istat
working
papers

N.2
2011

La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici, Epifania Fiorello, Manuela Marrone

Istat
Istituto Nazionale di Statistica

istat working papers

N.2
2011

La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici, Epifania Fiorello, Manuela Marrone

Istat

Istat Working Papers

La progettazione dei censimenti generali 2010-2011:
disegni campionari e stima di errori di campionamento

N. 2/2011

ISBN 978-88-458-1673-4

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici, Epifania Fiorello, Manuela Marrone

Sommario

*La Direzione centrale dei censimenti generali dell'Istat è stata fortemente impegnata nella progettazione dei censimenti del 2010 e 2011. Quest'attività ha rappresentato un'occasione unica per sperimentare nuove soluzioni sul versante metodologico, tecnologico e organizzativo; a riguardo, si è tenuto conto sia delle criticità e delle soluzioni che hanno caratterizzato gli ultimi censimenti, che delle esperienze internazionali che hanno destato maggiore interesse. Una delle innovazioni proposte per la realizzazione del 15° Censimento della Popolazione e delle Abitazioni riguarda l'adozione di una strategia campionaria basata sull'impiego di questionari in "forma ridotta" (*short form*) e in "forma completa" (*long form*) i cui contenuti sono in via di definizione. Il campionamento interesserà solo i comuni capoluogo di provincia e i comuni con almeno 20mila abitanti. In tali comuni la rilevazione sarà caratterizzata dall'osservazione di alcune variabili di natura strettamente socioeconomica solo su campioni di famiglie tramite* long form*, mentre le variabili contenute nello* short form *saranno rilevate in modo esaustivo. Nei comuni non coinvolti dalla strategia campionaria si procederà, invece, a rilevare l'informazione in modo completo tramite la somministrazione del questionario nella forma completa a tutte le famiglie residenti. Al fine di valutare la praticabilità di opportune strategie campionarie, in termini di disegno e di stimatore, per la produzione di stime affidabili su domini comunali e subcomunali, è stato messo in atto uno studio progettuale i cui risultati ottenuti sembrano supportare l'impiego delle tecniche di campionamento nel contesto censuario. Sono in corso ulteriori approfondimenti per migliorare la proposta metodologica e per valutare la qualità dell'informazione censuaria.*

Parole chiave: Censimento, *long form*, campionamento, calibrazione, accuratezza.

Abstract

The Italian National Institute of Statistics (Istat) is planning the 2011 population census and the main aims are to improve the efficiency of the survey operations and, in the meantime, to reduce the workload of the municipalities and the statistical burden for the people involved in the enumeration. In particular, Istat is going to use sampling techniques among many innovations in the censual strategy. It will consist in the simultaneous use of short and long forms: the first one regarding only a sub-set of variables, related to all demographic and a few socio-economic data; the second concerning the overall set of census variables. The sampling strategy will regard municipalities capital province and municipalities with population over 20.000 inhabitants where the whole set census data will be collected only on a sample of households. In municipalities smaller than 20.000 persons a traditional approach is planned by submitting the long form to the whole population. Following the need to adopt a very simple strategy, tests and studies were conducted in order to evaluate the efficiency of sampling estimates. The results seem to be favourable to the use of sampling techniques in order to adopt a short/long form data collection strategy for the 2011 Italian population census.

Keywords: Census, long form, sampling, calibrated estimators, accuracy.

Indice

1. Introduzione*

I censimenti costituiscono un momento conoscitivo unico e indispensabile in quanto portano alla costituzione di un patrimonio informativo di fondamentale importanza per la collettività. I dati raccolti attraverso il censimento garantiscono un elevato dettaglio territoriale, non deducibile da alcuna altra fonte né da altro tipo di indagine e vengono utilizzati ad ogni livello di governo e da una vasta e diversificata utenza a fini di valutazione, programmazione e decisione (Berntsen *et al.*, 2008).

Nonostante i contenuti dei censimenti abbiano subito un'evoluzione nel corso del tempo, esistono molteplici ragioni per proporre innovazioni rispetto alle modalità con cui le rilevazioni totali vengono condotte (Carbonetti *et al.*, 2008c). È forte la necessità di realizzare un censimento più "leggero" rispetto al passato, con l'auspicio da un lato di ridurre il carico di lavoro dei soggetti coinvolti nello svolgimento delle operazioni sul campo e dall'altro di limitare l'insieme delle informazioni richieste a tutta la popolazione.

Le innovazioni di carattere metodologico nelle rilevazioni censuarie (Crescenzi *et al.*, 2009) vanno verso due direzioni principali: l'uso integrato di dati provenienti da fonti amministrative e l'introduzione di tecniche campionarie per la rilevazione delle informazioni socio-economiche. La scelta di introdurre stime campionarie per alcune variabili storicamente censite è avvalorata anche dall'analisi delle esperienze estere (Abbatini *et al.*, 2007) dalla quale emergono realtà di Paesi (Canada, Usa, Francia, Germania, Israele, Olanda) in cui, adottando approcci non tradizionali per il censimento, si producono stime per le variabili non strettamente demografiche.

Da un preliminare studio delle soluzioni metodologiche (Cicchitelli *et al.*, 1992; Särndal *et al.*, 1992) praticabili per il contesto censuario in Italia, sono stati considerati disegni di campionamento *da lista* (per la possibilità di utilizzare i registri anagrafici) o *areali* (per l'opportunità di riferirsi alla lista delle sezioni di censimento delle Basi Territoriali Comunali).

La decisione di ottenere una parte dei dati tipici del censimento tramite campioni di famiglie deve però affiancarsi alla consapevolezza che questo tipo di approccio introduce un errore campionario, e alla capacità di convincere gli utilizzatori che con un buon campione si possono raggiungere risultati equivalenti, e per certi aspetti addirittura migliori, di quelli provenienti da una rilevazione totale.

Questo lavoro fa riferimento alla necessità di valutare alcune possibili conseguenze che l'introduzione della strategia campionaria produce sull'accuratezza dell'informazione censuaria prodotta e diffusa a livello comunale e sub-comunale.

Dopo la descrizione di alcune criticità del censimento della popolazione e delle abitazioni del 2001 e le principali soluzioni innovative che si intendono adottare per quello del 2011 (capitolo 2), viene presentata con maggior dettaglio (capitolo 3) la proposta dell'introduzione della tecnica di rilevazione basata sull'adozione di un questionario *long form*. I capitoli 4, 5 e 6, più a carattere metodologico, descrivono le strategie campionarie proposte e la procedura di calibrazione per la determinazione dei pesi di riporto all'universo. Nel capitolo 7 viene ampiamente descritta tutta la fase sperimentale mentre nel capitolo 8 viene illustrato l'ambiente informatico di supporto all'intero processo.

Dopo la presentazione dei risultati di tutte le sperimentazioni distinti in tre blocchi (capitolo 9), nei successivi capitoli 10, 11 e 12 vengono proposte alcune analisi condotte sui risultati conseguiti, mentre nel capitolo 13 vengono illustrati degli esempi relativi ad alcune delle variabili oggetto di analisi.

Nel capitolo 14, infine, si forniscono alcune conclusioni utili sia per le decisioni da prendere in merito alla scelta della strategia finale che per il proseguimento dello studio su altri fronti connessi alla tematica oggetto di questo lavoro.

*Il lavoro è frutto della collaborazione degli autori. G. Carbonetti ha curato i capitoli 1, 2, 3, 4, 5, 6, 10, 11, 14 e i paragrafi 7.1, 7.2, A.1, A.2, A.5, B.1, B.3; L. De Felici ha curato il capitolo 13 e i paragrafi 7.3, 7.4, 7.5, 9.1; F. Borrelli ha curato il capitolo 12 e i paragrafi 7.6, 9.3; E. Fiorello ha curato i paragrafi 9.2, A.3, A.4, B.2; M. Marrone ha curato il capitolo 8.

2. Le innovazioni proposte per il censimento del 2011

Il censimento della popolazione, riguardando la totalità dei residenti sul territorio nazionale, è per sua stessa natura assai impegnativo in termini di risorse economiche, di organizzazione e di lavoro sul campo (Fortini *et al.*, 2007). Dall'analisi sulla conduzione del passato censimento sono emerse alcune criticità relative all'organizzazione delle operazioni censuarie, tra cui: costituzione, coordinamento e mantenimento dell'imponente rete di rilevatori; gestione delle fasi di consegna e ritiro dei questionari. Va inoltre aggiunto che, a partire dalla tornata censuaria del 2011 il Censimento italiano, così come quello degli altri Paesi membri dell'Unione Europea, è sottoposto a Regolamento Europeo che pone vincoli[1] sui tempi (consegna dei dati entro la data del 1° Aprile 2014), sulle variabili obbligatorie (*core topics*), sulle classificazioni (*breakdowns*) e sulle tavole statistiche (*hypercubes*).

Al fine di migliorare l'efficienza delle operazioni censuarie sul campo e di rispettare gli obblighi sui tempi di rilascio dei risultati finali, sono state formalizzate diverse proposte di innovazione:

- diversificazione di metodi e organizzazione tra comuni di diversa classe di ampiezza demografica;
- formazione di aree di censimento sub-comunali (Astorri *et al.*, 2007) per la diffusione dei risultati;
- realizzazione di archivi comunali di numeri civici geocodificati alle sezioni di censimento;
- impiego di liste pre-censuarie desunte dalle anagrafi comunali per la spedizione postale dei questionari;
- uso congiunto di questionari ridotti (*short form*) e questionari completi (*long form*);
- consegna postale dei questionari;
- multicanalità per la raccolta dei questionari (postale, web, centri di raccolta comunali).

Tali azioni sono finalizzate ad ottenere una maggiore flessibilità dell'organizzazione sul territorio, una più elevata specializzazione degli organi coinvolti, una riduzione significativa del numero di rilevatori (*front-office*) con un contestuale rafforzamento delle capacità di coordinamento e controllo degli Uffici Comunali di Censimento (*back-office*).

3. La tecnica di rilevazione tramite questionari short/long form

Per ridurre il fastidio statistico sui rispondenti, in special modo di quelli residenti nei comuni più grandi, e per massimizzare il ritorno spontaneo dei questionari compilati, riducendo così le operazioni di sollecito da parte dei rilevatori, è stato considerato un questionario con un ridotto numero di quesiti. Inoltre, per mantenere inalterato il contenuto informativo e rispettare i vincoli internazionali (UNECE, 2006 ; United Nations, 2007) in merito alle variabili censuarie è stato suggerito di rilevare un più ampio insieme di variabili su campioni di famiglie. Si prevede quindi di adottare una strategia basata sull'utilizzo contemporaneo di un questionario in forma breve (*short form*), contenente pochi quesiti relativi alle abitazioni e alle caratteristiche familiari e demografiche, e di un questionario in forma estesa (*long form*), contenente tutte le variabili tradizionalmente osservate in occasione del censimento.

Il questionario *long form* verrebbe somministrato, nei comuni di almeno 20mila abitanti[2], solo a campioni rappresentativi di famiglie, mentre lo *short form* sarebbe sottoposto alle rimanenti

[1] Tali vincoli sono stabiliti nel "CES Recommendations for the 2010 Censuses of Population and Housing", preparato dall'UNECE (United Nations Economic Commission for Europe) in collaborazione con Eurostat (Statistical Office of European Communities) formalmente adottato a giugno 2006, in occasione della Conferenza degli Statistici Europei.

[2] È in corso di valutazione la possibilità di estendere questa strategia anche ai comuni tra 5mila e 20mila abitanti.

famiglie non incluse nel campione. Nei comuni di più piccola dimensione, invece, l'indicazione è quella di sottoporre il questionario in versione *long form* a tutte le famiglie. Con tale approccio, i dati demografici e familiari deriverebbero da un conteggio esaustivo, mentre, relativamente ai comuni sottoposti a campionamento, le informazioni di tipo socio-economico e il loro incrocio con le variabili demografiche sarebbero desunte da stime campionarie.

La proposta metodologica avanzata per i comuni superiori a 20mila abitanti considera l'adozione di disegni di campionamento pianificati per produrre stime per *aree di censimento di centro abitato* aventi le seguenti caratteristiche:

- aree di censimento con popolazione superiore ad una data soglia[3];
- aree di censimento dove le liste di famiglie o di sezioni siano disponibili e affidabili ai fini del campionamento (tipicamente quelle situate in zone più centrali del comune piuttosto che quelle periferiche o extra-urbane).

Vale dunque la pena di sottolineare che, sia nei comuni sotto i 20mila abitanti, sia nelle aree che non rispettano le condizioni sopra riportate, la rilevazione censuaria si avvarrebbe del solo questionario *long form* che sarebbe così consegnato a tutte le famiglie.

Questa strategia comporta, per i comuni di maggiore dimensione, una riduzione della mole dei dati da acquisire ed elaborare e permette di eseguire maggiori controlli sui dati raccolti a vantaggio della diminuzione degli errori di misura (Cocchi, 2007).

4. Disegni di campionamento proposti per l'estrazione del campione di famiglie da lista anagrafica

A partire dalla lista anagrafica delle famiglie residenti, disponibile per ciascun comune, è possibile estrarre campioni di famiglie, secondo schemi casuali semplici o stratificati (Cicchitelli *et al.*, 1992).

In questo capitolo saranno descritte le strategie campionarie ritenute adottabili per l'estrazione del campione di famiglie dalla lista anagrafica del comune e per la produzione delle stime dirette. La possibilità di effettuare una stratificazione delle unità di riferimento (le famiglie) è consentita dall'eventuale disponibilità di informazioni utilizzabili in fase di progettazione del campione. A riguardo si è proposto di stratificare le famiglie per "numero di componenti" o per "età del capofamiglia" (informazioni rilevabili sulle "schede di famiglia"). Poiché il dominio pianificato dal disegno è l'area di censimento di centro abitato, il campionamento si traduce nella selezione di un campione di famiglie, rappresentativo delle aree di censimento del comune, a cui sottoporre il questionario *long form*.

Per semplicità di esposizione è descritto solo lo schema di estrazione del campione, la definizione della probabilità di inclusione π_i della generica unità elementare i nel campione (utile per la determinazione del *peso di riporto all'universo diretto* o *da disegno*) e la formula dello *stimatore diretto* (noto in letteratura anche come *stimatore per espansione*) relativa alla stima dell'ammontare T_x della generica variabile *long form* X.

Fissiamo le seguenti notazioni iniziali:

N_F	dimensione della lista L di famiglie eleggibili;
$N_F(A)$	dimensione della sotto-lista $L(A)$ di famiglie dell'area A;
$\underline{c}_F(A)$	campione di famiglie relativo all'area A;
$n_F(A)$	dimensione del campione di famiglie estratto dall'area A.

[3] Per l'esercizio di questo lavoro è stata fissata la soglia di 5mila residenti; per aree con popolazione inferiore il campionamento non è praticabile in quanto la dimensione di un campione rappresentativo risulterebbe paragonabile a quella dell'area stessa.

Ora, per il totale della generica variabile X sulla famiglia *f*, sull'area *A* del comune e sul comune intero (dato dall'aggregazione delle sole aree sottoposte a campionamento), si può scrivere:

$T_{x,f}$ totale della variabile X sulla generica famiglia *f* ;

$T_x(A) = \sum_{f \in L(A)} T_{x,f}$ totale della variabile X sull'area *A* del comune;

$T_x = \sum_A T_x(A)$ totale della variabile X sul comune intero.

4.1 Disegno casuale semplice

Il disegno casuale semplice prevede l'estrazione di un campione di famiglie (con numerosità prefissata) da una lista unica secondo un procedimento casuale senza ripetizione e con probabilità di estrazione costante per ciascuna famiglia della lista. Detto ciò, con riferimento alla generica area di censimento di centro abitato *A* del comune preso in esame, si indica con:

$$\pi_f(A) = n_F(A) / N_F(A) \qquad (4.1)$$

la probabilità di inclusione della generica famiglia *f* appartenente all'area *A* (costante per tutte le famiglie appartenenti all'area in questione),

$$\hat{T}_x(A) = \sum_{f \in \underline{\mathcal{C}}_F} \left(T_{x,f} / \pi_f(A) \right) \qquad (4.2)$$

lo *stimatore diretto* del totale della variabile X sull'area *A* . Denotando poi con

$$r_f(A) = 1 / \pi_f(A) \qquad (4.3)$$

il *peso di riporto all'universo* (il peso che assume ogni unità campionaria nella stima dell'ammontare relativo all'intera popolazione), la (4.2) può essere riscritta come segue:

$$\hat{T}_x(A) = \sum_{f \in \underline{\mathcal{C}}_F} T_{x,f}\, r_f(A) \qquad (4.4).$$

4.2 Disegno casuale stratificato

Il disegno casuale stratificato prevede la partizione della popolazione in strati e l'estrazione da ogni strato di un campione casuale semplice. Quindi, nel caso in discussione, la lista anagrafica delle famiglie viene preliminarmente suddivisa, in base al criterio di stratificazione adottato, in più sottoliste; da ciascuna di esse si procederà poi all'estrazione di un campione di famiglie (con prefissata numerosità) secondo un procedimento casuale senza ripetizione e con probabilità di estrazione costante per ciascuna famiglia della sottolista.

Supponendo di suddividere la lista di famiglie in H strati ed indicando con *h* il generico strato (*h*=1, 2, … , H) le notazioni esposte in precedenza continuano a valere anche con l'aggiunta dell'indice di strato *h* (per cui, a seconda dei casi, si farà riferimento o allo strato sul comune, o alla parte di strato relativo all'area *A*).

Ne consegue che, per la generica area di censimento di centro abitato *A* del comune si indica con

$$\pi_{h,f}(A) = n_{h,F}(A) \big/ N_{h,F}(A) \qquad (4.5)$$

la probabilità di inclusione della generica famiglia *f* appartenente allo strato *h* dell'area *A* (costante per tutte le famiglie relative allo strato *h* dell'area *A*),

$$\hat{T}_x(A) = \sum_h \sum_{f \in \underline{c}_{h,F}} \left(T_{x,h,f} \big/ \pi_{h,f}(A) \right) \quad (4.6)$$

lo *stimatore diretto* del totale della variabile X sull'area *A* . Anche qui, ponendo con

$$r_{h,f}(A) = 1 \big/ \pi_{h,f}(A) \qquad (4.7)$$

il peso di riporto all'universo per le famiglie dell'area *A* che ricadono nello strato *h*, l'espressione (4.6) può essere riscritta come segue:

$$\hat{T}_x(A) = \sum_h \sum_{f \in \underline{c}_{h,F}} T_{x,h,f}\, r_{h,f}(A) \qquad (4.8).$$

In base alle informazioni desumibili dalle schede di famiglia, sono state individuate due diverse proposte per stratificare la lista anagrafica:

1) stratificazione per numero di componenti della famiglia;

2) stratificazione per età del capofamiglia.

Per la prima proposta si è ritenuto di individuare nella dimensione delle famiglie un possibile fattore discriminante per spiegare il comportamento delle variabili *long form*; di conseguenza, in fase di sperimentazione si è pensato di determinare gli strati secondo 4 classi dimensionali rispetto al numero di componenti della famiglia: 1, 2, 3, 4+ componenti. Nel secondo caso si è pensato all'età del capofamiglia quale fattore discriminante alternativo; la proposta di stratificazione ha fatto quindi riferimento all'età dell'intestatario della scheda anagrafica relativa alla famiglia. Per gli strati, sono state decise 4 classi di età: età fino a 30 anni compiuti, età compresa tra i 30 e i 45 anni compiuti, età compresa tra i 45 e i 60 anni compiuti, età superiore ai 60 anni.

I disegni campionari casuali stratificati hanno fatto quindi riferimento ai suddetti criteri di stratificazione; ciò sarà riproposto nel paragrafo 7.2 .

5. Disegni di campionamento areali proposti per l'estrazione del campione di sezioni di censimento

Il mancato aggiornamento della lista anagrafica delle famiglie potrebbe suggerire l'adozione di approcci alternativi per la formazione del campione di famiglie ai fini della rilevazione tramite *long form* al censimento della popolazione. La scelta è stata indirizzata verso disegni di tipo areale in cui la lista di riferimento è ricavabile dalle *basi territoriali comunali*[4]. Dall'elenco aggiornato delle sezioni di censimento si estrae un campione di sezioni (le unità areali) e, successivamente si sottopone il questionario *long form* a tutte le famiglie ivi residenti. Anche in questo caso si può eseguire una selezione secondo uno schema casuale semplice delle unità di riferimento (le sezioni di censimento) oppure in base ad un disegno stratificato. In questo lavoro sono suggeriti due procedimenti di stratificazione che utilizzano come variabile discriminante la dimensione delle sezioni in termini di popolazione residente.

Di seguito vengono presentate le strategie campionarie giudicate praticabili per l'estrazione di campioni di sezioni di censimento dalla lista desumibile dalle basi territoriali comunali, e per il conseguente calcolo delle stime dirette. Il completamento della fase di aggiornamento delle basi territoriali prima del censimento è fondamentale per l'impiego dei campioni areali nel contesto censuario.

Oltre allo schema casuale semplice, si propongono due possibili approcci di disegno di campionamento "areale a grappolo" per la formazione del gruppo di sezioni campione su cui effettuare la rilevazione tramite questionario *long form*; le proposte si differenziano tra di loro per il fatto che si effettui o meno una preliminare stratificazione delle sezioni di censimento dell'area.

Anche per questa soluzione sono illustrate lo schema di estrazione del campione, la definizione della probabilità di inclusione π_i della generica unità elementare i nel campione e l'espressione relativa allo *stimatore diretto* dell'ammontare T_x della generica variabile *long form* X .

Oltre alle notazioni esposte nel capitolo 4 fissiamo le seguenti notazioni aggiuntive:

N_S	dimensione della lista B di sezioni eleggibili;
$N_S(A)$	dimensione della sotto-lista $B(A)$ di sezioni dell'area A ;
$\underline{c}_S(A)$	campione di sezioni relativo all'area A ;
$n_S(A)$	dimensione del campione di sezioni estratto dall'area A .

Ora, con riferimento al totale della generica variabile X sulla sezione s, sull'area A del comune e sul comune intero (aggregazione delle aree campionate), si può scrivere:

$T_{x,s}$	totale della variabile X sulla generica sezione s ;
$T_x(A) = \sum_{s \in B(A)} T_{x,s}$	totale della variabile X sull'area A del comune;
$T_x = \sum_A T_x(A)$	totale della variabile X sul comune intero.

[4] Le basi territoriali suddividono il territorio comunale in località e sezioni di censimento e sono state poste a fondamento dell'organizzazione delle più recenti rilevazioni censuarie in Italia. Le basi dei censimenti del 2001 sono state realizzate nell'ambito del progetto CENSUS 2000, impiegando le più avanzate metodologie e tecnologie cartografiche secondo le norme tecniche definite dall'Istat (Art. 38 del Regolamento Anagrafico - D.P.R. 30 maggio 1989, n. 223).

Per ottenere un'adeguata omogeneità nella rappresentazione degli ambiti territoriali, ciascun comune è stato suddiviso in località di quattro tipologie: *centri abitati*, *nuclei abitati*, *località produttive* e *case sparse*. Ciascuna località è stata suddivisa in sezioni di censimento chiaramente e univocamente individuate, facendo coincidere i bordi sezionali con elementi fisici quali strade, ferrovie, fiumi e assumendo come linea di demarcazione la mezzeria di tali elementi. Per il censimento 2001 il territorio italiano è stato ripartito in 382.534 sezioni. In vista dei nuovi censimenti si procederà all'aggiornamento delle basi territoriali, mantenendo il più possibile inalterati il disegno e i codici delle sezioni del 2001 e permettendo variazioni solo in caso di espansione territoriale delle aree edificate.

5.1 Disegno casuale semplice di sezioni

Similmente a come descritto per il campione di famiglie, il campionamento casuale semplice prevede l'estrazione dalla lista delle unità eleggibili di un campione di sezioni (con prefissata numerosità) secondo un procedimento casuale senza ripetizione e con probabilità di estrazione costante per ciascuna sezione. Quindi, in relazione alla generica area di censimento di centro abitato *A* del comune preso in esame, si indica con:

$$\pi_s(A) = n_S(A) / N_S(A) \qquad (5.1)$$

la probabilità di inclusione della generica sezione *s* appartenente all'area *A* (costante su tutto l'insieme di sezioni appartenenti all'area in esame),

$$\hat{T}_x(A) = \sum_{s \in \underline{C}_S} \left(T_{x,s} / \pi_s(A) \right) \qquad (5.2)$$

lo *stimatore diretto* del totale della variabile X sull'area A . Fissando poi con

$$r_s(A) = 1 / \pi_s(A) \qquad (5.3)$$

il *peso di riporto all'universo* relativo alla generica sezione, la (5.2) descritta diventa:

$$\hat{T}_x(A) = \sum_{s \in \underline{C}_S} T_{x,s} r_s(A) \qquad (5.4).$$

5.2 Disegno casuale stratificato di sezioni

Analogamente a quanto proposto nel caso del campionamento stratificato di famiglie, la lista delle sezioni viene suddivisa, in base al criterio di stratificazione predefinito, in sottoliste da ciascuna delle quali si procederà alla selezione di un campione di sezioni (con fissata numerosità) secondo un procedimento casuale e con probabilità di estrazione costante per ciascuna sezione della sottolista.

Anche in questa situazione, si può supporre di suddividere la lista delle sezioni in H strati, per cui indicando con *h* il generico strato, le notazioni fino ad ora esposte continuano a valere con l'aggiunta dell'indice di strato *h* .

In base a ciò, per la generica area di censimento di centro abitato *A* del comune si indica con

$$\pi_{h,s}(A) = n_{h,S}(A) / N_{h,S}(A) \qquad (5.5)$$

la probabilità di inclusione della generica sezione *s* appartenente allo strato *h* dell'area *A* (costante per tutte le sezioni relative allo strato *h* dell'area *A*),

$$\hat{T}_x(A) = \sum_h \sum_{s \in \underline{C}_{h,S}} \left(T_{x,h,s} / \pi_{h,s}(A) \right) \quad (5.6)$$

lo *stimatore diretto* del totale della variabile X sull'area A .
Quindi, ponendo con

$$r_{h,s}(A) = 1 / \pi_{h,s}(A) \quad (5.7)$$

il peso di riporto all'universo per le sezioni dell'area A che ricadono nello strato h, lo stimatore (5.6) sopra descritto può essere riprodotto come segue:

$$\hat{T}_x(A) = \sum_h \sum_{s \in \underline{C}_{h,S}} T_{x,h,s} r_{h,s}(A) \quad (5.8).$$

Riguardo alla modalità di stratificazione della lista di sezioni sono state esaminate due diverse procedure che, per la loro attuazione, richiedono un preliminare ordinamento delle sezioni in base alla dimensione della popolazione residente osservata al Censimento del 2001:

1) stratificazione in 3 gruppi con ugual numero di individui;
2) stratificazione in 3 gruppi con ugual numero di sezioni.

