XXXII CONFERENZA ITALIANA DI SCIENZE REGIONALI

# METODOLOGIE DI STIMA PER PICCOLE AREE BASATE SU AUTOCORRELAZIONE SPAZIALE APPLICABILI A VARIABILI DI CENSIMENTO

Francesco BORRELLI[1], Giancarlo CARBONETTI[2], Epifania FIORELLO[3], Fabrizio SOLARI[4]

## SOMMARIO

Per la realizzazione del 15° Censimento generale della popolazione e delle abitazioni sono previste innovazioni di metodi, tecniche e organizzazione con gli obiettivi di semplificare l'impatto organizzativo sulle amministrazioni comunali, ampliare l'uso dei dati amministrativi, recuperare tempestività nella diffusione dei dati definitivi, ridurre il fastidio statistico sui rispondenti.

Tra le novità della strategia c'è l'impiego delle tecniche di campionamento per la rilevazione, nei comuni più grandi, di alcune delle principali caratteristiche di tipo socio-economico della popolazione. In particolare, si produrranno stime con riferimento al livello territoriale minimo coincidente con le aree di censimento definite in modo opportuno dall'Istat. Questa soluzione potrebbe però non garantire la continuità dei risultati censuari per tutti i domini sub-comunali non pianificati dal disegno di campionamento ma considerati dal piano di diffusione delle precedenti occasioni censuarie.

E' stata quindi proposta un metodologia per produrre stime riferite a tali domini con elevati livelli di accuratezza, che prevede l'utilizzo dei metodi di stima per piccole aree e tiene conto del rispetto della proprietà di coerenza tra stime riferite a livelli territoriali differenti. In questo lavoro, è stato dato maggior rilievo ad un metodo di stima che tiene conto dell'informazione territoriale, con particolare riferimento alla presenza di un effetto dovuto all'autocorrelazione spaziale tra le unità statistiche.

[1] Istituto Nazionale di Statistica, Via Adolfo Ravà 150, 00142, Roma, e-mail: borrelli@istat.it
[2] Istituto Nazionale di Statistica, Via Adolfo Ravà 150, 00142, Roma, e-mail: carbonet@istat.it
[3] Istituto Nazionale di Statistica, Via Adolfo Ravà 150, 00142, Roma, e-mail: fiorello@istat.it
[4] Istituto Nazionale di Statistica, Via Torino 6, 00184, Roma, e-mail: solari@istat.it

La metodologia è stata sperimentata, per la stima del *"numero di occupati"* e del *"numero di persone in cerca di occupazione"* in alcune unità territoriali sub-comunali del comune di Milano; i risultati sono infine confrontati con quelli ottenuti in un precedente studio che non teneva conto dell'informazione territoriale nell'approccio di stima.

## 1 INTRODUZIONE[5]

I dati raccolti in occasione dei censimenti costituiscono un patrimonio di conoscenza del territorio unico per l'elevato dettaglio informativo offerto e vengono utilizzati a tutti i livelli di governo, dalle imprese e dalle associazioni di categoria, per programmare e per pianificare attività e progetti, per gestire in modo efficiente l'offerta dei servizi di cui beneficiano i cittadini e per verificare i risultati delle politiche e degli interventi sul territorio.
Il censimento svolge inoltre una importante funzione per la conoscenza storica di un paese. Le informazioni raccolte, infatti, oltre a fornire una fotografia della situazione attuale, consentono valutazioni ed interpretazioni che si arricchiscono in una prospettiva di mutamento sociale, se confrontate con i risultati dei censimenti passati. A riguardo, la rilevazione censuaria, oltre alla caratteristica di universalità, deve permettere anche la confrontabilità, almeno per gli aspetti essenziali oggetto di osservazione.
Per l'appuntamento del 2011, l'Istat ha progettato il censimento della popolazione e delle abitazioni proponendo alcune soluzioni innovative che faranno abbandonare l'approccio tradizionale, in favore di un censimento basato sugli archivi anagrafici comunali (Crescenzi *et al.*, 2009). Una delle maggiori innovazioni prevede, nei comuni più grandi, la rilevazione esaustiva di tutte le principali caratteristiche demografiche e familiari e l'impiego delle tecniche di campionamento per raccogliere dati su alcune delle caratteristiche socio-economiche, solo su campioni di famiglie, attraverso l'uso contemporaneo di questionari in forma ridotta (*short*) e in forma estesa (*long*).
La strategia campionaria per il censimento permetterà da un lato di ridurre il disturbo statistico per le famiglie e il carico di lavoro dei soggetti coinvolti dalle operazioni sul campo, e dall'altro di migliorare la qualità e la tempestività dei risultati, per la minor mole di dati da sottoporre ad elaborazione statistica (Cocchi, 2007).
Con l'impiego delle tecniche di campionamento non si potrà però garantire la stessa ricchezza di risultati ottenibile nel caso di conduzione delle operazioni censuarie secondo un approccio di tipo tradizionale. Infatti, poiché le stime prodotte saranno riferite, al massimo dettaglio territoriale, solo a particolari aggregazioni di sezioni di censimento (le *aree di censimento di centro abitato*), si rischia di perdere il dettaglio informativo delle informazioni che saranno rilevate solo su campioni di famiglie, per tutti i domini sub-comunali non considerati in fase

[5] Il lavoro è frutto della collaborazione degli autori: F. Borrelli ha curato i paragrafi 3 e 7.1; G. Carbonetti i paragrafi 1, 2, 4, 8 e 9; E. Fiorello i paragrafi 6 e 7.2; F. Solari il paragrafo 5 .

di disegno del campione. In particolare, potrebbe non essere più possibile, diversamente da quanto avvenuto a conclusione delle passate tornate censuarie, diffondere dati di massimo dettaglio per ambiti territoriali definiti da suddivisioni del territorio comunale stabilite a livello locale per motivi amministrativi e/o funzionali.

Nell'ambito di questo lavoro si prosegue uno studio già avviato (Carbonetti e Fiorello, 2010) riguardante la possibilità di definire metodologie statistiche utili a calcolare stime affidabili relative ad alcune variabili di tipo socio-economico riferite a domini sub-comunali, incluse le unità territoriali minime coincidenti con le sezioni di censimento di centro, per i quali non sono stati pianificati campioni rappresentativi. L'approccio seguito in questo contesto si basa sull'impiego dei metodi di stima per piccole aree che tengono conto della componente spaziale delle unità in esame. La bontà del metodo è stata valutata in modo sperimentale e i risultati sono stati confrontati con quelli attesi nel caso di impiego di stimatori diretti e con quelli relativi all'adozione di stimatori che non tengono in considerazione l'informazione territoriale disponibile.

Dopo la descrizione della strategia campionaria che sarà impiegata nel contesto censuario (paragrafo 2), si illustrano alcuni elementi relativi all'impatto che il ricorso alle tecniche di campionamento pone sull'offerta di dati censuari sub-comunali (paragrafo 3).

Nei successivi paragrafi si tracciano le linee dell'approccio metodologico proposto: dapprima si analizza la proposta avanzata per produrre stime di dati riferiti ad ambiti territoriali sub-comunali con livelli di accuratezza accettabili (paragrafo 4); in seguito, si fa una trattazione metodologica dei modelli alla base degli stimatori per piccole aree oggetto di confronto nello studio (paragrafo 5).

Nel paragrafo 6 si descrive l'ambito dello studio sperimentale e nel paragrafo 7 si analizzano i risultati delle sperimentazioni svolte sull'impiego dello stimatore spaziale e il confronto con precedenti risultati che non hanno previsto l'uso dell'informazione territoriale.

Infine, nel paragrafo 8, sono riportate alcune valutazioni sull'impiego dello stimatore spaziale e, nel paragrafo 9, sono contenute le considerazioni finali del lavoro.

## 2 LA STRATEGIA DI CAMPIONAMENTO TRAMITE SHORT/LONG FORM AL CENSIMENTO DELLA POPOLAZIONE E DELLE ABITAZIONI DEL 2011

La rilevazione censuaria del 2011 sarà condotta secondo modalità innovative, per l'integrazione e l'utilizzo di dati di fonte amministrativa e l'adozione delle tecniche di campionamento. Tali soluzioni (Carbonetti e Fortini, 2008a; Carbonetti *et al.*, 2008b), sono state ampiamente studiate tenendo conto sia delle esperienze estere (Abbatini *et al.*, 2007) che delle opportunità offerte dal contesto nazionale.

Una delle innovazioni metodologiche introdotte per la realizzazione del prossimo censimento riguarda proprio l'adozione di una strategia campionaria tramite l'uso di questionari nelle

versioni *short* e *long*. Il campionamento interesserà solo i comuni capoluoghi di provincia e i comuni con almeno 20mila abitanti. In tali comuni si procederà alla somministrazione, solo ad un campione di famiglie, di un questionario esteso (*long*), contenente tutte le variabili solitamente osservate al censimento; un questionario in forma breve (*short*) per rilevare i dati demografici e familiari e poche informazioni di natura socio-economica, sarà sottoposto alla restante parte di popolazione. In tutti gli altri comuni, non coinvolti dalla strategia campionaria, si inoltrerà il questionario *long* a tutte le famiglie residenti, cosicché saranno rilevati in modo esaustivo tutti i dati censuari.

Nei comuni sottoposti a campionamento si attendono importanti vantaggi: per le famiglie residenti, che subiranno un minor disturbo statistico, in quanto la maggior parte riceverà il questionario breve che comporta un minor peso nella compilazione; per i comuni stessi, che vedranno ridotto il carico di lavoro sia nella fase di raccolta che nelle operazioni di revisione.

L'adozione di una strategia campionaria comporta però l'introduzione dell'errore di campionamento che può essere visto come il prezzo da pagare per l'uso dei campioni al censimento. Proprio per valutare l'errore campionario atteso e per la scelta della strategia campionaria praticabile nel contesto censuario in Italia (Carbonetti *et al.*, 2008b), è stato condotto un complesso studio sperimentale, su dati del censimento del 2001 (Borrelli *et al.*, 2011) per stabilire quali soluzioni proporre in termini di disegno di campionamento, stimatore, dominio minimo di stima, tasso di campionamento.

La valutazione dei risultati delle sperimentazioni ha portato a decidere in favore di una strategia campionaria, per il censimento della popolazione del 2011, che si baserà su:

1) selezione di campioni di famiglie secondo un *disegno di campionamento semplice* da lista anagrafica;
2) adozione di *stimatori di ponderazione vincolata* (Deville e Särndal, 1992) che garantiscono la coerenza tra dati stimati e dati osservati su tutta la popolazione;
3) disegno delle aree di censimento di centro abitato[6] con dimensione intorno a *15.000 unità* e indicazione a non procedere con il campionamento nelle aree troppo piccole;
4) scelta del *tasso di campionamento pari al 33%* che permette di ottenere stime con accuratezza accettabile fino al livello delle aree di censimento.

Questa strategia è stata supportata anche dai risultati di successivi studi effettuati per valutare la qualità attesa di tavole di diffusione dei risultati del censimento su cui impatterà il campionamento (Borrelli *et al.*, 2009; Carbonetti e Verrascina, 2009); i risultati hanno messo in evidenza che con l'impiego del tasso di campionamento del 33% si garantiscono elevati livelli di accuratezza anche per tavole statistiche riferite a contesti sub-comunali.

[6] Le *aree di censimento* (Astorri *et al.*, 2007) sono state definite dall'Istat tramite l'aggregazione di sezioni di censimento contigue e con limiti geografici che tengono conto dei limiti delle suddivisioni (meno fini) stabilite dai comuni a fini amministrativi o funzionali (suddivisioni amministrative dei comuni con almeno 250mila abitanti e delle unità non più piccole di 30mila).

