Serie Storiche

Laboratorio di Data Science in Economia CLEBA

Roberto Benedetti

Dipartimento di Economia, email benedett@unich.it

Argomanti trattati

- Concetti basilari
- I modelli ARMA
- La nonstazionarietà
- La previsione
- Applicazioni

SERIE STORICHE CON R

- ALCUNE OPERAZIONI PREVISTE NELLE STATISTICHE DI BASE DI R
- Installare PACKAGE : tseries

SERIE STORICHE CON R

Book
Springer Texts in Statistics
2011

Time Series Analysis and Its Applications

With R Examples

Authors: Robert H. Shumway, David S. Stoffer

ISBN: 978-1-4419-7864-6 (Print) 978-1-4419-7865-3 (Online)

DISPENSA di VITO RICCI:

https://cran.r-project.org/doc/contrib/Ricci-ts-italian.pdf

Definizione di serie storica

Definizione operativa non formale:

Una serie storica è una sequenza ordinata di osservazioni, dove l'ordine è dato dal tempo

I valori possono essere quantitativi o qualitativi

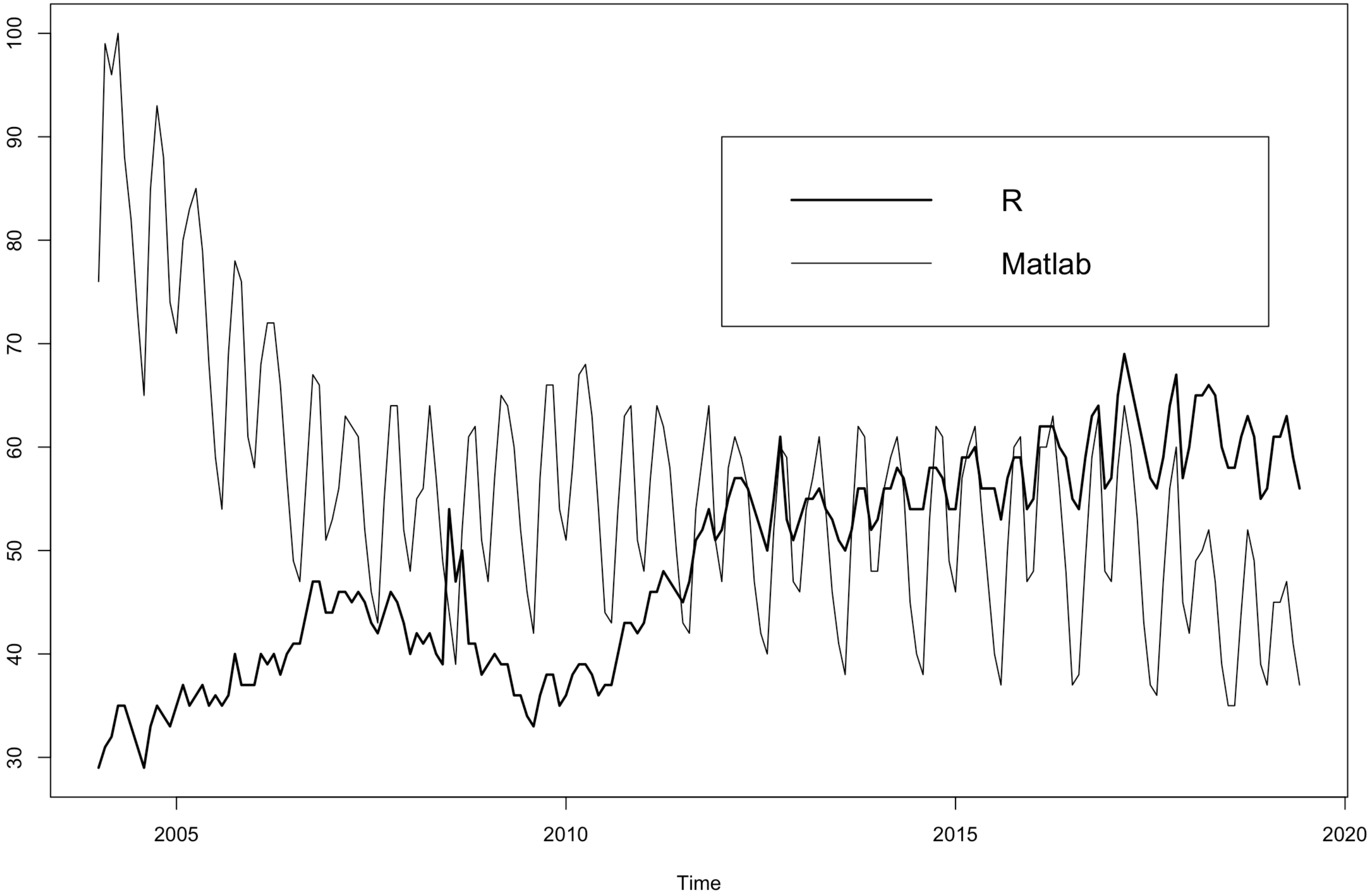
La serie storica può essere di natura discreta (nel tempo non nei valori) o continua, e può essere osservata ad intervalli regolari o no.

Noi analizziamo solo serie storiche discrete osservate ad intervalli regolari di valori quantitativi

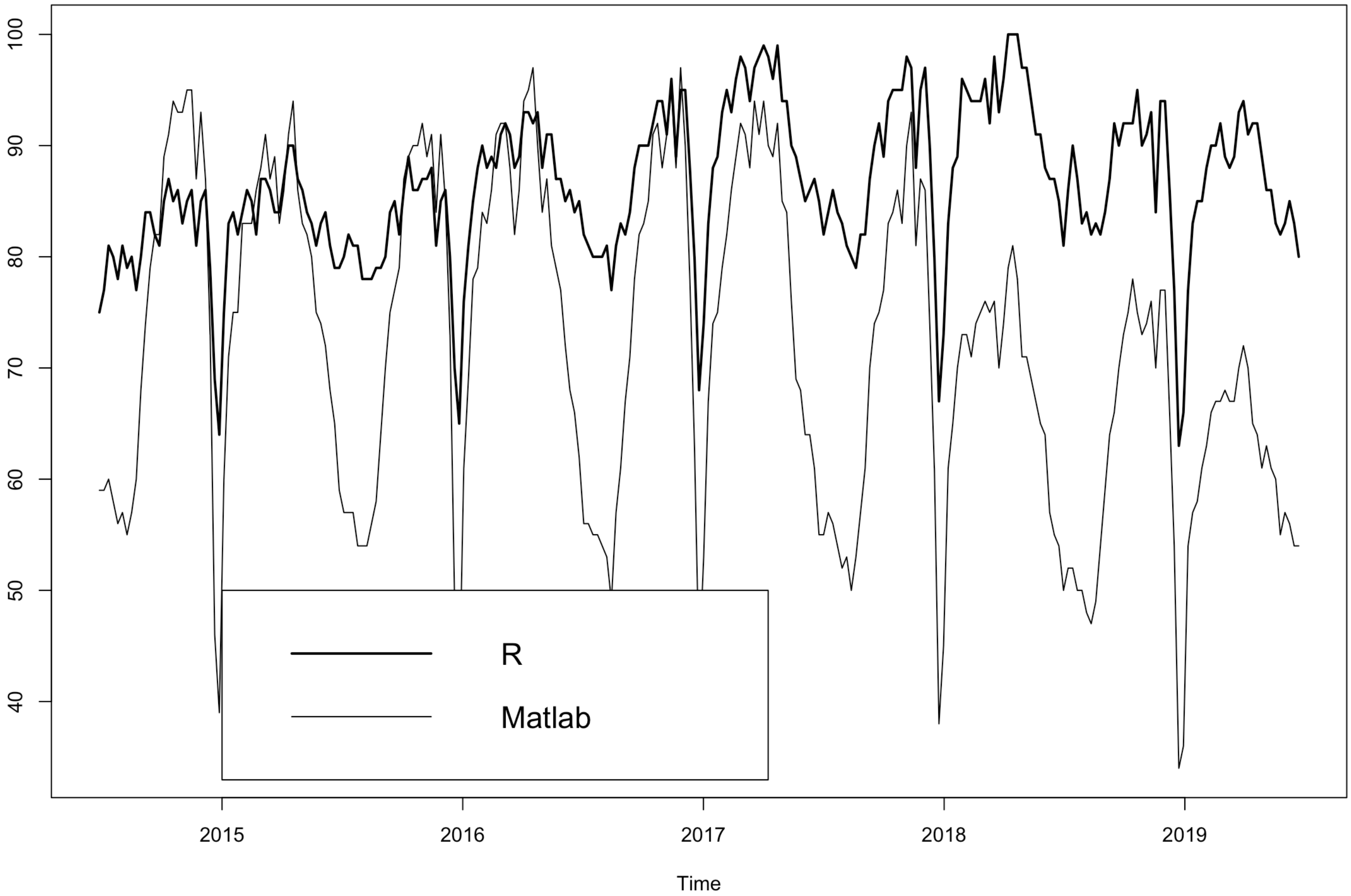
L'analisi delle serie storiche per...

1. Descrivere un fenomeno osservato
2. Prevedere valori futuri
3. Studiare impatto di shock
4. Analizzare dinamiche congiunturali
5. Destagionalizzare

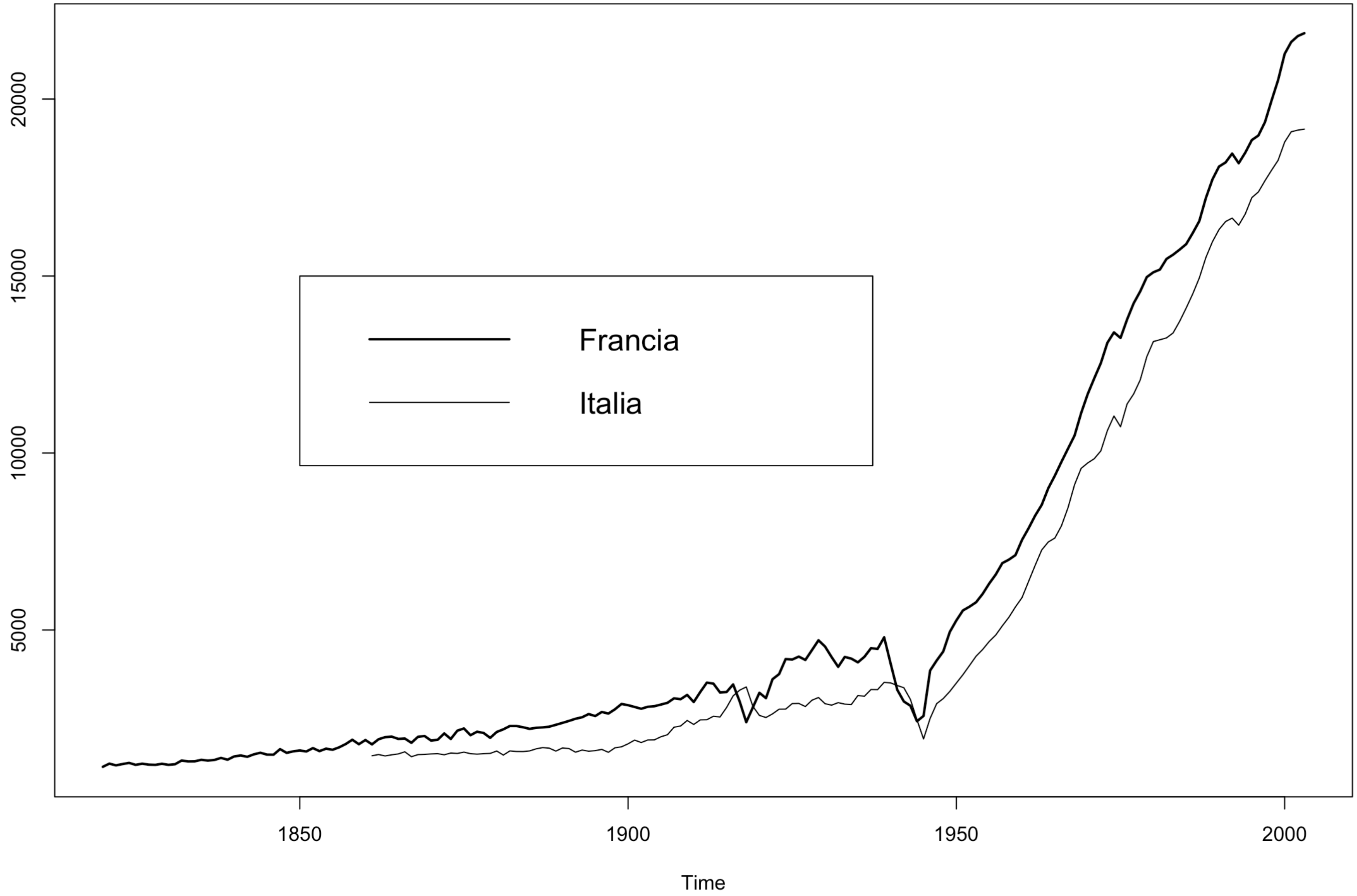
Esempio di serie storica: Google Trend Mensile



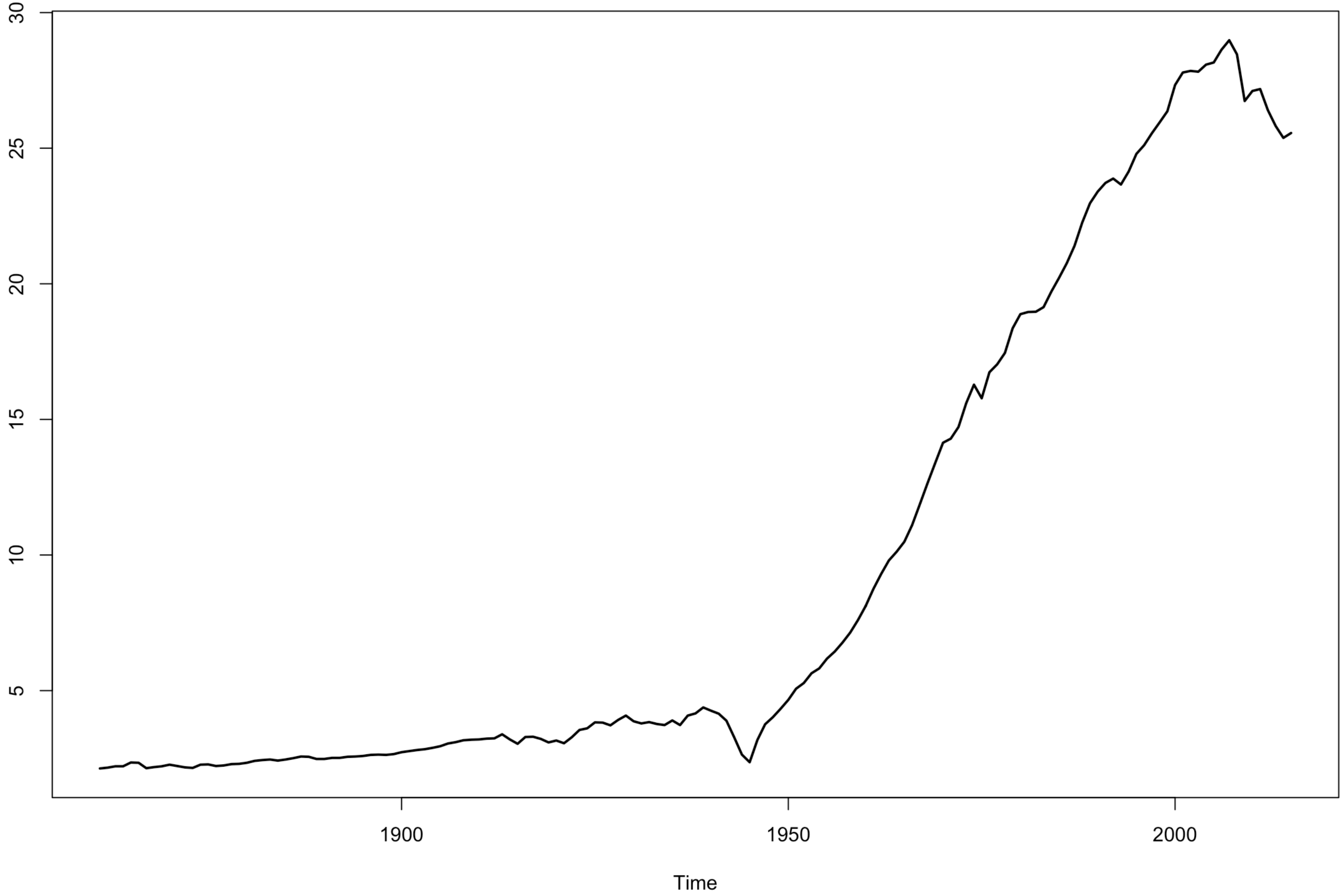
Esempio di serie storica: Google Trend Settimanale



