Quaderni di Statistica
Vol. 4, 2002

Una verifica della modellistica a differenza frazionaria per la serie delle portate del Tevere

Marcella Corduas e Domenico Piccolo
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
Centro per la Formazione in Economia e Politica dello Sviluppo Rurale
E-mail: marcella.corduas@unina.it; domenico.piccolo@unina.it

Summary: The class of *ARFIMA* models offers a flexible tool to describe long memory time series and it has a special role for modelling hydrological data. In this paper we review the statistical properties and the main estimation techniques for such models. An application to the daily flows of Tevere illustrates how this class of models can be applied in practice.

Key words: Fractional differencing, Long-term Persistence, Hydrological time series

1. Introduzione

I modelli a differenza frazionaria sono stati introdotti da Granger e Joyeaux (1980) e da Hosking (1981) per descrivere la dinamica di fenomeni che presentano una struttura di dipendenza che persiste nel tempo ed hanno trovato interesse ed applicazioni in numerosi e differenti ambiti disciplinari. In tutti i casi, essi si prestano alla interpretazione ed alla modellistica di processi le cui realizzazioni evidenziano correlazioni seriali significative anche se misurate in tempi molto distanti.

Applicazioni interessanti e discussioni su serie di interesse economico e finanziario sono state prodotte, tra gli altri, dai lavori di: Diebold e Rudenbusch (1989, 1991), Shea (1991), Cheung (1993), Ray (1993), Delgado e Robinson (1994), Hassler e Wolters (1995). A tale riguardo, può essere utile consultare la recente rassegna di Zaffaroni (2002). Di gran

lunga più numerose sono, invece, le applicazioni al caso di dati metereologici e idrologici, sin dai primi lavori di Hosking (1982, 1984); per il nostro Paese, un recente convegno ha affrontato tale problematica (Piccolo e Ubertini, 2001). Infine, per il settore delle telecomunicazioni, un contributo riferito a tali dinamiche è quello di De Giovanni (2002).

Per illustrare gli aspetti tipici di tale dinamica riportiamo nella Figura 1 la serie storica annuale dei livelli di minimo del fiume Nilo rilevati alla foce (come riportata da Tousson, 1925) negli anni 622-1284. I grafici rappresentano sia la serie originaria che le stime della funzione di autocorrelazione globale e del periodogramma, che forniscono una sintesi dei principali aspetti temporali e frequenziali presenti nelle osservazioni.

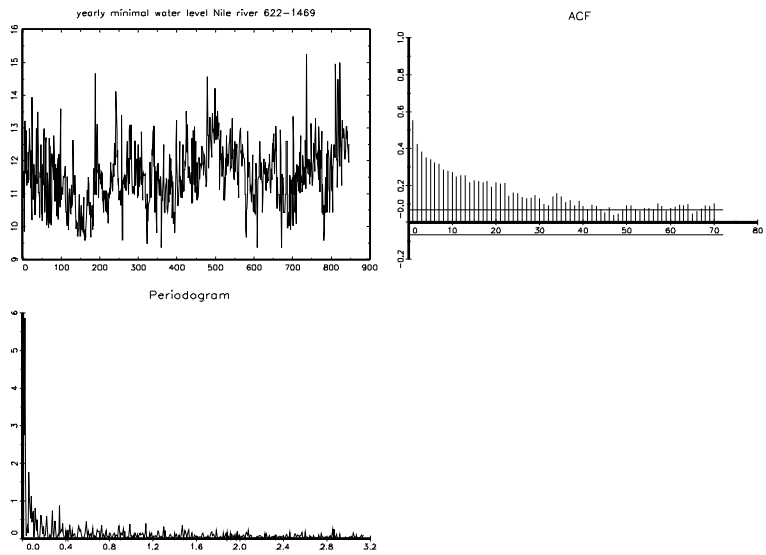

Figura 1. Livello minimo annuale del Nilo alla foce presso il Cairo

In particolare, si osserva che l'evoluzione della serie è sostanzialmente stazionaria in media anche se, localmente, presenta delle alternanze nei livelli che però non generano nè componenti cicliche nè trend locali. Nella letteratura questi effetti vengono spesso denominati *effetto Noè* ed *effetto Giuseppe*, intendendo, nel primo caso, l'effetto determinato

da valori, regolarmente e costantemente, sopra la media per lunghissimi periodi di tempo (con allusione all'episodio biblico del diluvio universale che ruota attorno alla figura di Noè) e riferendosi, nel secondo caso, ad alternanze di sequenze di osservazioni, sopra la media e sotto la media della serie, di uguale lunghezza (con allusione alla narrazione dei lunghi anni di abbondanza e carestie che ruota attorno alla figura biblica di Giuseppe).

D'altra parte, la funzione di autocorrelazione stimata mostra un andamento che decade a zero con estrema lentezza e con presenza di valori significativi anche a lags piuttosto elevati. Ad essa, sul piano frequenziale, corrisponde il periodogramma che conferma chiaramente come -nella spiegazione della variabilità complessiva del fenomeno- siano prevalenti le componenti a frequenza molto bassa.

Gli andamenti, qui evidenziati nel caso di una serie annuale, sono una costante per la gran parte delle serie idrologiche (e metereologiche), anche a prescindere dalla frequenza di rilevazione. Infatti, alle componenti sopra evidenziate, nel caso di serie mensili o giornalierie vanno aggiunte, rispettivamente, le corrispondenti frequenze stagionali (per il moto della Terra attorno al Sole che riproduce condizioni ambientali più o meno stabili ogni anno) ed un notevole aumento della variabilità stocastica (soprattutto per serie giornaliere, quando non si tratta di serie il cui comportamento è condizionato da fenomeni "pianificati" per esigenze collettive: dighe, serbatoi, sbarramenti periodici, etc.).

Appare quindi necessario, per adeguare la modellistica standard alla dinamica reale dei fenomeni idrologici, ampliare la classe dei processi lineari *ARIMA* (Box e Jenkins, 1970; Piccolo, 1990) introducendo una classe di processi che -alle tradizionali componenti stazionari ed invertibili di tipo *ARMA* e alla presenza di un operatore intero per spiegare trend di natura stocastica- aggiunga esplicitamente una componente di lunghissimo periodo, che spieghi i comportamenti, prima evidenziati, nella funzione di autocorrelazione globale e in quella spettrale.

In tale prospettiva, la classe dei modelli *ARFIMA (AutoRegressive Fractional Integrated Moving Average)* offre uno strumento flessibile ed ulteriormente generalizzabile per descrivere tale tipo di dinamica.

Il lavoro è così organizzato: dopo aver introdotto nel paragrafo 2 gli elementi strutturali che caratterizzano i modelli *ARFIMA*, richiameremo nel paragrafo 3 le problematiche della stima dei parametri e, successivamente, nel paragrafo 4 quelle della previsione. Infine, il paragrafo 5 sarà dedicato alla descrizione delle varie fasi per la costruzione di tali modelli

per la serie storica giornaliera delle portate del fiume Tevere. Tale esperienza viene qui presentata come una verifica empirica delle questioni metodologiche che una serie idrologica pone alla modellistica per serie storiche.

2. I processi ARFIMA

Un processo $Z_t \sim ARFIMA(p,d,q)$ di valor medio nullo è caratterizzato dalla seguente formulazione:

$$\phi(B)\nabla^d Z_t = \theta(B)a_t \tag{1}$$

dove a_t $\sim$WN(0,σ^2) è un processo White Noise Gaussiano, cioé una successione di variabili casuali (v.c.) Gaussiane di valore medio zero, omoschedastiche e incorrelate; B denota l'operatore ritardo: $B^k Z_t = Z_{t-k}$, $k = 0, 1, ...$; i polinomi: $\phi(B) = (1 - \phi_1 B - ... - \phi_p B^p)$ e $\theta(B) = (1 - \theta_1 B - ...\theta_q B^q)$ non hanno fattori comuni, mentre le $p + q$ radici dell'equazione $\phi(B)\theta(B) = 0$ sono tutte esterne al cerchio unitario. Inoltre, l'operatore alle differenze ∇^d è definito da:

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k =$$

$$= 1 - dB - \frac{1}{2}d(1-d)B^2 - \frac{1}{6}d(1-d)(2-d)B^3 - ... \tag{2}$$

Si dimostra (Hosking, 1981) che se $d < 0.5$, la formulazione $MA(\infty)$:

$$Z_t = \psi(B)a_t, \tag{3}$$

è ben definita avendo posto:

$$\psi(B) = \frac{\theta(B)}{\phi(B)(1-B)^d} = 1 + \sum_{k=1}^{\infty} \psi_k B^k, \tag{4}$$

con $\sum_{k=1}^{\infty} \psi_k^2 < \infty$.

