Serie storiche

Introduzione

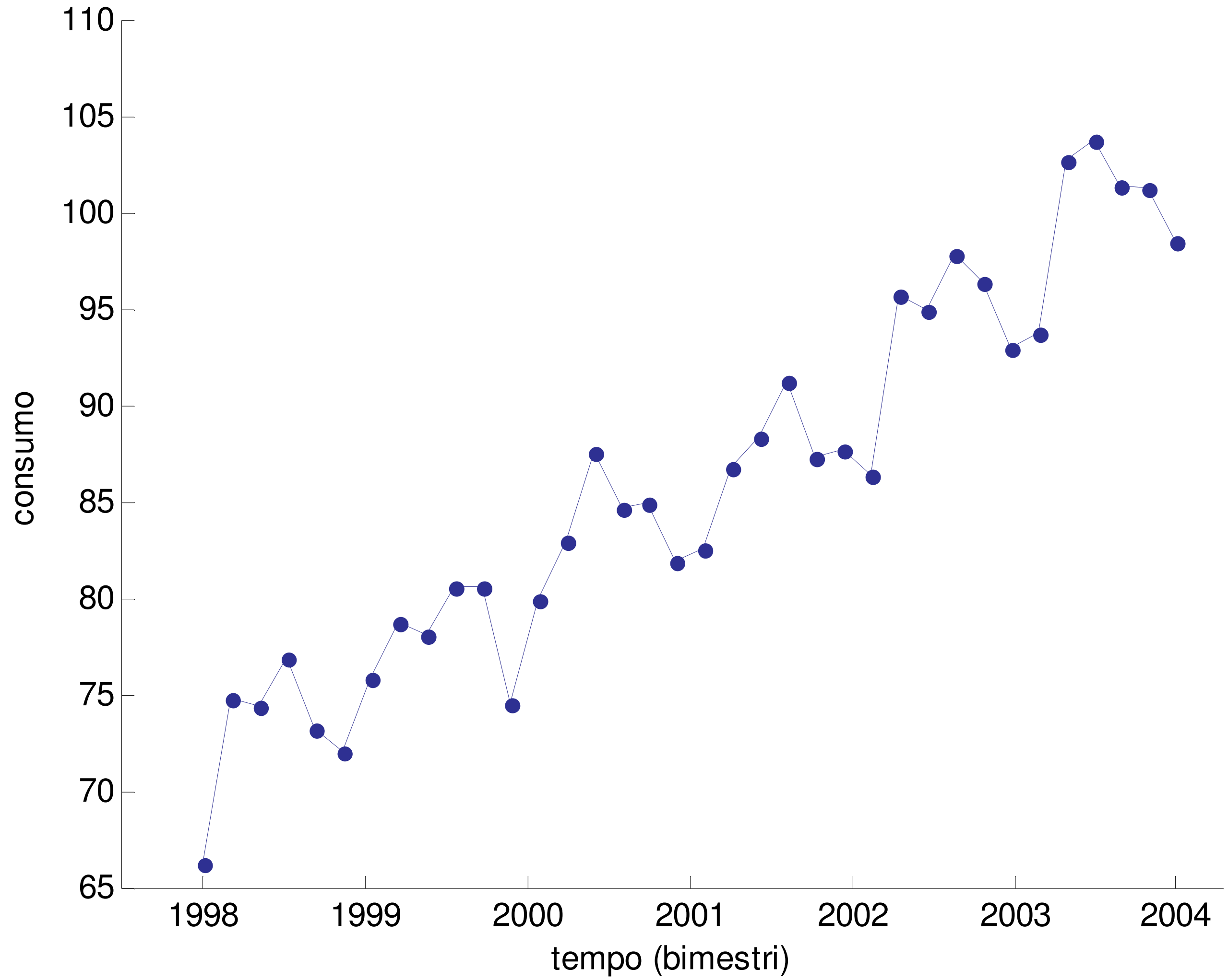
- Per alcuni dataset, l'attributo target è soggetto ad un'evoluzione temporale e risulta associato ad istanti di tempo successivi.
- I modelli di analisi delle serie storiche si propongono di identificare eventuali regolarità, con l'obiettivo di ricavare predizioni per il futuro.
- Vedremo alcuni esempi, i principali indicatori di valutazione, i metodi di scomposizione e quelli predittivi.