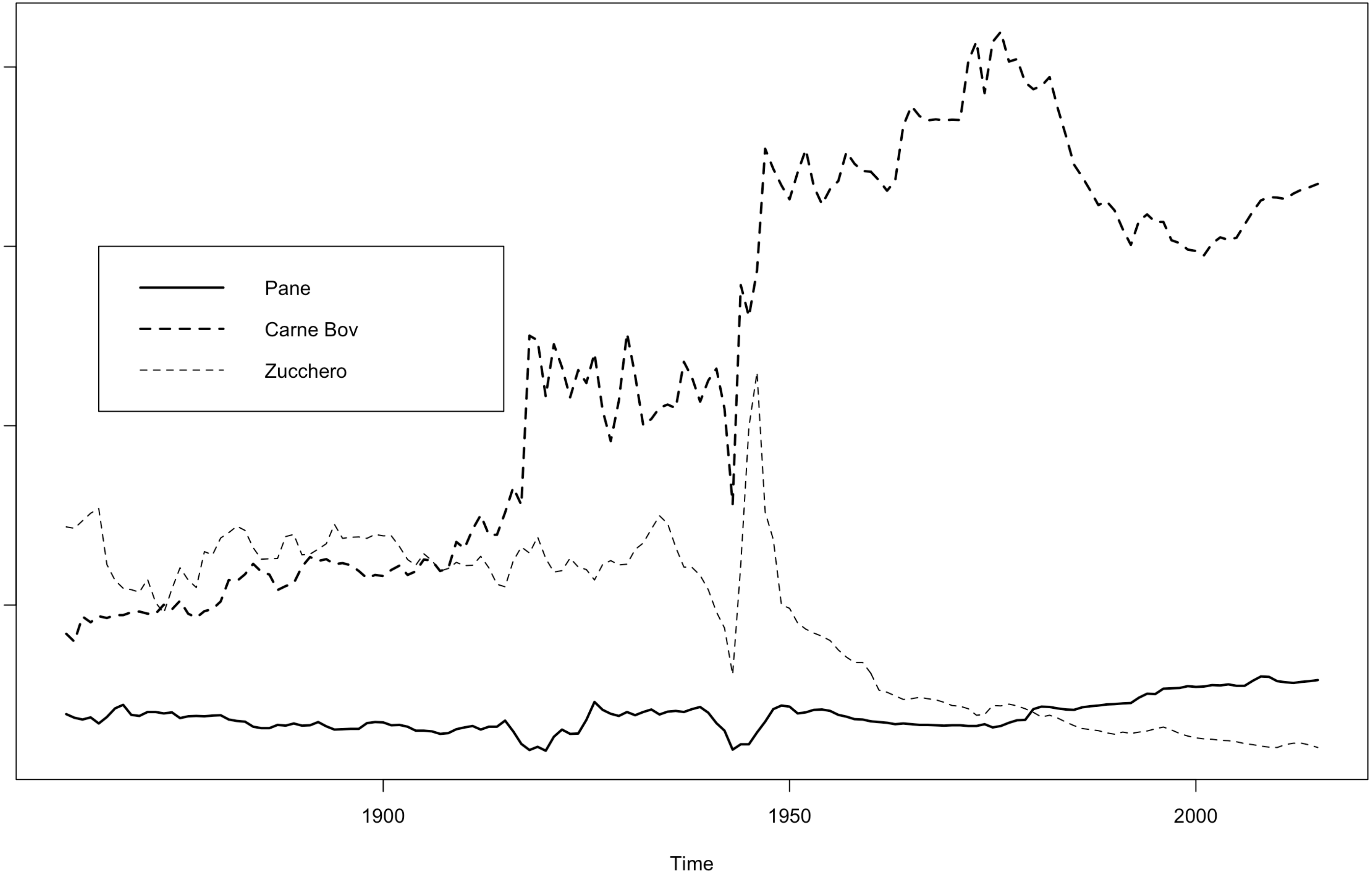
Esempio di serie storica: World Population, GDP and Per Capita GDP, 1-2003 AD (Angus Maddison)



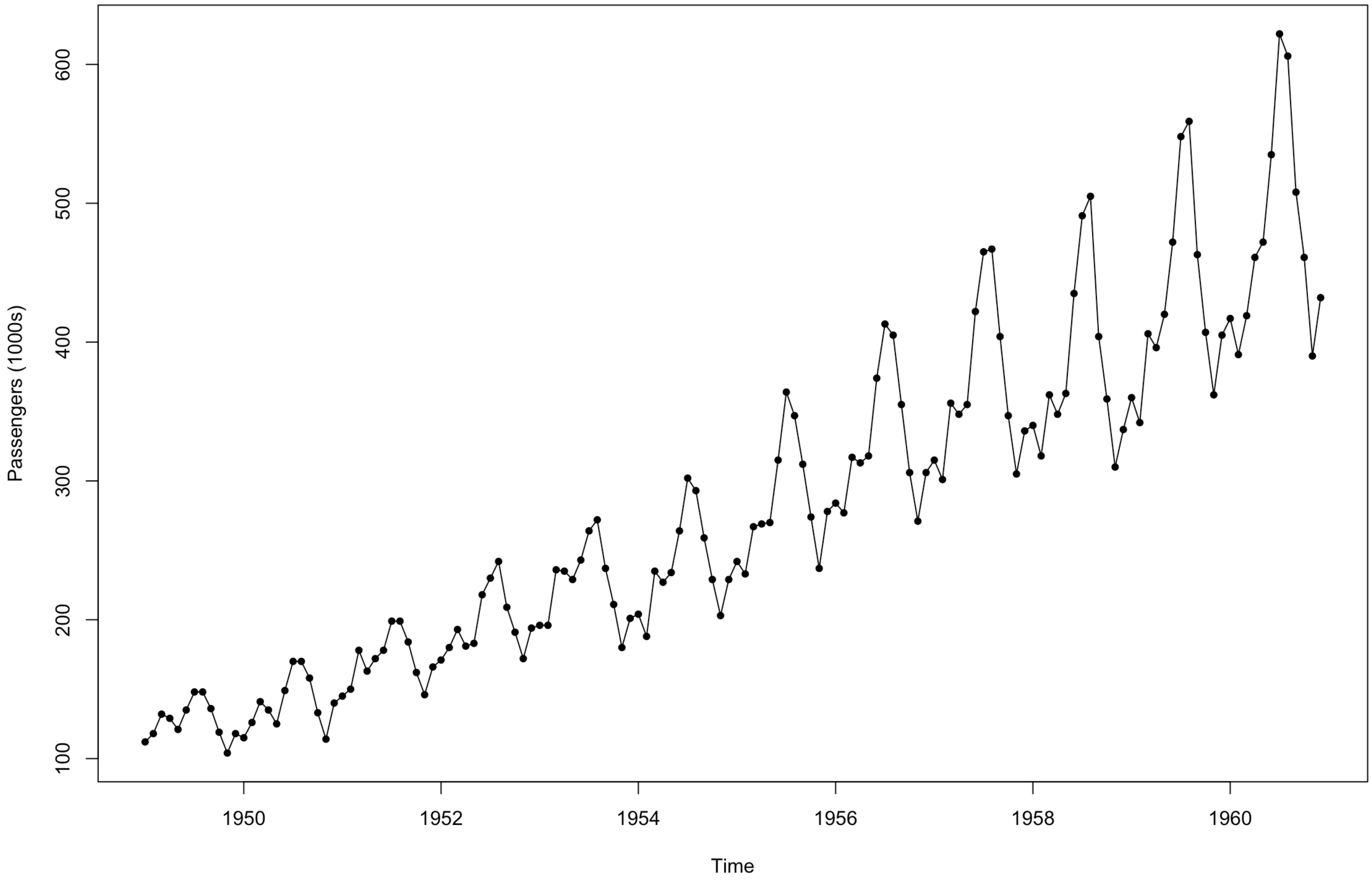
Esempio di serie storica: La ricostruzione Banca d'Italia-Istat 1861-2015 (da Statistiche storiche della Banca d'Italia)



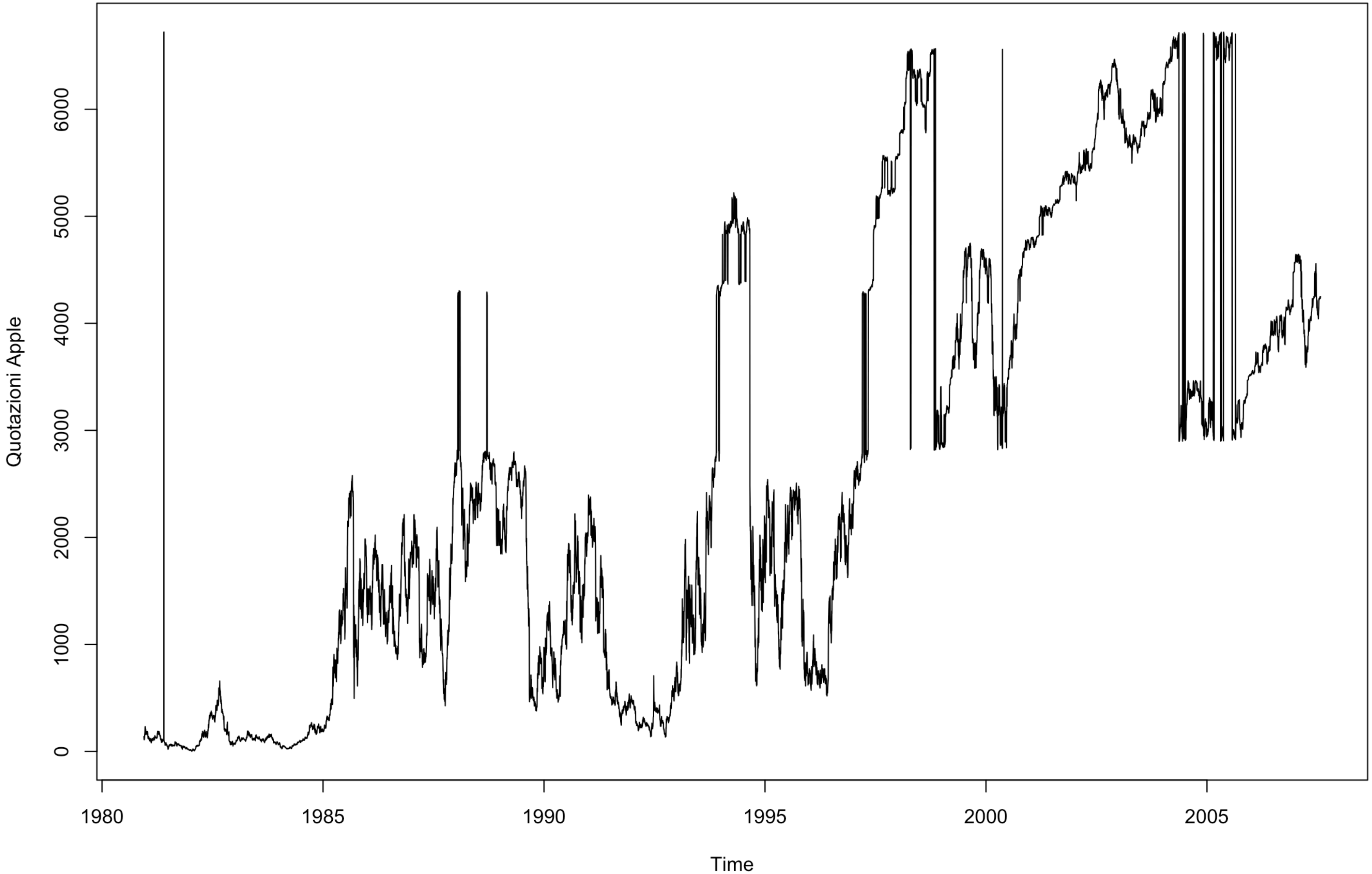
Esempio di serie storica: Prezzi medi al consumo di alcuni prodotti del comparto alimentare - Anni 1861-2015 (euro 2015)



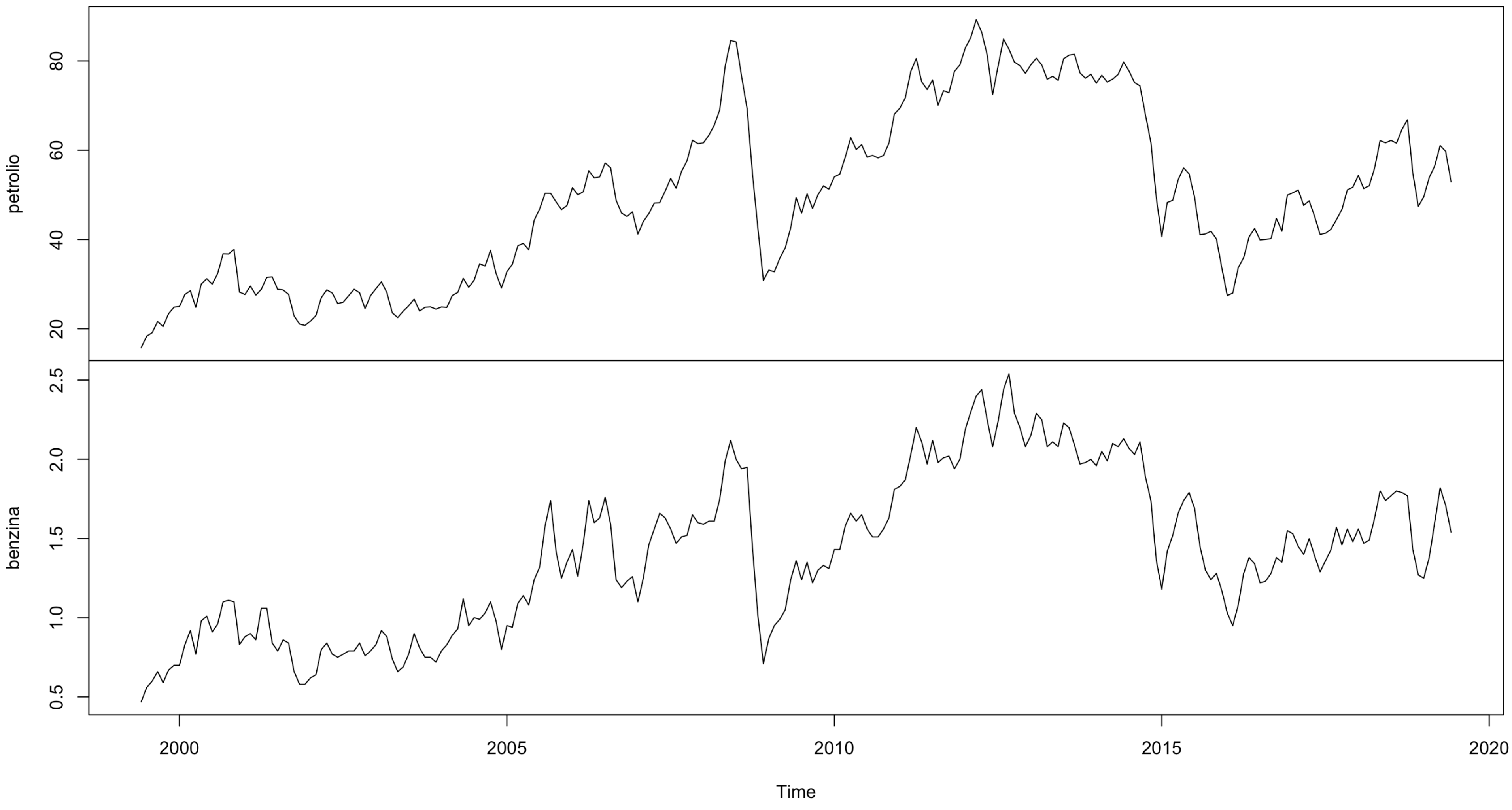
Esempio di serie storica: classici dati di Box & Jenkins su airolinee. Totali mensili dei passeggeri internazionali dal 1949 al 1960



