Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento Generale della Popolazione 2001

L. Di Consiglio, S. Falorsi
ISTAT

1. Introduzione

Le operazioni di conteggio effettuate dal Censimento possono essere soggette ad errori di copertura. Con l'indagine di copertura ci si prefigge di misurare tali errori, effettuando nuovamente il conteggio su un campione di sezioni censuarie. Il presente documento descrive le caratteristiche generali del disegno di campionamento dell'indagine di copertura del censimento generale della popolazione dell'anno 2001. In particolare viene delineata la prima fase di progettazione conclusasi nel giugno 2000 con l'estrazione del campione di primo stadio.
Si sottolinea che la progettazione della prima fase si è conclusa quando non erano ancora disponibili tutte le informazioni sulle nuove sezioni definite per il Censimento 2001; di questo si è dovuto tener conto nella definizione del disegno. D'altra parte per esigenze amministrative si rendeva necessario informare con largo anticipo i comuni della loro partecipazione all'indagine post-censuaria e pertanto non si poteva attendere che le sezioni censuarie fossero completamente formate
Il lavoro si articola come di seguito descritto. Nel paragrafo 2 vengono definiti, da un punto di vista generale, gli obiettivi di un'indagine di copertura e ne viene data una formalizzazione secondo il modello probabilistico di Petersen; vengono, quindi, precisati i domini di interesse specifici dell'indagine di Copertura del Censimento Generale della Popolazione 2001. Nel paragrafo 3 viene descritto il disegno campionario adottato tenendo conto dei vincoli operativi e di costo e delle informazioni ausiliarie disponibili. Viene, inoltre, descritta l'analisi che ha condotto alla scelta delle suddette variabili ausiliarie. Infine, sono illustrati i passi che hanno portato alla determinazione della numerosità campionaria complessiva e alla sua allocazione tra i domini territoriali di interesse.

2. Obiettivi dell'indagine di copertura

2.1 Definizione dei parametri di interesse

Obiettivo principale di un'indagine di copertura è fornire una valutazione dell'accuratezza delle operazioni di enumerazione censuaria delle unità della popolazione, ad esempio, facendo riferimento al Censimento Generale della Popolazione, la finalità dell'indagine di copertura è quella di valutare gli errori per eccesso (sovracopertura) o per difetto (sottocopertura) del conteggio di famiglie, individui residenti, individui presenti e abitazioni effettuato al censimento.

In particolare in questo documento l'attenzione è focalizzata sull'errore di sottocopertura. Il principale indice per la valutazione dell'accuratezza, in tal caso, è rappresentato dal *tasso di copertura* che, nell'ipotesi di assenza di *sovracopertura*, è ottenuto come rapporto tra il numero di unità enumerate al censimento e la numerosità effettiva N (incognita) della popolazione; equivalentemente il *tasso di sottocopertura* è dato dal rapporto tra il numero di unità sfuggite all'enumerazione censuaria e la numerosità effettiva N della popolazione. Poiché la numerosità effettiva della popolazione è incognita, al fine di poter valutare concretamente i suddetti tassi occorre formalizzare con un modello probabilistico la mancata enumerazione delle unità (che può verificarsi sia al censimento che all'indagine di copertura). Tali modelli sono caratterizzati da ipotesi più o meno stringenti. Nel presente lavoro si fa riferimento al modello di *omogeneità entro*

le liste o di Petersen che, opportunamente adattato, costituisce una ragionevole approssimazione della realtà.

Il modello di Petersen

Al fine di rendere chiara la seguente esposizione, è utile fare riferimento al caso del tutto ipotetico in cui anche l'indagine di copertura sia una rilevazione di tipo censuario e ripeta, pertanto, le operazioni censuarie su tutte le sezioni dell'intero territorio nazionale. Il censimento della popolazione e l'indagine di copertura danno allora luogo a due liste distinte della popolazione di interesse, entrambe soggette ad errore di copertura.

Indichiamo con C la lista ottenuta con il censimento e con I quella ottenuta con l'indagine.

Per la valutazione della numerosità *N* della popolazione è possibile utilizzare le due liste integrando l'informazione fornita da entrambe per mezzo di un modello di rappresentazione dell'errore di copertura.

Il modello di Petersen è caratterizzato dalle seguenti assunzioni (Wolter, 1986):

1. la popolazione di riferimento è chiusa e di dimensione fissata pari a *N* ;

2. l'evento che l'unità *j* appartenga o meno alla lista C e appartenga o meno alla lista I può essere rappresentato mediante una distribuzione multinomiale le cui probabilità congiunte e marginali sono riportate nel prospetto seguente:

		Lista I		
		Sì	*No*	
Lista C	*Sì*	p_{j11}	p_{j12}	p_{j1+}
	No	p_{j21}	p_{j22}	p_{j2+}
		p_{j+1}	p_{j+2}	1

3. le due liste C e I, possono essere considerate il risultato di *N* prove indipendenti dell'esperimento multinomiale del punto 2; per ogni singola unità *j* si definisce una variabile x_{jab} pari ad uno se l'unità *j* cade nella cella *ab* e zero altrimenti (*a*,*b*=1,2); per l'insieme delle unità della popolazione la situazione può essere rappresentata come segue:

		Lista I		
		Sì	*No*	
Lista C	*Sì*	x_{11}	x_{12}	x_{1+}
	No	x_{21}	x_{22}	x_{2+}
		x_{+1}	x_{+2}	*N*

dove

$$x_{ab} = \sum_{j=1}^{N} x_{jab} \ (a,b=1,2),\ x_{a+} = \sum_{j=1}^{N}\sum_{b=1}^{2} x_{jab}\ , (a=1,2) \text{ e } x_{+b} = \sum_{j=1}^{N}\sum_{a=1}^{2} x_{jab}\ , (b=1,2)$$

sono rispettivamente il numero di unità della cella *ab*, e delle celle marginali *a*+ e +*b*; ovviamente le quantità x_{22} e *N* non sono osservabili;

4. è possibile determinare senza errore quali unità registrate nella lista I sono o meno presenti nella lista C, ossia non sono presenti errori di *abbinamento*;

5. le liste sono depurate da errori di registrazione e duplicazione;

6. le probabilità che le unità siano incluse nella lista C, p_{j1+} (j=1,…, N) e le probabilità che le unità siano incluse nella lista I, p_{j+1} (j=1,…, N), sono costanti per ciascuna lista, ossia soddisfano le condizioni $p_{j1+}=p_{1+}$ e $p_{j+1}=p_{+1}$. D'altra parte le due probabilità p_{1+} e p_{+1} possono essere differenti.

