GIORNATE DELLA RICERCA IN ISTAT | 10-11 NOVEMBRE 2014

SESSIONE III
INTEGRAZIONE E USO DI DATI AMMINISTRATIVI PER FINI STATISTICI

Un'analisi multilivello dell'errore di copertura delle anagrafi comunali nel 15° Censimento della Popolazione e delle Abitazioni

Luca Mancini, Simona Toti e Alessandra Ronconi

Istat

Un'analisi multilivello dell'errore di copertura delle anagrafi comunali nel XV Censimento della Popolazione e delle Abitazioni

Luca Mancini Simona Toti e Alessandra Ronconi

Istat

lmancini@istat.it, toti@istat.it, alronconi@istat.it

Sommario

Il lavoro utilizza dati individuali del 2011 di fonte sia censuaria sia amministrativa per individuare, a fini predittivi, le determinanti dell'errore di copertura delle liste anagrafiche comunali (LAC), definito come scostamento tra la popolazione anagraficamente residente e la popolazione obiettivo del censimento in un dato comune. I risultati mostrano come la previsione della popolazione obiettivo sia decisamente accurata quando le probabilità individuali di sotto e sovracopertura sono stimate con modelli logistici multilivello che tengono conto esplicitamente della struttura gerarchica dei dati (individui entro comune). In questo modo è infatti possibile catturare l'eterogeneità non osservata della copertura delle LAC dei comuni italiani medio-piccoli. I risultati dell'analisi appaiono rilevanti sia per la progettazione delle indagini socioeconomiche Istat sulle famiglie che dal 2011 utilizzano le LAC come universo campionario sia per la pianificazione del prossimo censimento italiano della popolazione nel quale le LAC avranno un ruolo centrale.

Parole chiave: modelli multilivello, errore di copertura, censimento, dati amministrativi

Abstract

The paper uses 2011 census as well administrative data at the individual level to ascertain, for prediction purposes, the determinants of municipal population registers' (henceforth LACs) coverage error defined as the discrepancy between the registered and the census target populations in a given municipality. The results show how the prediction of the target population is significantly more accurate when the individual probabilities of under and overcounting are estimated via multilevel logistic models that account explicitly for the hierarchical structure of the data (individuals nested within municipalities). Indeed the model is able to factor in the unobserved heterogeneity of LACs'coverage for Italian medium-to-small sized municipalities. The results of the analysis are expected to be relevant not only for the design of Istat's socio-economic household surveys whose samples have been drawn from the LACs since 2011, but also for the set up of the next Italian population census where the LACs will play a pivotal role.

Key words:, multilevel model, coverage error model, population census, population registers

Introduzione

Le Liste Anagrafiche Comunali (LAC) hanno avuto un ruolo importante nell'ultima tornata censuaria. Milioni di famiglie hanno ricevuto per posta il questionario all'indirizzo di residenza anagrafica. Il nuovo sistema informatico di gestione della rilevazione ha, inoltre, reso possibile il confronto in tempo reale tra le informazioni individuali acquisite al censi-

mento e quelle presenti in anagrafe (Mancini, 2011). Ciò ha consentito di determinare per ciascun comune, in modo pressoché immediato, sia la popolazione abitualmente dimorante non presente nei registri anagrafici (sottocopertura delle LAC) sia la popolazione iscritta in anagrafe ma irreperibile al censimento (sovracopertura delle LAC).

Il confronto censimento-anagrafe ha evidenziato, a livello nazionale, un errore di copertura complessivo non superiore al 5% della popolazione legale. Questo risultato è stato accolto come un ritorno positivo dell'investimento fatto dall'Istituto Nazionale di Statistica per acquisire e rendere le LAC fruibili a fini censuari, tanto che dal 2011 l'Istituto ha scelto di utilizzarle anche per l'estrazione dei campioni di tutte le indagini socio-economiche sulle famiglie (Ceccarelli e Rosati, 2014). Alla luce delle note criticità di un censimento di tipo tradizionale decennale come l'obsolescenza informativa e gli alti costi di esercizio, questi risultati fanno presagire un ruolo sempre più centrale delle LAC anche a fini censuari. E' attualmente allo studio l'ipotesi di effettuare un censimento puramente da registro per quei comuni i cui archivi anagrafici forniscono adeguate garanzie di copertura della popolazione obiettivo (Calzaroni *et al.*, 2013).

