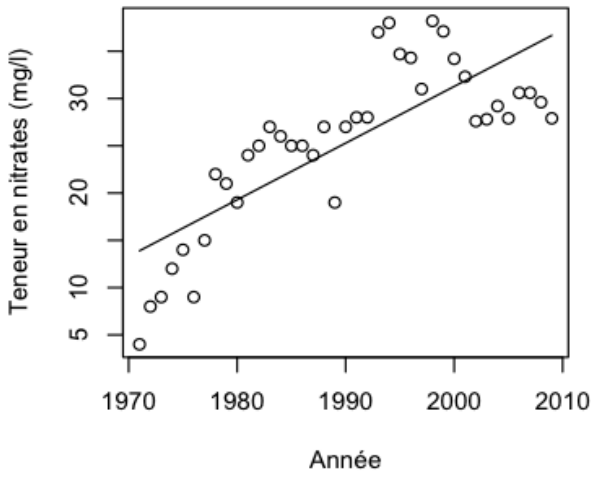


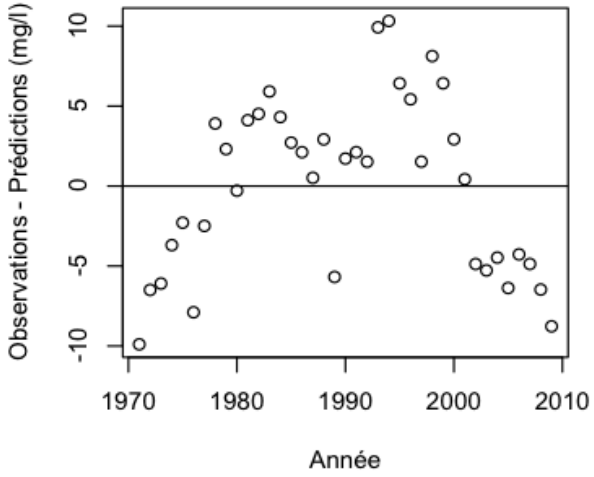


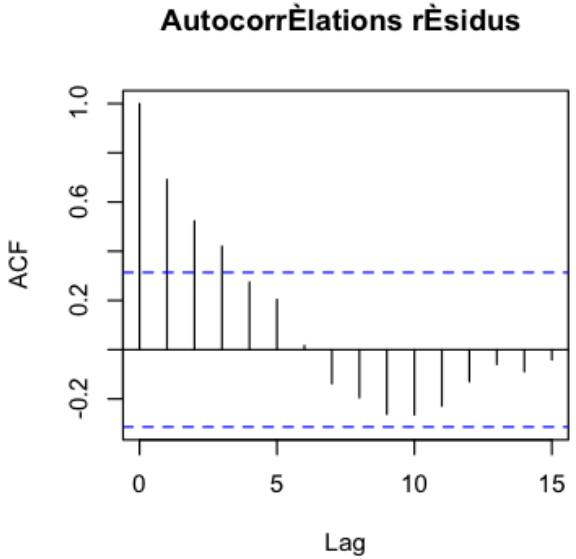


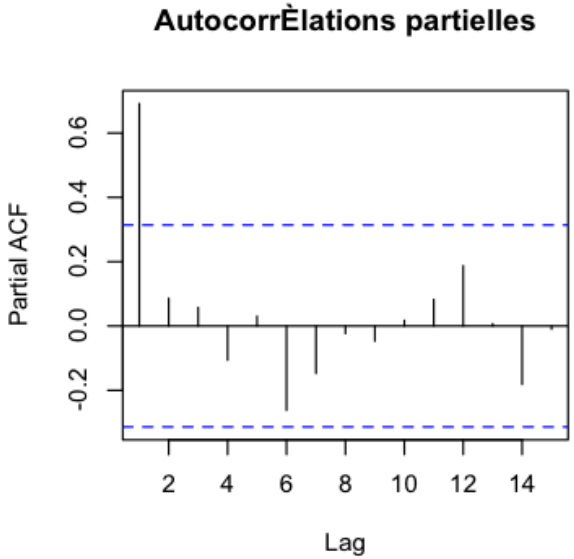
Stationnarité plausible ?