In base al primo criterio, si determinano 3 gruppi in cui ricade circa 1/3 della popolazione dell'area di riferimento; tale stratificazione generalmente produce strati aventi un numero differente di sezioni in quanto, essendo costituiti a partire dalla lista ordinata, include nel primo gruppo poche sezioni (molto grandi) e nel terzo gruppo tante sezioni (molto piccole). Secondo l'altro criterio si ottengono invece 3 strati aventi un numero di sezioni simile (pari a circa 1/3). Entrambi i criteri di stratificazione sono stati pensati ritenendo quale possibile fattore discriminante, per spiegare il comportamento delle variabili *long form,* la dimensione demografica della sezione; per tale motivo si è cercato di individuare delle modalità di stratificazione in modo da costituire gruppi di sezioni omogenei rispetto a tale caratteristica.

6. I pesi di riporto all'universo e la calibrazione

Ad ogni unità campionaria è associato un peso, chiamato *peso di riporto all'universo*, che esprime quante unità della popolazione sono rappresentate dall'unità in questione.

Una volta fissato il disegno campionario si determina il *peso base* (*peso diretto* o *da disegno*) di ciascuna unità pari all'inverso della probabilità di inclusione nel campione della generica unità (espressioni 4.3 e 4.7 per i pesi da disegno delle famiglie; espressioni 5.3 e 5.7 per i pesi da disegno delle sezioni).

Al fine di aumentare la rappresentatività del campione e la coerenza dei dati osservati con alcune informazioni note sulla popolazione di riferimento, si ricorre alla *calibrazione*. Tale procedura permette di determinare i pesi di riporto all'universo in modo tale che essi risultino il più vicino possibile ai pesi base secondo una prefissata misura di distanza, e allo stesso tempo soddisfino i seguenti *criteri di calibrazione*: le stime dei totali (medie) delle variabili ausiliarie ottenute con questi pesi devono essere uguali ai corrispondenti totali (medie) di popolazione.

La determinazione, quindi, dei pesi finali delle unità campionarie parte dal calcolo dei pesi base e prosegue, nel caso di una strategia che preveda l'utilizzo di stimatori di ponderazione vincolata

(Deville e Särndal, 1992), con la determinazione dei pesi finali (*pesi calibrati*); è proprio l'operazione di *calibrazione* che consente di migliorare la rappresentatività del campione e aumentare la precisione delle stime finali.

Nel presente studio si è decisa l'adozione di stimatori di ponderazione vincolata[5] mediante l'introduzione di vincoli a totali noti di popolazione: la costruzione dei pesi di riporto all'universo si è basata sulla risoluzione di un problema di minimo vincolato, nel quale si impone l'uguaglianza tra un prefissato insieme di totali riferiti a tutte le unità dell'area di censimento di centro abitato e relativi ad alcune variabili demografiche osservate nella medesima occasione censuaria (*totali noti*), e le corrispondenti stime determinate in base ai dati rilevati sul campione estratto dalla stessa area per la rilevazione con questionario *long form*.

È stata eseguita una pre-sperimentazione per determinare l'insieme dei totali noti relativi alle variabili demografiche che, risolvendo in maniera ottimale i problemi di convergenza dell'algoritmo sottostante al software utilizzato[6], rendono praticabile il procedimento di calibrazione utilizzando un numero elevato di vincoli. I risultati di questo studio sono esposti nel paragrafo 9.1.

7. Descrizione delle sperimentazioni

7.1 Obiettivi

Tenuto conto che l'accessibilità degli archivi comunali e la disponibilità delle basi territoriali permettono l'adozione delle strategie campionarie descritte nei capitoli 4 e 5, si è proceduto ad effettuare alcune sperimentazioni (Borrelli *et al.*, 2007; Carbonetti e De Vitiis, 2007; Carbonetti e Fortini, 2008b) per valutare la possibilità di impiego dei disegni di campionamento proposti. A tale scopo le stime di parametri relativi alle principali variabili rilevate in modo campionario tramite il questionario di tipo *long form* sono confrontate valutando i differenti livelli di efficienza campionaria nelle strategie campionarie prese in considerazione.

Gli obiettivi delle sperimentazioni sono riassunti nei seguenti 3 punti:

1. identificazione del *set* ottimale di vincoli per l'impiego dello stimatore di ponderazione vincolata;
2. confronti di efficienza di disegni da lista e di disegni areali (nell'ipotesi di una frazione sondata pari al 33%);
3. confronti di efficienza di disegni semplici da lista per differenti frazioni di campionamento (ipotesi del 10%, 15%, 20% e 33%).

Per ciascuno dei tre obiettivi è stato condotto uno specifico "blocco" di sperimentazioni utilizzando i dati relativi al censimento della popolazione e delle abitazioni del 2001.

L'ambito delle sperimentazioni ha richiesto la specificazione dei seguenti elementi:

a) i disegni di campionamento da confrontare;
b) l'insieme delle variabili di studio;
c) l'insieme dei vincoli di calibrazione;
d) i comuni da sottoporre a test;
e) i dati e le liste-universo per l'estrazione del campione;
f) la costruzione e l'implementazione dell'algoritmo di simulazione.

[5] Si fa presente che mentre lo stimatore diretto gode della proprietà della correttezza, lo stimatore di ponderazione vincolata è asintoticamente corretto. Dai risultati delle sperimentazioni però non emergono distorsioni sui valori attesi delle stime.

[6] Per la procedura di calibrazione si è fatto uso della *funzione di calibrazione* del software Genesees v3.0 sviluppato in Istat (Pagliuca, 2005).

7.2 Disegni di campionamento sperimentati

I disegni considerati nelle sperimentazioni sono quelli descritti con maggior dettaglio nei paragrafi 4.1, 4.2, 5.1 e 5.2, e che di seguito riassumiamo (in parentesi viene indicata una sigla identificativa del disegno che d'ora in avanti verrà utilizzata per brevità di esposizione anche nelle tabelle dei risultati):

- disegno casuale semplice di famiglie (**CCSFAM**);
- disegno casuale di famiglie stratificato per numero di componenti (**STRNCOMP**);
- disegno casuale di famiglie stratificato per età del capofamiglia (**STRETACAP**);
- disegno casuale semplice di sezioni (**CCSSEZ**);
- disegno casuale di sezioni ripartite in tre gruppi aventi un numero totale di individui simile (**STRSPOP**);
- disegno casuale di sezioni ripartite in tre gruppi aventi un numero totale di sezioni simile (**STRSSEZ**).

Al fine di mantenere quanto più possibile inalterato il contenuto informativo dei dati raccolti in occasione del futuro censimento (quelli rilevati in modo esaustivo e quelli osservati a campione), si è ritenuto necessario campionare circa 1/3 delle unità eleggibili (le famiglie appartenenti alle aree campionabili dei comuni superiori ai 20mila abitanti). Le prime verifiche sperimentali hanno quindi preso in considerazione la frazione di campionamento del 33%; successivamente, come verrà illustrato, sono state studiate strategie con frazioni di campionamento inferiori, al fine di valutare l'entità della perdita di efficienza attesa delle stime dovuta a riduzioni della dimensione del campione.

7.3 Variabili di studio

Per la scelta delle variabili di studio è stata preliminarmente effettuata un'analisi sul questionario del censimento 2001 per individuare le variabili che nel 2011 potrebbero essere rilevabili "solo" tramite il questionario nella versione *long*. Nella sperimentazione sono state quindi analizzate le *frequenze relative* per alcune combinazioni (denominati *incroci* nel seguito) di modalità (Tavola 7.1), per un totale pari a 90, delle variabili considerate e segnatamente:

- popolazione totale e popolazione maschile di età maggiore o uguale a 6 anni per "titolo di studio" (6 modalità), per un totale di 12 incroci;
- popolazione totale e popolazione maschile di età maggiore o uguale a 15 anni per "condizione professionale" (8 modalità), per un totale di 16 incroci;
- popolazione totale e popolazione maschile di età maggiore o uguale a 15 anni in condizione di occupati per "settore di attività economica" (14 modalità), per un totale di 28 incroci;
- popolazione totale e popolazione maschile di età maggiore o uguale a 15 anni in condizione di occupati per "posizione nella professione" (4 modalità), per un totale di 8 incroci;
- popolazione totale e popolazione maschile di età maggiore o uguale a 15 anni in condizione di occupati per "posizione nella professione" (4 modalità) e per "settore di attività economica" (3 modalità), per un totale di 24 incroci;
- popolazione totale che si sposta giornalmente nel comune di dimora abituale (1 variabile);
- popolazione totale che si sposta giornalmente fuori del comune di dimora abituale (1 variabile).

Tavola 7.1 - Descrizione delle modalità di classificazione delle variabili relative all'istruzione e al mercato del lavoro rilevabili tramite questionario long e oggetto di stima campionaria

Variabili rilevate con "long form"	Modalità di classificazione
Titolo di studio (6 modalità)	Laurea + Diploma universitario o terziario di tipo non universitario
	Diploma di scuola secondaria superiore
	Licenza di scuola media inferiore o di avviamento professionale
	Licenza di scuola elementare
	Alfabeti
	Analfabeti
Condizione professionale (8 modalità)	Occupati
	In cerca di occupazione
	In cerca di prima occupazione
	Totale Forze Lavoro
	Casalinghe/i
	Studenti
	Ritirati dal lavoro
	In altra condizione
Posizione nella professione (4 modalità)	Imprenditore e Libero professionista
	Lavoratore in proprio
	Coadiuvante familiare
	Dipendente o in altra posizione subordinata
Settore di attività economica (3 modalità)	Agricoltura
	Industria
	Altre attività
Settore di attività economica (14 modalità)	*Agricoltura Totale*
	Industria (Estrazione, Produzione energia)
	Industria (Manifatturiera)
	Industria (Costruzioni)
	Industria Totale
	Altre Attività (Commercio, riparazioni, Alberghi e Ristoranti)
	Altre Attività (Trasporti, Comunicazioni)
	Altre Attività (Intermediazione)
	Altre Attività (Immobiliari, Professionali, Imprenditoriali)
	Altre Attività (Pubblica Amm., Difesa, Ass. Sociale)
	Altre Attività (Istruzione)
	Altre Attività (Sanità, Servizi Sociali)
	Altre Attività (Servizi pubblici/domestici, Org. extraterritoriali)
	Altre Attività Totale

7.4 Comuni sottoposti a test

Per le sperimentazioni è stato definito un insieme di 40 comuni, scelti in modo ragionato in base all'ampiezza demografica[7] e alla collocazione geografica (Tavola 7.2). Da tale insieme sono stati scelti diversi gruppi di comuni (Tavola 7.3) i cui dati sono stati sottoposti ai differenti processi sperimentali descritti nel paragrafo 7.1.

Per lo studio sull'identificazione del *set* ottimale di vincoli per l'impiego dello stimatore di ponderazione vincolata (I° blocco di sperimentazioni), sono stati considerati solo 4 comuni in quanto l'esito relativo ai casi esaminati è stato ritenuto largamente soddisfacente.

Le sperimentazioni inerenti i confronti di efficienza dei disegni campionari da lista e dei disegni campionari areali (II° blocco), sono state condotte su un insieme di 25 comuni per ciascuno dei disegni proposti (tre per quello da lista e tre per quello areale); data l'ampia casistica di situazioni prese in esame, i risultati ottenuti non hanno richiesto ulteriori simulazioni su comuni differenti.

Infine, le simulazioni relative ai confronti di efficienza tra disegni semplici da lista per differenti frazioni di campionamento (III° blocco) hanno riguardato un numero inferiore di comuni (10 casi per ciascuna delle 4 frazioni sondate sottoposte a sperimentazione, di cui 5 in comune e 5 differenti). Tale scelta è stata inizialmente indotta da valutazioni di tipo computazionale, e successivamente confermata dalla stabilità dei risultati ottenuti sull'insieme dei casi esaminati.

Tavola 7.2 - Ripartizione dei 40 comuni sottoposti a sperimentazioni suddivisi per ripartizione geografica e ampiezza di popolazione

Ripartizioni Geografiche	Classi di ampiezza demografica *(popolazione legale 2001)*				Totale Ripartizione
	10.000-20.000	20.000-50.000	50.000-150.000	oltre 150.000	
Nord Occidentale	Tradate Piossasco	Aosta Sondrio	Novara Cuneo	Milano Brescia	8
Nord Orientale	Maranello Cittadella	Belluno Legnago	Rimini Bolzano	Bologna Padova	8
Centrale	Todi Grottammare	Macerata Pontedera	Perugia Pesaro	Firenze Livorno	8
Meridionale	Squinzano Melfi	Vibo Valentia Isernia	Salerno Matera	Napoli Foggia	8
Insulare	Porto Empedocle Sestu	Alghero Enna	Sassari Trapani	Palermo Cagliari	8
Totale classe	10	10	10	10	40

[7] Per valutare l'efficienza delle stime campionarie riferite anche a comuni di dimensione inferiore a 20mila unità, sono stati inclusi nelle sperimentazioni 10 piccoli comuni con popolazione compresa tra 10mila e 20mila.

Tavola 7.3 - Coinvolgimento dei 40 comuni nei differenti blocchi di sperimentazioni (nel caso del III° blocco, per la f.s. del 33%, pur essendo 25 il numero di comuni sottoposti a sperimentazione, è pari a 10 il numero di comuni considerati nell'analisi comparativa effettuata)

Codice Istat	Nome Comune	I° blocco	II° blocco						III° blocco			
		Studio sulla calibrazione	disegni da lista (f.s.=33%)			disegni areali (f.s.=33%)			disegni da lista (CCSFAM)			
			CCSFAM	STRNCOMP	STRETAC	CCSSEZ	STRSSEZ	STRSPOP	f.s.=10%	f.s.=15%	f.s.=20%	f.s.=33%
007003	Aosta	X	X	X	X	X	X	X				(X)
037006	Bologna	X	X	X	X	X	X	X	X	X	X	X
015146	Milano	X	X	X	X	X	X	X				X
054039	Perugia	X	X	X	X	X	X	X	X	X	X	X
090003	Alghero		X	X	X	X	X	X				(X)
004078	Cuneo		X	X	X	X	X	X	X	X	X	X
086009	Enna		X	X	X	X	X	X				X
044023	Grottammare		X	X	X	X	X	X				(X)
094023	Isernia		X	X	X	X	X	X				(X)
023044	Legnago		X	X	X	X	X	X				(X)
049009	Livorno		X	X	X	X	X	X				X
043023	Macerata		X	X	X	X	X	X				(X)
036019	Maranello		X	X	X	X	X	X				(X)
077014	Matera		X	X	X	X	X	X				(X)
076048	Melfi		X	X	X	X	X	X				(X)
028060	Padova		X	X	X	X	X	X				X
001194	Piossasco		X	X	X	X	X	X				(X)
084028	Porto Empedocle		X	X	X	X	X	X				(X)
099014	Rimini		X	X	X	X	X	X				X
092074	Sestu		X	X	X	X	X	X				(X)

Tavola 7.3 - Coinvolgimento dei 40 comuni nei differenti blocchi di sperimentazioni (nel caso del III° blocco, per la f.s. del 33%, pur essendo 25 il numero di comuni sottoposti a sperimentazione, è pari a 10 il numero di comuni considerati nell'analisi comparativa effettuata) (segue)

Codice Istat	Nome Comune	I° blocco	II° blocco						III° blocco			
		Studio sulla calibrazione	disegni da lista (f.s.=33%)			disegni areali (f.s.=33%)			disegni da lista (CCSFAM)			
			CCSFAM	STRNCOMP	STRETAC	CCSSEZ	STRSSEZ	STRSPOP	f.s.=10%	f.s.=15%	f.s.=20%	f.s.=33%
014061	Sondrio		X	X	X	X	X	X				(X)
075079	Squinzano		X	X	X	X	X	X	X	X	X	X
054052	Todi		X	X	X	X	X	X				(X)
012127	Tradate		X	X	X	X	X	X				(X)
081021	Trapani		X	X	X	X	X	X	X	X	X	X
025006	Belluno									X		
021008	Bolzano								X			
017029	Brescia								X			
092009	Cagliari										X	
028032	Cittadella								X			
048017	Firenze										X	
071024	Foggia									X		
063049	Napoli								X			
003106	Novara									X		
082053	Palermo									X		
041044	Pesaro									X		
050029	Pontedera										X	
065116	Salerno										X	
090064	Sassari								X			
102047	Vibo Valentia										X	

7.5 Dati utilizzati

Per le sperimentazioni è stata impiegata la base dati[8] relativa ai *record famiglia* del censimento della popolazione e delle abitazioni del 2001, successivamente aggregati per sezione di censimento. Riguardo le basi di campionamento utilizzate per l'estrazione dei campioni, come lista delle famiglie è stata considerata, in qualità di "ipotetica" lista anagrafica, la lista delle famiglie censite nel 2001 (nell'ipotesi di *invarianza* rispetto alla reale lista anagrafica presente negli archivi amministrativi del relativo comune); come lista delle sezioni di censimento è stata presa la base territoriale comunale del 2001.

Si specifica che i dati impiegati nell'intero processo di simulazione (Tavole 7.4 - 7.5) si riferiscono a:

- sezioni di censimento di tipo "centro" (non vuote);
- famiglie residenti;
- popolazione residente in famiglia

Tavola 7.4 - Numero di aree di censimento, sezioni di tipo "centro", famiglie residenti e individui coinvolti nelle sperimentazioni (Censimento della popolazione 2001)

	Unità campionate	Universo	%
Aree di censimento di centro abitato	498	3.347 (*)	14,85%
Sezioni di censimento di tipo "centro" *(non vuote)*	30.887	242.253	12,75%
Famiglie residenti	2.243.511	21.810.676	10,29%
Individui	5.537.582	56.594.021	9,78%

(*) Numero presunto nell'ipotesi di un disegno di aree di censimento con dimensione tra 5mila e 15mila unità.

[8] La fase di estrazione dati e creazione delle tabelle di riferimento per le sperimentazioni è ampiamente descritta nel successivo capitolo 8 .

Tavola 7.5 - Aree di censimento, sezioni di censimento di tipo “centro” (non vuote), popolazione residente in famiglia e famiglie residenti dei comuni coinvolti nelle sperimentazioni (Censimento della popolazione 2001)

Codice Istat	Nome Comune	Popolazione legale 2001	Popolazione residente in famiglia	Famiglie residenti	Aree di censimento *"disegnate"*	Aree di censimento sottoposte a campionamento	Numero di sezioni di tipo "centro" (non vuote)	Popolazione residente in famiglia appartenente alle aree campionate	Famiglie residenti appartenenti alle aree campionate
015146	Milano	1.256.211	1.243.745	588.197	111	111	5.621	1.239.346	586.475
003106	Novara	100.910	99.505	42.735	12	9	649	83.765	35.701
017029	Brescia	187.567	185.163	81.692	24	15	1.400	168.728	74.078
004078	Cuneo	52.334	51.411	22.082	6	3	373	32.990	14.731
007003	Aosta	34.062	33.679	15.096	3	3	234	32.151	14.468
014061	Sondrio	21.642	21.417	9.292	2	2	151	19.064	8.190
001194	Piossasco	16.138	16.106	6.236	1	1	23	14.926	5.800
012127	Tradate	15.960	15.896	6.479	1	1	48	15.501	6.335
037006	Bologna	371.217	366.617	177.680	35	32	1.883	352.461	171.561
028060	Padova	204.870	199.975	87.027	20	18	1.451	188.781	82.749
099014	Rimini	128.656	127.217	51.168	10	9	1.571	105.698	43.227
021008	Bolzano	94.989	93.726	41.361	10	7	222	84.530	37.230
025006	Belluno	35.050	34.757	14.873	4	2	245	21.797	9.379
023044	Legnago	24.274	24.040	9.405	3	1	193	13.166	5.038
028032	Cittadella	18.743	18.584	6.455	1	1	35	13.426	4.792
036019	Maranello	15.912	15.828	5.743	1	1	21	9.127	3.366
048017	Firenze	356.118	350.358	159.724	34	31	2.557	344.926	155.354
049009	Livorno	156.274	154.621	62.569	18	13	716	136.169	55.274
054039	Perugia	149.125	147.881	57.143	10	10	533	113.355	45.681
041044	Pesaro	91.086	90.503	35.138	8	7	810	74.847	29.525

Tavola 7.5 - Aree di censimento, sezioni di censimento di tipo "centro" (non vuote), popolazione residente in famiglia e famiglie residenti dei comuni coinvolti nelle sperimentazioni (Censimento della popolazione 2001) (segue)

Codice Istat	Nome Comune	Popolazione legale 2001	Popolazione residente in famiglia	Famiglie residenti	Aree di censimento *"disegnate"*	Aree di censimento sottoposte a campionamento	Numero di sezioni di tipo "centro" (non vuote)	Popolazione residente in famiglia appartenente alle aree campionate	Famiglie residenti appartenenti alle aree campionate
043023	Macerata	40.875	40.579	15.959	3	2	269	28.141	11.512
050029	Pontedera	24.971	24.856	9.781	2	2	49	17.690	7.058
044023	Grottammare	14.278	14.196	5.219	1	1	29	13.253	4.915
054052	Todi	16.704	16.458	6.175	1	1	150	10.750	4.153
063049	Napoli	1.004.500	999.641	337.787	102	87	3.646	968.299	325.383
071024	Foggia	155.203	154.041	50.778	13	13	979	145.261	47.685
065116	Salerno	138.188	137.557	46.747	19	11	873	115.308	39.631
077014	Matera	57.785	57.610	19.788	5	4	268	52.264	17.948
102047	Vibo Valentia	33.957	33.641	11.153	2	2	162	19.202	6.263
094023	Isernia	21.152	21.037	7.290	2	2	155	16.369	5.791
076048	Melfi	16.110	16.069	5.409	1	1	24	13.542	4.625
075079	Squinzano	15.355	15.313	5.173	1	1	27	15.313	5.173
082053	Palermo	686.722	682.116	233.557	68	59	2.494	657.626	224.746
092009	Cagliari	164.249	162.845	62.818	18	14	1.189	162.099	61.970
090064	Sassari	120.729	119.642	43.938	12	9	601	91.430	34.212
081021	Trapani	68.346	68.114	24.713	5	5	786	58.275	21.121
090003	Alghero	38.404	38.080	14.709	3	3	342	32.813	12.804
086009	Enna	28.983	28.918	10.823	2	2	52	25.022	9.418
084028	Porto Empedocle	15.957	15.941	5.444	1	1	32	15.769	5.381
092074	Sestu	15.233	15.217	5.045	1	1	24	14.402	4.768

7.6 Algoritmo di simulazione

Riguardo la procedura di simulazione, preliminarmente è stata fatta una attenta valutazione degli aspetti di processo (tempi di elaborazione e gestione fisica dei dati)[9] connessi ai metodi computazionali scelti per risolvere i vincoli metodologici imposti dalla sperimentazione. Il risultato di ciò ha portato alla compilazione di un articolato algoritmo che, lavorando in ambiente SAS (*Statistical Analysis System*), prevede, per ciascun comune scelto e per ciascun disegno campionario proposto, i seguenti passi:

1) l'estrazione di un campione $\underline{c}$ di unità (famiglie o sezioni);
2) la calibrazione (tramite il software Genesees) per il calcolo dei pesi di riporto all'universo;
3) il calcolo delle stime campionarie calibrate $\hat{p}_x(\underline{c})$ e $\hat{T}_x(\underline{c})$ rispettivamente delle percentuali p_x e dei totali T_x per ciascuna modalità di incrocio X presa in esame;
4) l'iterazione dei passi 1), 2) e 3) per un numero prefissato di volte (1.000 replicazioni campionarie[10]) per la simulazione dello spazio campionario con metodi di tipo Monte Carlo;
5) il calcolo, per ciascuna modalità di incrocio, delle medie $E(\hat{p}_x)$, $E(\hat{T}_x)$ e degli scarti quadratici medi $\sigma(\hat{p}_x(\underline{c}))$, $\sigma(\hat{T}_x(\underline{c}))$ delle stime $\hat{p}_x(\underline{c})$ e $\hat{T}_x(\underline{c})$ ottenute sulla distribuzione campionaria simulata tramite le 1.000 replicazioni campionarie generate.

Riassumendo, fissato il comune e il disegno campionario, l'algoritmo di simulazione calcola per tutte le 90 modalità di incrocio e per ciascuna area di censimento di centro abitato sottoposta a campionamento, il valore medio campionario (sull'insieme dei campioni simulati) e il σ campionario che esprime la variabilità delle stime sullo spazio campionario. Successivamente, si passa al calcolo del *coefficiente di variazione percentuale* che per *p* è dato da:

$$cv(\hat{p}_x) = \frac{\sigma(\hat{p}_x)}{E(\hat{p}_x)} \cdot 100 \qquad (7.1)$$

mentre per *T* è dato da:

$$cv(\hat{T}_x) = \frac{\sigma(\hat{T}_x)}{E(\hat{T}_x)} \cdot 100 \qquad (7.2)$$

[9] La complessità computazionale e i tempi di elaborazione necessari per completare le simulazioni hanno rappresentato una notevole difficoltà per il raggiungimento dei risultati di questo lavoro. Nonostante l'impiego contemporaneo di 4 Personal Computer e di un Server multi processore, i tempi impiegati per terminare l'intero processo sono stati molto elevati.

Ad esempio: per ognuno dei comuni sotto i 20mila abitanti, sono state sufficienti circa due ore di tempo-macchina per percorrere l'intero processo di simulazione relativamente ad un singolo disegno campionario; per ciascuno dei comuni di dimensioni tra 20mila e 150mila sono state impiegate circa 24 ore; infine, i singoli comuni più grandi hanno richiesto tempi di elaborazione più lunghi. In particolare, il comune di Milano ha rappresentato il caso limite in quanto sono stati necessari circa 30 giorni-macchina per completare tutte le elaborazioni.

L'utilizzo di un numero maggiore di macchine avrebbe consentito una forte riduzione dei tempi complessivi di esecuzione.

[10] La scelta pari a 1.000, indotta da vincoli computazionali, non ha influenzato la bontà dei risultati ottenuti (Appendice A).

che quantifica la misura[11] dell'errore che mediamente si commette con la stima campionaria. Data la grande mole di risultati ottenuti, al fine di permettere i necessari confronti tra le strategie sottoposte a test, si è proceduto ad una opportuna sintesi così come sarà esposto nel paragrafo 9.2.

8. Ambiente informatico di supporto al processo sperimentale

8.1 L'ambiente di consultazione ed analisi

In questo paragrafo è illustrata la composizione dell'ambiente di *data warehouse*[12] che è stato specificamente utilizzato nell'ambito dell'attività di sperimentazione e della successiva analisi statistica.

