SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

21 ottobre 2001

La qualità dei dati

14° Censimento generale della popolazione e delle abitazioni

Conoscere il censimento

Istat

Censimento 2001

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

La qualità dei dati

Conoscere il Censimento

14° Censimento generale
della popolazione e delle abitazioni

A cura di: Fernanda Panizon

Coordinamento redazionale: Patrizia Collesi e Giorgia Capacci

Per informazioni sul contenuto della pubblicazione
rivolgersi al Cont@ct Centre dell'Istat all'indirizzo:
https://contact.istat.it//

Eventuali rettifiche ai dati pubblicati saranno diffuse
all'indirizzo www.istat.it nella pagina di presentazione del volume

La qualità dei dati

Conoscere il Censimento

14° Censimento generale
della popolazione e delle abitazioni

ISBN 978-88-458-1624-6

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Realizzazione: Istat, Servizio Editoria

Stampato nel mese di novembre 2009
Per conto dell'Istat presso
Centro stampa e riproduzione S.r.l.
Via di Pietralata, 157 – Roma

Indice

Capitolo 3 – Le analisi per la stima dell'errore di risposta

Presentazione

Il Censimento è stata un'operazione statistica molto complessa e costosa, che ha coinvolto tutti gli 8.101 comuni italiani e tutta la popolazione residente e presente alla data di riferimento. In questo volume vengono illustrate e fornite alcune valutazioni sulle diverse dimensioni della qualità dei 14° Censimento generale della popolazione e delle abitazioni.

Nel primo capitolo vengono prese in esame le diverse fasi del processo produttivo dei dati censuari, che partendo dalla fase di acquisizione dei dati e attraverso le fasi di controllo e correzione, hanno portato alla costruzione di insiemi di dati validati, disponibili per l'elaborazione e la successiva diffusione. Per ciascuna fase considerata si fornisce una misurazione della qualità, considerando una serie di analisi e di indicatori e verificando in che modo le scelte operative via via adottate abbiano avuto effetti sui dati prodotti.

La qualità del censimento, come capacità di conteggiare esaustivamente la popolazione sul territorio, viene invece analizzata nel secondo capitolo, dedicato ai risultati dell'Indagine sul grado di copertura del 14° Censimento generale della popolazione, effettuata qualche settimana dopo il censimento per fornire una misura di questo aspetto della qualità dei dati censuari. Vengono qui descritte le principali caratteristiche dell'indagine, dalla progettazione alle operazioni sul campo, e quindi riportati i principali risultati.

Nel terzo capitolo si illustrano ulteriori analisi della qualità che, a partire dai dati raccolti con la predetta indagine di copertura, sono state condotte per valutare lo scostamento dei valori osservati di alcune variabili nelle due occasioni di indagine e per fornire una valutazione indiretta dell'errore di misura. I risultati presentati dimostrano la sostanziale buona qualità del Censimento del 2001, se pure mettono in evidenza qualche marginale criticità, ad esempio nella non perfetta comprensione di alcuni quesiti, da parte dei rispondenti, o nella difficoltà di censire correttamente alcuni specifici sottogruppi di popolazione.

Capitolo 1 - La qualità del processo

1.1 - L'approccio al problema della qualità e il piano di qualità

La qualità e la buona riuscita di una rilevazione statistica complessa come un censimento dipendono da numerosi fattori e competenze. Innanzitutto dalla capacità di progettare gli strumenti di rilevazione più idonei (questionari, modelli ausiliari, manuale di istruzioni), di pianificare le operazioni preliminari (stampa e invio dei modelli, istruzioni agli organi periferici), di organizzare i controlli della rilevazione sul campo (monitoraggio e ispezioni), di pianificare dettagliatamente tutte le operazioni di consegna, ritiro e revisione dei modelli, di verificare la fase di acquisizione dei dati attraverso accurati controlli di accettazione, di adottare un adeguato piano di controllo e correzione dei dati e, in definitiva, dipendono anche da quella esperienza che consente di prevedere i possibili scostamenti dalle regole date, per tentare di prevenirli e gestirli.

La strategia per la qualità del 14° Censimento della popolazione ha previsto un sistema di controllo della qualità tale da soddisfare tre aspetti fondamentali:

- monitoraggio di tutte le fasi, sia sui dati che sulle procedure, in modo da garantire interventi tempestivi ed efficaci nella gestione delle situazioni anomale;
- visione integrata di tutte le fasi censuarie, realizzabile con uno scambio di informazioni fra i diversi sistemi di monitoraggio ed elaborazione dei dati, resi coerenti fra loro (dalle operazioni sul campo alle basi territoriali, dal processo di elaborazione,alle basi di dati esterne);
- memorizzazione degli eventi, dei dati derivati dai controlli, dei risultati delle procedure in un "archivio di qualità", per consentire una valutazione finale dell'errore e fornire indicazioni sintetiche sulla qualità dei dati.

Per mettere a punto il sistema dei controlli sono state effettuate diverse attività preliminari:

1. *Definizione delle modalità di controllo di qualità sulla fase di acquisizione dei dati (lettura ottica e registrazione tradizionale)*

La maggior parte dei dati del 14° Censimento è stata acquisita in forma digitale attraverso la tecnologia della lettura ottica, utilizzando modelli di rilevazione appositamente predisposti e appaltando ad una società specializzata lo svolgimento delle attività connesse al processo di acquisizione ottica. I livelli di qualità attesi per questa fase sono stati definiti nel capitolato tecnico di gara, in cui sono state specificate le richieste e i requisiti del servizio da fornire. Alla società di lettura ottica sono state preventivamente rese note non solo le soglie di affidabilità stabilite dall'Istat, ma anche le modalità di verifica sull'effettivo raggiungimento di tali soglie, sono state comunicate.

Il piano dei controlli sulle operazioni di acquisizione ottica prevedeva l'estrazione di campioni di controllo per ogni consegna (sottoinsieme di dati provinciali), in modo da verificare sia la qualità delle immagini acquisite otticamente, sia la corrispondenza fra quanto indicato nel modello e le risultanze dell'acquisizione, in termini di dati registrati e di codifica delle variabili alfabetiche.

Considerata la mole di dati e di immagini da controllare, questa operazione di verifica sulla qualità non sarebbe stata effettuabile all'interno dell'Istat, pertanto essa è stata delegata ad una società esterna di monitoraggio, certificata Aipa (Autorità per l'informatica nella pubblica amministrazione), che doveva fornire per ogni consegna di dati e immagini (blocco), un dettagliato verbale sugli esiti dei controlli. Analizzando gli esiti di tali verifiche l'Istat poteva valutare se le soglie di qualità attesa fossero o meno rispettate e quindi decidere se accettare o rifiutare il blocco consegnato.

La registrazione tradizionale (con operatore) è stata scelta per un numero limitato di modelli censuari, non predisposti per la lettura ottica (i Modelli Istat CP.2 per la rilevazione delle convivenze, i Modelli Istat CP.10 relativi ai dati riepilogativi per sezione, i Modelli Istat CP.1 "integrativi", cioè i modelli inviati dai comuni successivamente alle date obiettivo, i Modelli Istat CP.1 in lingua slovena). Per questo tipo di acquisizione i

---

Il presente paragrafo è stato redatto da: Fernanda Panizon (parr. 1.1, 1.2, 1.3, 1.3.1, 1.3.2, 1.3.3, 1.3.4, 1.4, 1.5, 1.6, 1.7, 1.9, 1.10, 1.12, 1.14); Letizia Buzzi (par. 1.13); Giorgia Capacci (par. 1.3.5); Luana De Felici (par. 1.11). Le elaborazioni e le analisi di questo capitolo sono state condotte con la collaborazione di: Letizia Buzzi, Luana De Felici, Fabrizio Delli Priscoli, Epifania Fiorello.

controlli previsti sono stati quelli consueti di qualità campionari effettuati dall'Istat: i dati nuovamente registrati sono stati confrontati con quelli della fornitura esterna, per verificare se la percentuale di battute errate fosse in linea con la soglia contrattuale prefissata.

2. Progettazione ed implementazione della 'Base di Dati', (database di produzione)

Una delle innovazioni del Censimento 2001 rispetto a quelli precedenti è stata l'utilizzazione di un database relazionale per l'organizzazione e la gestione dei dati. Vista la complessità delle operazioni da effettuare sui dati, la molteplicità dei legami fra le unità e le variabili coinvolte, la mole di dati da trattare, la definizione e la costruzione di tale sistema ha richiesto un notevole impiego di risorse informatiche, in termini di hardware, di software e di professionalità informatiche.

Il controllo della fase di caricamento dei dati nel *database* è stato progettato per garantire la coerenza con le informazioni di supporto e la consistenza con le informazioni derivanti dai precedenti controlli di qualità effettuati nella fase di acquisizione.

3. Definizione dei criteri e degli schemi per il controllo della coerenza fra l'ammontare di microdati acquisiti nel database e i totali derivanti dai modelli riassuntivi provenienti dai Comuni

Il flusso delle operazioni di validazione, propedeutiche al rilascio dei dati, si è concentrato inizialmente sulla diffusione dei "primi risultati" (cioè dei dati aggregati forniti dai Comuni tramite il Modello Istat CP8.bis) e successivamente sulla pubblicazione della popolazione legale per comune. Queste priorità hanno concentrato gli sforzi iniziali di analisi e correzione dei dati sugli aspetti di verifica della coerenza quantitativa fra i totali risultanti da conteggi sui modelli di rilevazione, i totali risultanti da fonti statistiche esterne e i totali rilevati dai modelli ausiliari di censimento. La specificazione di quali macrodati mettere a confronto e delle relative soglie di accettabilità nelle differenze riscontrate, la definizione delle analisi e delle procedure da attivare per una eventuale rettifica (anche a livello di sezione di censimento per la definizione della popolazione legale) sono stati i primi passi da compiere in fase di controllo della qualità dei dati.

4. Definizione delle fasi (per unità di rilevazione e/o per gruppi di variabili) del processo di revisione e validazione dei dati, degli schemi di controllo e correzione (corrispondenti a ogni fase individuata) e infine il rilascio dei dati aggregati ai vari livelli territoriali

I questionari di censimento erano differenziati per le diverse unità di rilevazione e di analisi: le famiglie e le convivenze, suddivise nelle componenti di popolazione residente (che, come definita nelle istruzioni, corrisponde alle persone *abitualmente dimoranti nell'alloggio*) e non residente (i *temporaneamente dimoranti nell'alloggio*), gli alloggi, le abitazioni (occupate e non), gli edifici. Per ciascun sottoinsieme di quesiti, riferiti alla diverse unità, sono state predisposte procedure di controllo e correzione. Inoltre, dato che non esiste una procedura o uno strumento generalizzato che risolva completamente e simultaneamente il problema della correzione dei dati di questionari complessi, è stato necessario procedere alla partizione del questionario in blocchi (per unità d'analisi, per gruppi di variabili omogenee). Ogni blocco di variabili considerato è stato analizzato e corretto con una singola specifica procedura. La procedura è a sua volta suddivisa in moduli (per esempio, modulo deterministico, modulo probabilistico) o in passi, che risolvono un particolare aspetto. La suddivisione del questionario in blocchi, da trattare con procedure indipendenti, implica di norma che sia introdotta una sequenza o gerarchia nel trattamento e nella correzione dei blocchi e delle relative variabili. Deve accadere cioè che, una volta modificate o corrette le variabili di un certo gruppo o blocco, queste debbano risultare fisse (cioè non ulteriormente modificabili) per i successivi moduli di correzione.

In generale, per costruire una buona procedura di validazione dei dati e definirne i passi costituenti (per esempio, la sequenza o il tipo di approccio) è necessario valutare i risultati delle analisi preliminari sulla qualità dei dati grezzi, come i tassi di mancata risposta, gli indicatori di confronto con fonti esterne, le distribuzioni di frequenza delle variabili grezze (così come provengono dalla fase di acquisizione su supporto informatico), i valori modali delle incompatibilità rilevate, per identificare da un lato le eventuali sistematicità degli errori e dall'altro per definire quali siano le variabili più affidabili o più importanti, sulle quali cioè costruire il processo di correzione.

Una scelta ulteriore da compiere per costruire una solida procedura di controllo e correzione riguarda il livello territoriale di riferimento utilizzato: questa scelta si basa prevalentemente sulla dimensione dei *dataset* da trattare. Quindi, per esempio, i dati relativi ai 21 milioni di famiglie sono stati validati provincia per provincia,

mentre per le convivenze i dati da validare sono stati raggruppati nelle tre ripartizioni geografiche da trattare separatamente.

5. Definizione delle regole di incompatibilità fra i valori delle variabili e delle regole di correzione dei dati

Per la definizione delle regole di controllo e correzione, le diverse sezioni dei modelli di rilevazione, suddivise per gruppi di variabili omogenee o per area tematica, sono state considerate e analizzate separatamente dai ricercatori, responsabili degli specifici temi, i quali hanno specificato le regole di compatibilità e coerenza. Il piano di validazione ha previsto che fossero analizzate a priori le regole di compatibilità relative a ciascun gruppo di variabili, cioè che queste fossero testate sui dati grezzi, per verificare la qualità iniziale dei dati e la eventuale presenza di errori sistematici. In presenza di errore sistematico infatti è necessario procedere ad opportuni approfondimenti e ragionamenti per definire il miglior criterio di correzione (di tipo deterministico o probabilistico), in modo che assegnando nuovi valori non si produca distorsione nei dati. Ad esempio, avendo verificato che in molte occasioni lo "stato civile" dei bambini non era indicato sul modello, si procedeva forzatamente (in maniera deterministica) ad imporre lo stato civile "celibe/nubile", mentre per la mancata risposta su "stato civile" in altre classi di età la coerenza della correzione veniva cercata considerando i legami familiari o con una imputazione che seguisse una distribuzione data.

Per quanto riguarda gli errori che si presentano con minor frequenza, per i quali si può ipotizzare un'origine casuale, la correzione viene generalmente affidata a procedure di tipo probabilistico, che cioè modificano i dati errati assegnando valori casuali all'interno del *range* compatibile, oppure valori che dipendono dalle distribuzioni osservate nei dati esatti.

È importante sottolineare la delicatezza della fase di correzione dei dati di censimento, che per alcune variabili può essere soggetta a controlli strettissimi da parte dei Comuni, ove questi dispongano di informazioni aggiornate e coerenti nei loro archivi (per esempio sesso e anno di nascita dei residenti, cittadinanza, nati nel comune nel corso dell'anno, eccetera), o nei Comuni di piccole dimensioni che possono riuscire ad intercettare anche minime discrepanze fra dati censuari e caratteristiche note della propria popolazione.

6. Definizione degli indicatori di qualità e Data Warehouse di controllo della produzione.

Una volta che le procedure di correzione sono state implementate e testate, si è passati al loro utilizzo nel processo vero e proprio. Per tenere sotto controllo le procedure informatiche e gli effetti di queste sui dati sono stati predisposti indicatori, *report* e tabelle di verifica, da produrre relativamente a ciascuna fase attivata, in modo da evidenziare eventuali valori indesiderati o anomalie, scostamenti rilevanti nelle distribuzioni ottenute rispetto ad informazioni esterne, distorsione nei dati.

Allo scopo di monitorare il processo di validazione è stato progettato e implementato il *Data Warehouse* di controllo della produzione, nel quale sono stati caricati i dati di input delle procedure ed i dati di output per i confronti con gli indicatori di qualità, che servono a valutare l'impatto sui dati delle procedure di correzione (percentuale di modifiche introdotte, matrici di transizione, eccetera) e con i dati di fonte esterna (come l'indagine Posas ed il Censimento 1991, ai diversi livelli territoriali) per i confronti a livello macro.

È stata fondamentale l'implementazione di un ben preciso Sistema per il controllo della qualità dei dati, attivo nelle fasi di lavorazione e validazione, progettato in modo da integrare tutte le procedure e soprattutto le informazioni sui controlli e sulle analisi effettuate, in modo che le procedure e i dati di controllo fossero integrati con le altre attività di acquisizione e correzione dei dati. All'interno del più generale Sistema informativo del Censimento popolazione e abitazioni (l'insieme dell'organizzazione delle risorse e di tutte le attività il cui scopo è rilasciare i microdati finali per la diffusione) è possibile individuare quindi due sistemi paralleli: uno di "produzione", dedicato alla elaborazione dei dati, alla loro modifica e alla loro aggregazione; un altro di *quality management*, l'insieme di tutte le procedure adottate per controllare e valutare la qualità dei dati e per produrre alcuni indicatori finali di qualità. Di tale Sistema di *quality management* fanno parte tre importanti "sottosistemi", legati alle tre macrofasi delle attività di produzione dei dati censuari: le operazioni sul campo che coinvolgono gli organi periferici (Sottosistema di monitoraggio della rilevazione), la fase di acquisizione dei dati su supporto informatico (Sottosistema di monitoraggio per l'acquisizione dei dati, in generale, per tutte le operazioni connesse alla lettura ottica e alla registrazione dei dati, che sono state appaltate all'esterno) e la fase di controllo e correzione (Fig. 1.1) (Sottosistema di controllo della "produzione interna", cioè dell'elaborazione dei dati).

1.2 - Il Sistema di controllo della qualità dei dati nella lavorazione e validazione

Come nei passati censimenti, sono stati previsti controlli sui dati al fine di organizzare un'adeguata fase di correzione. La maggior parte dei controlli è avvenuta attraverso specifiche procedure informatiche ed il risultato dei controlli, eseguiti durante l'elaborazione dei dati, ha costituito un input per la scelta sull'azione correttiva da compiere. Altre analisi, volte alla produzione di indicatori atti a misurare la qualità del dato prodotto, sono state effettuate dopo il rilascio dei dati.

La necessità di coniugare un'ottica di *continuous quality improvement* con la specificità dell'indagine censuaria rende indispensabile e cruciale una grande flessibilità nella gestione dei controlli e nella possibilità di impostare nuove e specifiche analisi di qualità nel corso dell'elaborazione. Per le indagini statistiche correnti ripetute nel tempo si producono in genere dati sulla qualità con l'obiettivo di migliorare le indagini successive: l'impegno per la qualità si traduce quindi nel mettere a frutto nel tempo le esperienze maturate nelle tornate precedenti. Per i censimenti, l'intervallo tra due rilevazioni è talmente lungo che il contesto e la stessa struttura d'indagine si ripresentano in maniera molto differente, rendendo problematico l'utilizzo di dati di qualità retrospettivi. Per migliorare la qualità dei dati censuari è necessario quindi pianificare il più possibile le attività correnti di controllo e di valutazione, in modo che sia possibile intervenire tempestivamente sull'organizzazione e/o sui dati in corso d'opera (a seconda dell'errore riscontrato) e prendere le opportune decisioni sugli aggregati da diffondere.

In questo quadro si giustifica l'importanza e l'impostazione pensata per il Sistema di controllo della qualità dei dati per il Censimento 2001, cioè dello specifico sistema per la gestione di dati, di metadati e di procedure e per la realizzazione di un archivio di qualità. Le principali necessità prese in considerazione sono state le seguenti:

- identificare e assegnare una maggiore rilevanza alle varie fasi di controllo e di analisi dei dati, attraverso la creazione di un vero e proprio sistema, con propri peculiari requisiti e meccanismi procedurali. Se si considerano alcuni disegni proposti in letteratura per definire la produzione del dato statistico, le varie procedure di controllo sono usualmente suddivise nei vari sottosistemi, ma non sempre sono tra esse integrate. Questo approccio può ostacolare o limitare le analisi sui dati e in definitiva non consentire di avere "il controllo" complessivo sulla lavorazione. Creare un sistema di controllo significa inoltre organizzare in maniera adeguata le risorse, sia dal punto di vista degli strumenti informatici, sia delle figure professionali impiegate;
- predisporre una "libreria di controlli automatici", vale a dire di una serie di controlli sui dati nelle diverse fasi, che producono *report* sistematici, che aiutino a segnalare situazioni di incoerenze e/o di anomalia;
- rendere agevoli e possibili le analisi esplorative sui dati, al fine di consentire l'individuazione di errori non previsti e di analizzare le possibili cause di errore;
- produrre, per ogni fase, un registro di indicatori della qualità, dove memorizzare i controlli, gli esiti dei controlli, le azioni intraprese ed i risultati delle azioni. La disponibilità di una fonte informativa sulle modifiche che i dati hanno subìto all'interno del processo garantisce la possibilità di un *audit trail* (tracciabilità a ritroso) dei dati, molto utile ai fini del controllo in linea, per cercare di individuare con maggiore precisione l'istante e la fonte di generazione dell'errore (si pensi al caso di errori che ne possono generare altri a cascata). Per quanto riguarda la misurazione finale dell'accuratezza del dato (valutazione dell'errore) si provvede a far confluire i dati di controllo provenienti da questi registri nella struttura unitaria costituita dall'archivio di qualità, in grado di fornire sia indicatori analitici, sia indicatori sintetici sulla qualità complessiva dei dati e del processo.

Riguardo alle funzionalità previste dal sistema adottato, si può distinguere quindi tra la possibilità di visualizzare l'esito dei controlli automatici attraverso *browser* e report standard e la possibilità di condurre specifiche e approfondite analisi sui dati. In alcuni casi, inoltre, è stata prevista, sulla base degli esiti dei controlli, la funzione di "ritorno indietro" rispetto alle azioni di correzione compiute sui dati.

Figura 1.1 - Il Sistema di controllo della qualità e sistema di produzione dei dati

Le aree di analisi nel sistema di controllo della qualità vengono individuate sulla base delle possibili tipologie di errori (A. Chieppa, F. Panizon, 2001); per ciascuna di queste aree sono stati previsti sia controlli automatizzati sia possibilità di analisi "libere":

- errori nel calcolo dei primi dati di Censimento (dati provvisori), quindi analisi sulla validità/attendibilità delle comunicazioni degli Uffici di censimento comunali sui "primi risultati", per avere livelli di qualità necessari alla diffusione dei dati;
- errori attribuibili a smarrimento o alterazione di modelli di rilevazione, quindi analisi di controllo sulle quantità di modelli acquisiti, confrontando i dati desumibili dai diversi modelli ausiliari e i dati della fase di monitoraggio della rilevazione e della fase di acquisizione;
- errori di copertura, che comportano l'analisi di confronto tra dati censuari e dati che derivano da altre fonti demografiche;
- errori nella identificazione delle unità censite e pertanto analisi sui totali e su relazioni tra le diverse unità (ad esempio edifici e abitazioni, abitazioni e famiglie);
- errori nei valori relativi alle singole variabili rilevate, che implicano analisi delle distribuzioni delle variabili e delle violazioni delle regole (gli *edits*) di compatibilità fra le variabili;
- errori introdotti dalle procedure di elaborazione, attraverso analisi delle modificazioni introdotte sui dati.

1.3 - L'acquisizione dei dati con la tecnologia della lettura ottica

Le tecniche disponibili per l'acquisizione dei dati al momento della rilevazione censuaria erano diverse. Tenendo conto della tecnica di indagine adottata (intervista per autocompilazione) e del livello di complessità organizzativa che coinvolgeva tutti i Comuni italiani nella predisposizione di diverse tipologie di modelli, sono state adottate, oltre che le modalità tradizionali di registrazione dei dati, come il *data entry in-service*, anche la tecnologia della lettura ottica, affiancata a procedure di codifica automatica, l'acquisizione via *web* e l'acquisizione di modelli informatizzati all'origine (cioè modelli riepilogativi registrati direttamente su supporto informatico dagli uffici territoriali competenti).

La scelta di acquisire i dati del censimento tramite lettura ottica è stata dettata in primo luogo dall'esigenza di garantire un elevato livello della qualità dei dati, una standardizzazione di tale livello e la tempestività nella diffusione dei risultati. L'Istituto ha pertanto avviato, ad aprile 2000, le procedure di una gara di appalto per l'acquisizione di dati censuari tramite la tecnologia della lettura ottica, nominando la commissione tecnica[1] per la definizione del capitolato. In considerazione del fatto che nella provincia autonoma di Bolzano i modelli di rilevazione si sarebbero dovuti stampare in doppia lingua (italiano e tedesco), e quindi fuori dallo standard nazionale compatibile con la lettura ottica, la provincia di Bolzano ha quindi provveduto per conto proprio all'acquisizione dei dati censuari, tramite la registrazione tradizionale con operatore.

Il lavoro della commissione tecnica ha seguito, quale linea guida, il principio di affidare ad un unico fornitore di servizi (in seguito "Fornitore") la responsabilità della progettazione grafica del questionario (per renderlo compatibile con la lettura ottica), della stampa dei modelli, della loro movimentazione (consegna e ritiro presso i Comuni) e della acquisizione dei dati tramite lettura ottica.

L'unificazione delle responsabilità delle varie fasi in un unico referente ha voluto evitare che lavorazioni separate ed indipendenti creassero conflitti negli standard tecnici necessari all'acquisizione tramite lettura ottica. Pertanto il formato dei modelli, i codici di personalizzazione, l'impaginazione, il confezionamento, il trasporto, eccetera, sono stati organizzati ed effettuati con l'obiettivo di ottimizzare il processo della lettura ottica. L'affidamento della gestione e dell'integrazione delle fasi ad un unico fornitore, oltre a garantire una semplificazione amministrativa, intendeva minimizzare l'eventualità che la singola fase potesse pregiudicare la qualità finale dei dati, con uno scarico di responsabilità fra le imprese erogatrici dei diversi specifici servizi.

Oltre ai servizi sopra citati, al Fornitore si chiedeva anche la consegna ad Istat di un sistema informatico (hardware e software) per la gestione e per la consultazione delle immagini ("Sistema immagini") delle singole pagine acquisite tramite lettura ottica, in modo tale che fosse possibile utilizzarle nel processo di controllo della fornitura, ma anche durante la fase successiva di validazione dei dati. La conservazione delle immagini dei modelli su supporto informatico e la loro consultabilità ha consentito all'Istituto di evitare il deposito, stabilito da norme di legge, dei milioni di modelli cartacei in appositi (costosi ed onerosi) magazzini.

Il capitolato di gara definiva dettagliatamente i requisiti richiesti in termini di processo, di controllo di processo e di qualità. In particolare erano stabilite le modalità finali di consegna dei dati e delle immagini su supporto informatico, i livelli di qualità di acquisizione dei dati e di codifica della variabili alfabetiche, il tipo di controlli di accettazione che sarebbero stati effettuati.

Per garantire omogeneità nei controlli di qualità sulle consegne, il contratto prevedeva che il Fornitore effettuasse le lavorazioni per gruppi di comuni, aggregati a livello provinciale, e che in linea di massima i modelli di ogni provincia potessero dar luogo ad uno o più "blocchi" di dati e immagini, a seconda delle dimensioni (da un blocco per le province più piccole, fino a otto blocchi per la provincia di Roma) in modo che ogni blocco consegnato all'Istat fosse riferibile ad un massimo di 250 mila Modelli Istat CP.1 (famiglie/alloggi).

Un primo requisito di qualità, stabilito dal capitolato, era riferito alla congruenza fra dati e immagini, cioè alla corrispondenza fra dati presenti in ciascun record e le immagini delle pagine del modello generatore di quei dati, in pratica alla correttezza del codice identificativo univoco (*barcode*), che consentiva di associare i record dei dati alle corrispondenti immagini digitalizzate.

I livelli di qualità per l'acquisizione con lettura ottica richiesti contrattualmente dal capitolato tecnico, sono stati specificati diversamente a seconda del tipo di variabile, della sua importanza nel contesto della rilevazione,

[1] Commissione tecnica che entro e non oltre il 7 luglio 2000 approvi un capitolato tecnico e definisca le modalità di espletamento della gara volta alla scelta della società cui affidare i servizi relativi all'acquisizione dei dati del Censimento generale della popolazione e delle abitazioni del 2001, mediante tecniche di lettura ottica.

del tipo di riconoscimento ottico necessario.[2] Per esempio, i codici identificativi del modello prestampati come *barcode* (codice a barre) sono solitamente riconosciuti dalla lettura ottica con tassi di errore bassissimi, così pure le "biffature" delle risposte precodificate generalmente non danno luogo a problemi di riconoscimento ottico. Al contrario, tutto ciò che è scritto a mano dal compilatore del modello (date, orari, codici vari) è più problematico da identificare per il lettore ottico (*scanner*), per cui i livelli di tolleranza di errore ammesso sono stati tenuti più alti. Solo per "data di nascita" e "sezione di censimento", variabili cruciali per la validazione dei dati, la tolleranza sull'errore era ridotta (Tavola 1.1) e i livelli di qualità richiesta più alti.

Tavola 1.1 - Livelli minimi contrattuali di accuratezza da conseguire con l'acquisizione tramite lettura ottica per tipo di variabile *(valori percentuali)*

VARIABILI/ TIPOLOGIE DI DATI DA ACQUISIRE CON LETTURA OTTICA	Livello minimo di accuratezza richiesto nel capitolato di gara
Codici prestampati (identificativo di modello completo di numero di pagina, codici Istat di provincia e comune)	99,995
Variabili precodificate (biffature)	99,3
Variabili numeriche (esclusa "data di nascita" e "sezione di censimento")	98,0
Data di nascita	99,5
Sezione di censimento	99,5

Una differenziazione nei livelli di qualità attesa è stata imposta anche per la codifica delle variabili alfabetiche (Tavola 1.2), a seconda della difficoltà presunta di assegnazione del codice corrispondente, come risultante da alcuni test effettuati in precedenza.

Tavola 1.2 - Livelli minimi contrattuali di assegnazione e di accuratezza da conseguire con la codifica delle stringhe alfabetiche per tipo di variabile *(valori percentuali)*

VARIABILI ALFABETICHE DA CODIFICARE	Livello minimo di assegnazione del codice richiesto nel capitolato di gara	Livello minimo di accuratezza richiesto nel capitolato di gara
Comune	95,0	99,0
Stato estero	90,0	98,0
Titolo di studio	80,0	98,0

Il capitolato poneva anche l'accento sulla qualità delle immagini, cioè sulla loro leggibilità. Quest'ultimo attributo non presentava una modalità di verifica standardizzabile con parametri oggettivi, pertanto la sua valutazione veniva affidata alla leggibilità, così come rilevata da operatore.

Ottenuto il parere favorevole dell'Aipa, è stata predisposta la gara per aggiudicare la stampa, l'allestimento, la consegna ai Comuni dei modelli (corrispondenti a circa 1,6 miliardi di pagine, incluse le guide e i materiali ausiliari), il ritiro dei modelli dai Comuni e la scansione ottica di modelli per circa mezzo miliardo di pagine. Tale elevato numero di pagine è determinato dalla tecnologia della lettura ottica, che per garantire standard qualitativi soddisfacenti necessita in ogni pagina di spazi e distanze adeguate fra gli oggetti da sottoporre a scansione (caselle da biffare, cifre e lettere dell'alfabeto da identificare), comportando la necessità di un progetto grafico complesso dei modelli di rilevazione e incrementando il loro numero di pagine rispetto al modello tradizionale.

La gara è stata vinta da un raggruppamento temporaneo di imprese costituito da una società tipografica, che curava gli aspetti di stampa, una ditta di trasporto per la consegna ed il ritiro dei modelli e da più società informatiche per la fase di acquisizione con lettura ottica e la video codifica delle variabili alfabetiche.

1.3.1. - Le fasi preliminari all'acquisizione dei dati con la lettura ottica

Il primo problema affrontato con il Fornitore è stato quello del progetto grafico dei modelli: tenuto conto che una parte del modello doveva essere stampata in colori "ciechi" alla lettura ottica (il rosso), si è cercato di

[2] OMR *Optical Mark Recognition* per il riconoscimento delle biffature e dei codici, anche a barre, prestampati; OCR/ICR, *Optical/ Intelligent Character Recognition* per i caratteri numerici ed alfabetici.

tenere nella massima considerazione l'accettabilità del Modello Istat CP.1 presso i rispondenti (colore, dimensione e tipo dei caratteri tipografici utilizzati, spaziatura delle linee eccetera), in modo da agevolare la corretta compilazione delle risposte da parte dei cittadini. Inoltre, particolare attenzione è stata posta nel definire la qualità, lo spessore, la grammatura della carta, le caratteristiche degli inchiostri e di tutte le tecnologie di stampa che meglio rispondevano all'obiettivo di ottimizzare la qualità dell'acquisizione tramite lettura ottica.

I modelli sono stati personalizzati a livello comunale - stampati completi di denominazione di provincia e di comune e di corrispondente codice comunale Istat (anche in formato di codice a barre) - sia per migliorare la qualità della lettura ottica, sia per evitare ai comuni l'onere dell'apposizione manuale di tali codici, oppure il possibile uso di timbri ed inchiostri non adeguati all'acquisizione ottica.

Poiché per la fase successiva di acquisizione con lettura ottica, nel momento della preparazione alla scansione, i modelli cartacei in fascicoli dovevano essere ridotti in fogli singoli (con un taglio sul dorso) per poter alimentare correttamente le apparecchiature di scansione, è stato necessario prestampare opportuni codici identificativi univoci su tutti i fogli del singolo modello in modo da render possibile la corretta ricostruzione del modello nella sua sequenza di pagine. La tecnologia tipografica utilizzata è stata quella dei codici a barre a 16 *digit*, ripetuti sui singoli fogli dello stesso modello e dei codici a barre a quattro *digit* specifici per identificare le pagine del modello.

Conclusa la fase di stampa dei modelli di uno stesso comune, si procedeva al loro confezionamento in scatole e alla consegna ai competenti Uffici di censimento, tramite la ditta di trasporto del Fornitore.

Terminata la fase di rilevazione e di revisione dei modelli da parte dei comuni, i responsabili degli Uffici di censimento procedevano alla preparazione dei pacchi di modelli, separati in base al tipo (modello di famiglia, di edificio, di convivenza, modelli ausiliari). Le scatole preparate erano contrassegnate da una distinta, che esplicitava il tipo e il numero di modelli contenuti, e da un codice identificativo di scatola, che consentiva la tracciabilità della scatola stessa nelle fasi successive. Alla data concordata la ditta di trasporto incaricata si recava presso il Comune a ritirare le scatole e quindi i pacchi venivano smistati ai magazzini dei centri di produzione (Modello Istat CP.1, Modello Istat CP.ED) per la lettura ottica e all'Istat (Modello Istat CP.2 ed altri tipi di modelli) per le ulteriori lavorazioni.

1.3.2. - I due centri di deposito, scansione e acquisizione

Per la fase di acquisizione tramite lettura ottica, il Fornitore aveva organizzato due "centri di produzione", uno a Piacenza e uno a Pomezia (Rm) nei quali erano suddivise le lavorazioni, rispettivamente, dei comuni del Centro-Nord e del Centro-Sud. Questa divisione, nelle intenzioni del Fornitore doveva tutelare da eventuali emergenze o criticità, per cui un centro di lavorazione poteva supportare l'altro nei momenti di necessità. A posteriori invece la suddivisione logistica nei due centri ha prodotto l'effetto indesiderato di convogliare principalmente i modelli compilati al Centro-Sud in un unico centro, che ha dovuto affrontare un compito relativamente più difficile per arrivare ai livelli di qualità desiderata. Nel Sud infatti la cura posta nella compilazione dei modelli, nella revisione e nel confezionamento dei pacchi per il trasporto è stata mediamente inferiore a quella evidenziatasi al Centro-Nord, con un impatto negativo sulla qualità complessiva dell'acquisizione.

All'arrivo dei modelli presso il centro di lavorazione il Fornitore effettuava fasi preliminari di verifica del contenuto delle scatole, avvertendo tempestivamente l'Istat di qualsiasi anomalia riscontrata (soprattutto difformità fra dichiarazioni in distinta e quantità di modelli rilevate effettivamente).

Seguiva la fase di preparazione alla scansione. Lavorando i pacchi di modelli dello stesso Comune, per gruppi di Comuni della stessa provincia, si operava il taglio del dorso dei modelli per consentire una adeguata alimentazione degli *scanner*, tarati per l'acquisizione delle immagini. Un apposito software consentiva di tradurre il risultato della compilazione (biffature, numeri e stringhe alfabetiche) in file di dati corrispondenti.

A causa della elevata dimensione dei file generati dalle immagini digitalizzate, la parte strutturale fissa dei modelli (e quella in colore rosso) veniva rimossa dall'immagine durante la fase di scansione e memorizzazione, ma poteva poi essere ricostruita nel momento della visualizzazione con il Sistema immagini, in dotazione all'Istat, tramite il software sviluppato dal Fornitore.

Le stringhe alfabetiche, generate dal processo di acquisizione ottica, subivano un pretrattamento con procedure automatiche di codifica e successivamente erano sottoposte a video-correzione manuale da parte di

operatori. I file di codici e dati ottenuti dal processo venivano quindi masterizzati su cd-rom e le immagini su dvd. Tali supporti erano consegnati all'Istat in blocchi di comuni della stessa provincia.

1.3.3. - I controlli sulle consegne dei dati e delle immagini acquisiti con la lettura ottica

Il controllo della qualità, il monitoraggio del contratto stipulato con il Fornitore, la verifica della qualità del processo, dei dati e delle immagini acquisite sono stati delegati in parte ad una società esterna di monitoraggio certificata (il cosiddetto "Monitore"). A questa società esterna sono stati affidati diversi compiti legati al controllo della qualità, sia per quanto riguarda gli aspetti relativi agli adempimenti contrattuali e documentativi della fornitura, sia relativamente agli aspetti più propriamente legati alla qualità dell'acquisizione dei dati. Per ciascuna consegna del Fornitore erano previste sia verifiche sulla conformità fra dati e immagini (controlli quantitativi), sia controlli sui livelli contrattuali di accuratezza dell'acquisizione e della codifica (controlli qualitativi). Questi ultimi venivano effettuati tramite *data entry* e codifica manuale di un campione di modelli, estratti dal blocco consegnato dal Fornitore, confrontando i propri risultati con quelli pervenuti col blocco.

Tavola 1.3 - Obiettivi del monitoraggio in relazione ai servizi previsti dal contratto per la lettura ottica

ATTIVITÀ MONITORATE	Conduzione del progetto	Processo del fornitore	Qualità del fornitore
Registrazione e deposito dei modelli ritirati	X	X	
Preparazione dei modelli alla scansione	X	X	
Scansione dei modelli	X	X	
Cattura dei dati	X	X	
Codifica delle variabili alfabetiche	X	X	
Gestione delle eccezioni nella fase di acquisizione	X	X	
Procedura di monitoraggio del Fornitore	X	X	X
Approntamento dell'output (dati, immagini)	X	X	X
Consegna del materiale di output		X	
Sistema di gestione delle immagini	X	X	X

1.3.4. - I controlli preliminari e i controlli a campione

Obiettivo del controllo di qualità era quello di verificare che il materiale consegnato dal Fornitore fosse conforme ai livelli di qualità definiti nel contratto. La consegna di dati e immagini avveniva in blocchi, memorizzati su supporti separati e omogenei per provincia (159, più un blocco di *over-flow* contenente dati e immagini residuali o di recupero).

Per ciascuna consegna, dovevano essere sottoposti ai controlli di qualità da parte del Monitore sia i dati che le immagini dei questionari, in due passi successivi:

- controlli preliminari esaustivi sull'intero blocco (controlli quantitativi);
- controlli su un campione del blocco (controlli qualitativi).

Per ogni blocco consegnato il Monitore provvedeva ai controlli preliminari, che consistevano in una serie di operazioni di verifica sulle immagini contenute nei dvd e sui dati nei cd-rom. In primo luogo si procedeva al caricamento dei dvd, contenenti le immagini dei modelli acquisiti otticamente nel Sistema immagini e all'esecuzione dei controlli di leggibilità e di corrispondenza del contenuto dei supporti con quanto dichiarato dal Fornitore sulle rispettive etichette. Quindi i dati contenuti nei cd-rom venivano sottoposti ad elaborazione per il controllo quantitativo (in termini di numero di record per provincia, per comune e per tipologia di modello, di sequenza dei record generati dai Fogli di famiglia, della corretta sequenza delle pagine del modello). Veniva inoltre verificata la corrispondenza tra i dati e le immagini consegnate (presenza, completezza e non ridondanza reciproca di dati e immagini, verifica della coerenza dei codici di provincia e comune presenti sui record dei Fogli di famiglia e dei Questionari di edificio con i codici presenti nei record generati dalla registrazione delle distinta di scatola eccetera). Tutti i casi di assenza, incompletezza o duplicazione di dati o immagini venivano evidenziati dai controlli in una apposita lista, con segnalazione dei codici identificativi dei modelli interessati dall'anomalia.

Inoltre, in sede di controllo preliminare, il Monitore effettuava una serie di conteggi sul numero di caratteri registrati relativi a ciascuna tipologia di variabile specificata nel contratto, sul numero di campi alfabetici riempiti e sulla presenza della relativa codifica. Queste elaborazioni consentivano di pervenire alla quantificazione dell'universo di riferimento per l'estrazione del successivo campione di controllo e di valutare a priori il livello di assegnazione dei codici delle variabili quali *comune*, *stato estero* e *titolo di studio*. In presenza di stringhe alfabetiche, livelli non sufficienti di assegnazione dei codici rispetto al contratto potevano comportare il rifiuto immediato del blocco.

Nel caso in cui il blocco nella sua interezza non avesse superato la verifica preliminare, veniva rimandato dall'Istat al Fornitore per le operazioni di adeguamento della qualità, senza procedere ad ulteriori esami.

Se invece il blocco esaminato superava la verifica preliminare, si poteva passare alla fase successiva dei controlli qualitativi a campione, sui livelli di accuratezza del riconoscimento dei caratteri e di codifica. Per ciascun blocco il Monitore procedeva alla estrazione di un campione di (codici identificativi di) questionari da controllare, suddiviso in tre strati (Fogli di famiglia contenenti almeno un componente, Fogli di famiglia relativi ad abitazioni non occupate, Questionari di edificio), utilizzando le specifiche metodologiche fornite dall'Istat, tali da assicurare la significatività dei test statistici per tutte le tipologie di controlli da effettuare.

Solo per il controllo della codifica del campo *stato estero*, compilato prevalentemente dalla sub-popolazione degli stranieri, si identificavano a priori Fogli di famiglia in cui tale campo fosse compilato e si procedeva all'estrazione di un campione separato di modelli, in cui la stringa alfabetica *stato estero* fosse presente.

Il Monitore provvedeva, con proprio personale presso l'Istat ed utilizzando il "Sistema informatico di gestione delle immagini", alla estrazione dei dati dai cd-rom e delle immagini dai dvd corrispondenti ai codici selezionati nei passi precedenti. Procedeva quindi ai controlli sulla qualità dei dati e delle immagini del campione, con proprio personale e nella propria sede, sottoponendo ad acquisizione tramite *Kfi* (*Key From Image*)[3] le immagini estratte ed effettuando la codifica delle variabili *comune*, *stato estero* e *titolo di studio*, secondo la classificazioni e i dizionari forniti dall'Istat. A partire dai dati di controllo effettuava l'abbinamento, tramite i codici identificativi, con i record del Fornitore, identificava e verificava le discordanze, conteggiava le percentuali di errore per ciascun tipo di carattere/codifica provvedendo al calcolo delle stime dei parametri di qualità.

Il valore riscontrato nel campione (soglia operativa di accettazione/rifiuto) che discriminava fra esito positivo e negativo dei controlli, non era esattamente uguale al parametro/soglia contrattuale (Tavola 1.4), ma si poneva leggermente al di sotto dei livelli di qualità esplicitati nel capitolato, per tener conto della dimensione del campione di controllo e dei parametri ad esso collegati (come gli errori di prima e di seconda specie). Ad esempio, pur essendo da contratto il livello minimo di accuratezza della codifica delle *stato estero* uguale al 98 per cento, un valore pari o superiore al 97,043 per cento portava comunque ad una accettazione della lavorazione relativamente a quel controllo.

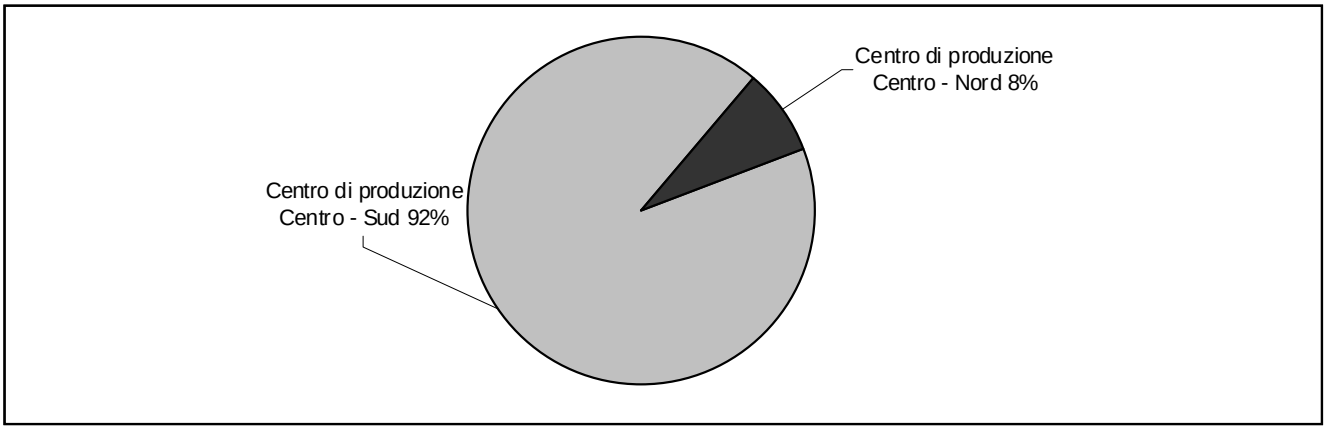
Qualora i risultati dei controlli rilevassero livelli di qualità inferiori a quelli stabiliti, l'Istat restituiva l'intero blocco al Fornitore affinché fosse rielaborato. Sulle successive consegne di materiale rielaborato il Monitore ripeteva daccapo tutti i controlli di qualità.

Tavola 1.4 - Soglie operative di accuratezza e di errore che comportavano la decisione di rifiuto dei blocchi di dati per tipologia di variabile sottoposta a controllo di qualità

TIPI DI VARIABILE	Soglia di errore nel capitolato	Soglia operativa di errore	Soglia di accuratezza nel capitolato	Soglia operativa di accuratezza
Campi qualitativi	0,70	1,033	99,30	98,966
Campi quantitativi	2,00	2,956	98,00	97,043
Data di nascita	0,50	0,738	99,50	99,261
Sezione Censimento	0,50	0,738	99,50	99,261
Comune	1,00	1,477	99,00	98,523
Stato estero	2,00	2,956	98,00	97,043
Titolo di studio	2,00	2,956	98,00	97,043

[3] I dati vengono registrati dagli operatori sulla base delle immagini visualizzate a schermo.

Figura 1.2 - Blocchi rifiutati per centro di produzione *(composizione percentuale)*



I blocchi accettati sono stati complessivamente 160. Le consegne effettuate dal Fornitore sono state pari a 254, i rifiuti che hanno comportato il rifacimento del blocco sono stati 75 (Tavola 1.5), mentre nella fase iniziale di test in 19 casi i blocchi consegnati non sono risultati adeguati e hanno richiesto una revisione dalla fornitura, ad esempio perché i blocchi si rivelavano incompleti o inadeguati già da una verifica preliminare.

Tavola 1.5 - Blocchi arrivati, rifiutati e accettati per centro di produzione

CENTRI DI PRODUZIONE	Numero di blocchi arrivati	Numero di rifiuti	Percentuale di rifiuti (per centro)	Numero di blocchi accettati
Centro di produzione Centro Nord	101	6	8,0	86
Centro di produzione Centro Sud	153	69	92,0	74
Totale	**254**	**75**	**100,0**	**160**

Tavola 1.6 - Blocchi arrivati, rifiutati o accettati per centro di produzione e decade di consegna *(valori assoluti)*

DECADI DI CONSEGNA	Centro di produzione Centro-Sud			Centro di produzione Centro-Nord		
	Blocchi arrivati	Rifiuti	Blocchi accettati	Blocchi arrivati	Rifiuti	Blocchi accettati
2ª – luglio 2002	0	0	0	3	0	3
3ª - luglio	1	0	1	2	0	2
1ª - agosto	1	0	1	4	0	4
2ª - agosto	0	0	0	5	0	5
3ª - agosto	7	6	1	12	0	12
1ª - settembre	16	9	7	11	2	9
2ª - settembre	2	1	1	11	1	10
3ª - settembre	12	4	8	12	0	12
1ª - ottobre	26	11	15	10	1	9
2ª - ottobre	28	15	13	19	1	18
3ª - ottobre	11	9	2	0	0	0
1ª - novembre	3	3	0	1	1	0
2ª - novembre	11	8	3	1	0	1
3ª - novembre	7	1	6	1	0	1
1ª - dicembre	5	0	5	1	0	1
2ª - dicembre	11	1	10	0	0	0
3ª - dicembre	2	0	2	0	0	0
1ª - gennaio 2003	1	0	1	0	0	0
Totale	**143**	**69**	**74**	**92**	**6**	**86**

In 60 casi i blocchi consegnati all'Istat sono stati rifiutati dopo il controllo a campione con *Kfi*. In 37 casi le motivazioni del rifiuto erano dovute alla insufficiente qualità di una singola variabile (soprattutto la codifica del *comune*), mentre in altri casi i parametri contrattuali di qualità erano violati a causa di più di una variabile, (*comune*, *titolo di studio, data di nascita*). Considerando nell'insieme i 98 motivi di rifiuto quindi, 48 erano imputabili alla codifica del *comune*, 25 alla codifica del *titolo di studio* e 19 alla *data di nascita*.

Nel corso delle lavorazioni la variabile per la quale si sono riscontrate le maggiori difficoltà nel raggiungimento dei parametri qualitativi contrattuali è stata il *titolo di studio,* in quanto il Fornitore non riusciva quasi mai a garantire livelli di accuratezza sufficienti (98 per cento da contratto) per l'accettazione del blocco e di fatto consegnava blocchi in cui il livello medio di accuratezza era circa del 96 per cento, pur ottenendo livelli di assegnazione superiori a quelli contrattuali (90-92 per cento contro l'80 per cento fissato nel capitolato).[4]

Una modifica tecnica sui criteri di valutazione, che premiava il maggior livello di assegnazione raggiunto a discapito di una lieve diminuzione della qualità di accuratezza, ha consentito di ottenere risultati di qualità anche per questa variabile, la cui codifica si presentava piuttosto complessa.

Tavola 1.7 - Blocchi rifiutati per dimensione della qualità che ha determinato il rifiuto

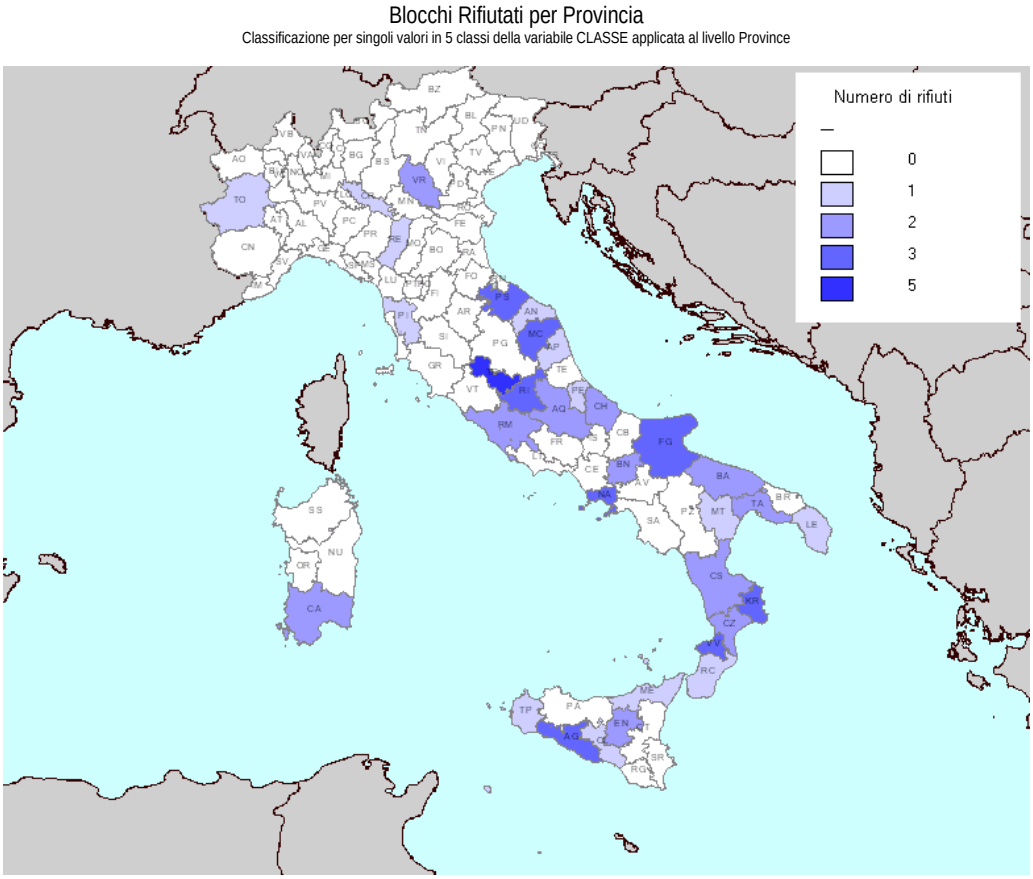
	Dimensione della qualità che ha determinato il rifiuto							
	Variabili qualitative	Variabili quantitative	Data di nascita	Sezione di censimento	Comune	Stato estero	Titolo di studio	Totale dei motivi
Rifiuti per motivo	1	1	19	2	48	2	25	98
Percentuale sul totale motivi di rifiuto	1,02	1,02	19,39	2,04	48,98	2,04	25,51	100,00

Tavola 1.8 - Livelli di qualità finale effettivi, in termini di mancata accuratezza, ottenuti con la lettura ottica e la codifica automatica

LIVELLO DI ERRORE	Variabili qualitative	Variabili quantitative	Data di nascita	Sezione di censimento	Comune	Stato estero	Titolo di studio
Previsto	0,7	2,0	0,5	0,5	1,0	2,0	2,0
Effettivo	0,125	0,756	0,359	0,141	0,637	0,612	3,070

[4] Nel capitolato tecnico, era previsto che il livello minimo di assegnazione del codice del campo "titolo di studio" (cioè il rapporto fra il numero di volte in cui era presente il valore testuale sul questionario ed il numero di codici presenti) fosse uguale all'80 per cento, a fronte di un livello minimo di accuratezza (cioè il rapporto fra il numero di variabili "titolo di studio" codificate correttamente e sul totale di quelle codificate) del 98 per cento.

Figura 1.3 - Blocchi acquisiti otticamente rifiutati al controllo di qualità per provincia



1.3.5. - Alcuni errori sistematici propri della fase di acquisizione: i "record impropri"

Uno degli aspetti problematici della lettura ottica è rappresentato dai possibili errori generati da una non corretta procedura di scansione delle immagini. Tali errori possono derivare da una serie di eventi o di fattori (Istat, 2006a):

- il carattere sul modello (vale principalmente per le parole, le lettere e i numeri scritti a mano) si discosta troppo, nella forma, dagli standard riconosciuti di scrittura e il motore di riconoscimento ottico lo identifica in modo errato;
- l'immagine del carattere è di bassa qualità e non risulta nitida al motore di riconoscimento;
- si sta sottoponendo a riconoscimento un'area della pagina che non contiene la variabile attesa;
- sono stati tracciati segni involontari su aree da riconoscere;
- la carta contiene impurità e punti di colore che possono dare luogo a interpretazione di caratteri fittizi.

Sono stati soprattutto gli ultimi due fattori ad influenzare la qualità della lettura ottica dei Fogli di famiglia del Censimento (Modello Istat CP.1). Questo modello era costituito da un fascicolo in cui le prime due pagine erano destinate ai quesiti sull'abitazione, seguito da insiemi di sei pagine destinati a contenere le risposte di ciascuno dei membri residenti (due o cinque rispondenti per modello) e quattro pagine dedicate ai temporaneamente dimoranti. L'acquisizione dei dati del Modello Istat CP.1 generava diversi tipi record, ad esempio un tipo record per la sezione abitazione, un tipo record per le persone residenti e un altro per le persone temporaneamente dimoranti. Le regole informatiche che descrivevano le modalità di esportazione dei dati su file, secondo tracciati record prestabiliti, prevedevano che, se le pagine relative al singolo individuo fossero state vuote, nessun record corrispondente sarebbe stato creato in output; al contrario, la presenza anche di una sola risposta nella sezione del questionario, riferita ai singoli componenti delle famiglia, determinava la creazione di un record specifico.

In alcuni casi è accaduto che un segno erroneamente tracciato con la penna su un qualunque quesito del questionario o un'impurità del foglio o la presenza di un qualsiasi campo valorizzato (biffatura, valore numerico o stringa) venissero letti dallo *scanner* come risposte effettivamente fornite, generando record individuali in cui però non era presente nessuna ulteriore informazione. In qualche caso si è osservato che, con eccesso di zelo, ritenendo magari di aver commesso qualche errore nel corso della compilazione del proprio questionario, i rispondenti avevano "abbandonato" le pagine già riempite, ricominciando nel gruppo di pagine destinato al componente successivo. Una sola "falsa" o "doppia" risposta acquisita per un questionario individuale (non compilato in altre parti) produceva comunque in output un record, in cui l'unico campo valorizzato conteneva un risultato falso. I record così erroneamente generati definivano individui che in realtà non erano parte della popolazione censita.

La gestione del problema delle false risposte è stata affrontata nella delicata fase dei controlli quantitativi. Nella fase di revisione e controllo relativa ai conteggi per la determinazione della popolazione legale si sono stabiliti i criteri per eliminare i record "impropri", creati erroneamente dal processo di scannerizzazione, al fine di conteggiare correttamente la popolazione legale comunale.

In presenza di record quasi totalmente vuoti, in cui non fossero indicate tutte le variabili demografiche, si è provveduto alla cancellazione del record stesso. In particolare, si è proceduto all'eliminazione di tutti record individuali che, contemporaneamente, non presentavano nessuna delle seguenti variabili: *sesso, giorno/mese/anno di nascita, luogo di nascita, cittadinanza, stato civile, mese/anno di matrimonio.*

Molta attenzione è stata posta inoltre a queste tipologie di record impropri nella fase di controllo quantitativo per la quale ci si è avvalsi dei Modelli ausiliari Istat CP.9 (Stato di sezione) e Istat CP.10 (Riepilogo di sezione). In alcuni casi si è proceduto anche alla consultazione delle immagini digitalizzate dei modelli.

Proprio dalla visualizzazione delle immagini si è compreso come, in molti casi, la creazione di record "impropri" fosse derivata da impurità, che presumibilmente si trovavano in uno specifico punto della superficie dello *scanner* e che determinava la valorizzazione di qualche variabile (quasi sempre la stessa per tutti i componenti del nucleo familiare) e il conseguente prodursi di uno o più record impropri da eliminare.

Problemi analoghi a quelli sin qui analizzati si sono verificati anche per i record impropriamente generati per le persone non abitualmente dimoranti, sottoposti anch'essi a correzioni automatiche e manuali.

Alcuni errori sistematici, dovuti alla fase di acquisizione, sono stati rilevati sia durante i controlli del Monitore, sia con l'analisi dei dati durante il processo di validazione. Gli errori di lettura ottica si sono rilevati soprattutto nei campi numerici, non correttamente interpretati in presenza di annerimenti o "sporcizie": alcune volte campi che dovevano essere letti come vuoti sono stati restituiti come "1" o "0". In particolare nei Questionari di edificio i numeri rilevati per alcuni campi numerici, compilati a mano dal rilevatore, presentavano dei picchi di frequenza ai numeri 10, 11, 101, eccetera.

1.4 - L'acquisizione dei dati con la registrazione tradizionale (non a lettura ottica)

Oltre ai modelli direttamente necessari alla rilevazione (delle famiglie, delle convivenze, degli edifici), sono stati utilizzati anche altri modelli, i cosiddetti "modelli ausiliari", utilizzati dagli uffici di censimento per tenere sotto controllo le operazioni sul campo e per verificare, nei giorni della rilevazione, che i risultati che si stavano accumulando fossero consistenti.

Alcuni di questi modelli, strumentali al controllo effettuato dai Comuni e alla definizione di dati consuntivi, sono stati fondamentali anche per le successive fasi di validazione dei dati.

Inoltre, dato che alcuni Comuni, soprattutto quelli di maggiore dimensione, avevano richiesto all'Istat, di poter consegnare alcuni modelli di famiglia (Modello Istat CP.1) in data successiva a quella prevista per il ritiro (già avvenuto a cura del Fornitore per i modelli destinati alla lettura ottica), si è concessa a questi una dilazione di qualche mese. I modelli cosiddetti integrativi, non potendo più essere trattati nel flusso della lettura ottica, sono stati raccolti presso l'Istat (circa 60 mila Modelli Istat CP.1, sia modelli base che aggiuntivi, più qualche centinaio di Modelli Istat CP.2 di convivenza), per essere avviati all'acquisizione tramite la registrazione tradizionale.

Completata la rilevazione censuaria i modelli non destinati alla lettura ottica sono stati confezionati dai Comuni ed inviati direttamente all'Istat, in scatole denominate CPMIX contenenti i Fogli di convivenza

(Modello Istat CP.2) ed altri modelli ausiliari di censimento (Istat CP.9, Istat CP.10, Istat CP.5, Istat CP.6, Istat CP6.ED). Le scatole CPMIX pervenute in Istat sono state 10.874 in 24 consegne.

Presso l'Istat è stato costituito un Ufficio ricezione, con il compito di monitorare e supervisionare le operazioni di magazzino inerenti la ricezione dei modelli ausiliari cartacei e/o informatizzati e dei Fogli di convivenza CP.2 pervenuti dai Comuni, da archiviare per le successive lavorazioni. Una volta arrivate in Istat, le scatole sono state sottoposte a verifica di conformità numerica con la distinta di trasporto acclusa.

