istat
working
papers

N. 8
2014

Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del Censimento 2011

Luca Mancini e Simona Toti

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 8
2014

Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del Censimento 2011

Luca Mancini e Simona Toti

Istat

Istat Working Papers

Dalla popolazione residente a quella abitualmente dimorante:
modelli di previsione a confronto sui dati del Censimento 2011

N. 8/2014

ISBN 978-88-458-1816-5

Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del censimento 2011

Luca Mancini e Simona Toti *

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. La capacità di riprodurre da modello il dato censuario da quello amministrativo viene valutata sotto ipotesi alternative sulla struttura di varianza delle variabili risposta, che identificano rispettivamente la sottocopertura (individui abitualmente dimoranti ma non iscritti in anagrafe) e la sovracopertura (individui non abitualmente dimoranti sebbene iscritti). I risultati mostrano come la previsione della popolazione obiettivo, tanto per i comuni campione quanto per i comuni non campionati, sia decisamente più 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). La superiorità di tali modelli deriva dalla loro capacità di 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 socio-economiche 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: censimento della popolazione, errore di copertura, archivi amministrativi, modelli logistici multilivello, modelli cattura-ricattura.

Abstract

The paper uses both 2011 census and 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 model's ability to deliver the census outcome from register data is assessed under competing assumptions on the variance structure of response variables, which identify the undercounts (unregistered individuals eligible for the census) and the overcounts (registered but ineligible), respectively. The results show how the prediction of the target population, for both the sampled municipalities and those out of the sample, 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). The superiority of these models stems from their ability 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 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.

Keywords: population census, coverage error model, population registers, multilevel logistic regression, capture-recapture models.

* DICA- MTO/B

Indice

1. Introduzione 9

2. I dati 10

3. Le determinanti dell'errore di copertura 11
3.1 Il modello 12
3.2 I risultati 13

4. La ricostruzione del censimento 19
4.1 Il modello di ricostruzione 19
4.2 Ricostruzione per i comuni campione 19
4.3 Ricostruzione per i comuni non appartenenti al campione 21

5. Discussione 29

6. Considerazioni conclusive 29

7. Ringraziamenti 30

Bibliografia 31

1. Introduzione

Uno degli obiettivi principali del censimento italiano della popolazione è quello di certificare la cosiddetta popolazione "legale", ovvero il conteggio per sesso, cittadinanza e classi d'età della popolazione abitualmente dimorante in ciascun comune. L'Italia, come numerosi altri paesi europei, dispone di registri di popolazione regolarmente aggiornati che già contengono informazioni a livello individuale su queste caratteristiche demografiche [1]. In particolare, ciascun comune italiano gestisce l'anagrafe della popolazione residente (APR), anche conosciuta come lista anagrafica comunale o LAC.

Il regolamento anagrafico prevede l'obbligo per il cittadino di stabilire la propria residenza nel comune di dimora abituale mediante l'iscrizione nella LAC di quel comune. Dunque, fissando arbitrariamente una data per il censimento, la popolazione anagrafica o residente e quella censuaria o abitualmente dimorante dovrebbero teoricamente coincidere, a meno di minimi aggiustamenti dovuti, ad esempio, a pratiche pendenti di trasferimento di residenza ad altro comune. Di fatto, per motivi di varia natura, ciò solitamente non avviene poiché le LAC sono notoriamente affette da errori cosiddetti "di copertura". Si definisce "sottocopertura delle LAC" la popolazione abitualmente dimorante ma "invisibile" ai registri anagrafici, mentre l'espressione "sovracopertura delle LAC" identifica la popolazione iscritta in anagrafe ma irreperibile al censimento.

Le LAC hanno avuto un ruolo importante nell'ultima tornata censuaria del 2011 [2]. Milioni di famiglie hanno ricevuto per posta il questionario all'indirizzo di residenza anagrafica. L'Istituto Nazionale di Statistica (Istat), in collaborazione con gli uffici di censimento comunali, ha inaugurato un sistema informatico di gestione della rilevazione censuaria che, tra le sue funzioni, prevedeva un confronto in tempo reale tra le informazioni individuali acquisite al censimento e quelle presenti in anagrafe. Ciò ha permesso di determinare, in modo pressoché immediato, la sottocopertura e la sovracopertura delle LAC, che a livello nazionale sono risultate essere rispettivamente pari all'1 e 4 per cento della popolazione censita. Questi risultati rappresentano un ritorno positivo dell'investimento fatto dall'Istat per 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 [3]. 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. L'imminente confluenza dei registri anagrafici comunali in un archivio unico nazionale della popolazione residente (ANPR) gestito centralmente dal Ministero degli Interni contribuisce a rafforzare la scelta delle LAC come eventuale sostituto naturale del censimento per la determinazione della popolazione legale.

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. Ad esempio, le LAC sono notoriamente carenti nel rappresentare la popolazione straniera abitualmente dimorante in Italia [5][6]. L'obiettivo della presente analisi è duplice: (1) 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; (2) valutare il potere predittivo di tali modelli inteso come capacità di riprodurre il dato censuario da quello anagrafico sia per i comuni campione sia per i comuni non campionati. Rispetto al primo obiettivo, questo studio 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*) [9][10]. I risultati dell'analisi sono inoltre utili per valutare il rischio di selezione del campione a cui sono esposte le indagini socio-economiche sulle famiglie laddove i nuclei vengono estratti dalle LAC. Un caso esemplare è rappresentato dall'indagine EUSILC [4] il cui obiettivo consiste nel documentare l'esclusione sociale, fenomeno che tende a colpire proprio quelle sottopopolazioni la cui presenza sul territorio le LAC non riescono adeguatamente a rappresentare. Con riferimento al secondo obiettivo, l'analisi prende spunto dai modelli cattura-ricattura con l'individuo "catturato" la prima volta dalla LAC e la seconda al censimento [17]. Sebbene l'applicazione di tali modelli al conteggio delle popolazioni umane si contraddistingua per

uno schema di campionamento di tipo areale, in questo lavoro si è scelto più semplicemente di selezionare campioni casuali di individui secondo le modalità descritte nel paragrafo successivo. Infatti, la simulazione del censimento 2011 qui presentata è un esercizio strumentale alla valutazione di modelli di previsione alternativi del rischio individuale di sotto e sovracopertura, sulla quale lo schema di campionamento adottato è stato verificato essere ininfluente.

Questo studio è parte di un progetto più ampio, denominato METOPOP, sulla pianificazione metodologica del censimento che verrà: il Censimento Permanente della Popolazione e delle Abitazioni. Uno dei cardini del prossimo censimento è rappresentato da un'indagine di copertura annuale dei registri anagrafici da somministrare a rotazione quinquennale su campioni di comuni, o su porzioni del loro territorio [7]. In particolare, si ipotizza che i comuni con una popolazione superiore a 50000 abitanti siano sondati annualmente su un campione opportunamente selezionato della popolazione residente. Per i comuni al di sotto di tale soglia demografica, invece, la strategia allo studio è quella di selezionare, sempre a rotazione, campioni bilanciati di comuni o di agglomerati di comuni per i centri più piccoli. L'utilizzo dei risultati dell'indagine di copertura è tuttora argomento di dibattito. Tra le ipotesi più accreditate c'è quella di effettuare un censimento puramente da registro per quei comuni i cui archivi anagrafici forniscono adeguate garanzie di copertura della popolazione obiettivo. Un'altra possibilità è utilizzare l'indagine di copertura per ridimensionare statisticamente le LAC affette da errore. Indipendentemente da quali saranno le soluzioni prescelte, ci si attende che l'analisi presentata nei paragrafi che seguono possa offrire degli spunti di riflessione utili alla progettazione del censimento permanente.

Il resto dell'articolo è organizzato come segue: il paragrafo 2 presenta i dati e la strategia di campionamento utilizzata nell'analisi; il paragrafo 3 discute le possibili determinanti dell'errore di copertura, descrive i modelli di stima e le variabili coinvolte (paragrafo 3.1) e, infine, presenta i risultati ottenuti (paragrafo 3.2). Il paragrafo 4 affronta la ricostruzione del censimento a partire dalle LAC. In particolare, il paragrafo 4.1 definisce il modello di previsione, mentre i paragrafi 4.2 e 4.3 presentano, rispettivamente, i risultati della ricostruzione per i comuni campione e per quelli non appartenenti al campione. Il paragrafo 5 sintetizza i risultati principali del lavoro evidenziandone i limiti e indicando possibili sviluppi futuri. Il paragrafo 6 conclude.

2. I 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. Restano pertanto esclusi dall'analisi sia i comuni più piccoli (circa duemila) che i 141 comuni più grandi. I primi non sono stati presi in considerazione per via dell'irrilevanza che la dimensione del fenomeno oggetto di studio assume percentualmente su scala nazionale. Per i comuni più grandi, risultati preliminari provenienti da sperimentazioni ISTAT parallele a quella qui descritta suggeriscono che modelli di previsione a livello di sezione di censimento forniscono risultati soddisfacenti nel ricostruire i conteggi censuari a partire dalle liste anagrafiche comunali. Per questo motivo, nonostante i comuni con una popolazione superiore a 50000 abitanti catturino oltre il 75% dell'errore di copertura totale delle LAC rispetto al censimento 2011, nel presente lavoro si è deciso di concentrare l'attenzione sui comuni di ampiezza demografica medio-piccola.

La stratificazione dei comuni oggetto di studio in tre classi risponde all'esigenza di considerare strati sufficientemente omogenei rispetto alla dimensione della popolazione comunale. Infatti l'errore di copertura tipicamente aumenta sia per l'effetto diretto legato al numero di residenti che per un effetto mediato dalla diversa composizione – in termini di profili demografici a rischio copertura- della popolazione residente all'aumentare della dimensione del comune. Infine, la soglia 1001-5000 ha anche una valenza operativa poiché si prospetta un possibile accorpamento dei comuni più piccoli in

Tabella 1 - Descrizione dei campioni e loro principali caratteristiche

	1001-5000	5001-10000	10001-50000	Totale Italia
Numero di comuni	3751	1187	1062	8092
Popolazione complessiva	8974000	8394302	20722806	59433744
Numero di comuni campionati	750	300	250	-
Popolazione nei comuni campionati (frame)	1871417	2226136	5145318	-
Popolazione nel campione	200371	200142	200122	-
Sottocopertura				
totale	53723	58121	179179	628741
frame	10058	13694	41254	
campione	1059	1252	1602	
CV	0,20	0,27	0,23	
Sovracopertura				
totale	185917	210604	668853	2379604
frame	35590	54016	165881	
campione	3735	4874	6429	
CV	0,11	0,16	0,14	
Sovrapposizione (sotto ∩ sovra)				
totale	19049	23804	65474	176084
frame	269	306	1347	
campione	52	25	52	

agglomerati di circa 5000 abitanti da considerare come unità campionarie di primo stadio nel disegno dell'indagine di copertura del censimento permanente.

Per ognuna delle tre classi di ampiezza demografica, è stato estratto un campione casuale a due stadi, con i comuni come unità di primo stadio e gli individui, selezionati con probabilità proporzionale alla dimensione della LAC comunale, come unità di secondo stadio. La Tabella 1 riporta, per ciascuna classe di ampiezza demografica, la dimensione della sotto e sovracopertura nel campione, nelle LAC dei comuni campionati (*frame*) e, infine, nelle LAC di tutti i comuni (campionati e non). I tre campioni estratti contengono circa 200 mila individui e assicurano, sotto ipotesi opportune in termini di rarità del fenomeno e di effetto del disegno, un coefficiente di variazione (*CV*) delle stime campionarie dell'errore di copertura compreso tra l'11 ed il 27%. La sottocopertura è un fenomeno circa quattro volte più raro rispetto alla sovracopertura. Le due componenti dell'errore di copertura hanno inoltre una sovrapposizione, a livello nazionale, pari a circa il 30%. Ciò significa che oltre 175000 individui sono stati censiti come nuovi individui (sottocopertura) in un comune italiano e dichiarati irreperibili (sovracopertura) in un altro comune. La dimensione di tale sovrapposizione all'interno dei campioni estratti è tuttavia molto esigua dato che la popolazione dei comuni campionati all'interno di ciascuna classe di ampiezza demografica oscilla tra il 3 ed il 9% del totale nazionale.