Sulla base delle ipotesi precedenti il tasso di copertura[1] è stimato da

$$\tilde{\tau} = (x_{11} / x_{+1}), \tag{1}$$

mentre una stima della numerosità[1] della popolazione è data da

$$\tilde{N} = x_{1+} + x_{1+}(x_{21} / x_{11}) = \frac{x_{1+} x_{+1}}{x_{11}}. \tag{2}$$

Poiché con l'indagine post-censuaria non si effettua un'enumerazione completa su tutte le sezioni censuarie ma solo su un campione di esse, le quantità x_{11} e x_{+1} sono note solo per le sezioni campione. Le precedenti relazioni (1) e (2) possono essere, quindi, adottate a livello di ciascuna sezione campione i. Si noti che le quantità x_{11i} e x_{+1i} sono osservabili essendo l'operazione di conteggio completa entro le sezioni campione.
Il tasso di copertura τ_i della sezione i è, quindi, stimato da

$$\tilde{\tau}_i = (x_{11i} / x_{+1i}).$$

Per quanto riguarda, invece, il tasso di copertura riferito ad un generico dominio di interesse (ad esempio, l'intero territorio nazionale, la regione o la classe di dimensione del comune), si possono stimare le quantità a numeratore e denominatore della (2) sulla base delle osservazioni campionarie. Si ha, quindi, che

$$\hat{\tau} = (\hat{x}_{11} / \hat{x}_{+1}), \tag{3}$$

in cui

$$\hat{x}_{11} = \sum_i x_{11i}\, k_i \quad \text{e} \quad \hat{x}_{+1} = \sum_i x_{+1i}\, k_i$$

sono rispettivamente le stime campionarie delle corrispondenti quantità x_{11} e $x_{+1,}$ essendo k_i il peso finale attribuito alla sezione campione i.
Analogamente la quantità (2) può essere stimata con

$$\hat{N} = \frac{x_{1+}\ \hat{x}_{+1}}{\hat{x}_{11}}. \tag{4}$$

2.2. Domini di interesse

L'indagine in esame ha la finalità principale di stimare il tasso di copertura a livello dell'intero territorio nazionale, delle cinque ripartizioni geografiche (Nord Ovest, Nord Est, Centro, Sud e Isole) e delle seguenti quattro classi di dimensione demografica dei comuni:

[1] Vedi Appendice A1.

- meno di 10.000 abitanti;
- tra 10.000 e 100.000 abitanti;
- oltre 100.000 esclusi i comuni metropolitani;
- comuni metropolitani: Torino, Genova, Milano, Venezia, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari.

L'indagine deve fornire anche stime per ciascuna classe determinata dall'intersezione delle modalità delle due variabili precedenti, ossia collocazione geografica e dimensione demografica dei comuni.
Sono, infine, d'interesse le stime del tasso di copertura riferite sia alle caratteristiche delle famiglie (ad esempio l'ampiezza della famiglia) che degli individui (ad esempio il sesso), che per domini territoriali maggiormente disaggregati (ad esempio le regioni e l'insieme di comuni con dimensione demografica vicina alle soglie di popolazione superate le quali il comune ottiene benefici di natura economica[2]) anche se queste ultime rappresentano un obiettivo secondario rispetto ai precedenti.
Si noti che mentre è possibile controllare la dimensione campionaria per i domini territoriali di maggiore interesse, quali la ripartizione e la dimensione demografica dei comuni, non è possibile determinare a priori il numero di unità campionarie e quindi la precisione delle stime per altri tipi di dominio legati alle caratteristiche delle famiglie e degli individui ad esse appartenenti. Di questo si potrà tenere conto nella definizione dello stimatore mediante opportune tecniche di post-stratificazione.

3. Disegno di campionamento

3.1 Caratteristiche generali

L'enumerazione censuaria fa riferimento al territorio che, per esigenze organizzative, viene ripartito in aree aventi determinate caratteristiche (dette sezioni di censimento). L'indagine di copertura, volta a valutare il conteggio effettuato dal censimento, ripete le operazioni di enumerazione su un campione casuale di dette sezioni che costituiscono, pertanto, le unità finali di campionamento. Non sarebbe, infatti, possibile definire come unità finali altre aree, quali segmenti o parti di sezione, poiché non si dispone di informazioni su aree all'interno delle sezioni, che sono indispensabili per effettuare l'abbinamento tra i dati del censimento con i dati rilevati dall'indagine di copertura (cfr par. 2.1).
Il disegno di campionamento adottato è a due stadi di selezione in cui al primo stadio si seleziona un campione di comuni e al secondo stadio - nell'ambito di ciascun comune selezionato al primo stadio – si seleziona un campione di sezioni, tutte le famiglie appartenenti alle sezioni campione vengono enumerate.
La scelta di selezionare i comuni al primo stadio è stata determinata dalla necessità di tenere sotto controllo, per esigenze organizzative e di costo, il numero di comuni coinvolti nella rilevazione; non si è ritenuta praticabile la scelta alternativa di selezionare direttamente un campione di sezioni censuarie perché ciò avrebbe comportato il coinvolgimento di un numero troppo elevato di comuni campione.
Il disegno proposto prevede, inoltre: la stratificazione delle unità di primo stadio; la stratificazione delle unità di secondo stadio; la selezione dei comuni campione senza reimmissione e con probabilità proporzionali all'ampiezza demografica e la selezione delle sezioni campione con probabilità uguali e senza reimmissione.

3.2. Scelta delle variabili di stratificazione

[2] si veda M. Dimitri '98.

I comuni sono ripartiti in venti strati T_1, K, T_h, K, T_H, (H=20), costituiti dalle classi formate dall'intersezione delle modalità delle due variabili, ripartizione geografica e classe di dimensione demografica, che costituiscono i principali domini di interesse descritti nel paragrafo 2.2. Definendo adeguate numerosità campionarie negli strati così definiti, è possibile garantire stime aventi prefissati livelli di precisione attesi per ciascuno dei principali domini di interesse.

Per la definizione della stratificazione dei comuni e delle sezioni all'interno di ogni T_h (h=1,…,20) è stata condotta un'analisi volta a valutare la relazione tra il tasso di copertura (desunto dall'indagine di copertura del Censimento della Popolazione 1991) con alcune variabili ottenute sia dal censimento 1991 che dalle indagini ISTAT sui dati di qualità delle anagrafi comunali dell'anno 1998. Le variabili considerate nello studio sono:

1. la collocazione geografica della sezione;
2. la dimensione demografica del comune;
3. un indicatore di vicinanza del comune alla soglie *critiche* di popolazione;
4. un indicatore della qualità del lavoro del comune in relazione alle rilevazioni statistiche in esso condotte;
5. la tipologia della sezione censuaria (centro, nucleo e case sparse).

E' necessario precisare che la collocazione geografica e la dimensione demografica del comune nei diversi modelli messi a confronto sono state considerate in forme differenti: regione o ripartizione per l'indicatore al punto 1 e popolazione intesa come variabile continua o suddivisa in classi di popolazione[3] per l'indicatore al punto 2.

L'indicatore al punto 3 è un indicatore definito dal Servizio delle Statistiche Demografiche dell'ISTAT al fine di identificare quei comuni aventi dimensione demografica prossima alla soglie di 5.000, 10.000,…, 50.000 abitanti oltre le quali il comune può accedere a finanziamenti e benefici definiti dalla legislazione vigente; si vuole infatti valutare se esiste un rischio di sovracopertura in tali casi. La variabile al punto 4 viene definita dal medesimo Servizio sulla base di una serie di informazioni (tra cui l'indicatore al punto 3), rilevate nel periodo intercensuario, riguardanti l'aggiornamento delle schede anagrafiche e in generale la qualità del lavoro statistico dei comuni. Come sintesi delle informazioni raccolte si attribuiscono ai comuni *punteggi* da 1 a 100, dove i valori inferiori del punteggio indicano una migliore qualità del lavoro statistico. Il suddetto punteggio viene utilizzato, inoltre, per la classificazione dei comuni in 4 tipologie secondo modalità connesse al rischio di commettere errori: rischio minimo, medio-basso, medio-alto e massimo (questa classificazione verrà indicata nel seguito come *voto*). Il punteggio e il voto utilizzati nelle analisi sono quelli attribuiti nell'anno 1999; questo fatto non dovrebbe avere un effetto distorsivo dal momento che gli indicatori non hanno subito variazioni sensibili nell'intervallo intercensuario.