Il compito di apertura e di controllo delle scatole di tipo CPMIX è stato affidato ad un gruppo di revisori, i quali hanno provveduto ad una prima verifica della conformità numerica dei modelli di censimento, rispetto a quanto dichiarato dal comune in apposita distinta, e alla suddivisione e conteggio dei modelli di diverso tipo, prima della loro archiviazione o del loro invio in registrazione.

1.4.1 - I modelli Istat CP.2 e CP.1 integrativi – registrazione con data entry

I Fogli di convivenza (Modello Istat CP.2) pervenuti direttamente all'Istat sono stati tutti sottoposti a revisione manuale (in particolare per controllarne i codici identificativi) ed integrati con i modelli delle convivenze militari, compilati a cura del Ministero della Difesa.

La registrazione dei Modelli Istat CP.2 è stata eseguita da una ditta specializzata di *data entry,* alla quale era richiesta contrattualmente una qualità, in termini di battute errate su battute utili, pari al 6 per mille.

Sono stati effettuati 18 invii alla registrazione tradizionale, per circa 90 mila Modelli Istat CP.2, di cui uno riguardante esclusivamente i Modelli Istat CP.2 integrativi.[5]

I Modelli Istat CP.2 sono stati registrati su file distinti per provincia, pertanto, in corrispondenza ad ogni invio, la ditta di registrazione restituiva tanti file quante erano le province incluse. Il controllo della qualità contrattuale sui dati registrati relativamente al numero di battute conteggiate e alla percentuale di battute errate (Tavole 1.9 e 1.10) è stato eseguito all'interno dell'Istat.

Prima del caricamento dei dati nel database di produzione è stato eseguito, a cura dell'Ufficio ricezione, il controllo quantitativo, al fine di individuare eventuali discordanze tra il numero di Modelli Istat CP.2 inviati alla registrazione *in-service* e quelli effettivamente registrati dalla ditta incaricata. Tale controllo ha consentito, tra l'altro, di individuare errori di assegnazione dei codici territoriali (provincia e comune) eventualmente commessi dalla ditta di registrazione.

Tavola 1.9 - Errori della registrazione dei Fogli di convivenza (Mod. Istat CP.2) riscontrati con i controlli di qualità a campione *(valori assoluti e percentuali)*

PROGRESSIVI DI INVIO IN REGISTRAZIONE	Modelli Istat CP.2 inviati	Battute conteggiate	Percentuale di battute errate (per mille)
1	1.484	1.264.464	4,92
2	8.396	7.138.678	0,68
3	7.252	6.344.421	4,85
4	6.749	5.853.613	2,47
5	7.375	6.319.786	1,94
6	10.213	9.471.277	0,17
7	7.760	7.599.240	0,30
8	5.888	5.701.029	1,68
9	5.420	4.888.026	0,10
10	4.940	5.046.776	0,39
11	4.441	4.220.785	0,90
12	2.284	2.317.816	6,85
13	3.027	3.030.155	0,32
14	5.762	5.688.180	6,38
15	2.257	2.313.992	0,50
16	6.055	6.082.555	1,37
17	334	301.563	3,64
18	199	206.171	1,88

[5] Sono quei modelli di convivenza che sono stati rinvenuti in scatole differenti da quelle CPMIX.

Tavola 1.10 - Errori della registrazione *in-service* dei Fogli di famiglia (Modello Istat CP.1) integrativi riscontrati con i controlli di qualità a campione *(valori assoluti e percentuali)*

PROGRESSIVI DI INVIO	Modelli CP.1 integrativi inviati	Modelli CP.1 aggiuntivi integrativi inviati	Battute conteggiate	Battute errate (per mille)
1	4.333	688	2.924.179	0,55
2	4.577	1.076	2.927.888	1,06
3	13.513	2.319	10.090.151	0,26
4	12.327	5.484	8.719.536	0,58
5	13.243	3.437	8.911.347	0,25
6	1.986	386	1.319.568	0,59
7	86	56	74.129	0,32
8	1.234	1.752	1.208.933	0,42

1.5 - Le fonti di errore nella compilazione: le caratteristiche dei rispondenti

La rilevazione censuaria prevede che siano acquisite informazioni sulla popolazione residente, sia in famiglia che in convivenza, e sulla popolazione presente alla data di censimento, sugli alloggi, sia occupati che non occupati e, per la prima volta nel 2001, anche sugli edifici.

La complessità della rilevazione richiede l'utilizzo di due diversi questionari per gli individui (uno per le famiglie ed uno per le convivenze). I modelli di rilevazione sono suddivisi in sezioni riservate alle diverse tipologie di unità rispondenti (residenti e non residenti). Nel questionario per le famiglie sono contenute anche le domande relative all'abitazione.

A ciascun tipo di modello corrisponde, in generale, un diverso compilatore su cui far ricadere la responsabilità delle risposte, quindi la qualità può essere riferita ed analizzata a partire dalle sue caratteristiche individuali (età, livello di istruzione, eccetera).

La persona che deve rispondere al modello specifico per le famiglie (Modello Istat CP.1) è l'intestatario del Foglio di famiglia (cioè la persona a cui è intestata la scheda di famiglia in Anagrafe). È questi che dovrebbe fornire le risposte ai quesiti relativi ai componenti della famiglia, alle caratteristiche dell'abitazione, ma anche alle domande che riguardano le persone temporaneamente o occasionalmente dimoranti. Anche se non è escluso che alcune risposte vengano fornite anche da altri componenti della famiglia, è di norma all'intestatario che si può attribuire il livello di qualità delle risposte del Foglio di famiglia.

Per quanto riguarda le convivenze è previsto che a rispondere siano i responsabili della convivenza, i quali devono fornire informazioni tanto sui residenti quanto sui non residenti.

Nel caso di abitazione non occupata (vuota) è il rilevatore ad essere incaricato di reperire le informazioni sulle caratteristiche dell'abitazione, ad esempio chiedendo ai vicini di casa. In questo caso la qualità delle risposte fornite è attribuibile alla scrupolosità del rilevatore nel reperire le notizie e nel compilare il Foglio di famiglia.

Ulteriore compito del rilevatore è la compilazione del Modello Istat CP.ED dedicato agli edifici e alle caratteristiche della costruzione. Infine un attore importante nel definire il livello di qualità è l'Ufficio comunale di censimento (Ucc), al quale sono assegnati i compiti di supervisione della rilevazione e di revisione dei diversi modelli.

Una analisi delle caratteristiche del rispondente e della sua "propensione" all'errore nella compilazione del modello è stata tentata con approcci successivi, allo scopo di determinare quale sia il profilo del rispondente che ha commesso un minor numero errori di compilazione. Elaborazioni esplorative sono state condotte per valutare la relazione fra gli errori per modello e le caratteristiche dell'intestatario del Foglio di famiglia.

Il numero di errori per modello è stato calcolato considerando il numero di differenze, variabile per variabile, dei dati presenti nei record iniziali o grezzi e i record finali, così come risultano dopo l'esecuzione delle procedure di correzione: l'assenza di eventi di correzione implica che i dati sono stati ritenuti esatti per quel modello e quindi la somma degli errori risulta nulla. Il totale di errori per modello (o meglio il totale degli errori per i record di famiglia relativi ai residenti) è influenzato dal numero di record (componenti) della famiglia e anche dal numero di risposte dovute o attese, dipendendo queste ultime dalle caratteristiche degli

individui (lavora o no, effettua spostamenti quotidiani per lavoro o per studio, è o meno cittadino straniero). La somma degli errori è difficilmente standardizzabile rispetto a queste dimensioni.

Per semplificare, invece di considerare la distribuzione del numero complessivo di errori, l'analisi si è concentrata sui modelli compilati esattamente (nessun errore rilevato della procedure e quindi nessun cambiamento apportato ai dati) quindi sui modelli con un totale di errori uguale a zero.

Prendendo in considerazione i Modelli Istat CP.1 (circa 21 milioni, relativi alle abitazioni occupate da almeno una persona residente) è stata osservata la distribuzione per età (Etac) degli intestatari complessiva (Figura 1.4) rispetto alla distribuzione per età degli intestatari che hanno compilato correttamente (Figura 1.5), calcolando il rapporto di incidenza dei secondi sui primi (Figura 1.6), cioè per ogni anno di età la percentuale di "migliori" compilatori sul totale degli intestatari di quell'età. Dalla figura 1.6 si evidenzia una più elevata percentuale di buona compilazione per i giovani 20-30 anni, ma anche per gli ultracinquantenni, mentre fra 30 e 50 anni è minore la percentuale di intestatari che non hanno commesso errori di compilazione.

Nella figura 1.7, in cui è rappresentata l'incidenza del fenomeno per numero di componenti (Ncomp), si osserva come l'assenza di errori di compilazione sia prerogativa del 33 per cento degli intestatari di famiglie unipersonali e del 18 per cento delle famiglie di due componenti e solo del 4 per cento delle famiglie di tre.

Figura 1.4 - Distribuzione per età degli intestatari del Modello Istat CP.1

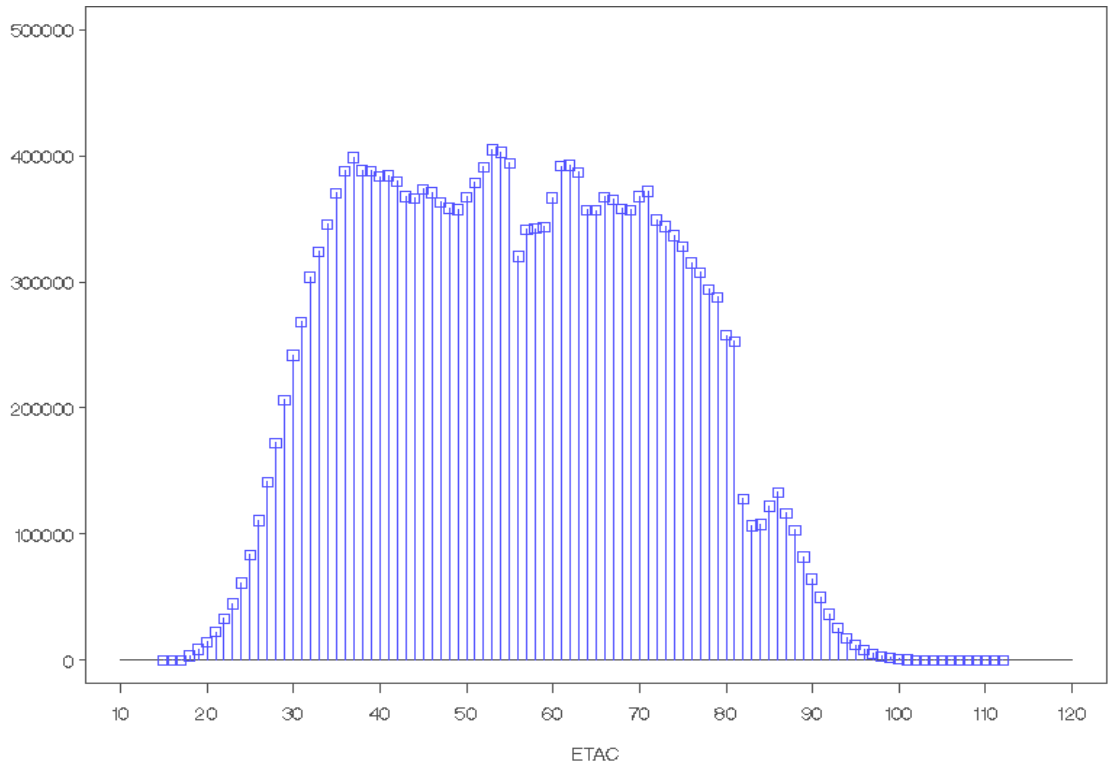


Figura 1.5 - Distribuzione per età degli intestatari che hanno compilato correttamente il Modello Istat CP.1 per la parte relativa agli individui residenti in famiglia

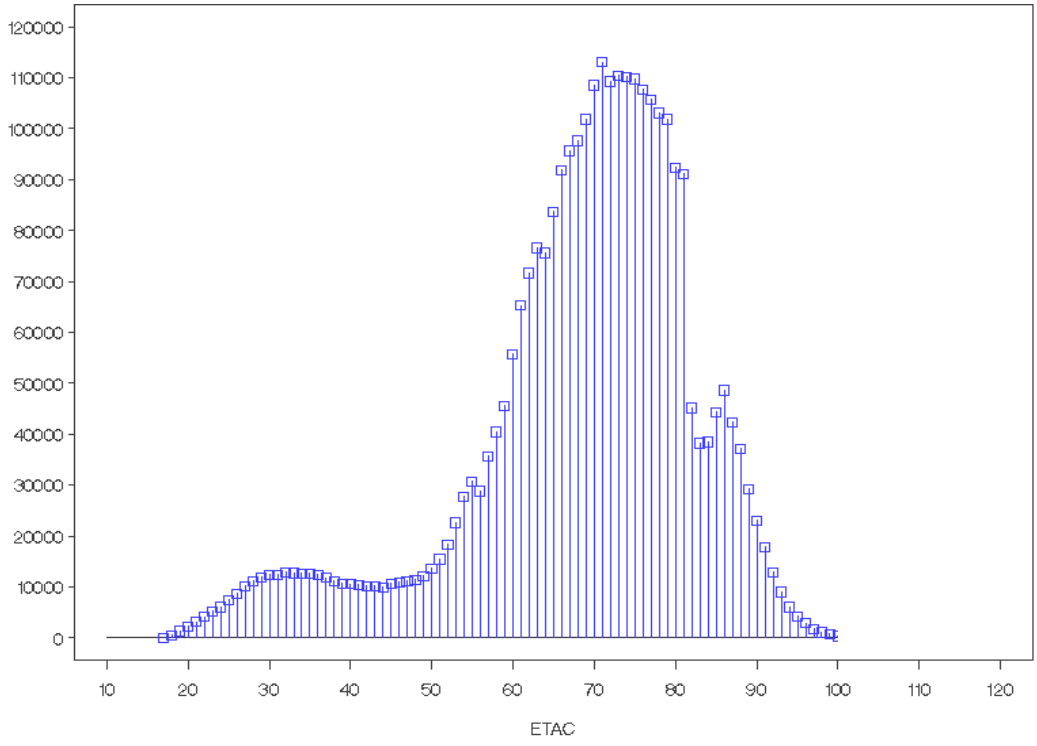


Figura 1.6 - Incidenza degli intestatari che hanno compilato correttamente il Modello Istat CP.1 per la parte relativa agli individui residenti, sul totale intestatari per anno di età

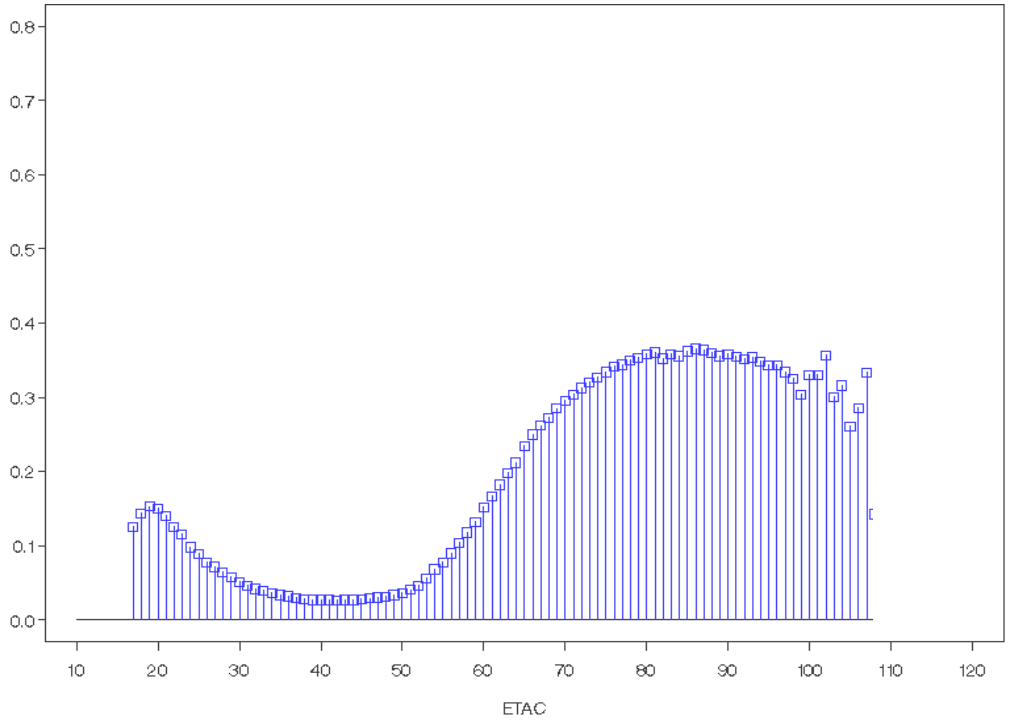


Figura 1.7 - Incidenza degli intestatari che hanno compilato correttamente il CP.1 sul totale degli intestatari per numero di componenti della famiglia

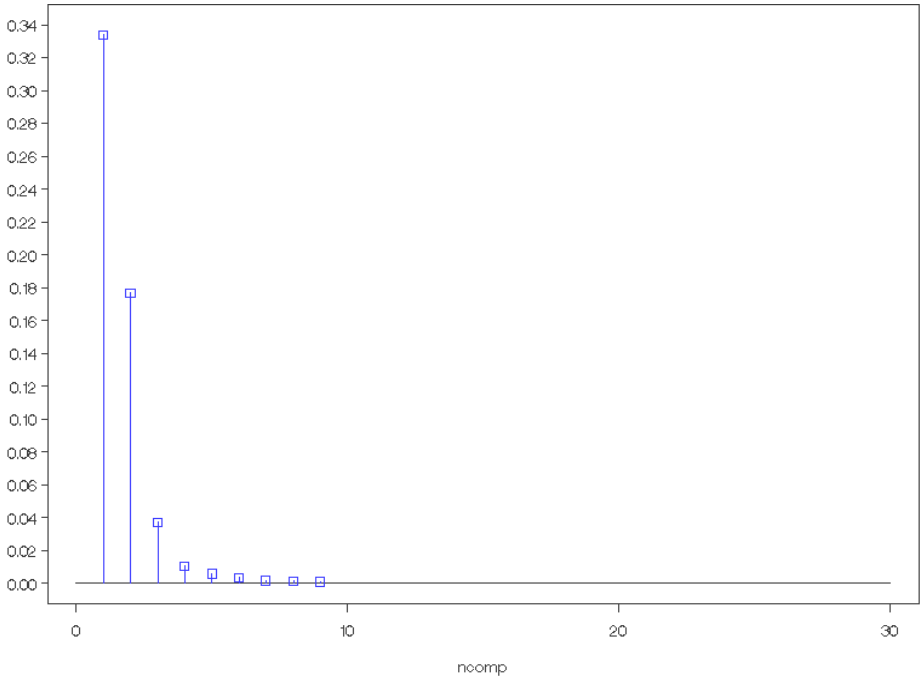


Figura 1.8 - Incidenza degli intestatari che hanno compilato correttamente il CP.1 sul totale degli intestatari per numero di risposte dovute

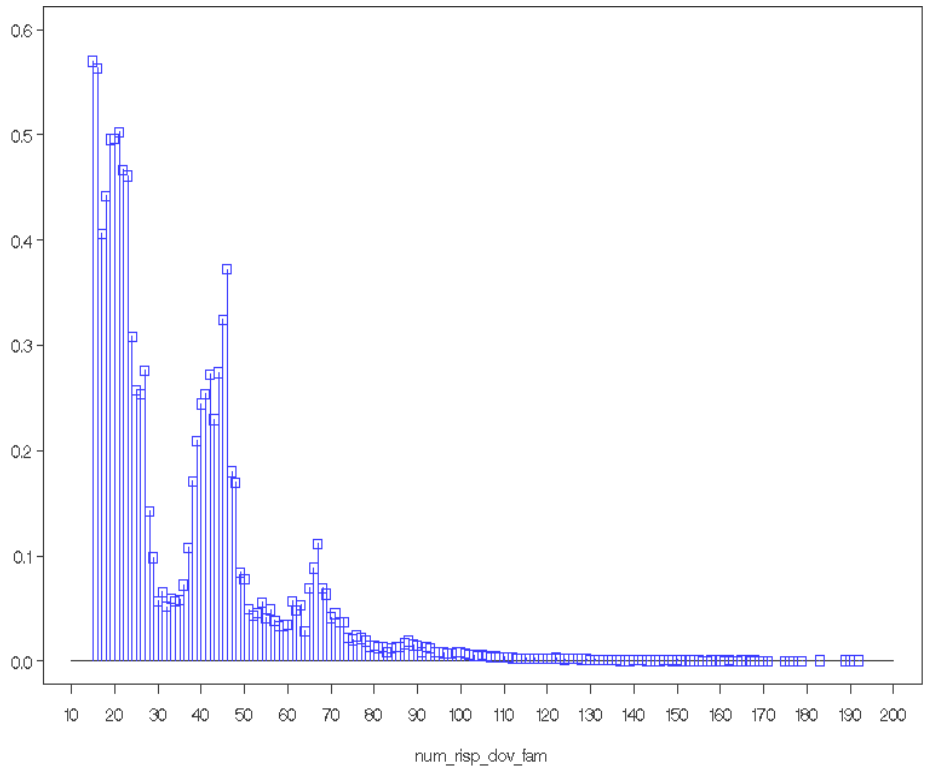
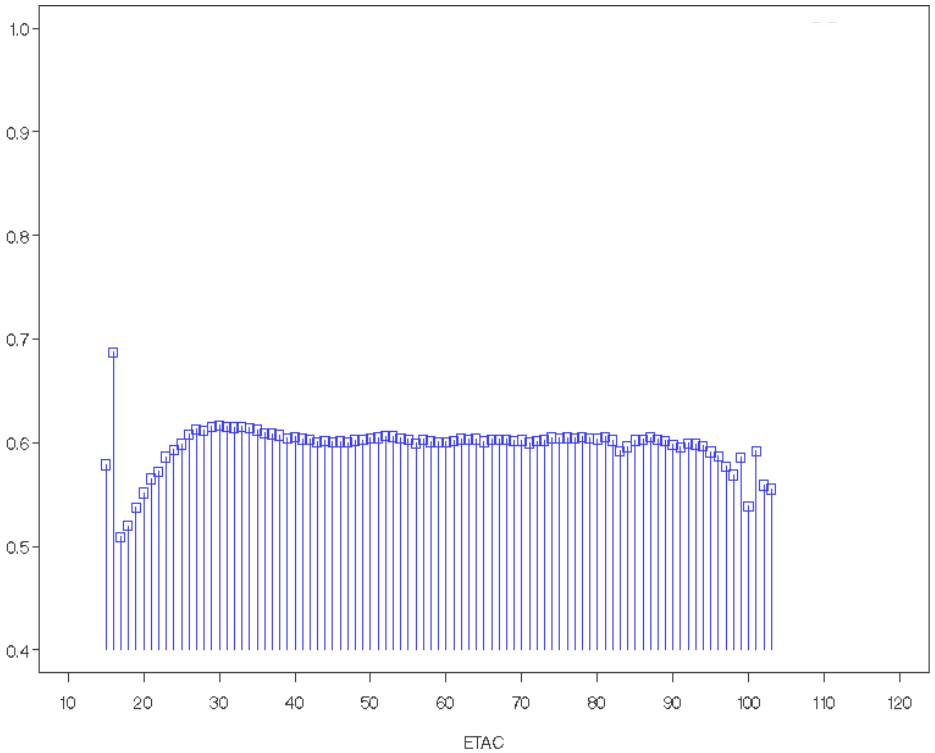


Figura 1.9 - Incidenza degli intestatari che hanno compilato correttamente il Modello Istat CP.1 per la sezione sull'abitazione occupata sul totale intestatari per anno di età



Anche l'effetto del numero di quesiti cui l'intestatario è chiamato a rispondere è valutabile graficamente (Figura 1.8), ma con un impatto meno diretto rispetto al numero di componenti: in funzione delle risposte "dovute" si evidenziano gruppi di eccellenza fino a 30 risposte e fra i 40 e 50, mentre per altre numerosità di risposte la qualità (misurata come assenza di errori) sembra appiattirsi ai livelli inferiori.

Questo risultato è collegato alle difficoltà nella compilazione di specifiche sezioni del questionario (gruppi di quesiti) che può portare a commettere uno o più errori. Per esempio chi ha risposto (per sé e per i familiari) solo alle sezioni "base" sui dati anagrafici è riuscito più facilmente a non commettere errori, mentre gli occupati, che devono rispondere in dettaglio alla sezione lavoro e/o alla sezione sul pendolarismo, devono seguire correttamente un maggior numero di percorsi del questionario e hanno una maggiore possibilità di commettere errori (mancate risposte, risposte non dovute, incoerenze).

Per eliminare gli effetti indiretti dovuti al numero di componenti e al numero di risposte dovute si è presa in considerazione la sezione relativa all'abitazione (occupata). La sua compilazione è infatti indipendente dalla dimensione della famiglia e dal numero di sezioni cui l'intestatario deve rispondere. La relazione fra età degli intestatari e la bontà delle risposte fornite è molto meno netta che quella evidenziata in precedenza e la figura 1.9 indica che solo fra i 20 e 30 anni e dopo i 90 la propensione all'errore è più alta. Per i molto anziani la frequenza del fenomeno è poco significativa, anche se si può pensare che la minore incidenza dell'errore sia dovuta all'aiuto di altre persone (i familiari o il rilevatore) che hanno fornito assistenza alla compilazione.

Tavola 1.11 - Modalità di compilazione del Modello Istat CP.1 per classi di età dell'intestatario *(valori assoluti e percentuali)*

ETÀ DELL'INTE-STATARIO	Il modello è stato compilato:							
	Senza aiuto del rilevatore		Con l'aiuto del rilevatore		D'ufficio		Totale (a)	
	Valori assoluti	Percentuale	Valori assoluti	Percentuale	Valori assoluti	Percentuale	Valori assoluti	Percentuale
15-20	11.466	57,81	7.207	36,33	1.162	5,86	19.835	100,00
21-30	638.846	68,33	268.165	28,68	27.885	2,98	934.896	100,00
31-40	2.248.112	71,95	820.377	26,26	55.981	1,79	3.124.470	100,00
41-50	2.337.441	71,95	866.333	26,67	44.850	1,38	3.248.624	100,00
51-60	2.284.429	70,64	912.888	28,23	36.613	1,13	3.233.930	100,00
61-70	2.104.987	65,04	1.098.291	33,94	32.991	1,02	3.236.269	100,00
71-80	1.655.633	59,34	1.101.056	39,46	33.607	1,20	2.790.296	100,00
81-90	601.400	55,95	450.331	41,89	23.245	2,16	1.074.976	100,00
91-100	79.412	53,43	63.171	42,50	6.048	4,07	148.631	100,00
101 e oltre	654	48,48	595	44,11	100	7,41	1.349	100,00
Totale	**11.962.380**	**67,15**	**5.588.414**	**31,37**	**262.482**	**1,47**	**17.813.276**	**100,00**

(a) Sono esclusi i valori mancanti della variabile relativa alla modalità di compilazione.

Le modalità di compilazione del modello sono influenzate dall'età dell'intestatario (Tavola 1.11), nel senso che il rilevatore ha aiutato il rispondente in molti più casi quando questi era un anziano o una persona molto giovane.

La tavola 1.11 prende in esame la sezione del Modello Istat CP.1 nella quale bisognava indicare in che modo fosse stato compilato il Foglio di famiglia: senza l'aiuto del rilevatore, con l'aiuto del rilevatore, oppure compilato d'ufficio. Come si vede, circa il 31,4 per cento degli intestatari ha usufruito di assistenza nella compilazione, ma tale percentuale è più alta per i molto giovani (36,3 per cento dai 15 ai 20 anni) e per i molto anziani (oltre 40 per cento per gli ultra ottantenni).

1.6 - Il carico del rilevatore

Il carico di lavoro dei rilevatori, in termini di numero di Modelli Istat CP.1 assegnanti, è stato calcolato a livello provinciale.

L'operazione di quantificazione ha presentato alcuni problemi determinati dall'inadeguato livello di qualità del "codice di rilevatore". Questo codice (in genere un numero progressivo), assegnato da ciascun Comune a

ciascun rilevatore doveva essere riportato sul modello dove erano presenti due campi, uno primario, di "base", e un altro di "recupero", da utilizzare per eventuale correzione o rettifica in caso di errore sul primo. Entrambi i codici sono stati registrati o acquisiti con la lettura ottica, ma non sono stati sottoposti alla fase di videocorrezione e controllo nella procedura standard. Per le elaborazioni sul carico di lavoro dei rilevatori si è pertanto provveduto ad una revisione sommaria dei codici di rilevatore, tale da consentire analisi e risultati coerenti.

In primo luogo si sono eliminati dai codici i caratteri anomali dovuti alla non precisa acquisizione tramite lettura ottica (asterischi, valori alfabetici, spazi bianchi e altri segni non dovuti), sia per il codice indicato nel campo primario, sia in quello di recupero. Nel caso il campo di recupero fosse valorizzato, esso è stato confrontato con quello primario e, se diverso da quello primario, considerato come valido, a meno di alcune eccezioni.

Come caso a parte è stata considerata l'elaborazione dei codici dei rilevatori del Comune di Napoli, che ha utilizzato codici di rilevatore non numerici ma misti, in cui cioè sono state utilizzate anche le lettere dell'alfabeto, che pertanto non è stato possibile eliminare come anomale.

Le statistiche sul numero di questionari assegnati e lavorati da ciascun rilevatore sono state calcolate a livello provinciale (Tavola 1.12), escludendo dalla distribuzione le "code" del carico di lavoro, considerando cioè che un carico inferiore a dieci modelli o superiore a mille fosse anomalo.

Tavola 1.12 - Rilevatori e modelli assegnati a ciascun rilevatore: numero medio, coefficiente di variazione, mediana e scarto interquartile per provincia e numero d'ordine nella graduatoria provinciale *(valori assoluti)*

PROVINCIA	Rilevatori	Modelli per rilevatore				
		Numero medio	Coefficiente di variazione	Numero mediano	Scarto interquartile	Numero d'ordine in graduatoria (rispetto alla media)
Torino	3.635	288,7	44,7	288	143	45
Vercelli	358	264,6	38,4	269	120	86
Novara	582	272,1	35,6	277	105	72
Cuneo	1.165	277,1	40,1	269	121	63
Asti	408	268,1	40,6	273	149	84
Alessandria	869	261,0	42,2	258	130	91
Aosta	327	299,0	41,5	299	144	29
Imperia	421	343,6	44,3	332	182	1
Savona	672	313,7	42,8	306	143	12
Genova	1.433	341,3	32,0	353	127	2
La Spezia	388	302,0	33,3	306	99	23
Varese	1.185	295,1	36,2	304	122	33
Como	884	279,3	35,1	278	111	58
Sondrio	344	335,5	43,3	315	177	3
Milano	5.310	292,9	44,6	303	129	38
Bergamo	1.529	296,5	37,6	299	125	31
Brescia	1.722	295,0	39,8	295	122	34
Pavia	876	270,5	43,1	267	124	78
Cremona	525	271,3	35,0	269	115	75
Mantova	509	305,2	29,0	309	90	20
Bolzano	887	223,3	43,6	217	104	103
Trento	920	309,8	36,9	315	128	16
Verona	1.281	277,3	39,3	276	126	62
Vicenza	1.177	292,0	36,7	295	116	39
Belluno	455	284,0	37,1	278	118	51
Treviso	1.171	271,0	36,5	275	118	76
Venezia	1.227	296,4	40,2	300	112	32
Padova	1.151	283,3	45,2	274	107	53
Rovigo	390	268,4	33,2	269	95	83
Udine	1.002	268,5	39,7	268	114	82
Gorizia	242	254,4	46,0	263	138	95
Trieste	416	297,9	37,0	299	88	30
Piacenza	527	260,9	35,5	250	89	92
Parma	693	285,4	38,5	288	110	47
Reggio nell'Emilia	704	285,7	32,1	287	96	46
Modena	1.041	284,6	34,0	285	88	48
Bologna	1.656	271,5	38,7	273	118	74
Ferrara	568	316,2	37,1	326	91	8
Ravenna	491	324,8	42,6	323	141	5
Forli-Cesena	565	284,1	36,7	275	93	50
Pesaro e Urbino	583	280,1	36,9	276	106	56
Ancona	752	260,8	41,4	258	132	93
Macerata	490	273,3	47,4	269	153	69
Ascoli Piceno	569	276,1	37,8	274	133	64
Massa-Carrara	389	262,4	42,5	239	103	90
Lucca	654	278,5	46,0	262	134	61
Pistoia	433	279,0	41,3	281	99	60
Firenze	1.270	313,1	41,7	315	141	13
Livorno	500	331,3	36,6	334	109	4
Pisa	606	270,7	42,0	270	124	77
Arezzo	501	284,1	35,3	289	108	49
Siena	418	279,1	33,8	284	117	59
Grosseto	430	307,8	46,6	289	141	17
Perugia	873	299,2	34,3	292	97	27
Terni	358	272,1	40,1	289	112	71
Viterbo	519	282,4	45,4	276	144	55
Rieti	310	294,9	46,1	303	183	35
Roma	5.257	314,6	53,8	307	161	11
Latina	749	301,1	52,4	296	174	25
Frosinone	777	274,7	38,7	268	116	66
Caserta	1.224	271,6	51,8	267	173	73
Benevento	477	248,1	47,7	243	158	100
Napoli	3.207	224,7	71,8	227	262	102
Avellino	716	268,8	50,0	262	161	81
Salerno	1.713	257,1	44,1	250	132	94
L'Aquila	637	299,9	44,6	292	154	26
Teramo	521	266,1	57,0	244	162	85
Pescara	448	275,0	47,0	270	175	65
Chieti	721	247,3	46,4	237	138	101
Campobasso	408	284,0	43,9	273	159	52

Tavola 1.12 segue **- Rilevatori e modelli assegnati a ciascun rilevatore: numero medio, coefficiente di variazione, mediana e scarto interquartile per provincia e numero d'ordine nella graduatoria provinciale** *(valori assoluti)*

PROVINCIA	Rilevatori	Modelli per rilevatore				
		Numero medio	Coefficiente di variazione	Numero mediano	Scarto interquartile	Numero d'ordine in graduatoria (rispetto alla media)
Foggia	1.122	269,5	52,2	264	165	80
Bari	1.906	315,9	42,8	305	136	9
Taranto	834	322,3	42,5	312	143	6
Brindisi	642	303,4	54,2	291	171	22
Lecce	1.387	272,5	47,7	265	145	70
Potenza	684	263,1	44,8	264	129	88
Matera	341	254,1	54,3	255	162	97
Cosenza	1.538	264,3	53,2	252	154	87
Catanzaro	694	274,2	52,5	260	174	67
Reggio di Calabria	992	270,0	45,7	266	160	79
Trapani	765	291,9	51,9	272	170	40
Palermo	1.684	317,7	49,6	306	168	7
Messina	1.335	251,8	58,0	231	169	99
Agrigento	817	304,8	45,8	290	173	21
Caltanissetta	435	309,9	47,0	296	173	15
Enna	294	289,7	44,9	282	132	42
Catania	1.782	262,6	46,1	244	142	89
Ragusa	510	315,6	50,5	280	166	10
Siracusa	604	310,6	45,4	292	155	14
Sassari	797	299,0	48,6	294	160	28
Nuoro	438	307,7	40,4	298	136	18
Cagliari	1.072	301,6	38,2	297	122	24
Pordenone	514	254,2	37,8	259	98	96
Isernia	188	252,6	51,6	233	155	98
Oristano	249	282,8	43,2	280	135	54
Biella	342	273,8	31,7	283	91	68
Lecco	512	294,7	38,8	298	123	36
Lodi	282	289,5	38,4	295	121	43
Rimini	446	290,8	40,5	292	127	41
Prato	314	279,4	36,7	273	65	57
Crotone	305	306,2	59,6	277	201	19
Vibo Valentia	305	289,4	48,5	293	187	44
Verbano-Cusio-Ossola	340	294,1	38,7	284	112	37
Italia	**90.686**	**285,4**	**45,6**	**283**	**140**	

Il numero medio di Modelli Istat CP.1 gestiti dal singolo rilevatore è pari a 285,4. Il carico del rilevatore è stato massimo nella provincia di Imperia (343,6 modelli a testa), Genova (341,3), Sondrio (335,5), Livorno (331,3), Ravenna (324,8). Al contrario la provincia dove il numero medio di Fogli di famiglia restituiti da ciascun rilevatore è stato minimo è stata Bolzano (223,3 modelli), seguita da Napoli (224,7), Chieti (247,3), Benevento (248,1), Messina (251,8).

1.7 - L'articolazione delle fasi di correzione per gruppi di variabili omogenee (le diverse fasi di Definizione valori)

Nel contesto del processo di validazione dei dati, la fase dei controlli e delle correzioni qualitative è stata denominata "Definizione valori" (DV). Per ciascuna unità di analisi sono state definite le procedure per l'*editing* (verifica della coerenza) dei dati, partendo da una serie di regole di compatibilità, che dovevano essere soddisfatte nell'insieme di dati finali.

In particolare le difficoltà connesse alla complessità della correzione delle variabili nei questionari sono state affrontate definendo una scomposizione del problema della validazione in aree disgiunte, risolvendo il problema della correzione per blocchi di variabili omogenee nelle sezioni del questionario.

Il primo livello di scomposizione è stato per tipo di questionario: Questionario di edificio, Foglio di famiglia e Foglio di convivenza.

Nel caso del modello relativo alle persone che vivono in famiglia, valutando che il questionario stesso è molto complesso e articolato (oltre 60 variabili individuali e 7 familiari) si è considerato opportuno suddividerlo ulteriormente in blocchi, a ciascuno dei quali è stata assegnata una specifica procedura (ad esempio, la procedura per correggere la cittadinanza, quella per il lavoro e quella per lo studio, quella per il pendolarismo).

La suddivisione del processo di validazione in procedure separate, che trattano sezioni del questionario, o gruppi di variabili omogenee, ha consentito l'ulteriore vantaggio di poter riutilizzare la stessa procedura sui sottoinsiemi di variabili analoghe, comuni a diverse unità di analisi (ad esempio, gli stranieri sia in famiglia che in convivenza, i pendolari sia residenti che non residenti, che sono stati trattati e corretti con la stessa metodologia).

Un altro punto cruciale nella definizione dei passi di "Definizione valori" ha riguardato la scelta delle priorità/precedenze di correzione delle variabili, tenendo conto della correlazione fra le variabili dei diversi gruppi, in modo da massimizzare l'indipendenza delle correzioni nelle procedure adottate, con l'obiettivo di correggere definitivamente in un certo passo ciascuna variabile e non modificarla più in un passo successivo.

Tavola 1.20 - Procedure di validazione utilizzate per la correzione delle diverse unità e per gruppi di variabili

Variabili relative ai residenti in famiglia	Variabili relative ai residenti in convivenza	Variabili relative ai non residenti in famiglia	Variabili relative ai non residenti in convivenza	Procedura di validazione di Definizione valori (DV)
Anagrafiche				DV1_FAM
	Anagrafiche			DV1_CONV
		Anagrafiche, cittadinanza, dimora, lavoro		DV_NON_RES_FAM
			Anagrafiche, cittadinanza, dimora	DV_NON_RES_CONV
Cittadinanza	Cittadinanza			DV2
Presenza e dimora	Presenza e dimora			DV3_DIMORA
Studio e lavoro	Studio e lavoro			DV3_STU_LAV
Pendolarismo		Pendolarismo		DV3_PENDOL

Nella tavola 1.20 sono schematizzati i gruppi di variabili corretti nelle varie procedure. Il primo passo, definito dalla procedura "DV1" ha corretto le variabili anagrafiche. La procedura "DV2" ha considerato le variabili di cittadinanza, in maniera analoga per residenti in famiglia e in convivenza. Con "DV3-studio-lavoro" sono state corrette le variabili di titolo di studio e condizione lavorativa dei residenti in famiglia e in convivenza; quindi "DV3-pendolarismo" ha corretto l'insieme di variabili relative al luogo di studio o lavoro, in maniera uguale per i residenti e per i non residenti in famiglia. Anche il gruppo di variabili riguardante la presenza e la dimora è stato corretto con procedure analoghe per residenti e non ("DV3-dimora").

Durante il processo di correzione sono stati monitorati gli effetti delle procedure sui dati, per valutare le regole di compatibilità più attivate, per tenere sotto controllo la coerenza dei dati via via prodotti, per misurare l'ammontare complessivo di modifiche introdotte e la direzione di tali modifiche sui dati. Il principale supporto ai controlli è stato fornito dal *data warehouse* di controllo, dove i dati grezzi e i dati puliti sono stati caricati ed organizzati per le analisi e le verifiche sulla correzione in corso. Gli esiti dei controlli di ciascun passo del processo sono a loro volta divenuti patrimonio del *data warehouse*, disponibili per le analisi sulla qualità dei passi successivi.

Alcuni degli indicatori generati in fase di validazione sono confluiti nell'"Archivio di qualità", costruito a scopo di documentazione.[8]

1.8 - L'*editing* e le regole maggiormente attivate

Si illustrano nel paragrafo i risultati delle analisi riguardanti il numero di regole di compatibilità che sono state attivate (*edit falliti*) e che hanno conseguentemente comportato interventi e modifiche sui dati al fine di ripristinare la coerenza delle risposte. Le regole di compatibilità hanno riguardato i diversi sottoinsiemi di popolazione e di unità: per consentire una valutazione adeguata degli interventi di correzione sugli aggregati coinvolti, nella tavola 1.21 si forniscono i totali dei diversi sottogruppi.

[6] Per ulteriori approfondimenti sull'Archivio di qualità si veda il paragrafo 1.14.

Tavola 1.21 - Frequenze di alcuni aggregati rilevati dal 14° Censimento *(valori assoluti)*

AGGREGATI		Valori assoluti
Famiglie		21.810.676
Residenti in famiglia		56.594.021
Temporaneamente dimoranti in famiglia		1.826.212
Nuclei familiari		16.130.368
Coppie		14.029.369
Coppie con figli		9.273.942
Famiglie unipersonali		5.427.621
Nuclei familiari costituiti da un solo genitore		2.100.999
Famiglie con 5 o più componenti		1.635.232
Convivenze		46.899
Totale persone residenti in convivenza		401.723
di cui	*Responsabile o dirigente della convivenza*	*11.989*
	Addetto all'assistenza sanitaria	*1.492*
	Addetto all'assistenza sociale o psicologica	*1.203*
	Addetto ai servizi amministrativi o ordinari	*1.833*
	Addetto ai servizi di manutenzione e di pulizia	*1.147*
	Religioso	*112.960*
	Detenuto con condanna definitiva	*15.252*
	Assistito in un centro di accoglienza per immigrati	*2.497*
	Ricoverato in un istituto di cura	*24.398*
	Assistito in istituto per anziani	*148.655*
	Assistito in istituto per minori	*4.113*
	Assistito in altro centro di accoglienza	*10.004*
	Altro	*66.180*
Totale persone temporaneamente dimoranti in convivenza		907.984
di cui	*Addetto ai servizi di assistenza (sanitaria, sociale, eccetera)*	*36.511*
	Addetto ad altri servizi	*14.740*
	Religioso	*10.282*
	Detenuto	*34.648*
	Assistito in un centro di accoglienza per immigrati	*4.441*
	Ricoverato, lungodegente in un istituto di cura	*219.539*
	Assistito in istituto per anziani	*64.699*
	Assistito in istituto per minori	*6.961*
	Collegiale	*37.222*
	Passeggero di nave e componente l'equipaggio	*21.723*
	Ospite di albergo	*285.682*
	Altro	*171.536*
Totale popolazione residente (in famiglia o in convivenza)		56.995.744
Totale popolazione temporaneamente dimorante (in famiglia o in convivenza)		2.734.196
Popolazione straniera residente (inclusi apolidi)		1.334.889
di cui	*Residenti in famiglia*	*1.306.999*
	Residenti in convivenza	*27.890*
Forze di lavoro		23.742.262
di cui	*Occupati*	*20.993.732*
	In cerca di occupazione	*2.748.530*
Non forze di lavoro		25.150.297
di cui	*Studenti*	*3.589.433*
	Casalinghe/i	*7.478.550*
	Ritirati dal lavoro	*10.089.487*
	In altra condizione	*3.992.827*
Popolazione residente in famiglia che si sposta giornalmente per motivi di lavoro		17.066.957
Popolazione residente in famiglia che si sposta giornalmente per motivi di studio		9.697.404
Abitazioni occupate		21.653.288
Abitazioni non occupate		5.638.705
Edifici		12.774.131
Complessi di edifici		38.397

Le regole di compatibilità della prima fase di controllo riguardano la **famiglia** censita nel suo insieme e le variabili individuali dei singoli componenti, variabili che incidono nella definizione della tipologia familiare, in particolare *sesso, anno di nascita, stato civile e relazione di parentela.* Se si considera il singolo individuo,

separatamente dagli altri membri della famiglia, le sue caratteristiche individuali hanno relazioni di compatibilità solo con le sue risposte individuali nel questionario e quindi un insieme di vincoli limitato. Considerare invece il contesto familiare determina un insieme più ampio di connessioni fra le caratteristiche anagrafiche del soggetto e quelle degli altri membri della famiglia, soprattutto attraverso la relazione di parentela, che implica legami di coppia e/o di discendenza. Si definiscono quindi vincoli di coerenza per le modalità della variabile *sesso,* che deve essere diverso per marito e moglie, sulla differenza di età fra genitori e figli, sull'anno di matrimonio che deve coincidere per la coppie coniugate, eccetera.

Le regole di compatibilità implementate nella procedura sulle famiglie sono state 229, di cui 77 (34 per cento) a livello individuale e 152 (66 per cento) che considerano i legami familiari. Le regole individuali sono quelle maggiormente attivate (14 milioni e mezzo di casi), mentre le regole familiari sono state attivate in 1 milione 470 mila casi. Alcune delle regole di controllo individuavano veri e propri errori, altre invece erano da considerarsi come *warning* o segnalazione di anomalia, in cui non necessariamente si procedeva a forzature di correzione. Queste intervenivano solo nel caso di implicazione congiunta di più *warning* o di compresenza di errori.

Le regole maggiormente attivate a livello individuale dalla procedura di correzione (Tavola 1.21) hanno riguardato lo *stato civile prima del matrimonio*, che doveva essere indicato dai coniugati o ex-coniugati (quasi 3 milioni di casi errati) e la mancanza di *mese* o di *anno di matrimonio*, cioè della data di matrimonio (circa 1,3 milioni di casi). Seguono le mancate risposte sulle altre variabili familiari: *relazione di parentela* (985 mila), *sesso* (759 mila), *stato civile* (564 mila), *anno di nascita* (353 mila).

Le regole strettamente familiari (in cui devono essere considerati i legami fra almeno due componenti) intervengono molto meno frequentemente (Tavola 1.22): in circa 198 mila casi nella coppia principale della famiglia (quella costituita da intestatario e il suo coniuge o convivente) le rispettive durate del matrimonio sono diverse ed in circa 183 mila casi il mese di matrimonio è diverso. Sempre per la coppia principale risultano errori sullo stato civile del partner, che risulta non coniugato pur essendo coniugato l'intestatario (137 mila casi), oppure l'intestatario è coniugato ma il partner si dichiara convivente coniugalmente (99 mila), oppure ancora si rilevano modalità di *stato civile* incompatibili per i due componenti della coppia principale. Altre regole di compatibilità frequentemente violate riguardano la differenza di età fra genitori e figli: l'intestatario risulta troppo giovane (cioè con una differenza di età inferiore ai 12 anni) per poter essere genitore del figlio della coppia in circa 98 mila casi, oppure è il coniuge dell'intestatario ad essere troppo giovane (83 mila casi) o ancora si riscontra una differenza di età con la madre superiore a 55 anni (56 mila casi).

Tavola 1.21 - Frequenza di attivazione di alcune regole di compatibilità a livello individuale definite nella procedura di correzione delle variabili familiari *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ (INDIVIDUALE)	Frequenza delle regole errate nei dati grezzi
Non celibe/nubile, ma senza *stato civile prima del matrimonio*	2.920.481
Non celibe/nubile, ma senza *mese di matrimonio*	1.366.531
Non celibe/nubile, ma senza *anno matrimonio*	1.270.100
Relazione di parentela non indicato	985.535
Sesso non indicato	759.376
Stato civile non indicato	564.072
Età non indicato	353.738
Sposato ad un'età minore di 14 anni	337.877
Coniugato che non indica *anno di matrimonio*	336.178
Celibe/nubile, ma con *stato civile prima del* matrimonio	275.976
Fuori *range* sul *mese di matrimonio*	205.031
Celibe/nubile, ma con *mese di matrimonio* valorizzato	200.519
Celibe/nubile, ma con *anno di matrimonio*	196.173
Altre regole	....

Tavola 1.22 - Frequenza di attivazione di alcune regole di compatibilità a livello familiare definite nella procedura di correzione delle variabili familiari *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ (FAMILIARE)	Frequenza delle regole errate nei dati grezzi
Coppia principale di coniugi con diversa durata di matrimonio	198.964
Coppia principale di coniugi con diverso mese di matrimonio	183.734
Coppia principale con intestatario coniugato e coniuge non sposato	137.203
Intestatario sposato ma il/la partner è convivente	99.413
Genitore (intestatario) con figlio che ha meno di 12 anni di differenza	98.059
Coppia principale con stato civile incompatibile	96.097
Genitore (coniuge o convivente dell'intestatario) con figlio che ha meno di 12 anni differenza	82.834
Coppia principale in cui lei ha almeno 19 anni più di lui	66.309
Madre (in coppia con intestatario) e figlio con più di 55 anni di differenza	55.552
Differenza di età nella coppia principale maggiore di 35 anni	43.927
Coppia principale di coniugi entrambi femmine	42.373
Differenza di età tra fratelli/sorelle maggiore di 48 anni	36.549
Differenza di età nella coppia principale maggiore di 35 anni	32.701
Coppia principale di coniugi entrambi maschi	26.562
Padre (intestatario) e figlio con più di 70 anni di differenza	19.577
Intestatario e figlio del solo intestatario con meno di 12 anni di differenza	18.682
Compresenza di due coniugi dell'intestatario	17.964
Altre regole	....

Per quanto riguarda le **convivenze** (Tavola 1.23) la procedura di correzione delle variabili strutturali si è sviluppata a partire da una trentina di regole di compatibilità. Le più frequentemente attivate hanno riguardato la mancata risposta sullo *stato civile* (circa 23 mila casi) e sul *motivo di permanenza nella convivenza* (15 mila). A seguire nella graduatoria la variabile *sesso* non era invece indicata in 4 mila casi, mentre la *data di nascita* non era indicata in 3.600 casi. Vere e proprie incompatibilità fra coppie di variabili sono verificate sempre al di sotto dei mille casi e riguardano sostanzialmente alcune incoerenze fra il *motivo di permanenza nella convivenza* e l'*età* o la *tipologia di convivenza*.

Tavola 1.23 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili anagrafiche per i residenti in convivenza *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Stato civile non indicato	23.098
Motivo convivenza non indicato	15.512
Sesso non indicato	4.101
Data di nascita non indicato	3.601
Motivo convivenza ="anziano" e 0<=*età*<=39	923
Stato civile non celibe/nubile e 0<=*età*<=17	723
Motivo convivenza ="minore" e 19<=*età*<=113	585
Motivo convivenza ="militare", "detenuto","minore", "anziano" e *tipo convivenza* = "altro tipo di accoglienza"	269
Motivo convivenza ="militare", "detenuto", "minore", "immigrato", "ospite di albergo" e *tipo convivenza* = "ospizio"	178
Motivo convivenza = "militare", "detenuto" e *tipo convivenza* = "convivenza ecclesiastica"	125
Sesso = "femmina" e *motivo convivenza* = "religioso" e 0<=*età*<=16	108
Motivo convivenza ="militare" e 71<=*età*<=113	100
Motivo convivenza ="militare" e 0<=*età*<=16	87
Motivo convivenza ="militare", "detenuto", "immigrato", " minore" e *tipo convivenza* = "centro per disabili"	70
Sesso = "maschio" e *motivo convivenza* = "religioso" e 0<=*età*<=18	68
Motivo convivenza = "religioso", "detenuto", "immigrato", " ricoverato lungodegente", "anziano", "minore" e *tipo di convivenza* = "caserma militare" e 66<=*età*<=113	60
Altre regole	....

La procedura di correzione delle variabili di **cittadinanza** per i **residenti in famiglia** è partita dalla definizione di 43 tipologie di errore o violazione delle regole di compatibilità. L'errore più frequentemente riscontrato (Tavola 1.24) è la mancata risposta al quesito esplicito sulla *cittadinanza* (3,6 milioni di casi), accompagnato da una risposta sulla *origine della cittadinanza italiana* (3 milioni circa), che viola una ulteriore regola. In pratica in molti casi i cittadini italiani, pur non avendo biffato la casella *cittadinanza italiana*, hanno

indicato subito dopo sono *italiano dalla nascita*. Viceversa sono circa 864 mila i casi di cittadini italiani che hanno dimenticato di indicare *l'origine della propria cittadinanza* (dalla nascita o acquisita). Altre regole frequentemente disattese riguardano le mancate risposte sul *luogo di nascita* o l'indicazione, fornita quando non richiesta, della provincia di nascita (887 mila) o anche del comune di nascita (631 mila). Al contrario in circa 384 mila hanno dimenticato di specificare la provincia e il comune di nascita quando necessario.

L'errore dei codici di cittadinanza non validi, (sia attuale, con 434 mila casi, che precedente all'acquisizione della cittadinanza italiana, 442 mila casi) riguarda quasi esclusivamente i dati delle provincia di Bolzano che, avendo registrato in proprio (fuori dal contratto per la lettura ottica) i dati raccolti, ha prodotto un tracciato record in cui, in caso di cittadinanza italiana, invece di non assegnare alcun codice di cittadinanza, come previsto dal tracciato standard, assegnava comunque un codice numerico (fittizio) di stato estero. In questo modo si sono attivate sistematicamente alcune regole di compatibilità per la presenza di tali codici.

Altri errori frequenti hanno riguardato il mancato rispetto dei "filtri" alle domande del questionario, per cui in molti hanno risposto al *motivo* e all'a*nno di trasferimento in Italia*, pur essendo italiani (dalla nascita o acquisiti) o non hanno risposto nei casi previsti, pur essendo stranieri.

Tavola 1.24 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili sulla cittadinanza per i residenti in famiglia *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Cittadinanza non indicato	3.653.965
Cittadinanza non indicato ma *Italiano dalla nascita* = sì	3.008.485
Luogo di nascita non indicato	1.653.222
Luogo di nascita = stesso comune o estero e *provincia di nascita* specificata	887.084
Cittadinanza = italiana e *italiano dalla nascita o acquisito* non indicato	864.602
Luogo di nascita = stesso comune o estero e *provincia di nascita* specificata	631.669
Codice di stato estero di cittadinanza precedente non valido	442.554
Codice di stato estero di cittadinanza attuale non valido	434.038
Cittadinanza = italiana e *italiano dalla nascita* = sì e *codice di cittadinanza precedente* valorizzato	426.729
Cittadinanza = italiana e *italiano dalla nascita* = sì e *codice di cittadinanza straniera attuale* valorizzato	423.789
Luogo di nascita = in altro comune e *comune di nascita* non indicato	390.443
Luogo di nascita = in altro comune *e provincia di nascita* non indicato	384.724
Cittadinanza = italiana e *italiano acquisito* = sì e *motivo di trasferimento in Italia* valorizzato	102.047
Cittadinanza = italiana e *italiano acquisito* = sì e *anno di trasferimento in Italia* valorizzato	97.633
Cittadinanza = straniera e *luogo di nascita* = estero e *motivo di trasferimento in Italia* non indicato	84.315
Cittadinanza = straniera e *luogo di nascita* = estero e *anno di trasferimento in Italia* non indicato	69.522
Comune di nascita fuori range	60.404
Cittadinanza = italiana e *italiano dalla nascita* = sì e *motivo di trasferimento in Italia* valorizzato	46.372
Luogo di nascita = estero e *stato estero di nascita* non indicato	36.673
Cittadinanza = italiana e *italiano dalla nascita* = sì e *anno di trasferimento in Italia* valorizzato	33.552
Cittadinanza = straniera e *luogo di nascita* = estero e *stato estero di nascita* non indicato	28.666
Cittadinanza = italiana e *italiano dalla nascita* = no e *stato estero di cittadinanza precedente* non indicato	27.519
Entrambi i genitori sono stranieri e il figlio minorenne è italiano	19.262
Altre regole	....

Infine, anche nell'ambito delle risposte sulla cittadinanza, si sono riscontrati errori di coerenza a livello familiare: 19 mila casi di genitori entrambi stranieri che hanno indicato di avere un figlio minorenne italiano (non conformemente alla legislazione corrente per la quale non è sufficiente essere nati in Italia per acquisire la cittadinanza italiana). Al contrario in circa 6 mila casi non è stata rispettata la regola (di legge) che assegna automaticamente la cittadinanza italiana al figlio di un italiano (anche se l'altro genitore è straniero).

Tavola 1.25 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili sulla cittadinanza per i residenti in convivenza *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Cittadinanza non indicato	28.590
Italiano dalla nascita o acquisito valorizzato e *cittadinanza* = non italiana	21.630
Luogo di nascita non indicato	7.703
Cittadinanza = italiana e *italiano dalla nascita o acquisito* non indicato	5.803
Cittadinanza = straniera e *luogo di nascita* = estero e *motivo di trasferimento in Italia* non indicato	3.310
Cittadinanza = straniera e *luogo di nascita* = estero e *anno di trasferimento in Italia* non indicato.	3.071
Luogo di nascita = altro comune italiano e *provincia di nascita* non indicato	1.616
Luogo di nascita = altro comune italiano e *comune di nascita* non indicato	1.597
Comune di nascita codice fuori range	1.418
Luogo di nascita = stesso comune o estero e *provincia di nascita* valorizzato	1.186
Luogo di nascita = stesso comune o estero e *comune di nascita* valorizzato	1.155
Cittadinanza = italiana e *italiano dalla acquisito* = sì e *motivo di trasferimento in Italia* valorizzato	937
Cittadinanza = italiana e *italiano dalla acquisito* = sì e *anno di trasferimento in Italia* valorizzato	907
Cittadinanza = italiana e *italiano dalla acquisito* = sì e *stato estero di cittadinanza precedente* non indicato	634
Cittadinanza = straniera e *luogo di nascita* = estero e *stato estero di cittadinanza* non indicato	511
Luogo di nascita = estero e *stato estero di nascita* non indicato	465
Altre regole	....

Gli errori verificati con la procedura **cittadinanza** sui 401.723 **residenti in convivenza** sono molto simili a quelli incontrati per le famiglie (Tavola 1.25). In particolare prevalgono: la mancata risposta su *cittadinanza* (28.590 casi), sull'*origine della cittadinanza italiana* (21.630) e sul *luogo di nascita* (7.703).

Per quanto riguarda il terzo passo di correzione denominato DV3. che prendeva in esame le variabili su titolo di studio, condizione professionale e pendolarismo per i residenti in famiglia (Tavola 1.26), le regole di compatibilità più frequentemente attivate hanno riguardato la condizione lavorativa e l'indicazione di avere lavorato per almeno un'ora nella settimana precedente (6 milioni 688 mila casi), eventualmente combinata con il settore di attività economica (5 milioni 859 mila casi). Altri errori molto frequenti (oltre 5 milioni di casi) sono quelli che riguardano la dichiarazione circa l'*iscrizione a corsi di studio* e a *corsi di formazione professionale* a cui dovevano rispondere i maggiori di sei anni. L'indicazione del filtro "Per i laureati" presente per i tre quesiti precedenti, ha probabilmente indotto molti rispondenti (non laureati) a saltare le risposte alle successive domande, relative al tema della formazione, e passare direttamente ai quesiti sul lavoro.

Per quanto riguarda le persone **non residenti in famiglia** gli errori di compilazione più frequentemente rilevati (Tavola 1.27) riguardano la mancata risposta alla *dimora nell'alloggio* alla data del censimento e alla *durata di dimora* in questo alloggio (652 mila e 394 mila rispettivamente). Un secondo tipo di errore riguarda la mancata risposta alle variabili anagrafiche (la *data di nascita*, la *cittadinanza*, lo *stato civile*) e la *condizione professionale*.