Se a livello aggregato la dimensione dell'errore di copertura delle LAC appare tollerabile, a livello disaggregato, sia territoriale che di specifiche sottopopolazioni (definite ad esempio per sesso, età e cittadinanza), essa manifesta una forte eterogeneità. E' noto, ad esempio, che le LAC sono carenti nel rappresentare la popolazione straniera abitualmente dimorante in Italia (Fortini e Gallo, 2009). Il primo obiettivo della ricerca è individuare, attraverso modelli di regressione, le principali determinanti dell'errore di copertura delle LAC per caratterizzare sia gli individui sia i comuni più a rischio di mancata o errata copertura. L'analisi si prefigge di contribuire alla letteratura sulle popolazioni difficili da contare o "*hard to count*" che tipicamente informa la progettazione delle indagini di copertura dei censimenti di popolazione (PES) (Alho *et al.*, 1993). Il secondo obiettivo è quello di valutare il potere predittivo di tali modelli inteso come capacità di riprodurre statisticamente la popolazione obiettivo del censimento a partire da quella anagrafica mediante modelli cattura-ricattura dove gli individui sono "catturati" la prima volta dalla LAC e la seconda al censimento (Wolter, 1986).

Il modello

L'errore di copertura delle anagrafi è legato sia alle scelte dei singoli individui sia alle caratteristiche del comune in cui essi risiedono. Ad esempio, una persona può decidere di vivere in un comune senza essere iscritta in quell'anagrafe o, per converso, mantenere la residenza anagrafica in un comune sebbene abiti stabilmente altrove. Può, altresì, accadere che per ragioni di natura geografica, economica, ambientale o demografica l'anagrafe di un certo comune sia particolarmente esposta a problemi di copertura. Infine, la qualità di un registro anagrafico dipende necessariamente dal modo in cui esso viene mantenuto dagli uffici comunali preposti, in termini sia di standard di completezza e accuratezza delle informazioni individuali sia di tempestività nell'aggiornamento delle variazioni anagrafiche.

Modelli multilivello gerarchici che considerano gli individui "annidati" entro il comune di appartenenza costituiscono, dunque, un punto di partenza naturale per lo studio del fenomeno. Formalmente,

$$\text{Logit } P(Y = 1 | X = x) = a + a_i + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \qquad (1)$$

dove Y è la variabile risposta dicotomica (alternativamente sotto o sovracopertura), ai l'intercetta casuale relativa al comune i-esimo, b1, …., bk il vettore dei coefficienti di re-

gressione relativi ai cosiddetti "effetti fissi" e X1, …., Xk il profilo o vettore delle modalità dei k regressori rilevati su un dato individuo e descritti nella Tavola 1.

Tavola 1 – Descrizione delle variabili

Variabile	Modalità e descrizione
sotto/sovra	1 se l'individuo è sotto/sovracoperto, 0 altrimenti
sesso	1 se femmina, 0 se maschio
cittad	1 se straniero, 0 se italiano
monocomponente	1 se individuo in famiglia anagrafica monocomponente, 0 altrimenti
eta_i	1 se l'individuo è nella classe di età i, 0 altrimenti; i=0-18,19-40,41-70,71+
distanza	distanza euclidea (km) del comune dal capoluogo di regione
lac	numero di individui iscritti nella LAC del comune
t_cf	tasso di codici fiscali errati o mancanti nella LAC del comune
t_citt	tasso di cittadini stranieri residenti nel comune
t_monocomponente	tasso di individui residenti nel comune in famiglia anagrafica monocomponente
t_anziani	tasso di individui residenti di età superiore a 70 anni
t_lavout	tasso di individui residenti che lavorano in altro comune
cittad*t_citt	interazione tra cittadinanza e tasso di stranieri residenti nel comune