Esempio di serie storica: quotazioni Apple Computer



Esempio di serie storiche: prezzi mensili di petrolio e benzina



Esempi di serie storica: R Script

```r
setwd("~/Documents/prova r pkg/lucidi del corso")
datigtm <- read.csv(file="RMatlab.csv")
datigtw <- read.csv(file="RMatlab week.csv")
tsgtm <- ts(datigtm[,2:3], frequency = 12, start = c(2004, 1)) # frequency 12 => Monthly Data
par(mar=c(4.5,4.5,1,1))
plot(tsgtm,cex.lab=2,main="",plot.type = "single",lty=1,lwd=c(2,1),ylab="")
legend(2012,90, c("R","Matlab"), lwd=c(2,1),cex=1.5)
library(lubridate)
tsgtw <- ts(datigtw[,2:3], freq=365.25/7, start=decimal_date(ymd("2014-06-29")))
par(mar=c(4.5,4.5,1,1))
plot(tsgtw,cex.lab=2,main="",plot.type = "single",lty=1,lwd=c(2,1),ylab="")
legend(2015,50, c("R","Matlab"), lwd=c(2,1),cex=1.5)
gdppc <- read.table(file="percapitagdp.csv",sep=";",dec=",",header = T)
tsgdp <- ts(gdppc[6:189,2:183], frequency = 1, start = c(1820))

plot(tsgdp[,c(5,7)],cex.lab=2,main="",plot.type = "single",lty=c(1,1),lwd=c(2,1),ylab="")
legend(1850,15000, c("Francia","Italia"), lty=c(1,1),lwd=c(2,1),cex=1.5)
```

Esempi di serie storica: R Script

```
gdppc_it_bi <- read.table(file="gdppc_it_bi.csv",sep=";",dec=",",header = T)
tsgdp_it_bi <- ts(gdppc_it_bi[,2:4], frequency = 1, start = c(1861))
plot(tsgdp_it_bi[,3],main="",lwd=2,ylab="")

data("AirPassengers")
AP <- AirPassengers
plot(AP, ylab="Passengers (1000s)", type="o", pch =19, cex=0.7)

prezzi <- read.table(file="prezzi.csv",sep=";",dec=",",header = T)
prezzi_istat <- ts(prezzi[,2:17], frequency = 1, start = c(1861))
par(mar=c(4.5,1,1,1))
plot(prezzi_istat[,c(1,5,16)],cex.lab=2,main="",plot.type =
"single",lty=c(1,2,2),lwd=c(2,2,1),ylab="")
legend(1865,15, c("Pane","Carne Bov","Zucchero"), lty=c(1,2,2),lwd=c(2,2,1),cex=1)

apple <- read.csv(file="AAPL.csv",header = T)
apple$Date <- as.Date(apple$Date)
tsapple <- ts(apple[,2:7], start = c(1980, as.numeric(format(inds[1], "%j"))),frequency = 365)
par(mar=c(4.5,4.5,1,1))
plot(tsapple[,5], main="",lwd=1,ylab="Quotazioni Apple")
```

Approcci di analisi

APPROCCIO DETERMINISTICO

- scomporre la serie nelle sue componenti

APPROCCIO STOCASTICO

- trovare processo generatore dei dati

Decemposizione di una serie storica

Ogni dato (Y_t) al tempo *t* di una serie storica può essere espresso come somma o prodotto di 3 componenti, ovvero Seasonality (S_t), Trend (T_t) ed Errore (ε_t) (detto White Noise).

Per le serie storiche additive,

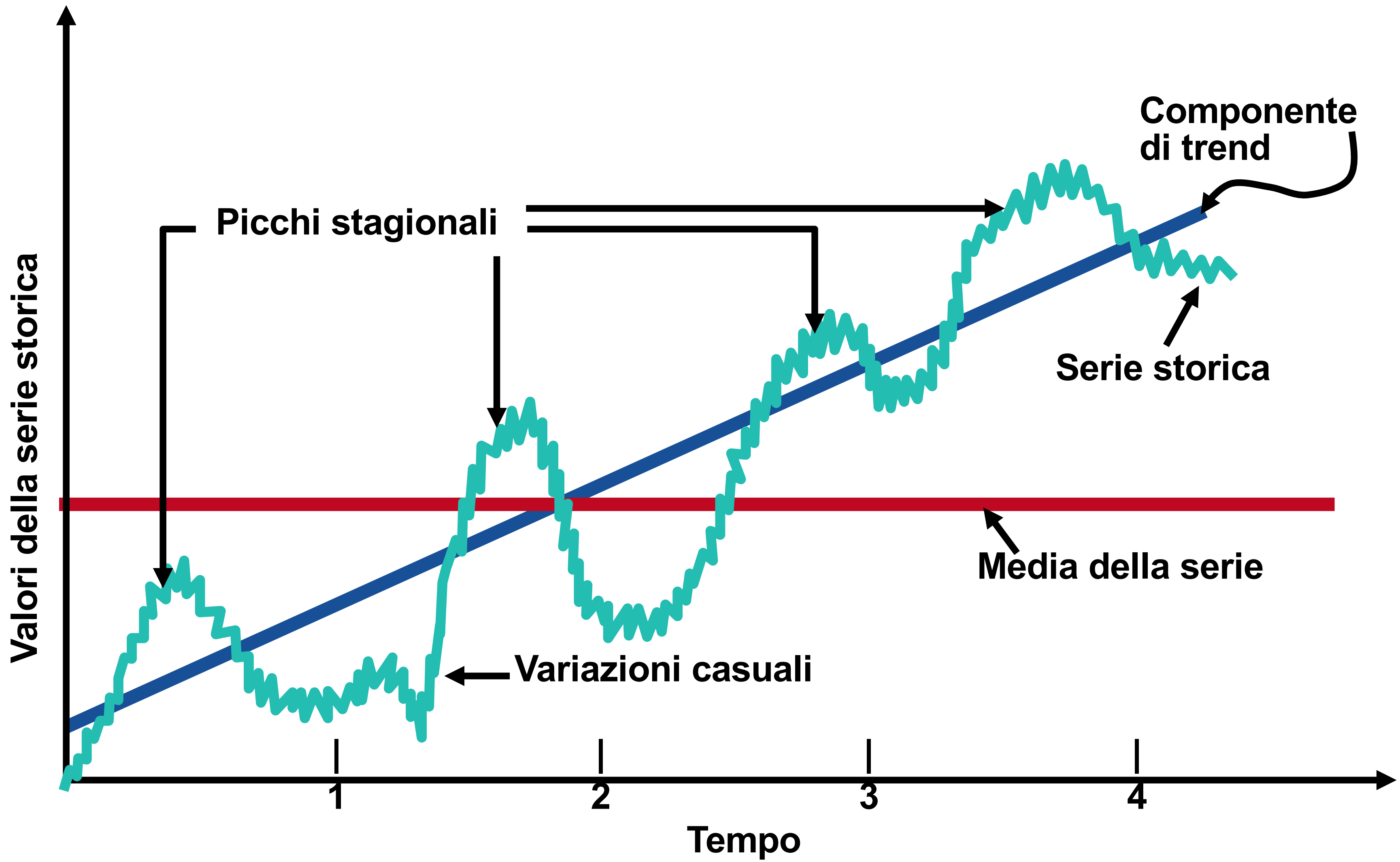
$$Y_t = S_t + T_t + \varepsilon_t$$

Per le serie storiche moltiplicative,

$$Y_t = S_t \times T_t \times \varepsilon_t$$

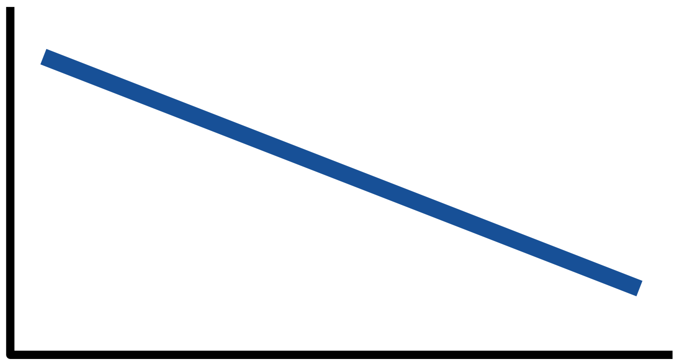
Una serie storica moltiplicativa può essere convertita in additiva calcolandone il logaritmo

Componenti di una serie storica



Componente trend

- Schema persistente, complessivamente verso l'alto o verso il basso
- Cambiamenti dovuti a popolazione, tecnologia, età, cultura, ecc.
- In genere per diverse occasioni



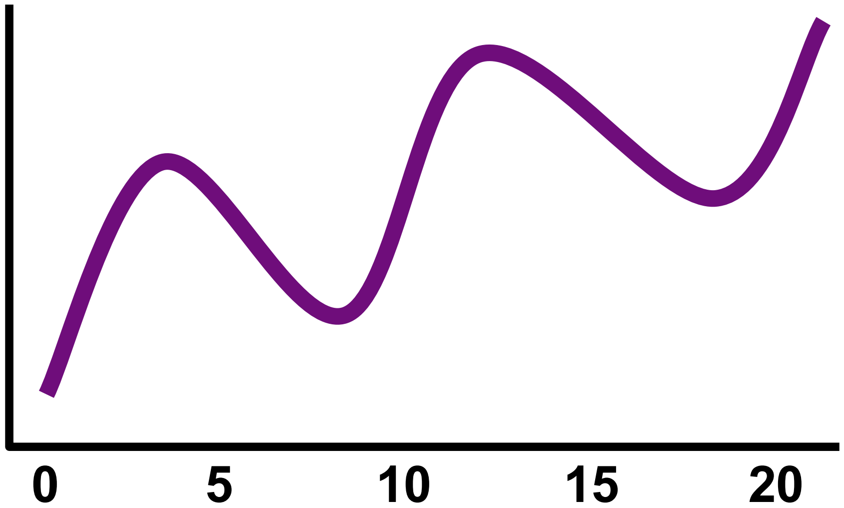
Componente stagionale

- Schema regolare delle fluttuazioni su e giù
- A causa del tempo, delle abitudini, ecc.
- Si verifica entro un periodo ben definito

Period	Length	Number of Seasons
Week	Day	7
Month	Week	4-4.5
Month	Day	28-31
Year	Quarter	4
Year	Month	12
Year	Week	52

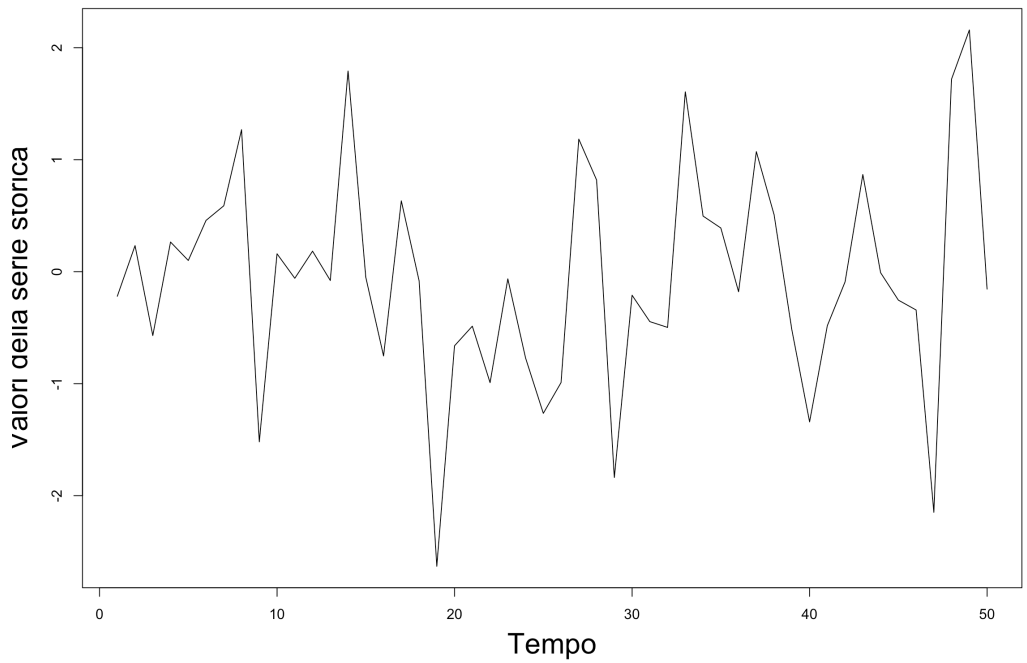
Componente ciclica

- Si ripetono i movimenti su e giù
- Interessato da ciclo economico, fattori politici ed economici
- Durata di medio lungo periodo (ad esempio più anni)



Componente casuale

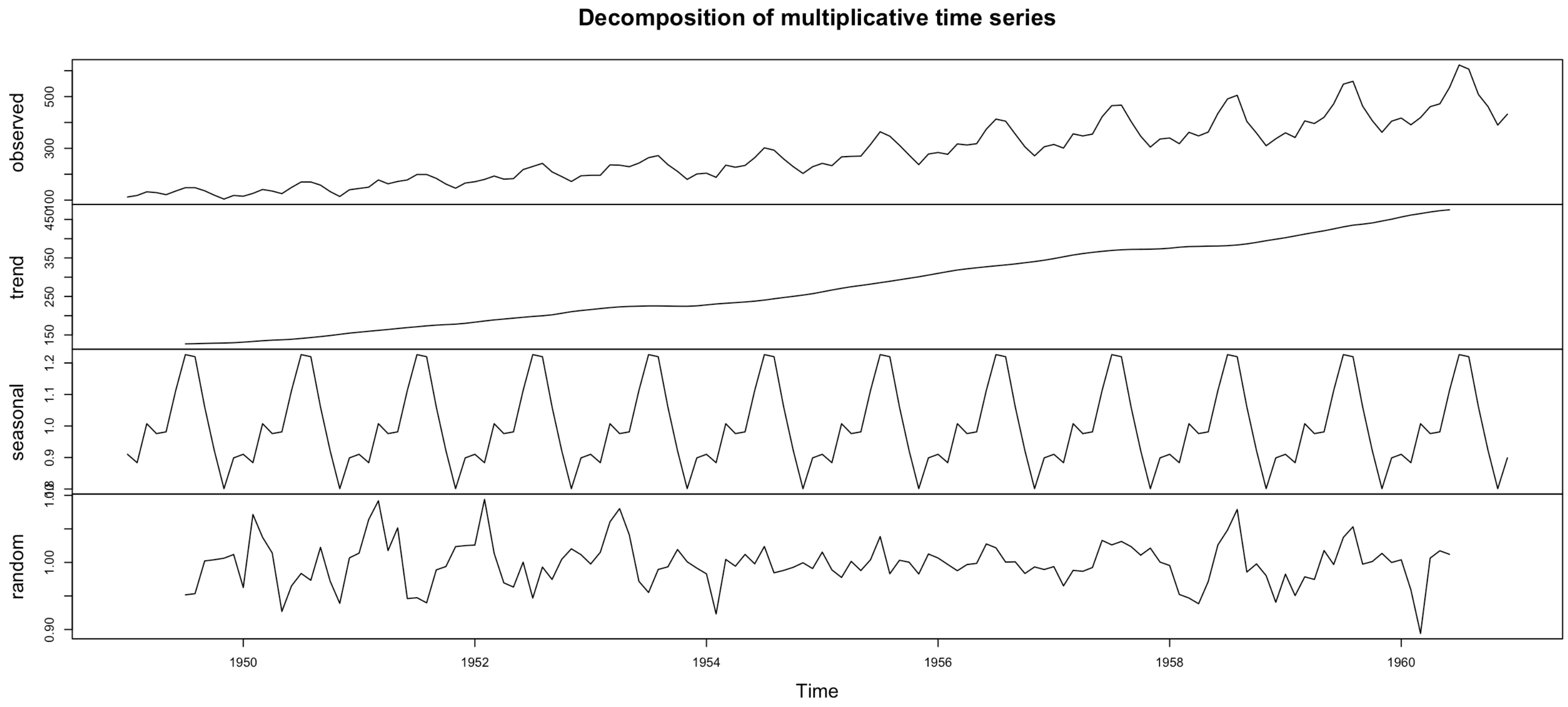
- Fluttuazioni "residue" irregolari, non sistematiche
- A causa di variazioni casuali o eventi imprevisti
- Breve durata e non ripetizione



