istat
working
papers

N. 3
2012

Metodologie di stima per piccole aree applicabili a variabili di censimento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici e Fabrizio Solari

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 3
2012

Metodologie di stima per piccole aree applicabili a variabili di censimento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici e Fabrizio Solari

Istat

Istat Working Papers

Metodologie di stima per piccole aree applicabili a variabili di censimento

N. 3/2012

ISBN 88-458-1707-5

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

Metodologie di stima per piccole aree applicabili a variabili di censimento

Francesco Borrelli, Giancarlo Carbonetti, Luana De Felici e Fabrizio Solari

Sommario

Durante la fase di progettazione del Censimento della Popolazione e delle Abitazioni del 2011, l'Istituto Nazionale di Statistica ha condotto uno studio per valutare la possibilità di affiancare all'usuale enumerazione delle principali caratteristiche demografiche, l'uso di campioni di famiglie per stimare alcune delle variabili tradizionalmente rilevate in modo esaustivo. L'adozione di metodi di stima comporta un costo in termini di errore di campionamento che pone vincoli sull'accuratezza delle stime prodotte. In particolare le stime basate su metodi diretti e riferite a contesti territoriali minimi o a piccoli aggregati di popolazione implicano errori di campionamento molto elevati. A seguito di ciò è stato proposto il ricorso a metodi indiretti che utilizzino al meglio le informazioni disponibili. L'obiettivo di questo lavoro è quello di valutare la praticabilità di tecniche di stima per piccole aree che si basano sull'uso di modelli lineari ad effetti misti riferiti a livello di unità o a livello di piccola area. I metodi studiati sono stati sottoposti a simulazioni su dati del censimento del 2001 e i risultati sono stati confrontati tramite indicatori sintetici, in termini di distorsione e variabilità.

Parole chiave: censimento, *long form*, campionamento, piccole aree

Abstract

The Italian National Institute of Statistics (ISTAT) has carried out a project study to evaluate the opportunity of using samples of households in the 2011 Population and Housing Census, in order to produce estimates for traditionally exhaustive variables. The usage of estimation methods, based on direct estimators, implies sampling errors affecting the accuracy of the estimates. In particular the estimates referred to minimal territorial contexts or to "rare" populations imply high sampling errors. Furthermore, the need of adopting indirect methods based on the use of the available information, has arisen. In this paper the direct estimator, the generalized regression estimator (GREG), two synthetic estimators and two estimators which are known in literature as EBLUPs ("Empirical Best Linear Unbiased Predictors") will be considered. These different estimators will be compared by means of a simulation study to assess their properties with overall indicators in terms of bias and mean squared error. The simulations will be carried out on 2001 census data.

Keywords: census, *long form*, sampling, small area

Indice

1. Introduzione[1]

La Direzione centrale dei censimenti generali dell'Istat è stata fortemente impegnata nella progettazione del Censimento della popolazione e delle abitazioni del 2011 con l'obiettivo di introdurre elementi di novità rispetto alle passate tornate censuarie. Una delle scelte di innovazione è relativa alla decisione di adottare le tecniche di campionamento per produrre stime riferite ad alcune delle variabili socio-economiche tradizionalmente rilevate in modo esaustivo.

La strategia stabilita è quella di somministrare, nei comuni capoluogo di provincia e in tutti i comuni di dimensione superiore ai 20mila abitanti, un questionario contenente tutte le variabili (long form) solo ad un campione di famiglie e un questionario in forma ridotta (short form), contenente solo le principali variabili demografiche e poche variabili socio-economiche, a tutte le famiglie non campionate. Nei comuni più piccoli, invece, l'indicazione è quella di sottoporre il questionario in versione long a tutte le famiglie in modo esaustivo secondo l'approccio del censimento tradizionale.

Il disegno campionario deciso per la formazione dei campioni di famiglie fa riferimento ad un procedimento di estrazione casuale semplice dalla lista anagrafica di ciascun comune eleggibile al campionamento.

Poiché l'adozione delle tecniche campionarie introduce un errore di campionamento, è stata fatta una preliminare ed attenta valutazione dell'accuratezza attesa delle stime producibili. Tale errore, misurato tramite verifica sperimentale, risulterà essere più evidente nel caso di stime riferite a domini territoriali "piccoli" o a sottopopolazioni aventi caratteristiche "rare" in quanto, in tali casi, il sottoinsieme di osservazioni campionarie aventi la caratteristica oggetto di stima non è quasi mai sufficientemente rappresentativo. Gli stimatori diretti forniscono infatti un elevato grado di affidabilità se la numerosità campionaria è adeguatamente ampia; invece, nel caso di dimensione ridotta, potrebbero comportare livelli di accuratezza non accettabili.

Si è riscontrata, quindi, la necessità di proporre metodi alternativi ai metodi classici al fine di determinare stime affidabili anche in presenza di una numerosità campionaria esigua. Tra questi, i più noti in letteratura sono i metodi di stima "per piccole aree" i quali permettono di utilizzare l'informazione proveniente dalle unità di un'area contenente la sottopopolazione oggetto di stima e/o da precedenti occasioni di indagine.

Nel presente lavoro si prendono in considerazione metodi di stima consolidati da un punto di vista applicativo e altri di recente sviluppo che si basano su modelli lineari ad effetti misti definiti a livello di unità o a livello di piccola area.

I metodi proposti sono stati sperimentati per la stima di alcuni parametri riferiti a sottopopolazioni di numerosità esigua. Per le verifiche sono stati utilizzati i dati del censimento della popolazione del 2001 e i risultati sono stati confrontati tramite indicatori riassuntivi, in termini di distorsione e variabilità, calcolati sulla distribuzione campionaria ottenuta tramite le simulazioni.

Dopo la descrizione della strategia campionaria decisa per il censimento del 2011 (Capitolo 2), si passa alla trattazione metodologica di alcuni metodi di stima per piccole aree, prima a carattere generale (Paragrafo 3.1), poi in maniera più puntuale con riferimento ai metodi sperimentati in questo lavoro (Paragrafo 3.2). Successivamente (Paragrafo 3.3) si illustrano i criteri impiegati per confrontare la "*performance*" degli stimatori e, in seguito, viene delineato l'ambito delle sperimentazioni (Capitolo 4) e i risultati delle analisi condotte (Capitolo 5). Infine, si riportano alcune considerazioni conclusive e indicazioni di sviluppo futuro (Capitolo 6).

[1] Il presente lavoro è stato realizzato nell'ambito del Gruppo di Lavoro "Metodologie di campionamento e di integrazione della sotto-copertura anagrafica del 15° Censimento generale della popolazione e delle abitazioni" costituito con delibera n° 137/DPTS del 19 novembre 2008 e successivamente prorogato con delibera n° 3/DPTS del 13 gennaio 2010. In particolare, sono esposti i risultati di uno studio condotto nel sottogruppo A "La strategia di campionamento delle famiglie cui somministrare il *long form*, con particolare riferimento alla produzione di stime per piccoli domini e di tabelle di risultati sparsi".
Ai fini dell'attribuzione delle singole parti si specifica che: i capitoli 1, 2, 6 e i paragrafi 5.3 e 5.4 sono interamente redatti da G. Carbonetti, il capitolo 3 da F. Solari, il capitolo 4 da L. De Felici, i paragrafi 5.1 e 5.2 da F. Borrelli.

2. Strategia campionaria per il censimento del 2011

Per valutare la praticabilità dell'impiego di strategie campionarie per produrre stime accurate riferite ad alcune delle variabili tradizionalmente rilevate in modo esaustivo, sono state fatte alcune considerazioni metodologiche sui disegni campionari e i metodi di stima adottabili. A tale scopo, la linea di principio seguita è stata indicata dall'esigenza di progettare disegni di campionamento semplici così da lasciare aperta la possibilità di adottare stimatori con struttura non standard come gli stimatori per piccole aree (Cocchi, 2007).

Sono state studiate differenti strategie campionarie riferite a disegni di campionamento casuali semplici, stratificati e a grappolo (areali) e metodi di stima diretti basati sull'impiego degli stimatori di ponderazione vincolata.[2] Le varie alternative prese in considerazione sono state sottoposte a sperimentazione, su insiemi di dati del censimento della popolazione del 2001, per stimare le frequenze relative ad un insieme di variabili, singole o di incrocio, connesse alle tematiche che interessano la rilevazione tramite long form (Borrelli *et al.*, 2011).

Le stime sono prodotte con riferimento sia al dominio territoriale comunale che a quello subcomunale rappresentato da specifiche aggregazioni di sezioni di censimento contigue (le *aree di censimento di centro abitato*) costruite, per lo studio in questione, con dimensione demografica compresa tra 5mila e 15mila unità (Bianchi *et al.*, 2010).

Le sperimentazioni sono state condotte sui dati relativi a 40 comuni, scelti per diversa ampiezza demografica e collocazione geografica, per i quali sono state disegnate complessivamente 498 aree di censimento di centro abitato (circa il 28% delle aree disegnate per il censimento 2011), e un numero di famiglie pari a 2.243.511 (poco più del 10% dell'universo delle famiglie residenti censite nel 2001).

I risultati delle simulazioni hanno messo in evidenza che la strategia di campionamento di famiglie da lista anagrafica è preferibile rispetto a quella di un campionamento areale a grappolo relativa alle sezioni di censimento. Tuttavia, i risultati delle sperimentazioni hanno mostrato che il campionamento areale rappresenta un'alternativa in grado di fornire stime qualitativamente soddisfacenti per via di un effetto *cluster* molto ridotto. 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 (Borrelli *et al.*, 2007; Carbonetti e De Vitiis, 2007).

In base a tali considerazioni, fissato il disegno casuale semplice di famiglie (CCSFAM), le sperimentazioni sono proseguite considerando diverse frazioni di campionamento per valutare i guadagni di efficienza ottenibili all'aumentare della frazione sondata (Carbonetti e Fortini, 2008), specialmente per la stima delle frequenze più piccole.

Nella tavola 1 sono presentati alcuni risultati delle sperimentazioni relativi a diverse frazioni di campionamento per la stima di frequenze relative p (riferite al dominio dell'area di censimento), confrontate in termini di coefficiente di variazione.[3]

[2] Tali metodi permettono di aumentare la rappresentatività del campione e la coerenza dei dati osservati con alcune informazioni note sulla popolazione di riferimento attraverso il procedimento della *calibrazione* (Deville e Särndal, 1992).

[3] Il coefficiente di variazione misura l'errore che mediamente si commette con la stima campionaria:

$$\mathrm{cv}(\hat{p}_x) = \frac{\sigma(\hat{p}_x)}{p_x} \cdot 100$$

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 .

Tavola 1 - Distribuzione dei cv mediani per classi di frequenze percentuali (parametro oggetto distima), riferite alle aree di censimento di centro abitato. Confronto per tre differenti frazioni di campionamento nel disegno CCSFAM

CLASSI DI P	Frazione di campionamento=10%	Frazione di campionamento=20%	Frazione di campionamento=33%
< 0,05%	220,51	142,00	98,21
0,05% ├ 0,1%	111,48	74,20	51,14
0,1% ├ 0,25%	75,57	49,97	34,76
0,25% ├ 0,5%	50,70	33,97	23,44
0,5% ├ 1%	35,54	23,74	16,56
1% ├ 2,5%	23,62	15,33	10,68
2,5% ├ 5%	15,50	10,09	7,04
5% ├ 10%	10,46	6,93	4,82
10% ├ 15%	7,06	4,40	3,13
15% ├ 20%	5,57	3,54	2,42
20% ├ 30%	4,50	2,84	1,93
≥ 30%	3,20	1,94	1,34

Dall'analisi dei valori emerge che il disegno di campionamento casuale semplice di famiglie con la frazione sondata del 33% presenta, come previsto, risultati migliori rispetto alle frazioni più piccole. A riguardo, con il passaggio dalla strategia campionaria che prevede una frazione di campionamento pari al 10% a quella con frazione di campionamento pari al 20% è attesa una riduzione percentuale dell'errore (misurato in termini di cv) compresa tra il 33% e il 38%; invece, nel caso di un incremento del campione fino alla frazione sondata del 33% la riduzione dell'errore sarà contenuta tra il 53% e il 58%.

