L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione

Fabio Bacchini[1*] Roberto Iannaccone[1]
Edoardo Otranto [2]

[1] Direzione statistiche congiunturali
Istituto Nazionale di Statistica

[2] Dipartimento di Economia, Impresa e Regolamentazione
Università degli Studi di Sassari

*Fabio Bacchini, Direzione statistiche congiunturali, Istituto Nazionale di Statistica, Via Tuscolana, 1782, 00173 Roma. E-mail: bacchini@istat.it

Abstract

For longitudinal data set nonresponse occurs for a unit at different points in time. Considering the missingness from a longitudinal perspective the non response in some period can be characterised as a set of *item nonresponse*. On the other side if each period is considered as a cross-section the missing pattern can be seen as a *unit nonresponse* problem. For the item nonreponse a method of imputation can be used to replace each missing value with a known acceptable value. Weighting methods are considered for unit nonresponse problem. Our present work deals with a longitudinal method of imputation based on a hot-deck donor for the estimation of missing values for the building permits survey (2000- 2002). The method is compared with an imputation method used for cross-section data and a weighting method based on the estimation of the non response probabilities.

Riassunto

Nel caso longitudinale le unità sono chiamate a fornire informazioni più volte nel corso del tempo e la mancata risposta può riferirsi a determinati periodi. In presenza di dati longitudinali la mancata risposta è rappresentata da unità che non rispondono in alcuni dei periodi che caratterizzano l'indagine. Da un punto di vista longitudinale la non risposta per alcuni periodi può essere vista come un insieme di mancate risposte parziali (*item nonresponse*) mentre da un punto di vista cross-section ciascuna mancata risposta può essere interpretata come una mancata risposta totale (*unit nonresponse*). Nel primo caso per la stima delle mancate risposte può essere appropriata una procedura di imputazione mentre nel secondo si può scegliere una strategia di ponderazione. In questo lavoro viene presentato un metodo di imputazione longitudinale basato sul donatore, applicato all'indagine sui permessi di costruzione per la stima del numero di nuove abitazioni nel triennio 2000- 2002. I risultati ottenuti sono stati confrontati sia con un metodo tradizionale di imputazione cross-section, la media, sia con uno di ponderazione basato sulla stima della probabilità di non risposta.

Keywords: Nonresponse; weighting adjustment; imputation; longitudinal data.

Indice

1 Introduzione

Nelle indagini statistiche sia censuarie sia campionarie la scelta della metodologia per il trattamento delle mancate risposte è associata alla distinzione tra mancate risposte totali (MRT) e mancate risposte parziali (MRP). La MRT viene comunemente trattata con procedure di ponderazione mentre il trattamento della MRP è realizzato attraverso tecniche di imputazione.

Nel caso longitudinale le unità sono chiamate a fornire informazioni più volte nel corso del tempo e la mancata risposta può riferirsi a determinati periodi. Da un punto di vista longitudinale questo tipo di mancata risposta può essere visto come un insieme di MRP in un record longitudinale suggerendo che l'imputazione possa essere la tecnica appropriata per la stima; da un punto di vista cross-section può essere interpretata come una MRT per la quale utilizzare una procedura di ponderazione (Kalton, 1986).

Sebbene nella recente letteratura sui panel la mancata risposta sia ampiamente trattata in termini di relazione tra il processo che la genera ed il modello di regressione di interesse (Robins e al., 1995, Vella, 1998), sono scarse le applicazioni che presentano confronti tra metodi di imputazione e di ponderazione nel caso longitudinale.

Nel presente lavoro si propone la stima della mancata risposta per la serie mensile dei permessi di costruzione rilasciati dai comuni italiani nel triennio 2000-2002. Una volta verificato che il processo di generazione della mancata risposta sia di tipo Missing at random (MAR), la stima è realizzata utilizzando un metodo di imputazione longitudinale basato sul donatore i cui risultati vengono confrontati sia con un metodo tradizionale di imputazione di tipo cross-section, la media, sia con uno di riponderazione basato sulla stima della probabilità di non risposta (Little e al., 1986). A conoscenza di chi scrive questo è uno dei primi lavori che presenta il confronto in modo sistematico estendendo così al caso panel il panorama dei contributi empirici disponibili per il caso cross-section (ad esempio Haziza et al., 2001, Chen e Shao, 2000).

Il lavoro è organizzato in 4 parti. Nella prima viene descritta in dettaglio la rilevazione dell'attività edilizia evidenziando le caratteristiche del pattern di collaborazione dei comuni e verificando l'ipotesi di MAR rispetto alle variabili di stratificazione. Nella seconda parte sono descritti i metodi di imputazione utilizzati mentre la terza parte è dedicata alla simulazione per la comparazione tra i metodi. Infine, nella quarta parte, i metodi vengono applicati al caso concreto.

2 La rilevazione dell'attività edilizia

In questa sezione si descrive la rilevazione dell'attività edilizia illustrando il tipo di mancata risposta che la caratterizza.

2.1 I dati

La rilevazione dell'attività edilizia rileva, mensilmente, le informazioni sui nuovi fabbricati residenziali e non residenziali e sugli ampliamenti di quelli preesistenti sulla base dei *Permessi di costruire* rilasciati dai Comuni. L'indagine è organizzata come un censimento mensile su tutti i Comuni italiani, cui si richiede, con un apposito modulo, anche la segnalazione di attività negativa. Alcune delle variabili raccolte mediante il modello (il numero di nuove abitazioni, la loro superficie e la superficie dei fabbricati non residenziali) fanno parte dell'insieme richiesto dal regolamento sulle statistiche congiunturali per il settore delle costruzioni (STS - Annex

B). Il regolamento richiede il totale nazionale delle variabili d'interesse rendendo necessaria la definizione di una procedura per la stima delle mancate risposte che caratterizzano la rilevazione. Nel presente lavoro i dati utilizzati si riferiscono al triennio 2000 - 2002, periodo per il quale l'Istat ha rilasciato la stima dei dati annuali sulla base dei risultati del lavoro presentato. Nel proseguo le analisi sono concentrate solo sul numero di nuove abitazioni che rappresenta la variabile più importante nel comparto dell'edilizia.

Nei tre anni considerati l'analisi della mancata risposta è stata condotta suddividendo gli $n = 8.100$ comuni (universo di riferimento) in due sottoinsiemi:

- i comuni capoluogo di provincia ed i non capoluogo con più di 50.000 abitanti ($n_1 = 160$) pari ad una popolazione di 20.998.197 abitanti ($36,4\%$ del totale Italia);

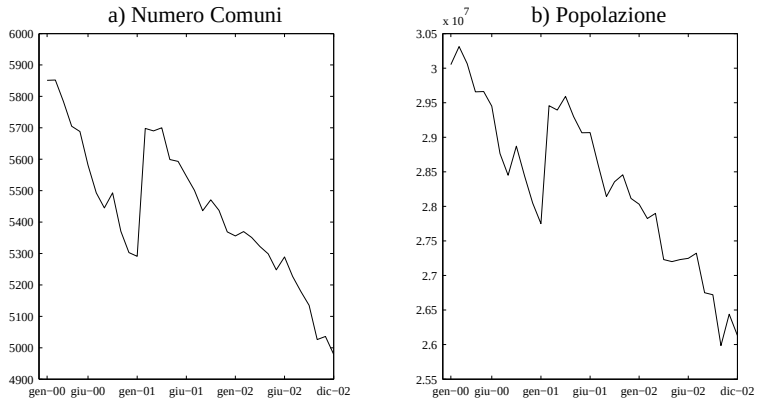
- i comuni non capoluogo con meno di 50.000 abitanti ($n_2 = 7.940$), in termini di popolazione 36.691.698 abitanti ($63,6\%$ del totale Italia).

