Il modello di regressione lineare con dati in serie storica

Il modello di regressione lineare può essere applicato senza problemi sia a dati cross-section, sia a dati in serie storica, purché le ipotesi di base vengano rispettate.

Con dati in serie storiche, l'ipotesi a rischio è quella di incorrelazione dei disturbi.

Tuttavia, con opportuni accorgimenti, è possibile risolvere il problema.

Definizione formale di *Serie Storica*

Una serie storica è una collezione ordinata di osservazioni su una variabile d'interesse:

$$Y_t, \; t = 0, 1, \ldots, n,$$

dove l'ordinamento è tipicamente costituito dal tempo ed è essenziale per comprendere il fenomeno esaminato.

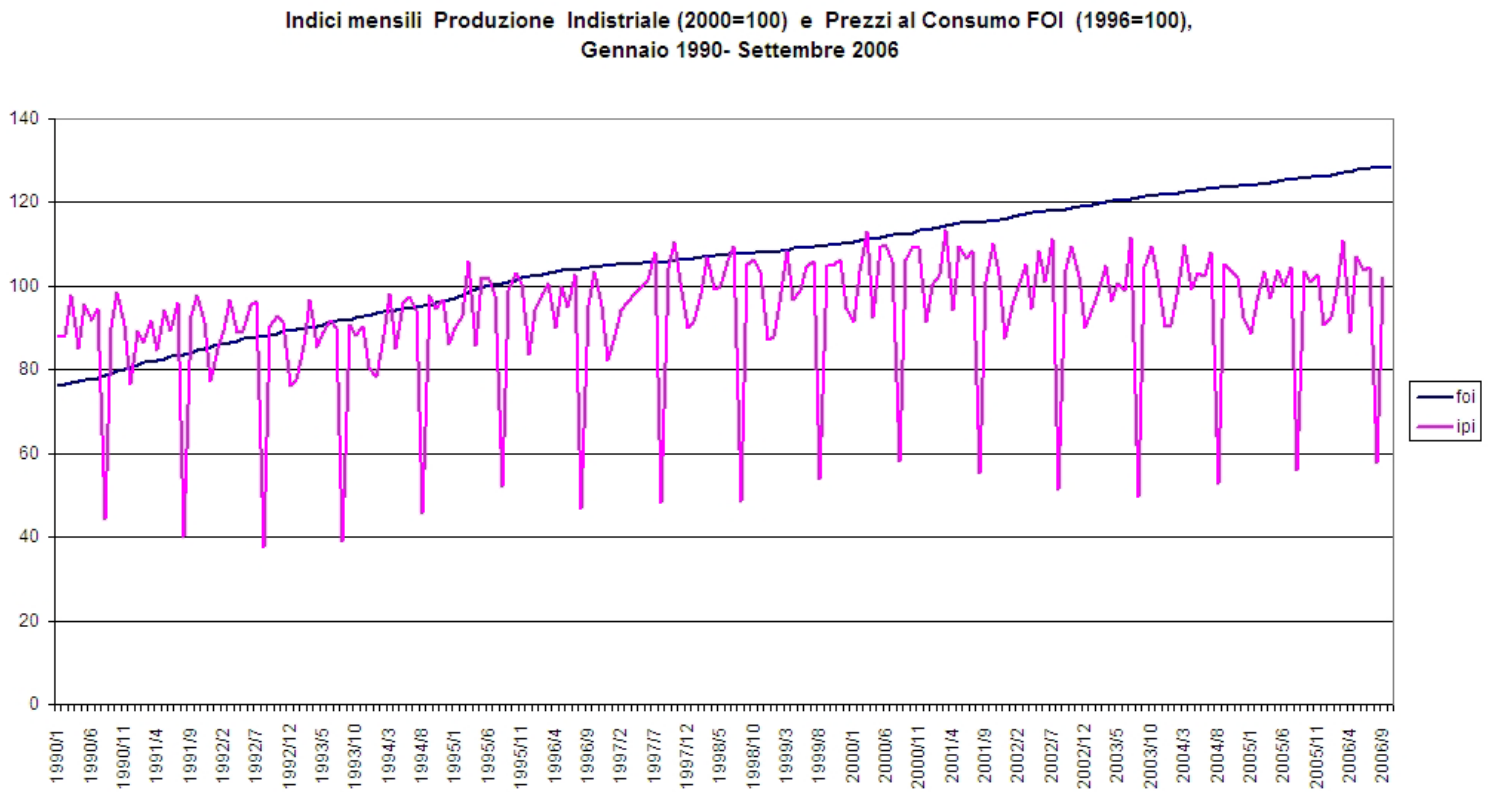
Serie Storiche 2

La maggior parte dei fenomeni reali si realizza nel corso del tempo e lo studio della loro dinamica diventa fondamentale per la comprensione dei fenomeni stessi.

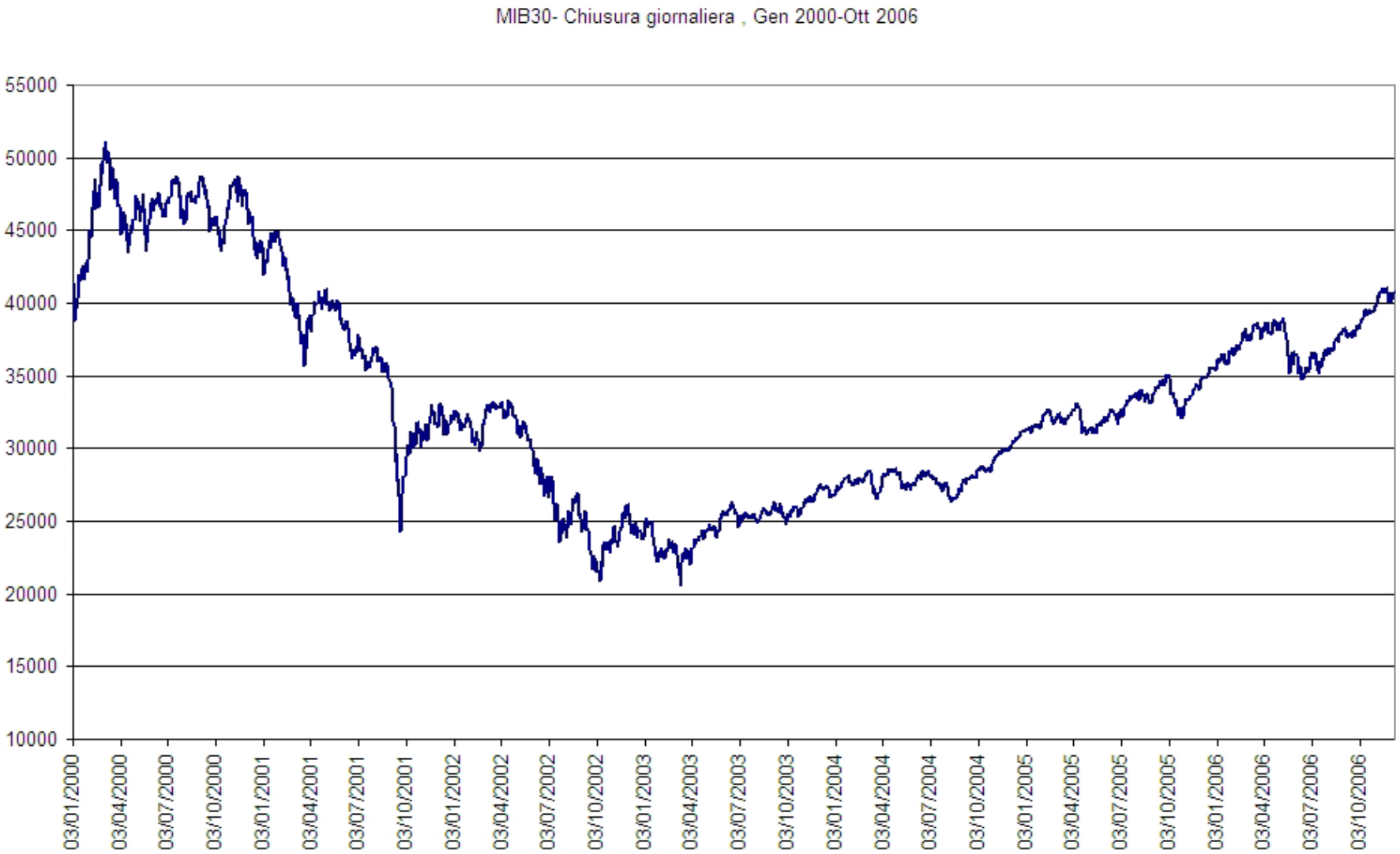
L'assunto da cui parte l'analisi delle serie storiche è che ciò che si realizzerà nel futuro, in qualche modo dipenda da quanto si è realizzato nel passato, secondo un principio generale di inerzia e stabilità.

