istat
working
papers

N. 15
2012

La progettazione dei censimenti generali 2010-2011: misure di accuratezza di tavole di diffusione per livelli territoriali e dettagli informativi

Francesco Borrelli, Giancarlo Carbonetti, Silvia Dardanelli, Luana De Felici, Epifania Fiorello, Manuela Marrone e Mariangela Verrascina

Istat
Istituto Nazionale di Statistica

istat
working
papers

N. 15
2012

La progettazione dei censimenti generali 2010-2011: misure di accuratezza di tavole di diffusione per livelli territoriali e dettagli informativi

Francesco Borrelli, Giancarlo Carbonetti, Silvia Dardanelli, Luana De Felici, Epifania Fiorello, Manuela Marrone e Mariangela Verrascina

Istat

Istat Working Papers

La progettazione dei censimenti generali 2010-2011: misure di accuratezza di tavole di diffusione per livelli territoriali e dettagli informativi

N. 15/2012

ISBN 88-458-1736-9

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

La progettazione dei censimenti generali 2010-2011: misure di accuratezza di tavole di diffusione per livelli territoriali e dettagli informativi

Francesco Borrelli, Giancarlo Carbonetti, Silvia Dardanelli, Luana De Felici, Epifania Fiorello, Manuela Marrone e Mariangela Verrascina

Sommario

La progettazione del Censimento della Popolazione e delle Abitazioni del 2011 ha offerto una straordinaria opportunità per proporre innovazioni di tipo metodologico, tecnologico e organizzativo. La scelta finale è stata quella di un censimento da lista anagrafica con l'adozione, nei comuni più grandi, di una strategia campionaria tramite short/long form per la raccolta dei dati censuari. La soluzione adottata prevede di rilevare in modo esaustivo solo i dati demografici, familiari e alcune delle principali informazioni socio-economiche tramite un questionario ridotto e riservare a campioni di famiglie la rilevazione dell'intero insieme di variabili tramite un questionario completo. Questa decisione da un lato comporta un costo statistico, in quanto si passa dall'osservazione completa di tutte le variabili a un sistema che prevede l'integrazione di dati esaustivi e dati stimati, dall'altro permette di migliorare l'efficienza delle operazioni sul campo, di ridurre il fastidio statistico sui rispondenti e di offrire maggiore qualità. Essendo la prima esperienza in Italia, si è proceduto in modo sperimentale su dati del 2001 per definire i livelli di accuratezza delle stime e per valutare i possibili riflessi sulla qualità dei risultati finali. Gli errori campionari sono stati testati su alcune tavole statistiche impiegate per diffondere i risultati del Censimento del 2001 e su alcuni ipercubi richiesti da Eurostat, per diversi livelli di dettaglio informativo e per differenti ambiti territoriali.

Parole chiave: censimento, long form, campionamento, ipercubi, accuratezza.

Abstract

The sampling strategy adopted by the Italian National Institute of Statistics (Istat) in 2011 General Population and Housing Censuses was based on the simultaneous use of short and long forms: the first one containing demographic variables and a few socio-economic data; the second including the overall set of census variables. In this way, demographic data on the entire population will be collected while information related to other variables will be surveyed only on a sample of households (private households). This sampling strategy regarded capital provinces and municipalities with population of over 20,000 inhabitants; for smaller municipalities a traditional approach was adopted and the long form was submitted to the entire population. Preliminary tests and studies have been conducted in order to evaluate the efficiency of sampling estimates and the accuracy of dissemination hypercubes (statistical tables obtained by cross-classification of census variables).

Keywords: census, long form, sampling, hypercubes, accuracy.

Indice

1. Introduzione[1]

Il censimento della popolazione e delle abitazioni ha da sempre rappresentato l'occasione per la costituzione di un patrimonio informativo unico e di fondamentale importanza per la collettività. I dati raccolti garantiscono una conoscenza ad un elevato grado di dettaglio territoriale non deducibile da alcuna altra fonte né da altro tipo di indagine e vengono richiesti ed impiegati ad ogni livello di governo e da un'ampia e diversificata utenza per fini di valutazione, programmazione e decisione (Berntsen *et al.*, 2008).

Nonostante i contenuti dei censimenti abbiano subito delle evoluzioni nel corso del tempo, sono state molteplici le ragioni che hanno spinto a proporre innovazioni rispetto alle modalità con cui la rilevazione è stata condotta nel passato. La necessità di realizzare un censimento più "leggero", con l'auspicio da un lato di ridurre il carico di lavoro dei soggetti coinvolti nelle operazioni sul campo e dall'altro di rilevare un insieme limitato di informazioni su tutta la popolazione, ha dato l'opportunità di cambiamenti che hanno investito l'impianto organizzativo, le scelte metodologiche e gli avanzamenti tecnologici.

Le innovazioni di carattere metodologico nell'ambito del censimento hanno seguito due principali direzioni: l'integrazione di dati provenienti da fonte amministrativa e l'introduzione delle tecniche campionarie per la rilevazione di alcune delle informazioni non strettamente demografiche solo su campioni di famiglie.

La decisione di ottenere una parte dei dati tipici del censimento tramite campioni deve però affiancarsi alla consapevolezza del costo statistico che questo tipo di soluzione implica e alla capacità di convincere gli utilizzatori che un buon campione permette di raggiungere risultati equivalenti e, per certi aspetti, addirittura migliori di quelli provenienti da una rilevazione totale.

Questo lavoro fa riferimento alla necessità di valutare alcuni possibili riflessi che l'introduzione della strategia campionaria produce sulla qualità dell'informazione censuaria prodotta e diffusa per differenti livelli territoriali.

Il documento inizia descrivendo alcune criticità del Censimento della popolazione e delle abitazioni del 2001 e le principali soluzioni innovative introdotte in quello del 2011 (Capitolo 2), continua presentando alcuni elementi caratterizzanti il piano di diffusione dei risultati censuari sia sul versante europeo che su quello italiano (Capitolo 3) e, successivamente, illustra i contenuti informativi dei questionari utilizzati per la rilevazione censuaria (Capitolo 4).

La trattazione prosegue esponendo nel dettaglio la strategia campionaria adottata per il 2011, basata sull'impiego di questionari di tipo *short* e *long* (Capitolo 5). Di seguito, si delinea la metodologia sviluppata per le valutazioni qualitative sull'accuratezza delle tavole di diffusione dei risultati censuari (Capitolo 6) e si descrivono le tavole statistiche, scelte per differenti livelli di dettaglio territoriale, prese a scopo di esercizio nello studio (Capitolo 7). Nel capitolo 8 si illustra l'ambiente informatico di supporto all'intera fase di analisi.

Il lavoro continua con la presentazione dei principali risultati delle valutazioni sull'accuratezza delle tavole statistiche scelte con riferimento al livello regionale (Capitolo 9) e ai livelli comunale e sub-comunale (Capitolo 10). Si indicano, inoltre, alcuni elementi sul possibile impatto derivante da un ampliamento del contenuto informativo del questionario in forma ridotta (Capitolo 11). Il lavoro si conclude con alcune considerazioni di sintesi e alcuni suggerimenti per il futuro (Capitolo 12).

[1] Il presente lavoro raccoglie alcuni elementi emersi nel Gruppo di Lavoro sugli Approfondimenti delle attività della Task Force Eurostat on Implementation of Legislation on Population and Housing Censuses e sull'analisi delle Recommendations for the 2010 Censuses of Population and Housing, costituito con delibera n. 116/DPTS del 12 dicembre 2007. In particolare, sono esposti i risultati di un insieme di studi condotti a supporto della fase di progettazione del 15° Censimento della Popolazione e delle Abitazioni, con lo scopo di produrre elementi oggettivi utili a decidere in merito alla strategia campionaria introdotta al censimento e alla struttura finale dei questionari di rilevazione.
La redazione del documento è frutto della collaborazione degli autori. Ai fini dell'attribuzione delle singole parti si specifica che: G. Carbonetti ha curato i capitoli 1, 2, 6, 12 e i paragrafi 5.3, 5.4, 9.1, 9.4; M. Verrascina ha redatto i paragrafi 3.1, 4.1, 4.2, 7.1 e 7.2; S. Dardanelli ha curato i paragrafi 3.2, 4.3 e 7.3; E. Fiorello ha curato i paragrafi 5.1, 5.2, 9.2 e 9.3; M. Marrone ha redatto il capitolo 8; L. De Felici ha curato i paragrafi 10.1, 10.2 e l'Appendice; F. Borrelli ha curato i paragrafi 10.3 e 10.4; il capitolo 11 è stato redatto da L. De Felici e F. Borrelli.

2. Le innovazioni introdotte nel Censimento della popolazione e delle abitazioni del 2011

Il censimento della popolazione, poiché si riferisce alla totalità della popolazione presente sul territorio nazionale, è la rilevazione più complessa e impegnativa in termini di risorse economiche e di pianificazione delle attività sul campo (Fortini *et al.*, 2007). Dall'analisi sulla conduzione del passato censimento sono emerse rilevanti criticità nella predisposizione delle operazioni censuarie, tra cui: costituzione, coordinamento e mantenimento della considerevole rete di rilevatori; gestione delle fasi di consegna e ritiro dei questionari. Occorre inoltre aggiungere che, a partire dalla tornata censuaria del 2011 il Censimento italiano, così come quello di tutti i Paesi membri dell'Unione Europea, è sottoposto a Regolamento Europeo che pone vincoli[2] sui tempi (consegna dei dati entro il 1° Aprile 2014), sulle variabili obbligatorie (core topics), sulle classificazioni (breakdowns) e sulle tavole statistiche (hypercubes).

Al fine di migliorare l'efficienza delle operazioni censuarie sul campo e di rispettare gli obblighi sui tempi di rilascio dei risultati finali, per la realizzazione del Censimento del 2011 è stata decisa una strategia caratterizzata dalle seguenti innovazioni:

- diversificazione di metodi e organizzazione tra comuni aventi diversa ampiezza demografica;
- disegno di aree di censimento sub-comunali (Astorri *et al*., 2007; Bianchi *et al*., 2010) per la diffusione dei risultati ad un più elevato livello di dettaglio territoriale;
- realizzazione di archivi comunali di numeri civici geocodificati alle sezioni di censimento;
- impiego di liste pre-censuarie derivate dalle anagrafi comunali per la spedizione postale dei questionari;
- uso congiunto di questionari ridotti (*short form*) e questionari completi (*long form*);
- consegna postale dei questionari;
- multicanalità per la raccolta dei questionari (postale, web, centri di raccolta comunali).

Le azioni descritte sono finalizzate a garantire una maggiore flessibilità dell'organizzazione sul territorio, una più elevata specializzazione degli organi interessati, una riduzione significativa del numero di rilevatori (front-office) con un contestuale rafforzamento delle capacità di coordinamento e controllo degli Uffici Comunali di Censimento (back-office) coinvolti nell'intero processo.

3. Il piano di diffusione dei risultati censuari

3.1 Il contesto europeo

3.1.1 La normativa internazionale

Analogamente al 2001, la Commissione Economica per l'Europa delle Nazioni Unite (UNECE) in cooperazione con l'Ufficio Statistico della Comunità Europea (Eurostat) ha redatto un documento contenente le *Recommendations for the 2010 Censuses of Population and Housing*, formalmente adottate a giugno 2006, in occasione della Conferenza degli Statistici Europei. Le Raccomandazioni internazionali forniscono consigli per la determinazione dei contenuti informativi dei censimenti demografici in termini di variabili da rilevare, definizioni e classificazioni. Contengono, infatti, indicazioni per una definizione chiara dei concetti, per una sincronizzazione delle operazioni di rilevazione e di quelle per la produzione dei dati censuari, al fine di garantire la comparabilità degli output nei diversi paesi.

Le variabili presentate nelle Raccomandazioni sono suddivise in *core topics* (da inserire nel piano di rilevazione obbligatoriamente) e *non core topics* (opzionali).

[2] Cfr. paragrafo 3.1.1.

La sezione dedicata alla Popolazione (*Population topics*) è suddivisa in diverse aree tematiche che riguardano: Popolazione da rilevare (il campo di osservazione), Caratteristiche geografiche, Caratteristiche demografiche, Caratteristiche economiche, Caratteristiche sull'istruzione, Migrazioni interne e internazionali, Caratteristiche delle famiglie, delle convivenze e dei nuclei familiari, Difficoltà nelle attività della vita quotidiana. La sezione dedicata agli Alloggi (*Housing topics*) è costituita da un'unica area tematica sulle caratteristiche relative agli alloggi e agli edifici ad uso residenziale.

Ciò che contraddistingue le Raccomandazioni UNECE 2010 rispetto alla precedente versione è un approccio di tipo *output oriented*: viene accuratamente definito ciò che ciascun paese deve fornire in termini di dati, lasciando però la libertà di scegliere la metodologia ritenuta più opportuna. La prima parte delle Raccomandazioni è dedicata ai possibili approcci metodologici per la raccolta dei dati: censimento tradizionale, censimento basato sui registri, combinazioni di queste con indagini campionarie, oppure indagini con campioni a rotazione (*rolling census*).

Infatti, negli ultimi anni diversi paesi hanno adottato metodi di conduzione dei censimenti alternativi a quello convenzionale (basato sulla rilevazione sul campo, esaustiva e periodica), orientandosi verso l'utilizzo dei dati amministrativi a fini statistici e verso l'impiego delle tecniche di campionamento per la rilevazione sul campo.

A differenza dei censimenti del 2001, per i quali i Paesi Membri avevano sottoscritto un *Gentlemen's Agreement*, la Commissione Europea ha deciso, per la tornata censuaria del 2010-2011, di procedere con la redazione di un *Framework Regulation*, Regolamento Quadro del Parlamento Europeo e del Consiglio relativo ai censimenti della popolazione e delle abitazioni.[3] Il Regolamento Quadro nasce dall'esigenza di garantire la conformità con le Raccomandazioni internazionali, armonizzare i contenuti, sincronizzare i tempi e assicurare maggiore qualità e comparabilità dei dati prodotti nei diversi paesi. Esso è contraddistinto dallo stesso approccio che caratterizza le nuove Raccomandazioni UNECE, volto a garantire l'uniformità dell'*output* delle rilevazioni censuarie, indipendentemente dalle tecniche e dai metodi utilizzati. Il *Framework Regulation* pone le basi per la definizione di un programma armonizzato di diffusione dei dati censuari a livello europeo, elencando in allegato i *Topics to be covered in the Population and Housing census*. Si tratta dei *topics* identificati come *core* nell'ambito delle *Recommendations 2010* e riguardano caratteristiche demografiche, sociali ed economiche delle persone, ma anche aspetti legati alle famiglie, ai nuclei familiari ed agli alloggi. Il Regolamento specifica, inoltre, quali di questi *topics* sono obbligatori fino al livello di dettaglio geografico LAU2 (comunale per l'Italia) e quali solo fino al livello NUTS2 (regionale). L'obiettivo prioritario è quello di garantire la coerenza nel contenuto dei *topics*, sia riguardo agli aspetti definitori che a quelli classificatori, al fine di rendere possibili le comparazioni tra gli Stati Membri.

In aggiunta al Regolamento sono stati predisposti anche alcuni *Implementing Regulations* (Regolamenti di attuazione), che riguardano le classificazioni e le specifiche tecniche (*Implementing Regulation on population and housing censuses as regards the technical specifications of the topics and their breakdowns*) e gli ipercubi (*Programme of the statistical data and of the metadata for population and housing censuses*).[4]

3.1.2 Le tavole previste per la diffusione europea

I due *Implementing Regulation* descrivono il piano di diffusione europeo per il Censimento della popolazione del 2011. In particolare, i *breakdowns* rappresentano le classificazioni che dovranno essere applicate ai *core topics* del Regolamento. Per ogni *topic* sono stati sviluppati uno o più *breakdowns* che si adattano a diffusioni per diversi livelli di dettaglio geografico (nazionale e regionale oppure provinciale e comunale) e informativo (più o meno fine). Per alcune caratteristiche

[3] Il Regolamento è stato adottato a maggioranza dal Parlamento Europeo a febbraio, approvato a luglio 2008 e pubblicato nella Gazzetta Ufficiale dell'Unione europea ad agosto 2008.

[4] A questi se ne aggiunge un altro che riguarda la qualità dei dati (*Implementing Regulation as regards the report on the quality of the transmitted data and the technical format for the data transmission*).

(ad esempio sesso, stato civile) è prevista un'unica classificazione che si applica a qualsiasi livello di dettaglio territoriale di diffusione. Altri *topics* sono presentati con due o tre classificazioni che differiscono per numero di modalità; l'uso dell'una piuttosto che dell'altra varia a seconda dell'incrocio in cui le variabili vengono proposte. Sulla base della loro dimensione, le classificazioni si distinguono in: *High breakdowns* (H), *Medium breakdowns* (M) e *Low breakdowns* (L), diminuendo progressivamente il numero di modalità. Le classificazioni sono state pensate in funzione del loro utilizzo all'interno degli ipercubi europei. Quelle più ampie ed articolate sono applicate, di norma, ai dati presentati per i livelli territoriali di minore dettaglio (nazionale e regionale); tuttavia la scelta del dettaglio classificatorio dipende anche dal numero delle variabili coinvolte nelle tavole da diffondere.

Gli *hypercubes* rappresentano il piano di diffusione dei dati censuari per il 2011, ovvero il piano degli incroci che ciascun Paese Membro dovrà rendere disponibile a Eurostat entro 27 mesi dalla fine dell'anno di riferimento della rilevazione censuaria (31 marzo 2014). La dimensione di un ipercubo è data dal prodotto del numero di modalità previste per ciascuna variabile in esso considerata. Sono stati definiti ipercubi con un numero ridotto di incroci, anche allo scopo di produrre tavole il cui contenuto sia sicuro per la diffusione. Il numero di celle (relative ai possibili incroci tra le modalità di diverse variabili) deve, infatti, essere limitato anche in considerazione del fatto che potrebbero comparire celle con un numero molto ridotto di osservazioni con possibili ripercussioni sulla riservatezza e sulla significatività statistica (nel caso in cui i dati provengano da operazioni di campionamento).

Data la complessità dei compiti assegnati, Eurostat, supportato dalla *Task Force on the Implementation of Legislation on Population and Housing Censuses in the European Union,*[5] ha deciso di focalizzare l'attenzione solo sulle variabili obbligatorie (*core topics*), le uniche riportate nell'allegato al *Framework Regulation* e lasciando su base volontaria la predisposizione di tavole aventi per oggetto i *non core topics* in relazione ai quali gli istituti nazionali di statistica non sono soggetti ad alcun vincolo di fornitura a Eurostat.

Gli ipercubi presenti nel Regolamento (Tavola 1) coprono tutti gli aspetti relativi agli individui (caratteristiche demografiche, economiche, sull'istruzione), ma anche relativi agli occupati al luogo di lavoro, alle famiglie, ai nuclei familiari, agli individui nelle famiglie e nei nuclei e relativi alle abitazioni e agli alloggi. Il dettaglio territoriale è prevalentemente quello regionale; il livello di classificazione NUTS3 (corrispondente a quello provinciale italiano) permette, rispetto a quello comunale, di dare maggiore flessibilità nel caso in cui si presentino, per alcuni incroci, problemi di riservatezza ovvero nel caso in cui ipercubi previsti a livello comunale abbiano molte celle con basse frequenze. Sono previsti, infine, alcuni ipercubi a livello nazionale che hanno come oggetto la popolazione residente totale e che incrociano esclusivamente variabili demografiche.

Tavola 1 - Classificazione delle tavole richieste da Eurostat per il Censimento del 2011

	Caratteristiche Demografiche	Caratteristiche Economiche	Occupati al luogo di lavoro	Caratteristiche sull'Istruzione	Famiglie e nuclei	Alloggi e abitazioni	Totale
National (nazionale)	9						9
NUTS2 (regionale)	31	47	17	11	17	15	138
NUTS3 (provinciale)	15				7	3	25
LAU2 (comunale)	1				2	2	5
Totale	**56**	**47**	**17**	**11**	**26**	**20**	**177**

[5] Composta da 7 paesi tra cui l'Italia.

3.2 Il contesto nazionale

3.2.1 L'esperienza italiana del 2001

Il sistema di diffusione dei dati del 14° Censimento generale della popolazione e delle abitazioni è stato innovativo rispetto al passato; infatti, i dati definitivi, oltre che attraverso i tradizionali fascicoli su base territoriale,[6] sono stati diffusi per la prima volta anche tramite un *Data Warehouse*, una banca dati accessibile via *internet* sia dal sito dell'Istat (www.istat.it) sia da quello dedicato ai censimenti (http://censimenti.istat.it). Il sistema informativo realizzato permette all'utente di navigare tra le tavole senza percorsi di consultazione predefiniti, nel rispetto dei vincoli di coerenza e significatività espressi dai dati, individuando autonomamente tutte le informazioni necessarie per i diversi livelli territoriali e con l'opportunità di trasferirle direttamente sul proprio *computer*. È presente anche un sistema di cartografia interattiva che consente di visualizzare cartogrammi tematici per alcune delle tavole accessibili e di effettuare operazioni sulle carte (ingrandimenti, spostamenti, ricerche, associazione di informazioni, eccetera). È possibile, inoltre, consultare i *report* (comunicati stampa, note per la stampa, eccetera) relativi ai vari rilasci effettuati nel corso degli anni e alcune basi di dati. La diffusione dei *report*, della cartografia e dei dati *on line*, che peraltro ha preceduto quella su supporto cartaceo, ha permesso, in linea con le strategie adottate in altri Paesi, di pubblicare i risultati definitivi "a moduli per aree tematiche", ovvero in date diverse in funzione delle variabili considerate per tutti i livelli territoriali, dal dettaglio nazionale fino al comunale (Carbonetti *et al.*, 2008a).

I piani di diffusione dei censimenti generali del 2000-2001 sono stati, dunque, caratterizzati dal rilascio dei dati *on line*. La scelta strategica di fornire la più ampia offerta informativa di dati censuari via *internet* ha seguito la politica generale fatta propria dall'Istituto Nazionale di Statistica orientata a incrementare e sviluppare le banche dati e i sistemi informativi attivi sul sito istituzionale dell'Istat attraverso i quali rilasciare all'utenza, in modo tempestivo e diretto, una parte sempre più consistente dei dati statistici prodotti (Berntsen *et al.*, 2008).

Per quanto riguarda le pubblicazioni cartacee, sono stati prodotti fascicoli territoriali regionali, provinciali e relativi ai grandi comuni, due volumi nazionali contenenti uno i risultati definitivi relativi alle variabili demografiche ed uno inerente le abitazioni e le variabili socio-economiche, un volume tematico sulla popolazione straniera residente in Italia al 21 ottobre 2001, un volume concernente il sistema di rilevazione e il processo di produzione dei dati, uno contenente tutta la documentazione predisposta per il 14° Censimento generale della popolazione e delle abitazioni, dagli atti a carattere normativo ai questionari e ai modelli ausiliari perfezionati per le indagini pilota, la rilevazione censuaria e l'indagine di copertura, uno sulla qualità dei dati.

In linea con quella che è una delle caratteristiche fondamentali di un censimento, ovvero la possibilità di fornire informazioni ad un elevato livello di dettaglio territoriale, nei fascicoli dei Grandi Comuni,[7] le tavole forniscono, oltre ai principali dati a livello comunale, anche indicazioni e rappresentazioni cartografiche per aree sub-comunali di tipo amministrativo e funzionale proprie dei comuni (che, a seconda dei comuni, sono chiamate: quartieri, circoscrizioni, zone urbane, eccetera). La diffusione di dati aggregati per unità territoriali molto fini permette di soddisfare una domanda di informazione qualificata e connessa alla dinamica delle più grandi città italiane (Carbonetti *et al.*, 2008a).

Come per i passati censimenti sono stati predisposti, inoltre, due "file per sezione di censimento", uno a 279 variabili riservato agli Enti facenti parte del Sistema statistico nazionale (Sistan) e uno a 205 variabili destinato a tutte le categorie di utenti. Sono stati altresì resi disponibili "file di record individuali", ovvero file di microdati che possono essere rilasciati agli Enti appartenenti al Sistan, previa autorizzazione da parte del Presidente dell'Istat, e un "file standard" contenente una

[6] Disponibili anche on line in formato acrobat.
[7] Comuni con popolazione superiore ai 150.000 abitanti.

collezione campionaria, all'1%, di dati elementari fruibili per fini di studio e di ricerca.

Con riferimento agli spostamenti pendolari, che costituiscono una delle tematiche più importanti oggetto delle rilevazioni censuarie (Berntsen *et al.*, 2008), è stata in seguito costruita una matrice, a livello comunale, che contiene informazioni dettagliate sulla mobilità giornaliera per motivi di studio o di lavoro.

Al fine di soddisfare particolari esigenze dell'utenza, le richieste di informazioni non diffuse tramite *web* e non presenti nei volumi pubblicati, né sui supporti informatici ad essi allegati, sono state, in molti casi, evase tramite elaborazioni personalizzate a cura dell'Istituto.