Per maggiori dettagli sulla definizione delle variabili ai punti 3 e 4 si veda M. Dimitri, 1998 *Analisi di qualità sul funzionamento delle anagrafi comunali*.

La finalità dello studio è stata quella di individuare l'insieme delle variabili tra quelle in esame che sono maggiormente legate al tasso di copertura da utilizzare per la stratificazione dei comuni. A tale

[3] Le classi di popolazione cui si fa riferimento sono quelle specificate nella definizione dei domini di interesse nel par. 2.2.

scopo sono stati studiati dei modelli che, per ciascuna sezione campione dell'indagine di copertura 1991, mettono in relazione il tasso di copertura con le informazioni sopra elencate.

Per valutare la validità dell'analisi sono state costruite alcune tabelle preliminari sui dati dell'indagine di copertura 1991 che, per le variabili qualitative in esame, riportano il numero di sezioni campione appartenenti a ciascuna classe. Dalle tabelle 1-5 (appendice A2) è possibile osservare che la Val D'Aosta non era presente nel campione della suddetta indagine e alcune regioni erano presenti con numerosità scarsa. Per quanto riguarda l'indicatore 3, il numero di sezioni appartenenti a comuni vicini alla soglia è esiguo, e pertanto è problematico trarre conclusioni utili riguardo a tale indicatore. Sono state costruite inoltre le tabelle delle distribuzioni congiunte per coppie di variabili di stratificazione, qui non riportate per economia di spazio. Si segnala che alcune celle di intersezione con la variabile *voto* del comune e la variabile *tipo* di sezione hanno numerosità scarsa. Ovviamente le distribuzioni triple evidenziano maggiormente il problema di scarsa numerosità nelle celle e pertanto non è stato possibile includere nei modelli interazioni triple delle variabili in esame.

Il modello adottato è del tipo seguente:

$$\log(\tilde{\tau}_i) = \alpha + \boldsymbol{\beta}\,\mathbf{X}_i + \varepsilon_i \qquad (5)$$

in cui *i* è l'indice di sezione, $\tilde{\tau}_i$ è il tasso di copertura della *i*-ma sezione calcolato secondo il modello di Petersen (cfr. paragrafo 2.1) e $\mathbf{X}_i$ sono le variabili esplicative costruite in base alle informazioni dei punti 1-5. Per le variabili di tipo categorico, ovviamente, la costruzione delle variabili corrispondenti nel vettore $\mathbf{X}_i$ è basata sull'utilizzo della forma disgiuntiva completa.

La procedura utilizzata per l'adattamento dei modelli è la PROC GENMOD del SAS. Per lo studio dei modelli proposti è stata adottata un'analisi di tipo sequenziale, che consiste nel valutare il contributo alla bontà di adattamento apportato da ciascuna variabile considerata secondo la sequenza. Nel prospetto 1 vengono riportate, per ciascun modello, le variabili esplicative (indicate con ×) e le variabili risultate significative (indicate con ××); non viene riportato, invece, l'ordine di introduzione delle variabili che può essere verificato con le elaborazioni prodotte dal SAS per i modelli stimati presentati in appendice A3.

Prospetto 1 *Variabili esplicative e variabili significative per ciascun modello*

Variabile	Modello 1	Modello 2	Modello 3	Modello 4	Modello 5	Modello 6	Modello 7
Ripartizione (x_1)	×	×	×	×	×		
Popolazione (x_2)		××	××				××
Classi di pop.(x_3)	××			××	××		××
Tipo sezione (x_4)	××	××	××	××	××		
Punteggio (x_5)			×				
Voto (x_6)	×	×		×	×		
Ind. soglia (x_7)						×	
$x_8 = x_1 * x_3$				××			
$x_9 = x_1 * x_4$				××			
$x_{10} = x_4 * x_6$				×			

Il simbolo × indica che la variabile è presente nel modello; il simbolo ×× indica che la variabile è significativa

Dall'analisi dei risultati ottenuti emerge che le variabili utili alla stratificazione sono la popolazione del comune e il tipo di sezione, inoltre risulta essere significativa anche la considerazione congiunta della ripartizione con la classe di popolazione e della ripartizione con il tipo di sezione.

Gli indicatori x_5, x_6, e x_7 e le variabili territoriali (regione e ripartizione) considerate marginalmente non risultano, invece, essere significative in nessuno dei modelli considerati; in particolare l'analisi

sull'indicatore x_7 risulta problematica sia per il numero esiguo di casi sia in quanto tale indicatore potrebbe essere eventualmente maggiormente legato al fenomeno della sovracopertura qui non considerato.

In conclusione, poiché le celle ottenute dall'intersezione delle modalità delle variabili ripartizione e classi di popolazione sono già incluse tra le variabili da utilizzare per la stratificazione, in quanto definiscono gli strati che rappresentano i domini principali di interesse per le stime dell'indagine, si è considerata, pertanto, solo la variabile tipo di sezione come ulteriore criterio di stratificazione delle sezioni all'interno di ciascuno strato T_h (h=1,2,…,20).

Si sottolinea, infine, che, come emerge dal modello 7, la classificazione della popolazione nelle quattro classi di popolazione prestabilite (variabile x_3) non esaurisce l'apporto della variabile ampiezza della popolazione. In tale modello, infatti, quest'ultima variabile risulta significativa anche se viene introdotta dopo che l'effetto della classificazione mediante x_3 è già stato preso in considerazione. Risulta quindi utile considerare una stratificazione più fine in termini di dimensione demografica dei comuni all'interno di ciascun dominio.

3.3 Criteri per la definizione del campione di primo stadio

Nell'ambito di ciascun dominio T_h, viene effettuata una stratificazione dei comuni in base alla loro dimensione demografica; dopo aver ordinato i comuni secondo una graduatoria decrescente in base alla popolazione residente, vengono suddivisi in strati. In tale processo è prevista l'identificazione di un sottoinsieme di unità primarie dette *Auto Rappresentative* (AR), ciascuna delle quali costituisce strato a se stante ed è inserita con certezza nel campione. Il sottoinsieme delle unità AR è formato dalle unità primarie di dimensione demografica superiore a una prefissata soglia di popolazione; le restanti unità primarie sono dette *Non Auto Rappresentative* (NAR) e sono suddivise in strati di uguale ampiezza in termini di popolazione totale.

Inoltre, nel sottoinsieme di unità AR sono stati inclusi anche tutti i comuni metropolitani con popolazione inferiore alla soglia demografica per garantire comunque la possibilità di effettuare per questi comuni stime dirette.

Il disegno è dunque di tipo composito, essendo costituito dall'unione di un disegno ad uno stadio stratificato in cui le unità primarie sono costituite dalle sezioni (parte AR) e da un disegno a due stadi stratificato (parte NAR).