Se pertanto si arriva a individuare la legge che lega il futuro al presente e al passato, diventano possibili sia la previsione sia il controllo.

Un paio di esempi:



Serie Storiche 4



Dovrebbe risultare evidente come per comprendere l'evoluzione del fenomeno sia fondamentale l'ordinamento temporale delle osservazioni:

→ *principio della non scambiabilità delle osservazioni*

Al contrario, l'analisi statistica classica si basa su campioni casuali di osservazioni indipendenti, ciascuna delle quali contiene informazioni sulla variabile casuale da cui è stata generata, ma non fornisce alcuna informazione sulle possibili determinazioni di un'osservazione successiva.

Al fine di tener conto del legame esistente tra le osservazioni di una successione temporale di dati, le varie tecniche di analisi delle serie storiche fanno riferimento alla teoria dei *processi stocastici*.

Una definizione di processo stocastico non rigorosa, ma intuitiva, può essere la seguente:

un processo stocastico è una sequenza infinitamente lunga di variabili casuali, ovvero, un vettore aleatorio di dimensione infinita

La conoscenza di tutte le osservazioni disponibili sino al presente diventa allora l'informazione a cui ci si può condizionare per ottenere valutazioni probabilistiche circa le realizzazioni future delle variabili casuali esaminate, qualunque sia l'obiettivo ultimo dell'analisi:

- Previsione
- Simulazione
- Individuazione di dati anomali, delle componenti elementari, …

Nei modelli STRUTTURALI tali obiettivi sono perseguibili condizionatamente all'individuazione e alla specificazione della struttura della serie storica osservata, costituita da componenti non osservabili ma direttamente interpretabili.

Le caratteristiche salienti di una serie, che vanno esplicitate coerentemente con una teoria, mediante un modello di comportamento, sono di seguito indicate.

Componenti di una serie storica:

- ***trend* (*T*)**: andamento di fondo, tendenza nel lungo periodo (crescente, decrescente; esponenziale, lineare, quadratico, ecc.);

- ***ciclo* (*C*)**: fluttuazioni periodiche di medio periodo di ampiezza rilevante;

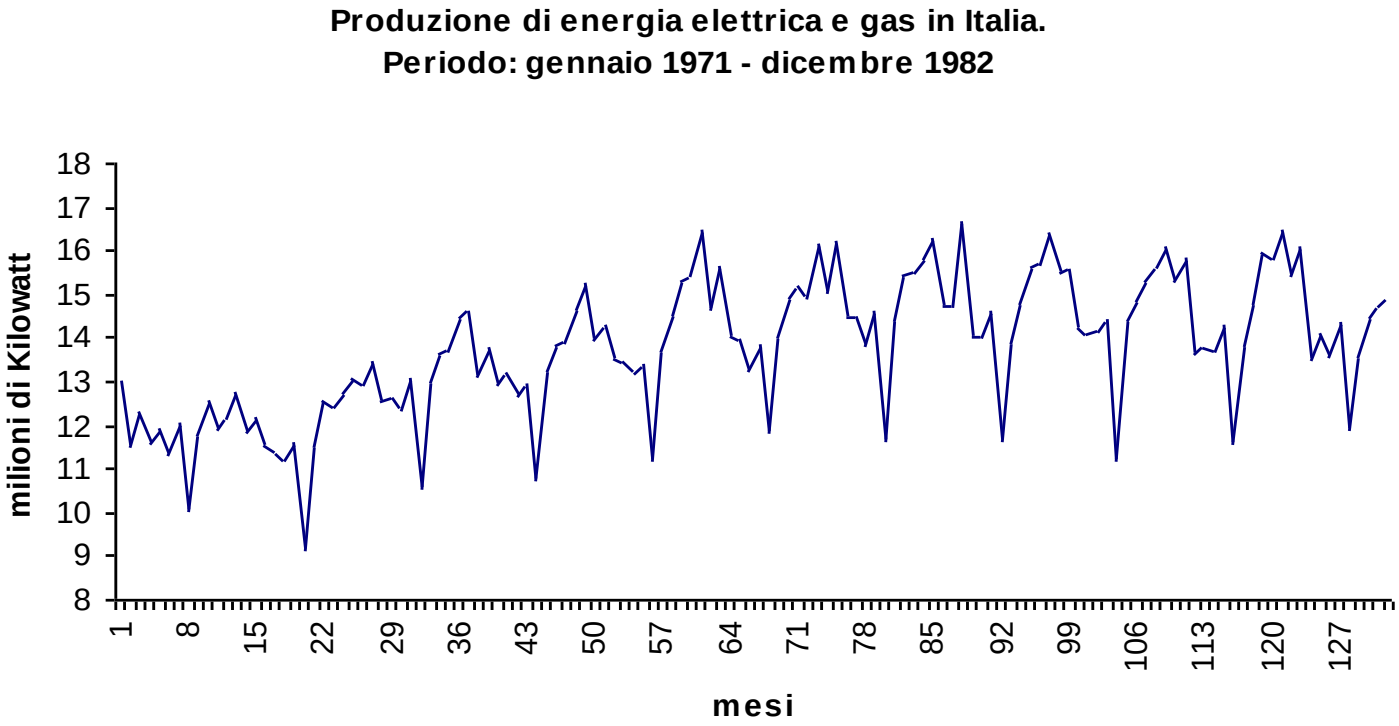
- ***stagionalità* (*S*)**: fluttuazioni periodiche più o meno regolari con cadenza infrannuale (trimestrale, mensile, settimanale, ecc.), spesso attribuibile all'alternarsi delle stagioni climatiche sul nostro pianeta;

- ***accidentale*** **(u_t)**: fluttuazioni di breve/brevissimo periodo, dovute all'insieme di cause non esplicitabili, assolutamente non prevedibili sulla base della storia passata e presente del fenomeno, anche se a volte spiegabili attraverso la conoscenza dell'azione di fenomeni esterni (catastrofi ambientali, guerre, amnistie, tassazioni 'una-tantum', ecc.).