3. Le determinanti dell'errore di copertura

In una LAC non affetta da errore di copertura la popolazione anagrafica del comune coincide con la popolazione obiettivo del censimento. Se ciascun individuo, ottemperando alle disposizioni del regolamento anagrafico, stabilisse la propria residenza nel comune in cui dimora abitualmente e comunicasse tempestivamente eventuali variazioni di residenza in modo tale da consentire agli uffici anagrafe di aggiornare regolarmente le LAC, gli scarti tra le due popolazioni dovrebbero essere trascurabili.

Nella realtà, tuttavia, le LAC sono spesso affette da un errore di copertura non trascurabile. Ciò dipende da una serie di fattori concomitanti e tra loro interagenti [8]:

- Le scelte individuali;
- Le caratteristiche del luogo di residenza/dimora abituale;
- La qualità dei registri anagrafici.

Iscriversi in anagrafe o cambiare residenza dipende dalle scelte più o meno consapevoli del cittadino che omette di comunicare all'ufficiale di anagrafe competente il proprio trasferimento in/da un dato

Tabella 2 - Descrizione delle variabili

Variabile	Modalità e descrizione
sotto	1 se l'individuo è sottocoperto, 0 altrimenti
sovra	1 se l'individuo è sovracoperto, 0 altrimenti
sesso	1 se femmina, 0 se maschio
cittad	1 se straniero, 0 se italiano
monocomponente	1 se l'individuo vive in famiglia anagrafica monocomponente, 0 altrimenti
eta_1	1 se l'individuo ha meno di 19 anni, 0 altrimenti
eta_2	1 se l'individuo ha tra i 19 e i 40 anni, 0 altrimenti
eta_3	1 se l'individuo ha tra i 41 e i 70 anni, 0 altrimenti
eta_4	1 se l'individuo ha oltre 70 anni, 0 altrimenti
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 ai 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

comune. Ciò può avvenire per motivi di convenienza fiscale soprattutto legati alla tassazione agevolata sulle "prime" rispetto alle "seconde" case di proprietà, ma anche per evitare l'iter burocratico di doversi recare in municipio specialmente per chi non conosce bene la lingua italiana o per chi ha un contratto di lavoro a termine e intravede la possibilità di doversi trasferire altrove. Alcune di queste caratteristiche concorrono tipicamente ad individuare le sottopopolazioni "difficili da contare" (*hard to count*) che alcuni istituti nazionali di statistica considerano espressamente come variabili di stratificazione nell'indagine di copertura post-censuaria [9][10].

Il fenomeno può dipendere altresì da fattori ambientali legati al territorio comunale come la sua geografia, l'economia locale e le opportunità occupazionali che esso offre, l'età media dei suoi abitanti così come la presenza più o meno forte di cittadini stranieri.

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.

3.1 Il modello

Sulla base delle considerazioni del paragrafo precedente, ci si aspetta che l'errore di copertura delle LAC sia influenzato sia da caratteristiche proprie dell'individuo sia da fattori legati direttamente al comune e al suo territorio. La Tabella 2 descrive le variabili utilizzate nella presente analisi[1]. Le grandezze oggetto di stima sono di fonte censuaria ("sotto" e "sovra") mentre quasi tutte le variabili esplicative presenti nel modello, sia a livello individuale che comunale, provengono dalle liste anagrafiche comunali[2].

Si assume che le variabili risposta dicotomiche sotto="l'individuo è/non è sottocopertura" e sovra="l'individuo è/non è sovracopertura" siano in relazione con le altre, tramite il seguente modello di regressione logistica:

$$\text{logit}\left\{ P(Y=1|X=x) \right\} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (1)$$

[1] Altre variabili come, ad esempio, la ripartizione territoriale ed un indice di vocazione turistica del comune, sono state inizialmente considerate nell'analisi ma poi escluse dalla specificazione finale del modello perché non statisticamente significative.

[2] Il tasso comunale di residenti che lavorano fuori comune proviene da altri archivi amministrativi appartenenti al Sistema Integrato dei Microdati (SIM) che riunisce archivi anagrafici, fiscali, sulla formazione, lavoro e *welfare* [11]. Le coordinate chilometriche dei centroidi comunali utilizzate per il calcolo della distanza euclidea del comune dal suo capoluogo di regione provengono dalle Basi Territoriali dell'ISTAT.

dove Y è la variabile risposta (alternativamente sotto o sovra), α l'intercetta generale, $\beta_1, \beta_2, \cdots, \beta_k$ il vettore dei coefficienti di regressione relativi ai cosiddetti "effetti fissi" e $X = X_1, X_2, \cdots, X_k$ il profilo o vettore delle modalità dei k regressori rilevati su un dato individuo [3].

L'assunzione di identica distribuzione ed indipendenza della variabile risposta (i.i.d.), condizionatamente al valore dei regressori, permette di ottenere la stima di massima verosimiglianza del modello e quella della probabilità di Y. Nel caso in esame, risulta però più verosimile supporre che la variabile risposta, condizionatamente al valore dei regressori, non sia i.i.d. Infatti, la variabilità di Y tra i soggetti sarà tipicamente diversa da comune a comune, così che l'ipotesi di omoschedasticità di Y resta verificata solo per individui dello stesso comune. Se si fa questa assunzione, l'equazione 1 relativa al valore atteso della variabile risposta come funzione logistica dei regressori resta valida, ma la varianza diventa specifica di comune. In questo caso, si parla di modello di regressione logistica "marginale" o "*population-average*", che richiede tecniche di stima generalizzate (*GEE*) [12].

Un ulteriore adattamento del modello logistico al caso della ricostruzione della sotto e sovracopertura, consiste nel prevedere che il livello di rischio di un individuo vari, a parità di valore dei suoi regressori, da comune a comune. L'eterogeneità dei comuni, diventa così essa stessa una grandezza d'interesse, da quantificate tra gli effetti stimati, e non più semplicemente una violazione dell'ipotesi di omoschedasticità tra unità di primo livello di cui tenere conto nella stima dei coefficienti e dei loro errori standard. Il modello di regressione utilizzato in questo quadro è quello logistico ad intercetta aleatoria o ad "effetti misti":

$$\text{logit}\left\{ P(Y = 1|X = x) \right\} = \alpha + \alpha_i + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{2}$$

con α_i intercetta casuale relativa al comune *i*-esimo. Dunque, oltre alle varianze, ora si avranno probabilità di Y specifiche di comune. Per questo modello si ricorre alle tecniche di stima proprie dei modelli multilivello [13][14][15].

La Tabella 3 mostra la distribuzione delle variabili individuali utilizzate nei modelli di stima per i "sottocoperti", per i "sovracoperti" e per tutti gli individui iscritti nelle LAC dei comuni campionati. La Tabella 4 riporta i valori medi e gli errori standard delle variabili comunali. Indipendentemente dalla classe di ampiezza demografica, risulta evidente come l'errore di copertura non si distribuisce in modo omogeneo nella popolazione ma che esistono categorie di individui – come i cittadini stranieri, le persone in età lavorativa e le famiglie anagrafiche monocomponente - sensibilmente più a rischio di essere esclusi dalla popolazione legale e/o di essere indebitamente inclusi o erroneamente attribuiti ad un dato comune qualora il censimento fosse basato esclusivamente sui registri anagrafici.

3.2 I risultati

I coefficienti di regressione con la relativa significatività statistica sono riportati nelle Tabelle 5 e 6 rispettivamente per i modelli di sotto e sovracopertura[4]. 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 del modello, 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

[3] Un modello simile è stato utilizzato da [10] per stimare le probabilità di copertura della *Post Enumeration Survey* del censimento statunitense della popolazione del 1990.

[4] I modelli sono stati stimati con STATA 13.0. In particolare, per il modello *GEE* è stato utilizzato il comando *xtgee* mentre per il modello a effetti misti e per la previsione è stato utilizzato il pacchetto *GLLAMM* (Generalized Linear Latent and Mixed Models)[16].

Tabella 3 - Caratteristiche individuali ed errore di copertura (*frame*)

		Sottocopertura		Sovracopertura		LAC	
	Caratteristica dell'individuo	n	%	n	%	n	%
1001-5000	italiano	7870	78,3	19637	55,2	1733813	93,2
	straniero	2188	21,8	15353	44,8	127490	6,9
	pluricomponente	6776	67,4	24373	68,5	1607822	86,4
	monocomponente	3282	32,6	11217	31,5	253532	13,6
	femmina	5019	50,8	16172	45,4	942337	50,6
	maschio	4949	49,2	19418	54,6	918966	49,4
	under 19	2544	25,3	5016	14,1	321594	17,3
	età 19-40	4380	43,6	16142	45,4	507092	27,2
	età 41-70	2680	26,7	11691	32,9	751012	40,4
	over 70	454	4,5	2741	7,7	281656	15,1
5001-10000	italiano	10325	75,4	28939	53,6	2041575	92,3
	straniero	3365	24,6	25077	46,4	170812	7,7
	pluricomponente	9598	70,1	38086	70,5	1948610	88,1
	monocomponente	4096	29,9	15930	29,5	263810	11,9
	femmina	6982	51,0	23813	44,1	1122741	50,8
	maschio	6712	49,0	30203	55,9	1089651	49,2
	under 19	2850	20,8	8476	15,7	402581	18,2
	età 19-40	6675	48,7	25482	47,2	622991	26,2
	età 41-70	3165	26,4	17050	31,6	889127	40,2
	over 70	554	4,1	3008	5,6	297721	13,4
10001-50000	italiano	30614	74,2	97808	59,0	4728096	92,6
	straniero	10640	25,8	68073	41,4	375757	7,4
	pluricomponente	28539	69,2	119534	72,1	4496733	88,1
	monocomponente	12715	30,8	46347	27,9	607264	11,9
	femmina	21034	51,0	75345	45,4	2610588	51,1
	maschio	20220	49,0	90536	54,6	2493305	48,9
	under 19	8934	21,7	29187	17,6	953185	18,7
	età 19-40	19590	47,5	75958	45,8	1452593	28,5
	età 41-70	11139	27,0	51822	31,2	2036096	39,9
	over 70	1591	3,9	8911	5,4	66210	13,0

Tabella 4 - Caratteristiche comunali (*frame*)

	1001-5000		5001-10000		10001-50000	
Caratteristica del comune	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
t_citt (%)	6,6	4,6	7,7	4,7	7,7	4,8
t_mono (%)	14,2	4,4	12,2	2,8	11,9	3,0
t_anziani (%)	15,5	4,0	13,4	2,8	12,9	2,9
t_lavout (%)	22,1	6,9	22,6	6,9	20,3	7,2
t_cf (%)	0,9	2,9	0,8	2,8	1,3	5,8
distanza (km)	71,2	40,0	63,4	42,9	53,9	43,8
lac (n)	2484,3	1099,4	7355,4	1432,4	20344,9	9834,8

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 Tabella 5 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. La Tabella 6 mostra un effetto simile sulla probabilità di sovracopertura, ma solo per i comuni con un numero di abitanti compreso tra 10000 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