Nel primo sottoinsieme la collaborazione risulta piuttosto elevata (tabella 1): l'$85,6\%$ dei comuni nel 2000 ha collaborato per 12 mesi (il $91,4\%$ in termini di popolazione). Il livello rimane elevato anche negli altri anni anche se in diminuzione.

La distribuzione degli $n_2 = 7.940$ comuni rispetto al numero di mesi di collaborazione presenta una alta concentrazione di comuni nelle code (tabella 1). Nei tre anni più del 50% dei comuni (67,9%, 69,1% e 61,9% in termini di popolazione rispettivamente per il 2000, il 2001 ed il 2002) hanno collaborato almeno 11 mesi. Tuttavia 1.384 comuni nel 2000, 1.626 nel 2001 e 1.875 nel 2002, non hanno mai collaborato alla rilevazione.

La progressiva diminuzione della collaborazione dei comuni risulta evidente anche dalla distribuzione dei rispondenti per mese. In particolare, si osserva un forte calo nel secondo semestre del 2002 (figura 1).

Figura 1: *Collaborazione alla rilevazione per numero di comuni (a) e popolazione (b)*



2.2 La mancata risposta

La mancata risposta nei panel è classificabile a seconda del suo modo di manifestarsi nel tempo. Supponendo per semplicità di osservare solo 3 mesi ed indicando con 1 la risposta e con 0 la non risposta, sono possibili $2^3 = 8$ pattern di risposta: 111; 110; 101; 011; 100; 010; 001;

Tabella 1: *Distribuzione dei comuni capoluogo e non capoluogo per numero mesi di collaborazione: Anni 2000-2002*

Valori assoluti

	Capoluogo						Non capoluogo					
	2000		2001		2002		2000		2001		2002	
Mesi coll.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.
0	4	229,777	6	316,057	7	527,518	1,384	3,978,814	1,626	4,690,383	1,875	6,090,853
1	-	-	-	-	1	35,328	203	550,889	196	688,489	183	543,943
2	-	-	-	-	-	-	136	403,221	104	418,368	104	366,838
3	1	35,328	2	123,765	2	124,559	127	339,577	92	274,607	113	530,646
4	1	56,355	2	200,149	1	55,856	108	338,073	106	331,605	95	387,298
5	1	54,035	1	54,035	2	168,022	136	686,557	90	435,422	111	478,991
6	-	-	2	1,066,564	5	358,571	125	453,637	123	602,857	113	410,378
7	-	-	1	91,729	3	181,389	140	619,916	141	609,951	148	565,572
8	1	64,596	3	388,699	2	433,586	177	660,907	148	781,962	170	893,223
9	3	324,123	1	93,454	3	1,319,183	303	1,373,867	213	975,971	279	1,500,045
10	5	408,384	4	490,520	7	441,893	525	2,357,283	307	1,493,471	453	2,202,343
11	7	636,185	9	880,176	24	2,089,624	1,146	5,076,565	582	2,805,252	871	4,582,977
12	137	19,179,414	129	17,283,049	103	15,252,668	3,430	19,852,392	4,212	22,583,360	3,425	18,138,591
Totale	160	20,988,197	160	20,988,197	160	20,988,197	7,940	36,691,698	7,940	36,691,698	7,940	36,691,698

Valori percentuali

	Capoluogo						Non capoluogo					
	2000		2001		2002		2000		2001		2002	
Mesi coll.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.	N.	Popol.
0	2.5%	1.1%	3.8%	1.5%	4.4%	2.5%	17.4%	10.8%	20.5%	12.8%	23.6%	16.6%
1	-	-	-	-	0.6%	0.2%	2.6%	1.5%	2.5%	1.9%	2.3%	1.5%
2	-	-	-	-	-	-	1.7%	1.1%	1.3%	1.1%	1.3%	1.0%
3	0.6%	0.2%	1.3%	0.6%	1.3%	0.6%	1.6%	0.9%	1.2%	0.7%	1.4%	1.4%
4	0.6%	0.3%	1.3%	1.0%	0.6%	0.3%	1.4%	0.9%	1.3%	0.9%	1.2%	1.1%
5	0.6%	0.3%	0.6%	0.3%	1.3%	0.8%	1.7%	1.9%	1.1%	1.2%	1.4%	1.3%
6	-	-	1.3%	5.1%	3.1%	1.7%	1.6%	1.2%	1.5%	1.6%	1.4%	1.1%
7	-	-	0.6%	0.4%	1.9%	0.9%	1.8%	1.7%	1.8%	1.7%	1.9%	1.5%
8	0.6%	0.3%	1.9%	1.9%	1.3%	2.1%	2.2%	1.8%	1.9%	2.1%	2.1%	2.4%
9	1.9%	1.5%	0.6%	0.4%	1.9%	6.3%	3.8%	3.7%	2.7%	2.7%	3.5%	4.1%
10	3.1%	1.9%	2.5%	2.3%	4.4%	2.1%	6.6%	6.4%	3.9%	4.1%	5.7%	6.0%
11	4.4%	3.0%	5.6%	4.2%	15.0%	10.0%	14.4%	13.8%	7.3%	7.6%	11.0%	12.5%
12	85.6%	91.4%	80.6%	82.3%	64.4%	72.7%	43.2%	54.1%	53.0%	61.5%	43.1%	49.4%
Totale	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

000. Little e David (1983) identificano i casi 110, 100 e 000 come casi di *attrition* (dove 000 rappresenta la mancata risposta totale), il comune risponde per un periodo continuativo e poi cessa definitivamente di collaborare; il pattern 101 costituisce il caso di rientro; i pattern 011 e 001 costituiscono l'entrata in ritardo; il pattern 010 rappresenta il caso di entrata in ritardo con successiva mancata collaborazione. La rilevazione dell'attività edilizia è caratterizza da un fenomeno di entrata/uscita dei comuni nella collaborazione (tabella 2) pari al $59,0\%$ dei 7.940 presenti nel triennio 2000-02, mentre 1.910 comuni hanno risposto sempre e 922 mai.

Tabella 2: *Distribuzione della mancata risposta per tipologia Anni 2000-2002*

	Valori assoluti	Percentuali
Mancata risposta totale	922	11,6
Attrition	378	4,8
Entrata in ritardo	46	0,6
Entrata/uscita	4.684	59,0
Collaborazione	1.910	24,1
Totali	7.940	100,0

In presenza di mancata risposta, 'per poter effettuare inferenza valida sui parametri del modello assunto per i dati senza specificare esplicitamente il meccanismo di mancata risposta è necessario fare l'assunzione che esso sia Missing at Random (MAR) cioè che la mancata risposta, condizionatamente ai valori osservati, non dipenda dai dati mancanti' (Manzari, 2004, Little e Rubin, 2002). Per verificare l'esistenza di un meccanismo di tipo MAR, considerando la popolazione e la ripartizione geografica come variabili ausiliari, si è stimato un modello di regressione logistica in cui la variabile di risposta y è definita come:

$$y = \begin{cases} 0 & \text{se numero di mesi di collaborazione} = 12 \\ 1 & \text{altrimenti} \end{cases} \tag{1}$$

La popolazione è considerata come variabile quantitativa; per la ripartizione geografica sono state provate diverse configurazioni: da una partizione più fine a 5 modalità (1=Nord-est, 2=Nord-ovest, 3=Centro, 4=Sud e 5=Isole) fino ad arrivare ad una partizione in due classi. Le stime del modello logistico risultano migliori con la scelta di 3 ripartizioni (1=Nord-est, 2=Nord-ovest, 3=Centro, Sud e Isole)[1]. Nella tabella 3 si riportano, separatamente per ciascun anno, il numero di comuni che non hanno risposto tutti e 12 i mesi. La distribuzione evidenzia tre distinti livelli di non risposta congiuntamente ad un aumento della collaborazione nel 2001 (anno di censimento): nel Nord-est le mancate risposte non superano il 40%, nel Nord-ovest raggiungono un picco del $50,7\%$ nel 2000 mentre nella ripartizione Centro-Mezzogiorno si registrano i valori più bassi di collaborazione con una punta del $73,1\%$ nel 2002.