Ajout d'une tendance linéaire



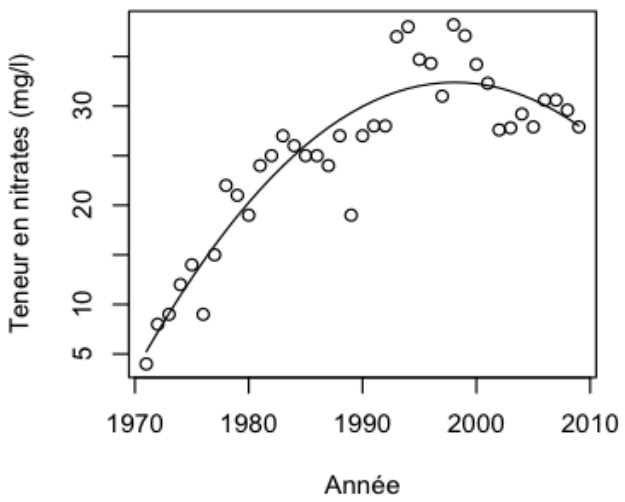


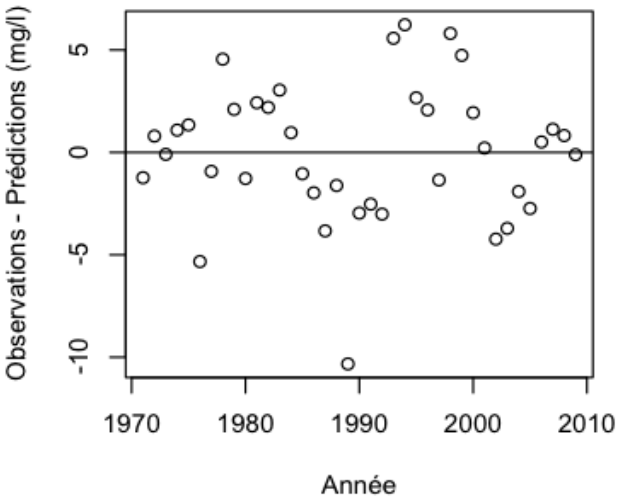


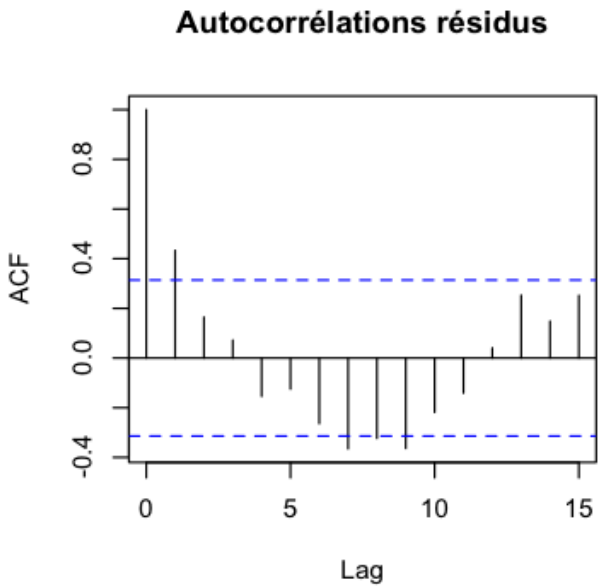


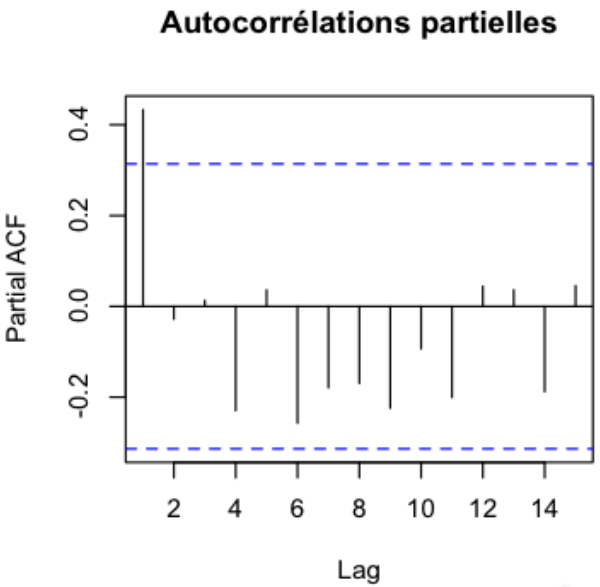
3. Processus stochastiques

Ajout d'une tendance quadratique









Modèle avec une tendance quadratique et des erreurs autocorrélées

$$Y_t = a + b \times t + c \times t^2 + e_t$$

$$e_t = \varphi \times e_{t-1} + \xi_t$$

$$E(\xi_t) = 0$$

$$\text{var}(\xi_t) = \sigma^2$$

Combien de paramètres à estimer ?

3. Processus stochastiques

$$\hat{a} = -149430.09, \quad \hat{b} = 149.61, \quad \hat{c} = -0.04$$

$$\hat{\varphi} = 0.43$$

AIC du modèle quadratique avec résidus autocorrélés = 205

AIC du modèle quadratique avec résidus indépendants = 211

Quel modèle choisir selon le critère AIC ?

Comment évaluer les erreurs de prédiction de ce modèle ?

Quelques variantes utiles

Ajout d'une composante saisonnière

$$Y_t = a + b.t + c.t^2 \quad \textit{(tendance)}$$

$$+ d_{mois(t)} \quad \textit{(composante saisonnière)}$$

$$+ e_t \quad \textit{(aléa)}$$

Variante possible

Exemple: composante saisonnière sinusoïdale

$$d_{mois(t)} = d.\sin(t/12) + e.\cos(t/12)$$

Une autre façon d'éliminer une tendance: la différenciation et le modèle ARIMA

Travailler sur des différences plutôt que sur les valeurs brutes

$$\nabla Y_t = Y_t - Y_{t-1} \longrightarrow \textbf{ARIMA(}\boldsymbol{p}\textbf{, 1, }\boldsymbol{q}\textbf{)}$$

Cette démarche se généralise.

