LE SERIE STORICHE

Una serie storica $y(t)$ è semplicemente la registrazione cronologica, non necessariamente con campionamento uniforme, di osservazioni sperimentali di una variabile: l'andamento dei prezzi delle materie prime, gli indici di borsa, lo spread BTP/BUND, il tasso di disoccupazione. Da questa serie di dati si vuole estrarre informazione per la caratterizzazione del fenomeno in osservazione e per la previsione di valori futuri.

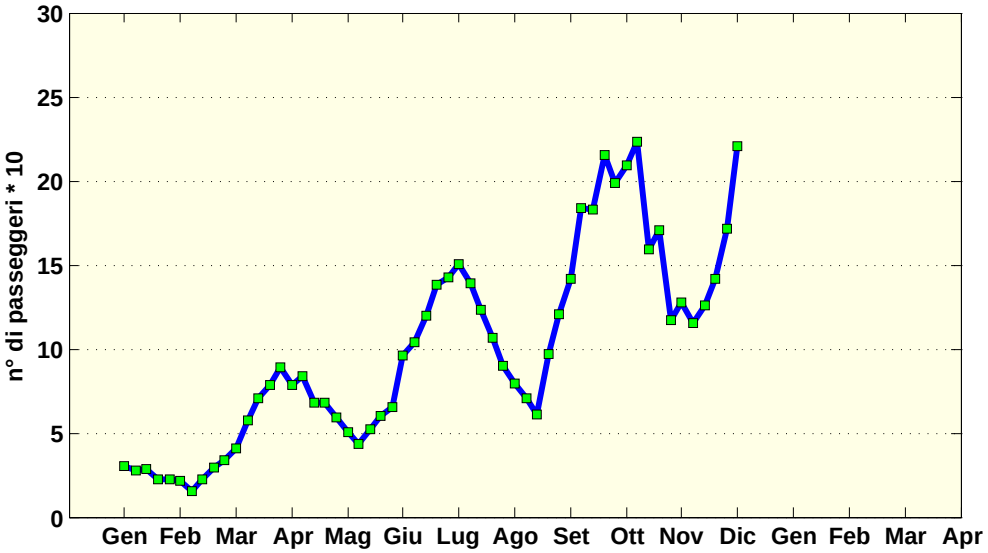

Figura 1: *rilevazione del numero di passeggeri in un piccolo aeroporto*

Dai dati di figura potremmo certo riconoscere che il numero dei passeggeri è in crescita (si nota un trend positivo), denotando tuttavia una certa variabilità (oscillazioni intorno ad una ipotetica linea di tendenza) che si va via via più accentuando al passare del tempo. Volendo tentare una predizione del numero di passeggeri nel prossimo Gennaio, potremmo ragionare nel

modo seguente: con i dati acquisiti potremmo tracciare la linea di tendenza e prolungarla fino al Gennaio successivo. In questo modo avremmo una valutazione di massima del numero di passeggeri, circa 200 (Fig.2), che ci dovremmo aspettare nell'immediato futuro. Tuttavia, avendo osservato una

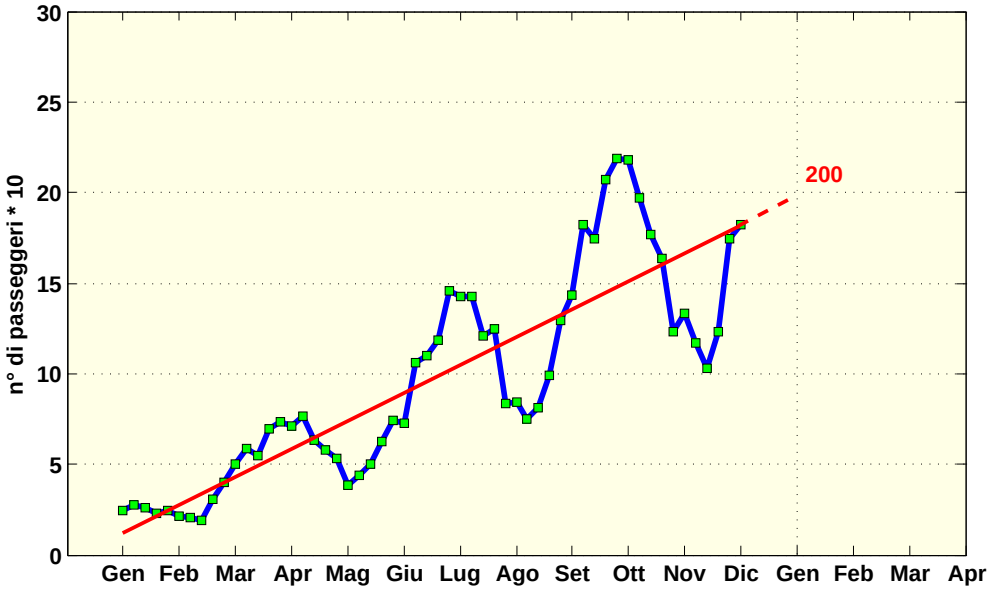

Figura 2: *linea di tendenza del numero di passeggeri*

certa variabilità dei dati, il numero effettivo di passeggeri potrebbe differire molto dalla stima precedente, sia in eccesso che in difetto. Per determinare un possibile intervallo di valori entro cui dovrebbe collocarsi con una certa confidenza il numero di passeggeri, potremmo tracciare le curve spezzate dei massimi e dei mimini (Fig.3), ottenendo in questo modo un intervallo di valori plausibili da 120 a 290, a cavallo della media di 200, dettata dalla tendenza.

Certamente la valutazione effettuata risulta abbastanza insoddisfacente, è molto grande la differenza tra il limite inferiore e quello superiore per prendere una qualche decisione affidabile, come ad esempio quanti impiegati destinare alle operazioni di check-in. Questo dipende sostanzialmente dal modello troppo semplice che si è adottato sia per la tendenza che per la variabilità della serie. In altra parole abbiamo caratterizzato la serie con descrittori con una scala temporale troppo grossolana. Nel caso della tendenza si è considerato il *trend lineare* su tutto l'intervallo d'osservazione, mentre per la variabilità si è considerata la proiezione lineare dei massimi e minimi della serie. Una valutazione migliore comporta l'analisi dei dati ad una scala

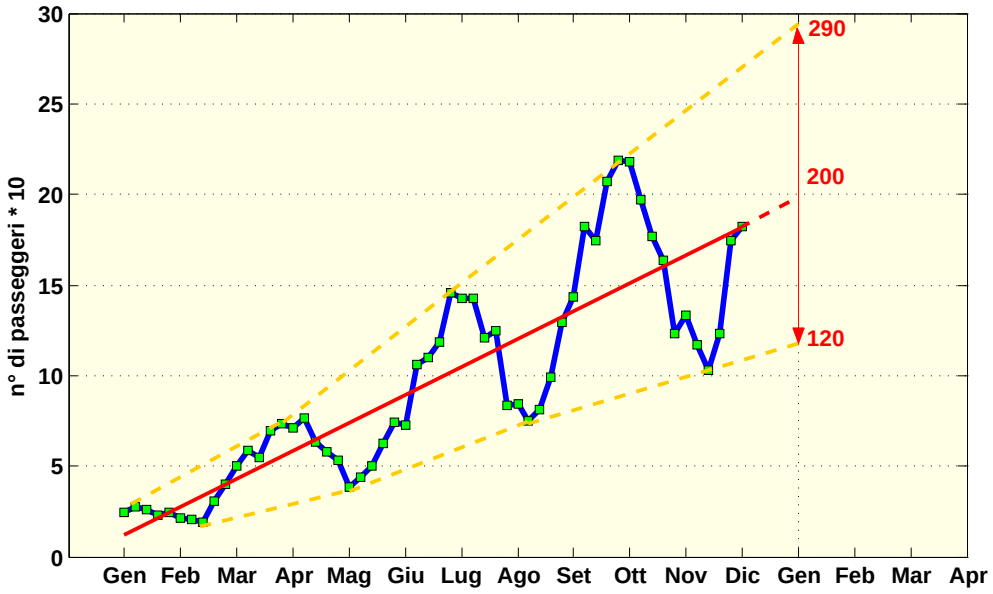

Figura 3: *linee di supporto del numero di passeggeri*

temporale più fine in modo da descrivere in modo più accurato il *movimento della serie*

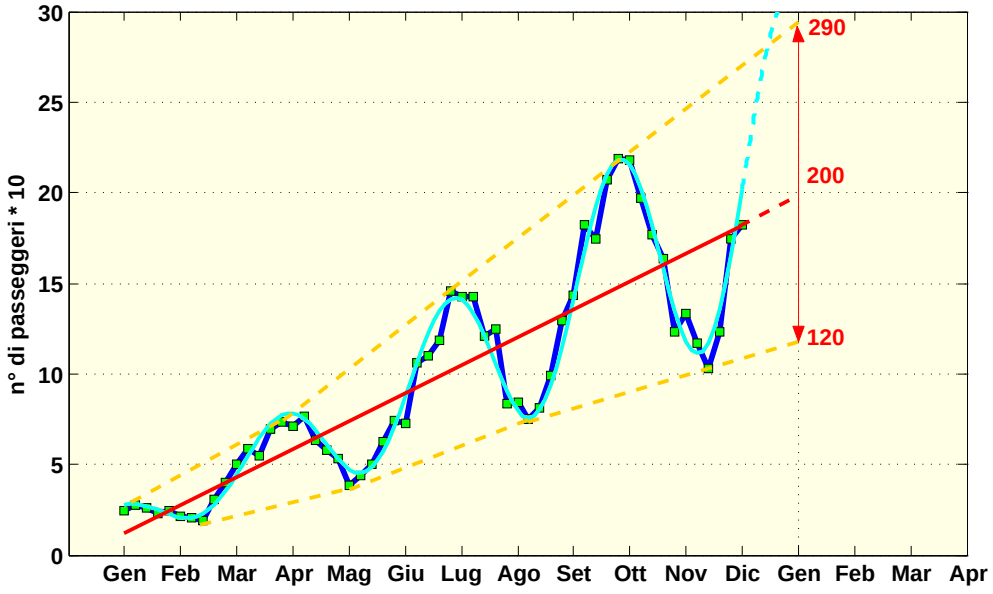

Figura 4: *stagionalità del numero di passeggeri*

La Fig.4 mostra una curva che insegue bene la variazione dei dati in ogni

punto di campionamento, evidenziando peraltro il carattere periodico della serie storica. Tale curva permette certamente una valutazione più realistica del numero di passeggeri nel prossimo Gennaio, propendendo per un valore vicino all'estremo superiore calcolato precedentemente (addirittura superiore a quello). Tale curva può essere pensata come la sovrapposizione del trend $\tau(t)$ e della componente periodica (o stagionalità) $S(t)$

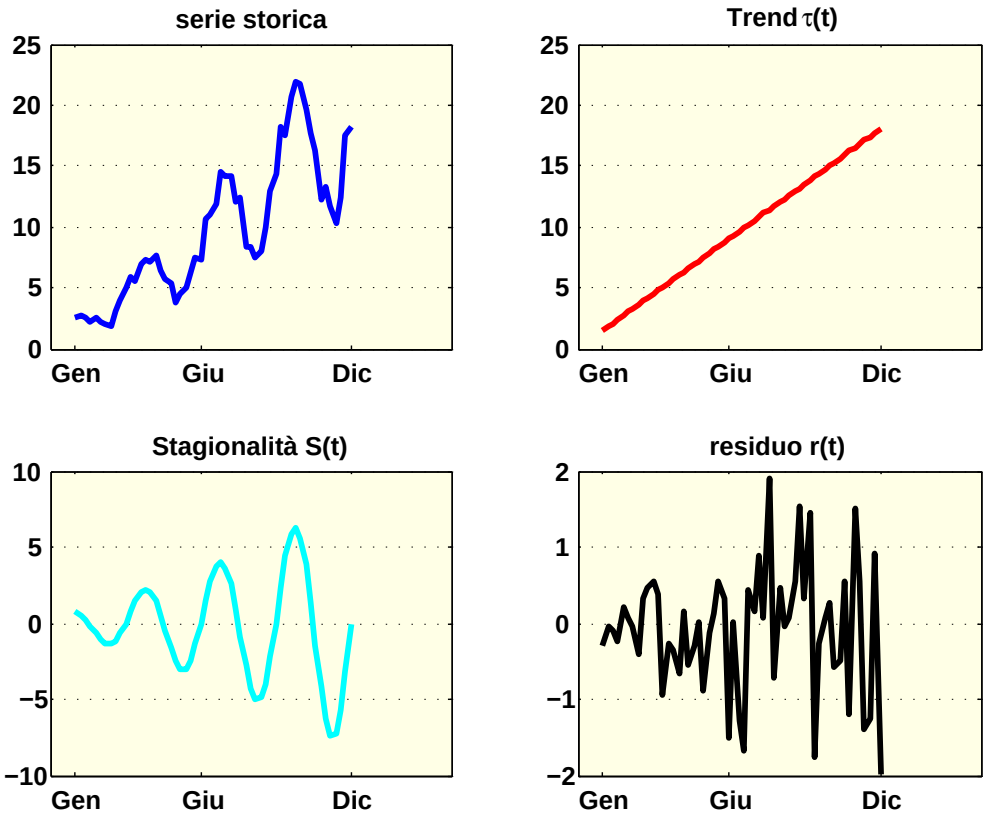

Figura 5: *analisi della serie storica*

Quello che resta nella serie di dati oltre al trend ed alla componente stagionale prende il nome di residuo $r(t)$. Questo, eventualmente, può essere ulteriormente analizzato $r(t) = \gamma(t) + \varepsilon(t)$ in una componente strutturata $\gamma(t)$ ed una sequenza *i.i.d.* (independent identically distributed random variables). Questa sequenza è completamente priva di informazione in quanto, essendo tutti i suoi valori indipendenti l'uno dall'altro, non è possibile prevedere in alcun modo il valore della sequenza in un qualunque punto di

campionamento anche se si conoscono i valori in tutti gli altri punti di campionamento. Tali sequenze sono anche dette *sequenze di rumore bianco*. In questo caso le componenti $\tau(t)$, $S(t)$ e $\gamma(t)$ intercettano tutta l'informazione presente nei dati $y(t)$.

