ESeC – Economic Statistics e-Center

Paola Maddalena Chiodini, Università degli Studi di Milano-Bicocca
Flavio Verrecchia, ESeC

Imputazione dei dati mancanti in basi dati economico-sociali per il forecasting regionale: il metodo ESeC-Rubin

In letteratura, per l'imputazione dei dati mancanti nelle serie storiche, si fa riferimento a statistiche applicate all'intera serie analizzata (e.g. media di tutti i termini della serie), ottenendo una costante d'imputazione generalmente adeguata per una specifica serie. Se le serie sono n ($n \to \infty$) è impossibile trovare un'unica funzione per le n costanti di imputazione dei missing.

Obiettivo

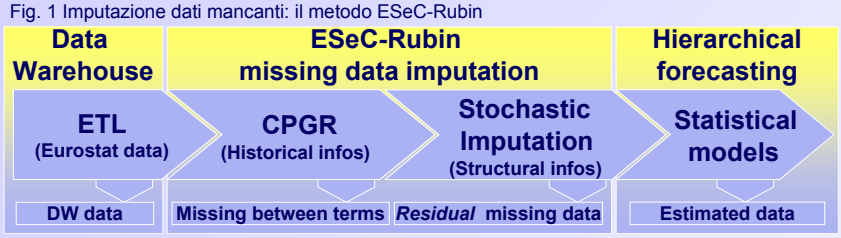
Obiettivo del lavoro è proporre un nuovo metodo di imputazione dei missing values - ESeC-Rubin - per basi dati gerarchiche, che trae spunto dalla teoria dei campioni, finalizzato alla modellistica temporale. In particolare, la ESeC-Rubin consente di ricostruire il dato mancante tenendo conto di una sequenza di metodi di imputazione e della naturale variabilità degli aggregati studiati. **L'applicazione prodotta con SAS Forecast Server su dati socio-economici di fonte Eurostat, consente di comparare i modelli (selezionati in automatico) a partire dalla base dati osservata con missing values e differenti tipologie di imputazione.**

Soluzione

La soluzione proposta (Fig. 1), si basa sull'uso sequenziale di vari metodi di imputazione sfruttando:

- l'**informazione temporale** - si applica la *CPGR* (Def. 1) solo ai casi mancanti interni a due termini della serie osservata;
- gli **indici a base mobile** (numeri puri indipendenti dal fattore di scala) sulle determinazioni delle singole serie;
- l'**informazione gerarchica** (e.g. territoriale) e l'ipotesi che i parametri empirici, per ogni t (tempo), siano i medesimi delle distribuzioni da cui sono stati generati i dati (e.g. media e deviazione standard degli indici a base mobile regionali delle singole nazioni) - si procede:
  - con l'estrazione casuale, dalle t distribuzioni ipotizzate normali (Fig. 4), al fine di imputare gli indici a base mobile mancanti;
  - con la ricostruzione dei missing values restanti attraverso l'applicazione degli indici a base mobile alle determinazioni esistenti.

Fig. 1 Imputazione dati mancanti: il metodo ESeC-Rubin



Def. 1 *Indice CPGR (Compound Periodical Growth Rate)*
la variazione media di periodo tra t e b (con t situazione confrontata e b situazione base), definita dalla seguente espressione

$$\left(e_t \cdot e_b^{-1} \right)^{(t-b)^{-1}}$$

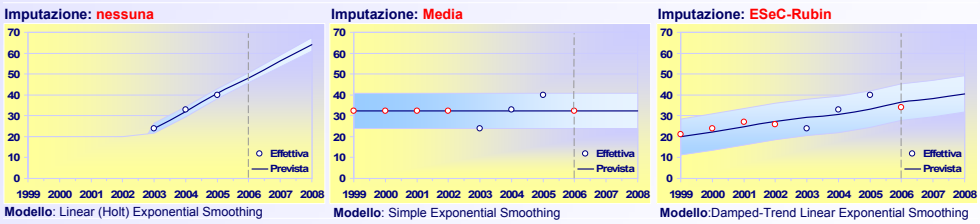
è detta tasso periodale medio di variazione (e nel caso annuale CAGR). Dove per **e** si intendono gli eventi elementari osservati.