L'ambiente si compone di:

Tabelle dei Metadati o delle dimensioni[13], che si distinguono in tabelle di descrizione del territorio e tabelle di descrizione degli oggetti e delle variabili d'analisi;

Tabelle dei Fatti o Datamart[14], volte a contenere le frequenze per sezione di censimento di ciascun oggetto d'analisi rispetto a determinate variabili di classificazione;

Tabelle delle Monovariate, volte a contenere per ciascuna unità d'analisi informazioni che permettano di classificarla rispetto a determinate caratteristiche inquadrate nell'ambito di specifiche variabili d'analisi.

8.1.1 Il database e le utenze

L'ambiente di data warehouse di supporto all'analisi statistica poggia su una *istanza di database Oracle*[15]. L'istanza di database Oracle POPDW si compone di diverse utenze; in particolare le utenze destinate al progetto in esame sono due:

• una utenza di sviluppo destinata alle attività di sviluppo *software* del gruppo informatico e che si compone delle seguenti strutture-dati:

- tabelle dei metadati o delle dimensioni;
- tabelle delle monovariate;
- tabelle dei fatti o Datamart.

• una utenza di lettura destinata alla produzione e all'analisi statistica che si compone in prevalenza di sinonimi creati su tutte le strutture-dati di rilevanza per l'attività di consultazione ed analisi svolta dall'utente statistico ad essa preposto.

L'utenza di lettura è sottoposta ad un periodico e costante allineamento rispetto all'utenza di sviluppo, per permettere all'utente di avere sempre un quadro completo delle tabelle a sua disposizione.

---

[11] In base al valore di *cv* si determina la quantità $\Delta_p = 1{,}96 \cdot p \cdot cv/100$ che rappresenta l'errore assoluto massimo a cui è mediamente esposta la generica stima di *p*. In base alla teoria dei campioni, infatti, sotto valide ipotesi di normalità, il vero valore della percentuale *p* oggetto di stima sarà compreso tra $(\hat{p} - \Delta_p)$ e $(\hat{p} + \Delta_p)$ con una probabilità pari a 0,95 .

Nel caso della stima dell'ammontare *T*, tramite *cv* si calcola l'errore assoluto $\Delta_T = 1{,}96 \cdot T \cdot cv/100$ che permette di definire l'intervallo di confidenza $\{(T - \Delta_T);(T + \Delta_T)\}$ che conterrà il vero valore di *T* con prob.=0,95 .

Esempio: per la stima di *T*=600, se il relativo *cv*=5,4% ne consegue che $\Delta_T = 1{,}96 \times 600 \times 5{,}4/100 \cong 64$. Quindi, il 95% dei campioni produrrà una stima compresa tra 536 e 664 .

[12] Il *data warehouse* è una raccolta di dati integrata, orientata al soggetto, variabile nel tempo e non volatile di supporto ai processi decisionali.

[13] Le *Tabelle dei Metadati o delle dimensioni* sono strutture-dati che contengono informazioni aggiuntive che descrivono e arricchiscono i dati contenuti nel data warehouse.

[14] Le *Tabelle dei Fatti o Datamart* sono raccoglitori di dati specializzati in un particolare soggetto. In termini più tecnici, un *datamart* è un sottoinsieme logico o fisico di un *data warehouse* di maggiori dimensioni.

[15] L'*istanza di database Oracle* è costituita da un set di processi di sistemi e strutture di memoria che interagiscono con i dati memorizzati.

8.2 Organizzazione ed utilizzo dei dati

8.2.1 Composizione ed utilizzo delle Tabelle dei Metadati o delle dimensioni

Le tabelle dei Metadati dell'ambiente di analisi hanno la funzione di fornire informazioni di carattere descrittivo relativamente alle "variabili" utilizzate. In particolare:

Tabella **SEZIONI** , che contiene le informazioni di decodifica del territorio e si compone di:
• campi relativi ai codici identificativi di ciascuna unità territoriale d'analisi:
- CODPRO, per il *codice provincia*;
- CODCOM, per il *codice comune*;
- NSEZ , per il *numero della sezione*;

• campi relativi alla descrizione di ciascuna unità territoriale:
- DZPRO, per la descrizione della provincia;
- DZCOM, per la descrizione del comune.

• un campo PROG_PROV, non utilizzato per fini consultivi ma esclusivamente procedurali.

Dal momento che il più basso livello territoriale rappresentato in questo ambiente è la sezione di censimento, la tabella SEZIONI contiene una riga per ogni sezione di ciascun comune italiano.

Tabella **SE2_MODALITA**, contiene le meta informazioni per la costruzione delle variabili monovariate d'interesse, e costituisce quindi sia il catalogo che i criteri (le regole) di mappatura delle informazioni prodotte relativamente alla tabella dei microdati sorgenti:
- un campo IND_MODALITA, volto a contenere i progressivi numerici assegnati alle modalità di ogni variabile di classificazione;
- un campo NOME_MODALITA, che contiene la descrizione di ciascuna modalità delle variabili di classificazione;
- un campo REGOLA, che contiene la regola per determinare specificamente ogni singola modalità della variabile;
- un campo DWP, che contiene il nome della tabella dei microdati sorgenti sul quale applicare la regola sopra descritta;
- un campo COD_CL, che contiene una sigla alfanumerica assegnata alla variabile di classificazione;
- un campo COD_OGG, che identifica, mediante un'apposita sigla, l'oggetto d'analisi per il quale deve essere determinata una specifica variabile di classificazione;
- un campo REGOLA_OGG, che contiene la regola da applicare sulla tabella dei microdati sorgenti per estrarre l'oggetto d'analisi;
- un campo STR_SELECT, volto a contenere l'operatore (generalmente di "COUNT(*)") da inserire nella *query*[16] da costruire dinamicamente ed eseguire sulla tabella dei microdati sorgenti per il popolamento dei datamart;

La tabella SE2_MODALITA contiene una riga per ogni modalità delle diverse variabili di classificazione definite relativamente ad uno specifico oggetto d'analisi.

[16] La *Query* è un' interrogazione per la consultazione, ricerca, selezione ed estrazione di particolari dati da un database. Viene espressa nel linguaggio SQL e può essere effettuata solamente con database compatibili con il linguaggio SQL (ad esempio database Oracle).

8.2.1.1 Esempi d'uso della tabella dei metadati "SE2_MODALITA"

Di seguito sono descritte alcune delle principali interrogazioni alla Tabella **SE2_MODALITA** con le corrispondenti istruzioni in linguaggio SQL:

interrogazione 1) - individuazione delle modalità di una variabile di classificazione, delle relative descrizioni e regole di calcolo specificando la sigla della classificazione interessata (nel caso della variabile sesso: '*ses2m'*):

```
select se2_modalita.cod_cl sigla_classificazione,
se2_modalita.ind_modalita indice_modalità_classifcazione,
se2_modalita.nome_modalita nome_modalità_classificazione,
se2_modalita.regola regola_modalità_classifcazione
from se2_modalita
where cod_cl ='ses2m'
order by se2_modalita.ind_modalita, se2_modalita.nome_modalita
```

interrogazione 2) – individuazione dell'oggetto d'analisi (la stima d'interesse) associato ad una specifica classificazione (nell'esempio '*ses2m*'), della tabella dei microdati sorgenti dalla quale l'oggetto é estratto e della relativa regola di estrazione:

```
select distinct se2_modalita.cod_cl sigla_classificazione,
se2_modalita.cod_ogg codice_oggetto_analisi,
se2_modalita.dwp nome_microdato_sorgente,
se2_modalita.regola_ogg regola_oggetto_analisi
from se2_modalita
where cod_cl='ses2m'
order by se2_modalita.cod_cl, se2_modalita.cod_ogg
```

8.2.2 Composizione ed utilizzo delle Tabelle dei Fatti o Datamart

Le Tabelle dei fatti o Datamart (struttura ***DATI2**_XXXXX*) sono state realizzate adottando una struttura omogenea composta da due parti

1) identificazione del territorio, in cui l'unità territoriale minima è la sezione di censimento e i campi che identificano le unità territoriali d'analisi sono: CODPRO, CODCOM, NSEZ;
2) variabili monovariate di interesse, i cui campi sono finalizzati a contenere le frequenze monovariate calcolate per ciascun indice di modalità riferita alla tabella dei meta dati relativi.

Il numero delle colonne del Datamart dipende dal numero dei possibili valori (*modalità di classificazione*) che la variabile può assumere. Ogni colonna di classificazione è identificata dal nome della variabile seguito dall'indice di modalità corrispondente: *cod_cl_ind_modalita*.

A ciascuna tabella dei fatti é stato assegnato una specifica *etichetta* che si compone delle sigle che riconducono alle informazioni oggetto d'analisi e alle variabili di classificazione considerate.

Per esempio, in **[DATI2**_cod_ogg_cod_cl**]** il *_cod_ogg_* rappresenta il codice dell'oggetto d'analisi e il *_cod_cl_* il codice che identifica la variabile di classificazione.

Il numero di tabelle dei fatti che si creano per ciascun oggetto d'analisi varia, poiché si possono identificare per uno stesso oggetto diverse variabili di classificazione per le quali determinare la relativa frequenza. In linea di massima lo standard adottato è quello di far corrispondere un Datamart

o tabella dei fatti ad ogni variabile di classificazione relativa ad uno specifico oggetto d'analisi.

Lo storage della tabella[17] dei Fatti (espresso in *Kbyte*[18]) é calcolato in una determinata percentuale (del 7%) sul prodotto del volume-dati che si prevede dovrà essere inserito nel Datamart e la somma della lunghezza delle colonne della tabella. E' importante evidenziare come il volume-dati dei Datamart corrisponde in questo caso al numero delle sezioni di ciascun comune che soddisfano la regola di calcolo dell'oggetto d'analisi.

8.2.3 Composizione ed utilizzo delle Tabelle delle Monovariate

Le tabelle delle monovariate si compongono di:

- un campo che identifica univocamente l'unità d'analisi: PU (al momento l'unità d'analisi presa in esame nel progetto è l'*Individuo residente in famiglia*);
- campi che identificano le unità territoriali d'analisi: CODPRO, CODCOM, NSEZ (il più basso livello territoriale richiesto è la sezione);
- laddove presente, un campo che identifica la famiglia di riferimento: CODFAM;
- i campi finalizzati a contenere, per ciascun indice di modalità delle diverse variabili di classificazione, un valore che può essere 1 o 0, a seconda se l'unità d'analisi in considerazione soddisfa o meno le caratteristiche associate a ciascun indice di modalità della variabile.

Le informazioni riportate nelle tabelle delle monovariate per ciascuna unità d'analisi si riferiscono alle sezioni dei comuni descritti nella Tabella 7.2.

Il numero di questi campi, così come avviene per i Datamart, corrisponde al numero dei possibili valori (*indici di modalità*) che le diverse variabili di classificazione prese in esame per una specifica unità d'analisi possono assumere.

Ogni colonna è identificata dal nome della variabile seguito dall'indice di modalità corrispondente: *cod_cl_ind_modalita*.

A ciascuna tabella delle monovariate é stato assegnato un nome dato dalla sintesi delle informazioni relative all'unità d'analisi di riferimento e il nome della tabella dei microdati sorgenti da cui é derivata: **MONO**_nome_tabella_microdati_sorgenti.

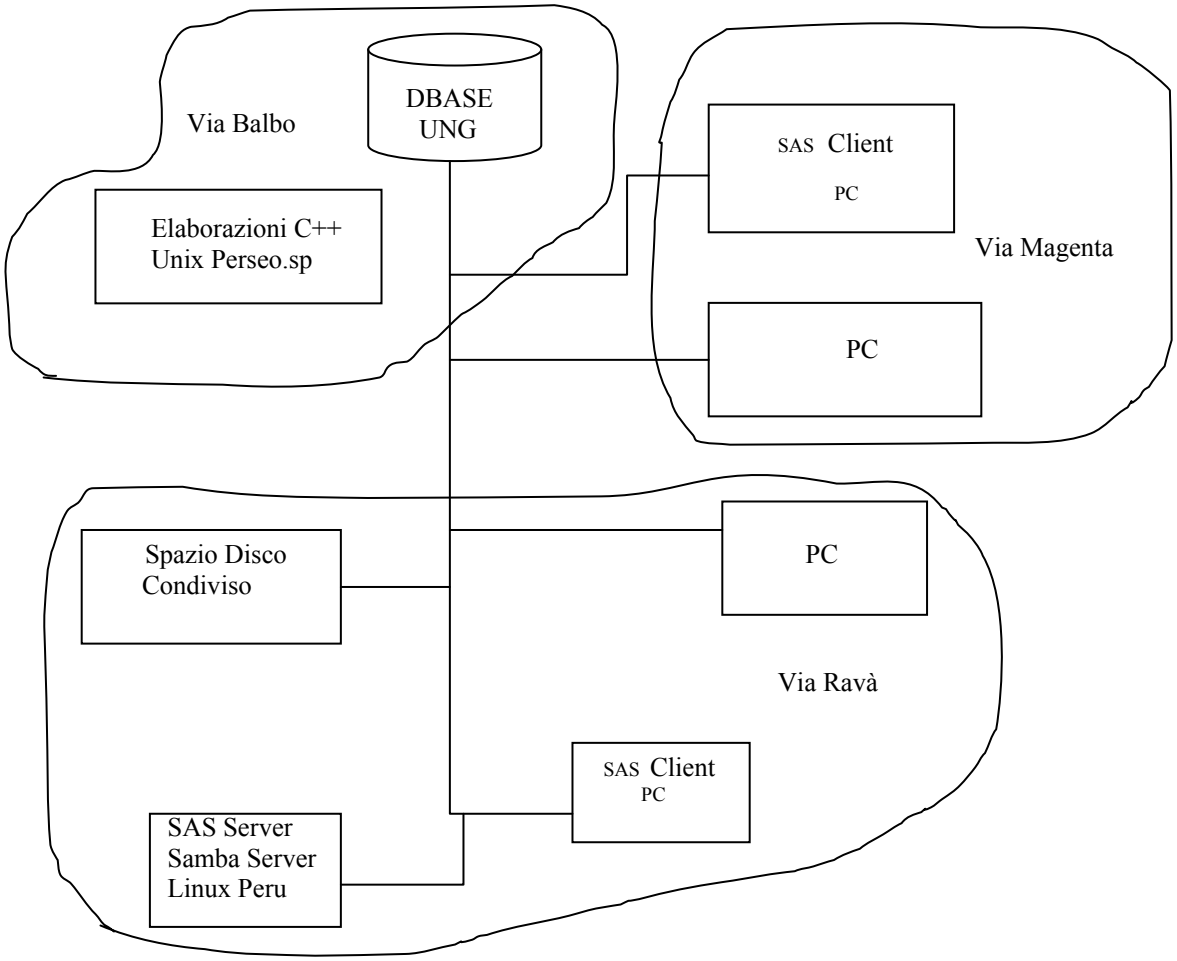
Il numero di tabelle delle monovariate che si creano per ciascuna unità d'analisi varia, poiché nell'attività di ri-classificazione di una specifica unità d'analisi possono essere coinvolte diverse tabelle dei microdati sorgenti, ciascuna delle quali presenta tipi di variabili di classificazione diversi.

Lo *storage della tabella* delle monovariate è calcolato in maniera analoga a quello delle Tabelle dei fatti. In questo caso, si evidenzia che il volume-dati delle tabelle delle monovariate corrisponde al numero complessivo di unità d'analisi presenti nelle diverse sezioni di censimento dei comuni selezionati per le sperimentazioni.

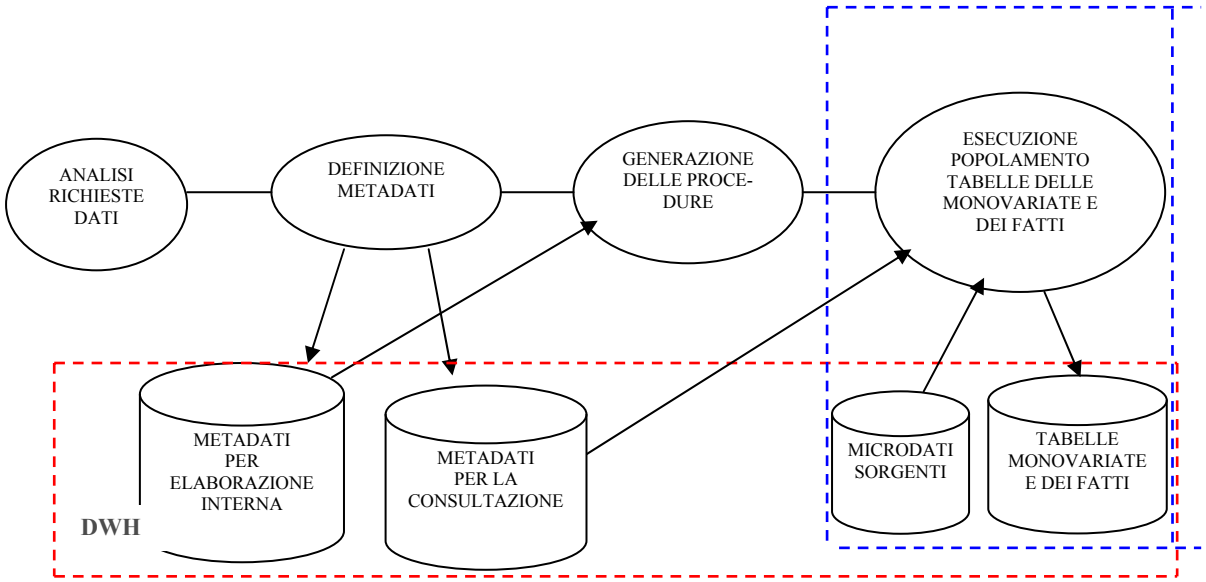
[17] Lo *storage della tabella* (più specificamente Oracle) è una clausola che può essere indicata nella creazione della tabella stessa e permette di impostare le caratteristiche relative al suo dimensionamento.

[18] Il *Kbyte* è un' unità di misura dell'informazione o della quantità di dati e fa parte dei vari multipli del byte.

8.2.4 Organizzazione delle elaborazioni e condivisione delle risorse informatiche



8.2.5 Logica del flusso di popolamento delle Tabelle delle Monovariate e dei Fatti



8.2.6 Strutture-dati del data warehouse

Di seguito è descritta la struttura fisica di alcune delle tabelle, distinte per tipologia, di cui si compone il data warehouse.

8.2.6.1 Tabelle dei metadati o delle dimensioni

A - Tabella degli Ambiti territoriali

Tabella **SEZIONI** (contiene una riga per ogni sezione all' interno di un determinato Comune e di una determinata Provincia).

NOME_COLONNA	FORMATO	DESCRIZIONE
CODPRO	NUMBER(3)	Codice Provincia
CODCOM	NUMBER(3)	Codice Comune
NSEZ	NUMBER(7)	Numero Sezione
DZPRO	VARCHAR2(40)	Denominazione della Provincia
DZCOM	VARCHAR2(70)	Denominazione del Comune
PROG_PROV	NUMBER(5)	Progressivo Provincia

B - Tabella di descrizione delle variabili e degli oggetti d'analisi e delle loro regole di calcolo

Tabella **SE2_MODALITA** (contiene una riga per ogni modalità di una determinata variabile di classificazione relativa ad uno specifico oggetto d'analisi).

NOME_COLONNA	FORMATO	DESCRIZIONE
IND_MODALITA	NUMBER(2)	Indice della modalità della variabile d'analisi
NOME_MODALITA	VARCHAR2(300)	Descrizione della modalità della variabile d'analisi
REGOLA	VARCHAR2(150)	Regola per calcolo della variabile d' analisi
DWP	VARCHAR2(35)	Nome della tabella dei microdati sorgenti
COD_CL	VARCHAR2(30)	Nome della variabile d'analisi
COD_OGG	VARCHAR2(20)	Codice dell' oggetto d' analisi
REGOLA_OGG	VARCHAR2(150)	Regola di estrazione dell'oggetto d'analisi
STR_SELECT	VARCHAR2(30)	Operatore per il calcolo dell' oggetto d'analisi

8.2.6.2 Tabelle dei fatti o datamart

Tabella **DATI2_PR_ETA16** (contiene una riga per ogni sezione di un determinato comune, con la frequenza della variabile di classificazione "eta16" calcolata con riferimento all'oggetto d'analisi "Popolazione residente in famiglia").

NOME_COLONNA	FORMATO	DESCRIZIONE
CODPRO	NUMBER(3)	Codice Provincia
CODCOM	NUMBER(3)	Codice Comune
NSEZ	NUMBER(7)	Numero Sezione
ETA16_1	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età < 5 anni
ETA16_2	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 5 - 9 anni
ETA16_3	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 10 - 14 anni
ETA16_4	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 15 - 19 anni
ETA16_5	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 20 - 24 anni
ETA16_6	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 25 - 29 anni
ETA16_7	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 30 - 34 anni
ETA16_8	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 35 - 39 anni
ETA16_9	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 40 - 44 anni
ETA16_10	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 45 - 49 anni
ETA16_11	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 50 - 54 anni
ETA16_12	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 55 - 59 anni
ETA16_13	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 60 - 64 anni
ETA16_14	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 65 - 69 anni
ETA16_15	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età 70 - 74 anni
ETA16_16	NUMBER(9)	Conteggio della Popolazione residente in famiglia - età > 74 anni

8.2.6.3 Tabelle delle monovariate

Tabella **MONO_DWP_INDIVIDUI**: contiene una riga per ogni individuo residente in famiglia e per ognuno di essi é indicato se soddisfa o meno le caratteristiche proprie di ciascuna modalità delle variabili d'analisi19. A scopo illustrativo è descritta la Tabella delle monovariate relative alle seguenti variabili: **eta16, sesso_eta16, sesso_staciv, ses2m, staciv5m**.

NOME_COLONNA	FORMATO	DESCRIZIONE
PU	NUMBER(10)	Identificatore Univoco della persona residente in famiglia
CODPRO	NUMBER(3)	Codice Provincia
CODCOM	NUMBER(3)	Codice Comune
NSEZ	NUMBER(7)	Numero Sezione
CODFAM	NUMBER(7)	Identificatore della Famiglia di appartenenza
ETA16_1	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età <5 anni
ETA16_2	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 5-9 anni
ETA16_3	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 10-14 anni
ETA16_4	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 15-19 anni
ETA16_5	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 20-24 anni
ETA16_6	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 25-29 anni
ETA16_7	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 30-34 anni
ETA16_8	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 35-39 anni
ETA16_9	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 40-44 anni
ETA16_10	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 45-49 anni
ETA16_11	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 50-54 anni
ETA16_12	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 55-59 anni
ETA16_13	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 60-64 anni
ETA16_14	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 65-69 anni
ETA16_15	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età 70-74 anni
ETA16_16	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di età >74 anni
SESSO_ETA16_1	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età <5 anni
SESSO_ETA16_2	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 5-9 anni
SESSO_ETA16_3	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 10-14 anni
SESSO_ETA16_4	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 15-19 anni
SESSO_ETA16_5	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 20-24 anni
SESSO_ETA16_6	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 25-29 anni
SESSO_ETA16_7	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 30-34 anni
SESSO_ETA16_8	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 35-39 anni
SESSO_ETA16_9	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 40-44 anni
SESSO_ETA16_10	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 45-49 anni
SESSO_ETA16_11	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 50-54 anni
SESSO_ETA16_12	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 55-59 anni
SESSO_ETA16_13	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 60-64 anni
SESSO_ETA16_14	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 65-69 anni
SESSO_ETA16_15	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età 70-74 anni
SESSO_ETA16_16	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e di età >74 anni
SESSO_STACIV_1	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e celibe
SESSO_STACIV_2	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e coniugato o separato di fatto
SESSO_STACIV_3	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e separato legalmente
SESSO_STACIV_4	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e vedovo
SESSO_STACIV_5	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile e divorziato
SES2M_1	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso maschile
SES2M_2	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia di sesso femminile
STACIV5M_1	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia celibe/nubile
STACIV5M_2	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia coniugato/a o separato/a di fatto
STACIV5M_3	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia separato/a legalmente
STACIV5M_4	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia vedovo/a
STACIV5M_5	NUMBER(1)	Variabile d'analisi: Individuo residente in famiglia divorziato/a

[19] La "variabile d'analisi" assume valore 1 se l'individuo soddisfa la caratteristica descritta, 0 se non la soddisfa.

9. Descrizione dei risultati delle sperimentazioni

9.1 Risultati del I° Blocco: identificazione del set ottimale di vincoli per l'impiego dello stimatore di ponderazione vincolata

Lo stimatore di ponderazione vincolata richiede[20] l'utilizzo di vincoli a totali noti di popolazione definiti in funzione di specifici obiettivi. Con riferimento all'ambito censuario, è richiesto che le stime calibrate riproducano il più possibile la struttura demografica della popolazione rilevata tramite i modelli *short* e *long form*.

Al fine di individuare l'insieme di vincoli relativi alle variabili demografiche (sesso, età, stato civile) da utilizzare nell'ambito del processo di calibrazione è stato condotto uno studio sperimentale che accertasse la convergenza dell'algoritmo di ponderazione[21]. L'obiettivo è quello di definire l'insieme di massimo dettaglio che garantisca la risoluzione esatta (convergenza completa) del problema di minimo vincolato che sottintende l'algoritmo impiegato nella calibrazione, così da definire un sistema di ponderazione che mantenga inalterata la struttura demografica della popolazione secondo il dettaglio informativo dei vincoli individuati.

Le valutazioni sono state effettuate per i disegni casuali semplici da lista e i disegni casuali di sezioni nel caso della frazione sondata del 33% e, successivamente, per il solo disegno casuale semplice da lista, per le frazioni di campionamento del 10%, 15% e 20% .

Dal punto di vista operativo la sperimentazione è stata condotta partendo dal livello di massima aggregazione per le variabili demografiche per passare poi in modo gerarchico a successive disaggregazioni secondo classificazioni riferite ad un crescente numero di modalità.

L'insieme di totali noti è costituito dalla distribuzione della popolazione per età (in classi) e sesso, e dalla distribuzione della popolazione per stato civile e sesso.

Le Tavole 9.1 e 9.2 descrivono le aggregazioni delle modalità delle variabili demografiche utilizzate per la definizione dei vincoli di calibrazione.