Tabella 5 - I coefficienti di regressione per i modelli di sottocopertura

	1001-5000			5001-10000			10001-50000		
Sotto	(1)[a,b,c]	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
femmina	0,13	0,13	*0,12*	*0,04*	*0,04*	*0,03*	*-0,02*	*-0,02*	*0,00*
eta_1	1,37	1,40	1,36	0,76	0,76	0,76	0,97	0,98	0,96
eta_2	0,99	1,00	0,99	0,91	0,92	0,92	0,87	0,87	0,85
eta_4	-1,31	-1,33	-1,30	-1,23	-1,23	-1,23	-0,94	-0,95	-0,98
monocomp.	1,94	1,97	1,93	1,52	1,53	1,54	1,67	1,68	1,69
cittad	1,77	1,75	1,78	1,23	1,27	1,29	1,58	1,55	1,52
citXtcit	-5,73	-5,21	-5,48	*0,22*	*-0,12*	*-0,17*	*-3,01*	*-2,63*	*-2,20*
t_citt	*1,47*	*0,74*	*0,99*	*-1,43*	*-0,98*	*-1,31*	*1,05*	*1,00*	*1,22*
t_mono	4,98	4,47	5,38	4,95	*2,81*	5,93	5,89	4,69	6,33
t_cf	0,03	0,03	0,03	0,03	0,03	0,03	*0,01*	*0,01*	*0,01*
t_anziani	-4,45	-4,57	-5,15	-4,82	*-3,08*	-5,32	-5,69	-5,07	-6,18
t_lavout	*-0,06*	*-0,24*	*-0,16*	*0,49*	*0,67*	*0,66*	1,84	2,08	2,43
distanza	0,00	0,00	0,00	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*	*0,00*	*0,00*	0,00
α	-6,35	-6,14	-6,57	-5,84	-5,90	-6,13	-6,26	-6,20	-6,60
AIC	11813		11642	13939		13761	16660		16087
Sensibilità	0,69	0,69	0,76	0,68	0,69	0,72	0,68	0,68	0,72
var(α_i)			0,61			0,39			0,25
ρ			0,19			0,16			0,13
N		195860			194551			191433	
n		746			299			247	

(a) (1): modello logistico ordinario; (2) modello logistico eteroschedastico (*GEE*); (3): modello logistico multilivello.
(b) I coefficienti statisticamente non significativi al 5% sono indicati in corsivo.
(c) Profilo di riferimento: cittadino italiano, maschio, di età compresa tra 41 e 70 anni, con nucleo familiare anagrafico pluricomponente.

una capacità predittiva, sia per la sotto che per la sovracopertura, compresa tra 68 e 71% per i due modelli ad effetti fissi. Il modello ad effetti misti, indipendentemente dalla componente dell'errore e dalla classe di ampiezza demografica, ha una sensibilità superiore di circa 5 punti percentuali rispetto ai primi due modelli. Inoltre, si nota come la varianza dell'intercetta aleatoria, e di conseguenza la quota di varianza (ρ) della variabile dipendente riconducibile all'eterogeneità tra comuni rispetto alla sotto e sovracopertura delle LAC , diminuisca all'aumentare della dimensione demografica e del numero di comuni inclusi nel campione.

I coefficienti di regressione riportati nelle Tabelle 5 e 6 forniscono indicazioni sia sulla direzione sia sulla forza dell'associazione tra i regressori e la variabile dipendente. Al fine di quantificare l'intensità di tale associazione, le Tabelle 7 e 8 riportano le probabilità di sotto e sovracopertura predette dal modello (3) per diversi profili sia individuali che municipali. 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 ($\alpha_i = 0$), 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%. E' interessante notare come facendo successivamente aumentare (diminuire) il solo tasso di cittadini stranieri del comune di una deviazione standard dal suo valor medio la probabilità stimata diminuisca (aumenti) di circa 1,5 (2) punti percentuali. Si noti, inoltre, il contributo dell'effetto casuale di comune: il valore di 8,6% più che raddoppia (si dimezza) facendo variare l'intercetta α_i di una deviazione standard sopra (sotto) il suo valor medio.

Le Tabelle 9 e 10 mostrano come i coefficienti di regressione di parte fissa (β) rimangano sostanzialmente stabili quando i modelli sono stimati su un campione di validazione di nuovi individui residenti in nuovi comuni estratto con un disegno identico al primo campione (di stima o calibrazione). Ciò è particolarmente evidente per i coefficienti delle caratteristiche individuali come sesso, età, cittadinanza e tipologia di nucleo familiare che rimangono pressoché invariati per dimensione, segno e significatività statistica indipendentemente dal tipo di modello o dalla classe di ampiezza demografica del comune. Gli effetti relativi alle variabili di livello comunale mostrano invece maggiore

Tabella 6 - I coefficienti di regressione per i modelli di sovracopertura

	1001-5000			5001-10000			10001-50000		
Sovra	(1)[a,b,c]	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
femmina	-0,16	-0,16	-0,17	-0,21	-0,21	-0,22	-0,22	-0,22	-0,24
eta_1	*-0,01*	*-0,02*	*-0,01*	0,12	0,12	0,13	0,24	0,24	0,25
eta_2	0,48	0,47	0,49	0,49	0,48	0,50	0,53	0,52	0,53
eta_4	-0,52	-0,52	-0,53	-0,61	-0,60	-0,61	-0,57	-0,56	-0,58
monocomp.	1,27	1,27	1,30	1,27	1,27	1,31	1,15	1,15	1,18
cittad	2,47	2,52	2,58	2,46	2,44	2,54	2,42	2,41	2,51
citXtcit	*-0,80*	*-1,34*	*-1,04*	*-0,28*	*-0,16*	*-0,58*	-1,95	-1,98	-2,32
t_citt	*-0,10*	*0,63*	*0,06*	*-0,04*	*-0,27*	*-0,39*	*0,42*	*0,46*	*1,35*
t_mono	4,00	2,40	3,60	9,41	8,49	9,13	4,45	*2,49*	3,73
t_cf	0,04	0,03	0,04	0,03	*0,02*	0,03	*0,00*	*0,00*	*-0,01*
t_anziani	-5,81	-4,47	-4,64	-9,84	-8,31	-8,05	-6,84	-3,17	-5,33
t_lavout	-2,14	-2,01	-1,64	-3,50	-2,98	-2,63	*-0,13*	*-0,17*	*-0,39*
distanza	0,00	0,00	0,00	*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*	0,00	0,00	0,00
α	-3,97	-4,04	-4,43	-3,93	-4,04	-4,39	-4,25	-4,22	-4,46
AIC	31293		30456	37797		36859	47767		46671
Sensibilità	0,71	0,71	0,77	0,68	0,68	0,76	0,68	0,68	0,73
var(α_i)			0,46			0,38			0,35
ρ			0,17			0,16			0,15
N		198524			198162			196203	
n		746			299			247	

(a) (1): modello logistico ordinario; (2) modello logistico eteroschedastico (*GEE*); (3): modello logistico multilivello.
(b) I coefficienti statisticamente non significativi al 5% sono indicati in corsivo.
(c) Profilo di riferimento: cittadino italiano, maschio, di età compresa tra 41 e 70 anni, con nucleo familiare anagrafico pluricomponente.

variabilità tra i due campioni, sia nella parte fissa che nella parte casuale. Ad esempio, l'effetto negativo sul rischio di sottocopertura dell'interazione tra cittadinanza dell'individuo e tasso comunale di cittadini stranieri evidenziato in precedenza per i comuni di ampiezza 1001-5000 diventa non statisticamente significativo nel campione di validazione. Un effetto interazione negativo e significativo sulla probabilità di sottocopertura è, invece, presente nei comuni di ampiezza 5001-10000, a differenza delle stime di calibrazione dove lo stesso effetto era debole e non statisticamente significativo.
Si noti che laddove si riscontrano variazioni significative nei coefficienti stimati di parte fissa anche la parte casuale, rappresentata dalla stima della varianza degli effetti casuali α_i, mostra delle differenze significative. Ad esempio, la varianza delle intercette casuali di sottocopertura per i comuni di ampiezza 1001-5000 estratti nel campione di validazione (0,80) è sensibilmente più elevata (e di conseguenza è più elevata la porzione di varianza totale (ρ) catturata dalla eterogeneità tra comuni) rispetto allo stesso valore stimato per i comuni di calibrazione (0,61).

Tabella 7 - Probabilità (%) stimata di sottocopertura per alcune categorie di individui

	1001-5000			5001-10000			10001-50000		
α_i	$-\sigma$	0	$+\sigma$	$-\sigma$	0	$+\sigma$	$-\sigma$	0	$+\sigma$
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
eta_2	1,0	2,3	5,3	1,3	2,7	5,6	1,6	3,2	6,3
cittad@($\overline{\text{t\_citt}}$)[b]	3,8	8,6	18,6	5,1	10,2	19,5	6,0	11,5	20,8
cittad@($\overline{\text{t\_citt}} + \sigma$)	3,1	7,1	15,6	3,9	8,0	15,7	5,7	11,0	20,0
cittad@($\overline{\text{t\_citt}} - \sigma$)	4,6	10,5	22,0	6,5	12,9	24,1	6,3	11,9	21,5
$\overline{\text{t\_mono}} + \sigma$	5,7	12,7	26,1	7,0	13,9	25,6	7,7	14,5	25,6
$\overline{\text{t\_lavout}} + \sigma$	5,6	12,6	25,8	7,3	14,5	26,5	9,1	16,8	29,0
$\overline{\text{t\_anziani}} + \sigma$	4,7	10,6	22,3	6,5	13,0	24,1	7,5	14,1	24,9
$\overline{\text{t\_cf}} + \sigma$	5,0	11,3	23,6	6,7	13,3	24,7	7,9	14,7	25,9
$\overline{\text{distanza}} + \sigma$	4,3	9,7	20,7	6,5	12,9	24,0	7,3	13,7	24,3
$\overline{\text{lac}} + \sigma$	4,4	9,9	21,0	6,3	12,5	23,4	8,0	14,9	26,2

(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@($\overline{\text{t\_citt}}$) 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.

Tabella 8 - Probabilità (%) stimata di sovracopertura per alcune categorie di individui

	1001-5000			5001-10000			10001-50000		
α_i	$-\sigma$	0	$+\sigma$	$-\sigma$	0	$+\sigma$	$-\sigma$	0	$+\sigma$
base[a]	0,3	0,7	1,5	0,4	0,9	1,9	0,6	1,3	2,8
femmina	0,2	0,6	1,3	0,3	0,7	1,5	0,5	1,1	2,2
monocomponente	0,9	2,0	4,4	1,0	2,2	4,6	1,6	3,3	6,9
eta_2	1,4	3,2	7,0	1,7	3,7	7,6	2,6	5,5	11,2
cittad@($\overline{\text{t\_citt}}$)[b]	15,2	28,9	48,1	16,8	30,4	48,4	22,0	37,9	56,8
cittad@($\overline{\text{t\_citt}} + \sigma$)	14,6	28,0	47,0	16,2	29,4	47,3	21,3	36,8	55,7
cittad@($\overline{\text{t\_citt}} - \sigma$)	15,8	29,9	49,3	17,5	31,3	49,6	22,9	39,0	60,0
$\overline{\text{t\_mono}} + \sigma$	17,8	33,1	53,0	19,0	33,6	52,1	25,3	42,2	61,1
$\overline{\text{t\_lavout}} + \sigma$	16,2	30,6	50,2	18,6	33,0	51,5	24,8	41,5	60,5
$\overline{\text{t\_anziani}} + \sigma$	14,0	27,0	45,8	16,5	29,8	47,8	21,6	37,3	56,1
$\overline{\text{t\_cf}} + \sigma$	15,3	29,2	48,4	16,2	29,4	47,4	20,8	36,3	55,1
$\overline{\text{distanza}} + \sigma$	14,3	27,5	46,4	15,8	28,8	46,5	20,3	35,4	54,2
$\overline{\text{lac}} + \sigma$	14,7	28,2	47,2	16,1	29,3	47,1	23,6	40,0	58,9

(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@($\overline{\text{t\_citt}}$) 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.