Decemposizione di una serie storica

La funzione decompose () divide le serie storiche in componenti stagionali, trend ed errori.

```
decAP <- decompose(AP, type = "multiplicative")
```

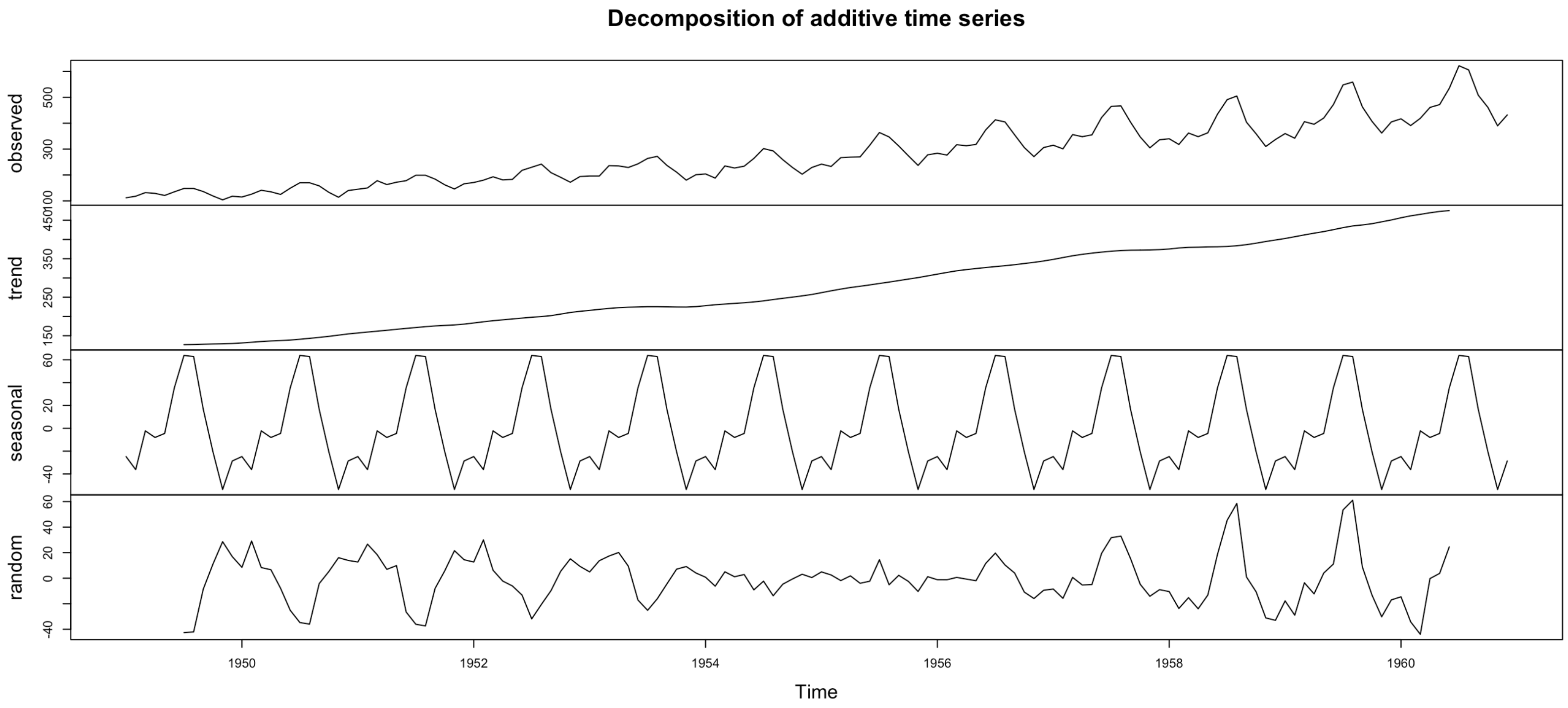


Decemposizione di una serie storica

La funzione decompose () divide le serie storiche in componenti stagionali, trend ed errori.

```
decAP <- decompose(AP, type = "additive")
```

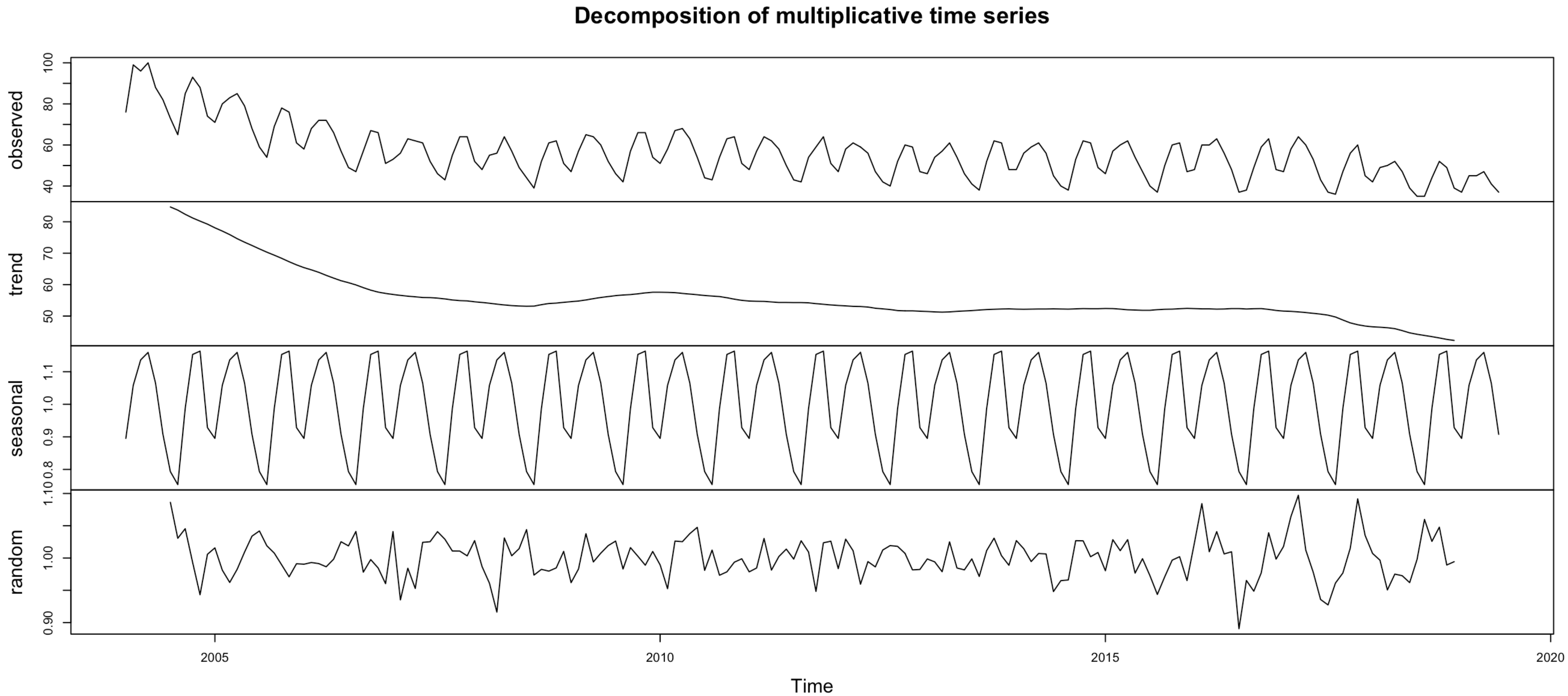
p



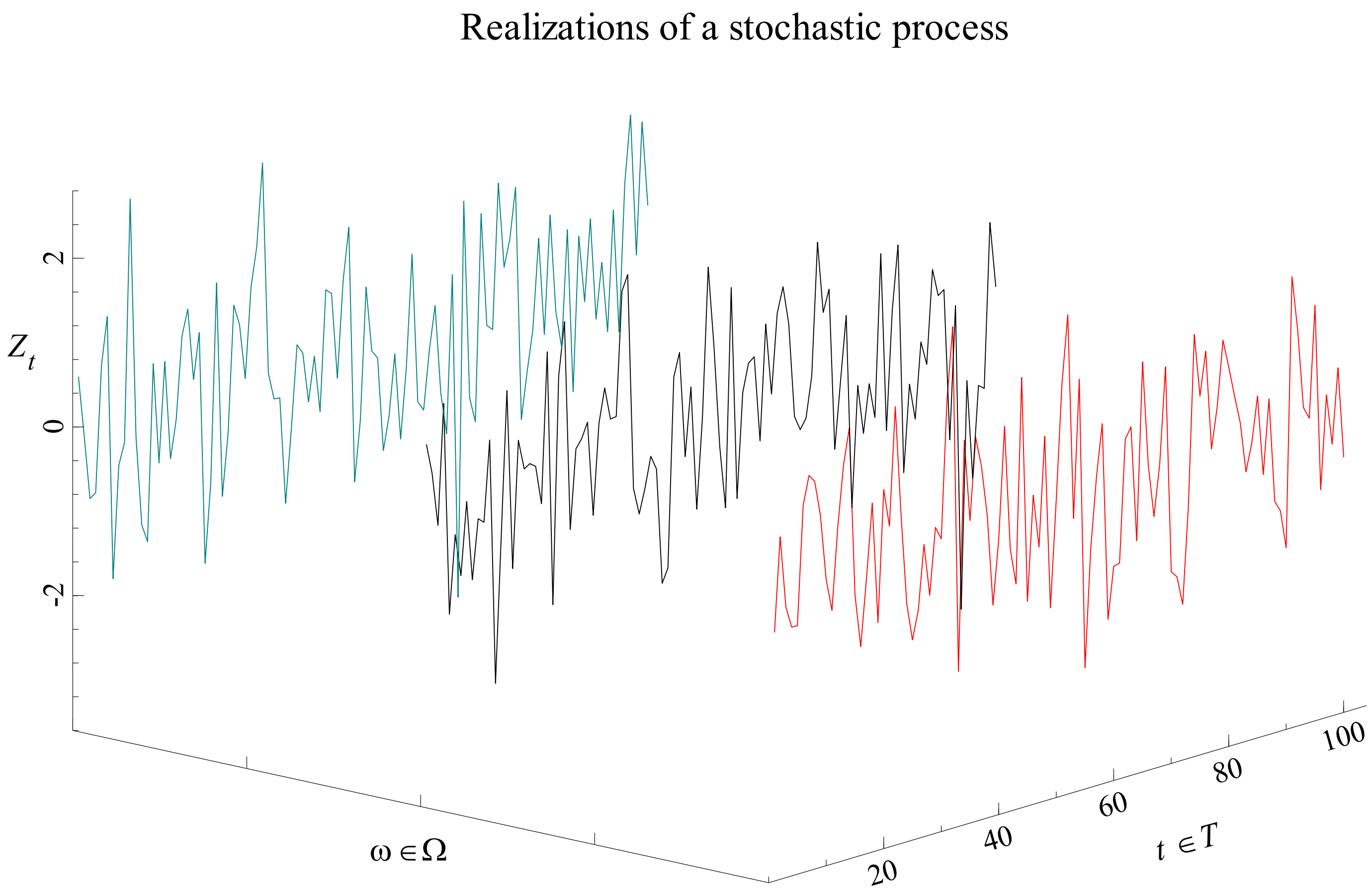
Decemposizione di una serie storica

```
decMatlab <- decompose(tsgtm[,2], type =
  "multiplicative")

plot(decMatlab)
```



Approccio Stocastico



Approccio Stocastico

- al variare di t e di ω $X(t,\omega)$ è un PROCESSO STOCASTICO
- al variare di t, $X(t,\omega)$ è una VARIABILE CASUALE
- al variare di ω, $X(t,\omega)$ è una REALIZZAZIONE
- fissati t e ω, $X(t,\omega)$ è un VALORE REALE

Approccio Stocastico

- Approccio stocastico
- Obiettivo determinare la legge che "governa" il processo
- Il passato spiega il presente e il futuro
- Modelli lineari che spiegano guardando il passato

Approccio Stocastico

L'obiettivo principale dell'analisi delle serie storiche è quello di sviluppare modelli matematici che riescano a descrivere in modo plausibile il campione osservato.

Punti adiacenti nel tempo sono correlati tra loro, pertanto il valore della serie osservato al tempo t dipenderà in qualche maniera dai valori passati

$$x_{t-1}, x_{t-2}, \ldots$$

Tale considerazione è alla base della costruzione e analisi delle serie storiche.

Modelli Lineari Univariati

- STIAMO LIMITANDO L'ATTENZIONE A MODELLI **LINEARI**
- STIAMO OSSERVANDO SOLO **UNA** VARIABILE X NEL TEMPO
- SPIEGHIAMO IL VALORE ATTUALE $\mathbf{X_t}$ OSSERVANDO IL PASSATO

➔ DEFINIZIONE DI VARIABILE RITARDATA

Variabile Ritardata

X_t	X_{t-1}	X_{t-2}	X_{t-3}
X_1	--	--	--
X_2	X_1	--	--
x_3	X_2	X_1	--
X_4	X_3	X_2	X_1
X_5	X_4	X_3	X_2

Variabile Ritardata

X_t	X_{t-1}	X_{t-2}	X_{t-3}
3,2	--	--	--
2,8	3,2	--	--
5,4	2,8	3,2	--
7,2	5,4	2,8	3,2
6,3	7,2	5,4	2,8

Operatore Ritardo

L'operatore ritardo è indicato con "**L**" (*Lag)* ed è definito come segue:

$$\mathbf{L^1 X_t = X_{t-1}}$$

$$\mathbf{L^2 X_t = X_{t-2}}$$

$$\mathbf{L^3 X_t = X_{t-3}}$$

A volte viene indicato con B (= Backshift)

Trasformazioni

Molto spesso nella elaborazione dei dati statistici le variabili vengono trasformate. Basti pensare all'analisi del PIL, il quale viene sempre analizzato, non nel valore, assoluto, ma in termini di variazione percentuale.

Le trasformazioni utili a fini interpretativi dei fenomeni sono:

- variazione assoluta
- variazione percentuale

Trasformazioni

Le trasformazioni

La variazione assoluta si ottiene immediatamente come differenza del valore di X al tempo t rispetto allo stesso un tempo precedente:

$$\Delta X_t = X_t - X_{t-1}$$

Il simbolo usato per la differenza assoluta è Δ.
Analizzando ΔX piuttosto che X, il modello specificato spiegherà la variazione di X.

Trasformazioni

Le trasformazioni

Il ritardo di una variabile può essere definito anche introducendo l'OPERATORE RITARDO:

$$LX_t = X_{t-1}$$

$$L^2X_t = X_{t-2}$$

$$\dots$$

$$L^pX_t = X_{t-p}$$

Avremo allora che

$$\Delta X_t = X_t - X_{t-1} = X_t - LX_t = (1-L)X_t$$

Trasformazioni

Le trasformazioni

L'altra trasformazione importante è quella che porta al tasso di variazione della serie $\frac{X_t - x_{t-1}}{X_{t-1}}$

Si dimostra facilmente che questa è ben approssimata dalla seguente espressione

$$\frac{X_t - X_{t-1}}{X_{t-1}} \approx \log X_t - \log X_{t-1} = \Delta \log X_t$$

Molto spesso quindi piuttosto che rilevare Yt, si studierà logYt e poi, successivamente, la sua variazione.
La trasformazione logaritmica viene ampiamente utilizzata in statistica.