Benefici

La **CPGR** consente un'imputazione basata sul tasso medio annuale di crescita tra due termini noti. Ad esempio, nel caso del Tirol (Tab. 1), per la disoccupazione 15-24, anni si hanno missing values tra il 2000 e il 2002. Con la media (3200 unità), si imputano dati del 50% superiori agli estremi dell'intervallo 1999-2003. Con la CPGR invece i dati imputati sono prossimi (ed interni) ai termini estremi osservati oltre che non pari ad una costante.

Con la **Stochastic Imputation** si imputano invece i casi mancanti agli estremi delle serie storica. La variabilità dei dati imputati è necessaria per la successiva fase di specificazione dei modelli. Ad esempio, per la regione Kärnten l'imputazione della media porta alla specificazione di un modello costante quando invece i dati osservati non sono stazionari in media (Fig. 2).

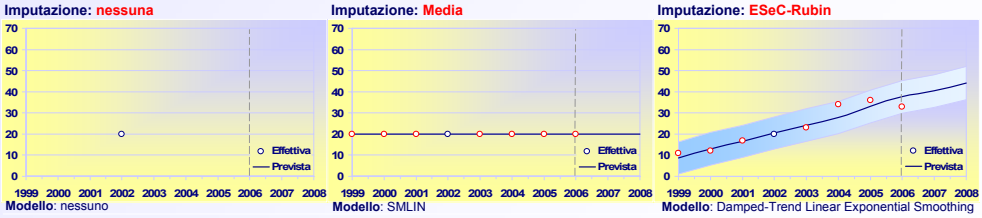
Fig. 2 Modelli e previsione della disoccupazione (in centinaia), classe d'età 15-24, Kärnten, 1999-2008



Modello: Linear (Holt) Exponential Smoothing — **Modello**: Simple Exponential Smoothing — **Modello**:Damped-Trend Linear Exponential Smoothing

Fonte: stime ESeC su dati Eurostat (Labour Force Survey). **Note**: Il MAPE (10.00) della ESeC-Rubin è naturalmente il più alto.

Fig. 3 Modelli e previsione della disoccupazione (in centinaia), classe d'età 15-24, Salzburg, 1999-2008



Modello: nessuno — **Modello**: SMLIN — **Modello**: Damped-Trend Linear Exponential Smoothing

Fonte: stime ESeC su dati Eurostat (Labour Force Survey). **Note**: Il MAPE (10.08) della ESeC-Rubin è naturalmente il più alto.

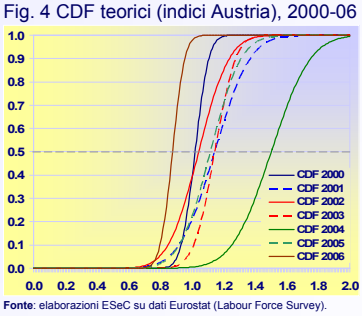
La **ESeC-Rubin** oltre ad utilizzare le informazioni temporali e strutturali disponibili, consente di specificare modelli della realtà che non producano previsioni indefinitamente costanti anche nel caso limite di una sola rilevazione per unità territoriale (Fig. 3).

Tab. 1 Disoccupazione (in centinaia), classe età 15-24, Tirol, 1999-06

	1999	2000	2001	2002	2003	2004	2005	2006
Serie	20	.	.	.	21	39	44	36
Media = 32	20	32	32	32	21	39	44	36
CPGR	20	20	20	21	21	39	44	36

Fonte: elaborazioni ESeC su dati Eurostat (Labour Force Survey) – dati disponibili al 12 marzo 2008.

Fig. 4 CDF teorici (indici Austria), 2000-06



Fonte: elaborazioni ESeC su dati Eurostat (Labour Force Survey).

Bibliografia essenziale

Brocklebank J. and Dickey D. (2003) *SAS for Forecasting Time Series*, SAS Institute, USA.

Little R.J.A. (1986) "Survey nonresponse adjustments for estimates of means", *Intern. Stat. Review*, 54, 139-157.

Martini M. (2001) *Numeri indice per il confronto nel tempo e nello spazio*, CUSL, Milano.

Rubin D.B. (1996) "Multiple imputation after 18+ year". *J. Am. Stat. Assoc.*, 91, 507-510.

Verrecchia F. (2005) "Théorie des nombres index: les Nombres Index généralisés (gIN)", Actes des JMS, INSEE, Paris.