Definizione di serie storica

- Una *serie storica* è una sequenza di valori $\{y_t\}$ assunti da una grandezza misurabile in corrispondenza di specifici istanti temporali t.
  - Istanti di tempo t discreti: serie *a tempo discreto*.
  - Altrimenti, serie *a tempo continuo.*

- Analizzeremo serie a tempo discreto e supporremo gli istanti t distribuiti uniformemente.

Consumo elettrico bimestrale



Modello di serie storica

- Supponiamo che ciascuna osservazione y_t rappresenti la realizzazione di una variabile casuale $Y_{t.}$
- Un modello della serie storica $\{y_t\}$ consiste nella definizione della distribuzione di probabilità congiunta della sequenza di variabili casuali $\{Y_t\}$.
- I modelli predittivi analizzano le serie per evidenziare le regolarità che si sono verificate nel passato, per effettuare previsioni.
- Detta f_t una predizione del valore casuale Y_t, cerchiamo F:

$$f_{t+1} = F(y_t, y_{t-1}, \ldots, y_{t-k+1})$$

idonea a rappresentare la serie storica.

Valutazione dei modelli

- Nella fase di identificazione e sviluppo di un modello, la misura dell'accuratezza permette di confrontare modelli diversi.

- Dopo averlo sviluppato e utilizzato, occorre svolgere periodicamente il monitoraggio dell'accuratezza, per evidenziare eventuali anomalie o inadeguatezze.
  - *Errore di predizione*:

$$e_t = y_t - f_t$$

  - *Errore di predizione percentuale*:

$$e_t^P = \frac{y_t - f_t}{y_t} \times 100$$

Misure di distorsione

- *Errore medio*

$$ME = \frac{\sum_{t=1}^{k} e_t}{k} = \frac{\sum_{t=1}^{k} (y_t - f_t)}{k}$$

- *Errore percentuale medio*

$$MPE = \frac{\sum_{t=1}^{k} e_t^P}{k}$$

- Tra due modelli si ritiene migliore quello con errore medio più prossimo a zero.

Misure di dispersione

- *Scarto medio assoluto (MAD)*

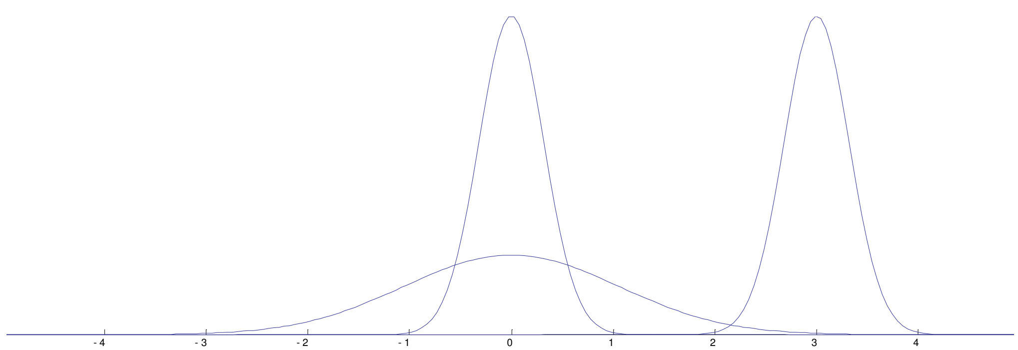
$$MAD = \frac{\sum_{t=1}^{k} |e_t|}{k} = \frac{\sum_{t=1}^{k} |y_t - f_t|}{k}$$