Nel caso appena illustrato si è effettuata un'analisi dei dati di tipo additivo

$$y(t) = \tau(t) + S(t) + r(t) \tag{1}$$

Di solito, tutto ciò che eccede il trend prende il nome di *componente ciclica* $c(t)$ della serie storica

$$c(t) = y(t) - \tau(t) \tag{2}$$

In altri casi risulta più indicata un'analisi di tipo moltiplicativo

$$y(t) = \tau(t) * S(t) * r(t) \tag{3}$$

Non c'è un modo sistematico per scegliere l'una o l'altra modalità, considerando inoltre che alcune serie storiche potrebbero richiedere contemporaneamente i due tipi di analisi; in altre parole le serie di dati sperimentali non sono semplicemente suddivise in serie con analisi additiva e serie con analisi moltiplicativa. Tuttavia i modelli additivo e moltiplicativo sono soddisfacenti nella stragrande maggioranza dei casi.

Nel caso appena esaminato per esempio, possiamo notare che la componente stagionale ha un'oscillazione la cui ampiezza cresce nel tempo, con una tendenza che sembrerebbe dipendere fortemente dal trend (crescita lineare). Anche il residuo denota lo stesso comportamento: un andamento erratico con ampiezza crescente. In questo caso sarebbe bene provare ad analizzare i dati con un modello moltiplicativo.

Vediamo invece i dati del caso di Fig.6, che rappresentano le rilevazioni sperimentali dell'indice NDVI (Normalized Difference Vegetation Index, indice della presenza di vegetazione ottenuto dal telerilevamento della riflettanza spettrale nel visibile, rosso, e nel vicino infrarosso). La componente stagionale (seasonal) ha un'oscillazione di ampiezza costante, ed anche il residuo varia all'interno di un intervallo pressoché costante di valori. Ne deduciamo che l'analisi mostrata è di tipo additivo. Tuttavia notiamo che la componente di trend non è un semplice trend lineare come nell'esempio della Fig.1, ma si è scelto di rappresentare la tendenza della serie storica su una scala fine del tempo, e non la tendenza globale riferita a tutto l'intervallo temporale di misura. In questo modo, descrivendo l'andamento medio della serie su una scala temporale più locale, si ottiene una curva di trend che segue in maniera più fedele la dinamica dei dati. E' da notare inoltre che probabilmente il

trend lineare su tutto l'intervallo di osservazione non avrebbe messo in luce alcuna tendenza significativa, e quindi sarebbe stato privo di informazione.

La scala a cui rilevare il trend dipende molto dal tipo di informazione richiesta dal problema allo studio. Nelle serie storiche di tipo finanziario per trend si intende quasi sempre il trend lineare (scala temporale lunga), salvo poi distinguere i trend secondari, terziari e quaternari, a scale temporali via via più locali.

Vediamo ora il prossimo caso.Nei grafici della Fig.7 vengono presentate due analisi della serie storica riguardanti il numero di soldati americani morti in Vietnam dal 1966 al 1971: la prima è su base annuale, la seconda su base trimestrale (quarterly).

L'analisi su base annuale mostra un andamento del trend molto regolare con un primo tratto crescente ed il tratto finale decrescente. Punto per punto questa curva fornisce l'andamento medio annuale dei dati. La componente stagionale ha un periodo di un anno con un ampiezza modulata dall'andamento del trend, prima crescente e poi decrescente. Per questo motivo il periodo non è proprio costante come per il caso della Fig.6; in questo caso la componente stagionale si dice *pseudo-periodica*. Osservando i dati si può pensare ad un'analisi in cui la componente stagionale si sovrappone al trend ma con un'ampiezza modulata da esso

$$\tau(t) + \tau(t) * S(t) \tag{4}$$

In questi casi è logico ritenere che anche il residuo risenta della modulazione dell'ampiezza in base al trend, ed ottenere la seguente analisi della

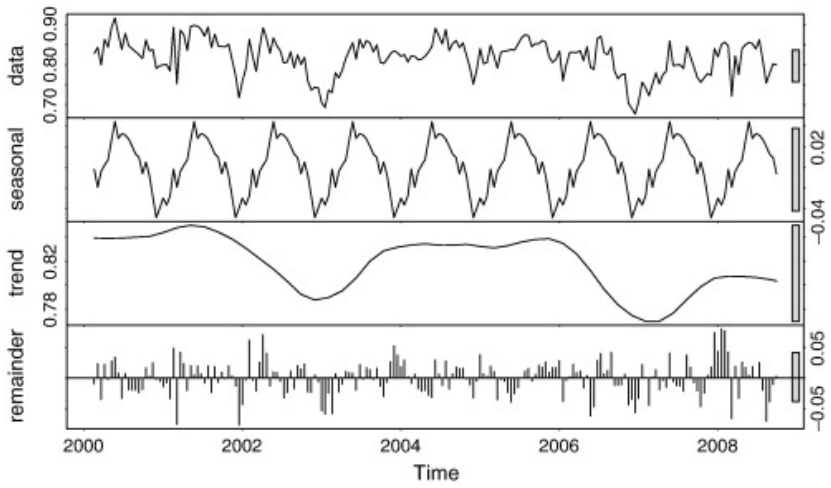

Figura 6: *analisi additiva dei dati del NDVI per una piantaggione di pini nel sud-est dell'Australia*

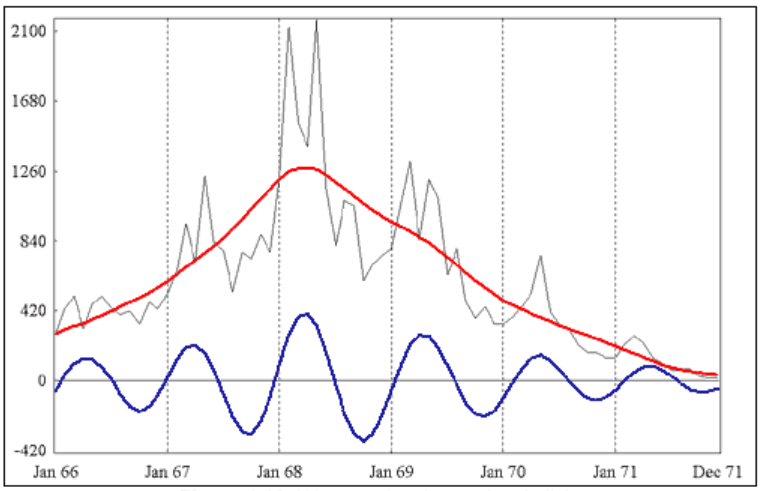

Figure 1.14: *War: trend and annual periodicity*

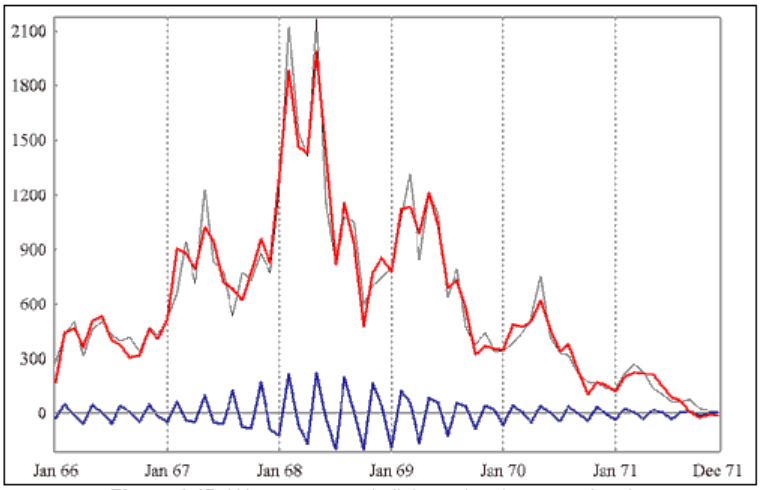

Figure 1.15: *War: quarter periodicity and series approximation*

Figura 7: *analisi a scale differenti del numero dei combattenti USA deceduti nella guerra in Indocina dal 1966 al 1971*

serie storica

$$y(t) = \tau(t) + \tau(t) * S(t)) + \tau(t) * r(t) \tag{5}$$

Un'analisi quindi di tipo misto: addizionale per quanto riguarda la sovrapposizione degli effetti delle tre componenti di trend, stagionalità e residuo, ma la stagionalità ed il residuo hanno un'ampiezza che varia nel tempo

secondo l'andamento del trend. Lo stesso tipo di comportamento dei dati lo avevamo osservato anche per il caso di Fig.1, per il quale è ipotizzabile lo stesso tipo di analisi mista.

Nel secondo grafico di Fig.7, il trend ha un andamento che denota una dinamica (variabilità) comparabile con quella dei dati. Esso infatti rappresenta l'andamento medio dei dati su base trimestrale, e quindi ad una scala temporale più fine del caso precedente, per cui *insegue* in modo più accurato i movimenti della serie di dati.

Il trend trimestrale cattura inoltre gran parte della dinamica che prima apparteneva alla componente stagionale, che infatti risulta essere di ampiezza molto più bassa rispetto a prima, con uno pseudo-periodo di tre mesi. Anche in questo caso l'ampiezza della componente stagionale sembra essere modulata in base all'ampiezza del trend, per cui è ipotizzabile un modello di analisi misto additivo/moltiplicativo, visto precedentemente. Le due analisi rappresentano la stessa serie di dati, ma c'è una differente ripartizione dell'informazione tra componente di trend e componente stagionale. Va sottolineato che la componente stagionale con pseudo-periodo trimestrale era già presente anche nell'analisi su base annua, probabilmente relegata nella componente di residuo, non rappresentata in Fig.7.

Analisi delle serie storiche

L'analisi di una serie storica consiste nel determinare le componenti che ne descrivono i caratteri utili alla formazione delle decisioni (tendenza, stagionalità) ed alla previsione. Tali componenti sono composte principalmente secondo un modello additivo od un modello moltiplicativo. Dal punto di vista algoritmico verrà trattata l'analisi solo per il modello additivo, in quanto questa può essere applicata anche al modello moltiplicativo, dopo un'opportuna trasformazione logaritmica dei dati

$$\begin{aligned} \ln y(t) &= \ln\left(\tau(t) * S(t) * r(t)\right) && (6)\\ &= \ln \tau(t) + \ln S(t) + \ln r(t) && (7)\\ &= \tau(t)' + S(t)' + r(t)' && (8) \end{aligned}$$

Dopodiché, una volta identificate le componenti $\tau(t)'$, $S(t)'$ e $r(t)'$ dall'analisi additiva del $\ln y(t)$, per l'analisi del modello moltiplicativo si ottiene

$$y(t) = \exp(\tau(t)' + S(t)' + r(t)') \tag{9}$$
$$= \exp(\tau(t)') * \exp(S(t)') * \exp(r(t)') \tag{10}$$

Nel caso invece del modello misto (5) incontrato in alcuni degli esempi trattati precedentemente si può procedere nel modo seguente. Si calcola il logaritmo dei dati

$$\ln y(t) = \ln \tau(t) + \ln(1 + S(t) + r(t)) \tag{11}$$

ottenendo in questo modo un'analisi additiva del $\ln y(t)$ come somma di una componente di trend

$$\tau(t)' = \ln \tau(t)$$

ed una parte ciclica

$$c(t)' = \ln(1 + S(t) + r(t))$$

Una volta identificato $\tau(t)'$ si determina $c(t)' = \ln y(t) - \tau(t)'$. Ora, ricordiamo che il termine $c(t)'$ è legato alle componenti di stagionalità e di residuo dell'analisi mista dalla relazione $c(t)' = \ln(1 + S(t) + r(t))$, per cui si ottiene

$$\exp(c(t)') = 1 + S(t) + r(t) \tag{12}$$

per cui basta eseguire un'analisi additiva del segnale $z(t) = \exp(c(t)') - 1$ per ottenere $S(t)$ ed $r(t)$. A questo punto l'analisi mista è data da

$$y(t) = \exp(\tau(t)') * (1 + S(t) + r(t)) \tag{13}$$

Per quanto visto quindi, nelle prossime sezioni verranno illustrati i possibili modelli per le componenti di trend, stagionale e di residuo dell'analisi additiva di una serie di dati, e contestualmente verranno descritti i metodi che permettono di identificarli dai dati sperimentali.