Trasformazioni

Le trasformazioni

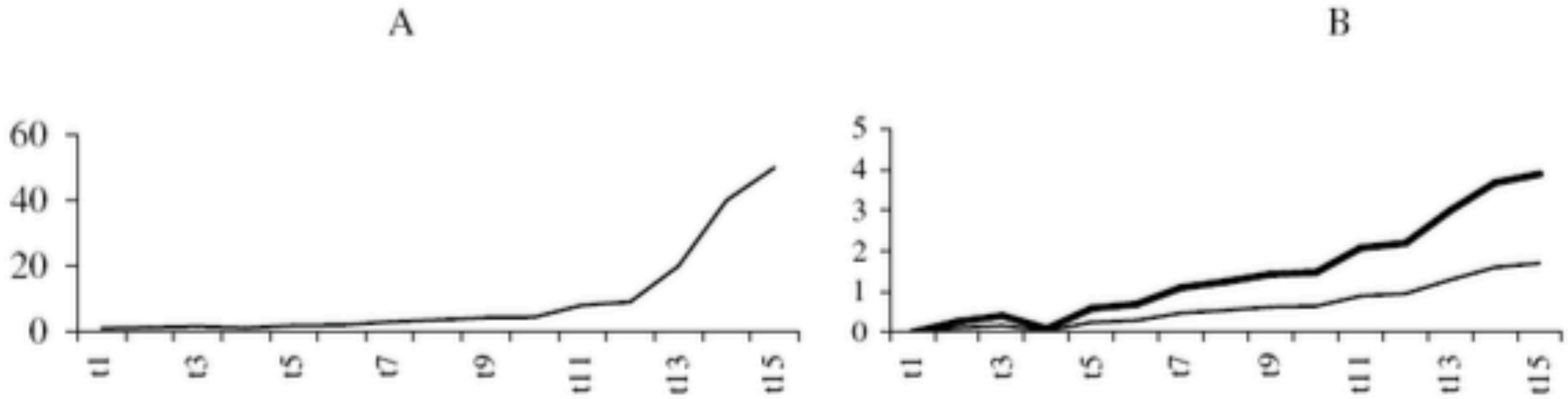


Fig. 3 – Trasformazione logaritmiche (B) di una serie storica (A).

Modello regressione

Modello di regressione:

- ✓ assunzione di independenza delle osservazioni
- ✓ La variabile dipendente Y è funzione di variabili esplicative $X_1, \ldots, X_k$
- ✓ Si usa il coefficiente di correlazione r

$$Y_t = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_t$$

Modello serie storiche

- Modelli Serie storiche:
  - ✓ Dipendenza tra osservazioni
  - ✓ X spiegata dal suo passato
  - ✓ Uso dell'autocorrelazione

$$X_t = \alpha + \phi_1 X_{t-1} + \cdots + \phi_k X_{t-k} + \varepsilon_t$$

Statistiche di base

MODELLO REGRESSIONE

- media
- varianza
- correlazione

MODELLO SERIE STORICHE

- media
- varianza
- autocorrelazione

Correlazione

La correlazione misura il legame lineare tra X_t e Y_t

Se ρ = 0: nessuna relazione lineare

Se ρ = +1: relazione lineare perfetta crescente

Se ρ = -1: relazione lineare perfetta decrescente

$$-1 \leq \rho_{X,Y} \leq +1$$

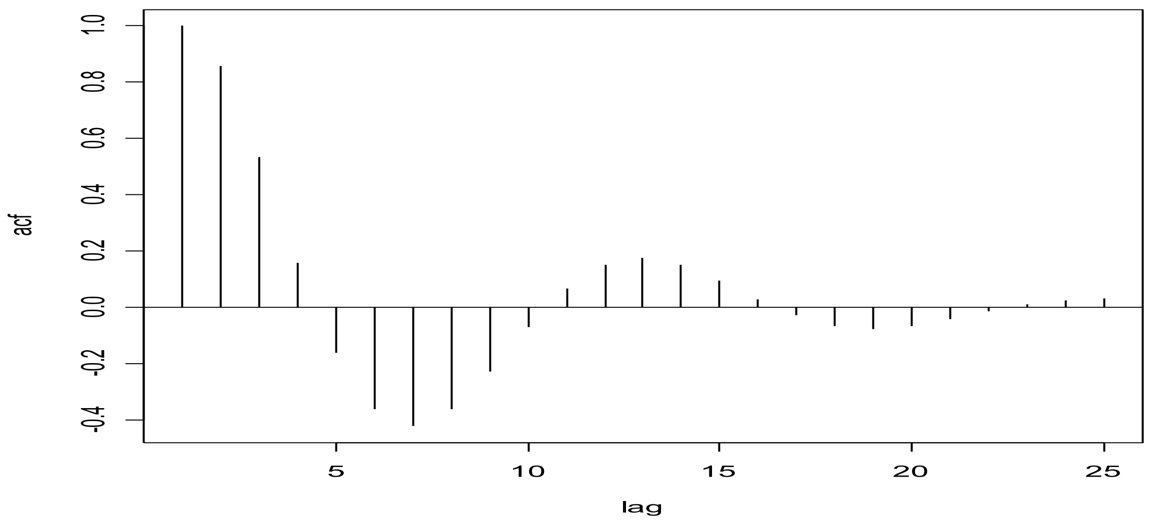
Autocorrelazione

Autocorrelazione (ACF) misura la dipendenza lineare tra X_t and X_{t-k}

$$\rho(k) = corr(X_{t-1}, X_{t-k})$$

Autocorrelogramma

- Il grafico dell'autocorrelazione al variare del ritardo k è detto correlogramma



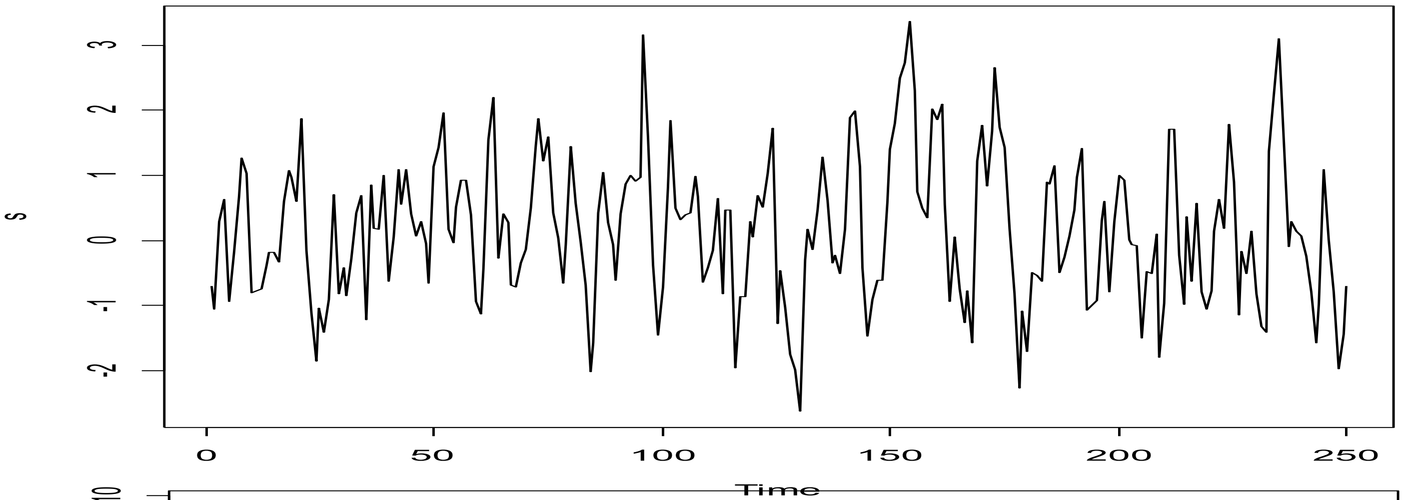
Stazionarietà

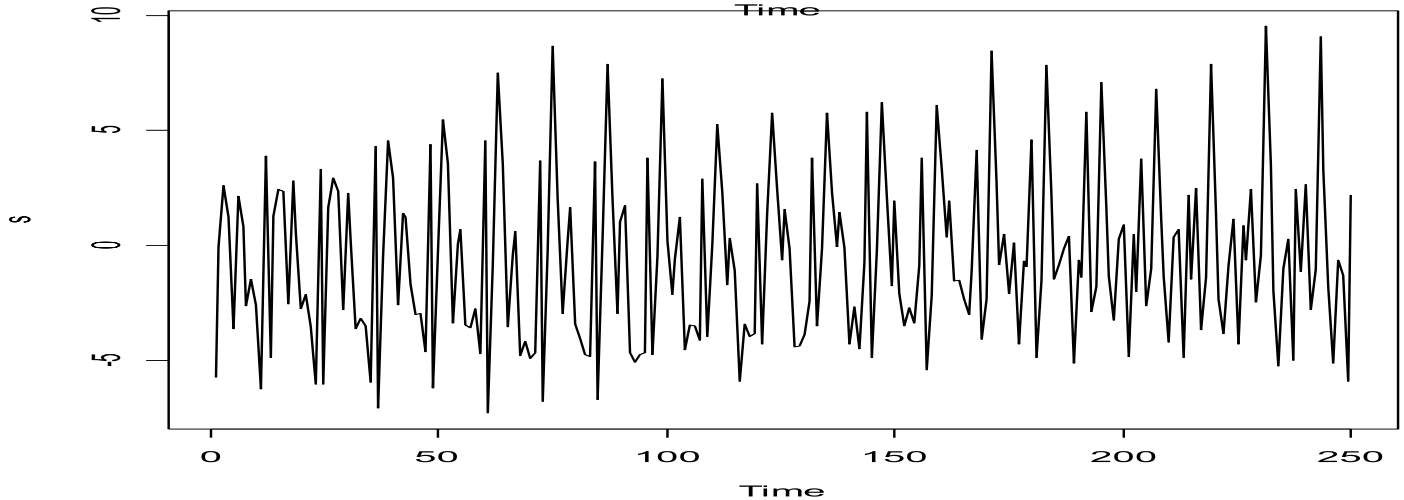
Necessario porre delle restrizioni su alcune proprietà fondamentali: stazionarietà debole

$$E\,(X_t) = \mu = const, \qquad \forall t$$
$$E\,(X_t - \mu)^2 = \sigma^2 = const < +\infty,$$
$$E\,[(X_t - \mu)\,(X_{t-k} - \mu)] = \gamma_k$$

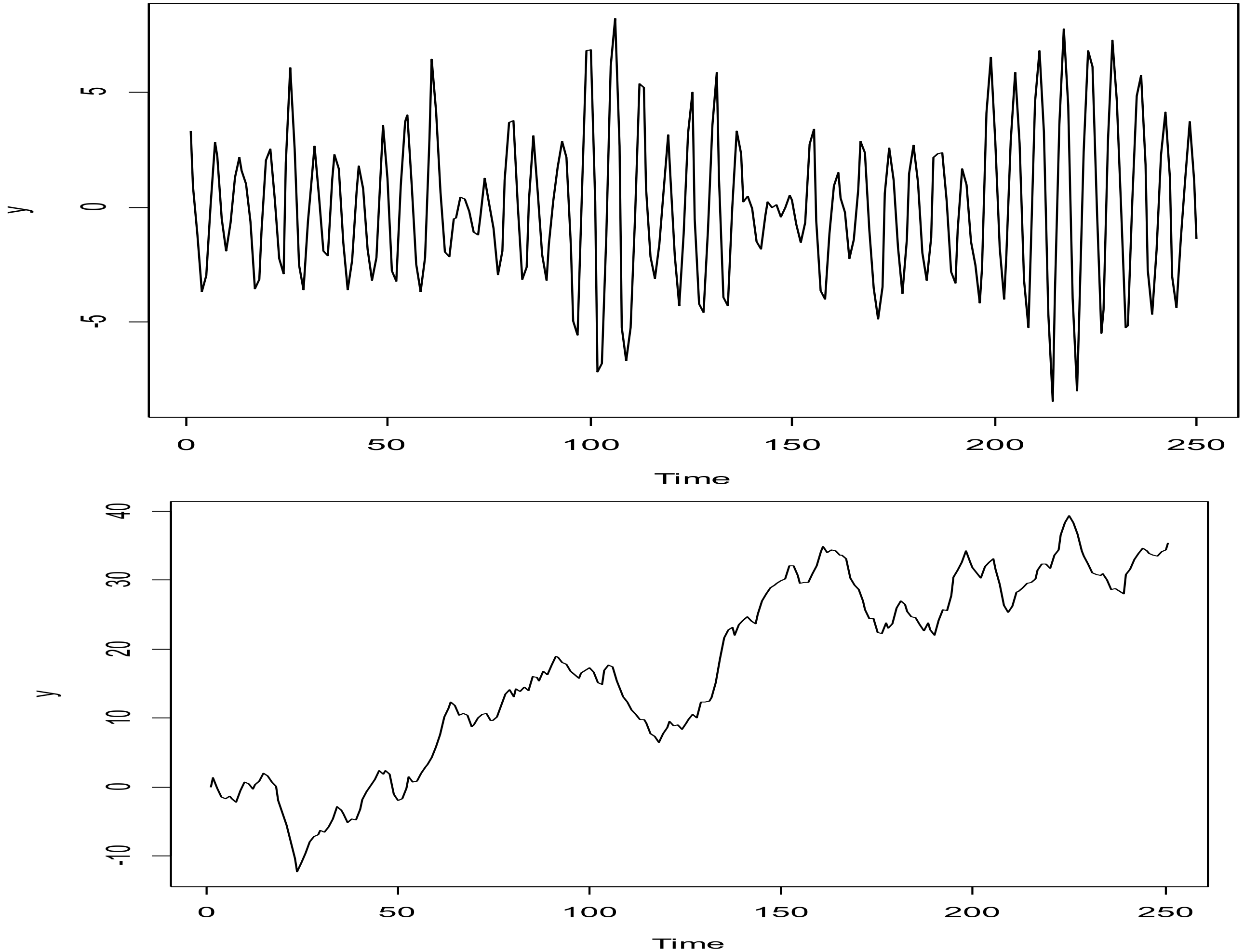
La media e la varianza sono costanti, l'autocorrelazione dipende solo da k e non da t.

Serie stazionarie





Serie non stazionarie



Modelli stazionari

Esistono diversi modelli per descrivere un processo stazionario:

- Moving average MA (*Media mobile)*
- Autoregressive Model AR

(*Autoregressivo)*

- Mixed Autoregressive Moving Average model ARMA

E poi esistono diversi modelli per descrivere processi non stazionari.

Modelli lineari

MA(q):

$$X_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

AR(p):

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t$$

Modelli lineari

ARMA(p,q):

Dove X è il fenomeno sotto osservazione e a sono gli shock (imprevedibili) che si verificano

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} + \ldots - \theta_q a_{t-q}$$

Modelli lineari

White Noise (WN):

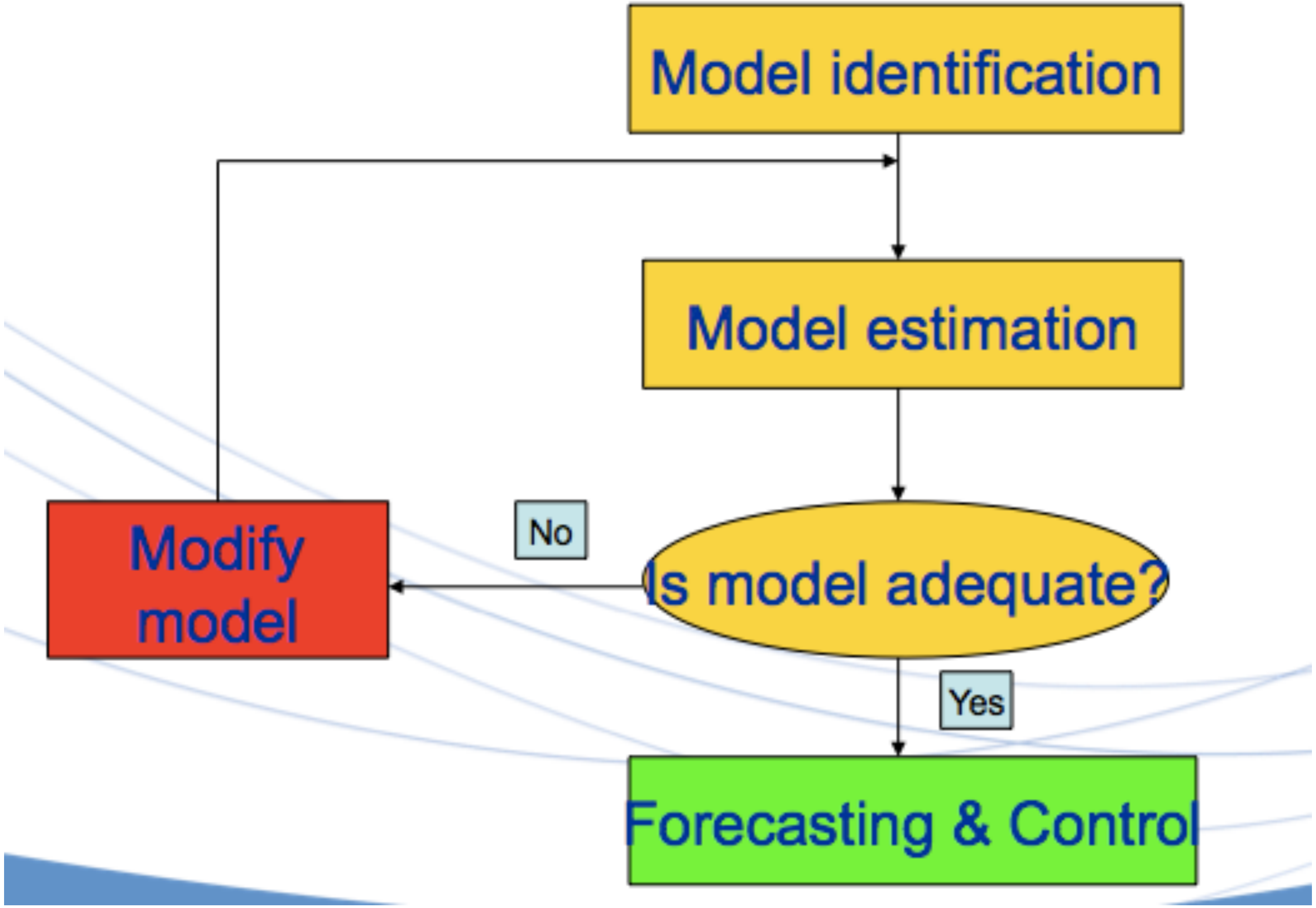
Dove at è uno shock casuale gaussiano, ossia:

$$X_t = a_t$$

$$a_t \approx N(0, \sigma)$$

$$\rho(k) = 0 = \lambda(k) \quad \forall k$$

BOX-JENKINS


Model identification
Model estimation
Is model adequate?
No
Modify model
Yes
Forecasting & Control

Identificazione

Necessario capire dai dati, quale è il modello più appropriato a descrive la nostra serie osservata (AR, MA, ARMA?)