Un altro importante risultato è quello descritto nella tavola 2 in cui si evidenzia come, all'aumentare della frazione di campionamento, la percentuale di stime che presentano errori elevati tende a diminuire; in particolare, all'aumentare della frazione sondata le stime tendono a disporsi nelle classi caratterizzate dai livelli di cv più bassi (Borrelli *et al.*, 2008).

Tavola 2 - Distribuzione delle stime delle frequenze percentuali riferite ad aree di censimento di centro abitato per classi di cv. Confronto per tre differenti frazioni di campionamento nel disegno CCSFAM

CLASSI DI CV	Frazione di campionamento=10%	Frazione di campionamento=20%	Frazione di campionamento=33%
< 2%	0,28	4,69	10,98
2% ├ 5%	10,72	17,54	21,41
5% ├ 10%	15,95	22,73	28,30
10% ├ 20%	26,62	26,56	18,36
20% ├ 50%	28,09	17,93	14,87
50% ├ 100%	10,76	7,00	4,19
100% ├ 200%	4,80	2,36	1,89
≥ 200%	2,77	1,19	-

In base ai dati della tavola 1 sono stati determinati per due prefissati livelli di cv "critico" i valori minimi delle frequenze percentuali *p* da stimare sotto il quale si "potrebbe" commettere un errore (valore atteso) maggiore di quello fissato (Tavola 3).

Tavola 3 - Valori minimi delle frequenze percentuali da stimare che presentano un cv atteso non superiore al 10% e al 20%. Confronto per tre differenti frazioni di campionamento nel disegno CCSFAM

CV "critico"	Frazione di campionamento=10%	Frazione di campionamento=20%	Frazione di campionamento=33%
10%	5%	2,5%	1%
20%	2,5%	1%	0,5%

Ad esempio, se si fissa il livello di cv critico del 10% con una frazione di campionamento del 10% tutte le stime di valori percentuali inferiori al 5% potrebbero comportare un errore (in termini di cv) superiore al 10%.

L'obiettivo finale di questo lavoro è dunque quello di ridurre l'insieme delle frequenze percentuali oggetto di stima che potrebbero comportare un livello di cv superiore a quello critico, cioè aumentare l'accuratezza complessiva delle relative stime diminuendo l'errore, a parità di disegno e frazione di campionamento, tramite l'impiego di metodi alternativi agli stimatori diretti.

Questo è il principale motivo per cui si ricorre ai metodi per piccole aree, che hanno proprio il vantaggio di ridurre la variabilità delle stime campionarie. Nel prossimo capitolo, dopo una breve introduzione alla metodologia di stima per piccole aree, verranno presentati in modo analitico i metodi che sono stati oggetto di successiva valutazione.

Nell'ambito di questo lavoro, le sperimentazioni sull'impiego dei metodi di stima per piccole aree sono state riferite al caso di campioni di famiglie estratti secondo il disegno casuale semplice da lista e la frazione di campionamento del 10%. È stata scelta la frazione sondata più bassa tra quelle prese in esame in quanto (Tavola 2) all'aumentare della dimensione del campione i risultati migliorano (la percentuale di stime con elevati cv diminuisce e quella con cv bassi aumenta); in conseguenza di ciò, i risultati ottenuti per tale frazione sono estendibili e sicuramente rafforzati nel caso di frazioni di campionamento più elevate.

3. Metodologia di stima per piccole aree

3.1 Introduzione alla stima per piccole aree

Negli ultimi anni l'esigenza di ottenere stime ad un livello il più disaggregato possibile ha determinato un interesse sempre più crescente verso l'impiego dei metodi di stima per piccole aree. A riprova di ciò non c'è solo la vasta letteratura prodotta negli ultimi dieci anni, ma anche i molteplici progetti riguardanti il problema di stima per piccole aree che hanno coinvolto alcuni dei principali Istituti Nazionali di Statistica e le più importanti strutture accademiche.

Restando nell'ambito europeo vanno ricordati i progetti co-finanziati da Eurostat: EURAREA, BIAS, AMELI, SAMPLE, BLU-ETS (i metodi di stima per piccole aree sono coinvolti solo in un *work-package*) e ESSnet SAE. A riguardo, l'Istat ha preso parte al progetto EURAREA, sta partecipando al progetto BLUE-ETS, ed è coordinatore del progetto ESSnet SAE.

Con il termine piccole aree si indica una suddivisione della popolazione individuata da aree geografiche e/o da classificazioni di tipo demografico o socio-economico per le quali non si è in grado di produrre stime dirette con un livello di precisione accettabile. In tali circostanze si fa, usualmente, ricorso ai metodi di stima per piccole aree. Tali metodi consentono di migliorare la precisione delle stime, avvalendosi anche dell'apporto delle osservazioni campionarie appartenenti ad un area più vasta (definita "macro-area") contenente la piccola area di interesse e/o relative ad altre occasioni d'indagine oltre a quella corrente.

Le modalità con cui tali informazioni supplementari concorrono alla determinazione delle stime avviene attraverso il riferimento ad un modello, implicito od esplicito, che stabilisce il legame esistente tra le osservazioni appartenenti alla stessa macro-area. L'introduzione di tali metodi, pur introducendo una componente distorsiva legata alla validità del modello ipotizzato, consente nella maggior parte dei casi, un sostanziale aumento della precisione delle stime ottenute rispetto alle stime dirette. Al fine di valutare l'entità della componente distorsiva introdotta nel processo di stima, è necessario effettuare studi simulativi basati su dati censuari o su dati provenienti da pseudo-popolazioni.

3.2 Metodologia alla base degli stimatori per piccole aree

In questo paragrafo si darà una descrizione formale dei principali stimatori diretti e dei più rilevanti stimatori per piccole aree. Tra gli stimatori diretti, saranno esaminati lo stimatore espansione e lo stimatore di regressione generalizzata GREG (*Generalized Regression Estimator*).

Successivamente, verrà esposto quello che può essere considerato il più diffuso metodo di stima per piccole aree, basato sull'esplicitazione formale di un modello: il predittore EBLUP (*Empirical Best Linear Unbiased Predictor*) basato su un modello lineare ad effetti misti a livello di unità elementare e l'analogo predittore basato su un modello lineare ad effetti misti a livello di area. Inoltre, per entrambi i modelli saranno presi in considerazione anche gli stimatori corrispondenti alla sola componente sintetica del predittore EBLUP.

Prima di procedere, è opportuno dare una definizione del parametro che si vuole stimare, o meglio predire, attraverso i metodi di stima per piccole aree. I parametri di interesse sono costituiti dai valor medi θ_d relativi della variabile di interesse y all'interno di ciascuna piccola area d. Formalmente, si ha:

$$\theta_d = \frac{1}{N_d}\sum_{i\in U_d} Y_i \quad , \quad d = 1,\ldots,D \quad , \tag{1}$$

dove U_d indica l'insieme delle unità della popolazione appartenenti alla piccola area d, Y_i è il valore osservato sull'i-esima unità della popolazione ($i=1,\ldots..,N_d$) per la variabile di interesse y.

3.2.1 Stimatore espansione

Lo stimatore espansione può essere formulato nel seguente modo:

$$\hat{\theta}_d = \frac{1}{N_d}\sum_{i\in s_d} w_i Y_i \quad , \tag{2}$$

in cui s_d denota l'insieme delle unità campionarie appartenenti alla piccola area d e w_i è il peso diretto di riporto all'universo associato all'i-esima unità campionaria nell'area d ($i=1,\ldots..,n_d$).

3.2.2 Stimatore di regressione generalizzata

Come noto in letteratura, uno stimatore diretto più efficiente dello stimatore espansione è lo stimatore GREG, che si basa sull'adozione di un modello lineare che lega la variabile di interesse ad alcune variabili ausiliarie. Il guadagno in termini di efficienza rispetto allo stimatore espansione risulta essere funzione del grado di correlazione esistente tra la variabile di interesse e le covariate considerate.

Lo stimatore GREG, che costituisce un caso particolare dello stimatore di calibrazione (Deville e Särndal, 1992), rientra nei metodi cosiddetti "assistiti da modello", in quanto si utilizza un modello per il calcolo delle stime, mentre le proprietà dello stimatore sono valutate all'interno del disegno di campionamento adottato.

L'espressione formale dello stimatore GREG può essere esplicitata aggiungendo allo stimatore espansione un termine di correzione, che dipende dalla differenza tra le medie a livello di popolazione delle variabili ausiliarie e le corrispondenti stime campionarie:

$$\hat{\theta}_d^{GREG} = \frac{1}{N_d}\sum_{i\in s_d} w_i y_i + \left(\overline{X}_d - \frac{1}{N_d}\sum_{i\in s_d} w_i x_i\right)^T \hat{\beta} \ . \tag{3}$$

Nell'espressione precedente il vettore $\overline{X}_d = \left(\overline{X}_{d,1},...,\overline{X}_{d,p}\right)^T$ denota l'insieme delle medie di popolazione delle p covariate, mentre $\hat{\beta}$ indica la stima dei coefficienti di regressione del modello lineare standard, ovvero:

$$y_{di} = x_{di}^T \beta + \varepsilon_{di} \quad , \tag{4}$$

con

$$E(\varepsilon_{di}) = 0 \quad , \quad Var(\varepsilon_{di}) = \sigma_\varepsilon^2 \quad , \quad \forall i = 1,\ldots,N_d \quad e \quad d = 1,\ldots,D \quad ,$$

dove $x_{di} = \left(x_{di,1},...,x_{di,p}\right)^T$ è il vettore delle osservazioni campionarie delle p covariate relativo all'i-ma unità dell'area d. La stima del coefficiente di regressione β si ottiene attraverso l'usuale metodo dei minimi quadrati ponderati, ossia:

$$\hat{\beta} = \left(\sum_{i \in s_d} w_i x_i x_i^T \right)^{-1} \sum_{i \in s_d} w_i x_i y_i \quad . \tag{5}$$

3.2.3 Predittore EBLUP basato su un modello a livello di unità

Come accennato nel paragrafo 3.1, un modello fondamentale nell'ambito dei problemi di stima per piccole aree è il modello lineare ad effetti misti specificato a livello di unità, con effetti casuali di area ed errori accidentali distribuiti normalmente e tra loro indipendenti (Battese *et al*., 1988).

Tale modello può essere esplicitato nel seguente modo:

$$y_{di} = x_{di}^T \beta + u_d + e_{di} \quad , \tag{6}$$

in cui per le componenti casuali del modello valgono le seguenti ipotesi:

$$u_d \sim \text{iid } N(0,\sigma_u^2) \quad , \quad e_{di} \sim \text{iid } N(0,\sigma_e^2) \quad , \quad \forall i = 1,\ldots,N_d \quad e \quad d = 1,\ldots,D \quad .$$

Una formulazione più compatta del modello (6) si ottiene facendo ricorso alla notazione matriciale

$$y = X\beta + Zu + e \quad , \tag{7}$$

dove y è il vettore delle osservazioni relative alla variabile di interesse, X la matrice dei valori rilevati sulle covariate, il vettore e rappresenta gli errori accidentali, Z la matrice di incidenza delle unità in ogni area e u il vettore delle componenti casuali di area.