- *Scarto percentuale medio assoluto (MAPE)*

$$MAPE = \frac{\sum_{t=1}^{k} |e_t^P|}{k}$$

- *Scarto quadratico medio (MSE)*

$$MSE = \frac{\sum_{t=1}^{k} |e_t^2|}{k} = \frac{\sum_{t=1}^{k} (y_t - f_t)^2}{k}$$



Segnale di tracking

- Indicatore utile in fase di monitoraggio del modello

$$TS_k = \frac{\sum_{t=1}^{k} e_t}{\sum_{t=1}^{k} |e_t|}.$$

  - $TS_k \approx 0 \Rightarrow$ Previsioni in linea con quelle osservate
  - $TS_k \approx -1 \Rightarrow$ Previsioni inferiori a quelle osservate
  - $TS_k \approx 1 \Rightarrow$ Previsioni superiori a quelle osservate

- Il segnale di tracking viene confrontato con una soglia e se la condizione è violata, viene segnalata un'anomalia.

Analisi delle componenti

$$Y_t = g(M_t, Q_t, \varepsilon_t)$$

- **Tendenza**. Andamento medio della serie nel tempo (M_t).
  - Si cerca di approssimare M_t con funzioni semplici.
- **Stagionalità**. Fluttuazioni di periodicità regolare (Q_t).
  - Oscillazioni Q_t legate ai cicli naturali.
- **Fluttuazione casuale**. Variazioni ε_t non spiegate dalle altre componenti.
  - In genere si richiede un rumore bianco (μ 0 e σ costante).
- Si possono considerare anche altre componenti, come ad esempio la ciclicità economica.
  - Periodo di lunga durata.

Media mobile

- La media mobile $m_t(h)$ è calcolata come media aritmetica di h osservazioni consecutive di una serie $\{y_t\}$.
- *Media mobile centrata (h dispari)*

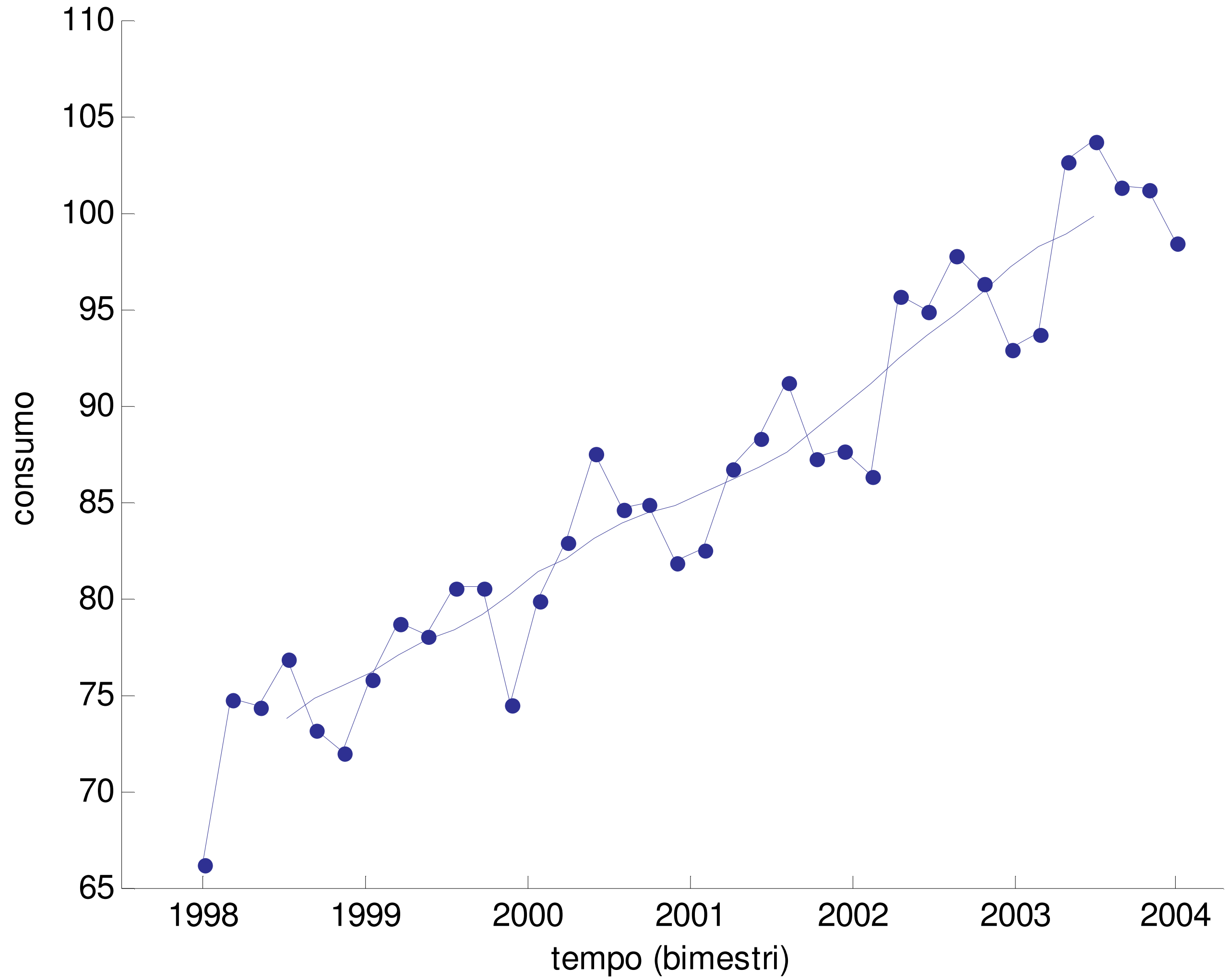
$$m_t(h) = \frac{y_{t+(h-1)/2} + y_{t+(h-1)/2-1} + \cdots + y_{t-(h-1)/2}}{h}$$