3.2.2 Il piano di diffusione italiano per il 2011

Il piano di diffusione italiano, sebbene non debba essere limitato a quello prefissato dalla Commissione europea, sarà influenzato dalle norme europee in materia di definizioni, classificazioni e specifiche tecniche delle variabili obbligatorie.

Per soddisfare il Regolamento Quadro, sono stati inseriti quesiti nuovi (o variazioni a quesiti già esistenti) nei questionari di rilevazione e testati in occasione della rilevazione pilota[8] del censimento svolta nell'autunno del 2009. Tra le novità del questionario si può citare la variabile *Ever resided abroad and year of arrival in the country*,[9] un nuovo *core topic* che focalizza l'attenzione su tutte le persone che hanno risieduto almeno una volta fuori dall'attuale Paese di dimora abituale, indipendentemente dal Paese di nascita, dalla cittadinanza e da eventuali altri trasferimenti di residenza avvenuti all'interno del Paese. Permetterà, dunque, di identificare la popolazione (anche italiana) che è stata oggetto di migrazione internazionale, con una variazione rispetto al precedente censimento italiano che rilevava l'anno di trasferimento in Italia solo per i cittadini stranieri e gli apolidi, se nati all'estero. Un'altra novità riguarda, per gli aspetti relativi agli alloggi, la variabile *Type of living quarters* (Tipo di alloggio) che classifica gli alloggi con almeno una persona residente (*living quarters*) in *Abitazioni*, *Altri tipi di alloggio* e *Strutture residenziali collettive*.[10] Le indicazioni per la prossima tornata censuaria impongono, così, delle modifiche alla classificazione riguardante il tipo di alloggio utilizzata fino al Censimento del 2001, la quale non considerava la *Struttura residenziale collettiva* come tipo di alloggio possibile anche per una famiglia. Ad esempio, per il Censimento italiano del 2001 le famiglie che avevano fissato la propria dimora abituale in stanze di albergo o in appartamenti in residence, venivano censite come dimoranti in "Altro tipo di alloggio" oppure come dimoranti in abitazione. Con la nuova rilevazione del tipo di alloggio adottata per il Censimento del 2011, invece, oltre ad individuare le famiglie in Abitazione ed in Altro tipo di alloggio si dovranno tenere distinte le famiglie individuate in Struttura residenziale collettiva.

I cambiamenti nei contenuti del questionario e nella strategia di realizzazione del censimento portano inevitabilmente a una ridefinizione del piano di diffusione prossimo, non solo per la pubblicazione dei risultati *on line* ed eventualmente di editoria tradizionale, ma anche per quanto concerne i file di dati aggregati per sezioni di censimento e la fornitura di elaborazioni personalizzate. In diversi casi sussistono sostanziali differenze (concettuali, di definizione/rilevazione e di classificazione) tra ciò che viene richiesto dal Regolamento europeo e ciò che dovrà essere diffuso in ambito nazionale. Dal punto di vista contenutistico, la principale differenza riguarderà la definizione e la conseguente classificazione dei "senza tetto". Secondo le Raccomandazioni internazionali e secondo il Regolamento di attuazione relativo alle classificazioni e specifiche tecniche emanato dalla Commissione Europea, queste persone rientrano nel concetto di *household* ma sono persone che non vivono né in famiglia (*private household*) né in convivenza (*institutional household*). Al Censimento del 2001, per persone senza tetto si intendevano le persone che non dimoravano in abita-

[8] Cfr. paragrafo 4.1.

[9] Eventuale residenza all'estero e anno di arrivo nel Paese.

[10] Struttura utilizzata per la dimora di ampi gruppi di persone e/o di una o più famiglie.

zione né in altro tipo di alloggio (persone che vivono per strada, sotto i ponti, eccetera).[11] Tradizionalmente in Italia i senza tetto sono rilevati con il questionario di famiglia (distinto dal questionario di convivenza utilizzato per la rilevazione delle convivenze e delle persone in convivenza). Pertanto, a differenza di quanto previsto nelle Raccomandazioni internazionali e nel Regolamento di attuazione europeo sui *breakdowns*, nell'ultima rilevazione censuaria italiana i senza tetto sono stati considerati all'interno delle famiglie, piuttosto che come popolazione che non vive né in famiglia né in convivenza. Nel conteggio delle famiglie, a Eurostat si fornirà un numero complessivo diverso rispetto a quello per la diffusione italiana che include anche le famiglie dei senza tetto. Tutti gli incroci che avranno come oggetto le famiglie, pertanto, avranno totali, subtotali e classificazioni differenti. Analoghe problematiche si avranno in corrispondenza dei nuclei familiari, calcolati all'interno delle famiglie e quindi con valori diversi (si parte da una base complessiva diversa) tra Eurostat e la diffusione italiana, ma anche per tutti gli incroci che hanno come oggetto la popolazione in famiglia che, per la diffusione italiana, includerà anche le persone "senza tetto".

Ci sono poi altri aspetti da considerare, non di carattere definitorio ma di classificazione. In alcuni casi le classificazioni richieste da Eurostat coincidono con quelle utilizzate in Italia. Rientrano in questo caso naturalmente il sesso, ma anche l'età in anni compiuti (calcolata rispetto alla data di riferimento del censimento), il Paese di nascita e il Paese di cittadinanza. In altri casi la classificazione da predisporre per Eurostat richiede aggregazioni, trasformazioni e/o derivazioni delle modalità necessarie alla diffusione nazionale. Un primo esempio della differenza tra la classificazione europea e quella italiana riguarda il grado di istruzione. Eurostat, per permettere il confronto tra i risultati di Paesi con differenti sistemi scolastici, richiede la classificazione internazionale standard dei titoli di studio (ISCED), mentre in Italia si adotterà la classificazione solitamente utilizzata per la diffusione dei dati censuari, con i necessari ampliamenti a seguito della riforma universitaria. Un altro esempio è quello relativo ai "separati legalmente", per l'Italia rilevati e diffusi distinti dai "coniugati" e dai "divorziati", per la diffusione europea, invece, aggregati ad altre modalità dello stato civile (non tutti i Paesi prevedono infatti nella loro legislazione la separazione legale).

Nei Regolamenti di attuazione si trovano indicazioni stringenti sulle classificazioni delle variabili *core* da diffondere, alle quali i Paesi Membri devono attenersi. Viceversa, non sono forniti indirizzi riguardo alle variabili *non core* perché, come specificato nelle Raccomandazioni internazionali, è lasciata ai singoli paesi la libertà di rilevarle o meno; se rilevate di renderle comunque disponibili. Nel caso in cui l'Italia decida di fornire a Eurostat anche le informazioni relative alle variabili *non core* rilevate, bisognerà confrontare quanto richiesto nelle Raccomandazioni in termini di classificazioni con quello che viene rilevato e conseguentemente diffuso in Italia. Diverse infatti sono le variabili *non core* inserite nel modello di rilevazione italiano. Le variabili toccano vari aspetti che riguardano informazioni demografiche (data di matrimonio) ma anche informazioni sull'attività lavorativa (numero di ore effettivamente lavorate nella settimana precedente la data del censimento). Per quanto riguarda aspetti relativi alle migrazioni, si richiede l'informazione sull'acquisizione della cittadinanza italiana e l'eventuale Stato estero di cittadinanza precedente e sul Paese di nascita dei genitori. Quest'ultima informazione rappresenta una novità per l'Italia; si è ritenuto infatti di inserire questi due nuovi quesiti per rendere possibile una valutazione del processo di integrazione degli immigrati e dei loro discendenti. Nell'ambito delle caratteristiche di famiglie e nuclei familiari, le Raccomandazioni internazionali per il 2010 pongono particolare attenzione ai nuclei ricostituiti, definiti sulla base della presenza di figli da precedenti unioni di almeno uno dei due partner. Anche nell'ultima diffusione censuaria italiana è stata data rilevanza ai nuclei ricostituiti, definiti e derivati non attraverso la presenza di figli da precedenti unioni ma sulla base dello stato civile precedente dei membri della coppia, ovvero costituiti dopo lo scioglimento, per vedovanza, separazione o divorzio, di una precedente unione coniugale di almeno uno dei due partner. Anche il numero

[11] Si tratta dunque di una definizione assimilabile a quella di *primary homelessness* (o *rooflessness*).

complessivo di famiglie "estese"[12] continuerà ad essere calcolato al prossimo censimento. Altri esempi, nella sezione relativa alle notizie sugli alloggi, sono: la variabile *Air-conditioning* per rilevare la presenza di impianti di aria condizionata; il quesito su *Number of cars available for the use of the household* per la disponibilità di automobili da parte della famiglia censita; la variabile *Availability of car parking* sul numero di posti auto di cui dispone la famiglia.

Inoltre, benché in alcuni casi non derivino direttamente dalle variazioni introdotte a livello europeo, è necessario, per completare il quadro dei quesiti contenuti nel questionario, definire alcune novità introdotte nel modello di rilevazione che avranno influenza nella diffusione dei risultati del Censimento del 2011. Ci saranno, ad esempio, nuove tavole con incroci che avranno come oggetto la popolazione che ha risieduto all'estero (risultati che permetteranno di individuare anche l'immigrazione di ritorno). La diffusione del 2011 risentirà anche delle variazioni introdotte nelle modalità di alcune domande di rilevazione. Un esempio è relativo al quesito sull'anno di arrivo in Italia per i nati all'estero e gli apolidi. Tale quesito, non più presente, è stato inglobato in quello sull'anno di trasferimento in Italia per chi ha risieduto all'estero. Naturalmente a questa domanda rispondono anche coloro che, nati all'estero, arrivano in Italia per la prima volta. Sarà quindi possibile continuare a diffondere l'informazione sull'anno di arrivo in Italia. Nel 2001 le domande sul tipo di impianto di riscaldamento e combustibile o energia utilizzata per il riscaldamento dell'abitazione non permettevano di individuare in modo univoco quale combustibile alimentava quale impianto. L'introduzione della domanda sotto forma di matrice, testata nella rilevazione pilota, consente di individuare, per ogni impianto di riscaldamento presente in un'abitazione, il combustibile o energia che lo alimenta e di ottenere un'informazione più puntuale rispetto al passato censimento.

Tra i quesiti inclusi nel questionario ve ne sono poi alcuni che non riguardano *core topics* o *non core topics* ma sono legati esclusivamente alla diffusione italiana. Tali variabili rivestono una particolare importanza per l'Italia e non sono citate nelle Raccomandazioni UNECE né nel Regolamento Quadro dell'Unione Europea. Ad esempio, lo "Stato civile prima dell'ultimo matrimonio" utilizzato per il calcolo delle famiglie ricostituite; mentre, per la sezione sugli alloggi, la variabile "Disponibilità di un impianto a energia rinnovabile per la produzione di energia elettrica", che delinea un particolare interesse verso l'adozione e la diffusione di nuove tecnologie per la produzione di energia elettrica sul territorio italiano. Infine, i risultati dell'analisi delle richieste e dell'utilizzo di dati censuari (Berntsen *et al.*, 2008) hanno costituito un prezioso supporto sia alla definizione dei contenuti informativi del Censimento del 2011, sia alla progettazione del piano di diffusione dei dati. In questa occasione, con l'obiettivo di snellire il questionario di rilevazione, sono stati eliminati quesiti che l'analisi suddetta ha dimostrato essere meno richiesti ed utilizzati dagli utenti.

3.2.3 Le implicazioni delle innovazioni di metodo

Il piano di diffusione italiano sarà influenzato anche dalle innovazioni di tipo metodologico e tecnologico introdotte dall'Istat per la conduzione del 15° Censimento generale della popolazione e delle abitazioni e in particolare dal più ampio utilizzo di fonti amministrative e dall'introduzione delle tecniche campionarie. Tali soluzioni, come introdotto nel capitolo 2, sono volte a ridurre il numero di rilevatori impiegati sul territorio (*front-office*), orientando maggiormente le risorse degli Uffici Comunali di Censimento (UCC) su attività di coordinamento e controllo (*back-office*). L'impiego delle Liste Anagrafiche Comunali (LAC) che guidano la rilevazione hanno permesso, per la prima volta, la distribuzione dei questionari attraverso un vettore postale. Per i rispondenti ci sono state possibilità differenziate di restituzione: spedizione per posta, compilazione via web, riconsegna presso centri di raccolta sul territorio.[13] La multicanalità dovrebbe favorire la restituzione spontanea dei questionari da parte delle famiglie, con un risparmio nel lavoro sul campo dei rileva-

[12] La tipologia familiare all'interno della quale si individuavano almeno due nuclei (coppia o nuclei monogenitore) oppure un solo nucleo con altre persone residenti.

[13] Come ultima possibilità i rispondenti, nella fase di "recupero delle mancate risposte", hanno potuto consegnare il questionario al rilevatore che è passato presso le famiglie che non hanno risposto spontaneamente.

tori per il recupero delle mancate risposte. A riguardo, i risultati della rilevazione pilota del 2009 hanno confermato questa ipotesi.[14]

È stato previsto anche l'impiego di nuovi strumenti territoriali volti a migliorare il riferimento geografico delle unità di rilevazione. Nei comuni di maggiore ampiezza demografica, sono state definite nuove unità territoriali sub-comunali, le *aree di censimento di centro abitato* (insiemi di sezioni di censimento di tipo "centro" contigue, della dimensione di circa 15mila abitanti) che, utilizzate come unità territoriali di minimo riferimento per la rilevazione campionaria, garantiranno la produzione di dati rappresentativi a livello sub-comunale (Astorri *et al.*, 2007; Bianchi *et al.*, 2010).

Con la nuova strategia censuaria, tutte le variabili rilevate con il questionario in forma ridotta (contenute anche nella versione completa) potranno essere diffuse a livello di sezione di censimento. Per le altre variabili, contenute solo nel modello *long* e soggette a stima campionaria, saranno diffusi dati per aree di censimento e non per sezioni di censimento. La scelta di inserire nel questionario ridotto poche informazioni socio-economiche (grado di istruzione, condizione professionale o non professionale e spostamenti quotidiani per motivi di studio o di lavoro)[15] permetterà di disporre di dati a livello di sezione di censimento, seppur con un minor livello di dettaglio classificatorio rispetto al questionario *long*.

L'introduzione delle due differenti versioni di questionario e la stima per aree sub-comunali delle variabili socio-economiche rilevate nel questionario *long* porterà, quindi, ad una variazione nella diffusione di alcuni prodotti specifici del censimento (file per sezioni, file di microdati). Rispetto al 2001 rimane invariata l'offerta informativa a livello comunale. Il nuovo disegno censuario determinerà anche, e prevalentemente, differenze nella fornitura dei dati agli utenti esterni attraverso le richieste di elaborazioni personalizzate. Più che in passato (accadeva già per il Censimento del 2001) sarà necessario valutare caso per caso le richieste, anche in base all'errore campionario delle frequenze assolute riferite a tavole con un elevato numero di incroci.

Ciò che accomuna le due diffusioni (italiana ed europea) è l'attenzione verso il soddisfacimento delle esigenze degli utenti finali dei dati censuari. A livello internazionale, la diffusione dei risultati censuari consentirà all'utente un facile accesso ad una vasta gamma di incroci comparabili tra tutti gli Stati Membri (mediante *SDMX - Census Hub*). A livello nazionale sarà previsto, come già nel 2001, un sistema di diffusione *on line* che permetterà all'utente di navigare tra tavole non rigidamente strutturate ma scegliendo le classificazioni e i dettagli territoriali secondo i propri interessi.

4. I contenuti informativi dei questionari di censimento

4.1 La progettazione

Nella progettazione dei contenuti informativi dei questionari di rilevazione censuaria è stato necessario garantire la conformità con i Regolamenti dell'Unione Europea ed il rispetto delle definizioni e delle classificazioni imposte dalla normativa nazionale e internazionale; si è inoltre tenuto conto della necessità di garantire la continuità con le passate rilevazioni censuarie e di soddisfare specifiche esigenze informative degli utilizzatori italiani.

Come descritto nel precedente capitolo, la nuova strategia di rilevazione, che prevede l'introduzione di tecniche di campionamento per l'acquisizione di una parte delle informazioni censuarie, ha richiesto la definizione di due versioni di questionario: una versione in forma ridotta e

[14] Una prima analisi dei dati sulla restituzione spontanea, ovvero relativa ai questionari restituiti non al rilevatore e con esito postale positivo, ha portato ad un tasso di restituzione spontanea pari al 50,0%. Complessivamente, il 9,1% delle famiglie ha compilato il questionario via web, il 40,8% ha optato per la restituzione postale, il 12,6% ha portato il questionario compilato al Centro Comunale di Raccolta il 37,5% ha consegnato il questionario direttamente al rilevatore.

[15] Il numero di variabili contenute nello *short form* per ciascuno degli argomenti è ridotto. In particolare, si tratta di 2 quesiti sulla acquisizione di cittadinanza, 4 quesiti che riguardano il titolo di studio più elevato conseguito (con un minor numero di modalità rispetto alla versione *long*), i corsi di formazione professionale regionale e i titoli di studio post-laurea, 4 quesiti sulla condizione professionale che consentono di quantificare le forze di lavoro e le non forze di lavoro, e 4 quesiti relativi agli spostamenti quotidiani al luogo di lavoro o di studio.

una in forma completa. La prima include informazioni che verranno rilevate in maniera esaustiva sull'intera popolazione italiana e comprendono tutte le variabili demografiche necessarie per la produzione delle tavole statistiche (*hypercubes*) che dovranno essere rese disponibili a Eurostat a livello comunale. La versione completa (*long form*) contiene, oltre ai quesiti della forma ridotta, tutte le altre variabili previste nel piano di rilevazione, che vengono dunque riservate a un campione di famiglie residenti. Questa strategia si pone l'obiettivo di ridurre il carico statistico su una parte di rispondenti, diminuire il numero di pagine da stampare e acquisire, e limitare l'attività di revisione qualitativa dei modelli da parte dei comuni; consente, allo stesso tempo, di mantenere inalterato il contenuto informativo della rilevazione censuaria e rispettare i vincoli internazionali in merito alle variabili da diffondere.

Al fine di mettere a punto l'organizzazione, i metodi e le tecniche da adottare in occasione dell'appuntamento del 2011, si è svolta ad ottobre 2009 la rilevazione pilota del 15° Censimento della popolazione e delle abitazioni con l'obiettivo di testare molte delle innovazioni previste. In particolare, dal punto di vista dei contenuti informativi, ha costituito l'occasione per sperimentare differenti versioni del questionario di rilevazione e, dunque, per verificare quale fosse la soluzione migliore da adottare durante la rilevazione censuaria vera e propria. Sono state sperimentate due forme ridotte di questionario, le cosiddette *short form* (13 domande) e *medium form* (30 domande), utilizzate alternativamente e in abbinamento con la *long form* (71 domande).

4.2 Le versioni short, medium e long testate con la rilevazione pilota

Tutte le versioni del questionario sono composte da: una "Lista A" contenente l'elenco delle persone abitualmente dimoranti nell'alloggio (persone della famiglia); una "Lista B" contenente l'elenco delle persone non abitualmente dimoranti nell'alloggio ma temporaneamente o occasionalmente presenti nell'alloggio; una "Sezione I" con le notizie su famiglia e alloggio; una "Sezione II" con le notizie sulle persone che hanno dimora abituale nell'alloggio (notizie sui singoli componenti della famiglia).

Le Liste A e B includono le stesse informazioni in tutti i tipi di questionario.

La versione *short* del modello di rilevazione include il minimo numero di quesiti. In particolare si tratta, per le notizie su famiglia e alloggio, di informazioni sul tipo di alloggio, lo stato di occupazione, le famiglie coabitanti e la superficie dell'abitazione; mentre per le notizie sulle persone che hanno dimora abituale nell'alloggio si tratta solamente di caratteristiche demografiche (relazione di parentela, sesso, età, luogo di nascita, stato civile, data di matrimonio, stato civile prima dell'ultimo matrimonio, cittadinanza, dimora un anno prima del censimento).

L'offerta informativa del modello di rilevazione di tipo *long* è notevolmente più ampia, riguardando, oltre ai quesiti sopra citati, tutte le altre variabili di natura socio-economica tradizionalmente acquisite in occasione del censimento, che sono relative all'istruzione e alla formazione, alla condizione professionale, all'attività lavorativa ed al luogo di studio o di lavoro. Per le notizie su famiglia e alloggio si aggiungono quesiti su: l'acqua e l'impianto igienico-sanitari, l'impianto di climatizzazione, l'auto e il posto auto, il telefono e la connessione a internet.

I quesiti presenti esclusivamente nella versione *long* del modello rispondono ad esigenze differenti: in alcuni casi rappresentano *core topics* imposti da Eurostat (ad esempio le variabili sull'eventuale residenza all'estero e l'anno di arrivo nel Paese), in altri corrispondono a *non core topics* ritenuti particolarmente interessanti in ambito italiano (è il caso, ad esempio, del Paese di nascita dei genitori che costituisce una novità per l'Italia); vi sono poi variabili che non vengono segnalate in ambito europeo ma rivestono esclusivamente interesse nazionale (ad esempio, tra quelle relative al lavoro si può citare la frequenza di corsi di formazione/aggiornamento professionale, il tipo di rapporto di lavoro e la tipologia dei contratti di lavoro a tempo determinato). Nel *long form* sono inoltre inserite, al fine di ampliare l'offerta dei dati sui flussi pendolari, le informazioni sul luogo di studio (degli studenti) e di lavoro (degli occupati), sul mezzo di trasporto, sulla distanza percorsa e sul tempo impiegato per recarsi al luogo di lavoro o di studio.

Secondo la strategia di tipo *short/long form*, la rilevazione esaustiva sarebbe limitata alle caratteristiche demografiche degli individui, mentre le informazioni di natura socio-economica rimar-

rebbero tutte oggetto di stima campionaria. La proposta *medium/long form* differisce dalla precedente in quanto la versione *medium* è un questionario caratterizzato, oltre che dalle variabili strettamente demografiche, anche da poche informazioni di carattere socio-economico (non presenti nella versione *short*), rimandando, anche in questo caso, al questionario *long* per una maggiore ricchezza informativa.

Il questionario di tipo *medium* contiene, oltre alle variabili della *short form*, anche un insieme di quesiti relativi alla cittadinanza acquisita, al grado di istruzione, alle forze di lavoro e agli spostamenti pendolari. In particolare, sono stati aggiunti: 2 quesiti relativi all'acquisizione della cittadinanza; 4 domande che riguardano il titolo di studio (con un minor numero di modalità di classificazione rispetto al questionario *long*), i corsi di formazione professionale regionale e i titoli di studio post-laurea; 4 quesiti sulla condizione professionale che consentono di quantificare le forze di lavoro e le non forze di lavoro; 4 quesiti relativi agli spostamenti quotidiani e al luogo di lavoro. I contenuti informativi dei questionari *short*, *medium* e *long* sono riassunti, per sezioni e sotto-sezioni nella tavola 2.

Tavola 2 - Distribuzione dei quesiti per sezioni nelle tre versioni di questionario: short, medium, long

SEZIONI	Sotto-sezioni	Numero di quesiti rilevati esaustivamente con lo SHORT	Numero di quesiti rilevati Esaustivamente con il MEDIUM	Numero di quesiti rilevati a campione con il LONG
Sezione I – Notizie su famiglia e alloggio				
	Tipo di alloggio e famiglia	3	4	4
	Proprietà e struttura dell'abitazione	1	3	5
	Acqua e impianto igienico-sanitari			6
	Impianto di climatizzazione			4
	Auto e posto auto			2
	Telefono e connessione ad internet			2
Sezione II – Fogli individuali				
	Notizie anagrafiche	4	4	4
	Stato civile e matrimonio	3	3	3
	Cittadinanza	1	3	5
	Dimora precedente	1	1	4
	Istruzione e formazione		4	12
	Condizione professionale		4	5
	Attività lavorativa			8
	Luogo di studio o di lavoro		4	7
Totale		**13**	**30**	**71**

4.3 I questionari per il Censimento del 2011

Sulla base dei risultati della rilevazione pilota si è scelto di adottare per il Censimento del 2011, oltre alla forma completa del questionario, la versione *medium* come forma ridotta abbandonando la *short form* e optando, dunque, per una maggiore ricchezza informativa. Tale scelta è stata avvalorata dal fatto che la lunghezza del questionario non ha influenzato le famiglie nella propensione a rispondere: per i tre modelli *short*, *medium* e *long*, infatti, il tasso di risposta è stato analogo.

In questo modo sarà possibile disporre di un maggior numero di informazioni a livello di sezione di censimento; il vantaggio è dunque quello di poter diffondere, oltre ai dati delle variabili strettamente demografiche, anche alcune informazioni di carattere socio-economico per sezione di censimento, seppur con un minimo livello di dettaglio classificatorio. Inoltre, l'approccio *medium/long form* porterà a disporre di maggiore informazione benchmark (rilevata esaustivamente sulla popolazione) da utilizzare in fase di stima delle variabili inserite solo nei modelli *long form* e rilevate su campioni di famiglie.