Nell'ipotesi che le componenti di varianza σ_e^2 e σ_u^2 siano note, il miglior stimatore lineare non distorto di β, ottenuto attraverso il metodo dei minimi quadrati ordinari, è dato da

$$\hat{\beta} = \left(X^T V X\right)^{-1} X^T V\, y \quad , \tag{8}$$

in cui la matrice di varianza V di y è esprimibile tramite la seguente espressione:

$$V = \sigma_e^2 I + \sigma_u^2 Z Z^T \quad . \tag{9}$$

Quando le componenti di varianza σ_e^2 e σ_u^2 sono incognite, la stima $\hat{\beta}$ del vettore dei coefficienti di regressione β è ottenuta attraverso l'algoritmo di Newton-Raphson. Più precisamente si adotta un processo di stima iterativo che ad ogni passo, dopo aver sostituito nella (9) le stime $\hat{\sigma}_e^2$ e $\hat{\sigma}_u^2$ delle componenti di varianza ottenuta nell'iterazione precedente, prevede un nuovo calcolo dei coefficienti di regressione mediante la (8) e l'aggiornamento delle componenti di varianza fino ad ottenere la convergenza delle stime. L'aggiornamento delle componenti di varianza del modello può essere fatto massimizzando la funzione di verosimiglianza completa o la funzione di verosimiglianza ristretta (Cressie, 1992).

Il miglior predittore lineare empirico (EBLUP_UL) è uno stimatore di tipo composto che, trascurando il fattore di correzione per popolazioni finite, è dato da:

$$\hat{\theta}_d^{\,EBLUP\_UL} = \gamma_d \left[\bar{y}_d + \left(\bar{X}_d^T \hat{\beta} - \bar{x}_d^T \hat{\beta}\right)\right] + \left(1 - \gamma_d\right) \bar{X}_d^T \hat{\beta} \quad , \tag{10}$$

in cui

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d} \tag{11}$$

è il peso associato alla componente diretta del predittore, $\bar{y}_d$ e $\bar{x}_d$ denotano la media campionaria e il vettore delle medie campionarie rispettivamente della variabile di interesse y e delle covariate nell'area d, $\bar{X}_d$ è il vettore dei valori medi di popolazione delle covariate nell'area d.

La seconda parte del predittore (10), ossia $\bar{X}_d^T \hat{\beta}$, è detta componente sintetica del predittore (di qui in seguito verrà indicata con SINT_UL).

3.2.4 Predittore EBLUP basato su un modello a livello di area

Il modello che verrà descritto in questo paragrafo è definito a livello aggregato. Più precisamente, si definisce una relazione lineare tra il parametro di interesse θ_d e le medie delle variabili ausiliarie nelle piccole aree, ossia:

$$\theta_d = \overline{X}_d^T \beta + u_d \quad , \tag{12}$$

in cui

$$u_d \sim \text{iid } N\left(0,\sigma_u^2\right) \quad , \quad \forall d = 1,\ldots,D \quad ,$$

dove $\overline{X}_d^T$ è il vettore delle medie di popolazione delle p variabili ausiliarie nell'area d e σ_u^2 è la varianza dell'effetto casuale di area u_d .

Si supponga, inoltre, di disporre dell'insieme delle stime dirette $\hat{\theta}_d$ per le aree $d = 1,\ldots,D$, ossia:

$$\hat{\theta}_d = \theta_d + \overline{e}_d \quad , \tag{13}$$

in cui

$$\overline{e}_d \sim \text{iid } N\left(0, \varphi_d\right) \quad ,$$

con φ_d che indica la varianza campionaria associata a $\hat{\theta}_d$.
Combinando la (12) e la (13), si ottiene

$$\hat{\theta}_d = \overline{X}_d^T \beta + u_d + \overline{e}_d \quad , \tag{14}$$

in cui per u_d e $\overline{e}_d$ valgono le ipotesi fatte in precedenza. In forma matriciale il modello (14) può essere riscritto nel modo seguente:

$$\hat{\theta} = \overline{X}\beta + u + \overline{e} \quad , \tag{15}$$

con

$$u \sim MN\left(0,\sigma_u^2 I\right) \quad , \quad \overline{e} \sim MN\left(0, D\right)$$

e

$$D = \begin{pmatrix} \varphi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \varphi_D \end{pmatrix} .$$

Al fine di evitare problemi di identificabilità, le varianze campionarie φ_d $(\forall d = 1,\ldots,D)$ vengono assunte note. Nel caso in cui si non si disponga delle varianze campionarie, disponendo delle informazioni a livello individuale è possibile ottenere delle loro stime attraverso il modello (6), o la sua riformulazione matriciale (7). Nell'ipotesi di omoschedasticità della componente accidentale e nel modello (6), una stima $\hat{D}$ di D si ottiene stimando la varianza campionaria φ_d relativa all'area d mediante l'espressione

$$\hat{\varphi}_d = \frac{1}{n_d}\tilde{\sigma}_e^2 \quad , \tag{16}$$

con

$$\tilde{\sigma}_e^2 = \frac{1}{n - n^{(D)}}\sum_i\sum_d \left(y_{di} - \bar{y}_d\right)^2 \quad ,$$

in cui n è il numero di individui appartenenti al campione ed $n^{(D)}$ il numero complessivo di aree presenti nel campione. Nel presente lavoro si è preferito adottare le varianze delle stime dirette date dalla (16) in modo da garantire che sia i risultati ottenuti con un modello a livello di unità sia quelli ottenuti con un modello a livello di area avessero un unico modello come comune denominatore.

Lo stimatore dei minimi quadrati ponderati del vettore dei coefficienti di regressione β nella (15) è dato da:

$$\hat{\beta} = (\bar{X}^T V^{-1}\bar{X})^{-1}\bar{X}^T V^{-1}\bar{y} \quad , \tag{17}$$

dove $\bar{y}$ è il vettore delle medie campionarie per la variabile di interesse y, $\bar{X}$ è la matrice composta dalle righe $\bar{X}_d^T$, la matrice di covarianza $V = \sigma_u^2 I + \hat{D}$ è una matrice diagonale con elementi pari a $\sigma_u^2 + \hat{\varphi}_d$. Nel caso in cui la varianza degli effetti casuali di area σ_u^2 non sia nota, σ_u^2 e β sono stimati in modo iterativo, in modo analogo a quanto specificato nel precedente paragrafo per il modello a livello di unità, sostituendo $\hat{V} = \hat{\sigma}_u^2 I + \hat{D}$ a V nella (17).

L'espressione finale del predittore EBLUP a livello di area (EBLUP_AL) è data da:

$$\hat{\theta}_d^{EBLUP\_AL} = \gamma_d\hat{\theta}_d + (1-\gamma_d)\bar{X}_d^T\hat{\beta} \quad , \tag{18}$$

dove

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_e^2} \tag{19}$$

è il peso relativo allo stimatore diretto.

Come per il predittore EBLUP_UL, la seconda parte del predittore (18), vale a dire $\bar{X}_d^T\hat{\beta}$, è detta componente sintetica del predittore (nel proseguo del lavoro sarà denotato con SINT_AL).

3.3 Criteri di valutazione degli stimatori

Un'analisi comparativa tra gli stimatori considerati si rende necessaria per la scelta del metodo di stima per piccole aree ritenuto migliore nel caso in esame. Occorre, quindi, definire degli opportuni criteri in grado di fornire una misura dell'efficienza dei metodi e stabilirne una graduatoria. In letteratura si distingue tra criteri che si basano su studi simulativi condotti sulla vera popolazione, su una pseudo-popolazione o su una popolazione artificiale, e criteri che dipendano da analisi eseguite per mezzo di un unico campione.

Nel caso in cui si utilizzi la vera popolazione, è necessario avere a disposizione per tutte le unità della popolazione i valori delle variabili di interesse e delle variabili ausiliarie ad una data il più possibile vicina a quella a cui si riferisce l'indagine. Qualora non si disponga di tali valori, è possibile ovviare in due modalità: la popolazione può essere ottenuta utilizzando i dati provenienti da uno o più campioni sui quali sono stati osservati i valori sia della variabile di interesse che delle variabili ausiliarie, replicando le informazioni relative a ciascuna unità campionaria in modo proporzionale al proprio peso di riporto all'universo (pseudo-popolazione), oppure si possono ipotizzare sia per le variabili oggetto di stima che per le variabili ausiliarie specifiche forme distribuzionali, dalle quali generare i valori della variabile di interesse e delle covariate relativamente alla generica unità (popolazione artificiale).

L'analisi delle proprietà empiriche degli stimatori considerati si basa sul metodo Monte Carlo, ossia selezionando un determinato numero R di campioni dalla popolazione di interesse (reale, pseudo o artificiale) in conformità al disegno di campionamento adottato per l'indagine. Quindi, si costruiscono le distribuzioni empiriche degli stimatori che si desidera porre a confronto, applicando ognuno di tali metodi di stima all'insieme degli R campioni. Le informazioni che derivano da tali distribuzioni empiriche, possono essere condensate sotto forma di indici sintetici, utili a misurare la distorsione e la variabilità degli stimatori. Tra le diverse possibili misure sintetiche quelle maggiormente utilizzate fanno riferimento ai principali quantili della distribuzione. Nel presente lavoro, per misurare la prestazione dei diversi metodi analizzati in termine di distorsione ed errore quadratico medio, sono state adottate le seguenti misure (EURAREA Consortium, 2004):

Average Absolute Relative Bias:

$$\mathrm{AARB} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\frac{{}_r\hat{\theta}_d - \theta_d}{\theta_d}\right| \times 100 \quad , \tag{20}$$

Average Relative Root Mean Squared Error:

$$\mathrm{ARRMSE} = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{{}_r\hat{\theta}_d - \theta_d}{\hat{\theta}_d}\right)^2} \times 100 \quad . \tag{21}$$

I criteri di valutazione (20) e (21) misurano rispettivamente la distorsione relativa e la radice quadrata dell'errore quadratico medio relativo.

Una misura delle proprietà degli stimatori di tipo più conservativo si ottiene considerando il valore massimo dei valori assoluti della distorsione relativa ed il valore massimo della radice dell'errore quadratico medio relativo rispetto alle piccole aree. Più precisamente:

Maximum Absolute Relative Bias:

$$\mathrm{MARB} = \max_{d}\left|\frac{1}{R}\sum_{r=1}^{R}\frac{{}_{r}\hat{\theta}_{d}-\theta_{d}}{\theta_{d}}\right| \times 100 \quad , \tag{22}$$

Maximum Relative Root Mean Squared Error:

$$\mathrm{MRRMSE} = \max_{d}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{{}_{r}\hat{\theta}_{d}-\theta_{d}}{\theta_{d}}\right)^{2}} \times 100 \quad . \tag{23}$$

Come accennato in precedenza, le proprietà degli stimatori possono essere valutate per mezzo di un singolo campione i cui dati sono riferiti ad un periodo temporale il più vicino possibile a quello relativo ai valori noti del parametro. Due misure usualmente adottate in letteratura (Ghosh e Rao, 1994), che possono essere intese come misure di variabilità, sono l'errore relativo medio e l'errore quadratico medio:

Average Relative Error:

$$\mathrm{ARE} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{\hat{\theta}_{d}-\theta_{d}}{\theta_{d}}\right| \quad , \tag{24}$$

Absolute Squared Error:

$$\mathrm{ASE} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_{d}-\theta_{d}\right)^{2} \quad . \tag{25}$$

4. Sperimentazioni

4.1 Premessa

La necessità di produrre dati ad elevato dettaglio informativo ha posto l'esigenza di predisporre uno studio atto a valutare la possibilità di produrre stime affidabili per prefissati livelli territoriali e per specifiche sottoclassi di popolazione.

Le valutazioni effettuate in questo lavoro riguardano la possibilità di adottare stimatori per piccole aree nel contesto censuario e si propongono di rispondere a due importanti quesiti:

Q1) è possibile aumentare l'accuratezza delle stime di frequenze percentuali riferite a "popolazioni rare" per domini relativi alle aree di censimento di centro abitato?

Q2) è praticabile una strategia campionaria basata sull'adozione di stimatori per piccole aree anche nei comuni con dimensione compresa tra 5mila e 20mila abitanti?

Riguardo il punto Q1), risulta evidente che il ricorso a metodi di stima capaci di ridurre l'errore campionario permetterebbe di innalzare la qualità complessiva delle stime prodotte. In particolare, si potrebbe fortemente ridurre l'errore campionario di stime riferite ai domini di massimo dettaglio territoriale (le aree di censimento).