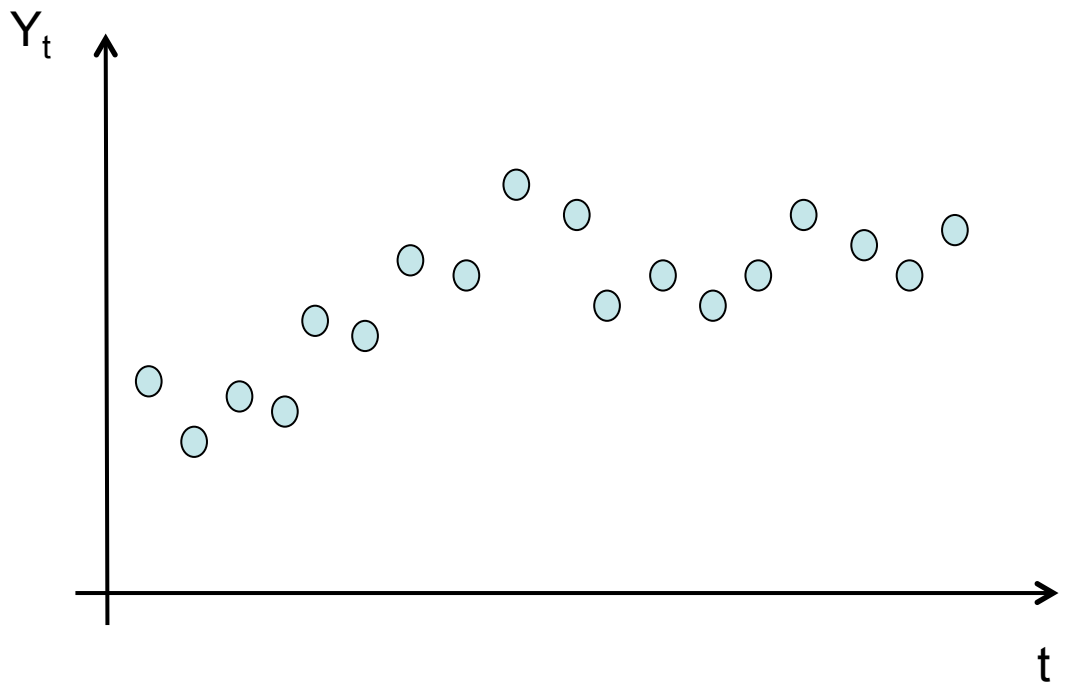
Différenciation d'ordre *d* $\longrightarrow$ **ARIMA(*p*, *d*, *q*)**