- *Media mobile centrata (h pari)*

$$m_t(h) = \frac{y_{t+h/2} + y_{t+h/2-1} + \cdots + y_{t-h/2+1}}{2h} + \frac{y_{t+h/2-1} + y_{t+h/2-2} + \cdots + y_{t-h/2}}{2h}$$

  - Media delle medie mobili centrate in (t-1/2) e (t+1/2) su h/2 osservazioni
- Può essere impiegata per effettuare previsioni.

Media mobile per consumo energia



Scomposizione di una serie

- Attività di analisi e comprensione del fenomeno.

- Si postula la forma della funzione g che indica la dipendenza della serie dalle sue componenti.
  - *Modello additivo*

$$Y_t = M_t + Q_t + \varepsilon_t,$$

  - *Modello moltiplicativo*

$$Y_t = M_t \times Q_t \times \varepsilon_t,$$

- Altri modelli possibili
  - Es. Modello di Winters

1. Rimozione della tendenza

- Si calcola la media mobile centrata con h uguale alla stagionalità L del fenomeno, da cui la relazione:

$$M_t \approx m_t(h)$$

- In un modello moltiplicativo, si identifica la componente congiunta di stagionalità e fluttuazione casuale mediante il calcolo di $\{B_t\}$:

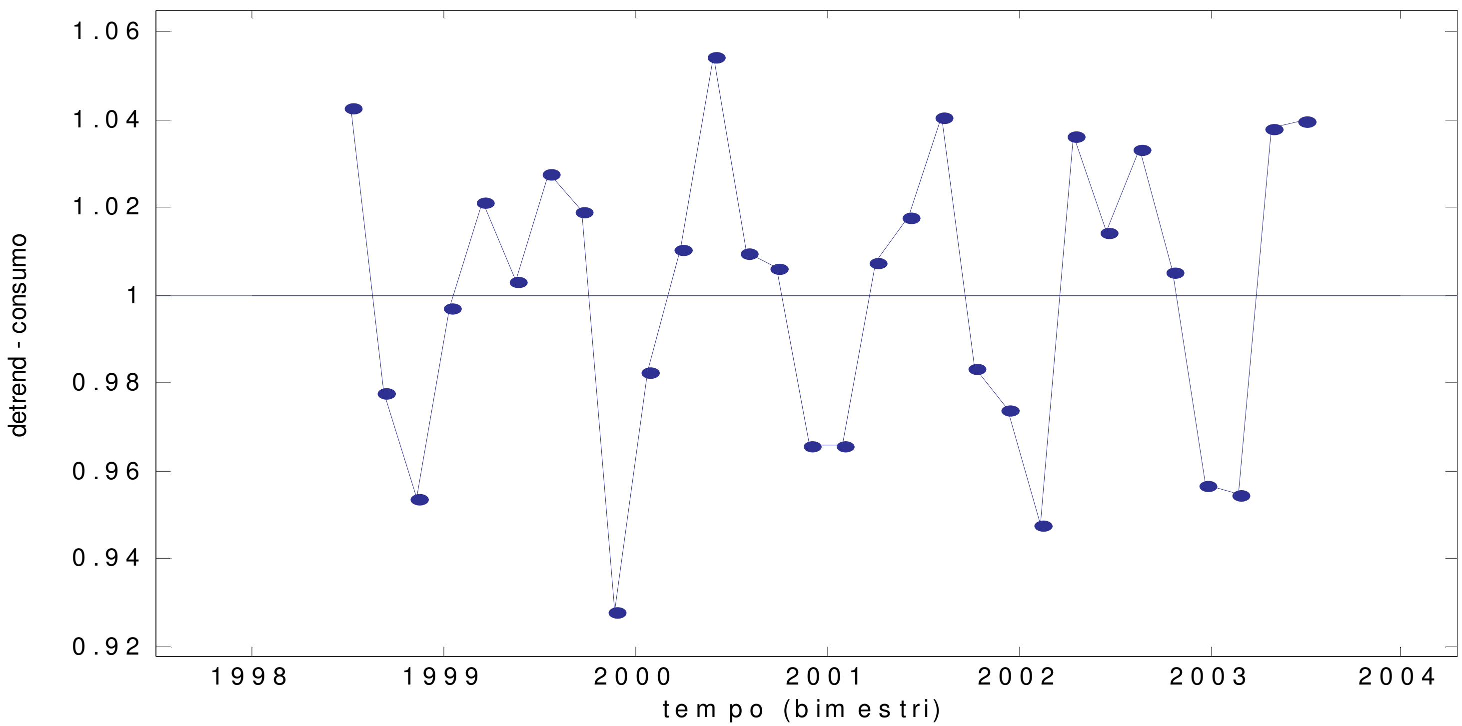
$$B_t = Q_t \epsilon_t = \frac{Y_t}{M_t} \approx \frac{Y_t}{m_t(L)}$$

- In alternativa, si può rimuovere la tendenza:
  - Differenziando

$$D_t = Y_t - Y_{t-1}, \quad t \geq 2.$$

  - Attraverso un modello di regressione della M_t.

1. Rimozione della tendenza



Rimozione moltiplicativa della tendenza della serie storica del consumo di energia elettrica

2. Identificazione della stagionalità

- In primo luogo bisogna valutare se la serie presenta una regolarità periodica.
  - L'esame visivo del grafico costituisce il punto di partenza.

- Il valore L della stagionalità è spesso suggerito dal fenomeno.
  - Per osservazioni mensili L = 12, dati giornalieri, L = 7, …

2. Identificazione della stagionalità

- Si calcolano gli indici di stagionalità Q_l, $l=1,\ldots,L$, ottenuti come media dei $Q_t\varepsilon_t$ per i periodi omologhi ad l.
  - Se la stagionalità è in mesi, i periodi omologhi al mese di gennaio, sono tutti i gennaio compresi nell'orizzonte temporale.