Per la variabile risposta binaria y e le variabili ausiliarie $\mathbf{X}$, indichiamo con $\pi(\mathbf{x})$ la probabilità di che un comune abbia collaborato meno di 12 mesi ($y = 1$) quando il vettore $\mathbf{X}$ assume il valore $\mathbf{x}$. Il modello di regressione logistica che utilizziamo nel seguito considera per il logit della probabilità la seguente forma lineare:

$$\text{logit}[\pi(\mathbf{x})] = \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{2}$$

[1]Per la scelta della ripartizione è stato utilizzato il valore dello Schwartz Criterion della stima del modello logit

Tabella 3: *Distribuzione dei comuni non rispondenti 12 mesi per ripartizione - Anni 2000-2002*

Ripartizione	**N. comuni**	**Frequenze assolute**			**Frequenze percentuali**		
		2000	2001	2002	2000	2001	2002
Nord-ovest	3.027	1.536	1.079	1.420	50,7%	35,6%	46,9%
Nord-est	1.453	569	498	566	39,2%	34,3%	39,0%
Centro-Sud	3.460	2.405	2.151	2.529	69,5%	62,2%	73,1%
Totale	7.940	4.510	3.728	4.515			

dove x_1 e x_2 rappresentano le due variabili dummies associate alle 3 ripartizioni territoriali mentre x_3 è la popolazione.

Nella stima del modello logit (tabella 4), per tutti i parametri si verifica che la $Pr > \chi^2$ è inferiore a $0,0001$. Nei 3 anni la probabilità di non risposta risulta correlata negativamente con i comuni appartenenti alla ripartizione Nord-ovest e Nord-est (assunto come baseline) nonchè con la popolazione, indicando una maggiore collaborazione da parte dei comuni situati al nord con dimensioni demografiche più ampie.

Tabella 4: *Stima dei parametri della regressione logistica: Anni 2000-2002*

	2000		**2001**		**2002**	
	Stima	Wald χ^2	Stima	Wald χ^2	Stima	Wald χ^2
β_0	0,433	193,4	-	-	0,346	125,3
β_1	-0,198	34,9	-0,421	161,4	-0,329	96,2
β_2	0,773	515,4	0,819	612,2	0,917	708,5
β_3	-0,00006	242,4	-0,00006	271,3	-0,00004	130,2

La stima del modello logistico fornisce anche indicazioni sull'esistenza di una relazione tra la probabilità di osservare una mancata risposta e le variabili ausiliarie: la mancata risposta non è quindi indipendente dai valori osservati per le variabili ausiliarie considerate, avallando l'ipotesi di un meccanismo di tipo MAR.

3 I metodi di imputazione

L'obiettivo del lavoro è quello di valutare la performance del metodo longitudinale di imputazione proposto, basato sull'individuazione di un donatore di minima distanza, per stimare il valore totale delle abitazioni mensili riferite ai permessi di costruzione dei 7.940 comuni appartenenti al secondo gruppo. I risultati ottenuti sono stati confrontati sia con un tradizionale metodo di imputazione cross-section, la media, sia con una strategia di ponderazione effettuata mese per mese. Nel paragrafo si descrivono i 3 metodi utilizzati.

3.1 Imputazione longitudinale: il donatore

I metodi di imputazione longitudinale sono disegnati per utilizzare l'informazione disponibile nei diversi periodi ai fini dell'imputazione. Heeringa e Lepkowski (1986) propongono 5 classi generali per l'imputazione longitudinale: i) diretta sostituzione longitudinale; ii) imputazione deterministica delle variazioni; iii) imputazione attraverso regressione longitudinale; iv) hot-deck longitudinale; v) hot-deck longitudinale delle variazioni.

Il metodo proposto nel lavoro combina la diretta sostituzione con l'individuazione di un donatore di minima distanza. Come è noto, il metodo del donatore (nearest neighbor imputation, NNI) produce, sotto determinate condizioni, stime asintoticamente corrette e stimatori consistenti del totale della popolazione, della sua distribuzione e dei suoi quantili (Chen e Shao, 2000). L'applicazione del metodo è assai diffusa (Statistics Canada, U.S. Census Bureau) e in continuo sviluppo in associazione con la diffusione di software generalizzati per l'editing e l'imputazione dei dati (si veda ad esempio i risultati del progetto Euredit in Chambers, 2000). Per descrivere il metodo si supponga di essere in un caso cross-section e di avere provato a rilevare il fenomeno y su n unità, ma di disporre di solo r risposte e, conseguentemente di $n-r$ mancate risposte mentre i valori delle variabili ausiliarie x sono disponibili per tutte le unità. Supponendo per semplicità che $y_{r+1}, ..., y_n$ sono le mancate risposte, il metodo NNI imputa y_j, $r+1 \leq j \leq n$, con y_i, $1 \leq i \leq r$, dove i è il donatore più vicino a j in termini della distanza misurata rispetto alla variabile ausiliaria x:

$$|x_i - x_j| = \min_{1 \leq k \leq r} |x_k - x_j|$$

Nel caso in cui ci sia più di un donatore per l'unità j allora viene effettuata una scelta casuale tra le unità con il valore minimo.

Nel caso longitudinale di nostro interesse, considerando come periodo di riferimento l'anno, sono considerate due possibili situazioni di mancate risposta.

Nel primo caso, quello più frequente nella realtà empirica, il comune risponde almeno in uno dei dodici mesi di riferimento. All'interno di ciascun strato, il donatore viene individuato minimizzando, per ciascuna unità j che risulta non rispondente in uno o più mesi dell'anno, la seguente funzione di distanza:

$$|x_i - x_j| = \min_{1 \leq k \leq r_h} \sum_{m \in M} |x_k^m - x_j^m|$$

dove M indica l'insieme dei mesi in cui l'unità j ha risposto nel corso dell'anno (i mesi possono non essere temporalmente contigui) e r_h il numero dei rispondenti 12 mesi nello strato specificato.

Nel secondo caso, in cui il comune non ha risposto in nessuno dei 12 mesi dell'anno, la selezione del donatore avviene estraendo casualmente un comune dall'insieme dei comuni rispondenti 12 mesi nello strato.

In entrambe le situazioni il donatore individuato viene utilizzato per imputare congiuntamente tutti i mesi mancanti al fine di preservare la stagionalità del fenomeno ovvero, più in generale, la sua autocorrelazione. Tra i vantaggi del metodo si segnala infatti l'assenza di distorsioni significative nella distribuzione del fenomeno.