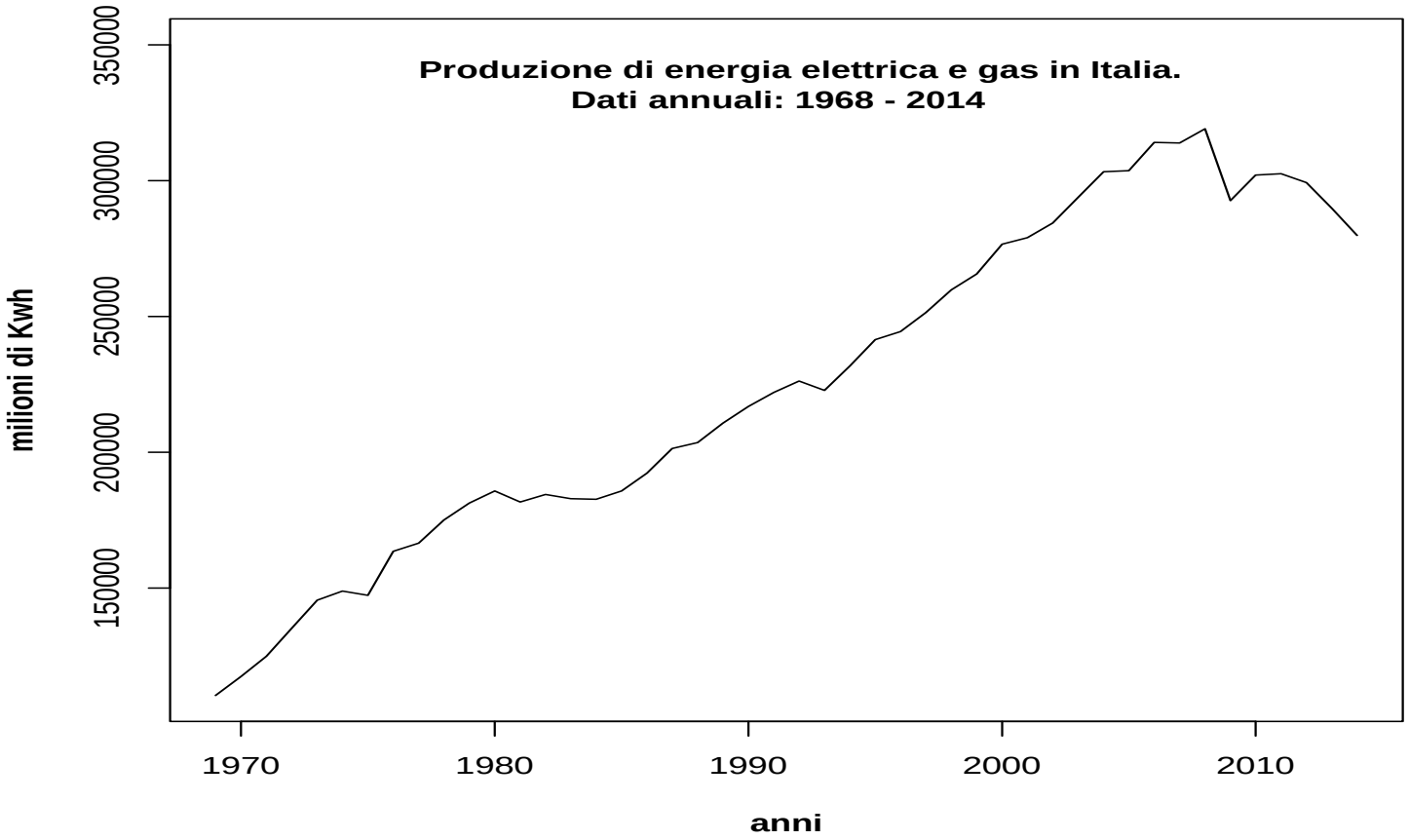
N.B.: Molto spesso le due componenti di trend e ciclo sono accorpate in un'unica componente di ciclo-trend, alla quale imputare l'andamento di medio-lungo periodo.

Se inoltre i dati derivano da una misurazione aggregata entro un intervallo (sono cioè dati di "accumulo" o di stock), o vengono rilevati con una frequenza infra-annuale abbastanza elevata (dati orari, giornalieri, settimanali), diventa fondamentale tener conto anche di una componente calendario, che spieghi gli effetti del diverso numero di giornate lavorative (e quindi dei week-end, delle festività infrasettimanali, ecc.) sui dati osservati.

Esempio: si noti la notevole differenza, in termini di ricorrenze, tra la serie *mensile* della "Produzione di energia elettrica e gas in Italia"



e quella annuale:



Un modello strutturale per serie storiche deve essere in grado di “catturare” le componenti strutturali caratteristiche, consentendo la loro stima sulla base delle osservazioni disponibili.

Esso può essere specificato come un qualsiasi modello di regressione, inserendo le opportune componenti tra i regressori.

Come per tutti i modelli di regressione, esistono due grosse categorie di modelli: i modelli additivi e quelli moltiplicativi.

Nel primo caso, si fa riferimento alla seguente scomposizione di una serie storica:

serie osservata = trend + ciclo + stagionalità + componente accidentale: $Y_t = T_t + C_t + S_t + u_t$

Nel secondo caso, le componenti strutturali agiscono in forma moltiplicativa:

serie osservata = trend $\times$ ciclo $\times$ stagionalità $\times$ componente accidentale: $Y_t = T_t \times C_t \times S_t \times u_t$

Ovviamente, un modello moltiplicativo può sempre essere ricondotto ad uno additivo, semplicemente passando dalla serie osservata a quella dei logaritmi.

Al fine di individuare le componenti strutturali eventualmente presenti nella serie osservata e la loro tipologia, anche in questo caso risulta di estrema utilità il ricorso ad una preliminare analisi esplorativa dei dati, consistente in opportune trasformazioni della serie che ne mettano in luce le principali caratteristiche.

Un criterio intuitivo per evidenziare le componenti evolutive di una serie storica a cadenza infrannuale è quello di riportare i dati in una tabella a doppia entrata, ponendo gli S periodi (trimestri, mesi, ecc.) sulle colonne e gli A anni sulle righe (con $n=A \times S$):

Anni \ Stagioni	1	2	…	S
1	Y_1	Y_2	…	Y_S
2	Y_{S+1}	Y_{S+2}	…	Y_{2S}
…	…			
A	$Y_{S(H-1)+1}$		…	Y_n

Se si confrontano i dati appartenenti alla medesima colonna, essendo essi riferiti allo stesso periodo stagionale, le variazioni riscontrate potranno essere imputate alla tendenza evolutiva di fondo, cioè al trend.