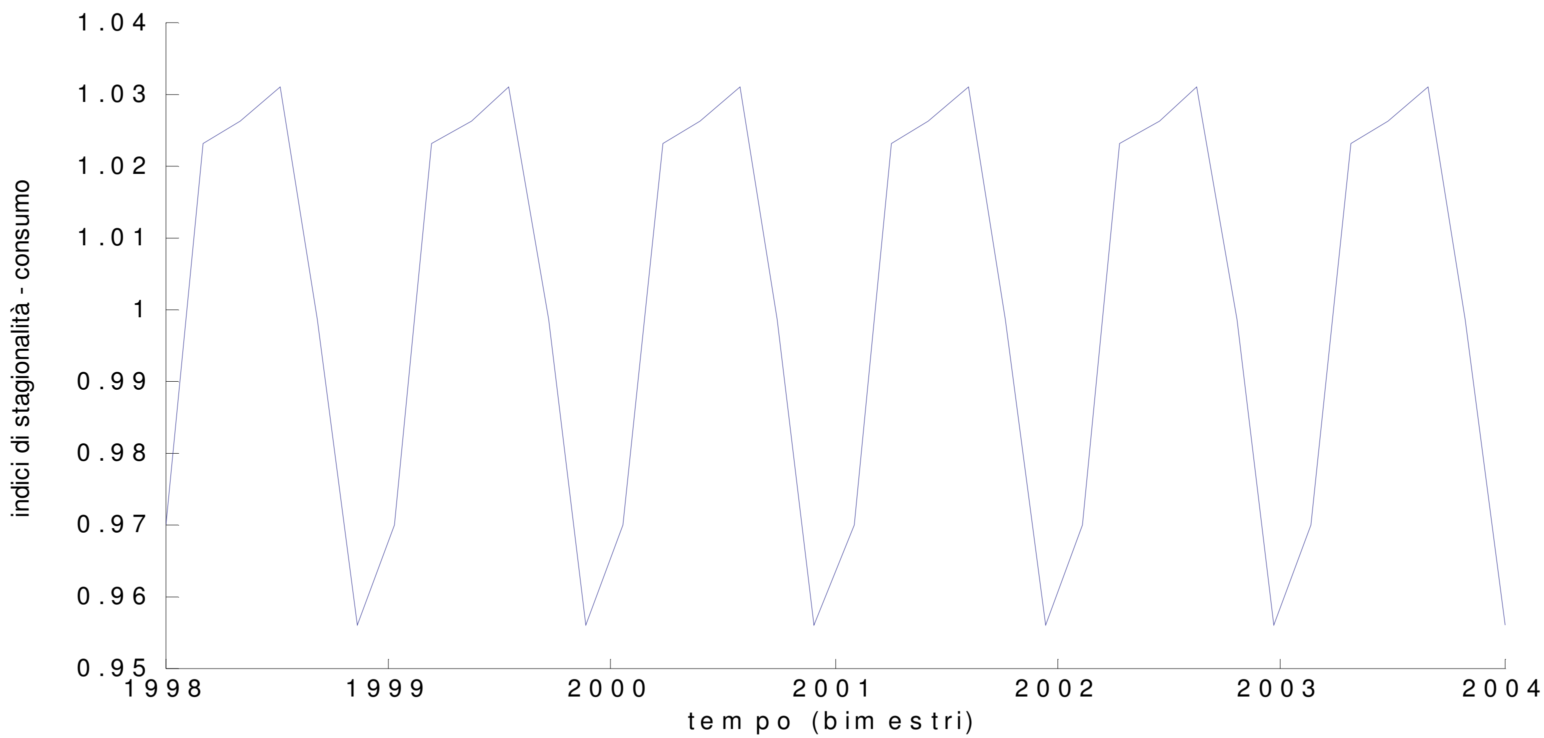
- Posto:

$$Q_l = \frac{\sum_{t \in z_l} Q_t \epsilon_t}{|z_l|}$$

con z_l insieme degli indici dei periodi omologhi a l.