Capire l'ordine del modello (p, q?)

Tale fase si basa sullo studio di alcune statistiche campionarie:

ACF (= ρ_k), PACF (=λ_k)

Identificazione

Ogni modello presenta un andamento teorico di ACF e PACF ben preciso.

Sulla base della corrispondenza tra ACF e PACF campionaria e teorica è possibile abbinare l'appropriato modello ai dati osservati.

Solo dopo attenta lettura e interpretazione di ACF e PACF si passa alla stima del modello.

Identificazione AR(1)

Modello AR(1):

$$X_t = \phi_1 X_{t-1} + a_t$$

Mi aspetto che ci sia autocorrelazione solo per k = 1

Invece NO!!

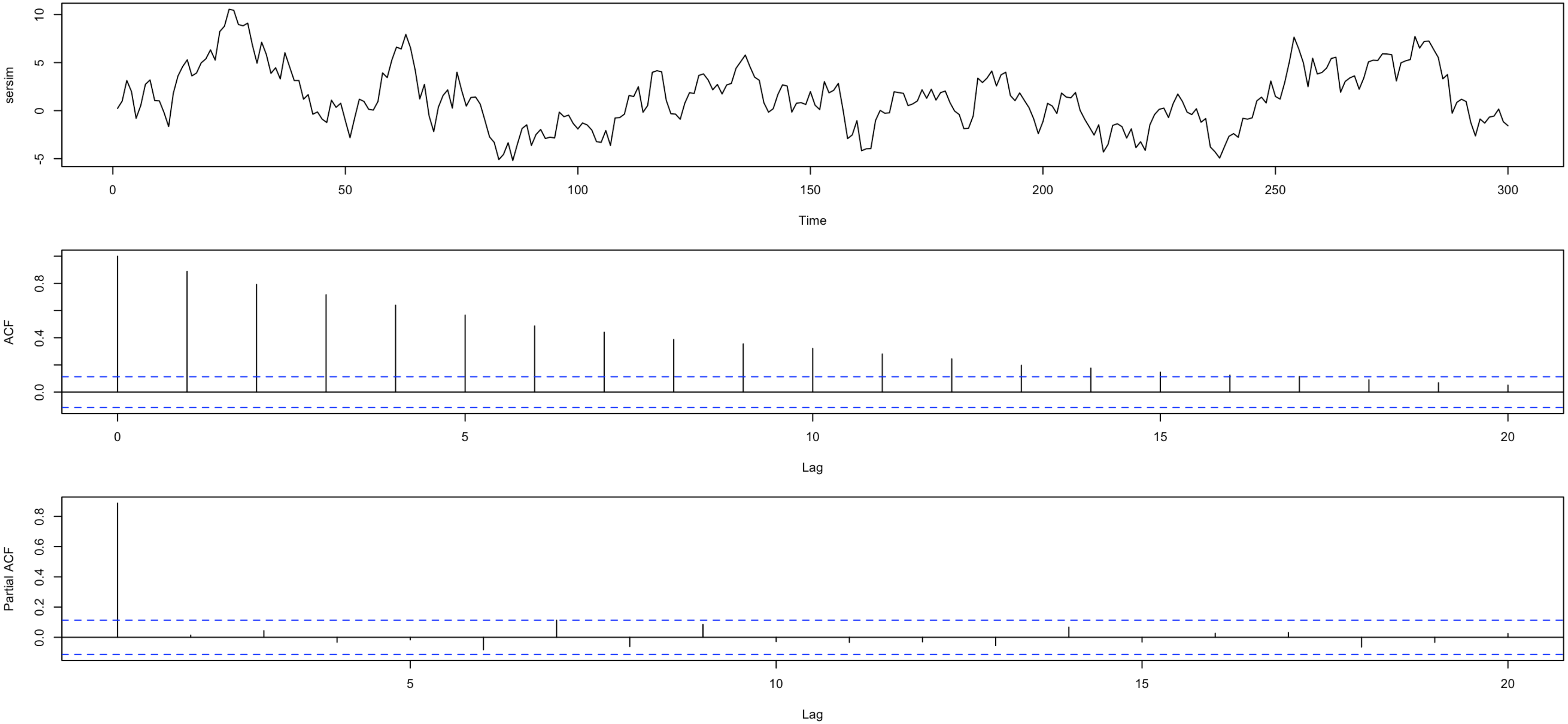
X_{t-1} sarà influenzata da X_{t-2} e quindi ci sarà una relazione indiretta anche tra X_t e X_{t-2}, etc.

Identificazione AR(1)

Correlogramma ACF di un processo AR(1):

Introduciamo la Partial AutoCorrelation Function (PACF):

= La correlazione tra X_t e X_{t-k}, tolti tutti i legami intermedi indotti

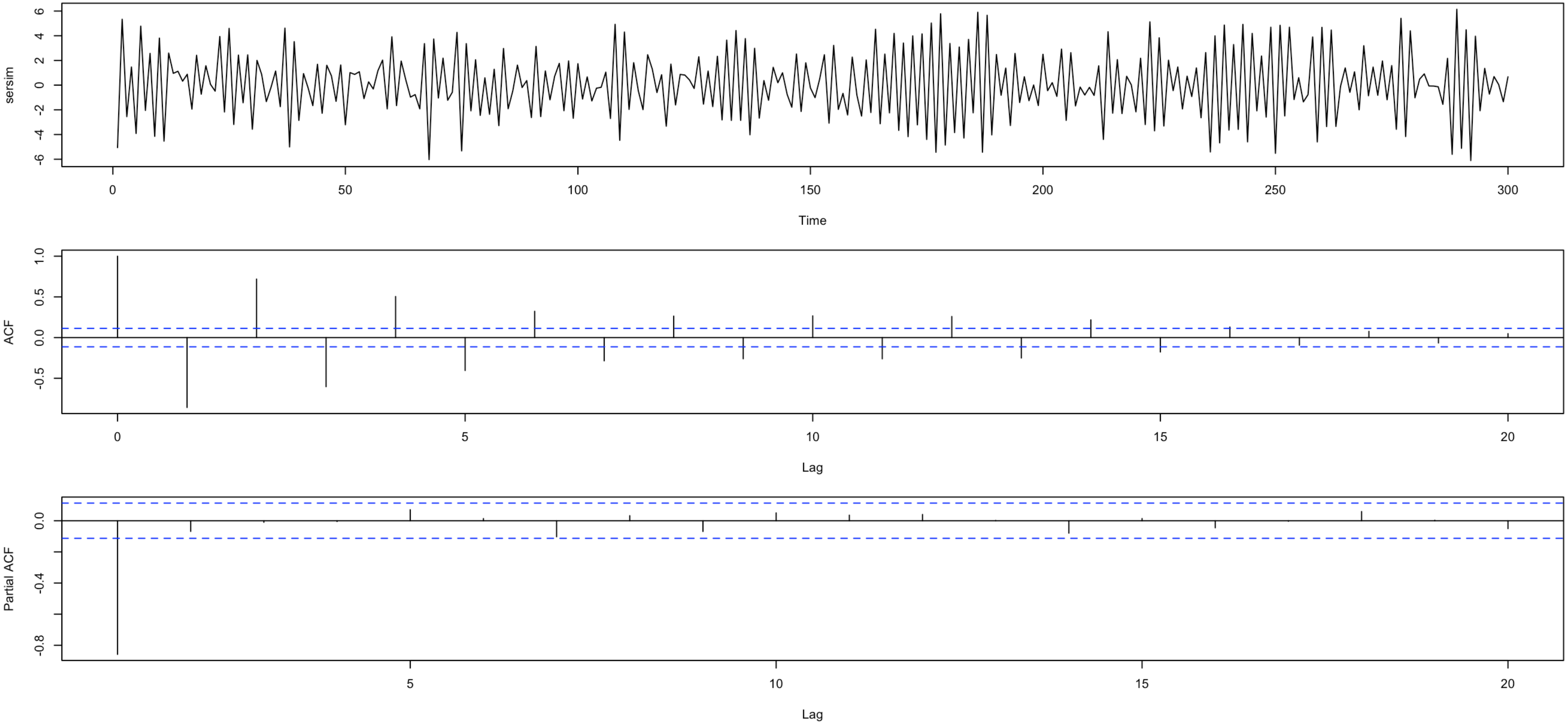


Identificazione AR(1) - simulazione

```
set.seed(12345)
par(mfrow=c(3,1),mar=c(4.5,4.5,1,1))
sersim <- arima.sim(n = 300, list(ar = c(0.9,
+          0), ma = c(0, 0)), sd = sqrt(2))
plot(sersim)
acfsim <- acf(sersim,20,type="correlation")
pacfsim <- pacf(sersim,20)
```

Identificazione AR(1)

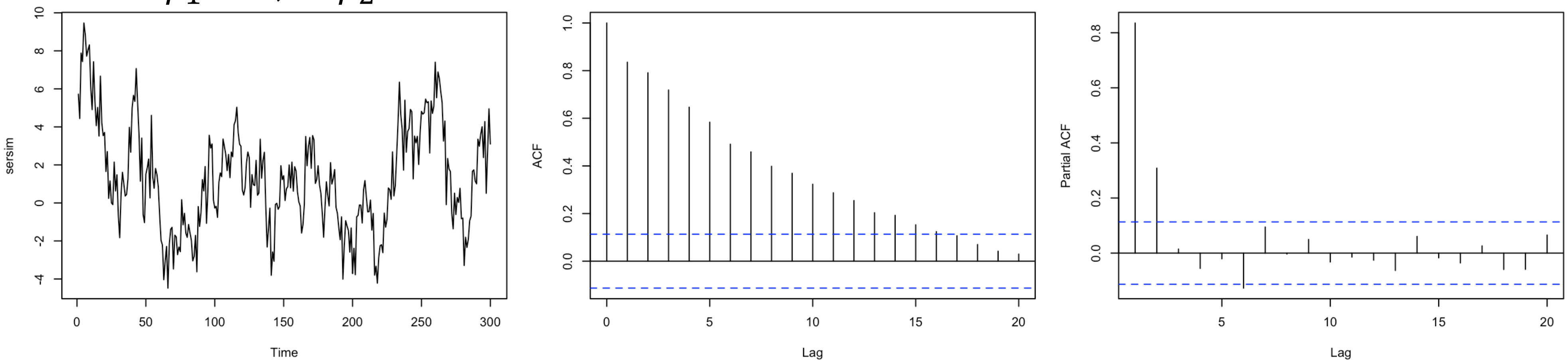
ACF e PACF di un AR(1) con $\phi_1 < 0$, in questo caso pari a -0.9

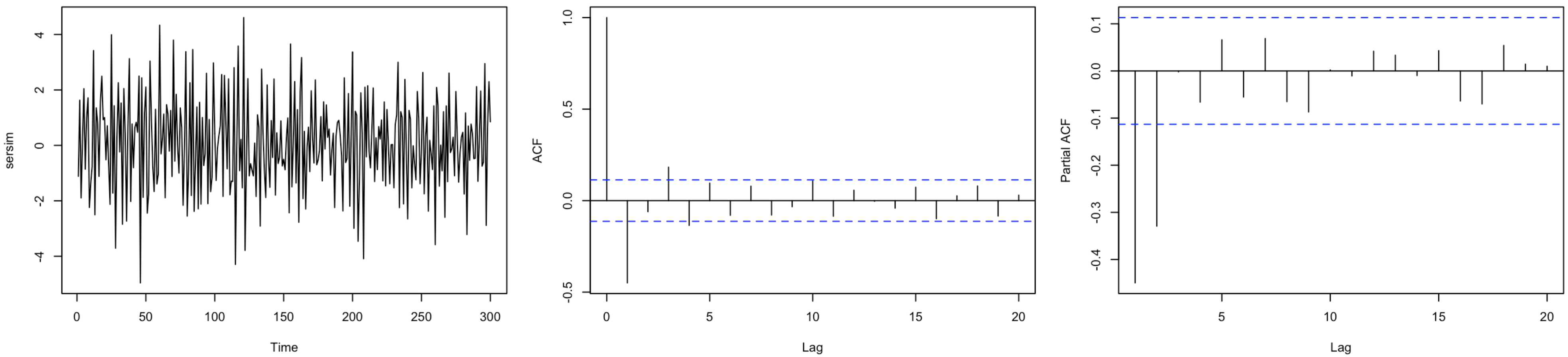


Identificazione AR(2)

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + a_t$$

$\varphi_1 > 0, \quad \varphi_2 > 0$



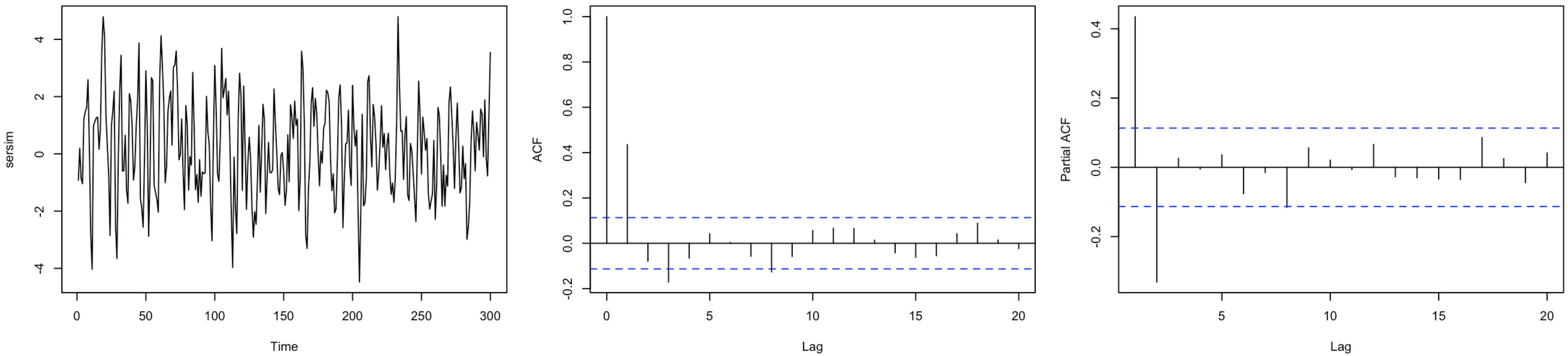


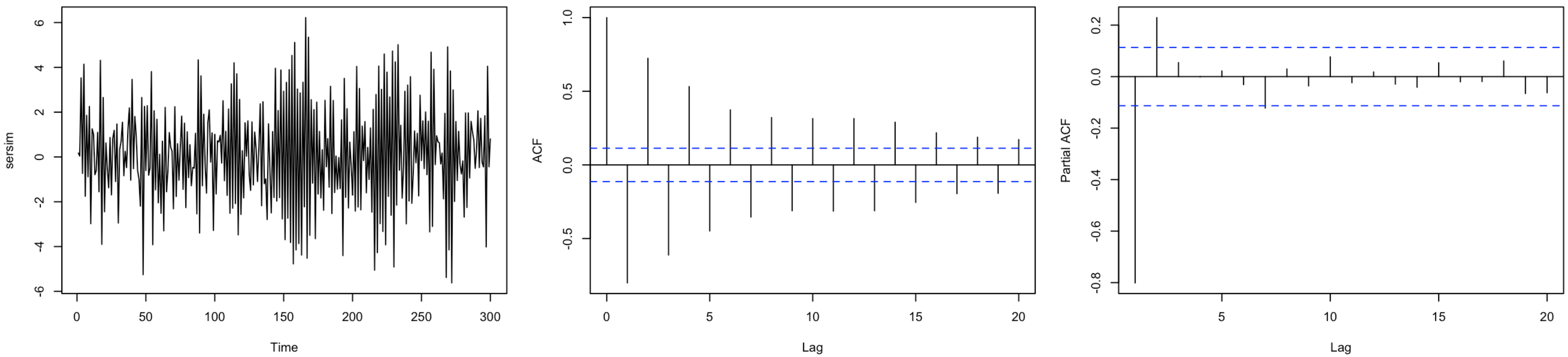
$\varphi_1 < 0, \quad \varphi_2 < 0$

Identificazione AR(2)