I livelli di efficienza attesi delle stime ottenibili in base alla strategia campionaria decisa sono stati misurati nell'ambito del prima citato studio sperimentale. Nella Tabella 1 sono presentati i livelli di efficienza attesa (misurati tramite il coefficiente di variazione[7]) delle stime di frequenze assolute riferite a variabili rilevate a campione tramite questionario *long*, nel caso di una strategia campionaria basata su un disegno casuale semplice per estrarre campioni di famiglie dalla lista anagrafica (con un tasso di campionamento del 33%) e l'utilizzo di stimatori di ponderazione vincolata[8].

*Tabella 1* - Coefficiente di variazione CV minimo, mediano e massimo delle stime di frequenze assolute riferite alle aree di censimento. Disegno casuale semplice di famiglie con frazione sondata pari al 33%.

| Classi di frequenza assoluta [a)] | Disegno casuale semplice - f.s.=33% | | |
|---|---|---|---|
| | CV minimo | **CV mediano** | CV massimo |
| <10 | 53,1 | **66,5** | 95,8 |
| 10 ├30 | 31,3 | **33,8** | 38,5 |
| 30 ├50 | 20,8 | **23,4** | 25,6 |
| 50 ├100 | 15,7 | **17,4** | 19,1 |
| 100 ├250 | 9,4 | **11,4** | 12,8 |
| 250 ├500 | 6,6 | **7,5** | 8,1 |
| 500 ├1.000 | 4,9 | **5,3** | 5,9 |
| 1.000 ├2.500 | 2,7 | **3,3** | 3,9 |
| 2.500 ├5.000 | 1,5 | **2,0** | 2,5 |
| 5.000 ├10.000 | 0,8 | **1,3** | 1,9 |

a) Poiché in questa tabella le stime sono riferite a domini non superiori a 15mila abitanti, le frequenze assolute oggetto di stima non sono mai risultate superiori a 10mila unità.

In particolare, dall'analisi dei dati della Tabella 1 si evidenzia che: per la stima di frequenze assolute pari a 1.000 unità il CV atteso è vicino al 4%; per la stima di frequenze di 100 unità il CV atteso è intorno al 13%; per la stima di frequenze di 10 unità il CV atteso è di circa il 40%.

[7] Il coefficiente di variazione CV misura l'errore che mediamente si commette con la stima campionaria. Il valore di CV, riferito alla stima della generica frequenza assoluta T, consente di determinare la quantità $\Delta_T = 1{,}96 \cdot T \cdot CV/100$ che rappresenta l'errore assoluto massimo a cui è mediamente esposta la stima. In base alla teoria dei campioni, infatti, sotto valide ipotesi di normalità, il vero valore (incognito) di T oggetto di stima sarà compreso nell'intervallo di confidenza $\{(\hat{T} - \Delta_T); (\hat{T} + \Delta_T)\}$ con una probabilità pari a 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 .

[8] Per la fase di calibrazione è stato impiegato un sistema di vincoli riferiti alle seguenti strutture di popolazione (40 totali noti): distribuzione della popolazione per sesso e per 16 classi di età; distribuzione della popolazione per sesso e per 4 modalità di stato civile.

# 3 L'IMPATTO DEL CAMPIONAMENTO SULLA PRODUZIONE DI DATI DEL CENSIMENTO RIFERITI A DOMINI SUB-COMUNALI

L'introduzione delle tecniche di campionamento in ambito censuario, nei comuni più grandi, riducendo la mole di informazioni raccolte, avrà un impatto sul piano di diffusione dei risultati, in particolare sulla diffusione dei risultati per tutti i domini di elevato dettaglio territoriale.
Come evidenziato nel paragrafo 2, per i comuni coinvolti dalla strategia di campionamento, solo i dati sulle caratteristiche individuali e familiari e per alcune delle principali variabili socio-demografiche continueranno ad essere garantiti per tutti i livelli territoriali. Le informazioni deducibili dalle domande non incluse nel questionario in forma breve, saranno invece oggetto di stima fino al livello minimo dell'area di censimento, con i livelli di accuratezza attesa riportati nella Tabella 1.
La nuova strategia potrebbe quindi avere un impatto non favorevole, nei comuni a campione, sulla produzione dei dati censuari riferiti ai livelli territoriali sub-comunali differenti da quello delle aree di censimento; invece, per i comuni non coinvolti dal campionamento si potrà proseguire la diffusione dei dati in maniera completa fino al livello della sezione di censimento.

Una importante implicazione del campionamento è dunque legata all'opportunità di produrre stime accurate riferite a tutti quei domini sub-comunali che non sono stati pianificati dal disegno di indagine; questo tema è molto rilevante di fronte alla continua e crescente richiesta di dati di massimo dettaglio territoriale.
Al fine di continuare ad offrire tutti i risultati censuari diffusi in occasione del censimento del 2001, in questo lavoro verranno esposti avanzamenti nello sviluppo di metodologie di stima utili a garantire le informazioni riferite a tutti i domini territoriali sub-comunali che con la nuova strategia censuaria si rischierebbe di non poter più offrire.

# 4 UNA PROPOSTA METODOLOGICA PER STIMA DI DATI RIFERITI A DOMINI SUB-COMUNALI NON PIANIFICATI DAL DISEGNO D'INDAGINE

## *4.1 La disaggregazione dell'informazione statistica censuaria*

In base al disegno di campionamento scelto, si definiscono campioni di famiglie (residenti) rappresentativi per i domini sub-comunali coincidenti con le aree di censimento di centro abitato utili a produrre, su tali domini, stime per tutte le variabili oggetto di rilevazione solo tramite questionario *long*. Per soddisfare la domanda di informazione censuaria riferita ad

ambiti sub-comunali differenti dalle aree di censimento non previsti dal disegno campionario, si dovrà ricorrere ad opportune metodologie statistiche; tali stime dovranno essere da un lato coerenti con le stime riferite ad ambiti territoriali superiori e dall'altro soddisfare prefissati livelli di precisione.
La soluzione proposta si basa sulla possibilità di "disaggregare" l'informazione relativa ad una data area di censimento, in dati riferiti alle sezioni di censimento che la compongono. Per tale obiettivo si propone l'impiego dei metodi di stima per piccole aree in cui le "piccole aree" coincidono proprio con le sezioni di censimento.
Le stime per sezione ottenute con l'applicazione di tali metodi indiretti dovranno però riprodurre in modo "esatto", in caso di riaggregazione delle sezioni, l'informazione riferita all'area di censimento che le contiene. Questo approccio permette di passare da stime aggregate calcolate con riferimento alle aree di censimento a stime disaggregate riferite alle sezioni di censimento secondo un criterio di coerenza.

### *4.2 I metodi di stima per piccole aree*

L'impiego dei metodi di stima per piccole aree, rispetto ai metodi di stima diretti, permette di aumentare l'accuratezza delle stime finali, utilizzando non solo i valori della variabile d'interesse osservati sulla piccola area, ma anche sulle unità campionarie di un'area molto ampia (definita *macro-area*) che include la piccola area. Infatti, il riferimento alla macro-area permette di incrementare la numerosità campionaria effettiva su cui sono calcolate le stime.
Il processo inferenziale alla base degli stimatori per piccole aree è generalmente relativo ad un modello più o meno esplicito che esprime il legame tra le osservazioni relative alle piccole aree appartenenti alla macro-area, sfruttando la conoscenza di informazioni ausiliarie (covariate), correlate alla variabile di interesse.
L'accuratezza delle stime prodotte può inoltre aumentare nel caso di impiego di modelli che tengano conto di una struttura di autocorrelazione spaziale tra le osservazioni considerando anche la distanza tra i domini di interesse. Questo approccio presuppone l'ipotesi verosimile che le osservazioni rilevate in uno specifico dominio siano legate a quelle rilevate nei domini geograficamente vicini.
I metodi di stima per piccole aree, se da un lato introducono una componente distorsiva legata alla validità del modello ipotizzato, dall'altro consentono una notevole riduzione della variabilità delle stime prodotte rispetto a quella ottenibile con i metodi di tipo diretto, specialmente nella stima di fenomeni rari o riferiti a domini molto piccoli.
In questo lavoro si confronterà il comportamento atteso di stimatori sintetici e stimatori EBLUP (Empirical Best Linear Unbiased Predictor), basati su un modello lineare ad effetti misti, e di stimatori che considerano una componente correlata spazialmente.

La differenza tra gli stimatori sintetici e gli stimatori EBLUP è che nei primi si tiene conto della sola componente fissa del modello mentre nei secondi si introduce una componente casuale di area allo scopo di tener conto di una fonte di variabilità specifica di area. Nel modello spaziale la struttura di autocorrelazione tra le unità è basata su una opportuna distanza tra i centroidi delle unità, definita in base alle loro coordinate geografiche.

Nella specificazione del modello alla base di tali stimatori si assume che gli effetti casuali di area siano normalmente distribuiti, indipendenti ed omoschedastici.

### *4.3 La condizione di coerenza delle stime riferite a differenti livelli territoriali*

Come descritto nel paragrafo 4.1, il metodo proposto dovrà essere in grado di produrre stime riferite a domini territoriali di massimo dettaglio che, in caso di aggregazione in una delle aree di censimento pianificate dal disegno campionario, riproducano la stima precedentemente calcolata per tale ambito (Carbonetti e Fiorello, 2010).

Indicando con:

A la generica area di censimento;

$s_j$ la generica sezione di censimento contenuta in A, per cui $s_j \subset A \quad , \quad j = 1,...,J_A$ ;

$\hat{Y}_{sc}(A)$ la stima *calibrata* riferita alla variabile Y sulla generica area di censimento A;

$\hat{Y}_{sg}(s_j)$ la stima ottenuta con un metodo qualsiasi e riferita alla variabile Y sulla generica sezione di censimento $s_j$;

le stime calcolate per le sezioni di censimento saranno "coerenti" con quelle definite per le aree di censimento se è soddisfatta la seguente **condizione di coerenza**:

$$\hat{Y}_{sc}(A) = \sum_{s_j \subset A} \hat{Y}_{sg}(s_j) \qquad (4.1)$$

cioè, se si verifica che la somma delle stime sulle sezioni di censimento di una generica area di censimento, ottenute tramite l'impiego di un generico stimatore, coincide con la stima calcolata per quell'area e definita in base allo stimatore che si è scelto di adottare per il censimento della popolazione del 2011 (lo stimatore di ponderazione vincolata).

Il soddisfacimento della condizione (4.1) richiede, quindi, un *riproporzionamento* delle stime per sezione; questa operazione si realizza[9] tramite il prodotto del valore della stima di ciascuna sezione dell'area A con la seguente quantità:

[9] Va precisato che l'operazione può essere facilmente garantita per livelli territoriali gerarchicamente connessi; nei casi in cui tale relazione geografica non è soddisfatta ci potrebbero essere problemi nel calcolo del coefficiente di riproporzionamento.

$$\delta_s(A) = \frac{\hat{Y}_{sc}(A)}{\sum_{s_j \subset A} \hat{Y}_{sg}(s_j)} \qquad (\textit{costante} \text{ per ogni sezione } s_j \subset A) . \qquad (4.2)$$

La (4.2) rappresenta il *coefficiente di riproporzionamento* (specifico per ciascuna area di censimento) tramite il quale si ottengono stime finali della variabile di interesse Y sulle sezioni di censimento in modo coerente con il valore della stima della stessa variabile calcolata sulla relativa area di censimento di appartenenza.