Tavola 9.1 - Strutture della popolazione per età e sesso prese in esame per la calibrazione

Struttura	Descrizione dei vincoli di calibrazione: Età X Sesso	Numero di vincoli
5m5f	(<15 ; 15-29 ; 30-44 ; 45-64 ; >64) X (M ; F)	10
10m10f	(<5 ; 5-14 ; 15-24 ; 25-29 ; 30-34 ; 35-39 ; 40-44 ; 45-49 ; 50-64 ; >64) X (M ; F)	20
14m14f	(<10 ; 10-19 ; 20-24 ; 25-29 ; 30-34 ; 35-39 ; 40-44 ; 45-49 ; 50-54 ; 55-59 ; 60-64 ; 65-69 ; 70-74 ; >74) X (M ; F)	28
16m16f	(<5 ; 5-9 ; 10-14 ; 15-19 ; 20-24 ; 25-29 ; 30-34 ; 35-39 ; 40-44 ; 45-49 ; 50-54 ; 55-59 ; 60-64 ; 65-69 ; 70-74 ; >74) X (M ; F)	32

[20] Vedi capitolo 6.

[21] Le verifiche sono state eseguite su un numero limitato di simulazioni campionarie e sui dati del censimento 2001 relativi ai comuni di Aosta, Bologna, Milano e Perugia (per un totale di 156 aree di censimento).

Tavola 9.2 - Strutture della popolazione per stato civile e sesso prese in esame per la calibrazione

Struttura	Descrizione dei vincoli di calibrazione: Stato Civile X Sesso	Numero di vincoli
3cm3cf	(celibi ; coniugati e separati di fatto ; separati legalmente + vedovi + divorziati) X (M ; F)	6
4cm4cf	(celibi ; coniugati e separati di fatto ; separati legalmente + divorziati; vedovi) X (M ; F)	8
5cm5cf	(celibi ; coniugati e separati di fatto ; separati legalmente ; vedovi ; divorziati) X (M ; F)	10

Oltre alle classificazioni prese singolarmente sono state considerate anche alcune combinazioni delle stesse: per esempio, la notazione **16m16f4cm4cf** fa riferimento alla classificazione della popolazione relativa a maschi e femmine in 16 classi di età (32 vincoli) e alla suddivisione della stessa popolazione in maschi e femmine per quattro modalità di stato civile (8 vincoli), per un totale di 40 vincoli.

Nelle Tavole 9.3 e 9.4 sono riportati i risultati della verifica della convergenza dell'algoritmo di calibrazione per gli insiemi di vincoli testati alle diverse frazioni campionarie e ai due disegni di campionamento presi in esame. Sono considerati ottimali quegli insiemi che soddisfacendo la condizione di convergenza (a livello di area di censimento di centro abitato) presentano la massima disaggregazione di classificazione:

- disegno casuale semplice da lista, frazioni del 33% e del 20%: 40 totali noti (**16m16f4cm4cf**);
- disegno casuale semplice di sezioni, frazione del 33%: 40 totali noti (**16m16f4cm4cf**);
- disegno casuale semplice da lista, frazioni del 10%, 15% e 20%: 34 totali (**14m14f3cm3cf**).

Tavola 9.3 - Esito del procedimento di convergenza dell'algoritmo di calibrazione nei disegni causali semplice da lista e semplice di sezioni con frazione di campionamento del 33%

Struttura	Numero di vincoli	Disegno casuale semplice da lista	Disegno casuale semplice di sezioni
		f.s.=33%	f.s.=33%
5m5f	10	OK	OK
5m5f3cm3cf	16	OK	OK
10m10f	20	OK	OK
10m10f3cm3cf	26	OK	OK
16m16f	32	OK	OK
16m16f3cm3cf	38	OK	OK
16m16f4cm4cf	40	OK	OK
16m16f5cm5cf	42	NO	NO

Tavola 9.4 - Esito del procedimento di convergenza dell'algoritmo di calibrazione nel disegno causale semplice da lista per differenti frazioni di campionamento

Struttura	Numero di vincoli	Disegno casuale semplice da lista		
		f.s.=10%	f.s.=15%	f.s.=20%
5m5f	10	OK	OK	OK
5m5f3cm3cf	16	OK	OK	OK
10m10f	20	OK	OK	OK
10m10f3cm3cf	26	OK	OK	OK
14m14f	28	OK	OK	OK
16m16f	32	NO	NO	OK
14m14f3cm3cf	34	OK	OK	OK
16m16f3cm3cf	38	NO	NO	OK
16m16f4cm4cf	40	NO	NO	OK
16m16f5cm5cf	42	NO	NO	NO

Dai risultati (Tavola 9.3) si evidenzia che, nel caso della frazione del 33% e per entrambi i disegni campionari presi in esame, la struttura di popolazione ottimale individuata per la calibrazione è la stessa (**16m16f4cm4cf**). Inoltre, è stato provato che tale insieme di totali noti garantisce la convergenza dell'algoritmo anche per i disegni casuali stratificati (con frazione al 33%) introdotti nei paragrafi 4.2 e 5.2.

Nel caso del disegno casuale semplice da lista (Tavola 9.4), la struttura ottimale di totali noti rimane identica scendendo dalla frazione sondata del 33% a quella del 20%, mentre per quelle del 10% e del 15% la convergenza è raggiunta solo tramite una riduzione dell'insieme di modalità relative all'età e allo stato civile (struttura **14m14f3cm3cf**): si è reso necessario aggregare le prime quattro classi di età nelle due "0-9" e "10-19" e accorpare la modalità di stato civile dei "vedovi" con quella dei "separati legalmente o divorziati"[22].

La mancata considerazione del set di vincoli **16m16f5cm5cf** nel caso del disegno casuale semplice da lista per le frazioni sondate del 33% e del 20% è dovuta alla non completa convergenza dell'algoritmo di calibrazione, in quanto in alcune aree si osservano delle differenze tra i valori stimati e i corrispondenti valori veri relativi alle variabili di calibrazione. Poiché le differenze sono di piccola entità e riferite a pochi casi, si ritiene che la struttura di vincoli possa essere validamente impiegata, magari modificando i parametri dell'algoritmo di calibrazione affinché questo possa raggiungere la completa convergenza.

In conclusione, i sistemi di vincoli sulle strutture demografiche della popolazione garantiscono da un lato la piena applicabilità dell'algoritmo di calibrazione e dall'altro il rispetto delle distribuzioni marginali riferite alle variabili demografiche delle tavole statistiche che saranno previste dal piano di diffusione dei dati del censimento della popolazione e delle abitazioni del 2011.

[22] La scarsa numerosità in termini di unità osservate per alcune modalità comporta infatti la mancata convergenza del processo di calibrazione.

9.2 Risultati del II° Blocco: confronti di efficienza di disegni da lista e di disegni areali (nell'ipotesi di una frazione sondata pari al 33%)

In questo contesto sono presentati i risultati ottenuti con le sperimentazioni (II° blocco), utili ai confronti di efficienza dei disegni campionari da lista e di quelli di tipo areale suggeriti per la strategia campionaria al futuro censimento. I confronti sono stati effettuati in termini di coefficiente di variazione (*cv*) che, come detto, fornisce una misura dell'errore che mediamente si commette con le stime campionarie.

Le analisi hanno interessato i dati (Tavole 9.5-9.6) di 25 comuni per un totale di 229 aree di censimento (dominio di stima pianificato nel disegno campionario) sulle quali sono state stimate le frequenze relative delle modalità delle variabili socio-demografiche del questionario *long* su cui si è fatta la sperimentazione, prese singolarmente o incrociate con le variabili demografiche (vedi paragrafo 7.3)

Con riferimento alle frequenze oggetto di stima si è quindi studiata la distribuzione dei *cv* attesi delle stime, classificandoli[23] per valori delle percentuali osservate a livello di area di censimento[24].

Tavola 9.5 - Distribuzione delle aree di censimento sottoposte a sperimentazione (II° Blocco) per dimensione demografica

II° Blocco di sperimentazioni: aree di censimento dei 25 comuni coinvolti.			
Aree con popolazione tra 5.000 e 10.000	Aree con popolazione tra 10.000 e 12.000	Aree con popolazione tra 12.000 e 15.000	Totale
83	39	107	229

Tavola 9.6 - Distribuzione delle frequenze relative da stimare per classi di p e per dimensione di area di censimento (relative al II° Blocco di sperimentazioni)

II° Blocco di sperimentazioni: frequenze relative oggetto di stima sulle aree di censimento dei 25 comuni coinvolti.				
Classi di p	Aree con popolazione tra 5.000 e 10.000	Aree con popolazione tra 10.000 e 12.000	Aree con popolazione tra 12.000 e 15.000	Totale
< 0,05%	408	179	501	1088
0,05% ├0,1%	210	105	271	586
0,1% ├0,25%	500	221	590	1311
0,25% ├0,5%	528	252	680	1460
0,5% ├1%	516	272	748	1536
1% ├2,5%	1286	596	1628	3510
2,5% ├5%	1259	590	1614	3463
5% ├10%	1022	475	1365	2862
10% ├15%	460	238	585	1283
15% ├20%	281	116	370	767
20% ├30%	483	240	641	1364
≥ 30%	517	226	637	1380
Totale	7470	3510	9630	20610

Dapprima si è valutato l'effetto sulle stime dovuto all'adozione di un disegno casuale semplice di famiglie rispetto a quello areale di sezioni. Successivamente, per ciascuno dei due approcci, si è verificato se l'introduzione di differenti modi di stratificare le unità di rilevazione apportasse riduzioni significative al coefficiente di variazione.

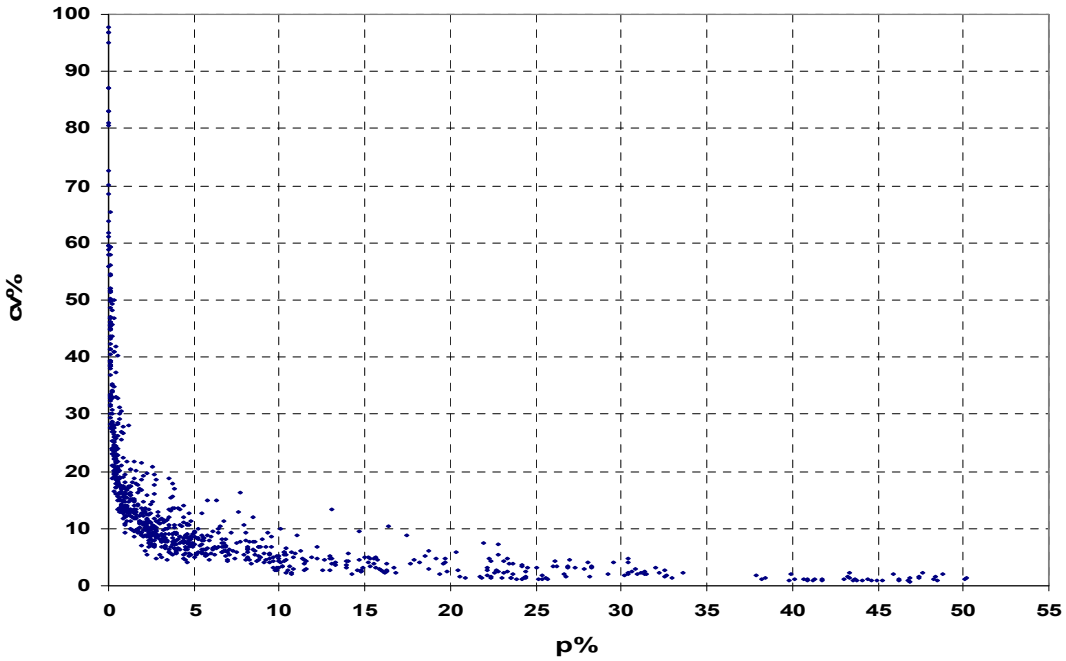
Il primo risultato generale che emerge dall'analisi è che, indipendentemente dal disegno adottato, il *cv* mediano[25] decresce all'aumentare del livello di *p* da stimare (esempio – Figura 9.1).

[23] Le frequenze percentuali *p* oggetto di stima sono state classificate in 12 classi: $p<0,05\%$; $0,05\%\leq p<0,1\%$; $0,1\%\leq p<0,25\%$; $0,25\%\leq p<0,5\%$; $0,5\%\leq p<1\%$; $1\%\leq p<2,5\%$; $2,5\%\leq p<5\%$; $5\%\leq p<10\%$; $10\%\leq p<15\%$; $15\%\leq p<20\%$; $20\%\leq p<30\%$; $p\geq 30\%$.

[24] Poiché sono state escluse dall'analisi tutte le modalità di incrocio per le quali si sono osservate percentuali $p=0$, la somma del numero di stime per area non sempre raggiunge il numero teorico massimo dato dal prodotto tra il numero di modalità di incrocio (90) e il numero di aree.

[25] Per la distribuzione dei valori *cv* si prende a riferimento il valore mediano in quanto misura di centralità relativamente più robusta della media rispetto ai valori estremi.

Figura 9.1 - Grafico di dispersione dei cv rispetto ai valori di p (stime per area). Disegno CCSSEZ. Stime riferite alle aree di censimento del comune di Perugia



9.2.1 Confronto tra il campionamento da lista e il campionamento areale

Dai valori delle distribuzioni dei *cv* mediani emerge che la strategia di campionamento per famiglie da lista garantisce un errore mediamente inferiore a quella per sezioni. Questa tendenza attesa, dovuta all'effetto *cluster* indotto dal campionamento areale, risulta maggiormente evidente per frequenze percentuali *p* superiori allo 0,25%. Anche dall'analisi del valore massimo osservato del *cv* il campionamento di famiglie da lista risulta più affidabile (Tavola 9.7). La differenza del *cv* mediano rimane comunque mediamente inferiore ai 2 punti percentuali; ciò fa ritenere il disegno areale una valida alternativa nel caso in cui le liste anagrafiche comunali non siano impiegabili per qualità non accettabile. Sono stati messi poi in relazione i livelli di *cv* osservati e la tipologia delle aree per dimensione di popolazione[26]. Dal confronto emerge che, a parità di disegno, si registra un notevole guadagno di accuratezza delle stime sulle aree più grandi: il risultato è maggiormente vero per la stima di percentuali inferiori all'1%, mentre all'aumentare di *p* tali differenze sono di entità minore (Tavole 9.8-9.10).

[26] Tre classi: tra 5.000 e 10.000 ; tra 10.000 e 12.000 ; tra 12.000 e 15.000 .

Tavola 9.7 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSFAM e CCSSEZ nel caso della frazione sondata del 33%

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=33% 25 comuni 229 aree			Disegno **CCSSEZ** f.s.=33% 25 comuni 229 aree		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	51,72	96,52	154,50	45,13	94,19	158,12
0,05% ├ 0,1%	36,16	51,02	84,55	30,01	51,41	110,25
0,1% ├ 0,25%	23,31	34,59	98,26	19,46	34,87	111,55
0,25% ├ 0,5%	15,78	23,33	41,37	13,62	24,30	80,53
0,5% ├ 1%	11,12	16,37	33,04	9,29	18,09	89,68
1% ├ 2,5%	4,74	10,65	20,33	4,41	12,25	64,02
2,5% ├ 5%	3,47	7,01	12,75	3,49	8,43	65,76
5% ├ 10%	1,91	4,80	8,91	2,10	5,94	52,52
10% ├ 15%	1,39	3,10	6,12	1,55	4,31	20,44
15% ├ 20%	1,14	2,41	4,81	1,42	3,39	22,07
20% ├ 30%	0,85	1,93	3,85	0,80	2,67	13,54
≥ 30%	0,73	1,34	2,90	0,67	1,74	11,84

Tavola 9.8 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSFAM e CCSSEZ nel caso della frazione sondata del 33%, per stime riferite ad aree di censimento tra 5mila e 10mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=33% 25 comuni 83 aree con popolazione tra 5.000 e 10.000			Disegno **CCSSEZ** f.s.=33% 25 comuni 83 aree con popolazione tra 5.000 e 10.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	68,25	102,20	154,50	56,63	100,91	155,39
0,05% ├ 0,1%	45,39	61,15	84,55	37,99	61,63	110,25
0,1% ├ 0,25%	27,74	40,22	98,26	22,30	40,73	111,55
0,25% ├ 0,5%	20,29	27,31	41,37	15,69	29,04	80,53
0,5% ├ 1%	14,10	19,55	33,04	13,10	21,60	89,68
1% ├ 2,5%	5,46	12,43	20,33	4,90	14,61	64,02
2,5% ├ 5%	3,89	8,25	12,75	3,92	10,07	65,76
5% ├ 10%	2,38	5,65	8,91	2,51	7,13	52,52
10% ├ 15%	1,58	3,66	6,12	1,56	4,96	20,44
15% ├ 20%	1,55	2,84	4,81	1,81	3,90	22,07
20% ├ 30%	1,01	2,27	3,85	0,92	3,20	13,54
≥ 30%	0,89	1,57	2,90	0,80	2,15	11,84

Questo risultato suggerisce di adottare aree di censimento intorno alle 15mila unità, anche perché al crescere della dimensione delle aree si riduce la differenza di efficienza tra il disegno casuale di famiglie e quello di sezioni; quest'ultimo quindi risulta un'alternativa capace di fornire stime soddisfacenti soprattutto nelle aree più grandi.

Tavola 9.9 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSFAM e CCSSEZ nel caso della frazione sondata del 33%, per stime riferite ad aree di censimento tra 10mila e 12mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=33% 25 comuni 39 aree con popolazione tra 10.000 e 12.000			Disegno **CCSSEZ** f.s.=33% 25 comuni 39 aree con popolazione tra 10.000 e 12.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	61,77	95,60	153,27	56,95	92,67	145,06
0,05% ├ 0,1%	41,96	51,39	79,19	37,64	50,99	86,02
0,1% ├ 0,25%	26,85	34,58	54,97	21,98	34,67	62,40
0,25% ├ 0,5%	18,92	23,36	29,79	13,65	24,00	55,71
0,5% ├ 1%	12,67	16,22	26,82	10,86	17,71	44,72
1% ├ 2,5%	5,26	10,62	14,32	4,92	12,03	31,05
2,5% ├ 5%	4,08	7,05	9,54	3,66	8,31	23,68
5% ├ 10%	2,13	4,81	6,43	2,24	5,92	21,30
10% ├ 15%	1,48	3,12	4,64	1,79	4,10	11,46
15% ├ 20%	1,25	2,31	3,47	1,44	3,14	10,44
20% ├ 30%	1,01	2,02	3,00	0,87	2,73	9,28
≥ 30%	0,73	1,32	2,16	0,82	1,63	5,75

Tavola 9.10 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSFAM e CCSSEZ nel caso della frazione sondata del 33%, per stime riferite ad aree di censimento tra 12mila e 15mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=33% 25 comuni 107 aree con popolazione tra 12.000 e 15.000			Disegno **CCSSEZ** f.s.=33% 25 comuni 107 aree con popolazione tra 12.000 e 15.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	51,72	82,31	153,17	45,13	80,05	158,12
0,05% ├ 0,1%	36,16	46,19	68,23	30,01	45,06	77,96
0,1% ├ 0,25%	23,31	30,62	52,63	19,46	30,94	63,81
0,25% ├ 0,5%	15,78	20,98	34,23	13,62	22,02	61,88
0,5% ├ 1%	11,12	14,76	21,23	9,29	16,22	36,43
1% ├ 2,5%	4,74	9,49	13,74	4,41	10,99	35,18
2,5% ├ 5%	3,47	6,35	9,01	3,49	7,57	39,74
5% ├ 10%	1,91	4,34	6,01	2,10	5,29	34,17
10% ├ 15%	1,39	2,81	4,07	1,55	3,87	13,66
15% ├ 20%	1,14	2,18	3,34	1,42	3,07	14,94
20% ├ 30%	0,85	1,79	2,77	0,80	2,37	11,99
≥ 30%	0,73	1,18	2,08	0,67	1,53	6,22

9.2.2 Valutazione dell'introduzione della stratificazione (effetto disegno)

Il confronto tra i *cv* mediani dei disegni campionari per famiglia porta a ritenere che la possibilità di stratificare la lista delle famiglie, o per numero di componenti o per classi di età del capofamiglia, non produce effetti rilevanti in termini di una riduzione dell'errore (Tavola 9.11).

Tavola 9.11 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSFAM, STRNCOMP e STRETACAP nel caso della frazione sondata del 33%

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=33%			Disegno **STRNCOMP** f.s.=33%			Disegno **STRETACAP** f.s.=33%		
	25 comuni 229 aree			25 comuni 229 aree			25 comuni 229 aree		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	51,72	96,52	154,50	52,01	96,72	157,78	51,35	96,29	158,90
0,05% ├ 0,1%	36,16	51,02	84,55	36,75	50,99	82,72	37,02	51,08	86,37
0,1% ├ 0,25%	23,31	34,59	98,26	22,96	34,69	90,14	23,14	34,71	98,44
0,25% ├ 0,5%	15,78	23,33	41,37	15,68	23,26	41,71	16,18	23,33	40,43
0,5% ├ 1%	11,12	16,37	33,04	11,33	16,35	33,65	11,27	16,33	33,16
1% ├ 2,5%	4,74	10,65	20,33	4,65	10,63	21,47	4,80	10,60	20,84
2,5% ├ 5%	3,47	7,01	12,75	3,51	6,99	12,58	3,51	6,99	13,06
5% ├ 10%	1,91	4,80	8,91	1,90	4,81	9,39	1,87	4,79	9,27
10% ├ 15%	1,39	3,10	6,12	1,32	3,09	6,07	1,37	3,08	6,25
15% ├ 20%	1,14	2,41	4,81	1,18	2,39	4,66	1,17	2,41	4,71
20% ├ 30%	0,85	1,93	3,85	0,87	1,93	3,77	0,86	1,93	3,78
≥ 30%	0,73	1,34	2,90	0,73	1,34	2,99	0,71	1,34	2,95

Invece nell'ambito dei disegni campionari di tipo areale, le stratificazioni delle sezioni proposte producono effetti moderati sull'efficienza attesa delle stime; in particolare, il disegno STRSPOP risulta essere generalmente migliore rispetto ai disegni CCSSEZ e STRSSEZ per le stime riferite al livello di area (Tavola 9.12).

Tavola 9.12 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nei disegni CCSSEZ, STRPOP e STRSEZ nel caso della frazione sondata del 33%

Coefficiente di Variazione	Disegno **CCSSEZ** f.s.=33%			Disegno **STRPOP** f.s.=33%			Disegno **STRSEZ** f.s.=33%		
	25 comuni 229 aree			25 comuni 229 aree			25 comuni 229 aree		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	45,13	94,19	158,12	33,81	89,60	154,88	47,61	92,63	159,24
0,05% ├ 0,1%	30,01	51,41	110,25	28,41	50,29	97,56	26,94	50,73	113,69
0,1% ├ 0,25%	19,46	34,87	111,55	10,96	33,81	91,48	16,17	34,69	116,47
0,25% ├ 0,5%	13,62	24,30	80,53	12,75	23,64	83,27	10,89	24,01	93,96
0,5% ├ 1%	9,29	18,09	89,68	8,45	17,54	73,16	7,76	17,90	63,30
1% ├ 2,5%	4,41	12,25	64,02	4,24	11,96	65,18	4,37	12,12	70,81
2,5% ├ 5%	3,49	8,43	65,76	3,42	8,17	66,26	3,78	8,37	60,77
5% ├ 10%	2,10	5,94	52,52	1,74	5,74	52,93	1,92	5,90	48,63
10% ├ 15%	1,55	4,31	20,44	1,50	4,19	20,63	1,46	4,25	19,38
15% ├ 20%	1,42	3,39	22,07	1,34	3,27	24,20	1,38	3,31	23,34
20% ├ 30%	0,80	2,67	13,54	0,80	2,55	16,55	0,81	2,61	14,14
≥ 30%	0,67	1,74	11,84	0,62	1,65	11,52	0,64	1,68	8,74

9.3 Risultati del III° Blocco: confronti di efficienza di disegni semplici da lista per differenti frazioni di campionamento (ipotesi del 10%, 15%, 20% e 33%)

In considerazione dei risultati precedenti che hanno evidenziato un vantaggio nell'utilizzo del disegno casuale semplice da lista di famiglie, si è scelto di valutare l'efficienza delle stime per differenti frazioni di campionamento (10%, 15%, 20% e 33%) solo per tale disegno. I confronti

hanno riguardato i risultati delle sperimentazioni condotte sui dati di 25 comuni in totale (Tavole 9.13-9.14). Per ciascuna delle 4 frazioni di campionamento testate sono stati presi in considerazione i risultati relativi a 10 comuni di cui 5 in comune e 5 differenti (Tavola 7.3).

Tavola 9.13 - Distribuzione delle aree di censimento per dimensione demografica, per le frazioni sondate sperimentate con il disegno CCSFAM (III° Blocco)

	III° Blocco di sperimentazioni: aree di censimento dei comuni coinvolti			
Disegno **CCSFAM**	Aree con popolazione tra 5.000 e 10.000	Aree con popolazione tra 10.000 e 12.000	Aree con popolazione tra 12.000 e 15.000	Totale
f.s.=10% (10 comuni)	57	37	76	170
f.s.=15% (10 comuni)	48	38	55	141
f.s.=20% (10 comuni)	41	23	47	111
f.s.=33% (10 comuni)	78	32	94	204

Tavola 9.14 - Distribuzione delle frequenze relative da stimare per classi di p e per dimensione di area di censimento, per le frazioni sondate sperimentate con il disegno CCSFAM (III° Blocco)

	III° Blocco di sperimentazioni: frequenze relative oggetto di stima sulle aree di censimento dei comuni coinvolti							
Classi di p	Disegno **CCSFAM**; f.s.=10% (10 comuni)				Disegno **CCSFAM**; f.s.=15% (10 comuni)			
	Aree 5-10mila	Aree 10-12mila	Aree 12-15mila	Totale	Aree 5-10mila	Aree 10-12mila	Aree 12-15mila	Totale
< 0,05%	313	223	432	968	271	179	325	775
0,05% ├ 0,1%	132	97	199	428	130	89	146	365
0,1% ├ 0,25%	337	216	417	970	266	211	287	764
0,25% ├ 0,5%	346	239	481	1066	323	231	363	917
0,5% ├ 1%	443	302	563	1308	369	273	419	1061
1% ├ 2,5%	917	597	1180	2694	838	570	906	2314
2,5% ├ 5%	895	563	1204	2662	762	533	856	2151
5% ├ 10%	617	433	813	1863	534	404	564	1502
10% ├ 15%	340	232	434	1006	303	228	312	843
15% ├ 20%	223	133	277	633	176	121	210	507
20% ├ 30%	300	201	401	902	289	214	294	797
≥ 30%	267	184	349	800	239	187	268	694
Totale	5130	3420	6750	15300	4500	3240	4950	12690

Tavola 9.14 - Distribuzione delle frequenze relative da stimare per classi di p e per dimensione di area di censimento, per le frazioni sondate sperimentate con il disegno CCSFAM (III° Blocco) (segue)

III° Blocco di sperimentazioni: frequenze relative oggetto di stima sulle aree di censimento dei comuni coinvolti *(segue)*

Classi di p	Disegno **CCSFAM**; f.s.=20% (10 comuni)				Disegno **CCSFAM**; f.s.=33% (10 comuni)			
	Aree 5-10mila	Aree 10-12mila	Aree 12-15mila	Totale	Aree 5-10mila	Aree 10-12mila	Aree 12-15mila	Totale
< 0,05%	195	109	220	524	385	157	450	992
0,05% ├ 0,1%	111	60	109	280	200	85	243	528
0,1% ├ 0,25%	263	149	278	690	473	188	533	1194
0,25% ├ 0,5%	259	152	274	685	488	208	598	1294
0,5% ├ 1%	250	153	289	692	477	204	618	1299
1% ├ 2,5%	686	371	688	1745	1210	499	1422	3131
2,5% ├ 5%	660	370	737	1767	1182	470	1423	3075
5% ├ 10%	474	292	526	1292	971	394	1221	2586
10% ├ 15%	271	150	272	693	433	192	502	1127
15% ├ 20%	114	70	127	311	264	97	321	682
20% ├ 30%	276	148	288	712	449	195	555	1199
≥ 30%	221	136	242	599	488	191	574	1253
Totale	3780	2160	4050	9990	7020	2880	8460	18360

I livelli più bassi di *cv* si registrano, come ovvio, per la frazione di campionamento pari al 33% in quanto il campione ha una numerosità più elevata.