Il trend

Il trend $\tau(t)$ di una serie storica descrive l'andamento medio della stessa riferito ad un'opportuna scala temporale. Nella maggior parte dei casi per trend si intende il trend lineare su tutto l'intervallo di osservazione della serie. In questo caso la componente ciclica ha valor medio nullo su tutto l'intervallo di osservazione. Quindi se stimassimo la media campionaria di $c(t)$ dovremmo ottenere un valore pressoché nullo

$$\hat{\mu}_c = \frac{1}{N}\sum_{i=1}^{N} c(t_i) \simeq 0$$

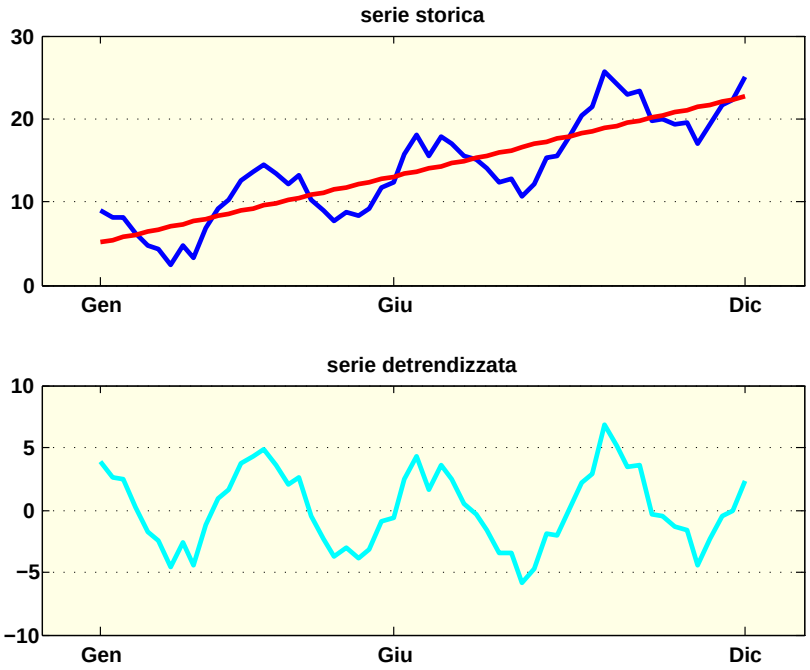

Figura 8: *serie storica, trend lineare, componente ciclica (serie detrendizzata)*

Il modello del trend lineare è un polinomio di primo grado della variabile temporale t

$$\tau(t) = \theta_0 + \theta_1 t \tag{14}$$

dove i parametri sono costanti, in quanto tale modello vale per ogni t appartenente all'intervallo di osservazione dei dati. Tali parametri possono quindi essere stimati risolvendo un problema dei minimi quadrati

$$\left(\hat{\theta}_0, \hat{\theta}_1\right) = \text{argmin}\left[\sum_{i=1}^{N}\left(y(t_i) - \theta_0 - \theta_1 t_i\right)^2\right]$$

ottenendo $\hat{\theta}_0 = 3.4320$, $\hat{\theta}_1 = 1.6056$ per i dati di Fig.8. Ovviamente, nella soluzione di questo semplice programma, si sono adottate tutte le tecniche più volte discusse per eliminare il malcondizionamento del problema. Sottolineiamo che in questo caso non ha senso valutare l'R^2 del modello identificato in

quanto, rappresentando questo solo l'andamento medio dei dati, certamente non spiega granché della varianza dei dati, che invece è praticamente tutta contenuta nella componente ciclica. In particolare, i valori di quest'ultima possono essere stimati sottraendo il trend dai dati

$$\hat{c}(t_i) = y(t_i) - \hat{\theta}_0 - \hat{\theta}_1(t_i), \quad i = 1, \ldots, N$$

e sono graficati in Fig.8.

In casi più generali è possibile che il trend non abbia semplicemente un andamento lineare, ma segua una legge di variazione temporale più complessa. In queste situazioni si può ricorrere ad un modello polinomiale di grado più elevato

$$\tau(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \ldots + \theta_m t^m \tag{15}$$

L'identificazione parametrica di tale modello si ottiene risolvendo un programma dei minimi quadrati del tutto simile al caso lineare appena trattato, per cui non verrà ulteriormente discusso. Invece dobbiamo capire come scegliere il grado m del polinomio. Dato che anche in questo caso il trend deve solo descrivere l'andamento medio dei dati, non è di alcun aiuto valutare l'R^2 del modello identificato, che comunque risulterà basso. Quindi, al solito, la scelta del grado va fatta per tentativi e sfruttando caratteristiche generali individuabili per ispezione visiva della serie storica. Ad esempio, per i dati di Fig.8, si nota che l'andamento medio su tutto l'intervallo di osservazione non denota una variazione della curvatura, per cui si può pensare al più ad un polinomio di secondo grado. In effetti, risolvendo il problema dei minimi quadrati per il trend quadratico

$$\tau(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \theta_2 t^2$$

si ottiene $\hat{\theta}_0 = 3.7924$, $\hat{\theta}_1 = 1.3846$, $\hat{\theta}_2 = 0.0197$. Trend e componente ciclica sono graficati in Figura 9. Si nota in effetti una leggera curvatura positiva ($\hat{\theta}_2 > 0$) ma non così evidente, per cui potremmo senz'altro ritenere accettabile il trend lineare precedente.

Come regola generale per determinare l'ordine del trend si potrebbe stabilire di fare vari tentativi aumentando di volta in volta il grado del polinomio che lo rappresenta fino a che il termine di grado massimo $\hat{\theta}_m t^m$ non dia un contributo considerato trascurabile rispetto alla somme dei termini fino al grado $m-1$. Naturalmente questo dipende molto dall'estensione dell'intervallo di osservazione dei dati. Nel caso appena trattato, se l'intervallo di osservazione fosse molto più esteso (t molto più grande), allora il termine $\hat{\theta}_2 t^2$ certamente darebbe un contributo via via più consistente al crescere di

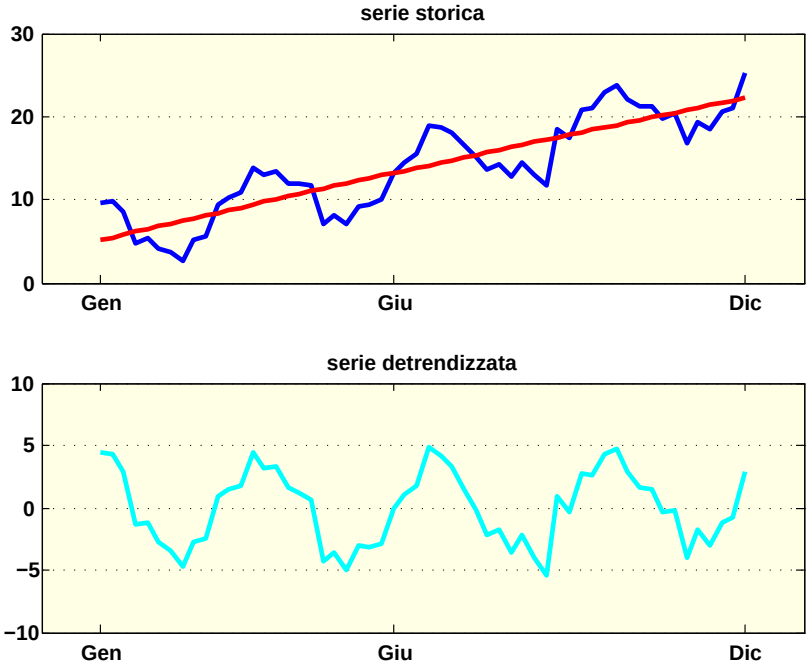

Figura 9: *serie storica, trend lineare, componente ciclica (serie detrendizzata)*

t, tanto da far preferire il modello del secondo ordine rispetto al trend lineare. La complessità del modello polinomiale può essere determinata in modo sistematico ricorrendo al criterio di Akaike.

Il modello polinomiale con coefficienti costanti, permette quindi di rappresentare il trend della serie storica riferito a tutto l'intervallo temporale di osservazione. Quale modello potremmo scegliere se invece della tendenza generale della serie volessimo un andamento medio che seguisse al meglio la dinamica dei dati istante per istante? Questo ad esempio è quello che viene mostrato nel secondo grafico della Fig.7. In questo caso, nel generico istante t, l'andamento medio richiesto deve rappresentare la tendenza media dei dati in un intorno ristretto dell'istante considerato. Come si nota nella Fig.10, nell'intorno degli istanti scelti il valore del trend globale (linea rossa) è molto differente dai valori della serie, in quanto esso è funzione dei valori dei dati su tutto l'intervallo di osservazione. Il trend locale rappresentato dai tratti di linea verde rappresenta abbastanza bene la tendenza locale dei dati.

Prendendo lo spunto dalla Fig.10, si potrebbe pensare quindi di descrivere il trend locale con una sequenza di tratti lineari che, istante per istante, cambino pendenza in modo da adattarsi alla media locale dei dati

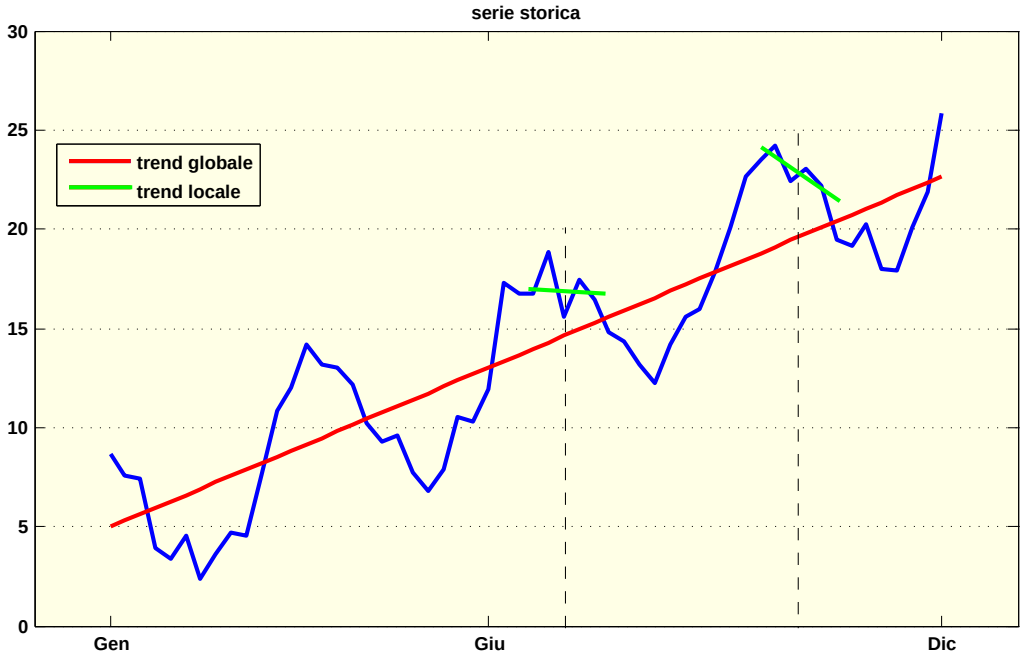

Figura 10: *serie storica, trend globale, trend locale*

$$\tau(t) = \theta_0(t) + \theta_1(t)t, , \quad t \in (t - t_m, t) \tag{16}$$

Il modello precedente è ancora un modello del primo ordine, ma i suoi coefficienti non sono costanti, ma variano ad ogni istante dipendentemente dai dati che si trovano in un intorno $((t - t_m, t)$ più o meno esteso dell'istante corrente t. Il modello del trend locale quindi è definito una volta che si conoscano i suoi parametri $\theta_0(t)$ e $\theta_1(t)$, per ogni t. Vediamo ora come sia possibile stimare dai dati i valori di $\theta_0(t)$ e $\theta_1(t)$ al variare del tempo, usando ancora il metodo dei minimi quadrati. Considerando che i dati sono campionati in istanti discreti del tempo, indichiamo con t_k l'istante corrente, e con t_1 il primo istante di campionamento. Supponiamo di aver collezionato i dati fino all'istante t_k. Se risolvessimo il seguente programma

$$(\theta_0(t_k), \theta_1(t_k)) = \text{argmin} \left[\sum_{i=1}^{k} \Big(y(t_i - \theta_0(t_k) + \theta_1(t_k)t_i \Big)^2 \right]$$

otterremmo un modello lineare (16) su tutto l'intervallo (t_1, t_k), rappresentato dal tratto di retta di color rosso in Fig.11. Si avrebbe quindi un andamento medio lineare, rappresentativo di tutto l'intervallo di osservazione, e non dell'andamento dei dati solo nelle vicinanze dell'istante corrente t_k. Questo