Analogamente, se $d > -0.5$, è ben definita la formulazione $AR(\infty)$:

$$\pi(B)Z_t = a_t, \tag{5}$$

dove abbiamo posto:

$$\pi(B) = \frac{\phi(B)(1-B)^d}{\theta(B)} = 1 - \sum_{k=1}^{\infty} \pi_k B^k, \tag{6}$$

con $\sum_{k=1}^{\infty} \pi_k^2 < \infty$.

Le caratteristiche di processi appartenenti a tale classe possono essere sintetizzate dall'andamento della funzione di autocorrelazione globale $\rho(k)$, e dualmente -sul piano frequenziale- della funzione di densità spettrale (non-normalizzata) $f(\omega)$. A tal fine, esaminiamo in dettaglio un processo $Z_t \sim ARFIMA(0, d, 0)$, che descrive la sola dinamica di lunga memoria, ovvero:

$$\nabla^d Z_t = a_t. \tag{7}$$

In particolare avremo:

i) per $d < 0.5$, la formulazione $MA(\infty)$ descritta da: $Z_t = \sum_{k=0}^{\infty} \psi_k a_{t-k}$ è caratterizzata dai pesi MA: $\psi_k = \frac{(k+d-1)!}{k!(d-1)!}$, $k = 0, 1, ...$ i quali decadono a zero ad un tasso iperbolico essendo per $k \to \infty$, $\psi_k \sim k^{d-1}/(d-1)!$;

ii) per $d > -0.5$, la formulazione $AR(\infty)$: $Z_t = \sum_{k=1}^{\infty} \pi_k Z_{t-k} + a_t$, è caratterizzata dai pesi AR: $\pi_k = (-1)\frac{(k-d-1)!}{k!(-d-1)!}$, $k = 1, 2, ...$ tali che per $k \to \infty$, $\pi_k \sim (-1)\, k^{d-1}/(-d-1)!$

Si osservi che, invece, per i processi ARMA stazionari e invertibili sia i coefficienti AR $\{\pi_k\}$ che i coefficienti MA $\{\psi_k\}$ decadono a zero con un andamento esponenziale, il che implica un più rapido esaurirsi dei legami di correlazione seriale al crescere del ritardo temporale.

Inoltre, la funzione di autocorrelazione globale per il processo $Z_t \sim ARFIMA(0, d, 0)$ è la seguente:

$$\rho(k) = \frac{(-d)!(k+d-1)!}{(d-1)!(k-d)!}, \; k = 0, \pm 1, ... \tag{8}$$

Essa assume valori tutti negativi per $-0.5 < d < 0$ e valori tutti positivi per $0 < d < 0.5$ con una più lenta decadenza a zero rispetto a quanto accade per i processi ARMA, essendo per $k \to \infty$, $\rho(k) \sim k^{2d-1}(-d)!/(d-1)!$

Vale la pena di sottolineare che la funzione di autocorrelazione parziale per tale processo è particolarmente semplice, essendo espressa da:

$$\pi(k) = \frac{d}{(k-d)}, \quad k = \pm 1, ...; \quad \pi(0) = 1. \tag{9}$$

Infine, la funzione di densità spettrale (non-normalizzata) è definita come:

$$f(\omega) = \frac{\sigma^2}{2\pi}|1 - e^{-i\omega}|^{-2d} = \frac{\sigma^2}{2\pi}[2sen(\omega/2)]^{-2d}, \quad -\pi < \omega \leq \pi. \tag{10}$$

Per $-0.5 < d < 0$, $f(\omega) \to 0$ per $\omega \to 0$; la densità spettrale è dominata dalle componenti ad alta frequenza, e individua così un processo con una dinamica estremamente alternata nei valori e con correlazioni negative, definito *antipersistente*.

Per $0 < d < 0.5$, $f(\omega) \to \infty$ per $\omega \to 0$; pertanto, la densità spettrale è dominata dalle componenti a bassa frequenza e il processo è caratterizzato da una dinamica di *lunga memoria*.

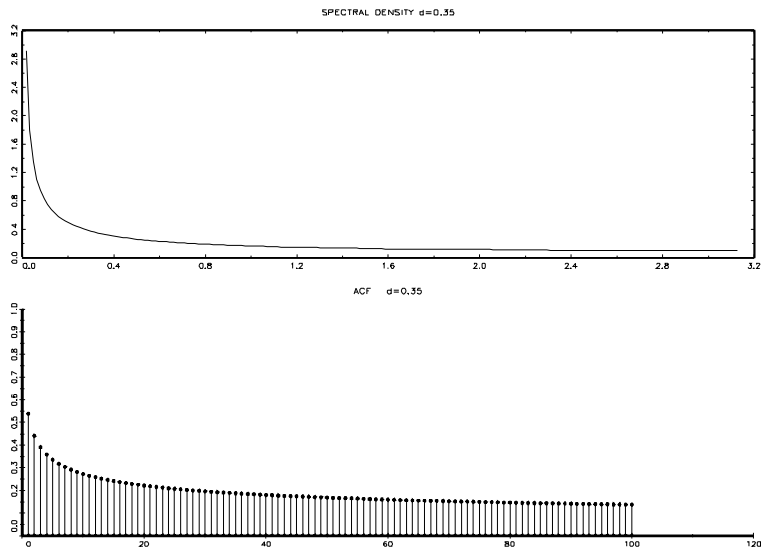

Figura 2. Funzione di densità spettrale e funzione di autocorrelazione globale del processo ARFIMA(0,d,0) per d=0.35

Tenuto conto di tali risultati, nella misura in cui questo lavoro è finalizzato prioritariamente alla applicazione della modellistica *ARFIMA* a serie idrologiche, nel seguito (ove non specificato diversamente), limiteremo la nostra trattazione al solo caso in cui $0 < d < 0.5$. Infatti, le serie che generalmente si registrano in tale ambito -per quanto discusso in precedenza- sono quasi esclusivamente specificate da componenti di lunga memoria, assieme a quelle più tradizionali. Ad esempio, la Figura 2 illustra le funzioni teoriche di autocorrelazione globale e della densità spettrale (non-normalizzata) per un processo *ARFIMA(0,d,0)* quando $d = 0.35$. Pur trattandosi di schemi teorici, questi andamenti risultano molto spesso adeguati a rappresentare componenti di lunga memoria presenti nelle serie idrologiche.

Ritornando alla formulazione del processo $Z_t \sim ARFIMA\,(p, d, q)$ dovrebbe essere ora evidente il ruolo dei corrispondenti operatori. Difatti, mentre l'operatore a differenza frazionaria descrive le componenti di lunga memoria, che si esplicitano e risultano dominanti nel lungo periodo, i polinomi AR e MA rappresentano ed esauriscono le componenti con memoria breve, che spiegano il loro effetto prevalentemente ai primi lags della struttura di correlazione seriale.

Va inoltre sottolineato che la modellistica *ARFIMA* è definita per qualsiasi $d > -0.5$ reale, perchè è sempre possibile trovare un valore intero $d_1 \geq 0$ e un valore reale $d_2 \in (-0.5, 0.5)$ tali che: $d = d_1 + d_2$. In sostanza, è sempre possibile differenziare il processo con operatori alle differenze di grado intero e ritrovare la parte frazionaria tramite un operatore alle differenza frazionaria. Per fenomeni sostanzialmente stazionari a livello di epoche storiche, come quelli idrologici e metereologici, è ragionevole assumere (come noi faremo, nel seguito) che $d \in [0, 0.5)$.