Quando l'imputazione si riferisce a più di un anno è possibile scegliere tra l'imputazione separata per ciascun anno e quella su tutto il triennio considerato. Una volta effettuata la selezione del panel dei rispondenti nel triennio, le due diverse opzioni sono considerate come due diverse strategie di imputazione longitudinale.

3.2 Imputazione cross-section: la media

Tra i metodi di imputazione considerati per il confronto è stato scelto uno dei metodi deduttivi di più semplice implementazione, la media, consigliabile quando lo scopo dell'analisi è limitato alla stima di medie e totali (Grande e Luzi, 2003). L'applicazione del metodo è di tipo cross-section: per ogni mese vengono imputate le mancate risposte utilizzando solo le informazioni sui rispondenti dello stesso mese.

In particolare per ciascun mese, si calcola il valore medio della variabile di interesse, il numero di abitazioni, sul totale dei rispondenti. Il valore medio viene poi assegnato come valore ai comuni non rispondenti all'interno dello strato:

$$y_h^i = \frac{1}{n_h^r} \sum_{j=1}^{n_h^r} y_j$$

dove con y_h^i si indica la i-esima unità non rispondente nello strato h in un determinato mese. Come noto (ad esempio Grande Luzi, 2003) questo metodo ha tra i suoi vantaggi la semplicità di applicazione una volta definite le classi di imputazione ma provoca un'attenuazione della variabilità del fenomeno e una seria distorsione nella distribuzione della variabile.

3.3 Ponderazione

Nel caso longitudinale, per ogni unità, la mancata risposta in più di un periodo può essere interpretata come un insieme di mancate risposte parziali, suggerendo che, per ciascun periodo, la ponderazione possa essere la tecnica appropriata per la stima. In letteratura, la ponderazione dei dati è realizzata mediante un sistema di pesi che tiene conto della diversa propensione a rispondere dei comuni. In tal senso definendo, per ciascun mese t, con π_i la probabilità di risposta del generico comune i, tale probabilità viene stimata attraverso il seguente modello di regressione logistica:

$$logit[\pi_i(x)] = \log\left(\frac{\pi_i(\mathbf{x})}{1-\pi_i(\mathbf{x})}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \tag{3}$$

dove x_{i1} e x_{i2} rappresentano le variabili dicotomiche relative alle 3 ripartizioni geografiche, x_{i3} la variabile popolazione e x_{i4} una variabile dicotomica che assume valore 1 se il comune ha risposto al mese $t-12$ e valore 0 altrimenti.

Tuttavia, per ridurre la distorsione da mancata risposta, le probabilità stimate nel modello (3) non vengono utilizzate direttamente come pesi delle unità rispondenti ma si ricorre alla definizione di *classi di aggiustamento*. Se le classi sono formate in modo efficiente, le unità appartenente a ciascuna di esse presenteranno una probabilità di risposta omogenea. In analogia a quanto proposto in letteratura (Little, 1986, Eltinge e Yansaneh, 1997) per ridurre la variabilità dei coefficienti di ponderazione delle singole osservazioni le classi di aggiustamento vengono individuate in base ai $k, k = 2, \ldots, n$ quantili della distribuzione delle probabilità π_i. Quindi considerando k quantili la stima del totale di una variabile Y è data da:

$$\hat{Y}^k = \sum_{h=1}^{k+1} \frac{n_{s_h}}{r_{s_h}} \sum_{i \in s_h} y_i \tag{4}$$

dove n_{s_h} e r_{s_h} rappresentano, rispettivamente, il numero di unità ed il numero di rispondenti della generica classe s_h. Confrontando i valori della stima al crescere di k, è possibile sia individuare il valore di k oltre il quale l'aumento del numero di classi produce variazioni minime del totale stimato sia utilizzare uno dei test presentati da Eltinge e Yansaneh.

Nell'applicazione la stima del modello e quindi la scelta della soglia k, sulla base di una ripartizione in quantili, è realizzata per ogni mese.

4 La simulazione

La performance dei metodi di imputazione e ponderazione descritti nel paragrafo precedente è stata esplorata in simulazione. Prima dell'applicazione dei metodi si è proceduto alla identificazione di classi omogenee costruite a partire da variabili ausiliarie ovvero a partire dalla stima delle probabilità di non risposta. Questo aspetto viene affrontato nella prima parte del paragrafo mentre nella seconda parte si descrive lo schema della simulazione e nella terza i risultati.

4.1 L'identificazione delle celle di imputazione

Nello studio sono state individuate due variabili ausiliare, la popolazione e la ripartizione geografica. Per i due metodi di imputazione considerati, l'individuazione di classi omogenee per le due variabili è avvenuta in due step: la suddivisione della ripartizione geografica è avvenuta utilizzando i risultati del modello logistico stimato nel paragrafo 2 che, in base allo Schwartz criterion, identificava 3 sottoinsiemi (Nord-ovest, Nord-est, Centro e Mezzogiorno). Per la scelta di classi omogenee di popolazione, considerata come variabile continua nella stima del modello logistico, si è invece proceduto calcolando la varianza within tra i gruppi individuali. In particolare si indichi con y_{ij} il valore assunto dalla variabile numero di abitazioni residenziali del comune j-esimo appartenente all'i-esimo gruppo ($i = 1, \ldots, g; j = 1, \ldots, n_i$), con $\overline{y}_i$ la media del gruppo i e $\overline{y}$ la media generale:

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{5}$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} \tag{6}$$

Il problema di individuazioni delle classi si traduce nella scelta del numero di classi g e della loro numerosità interna n_i tale che i comuni appartenenti a ciascuna classe risultino il più possibile omogenei tra loro. Come criterio di omogeneità è stata considerata la variabilità entro i gruppi ovvero la minimizzazione della varianza within (SSW)

$$\frac{SSW}{n-g}$$

La variabile popolazione è una variabile quantitativa continua e, quindi, il problema in questo contesto è l'individuazione degli estremi $l_1, l_2, \ldots, l_h$ delle classi in cui suddividere l'intervallo 0-50.000 abitanti, condizionatamente alle 3 ripartizioni geografiche individuate in precedenza. La procedura è stata effettuata in maniera iterativa. In particolare, al primo stadio si sono considerate diverse possibili suddivisioni in due gruppi della popolazione, usando come valore di individuazione delle due classi $l_1 = 10.000,\ 20.000,\ 30.000$ e 40.000. Per $l_1 = 10.000$ si ha la maggiore riduzione della varianza within e viene perciò fissato come primo estremo.

Al passo successivo, si individuano i valori l_1 e l_2 che consentono di ottenere 3 classi di popolazione (9 strati in totale), suddividendo dapprima la classe $0 - 10.000$ usando vari valori soglia (con incrementi di 2.000 unità) e poi la classe $10.000 - 50.000$ usando differenti valori soglia con incrementi di 5.000. Come risulta dalla tabella 5, i valori di ripartizione in 3 gruppi individuati sono 7.000 e 25.000. Tale ripartizione produce una riduzione della varianza within rispetto alla ripartizione in due gruppi. Ulteriori diminuzioni si ottengono nel passaggio a 3,

Tabella 5: *Classi di popolazione*

l_h					g	WITHIN
-	-	-	-	-	1	1.854,63
-	-		10.000		2	1.320,13
-		7.000	-	25.000	3	1.156,71
-	3.000	7.000		25.000	4	1.109,50
-	3.000	7.000	13.000	25.000	5	1.086,96
1.500	3.000	7.000	13.000	25.000	6	1.082,84

4, 5, 6 gruppi. Tuttavia nel passaggio da 5 a 6 gruppi, anche se la varianza within continua a diminuire, vengono individuati strati con poche unità, inutilizzabili per imputare più dati. Ai fini dell'imputazione i comuni vengono quindi raggruppati in 15 classi omogenee identificate da 5 classi di popolazione e 3 ripartizioni geografiche.

