NOTA METODOLOGICA[1]

1. Le conseguenze del cambiamento della normativa di dichiarazione delle nascite

La legge 127/97 del 17/05/1997, detta comunemente "Bassanini-bis" o sulla "semplificazione amministrativa", stabilisce nuove modalità in materia di dichiarazione di nascita. La dichiarazione di nascita (solo per i nati vivi) può essere resa non più esclusivamente all'ufficiale di Stato Civile del Comune nel cui territorio si è verificato l'evento, bensì anche – in alternativa – all'ufficiale di Stato Civile del Comune di residenza dei genitori (se i genitori risiedono in Comuni diversi e in mancanza di accordo, nel Comune di residenza della madre) oppure direttamente presso la direzione sanitaria del centro di nascita presso il quale la nascita è avvenuta. In quest'ultimo caso la dichiarazione di nascita viene trasmessa o al Comune nel cui territorio è situato il centro di nascita o, su richiesta dei genitori, al Comune di residenza.

L'entrata in vigore immediata, a metà 1997, della normativa ha determinato una situazione di improvvisa inadeguatezza dell'apparato di generazione del dato elementare relativo agli eventi di nascita. A causa della possibilità di dichiarare la nascita non solo nel luogo di evento, ma anche presso il Comune di residenza, la rilevazione, prima concentrata nei Comuni contenenti i centri di nascita (all'incirca 600 in tutto il territorio), si è estesa a tutti gli 8100 Comuni italiani. Dal mese di giugno 1997, quindi, anche i Comuni che non ospitavano un centro di nascita si sono trovati a dover effettuare una rilevazione senza disporre degli appositi modelli. Infatti i modelli utilizzati dagli ufficiali di stato civile venivano forniti dall'Istat secondo un piano di spedizione che teneva conto degli eventi verificatisi negli anni precedenti.

Questa situazione ha comportato un sottodimensionamento delle nascite del 1997 (considerando esclusivamente i nati vivi) dal mese di giugno in poi (con la registrazione di 496.829 eventi contro i 528.103 del 1996).

Un sottodimensionamento del numero dei nati vivi si è manifestato anche nel 1998 (gli eventi registrati sono 479.463), nonostante l'adeguamento del modello di rilevazione e la rivisitazione del piano di spedizione dei modelli ai Comuni.

Sono state eseguite alcune analisi per verificare se, oltre al numero assoluto di eventi, risultassero intaccate anche le strutture secondo le variabili presenti sul modello. Dal confronto con i dati relativi al 1995 e al 1996, le strutture dei nati sono risultate inalterate per quel che riguarda le seguenti variabili:

- Età della madre al parto;
- Età del padre;
- Genere del parto;
- Ordine di nascita;
- Tipo di parto;
- Luogo del parto;
- Filiazione;

Le strutture sono invece risultate alterate per quel che riguarda le seguenti variabili:

- Grado di istruzione della madre;
- Grado di istruzione del padre;
- Posizione professionale della madre;
- Posizione professionale del padre;
- Condizione professionale della madre;
- Condizione professionale della madre;

[1] I testi sono stati redatti da Marco Battaglini (parr. 2.1, 2.2, 2.3, 2.4) e Claudia Iaccarino (parr. 1, 3, 3.1, 3.2)

- Ramo di attività economica della madre;
- Ramo di attività economica del padre.

Per quanto precedentemente esposto si è reso necessario procedere ad una previsione (per regione di nascita) dei nati vivi nel 1997 e nel 1998. Disponendo della serie mensile completa delle nascite a partire dal 1969, la previsione è stata realizzata utilizzando tecniche di analisi delle serie storiche, mediante il pacchetto TRAMO (cfr. par. 2).