Tabella 9 - I coefficienti di regressione per i modelli di sottocopertura (campione di validazione)

	1001-5000			5001-10000			10001-50000		
Sotto	(1)[a,b,c]	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
femmina	0,19	0,19	0,19	0,13	0,13	*0,12*	*0,01*	*0,05*	*0,00*
eta_1	1,08	1,09	1,08	0,90	0,91	0,90	0,82	0,82	0,82
eta_2	0,76	0,78	0,78	0,94	0,95	0,95	0,86	0,87	0,87
eta_4	-1,13	-1,13	-1,13	-1,14	-1,15	-1,14	-1,10	-1,10	-1,09
monocomp.	1,60	1,61	1,61	1,54	1,56	1,56	1,64	1,65	1,64
cittad	1,40	1,38	1,40	1,90	1,88	1,87	1,47	1,48	1,51
citXtcit	*-2,32*	*-1,83*	*-1,91*	-6,82	-6,34	-6,23	*-1,21*	*-1,18*	*-1,39*
t_citt	*1,23*	*1,00*	*1,15*	*0,02*	*-0,14*	*0,17*	*0,01*	*-0,03*	*0,36*
t_mono	3,49	3,36	3,86	*3,10*	*1,88*	2,97	5,07	4,64	5,10
t_cf	*0,01*	*0,01*	0,01	*0,02*	*-0,01*	*0,02*	*0,01*	*0,02*	*0,02*
t_anziani	-5,52	-4,75	-5,20	-4,90	-4,47	-4,26	-7,78	-8,18	-7,59
t_lavout	*-0,37*	*-0,09*	*-0,21*	*0,44*	*0,26*	*0,60*	2,57	2,40	1,88
distanza	*0,00*	*0,00*	*0,00*	*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*	*0,00*	*0,00*	*0,00*
α	-5,63	-5,80	-6,30	-5,67	5,50	-5,93	6,00	-5,89	-6,01
AIC	11627		11534	14035		13875	16395		16260
Sensibilità	0,66	0,66	0,75	0,67	0,67	0,72	0,66	0,66	0,70
var(α_i)			0,80			0,33			0,25
ρ			0,21			0,15			0,13
N		194533			195391			193323	
n		742			300			249	

(a) (1): modello logistico ordinario; (2) modello logistico eteroschedastico (*GEE*); (3): modello logistico multilivello.
(b) I coefficienti statisticamente non significativi al 5% sono indicati in corsivo.
(c) Profilo di riferimento: cittadino italiano, maschio, di età compresa tra 41 e 70 anni, con nucleo familiare anagrafico pluricomponente.

Tabella 10 - I coefficienti di regressione per i modelli di sovracopertura (campione di validazione)

	1001-5000			5001-10000			10001-50000		
Sovra	(1)[a,b,c]	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
femmina	-0,21	-0,22	-0,23	-0,26	-0,26	-0,27	-0,27	-0,27	-0,27
eta_1	*-0,03*	*-0,03*	*0,04*	0,14	0,15	0,16	0,14	0,14	0,15
eta_2	0,47	0,45	0,48	0,50	0,50	0,52	0,47	0,47	0,48
eta_4	-0,36	-0,37	-0,38	-0,63	-0,61	-0,62	-0,63	-0,64	-0,64
monocomp.	1,24	1,24	1,28	1,33	1,34	1,38	1,23	1,24	1,26
cittad	2,46	2,51	2,55	2,57	2,60	2,58	2,12	2,22	2,30
citXtcit	*-0,61*	*-1,44*	*-0,76*	*-1,74*	-2,16	*-1,22*	2,25	*1,19*	*0,98*
t_citt	*-0,01*	*1,52*	*1,04*	-2,04	*-0,90*	-1,92	-3,01	-1,85	-2,34
t_mono	2,34	2,19	2,38	6,26	*2,67*	5,97	7,19	3,21	7,64
t_cf	*0,01*	0,01	0,02	*0,01*	*0,03*	*0,02*	*0,02*	*0,02*	*0,01*
t_anziani	-7,11	-6,77	-6,58	-8,36	-5,32	-8,19	-9,28	-6,12	-9,58
t_lavout	-2,12	-2,53	-2,14	-2,34	-2,10	-2,35	*-0,12*	*-0,39*	*-0,50*
distanza	0,00	*0,00*	*0,00*	*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*	*0,00*	*0,00*	*0,00*
α	-3,65	-3,57	-3,96	-3,81	-3,91	-3,90	-3,68	-3,53	-3,82
AIC	31671		30735	37596		36841	46892		45792
Sensibilità	0,71	0,72	0,77	0,72	0,72	0,75	0,69	0,69	0,74
var(α_i)			0,51			0,29			0,32
ρ			0,18			0,14			0,15
N		197294			198899			198020	
n		742			300			249	

(a) (1): modello logistico ordinario; (2) modello logistico eteroschedastico (*GEE*); (3): modello logistico multilivello,
(b) I coefficienti statisticamente non significativi al 5% sono indicati in corsivo.
(c) Profilo di riferimento: cittadino italiano, maschio, di età compresa tra 41 e 70 anni, con nucleo familiare anagrafico pluricomponente.

4. 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.

4.1 Il modello di ricostruzione

I tre modelli logistici (standard, marginale e ad intercetta aleatoria) introdotti nel paragrafo precedente forniscono una stima delle probabilità che un individuo risulti sottocoperto o sovracoperto nel registro anagrafico del suo comune dato il suo profilo definito dai valori delle covariate (X) considerate e, per il modello (3), anche dal valore dell'intercetta casuale di comune.

Indicato con $\hat{p}_{\text{sotto},\underline{x}_i}$ e $\hat{p}_{\text{sovra},\underline{x}_i}$ il rischio stimato per un individuo di essere rispettivamente sottocoperto o sovracoperto, dato che egli presenta determinati valori delle variabili di regressione, la popolazione abitualmente dimorante in un certo comune viene calcolata riponderando l'ammontare fornito dall'anagrafe di quel comune, secondo la seguente espressione:

$$\hat{N}_i = \sum_X \hat{N}_{\underline{x}_i} = \sum_X \hat{w}_{\underline{x}_i} \cdot \text{LAC}_{\underline{x}_i} = \sum_X \frac{1 - \hat{p}_{\text{sovra},\underline{x}_i}}{1 - \hat{p}_{\text{sotto},\underline{x}_i}} \cdot \text{LAC}_{\underline{x}_i} \tag{3}$$

dove:

- $\underline{x}_i$ è il profilo o vettore $x_1, x_2, \cdots, x_k$ 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;
- $\text{LAC}_{\underline{x}_i}$ è il numero di individui iscritti in anagrafe nel comune *i*-esimo con profilo $\underline{x}_i$.

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 3 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 [17], mentre per un'applicazione specifica del modello logistico a dati individuali di fonte censuaria si veda [10] [18].

Di seguito vengono affrontate due tipologie distinte di ricostruzione [19]. La prima, trattata nel paragrafo 4.2, utilizza le probabilità individuali di sotto e sovracopertura predette dai tre modelli logistici per tutti gli individui (campionati e non campionati) residenti nei comuni campione. La seconda, descritta nel paragrafo 4.3, utilizza invece le probabilità predette per individui residenti in comuni non appartenenti al campione. Questa seconda tipologia di ricostruzione è, dunque, basata su una previsione "fuori campione" (nuovi individui, nuovi comuni) ed assume una particolare rilevanza per i comuni medio-piccoli considerati in questo studio. Infatti, a differenza dei grandi centri urbani, questi comuni potrebbero essere coinvolti nell'indagine di copertura del nuovo censimento permanente una volta ogni cinque anni e pertanto sarebbe necessario ricorrere a stime indirette della popolazione obiettivo negli anni in cui essi non partecipano all'indagine.

4.2 Ricostruzione per i comuni campione

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 $\mid \hat{EC} \mid = \mid \hat{N}_i - N_i \mid$ con $\mid EC \mid = \mid LAC_i - N_i \mid$. 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. Nel paragrafo successivo, i risultati verranno presentati in termini di conteggi oltre che di scarti tra le diverse popolazioni. Le Figure 1-3 confrontano, attraverso grafici di dispersione, l'errore di copertura osservato con l'errore corretto da modello. 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 è più vicino al censimento di quanto non lo sia la LAC. Si nota dalle figure come almeno l'80% dei comuni campionati si trovi sotto la bisettrice, indipendentemente dalla loro dimensione demografica. Inoltre, laddove il modello è più distante della LAC dal censimento (punti sopra la diagonale), si tratta di comuni con un errore di copertura relativamente piccolo. Questo è vero in particolare per il modello (3).

E' evidente la superiorità della previsione basata sul modello ad effetti misti (3) rispetto ai due modelli ad effetti fissi (1) e (2). Sebbene anche questi ultimi riducano generalmente l'errore di copertura osservato, non mancano casi di comuni, soprattutto nella fascia 10001-50000, per i quali i modelli ad effetti fissi tendono a sovra-correggere le LAC in misura significativa. Ciò implica, come del resto evidenziato dai valori di sensibilità riportati in fondo alle Tabelle 5 e 6, che le covariate, sia individuali che comunali, considerate nei modelli di previsione, spieghino 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. Tale conclusione è valida tanto per i comuni di 1000 abitanti quanto per quelli di 50000 abitanti, nel senso che il contributo degli effetti casuali specifici di comune α_i rimane decisivo.

Figura 1 - $|\hat{EC}|$ **contro** $|EC|$ **per modello: comuni con 1001-5000 abitanti**

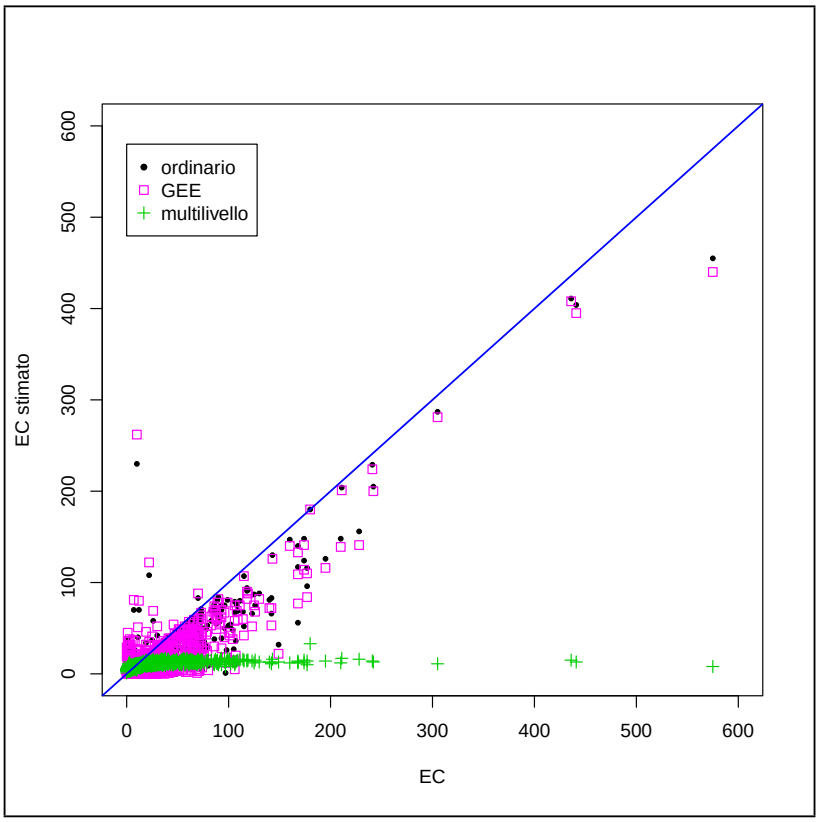
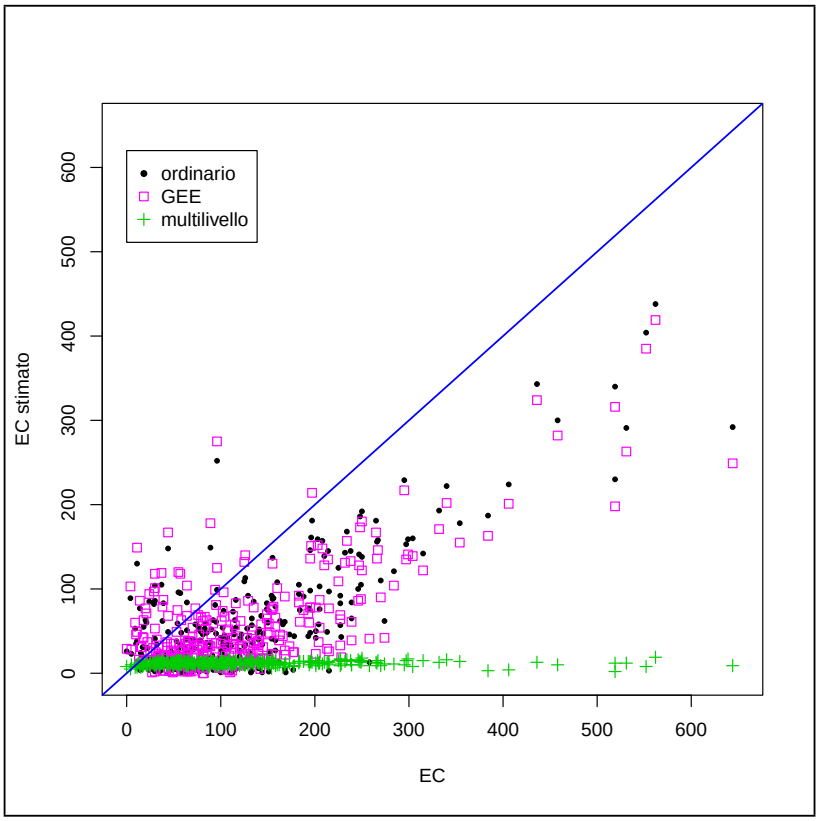