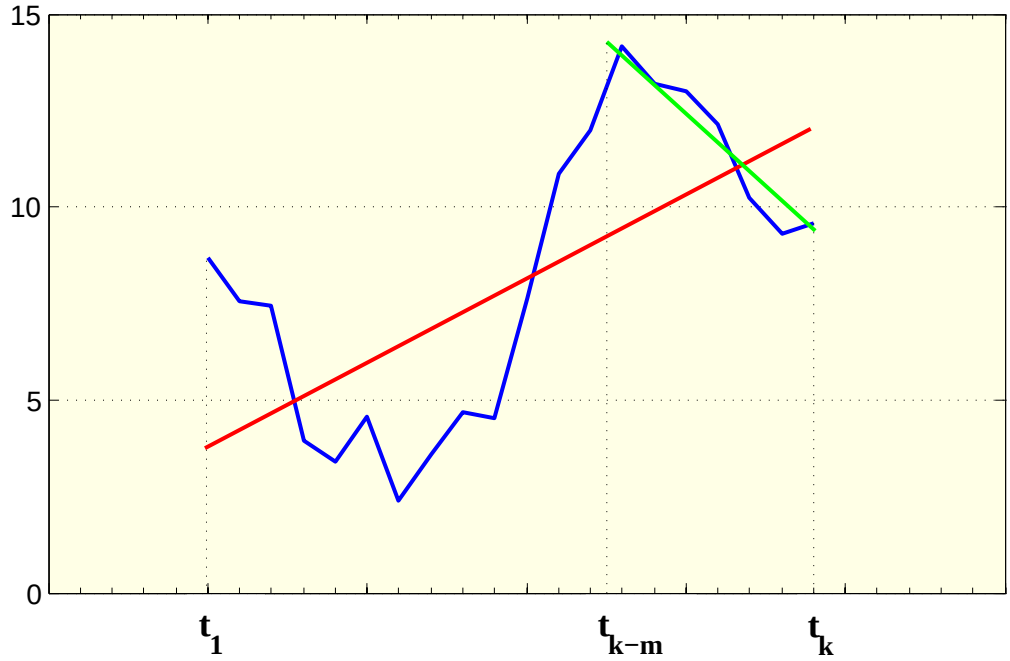

Figura 11: *trend globale, trend locale*

perché nell'indice di costo del programma vengono inclusi tutti i dati da t_1 a t_k. Per ovviare a questo inconveniente, basta introdurre dei pesi nell'indice di costo che diano maggiore importanza ai termini più vicini all'istante corrente e rendano trascurabili i termini distanti da esso

$$(\theta_0(t_k), \theta_1(t_k)) = \text{argmin} \left[\sum_{i=1}^{k} \mu^{k-i} \Big(y(t_i - \theta_0(t_k) + \theta_1(t_k) t_i \Big)^2 \right], \quad \mu \in (0,1)$$

che viene detto problema dei minimi quadrati con pesi esponenziali (Exponential Weighted Least Square Estimate, EWLSE). In questo modo è possibile ottenere il modello lineare rappresentato dalla linea di colore verde di Fig.11. Il parametro μ determina quanto è esteso l'intorno dell'istante corrente t_k in cui risulta valido il modello locale: come si vede i termini della sommatoria sono moltiplicati per μ^{k-i} con μ positivo e minore di 1, per cui per $i << k$ cioè per istanti molto distanti da quello corrente, il peso diventa veramente piccolo e praticamente cancella il termine corrispondente $(y(t_i - \theta_0(t_k) + \theta_1(t_k)t_i)^2$ dalla sommatoria. Per fare un esempio, scegliamo

$\mu = 0.95$; per $i = k, k-1, k-2, ..., 1$ si ha

1 0.95 0.9025 0.8574 0.8145 0.7738 0.7351 0.6983 0.6634 0.6302 0.5987...

per cui a distanza di 7 passi dall'istante corrente t_k il peso diventa circa 0.7 e via via diminuisce indebolendo l'influenza dei temini corrispondenti della sommatoria. Per $\mu = 0.8$ si ha

1 0.8 0.64 0.512 0.4096 0.3277 0.2621 0.2097 0.1678 0.1342 0.1074...

ed in questo caso già a 3 passi dall'istante corrente il peso vale circa 0.5, ottenendo quindi un algoritmo di stima in cui contano solo i dati entro una finestra di 3 passi dall'istante corrente, a differenza del primo caso in cui la finestra era circa di 7 passi.

Facciamo ora un passo in più. Supponiamo di aver correttamente determinato il trend locale all'istante t_k, e viene prelevato un nuovo dato ad un istante successivo t_{k+1}. Per calcolare il trend locale aggiornato al nuovo istante corrente

$$\tau(t) = \theta_0(t_{k+1}) + \theta_1(t_{k+1})t, \quad t \in (t_{k+1-m}), t_{k+1}$$

dovremmo risolvere il seguente programma

$$\Big(\hat{\theta}_0(t_{k+1}), \hat{\theta}_1(t_{k+1})\Big) = \text{argmin} \left[\sum_{i=1}^{k+1} \mu^{k-i}\Big(y(t_i - \theta_0(t_{k+1}) + \theta_1(t_{k+1})t_i\Big)^2\right], \mu \in (0,1)$$

e rielaborare daccapo tutti i dati da t_1 a t_{k+1} per ottenere le stime dei parametri del modello aggiornate all'istante corrente t_{k+1}. In altre parole bisogna di volta in volta rieseguire tutto il calcolo dall'inizio, su un campione di dati di dimensione via via crescente. Questo può essere evitato mediante un algoritmo ricorsivo di soluzione del programma EWLSE, che calcola la soluzione del problema all'istante t_{k+1} in funzione della soluzione al passo precedente t_k. Sia

$$\theta(t_k) = \begin{bmatrix} \theta_0(t_k) \\ \theta_1(t_k) \end{bmatrix}, \quad \ell(t_k) = [1 \quad t_k], \quad L(t_k) = \begin{bmatrix} \ell(t_1) \\ \vdots \\ \ell(t_k) \end{bmatrix}$$

Si ottiene

$$G(t_{k+1}) = \frac{S(t_k)\ell(t_{k+1})^T}{\mu + \ell(t_{k+1})S(t_k)\ell(t_{k+1})^T} \tag{17}$$

$$\hat{\theta}(t_{k+1}) = \hat{\theta}(t_k) + G(t_{k+1})\Big(y(t_{k+1}) - \ell(t_{k+1})\hat{\theta}(t_k)\Big) \tag{18}$$

$$S(t_{k+1}) = \frac{1}{\mu}\Big(I - G(t_{k+1})\ell(t_{k+1})\Big)S(t_k) \tag{19}$$

L'algoritmo va opportunamente inizializzato. Per questo, a partire da un

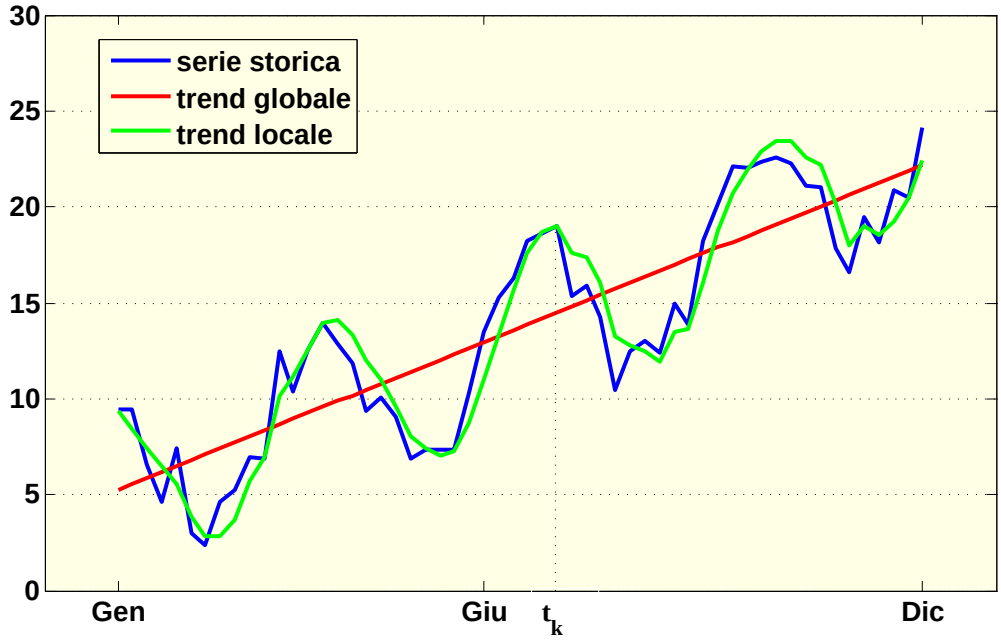

Figura 12: *trend globale, trend locale del primo ordine* $\tau(t) = \theta_0(t) + \theta_1(t)t$

istante t_m, si risolve il problema utilizzando tutti i dati da t_1 a t_m, ottenendo

$$\hat{\theta}(t_m) = S(t_m)L(t_m)^T W(t_m)Y(t_m), \quad S(t_m) = \big(L(t_m)^T W(t_m)L(t_m)\big)^{-1}$$

$$Y(t_m) = \begin{bmatrix} y(t_1) \\ \vdots \\ y(t_m) \end{bmatrix}, \quad W(t_m) = \begin{bmatrix} \mu^{m-1} & 0 & \dots & \dots & 0 \\ 0 & \mu^{m-2} & 0 & \dots & 0 \\ 0 & \dots & \ddots & \dots & 0 \\ 0 & \dots & \dots & \mu & 0 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}$$

A questo punto con $t_k = t_m$, $\hat{\theta}(t_k) = \hat{\theta}(t_m)$ ed $S(t_k) = S(t_m)$ è possibile innescare l'algoritmo ricorsivo e calcolare in successione l'aggiornamento dei

parametri del trend locale secondo le (17), (18) e (19), elaborando solo un dato alla volta. La Fig.12 mostra l'andamento dei dati, del trend lineare (globale) e del trend locale del primo ordine secondo l'algoritmo (17)-(19), con $\mu = 0.8$. Possiamo subito notare un effetto dell'elaborazione ricorsiva dei dati: il trend locale segue con un certo ritardo i dati. Ciò si verifica in quanto il modello locale, ad es. nell'istante t_k in Fig.12, risente solo dei dati precedenti a t_k e non di quelli futuri. La pendenza del trend locale cambierà gradualmente man mano che l'istante corrente si inoltra nel tratto dei dati successivo a t_k (di qui il ritardo), e rappresenterà bene la tendenza locale solo quando la finestra di dati sui cui il modello locale viene stimato sarà tutta compresa nell'intervallo di tempo a destra di t_k. Questo effetto è tanto più evidente quanto più la memoria dell'algoritmo è grande. In questo caso la memoria dell'algoritmo dipende dal valore del parametro μ. Per

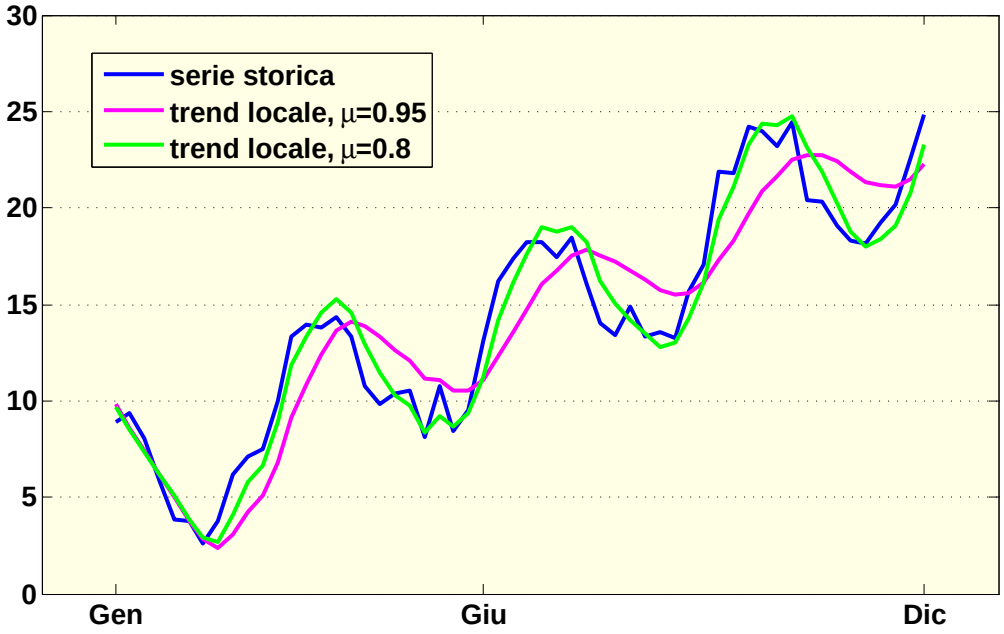

Figura 13: *trend locale con differenti valori di* μ

valori di μ più grandi, l'algoritmo ha una memoria più grande, per cui la stima dei parametri del modello locale dipende da un numero più esteso di dati, anche un pò lontani dall'istante corrente. Come si nota dalla Fig.13 (linea magenta) questo determina una curva di trend locale molto regolare (smooth), ma con un certo ritardo rispetto ai dati. Per valori minori di μ invece, la memoria dell'algoritmo si accorcia, interessando quindi solo i dati pi prossimi all'istante corrente. Ne risulta una curva (linea verde) meno

regolare, che però segue i dati meglio che nel caso precedente, con un ritardo inferiore.

Sin qui si sono forniti dei modelli analitici per il trend, sia globale che locale. In altre parole si sono fornite in forma analitica possibili leggi temporali che rappresentassero al meglio la tendenza dei dati. Questi modelli hanno la loro importanza in quanto i loro parametri quantificano alcune caratteristiche fondamentali dei dati: ad esempio il parametro θ_1 misura la pendenza della serie, e permette di dire se siamo in un peridodo di trend al rialzo o al ribasso. Il parametro θ_2 misura la convessità della serie, per cui permette di stabilire se il trend continuerà con il segno attuale, ad esempio permane il trend al rialzo ($\theta_1 > 0$ e $\theta_2 > 0$) ovvero ci si avvia verso un trend al ribasso ($\theta_1 > 0$ e $\theta_2 < 0$).