### *4.4 La metodologia proposta per la determinazione di stime riferite a domini sub-comunali*

L'approccio preso in esame per determinare stime di variabili su domini sub-comunali non pianificati dal disegno di indagine, sfrutta da un lato le possibilità offerte dai metodi di stima per piccole aree per produrre stime relative a piccoli ambiti territoriali e dall'altro la circostanza di decomporre i dati riferiti alle stesse variabili e noti in forma aggregata su un'area più ampia, secondo la modalità descritta nel paragrafo precedente.
Di seguito sono descritti i passi della metodologia necessari all'applicazione della procedura:

1) *calcolo della stima campionaria riferita all'area di censimento;*
2) *impiego dei metodi di stima per piccole aree per ottenere stime riferite alle sezioni di censimento;*
3) *applicazione della procedura di riproporzionamento alle stime riferite alle sezioni di censimento per renderle coerenti con quella calcolata per l'area di censimento.*

Così facendo, si potrà procedere da dati stimati ad un livello più aggregato (le aree di censimento) a stime coerenti riferite a domini più piccoli (le sezioni di censimento) che rispettano la condizione espressa dalla (4.1). Si potranno quindi determinare stime sia per sezione di censimento che per domini sub-comunali risultanti da una qualunque aggregazione delle sezioni stesse; di conseguenza, sarà possibile produrre informazioni censuarie per tutti quegli ambiti sub-comunali non pianificati dal disegno di indagine che, non godendo di campioni rappresentativi, pongono problemi per l'adozione di stimatori classici.
Riguardo ai metodi di stima per il passo 2) della procedura, in questo lavoro sono stati presi in esame due metodi diretti che fanno riferimento ad un approccio basato sul disegno, lo stimatore espansione e lo stimatore di regressione generalizzata (GREG) (Generalized Regression Estimator) e i metodi indiretti[10] introdotti nel paragrafo 4.2 .
Per motivi computazionali (tempi di elaborazione inferiori per matrici di dati di dimensioni ridotte), gli stimatori per piccole aree (diretti ed indiretti) considerati si basano su un modello

[10] I metodi presi in considerazione nello studio sperimentale sono quelli impiegati nell'ambito del Progetto Europeo EURAREA (EURAREA Consortium, 2004).

ad “effetti singoli”, in cui le distribuzioni relative alle variabili ausiliarie sono considerate in modo disgiunto[11].

# 5 METODOLOGIE DI STIMA PER PICCOLE AREE

## *5.1 Introduzione alla stima per piccole aree*

Le indagini campionarie su larga scala hanno generalmente la finalità di stimare un vasto insieme di parametri non solo relativi all’intera popolazione oggetto di studio ma anche a sottopopolazioni riferite ad aree geografiche oppure indotte da classificazioni di tipo demografico o socio-economico. Le stime dirette dei parametri relativi ad una data sottopopolazione sono generalmente basate unicamente sui dati osservati sulle unità campionarie ad essa appartenenti. Tuttavia, nella maggior parte delle indagini reali, la numerosità campionaria complessiva non è tale da garantire l’attendibilità delle stime dirette per tutte le sottopopolazioni di interesse. Si utilizza il termine “piccola area” per indicare ogni sottopopolazione per la quale non è possibile produrre stime dirette con una adeguata precisione campionaria.

Lo studio di metodologie statistiche per la produzione di stime per piccole aree sta assumendo sempre maggiore rilevanza sia a livello nazionale che internazionale; infatti, nel corso degli ultimi anni è cresciuta l'esigenza di adeguare le strutture e le procedure mediante le quali si attuano, ai vari livelli di governo, le scelte politiche e gli adempimenti di natura amministrativa (D’Alò *et al.*, 2008). Ciò ha determinato un ampliamento ed una specializzazione della domanda di statistiche riferite a piccole aree.

Gli stimatori per piccole aree più rilevanti dal punto di vista teorico e di maggiore diffusione applicativa possono essere classificati sia in funzione del contesto di rilevazione a cui si riferiscono, indagine occasionale o ripetuta nel tempo, sia in funzione dell’approccio inferenziale alla base della loro costruzione, approccio basato sul disegno o basato sul modello.

Rispetto ai metodi di stima diretti, gli stimatori per piccole aree permettono di migliorare il livello di precisione, utilizzando i valori della variabile d'interesse osservati sulle unità campionarie di un'area, detta macro-area, contenente la piccola area e/o relativi ad altre occasioni d'indagine oltre a quella corrente. Infatti, il riferimento alla macro-area ed

[11] Il modello ad “effetti singoli” ha impiegato le seguenti distribuzioni:
- 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).

eventualmente alle precedenti occasioni di indagine porta ad aumentare la numerosità campionaria effettiva su cui sono calcolate le stime.
L'inferenza si basa generalmente su un modello che esprime il legame tra le osservazioni relative alle piccole aree appartenenti alla macro-area e/o riferite alle precedenti occasioni di indagine, sfruttando la conoscenza di informazioni ausiliarie, correlate alla variabile di interesse, desunte dal censimento oppure da archivi di tipo amministrativo. Tali metodi, pur introducendo una certa componente distorsiva legata alla validità del modello ipotizzato, consentono generalmente una riduzione della variabilità delle stime prodotte rispetto a quella ottenibile con i metodi di tipo diretto.
Per la loro effettiva utilizzazione è necessario tener conto di alcuni problemi di natura teorica ed applicativa. Poiché non si verifica mai una perfetta aderenza tra il modello ipotizzato ed i rispettivi fenomeni rilevati, gli stimatori in questione sono sempre soggetti a distorsioni di difficile misurazione. Uno dei problemi riguarda quindi la robustezza dei metodi, in particolare l'individuazione di opportuni criteri per la diagnosi della validità delle ipotesi alla base degli stimatori; un ulteriore problema è la costruzione di stimatori che tengano conto di situazioni più complesse e più aderenti alla realtà e l'individuazione di tutte le fonti che possono fornire informazioni qualitativamente affidabili per la scelta delle variabili ausiliarie. Infine, per valutare le proprietà empiriche dei metodi di stima in esame nei reali contesti di indagine, è necessario effettuare verifiche su dati censuari o su dati provenienti da pseudo-popolazioni.

## *5.2 Metodi diretti*

Per primi saranno descritti alcuni stimatori diretti, più precisamente lo stimatore espansione e lo stimatore di regressione generalizzata GREG (Generalized Regression Estimator).
Il parametro di interesse relativo alla piccola area d può essere scritto nel modo seguente

$$\theta_d = \frac{1}{N_d} \sum_{i \in U_d} Y_i \quad , \qquad (5.1)$$

in cui $U_d$ è l'insieme delle unità appartenenti alla piccola area d, $Y_i$ è il valore della variabile di interesse osservato sull'i-esima unità della popolazione, con i=1,…..,$N_d$.

### *5.2.1 Stimatore espansione*

Lo stimatore diretto può essere espresso come:

$$\hat{\theta}_d = \frac{1}{N_d} \sum_{i \in s_d} w_i Y_i \quad , \qquad (5.2)$$

dove $s_d$ indica l'insieme delle unità campionarie appartenenti alla piccola area d e $w_i$ è il peso assegnato all'i-esima unità del campione (i=1,…..,$n_d$).

### *5.2.2 Stimatore di regressione generalizzata (GREG)*

Lo stimatore GREG utilizza in modo efficiente l'informazione ausiliaria, calibrando le stime rispetto alle covariate considerate all'interno del modello. Nella stima di parametri riferiti alla popolazione obiettivo o a sotto-popolazioni della stessa, lo stimatore GREG da luogo a guadagni in termini di efficienza rispetto allo stimatore espansione, in funzione del grado di correlazione esistente tra la variabile di interesse e le covariate considerate.
Lo stimatore GREG costituisce un caso particolare degli stimatori di calibrazione (Deville e Särndal 1992) e può essere esplicitato aggiungendo allo stimatore espansione un termine di aggiustamento, funzione della differenza calcolata tra le medie delle singole covariate note nella popolazione e le rispettive stime calcolate attraverso le osservazioni campionarie:

$$\hat{\theta}_d^{\mathrm{GREG}} = \frac{1}{N_d}\sum_{i\in s_d} w_i y_i + \left(\overline{\mathbf{X}}_d - \frac{1}{N_d}\sum_{i\in s_d} w_i \mathbf{x}_i\right)^T \hat{\boldsymbol{\beta}} \quad . \tag{5.3}$$

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

$$y_{di} = \mathbf{x}_{di}^T\boldsymbol{\beta} + \varepsilon_{di} \quad , \tag{5.4}$$

in cui

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

dove $\mathbf{x}_{di} = \left(x_{di,1},...,x_{di,p}\right)^T$ è il vettore delle osservazioni campionarie delle p covariate relativa all'i-ma unità sull'area d. Stimato il coefficiente di regressione $\boldsymbol{\beta}$ mediante l'usuale metodo dei minimi quadrati ponderati:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in s_d} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i\in s_d} w_i \mathbf{x}_i y_i \quad , \tag{5.5}$$

l'espressione dello stimatore (5.3) è equivalente alla seguente:

$$\hat{\theta}_d^{\text{GREG}} = \frac{1}{N_d}\sum_{i\in s_d} w_i\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right) + \overline{\mathbf{X}}_d^T\hat{\boldsymbol{\beta}} \qquad (5.6)$$

### *5.3 Metodi basati su modelli lineari ad effetti misti*

In questo paragrafo si descrive la struttura formale di alcuni metodi di stima per piccole aree basati su un approccio predittivo: lo stimatore sintetico di regressione e il predittore EBLUP (Empirical Best Linear Umbiased Predictor) basati su un modello lineare ad effetti misti a livello di unità elementare e gli analoghi stimatori basati su un modello lineare ad effetti misti a livello di area. La differenza tra gli stimatori sintetici e i corrispondenti stimatori EBLUP è che nei primi, nell'espressione del predittore, si tiene conto della sola componente fissa del modello mentre nei secondi l'introduzione della componente casuale specifica di area, all'interno del predittore, si combina con gli effetti fissi, permettendo di definire stimatori che tengono conto di una fonte di variabilità specifica di area e non spiegata dagli effetti fissi del modello.
Nella specificazione del modello alla base di tali stimatori si assume che gli effetti casuali di area sono normalmente distribuiti, indipendenti ed omoschedastici.

#### *5.3.1 Stimatore sintetico basato su un modello a livello di unità (SYN_A)*

Tale stimatore è costruito sulla base di un modello lineare ad effetti misti con variabili ausiliarie specificate a livello di unità, effetti casuali di area ed errori accidentali distribuiti normalmente e tra loro indipendenti (Battese et al., 1988), che può essere formalizzato nel seguente modo:

$$y_{di} = x_{di}^T\boldsymbol{\beta} + u_d + e_{di} \quad , \qquad (5.7)$$

dove

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

Il modello (5.7) può essere scritto in forma più compatta mediante una rappresentazione matriciale. In tal caso, considerando la formulazione relativa alle sole unità campionarie, risulta:

$$\mathbf{y}_s = \mathbf{x}_s\boldsymbol{\beta} + \mathbf{z}_s\mathbf{u} + \mathbf{e}_s \quad , \qquad (5.8)$$

dove $\mathbf{y}_s$ è il vettore n-dimensionale delle osservazioni campionarie, $\mathbf{x}_s$ è la matrice delle covariate osservate sulle unità campionarie di dimensione $n \times p$ , $\mathbf{e}_s$ è il vettore n-dimensionale degli errori accidentali, $\mathbf{z}_s$ è la matrice di incidenza delle unità campionarie

in ogni area di dimensione $n \times D$, $\mathbf{u}$ il vettore D-dimensionale delle componenti casuali di area. Determinato lo stimatore dei minimi quadrati ponderati di $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}_s \hat{\mathbf{V}}_s \mathbf{x}_s^T\right)^{-1} \mathbf{x}_s^T \hat{\mathbf{V}}_s \, \mathbf{y}_s \quad , \tag{5.9}$$

in cui la stima $\hat{\mathbf{V}}_s$ della matrice di varianza di $\mathbf{y}_s$

$$\hat{\mathbf{V}}_s = \hat{\sigma}_e^2 \mathbf{I}_s + \hat{\sigma}_u^2 \mathbf{z}_s \mathbf{z}_s^T \tag{5.10}$$

è ottenuta stimando iterativamente le componenti di varianza $\sigma_e^2$ e $\sigma_u^2$, lo stimatore sintetico (SYN_A) basato su un modello lineare ad effetti misti definito a livello di unità è:

$$\hat{\theta}_d^{\text{SYN\_A}} = \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} \quad , \tag{5.11}$$

in cui $\bar{\mathbf{x}}_d = \left(\bar{x}_{d,1},...,\bar{x}_{d,p}\right)^T$ indica il vettore delle medie campionarie delle p variabili ausiliarie.