Con riferimento al secondo obiettivo, il modello di ricostruzione della popolazione censita a partire da quella anagrafica è specificato come segue:

$$\hat{N}_i = \sum_{\underline{x}} \hat{N}_{\underline{x_i}} = \sum_{\underline{x}} \hat{w}_{\underline{x_i}} * LAC_{\underline{x_i}} = \sum_{\underline{x}} \frac{1-\hat{p}_{sovra,\underline{x_i}}}{1-\hat{p}_{sotto,\underline{x_i}}} * LAC_{\underline{x_i}} \qquad (2)$$

dove:
- $\hat{p}_{sotto,\underline{x_i}}$ e $\hat{p}_{sovra,\underline{x_i}}$ sono rispettivamente il rischio stimato per un individuo di essere "sottocoperto" o "sovracoperto", dato che egli presenta determinati valori delle variabili di regressione;
- $\underline{x_i}$ è il profilo o vettore $x_1, x_2, \ldots, x_k$ delle modalità dei *k* regressori, per il comune i-esimo;
- $\hat{N}_{\underline{x_i}}$ è il numero stimato di individui abitualmente dimoranti nel comune i-esimo con profilo $\underline{x_i}$;
- $\hat{w}_{\underline{x_i}}$ è il peso stimato del profilo $\underline{x_i}$ in funzione delle probabilità di sotto e sovracopertura;
- $LAC_{\underline{x_i}}$ è il numero di individui iscritti in anagrafe nel comune i-esimo con profilo $\underline{x_i}$.

Così se, ad esempio, un individuo ha - condizionatamente al proprio profilo in termini di sesso, età, cittadinanza, composizione del nucleo familiare anagrafico- una probabilità relativamente elevata di non essere registrato in anagrafe pur essendo stato censito nel comune, gli verrà attribuito un peso superiore a 1, ritoccando così verso l'alto la popolazione residente del comune. Analogamente per la sovracopertura dove le LAC sono, però, corrette verso il basso. L'equazione 2 si ispira ai modelli di tipo cattura-ricattura o dual system originariamente concepiti per stimare la consistenza ignota delle popolazioni animali. Per una rassegna di questi modelli si rimanda a Wolter (ibid.), mentre per un'applicazione specifica del modello logistico a dati individuali di fonte censuaria si veda (Alho, 1986) e (Alho *et al.*, ibid.).

Dati

I dati utilizzati in questo lavoro si riferiscono a informazioni a livello individuale e municipale sia di fonte censuaria che amministrativa. L'analisi è circoscritta ai comuni italiani con una popolazione censita di almeno 1000 e non superiore a 50000 abitanti suddivisi in tre gruppi secondo le seguenti soglie di ampiezza demografica: 1001-5000, 5001-10000 e 10001-50000 abitanti. Per ciascuna classe di ampiezza è stato estratto un campione casuale a due stadi di individui (unità finali) entro comune (unità primarie).