La versione *long* del modello di rilevazione censuaria presenta alcune differenze rispetto ai contenuti dei questionari di rilevazione testati con l'indagine pilota: sono stati inseriti alcuni quesiti

aggiuntivi che riguardano, in particolare, l'iscrizione nell'Anagrafe comunale, la presenza alla data del censimento, la dimora abituale cinque anni prima del censimento. È stata inoltre aggiunta, rispetto alla rilevazione pilota, l'intera sotto-sezione 8 "Difficoltà nelle attività della vita quotidiana" comprensiva di 4 quesiti (senza obbligo di risposta) volti a rilevare le eventuali difficoltà incontrate nello svolgere alcune attività a causa di problemi di salute; riguardano, in particolare, difficoltà nel vedere, sentire, camminare e ricordare o concentrarsi.

Nella versione ridotta definitiva del modello di rilevazione censuaria, rispetto alla *medium form* adottata per la rilevazione pilota, sono stati aggiunti i quesiti sull'iscrizione nell'Anagrafe comunale, la presenza alla data del censimento, la dimora abituale cinque anni prima del censimento e la condizione non professionale.

5. La strategia campionaria tramite questionari short e long

5.1 La strategia di campionamento

La strategia prevede la somministrazione del questionario in forma completa (*long form*) nei comuni sopra i 20mila abitanti e in tutti i comuni capoluogo di provincia, solo a campioni rappresentativi di famiglie residenti presenti nella relativa lista anagrafica comunale; la versione in forma ridotta (*short form*) a tutte le famiglie presenti in lista e non estratte per il campione. Nei comuni più piccoli, non eleggibili per il campionamento, invece, la decisione è quella di sottoporre il questionario completo a tutte le famiglie residenti. In base a tale approccio, i dati relativi alle domande contenute in entrambi i questionari deriveranno da un conteggio esaustivo, mentre le informazioni osservabili solo sui campioni e il loro incrocio con le variabili esaustive saranno desunte da stime campionarie. Dal punto di vista tecnico-metodologico, questa strategia comporterà una riduzione della mole dei dati da acquisire ed elaborare così da permettere maggiori controlli a vantaggio di una diminuzione dell'errore di misura (Cocchi, 2007).

La scelta di introdurre il campionamento nel censimento italiano è avvalorata anche dall'analisi delle esperienze estere (Abbatini *et al.*, 2007) dalla quale emergono realtà di Paesi (Canada, Usa, Francia, Germania, Israele, Olanda) in cui, adottando approcci non convenzionali per il censimento, si producono stime per le variabili non strettamente demografiche.

5.2 Valutazioni sperimentali per la definizione della strategia

In generale, l'adozione di una strategia campionaria richiede decisioni metodologiche connesse al disegno di campionamento, ai domini per i quali produrre le stime, alle variabili oggetto di rilevazione campionaria, ai parametri da stimare e allo stimatore da utilizzare.

Da un preliminare studio delle soluzioni metodologiche (Cicchitelli *et al.*, 1992; Särndal *et al.*, 1992) praticabili per il contesto censuario in Italia, sono stati considerati disegni di campionamento da lista (per la possibilità di utilizzare i registri anagrafici comunali) o areali (per l'opportunità di riferirsi alla lista delle sezioni di censimento delle Basi Territoriali). Inoltre è stato definito un insieme di possibili stimatori tra cui individuare quello che potrebbe offrire le migliori garanzie in termini di distorsione e di variabilità campionaria.

L'impianto di base è stato quello di uno schema semplice di selezione del campione di famiglie, secondo modalità tali da comportare un basso impatto sulle operazioni sul campo e, casomai, diversificare la scelta dello stimatore, anche in favore di metodi indiretti, con il duplice obiettivo di ottenere stime con elevati livelli di accuratezza e garantire coerenza tra dati esaustivi e valori stimati.

Per gli scopi sopraindicati, sono state condotte delle sperimentazioni[16] al fine di individuare, tra le possibili soluzioni metodologiche, quelle più facilmente praticabili nel contesto censuario da un

[16] Queste hanno riguardato l'estrazione di 1.000 campioni di famiglie per la simulazione dello spazio campionario e la produzione di tavole di frequenze relative e assolute riferite ad aree di censimento.

punto di vista organizzativo e più rispondenti alle esigenze di precisione e qualità. A riguardo, sono state effettuate simulazioni su dati relativi al Censimento della popolazione e delle abitazioni del 2001 (Borrelli *et al.*, 2007; Carbonetti e De Vitiis, 2007; Carbonetti e Fortini, 2008b) al fine di misurare l'accuratezza attesa di stime inerenti le frequenze relative e assolute per le modalità delle variabili di *long form*, prese singolarmente e/o incrociate con le modalità delle variabili esaustive.

Le sperimentazioni hanno interessato dati relativi a circa il 10% delle famiglie residenti nel 2001, appartenenti a 498 aree di censimento di centro abitato appositamente disegnate (con popolazione compresa tra 5mila e 15mila unità) su 40 comuni scelti per diversa ampiezza demografica in differenti regioni italiane (Borrelli *et al.*, 2011a). In generale, riguardo ai possibili effetti sulla produzione del dato, dai risultati delle sperimentazioni si è osservato che:

- le stime comportano un errore che, espresso in termini percentuali, diminuisce al crescere della frequenza assoluta della variabile (singola o di incrocio) cui fa riferimento;
- errori più grandi sono attesi per la stima delle frequenze assolute più piccole; a tal riguardo, sono stati valutati alcuni metodi indiretti di stima (Borrelli *et al.*, 2008; Borrelli *et al.*, 2012) proponibili per aumentare la precisione e quindi l'affidabilità dei dati finali.

I risultati sperimentali hanno portato a scegliere una strategia campionaria basata sull'adozione del disegno di campionamento casuale semplice di famiglie da lista anagrafica e sull'uso degli stimatori di ponderazione vincolata (Deville e Särndal, 1992) che garantiscono una migliore rappresentatività dell'informazione osservata sulle unità del campione estratto. Inoltre, come anticipato nel paragrafo 3.2.3, le stime di massimo dettaglio territoriale saranno riferite a domini sub-comunali coincidenti con le *Aree di Censimento di centro abitato* (Astorri *et al.*, 2007; Bianchi *et al.*, 2010) definite dall'Istat, per i comuni interessati dal campionamento, tramite l'aggregazione delle sezioni di censimento di tipo "centro" con il vincolo della contiguità. Le aree di censimento sono disegnate in modo tale da avere una dimensione media di circa 15mila unità e da rispettare i limiti geografici[17] delle suddivisioni (meno fini) predefinite dai comuni a scopi amministrativi o funzionali.

5.3 Accuratezza attesa delle stime di frequenze assolute riferite a domini interamente sottoposti a campionamento

La tavola 3 contiene i valori attesi del coefficiente di variazione percentuale[18] (cv) per la stima di frequenze assolute riferite a domini interamente sottoposti a campionamento. Tali valori sono stati determinati dai risultati delle sperimentazioni che hanno considerato una strategia campionaria basata sul disegno casuale semplice, per differenti frazioni sondate e l'impiego dello stimatore di ponderazione vincolata. Si fa presente che, non essendo stato possibile derivare la funzione esatta che esprime il valore atteso del cv per valori puntuali delle quantità oggetto di stima, si è proceduto a sintetizzare i risultati tramite una distribuzione di valori mediani, per classi di frequenza assoluta.

[17] Il disegno delle aree di censimento prevede il rispetto dei limiti delle suddivisioni amministrative dei comuni con almeno 250mila abitanti e delle unità territoriali non più piccole di 30mila.

[18] Con riferimento alla stima $\hat{\theta}$ di un generico parametro θ il *coefficiente di variazione percentuale* è dato dal valore percentuale del rapporto tra lo scarto quadratico medio campionario e il valore atteso:

$$cv(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{E(\hat{\theta})} 100 \quad .$$

Nel caso della stima della generica frequenza assoluta T, in base al valore di cv si calcola l'errore assoluto $\Delta_T = 1{,}96\, T \cdot cv/100$ che permette di determinare l'intervallo di confidenza $\{(\hat{T}-\Delta_T);(\hat{T}+\Delta_T)\}$ che conterrà il vero valore (incognito) della frequenza T con probabilità pari a 0,95 .

Esempio: per la stima della frequenza assoluta T=600, nel caso di un disegno semplice con frazione sondata del 33%, il cv atteso è pari al 5,4% (cfr. Tavola 3); ne consegue un errore assoluto medio $\Delta_T = 1{,}96 \times 600 \times 5{,}4/100 \cong 64$. Quindi, per il 95% dei campioni estraibili secondo tale schema di campionamento la stima sarà compresa tra 536 e 664.

Tavola 3 - Distribuzione del coefficiente di variazione atteso (valore mediano) **per classi di frequenza assoluta relative a domini interamente sottoposti a campionamento** (disegno casuale semplice; frazione di campionamento del 10%, del 20% e del 33%)

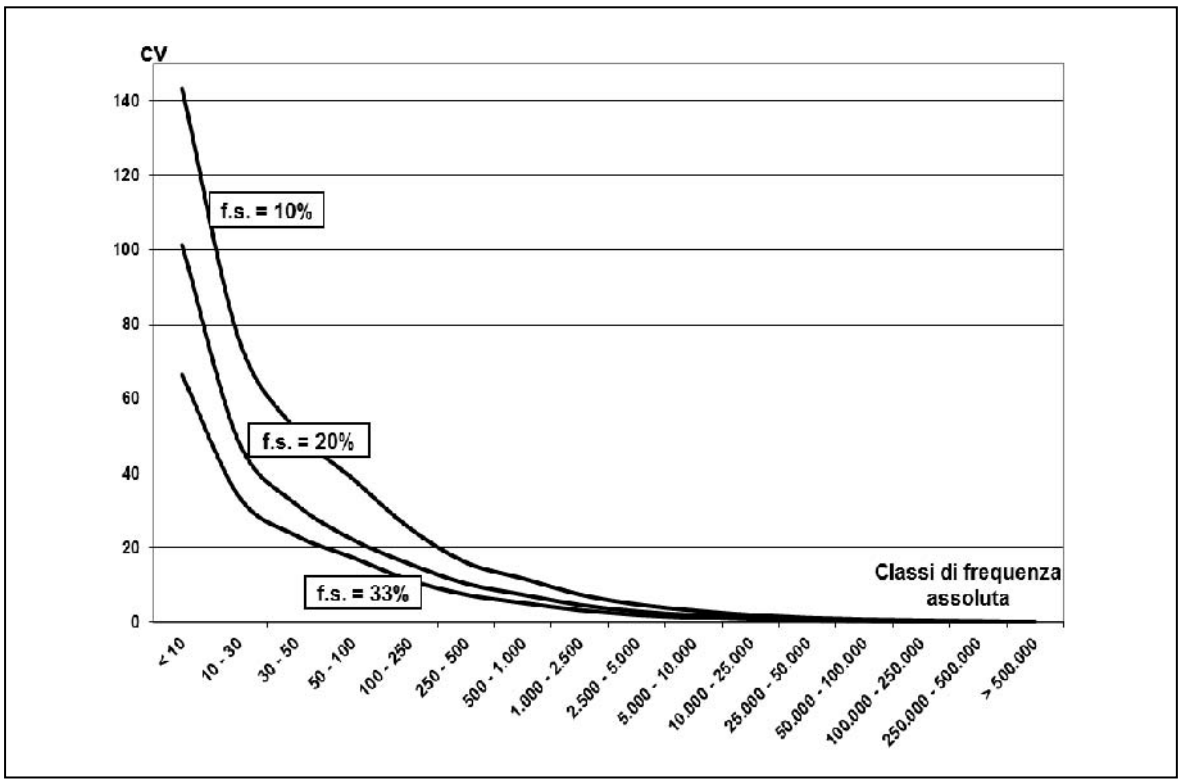
CLASSI DI FREQUENZA ASSOLUTA T	Frazione sondata = 10%	Frazione sondata = 20%	Frazione sondata = 33%
	cv atteso	cv atteso	cv atteso
< 10	143.3	101,4	66,5
10 ├ 30	75.9	48,4	33,8
30 ├ 50	51.8	31,8	23,4
50 ├ 100	38.6	22,3	17,4
100 ├ 250	25.4	15,7	11,4
250 ├ 500	16.1	10,4	7,5
500 ├ 1.000	11.8	7,5	5,4
1.000 ├ 2.500	7.5	4,7	3,3
2.500 ├ 5.000	4.9	3,0	2,0
5.000 ├ 10.000	3.2	2,0	1,3
10.000 ├ 25.000	2,0	1,4	0,9
25.000 ├ 50.000	1,2	0,8	0,6
50.000 ├ 100.000	0,8	0,5	0,4
100.000 ├ 250.000	0,6	0,3	0,3
250.000 ├ 500.000	0,4	0,2	0,2
≥ 500.000	0,2	0,1	0,1

Si tiene a precisare che i risultati della tavola 3 sono relativi a stime di frequenze riferite ad aree di censimento campionabili (con almeno 5mila abitanti) o a domini composti dalla aggregazione di aree di censimento tutte eleggibili per il campionamento. Per le valutazioni dell'efficienza di stime relative ad una singola area di censimento bisogna riferirsi solo alle classi di valori non superiori a 10mila unità; i valori superiori a 10mila sono, invece, specifici per domini dati dall'aggregazione di più aree.

La Figura 1 descrive l'andamento delle curve degli errori campionari risultanti dalle sperimentazioni per la stima di frequenze assolute con un disegno di campionamento casuale semplice di famiglie da lista e relative a tre differenti frazioni di campionamento sperimentate (10%, 20% e 33%). L'andamento delle curve, basate solo su riscontro empirico, mostra per tutte un andamento monotono decrescente.

Gli errori campionari più bassi sono attesi nel caso della frazione di campionamento pari al 33% per la presenza di un campione con una numerosità molto elevata. Un risultato che emerge dall'analisi dei dati riportati nella tavola 3 è che raddoppiando la dimensione del campione, dalla frazione sondata del 10% a quella del 20%, si ottiene una riduzione dell'errore campionario nell'ordine del 33-38%; incrementando invece il campione di più di tre volte, dalla frazione del 10% alla frazione del 33%, si ottiene un guadagno nell'ordine del 53-58%. Queste indicazioni sono state di grande utilità nella scelta della frazione di campionamento, anche congiuntamente a valutazioni effettuate a più ampio raggio, inclusa quella del costo dell'intera operazione censuaria.

Figura 1 - Curve degli errori di campionamento attesi (misurate dal cv) **nel caso del disegno casuale semplice** (per le frazioni sondate del 10%, del 20% e del 33%)



5.4 Accuratezza attesa delle stime di frequenze assolute riferite a domini parzialmente sottoposti a campionamento

In generale, i domini di diffusione dei risultati censuari, a partire da quello comunale, non saranno quasi mai interamente sottoposti a campionamento; infatti, questi possono includere sottodomini non campionabili (aree di censimento sotto la soglia di campionabilità; zone periferiche; spazi rurali) oppure, nel caso di realtà territoriali sovra-comunali, comprendono comuni non eleggibili alla strategia campionaria (comuni con dimensione inferiore a 20mila abitanti e non capoluogo di provincia). Quindi, è solo su una parte di tali domini che si procederà con la rilevazione campionaria, somministrando il questionario *long* a campioni di famiglie e il questionario *short* alle famiglie eleggibili non selezionate; invece, alle famiglie residenti nel resto del dominio, non coinvolte dalla strategia campionaria, verrà distribuito il questionario in forma completa.

In forza di ciò, il dato finale relativo alle variabili rilevate solo tramite il *long form* sarà pari alla somma di una componente stimata (riferita alla parte del dominio che è campionato) e di una componente conteggiata in modo esaustivo (derivante dall'osservazione effettuata sulla parte del dominio non coinvolta dal campionamento), per cui questo continuerà ad essere il risultato di una stima anche se con un errore campionario ridotto. Ne consegue che, le stime relative a contesti territoriali superiori all'area di censimento di centro abitato (comune, provincia, regione, …) beneficeranno di riduzioni dell'errore campionario in misura del fatto che una parte del territorio non è sottoposta a campionamento (ma si procede con la rilevazione esaustiva di tutte le variabili).

Nel seguito si illustra la misura della riduzione dell'errore di campionamento della stima campionaria di una frequenza assoluta riferita ad un dominio qualsiasi e dovuta alla presenza, nel dominio stesso, di una parte non rilevata a campione.

Si consideri dapprima il caso in cui, nel generico dominio R, di dimensione N, si impiega la strategia campionaria tramite questionari *short* e *long* su tutte le unità. Supposto di aver estratto un campione $\underline{s}$ dal dominio R (in base ad un prefissato disegno di campionamento), si proceda

all'osservazione di una qualunque variabile x (rilevabile solo con il questionario completo) per ciascuna unità campionaria di $\underline{s}$. L'obiettivo della rilevazione è determinare, per la variabile in questione, una stima della frequenza assoluta (incognita) riferita all'intero dominio

$$T_x = \sum_{i \in R} x_i = N\, p_x \tag{1}$$

dove p_x è la corrispondente frequenza relativa.

Dai risultati delle osservazioni rilevate sul campione $\underline{s}$, una stima dell'ammontare (1) è data dalla seguente quantità

$$\hat{T}_x = N\, \hat{p}_x \tag{2}$$

espressione in cui è utilizzata una stima ("corretta" o "asintoticamente corretta") della frequenza relativa p_x definita da

$$\hat{p}_x = \frac{1}{N} \sum_{i \in \underline{s}} x_i w_i \tag{3}$$

in cui w_i rappresenta il peso di riporto all'universo ("da disegno" o "da calibrazione") associato alla generica unità del campione $\underline{s}$.

La stima (2) di T_x è una variabile casuale campionaria che, tenendo conto della (3), ha media e varianza rispettivamente pari a:

$$E\left(\hat{T}_x\right) \cong N\, p_x \tag{4}$$

$$Var\left(\hat{T}_x\right) = N^2\, Var(\hat{p}_x) \tag{5}$$

In base alla (4) e alla (5) si può facilmente derivare il coefficiente di variazione (cv) che misura l'errore di campionamento che si commette con la stima (2). Tale misura, data dal rapporto tra lo scarto quadratico medio e il valore atteso della stima campionaria è pari alla seguente espressione:

$$cv\left(\hat{T}_x\right) = \frac{\sqrt{Var\left(\hat{T}_x\right)}}{E\left(\hat{T}_x\right)} 100 \cong \frac{\sigma\left(\hat{p}_x\right)}{p_x} 100 = cv\left(\hat{p}_x\right) \tag{6}$$

La (6) evidenzia che il valore del cv della stima di una frequenza assoluta, riferita ad un dominio qualsiasi, è equivalente a quello della stima della frequenza relativa corrispondente.

Si ipotizzi ora la situazione in cui nello stesso dominio R solo una parte, indicata con R_c (di

ampiezza $N_c < N$) sia sottoposta a campionamento, mentre nella parte residua R_{nc} (di ampiezza $N_{nc} = N - N_c$) si proceda somministrando il questionario completo a tutte le famiglie residenti $\left(R = R_c \cup R_{nc} \;\; ; \;\; R_c \cap R_{nc} = \varnothing\right)$. In questo caso, poiché il campione (denotato $\underline{s}_c$) verrebbe estratto solo dal dominio R_c, si procederà all'osservazione della stessa generica variabile x sia su ciascuna unità campionaria di $\underline{s}_c$, sia su tutte le unità del dominio non campionato R_{nc}.

In base ai dati osservati è possibile determinare una stima riferita all'intero dominio R, del medesimo ammontare (1), come somma del valore stimato sulla parte campionata e della frequenza assoluta determinata sulla parte residua in modo esaustivo:

$$\hat{T}_x^* = \hat{T}_{x,c} + T_{x,nc} = N_c \, \hat{p}_{x,c} + N_{nc} \, p_{x,nc} \tag{7}$$

dove la frequenza relativa $p_{x,c}$ riferita al dominio R_c è stimata, in modo analogo alla (3), da

$$\hat{p}_{x,c} = \frac{1}{N_c} \sum_{i \in \underline{s}_c} x_i w_i^* \tag{8}$$

in cui w_i^* continua ad esprimere il peso di riporto all'universo. Inoltre, tornando alla (7), la frequenza relativa $p_{x,nc}$ riferita al dominio R_{nc}, non sottoposto a campionamento, è calcolata in maniera esatta.

La stima (7) di T_x è anch'essa una variabile casuale campionaria che, in base alla (8), assume media e varianza rispettivamente pari a:

$$E\left(\hat{T}_x^*\right) \cong N_c \, p_{x,c} + N_{nc} \, p_{x,nc} \tag{9}$$

$$Var\left(\hat{T}_x^*\right) = N_c^2 \, Var(\hat{p}_{x,c}) \tag{10}$$

In questa situazione, tenendo conto della (9) e della (10), il coefficiente di variazione associato alla stima (7) è pari dalla seguente espressione:

$$cv\left(\hat{T}_x^*\right) \cong \frac{N_c \, \sigma\left(\hat{p}_{x,c}\right)}{N_c \, p_{x,c} + N_{nc} \, p_{x,nc}} 100 \tag{11}$$

Ora, ritenendo plausibile l'omogeneità di comportamento del fenomeno associato alla variabile x nel dominio R_c sottoposto a campionamento e nel dominio R_{nc} rilevato in modo esaustivo ($p_{x,c} \approx p_x \approx p_{x,nc}$) e nell'ipotesi di una variabilità campionaria pressoché simile delle stime $\hat{p}_x$ e

$\hat{p}_{x,c}$, per l'adozione di uno stesso disegno di campionamento, la (11) diventa:

$$cv\left(\hat{T}_x^*\right) \cong \frac{N_c\ \sigma\left(\hat{p}_x\right)}{N\ p_x} 100 = \gamma\ cv\left(\hat{p}_x\right) \tag{12}$$

dove γ indica la quota di popolazione di R eleggibile al campionamento ($\gamma = N_c / N$).

Infine, tenendo conto di quanto emerso dalla (6), si giunge alla seguente relazione

$$cv\left(\hat{T}_x^*\right) \cong \gamma\ cv\left(\hat{T}_x\right) \tag{13}$$

dalla quale si conclude che l'errore di campionamento per la stima di una frequenza assoluta su un domino R non interamente a campione è, all'incirca, una quota γ dell'errore atteso per la stima della stessa quantità riferita ad un dominio della stessa dimensione di R, dove però tutte le unità sono eleggibili per il campionamento. La riduzione dell'errore è legata, tramite il parametro γ, proprio alla dimensione della parte del dominio R interessata dal campionamento. Si precisa che la relazione (13) è stata successivamente verificata sui risultati delle sperimentazioni, per stime riferite sia a domini interamente campionati che a domini parzialmente campionati.

In base alla (13), la riduzione percentuale attesa dell'errore campionario è così espressa:

$$rid_{cv}\% = \frac{cv(\hat{T}_x) - cv(\hat{T}_x^*)}{cv(\hat{T}_x)} \cdot 100 \cong (1-\gamma) \cdot 100 \tag{14}$$

Riassumendo, l'errore di campionamento misurato dal cv della stima di una frequenza assoluta su un generico dominio ha una riduzione in ragione del parametro γ rispetto al medesimo errore di stima relativo allo stesso valore di T_x riferito però ad un dominio interamente interessato dalla strategia campionaria. La relazione (13) permette così di impiegare i risultati sperimentali contenuti nella tavola 3 per calcolare,[19] con sufficiente approssimazione, l'errore campionario atteso della stima di una frequenza assoluta riferita ad un qualsiasi dominio non interamente campionato e di cui si conosce il valore del relativo parametro γ .

A riguardo, il parametro γ esprime il "grado di coinvolgimento" della popolazione del dominio preso a riferimento, nella strategia campionaria definita per il censimento della popolazione.

In particolare, γ assume valori nell'intervallo [0;1]:

- $\gamma = 1$ nei casi in cui il dominio coincide con un ambito territoriale costituito solo da aree di censimento di centro abitato campionabili (rilevazione campionaria tramite l'impiego di questionari *short* e *long* in tutto il territorio);
- $\gamma = 0$ nei casi in cui nel dominio nessuna parte è coinvolta dalla strategia campionaria (rilevazione esaustiva tramite *long form* in tutto il territorio).

[19] Supposto di aver stimato la frequenza assoluta T relativa ad una cella di incrocio di modalità di una tavola qualunque riferita ad un dominio territoriale non interamente sottoposto a campionamento, il cv atteso della stima di tale frequenza è ridotto di una quota individuata tramite il valore del parametro γ (riferito al dominio di stima) rispetto allo stesso valore stimato su un dominio interamente sottoposto a campionamento.