Tuttavia, laddove non sia necessaria una descrizione analitica della tendenza della serie storica, è possibile ricorrere a metodi che *calcolino* direttamente i valori di $\tau(t)$ su tutto l'intervallo di osservazione dei dati. Uno tra i più utilizzati è il filtro di Prescott-Hodrick. Secondo questo metodo, vengono calcolati simultaneamente i valori $\tau(t_i)$ su tutto l'intervallo di osservazione dei dati, risolvendo il seguente programma

$$\{\tau(t_i)\} = \text{argmin}\left[\sum_{i=1}^{N}\Big(y(t_i) - \tau(t_i)\Big)^2 + \lambda\sum_{i=2}^{N-1}\Big(\tau(t_{i+1}) - 2 * \tau(t_i) + \tau(t_{i-1})\Big)^2\right] \tag{20}$$

Il primo termine dell'indice di costo misura il fit con cui la sequenza $\tau(t_i)$ rappresenta bene la sequenza dei dati $y(t_i)$. Il secondo termine è invece un termine di penalizzazione che misura la derivata seconda della sequenza di trend (ogni addendo della seconda sommatoria è il quadrato dell'approssimazione numerica della derivata seconda di $\tau(t)$ nel generico istante t_i). Più λ è grande e più verranno selezionate sequenze $\tau(t_i)$ con derivata seconda con ampiezza piccola (quindi molto regolari); al limite per λ molto grande la soluzione potrebbe assomigliare al trend lineare globale visto precedentemente (derivata seconda nulla). Per valori di λ più piccoli invece il programma (20) rende ammissibili sequenze che siano meno regolari, e che quindi possano seguire al meglio, anche localmente, la dinamica dei dati.

La Fig.14, oltre che mostrare il comportamento annunciato della stima del trend al variare del parametro λ, mostra anche l'assenza del fenomeno di ritardo che avevamo osservato nei modelli stimati con il metodo ricorsivo. Questo però non deve trarre in inganno: l'elaborazione del filtro non è in tempo reale, perché la stima degli N valori $\tau(t_i)$ è ottenuta elaborando tutti i dati contemporaneamente, per cui essa non può essere prodotta se non dopo aver acquisito tutti i dati, di fatto con un ritardo massimo pari al tempo necessario ad acquisire tutti i dati. C'è inoltre da osservare che, se

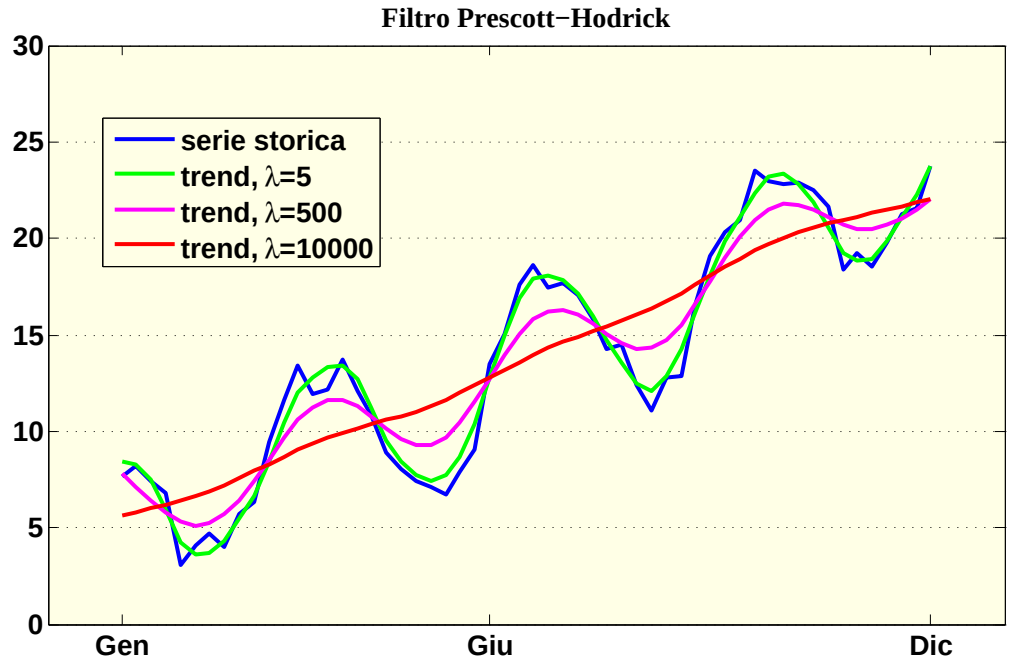

Figura 14: *Stima del trend con il filtro di Prescott-Hodrick con differenti valori di* λ

venisse acquisito un dato ulteriore $y(t_{N+1})$, la stima del trend di Prescott-Hodrick andrebbe ricalcolata risolvendo di nuovo il programma (20) con $i = 1, \ldots, N, N+1$.

Si può ovviare a questo inconveniente con un algoritmo ricorsivo che stima istante pet istante il trend locale $\tau(t_i)$ della serie di dati secondo la seguente formula

$$\tau(t_i) = \frac{1}{m}\left[(y(t_i) + y(t_{i-1}) + +y(t_{i-2}) + \ldots + +y(t_{i-m+1})\right] \qquad (21)$$

Per ogni istante t_i, la stima $\tau(t_i)$ è la media aritmetica degli ultimi m valori della serie, per tale motivo viene detta *media mobile* (moving average). La sequenza stimata $\tau(t_i)$ è tanto più regolare quanto più la memoria m è grande, come si nota facilmente nella Fig.15. Come per tutti gli algoritmi ricorsivi ovviamente è presente un ritardo con cui il trend locale segue la dinamica dei dati, che si accentua, all'aumentare di m.

La media mobile è un algoritmo molto semplice ed efficiente, anche se la tima del trend che si ottiene si può dimostrare essere meno accurata di quella ottenibile con il modello analitico (16). È la stima più utilizzata in ambito finanziario, in quanto permette di ottenere in modo semplice la tendenza dei dati. In ambito finanziario le serie storiche possono essere soggette a shock

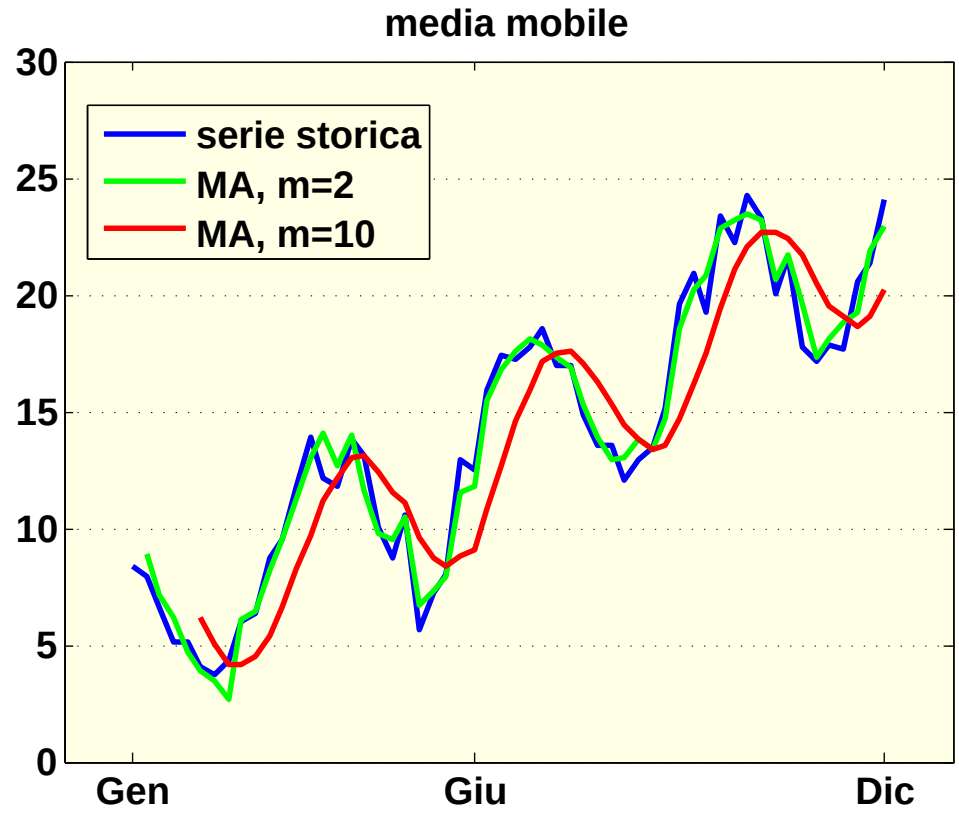

Figura 15: *Stima del trend con la media mobile per diversi valori di m*

(variazioni) con dinamiche molto veloci, si pensi ad esempio al fixing delle valute, per cui l'algoritmo di stima del trend deve raggiungere un compromesso tra il catturare la tendenza dei dati con una certa fedeltà e senza un ritardo eccessivo (il fixing monetario avviene su una scala di secondi) ed il filtrare gli shock dovuti alla volatilità del mercato. Agendo solo sulla memoria m dell'algoritmo il più delle volte non si raggiunge un buon compromesso tra velocià di risposta e filtraggio degli shock. Per questo motivo la media mobile semplice (21)(SMA, simple moving average) viene modificata in modo da dare più peso al dato corrente

$$\tau(t_i) = \alpha y(t_i) + (1-\alpha)\tau(t_{i-1}), \quad \alpha = \frac{2}{m+1} \tag{22}$$

Tale algoritmo si chiama *media mobile esponenziale* (EMA, exponential moving average); la scelta di α indicata nella (22) garantisce generalmente il miglior compromesso. Come tutti gli algoritmi ricorsivi, la EMA va inizializzata, calcolandone un primo campione $\tau(t_m)$ come media mobile semplice

dei primi m dati, poi per $i = m+1, m+2, \ldots$ si usa l'algoritmo (22). Nella

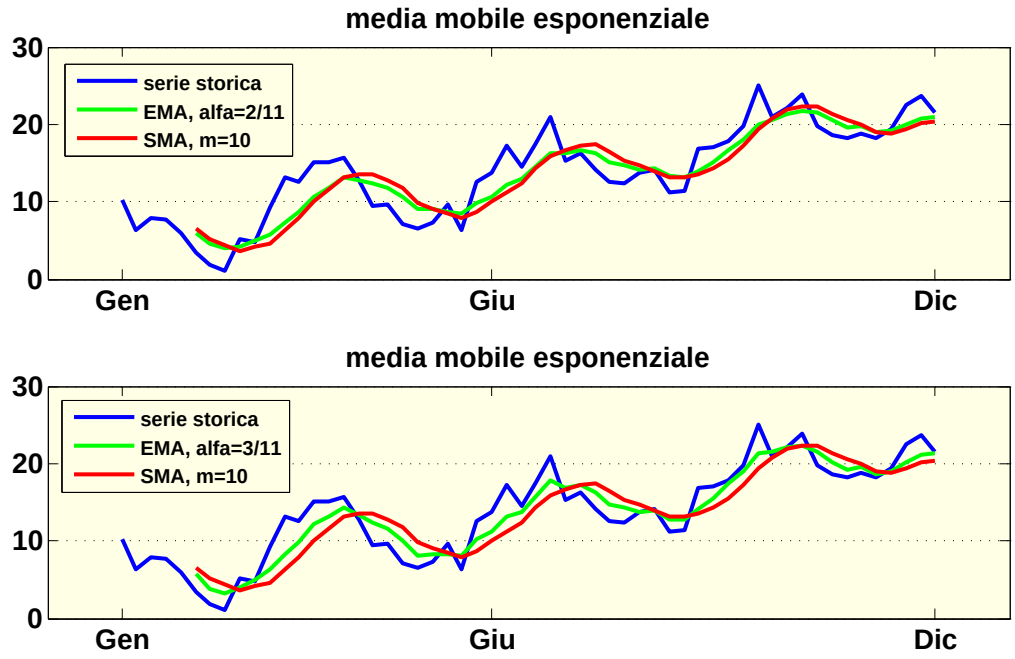

Figura 16: *Stima del trend con la media mobile esponenziale*

Fig.16 si nota come la EMA anticipi la SMA, cioè ha una maggiore velocità di risposta in quanto si accorge prima dei tratti della serie storica sia con tendenza al rialzo che che al ribasso. Inoltre al crescere di α rispetto al valore generalmente consigliato si nota come l'algoritmo migliori in termini di prontezza di risposta.