3. La stima dei parametri

In letteratura sono stati proposti vari metodi di stima per i parametri dei modelli a differenza frazionaria. La particolare struttura di correlazione determina, infatti, molteplici problemi computazionali nel calcolo della funzione di verosimiglianza esatta e nella conseguente ricerca delle stime di massima verosimiglianza dei parametri.

Tali problematiche, ovviamente comuni ad altri ambiti dell'inferenza

statistica, nel caso della modellistica *ARFIMA* diventano piuttosto gravosi. Difatti, affinché gli effetti di lunga memoria siano osservabili e siano statisticamente discernibili da altre componenti, pure presenti nei dati, occorre esaminare serie storiche con numerosità superiore (spesso, di gran lunga superiore) al migliaio di osservazioni.

La letteratura sull'argomento è piuttosto ampia ed è necessario farvi riferimento per gli opportuni approfondimenti[1]; nel seguito, ci limitiamo a richiamare brevemente i due approcci piu' comunemente utilizzati:

i) la *stima semiparametrica* (Geweke e Porter-Hudak, 1983) che ha dato origine a molteplici varianti e successivi miglioramenti;

ii) la *stima di massima verosimiglianza*, sviluppata sia nel dominio temporale (Sowell, 1992) che nel dominio frequenziale (Fox e Taqqu, 1986).

• Per la stima semiparametrica, il punto di partenza è la funzione di densità spettrale (non-normalizzata) di $Z_t \sim ARFIMA\,(p,d,q)$:

$$f_F(\omega) = |1 - e^{-i\omega}|^{-2d} f_{ARMA}(\omega), \quad -\pi < \omega \leq \pi \tag{11}$$

dove abbiamo indicato con $f_{ARMA}(\omega)$ la funzione di densità spettrale del processo $X_t \sim ARMA\,(p,q)$, definito da: $X_t = \nabla^d Z_t$, ovvero:

$$f_{ARMA}(\omega) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\omega})|^2}{|\phi(e^{-i\omega}|^2}, \ -\pi < \omega \leq \pi. \tag{12}$$

Considerando il logaritmo della (11), e ricordando la (10), si avrà:

$$\ln(f_F(\omega)) = -d \ln(4sen^2(\omega/2)) + \ln(f_{ARMA}(\omega)) \tag{13}$$

e quindi, addizionando e sottraendo $\ln(f_{ARMA}(0))$, si ottiene:

$$\begin{aligned} \ln(f_F(\omega)) = \ln(f_{ARMA}(0)) - d \ln(4sen^2(\omega/2)) + \\ + \ln(f_{ARMA}(\omega)/f_{ARMA}(0)) \end{aligned} \tag{14}$$

Dalla serie storica osservata $\{z_t,\ t = 1, 2, ..., n\}$, posto $\widetilde{z}_t = z_t - \overline{z}$ la serie centrata, è possibile stimare il periodogramma:

[1]Una recente rassegna dei contributi più significativi è stata curata da Marinucci (2002). Per altri approcci, si vedano anche: Beran, 1995; Cancelliere et al. (2001); Corduas e Piccolo (2001).

$$I_n(\omega_j) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} \tilde{z}_t e^{-i\omega_j t} \right|^2 \tag{15}$$

alle frequenze di Fourier: $\omega_j = 2\pi j/n$, $j = 1, 2, ..., n$. Pertanto, la (14) può essere riferita alle frequenze prossime a zero: $\omega_j,\ j = 1, 2...g(n)$, e riscritta come segue:

$$\begin{aligned} \ln\{I_n(\omega_j)\} &= \ln(f_{ARMA}(0)) - d\ln\{4sen^2(\omega_j/2)\} + \\ &+ \ln\{f_{ARMA}(\omega_j)/f_{ARMA}(0)\} + \ln\{I_n(\omega_j)/f_F(\omega_j)\} \end{aligned} \tag{16}$$

Posto che $\ln\{f_{ARMA}(\omega_j)/f_{ARMA}(0)\}$ è una quantità trascurabile se si concentra l'attenzione alle sole frequenze prossime a zero, Geweke e Porter-Hudak (1983) interpretano tale relazione come un modello di regressione nel quale:

- $\ln\{I_n(\omega_j)\}$ è la variabile dipendente;
- $\ln\{4sen^2(\omega_j/2)\}$ è la variabile esplicativa;
- $\ln\{I_n(\omega_j)/f_F(\omega_j)\} + \eta$ è la v.c. errore;
- $\ln(f_{ARMA}(0)) - \eta$ è l'intercetta;
- il parametro d è il coefficiente angolare.

La presenza della costante di Eulero, η, è richiesta dal fatto che la sua aggiunta nel termine di errore consente di avere v.c. il cui valore medio è asintoticamente pari a zero, come richiesto dai modelli di regressione[2]. La stima del parametro d è quindi ottenuta mediante il metodo dei minimi quadrati assumendo che gli errori siano indipendenti ed identicamente distribuiti. Nel lavoro originario, inoltre, Geweke e Porter-Hudak suggeriscono di considerare $g(n) = n^{1/2}$.

I successivi sviluppi sono stati diretti alla soluzione di alcuni problemi che questa procedura implicitamente propone per il venir meno di alcune delle assunzioni classiche su cui è fondato il modello di regressione lineare. Robinson (1995) e Hurvich e Beltrao (1993) hanno infatti dimostrato che le ordinate del periodogramma normalizzato $I_n(\omega_j)/f_F(\omega_j)$,

[2] Ai fini pratici, come si vede, la costante di Eulero non compare esplicitamente nella (16); essa dovrebbe essere presa in considerazione qualora si fosse interessati alla stima dell'intercetta ma, come spiegato nel testo, il solo uso del modello di regressione (16) consiste nella stima del coefficiente angolare d, sul quale evidentemente la presenza della costante di Eulero è irrilevante.

per j fissato, asintoticamente sono distorte, tra loro correlate e non identicamente distribuite. Hurvich et al.(1998), poi, hanno derivato l'errore quadratico medio asintotico dello stimatore di Geweke e Porter-Hudack come funzione del numero di ordinate del periodogramma che sono utilizzate nel modello di regressione, giungendo alla conclusione che $g(n) = O(n^{4/5})$ costituisce un valore ottimale che minimizza l'errore quadratico medio.

Inoltre, molti lavori si sono concentrati sulla delimitazione delle frequenze ω_j realmente utili nella determinazione della stima del parametro d: l'orientamento prevalente è a favore della inclusione nel modello di regressione (16) di un numero estremamente ridotto di frequenze angolari vicine allo zero.

Infine, va aggiunto che tale approccio, anche con le varianti ed i miglioramenti introdotti successivamente alla proposta originaria, consente la stima del solo parametro d, per cui subordinatamente a tale stima la serie storica va filtrata (troncando la formulazione AR esplicitata nella (6) ad un lag conveniente) per pervenire alla serie x_t su cui identificare e, quindi, stimare i parametri della componente $ARMA$ del modello *ARFIMA*.