2. La metodologia di previsione delle nascite

2.1 I dati

Lo scopo è quello di stimare il numero dei nati in ogni mese e per ogni singola regione, per ovviare alla carenza di dati. Per ottenere questa stima sono state utilizzate le tecniche di analisi delle serie storiche, che spesso si sono dimostrate molto utili per le previsioni a breve termine.
La serie dei nati analizzata parte dal 1969 ed arriva al 1996. Si è scelto di lavorare con i valori assoluti dei nati e di tralasciare gli indici di natalità (nati in complesso/popolazione), perché nelle previsioni è importante conoscere l'ammontare complessivo delle nascite e perché, su dati mensili, appare problematico definire il denominatore, la popolazione di riferimento, in modo univoco e corretto.
La scelta di operare con serie mensili permette di tenere esplicitamente conto, nel processo di previsione, di tutte le componenti sistematiche, ciclo-trend e stagionalità, che determinano l'evoluzione futura del fenomeno. Operando invece con dati annuali è inevitabile restringere l'attenzione alla sola componente ciclo tendenziale e alla sua dinamica inter-annuale, escludendo così a priori la possibilità di analizzare e proiettare il ciclo-trend anche a livello infra-annuale.
Le 341 osservazioni mensili di cui è composta ciascuna serie regionale sono più che sufficienti per poter effettuare una corretta applicazione dei moderni metodi di analisi delle serie storiche, anche nel caso di piccole popolazioni. Con serie di tale lunghezza si riesce a tener conto in maniera soddisfacente sia della stagionalità del fenomeno che del suo ciclo-trend più recente, in modo da poter ottenere poi accurate previsioni, almeno di breve periodo, delle nascite mensili.

2.2 I Modelli ARIMA

La presenza del trend e della stagionalità in una serie comporta, in generale, un legame di dipendenza stocastica non solo tra termini immediatamente successivi, ma anche tra quelli separati da tanti istanti temporali quanti sono quelli del periodo stagionale ed, eventualmente, da multipli di esso. In altri termini se x_t è il valore della serie al tempo *t* ed *s* il periodo stagionale, ci si aspetta un legame non solo tra x_t e $x_{t-1}, x_{t-2}, \ldots\ldots,$ ma anche tra x_t e $x_{t-s}, x_{t-2s}, \ldots\ldots$. I modelli ARIMA stagionali traducono questo concetto ipotizzando che i legami stocastici siano di tipo lineare.
Scritto in formule si ha:

$$\Psi(B)X_t = \Theta(B)Z_t$$

dove:

- $\Psi(B)$ è l'operatore autoregressivo generalizzato, cioè un polinomio in B e B^s avente radici sia fuori che sul cerchio di raggio unitario $(1-B)^d(1-B^s)^D\Phi(B)$;
- B è l'operatore di retrotraslazione: $Bx_t = x_{t-1}$;
- $(1-B)^d$ e $(1-B)^D$ sono rispettivamente le differenze d-me (∇^d) tra i termini successivi, e le differenze D-me (∇_s^D) tra i termini separati da *s* tempi;

- Z_t rappresenta un a successione di variabili casuali indipendenti, normalmente distribuite con media nulla e varianza σ^2.

Se alla serie originaria vengono applicate le differenze finite d-me e successivamente le differenze finite D-me, si ottiene una serie stazionaria della classe autoregressiva-media mobile (ARMA). Quando accade che sia $\Phi(B)$ che $\Theta(B)$ si fattorizzino in polinomi distinti in B e B^s di ordine *p* e *P*, *q* e *Q*, cioè

$$\Phi(B) = \Phi_p(B)\Phi_p(B^s) \text{ e } \Theta(B) = \Theta_q(B)\Theta_q(B^s)$$

il modello è detto moltiplicativo di ordine $(p,d,q)\times(P,D,Q)$.

2.3 I Test di Verifica

Come verifica dei modelli trovati, prima di fare le previsioni è stata fatta un'analisi di adattamento delle previsioni su dati reali: utilizzando la serie per gli anni 1969-1994 si è fatta la previsione per gli anni 1995-1996 ed in seguito si è verificata la bontà delle previsioni con lo studio degli scarti relativi tra osservazioni e stime.
Il software utilizzato per questa analisi è il programma TRAMO, acronimo di "**T**ime Series **R**egression whit **A**RIMA Noise, **M**issing, Observation and **O**utliers". Questo programma è stato creato per poter stimare, prevedere o interpolare delle serie storiche attraverso modelli di tipo ARIMA.
Utilizzando un'istruzione ben precisa il programma è in grado di stimare i parametri di un modello ARIMA (*p,d,q*)×(*P,D,Q*) in maniera automatica per la serie storica in considerazione. Il test che verifica la scelta la scelta del modello con la miglior struttura parametrica è l'indice AIC (Asymptotic Information Criterion) introdotto nel 1973 da Akaike. Dati più modelli per gli stessi dati si preferisce quello che minimizza l'indice AIC, espresso in funzione dei parametri del modello. A tale riguardo occorre tener conto di due componenti opposte: al crescere del numero dei parametri di un determinato modello la varianza dei residui diminuisce (perché migliora l'adattamento), ma aumentano i vincoli imposti dagli stessi parametri, e quindi peggiora la parsimonia. Il test AIC è così definito:

$$AIC(p+1) = n\log(S^2) + 2(p+1).$$

Le serie di tutte le regioni e quella relativa all'Italia sono state fatte girare con questo comando. L'output è in grado di fornire i parametri del modello e i test di verifica sui residui di Durbin-Watson e di Ljung-Box.
Il confronto tra la varianza residua e la varianza della serie resa stazionaria è uno dei criteri per poter testare la variabilità del fenomeno. Una prima analisi consiste nel verificare se la varianza dei residui è sensibilmente inferiore alla varianza della serie resa stazionaria (Piccolo 1998).
Le analisi sui residui di un modello ARIMA stimato hanno due scopi principali. Il primo è quello di verificare la bontà del modello stimato non rifiutando nessuna delle ipotesi che caratterizzano un processo WN a_t: normalità, omoschedasticità, incorrelazione tra valori consecutivi. Il secondo è quello di suggerire, in caso di rifiuto di un modello, un altro alternativo, considerando i residui stimati come una nuova serie storica su cui ripetere la fase di identificazione, stima e verifica.
Una delle analisi più comuni per quanto riguarda l'aspetto della verifica della bontà di un modello riguarda lo studio delle funzioni di autocorrelazione dei residui. Considerando i residui stimati $\hat{a}_t$ come una serie storica di cui si vuole verificare la correlazione seriale, che deve essere identicamente nulla, si calcolano le stime delle autocorrelazioni globale e parziale e si verifica se tutti i valori dei lag *k*>0 non sono significativamente diversi da zero, cioè se non superano in valore assoluto $\pm 2n^{-1/2}$.
Il test che si utilizza è fondato sulla statistica:

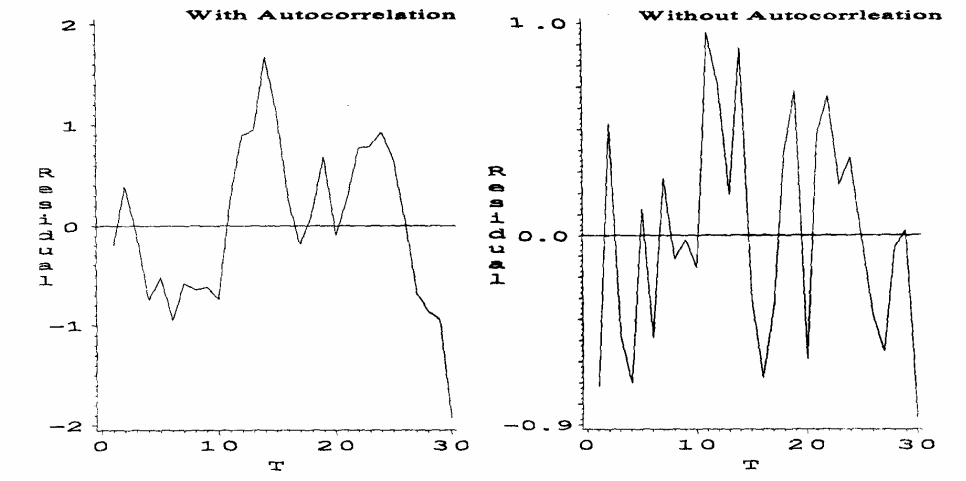
$$Q = n\sum_{k=1}^{m} W_k \left[\hat{\rho}_{\hat{a}}(k)\right]^2$$

in cui i pesi possono assumere i valori:

- $W_k \equiv 1$, secondo il test di Box-Pierce (1970);
- $W_k \equiv \frac{n+2}{n-k}$, secondo il test di Ljung-Box (1979).

