Applicazione della Metodologia Box-Jenkins per la previsione della progettazione di edilizia non residenziale(*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di Economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Riassunto.
La metodologia Box-Jenkins è applicata alla serie storica trimestrale della progettazione di edilizia non residenziale. Pur nell'ambito ristretto determinato dal ridotto di osservazioni si ottiene un modello ragionevole di previsione. I suoi risultati, vanno certo usati con molta cautela, ma sul piano dei tests statistici non presenta limiti di rilievo

Keywords: serie storiche, stagionalità, trasformazione Box-Cox

Lavoro inserito come Appendice B al capitolo IV del libro "Investimenti in costruzioni e mercato del lavoro in Emila-Romagna. Un modello per la previsione degli effetti del piano decennale. Quaderni del Cress n.2,1979, pp. 151-157, Marsilio Editori, Venezia.

APPENDICE B

APPLICAZIONE DELLA METODOLOGIA BOX-JENKINS PER LA PREVISIONE DELLA PROGETTAZIONE DI EDILIZIA NON RESIDENZIALE

1. I modelli ARIMA stagionali

La formula generale per un modello ARIMA *(Autoregressive Integrated Moving Average)* viene espressa con la notazione

$$[1] \qquad \Phi_p(B)\, \Phi_P(B^s)\, (1 - B)^d (1 - B^{sD}) W_t = \theta_o + \Theta_q(B)\, \Psi_Q(B^s)\, a_t$$

dove *B* è l'operatore retroattivo definito da $B_j X_t = X_{t-j}$; *s* è la periodicità della serie W_t (per modelli trimestrali s = 4); $\Phi_p(B)$, $\Phi_p(B^s)$, $\Theta_q(B)$, $\Psi_q(B^s)$ sono polinomi di grado p, P, q, Q negli argomenti *B*, B^s, *B*, B^s rispettivamente; d e D sono degli interi non negativi che indicano il grado di differenziazione richiesto per indurre la stazionarietà nelle W_t. In particolare, si assume che le radici dei due polinomi

$$[2] \qquad \Phi_p(B)\, \Phi_P(B^s)\, (1 - B)^d (1 - B^{sD}) = 0$$

$$[3] \qquad \Theta_q(B)\, \Psi_Q(B^s) = 0$$

abbiamo radici al di fuori del circolo unitario. Questo assicura che il modello [1] soddisfi le condizioni di stazionarietà nella componente autoregressiva e le condizioni di invertibilità nella componente *Moving Average*. θ_o è una costante da stimare che rappresenta la possibile presenza di uno *shift* sul tempo. Le a_t sono degli *shocks* indipendenti e identicamente distribuiti secondo una legge normale di media nulla e con varianza costante. Le W_t sono legate alle osservazioni originali Z_t dalla funzione

$$[4] \qquad W_t = \begin{cases} Z_t^{\lambda} & se \;\; \lambda \neq 0 \\ Ln(Z_t) & se \;\; \lambda = 0 \end{cases}$$

I modelli della classe [1] consentono da un Iato di ottenere delle previsioni accurate senza necessariamente ricorrere a un esame approfondito della teoria del fenomeno analizzato e d'altra parte possono essere impiegati in situazioni in cui non si hanno dati sufficienti o adeguati per adattare loro, secondo le tecniche econometriche classiche, un modello teorico fondato su un insieme di assunzioni.

L'ottenimento delle previsioni avviene alla fine di una procedura iterativa che non è molto agile, particolarmente nei modelli stagionali: si ha una prima fase di determinazione del modello più adeguato fra quelli della classe [1] (questa fase viene detta di identificazione). Segue poi una fase di stima dei parametri che il modello contiene e una fase di controllo in cui si effettua la verifica dell'adeguatezza del modello prescelto alla rappresentazione dei dati osservati per la Z_t. Solo se il modello supera questa fase si passerà alle previsioni.

2. Una classe ristretta di modelli

La costruzione di modelli della classe [1], a partire dalla serie delle progettazioni di edilizia non residenziale APNR, richiederebbe un numero di osservazioni almeno doppio di quello effettivamente disponibile. È indispensabile quindi procedere a delle semplificazioni in modo da sfoltire drasticamente la classe di modelli su cui effettivamente operare la scelta, badando di conservare solo quelli più semplici.