Per la ponderazione, le classi omogenee basate sulla probabilità di mancata risposta sono costruite a partire dalle stime del modello logit individuate in precedenza in cui, per la popolazione, si sono utilizzate le 5 classi di cui sopra.

Tutti i metodi utilizzano quindi la stessa informazione ausiliaria trattandola in modo differente.

4.2 Lo schema della simulazione

Per valutare le diverse strategie proposte per il trattamento della mancata risposta si è condotto uno studio simulando i valori mancanti: ciascun metodo di imputazione è applicato ad un insieme di valori reali osservati nei quali la mancata risposta è creata in modo artificiale mediante un meccanismo di tipo MAR. Ogni simulazione è quindi individuata da un ciclo comprendente la generazione dei dati mancanti, l'applicazione dei metodi di imputazione ed il calcolo degli indicatori per valutare le differenti strategie di imputazione.

Tabella 6: *Comuni rispondenti 36 mesi*

Popolazione	Nord-Ovest	Nord-Est	Centro	Sud-Isole	Totale
$x \leq 3.000$	606	212	39	83	940
$3.000 < x \leq 7.000$	192	150	45	64	451
$7.000 < x \leq 13.000$	87	107	31	51	276
$13.000 < x \leq 25.000$	50	64	17	26	157
$x > 25.000$	21	20	22	23	86
Totale	956	553	154	247	1.910

Per tenere conto della variabilità del meccanismo di mancata risposta ed evitare quindi possibili effetti dovuti alla selezione di un particolare campione sono state realizzate 50 simulazioni. Per ognuna delle 50 simulazioni abbiamo considerato il panel dei 1.910 comuni rispondenti 36 mesi. Per tali comuni si conoscono, quindi, sia il numero di abitazioni per ciascun mese dell'anno sia le variabili ausiliari, popolazione residente al 31 dicembre 1999 e ripartizione geografica.

Il panel dei comuni selezionati presenta, coerentemente con i risultati sulla collaborazione illustrati nel paragrafo 2, una forte asimmetria nella distribuzione per ripartizione geografica con una significativa prevalenza di comuni del Nord (tabella 6). Le simulazioni effettuate non hanno,

quindi, permesso di individuare la partizione geografica ottimale secondo gli indicatori utilizzati per la valutazione dei metodi di imputazione. Questo aspetto è stato esplorato nella applicazione confrontando i risultati ottenuti nelle diverse ipotesi di ripartizione con quelli stimati tramite ponderazione. Nella simulazione sono state, quindi, utilizzate le 15 celle di imputazione definite precedentemente.

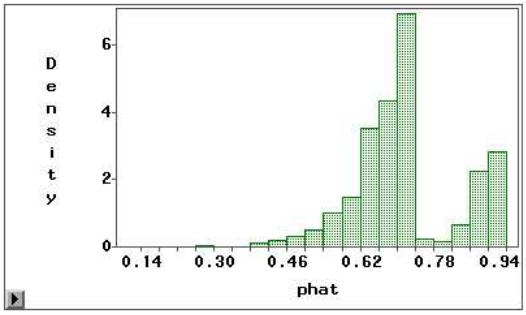
4.2.1 La generazione dei dati mancanti

Al fine di riprodurre un pattern di mancata risposta simile a quello osservato nel dataset di partenza, la generazione dei dati mancanti procede in due differenti passi: selezione casuale di un campione di comuni da rendere non rispondenti e selezione di un pattern infrannualedi mancata risposta.

Al primo passo, considerando le distribuzioni in tabella 1, è stato selezionato un campione casuale di 1.450 comuni (corrispondenti al 76% dei comuni appartenenti al panel) sui quali generare la mancata risposta. La selezione è stata effettuata con una probabilità di selezione proporzionale alla probabilità di mancata risposta stimata mediante il modello di regressione logistica del paragrafo 2. Come illustrato dalla tabella 7, nella quale si riportano i risultati di una singola simulazione, questa procedura ha permesso di massimizzare, tra i comuni da imputare, la presenza di quelli con un alto valore della probabilità di mancata risposta.

Tabella 7: *Distribuzione della probabilità di mancata risposta*

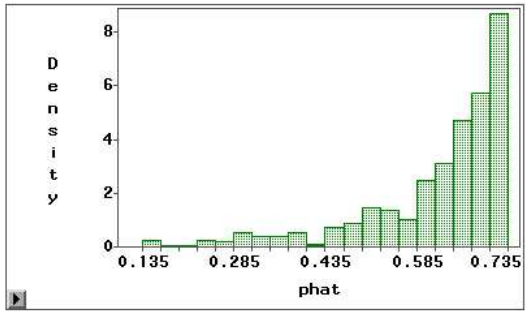
Non rispondenti



Moments			
N	1430.0000	Sum Wgts	1430.0000
Mean	0.7129	Sum	1019.4110
Std Dev	0.1195	Variance	0.0143
Skewness	-0.1151	Kurtosis	0.7171
USS	747.1219	CSS	20.4094
CV	16.7643	Std Mean	0.0032

Quantiles			
100% Max	0.9176	99.0%	0.9154
75% Q3	0.7257	97.5%	0.9141
50% Med	0.7050	95.0%	0.9112
25% Q1	0.6495	90.0%	0.9029
0% Min	0.1597	10.0%	0.5819
Range	0.7579	5.0%	0.5278
Q3-Q1	0.0762	2.5%	0.4678
Mode	0.7195	1.0%	0.4051

Rispondenti



Moments			
N	480.0000	Sum Wgts	480.0000
Mean	0.6177	Sum	296.4996
Std Dev	0.1219	Variance	0.0149
Skewness	-1.6805	Kurtosis	2.4593
USS	190.2698	CSS	7.1199
CV	19.7372	Std Mean	0.0056

Quantiles			
100% Max	0.7278	99.0%	0.7239
75% Q3	0.7067	97.5%	0.7231
50% Med	0.6658	95.0%	0.7217
25% Q1	0.5821	90.0%	0.7190
0% Min	0.1445	10.0%	0.4507
Range	0.5833	5.0%	0.3277
Q3-Q1	0.1246	2.5%	0.2743
Mode	0.5239	1.0%	0.1874

Per il campione di comuni selezionati, sono stati cancellati i dati relativi ai mesi di ciascun anno in modo tale che la distribuzione percentuale dei comuni con collaborazione da 0 a 35 mesi sia uguale a quella osservata nell'universo. Per esempio, nel triennio dal 2000 al 2002 il $12,4\%$ dei comuni ha avuto una collaborazione pari a 35 mesi. Quindi, per 234 comuni (il $12,4\%$ dei 1.910 comuni selezionati al passo precedente) è stato cancellato un mese mediante la scelta casuale di un profilo di risposta 0 e 1 tra i 987 comuni dell'universo che hanno collaborato 35 mesi.