Tavola 1.26 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili su titolo di studio, condizione professionale e pendolarismo per i residenti in famiglia *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Incompatibilità tra *condizione professionale* ed *effettuazione ore lavoro* nella settimana precedente	6.688.176
Incompatibilità tra *condizione professionale, ore lavoro* e *settore di attività economica*	5.859.340
Iscrizione ad un corso regolare di studi non indicato ed *età* > 6	5.195.558
Incompatibilità tra *età e corso di formazione professionale*	5.167.440
Incompatibilità tra *condizione professionale, ore lavoro, settore di attività economica* e *attività lavorativa svolta*	4.769.556
Incompatibilità tra *titolo di studio* e *codice* di titolo di studio	3.807.087
Incompatibilità tra *condizione professionale*, *posizione nella professione* ed *effettuazione ore lavoro* nella settimana precedente	3.785.712
Incompatibilità tra *condizione professionale* ed *effettuazione ore lavoro* nella settimana precedente sotto la condizione che dichiara un luogo di lavoro	3.706.517
Incompatibilità tra *età* e *corso di formazione professionale* e specificazione del *tipo di corso*	3.688.178
Età <14 e frequenza di *corso di formazione professionale*	3.628.904
Incompatibilità *effettuazione ore lavoro* nella settimana precedente e *luogo di lavoro* con *condizione non professionale*	2.805.729
Incompatibilità tra *luogo di lavoro* e *alloggio di uscita*	2.547.480
Incompatibilità *durata* del corso di studi superiori e *titolo di studio*	2.463.544
Incompatibilità *durata* del corso di studi superiori e *titolo di studio* e *codice* di titolo di studio	2.130.752
Età < 15 e *condizione professionale* valorizzata	2.094.627
Incompatibilità tra *titolo di studio* ed *anno di nascita*	1.919.910
Condizione non professionale con *ore di lavoro* di persona che dichiara di non avere mai lavorato	1.385.775
Incompatibilità tra *posizione nella professione* e *tipo di rapporto di lavoro*	1.299.942
Incompatibilità tra *specializzazione post laurea* e *codice* di titolo di studio	1.160.019
Effettuazione ore di lavoro e presenza di risposte dovute solo da chi non ha effettuato ore	1.148.737
Incompatibilità tra *specializzazione post laurea* e *titolo di studio*	1.096.069
Incompatibilità *ore di lavoro* e *luogo di lavoro*	1.072.479
Altre regole	**....**

Tavola 1.27 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili relative ai non residenti in famiglia *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Dimora abituale alla data di censimento non indicato	652.637
Durata dimora in questo alloggio non indicato	394.351
Età < 0 o non indicata	360.681
Anno di nascita non indicato	347.578
Mese di nascita non indicato	343.846
Giorno di nascita non indicato	342.238
Cittadinanza non indicato	298.906
Età < 16 e *stato civile* diverso da celibe/nubile	225.024
Età <15 e *condizione professionale* valorizzata	215.367
Luogo di presenza giorno censimento non indicato	182.147
Durata dimora in questo alloggio < 90 gg e *condizione professionale* valorizzata (risposta non dovuta)	152.488
Sesso non indicato	127.964
Durata dimora in questo alloggio >91 gg e *motivo di utilizzo dell'alloggio* non indicato	116.205
Stato civile non indicato	114.220
Altre regole	**....**

Anche per i quasi 908 mila **temporaneamente dimoranti in convivenza** (Tavola 1.28) le mancate risposte sono concentrate sulle variabili relative alla residenza ed in particolare alla *durata della dimora* in questo alloggio (307.284 casi) e al *luogo di presenza* il giorno del censimento (245.908). In oltre 153 mila casi non è

indicato lo *stato civile*, e in 82 mila casi la *cittadinanza*. Al contrario in 119 mila casi viene fornita una risposta non dovuta alla *condizione professionale*, da coloro che dimorano nell'alloggio da meno di 90 giorni.

Gli **edifici** (Tavola 1.29) rilevati sono stati poco meno di 13 milioni (12.812.528, inclusi i complessi di edifici). In oltre 2 milioni di casi si sono evidenziati valori anomali (rispetto alla media) nella relazione fra numero di scale e numero di piani e numero di interni dell'edificio. Altri errori riscontrati riguardano soprattutto le mancate risposte sul numero di scale (1,2 milioni di casi) e sul numero di interni (circa 1 milione). Incongruenze fra numero di scale, livelli, numero di piani fuori terra e numero di interni si sono verificate in circa 500 mila casi per gli edifici costituiti da un solo interno e in circa altrettanti nel caso di edifici includenti più di un interno.

Per le **abitazioni** nella fase di controllo e correzione si sono considerate separatamente quelle **occupate** (21.653.288), il cui modello è stato compilato da una persona che vive nell'abitazione, da quelle **non occupate** (5.638.705), per le quali le informazioni sono state fornite dal rilevatore, che poteva trascurare le risposte ad alcuni quesiti. Gli errori più frequenti per le abitazioni occupate (Tavola 1.30) hanno riguardato le mancate risposte sulla presenza di angolo cottura (2 milioni 944 mila casi), di cucinino (2 milioni 587 mila) e di cucina (1 milione 309 mila); la mancata indicazione del numero di stanze ad uso professionale (1 milione 412 mila) o di disponibilità di posto auto in garage o in cortile.

In molti casi (4 milioni 481 mila) non sono state rispettate le indicazioni del questionario relativamente alle domande sul tipo di combustibile o di energia usata per riscaldare l'acqua. Nel modello erano previste due domande separate sul tipo di energia o combustibile, una per l'impianto per la produzione di acqua calda, una per l'impianto di riscaldamento dell'abitazione. Nel caso in cui l'impianto per produrre acqua calda coincidesse con quello destinato al riscaldamento dell'abitazione era previsto che la risposta sul tipo di combustibile fosse fornita una sola volta (relativamente al tipo di combustibile usato per il riscaldamento domestico). È accaduto invece che in quest'ultimo caso, le informazioni sul tipo di combustibile siano state fornite due volte, rispondendo ai due quesiti separatamente.

Tavola 1.28 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili per i non residenti in convivenza *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Durata dimora in questo alloggio non indicato	307.284
Luogo di presenza giorno censimento non indicato	245.908
Stato civile non indicato	153.562
Durata dimora in questo alloggio < 90 gg e *condizione professionale* valorizzata (risposta non dovuta)	119.061
Cittadinanza non indicato	82.455
Età <0 o *età* >113	71.552
Mese di nascita non indicato	69.332
Giorno di nascita non indicato	69.325
Anno di nascita non indicato	69.134
Sesso non indicato	63.854
Cittadinanza = straniero, *dimora abituale* = estero, *mese da quando è in Italia* non indicato	41.419
Cittadinanza = straniero, *dimora abituale* = estero, *anno da quando è in Italia* non indicato	41.268
Dimora abituale = Italia, *motivo presenza in Italia* = non indicato	33.482
Durata dimora in questo alloggio < 90 gg e *settore attività economica* valorizzato (risposta non dovuta)	32.198
Cittadinanza = straniero, *dimora abituale* = estero, *motivo presenza in Italia* non indicato	31.638
Condizione professionale = occupato e *settore di attività economica* non indicato	29.681
Cittadinanza = italiana e *motivo della presenza in Italia* valorizzato	21.591
Dimora abituale = Italia e *mese di inizio di presenza in Italia* valorizzato	15.134
Dimora abituale = Italia e *anno di inizio di presenza in Italia* valorizzato	15.062
Altre regole	....

Tavola 1.29 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili per edifici (a) *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Rapporto fra *numero di scale* ed il *numero di piani* e*l interni* è anomalo	2.102.988
Numero di scale fuori *range* o non indicato	1.208.629
Numero di interni fuori *range*	1.046.525
Proporzione anomala fra *numero di piani* (=1), *numero di interni*, *livelli*	553.303
Proporzione anomala fra *numero di piani* (>1), *numero di interni*, *livelli*	532.261
Numero di interni non indicato	530.256
Totale interni = 1 e proporzione con *numero di piani* eccessivo	363.366
Presenza di scale in assenza di piani superiori	332.846
Edificio non utilizzato e *numero di piani* valorizzato	310.222
Edificio utilizzato per abitazione e *presenza di piani* non indicato	300.747
Edificio non utilizzato e *numero di interni* valorizzato	292.770
Edificio non utilizzato e contiguità valorizzato	286.792
Altre regole	....

(a) Sono inclusi i complessi di edifici.

Tavola 1.30 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili per abitazioni occupate *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Impianto per acqua calda in comune con impianto riscaldamento, ma è specificato il tipo di energia che lo alimenta (risposta non dovuta)	4.481.181
Presenza angolo cottura non indicato	2.944.250
Presenza cucinino non indicato	2.587.305
Numero stanze e numero stanze uso ufficio non indicati	1.412.061
Numero cucine non indicato	1.309.876
Disponibilità posto auto in cortile non indicato	936.151
Disponibilità posto auto garage non indicato	896.495
Superficie in metri quadri non indicato	843.332
Disponibilità box privato non indicato	724.357
Interventi strutturali non indicato	632.048
Proprietà non indicato	533.637
Ristrutturazione interni non indicato	485.427
Numero di piani non indicato	424.015
Altre regole	....

Per le **abitazioni non occupate** (Tavola 1.31) si osservano numerosi interventi di correzione per eliminare le risposte non dovute ma ugualmente fornite dal rilevatore: disponibilità di box o garage o cortile (oltre 1,8 milioni di casi), interventi di ristrutturazione (oltre 1,7 milioni di casi), presenza di telefono fisso (1,5 milioni), titolo di godimento dell'abitazione (1,3 milioni). In 1 milione 280 mila casi il rilevatore invece non ha saputo indicare il tipo di proprietà dell'abitazione o non ha fornito risposte ai quesiti relativi alla presenza di angolo cottura (927 mila), cucinino (925 mila), numero di cucine (742 mila).

Tavola 1.31 - Frequenza di attivazione di alcune regole di compatibilità definite nella procedura di correzione delle variabili per abitazioni non occupate *(valori assoluti)*

SPIEGAZIONI DELLA REGOLA DI COMPATIBILITÀ	Frequenza delle regole errate nei dati grezzi
Disponibilità box privato valorizzato (risposta non dovuta)	1.848.025
Disponibilità posto auto cortile valorizzato (risposta non dovuta)	1.826.464
Disponibilità posto auto garage valorizzato (risposta non dovuta)	1.816.901
Ristrutturazione impianti valorizzato (risposta non dovuta)	1.781.906
Interventi strutturali valorizzato (risposta non dovuta)	1.774.980
Ristrutturazione interni valorizzato (risposta non dovuta)	1.742.238
Disponibilità telefono fisso valorizzato (risposta non dovuta)	1.578.346
Titolo godimento abitazione valorizzato (risposta non dovuta)	1.325.688
Proprietà non indicato	1.280.393
Presenza angolo cottura non indicato	927.591
Presenza cucinino non indicato	925.488
Numero cucine non indicato	742.819
Numero stanze e numero stanze uso ufficio mancanti	725.620
Superficie in metri quadri non indicato	723.806
Altre regole	....

1.9 - Le modifiche introdotte dai processi di correzione e di imputazione sulle principali variabili

Un criterio per valutare la qualità dei dati è quello di considerare le variazioni che sono state introdotte dalle procedure di correzione. L'ipotesi è che i dati esatti siano quelli validati, e quindi la valutazione degli interventi effettuati sui dati originari possa fornire una misura di quanto questi fossero distanti da quelli finali "buoni". Dal confronto fra i dati in ingresso nelle procedure di correzione (grezzi) e i dati in uscita (puliti) si possono calcolare i tassi di variazione totali e specifici. È così possibile quantificare il tasso di mancata risposta per ogni variabile, ma anche la percentuale di imputazioni o correzioni dovute alle incompatibilità fra variabili, ai fuori *range*, alle risposte non dovute.

Vengono presentati sinteticamente i risultati, relativi alle singole variabili, dei processi di imputazione e correzione (Tavola 1.32), per evidenziare e quantificare quale sia stato l'impatto delle procedure utilizzate.

La tavola 1.32 evidenzia quali sono le variabili relative ai **residenti in famiglia** che sono state maggiormente modificate dalle procedure di correzione, riportando le variabili in ordine crescente per tassi di riposta esatta. Si va dal 79,28 per cento di valori esatti della variabile precodificata sul tipo di attività lavorativa svolta (Tipatt), al 99,93 per cento di valori esatti per il codice di stato estero, indicato da coloro che hanno dimorato all'estero l'anno precedente. La prima delle variabili elencate è stata quindi soggetta a numerose modifiche (22,72 per cento sul totale), in quanto entra in numerose regole di compatibilità, ma anche perché nella fase di correzione è risultata subordinata ad altre variabili: in presenza di una eventuale incoerenza con altre variabili essa è stata pertanto modificata preferendo mantenere invariate le altre risposte fornite ritenute più importanti, come ad esempio la condizione professionale e il titolo di studio, secondo il criterio del minimo cambiamento.

Tavola 1.32 - Modifiche introdotte dal processo di correzione sulle variabili relative ai residenti in famiglia *(valori percentuali)*

VARIABILI	Sul totale degli individui				Sul totale di coloro che dovevano rispondere			Sul totale di coloro che non dovevano rispondere
	Esatte	Imputate	Di cui mancanti	Di cui incompatibili	Esatte	Imputate	Di cui mancanti	Eliminate
Tipo di attività lavorativa svolta	79,28	20,72	2,88	17,84	44,44	55,56	7,79	0,32
Effettuazione ore lavoro nella settimana precedente	85,81	14,19	4,90	9,29	89,76	10,24	10,24	17,82
Frequenza di corso di formazione professionale	87,88	12,12	10,34	1,78	87,44	12,56	10,95	4,68
Iscrizione a scuola e università	89,43	10,57	10,14	0,43	89,13	10,87	10,73	5,44
Studio all'estero	89,76	10,24	8,61	1,63	89,75	10,25	10,25	10,14
Lavoro nel corso della vita	90,18	9,82	3,48	6,34	87,80	12,20	7,34	7,67
Ricerca di lavoro	91,93	8,07	3,86	4,21	91,86	8,14	8,14	8,01
Disponibilità al lavoro	92,05	7,95	3,89	4,06	91,79	8,21	8,20	7,72
Condizione professionale	92,63	7,37	1,30	6,07	95,71	4,29	1,51	25,86
Cittadinanza	93,33	6,67	6,46	0,21	—	—	—	—
Studio o lavoro fuori casa	94,17	5,83	3,36	2,47	94,40	5,60	3,46	13,94
Stato civile precedente al matrimonio	94,35	5,65	5,38	0,27	91,10	8,90	8,90	0,68
Gruppo settore di attività	94,62	5,38	4,87	0,51	86,03	13,97	13,18	0,35
Gruppo titolo di studio	94,73	5,27	3,11	2,16	94,56	5,44	3,29	2,33
Effettuazione ore lavoro nella settimana precedente	95,21	4,79	2,16	2,63	94,05	5,95	5,85	4,11
Anno di ritiro dal lavoro	95,45	4,55	3,78	0,77	85,78	14,22	13,94	0,95
Proseguimento altra dimora	95,76	4,24	0,19	4,05	97,28	2,72	2,71	4,36
Alloggio di rientro da luogo di lavoro/studio	95,81	4,19	2,75	1,44	94,01	5,99	5,77	2,54
Alloggio di uscita verso il luogo di lavoro/studio	95,88	4,12	2,69	1,43	93,97	6,03	5,22	2,10
Luogo di nascita (Italia o estero)	95,98	4,02	2,92	1,10	—	—	—	—
Luogo di studio o lavoro	96,03	3,97	1,55	2,42	95,53	4,47	3,28	3,52
Relazione di parentela	96,58	3,42	1,74	1,68	—	—	—	—
Durata del rapporto di lavoro	96,64	3,36	1,02	2,34	95,06	4,94	3,76	2,77
Posizione nella professione	96,81	3,19	1,29	1,90	91,81	8,19	3,50	0,27
Mese di matrimonio	96,88	3,12	2,43	0,69	94,93	5,07	4,03	0,15
Anno di matrimonio	96,97	3,03	2,25	0,78	95,06	4,94	3,73	0,12
Mezzo di trasporto utilizzato	97,01	2,99	1,06	1,93	97,35	2,65	2,23	3,30
Origine della cittadinanza italiana	97,05	2,95	2,62	0,33	97,24	2,76	2,68	10,68
Dimora un anno fa	97,23	2,77	2,22	0,55	97,52	2,48	2,25	33,01
Tempo impiegato per recarsi al luogo di studio/lavoro	97,34	2,66	0,74	1,92	97,70	2,30	1,55	2,98
Altra dimora ultimo anno	97,40	2,60	2,19	0,41	—	—	—	—
Numero di giorni altra dimora	97,50	2,50	1,06	1,44	81,15	18,85	9,44	0,42
Tipo di rapporto di lavoro	97,61	2,39	0,64	1,75	80,19	19,81	14,86	1,60
Specializzazione post-laurea	97,64	2,36	0,83	1,53	86,32	13,68	13,64	1,63
Durata dell'attività	97,65	2,35	2,22	0,13	93,98	6,02	6,01	0,20
Numero di ore lavorate	98,30	1,70	1,25	0,45	95,50	4,50	3,56	0,19
Motivo altra dimora	98,36	1,64	0,47	1,17	92,37	7,63	6,61	1,19
Motivo di eventuale mancanza di ore lavorate	98,39	1,61	0,17	1,44	89,88	10,12	9,64	1,46
Stato civile	98,52	1,48	1,00	0,48	—	—	—	—
Sesso	98,54	1,46	1,34	0,12	—	—	—	—
Presenza di dipendenti	98,72	1,28	0,97	0,31	88,86	11,14	11,13	0,33
Anno di nascita	98,83	1,17	0,62	0,55	—	—	—	—
Mese di nascita	98,87	1,13	0,20	0,93	93,15	6,85	3,71	0,81
Giorno di nascita	99,05	0,95	0,62	0,33	—	—	—	—
Mese di nascita	99,08	0,92	0,61	0,31	—	—	—	—
Ubicazione altra dimora	99,11	0,89	0,39	0,50	94,04	5,96	5,45	0,50
Presenza alla data di censimento	99,11	0,89	0,89	0,00	—	—	—	—
Motivo di trasferimento in Italia (per gli stranieri nati all'estero)	99,51	0,49	0,20	0,29	86,90	13,10	9,80	0,22
Anno di trasferimento in Italia (per gli stranieri nati all'estero)	99,60	0,40	0,17	0,23	89,73	10,27	8,22	0,19
Codice stato estero di cittadinanza (per gli stranieri)	99,63	0,37	0,32	0,05	85,40	14,60	13,86	0,03
Stato estero di nascita	99,75	0,25	0,22	0,03	93,63	6,37	5,62	0,01
Codice di stato estero di cittadinanza precedente (per gli italiani acquisiti)	99,77	0,23	0,03	0,20	93,40	6,60	5,20	0,20
Codice stato estero dimora abituale anno precedente	99,93	0,07	0,04	0,03	87,83	12,17	11,22	0,03

È da notare che, se il riferimento dei tassi di imputazione è effettuato rispetto a coloro che avevano l'obbligo di rispondere a quella domanda e non sul totale della popolazione, i tassi di imputazione per la variabile Tipatt risultano ancora più elevati. Gli interventi sulla variabile, che indica il tipo di attività svolta, hanno comunque salvaguardato la sua distribuzione, come evidenziato dagli indicatori sulla dissomiglianza (Cfr. par. 1.13) delle distribuzioni prima e dopo il cambiamento nei dati.

Le variabili maggiormente modificate dal processo di imputazione sono inoltre quelle legate alla condizione professionale (ha lavorato nella settimana precedente) e quelle sulla formazione (iscrizione a corsi di formazione professionale o a scuola o università), con mancate risposte che è stato necessario imputare per oltre il 10 per cento di casi.

Si può osservare, analizzando l'ultima colonna della tavola 1.32, che in alcuni casi sono state fornite risposte anche quando non dovute, ad esempio su: dimora nell'anno precedente, condizione professionale, studio all'estero, ore di lavoro nella settimana precedente. Questo indica o una non chiara interpretazione delle istruzioni nel questionario che definiva le regole di risposta o una propensione a fornire comunque risposte, ritenendo che queste potessero essere utili all'indagine.

Nella tavola 1.33, relativa ai **residenti in convivenza**, si osserva come le variabili su cui è stato necessario intervenire di più sono state quelle relative ai corsi di formazione professionale (21,73 per cento), l'anno di arrivo nella convivenza (19,34 per cento), l'attività lavorativa nel corso della vita (17,87 per cento) ed altre variabili connesse alla condizione professionale, soprattutto a causa di mancata risposta.

Per quanto riguarda le risposte non dovute esse si concentrano sulle variabili di dimora nell'anno precedente.

Tavola 1.33 - Modifiche introdotte dal processo di correzione sulle variabili relative ai residenti in convivenza *(valori percentuali)*

VARIABILI	Sul totale degli individui				Sul totale di coloro che dovevano rispondere			Sul totale di coloro che non dovevano rispondere
	Esatte	Imputate	Di cui mancanti	Di cui incompatibili	Esatte	Imputate	Di cui mancanti	Eliminate
Frequenza di corso formazione professionale	78,27	21,73	0,24	21,49	78,21	21,79	0,24	2,53
Anno di arrivo nella convivenza	80,66	19,34	6,06	13,28	—	—	—	—
Lavoro nel corso della vita	82,13	17,87	12,43	5,44	71,12	28,88	21,73	3,14
Anno ritiro dal lavoro	86,07	13,93	13,10	0,83	58,50	41,50	41,27	1,12
Disponibilità al lavoro	86,82	13,18	11,86	1,32	79,27	20,73	20,72	3,08
Ricerca di lavoro	86,96	13,04	11,68	1,36	79,58	20,42	20,41	3,18
Effettuazione ore di lavoro nella settimana precedente	87,47	12,53	11,35	1,18	80,37	19,63	19,60	2,75
Iscrizione a scuola o università	89,54	10,46	10,40	0,06	89,51	10,49	10,42	0
Titolo di studio	90,50	9,50	7,63	1,87	90,48	9,52	7,65	3,35
Studio all'estero	91,68	8,32	8,31	0,01	90,02	9,98	9,98	0,05
Tipo di dimora un anno fa	91,81	8,19	4,86	3,33	91,82	8,18	4,86	18,88
Mese di arrivo nella convivenza	92,52	7,48	6,23	1,25	—	—	—	—
Cittadinanza	92,76	7,24	7,12	0,12	—	—	—	—
Motivo di permanenza nella convivenza	93,43	6,57	3,86	2,71	—	—	—	—
Stato civile	94,13	5,87	5,75	0,12	—	—	—	—
Proseguimento altra dimora	94,39	5,61	0,39	5,22	97,48	2,52	2,47	6,19
Altra dimora nell'ultimo anno	94,50	5,50	4,45	1,05	—	—	—	—
Numero di giorni altra dimora	94,67	5,33	2,12	3,21	80,28	19,72	8,38	0,46
Ore di lavoro settimana precedente	94,91	5,09	3,99	1,10	82,68	17,32	17,13	1,38
Durata dell'attività	95,41	4,59	4,28	0,31	81,59	18,41	18,37	0,39
Posizione nella professione	95,41	4,59	3,95	0,64	81,53	18,47	16,96	0,38
Dimora un anno fa	95,45	4,55	2,85	1,70	95,47	4,53	2,86	23,69
Numero di ore lavorate	96,17	3,83	3,31	0,52	83,93	16,07	14,91	0,33
Durata del rapporto di lavoro	96,52	3,48	2,52	0,96	87,24	12,76	12,50	1,14
Motivo di altra dimora	96,57	3,43	1,44	1,99	90,13	9,87	9,17	2,23
Origine della cittadinanza italiana	96,78	3,22	3,05	0,17	96,62	3,38	3,28	1,16
Specializzazione post-laurea	96,95	3,05	0,88	2,17	84,03	15,97	15,83	2,29
Luogo di nascita	96,95	3,05	1,92	1,13	—	—	—	—
Tipo di rapporto di lavoro	97,70	2,30	0,90	1,40	63,16	36,84	28,87	1,18
Ubicazione altra dimora	97,74	2,26	1,64	0,62	88,11	11,89	10,41	0,46
Motivo di trasferimento in Italia	98,10	1,90	1,02	0,88	78,22	21,78	14,88	0,45
Presenza di dipendenti	98,11	1,89	1,53	0,36	34,62	65,38	63,64	0,33
Condizione professionale	98,27	1,73	0	1,73	98,26	1,74	0	0
Presenza alla data di censimento	98,43	1,57	1,48	0,09	—	—	—	—
Anno di trasferimento in Italia	98,65	1,35	0,96	0,39	85,81	14,19	14,11	0,41
Anno di nascita	98,77	1,23	0,81	0,42	—	—	—	—
Motivo di eventuale mancanza di ore lavorate	98,84	1,16	0,14	1,02	86,72	13,28	13,12	1,03
Sesso	98,98	1,02	1,02	0	—	—	—	—
Giorno di nascita	99,12	0,88	0,81	0,07	—	—	—	—
Mese di nascita	99,13	0,87	0,81	0,06	—	—	—	—
Codice di stato estero di cittadinanza (stranieri)	99,59	0,41	0,38	0,03	94,28	5,72	5,45	0,01
Stato estero di nascita	99,65	0,35	0,31	0,04	96,26	3,74	3,34	0,01
Codice di stato estero di cittadinanza (italiani acquisiti)	99,87	0,13	0,07	0,06	89,30	10,70	9,03	0,04
Frequenza asilo	99,98	0,02	0,01	0,01	93,83	6,17	4,59	0

Tavola 1.34 - Modifiche introdotte dal processo di correzione sulle variabili relative alle abitazioni *(valori percentuali)*

VARIABILE	Abitazioni occupate/non occupate	Esatte	Imputate	Di cui mancanti	Di cui incompatibili
Titolo di godimento dell'abitazione	Abitazioni occupate	98,17	1,83	1,83	0,00
	Abitazioni non occupate	–	–	–	0,00
Proprietà dell'abitazione	Abitazioni occupate	97,56	2,44	2,44	0,00
	Abitazioni non occupate	77,96	22,04	22,04	0,00
Numero di stanze	Abitazioni occupate	99,01	0,99	0,37	0,62
	Abitazioni non occupate	89,81	10,19	9,41	0,78
Numero di stanze uso ufficio (fino a 3)	Abitazioni occupate	93,25	6,75	6,58	0,17
	Abitazioni non occupate	87,18	12,82	12,64	0,18
Numero di stanze uso ufficio (oltre 3)	Abitazioni occupate	99,99	0,01	–	0,01
	Abitazioni non occupate	–	–	–	–
Numero di cucine	Abitazioni occupate	93,62	6,38	5,98	0,40
	Abitazioni non occupate	86,38	13,62	13,13	0,49
Presenza di cucinino	Abitazioni occupate	88,10	11,90	11,83	0,07
	Abitazioni non occupate	83,71	16,29	16,19	0,10
Presenza di angolo cottura	Abitazioni occupate	86,47	13,53	13,46	0,07
	Abitazioni non occupate	83,70	16,30	16,19	0,11
Numero di piani/livelli dell'abitazione	Abitazioni occupate	97,95	2,05	1,94	0,11
	Abitazioni non occupate	88,58	11,42	11,33	0,09
Superficie	Abitazioni occupate	95,17	4,83	3,85	0,98
	Abitazioni non occupate	86,08	13,92	12,83	1,09
Disponibilità di acqua potabile	Abitazioni occupate	98,38	1,62	0,87	0,75
	Abitazioni non occupate	88,94	11,06	10,76	0,30
Numero di vasche o docce (fino a 3)	Abitazioni occupate	99,01	0,99	0,84	0,15
	Abitazioni non occupate	88,40	11,60	11,45	0,15
Numero di vasche o docce (oltre 3)	Abitazioni occupate	99,99	0,01	0,01	0,00
	Abitazioni non occupate	–	–	–	–
Numero di gabinetti (fino a 3)	Abitazioni occupate	98,40	1,60	1,02	0,58
	Abitazioni non occupate	88,14	11,86	11,22	0,64
Numero di gabinetti (oltre 3)	Abitazioni occupate	99,98	0,02	0,01	0,01
	Abitazioni non occupate	–	–	–	–
Disponibilità di acqua calda	Abitazioni occupate	98,28	1,72	1,33	0,39
	Abitazioni non occupate	87,22	12,78	12,31	0,47
Riscaldamento comune con acqua calda	Abitazioni occupate	94,83	5,17	2,02	3,15
	Abitazioni non occupate	87,94	12,06	11,26	0,80
Fonte energetica per la produzione di acqua calda	Abitazioni occupate	99,52	0,48	0,48	0,00
	Abitazioni non occupate	95,04	4,96	4,96	0,00
Tipo di impianto di riscaldamento	Abitazioni occupate	99,01	0,99	0,76	0,23
	Abitazioni non occupate	86,85	13,15	12,71	0,44
Combustibile per il riscaldamento	Abitazioni occupate	97,05	2,95	0,88	2,07
	Abitazioni non occupate	89,20	10,80	10,31	0,49
Ristrutturazione impianti negli ultimi 10 anni	Abitazioni occupate	98,13	1,87	1,87	0,00
	Abitazioni non occupate	–	–	–	–
Interventi strutturali negli ultimi 10 anni	Abitazioni occupate	96,32	3,68	2,68	1,00
	Abitazioni non occupate	–	–	–	–
Ristrutturazione interni negli ultimi 10 anni	Abitazioni occupate	97,78	2,22	2,22	0,00
	Abitazioni non occupate	–	–	–	–
Disponibilità box	Abitazioni occupate	96,69	3,31	3,31	0,00
	Abitazioni non occupate	–	–	–	–
Disponibilità posto auto garage	Abitazioni occupate	95,90	4,10	4,10	0,00
	Abitazioni non occupate	–	–	–	–
Disponibilità posto auto cortile	Abitazioni occupate	95,73	4,27	4,27	0,00
	Abitazioni non occupate	–	–	–	–
Disponibilità telefono fisso	Abitazioni occupate	99,03	0,97	0,97	0,00
	Abitazioni non occupate	–	–	–	–

Nella tavola 1.34 si osserva in generale come le imputazioni si sono concentrate soprattutto **sulle abitazioni non occupate**, per tutti quei quesiti a cui il rilevatore (cui era delegata la risposta in assenza di una persona

occupante l'abitazione) non è stato in grado di fornire risposta. Il numero di stanze o di zone adibite alla cottura di cibi (cucina, cucinino, angolo cottura), il tipo di riscaldamento o di combustibile usato per alimentare l'impianto di riscaldamento, sono caratteristiche dell'abitazione non visibili dall'esterno e probabilmente su questi aspetti il rilevatore non è stato in grado di ottenere risposte, seppure approssimative o indirette, neppure dai vicini di casa.

Per quanto riguarda le **abitazioni occupate** le mancate risposte si addensano ancora sui quesiti relativi al numero di cucinini e di angoli cottura, o al numero di stanze ad uso professionale.

La rilevazione degli **edifici** è stata affidata completamente ai rilevatori, che hanno talvolta lasciato in bianco le risposte (Tavola 1.35) relative al numero di scale (12,19 per cento di mancate risposte), al numero totale di interni (6,77 per cento) e alla presenza di piani interrati (5,24 per cento).

La variabile numero di *piani fuori terra* dell'edificio è stata imputata molte volte (15,80 per cento), non tanto per una mancata risposta, ma piuttosto in conseguenza di errori, dovuti anche alla difficoltà della lettura ottica di interpretare i valori numerici scritti a mano dai rilevatori.

Tavola 1.35 - Modifiche introdotte dal processo di correzione sulle variabili relative agli edifici *(valori percentuali)*

VARIABILI	Sul totale degli edifici			
	Esatte	Imputate	Di cui mancanti	Di cui incompatibili
Tipo di costruzione	98,28	1,72	1,24	0,48
Utilizzazione edificio	95,47	4,53	2,96	1,57
Tipo di edificio	95,57	4,43	2,90	1,53
Contiguità con altri edifici	96,40	3,60	3,60	0,00
Materiale di costruzione	95,57	4,43	4,02	0,42
Epoca di costruzione	95,50	4,50	4,32	0,18
Presenza di ascensore	95,53	4,47	4,41	0,06
Stato di conservazione	95,97	4,03	4,02	0,00
Numero di piani fuori terra	84,20	15,80	3,52	12,28
Presenza di piani interrati	94,76	5,24	5,24	0,00
Numero di scale	82,49	17,51	12,19	5,32
Totale interni dell'edificio	87,74	12,26	6,77	5,49

1.10 - Una valutazione degli errori presenti nei dati sulla base delle modifiche introdotte dalle procedure e alcuni confronti con il Censimento del 1991

Una valutazione complessiva della qualità dei dati del Censimento del 2001 si può ottenere a partire dalla tavola 1.36, in cui sono riportati i tassi di correzione e di mancata risposta per le principali variabili relative alla popolazione residente (in famiglia o in convivenza). Si osserva come gli interventi di correzione sui dati raccolti siano derivati in massima parte dalla necessità di imputare le mancate risposte, e molto meno da incoerenze o incompatibilità fra le risposte fornite. I tassi di correzione riscontrati si mantengono su livelli modesti, a dimostrazione della buona qualità complessiva della rilevazione censuaria.

Con le dovute cautele è possibile effettuare un confronto fra i due censimenti consecutivi, quelli del 1991 e del 2001, esaminando i tassi di correzione per un sottoinsieme delle principali variabili rilevate in entrambi. Le differenze risentono ovviamente delle diverse formulazioni dei quesiti e delle diverse metodologie utilizzate per la correzione. Nelle tavole 1.37 e 1.38 si osserva che generalmente la percentuale di imputazioni è stata inferiore nel 2001 rispetto al 1991 e che le differenze sono per lo più trascurabili o attribuibili a un diverso numero di modalità di risposta nei due casi.

Tavola 1.36 - Tassi di imputazione e tassi di mancata risposte di alcune variabili relative alla popolazione residente in Italia (famiglie e convivenze) *(valori percentuali)*

VARIABILI	Imputate	Di cui dati mancanti
Relazione di parentela con l'intestatario del Foglio di famiglia	3,42	1,74
Motivo di permanenza nella convivenza	6,57	3,86
Sesso	1,46	1,34
Data di nascita	1,58	0,69
Stato civile	1,51	1,03
Stato civile precedente l'ultimo matrimonio	5,38	5,38
Data di matrimonio	3,64	2,48
Luogo di nascita	3,02	2,91
Stato estero di nascita	0,25	0,22
Cittadinanza	6,67	6,46
Origine della cittadinanza italiana	2,70	2,62
Stato estero di cittadinanza	0,34	0,32
Stato estero di cittadinanza precedente l'acquisizione	0,03	0,03
Anno di trasferimento in Italia	0,21	0,17
Motivo di trasferimento in Italia	0,27	0,20
Presenza alla data del censimento	0,89	0,89

Tavola 1.37 - Confronto fra tassi di correzione al Censimento 2001 e al Censimento 1991 per alcune variabili relative agli individui residenti

VARIABILI	Percentuale di risposte imputate nel 2001	Percentuale di risposte imputate nel 1991	Differenza delle percentuali (2001-1991)	Note
Relazione di parentela	3,42	3,09	0,33	
Sesso	1,46	1,76	-0,30	
Luogo di nascita	3,02	5,21	-2,19	
Giorno di nascita	0,95	2,12	-1,17	
Mese di nascita	0,92	2,03	-1,11	
Anno di nascita	1,17	2,65	-1,48	
Stato civile	1,48	3,07	-1,59	
Mese di matrimonio	3,12	3,27	-0,15	
Anno di matrimonio	3,03	3,14	-0,11	
Cittadinanza	6,67	1,11	5,56	
Anno di trasferimento	0,40	0,38	0,02	
Presenza alla data censimento	0,89	1,14	-0,25	
Titolo di studio	5,27	9,07	-3,80	
Specializzazione post laurea	2,36	1,42	0,94	
Iscrizione ad un corso di studi	1,13	26,29	-25,16	Nel 1991 è differente rispetto al 2001 (4 alternative invece di un si/ no nel 2001)
Frequenza corso professionale	12,12	21,47	-9,35	
Condizione professionale	7,37	4,92	2,45	Nel 1991 rispetto al 2001 si hanno solo 8 alternative invece di 10
Posizione nella professione	3,19	5,41	-2,22	Nel 1991 14 alternative, nel 2001 12 alternative
Lavoratori retribuiti alle dipendenze	1,28	2,20	-0,92	Nel 1991 è differente rispetto al 2001 perché oltre ai lavoratori retribuiti si considerano anche gli apprendisti
Rientro giornaliero (pendolarismo)	4,19	7,68	-3,49	
Orario uscita (pendolarismo)	2,44	3,40	-0,96	Nel 1991 è differente rispetto al 2001 perché nel 2001 è chiesto l'ora di uscita mentre nel 1991 vi è un raggruppamento in 8 classi
Tempo impiegato	2,66	3,07	-0,41	Nel 1991 rispetto al 2001 si hanno solo 4 alternative invece di 5
Mezzo di trasporto	2,99	3,37	-0,38	Nel 1991 rispetto al 2001 si hanno solo 10 alternative invece di 12

Tavola 1.38 - Confronto fra tassi di correzione al Censimento 2001 e al Censimento 1991 per alcune variabili relative alle abitazioni occupate

QUESITI	Percentuale di risposte modificate 2001	Percentuale di risposte modificate 1991	Differenza delle percentuali 2001-1991	Note
Struttura portante	5,82	6,25	-0,43	
Epoca di costruzione del fabbricato	5,87	4,81	1,06	
Numero di piani fuori terra	17,16	3,27	13,89	Nel 2001 viene chiesto il numero esatto dei piani fuori terra, mentre nel 1991 le modalità di risposta sono in classi
Presenza di ascensore	5,82	5,46	0,36	
Proprietario dell'abitazione	2,44	1,33	1,11	
Titolo di godimento	1,83	1,77	0,06	
Numero di stanze adibite ad abitazione	0,99	1,51	-0,52	
Numero di stanze adibite ad altro uso	6,75	0,09	6,66	
Numero di cucine con caratteristiche di stanza	6,38	7,43	-1,05	
Superficie dell'abitazione	4,83	16,96	-12,13	
Acqua potabile	1,62	1,04	0,58	
Numero di gabinetti	1,6	1,41	0,19	
Presenza di vasca o doccia	0,99	1,73	-0,74	
Tipo di impianto di riscaldamento	0,99	2,45	-1,46	
Tipo combustibile per riscaldamento	3,49	6,05	-2,56	
Disponibilità di acqua calda per uso igienico-sanitario	1,72	2	-0,28	
Impianto comune o meno a quello di riscaldamento	5,34	5,73	-0,39	
Disponibilità di telefono	0,97	1,14	-0,17	

1.11 - Un'analisi multidimensionale della qualità dei dati censuari a livello provinciale

Allo scopo di evidenziare eventuali comportamenti omogenei delle province italiane riguardo agli indicatori di qualità calcolati sui tassi di risposte esatte (non modificate dalle procedure di correzione), si è effettuata un'analisi di tipo multivariato.

Si è considerato un sottoinsieme delle risposte ai quesiti del Modello Istat CP.1, cioè quelle relative alle variabili anagrafiche, alla cittadinanza e al titolo di studio. La matrice unità-variabili, costruita per l'analisi, è composta da 103 righe, le unità statistiche che corrispondono alle province italiane, e dalle colonne, che corrispondono alla percentuale di risposte esatte per ciascuna delle 20 variabili considerate.

Per avere una sintesi degli indicatori la tecnica di analisi multidimensionale ritenuta più appropriata è l'analisi in componenti principali, che consente di ottenere un numero limitato di nuove variabili, come "fattori non osservabili".

La matrice di correlazione tra gli indicatori considerati evidenzia che le variabili originarie hanno scarsa correlazione tra loro, così da non poterne escludere a priori nessuna.

I primi due autovalori (Tavola 1.39) spiegano circa il 77 per cento dell'inerzia totale della nuvola dei "punti-province" nello spazio dei 20 indicatori considerati. La prima componente principale spiega circa il 53 per cento dell'inerzia complessiva, mentre la seconda componente spiega un ulteriore 24 per cento circa, il che induce a scegliere le prime due componenti principali come sintesi delle variabili di partenza.

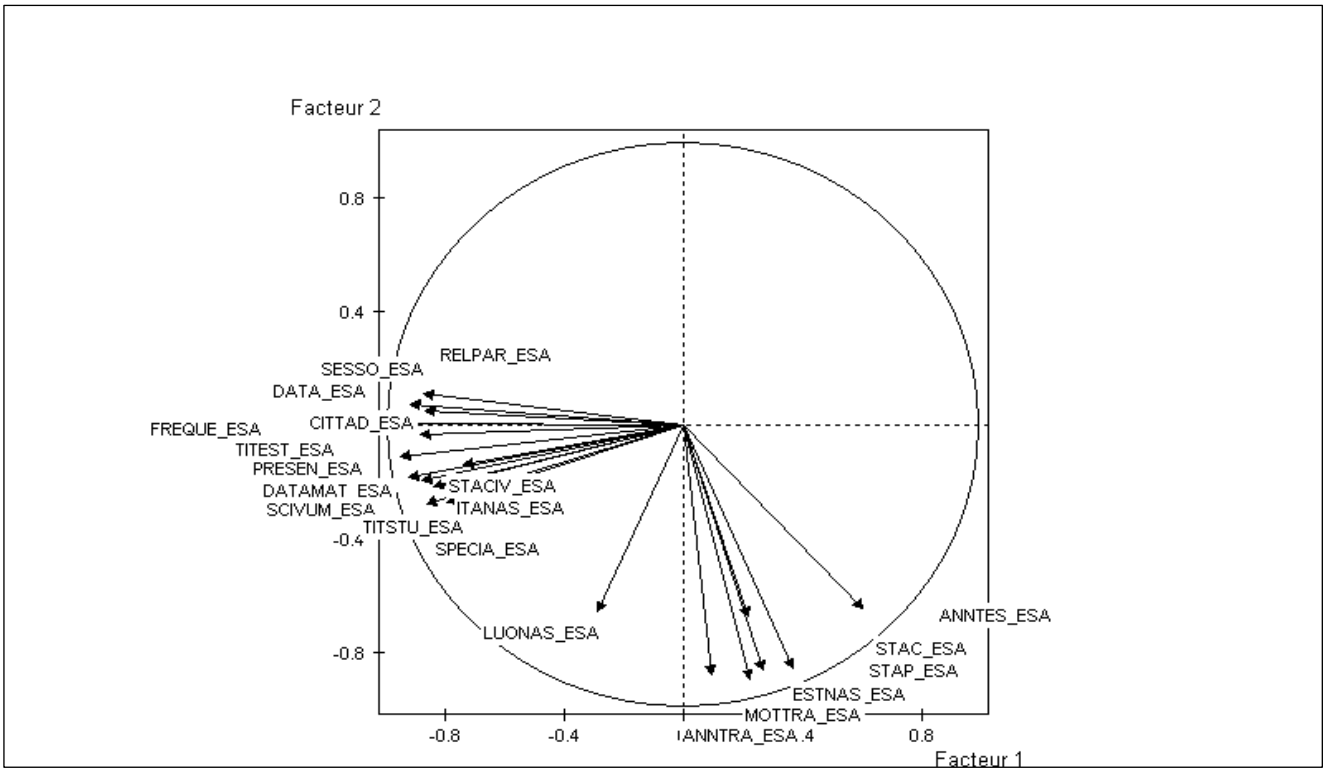
Tavola 1.39 - Tavola degli autovalori

NUMERO COMPONENTE	Autovalore λ_i	Percentuale di inerzia spiegata	Percentuale di inerzia cumulata
1	10,6288	53,14	53,14
2	4,7352	23,68	76,82
3	1,5020	7,51	84,33
4	0,7923	3,96	88,29
5	0,5030	2,52	90,81

20	0,0161	0,08	100,00

Per interpretare le componenti principali si analizzano le coordinate degli indicatori originari[7] (dove il suffisso ESA ricorda che si sta lavorando sulla percentuale di risposte "esatte"), rispetto ai due assi fattoriali, che corrispondono ai coefficienti di correlazione fra ciascun indicatore e ciascun asse, come evidenziate dal "cerchio delle correlazioni" (Figura 1.11).[8] L'interpretazione dell'asse viene fatta osservando quali variabili si concentrano su una polarità e quali su quella opposta.

Figura 1.11 - Cerchio delle correlazioni



La prima componente principale risulta fortemente correlata, in senso negativo, con le variabili inerenti ai quesiti anagrafici e con quelli relativi al titolo di studio (13 variabili su 20). Tale componente spiega la maggior parte del modello.

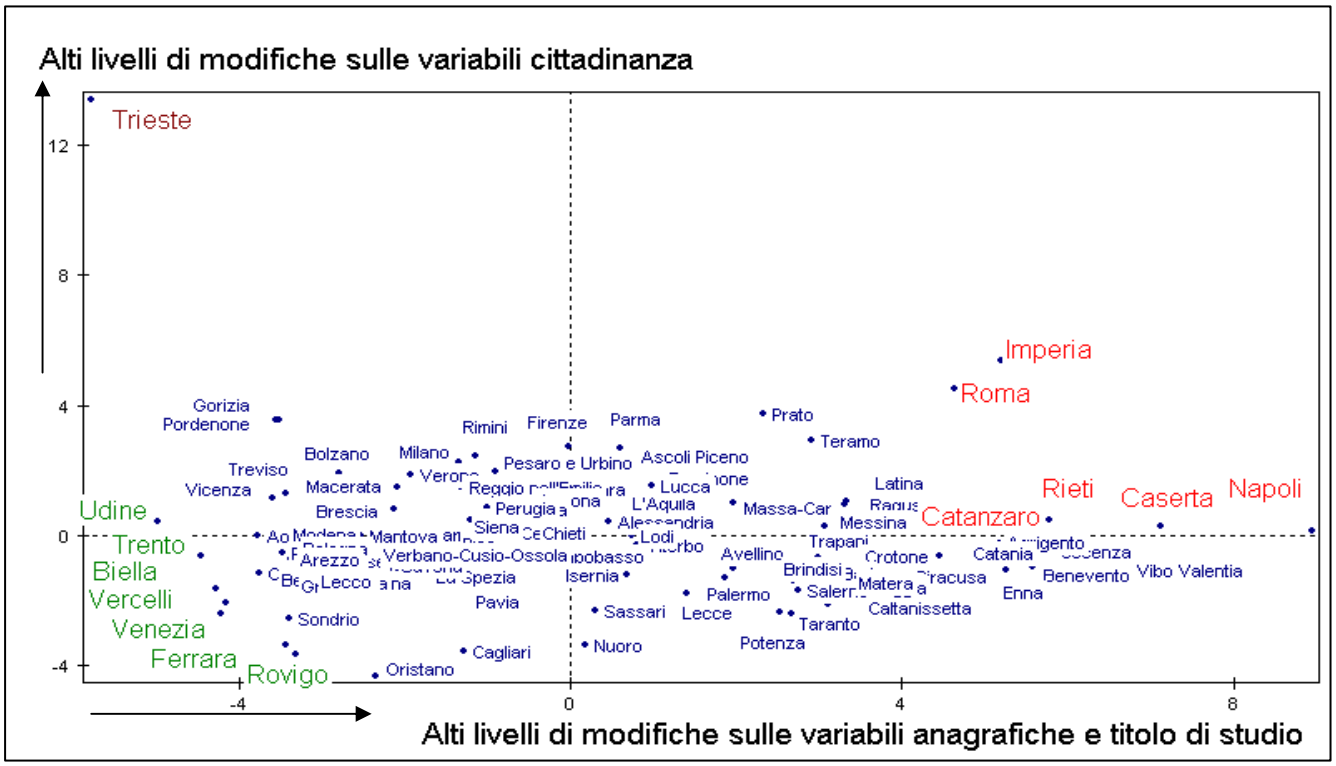
Tenuto conto del segno dei coefficienti di correlazione tra gli indicatori e il primo asse principale e muovendosi lungo tale asse, da sinistra verso destra, si osservano alti livelli di modifiche (correzioni) sulle variabili anagrafiche e titolo di studio.

[7] Per il significato dei nomi delle variabili si veda l'appendice A.
[8] Tanto più una variabile forma un angolo piccolo con la dimensione fattoriale, tanto più è correlata con il fattore e determina l'interpretazione dell'asse.

Il secondo asse principale risulta invece influenzato maggiormente dagli indicatori relativi ai quesiti inerenti la "condizione di cittadinanza straniera o di nascita all'estero" (cittadino italiano nato all'estero o cittadino straniero). Tenendo conto del segno negativo dei coefficienti di correlazione e muovendosi lungo tale asse, dal basso verso l'alto, si osserva come tale condizione abbia comportato un maggior numero di interventi di correzione su tali variabili, quindi "alti livelli di modifiche sulle variabili relative alla cittadinanza".

In base al comportamento degli indicatori, che i singoli assi rappresentano, e alla posizione delle province sul piano principale (Figura 1.12) si notano, procedendo in senso orario a partire dal primo quadrante, in alto a destra, le province con i livelli più alti di modifiche relative ai quesiti rappresentati sul secondo asse principale (cittadinanza) e con valori inferiori nelle percentuali di modifiche relativamente alle variabili rappresentate sul primo asse principale (anagrafiche e titolo di studio). Analogamente possono esser interpretati gli altri quadranti.

Figura 1.12 - Proiezione delle 103 province italiane al Censimento 2001 sul piano principale



Tenuto conto del significato degli assi è possibile affermare che le province di Imperia, Roma, Rieti, Caserta, Napoli e Catanzaro sono quelle che si sono comportate peggio, essendo state maggiormente interessate da interventi nella fase di correzione. Al contrario Udine, Trento, Biella, Ferrara, Venezia, Vercelli e Rovigo sono quelle i cui dati hanno subìto il minor numero di imputazioni/correzioni, e per le quali si può parlare di una maggiore attenzione e cura nella compilazione del questionario.

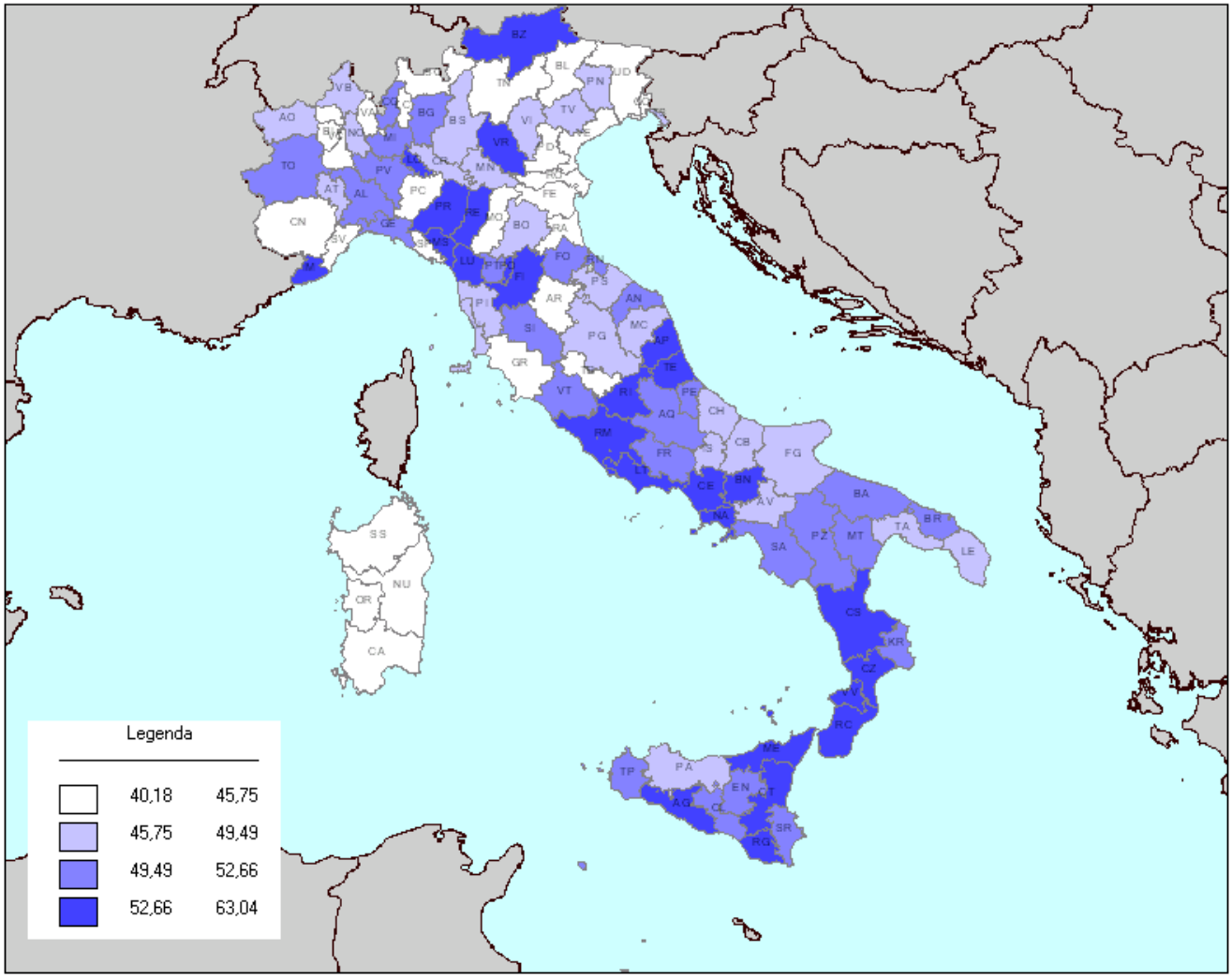
La provincia di Trieste si distingue per valori particolarmente elevati nelle percentuali di modifiche sulle variabili inerenti al luogo di nascita. Tale anomalia è giustificata dal fatto che molti dei residenti in questa città (ad esempio di origine istriana o dalmata) hanno dichiarato di essere nati in Italia, perché nati in comuni allora appartenenti al territorio italiano, anche se questi sono stati successivamente ceduti alla ex-Jugoslavia. Avrebbero dovuto dichiarare invece di essere nati all'estero, come richiesto dalla norma di compilazione che chiede il riferimento alla situazione territoriale del 2001. Di conseguenza per questi casi, nella fase di imputazione/correzione dei dati delle variabili connesse alla cittadinanza, è stato necessario valorizzare la variabile stato di nascita estero (come ex-Jugoslavia).

1.12 - Gli errori per record

Considerando complessivamente tutte le variabili del modello relativo alle persone residenti si sono conteggiate le modifiche intervenute nei dati nel corso del processo di validazione.

Una prima valutazione sulla presenza di errori è stata effettuata considerando, per ogni provincia, la percentuale di casi in cui si sono riscontrati due o più errori per record. La graduatoria delle province è stata evidenziata nel cartogramma (Figura 1.13) in cui sono rappresentati quattro gruppi di province, quelle colorate di bianco hanno due o più errori per record in una percentuale fra 40,18 per cento e 45,75 per cento, mentre quelle più scure fra 52,66 e 63,04 per cento, quindi con più di errori per record.

Figura 1.13 - Cartogramma del numero di modifiche per record dei residenti in famiglia per provincia



Si è anche calcolato il numero medio di modifiche a livello provinciale come rapporto fra la somma, per tutti i record della provincia, del numero di modifiche di correzione ed il totale delle risposte (su 59 variabili considerate per le persone residenti in famiglia). Tale rapporto assume in media nazionale un valore pari a 2,48 interventi (numero standard di errori sanati) per record, mentre a livello provinciale è un buon indicatore sintetico della qualità della compilazione dei modelli. Le province di Roma, Napoli, Rieti, Imperia, Benevento, Caserta (Tavola 1.40) hanno un valore medio di modifiche per record più elevato delle altre (oltre tre errori per record), mentre le province più virtuose sono Biella, Oristano, Udine, Trento, Ferrara, Venezia, con indicatori molto più bassi.

Analoghe considerazioni emergono dalla graduatoria per provincia esaminata nella tavola 1.41, dove si osserva che le province che presentano le più alte percentuali di record con oltre due errori sono Roma, Rieti, Prato, Napoli, eccetera, mentre le province in cui tale percentuale è minore sono Oristano, Venezia, Grosseto, Ferrara.

Tavola 1.40 - Numero medio di errori per record dei residenti in famiglia

PROVINCE PEGGIORI	Numero medio di errori per record	PROVINCE MIGLIORI	Numero medio di errori per record
Roma	3,41	Venezia	1,75
Napoli	3,34	Ferrara	1,77
Rieti	3,23	Trento	1,81
Imperia	3,18	Udine	1,82
Benevento	3,07	Oristano	1,82
Caserta	3,06	Biella	1,83

Tavola 1.41 - Graduatoria delle province per percentuale di errori *(ordinate su due errori o più, in senso decrescente per le peggiori, e in senso crescente per le migliori)*

PROVINCE PEGGIORI	0 errori	1 errore	2 errori o più	PROVINCE MIGLIORI	0 errori	1 errore	2 errori o più
Roma	18,67	18,29	63,04	Oristano	34,95	24,87	40,18
Rieti	20,23	18,86	60,91	Venezia	32,74	25,91	41,35
Prato	20,04	20,55	59,41	Grosseto	33,94	24,46	41,59
Napoli	22,55	18,69	58,76	Ferrara	32,96	25,2	41,84
Imperia	22,68	19,13	58,19	Cagliari	32,76	25,37	41,87
Benevento	22,17	19,85	57,98	Gorizia	32,67	25,38	41,95
Parma	21,43	20,74	57,83	Udine	32,22	25,61	42,17
Teramo	22,22	20,86	56,93	Trento	30,91	26,52	42,56
Ragusa	23,90	19,73	56,37	Savona	34,25	23,03	42,72
Catanzaro	23,63	20,29	56,08	Rovigo	31,61	25,41	42,98
Caserta	24,60	20,01	55,39	La Spezia	32,44	24,25	43,3
Lodi	23,48	21,15	55,37	Biella	31,95	24,52	43,52
Catania	24,65	20,28	55,07	Arezzo	30,87	25,5	43,63
Reggio nell'Emilia	22,78	22,40	54,82	Belluno	31,31	25	43,69
Reggio di Calabria	24,71	20,59	54,70	Vercelli	32,2	23,92	43,88
Cosenza	24,48	21,02	54,50	Terni	31,7	24,12	44,19
Messina	25,11	20,79	54,10	Ravenna	30,87	24,78	44,35
Latina	24,95	21,10	53,95	Padova	29,17	26,3	44,53
Ascoli Piceno	24,61	21,70	53,69	Nuoro	30,76	24,5	44,73

1.13 - Gli indici di dissomiglianza

Per la valutazione dell'impatto delle procedure di controllo e correzione sui dati sono stati elaborati opportuni indici di dissomiglianza, indicativi delle differenze nelle distribuzioni di frequenza dei dati iniziali (grezzi) rispetto ai dati finali (puliti).

Per studiare il cambiamento sulla forma delle distribuzioni introdotto dalle procedure di correzione, e mettere quindi in evidenza eventuali effetti sistematici e/o di distorsione, si sono considerati inizialmente diversi indici di dissomiglianza. Gli indici esaminati, tutti indici relativi, considerano le due distribuzioni: A, la distribuzione iniziale grezza, e B, la distribuzione finale pulita. Se *k* indica il numero di modalità della variabile esaminata, "f" indica la frequenza relativa con cui la i-esima delle *k* modalità è rappresentata nella popolazione ed *r* è il valore dell'esponente, due indici possibili sono espressi nelle formule seguenti:

$$z = (½)\ (\sum_i |f_{Ai} - f_{Bi}|^r) \qquad (1)$$

$$d'_s = 1-(1/k) \sum_i f'_{(A,B)i} / f_{(A,B)i} \quad (2)$$

dove $f(A,B)_i = \max(fA_i, fB_i)$ e i $f'(A,B)_i = \min(fA_i, fB_i)$.

Considerando come varia il primo indice al variare dell'esponente *r*, si evidenzia che, già per $r \geq 2$ i valori assunti dall'indice corrispondono a numeri molto piccoli, talvolta difficilmente interpretabili.

Il secondo indice è basato su una ponderazione del rapporto fra frequenza minima e massima per ogni modalità delle due distribuzioni considerate ed è meno sensibile, rispetto al precedente, al numero di modalità della variabile, ma è più complesso da calcolare.

In conclusione, anche per la leggerezza di calcolo, si è adottato l'indice semplice di dissomiglianza ponendo l'esponente $r = 1$.

$$z = (½) (\sum_i |f_{Ai} - f_{Bi}|)$$

Tavola 1.42 - Indicatori sulla diversità delle distribuzioni della singola variabile prima e dopo la fase di correzione

MODALITÀ DI UNA GENERICA VARIABILE	Frequenza percentuale sui dati (validi) della variabile grezza	Frequenza percentuale sui dati (validi) della variabile corretta	Differenza delle percentuali in valore assoluto
1ª	fA_1	fB_1	$\lvert fA_1 - fB_1 \rvert$
2ª	fA_2	fB_2	$\lvert fA_2 - fB_2 \rvert$
3ª	fA_3	fB_3	$\lvert fA_3 - fB_3 \rvert$
i-esima	fA_i	fB_i	$\lvert fA_i - fB_i \rvert$
N-esima	fA_N	fB_N	$\lvert fA_N - fB_N \rvert$
Blank	$fA_{blank} = 0$		
Fuori *range*	$fA_{f.r.} = 0$		
Totale	**100**	**100**	**z = somma degli scarti in valore assoluto/ 2**

Il valore che l'indice *z* può assumere è compreso tra 0 (somiglianza perfetta delle frequenze relative nelle due distribuzioni) e 1 (massima dissomiglianza). Per renderne più agevole la lettura, gli indici di dissomiglianza sono riportati in percentuale nelle tabelle allegate, dunque con valori tra 0 e 100.

Per il calcolo dell'indice di dissomiglianza si è valutato se fosse il caso o meno di considerare anche il *blank* fra le modalità messe a confronto, cioè la non risposta.

L'insieme delle variabili rilevate può esser suddiviso in due gruppi: quelle che non ammettono il *blank* tra i valori possibili, e sono quindi obbligatorie per tutti (ad esempio il sesso, la relazione di parentela, la cittadinanza) e quelle che lo ammettono, e sono quindi non obbligatorie (ad esempio lo stato civile prima del matrimonio per chi dichiara di essersi sposato, il motivo di trasferimento in Italia per chi dichiara di non essere italiano).

Nel secondo caso, di variabili non obbligatorie, il numero di *blank* nei dati grezzi comprende sia i casi esatti di chi non doveva rispondere, sia i casi errati in cui la risposta non è stata fornita, anche se teoricamente dovuta. Per i casi ritenuti errati la procedura di correzione modificherà i *blank*, che sono vere e proprie mancate risposte (errori) in valori validi (non nulli) e manterrà inalterata la quota di *blank* esatti. Nel caso di risposte non obbligatorie, che devono esser fornite solo da insiemi ristretti di popolazione (ad esempio gli stranieri) la quota di *blank* esatti può essere molto elevata (sia nei dati grezzi che nei dati puliti), e conseguentemente le frequenze relative delle modalità diverse da *blank* possono essere piccole. In queste situazioni, l'indice di dissomiglianza calcolato considerando anche il *blank* fra le modalità, può risultare comunque vicino a zero, anche nel caso di un piano di correzione che modifichi in modo sostanziale le distribuzioni prima/dopo delle modalità diverse da *blank*.

Nel caso di variabili con risposta obbligatoria i *blank* non sono ammessi e certamente corrispondono, nella distribuzione grezza a mancate risposte, cioè a errori che verranno corretti, in modo da escludere la modalità *blank* dalla distribuzione dei dati "dopo" la correzione.