La componente stagionale

Consideriamo ancora la serie storica di Fig.9. Il grafico in basso mostra l'andamento della serie di dati una volta che da essa venga sottratto il trend $\tau(t)$. Si ottiene quindi la componente ciclica che denota chiaramente un comportamento periodico. È quindi logico ritenere che sia possibile analizzare questa componente ciclica separando la componente stagionale dal residuo

$$c(t) = S(t) + r(t) \tag{23}$$

La componente stagionale è la parte periodica di $c(t)$, il suo grafico è quindi una curva che assume gli stessi valori ad intervalli regolari di tempo (si veda ad esempio anche la Fig.6)

$$S(t) = S(t+T) = S(t+2T) = ... = S(t+kT) = ... \tag{24}$$

Il parametro T prende il nome di periodo. Per capire come stimare $S(t)$ dai dati della componente ciclica $c(t)$, consideriamo il modello per eccellenza di una funzione periodica, un segnale sinusoidale con un periodo ad esempio pari a $5\,sec.$ (linea rossa Fig.17) Nella stessa figura sono riportate alcune

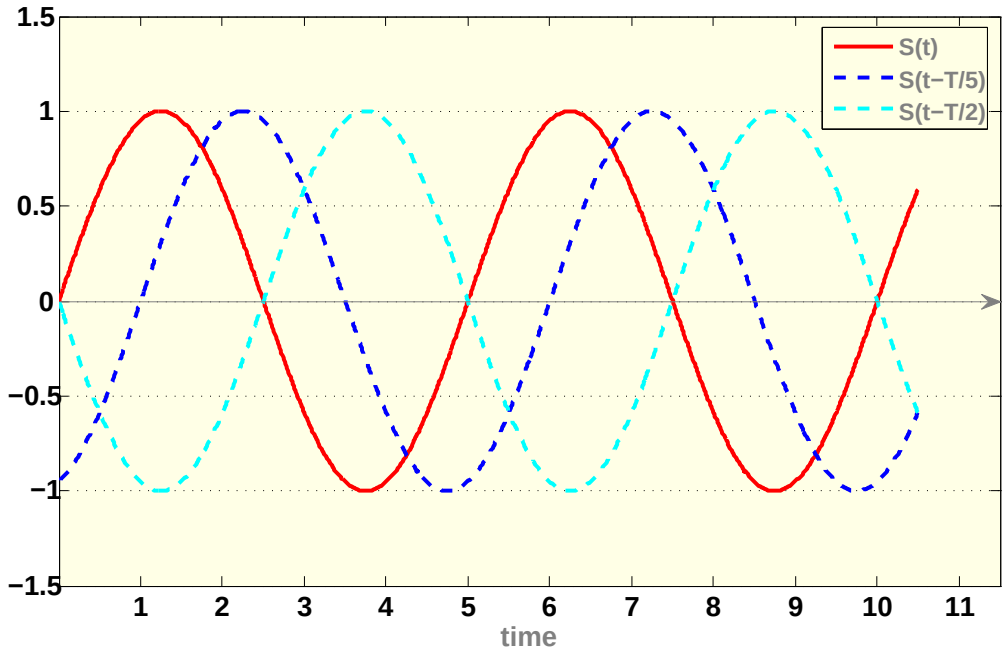

Figura 17: *andamento sinusoidale, T=5 sec*

repliche di $S(t)$: la prima (linea blu) ha un ritardo δ pari a $1\,sec$ rispetto a $S(t)$, mentre la seconda ritardata di mezzo periodo $\delta = 2.5\,sec.$ Confrontando $S(t)$ con $S(t-\delta)$ si nota come, istante per istante, i valori assunti dai due segnali siano abbastanza differenti, e questa differenza si accentua o meno dipendentemente dal valore del ritardo δ. In particolare per $\delta = 2.5 = T/2$ le due curve hanno stessi valori in modulo ma di segno contrario, mentre se ponessimo $\delta = 5 = T$, cioè per un ritardo pari ad un periodo, otterremmo una curva esattamente uguale a $S(t)$. Da quanto detto, il metodo per estrarre la componente stagionale dalla componente ciclica consiste proprio nel misurare la somiglianza di $c(t)$ con le sue versioni ritardate $c(t-\delta)$, al variare di δ,

ed individuare i valori del ritardo per cui tale somiglianza è massima. Tale somiglianza viene misurata dalla *funzione di autocorrelazione*

$$\phi(k) = \frac{\sum_{i=1}^{N-k} c(t_i) * c(t_{i+k})}{\sqrt{\sum_{i=1}^{N-k} c(t_i)^2}\sqrt{\sum_{i=1}^{N-k} c(t_{i+k})^2}} \tag{25}$$

L'espressione (25) si riferisce ovviamente alla versione campionaria della funzione di autocorrelazione, in cui il ritardo δ è dato dal *numero* di campioni che separano l'istante corrente t_i e quello ritardato t_{i+k}. La (25) fornisce il risultato corretto nell'ipotesi che i dati a disposizione siano molti, in altre parole deve risultare $N >> k$. In questo modo risulterà sempre che $c(t_i)$, $i \in [1, N-k]$ ha valor medio nullo. Tuttavia in quasi tutti i casi pratici in cui N non sia molto grande, è bene ricalcolare il valor medio $\mu_{c,0}(k) = \sum_{i=1}^{N-k} c(t_i)/(N-k)$ della componente ciclica sul sottoinsieme dei dati $[1, N-k]$ ed il valor medio $\mu_{c,1}(k) = \sum_{i=1}^{N-k} c(t_{i+k})/(N-k)$ della componente ciclica sul sottoinsieme dei dati $[k+1, N]$, e modificare la (25) nel seguente modo

$$\phi(k) = \frac{\sum_{i=1}^{N-k} \bar{c}(t_i) * \bar{c}(t_{i+k})}{\sqrt{\sum_{i=1}^{N-k} \bar{c}(t_i)^2}\sqrt{\sum_{i=1}^{N-k} \bar{c}(t_{i+k})^2}}$$

dove $\bar{c}(t_i) = c(t_i) - \mu_{c,0}(k)$ e $\bar{c}(t_{i+k}) = c(t_{i+k}) - \mu_{c,1}(k)$

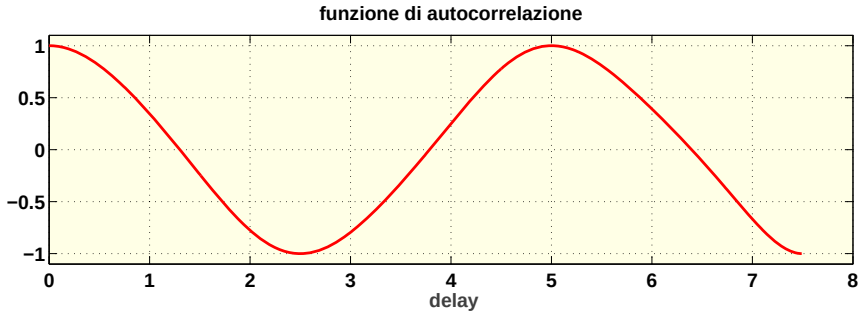

Figura 18: *funzione di autocorrelazione di* $S(t) = sin(2 * \pi * t/T)$

A titolo di esempio, calcoliamo la funzione di autocorrelazione per la funzione $S(t)$ di Fig.17 e grafichiamo il risultato. Come si nota dalla Fig.18 i massimi locali della funzione di autocorrelazione si hanno in corrispondenza al ritardo nullo, ovviamente, e al ritardo $\delta = 5\,sec$ (un periodo), come ci si aspettava. Si ha inoltre un minimo locale $\delta = 2.5\,sec$ (mezzo periodo), che denota che la curva ritardata di tale entità ha lo stesso andamento di $S(t)$

ma con segno opposto, come avevamo già osservato nel comportamento della sinusoide.

Naturalmente le cose sarebbero meno evidenti nel caso di presenza di più di una componente sinusoidale, per cui consideriamo la seguente componente ciclica $c(t) = sin(2\pi t/5) + 1.5sin(2\pi t/2.5)$ in assenza di residuo.

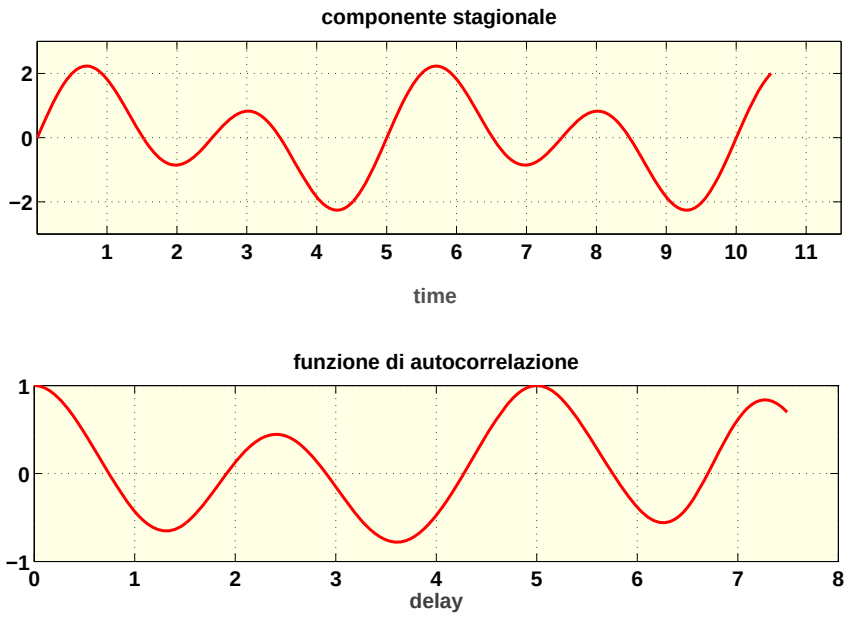

Figura 19: *componente ciclica* $c(t) = sin(2\pi t/5)+1.5sin(2\pi t/2.5)$ *e sua funzione di autocorrelazione*

Osserviamo in Fig.19 che la funzione di autocorrelazione ha un massimo locale (positivo) ben visibile per un ritardo pari a 5 *sec* che segnala la presenza di una componente stagionale con periodo pari proprio a 5 *sec*. Sottolineiamo che il massimo locale a $T = 5$ non segnala la presenza della sola componente sinusoidale con periodo $T = 5$ sec, ma la presenza di una componente stagionale (data dalla combinazione delle due sinusoidi con $T1 = 2.5\,sec$ e $T2 = 5\,sec$) con periodo $T = 5$ sec.

Questo è dovuto ad un risultato generale in base al quale se sommiamo più segnali periodici, i cui periodi hanno a due a due un rapporto dato da un numero razionale (questo si ottiene ad esempio se tutti i periodi sono dei numeri interi, oppure se sono l'uno un multiplo intero dell'altro, come nell'esempio precedente delle due sinusoidi), allora il segnale risultante è un segnale periodico con periodo pari al minimo comune multiplo dei periodi

delle componenti. Quindi, la composizione di un segnale periodico con periodo $3\,sec$ con uno con periodo $5\,sec$ darebbe luogo ad un segnale periodico complessivo con periodo pari a $15\,sec$. Nella funzione di autocorrelazione quindi noteremmo un massimo locale molto evidente per un ritardo pari a $15\,sec$.

In generale quindi, i massimi della funzione di autocorrelazione denotano la presenza di segnali periodici, non necessariamente sinusoidali, cioè di pattern ripetitivi che soddisfano la (24). In questo caso, la stima della componente stagionale deve sfruttare la relazione (24). Quindi, un segnale periodico di periodo T è individuato dai valori $S(t_i)$, $i = 1, \dots M$ che ne compongono un periodo, e che quindi si ripetono tali e quali in ogni intervallo di tempo pari ad un periodo T. Il numero di questi valori dipende dal passo di campionamento del segnale: se ad es. avessi un passo di campionamento pari a $1\,sec$ ed un periodo $T = 10\,sec$, avremmo un numero $M = 10$ di valori differenti del segnale all'interno di un periodo T. Tali valori incogniti possono essere stimati risolvendo il seguente programma

$$\left\{\hat{S}_1, \dots, \hat{S}_M\right\} = \text{argmin}\left[\sum_{j=1}^{N}\sum_{i=1}^{M}\Big(c(t_{(j-1)*T+i}) - S_i\Big)^2\right] \tag{26}$$

dove N è il numero di periodi T compresi nella serie storica analizzata. La funzione obiettivo del programma (26) è una semplice funzione quadratica delle incognite, strettamente convessa, per cui la soluzione ottima se esiste è unica, e si ricava dall'annullamento del gradiente della funzione obiettivo. Si ottiene facilmente che

$$\hat{S}_i = \frac{1}{N}\sum_{j=1}^{N} c(t_{(j-1)*T+i}), \quad i = 1, \dots, M \tag{27}$$

Con il metodo dei minimi quadrati si ottiene che ogni valore $\hat{S}_i$ è semplicemente la media dei campioni di $c(t)$ all'i-esimo istante all'interno di ciascuno degli N periodi di ampiezza T che la compongono.