• Per quanto concerne gli stimatori di massima verosimiglianza dei parametri di un modello *ARFIMA* conviene distinguere le procedure che procedono nel dominio temporale da quelle che procedono nel dominio frequenziale.

i) dominio temporale

Posto $\mathbf{z} = (\tilde{z}_1, ..., \tilde{z}_n)'$ il vettore delle osservazioni centrate rispetto alla media, e definita la matrice: $\mathbf{\Sigma}_n = \{\gamma_{|i-j|}, i = 1, .., n; j = 1, ..., n\}$, dove $\gamma_k = Cov(Z_t, Z_{t-k})$ è la funzione di autocovarianza, la funzione di log-verosimiglianza è espressa da:

$$l(d, \phi, \theta, \sigma^2; \mathbf{z}) = -\frac{n}{2} \ln (2\pi) - \frac{1}{2} \ln |\mathbf{\Sigma}_n| - \frac{1}{2} \mathbf{z}' \mathbf{\Sigma}_n^{-1} \mathbf{z}. \qquad (17)$$

I problemi connessi con l'utilizzo di tale formulazione derivano dal calcolo delle autocovarianze $\{\gamma_0, ..., \gamma_{n-1}\}$ che intervengono nella costruzione della matrice $\mathbf{\Sigma}_n$ e dalla usuale elevata numerosità dei dati (che richiede specifiche tecniche per il trattamento di matrici di dimensioni

elevate). Sowell (1992) ha fornito una prima soluzione a tali problemi, sfruttando una formulazione delle autocovarianze che ne consente la valutazione mediante il calcolo ricorsivo di funzioni ipergeometriche. Il lavoro di Doornik e Ooms (2001) presenta una aggiornata rassegna delle varie tecniche di calcolo applicabili in tale contesto, mentre un ulteriore miglioramento computazionale è stato recentemente proposto da Bertelli e Caporin (2002).

ii) dominio delle frequenze

Il punto di partenza di tutti i risultati successivi è la (17) nella quale si utilizzano due importanti risultati asintotici, ben noti nella letteratura classica delle serie storiche e -nel nostro contesto- accettabili stante la elevata numerosità delle osservazioni. Essi trasformano le espressioni che coinvolgono il determinante e l'inversa della matrice delle varianze e covarianze (che dipende dal vettore $\beta = (d, \phi, \theta, \sigma^2)'$ dei parametri da stimare) in quantità che, invece, coinvolgono la densità spettrale, nel modo seguente:

$$ln|\mathbf{\Sigma_n}(\beta)| \rightarrow \sum_{j=1}^{n} \ln\, f(\omega_j; \beta); \tag{18}$$

$$\mathbf{z}'\left[\mathbf{\Sigma_n}(\beta)\right]^{-1}\mathbf{z} \rightarrow \sum_{j=1}^{n} \frac{I_n(\omega_j)}{f(\omega_j; \beta)}. \tag{19}$$

Ora, se si indica con γ il vettore dei parametri ad esclusione della varianza σ^2 del processo WN, cioè se si pone $\gamma = (d, \phi, \theta)$ tenendo presente che: $f(\omega_j; \beta) = \left(\frac{\sigma^2}{2\pi}\right) g(\omega_j; \gamma)$, sostituendo le espressioni (18) e (19) nella (17), si ottiene una approssimazione asintotica alla funzione di verosimiglianza, detta di Whittle, cioè:

$$l(d, \phi, \theta, \sigma^2; \mathbf{z}) = -\frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\sum_{j=1}^{n}\left\{\ln\, g(\omega_j; \gamma) + \frac{2\pi}{\sigma^2}\,\frac{I_n(\omega_j)}{g(\omega_j; \gamma)}\right\}. \tag{20}$$

Da essa, è agevole dedurre la stima di σ^2 che vale:

$$\widehat{\sigma}^2 = \frac{2\pi}{n}\sum_{j=1}^{n}\frac{I_n(\omega_j)}{g(\omega_j; \gamma)}. \tag{21}$$

Se si sostituisce la stima (21) nella funzione di log-verosimiglianza (20) (cioè se si "concentra" rispetto a σ^2), si ottiene una funzione che dipende dai soli parametri espliciti del modello *ARFIMA*, cioè:

$$l(d,\phi,\theta;\mathbf{z}) \propto -\frac{n}{2}\ln\left[\sum_{j=1}^{n}\frac{I_n(\omega_j)}{g(\omega_j;\gamma)}\right] - \frac{1}{2}\sum_{j=1}^{n}\ln\, g(\omega_j;\gamma) \qquad (22)$$

Mediante tale approccio, D'Elia e Piccolo (2002) hanno derivato la stima congiunta del parametro d della differenza frazionaria e del parametro λ della trasformazione normalizzante di Box e Cox. Essa è diffusa come trasformazione preliminare anche nell'analisi dei dati idrologici, perchè essi spesso derivano da distribuzioni marcatamente asimmetriche.

4. La previsione

Una delle applicazioni tipiche dei modelli per serie storiche è costituita dalla previsione, che giustifica l'ingente sforzo di specificazione e stima che la costruzione di un modello spesso implica. In effetti, la generazione di previsioni da modelli *ARFIMA* comporta delle ovvie approssimazioni determinate proprio dalla struttura di 'lunga memoria' del processo generatore e dalla inevitabile limitatezza della realizzazione a disposizione.

Il previsore $F_{n,k}$ per Z_{n+k}, $k>0$, basato sull'insieme di informazioni $I_n=\{Z_{n-j}, j=0,1,2,...\}$ disponibili su Z_t al tempo $t=n$, è ottenuto in maniera analoga a quanto ben noto per la più consolidata modellistica $ARIMA$.

Pertanto, posto $F_{n,k}=E(Z_{n+k}|I_n)$, sfruttando la formulazione $AR(\infty)$ definita nella (4), avremo:

$$F_{n,k}=\sum_{j=1}^{\infty}\pi_j\,E(Z_{n+k-j}|I_n)+E(a_{n+k}|I_n)=\sum_{j=1}^{\infty}\pi_j\,\widehat{Z}_{n+k-j} \qquad (23)$$

essendo $E(a_{n+k}|I_n)=0$, e dove abbiamo posto:

$$\widehat{Z}_{n+k-j}=\begin{cases} Z_{n+k-j}, & \text{per } k\le j \\ F_{n,k-j}, & \text{per } k>j. \end{cases} \qquad (24)$$

Se, invece, si parte dalla formulazione $MA(\infty)$ definita nella (3), ricordando che $\psi_0 = 1$, è possibile determinare l'errore di previsione:

$$e_{n,k} = Z_{n+k} - F_{n,k} = \sum_{j=0}^{k-1} \psi_j \, a_{n+k-j} \tag{25}$$

e la relativa varianza:

$$Var(e_{n,k}) = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2. \tag{26}$$

Diversamente da quanto accade allorchè l'esponente dell'operatore alle differenze è un numero intero, nel caso di d frazionario non è possibile, partendo dall'equazione alle differenze, individuare una formula per produrre e aggiornare le previsioni che sia espressa con un numero finito di termini. In pratica, quando si dispone di una realizzazione rilevata per $t = 1, 2, ..., n$, bisognerà troncare opportunamente lo sviluppo AR indicato nella (19) per stimare la previsione.

A tale riguardo, va ricordato che il problema della previsione (anche per i processi *ARFIMA*) va correttamente posto per l passi avanti, $l = 1, 2, ...$; alcune proposte di soluzione per la modellistica che qui ci interessa sono state affrontate recentemente da Bhansali e Kokoszka (2002).

5. La simulazione e la generazione di serie sintetiche

Nell'ambito dello studio di dati idrologici, assumono particolare interesse le tecniche per simulare e generare serie sintetiche da un modello prefissato.

Una tecnica di simulazione efficace sfrutta le relazioni ricorsive di Durbin e Levinson per determinare i coefficienti $\phi_{k,j}$, $j = 1, 2, ..., k-1$, e le autocorrelazioni parziali $\pi(k) = \phi_{k,k}$ (si veda Brockwell e Davies, 1987, pag.162 e segg.).

Inoltre, come già anticipato in precedenza per qualsiasi processo stazionario:
i) il *previsore lineare un passo avanti* per Z_t, condizionato alla informazione disponibile I_{t-1}, è espresso da:

$$F_{t-1,1} = E(Z_t|I_{t-1}) = \sum_{j=1}^{t} \phi_{t,j} Z_{t-j}, \quad t \geq 1, \tag{27}$$

che per brevitá indicheremo con m_t;

ii) la *varianza dell'errore di previsione* è espressa da:

$v_t = E(Z_t - F_{t-1,1})^2 = v_0 \prod_{j=1}^{t}(1 - \phi_{j,j}^2)$, per $t \geq 1$.

essendo v_0 la varianza del processo: v_0=var(Z_t)=$\sigma^2(-2d)!/\{(-d)!\}^2$.