6. Misure di accuratezza di tavole statistiche determinate con dati provenienti da short/long form

6.1 Premessa

La principale condizione per la definizione di una strategia di campionamento è quella di garantire livelli di accuratezza accettabili delle stime calcolate ai differenti livelli territoriali; a riguardo, più ampio è l'ambito del dominio di riferimento, maggiore dovrà essere la precisione delle stime che si intende produrre a quel livello territoriale.

Da quanto messo in evidenza nel capitolo 5, l'introduzione delle tecniche di campionamento al censimento delle popolazione e delle abitazioni potrebbe comportare il rischio di confrontarsi con stime per incroci molto fini, con livelli di qualità critici. Questo rischio è tanto più elevato quanto più ridotta è la frazione sondata scelta per la formazione dei campioni di famiglie e quanto più diffuso è, per un dato dominio, il coinvolgimento della popolazione nella strategia campionaria. Del resto, i risultati del censimento che saranno dapprima prodotti e successivamente diffusi sono, per tradizione, tali che al diminuire del dettaglio territoriale aumenta l'insieme di incroci proposti, con possibili conseguenze sulla precisione.

L'interrogativo di partenza è stato, dunque, il seguente: fissata la strategia campionaria, come è possibile valutare l'accuratezza complessiva dell'informazione rappresentata da una tavola statistica definita per un prefissato dettaglio informativo e riferita ad un dato livello territoriale?

In questo ambito è descritta la metodologia impiegata proprio per studiare l'impatto della strategia campionaria sulla qualità delle tavole di diffusione che saranno oggetto del futuro piano di diffusione nazionale ed internazionale, dei risultati censuari.

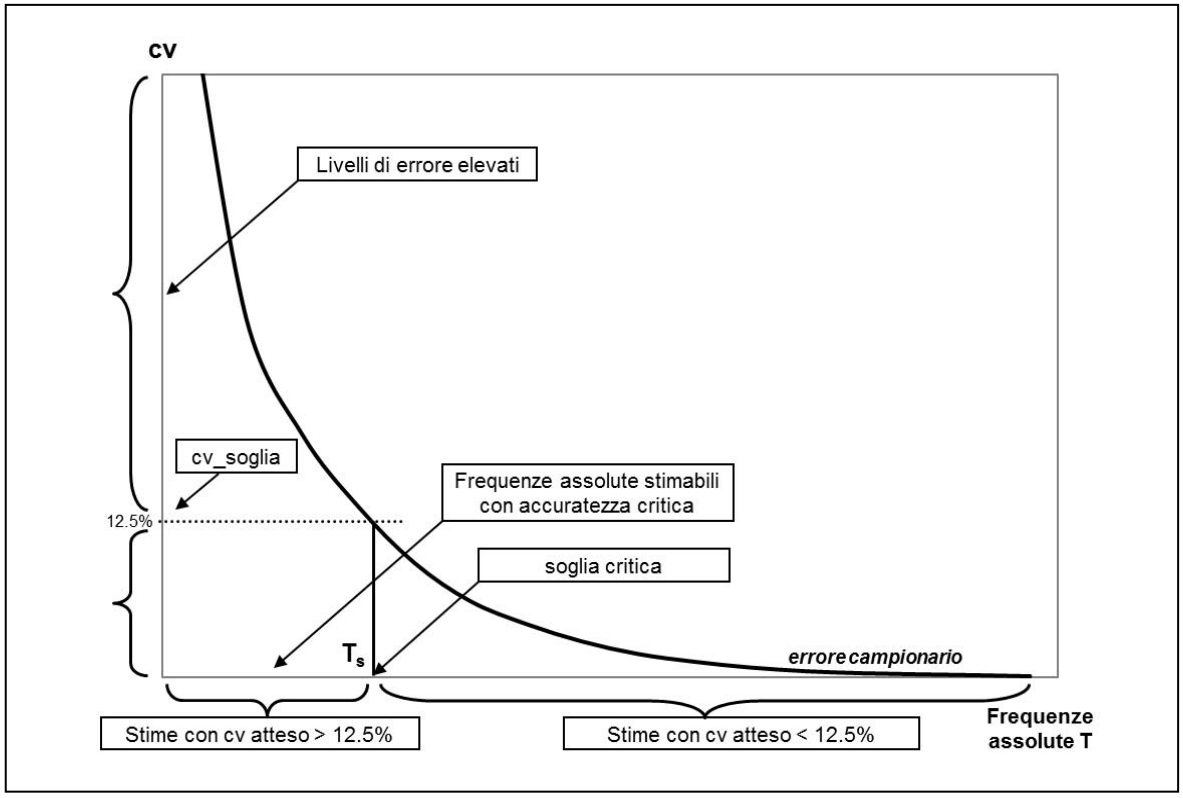
6.2 Metodologia

L'approccio suggerito si è basato sull'idea che, fissato un livello di errore campionario massimo accettabile (pari a un valore "soglia" giudicato critico del coefficiente di variazione) le stime con errori superiori presentano livelli di accuratezza non soddisfacenti in quanto, per l'elevata ampiezza del corrispondente intervallo di confidenza, comporterebbero elevati margini di indecisione. In tal modo, è possibile esprimere giudizi qualitativi sulla bontà della procedura di stima valutando la dimensione dell'insieme delle frequenze assolute riferite alla stessa tavola statistica, la cui stima potrebbe comportare un errore superiore alla soglia critica prefissata.

Quindi, una volta definiti la strategia di campionamento, in termini di disegno e di stimatore, il dominio territoriale di riferimento per la produzione delle stime e il livello di errore campionario soglia (cv_soglia) è possibile derivare (Carbonetti *et al.*, 2008a), sulla corrispondente curva dei valori percentuali attesi del coefficiente di variazione, la *soglia critica* T_S : tutte le frequenze assolute oggetto di stima inferiori alla soglia T_S saranno stimate con un errore campionario atteso non inferiore al cv_soglia fissato; viceversa, per tutte le frequenze assolute da stimare superiori a T_S la relativa stima comporterà un errore inferiore all'errore critico.

Questo tipo di analisi è rappresentato nella Figura 2, in cui è tracciato il grafico di una generica curva degli errori specificata in relazione ad una strategia campionaria che considera una frazione sondata qualunque e il caso di stime di frequenze assolute riferite ad un dominio territoriale qualsiasi. La figura mostra come è possibile individuare l'insieme di stime con accuratezza critica: per un dato valore di cv (per esempio, pari al 12,5%) tramite la curva degli errori campionari attesi è possibile individuare la frequenza assoluta soglia T_S sotto la quale tutte le frequenze assolute saranno stimate con livelli di cv maggiori del valore critico fissato. In tal modo, le frequenze più piccole di T_S potranno essere considerate casi critici in quanto potrebbero comportare elevati errori campionari; invece, per le frequenze maggiori della soglia T_S l'accuratezza potrà essere ritenuta soddisfacente perché sono attesi errori campionari più bassi del cv_soglia.

Figura 2 - Esempio di individuazione, su una ipotetica curva degli errori campionari attesi, della soglia critica e del relativo insieme di frequenze assolute stimabili con accuratezza critica, in relazione ad un prefissato livello di errore di campionamento



Si fa presente che la curva degli errori di campionamento, disegnata in corrispondenza di stime determinate con una prefissata strategia e riferite ad uno specifico dominio territoriale, può assumere forme più schiacciate verso il basso per effetto di due possibili condizioni:

a) per valori elevati della frazione sondata, perché il campione risulta numericamente più ampio e le stime più accurate per diminuzioni della varianza campionaria e quindi del cv;
b) per quote elevate di popolazione del dominio non coinvolta dalla strategia campionaria, in quanto, come è stato illustrato nel paragrafo 5.4, queste situazioni conducono a valori bassi del parametro γ a vantaggio di riduzioni dell'errore campionario (relazioni (13) e (14)).

Di conseguenza, in corrispondenza di un prefissato livello soglia di errore, curve degli errori più schiacciate verso il basso comportano riduzioni del valore soglia T_S e dell'insieme delle frequenze assolute oggetto di stima per le quali l'errore atteso potrebbe essere superiore a quello fissato come critico, con vantaggi per l'accuratezza complessiva della tavola.

6.3 Indicatori di accuratezza

In base a quanto illustrato nel precedente paragrafo, per un determinato dominio e in riferimento a una fissata tavola di risultati censuari oggetto di stima, è possibile calcolare sia la percentuale di celle le cui frequenze assolute stimate sono inferiori alla soglia critica, sia la percentuale di unità classificate in quelle celle (ritenute critiche). In particolare, quest'ultima quantifica l'ammontare di dati stimati con un basso livello di accuratezza; basse percentuali di unità classificate nelle celle critiche indicano una buona qualità dell'informazione riferita a quella tavola (per esempio, il 10% potrebbe essere ritenuto un valore di difettosità massimo accettabile).

Sono quindi stati introdotti tre indicatori che permettono di esprimere valutazioni oggettive sull'accuratezza globale di una tavola di diffusione oggetto di stima. Tali indicatori sono:

FC → *Frequenza Critica*: valore soglia sotto la quale la frequenza assoluta è stimabile con un errore atteso superiore ad un dato livello di errore critico fissato sulla corrispondente curva degli errori;

CC → *Celle Critiche* (espresso in valori percentuali): percentuale di celle non vuote la cui frequenza assoluta è inferiore alla soglia FC rispetto a tutte le celle non vuote della tavola;

UC → *Unità Critiche* (espresso in valori percentuali): percentuale di unità classificate in celle la cui frequenza assoluta è inferiore alla soglia FC rispetto a tutte le unità classificate nella tavola.

6.4 Relazione tra la quota di popolazione eleggibile al campionamento e la frequenza critica

Prima di passare a descrivere le analisi sperimentali condotte e i risultati raggiunti, è importante illustrare la relazione che lega la quota γ di popolazione di un dato dominio che è interessata dalla strategia di campionamento, al valore soglia della frequenza assoluta sotto la quale le stime potrebbero comportare errori superiori a quello fissato.

Se per la stima di una prefissata frequenza assoluta T^*, riferita ad un dominio qualsiasi non interamente sottoposto a campionamento, si richiede un prestabilito livello di errore cv^* (per esempio, si può ritenere accettabile un cv pari al 12,5%), in base alla relazione (13) illustrata nel paragrafo 5.4 è possibile ricavare il valore ideale γ^* del parametro che esprime la quota di popolazione del dominio eleggibile al campionamento, che soddisfa la condizione richiesta:

$$\gamma^* \cong \frac{cv^*}{cv(T^*)} \tag{15}$$

in cui il valore del denominatore $cv(T^*)$ è ricavabile dalla curva degli errori campionari disegnata con riferimento a un dominio interamente campionato (caso $\gamma = 1$).

Il principale risultato che emerge dalla (15) è che, per ogni dominio con un valore di $\gamma \leq \gamma^*$ l'errore atteso di ogni valore stimato maggiore di T^* sarà non superiore al cv^* fissato.

Dal momento che i livelli di errore campionario sono stati rappresentati, in questo lavoro, tramite una distribuzione in classi di frequenze assolute (Tavola 3), si può facilmente estendere il ragionamento al caso di una classe di valori da stimare.

Si prenda ad esempio, sulla tavola 3, la classe [<10]; con riferimento alla frazione sondata del 33%, l'errore atteso di una qualunque frequenza assoluta inferiore a 10 è misurato, come valore mediano nella classe, da un cv pari al 66,5%. Supponendo di fissare il livello di errore massimo tollerabile per la stima delle frequenze assolute appartenenti a questa classe, pari alla soglia del 12,5%, applicando la (15) si ottiene il seguente valore ottimale della quota γ :

$$\gamma^* = \frac{12,5}{66,5} = 0,188$$

per cui, solo per i domini di stima con valori della quota γ inferiori a 0,188 il processo inferenziale garantirà livelli di errore non superiori alla soglia del 12,5% già a partire dalla prima classe di frequenze assolute [<10]; in virtù di ciò, tutti i valori riferiti a tali domini saranno stimati con errori non superiori alla soglia fissata.

Se ora si considera la successiva classe [10├30], sempre con riferimento al disegno di campionamento che impiega la frazione sondata del 33%, l'errore atteso (come valore mediano) di una qualunque frequenza assoluta compresa tra 10 e 30 è pari al 33,8%. Fissando sempre l'obiettivo di un errore massimo accettabile del 12,5% per la stima delle frequenze assolute di questa classe, tramite la (15) si ottiene la quota ottimale:

$$\gamma^* = \frac{12,5}{33,8} = 0,370$$

che porta ad affermare che, per tutti i domini territoriali con valori della quota γ inferiori a 0,370, la procedura di stima assicurerà errori non superiori alla soglia del 12,5% per tutte le frequenze assolute a partire da quelle della classe [10├30]. Di conseguenza, per tali domini, la frequenza assoluta pari a 10 può essere ritenuta critica in corrispondenza dell'errore soglia del 12,5%, nel senso che per la stima di ogni valore più piccolo di 10 è atteso un errore superiore al livello soglia. Però, per quanto verificato in precedenza nel caso della classe di valori [<10], se i domini in questione hanno un valore della quota γ inferiore a 0,188, quest'ultimo risultato decade perché per tali domini tutte le stime avranno errori inferiori alla soglia critica.

Queste valutazioni portano a concludere che: nel caso della frazione sondata del 33%, per i domini con una quota γ compresa tra 0,188 e 0,370 la frequenza critica FC oltre la quale gli errori attesi sono non superiori al 12,5% è pari a 10 (mentre per frequenze più piccole sono attesi errori maggiori); per i domini con una quota γ inferiore a 0,188 non esiste una frequenza critica (FC=0) e tutte le stime avranno errori attesi non superiori al 12,5%.

Estendendo il ragionamento alle altre classi di frequenze assolute e in base ai livelli di errore riportati nella tavola 3, viene definita (in modo empirico) la tavola 4 che riassume, per le tre frazioni sondate esaminate, la relazione tra il valore della quota γ e le frequenze critiche FC determinate per l'errore percentuale soglia del 12,5% (tale relazione può essere riprodotta in modo simile per differenti livelli dell'errore soglia).

Tavola 4 - Classificazione dei valori della quota γ per valori della frequenza critica FC corrispondenti al livello di errore soglia del 12,5% (in termini di coefficiente di variazione percentuale)**, con riferimento al disegno casuale semplice di famiglie per le frazioni di campionamento del 10%, del 20% e del 33%**

CLASSI DI QUOTE γ	Frequenze critiche (FC) relative al livello errore (cv) soglia del 12,5%						
FRAZIONE SONDATA	0	10	30	50	100	250	500
10%	< 0,087	0,087 ├ 0,165	0,165 ├ 0,241	0,241 ├ 0,324	0,324 ├ 0,493	0,493 ├ 0,778	≥ 0,778
20%	< 0,123	0,123 ├ 0,258	0,258 ├ 0,393	0,393 ├ 0,561	0,561 ├ 0,794	≥ 0,794	
33%	< 0,188	0,188 ├ 0,370	0,370 ├ 0,534	0,534 ├ 0,717	≥ 0,717		

Le celle della tavola 4 descrivono le classi delle quote γ corrispondenti alle frequenze critiche per ciascuna frazione sondata esaminata; pertanto, per un dominio con un dato valore γ e per una prefissata frazione sondata, è possibile individuare la soglia di frequenze assolute oltre la quale le stime riferite a quel dominio saranno prodotte con errori non superiori alla soglia del 12,5%.

Per esempio, se per un dominio di diffusione qualsiasi (regione, provincia, comune) si registra un valore della quota γ pari a 0,67, il valore della frequenza critica FC nel caso della frazione sondata del 10% è pari a 250. Per quel valore di γ e per quella frazione sondata, le frequenze assolute stimate con valori più piccoli di 250 potrebbero essere affette (con elevata probabilità) da un errore

superiore alla soglia del 12,5% . In corrispondenza dello stesso valore di γ, per effetto dello "schiacciamento" della curva degli errori che avviene al crescere della frazione sondata (cfr. paragrafo 5.3), con la frazione sondata del 20% la frequenza critica si riduce a 100, mentre nel caso della frazione sondata pari al 33% scende a 50.

Per domini con valori molto piccoli di γ (quote molto basse di popolazione residente coinvolta nell'operazione campionaria) non esiste una frequenza critica e quindi tutte le frequenze assolute sono stimabili con margini di errore inferiori al valore fissato (12,5%).

L'analisi è proseguita con specifiche valutazioni, effettuate tramite il calcolo degli indicatori proposti nel paragrafo 6.3, dell'accuratezza complessiva di alcune tavole (illustrate nel prossimo capitolo) con dati del Censimento del 2001, aventi diverso dettaglio informativo e relative a differenti livelli territoriali: regionale, comunale e sub-comunale (area di censimento di centro abitato).[20] Nei capitoli successivi (9 e 10) sono esposti i risultati dello studio che ha considerato la prima ipotesi progettuale relativa all'impiego di questionari di tipo *short* e *long* (secondo le versioni descritte nel paragrafo 4.2). Nel capitolo 11, a seguito della scelta definitiva di una versione più ampia del questionario in forma ridotta rispetto alla versione proposta nella rilevazione pilota, sono illustrati alcuni possibili riflessi sull'efficienza delle stime e sull'accuratezza delle tavole statistiche di diffusione.

7. Le tavole statistiche oggetto della sperimentazione

7.1 I criteri per la scelta

La scelta delle tavole statistiche su cui procedere con le valutazioni dell'accuratezza attesa dei risultati censuari prodotti con l'impiego dei metodi di stima, ha risposto sia all'esigenza di rappresentare quanto richiesto da Eurostat, sia alle crescenti esigenze degli utenti nazionali di dati con elevato dettaglio territoriale.

Le tavole oggetto della sperimentazione sono state selezionate cercando di diversificare il contenuto informativo, in termini di variabili di incrocio e di classificazioni, e di riferire i dati a diversi livelli territoriali. Riguardo quest'ultimo aspetto, è stato considerato il livello regionale in relazione al piano di diffusione europeo, mentre livelli territoriali più fini (comunale e sub-comunale) sono stati esaminati per gli obiettivi di interesse nazionale.

L'individuazione delle tavole per la sperimentazione è stata, inoltre, subordinata alla possibilità di ricondurre le variabili di incrocio e le relative classificazioni ai dati del passato censimento della popolazione; solo per le tavole che hanno soddisfatto questa condizione è stato possibile procedere all'operazione di "popolamento" con i dati del 2001 (cfr. capitolo 8) e ai successivi esercizi sperimentali.

7.2 Le tavole selezionate per il dettaglio regionale

Da una versione provvisoria del piano di diffusione[21] europeo sono stati selezionati 8 ipercubi contenenti unicamente variabili rilevate al Censimento italiano del 2001 e con classificazioni di diffusione che si avvicinavano, sia in termini di numerosità che di contenuto informativo, alle classificazioni previste per la prossima tornata censuaria. Così, ad esempio, sono stati esclusi gli ipercubi contenenti la variabile "*ever resided abroad*" che l'Italia rileva per la prima volta nel 2011, ma anche variabili e classificazioni relative a famiglie e nuclei in alcuni casi classificate, in Italia, in modo diverso da quello previsto a livello internazionale. L'utilizzo di ipercubi relativi a famiglie, nu-

[20] Per l'analisi i totali riferiti sia al dominio comunale che a quello di area di censimento sono ottenuti come aggregazione dei relativi dati prodotti a livello di sezione di censimento; si precisa che le informazioni relative al dettaglio comunale sono state calcolate considerando tutte le sezioni appartenenti al comune stesso e non solo quelle costituenti le aree di censimento campionate.

[21] All'avvio delle sperimentazioni il piano di diffusione europeo era ancora in corso di definizione.

clei familiari, popolazione in famiglia e in nuclei, avrebbe comportato la ridefinizione e il ricalcolo delle classificazioni relative in quanto differenti da quelle previste nella diffusione per l'Italia e ciò avrebbe allungato i tempi della sperimentazione.

Non sono stati altresì presi in considerazione gli ipercubi richiesti da Eurostat a livello comunale (LAU2) perché le variabili previste a tale livello, essendo solo di natura demografica, non sono coinvolte nelle procedure di stima, ma inserite sia nel questionario *short* che in quello *long* e pertanto rilevate su tutta la popolazione. Ci si è perciò concentrati sulle variabili richieste solo a livello regionale e quindi sugli ipercubi riferiti al livello territoriale NUTS2; sono stati selezionati ipercubi che includevano solo le caratteristiche della popolazione, diversi tra loro e che coprivano differenti tematiche previste per la rilevazione campionaria (Tavola 5).

Tavola 5 - Descrizione degli ipercubi Eurostat riferiti al livello territoriale regionale (NUTS2) oggetto della sperimentazione

CODICE	Nome	Universo di riferimento	Variabili di incrocio (numero di modalità della classificazione)
H.B1.E0.R1	Single ages – "Current activity status"	Popolazione totale	Sesso (2) Età (101) Condizione professionale (6)
H.B1.E0.R2	Single ages – "Occupation"	Popolazione totale	Sesso (2) Età (101) Professione (10)
H.B1.E0.R3	Single ages – "Industry"	Popolazione totale	Sesso (2) Età (101) Sezioni di attività economica (17)
H.B1.E0.R4	Single ages – "Status in employment"	Occupati	Sesso (2) Età (101) Posizione nella professione (6)
H.B1.E0.R5	Single ages – "Educational attainment"	Popolazione totale	Sesso (2) Età (101) Grado di istruzione (7)
H.B1.E1.R2	Employment – "Occupation"	Popolazione totale	Sesso (2) Età (21) Condizione professionale (8) Professione (10) Grado di istruzione (7)
H.B1.E1.R3	Employment – "Industry"	Popolazione totale	Sesso (2) Età (21) Condizione professionale (6) Sezioni di attività economica (17) Grado di istruzione (7)
H.B1.E1.R4	Employment – "Occupation" by "Industry"	Popolazione totale	Sesso (2) Età (13) Professione (10) Sezioni di attività economica (17) Grado di istruzione (7)

Si noti che, un insieme minimo di variabili è comune a tutti gli ipercubi esaminati e include "sesso" ed "età" anche se per quest'ultima ci sono alcuni ipercubi con una classificazione per singolo anno (a 101 modalità), altri con classificazioni più aggregate (a 13 o a 21 classi). Nei primi cinque ipercubi le variabili socio-economiche sono considerate una per volta e l'incrocio con sesso ed età avviene al massimo dettaglio (2x101 modalità); negli altri tre, le variabili sono combinate tra loro e l'incrocio con sesso ed età è proposto a minori dettagli (2x21 o 2x13 modalità).

Nell'ambito della Task Force[22] era stato inizialmente previsto, per ogni ipercubo, un numero di celle calcolato come prodotto del numero di modalità di classificazione di ciascuna variabile conte-

[22] Vedi paragrafo 3.1.2.

nuta nell'ipercubo. È stato poi introdotto un nuovo concetto di ampiezza per gli ipercubi, la "*relevant size*", calcolata escludendo dal primo conteggio il numero di modalità che corrispondono ai totali e ai sub-totali.

Degli ipercubi selezionati, la "*relevant size*" (Tavola 6) è relativa al numero di celle potenziali ottenuto come prodotto delle modalità di diffusione previste per il Censimento italiano del 2001. Sono stati poi esclusi gli "zeri strutturali", ovvero, le frequenze certamente uguali a zero perché corrispondenti a risultati "impossibili" (per esempio, l'età "15 anni" con il titolo di studio "ISCED5-laurea"); si giunge così al numero di celle possibili.

Tavola 6 - Numero di "celle potenziali" e numero di "celle possibili" degli 8 ipercubi Eurostat

CODICE	Numero di celle potenziali	Numero di celle possibili
H.B1.E0.R1	1.212 (2x101x6)	1.062
H.B1.E0.R2	2.020 (2x101x10)	1.922
H.B1.E0.R3	3.434 (2x101x17)	3.126
H.B1.E0.R4	1.212 (2x101x6)	1.032
H.B1.E0.R5	1.414 (2x101x7)	1.342
H.B1.E1.R2	23.520 (2x21x8x10x7)	3.810
H.B1.E1.R3	29.988 (2x21x6x17x7)	5.574
H.B1.E1.R4	30.940 (2x13x10x17x7)	26.350

È immediato verificare come i primi cinque ipercubi, che incrociano le variabili demografiche con una sola variabile socio-economica, sono molto gestibili in quanto presentano dimensioni ridotte anche solo per il numero di celle potenziali. Tale ampiezza si riduce, anche se non di molto, quando si vanno a considerare solo le celle possibili, cioè quelle per le quali è verosimile attendersi una frequenza non nulla. Passando agli altri ipercubi, invece, anche se si riduce la classificazione dell'età (non più per singoli anni ma in classi quinquennali o decennali), aumenta il numero delle variabili incrociate e aumentano le variabili socio-economiche coinvolte; di conseguenza, cresce in modo evidente l'ampiezza potenziale degli ipercubi ma, analizzando solo i casi di possibile frequenza delle celle, l'aumento della dimensione risulta più contenuto. L'ipercubo con la dimensione più grande è l'ultimo (H.B1.E1.R4), che incrocia, oltre a "sesso" ed "età" (in 13 classi), anche "professione", "sezioni di attività economica" e "grado di istruzione".