4.2.2 Gli indicatori per la valutazione

Per ciascuna simulazione è possibile identificare degli indicatori per la valutazione delle diverse strategie di stima delle mancate risposte. L'insieme degli indicatori è ampio nel caso del confronto tra i tre metodi di imputazione, donatore longitudinale (sul singolo anno o sul triennio) e media: è possibile valutare le differenze sia tra le stime degli aggregati (media, varianza) sia di quelle tra i microdati. Il confronto con la ponderazione può, invece, avvenire solo in termini della stima del totale mensile ovvero delle sue variazioni nel tempo attraverso il calcolo delle variazione tendenziale e della autocorrelazione.

In particolare, indicando con $\mathbf{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)$ e con $\bar{\mathbf{Y}} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n)$ i vettori dei valori, rispettivamente, veri e imputati della variabile Y, una prima misura della bontà del metodo di imputazione può essere valutata calcolando alcuni momenti delle due distribuzioni quali la media e la varianza.

Inoltre, per valutare la vicinanza tra le due distribuzioni, si sono calcolate le distribuzioni empiriche:

$$F_{\bar{Y}_n}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\bar{Y}_i \leq t) \tag{7}$$

e

$$F_{Y_n^*}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i^* \leq t) \tag{8}$$

si è misurata la loro differenza mediante la distanza di Kolmogorov-Smirnov:

$$d_{KS}(F_{\bar{Y}_n}, F_{Y_n^*}) = max_t(|F_{\bar{Y}_n}(t) - F_{Y_n^*}(t)|) \tag{9}$$

Utilizzando i microdati è possibile calcolare la distanza tra i valori veri e quelli imputati per ciascun comune. Tuttavia la presenza di una alta percentuale di valori pari a 0 non rende possibile considerare alcuni dei metodi proposti in letteratura, quali, ad esempio, la regressione dei valori veri sui valori imputati. Per i confronti è stata scelta la seguente distanza:

$$d_1(\mathbf{Y}^*, \bar{\mathbf{Y}}) = \frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_i - Y_i^*| \tag{10}$$

I risultati ottenuti con la ponderazione sono stati valutati attraverso il calcolo del MAPE (*Mean Absolute Percentage Error*). Considerando tutti gli n_2 valori del dataset iniziale e, non solamente, gli n valori imputati, è stata calcolata la statistica:

$$d_2 = 100 * \frac{|\sum_{i=1}^{n_2} \bar{Y}_i - \sum_{i=1}^{n_2} Y_i^*|}{\sum_{i=1}^{n_2} Y_i^*} \tag{11}$$

Oltre a questi indicatori, tradizionalmente utilizzati in ambito cross-section, la disponibilità di dati aggregati per tutti i mesi del triennio, rende possibile il confronto tra i metodi sia in termini di variazioni tendenziali sia di calcolo della funzione di autocovarianza.

4.3 I risultati

Gli indicatori presentati nella sezione precedente sono stati calcolati a partire dai risultati stimati in ciascuna delle 50 simulazioni realizzate. Per i tre metodi di imputazione, longitudinale annuale, longitudinale sul triennio e media, sono stati rispettivamente calcolati: i) le distanze tra

la media (varianza) stimata e quella vera; ii) la distanza d_1; iii) il test di Kolmogorov Smirnov. In tutti e tre i casi gli indicatori sono elaborati per ogni mese e per ciascuna simulazione identificando una matrice di 36x50 elementi la cui rapprentazione grafica è stata ottenuta calcolando la media dei valori per il triennio. La distribuzione dei 50 valori ottenuti è stata rappresentata attraverso il grafico boxplot. Dato un vettore di dati, attraverso la rappresentazione boxplot è possibile identificare immediatamente i quartili, l'estensione della distribuzione e gli eventuali outlier.

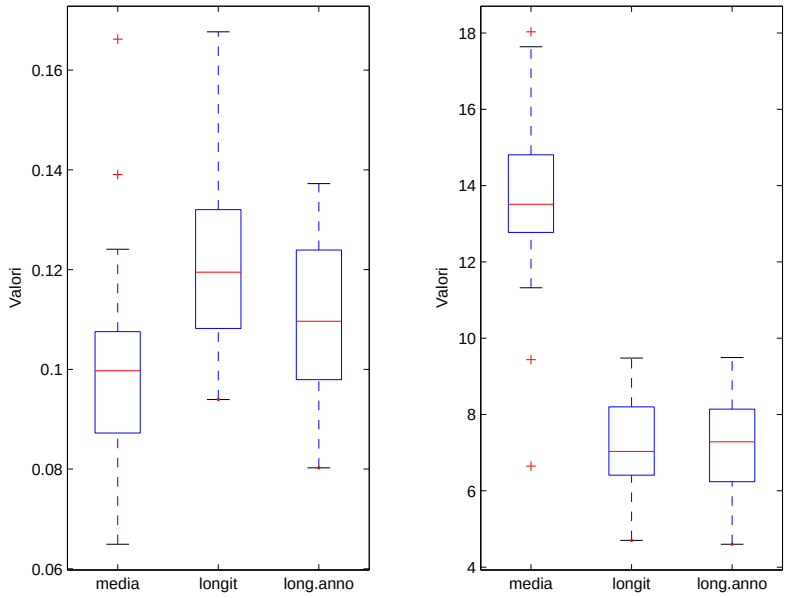
I risultati della ponderazione sono stati confrontati calcolando la distanza d_2, la variazione tendenziale e la funzione di autocovarianza. Per l'analisi della variazione tendenziale e dell'autocorrelazione, le distribuzioni si riferiscono al coefficiente di correlazione tra la dinamica espressa dai dati *veri* ed in quelli imputati ovvero ponderati.

Nella presentazione dei risultati l'imputazione attraverso la media è indicata con *media*, con *longit* si indica l'imputazione attraverso il donatore su 36 mesi ed infine con *long.anno* il donatore longitudinale sul singolo anno. La ponderazione è indicata con *ripond*.

Metodi di imputazione

Nella figura 2 si riportano i risultati del confronto tra le medie e le varianze stimate. Per ogni simulazione e per ogni mese è stato calcolato il valore assoluto della differenza tra la media della popolazione e quella stimata con i diversi metodi. Come noto, l'imputazione attraverso il metodo *media* tende a riprodurre il valore medio della popolazione con più precisione rispetto ai due metodi del donatore; risultati opposti si ottengono nel confronto tra le varianze. A parità di performance nel confronto tra le varianze, il donatore *long.anno* mostra degli scostamenti più contenuti del donatore *longit* in termini di media.

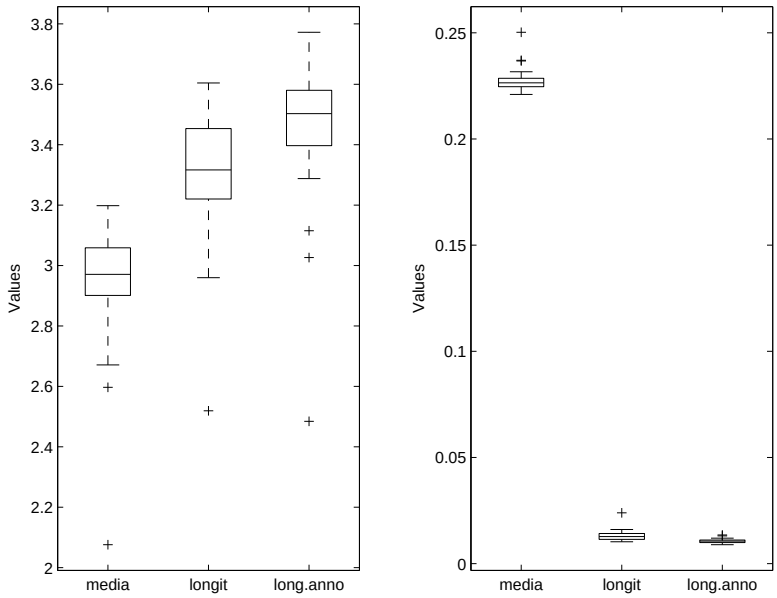
Figura 2: *Scostamenti dalla media (a) e dalla varianza (b) per simulazione*



Dal panel b) della figura 3 risulta evidente come il metodo di imputazione mediante la media alteri le caratteristiche della distribuzione dei dati osservati. Infatti la distanza di Kolmogorov-Smirnov è di gran lunga maggiore delle distanze calcolate per i due metodi con il donatore,

pressochè equivalenti tra loro (con una minore variabilità nel caso di *long.anno*). Come atteso la media presenta risultati migliori se si considera la distanza d_1.