1. Définition et domaines d’application
2. Lissage
3. Processus stochastiques
4. Modèle linéaire dynamique

4. Modèles linéaires dynamiques

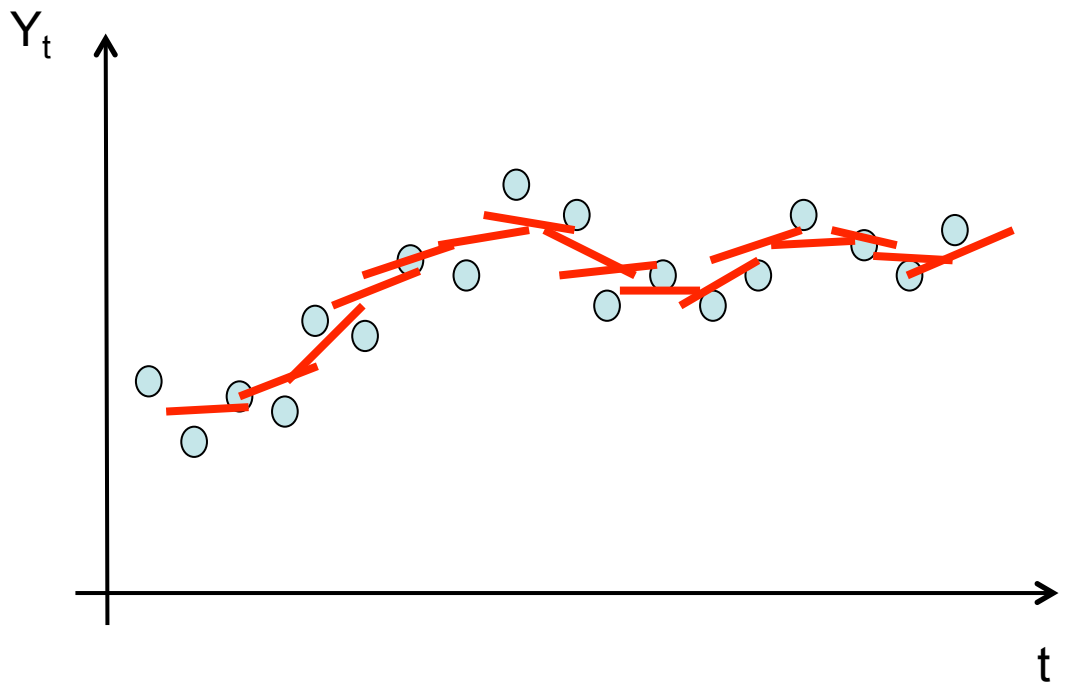
Principe

Modèle linéaire à coefficients variables

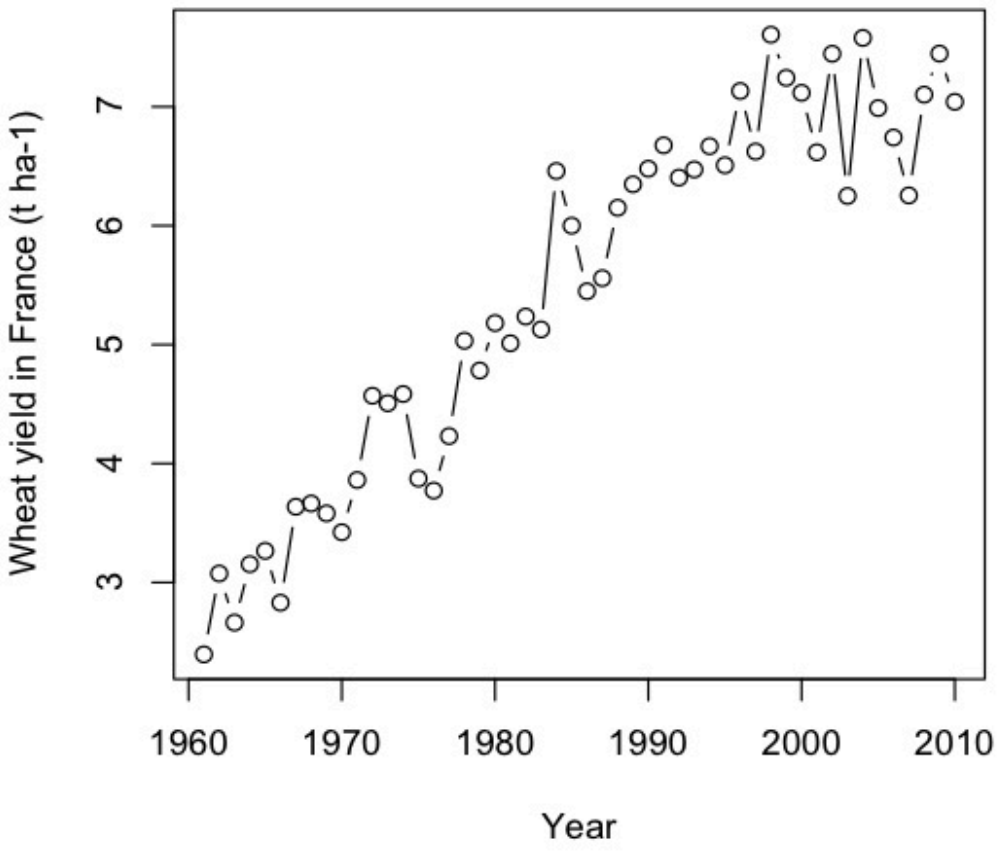


Principe

Modèle linéaire à coefficients variables



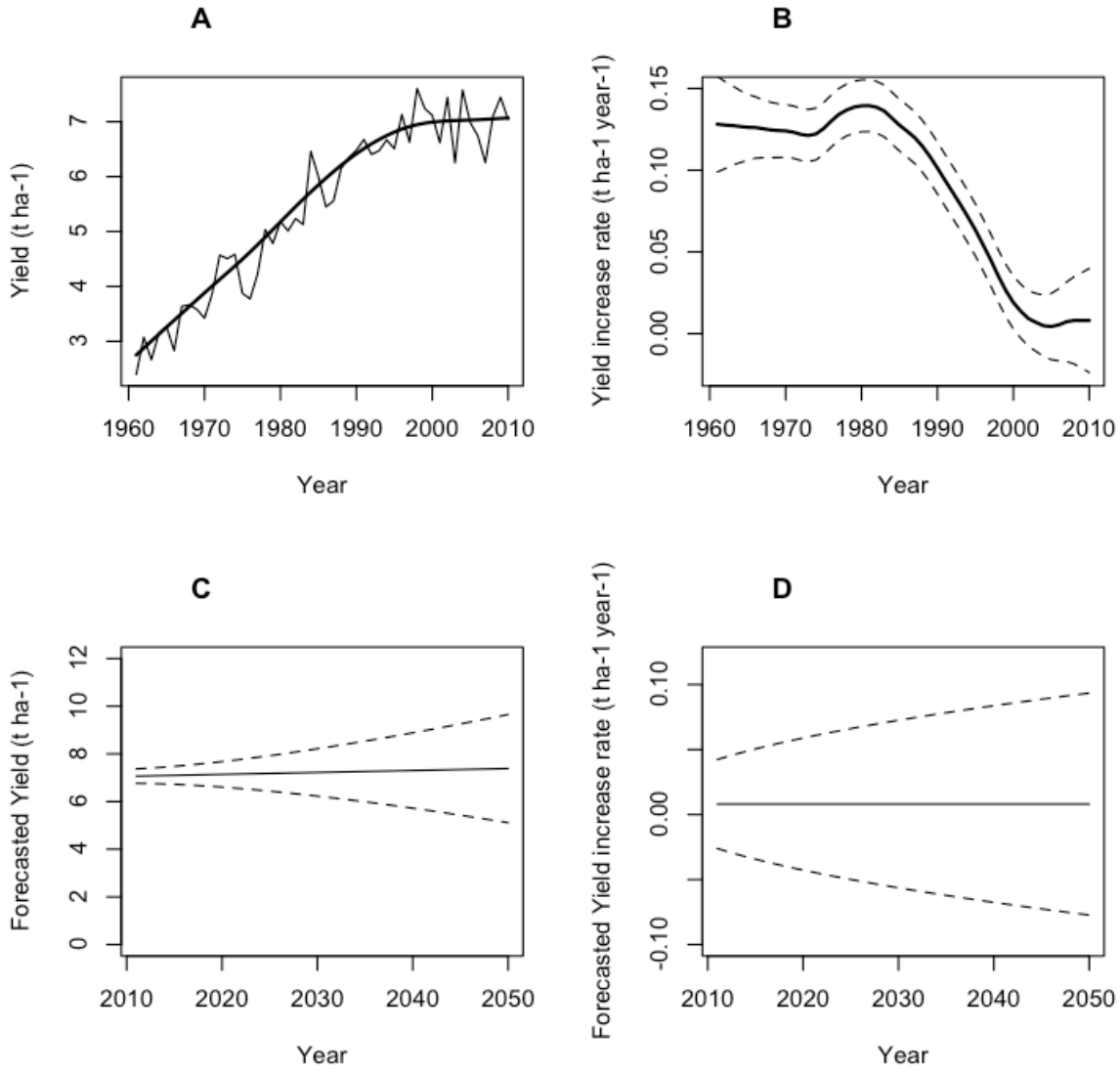
Le rendement plafonne-t-il ?