Con riferimento al generico dominio T_h, si è indicato[4] con: h (h=1,…, H=20) l'indice di dominio, l (l=1,…. L_h) l'indice di strato di comune e con c l'indice di comune. Inoltre le quantità S_{hlc}, Q_{hlc} e N_{hlc} denotano rispettivamente il numero di sezioni, il numero di famiglie e di individui residenti nel generico comune c dello strato l, mentre S_h, N_h e Q_h indicano le analoghe quantità marginali riferite al dominio T_h. Siano inoltre M_{hl} il numero di comuni appartenenti allo strato l, $\overline{\overline{Q}}_h = Q_h / N_h$ il numero medio di componenti per famiglia del dominio, $\overline{N}_h = N_h / S_h$ il numero medio di famiglie residenti per sezione, m_{hl} il numero di comuni campione nello strato l ($m_{hl} = 1$ per gli strati AR ed $m_{hl} = m_h$ per gli strati NAR) e n_h il numero pianificato di famiglie campione. Sia, infine

$$n = \sum_{h=1}^{20} n_h \qquad (6)$$

[4] Adottiamo qui e nel seguito la convenzione che laddove un indice è assente, la quantità deve essere intesa come marginale rispetto a quell'indice

il numero pianificato di famiglie campione a livello totale .

Ciò premesso, le caratteristiche principali del disegno campionario di primo stadio inerente a ciascun dominio T_h (h=1,…20) sono:

a) la stratificazione dei comuni in funzione della sola dimensione demografica degli stessi;
b) l'autoponderazione del campione al livello di ciascun dominio T_h;
c) la selezione di $m_h = 2$ comuni campione per ciascuno strato NAR;
d) la definizione di un numero minimo pianificato, $\bar{n}_h$, di famiglie da intervistare in ciascun comune campione;
e) la formazione di strati di comuni aventi ampiezza approssimativamente costante in termini di popolazione residente.

In termini operativi, determinate le numerosità *n* ed n_h (*h*=1,…,20) in base agli errori attesi della stima del tasso di copertura (si veda il successivo paragrafo 3.5 sulla valutazione degli errori), per la formazione del campione si procede come segue:

a) si definisce il numero minimo pianificato di famiglie, $\bar{n}_h$, da intervistare in ciascun comune campione di T_h $(\mathrm{h}=1,\mathrm{K}\ \mathrm{K},\mathrm{H})$. Tale definizione dipende sia dalla scelta delle quantità *n* ed n_h (*h*=1,…20) che dal numero complessivo di comuni che si intende far partecipare all'indagine. Inoltre, in considerazione del fatto che il disegno di secondo stadio è a grappoli - prevede, infatti, la selezione di un campione di sezioni e l'enumerazione di tutte le famiglie appartenenti alle sezioni estratte - il numero prefissato per $\bar{n}_h$ deve essere necessariamente maggiore o uguale al numero medio previsto[5] di famiglie residenti per sezione, $\overline{N}_h^*$. Sulla base della numerosità complessiva (prefissata in termini di numero pianificato di famiglie M), tenendo conto anche del fatto che per ragioni operative e di costo il numero complessivo di comuni campione non deve essere superiore a 100, si è posto $\bar{n}_h$ pari a 600;
b) si attua il processo di stratificazione dei comuni:
   b1) si determina il valore della soglia G_h, mediante la relazione $G_h = \bar{n}_h \overline{\overline{Q}}_h \left(N_h / n_h \right)$;
   b2) si ordinano i comuni di ciascun dominio secondo la dimensione (demografica) decrescente, i comuni di tipo AR sono quelli la cui dimensione è superiore o uguale a G_h e i comuni metropolitani; i restanti comuni vengono definiti come NAR;
   b3) i comuni di tipo NAR sono suddivisi in L_h strati la cui dimensione è approssimativamente uguale al prodotto $m_h \times G_h$;

[5] Tale numero dipende dai criteri con i quali sono state formate le sezioni censuarie per il Censimento Generale della popolazione2001. Dalla conoscenza di tali criteri è possibile, infatti, effettuare una previsione del numero medio di famiglie per sezione.

c) si selezionano $m_h = 2$ comuni campione da ciascuno degli strati[6] NAR con probabilità di selezione espresse da $z_{hlc} = Q_{hlc} / Q_{hl}$;

d) per ognuno dei comuni selezionati si definisce il numero atteso di famiglie campione mediante la relazione[7] $n_{hlc} = \frac{1}{m_h} \frac{n_h}{N_h} N_{hl}$, basata sulla condizione di autoponderazione.

Il campione di primo stadio è stato progettato in base alle ultime informazioni disponibili sulla popolazione residente dei comuni che risalivano a gennaio 1999. Tali informazioni sono state utilizzate sia per la stratificazione dei comuni che per la selezione dei comuni campione con probabilità variabili. Inoltre, poiché il campione di primo stadio è stato definito e selezionato prima che le operazioni di formazione delle nuove sezioni censuarie fossero terminate, non si disponeva, a tale data, di importanti informazioni relative alle nuove sezioni di censimento utili per la definizione del campione. Tali informazioni riguardano, in particolare, il numero di nuove sezioni appartenenti a ciascun comune e l'ampiezza media prevista, $\overline{N}_h^*$, delle nuove sezioni in termini di individui e famiglie. Tale circostanza ha portato a definire il campione di primo stadio – in termini di stratificazione dei comuni e selezione dei comuni campione - sulla base dell'ipotesi semplificatrice di dover selezionare al secondo stadio un campione casuale semplice di famiglie, definito in termini di numerosità attesa di famiglie, anziché un campione casuale di sezioni. Tuttavia come illustrato nel par. 3.1, il disegno effettivo prevede, al secondo stadio, la selezione di un campione di sezioni all'interno delle quali la numerosità di famiglie da enumerare è casuale. Pertanto, anche se il numero di sezioni campione è stato successivamente definito sotto il vincolo che il numero effettivo di famiglie enumerate sia il più possibile prossimo a quello definito nella fase di determinazione del disegno di I stadio, il numero effettivo di famiglie campione e quello atteso possono risultare differenti.

3.4 Cenni sulla definizione del campione di secondo stadio

Come emerso dal par. 3.2, la tipologia di sezione risulta essere legata al tasso di copertura, si è quindi ritenuto importante considerarla come criterio di stratificazione delle unità di campionamento di secondo stadio.
Con riferimento al generico comune campione *c* dello strato *l* nel dominio T_h, denotiamo[8] con *g* la tipologia di sezione (1=nucleo, 2=centro, 3=case sparse), siano, quindi, S_{hlcg} il numero di sezioni di tipo *g*, N_{hlcg} il numero di famiglie residenti nelle sezioni di tipo *g*, s_{hlcg} il numero di sezioni campione di tipo *g*. Consideriamo, inoltre, relativamente al dominio T_h, le seguenti quantità:

---

[6] Per l'estrazione delle UP si può ricorrere alla procedura di selezione sistematica, suggerita da Madow (1949) e Murthy (1967), che presenta le seguenti caratteristiche: (i) assegna una probabilità di inclusione espressa nella forma $\pi_{hlc} = m\, z_{hlc}$; (ii) la sua implementazione è estremamente semplice; (iii) permette di ottenere stime generalmente più efficienti rispetto a quelle che si ottengono con altre procedure di selezione (Fabbris, 1991).

[7] Negli strati AR si ha m_h e $z_{hlc} = 1$.