In primo luogo, data la presenza di una marcata stagionalità trimestrale (in entrambe le serie l'autocorrelazione di *lag 4* per le Z_t è superiore a 0.5) si porrà subito s = 4. Per quanto attiene alle differenziazioni decidiamo di utilizzarne solo una di tipo stagionale $(D = 1)$ in modo da eliminare il trend crescente che è presente fra trimestri dello stesso anno. I valori ammessi per d sono invece $d = 0, 1, 2$ dato che non appaiono forti tendenze crescenti o decrescenti fra trimestri di anni diversi.

Sintetizziamo queste limitazioni con la disuguaglianza

$$d + D \leq 3 \tag{5}$$

Data la [5] rimangono effettivamente disponibili per la stima da 31 a 33 osservazioni ed a queste possiamo ragionevolmente richiedere la stima di almeno due parametri. Poniamo quindi il vincolo

$$\delta + p + P + q + Q \leq 2 \tag{6}$$

dove $\delta = 1$ se θ_o è incluso fra i parametri da stimare e $\delta = 0$ se $\theta_o = 0$. Dato che λ non potrà essere stimato dai dati esso non sarà quel valore che rende minima la somma dei quadrati dei residui prodotti dal modello già identificato e stimato. Piuttosto sarà scelto nell'insieme

[7] $$\lambda \in \{0.0; 0.25; 0.50; 0.75, 1.00\}$$

La classe ridotta di modelli con cui si cerca di rappresentare i dati delle due serie sarà

[8] $$(1-\phi B)(1-\Phi B^4)(1-B)^d(1-B)(1-B^4)W_t = (1-\theta B)(1-\tau B^4)a_t$$

3. Identificazione

In questa fase i punti di riferimento più affidabili sono:
Le autocorrelazioni semplici date da

[9] $$r_k = \frac{\sum_{j=k+1}^{n}\left(W_j - \bar{w}\right)\left(W_{t-j} - \bar{w}\right)}{\sum_{j=1}^{n}\left(W_j - \bar{w}\right)}, \quad k = 1,2,\ldots,K$$

con $K = n/3$ ed $n = N - d - sD$. Le autocorrelazioni parziali, ovvero i parametri $r_k^{'}$ nel modello di regressione lineare

[10] $$W_t = \sum_{j=1}^{K} r_k^{'} W_{t-k}, \quad k = 1,2,\ldots,K$$

La statistica Q definita da

[11] $$Q = n(n+2)\sum_{k=1}^{K}\frac{r_k^2}{(n-k)}$$

La Q per una serie che verifica le condizioni poste sulle a_t, ha una distribuzione asintotica χ^2 con $(k - 1)$ gradi di libertà. Il suo valore ci fornirà in fase di identificazione, indicazioni sia sul grado di stazionarietà raggiunto da (1 - B) $(1 - B^4)W_t$ sia su quanto questa differisce da una serie di *random shocks*.

La tabella che segue ci consente di determinare λ, e *d*

Tabella IV.B1. Valori della statistica Q.

λ	d	D	n	APNR
1.00	0	1	33	114.108
0.75	0	1	33	112.743
0.50	0	1	33	117.941
0.25	0	1	33	122.110
0.00	0	1	33	124.863
1.00	1	1	32	47.721
0.75	1	1	32	46.908
0.50	1	1	32	44.660
0.25	1	1	32	45.113
0.00	1	1	32	46.416
1.00	2	1	31	55.113
0.75	2	1	31	57.319
0.50	2	1	31	56.826
0.25	2	1	31	56.903
0.00	2	1	31	61.170

I valori più piccoli di *Q* si registrano per il valore *d = 1* segno che nelle due serie originali vi era un *trend* che va eliminato per arrivare alla rappresentazione con un modello più stabile. Il valore di λ non risulta particolarmente discriminante anche se si può escludere il valore di λ è di 0.75.

Nella tabella che segue riportiamo i valori delle autocorrelazioni semplici e parziali relativi ai valori più interessanti di λ.