Données FAO-Stat

4. Modèles linéaires dynamiques

Rendement et accroissement du rendement en France



Autres méthodes

- Modéliser la dynamique de la variance
  - ARCH, GARCH
  - Stochastic volatility modelling
- Séries chronologiques multivariables
- Analyse spectrale
- Analyse spatio-temporelle

Code R

Décomposition et lissage

#Transformation d'un vecteur de données en un object « série chronologique »

ts(data, start, end, frequency…)

#Décomposition d'une série chronologique en une tendance, un effet saisonnier, et un résidu

decompose(x, type = c("additive", "multiplicative"), filter = NULL)

#Autre fonction plus complexe pour la décomposition: stl

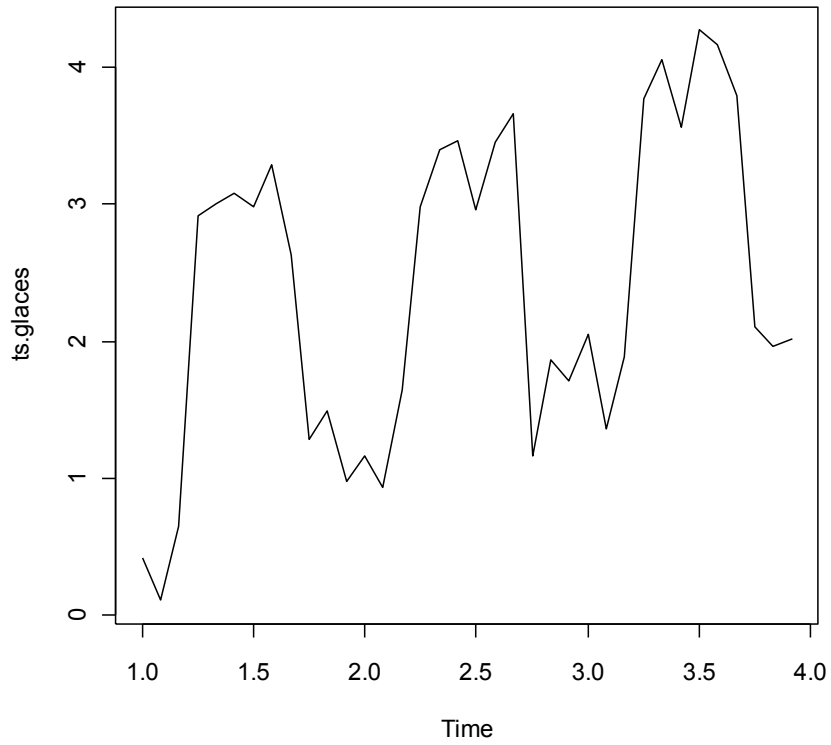
#Application de la moyenne mobile

filter(x, filter=rep(1,5), sides = 2…)

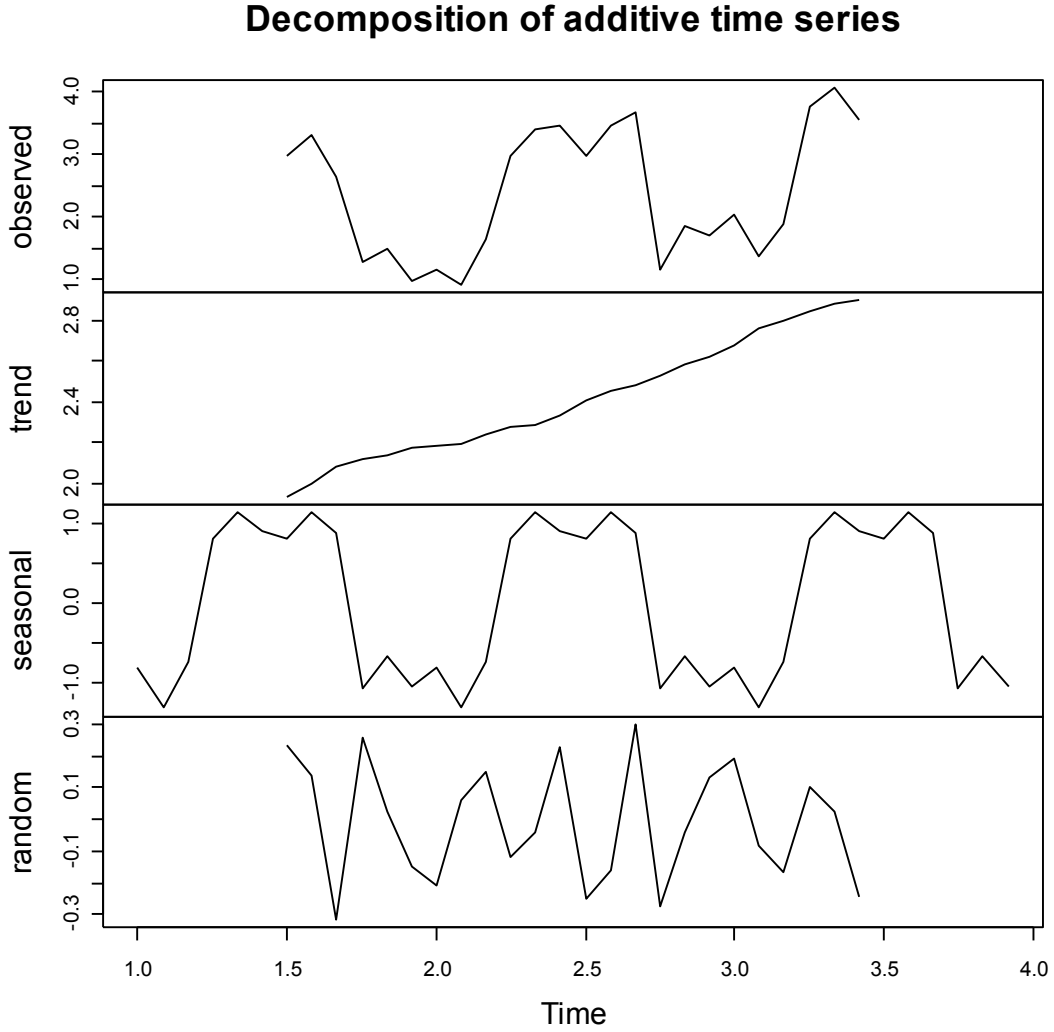
#Méthode de Holt-Winters (package tseries)