[8] Si ricorda che, al fine di non appesantire la notazione introdotta, si adotta la convenzione che laddove un indice è assente la quantità deve essere intesa come marginale rispetto allo stesso indice.

$S_{hg} = \sum_{l=1}^{L} \sum_{c=1}^{M_{hl}} S_{hlcg}$ numero di sezioni di tipo *g*; $N_{hg} = \sum_{l=1}^{L} \sum_{c=1}^{M_{hl}} N_{hlcg}$ numero di famiglie residenti in sezioni di tipo *g*; $s_{hg} = \sum_{l=1}^{L} \sum_{c=1}^{M_{hl}} s_{hlcg}$ numero di sezioni campione di tipo *g*.

Se si selezionasse un campione di sezioni da ciascuno strato g (g=1,2,3) definito all'interno dei comuni estratti, diverrebbe problematico limitare il numero di sezioni campione (e conseguentemente il numero di famiglie campione) .
Per ovviare a tale inconveniente è possibile adottare un disegno di campionamento analogo al *metodo di stratificazione a due vie* proposto da Bryant, Hartley e Jessen (1960) .

Adattando lo schema di selezione di Bryant, Hartley e Jessen (1960) al caso in esame, non si definisce un campione di sezioni di numerosità s_{hlcg} per ciascuno strato g (g=1,2,3) di ogni comune campione, ma si adotta un meccanismo di selezione che determina casualmente le numerosità campionarie s_{hlcg} (eventualmente pari a zero), in modo da rispettare le numerosità campionarie prefissate a livello marginale, ossia il numero di sezioni per tipologia di sezione g (g=1,2,3) e il numero di sezioni per comune. Quest'ultimo metodo consente, quindi, di tenere sotto controllo sia il numero di sezioni campione totale che il numero sezioni campione per tipologia di sezione.
Al fine di illustrare il metodo di selezione di Bryant, Hartley e Jessen, si supponga che nel generico dominio T_h siano stati formati L_h=4 strati di comuni e che sia stato definito il numero di sezioni, s_{hlc} (c=1,2; l=1,…, L) con riferimento a ciascuno dei $2L_h$= 8 comuni campione selezionati al primo stadio. Si supponga, inoltre, di aver definito il numero di sezioni, s_{hg} (g=1,2,3), per ciascuna tipologia di sezione *g*. Tali numerosità campionarie rappresentano rispettivamente le distribuzioni marginali di riga e di colonna del prospetto 2.
Come è possibile osservare non tutte le celle del prospetto sono piene in quanto solo per gli strati selezionati, *hlcg*, vengono definite le numerosità campionarie s_{hlcg}.

Prospetto 2: *Esempio di selezione in una stratificazione a due vie*

		Tipologia di sezione			
Strati di comuni	**Comuni campione**	g=1	g=2	g=3	
1	c=1	s_{h111}	-	-	s_{h11}
	c=2	-	-	s_{h123}	s_{h12}
2	c=1	-	s_{h212}		s_{h21}
	c=2	s_{h221}	-	s_{h223}	s_{h22}
3	c=1	-	s_{h212}	-	s_{h21}
	c=2	s_{h321}	-	-	s_{h32}
4	c=1	s_{h411}	s_{h412}	s_{h413}	s_{h41}
	c=2	s_{h421}	-	-	s_{h42}
		s_{h1}	s_{h2}	s_{h3}	s_h

Per la definizione delle quantità marginali del prospetto precedente in termini di sezioni - così come l'allocazione del campione tra gli L_h strati di comuni (nel prospetto la marginale di riga) è proporzionale alla dimensione in termini di popolazione degli strati - in assenza di informazioni

aggiornate ed affidabili sulle varianze di strato, anche l'allocazione tra le tre tipologie di sezione (nel prospetto la marginale di colonna) è stata effettuata in modo proporzionale alla dimensione demografica, secondo l'espressione $n_{hg} = n_h Q_{hg} / Q_h = n_h \; N_{hg} / N_h$, g=1,2,3.

3.5 Definizione della numerosità campionaria e allocazione

3.5.1 Valutazione della variabilità attesa

Per procedere alla definizione della numerosità campionaria complessiva e all'allocazione della dimensione campionaria totale tra i domini di stima dell'indagine, in primo luogo, è stata effettuata un'analisi della variabilità attesa delle stime campionarie prodotte dall'indagine di copertura.

La variabilità delle stime (3) e (4) ha una componente dovuta al modello e una componente dovuta al disegno. In questo lavoro ci siamo concentrati sulla valutazione della seconda componente che ha rilievo per la specificazione del disegno di campionamento.

Al fine di valutare la variabilità attesa della stima del tasso di copertura è utile definire tale variabilità mediante una relazione che la lega al valore del tasso di copertura e all'effetto del disegno di campionamento progettato, ossia, il rapporto tra la varianza del campione complesso adottato e la varianza di un ipotetico campione casuale semplice di pari numerosità in termini di individui. L'effetto del disegno può essere valutato utilizzando una relazione che lo lega ad alcune quantità caratteristiche sia del disegno progettato che della variabile oggetto di indagine. In particolare, tali quantità sono rappresentate dalle numerosità della popolazione e del campione in termini di individui suddivisi in parte AR e in parte NAR, dal numero medio di individui per unità primaria (che è costituita dalla sezione per la parte AR e dal comune per la parte NAR) e dal coefficiente di correlazione intraclasse all'interno dell'unità primaria. Con riferimento al generico dominio T_h, tralasciando per semplicità l'indice di dominio h, la relazione suddetta è data da

$$deff \;\; = deff_{NAR} \;\; + \; deff_{NAR} \qquad (7)$$

con

$$deff_{AR} = \frac{q}{Q^2} \frac{Q_{AR}^2}{q_{AR}^2} \left[1 + \rho_{AR} \left(b_{AR} - 1\right)\right] \qquad (8)$$

e

$$\mathrm{deff}_{\mathrm{NAR}} = \frac{\mathrm{q}}{\mathrm{Q}^2} \frac{\mathrm{Q}_{\mathrm{NAR}}^2}{\mathrm{q}_{\mathrm{NAR}}^2} \left[1 + \rho_{\mathrm{NAR}} \left(\mathrm{b}_{\mathrm{NAR}} - 1\right)\right] \qquad (9)$$

dove q, Q, sono rispettivamente la numerosità campionaria e la numerosità della popolazione (in termini di individui) per l'intero dominio, mentre q_{AR}, Q_{AR} e q_{NAR}, Q_{NAR} sono le medesime quantità per la parte AR e per la parte NAR. Inoltre, ρ_{AR} e ρ_{NAR} sono i coefficienti di correlazione intraclasse per la parte AR e per la parte NAR, b_{AR} è il numero medio di individui per sezione nei comune AR e b_{NAR} è il numero medio di individui campione dei comuni NAR.

I coefficienti di correlazione intraclasse sono stati ricavati a partire dai dati dell'indagine di copertura 1991. Questi non dovrebbero aver subito variazioni sostanziali, nonostante il disegno di campionamento dell'indagine 1991 differisca da quello progettato per l'indagine di copertura del 2001 (si veda Abbate *et al.* 1993) e l'intervallo di tempo trascorso tra le due indagini sia ampio.