La statistica Q è asintoticamente distribuita, sotto l'ipotesi nulla H_0 (secondo la quale i residui di un modello ARIMA stimato sono realizzazioni di un processo WN), come una variabile casuale χ^2 con *g=m-k* gradi di libertà, essendo *r* il numero totale dei parametri stimati nel modello di cui $\hat{a}_t$ sono i residui.
Il test utilizzato è quello di Ljung-Box in quanto più selettivo. Pertanto l'ipotesi nulla H_0, secondo la quale i residui stimati provengano da un processo WN, viene rifiutata se $Q > \chi^2_{\alpha,g}$, dove $\chi^2_{\alpha,g}$ è il $100(1-\alpha)\%$ percentile di una variabile casuale χ^2 con *g* gradi di libertà.
Gli errori correlati positivamente tendono a creare una lunga serie di residui che hanno lo stesso segno. Questo modello è di solito evidente in una distribuzione residuale dove i residui sono distribuiti in funzione della variabile tempo. Questo andamento per i dati di una serie storica simulata è mostrato per entrambi gli errori, correlati e non correlati, nella figura (Figura n. 1). L'ampiezza dell'errore è più ampia nel caso degli errori correlati.

Figura n. 1



Inoltre, come è noto, gli errori autocorrelati producono una tendenza per le serie più lunghe di residui che tende a zero.
Un test per trend di questo tipo è fornito da i "runs test" che possono essere usati per qualsiasi andamento non casuale.
Per lo studio di questi tipi di modelli, causati da autocorrelazione di primo ordine, si usa il test di Durbin-Watson, che è più specificatamente opportuno per testare questo genere di errori correlati. Questo test usa i residui della regressione ordinaria stimata attraverso il metodo dei minimi quadrati. La statistica D del test di Durbin-Watson è così calcolata:

$$D = \frac{\sum_{t>2}(e_t - e_{t-1})}{\sum_{t>2} e_t^2}$$

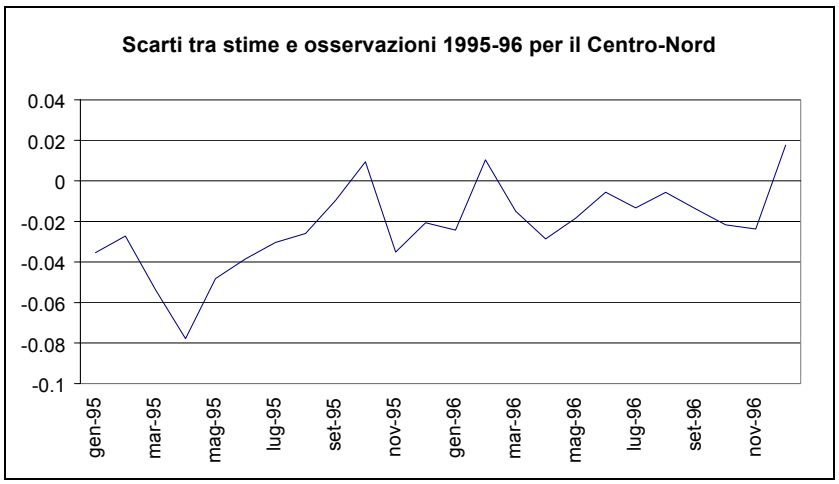
dove e_t è il residuo della t-esima osservazione.

Poiché e_{t-1} non può essere calcolato per la prima osservazione, entrambe le somme partono dall'osservazione numero due.
La distribuzione campionaria di questa statistica è piuttosto insolita. Il range della distribuzione è tra 0 e 4, e sotto l'ipotesi nulla H_0 di non autocorrelazione, la media di questa distribuzione è all'incirca due. Le autocorrelazioni positive creano piccole differenze vicine, per questo tendono a ridurre il numeratore. Quindi la regione di rifiuto per le correlazioni positive è nella coda più piccola della distribuzione. Inoltre il calcolo dei valori critici della distribuzione non dipende solo dall'ampiezza campionaria, ma anche dal numero delle variabili indipendenti e dall'andamento delle variabili indipendenti; i valori critici non sono esatti, ma sono solo una buona approssimazione (Piccolo 1994).