Vediamo subito un esempio reale, non simulato. Prendiamo la serie di dati mostrata in Fig.20. Si hanno ha disposizione rilievi mensili su 6 anni (72 campioni), la funzione di autocorrelazione è stata calcolata per un ritardo da 0 a 48 mesi. Il massimo relativo più grande della funzione di autocorrelazione si ha per un ritardo pari 36 mesi. In base a quanto discusso precedentemente questo segnalerebbe la presenza di un segnale periodico $S(t)$ con periodo $T = 36$ mesi. Tuttavia una componente stagionale con periodo così lungo sarebbe stimata in modo poco affidabile con soli 72 dati: si avrebbero a

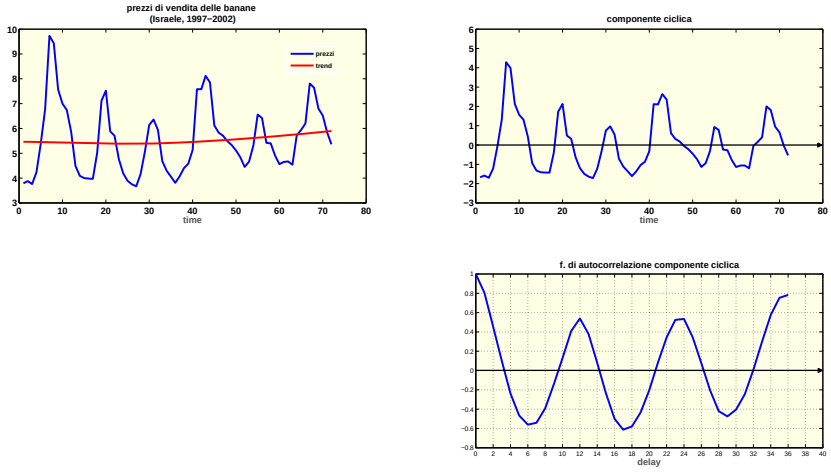

Figura 20: *serie storica dei prezzi delle banane in Israele (1997-2002): serie dei prezzi e trend (Filtro di Prescott-Hodrick,* $\lambda = 10^5$ *), componente ciclica, funzione di autocorrelazione della componente ciclica*

disposizione solo $N = 2$ periodi, e conseguentemente le stime ottenibili dalla (27) sarebbero le medie solo su due campioni di $c(t)$, distanti 36 mesi uno dall'altro. C'è inoltre da considerare che, per quanto riguarda i prodotti agricoli, la periodicità è tipicamente data dalle stagioni, per cui ci si deve attendere periodicità annuali (il ripetersi dello stesso clima) o trimestrali (il succedersi delle stagioni). Per cui nel caso in esame scegliamo certamente il massimo relativo corrispondente al ritardo pari a 12 mesi. Dato che le rilevazioni dei prezzi sono mensili, in un periodo si hanno esattamente 12 valori del prezzo S_1, S_2, ..., S_{12} che, in accordo alla (24), si ripetono tali e quali ogni 12 mesi. Per poter stimare questi valori risolviamo il programma (26)

$$\left\{\hat{S}_1, \ldots, \hat{S}_{12}\right\} = \text{argmin}\left[\sum_{j=1}^{6}\sum_{i=1}^{12}\Big(c(t_{(j-1)*12+i}) - S_i\Big)^2\right] \tag{28}$$

con $M = 6$ periodi di $T = 12$ mesi compresi nella serie storica analizzata (la serie si compone di 72 campioni). Dalla (27) si ottiene facilmente che

$$\hat{S}_i = \frac{1}{6}\sum_{j=1}^{6} c(t_{(j-1)*12+i}), \quad i = 1, \ldots, 12 \tag{29}$$

La Fig.21 mostra l'andamento di questa componente stagionale $S(t)$, della componente ciclica depurata di $S(t)$ (componente residua), e della funzione di autocorrelazione risultante. La componente stagionale $S(t)$ ha un anda-

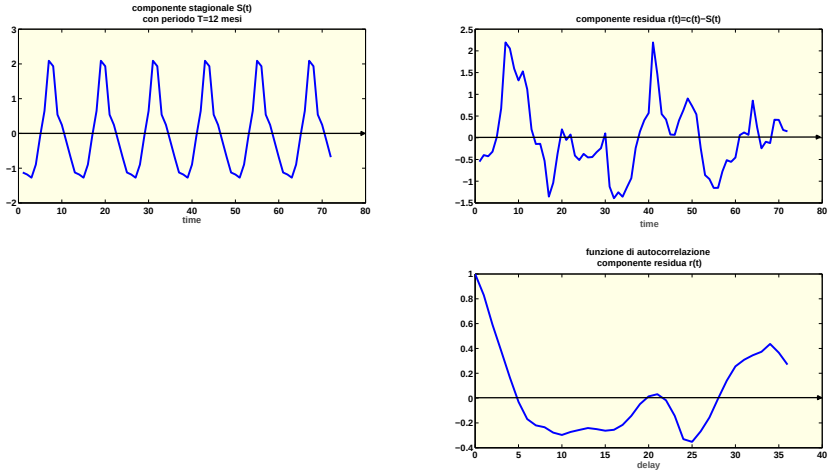

Figura 21: *componente stagionale $S(t)$, componente residua $r(t) = c(t) - S(t)$ e sua funzione di autocorrelazione*

mento periodico con periodo $T_1 = 12$ mesi, ma non è sinusoidale, come ci attendevamo. Nella funzione di autocorrelazione della componente residua $r(t)$ non si evidenziano più massimi locali isolati, significativi, ed infatti $r(t)$ ha un andamento piuttosto erratico, che peraltro è suscettibile di analisi ulteriore, come vedremo nella prossima sezione. La Fig.22 mostra tutte le componenti dell'analisi effettuata del prezzo di vendita delle banane.

Va precisato che le scelte effettuate nell'analisi svolta sono solo indicative, indicano cioè una possibile procedura di analisi. Per esempio, il fatto che la componente di trend sia stata ottenuta con un filtro di Prescott-Hodrick è stata una delle possibili scelte; tra l'altro il valore del parametro λ è stato anch'esso selezionato in modo euristico, provando alcuni valori e scegliendo quello per cui si ottenevano risultati accettabili. Alternativamente, avremmo potuto seguire una via un pò più sistematica, modellando $\tau(t)$ con un polinomio di ordine crescente e scegliendo l'ordine migliore in base al criterio di Akaike. Avremmo ottenuto i risultati riportati in Fig.23. Si nota un notevole abbassamento della figura di merito del criterio di Akaike per $m = 6$, per poi risalire. Questo indica che il miglior compromesso tra fitting e complessità del modello si ha per un polinomio di sesto grado. Procedendo poi alla stima della componente stagionale, esattamente come nel caso precedente, otterremmo gli andamenti di Fig.24. In questa si nota che la componente residua, pur mantenendo un andamento erratico come nel caso precedente, ha tuttavia una funzione di autocorrelazione con dei massimi locali significativi per un ritardo $T = 20$ anomalo. Questo indica semplicemente che

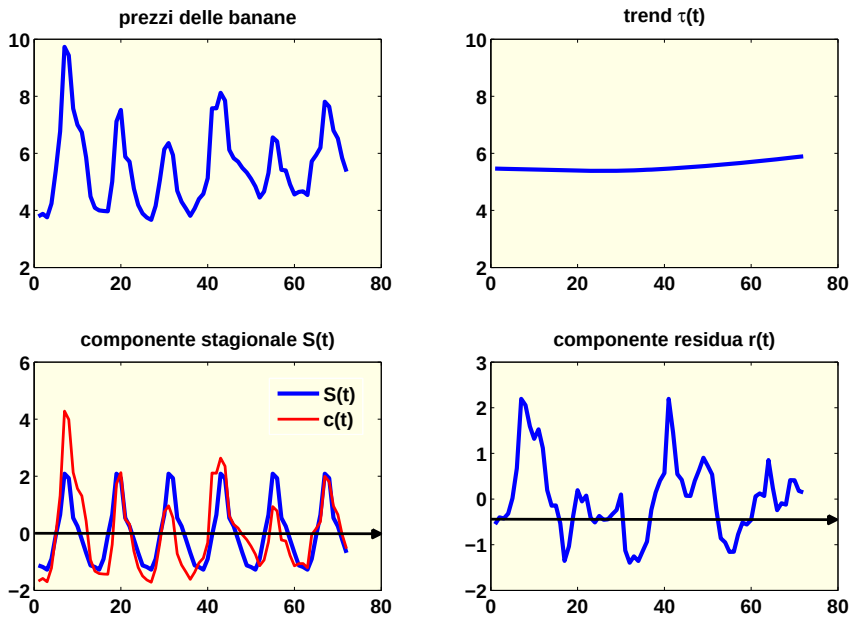

Figura 22: *analisi del prezzo delle banane: trend, componente stagionale, componente residua*

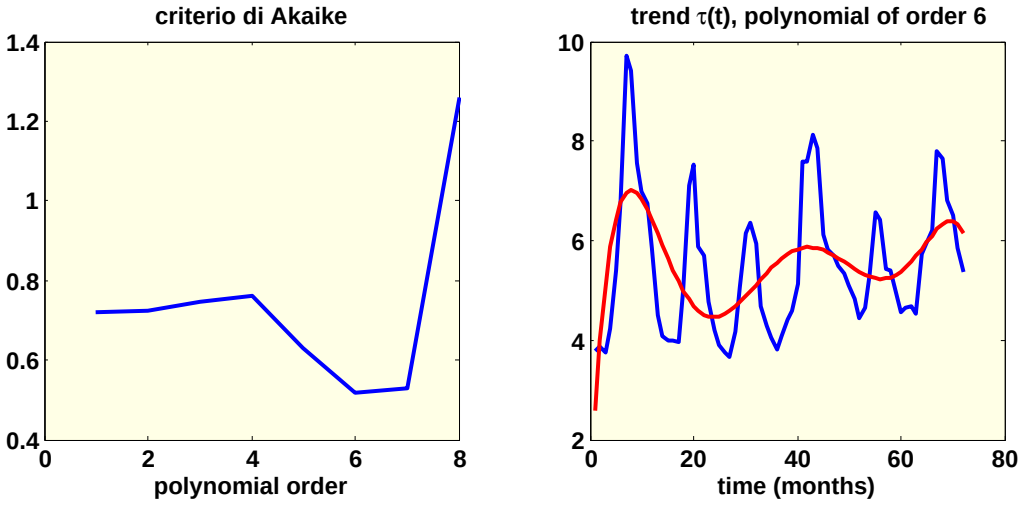

Figura 23: *stima del trend del prezzo delle banane con il criterio di Akaike*

seguire la strategia di selezionare la componente di trend in base a criteri di ottimalità non ha portato ad un buon risultato per quanto concerne l'analisi

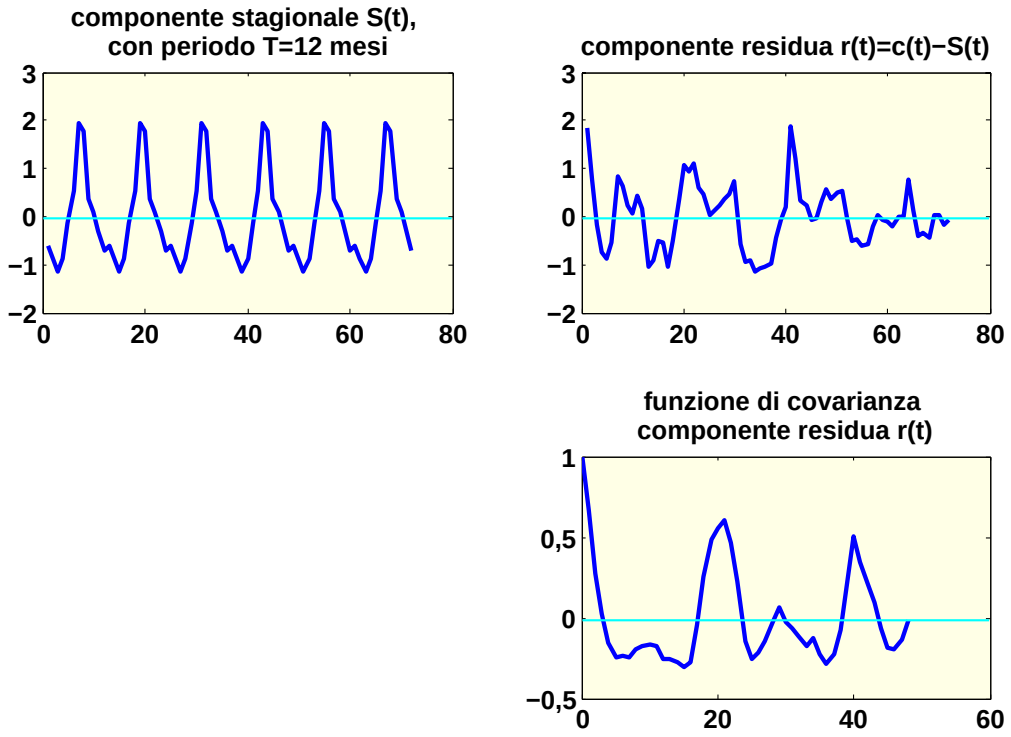

Figura 24: *componente stagionale, componente residua e sua funzione di autocorrelazione, della serie storica dei prezzi delle banane in Israele (1997-2002) tolto il trend polinomiale di sesto ordine*

della componente stagionale. Questo perchè il criterio di ottimalità di Akaike va utilizzato con una certa cautela, in quanto cerca il miglior compromesso tra un buon fitting e la complessità del modello di trend; ma la componente di trend non deve in alcun modo dare un buon fitting, deve solo indicare la tendenza generale della serie. Infatti se ripetiamo l'analisi con un polinomio di ordine 2, si ottiene l'analisi mostrata in Fig.25. Come si vede si ottengono praticamente gli stessi risultati ottenuti stimando il trend con il filtro di Prescott-Hodrick (si confrontino in particolare gli andamenti della funzione di autocorrelazione della componente residua in Fig.25 ed in Fig.21), e lo stesso si sarebbe ottenuto se si fosse scelto un polinomio del quarto ordine. Certo, disporre della forma analitica del trend della serie dei prezzi offre il vantaggio di poter predire la tendenza futura del prezzo, anche se solo per pochi mesi immediatamente successivi all'ultimo rilevamento dei dati.