Viceversa, se si confrontano i dati appartenenti alla medesima riga, riferiti quindi a diversi periodi stagionali, le variazioni riscontrate potranno essere imputate alla componente stagionale di breve periodo.

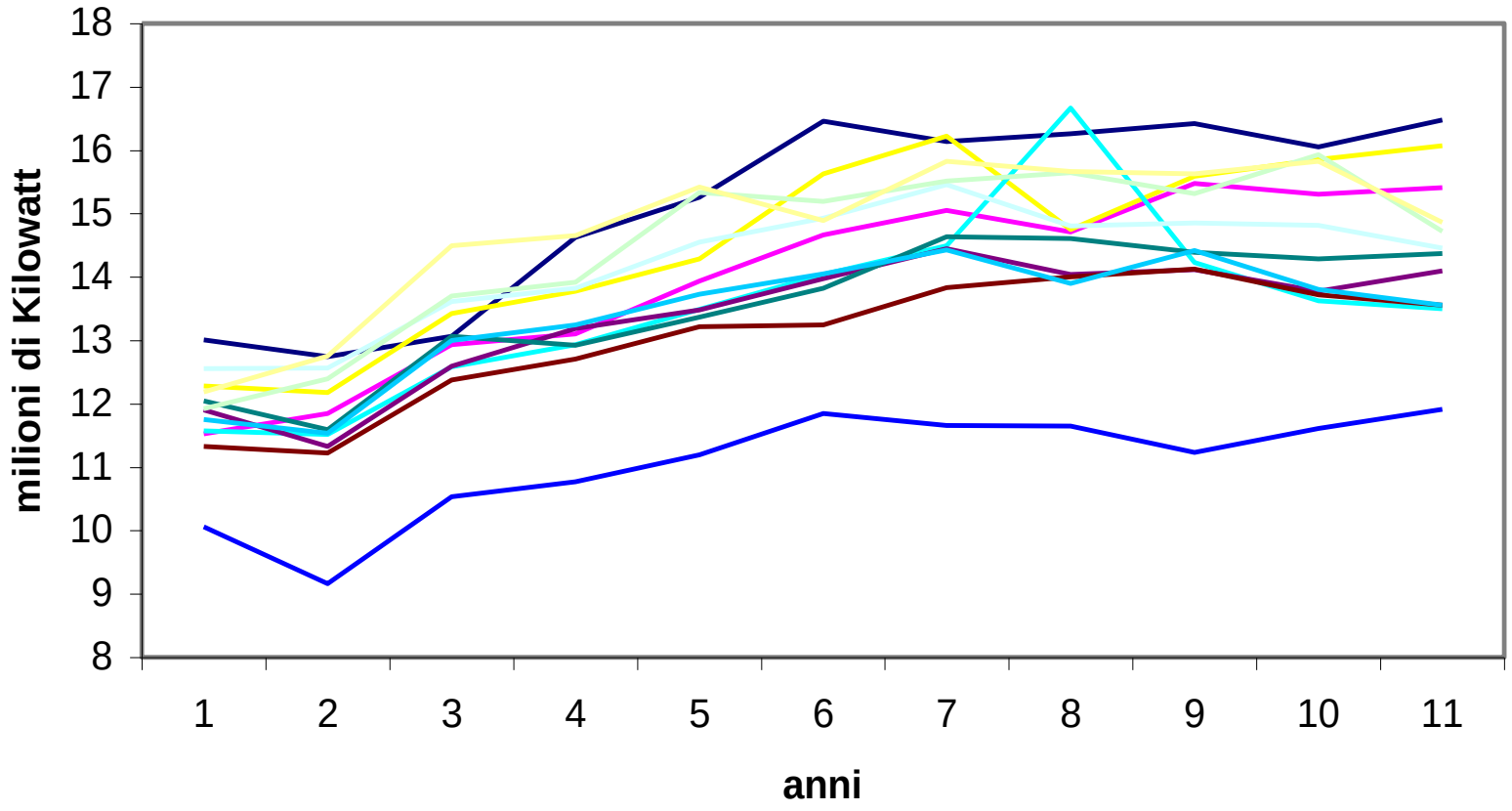
Graficamente tali confronti possono essere fatti riportando su uno stesso diagramma:

a) le *S* sottoserie corrispondenti alle colonne della tabella;

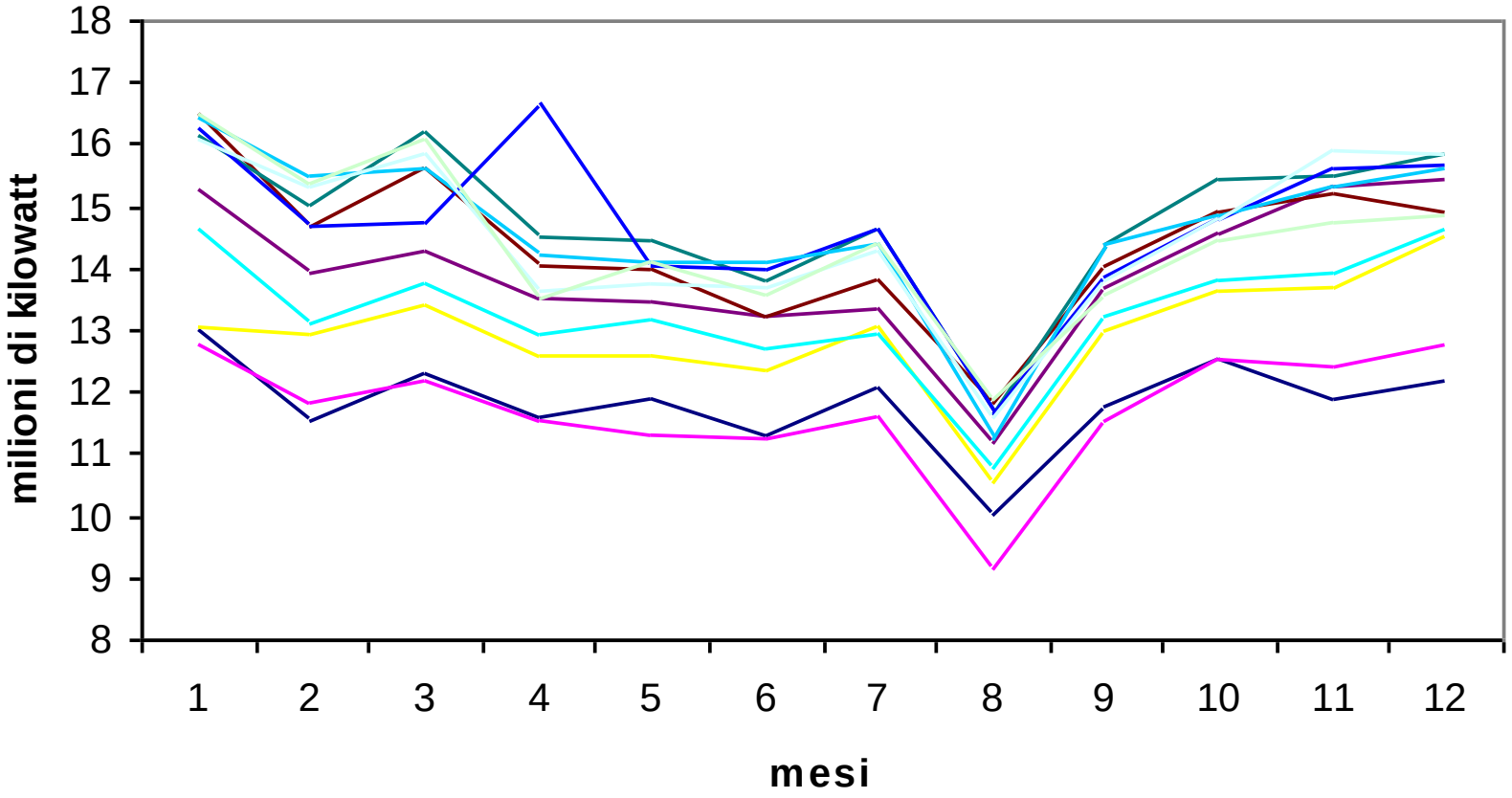
b) le *A* sottoserie corrispondenti alle righe della tabella.