Figura 2 - $|\hat{EC}|$ **contro** $|EC|$ **per modello: comuni con 5001-10000 abitanti**

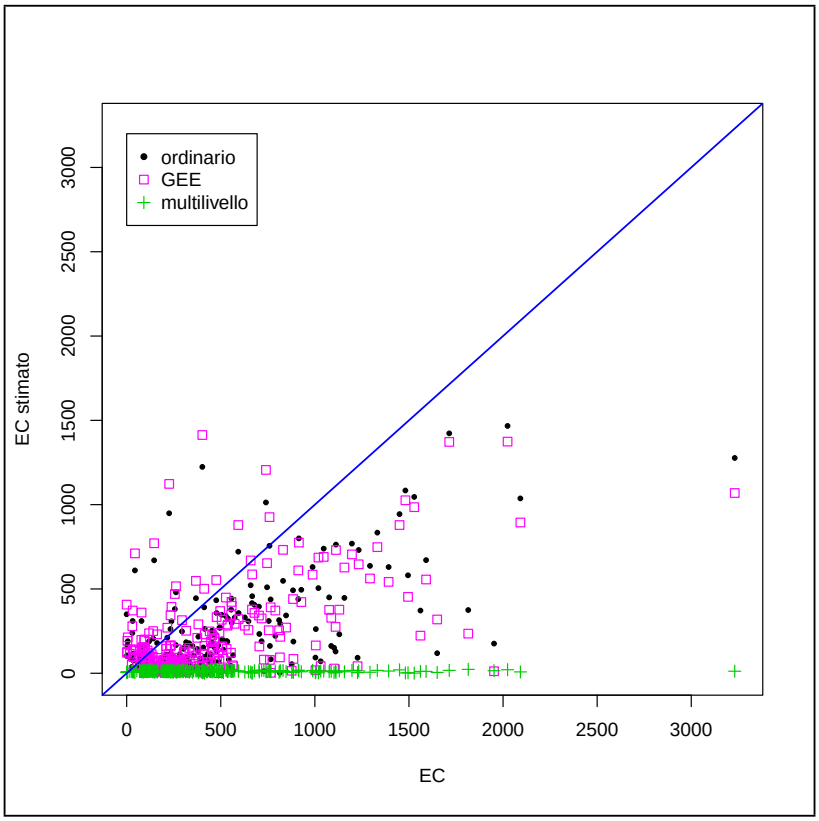


4.3 Ricostruzione per i comuni non appartenenti al campione

I risultati presentati nel paragrafo precedente mostrano inequivocabilmente la superiorità del modello (3) come strumento di stima diretta della popolazione obiettivo nei comuni campione. Tuttavia, il modello ad effetti misti non è immediatamente utilizzabile per stime indirette, ossia a fini predittivi per i comuni non compresi nel campione di stima, poiché esso dipende in modo cruciale da effetti casuali α_i specifici del comune su cui essi sono stimati e, dunque, non direttamente attribuibili ad altri comuni. Al contrario, sia il modello logistico standard (1) sia il modello marginale (2) possono essere utilizzati per fare previsione di tipo indiretto poiché non dipendono in modo così esclusivo dal campione di stima. Una strategia possibile per fare previsione fuori campione (nuovi individui, nuovi comuni) che coniughi i punti di forza del modello ad effetti misti con quelli ad effetti fissi consiste nell'utilizzare le intercette casuali per classificare i comuni in classi omogenee rispetto alla dimensione (e al segno) di α_i per poi applicare il modello ad effetti fissi entro classe. Ciò equivale a supporre che esista una variabile latente che assegna ciascun comune ad una classe con una certa probabilità.

Nella letteratura economica le stime delle intercette aleatorie sono state utilizzate per ordinare/classificare le unità di secondo livello. Ad esempio, [15] utilizzano le intercette casuali come misura del "valore aggiunto" della scuola (unità di secondo livello) sul rendimento scolastico degli studenti (unità di primo livello) al netto della loro abilità all'entrata. Nel nostro contesto, le intercette aleatorie possono essere interpretate come il contributo del singolo comune al rischio individuale di sotto e sovracopertura in aggiunta all'effetto delle caratteristiche osservate descritte dai regressori utilizzati nel modello. Le Figure 4 e 5 riportano la stima "*empirical Bayes*" [14] con il relativo intervallo di confidenza del 95% delle intercette casuali rispettivamente per i modelli di sottocopertura e sovracopertura dei 247 comuni di ampiezza 10001-50000. I grafici "a bruco" mostrano come alcuni intervalli non comprendono lo zero. Ad esempio, nella Figura 4, per 21 (3) comuni l'intervallo risulta tutto positivo (negativo). Dunque, un individuo risulterà a rischio più alto (basso) di sotto e sovraco-

Figura 3 - $|\hat{EC}|$ **contro** $|EC|$ **per modello: comuni con 10001-50000 abitanti**



pertura indipendentemente dalle proprie caratteristiche anagrafiche, per il solo fatto di abitare in uno di quei comuni.

E' sempre possibile suddividere i comuni in tre gruppi per ciascuna delle due componenti dell'errore di copertura: a) neutri (0), ovvero quelli con intercetta casuale non statisticamente diversa da zero, b) positivi (+), con intercetta casuale significativamente maggiore di zero, e c) negativi (-), con intercetta casuale significativamente minore di zero. Questa classificazione genera 9 gruppi di comuni definiti dalle possibili permutazioni delle classi suddette [5]. Le probabilità predette di sotto e sovracopertura non sono più condizionate ad un valore unico dell'intercetta casuale (pari a 0 per il modello (1) o al valore "*population average*" per il modello (2)) ma ad un valore medio specifico di classe. Ai fini della stima indiretta della popolazione abitualmente dimorante, assegnare un effetto casuale stimato su un comune campione i nell'anno t_0 ad un altro comune j al tempo t_1 (ma anche allo stesso comune i in t_1) equivale a fare un'assunzione di stabilità nel tempo degli effetti casuali α_i molto più forte rispetto ad assumere che sia l'appartenenza ad una certa classe a rimanere stabile. In particolare, è plausibile ritenere che un comune rimanga nella propria classe di assegnazione al tempo t_0 per lo meno fino al tempo t_4, ossia fino a conclusione del ciclo censuario quinquennale previsto per il nuovo censimento permanente italiano.

Per verificare la capacità predittiva del modello a "effetti fissi entro classe" o "ibrido" descritto sopra sono stati utilizzati i tre campioni di validazione già introdotti in precedenza, sui quali è stato stimato in modo indipendente il modello ad effetti misti. I comuni sono stati, dunque, assegnati a ciascuna delle 9 classi sulla base della propria intercetta casuale. Si è, quindi, proceduto a ricostruire con l'equazione 3 la popolazione censuaria di ciascuno dei nuovi comuni estratti utilizzando le stime

[5] (+/+),(+/0),(+/-),(0/+),(0/0),(0/-),(-/+),(-/0),(-/-).

Figura 4 - Stima delle intercette aleatorie per i comuni con 10001-50000 abitanti: sottocopertura

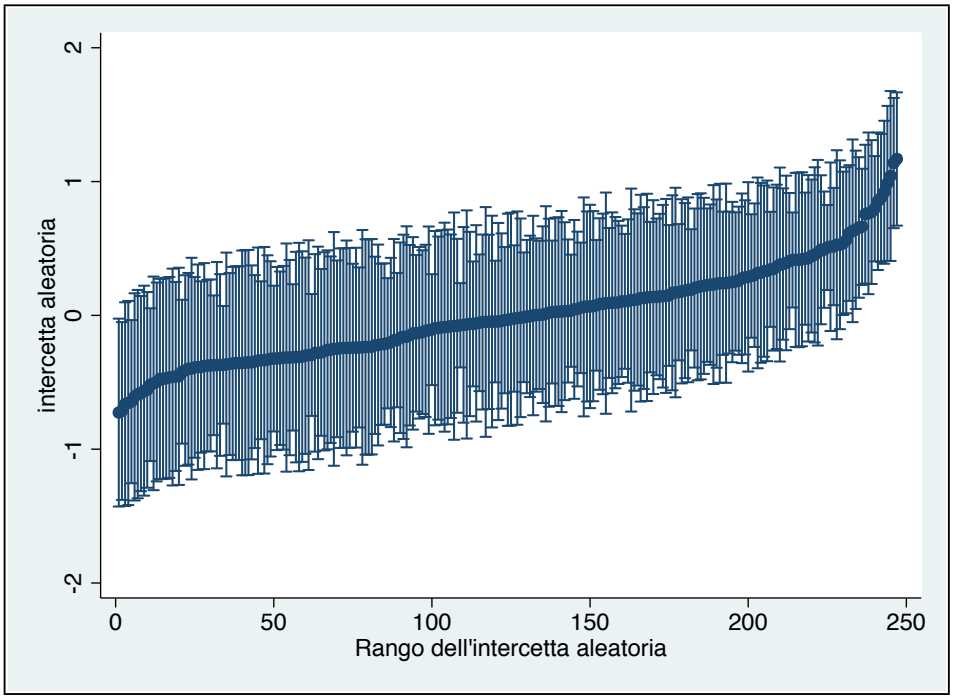


Figura 5 - Stima delle intercette aleatorie per i comuni con 10001-50000 abitanti: sovracopertura