E’ stato inoltre ipotizzato per i comuni AR che il numero di sezioni selezionato sia pari a 10.

Inoltre, sono stati calcolati gli errori di campionamento dell’indagine 1991, che rappresentano una prima valutazione dell’errore campionario ottenibile con una numerosità in termini di individui pari a quella della suddetta rilevazione.

3.5.2 *Valutazione degli errori 1991*

Gli stimatori del tasso di copertura e della numerosità di popolazione (3) e (4) sono entrambi funzione non lineare dei dati. La varianza può essere calcolata utilizzando i metodi standard per la stima della varianza applicati sulle *trasformate di Woodruf* a livello di sezione.

Con riferimento agli stimatori (3) e (4) le variabili linearizzate risultano essere rispettivamente:

$$y_{1,hlci} = x_{11,hlci} / \hat{x}_{+1,h} - \hat{x}_{11,h}\ x_{+1,hlci} / \hat{x}^2_{+1,h} \tag{10}$$

$$y_{2,hlci} = x_{1+,h} x_{+1,hlci} / \hat{x}_{11,h} - x_{1+,h}\ \hat{x}_{+1,h}\ x_{11,hlci} / \hat{x}^2_{11,h}\,. \tag{11}$$

La varianza dello stimatore (3) è quindi data da:

$$\hat{Var}(\hat{\tau}) = \hat{Var}_{AR}(\hat{\tau}) + \hat{Var}_{NAR}(\hat{\tau}), \tag{12}$$

con

$$\hat{Var}_{AR}(\hat{\tau}) = \sum_{h=1}^{20} \sum_{c=1}^{C_{h,AR}} S_{hc} \frac{(S_{hc} - s_{hc})}{s_{hc}(s_{hc}-1)} \sum_{i=1}^{s_{hc}} (y_{1hci} - \frac{1}{s_{hc}} \sum_{i=1}^{s_{hc}} y_{1hci})^2\,, \tag{12a}$$

$$\hat{Var}_{NAR}(\hat{\tau}) = \sum_{h=1}^{20} \sum_{l=1}^{L_{h,NAR}} \sum_{c=1}^{2} \frac{m_{hl}}{m_{hl}-1} \sum_{c=1}^{m_{hl}} (\hat{y}_{1hlc} - \frac{1}{m_{hl}} \sum_{c=1}^{m_{hl}} \hat{y}_{1hlc})^2\,. \tag{12b}$$

Inoltre la varianza dello stimatore (4) è espressa da

$$\hat{Var}(\hat{N}) = \hat{Var}_{AR}(\hat{N}) + \hat{Var}_{NAR}(\hat{N}), \tag{13}$$

dove

$$\hat{Var}_{AR}(\hat{N}) = \sum_{h=1}^{20} \sum_{c=1}^{C_{h,AR}} S_{hc} \frac{(S_{hc} - s_{hc})}{s_{hc}(s_{hc}-1)} \sum_{i=1}^{s_{hc}} (y_{2hci} - \frac{1}{s_{hc}} \sum_{i=1}^{s_{hc}} y_{1hci})^2\,, \tag{13a}$$

$$\hat{Var}_{NAR}(\hat{N}) = \sum_{h=1}^{20} \sum_{l=1}^{L_{h,NAR}} \sum_{c=1}^{2} \frac{m_{hl}}{m_{hl}-1} \sum_{c=1}^{m_{hl}} (\hat{y}_{2hlc} - \frac{1}{m_{hl}} \sum_{c=1}^{m_{hl}} \hat{y}_{2hlc})^2\,, \tag{13b}$$

in cui, per semplicità, nella parte AR è stato tralasciato l’indice l di strato, $C_{h,AR}$ è il numero dei comuni AR, $L_{h,\mathrm{NAR}}$ è il numero di strati NAR del dominio h, $\hat{y}_{1hlc} = \sum_{i=1}^{s_{hlc}} y_{1hlci}\, k_{\mathrm{hlci}}$ e $\hat{y}_{2hlc} = \sum_{i=1}^{s_{hlc}} y_{2hlci}\, k_{\mathrm{hlci}}$, essendo k_{hlci} il peso finale attribuito alla sezione campione i del comune c dello strato l del dominio T_h.

Per il dominio costituito dai comuni della ripartizione CENTRO con numero di abitanti tra i 100.000 e 350.000, per il quale risultava una sola unità primaria campione non autorappresentativa,

il metodo precedente non può essere applicato. Si è scelto, quindi, di stimare la variabilità utilizzando modelli regressivi ricavati sulla base degli errori ottenuti per i restanti domini.

Entrando nel dettaglio, sono state dapprima valutate le varianze del numeratore e denominatore del tasso di copertura per il dominio in questione, utilizzando i modelli:

$$\log(\mathrm{var}(\hat{x}_{11})) = a\log(\hat{x}_{11}) + b\log(s)\,, \tag{14}$$

$$\log(\mathrm{var}(\hat{x}_{+1})) = a\log(\hat{x}_{+1}) + b\log(s)\,, \tag{15}$$

dove *s* è il numero di sezioni campione del dominio di riferimento.

Si è poi determinata la varianza della stima di copertura sulla base del modello:

$$\log(\mathrm{var}(\hat{\tau})) = a\log(\mathrm{var}(\hat{x}_{11})) + b\log(\mathrm{var}(\hat{x}_{+1})) + c\log(\hat{\tau})\,. \tag{16}$$

L'adattamento dei modelli (14), (15) e (16) è risultato superiore al 99%.

Gli errori di campionamento ottenuti sono riportati nell'Appendice A4.

3.5.3 *Allocazione del campione e errori attesi*

Per ragioni di tipo operativo e di costo si è stabilito che il numero di comuni campione non sia superiore a 100 e che la numerosità in termini di famiglie sia approssimativamente pari a 65.000. Come è risultato dalla valutazione degli errori relativi all'indagine 1991, le prefissate numerosità per l'indagine 2001, che sono simili a quelle dell' indagine 1991, garantiscono stime attendibili del tasso di copertura a livello nazionale.
Per quanto riguarda la definizione delle numerosità campionarie in ciascuno dei 20 domini di interesse, in assenza di informazioni aggiornate sulla variabilità del tasso di copertura tra i domini non si è ritenuto possibile definire tali numerosità secondo l'allocazione ottima di Neyman. Come noto, sotto l'ipotesi di costanza del tasso di copertura, la migliore allocazione tra i domini per la stima a livello nazionale è l'allocazione proporzionale alla dimensione demografica; tuttavia tale allocazione comporta coefficienti di variazione della stima troppo diversi tra i differenti domini. Al contrario, al fine di rendere approssimativamente uguali gli errori relativi del tasso di copertura riferito ai diversi domini , l'allocazione più opportuna è ovviamente l'allocazione uniforme tra questi. Tuttavia, in tal caso, ci si allontana notevolmente dall'allocazione proporzionale. Per le ragioni sopra esposte, al fine di garantire un certa comparabilità degli errori dei differenti domini senza penalizzare troppo la stima nazionale, è stata adottata un'allocazione di compromesso tra l'allocazione uniforme e l'allocazione proporzionale.
In appendice A5 sono riportati gli errori relativi attesi risultati adottando tale allocazione. Gli errori sono stati valutati ipotizzando per tutti i domini un tasso di copertura costante pari a 0,99 e applicando l'effetto stimato del disegno dell'indagine progettata (si veda par. 3.5.1) alla varianza del campione casuale semplice.
Le tabelle riportano anche le numerosità in termini di individui, tali numerosità sono basate sulla numerosità media delle sezioni censuarie al 1991. Il numero delle sezioni (pari a circa 730 per il campione estratto) è quello risultante al momento dell'estrazione dei comuni e non tiene pertanto conto della definizione delle sezioni censuarie del censimento 2001 e della loro dimensione media, elementi non ancora disponibili al momento della suddetta operazione di estrazione.