Se in questo caso si considerassero i *blank* fra le modalità delle due distribuzioni da confrontare per il calcolo dell'indice di dissomiglianza potrebbe accadere che, semplicemente a causa di un elevato tasso di mancate risposte (molti *blank* nella distribuzione grezza), l'indice potrebbe risultare alto, pur in presenza di una procedura di correzione che non altera le distribuzioni relative alle modalità non *blank*.

Per mettere meglio in evidenza il vero scostamento delle distribuzioni prima e dopo la correzione, il calcolo dell'indice di dissomiglianza scelto è stato eseguito escludendo i valori *blank* dalle distribuzioni in entrambi i casi di variabili obbligatorie e non obbligatorie e considerando le sole modalità diverse da *blank*.

Gli indici di dissomiglianza, al netto della modalità *blank*, sono stati calcolati per le variabili contenute nei Fogli di famiglia (Modello Istat CP.1) e di convivenza (Modello Istat CP.2) considerando le distribuzioni a livello Italia, a livello di provincia e per i comuni capoluogo (Italia, 103 province e 103 comuni capoluogo di provincia) e memorizzate nell'archivio di qualità.

L'analisi si articola per gruppi di variabili (residenti in famiglia, residenti in convivenza, temporaneamente dimoranti in alloggio, temporaneamente dimoranti in convivenza, edifici, abitazioni occupate, abitazioni non occupate). Il significato dei nomi delle variabili è riportato nell'appendice A.

Gli indici di dissomiglianza, in ordine decrescente per le variabili del questionario di rilevazione Modello Istat CP.1 riguardanti informazioni sulla persona, sono riportati nella tavola 1.43.

I valori degli indicatori sono generalmente molto piccoli, esprimendo complessivamente un effetto neutrale dei piani di correzione, che in definitiva hanno riprodotto nei dati finali la stessa distribuzione dei valori iniziali esatti.

Tavola 1.43 - Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili dei residenti in famiglia (per il significato delle variabili si veda l'Appendice A) *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Motnes	23,97	Corfor	1,29
Contin	19,35	Sireca	1,18
Ohow	9,93	Setatt	1,08
Hasvol	6,63	Annter	1,02
Oresp	6,48	Specie	0,90
Condii	5,46	Iscriz	0,86
Posiz	5,25	Rientr	0,79
Motivo	4,14	Situat	0,71
Stap	4,01	Mezzot	0,59
Novant	3,67	Luonas	0,54
Anntra	3,62	Tipcor	0,51
Tiprap	3,21	Cercat	0,50
Orelav	3,12	Vissut	0,41
Stac	2,88	Amat	0,34
Mottra	2,80	Itanas	0,25
Hadip	2,45	Titest	0,18
Tipatt	2,42	Anas	0,18
Rappor	2,21	Temimp	0,18
Freque	2,20	Cittad	0,16
Luosl	2,20	Staciv	0,13
Titstu	2,18	Dimanp	0,09
Dadove	1,98	Dueset	0,07
Estnas	1,91	Sesso	0,05
Relpar	1,68	Scivum	0,05

Il caso di massimo indice di dissomiglianza si osserva per la variabile Motnes (motivo per cui non si sono effettuate ore di lavoro), per la quale gli interventi di correzione hanno modificato la distribuzione di frequenza finale rispetto a quella iniziale. Confrontando i dati finali con quelli iniziali si osserva un incremento dei *blank* derivanti dalla eliminazione diffusa di tutte le modalità originariamente presenti, ma soprattutto per l'eliminazione della modalità voce "altro", che nell'1,08 per cento dei casi è stata cancellata dalle procedure di correzione.

Per meglio evidenziare gli effetti dei piani di compatibilità e correzione si possono esaminare le matrici o tabelle di transizione, per osservare su quali modalità sono intervenute prevalentemente le procedure di correzione. In questo tipo di matrice, per ogni specifica variabile, si incrociano le frequenze delle modalità

iniziali con quella delle modalità finali. Nelle generica cella *ij* è contenuta la frequenza con cui la modalità *i* è stata modificata dalla procedura nella modalità *j*.

La casella *ii* della diagonale principale definisce le frequenze delle "permanenze", cioè per quante volte la procedura ha lasciata inalterata la modalità *i* della variabile in esame. Nella matrice di transizione (al contrario di quanto si è fatto per il calcolo degli indici di dissomiglianza) vengono considerati anche i *blank* fra le modalità, per poter esaminare se il processo di imputazione delle mancate risposte ha agito in modo da rispettare la distribuzione della variabile, per come essa si presenta nei casi esatti.

Per meglio evidenziare l'effetto netto della correzione la matrice di transizione viene depurata dalle "permanenze" e contiene i soli casi in cui si sono apportate modifiche ai dati iniziali (quindi escludendo la diagonale principale della tavola, che rappresenta i casi non modificati). Per la variabile Motnes si osserva come molte risposte non dovute siano state modificate in *blank* e che invece i *blank* originari considerati errati siano stati imputati principalmente alle voci "1=ferie" o "2=malattia" oppure "7=altro", quindi seguendo una distribuzione di frequenza diversa da quella dei dati originari presenti.

Tavola 1.44 - Effetti dell'imputazione sulla variabile Motnes inclusi i *blank* *(valori assoluti e percentuali)*

VARIABILE MOTNES	Frequenze assolute "prima"	Frequenze assolute "dopo"	Frequenze assolute (dopo–prima)	Frequenze percentuali "prima"	Frequenze percentuali "dopo"	Frequenze percentuali (dopo–prima)
Blank	54.862.959	55.575.552	712.593	96,94	98,20	1,26
1. Ferie	220.751	213.666	-7.085	0,39	0,38	-0,01
2. Malattia	260.540	236.838	-23.702	0,46	0,42	-0,04
3. Maternità	201.841	185.685	-16.156	0,36	0,33	-0,03
4. Aspettativa	45.346	28.682	-16.664	0,08	0,05	-0,03
5. Cassa integrazione guadagni	44.183	29.696	-14.487	0,08	0,05	-0,03
6. Mancanza commesse	65.616	42.772	-22.844	0,12	0,08	-0,04
7. Altro	892.785	281.130	-611.655	1,58	0,50	-1,08
Totale (con *blank*)	**56.594.021**	**56.594.021**	**0**	**100,00**	**100,00**	**0,00**

Tavola 1.45 - Effetti dell'imputazione sulla variabile Motnes, esclusi i *blank* *(valori assoluti e percentuali)*

VARIABILE MOTNES	Frequenze assolute "prima"	Frequenze assolute "dopo"	Frequenze percentuali "prima"	Frequenze percentuali "dopo"	Frequenze percentuali (dopo–prima)
1. Ferie	220.751	213.666	12,75	20,98	8,23
2. Malattia	260.540	236.838	15,05	23,25	8,20
3. Maternità	201.841	185.685	11,66	18,23	6,57
4. Aspettativa	45.346	28.682	2,62	2,82	0,20
5. Cassa integrazione guadagni	44.183	29.696	2,55	2,92	0,36
6. Mancanza commesse	65.616	42.772	3,79	4,20	0,41
7. Altro	892.785	281.130	51,57	27,60	-23,97
Totale (senza *blank*)	**1.731.062**	**1.018.469**	**100,00**	**100,00**	**0,00**

Nella tavola di cui sopra (Tavola 1.45), considerando la semisomma dei valori assoluti dell'ultima colonna si ottiene l'indice di dissomiglianza (23,97 per cento).

Tavola 1.46 - Matrice di transizione sugli effetti dell'imputazione per la variabile Motnes, esclusa la diagonale principale, corrispondente ai valori non modificati dalla procedure di correzione

MOTNES PRIMA	Motnes dopo								
	Blank	Ferie	Malattia	Maternità	Aspet-tativa	Cassa integrazione guadagni	Mancanza commesse	Altro	Totale complessivo
Blank		16,21	3,24	0,11	0,02	0,03	0,03	80,37	100,00
1. Ferie	99,98		0,01	0,00	0,00	0,00	0,00	0,01	100,00
2. Malattia	99,98	0,00		0,00	0,00	0,00	0,00	0,001	100,00
3. Maternità	99,48	0,50	0,00		0,00	0,00	0,00	0,01	100,00
4. Aspettativa	80,77	0,00	0,00	0,00		0,00	0,00	19,23	100,00
5. Cassa integrazione guadagni	89,38	0,10	0,00	0,00	0,00		0,00	10,53	100,00
6. Mancanza commesse	99,91	0,09	0,00	0,00	0,00	0,00		0,00	100,00
7. Altro	100,00	0,00	0,00	0,00	0,00	0,00	0,00		100,00
Totale	**88,66**	**1,76**	**0,35**	**0,01**	**0,00**	**0,00**	**0,00**	**9,20**	**100,00**

Anche la variabile Contin (continua a vivere in alloggi diversi da quello di residenza, per chi ha già indicato una permanenza in altro alloggio o convivenza per un periodo superiore a tre mesi) presenta un indice di dissomiglianza piuttosto elevato (19,35 per cento), dovuto sostanzialmente all'effetto dell'eliminazione di una quota molto rilevante di risposte non dovute ("No").

Gli indici di dissomiglianza delle distribuzioni nazionali per le variabili corrispondenti alle risposte al Modello Istat CP.2, riguardanti informazioni sulle persone **residenti in convivenza**, sono riportati nella Tavola 1.47.

Anche in questo caso la variabile che ha maggiormente modificato la sua distribuzione in seguito alle procedure di correzione è Motnes, seguita da Hadip (variabile che verifica la presenza di dipendenti per i lavoratori autonomi o imprenditori) e Corfor (corsi di formazione professionale).

Anche in questo caso per la variabile Motnes risulta che il numero di cancellazioni prevale sul numero di imputazioni e le cancellazioni riguardano soprattutto la modalità "7= altro". L'aumento di frequenza percentuale delle modalità "1=ferie" e "2=malattia" è dovuto, essenzialmente, proprio a questa diminuzione di numerosità dell'insieme di risposte valorizzate.

Per la variabile Hadip, guardando alla matrice di transizione, si osserva che il numero di eliminazioni della modalità "2=non ha dipendenti retribuiti" è pressoché uguale a quelli della modalità "1=ha dipendenti retribuiti", mentre il numero di imputazioni da *missing* a valore è nettamente superiore verso la modalità "2".

Corfor è la variabile che esprime la frequenza ad un corso di formazione professionale alla data del censimento ("1=sì ", "2=no") e si riferisce a persone di sei anni o più. Le imputazioni (da *blank* a valore) sono più numerose delle eliminazioni (da valore a *blank*). La matrice di transizione mostra che l'aumento della percentuale di risposta "1" è legato non tanto all'imputazione di *missing* quanto a trasformazioni di "2" in "1".

Tavola 1.47 - Dissomiglianza delle distribuzioni a livello Italia per le variabili dei residenti in convivenza (graduatoria decrescente) *(valori percentuali)*

VARIABILI	Dissomiglianza (%)	VARIABILI	Dissomiglianza (%)
Motnes	28,01	Estnas	2,43
Hadip	20,71	Rappor	2,31
Corfor	20,62	Condii	1,97
Orelav	9,49	Specie	1,73
Contin	8,74	Luonas	1,52
Hasvol	8,32	Stac	1,48
Mottra	7,79	Mdac	1,27
Freque	7,06	Dimapc	1,08
Ohow	6,58	Situat	1,07
Oresp	6,33	Titest	1,04
Tiprap	5,18	Novant	0,52
Adac	4,98	Tempp	0,51
Annter	4,67	Staciv	0,42
Titstu	4,22	Cercat	0,37
Iscriz	4,04	Dueset	0,31
Dimcop	4,00	Cittad	0,31
Stap	3,98	Anas	0,27
Posiz	3,54	Vissut	0,21
Anntra	3,37	Itanas	0,18
Motivo	2,96	Presen	0,18
Moperc	2,85	Sesso	0,04

Nella tavola 1.48 sono riportati gli indici di dissomiglianza per le variabili corrispondenti alle risposte della sezione III del questionario di rilevazione, Modello Istat CP.1, che riguarda le informazioni sulle persone **non abitualmente dimoranti nell'alloggio**.

Tavola 1.48 - Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili dei non abitualmente dimoranti in famiglia *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Apres	13,44	Mpres	2,81
Luosl	9,41	Stac	2,38
Condsp	6,31	Setatp	1,92
Rientr	5,06	Dimanc	1,48
Anas	3,86	Mezzot	1,24
Motpre	3,59	Dadove	1,22
Motal	3,50	Temimp	1,22
Tmpvis	3,23	Presen	0,98
Sireca	2,97	Cittad	0,97

In questo caso valori superiori al 10 per cento si registrano per la variabile Apres, che indica da quale anno un cittadino straniero (o apolide) residente all'estero è presente in Italia.

La distribuzione di Apres non è stata fortemente alterata dalle procedure di correzione e la sua moda resta l'anno 2001, anche se le modifiche introdotte con le correzioni hanno principalmente riguardato l'eliminazione della risposta non dovuta relativa proprio alla modalità "2001".

Nella tavola 1.49 sono riportati gli indici di dissomiglianza per le variabili della sezione III del Modello di rilevazione Istat CP.2 e riguardanti le informazioni sulle **persone non abitualmente dimoranti in convivenza**.

Tavola 1.49 - Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili dei non abitualmente dimoranti in convivenza *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Setatp	10,64	Apres	1,98
Tmpvis	10,51	Staciv	1,41
Dimanc	9,72	Anas	1,39
Motpre	4,10	Sesso	0,75
Presen	3,80	Moperc	0,69
Condsp	2,93	Mpres	0,65
Cittad	2,03	Stac	0,24

L'indice di dissomiglianza maggiore si osserva in corrispondenza della variabile Setatp (settore di attività economica). Si tratta comunque di un valore non molto elevato. La matrice di transizione evidenzia come le modalità "4" (commercio, riparazioni, pubblici esercizi, trasporti, comunicazioni) e "6" (servizi sociali alle persone, esclusi i servizi domestici presso famiglie e convivenze) siano quelle più coinvolte nelle correzioni. Si tratta principalmente di eliminazione (per incompatibilità) della modalità "4" e di imputazioni della modalità "6" (per le mancate risposte).

Nella tavola 1.50 sono riportati gli indici di dissomiglianza per le variabili relative al Modello Istat CP.ED, che riguardava il censimento degli **edifici**.

Tavola 1.50 - **Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili degli edifici** *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Npiaft	9,10	Tipmat	0,45
Scale	5,31	Piaint	0,40
Edutil	1,75	Epocos	0,28
Totint	1,16	Contig	0,12
Tipuso	1,10	Ascens	0,11
Tipedi	0,49	Stcons	0,10

In questo caso i valori degli indici di dissomiglianza sono molto contenuti e nessuno supera il dieci per cento. Per la variabile Npiaft (che riporta il numero di piani fuori terra dell'edificio) la moda resta la modalità "2 piani", con una frequenza percentuale che supera il 50 per cento dopo la correzione. La matrice di transizione rivela che il maggior numero di trasformazioni non diagonali (701.484 casi) è avvenuto da "3 piani" a "2 piani".

I valori degli indici di dissomiglianza per le variabili sulle **abitazioni occupate** (sezione I del Modello Istat CP.1) sono forniti nella tavola 1.51.

La distribuzione maggiormente modificata risulta quella della variabile Stusup, che indica se il numero di stanze dell'abitazione sia superiore a tre, e per la quale si osserva un elevato numero di eliminazioni.

La forma originaria della distribuzione dati presentava una discontinuità con un incremento sospetto per il valore "11", dovuto sostanzialmente ad imprecisione della lettura ottica, che in presenza di cifre non chiaramente interpretabile assegnava il valore "1". Il valore modale resta pari a "4 stanze".

La matrice di transizione delle correzioni mostra la presenza di errori sistematici legati alle modalità “40 stanze”, “50 stanze”, “60 stanze”, “70 stanze”, eccetera. Anche in questo caso si tratta di un errori sistematici dovuti alla lettura ottica che la correzione ha annullato.

Tavola 1.51 - Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili delle abitazioni occupate *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Stusup	21,71	Acqcal	0,40
Enriac	19,15	Superf	0,35
Risac	3,70	Gabin	0,35
Cucsta	3,28	Stanuf	0,24
Ticomb	2,17	Titgod	0,22
Cucini	2,16	Imprid	0,20
Gasup	1,70	Nstab	0,19
Angcot	1,54	Nstab	0,19
Cortil	1,19	Npiani	0,15
Opestr	1,14	Propr	0,07
Vadosu	1,07	Openos	0,05
Garage	0,86	Telfis	0,05
Boxpri	0,85	Vasdoc	0,01
Fontac	0,62	Opeimp	0,00

Per le **abitazioni non occupate** (sempre sezione I del Modello Istat CP.1) solo la variabile Stusup ha un indice di dissomiglianza il cui valore si discosta dagli altri (Tavola 1.52). Restano valide le stesse considerazioni svolte per il caso delle abitazioni occupate.

Tavola 1.52 - Modifiche introdotte nei dati dalle procedure di correzione. Dissomiglianza delle distribuzioni a livello Italia per le variabili delle abitazioni non occupate *(valori percentuali)*

VARIABILI	Dissomiglianza	VARIABILI	Dissomiglianza
Stusup	21,77	Fontac	0,79
Gasup	4,75	Fontac	0,79
Vadosu	4,65	Gabin	0,50
Enriac	2,59	Risacq	0,50
Enriac	2,59	Nstab	0,36
Ticomb	2,41	Cucsta	0,33
Ticomb	2,41	Npiani	0,31
Angcot	1,79	Stanuf	0,27
Imprid	1,48	Propr	0,22
Imprid	1,39	Vasdoc	0,19
Superf	1,06	Acqcal	0,13
Cucini	0,92		

1.14 - L’archivio di qualità

L’archivio di qualità nasce dalla necessità di documentare gli effetti dei controlli e delle correzioni stabiliti in fase di progettazione ed eseguiti in fase di produzione e di validazione dei dati. Esso è costituito da uno specifico *database*, che contiene le tabelle di dati aggregati relativi ai confronti fra i dati prima della correzione e dopo la correzione (somme e distribuzioni), alle tabelle di alcuni indici calcolati su microdati modificati dalle

procedure e agli indicatori di dissomiglianza fra le distribuzioni dei dati grezzi e puliti. Attraverso l'elaborazione delle tabelle riassuntive e tramite l'analisi degli indicatori si può procedere alla valutazione *ex post* della qualità dei dati.

L'iter di lavorazione dei dati del 14° Censimento è iniziato con il caricamento dei dati nel *database* di produzione, dove sono avvenute le prime verifiche e correzioni dei codici territoriali di provincia e comune. Si è passati quindi alla correzione dei codici identificativi e alla correzione dei codici di *linkage*. Quindi le procedure di controllo hanno cercato di individuare e di separare gli errori sistematici e gli errori casuali, che dovevano essere opportunamente corretti con l'assegnazione di nuovi valori alle variabili, tali da garantire la coerenza formale e sostanziale dei dati stessi. Infine, per la produzione dei dati aggregati di diffusione, si sono utilizzate procedure di controllo a livello macro, anche con l'utilizzo di dati da fonti esterne.

Le procedure di *editing* hanno, com'è noto, lo scopo di controllare la coerenza dei microdati e di intervenire su di essi in modo da ricostruire coerentemente tutte le informazioni del questionario. I tipi di correzione effettuata sui dati sono fondamentalmente di tre tipi: imputazione di mancata risposta, eliminazione di risposta non dovuta, imputazione di valore ritenuto non coerente o incompatibile (con quello di una o più variabili) o fuori *range*.

Per valutare l'impatto delle procedure di correzione si può ricorrere ad alcuni indicatori, che sintetizzano l'ammontare di interventi effettuati sui dati, ed in particolare si possono calcolare i tassi di correzione e/o imputazione secondo le tre tipologie menzionate: tassi di imputazione di mancata risposta, tassi di imputazione a nuovo valore valido, tassi di eliminazione di valori ritenuti non coerenti. In prevalenza si tratta di indicatori che quantificano l'errore (errore di compilazione del questionario, errore nella fase di raccolta e di acquisizione) rimosso dalle procedure di controllo e correzione. L'ipotesi sottostante è che i valori finali ottenuti siano quelli validi e che pertanto ciò che è stato modificato sia stato errato in origine. Maggiore è il tasso calcolato, maggiore è stato l'effetto delle procedure sui dati, ma anche, presumibilmente, peggiore era la qualità di partenza dei dati pervenuti.

La mancata risposta deve sempre essere sostituita con un valore imputato se la variabile in esame è *obbligatoria* (Cfr. par. 1.13), cioè è relativa ad un quesito a cui devono rispondere tutti, non dipendendo da alcuna variabile filtro che definisce un "salto" nel questionario e quindi non essendo specifica per particolari sottogruppi di popolazione.

Al contrario le variabili *non obbligatorie*, che dipendono da uno o più filtri, o sono relative a sottoinsiemi di popolazione, ammettono anche il valore *blank* fra le risposte esatte.

Nel calcolare gli indicatori di qualità sul numero di modifiche si può tenere conto di questa diversa prospettiva. Per ciascuna variabile (obbligatoria e non) si possono quindi analizzare i tassi lordi (ad esempio di risposte mancanti) sul totale dei rispondenti, mentre (solo) nel caso delle variabili non obbligatorie si possono analizzare anche i tassi netti, calcolati sul totale della risposte effettivamente dovute per la specifica sotto popolazione di riferimento.

Nelle tabelle dell'archivio di qualità sono contenuti, variabile per variabile, i totali necessari al calcolo dei tassi di correzione, quindi il numero di modifiche (per tipo: *missing*, fuori *range*, valori ammissibili e non) ed il totale dei rispondenti potenziali.

I nomi delle tabelle dell'archivio fanno riferimento alle diverse unità di analisi definite nei modelli. Quindi le tabelle relative alla popolazione residente in famiglia contengono nel nome la sigla "pad_alg" (come Popolazione Abitualmente Dimorante in Alloggio), quelle per i residenti in convivenza la sigla "pad_conv", quelle per i non residenti in famiglia la sigla "nad_alg" (Non Abitualmente Dimoranti in Alloggio) e per i non residenti in convivenza "nad_conv". Per gli edifici nel nome delle tabelle compare la sigla "edi", mentre per le abitazioni il nome delle tabella distingue se si tratta di abitazioni occupate "abiocc" o non occupate "abinocc". Nel nome della tabella è incluso anche il nome della variabile, assegnato nelle procedure informatiche, a cui si riferiscono i conteggi.

I dati di base (somme) per il calcolo dei tassi di correzione sono stati calcolati per tutte le variabili al livello territoriale più fine, cioè per comune, in modo che sia possibile da questo arrivare ad aggregazioni superiori: per provincia, per regione o per dimensione demografica.

Si utilizzano le seguenti notazioni per denominare le colonne nelle tabelle dell'archivio:

CODPRO = codice di provincia

CODCOM = codice di comune

TOTCOM = totale comunale di quella unità di analisi

V0 = somma dei dati modificati da valore a *blank*
B2 = somma dei dati modificati da *blank* a valore
V2 = somma dei dati modificati da valore ad altro valore
F2 = somma dei dati modificati da fuori *range* a valore
F0 = somma dei dati modificati da fuori *range* a *blank*
V1 = somma dei valori validi non modificati
B1 = somma dei *blank* validi non modificati
FILTRO1 = somma delle risposte non dovute (solo per variabili non obbligatorie)
FILTRO2 = somma delle risposte dovute (coincide con TOTCOM per le variabili obbligatorie)

Ad esempio una tabella relativa alla variabile obbligatoria "sesso" contiene tante righe quanti sono i comuni ed è strutturata nelle seguenti colonne: CODPRO, CODCOM, TOTCOM, B2, F2, V2, V1, seguita da eventuali percentuali calcolate a livello comunale.

La variabile "stato estero di cittadinanza" invece non è obbligatoria (riguarda i solo cittadini stranieri) e quindi la tabella dell'archivio, che contiene i relativi conteggi delle modifiche introdotte di correzione a livello comunale, è definita a partire dalle colonne: CODPRO, CODCOM, TOTCOM, B2, F2, V2, V1, FILTRO2, F0, V0, B1, FILTRO1.

I tassi principali sugli effetti delle correzioni calcolabili per le *variabili obbligatorie* sul totale della popolazione (comunale) sono:

V1/ TOTCOM * 100 = percentuale di risposte esatte all'origine
B2 / TOTCOM * 100= percentuale di risposte mancanti
(V2+F2)/ TOTCOM *100 = percentuale di risposte incompatibili
(B2+V2+F2)/ TOTCOM *100 = percentuale di modifiche.

Per le *variabili non obbligatorie*:

(V1+B1)/ TOTCOM * 100 = percentuale di risposte esatte all'origine, inclusi *blank* validi
B2 / TOTCOM * 100 = percentuale di risposte mancanti
(V2+F2+ V0+F0)/ TOTCOM *100 = percentuale di risposte incompatibili
(B2+V2+F2+V0+F0)/ TOTCOM *100 = percentuale di modifiche

e per queste si possono anche calcolare i tassi netti, sulle risposte dovute e non dovute

B2/ FILTRO2 * 100 = percentuale di risposte mancanti su risposte dovute
(V2+F2)/ FILTRO2 * 100 = percentuale di risposte incompatibili imputate su risposte dovute
(V0+F0)/ FILTRO1 * 100 = percentuale di risposte eliminate su risposte non dovute

Per le variabili obbligatorie quindi nel primo caso il tasso di variazione dovuto al processo di validazione viene calcolato ponendo al denominatore il totale di popolazione, mentre nel secondo si fa riferimento al totale della popolazione che "deve" rispondere (come risulta dai dati puliti), tenendo conto del filtro esatto.

L'archivio contiene anche la tabelle riassuntive (il cui nome è definito dal prefisso *disso_nome dell'unità di analisi)* degli indicatori di dissomiglianza, calcolati sui risultati delle correzioni, che mettono in evidenza gli scostamenti fra le distribuzione di frequenza grezza e la distribuzione di frequenza corretta, come già esplicitato nel paragrafo relativo alla dissomiglianza.

Per quanto riguarda le fasi di *editing,* per ogni fase di controllo e correzione, considerando le regole di compatibilità indicate dai ricercatori sulla base della loro esperienza, è stato calcolato, sui dati grezzi, il numero di attivazioni delle regole. Nell'archivio sono memorizzati i risultati di questi conteggi, che hanno permesso l'elaborazione delle tabelle relative all'*editing*. La legenda, che spiega il significato dei singoli errori (indicati con un numero progressivo nelle tabelle di conteggio) e che descrive le regole di compatibilità è presente nell'archivio per ogni unità di analisi.

Capitolo 2 - L'indagine sul grado di copertura del 14° Censimento della popolazione

2.1 - Introduzione

Il 14° Censimento generale della popolazione[1] ha avuto, tra i suoi obiettivi principali, quello di censire (o, con lo stesso significato, enumerare) la popolazione residente o presente sul territorio italiano alla data di riferimento del 21 ottobre 2001. Di ciascuno degli individui residenti è stato rilevato anche il legame di parentela con gli altri componenti della famiglia, in modo da ricostruire i nuclei familiari e conteggiare il numero di famiglie. Il livello territoriale più fine rispetto al quale sono riferiti i dati raccolti è costituito dalle sezioni di censimento.

In questo capitolo si descrive l'indagine di controllo predisposta per stimare l'incidenza degli errori di copertura, una particolare tipologia di errori possibili durante le operazioni di censimento, la cui presenza conduce a errori nel conteggio della popolazione residente. L'indagine prende il nome di Indagine sul grado di copertura del 14° Censimento generale della popolazione (Idc) ed è stata condotta su un campione di sezioni di censimento dopo il termine delle operazioni censuarie.

In questo capitolo si illustrano le principali caratteristiche dell'indagine e i suoi risultati più importanti, rimandando ai paragrafi successivi per l'approfondimento dei singoli argomenti di interesse. In questo modo si cerca di fornire al lettore un quadro di insieme, lasciandolo libero di approfondire successivamente gli argomenti che più lo interessano.

La popolazione obiettivo dell'Idc è costituita dagli individui e dalle famiglie residenti in abitazioni collocate nel territorio nazionale; la principale stima prodotta è rappresentata dall'errore di sottocopertura, cioè dalla distorsione per difetto del reale ammontare dei conteggi di popolazione a causa della pratica impossibilità di enumerare tutti gli individui. La stima dell'errore di sottocopertura è ottenuta tramite il modello statistico noto come cattura-ricattura[2] o anche con il nome di *dual-system*, quando applicato a popolazioni umane. Il metodo si avvale di più occasioni di conteggio, tutte affette da possibile sottonotifica, al fine di stimare il reale ammontare di una popolazione. Nella forma utilizzata in questa sede le occasioni di cattura sono due (Censimento e indagine post-censuaria), delle quali la seconda ha carattere campionario.[3] Il metodo cattura-ricattura è soggetto a una serie di ipotesi le quali, se non verificate, possono portare a distorsioni nelle stime. Fra queste, rivestono particolare importanza ai fini dell'indagine, l'indipendenza statistica tra le due occasioni di rilevazione, l'omogeneità della probabilità di cattura tra tutti gli individui rilevati in una stessa occasione e la capacità di identificare senza errore quante volte un individuo è stato "catturato". Se rispettate, le ipotesi sottostanti il modello permettono di stimare l'ammontare di popolazione come prodotto dei totali degli individui conteggiati in ciascuna delle due occasioni, rapportato al totale degli individui che risultano conteggiati in entrambe le occasioni.

L'errore di sovracopertura, di segno opposto a quello di sottocopertura, è invece solitamente considerato di minore entità a causa delle modalità con le quali si svolge il Censimento italiano, e per questo motivo ne è stata prodotta una stima sperimentale per la prima volta in questa occasione di indagine. Maggiori dettagli sulle caratteristiche della popolazione obiettivo e sulle tipologie di stima prodotte dall'indagine sono esposti nel paragrafo 2.2 del presente capitolo.

Il presente capitolo è stato redatto da Giancarlo Carbonetti (parr. 2.3.4, 2.6.4, 2.6.5, 2.10.1), Claudia Cianfarani (parr. 2.7.1, 2.7.3, 2.7.5, 2.7.6), Nicoletta Cibella (parr. 2.9, 2.9.1, 2.9.2, 2.9.3), Fabrizio Delli Priscoli (parr. 2.7.2, 2.7.4, 2.7.7), Loredana Di Consiglio (parr. 2.3.2, 2.3.3, 2.10.5), Stefano Falorsi (par. 2.3.1), Epifania Fiorello (par. 2.9.4), Marco Fortini (parr.2.1, 2.2, 2.10.2, 2.10.3, 2.10.4, 2.11, 2.13), Patrizia Leonardi (par. 2.12), Nadia Mignolli (parr. 2.4, 2.5, 2.6.1, 2.6.2, 2.6.3), Alessandra Nuccitelli (par. 2.8). Hanno inoltre collaborato alle fasi operative dell'indagine di copertura, con particolare riferimento all'esecuzione delle procedure di abbinamento tra i dati individuali dell'Idc e quelli del censimento: Maria Rita Lisandrelli, Silvia Melfi e Laura Monacelli.

[1] Istat (2005); *14° Censimento della popolazione e delle abitazioni 2001, struttura demografica e familiare della popolazione residente – Italia,* Collana Censimenti, Istat, Roma.

[2] Sekar, C. Chandra, e W. Edwards Deming. "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association,* 44, n. 245, (1949): 101-115.

[3] Wolter K. M., (1986); **"**Some Coverage Error Models for Census Data", *Journal of the American Statistical Association,* Vol. 81, No. 394 , pp. 338-346.

L'Idc adotta un disegno di campionamento di tipo areale a due stadi, in cui il primo stadio è costituito dai comuni, stratificati in base alla ripartizione geografica di appartenenza (a cinque categorie) e a una classificazione a quattro categorie di dimensione demografica.[4] Le unità di secondo stadio sono invece costituite dalle sezioni di censimento e su esse è stata applicata una stratificazione basata sulla tipologia di località di appartenenza[5]. Nel complesso, il campione estratto ha riguardato 98 comuni e 1.102 sezioni di censimento tra le quali quattro sono poi risultate non abitate. Nelle 1.098 sezioni risultate abitate da almeno un individuo sono quindi stati intervistati 179.886 individui appartenenti complessivamente a 68.310 famiglie. Una descrizione dettagliata delle caratteristiche del disegno di campionamento è riportata nel paragrafo 2.3.

La data di riferimento dell'indagine è, analogamente al Censimento, quella del 21 ottobre 2001. Tuttavia, per intervenire sulla stessa popolazione di riferimento, evitando al contempo di interferire con le normali operazioni censuarie, si è scelto di procedere all'Idc in un momento immediatamente successivo alla conclusione del censimento nelle sezioni campione. Anche dal punto di vista della tecnica d'indagine l'Idc si è svolta in modo conforme al censimento, per mezzo dell'enumerazione esaustiva delle famiglie e degli individui nelle sezioni di interesse da parte dei rilevatori.[6] Questi avevano il compito di consegnare il questionario di indagine alle famiglie nelle abitazioni e provvedere poi a ritirarlo a domicilio dopo l'avvenuta compilazione a cura dei rispondenti.

Il questionario è stato progettato in modo da essere il più simile possibile, nel formato e nel testo delle domande, a quello proposto in occasione del Censimento, evitando, però, che potesse in alcun modo essere confuso con questo. Tale risultato è stato ottenuto attraverso l'adozione di modalità grafiche molto simili, affiancate però a una scelta cromatica differente e caratteristica. La confrontabilità tra i questionari adottati nelle due occasioni era del resto particolarmente importante soprattutto perché, per mezzo di ulteriori elaborazioni sui dati dell'Idc, sono state anche effettuate analisi sull'errore di risposta per le principali variabili rilevate sugli individui censiti. Infatti, per ragioni di costo/beneficio, in questa occasione si è preferito non effettuare un'apposita indagine di qualità come invece fu fatto per i censimenti del 1981 e del 1991. In alternativa si è scelto di utilizzare i dati riguardanti gli individui censiti, insieme ai corrispondenti contenuti sui record a essi abbinati in occasione dell'Idc, per valutare la componente di variabilità indotta dall'errore di misurazione. Ai dettagli su tali elaborazioni è dedicato il terzo capitolo del presente volume, mentre nei paragrafi 2.4 e 2.5 di questo capitolo sono descritte rispettivamente la tecnica d'indagine dell'Idc e il questionario adottato, fornendo inoltre informazioni sull'esito delle operazioni sul campo.

L'architettura informatica e le procedure utilizzate per la gestione dei dati sono descritte nel paragrafo 2.7. In questa sede ci si limita a richiamare l'attenzione su questa componente dell'indagine, dato che il processo di produzione adottato per l'Idc si è basato sull'integrazione di due fonti distinte e, per questo, ha richiesto di gestire e far comunicare due flussi di dati complessi e tra loro separati.

Come appena rimarcato, uno dei punti cruciali dell'Idc è costituito dalla fase di abbinamento tra i record individuali raccolti in occasione dell'indagine, con quelli rilevati nelle stesse sezioni durante il Censimento.[7] Tale operazione è indispensabile per applicare il metodo cattura-ricattura che utilizza il numero di individui rilevati in entrambe le occasioni di indagine per stimare l'ammontare della popolazione. La qualità dell'operazione di abbinamento influisce direttamente sull'affidabilità delle stime; infatti per ogni punto percentuale di errore nell'identificare correttamente gli individui rilevati in entrambe le occasioni risulta corrispondere una sottostima di dimensione paragonabile nel grado di copertura. Per questo motivo si è cercato di ridurre virtualmente a zero la possibilità di introdurre errori di abbinamento, applicando metodologie basate sul *linkage* probabilistico e acquisendo, a cura degli Uffici comunali di statistica (Ucs) coinvolti nell'Idc e per le sole sezioni campione, i nominativi degli individui rilevati in ciascuna delle due occasioni, in modo da fruire di chiavi d'aggancio altamente affidabili. Alla descrizione approfondita del complesso processo d'integrazione tra i dati raccolti con l'Indagine di copertura e i quelli censuari è dedicato il paragrafo 2.8.

---

[4] Di Consiglio L., Falorsi S. (2003); *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento generale della popolazione 2001*, Collana Documenti (http://www.istat.it/dati/pubbsci/documenti/), Istat, Roma.

[5] In base alle definizioni del Censimento le località sono definite come raggruppamenti di sezioni di censimento e classificate come di centro, di nucleo e di case sparse, in funzione della loro collocazione all'interno del comune di appartenenza e della loro dotazione in termini di servizi essenziali.

[6] Per assicurare l'indipendenza tre le due occasioni di rilevazione, i rilevatori impiegati dall'Indagine di copertura, scelti tra quelli più esperti, sono stati inviati in sezioni differenti da quelle che avevano enumerato per il censimento.

[7] Nuccitelli A. (2005). "La strategia di abbinamento dei dati del 14° Censimento della popolazione con i dati dell'indagine di copertura". In P. D. Falorsi, A. Pallara, A. Russo (a cura di). *L'integrazione di dati di fonti diverse - Tecniche e applicazioni del record linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*, Franco Angeli, Milano: 61-91.

Terminate le elaborazioni relative alla fase di abbinamento sono stati necessari ulteriori aggiustamenti nei dati, in modo tale da assicurare che tutti i record individuali fossero classificabili nelle modalità rispetto alle quali i tassi di copertura sono prodotti. Ad esempio, per poter produrre correttamente i tassi di copertura riferiti al genere maschile e a quello femminile è opportuno che per tutti gli individui sia attribuibile la variabile sesso. Per ottenere questo risultato i dati da analizzare sono stati preventivamente sottoposti a una procedura di revisione e imputazione tale da attribuire un valore plausibile ai record che presentavano valori mancanti o non ammissibili. L'ammontare dei record sottoposti a correzione è stato comunque di modeste dimensioni se rapportato al totale delle unità eleggibili, essendo risultato inferiore all'1 per cento per le principali variabili (anno di nascita, sesso, stato civile e posizione nella professione) e comunque inferiore al 5 per cento per le altre variabili analizzate (titolo di studio 1,7 per cento; cittadinanza 4 per cento; condizione professionale 4,3 per cento). Tale quantità è certamente trascurabile rispetto alla dimensione dei dati trattati e non tale da produrre modifiche apprezzabili alle stime del grado di copertura. Il paragrafo 2.9 è dedicato alla descrizione di questa attività.

Lo stimatore del tasso di copertura deve tenere conto delle caratteristiche del modello statistico utilizzato, del disegno di campionamento adottato e della possibilità di usufruire di informazioni ausiliarie tali da aumentare l'accuratezza delle stime. In particolare è stata adottata una post-stratificazione sulla popolazione, basata sul sesso e su una classificazione a 13 categorie dell'età, per garantire l'ipotesi di omogeneità delle probabilità di cattura degli individui entro i post-strati, richiesta dal modello cattura-ricattura per le due occasioni di rilevazione. Le caratteristiche del disegno di campionamento sono state invece considerate per la predisposizione dei pesi di riporto all'universo. I coefficienti sono stati quindi corretti sfruttando la conoscenza dei totali di popolazione censita in alcuni sottoinsiemi di popolazione. Questo risultato è stato ottenuto applicando un algoritmo di ponderazione vincolata[8] agli individui censiti nelle sezioni campione, perché questi restituissero i totali noti con la minima variazione possibile dei pesi originali. I dettagli riguardanti le caratteristiche degli stimatori utilizzati sono contenuti nel paragrafo 2.10, insieme all'esposizione delle modalità di calcolo della variabilità delle stime prodotte.

La fase di analisi dei dati ha riguardato 173.109 individui Idc, selezionati scartando gli individui che alla data di censimento non erano ancora nati o non erano residenti all'interno delle sezioni nelle quali era effettuata l'intervista. L'abbinamento con i 182.519 censiti nelle sezioni campione ha quindi permesso di determinare in 169.980 il numero degli individui rilevati in entrambe le occasioni.

I risultati evidenziano un tasso di copertura per gli individui del 98,55 per cento a livello Italia, con un andamento abbastanza uniforme rispetto alle ripartizioni geografiche e, invece, marcatamente decrescente all'aumentare della dimensione comunale, dove la copertura risulta sensibilmente più bassa in corrispondenza dei comuni urbani (95,89 per cento). Per quanto riguarda le caratteristiche degli individui, mentre non si notano differenze sostanziali nel tasso di copertura in relazione al sesso, è da segnalare una forte associazione con l'età. Si nota infatti che i valori minimi della copertura si collocano nella fascia d'età tra 0 e 5 anni (97,92 per cento) e in quella tra 20 e 29 anni (97,82 per cento). Il tasso, invece, cresce uniformemente dopo i 30 anni per giungere al suo apice nella fascia d'età 75-84 (99,23 per cento). In effetti il valore riscontrato per la copertura nella fascia d'età più bassa è quello più preoccupante in quanto sembra rivelare un particolare problema di sottonotifica per la fascia di popolazione più giovane. Inoltre tale dato si mostra costante, con la sola eccezione dell'Italia insulare, sia in relazione al territorio sia alla dimensione comunale. Al contrario, la depressione osservata nel tasso di copertura per la classe d'età 20-29, sebbene non desiderabile, è più attesa e già verificata anche in un contesto internazionale.

Rispetto alle famiglie, definendo censita una famiglia inclusa nel campione dell'Indagine sul grado di copertura,[9] se almeno un suo componente è risultato censito, il grado di copertura risulta sensibilmente più basso per le famiglie monocomponente (96,91 per cento) rispetto a quelle di due o più componenti, per le quali si va da un minimo di 98,66 per cento (classe otto o più componenti) a un massimo di 99,52 per cento (classe sei o sette componenti). Questo risultato è atteso proprio per la definizione adottata di famiglia censita, visto che una famiglia di un solo componente ha molte più probabilità di non essere censita di quelle più numerose. Il dato segnala comunque un problema nell'identificazione dei nuclei familiari piccoli, che, di contro, sono in costante

[8] Deville, J. C., e C.-E. Särndal. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association*, 87, 418 (1992): 376-382.

[9] Nel seguito le famiglie e gli individui che fanno parte del campione dell'Indagine sul grado di copertura verranno indicati come "famiglie Idc" e "individui Idc", quelle riferite al Censimento della popolazione come "famiglie Cen" e "individui Cen".

aumento. D'altro canto, anche analizzando il tasso di copertura individuale in funzione del numero di componenti della propria famiglia di appartenenza, si osserva un valore del tasso di copertura sensibilmente inferiore per gli individui appartenenti alle famiglie monocomponente, che cresce fino alla classe degli individui appartenenti a famiglie di quattro o cinque componenti e torna poi a diminuire. Anche questo fenomeno testimonia l'importanza della dimensione del nucleo familiare nella probabilità di essere censiti.

Dai risultati esposti in questa introduzione si nota che i guadagni maggiori nel tasso di copertura potrebbero essere ottenuti investendo soprattutto nei grandi centri urbani e ponendo in atto comportamenti di rilevazione particolarmente dedicati all'identificazione dei bambini nella classe 0-5 anni, dei giovani nella fascia d'età 20-29 e, in generale, degli individui che vivono da soli.

I risultati che emergono dall'Idc si mostrano in linea con quanto riscontrato nei maggiori paesi. Infatti, anche se un confronto diretto delle varie statistiche non è agevole a causa delle diverse modalità con le quali sono condotti i censimenti nelle varie nazioni, il tasso di copertura riscontrato in Italia è certamente paragonabile a quello conseguito in altri grandi paesi quali il Regno Unito (98,00 per cento), gli Stati Uniti (98,82 per cento) o l'Australia (98,90 per cento). Anche i confronti con i passati censimenti italiani confermano il senso di una evoluzione della qualità. Infatti, nel 1981[10] il dato del tasso di copertura si attestava sul 96,45 per cento mentre il dato ufficiale del 1991[11] (99,30 per cento) più elevato di quello attuale, è comunque da considerare come un limite superiore in quanto si fonda sulla semplice risposta degli intervistati alla richiesta di specificare se erano stati censiti o meno e non su un rigoroso *record linkage* con i dati censuari. È infatti noto che questa metodica tende a sovrastimare il tasso di copertura, dal momento che gli intervistati sono inclini a non segnalare l'eventuale mancato censimento per timore di incorrere in sanzioni di qualche tipo o, comunque, di perdere del tempo. A questo proposito si cita un'analisi effettuata per ovviare a questo problema,[12] condotta attraverso l'impiego aggiuntivo di informazioni legate al *linkage*, eseguito anche nel '91, ma affetto da errori che ne impedirono l'uso a fini diretti di stima. I risultati di questo studio considerarono plausibile il ridimensionamento del tasso copertura del 1991 fino al valore di 98,20 per cento riportandolo su valori del tutto confrontabili con i risultati attuali. Maggiori dettagli nell'analisi dei dati sono fornite nel paragrafo 2.11, mentre dei confronti internazionali e con i passati censimenti italiani sono svolte nel paragrafo 2.12.

2.2 - La popolazione obiettivo, gli errori di copertura e i domini di interesse

Oggetto dell'Idc è la stima degli errori di copertura, cioè di quegli errori commessi nella fase di enumerazione degli individui durante il censimento che possono provocare distorsioni per eccesso o per difetto del conteggio della popolazione. La popolazione di interesse è quella dei residenti. Il residente, secondo la definizione in uso al censimento, è "un individuo enumerato nella sua dimora abituale, dove già possiede o intende spostare nel futuro la sua residenza anagrafica". Tale definizione non richiede che un individuo possieda la residenza anagrafica nell'indirizzo in cui risiede, ma soltanto che tale indirizzo venga da lui identificato come sua dimora abituale.

Non sono invece oggetto di interesse dell'Idc: la popolazione presente, quella senza fissa dimora e quella residente in comunità o istituzioni. Infatti, mentre le prime due sono considerate troppo mutevoli per essere correttamente stimate dall'Idc, la popolazione residente in comunità è relativamente ridotta e generalmente sottoposta a specifici controlli tali da rendere non necessaria l'esecuzione di una indagine post-censuaria.

Gli errori di copertura possono essere distinti in due grandi categorie:

- gli errori che generano sovracopertura, costituiti dall'enumerazione nella popolazione residente di individui che non ne fanno parte (non residenti) e quindi provocano un eccesso nel conteggio della popolazione residente;

[10] Terra Abrami, V., e M. Masselli. "L'indagine di controllo di copertura del censimento della popolazione", In *Atti del Convegno intermedio della Società italiana di statistica: La qualità dei dati statistici*, Trieste: 21-23 aprile 1983.

[11] Abbate, C.C., M. Masselli, e M. Signore. "A combined post-enumeration survey of the 1991 population and industrial census", *Proceedings of ISI*, 2, 16.3 (1993).

[12] Fortini, M. "Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione". In *Atti della XXXVII Riunione Scientifica della Società italiana di statistica*, San Remo: 6-8 aprile 1994.

- gli errori che generano sottocopertura, costituiti dalla mancata enumerazione nella popolazione residente di individui che invece ne fanno parte (residenti) e quindi portano a un difetto nel conteggio della popolazione residente.

Dal momento che il concetto di popolazione residente riguarda il territorio, attraverso il legame esistente fra l'individuo e l'indirizzo presso il quale questo risiede, anche i concetti di sovra e sottocopertura dipendono dalla porzione di territorio che viene considerato ai fini della stima.

Infatti se a livello nazionale l'enumerazione di un individuo a un indirizzo differente da quello dove dimora abitualmente (purché entrambi collocati sul territorio italiano) non provoca problemi di copertura in quanto l'individuo viene comunque conteggiato, quando ci si riferisce a porzioni di territorio (ad esempio ripartizioni geografiche o regioni) anche un errore di localizzazione può causare problemi se la vera dimora abituale si trova in una zona diversa da quella nella quale viene erroneamente collocata. È quindi importante considerare che, quando ci si riferisce a porzioni del territorio nazionale, gli errori di sovracopertura in un determinato luogo possono, anche se non necessariamente, corrispondere a errori di sottocopertura in un altro luogo e viceversa.

La casistica degli errori possibili prevede quindi: duplicazioni, costituite da individui che sono enumerati in più di un indirizzo; erronee enumerazioni, costituite da individui non residenti sul territorio nazionale che però vengono lo stesso censiti; omissioni, costituite da individui residenti che però non sono enumerati in alcun luogo. Mentre i primi due casi provocano sempre errori di sovracopertura, la terza situazione genera sempre un caso di sottocopertura.

Ai fini pratici verranno prese in considerazione le seguenti circostanze:

- individui residenti sul territorio nazionale ma non censiti;
- individui residenti sul territorio nazionale censiti in un luogo diverso da quello di dimora abituale;
- individui residenti sul territorio nazionale enumerati per due volte, delle quali una all'indirizzo corretto;
- individui non residenti sul territorio nazionale enumerati per errore a un indirizzo collocato sul territorio nazionale.

Altri eventi, come quelli di individui non residenti enumerati due o più volte a indirizzi errati e non all'indirizzo corretto possono naturalmente essere considerati. Tuttavia tali casi non verranno presi in esame in quanto considerati rari e comunque molto difficili da individuare mediante un'indagine del tipo previsto.

La misura dell'entità degli errori di copertura per famiglie e individui sarà quindi ottenuta mediante gli stimatori introdotti nel paragrafo 2.10, i quali faranno uso dei risultati del processo di integrazione tra i dati dell'Idc e quelli del Censimento come verrà illustrato nel paragrafo 2.8.

In breve, la metodologia utilizzata per produrre tali risultati si basa sul metodo cattura-ricattura, introdotto per la stima della dimensione ignota di popolazioni di origine animale[13] e successivamente applicato anche alle popolazioni umane,[14] ambito in cui il metodo è anche noto come *dual system*.

L'approccio *dual system* assume che l'ammontare di popolazione residente in una data area possa essere stimato attraverso un primo conteggio esaustivo (cattura) della popolazione in occasione del censimento, una ripetizione della conta su un campione di sezioni di censimento[15] e una successiva integrazione finale tra i due insiemi di dati per calcolare la proporzione di "ricatturati", ovvero di coloro che sono enumerati sia in occasione del censimento, che in occasione della successiva indagine post-censuaria.

Gli errori di copertura possono portare a un conteggio della popolazione residente distorto, per eccesso o per difetto. Se indichiamo con CEN il totale della popolazione residente conteggiato al Censimento e, rispettivamente, con SOV e SOT il numero di casi di sovra e sottocopertura, possiamo risalire al valore reale della popolazione residente (POP) applicando la relazione POP=CEN-SOV+SOT dove la quantità SOV-SOT viene indicata come errore netto di copertura. La popolazione misurata dal censimento corrisponde al vero valore solo nel caso, improbabile, in cui l'errore netto di copertura sia pari a zero. Tali considerazioni restano valide anche se riferite a porzioni dell'intero territorio, come le regioni, a sottoinsiemi della popolazione (maschi, occupati, eccetera) o a loro intersezioni.

In realtà la stima dell'esatto ammontare di popolazione, allo stato attuale dell'esperienza nel nostro paese, esula dagli obiettivi dell'indagine. Infatti, la dimensione del campione utilizzato e le numerose e delicate ipotesi sottostanti il modello di stima non forniscono adeguate garanzie di accuratezza per le stime di ammontari così

[13] Pollock, K. H., J. D. Nichols, C. Brownie, e J. E. Hines, (1990): *Statistical inference for capture-recapture experiments*, Wildlife Monographs 107.
[14] Sekar C.C., Deming W. E., (1949): "On a Method of Estimating Birth and Death Rates and the Extent of Registration", *Journal of the American Statistical Association*, Vol. 44, No. 245, pp. 101-115.
[15] Wolter K. M., (1986): **"**Some Coverage Error Models for Census Data".

importanti, non solo dal punto di vista statistico, ma anche da quelli politico, sociale ed economico. Per questo motivo si preferisce fornire stime per i soli rapporti

$$1-\frac{SOT}{POP} \quad \text{e} \quad \frac{SOV}{POP}$$

i quali, pur possedendo le medesime proprietà di accuratezza degli ammontari, subiscono di meno le implicazioni esterne all'interpretazione puramente statistica dei dati.

I domini rispetto ai quali si vogliono produrre le informazioni sono diversi. È certo importante riferire la copertura a porzioni del territorio nazionale oltre che al suo complesso. Per questo motivo i risultati dell'Indagine di copertura saranno riferiti anche alle cinque ripartizioni geografiche. Altra dimensione importante da considerare ai fini dello studio della copertura sul territorio è legata alla dimensione dei comuni italiani. Come infatti riscontrato nell'Indagine di copertura del 1991,[16] la minore dimensione demografica del comune sembra essere associata positivamente all'entità degli errori di sottocopertura, probabilmente per la difficoltà di controllare il territorio e le operazioni di censimento nei comuni più grandi. Inoltre, per valutare come gli errori di copertura siano collegati alla densità di popolazione sul territorio, il tasso di copertura sarà riferito alle tipologie di sezioni classificate dal Censimento in "nuclei", "centri abitati" e "case sparse", a seconda della densità abitativa attesa nel loro territorio.

Altre caratteristiche interessanti per l'analisi della copertura sono quelle che descrivono la popolazione di individui e di famiglie. A tale scopo si citano la dimensione della famiglia, l'età e il sesso degli individui, il loro stato civile, la loro occupazione o il loro grado di istruzione. È appena il caso di osservare che, poiché la copertura (netta o relativa) sarà stimata su un campione areale costituito da sezioni di censimento, l'efficienza delle stime relativamente alle categorie considerate su ciascuno dei domini citati, nonché su loro eventuali combinazioni, dipende dai domini programmati nel disegno e dalla massima dimensione sostenibile del campione considerato. Maggiori dettagli sui domini pianificati nel disegno campionario saranno forniti nel prossimo paragrafo.

2.3 - Il disegno di campionamento

2.3.1 - Caratteristiche generali

L'enumerazione censuaria fa riferimento al territorio che, per esigenze organizzative, viene ripartito in aree aventi determinate caratteristiche (dette sezioni di censimento). L'Indagine di copertura, volta a valutare il conteggio effettuato dal Censimento, ripete le operazioni di enumerazione su un campione casuale di dette sezioni che costituiscono, pertanto, le unità finali di campionamento.

Il disegno di campionamento adottato è stato di tipo composito, essendo costituito dall'unione di un disegno a uno stadio stratificato nei comuni più grandi inclusi con certezza nel campione (parte autorappresentativa), in cui si sono estratte direttamente le sezioni di censimento, e da un disegno a due stadi stratificato (parte non autorappresentativa) in cui al primo stadio si seleziona un campione di comuni e al secondo stadio – nell'ambito di ciascun comune selezionato al primo stadio – si è selezionato un campione di sezioni.

In entrambi i casi tutte le famiglie appartenenti alle sezioni campione sono state enumerate.

La scelta di selezionare i comuni al primo stadio è stata determinata dalla necessità di tenere sotto controllo, per esigenze organizzative e di costo, il numero di comuni coinvolti nella rilevazione; non si è ritenuta praticabile la scelta alternativa di selezionare direttamente un campione di sezioni censuarie perché ciò avrebbe comportato il coinvolgimento di un numero troppo elevato di comuni campione.

Il disegno utilizzato ha inoltre previsto: la stratificazione delle unità di primo stadio; la stratificazione delle unità di secondo stadio; la selezione dei comuni campione senza reimmissione e con probabilità proporzionali all'ampiezza demografica; la selezione delle sezioni campione con probabilità uguali e senza reimmissione.

Per ragioni di tipo operativo e di costo è stato stabilito che la numerosità campionaria in termini di comuni fosse pari a circa 100 e che la numerosità di famiglie fosse in media circa 65 mila (ricordiamo che la numerosità delle famiglie nell'indagine è una variabile aleatoria la cui media dipende dalla tipologia delle sezioni); le

[16] Abbate C., Masselli M., Signore M. (1993): "A combined post-enumeration survey of the 1991 population and industrial censuses".

suddette numerosità, simili a quelle della precedente Indagine di copertura del 1991, da uno studio[17] del 2003 garantivano stime attendibili.

2.3.2 - La scelta delle variabili di stratificazione

I comuni sono stati ripartiti in 20 strati $T_1,\ldots,T_h,\ldots,T_H$, (H=20), costituiti dalle classi formate dall'intersezione delle modalità delle due variabili, ripartizione geografica e classe di dimensione demografica, che costituiscono i principali domini di interesse.

Si ricorda che le classi di dimensione demografica sono state definite come:

- comuni con meno di 10 mila abitanti;
- comuni tra 10 mila e 100 mila abitanti;
- comuni con oltre i 100 mila abitanti esclusi i comuni metropolitani;
- 13 grandi comuni (i comuni metropolitani più *Messina*): Torino, Genova, Milano, Venezia, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari.

Assegnando adeguate numerosità campionarie negli strati così definiti, è stato possibile garantire stime aventi prefissati livelli di precisione attesi per ciascuno dei principali domini di interesse.

Per la definizione di ulteriori criteri di stratificazione, sia dei comuni sia delle sezioni in ciascun dominio di stima T_h (h=1,…,20), è stata condotta un'analisi volta a valutare la relazione tra il tasso di copertura (desunto dall'Indagine di copertura del Censimento della popolazione 1991) e alcune variabili ottenute, sia dal Censimento 1991 che dalle indagini Istat, sui dati di qualità delle anagrafi comunali dell'anno 1998.[18]

Le variabili considerate nello studio sono state:

- la collocazione geografica della sezione;
- la dimensione demografica del comune;
- un indicatore di vicinanza del comune alle soglie critiche di popolazione;
- un indicatore della qualità del lavoro del comune in relazione alle rilevazioni statistiche in esso condotte;
- la tipologia della sezione censuaria (centro, nucleo e case sparse).

Dall'analisi dei risultati ottenuti è emerso che le variabili maggiormente utili alla stratificazione sono la dimensione demografica del comune e la tipologia di sezione (sezioni di centro; sezioni di nucleo; sezioni di case sparse), inoltre risulta essere significativa anche la considerazione congiunta della ripartizione con la classe di popolazione e della ripartizione con il tipo di sezione.

In conclusione, poiché le celle ottenute dall'intersezione delle modalità delle variabili ripartizione e classi di popolazione sono già incluse tra le variabili da utilizzare per la stratificazione, in quanto definiscono gli strati che rappresentano i domini principali di interesse per le stime dell'indagine, si è considerata solo la variabile tipologia di sezione come ulteriore criterio di stratificazione delle sezioni all'interno di ciascuno dominio di stima T_h (h=1,2,…,20). È inoltre emerso che la classificazione della popolazione nelle quattro classi di popolazione prestabilite non esaurisce l'apporto della variabile ampiezza della popolazione, ma è sembrato utile considerare una stratificazione più fine in termini di dimensione demografica dei comuni all'interno di ciascun dominio.

2.3.3 - I criteri per la definizione del campione di primo stadio

Le principali caratteristiche del disegno campionario di primo stadio, all'interno di ciascun dominio di stima T_h (h=1,…,20), sono state la stratificazione dei comuni in funzione della dimensione demografica degli stessi e l'autoponderazione del campione.

Sulla base di questi principi, la stratificazione dei comuni è stata effettuata raggruppando comuni di ampiezza omogenea in modo tale che l'ampiezza complessiva dello strato sia approssimativamente costante in termini di popolazione residente.

Il processo di stratificazione delle unità primarie comporta anche l'identificazione di un sottoinsieme di unità primarie la cui dimensione demografica è superiore a una prefissata soglia di popolazione. Tali unità sono

[17] Di Consiglio, L., e S. Falorsi. *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento generale della popolazione 2001*. Roma: Istat, 2003. (Documenti, n.11). http://www.istat.it/dati/pubbsci/documenti/Doc_anno2003.htm

[18] Op. cit., pag. 5.

dette *Auto Rappresentative* (AR) in quanto costituiscono strato a sé stante e sono inserite con certezza nel campione.

Nel sottoinsieme di unità AR sono inclusi, inoltre, anche tutti i comuni metropolitani con popolazione inferiore alla soglia demografica per garantire comunque la possibilità di effettuare stime dirette per questi comuni.

Le restanti unità primarie, dette *Non Auto Rappresentative* (NAR), sono suddivise in strati approssimativamente di uguale ampiezza in termini di popolazione totale.

Le informazioni disponibili per la stratificazione dei comuni e per la selezione dei comuni campione con probabilità variabili sono relative alla popolazione residente dei comuni al gennaio 1999.

Inoltre, poiché il campione di primo stadio è stato definito e selezionato prima che le operazioni di formazione delle nuove sezioni censuarie fossero terminate, al momento della progettazione ed estrazione non erano disponibili alcune informazioni rilevanti per la definizione del campione. In particolare, quindi, era incognito il numero di nuove sezioni appartenenti a ciascun comune e l'ampiezza media effettiva delle nuove sezioni in termini di individui e famiglie.

La mancanza di informazioni relative al numero di sezioni per ciascun comune e all'ampiezza media di queste ha portato a definire il campione di primo stadio - in termini di stratificazione dei comuni e selezione dei comuni campione - sulla base dell'ipotesi semplificatrice di dover selezionare al secondo stadio un campione casuale semplice di famiglie, anziché un campione casuale di sezioni, in termini di numerosità attese previste di famiglie.

Ciò nonostante il disegno effettivo prevede, al secondo stadio, la selezione di un campione di sezioni all'interno delle quali la numerosità di famiglie da enumerare è casuale.

In termini operativi, fissate la numerosità complessiva attesa di famiglie campione, *n*, e la numerosità attesa di famiglie per ciascun dominio di stima, n_h, in base agli errori campionari attesi della stima del tasso di copertura, si definisce il numero minimo pianificato di famiglie, $\overline{n}_h$, da intervistare in ciascun comune campione del dominio T_h.

Tale definizione dipende sia dalla scelta delle quantità *n* ed n_h sia dal numero complessivo di comuni che si vuole far partecipare all'indagine.

Inoltre, in considerazione del fatto che il disegno di secondo stadio è a grappoli, ossia contempla la selezione di un campione di sezioni e l'enumerazione di tutte le famiglie appartenenti alle sezioni estratte, il numero minimo di famiglie per comune campione prefissato, $\overline{n}_h$, deve essere necessariamente maggiore o uguale al numero medio previsto di famiglie residenti per sezione, $\overline{N}_h^*$. Sulla base della numerosità complessiva e tenendo conto del fatto che per ragioni operative e di costo il numero complessivo di comuni campione non deve essere superiore a 100, si è posto il numero minimo di famiglie per comune, $\overline{n}_h$, pari a 600.