Per quanto concerne il punto Q2), pur avendo già deciso l'adozione della strategia campionaria, per il censimento 2011, solo nei comuni sopra i 20mila abitanti e in tutti i comuni capoluogo di provincia, le valutazioni finali potranno portare a ritenere il campionamento proponibile al censimento della popolazione anche nei comuni con ampiezza demografica compresa tra 5mila e 20mila abitanti, intesi come domini minimi di stima. Si potranno, infatti, avere indicazioni utili per valutare se si potrà procedere, in future occasioni censuarie, con la strategia campionaria anche nei comuni di tale dimensione, prevedendo l'impiego dei metodi per piccole aree, oppure non si potrà prescindere, in tali realtà comunali, dalla rilevazione completa ed esaustiva.

I metodi di stima per piccole aree descritti nel capitolo 3 sono stati sottoposti a simulazioni al fine di valutare l'accuratezza delle stime prodotte ed effettuare una valutazione comparativa tra i metodi diretti ed indiretti. Le valutazioni si sono basate sui risultati conseguiti tramite opportune sperimentazioni condotte impiegando dati del censimento della popolazione e delle abitazioni del 2001.

Per l'ambito della sperimentazione è stato necessario:

a) fissare il disegno di campionamento;
b) definire la macro-area da sottoporre a test;
c) stabilire l'insieme delle variabili di studio;
d) determinare l'insieme di variabili ausiliarie;
e) scegliere il dominio di stima;
f) costruire l'algoritmo di simulazione.

4.2 Il disegno di campionamento

Nella sperimentazione è stato considerato il disegno casuale semplice di famiglie (CCSFAM) che, in base a quanto illustrato nel capitolo 2, è stato deciso per l'adozione della strategia campionaria tramite questionari short e long form al censimento della popolazione e delle abitazioni del 2011.

Riguardo la dimensione dei campioni simulati è stata presa in esame la frazione del 10% con l'obiettivo di studiare i casi con più elevati errori di campionamento atteso; in tal modo, i risultati osservati potranno essere ritenuti più affidabili nei casi di frazioni campionarie più ampie (per esempio, per la frazione del 33% decisa per la strategia censuaria del 2011). In tali situazioni, inoltre, l'impiego dei metodi indiretti potrebbe riguardare un minor numero di casi in quanto, al crescere della frazione di campionamento, si riduce l'insieme di frequenze percentuali le cui stime saranno affette da elevata variabilità campionaria (Tavole 2 e 3).

4.3 Le macro-aree

Le macro-aree sono state definite in funzione dei due diversi obiettivi fissati (Tavola 4). Per l'obiettivo Q1 ("aumentare l'efficienza delle stime campionarie riferite alle aree di censimento") è stato deciso di sottoporre a sperimentazione macro-aree composte da aggregazioni di comuni omogenei rispetto all'ampiezza demografica e scelti in base alla disponibilità dei dati.[4] Le macro-aree relative ai test Q1_1, Q1_2 e Q1_3 (Tavola 4) sono state disegnate in modo da valutare l'eventuale differenza dei livelli di efficienza per ambiti riferiti a comuni più grandi o a comuni di più piccola dimensione.

[4] Per la definizione delle macro-aree è stata considerata la lista dei comuni impiegati nelle sperimentazioni sull'efficienza delle stime campionarie tramite short/long form (Borrelli *et al.*, 2011).

Per l'obiettivo Q2 ("valutare la praticabilità di una strategia campionaria tramite l'uso di short e long form nei comuni con dimensione tra 5mila e 20mila abitanti") si è fatto dapprima un tentativo di valutare macro-aree composte da comuni tra 5mila e 20mila unità (relative ai test Q2_1, Q2_2 e Q2_3 descritti nella Tavola 4) appartenenti alla stessa ripartizione territoriale; successivamente, si è provato a definire macro-aree riferite a contesti territoriali sub-ripartizionali al fine di valutare l'impatto della dimensione delle macro-aree, in termini di numerosità di piccole aree, sull'accuratezza delle stime (macro-aree dei test Q2_4, Q2_5, Q2_6 e Q2_7 della Tavola 4).

Tavola 4 - Dettaglio descrittivo delle sperimentazioni eseguite

CONTESTO TERRITORIALE	Test	Macro-area	Piccole aree	Numero delle piccole aree	Numero delle replicazioni campionarie
Comuni-Campione scelti per le sperimentazioni sull'efficienza delle stime campionarie tramite short/long form	Q1_1	Aggregazione di 2 comuni con popolazione maggiore di 500mila abitanti[(a)]	Aree di censimento di centro	146	500
	Q1_2	Aggregazione di 8 comuni con popolazione compresa tra 100mila e 500mila abitanti[(b)]	Aree di censimento di centro	137	500
	Q1_3	Aggregazione di 25 comuni con popolazione compresa tra 10mila e 100mila abitanti[(c)]	Aree di censimento di centro	57	500
Comuni di dimensione compresa tra 5mila e 20mila abitanti di alcune Ripartizioni dell'Italia	Q2_1	Ripartizione Centrale	Comuni 5-20mila	267	100
	Q2_2	Ripartizione Nord - Occidentale	Comuni 5-20mila	475	100
	Q2_3	Ripartizione Insulare	Comuni 5-20mila	182	500
Comuni di dimensione compresa tra 5mila e 20mila abitanti di alcune Regioni dell'Italia	Q2_4	Lazio	Comuni 5-20mila	84	500
	Q2_5	Toscana	Comuni 5-20mila	111	500
	Q2_6	Lazio + Toscana	Comuni 5-20mila	195	500
	Q2_7	Lazio + Toscana + Umbria	Comuni 5-20mila	215	500

(a) Napoli, Palermo.
(b) Bologna, Brescia, Firenze, Livorno, Novara, Padova, Perugia, Rimini.
(c) Alghero - Aosta - Belluno - Bolzano - Cittadella - Cuneo - Enna - Grottammare - Isernia - Legnago - Macerata - Maranello - Matera - Melfi - Pesaro - Piossasco - Pontedera - Porto Empedocle - Sestu - Sondrio - Squinzano - Todi - Tradate - Trapani - Vibo Valentia.

4.4 Le variabili di studio

Al fine di valutare la praticabilità dei metodi per piccole aree per la stima di percentuali riferite a sottopopolazioni aventi caratteristiche "rare", sono state selezionate alcune modalità che, con riferimento ai dati del censimento del 2001, hanno presentato frequenze molto piccole nella popolazione italiana. Nello specifico, sono state analizzate le seguenti 5 variabili (in parentesi è indicata una sigla identificativa):

1) "Coadiuvanti in Agricoltura" (posprof_setteco_3);
2) "Imprenditori e liberi professionisti in altre attività" (posprof_setteco_9);
3) "Popolazione totale di età maggiore o uguale a 15 anni occupata in Agricoltura" (ateco35m_1);
4) "Analfabeti" (titolo_6_6);
5) "Popolazione totale che si sposta giornalmente fuori del comune di dimora abituale" (luodst2m_2).

Le variabili "posprof_setteco_3" e "posprof_setteco_9" hanno fatto registrare al censimento del 2001, nei domini di studio, frequenze percentuali mediamente inferiori allo 0,1%; per le variabili "ateco35m_1" e "titolo_6_6" si sono osservati livelli di percentuale compresi tra lo 0,1% e l'1%; la variabile "luodst2m_2" ha presentato frequenze percentuali sempre superiori all'1%.

Si precisa che, sulla base dei contenuti informativi presenti nel questionario in forma ridotta (short) adottato per il censimento 2011, solo le variabili "posprof_setteco_3", "posprof_setteco_9" e "ateco35m_1" saranno oggetto di stima; le altre variabili, impiegate in questo lavoro a scopo di esercizio, saranno osservate nella rilevazione del 2011 in modo esaustivo su tutta la popolazione.

4.5 Le variabili ausiliarie

4.5.1 Individuazione dell'insieme delle covariate

La scelta delle variabili ausiliarie da impiegare nei modelli alla base degli stimatori per piccole aree considerati in questo lavoro, è stata indirizzata verso le principali variabili oggetto di osservazione esaustiva: *sesso*, *età*, *stato civile* e *numero di componenti in famiglia*.

Per ognuna delle variabili di interesse (Paragrafo 4.4) è stata studiata l'associazione con le sopra indicate variabili ausiliarie con lo scopo di identificare quelle che maggiormente spiegano il comportamento dei parametri da stimare. In particolare, è stato condotto uno studio di regressione logistica[5] in cui il comportamento di ciascuna variabile oggetto di stima è stato messo in relazione con le differenti covariate secondo una procedura di tipo *stepwise*.

L'analisi è stata condotta sui dati del comune di Bologna (352.448 record individui); per tale ambito sono state analizzate le statistiche di adattamento[6] dei modelli di regressione logistica definiti per successivi inserimenti delle covariate prese in considerazione, al fine di individuare quello che meglio si adattava. I risultati sono stati supportati anche dall'esito delle verifiche sul grado di significatività sia dell'intero modello[7] che delle singole covariate.[8]

Nello specifico, l'analisi condotta con riferimento alla variabile oggetto di stima "numero di analfabeti" (titolo_6_6) tutte le covariate prese in esame entrano nel modello; per la variabile "numero di coadiuvanti in agricoltura" (posprof_setteco_3) si osserva una associazione significativa solo con le variabili "sesso" ed "età".

In conclusione, si sono osservati casi in cui i contributi delle covariate scelte erano simili e casi in cui la variabilità dei parametri da stimare era spiegata solo da alcune delle covariate prese in esame; inoltre, per le covariate più significative, si sono evidenziati livelli di associazione più forti per alcune modalità rispetto ad altre e tali insiemi sono risultati differenti tra le variabili oggetto di stima.

È evidente che i risultati spingerebbero a definire insiemi diversi di covariate per i vari obiettivi di stima; una ulteriore diversificazione potrebbe essere disegnata anche in riferimento a contesti territoriali differenti. Per esempio, l'insieme di covariate per la stima di uno stesso parametro potrebbe essere diverso se riferito ad un dominio della ripartizione geografica "Nord-Ovest" rispetto ad uno della ripartizione "Sud". Così facendo si arriverebbe a determinare modelli specifici per ciascun obiettivo di stima, sia in termini di variabile che di dominio. Questo approccio porterebbe però ad una soluzione difficilmente praticabile nel contesto censuario sia per la molteplicità di obiettivi di stima per i quali si propone il ricorso ai metodi per piccole aree che per i numerosi riferimenti territoriali.

Questa considerazione ha necessariamente indirizzato la decisione, ai fini di questo lavoro, verso l'impiego di uno stesso insieme di covariate, scelto nel modo più ampio ("sesso", "età", "stato civile", "numero di componenti in famiglia"), per tutte le variabili oggetto di stima. Tale opzione

[5] Per l'analisi è stata impiegata la *"proc logistic"* nell'ambito del software statistico SAS.

[6] Sono stati considerati i seguenti criteri:
- "Criterio del logaritmo della funzione di verosimiglianza" (-2 Log L);
- "Akaike Information Criterion" (AIC);
- "Schwarz Criterion" (SC).

[7] Sono stati considerate le seguenti statistiche test:
- "Rapporto di verosimiglianza";
- "Score";
- "Wald".

[8] Tramite il test di "Wald".

presenta l'ulteriore vantaggio di garantire l'omogeneità con le variabili di calibrazione impiegate nell'ambito degli stimatori di ponderazione vincolata, scelti per il calcolo delle stime riferite ai domini pianificati dal disegno d'indagine (Borrelli *et al.*, 2011); in tal modo, si attende che le stime indirette garantiscano maggiormente la coerenza con le variabili esaustive.