Risultati

I coefficienti di regressione, con la relativa significatività statistica, stimati dal modello (1) sono riportati nella Tavola 2. Il segno dei coefficienti e dunque la direzione in cui le caratteristiche, individuali e comunali, influenzano l'errore di copertura, generalmente non cambia al variare della classe d'ampiezza demografica del comune e della componente dell'errore di copertura. Tuttavia, emergono differenze rilevanti in termini sia di ampiezza dei coefficienti sia della loro significatività statistica. Ad esempio, essere femmina diminuisce significativamente il rischio di sovracopertura ma non quello di sottocopertura, dove l'effetto è anzi di segno opposto anche se solo per i comuni della classe 1001-5000. Gli individui di età inferiore a 19 anni hanno una probabilità di sottocopertura sensibilmente più alta rispetto ad individui di età compresa tra 41 e 70 anni (categoria di riferimento, omessa) ma lo stesso effetto appare molto più debole sia per intensità che per significatività tra le determinanti della sovracopertura, in particolare nei comuni della classe 1001-5000. Il tasso comunale di individui che lavorano fuori comune ha un effetto positivo sul rischio di sottocopertura, ma solo per i comuni della classe 10001-50000. L'effetto cambia di segno sul rischio di sovracopertura, anche se solo per i comuni con popolazione compresa tra 1001 e 10000. E' interessante notare l'effetto dell'interazione tra la cittadinanza dell'individuo ed il tasso comunale di cittadini stranieri. La Tavola 2 suggerisce che, per la classe di ampiezza 1001-5000, vivere in un comune con una presenza straniera relativamente elevata riduce notevolmente il rischio di sottocopertura per un cittadino straniero. L'effetto è simile sulla probabilità di sovracopertura, ma solo per i comuni con un numero di abitanti compreso tra 10001 e 50000. Ciò potrebbe suggerire come la presenza di altri concittadini o più in generale di una o più comunità di cittadini stranieri residenti, non necessariamente strutturate sotto forma di associazioni formalmente riconosciute ma anche costituite da reti spontanee di contatti, contribuisca a rendere più visibile e stabile la presenza straniera in un dato territorio comunale.
I risultati sulla sensibilità (proporzione di "sotto" e "sovracoperti" correttamente classificati) indicano una capacità predittiva, sia per la sotto che per la sovracopertura, compresa tra 72 e 76 per cento. Inoltre, si nota come la varianza dell'intercetta aleatoria, e di conseguenza la quota di varianza della variabile dipendente riconducibile all'eterogeneità tra comuni rispetto alla sotto e sovracopertura delle LAC (ρ), diminuisca all'aumentare della dimensione demografica e del numero di comuni inclusi nel campione.

Tavola 2 – I coefficienti di regressione

	sottocopertura			sovracopertura		
Profilo	1001-5000	5001-10000	10001-50000	1001-5000	5001-10000	10001-50000
Femmina[(a)]	*0,12*	*0,03*	*0,00*	-0,17	-0,22	-0,24
eta_1	1,36	0,76	0,96	*-0,01*	0,13	0,25
eta_2	0,99	0,92	0,85	0,49	0,50	0,53
eta_4	-1,30	-1,23	-0,98	-0,53	-0,61	-0,58

monocomp.	1,93	1,54	1,69	1,30	1,31	1,18
cittad	1,78	1,29	1,52	2,58	2,54	2,51
citXtcit	-5,48	*-0,17*	*-2,20*	*-1,04*	*-0,58*	-2,32
t_citt	*0,99*	*-1,31*	*1,22*	*0,06*	*-0,39*	*1,35*
t_mono	5,38	5,93	6,33	3,60	9,13	3,73
t_cf	0,03	0,03	0,01	0,04	0,03	*-0,01*
t_anziani	-5,15	-5,32	-6,18	-4,64	-8,05	-5,33
t_lavout	*-0,16*	*0,66*	2,43	-1,64	-2,63	*-0,39*
distanza	0,00	0,00	*0,00*	0,00	*0,00*	*0,00*
lac	*0,00*	*0,00*	0,00	*0,00*	*0,00*	0,00
a	-6,57	-6,13	-6,60	-4,43	-4,39	-4,46
AIC	11642	13761	16087	30456	36859	46671
sensibilità	0,76	0,72	0,72	0,77	0,76	0,73
var(a_i)	0,61	0,39	0,25	0,46	0,38	0,35
ρ	0,19	0,16	0,13	0,17	0,16	0,15
N	195860	194551	191433	198524	198162	196203
n	746	299	247	746	299	247

(a) I coefficienti statisticamente non significativi al 5% sono indicati in corsivo.