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + a_t$$

$\varphi_1 > 0, \quad \varphi_2 < 0$





$\varphi_1 < 0, \quad \varphi_2 > 0$

Identificazione AR(p)

ACF: decresce lentamente all'aumentare di k

PACF: è nulla per k > p

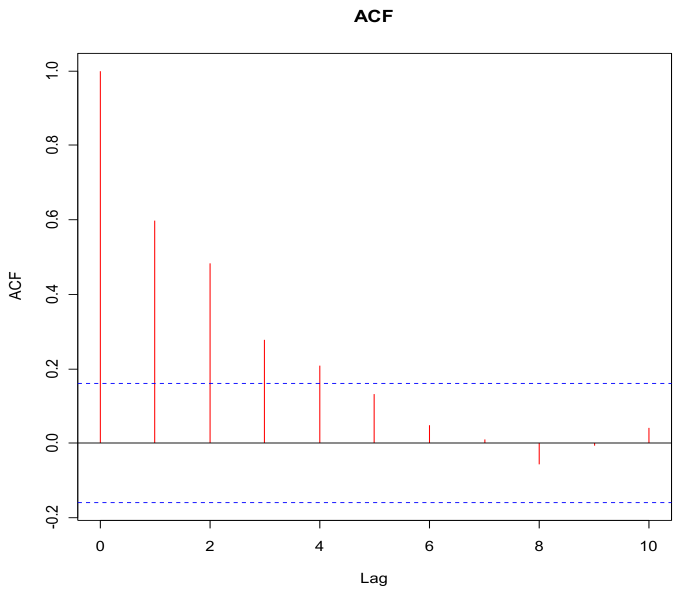
Riportare sul grafico ACF e PACF le linee di significatività, per serie osservate Xt

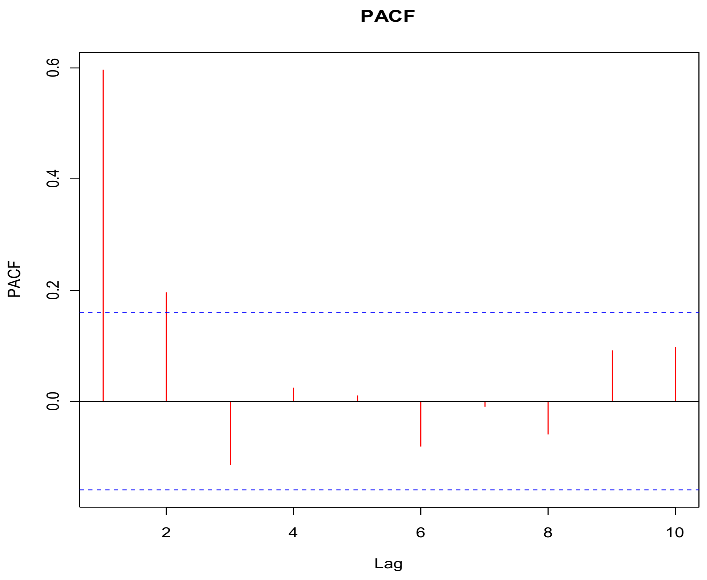
Identificazione AR(p)

Per verificare se ACF e PACF sono significativamente diversi da zero, si usa la regola empirica:

$$|\hat{\rho}| > 2\sqrt{\frac{1}{n}}$$

Esempio: ACF e PACF con livelli di significativià





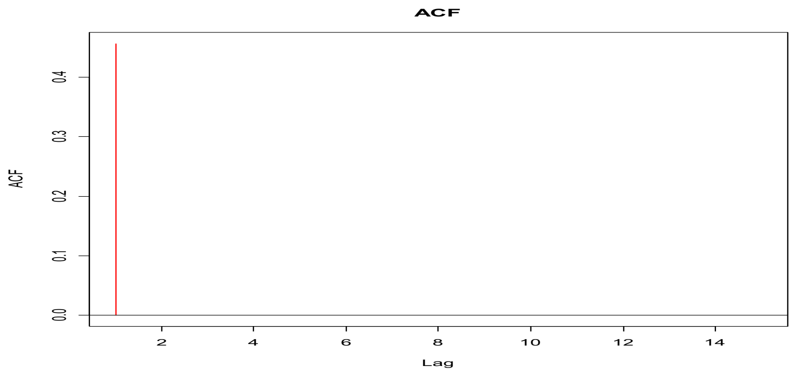
Identificazione MA(1)

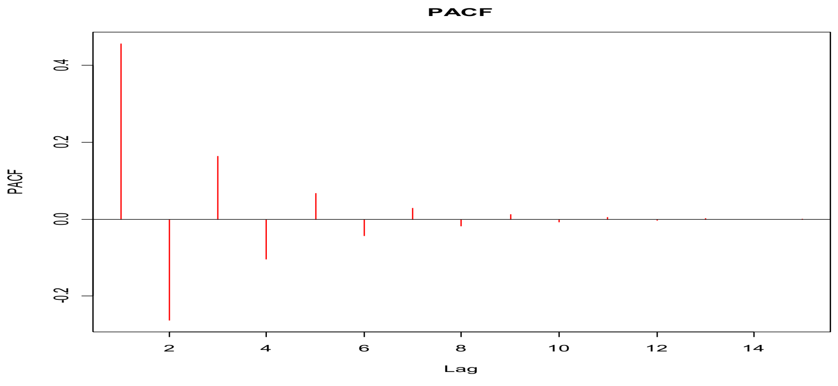
Modello MA(1):

$$X_t = a_t - \theta_1 a_{t-1}$$

ACF

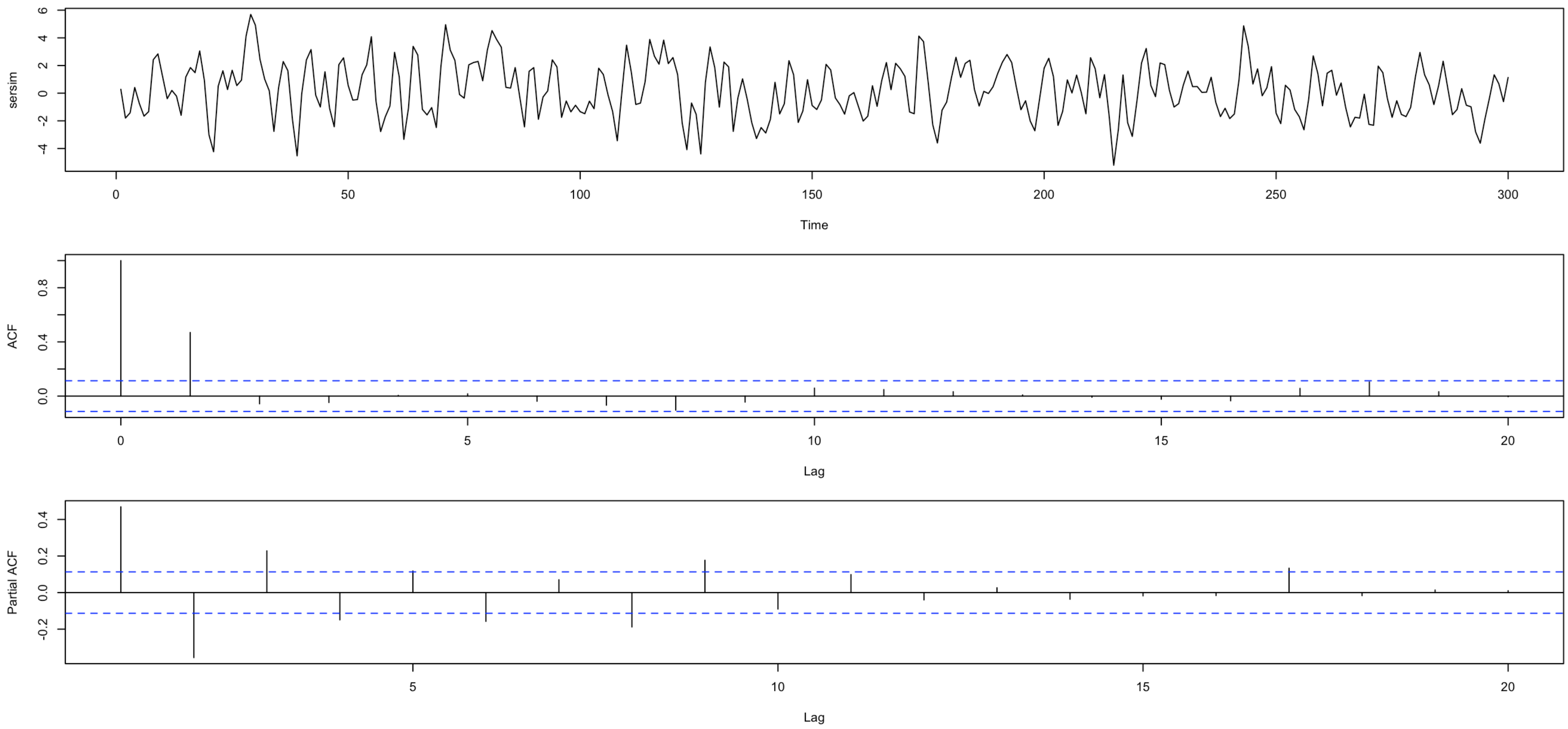
PACF





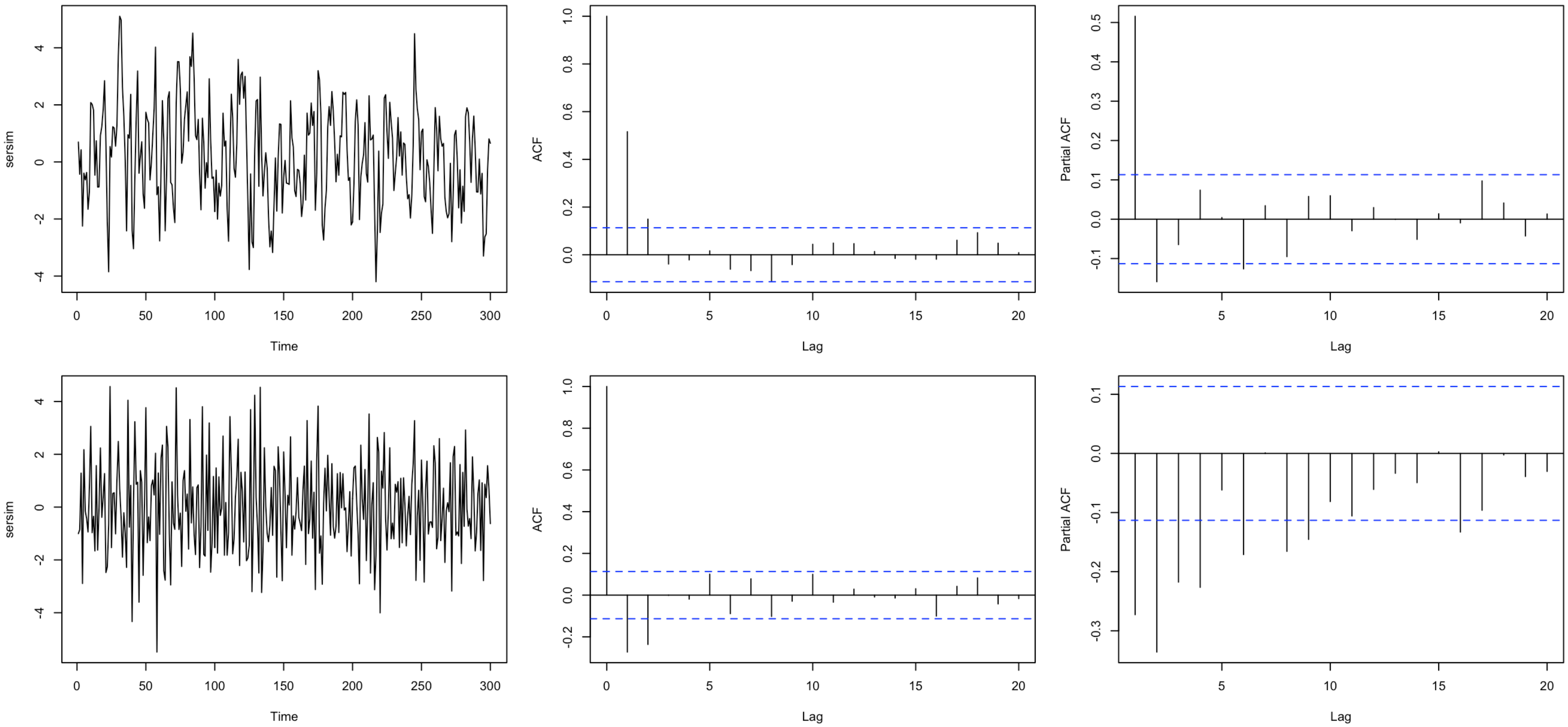
Identificazione MA(1)

$$Z_t = \varepsilon_t + 0.9\varepsilon_{t-1}$$



Identificazione MA(1)

$$\theta_1 > 0, \quad \theta_2 > 0$$



$$\theta_1 < 0, \quad \theta_2 < 0$$

Identificazione MA(q)

ACF: è nulla per $k > q$

PACF: decresce lentamente all'aumentare di k

Riportare sul grafico ACF e PACF le linee di significatività, per serie osservate X_t

Modelli ARMA(p,q)

$$Z_t = \overbrace{\varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \ldots + \varphi_p Z_{t-p}}^{AR(p)} + a_t \underbrace{- \vartheta_1 a_{t-1} - \vartheta_2 a_{t-2} + \ldots - \vartheta_q a_{t-q}}_{MA(q)},$$