HoltWinters(x, beta=FALSE, gamme=FALSE…)

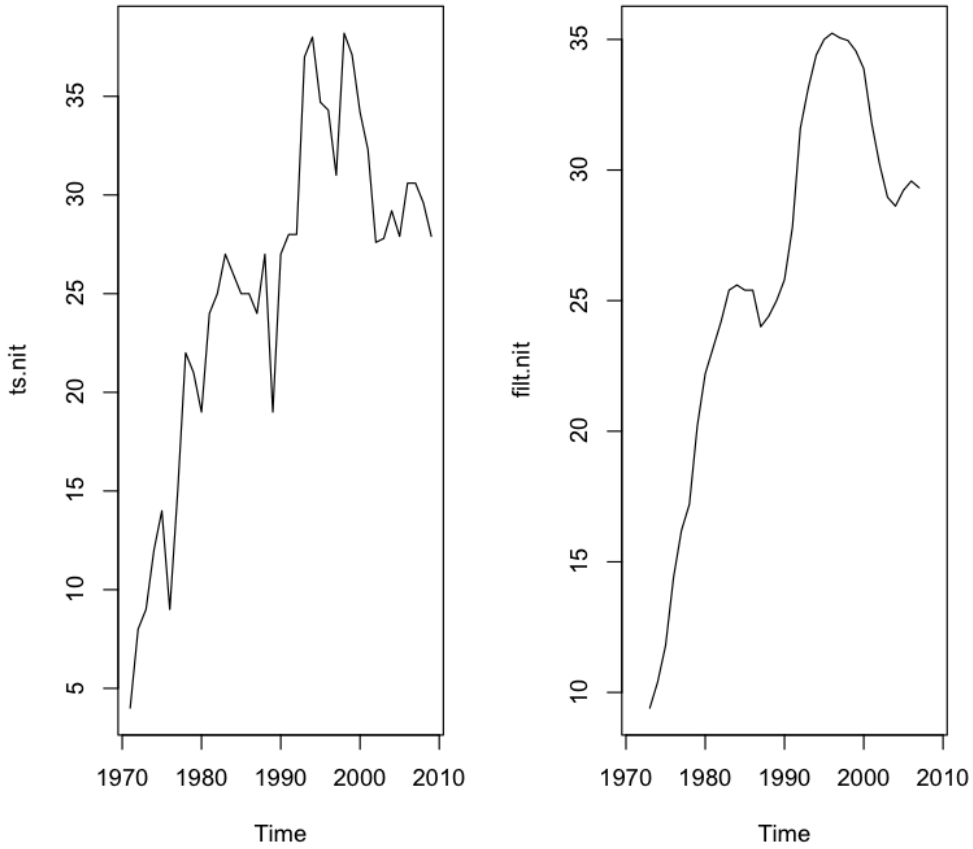
```
Glaces<- c(0.42,0.11,0.65,2.92,3.01,3.08,2.98,3.29,2.63,1.29,1.49,
0.98,1.16,0.93,1.65,2.98,3.40,3.46,2.96,3.45,3.66,1.17,1.87,1.71,2.05,1.36,1.89,3.
77,4.06,3.56,4.27,4.17,3.79,2.11,1.96,2.02)

ts.glaces<-ts(Glaces, start=1, frequency=12)

plot(ts.glaces)
```



```
glaces.dec<-decompose(ts.glaces)

plot(glaces.dec)
```



```
par(mfrow=c(1,2))

TAB<-read.table("Nitrates.txt",sep="\t")

ts.nit<-ts(TAB[,2], start=1971, frequency=1)
plot(ts.nit)

filt.nit<-filter(ts.nit,filter=rep(1/5,5))
plot(filt.nit)
```