Con riferimento alla serie mensile della Produzione di energia elettrica e gas (PE) si ha:

a) Diagramma cartesiano della serie PE rispetto agli 11 anni di rilevazione



b) Diagramma cartesiano della serie PE rispetto ai 12 mesi di rilevazione

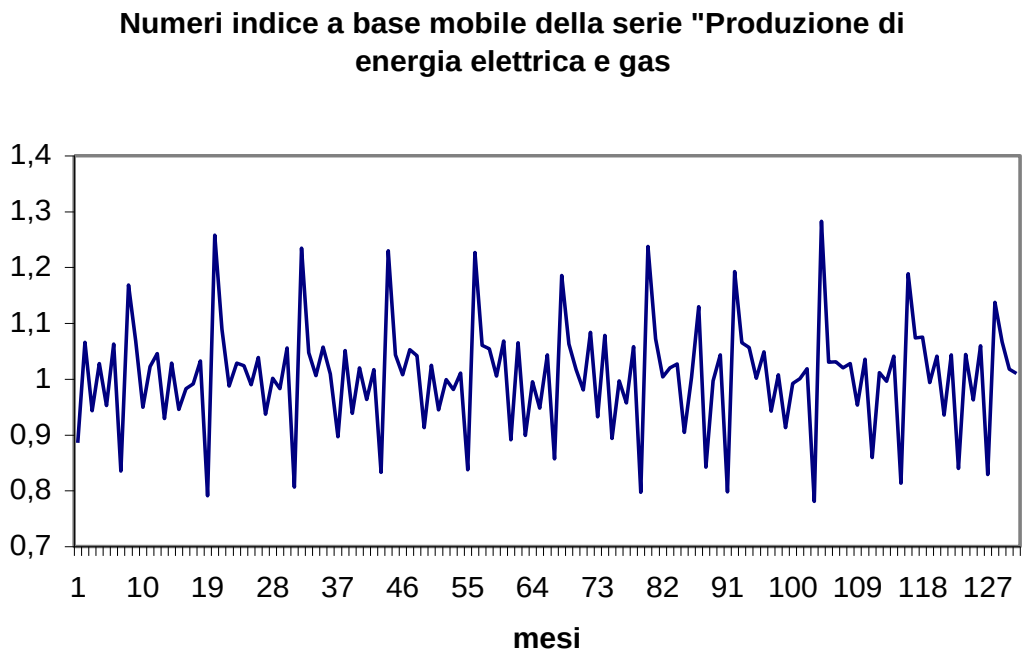


Per una più facile individuazione delle componenti, può essere utile ricorrere ad opportune trasformazioni dei dati.

A questo fine, le principali trasformazioni sono i rapporti, cioè i numeri indici (a base fissa, Y_t / Y_b, e a base mobile, Y_t / Y_{t-1}) e le differenze (prime, $Y_t - Y_{t-1,}$ o di ordine d, $Y_t - Y_{t-d}$).

Le serie dei numeri indici, essendo adimensionali (cioè indipendenti dall'unità di misura) consentono il confronto tra più serie. Inoltre, con la base fissa l'andamento di ciascuna serie rimane invariato.

Con le serie a base mobile, poiché il confronto avviene fra due osservazioni consecutive, si evidenziano variazioni di BREVE PERIODO:



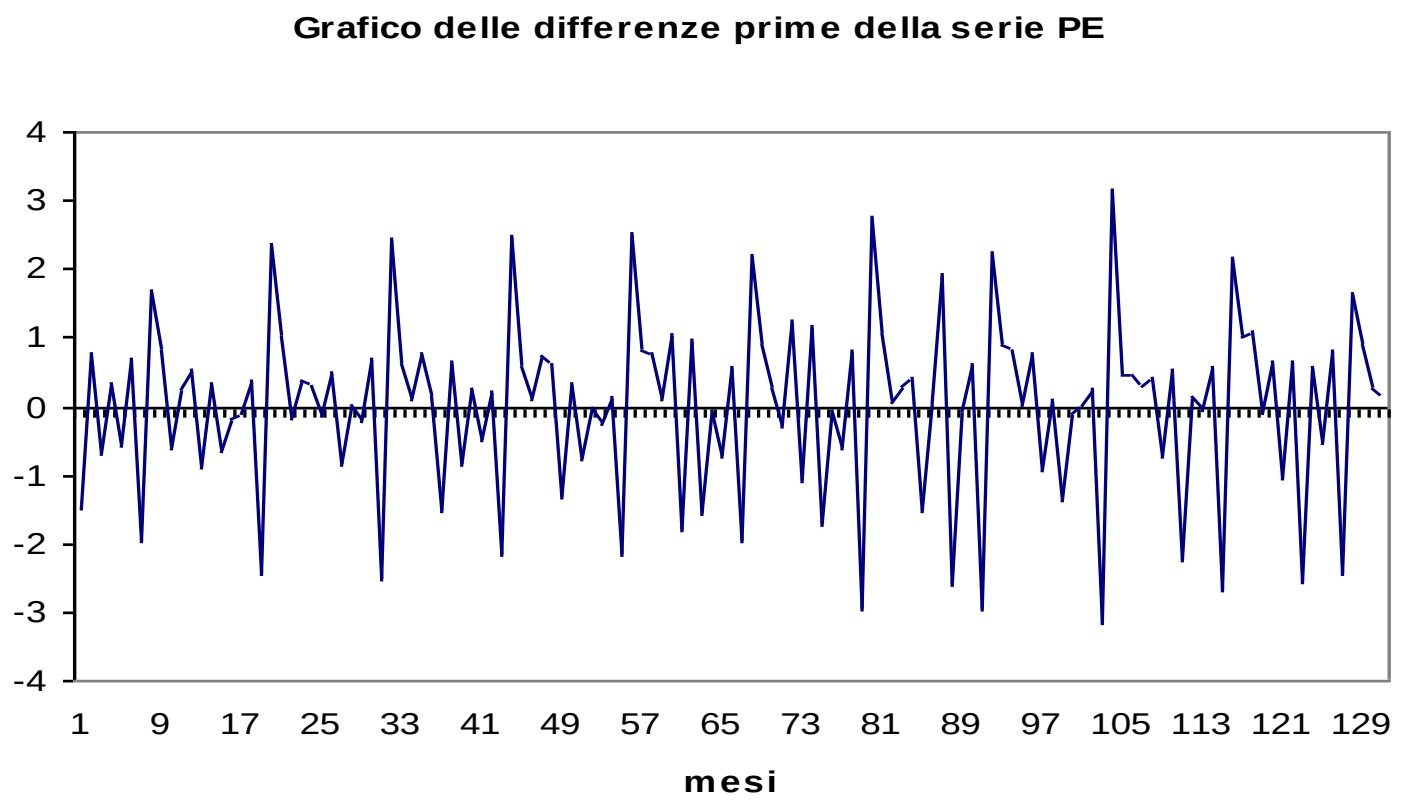
Se tale serie risulta stazionaria, mentre in quella originaria si riscontra un evidente trend, si ha un indizio di *TREND ESPONENZIALE*:

$$T_t = \alpha \times \gamma^t$$

dove:

- se $\gamma > 1 \Longrightarrow$ il trend è CRESCENTE;
- se $\gamma = 1 \Longrightarrow$ il trend è ASSENTE;
- se $\gamma < 1 \Longrightarrow$ il trend è DECRESCENTE.

Anche le differenze prime, al pari degli indici a base mobile, evidenziano variazioni di breve periodo:



Se la corrispondente serie risulta stazionaria, mentre in quella originaria si riscontra un evidente trend, si ha un indizio di *TREND LINEARE*:

$$T_t = \alpha + \gamma \times t$$

dove:

- se $\gamma > 0 \Longrightarrow$ il trend è CRESCENTE;
- se $\gamma = 0 \Longrightarrow$ il trend è ASSENTE;
- se $\gamma < 1 \Longrightarrow$ il trend è DECRESCENTE.

Le differenze prime possono servire anche a individuare la presenza di valori eccezionali in Y_t, spesso "mascherati" dall'eventuale trend.

Infatti, ammesso che in t^* vi sia un valore eccezionale, nel grafico delle differenze prime si osserveranno due picchi, di entità più o meno equivalente, uno in t^* e l'altro in t^*+1.

Inoltre, se Y_{t^*} è eccezionalmente elevato, il primo picco sarà positivo e il secondo negativo; viceversa, in caso di Y_{t^*} eccezionalmente piccolo.

Le differenze prime possono inoltre servire a individuare un cambiamento di livello in Y_t, meno evidente in presenza di un forte trend.

Infatti, ammesso che in t^* vi sia tale cambiamento, a partire da questo istante temporale la serie delle differenze si attesterà su un livello differente (più basso o più elevato, coerentemente con il cambiamento in Y_{t^*}).

In presenza di un trend quadratico nella serie Y_t:

$$T_t = \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2,$$

la serie delle differenze prime non risulterebbe stazionaria, ma presenterebbe un trend lineare.

Infatti, supponendo per semplicità che $Y_t = T_t + u_t$

e indicando con il simbolo ∇ indica l'operatore differenza prima, cioè $\nabla(Y_t) = Y_t - Y_{t-1}$, si ha:

$$\nabla(Y_t) = \nabla(T_t) + \nabla(u_t) =$$

$$= \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2 - [\gamma_0 + \gamma_1 \times (t-1) + \gamma_2 \times (t-1)^2] + \nabla(u_t) = \gamma_1 - \gamma_2 + 2\gamma_2 t + \nabla(u_t),$$

che è appunto una funzione lineare del tempo.

Risulterebbe invece stazionaria la serie delle differenze seconde (differenze prime delle differenze prime), così ottenibili:

$$\nabla^2 Y_t = \nabla(Y_t - Y_{t-1}) = \nabla(Y_t) - \nabla(Y_{t-1}) =$$
$$= Y_t - Y_{t-1} - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} =$$
$$= \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2 - 2[\gamma_0 + \gamma_1 \times (t-1) +$$
$$+ \gamma_2 \times (t-1)^2] + [\gamma_0 + \gamma_1 \times (t-2) + \gamma_2 \times$$
$$(t-2)^2] + (u_t - 2u_{t-1}u_{t-2}) = 2\gamma_2,$$

che è appunto indipendente dal tempo.

E’ possibile infine definire le differenze stagionali:

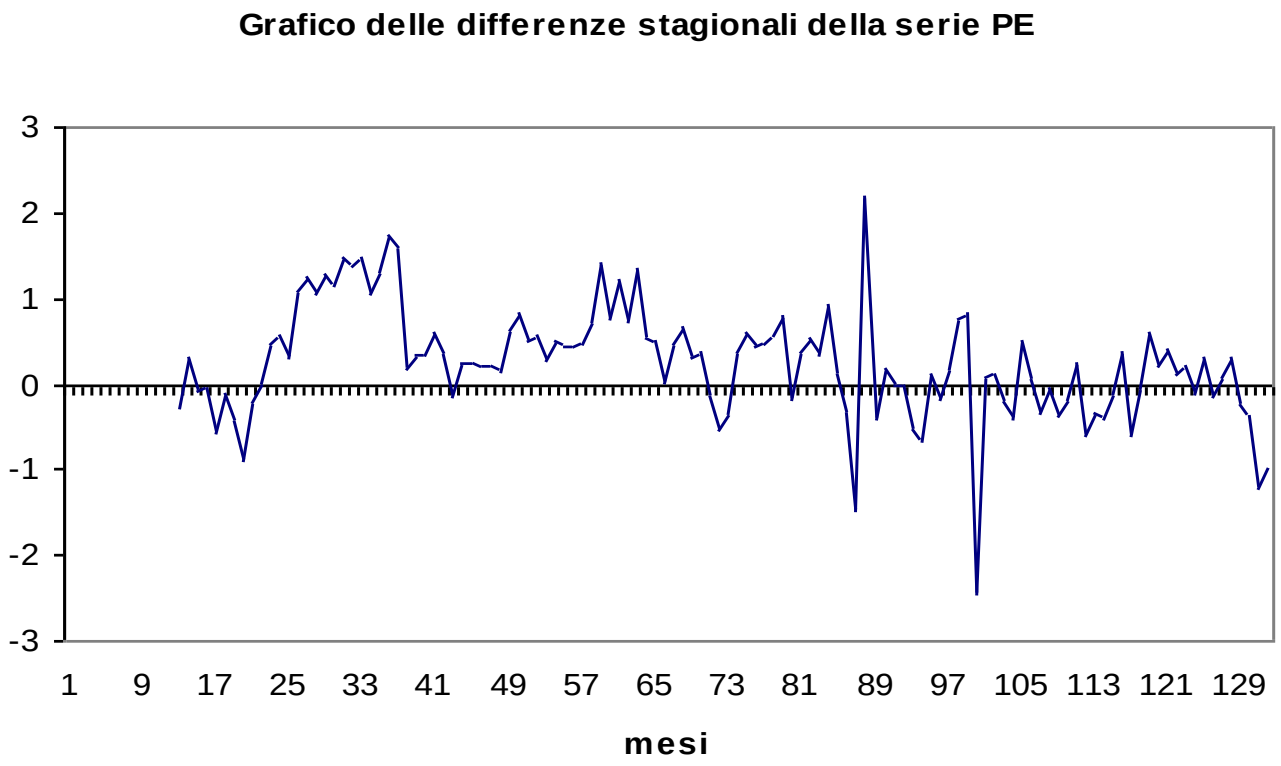
$$\nabla_s Y_t = Y_t - Y_{t-s}, \quad t=s+1, \ \ldots, n,$$

dove s indica il periodo della stagionalità.