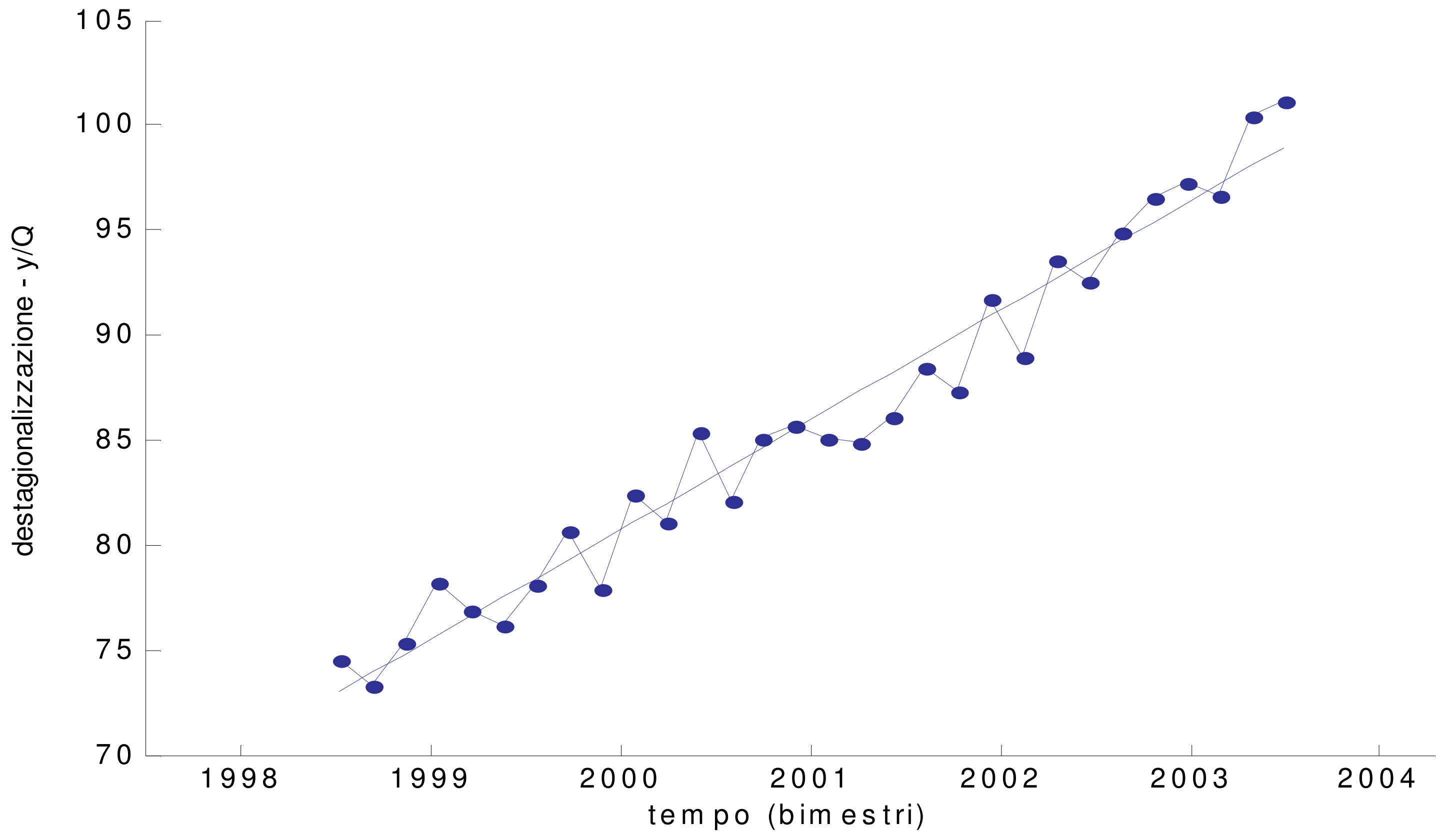
- La destagionalizzazione si calcola mediante $y_t/Q_{l(t)}$
  - $l(t)$ indica il tipo di periodo corrispondente a t.