Ciò premesso, allora, l'algoritmo di simulazione (Hosking, 1982) effettua le seguenti operazioni:

- genera un numero pseudo casuale, z_0, da una v.c. Normale $Z \sim \mathcal{N}(0, v_0)$;
  - per $t = 1, ..., n$;
    - ▷ calcola iterativamente per $j = 1, ..., t$:
      $\phi_{t,j} = \phi_{t-1,j} - \phi_{t,t}\phi_{t-1,j-1}, \quad j = 1, ..., t-1$;
      $\phi_{t,t} = \frac{d}{(t-d)}$;
      $m_t = \sum_{j=1}^{t} \phi_{t,j}\, z_{t-j}$;
      $v_t = v_{t-1}(1 - \phi_{t,t}^2)$;
    - ▷ genera un numero pseudo casuale, z_t, da una v.c. $\mathcal{N}(m_t, v_t)$.

Tecniche alternative per la simulazione (si veda Bardet et al., 2001, per una rassegna aggiornata) sono state elaborate sfruttando, ad esempio, la Fast Fourier Transform (Janaceck, 1993; Davies and Harte, 1987) e l'approssimazione di un processo *ARFIMA(0,d,0)* quale aggregazione di processi $AR(1)$, come proposto originariamente da Granger (1980).

L'interesse dell'idrologo per la generazione di serie sintetiche è in genere giustificato dalla necessità di studiare scenari alternativi che possono verificarsi in relazione alla dinamica di un particolare corso d'acqua, per finalità di controllo e di previsione. Ciò consente, ad esempio, di simulare la successione di livelli che esondano da un corso d'acqua e valutare così l'impatto territoriale delle fasi di piena, di gestire il deflusso delle acque da piccole dighe, di determinare i parametri di progettazione di grandi serbatoi artificiali, e così via.

In tale contesto, l'applicazione di un mero algoritmo di simulazione -come quello appena descritto- potrebbe talora non garantire risultati adeguati. La simulazione difatti va condotta sostituendo nell'algoritmo prece-

dente i parametri stimati sulla base di una specifica serie osservata. Bisogna però considerare che anche piccoli scostamenti dall'assunzione di Normalità possono in fase di simulazione produrre effetti considerevoli, per esempio in relazione alla valutazione dei massimi di una serie storica. Pertanto, una soluzione più efficace, rispetto all'utilizzo finale, potrebbe essere rappresentata dall'utilizzo di una tecnica bootstrap.

La serie sintetica viene generata come una replicazione bootstrap sfruttando il modello stimato per la realizzazione osservata $z_t,\ t = 1, ..., n$.

In particolare, indichiamo con $\widehat{a}_t$, $t = L + 1, ...n$ i residui ottenuti da tale modello e con $\widehat{\psi}_k$ i pesi della corrispondente formulazione $MA(\infty)$.[3]

La generazione di una serie sintetica viene effettuata in base alla seguente procedura:

i) si costruisce una realizzazione bootstrap dei residui che indicheremo con $\{\widehat{a}_t^*,\ t = -L + 1, ..., n\}$ estraendo con ripetizione elementi dalla successione di residui stimati $\{\widehat{a}_t\}$;

ii) si ricostruisce la serie sintetica mediante la relazione:

$$Z_t^* = a_t^* + \sum_{k=1}^{L} \widehat{\psi}_k\, a_{t-k}^*, \qquad t = 1, ..., n, \tag{28}$$

Nelle Figure 3-4 abbiamo fornito alcune esemplificazioni di serie sintetiche generate con tale tecnica e, per ciascuna di esse, abbiamo riportato la corrispondente stima delle funzioni di autocorrelazione globale. Le evidenze empiriche sui dati generati ed i corrispondenti comportamenti delle funzioni di autocorrelazione stimate mostrano che la procedura rispetta sostanzialmente sia la dinamica temporale del fenomeno che la struttura interna di correlazione seriale implicata e richiesta per un processo a lunga memoria. Per verificare la coerenza statistica della procedura, nel caso esaminato, abbiamo simulato 1000 serie ed utilizzato il test di Kolmogorov-Smirnov cosí come descritto da Anderson(1993) per verificare l'ipotesi $H_0 : \tilde{f}(\omega) = \tilde{f}_0(\omega)$, essendo essendo $\tilde{f}(.)$ la densitá spettrale normalizzata di un processo Gaussiano.

[3]Ricordiamo che, nelle applicazioni di tipo idrologico, la numerosità dei dati è sempre considerevole; pertanto, l'effetto dei valori iniziali che si perdono con tale troncamento è a tutti gli effetti trascurabile.

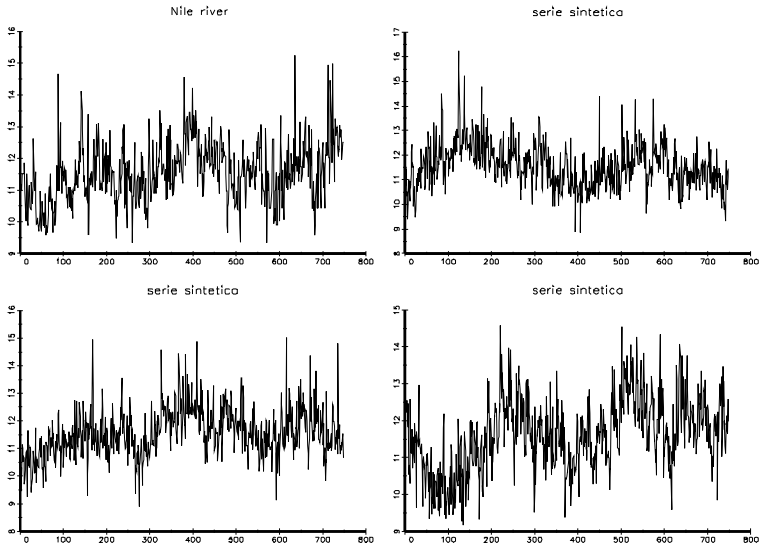

Figura 3. Serie storica del livello minimo del Nilo e serie sintetiche

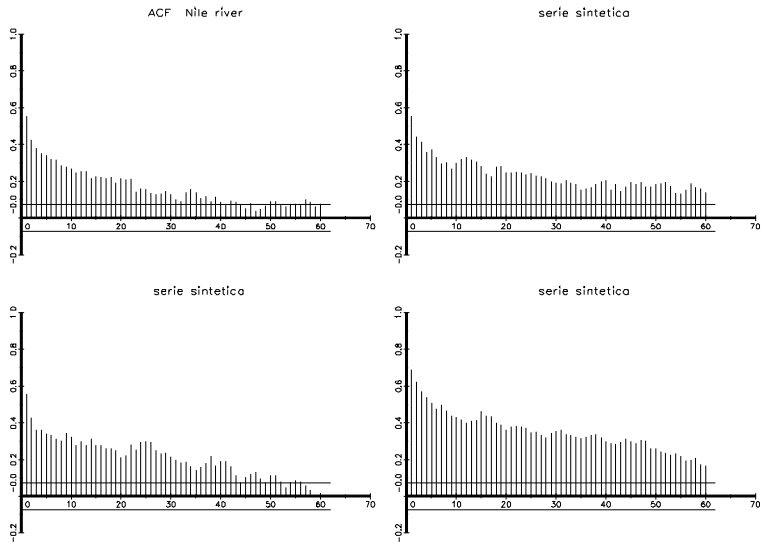

Figura 4. Funzioni di autocorrelazione globale della serie storica del livello minimo del Nilo e di alcune serie sintetiche

In particolare, si è assunto il processo *ARFIMA(0,d,0)* con *d* pari al valore stimato sulla serie osservata come processo generatore. Ebbene, nel caso esaminato, su 1000 serie simulate 84 vengono rifiutate perché non conformi all'ipotesi nulla. Stante l'approssimazione indotta dalla procedura di troncamento si ritiene accetabile tale risultato.