2.4 L'Elaborazione

Come primo passo si è proceduto alla stima dei nati per gli anni 1995-1996 per la serie di ogni regione che partiva dall'anno 1969. In seguito è stato effettuato il confronto tra le osservazioni degli anni 1995-1996 e le stime per gli anni stessi per verificare la bontà di adattamento del modello. Inoltre, con lo scopo di ricercare il modello che meglio rappresentasse le serie di ogni regione, è stato studiato il trend seriale e si è provato ad eliminare la parte decrescente sotto l'ipotesi che negli anni '90 i tassi di fecondità totali non risentissero più del forte calo che ha caratterizzato gli anni '70.
Si è così proceduto alla stima delle nascite per gli anni 1995-1996 con le serie che partivano dal 1983. E' stato scelto l'anno 1983 poiché dall'inizio degli anni '80 la serie dei nati tende ad avere un trend stabile, e anche perché questo particolare anno assicura 144 osservazioni che sono sufficienti per procedere alla stima dei parametri dei modelli ARIMA.
Dall'analisi delle elaborazioni è risultato che per le regioni del Centro-Nord, nelle quali è inclusa anche la Sardegna (che ha un andamento riproduttivo simile a quello delle regioni settentrionali), i modelli che assicurano una bontà migliore sono quelli derivanti dalle serie depurate dal trend decrescente 1983-1994, mentre per le regioni meridionali il confronto osservazioni stime risulta migliore con modelli costruiti dalle serie più lunghe 1969-1994.
Per il Centro-Nord gli scarti tra le stime ed i valori osservati per gli anni 1995-1996 sono contenuti tra il –0.04% e lo 0.02% tranne che per due casi, marzo e aprile 1995 dove arrivano anche al –0.08 (Figura n. 2).

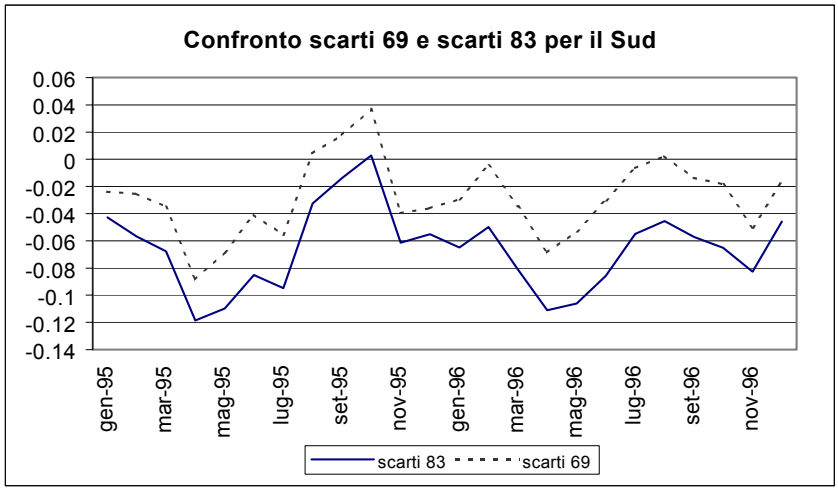
Figura n. 2



Lo scarto sul totale anno è del –0.3213% per il 1995 e del –0.01186 per il 1996.
Per le regioni del Sud gli scarti relativi alle stime ottenute con serie partenti dal 1983 sono molto più alte, con punte anche del –0.12%. La situazione migliora con le stime ottenute dalle serie che

partono dal 1969. Gli scarti, tranne che per due casi, sono contenuti tra il –0.06% e lo 0.04% (Figura n. 3).

Figura n. 3



Il miglioramento è visibile anche per il totale anno, si passa da uno scarto di –0.060212 per il 1995 e di –0.07423 per il 1996 con stime ottenute dalle serie troncate al 1983 a dei valori accettabili pari a –0.028068 per il 1995 e a –0.026301 per il 1996 con le serie complete che partono dal 1969.
Questa differenziazione tra Nord-Centro e Sud era prevedibile anche a priori. Infatti la riproduzione in Italia è sempre stata caratterizzata da forti disuguaglianze territoriali. Non è corretto considerare la fecondità italiana come un fenomeno univoco, i dati disaggregati hanno sempre evidenziati più di un modello evolutivo che ha mostrato l'esistenza di almeno due Italie: una formata dalle regioni del Centro-Nord e dalla Sardegna, che da tempo sono al di sotto del livello di sostituzione, e l'altra formata dalle regioni del Meridione, caratterizzata da una fecondità in declino ma che ancora non ha raggiunto i livelli delle regioni settentrionali (Figura n. 4a e 4b).