I coefficienti di regressione riportati nella Tavola 2 forniscono indicazioni sia sulla direzione sia sulla dimensione dell'associazione tra i regressori e la variabile dipendente. Al fine di quantificare l'intensità di tale associazione, la Tavola 3 riporta le probabilità di sottocopertura predette dal modello (1) per alcuni profili specifici. Ad esempio, un cittadino italiano di sesso maschile, di età compresa tra 41 e 70 anni che vive in famiglia anagrafica pluricomponente in un comune di classe 1001-5000 abitanti caratterizzato da valori campionari medi delle variabili di secondo livello sia fisse che casuali ($a_i = 0$), ha una probabilità stimata pressoché nulla di essere "sottocoperto". Variando in modo cumulativo le sole caratteristiche individuali –ovvero considerando una donna straniera di età compresa tra 19 e 40 anni che vive in famiglia anagrafica monocomponente nello stesso comune dell'individuo di riferimento - il rischio di sottocopertura sale a 8,6%. Si noti, inoltre, il contributo dell'effetto casuale di comune: il valore di 8,6% più che raddoppia (si dimezza) facendo variare l'intercetta a_i di una deviazione standard (σ) sopra (sotto) il suo valor medio. Gli effetti sulle probabilità di sovracopertura (non riportate), pur con alcune eccezioni di rilievo, sono sostanzialmente simili sia per direzione che per significatività statistica a quelli stimati per la sottocopertura.

Tavola 3 – Probabilità (%) stimata di sottocopertura per alcune categorie di individui

	1001-5000			5001-10000			10001-50000		
Profilo	-σ	0	+σ	-σ	0	+σ	-σ	0	+σ
Base[a]	0,0	0,1	0,3	0,0	0,2	0,4	0,1	0,2	0,5
Femmina	0,0	0,1	0,3	0,1	0,2	0,5	0,1	0,3	0,5
Monocomponente	0,4	0,8	2,1	0,5	1,1	2,2	0,7	1,4	2,8
Età 19-40	1,0	2,3	5,3	1,3	2,7	5,6	1,6	3,2	6,3
cittad@(t_citt)[b]	3,8	8,6	18,6	5,1	10,2	19,5	6,0	11,5	20,8

(a) Italiano maschio di età 41-70 con nucleo familiare anagrafico pluricomponente. Le altre caratteristiche sono poste pari al loro valore medio nel campione. Le probabilità sono tutte statisticamente significative all'1%.
(b) La notazione sintetica cittad@(t_citt) indica l'effetto individuale di essere cittadino straniero più la sua interazione con il tasso comunale di cittadini stranieri impostato al suo valor medio campionario.