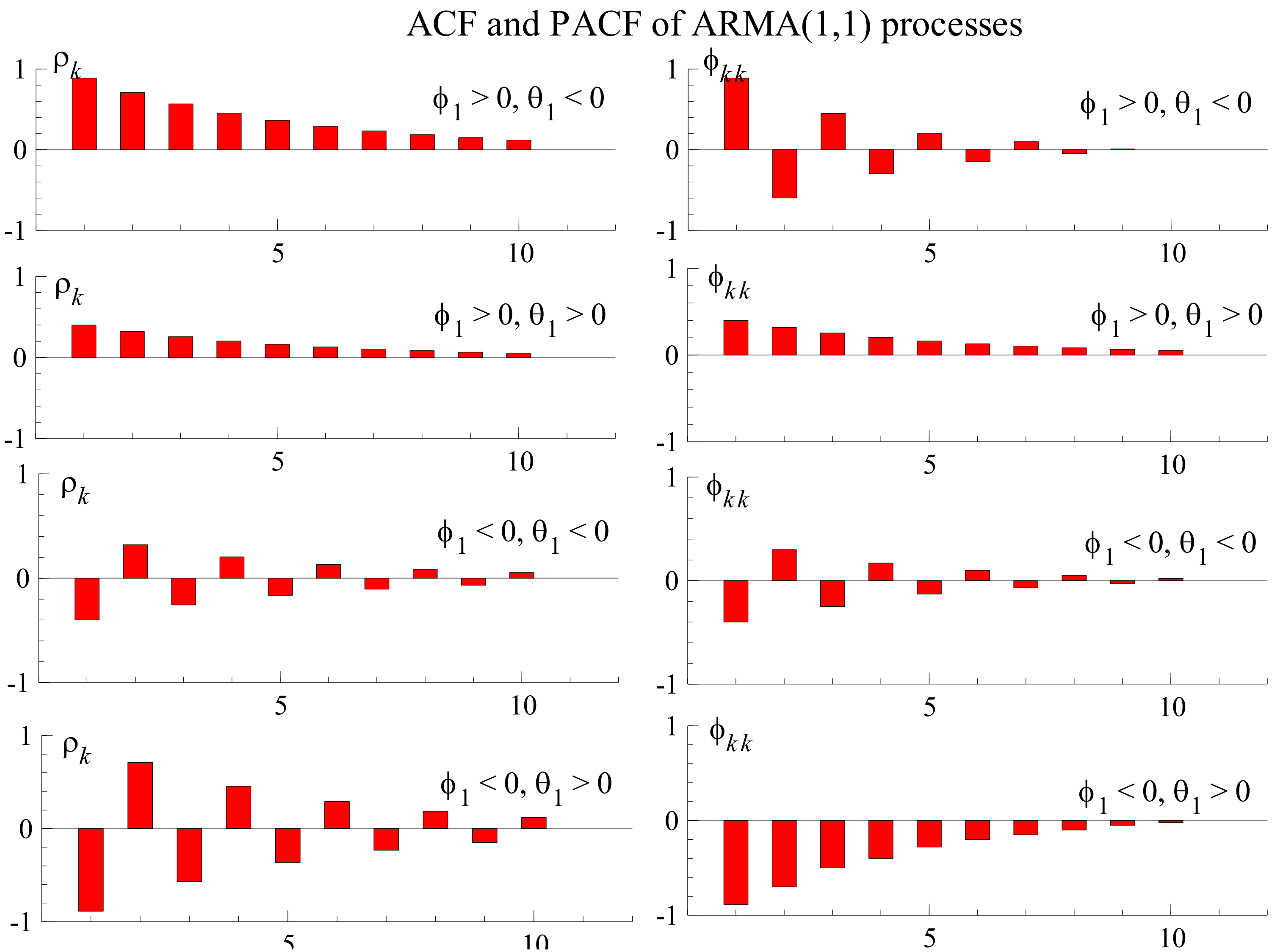
Questo è il modello Autoregressivo a Media Mobile ARMA (p, q). Casi particolari sono:

$$\text{ARMA(p,0)} = AR(p) \quad : Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_p Z_{t-p} + a_t$$

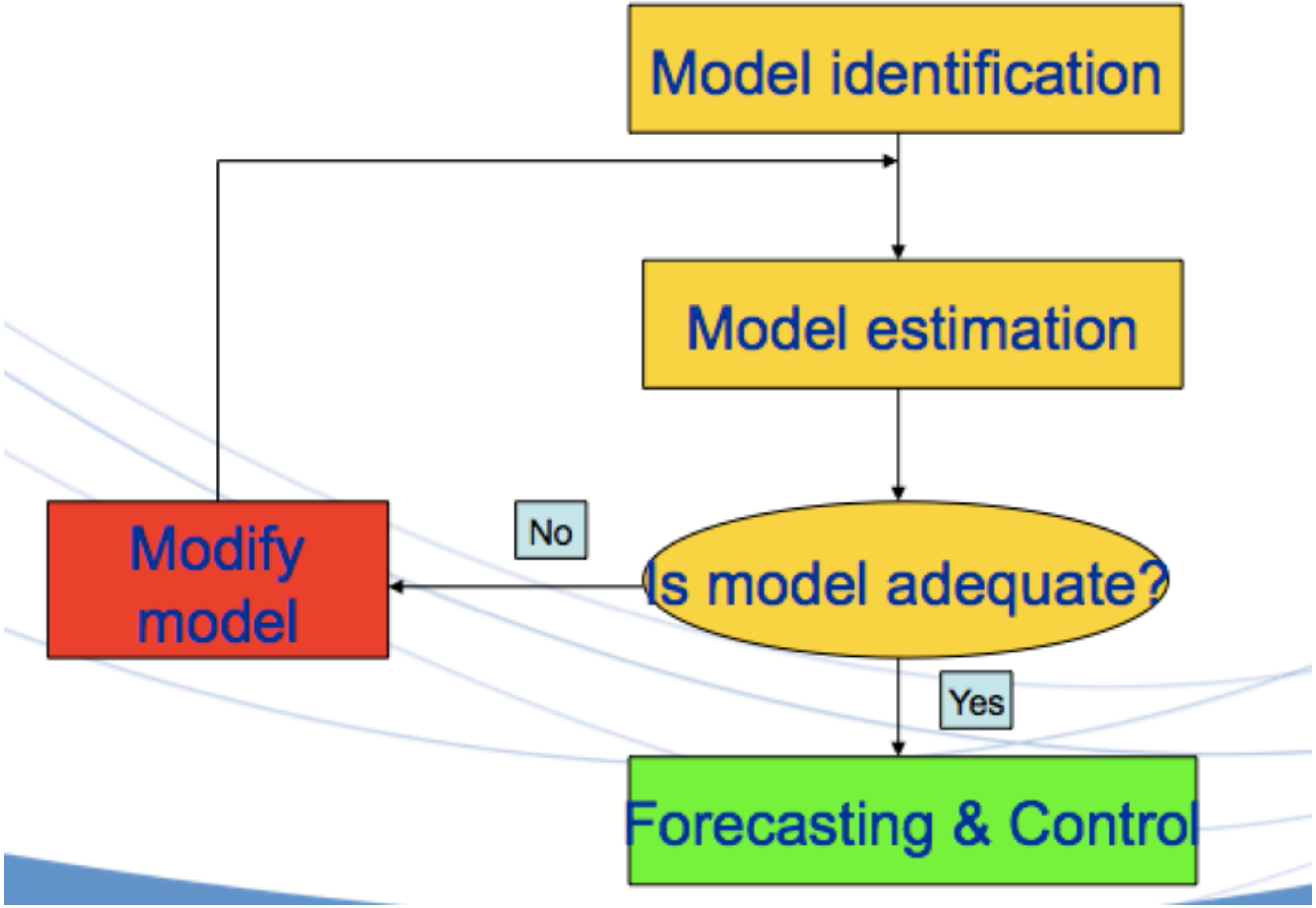
$$ARMA(0,q) = MA(q) \quad : Z_t = a_t - \vartheta_1 a_{t-1} + \cdots - \vartheta_q a_{t-q}$$

$$ARMA(1,1) \quad : Z_t = \varphi_1 Z_{t-1} + a_t + \vartheta_1 a_{t-1}$$

Modelli ARMA(1,1)



BOX-JENKINS


Model identification
Model estimation
Is model adequate?
No
Modify model
Yes
Forecasting & Control

Identificazione ARMA(p,q)

Modello	ACF	PACF
AR (p)	Diminuisce esponenzialmente	nulla per k > p
MA (q)	Nulla per k > q	Diminuisce esponenzialmente
ARMA (p,q)	Nulla per k > (q-p)	Nulla per k > (p-q)

Passaggi per identificazione modello ARMA(p,q)

"The art of a time series analyst's model identification is very much like the method of an FBI agent's criminal search. Most criminals disguise themselves to avoid being recognized. This is also true of ACF and PACF" (Wei, 1990)

Step1: Grafico della serie e trasformazioni (log, differenza, differenza percentuale,...)

Step 2: Calcolare ACF e PACF campionarie

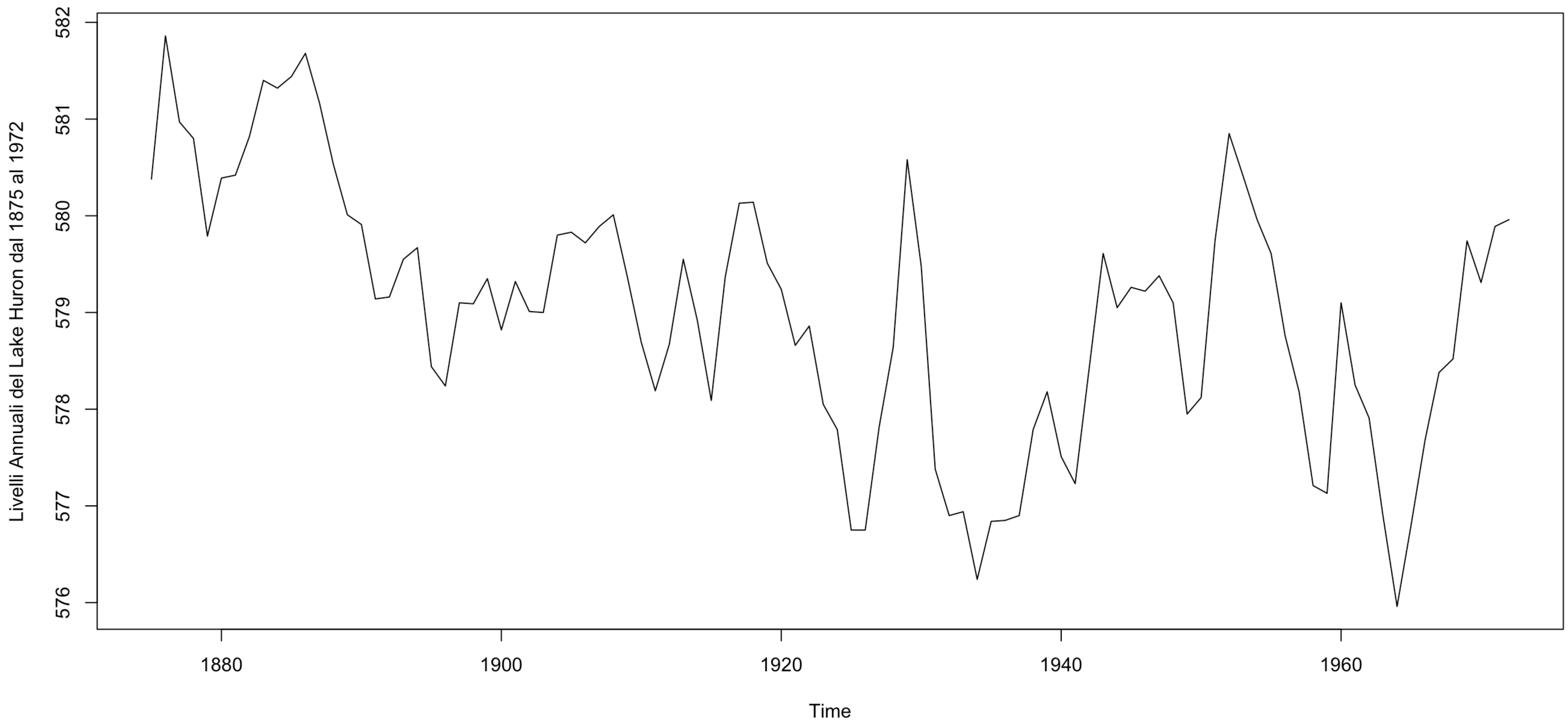
Step 3: Identificare il modello ARMA(p,q) da stimare

Stima modello ARMA(p,q)

1. AR(p) OLS corretto consistente ed efficiente, quindi MLE
2. MA(q) OLS non è possibile perché I regressori sono incogniti. Si utilizza la stima MLE che richiede un’ipotesi sulla distribuzione dei termini di disturbo e procedure numeriche di massimizzazione (Algoritmo di Newton)
3. ARMA(p,q) vedi punto 2

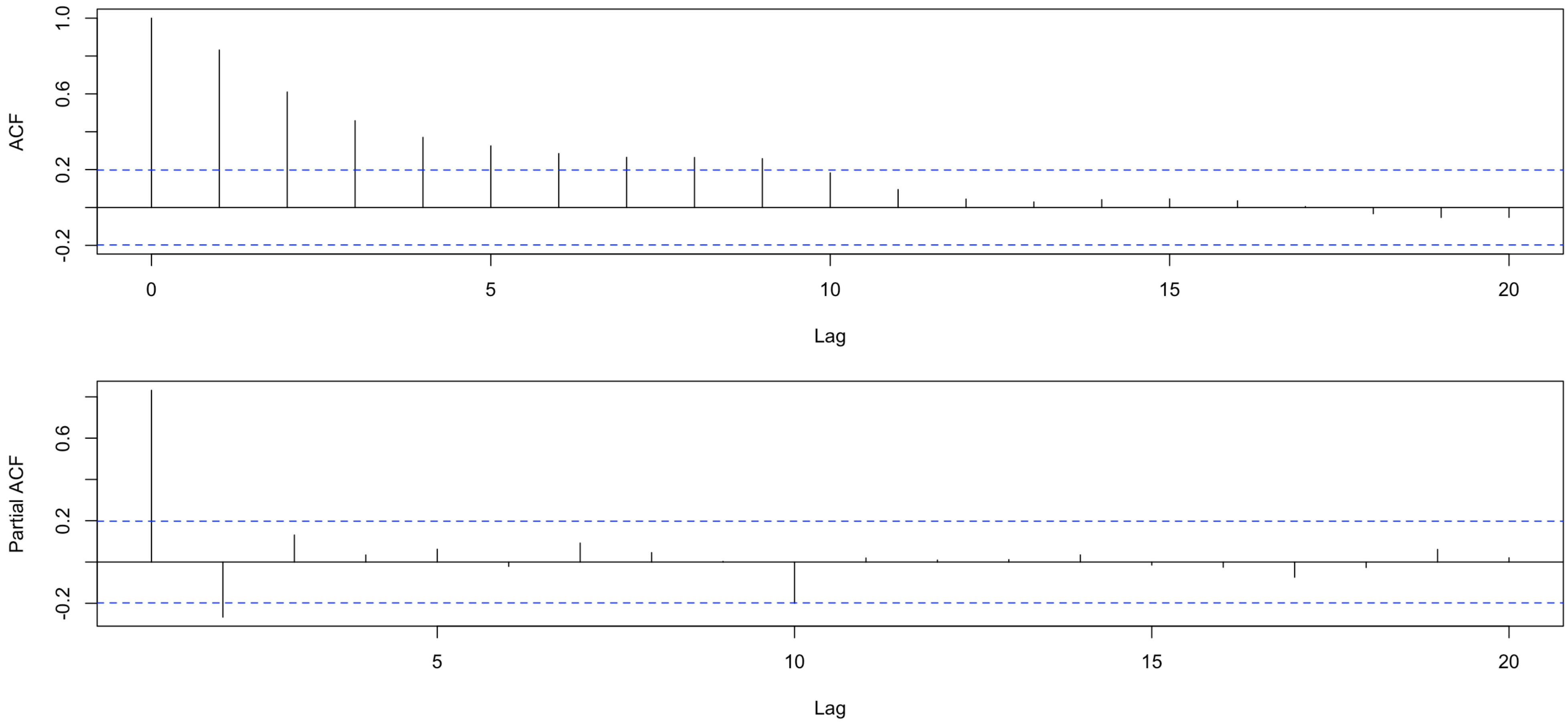
Esempio: Lake Huron

- Utilizziamo la serie "Annual levels of the Lake Huron from 1875 to 1972"



Esempio: Lake Huron

- Calcoliamo ACF e PACF, k=20



Esempio: Lake Huron

Stimiamo il modello AR(2):

1. I valori dei coefficienti
2. Lo standard error σ
3. La t-student (= coefficiente / s.e.)
4. Il p-value

```
arima(LakeHuron,order=c(2,0,0))
Coefficients:
         ar1      ar2  intercept
      1.0436  -0.2495   579.0473
s.e.  0.0983   0.1008     0.3319
sigma^2 estimated as 0.4788:
log likelihood = -103.63,  aic = 215.27
library(lmtest)
coeftest(stilak)
z test of coefficients:
            Estimate Std. Error   z value Pr(>|z|)
ar1         1.043614   0.098283   10.6185  < 2e-16 ***
ar2        -0.249498   0.100792   -2.4754  0.01331 *
intercept 579.047322   0.331876 1744.7727  < 2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostic checking e criteri di selezione

- Dopo aver stimato il modello, dobbiamo verificare la bontà del modello stesso.

- L'assunzione è che il residuo stimato sia un "White noise"
  - No autocorrelazione
  - Distribuzione Normale

- Se abbiamo più modelli possibili → applicare criterio di selezione

Diagnostic checking: Lake Huron