Figura n. 4a

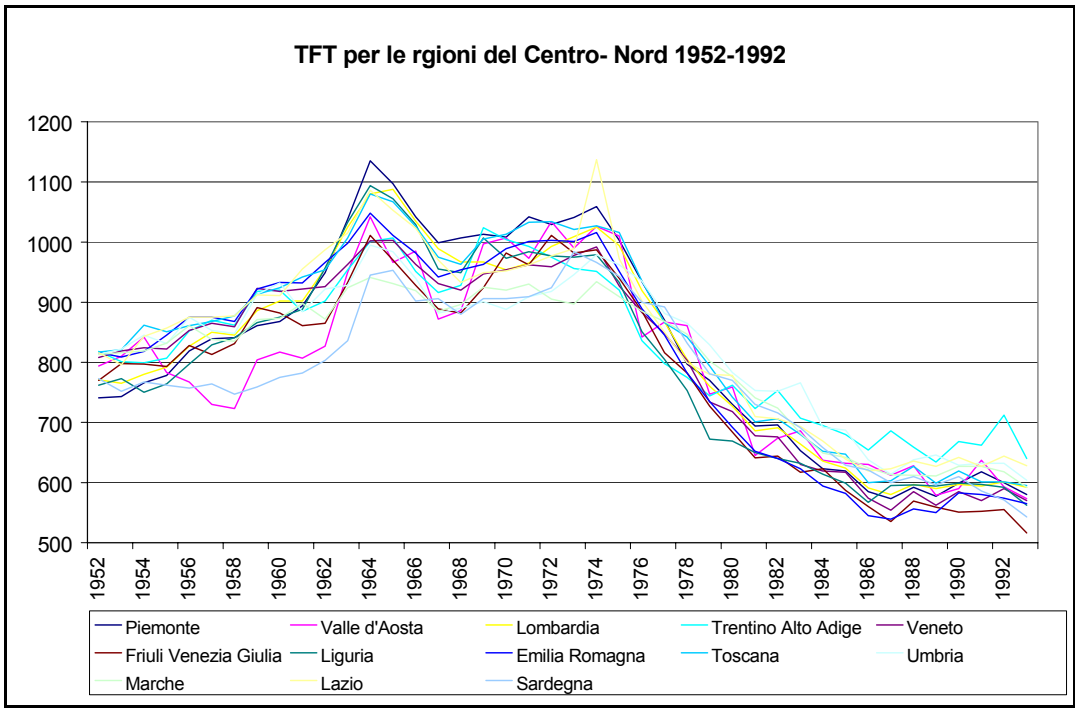
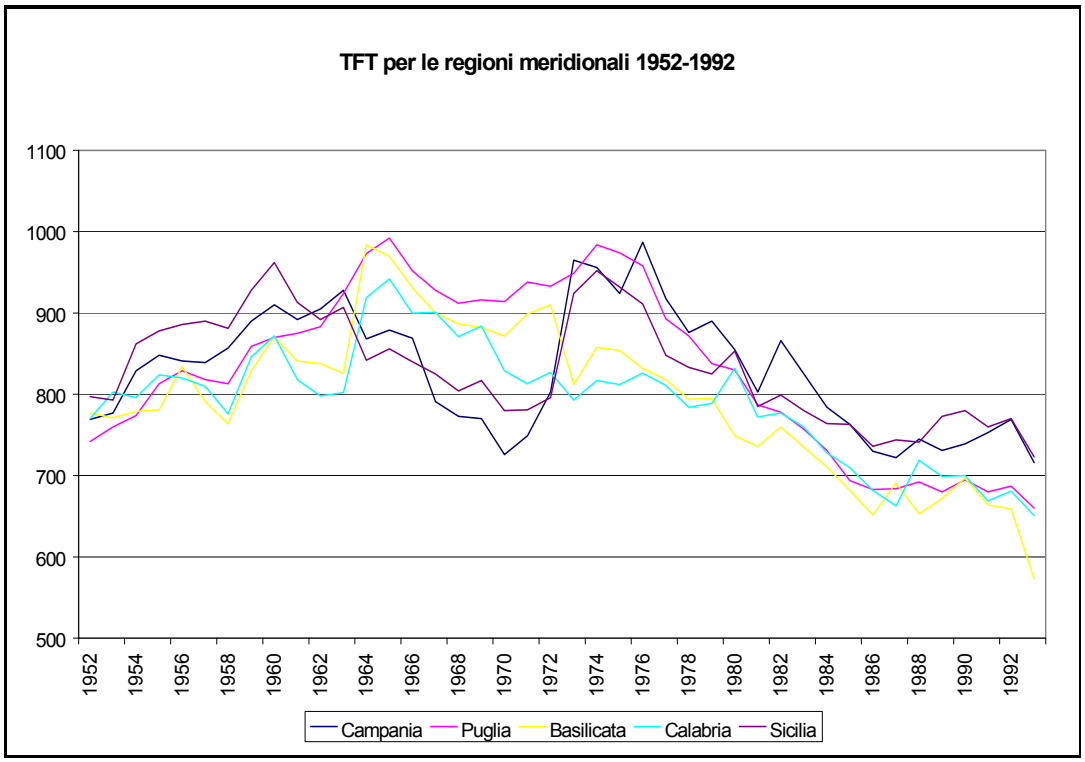