6. Un'applicazione: le portate giornaliere del Tevere

In questo paragrafo presentiamo il modello statistico per le portate giornaliere del fiume Tevere (espresse in m^3/sec). La serie storica, rilevata a Roma Ripetta, costituisce un importante data set di osservazioni giornaliere le cui prime misurazioni risalgono al 1180 (per una disamina storica, si veda: Margaritora e Magnaldi, 2000). Dal 1782, l'abate Calandrelli (direttore dell'Osservatorio Metereologico del Collegio Romano) iniziò a registrare con regolarità il livello del fiume in località Roma Ripetta e, dal 1822 iniziarono le osservazioni giornaliere. Dati omogenei, con frequenza giornaliera, come quelli di cui si tratta in questo lavoro, sono disponibili negli annali idrologici a partire dal 1921.

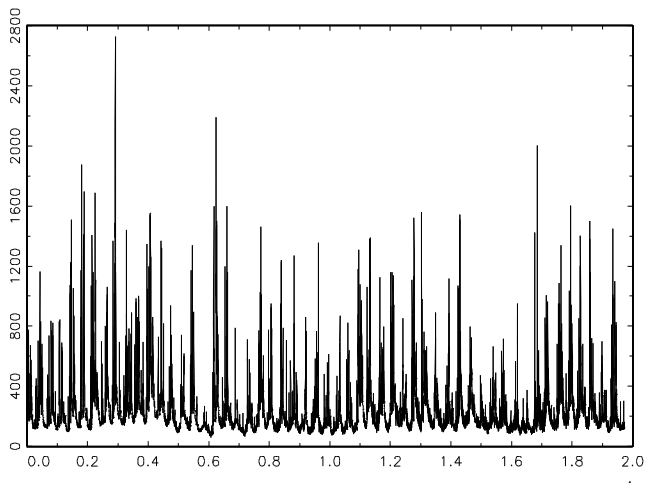

Figura 5. Serie storica delle portate giornaliere del Tevere (rilevate a Roma Ripetta), dal 1/1/1930 al 31/12/1983

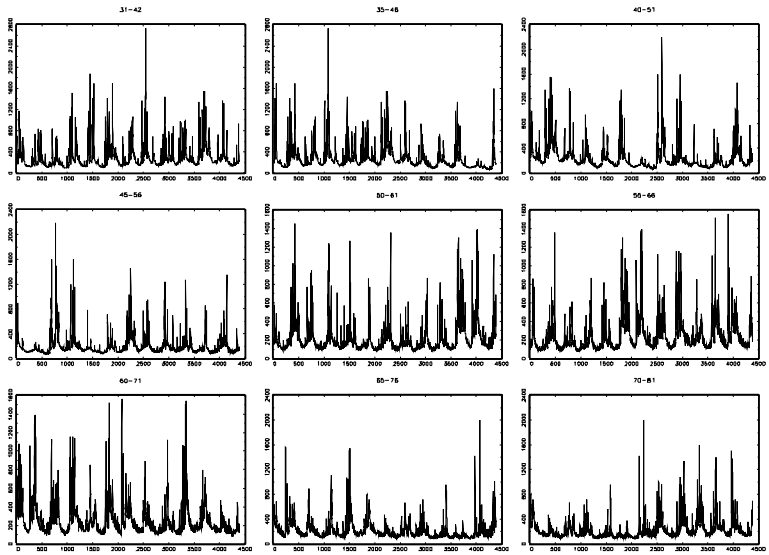

Figura 6. Portate giornaliere del Tevere suddivise in 9 sottoserie

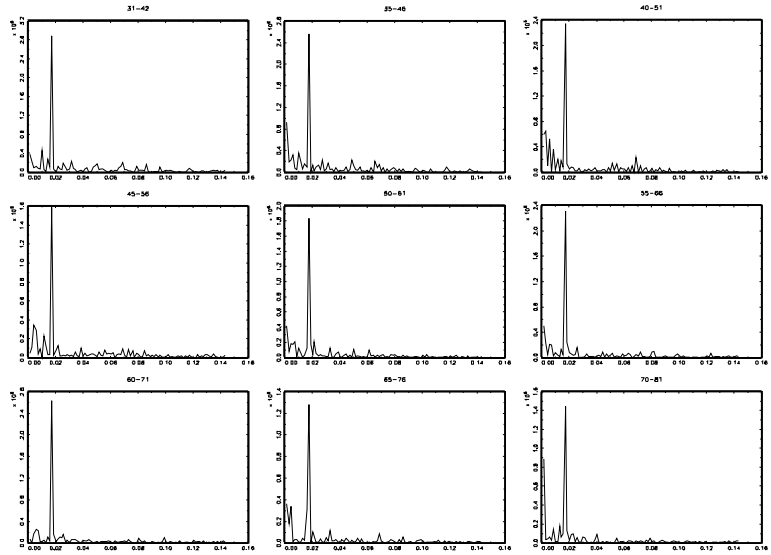

Figura 7. Periodogramma delle 9 sotto-serie

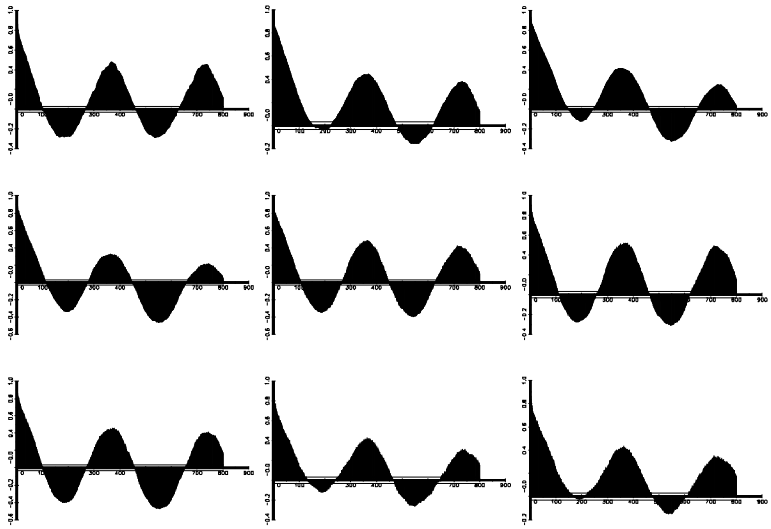

Figura 8. Stima delle funzioni di autocorrelazione globale delle 9 sotto-serie

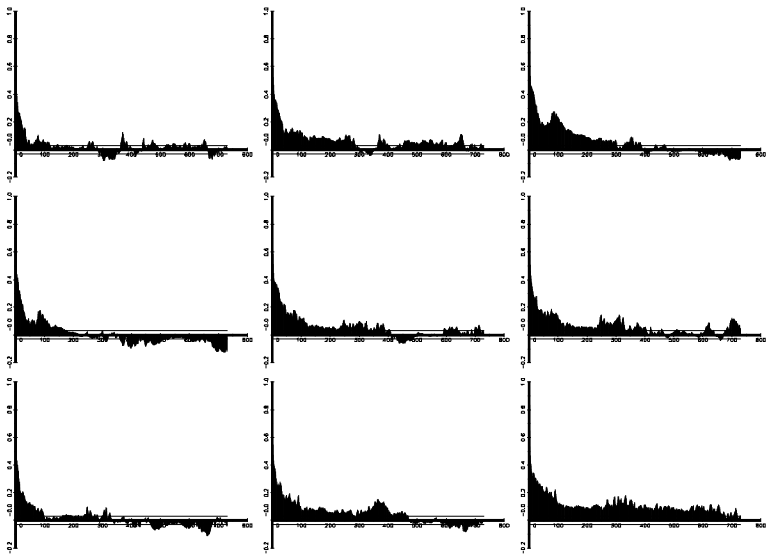

Figura 9. Stima delle funzioni di autocorrelazione dei residui dal modello di regressione per le 9 sotto-serie

Nel seguito, a parte la Figura 5 - che considera il periodo 1930-1983- esamineremo la serie storica (e le sottoserie) delle portate giornaliere del fiume Tevere dal 1/1/1931 al 31/12/1981. I grafici delle serie e delle funzioni corrispondenti confermano una sostanziale omogeneità di comportamento nei vari sottoperiodi, pur essendo essi caratterizzati, talora, dall'alternanza di periodi che si allontanano anche significativamente dalla portata media del fiume.