### *5.3.2 Stimatore sintetico basato su un modello a livello di area (SYN_B)*

Il modello alla base dello stimatore sintetico in questione è definito a livello aggregato. In tal caso si definisce una relazione tra le stime dirette del parametro di interesse e le medie delle variabili ausiliarie riferite alle piccole aree. Il modello può essere espresso nel seguente modo:

$$\hat{\theta}_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d + \bar{e}_d \quad , \tag{5.12}$$

in cui

$$u_d \sim iid\ N\left(0, \sigma_u^2\right),\ \bar{e}_d \sim iid\ N\left(0, \sigma_e^2 / n_d\right) \quad ,$$

dove $n_d$ è la dimensione campionaria nell'area d e $\bar{\mathbf{X}}_d^T$ è il vettore delle medie di popolazione delle p variabili ausiliarie nell'area d. In forma matriciale il modello (5.12) può essere riscritto nel modo seguente:

$$\bar{\mathbf{y}} = \bar{\mathbf{X}} \boldsymbol{\beta} + \mathbf{u} + \bar{\mathbf{e}} \quad , \tag{5.13}$$

con

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

e

$$\mathbf{D} = \begin{pmatrix} \sigma_e^2 / n_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_e^2 / n_D \end{pmatrix} \quad .$$

Nei modelli di tipo aggregato, al fine di evitare problemi di identificabilità, le varianze campionarie si assumono generalmente note, tuttavia, disponendo delle informazioni a livello individuale e nell'ipotesi di omoschedasticità della componente accidentale, la varianza $\sigma_e^2$ può essere stimata mediante la seguente espressione:

$$\hat{\sigma}_e^2 = \frac{1}{n - n^{(\mathrm{D})}} \sum_i \sum_d \left( y_{di} - \bar{y}_d \right)^2 \quad , \tag{5.14}$$

dove n è il numero di individui appartenenti al campione complessivo ed $n^{(D)}$ è il numero di aree presenti nel campione.
Lo stimatore dei minimi quadrati ponderati del vettore dei coefficienti di regressione $\boldsymbol{\beta}$ è dato da:

$$\hat{\boldsymbol{\beta}} = (\overline{\mathbf{X}}^T \hat{\mathbf{V}}^{-1} \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T \hat{\mathbf{V}}^{-1} \overline{\mathbf{y}} \quad , \tag{5.15}$$

dove $\overline{\mathbf{y}}$ è il vettore delle medie campionarie, $\overline{\mathbf{X}}$ è la matrice composta dalle righe $\overline{\mathbf{X}}_d^T$, $\hat{\mathbf{V}} = \hat{\sigma}_u^2 \mathbf{I} + \hat{\mathbf{D}}$ è una matrice diagonale con elementi pari a $\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d$, in cui $\hat{\sigma}_u^2$ e $\boldsymbol{\beta}$ sono stimati in modo iterativo.
L'espressione finale dello stimatore sintetico (SYN_B) basato su un modello definito a livello di area è:

$$\hat{\theta}_d^{\mathrm{SYN\_B}} = \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} \quad . \tag{5.16}$$

### *5.3.3 Stimatore EBLUP basato su un modello a livello di unità (EBLUP_A)*

Lo stimatore in questione, così come lo stimatore SYN_A, è basato sul modello lineare misto definito nella (5.7). Nell'ambito dell'approccio predittivo, il miglior predittore lineare corretto (BLUP) è ottenuto minimizzando gli errori quadratici medi all'interno della classe degli stimatori lineari non distorti. Lo stimatore BLUP dipende dalle componenti di varianza $\sigma_u^2$ e $\sigma_e^2$ che sono generalmente incognite; è necessario, quindi, calcolare una loro stima. Stimati i coefficienti di regressione tramite l'espressione (5.9), le componenti di varianza del modello possono essere stimate mediante differenti metodi, tra cui quello della massima verosimiglianza (ML) o della massima verosimiglianza ristretta (REML) (si veda ad esempio Cressie, 1992). Il miglior predittore lineare empirico (EBLUP_A) è uno stimatore di tipo composto che, trascurando il fattore di correzione per popolazioni finite, è dato da:

$$\hat{\theta}_d^{EBLUP\_A} = \gamma_d \left[ \bar{y}_d + \left( \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} \right) \right] + (1 - \gamma_d) \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} \quad , \tag{5.17}$$

dove

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d} \tag{5.18}$$

è il peso associato alla componente diretta, $\bar{y}_d$ e $\bar{\mathbf{x}}_d^T$ sono rispettivamente il vettore delle medie campionarie della variabile di interesse $y$ e delle covariate nell'area d, $\overline{\mathbf{X}}_d$ è il vettore

dei valori medi di popolazione delle covariate, $\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$ sono le stime dei parametri del modello lineare definito a livello di unità.

### *5.3.4 Stimatore EBLUP basato su un modello a livello di area (EBLUP_B)*

Tale stimatore, al pari di quello SYN_B, si basa sul modello lineare normale ad effetti misti definito a livello di piccola area, dato dalla (5.12). Dopo aver stimato le componenti di varianza mediante i metodi ML o REML ed i coefficienti di regressione, con l'espressione (5.15), il miglior predittore lineare corretto ottenuto sulla base del modello (5.12) è pari alla combinazione ponderata dello stimatore diretto e dello stimatore SYN_ B:

$$\hat{\theta}_d^{\text{EBLUP\_B}} = \gamma_d \hat{\theta}_d + (1-\gamma_d)\overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} \quad , \tag{5.19}$$

dove

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \tag{5.20}$$

è il peso relativo allo stimatore diretto.

## *5.4 Metodi basati su modelli con correlazione spaziale*

Nel paragrafo 5.3.1 è stato descritto il modello lineare ad effetti misti definito a livello di unità. In tale modello gli effetti casuali di area sono stati ipotizzati incorrelati. In realtà, è possibile complicare ulteriormente il modello (5.7) considerando effetti di area $u_d$ che abbiano una struttura di autocorrelazione spaziale, ossia ipotizzare che il legame tra le osservazioni rilevate nelle aree di interesse e quelle rilevate nelle altre aree sia funzione della loro distanza geografica.
Considerando il modello relativo alle sole unità campionarie si può scrivere

$$y_s = x_s\beta + z_s u + e_s \quad , \tag{5.21}$$

in cui $e_s$ ed u hanno una distribuzione multinormale con vettore delle medie nullo e matrice di varianza rispettivamente uguali a $\sigma_e^2 I_n$ e $\sigma_u^2 A$, dove, posto $\delta_{dd'} = 0$ se $d = d'$ e $\delta_{dd'} = 1$ altrimenti, la matrice **A** è data da

$$A = [a_{dd'}] = \left\{ \left[ 1 + \delta_{dd'} \exp\left( \frac{\text{dist}(d, d')}{\alpha} \right) \right]^{-1} \right\} \tag{5.22}$$

dove $\text{dist}(d, d')$ è la distanza tra le aree d e d' ed $\alpha$ è un parametro di scala incognito.
Di conseguenza, la matrice di varianza e covarianza di $y_s$ è pari a $\sigma_e^2 V_s$, con $V_s = I_n + \varphi z_s A z_s^T$, in cui $\varphi = \sigma_u^2 / \sigma_e^2$ .
L'espressione dello stimatore EBLUP per la generica area d può essere scritto come

$$\hat{\theta}_d^{EBLUP\_S} = \frac{1}{P_d}\left( y_d + \left[ (X_d - x_d)\hat{\beta} + (P_d - n_d)\sum_{d'=1}^{D} \left[ y_{d'} - x_{d'}^{'}\hat{\beta} \right] \tau_{d',d} \right] \right), \qquad (5.23)$$

dove $y_d$ è il numero totale di famiglie (aventi la caratteristica in esame) osservate nel sottocampione relativo all'area d, $\mathbf{X}_d$ è il vettore dei totali di popolazione delle covariate nella piccola area d, $x_d$ è il vettore dei totali campionari delle covariate nella piccola area d, $\hat{\beta}$ è l'usuale stimatore dei minimi quadrati ponderati dei coefficienti di regressione, $P_d - n_d$ è la differenza tra il numero unità elementari della popolazione appartenenti all'area d e il corrispondente numero unità osservate nel campione e, infine, $\tau_{d',d}$ è il generico elemento della seguente matrice

$$T^* = [\tau_{d',d}] = \left( z_s^{'} z_s + \varphi^{-1} A^{-1} \right)^{-1} = \left( \text{Diag}[n_d] + \varphi^{-1} A^{-1} \right)^{-1} \quad . \qquad (5.24)$$

Analogamente a quanto visto per il predittore EBLUP (5.17) basato sul modello (5.7), l'espressione (5.23) è ottenuta stimando le componenti di varianza mediante il metodo della verosimiglianza ristretta, utilizzando un algoritmo di stima iterativo descritto in EURAREA Consortium (2004).

# 6 AMBITO DELLO STUDIO SPERIMENTALE

## *6.1 Obiettivi*

Questo studio prosegue una sperimentazione già avviata (Carbonetti e Fiorello, 2010) con lo scopo di valutare la praticabilità di stimatori indiretti per produrre stime accurate riferite a domini sub-comunali differenti da quelli pianificati dal disegno d'indagine.
In questo ambito, si valuta la bontà dei risultati ottenibili con l'impiego dei metodi di stima per piccole aree, sia nell'approccio basato sul modello ad effetti lineari misti (paragrafo 5.3), sia nell'approccio che considera anche la componente spaziale (paragrafo 5.4) per spiegare il comportamento delle variabili oggetto di studio. In particolare, i diversi metodi saranno valutati nell'ottica di produrre informazione censuaria sia con riferimento a domini di

dimensioni paragonabili a quelle delle aree di censimento, che per domini di elevato dettaglio territoriale, coincidenti con le sezioni di censimento.
Le sperimentazioni condotte in questo specifico lavoro hanno inoltre permesso di fare importanti valutazioni in merito al contributo fornito dall'informazione spaziale delle unità considerate.

### *6.2 Dati*

Il caso di studio, riprendendo in modo identico quello del precedente lavoro di Carbonetti e Fiorello (2010), si riferisce al comune di Milano e alle sue suddivisioni territoriali. Per la fase sperimentale sono stati considerati dati del censimento della popolazione e delle abitazioni del 2001 relativi agli individui residenti in famiglia. Come base di campionamento per l'estrazione dei campioni di famiglie, è stata considerata la lista delle famiglie censite nel 2001, in qualità di "ipotetica" lista anagrafica[12].
Per le variabili oggetto di stima, è stata confermata la scelta relativa al "numero di occupati" e al "numero di persone in cerca di occupazione", aventi frequenze percentuali differenti (sui dati 2001 e per le unità territoriali in esame, si osservano valori compresi tra il 43% e il 47% per la prima, mentre si registrano frequenze comprese tra l'1% e il 3% per la seconda).
Per quanto riguarda i domini territoriali, sono state considerate le suddivisioni territoriali del comune di Milano relative alle seguenti unità:
- le *zone di decentramento* valide in occasione della diffusione dei risultati del censimento 2001 e riproposte senza modifiche per il 2011 (9 zone);
- le *aree di censimento di centro abitato* proposte per il censimento del 2011 (134 aree) e disegnate nel rispetto dei limiti territoriali delle zone di decentramento;
- i *nuclei di identità locale*[13] aventi valenza amministrativa/funzionale (83 nuclei);
- le *sezioni di censimento di centro* delle Basi Territoriali Comunali 2001 (5.933 sezioni).

Si specifica che le suddivisioni territoriali del comune di Milano in *aree di censimento* e in *nuclei di identità locale* utilizzate in questo lavoro non sono quelle definitive ma fanno riferimento ad una proposta di "disegno" (Aprile 2010).