Sulla base dei vincoli sulle numerosità, si determina il valore della soglia, G_h, che definisce le unità AR e NAR e la dimensione degli strati di unità primarie mediante la relazione $G_h = \overline{n}_h \overline{\overline{Q}}_h \left(N_h / n_h \right)$, con N_h il numero totale di famiglie e $\overline{\overline{Q}}_h$ la numerosità media della famiglia all'interno del dominio.

In altri termini, la soglia è data dal rapporto tra il numero minimo atteso di individui campione e la frazione di campionamento, $\left(n_h / N_h \right)$, costante all'interno del dominio.

Definita la soglia, i comuni sono distinti in AR, se di dimensione superiore o uguale al valore G_h (o comunque se comuni metropolitani), e in comuni NAR, se di dimensione inferiore a detta soglia di popolazione.

La stratificazione effettiva dei comuni viene, quindi, realizzata ordinando i comuni di ciascun dominio secondo la dimensione (demografica) decrescente e suddividendo i comuni NAR in strati di dimensione approssimativamente uguale al prodotto $2 \times G_h$.

Infine, il campione di unità di primo stadio è realizzato selezionando due comuni campione da ciascuno degli strati NAR con probabilità di selezione espresse dal rapporto tra gli individui residenti nel comune *c* e il numero complessivo di individui dello strato *l* pari a $z_{hlc} = Q_{hlc} / Q_{hl}$.

Per ognuno dei comuni selezionati il numero atteso di famiglie campione è definito mediante la relazione $n_{hlc} = \frac{1}{2}\frac{n_h}{N_h} N_{hl}$ per gli strati NAR e $n_{hlc} = \frac{n_h}{N_h} N_{hl}$ per gli strati AR.

2.3.4 - I criteri per la definizione del campione di secondo stadio

In base a quanto emerso nel paragrafo 2.3.2 sul forte legame tra il tasso di copertura e la tipologia di sezione, si è ritenuto utilizzare quest'ultima caratteristica come criterio di stratificazione delle unità di campionamento di secondo stadio.

Con riferimento al generico comune campione *c* dello strato *l* e del dominio T_h, si denota[19] con *g* la tipologia di sezione (1=nucleo, 2=centro, 3=case sparse). Siano quindi, S_{hlcg} il numero di sezioni di tipo *g*, N_{hlcg} il numero di famiglie residenti nelle sezioni di tipo *g*, s_{hlcg} il numero di sezioni campione di tipo *g*. Si considerino, inoltre, relativamente al dominio T_h, le seguenti quantità: $S_{hg} = \sum_{l=1}^{L_h}\sum_{c=1}^{M_{hl}} S_{hlcg}$, il numero di sezioni di tipo *g*; $N_{hg} = \sum_{l=1}^{L_h}\sum_{c=1}^{M_{hl}} N_{hlcg}$, il numero di famiglie residenti in sezioni di tipo *g*; $s_{hg} = \sum_{l=1}^{L_h}\sum_{c=1}^{M_{hl}} s_{hlcg}$, il numero di sezioni campione di tipo *g* (per il dominio T_h si è indicato con L_h il numero di strati presenti e con M_{hl} il numero totale di comuni appartenenti allo strato *l*).

Si indica infine:

f_{hg}, la quota di famiglie dei comuni dello strato *h* che ricadono in sezioni di tipo *g* (*g*=1,2,3), data dalla seguente espressione:

$$f_{hg} = \frac{\sum_{c\in h} N_{hcg}}{\sum_{g=1}^{3}\sum_{c\in h} N_{hcg}} = \frac{N_{hg}}{\sum_{g=1}^{3} N_{hg}} \qquad (2.3.1)$$

$n_{hc} = \sum_{l} n_{hlc}$, il numero atteso di famiglie campione per il comune *c;*

n_{hcg}, il numero atteso di famiglie campione per il comune *c* che ricadono in sezioni di tipo *g* e dato da:

$$n_{hcg} = f_{hg}\, n_{hc} \qquad (2.3.2)$$

$\overline{N}_{hcg}$, il numero di famiglie mediamente presenti in una sezione di tipo *g* del comune *c* e del generico strato *h*;

$\bar{s}_{hcg}$, il numero di sezioni di tipo *g* da estrarre dal comune *c* del generico strato *h* e mediamente necessarie per ottenere il numero n_{hcg} di famiglie campione, data da:

$$\bar{s}_{hcg} = \frac{n_{hcg}}{\overline{N}_{hcg}} \qquad (2.3.3)$$

Nell'applicazione pratica la quantità $\bar{s}_{hcg}$ spesso non coincide con un numero intero, quindi, per ovvie ragioni, si considera la quantità s^*_{hcg} pari all'intero più vicino a $\bar{s}_{hcg}$. Il numero s^*_{hcg} rappresenta la dimensione sub-campionaria del campione di secondo stadio, ed esprime il numero di sezioni campione di tipo *g* che

[19] Si fa presente che, al fine di non appesantire la notazione introdotta, si adotta la convenzione secondo la quale laddove un indice è assente la quantità deve essere intesa come marginale rispetto allo stesso indice.

mediamente vanno estratte dal comune campione *c* dello strato *h* per ottenere il numero atteso di famiglie in quel medesimo contesto.

Sia inoltre:

$$n^*_{hcg} = s^*_{hcg}\,\overline{N}_{hcg} \qquad (2.3.4)$$

il numero di famiglie campione che si attende di osservare durante l'indagine nelle sezioni campione di tipo *g* del comune *c*; sommando poi rispetto a *g* si ottiene:

$$n^*_{hc} = \sum_{g=1}^{3} n^*_{hcg} \qquad (2.3.5)$$

il numero totale di famiglie campione atteso per il generico comune campione *c*.

Passando ora al disegno di secondo stadio, si fa presente che la questione è stata affrontata nella pratica come un problema di allocazione delle famiglie campione pianificate per i singoli comuni campione tra le tre tipologie possibili di sezioni.

Al fine di garantire quanto più possibile il numero atteso di unità da estrarre in ciascun comune campione, lo schema di campionamento adottato ha previsto, per ciascuna unità di primo stadio, l'estrazione di un campione di sezioni in modo stratificato per tipologia di sezione e con numerosità tali da riprodurre quanto più possibile, nelle sezioni campione, la distribuzione percentuale delle famiglie presenti nelle tipologie di sezioni del dominio (distribuzione marginale) in cui il comune ricadeva.

L'adozione di questa tecnica ha garantito da un lato la rappresentatività del campione e dall'altro ha permesso di tenere sotto controllo il numero previsto di famiglie campione evitando una esplosione del campione con gravi ripercussioni sui tempi e sui costi.

Si fissi l'attenzione, in termini generici, sul dominio *h* per il quale $(f_{h1}; f_{h2}; f_{h3})$, in base anche alla (2.3.1), rappresenta la distribuzione percentuale marginale delle famiglie dei comuni appartenenti al dominio in questione, tra le tre tipologie di sezioni. Applicando al numero atteso di famiglie campione n_{hc} del generico comune campione *c* del dominio *h*, la sopracitata distribuzione e tramite la (2.3.2) si ottiene la ripartizione $(n_{hc1}; n_{hc2}; n_{hc3})$ delle famiglie campione tra le tre tipologie di sezioni. Infine, note le quantità $(\overline{N}_{hc1}; \overline{N}_{hc2}; \overline{N}_{hc3})$ per il comune in questione, tramite la (2.3.3) si ottiene $(\overline{s}_{hc1}; \overline{s}_{hc2}; \overline{s}_{hc3})$ e di conseguenza $(s^*_{hc1}; s^*_{hc2}; s^*_{hc3})$; queste ultime rappresentano le numerosità sub-campionarie del campione di sezioni da impiegare per il comune *c* per tipologia, e dovranno essere tali che la numerosità calcolata tramite la (2.3.5) non si discosti sensibilmente da n_{hc}.

2.4 - La tecnica di rilevazione

2.4.1. - Premessa

Per quanto concerne l'insieme delle modalità di contatto delle unità statistiche interessate dall'Indagine di copertura, la tecnica di indagine è consistita in una vera e propria ripetizione delle operazioni di censimento all'interno delle sezioni campione. Di conseguenza, in modo sintetico, si è trattato di una enumerazione di edifici, abitazioni e famiglie con la consegna di un questionario da parte dei rilevatori da completarsi attraverso autocompilazione; tutte le operazioni necessarie hanno seguito una regolamentazione e una sequenza ben precise e sono state coordinate dall'Istat insieme ai responsabili presso i diversi Uffici di censimento, con l'obiettivo primario di assicurare una completa indipendenza tra l'Idc e il censimento stesso e, in tal modo, di orientare i risultati verso la più elevata qualità possibile.

2.4.2 - Il ruolo svolto dagli Uffici di censimento comunali

I responsabili degli Uffici di censimento dei 98 Comuni (Ucc) interessati dall'Indagine di copertura, anche con funzioni di controllo e attraverso il coordinamento centralizzato da parte dell'Istituto nazionale di statistica, hanno svolto le seguenti operazioni:

- accertamento del rispetto della tempistica;
- presa in carico di tutta la modulistica necessaria alla raccolta delle informazioni;
- selezione dei rilevatori;
- istruzione dei rilevatori;
- rotazione dei rilevatori nelle sezioni di censimento;
- fotocopia e invio all'Istat delle liste dei componenti della famiglia contenute nei Fogli di famiglia raccolti in occasione del censimento;
- controlli di completezza degli itinerari e delle liste provvisorie degli edifici;
- stesura della lista definitiva degli edifici;
- controlli di completezza dei questionari dopo il ritiro;
- telefonate alle famiglie rilevate per controllare a campione l'operato dei rilevatori.

In relazione al rispetto della tempistica, per evitare che eventuali spostamenti della popolazione sul territorio ne potessero modificare in maniera sostanziale la dislocazione rispetto a quella registrata dal censimento, il periodo di riferimento dell'Indagine di copertura è stato fissato molto a ridosso del censimento stesso: in caso contrario, si sarebbero seriamente compromessi i risultati della nuova indagine e, inoltre, si sarebbe acuito l'*effetto memoria* nelle risposte alle domande del questionario, riferite sempre al giorno del censimento (21 ottobre 2001). Di conseguenza, l'arco temporale è stato stabilito tra la metà di novembre e la metà di dicembre 2001, anche se in modo per così dire orientativo poiché, in ogni caso, l'Indagine di copertura non poteva tassativamente iniziare fino a che i rilevatori del censimento non avessero terminato il loro lavoro con il ritiro dei modelli nelle sezioni campione interessate; ciò soprattutto al fine di rendere improbabile qualsiasi sovrapposizione tra le operazioni sul campo delle due rilevazioni. In effetti, questo ha portato inevitabilmente ad alcuni slittamenti, che seppur senza evidenti ricadute sulla qualità, hanno comunque posticipato l'inizio delle operazioni dell'Indagine di copertura alla fine del mese di gennaio 2002.

Per quanto riguarda la modulistica per la raccolta delle informazioni dell'Indagine di copertura, oltre al questionario, di cui si tratterà diffusamente più avanti, sono stati inviati ai Comuni dall'Istat e distribuiti dai responsabili degli Uffici comunali ai rilevatori i seguenti modelli:

- gli itinerari di sezione (Modello Istat CP.5), quando possibile insieme alla cartina planimetrica della sezione, che devono coincidere con quelli già utilizzati per il censimento;
- i quaderni del rilevatore (Modello Istat COPCP.3), che rappresentano il documento comprovante l'individuazione in loco delle unità da rilevare, e rivestono una particolare importanza come strumento di rilevazione e di controllo;
- i modelli per la compilazione della lista provvisoria degli edifici (Modello Istat COPCP.4);
- i modelli per la compilazione della lista definitiva degli edifici (Modello Istat COPCP.4bis);
- i modelli per la registrazione delle abitazioni non occupate (Modello Istat COPCP.2).

Ai responsabili degli Uffici di censimento, inoltre, è spettato il compito di individuare accuratamente i rilevatori per l'Indagine di copertura, attraverso la selezione dei più meritevoli tra quelli che avevano appena svolto le operazioni di censimento.

Successivamente alla nomina dei rilevatori per le nuove operazioni, i responsabili hanno proceduto alla loro formazione utilizzando l'apposito manuale di istruzione fornito dall'Istat, con il costante supporto dei referenti dell'Istituto stesso per le questioni rimaste poco chiare o per dipanare situazioni dubbie.

I responsabili degli Uffici di censimento hanno assegnato ai rilevatori dell'Indagine di copertura una sezione diversa rispetto a quella percorsa durante il censimento, in modo da contribuire all'indipendenza tra le due rilevazioni e garantire la significatività dei risultati della copertura, tenendo sotto controllo il rischio di possibili influenze dei rilevatori sulle risposte delle famiglie al nuovo questionario attraverso il ricorso a informazioni già precedentemente raccolte.

In aggiunta, i responsabili degli Uffici di censimento hanno provveduto alla delicata operazione della fotocopia dei lembi con le liste dei componenti della famiglia, insieme a tutte le relative informazioni (*indirizzo* della famiglia, *nome*, *cognome*, *data di nascita*, *sesso*, *luogo di nascita* di ogni componente), contenuti nei Fogli

di famiglia raccolti in occasione del Censimento nelle sezioni di interesse per l'Indagine di copertura. Tale operazione, di estrema importanza ai fini del buon esito della fase di "abbinamento esatto" (o *record linkage*) dei dati elementari del censimento con quelli dell'Idc, è stata svolta presso gli Uffici di censimento con l'intervento di supporto dei coordinatori provinciali in veste di supervisori. I lembi sopra citati, che di solito vengono trattenuti esclusivamente presso il Comune, sono stati fotocopiati al momento in cui si è proceduto al loro distacco dal resto del Foglio di famiglia e le fotocopie sono state inviate all'Istat per la registrazione insieme al restante materiale relativo all'Indagine di copertura.

In relazione ai controlli di completezza, i responsabili degli Uffici di censimento hanno provveduto a verificare l'esattezza dei confini della sezione di censimento secondo quanto indicato sull'itinerario di sezione (Modello Istat CP.5), eventualmente fornendo tutti i chiarimenti necessari per limitare il più possibile il rischio, da parte dei rilevatori, di non individuare con precisione il territorio di propria competenza e, quindi, di sconfinare nel territorio di altre sezioni, o di non considerare nella rilevazione intere aree, ipotizzando erroneamente una loro appartenenza a una sezione confinante. Si ricorda che lo stadio di aggiornamento di tali itinerari di sezione doveva coincidere fedelmente con l'originale utilizzato durante le operazioni di censimento.

Procedendo nelle operazioni, i responsabili hanno controllato la conformità degli elenchi provvisori degli edifici compilati dai rilevatori e hanno curato la stesura delle corrispettive liste definitive. Immediatamente dopo il ritiro dei questionari dalle famiglie, i responsabili hanno provveduto alla disamina dei quaderni dei rilevatori e al conseguente ulteriore controllo dell'esattezza e della completezza del percorso di competenza anche attraverso il confronto degli indirizzi delle famiglie rilevate; inoltre, hanno proceduto al controllo delle informazioni raccolte nei questionari con il ricorso, quando necessario, a un ulteriore contatto diretto delle famiglie in caso di incongruenza nelle risposte date. A tale proposito, si sottolinea che l'Indagine di copertura prevedeva in ogni caso una fase di telefonate di controllo alle famiglie da parte degli Uffici di censimento, con l'obiettivo di verificare l'impatto del contenuto del questionario, chiarire eventuali problemi ancora aperti, monitorare il lavoro e i tempi impiegati dai rilevatori nelle operazioni di consegna e ritiro dei modelli.

Infine, i responsabili hanno proceduto all'invio all'Istat di tutto il materiale utilizzato in corso di indagine, con l'accortezza di tenere separati i modelli ausiliari (schede di riepilogo, di conteggio complessivo e di controllo) non destinati alla registrazione.

2.4.3 - I rilevatori e le operazioni sul campo

Nel corso dell'indagine sul grado di copertura il ruolo svolto dai rilevatori è stato senz'altro di alta responsabilità, soprattutto se si pensa alla necessaria rigidità della tempistica da osservare e all'importanza del rispetto della sequenza dei diversi interventi di ricognizione sul territorio delle sezioni di competenza, al fine di individuare, affrontare e provare a risolvere gli eventuali problemi in un momento anteriore alla consegna dei modelli presso le famiglie, evitando deleterie interruzioni del lavoro e cercando di raggiungere un elevato grado di completezza dei risultati.

Nello specifico, i rilevatori hanno seguito direttamente le seguenti fasi:

- ricognizione dell'area e verifica dei limiti territoriali delle sezioni di competenza;
- presa di contatto con le realtà esistenti nel territorio della sezione, per annotare casi particolari e prevenire eventuali difficoltà durante le successive fasi della rilevazione;
- rilevazione e enumerazione degli edifici;
- individuazione delle abitazioni occupate;
- individuazione e raccolta delle informazioni sulle abitazioni non occupate all'interno degli edifici;
- consegna dei questionari di indagine a tutte le famiglie e individui della sezione;
- ritiro dei questionari autocompilati; primi controlli sulla completezza e sull'omogeneità dei contenuti del questionario.

La ricognizione delle aree territoriali di competenza dei rilevatori ha rappresentato il punto di partenza dell'Idc ed è avvenuta essenzialmente attraverso l'utilizzo dei relativi itinerari di sezione. In sede di questa verifica nelle sezioni di competenza, i rilevatori non hanno tralasciato di annotare eventuali casi particolari, in modo da prevenire possibili difficoltà e rallentamenti durante le successive fasi di enumerazione di edifici, famiglie e individui. Ad esempio, riscontrando in anticipo una rilevante presenza di cittadini stranieri sul territorio, è stato possibile non trascurare la consegna dei *fac-simile* dei questionari, tradotti in inglese e francese e, in tal modo, agevolare la comprensione e la risposta alle domande..

A seguire, la rilevazione degli edifici[20] è stata la fase nella quale i rilevatori hanno individuato con precisione le unità di rilevazione comprese nel campo di osservazione, valutando cosa includere o escludere e compilato direttamente la lista provvisoria degli edifici (Modello Istat COPCP.4). In effetti, se per il censimento sono stati compilati dei veri e propri modelli per ogni singolo edificio, nell'Indagine di copertura la rilevazione degli edifici ha avuto come obiettivo la sola predisposizione di una lista, dapprima provvisoria, poi definitiva. Sulla lista sono stati indicati tutti gli indirizzi, ponendo la massima attenzione al fine di evitare doppi conteggi in caso, ad esempio, di più ingressi. La corretta identificazione degli edifici, prima dell'enumerazione delle famiglie e degli individui, ha semplificato e migliorato la qualità dell'Indagine di copertura.

I rilevatori hanno proceduto quindi all'individuazione delle abitazioni occupate, o delle altre tipologie di alloggio, e all'annotazione nel quaderno del rilevatore della tipologia e dell'ammontare delle unità di rilevazione presenti presso ogni interno dell'edificio.

Contestualmente, sono state registrate le notizie riguardanti le abitazioni non occupate, compilando l'apposito modello (Modello Istat COPCP.2). Tale compilazione è avvenuta solo dopo aver accertato che l'abitazione fosse realmente disabitata, tornando più volte, e in orari differenti, presso le abitazioni che risultavano non occupate e cercando di reperire informazioni e conferme presso i vicini e i portieri degli stabili.

Le convivenze, e quindi anche gli interi edifici adibiti a convivenza, le persone senza abitazione e le persone senza tetto non dovevano essere considerate nell'ambito dell'Indagine di copertura.

Nella successiva fase di consegna dei questionari alle famiglie (Modello Istat COPCP.1), i rilevatori hanno distribuito un solo modello per famiglia se il numero di componenti non superava i cinque, un numero più elevato di modelli in caso di famiglie più numerose o di diverse famiglie coabitanti nel medesimo alloggio, ponendo la massima attenzione nell'assegnazione dei numeri d'ordine identificativi, per non generare erronee e pericolose duplicazioni. Inoltre, in presenza di famiglie coabitanti è stato necessario individuare e segnalare la Famiglia principale, sulla base di due criteri di predominanza: il più elevato tempo di occupazione dell'abitazione o la maggiore estensione della superficie occupata.

Per ogni questionario consegnato e per ogni modello di abitazione non occupata compilato, i rilevatori hanno provveduto a documentare tutte le operazioni utilizzando l'apposito quaderno (Modello Istat COPCP.3) che ha rappresentato un vero e proprio strumento di rilevazione e controllo sul campo dell'Idc.

Nella fase finale di raccolta dei questionari, i rilevatori hanno provveduto ai controlli di completezza delle informazioni raccolte, ricorrendo a ulteriori contatti, anche telefonici, con le famiglie in caso di risposte dubbie o non omogenee; inoltre, hanno posto attenzione particolare all'operazione di recupero dei modelli consegnati, reperendoli anche presso i nuovi indirizzi, nei casi di trasferimento delle famiglie risultate in precedenza eleggibili ai fini dell'Idc.

Se, da una parte, non è stato necessario consegnare ulteriori modelli a famiglie venute ad abitare successivamente nelle sezioni campione, in quanto non di interesse per l'Idc, dall'altra i rilevatori hanno provveduto a consegnare e ritirare seduta stante i modelli a famiglie eleggibili per l'Idc ma non contattate durante la fase preliminare di consegna.

2.5 - Lo strumento di rilevazione: il questionario dell'indagine

2.5.1 - Gli aspetti innovativi e la struttura del questionario

Per l'Indagine di copertura è stato progettato un questionario,[21] sullo schema di quello utilizzato in occasione del Censimento della popolazione, anche se in forma molto semplificata e snellita. Questo per tenere conto del fatto che l'Idc si poneva il duplice obiettivo di rilevare informazioni per la misura dell'errore di copertura del censimento ma anche le variabili per le analisi sulla qualità delle risposte (Cfr. Capitolo 3).

[20] La definizione di edificio adottata è stata la stessa utilizzata per il Censimento: "una costruzione di regola di concezione ed esecuzione unitaria, dotata di una propria indipendente struttura, contenente spazi utilizzabili stabilmente da persone per usi destinati all'abitazione e/o alla produzione di beni e/o di servizi, con le eventuali relative pertinenze, delimitata da pareti continue, esterne o divisorie, e da coperture, dotata di almeno un accesso dall'esterno".

[21] Per una consultazione del modello di rilevazione utilizzato nell'Idc si rinvia alla seguente pubblicazione: Istat. *I documenti - 14° Censimento generale della popolazione e delle abitazioni*, 2006.

Inoltre, l'esigenza dettata dall'autocompilazione del questionario ha portato a una serie di scelte pensate con una particolare attenzione ai rispondenti. Da qui l'organizzazione del testo nel questionario di rilevazione (il cosiddetto *wording*) e, quindi, la selezione e l'ordine a imbuto in cui sono stati presentati i quesiti, partendo dai concetti più generali fino al dettaglio degli argomenti più strettamente legati alla copertura, la cura particolare che si è voluta dare alla formattazione del modello, con l'uso studiato delle gradazioni di colore, la collocazione dei filtri, il richiamo a chiarimenti e avvertenze. In tal modo, è stato predisposto uno strumento il più possibile autoesplicativo, comprensibile in ogni sua parte, omogeneo, logico nella sequenza delle domande proposte.

Nello specifico, il questionario dell'Idc è stato strutturato nelle seguenti parti:

- un frontespizio a carattere esplicativo, con un prospetto di competenza del rilevatore;
- una pagina di chiusura di competenza del rilevatore;
- la lettera informativa firmata dal Presidente dell'Istat;
- la lista delle persone della famiglia;
- le notizie sulle persone della famiglia, distinte in quattro sezioni.

Si ricorda che il questionario di copertura era rivolto alle famiglie e agli individui con dimora abituale nell'abitazione della sezione campione, o che vivevano temporaneamente nell'alloggio, anche se assenti al momento della rilevazione; gli eventuali ospiti e, in generale, le persone presenti per un breve periodo, per motivi di turismo, vacanza, visita a parenti e amici, non dovevano essere considerati nel questionario.

A tale proposito, è sembrato utile riportare nel frontespizio del modello (Figura 2.1) una descrizione sintetica del significato e dei principali obiettivi dell'Idc e, soprattutto, inserire alcune indicazioni di massima sulle diverse categorie di rispondenti. Secondo quanto stabilito, il modello doveva essere compilato dall'intestatario del questionario, ossia dalla persona di riferimento della famiglia, che in genere coincide con l'intestatario della scheda anagrafica. Negli altri casi il compilatore era un'altra persona con dimora abituale o temporanea nel medesimo alloggio che avesse una buona conoscenza degli altri componenti.

La definizione di famiglia adottata è la stessa utilizzata per il Censimento. Nell'indagine sul grado di copertura è stato stabilito che in caso di famiglie coabitanti i rilevatori consegnassero più questionari, uno per famiglia.

La compilazione del riquadro posto in fondo al frontespizio era a carico dei rilevatori, che hanno provveduto all'inserimento del nome e del cognome dell'intestatario del questionario, dell'indirizzo completo e del numero di telefono, utile per eventuali successivi contatti ai fini di chiarimenti per le incompletezze riscontrate. Nel riquadro A del frontespizio i rilevatori erano tenuti a inserire:

- il codice della sezione;
- il codice provvisorio di edificio;
- il numero d'ordine provvisorio della famiglia, da riportare anche nel relativo quaderno del rilevatore (Modello Istat COPCP.3), vale a dire un numero progressivo assegnato univocamente nell'ambito della sezione in modo da rispecchiare l'ordine di consegna, da facilitare e rendere più ordinate le operazioni di ritiro;
- il codice identificativo del rilevatore.

Figura 2.1 - Frontespizio del questionario dell'indagine sul grado di copertura

Mod. ISTAT **COPCP. 1**

ISTAT

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Rilevazione sul grado di copertura del 14° Censimento generale della Popolazione: Questionario

Questa rilevazione è molto importante per capire se durante il recente censimento tutte le famiglie sono state censite e se le informazioni sono state raccolte in modo corretto. La preghiamo pertanto di leggere con attenzione ogni singola domanda, comprese le note evidenziate o riportate tra parentesi, e di fornire quindi le informazioni richieste.

Nelle risposte per cui è richiesta la scrittura di parole o numeri La preghiamo di scrivere in **stampatello maiuscolo** nel modo più chiaro possibile.

Le informazioni raccolte saranno utilizzate dall'Istat soltanto sotto forma di tabelle di dati aggregati. Esse non potranno essere utilizzate in alcun modo da qualsiasi altro soggetto per effettuare variazioni anagrafiche o qualsiasi altra procedura di tipo amministrativo.

RingraziandoLa per la cortese collaborazione La invitiamo a leggere le informazioni riportate nel seguito che La aiuteranno nella compilazione.

Chi compila il questionario
Il questionario va compilato dall'intestatario del questionario stesso (cioè dalla persona alla quale è intestata la scheda di famiglia in Anagrafe) o, se ciò non è possibile, da un'altra delle persone che vivono abitualmente o temporaneamente nell'alloggio.

Questo questionario deve essere compilato:
- per ogni persona della famiglia che ha dimora abituale nell'alloggio (anche se al momento è assente);
- per ogni persona della famiglia che non ha dimora abituale nell'alloggio, ma ci vive temporaneamente (anche se al momento è assente).

Questo questionario NON deve essere compilato per le persone occasionalmente ospiti dell'alloggio, cioè quelle presenti per un breve periodo (ad esempio per turismo, vacanza o visita a parenti e amici).

È molto importante che i rispondenti compilino il questionario facendo sempre riferimento al 21 ottobre 2001 (data del Censimento).

Se in questo alloggio dimorano più famiglie (famiglie coabitanti), ciascuna di esse deve compilare un diverso questionario.

Per famiglia si intende
Un insieme di persone, legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti ed aventi dimora abituale nello stesso comune (anche se non sono ancora iscritte nell'Anagrafe della popolazione residente del comune medesimo).
Una famiglia può essere costituita anche da una sola persona.

Cosa contiene il questionario
Pagina 2 – Lettera di presentazione del Presidente dell'Istat e riferimenti normativi.
Pagina 3 – Lista delle persone della famiglia che hanno la loro dimora abituale o temporanea nell'alloggio.
Pagine da 4 a 23 – Notizie sulle persone che hanno dimora abituale o temporanea nell'alloggio.

Riservato al Rilevatore e all'Ufficio di censimento comunale

Dati dell'intestatario del questionario *(scrivere in stampatello)*
Cognome
Nome
Indirizzo *(via, piazza, viale, località, ecc.)*
Palazzina Scala Piano Interno
Telefono

A
Sezione di censimento
Edificio (codice provvisorio)
Num.d'ordine provvisorio
Rilevatore

Data della consegna
..... / ..... / .....

Firma del rilevatore

Infine, anche con valore di riconoscimento della propria responsabilità nei confronti dei modelli di competenza, i rilevatori hanno apposto la propria firma e indicato la data di consegna, elemento fondamentale per i controlli di qualità del funzionamento del sistema di monitoraggio (Cfr. par. 2.6).

In relazione alla pagina di chiusura del modello (Figura 2.2), i rilevatori dovevano verificare la presenza e la correttezza dei codici Istat di provincia e comune, originariamente a cura dei responsabili degli Uffici di censimento.

Sempre nell'ultima pagina, nel riquadro B1, i rilevatori ricopiavano le informazioni riportate nel riquadro A del frontespizio. In alcuni casi i responsabili degli Ucc avevano stabilito di recuperare le eventuali famiglie non censite che fossero state individuate nel corso nell'Idc. Per tenere sotto controllo questi casi il codice identificativo (codice a barre) del Foglio di famiglia consegnato doveva essere trascritto nel *riquadro B2*. In questa eventualità la famiglia aveva già risposto al questionario dell'Idc definendosi come non censita a ottobre 2001.

Nel riquadro C i rilevatori dovevano trascrivere il numero di stanze a uso abitativo senza considerare nel conteggio i bagni, le cucine, i cucinini, i vani accessori e le pertinenze (balconi, terrazze, verande, cantine, soffitte, garage), ma, soprattutto, dovevano specificare la tipologia dell'alloggio (distinguendo l'*abitazione occupata* da *altro tipo di alloggio).*

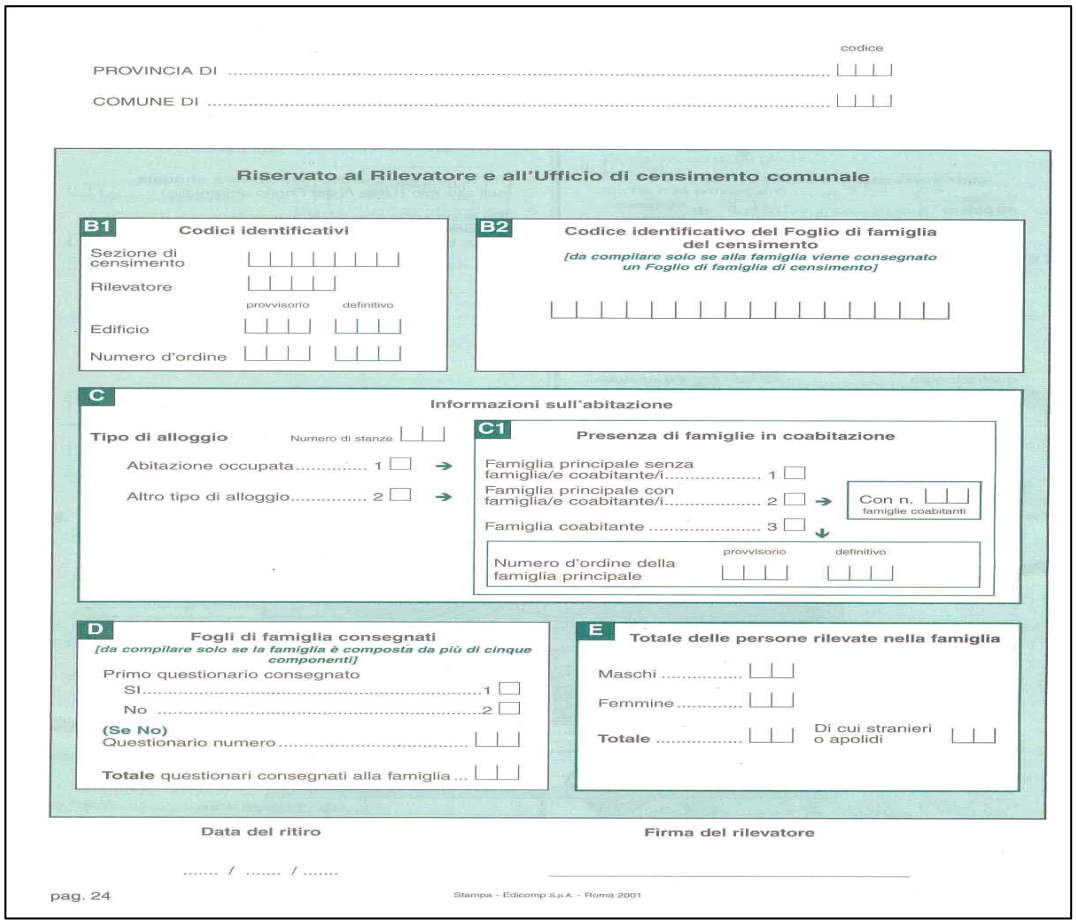
Il riquadro C1 è stato utilizzato per indicare l'eventuale presenza e il numero di famiglie coabitanti nell'alloggio, con l'inserimento dei codici identificativi per l'aggancio con la famiglia principale.

Nel caso di famiglie con più di cinque componenti, doveva essere compilato anche il riquadro D, barrando la casella SI per identificare il primo questionario, indicando il numero d'ordine corrispondente dei questionari aggiuntivi successivi al primo e, infine, riportando su ogni questionario il numero complessivo dei modelli consegnati alla stessa famiglia.

Nel riquadro E di riepilogo veniva riportato l'ammontare totale dei componenti della famiglia, con la distinzione per sesso e il numero di stranieri o apolidi. Anche su questa ultima pagina il rilevatore doveva

apporre la firma e indicare la data del ritiro del questionario, informazione utile nella successiva fase di monitoraggio.

Figura 2.2 - Pagina di chiusura del questionario dell'indagine sul grado di copertura



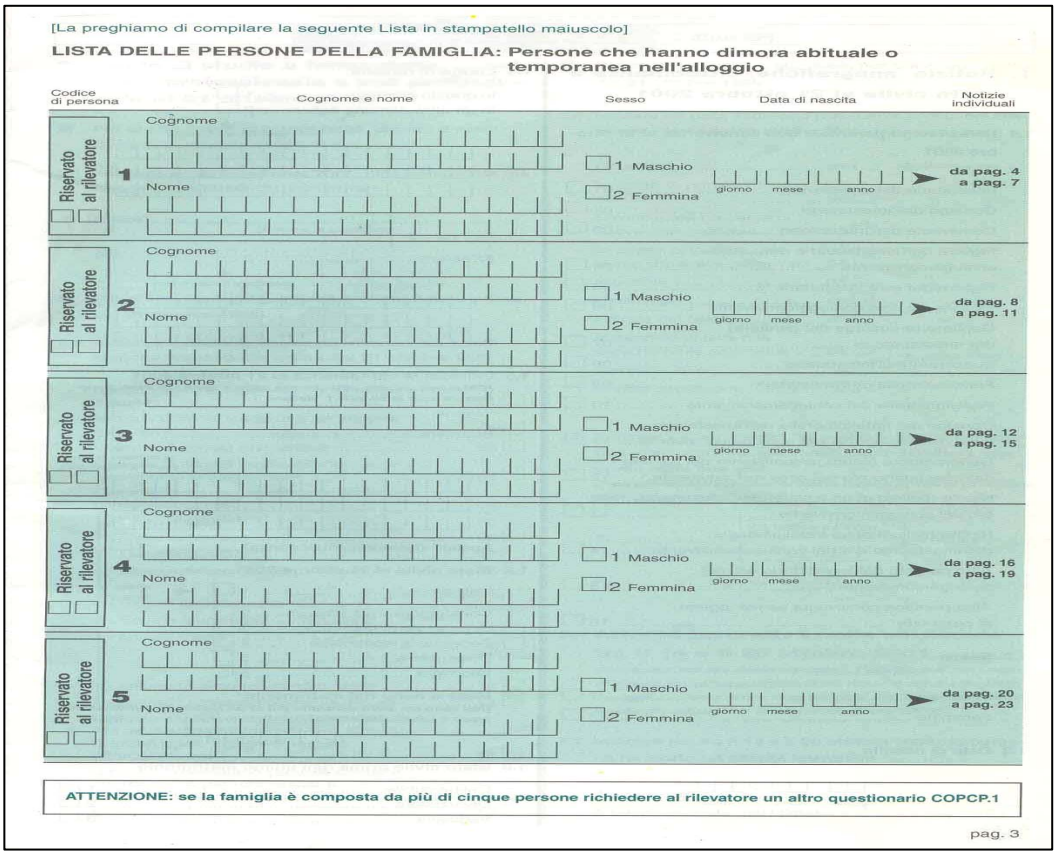
Nella seconda pagina del questionario dell'Idc è stata riportata la lettera informativa a firma del Presidente dell'Istat, con la descrizione delle fasi della rilevazione, con l'invito alle famiglie alla più ampia collaborazione nella compilazione, con i riferimenti normativi in tema di segreto statistico e riservatezza dei dati. Nella lettera era anche indicato il numero verde a disposizione dei rispondenti per eventuali chiarimenti.

Il questionario dell'Idc prosegue nella terza pagina con la *lista delle persone della famiglia* (Figura 2.3). Per ciascun componente dovevano essere indicati in stampatello maiuscolo il nome e cognome. Doveva essere biffata la casella relativa al sesso e trascritta la data di nascita nel formato numerico giorno, mese e anno. Queste informazioni, regolarmente registrate, sono risultate indispensabili nelle successive operazioni di *record linkage*.

Nell'elencare le singole persone della famiglia doveva essere scrupolosamente osservato il seguente ordine:

- intestatario del questionario (secondo le definizioni già illustrate in precedenza);
- coniuge o convivente con l'intestatario del questionario;
- figli non sposati (in ordine decrescente di età);
- figli sposati e loro familiari;
- altri parenti o affini dell'intestatario del questionario (padre, fratello, suocero, zio, nipote, cognato);
- altre persone eventualmente conviventi.

Figura 2.3 - Lista delle persone della famiglia nel questionario dell'indagine sul grado di copertura


[La preghiamo di compilare la seguente Lista in stampatello maiuscolo]

LISTA DELLE PERSONE DELLA FAMIGLIA: Persone che hanno dimora abituale o temporanea nell'alloggio

Codice di persona | Cognome e nome | Sesso | Data di nascita | Notizie individuali

Riservato al rilevatore 1	Cognome / Nome	1 Maschio, 2 Femmina	giorno mese anno	da pag. 4 a pag. 7
Riservato al rilevatore 2	Cognome / Nome	1 Maschio, 2 Femmina	giorno mese anno	da pag. 8 a pag. 11
Riservato al rilevatore 3	Cognome / Nome	1 Maschio, 2 Femmina	giorno mese anno	da pag. 12 a pag. 15
Riservato al rilevatore 4	Cognome / Nome	1 Maschio, 2 Femmina	giorno mese anno	da pag. 16 a pag. 19
Riservato al rilevatore 5	Cognome / Nome	1 Maschio, 2 Femmina	giorno mese anno	da pag. 20 a pag. 23

ATTENZIONE: se la famiglia è composta da più di cinque persone richiedere al rilevatore un altro questionario COPCP.1

pag. 3

Nella lista, per ogni componente è stato evidenziato il riferimento al numero delle pagine del questionario da completare. Nella caselle riservate al rilevatore doveva essere invece aggiornato il numero d'ordine delle persone in caso di famiglie con più di cinque componenti.

Ciascun membro della famiglia doveva compilare il proprio questionario individuale suddiviso in quattro sezioni, relativo a:

1. Notizie anagrafiche, cittadinanza e stato civile al 21 ottobre 2001;
2. Titolo di studio e formazione, condizione professionale e non professionale al 21 ottobre 2001;
3. Attività lavorativa nella settimana tra il 14 e il 20 ottobre 2001;
4. Notizie sullo stato abitativo al 21 ottobre 2001.

2.5.2 - Le sezioni del questionario

Nell'ambito della prima sezione sono compresi otto quesiti (da 1.1 a 1.8), che riguardano notizie anagrafiche e cittadinanza. Per la relazione di parentela o di convivenza si dovevano classificare tutte le persone appartenenti alla lista descritta nel paragrafo precedente. In rispetto dell'ordine di cui si è già detto la persona 1 della lista doveva classificarsi come Intestatario del questionario.

In relazione al luogo di nascita, per l'Italia andava indicata la denominazione del Comune e non quella della località (frazione o centro abitato); per i nati all'estero doveva essere dichiarata la denominazione attuale dello Stato estero entro i cui confini era compreso il luogo di nascita. Nell'indicazione della cittadinanza è stato messo in evidenza che, nel caso di più cittadinanze oltre a quella italiana, si doveva barrare solo la casella *Italiana*, mentre i cittadini stranieri erano tenuti a specificare lo Stato estero di cittadinanza.

Le persone con lo stato civile di celibe o nubile saltavano gli ultimi due quesiti della prima sezione per passare alla seconda.

Nella seconda sezione le domande sono distinte in base all'età: tra i minori di sei anni e chi ha sei anni e più. Si richiedeva ai primi di dichiarare se al 21 ottobre 2001 frequentassero la scuola elementare mentre chi aveva più di sei anni doveva invece indicare il più alto titolo di studio conseguito alla stessa data.

Il quesito sulla condizione professionale o non professionale è stato invece rivolto esclusivamente alle persone di 15 anni e oltre. Per rispondere adeguatamente era necessario attenersi alla classificazione e alle definizioni utilizzate nel corso del Censimento. Seguiva una domanda sul lavoro retribuito svolto nel corso della vita.

La terza sezione del questionario di copertura è stata compilata da tutti coloro che si sono dichiarati occupati in quella precedente, facendo riferimento esclusivamente all'attività lavorativa definita come principale, cioè quella alla quale nella settimana di riferimento precedente il Censimento (14-20 ottobre 2001) è stato dedicato il maggior numero di ore di lavoro.

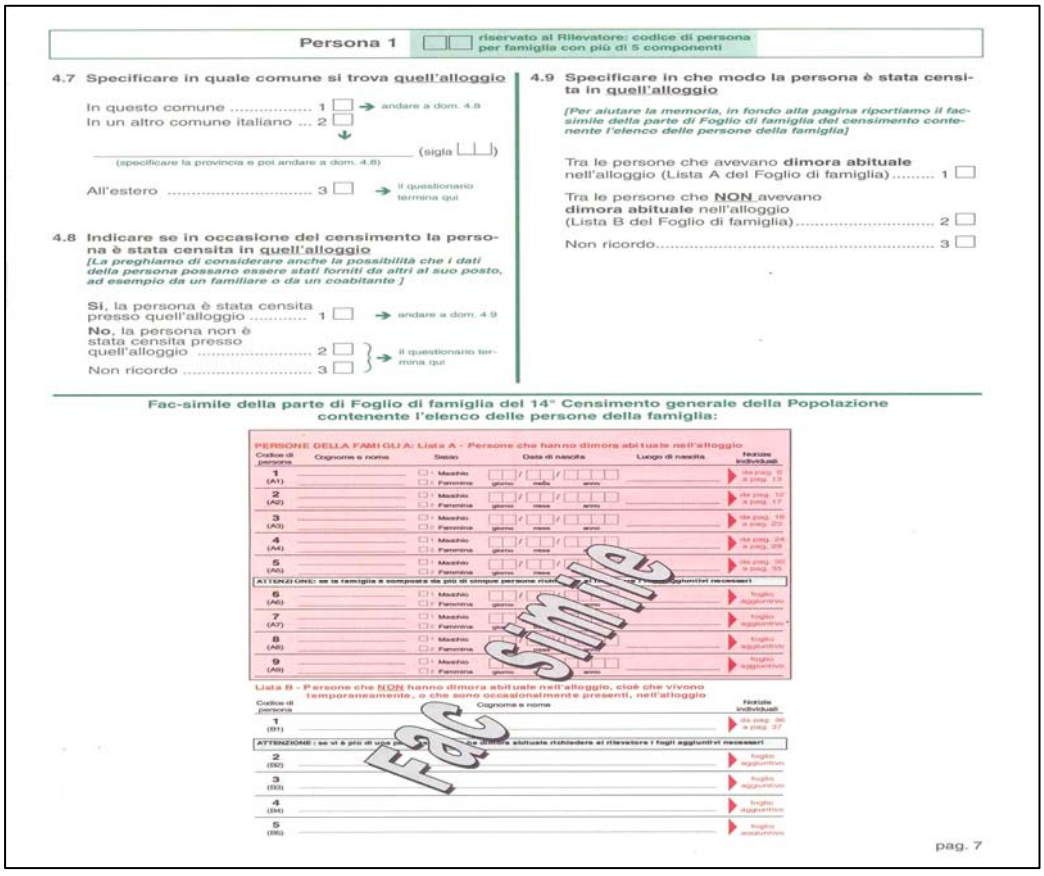
Il primo quesito ha riguardato l'orario di lavoro svolto, con la possibilità di indicare il tempo pieno o quello parziale. Il secondo ha riguardato la posizione nella professione (dipendente, autonomo, eccetera), secondo la classificazione del Censimento, e quindi il tipo di rapporto di lavoro.

In caso di lavoro a tempo determinato era necessario specificare il tipo di contratto.

Nella prima parte del questionario i quesiti hanno seguito lo schema utilizzato nel censimento; la quarta sezione invece è specifica dell'Idc. Questa sezione doveva accertare se tutte le persone fossero state censite in modo corretto ricostruendo i dettagli relativi all'alloggio e al luogo di dimora abituale alla data del censimento. Il primo quesito della sezione distingue tra due sottogruppi: coloro che anche al censimento dimoravano abitualmente nell'alloggio dove sono intervistati per l'Idc e coloro che al 21 ottobre dimoravano abitualmente in un alloggio diverso. Questo secondo gruppo doveva indicare in quale comune fosse la dimora abituale e per quanti giorni dal 21 ottobre 2000 al 20 ottobre 2001 aveva utilizzato l'alloggio attuale.

Per il primo gruppo invece è stata prevista una batteria di domande attraverso le quali capire se la persona sia stata o meno censita correttamente. In particolare, si chiede se la persona sia stata censita nell'alloggio e, in caso di risposta affermativa, se sia stata censita come residente o come temporaneamente presente. Per agevolare la risposta a fondo pagina è riportato il *fac-simile* delle liste A e B relative al questionario del Censimento della popolazione destinate rispettivamente ai residenti e non residenti (Figura 2.4). Seguono le domande sull'utilizzo di alloggio diverso da questo (per quanto tempo e in quale comune), se la persona sia stata censita in questo secondo alloggio e con quale modalità (residente o temporaneamente dimorante).

Figura 2.4 - Foglio di famiglia dell'Idc contenente il fac-simile delle liste A e B del Censimento della popolazione



2.5.3 - Alcuni risultati sul questionario dell'Indagine di copertura

Si illustrano, di seguito, le tipologie di risposta raccolte anche per delineare le principali difficoltà incontrate dai rispondenti e procedere a un bilancio.

L'Indagine di copertura ha interessato un totale di 179.886 individui con una mancata risposta parziale, relativa alla non compilazione di tutta la quarta sezione del questionario, pari allo 0,5 per cento. Dei 179.034 rispondenti il 97,4 per cento (174.291 individui) si è dichiarato "abitualmente dimorante" al Censimento nel medesimo alloggio dove è stato rilevato al momento della copertura e rientra pertanto, secondo quanto già affermato in precedenza, nell'universo strettamente di interesse dell'indagine.

La procedura di *record linkage* tra i dati dell'Idc e quelli del Censimento ha mostrato che la dichiarazione di essere stato "abitualmente dimorante al Censimento", fornita in fase di Indagine di copertura, è stata confermata nel 97,5 per cento degli individui (Tavola 2.1), mentre lo 0,7 per cento, pur dichiarandosi "abitualmente dimorante alla data del Censimento", in realtà aveva compilato durante la rilevazione censuaria la sezione relativa a un "non abitualmente dimorante".

L'incoerenza mostrata nel classificarsi al momento dell'Idc da parte dei 1.148 individui di cui sopra può essere legata a un effetto memoria, oppure ad alcune difficoltà nella comprensione del concetto stesso di "dimora abituale".

Soffermandosi sull'insieme dei rispondenti che si sono dichiarati "non ero abitualmente dimorante alla data del Censimento" in occasione dell'Indagine di copertura (nel complesso 4.743 individui), e che, quindi, sono stati esclusi dall'analisi, il 26,5 per cento attraverso il *record linkage* si collega in modo coerente ai "non abitualmente dimoranti", mentre la percentuale più significativa (37,3 per cento) si abbina a individui "abitualmente dimoranti" al Censimento. Tale incongruenza sembra spiegarsi con il disturbo statistico a carico dei rispondenti che si è venuto a creare in sede di Indagine di copertura.

Tavola 2.1 - Record linkage degli individui che hanno risposto alla IV sezione del questionario dell'Indagine di copertura con il Censimento della popolazione *(valori assoluti e percentuali)*

CENSIMENTO DELLA POPOLAZIONE	Indagine sul grado di copertura					
	Aventi dimora abituale	Valore %	Non aventi dimora abituale	Valore %	Totale	Valore %
Abbinati	171.136	98,19	3.023	63,73	174.159	97,28
Aventi dimora abituale	*169.988*	*97,53*	*1.767*	*37,25*	*171.755*	*95,93*
Non aventi dimora abituale	*1.148*	*0,66*	*1.256*	*26,48*	*2.404*	*1,34*
Non abbinati	3.155	1,81	1.720	36,26	4.875	2,72
Totale	**174.291**	**100,00**	**4.743**	**100,00**	**179.034**	**100,00**

In relazione al complesso degli individui che non è stato possibile abbinare nel corso delle operazioni di *record linkage* (4.875, pari al 2,7 per cento dei rispondenti all'Idc), per la maggior parte (64,7 per cento) si tratta di residenti in qualche modo sfuggiti al Censimento o, comunque, che si sono stabiliti in quella dimora in un momento immediatamente successivo; nel 35,3 per cento dei casi, invece, si sono dichiarati "non abitualmente dimoranti".

Focalizzando l'attenzione sui percorsi di risposta alla quarta sezione del questionario di copertura (Tavola 2.2), è possibile ricavare informazioni sul comportamento degli intervistati, sulle difficoltà incontrate nella comprensione e compilazione del questionario, sui disturbi introdotti a vari livelli nel corso di un'indagine svolta comunque a una certa distanza temporale dal Censimento.

Tavola 2.2 - Percorsi di risposta alla IV sezione del questionario Idc degli individui campione che si sono classificati "abitualmente dimoranti" al Censimento e sono stati rilevati nello stesso luogo al momento dell'Indagine

PERCORSI DI RISPOSTA	Rispondono a tutte le domande	Rispondono "non ricordo" ad alcune domande	Non rispondono ad alcune domande	Totale
Censiti correttamente	96,00	0,31	0,94	97,25
Censiti non correttamente	0,69	0,05	0,03	0,76
Confusione sul concetto di dimora abituale	*0,39*	*0,05*	*0,03*	*0,47*
Censiti come abitualmente dimoranti in due abitazioni distinte	*0,29*	-	-	*0,29*
Sfuggiti al Censimento	0,38	0,05	0,03	0,46
Non forniscono informazioni su	-	0,67	0,86	1,53
L'essere stati censiti	-	*0,24*	*0,57*	*0,82*
Come sono stati censiti	-	*0,43*	*0,29*	*0,72*
Totale	**97,07**	**1,08**	**1,85**	**100,00**

Considerando l'insieme degli individui di interesse per l'Indagine di copertura, emerge che circa il 97,1 per cento ha compilato la quarta sezione del questionario in ogni sua parte, fornendo tutte le informazioni richieste, l'1,9 per cento non ha risposto ad alcuni quesiti e l'1,1 per cento ha riscontrato alcune difficoltà, probabilmente legate a problemi di memoria.

La maggior parte dei *missing* si concentra sui quesiti relativi all'alloggio diverso da quello abituale e riguarda lo 0,9 per cento del totale degli individui di interesse; una percentuale simile si registra in corrispondenza della mancanza di informazioni quali l'essere stati censiti e il come si è stati censiti ("abitualmente dimoranti"; "non abitualmente dimoranti"): questo secondo aspetto fa emergere alcune difficoltà di comprensione del concetto di "dimora abituale".

Tra gli individui dell'Idc che non hanno ricordato in che modo sono stati censiti nella dimora abituale (lo 0,4 per cento del complesso degli "abitualmente dimoranti") si riscontra una certa difficoltà, che si spiega con il fatto che il questionario di censimento poteva essere compilato anche da una persona diversa dal rispondente all'Idc, per esempio un altro componente della famiglia. Lo 0,2 per cento ha dichiarato di non ricordare di essere stato censito (e questo solo in parte si giustifica con la presenza di più alloggi) e, infine, lo 0,3 per cento, pur affermando di essere stato censito correttamente nell'abitazione principale, ha risposto "non ricordo" ai quesiti relativi alla sua situazione rispetto alla seconda abitazione.

Nel complesso, circa il 97,3 per cento degli individui di interesse per la copertura ha dichiarato di essere stato censito in modo corretto; in tale percentuale è compresa quella residuale di coloro che non è stato possibile classificare in relazione alla seconda abitazione ("non ricordo" e *missing* su alcuni quesiti e, nello specifico, sull'abitazione diversa da quella abituale). Di conseguenza, almeno per l'abitazione principale, tutti questi individui hanno compilato la lista A del Foglio di famiglia del censimento, riferita alle persone abitualmente dimoranti nell'alloggio; nella quasi totalità dei casi si tratta di individui che non avevano altre abitazioni oltre alla dimora abituale (94,5 per cento degli individui censiti correttamente), contro il 4,5 per cento che ha dichiarato di possedere altre abitazioni in Italia e l'1 per cento con abitazioni all'estero.

Circa lo 0,5 per cento degli individui di interesse ha evidenziato un certo grado di confusione rispetto al concetto di dimora abituale e di conseguenza ha dichiarato di essere stato censito non correttamente: sia come "non abitualmente dimorante" nell'alloggio definito "abituale" (lista B del Foglio di famiglia del censimento, relativa alle persone temporaneamente o occasionalmente presenti nell'alloggio), sia come "abitualmente dimorante" (lista A) in un'altra abitazione indicata come "non abituale". Analizzando più in profondità questo piccolo gruppo di persone che hanno confuso il concetto di dimora, si rileva che quasi il 41 per cento ha dichiarato di non avere altre abitazioni di riferimento in Italia, mentre il 23,5 per cento, pur in presenza di altre abitazioni, ha ammesso di non essere stato censito altrove; poco più del 6 per cento ha dichiarato di non essere stato censito nella dimora abituale e di essersi classificato in modo errato in un altro alloggio, contro quasi il 5,2 per cento che ha selezionato la lista B del Foglio di famiglia sia nella dimora abituale sia in eventuali altri alloggi.

Tornando al complesso degli individui di interesse, sempre sulla base di quanto dichiarato in sede di Indagine di copertura, la quota di coloro che sembrano sfuggiti al Censimento non raggiunge lo 0,5 per cento, mentre risulta pari a poco meno dello 0,3 per cento la percentuale di coloro che hanno risposto di essere stati censiti come "abitualmente dimoranti" in abitazioni diverse (in totale rappresentano circa lo 0,8 per cento coloro che hanno affermato di non essere stati censiti in modo corretto). Sulla base dei risultati effettivi derivanti dalle operazioni di *record linkage* (Par. 2.8), tale quesito non può essere considerato affidabile ai fini della stima dell'errore di copertura, ma può essere utilizzato esclusivamente come ausilio al monitoraggio dei dati.

2.6 - Il sistema di monitoraggio

2.6.1 - Premessa

Nel 2001, in occasione della terza edizione dell'Idc, è stato introdotto un accurato sistema di monitoraggio che ha rappresentato una significativa innovazione di processo, finalizzata al raggiungimento di un soddisfacente grado di qualità dell'indagine. In effetti, l'obbiettivo principale è stato quello di controllare l'indipendenza dell'Idc e del Censimento attraverso una completa separazione dei momenti delle rispettive fasi di indagine. Di conseguenza, sono stati tenuti sotto controllo i rallentamenti nell'avvio delle operazioni sul campo dell'Idc che, se portati all'estremo, avrebbero causato eccessive modificazioni nella distribuzione territoriale dei residenti e difficoltà nel ricordare alcune notizie, con un acuirsi dell'effetto memoria sulle risposte.

Per questo, l'Istat ha coordinato il supporto tecnico e organizzativo di tutte le fasi della rilevazione, operando sia attraverso i contatti giornalieri diretti - telefonici, via e-mail e/o fax - con i responsabili dei 98 Ucc partecipanti ai lavori, sia attraverso i propri uffici territoriali, con il coinvolgimento delle 17 Regioni interessate dall'indagine (si ricorda che per Basilicata, Molise e Valle d'Aosta nessun comune è entrato a far parte del campione) e l'Astat (Istituto provinciale di statistica della provincia autonoma di Bolzano), che hanno visitato regolarmente i comuni di competenza e supervisionato alle varie operazioni.

Indicazioni utili per orientare in modo più corretto le azioni di monitoraggio, soprattutto in relazione all'operato dei rilevatori, sono state tratte anche dalle chiamate al numero verde, che è stato attivo dal 12 ottobre 2001 alla fine di marzo 2002, sia per il Censimento sia per l'Idc. In generale, i quesiti posti hanno riguardato nella maggior parte dei casi (circa il 68 per cento) la compilazione dei questionari, per il 27 per cento le modalità organizzative, soprattutto i tempi di consegna e il ritiro dei questionari, per il 4,5 per cento informazioni generali su obiettivi, normativa, obbligo di risposta, tutela del segreto statistico, mentre reclami (comportamento del rilevatore, mancato ritiro/consegna del questionario) e obiezioni hanno interessato complessivamente lo 0,5 per cento delle chiamate.

2.6.2 - La gestione dei contatti con i Comuni

Il sistema di monitoraggio ha interessato tutte le fasi dell'Idc, ampiamente illustrate nei paragrafi precedenti, che si sono svolte secondo il calendario riportato nella successiva tavola.

Tavola 2.3 - Calendario delle fasi dell'Idc interessate dal sistema di monitoraggio

FASI DELL'IDC INTERESSATE DAL MONITORAGGIO	Periodo
Preavviso agli Ucc dei comuni estratti sulla partecipazione all'Idc	Luglio 2001 (Circolare n. 16 del 13 luglio 2001)
Comunicazione agli Ucc delle modalità di svolgimento dell'Idc, del calendario delle operazioni, del trattamento economico	Ottobre 2001 (Circolare n. 30 dell'11 ottobre 2001)
Comunicazione agli Ucc delle sezioni di Censimento estratte	Dal 15 al 23 novembre 2001
Invio materiale dall'Istat agli Ucc	
Verifica da parte degli Ucc delle sezioni campione e comunicazione degli itinerari sezione	
Fotocopia dei fogli famiglia e dei modelli ausiliari del Censimento	Dal 23 novembre al 3 dicembre 2001
Comunicazione rilevatori estratti, assegnazione sezione e verifica rotazione dei rilevatori; formazione dei rilevatori	
Ricognizione edifici: preparazione lista provvisoria edifici	Dal 3 al 7 dicembre 2001
Consegna alle famiglie dei questionari e della guida alla compilazione	Dal 10 al 18 dicembre 2001
Ritiro dei questionari e computi giornalieri di sezione	Dal 21 dicembre 2001 al 9 gennaio 2002
Controllo effettuazione interviste attraverso contatti telefonici	
Revisione questionari; assegnazione codici definitivi edifici e numero d'ordine; compilazione stati definitivi di sezione	Entro il 18 gennaio 2002
Confezione e invio materiale all'Istat tramite corriere	Dal 21 al 31 gennaio 2002
Pagamento contributi	Da febbraio 2002

Tale calendario tiene conto di un ritardo complessivo di circa tre settimane nelle date originariamente programmate, in conseguenza dell'analoga proroga dei termini concessi ai comuni per il completamento delle operazioni di raccolta dei Fogli di famiglia del Censimento. Lo slittamento si è reso dunque necessario per evitare nel modo più assoluto la compresenza nell'ambito della stessa sezione del rilevatore del Censimento e di quello dell'Indagine di copertura. Questa eventualità, come più volte sottolineato, avrebbe causato seri problemi organizzativi e metodologici, con gravi conseguenze sul livello della qualità di entrambe le rilevazioni.

Come si evince dalla tavola 2.3, l'attività di monitoraggio ha avuto inizio a luglio 2001, in concomitanza con la prima circolare Istat relativa all'Idc, a carattere informativo e contenente l'elenco dei 98 comuni estratti per partecipare all'Indagine di copertura.

A seguire, la seconda circolare (11 ottobre 2001) ha riguardato le date di svolgimento dell'indagine, le istruzioni su come eseguire le operazioni di consegna, raccolta, revisione e restituzione del materiale, e le informazioni sui contributi economici ai Comuni. Successivamente, a partire dal 15 novembre 2001, si è provveduto a inviare ai Comuni una e-mail (o un fax, quando necessario) contenente, oltre alla lista delle sezioni estratte e ad alcune precisazioni sulle operazioni da eseguire, la nuova programmazione delle date di esecuzione dell'indagine con le posticipazioni di cui si è detto.

Contemporaneamente a tale comunicazione è stato spedito il materiale necessario a eseguire le diverse operazioni dell'Idc, insieme alle scorte destinate agli uffici regionali dell'Istat, nel caso singoli Comuni avessero dovuto ricorrere a materiale aggiuntivo. Per quanto riguarda i tempi, la consegna è stata effettuata per quasi tutti i Comuni entro la data prevista del 23 novembre 2001, con qualche eccezione in cui si è verificato uno slittamento tra l'ultima settimana di novembre e la prima settimana di dicembre a causa di alcuni problemi in fase di stampa e consegna dei modelli. In ogni caso, i margini di sicurezza previsti per le quantità da inviare a ciascun Comune hanno garantito dappertutto l'arrivo di materiale sufficiente alla gestione del lavoro.