Infine, nell'ottica di differenziare l'insieme di covariate per ciascun obiettivo di stima, la prospettiva di impiegare un questionario in forma ridotta, che presenti un numero maggiore di quesiti, potrebbe offrire la possibilità di disporre di un insieme più ampio di covariate dove selezionare quelle più correlate alle variabili di stima, da implementare nel modello di stima per piccole aree che si potrebbe eventualmente decidere di prendere in considerazione in fase di produzione.

4.5.2 Effetti singoli ed effetti misti

Una volta definito l'insieme delle variabili ausiliarie è stata valutata la possibilità di considerare tali variabili in modo *singolo* o *congiunto* (per la scelta di un modello basato su effetti singoli o su effetti congiunti):

a) il primo ad "effetti congiunti" considera le distribuzioni date da:
   - combinazione della popolazione per "sesso" (2 modalità) e per "classe di età" (16 classi: <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) per **32 modalità** in totale;
   - combinazione della popolazione per "sesso" (2 modalità) e per "stato civile" (5 classi: celibi; coniugati o separati di fatto; separati legalmente; divorziati; vedovi), per **10 modalità** totali;
   - famiglie per numero di componenti (**4 modalità** di classificazione: 1, 2, 3, 4 e più componenti).

b) il secondo ad "effetti singoli" considera le distribuzioni marginali seguenti:
   - popolazione per "sesso" (**2 modalità**);
   - popolazione per "classe di età" (**16 modalità** relative alle seguenti classi: <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);
   - popolazione per "stato civile" (**5 modalità**: celibi; coniugati o separati di fatto; separati legalmente; divorziati; vedovi);
   - famiglie per numero di componenti (**4 modalità** di classificazione: 1, 2, 3, 4 e più componenti).

Tali insiemi sono stati confrontati in termini di prestazione degli stimatori relativamente alla simulazione Q1_1 (Tavola 4). In particolare sono stati calcolati gli indicatori sintetici ARE e ASE descritti dalle espressioni (24) e (25).

I risultati sono simili per i due insiemi di variabili esplicative (Tavole 5 e 6); in base a ciò, si è scelto l'insieme relativo alle distribuzioni singole per motivi computazionali (tempi di elaborazione significativamente inferiori dovuti a matrici di dati di dimensioni ridotte).

Tavola 5 - Indicatori ARE e ASE calcolati per le variabili prese in esame nel caso di metodi basati su modelli ad effetti singoli

VARIABILI (media ponderata tra le aree di censimento)	Stimatori	ARE	ASE
posprof_setteco_3 (0,01%)	Espansione	1,00	1,07E-07
	GREG	0,97	1,07E-07
	SINT_UL	0,25	1,58E-08
	SINT_AL	0,49	2,50E-08
	EBLUP_UL	0,25	1,41E-08
	EBLUP_AL	0,49	2,50E-08
ateco35m_1 (0,55%)	Espansione	0,34	5,20E-06
	GREG	0,34	5,15E-06
	SINT_UL	0,29	3,61E-06
	SINT_AL	0,26	3,01E-06
	EBLUP_UL	0,22	2,03E-06
	EBLUP_AL	0,23	2,28E-06
titolo_6_6 (1,72%)	Espansione	0,24	1,80E-05
	GREG	0,35	2,52E-05
	SINT_UL	1,15	1,73E-04
	SINT_AL	0,42	4,51E-05
	EBLUP_UL	0,25	2,08E-05
	EBLUP_AL	0,23	1,54E-05
posprof_setteco_9 (1,93%)	Espansione	0,23	2,13E-05
	GREG	0,41	3,19E-05
	SINT_UL	1,74	3,13E-04
	SINT_AL	0,43	3,70E-05
	EBLUP_UL	0,37	2,37E-05
	EBLUP_AL	0,22	1,38E-05
luodst2m_2 (2,59%)	Espansione	0,17	2,84E-05
	GREG	0,19	3,06E-05
	SINT_UL	0,94	4,03E-04
	SINT_AL	0,43	1,70E-04
	EBLUP_UL	0,21	2,95E-05
	EBLUP_AL	0,17	2,59E-05

Tavola 6 - Indicatori ARE e ASE calcolati per le variabili prese in esame nel caso di metodi basati su modelli ad effetti congiunti

VARIABILI (media ponderata tra le aree di censimento)	Stimatori	ARE	ASE
posprof_setteco_3 (0,01%)	Espansione	1,00	1,07E-07
	GREG	0,97	1,07E-07
	SINT_UL	0,25	1,58E-08
	SINT_AL	0,62	3,49E-08
	EBLUP_UL	0,25	1,41E-08
	EBLUP_AL	0,62	3,49E-08
ateco35m_1 (0,55%)	Espansione	0,34	5,20E-06
	GREG	0,34	5,16E-06
	SINT_UL	0,29	3,58E-06
	SINT_AL	0,28	3,26E-06
	EBLUP_UL	0,22	2,02E-06
	EBLUP_AL	0,26	2,90E-06
titolo_6_6 (1,72%)	Espansione	0,24	1,80E-05
	GREG	0,35	2,54E-05
	SINT_UL	1,15	1,74E-04
	SINT_AL	0,41	2,93E-05
	EBLUP_UL	0,25	2,10E-05
	EBLUP_AL	0,25	1,49E-05
posprof_setteco_9 (1,93%)	Espansione	0,23	2,13E-05
	GREG	0,41	3,16E-05
	SINT_UL	1,73	3,14E-04
	SINT_AL	0,43	3,02E-05
	EBLUP_UL	0,37	2,35E-05
	EBLUP_AL	0,25	1,48E-05
luodst2m_2 (2,59%)	Espansione	0,17	2,84E-05
	GREG	0,19	3,04E-05
	SINT_UL	0,94	4,05E-04
	SINT_AL	0,39	1,48E-04
	EBLUP_UL	0,21	2,93E-05
	EBLUP_AL	0,17	2,63E-05

4.6 I domini di stima

Riguardo la scelta del dominio di stima, nell'ambito delle sperimentazioni effettuate, le stime delle variabili di interesse vengono prodotte con riferimento al dominio territoriale coincidente con l'area di censimento di centro abitato, per le simulazioni rispondenti all'obiettivo Q1 e al comune di dimensione compresa tra 5mila e 20mila abitanti per quelle relative all'obiettivo Q2.

4.7 L'algoritmo di simulazione

Per l'esecuzione delle sperimentazioni è stato definito un algoritmo che, lavorando in ambiente SAS (Statistical Analysis System), ha previsto i seguenti passi:

1) estrazione di un campione di famiglie secondo un disegno casuale semplice con frazione sondata pari al 10%;
2) calcolo delle stime campionarie delle frequenze relative riferite a ciascuna delle variabili prese in esame, utilizzando gli stimatori diretti (espansione, GREG) e gli stimatori indiretti (SINT_UL, SINT_AL, EBLUP_UL, EBLUP_AL) oggetto di valutazione;
3) iterazione dei passi 1) e 2) per un numero prefissato di volte (al massimo 500 replicazioni campionarie);[9]
4) calcolo di indicatori sintetici, in termini di distorsione e variabilità, sia per ciascuna piccola area che globali, sulla distribuzione campionaria simulata degli stimatori relativamente all'insieme delle replicazioni campionarie generate.

5. Descrizione dei risultati delle sperimentazioni

In questo capitolo sono presentati i risultati delle elaborazioni e le relative sintesi per valutare:

- il confronto tra i livelli di accuratezza delle stime basate sull'impiego degli stimatori per piccole aree rispetto ai metodi diretti per diversi valori percentuali oggetto di stima;
- l'effetto derivante dalla scelta della macro-area, con riferimento sia alla sua collocazione geografica che alla sua dimensione, sui livelli di efficienza dei differenti metodi di stima indiretti sperimentati.

Per le valutazioni sono stati calcolati gli indicatori di distorsione AARB e MARB, secondo le espressioni (20) e (22), e di variabilità ARRMSE e MRRMSE, in base alle formule (21) e (23).

5.1 Valutazioni sull'efficienza degli stimatori per piccole aree per la stima di frequenze relative riferite alle aree di censimento

Nell'ambito di questo paragrafo è stato valutato il guadagno di accuratezza delle stime riferite alle aree di censimento di centro abitato, derivante dall'impiego dei metodi di stima per piccole aree rispetto ai metodi diretti espansione e GREG. Le valutazioni hanno riguardato differenti valori percentuali oggetto di stima a parità della macro-area definita per la sperimentazione (Tavole 7-9).

[9] Come si può evincere dalla Tavola 4, le simulazioni Q2_1 e Q2_2 hanno riguardato un numero ridotto di replicazioni campionarie pari a 100 perché i tempi di elaborazione, comunque sempre consistenti, si sono rivelati particolarmente onerosi in tali situazioni, arrivando a richiedere circa 10 giorni di tempo-macchina per ciascuna simulazione.

Tavola 7 - Risultati delle simulazioni relative al test Q1_1 (macro-area riferita all'aggregazione di 2 comuni superiori ai 500mila abitanti; piccole aree date dalle aree di censimento di centro abitato). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra le aree di censimento)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,01%)	Espansione	3,87	242,60	29,61	311,83
	GREG	6,45	242,71	32,44	312,78
	SINT_UL	37,84	46,95	80,62	81,57
	SINT_AL	42,82	100,71	154,38	228,18
	EBLUP_UL	34,51	52,33	75,93	85,91
	EBLUP_AL	42,82	100,71	154,38	228,18
ateco35m_1 (0,55%)	Espansione	1,81	41,92	18,60	85,67
	GREG	3,25	42,01	16,83	86,38
	SINT_UL	29,66	30,29	189,94	190,31
	SINT_AL	22,03	29,40	100,99	107,51
	EBLUP_UL	17,34	26,32	129,99	133,85
	EBLUP_AL	17,78	27,41	84,71	89,43
titolo_6_6 (1,72%)	Espansione	1,46	29,78	11,08	93,59
	GREG	26,96	41,00	140,53	163,85
	SINT_UL	113,41	113,60	1126,60	1127,01
	SINT_AL	40,19	44,92	368,65	371,46
	EBLUP_UL	12,11	30,45	66,10	112,37
	EBLUP_AL	11,96	26,91	95,76	103,31
posprof_setteco_9 (1,93%)	Espansione	1,59	29,43	21,36	113,41
	GREG	40,46	51,84	383,53	406,79
	SINT_UL	174,39	174,49	2024,82	2025,20
	SINT_AL	41,53	46,98	761,79	778,08
	EBLUP_UL	34,25	46,34	397,54	414,61
	EBLUP_AL	16,35	28,36	241,82	254,20
luodst2m_2 (2,59%)	Espansione	1,15	21,98	10,07	48,05
	GREG	5,29	22,73	25,90	53,02
	SINT_UL	81,29	81,41	527,66	527,77
	SINT_AL	46,41	48,36	402,26	405,79
	EBLUP_UL	12,53	24,75	73,97	83,89
	EBLUP_AL	7,50	21,06	48,73	61,92