Tabella IV.B.2. Autocorrelazioni semplici e parziali

λ= 0.50		λ= 0.25		λ= 0.00	
r	r'	r	r'	r	r'
-0.297	-0.297	-0.280	-0.280	-0.277	-0.277
0.134	0.050	0.144	0.071	0.166	0.097
0.113	0.182	0.197	0.169	0.074	0.156
-0.232	-0.180	-0.212	-0.175	-0.196	-0.175
-0.016	-0.186	-0.031	-0.188	-0.037	-0.188
-0.056	-0.093	-0.081	-0.117	-0.108	-0.130
-0.088	-0.104	-0.042	-0.013	-0.010	0.013
-0.133	-0.194	-0.168	-0.209	-0.207	-0.226
0.037	-0.084	0.092	-0.038	0.118	-0.029
-0.087	-0.098	-0.103	-0.091	-0.120	-0.096
-0.042	0.117	-0.163	-0.281	-0.177	-0.314

L'applicazione a W_t dell'operatore *(1 - B) (1 - B^4)* ha prodotto una cancellazione sostanziale della memoria dei valori passati W_{t-1}, W_{t-2} etc. Non si hanno infatti valori degli r e degli r' particolarmente elevati ed anche le statistiche Q sono relativamente basse. Applicheremo dunque alla serie il modello

[12] $$\tilde{W}_t = (1 - \theta B)(1 - \tau B^4)a_t$$

dove $\tilde{W}_t = (1 - B)(1 - B^4)$.

La funzione di autocorrelazione per il modello [12] prevede le relazioni

[13] $$\rho_1 = -\frac{\theta}{\left(1+\theta^2\right)}; \quad \rho_2 = 0; \quad \rho_4 = \frac{-\tau}{1+\tau^2}$$
$$\rho_3 = \rho_5 = \rho_1 * \rho_2; \quad \rho_k = 0, \ k \geq 6$$

In nessuna delle colonne della Tab. Iv.B, 2 relative alle autocorrelazioni semplici si verifica una struttura molto simile alla [13]. Lo scostamento non è comunque così accentuato come potrebbe sembrare a prima vista In particolare la dissomiglianza più fastidiosa è la diversità non solo in valore, ma anche in segno delle autocorrelazioni di *lag 3* e di *lag 5*. Si può comunque attribuire il *bad behavior* di r_5 a fluttuazioni campionarie il che è molto verosimile dato che r_5 viene calcolato solo su sei termini. Alla stessa ragione sono da attribuire autocorrelazioni alla soglia della significatività per *lag* maggiori di 5.

4. Stima e verifica

Una volta identificato il modello che sembra più adatto il passo successivo è di ottenere le stime di massima verosimiglianza per i parametri θ e τ. La determinazione di queste stime richiede una procedura non-lineare che non può essere effettuata senza l'ausilio del computer.

L'applicazione corretta della metodologia Box Jenkins è molto legata alla disponibilità di un adeguato package di programmi che consenta in effetti di mettere in pratica la teoria sviluppata. La disponibilità non è ancora abbastanza generalizzata come si potrebbe sperare. Fortunatamente per questa applicazione è stato possibile utilizzare il package prodotto al Madison Academic Computing Center.

Tabella IV.B.3: Stima dei parametri per la serie delle progettate non residenziali

Parametri	Valori iniz.	Valori finali	Lim. Inf.	Lim. Sup.
θ	0.30	0.254	-0.1003	0.6082
τ	0.20	0.386	0.0427	0.7302
Somma dei residui quadrati		3.0548		
Mean Square Errore		0.1018		
Error Std. dei residui		0.3191		
Correlazione fra θ e τ		0.1126		
Statistica Q sui residui		9.357		
Max $\lvert r_k \rvert$ nei residui		0.20		

I controlli solitamente effettuati sui residui non rivelano particolari carenze per il modello che si è stimato. In particolare nessuna delle autocorrelazioni nei residui è risultata significativa e la statistica *Q* conferma questa ipotesi se valutata di fianco al valore dato sulle tavole della distribuzione χ^2 con 11 gradi di libertà.

5. Previsioni

Le previsioni basate sul modello [12] per le due serie sono state calcolate per 12 periodi in avanti a partire dal primo trimestre del 1979 e utilizzate per la previsione della produzione edilizia non residenziale della prima colonna della Tav. IV, 12.

Trimestre	1979.1	1979.2	1979.3	1979.4
m^3 di Pr.Ed.NR.	2318.11	2298.43	2509.68	2744.17
	1980.1	1980.2	1980.3	1980.4
	2820.15	2779.27	2944.55	3132.56
	1981.1	1981.2	1981.3	1981.4
	3161.46	3068.14	2459.73	2496.65

Più che approfondire l'esame delle previsioni ottenute che del resto avrebbe poco da aggiungere dato che l'intera tecnica Box-Jenkins è stata applicata solo su 37 osservazioni si può concludere dicendo che le previsioni ottenute sono affidabili solo se usate come informazioni su cui poggiare una qualche altra tecnica econometrica meno vincolata dal numero di osservazioni.

.