Contatti telefonici continui, effettuati direttamente dall'Istat o mediati dagli uffici regionali, sono stati disposti per tenere sotto controllo lo stato di completamento delle operazioni relative all'Indagine di copertura da parte dei comuni. Nel complesso, il monitoraggio ha evidenziato che, rispetto alle date prefissate per l'inizio della rilevazione sul campo (Tavola 2.3), sono stati registrati ritardi aggiuntivi dovuti principalmente all'ulteriore protrarsi delle operazioni di censimento, soprattutto nei grandi comuni campione e in quelli metropolitani. Dal monitoraggio risulta che 58 comuni hanno iniziato le operazioni di contatto delle famiglie secondo quanto stabilito dal calendario e quindi in epoca precedente al 24 dicembre 2001, 40 Comuni hanno registrato un ritardo: 21 Comuni hanno cominciato tra il 24 dicembre e il 7 gennaio 2002 e 19 dopo tale data.

Nello specifico, alla fine di marzo 2002, le attività di raccolta presso le famiglie dei questionari compilati si erano concluse per circa il 74 per cento dei comuni campione e poco più del 16 per cento aveva già inviato tutto il materiale all'Istat. La raccolta dei questionari è avvenuta più speditamente nel Nord (Tavola 2.4) e nei comuni con popolazione inferiore ai 100 mila abitanti.

Tavola 2.4 - Distribuzione dei comuni campione che hanno concluso le operazioni di raccolta dei questionari Idc entro il mese di marzo 2002 e della relativa percentuale di famiglie campione, per ripartizione geografica e ampiezza demografica

RIPARTIZIONI GEOGRAFICHE E AMPIEZZA DEMOGRAFICA	Numero di Comuni	% Comuni campione nello stesso strato	% Famiglie campione nello stesso strato
Nord	33	82,5	79,7
Centro	15	75,0	59,0
Sud e Isole	24	63,2	61,3
< 10.000 abitanti	22	73,3	71,1
10.001 |–| 100.000 abitanti	34	89,5	89,6
> 100.000 abitanti (a)	11	61,1	59,4
Comuni metropolitani	5	41,7	39,0
Totale	**72**	**73,5**	**68,4**

(a) Esclusi i comuni metropolitani.

Nel complesso, dato che l'inizio delle fasi dell'Idc era subordinato al termine della raccolta dei modelli di censimento, i ritardi che si sono verificati sono stati influenzati dall'andamento delle operazioni di quest'ultimo e hanno interessato soprattutto i comuni cosiddetti metropolitani. Tra questi, Firenze, Roma, Palermo, Messina e Cagliari alla fine di marzo 2002 non avevano ancora iniziato le operazioni di raccolta dei questionari presso le famiglie.

2.6.3 - La permanenza dei questionari presso le famiglie

Insieme alla durata complessiva delle operazioni dell'Indagine di copertura, sembrano utili alcune considerazioni sui tempi di permanenza effettivi dei questionari presso le famiglie, che si possono trarre dall'indicazione da parte del rilevatore delle date di consegna e ritiro.

Prendendo in esame solo i questionari con le informazioni complete relativamente a entrambe le date (65.640, il 94 per cento del totale dei questionari compilati e inviati all'Istat), questi sono rimasti mediamente a disposizione delle famiglie per poco meno di una settimana (Tavola 2.5). Nella maggior parte dei casi (poco più dell'87 per cento) si tratta di una permanenza che si attesta entro le due settimane, con un numero medio di giorni pari a poco più di 4. Circa l'11 per cento dei questionari è stato ritirato nel corso dei primi 30 giorni dalla consegna (in questo caso la durata media è risultata pari a 21 giorni), mentre nell'1,7 per cento dei casi sono stati superati i 30 giorni, con una permanenza media di quasi 44 giorni.

La durata della permanenza è strettamente legata all'ampiezza demografica dei comuni campione, con un massimo in quelli metropolitani (i questionari sono stati ritirati in media dopo circa 9 giorni) e un minimo di 5 giorni nei comuni tra i 10 mila e i 100 mila abitanti. Nei comuni campione di dimensione demografica più contenuta (inferiore a 10 mila abitanti) il numero medio di giorni intercorso tra consegna e ritiro dei questionari è risultato pari a 6. In relazione alle classi di permanenza considerate, se fino ai 30 giorni si nota comunque una certa omogeneità tra i comuni campione nell'organizzare le operazioni di ritiro, differenze più marcate si registrano nell'ultima classe con una permanenza media che arriva a circa 53 giorni proprio nei comuni tra i 10 mila e i 100 mila abitanti.

Per quanto riguarda la situazione per ripartizione geografica (Tavola 2.6), tempi di permanenza più prolungati dei questionari presso le famiglie si sono riscontrati nel Centro e nel Sud (rispettivamente una media di circa 8 e 9 giorni), per tutti i comuni campione indipendentemente dall'ampiezza demografica, con l'eccezione dei comuni con popolazione superiore ai 100 mila abitanti. Rispetto a questi ultimi, il Sud ha registrato una media di 10 giorni e il Nord-ovest di circa 7.

I Comuni delle Isole hanno registrato intervalli di tempo più brevi tra consegna e ritiro dei modelli di rilevazione (in media poco più di 5 giorni), anche se per i comuni al di sotto dei 10 mila abitanti il Nord-est ha mostrato una rapidità maggiore nel ritiro dei questionari, attestandosi sui 4 giorni.

2.6.4 - I problemi di lista del campione di sezioni

Nella fase di monitoraggio l'attività più impegnativa è stata senz'altro quella di assistenza agli uffici comunali, affinché la struttura del campione dell'Indagine di copertura fosse il più possibile rispettata.

Questa necessità è stata indotta dal fatto che il campione programmato ha utilizzato, per ciascuno dei comuni coinvolti, da un lato informazioni non ancora definitive sulle basi territoriali comunali e dall'altro una stima del numero di famiglie residenti per sezione. Infatti, al momento dell'indagine il lavoro di definizione delle sezioni di censimento non si era ancora completamente concluso, e i relativi dati numerici delle sezioni, distinti per tipologia di località abitata, non erano stati ancora resi definitivi per tutti i comuni. In conseguenza si è verificato un numero complessivo di famiglie campione sostanzialmente differente da quello programmato in fase di disegno campionario; si sono, inoltre, riscontrate anomalie ed errori di diversa natura nelle liste dei codici identificativi delle sezioni utilizzate come base per l'estrazione del campione di sezioni.

Per far fronte all'indisponibilità, per molti dei comuni campione, della lista definitiva delle sezioni di censimento (si precisa che l'estrazione del campione di sezioni è avvenuta alcuni mesi prima dell'inizio della rilevazione censuaria), si è deciso di utilizzare la lista più aggiornata al momento disponibile e di lasciare al confronto con gli organi comunali in fase di monitoraggio dell'indagine la revisione del campione e ancora, in fase di revisione post-censuaria delle basi territoriali, l'aggiornamento delle liste.

Per tale motivo, si è deciso di fornire a ciascuno dei comuni estratti per l'indagine, in aggiunta all'elenco delle sezioni campione distinte per tipologia di località abitata, un elenco di sezioni sostitutive a cui far ricorso per le cadute dovute ai sopradescritti problemi di lista.

Tavola 2.5 - Questionari Idc per ampiezza demografica dei comuni campione e durata della permanenza presso le famiglie (in giorni)

GIORNI DI PERMA-NENZA	Ampiezza demografica dei comuni												Totale		
	Comuni metropolitani			>100.000 abitanti			10.001|-|100.000 abitanti			<10.000 abitanti					
	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni
0-15	13.072	79,2	5,1	20.212	88,3	4,2	10.778	92,0	3,3	13.230	91,0	4,4	57.292	87,3	4,3
16-30	3.083	18,7	20,9	2.175	9,5	21,8	758	6,5	20,0	1.198	8,2	20,2	7.214	11,0	21,0
>30	355	2,2	40,2	491	2,1	42,2	177	1,5	52,6	111	0,8	45,8	1.134	1,7	43,6
Totale	**16.510**	**100,0**	**8,8**	**22.878**	**100,0**	**6,7**	**11.713**	**100,0**	**5,1**	**14.539**	**100,0**	**6,1**	**65.640**	**100,0**	**6,8**

Tavola 2.6 - Questionari Idc per ampiezza demografica dei comuni campione, durata media della permanenza presso le famiglie (in giorni) e ripartizione geografica

RIPARTI-ZIONI GEOGRA-FICHE	Ampiezza demografica dei comuni												Totale		
	Comuni metropolitani			>100.000 abitanti			10.001|-|100.000 abitanti			<10.000 abitanti					
	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni	Questio-nari	Valore %	Numero medio di giorni
Nord-est	4.181	25,3	7,1	5.217	22,8	5,3	1.806	15,4	5,1	3.087	21,2	4,1	14.291	21,8	5,5
Nord-ovest	3.317	20,1	9,6	3.542	15,5	6,6	2.684	22,9	3,5	1.848	12,7	6,2	11.391	17,4	6,7
Centro	3.092	18,7	11,0	3.565	15,6	4,8	2.849	24,3	6,7	3.872	26,6	7,8	13.378	20,4	7,5
Sud	2.511	15,2	10,6	6.861	30,0	10,1	2.271	19,4	5,6	2.917	20,1	6,3	14.560	22,2	8,7
Isole	3.409	20,6	6,8	3.693	16,1	4,3	2.103	18,0	4,0	2.815	19,4	5,4	12.020	18,3	5,2
Totale	**16.510**	**100,0**	**8,8**	**22.878**	**100,0**	**6,7**	**11.713**	**100,0**	**5,1**	**14.539**	**100,0**	**6,1**	**65.640**	**100,0**	**6,8**

2.6.5 - Le operazioni di controllo del campione in corso d'opera

I problemi di lista sopra descritti hanno causato, per alcuni comuni, conseguenze non trascurabili sulla composizione del campione di sezioni, per cui si è reso necessario un contatto diretto con l'Istat per apportare le modifiche necessarie al campione di sezioni al fine di renderlo compatibile con quello progettato.

In generale, i cambiamenti nella struttura del campione di sezioni, avvenuti nella fase precedente la raccolta dei dati in campo, hanno riguardato quattro tipi di variazioni:

a) riduzione del numero di sezioni campione;
b) aumento del numero di sezioni campione;
c) sostituzione di alcune sezioni campione;
d) nuova estrazione del campione di sezioni.

Nei casi a) e b) non ci sono state sostituzioni delle sezioni originariamente estratte per la formazione del campione, mentre nei casi c) e d) si sono verificate sostituzioni parziali o totali dell'insieme del campione di sezioni.

La tavola successiva (Tavola 2.7) riassume in maniera più dettagliata i motivi che hanno indotto variazioni nella struttura del campione iniziale di sezioni.

Tavola 2.7 - Motivi della variazione della struttura del campione iniziale di sezioni

OPERAZIONI	Motivi	Variazioni
Non sostituzione	Presenza di un numero troppo elevato di famiglie nelle sezioni del campione iniziale	Riduzione del campione di Sezioni
	Presenza di sezioni che non contengono unità da rilevare (sezioni vuote)	
	Assenza nella base territoriale di un dato comune campione, di sezioni appartenenti a una data tipologia di località	
	Necessità di aumentare la precisione dei dati di input (numero di famiglie - informazioni sulle basi territoriali) utilizzati per la formazione del campione iniziale	
	Esiguità del numero di famiglie nelle sezioni del campione iniziale	Aumento del Campione di Sezioni
Sostituzione	Presenza di un numero troppo elevato di famiglie nelle sezioni del campione iniziale	Sostituzione di alcune Sezioni campione
	Esiguità del numero di famiglie nelle sezioni del campione iniziale	
	Presenza di sezioni nel campione iniziale che appartengono a località con tipologia diversa da quella preventivata	
	Presenza di sezioni che non contengono unità da rilevare (sezioni vuote)	
	Presenza nel campione iniziale di codici di sezione errati che non corrispondono a nessuna sezione per la data tipologia di località	
	Presenza nel campione iniziale di alcune sezioni con problemi logistici che ne impediscono la rilevazione	
	Casi in cui si registrano gravi errori nei dati di input (numero di famiglie - informazioni sulle basi territoriali) utilizzati per la formazione del campione iniziale	Estrazione di un nuovo campione di Sezioni

La variazione alla composizione del campione iniziale di sezioni ha riguardato circa il 35 per cento dei comuni campione (35 su 98).

Le modifiche apportate sono state tali per cui per 22 dei 35 comuni interessati alla modifica è rimasto inalterato il numero previsto di sezioni su cui effettuare l'indagine, in 10 casi su 35 si è avuta una riduzione del numero di sezioni mentre solo in un comune c'è stato un aumento del numero di sezioni campione.

La tavola 2.8 riassume il numero di comuni per i quali è stato modificato il campione iniziale, classificati sia in base al tipo di modifica intervenuta sul numero di sezioni campione che sul tipo di intervento avvenuto sulle sezioni campione iniziali.

Tavola 2.8 - Casi di variazione del campione per tipologia di modifica e per tipo di intervento sul campione iniziale

TIPI DI VARIAZIONE	Comuni campione coinvolti		Tipo di intervento sulle sezioni campione
Riduzione	9	10	Non sostituzione
Aumento	1		
Sostituzione	19	25	Sostituzione
Nuova estrazione	6		

A conclusione delle operazioni, le variazioni nel numero di sezioni hanno condotto a una riduzione del numero finale di sezioni campione dalle 1.154 programmate alle 1.099 definitive con una variazione pari a -5,0 per cento, mentre il numero di famiglie ha subito una diminuzione più contenuta e pari complessivamente a -0,7 per cento.

Nella tavola 2.9 sono riportati i dati relativi al numero di sezioni campione e di famiglie interessate dalla rilevazione, distinti anche in base alla tipologia di località, nel campione programmato e nel campione finale osservato e risultante dalle modifiche intervenute in seguito alla revisione avvenuta prima dell'inizio dell'indagine, come descritto.

Tavola 2.9 - Struttura del campione – Sezioni e Famiglie – per tipologia di località abitata. Confronto tra campione programmato e campione osservato

UNITÀ CAMPIONARIE	Tipo di località						Totale	
	Centri		Nuclei		Case sparse/ Località produttive			
	Sezioni	Famiglie	Sezioni	Famiglie	Sezioni	Famiglie	Sezioni	Famiglie
Campione programmato	817	63.182	112	2.037	225	3.638	1.154	68.857
Campione osservato	805	63.051	108	1.832	186	3.491	1.099	68.374
Variazione percentuale	-1,5	-0,2	-3,7	-11,2	-21,0	-4,2	-5,00	-0,7

Risulta chiaramente che le differenze più forti si sono registrate per le sezioni del tipo case sparse/località produttive (-21,0 per cento), dove l'esiguo numero di famiglie e individui ha reso difficile la pianificazione in sede di disegno campionario.

In relazione alle variazioni nel numero delle famiglie, nel complesso la situazione risulta decisamente migliore, molto buona per il centro (pari ad appena -0,2 per cento) dove la pianificazione campionaria ha avuto una risposta estremamente precisa, leggermente meno per i nuclei (-3,7 per cento) e le case sparse (-4,2 per cento) a causa della maggiore variabilità del numero delle famiglie.

D'altra parte, come è stato più sopra sottolineato, il disegno campionario ha portato a rilevare un numero totale di famiglie estremamente vicino a quanto pianificato in fase di progettazione del campione (con una differenza che rimane al di sotto dell'1 per cento) e ciò soprattutto grazie al notevole lavoro di supporto agli organi periferici, ai responsabili comunali, durante l'intero svolgimento dell'indagine al fine di rendere la struttura del campione il più possibile coerente con quello definito inizialmente.

In conclusione, c'è stato un intenso lavoro di contatto quasi giornaliero con molti dei Comuni interessati dall'indagine affinché la struttura delle sezioni campione fosse il più possibile rispettata; per alcuni Comuni ci sono stati particolari problemi per la definizione del campione finale di sezioni, senza che ciò abbia impedito la

definizione del campione finale e l'inizio dell'indagine da parte dei rilevatori. Inoltre, durante il contatto con i Comuni si è provveduto, in alcuni casi, ad aggiornare le liste delle sezioni universo, e a individuare, escludere e sostituire le sezioni non eleggibili perché prive di unità statistiche oggetto di rilevazione (famiglie e/o abitazioni non occupate).

2.7 - L'architettura informatica

2.7.1- Premessa

Nel paragrafo si descrive l'architettura informatica del sistema informativo predisposto per l'Indagine di copertura, riferendosi all'organizzazione dell'intera applicazione in termini di componenti e dati, nonché alle caratteristiche e ai vincoli tra le componenti e i dati stessi.

L'Indagine di copertura si è avvalsa di un sistema informativo di supporto alle fasi di lavorazione dei dati, al fine di garantire un elevato livello di qualità dei dati prodotti.

Il trattamento strutturato e proceduralizzato dei dati è stato adottato per garantire la ripetibilità del risultato e, quindi, la possibilità di controllare la qualità dei risultati intermedi e di intervenire opportunamente in caso si riscontrino anomalie.

Il sistema informatico a supporto dell'indagine è stato progettato prendendo in esame alcuni requisiti che vengono descritti in dettaglio di seguito.

2.7.2 - La tipologia di utenti del sistema

Le fasi di lavorazione dei dati dell'indagine hanno previsto l'impiego di cinque operatori, i quali hanno avuto accesso ai dati attraverso postazioni contigue di personal computer.

L'esiguo numero di operatori e la loro collocazione sono stati sicuramente fattori che hanno permesso di operare una scelta orientata a un sistema *client/server* tradizionale; in tal caso, infatti, è stato realisticamente possibile effettuare installazioni della parte *client* dell'applicazione su ciascun pc, essendo in numero ridotto e prossimi tra loro.

Inoltre, la fase di collaudo e messa a punto del sistema - che ha coinvolto gli operatori su due versioni successive dell'applicazione stessa - è stata condotta nello stesso edificio degli operatori rendendo possibile la rapida installazione sui loro pc della versione più aggiornata.

2.7.3 - La natura dei dati

I dati gestiti dal sistema prevedevano quattro tipi principali di record, archiviati attraverso un Rdbms (*Relational Data Base Management System*) Oracle, rispettando inizialmente la loro struttura di registrazione:

- record relativi a informazioni della famiglia rilevati direttamente dall'Indagine di Copertura, dell'ordine di circa 70 mila;
- record relativi agli individui che compongono le famiglie di cui al punto precedente, dell'ordine di circa 180 mila;
- record relativi a informazioni della famiglia e delle abitazioni rilevati dal Censimento della popolazione, dell'ordine di circa 84 mila record;
- record relativi a informazioni degli individui facenti parte delle famiglie di cui al punto precedente, rilevati dal Censimento della popolazione, dell'ordine di circa 182 mila record;
- record relativi alle abitazioni non occupate, dell'ordine di circa 15.500 record.

Le fasi successive di lavorazione dei dati hanno previsto l'utilizzo di nuove tabelle, che descrivono in modo semplice ed efficiente la struttura familiare, sia per quanto concerne i dati dell'Indagine di copertura che quelli del Censimento.

Tali tabelle, valorizzate attraverso passi funzionali di pulizia e completamento di dati mancanti, hanno costituito la sorgente per le funzioni di *merge* e di *record linkage*.

Si è stabilito di adottare un criterio di ridondanza per la memorizzazione dei dati durante le varie fasi di lavorazione per favorire le prestazioni dei programmi di *record linkage* predisponendo strutture di dati

ottimizzate per il loro input: ciò ha richiesto una replicazione di dati, con un costo di spazio disco del tutto accettabile e con uno sforzo per il mantenimento del loro allineamento del tutto trascurabile.

2.7.4 - I controlli preliminari

Durante la fase di ricezione e in fase di inserimento dei primi dati si è notato uno scostamento degli stessi rispetto a quanto previsto dal modello concettuale per il sistema stabilito in fase di progettazione.

Pertanto si è deciso di approntare alcuni controlli preliminari da eseguire a ogni ricezione dei dati dalla società di registrazione e in particolare:

- verifica della presenza di record aventi il numero d'ordine nullo;
- verifica dell'uguaglianza dei codici sezione riportati in sezioni diverse dello stesso questionario;
- controllo della presenza di record aventi numeri d'ordine duplicati appartenenti a famiglie a cui è stato consegnato un solo questionario.

Per quanto riguarda le famiglie pluriquestionario, quelle cioè a cui è stato consegnato più di un questionario a causa della elevata numerosità della famiglia, si è controllato che fosse valorizzato il campo identificativo del primo questionario.

Si è proceduto, inoltre, a creare un piccolo archivio di indirizzi con denominazione stabilita a priori per ogni sezione interessata dall'indagine: ciò ha permesso la normalizzazione degli indirizzi effettivamente disponibili sui questionari.

L'archivio è stato realizzato utilizzando gli itinerari di sezione che i Comuni hanno inviato in fase di svolgimento dell'indagine stessa.

2.7.5 - L'architettura client/server a due livelli

Le funzionalità offerte dall'applicazione possono essere classificate in:

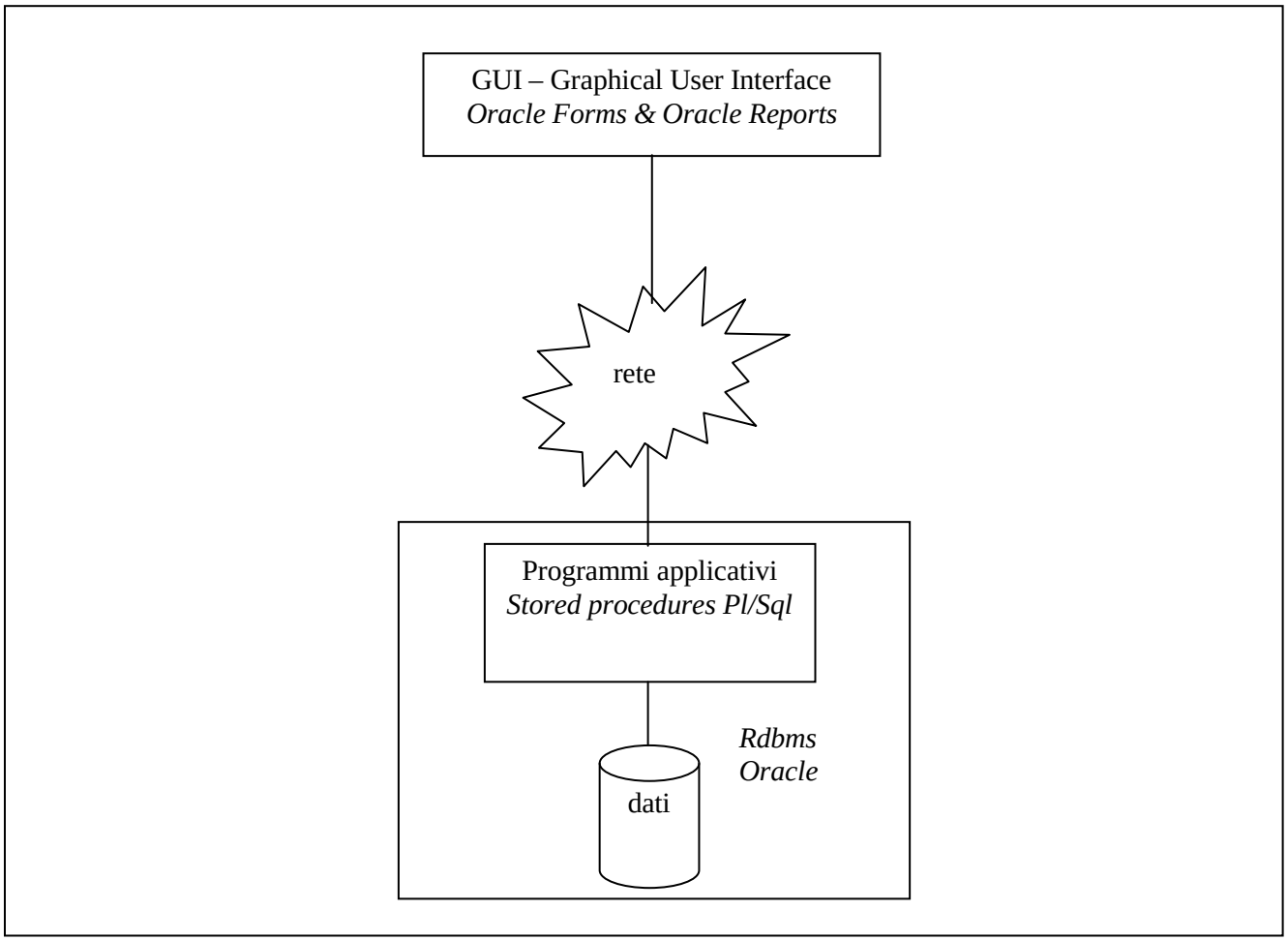
- funzionalità di interfaccia utente;
- funzionalità di elaborazione;
- funzionalità di gestione permanente dei dati.

A seconda dell'allocazione delle funzionalità a processi distinti, si distinguono le architetture *client/server* a due o tre livelli. L'architettura adottata per il sistema è stata di tipo tradizionale *client/server* a due livelli (Figura 2.5), tale per cui la funzionalità di interfaccia utente risiede sul pc, mentre le funzionalità di elaborazione (o applicativa) e quella di gestione permanente sui dati risiedono sul server.

In particolare, la scelta della base dati Oracle ha permesso di utilizzare le *stored procedures Pl/Sql* per la realizzazione delle funzionalità applicative, ovvero procedure e funzioni che sono memorizzate e gestite direttamente dal Rdbms, garantendo così efficienza di interazione con i dati.

Di fatto esiste una minima porzione di logica applicativa anche sulla parte *client* del sistema, finalizzata per lo più a controlli preliminari sulla corretta acquisizione dati da parte dell'utente.

Figura 2.5 - Architettura client/server a due livelli



2.7.6 - Gli strumenti informatici standard in Istat e l'ambiente di sviluppo Oracle Forms

La scelta degli strumenti informatici è stata dettata, soprattutto, dalla disponibilità degli stessi all'interno dell'Istituto nazionale di statistica.

La base dati è stata predisposta su ambiente Oracle, in quanto ciò garantiva la disponibilità del servizio di *backup* gestito dai sistemisti dell'Istat. Inoltre, tale base dati garantiva la centralità dei dati e la loro accessibilità da parte di tutti gli operatori.

La funzionalità di interfaccia utente è stata sviluppata in ambiente Oracle Developer, utilizzando i *tool* Oracle Forms e Oracle Report: tale scelta è stata dettata dalla diffusa conoscenza dello strumento, dalla sua estrema compatibilità con la base dati Oracle, dalla sua versatilità e rapidità di sviluppo e per la sua semplicità di utilizzo.

2.7.7 - Le fasi di lavorazione assistite dal sistema

Le operazioni di trattamento dei dati hanno previsto la lavorazione, da parte dell'operatore designato, di una sezione di censimento alla volta. Una volta selezionata una sezione da lavorare, l'applicazione informatizzata proponeva una schermata base dalla quale l'operatore accedeva a ciascuna fase di lavorazione selezionando, in una sequenza prestabilita, le diverse schede previste e, al loro interno, i pulsanti di attivazione delle procedure.

In fase di progettazione del software si è deciso di limitare al minimo la possibilità di intervento degli operatori sul programma in modo da limitare al minimo gli errori causati sia dalla complessità del software sia

dalla preparazione informatica degli operatori. Per questo, la gran parte dei controlli effettuati sono eseguiti in maniera semi-automatica; nella maggior parte dei casi, cioè, all'operatore è stato richiesto di scegliere un pulsante sulla maschera per avviare le procedure di controllo e, al termine, di controllare che l'esito delle procedure fosse conforme alle attese.

Le principali fasi di lavorazione eseguite hanno riguardato il caricamento delle basi dati da elaborare, l'esecuzione di una serie di controlli quali/quantitativi sui dati, la preparazione dei dati alle operazioni di abbinamento e l'esecuzione vera e propria dei vari passi di abbinamento.

Al solo scopo di dare un'idea della complessità del sistema, l'elenco che segue riporta l'intero insieme delle sottofasi di elaborazione, suddiviso nelle relative schede da attivare in cascata, normalmente eseguite su una sezione di censimento. Nella successiva figura (Figura 2.6) è invece riportata una schermata del programma di interfaccia utente, in cui è evidenziata la scheda "famiglia" la quale, insieme alla scheda "individuo", rappresenta il nucleo centrale delle elaborazioni, dove si svolge il *record linkage* tra i record riferiti all'Idc e quelli riferiti al Censimento.

Scheda "Info sezione"
- Caricamento delle tabelle di lavoro per la sezione di censimento da elaborare
- Informazioni di riepilogo sui dati Idc per la sezione caricata
- Informazioni di riepilogo sui dati Cen per la sezione caricata

Scheda "Procedure Idc"
- Controllo di integrità dei codici identificativi delle tabelle dati
- Normalizzazione dei cognomi e dei nomi degli individui
- Normalizzazione manuale degli indirizzi delle abitazioni
- Normalizzazione automatica degli indirizzi delle abitazioni
- Ricostruzione e controlli di congruenza sugli individui appartenenti alla stessa famiglia
- Creazione delle variabili chiave per l'esecuzione del *merge*/*record linkage*
- Caricamento tabelle per il *merge/record linkage*

Scheda "Procedure Cen"
- Controllo di integrità dei codici identificativi delle tabelle dati
- Normalizzazione dei cognomi e dei nomi degli individui
- Normalizzazione manuale degli indirizzi delle abitazioni
- Normalizzazione automatica degli indirizzi delle abitazioni
- Ricongiungimento automatico delle informazioni anagrafiche sulla famiglia con i corrispondenti record di Censimento
- Stampe per la riconciliazione di situazioni anomale riscontrate al passo precedente
- Ricongiungimento manuale delle informazioni anagrafiche sulla famiglia con i corrispondenti record di Censimento
- Creazione delle variabili chiave per l'esecuzione del *merge/record linkage*
- Caricamento tabelle per il *merge/record linkage*

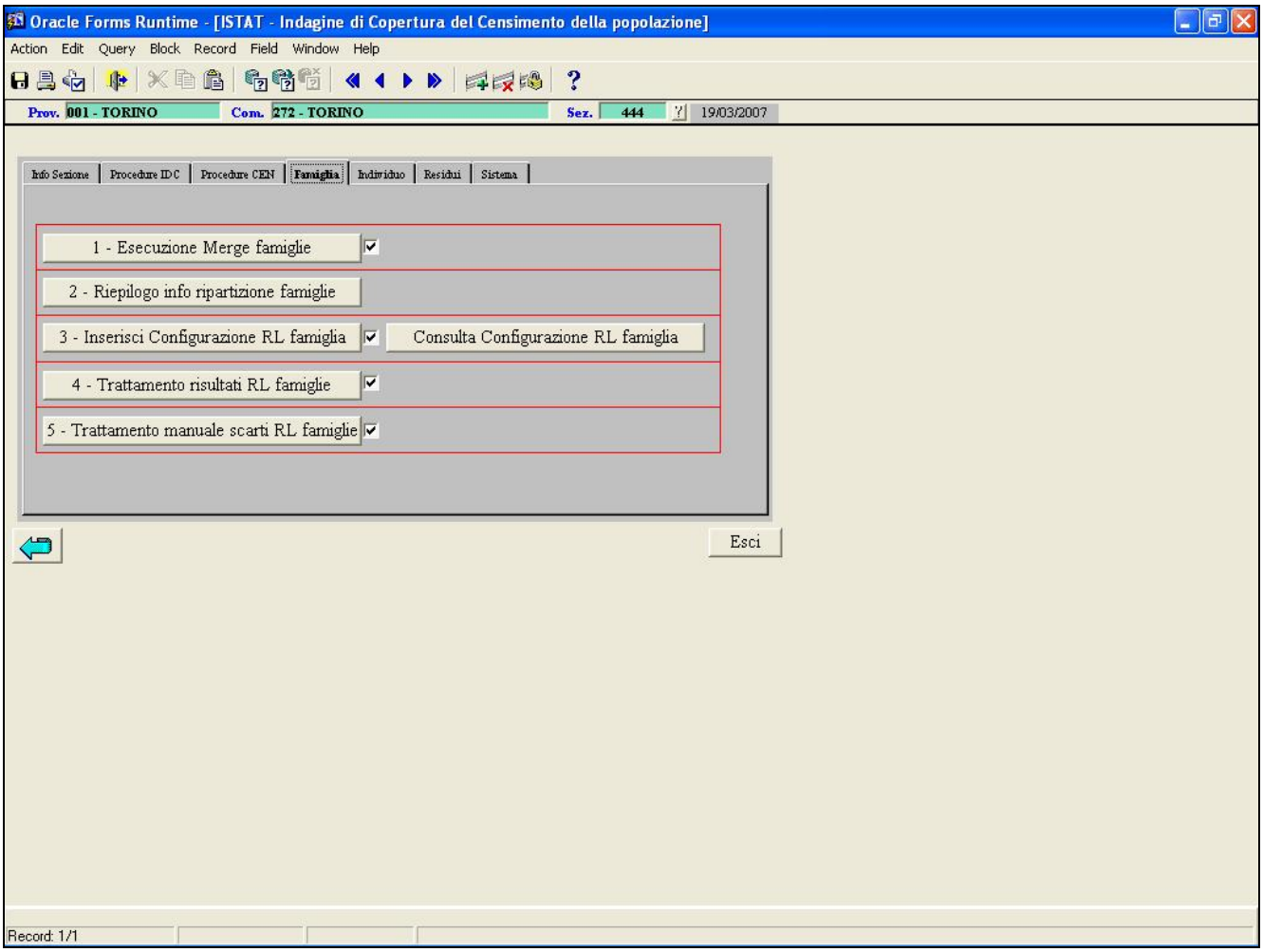
Scheda "Famiglia"
- Esecuzione *merge famiglie*
- Informazioni di riepilogo sulla frequenza delle famiglie nella sezione di censimento secondo l'indirizzo e la prima lettera del cognome dell'intestatario del Foglio di famiglia
- Attivazione delle procedure di *record linkage* probabilistico delle famiglie
- Trattamento risultati *record linkage* famiglie
- Trattamento manuale scarti *record linkage* famiglie: esecuzione del *merge* degli individui all'interno delle famiglie abbinate

Scheda "Individuo"
- Gestione degli scarti del *merge individui in famiglie abbinate*
- Esecuzione *merge individui residui*
- Informazioni di riepilogo sulla frequenza degli Individui nella sezione di censimento secondo l'indirizzo e la prima lettera del cognome
- Attivazione delle procedure di *record linkage* probabilistico degli individui
- Trattamento risultati *record linkage* individui residui
- Trattamento scarti *record linkage* individui residui

Scheda "Residui"

- Stampa della lista degli individui non abbinati per Idc
- Stampa della lista degli individui non abbinati per Cen
- Abbinamento manuale degli individui residui di Idc e Cen
- Riepilogo della sezione
- Stampa della lista degli individui non abbinati Idc
- Stampa della lista degli individui non abbinati Cen

Figura 2.6 - Esempio di schermata del programma di interfaccia utente per la fase di trattamento dei dati



2.8 - Il processo di integrazione dei dati

2.8.1 - Premessa

Le stime del grado di copertura del Censimento sono state prodotte tramite abbinamento esatto[22] (o *record linkage*) dei dati elementari del Censimento con quelli dell'Indagine di copertura. L'obiettivo del processo di abbinamento (o integrazione) è stato determinare il numero degli individui e delle famiglie rilevati in entrambe le occasioni e, per differenza, il numero degli individui e delle famiglie sfuggiti a una delle due rilevazioni. L'utilizzo del metodo *dual system* per la stima dell'errore di sottocopertura presuppone che il riconoscimento

[22] Il termine "abbinamento esatto" (o "*record linkage*") si riferisce all'uso di tecniche algoritmiche per identificare record relativi ad una stessa unità statistica e contenuti in archivi diversi (per un'introduzione si veda, ad esempio, Winkler W. E., 2001).

delle unità enumerate in entrambe le occasioni avvenga senza errori. Dal momento che errori di abbinamento,[23] anche molto contenuti, potevano compromettere l'affidabilità delle stime, l'accuratezza del processo di abbinamento ha assunto un'importanza cruciale al fine di fornire un'informazione corretta sul grado di copertura del Censimento della popolazione.

Le procedure utilizzate per l'abbinamento si basano generalmente sul confronto delle modalità assunte da un sottoinsieme di variabili comuni (*variabili di abbinamento*) agli archivi da integrare. Per garantire la possibilità di utilizzare variabili di abbinamento con un elevato potere identificativo nei confronti delle unità osservate, si è deciso di acquisire, per la prima volta in Italia da quando si svolge il Censimento della popolazione, i nominativi e gli indirizzi degli individui rilevati sia al Censimento che all'Indagine di copertura.

Per quanto riguarda i dati censuari, l'acquisizione di tali informazioni è avvenuta mediante il recupero, per le sole sezioni campione, dei cosiddetti lembi,[24] riportanti l'indirizzo della famiglia e nome, cognome, data di nascita, sesso, luogo di nascita di ogni componente. Questi lembi che, nel rispetto delle norme che tutelano la riservatezza, sono stati staccati dagli altri fogli del questionario e trattenuti presso i Comuni durante le operazioni censuarie, sono stati fotocopiati e inviati all'Istat per la registrazione su supporto magnetico.

Il ricongiungimento dei record contenenti le informazioni desunte dai lembi con i rispettivi record contenenti le altre informazioni provenienti dai questionari di Censimento è avvenuto - prima dell'abbinamento con i dati dell'Indagine di copertura - utilizzando il codice a barre riportato sia sul lembo che su ogni pagina del corrispondente questionario.

La base informativa su cui ha operato il processo di integrazione dei dati è costituita da tre archivi diversi, rispettivamente costituiti da:

1. i record dell'Idc, contenenti, oltre alle informazioni utili per la stima del grado di copertura, anche i dati identificativi[25] individuali che sono serviti per effettuare l'abbinamento con i record di censimento;
2. i record relativi ai lembi dei questionari di censimento sui quali sono stati riportati i dati identificativi[26] degli individui censiti. Come già osservato, tali lembi, staccati dai corrispondenti questionari per essere utilizzati dai Comuni nell'operazione di adeguamento dei registri anagrafici, sono stati recuperati e inviati in fotocopia all'Istat per le sole sezioni campione;
3. i record relativi ai questionari di censimento privati dei lembi, per le sole sezioni campione.

Gli archivi 1 e 2 sono stati ottenuti mediante registrazione manuale delle informazioni su supporto magnetico. I dati dell'archivio 3 sono stati acquisiti tramite lettura ottica; sono affluiti dall'ufficio preposto alla loro gestione solo dopo aver subìto il processo di revisione finalizzato alla definizione della popolazione legale e altri ulteriori controlli volti a garantire l'esatta attribuzione degli individui alle famiglie e alle sezioni campione di appartenenza.

Il trattamento preliminare che i dati negli archivi 1, 2 e 3 hanno subìto prima di essere sottoposti all'abbinamento è rappresentato schematicamente nel diagramma di flusso successivo (Figura 2.7).

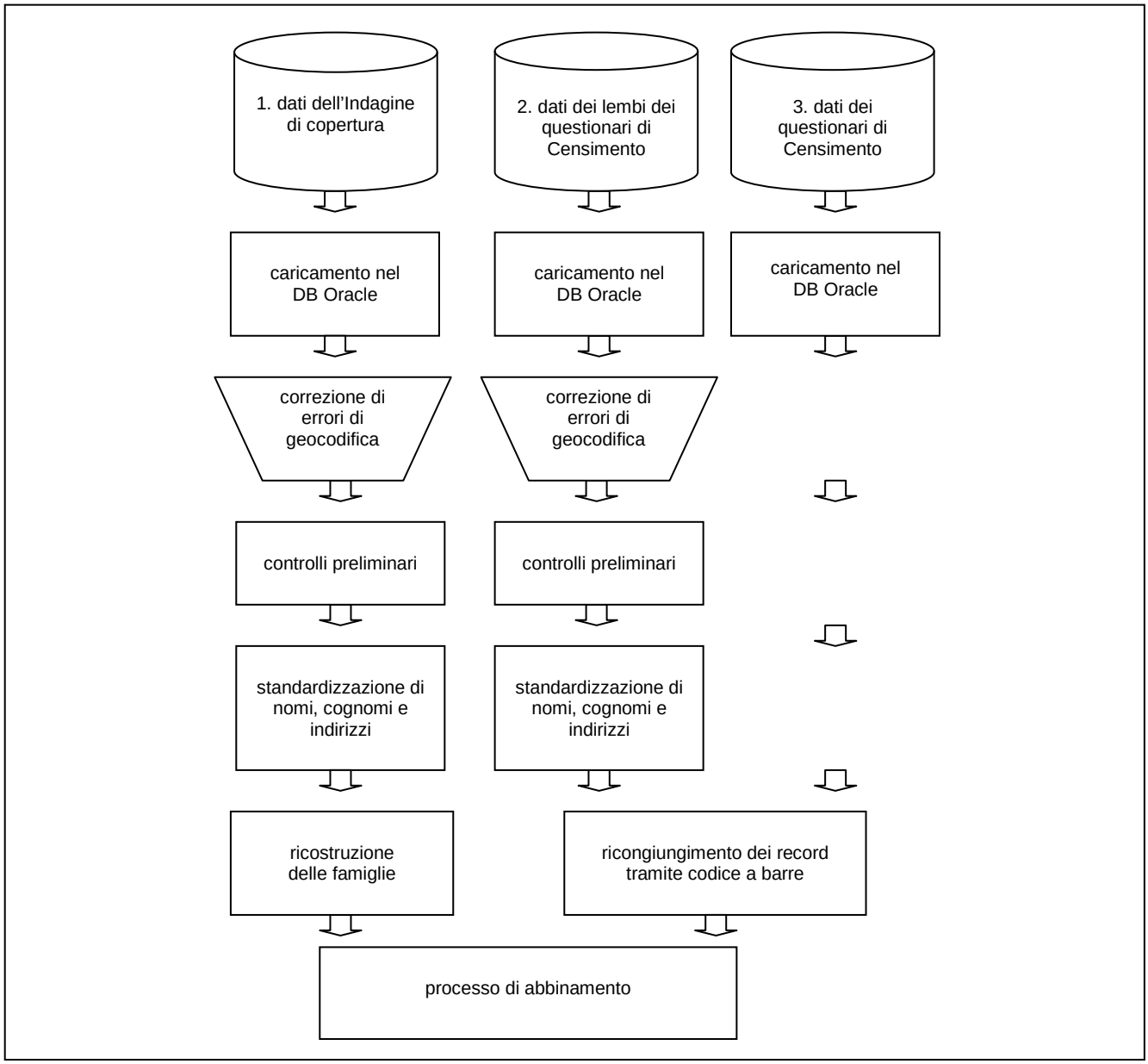
[23] Nell'abbinamento è possibile commettere i seguenti due tipi di errore:
- record che si riferiscono ad una stessa unità possono essere erroneamente non abbinati (mancanti abbinamenti);
- record che si riferiscono ad unità diverse possono essere erroneamente abbinati (falsi abbinamenti).

[24] Per lembo si intende il primo foglio del questionario di censimento.

[25] Dalla prima pagina del questionario dell'Indagine di copertura provengono: nome e cognome dell'intestatario e indirizzo della famiglia. Dalla lista (ovvero da pagina 3 del questionario dell'Indagine di copertura) provengono nome, cognome, sesso e data di nascita di ogni componente del nucleo familiare. Il sesso e la data di nascita di ogni persona sono presenti anche nel dettaglio (vale a dire all'interno del questionario dell'Indagine di copertura) insieme a tutte le altre notizie individuali, tra cui il luogo di nascita e lo stato civile.

[26] Dalla prima facciata del lembo provengono*:* nome, cognome dell'intestatario e indirizzo della famiglia. Nella seconda facciata del lembo sono presenti, per ogni componente del nucleo familiare, le stesse informazioni presenti nella lista dell'Indagine di copertura. Il sesso e la data di nascita di ogni persona sono riportati anche nel questionario di Censimento insieme ad altre notizie individuali, tra cui il luogo di nascita e lo stato civile.

Figura 2.7 - Trattamento dei dati preliminare al processo di abbinamento



Sia il trattamento preliminare che il processo di integrazione sono stati effettuati per i dati di una sezione campione[27] alla volta. Eventuali errori di geocodifica, presenti negli archivi 1 e 2, sono stati evidenziati e risolti subito dopo il loro caricamento nella base di dati Oracle e prima di qualsiasi altra operazione.

Per ulteriori approfondimenti sui controlli e sulla standardizzazione delle informazioni in formato libero si vedano i sottoparagrafi 2.8.2 e 2.8.3. Il processo di integrazione dei dati è descritto nel sottoparagrafo 2.8.4; alcuni risultati relativi ai dati abbinati vengono riportati nel sottoparagrafo 2.8.5.

2.8.2 - I controlli preliminari

Per quanto riguarda l'archivio dell'Indagine di copertura, è necessario premettere che il questionario è stato predisposto per accogliere i dati di una famiglia composta da al più cinque persone; nel caso in cui una famiglia

[27] Una sezione di censimento poteva arrivare a comprendere circa 600 famiglie.

risultasse più numerosa, a questa sono stati consegnati, in fase di rilevazione, uno o più questionari aggiuntivi dello stesso tipo. Prima di procedere con l'abbinamento, è stato necessario quindi riconoscere tutte le persone appartenenti a uno stesso nucleo familiare (operazione di ricostruzione delle famiglie); questa operazione è stata effettuata utilizzando il numero d'ordine definitivo, assegnato dal rilevatore a ogni famiglia della sezione campione.

Le attività preliminari di controllo e correzione dei dati dell'Indagine di copertura sono state volte, pertanto, a garantire:

- l'attribuzione univoca di numeri d'ordine definitivi alle famiglie, quando mancanti o duplicati nell'ambito della stessa sezione campione;
- l'esatta attribuzione degli individui alle famiglie di appartenenza, soprattutto nel caso di famiglie costituite da più di cinque persone;
- il riconoscimento e l'eliminazione di eventuali record duplicati.

Inoltre, in questo stadio del trattamento dei dati dell'Idc, sono stati calcolati alcuni indicatori[28] utili a evidenziare la presenza di valori mancanti per ognuna delle variabili di abbinamento utilizzate successivamente ed eventuali incoerenze tra informazioni riportate nella lista e informazioni di dettaglio (ovvero interne al questionario).

Per quanto riguarda le informazioni desunte dai lembi, le attività preliminari di controllo e correzione hanno avuto lo scopo di garantire il corretto ricongiungimento automatico dei record dell'archivio 2 con quelli dell'archivio 3 mediante l'utilizzo del codice a barre, presente sia sul lembo che su ogni pagina del corrispondente questionario di censimento. Tali attività sono consistite principalmente nel riconoscimento e nella correzione di codici a barre duplicati, mancanti o con anomalie.

Nei casi rari in cui il ricongiungimento automatico è fallito a causa di incongruenze tra i codici a barre negli archivi 2 e 3, altre informazioni comuni, quali il codice dell'edificio, il numero d'ordine definitivo della famiglia al Censimento, la data di nascita e il sesso dei componenti, hanno garantito il riconoscimento e l'eventuale aggancio dei lembi con i rispettivi questionari di censimento.

Dopo i controlli preliminari si è proceduto con la standardizzazione delle informazioni in formato libero (le variabili nome, cognome e indirizzo presenti negli archivi 1 e 2), al fine di poterle utilizzare in modo efficiente nel processo di abbinamento.

2.8.3 - La standardizzazione delle informazioni in formato libero

La standardizzazione di stringhe quali nomi, cognomi e indirizzi, permette di ridurre notevolmente gli errori di abbinamento. Infatti, spesso tali campi, pur essendo relativi a una stessa unità statistica, possono non coincidere a causa della presenza di parole non significative, sigle e abbreviazioni (ad esempio, *sig.*, *dott.*, *prof.* per gli individui, *v.*, *p.zza*, *v.le* per gli indirizzi), banali errori di trascrizione e/o registrazione, o modi alternativi utilizzati per denominare una stessa entità. Per tale motivo, la standardizzazione di nomi, cognomi e indirizzi è un'elaborazione che ha preceduto il processo di integrazione dei dati.

Per la standardizzazione del cognome e del nome è stato adottato un algoritmo che eliminava i caratteri speciali e le parole di supporto e restituiva in output due stringhe di tre caratteri, una per il cognome e una per il nome, secondo le regole adottate per la costruzione del codice fiscale di un individuo. In presenza di caratteri non alfabetici nel cognome (o nel nome), al cognome standardizzato (o al nome standardizzato) veniva assegnato valore mancante.

La standardizzazione degli indirizzi è stata effettuata mediante l'utilizzo di un dizionario, appositamente costruito, comprendente tutte le strade facenti parte degli itinerari percorsi da ciascun rilevatore nelle sezioni campione. La procedura si componeva di una parte automatica e di una manuale. La parte automatica prendeva come input la stringa dell'indirizzo da standardizzare, la ricercava tra quelle presenti nel dizionario e, nel caso venisse trovata, ne restituiva la forma standard presente nel dizionario. Quando l'indirizzo non veniva

[28] Gli indicatori calcolati per i dati dell'archivio 1 riguardavano:
- il numero di questionari con nome o cognome dell'intestatario mancante;
- il numero di questionari con indirizzo mancante;
- il numero di persone risultanti dalla lista e dal dettaglio;
- il numero di persone con valori mancanti per ognuna delle variabili nome, cognome, sesso e data di nascita;
- il numero di persone con valore discordante nella lista e nel dettaglio per ognuna delle variabili sesso e data di nascita.

riconosciuto automaticamente, si effettuava un recupero manuale associando alla stringa da standardizzare la forma standard più simile tra quelle presenti nel dizionario; era presente anche una modalità apposita ("non riconosciuto", che equivaleva ad attribuire valore mancante all'indirizzo standardizzato) per etichettare, se esistenti, gli indirizzi che non era possibile standardizzare a causa della loro marcata dissimilarità con le forme standard presenti nel dizionario.

2.8.4 - Il processo di integrazione dei dati

Il processo di integrazione dei dati è stato avviato solo dopo aver eseguito le operazioni di ricostruzione delle famiglie dell'Idc e di ricongiungimento dei lembi con i rispettivi questionari di censimento, secondo i criteri menzionati in precedenza.

Vista la criticità del processo di integrazione e la dimensione degli archivi coinvolti (ognuno è costituito da circa 180 mila record individuali), sono state utilizzate procedure che abbinassero automaticamente la maggior parte dei record, in modo che i restanti casi, più ambigui, potessero essere risolti in maniera agevole mediante ispezione manuale.

Il problema di riconoscere in maniera esatta record relativi a una stessa unità è stato risolto in prima battuta mediante un'operazione automatica di abbinamento deterministico (o *merge*), che produceva abbinamenti ogni qualvolta il valore della chiave utilizzata fosse unico e coincidesse perfettamente nelle due occasioni, Censimento e Indagine di copertura.

Tuttavia, spesso, anche quando i record in due archivi si riferiscono a una stessa unità, la coincidenza dei valori delle variabili di abbinamento non si verifica; nel contesto specifico, ciò poteva avvenire a causa di:

- mancate risposte;
- variazioni avvenute durante il periodo di tempo trascorso tra le due rilevazioni;
- errori nella fase di rilevazione dei dati;
- errori nella fase di registrazione dei dati;
- errori nel trattamento preliminare che i dati hanno subìto prima di essere sottoposti all'abbinamento.

In questo caso, si è fatto ricorso a una procedura di *record linkage* basata su un metodo di tipo probabilistico. Il metodo probabilistico utilizzato è stato sviluppato secondo un'impostazione bayesiana del problema dell'abbinamento esatto:[29] l'informazione fornita dai dati osservati è stata combinata con quella a priori in modo da ottenere una distribuzione a posteriori sull'insieme di tutte le possibili coppie di record. Il metodo è stato implementato in un software che, nella sua versione attuale, esegue l'abbinamento esatto di record contenuti in due diverse tabelle di una base di dati Oracle residente su server Unix.[30]

Il processo di integrazione è stato complesso e articolato[31] e può essere schematizzato nei seguenti passi, il cui ordine di svolgimento ha ridotto sensibilmente la possibilità di commettere errori:

1) abbinamento deterministico di famiglie utilizzando la chiave costituita da variabili di abbinamento riguardanti la famiglia (indirizzo standardizzato e numero civico) e l'intestatario (nome e cognome);
2) abbinamento probabilistico di famiglie, non abbinate al passo precedente, utilizzando la chiave costituita da variabili di abbinamento riguardanti la famiglia (indirizzo standardizzato, numero civico, numero di maschi nella famiglia, numero di femmine nella famiglia) e l'intestatario (nome standardizzato, cognome standardizzato, sesso, giorno di nascita, mese di nascita, anno di nascita);
3) abbinamento deterministico di individui, appartenenti a famiglie abbinate al passo 1 o 2, utilizzando la chiave costituita da nome standardizzato, cognome standardizzato, sesso, anno di nascita oppure nome standardizzato, giorno di nascita, mese di nascita, anno di nascita;
4) abbinamento mediante ricerca manuale di individui, non abbinati al passo precedente e appartenenti a famiglie abbinate al passo 1 o 2, utilizzando informazioni relative a nome, cognome, sesso, giorno di nascita, mese di nascita, anno di nascita, luogo di nascita;

---

[29] Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001). "On Bayesian record linkage". *Research in Official Statistics*, 4: 185-198; Nuccitelli, A. *Integrazione di dati mediante tecniche di abbinamento esatto: una rassegna critica ed una proposta in ambito bayesiano*. Tesi di dottorato in "Metodi statistici per l'economia e l'impresa". Roma: Università degli Studi di Roma Tre, 2001.

[30] Nuccitelli, A., F. Bosio, e L. Fioriti. *L'applicazione Reclink per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*. Roma: Istat, 2004. (Documenti, n.10).

[31] Nuccitelli, A. "La strategia di abbinamento dei dati del 14° Censimento della popolazione con i dati dell'indagine di copertura". In *L'integrazione di dati di fonti diverse - Tecniche e applicazioni del record linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*, a cura di P.D. Falorsi, A. Pallara, A. Russo. 61-91. Milano: Franco Angeli, 2005.

5) abbinamento deterministico di individui, non abbinati ai passi 3 o 4, utilizzando la chiave costituita da indirizzo standardizzato, numero civico, nome standardizzato, cognome standardizzato, sesso, giorno di nascita, mese di nascita, anno di nascita;
6) abbinamento probabilistico di individui, non abbinati ai passi 3, 4 o 5, in base alla chiave costituita da indirizzo standardizzato, numero civico, nome standardizzato, cognome standardizzato, sesso, giorno di nascita, mese di nascita, anno di nascita, luogo di nascita, stato civile;
7) abbinamento mediante ricerca manuale di individui, non abbinati ai passi 3, 4, 5 o 6, utilizzando tutte le informazioni disponibili (anche su supporto cartaceo).

Nei passi 2 e 6 di *record linkage* probabilistico, l'esecuzione della procedura automatica è stata sempre seguita da un'ispezione manuale di un sottoinsieme delle coppie soluzione fornite in output, al fine di cautelarsi il più possibile dal rischio di commettere errori di abbinamento. Più precisamente, dal momento che per ogni coppia si disponeva della probabilità a posteriori che i record si riferissero alla stessa unità statistica, le coppie caratterizzate da una probabilità a posteriori relativamente bassa venivano sottoposte a revisione per essere confermate o meno, mediante il ricorso a informazioni supplementari.

Vista la complessità di tutto il processo di elaborazione dei dati, opportune interfacce utente sono state predisposte per agevolare il più possibile sia il trattamento preliminare dei dati (controlli, standardizzazione di nomi, cognomi e indirizzi, ricongiungimento dei lembi ai rispettivi questionari del censimento, ricostruzione delle famiglie dell'Indagine di copertura), che lo svolgimento di tutti i passi di abbinamento. Particolare attenzione è stata rivolta soprattutto alla predisposizione delle interfacce relative:

- all'immissione dei dati di input necessari per l'esecuzione del programma di *record linkage* probabilistico e alla revisione dell'output da questo generato;
- all'esecuzione dell'abbinamento mediante ricerca manuale.

Inoltre, per ogni sezione campione, il processo di integrazione è stato documentato da un report che consentiva di supervisionare i risultati ottenuti a ogni passo e valutare la qualità del lavoro svolto dalle persone che utilizzavano il software predisposto.

2.8.5 - Alcuni risultati del processo di integrazione dei dati

Nella tavola 2.10 è riportata la distribuzione di frequenza - espressa in valori percentuali - degli individui abbinati, per passo di abbinamento.

Come si può notare, poco più del 90 per cento degli abbinamenti a livello individuale è stato ottenuto automaticamente mediante un'operazione di *merge* (passi 3 e 5). Il ricorso ad altre modalità di riconoscimento delle unità ha permesso di contenere al massimo gli errori di abbinamento: il 7,83 per cento degli abbinamenti a livello individuale è stato effettuato mediante ricerca manuale (passi 4 e 7) e l'1,85 per cento utilizzando la procedura di *record linkage* probabilistico (passo 6). Inoltre, circa il 96 per cento degli individui è stato abbinato all'interno di famiglie abbinate (passi 3 e 4).

Tavola 2.10 - Distribuzione di frequenza degli individui abbinati per passo di abbinamento *(valori percentuali)*

PASSO DI ABBINAMENTO	Individui abbinati
3) Abbinamento deterministico di individui appartenenti a famiglie abbinate al passo 1) o 2)	89,01
4) Abbinamento mediante ricerca manuale di individui non abbinati al passo 3) e appartenenti a famiglie abbinate al passo 1) o 2)	6,97
5) Abbinamento deterministico di individui non abbinati ai passi 3) o 4)	1,31
6) Abbinamento probabilistico di individui non abbinati ai passi 3), 4) o 5)	1,85
7) Abbinamento mediante ricerca manuale di individui non abbinati ai passi 3), 4), 5) o 6)	0,86
Totale: 3) + 4) + 5) + 6) + 7)	**100,00**

Nella tavola 2.11 è riportata la distribuzione di frequenza - espressa in valori percentuali - degli individui abbinati, per numero di valori concordanti di dieci variabili di abbinamento[32] all'Indagine di copertura e al Censimento.

La percentuale di abbinamenti per cui tutte o quasi tutte le variabili considerate presentano valori concordanti - per il 96 per cento degli individui abbinati si verificano almeno otto concordanze - indica l'elevata qualità di tutto il processo di integrazione. Inoltre, solo per lo 0,04 per cento degli abbinamenti si osservano meno di quattro concordanze.

Tavola 2.11 - Distribuzione di frequenza degli individui abbinati, per numero di valori concordanti delle variabili di abbinamento all'Indagine di copertura e al Censimento *(valori percentuali)* (a)

NUMERO DI VALORI CONCORDANTI DELLE VARIABILI DI ABBINAMENTO ALL'INDAGINE DI COPERTURA E AL CENSIMENTO	Individui abbinati
10	59,90
9	29,43
8	6,67
7	1,16
6	1,84
5	0,81
4	0,14
Meno di 4	0,04
Totale	**100,00**

(a) Gli arrotondamenti delle cifre sono stati effettuati direttamente dal software che elabora i dati, pertanto non sempre si trova realizzata la quadratura verticale e/o orizzontale.

Scendendo nel dettaglio delle singole variabili di abbinamento, nella tavola 2.12 è riportata la distribuzione di frequenza degli individui abbinati secondo il tipo di relazione tra i valori che ciascuna variabile assume all'Idc e al Censimento (concordanza, discordanza, presenza di valore mancante all'Idc e non al Censimento, presenza di valore mancante al Censimento e non all'Idc, valore mancante sia all'Idc che al Censimento).

In aggiunta alle dieci variabili considerate per la determinazione dei valori contenuti nella tavola 2.11, nella tavola 2.12 sono presi in esame anche indirizzo, nome, cognome (anche se l'indirizzo in forma non standardizzata non è stato mai utilizzato come variabile di abbinamento).

Come si può vedere, la variabile sesso presenta il livello più elevato di concordanza (99,33 per cento). Al contrario, per la variabile indirizzo il grado di concordanza è minimo (48,85 per cento); tuttavia, la standardizzazione, effettuata secondo le modalità descritte, ha permesso di recuperare buona parte delle discordanze (la percentuale di valori concordanti passa dal 48,85 per cento al 94,77 per cento).

Inoltre, la minore percentuale di valori mancanti all'Indagine di copertura rispetto al Censimento - riscontrabile per gran parte delle variabili di abbinamento - è indicativa del maggiore grado di accuratezza raggiunto nella fase di rilevazione sul campo dell'indagine post-censuaria.

[32] Le variabili di abbinamento considerate per la determinazione dei valori contenuti nella tavola 2.11 sono le seguenti: indirizzo standardizzato, numero civico, nome standardizzato, cognome standardizzato, sesso, giorno di nascita, mese di nascita, anno di nascita, luogo di nascita, stato civile.