Figura 3: *Distanza d_1 (a) e Kolmogorov-Smirnov (b) per simulazione*



Ponderazione

I risultati ottenuti mediante il donatore longitudinale sono stati confrontati con un metodo di ponderazione in termini di stima dei valori aggregati mensili.

Come già descritto, la ponderazione prevede la costruzione di classi di aggiustamento definite a partire dai quantili della distribuzione della probabilità di risposta. In tabella 9 sono riportati, per una delle simulazioni, le stime ottenute al variare del numero di classi riportato nella prima colonna: per ciascuna delle simulazioni, è stato scelto un numero di classi pari a 8, valore per cui la stima mostra una maggiore stabilità.

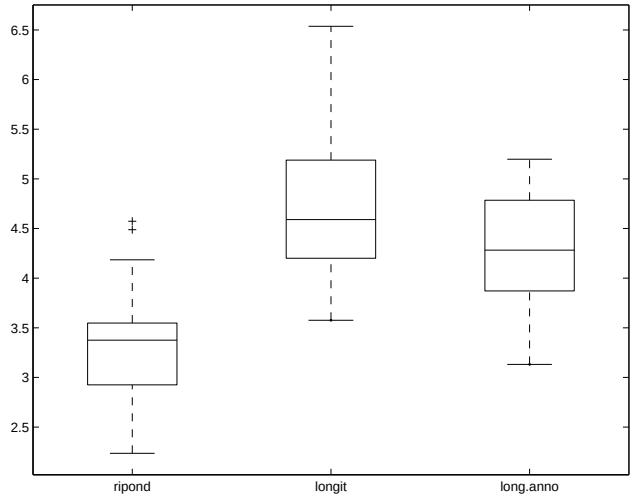
In figura 4 sono riportati i boxplot per la distanza d_2. La ponderazione *ripond* ha un comportamento leggermente migliore rispetto all'imputazione *long.anno*. I risultati ottenuti con il metodo *longit* sono peggiori sia in termini di mediana sia per la presenza di valori particolarmente elevati.

L'analisi aggregata è stata condotta anche sulle variazioni tendenziali e sull'autocorrelazione. Per ogni simulazione è stata calcolata, per ogni metodo, la correlazione tra la variazione tendenziale e l'autocovarianza stimata e quella vera (figura 5). Per tutti e tre i metodi si ottengono ottimi risultati.

5 Applicazione

Il metodo longitudinale annuale, preferito a quello longitudinale triennale, è stato applicato per la stima del numero di nuove abitazioni per i 7.940 comuni non capoluogo con meno di 50.000 abitanti rilevati dalla indagine dell'attività edilizia. Poiché il processo di autoselezione

Figura 4: *Distanza d_2 per simulazione*



dei comuni utilizzati nel panel per le simulazioni non ha reso possibile confrontare tra loro diverse classificazioni del territorio, questo aspetto è stato approfondito nell'applicazione. Data la suddivisione della popolazione in 5 classi, l'imputazione è stata realizzata ipotizzando 3 diverse aggregazioni delle ripartizioni geografiche: 2 ripartizioni, isolando il Nord-est; 3 ripartizioni, Nord-est, Nord-ovest, Centro e Mezzogiorno; 4 ripartizioni, separando il Centro dal Mezzogiorno.

In tabella 8 sono riportati i totali annuali per il 2000, il 2001 ed il 2002. Le variazioni annuali, mostrano, nel periodo 2000-01 risultano simili tra l'ipotesi di 10 strati (4,7%) quella a 20 strati (4,8%) e la ponderazione (4,9%) mentre nell'ipotesi a 15 strati la variazione è più elevata (6,2%); nel periodo 2001-02 la crescita segnalata dalla ponderazione (10,7%) è meno elevata di quella segnalata dai diversi criteri di ponderazione che oscillano dal $11,2\%$ (ponderazione a 20 strati) al $12,0\%$. In termini di livello la stima ottenuta con la ponderazione è più elevata di quella ottenuta con l'imputazione. Le differenze risultano più contenute per i 10 strati (rispettivamente 1,5%, 1,7% e 0,5% nel 2000, 2001 e 2002) rispetto ai 15 strati (rispettivamente 4,3%, 3,0% e 2,0%) ed ai 20 strati (rispettivamente 3,8%, 3,8% e 3,4%).

Tabella 8: *Stima del numero di abitazioni: donatore e ponderazione*

	Valori assoluti				**Variazioni**			
	10 strati	15 strati	20 strati	Ponder.	10 strati	15 strati	20 strati	Ponder.
2000	144.595	140.562	141.258	146.850	-	-	-	-
2001	151.335	149.322	148.059	153.976	4,7	6,2	4,8	4,9
2002	169.543	167.016	164.628	170.439	12,0	11,8	11,2	10,7

In figura 6, invece, sono riportate le variazioni tendenziali calcolate sui valori trimestrali.

Le variazioni tendenziale ottenute attraverso l'imputazione sono assimilabili a quelle ottenute con la ponderazione, anche se la variazione del IV trim. 02, ottenuta utilizzando 15 strati si discosta dalle altre.

A seguito dei risultati ottenuti nell'applicazione si è scelto di utilizzare la suddivisione in 10 strati per l'imputazione delle nuove abitazioni in edilizia nel periodo 2000-2002. Questa

Figura 5: *Variazioni tendenziali (a) e autocorrelazione (b) per simulazione*

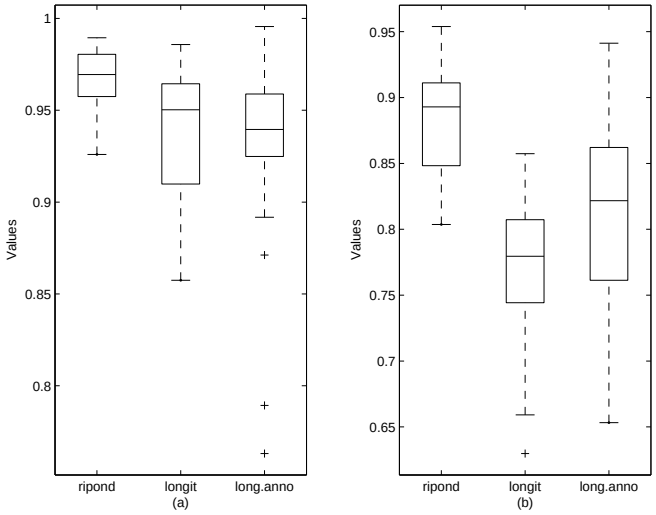
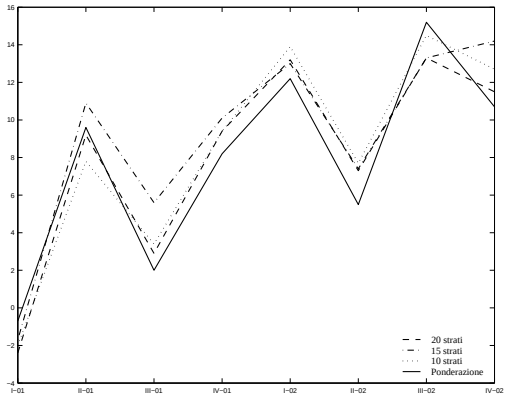


Figura 6: *Variazioni tendenziali*



strategia offre la possibilità di ricostruire i microdati garantendo risultati aggregati in linea con quelli ottenuti con la ponderazione.