7.3 Le tavole selezionate per il dettaglio comunale e sub-comunale

Per le valutazioni sull'accuratezza attesa dei risultati censuari riferiti ad ambiti territoriali comunali e sub-comunali, le tavole statistiche oggetto della sperimentazione sono state selezionate dal piano di diffusione italiano del 2001: contengono variabili rilevate a campione tramite *long form* incrociate con variabili demografiche quali "sesso", "età", "cittadinanza" e "stato civile", prese singolarmente o in modo congiunto fino a un massimo di due.

In particolare, dall'analisi dei dati disponibili sul *Data Warehouse* del Censimento della popolazione e delle abitazioni del 2001 (DAWINCI) sono state individuate 15 tavole, oggetto di diffusione a partire dal livello comunale, contenenti variabili rilevabili nel Censimento del 2011 solo tramite il questionario in forma completa incrociate con dati esclusivamente di carattere demografico. Nello specifico, sono stati proposti due differenti gruppi di tavole statistiche: nel primo gruppo[23] (blocco 1) le tematiche socio-economiche incrociano le variabili "sesso" e/o "età" (Tavola 7); nel secondo gruppo[24] (blocco 2) le tematiche socio-economiche incrociano le variabili relative alla "cittadinanza"

[23] La scelta di tali tavole ha avuto il fine di dare maggiore attenzione all'incrocio delle variabili socio-economiche con le variabili demografiche relative a sesso e classe di età.

[24] Per la scelta di tali tavole il criterio seguito è stato quello di considerare l'incrocio con variabili demografiche diverse da quelle del blocco 1 e comunque oggetto del piano di diffusione nazionale del Censimento 2001; a riguardo, si precisa che alcune di queste tavole sono state diffuse solo per i Grandi Comuni (con popolazione residente superiore a 150mila unità).

o allo "stato civile" (Tavola 8). Successivamente si è proceduto a riprodurre, con i dati del 2001, le tavole scelte per i 40 comuni considerati nelle sperimentazioni citate nel paragrafo 5.2 e per i relativi domini sub-comunali (le 498 aree di censimento).

Tavola 7 - Descrizione delle tavole statistiche del piano di diffusione italiano riferite ai livelli territoriali comunale e sub-comunale prese in considerazione per le analisi qualitative sperimentali. Blocco 1

CODICE	Nome	Universo di riferimento	Variabili di incrocio (numero di modalità di classificazione)
GRA.IST.1	Popolazione residente in famiglia di 6 anni e più per sesso e grado di istruzione	Popolazione di 6 anni e più	Sesso (2) Grado di istruzione (8)
ATT.ECO.1	Occupati per sesso, età ed attività economica	Occupati	Sesso (2) Età (4) Attività economica (3)
ATT.ECO.2	Occupati per sesso e sezioni di attività economica	Occupati	Sesso (2) Sezioni di attività economica (17)
POS.PRO.1	Occupati per sesso, posizione nella professione e attività economica	Occupati	Sesso (2) Posizione nella professione (5) Attività economica (3)
CON.PRO.1	Popolazione residente in famiglia di 15 anni e più per sesso e condizione professionale	Popolazione di 15 anni e più	Sesso (2) Condizione professionale (6)
PENDOL.1	Popolazione residente in famiglia che si sposta giornalmente per sesso e luogo di destinazione	Popolazione totale	Sesso (2) Luogo di destinazione (2)

Tavola 8 - Descrizione delle tavole statistiche del piano di diffusione italiano riferite ai livelli territoriali comunale e sub-comunale prese in considerazione per le analisi qualitative sperimentali. Blocco 2

CODICE	Nome	Universo di riferimento	Variabili di incrocio (numero di modalità di classificazione)
GRA.IST.2	Popolazione residente in famiglia di 6 anni e più per cittadinanza e grado di istruzione	Popolazione di 6 anni e più	Cittadinanza (2) Grado di istruzione (8)
GRA.IST.3	Popolazione residente in famiglia di 6 anni e più per stato civile e grado di istruzione	Popolazione di 6 anni e più	Stato civile (5) Grado di istruzione (8)
ATT.ECO.3	Occupati per stato civile ed attività economica	Occupati	Stato civile (5) Attività economica (3)
ATT.ECO.4	Occupati per cittadinanza ed attività economica	Occupati	Cittadinanza (2) Attività economica (3)
POS.PRO.2	Occupati per stato civile e posizione nella professione	Occupati	Stato civile (5) Posizione nella professione (5)
CON.PRO.2	Popolazione residente in famiglia di 15 anni e più per cittadinanza e condizione professionale	Popolazione di 15 anni e più	Cittadinanza (2) Condizione professionale (6)
CON.PRO.3	Popolazione residente in famiglia di 15 anni e più per stato civile e condizione professionale	Popolazione di 15 anni e più	Stato civile (5) Condizione professionale (6)
CON.POS.1	Condizione e posizione nella professione della persona di riferimento del nucleo familiare per numero di figli	Nuclei familiari	Condizione e posizione nella professione (13) Numero di figli del nucleo (4)
CON.POS.2	Condizione e posizione nella professione della persona di riferimento del nucleo familiare per il suo stato coniugale	Nuclei familiari	Condizione e posizione nella professione (13) Stato coniugale (2)

8. Ambiente informatico a supporto delle analisi qualitative

8.1 Consultazione ed analisi

In questo capitolo è illustrata la composizione dell'ambiente di *data warehouse*[25] che è stato specificamente utilizzato a supporto dell'analisi statistica di tipo qualitativo, oggetto dello studio presentato in questo documento. L'ambiente si compone di:

- *Tabelle dei Metadati o delle Dimensioni*,[26] organizzate in tabelle di descrizione degli ambiti territoriali di riferimento e in tabelle di descrizione delle variabili d'analisi;
- *Tabelle dei Fatti o Datamart*,[27] volte a contenere le frequenze assolute per domini territoriali predefiniti (nell'esercizio di questo lavoro sono stati presi in considerazione le sezioni di censimento e le regioni) di ciascuna variabile d'analisi rispetto a determinate modalità di classificazione.

8.1.1 La base di dati e le utenze

L'ambiente di *data warehouse* di supporto alle analisi statistiche qualitative si poggia su una *istanza di database Oracle*.[28] L'istanza di database Oracle UNG si compone di diverse utenze; a riguardo, le utenze destinate al lavoro sono due:

- una utenza di sviluppo rivolta alle attività di sviluppo software del gruppo informatico che si compone delle seguenti strutture-dati:
  - tabelle dei Metadati (o delle Dimensioni);
  - tabelle dei Fatti (o Datamart).
- una utenza di lettura destinata alla produzione e alle successive analisi statistiche; questa si configura in prevalenza di sinonimi creati su tutte le strutture-dati di rilevanza ai fini dell'attività di consultazione e analisi.

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

8.2 Gestione ed utilizzo dei dati

8.2.1 Le tabelle dei Metadati

Le tabelle dei Metadati dell'ambiente di analisi svolgono la funzione di dare informazioni di carattere descrittivo per le "variabili" utilizzate.

La tabella **REGIONI** contiene i codici territoriali delle regioni italiane e si compone di:

- campi relativi ai codici identificativi di ciascuna unità territoriale d'analisi:
  - CODICE_REG, per il codice di regione;
  - CODICE_RIP, per il codice di ripartizione geografica.
- campi relativi alla descrizione di ciascuna unità territoriale:
  - DENOMINAZIONE, per la descrizione della regione.

La tabella REGIONI contiene una riga per ogni regione italiana appartenente a una specifica ripartizione del territorio nazionale.

[25] Il *data warehouse* è una raccolta di dati integrata, orientata all'utente, variabile nel tempo e non volatile, di supporto ai processi decisionali.
[26] Le tabelle dei Metadati o delle Dimensioni sono strutture-dati che contengono informazioni aggiuntive che descrivono e arricchiscono i dati contenuti nel *data warehouse*.
[27] Le tabelle dei Fatti o Datamart sono raccoglitori di dati specializzati in un particolare soggetto. In termini più tecnici, un Datamart è un sottoinsieme logico o fisico di un *data warehouse* di maggiori dimensioni.
[28] L'istanza di database Oracle è costituita da un set di processi di sistemi e strutture di memoria che interagiscono con i dati memorizzati.

La tabella **SEZIONI** include i codici territoriali delle sezioni di censimento e si compone di:

- campi relativi ai codici identificativi di ciascuna unità territoriale:
  - CODPRO, per il codice di provincia;
  - CODCOM, per il codice di comune;
  - NSEZ, per il numero della sezione.
- campi relativi alla descrizione di ciascuna unità territoriale:
  - DZPRO, per la descrizione della provincia;
  - DZCOM, per la descrizione del comune.
- un campo PROG_PROV, non utilizzato per fini consultativi ma esclusivamente procedurali.

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

La tabella **MODALITA**, comprende le meta informazioni costituendo il catalogo delle informazioni prodotte in relazione alla tabella dei microdati sorgenti; essa si compone dei seguenti "campi":

- COD_CL, che contiene una sigla alfanumerica assegnata alla variabile di classificazione;
- IND_MODALITA, volto a comprendere i progressivi numerici assegnati alle modalità di ogni variabile di classificazione;
- NOME_MODALITA, che include la descrizione di ciascuna modalità delle variabili di classificazione;
- LIV_GERARCHICO, che contiene il livello di gerarchia cui appartiene ciascuna modalità della variabile di classificazione.[29]

La tabella MODALITA comprende una riga per ogni modalità delle differenti variabili di classificazione considerate.

Di seguito è descritta, a titolo di esempio, una delle principali interrogazioni alla tabella **MODALITA** con le corrispondenti istruzioni in linguaggio SQL:

- *interrogazione* - individuazione delle modalità di una variabile di classificazione e delle relative descrizioni, specificando la sigla della classificazione interessata (nel caso della variabile "Classe di età", con 13 modalità: "eta13"):

```
select modalita.cod_cl          sigla_classificazione,
    modalita.ind_modalita       indice_modalità_classificazione,
    modalita.nome_modalita      nome_modalità_classificazione,
    modalita.liv_gerarchico     gerarchia_modalità_classificazione
from modalita
where cod_cl ='eta13'
order by modalita.ind_modalita;
```

8.2.2 Le tabelle dei Fatti

Le tabelle dei Fatti o Datamart (struttura **DATI2**_XXXXX) sono state costruite tramite l'adozione di una struttura omogenea composta da quattro parti, di seguito illustrate:

1. Identificazione dell'incrocio, in cui ogni combinazione delle modalità di classificazione oggetto d'analisi è individuabile univocamente mediante un progressivo numerico che è contenuto nel campo: ID_INCROCIO;

[29] Se non ci sono gerarchie tra le modalità questo campo assumerà valore 1 per tutte le modalità della variabile, se invece, per esempio, la variabile presenta dei "di cui" di modalità principali, tutte le modalità corrispondenti al "di cui" assumeranno un diverso livello gerarchico rispetto a quello della modalità principale di cui costituiscono un sottogruppo.

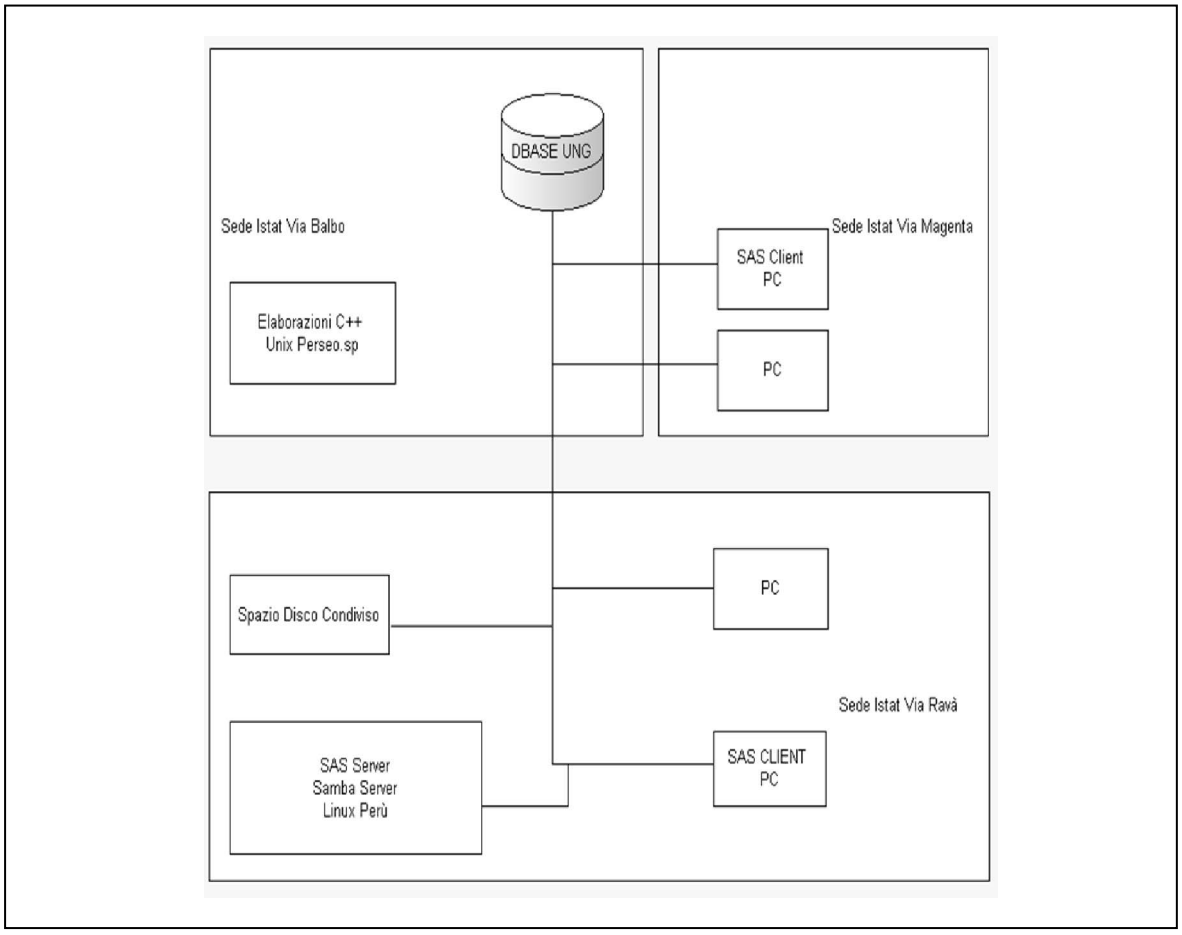
2. Identificazione del territorio, per due tipologie di Datamart, una in cui l'unità territoriale minima è la sezione di censimento e una in cui è la regione; i campi che individuano le unità territoriali d'analisi sono rispettivamente:
   - per i Datamart a livello di sezione di censimento: CODPRO; CODCOM; NSEZ;
   - per i Datamart a livello regionale: COD_REGIONE.
3. Modalità di classificazione, i cui campi sono finalizzati a contenere gli indici di modalità di ciascuna classificazione oggetto d'analisi. Il numero dei campi del Datamart che identificano le variabili qualitative dipende dal numero delle modalità di classificazione coinvolte nella tavola statistica da produrre. Ogni colonna di classificazione è identificata dal nome della variabile ("cod_cl") a essa assegnata, così come risulta dalla tabella dei metadati "MODALITA";
4. Conteggio della popolazione residente determinato per il livello territoriale d'analisi e identificato dal campo: FREQUENZA.

A ciascuna tabella dei Fatti è stata assegnata una specifica *etichetta*:

- per i Datamart per sezione di censimento, si compone delle sigle che riconducono alle informazioni oggetto d'analisi. Per esempio, in [**DATI2**_cod_oggprog] il "_cod_ogg" rappresenta il codice dell'oggetto d'analisi e il "prog" è soltanto un progressivo numerico per distinguere Datamart relativi a uno stesso oggetto;
- per i Datamart per regione, si compone della sigla attribuita a ciascuna tavola statistica da produrre. Ad esempio, in **[DATI2**_nometavola**]** il "_nometavola" rappresenta la sigla della tavola statistica.

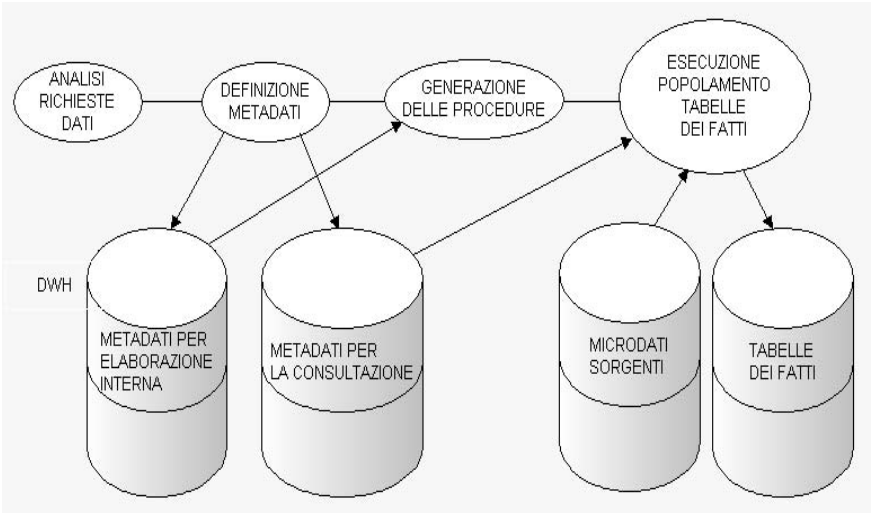
8.2.3 Elaborazioni e condivisione delle risorse informatiche

Figura 3 - Schema dell'organizzazione territoriale delle risorse informatiche



La Figura 4 descrive il flusso logico del processo di definizione e popolamento dei meta-dati e dei macro-dati.

Figura 4 - Rappresentazione del flusso di popolamento dei meta-dati e macro-dati



Il processo ha presupposto un'attività di analisi sui micro dati sorgenti di produzione e di diffusione del Censimento della Popolazione e delle Abitazioni del 2001. A completamento dell'analisi, sono stati definiti i metadati specifici sulla base dei quali sono state create le procedure di popolamento delle tabelle dei Fatti.

8.2.4 Strutture del Data Warehouse

In questo paragrafo è descritta la struttura fisica di alcune delle tabelle, distinte per tipologia, di cui si è composto il *data warehouse* di supporto alle analisi statistiche.

8.2.4.1 Struttura delle tabelle dei Metadati

Tabelle degli Ambiti territoriali

La tabella **REGIONI** contiene una riga per ogni regione all'interno di una specifica ripartizione geografica italiana.

NOME_COLONNA	FORMATO	DESCRIZIONE
COD_RIPARTIZIONE	NUMBER(1)	Codice di Ripartizione
COD_REGIONE	NUMBER(2)	Codice di Regione
DESCRIZIONE	VARCHAR2(40)	Denominazione della Regione

La tabella **SEZIONI** contiene una riga per ogni sezione di censimento all'interno di un fissato comune e di una determinata provincia.

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

Tabelle di descrizione delle variabili

La tabella **MODALITA** comprende una riga per ogni modalità di classificazione di una variabile.

NOME_COLONNA	FORMATO	DESCRIZIONE
COD_CL	VARCHAR2(30)	Nome della variabile d'analisi
IND_MODALITA	NUMBER(2)	Indice della modalità della variabile d'analisi
NOME_MODALITA	VARCHAR2(300)	Descrizione della modalità della variabile d'analisi
NOME_MODALITA_ENG	VARCHAR2(300)	Descrizione in lingua inglese della modalità della variabile d'analisi
LIV_GERARCHICO	NUMBER	Livello gerarchico cui appartiene ciascuna modalità della variabile d'analisi

8.2.4.2 Struttura delle tabelle dei Fatti

Datamart per regione

La tabella **DATI2_HB1E1R4**, presa ad esempio tra quelle richieste da Eurostat e prodotte, a scopo di esercizio, per le analisi riferite al livello regionale, contiene una riga per ogni possibile incrocio delle modalità di classificazione delle variabili "ses2m", "eta13", "attlavs", "ateco17m", "titstu7m" e una per la relativa frequenza calcolata per regione con riferimento all'oggetto d'analisi "Popolazione residente in famiglia".

NOME_COLONNA	FORMATO	DESCRIZIONE
ID_INCROCIO	NUMBER(6)	Identificativo dell' incrocio
COD_REGIONE	NUMBER(2)	Codice di Regione
ETA13	NUMBER(9)	Classe di età (13 modalità)
ATTLAVS	NUMBER(9)	Attività lavorativa svolta (10 modalità)
ATECO17M	NUMBER(9)	Sezioni di attività economica (17 modalità)
TITSTU7M	NUMBER(9)	Grado di istruzione (7 modalità)
FREQUENZA	NUMBER(9)	Conteggio della Popolazione residente in famiglia per regione

Datamart per sezione di censimento

La tabella **DATI2_PRF9**, presa a scopo illustrativo tra quelle prodotte per le analisi a livello comunale e sub-comunale, contiene una riga per ogni possibile combinazione delle diverse modalità di classificazione delle variabili "stcon", "condpos13m" e una per la corrispondente frequenza calcolata a livello di sezione di censimento riferita all'oggetto d'analisi "Popolazione residente in famiglia".

NOME_COLONNA	FORMATO	DESCRIZIONE
ID_INCROCIO	NUMBER(6)	Identificativo dell' Incrocio
CODPRO	NUMBER(3)	Codice di Provincia
CODCOM	NUMBER(3)	Codice di Comune
NSEZ	NUMBER(7)	Numero della sezione di sensimento
STCON	NUMBER(1)	Stato coniugale della coppia (2 modalità)
CONDPOS13M	NUMBER(2)	Condizione e posizione nella professione (13 modalità)
FREQUENZA	NUMBER(9)	Conteggio della Popolazione residente in famiglia per sezione di censimento

9. Accuratezza di tavole statistiche riferite al livello regionale

9.1 Premessa

In questo ambito si è proceduto a valutare il possibile impatto del campionamento (Carbonetti, 2009a) sulla qualità dei risultati del censimento, con riferimento a tavole statistiche definite per il livello territoriale regionale (Carbonetti e Verrascina, 2009b; Carbonetti e Verrascina, 2010a). A tale scopo, sono stati presi in esame alcuni ipercubi (descritti nel paragrafo 7.2)

relativi al piano di diffusione europeo aventi differenti dettagli informativi ed elevati livelli di granularità.[30]

Per la fase sperimentale, come prima operazione, gli ipercubi regionali scelti (Tavola 5) sono stati popolati con i dati relativi al Censimento del 2001 per ciascuna delle 20 regioni italiane; successivamente, nell'ipotesi di una strategia campionaria basata sull'impiego di questionari *short* e *long* (nelle versioni descritte nel paragrafo 4.2) sono stati calcolati, per ciascun ipercubo e per ciascuna regione, gli indicatori FC, CC e UC (cfr. paragrafo 6.3).

I risultati sono stati dapprima presentati per singoli casi, con riferimento a due ipercubi e per tre regioni, poi sono stati valutati in modo complessivo sull'intero insieme dei domini regionali. Le valutazioni sono state ripetute, a parità di disegno campionario (casuale semplice da lista), per i tre differenti livelli di frazione sondata già oggetto di confronto nelle precedenti sperimentazioni campionarie (10%, 20%, 33%). Il tasso di campionamento, infatti, influisce in maniera significativa sulla curva degli errori (riporta errori più bassi con la frazione sondata più ampia) e, di conseguenza, sulla frequenza critica FC e sulle misure CC e UC calcolate per le tavole statistiche oggetto della sperimentazione.

In termini operativi, dopo aver definito, per ciascuna regione italiana, il valore del parametro γ (introdotto nel paragrafo 5.4) in base ad alcune ipotesi riguardo la presumibile quota di popolazione residente nella regione che potrebbe essere sottoposta a rilevazione censuaria tramite campionamento, è stata identificata la corrispondente frequenza critica FC (per un errore soglia del cv pari al 12,5%) sulla tavola 4.

9.2 Livelli di accuratezza attesa di due ipercubi di diffusione europea

In questo paragrafo sono presentati, a titolo illustrativo, i risultati delle misure di accuratezza attesa relativamente a due ipercubi richiesti da Eurostat per il livello territoriale NUTS2 (coincidente in Italia con le Regioni), tra i più complessi, in termini di numero di celle di classificazione, tra quelli considerati nello studio (cfr. paragrafo 7.2) e riprodotti con i risultati del Censimento della popolazione e delle abitazioni del 2001 secondo le modalità informatiche descritte nel capitolo 8.

Il primo esempio si riferisce all'ipercubo H.B1.E1.R3 che classifica gli "occupati" per "sesso", "età", "grado di istruzione" (*Educational Attainment*), "condizione professionale" (*Current Activity Status*) e "sezioni di attività economica" (*Industry*). Il numero totale di celle possibili per l'ipercubo in esame è 5.574 (non sono considerati gli "zero strutturali").