Ad esempio, con dati mensili $s=12$, mentre con dati trimestrali $s=4$.

Con tale trasformazione si opera un confronto tra le osservazioni riferite alla stessa stagione di due anni consecutivi: la variazione registrata sarà quindi imputabile principalmente all’evoluzione di fondo del fenomeno.

Le differenze stagionali pertanto eliminano la stagionalità e evidenziano il trend.



Tuttavia, se il trend presente nella serie è di tipo lineare, le differenze stagionali eliminano anche il trend.

Se una serie presenta sia un forte trend, sia una forte stagionalità, per ottenere una serie stazionaria può essere necessario ricorrere alla doppia differenziazione (differenza prima delle differenze stagionali):

$$\nabla\nabla_s Y_t = \nabla(Y_t - Y_{t-s}) = Y_t - Y_{t-1} - Y_{t-s} + Y_{t-s-1},$$

$$t = s+2, \ldots, n.$$

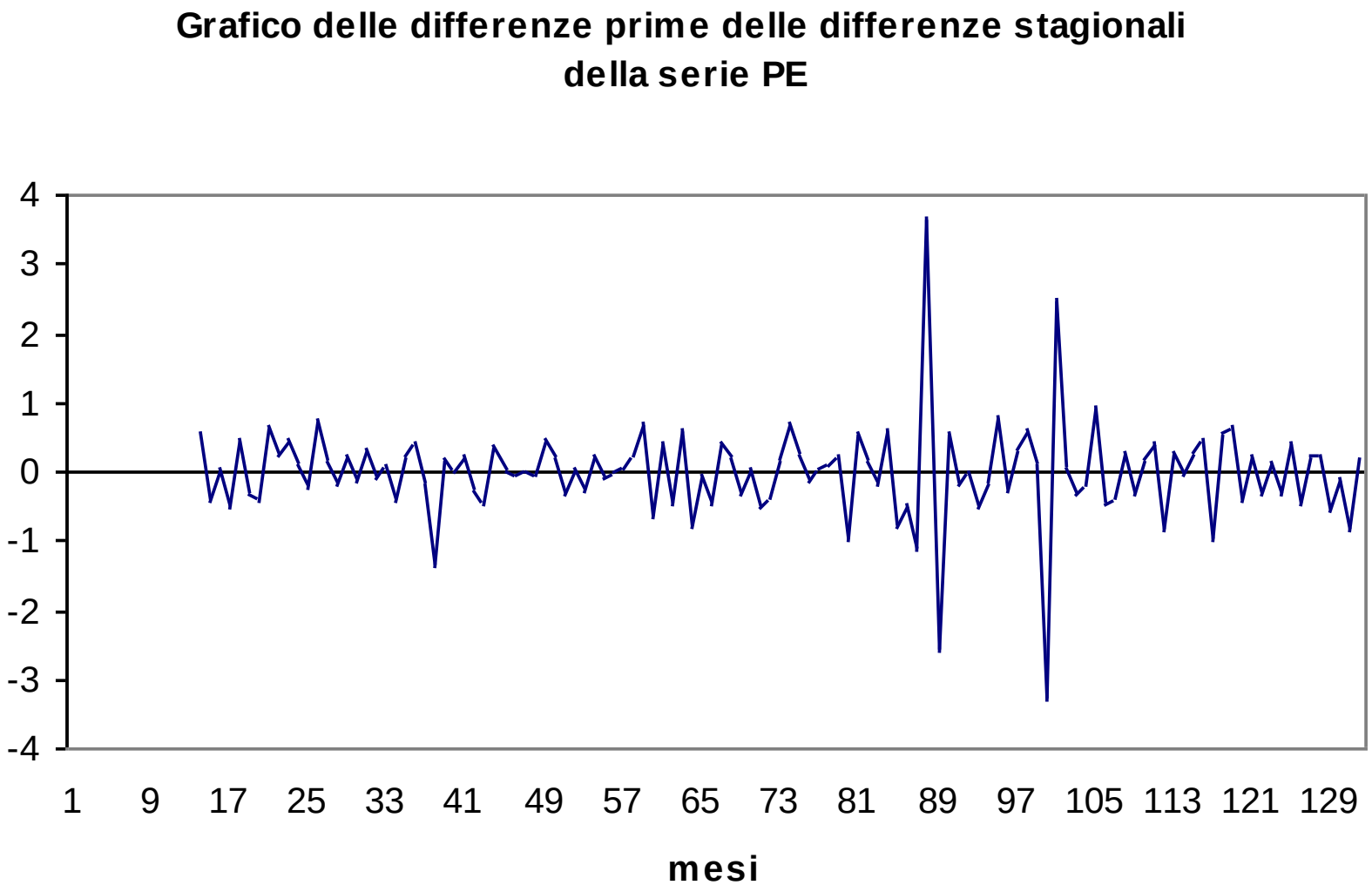


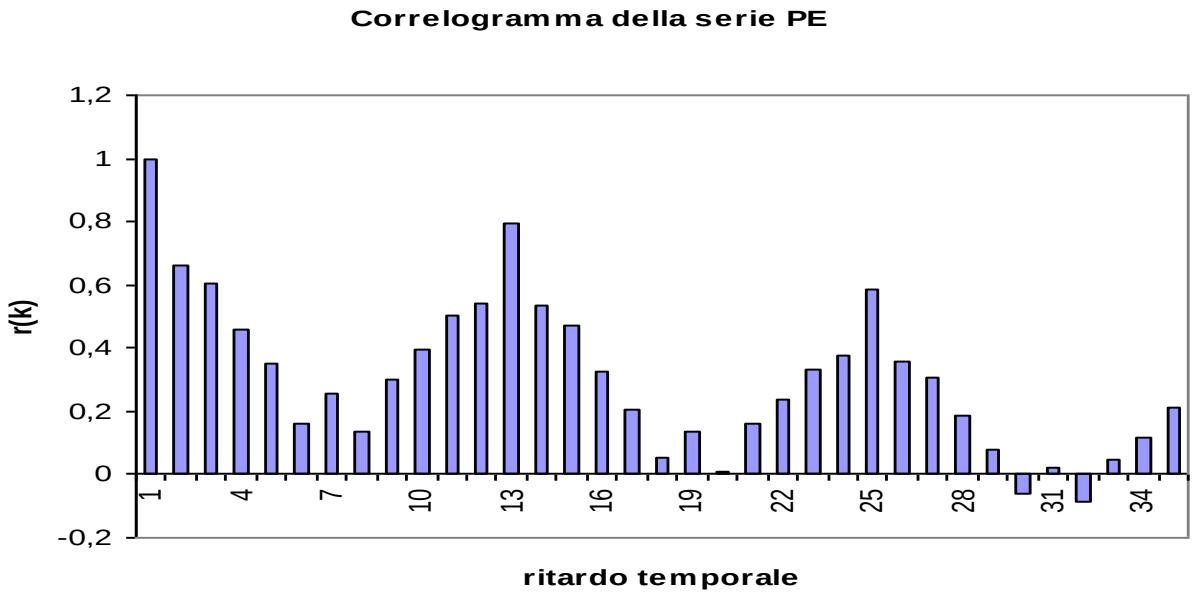
Grafico delle differenze prime delle differenze stagionali della serie PE
4
3
2
1
0
-1
-2
-3
-4
1
9
17
25
33
41
49
57
65
73
81
89
97
105
113
121
129
mesi