Emerge, in ogni caso, che la serie storica che stiamo considerando (sia nella sua interezza che nelle sottoserie prima individuate) è la risultante di alcune componenti:

i) il livello medio della serie;

ii) una componente stagionale sufficientemente regolare dovuta alla alternanza delle stagioni e, quindi, alla piovosità che alimenta il bacino idrografico;

iii) elementi di breve memoria ed occasionali dovuti alla dinamica propria di tali fenomeni che inducono correlazione seriale ai primi lags;

iv) una componente di lunga memoria tipica di tali fenomeni e già ben evidente dai grafici precedenti.

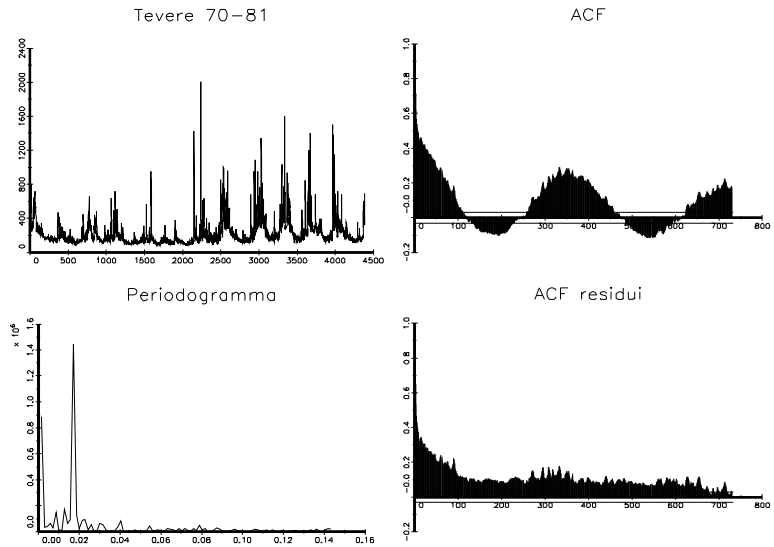

Figura 10. Portate giornaliere del Tevere dal 1/1/1970 al 31/12/1981

Per motivi pratici, nel seguito, concentreremo l'attenzione sulla ultima delle sotto-serie relativamente ai dati giornalieri compresi negli anni

1970-1981. Tali dati sono riportati nella Figura 10, assieme alla stima della funzione di autocorrelazione globale e al periodogramma.

Per eliminare la componente periodica, che risulta essere praticamente deterministica con periodo pari a $P = 365.25$ giorni (per la presenza dell'anno bisestile ogni quattro anni), posto X_t il processo originario, si deriva il seguente modello di regressione:

$$x_t - 224.45 = \\ = 65.088\, cos\left(\frac{2\pi t}{365.25}\right) + 63.518\, sen\left(\frac{2\pi t}{365.25}\right) + \hat{z}_t. \qquad (29)$$

Tutte le successive analisi relative alla modellistica *ARFIMA* saranno condotte sulla serie storica "residua" $\hat{z}_t$, che è depurata dalla componente stagionale deterministica[4]. Le stime dei parametri del modello (29) risultano tutte significative; la componente stagionale spiega circa il 20% della variabilità originaria dei dati, essendo, per il modello di regressione descritto, $R^2 = 0.19152$.

Tenuto conto delle componenti inerziali e di una componente di lunga memoria, non necessitando più della costante (perchè trattasi di una serie residua dal modello di regressione), la fase di specificazione conduce, per la serie $\hat{z}_t$, ad un modello *ARFIMA(2,d,0)* la cui stima risulta essere la seguente:

$$\left(1 - 0.607B + 0.172B^2\right) \nabla^{0.359}\, \hat{z}_t = a_t \qquad (30)$$

I parametri sono tutti significativi e l'onda pseudo-periodica indotta dalla componente $AR(2)$ con radici complesse e coniugate deriva da una struttura di correlazione seriale[5] che possiede un periodo stocastico stimato approssimativamente in: $2\pi / arc\, cos\left\{\frac{0.607}{2\sqrt{0.172}}\right\} = 8.379$. Questo ciclo, non attribuibile certamente a fattori metereologici, potrebbe forse

[4]Per un approccio modellistico alternativo della componente stagionale, si veda Grimaldi (2001).

[5]Si noti che il modello AR(2) stimato presenta parametri collocati in quella particolare regione dello spazio parametrico ammissibile ove, pur esistendo una struttura di correlazione seriale assimilabile ad una sinusoidale smorzata (con il periodo stocastico calcolato nel testo), non esiste un accumulo di varianza attorno ad una specifica frequenza angolare, perchè lo spettro corrispondente a quei parametri non ammette sul campo di variazione $(-\pi, +\pi]$ un massimo differente da $\omega = 0$. Per una discussione di queste problematiche, si veda Piccolo (1990).

derivare da sistemi di regolazione degli afflussi del Tevere verso la capitale che tengono contono dell'apertura e della chiusura di sbarramenti a valle e di fenomeni di prelievo per irrigazione che, molto approssimativamente e con notevole variabilitá, si possono far derivare da ritmi quasi settimanali.

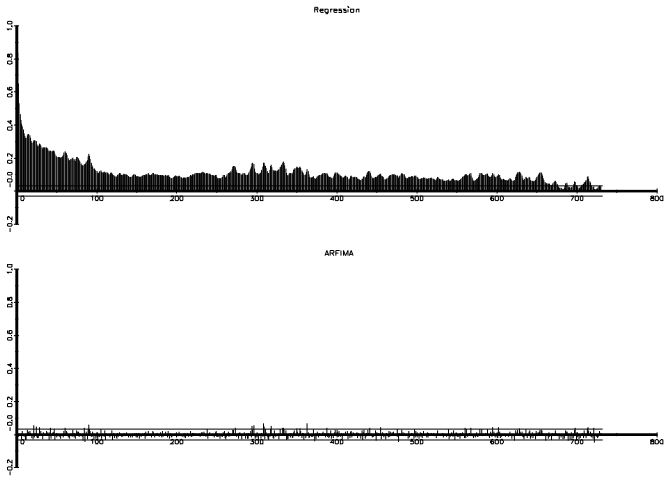

Figura 11. Stima delle funzioni di autocorrelazione globali della serie $\hat{z}_t$ e dei residui stimati dal modello ARFIMA

La tabella seguente presenta la stima dei parametri e l'errore standard asintotico a conferma della elevata significatività dei risultati ottenuti:

Parametri	Stima	Errore standard
β_1	65.088	2.8247
β_2	63.518	2.8247
ϕ_1	0.607	0.0296
ϕ_2	-0.172	0.0149
d	0.359	0.0271

Inoltre, riportiamo di seguito le varianze della serie originaria x_t, della serie $\hat{z}_t$ al netto della componente stagionale deterministica e della serie $\hat{a}_t$ dei residui stimati dal modello *ARFIMA*.

Serie	Varianza	Riduzione% progressiva	Riduzione % complessiva
x_t	21618.79	===	===
$\widehat{z}_t$	17478.36	19.152 %	19.152 %
$\widehat{a}_t$	4839.57	72.311 %	77.614 %

La riduzione complessiva in varianza (pari a circa l'80 %) per una serie storica senza trend, ottenuta mediante la stima di soli 5 parametri espliciti, va giudicata importante. Soprattutto, la dinamica dei residui stimati (così come sintetizzata dalla stima della funzione di autocorrelazione, di cui al secondo grafico della Figura 11) è coerente con le realizzazioni tipiche dei processi WN, indicando un buon adattamento complessivo del modello alle realizzazioni.