In relazione alla possibilità di estrarre campioni di famiglie con un disegno casuale semplice da lista anagrafica per diverse frazioni di campionamento (10%, 20%, 33%), sono state calcolate le misure di accuratezza dell'ipercubo H.B1.E1.R3 per tre regioni italiane scelte con differente dimensione: il Molise, una delle più piccole; le Marche, con una dimensione media tra le regioni italiane; la Sicilia, una delle più grandi. Per tali regioni il valore di γ impiegato ai fini dell'esercizio è stato presunto pari a 0,46 per il Molise, 0,67 per le Marche e 0,86 per la Sicilia.

Per una soglia di errore massimo tollerabile (in termini di cv) pari al 12,5%, è stata individuata la frequenza critica FC (Tavola 9) e sono state calcolate le corrispondenti misure di accuratezza CC (celle critiche) e UC (unità critiche), espresse in termini percentuali.

[30] Presenza di un grande numero di celle di classificazione in una tavola statistica. Questa situazione potrebbe portare a un'elevata dispersione delle unità classificate (tante celle con frequenze molto piccole).

Tavola 9 - Valori della Frequenza critica FC corrispondenti al livello di errore soglia del 12,5% (in termini di coefficiente di variazione percentuale) **con riferimento al disegno casuale semplice di famiglie per le frazioni sondate del 10%, del 20% e del 33%, per le regioni Molise, Marche e Sicilia**

	Molise	Marche	Sicilia
FRAZIONE SONDATA	γ=0,46	γ=0,67	γ=0,86
10%	100	250	500
20%	50	100	250
33%	30	50	100

I risultati per l'ipercubo considerato (Tavola 10) evidenziano che i livelli di accuratezza migliori sono attesi con la strategia campionaria che adotta la frazione di campionamento più elevata (33%). Se poi si fissa una qualunque delle frazioni sondate esaminate, si osservano livelli di accuratezza crescenti all'aumentare della dimensione del dominio di riferimento. Nello specifico, per la regione Sicilia e nel caso della frazione sondata del 33%, la frequenza critica FC è pari a 100, il 59,4% di celle hanno una frequenza assoluta inferiore a 100, ma in queste celle (ritenute "critiche") si classifica solo l'1% delle unità classificabili nell'ipercubo.

Tavola 10 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC, calcolati per un cv soglia pari al 12,5%, sull'ipercubo H.B1.E1.R3. per le regioni Molise, Marche e Sicilia (Censimento della popolazione 2001)

	Molise			Marche			Sicilia		
FRAZIONE SONDATA	FC	CC	UC	FC	CC	UC	FC	CC	UC
10%	100	79,2	10,7	250	78,8	6,9	500	75,9	4,2
20%	50	71,0	5,8	100	68,4	3,0	250	68,7	2,1
33%	30	63,6	3,4	50	59,8	1,5	100	59,4	1,0

La tavola 11 mostra i risultati relativi a un ipercubo più complesso di quello appena trattato. In questo secondo esempio è stata studiata l'accuratezza attesa per l'ipercubo H.B1.E1.R4 che classifica gli "occupati" per "sesso", "età", "grado di istruzione" (*Educational Attainment*), "professione" (*Occupation*) e "sezioni di attività economica" (*Industry*); per questo ipercubo il numero totale di celle possibili è 26.350 (sono esclusi gli "zero strutturali").

Le considerazioni a cui si giunge per l'ipercubo H.B1.E1.R4 sono simili a quelle evidenziate dall'analisi riferita al primo caso preso in esame. La differenza significativa riguarda l'osservazione di criticità maggiori che portano a livelli di accuratezza più bassi; ciò è giustificato dal fatto che questo secondo ipercubo presenta un numero più elevato di celle di classificazione (maggiore granularità della tavola). Comunque, pur in presenza di una leggera riduzione dell'accuratezza attesa, il giudizio sulla qualità attesa per la stima dell'intera tavola rimane favorevole.

Tavola 11 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC, calcolati per un cv soglia pari al 12,5%, sull'ipercubo H.B1.E1.R4. per le regioni Molise, Marche e Sicilia (Censimento della popolazione 2001)

	Molise			Marche			Sicilia		
FRAZIONE SONDATA	FC	CC	UC	FC	CC	UC	FC	CC	UC
10%	100	91,9	14,9	250	91,1	11,2	500	91,8	7,3
20%	50	86,5	9,3	100	84,4	6,0	250	87,6	4,5
33%	30	81,3	6,5	50	77,1	3,4	100	79,4	2,2

9.3 Accuratezza attesa per ipercubi di diffusione europea a livello regionale

Si presentata ora una sintesi delle misure di accuratezza, definite tramite il calcolo della percentuale di individui classificati in celle critiche (indicatore UC) relativamente a tutte le 20 regioni italiane, per gli 8 ipercubi presi in considerazione tra quelli richiesti da Eurostat e per le tre strategie campionarie che considerano differenti frazioni sondate. I risultati sono riassunti nella tavola 12 in cui sono esposti, per i differenti casi esaminati, il numero delle regioni italiane distribuite (in classi) in base al corrispondente valore percentuale dell'indicatore UC calcolato.

Tavola 12 - Distribuzione delle regioni italiane in base al valore atteso dell'indicatore UC calcolato per alcuni ipercubi Eurostat (Censimento della popolazione 2001)**, in riferimento a un cv soglia del 12,5% e per tre strategie campionarie differenti per il livello della frazione sondata**

NUMERO DI REGIONI	f.s. = 33%		f.s. = 20%			f.s. = 10%				
IPERCUBI EUROSTAT (*celle possibili*)	UC - Classi di valori dell'indicatore UC (% di individui classificati in celle critiche)									
	<5%	5-10%	<5%	5-10%	10-15%	<5%	5-10%	10-15%	15-20%	>20%
H.B1.E0.R1 (1.062)	20	0	20	0	0	19	1	0	0	0
H.B1.E0.R2 (1.922)	20	0	20	0	0	15	4	1	0	0
H.B1.E0.R3 (3.126)	20	0	17	3	0	11	4	4	1	0
H.B1.E0.R4 (1.032)	20	0	16	4	0	7	8	5	0	0
H.B1.E0.R5 (1.342)	20	0	20	0	0	15	5	0	0	0
H.B1.E1.R2 (3.810)	20	0	18	2	0	12	6	2	0	0
H.B1.E1.R3 (5.574)	20	0	17	3	0	10	5	5	0	0
H.B1.E1.R4 (26.350)	15	5	8	10	2	1	9	6	3	1

Dalle frequenze indicate nella tavola 12 è possibile osservare che, nel caso della strategia campionaria che impiega la frazione sondata più elevata (33%), per 7 degli 8 ipercubi esaminati la percentuale di unità classificate in celle critiche (UC) è inferiore al 5% per tutte le regioni italiane, mentre solo per l'ipercubo più complesso (H.B1.E1.R4) il valore dell'indicatore è inferiore al 5% per 15 regioni ed è compreso tra il 5% e il 10% nelle rimanenti 5 (quelle più piccole).

La strategia campionaria basata sulla frazione sondata del 20% comporterebbe una riduzione lieve dell'accuratezza attesa in quanto i valori dell'indicatore UC riferiti alle differenti regioni italiane sono tutti inferiori al 10% a eccezione di quelli calcolati con riferimento sempre all'ipercubo H.B1.E1.R4 dove, in due regioni (le più piccole: Valle d'Aosta e Molise), i valori dell'indicatore sono compresi tra il 10% e il 15%.

L'accuratezza attesa degli ipercubi di diffusione europea presi in esame potrebbe essere ritenuta accettabile anche nel caso della frazione di campionamento del 10%. In questo caso, per un gran numero di regioni italiane i valori attesi degli indicatori di accuratezza risulterebbero inferiori alla soglia del 10%; alcune criticità potrebbero essere riscontrate per gli ipercubi più complessi relativamente alle regioni più piccole, per la possibilità di ottenere valori di UC superiori al 15%.

9.4 Conclusioni

I risultati degli esperimenti condotti per valutare i possibili riflessi dell'operazione campionaria sull'accuratezza di alcuni ipercubi richiesti da Eurostat a livello territoriale regionale (NUTS2) hanno messo in evidenza un basso impatto della nuova strategia per il Censimento della popolazione del 2011 sui dati finali.

Il principale risultato emerso è che gli ipercubi che presentano elevata granularità (più di 20mila celle di classificazione) possono essere stimati con basse percentuali di unità assegnate in celle critiche anche nel caso di un disegno campionario che impieghi la frazione più piccola tra quelle prese in esame (10%). I rischi di minore accuratezza sono sempre maggiori per le regioni più piccole.

Poiché per la diffusione italiana, oltre a quanto prodotto per Eurostat, si prevede la definizione di ulteriori tavole con dettagli informativi mediamente più ridotti, i risultati esposti nel capitolo portano a ritenere che i livelli di accuratezza saranno migliori per le tavole a livello regionale che saranno definite per il completamento del piano di diffusione dei risultati del Censimento 2011.

10. Accuratezza di tavole statistiche riferite al livello comunale e sub-comunale

10.1 Premessa

In questo capitolo proseguono le valutazioni empiriche dei possibili effetti del campionamento sull'accuratezza dei dati finali del censimento in relazione, stavolta, a tavole statistiche con un più elevato dettaglio territoriale (Borrelli *et al.*, 2009). A riguardo, sono state prese in esame, dal piano di diffusione italiano del 2001, alcune tavole statistiche (descritte nel paragrafo 7.3) aventi differenti contenuti informativi e riferibili ad ambiti comunali e sub-comunali.

Le tavole di diffusione scelte per le analisi (Tavole 7 e 8) sono state popolate con i dati del Censimento 2001, per i 40 comuni e per le relative 498 aree di censimento di centro abitato, già considerati nelle precedenti sperimentazioni campionarie (Borrelli *et al.*, 2011a). In particolare, le frequenze assolute riferite sia ai domini comunali che alle aree di censimento sono stati ottenuti come aggregazione dei relativi dati prodotti a livello di sezione di censimento; si precisa, inoltre, che le informazioni relative al dettaglio comunale sono state calcolate considerando tutte le sezioni appartenenti al comune stesso e non solo quelle costituenti le aree campionate.

Anche in questo contesto, così come in quello delle analisi regionali descritte nel capitolo 9, è stata ipotizzata l'adozione della strategia campionaria basata sull'impiego di questionari *short* e *long* secondo le versioni presentate nel paragrafo 4.2.

Per le analisi, gli indicatori FC, CC e UC (descritti nel paragrafo 6.3) sono stati calcolati, per ciascuna tavola e per ciascun dominio, sempre per un errore soglia del 12,5% (in termini di cv). Si fa presente che le soglie critiche FC sono state individuate sulla curva degli errori campionari nell'ipotesi accettabile di un valore della quota γ di popolazione rilevata a campione tramite *short/long form* pari a 1; infatti, se a livello di area di censimento tutta la popolazione è interessata dalle operazioni di campionamento ($\gamma=1$), a livello di comune la quota γ è molto vicina a 1 (rimangono escluse piccole porzioni di territorio extra-urbano o di nuova edificazione che non sono considerate nel disegno delle aree di censimento).

Poiché al diminuire del valore di γ si determinano abbassamenti della curva degli errori con vantaggi in termini di accuratezza attesa delle stime, la soglia critica FC delle stime riferite alle tavole comunali sarebbe inferiore a quella impiegata nell'esercizio (determinata nell'ipotesi $\gamma=1$) e i reali risultati risulterebbero certamente migliori di quelli descritti nei prossimi paragrafi.

Le valutazioni di sintesi sono presentate prima con riferimento al livello comunale e poi per il livello sub-comunale; in entrambi i casi le valutazioni sono state fatte anche per classe di ampiezza demografica[31] dei domini sottoposti a misurazione. Inoltre, le sintesi sono proposte per le frazioni di campionamento del 10%, del 20% e del 33% con lo scopo di valutare gli effetti indotti dall'abbassamento della curva degli errori dovuto all'aumento della frazione sondata.

[31] In questo ambito le valutazioni hanno riguardato anche i comuni con dimensione di popolazione compresa tra 10mila e 20mila; anche se il disegno progettuale non prevede il campionamento per i comuni con popolazione inferiore a 20mila unità, i risultati possono essere riferibili a domini sub-comunali di dimensioni più ampie di quelli sottoposti ad esercizio in questo studio (tra 5mila e 15mila).

10.2 Accuratezza attesa di tavole di dati per comune

Le Tavole presentate in questo paragrafo riportano in dettaglio i valori mediani degli indicatori riferiti ai 40 comuni presi in considerazione e calcolati sulle tavole statistiche scelte per le analisi a livello di comune, distintamente per il blocco 1 e il blocco 2 (cfr. paragrafo 7.3). In particolare, sono presentati i valori della frequenza critica FC per ciascuna delle frazioni sondate e le corrispondenti percentuali di casi critici, sia in termini di celle (CC) che di individui (UC), relativi a frequenze assolute (non nulle) inferiori alla soglia critica.

Come evidenziato nel capitolo 6, a parità di errore di campionamento ammissibile, al crescere della frazione sondata si riduce la frequenza critica con conseguente riduzione del numero di celle critiche e aumento dell'accuratezza attesa (diminuzione dei valori degli indicatori CC e UC).

Gli elevati valori dell'indicatore CC sono principalmente dovuti alla presenza di molte celle di incrocio con frequenze assolute piccole la cui stima comporta errori di campionamento grandi. Questo, però, non implica necessariamente un basso livello di accuratezza se in tali celle sono classificate poche unità. Tale livello di accuratezza è invece rappresentato in maniera più significativa dal valore dell'indicatore UC che esprime la percentuale di unità (individui) che ricadono in celle stimabili con un elevato grado di imprecisione.

Nella tavola 13 si osservano[32] valori degli indicatori CC e UC più grandi per le tavole statistiche ATT.ECO.1, ATT.ECO.2 e POS.PRO.1, quelle con il maggiore numero di incroci: in particolare, per l'indicatore UC nel caso della frazione sondata del 10% si evidenziano valori rispettivamente pari all'8,8%, al 21,5% e al 12,6%; valori più contenuti si registrano per le frazioni di campionamento più ampie.

Tavola 13 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme dei 40 comuni considerati nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 1 riferite al livello comunale** (Censimento della popolazione 2001)

FRAZIONE SONDATA	FC	Indicatori %	GRA.IST.1	CON.PRO.1	ATT.ECO.1	ATT.ECO.2	POS.PRO.1	PENDOL.1
f.s.=10%	500	CC	31,3	8,3	62,5	60,3	76,3	0,0
		UC	1,6	0,1	8,8	21,5	12,6	0,0
f.s.=20%	250	CC	18,8	8,3	53,2	36,8	66,7	0,0
		UC	0,7	0,1	4,3	4,9	5,9	0,0
f.s.=33%	100	CC	3,1	8,3	41,7	23,5	55,0	0,0
		UC	0,0	0,0	1,9	1,0	2,9	0,0

I risultati presentati nella tavola 14 sono relativi alle tavole statistiche che incrociano le variabili demografiche "stato civile" e "cittadinanza". Si osservano livelli di accuratezza attesa su valori di poco superiori a quelli esposti nella tavola 13 (in cui le variabili demografiche di incrocio sono "sesso" ed "età"); si registrano misure dell'indicatore UC di poco superiori al 10% solo nel caso della frazione di campionamento più piccola e per le tavole statistiche POS.PRO.2 (UC=10,8%) e CON.POS.1 (UC=11,6%).

[32] Esempio: nel caso della strategia di campionamento che prevede la frazione sondata del 10%, la frequenza critica FC individuata sulla curva degli errori campionari (misurati dal coefficiente di variazione percentuale), in relazione ad un errore massimo del 12,5%, è pari a 500. Con riferimento al valore mediano della distribuzione degli indicatori CC e UC calcolati per la tavola statistica GRA.IST.1 sui 40 comuni, per il 31,3% di celle sono attese stime inferiori a 500 che comporteranno errori campionari non inferiori al 12,5%; in tali celle critiche è atteso che ricadrà l'1,6% degli individui classificati nella tavola.

Tavola 14 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme dei 40 comuni considerati nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 2 riferite al livello comunale** (Censimento della popolazione 2001)

FRAZIONE SONDATA	FC	Indicatori %	GRA.IST.2	GRA.IST.3	CON.PRO.2	CON.PRO.3
f.s.=10%	500	CC	50,0	64,5	50,0	59,7
		UC	1,8	6,4	0,7	5,4
f.s.=20%	250	CC	44,5	50,0	41,7	46,5
		UC	0,6	3,0	0,6	1,9
f.s.=33%	100	CC	33,3	35,9	36,7	38,0
		UC	0,2	0,9	0,4	0,9

Tavola 14 segue

FRAZIONE SONDATA	FC	Indicatori %	ATT.ECO.3	ATT.ECO.4	POS.PRO.2	CON.POS.1	CON.POS.2
f.s.=10%	500	CC	50,0	66,7	80,0	65,1	55,8
		UC	1,9	5,2	10,8	11,6	4,7
f.s.=20%	250	CC	50,0	56,7	68,0	53,4	44,2
		UC	0,6	3,3	5,2	4,5	1,9
f.s.=33%	100	CC	33,3	40,0	56,0	37,5	34,6
		UC	0,2	0,8	2,1	1,9	1,0

Nelle Tavole 15 e 16 sono riportati i risultati dell'analisi condotta con riferimento alla dimensione demografica dei comuni, suddivisi per l'ampiezza della popolazione in 4 classi.[33]

Nella tavola 15, si notano valori più elevati dell'indicatore UC (superiori al 10%) per le stesse tavole statistiche (ATT.ECO.1, ATT.ECO.2 e POS.PRO.1) che hanno mostrato in precedenza le maggiori criticità; dall'analisi per dimensione comunale, i casi più problematici si osservano per i comuni inferiori ai 50mila abitanti e per le strategie campionarie con le frazioni sondate del 10% e del 20%.

Nel caso della strategia che campiona al 33%, le misure dell'indicatore UC si attestano su valori sempre inferiori al 10% per tutte le tavole, tranne il caso della tavola ATT.ECO.2 che, solo per la classe dei comuni più piccoli, mostra un valore pari al 10,7%.

[33] Le classi di popolazione prese in considerazione sono: 10mila-20mila; 20mila-50mila; 50mila-150mila; ≥ 150mila. In ciascuna classe sono classificati 10 comuni.

Tavola 15 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme dei 40 comuni considerati nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 1 riferite al livello comunale** (Censimento della popolazione 2001)**. Comuni classificati per dimensione demografica**

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	GRA.IST.1	CON.PRO.1	ATT.ECO.1	ATT.ECO.2	POS.PRO.1	PENDOL.1
f.s.=10%	500	10.000 ├ 20.000	CC	43,8	41,7	86,3	93,9	87,8	0,0
			UC	8,8	8,0	34,6	66,8	28,6	0,0
		20.000 ├ 50.000	CC	31,3	20,8	74,5	80,0	84,5	0,0
			UC	1,9	2,1	16,7	40,6	20,6	0,0
		50.000 ├ 150.000	CC	21,9	8,3	55,3	39,7	66,7	0,0
			UC	0,9	0,1	3,3	5,6	6,2	0,0
		≥150.000	CC	9,4	8,3	47,9	23,5	53,3	0,0
			UC	0,2	0,1	1,7	0,9	2,2	0,0
f.s.=20%	250	10.000 ├ 20.000	CC	31,3	16,7	71,7	80,0	79,3	0,0
			UC	2,5	1,5	18,0	35,0	18,9	0,0
		20.000 ├ 50.000	CC	31,3	8,3	60,9	60,0	74,2	0,0
			UC	1,6	0,1	6,6	19,2	10,5	0,0
		50.000 ├ 150.000	CC	6,3	8,3	47,8	26,5	60,0	0,0
			UC	0,3	0,1	2,8	1,3	3,6	0,0
		≥150.000	CC	6,3	8,3	31,3	22,1	40,0	0,0
			UC	0,1	0,0	0,7	0,6	1,0	0,0
f.s.=33%	100	10.000 ├ 20.000	CC	15,6	8,3	52,2	52,3	67,2	0,0
			UC	0,5	0,1	5,8	10,7	6,9	0,0
		20.000 ├ 50.000	CC	18,8	8,3	52,2	33,3	62,1	0,0
			UC	0,4	0,1	3,2	3,1	3,8	0,0
		50.000 ├ 150.000	CC	0,0	8,3	33,3	20,6	40,0	0,0
			UC	0,0	0,0	0,9	0,6	1,1	0,0
		≥150.000	CC	0,0	0,0	18,8	13,2	28,3	0,0
			UC	0,0	0,0	0,2	0,2	0,4	0,0

La tavola 16, riferita al blocco 2 delle tavole statistiche esaminate, evidenzia risultati analoghi e in alcuni casi migliori di quelli mostrati nella tavola 15.

L'impiego della frazione sondata del 10% conduce ad alcune situazioni critiche per i comuni delle due classi inferiori a 50mila abitanti. In particolare, per la tavola CON.POS.1 si hanno i valori più elevati dell'indicatore UC, pari al 19,0% per la classe dei comuni tra 20mila e 50mila e al 33,7% per quella dei comuni tra 10mila e 20mila.

Con la frazione campionaria del 20% le verifiche conducono a valori critici di UC superiori al 10% solo per le tavole POS.PRO.2 e CON.POS.1 ed unicamente per la classe dei comuni tra 10mila e 20mila abitanti (l'indicatore UC vale rispettivamente 13,2% e 18,2%).

Infine, nel caso della strategia campionaria che impiega la frazione sondata del 33% il valore dell'indicatore UC è inferiore al 10% per tutte le tavole considerate e per qualunque classe dimensionale dei comuni.

Tavola 16 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme dei 40 comuni considerati nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 2 riferite al livello comunale** (Censimento della popolazione 2001)**. Comuni classificati per dimensione demografica**

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	GRA.IST.2	GRA.IST.3	CON.PRO.2	CON.PRO.3
f.s.=10%	500	10.000 ├ 20.000	CC	60,8	77,4	58,3	77,8
			UC	4,4	16,3	4,2	14,2
		20.000 ├ 50.000	CC	60,0	70,9	50,0	65,0
			UC	2,7	10,6	1,7	8,4
		50.000 ├ 150.000	CC	48,3	51,3	41,7	48,3
			UC	0,8	2,9	0,9	2,5
		≥150.000	CC	25,0	37,5	20,8	33,9
			UC	0,3	0,7	0,2	0,8
f.s.=20%	250	10.000 ├ 20.000	CC	57,1	66,7	50,0	66,1
			UC	2,4	8,8	1,5	7,8
		20.000 ├ 50.000	CC	60,0	61,1	50,0	52,7
			UC	2,0	5,3	0,8	3,3
		50.000 ├ 150.000	CC	39,0	38,8	37,5	39,7
			UC	0,5	1,0	0,5	1,3
		≥150.000	CC	15,6	23,8	8,3	21,7
			UC	0,1	0,2	0,1	0,3
f.s.=33%	100	10.000 ├ 20.000	CC	43,3	57,1	41,7	48,2
			UC	1,2	3,9	0,8	2,3
		20.000 ├ 50.000	CC	45,2	43,5	41,7	41,1
			UC	0,7	1,8	0,6	1,4
		50.000 ├ 150.000	CC	21,9	23,8	16,7	22,0
			UC	0,1	0,2	0,1	0,3
		≥150.000	CC	12,5	20,0	0,0	11,8
			UC	0,0	0,1	0,0	0,0

Tavola 16 segue

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	ATT.ECO.3	ATT.ECO.4	POS.PRO.2	CON.POS.1	CON.POS.2
f.s.=10%	500	10.000 ├ 20.000	CC	66,7	73,3	87,0	81,0	78,5
			UC	6,6	11,3	17,1	33,7	19,4
		20.000 ├ 50.000	CC	66,7	73,3	83,3	73,3	65,4
			UC	3,5	9,6	14,4	19,0	11,1
		50.000 ├ 150.000	CC	33,3	53,3	68,0	52,4	46,1
			UC	0,4	3,6	5,6	4,4	2,5
		≥150.000	CC	16,7	36,7	54,0	39,4	23,1
			UC	0,2	0,8	2,3	2,0	0,7
f.s.=20%	250	10.000 ├ 20.000	CC	50,0	69,1	84,0	70,1	61,5
			UC	3,3	7,7	13,2	18,2	7,8
		20.000 ├ 50.000	CC	55,0	66,7	77,8	61,2	50,0
			UC	2,8	4,9	9,3	9,0	3,2
		50.000 ├ 150.000	CC	16,7	36,7	58,0	45,2	38,5
			UC	0,2	0,9	3,0	2,6	1,3
		≥150.000	CC	16,7	23,3	28,0	25,0	17,3
			UC	0,1	0,2	0,3	0,6	0,3
f.s.=33%	100	10.000 ├ 20.000	CC	50,0	65,5	75,0	54,1	46,1
			UC	0,9	5,0	6,6	6,0	2,6
		20.000 ├ 50.000	CC	33,3	50,0	62,3	48,0	42,3
			UC	0,7	2,3	3,5	2,9	1,6
		50.000 ├ 150.000	CC	16,7	26,7	32,0	30,8	19,2
			UC	0,1	0,4	0,5	0,8	0,2
		≥150.000	CC	8,3	20,0	24,0	17,3	9,6
			UC	0,0	0,1	0,2	0,2	0,1

10.3 Accuratezza attesa di tavole di dati per area di censimento di centro abitato

In questo contesto si presentano i risultati delle analisi svolte in modo analogo a quanto illustrato nel precedente paragrafo, ma riferite al livello territoriale sub-comunale coincidente con quello delle aree di censimento di centro abitato. Anche in questa situazione i risultati sono esposti in modalità separata per le tavole statistiche relative al blocco 1 e al blocco 2 (cfr. paragrafo 7.3).