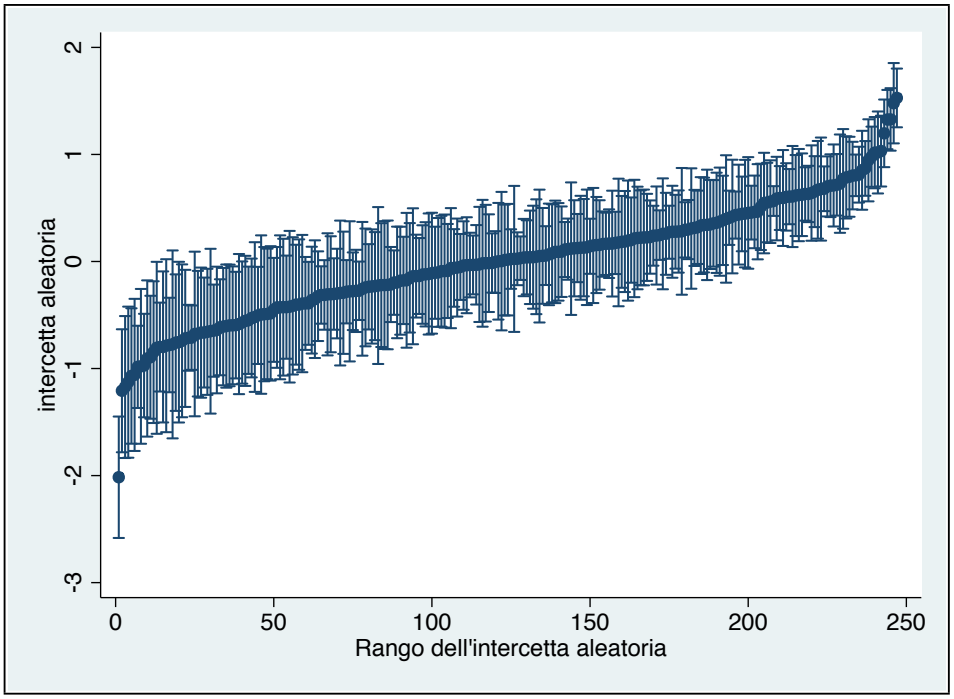


Tabella 11 - Intercetta media e sua variabilità, per classe e per tipo di campione

			Sovracopertura				Sottocopertura			
	Classe	Campione	$\overline{\hat{\alpha}}_i$	σ	N	%	$\overline{\hat{\alpha}}_i$	σ	N	%
1001-5000	Positivi	Calibrazione	1,02	0,34	79	11%	1,23	0,30	33	4%
		Validazione	1,02	0,38	91	12%	1,40	0,55	47	6%
	Negativi	Calibrazione	-0,99	0,14	13	2%	-1,15	0,00	1	0%
		Validazione	-1,08	0,19	17	2%	-	-	0	0%
5001-10000	Positivi	Calibrazione	0,75	0,27	64	21%	0,90	0,25	29	10%
		Validazione	0,70	0,23	55	18%	0,96	0,27	20	7%
	Negativi	Calibrazione	-0,82	0,21	40	13%	-0,94	0,05	4	1%
		Validazione	-0,76	0,14	30	10%	-	-	0	0%
10001-50000	Positivi	Calibrazione	0,71	0,28	54	22%	0,75	0,22	21	9%
		Validazione	0,70	0,27	56	22%	0,75	0,19	21	8%
	Negativi	Calibrazione	-0,78	0,30	39	16%	-0,70	0,04	3	1%
		Validazione	-0,71	0,27	37	15%	-0,72	0,08	3	1%

ottenute sul primo campione (di calibrazione). In particolare, per la parte fissa, sono stati usati i coefficienti del modello (1) mentre, per la parte casuale, l'intercetta media della classe di appartenenza del comune. La Tabella 11 mostra come sia l'intercetta media e la sua variabilità sia il numero e la percentuale di comuni per classe calcolata sul totale dei comuni campionati rimangano relativamente stabili tra un campione e l'altro sia per la sotto che per la sovracopertura.

Le Figure 6-8 confrontano il potere predittivo del modello ibrido sia con quello del modello ad effetti fissi sia con quello ad effetti misti stimato sul campione di validazione. Si nota come il modello ibrido ricostruisce la popolazione legale decisamente meglio non solo rispetto all'anagrafe (la maggior parte dei punti si trovano sotto la bisettrice) ma anche rispetto al modello ad effetti fissi. Nel caso della fascia 10001-50000, ad esempio, il valore ricostruito dal modello ibrido è più vicino al censimento di quanto non lo sia la LAC 90 volte su 100, ma la percentuale sale al 98% quando l'errore di copertura delle LAC è superiore a 250 individui. Rispetto al modello ad effetti fissi, il modello ibrido ricostruisce meglio 84 volte su 100, che salgono a 91 quando l'errore di copertura delle LAC è superiore a 250 individui.

Le ricostruzioni censuarie ottenute tramite l'equazione 3 costituiscono una stima puntuale non corredata di una misura di variabilità. Per valutarne la stabilità è stata prodotta una stima intervallare. Confrontando i coefficienti di regressione di parte fissa (β) stimati sul campione di calibrazione (Tabelle 5 e 6) con quelli di validazione (Tabelle 9 e 10) si nota come i valori rimangono sostanzialmente stabili. Dunque, con buona approssimazione, la variabilità associata alle probabilità individuali di sotto e sovracopertura è determinata in modo predominante dalla variabilità delle intercette aleatorie (α_i). Per tale motivo l'intervallo della ricostruzione censuaria di comune ($\hat{N}_i$) è stato calcolato esclusivamente in funzione dell'incertezza sull'intercetta casuale. Come stimatore di tale intercetta si è scelta la media delle intercette aleatorie di classe (Tabella 11). Sotto l'ipotesi di normalità degli effetti casuali medi di classe sono stati costruiti gli intervalli di confidenza del 95% i cui estremi sono poi stati utilizzati nell'equazione 3 per giungere ad una stima intervallare di ($\hat{N}_i$), calcolata come segue:

$$\hat{N}_i^H = \sum_X \hat{N}_{\underline{x}_i}^H = \sum_X \hat{w}_{\underline{x}_i}^H \cdot \text{LAC}_{\underline{x}_i} = \sum_X \frac{1 - \hat{p}^L_{\text{sovra},\underline{x}_i}}{1 - \hat{p}^H_{\text{sotto},\underline{x}_i}} \cdot \text{LAC}_{\underline{x}_i} \tag{4}$$

$$\hat{N}_i^L = \sum_X \hat{N}_{\underline{x}_i}^L = \sum_X \hat{w}_{\underline{x}_i}^L \cdot \text{LAC}_{\underline{x}_i} = \sum_X \frac{1 - \hat{p}^H_{\text{sovra},\underline{x}_i}}{1 - \hat{p}^L_{\text{sotto},\underline{x}_i}} \cdot \text{LAC}_{\underline{x}_i} \tag{5}$$

dove

$$\hat{p}^H_{\text{sovra},\underline{x}_i} = \text{invlogit}\left\{ \hat{\alpha} + (\overline{\hat{\alpha}}_i + 1,96 \cdot \sigma) + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \right\} \tag{6}$$

Figura 6 - | $\hat{EC}$ | **contro** | EC | **per modello: comuni con 1001-5000 abitanti**

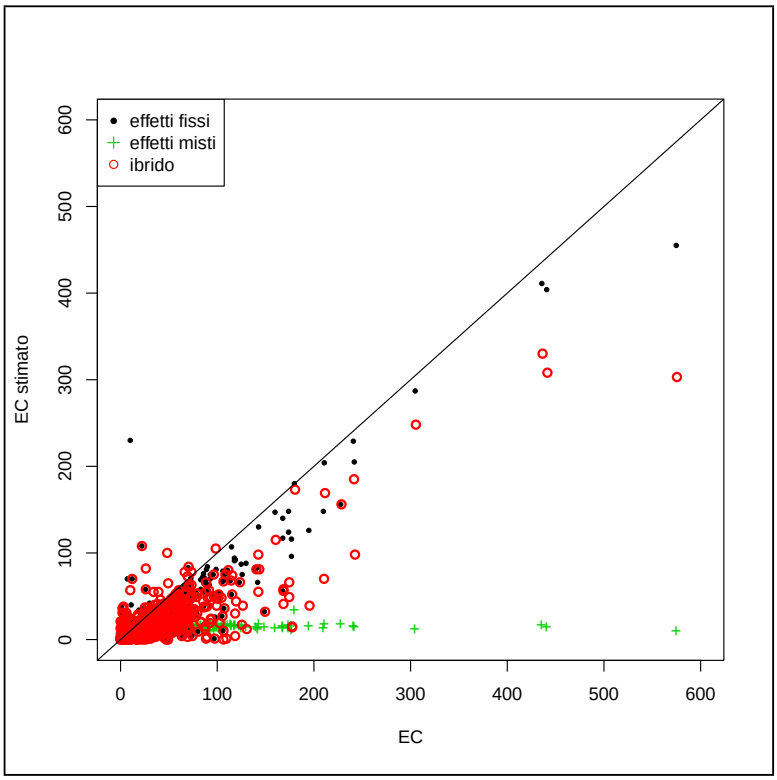


Figura 7 - | $\hat{EC}$ | **contro** | EC | **per modello: comuni con 5001-10000 abitanti**

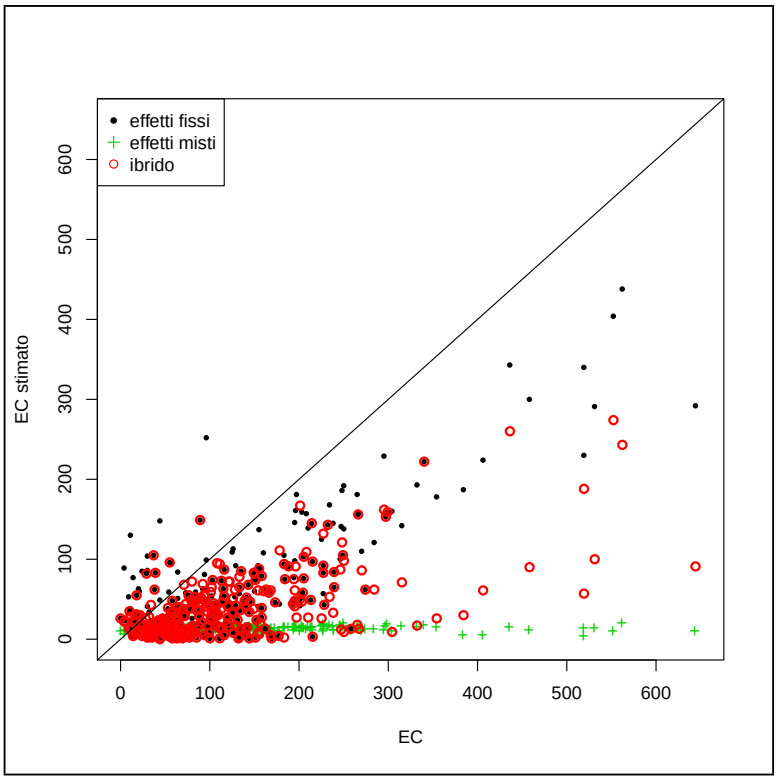
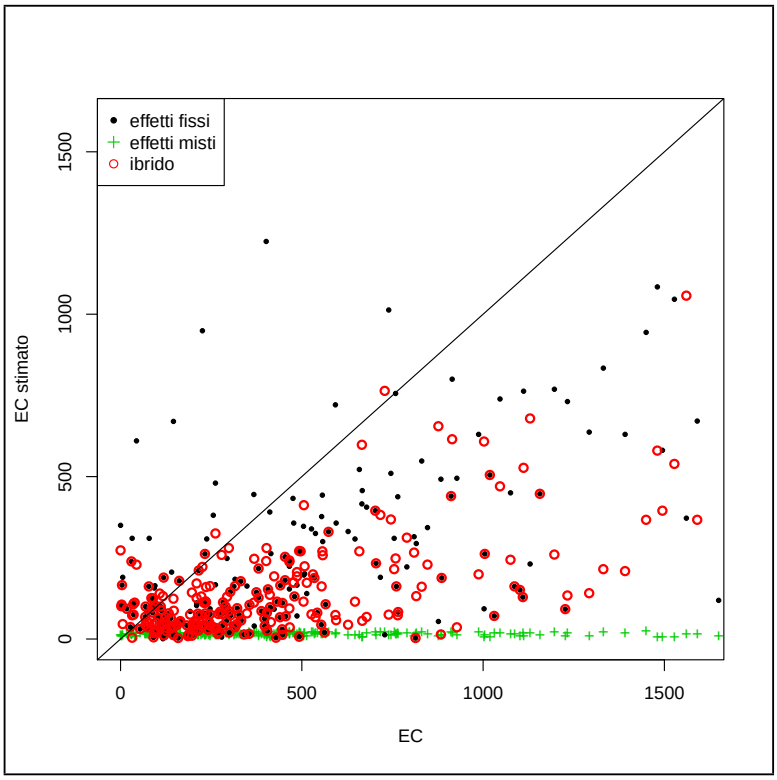


Figura 8 - $|\hat{EC}|$ **contro** $|EC|$ **per modello: comuni con 10001-50000 abitanti**



$$\hat{p}^{L}_{\text{sovra},\underline{x}_i} = \text{invlogit}\left\{\hat{\alpha} + (\overline{\hat{\alpha}}_i - 1,96 \cdot \sigma) + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k\right\} \tag{7}$$

con $\overline{\hat{\alpha}}_i$ le intercette medie di classe stimate sul campione di calibrazione. Le equazioni 4-7 evidenziano come si sia scelta la combinazione di valori delle probabilità di sotto e sovracopertura cui corrisponde il più ampio intervallo per $\hat{N}_i$ dato l'intervallo di confidenza del 95% sulle $\hat{\alpha}_i$.