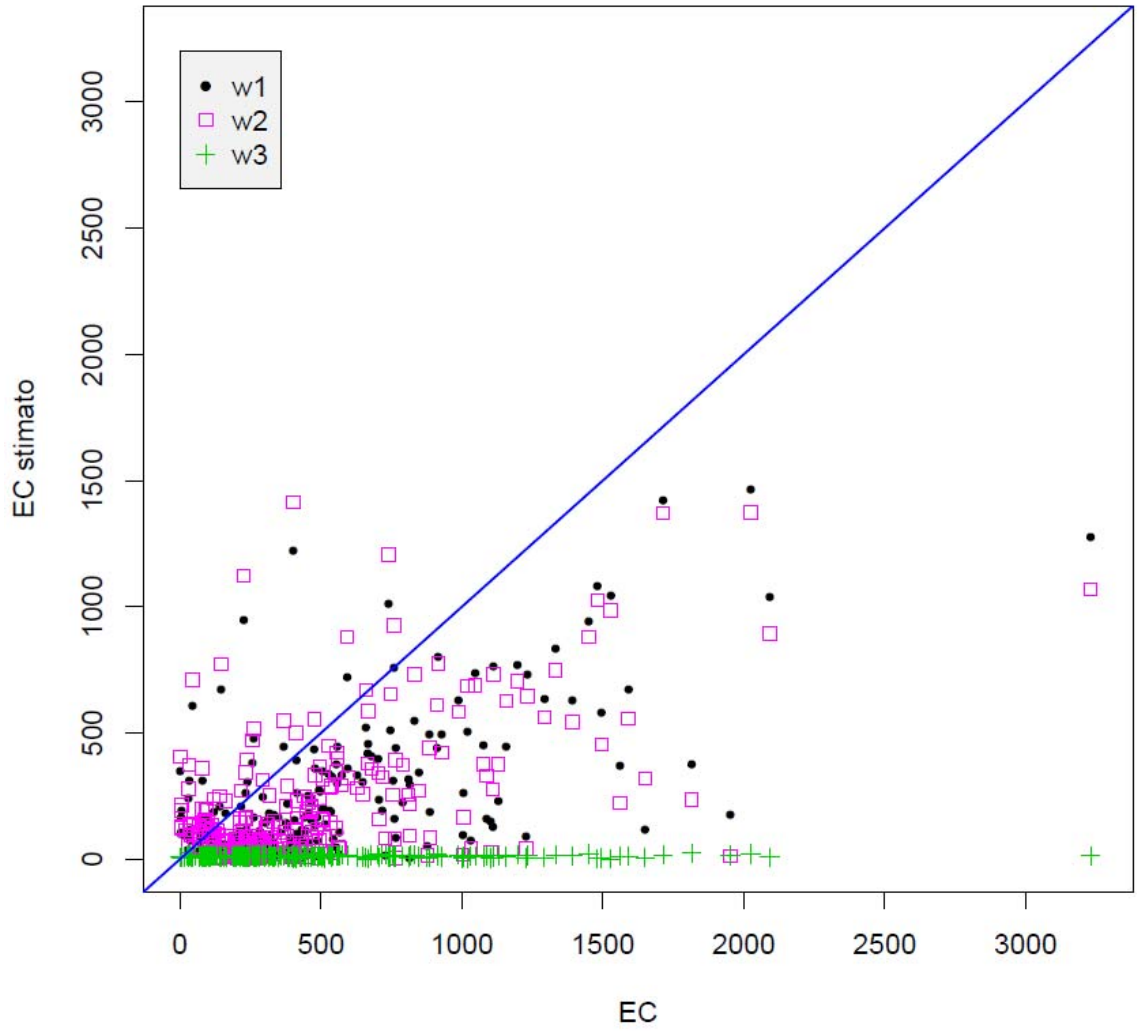
La ricostruzione del censimento

Il secondo obiettivo del presente studio è quello di ricostruire la popolazione censita da quella anagrafica, opportunamente riponderata sulla base delle probabilità individuali di sotto e sovracopertura predette dal modello (1). In questo contributo riportiamo i risultati della ricostruzione per i comuni campione. Per una previsione sui comuni non campionati rimandiamo all'articolo esteso (Mancini e Toti, 2014).

L'obiettivo è capire se le LAC "pesate" ($\hat{N}_i$) siano più vicine al censimento (N_i) di quanto non lo siano le LAC originarie. Ciò è equivalente a verificare se l'errore di copertura delle LAC corrette da modello è, in valore assoluto, inferiore a quello osservato al confronto censimento-anagrafe. In termini formali si vuole confrontare $|\hat{EC}| = |\hat{N}_i - N_i|$ con $|EC| = |LAC_i - N_i|$. Si è scelto di mostrare i risultati sotto forma di scarti tra conteggi di popolazione (errore) per i vantaggi che tale rappresentazione offre in termini grafici. La Figura 1 confronta, attraverso grafici di dispersione, l'errore di copertura osservato con l'errore corretto da modello per i comuni nella fascia 10001-50000. I risultati per le altre due classi sono simili. La bisettrice rappresenta il luogo dei punti (comuni) per i quali il modello e la LAC si equivalgono, mentre i punti sotto la diagonale sono comuni per i quali il modello "fa meglio" della LAC. A fini comparativi, la Figura 1 riporta inoltre la capacità predittiva di due modelli ad effetti fissi, uno rappresentato da una regressione logistica standard (variabile risposta identicamente ed indipendentemente distribuita, condizionatamente al valore dei regressori) e l'altro da una regressione logistica "marginale" o "*population-average*" dove la variabile risposta è i.i.d. solo per gli individui di uno stesso comune.