Un risultato da evidenziare è che all'aumentare della frazione sondata corrisponde un guadagno di efficienza meno che proporzionale. In particolare, triplicando la frazione sondata dal 10% al 33%, la precisione delle stime si incrementa in misura poco più che doppia (Tabella 9.15).

Queste indicazioni potranno essere utili nella scelta della frazione di campionamento, nel caso in cui questa derivasse da una soluzione di compromesso tra costo statistico e costo finanziario.

Le Tavole 9.16-9.18 descrivono i risultati dei livelli di efficienza osservati nella sperimentazione del disegno CCSFAM per le frazioni di campionamento fissate, per stime riferite ad aree di censimento aventi differenti dimensioni (5.000-10.000 ; 10.000-12.000 ; 12.000-15.000).

Le considerazioni sono analoghe rispetto a quelle già evidenziate nel precedente paragrafo; i risultati sono uniformemente migliori sulle aree più grandi.

Tavola 9.15 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nel caso del disegno CCSFAM per differenti frazioni sondate (10%, 15%, 20% e 33%)

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=10%			Disegno CCSFAM f.s.=15%			Disegno **CCSFAM** f.s.=20%			Disegno **CCSFAM** f.s.=33%		
	10 comuni 170 aree			10 comuni 141 aree			10 comuni 111 aree			10 comuni 204 aree		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	113,48	220,51	845,23	87,43	157,20	322,46	75,87	142,00	232,53	51,72	98,21	154,50
0,05% ├ 0,1%	81,13	111,48	196,37	63,79	87,22	157,85	51,86	74,20	129,82	36,16	51,14	84,55
0,1% ├ 0,25%	49,31	75,57	184,49	39,14	59,83	122,29	32,76	49,97	102,56	23,31	34,76	98,26
0,25% ├ 0,5%	35,28	50,70	92,19	27,71	39,92	71,71	23,54	33,97	57,62	15,78	23,44	41,37
0,5% ├ 1%	24,28	35,54	68,12	19,10	28,10	54,84	16,33	23,74	42,85	11,12	16,56	33,04
1% ├ 2,5%	13,07	23,62	45,85	10,83	18,56	37,08	6,82	15,33	31,65	4,74	10,68	20,33
2,5% ├ 5%	7,52	15,50	31,61	5,92	12,29	22,56	5,18	10,09	19,11	3,47	7,04	12,75
5% ├ 10%	4,22	10,46	23,36	3,25	8,26	16,01	2,83	6,93	14,86	1,91	4,82	8,91
10% ├ 15%	2,94	7,06	13,92	2,38	5,40	11,04	1,99	4,40	9,00	1,39	3,13	6,12
15% ├ 20%	2,50	5,57	12,52	2,08	4,27	7,95	1,75	3,54	7,03	1,14	2,42	4,81
20% ├ 30%	1,88	4,50	9,88	1,51	3,48	7,43	1,31	2,84	5,96	0,87	1,93	3,85
≥ 30%	1,64	3,20	7,10	1,26	2,42	5,39	1,04	1,94	4,57	0,73	1,34	2,90

Tavola 9.16 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nel caso del disegno CCSFAM per differenti frazioni sondate (10%, 15%, 20% e 33%) per stime riferite ad aree di censimento con dimensione compresa tra 5mila e 10mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=10%			Disegno **CCSFAM** f.s.=15%			Disegno **CCSFAM** f.s.=20%			Disegno **CCSFAM** f.s.=33%		
	10 comuni 57 aree con popolazione tra 5.000 e 10.000			10 comuni 48 aree con popolazione tra 5.000 e 10.000			10 comuni 41 aree con popolazione tra 5.000 e 10.000			10 comuni 78 aree con popolazione tra 5.000 e 10.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	149,46	233,15	775,52	113,49	177,68	283,50	100,57	149,83	232,53	68,25	102,20	154,50
0,05% ├ 0,1%	100,83	136,29	196,37	76,29	109,01	157,85	66,07	88,57	129,82	45,39	61,14	84,55
0,1% ├ 0,25%	63,99	90,00	184,49	48,32	71,27	122,29	41,15	59,48	102,56	27,74	40,20	98,26
0,25% ├ 0,5%	42,60	60,88	92,19	35,39	47,23	71,71	29,80	40,24	57,62	20,29	27,29	41,37
0,5% ├ 1%	30,59	43,11	68,12	24,13	33,50	54,84	20,03	28,97	42,85	14,10	19,54	33,04
1% ├ 2,5%	16,58	28,85	45,85	13,45	21,82	37,08	9,07	18,11	31,65	5,46	12,42	20,33
2,5% ├ 5%	10,30	19,12	31,61	7,94	14,68	22,56	6,30	12,22	19,11	3,89	8,25	12,75
5% ├ 10%	5,43	12,70	23,36	4,30	9,77	16,01	3,54	8,14	14,86	2,41	5,66	8,91
10% ├ 15%	3,45	8,78	13,92	2,69	6,44	11,04	2,23	5,22	9,00	1,58	3,67	6,12
15% ├ 20%	3,62	6,79	12,52	2,74	5,03	7,95	2,27	4,31	7,03	1,55	2,85	4,81
20% ├ 30%	2,43	5,46	9,88	2,02	4,16	7,43	1,66	3,44	5,96	1,01	2,27	3,85
≥ 30%	2,08	3,86	7,10	1,62	2,84	5,39	1,29	2,32	4,57	0,89	1,56	2,90

Tavola 9.17 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nel caso del disegno CCSFAM per differenti frazioni sondate (10%, 15%, 20% e 33%) per stime riferite ad aree di censimento con dimensione compresa tra 10mila e 12mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=10% 10 comuni 37 aree con popolazione tra 10.000 e 12.000			Disegno **CCSFAM** f.s.=15% 10 comuni 38 aree con popolazione tra 10.000 e 12.000			Disegno **CCSFAM** f.s.=20% 10 comuni 23 aree con popolazione tra 10.000 e 12.000			Disegno **CCSFAM** f.s.=33% 10 comuni 32 aree con popolazione tra 10.000 e 12.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	135,27	225,61	363,13	106,18	145,75	265,21	90,31	123,95	216,64	61,77	88,14	153,27
0,05% ├ 0,1%	89,49	107,97	160,62	70,97	85,49	126,04	58,97	72,90	107,32	42,05	50,61	79,19
0,1% ├ 0,25%	58,39	76,23	108,17	44,36	60,03	86,90	37,72	50,45	73,86	26,85	34,41	49,64
0,25% ├ 0,5%	40,33	50,87	62,67	32,41	39,73	52,22	27,50	34,10	40,97	18,92	23,40	29,30
0,5% ├ 1%	27,87	35,08	48,12	21,64	27,46	38,78	18,00	22,99	31,87	13,04	16,48	26,82
1% ├ 2,5%	15,36	23,37	31,98	11,28	18,00	26,81	7,63	14,92	20,61	5,26	10,62	14,32
2,5% ├ 5%	8,60	15,58	20,77	7,22	12,34	16,38	5,80	9,90	13,61	4,08	7,08	9,54
5% ├ 10%	4,90	10,36	15,84	3,46	8,01	11,10	3,02	6,90	9,34	2,23	4,82	6,43
10% ├ 15%	3,35	7,00	10,58	2,53	5,46	7,78	2,13	4,41	6,59	1,48	3,13	4,64
15% ├ 20%	2,72	5,47	8,35	2,15	4,18	6,11	1,83	3,41	5,07	1,25	2,32	3,47
20% ├ 30%	2,16	4,57	6,86	1,73	3,42	5,14	1,51	2,94	3,96	1,01	2,01	3,00
≥ 30%	1,74	3,14	4,70	1,33	2,30	3,60	1,12	1,88	3,13	0,73	1,32	2,16

Tavola 9.18 - Confronto tra i livelli di cv (minimo, mediano e massimo) osservati nel caso del disegno CCSFAM per differenti frazioni sondate (10%, 15%, 20% e 33%) per stime riferite ad aree di censimento con dimensione compresa tra 12mila e 15mila unità

Coefficiente di Variazione	Disegno **CCSFAM** f.s.=10% 10 comuni 76 aree con popolazione tra 12.000 e 15.000			Disegno **CCSFAM** f.s.=15% 10 comuni 55 aree con popolazione tra 12.000 e 15.000			Disegno **CCSFAM** f.s.=20% 10 comuni 47 aree con popolazione tra 12.000 e 15.000			Disegno **CCSFAM** f.s.=33% 10 comuni 94 aree con popolazione tra 12.000 e 15.000		
Classi di *p*	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo	minimo	mediana	massimo
< 0,05%	113,48	184,05	845,23	87,43	141,61	322,46	75,87	117,45	209,98	51,72	82,77	153,17
0,05% ├ 0,1%	81,13	100,88	138,21	63,79	79,37	108,91	51,86	64,90	103,55	36,16	46,30	68,23
0,1% ├ 0,25%	49,31	66,65	97,81	39,14	53,04	77,28	32,76	43,52	63,62	23,31	30,58	52,63
0,25% ├ 0,5%	35,28	45,22	58,97	27,71	35,68	47,81	23,54	29,94	38,72	15,78	21,01	34,23
0,5% ├ 1%	24,28	31,25	43,27	19,10	24,95	34,40	16,33	20,97	31,60	11,12	14,85	21,23
1% ├ 2,5%	13,07	21,10	31,20	10,83	16,46	23,74	6,82	13,45	18,68	4,74	9,48	13,59
2,5% ├ 5%	7,52	14,00	18,99	5,92	10,98	15,17	5,18	9,06	12,23	3,47	6,35	8,53
5% ├ 10%	4,22	9,37	13,57	3,25	7,43	10,77	2,83	6,21	8,54	1,91	4,35	6,01
10% ├ 15%	2,94	6,29	10,02	2,38	4,79	7,18	1,99	3,89	5,65	1,39	2,83	4,03
15% ├ 20%	2,50	4,99	7,76	2,08	3,80	5,52	1,75	3,09	4,77	1,14	2,17	3,34
20% ├ 30%	1,88	4,05	6,26	1,51	3,21	4,79	1,31	2,59	4,04	0,87	1,77	2,77
≥ 30%	1,64	2,83	4,52	1,26	2,17	3,44	1,04	1,74	2,94	0,73	1,19	2,08

La Tavola 9.19 fornisce una sintesi del confronto tra i livelli di efficienza attesi delle stime osservati nella sperimentazione per il disegno CCSFAM e per le differenti frazioni sondate. In particolare:

- l'incremento della frazione sondata dal 10% al 15% porta ad una riduzione dell'errore mediamente compreso tra il 21% e il 28%;
- il raddoppio del campione dalla frazione sondata del 10% alla frazione sondata del 20% porta ad una guadagno di efficienza compreso tra il 33% e il 39%;
- il passaggio dalla frazione sondata del 10% a quella più elevata del 33% conduce ad una riduzione del coefficiente di variazione atteso delle stime compreso tra il 53% e il 58%.

Le Figure 9.2-9.4 descrivono l'andamento delle curve di livello dei *cv* mediani attesi nel caso del disegno CCSFAM (per le frazioni sondate del 10%, 15%, 20% e 33%) delle stime rispettivamente per le aree di censimento tra 5mila e 10mila abitanti, per le aree di censimento tra 10mila e 12mila abitanti e per le aree di censimento tra 12mila e 15mila abitanti. A riguardo:

- il guadagno di efficienza (in termini di *cv*) per aree di censimento tra 10mila e 12mila abitanti rispetto alle aree tra 5mila e 10mila abitanti è compreso tra il 14% e il 20%;
- il guadagno di efficienza per aree di censimento tra 12mila e 15mila abitanti rispetto alle aree tra 5mila e 10mila abitanti è invece compreso tra il 22% e il 28%.

Tavola 9.19 - Guadagno di efficienza (dato dalla riduzione dell'errore campionario misurato tramite il cv atteso) delle stime riferite all'area di censimento nel caso di incremento della frazione sondata dal 10% rispettivamente al 15%, al 20% e al 33% (disegno CCSFAM)

Riduzione percentuale(*) del valore di *cv* per l'incremento della f.s. rispetto al livello del 10% (disegno **CCSFAM**)

Classi di *p*	Incremento della f.s. dal 10% al 15%	Incremento della f.s. dal 10% al 20%	Incremento della f.s. dal 10% al 33%
< 0,05%	28,71	35,60	55,46
0,05% ├ 0,1%	21,76	33,44	54,13
0,1% ├ 0,25%	20,83	33,88	54,00
0,25% ├ 0,5%	21,26	33,00	53,77
0,5% ├ 1%	20,93	33,20	53,40
1% ├ 2,5%	21,42	35,10	54,78
2,5% ├ 5%	20,71	34,90	54,58
5% ├ 10%	21,03	33,75	53,92
10% ├ 15%	23,51	37,68	55,67
15% ├ 20%	23,34	36,45	56,55
20% ├ 30%	22,67	36,89	57,11
≥ 30%	24,38	39,38	58,13

(*) Calcolata da: [*cv*(CCSFAM_f.s. 10%)-*cv*(CCSFAM_f.s. K%)]x100/[*cv*(CCSFAM_f.s. 10%)] , dove K=15, 20, 33 a seconda della situazione presa in esame.

Figura 9.2 - Curve di livello dei cv attesi (valori mediani) nel caso del disegno CCSFAM (per le frazioni sondate del 10%, 15%, 20% e 33%) delle stime per aree di censimento tra 5mila e 10mila abitanti

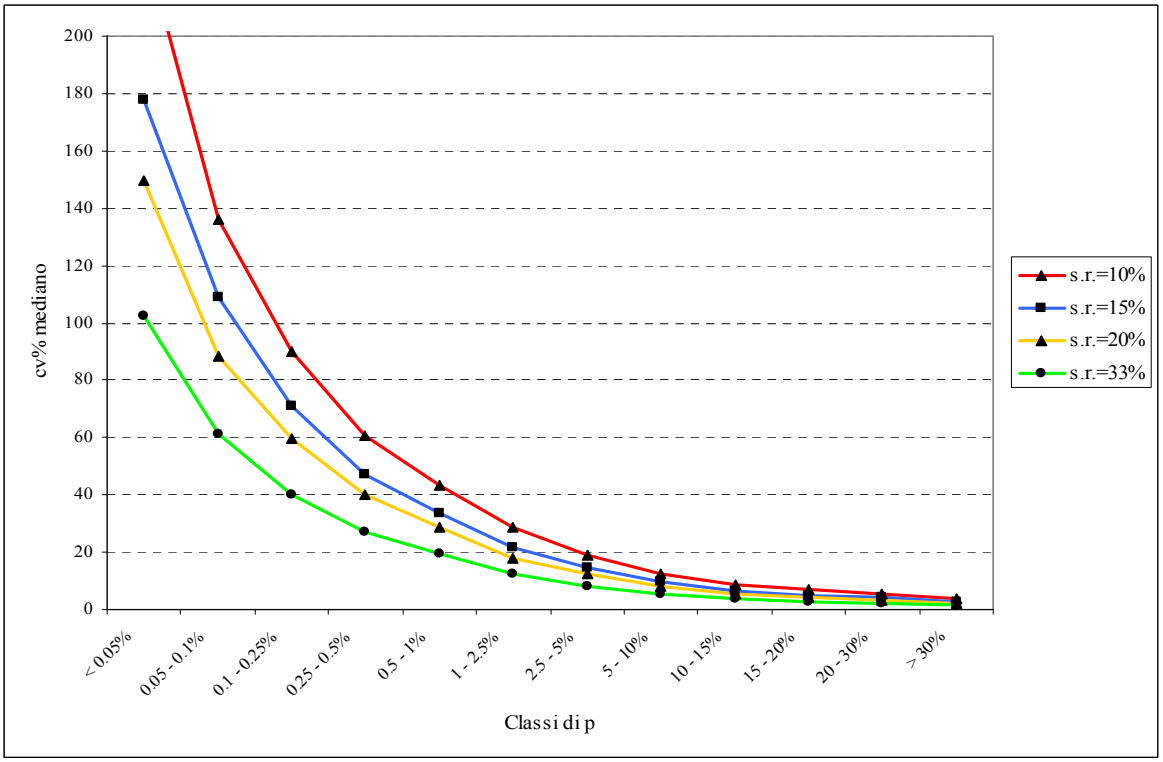


Figura 9.3 - Curve di livello dei cv attesi (valori mediani) nel caso del disegno CCSFAM (per le frazioni sondate del 10%, 15%, 20% e 33%) delle stime per aree di censimento tra 10mila e 12mila abitanti

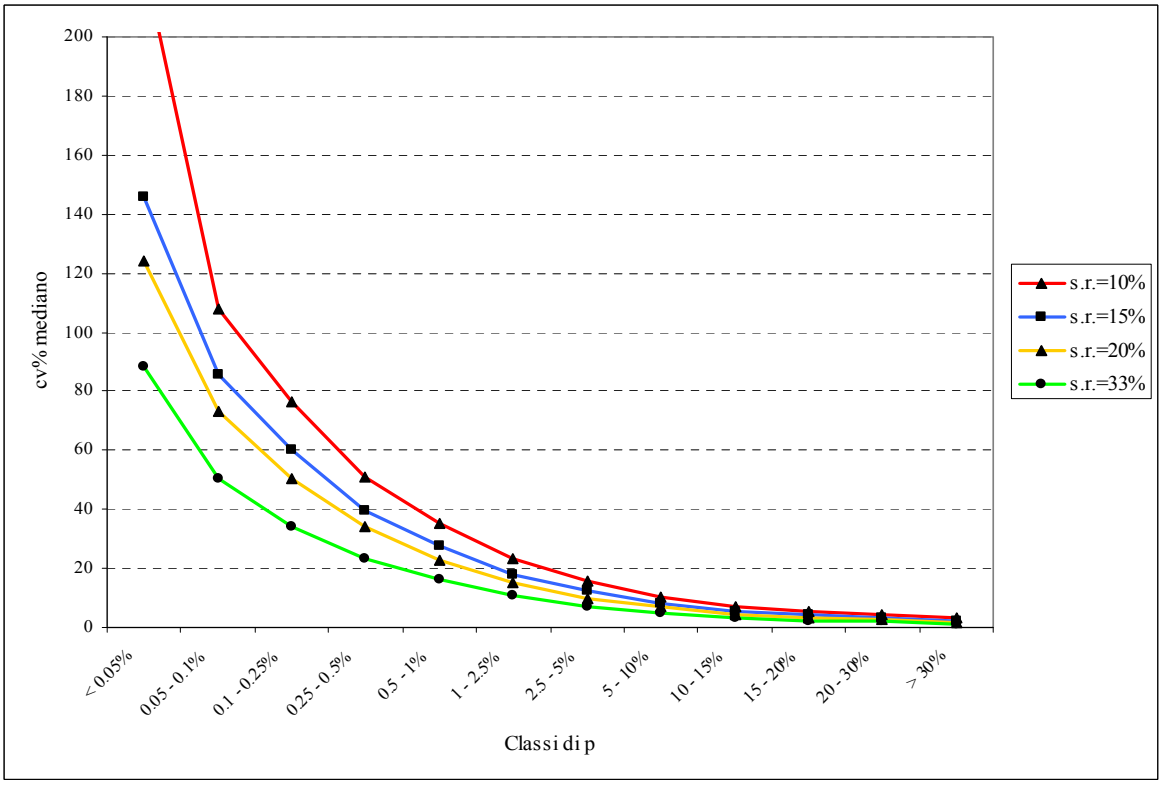
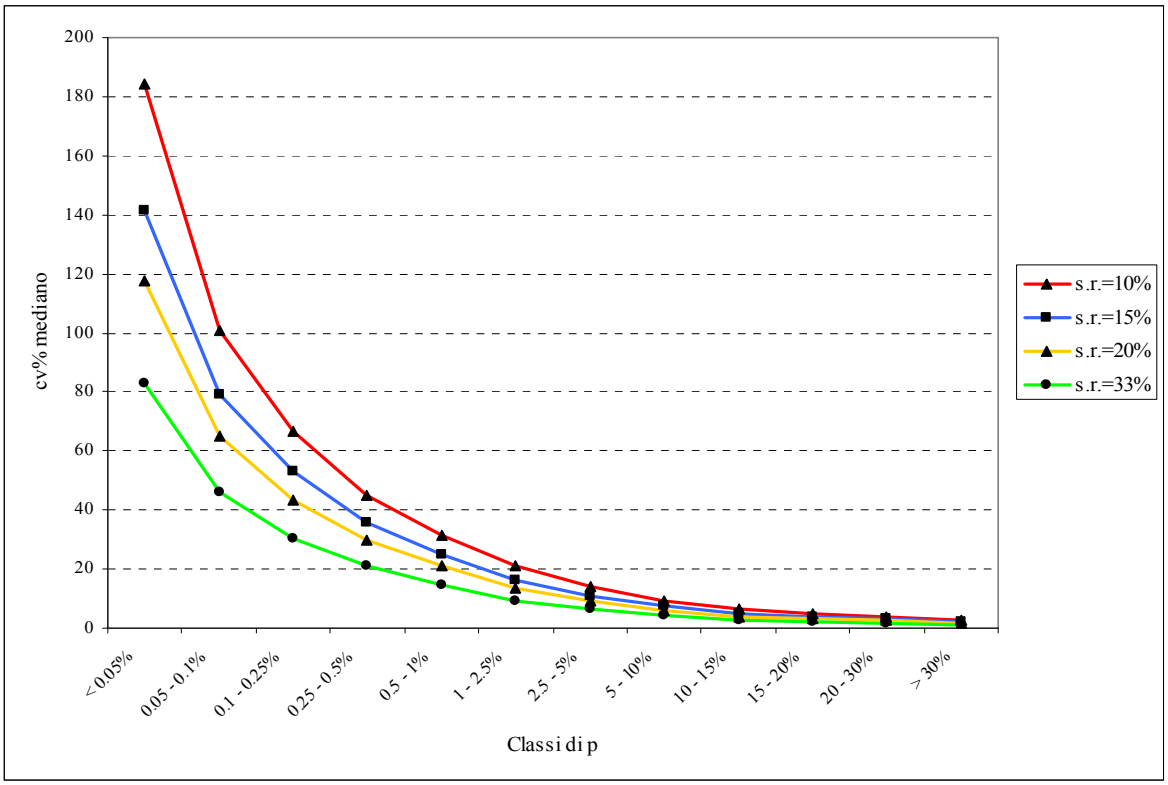


Figura 9.4 - Curve di livello dei cv attesi (valori mediani) nel caso del disegno CCSFAM (per le frazioni sondate del 10%, 15%, 20% e 33%) delle stime per aree di censimento tra 12mila e 15mila abitanti



10. Considerazioni sull'efficienza delle stime riferite alle aree di censimento più grandi

In questo capitolo viene presentata un'analisi sull'efficienza delle stime riferite solo alle aree di censimento con dimensione compresa tra 12mila e 15mila unità (Carbonetti e Fortini, 2008b). A tal scopo, a partire dai risultati delle sperimentazioni ottenuti con il disegno CCSFAM e per ciascuna delle quattro frazioni di campionamento testate (10%, 15%, 20%, 33%), si è preliminarmente proceduto a classificare le stime delle frequenze percentuali oggetto di studio in base al valore del coefficiente di variazione desunto dalla distribuzione delle stime campionarie simulata.

Tavola 10.1 - Distribuzione delle stime riferite alle aree di censimento comprese tra 12mila e 15mila abitanti, per classi di cv. Confronto delle frequenze percentuali per quattro frazioni di campionamento nel disegno CCSFAM

Classi di *cv%*	Frazione di campionamento			
	10%	15%	20%	33%
< 2%	0,57	2,69	6,39	13,14
2% ├ 5%	13,04	17,53	18,40	23,64
5% ├ 10%	16,18	18,02	26,28	28,64
10% ├ 20%	29,14	30,16	23,54	16,20
20% ├ 50%	25,09	19,71	16,75	13,32
50% ├ 100%	9,32	7,21	5,69	3,44
100% ├ 200%	4,40	3,65	2,00	1,61
≥ 200%	2,25	1,03	0,95	-

La Tavola 10.1 descrive le distribuzioni delle stime delle percentuali, raggruppate in classi di *cv* per ogni frazione di campionamento testata nel disegno CCSFAM.

Esempio: nel caso della frazione sondata più bassa (10%) il 16,18% delle stime delle frequenze relative coinvolte nelle sperimentazioni mostrano un *cv* compreso tra il 5% e il 10%.

In generale, si osserva che all'incremento del campione (dalla f.s.=10% alla f.s.=33%) fa seguito un uniforme aumento dei casi delle classi con livelli di *cv* inferiori alla soglia del 10%; queste classi (< 2%; 2%├ 5%; 5%├ 10%) potrebbero rappresentare stime con *livelli alti* di accuratezza. Differenti livelli di accuratezza potrebbero essere considerati per le altre classi: *livelli medi* per le classi di *cv* tra il 10% e il 50% (10%├ 20%; 20%├ 50%); *livelli bassi* per le classi di *cv* superiori al 50% (50%├ 100%; 100%├ 200%; ≥ 200%).

La Tavola 10.2 riassume in forma aggregata gli stessi risultati della Tavola 10.1: in questo caso i risultati sono presentati in 3 classi di *cv* relative ai 3 livelli di accuratezza proposti (*alta*, *media*, *bassa*).