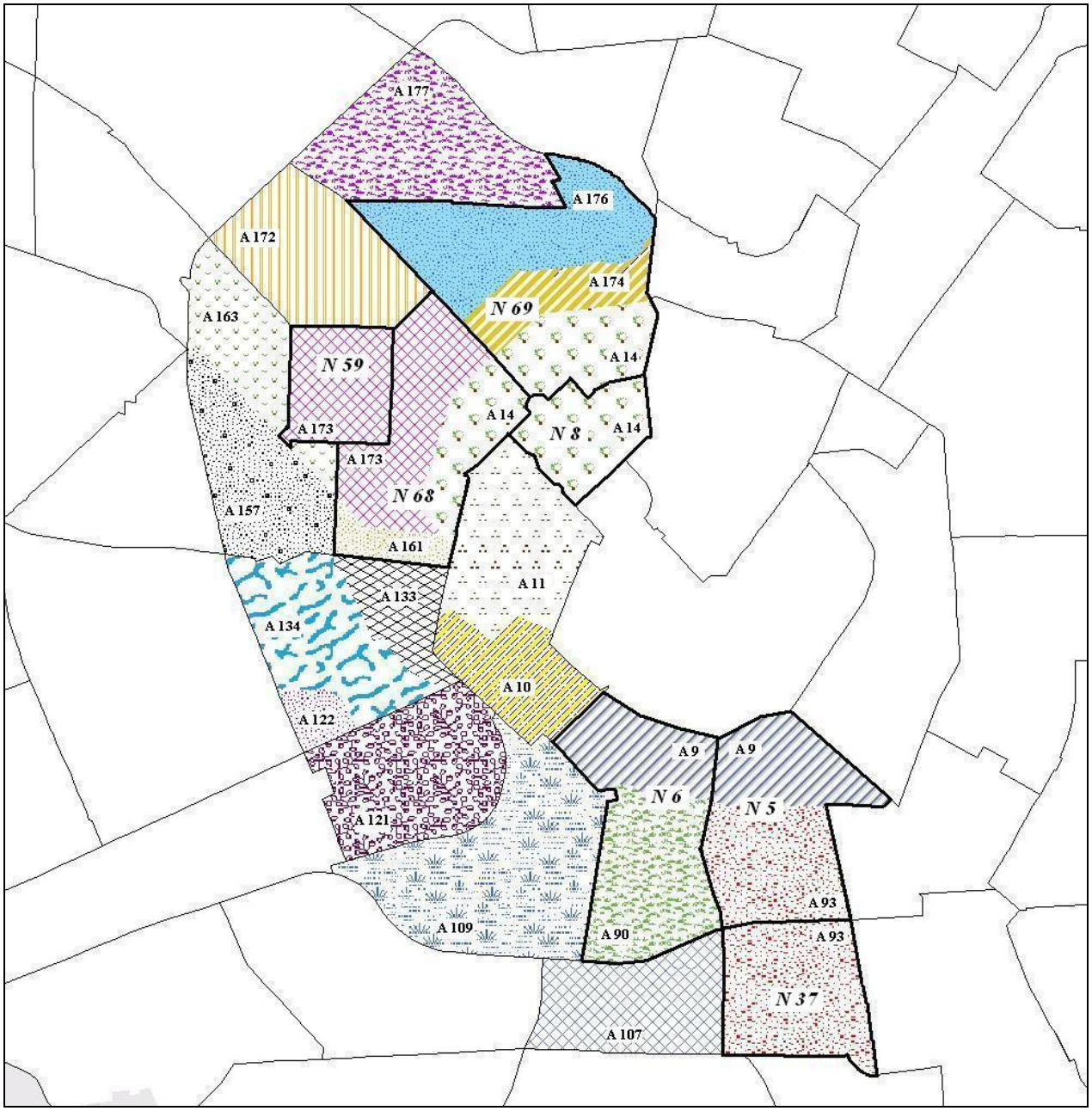
### *6.3 Ambito territoriale considerato per le sperimentazioni*

Le sperimentazioni sono state condotte con riferimento ad una porzione del territorio di Milano definita da una opportuna *macro-area* scelta nel centro del comune (Figura 1). La

[12] Sotto l'ipotesi di *invarianza* rispetto alla lista anagrafica presente negli archivi amministrativi del comune.
[13] Sono unità territoriali proposte come ulteriore suddivisione sub-comunale, geograficamente più fine di quella in zone di decentramento, e alternativa a quella in vigore nel 2001 riferita alle *aree funzionali*. I confini di tali unità non sempre sono contenuti interamente nelle zone di decentramento (vedi Appendice).

dimensione di tale macro-area è stata individuata proprio in funzione della possibilità di impiegare i metodi di stima per piccole aree; essa accorpa frazioni di 4 zone di decentramento ed è costituita, in riferimento ai diversi livelli territoriali, da 20 aree di censimento, 15 nuclei di identità locale e 953 sezioni di censimento di centro.

*Figura 1* - Cartografia della *macro-area* del comune di Milano presa in esame nelle sperimentazioni: i limiti dei poligoni distinguono i nuclei di identità locale; i poligoni in *tessitura* identificano le aree di censimento.




*I nuclei e le aree oggetto di studio sono identificati nel cartogramma rispettivamente dalle lettere N ed A seguite dal relativo codice numerico. Sono considerati ai fini delle analisi dei risultati i 7 nuclei disegnati con i limiti di confine più marcati.*

Nell'ambito territoriale preso in esame, le stime sono state determinate per tutte le unità territoriali presenti al livello delle aree di censimento, dei nuclei di identità locale e delle sezioni di censimento.
Con riferimento all'opportunità di produrre dati sub-comunali relativi ai nuclei di identità locale, per le valutazioni è stata posta l'attenzione solo su 7 dei 15 nuclei della macro-area (quelli con codice: "5", "6", "8", "37", "59", "68", "69") perché pongono reali problemi di stima; gli altri nuclei della macro-area, coincidenti con una singola area di censimento (5 casi) o ottenuti dall'aggregazione di più aree (3 casi), non sono stati considerati nell'analisi in quanto non comportano problemi di stima.
Il dettaglio delle zone di decentramento non è stato preso in esame in quanto, essendo domini riconducibili ad aggregazioni di aree di censimento, la produzione di dati censuari è garantita con livelli di accuratezza paragonabili a quelli attesi per le aree di censimento.
Infine, con riferimento alle stime riferite ai domini di massimo dettaglio territoriale, sono state considerate tutte le 953 sezioni di censimento della macro-area; le valutazioni sono state inoltre condotte sia globalmente sull'intero insieme che per gruppi di sezioni, classificate secondo opportuni criteri.

### *6.4 Algoritmo di simulazione*

In questo lavoro, l'operazione di simulazione effettuata, pur riferendosi solo al metodo di stima indiretto basato sull'approccio che tiene conto dell'informazione spaziale, ha ripercorso gli stessi passi dell'algoritmo (in ambiente SAS, *Statistical Analysis System*) seguito nella precedente sperimentazione:

1) estrazione di un campione di famiglie da lista anagrafica per ciascuna area di censimento secondo un disegno casuale semplice con frazione sondata pari al 33%;
2) calcolo delle stime calibrate[14] per ciascuna area di censimento;
3) applicazione del metodo di stima per piccole aree basato su autocorrelazione spaziale per determinare le stime riferite alle sezioni di censimento;
4) riproporzionamento delle stime per sezione per renderle *coerenti* con la stima relativa all'area di censimento in cui la sezione ricade;
5) aggregazione delle stime per sezione sui domini sub-comunali nelle quali ricadono;
6) iterazione dei passi 1)-5) per un numero prefissato di volte (500 replicazioni campionarie).

[14] Per la procedura di calibrazione si è fatto uso della *funzione di calibrazione* del software Genesees V3.0 sviluppato in Istat (2005). I vincoli di calibrazione utilizzati in questa sperimentazione si riferiscono ad una struttura più ampia rispetto a quella considerata in precedenti sperimentazioni e di cui sono presentati alcuni risultati nella Tabella 3. In particolare, in questo studio, le stime calibrate hanno impiegato il sistema di vincoli relativo alle seguenti strutture demografiche (42 totali noti): distribuzione della popolazione per sesso e per 16 classi di età; distribuzione della popolazione per sesso e per 5 modalità di stato civile.

# 7 RISULTATI

## *7.1 Confronto dei livelli di distorsione e variabilità delle stime riferite ai "nuclei di identità locale"*

In questo primo paragrafo si considerano i risultati delle sperimentazioni relative alla produzione di stime per domini sub-comunali di dimensioni paragonabili a quelle delle aree di censimento. Per lo studio condotto sui dati di Milano, sono esposti i risultati dei livelli di efficienza delle stime riferite ai "nuclei di identità locale" (**NIL**) che, tra quelli della macro-area presa in esame, presentano particolari problemi di stima (paragrafo 6.3); dall'insieme iniziale è stato escluso il nucleo con codice "8" in quanto numericamente inconsistente.
Le proprietà dei differenti stimatori considerati nell'approccio metodologico proposto sono state studiate tramite opportuni indici calcolati sulla distribuzione empirica definita in base ai risultati osservati sulle **R**=500 replicazioni campionarie generate nella simulazione.
Per misurare la distorsione e la variabilità attesa dall'adozione dei differenti stimatori per ciascuno dei domini di stima presi in esame, sono stati calcolati i seguenti indici sintetici:

$$ARB_d = \frac{1}{R}\sum_{r=1}^{R}\frac{\left|{}_d\hat{T}_r - {}_dT\right|}{{}_dT}\times 100 \qquad (7.1)$$

$$RRMSE_d = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{{}_d\hat{T}_r - {}_dT}{{}_dT}\right)^2}\times 100 \qquad (7.2)$$

$$CV_d = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left({}_d\hat{T}_r - {}_dT\right)^2}}{E({}_d\hat{T}_r)}\times 100 \qquad (7.3)$$

in cui ${}_d\hat{T}_r$ indica la stima della frequenza assoluta ${}_dT$ (incognita) riferita al dominio d e determinata in corrispondenza della realizzazione del generico campione r .
La *Distorsione Relativa* (7.1) e la *Radice dell'Errore Quadratico Medio Relativo* (7.2) permettono di misurare rispettivamente la distorsione e la variabilità attesa delle stime; il *Coefficiente di Variazione*[15] (7.3) esprime, invece, una misura sintetica dell'errore di campionamento atteso.

[15] Poiché gli stimatori utilizzati non sono corretti, il calcolo del coefficiente di variazione viene definito come rapporto tra la radice quadrata dell'errore quadratico medio (MSE) e il valore atteso delle stime. Nel caso di

### *7.1.1 Confronti di efficienza delle stime riferite ai "nuclei di identità locale" per la variabile "numero di occupati"*

La Tabella 2 contiene i valori degli indicatori ARB, RRMSE e CV, misurati per i 6 nuclei presi in esame, per la stima della variabile *"numero di occupati"*, con i differenti approcci sperimentati che impiegano metodi diretti ed indiretti.