Per associare alle LAC "pesate" una misura della loro variabilità, le Figure 9-11 mostrano la stima intervallare, oltre che puntuale, della ricostruzione censuaria da modello ibrido confrontandola sia con le LAC che con il censimento. Si nota, in termini di stima puntuale, come la ricostruzione (linea verde) sia sensibilmente più vicina al censimento (puntini neri) rispetto alle LAC (asterischi fucsia). In termini di stima intervallare, il valore censuario è, inoltre, quasi sempre interno all'intervallo di confidenza della sua ricostruzione, mentre i conteggi di LAC sono quasi sempre a ridosso dell'estremo superiore dell'intervallo e, talvolta, addirittura esterni ad esso.

Come evidenziato nei paragrafi precedenti, l'errore di copertura delle LAC, in entrambe le sue componenti, non è un fenomeno distribuito uniformemente sulla popolazione obiettivo ma influenza alcune sottopopolazioni più pesantemente di altre. Ad esempio, l'ultimo censimento ha confermato come la popolazione straniera che vive in Italia è particolarmente a rischio di non essere adeguatamente rappresentata dalla popolazione anagrafica. La Figura 12 mostra la ricostruzione censuaria presentata nelle Figure 9-11, ma per i soli cittadini stranieri. Per completezza è stata riportata anche la previsione $\hat{N}_i$ ottenuta dal modello ad effetti misti stimato sui nuovi comuni (linea grigia tratteggiata). Si può notare come, rispetto alla ricostruzione per l'intera popolazione, il modello ibrido è ancora più vicino al valore censuario mentre i conteggi da LAC sono spesso esterni agli estremi dell'intervallo di variabilità della stima ibrida.

Figura 9 - Stima intervallare della ricostruzione censuaria, LAC e censimento: 1001-5000 abitanti

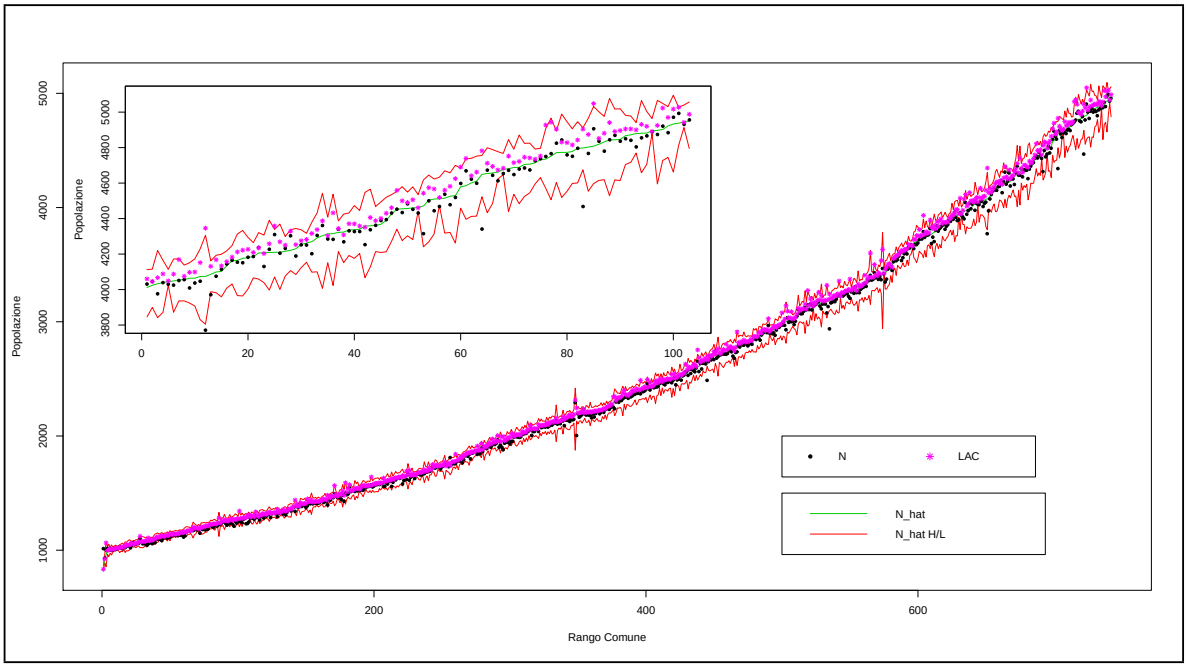


Figura 10 - Stima intervallare della ricostruzione censuaria, LAC e censimento: 5001-10000 abitanti

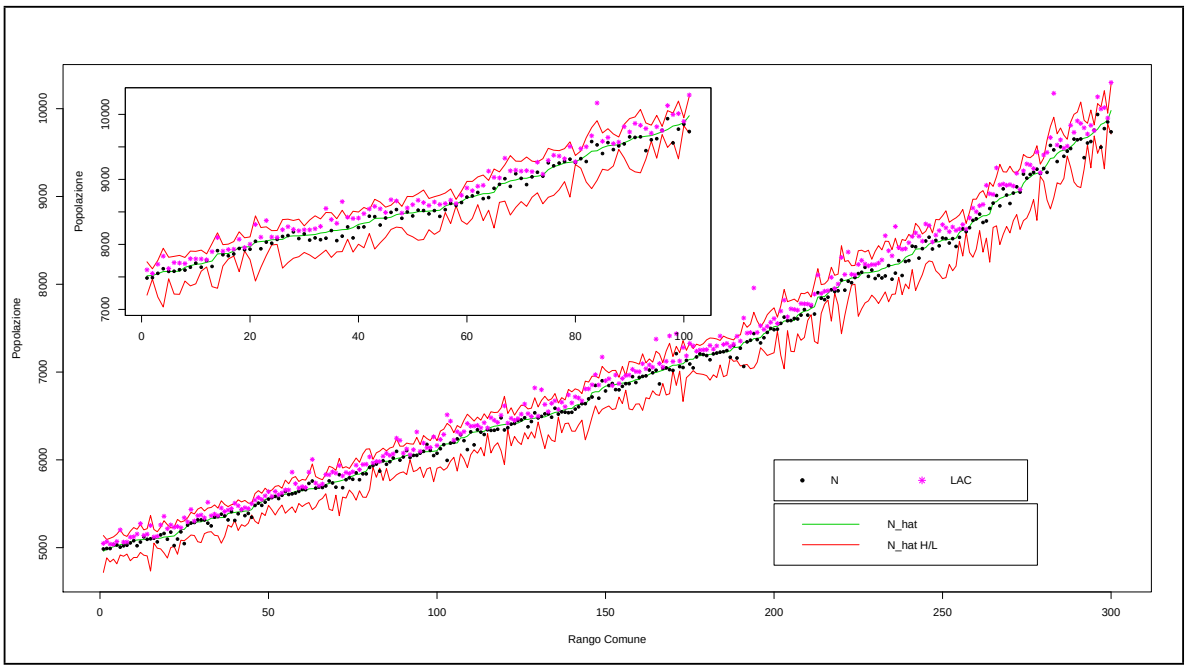


Figura 11 - Stima intervallare della ricostruzione censuaria, LAC e censimento: 10001-50000 abitanti

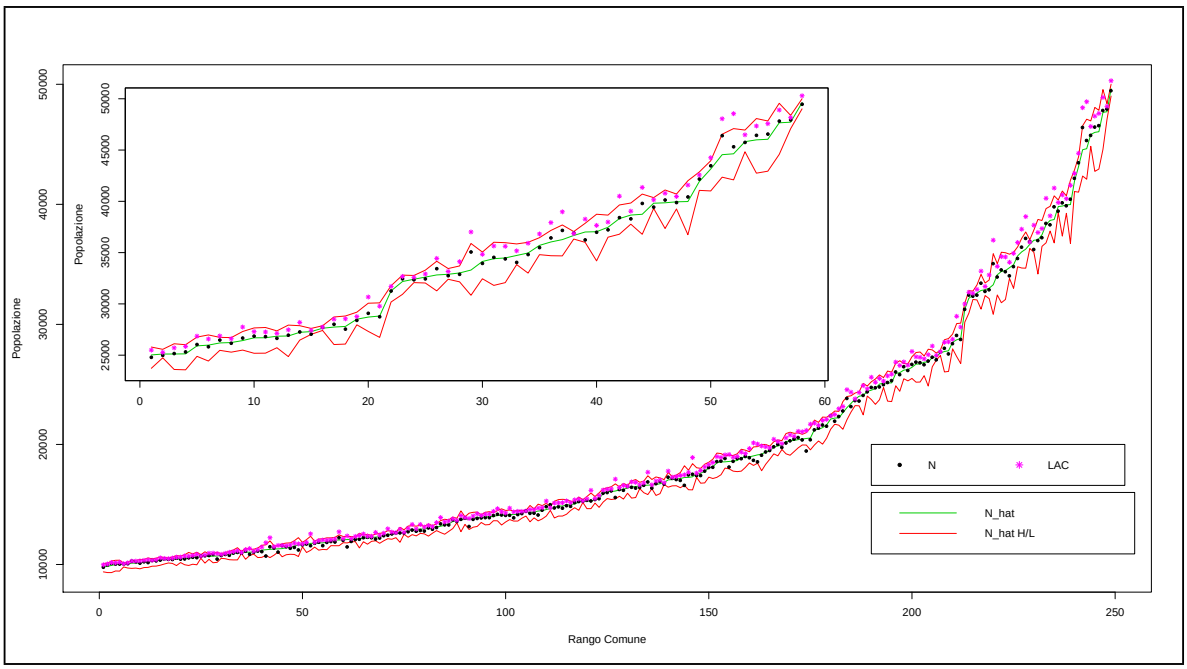
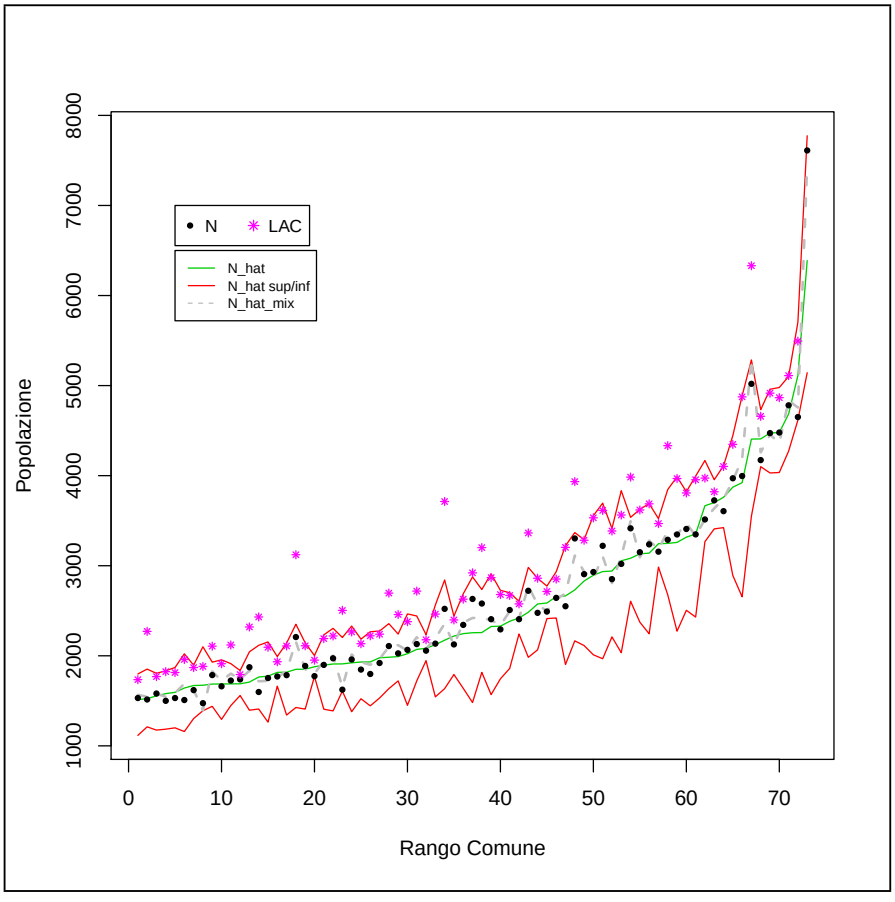


Figura 12 - Ricostruzione censuaria, LAC e censimento per i cittadini stranieri.