Tavola 2.12 - Distribuzione di frequenza degli individui abbinati secondo la relazione tra i valori che ciascuna variabile di abbinamento assume all'Indagine di copertura e al Censimento *(valori percentuali)* (a)

RELAZIONE TRA I VALORI DELLA VARIABILE DI ABBINAMENTO ALL'INDAGINE DI COPERTURA E AL CENSIMENTO	Individui abbinati
INDIRIZZO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	48,85
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	48,62
Valore mancante all'indagine di copertura e non mancante al Censimento	0,03
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
INDIRIZZO STANDARDIZZATO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	94,77
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	2,69
Valore mancante all'indagine di copertura e non mancante al Censimento	0,04
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
NUMERO CIVICO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	74,55
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	10,19
Valore mancante all'indagine di copertura e non mancante al Censimento	3,96
Valore non mancante all'indagine di copertura e mancante al Censimento	5,40
Valori mancanti all'indagine di copertura e al Censimento	5,90
Totale	**100,00**
NOME	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	91,91
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	5,49
Valore mancante all'indagine di copertura e non mancante al Censimento	0,11
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
NOME STANDARDIZZATO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	94,22
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	3,16
Valore mancante all'indagine di copertura e non mancante al Censimento	0,12
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,01
Totale	**100,00**
COGNOME	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	91,25
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	6,15
Valore mancante all'indagine di copertura e non mancante al Censimento	0,10
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,01
Totale	**100,00**

Tavola 2.12 segue **- Distribuzione di frequenza degli individui abbinati secondo la relazione tra i valori che ciascuna variabile di abbinamento assume all'Indagine di copertura e al Censimento** *(valori percentuali)* (a)

RELAZIONE TRA I VALORI DELLA VARIABILE DI ABBINAMENTO ALL'INDAGINE DI COPERTURA E AL CENSIMENTO	Individui abbinati
COGNOME STANDARDIZZATO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	94,92
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	2,45
Valore mancante all'indagine di copertura e non mancante al Censimento	0,12
Valore non mancante all'indagine di copertura e mancante al Censimento	2,50
Valori mancanti all'indagine di copertura e al Censimento	0,01
Totale	**100,00**
SESSO	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	99,33
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	0,50
Valore mancante all'indagine di copertura e non mancante al Censimento	0,08
Valore non mancante all'indagine di copertura e mancante al Censimento	0,08
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
GIORNO DI NASCITA	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	98,01
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	1,80
Valore mancante all'indagine di copertura e non mancante al Censimento	0,15
Valore non mancante all'indagine di copertura e mancante al Censimento	0,03
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
MESE DI NASCITA	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	98,75
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	1,09
Valore mancante all'indagine di copertura e non mancante al Censimento	0,14
Valore non mancante all'indagine di copertura e mancante al Censimento	0,02
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
ANNO DI NASCITA	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	98,40
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	1,45
Valore mancante all'indagine di copertura e non mancante al Censimento	0,13
Valore non mancante all'indagine di copertura e mancante al Censimento	0,02
Valori mancanti all'indagine di copertura e al Censimento	0,00
Totale	**100,00**
LUOGO DI NASCITA	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	91,56
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	5,53
Valore mancante all'indagine di copertura e non mancante al Censimento	0,63
Valore non mancante all'indagine di copertura e mancante al Censimento	2,22
Valori mancanti all'indagine di copertura e al Censimento	0,06
Totale	**100,00**

Tavola 2.12 segue **- Distribuzione di frequenza degli individui abbinati secondo la relazione tra i valori che ciascuna variabile di abbinamento assume all'Indagine di copertura e al Censimento** *(valori percentuali)* (a)

RELAZIONE TRA I VALORI DELLA VARIABILE DI ABBINAMENTO ALL'INDAGINE DI COPERTURA E AL CENSIMENTO	Individui abbinati
STATO CIVILE	
Valori non mancanti e concordanti all'indagine di copertura e al Censimento	96,65
Valori non mancanti e discordanti all'indagine di copertura e al Censimento	1,80
Valore mancante all'indagine di copertura e non mancante al Censimento	0,73
Valore non mancante all'indagine di copertura e mancante al Censimento	0,80
Valori mancanti all'indagine di copertura e al Censimento	0,01
Totale	**100,00**

(a) Gli arrotondamenti delle cifre sono stati effettuati direttamente dal software che elabora i dati, pertanto non sempre si trova realizzata la quadratura verticale e/o orizzontale.

2.9 - La revisione e correzione dei record individuali

Dopo aver realizzato l'abbinamento tra i record del Censimento e i record dell'Indagine di copertura è stato necessario utilizzare procedure di controllo e correzione dei microdati dell'indagine per alcune variabili di interesse. L'obiettivo era quello di stabilire valori validi per ogni individuo, eliminando gli errori e le incongruenze al fine di non compromettere la correttezza e l'attendibilità delle stime di copertura. Le suddette stime sono fornite, infatti, per alcune caratteristiche sociodemografiche che sono, nello specifico: età, sesso, stato civile, cittadinanza, titolo di studio, condizione professionale e posizione professionale. Tali variabili sono diffuse con un diverso livello di dettaglio territoriale, secondo quanto stabilito nel disegno e nel piano di indagine.

Nella revisione dei microdati si è attribuito un ordine gerarchico alle variabili, per tener conto dell'impatto che l'avvenuta correzione di un campo poteva avere sul valore di un altro a essa legato.

Si è corretto, in primo luogo, l'anno di nascita per due principali motivi:

- in funzione della data di nascita si stabilisce uno dei criteri di eleggibilità dell'unità al campione dell'indagine. Si ricorda infatti che sono eleggibili per l'indagine tutti e solo quegli individui che (a) sono nati prima del 21 ottobre 2001 e che (b) non sono risultati abbinati con un individuo non abitualmente dimorante (Nad) del Censimento;
- in base all'età calcolata sulla data di nascita che si deve o meno fornire una risposta in merito al titolo di studio (solo per chi ha 6 anni o più) e alla condizione professionale (solo per chi ha 15 anni o più).

Una volta revisionata, la variabile anno di nascita ha permesso di stabilire l'eleggibilità dell'unità per l'indagine.

Determinati i record eleggibili, se sono corrette e, in presenza di valori mancanti, imputate, le variabili di interesse con procedure sia di tipo deterministico che probabilistico, tenendo conto della diversa natura delle variabili esaminate, in modo tale che ogni individuo dell'Indagine di copertura presentasse valori compatibili per le sette variabili oggetto di stima.

2.9.1 - La variabile anno di nascita e la variabile sesso

Nella revisione della variabile anno di nascita, come anche per la variabile sesso, si è considerato, in primo luogo, il fatto che tale informazione, in caso di abbinamento copertura-censimento, doveva essere riportata per tre volte: sulla lista del modello di rilevazione dell'Idc, all'interno del questionario Idc vero e proprio e all'interno del corrispettivo questionario dell'indagine censuaria.

A differenza di quanto è stato fatto per le procedure di *record linkage*, nella revisione dei microdati si è considerata, dapprima, l'informazione riportata nell'interno del questionario Idc; questa, a patto che avesse un valore ammissibile, veniva considerata come esatta, anche se discordante da quella indicata nel questionario dell'indagine censuaria.

In primo luogo, quindi, sono stati analizzati i valori dei 173.143 individui eleggibili della copertura in base alla condizione (b) e sono stati azzerati i campi dei valori fuori dominio (12 casi).

In 171.701 record (il 99,2 per cento del totale) la variabile anno di nascita era dichiarata e presentava valori coerenti con le modalità indicabili nel questionario.

Per imputare il valore ai restanti record si sono seguiti i seguenti passi:

1. agli individui con valore mancante nell'interno del questionario ma con valore ammissibile nella lista veniva imputato il valore della variabile riportato sulla lista (1.228 casi);
2. agli individui che non presentavano né il valore dell'interno del questionario, né della lista, ma si erano abbinati e avevano un anno di nascita valorizzato e ammissibile nel questionario del Censimento, veniva imputato il corrispettivo valore (82 unità);
3. ai restanti individui (24 casi) veniva imputato casualmente un anno di nascita in modo che l'insieme dei record, imputati e non, riproducesse la distribuzione della variabile anno di nascita calcolata sui soli record che avevano tale informazione corretta o nella lista o all'interno del questionario.

Sono stati seguiti gli stessi passi (passi 1-3) anche per correggere e imputare il giorno e il mese di nascita; tali informazioni individuali erano necessarie per il calcolo dell'età in anni compiuti e per poter stabilire l'eleggibilità dell'unità per l'Indagine.

Dopo le procedure di controllo e correzione dell'anno di nascita sono state considerate eleggibili per l'indagine 173.109 unità poiché 34 avevano una data di nascita posteriore al 20 ottobre 2001.

Anche per la variabile sesso, riportata fino a tre volte nel record individuale, si sono seguiti i passi 1, 2 e 3 della procedura di correzione.

In questo caso sono stati corretti e imputati, in dettaglio:

- 70 record al passo 1;
- 1.333 al passo 2;
- 62 record al passo 3.

Si ritiene utile precisare che si è cercato di ricorrere il meno possibile alle informazioni riportate nel file del Censimento, perché i dati in esso contenuti non erano ancora stati corretti al momento della revisione dell'Indagine di copertura e inoltre l'informazione censuaria era disponibile solo nel caso dei record abbinati.

In generale, la procedura di imputazione per l'anno di nascita e per il sesso ha riguardato una parte marginale dei record (meno dell'1 per cento del totale dei record della copertura).

2.9.2 - La variabile stato civile

Corrette le variabili anno di nascita e sesso si è proseguito con il controllo e la correzione della variabile stato civile, per i soli record eleggibili per l'indagine, ossia 173.109 individui.

Lo stato civile era mancante, o aveva un valore non ammissibile, in 1.239 casi, lo 0,72 per cento del totale dei record di indagine.

In 404 casi si era in presenza di minori a cui è stata imputata con un criterio deterministico la modalità celibe/nubile.

In 501 casi lo stato civile era non dichiarato, ma era presente la variabile anno di matrimonio; a questi individui è stato imputato un valore dello stato civile in modo tale che si riproducesse (sull'insieme dei record non imputati e imputati) la distribuzione dello stato civile condizionatamente alla presenza della data di matrimonio sui soli individui che avevano dichiarato sia lo stato civile che la data di matrimonio.

Ai 334 casi rimanenti è stato imputato in modo probabilistico un valore ammissibile facendo in modo che si rispettasse la distribuzione dello stato civile calcolata sui restanti 172.775 record; è evidente che il valore imputato doveva essere coerente con le altre informazioni contenute nel record.

2.9.3 - La variabile cittadinanza

Un approccio diverso è stato adottato nel caso della variabile cittadinanza che, come atteso, presentava un gran numero di casi mancanti o valori non ammissibili. Dopo aver eliminato i "fuori campo", risultavano essere *missing* 7.568 record su 173.109 totali, oltre il 4 per cento. È importante precisare che si sono scelte regole conservative, che tendessero a imputare il valore "straniero" solo nei casi in cui si avesse la quasi certezza della cittadinanza stranera; questo per non introdurre distorsioni nella stima del tasso di copertura, infatti è probabile che gli stranieri abbiano una maggiore probabilità di sfuggire alla conta censuaria.

Per poter attribuire la cittadinanza ai 7.568 individui si sono seguite diverse procedure di imputazione cercando, in prima istanza di dedurre l'informazione mancante dalle altre variabili contenute nel record individuale. Il modello di rilevazione dell'Idc richiedeva, per coloro che avevano indicato di possedere la cittadinanza straniera, lo stato estero di cittadinanza, che andava riportato, per esteso, in formato di stringa alfabetica. Inoltre veniva richiesto a tutti il luogo di nascita, distinguendo se nel comune d'indagine o in un altro comune italiano (nel caso andava indicato quale fosse), oppure all'estero (nel caso andava indicato quale fosse lo stato estero di nascita).

La variabile stato estero di cittadinanza, se presente e ammissibile, ha permesso di risalire alla cittadinanza qualora non fosse stata indicata. In 16 casi si è recuperato il valore della cittadinanza in questo modo. Si è poi fatto ricorso alla variabile stato estero di nascita; questa variabile era valorizzata con valori ammissibili in 114 casi. Si è usato il paese di nascita come *proxy* della cittadinanza, considerando italiani i cittadini provenienti dai tradizionali paesi di immigrazione o dai paesi a sviluppo avanzato limitrofi all'Italia (53 casi) mentre negli altri casi (61) si è imputato il valore "straniero".

I restanti 7.438 casi non avevano informazioni individuali che potessero far risalire alla cittadinanza essendo mancanti anche le altre variabili a essa collegate.

Si sono considerate, allora, le famiglie a cui gli individui appartenevano. Per le famiglie con più componenti si sono considerate le cittadinanze dichiarate all'interno della famiglia e si è imputato il valore prevalente nella famiglia nel caso in cui almeno la metà dei componenti della stessa ne avesse dichiarata una (2.603 casi); imputando, così, la cittadinanza straniera a poco più dell'1 per cento del totale dei record.

In tutti i restanti casi e per le famiglie unipersonali si è imputato il valore "italiano" alla variabile cittadinanza.

È importante sottolineare che per la correzione di questa variabile non si è fatto mai ricorso ai record "grezzi" del Censimento per evitare che si potesse imputare un valore non corretto.

2.9.4 - La correzione probabilistica con Scia: il titolo di studio, la condizione occupazionale e la posizione nella professione

Per la correzione e imputazione delle variabili titolo di studio, condizione occupazionale e posizione nella professione si è usato principalmente il modulo Scia (Sistema di controllo e imputazione automatica) che implementa la metodologia di Fellegi-Holt nel caso di variabili categoriche. Le variabili in questione sono legate poiché la posizione nella professione deve essere valorizzata solo da chi si è dichiarato occupato; inoltre solo alcuni titoli di studio permettono di accedere a talune professioni.

Prima delle imputazioni con Scia si sono fatte alcune correzioni preliminari, mentre altre procedure di correzione sono state implementate a posteriori.

Con l'impiego di Scia i valori da attribuire ai dati mancanti o incongruenti sono desunti dalle informazioni esatte presenti in altri dati, "somiglianti" quanto più possibile in termini di caratteristiche possedute ai dati errati, nel rispetto di certe regole di compatibilità formulate a priori.

Sulla base di queste regole, Scia individua e separa due insiemi di dati, errati ed esatti a seconda che violino o meno le regole, e utilizza i secondi come donatori di informazione ai primi.

Sono stati fatti diversi tentativi prima di giungere al *set* finale di regole da inserire nel sistema che hanno coinvolto, oltre alle tre variabili di interesse, anche le variabili: "durata del corso di studi", "ha svolto attività lavorativa retribuita nel corso della sua vita?", "ha dipendenti retribuiti?" e soprattutto una variabile (CLANAS) ottenuta come classificazione della variabile "anno di nascita"; le regole di compatibilità tenevano conto anche dell'analisi del questionario e di quelle che erano state le norme adottate nel caso delle procedure di correzione dei dati del Censimento.

La variabile CLANAS è stata assunta come variabile chiave (o di strato) affinché nella fase di correzione il sistema rispettasse la distribuzione per età della popolazione, evitando in tal modo di generare ad esempio "studenti troppo anziani" e "pensionati troppo giovani".

Sulla base delle regole stabilite, dopo le diverse simulazioni del processo di imputazione, il *set* di regole adottato ha prodotto, su un totale di 173.109 record riferiti alla totalità di unità eleggibili, 13.463 record classificati come errati e 159.646 come esatti.

Le modifiche introdotte nei dati, alla fine della correzione, sono riassunte nella tavola seguente (Tavola 2.13).

Tavola 2.13 - Totali valori imputati, modificati e sbiancati con la procedura di correzione probabilistica

VARIABILI	Valori *blank* imputati	Valori modificati	Valori sbiancati
Condizione lavorativa	831	3	6.590
Titolo di studio	2.116	484	443
Posizione lavorativa	747	632	1

Dopo la correzione con Scia si è intervenuti per correggere alcuni casi (181) di imputazioni del titolo di studio ritenute poco verosimili (in quanto troppo basso), partendo dalle distribuzioni età e titolo di studio dei record esatti in partenza.

Inoltre, dato che il software in presenza di titolo di studio non elevato e posizione nella professione "libero professionista o imprenditore", modificava spesso il titolo di studio (assumendo che per essere "libero professionista o imprenditore" si deve essere in possesso almeno della laurea) si è applicata una regola deterministica per cui, qualora il titolo di studio fosse non elevato e la posizione nella professione nelle modalità "imprenditore" o "libero professionista", la posizione nella professione doveva essere modificata in "lavoratore in proprio", senza cambiare titolo di studio.

2.10 - Il procedimento di stima

2.10.1 - I pesi di riporto all'universo

I pesi sono utilizzati nel procedimento di stima e sono stati applicati agli individui e alle famiglie osservati nell'indagine. Gli individui e le famiglie ereditano il peso della sezione e del comune a cui appartengono.

La determinazione dei pesi delle unità finali dell'indagine procede dal calcolo dei pesi iniziali (pesi diretti) e prosegue, tramite il procedimento di riponderazione, nella determinazione dei pesi finali (pesi calibrati); quest'ultima operazione ha lo scopo di migliorare la rappresentatività del campione e migliorare la precisione delle stime.

I pesi diretti dipendono dal disegno campionario adottato e sono riferiti alle unità di primo stadio (comuni campione) e alle unità di secondo stadio (sezioni campione).

I pesi delle unità di primo stadio (comuni) sono direttamente proporzionali all'ammontare di popolazione di tutti i Comuni dello strato di appartenenza del comune campione e inversamente proporzionali alla dimensione demografica del comune campione e alla dimensione del sub-campione estratto dallo strato (=1 o =2 a seconda che sia costituito da comuni auto rappresentativi o non auto rappresentativi); indicato con P_{hj} la popolazione residente nel comune *j* dello strato *h*, il peso del comune è dato da:

$$W_{hj1} = \frac{\sum_{j \in h} P_{hj}}{P_j} \frac{1}{n_h} \qquad (2.10.1)$$

dove $n_h = 1$ se il comune è Ar, $n_h = 2$ se il comune è Nar.

I pesi delle unità di secondo stadio (sezioni) sono calcolati a partire dalle informazioni fornite dalle Basi territoriali del Censimento 2001 e sono pari all'inverso delle probabilità di inclusione del primo ordine calcolate per il disegno semplice senza ripetizione; per ciascuna tipologia di sezione (centro; nucleo; case sparse/località produttiva) sono direttamente proporzionali al numero complessivo di sezioni di quella tipologia presenti nel dato comune campione e inversamente proporzionali al numero di sezioni campione estratte da quel comune, sempre per la stessa tipologia; indicato con $S_{hj}^{(g)}$ il numero totale di sezioni di censimento del comune *j* dello strato *h* che appartengono alla tipologia *(g)*, e con $s_{hj}^{(g)}$ il corrispondente numero di sezioni campione, il peso in questione è dato da:

$$W_{hj2}^{(g)} = \frac{S_{hj}^{(g)}}{s_{hj}^{(g)}} \qquad (2.10.2)$$

Quindi, il peso diretto o peso da disegno della generica sezione campione, di tipo *(g)*, appartenente al comune campione *j*, dello strato *h*, è pari al prodotto dei sopra specificati pesi:

$$W_{hj}^{(g)} = W_{hj1} W_{hj2}^{(g)} \qquad (2.10.3)$$

Al vettore dei pesi diretti viene applicata una procedura di riponderazione per ottenere i pesi finali che corregge il peso originario da disegno in funzione di esigenze e vincoli provenienti da fonti esterne.

Il procedimento adottato è contraddistinto dalle seguenti due caratteristiche:

- pesi calibrati per il campione di dati del Censimento 2001;
- doppio passo di calibrazione.

Riguardo al primo punto, la riponderazione è stata fatta in modo tale che, applicando i pesi finali ai dati del Censimento relativi alle sezioni campione dell'Indagine di copertura, si potessero ottenere stime di prefissati totali del Censimento 2001, tali da riprodurre il più possibile i corrispondenti valori osservati sull'intera popolazione.

Per quanto concerne il secondo punto, sono stati eseguiti due passi di calibrazione, che di seguito vengono illustrati e opportunamente descritti negli schemi allegati.

Nel primo passo (Figura 2.8) è stata fatta una correzione in base alla ripartizione della popolazione censita al 2001 residente nelle sezioni di centro e quella residua (concentrata nelle sezioni di nucleo, case sparse e località produttive), per ciascuno dei 56 strati definiti dal disegno campionario (2 totali noti per 56 domini). La scelta in tal senso è stata dettata sia per dare maggiore rappresentatività al dato di popolazione che per ridurre i possibili effetti distorsivi riconducibili al mancato aggiornamento delle basi territoriali 2001, al momento di estrazione del campione di sezioni.

Si fa inoltre presente che, in questa fase, l'algoritmo di calibrazione, pur non garantendo convergenza completa in tutti i domini, ha fornito risultati soddisfacenti in quanto, laddove non c'è stata convergenza, la distanza tra i totali noti e i valori stimati è risultata trascurabile.

Nel secondo passo (Figura 2.9) è stato fatto un duplice aggiustamento per far coincidere sia la distribuzione per sesso ed età (13 classi di età) degli individui censiti nelle sezioni campione dell'Idc con quella del Censimento della popolazione 2001 che la distribuzione per sesso dei cittadini stranieri residenti con i corrispondenti totali osservati al Censimento 2001, per ripartizione territoriale (28 totali noti x 5 domini).

Il risultato del doppio passo di calibrazione porta al sistema di pesi finale che di seguito verrà utilmente impiegato nell'ambito di una strategia di stimatori vincolati.

Figura 2.8 - Primo passo di ponderazione

PRIMO PASSO DI PONDERAZIONE	Censimento della popolazione 2001. Dati aggregati per sezione di censimento	
	Popolazione residente	
Dominio *(56 Strati definiti dal disegno campionario)*	Popolazione residente nelle sezioni di Centro	Popolazione residente nelle sezioni di Nucleo, Case Sparse e Località Produttive
111		
112		
113		
…		
542		
543		

Figura 2.9 - Secondo passo di ponderazione

SECONDO PASSO DI PONDERAZIONE	Censimento della popolazione 2001																											
	Popolazione residente per sesso e classi di età																										Popolazione residente straniera	
Dominio *(Ripartizioni Territoriali)*	PopRes_0-5_M	PopRes_6-13_M	PopRes_14-17_M	PopRes_18-19_M	PopRes_20-24_M	PopRes_25-29_M	PopRes_30-34_M	PopRes_35-44_M	PopRes_45-54_M	PopRes_55-64_M	PopRes_65-74_M	PopRes_75-84_M	PopRes_85+_M	PopRes_0-5_F	PopRes_6-13_F	PopRes_14-17_F	PopRes_18-19_F	PopRes_20-24_F	PopRes_25-29_F	PopRes_30-34_F	PopRes_35-44_F	PopRes_45-54_F	PopRes_55-64_F	PopRes_65-74_F	PopRes_75-84_F	PopRes_85+_F	Stranieri_M	Stranieri_F
Nord-ovest																												
Nord-est																												
Centro																												
Sud																												
Isole																												

2.10.2 - Lo stimatore della copertura per gli individui

Mediante l'operazione di *record linkage* si riesce ad associare a ogni individuo rilevato in occasione dell'Idc l'etichetta di "censito" o di "non censito" da utilizzare nel conteggio dell'ammontare (x_{11}) degli individui enumerati anche in occasione del Censimento. Tale quantità, insieme al totale degli individui censiti ($x_{1\cdot}$) e al totale degli individui intervistati dall'Idc ($x_{\cdot 1}$), costituisce la base per l'applicazione di un modello di tipo cattura-ricattura finalizzato alla stima dell'ammontare ignoto della popolazione $\hat{N}$ mediante il modello di Petersen[33].

Nel caso dell'Idc, essendo $x_{\cdot 1}$ e x_{11} ottenuti da una stima su un campione di sezioni[34], tali quantità saranno nel seguito più correttamente indicate con $\hat{x}_{\cdot 1}$ e $\hat{x}_{11}$ rispettivamente.

Inoltre, una delle condizioni affinché la stima $\hat{N}$ della popolazione ignota risulti non distorta, impone che siano valide le relazioni $p_{1\cdot}=p_{i1\cdot}$ e $p_{\cdot 1}=p_{i\cdot 1}$, $\forall$ i, cioè che la probabilità di cattura sia costante in ogni individuo *i* in ciascuna delle due catture. Il modo più agevole di rispettare tale ipotesi è costituito dal ricorso a una stratificazione delle unità statistiche per un insieme di variabili scelte allo scopo di assicurare la suddetta omogeneità delle probabilità di cattura, e calcolare quindi la popolazione ignota $\hat{N}$ come somma delle popolazioni stimate per ciascuno strato, ovvero

$$\hat{N} = \sum_h \frac{\hat{x}_{h\cdot 1} x_{h1\cdot}}{\hat{x}_{h11}}, \qquad (2.10.4)$$

dove[35]

$$\hat{x}_{h\cdot 1} = \sum_i \omega_{hi} \pi_{hi} x_{hi\cdot 1},$$

$$\hat{x}_{h11} = \sum_i \omega_{hi} \pi_{hi} x_{hi11},$$

[33] Petersen, C. G. J. (1896). *The yearly immigration of young plaice into Limfjord from the German Sea*. Report of the Danish Biological Station 6:1Ð48.
[34] Wolter K. M., (1986); **"**Some Coverage Error Models for Census Data".
[35] Nella formalizzazione degli stimatori sono stati omessi gli indici riferibili al disegno campionario al fine di non appesantire la notazione. Si deve quindi intendere che i pesi ω_{hi} e π_{hi} contengano al loro interno tutte le informazioni riferibili alla complessità del disegno di campionamento.

e si indica con:

h indice dello strato di appartenenza;

i indice dell'individuo rilevato in occasione dell'Idc;

ω_{hi} pesi diretti di riporto all'universo per l'i-mo individuo nello strato h;

π_{hi} pesi di calibrazione per l'i-mo individuo nello strato h,

$x_{i.1}$ variabile indicatrice che assume valore 1 se l'i-mo individuo dichiara di essere residente nella sezione il 21 ottobre 2001; assume il valore 0 in tutti gli altri casi;

x_{i11} variabile indicatrice che assume valore 1 se l'i-mo individuo dichiara di essere residente nella sezione il 21 ottobre 2001 e risulta abbinato con un corrispondente record rilevato all'epoca del Censimento; assume il valore 0 in tutti gli altri casi.

Il tasso di copertura viene infine stimato come,

$$\hat{r} = \frac{x_{1\cdot} + R}{\hat{N}} \qquad (2.10.5)$$

dove R rappresenta l'ammontare degli individui recuperati dal Censimento, ovvero quelli conteggiati dai comuni dopo la normale chiusura delle operazioni di rilevazione sul campo.

Per quanto riguarda i pesi diretti ω_{hi}, essi dipendono dalle probabilità di selezione determinate dal disegno di campionamento areale adottato. Tali probabilità di selezione derivano a loro volta dal prodotto fra il peso riferito alle unità di primo stadio (comuni), proporzionale alla dimensione demografica del comune campione rispetto alla popolazione totale del proprio strato di appartenenza, e quello riferito alle unità di secondo stadio (sezioni), stratificate secondo la loro tipologia ed estratte con campionamento casuale semplice tra quelle appartenenti al comune. È appena il caso di osservare che, dal momento che all'interno delle sezioni campione la rilevazione delle famiglie e degli individui è esaustiva, tutte le famiglie e gli individui di una data sezione ereditano il peso di riporto all'universo della propria sezione di appartenenza.

La presenza dei pesi π_{hi} è invece finalizzata alla calibrazione delle stime mediante vincoli del tipo $\hat{x}_{h1\cdot}=x_{h1\cdot}$ tali da assicurare che le stime $\hat{x}_{h1\cdot}$ della popolazione censita, ottenute a partire dai soli record di censimento riferiti alle sezioni campione dell'Idc, corrispondano ai totali di censimento. Questi sono calcolati come

$$\pi_h = \frac{x_{h1\cdot}}{\sum_i \omega_{hi} x_{hi1\cdot}} \qquad (2.10.6)$$

dove

$x_{hi1\cdot}$ è la variabile indicatrice che assume valore 1 se l'i-mo individuo dichiara di essere residente nella sezione il 21 ottobre 2001; assume il valore 0 in tutti gli altri casi.

2.10.3 - Lo stimatore della copertura per le famiglie

A partire dalle informazioni relative al risultato dell'abbinamento degli individui Idc con quelli del Censimento, e attraverso il raggruppamento degli individui in famiglie, è possibile fornire stime di copertura anche per queste ultime.

Se si assume che una famiglia Idc deve considerarsi censita quando almeno un suo componente è stato censito, allora un tasso di copertura per le famiglie può essere stimato come,

$$\hat{r}_{fam} = \frac{{}_{fam}x_{1\cdot}}{\sum_h \frac{{}_{fam}x_{h1\cdot}\ {}_{fam}\hat{x}_{h\cdot 1}}{{}_{fam}\hat{x}_{h11}}} \qquad (2.10.7)$$

dove le quantità riportate sopra, riferite alle famiglie, possono essere definite e stimate in modo analogo a quanto fatto per gli individui.

Tale stimatore è però affetto da una serie di potenziali problemi dipendenti dall'incidenza degli errori di rilevazione nei dati. In particolare, preoccupa la possibilità che, contravvenendo alle definizioni fra loro omogenee fornite in occasione del Censimento e dell'Idc, alcune famiglie composte da più nuclei tra loro imparentati, in una delle due rilevazioni si siano dichiarate come una sola famiglia, mentre nell'altra rilevazione

abbiano erroneamente riportato ciascun nucleo come una singola famiglia. Come esempio di tale fenomeno può essere considerato il caso della convivenza sotto lo stesso tetto di due coppie sposate e fra loro apparentate, come una coppia di genitori che convivono con il loro figlio e la sua sposa sotto lo stesso tetto. In base alle istruzioni fornite al Censimento e all'Idc tale insieme di persone costituisce una sola famiglia, ma sussiste il rischio che in un delle due rilevazioni si possano essere registrati come due famiglie coabitanti.

La presenza di questo fenomeno determina la violazione di una delle ipotesi fondamentali per il modello cattura-ricattura, in quanto corrisponde a considerare alcune unità non eleggibili nel computo totale delle famiglie censite e/o in quelli delle famiglie Idc. In altre parole la presenza di tale fenomeno porterebbe a una sottostima del tasso di copertura per la presenza di errori di sovracopertura in almeno una delle due occasioni di rilevazione.

Per ovviare a questo problema si è ricorso a uno stimatore del tasso di copertura più semplice, dato da

$$\hat{r}'_{fam} = \frac{{}_{fam}\hat{x}_{11}}{{}_{fam}\hat{x}_{\cdot 1}} \qquad (2.10.8)$$

dove:

${}_{fam}\hat{x}_{\cdot 1}$ è il totale delle famiglie Idc riportate all'universo;

${}_{fam}\hat{x}_{11}$ è il totale delle famiglie Idc, riportate all'universo, per le quali almeno un componente risulta abbinato.

Purtroppo, questo stimatore non può essere convenientemente stratificato per assicurare che le famiglie siano omogenee rispetto alla loro probabilità di cattura e quindi è più facile possa risultare distorto per qualche dominio di interesse. In particolare, essendo sicuramente le probabilità di cattura fortemente dipendenti dalla dimensione della famiglia, i tassi di copertura sono prodotti condizionatamente alla dimensione delle famiglie. Tuttavia i confronti, operati tra lo stimatore del tasso di copertura individuale che utilizza i post-strati con quello che non li utilizza, hanno mostrato discrepanze al di sotto dei due decimi di punto percentuale e questo sembra testimoniare a favore di una buona accuratezza delle stime sulle famiglie, prodotte con la versione semplificata dello stimatore.

2.10.4 - Lo stimatore della sovracopertura per gli individui

Fra le innovazioni introdotte dall'attuale Idc, rispetto a quelle svolte nel 1981 e 1991, c'è quella della stima dell'entità della sovracopertura. Tale fenomeno è stato indagato sottoponendo una serie di domande sul questionario Idc a un sottoinsieme degli individui partecipanti all'indagine di controllo. In tale sottoinsieme sono stati inclusi tutti coloro i quali, avendo già dichiarato di essere già residenti nell'alloggio alla data di censimento, hanno successivamente dichiarato di aver vissuto anche in uno o più altri alloggi durante i 365 giorni precedenti al censimento.

Le domande poste ai rispondenti eleggibili per l'analisi della sovracopertura miravano quindi a scoprire se essi, nel corso dell'anno precedente, ricordassero di essere stati censiti fra gli abitualmente dimoranti, anche nell'abitazione utilizzata come residenza occasionale, determinando in tal caso un errore di sovracopertura.

Attraverso tali domande il tasso di sovracopertura sugli individui può essere stimato dal rapporto:

$$\hat{r}_{sov} = \frac{\hat{x}_{sov}}{\hat{x}_{\cdot 1}} \qquad (2.10.9)$$

dove

$$\hat{x}_{sov} = \sum_i \pi_i \omega_i x_i^{(sov)}$$

e i pesi di riporto all'universo sono definiti come illustrato in precedenza, mentre $x_i^{(sov)}$ è una variabile indicatrice che assume valore 1 se l'i-mo individuo si dichiara residente nella sezione il 21 ottobre 2001 e dichiara inoltre di essere stato erroneamente censito come residente altrove; assume il valore 0 negli altri casi.

La stima della sovracopertura effettuata in questo modo permette per la prima volta di confrontare tale fenomeno con quello della sottocopertura, verificando come la sua portata sia effettivamente marginale. Inoltre, si tenta di analizzare in questo modo il problema di sovracopertura su sottodomini territoriali dovuti a errori di localizzazione dei residenti sul territorio.

Va da sé che, essendo l'informazione fondata sulla libera dichiarazione dei rispondenti, questa può essere affetta da problemi di comprensione, di memoria oppure dalla esplicita volontà di dichiarare il falso. Per questo motivo l'inserimento della batteria di domande rivolte alla stima del fenomeno della sovracopertura riveste un carattere sperimentale.

Infine, per come è stata costruita, l'informazione raccolta non mira all'identificazione degli eventi di sovracopertura determinati da individui presenti sul suolo nazionale ed enumerati per errore. Infatti si ricorda che vengono sottoposti alle domande sulla sovracopertura solo coloro i quali in precedenza dichiarano di possedere la residenza nell'abitazione contattata per l'Idc da prima della data del 21 ottobre 2001. La scelta di non indagare tale fenomeno è stata determinata da un giudizio sulla sua rarità in rapporto alla sua difficoltà di misurazione.

2.10.5 - La valutazione dell'errore campionario delle stime

La valutazione dell'efficienza campionaria delle stime prodotte dall'indagine di copertura è stata effettuata ricorrendo ad approssimazioni basate sulla linearizzazione degli stimatori utilizzati.

Al fine di illustrare la procedura utilizzata, ricordiamo brevemente le caratteristiche dello stimatore del tasso di copertura; richiamando la (2.10.5) lo stimatore del tasso di copertura è dato da:

$$\hat{r} = \frac{x_{1\cdot} + R}{\hat{N}},$$

in cui il denominatore, definito dalla (2.10.4), $\hat{N} = \sum_h \frac{\hat{x}_{h\cdot 1} x_{h1\cdot}}{\hat{x}_{h11}}$, è funzione delle quantità stimate $\hat{x}_{h\cdot 1} = \sum_i \omega_{hi}\pi_{hi}x_{hi\cdot 1}$ e $\hat{x}_{h11} = \sum_i \omega_{hi}\pi_{hi}x_{hi11}$.

Dal momento che è agevole mostrare che il coefficiente di variazione dello stimatore del tasso di copertura è approssimativamente uguale al coefficiente di variazione dello stimatore della numerosità di popolazione $\mathrm{CV}(\hat{r}) \approx \hat{N}\sqrt{\mathrm{VAR}(\hat{\mathrm{N}})/\hat{\mathrm{N}}^{\,4}} = \mathrm{CV}(\hat{N})$, abbiamo sviluppato una linearizzazione di quest'ultima quantità.

Ricordiamo che gli stimatori utilizzati nell'espressione della stima del tasso di copertura, $\hat{x}_{h\cdot 1}$ e $\hat{x}_{h11}$ (h=1…H) sono stimatori di calibrazione per i quali, al fine della valutazione della varianza, è utile applicare l'equivalenza asintotica agli stimatori di regressione generalizzata (GREG).

Si indichi con ${}_{\omega}\hat{Y}$ il generico stimatore espansione, ossia lo stimatore che utilizza i pesi iniziali pari al reciproco delle probabilità di inclusione, ${}_{\omega}\hat{Y} = \sum_i \omega_i y_i$, allora lo stimatore GREG può essere espresso come ${}_{G}\hat{Y} = {}_{\omega}\hat{Y} + (Z - {}_{\omega}\hat{Z})\hat{B}$ dove Z sono le informazioni ausiliarie utilizzate per la determinazione dei pesi finali, ${}_{\omega}\hat{Z}$ è lo stimatore di espansione di Z e $\hat{B} = \hat{T}_1^{-1}\hat{T}_2$ la stima campionaria del coefficiente di regressione con $\hat{T}_1 = \sum_{hk} z_{hk} z'_{hk} w_{hk}$ e $\hat{T}_2 = \sum_{hk} z_{hk} y_{hk} w_{hk}$, essendo z_{hk} il vettore delle variabili ausiliarie dello stimatore GREG sull'unità *k*. Nel presente caso le variabili ausiliarie sono rappresentate dai totali noti descritti nel precedente paragrafo 2.10.1.

Tenendo conto della approssimazione degli stimatori di calibrazione e della rappresentazione data dello stimatore GREG, è quindi possibile applicare la seguente approssimazione:

$$\sum_h \frac{{}_{G}\hat{x}_{h\cdot 1} x_{h1\cdot}}{{}_{G}\hat{x}_{h11}} = \sum_h \frac{{}_{\omega}\hat{x}_{h\cdot 1} + (Z - {}_{\omega}\hat{Z})\,{}_{\omega}\hat{B}_1}{{}_{\omega}\hat{x}_{h11} + (Z - {}_{\omega}\hat{Z})\,{}_{\omega}\hat{B}_2} x_{h1\cdot} \qquad (2.10.10)$$

con ${}_{\omega}\hat{B}_1$ e ${}_{\omega}\hat{B}_2$ stime dei coefficienti di regressione delle variabili ausiliarie rispettivamente per $x_{\cdot 1}$ e $x_{11\cdot}$.

Per la determinazione della varianza della precedente quantità (2.10.10), è possibile ricorrere al metodo di linearizzazione di Woodroof [36].

Sia $\hat{Y} = G(\hat{\mathbf{T}})$ uno stimatore del parametro $Y = G(\mathbf{T})$, in cui G è una funzione[37] non lineare, essendo $\hat{\mathbf{T}} = (\hat{T}_1,\ldots,\hat{T}_L)'$ e $\mathbf{T}=(T_1,\ldots,T_L)'$ due vettori L-dimensionali il cui generico elemento $\hat{T}_l$ è uno stimatore lineare non distorto del totale T_l (l=1,…,L). Il metodo di linearizzazione suddetto permette di stimare la varianza campionaria $V(\hat{Y})$ di $\hat{Y}$, attraverso una stima della varianza campionaria $V(\hat{Y}_e)$ dello stimatore linearizzato $\hat{Y}_e = \sum_{k\in s} \omega_k^{-1} e_k$, dove ω_k è la probabilità di inclusione dell'unità k e $e_k = \sum_{l=1}^{L} g_l(\mathbf{T})\,\theta_{lk}$ la trasformata di Woodruff, essendo θ_{lk} il valore assunto dalla variabile θ_l sull'unità k e $g_l(\mathbf{T}) = \left[\frac{\partial G(\mathbf{T})}{\partial T_l}\right]_{T=\hat{T}}$ la corrispondente derivata parziale.

L'applicazione del metodo alla (2.10.10) permette di ottenere la variabile linearizzata

$$E_{hi} = \frac{\hat{x}_{h11} E_{\cdot 1,hi} - \hat{x}_{\cdot 1} E_{11,hi}}{\hat{x}_{h11}^2} \qquad (2.10.11)$$

con $E_{\cdot 1,hi} = \left(x_{\cdot 1,hi} - \hat{B}_1 z_{hi}\right)$ e $E_{11,hi} = \left(x_{11,hi} - \hat{B}_2 z_{hi}\right)$, che risultano essere le trasformate di Woodruff per i rispettivi stimatori GREG dei totali considerati, $x_{h\cdot 1}$ e x_{h11} di cui la (2.10.10) è funzione.

Si noti che è necessario applicare la linearizzazione (2.10.11) distintamente per ciascuno dei domini di stima pianificati (quali Italia, ripartizione, dimensione comunale, intersezione di ripartizione e dimensione) essendo le quantità coinvolte funzione di tali domini. Valutata l'espressione (2.10.11) è possibile determinare la stima della varianza per mezzo del software *Genesees*.[38]

Per la valutazione della precisione delle stime del tasso di copertura per le sottoclassi S, (ossia quei domini di stima non pianificati, quali ad esempio il sesso, la classe di età o il titolo di studio) è invece sufficiente utilizzare le precedenti espressioni, avendo prima definito opportune variabili $x^*_{h\cdot 1i}$ e x^*_{h11i} pari a 0 al di fuori del dominio S e pari al valore di $x_{h\cdot 1i}$ e x_{h11i} all'interno del dominio.

Infine per la valutazione della varianza della stima del tasso di copertura relativo alle famiglie (2.10.8), così come per la valutazione della precisione del tasso di sovracopertura (2.10.9), la linearizzazione applicata è l'usuale linearizzazione per la stima di un rapporto; si noti comunque che, data la natura assolutamente stabile dei denominatori che compaiono nelle stime, i risultati non differiscono dalle stime prodotte utilizzando le varianze standard prodotte dal software *Genesees* sui rispettivi numeratori.

2.11 - I risultati dell'indagine

In questo paragrafo sono illustrati i risultati dell'Indagine di copertura a partire dal tasso di copertura per gli individui a livello nazionale, delle cinque ripartizioni territoriali, delle quattro classi di dimensione demografica dei comuni di residenza e dei venti domini costituiti dall'incrocio delle due classificazioni. Oltre i risultati riferiti a questi domini, pianificati in fase di progettazione dell'indagine, sono riportati i tassi di copertura individuali riferiti anche ad alcune variabili sociodemografiche e segnatamente: sesso, classe di età, stato civile, titolo di studio (per i maggiori di 6 anni), condizione professionale (per i maggiori di 15 anni) e posizione nella

[36] Woodruff R.S. (1974) "*A Simple method for Approximating the Variance of a Complicated Estimate*" Journal of the American Statistical Association, vol.66, n.334, pp411-414.

[37] Per applicare il metodo è necessario che G sia differenziabile almeno fino al secondo ordine in un intorno sufficientemente ampio del punto $\mathbf{T}=(T_1,\ldots,T_L)'$.

[38] Pagliuca, D. *Genesees v.3.0, Funzione di Stime ed Errori. Manuale utente ed aspetti metodologici.* 2005. (Tecniche strumenti Istat, n.3).

professione (per gli occupati). Inoltre, sempre per i tassi di copertura individuali, sono esposti i dati relativi alla tipologia di sezione di censimento di residenza e quelli relativi ai 12 comuni metropolitani.

Con riferimento alle famiglie sono invece riportati i tassi di copertura secondo la dimensione della famiglia e i tassi di copertura individuali secondo il numero di componenti della famiglia di appartenenza.

Infine sono presentati e discussi i risultati relativi alla stima della sovracopertura potenzialmente prodotta dagli individui che, possedendo una dimora in più abitazioni per motivi di lavoro, studio o altro, possono essere stati censiti più volte tra i residenti.

La fase di analisi dei dati ha riguardato i 173.109 individui risultati eleggibili tra i 179.886 rilevati all'Idc. La selezione in base all'eleggibilità è stata effettuata per scartare gli individui che alla data di censimento non erano ancora nati o non erano residenti all'interno delle sezioni nelle quali era stata effettuata l'intervista.[39] Dal punto di vista operativo, l'appartenenza dei soggetti all'insieme degli individui eleggibili è stata determinata principalmente in base alle risposte degli interessati alle domande sulla data di nascita e sulla residenza alla data di censimento. In aggiunta è stato compiuto un controllo sugli individui Idc non abbinati, per i quali si è tentato l'abbinamento con quelli non abitualmente dimoranti (Nad) al censimento. In seguito a tale controllo tutti gli individui Idc risultati abbinati a dei Nad sono stati considerati non residenti alla data del censimento anche se all'Idc avevano dichiarato di esserlo, e per questo collocati tra i non eleggibili. Le selezioni effettuate hanno permesso di escludere dal calcolo dalle stime un insieme di 6.777 individui, i quali avrebbero provocato una sottostima del tasso di copertura in quanto, per costruzione, non abbinabili con individui censiti nelle sezioni campione.

Gli individui rilevati dal censimento nelle sezioni campione sono stati 182.519. Di questi, 400 individui rilevati dopo il termine delle normali operazioni di Censimento, in occasione delle successive procedure di recupero basate sulle anagrafi e su altre informazioni possedute localmente dai Comuni, non sono stati considerati nelle operazioni di *linkage*. Infatti per tali individui non sarebbero stati disponibili gli identificativi personali (nome, cognome e indirizzo); ciò avrebbe pregiudicato il loro abbinamento con i corrispondenti record riferiti all'Idc e comportato una sovrastima dell'errore di copertura. Le operazioni di *linkage* effettuate tra gli individui selezionati rispettivamente per l'Idc e il Censimento hanno quindi individuato 169.980 soggetti risultati come rilevati in entrambe le occasioni.

Tavola 2.14 - Tassi di copertura (e deviazione standard) per la popolazione italiana nel complesso, per ripartizione geografica, classe di dimensione demografica dei comuni e per le combinazioni delle due classificazioni

CLASSI DI DIMENSIONE DEMOGRAFICA DEI COMUNI	Ripartizione geografiche					
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole
Tutti i comuni	98,55	98,44	99,35	97,47	98,81	98,76
	(0,112)	(0,234)	(0,151)	(0,325)	(0,234)	(0,252)
0-10 mila	99,30	98,69	99,60	99,45	99,52	99,98
	(0,188)	(0,525)	(0,219)	(0,290)	(0,161)	(0,112)
10 mila -100 mila	99,00	99,32	99,51	98,20	99,04	98,94
	(0,196)	(0,177)	(0,275)	(0,620)	(0,415)	(0,537)
100 mila + (esclusi metropolitani)	98,21	98,23	98,97	97,46	97,36	97,76
	(0,446)	(0,251)	(0,396)	(2,162)	(0,299)	(0,586)
Metropolitani	95,89	96,24	97,89	94,68	96,15	96,80
	(0,156)	(0,320)	(0,230)	(0,344)	(0,256)	(0,128)

Dalla tavola 2.14 si osserva che il grado di copertura a livello nazionale è soddisfacente, attestandosi su un valore superiore al 98,5 per cento e che i valori per ripartizione territoriale variano tra un massimo di 99,35 per cento del Nord-est a un minimo del 97,47 per cento del Centro. Per quanto riguarda il grado di copertura relativo alla dimensione del comune di residenza degli individui, si osserva una marcata dipendenza del fenomeno in relazione all'ampiezza del comune, dato che si passa da un valore massimo di 99,30 per cento per i comuni sotto i 10 mila abitanti a un valore di 95,89 per cento per i comuni metropolitani. Osservando gli incroci tra le classi dimensionali e le ripartizioni territoriali, si osserva che il dato migliore è conseguito per i comuni

[39] È appena il caso di osservare che l'insieme opposto degli individui morti o trasferiti nel tempo intercorso tra il 21 ottobre e la data di rilevazione dell'Idc non costituiscono un problema per l'applicazione del modello cattura-ricattura in quanto, in base all'ipotesi di indipendenza tra le due occasioni di rilevazione, essi tendono a distribuirsi casualmente tra i censiti e i non censiti e pertanto non comportano distorsioni nelle stime.

più piccoli delle Isole (99,98 per cento), mentre quello peggiore si riscontra in corrispondenza dei comuni metropolitani del Centro (94,68 per cento), costituito da Roma e Firenze.

Nella tavola 2.15 sono riportati, a livello Italia, per ripartizione geografica e classe di dimensione demografica dei comuni, i valori dei tassi di copertura secondo il sesso, per gli stranieri e secondo le classi di età degli individui. Si può osservare che non sussistono differenze significative nel grado di copertura del Censimento in relazione al sesso, mentre ben più marcate sono le differenze in rapporto all'età degli individui. Infatti, si nota chiaramente che, per tutte le ripartizioni e tutte le classi di ampiezza demografica, le età tra i 20 e i 29 anni verificano un calo nella copertura. Questo effetto è atteso, in analogia con quanto si verifica anche nei dati censuari di nazioni statisticamente avanzate come Usa e Gran Bretagna, e collegato con una maggiore mobilità degli individui appartenenti a tale fascia d'età rispetto al resto della popolazione. Interessante è anche il dato che riguarda il grado di copertura per la popolazione appartenente alla classe d'età 0-5 anni, il quale risulta più basso di quello generale, oltre che a livello nazionale, anche per il Centro e il Mezzogiorno, e al crescere della dimensione demografica dei comuni. Questo effetto sembra suggerire una tendenza a non dichiarare i più piccoli nel Foglio di famiglia soprattutto nelle situazioni in cui l'errore di copertura è già alto di per sé e consiglia un aumento dell'attenzione da parte dei rilevatori rispetto a questo fenomeno. Il grado di copertura, toccato un limite inferiore per la classe d'età 20-29, torna poi a crescere e tende a mantenersi pressoché costante in corrispondenza delle età via via più anziane, per poi diminuire di nuovo per gli individui di 84 anni o più. Questo andamento, seppure ovviamente con diversi valori di base, si riproduce in tutte le ripartizioni territoriali e per tutte le classi dimensionali dei comuni.

Tavola 2.15 - Tassi di copertura (e deviazioni standard) per sesso, cittadinanza e classe di età degli individui per il totale Italia, ripartizione geografica e classe di dimensione demografica dei comuni

SESSO E CLASSI DI ETÀ	Ripartizioni geografiche					Classi di dimensione demografica dei comuni				
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più[a]	Metropolitani
Maschi	98,48	98,27	99,35	97,50	98,71	98,65	99,16	98,92	98,06	95,91
	(0,142)	(0,362)	(0,149)	(0,369)	(0,269)	(0,252)	(0,274)	(0,223)	(0,393)	(0,247)
Femmine	98,62	98,61	99,35	97,45	98,90	98,86	99,43	99,08	98,34	95,88
	(0,100)	(0,150)	(0,166)	(0,314)	(0,217)	(0,259)	(0,115)	(0,183)	(0,405)	(0,219)
Stranieri	89,66	88,76	97,29	81,15	94,70	91,16	93,83	94,75	94,18	73,92
	(2,854)	(7,355)	(1,066)	(3,911)	(1,934)	(0,871)	(2,925)	(4,283)	(0,823)	(8,304)
0 – 5	97,92	98,41	99,48	95,86	97,21	99,35	98,68	98,43	98,60	94,29
	(0,238)	(0,486)	(0,232)	(0,722)	(0,535)	(0,284)	(0,332)	(0,362)	(0,550)	(0,815)
6 – 13	98,34	97,78	99,36	97,07	98,69	98,92	98,87	99,07	98,24	94,82
	(0,248)	(0,862)	(0,254)	(0,542)	(0,357)	(0,196)	(0,599)	(0,262)	(0,536)	(0,580)
14 – 19	98,62	98,51	98,86	98,18	98,86	98,52	99,50	99,13	97,29	95,49
	(0,222)	(0,664)	(0,565)	(0,572)	(0,308)	(0,259)	(0,428)	(0,305)	(0,534)	(0,621)
20 – 29	97,82	97,28	98,47	96,41	98,73	98,06	99,21	98,70	97,16	92,66
	(0,174)	(0,339)	(0,379)	(0,483)	(0,314)	(0,505)	(0,190)	(0,274)	(0,810)	(0,556)
30 – 44	98,42	98,36	99,32	97,06	98,82	98,55	99,21	99,05	97,94	95,25
	(0,137)	(0,322)	(0,155)	(0,403)	(0,261)	(0,203)	(0,237)	(0,202)	(0,466)	(0,376)
45 – 54	98,83	98,93	99,41	97,66	99,26	98,82	99,51	99,13	98,13	96,95
	(0,131)	(0,242)	(0,222)	(0,270)	(0,203)	(0,156)	(0,182)	(0,213)	(0,415)	(0,389)
55 – 64	98,92	98,76	99,72	98,43	98,81	98,99	99,28	99,19	98,87	97,54
	(0,126)	(0,229)	(0,091)	(0,365)	(0,365)	(0,493)	(0,114)	(0,220)	(0,514)	(0,362)
65 – 74	99,02	99,02	99,68	98,09	99,17	99,38	99,61	99,07	98,78	97,87
	(0,137)	(0,292)	(0,131)	(0,452)	(0,211)	(0,227)	(0,097)	(0,271)	(0,286)	(0,382)
75 – 84	99,23	99,04	99,97	98,78	99,19	99,33	99,63	99,05	99,35	98,80
	(0,140)	(0,359)	(0,080)	(0,384)	(0,244)	(0,177)	(0,223)	(0,224)	(0,235)	(0,405)
85 +	98,80	98,54	99,82	97,77	99,39	98,47	99,79	98,92	98,86	96,34
	(0,355)	(0,835)	(0,116)	(0,824)	(0,651)	(1,519)	(0,256)	(0,663)	(0,539)	(1,255)

(a) Esclusi i comuni metropolitani

Per quanto riguarda gli stranieri, i dati mostrano una sensibile diminuzione del grado di copertura in confronto con quanto accade per la popolazione generale. Il dato sembra assumere una particolare rilevanza nei comuni metropolitani, soprattutto se si considera che la maggior parte degli stranieri rilevati in occasione del Censimento sono probabilmente quelli più stabili sul territorio.

Nella tavola 2.16 sono riportati i tassi di copertura secondo lo stato civile. Si osserva che i maggiori livelli di copertura sono ottenuti per la modalità coniugato, mentre si registrano problemi, anche rilevanti, per gli individui separati, i quali sono comunque in numero esiguo. L'andamento di tali tassi per ripartizione geografica e per classe di dimensione non presentano marcate differenze rispetto ai valori registrati per la popolazione generale. L'andamento abbastanza erratico dei tassi per separati e divorziati è probabilmente determinato dalla scarsa numerosità del campione per tali classi di individui.

Tavola 2.16 - Tassi di copertura (e deviazioni standard) per stato civile degli individui per il totale Italia per ripartizione geografica e classe di dimensione demografica dei comuni

STATO CIVILE	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitani
Celibe/Nubile	98,05 (0,153)	97,87 (0,411)	98,93 (0,211)	96,58 (0,409)	98,46 (0,259)	98,45 (0,267)	99,07 (0,298)	98,74 (0,239)	97,59 (0,434)	94,33 (0,290)
Coniugato/a	98,93 (0,089)	98,95 (0,133)	99,67 (0,130)	98,00 (0,274)	99,07 (0,206)	99,02 (0,219)	99,51 (0,103)	99,22 (0,158)	98,78 (0,344)	96,86 (0,226)
Separato/a	94,65 (0,480)	91,76 (0,964)	98,83 (0,852)	94,12 (1,193)	95,23 (0,742)	96,80 (0,429)	95,84 (0,967)	95,18 (0,405)	93,87 (1,505)	92,64 (1,463)
Divorziato/a	96,69 (0,719)	96,56 (1,153)	97,36 (1,648)	97,33 (1,651)	93,12 (2,000)	99,25 (0,696)	96,79 (1,462)	99,38 (0,411)	90,20 (3,195)	95,47 (1,772)
Vedovo/a	98,59 (0,193)	98,68 (0,350)	99,66 (0,188)	97,18 (0,620)	98,93 (0,424)	98,37 (0,404)	99,37 (0,232)	98,71 (0,376)	98,82 (0,348)	96,59 (0,494)

(a) Esclusi i comuni metropolitani

Nella tavola 2.17 sono illustrati i dati riguardanti il grado di copertura per gli individui di almeno 6 anni secondo il titolo di studio conseguito. A livello nazionale si nota come il grado di copertura si mantenga su valori prossimi a quelli della popolazione generale per i titoli fino al diploma, per poi scendere in modo significativo per i laureati (il dato riferito ai possessori di diploma di laurea è inficiato dall'esiguità dei casi nel campione). Tale andamento tende a riprodursi anche quando si considerano le differenti ripartizioni territoriali e le classi di dimensione demografica, anche se in corrispondenza della classe "nessun titolo" si osserva una marcata flessione nel Centro, nella classe dei comuni sopra i 100 mila abitanti e, soprattutto, in quelli metropolitani.

Tavola 2.17 - Tassi di copertura (e deviazioni standard) per titolo di studio degli individui per il totale Italia per ripartizione geografica e classe di dimensione demografica dei comuni

TITOLO DI STUDIO	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitani
Nessuno	98,55 (0,213)	98,88 (0,286)	99,33 (0,310)	96,81 (0,987)	98,66 (0,260)	99,15 (0,227)	99,58 (0,164)	99,19 (0,228)	97,77 (0,463)	93,36 (1,378)
Licenza elementare	98,54 (0,172)	98,43 (0,549)	99,81 (0,070)	97,89 (0,265)	98,09 (0,258)	98,58 (0,178)	99,00 (0,142)	98,99 (0,166)	98,83 (0,349)	95,60 (1,086)
Licenza media	98,56 (0,141)	98,41 (0,344)	99,27 (0,210)	97,75 (0,437)	98,77 (0,216)	98,58 (0,185)	99,18 (0,271)	98,98 (0,201)	98,02 (0,582)	96,09 (0,298)
Diploma	98,33 (0,165)	98,22 (0,368)	99,05 (0,236)	96,83 (0,401)	99,04 (0,362)	98,82 (0,361)	99,36 (0,302)	98,99 (0,265)	97,75 (0,562)	95,31 (0,331)
Diploma universitario o laurea	97,37 (0,288)	97,16 (0,753)	97,90 (0,471)	95,89 (0,616)	98,95 (0,223)	97,10 (0,982)	97,96 (0,768)	98,51 (0,399)	97,82 (0,225)	95,23 (0,641)

(a) Esclusi i comuni metropolitani

Nella tavola 2.18 sono riportati i risultati relativi alla condizione professionale per gli individui di almeno 15 anni. A livello nazionale, il valore più elevato risulta corrispondere alla categoria “Ritirato”, seguita da quella di “Casalinga”. Questo risultato è atteso, come attesi sono i valori più bassi per i tassi riferiti a “Occupato” e “Studente”. Nel complesso non sembrano emergere situazioni peculiari a particolari aree geografiche o in funzione della dimensione comunale, per le quali risultano confermate le tendenze già riscontrate nella popolazione in generale. Come al solito le categorie rappresentate da un minor numero di individui mostrano valori erratici del tasso e quindi non sono interpretabili facilmente.

Tavola 2.18 - Tassi di copertura (e deviazioni standard) secondo la condizione professionale degli individui per il totale Italia, per ripartizione geografica e classe di dimensione demografica dei comuni

CONDIZIONE PROFESSIONALE O NON PROFESSIONALE	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitani
Occupato	97,99 (0,193)	97,26 (0,436)	99,13 (0,184)	96,91 (0,387)	98,98 (0,554)	97,96 (0,367)	99,15 (0,187)	98,35 (0,387)	97,60 (0,486)	94,73 (0,483)
In cerca di occupazione	97,31 (0,114)	95,74 (0,221)	98,44 (0,152)	96,33 (0,405)	98,15 (0,177)	98,23 (0,253)	99,19 (0,127)	98,13 (0,170)	95,69 (0,327)	91,27 (0,419)
Ritirato	99,07 (0,099)	98,89 (0,183)	99,76 (0,072)	98,36 (0,329)	99,07 (0,237)	99,49 (0,157)	99,49 (0,139)	99,12 (0,175)	99,01 (0,147)	97,82 (0,309)
Studente	97,77 (0,203)	96,82 (0,608)	97,83 (0,378)	97,15 (0,454)	98,57 (0,280)	98,66 (0,375)	98,79 (0,415)	98,71 (0,274)	96,53 (0,357)	93,50 (0,565)
Casalinga	98,68 (0,244)	98,80 (0,553)	99,49 (0,497)	97,62 (0,836)	98,67 (0,350)	98,90 (0,268)	99,22 (0,306)	98,90 (0,364)	98,31 (1,104)	97,30 (0,834)
In altra condizione	94,40 (0,340)	90,63 (0,726)	97,69 (0,941)	97,29 (0,865)	92,94 (0,337)	88,13 (0,418)	97,14 (0,492)	94,36 (0,570)	98,36 (0,845)	98,46 (1,056)

(a) Esclusi i comuni metropolitani

Nella tavola 2.19 si mostrano i tassi relativi alla posizione professionale per il sottoinsieme degli occupati. A livello nazionale i valori più elevati di copertura sono ottenuti per i dipendenti e i lavoratori in proprio. Su un valore più basso si collocano invece i lavoratori nella classe “imprenditore o libero professionista”, mentre le ultime due categorie sono più difficili da interpretare in quanto marginali. Le relazioni tra le tre classi principali di individui, in termini di tasso di copertura, non variano considerando le ripartizioni territoriali e le classi di dimensione dei comuni. È tuttavia interessante notare che, in corrispondenza dei comuni metropolitani, nei quali la copertura tende ad abbassarsi in modo sensibile, le differenze tra i tassi di copertura sono meno marcate.