Bibliografia

Abbate C., Masselli M. e Signore M. (1993): "A combined post-enumeration survey of the 1991 population and industrial censuses", *Proceedings of ISI*, Vol. 2, 16.3, Florence.

Bryant E. C., Hartley H. O. e Jessen R. J. (1960) “Design and Estimation in a Two-Way Stratification”, *Jour Amer. Stat. Assoc*., 55, 105-124;

Dimitri M. (1998) “Analisi di qualità sul funzionamento delle anagrafi comunali”. *Documento interno ISTAT*.

Fabbris L. (1991) “Campioni di numerosità due o tre per strato selezionati con probabilità variabili: valutazione empirica di alcune stime di frequenze assolute”, in *Atti della giornata di studio sul campionamento statistico*. Annali di statistica, Serie IX, ISTAT.

Madow W. G. (1949) “On theory of systematic sampling”, *Annals of Mathematical Statistics*, 20.

Murthy M.N. (1967) “Sampling theory and methods”, *Statistical Publishing Society*, Calcutta.

Wolter K.M. (1986) “Some Coverage Error models for Census Data”, *Journal of the American Statistical Association*, **81**, 394, pag. 338-346

Appendice A1: Derivazione delle stime delle quantità incognite nel modello di Petersen

Sotto il modello di Petersen descritto nel paragrafo 2.2, dal momento che si assume indipendenza tra le due liste e omogeneità tra le unità all'interno di ciascuna lista le probabilità possono essere scritte

$p_{i1+} = p_{1+}$,
$p_{i+1} = p_{+1}$,
$p_{ijk} = p_{ij+}p_{i+k} = p_{j+}p_{+k}$, jk=1,2

per cui la funzione di verosimiglianza risulta essere

$$L(N,\mathbf{p}) = L(N,p_{1+},p_{+1}) = \binom{N}{x_{11}\, x_{12}\, x_{21}} p_{1+}^{x_{1+}}\ p_{+1}^{x_{+1}}\ (1-p_{1+})^{N-x_{1+}}\ (1-p_{+1})^{N-x_{+1}};$$

da cui deriva il sistema di massima verosimiglianza:

$$\sum_{k=0}^{x_1-1}\log(N-k) + N\log(1-p_{1+})(1-p_{+1}) - \sum_{k=0}^{x_1-1}\log(N-1-k) + (N-1)\log(1-p_{1+})(1-p_{+1}) =$$
$$\log(N) - \log(N-x_1) + \log(1-p_{1+})(1-p_{+1}) = 0$$

$$\frac{x_{1+}}{p_{1+}} - \frac{N-x_{1+}}{1-p_{1+}} = 0$$

$$\frac{x_{+1}}{p_{+1}} - \frac{N-x_{+1}}{1-p_{+1}} = 0$$

dove $x_1 = x_{11} + x_{12} + x_{21}$ ossia è la somma dei conteggi osservati in entrambe le occasioni di enumerazione.

Le tre equazioni precedenti danno luogo al sistema

$p_{1+} = x_{1+} / N$
$p_{+1} = x_{+1} / N$
$(N - x_1)/N = (1 - x_{1+}/N)(1 - x_{+1}/N)$

da cui risultano gli stimatori di massima verosimiglianza:

$$\tilde{N} = \frac{x_{1+}x_{+1}}{x_{11}},\ \tilde{p}_{1+} = \frac{x_{11}}{x_{+1}} \text{ e } \tilde{p}_{1+} = \frac{x_{11}}{x_{1+}}.$$

Il parametro p_{1+} corrisponde al parametro che nel testo è stato definito come tasso di copertura, τ.

Appendice A2: Analisi sulle distribuzioni del campione di copertura 1991

Tabella 1: Distribuzione delle sezioni campione tra le regioni

Regione	*Frequenza*	*Percentuale*	*Frequenza cumulata*	*Percentuale Cumulata*
Piemonte	49	7,8	49	7,8
Lombardia	50	8	99	15,8
Trentino A.A.	12	1,9	111	17,7
Veneto	42	6,7	153	24,4
Friuli V.G.	6	1	159	25,4
Liguria	31	5	190	30,4
Emilia R.	47	7,5	237	37,9
Toscana	53	8,5	290	46,3
Umbria	6	1	296	47,3
Marche	19	3	315	50,3
Lazio	83	13,3	398	63,6
Abruzzo	28	4,5	426	68,1
Molise	1	0,2	427	68,2
Campania	30	4,8	457	73
Puglia	31	5	488	78
Basilicata	2	0,3	490	78,3
Calabria	49	7,8	539	86,1
Sicilia	53	8,5	592	94,6
Sardegna	34	5,4	626	100

Tabella 2: Distribuzione delle sezioni campione secondo la variabile IND14

Indicatore Soglia Popolazione	*Frequenza*	*Percentuale*	*Frequenza cumulata*	*Percentuale Cumulata*
0	620	99	620	99
1	6	1	626	100

Tabella 3: Distribuzione delle sezioni campione secondo l'indicatore di rischio

Voto	*Frequenza*	*Percentuale*	*Frequenza cumulata*	*Percentuale Cumulata*
1	67	10,7	67	10,7
2	143	22,8	210	33,5
3	153	24,4	363	58
4	263	42	626	100

Tabella 4: Distribuzione delle sezioni campione secondo il tipo di sezione (centro, nucleo, case sparse)

Tipo	*Frequenza*	*Percentuale*	*Frequenza cumulata*	*Percentuale Cumulata*
Centro	544	87,5	544	87,5
Nucleo	34	5,5	578	92,9
Case sparse	44	7,1	622	100

Tabella 5: Distribuzione delle sezioni campione tra le 5 ripartizioni geografiche

Ripartizione	*Frequenza*	*Percentuale*	*Frequenza cumulata*	*Percentuale Cumulata*
Nord Ovest	130	20,8	130	20,8
Nord Est	107	17,1	237	37,9
Centro	161	25,7	398	63,6
Sud	141	22,5	539	86,1
Isole	87	13,9	626	100

Appendice A3: Studio dei modelli per la scelta delle variabili di stratificazione - dati indagine di copertura 1991

Output PROC GENMOD SAS funzione link= log, distribuzione = normale.