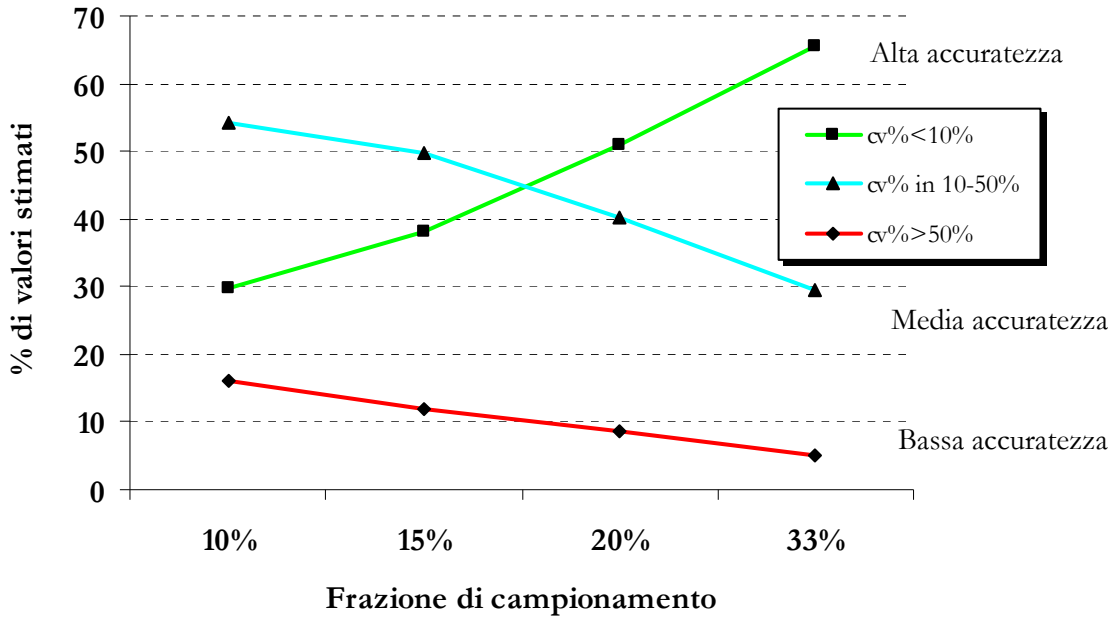
Esempio: al crescere della dimensione del campione dalla frazione sondata del 10% a quella del 33%, le stime con elevata accuratezza (*cv%*<10%) aumentano dal 29,80% al 65,42% dei casi studiati; invece, i casi di stime con bassa accuratezza (*cv%*>50%) si riducono dal 15,97% al 5,05%.

Questi risultati sono descritti in modo più evidente nel Figura 10.1.

Tavola 10.2 - Distribuzione delle stime riferite alle aree di censimento comprese tra 12mila e 15mila abitanti, per livelli di accuratezza. Confronto delle frequenze percentuali per 4 frazioni di campionamento nel disegno CCSFAM

Livelli di accuratezza *(classe di cv%)*	frazione di campionamento			
	10%	15%	20%	33%
Alta *(cv%<10%)*	29,80	38,25	51,07	65,42
Media *(cv% in 10-50%)*	54,23	49,87	40,29	29,52
Bassa *(cv%>50%)*	15,97	11,89	8,64	5,05

Figura 10.1 - Andamento dei livelli ai accuratezza delle stime al crescere della frazione di campionamento



11. L'errore di campionamento atteso delle stime delle frequenze relative per domini superiori all'area di censimento

Le valutazioni fino ad ora esposte hanno riguardato gli errori di campionamento attesi di stime di percentuali riferite alle aree di censimento di centro abitato. In questo capitolo, sulla base dei risultati delle sperimentazioni ottenute nel caso del disegno CCSFAM e della frazione di campionamento del 33%, si è valutato l'errore atteso di stime riferite a contesti territoriali di più ampie dimensioni (comune, provincia, regione,…) e per i quali sono diffusi la maggior parte dei risultati censuari.

In generale (Carbonetti e De Vitiis, 2007), un qualunque territorio R può essere "immaginato" come in uno dei seguenti 2 casi:

1) aggregazione di aree di censimento sottoposte a campionamento (pari al numero K>1);
2) aggregazione di aree di censimento sottoposte a campionamento e di aree non sottoposte a campionamento (per esempio: le aree sotto i 5mila abitanti o le aree di pertinenza delle zone extra-urbane e periferiche).

Nell'ipotesi 1), poiché il territorio R=R_C è aggregazione solo di aree campionabili, la stima di p può essere espressa nella forma seguente:

$$\hat{p}_{R_C}(x) = \sum_{a \in R_C} w_a \hat{p}_a \qquad (11.1)$$

dove $w_a = N_a \big/ N_{R_C}$ esprime il peso della generica area a sul territorio R_C in termini di popolazione.

In base a semplici sviluppi teorici, supportati dai riscontri empirici, si dimostra che, per un dato livello di percentuale p (identico sia per l'area di censimento che per il territorio R_C) il coefficiente di variazione cv della sua stima campionaria (con il disegno CCSFAM e una frazione di campionamento del 33%) è sviluppabile nel seguente modo:

$$cv(\hat{p}_{R_C}) \cong \frac{1}{\sqrt{K}} cv(\hat{p}_a) \qquad (11.2)$$

dove K è il numero di aree di censimento sottoposte a campionamento. Per il livello di cv è quindi attesa la seguente riduzione percentuale:

$$rid\% \cong \left(1 - \frac{1}{\sqrt{K}}\right) \cdot 100 \qquad (11.3)$$

La precedente espressione mette in evidenza che la stima di una prefissata percentuale subisce una riduzione dell'errore di campionamento (in termini di cv) nella misura del numero di aree di censimento campionabili che compongono il territorio R_C e a cui la percentuale da stimare si riferisce. Questa formula permette, a partire dall'errore atteso per la stima di un dato livello di p riferito al dominio minimo (l'area di censimento), di determinare l'errore atteso per la stima dello stesso livello di p riferito però ad un dominio più ampio comprendente un determinato numero di aree (K) tutte campionate.

Nell'ipotesi 2) di un territorio R costituito in parte (R_C) dalla presenza di aree campionabili e in parte (R_{NC}) da aree non campionabili ($R = R_C \cup R_{NC}$), indicando con γ la quota di popolazione di R soggetta a campionamento per la rilevazione con *long form* ($\gamma = N_C / N$), la stima di p su R sarà data da:

$$\hat{p}_R = \gamma \hat{p}_C + (1-\gamma) p_{NC} \qquad (11.4)$$

In questo caso il coefficiente di variazione può essere espresso come segue:

$$cv(\hat{p}_R) \cong \gamma \frac{1}{\sqrt{K}} cv(\hat{p}_a) \qquad (11.5)$$

con una riduzione percentuale attesa pari a:

$$rid\% \cong \left(1 - \frac{\gamma}{\sqrt{K}}\right) \cdot 100 \qquad (11.6)$$

La precedente espressione mostra che la riduzione dell'errore di campionamento delle stima di una data percentuale è dipendente oltre che dal numero di aree di censimento campionabili che sono presenti nel territorio R, anche dalla loro quota γ di popolazione complessiva rispetto alla popolazione totale di R. Anche questa formula permette di stimare l'errore atteso della stima di una percentuale p riferita però ad un territorio in cui ricadono K aree di censimento e su cui è presente la quota γ di popolazione rispetto a quella di R.

Le Tavole 11.1-11.2 riportano alcuni esempi di coefficienti di variazione attesi determinati nel caso di stime riferite a domini superiori all'area di censimento per alcuni livelli di percentuale.

Per il dominio comunale, gli esempi presentati nella Tavola 11.1 mettono in evidenza una riduzione percentuale del *cv* della stima riferita al livello di comune, riduzione tanto più grande quanto maggiore è il numero di aree campionabili K presenti nel comune: per un qualsiasi comune costituito solo da aree campionabili (*caso 1*), la riduzione di *cv* passerebbe dal 42,26% per il comune di Alghero (3 aree) al 90,51% per il comune di Milano (111 aree); nel caso invece di un comune composto da aree campionabili e aree non campionabili e nell'ulteriore ipotesi che il 90% della popolazione ($\gamma=0,9$) ricada nelle aree sottoposte a campionamento (*caso 2a*), la riduzione di *cv* passerebbe dal 48,04% per Alghero al 91,46% per Milano. E' evidente che la presenza di una porzione di territorio non sottoposta a campionamento favorisce un ulteriore riduzione dell'errore campionario atteso a vantaggio dell'efficienza delle stime finali.

Supponendo poi, ad esempio, di dover stimare una percentuale pari all'1%, si osserva che: per una generica area di censimento è atteso un *cv* pari al 13,6%; per il comune di Alghero lo stesso valore della percentuale comporterà un *cv* pari a 7,85% nel *caso 1* e pari a 7,07% nel *caso 2a*; per il comune di Milano invece il *cv* atteso sarà pari a 1,29% nel *caso 1* e a 1,16% nel *caso 2a*.

Tavola 11.1 - Coefficiente di variazione atteso delle stime di frequenze relative riferite a domini comunali (disegno CCSFAM ; f.s.=33%). Alcuni esempi dei comuni sottoposti a sperimentazioni

cv attesi delle stime riferite a domini comunali		ALGHERO (3 aree)	TRAPANI (5 aree)	PERUGIA (10 aree)	BOLOGNA (32 aree)	MILANO (111 aree)
		Riduzione % del coefficiente di variazione				
Caso 1		42,26%	55,28%	68,38%	82,32%	90,51%
Caso 2a		48,04%	59,75%	71,54%	84,09%	91,46%
percentuale da stimare	*cv* mediano delle stime riferite alle aree di censimento	*cv* della stima riferita al dominio comunale (nei 2 casi)				
0,1%	45,5	26,27	20,35	14,39	8,04	4,32
		23,64	18,31	12,95	7,24	3,89
0,5%	20,6	11,89	9,21	6,51	3,64	1,96
		10,70	8,29	5,86	3,28	1,76
1%	13,6	7,85	6,08	4,30	2,40	1,29
		7,07	5,47	3,87	2,16	1,16
2%	9,7	5,60	4,34	3,07	1,71	0,92
		5,04	3,90	2,76	1,54	0,83
5%	5,5	3,18	2,46	1,74	0,97	0,52
		2,86	2,21	1,57	0,88	0,47
10%	3,8	2,19	1,70	1,20	0,67	0,36
		1,97	1,53	1,08	0,60	0,32
20%	2,4	1,39	1,07	0,76	0,42	0,23
		1,25	0,97	0,68	0,38	0,21
30%	1,9	1,10	0,85	0,60	0,34	0,18
		0,99	0,76	0,54	0,30	0,16

Caso 1: Territorio composto solo da aree campionabili.

Caso 2a: Territorio composto da aree campionabili e non campionabili (ipotesi γ=0,9)

La determinazione del coefficiente di variazione atteso per stime riferite a domini sovra-comunali ha richiesto una preliminare stima del numero di aree che potrebbero essere "disegnate" secondo le stesse specifiche progettuali che hanno permesso il disegno delle aree di censimento utili alle sperimentazioni presentate in questo documento.

Nella Tavola 11.2 sono descritti alcuni esempi di domini superiori all'ambito del comune: la *Provincia di Milano*, la *Regione Lombardia*, la *Ripartizione Territoriale Nord-Ovest* e l'intero territorio dell'*Italia* (nella Tavola è indicato il numero "presunto" di aree di censimento campionabili).

Così come visto nel caso del dominio comunale, la riduzione percentuale del *cv* della stima riferita al livello di area sarà tanto più elevata quanto maggiore è il numero di aree campionabili *K* presenti nel dominio preso in considerazione. In particolare, nei territori presi in esame la riduzione risulta superiore al 90% sia nel caso in cui il dominio sia costituito solo da aree campionabili (*caso 1*), passando dal 93,97% del contesto provinciale al 98,27% del contesto nazionale, sia nel caso in cui il dominio sia composto da aree campionabili e aree non campionabili e nell'ulteriore ipotesi che il 60% della popolazione ($\gamma=0,6$) ricada nelle aree sottoposte a campionamento (*caso 2b*), incrementando dal 96,38% dell'ambito provinciale al 98,96% di quello nazionale.

Tavola 11.2 - Coefficiente di variazione atteso delle stime di frequenze relative riferite a domini sovra-comunali (disegno CCSFAM ; f.s.=33%). Alcuni esempi

cv attesi delle stime relative a domini sovra-comunali		Provincia di Milano (275 aree*)	Regione Lombardia (446 aree*)	Ripartiz. Territoriale Nord-Ovest (763 aree*)	ITALIA (3347 aree*)
		Riduzione % del coefficiente di variazione			
Caso 1		93,97%	95,26%	96,38%	98,27%
Caso 2b		96,38%	97,16%	97,83%	98,96%
percentuale da stimare	*cv* mediano delle stime riferite alle aree di censimento	*cv* della stima riferita al dominio sovra-comunale (nei 2 casi)			
0,1%	45,5	2,74	2,15	1,65	0,79
		1,65	1,29	0,99	0,47
0,5%	20,6	1,24	0,98	0,75	0,36
		0,75	0,59	0,45	0,21
1%	13,6	0,82	0,64	0,49	0,24
		0,49	0,39	0,30	0,14
2%	9,7	0,58	0,46	0,35	0,17
		0,35	0,28	0,21	0,10
5%	5,5	0,33	0,26	0,20	0,10
		0,20	0,16	0,12	0,06
10%	3,8	0,23	0,18	0,14	0,07
		0,14	0,11	0,08	0,04
20%	2,4	0,14	0,11	0,09	0,04
		0,09	0,07	0,05	0,02
30%	1,9	0,11	0,09	0,07	0,03
		0,07	0,05	0,04	0,02

Caso 1: Territorio composto solo da aree campionabili

Caso 2b: Territorio composto da aree campionabili e non campionabili (ipotesi γ=0,6)

* Numero presunto nell'ipotesi di un disegno di aree di censimento con dimensione tra 5mila e 15mila unità

Si nota in questi esempi che la presenza di una parte del dominio non sottoposta a campionamento favorisce una ulteriore riduzione dell'errore campionario atteso anche se di minima entità, in quanto già ampiamente ridotto per la presenza di un numero molto elevato di aree campionabili.

Fissando ora l'attenzione alla stima della percentuale dell'1% si osserva che: su una generica area di censimento è atteso un *cv* pari al 13,6%; per la provincia di Milano la stima dello stesso valore della percentuale comporterà un *cv* pari a 0,82% nel *caso 1* e pari a 0,49% nel *caso 2b*; per la regione Lombardia il *cv* atteso sarà pari a 0,64% nel *caso 1* e a 0,39% nel *caso 2b*; nella ripartizione Nord-Ovest il *cv* sarà 0,49% nel *caso 1* e 0,30% nel *caso 2b*; infine, per il contesto nazionale il *cv* atteso sarà pari a 0,24% nel *caso 1* e a 0,14% nel *caso 2b*.

12. L'errore di campionamento atteso delle stime delle frequenze assolute

La successiva analisi ha riguardato la valutazione dell'errore di campionamento atteso per la stima delle frequenze assolute (totali). Quindi, sulla base dei risultati delle sperimentazioni (per quanto descritto nel paragrafo 7.6) si è determinato, per ciascuna delle 90 modalità di incrocio (relative alle variabili di studio), il coefficiente di variazione delle stime delle frequenze assolute (i totali T_x) sulle aree di censimento.

In particolare, sono stati considerati nello studio i risultati delle sperimentazioni relativi all'adozione del disegno CCSFAM e per le frazioni sondate del 10%, 20% e 33%. Quindi, dopo

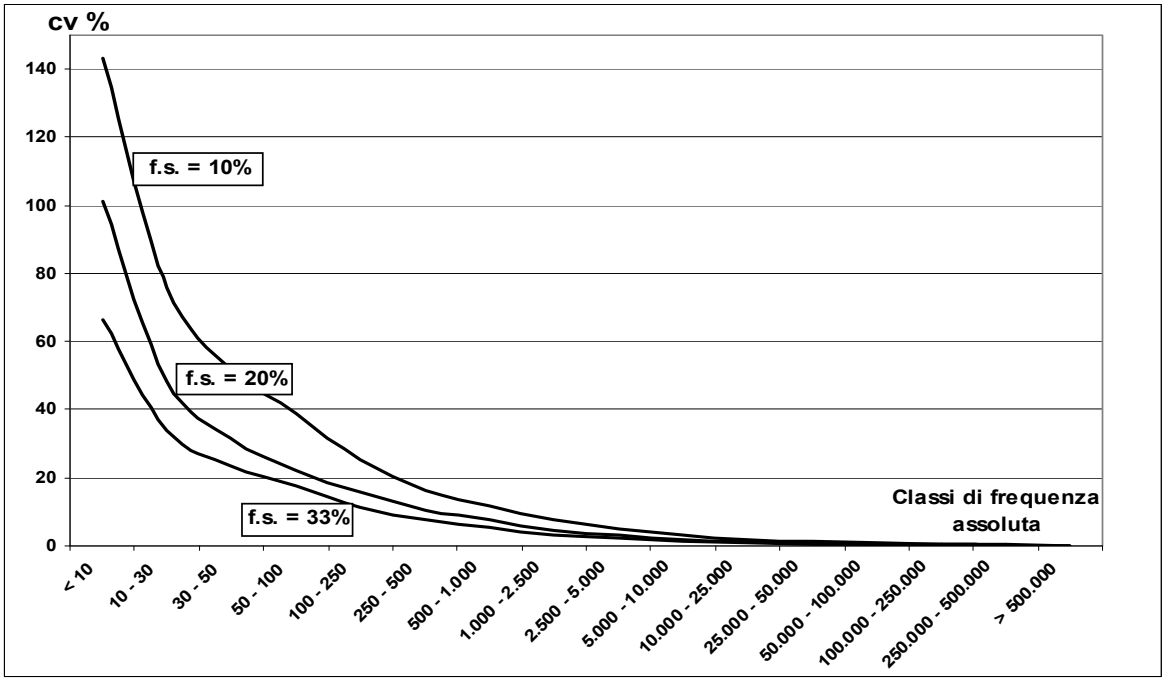
aver classificato le frequenze assolute in 10 classi[27] è stato determinato, per ciascuna classe di *T*, il *cv* mediano e massimo. Con riferimento alle strategie campionarie prese in considerazione, nella Tavola 12.1 sono presentati i risultati ottenuti.

Tavola 12.1 - Distribuzione del coefficiente di variazione Mediano e Massimo per classi di stime di frequenze assolute relative alle aree di censimento (disegno CCSFAM ; frazioni di campionamento del 10%, del 20% e del 33%)

Classi di frequenze assolute T	f.s. = 10%		f.s. = 20%		f.s. = 33%	
	cv% mediano	*cv*% massimo	*cv*% mediano	*cv*% massimo	*cv*% mediano	*cv*% massimo
<10	143,3	191,8	101,4	123,7	66,5	95,8
10 ├ 30	75,9	85,1	48,4	54,6	33,8	38,5
30 ├ 50	51,8	57,1	31,8	37,1	23,4	25,6
50 ├ 100	38,6	41,3	22,3	28,4	17,4	19,1
100 ├ 250	25,4	28,5	15,7	19,6	11,4	12,8
250 ├ 500	16,1	18,3	10,4	12,5	7,5	8,1
500 ├ 1.000	11,8	12,8	7,5	8,2	5,3	5,9
1.000 ├ 2.500	7,5	8,9	4,7	5,9	3,3	3,9
2.500 ├ 5.000	4,9	5,4	3,0	3,6	2,0	2,5
5.000 ├ 10.000	3,2	3,8	2,0	2,5	1,3	1,9

Lo stesso esercizio è stato effettuato anche con riferimento ai domini comunali; in questo modo è stato possibile valutare l'errore atteso di campionamento anche per frequenze assolute superiori a 10.000 unità. Infatti, per nessuna area di censimento è risultata una frequenza assoluta superiore a tale soglia. Il Figura 12.1 riporta le curve degli errori campionari risultanti dalle sperimentazioni.

Figura 12.1 - Curve degli errori di campionamento attesi (misurate dal cv mediano) nel caso del disegno CCSFAM (per le frazioni sondate del 10%, del 20% e del 33%)



[27] Le frequenze assolute *T* oggetto di stima sono state classificate in 10 classi: *T*<10 ; 10≤ *T*<30 ; 30≤ *T*<50 ; 50≤ *T*<100 ; 100≤ *T*<250 ; 250≤ *T*<500 ; 500≤ *T*<1.000 ; 1.000≤ *T*<2.500 ; 2.500≤ *T*<5.000 ; 5.000≤ *T*<10.000 . Si fa presente che, poiché le stime sono riferite alle aree di censimento (domini con popolazione tra 5.000 e 15.000), le frequenze assolute oggetto di stima non sono mai risultate superiori alle 10.000 unità.

13. Alcuni esempi

In questo capitolo sono presentati (Tavole 13.1-13.6) alcuni esempi di stima per intervalli di confidenza delle frequenze assolute per alcune delle variabili oggetto di analisi e riferite sia a livello di area di censimento (prendendo a riferimento la media tra le aree) che a livello di comune (inteso come aggregazione delle relative aree di censimento sottoposte a campionamento e quindi considerate nella sperimentazione). Le tavole riportano, con riferimento ai comuni di *Milano*, *Bologna* e *Trapani*, i risultati per i disegni ritenuti migliori nelle due strategie proposte: CCSFAM per l'estrazione di campioni di famiglie direttamente dalla lista anagrafica; STRSPOP per la formazione di campioni areali di sezioni di censimento (con stratificazione delle unità finali secondo lo schema descritto nel paragrafo 7.2). Per entrambi i disegni è stata presa in considerazione la frazione sondata più ampia (33%).

In particolare, in corrispondenza della frequenza assoluta T (dati osservati al Censimento 2001) riferita ad una generica variabile rilevabile a campione tramite *long form* al prossimo censimento, è stato definito il livello di errore campionario atteso (in termini di *cv*) e l'errore assoluto atteso ($\Delta_T = 1{,}96 \cdot T \cdot cv/100$) in base al quale è possibile determinare i limiti dell'intervallo di confidenza[28] $\{(\hat{T} - \Delta_T);(\hat{T} + \Delta_T)\}$, che conterrà il vero valore di T con probabilità pari a 0,95. Si osserva che, nei casi di frequenze percentuali p piccole, pur in presenza di un *cv* molto alto, l'intervallo di confidenza determinato per la stima della corrispondente frequenza assoluta, risulta abbastanza ristretto a vantaggio della precisione della stima.

[28] Si precisa che negli esempi riportati si è assunta l'ipotesi della normalità della distribuzione delle stime campionarie. Alcune considerazioni di merito sono esposte nell'Appendice B.

Tavola 13.1 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite alle aree di censimento (media sulle aree) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata pari al 33%). Comune di Milano (Censimento 2001)

MILANO (111 aree)			CCSFAM				STRSPOP			
VARIABILI LONG FORM	Totale medio sulle aree	Livello medio di p sulle aree	cv	Δ	estremo inferiore	estremo superiore	cv	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	54	0,5	19,87	22	32	76	22,18	24	30	78
Pop 15+ Occupata in Industria	1.092	9,8	4,04	87	1.005	1.179	4,45	96	996	1.188
Pop 15+ Occupata in Altra attività	3.696	33,1	1,65	120	3.576	3.816	2,09	152	3.544	3.848
Pop che si sposta giornalm. nel comune di DA	4.872	43,6	1,37	132	4.740	5.004	1,76	169	4.703	5.041
Pop che si sposta giornalm. fuori dal comune di DA	789	7,1	5,00	78	711	867	5,56	86	703	875
Lavoratori in proprio	603	5,4	5,84	70	533	673	6,52	78	525	681
Lavoratori dipendenti	3.499	31,3	1,70	117	3.382	3.616	2,19	150	3.349	3.649
Pop 15+ Occupata	4.842	43,4	1,10	105	4.737	4.947	1,48	141	4.701	4.983
Pop 15+ Disoccupata in cerca di nuova occupazione	226	2,0	9,95	45	181	271	12,62	56	170	282
Pop 15+ Disoccupata in cerca di prima occupazione	58	0,5	19,60	23	35	81	19,76	23	35	81
Pop. media res. in fam. nelle aree di centro per LF	*11.165*									

Tavola 13.2 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite al comune (aggregazione di aree campionabili) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata pari al 33%). Comune di Milano (Censimento 2001)

MILANO			CCSFAM				STRSPOP			
VARIABILI LONG FORM	Totale	*p*	*cv*	Δ	estremo inferiore	estremo superiore	*cv*	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	6.007	*0,5*	1,89	223	5.784	6.230	2,21	261	5.746	6.268
Pop 15+ Occupata in Industria	121.223	*9,8*	0,39	927	120.296	122.150	0,45	1070	120.153	122.293
Pop 15+ Occupata in Altra attività	410.223	*33,1*	0,16	1287	408.936	411.510	0,23	1850	408.373	412.073
Pop che si sposta giornalm. nel comune di DA	540.765	*43,6*	0,13	1378	539.387	542.143	0,19	2014	538.751	542.779
Pop che si sposta giornalm. fuori dal comune di DA	87.556	*7,1*	0,48	824	86.732	88.380	0,55	944	86.612	88.500
Lavoratori in proprio	66.902	*5,4*	0,56	735	66.167	67.637	0,63	827	66.075	67.729
Lavoratori dipendenti	388.358	*31,3*	0,17	1295	387.063	389.653	0,22	1675	386.683	390.033
Pop 15+ Occupata	537.453	*43,4*	0,10	1054	536.399	538.507	0,16	1686	535.767	539.139
Pop 15+ Disoccupata in cerca di nuova occupazione	25.097	*2,0*	0,94	463	24.634	25.560	1,36	669	24.428	25.766
Pop 15+ Disoccupata in cerca di prima occupazione	6.418	*0,5*	1,84	232	6.186	6.650	2,08	262	6.156	6.680
Pop. Residente in fam. nelle aree di centro per LF	*1.239.346*									

Tavola 13.3 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite alle aree di censimento (media sulle aree) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata pari al 33%). Comune di Bologna (Censimento 2001)