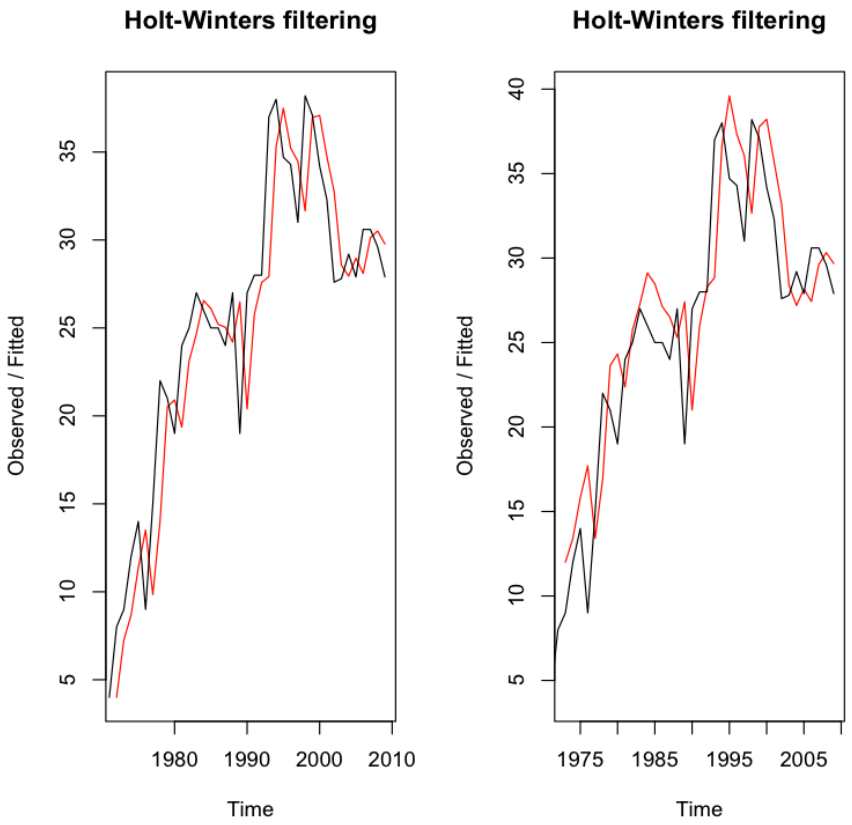


#Lissage exponentiel

```
par(mfrow=c(1,2))

HW.Nit<-HoltWinters(ts.nit, beta=FALSE, gamma = FALSE)
plot(HW.Nit)
```

#Holt-Winters sans composante saisonnière

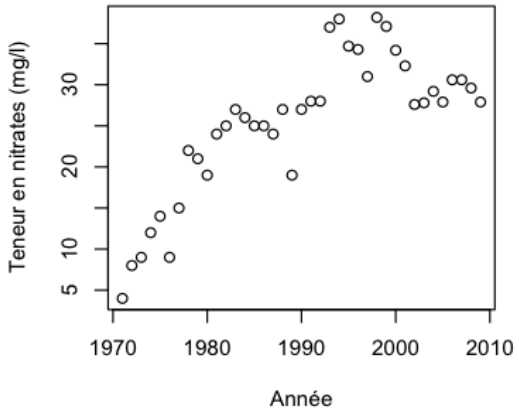
```
HW.Nit<-HoltWinters(ts.nit,  gamma = FALSE)
plot(HW.Nit)
```

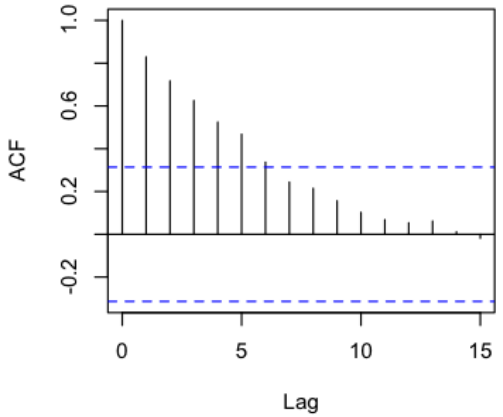


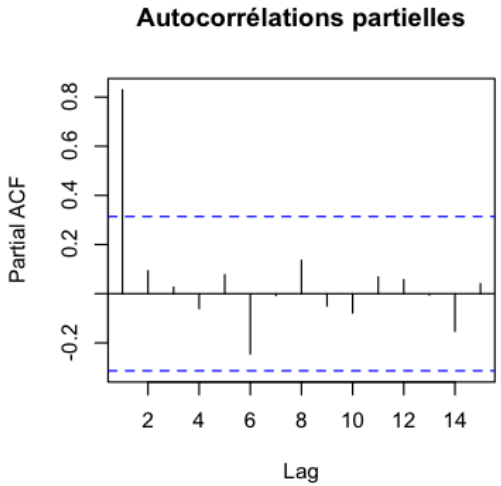
Autocorrélations et processus stochastiques

```
#Graphique des autocorrélations
acf(x, …)

#Régression avec résidus auto-corrélés
gls(Nit~An+An2, data=TAB2,correlation = corAR1(), method="ML")

#ARMA, ARIMA
arima(x, order = c(0, 0, 0), …)
```

```
par(mfrow=c(2,2))
TAB<-read.table("Nitrates.txt",sep="\t")

ts.nit<-ts(TAB[,2], start=1971)
plot(TAB[,1],TAB[,2], xlab="Année", ylab="Teneur en nitrates (mg/l)")

acf(ts.nit, main="Autocorrélations")
acf(ts.nit, type=c("partial"),
 main="Autocorrélations partielles")
```