5. Discussione

L'analisi sulla ricostruzione del censimento a partire dalle anagrafi comunali rivela una netta superiorità predittiva del modello ad effetti misti rispetto ai modelli ad effetti fissi quando la previsione si riferisce a nuovi individui nei comuni campione. Tuttavia, quando i nuovi individui appartengono a comuni non campionati il modello ad effetti misti non è applicabile direttamente. Per migliorare il potere predittivo del modello ad effetti fissi in questi casi è stato proposto l'utilizzo di un modello ibrido ovvero di un modello ad effetti fissi applicato entro classi di comuni grossomodo omogenei rispetto ai valori delle rispettive intercette aleatorie.

La validità di un modello ibrido così costruito si regge su alcune assunzioni che potrebbero non essere sempre verificate. In primo luogo, la sua applicazione presuppone l'invarianza nel tempo della classificazione dei comuni. Sarà possibile verificare questa ipotesi con l'indagine pilota del censimento permanente che tornerà su alcuni comuni italiani nel corso del 2015. In particolare, si potrà verificare se le intercette casuali stimate sul campione pilota sono coerenti con quelle stimate al censimento. Più in generale, per valutare la stabilità del modello nel tempo, sarà possibile ripetere le stime sul campione pilota per ricostruire (a ritroso) i conteggi del censimento 2011 sia per i comuni campione che per comuni non campionati[6].

In secondo luogo, l'applicazione del modello ibrido presuppone di conoscere i valori $\hat{\alpha}_i$ per i nuovi comuni così da poterli assegnare ad una determinata classe. Nel presente studio si è ristimato il modello ad effetti misti su un campione di validazione estratto con un disegno identico al campione di calibrazione. A fini applicativi, è preferibile stimare le $\hat{\alpha}_i$ sull'intera popolazione delle unità di secondo livello estraendo un campione unico di individui.

In terzo luogo, come valore rappresentativo di classe si è scelto di utilizzare il valor medio delle intercette casuali $\overline{\hat{\alpha}}_i$. In questo modo è stata mimata la procedura di stima di un modello di variabile latente rappresentata dalla classe di appartenenza del comune. In futuro si potrebbe introdurre esplicitamente la variabile latente nel modello ottenendo stime di $\overline{\hat{\alpha}}_i$ più rigorose. Inoltre, l'introduzione della variabile latente permetterebbe di considerare anche la variabilità dell'appartenenza di un comune ad una certa classe. Nella presente analisi, infatti, l'assegnazione dei comuni alle classi è deterministica. Un'assegnazione di tipo "*fuzzy*" consentirà una migliore gestione dei casi di frontiera in cui l'attribuzione di un comune ad una certa classe è più ambigua e per i quali il modello ibrido può ricostruire il censimento in modo inaccurato.

In generale, il potere predittivo del modello potrebbe essere migliorato dall'aggiunta di nuovi regressori, soprattutto a livello individuale e familiare. Per le finalità soprattutto metodologiche del presente studio, si è scelta deliberatamente una specificazione parsimoniosa del modello di stima per i vantaggi sia computazionali sia interpretativi che essa offre. Tuttavia, l'inclusione di altre informazioni come, ad esempio, se un individuo lavora fuori comune e da quanto tempo, se l'unità abitativa in cui risiede è occupata da una sola o da più famiglie o se l'alloggio è di proprietà o in locazione/comodato d'uso, in quanto presumibilmente legate al fenomeno della copertura dei registri anagrafici, dovrebbe portare ad un aumento dei valori di sensibilità e specificità delle stime ad effetti fissi.

6. Considerazioni conclusive

Le liste anagrafiche comunali (LAC) hanno avuto un ruolo importante nell'ultima tornata censuaria del 2011. Il confronto tra la popolazione censita e quella anagrafica ha avuto un esito incoraggiante evidenziando una sostanziale sovrapposizione tra le due popolazioni a livello nazionale. Questo ri-

[6] L'indagine pilota 2015 prevede un campionamento di tipo areale basato sulle sezioni di censimento. Sebbene l'analisi qui presentata non tenga conto della sezione, il modello multilivello utilizzato può essere facilmente esteso ad un disegno di campionamento areale, prevedendo un ulteriore livello (intermedio) rappresentato dalla sezione di censimento. Inoltre, è stato verificato (risultati non riportati ma disponibili su richiesta) che, replicando l'analisi su campioni areali estratti secondo il disegno base previsto per l'indagine pilota, le stime ottenute sono molto simili a quelle presentate in questo lavoro.

sultato ha spinto l'Istituto Nazionale di Statistica ad utilizzare le LAC per l'estrazione dei campioni di tutte le indagini socio-economiche a partire dal 2011 e ne fa presagire un ruolo sempre più centrale anche a fini censuari: un censimento non più semplicemente assistito da archivi amministrativi ma potenzialmente sostituito da essi per una parte rilevante della rilevazione. La riforma annunciata del censimento italiano della popolazione da decennale a permanente e gli investimenti fatti negli ultimi anni dall'Istat sul potenziamento dei propri archivi come il Sistema Integrato dei Microdati e l'Archivio Nazionale dei Numeri Civici puntano decisamente in questa direzione.

Il confronto censimento-anagrafe del 2011 ha, tuttavia, evidenziato differenze significative, sia per conteggio che per composizione, tra le due popolazioni per domini territoriali sub-nazionali, soprattutto a livello comunale. L'analisi presentata in questo articolo ha investigato tali divergenze utilizzando dati individuali sia di fonte censuaria sia amministrativa per un campione di comuni italiani suddivisi per fasce di ampiezza demografica. A tal fine sono stati stimati modelli di regressione logistica sotto ipotesi alternative sulla struttura di varianza delle variabili risposta, definite rispettivamente come sottocopertura (individuo abitualmente dimorante ma non anagraficamente residente nel comune) e sovracopertura (individuo iscritto ma non abitualmente dimorante).

I risultati confermano come alcune categorie di individui (i cittadini stranieri, gli individui giovani, i nuclei familiari monocomponenti) sono sensibilmente più a rischio di mancata o errata copertura nelle LAC. L'analisi rivela, 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.

7. Ringraziamenti

L'articolo ha beneficiato del supporto e della condivisione di basi infomative da parte dei membri del gruppo di lavoro METOPOP. Ringraziamo, poi, Gerardo Gallo per i commenti a versioni precedenti dell'articolo che hanno contribuito a migliorarne la qualità. Siamo, inoltre, grati ad Anna Pezone (LAC) e Rossella Molinaro (Basi Territoriali ISTAT) per aver condiviso con noi alcune delle informazioni di tipo amministrativo utilizzate nell'analisi. Ringraziamo, infine, Roberta Radini e Maria Pia Di Maio per il supporto informatico sull'estrazione dei record individuali dal database del censimento 2011. La responsabilità per quanto scritto è esclusivamente degli autori.

Bibliografia

Poulain, M. and Herm, A. "Le registre de population centralisé, source de statistiques démographiques en Europe", *Population* no. 2 (vol. 68) 2013: 183-212.

Mancini, A. "Latest Innovation of Italian Population Census", *Rivista Italiana di Economia Demografia e Statistica*, no. LXV, no. 2 April-June 2011.

Ceccarelli, C. e Rosati, S. "L'utilizzo delle Liste Anagrafiche Comunali", Seminario su *Le innovazioni metodologiche nelle indagini socio-economiche sulle famiglie*, Roma, 20 maggio 2014.

Atkinson, A. B. and Marlier, E. *Income and Living Conditions in Europe*, Eurostat Statistical Books, Luxemburg: Publication Office of the European Union.

Mancini, L., Fortini, M., Marcone, L., Borrelli, F. and Ronconi, A. "Assessing the Effectiveness of Administrative Registers in Reducing Under-coverage Errors in a Population Census: Evidence from the 2009 Italian Census Pilot Study", *Statistics in the 150 years from Italian Unification. 2011 Statistical Conference Bologna*, 8-10 June 2011. Book of Abstracts, p. 59-60, Bologna, Italy, Alma Mater Studiorum Acta.

Fortini, M., Mancini, L., Marcone, L., Mussino, E. and Paluzzi, E. "Who Settles Down in Italy? Transition to Residency of non-EU Migrants", *Rivista Italiana di Economia Demografia e Statistica*, no. LXVII, no. 3/4 July-December 2013.

Calzaroni, M., Crescenzi, F., Fortini, M., Mancini, A. e Sindoni, G. "Linee strategiche su metodi, tecniche e organizzazione del Censimento permanente della popolazione e delle abitazioni" (2013). Technical Report to the Steering Committee for the design of the Italian Permanent Census.

Gallo, G. e Tamburrano, T. "Popolazione censita e popolazione anagrafica al 2011: valutazioni sulle differenze quantitative tra le due fonti", 2013. *Unpublished manuscript*.

Office of National Statistics (2010). "The 2001 Hard to Count Index", *One Number Census Steering Committee Key Paper* no. 00/15.

Alho, J. M., Mulry, M.H., Wurdeman, K. and Kim, J. "Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation", *Journal of the American Statistical Association* no.88-423(1993):1130-1136.

Garofalo, G. "Archimede e la Statistica Comunale", Convegno *USCI*, Messina, 26/27 Settembre 2013. (http://www.usci.it/file.pdf/Convegno2013/garofalo.pdf)

Liang, K.-Y., and Zeger, S.L. "Longitudinal data analysis using generalized linear models", *Biometrika* no.73 (1986):13–22.

Skrondal, A. and Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton: Chapman and Hall–CRC, 2004.

Snijders, T. A. B. and Bosker, R. J. *Multilevel Analysis*, 2nd edition. Sage Publication, London, 2011.

Grilli, L. and Rampichini, C. "A review of random effect modelling using Stata. A review written for the Multilevel Modelling Software Reviews of The Centre for Multilevel Modelling", *University of Bristol* (2006). (http://www.bristol.ac.uk/cmm/learning/mmsoftware/gllamm.html).

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. "Generalized Multilevel Structural Equation Modelling", *Psychometrica*, no.69(2) (2004):167-190.

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

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

Skrondal, A. and Rabe-Hesketh, S. “Prediction in Multilevel Generalized Linear Models”, *Journal of the Royal Statistical Society*, Series A (Statistics in Society), no. 172-3 (2009):659-687.

Goldstein, H. and Healy, M.J. “The Graphical Presentation of a Collection of Means”, *Journal of the Royal Statistical Society*, Series A (Statistics in Society), no.158-1(1995):175-177.

Informazioni per gli autori

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

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

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

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Salvo diversa indicazione la riproduzione è libera, a condizione che venga citata la fonte.