Figura n. 4b



Questa situazione ha fatto sì che le regioni del Centro-Nord non risentissero più del forte calo della fecondità già a partire dai primi anni '80 e che quindi i modelli stimati con le serie troncate non risentissero più del trend fortemente decrescente degli anni settanta. Invece per le regioni del Sud non si è ancora raggiunta una stabilità della fecondità pertanto i modelli migliori risultano quelli ricavati dalle serie più lunghe in cui le ripercussioni del trend fortemente decrescente sono ancora attuali.
Le previsioni delle nascite per l'Italia sono state effettuate non applicando il modello alla serie dei nati dal 1969, ma come somma delle previsioni ottenute per le singole regioni.

3. La metodologia di stima dei dati aggregati

I dati aggregati diffusi dall'Istat relativamente al 1997 e al 1998 sono stati stimati, secondo la metodologia descritta di seguito, sulla base delle previsioni ottenute come descritto nel par. 2.
Poiché le suddette previsioni sono a livello regionale, non si è ritenuto opportuno diffondere dati per disaggregazioni subregionali (per provincia e/o per comune), non essendo possibile stimare i dati ad un tale livello di dettaglio.

3.1 La stima utilizzata per i dati del 1997

Nel 1997 si è manifestata un'ulteriore distorsione nella registrazione dei nati vivi naturali. Questi ultimi sono, infatti, risultati pari a 31.092 unità (contro le 43.758 del 1996). La causa di questo calo è da ricercarsi nel fatto che, con molta probabilità, si sono maggiormente avvalse della nuova legge, registrando i propri figli direttamente presso la direzione sanitaria del centro di nascita, le madri di bambini naturali, che sarebbero invece dovute andare col padre del bambino all'ufficio di Stato Civile nei giorni immediatamente successivi al parto. La mancanza di organizzazione fra l'Istat e gli ospedali per il recupero dei modelli, ha fatto quindi si che, per il 1997, i nati vivi naturali fossero fortemente inferiori rispetto al valore reale.

Per questo motivo si è provveduto ad effettuare un'ulteriore stima della nascite naturali in base al seguente criterio:

Per regione di nascita

- Sono state calcolate le percentuali di nati vivi naturali per i primi 6 mesi del 1997. Queste percentuali sono infatti risultate in linea con quelle registrate negli anni precedenti (8,7% a livello nazionale), poiché la "Bassanini-bis" non era ancora entrata in vigore;
- Sono state applicate le percentuali alle previsioni dei nati vivi, ottenendo una stima dei nati naturali;
- La stima dei nati legittimi è stata ottenuta come differenza fra la previsione del numero di nati vivi e la stima dei nati naturali;
- Sono state calcolate due serie di pesi, la prima da utilizzare per i nati vivi legittimi, la seconda per i nati vivi naturali. I pesi sono stati calcolati nel seguente modo:

PESO NATI VIVI LEGITTIMI=NVL_S/NVL_E

dove

NVL_S=Nati vivi legittimi stimati;
NVL_E=Nati vivi legittimi effettivi (estratti dall'archivio dei nati del 1997 e relativi all'intero anno);

PESO NATI VIVI NATURALI=NVN_S/NVN_E

dove

NVN_S=Nati vivi naturali stimati;
NVN_E=Nati vivi naturali effettivi (estratti dall'archivio dei nati del 1997 e relativi all'intero anno);

I dati aggregati sono stati ottenuti moltiplicando i dati estratti dall'archivio per i suddetti pesi.

3.2 La stima utilizzata per i dati del 1998

Per i dati del 1998 è stata calcolata una serie di pesi in base alla formula:

PESO NATI VIVI=NV_S/NV_E

dove

NV_S=Nati vivi previsti;
NV_E=Nati vivi effettivi (estratti dall'archivio dei nati del 1998 e relativi all'intero anno);

I dati aggregati sono stati ottenuti moltiplicando i dati estratti dall'archivio per i suddetti pesi.