7. Considerazioni finali

Il lavoro ha discusso sul piano dei processi e degli aspetti inferenziali le principali questioni poste dalla modellistica per serie idrologiche, per le quali è presente, spesso con un ruolo rilevante, una componente di lunga memoria. Nell'analisi empirica della serie delle portate del Tevere è emerso, poi, come la modellistica debba includere anche una componente deterministica stagionale, che contribuisce in maniera rilevante alla spiegazione e previsione di tale fenomeno.

Ringraziamenti: Il lavoro è stato parzialmente finanziato dal MIUR (Progetto 2001-2002) e CNR. Una versione preliminare è stata presentata alla Scuola della Società Italiana di Statistica su "Modelli lineari e non lineari, reti neurali e algoritmi genetici per l'analisi e la previsione", svoltasi a Treviso dal 10 al 15 giugno 2002. Si ringrazia l'ing. Salvatore Grimaldi del CNR-IRPI di Perugia per aver fornito i dati e le informazioni relative allo studio del bacino del Tevere.

Riferimenti bibliografici

Anderson T.W. (1993) Goodness of Fit Tests for Spectral Distributions, *Annals of Statistics*, 21, 830-847.

Bardet J.M., Lang G., Oppenheim G., Philippe A., Taqqu M.S. (2001) Generators of Long Range Dependent Processes: a survey, in: Taqqu M.S., Oppenheim G., Doukhan P. Surgailis, D., Doukhan, P., Lang, G., editors (2001), *Asymptotics of empirical processes of linear random fields with long-range dependence*, preprint, Birkhaüser.

Beran J. (1995) *Statistics for Long Memory Processes*, Chapman and Hall, New York.

Bertelli S., Caporin M. (2002) Autocovariances of Long-Memory Processes, *Journal of Time Series Analysis*, 23, 503-508.

Bhansali R.J., Kokoszka P.S. (2002) Computation of the Forecast Coefficients for Multistep Prediction of Long-range Dependent Time Series, *International Journal of Forecasting*, 18, 181-206.

Box G.E.P., Jenkins G.M. (1970) *Time Series Analysis: Forecasting and Control*, (Revised edition, 1976), Holden-Day, San Francisco.

Brockwell P.J., Davies R.A. (1987) *Time Series: Theory and Methods*, Springer-Verlag, New York.

Cancelliere A., Salas J.D., Boes D.C. (2001) Maximum Likelihood of Low Order fractionally Integrated Autoregressive Moving Average Models, *in: Piccolo and Ubertini, editors (2001)*, 21-32.

Cheung Y. W. (1993) Long Memory in Foreign Exchange Rates, *Journal of Business and Economic Statistics*, 11, 93-101.

Corduas M., Piccolo D. (2001) Fractional Differencing Model Estimation: some new approaches, *in: Piccolo e Ubertini*, editors, 73-79.

Davies R.B., Harte D.S. (1987) Test for Hurst Effect, *Biometrika*, 74, 95-102.

De Giovanni L. (2002) Long Memory Statistical Models in Telecommunication Networks, *Atti XLI Riunione Scientifica SIS Milano-Bicocca*, CLEUP, Padova, vol. I, 167-178.

D'Elia A., Piccolo D. (2002) The Effects of Instantaneous Transformations on the Fractional Differencing Parameter, *Atti XLI Riunione Scientifica SIS Milano-Bicocca*, CLEUP, Padova, vol. II, 409-412.

Delgado M.A., Robinson P.M. (1994) New Methods for the Analysis of Long-Memory Time Series: Application to Spanish Inflation, *Journal of Forecasting*, 13, 97-107.

Diebold F.X., Rudenbusch G.D.(1989) Long Memory and Persistence in Aggregate Output, *Journal of Monetary Economics*, 24, 189-209.

Diebold F.X., Rudenbusch G.D.(1991) Is consumption too smooth? Long memory and the Deaton Paradox, *The Review of Economics and Statistics*, 73, 1-9.

Doornick J.A., Ooms M. (2001) Computational Aspects of Maximum Likelihood Estimation of Autoregressivve Fractionally Integrated Moving Average Models, *Nuffield College Economic Working Papers*, 2001-W27, in corso di pubblicazione in *Computational Statistics and Data Analysis*.

Fox R., Taqqu M.S. (1986) Large Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series, *Annals of Statistics*, 14, 517-532.

Geweke J., Porter-Hudak S. (1983) The Estimation and Application of Long-Memory Time Series, *Journal of Time Series Analysis*, 4, 231-238.

Granger C.W.J. (1980) Long Memory Relationships and the Aggregation of Dynamics Models, *Econometrica*, 34, 150-161.

Granger C.W.J., Joyeaux J. (1980) An Introduction to Long Memory Time Series Models and Fractional Differencing, *Journal of Time Series Analysis*, 1, 15-30.

Grimaldi S. (2001) Modelli parametrici lineari per serie idrologiche giornaliere, *Quaderni di Statistica*, 3, 83-105.

Hassler U., Wolters J. (1995) Long Memory in Inflation Rates: International Evidences, *Journal of Business and Economic Statistics*, 13, 37-45.

Hosking J.R.M. (1981) Fractional Differencing, *Biometrika*, 68, 165-176.

Hosking J.R.M. (1982) Some Models on Persistence in Time Series, in: *Time Series Analysis: Theory and Practice*, (O.D. Anderson ed.), vol. 1, North Holland, 641-653.

Hosking J.R.M. (1984) Modeling Persistence in Hydrological Time Series Using Fractional Differencing, *Water Resources Research*, 20, 1898-1908.

Hurvich C.M., Beltrao K.I. (1993) Asymptotics for the Low-frequency Ordinates of the Periodogram of Long Memory Time Series, *Journal of Time Series Analysis*, 5, 455-472.

Hurvich C.M., Deo R., Brodsky J. (1998) The Mean Squared Eroor of Geweke and Porter-Hudak's Estimator of the Memory Parameter of a

Long-memory Time Series, *Journal of Time Series Analysis*, 19, 19-46.

Janaceck G.J. (1993) A Spectral Approach to Long Memory Time Series, in: Subba Rao T. editor (1993), 164-179.

Lawrance A.J., Kottegoda N.T. (1977) Stochastic Modelling of River-flow Time Series, *Journal of the Royal Statistical Society*, A, 140, 1-47.

Marinucci D. (2002) Some Aspects of Statistical Inference for Long Memory Processes, *Atti XLI Riunione Scientifica SIS Milano-Bicocca*, CLEUP, Padova, vol. I, 157-166.

Margaritora G., Magnaldi S. (2000) Il Tevere nella Storia di Roma, in: *Conferenze e Seminari dell'area di Costruzioni Idrauliche e Marittime*, D.I.T.S., Università di Roma "La Sapienza", 147-163.

Piccolo D. (1990) *Introduzione all'Analisi delle Serie Storiche*, La Nuova Italia Scientifica, Roma.

Piccolo D., Ubertini L. editors (2001) *Metodi statistici e matematici per l'analisi delle serie idrologiche*, CNR-GNDCI, n.2136, Roma.

Ray B.K. (1993) Long Range Forecasting of IBM Product Revenue using a Seasonal Fractionally Integrated Differenced ARMA Model, *International Journal of Forecasting*, 9, 255-269.

Robinson P.M. (1995) Log-periodogram Regression of Time Series with Long Range Dependence, *Annals of Statistics*, 23, 1048-1072.

Shea G.S. (1991) Uncertainity and Implied Variance Bounds in Long Memory Models of Interest Rate Terms in Structure.*Empirical Economics*, 16, 287-312.

Sowell F. (1992) Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models, *Journal of Econometrics*, 53, 165-188.

Subba Rao T. editor (1993) *Developments in Time Series Analysis*, Chapman and Hall, London.

Tousson O. (1925) Mèmoire sur l'Histoire du Nil, *Mèmoires de L' Institute D'Egypte*, 366-385.

Zaffaroni P. (2002) The Long Range Dependence Paradigm for Finance, *Atti XLI Riunione Scientifica SIS Milano-Bicocca*, CLEUP, Padova, vol. I, 179-188.