Tavola 8 - Risultati delle simulazioni relative al test Q1_2 (macro-area riferita all'aggregazione di 8 comuni tra 100mila e 500mila abitanti; piccole aree date dalle aree di censimento di centro abitato). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra le aree di censimento)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,017%)	Espansione	6,57	231,03	100,00	377,70
	GREG	19,50	233,00	129,38	376,45
	SINT_UL	50,53	60,27	142,63	152,19
	SINT_AL	52,30	121,73	320,83	381,28
	EBLUP_UL	41,42	68,37	104,56	131,12
	EBLUP_AL	52,30	121,73	320,83	381,28
ateco35m_1 (0,48%)	Espansione	1,94	46,69	20,38	77,69
	GREG	8,16	47,50	27,02	79,04
	SINT_UL	26,61	27,68	129,24	129,62
	SINT_AL	22,90	32,20	92,26	97,40
	EBLUP_UL	15,19	26,41	79,62	84,57
	EBLUP_AL	17,40	30,00	74,32	81,05
titolo_6_6 (0,42%)	Espansione	2,14	53,46	15,26	129,84
	GREG	45,28	71,20	166,48	213,04
	SINT_UL	39,40	41,69	208,66	211,20
	SINT_AL	30,74	42,48	204,64	213,00
	EBLUP_UL	32,70	46,68	63,52	90,69
	EBLUP_AL	22,71	38,17	152,38	162,03
posprof_setteco_9 (3,50%)	Espansione	0,83	18,41	5,16	37,15
	GREG	7,24	19,83	25,80	42,60
	SINT_UL	56,97	57,05	294,37	294,43
	SINT_AL	23,95	26,28	225,94	226,76
	EBLUP_UL	8,69	19,40	41,58	52,55
	EBLUP_AL	8,20	16,60	58,49	62,37
luodst2m_2 (8,93%)	Espansione	0,54	10,15	4,26	17,79
	GREG	1,57	9,96	5,99	17,70
	SINT_UL	23,41	23,48	126,60	126,61
	SINT_AL	15,26	16,33	79,55	79,93
	EBLUP_UL	3,23	9,45	18,30	23,39
	EBLUP_AL	3,62	9,26	16,87	19,97

Tavola 9 - Risultati delle simulazioni relative al test Q1_3 (macro-area riferita all'aggregazione di 25 comuni tra 10mila e 100mila abitanti; piccole aree date dalle aree di censimento di centro abitato). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra le aree di censimento)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,04%)	Espansione	4,09	180,10	46,16	304,89
	GREG	43,76	188,86	148,48	348,45
	SINT_UL	87,83	99,33	352,25	370,07
	SINT_AL	79,99	186,04	494,13	597,09
	EBLUP_UL	51,63	95,96	183,53	254,43
	EBLUP_AL	79,99	186,04	494,13	597,09
ateco35m_1 (0,97%)	Espansione	1,37	35,48	6,87	69,24
	GREG	16,84	39,65	40,39	80,18
	SINT_UL	53,54	54,21	191,05	191,51
	SINT_AL	43,68	53,08	314,07	318,36
	EBLUP_UL	8,52	31,31	32,07	60,13
	EBLUP_AL	17,28	35,31	125,28	134,47
titolo_6_6 (1,02%)	Espansione	1,32	42,11	7,96	112,27
	GREG	37,26	58,40	219,17	253,89
	SINT_UL	157,16	157,86	1276,18	1278,64
	SINT_AL	46,98	63,51	375,33	418,72
	EBLUP_UL	9,23	41,28	26,77	119,68
	EBLUP_AL	25,52	45,61	206,59	240,01
posprof_setteco_9 (2,45%)	Espansione	0,89	20,53	7,15	32,80
	GREG	8,94	22,56	24,91	41,72
	SINT_UL	50,32	50,60	229,19	229,39
	SINT_AL	22,59	27,90	147,76	151,19
	EBLUP_UL	10,23	21,41	43,99	53,33
	EBLUP_AL	11,34	20,32	75,35	81,09
luodst2m_2 (9,97%)	Espansione	0,64	10,64	3,24	16,74
	GREG	6,29	12,41	15,02	23,61
	SINT_UL	68,04	68,13	235,27	235,31
	SINT_AL	50,46	51,80	174,05	174,42
	EBLUP_UL	6,48	12,47	17,14	24,80
	EBLUP_AL	1,82	10,63	6,71	16,71

In generale, a parità di disegno campionario e di frazione di campionamento, le stime ottenute con i metodi diretti hanno una distorsione inferiore a quelle ottenute con i metodi per piccole aree presi in esame. Queste ultime, pur mostrando la presenza di una distorsione non trascurabile, evidenziano il vantaggio di un ridotto livello dell'errore quadratico medio. Il guadagno di efficienza che si ottiene con i metodi per piccole aree non è però sempre della stessa entità, ma dipende dal valore del parametro di interesse. A riguardo, emerge che per le variabili con frequenza percentuale inferiore al 2% il passaggio ai metodi indiretti porta ad una riduzione della variabilità rispetto ai metodi diretti. Per valori percentuali da stimare superiori a tale soglia invece è preferibile il ricorso ai metodi diretti rispetto agli stimatori per piccole aree in quanto, per questi ultimi, la riduzione della variabilità non è in grado di compensare l'aumento della distorsione.

Confrontando tra loro i metodi indiretti si evince che gli stimatori basati su un modello lineare ad effetti misti costruiti a livello di unità e a livello di area (EBLUP_UL ed EBLUP_AL) danno luogo ai risultati migliori. Gli stimatori sintetici (SINT_UL e SINT_AL) forniscono invece risultati meno soddisfacenti per la presenza di elevati livelli di distorsione.

Lo stimatore EBLUP_UL presenta risultati migliori in termini di variabilità per la stima di frequenze percentuali inferiori all'1%; invece gli stimatori EBLUP_AL, basati su un modello definito a livello di area, forniscono, sia in media che con riferimento ai valori massimi, risultati più vantaggiosi quando il parametro di interesse assume frequenze percentuali comprese tra l'1% e il 2%.

I risultati di efficienza osservati per la stima di frequenze inferiori all'1% (corrispondenti ad un massimo di 50-150 unità riferite alle aree di censimento), evidenziano che il metodo di stima EBLUP_UL porta ad un guadagno fino all'80%, in termini di errore quadratico medio, rispetto allo stimatore diretto di tipo espansione.

Per la stima di frequenze percentuali comprese tra l'1% e il 2% (corrispondenti a frequenze assolute comprese tra 50 e 300 unità sulle aree di censimento) lo stimatore EBLUP_AL mostra valori più bassi dell'errore quadratico medio, con guadagni rispetto allo stimatore espansione fino al 30%.

Inoltre, si è osservato che, a parità di frequenza percentuale da stimare, i livelli di errore rimangono pressoché invariati sia nel caso di una macro-area data dall'aggregazione delle aree di censimento di grandi comuni che nel caso di aggregazione di aree riferite ad un insieme di comuni di dimensioni inferiori.

I risultati appena presentati sono confermati anche nel caso dell'analisi svolta dopo aver classificato le aree di censimento di centro abitato in base alla loro dimensione demografica: tra 5mila e 10mila abitanti; tra 10mila e 12mila abitanti; tra 12mila e 15mila abitanti. Nella tavola 10 gli indicatori di distorsione e di variabilità sono presentati per la variabile "ateco35m_1" (con valori percentuali compresi tra 0,5% e 1% sui domini di interesse) separatamente per le tre classi dimensionali delle aree di censimento.

Si nota che, a prescindere dalla dimensione media dell'area di censimento, per la stima del parametro di interesse è da preferire l'adozione di uno tra i due metodi di tipo indiretto EBLUP, con un guadagno di efficienza, rispetto ai metodi di tipo diretto, compresi tra il 30% e il 40%. In particolare, il modello unit-level conduce a stime più efficienti sulle aree di censimento relative alle prime due classi, mentre per le aree più grandi il comportamento atteso degli stimatori EBLUP_UL ed EBLUP_AL è pressoché equivalente.

Tavola 10 - Risultati delle simulazioni relative al test Q1_1 (macro-area riferita all'aggregazione di 2 comuni superiori ai 500mila abitanti; piccole aree date dalle aree di censimento di centro abitato). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per la variabile "ateco35m_1" per tre classi di dimensione demografica delle aree di censimento

"ateco35m_1"	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
Aree di censimento con popolazione tra 5mila e 10mila	Espansione	2,08	48,88	18,60	85,67
	GREG	3,58	48,91	16,83	86,38
	SINT_UL	29,77	30,56	189,94	190,31
	SINT_AL	24,31	32,64	99,90	103,56
	EBLUP_UL	19,82	28,45	129,99	133,85
	EBLUP_AL	20,60	30,91	84,71	89,43
Aree di censimento con popolazione tra 10mila e 12mila	Espansione	1,68	40,74	8,55	56,32
	GREG	3,29	40,86	7,32	56,69
	SINT_UL	30,74	31,15	80,25	80,57
	SINT_AL	23,26	29,39	100,99	107,51
	EBLUP_UL	17,56	25,85	43,16	48,82
	EBLUP_AL	18,66	27,19	81,64	89,28
Aree di censimento con popolazione tra 12mila e 15mila	Espansione	1,64	36,87	10,83	54,91
	GREG	2,98	37,01	7,41	54,84
	SINT_UL	29,16	29,75	103,92	104,26
	SINT_AL	19,75	26,83	89,54	92,78
	EBLUP_UL	15,30	24,81	53,02	58,89
	EBLUP_AL	15,20	24,74	70,43	75,58

5.2 Valutazioni sull'efficienza degli stimatori per piccole aree per la stima di frequenze relative riferite ai comuni di dimensione tra 5mila e 20mila abitanti

In questo ambito le analisi hanno riguardato il confronto tra i livelli di accuratezza attesa delle stime dirette ed indirette riferite ai comuni con popolazione compresa tra 5mila e 20mila unità. Le valutazioni sono state condotte sempre per diverse frequenze percentuali da stimare e a parità della macro-area scelta per l'applicazione degli stimatori indiretti (Tavole 11-17).

I risultati delle simulazioni conducono a considerazioni sostanzialmente analoghe a quelle fatte nell'analisi precedente. Anche in questo caso, dai valori degli indicatori riportati nelle tavole 11-17, si evince che gli stimatori di tipo indiretto portano ad una riduzione della variabilità per la stima di frequenze percentuali inferiori al 2%; tra questi i guadagni maggiori in termini di efficienza si riscontrano per gli stimatori EBLUP_UL ed EBLUP_AL.

In particolare, è attesa una riduzione compresa tra il 10% e il 50% dell'errore quadratico medio dello stimatore EBLUP_UL rispetto allo stimatore espansione, nel caso di stime di frequenze percentuali inferiori all'1% (valori corrispondenti a 50-200 unità per i comuni di dimensione compresa tra 5mila e 20mila abitanti); tale diminuzione è leggermente inferiore rispetto a quanto osservato nel caso di stime riferite alle aree di censimento di centro abitato (si veda paragrafo 5.1).

Tavola 11 - Risultati delle simulazioni relative al test Q2_1 (macro-area riferita alla ripartizione dell'Italia centrale; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,14%)	Espansione	2,25	57,94	38,78	163,32
	GREG	15,45	68,88	121,52	300,27
	SINT_UL	24,63	54,41	256,00	564,55
	SINT_AL	30,28	74,39	345,83	773,25
	EBLUP_UL	7,33	43,87	41,09	140,28
	EBLUP_AL	17,66	53,21	185,00	422,76
ateco35m_1 (2,15%)	Espansione	0,74	12,83	8,44	38,39
	GREG	3,55	15,17	18,17	45,46
	SINT_UL	13,82	30,34	115,81	253,98
	SINT_AL	13,10	29,55	123,19	271,25
	EBLUP_UL	3,00	13,71	12,51	34,64
	EBLUP_AL	1,42	12,24	22,75	58,27
titolo_6_6 (1,02%)	Espansione	0,90	16,66	8,15	50,65
	GREG	3,99	18,79	15,05	48,84
	SINT_UL	8,36	18,46	108,88	238,81
	SINT_AL	6,52	16,04	40,85	90,06
	EBLUP_UL	3,12	14,50	19,85	51,25
	EBLUP_AL	3,31	12,81	21,75	51,98
posprof_setteco_9 (1,87%)	Espansione	0,65	12,06	4,10	28,02
	GREG	1,85	12,69	7,75	32,51
	SINT_UL	6,74	14,84	42,57	93,36
	SINT_AL	3,79	9,51	20,90	47,07
	EBLUP_UL	3,23	11,08	21,57	50,10
	EBLUP_AL	2,36	8,17	14,57	33,41
luodst2m_2 (22,73%)	Espansione	0,18	3,11	0,75	7,61
	GREG	0,48	3,05	2,82	12,83
	SINT_UL	6,82	14,97	136,73	299,80
	SINT_AL	4,46	9,91	119,42	261,93
	EBLUP_UL	0,65	3,16	6,69	18,29
	EBLUP_AL	0,39	3,04	7,20	17,34