*Tabella 2* - Valori degli indici ARB, RRMSE e CV degli stimatori sperimentati per la stima della variabile *"numero di occupati"* di alcuni Nuclei di Identità Locale (NIL) del comune di Milano. (Censimento 2001)

| Codice NIL | 59 | 37 | 5 | 68 | 6 | 69 |
|---|---|---|---|---|---|---|
| Popolazione dei NIL (2001) | 1375 | 5263 | 13386 | 16044 | 20343 | 27894 |
| *Occupati* (2001) | 609 | 2238 | 5943 | 7144 | 9589 | 13026 |
| | Misura della distorsione relativa assoluta (**ARB**) | | | | | |
| Stimatore Espansione | 3,15 | 1,61 | 0,93 | 0,88 | 0,69 | 0,57 |
| Stimatore Greg | 2,90 | 1,46 | 2,28 | 1,00 | 1,51 | 0,66 |
| Stimatore Sintetico *unit-level* | 0,95 | 3,14 | 0,57 | 0,66 | 0,82 | 0,51 |
| Stimatore EBLUP *unit-level* | 1,03 | 2,56 | 0,72 | 0,69 | 0,92 | 0,52 |
| Stimatore Sintetico *area-level* | 1,05 | 2,26 | 1,70 | 0,78 | 0,77 | 0,57 |
| Stimatore EBLUP *area-level* | 1,06 | 2,22 | 1,68 | 0,78 | 0,77 | 0,58 |
| Stimatore Spaziale *unit-level* | 1,46 | 1,22 | 0,75 | 0,69 | 0,62 | 0,54 |
| | Misura della variabilità (**RRMSE**) | | | | | |
| Stimatore Espansione | 3,97 | 2,03 | 1,15 | 1,09 | 0,87 | 0,72 |
| Stimatore Greg | 3,62 | 1,86 | 2,48 | 1,24 | 1,70 | 0,82 |
| Stimatore Sintetico *unit-level* | 1,19 | 3,33 | 0,73 | 0,82 | 0,99 | 0,64 |
| Stimatore EBLUP *unit-level* | 1,28 | 2,80 | 0,89 | 0,85 | 1,11 | 0,65 |
| Stimatore Sintetico *area-level* | 1,31 | 2,53 | 1,85 | 0,98 | 0,94 | 0,71 |
| Stimatore EBLUP *area-level* | 1,33 | 2,49 | 1,83 | 0,98 | 0,94 | 0,72 |
| Stimatore Spaziale *unit-level* | 1,85 | 1,50 | 0,96 | 0,87 | 0,77 | 0,67 |
| | Coefficiente di variazione (**CV**) | | | | | |
| Stimatore Espansione | 3,97 | 2,03 | 1,15 | 1,09 | 0,87 | 0,72 |
| Stimatore Greg | 3,58 | 1,85 | 2,42 | 1,24 | 1,73 | 0,83 |
| Stimatore Sintetico *unit-level* | 1,19 | 3,23 | 0,73 | 0,82 | 1,00 | 0,64 |
| Stimatore EBLUP *unit-level* | 1,27 | 2,74 | 0,88 | 0,85 | 1,12 | 0,65 |
| Stimatore Sintetico *area-level* | 1,31 | 2,47 | 1,88 | 0,98 | 0,93 | 0,71 |
| Stimatore EBLUP *area-level* | 1,33 | 2,44 | 1,86 | 0,99 | 0,93 | 0,71 |
| Stimatore Spaziale *unit-level* | 1,85 | 1,49 | 0,96 | 0,87 | 0,77 | 0,67 |

stimatori corretti, poiché l'MSE si riduce alla varianza campionaria delle stime e il valore atteso riproduce il valore vero oggetto di stima, il coefficiente di variazione si traduce nel rapporto tra lo scarto quadratico medio e il valore vero.

A riguardo si nota che:

- ✓ per il dominio più piccolo (NIL "59"), i metodi indiretti conducono ai risultati migliori, e tra questi l'impiego dello stimatore sintetico *unit-level* fa registrare i valori più bassi degli indicatori calcolati rispetto agli altri metodi confrontati;
- ✓ i risultati osservati sul dominio "37" evidenziano come lo stimatore spaziale risolve i livelli di inefficienza osservati con gli altri metodi indiretti rispetto ai metodi diretti;
- ✓ sui NIL "5" e "68", di dimensioni paragonabili a quelle delle aree di censimento più grandi, sono attesi comportamenti più favorevoli verso l'utilizzo dei metodi indiretti *unit-level*, con preferenza verso lo stimatore sintetico, anche se il metodo spaziale si propone come una valida alternativa;
- ✓ sui domini più grandi (NIL "6" e "69"), superiori ai 20mila abitanti, il favore si sposta decisamente verso l'impiego dello stimatore spaziale che sembra offrire vantaggi anche rispetto all'impiego dello stimatore espansione.

### *7.1.2 Confronti di efficienza delle stime riferite ai "nuclei di identità locale" per la variabile "numero di persone in cerca di occupazione"*

La Tabella 3 contiene le misure degli indicatori di distorsione e variabilità attesa riferite alla stima della variabile *"numero di persone in cerca di occupazione"*.

Per la stima di questa variabile si osserva che:

- ✓ sul dominio più piccolo (NIL "59"), dai valori del CV osservati si evidenzia che l'implementazione dello stimatore per piccole aree basato sulla componente spaziale conduce a livelli di efficienza migliori rispetto a tutti gli altri metodi di stima sperimentati, sia diretti che indiretti;
- ✓ sul NIL "37", di dimensioni intorno a 5mila unità, si osservano valori migliori con i metodi indiretti *unit-level*, incluso quello spaziale, con una leggera preferenza verso lo stimatore EBLUP;
- ✓ per tutti gli altri domini esaminati, superiori a 10mila abitanti e nei quali il valore della frequenza assoluta oggetto di stima è compresa tra 250 e 650 unità, i metodi indiretti di tipo *area-level* e il metodo spaziale risultano più efficienti.

*Tabella 3* - Valori degli indici ARB, RRMSE e CV degli stimatori sperimentati per la stima della variabile *"numero di persone in cerca di occupazione"* di alcuni Nuclei di Identità Locale (NIL) del comune di Milano. (Censimento 2001)

| Codice NIL | 59 | 37 | 5 | 68 | 6 | 69 |
|---|---|---|---|---|---|---|
| Popolazione dei NIL (2001) | 1375 | 5263 | 13386 | 16044 | 20343 | 27894 |
| *In cerca di occupazione* (2001) | 14 | 148 | 297 | 279 | 539 | 642 |
| | Misura della distorsione relativa assoluta (**ARB**) | | | | | |
| Stimatore Espansione | 34,93 | 9,49 | 7,02 | 7,42 | 4,92 | 4,88 |
| Stimatore Greg | 35,88 | 10,71 | 7,62 | 7,66 | 5,20 | 4,98 |
| Stimatore Sintetico *unit-level* | 70,57 | 7,32 | 8,20 | 6,93 | 5,31 | 5,31 |
| Stimatore EBLUP *unit-level* | 63,26 | 7,15 | 7,11 | 6,80 | 5,12 | 5,11 |
| Stimatore Sintetico *area-level* | 75,55 | 8,35 | 5,40 | 6,52 | 4,57 | 4,76 |
| Stimatore EBLUP *area-level* | 75,31 | 8,37 | 5,41 | 6,52 | 4,58 | 4,76 |
| Stimatore Spaziale *unit-level* | 44,18 | 7,86 | 5,85 | 6,58 | 4,68 | 4,75 |
| | Misura della variabilità (**RRMSE**) | | | | | |
| Stimatore Espansione | 43,34 | 11,83 | 8,88 | 9,23 | 6,20 | 6,09 |
| Stimatore Greg | 44,31 | 13,39 | 9,63 | 9,52 | 6,56 | 6,23 |
| Stimatore Sintetico *unit-level* | 73,14 | 9,07 | 9,64 | 8,65 | 6,51 | 6,61 |
| Stimatore EBLUP *unit-level* | 65,72 | 8,93 | 8,64 | 8,42 | 6,31 | 6,40 |
| Stimatore Sintetico *area-level* | 78,28 | 10,30 | 6,64 | 8,01 | 5,77 | 5,95 |
| Stimatore EBLUP *area-level* | 78,04 | 10,32 | 6,66 | 8,02 | 5,77 | 5,94 |
| Stimatore Spaziale *unit-level* | 49,48 | 9,90 | 7,32 | 8,18 | 5,81 | 5,92 |
| | Coefficiente di variazione (**CV**) | | | | | |
| Stimatore Espansione | 43,04 | 11,82 | 8,89 | 9,25 | 6,23 | 6,11 |
| Stimatore Greg | 46,54 | 12,91 | 9,62 | 9,40 | 6,67 | 6,28 |
| Stimatore Sintetico *unit-level* | 42,88 | 9,47 | 9,00 | 8,35 | 6,74 | 6,84 |
| Stimatore EBLUP *unit-level* | 40,26 | 9,20 | 8,19 | 8,20 | 6,49 | 6,59 |
| Stimatore Sintetico *area-level* | 44,59 | 10,97 | 6,52 | 8,11 | 5,77 | 6,04 |
| Stimatore EBLUP *area-level* | 44,51 | 10,99 | 6,54 | 8,12 | 5,77 | 6,03 |
| Stimatore Spaziale *unit-level* | 34,38 | 10,23 | 7,18 | 8,19 | 5,85 | 6,00 |

### *7.2 Confronto dei livelli di distorsione e variabilità delle stime riferite alle sezioni di censimento*

In questo paragrafo, si esaminano i risultati delle sperimentazioni relative al calcolo delle stime riferite ai domini di massimo dettaglio sub-comunale identificati dalle sezioni di censimento di centro. I risultati presentati riguardano i livelli di efficienza delle stime riferite alle **S**=953 sezioni di censimento di tipo "centro" (non vuote) incluse nella macro-area presa in esame per il comune di Milano (paragrafo 6.3).

Diversamente da quanto illustrato nel paragrafo 7.1, in cui le valutazioni sono state fatte in modo distinto per ciascun dominio di studio, in questo ambito è stata effettuata una sintesi degli indici di distorsione e di variabilità calcolati per tutte le sezioni di censimento prese in esame, tramite il calcolo del valore medio e del valore massimo. Quindi, con riferimento alla stima ${}_s\hat{T}_r$ della frequenza assoluta ${}_sT$ (incognita) sulla generica sezione di censimento **s** calcolata in base ai valori osservati sul campione **r**, per ciascun metodo impiegato nella procedura, sono stati calcolati i seguenti indici di distorsione e di variabilità attesa:

$$AARB = \frac{1}{S}\sum_s \left\{ \frac{1}{R} \left| \sum_r \left( \frac{{}_s\hat{T}_r - {}_sT}{{}_sT} \right) \right| \right\} \qquad (7.4)$$

$$ARRMSE = \frac{1}{S}\sum_{s=1}^{S} \left( RRMSE_s \right) \qquad (7.5)$$

dove, in particolare la (7.5) tiene conto dell'espressione (7.2). Quindi, le sopra descritte misure si intendono, rispettivamente, media della *Distorsione Relativa* (7.4) e media della *Radice dell'Errore Quadratico Medio Relativo* (7.5) sull'insieme delle 953 sezioni di censimento della macro-area presa a riferimento. Inoltre, al fine di tenere in considerazione il valore massimo assunto dagli indici di distorsione e di variabilità sull'insieme considerato, sono stati calcolati anche i seguenti indicatori:

$$MARB = \max_s \{ ARB_s \} \qquad (7.8)$$

$$MRRMSE = \max_s \{ RRMSE_s \} \quad . \qquad (7.9)$$

### *7.2.1 Analisi globale dei livelli di efficienza delle stime "coerenti" riferite alle sezioni di censimento*

La Tabella 4 riassume, per ciascuno degli stimatori diretti ed indiretti considerati nell'analisi e per le due variabili di studio prese in esame, le sintesi degli indici di distorsione e variabilità delle stime riferite alle sezioni di censimento, ottenute tramite l'approccio che riproporziona le stime finali in modo da renderle coerenti con il valore relativo all'area di censimento in cui le sezioni sono contenute. Il soddisfacimento della proprietà di coerenza implica, infatti, la coincidenza tra il valore dato dalla somma delle stime per sezione, estesa a tutte le sezioni dell'area di censimento, e il dato precedentemente stimato sull'area stessa.