Un'ulteriore trasformazione dei dati, che ci aiuta a individuare la presenza di trend e/o stagionalità nella serie originaria Y_t, è la funzione di autocorrelazione, già introdotta per l'analisi dei residui del modello di regressione:

$$r_h = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \in (0,1), \qquad h=1,2,\ldots,H.$$

Tale indice misura l'intensità dei legami lineari tra i valori della serie distanti *h* tempi. Poiché all'aumentare di *h* si riducono gli addendi della sommatoria, per alti valori di *h* r_h diviene sempre meno affidabile ⇒ è buona norma fissare $H \leq n / 4$.

Se si rappresentano graficamente le autocorrelazioni si ottiene il cosiddetto (AUTO)CORRELOGRAMMA, che riporta le coppie di punti (h, r_h), $h = 1, 2, \ldots, H$, in un sistema di assi cartesiani:



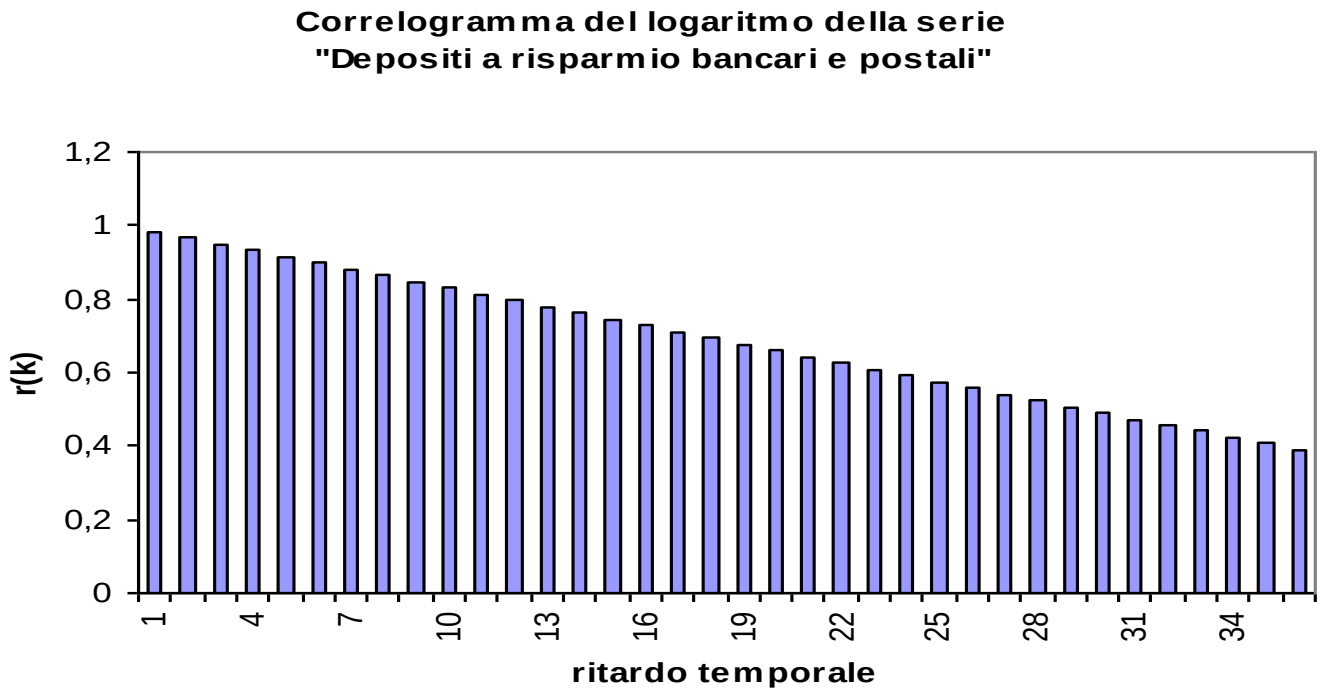
L'ispezione visiva del correlogramma ci fornisce una serie di informazioni sulla struttura della serie:

- Se in Y_t vi è un forte trend che maschera le altre componenti, allora il suo correlogramma decresce linearmente verso lo zero, al crescere di k, e viceversa.
- Se in Y_t vi è una forte componente stagionale di periodo s, che domina sulle altre, allora il correlogramma della serie ha un chiaro andamento sinusoidale di periodo s, e viceversa.

- Una situazione intermedia tra le precedenti due si ha se in Y_t sono presenti contemporaneamente forte trend e marcata stagionalità, come visto per la serie PE.
- Se, infine, Y_t è una serie puramente accidentale (con assenza, cioè, di una qualsiasi struttura al suo interno), allora $|r_h| \cong 0$, per $h = 1, 2, \ldots, H$.

  Più precisamente, $\forall h$, si avrà: $|r_h| \leq 2/\sqrt{n}$.

Esempio di correlogramma per una serie che presenta un forte trend:



Utilizzo delle componenti strutturali di una serie storica nel MLC

Quanto finora visto, fornisce elementi utili a gestire l'eventuale mal specificazione di un modello di regressione, qualora questa sia dovuta all'incapacità dei regressori utilizzati di spiegare tutta la variabilità temporale del fenomeno Y.

Infatti, se i regressori X_{tj} non riescono a spiegare una delle componenti strutturali di Y, questa si scarica sul disturbo ε_t e quindi sui residui e_t.

Componenti strutturali nel MLC 2

In particolare, supponendo di specificare il modello

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \varepsilon_t$$

e che i regressori *X* non siano in grado di cogliere l'andamento di lungo periodo, il plot dei residui stimati mostrerebbe la presenza di un trend.

Potrebbe essere allora utile integrare la funzione f(X) con la componente di trend. Questa andrebbe opportunamente specificata, a seconda del tipo di trend suggerito dall'analisi esplorativa.

Componenti strutturali nel MLC 3

Analogamente, se i regressori X non fossero in grado di cogliere le fluttuazioni di breve periodo eventualmente presenti nella serie Y_t, il plot dei residui stimati mostrerebbe la presenza di una stagionalità e bisognerebbe allora integrare la funzione f(X) con una opportuna componente stagionale.

Spesso, per gestire la stagionalità, è sufficiente inserire S-1 variabili dummy, che tengano conto delle peculiarità specifiche a ogni 'stagione'.