Si nota, in generale, come almeno l'80% dei comuni campionati si trovi sotto la bisettrice, indipendentemente dalla loro dimensione demografica. Inoltre, laddove il modello "fa peggio", si tratta di comuni con un errore di copertura relativamente piccolo. E' inoltre evidente la superiorità della previsione basata sul modello ad effetti misti (1) rispetto ai due modelli ad effetti fissi. Sebbene anche questi ultimi riducano generalmente l'errore di copertura osservato, non mancano casi di comuni per i quali i modelli ad effetti fissi tendono a sovra-correggere le LAC in misura significativa. Ciò implica che le covariate, sia individuali che comunali, considerate nei modelli di previsione, spiegano in modo parziale l'errore di copertura dei registri anagrafici. Esiste, in altre parole, una forte eterogeneità non osservata nella capacità dei registri anagrafici comunali di rappresentare in modo fedele, in un dato istante temporale, la popolazione abitualmente dimorante in un dato comune.

Figura 1 – $|\hat{EC}|$ contro $|EC|$ per modello: comuni con 10001-50000 abitanti: con w1 sono indicate le stime del modello standard, con w2 quelle del modello *population average*, con w3 quelle del modello multilivello.



Considerazioni conclusive

Il lavoro si propone di individuare le determinanti dell'errore di copertura dei registri anagrafici rispetto alla popolazione obiettivo del censimento. Modelli logistici multilivello stimati con dati individuali sia di fonte amministrativa che censuaria rivelano come alcune sottopopolazioni (i cittadini stranieri, gli individui giovani, i nuclei familiari monocomponenti) sono sensibilmente più a rischio di mancata o errata copertura. L'analisi evidenzia, inoltre, come la qualità del dato anagrafico possa essere notevolmente migliorata attraverso una opportuna riponderazione della popolazione residente che tenga conto dei profili di rischio individuale di sotto e sovracopertura. In particolare, modelli multilivello che considerano esplicitamente la struttura gerarchica dei dati (individui entro comune) assicurano una capacità predittiva decisamente più alta. La superiorità di questi modelli deriva dal contributo delle intercette aleatorie interpretabili come una misura del "valore aggiunto" del singolo comune al rischio di sotto e sovracopertura altrimenti attribuibile alle sole caratteristiche socio-demografiche dei suoi abitanti. Nell'ipotesi plausibile che tale "effetto comune" rimanga stabile nel tempo, l'utilizzo di modelli gerarchici di previsione come strumento per verificare la capacità dei registri anagrafici comunali di rispecchiare più o meno fedelmente la popolazione abitualmente dimorante nel territorio appare molto promettente.

Bibliografia

Alho, Juha M., Mulry, Mary H., Wurdeman, Kent and Kim, Jay. Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. ASA, 1993. (Journal of the American Statistical Association no.88-423:1130-1136.)

Alho, J. M. Analysis of Sample Based Capture-Recapture Experiments, Journal of Official Statistics no.10-3(1994):245-256.

Calzaroni, Manlio, Crescenzi, Fabio, Fortini, Marco, Mancini, Andrea e Sindoni, Giuseppe. *Linee strategiche su metodi, tecniche e organizzazione del Censimento permanente della popolazione e delle abitazioni*. Roma: ISTAT, 2013. (Relazione per il Comitato Scientifico del Censimento Permanente.)

Ceccarelli, Claudio e Rosati, Simona. *L'utilizzo delle Liste Anagrafiche Comunali*. Roma: ISTAT, 2014. (Seminario su Le innovazioni metodologiche nelle indagini socio-economiche sulle famiglie, 20 maggio).

Fortini, Marco e Gallo, Gerardo. *Misure di sottocopertura anagrafica in base alla revisione post-censuaria del 2001*. Pescara: SIS, 2009. (Convegno su Statistical Methods for the Analysis of Large Data-Sets, 23-25 settembre).

Mancini, Andrea. Latest Innovation of Italian Population Census. Roma: SIEDS, 2011. (Rivista Italiana di Economia Demografia e Statistica, no. LXV, no. 2 Aprile-Giugno: 43-56.)

Mancini, L. e Toti, S. (2014). *Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del censimento 2011.* Articolo accettato per la pubblicazione sulla collana Istat Working Papers.

Wolter, Kirk M. *Some Coverage Error Models for Census Data*. ASA, 1986. (Journal of the American Statistical Association no.81-394 (1986):338-346.)