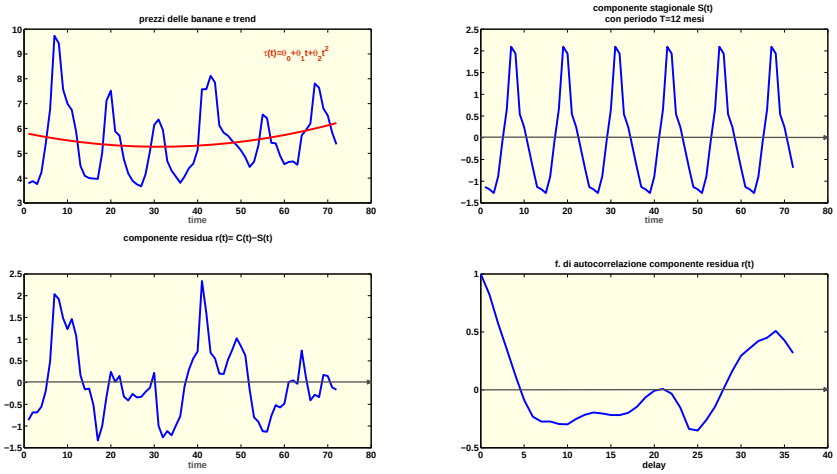

Figura 25: *serie dei prezzi e trend (polinomio di secondo ordine), componente stagionale, componente residua e sua funzione di autocorrelazione*

La componente residua

Individuati il trend $\tau(t)$ e la componente stagionale $S(t)$, il residuo

$$r(t) = y(t) - \tau(t) - S(t) \tag{30}$$

ha un andamento erratico, suscettibile di ulteriore analisi. Si può generalmente ipotizzare che la componente residua $r(t)$ sia una sequenza stazionaria secondo le statistiche del secondo ordine (sequenza debolmente stazionaria)

- la sequenza ha valor medio costante $E[r(t)] = \text{cost.}$, nel nostro caso pari a 0;
- la sequenza ha funzione di covarianza stazionaria

$$\phi(t+h, t) = E[r(t+h)r(t)] = \gamma(h), \quad \forall t, h$$

La seconda relazione indica che la correlazione tra due campioni qualunque della sequenza dipende non esplicitamente dagli istanti di tempo in cui sono presi, ma solo dalla differenza dei due istanti di tempo. Questo peraltro indica che la sequenza ha varianza costante $\gamma(0) = E[r(t)^2] = \sigma^2$. Per questo tipo di sequenze l'informazione è contenuta nella correlazione seriale dei suoi campioni: in altre parole è in genere possibile esprimere il generico valore $r(t)$ in funzione dei valori passati $r(t-1)$, $r(t-2)$, ... fino ad un certo ritardo m

$$r(t) = f(r(t-1), \ldots, r(t-m)) + \epsilon(t) \tag{31}$$

con la funzione $f(\cdot)$ che non dipende esplicitamente dal tempo. Il termine $\hat{r}(t) = f(r(t-1), \ldots, r(t-m))$ prende il nome di predizione ad un passo di $r(t)$, e quindi $\epsilon(t)$ è l'errore di predizione $r(t) - \hat{r}(t)$. Ogni trasformazione dei dati precedenti costituisce una predizione di $r(t)$ cui, dipendentemente dalla forma di $f(\cdot)$ e dall'ordine m, corrisponderà un errore di predizione differente. La predizione migliore si ha quando essa cattura tutta l'informazione contenuta nei campioni precedenti $r(t-1)$, $r(t-2)$, ... e determina un errore $\epsilon(t)$ *privo di informazione*. Una sequenza è priva di informazione quando i suoi campioni sono tutti tra loro indipendenti, per cui nessun insieme di essi può fornire informazione circa nessun'altro campione. Tali sequenze sono indicate come *sequenze i.i.d* (independent identically distributed) o *sequenze di rumore bianco*. La predizione ottima è quindi quella per cui l'errore di predizione $\epsilon(t)$ è una sequenza i.i.d..

L'espressione della predizione ottima può essere ottenuta in vari modi. Analizzeremo i modelli che si incontrano più frequentemente in pratica.

Modello AR(n)

Viene detto modello autoregressivo in quanto la predizione è ottenuta come media pesata dei campioni passati della serie fino ad un ritardo pari ad n, che definisce l'ordine del modello

$$r(t) = a_1\, r(t-1) + a_2\, r(t-2) + \ldots + a_n\, r(t-n) + \epsilon(t) \tag{32}$$

Si tratta quindi di determinare l'ordine n della regressione ed il valore dei parametri $a_1, \ldots, a_n$ per cui l'errore $\epsilon(t)$ risulti i.i.d.. Fissato l'ordine n, il valore dei parametri del modello si ottiene risolvendo il seguente programma di stima dei minimi quadrati

$$\{\hat{a}_1, \ldots, \hat{a}_n\} = \operatorname{argmin} \left[\sum_{j=1}^{N} \Big(r(t+j+n) - \ell(t+j+n)\theta \Big)^2 \right], \tag{33}$$

con $\theta = [a_1\, a_2 \ldots a_n]^T$ e $\ell(t) = [r(t-1)r(t-2)\ldots r(t-n)]$. L'identificazione del modello AR(n) viene quindi effettuata per iterazione, partendo ad esempio con $n = 1$, si calcolano i parametri del modello risolvendo il programma (33), si calcola l'errore di predizione e si effettua un test di bianchezza. Se il test fallisce si incrementa l'ordine n e si ricomincia. Man mano che aumentiamo la complessità del modello, controlliamo con il criterio di Akaike se non si sia raggiunta o meno la complessità ottima: se questo succede, ma l'errore del modello non soddisfa ancora il test di bianchezza, allora vuol dire che la sequenza in esame non è descrivibile mediante un modello AR(n). La

conclusione è la stessa anche se ci si vede costretti ad aumentare molto l'ordine del modello, cioè anche se non ri raggiunge la complessità ottima entro un numero limitato di valori di n.

Modello ARMA(n,p)

In questo modello si ha una componente autoregressiva di ordine n, come nel caso precedente, ma l'errore di modello è espresso come una media mobile di p valori di una sequenza i.i.d.

$$r(t) = a_1\, r(t-1)+a_2\, r(t-2)+,\ldots,+a_n\, r(t-n)+\epsilon(t)+b_1\, \epsilon(t-1)+\ldots+b_p\, \epsilon(t-p) \tag{34}$$

In questo caso il predittore ottimo si ottiene come una funzione $f(r(t-1),\ldots,r(t-n),\theta)$, $\theta = [a_1\, a_2 \ldots a_n\, b_1\, b_2 \ldots b_p]^T$, lineare nei dati $[r(t-1)r(t-2)\ldots r(t-n)]$ ma non lineare nei parametri θ. La stima dei parametri incogniti del modello si ottiene quindi risolvendo un problema di ottimo non lineare con vincoli

$$\hat{\theta} = \min_{\theta \in D} \left[\sum_{j=1}^{N} \Big(r(t+j+n) - f(r(t+j+n-1),\ldots,r(t+j),\theta) \Big)^2 \right], \tag{35}$$

dove l'iniseme ammissibile è definito come i valori di $b_1\, b_2 \ldots b_p$ per cui il polinomio $z^p + b_1\, z^{p-1} + \ldots + b_p$ ha tutte radici interne al cerchio di raggio unitario. Esistono vari applicativi in grado di risolvere il programma (35) con opportuni algoritmi. Per cui l'identificazione procede in modo iterativo come per i modelli AR(n), tenendo conto che la complessit del modello in questo caso è pari a $n + p$, e l'aggiornamento del modello si ottiene aumentando l'ordine della parte autoregressiva e/o l'ordine della parte di media mobile. Il procedimento termina nel momento in cui l'errore di predizione $\epsilon(t) = r(t) - f(r(t-1),\ldots,r(t-n),\hat{\theta})$ soddisfa il test di bianchezza. Come nel caso precedente, qualora si raggiunga la complessità massima secondo Akaike ed il test di bianchezza non sia ancora soddisfatto, si deve ritenere che anche la classe di modelli ARMA(n,p) non sia adatta a rappresentare la componente residua in esame.

Le classi di modelli adottate per sequenze debolmente stazionarie, sono casi particolari della famiglia di modelli Box-Jenkins. Esistono poi altre famiglie in grado di descrivere altre cause di non stazionarietà della sequenza, oltre al trend ed alla componente stagionale, che consiste ad esempio nel

fenomeno del *volatility clustering*, tipico delle sequenze dei *returns* di serie finanziarie: nella sequenza si individuano sottosequenze (cluster) in cui la varianza è costante, ma varia molto da cluster a cluster. Questi sono i modelli ARCH(n) (autoregressive conditionally heteroschedastic)

$$\begin{aligned} r(t) =& \mu + \sigma(t)\,\epsilon(t) \\ \sigma(t)^2 =& \alpha_0 + \alpha_1\,\epsilon(t-1)^2 + \ldots + \alpha_n\,\epsilon(t-n)^2 \end{aligned}$$

ed i modelli GARCH(n,p) (generalized autoregressive conditional heteroskedastic)

$$\begin{aligned} r(t) =& \mu + \sigma(t)\,\epsilon(t) \\ \sigma(t)^2 =& \alpha_0 + \alpha_1\,\epsilon(t-1)^2 + \ldots + \alpha_n\,\epsilon(t-n)^2 \\ & + \beta_1\,\sigma(t-1)^2 + \ldots + \beta_p\,\sigma(t-p)^2 \end{aligned}$$

L'identificazione di questi tipi di modelli comporta la soluzione di un problema di ottimizzazione non lineare vincolato. Algoritmi adatti alla soluzione del problema sono disponibili negli applicativi di largo uso come Matlab, R^2, Eviews, SAS, SPSS, ...

Analizziamo ora la componente residua (30) dell'analisi del prezzo delle banane. Iniziamo con modellare $r(t)$ come una sequenza AR(n) con ordine $n = 1, 2, \ldots$ ed applichiamo la procedura iterativa di identificazione di cui si discusso nel relativo paragrafo. Risolvendo il problema (33) si ottiene

$$\begin{aligned} n =& 1, \quad a_1 = 0.8266, \quad FPE = 0.21848 \\ n =& 2 \quad a_1 = 1.076,\ a_2 = -0.302, \quad FPE = 0.205444 \\ n =& 3 \quad a_1 = 1.07,\ a_2 = -0.2804,\ a_3 = -0.02011, \quad FPE = 0.212812 \end{aligned}$$

Come si nota nel passaggio da $n = 1$ a $n = 2$ il valore di FPE (Final Prediction Error del criterio di Akaike) diminuisce, indicando quindi che il modello di ordine 2 è migliore di quello di ordine 1. Tuttavia quando si passa al modello di ordine 3, il valore di FPE aumenta, indicando quindi che $n = 2$ fornisce il modello con la complessità migliore. A questo punto eseguiamo il test di bianchezza di Ljung-Box sulla sequenza di errore

$$\epsilon(t) = r(t) - 1.076r(t-1) + 0.302r(t-2)$$

per ritardi pari a 2, 3, 4 e 5. Si ottiene che l'ipotesi nulla che la sequenza di errore $\epsilon(t)$ non abbia correlazione seriale, sia quindi priva di informazione,

deve essere accettata per tutti e quattro i valori del ritardo. L'errore quindi è una sequnza i.i.d. per cui la componente residua $r(t)$ ha una struttura autoregresiva di ordine 2. La Figura 26 mostra infine tutte le componenti dell'analisi additiva effettuata.

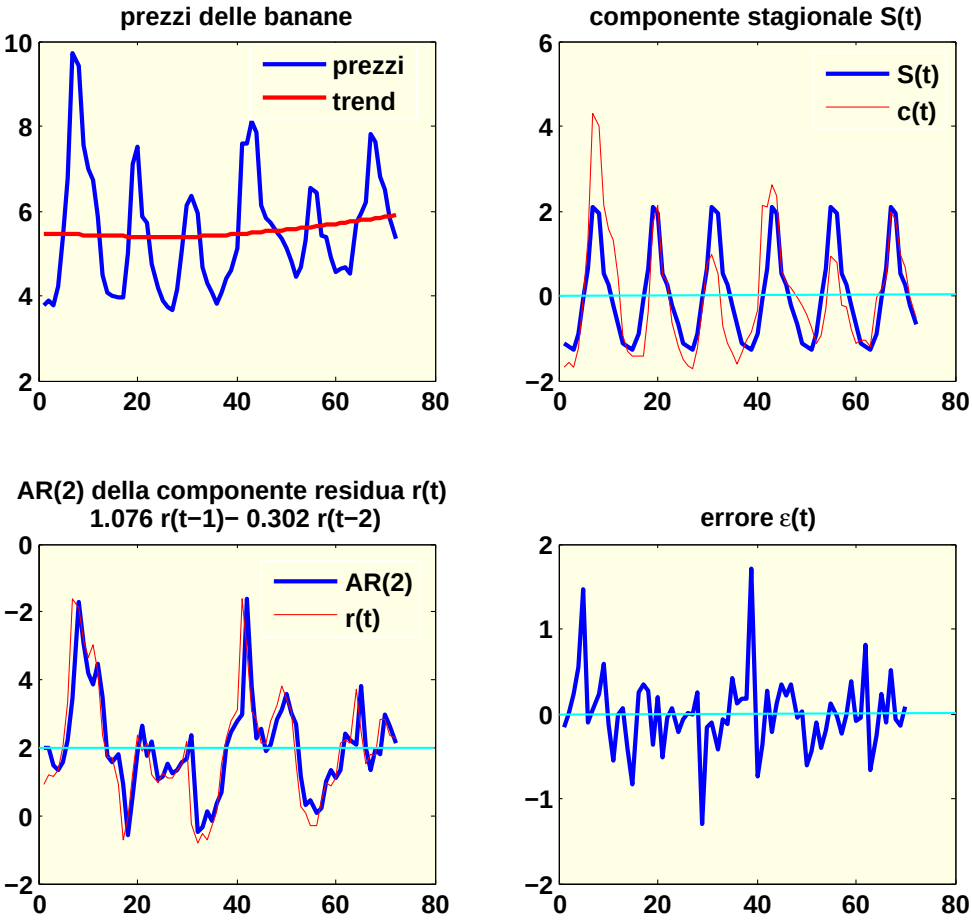

Figura 26: *serie dei prezzi e trend (polinomio di secondo ordine), componente stagionale, componente residua ed errore del modello AR(2) della componente residua*