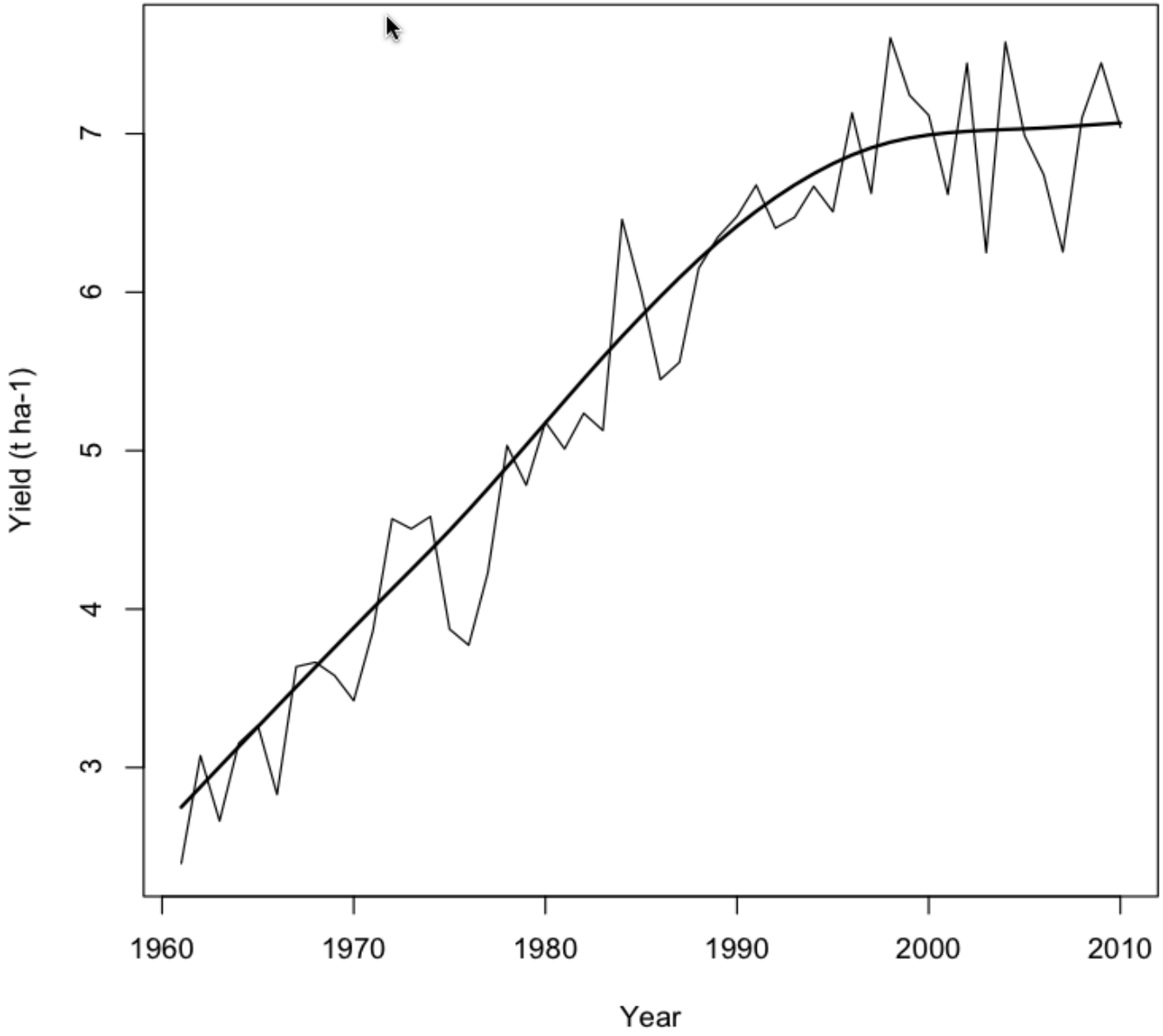


```
par(mfrow=c(2,2))
library(nlme)

#Régression quadratique
Nit<-TAB[,2]
An<-TAB[,1]
An2<-TAB[,1]^2
TAB2<-data.frame(An, An2, Nit)

#Régression quadratique avec processus auto-régressif

reg.1<-gls(Nit~An+An2, data=TAB2,correlation = corAR1(), method="ML")
print(summary(reg.1))
```

Modèle linéaire dynamique

```
#Lecture des données

TAB_Yield<-read.table("YieldFAO_France.txt",header=T)

Year<-TAB_Yield[,1]
Yield<-TAB_Yield[,2]/10000

#Definition du modèle

MyModel<-function(x) {
        return(dlmModPoly(2, dV=exp(x[1]), dW=c(exp(x[2]), exp(x[3]))))
        }

#Estimation des paramètres du modèle

fitMyModel<-dlmMLE(Yield,parm=c(0,0,0), build=MyModel, hessian=T)
print(fitMyModel)
```

```
FittedModel<-MyModel(fitMyModel$par)

#Filtrage
YieldFilter<-dlmFilter(Yield, FittedModel)

#Lissage
YieldSmooth<-dlmSmooth(Yield, FittedModel)

#Plot
plot(Year,Yield,ylab="Yield (t ha-1)", type="l",lwd=1)
lines(Year,YieldSmooth$s[,1][-1]+YieldSmooth$s[,2][-1],lwd=2)

#Prediction
foreYield<-dlmForecast(YieldFilter,nAhead=40)
```


Yield (t ha-1)
3
4
5
6
7
1960
1970
1980
1990
2000
2010
Year

Quelques références

Mots clés en anglais: time series, forecasting

Livres:

Brockwell P.J., David R.A. 2002. Introduction to time series and forecasting. Springer.

Harvey A.C. 1991. Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Makowski D, Monod H. 2011. Analyse statistique des risques agro-environnementaux. Springer.

Petris G. 2010. A R package for dynamic linear models. Journal of Statistical Software 36, 1-14.

Venables W.N., Ripley B.D. 2002. Modern applied statistics with S. Springer. Chapitre 14.

Vose D. 2000. Risk analysis, a quantitative guide. Wiley (2nd edition). Chapitre 12.