Il primo risultato che emerge è un aumento dei valori degli indicatori CC e UC riferiti alle aree di censimento, rispetto alle misure precedentemente calcolate per i domini comunali, a svantaggio dell'accuratezza attesa delle tavole statistiche producibili a quel livello territoriale sub-comunale.

Come si nota nella tavola 17, gli indicatori di accuratezza assumono[34] i valori più elevati, così come già evidenziato nell'analisi comunale, per le tavole statistiche ATT.ECO.1, ATT.ECO.2 e

[34] Esempio: nel caso della strategia di campionamento che considera la frazione sondata del 10%, la frequenza critica FC determinata sulla curva degli errori campionari (misurati dal coefficiente di variazione percentuale), per un valore di cv massimo accettabile del 12,5%, è pari a 500. Con riferimento al valore mediano della distribuzione degli indicatori UC e CC calcolati in relazione alla tavola statistica GRA.IST.1 sulle 498 aree di censimento di centro abitato, per il 50,0% di celle sono attese stime inferiori a 500 che implicheranno errori non inferiori al 12,5%; in tali celle, ritenute critiche, sarà classificato il 10,5% degli individui totali della tavola.

POS.PRO.1. Per tali tavole l'indicatore UC si attesta su valori significativamente alti anche con la frazione sondata del 20%. Nel caso della frazione campionaria del 33% si osserva un valore di UC superiore al 10% solo in corrispondenza della tavola ATT.ECO.2 (UC=16,1%).

Tavola 17 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme delle 498 aree di censimento considerate nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 1 riferite al livello sub-comunale** (Censimento della popolazione 2001)

FRAZIONE SONDATA	FC	Indicatori %	GRA.IST.1	CON.PRO.1	ATT.ECO.1	ATT.ECO.2	POS.PRO.1	PENDOL.1
f.s.=10%	500	CC	50,0	58,3	90,5	100,0	91,7	50,0
		UC	10,5	13,6	41,2	100,0	36,5	7,0
f.s.=20%	250	CC	37,5	25,0	81,4	87,1	85,2	25,0
		UC	4,1	4,4	24,8	58,4	22,2	2,1
f.s.=33%	100	CC	31,3	8,3	61,9	55,0	72,4	0,0
		UC	1,3	0,1	7,0	16,1	9,6	0,0

Nella tavola 18 sono riportati i valori degli indicatori CC e UC calcolati sulle tavole statistiche che prevedono incroci con le variabili "stato civile" e "cittadinanza"; anche per questi casi, analogamente a quanto emerso dalle analisi riferite al livello comunale, sono attesi livelli di accuratezza superiori rispetto a quelli prevedibili per le tavole che incrociano le variabili "sesso" ed "età".

Nel caso della frazione sondata del 33% i valori dell'indicatore UC risultano quasi sempre inferiori al 10%; fa eccezione la tavola CON.POS.1 con UC=10,3%. Per le frazioni di campionamento più basse l'accuratezza attesa delle tavole peggiora in modo evidente: per la frazione del 20% le misure di UC sono, in molte situazioni, maggiori del 10% e superano addirittura la soglia del 25% per alcune tavole nel caso della frazione sondata del 10%.

Tavola 18 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme delle 498 aree di censimento considerate nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 2 riferite al livello sub-comunale** (Censimento della popolazione 2001)

FRAZIONE SONDATA	FC	Indicatori %	GRA.IST.2	GRA.IST.3	CON.PRO.2	CON.PRO.3
f.s.=10%	500	CC	66,7	80,6	58,3	78,6
		UC	7,6	22,4	7,2	21,5
f.s.=20%	250	CC	62,0	72,2	50,0	69,0
		UC	4,2	12,3	2,0	12,2
f.s.=33%	100	CC	50,0	60,0	41,7	51,8
		UC	1,8	5,9	1,1	4,1

Tavola 18 segue

FRAZIONE SONDATA	FC	Indicatori %	ATT.ECO.3	ATT.ECO.4	POS.PRO.2	CON.POS.1	CON.POS.2
f.s.=10%	500	CC	66,7	84,6	91,3	89,8	80,8
		UC	5,9	21,6	29,5	58,3	27,5
f.s.=20%	250	CC	66,7	73,3	85,7	75,0	68,0
		UC	3,9	10,7	18,6	23,7	14,3
f.s.=33%	100	CC	50,0	66,7	75,0	61,2	50,0
		UC	2,4	5,7	9,1	10,3	4,3

L'analisi dei risultati prosegue tramite la classificazione delle aree di censimento in base alla loro dimensione demografica.[35] La tavola 19 riporta l'esito delle verifiche relative al primo blocco di tavole statistiche.

Si evidenziano livelli di accuratezza accettabili solo con la frazione sondata del 33% e per le aree di censimento con più di 12mila abitanti; il valore più elevato dell'indicatore UC (10,4%) si osserva per la tavola ATT.ECO.2. Per i domini sub-comunali con popolazione inferiore a 12mila unità sono attesi livelli di accuratezza più bassi e in modo più evidente sempre per la tavola statistica ATT.ECO.2, quella che presenta il maggior numero di incroci tra le tavole del blocco 1.

Per le frazioni campionarie più piccole l'indicatore UC mostra valori ovunque peggiori; maggiormente critici appaiono quelli relativi alle tavole ATT.ECO.1, ATT.ECO.2 e POS.PRO.1. Le problematicità risultano crescenti al diminuire della dimensione delle aree di censimento.

Tavola 19 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme delle 498 aree di censimento considerate nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 1 riferite al livello comunale** (Censimento della popolazione 2001)**. Aree classificate per dimensione demografica**

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	GRA.IST.1	CON.PRO.1	ATT.ECO.1	ATT.ECO.2	POS.PRO.1	PENDOL.1
f.s.=10%	500	<10.000	CC	62,5	58,3	90,9	100,0	92,3	50,0
			UC	16,7	16,4	46,1	100,0	42,7	10,8
		10.000 ├ 12.000	CC	50,0	58,3	90,5	97,1	91,7	50,0
			UC	10,2	13,8	40,5	89,6	35,0	7,3
		≥12.000	CC	50,0	50,0	87,0	96,9	89,3	25,0
			UC	8,3	11,8	37,0	85,4	32,4	5,4
f.s.=20%	250	<10.000	CC	46,7	41,7	85,7	93,8	88,0	25,0
			UC	6,3	8,9	32,5	74,9	28,3	4,4
		10.000 ├ 12.000	CC	37,5	33,3	81,0	84,4	85,2	0,0
			UC	3,8	4,6	24,8	55,4	22,1	0,0
		≥12.000	CC	37,5	16,7	73,9	78,1	82,1	0,0
			UC	2,6	2,3	17,3	41,6	17,8	0,0
f.s.=33%	100	<10.000	CC	31,3	8,3	66,7	63,6	76,9	0,0
			UC	1,7	0,2	9,1	25,6	12,9	0,0
		10.000 ├ 12.000	CC	31,3	8,3	60,9	54,5	74,1	0,0
			UC	1,4	0,1	6,7	14,8	9,1	0,0
		≥12.000	CC	31,3	8,3	57,1	47,1	69,2	0,0
			UC	1,1	0,1	4,6	10,4	7,2	0,0

Infine, nella tavola 20 sono presentate le misure di accuratezza relative al secondo blocco di tavole statistiche, sempre distinte per classe di ampiezza demografica delle aree di censimento. I valori degli indicatori conducono a conclusioni simili a quelle emerse dall'analisi dei risultati riportati nella tavola 19: per la frazione di campionamento del 33% si osservano valori dell'indicatore UC inferiori al 10% per le aree di censimento più grandi; per le altre due classi dimensionali i valori sono al più di poco superiori al 15% (più precisamente, UC=15,4% per la classe delle aree più piccole con riferimento alla tavola statistica CON.POS.1).

Nel caso della frazione sondata del 20%, indipendentemente dalla dimensione delle aree di censimento, l'indicatore UC presenta valori diffusamente più alti ma sempre inferiori al 25%, ad ecce-

[35] Sono state considerate tre classi di ampiezza di popolazione: 5mila-10mila (184 aree); 10mila-12mila (95 aree); 12mila-15mila (219 aree).

zione della tavola CON.POS.1 riferita ai domini inferiori a 10mila unità (UC=36,3%). I problemi di bassa accuratezza si fanno più evidenti nel caso della frazione sondata più piccola, anche per le aree più grandi che, per costruzione, dovrebbero garantire stime più affidabili.

Tavola 20 - Frequenza critica FC e indicatori percentuali di accuratezza CC e UC (valori mediani determinati sull'insieme delle 498 aree di censimento considerate nelle sperimentazioni campionarie) **calcolati per un cv soglia del 12,5%, sulle tavole statistiche del blocco 2 riferite al livello comunale** (Censimento della popolazione 2001)**. Aree classificate per dimensione demografica**

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	GRA.IST.2	GRA.IST.3	CON.PRO.2	CON.PRO.3
f.s.=10%	500	<10.000	CC	71,4	83,6	66,7	82,8
			UC	10,1	29,7	11,1	29,5
		10.000 ├ 12.000	CC	66,7	81,1	58,3	78,6
			UC	6,1	23,3	5,9	21,5
		≥12.000	CC	66,7	77,8	50,0	75,9
			UC	6,3	18,5	5,1	17,9
f.s.=20%	250	<10.000	CC	64,3	75,3	50,0	74,1
			UC	5,5	15,1	4,1	15,4
		10.000 ├ 12.000	CC	62,5	72,2	50,0	70,4
			UC	4,0	12,1	2,3	12,1
		≥12.000	CC	60,0	71,1	50,0	65,5
			UC	3,8	11,0	1,7	10,4
f.s.=33%	100	<10.000	CC	53,3	64,9	41,7	57,1
			UC	2,7	8,0	1,3	6,1
		10.000 ├ 12.000	CC	50,0	60,5	41,7	51,9
			UC	1,9	6,2	1,0	4,0
		≥12.000	CC	43,7	56,4	41,0	48,3
			UC	1,4	4,6	1,0	3,3

Tavola 20 segue

FRAZIONE SONDATA	FC	Dimensione demografica	Indicatori %	ATT.ECO.3	ATT.ECO.4	POS.PRO.2	CON.POS.1	CON.POS.2
f.s.=10%	500	<10.000	CC	66,7	85,7	91,7	95,9	80,8
			UC	9,4	30,1	33,7	81,6	30,2
		10.000 ├ 12.000	CC	66,7	84,6	91,3	89,6	80,8
			UC	4,9	20,2	29,6	55,2	27,8
		≥12.000	CC	66,7	80,0	90,5	84,0	80,8
			UC	5,2	17,4	27,2	41,1	25,7
f.s.=20%	250	<10.000	CC	66,7	76,9	87,5	80,6	76,0
			UC	4,4	11,9	23,2	36,3	21,3
		10.000 ├ 12.000	CC	66,7	73,3	85,0	76,0	68,0
			UC	4,1	10,7	17,9	22,8	13,7
		≥12.000	CC	66,7	73,3	84,0	74,0	61,5
			UC	3,7	10,3	16,6	20,0	10,0
f.s.=33%	100	<10.000	CC	60,0	69,2	79,0	68,1	56,0
			UC	3,0	7,7	12,5	15,4	6,5
		10.000 ├ 12.000	CC	60,0	64,3	73,7	61,2	50,0
			UC	2,8	5,5	8,7	10,5	4,1
		≥12.000	CC	50,0	60,0	68,2	56,0	48,0
			UC	2,0	4,5	7,6	7,9	3,5

10.4 Conclusioni

In questo capitolo sono state fatte alcune valutazioni sull'accuratezza attesa di tavole statistiche di risultati censuari riferiti a domini comunali e sub-comunali.

Il primo risultato emerso in maniera evidente è che minore è il dettaglio territoriale preso in esame migliore è l'accuratezza delle tavole statistiche; infatti, i valori degli indicatori considerati a livello comunale sono sempre inferiori a quelli definiti a livello di area di censimento; questo risultato è chiaramente indotto dal fatto che il numero di unità classificate nelle tavole definite a livello di area è maggiormente ridotto rispetto a quelle dell'intero comune.

La seconda conclusione a cui si giunge è che i livelli di accuratezza complessiva tendono a migliorare all'aumentare della frazione di campionamento. Ad esempio, considerando la tavola statistica ATT.ECO.2 a livello di comune, si nota che l'indicatore UC subisce una forte riduzione, passando dal valore del 21,5% con la frazione campionaria del 10%, al valore dell'1,0% nel caso della frazione del 33% (Tavola 13).

L'analisi mostra in modo evidente anche il legame tra i livelli di accuratezza delle tavole e la dimensione demografica del dominio a cui si riferisce (il comune o l'area di censimento); infatti, maggiore è la dimensione demografica migliori sono i valori attesi degli indicatori di accuratezza. Prendendo ad esempio ancora la tavola ATT.ECO.2 per diverse classi di ampiezza demografica del comune e per la frazione sondata del 33%, il valore percentuale di UC passa dal 10,7% per i comuni di dimensione compresa tra 10mila e 20mila allo 0,2% per i comuni con più di 150mila abitanti (Tavola 15); in particolare, è attesa un'elevata accuratezza delle tavole (valori di UC inferiori al 10%) per tutte le classi dimensionali dei comuni nel caso della strategia campionaria che adotta la frazione sondata del 33%. Invece, nei casi di frazione campionaria più piccola, questo risultato vale solo per i comuni con popolazione superiore ai 50mila abitanti.

A livello di area di censimento e per la frazione sondata maggiore, l'indicatore UC assume valori minori del 10% solo nel caso dei domini con popolazione superiore ai 12mila abitanti. Questo risultato è stato molto utile per suggerire la dimensione per il disegno delle aree di censimento ai fini della determinazione della loro eleggibilità al campionamento (Bianchi *et al.*, 2010): è stato indicato di costruire, ove possibile, aree con popolazione compresa tra 13mila e 18mila unità.

Si fa inoltre osservare che, indipendentemente dalla scelta della frazione sondata, la riduzione del dettaglio informativo causata dalla diminuzione del numero di modalità di incrocio, comporta un aumento dell'accuratezza della tavola a vantaggio della qualità complessiva dell'informazione prodotta. Ad esempio, sia a livello comunale che a livello di area di censimento, le tavole di diffusione che potrebbero avere livelli di qualità critica sono quelle che presentano i maggiori dettagli informativi, in termini di numero di incroci (ATT.ECO.1, ATT.ECO.2, POS.PRO.1, POS.PRO.2 e CON.POS.1); al contrario per la tavola PENDOL, osservando i valori più bassi degli indicatori, si giunge a valutazioni molto positive.

Il legame tra il dettaglio informativo e il livello di accuratezza è molto più evidente quando il confronto viene effettuato su tavole concernenti la stessa tematica (ad esempio, solo le tavole relative al grado di istruzione: GRA.IST.1, GRA.IST.2, GRA.IST.3). Se il confronto invece è effettuato tra due tematiche diverse (ad esempio: grado di istruzione e attività economica) questo tipo di legame non è sempre confermato; questo aspetto è giustificato dal fatto che le tavole statistiche si riferiscono a sotto-universi di popolazione tra loro differenti (rispettivamente la "popolazione residente di 6 anni e più" e la "popolazione residente occupata di 15 anni più").

11. Riflessi dell'ampliamento del questionario in forma ridotta

Come anticipato nel paragrafo 4.3, in occasione della rilevazione censuaria del 2011 si è scelto di impiegare, oltre al questionario completo, la versione *medium* come forma ridotta, rinunciando alla precedente proposta di questionario *short*, così da raccogliere in maniera esaustiva un insieme più esteso di dati.

L'adozione di una versione più ampia del questionario ridotto è stata fortemente richiesta dai principali soggetti istituzionali per l'opportunità di disporre di informazioni a livelli di maggiore dettaglio e per ambiti territoriali molto fini. La decisione finale è comunque stata presa anche con il conforto dei risultati dell'indagine pilota del 2009, in particolare per l'omogeneità dei tassi di risposta spontanei osservati per i diversi questionari testati nella rilevazione.

Il questionario ridotto di tipo *medium*, unitamente al modello completo, permette di rilevare e disporre di un più ricco insieme di informazioni già a partire dal livello di sezione di censimento; questa soluzione offrirà la possibilità di diffondere, con riferimento alle unità territoriali più fini, oltre ai dati di natura strettamente demografica, anche alcune informazioni di carattere socio-economico, seppur con classificazioni meno dettagliate.

In generale, dal lato della produzione statistica, l'aumento del numero di variabili rilevate su tutta la popolazione comporta la riduzione dell'insieme delle variabili osservate solo su campioni di famiglie e, quindi, dell'insieme di tavole statistiche di diffusione che saranno interessate dalla procedura di stima.

Dal punto di vista inferenziale, invece, i vantaggi attesi con l'adozione del questionario ridotto più ampio riguardano due principali aspetti:

1. la diminuzione della variabilità campionaria delle stime;
2. il miglioramento dell'accuratezza delle tavole statistiche prodotte.

Riguardo il primo punto, la versione *medium* del questionario ridotto permette di arricchire l'insieme delle variabili esaustive che potrebbero essere utilmente considerate come benchmark (informazione ausiliaria) nel processo di stima che prevede il calcolo dei pesi di riporto all'universo secondo la procedura di riponderazione vincolata (calibrazione).

Con riferimento a ciò, è stato condotto uno studio sperimentale su dati del Censimento del 2001 (Borrelli *et al.*, 2010) finalizzato a misurare il guadagno di efficienza delle stime campionarie ottenibile con la strategia tramite *medium/long form* rispetto a quella basata su *short/long form*. Per la sperimentazione è stato considerato il disegno casuale semplice di famiglie, esaminando differenti frazioni di campionamento (10%, 20% e 33%) e l'impiego dello stimatore di ponderazione vincolata. Per una corretta confrontabilità degli errori campionari attesi con le strategie *short/long* e *medium/long*, è stato preso in esame uno stesso insieme di variabili da sottoporre a stima.

Il primo vantaggio derivante dalla strategia *medium/long form* è quello di poter disporre di un insieme di vincoli di calibrazione più ampio rispetto a quello praticabile con la strategia *short/long*. Gli insiemi di vincoli individuati[36] sono stati più numerosi nel caso delle frazioni di campionamento più elevate; con la frazione più piccola (10%) la ridotta numerosità campionaria ha spesso posto problemi per la convergenza dell'algoritmo e, quindi, per la determinazione dei pesi di riporto all'universo secondo una soluzione che producesse risultati coerenti con i vincoli imposti (condizione non trascurabile per gli scopi inferenziali previsti).

Dall'analisi dei valori delle distribuzioni degli errori campionari attesi (espressi tramite il cv), si è potuto desumere che, per qualunque frazione di campionamento impiegata, la strategia basata su *medium/long* comporta, per ogni classe di valori da stimare, un errore mediamente inferiore a quello atteso nel caso della strategia *short/long form*.

Per quantificare il guadagno di efficienza ottenuto con la strategia *medium/long*, è stata misurata la riduzione percentuale del valore atteso di cv;[37] tale operazione ha messo in evidenza che, a parità di tasso di campionamento, il guadagno di efficienza aumenta al crescere del valore della frequenza assoluta oggetto di stima fino a un massimo del 27% (guadagno osservato per le frequenze più

[36] L'identificazione degli insiemi dei vincoli di calibrazione (definiti a livello di area di censimento) è stata basata sulla verifica della convergenza dell'algoritmo di calibrazione. In particolare (Borrelli *et al.*, 2010), nella strategia *short/long form* i vincoli di calibrazione impiegati sono stati 42 con la frazione sondata del 10% e 44 con le frazioni più grandi. Invece, per la strategia *medium/long form*, dal momento che era possibile impiegare insiemi più numerosi di variabili di calibrazione, sono stati considerati 56 vincoli con la frazione sondata del 10%, 58 con quella del 20% e 62 con la frazione del 33%. Insiemi più ampi, nei differenti casi esaminati, hanno evidenziato problemi di convergenza, pertanto non sono stati presi in considerazione.

[37] È stata calcolata la seguente misura percentuale: $rid(cv) = [(cv_{sh} - cv_{med})/cv_{sh}] \times 100$

grandi) per tutte le frazioni sondate considerate (Borrelli *et al.*, 2010). In particolare, i guadagni maggiori si osservano con la frazione di campionamento più piccola (10%) per le frequenze assolute da stimare più grandi (mediamente superiori a 100); invece per la stima di valori molto piccoli i guadagni più consistenti si rilevano nel caso della strategia che impiega la frazione di campionamento maggiore.

Quest'ultimo risultato evidenzia il notevole vantaggio che si ottiene dall'adozione della strategia *medium/long form* per la stima delle frequenze più piccole (quelle più a rischio di comportare errori relativi molto elevati) proprio con il disegno di campionamento basato sulla frazione sondata del 33%, scelto per l'operazione campionaria del Censimento del 2011.

Dal successivo confronto tra le distribuzioni empiriche delle misure degli errori campionari (cv) relative alle due strategie, è emerso che la disponibilità di un numero maggiore di variabili ausiliarie impiegabili nel processo di stima si traduce in una diminuzione dell'errore di campionamento e in una traslazione della distribuzione degli errori campionari verso livelli mediamente più bassi con l'ulteriore effetto di provocare un maggiore addensamento sulla coda di sinistra e una riduzione della concentrazione sulla coda di destra.

Questo risultato è più evidente in alcuni casi, come è possibile vedere nella Figura 5 relativa alla stima di frequenze assolute grandi.

Per altre situazioni, invece, non si hanno vantaggi della stessa evidenza dall'impiego di un numero maggiore di variabili di tipo benchmark nel procedimento di stima; a riguardo, sono stati rilevati guadagni assoluti minimi con la strategia *medium/long* in riferimento alla stima dei valori mediamente più piccoli (Figura 6).