BOLOGNA (32 aree)			CCSFAM				STRSPOP			
VARIABILI LONG FORM	Totale medio sulle aree	Livello medio di p sulle aree	cv	Δ	estremo inferiore	estremo superiore	cv	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	43	0,4	23,32	20	23	63	22,02	19	24	62
Pop 15+ Occupata in Industria	1.127	10,2	4,12	91	1.036	1.218	4,44	99	1.028	1.226
Pop 15+ Occupata in Altra attività	3.589	32,6	1,67	118	3.471	3.707	1,74	123	3.466	3.712
Pop che si sposta giornalm. nel comune di DA	4.373	39,7	1,47	127	4.246	4.500	1,57	135	4.238	4.508
Pop che si sposta giornalm. fuori dal comune di DA	1.038	9,4	4,32	88	950	1.126	4,34	89	949	1.127
Lavoratori in proprio	693	6,3	5,39	74	619	767	5,93	81	612	774
Lavoratori dipendenti	3.452	31,3	1,65	112	3.340	3.564	1,99	135	3.317	3.587
Pop 15+ Occupata	4.759	43,2	1,04	98	4.661	4.857	1,83	171	4.588	4.930
Pop 15+ Disoccupata in cerca di nuova occupazione	179	1,6	11,46	41	138	220	11,98	43	136	222
Pop 15+ Disoccupata in cerca di prima occupazione	38	0,3	24,39	19	19	57	23,73	18	20	56
Pop. media res. in fam. nelle aree di centro per LF	*11.014*									

Tavola 13.4 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite al comune (aggregazione di aree campionabili) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata pari al 33%). Comune di Bologna (Censimento 2001)

BOLOGNA			CCSFAM				STRSPOP			
VARIABILI LONG FORM	Totale	*p*	*cv*	Δ	estremo inferiore	estremo superiore	*cv*	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	1.385	*0,4*	3,92	107	1.278	1.492	4,22	115	1.270	1.500
Pop 15+ Occupata in Industria	36.048	*10,2*	0,71	503	35.545	36.551	0,77	542	35.506	36.590
Pop 15+ Occupata in Altra attività	114.861	*32,6*	0,28	640	114.221	115.501	0,36	809	114.052	115.670
Pop che si sposta giornalm. nel comune di DA	139.951	*39,7*	0,26	708	139.243	140.659	0,31	859	139.092	140.810
Pop che si sposta giornalm. fuori dal comune di DA	33.221	*9,4*	0,74	485	32.736	33.706	0,81	525	32.696	33.746
Lavoratori in proprio	22.184	*6,3*	0,98	427	21.757	22.611	1,10	480	21.704	22.664
Lavoratori dipendenti	110.451	*31,3*	0,30	646	109.805	111.097	0,38	827	109.624	111.278
Pop 15+ Occupata	152.294	*43,2*	0,19	565	151.729	152.859	0,24	714	151.580	153.008
Pop 15+ Disoccupata in cerca di nuova occupazione	5.737	*1,6*	2,01	227	5.510	5.964	2,42	272	5.465	6.009
Pop 15+ Disoccupata in cerca di prima occupazione	1.212	*0,3*	4,10	98	1.114	1.310	4,22	101	1.111	1.313
Pop. Residente in fam. nelle aree di centro per LF	*352.461*									

Tavola 13.5 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite alle aree di censimento (media sulle aree) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata peri al 33%). Comune di Trapani (Censimento 2001)

TRAPANI *(5 aree)*	Totale medio sulle aree	*Livello medio di p sulle aree*	CCSFAM				STRSPOP			
VARIABILI LONG FORM			*cv*	Δ	estremo inferiore	estremo superiore	*cv*	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	154	*1,3*	12,98	40	114	194	13,79	42	112	196
Pop 15+ Occupata in Industria	623	*5,3*	5,48	67	556	690	6,66	82	541	705
Pop 15+ Occupata in Altra attività	2.550	*21,9*	2,17	109	2.441	2.659	5,01	251	2.299	2.801
Pop che si sposta giornalm. nel comune di DA	3.982	*34,2*	1,68	132	3.850	4.114	2,97	232	3.750	4.214
Pop che si sposta giornalm. fuori dal comune di DA	902	*7,7*	4,89	87	815	989	6,58	117	785	1.019
Lavoratori in proprio	452	*3,9*	6,40	57	395	509	7,55	67	385	519
Lavoratori dipendenti	2.596	*22,3*	2,10	107	2.489	2.703	3,30	168	2.428	2.764
Pop 15+ Occupata	3.326	*28,5*	1,65	108	3.218	3.434	3,13	205	3.121	3.531
Pop 15+ Disoccupata in cerca di nuova occupazione	545	*4,7*	6,43	69	476	614	8,39	90	455	635
Pop 15+ Disoccupata in cerca di prima occupazione	394	*3,4*	7,40	58	336	452	11,02	86	308	480
Pop. media res. in fam. nelle aree di centro per LF	*11.655*									

Tavola 13.6 - Coefficienti di variazione (cv mediani), errori assoluti (Δ) e estremi degli intervalli di confidenza (al 95%) di stime riferite al comune (aggregazione di aree campionabili) di alcune variabili long form, per i disegni CCSFAM e STRSPOP (frazione sondata pari al 33%). Comune di Trapani (Censimento 2001)

TRAPANI	Totale	*p*	CCSFAM				STRSPOP			
VARIABILI LONG FORM			cv	Δ	estremo inferiore	estremo superiore	*cv*	Δ	estremo inferiore	estremo superiore
Pop 15+ Occupata in Agricoltura	769	*1,3*	5,38	82	687	851	6,81	103	666	872
Pop 15+ Occupata in Industria	3.115	*5,3*	2,68	164	2.951	3.279	2,94	180	2.935	3.295
Pop 15+ Occupata in Altra attività	12.748	*21,9*	1,08	270	12.478	13.018	2,15	538	12.210	13.286
Pop che si sposta giornalm. nel comune di DA	19.908	*34,2*	0,79	309	19.599	20.217	1,52	595	19.313	20.503
Pop che si sposta giornalm. fuori dal comune di DA	4.510	*7,7*	2,17	192	4.318	4.702	3,00	266	4.244	4.776
Lavoratori in proprio	2.262	*3,9*	2,99	133	2.129	2.395	3,38	150	2.112	2.412
Lavoratori dipendenti	12.982	*22,3*	0,98	249	12.733	13.231	1,55	395	12.587	13.377
Pop 15+ Occupata	16.632	*28,5*	0,77	250	16.382	16.882	1,42	464	16.168	17.096
Pop 15+ Disoccupata in cerca di nuova occupazione	2.727	*4,7*	2,92	157	2.570	2.884	3,96	212	2.515	2.939
Pop 15+ Disoccupata in cerca di prima occupazione	1.970	*3,4*	3,48	135	1.835	2.105	6,10	236	1.734	2.206
Pop. Residente in fam. nelle aree di centro per LF	*58.275*									

14. Considerazioni conclusive

La possibilità di introdurre nel censimento della popolazione tecniche di campionamento per l'osservazione di alcune delle variabili socio-economiche tramite l'impiego simultaneo di questionari in versione *short* e *long*, ha spinto l'Istat ad avviare uno studio al fine di proporre strategie campionarie che consentano di produrre informazione statistica riferita a domini territoriali minimi con livelli di accuratezza accettabili.

Sono state suggerite due differenti strategie: una presuppone l'utilizzo dei registri anagrafici comunali per la formazione di campioni di famiglie; l'altra si riferisce all'impiego delle basi territoriali, messe a disposizione dall'Istat ai comuni, come lista di riferimento per l'estrazione di campioni di sezioni. Riguardo l'estrazione delle famiglie dai registri anagrafici, in fase di sperimentazione, si è utilizzata, al posto della lista anagrafica, la lista censuaria del 2001; per quanto attiene invece all'estrazione delle sezioni, la fase sperimentale ha preso a riferimento le basi territoriali comunali disegnate per il censimento del 2001.

Per ciascuna strategia sono stati proposti tre diversi disegni di campionamento, per i quali sono state realizzate specifiche simulazioni campionarie basate su metodi di tipo Monte Carlo. Le sperimentazioni, condotte complessivamente per 40 comuni scelti in modo opportuno tra le varie aree geografiche del paese e con diverso peso demografico, hanno riguardato la valutazione dell'efficienza di stime relative a variabili che potrebbero essere oggetto della rilevazione campionaria tramite *long form*: titolo di studio, condizione lavorativa, posizione professionale, settore di attività economica, spostamenti giornalieri. Va inoltre aggiunto che i disegni di campionamento sono stati pianificati per produrre stime riferite ai domini sub-comunali delle aree di censimento di centro abitato, aventi le caratteristiche per essere ritenute campionabili.

I risultati delle sperimentazioni relativi ai diversi disegni sono stati confrontati in termini di coefficiente di variazione, che fornisce una misura dell'errore che mediamente si commette con le stime campionarie.

Le conclusioni dell'analisi svolta sull'efficienza delle stime sembrano essere favorevoli all'uso dei metodi campionari; a tal riguardo si evidenzia che la strategia di campionamento da lista delle famiglie è da preferire rispetto a quella di campionamento areale relativa alle sezioni di censimento. Tuttavia, i risultati forniti dalla sperimentazione mostrano che il campione areale rimane una valida alternativa, in quanto permette di fornire stime con un ridotto aumento dell'errore e quindi qualitativamente soddisfacenti.

Riguardo la possibilità di ridurre la variabilità delle stime con la stratificazione delle famiglie (per numero di componenti o per età del capofamiglia) il campionamento da lista non sembra ricevere vantaggi consistenti. Viceversa, per il campionamento areale si osserva un piccolo guadagno con l'introduzione della stratificazione delle sezioni, in particolare, dall'adozione dello schema che prevede la stratificazione delle sezioni in tre gruppi aventi un numero simile di individui (disegno STRSPOP).

Per quanto concerne la frazione di campionamento, i risultati fanno prevalere l'adozione di quello più ampio del 33% in quanto in grado di soddisfare maggiormente il requisito dell'accuratezza delle stime rispetto alle frazioni di campionamento più basse, specialmente quando le stime sono riferite a domini di più piccole dimensioni. La scelta finale dovrà comunque tener conto, oltre che di aspetti statistici, anche di valutazioni di costo finanziario.

Dal confronto tra i risultati per tipologia di aree di censimento, con riferimento alla dimensione di popolazione, si osserva che il guadagno di accuratezza delle stime sulle aree più grandi è così evidente da suggerire fortemente il disegno di aree intorno alle 15mila unità. Questo risultato potrebbe fornire indicazioni utili sia sulla metodologia per la costruzione delle aree di censimento che sulla scelta della soglia dimensionale, riferita alle aree stesse, sotto la quale decidere di non selezionare campioni, ma di procedere ad una rilevazione esaustiva tramite *long form*.

Un ulteriore risultato riguarda il guadagno di efficienza delle stime che si osserva nel passaggio dal dominio più piccolo dell'area di censimento a domini più estesi riferiti ad ambiti comunali, provinciali, regionali, eccetera. Fissata la strategia campionaria e il livello della frequenza percentuale da stimare, la riduzione dell'errore di campionamento atteso è più ampia sui domini più

grandi (con un elevato numero di aree campionabili) per la maggiore dimensione del campione complessivamente osservato; si riscontra un ulteriore guadagno di efficienza nel caso in cui una parte del contesto territoriale di riferimento è sottoposta a rilevazione tramite *long form* in modo esaustivo.

Riguardo, infine, la scelta dello stimatore, nelle sperimentazioni è stato impiegato lo stimatore di ponderazione vincolata che favorisce una maggiore rappresentatività dell'esito campionario. Si potrebbero comunque, adottare anche tecniche di stima alternative basate, per esempio, sui *metodi di stima per piccole aree*, nel caso di stime riferite a piccoli domini territoriali o a particolari sottogruppi della popolazione (Borrelli *et al.*, 2008).

Alla luce delle analisi comparative sull'efficienza delle stime illustrate in questo lavoro, si ritiene che le strategie di campionamento da lista e campionamento areale potrebbero anche coesistere in una proposta complessiva di rilevazione tramite questionari in formato *short* e *long* al prossimo censimento. Comunque, la scelta della strategia dovrà rispondere alla specificità del contesto censuario che richiede semplicità (obiettivo rilevante nelle indagini su larga scala dove si ha a che fare con grandi moli di dati) per la pluralità degli obiettivi di indagine (che si manifesta attraverso la stima dei totali riferiti a più variabili), e dovrà tenere in considerazione anche l'eventuale disponibilità di informazioni ausiliarie provenienti da fonti amministrative.

Una volta definita la strategia campionaria e i relativi livelli di accuratezza delle stime che è possibile garantire con l'impianto metodologico scelto, il successivo passo sarà quello di valutare attentamente il grado di compatibilità tra il piano di diffusione dei risultati censuari e l'errore di campionamento atteso delle stime prodotte. A riguardo, tenuto conto del fatto che al crescere del contesto territoriale di riferimento aumenta l'insieme di informazioni censuarie da rendere disponibili, è stato di recente avviato uno studio (Carbonetti *et al.*, 2008a ; Borrelli *et al.*, 2009 ; Carbonetti, 2009) volto a valutare il possibile impatto della strategia campionaria sull'accuratezza di tavole di diffusione dei risultati censuari per differenti livelli territoriali e dettagli informativi.

APPENDICE A – Studio sulla robustezza dei risultati delle simulazioni

A.1. Premessa

L'affidabilità delle stime campionarie è stata valutata in modo empirico attraverso misure (stime) del coefficiente di variazione facendo riferimento ad uno spazio campionario simulato tramite la replicazione di 1.000 campioni, generati secondo il prefissato disegno di campionamento (i dettagli dell'algoritmo sono descritti nel paragrafo 7.6).

Poiché la scelta del numero di replicazioni pari a 1.000 è stata principalmente indotta da considerazioni di tipo computazionale, è sorta la necessità di valutare la *robustezza* dei risultati ottenuti sia incrementando il numero (analisi per dimensione) che differenziando l'insieme (analisi per struttura) dei campioni impiegati per le valutazioni.

A.2. Metodologia

Per valutare la robustezza delle misure dell'errore campionario effettuate con riferimento ad un universo di 1.000 campioni, è stato deciso di aumentare il numero delle replicazioni campionarie fino a 20.000, in modo da accertare se ad un ampliamento dello spazio campionario corrispondesse una variazione significativa del valore del coefficiente di variazione.

Con riferimento allo spazio campionario simulato esteso a 20.000 replicazioni, sono state calcolate le stime secondo due diversi disegni di campionamento: il disegno casuale semplice di famiglie (CCSFAM) ed il disegno areale stratificato di sezioni (STRSPOP), entrambi con la frazione di campionamento pari al 33%. Le valutazioni hanno previsto l'impiego dei dati del Censimento della popolazione del 2001 relativi al comune di Aosta: per ognuno dei 20.000 campioni sono state calcolate le stime delle frequenze percentuali p e assolute T relative alle modalità di incrocio delle variabili descritte nel paragrafo 7.3, con riferimento al dominio dell'area di censimento di centro abitato. Delle tre aree del comune di Aosta ne sono state prese in esame solo due: la più piccola (circa 8.900 ab.) e la più grande (circa 12.100 ab.).

Le valutazioni critiche di questo studio si muovono però solo con riferimento a quattro modalità di incrocio, scelte per differenti valori della frequenza percentuale osservata al 2001:

1) *ateco35m_14* → Popolazione residente in famiglia di 15 anni e più occupata per sezione di attività economica - Altre attività Totale (p%=29,91% sull'area più piccola; p%=33,07% sull'area più grande);
2) *pos_set_sesso_8* → Popolazione residente in famiglia – Maschi – Lavoratori dipendenti in Industria (p%=5,94% sull'area più piccola; p%=5,62% sull'area più grande);
3) *ateco35m_1* → Popolazione residente in famiglia di 15 anni e più occupata per sezione di attività economica A, B – Agricoltura Totale (p%=0,45% sull'area più piccola; p%=0,30% sull'area più grande);
4) *posprof_setteco_2* → Popolazione residente in famiglia – Lavoratori in proprio in Agricoltura (p%=0,04% sull'area più piccola; p%=0,06% sull'area più grande).

Per ogni modalità di incrocio è stato dunque calcolato il coefficiente di variazione con riferimento ad uno spazio campionario diverso in composizione e/o in ampiezza da quello considerato per lo studio esposto nel documento, così da valutare le eventuali variazioni di *cv* riconducibili ad un differente insieme di campioni simulati.

A.3. Analisi della robustezza per dimensione

In questa prima analisi si è cercato di valutare l'effetto del numero di replicazioni campionarie sulla determinazione dello spazio campionario simulato e sul relativo calcolo del coefficiente di variazione. A tale scopo, sono stati generati 20 blocchi distinti tra loro e composti da 1.000 campioni casuali. Successivamente sono stati misurati i livelli di *cv* delle stime sugli spazi campionari che venivano a configurarsi per successive aggregazioni dei blocchi: 20 misure relative agli spazi campionari determinati per incrementi di 1.000 replicazioni distinte alla volta, fino allo spazio campionario più esteso comprensivo di tutti i 20.000 campioni generati. Tale operazione ha permesso di analizzare il comportamento del *cv* anche al crescere della dimensione dello spazio campionario, per un ampliamento graduale dell'universo dei campioni. I risultati delle simulazioni sono presentati nelle Tavole A.1-A.4.

Nella Tavola A.1 sono riportati, con riferimento all'area più piccola, i *cv* ottenuti all'incrementare dello spazio campionario nel caso del disegno CCSFAM (f.s.=33%). Dall'analisi dei valori è possibile notare che per stime di frequenze assolute molto piccole (*posprof_setteco_8*) la differenza di errore tra 1.000 e 20.000 campioni è pari a 0,8 infatti, il *cv* passa da 73,69 a 72,90 , ma in realtà ciò non porta ad un reale miglioramento dei risultati trattandosi di ammontari molto piccoli. Per valori di *T* più grandi invece, all'aumentare delle replicazioni campionarie, (*ateco35m_14, pos_set_sesso_8, ateco35m_1*) il *cv* sembrerebbe attestarsi su valori più alti; in realtà, essendo il differenziale molto esiguo emerge una sostanziale robustezza dei risultati ottenuti.

Ad analoghe considerazioni si giunge esaminando le Tavole A.2-A.4 riferite rispettivamente al disegno CCSFAM per l'area più grande e al disegno STRSOPOP per la più piccola e la più grande area del comune di Aosta.

Tavola A.1 – Coefficiente di variazione cv delle stime delle frequenze assolute T di quattro incroci rilevabili tramite long form e riferiti alla più piccola area del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33%

AOSTA Area piccola	Disegno **CCSFAM** f.s.=33%							
Numero di replicazioni campionarie	ateco35m_14		pos_set_sesso_8		ateco35m_1		posprof_setteco_2	
	stima_T	cv	stima_T	cv	stima_T	cv	stima_T	cv
1.000	2678	2,03	532	5,49	40	22,53	4	73,69
2.000	2680	2,01	532	5,56	40	22,84	4	74,23
3.000	2679	2,01	532	5,66	40	22,72	4	73,27
4.000	2681	2,02	533	5,68	40	22,95	4	73,47
5.000	2681	2,02	533	5,70	40	23,06	4	73,24
6.000	2680	2,01	532	5,71	40	23,08	4	72,92
7.000	2681	2,02	532	5,74	40	23,10	4	72,69
8.000	2681	2,14	532	5,75	40	23,18	4	72,67
9.000	2681	2,14	532	5,75	40	23,24	4	72,78
10.000	2681	2,13	532	5,75	40	23,24	4	72,97
11.000	2680	2,12	532	5,75	40	23,19	4	72,96
12.000	2680	2,11	532	5,75	40	23,24	4	72,92
13.000	2680	2,10	532	5,75	40	23,22	4	73,01
14.000	2680	2,10	532	5,75	40	23,26	4	73,08
15.000	2681	2,09	532	5,76	40	23,25	4	72,92
16.000	2680	2,09	532	5,77	40	23,17	4	72,72
17.000	2680	2,08	532	5,77	40	23,14	4	72,80
18.000	2680	2,08	532	5,77	40	23,17	4	72,87
19.000	2680	2,07	532	5,77	40	23,19	4	72,91
20.000	2681	2,07	532	5,76	40	23,21	4	72,90

Tavola A.2 – Coefficiente di variazione cv delle stime delle frequenze assolute T di quattro incroci rilevabili tramite long form e riferiti alla più grande area del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33%

AOSTA Area grande	Disegno **CCSFAM** f.s.=33%							
Numero di replicazioni campionarie	ateco35m_14		pos_set_sesso_8		ateco35m_1		posprof_setteco_2	
	stima_T	cv	stima_T	cv	stima_T	cv	stima_T	cv
1.000	4015	1,50	683	5,12	37	23,21	7	54,78
2.000	4016	1,54	683	5,11	37	22,96	7	53,86
3.000	4015	1,58	683	5,12	37	23,47	7	53,59
4.000	4015	1,57	683	5,15	37	23,29	7	53,13
5.000	4016	1,56	683	5,15	37	23,33	7	53,23
6.000	4016	1,56	683	5,18	37	23,32	7	53,13
7.000	4016	1,57	683	5,16	37	23,30	7	53,22
8.000	4016	1,75	683	5,23	37	23,23	7	53,15
9.000	4016	1,73	683	5,21	37	23,29	7	53,16
10.000	4016	1,71	683	5,22	37	23,28	7	53,15
11.000	4016	1,70	683	5,20	37	23,28	7	53,20
12.000	4016	1,69	683	5,19	37	23,36	7	53,37
13.000	4017	1,69	683	5,22	37	23,36	7	53,62
14.000	4017	1,68	683	5,21	37	23,34	7	53,54
15.000	4017	1,67	683	5,21	37	23,33	7	53,47
16.000	4016	1,66	683	5,21	37	23,33	7	53,51
17.000	4017	1,66	683	5,21	37	23,30	7	53,53
18.000	4016	1,66	683	5,20	37	23,36	7	53,54
19.000	4016	1,65	683	5,20	37	23,36	7	53,65
20.000	4016	1,65	683	5,19	37	23,35	7	53,62

Tavola A.3 – Coefficiente di variazione cv delle stime delle frequenze assolute T di quattro incroci rilevabili tramite long form e riferiti alla più piccola area del comune di Aosta. Disegno STRSPOP con frazione di campionamento del 33%

AOSTA Area piccola	Disegno STRSPOP f.s.=33%							
Numero di Replicazioni campionarie	ateco35m_14		pos_set_sesso_8		ateco35m_1		posprof_setteco_2	
	stima_T	cv	stima_T	cv	stima_T	cv	stima_T	cv
1.000	2679	3,63	533	7,55	40	20,82	4	69,13
2.000	2678	3,61	534	7,43	40	20,72	4	67,23
3.000	2679	3,63	533	7,57	40	20,65	4	67,93
4.000	2681	3,65	533	7,64	40	20,61	4	68,25
5.000	2681	3,65	532	7,64	40	20,59	4	68,43
6.000	2681	3,65	532	7,63	40	20,57	4	68,02
7.000	2680	3,66	532	7,66	40	20,51	4	67,85
8.000	2681	3,66	532	7,62	40	20,46	4	67,80
9.000	2681	3,66	532	7,62	40	20,49	4	67,73
10.000	2681	3,66	532	7,61	40	20,52	4	68,08
11.000	2680	3,65	532	7,59	40	20,52	4	68,28
12.000	2681	3,65	532	7,58	40	20,48	4	68,31
13.000	2681	3,64	532	7,57	40	20,50	4	68,37
14.000	2680	3,64	532	7,57	40	20,57	4	68,38
15.000	2681	3,64	532	7,55	40	20,55	4	68,40
16.000	2681	3,64	532	7,55	40	20,54	4	68,41
17.000	2680	3,64	532	7,56	40	20,54	4	68,28
18.000	2680	3,64	532	7,56	40	20,51	4	68,06
19.000	2680	3,64	532	7,56	40	20,48	4	68,22
20.000	2680	3,64	533	7,54	40	20,41	4	68,10

Tavola A.4 – Coefficiente di variazione cv delle stime delle frequenze assolute T di quattro incroci rilevabili tramite long form e riferiti alla più grande area del comune di Aosta. Disegno STRSPOP con frazione di campionamento del 33%

AOSTA Area grande	Disegno STRSPOP f.s.=33%							
Numero di replicazioni campionarie	ateco35m_14		pos_set_sesso_8		ateco35m_1		posprof_setteco_2	
	stima_T	cv	stima_T	cv	stima_T	cv	stima_T	cv
1.000	4012	2,17	684	6,45	37	21,11	7	47,33
2.000	4013	2,18	684	6,42	37	21,10	7	47,18
3.000	4014	2,18	683	6,48	37	20,94	7	47,26
4.000	4015	2,20	683	6,50	37	20,90	7	47,50
5.000	4015	2,18	683	6,48	37	20,78	7	47,54
6.000	4015	2,18	683	6,46	37	20,88	7	47,63
7.000	4015	2,18	683	6,46	37	20,75	7	47,39
8.000	4016	2,17	683	6,46	37	20,75	7	47,49
9.000	4015	2,17	683	6,43	37	20,72	7	47,51
10.000	4015	2,17	683	6,44	37	20,70	7	47,45
11.000	4015	2,17	683	6,43	37	20,70	7	47,71
12.000	4015	2,18	684	6,45	37	20,69	7	47,69
13.000	4015	2,18	684	6,45	37	20,66	7	47,85
14.000	4015	2,17	684	6,43	37	20,66	7	47,86
15.000	4015	2,17	684	6,42	37	20,64	7	47,86
16.000	4015	2,17	684	6,43	37	20,64	7	47,80
17.000	4015	2,17	684	6,42	37	20,66	7	47,77
18.000	4015	2,17	684	6,43	37	20,70	7	47,78
19.000	4015	2,17	684	6,43	37	20,70	7	47,69
20.000	4015	2,17	684	6,43	37	20,71	7	47,72

A.4. Analisi della robustezza per struttura

In questa seconda analisi si è studiata la robustezza dei risultati al variare della composizione dell'insieme delle 1.000 replicazioni campionarie. Per tale obiettivo è stato utilizzato il "serbatoio" dei 20.000 campioni da cui sono stati estratti casualmente 100 blocchi di 1.000 campioni, per ciascuno dei quali è stato determinato il *cv* relativo al corrispondente spazio campionario. In seguito, sull'insieme dei 100 blocchi, è stata calcolata una misura di dispersione Δ data dalla differenza tra il *cv* massimo e il *cv* minimo (valori osservati sui 100 livelli di *cv* calcolati): minore sarà il valore osservato della misura Δ e maggiore sarà la robustezza attesa dei risultati ottenuti.

Le Tavole A.5-A.8 riportano i valori di Δ per le quattro modalità di incrocio presi in esame, per le due aree di Aosta scelte e per i due disegni di campionamento fissati.

Dall'esame delle tabelle emerge che, qualunque sia il disegno adottato e l'area presa in considerazione, per tre delle modalità di incrocio esaminate (*ateco35m_1, ateco35m_14, pos_set_sesso_8*) il valore di Δ è al più pari a 2,62 , ciò significa che l'errore che si commette rispetto al "vero valore di *cv*" (inteso come valore asintotico del *cv* quello osservato sullo spazio campionario simulato tramite i 20.000 campioni), è inferiore alla soglia definita dalla quantità Δ. Valori di Δ più elevati si registrano solo nel caso della modalità *posprof_setteco_2* che si riferisce a valori di p (e quindi di T) molto piccoli. Questa situazione però non preoccupa in quanto, pur in presenza di errori percentuali più grandi, gli errori assoluti risultano sempre molto ridotti.