*Tabella 4* - Valori degli indici AARB, ARRMSE, MARB e MRRMSE degli stimatori sperimentati per la stima delle variabili *"numero di occupati"* e *"numero di persone in cerca di occupazione"* riferite alle sezioni di censimento non vuote (953) di una macro-area del comune di Milano. Caso di stime per sezioni **coerenti** con le stime riferite alle aree di censimento di appartenenza. (Censimento 2001)

| *Stime per sezione di censimento **coerenti** con le stime per area di censimento* | *"Occupati"* | | | | *"Persone in cerca di occupazione"* | | | |
|---|---|---|---|---|---|---|---|---|
| | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** |
| Stimatore Espansione | 0,75 | 13,05 | 27,30 | 83,61 | 2,88 | 75,75 | 25,27 | 211,40 |
| Stimatore Greg | 8,86 | 14,83 | 229,04 | 231,93 | 16,03 | 89,48 | 122,44 | 242,91 |
| Stimatore Sintetico *unit-level* | 11,33 | 11,48 | 263,64 | 263,75 | 55,84 | 58,38 | 745,54 | 747,76 |
| Stimatore EBLUP *unit-level* | 10,79 | 11,17 | 261,27 | 261,37 | 44,49 | 51,04 | 495,86 | 501,51 |
| Stimatore Sintetico *area-level* | 6,76 | 7,41 | 168,16 | 174,76 | 53,00 | 57,66 | 642,91 | 647,79 |
| Stimatore EBLUP *area-level* | 6,10 | 6,97 | 89,16 | 122,89 | 52,40 | 57,23 | 639,53 | 644,45 |
| Stimatore Spaziale *unit-level* | 3,97 | 6,33 | 126,65 | 151,78 | 29,66 | 48,65 | 374,46 | 385,57 |

Con riferimento alla distorsione assoluta, l'impiego dello stimatore diretto di tipo espansione offre migliori garanzie per entrambe le variabili; per quanto riguarda la variabilità, lo stimatore indiretto che tiene conto della componente spaziale offre, invece, la possibilità di ottenere stime mediamente più efficienti.
Nel confronto tra gli stimatori indiretti, si conferma che il metodo basato sullo stimatore spaziale è meno distorto, però, presenta una variabilità più ampia se si effettua la misura come differenza tra l'errore quadratico medio e la distorsione.

### *7.2.2 Analisi globale dei livelli di efficienza delle stime "non coerenti" riferite alle sezioni di censimento*

La Tabella 5 descrive le misure degli stessi indicatori della Tabella 4; in questo caso però le stime ottenute sono il risultato dell'applicazione dei relativi stimatori (diretti ed indiretti), senza il passaggio per l'operazione di riproporzionamento. Poiché si sono eseguiti solo i primi due passi della procedura descritta nel paragrafo 4.4, le stime finali risultano non coerenti con il dato riferito all'area di censimento che la contiene.
Dall'analisi dei risultati attesi per i differenti stimatori, sia in termini di distorsione assoluta che di variabilità, si giunge a conclusioni identiche a quelle esposte nel caso delle stime coerenti (paragrafo 7.2.1).

*Tabella 5* - Valori degli indici AARB, ARRMSE, MARB e MRRMSE degli stimatori sperimentati per la stima delle variabili *"numero di occupati"* e *"numero di persone in cerca di occupazione"* riferite alle sezioni di censimento non vuote (953) di una macro-area del comune di Milano. Caso di stime per sezioni **non coerenti** con le stime riferite alle aree di censimento di appartenenza. (Censimento 2001)

| *Stime per sezione di censimento* ***non coerenti*** *con le stime per area di censimento* | *"Occupati"* | | | | *"Persone in cerca di occupazione"* | | | |
|---|---|---|---|---|---|---|---|---|
| | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** |
| Stimatore Espansione | 0,75 | 13,09 | 27,32 | 83,56 | 2,88 | 75,56 | 25,44 | 217,00 |
| Stimatore Greg | 10,32 | 15,39 | 211,92 | 214,69 | 19,57 | 79,14 | 118,64 | 214,95 |
| Stimatore Sintetico *unit-level* | 12,66 | 12,70 | 235,83 | 235,90 | 51,03 | 51,62 | 521,53 | 522,08 |
| Stimatore EBLUP *unit-level* | 12,20 | 12,41 | 235,28 | 235,35 | 41,56 | 45,81 | 413,80 | 416,25 |
| Stimatore Sintetico *area-level* | 6,90 | 7,41 | 169,78 | 171,57 | 54,22 | 57,90 | 575,83 | 578,00 |
| Stimatore EBLUP *area-level* | 6,24 | 6,97 | 90,03 | 124,06 | 53,59 | 56,62 | 573,13 | 575,26 |
| Stimatore Spaziale *unit-level* | 4,00 | 6,27 | 127,75 | 152,93 | 30,31 | 48,33 | 328,36 | 338,46 |

### *7.2.3 Confronto tra i livelli di efficienza delle stime "coerenti" e "non coerenti" riferite alle sezioni di censimento*

Altro aspetto analizzato è stato valutare se l'operazione di riproporzionamento delle stime (terzo passo della metodologia proposta nel paragrafo 4.4), necessaria a garantire la coerenza tra le stime per sezione e le stime per area di censimento, produce effetti significativi sull'accuratezza dei metodi sperimentati.

In particolare, confrontando i valori degli indicatori AARB e AARMSE osservati con i vari metodi, nel caso di stime "coerenti" (Tabella 4) e di stime "non coerenti" (Tabella 5), si osserva che la fase di riproporzionamento fa ridurre, in modo lieve, l'accuratezza delle stime; ciò è spiegato dal fatto che il procedimento di riproporzionamento tende a "sporcare" le stime e, quindi, a far incrementare leggermente la variabilità.

### *7.2.4 Analisi dei livelli di efficienza delle stime riferite alle sezioni di censimento classificate per dimensione di popolazione*

In questo ambito, le misure di distorsione e variabilità relative alle stime "coerenti" a livello di sezione di censimento sono sintetizzate per classi di ampiezza demografica delle sezioni. L'analisi comparativa viene effettuata tra i risultati dell'impiego dello stimatore spaziale e dello stimatore ritenuto migliore tra quelli che non considerano la componente spaziale (Carbonetti e Fiorello, 2010), per ciascuna delle due variabili prese in esame: per la stima del

*"numero di occupati"* lo stimatore più efficiente è risultato l'EBLUP *area-level*, mentre, per la stima del *"numero di persone in cerca di occupazione"* è risultato l'EBLUP *unit-level*.

Dal confronto relativo alla variabile *"numero di occupati"* si osserva (Tabella 6) un vantaggio uniforme, sia per minore distorsione che per maggiore efficienza, dall'introduzione della componente spaziale rispetto allo stimatore EBLUP *area-level*; tale vantaggio tende però a ridursi al crescere della dimensione media delle sezioni di censimento.

*Tabella 6* - Valori degli indici AARB, ARRMSE, MARB e MRRMSE dei metodi indiretti, EBLUP (*area-level*) e Spaziale, sperimentati per la stima della variabile *"numero di occupati"* riferita alle sezioni di censimento non vuote (953) di una macro-area del comune di Milano, classificate per dimensione di popolazione. (Censimento 2001)

| *"Occupati"* | *Stimatore Spaziale unit-level* | | | | *Stimatore EBLUP area-level* | | | |
|---|---|---|---|---|---|---|---|---|
| Classi di popolazione (censimento 2001) | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** |
| 0-15 | 21,42 | 27,06 | 126,65 | 151,78 | 26,01 | 30,79 | 89,16 | 122,89 |
| 16-25 | 15,42 | 18,66 | 57,23 | 62,65 | 22,08 | 23,29 | 73,41 | 74,68 |
| 26-50 | 6,86 | 10,09 | 28,47 | 31,61 | 10,08 | 10,97 | 37,78 | 38,13 |
| 51-75 | 4,43 | 7,24 | 14,35 | 17,05 | 7,62 | 8,46 | 19,50 | 19,65 |
| 76-100 | 4,46 | 6,77 | 12,49 | 13,94 | 7,07 | 7,70 | 19,11 | 19,41 |
| 101-150 | 3,20 | 5,58 | 14,82 | 15,73 | 5,37 | 6,05 | 20,80 | 20,96 |
| 151-200 | 2,70 | 4,90 | 13,10 | 14,25 | 4,77 | 5,38 | 20,82 | 20,93 |
| 201-250 | 2,79 | 4,77 | 17,28 | 18,27 | 5,12 | 5,63 | 25,43 | 25,56 |
| 251-300 | 2,64 | 4,51 | 20,11 | 21,01 | 4,82 | 5,29 | 33,30 | 33,47 |
| 301-350 | 2,17 | 4,17 | 13,22 | 14,33 | 3,79 | 4,35 | 19,17 | 19,39 |
| 351-400 | 2,33 | 4,14 | 13,43 | 14,22 | 4,48 | 4,95 | 17,11 | 17,44 |
| Oltre 400 | 1,42 | 3,45 | 5,18 | 6,37 | 3,24 | 3,71 | 13,16 | 13,36 |

Riguardo la variabile *"numero di persone in cerca di occupazione"* (Tabella 7) emerge un risultato simile al caso precedente[16]. Si assiste ad un diffuso miglioramento, sia in termini di distorsione che di variabilità, dall'adozione dello stimatore spaziale rispetto allo stimatore EBLUP *unit-level*; l'entità del vantaggio sembrerebbe però dipendere molto dalla dimensione media della sezione di censimento, secondo un andamento non regolare.

[16] Si fa presente che, per la variabile *"numero di individui in cerca di occupazione"*, i valori degli indici riferiti alle prime due classi ("0-15" ; "16-25") non sono ritenuti rappresentativi per l'esiguo numero di casi osservati.

*Tabella 7* - Valori degli indici AARB, ARRMSE, MARB e MRRMSE dei metodi indiretti, EBLUP (*unit-level*) e Spaziale, sperimentati per la stima della variabile *"numero di persone in cerca di occupazione"* riferita alle sezioni di censimento non vuote (953) di una macro-area del comune di Milano, classificate per dimensione di popolazione. (Censimento 2001)

| *"Persone in cerca di occup."* | *Stimatore Spaziale unit-level* | | | | *Stimatore EBLUP unit-level* | | | |
|---|---|---|---|---|---|---|---|---|
| Classi di popolazione (censimento 2001) | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** |
| 0-15 | 52,83 | 65,14 | 60,42 | 67,55 | 85,45 | 85,47 | 93,67 | 93,68 |
| 16-25 | 48,83 | 65,12 | 53,25 | 67,14 | 70,61 | 70,71 | 75,75 | 75,83 |
| 26-50 | 21,49 | 51,44 | 47,23 | 60,70 | 38,49 | 40,36 | 79,90 | 80,00 |
| 51-75 | 25,91 | 49,60 | 57,83 | 83,76 | 34,97 | 38,04 | 72,41 | 74,28 |
| 76-100 | 28,36 | 52,72 | 67,75 | 86,66 | 46,04 | 48,82 | 105,68 | 108,45 |
| 101-150 | 40,76 | 60,94 | 187,41 | 197,49 | 61,01 | 64,90 | 287,25 | 289,59 |
| 151-200 | 36,19 | 55,03 | 225,33 | 239,92 | 55,05 | 60,17 | 358,10 | 360,42 |
| 201-250 | 38,25 | 56,58 | 276,74 | 288,31 | 56,88 | 63,43 | 420,00 | 423,93 |
| 251-300 | 23,64 | 42,23 | 374,46 | 385,57 | 35,39 | 43,33 | 495,86 | 501,51 |
| 301-350 | 21,00 | 39,45 | 114,36 | 127,07 | 30,07 | 38,72 | 196,27 | 200,05 |
| 351-400 | 22,24 | 39,38 | 144,75 | 155,58 | 32,55 | 41,17 | 235,55 | 239,14 |
| Oltre 400 | 20,84 | 37,96 | 165,78 | 177,07 | 30,41 | 39,31 | 215,62 | 221,89 |

# 8 CONSIDERAZIONI SULL'IMPIEGO DELLO STIMATORE PER PICCOLE AREE BASATO SU AUTOCORRELAZIONE SPAZIALE

Questo lavoro poggia sull'ipotesi verosimile che le manifestazioni di un fenomeno in un'area siano interconnesse a quelle relative alle aree limitrofe, più di quanto possano esserlo con quelle relative alle aree più distanti. In tal caso, è sembrato opportuno studiare la possibilità di impiegare metodi di stima indiretta basati su modelli che tengono conto di una struttura di autocorrelazione spaziale tra le osservazioni considerando la distanza tra i domini di interesse per le stime; in tal modo, le osservazioni rilevate in uno specifico dominio saranno considerate legate a quelle rilevate nei domini geograficamente più vicini, secondo un'opportuna struttura di autocorrelazione spaziale.
A tal riguardo è stata implementata, come soluzione alternativa, la metodologia di stima per piccole aree basata sulla componente spaziale, per valutare la eventualità di determinare stime più affidabili su domini di elevato dettaglio territoriale.