Tavola 12 - Risultati delle simulazioni relative al test Q2_2 (macro-area riferita alla ripartizione dell'Italia nord-occidentale; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 *(0,157%)*	Espansione	11,26	141,21	454,00	479,09
	GREG	111,79	186,79	790,68	861,81
	SINT_UL	183,21	184,13	1518,75	1522,09
	SINT_AL	222,20	238,07	2570,01	2590,13
	EBLUP_UL	57,23	127,48	476,25	565,44
	EBLUP_AL	144,59	169,27	1618,70	1635,75
ateco35m_1 *(1,405%)*	Espansione	4,64	39,66	45,65	95,28
	GREG	29,08	50,04	116,71	147,62
	SINT_UL	95,78	95,90	492,21	492,34
	SINT_AL	97,00	98,83	1061,06	1061,85
	EBLUP_UL	21,14	43,03	81,41	99,49
	EBLUP_AL	10,98	38,01	147,52	169,77
titolo_6_6 *(0,476%)*	Espansione	6,81	56,88	92,52	148,13
	GREG	29,24	64,210	156,22	215,81
	SINT_UL	39,88	40,14	365,06	365,29
	SINT_AL	40,48	44,02	391,22	395,25
	EBLUP_UL	23,14	37,49	169,23	179,66
	EBLUP_AL	25,85	37,99	285,59	291,75
posprof_setteco_9 *(1,789%)*	Espansione	3,30	27,15	20,35	50,09
	GREG	4,43	27,36	27,20	52,30
	SINT_UL	26,96	27,06	128,23	128,26
	SINT_AL	18,08	19,80	110,44	111,17
	EBLUP_UL	11,18	21,60	69,73	74,07
	EBLUP_AL	10,48	17,52	73,55	75,72
luodst2m_2 *(30,283%)*	Espansione	0,76	5,70	3,57	23,65
	GREG	2,98	5,90	48,68	64,54
	SINT_UL	25,71	25,71	1459,34	1459,34
	SINT_AL	14,75	14,91	571,44	572,44
	EBLUP_UL	3,44	6,13	136,76	142,44
	EBLUP_AL	1,80	5,55	93,48	95,66

Tavola 13 - Risultati delle simulazioni relative al test Q2_3 (macro-area riferita alla ripartizione dell'Italia insulare; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,055%)	Espansione	3,98	178,54	55,19	356,69
	GREG	32,92	182,72	120,92	357,37
	SINT_UL	98,05	101,03	478,07	483,79
	SINT_AL	96,87	142,93	910,93	952,90
	EBLUP_UL	51,67	90,77	203,26	260,56
	EBLUP_AL	96,87	142,93	910,93	952,90
ateco35m_1 (2,9%)	Espansione	1,19	23,69	16,28	53,25
	GREG	4,46	24,18	23,96	59,03
	SINT_UL	74,58	74,70	746,50	746,65
	SINT_AL	61,35	63,17	584,13	586,88
	EBLUP_UL	4,31	22,62	30,77	54,63
	EBLUP_AL	6,08	22,90	46,01	67,21
titolo_6_6 (2,71%)	Espansione	1,14	23,77	6,27	49,07
	GREG	4,48	23,47	20,16	55,84
	SINT_UL	46,41	46,57	523,16	523,32
	SINT_AL	29,66	32,17	213,04	214,27
	EBLUP_UL	6,47	20,56	52,52	69,65
	EBLUP_AL	9,73	20,69	55,61	62,84
posprof_setteco_9 (1,05%)	Espansione	1,82	36,14	49,24	69,67
	GREG	17,05	40,13	77,45	91,87
	SINT_UL	47,22	47,46	171,97	172,13
	SINT_AL	26,87	31,68	120,36	121,56
	EBLUP_UL	23,69	35,64	83,25	89,12
	EBLUP_AL	16,26	26,69	80,65	84,03
luodst2m_2 (14,17%)	Espansione	0,51	11,52	3,31	298,51
	GREG	27,27	39,09	4073,58	5104,69
	SINT_UL	604,86	604,91	97028,98	97032,10
	SINT_AL	217,32	218,98	30348,56	30553,67
	EBLUP_UL	44,32	53,11	6930,94	7524,15
	EBLUP_AL	12,72	20,54	1868,78	1902,24

Tavola 14 - Risultati delle simulazioni relative al test Q2_4 (macro-area riferita alla regione Lazio; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 *(0,12%)*	Espansione	3,45	135,12	31,66	308,81
	GREG	55,91	149,48	347,06	482,67
	SINT_UL	120,69	124,52	1190,06	1201,12
	SINT_AL	145,11	190,23	1403,87	1524,40
	EBLUP_UL	20,98	97,62	60,58	277,07
	EBLUP_AL	115,70	161,69	1102,05	1191,68
ateco35m_1 *(2,25%)*	Espansione	1,27	28,32	5,64	67,02
	GREG	11,70	31,04	55,13	90,09
	SINT_UL	77,96	78,21	486,57	486,97
	SINT_AL	67,71	72,22	335,01	337,32
	EBLUP_UL	9,02	28,32	25,28	67,47
	EBLUP_AL	7,60	27,48	31,43	62,58
titolo_6_6 *(1,28%)*	Espansione	1,67	35,18	13,96	79,41
	GREG	13,27	37,47	43,35	94,14
	SINT_UL	53,32	53,86	289,19	289,61
	SINT_AL	35,77	43,81	179,04	184,38
	EBLUP_UL	7,36	30,30	34,51	73,64
	EBLUP_AL	19,02	32,51	117,35	126,00
posprof_setteco_9 *(1,62%)*	Espansione	1,27	28,07	7,24	50,75
	GREG	12,72	30,88	25,95	57,08
	SINT_UL	39,49	39,86	172,15	172,35
	SINT_AL	17,39	24,53	68,24	69,57
	EBLUP_UL	19,58	28,49	80,70	86,50
	EBLUP_AL	13,77	22,51	53,16	61,74
luodst2m_2 *(22,89%)*	Espansione	0,40	6,76	1,52	13,70
	GREG	2,18	6,64	5,97	15,14
	SINT_UL	25,09	25,13	129,10	129,12
	SINT_AL	16,25	17,08	80,84	81,09
	EBLUP_UL	2,78	6,74	14,56	19,57
	EBLUP_AL	1,60	6,52	7,87	13,12

Tavola 15 - Risultati delle simulazioni relative al test Q2_5 (macro-area riferita alla regione Toscana; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,15%)	Espansione	3,62	121,72	25,52	304,23
	GREG	82,68	153,85	642,75	727,52
	SINT_UL	140,31	143,55	1331,55	1341,35
	SINT_AL	165,40	203,27	2042,13	2107,42
	EBLUP_UL	47,13	108,19	322,51	438,80
	EBLUP_AL	92,44	138,34	1203,35	1259,67
ateco35m_1 (2,1%)	Espansione	1,32	28,28	7,86	60,21
	GREG	17,00	33,49	79,80	95,73
	SINT_UL	73,20	73,40	550,28	550,58
	SINT_AL	58,43	62,78	315,48	317,92
	EBLUP_UL	15,81	31,40	63,42	81,54
	EBLUP_AL	5,97	27,13	35,49	65,10
titolo_6_6 (0,89%)	Espansione	2,24	37,99	49,38	83,34
	GREG	23,24	44,50	80,42	115,69
	SINT_UL	37,71	38,17	361,03	361,56
	SINT_AL	21,88	30,22	97,87	100,18
	EBLUP_UL	20,97	31,34	137,11	145,27
	EBLUP_AL	15,87	27,29	73,12	81,24
posprof_setteco_9 (2,06%)	Espansione	1,20	24,29	6,26	48,20
	GREG	6,39	25,05	18,65	52,51
	SINT_UL	27,16	27,47	201,92	202,07
	SINT_AL	15,60	20,35	75,54	78,09
	EBLUP_UL	11,94	21,13	93,08	98,06
	EBLUP_AL	10,62	17,90	53,59	59,59
luodst2m_2 (23,9%)	Espansione	0,35	6,69	1,93	16,22
	GREG	2,21	6,59	13,86	28,14
	SINT_UL	40,82	40,84	716,88	716,90
	SINT_AL	20,80	21,43	422,99	423,52
	EBLUP_UL	3,57	7,16	34,32	41,79
	EBLUP_AL	1,75	6,67	29,54	33,43

Tavola 16 - Risultati delle simulazioni relative al test Q2_6 (macro-area riferita all'aggregazione delle regioni Lazio e Toscana; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,14%)	Espansione	3,65	127,35	82,67	304,23
	GREG	72,20	151,04	538,39	628,07
	SINT_UL	133,27	135,15	1325,39	1330,83
	SINT_AL	152,99	178,21	1570,70	1631,45
	EBLUP_UL	34,18	102,05	211,69	342,88
	EBLUP_AL	84,21	123,11	842,55	891,95
ateco35m_1 (2,15%)	Espansione	1,38	28,36	7,86	68,13
	GREG	14,98	32,48	74,97	102,37
	SINT_UL	74,95	75,07	562,82	562,99
	SINT_AL	72,39	74,63	611,37	614,05
	EBLUP_UL	13,17	30,01	50,62	75,21
	EBLUP_AL	5,98	27,37	65,61	90,19
titolo_6_6 (1,04%)	Espansione	2,19	36,64	49,38	83,34
	GREG	17,21	40,36	64,68	107,47
	SINT_UL	44,71	45,08	568,89	569,19
	SINT_AL	30,27	35,29	180,01	183,02
	EBLUP_UL	12,40	31,98	90,91	111,86
	EBLUP_AL	14,80	27,95	100,83	107,71
posprof_setteco_9 (1,89%)	Espansione	1,43	25,97	8,06	51,07
	GREG	8,73	27,39	25,23	55,67
	SINT_UL	34,64	34,85	206,01	206,10
	SINT_AL	18,12	21,33	109,60	111,63
	EBLUP_UL	15,16	24,70	89,60	94,95
	EBLUP_AL	11,35	18,24	77,45	81,15
luodst2m_2 (23,36%)	Espansione	0,37	6,72	1,93	16,22
	GREG	2,36	6,66	14,72	27,84
	SINT_UL	34,66	34,67	671,90	671,91
	SINT_AL	22,21	22,52	589,71	589,89
	EBLUP_UL	3,39	7,06	34,58	41,54
	EBLUP_AL	1,68	6,61	35,96	39,11

Tavola 17 - Risultati delle simulazioni relative al test Q2_7 (macro-area riferita all'aggregazione delle regioni Lazio, Toscana e Umbria; piccole aree date dai comuni tra 5mila e 20mila abitanti). Indicatori AARB, ARRMSE, MARB e MRRMSE calcolati per le variabili prese in esame

VARIABILI (media ponderata tra i comuni 5mila-20mila)	Stimatori	AARB	ARRMSE	MARB (max ARB)	MRRMSE (max RRMSE)
posprof_setteco_3 (0,14%)	Espansione	4,01	126,72	79,07	304,23
	GREG	73,07	151,09	544,89	631,33
	SINT_UL	124,69	126,50	1260,98	1266,16
	SINT_AL	147,65	171,65	1556,37	1605,33
	EBLUP_UL	35,29	100,59	210,33	337,19
	EBLUP_AL	82,10	120,19	770,16	822,10
ateco35m_1 (2,18%)	Espansione	1,26	27,94	10,95	71,02
	GREG	16,08	32,67	75,85	100,29
	SINT_UL	71,74	71,86	566,67	566,81
	SINT_AL	68,39	70,59	617,60	620,44
	EBLUP_UL	13,73	29,81	52,52	73,12
	EBLUP_AL	5,69	26,88	80,83	102,92
titolo_6_6 (1,05%)	Espansione	2,06	36,11	49,38	85,47
	GREG	17,71	40,08	68,46	109,70
	SINT_UL	42,05	42,44	563,46	563,75
	SINT_AL	28,85	33,46	175,29	177,87
	EBLUP_UL	13,49	31,51	98,54	116,85
	EBLUP_AL	14,77	27,07	105,02	112,39
posprof_setteco_9 (1,89%)	Espansione	1,30	25,97	9,36	54,71
	GREG	9,01	27,47	26,41	60,32
	SINT_UL	33,77	33,94	208,65	208,72
	SINT_AL	18,22	21,29	113,51	115,38
	EBLUP_UL	15,46	24,63	78,76	101,63
	EBLUP_AL	11,14	18,13	96,20	82,17
luodst2m_2 (23,01%)	Espansione	0,37	6,77	1,93	16,22
	GREG	2,26	6,66	14,09	27,33
	SINT_UL	33,84	33,85	664,02	664,03
	SINT_AL	21,77	22,08	585,92	586,11
	EBLUP_UL	3,23	7,00	33,78	40,77
	EBLUP_AL	1,64	6,63	36,42	39,58

5.3 Valutazioni sull'effetto della definizione della macro-area sui livelli di efficienza degli stimatori per piccole aree

5.3.1 Studio dell'effetto dovuto alla collocazione geografica della macro-area

Per valutare l'effetto della scelta, in termini di collocazione geografica, della macro-area di riferimento per l'applicazione dei metodi per piccole aree sui livelli di efficienza delle stime, sono stati confrontati i risultati descritti nelle tavole 11-13. Queste fanno riferimento a macro-aree definite dall'unione dei comuni di dimensione compresa tra 5mila e 20mila abitanti appartenenti rispettivamente alle ripartizioni dell'Italia centrale, nord-occidentale e insulare.