2. Identificazione della stagionalità



Indici di stagionalità per la serie storica del consumo di energia elettrica

Rimozione della stagionalità



Destagionalizzazione e regressione per la serie storica del consumo di energia elettrica

Modelli di predizione

- I modelli di smoothing esponenziale sono tra i modelli predittivi per l'analisi di serie storiche più versatili ed accurati.

- Si sono dimostrati particolarmente adatti per la previsione di fenomeni di natura economica.

- Originariamente formulati su base empirica, hanno trovato adeguate giustificazioni teoriche.

Modello di Brown

- *Modello di smoothing esponenziale semplice*:

$$s_t = \alpha\, y_{t-1} + (1-\alpha)\, s_{t-1}$$
$$s_1 = y_1,\ \alpha \in [0,\ 1]$$

- Si dimostra che:

$$f_{t+1} = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots + \alpha(1-\alpha)^{k-1} y_{t-k+1}$$

- Il valore del parametro α è indice dell'inerzia del modello.

Modello di Holt

- Il modello di Brown non è in grado di cogliere la tendenza presente tra le componenti della serie.
- La relazione ricorsiva viene quindi modificata con:

$$s_t = \alpha y_t + (1-\alpha)(s_{t-1} + m_{t-1})$$

- Un’analoga relazione ricorsiva governa l’aggiornamento della tendenza smorzata:

$$m_t = \beta (s_t - s_{t-1}) + (1 - \beta)\, m_{t-1}$$

- La predizione per l’istante $t+1$ è:

$$f_{t+1} = s_t + m_t$$

Modello di Winters

- Se nella serie è presente anche la stagionalità è necessario estendere i modelli precedenti.

- Se ogni ciclo è costituito da L periodi, si modificano gli aggiornamenti di m_t e s_t introducendo un aggiornamento della stagionalità:

$$s_t = \alpha \frac{y_t}{q_{t-L} - L} + (1-\alpha)(s_t - 1 + m_{t-1})$$

$$m_t = \beta(s_t - s_{t-1}) + (1-\beta)m_{t-1}$$

$$q_t = \gamma \frac{y_t}{s_t} + (1-\gamma)q_{t-L}$$

- La previsione diventa quindi:

$$f_{t+1} = (s_t + m_t)q_{t\text{-}L+1}.$$

Esercizi

- Generare un dataset sintentico per il consumo di bottiglie di vino *Ponter* negli anni 2000 2007 che abbia le seguenti caratteristiche moltiplicative:
  - Nel primo Trimestre ne sono state consumate 10K unità
  - Trend di crescita lineare del 10% all'anno
  - Stagionalità trimestrale tra [0.96, 1.03] con un aumento del 7% tra il primo e secondo mese e diminuzione del 4% tra secondo e terzo.
  - Errore bianco con ampiezza in [0.96, 1.04]
- Separare le componenti e valutare il modello determinato.
- Predirre il consumo di Ponter nel primo e secondo trimestre 2008.

Sommario

- Abbiamo visto come alcuni fenomeni vengano descritti da serie storiche
- Abbiamo introdotto gli indici di valutazione
- Abbiamo scomposto le serie storiche in componenti
- Abbiamo derivato modelli di previsione
  - Modello di Brown
  - Modello di Holt
  - Modello di Winters

Nella prossima lezione

- Classificazione
  - Valutazione
  - Alberi di classificazione
  - Metodi bayesiani
  - Regressione logistica
  - Reti neurali
  - Support vector machines