Le sperimentazioni hanno interessato due differenti livelli territoriali: uno relativo ai domini sub-comunali a valenza amministrativo/funzionale non pianificati dal disegno di indagine,

l'altro alle sezioni di censimento di centro. I risultati hanno messo in evidenza che la disponibilità di informazioni di natura territoriale offre un contributo rilevante per migliorare l'accuratezza delle stime, anche se con modalità diverse.
Nella prima situazione (Tabelle 2-3), il valore della frequenza assoluta da stimare e la dimensione del dominio di riferimento sembrano influenzare i possibili vantaggi inferenziali derivanti dall'introduzione della componente spaziale nel modello di stima alla base del metodo per piccole aree.
Nel caso di stime riferite alle sezioni di censimento (Tabelle 4-5) lo stimatore spaziale permette di determinare stime diffusamente più accurate, per una forte riduzione della distorsione dovuta all'impiego dell'informazione territoriale. Calcolando, però, la variabilità come differenza tra l'MSE e la distorsione, il metodo spaziale risulta un pò più variabile rispetto agli altri metodi indiretti, per l'incremento di variabilità dovuto alla presenza della struttura di autocorrelazione spaziale da stimare nel modello. Questi dati sono confermati anche da risultati ottenuti in altri studi (D'Alò *et al.*, 2011).

Inoltre, la procedura di riproporzionamento sembra portare ad un leggero incremento della variabilità, di entità però tale da non influire sulle indicazioni date fino ad ora sull'impiego dello stimatore spaziale.

Riguardo infine le analisi condotte per dimensione delle sezioni (Tabelle 6-7), si osserva un comportamento regolare per la stima del *"numero di occupati"* al crescere della dimensione del dominio (in termini di popolazione). Invece, per la stima del *"numero di persone in cerca di occupazione"* (variabile che fa registrare generalmente poche unità a livello di sezione di censimento) si osserva un comportamento irregolare rispetto alla dimensione media delle sezioni. Poiché questo risultato potrebbe far pensare ad un comportamento anomalo dello stimatore, si è ritenuto doveroso ripercorrere le analisi classificando le sezioni in base al valore da stimare, anziché alla dimensione di popolazione.

La Tabella 8 ripropone le misure di distorsione e variabilità relative alle stime "coerenti" a livello di sezione di censimento, sintetizzando però i risultati della sperimentazione in classi di frequenza assoluta (oggetto di stima). Dall'analisi dei valori continua a non emergere una certa regolarità di comportamento delle misure calcolate per valutare le proprietà statistiche prese a riferimento; se però si desume la variabilità come differenza tra l'errore quadratico medio e la distorsione (misura fornita dalla quantità [ARRMSE-AARB]) si osserva una riduzione molto più regolare della variabilità all'aumentare della frequenza assoluta da stimare.

*Tabella 8* - Valori degli indici AARB, ARRMSE, MARB e MRRMSE del metodo Spaziale, sperimentato per la stima della variabile *"numero di persone in cerca di occupazione"* riferita alle sezioni di censimento non vuote (953) di una macro-area del comune di Milano, classificate per valore della frequenza oggetto di stima. (Censimento 2001)

| *"Persone in cerca di occupazione"* | *Stimatore Spaziale unit-level* | | | |
|---|---|---|---|---|
| Classi di frequenza assoluta (censimento 2001) | **AARB** | **ARRMSE** | **MARB** | **MRRMSE** |
| 1-2 | 53,62 | 374,46 | 72,59 | 385,57 |
| 3-5 | 16,13 | 91,03 | 37,63 | 102,42 |
| 6-8 | 11,07 | 39,01 | 30,31 | 50,96 |
| 9-12 | 12,26 | 36,77 | 27,66 | 42,91 |
| 13-16 | 20,72 | 32,68 | 29,84 | 36,40 |
| 17-20 | 22,67 | 32,84 | 29,97 | 36,77 |
| > 20 | 25,70 | 28,40 | 31,40 | 33,71 |

## 9 CONCLUSIONI

La nuova strategia metodologica scelta per l'esecuzione del 15° Censimento generale della popolazione ha richiesto un'attenta valutazione delle conseguenze sulla produzione e sulla diffusione dei risultati finali. In particolare, l'adozione delle tecniche di campionamento, nei comuni più grandi, porterà alla determinazione di stime affidabili riferite solo alle aree di censimento, domini sub-comunali di elevato dettaglio territoriale disegnati secondo criteri e, in molti casi, con limiti geografici differenti rispetto ai domini sub-comunali definiti dai comuni stessi per motivi amministrativi o funzionali.

La necessità di rispondere ad un fabbisogno di informazione statistica censuaria anche per i domini non considerati dal piano di campionamento, ha spinto l'interesse verso lo studio di soluzioni metodologiche con l'obiettivo di garantire la continuità dell'informazione riferita a tali domini, rispetto a quanto è stato prodotto e diffuso nel censimento del 2001.

L'approccio metodologico seguito tiene conto sia dell'esigenza di assicurare la coerenza tra le stime riferite a livelli territoriali differenti che dell'opportunità offerta dai metodi di stima per piccole aree, che risolvono i problemi di un'eventuale non rappresentatività del campione nei domini non pianificati dal disegno e per i quali si vuole fare inferenza.

In questo lavoro, oltre ai metodi per piccole aree basati su effetti lineari misti (sintetico, EBLUP), è stata approfondita l'analisi di un metodo che estende tale classe di modelli al caso in cui si introduce una componente aggiuntiva che tiene conto della correlazione spaziale esistente tra le unità che entrano nel modello.

Il ricorso ai metodi di stime per piccole aree comporta il fatto di accettare un certo livello di distorsione delle stime compensato però da una diminuita varianza e conseguentemente da un livello più basso dell'errore quadratico medio. Quindi, proprio per studiare le proprietà statistiche delle stime ottenute con l'impiego di tali metodi, si è proceduto ad uno studio sperimentale su dati del censimento della popolazione del 2001 e riferiti ad una macro-area delimitata nel comune di Milano. L'ambito di analisi è stato fissato in modo opportuno per valutare il comportamento campionario delle stime riferite a due variabili legate al mercato del lavoro (*"numero di occupati"* e *"numero di persone in cerca di occupazione"*), per due livelli territoriali di analisi (i *"nuclei di identità locale"* e le sezioni di censimento).

L'impiego dei metodi indiretti nella metodologia proposta offre la possibilità di ottenere stime con livelli di accuratezza migliori di quelli ottenibili con i metodi standard. I maggiori vantaggi si osservano nel caso di stime riferite a domini di dimensioni ridotte o per frequenze assolute piccole; nel caso di stime riferite a domini ampi o per frequenze assolute grandi, l'adozione dei metodi diretti, in presenza di una frazione di campionamento molto elevata, risulta quasi sempre la soluzione più vantaggiosa per una ridotta distorsione.
Riguardo le stime riferite alle sezioni di censimento, sono stati evidenziati i migliori livelli di efficienza con l'impiego degli stimatori indiretti; in particolare, il metodo spaziale si evidenza come la migliore soluzione per produrre stime, riferite ai domini più piccoli, sufficientemente accurate.

Lo studio presentato in questo documento rappresenta un ulteriore passo in avanti verso la messa a punto di una proposta metodologica che risponda alla necessità di produrre stime affidabili per tutti i domini non considerati dal disegno di campionamento che si adotterà nell'ambito della realizzazione del censimento del 2011.
I risultati ottenuti evidenziano le grandi opportunità offerte dall'utilizzo dell'informazione territoriale e dalla possibilità di geo-codificare le unità nelle sezioni di censimento; i dati territoriali, offrono così la possibilità di migliorare la qualità dell'informazione prodotta per i domini non previsti dal disegno di indagine, anche nei casi in cui questa riguarda fenomeni rari o se è riferita ad ambiti territoriali molto piccoli.

In conclusione, anche se la nuova strategia censuaria non procederà alla raccolta dei dati in modo completo su tutti gli individui, si dovranno sviluppare ulteriori soluzioni metodologiche rivolte a produrre stime con accuratezza accettabile per tutti i livelli territoriali; questo porterà così ad una più ampia e diversificata disponibilità dei risultati del censimento.

ABSTRACT

Italian National Institute of Statistics has considered to use sampling techniques in the strategy for the 2011 Italian Population and Housing Census. The approach will consist in the simultaneous use of two different forms: the short form collecting only demographic and housing variables; the long form collecting the overall set of census variables.
The estimation procedure is based on a simple random design for the selection of private households samples from population registers and on calibrated estimators. The sampling strategy will regard municipalities with population over 20,000 inhabitants and the sample will be designed to produce very accurate estimates referred to Census Areas, sub-municipal domains with about 15,000 people.
In the 2001 census, sub-municipal data have been provided for domains defined by local government for administrative or functional aims.
The 2011 census strategy could assume some problems to provide data referred to the same sub-municipal domains, different to the census areas.
For this aim, a methodological approach has been proposed in order to produce estimates to different sub-municipal levels with good accuracy. The approach involves the adoption of small area estimators and takes into account the need to respect a constraint of coherence for estimates referred to different territorial levels.
In particular, the improvement given by small area estimators through the introduction of spatial relationship in the linear mixed model has been studied in this work. The variance-covariance matrix of the random effects has been built up with an auto-correlation structure, depending on the distance between areas.
A simulation study has carried out on 2001 census data, in order to assess the accuracy levels of estimates referred to some sub-municipal domains chosen in the municipality of Milan.

## 10 BIBLIOGRAFIA


Abbatini D., Cassata L., Martire F., Reale A., Ruocco G., 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 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., 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 2007.

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

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.

Borrelli F., Carbonetti G., De Felici L., Fiorello E., Marrone M. (2011) La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento *Istat Working Papers*, N. 2/2011. http://www.istat.it/dati/pubbsci/working_papers/wp/wp_2011/Istat_Working_Papers_2_2011.pdf

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

Carbonetti G., Fortini M., F. Solari (2008b) Innovations on methods and survey process for the 2011 Italian population census, *Proceedings of the European Conference on Quality in Official Statistics*, Roma 8-11 Luglio 2008.

Carbonetti G., Verrascina M. (2009) Accuracy evaluation of Nuts level 2 hypercubes with the adoption of a sampling strategy in the 2011 Italian population census. Group of Experts on Population and Housing Censuses. UN, Geneva (Switzerland), October 2009. ECE/CES/GE.41/2009/10.

Carbonetti G., Fiorello E. (2010) La produzione di informazione statistica a livello territoriale sub-comunale: possibili cambiamenti indotti dalla strategia proposta per il censimento della popolazione e delle abitazioni del 2011. Relazione presentata alla *XXXI Conferenza Italiana di Scienze Regionali*, Aosta 20-22 Settembre 2010.

Cocchi D. (2007) Uso dei campioni nelle rilevazioni censuarie, *Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni"*. CNR, Roma, Novembre 2007.

Crescenzi F., Fortini M., Gallo G., 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
http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2009/doc6_2009.pdf

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

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

D'Alò M., Di Consiglio L., Falorsi S., Ranalli, M.G., Solari F. (2011). Use of spatial information for unemployment rate at sub-provincial areas in Italy. Journal of the Indian Society of Agricultural Statistics (in corso di pubblicazione)

Deville J.C., Särndal, C.E. (1992) *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, vol. 87, pp. 367-382

EURAREA Consortium (2004) PROJECT REFERENCE VOLUME, Vol. 1
http://www.statistics.gov.uk/eurarea/

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