Figura 5 - Confronto delle distribuzioni degli errori di campionamento attesi (misurati dal cv)**, riferite alle strategie short/long e medium/long, per frequenze appartenenti alla classe di valori compresi tra 1.500 e 2.000. Caso del disegno casuale semplice** (frazione sondata del 33%)

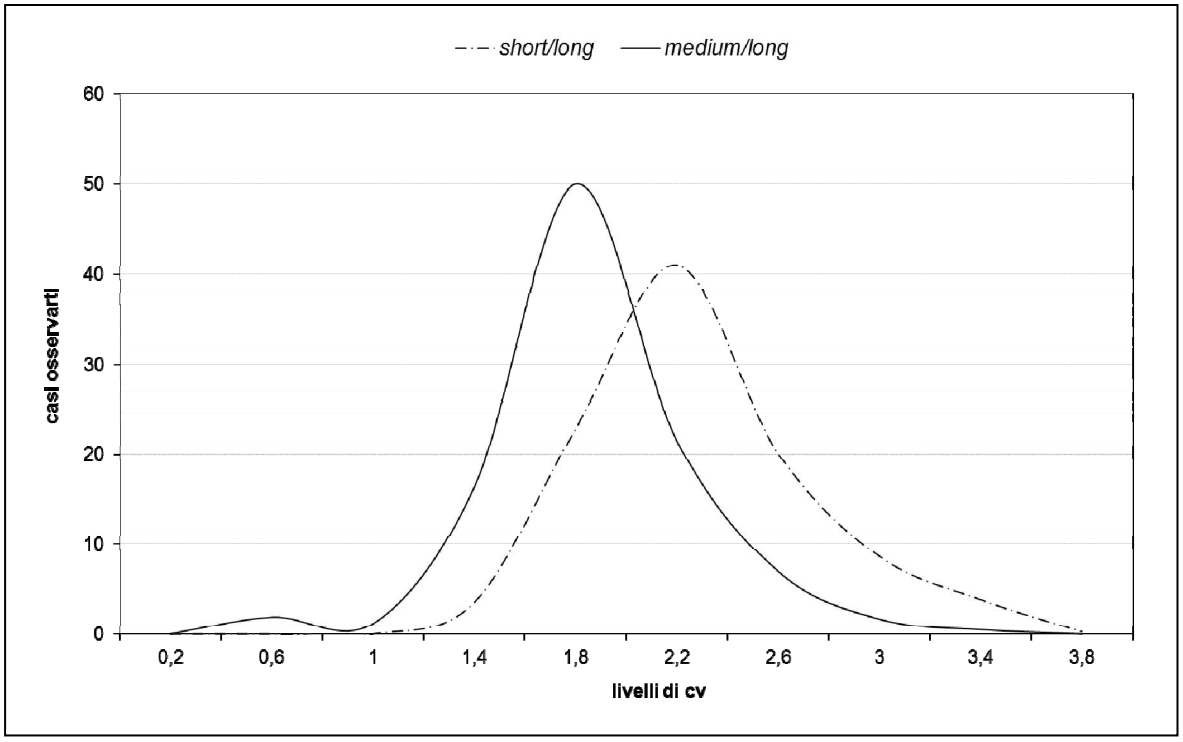
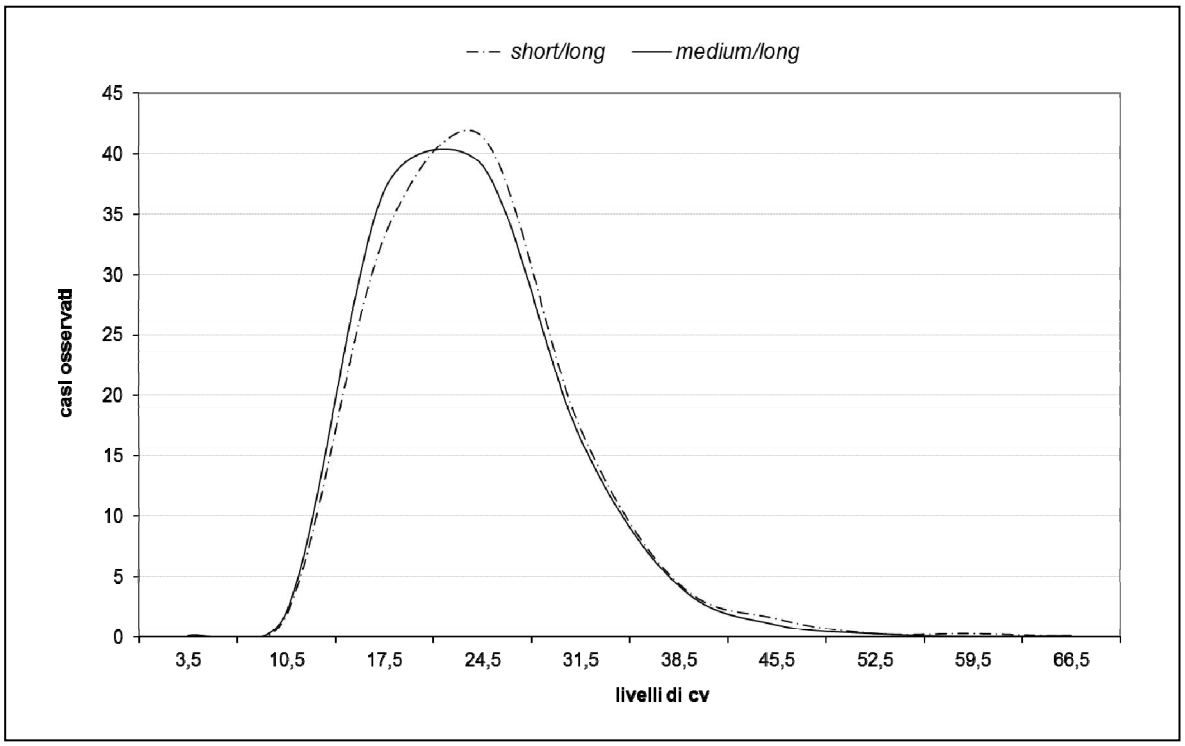


Figura 6 - Confronto delle distribuzioni degli errori di campionamento attesi (misurati dal cv)**, riferite alle strategie short/long e medium/long, per frequenze appartenenti alla classe di valori compresi tra 25 e 50. Caso del disegno casuale semplice** (frazione sondata del 33%)



In generale, il maggior dettaglio informativo derivante dall'utilizzo del questionario ridotto in forma ampliata comporta un miglioramento delle stime per tutte le frazioni di campionamento e, in particolare, tale guadagno aumenta al crescere del valore della frequenza da stimare.

Passando al secondo punto, per quanto finora detto, si può facilmente derivare che la riduzione della variabilità campionaria, indotta dall'impiego della strategia *medium/long form*, porta a far abbassare la curva degli errori di campionamento in maniera più marcata nei casi di stime riferite alle frequenze più grandi e un po' meno per le classi di valori più piccoli.

In conseguenza di questo risultato, sono attesi livelli di accuratezza migliori per la stima delle tavole statistiche che incrociano le variabili rilevate a campione. Infatti, seguendo l'impostazione metodologica illustrata nel capitolo 6, fissato un livello di errore campionario ammissibile, con la strategia *medium/long* si osserveranno valori ridotti della frequenza critica FC sotto la quale sono attesi errori campionari maggiori di quello critico accettabile, con vantaggi sull'accuratezza (misurabile tramite gli indicatori CC e UC) delle tavole statistiche che saranno prodotte con dati stimati.

12. Considerazioni conclusive

La progettazione del 15° Censimento della popolazione e delle abitazioni ha rivolto particolare attenzione a soluzioni innovative sia sul versante delle tecnologie informatiche che su quello delle metodologie statistiche, con lo scopo di migliorare l'efficienza delle operazioni sul campo e ridurre il disturbo statistico alle famiglie.

Tra gli elementi di novità rispetto alle passate rilevazioni, il Censimento del 2011 si caratterizza per l'uso di dati di fonte amministrativa e per l'impiego delle tecniche di campionamento. La strategia campionaria prevede, nei comuni più grandi, la rilevazione tramite il questionario in forma completa solo su campioni di famiglie e, sul resto della popolazione residente nel comune, l'impiego di un questionario in forma ridotta; invece, nei comuni non interessati dall'operazione campionaria, la raccolta dei dati avviene unicamente con il questionario completo.

La scelta di adottare la tecnica campionaria nel contesto del censimento causa inevitabili riflessi sulla qualità dei dati finali. Pur nella convinzione che un procedimento campionario può portare a risultati migliori di quelli ottenibili con una rilevazione totale, è stato ritenuto fondamentale valutare la qualità attesa dei dati, sia in termini di efficienza delle stime producibili che di accuratezza dei risultati rappresentati tramite le tavole statistiche di diffusione.

Dopo uno studio finalizzato a misurare, per differenti disegni campionari, i livelli di efficienza attesa delle stime relative alle variabili di censimento oggetto di rilevazione solo su campioni di famiglie residenti, si è proceduto ad un insieme di valutazioni quantitative dell'accuratezza attesa dell'informazione di censimento producibile e diffondibile per differenti livelli territoriali. A riguardo, è stato studiato l'impatto della strategia campionaria proposta per il contesto censuario italiano, su tavole statistiche riferite sia a livello regionale, secondo quanto richiesto da Eurostat, che a livello comunale e sub-comunale, di maggiore interesse per gli utenti nazionali, per il fabbisogno di dati con elevato dettaglio territoriale.

Gli esercizi sperimentali sono stati condotti con l'impiego dei dati del Censimento del 2001. Per la determinazione dei livelli di accuratezza è stato seguito un approccio che ha suggerito la definizione di due specifici indicatori di accuratezza, utili a fornire indicazioni sulla qualità attesa dei risultati censuari contenuti nelle tavole statistiche interessate dal procedimento di stima.

Le analisi hanno portato a ritenere trascurabile l'impatto della strategia di campionamento sull'accuratezza prevedibile degli ipercubi richiesti da Eurostat a livello regionale, anche nel caso in cui questi si configurano con molte celle di classificazione. Infatti, risulta proponibile perfino il disegno campionario che impiega la frazione sondata del 10%, la più bassa tra quelle esaminate. Solo a causa delle regioni italiane più piccole, per le quali si sono osservati maggiori rischi di bassa accuratezza, si suggerisce l'adozione di un tasso di campionamento più elevato.

Le valutazioni condotte sulle tavole statistiche definite per contesti comunali e sub-comunali hanno messo in evidenza che l'accuratezza attesa dipende sia dal dettaglio informativo della tavola presa in esame che dalla dimensione demografica del dominio territoriale di riferimento. Infatti, è stato osservato che più piccolo è il dominio preso in considerazione, minore è l'accuratezza attesa delle tavole statistiche; inoltre, il risultato peggiora nei casi di tavole che hanno previsto un elevato numero di modalità di classificazione. Tali effetti sono però fortemente ridotti se si fa riferimento alla frazione di campionamento del 33%, la più elevata tra quelle valutate in fase sperimentale.

In sintesi, le diverse analisi empiriche hanno messo in evidenza che l'adozione della strategia campionaria tramite *short/long form* al 15° Censimento della popolazione e delle abitazioni porta a risultati con elevati livelli di accuratezza. È possibile riscontrare delle criticità per stime relative ad incroci di modalità che classificano poche unità o a tavole riferite a piccoli domini; il rischio di non poter evitare stime poco accurate è tanto più elevato quanto più fine è il dominio di riferimento e più dettagliato è il livello di classificazione delle variabili di incrocio. Tale rischio è sicuramente ridotto con l'impiego di una frazione di campionamento molto ampia, in grado di garantire livelli di accuratezza soddisfacenti per stime ad elevato dettaglio informativo anche per i domini più piccoli coincidenti con le aree di censimento di centro abitato.

I risultati conseguiti sono stati molto utili al processo decisionale per le rilevanti implicazioni sia sulla scelta della metodologia sia sulla messa a punto dello strumento di rilevazione.

Le prime decisioni sono state quelle di favorire la strategia campionaria che impiega la frazione sondata del 33% e di proporre il disegno dei domini di campionamento, le aree di censimento di centro abitato, in modo da rappresentare ambiti sub-comunali di dimensione demografica intorno alle 15mila unità.

Poiché le valutazioni relative all'errore di campionamento atteso e all'impatto delle procedure di stima sull'accuratezza delle tavole statistiche di diffusione hanno ipotizzato l'utilizzo del questionario ridotto con poche domande, è stato condotto un successivo studio volto a valutare i riflessi, sul livello di accuratezza atteso dei risultati, dall'adozione di una versione ampliata del questionario ridotto caratterizzato, oltre che dalle domande di natura strettamente demografica, anche da alcuni quesiti di carattere socio-economico (a un minimo livello di classificazione). I risultati finali di questo studio hanno evidenziato riduzioni dell'errore di campionamento e vantaggi per la stima delle frequenze più piccole, in modo più significativo nel caso della frazione sondata del 33%. Tali

conclusioni hanno avvalorato la decisione di impiegare la versione ampliata del questionario in forma ridotta. In questo modo, la maggiore disponibilità di dati rilevati in modo esaustivo sulla popolazione offre la possibilità di migliorare l'efficienza delle stime finali e l'accuratezza complessiva dei risultati contenuti nelle tavole statistiche di diffusione.

La produzione dei risultati definitivi del 15° Censimento della popolazione e delle abitazioni con livelli di accuratezza e affidabilità soddisfacenti dovrà procedere attraverso scelte condivise sia sulla procedura di riponderazione per il calcolo dei pesi di riporto all'universo che nella definizione delle tavole statistiche per la diffusione italiana.

In relazione all'operazione di stima, il sistema di vincoli da implementare nell'algoritmo di riponderazione dovrà garantire la coerenza tra i dati stimati e i dati esaustivi delle tavole statistiche per cui è stata progettata la calibrazione; di fatto, il grado di soddisfacimento della proprietà di coerenza contribuisce in modo rilevante alla qualità finale dei dati diffusi.

Il sistema di vincoli sarà tanto più completo quanto più ampio sarà l'insieme di marginali di stima riferite a tavole differenti per le quali si garantisce la coerenza. Inoltre, tale proprietà dovrà essere soddisfatta sia in senso "orizzontale" per tavole riferite a uno stesso dominio, che in senso "verticale" per tavole relative a differenti livelli territoriali.

Passando al secondo aspetto connesso all'accuratezza finale dei risultati censuari, è necessario definire le tavole che incrociano le variabili di stima in modo da raggiungere un compromesso tra l'informazione censuaria che si vuole offrire agli utenti e l'affidabilità del dato stesso che risulta dal procedimento di stima. L'obiettivo di rendere possibile la stima di una qualsiasi tavola deve, infatti, tenere conto del rischio di produrre risultati poco accurati, specialmente per distribuzioni multivariate con molte celle. Inoltre, la possibilità di rendere fruibile una qualsiasi tavola dei dati può pregiudicare il rispetto delle regole di riservatezza per qualche incrocio di variabili.

Quindi, in relazione alla definizione del piano di diffusione italiano, è evidente che la scelta dei dettagli classificatori, per i differenti livelli territoriali, potrebbe riflettersi in modo rilevante sull'accuratezza delle stime finali. A riguardo, per offrire dati con livelli di accuratezza comparabili, potrebbe essere opportuno impiegare classificazioni con un numero limitato di modalità su domini territoriali più fini e rimandare un maggiore dettaglio informativo, sia in termini di variabili di incrocio che di numero di modalità di classificazione, ad ambiti territoriali più estesi.

In prospettiva, per aumentare l'accuratezza delle stime, si dovrà tener conto della possibilità di ricorrere a soluzioni metodologiche alternative, in particolare sulla scelta dello stimatore favorendo l'impiego di metodi di stima indiretta, specialmente per la stima di frequenze riferite a piccoli domini o a popolazioni rare, casi per i quali il campione potrebbe non essere sufficientemente rappresentativo. Alcune soluzioni potrebbero derivare dall'impiego dei metodi di stima per piccole aree che, sotto opportune ipotesi di validità dei modelli sottostanti, potrebbero portare ad un recupero di efficienza a vantaggio di un miglioramento dell'accuratezza delle stime. Inoltre, tale guadagno potrebbe essere più consistente con l'impiego della versione più ampia del questionario ridotto, per la maggiore disponibilità di variabili ausiliarie.

Al fine di valutare la praticabilità dei metodi per piccole aree in ambito censuario, sono stati condotti alcuni studi, sia a carattere metodologico che sperimentale (Carbonetti e Fiorello, 2010b; Borrelli *et al.*, 2011b; Borrelli *et al.*, 2012) i cui risultati incoraggiano a proseguire gli approfondimenti su questo versante.

Appendice

Prospetto 1: Classificazione delle variabili

Sesso	maschio
	femmina
Classe di età da 15 anni in poi (4 modalità)	da 15 a 19
	da 20 a 29
	da 30 a 54
	55 e più
Classe di età (13 modalità)	fino a 5
	da 6 a 13
	da 14 a 17
	da 18 a 19
	da 20 a 24
	da 25 a 29
	da 30 a 34
	da 35 a 44
	da 45 a 54
	da 55 a 64
	da 65 a 74
	da 75 a 84
	85 e più
Classe di età quinquennale (21 modalità)	meno di 5
	da 5 a 9
	da 10 a 14
	da 15 a 19
	da 20 a 24
	da 25 a 29
	da 30 a 34
	da 35 a 39
	da 40 a 44
	da 45 a 49
	da 50 a 54
	da 55 a 59
	da 60 a 64
	da 65 a 69
	da 70 a 74
	da 75 a 79
	da 80 a 84
	da 85 a 89
	da 90 a 94
	da 95 a 99
	100 e più
Età per singolo anno (101 modalità)	singoli anni di età da 0 a 99; 100 e più

Stato coniugale	coniugati
	non coniugati
Stato civile	celibi/nubili
	coniugati/e
	separati/e legalmente
	divorziati/e
	vedovi/e
Cittadinanza	italiana
	straniera
Condizione professionale (6 modalità)	occupati
	in cerca di occupazione
	studenti
	casalinghe/i
	ritirati dal lavoro
	in altra condizione
Condizione professionale (8 modalità)	occupati
	in cerca di prima occupazione
	disoccupati
	altre persone in cerca di lavoro
	studenti
	casalinghe/i
	ritirati dal lavoro
	in altra condizione
Posizione nella professione (5 modalità)	imprenditore e libero professionista
	lavoratore in proprio
	socio di cooperativa
	coadiuvante familiare
	dipendente o in altra posizione subordinata
Posizione nella professione (6 modalità)	imprenditore
	libero professionista
	lavoratore in proprio
	socio di cooperativa
	coadiuvante familiare
	dipendente o in altra posizione subordinata

Condizione e posizione nella professione	FORZE DI LAVORO
	Occupati
	imprenditore e libero professionista
	lavoratore in proprio
	socio di cooperativa
	coadiuvante familiare
	dipendente o in altra posizione subordinata
	in cerca di occupazione
	NON FORZE DI LAVORO
	studenti
	casalinghe/i
	ritirati dal lavoro
	in altra condizione
Professione	lavoratore operaio o di servizio non specializzato
	addetto a impianti fissi di produzione, a macchinari, a linee di montaggio o conduce veicoli
	svolge un'attività operaia qualificata
	coltiva piante e/o alleva animali
	svolge un'attività di vendita al pubblico o di servizio alle persone
	svolge un'attività impiegatizia di tipo non tecnico
	svolge un'attività tecnica, amministrativa, sportiva o artistica a media qualificazione
	svolge un'attività organizzativa, tecnica, intellettuale, scientifica o artistica ad elevata specializzazione
	gestisce un'impresa o dirige il lavoro di strutture organizzative complesse
	lavora come ufficiale, sottufficiale, allievo o volontario nelle Forze Armate
Attività economica (3 modalità)	agricoltura
	industria
	altre attività
Sezioni di attività economica (17 modalità)	agricoltura, caccia e silvicoltura
	pesca, piscicoltura e servizi connessi
	estrazione di minerali
	attività manifatturiere
	produzione e distribuzione di energia elettrica, gas e acqua
	costruzioni
	commercio all'ingrosso e al dettaglio; riparazione di autoveicoli, motocicli e di beni personali e per la casa
	alberghi e ristoranti
	trasporti, magazzinaggio, e comunicazioni
	intermediazione monetaria e finanziaria
	attività immobiliari, noleggio, informatica, ricerca, altre attività professionali e imprenditoriali
	pubblica amministrazione e difesa; assicurazione sociale obbligatoria
	istruzione
	sanità e altri servizi sociali
	altri servizi pubblici, sociali e personali
	servizi domestici presso famiglie e convivenze
	organizzazioni ed organismi extraterritoriali

Grado di istruzione (8 modalità)	laurea
	diploma di scuola secondaria superiore
	licenza di scuola media inferiore o di avviamento professionale
	licenza di scuola elementare
	alfabeti privi di titoli di studio
	di cui: in età da 65 anni in poi
	analfabeti
	di cui: in età da 65 anni in poi
Grado di istruzione (7 modalità)	laurea
	diploma universitario o terziario di tipo non universitario
	diploma di scuola secondaria superiore
	licenza di scuola media inferiore o di avviamento professionale
	licenza di scuola elementare
	alfabeti privi di titoli di studio
	analfabeti
Numero di figli del nucleo	nessun figlio
	1 figlio
	2 figli
	3 o più figli
Luogo di destinazione	nello stesso comune di dimora abituale
	fuori dal comune

Riferimenti bibliografici

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

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

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

Bianchi G., F. Di Pede, A. Reale e S. Talice. 2010. Aree di censimento, nuove suddivisioni sub-comunali per la raccolta campionaria di informazioni aggiuntive durante il prossimo censimento della popolazione: applicazioni nella regione Marche. Atti della XXXI Conferenza Italiana di Scienze Regionali, Aosta 20-22 settembre.

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

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

Borrelli F., G. Carbonetti e L. De Felici. 2009. Problemi di accuratezza delle stime da campioni di famiglie in un contesto censuario. Giornate di Studio sulla Popolazione, VIII Edizione, Milano 2-4 febbraio.
http://www.unicatt.it/convegno/gsp09/allegati/pdf%20link%20Programma%20relazioni%20estese/Borrelli_Carbonetti_DeFelici_esteso_rev1.pdf

Borrelli F., G. Carbonetti, S. Dardanelli e L. De Felici. 2010. Una nuova strategia per il Censimento della popolazione e delle abitazioni del 2011: confronto tra tecniche per la produzione di informazione territoriale di qualità. Atti della XXXI Conferenza Italiana di Scienze Regionali, Aosta 20-22 settembre.

Borrelli F., G. Carbonetti, L. De Felici, E. Fiorello e M. Marrone. 2011a. *La progettazione dei censimenti generali 2010-2011: disegni campionari e stima di errori di campionamento*. Istat Working Papers, Collana Scientifica dell'Istituto Nazionale di Statistica. 2/2011. Roma.
http://www.istat.it/it/files/2011/06/Istat_Working_Papers_2_2011.pdf

Borrelli F., G. Carbonetti, E. Fiorello e F. Solari. 2011b. Metodologie di stima per piccole aree basate su autocorrelazione spaziale applicabili a variabili di censimento. Atti della XXXII Conferenza Italiana di Scienze Regionali, Torino 15-17 settembre.

Borrelli F., G. Carbonetti, L. De Felici, F. Solari. 2012. *Metodologie di stima per piccole aree applicabili a variabili di censimento*. Istat Working Papers, Collana Scientifica dell'Istituto Nazionale di Statistica. 3/2012. Roma.
http://www.istat.it/it/files/2012/02/Istat_Working_Papers_n3_2012.pdf

Carbonetti G. e C. De Vitiis. 2007. Efficienza di stime campionarie relative ad un sottoinsieme di variabili di censimento. Atti della Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni". CNR, Roma novembre.

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

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

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

Carbonetti G. 2009a. Use of sampling strategy in the Italian population census and accuracy of estimates for different territorial domains. ITACOSM09 - First Italian Conference on Survey Methodology, Siena 10-12 giugno.

Carbonetti G. e M. Verrascina. 2009b. Accuracy evaluation of Nuts level 2 hypercubes with the adoption of a sampling strategy in the 2011 Italian population census. Group of Experts on Population and Housing Census. UN, Ginevra ottobre. ECE/CES/GE.41/2009/10 http://unstats.un.org/unsd/censuskb20/Attachments/2009ITA_ECE_Quality-GUIDae523c3db60f450dab82cbc01c707580.pdf

Carbonetti G. e M. Verrascina. 2010a. Accuracy evaluation of dissemination data adopting a sampling strategy in the 2011 Italian Population Census. Group of Experts on Population and Housing Census. Proceedings of the European Conference on Quality in Official Statistics, Helsinki.

Carbonetti G. e E. Fiorello. 2010b. La produzione di informazione statistica a livello territoriale sub-comunale: possibili cambiamenti indotti dalla strategia proposta per il Censimento della popolazione delle abitazioni del 2011. Atti della XXXI Conferenza Italiana di Scienze Regionali, Aosta 20-22 settembre.

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

Cocchi D. 2007. Uso dei campioni nelle rilevazioni censuarie. Atti della Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni". CNR, Roma novembre.

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

Dardanelli S., S. Mastroluca, A. Sasso e M. Verrascina. 2008. *La progettazione dei censimenti generali 2010–2011. Novità di regolamentazione internazionale per il 15° Censimento generale della popolazione e delle abitazioni*. Istat, Dati e prodotti, Pubblicazioni scientifiche, Documenti n. 1/2009. Roma.

Dardanelli S., A. Sasso e M. Verrascina. 2009. Comparabilità dell'output censuario a livello europeo: dall'esperienza della precedente tornata dei censimenti demografici alla definizione degli hypercubes per la prossima. Atti della XXX Conferenza Italiana di Scienze Regionali, Firenze 9-11 settembre.

Dardanelli S., A. Sasso e M. Verrascina. 2010. La diffusione dei dati del Censimento della popolazione e delle abitazioni del 2011 alla luce di alcune novità introdotte a livello nazionale e internazionale. Atti della XXXI Conferenza Italiana di Scienze Regionali, Aosta 20-22 settembre.

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

European Community. 2008. Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses

Ferruzza A., S. Mastroluca e D. Zindato. 2007. I censimenti esteri: modelli a confronto alla luce dei regolamenti internazionali, Conferenza Nazionale di Statistica: "Censimenti generali 2010-2011. Criticità e innovazioni". CNR, Roma, Novembre 2007

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

Istat. 2006a. *Il Piano di rilevazione e il Sistema di produzione*. Roma.

Istat. 2006b. *I Documenti*. Roma.

Istat. 2009. *La qualità dei dati*. Roma.

ONU. 2008. United Nations, Statistics Division, Department of Economic and Social Affairs. *Principles and Reccommendations for Population and Housing Censuses Revision 2*. New York.

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

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

UNECE. 1998. United Nations Economic Commission for Europe and the Statistical Office of the European Communities. *Recommendations for the 2000 Censuses of Population and Housing in the Ece Region*. Statistical Standards and Studies No. 49. United Nations Publication.

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

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

Informazioni per gli autori

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

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

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

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