Modello 1

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	609	1.3135	0.0022
Scaled Deviance	609	622.0000	1.0213
Pearson Chi-Square	609	1.3135	0.0022
Scaled Pearson X2	609	622.0000	1.0213
Log Likelihood	.	1033.2571	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
RIP	1.3626	4	3.5237	0.4743
CLAPOP	1.3307	3	14.7430	0.0020
TIPO	1.3181	2	5.9107	0.0521
VOTO	1.3135	3	2.1670	0.5385

Modello 2

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	611	1.3205	0.0022
Scaled Deviance	611	622.0000	1.0180
Pearson Chi-Square	611	1.3205	0.0022
Scaled Pearson X2	611	622.0000	1.0180
Log Likelihood	.	1031.5940	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
RIP	1.3626	4	3.5237	0.4743
POP	1.3365	1	12.0245	0.0005
TIPO	1.3234	2	6.1385	0.0465
VOTO	1.3205	3	1.3315	0.7217

Modello 3

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	613	1.3219	0.0022
Scaled Deviance	613	622.0000	1.0147
Pearson Chi-Square	613	1.3219	0.0022
Scaled Pearson X2	613	622.0000	1.0147
Log Likelihood	.	1031.2761	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
RIP	1.3626	4	3.5237	0.4743
POP	1.3365	1	12.0245	0.0005
TIPO	1.3234	2	6.1385	0.0465
INDTOT	1.3219	1	0.6956	0.4043

Modello 4

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	581	1.2155	0.0021
Scaled Deviance	581	622.0004	1.0706
Pearson Chi-Square	581	1.2155	0.0021
Scaled Pearson X2	581	622.0004	1.0706
Log Likelihood	.	1057.3657	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
CLAPOP	1.3417	3	13.1484	0.0043
TIPO	1.3296	2	5.6122	0.0604
VOTO	1.3217	3	3.7018	0.2955
RIP	1.3135	4	3.8819	0.4222
RIP*CLAPOP	1.2741	12	18.9216	0.0904
RIP*TIPO	1.2247	8	24.6316	0.0018
VOTO*CLAPOP	1.2155	8	4.6638	0.7928

Modello 5

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	609	1.3135	0.0022
Scaled Deviance	609	622.0000	1.0213
Pearson Chi-Square	609	1.3135	0.0022
Scaled Pearson X2	609	622.0000	1.0213
Log Likelihood	.	1033.2571	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
VOTO	1.3596	3	4.8739	0.1813
CLAPOP	1.3349	3	11.4121	0.0097
TIPO	1.3217	2	6.1765	0.0456
RIP	1.3135	4	3.8819	0.4222

Modello 6

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	620	1.3695	0.0022
Scaled Deviance	620	622.0000	1.0032
Pearson Chi-Square	620	1.3695	0.0022
Scaled Pearson X2	620	622.0000	1.0032
Log Likelihood	.	1020.2688	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
IND14	1.3695	1	0.3678	0.5442

Modello 7

Criteri per valutare la bontà dell'adattamento

Criterio	G.d.L.	Valore	Valore/G.d.L.
Deviance	617	1.3327	0.0022
Scaled Deviance	617	622.0000	1.0081
Pearson Chi-Square	617	1.3327	0.0022
Scaled Pearson X2	617	622.0000	1.0081
Log Likelihood	.	1028.7529	.

Statistiche LR per Analisi Tipo 1 (sequenziale)

Fonte	Devianza	G.d.L.	ChiSquare	Pr>Chi
INTERCEPT	1.3703	0	.	.
CLAPOP	1.3417	3	13.1484	0.0043
POP	1.3327	1	4.1875	0.0407

Appendice A4: Stime ed errori campionari dell'indagine copertura 1991

Nelle tabelle che seguono sono riportate le stime del tasso di copertura e gli errori relativi percentuali di tali stime per i domini costituiti da ripartizione e classe di popolazione (Censimento 91).
I domini sono indicati con un codice la cui prima cifra indica la ripartizione (1=Nord Ovest, 2=Nord Est, 3=Centro, 4=Sud, 5=Isole) e la seconda classe di popolazione (1 =meno di 10000 abitanti, 2 = tra 10.000 e 100.000 abitanti, 3 = tra 100.000 e 350.000 abitanti, 4 = più di 350.000 abitanti).

Tabella 1. Stime dei tassi di copertura ed errori relativi percentuali per ripartizione × classe di popolazione

Dominio	Stima della copertura	Errore della stima di copertura
11	0,998	0,116
12	0,994	0,259
13	0,995	0,345
14	0,984	0,305
21	0,998	0,151
22	0,989	0,198
23	0,982	0,374
24	0,986	0,843
31	0,994	0,300
32	0,991	0,253
33	0,969	0.667
34	0,969	0,780
41	0,996	0,345
42	0,993	0,246
43	0,970	0,339
44	0,989	0,355
51	0,997	0,204
52	0,980	0,846
53	0,993	0,167
54	0,986	0,641
Totale	0,990	0,089

Tabella 2. Stime dei tassi di copertura ed errori relativi percentuali per ripartizione

Ripartizione	Stima della copertura	Errore della Stima di copertura
1	0,994	0,135
2	0,991	0,155
3	0,984	0,255
4	0,991	0,196
5	0,988	0,382

Tabella 3. Stime dei tassi di copertura ed errori relativi percentuali per classe di dimensione del comune

Classe	Stima della Copertura	Errore della Stima di copertura
1	0,997	0,104
2	0,991	0,152
3	0,981	0,207
4	0,979	0,323

Appendice A5: Numerosità attese ed errori campionari attesi dell'indagine di copertura 2001

Nelle tabelle che seguono sono riportati le numerosità attese del campione e gli errori relativi attesi percentuali per le stime del tasso di copertura nei domini costituiti da ripartizione e classe di popolazione.
I domini sono indicati con un codice la cui prima cifra indica la ripartizione (1=Nord Ovest, 2=Nord Est, 3=Centro, 4=Sud, 5=Isole) e la seconda classe di popolazione (1 =meno di 10000 abitanti, 2 = tra 10.000 e 100.000 abitanti, 3 = oltre 100.000 abitanti esclusi comuni metropolitani, 4 = comuni metropolitani).

Tabella 1. Numerosità campionarie attesa in termini di individui e famiglie ed errori relativi percentuali per ripartizione × classe di popolazione

Dominio	Numerosità attesa camp. individui	Numerosità attesa camp. Famiglie	Errore della stima di copertura
11	12664	5071	0,089
12	11793	4769	0,092
13	4418	1924	0,480
14	6718	3238	0,413
21	10606	3943	0,359
22	9930	3818	0,360
23	6274	2691	0,441
24	4398	2005	0,704
31	8022	3025	0,363
32	11731	4365	0,313
33	5356	2025	0,445
34	8402	3330	0,497
41	11473	4065	0,358
42	17970	5956	0,278
43	5609	2078	0,486
44	6532	2388	0,698
51	7299	2714	0,439
52	9733	3385	0,360
53	5409	1912	0,478
54	6382	2298	0,698
Totale (Italia)	170719	65000	0,092

Tabella 2. Numerosità campionaria attesa in termini di individui e famiglie ed errori relativi percentuali per ripartizione

Ripartizione	Numerosità attesa camp. Individui	Numerosità attesa camp. famiglie	Errore della stima di copertura
1	35593	15002	0,196
2	31208	12457	0,211
3	33511	12745	0,247
4	41584	14487	0,191
5	28823	10309	0,228

Tabella 3. Numerosità campionaria attesa in termini di individui e famiglie ed errori relativi per classe di dimensione del comune

Classe demografica	Numerosità attesa camp. individui	Numerosità attesa camp. famiglie	Errore della stima di copertura
1	50064	18818	0,045
2	61157	22293	0,143
3	27066	10630	0,212
4	32432	13259	0,292