A riguardo è stato verificato che rimane inalterato l'esito del confronto tra i metodi di stima diretti ed indiretti; si nota però un effetto geografico sul livello atteso di efficienza delle stime, in quanto i valori dell'errore quadratico medio sono diversi per le tre macro-aree ripartizionali prese in considerazione per valori simili della frequenza percentuale da stimare. Ciò fa ritenere differente il potere esplicativo delle covariate scelte per spiegare la variabilità dei fenomeni oggetto di studio nelle macro-aree prese in esame; questo dovrebbe spingere a diversificare la scelta delle variabili ausiliarie tra le macro-aree per massimizzare il guadagno derivante dall'impiego dei metodi per piccole aree.

5.3.2 Studio dell'effetto dovuto alla dimensione della macro-area

Successivamente, sulla base dei risultati presentati nelle tavole 14-17, si è cercato di valutare l'effetto della dimensione della macro-area, in termini di numero dei piccoli domini che la compongono, sui livelli di efficienza attesa dei metodi di stima sperimentati. In particolare, è stato studiato il comportamento dell'errore quadratico medio al crescere della macro-area di riferimento: i test Q2_4 e Q2_5 si riferiscono a due regioni (rispettivamente Lazio e Toscana) prese singolarmente; il test Q2_6 considera le due regioni nella stessa macro-area; il test Q2_7 si riferisce ad una macro-area più ampia che considera, oltre a Lazio e Toscana, anche la regione Umbria. Le sperimentazioni hanno evidenziato che, indipendentemente dal metodo di stima adottato, per le macro-aree coincidenti con le regioni Lazio e Toscana prese singolarmente, i livelli di efficienza sono differenti; invece, nei due casi di aggregazione di regioni gli errori tendono a stabilizzarsi su livelli di efficienza migliori. Si è inoltre osservato che, in alcuni casi, la definizione della macro-area può influire sull'indicazione del metodo di stima più efficiente.

5.4 Conclusioni

In sintesi, gli stimatori per piccole aree possono essere considerati alternative soddisfacenti rispetto agli stimatori diretti per ottenere stime più efficienti di frequenze relative particolarmente piccole e riferite a domini dimensione compresa tra 5mila e 20mila unità (le aree di censimento di centro abitato o i piccoli comuni). La scelta del metodo e i livelli attesi di efficienza potrebbero, inoltre, essere condizionati dalla scelta della macro-area di riferimento per l'applicazione dei metodi sperimentati in questo studio.

6. Considerazioni finali

Nell'ottica di rilevare al censimento della popolazione e delle abitazioni del 2011 alcune informazioni relative a variabili non strettamente demografiche tramite campioni di famiglie, sono stati studiati diversi approcci metodologici, sia in termini di disegno che di stimatore.

In base ai risultati di un articolato studio sperimentale, è stata definita una strategia campionaria basata sull'adozione di un disegno casuale semplice da lista anagrafica per la selezione di campioni di famiglie e sull'impiego dello stimatore di ponderazione vincolata per la produzione dei risultati finali.

Al fine di migliorare l'efficienza delle stime è stata studiata la possibilità di utilizzare alcuni metodi indiretti e i risultati emersi da alcune sperimentazioni mostrano che l'impiego degli stimatori per piccole aree è appropriato per la stima di frequenze relative molto piccole e riferite a domini di elevato dettaglio territoriale. Il ricorso a tali metodi comporta infatti una significativa riduzione della variabilità campionaria, ma richiede giuste attenzioni per l'introduzione di una componente aggiuntiva di distorsione.

Dall'analisi dei livelli di efficienza attesa degli stimatori per piccole aree presi in esame, è risultato che la scelta del metodo più accurato dipende fortemente dal valore della variabile oggetto di stima; saranno quindi necessari ulteriori approfondimenti con lo scopo di individuare opportune soglie (in termini di frequenza relativa o assoluta) utili a decidere a priori tra i differenti approcci proponibili. Infatti, è stato evidenziato che, oltre un determinato valore della frequenza da stimare, non si ha nessun vantaggio dall'impiego di metodi per piccole aree analizzati in questo studio rispetto ai metodi diretti; inoltre, tra gli stimatori esaminati, esiste una soglia utile a discriminare tra la scelta verso un metodo *unit-level* rispetto ad un metodo *area-level*.

Le considerazioni a cui si è giunti fino ad ora forniscono alcune prime indicazioni riguardanti da un lato il vantaggio offerto dall'impiego degli stimatori per piccole aree, nell'ambito della strategia del censimento del 2011, utili a produrre informazioni di qualità riferite ai domini territoriali subcomunali (le aree di censimento di centro abitato), dall'altro la possibilità di introdurre, in future occasioni censuarie, le tecniche di campionamento anche nei comuni con popolazione compresa tra 5mila e 20mila unità. Il miglioramento dell'efficienza attesa con l'impiego di tali metodi richiede il proseguimento degli studi, con riferimento sia alle scelte metodologiche che alla fase applicativa.

Da un punto di vista strettamente metodologico, si prevede in futuro di concentrare l'attenzione sulle seguenti proposte di modelli alternativi a quelli su cui si basano gli stimatori indiretti presi in esame in questo lavoro:

- Modelli di tipo logistico al posto di quelli lineari sperimentati in questo studio. Tali modelli sono più idonei per la stima di variabili di tipo dicotomico (presenza/assenza della caratteristica); inoltre forniscono la garanzia di stime non negative;
- Modelli che tengano conto di una struttura di autocorrelazione spaziale tra le osservazioni considerando anche la distanza tra le aree di interesse. Ciò presuppone l'ipotesi verosimile che le osservazioni rilevate nei domini di interesse siano legate maggiormente a quelle rilevate in domini geograficamente vicini piuttosto che a quelle osservate in domini lontani;
- Modelli che stimano le variabili di interesse in modo congiunto e non più separatamente, sfruttando la correlazione esistente tra queste.

Un'altra importante scelta metodologica è relativa alla caratterizzazione dell'insieme delle covariate e della macro-area di riferimento per l'applicazione del metodo di stima. Andrebbe infatti individuato, per ogni variabile oggetto di stima, un insieme idoneo di covariate, selezionate tra quelle maggiormente correlate con la variabile incognita. Nello stesso modo si potrebbero costruire macro-aree diverse per ogni variabile di interesse. Tali soluzioni però, potrebbero porre problemi di operatività per il contesto censuario, per la presenza di una molteplicità di parametri da stimare.

Per quanto concerne la scelta della macro-area, una proposta percorribile potrebbe essere quella di differenziare il disegno della macro-area, sia per la collocazione geografica che per il numero di domini che la compongono, in modo tale da massimizzare il contributo informativo delle covariate considerate per i diversi contesti territoriali.

Inoltre, indipendentemente dalla scelta della metodologia di stima per piccole aree, si dovranno affrontare alcune non trascurabili problematiche legate all'uso di tali metodi nel contesto censuario:

- complessità computazionale;
- coerenza tra stime ottenute con metodi diretti ed indiretti presenti in una stessa tavola;
- identificazione degli zero non-strutturali (dovuti all'aleatorietà del risultato campionario).

Un ultimo punto che andrà esaminato riguarda il processo di stima di frequenze piccole riferite a domini territoriali di dimensione superiore (per esempio, i comuni con popolazione compresa tra 20mila e 50mila abitanti) e le sue implicazioni. L'impiego dei metodi indiretti per stime su domini più grandi richiede la determinazione di un modello diverso da quello impiegato per i domini più piccoli e il soddisfacimento della coerenza tra stime riferite ad ambiti territoriali differenti. Infatti, in fase di definizione dello stimatore, si dovrà tener conto della relazione gerarchica esistente tra i livelli territoriali di riferimento dei domini di stima; in particolare, si dovrà decidere se procedere vincolando le stime calcolate ai livelli territoriali superiori a quelle riferite ai livelli inferiori (per esempio, per le aree di censimento di centro abitato) o viceversa.

Riferimenti bibliografici

Battese G.E., R.M. Harter and W.A. Fuller. 1988. An error-components model for prediction of county crops using survey and satellite data. *Journal American Statistical Association*, vol. 83: 28-36.

Bianchi G., F. Di Pede, A. Reale e S. Talice. 2010. Aree di censimento, nuove suddivisioni subcomunali per la raccolta campionaria di informazioni aggiuntive durante il prossimo censimento della popolazione: applicazione nella regione Marche. Relazione presentata alla XXXI Conferenza Italiana di Scienze Regionali, Aosta 20-22 Settembre.

Borrelli F., G. Carbonetti e L. De Felici. 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., G. Carbonetti, L. De Felici e F. Solari. 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., G. Carbonetti, L. De Felici, E. Fiorello e M. Marrone. 2011. La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento. Istat Working Papers, Collana Scientifica dell'Istituto Nazionale di Statistica. 2/2011. Roma. http://www.istat.it/it/files/2011/04/Istat_Working_Papers_2_2011.pdf

Carbonetti G. e C. De Vitiis. 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.

Carbonetti G. e M. Fortini. 2008. 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/fileadmin/DAM/stats/documents/ece/ces/ge.41/2008/4.e.pdf

Cicchitelli G., A. Herzel e G. E. Montanari. 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.

Cressie N. 1992. REML Estimation in Empirical Bayes Smoothing of Census Undercount. *Survey Methodology*, vol. 18: 75-94.

D'Alò M., L. Di Consiglio, S. Falorsi e F. Solari. 2008. Small area estimation methods for socio-economic indicators in household surveys. *Rivista Internazionale di Scienze Sociali*, n. 4: 419-442.

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

EURAREA Consortium, 2004. PROJECT REFERENCE VOLUME, vol. 1

Fay R.E. and R.A. Herriot. 1979. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the american statistical association*. 74 269-277.

Rao J.N.K. 2003 Small area estimation, John Wiley & Sons, Hoboken, New Jersey.

Ghosh M. and J.N.K. Rao. 1994. Small Area Estimation: an appraisal. *Statistical Science*, vol. 9, n. 1: 55-76.

Särndal C.E. 1984. Design-Consistent Versus Model-Dependent Estimators for Small Domains. *Journal of the american statistical association*, vol. 79: 624-631.

Särndal C.E., B. Swensson and J. Wretman. 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

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.