6 Conclusioni

In questo lavoro è stato presentato un metodo di imputazione longitudinale basato sul donatore per la stima delle mancate risposte del numero di abitazioni mensili rilevate mediante l'indagine sui permessi di costruzione. Le sue performance sono state valutate rispetto ad un altro classico metodo di imputazione cross-section ed una tecnica di ponderazione. Il metodo proposto preserva la distribuzione dei valori ed è in grado di approssimare la variazione tendenziale e la autocorrelazione ottenute attraverso la ponderazione. Tale metodo è stato utilizzato per il rilascio dei dati annuali dell'Attività Edilizia per gli anni 2000, 2001 e 2002.

Tabella 9: *Stima del numero di abitazione mediante ponderazione al variare del numero di classi: Anni 2000-2002*

Valore stimato - Valore vero														$Valorestimato_c$ - $Valorestimato_{c-1}$											
	2000													**2000**											
celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic	celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic
1	375	555	502	604	292	749	293	426	671	696	579	232	2	-283	-341	-293	-348	-355	-362	-292	-254	-382	-429	-391	-295
2	92	215	209	256	-63	387	1	172	289	268	188	-63	3	2	-15	-66	6	-8	31	-48	-5	-48	-47	-7	-3
3	93	200	144	262	-71	418	-48	167	241	220	181	-66	4	8	-27	-3	-19	-14	-43	3	-29	13	-2	36	-9
4	102	172	140	243	-86	375	-45	138	254	218	217	-75	5	-33	-49	-61	-34	-62	-120	-29	-50	-59	-84	-93	-8
5	69	123	79	209	-147	255	-74	87	194	134	124	-83	8	-44	-15	-41	-31	-27	12	-40	-24	-27	-52	19	-46
8	25	109	38	178	-174	267	-114	63	167	82	143	-129	9	4	14	17	-3	-12	8	-32	-8	-36	5	-37	1
9	29	123	55	175	-186	275	-147	55	131	87	106	-128	10	-15	-23	-27	-18	-19	-59	1	-34	16	-7	-37	21
10	14	99	28	156	-205	216	-146	21	147	79	70	-107	15	-5	-5	-19	-20	11	-11	-27	2	-13	-25	38	-14
15	9	94	9	136	-194	205	-172	22	134	55	108	-121													
	2001													**2001**											
celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic	celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic
1	570	310	474	641	1060	677	752	616	731	510	824	616	2	-341	-289	-301	-352	-457	-465	-439	-329	-432	-380	-456	-392
2	229	20	173	289	603	211	313	287	299	131	368	224	3	33	-7	-55	-53	-41	-17	-37	15	-86	-1	-6	-50
3	263	14	118	236	562	194	277	303	213	130	362	174	4	-78	-33	-21	-18	15	-74	-31	-53	-32	-53	-110	-26
4	185	-19	96	218	576	120	245	249	181	77	251	148	5	-21	-31	-35	-67	-58	-63	-106	-58	-91	-110	-75	-74
5	164	-50	61	151	519	57	140	192	90	-33	176	74	8	-57	-45	-83	-128	-100	-109	-114	-47	-39	-43	-59	-50
8	107	-95	-21	23	419	-52	25	145	51	-76	118	24	9	-40	22	17	66	-12	49	19	-15	17	37	-31	-2
9	67	-73	-4	89	407	-3	44	130	68	-39	86	22	10	39	-18	-7	-5	0	14	-4	24	-33	-56	5	35
10	106	-90	-12	84	407	11	40	154	35	-95	92	57	15	-21	-2	-40	-68	1	-95	-48	-22	-43	9	-6	-37
15	85	-92	-52	16	408	-84	-8	132	-8	-86	86	20													
	2002													**2002**											
celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic	celle	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic
1	502	583	616	855	1175	1301	1531	219	1066	1094	1313	1164	2	-322	-412	-507	-419	-592	-599	-710	-301	-565	-592	-720	-507
2	179	171	109	436	583	702	820	-83	502	501	593	657	3	-43	-7	20	-29	3	-40	-39	8	-9	-77	15	-27
3	136	164	130	407	586	663	781	-74	493	425	608	630	4	-54	-88	-79	-17	-97	-73	-56	-49	-90	-67	-132	-44
4	82	76	50	391	489	590	725	-123	403	358	477	586	5	-44	-49	-23	-74	-77	-129	-146	-34	-91	-51	-32	-107
5	38	26	27	317	412	461	579	-157	312	307	445	479	8	-47	-35	-128	-29	-22	-58	-96	-45	-75	-21	-113	-90
8	-9	-9	-101	288	390	403	483	-202	237	286	332	390	9	-20	-15	8	-51	-79	25	-44	21	0	-57	-27	28
9	-29	-24	-92	237	310	428	438	-181	237	229	305	418	10	18	33	51	7	37	-44	37	-14	4	18	23	-16
10	-11	9	-41	244	347	384	476	-195	241	246	328	401	15	-25	-51	-38	-10	-60	-43	-42	3	-35	-6	-49	-11
15	-36	-42	-79	235	287	341	433	-192	206	240	279	390													

Riferimenti bibliografici

[1] Chambers, R. (2001). Evaluation criteria for statistical editing and imputation *National Statistics Methodological Series*, n. 28, National Statistics, London.

[2] Chen, J. e Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 113- 131.

[3] Grande, E. e Luzi, O. (2003). Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat. *Contributi Istat*, n. 6, Istituto Nazionale di Statistica, Roma.

[4] Eltinge, J.L. e Yansaneh, I.S. (1997). Diagnostic for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, 23, 33- 40.

[5] Heeringa, G.S. e Lepkowski (1986). Longitudinal imputation for the SIPP. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 206- 219.

[6] Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303- 314.

[7] Kalton, G. e Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1- 16.

[8] Little, R.J.A. e Rubin, D.B. (1987). *Statistical analysis with missing data.* John Wiley & Sons, New York.

[9] Manzari, A. (2004). Valutazione Comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi. *Contributi Istat*, n. 18, Istituto Nazionale di Statistica, Roma.

[10] Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New York.

[11] Schafer, J.L. (1987). *Analysis of incomplete multivariate data.* Chapman and Hall, New York.

[12] Vartivarian, S. e Little, R. (2003). On the formation of weigthing adjustment cells for unit nonresponse. *Department of Biostatistics Working Paper Series*, n. 10, University of Michigan.

[13] Zhang, L.- C. (2001). A method of weighting adjustment for survey data subject to non-ignorable nonresponse *Statistics Norway Discussion Papers*, n. 311, Statistics Norway, Kongsvinger.