Tavola A.5 – Coefficiente di variazione minimo e massimo relativi a 100 blocchi di estrazioni di 1.000 campioni, per la stima delle frequenze percentuali p e assolute T di quattro incroci rilevabili tramite long form e riferiti alla più piccola area del comune di Aosta. Disegno CCSFAM (f.s.= 33%)

AOSTA Area piccola	Disegno **CCSFAM** f.s.=33%					
Modalità di incrocio	Stima_T Media delle stime relative ai 20.000 campioni	Stima_p Media delle stime relative ai 20.000 campioni	*cv* calcolato sullo spazio campionario indotto dai 20.000 campioni	*cv* min sulle 100 estrazioni dei blocchi di 1.000 campioni	*cv* max sulle 100 estrazioni dei blocchi di 1.000 campioni	Δ *cv*max-*cv*min
POSPROF_SETTECO_2	3,99	0,04	72,90	68,79	78,90	10,11
ATECO35M_1	39,91	0,45	23,21	21,82	24,41	2,59
ATECO35M_14	2680,53	29,91	2,07	1,91	2,89	0,98
POS_SET_SESSO_8	532,22	5,94	5,76	5,39	6,17	0,79

Tavola A.6 – Coefficiente di variazione minimo e massimo relativi a 100 blocchi di estrazioni di 1.000 campioni, per la stima delle frequenze percentuali p e assolute T di quattro incroci rilevabili tramite long form e riferiti alla più grande area del comune di Aosta. Disegno CCSFAM (f.s.= 33%)

AOSTA Area grande	Disegno **CCSFAM** f.s.=33%					
Modalità di incrocio	Stima_T Media delle stime relative ai 20.000 campioni	Stima_p Media delle stime relative ai 20.000 campioni	*cv* calcolato sullo spazio campionario indotto dai 20.000 campioni	*cv* min sulle 100 estrazioni dei blocchi di 1.000 campioni	*cv* max sulle 100 estrazioni dei blocchi di 1.000 campioni	Δ *cv*max-*cv*min
POSPROF_SETTECO_2	6,99	0,06	53,62	49,81	57,13	7,31
ATECO35M_1	37,07	0,31	23,35	22,11	24,73	2,62
ATECO35M_14	4016,19	33,06	1,65	1,48	2,73	1,25
POS_SET_SESSO_8	682,91	5,62	5,19	4,89	5,80	0,91

Tavola A.7 – Coefficiente di variazione minimo e massimo relativi a 100 blocchi di estrazioni di 1.000 campioni, per la stima delle frequenze percentuali p e assolute T di quattro incroci rilevabili tramite long form e riferiti alla più piccola area del comune di Aosta. Disegno STRSPOP (f.s.= 33%)

AOSTA Area piccola	Disegno STRSPOP f.s.=33%					
Modalità di incrocio	Stima_T Media delle stime relative ai 20.000 campioni	Stima_p Media delle stime relative ai 20.000 campioni	*cv* calcolato sullo spazio campionario indotto dai 20.000 campioni	*cv* min sulle 100 estrazioni dei blocchi di 1.000 campioni	*cv* max sulle 100 estrazioni dei blocchi di 1.000 campioni	Δ *cv*max-*cv*min
POSPROF_SETTECO_2	3,98	0,04	68,10	64,69	72,03	7,35
ATECO35M_1	40,26	0,45	20,41	19,38	21,64	2,25
POS_SET_SESSO_8	532,53	5,94	7,54	7,11	7,91	0,79
ATECO35M_14	2680,05	29,91	3,64	3,45	3,87	0,42

Tavola A.8 – Coefficiente di variazione minimo e massimo relativi a 100 blocchi di estrazioni di 1.000 campioni, per la stima delle frequenze percentuali p e assolute T di quattro incroci rilevabili tramite long form e riferiti alla più grande area del comune di Aosta. Disegno STRSPOP (f.s.= 33%)

AOSTA Area grande	Disegno STRSPOP f.s.=33%					
Modalità di incrocio	Stima_T Media delle stime relative ai 20.000 campioni	Stima_p Media delle stime relative ai 20.000 campioni	*cv* calcolato sullo spazio campionario indotto dai 20.000 campioni	*cv* min sulle 100 estrazioni dei blocchi di 1.000 campioni	*cv* max sulle 100 estrazioni dei blocchi di 1.000 campioni	Δ *cv*max-*cv*min
POSPROF_SETTECO_2	7,01	0,06	47,72	44,64	50,06	5,42
ATECO35M_1	37,16	0,31	20,71	19,46	21,57	2,11
POS_SET_SESSO_8	683,75	5,63	6,43	6,08	6,79	0,70
ATECO35M_14	4015,00	33,05	2,17	2,08	2,29	0,22

A.5. Conclusioni

Dall'analisi della robustezza per dimensione, condotta tramite l'estrazione di 20.000 campioni, è emerso che per entrambi i disegni studiati (CCSFAM e STRASPOP) il coefficiente di variazione, per le quattro modalità di incrocio prese in esame, non sembra mostrare variazioni significative all'ampliamento dello spazio campionario e questo risultato sembra confermato in entrambe le aree del comune di Aosta (Tavole A.1-A.4) scelte con differente ampiezza. Quindi, si può ritenere appropriato il numero di replicazioni pari a 1.000 per la simulazione degli spazi campionari proposti e, di conseguenza, affidabili i risultati relativi alla determinazione del coefficiente di variazione, indipendentemente dal valore della frequenza da stimare e dalla dimensione del dominio a cui la stima è riferita.

Dall'analisi della robustezza per struttura è invece emerso che la composizione dei campioni simulati può avere una leggera influenza nella determinazione del *cv* solo nel caso in cui questo si riferisca alla stima di frequenze molto piccole. Per questi casi però, essendo l'errore campionario molto elevato, le distanze osservate del valore stimato di *cv* dal valore vero non sono da ritenersi significative a vantaggio dell'affidabilità del risultato per le valutazioni finali.

In conclusione, lo studio permette di assegnare un grado di robustezza molto elevato ai risultati delle simulazioni utilizzati nel documento per lo studio sull'efficienza delle stime campionarie in ambito censuario.

APPENDICE B – Studio sulla distribuzione campionaria delle stime riferite a variabili rilevabili tramite long form

B.1. Premessa

In molte occasioni, oltre alle stime "puntuali" vengono richieste anche stime "intervallari", vale a dire valori stimati tenendo conto di un margine di *errore ammissibile*. A riguardo, si tende a fare delle ipotesi teoriche di base per la costruzione degli intervalli di confidenza; generalmente si ipotizza a priori che le stime campionarie tendono a distribuirsi secondo una distribuzione *Normale*, ipotesi che però non sembra essere sempre confermata.

In questo studio sono state fatte alcune prime valutazioni sull'andamento delle distribuzioni delle stime delle frequenze percentuali riferite alle variabili *long form* prese in esame nel documento (paragrafo 7.3), utilizzando l'impianto di simulazione impiegato a supporto dello studio descritto nell'Appendice A (20.000 replicazioni campionarie; disegno CCSFAM con frazione di campionamento del 33% ; stime riferite all'area più piccola e all'area più grande del comune di Aosta).

B.2. Risultati

Dall'esame grafico delle distribuzioni delle stime emergono differenti comportamenti a seconda del valore della frequenza relativa *p* da stimare. La distribuzione delle stime campionarie (determinata tramite le 20.000 replicazioni) si discosta significativamente dalla distribuzione Normale mostrando forte asimmetria per frequenze relative estreme (*positiva* per valori di *p* bassi; *negativa* per valori di *p* alti); invece, si osserva una tendenza secondo una distribuzione Normale per valori di *p* intermedi.

In particolare, per le distribuzioni campionarie determinate per ciascuna delle 90 modalità di incrocio prese in esame, e con riferimento alle due aree del comune di Aosta, sono stati calcolati gli indici di asimmetria e di curtosi, nonché il test di normalità di Kolmogorov-Sminorv, utili per valutare sia la *forma* che la normalità della distribuzione delle stime.

I risultati sono descritti nelle Tavole B.1-B.2, in cui sono presentati per ciascuna delle frequenze relative *p*, ordinate in modo crescente, i valori degli indici (*skewness* e *kurtosis*), la statistica relativa al test di normalità e il rispettivo *p_value* utile a decidere sull'accettazione dell'ipotesi di normalità (in questo esercizio tale ipotesi è stata ritenuta accettabile nei casi in cui si registra *p_value*>0,05).

Con riferimento all'area più piccola (Tavola B.1) si osserva che il test porta a rifiutare l'ipotesi di normalità generalmente per frequenze relative inferiori al 2,5% e per valori superiori al 10% (salvo qualche raro caso); invece, per livelli di *p* mediamente compresi tra questi due valori la normalità è sufficientemente provata dai risultati del test. A supporto di ciò, l'analisi dei valori dello *skewness* mostra forti asimmetrie proprio per i livelli di *p* per i quali l'ipotesi della normalità è rifiutata.

Si evidenziano inoltre valori molto elevati della curtosi per le frequenze percentuali più grandi che danno ulteriori dimostrazioni di situazioni distanti dalla normalità, per la maggiore concentrazione dei valori sulle code della distribuzione.

Risultati analoghi si osservano con riferimento all'area più grande (Tavola B.2): in tale caso il test di normalità è quasi sempre rifiutato per percentuali inferiori all'1,25% e superiori al 13%, mentre è diffusamente accettato per livelli di *p* compresi tra tali valori.

Tavola B.1 – Risultati dei test di asimmetria, curtosi e normalità sull'area più piccola del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33%

AOSTA (**area piccola**) - Disegno **CCSFAM** ; f.s.=33%

Frequenza relativa *p* oggetto di stima	Indice di asimmetria *(skewness)*	Indice di curtosi *(kurtosi)*	Test di normalità di kolmogoroff (statistica test)	Test di normalità di kolmogoroff *(p_value)*	Accettazione dell'ipotesi di normalità
0,022	0,605	-0,508	0,284	0,010	
0,045	0,456	-0,214	0,113	0,010	
0,078	0,388	-0,315	0,076	0,010	
0,101	0,343	-0,204	0,057	0,010	
0,190	0,199	-0,023	0,017	0,010	
0,223	0,173	-0,093	0,017	0,010	
0,246	0,187	-0,068	0,016	0,010	
0,256	0,172	-0,092	0,018	0,010	
0,268	0,150	-0,076	0,017	0,010	
0,290	0,179	-0,062	0,016	0,010	
0,324	0,174	-0,065	0,016	0,010	
0,325	0,160	-0,060	0,015	0,010	
0,390	0,144	-0,038	0,015	0,010	
0,445	0,142	-0,048	0,013	0,010	
0,446	0,138	-0,014	0,012	0,010	
0,502	0,104	-0,053	0,012	0,010	
0,515	0,137	-0,105	0,015	0,010	
0,557	0,085	-0,098	0,011	0,010	
0,615	0,129	-0,054	0,011	0,010	
0,637	0,107	-0,064	0,012	0,010	
0,706	0,084	0,007	0,011	0,010	
0,794	0,134	0,008	0,014	0,010	
0,860	0,099	0,021	0,010	0,010	
1,103	0,087	0,027	0,011	0,010	
1,127	0,075	0,072	0,010	0,010	
1,165	0,106	0,031	0,010	0,010	
1,213	0,071	-0,015	0,011	0,010	
1,250	0,064	0,036	0,009	0,010	
1,302	0,073	-0,005	0,012	0,010	
1,307	0,086	0,050	0,008	0,010	
1,417	0,089	0,078	0,010	0,010	
1,752	0,053	0,077	0,008	0,010	
1,798	0,072	0,093	0,008	0,010	
1,899	0,048	0,078	0,006	0,042	
2,167	0,058	0,057	0,007	0,016	
2,285	0,050	0,097	0,008	0,010	
2,355	0,057	0,100	0,008	0,010	
2,477	0,028	0,221	0,006	0,052	*
2,678	0,002	0,382	0,006	0,057	*
2,901	0,066	0,107	0,010	0,010	
2,910	0,054	0,193	0,009	0,010	
2,913	0,006	0,081	0,007	0,040	
2,945	0,006	0,380	0,006	0,044	

Tavola B.1 – Risultati dei test di asimmetria, curtosi e normalità sull'area più piccola del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33% (segue)

AOSTA (**area piccola**) - Disegno **CCSFAM** ; f.s.=33% *(segue)*

Frequenza relativa *p* oggetto di stima	Indice di asimmetria *(skewness)*	Indice di curtosi *(kurtosi)*	Test di normalità di kolmogoroff (statistica test)	Test di normalità di kolmogoroff *(p_value)*	Accettazione dell'ipotesi di normalità
3,194	0,010	0,262	0,006	0,104	*
3,401	0,029	0,147	0,008	0,010	
3,436	0,063	0,177	0,011	0,010	
3,448	0,017	0,130	0,006	0,056	*
3,461	0,047	0,123	0,007	0,012	
3,549	-0,014	0,417	0,006	0,060	*
4,028	-0,032	0,366	0,004	0,150	*
4,195	0,015	0,168	0,006	0,122	*
4,307	-0,018	0,343	0,005	0,150	*
4,378	-0,021	0,396	0,005	0,150	*
4,639	-0,008	0,343	0,004	0,150	*
4,808	0,020	0,231	0,007	0,033	
5,601	-0,129	1,727	0,005	0,150	*
5,939	-0,057	0,712	0,006	0,140	*
6,491	-0,017	0,582	0,005	0,150	*
6,503	-0,013	0,409	0,005	0,150	*
6,809	-0,069	0,976	0,004	0,150	*
7,298	-0,068	1,381	0,006	0,052	*
7,766	-0,111	1,419	0,004	0,150	*
8,074	-0,036	1,189	0,008	0,010	
8,781	-0,094	1,860	0,006	0,129	*
8,825	-0,119	1,577	0,006	0,088	*
10,022	-0,184	3,302	0,007	0,020	
10,693	-0,196	3,171	0,007	0,020	
11,390	-0,391	8,244	0,006	0,096	*
12,743	-1,721	55,102	0,008	0,010	
13,842	-0,482	10,621	0,007	0,025	
15,174	-0,481	10,427	0,007	0,020	
15,824	-0,532	12,036	0,007	0,013	
16,184	-0,832	21,303	0,007	0,027	
21,252	-0,447	9,987	0,006	0,075	*
21,708	-1,586	49,495	0,009	0,010	
21,864	-3,022	118,416	0,013	0,010	
23,181	-4,255	186,270	0,016	0,010	
23,547	-0,820	21,662	0,005	0,150	*
27,613	-1,041	29,203	0,007	0,017	
29,913	-1,553	48,492	0,008	0,010	
30,746	-1,903	64,165	0,009	0,010	
32,467	-1,378	40,846	0,007	0,010	
38,982	-2,909	113,053	0,010	0,010	
39,184	-4,904	225,363	0,016	0,010	
42,106	-6,763	347,615	0,020	0,010	
47,122	-8,978	505,568	0,024	0,010	

Tavola B.2 – Risultati dei test di asimmetria, curtosi e normalità sull'area più grande del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33%

AOSTA (**area grande**) - Disegno **CCSFAM** ; f.s.=33%

Frequenza relativa *p* oggetto di stima	Indice di asimmetria *(skewness)*	Indice di curtosi *(kurtosi)*	Test di normalità di kolmogoroff (statistica test)	Test di normalità di kolmogoroff *(p_value)*	Accettazione dell'ipotesi di normalità
0,025	0,443	-0,456	0,183	0,010	
0,033	0,382	-0,369	0,127	0,010	
0,033	0,401	-0,308	0,128	0,010	
0,049	0,326	-0,192	0,103	0,010	
0,050	0,303	-0,267	0,091	0,010	
0,058	0,295	-0,176	0,090	0,010	
0,148	0,161	-0,087	0,027	0,010	
0,156	0,189	-0,038	0,024	0,010	
0,157	0,179	-0,086	0,019	0,010	
0,182	0,150	-0,069	0,023	0,010	
0,206	0,170	-0,046	0,019	0,010	
0,231	0,151	0,006	0,017	0,010	
0,280	0,142	0,010	0,015	0,010	
0,305	0,137	-0,026	0,015	0,010	
0,313	0,146	-0,027	0,014	0,010	
0,338	0,134	-0,019	0,014	0,010	
0,387	0,130	-0,022	0,013	0,010	
0,427	0,129	0,008	0,013	0,010	
0,485	0,126	-0,009	0,012	0,010	
0,545	0,128	-0,035	0,014	0,010	
0,595	0,129	-0,045	0,013	0,010	
0,715	0,086	-0,028	0,011	0,010	
0,790	0,089	-0,039	0,010	0,010	
0,899	0,058	-0,042	0,007	0,019	
1,028	0,043	0,076	0,008	0,010	
1,152	0,054	0,029	0,012	0,010	
1,269	0,037	0,031	0,005	0,150	*
1,283	0,072	-0,106	0,010	0,010	
1,285	0,029	0,024	0,006	0,112	*
1,404	0,032	0,118	0,007	0,019	
1,417	0,035	0,062	0,006	0,064	*
1,646	0,064	0,009	0,008	0,010	
1,772	0,027	0,074	0,006	0,087	*
1,976	0,044	0,069	0,008	0,010	
2,116	0,041	0,185	0,006	0,050	*
2,148	0,001	0,134	0,006	0,048	
2,313	-0,002	0,322	0,009	0,010	
2,428	0,027	0,168	0,007	0,029	
2,703	-0,010	0,287	0,006	0,112	*
2,719	-0,109	1,560	0,007	0,025	
2,741	-0,026	0,521	0,005	0,150	*
2,881	0,041	0,201	0,008	0,010	
3,174	-0,022	0,477	0,005	0,150	*
3,399	0,023	0,332	0,006	0,132	*

Tavola B.2 – Risultati dei test di asimmetria, curtosi e normalità sull'area più grande del comune di Aosta. Disegno CCSFAM con frazione di campionamento del 33% (segue)

AOSTA (**area grande**) - Disegno **CCSFAM** ; f.s.=33% *(segue)*

Frequenza relativa *p* oggetto di stima	Indice di asimmetria *(skewness)*	Indice di curtosi *(kurtosi)*	Test di normalità di kolmogoroff (statistica test)	Test di normalità di kolmogoroff *(p_value)*	Accettazione dell'ipotesi di normalità
3,418	0,006	0,399	0,007	0,035	
3,527	-0,025	0,453	0,007	0,024	
3,581	0,015	0,326	0,006	0,092	*
3,631	-0,013	0,730	0,007	0,013	
3,795	-0,071	1,020	0,006	0,141	*
3,883	-0,024	0,649	0,005	0,150	*
3,911	-0,001	0,528	0,004	0,150	*
3,970	0,028	0,529	0,010	0,010	
4,370	-0,046	0,651	0,004	0,150	*
4,377	-0,021	0,748	0,007	0,021	
4,541	-0,033	0,808	0,006	0,073	*
4,561	-0,047	0,919	0,006	0,046	
5,358	-0,056	1,295	0,007	0,031	
5,619	-0,297	5,324	0,008	0,010	
5,622	-0,041	1,375	0,007	0,022	
6,462	-0,072	1,684	0,006	0,061	*
6,595	-0,133	2,541	0,006	0,040	
7,003	-0,170	2,633	0,004	0,150	*
7,393	-0,129	2,851	0,006	0,045	
8,455	-0,162	3,363	0,008	0,010	
8,534	-0,132	2,119	0,005	0,150	*
8,818	-0,236	4,319	0,005	0,150	*
9,287	-0,333	6,829	0,007	0,039	
9,489	-0,138	2,495	0,005	0,150	*
11,169	-0,590	13,500	0,006	0,073	*
11,857	-2,002	67,254	0,011	0,010	
12,141	-0,451	10,108	0,006	0,084	*
12,956	-0,557	12,809	0,005	0,150	*
15,411	-0,819	22,040	0,006	0,046	
15,737	-1,285	37,849	0,007	0,010	
16,947	-1,866	63,104	0,010	0,010	
21,974	-2,228	80,126	0,011	0,010	
22,524	-1,015	27,693	0,006	0,049	
23,360	-6,941	360,005	0,021	0,010	
24,524	-1,059	30,147	0,006	0,073	*
24,545	-8,920	502,153	0,025	0,010	
25,392	-1,996	67,879	0,011	0,010	
31,341	-1,942	66,589	0,010	0,010	
32,061	-3,885	165,840	0,014	0,010	
33,063	-3,670	152,942	0,013	0,010	
41,291	-4,803	219,524	0,015	0,010	
41,824	-10,231	604,023	0,027	0,010	
44,129	-11,010	664,218	0,027	0,010	
44,309	-12,629	798,296	0,031	0,010	

B.3. Conclusioni

L'idea alla base di questo lavoro era quello di fornire un giudizio circa l'ipotesi della normalità delle distribuzioni delle stime riferite a variabili che potrebbero essere oggetto di rilevazione campionaria tramite *long form* in occasione del censimento della popolazione.

Alla luce delle considerazioni fatte emerge che tale l'ipotesi, utile nella fase di costruzione di stime per intervalli di confidenza, viene a cadere in presenza di frequenze percentuali molto piccole o molto grandi, mentre per valori intermedi la normalità è generalmente ammessa dai risultati dei test. Inoltre, si è osservato che la soglia minima e massima per l'accettazione della normalità delle stime campionarie, dipende molto dalla dimensione demografica del dominio di stima.

In futuro si potranno fare ulteriori approfondimenti sulle distribuzioni campionarie delle stime di frequenze relative e assolute, specialmente nei casi in cui queste assumano valori estremi; i risultati potranno risultare strategici per la produzione di stime di intervallo maggiormente accurate, in quanto basate su ipotesi distribuzionali più pertinenti.

Riferimenti bibliografici

Abbatini D., Cassata L., Martire F., Reale A., Ruocco G. e Zindato D. (2007). La progettazione dei censimenti generali 2010-2011. Analisi comparativa di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani. ISTAT, Dati e prodotti, Pubblicazioni scientifiche, Documenti n. 9/2007. Roma.
http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2007/2007_9.pdf

Astorri P., Bianchi G., Di Pede F., Esposito N., Patruno E., Reale A., Ronchi I. e Talice S. (2007). Metodi di determinazione delle aree di censimento a livello sub comunale. Relazione presentata alla XXVIII Conferenza Italiana di Scienze Regionali, Bolzano 26-28 settembre.

Berntsen E., De Angelis S. e Mastroluca S. (2008). *La progettazione dei censimenti generali 2010-2011. L'uso dei dati censuari del 2000-2001: alcune evidenze empiriche*. ISTAT, Dati e prodotti, Pubblicazioni scientifiche, Documenti n. 2/2008. Roma.
http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2008/doc2_2008.pdf

Borrelli F., Carbonetti G. e De Felici L. (2007). Strategie campionarie per la stima di variabili di censimento con long form. Atti della XXVIII Conferenza Italiana di Scienze Regionali, Bolzano 26-28 settembre.

Borrelli F., Carbonetti G., De Felici L. e Solari F. (2008). Metodologie di stima per piccole aree applicabili a variabili di censimento rilevabili tramite questionario long form. Atti della XXIX Conferenza Italiana di Scienze Regionali, Bari 24-26 settembre.

Borrelli F., Carbonetti G. e De Felici L. (2009). Problemi di accuratezza delle stime da campioni di famiglie in un contesto censuario. Giornate di Studio sulla Popolazione, VIII Edizione, Milano 2-4 febbraio.

Carbonetti G. e De Vitiis C. (2007). Efficienza di stime campionarie relative ad un sottoinsieme di variabili di censimento. Atti della Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni". CNR, Roma novembre.
http://www.istat.it/istat/eventi/2007/interconferenza/interventi/Carbonetti_DeVitiis.pdf

Carbonetti G., Dardanelli S., Fiorello E., Mastroluca S. e Verrascina M., (2008a). Ipotesi di innovazione per il censimento della popolazione del 2011: una valutazione degli effetti su un possibile piano di diffusione. Atti della XXIX Conferenza Italiana di Scienze Regionali, Bari 24-26 settembre.

Carbonetti G. e Fortini M. (2008b). Sample results expected accuracy in the Italian population and housing census. Joint UNECE/Eurostat Meeting on Population and Housing Censuses. UN, Ginevra maggio. ECE/CES/AC.6/2008/4
http://www.unece.org/stats/documents/ece/ces/ge.41/2008/4.e.pdf

Carbonetti G., Fortini M. e Solari F. (2008c). Innovations on methods and survey process for the 2011 Italian population census. Proceedings of the European Conference on Quality in Official Statistics, Roma.

Carbonetti G. (2009). Use of sampling strategy in the Italian population census and accuracy of estimates for different territorial domains. ITACOSM09 - First Italian Conference on Survey Methodology, Siena (Italy) june.

Cicchitelli G., Herzel A. e Montanari G. E. (1992). *Il campionamento statistico*. Bologna: il Mulino.

Cocchi D. (2007). Uso dei campioni nelle rilevazioni censuarie. Atti della Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni". CNR, Roma novembre.
http://www.istat.it/istat/eventi/2007/interconferenza/interventi/Cocchi.pdf

Crescenzi F., Fortini M., Gallo G. e Mancini A. (2009). La progettazione dei censimenti generali 2010-2011. Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione. ISTAT, Dati e prodotti, Pubblicazioni scientifiche, Documenti n. 6/2009. Roma.
http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2009/doc6_2009.pdf

Deville J.C. e Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the american statistical association*, vol. 87: 367-382.

Fortini M., Gallo G., Paluzzi E., Reale A. e Silvestrini A. (2007). La progettazione dei censimenti generali 2010-2011. Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento. ISTAT, Dati e prodotti, Pubblicazioni scientifiche, Documenti n. 10/2007. Roma.
http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2007/2007_10.pdf

Pagliuca D. (2005). *Genesees v.3.0., Funzione Riponderazione. Manuale utente ed aspetti metodologici.* Tecniche e Strumenti, Istat, n. 2. Roma.

Särndal C.E., Swensson B.e Wretman J. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag.

UNECE United Nations Economic Commission for Europe and Statistical Office of the European Communities 2006. Conference of European Statisticians. *Recommendations for the 2010 Censuses of Population and Housing.* ECE/CES/STAT/NONE/2006/4

United Nations (2007). *Principles and Recommendations for Population and Housing Censuses - Revision 2*. Expert Group Meeting on the 2010 World Programme on Population and Housing Censuses.