Tavola 2.19 - Tassi di copertura (e deviazioni standard) per posizione nella professione degli individui per il totale Italia, per ripartizione geografica e classe di dimensione demografica dei comuni

POSIZIONI NELLA PROFESSIONE	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitano
Dipendente	98,56 (0,077)	98,47 (0,143)	99,30 (0,208)	97,28 (0,185)	99,30 (0,093)	98,40 (0,179)	99,33 (0,162)	99,15 (0,097)	98,15 (0,136)	95,58 (0,218)
Imprenditore o Libero professionista	96,46 (0,303)	96,64 (0,450)	97,65 (0,236)	93,49 (0,335)	98,79 (1,458)	95,60 (0,668)	96,95 (0,242)	96,74 (0,245)	97,06 (0,488)	94,64 (1,473)
Lavoratore in proprio	98,61 (0,261)	98,14 (0,696)	99,16 (0,063)	98,27 (0,792)	98,91 (0,037)	98,92 (0,207)	99,28 (0,445)	99,05 (0,238)	97,30 (0,774)	95,18 (1,357)
Altro	95,25 (1,813)	91,57 (2,765)	99,78 (2,442)	94,43 (5,068)	97,36 (6,051)	94,12 (5,257)	98,70 (1,920)	98,95 (3,401)	89,07 (7,217)	81,21 (4,877)

(a) Esclusi i comuni metropolitani

Tavola 2.20 - Tassi di copertura (e deviazioni standard) per tipo di località di residenza degli individui per il totale Italia, per ripartizione geografica e classe di dimensione demografica dei comuni

TIPI DI LOCALITÀ	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitani
Centro	98,64	98,64	99,33	97,59	98,87	98,80	99,45	99,18	98,34	95,96
	(0,095)	(0,145)	(0,148)	(0,291)	(0,207)	(0,288)	(0,114)	(0,176)	(0,431)	(0,160)
Nucleo	98,31	98,29	99,96	96,99	97,42	95,55	99,54	98,09	90,05	89,96
	(0,329)	(0,627)	(0,106)	(0,985)	(0,612)	(3,626)	(0,184)	(0,360)	(8,511)	(3,189)
Case sparse	95,74	91,08	99,15	94,75	97,27	97,78	96,49	95,68	94,56	62,66
	(0,990)	(4,172)	(0,165)	(1,245)	(0,539)	(0,564)	(1,756)	(0,851)	(0,356)	(8,081)

(a) Esclusi i comuni metropolitani

La tavola 2.20 riporta i dati relativi agli individui secondo la tipologia della sezione di censimento in cui si trova la dimora abituale. Come atteso, per le sezioni di tipo case sparse si registrano i valori di copertura più bassi, con un minimo di valore critico per le sezioni di case sparse situate nei comuni metropolitani. Le stime relative alle case sparse sono tuttavia più erratiche a causa del loro minor numero assoluto fra le sezioni abitate e della loro generale minor popolosità. Il dato riferito alle sezioni sembra comunque suggerire che gli errori di copertura si concentrino maggiormente nelle aree periferiche, alle quali si dovrebbe quindi dedicare maggiore cura in fase di rilevazione.

L'analisi della copertura del Censimento ha considerato anche l'enumerazione delle famiglie. La base dati utilizzata per le analisi è consistita in 66.384 famiglie eleggibili rilevate dall'Idc. È bene premettere che il numero assoluto di famiglie rilevate dall'Idc è stato superiore di quasi il 10 per cento rispetto alle famiglie enumerate al Censimento nelle corrispondenti sezioni. Tale risultato non è coerente con il minor numero di individui rilevati dall'Idc a confronto con quelli enumerati al Censimento nelle sezioni campione ed è compatibile con l'ipotesi che le famiglie all'Idc si siano frammentate per effetto, ad esempio, di una diversa impaginazione dei questionari[40] nelle due occasioni di indagine. La conseguente incomparabilità nella definizione di famiglia per le due occasioni di indagine ha consigliato di non applicare il modello cattura-ricattura per stimare il tasso di copertura delle famiglie e di fare invece riferimento alle sole famiglie definite all'Idc. Il tasso di copertura per le famiglie è stato quindi stimato come il rapporto delle famiglie per le quali almeno un componente sia stato censito, sul totale delle famiglie intervistate, sotto l'ipotesi che l'effetto di frammentazione delle famiglie sia stato indipendente dall'errore di copertura.

Il risultato di tale analisi, riportato nella tavola 2.21, mostra che il tasso di copertura è sensibilmente più basso per le famiglie monocomponente, verifica una lieve crescita per le famiglie di 2-3, 4-5 e 6-7 componenti rispettivamente, per poi tornare a diminuire sensibilmente per le famiglie più numerose. Tuttavia, della diminuzione del tasso di copertura per le famiglie con più di otto componenti sembrano responsabili il Centro e i comuni metropolitani, facendo quindi supporre che, in questo effetto, sia prevalente il ruolo di Roma.

Tavola 2.21 - Tasso di copertura (e deviazioni standard) per le famiglie secondo il numero di componenti (a)

COMPONENTI DELLA FAMIGLIA	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (b)	Metropolitani
1	96,91	96,63	98,77	94,27	97,84	97,23	99,24	98,07	97,35	94,35
	(0,258)	(0,558)	(0,558)	(0,558)	(0,558)	(0,558)	(0,442)	(0,529)	(0,287)	(0,501)
2-3	98,97	99,11	99,65	98,08	98,94	98,94	99,70	99,25	99,17	97,93
	(0,103)	(0,180)	(0,121)	(0,383)	(0,166)	(0,180)	(0,073)	(0,157)	(0,624)	(0,249)
4-5	99,12	99,13	99,75	98,32	99,27	99,17	99,62	99,26	99,31	98,12
	(0,116)	(0,224)	(0,144)	(0,345)	(0,212)	(0,362)	(0,140)	(0,202)	(0,449)	(0,334)
6-7	99,52	100,00	98,44	99,91	99,71	99,29	100,00	100,00	99,13	99,72
	(0,307)	(1,024)	(1,637)	(0,103)	(0,329)	(0,729)	(0,012)	(0,575)	(0,495)	(0,674)
8+	98,66	100,00	100,00	90,61	100,00	100,00	100,00	100,00	100,00	93,46
	(0,877)	(8,918)	(6,621)	(6,162)	(1,127)	(1,318)	(0,245)	(0,509)	(4,649)	(4,724)

(a) Una famiglia è considerata *linked* se almeno un suo componente è *linked.*
(b) Esclusi i comuni metropolitani

[40] Fogli di famiglia di due o cinque componenti, con eventuali moduli individuali aggiuntivi, per il censimento; un Foglio di famiglia di cinque componenti, con eventuale compilazione di più moduli per le famiglie numerose, all'Indagine di copertura.

Se il risultato della tavola 2.21 si può ritenere in qualche misura atteso per il modo stesso di definire le famiglie censite, la tavola 2.22 riporta un risultato, sempre connesso alla dimensione familiare, che sembra decisamente interessante. Infatti in tale tavola sono riportati i tassi di copertura riferiti agli individui, condizionatamente alla dimensione della famiglia di appartenenza. Si nota come l'aggregazione degli individui in famiglie favorisca la probabilità di essere censiti solo fino a che la dimensione della famiglia non cresce oltre i cinque individui. Dopo tale numerosità infatti il valore del tasso di copertura scende in modo sensibile. Questo effetto, oltre a una componente fisiologica insita nella dimensione delle famiglie, potrebbe essere spiegato anche dal numero di componenti previsti dal questionario Idc, pari a cinque, che a fronte di un risparmio economico per la conduzione dell'indagine, potrebbe aver determinato una perdita di qualità.

Sembra tuttavia evidente da questa analisi che le componenti più a rischio di sottocopertura siano gli individui che vivono in famiglie monocomponente. Altresì non è assolutamente ipotizzabile, come fu fatto in occasione dell'Indagine di copertura del 1991, che tutti gli individui appartenenti a una famiglia censita siano stati a loro volta censiti.

Tavola 2.22 - Tasso di copertura (e deviazioni standard) per gli individui secondo il numero di componenti

COMPONENTI DELLA FAMIGLIA	Ripartizioni geografiche						Classi di dimensione demografica dei comuni			
	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole	Fino a 10.000	10.000-100.000	100.000 e più (a)	Metropolitani
1	96,91 (0,258)	96,63 (0,558)	98,77 (0,152)	94,27 (0,769)	97,84 (0,507)	97,23 (0,387)	99,24 (0,442)	98,07 (0,529)	97,35 (0,287)	94,35 (0,501)
2-3	98,57 (0,123)	98,73 (0,219)	99,43 (0,120)	97,45 (0,443)	98,48 (0,226)	98,62 (0,227)	99,45 (0,132)	98,91 (0,186)	98,84 (0,686)	97,22 (0,270)
4-5	98,64 (0,137)	98,77 (0,324)	99,39 (0,235)	97,75 (0,400)	98,69 (0,197)	98,70 (0,423)	99,38 (0,188)	98,81 (0,234)	98,85 (0,468)	97,37 (0,362)
6-7	98,27 (0,404)	98,06 (1,299)	98,47 (1,603)	98,90 (0,517)	97,67 (0,566)	99,01 (0,722)	99,64 (0,397)	98,96 (0,638)	98,13 (0,762)	96,85 (1,542)
8+	95,09 (1,607)	94,66 (9,395)	92,66 (6,680)	91,14 (5,874)	95,72 (1,385)	97,99 (1,422)	100,00 (0,765)	96,95 (2,136)	95,93 (4,828)	87,08 (5,483)

(a) Esclusi i comuni metropolitani

Dalla tavola 2.23 si osserva che, nei 13 comuni metropolitani, valori fortemente bassi del tasso di copertura nazionale sono riscontrati nelle città di Firenze, Catania, Bari e Genova; valori intermedi, anche se comunque inferiori al dato nazionale, sono riscontrati nei comuni di Roma, Milano, Napoli, Messina e Venezia, mentre valori in linea con il dato nazionale risultano per il comuni di Bologna, Palermo, Cagliari e Torino.

Tavola 2.23 - Tassi di copertura (e deviazioni standard) per i 13 comuni metropolitani

COMUNE	Tasso %	Deviazione Standard
Torino	99,44	0,133
Genova	93,84	0,711
Milano	95,30	1,000
Venezia	97,61	0,314
Bologna	98,09	0,351
Firenze	91,95	0,652
Roma	95,03	0,387
Napoli	97,08	0,322
Bari	93,31	0,409
Palermo	98,59	0,199
Messina	97,17	0,309
Catania	92,14	0,791
Cagliari	98,75	0,331

L'ultima tavola presentata riguarda il fenomeno della sovracopertura del censimento, stimata in Italia per la prima volta in questa tornata censuaria, e i cui risultati, in questa prima occasione, devono essere considerati solo sperimentali. Infatti, la stima di questa componente dell'errore di copertura non è stata ottenuta attraverso un *record linkage*, ma ricorrendo a una serie di domande somministrate direttamente agli individui, dal momento che sarebbe altrimenti stato necessario disporre dei nominativi di tutti i censiti e non solo di quelli enumerati nelle sezioni campione per l'Idc. Nella tavola 2.24 si riportano, per il totale Italia e per le cinque ripartizioni territoriali, il tasso di sovracopertura (t) e quelli relativi agli individui dimoranti rispettivamente tra 1 e 89 giorni [t(1-89)] e per 90 giorni o più [t(90+)] in una dimora alternativa a quella abituale. Sono inoltre riportati: la percentuale dei casi di sovracopertura sul totale degli individui che hanno fatto uso di una dimora alternativa per almeno un giorno nell'anno precedente al 21 ottobre 2001 (tcond), la percentuale dei "non ricordo" [t(Non ric.)] e quella delle mancate risposte [t(Manc.)] sul totale degli individui Idc.

Tavola 2.24 - Tasso di sovracopertura t (e sua deviazione standard [SQM(t)])

TASSI (a)	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole
t	0,32	0,33	0,24	0,39	0,18	0,59
(Standard error)	(0,062)	(0,096)	(0,115)	(0,095)	(0,042)	(0,338)
t(1-89)	0,08	0,08	0,03	0,12	0,05	0,16
t(90+)	0,24	0,26	0,21	0,27	0,13	0,43
tcond	6,18	5,47	4,17	7,04	5,39	11,18
t(Non ric.)	0,06	0,04	0,04	0,13	0,04	0,04
t(Manc.)	0,53	0,55	0,59	0,57	0,42	0,56

(a) Per il significato dei tassi di sovracopertura si vedano le spiegazioni nel testo.

Il tasso di sovracopertura, definito nel paragrafo 2.10.4, è risultato pari a 0,32 per cento, registrando quindi un valore sensibilmente più basso di quello ottenuto per il tasso di sottocopertura. Tenendo conto del tasso di sovracopertura si ottiene un tasso netto di copertura a livello nazionale del 98,88 per cento anziché il 98,55 per cento restituito tenendo conto della sola sottocopertura. Tuttavia l'accuratezza attesa per la stima del grado di sovracopertura sconsiglia di farne un tale uso dato che, come si vede nelle ultime due righe della tavola 2.24, le percentuali di coloro che non ricordano o non forniscono la risposta sono dello 0,06 per cento e dello 0,53 per cento rispettivamente. Essendo l'entità di tali valori simile a quella ottenuta per il tasso di sovracopertura è, possibile, almeno in teoria, che una parte consistente di tale percentuale possa riversarsi sulla sovracopertura. D'altro canto, studi basati sul confronto delle analoghe domande dirette all'identificazione dei casi di mancato Censimento e l'esito del *record linkage*,[41] suggeriscono che gli individui, almeno per quanto riguarda la misura del fenomeno della sottocopertura, tendano a dare una risposta affermativa anche quando l'evento non si sia verificato. Dalle considerazioni esposte sussistono elementi per sospettare che il tasso di sovracopertura stimato possa essere distorto, sia per difetto sia per eccesso, e che quindi occorrerebbero evidenze ulteriori per certificarne l'accuratezza.

Se il totale dei casi di sovracopertura riportati si rapporta al solo ammontare degli individui che dimorano anche in luoghi diversi dalla loro residenza abituale, anziché al complesso degli individui eleggibili, si nota che tale rapporto cresce fino a un valore superiore al 6 per cento. Inoltre, a testimoniare dell'affidabilità della risposta, si nota come gli errori siano concentrati soprattutto fra gli individui che dimorano per un tempo più lungo (sopra i 90 giorni) in un luogo diverso quello della propria residenza.

I risultati mostrati, anche se evidenziano stime di sovracoperture ancora non abbastanza affidabili, testimoniano comunque la necessità di valutare attentamente, durante un censimento, se il rispondente è effettivamente dimorante nell'alloggio al fine di incorrere nel minor numero possibile di casi di sovracopertura.

Le considerazioni esposte spingono a considerare la stima della sovracopertura effettuata in questa sede come una indicazione di massima dell'importanza di tale fenomeno, ma consigliano per il futuro il ricorso a una tecnica di misurazione con maggiori caratteristiche di oggettività.

[41] Fortini, M. "Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione". In *Atti della XXXVII Riunione Scientifica della Società italiana di statistica*, San Remo: 6-8 aprile 1994.

2.12 - I confronti nel tempo e internazionali

2.12.1 - L'indagine sul grado di copertura in Italia: un quadro storico

In Italia l'indagine sul grado di copertura è stata condotta in occasione degli ultimi tre Censimenti della popolazione e delle abitazioni (1981, 1991, 2001). I tassi percentuali prodotti forniscono una misura della copertura relativamente agli universi delle famiglie (Tavola 2.25) e degli individui residenti (Tavola 2.26). Tali risultati risentono delle differenti modalità con cui è stata svolta l'indagine da un decennio all'altro e vanno quindi letti con la dovuta cautela. Il quadro storico che viene qui presentato vuole documentare i cambiamenti metodologici via via introdotti in risposta ai problemi emersi nelle varie occasioni censuarie.

Tavola 2.25 - Tassi di copertura riferiti alle famiglie per ripartizione geografica e anno di censimento *(valori percentuali)*

RIPARTIZIONI GEOGRAFICHE	1981	1991	2001
Nord	96,52	99,28	98,85
Nord-ovest	-	-	*98,45*
Nord-est	-	-	*99,41*
Centro	95,50	98,66	97,20
Sud e Isole	97,01	99,16	98,79
Sud	**-**	**-**	*98,84*
Isole	**-**	**-**	*98,68*
Italia	**96,45**	**99,10**	**98,51**

Tavola 2.26 - Tassi di copertura riferiti agli individui residenti per ripartizione geografica e anno di censimento *(valori percentuali)*

RIPARTIZIONI GEOGRAFICHE	1991	2001
Nord	99,47	98,83
Nord-ovest	-	*98,44*
Nord-est	-	*99,35*
Centro	98,87	97,47
Sud e Isole	99,34	98,78
Sud	-	*98,81*
Isole	-	*98,76*
Italia	**99,30**	**98,55**

L'indagine del 1981, che ha coinvolto 66 comuni campione rilevando 120 mila famiglie e 15 mila abitazioni non occupate, ha fornito tassi percentuali di copertura solo per l'universo delle famiglie. In quella occasione, infatti, la stima della sottocopertura è stata fatta abbinando le unità rilevate al Censimento e all'Indagine di copertura mediante il nominativo della persona di riferimento del Foglio di famiglia. Errori di percorso e assenza dei suddetti nominativi hanno rappresentato i maggiori problemi, che hanno inciso nella stima dei tassi di copertura. Per il solo 1981, inoltre, i risultati illustrati nella tavola 2.25 non sono riportati all'universo, ma sono riferiti al campione di sezioni costituito a partire dall'insieme iniziale delle sezioni di censimento originariamente estratto (unità di secondo stadio) e sostituendo sia quelle risultate vuote sia quelle con un numero non significativo di famiglie.

Per l'indagine del 1991 è stato adottato un campione di tipo areale costituito da 638 sezioni, selezionate con probabilità di inclusione costante in 85 comuni campione. I comuni con popolazione superiore ai 500 mila abitanti sono stati considerati autorappresentativi, gli altri sono stati stratificati in base alla ripartizione

geografica e all'ampiezza demografica. Complessivamente sono state rilevate dall'indagine 52.844 famiglie, a cui si riferiscono i risultati, e 11.157 abitazioni non occupate.

Nel 1991, nella fase di *record linkage* tra i dati censuari e quelli dell'indagine, gli identificativi utilizzati, poiché si riferivano alle famiglie e non agli individui, non sono risultati abbastanza discriminanti; a ciò si aggiunga, come ulteriore problema, il fatto che non si faceva riferimento al nominativo della persona di riferimento del Foglio di famiglia, bensì alle seguenti chiavi: provincia, comune, sezione di censimento, via, numero civico, piano, interno.

In molti casi, non potendo fare abbinamenti esatti, si è fatto riferimento esclusivamente alle dichiarazioni delle famiglie, e ciò sembra aver contribuito a una sottostima dell'errore di copertura. In effetti, per le dichiarazioni mendaci sia di alcuni intervistatori sia di famiglie reticenti a partecipare al Censimento, molte famiglie sono risultate censite pur non essendolo realmente.

L'indagine del 2001 ha coinvolto 98 comuni (tra cui 14 autorappresentativi), 1.154 sezioni campione (di cui 1.098 con "unità eleggibili") stratificate secondo la tipologia di località abitata (centri; nuclei; case sparse/località produttive) e oltre 68 mila famiglie. Di seguito vengono presentati alcuni dei principali cambiamenti introdotti e i motivi che hanno portato alla decisione di adottarli.

- Al fine di avere identificativi altamente discriminanti per l'individuazione delle unità enumerate nelle due occasioni ed evitare il più possibile errori di abbinamento che, anche se contenuti, compromettono la qualità delle stime di copertura, nel 2001, per la prima volta, si è deciso di acquisire i nominativi e gli indirizzi degli individui rilevati al Censimento e all'Indagine di copertura. Inoltre, rispetto alle precedenti rilevazioni, nel 2001 le stime si riferiscono direttamente agli individui e non solo alle famiglie.
- Per la prima volta sono state prodotte stime della sovracopertura. Tali stime, ottenute dai risultati dei questionari autocompilati in occasione dell'Indagine di copertura e non tramite il *record linkage*, devono però ritenersi solo indicative del fenomeno e non essere considerate pienamente affidabili.
- Nel 2001 si è giunti a misure di copertura che si riferiscono non solo all'intero territorio nazionale e alle cinque ripartizioni, ma anche a quattro classi di dimensione demografica dei comuni. Infatti, la difficoltà nel controllo delle operazioni censuarie nei grandi comuni generalmente comporta una maggiore sottocopertura in tale insieme. Non a caso, la presenza del comune di Roma, dove lo svolgimento del Censimento risulta di norma particolarmente problematico, fa sì che in tutte e tre le indagini il tasso di copertura relativo all'Italia centrale risulti più contenuto rispetto alle altre ripartizioni geografiche (Tabelle 2.25 e 2.26).
- Sempre nel 2001, per valutare il legame tra tasso di copertura e densità della popolazione, si è scelto di stratificare le sezioni campione e di classificare le stime dei tassi di copertura secondo la tipologia di località abitata.
- Infine, si è pensato di ampliare le informazioni relative alla dimensione dei nuclei familiari e alle caratteristiche socio-demografiche degli individui: sesso, classe di età, stato civile, cittadinanza, titolo di studio, condizione professionale, relazione di parentela con l'intestatario.

2.12.2 - I diversi approcci al censimento e i risultati dell'Indagine di copertura: una rassegna internazionale

Un confronto internazionale fornisce un riferimento utile per valutare l'esito del Censimento generale della popolazione e delle abitazioni in Italia e interpretare i risultati dell'Indagine di copertura. Da una rassegna della letteratura internazionale risulta che gli approcci al censimento sono stati diversi nei vari paesi.

Anzitutto, non tutti i paesi conducono censimenti in senso tradizionale, intesi come la completa enumerazione della popolazione riferita a un particolare giorno, ripetuta periodicamente negli anni (in genere ogni cinque o dieci).

Tra gli Stati del continente europeo, Grecia, Italia, Spagna, Portogallo, Irlanda, Regno Unito, Malta, Cipro, Repubblica Ceca, Estonia, Lettonia, Polonia, Ungheria, Slovacchia, Turchia conducono un Censimento della popolazione di tipo tradizionale utilizzando fonti amministrative e registri come strumenti integrativi e di supporto.

L'elevato costo del Censimento ha spinto molti paesi a trovare soluzioni alternative per la raccolta di dati demografici e socio-economici, anche con l'uso di dati provenienti da fonti esterne: paesi dell'Europa settentrionale, quali Danimarca, Finlandia, Islanda, Svezia, Norvegia, si basano esclusivamente su registri amministrativi (inclusi i registri della popolazione). Un altro gruppo di paesi (Austria, Belgio, Svizzera, Lussemburgo) sta passando dal censimento tradizionale ai registri della popolazione.

Tra i paesi che stanno applicando nuove soluzioni, i Paesi Bassi integrano i dati di fonte amministrativa e di indagini campionarie; la Germania integra i registri comunali della popolazione con dati di altre fonti amministrative ed effettua ogni anno un microcensimento; la Francia sta sperimentando un nuovo modello che prevede, nell'arco temporale di cinque anni, un vero e proprio Censimento a rotazione per i comuni con meno di 10 mila abitanti insieme a un'indagine campionaria, che ogni anno copre l'8 per cento della popolazione (il 40 per cento della popolazione nell'arco dei cinque anni), per tutti gli altri comuni.

Eccezion fatta per i paesi che si basano totalmente sui registri amministrativi, quasi tutti conducono studi di qualità e/o copertura.

Passando da un contesto strettamente europeo a un contesto mondiale, nella tavola 2.27 sono illustrati alcuni risultati relativi alle indagini di copertura dei paesi che li hanno resi disponibili.

Tavola 2.27 - Tassi di errore di copertura per alcuni paesi esteri – Censimenti 2000-2001 *(valori percentuali)*

PAESI	Anno	Err. %
Regno Unito	2001	2,0
City of London (a)		*16,0*
Svizzera	2000	1,4
Zurigo		*1,5*
Estonia	2000	1,2
Usa	2000	1,2
Canada	2001	3,0
Australia	2001	1,1
Nuova Zelanda	2001	2,2
Repubblica Popolare cinese	2000	1,9

Fonte: ONS, Swiss Federal Statistical Office, Statistical Office of Estonia, U.S. Census Bureau, Statistics Canada, Australian Bureau of Statistics, Statistics New Zealand, China's National Bureau of Statistics
(a) I dati non tengono conto dei recuperi.

Nella tavola 2.28 sono riportate la periodicità e la modalità di rilevazione del censimento. Infatti, l'analisi dei dati di copertura e il loro confronto internazionale deve tener conto, oltre a eventuali differenze nel disegno campionario e nelle definizioni adottate, anche delle diverse tecniche di rilevazione al censimento (questionari lasciati e/o ritirati da rilevatori, spediti e/o ritirati via posta o via internet) che influenzano i tassi di risposta. Infine, sempre dalla tavola 2.28 si può notare che alcuni paesi hanno usato i risultati dell'Indagine di copertura per correggere i dati del censimento e ricalcolare la popolazione residente resa come dato ufficiale di popolazione.

Tavola 2.28 - Il censimento in alcuni paesi esteri

PAESI	Cadenza Censimento (anni)	Modalità enumerazione	Correzione dati censuari
Regno Unito	10	Questionari prevalentemente consegnati da rilevatore e raccolti per posta. Interviste nei casi ritenuti critici	Si, basandosi interamente sull'Indagine di copertura
Svizzera	10	Mista a seconda della scelta del comune	No
Estonia	10	Intervista	No
Usa	10	Questionari prevalentemente recapitati e raccolti per posta. Interviste nei casi ritenuti critici	No
Canada	5	Questionari consegnati e raccolti da rilevatore	No
Australia	5	Questionari consegnati e raccolti da rilevatore	Uso dell'Indagine di copertura e di altre fonti amministrative per la stima della popolazione residente
Nuova Zelanda	5	Questionari consegnati e raccolti da rilevatore	No
Repubblica Popolare Cinese	10	Intervista	Si, basandosi interamente sull'Indagine di copertura

Fonte: ONS, Swiss Federal Statistical Office, Statistical Office of Estonia, U.S. Census Bureau, Statistics Canada, Australian Bureau of Statistics, Statistics New Zeland, China's National Bureau of Statistics

Regno Unito: ufficialmente il Censimento del 2001 ha coperto l'intera popolazione residente. In realtà l'Indagine di copertura, Census coverage survey (Ccs), che ha riguardato le 101 aree geografiche in cui è stato suddiviso il territorio, ognuna con 500 mila abitanti, per un campione di circa 320 mila abitazioni, ha stimato un tasso di copertura del 98 per cento con un tasso di risposta al Censimento del 94 per cento. L'errore di copertura non viene diffuso quantificando distintamente la sottocopertura dalla sovracopertura. Il dato riferito a Londra, pari al 16 per cento, risulta molto elevato se non si tiene conto che si riferisce ai soli questionari restituiti compilati, senza considerare eventuali recuperi. Applicando però a Londra la proporzione di recuperi dichiarati a livello nazionale, la stima dell'errore di copertura si riduce al 5 per cento. La difficoltà di leggere chiaramente i risultati della Ccs deriva dall'impostazione adottata nel Censimento del 2001. Il Regno Unito, infatti, sembra essere il primo paese che ha corretto i risultati del censimento ufficiale esclusivamente sulla base dell'Indagine di copertura. La Ccs infatti è inserita nel progetto "*One number census*" che, integrando i dati del censimento con la stima di sottocopertura, perviene a un'unica stima della popolazione residente.

Svizzera: il Censimento ha periodicità decennale. Quello del Duemila ha enumerato sia la popolazione presente sia quella residente. I comuni, che gestiscono la raccolta dei dati censuari, ne hanno scelto la modalità. Il risultato è stato un metodo di raccolta misto: intervista, questionari inviati e raccolti per posta, inviati per posta e raccolti dal rilevatore, Internet. La Svizzera è uno di quei paesi che sta valutando il passaggio dal censimento tradizionale ad altre soluzioni alternative per la raccolta di dati demografici e socio-economici. L'obiettivo dell'Indagine di copertura, svoltasi per la prima volta in occasione del Censimento del Duemila, è stato quello di valutare e migliorare la qualità del censimento; in quest'ottica di analisi, diversi sono i risultati prodotti e rilasciati. Oltre all'entità della sottocopertura (1,6 per cento) e della sovracopertura (0,4 per cento), infatti, la Svizzera conta di fornire l'errore di copertura non solo per particolari domini territoriali e caratteristiche socio-demografiche degli individui, ma anche secondo la modalità di raccolta dati.

Estonia: ha condotto nel Duemila un censimento tradizionale con interviste faccia a faccia. L'Indagine di copertura, che ha riguardato circa l'1 per cento della popolazione, ha stimato un tasso di sottocopertura dell'1,2 per cento.

Usa: l'ultimo censimento, a cadenza decennale, si è svolto nel Duemila e ha enumerato le persone secondo la residenza abituale. Gran parte dei modelli sono stati spediti per posta, solo nelle zone remote o rurali sono stati consegnati dai rilevatori. I questionari compilati sono stati raccolti essenzialmente per posta, nei casi restanti tramite rilevatori. L'Indagine di copertura (Accuracy and coverage evaluation - Ace), condotta su 300 mila abitazioni, seguendo la metodologia del *dual system*, ha stimato un errore di copertura dell'1,2 per cento. L'indagine doveva essere usata per aggiustare i risultati del censimento. Una successiva analisi demografica ha mostrato che l'Ace non è riuscita a misurare adeguatamente un numero significativo di enumerazioni errate (molti duplicati), sovrastimando così l'ammontare della popolazione di circa 3 milioni di persone. C'è stata quindi una revisione dell'Indagine di copertura e comunque i risultati del Censimento non sono stati più modificati.

Canada: il Censimento ha cadenza quinquennale; l'ultimo, del 2001, ha enumerato gli individui nella loro residenza abituale. Nel 98 per cento dei casi i questionari distribuiti dal rilevatore sono stati autocompilati e restituiti per posta; nel 2 per cento dei casi, dove si prevedeva una mancata risposta, è stata condotta un'intervista. L'errore netto di copertura è pari quasi al 3 per cento, con un tasso di sottocopertura di oltre il 3,9 per cento e un tasso di sovracopertura superiore allo 0,9 per cento. Il Canada ha fornito l'errore di copertura per le province che lo compongono e per le caratteristiche socio-demografiche degli individui. Tali risultati sono il frutto di analisi specifiche dedicate da un lato a quantificare e a correggere i dati censuari tenendo conto degli errori di copertura derivanti da un'errata classificazione delle abitazioni (occupata, non occupata), dall'altro volte interamente a misurare l'entità della sovracopertura. La particolare cura rivolta alla classificazione delle abitazioni deriva dall'esperienza dei passati censimenti poiché un'errata classificazione sembrerebbe aver inciso significativamente sulla crescita dell'errore di copertura.

Australia: il censimento, a cadenza quinquennale, enumera la popolazione presente. Ai rispondenti vengono chieste informazioni sul luogo di dimora abituale. I questionari vengono lasciati alle famiglie e successivamente raccolti dai rilevatori. L'Indagine di copertura ha stimato un tasso netto di sottocopertura dell'1,8 per cento. In particolare, il Censimento non ha enumerato il 2,7 per cento delle persone presenti in Australia la notte del 7 agosto 2001, mentre la sovracopertura è pari allo 0,9 per cento. L'Australia usa l'Indagine di copertura e altre fonti amministrative per stimare il dato ufficiale della popolazione residente.

Nuova Zelanda: il censimento viene svolto ogni cinque anni ed enumera la popolazione presente mediante la consegna e la raccolta dei questionari da parte dei rilevatori. L'Indagine di copertura, condotta per la prima volta nel 1996, nel 2001 ha riguardato 11 mila abitazioni (lo 0,7 per cento delle abitazioni totali) e 25 mila individui, stimando un tasso netto di sottocopertura del 2,2 per cento. I risultati dell'Indagine di copertura non sono utilizzati per correggere i dati del Censimento, bensì per le stime della popolazione intercensuaria.

Repubblica popolare cinese: il quinto Censimento della popolazione, svoltosi nel Duemila, per la prima volta ha enumerato tutta la popolazione, sia residente sia temporaneamente presente, nelle 31 province, regioni autonome e municipalità (escluse le isole di Jinmen e Mazu nella provincia di Fujian) del territorio cinese. Un elemento di assoluta novità è rappresentato dal fatto che i rilevatori hanno distribuito due diverse versioni di questionario: una breve (*short-form questionnaire)* al 90 per cento della popolazione, una più dettagliata (*long-form questionnaire)* al restante 10 per cento. L'Indagine di copertura, condotta su 602 distretti di Censimento, ha stimato un tasso di sottocopertura del 1,81 per cento. Per la prima volta il China's national bureau of statistics ha corretto la popolazione residente al Censimento esclusivamente sulla base dell'Indagine di copertura.

In generale, dal confronto internazionale emerge che la quota dei censiti rispetto alla popolazione stimata dall'Indagine di copertura varia dal 97,1 per cento in Canada al 98,8 per cento negli Usa; il tasso di copertura in Cina, pari al 98,2 per cento, è in linea con gli altri paesi, anche se giudicato dal China's national bureau of statistics più alto rispetto alle precedenti occasioni censuarie; in Italia il 14° Censimento della popolazione e della abitazioni ha coperto oltre il 98,5 per cento della popolazione. Purtroppo, al momento, anche a causa della scarsa disponibilità di informazioni circa i tassi di risposta, non è immediato analizzare l'influenza delle diverse tecniche di rilevazione, riportate nella tavola 2.27, sul buon esito del Censimento.

Concludendo, si sottolinea una generale difficoltà nel reperire i dati di copertura a fronte di un dato censuario ampliamente diffuso. Inoltre, sussistono senz'altro delle differenze tra i paesi nelle modalità di rendere disponibili e più facilmente leggibili i risultati dell'Indagine di copertura, che sembrerebbero legate alle diverse finalità assunte dall'indagine. In effetti, come emerge anche dalle analisi sopra riportate, i paesi che hanno effettuato l'Indagine di copertura per apportare miglioramenti ai successivi censimenti hanno dato completo accesso alle informazioni, riportando anche le metodologie utilizzate, e ciò ha permesso una più immediata interpretazione dei risultati ottenuti.

2.13 - Conclusioni

L'indagine sul grado di copertura del 14° Censimento della popolazione è stata caratterizzata da importanti innovazioni rispetto alle passate esperienze. Tra queste il fatto di essere stata per la prima volta effettuata sugli individui anziché sui nuclei familiari, di aver usufruito di nomi e cognomi dei rispondenti per una migliore identificazione del loro stato di censiti/non censiti, di aver applicato tecniche di *record linkage* probabilistico e aver stimato, a livello sperimentale, il grado di sovracopertura. I risultati mostrano che il tasso di copertura a livello nazionale è compatibile con gli standard internazionali, coerente con il livelli attesi e confrontabile con i dati anagrafici. La copertura risulta maggiore nel Nord-est del paese, nelle Isole e nei centri sotto i 10 mila abitanti, mentre i valori più bassi si riscontrano nel Centro Italia e nei comuni metropolitani. Non si segnalano significative differenze di genere, mentre si riscontrano maggiori problemi nella copertura della classe d'età 19-29 anni e, sorprendentemente, in quella 0-5. Per questo secondo caso i problemi sembrano concentrarsi soprattutto nei comuni metropolitani, nel Centro e nel Sud Italia. Problemi si presentano per la copertura degli stranieri, anche in questo caso proporzionalmente maggiori nei comuni metropolitani. Per quanto riguarda le altre caratteristiche sociodemografiche considerate, si riscontra che i maggiori problemi di copertura si presentano per i divorziati, per le classi di istruzione superiore (nei comuni metropolitani, però, anche per i senza titolo), per gli occupati e gli studenti, per i liberi professionisti o lavoratori in proprio.

L'analisi del grado di copertura delle famiglie ha mostrato una decisa prevalenza dell'errore per le famiglie di un solo componente, ma ha anche rivelato problemi per le famiglie sopra i cinque componenti. In generale, viene confermata la convenienza di concentrare l'analisi sugli individui, non essendo valida l'ipotesi che una volta identificata una famiglia, tutti i propri componenti si possano considerare censiti.

La stima della sovracopertura, calcolata per la prima volta a livello sperimentale, ha riguardato esclusivamente coloro i quali all'epoca del Censimento possedevano più di un'abitazione di riferimento. Si teme, infatti, che il contemporaneo conteggio delle popolazioni dei temporaneamente e degli abitualmente

dimoranti possa aver causato duplicazioni in questa seconda classe di soggetti, qualora parte di essi si fossero erroneamente registrati nella lista A (persone residenti) anche nel luogo dove alloggiavano temporaneamente.

La stima della sovracopertura si è basata sulla dichiarazione degli interessati. I risultati testimoniano un livello di sovracopertura trascurabile rispetto a quello della sottocopertura, nell'ordine di 5 volte inferiore (1,45 per cento il tasso di sottocopertura contro lo 0,3 per cento del grado di sovracopertura). Tuttavia il sospetto di distorsioni causate da errori non campionari, dovuti al fatto che gli individui possano ricordare di essere stati censiti in altri alloggi, anche quando non è vero e soprattutto per la presenza relativamente consistente di mancate risposte, non consente di valutare l'accuratezza di tali stime e consiglia per il futuro di utilizzare mezzi differenti dalla domanda diretta agli individui.

A fronte delle innovazioni introdotte, dei risultati conseguentemente ottenuti e alla luce dell'esperienza acquisita, possono essere considerati i possibili sviluppi futuri. In questo contesto si possono quindi identificare futuri spazi di miglioramento, anche consistenti, soprattutto in termini di tempestività, senza che questo abbia a incidere con l'accuratezza dei dati. Ciò potrebbe essere ottenuto soprattutto limitando al massimo l'acquisizione di dati micro e macro dal Censimento e automatizzando le procedure di *linkage*, dimostratesi particolarmente affidabili, così da ridurre il ricorso a interventi manuali. Risultati apprezzabili si potrebbero inoltre ottenere anche migliorando l'organizzazione interna e aumentando la dotazione di personale impiegato a tempo pieno.

Infine, dal punto di vista dell'integrazione dei dati dell'Indagine di copertura con gli altri disponibili sull'argomento, si evidenziano potenziali grandi margini di miglioramento derivanti da un maggiore coordinamento con le analisi, attualmente svolte in modo completamente indipendente, che riguardano il confronto tra censimento e anagrafe.

Capitolo 3 - Le analisi per la stima dell'errore di risposta

3.1 - Introduzione

La valutazione della variabilità dei risultati riconducibile agli errori di risposta normalmente richiede il confronto delle risposte fornite al censimento con quelle ottenute a una successiva indagine di qualità, condotta su un campione di individui censiti .[42] In occasione del Censimento della popolazione e delle abitazioni (Cpa) del 2001 non è stata condotta una indagine ad hoc, ma sono stati utilizzati i dati derivanti dell'Indagine di copertura (Idc), il cui disegno è stato opportunamente modificato per rispondere al duplice obiettivo, di stimare la copertura e di valutare l'errore di risposta. In breve, al questionario di reintervista dell'indagine di copertura sono stati aggiunti alcuni dei quesiti presenti sul questionario del censimento. Per tali quesiti è stato possibile costruire le stime della varianza semplice di risposta (Vsr), confrontando le risposte alle due occasioni di rilevazione, dopo aver effettuato la procedura di *record linkage* tra gli individui trovati al censimento e quelli rintracciati nella indagine di copertura. A causa del disegno adottato, non è stato possibile derivare stime della distorsione di risposta, ma solo della variabilità. Inoltre, per evitare un eccessivo appesantimento del questionario Idc, si è deciso di considerare solo alcuni dei quesiti rilevati al censimento, escludendo in particolare i quesiti sul pendolarismo. La stima della Vsr riferita a ciascuna delle variabili considerate è stata rapportata alla stima variabilità complessiva delle stesse ottenendo così l'indice di inconsistenza (*I*). Una regola empirica per l'interpretazione di *I* considera come critici quei quesiti per i quali l'indice superi 0,50; valori tra 0,20 e 0,50 indicano invece situazioni di variabilità moderata.[43]

Le stime sono state calcolate pesando ciascun individuo abbinato in una data sezione campione con il peso finale assegnato alla stessa al termine della fase di aggiustamento dei pesi base della Idc, trasformato in modo tale che la somme dei pesi delle sezioni campione eleggibili fosse uguale alla loro numerosità (1.104 sezioni con almeno un individuo censito).

Sono state calcolate le stime per l'Italia e per alcuni dei principali domini di stima. Si noti che per la valutazione dell'errore sono stati utilizzati i dati disponibili prima delle procedure di controllo e correzione (le due rilevazioni hanno previsto modalità diverse di acquisizione dei dati, per il Cpa si è utilizzata la lettura ottica mentre per l'Idc la registrazione è stata condotta manualmente) per questo motivo parte della variabilità riscontrata potrebbe essere attribuibile ad errori di registrazione (o di lettura ottica) dei questionari.

Nel seguito sono presentate in dettaglio le scelte intraprese per adattare la Idc ai fini della valutazione della accuratezza dei dati raccolti al Cpa, le metodologie utilizzate e i principali risultati ottenuti. Nelle appendici B e C si riportano una sintesi della teoria che sta alla base delle metodologia di valutazione degli errori di risposta nelle indagini di qualità, le formule utilizzate per le stime nel caso in questione e le tabelle dati ottenute incrociando le risposte alla Idc con quelle del Cpa.

3.2 - Il disegno dell'Indagine di copertura e la valutazione dell'errore di risposta

La decisione di utilizzare l'Indagine di copertura per valutare l'impatto degli errori di risposta sulle stime finali dei parametri di interesse ha posto alcune limitazioni. A questo proposito conviene illustrare brevemente le caratteristiche essenziali di un'indagine di qualità.

Un'indagine di qualità mira a stimare l'impatto degli errori di risposta sulle stime finali di una data indagine. Gli errori di risposta si hanno quando il valore osservato su un dato individuo per un certo fenomeno risulta essere diverso dal "vero" valore. Tale errore può essere dovuto al caso, portando a valori a volte maggiori a volte minori di quello vero, oppure a fattori sistematici, tali che l'errore si manifesta sempre nella stessa direzione. La diversa natura degli errori di risposta ha implicazioni diverse sulle stime finali. L'errore casuale

Il presente capitolo è stato redatto da Giovanna Brancato (parr. 3.1, 3.2) e Marcello D'Orazio (parr. 3.3, 3.4)

[42] US Bureau of the Census, 1985, 2003.

[43] US Bureau of the Census, 1985.

comporta un aumento della variabilità associata alle stime prodotte dall'indagine, per cui si parla di "varianza di risposta". A sua volta la varianza di risposta si compone di due termini: la varianza semplice di risposta (*Simple response variance*: Srv) e la varianza correlata di risposta (*Correlated response variance*: Crv). La varianza semplice di risposta (Srv) riflette quanto le diverse misurazioni, condotte (nelle medesime condizioni) sulla stessa unità, possono variare tra loro per fattori legati esclusivamente al caso. La componente correlata della varianza di risposta (Crv) è dovuta alla somiglianza tra i valori osservati per un fenomeno su un gruppo di unità distinte, ad esempio gli individui intervistati da uno stesso intervistatore. Tale componente si ritiene trascurabile nel caso di questionari autocompilati (la raccolta dei dati non è condotta da rilevatori) come è prevalentemente avvenuto per il Cpa e, pertanto, viene assunta nulla nei modelli che si utilizzeranno (si veda l'appendice C).

La componente sistematica dell'errore di risposta introduce distorsione nella stima di un parametro della popolazione; in pratica, il valore stimato è sistematicamente al di sopra o al di sotto del vero valore del parametro; tecnicamente si parla di "distorsione da risposta".

In un'indagine di qualità solitamente la stima della varianza e della distorsione da risposta implica il ricorso a due diverse strategie di raccolta dei dati.[44] In genere, si seleziona un campione casuale delle unità osservate al censimento e lo si suddivide casualmente in due sub-campioni distinti (solitamente di ampiezza diversa). Le unità del sub-campione di ampiezza più piccola sono sottoposte ad una semplice reintervista. In pratica, ad esse viene sottoposto nuovamente un sottoinsieme delle domande del questionario dell'indagine originaria, seguendo quanto più possibile le stesse modalità di intervista adottate nell'intervista principale. Alle restanti unità (sub-campione più ampio di rispondenti all'indagine originaria) si somministra un'intervista con riconciliazione. In pratica, si reintervistano gli individui e qualora emerga una discrepanza tra le risposte fornite alla stessa domanda se ne chiede conto cercando di stabilire quale sia la "vera" risposta.

Il sub-campione di unità sottoposto a semplice reintervista viene utilizzato per stimare la Srv attraverso il confronto delle risposte ottenute alla reintervista con quelle della intervista originaria. Se, ad esempio, si considera una domanda che preveda una risposta dicotomica allora si costruisce la tavola del tipo (Tavola 3.1):

Tavola 3.1 - Esempio della tabella da costruire ai fini della stima della varianza semplice di risposta (Svr)

REINTERVISTE	Intervista originaria			
	"Sì"	"No"	Missing	Totale
"Sì"	$n_{SI,SI}$	$n_{SI,NO}$	$n_{SI,Miss.}$	$n_{SI,\bullet}$
"No"	$n_{NO,SI}$	$n_{NO,NO}$	$n_{NO,Miss.}$	$n_{NO,\bullet}$
Missing	$n_{Miss.,SI}$	$n_{Miss.,NO}$	$n_{Miss.,Miss.}$	$n_{Miss.,\bullet}$
Totale	$n_{\bullet,SI}$	$n_{\bullet,NO}$	$n_{\bullet,Miss.}$	$n_{\bullet\bullet}$

e, nell'ipotesi che il campione per la reintervista sia stato selezionato mediante campionamento casuale semplice si calcola:

$$\hat{SRV} = \frac{1}{2} \cdot \frac{n_{SI,NO} + n_{NO,SI}}{n_{SI,SI} + n_{SI,NO} + n_{NO,SI} + n_{NO,NO}} = \frac{1}{2} g$$

in cui *g* è noto come tasso lordo di discrepanza (nella letteratura anglosassone è noto come *Gross difference rate* - Gdr). Tale indice non è altro che il tasso delle unità del campione che hanno risposto differentemente alla domanda in questione nelle due occasioni di indagine (tasso di risposte discrepanti). Come si può notare, ai fini della stima della varianza di risposta non si tiene conto dei valori mancanti ad una o ad entrambe le occasioni di indagine.

Di solito per avere un'idea dell'entità della varianza di risposta conviene confrontarla con la varianza complessiva calcolando l'indice di inconsistenza:

[44] Biemer P. P., Forsman G. "On the Quality of Reinterview Data with Application to the Current Population Survey". *Journal of the American Statistical Association,* 87, n.420 (1992): 915-923; US Census Bureau. "Evaluating Censuses of population and Housing". *Statistical Training Document* ISP-TR-5 (1985).

$$I = \frac{SRV}{V_{tot.}} = \frac{SRV}{SRV + S^2_{Ey}}$$

in cui S^2_{Ey} è la variabilità di un insieme di singole misurazioni condotte su tutte le unità della popolazione in assenza di errori di misurazione dovuti al caso. Per stimarlo basta sostituire le quantità che lo compongono con le corrispondenti stime.

Per quel che riguarda l'interpretazione di *I* vale la seguente regola empirica[45]:

$I \leq 0{,}20$	Varianza di risposta bassa
$0{,}20 < I < 0{,}50$	Varianza di risposta moderata
$I \geq 0{,}50$	Varianza di risposta elevata

Nella valutazione dell'indice di inconsistenza bisogna usare una certa cautela in presenza di categorie rare, la cui frequenza relativa sia inferiore a 0,05.[46] In tal caso, infatti, la stima dell'indice di inconsistenza può risultare piuttosto elevata anche solo in presenza di poche coppie di individui che abbiano fornito risposte diverse per la stessa domanda nelle due occasioni di indagine. Più in generale, in presenza di categorie rare, la varianza complessiva tenderà ad essere bassa per cui si potranno ottenere stime elevate per *I* anche quando *g* è basso. Molta cautela deve essere usata anche per quelle domande a cui abbiano risposto solo pochi individui. In tali situazioni, le stime per gli intervalli di confidenza degli indici potrebbero risultare instabili.

Per la stima della distorsione da risposta si considera la sola parte del campione in cui la reintervista è stata condotta con riconciliazione. In particolare, posto di essere interessati alla percentuale di quanti hanno risposto "Sì" alla domanda in questione, la stima della distorsione da risposta è data dalla differenza tra la stima ottenuta sulle unità del campione alla prima occasione di indagine ($p_{SI,1}$) e la stima della percentuale delle persone per le quali il vero valore della risposta è "Sì", $p_{SI,3}$, (valori desunti da riconciliazione):

$$\hat{NDR} = p_{SI,1} - p_{SI,3}$$

Ndr (deriva da *Net difference rate*) è il tasso netto di discrepanza.

L'Indagine di copertura, come illustrato nel capitolo 2, prevede la selezione di un campione areale di sezioni di censimento, anziché un campione di individui. In pratica, ogni sezione campione viene censita nuovamente e attraverso la procedura di *record linkage* si individuano le persone "coperte" da entrambe le rilevazioni.

Per poter utilizzare i dati della Idc ai fini della valutazione dell'incidenza degli errori di risposta, in primo luogo si è provveduto ad ampliare il questionario di questa indagine in modo da rilevare nuovamente alcuni fenomeni già osservati al Cpa. In particolare, l'attenzione è stata posta sulle seguenti variabili:

- Relazione di parentela
- Sesso
- Data di nascita
- Luogo di nascita
- Cittadinanza
- Stato civile
- Titolo di studio
- Condizione professionale
- Posizione nella professione

Ovviamente, per assicurare la comparabilità tra dati Idc e Cpa, le domande incluse nel questionario della Idc dovevano risultare identiche a quelle del questionario Cpa. Analogamente si sono considerate le medesime modalità di risposta, nello stesso ordine, con la medesima relativa codifica.

Nell'ipotesi che la Idc sia una replicazione su scala ridotta del Cpa, si può pensare di utilizzare i dati raccolti in essa ai fini della stima della Srv, sebbene limitatamente al sottoinsieme di individui che sono stati

[45] US Census Bureau. "Evaluating Censuses of population and Housing". *Statistical Training Document* ISP-TR-5 (1985)

[46] US Census Bureau. "Census 2000 Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by reinterview". *Census 2000 Evaluation, Final Report* B.5 (2003): 9.

"coperti" da entrambe le rilevazioni. Ovviamente, per come è strutturata la Idc, non è possibile condurre la reintervista con riconciliazione. La Idc, infatti, censisce nuovamente le unità di una sezione campione andando a verificare a posteriori se una unità è stata "coperta" anche dal Cpa. Questa limitazione rende impossibile stimare la distorsione da risposta attraverso i dati della Idc, con metodologia classica.

Il fatto che la Idc preveda la selezione di un campione di sezioni di censimento, anziché un campione di individui, è suscettibile di introdurre anche alcuni problemi ai fini della stima della varianza di risposta. Infatti, l'individuazione delle unità di una sezione campione osservate, sia al Cpa che nella Idc, passa attraverso la complessa procedura di accoppiamento, detta *record linked* (illustrata nel precedente capitolo). In pratica, per una data sezione campione può accadere che il numero di individui accoppiati sia inferiore al numero di individui censiti. In tal caso, la stima della Srv condotta solo sugli abbinati (*linked*) può ritenersi attendibile solo se il sotto-insieme dei *linked* può configurarsi come un campione casuale degli individui censiti. Se, invece, si ha motivo di ritenere che gli individui censiti, ma non *linked*, siano più propensi a commettere errori di risposta, allora ci si attende che la stima dell'Srv sul solo insieme dei *linked* possa essere inferiore al valore effettivo. Per certi versi, il sottoinsieme degli individui *linked* può essere considerato alla stregua dell'insieme di rispondenti ad un'indagine; com'è noto in tali casi la distorsione delle stime finali si può ritenere trascurabile in presenza di frazione molto bassa di unità non rispondenti. Nel nostro caso, la bassa percentuale di unità censite e non *linked*, che si attesta intorno al 5 per cento (4,94 per cento considerando i pesi finali della Idc; senza pesi la percentuale di non *linked* sale 5,42 per cento), lascia pensare che la distorsione determinata dal fatto di stimare la Srv solo sui *linked* possa essere ritenuta trascurabile.

Un altro possibile problema, legato all'utilizzo dei dati della Idc per la stima della Srv, è rappresentato dalla presenza di "falsi *link*", ossia record accoppiati dalla procedura di *record linkage*, ma che nella realtà fanno riferimento a due individui distinti. Una simile eventualità può incidere negativamente sulla stima della Srv poiché il confronto di record che fanno riferimento a due individui distinti verosimilmente tenderà a produrre discrepanze su gran parte delle variabili comuni alle due indagini, conducendo quindi ad una sovrastima della Srv effettiva. L'incidenza degli errori dovuti al *record linkage* è stata studiata e si è verificato che fosse trascurabile, rassicurando sulla correttezza delle stime del Srv[47].

In definitiva, la stima della variabilità determinata dagli errori di risposta è stata condotta limitando l'attenzione ai soli individui accoppiati per ciascuna sezione censuaria selezionata nel campione. Escludendo anche le sezioni campione per le quali non vi è alcun individuo accoppiato, alla fine le analisi sono state condotte su 1.094 sezioni campione per un totale di 172.620 individui accoppiati (Tavola 3.2).

Tavola 3.2 - Principali risultati della procedura di *record linkare*

	.Sezioni	Individui censiti	Individui alla Idc	Individui *linked*
Sezioni con almeno un *linked*	1.094	182.492	180.133	172.620
Sezioni con nessun *linked*	5	27	29	-
Sezioni con abitazioni non occupate	8	-	-	-
Totale	**1.107**	**182.519**	**180.162**	**172.620**

Nelle tabelle che seguono, l'aumento di variabilità dovuto alla presenza di errori di risposta è stato misurato stimando, per ciascuna domanda comune alle due indagini, il Gdr (*g*) calcolato sia per ciascuna modalità di risposta che a livello aggregato, e l'indice di consistenza *I*, anch'esso calcolato per ciascuna modalità di risposta e a livello aggregato. Ad essi è stata affiancata anche la stima del Ndr, calcolato confrontando le stime delle percentuali osservate per ciascuna modalità di risposta nelle due occasioni di indagine. Infatti l'Ndr in questo caso può essere utilizzato per confrontare le distribuzioni delle risposte fornite a ciascuna domanda alle due indagini. Tale confronto è utile per verificare indirettamente se è lecito ipotizzare che le due rilevazioni siano state condotte nelle medesime condizioni e se, pertanto, il Gdr possa fornire una stima attendibile della Srv. Le formule di stima sono riportate nella Appendice C. Per ciascuna statistica sono stati calcolati i corrispondenti intervalli di confidenza al 95 per cento.

[47] Brancato, Giovanna, Marcello D'Orazio e Marco Fortini. "Response Error Estimation in Presence of Record Linkage Errors: the case of the Italian Population Census". In *Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004)*. Mainz, 24-26 May 2004. Wiesbaden: Statistisches Bundesamt, 2004.

Si noti che le stime sono state calcolate pesando ciascun individuo accoppiato, di una data sezione campione, con il peso finale assegnato alla stessa al termine della fase di aggiustamento dei pesi base della Idc. Nel nostro caso, il peso finale è stato trasformato in modo tale che la somma dei pesi delle sezioni campione eleggibili fosse uguale alla loro numerosità, vale a dire 1.107.

Nelle tabelle sono riportate le stime per Italia e per alcuni dei principali domini di stima:

- le cinque ripartizioni geografiche: Nord-ovest; Nord-est; Centro; Sud; Isole;
- l'ampiezza demografica dei comuni: con meno di 10 mila abitanti; da 10 mila a 100 mila abitanti; con più di 100 mila abitanti con esclusione dei comuni metropolitani; comuni metropolitani (Bari, Bologna, Cagliari, Catania, Firenze, Genova, Roma, Milano, Napoli, Palermo, Torino, Venezia).

Prima di passare all'analisi dei risultati relativi alla stima delle statistiche in questione conviene tener presente che le analisi sono state condotte sui dati raccolti in entrambe le indagini, Cpa e Idc, prima che gli stessi fossero sottoposti alla fase di controllo e correzione. A tal proposito, vale la pena osservare che le due indagini hanno previsto diverse modalità di registrazione dei dati; il Censimento ha utilizzato la lettura ottica dei questionari cartacei. Per la Idc i questionari cartacei sono stati sottoposti a registrazione manuale e pertanto esiste la possibilità che siano stati introdotti degli errori di registrazione (ad esempio in alcuni casi è stata riscontrata la presenza di valori al di fuori del campo di definizione di una variabile).

Nel paragrafo che segue sono riportate le stime delle statistiche sugli errori di risposta, le tabelle in cui si confrontano i valori raccolti nella Idc rispetto a quelli del Cpa sono riportate nell'appendice B.

3.3 - I risultati delle analisi per la stima della variabilità di risposta

3.3.1 - I principali risultati ottenuti per il Gross difference rate e per l'indice di inconsistenza

Nella tavola che segue sono sintetizzati i risultati su tutte le variabili prese in esame in relazione alla misura del *Gross difference rate*, ossia il tasso di discordanza nelle risposte alle due rilevazioni, per ampiezza demografica dei comuni e ripartizione territoriale e per il totale Italia (Tavola 3.3). Se si considera la variabile sesso come un valore di riferimento delle discrepanze che sono connaturate nei dati, si osserva che anche altre variabili, quali la cittadinanza e lo stato civile prima dell'ultimo matrimonio, hanno un livello di disaccordo molto limitato. Si pongono in un livello di errore intermedio variabili quali relazione di parentela, età calcolata, luogo di nascita, stato civile, attività lavorativa a tempo pieno o parziale, tipo di rapporto di lavoro. Decisamente maggiore, attestandosi su più del 10 per cento di valori discordanti, è l'errore per le variabili titolo di studio e condizione professionale. L'errore è quasi sistematicamente maggiore nei comuni metropolitani e nella ripartizione centrale, a indicazione di una minore accuratezza delle risposte, forse dovuta ad una minore disponibilità di tempo per la compilazione del questionario.

Per l'indice di inconsistenza, si fa riferimento alla regola empirica riportata nel paragrafo precedente. Nell'interpretare i risultati di questo indice, bisogna però ricordare che, se da una parte è una misura che consente confronti tra variabili diverse perché è "aggiustata" rispetto alla variabilità del fenomeno, dall'altra risulta molto instabile in presenza di quesiti con variabilità molto bassa o molto alta (Tavola 3.4).

Secondo la suddetta regola empirica presentano errore basso quei quesiti per i quali l'indice è inferiore a 20 per cento e quindi tutte le variabili prese in considerazione, ad eccezione dello stato civile prima dell'ultimo matrimonio e del tipo di rapporto di lavoro. Questo risultato sembrerebbe in contraddizione con quello precedente, ma si giustifica solo considerando la bassissima variabilità delle risposte a questo quesito, come è desumibile dall'analisi dettagliata condotta sugli stessi (Cfr. sottoparagrafo 3.3.2).

Tavola 3.3 - Stime del *Gross difference rate* per ciascuna delle domande del censimento prese in considerazione *(valori percentuali)*

DOMANDE	Stime per ampiezza demografica dei comuni				Stime per ripartizione territoriale					
	Fino a 10.000	10.000-100.000	100.000 e più (a)	Comuni metropolitani	Nord-ovest	Nord-est	Centro	Sud	Isole	**Italia**
Relazione di parentela	4,48	4,70	5,32	8,69	4,87	5,04	7,09	4,77	4,52	**5,26**
Sesso	0,85	0,78	0,70	0,88	0,62	0,81	0,89	0,93	0,85	**0,81**
Data di nascita (stringa gg/mm/aaaa)	6,28	6,63	7,23	7,64	5,65	6,15	7,54	7,28	7,44	**6,71**
Data di nascita (età calcolata)	2,17	2,25	2,29	2,49	1,97	1,95	2,44	2,59	2,45	**2,26**
Luogo di nascita	1,52	2,04	2,20	2,00	1,37	1,85	2,11	2,42	1,58	**1,88**
Cittadinanza	0,22	0,14	0,15	0,40	0,27	0,21	0,27	0,14	0,07	**0,21**
Stato civile	1,77	1,65	1,76	2,45	1,82	1,71	2,02	1,81	1,65	**1,81**
Data di matrimonio (stringa mm/aaaa)	7,89	8,36	10,34	13,02	7,71	8,98	10,31	9,45	9,28	**9,02**
Stato civile prima dell'ultimo matrimonio	0,80	0,84	1,12	1,42	1,10	0,85	0,93	0,92	0,64	**0,93**
Titolo di studio	10,34	9,90	10,90	13,86	11,67	13,68	14,28	12,36	11,83	**12,73**
Condizione professionale	10,96	11,74	11,96	15,76	9,54	9,65	13,47	14,11	15,70	**12,09**
Attività lavorativa a tempo pieno o parziale	3,25	3,35	3,61	5,05	3,13	3,21	3,68	4,53	3,85	**3,57**
Posizione nella professione	7,22	7,62	9,16	9,07	6,42	8,38	9,68	6,91	8,52	**7,81**
Tipo rapporto di lavoro	5,95	5,76	5,27	5,61	4,96	5,77	5,96	6,43	6,83	**5,76**

(a) Esclusi i comuni metropolitani.

Tavola 3.4 - Stime dell'Indice di inconsistenza per ciascuna delle domande del censimento prese in considerazione *(valori percentuali)*

DOMANDE	Stime per ampiezza demografica dei comuni				Stime per ripartizione territoriale					
	Fino a 10.000	10.000-100.000	100.000 e più (a)	Comuni metropolitani	Nord-Ovest	Nord-Est	Centro	Sud	Isole	**Italia**
Relazione di parentela	6,33	6,63	7,55	12,23	6,96	7,09	9,86	6,78	6,42	**7,41**
Sesso	1,69	1,56	1,40	1,77	1,25	1,63	1,77	1,85	1,70	**1,62**
Data di nascita (stringa gg/mm/aaaa)	-	-	-	-	-	-	-	-	-	**-**
Data di nascita (età calcolata)	2,32	2,41	2,46	2,66	2,11	2,09	2,62	2,77	2,62	**2,42**
Luogo di nascita	3,29	3,91	4,28	4,10	2,81	3,57	3,93	4,64	3,19	**3,57**
Cittadinanza	4,26	4,71	3,68	7,41	4,99	3,64	5,88	7,90	3,02	**4,93**
Stato civile	2,97	2,83	2,92	3,97	3,01	2,84	3,41	3,13	2,84	**3,06**
Data di matrimonio (stringa mm/aaaa)	-	-	-	-	-	-	-	-	-	**-**
Stato civile prima dell'ultimo matrimonio	22,74	26,79	22,65	24,15	23,96	21,69	23,99	32,07	21,00	**24,51**
Titolo di studio	13,62	12,74	13,81	17,71	12,99	14,99	15,15	13,31	12,59	**13,78**
Condizione professionale	14,95	15,66	16,58	20,72	13,71	14,48	18,60	17,44	19,48	**16,24**
Attività lavorativa a tempo pieno o parziale	19,94	19,67	22,16	26,61	17,31	19,18	21,58	28,51	24,29	**20,98**
Posizione nella professione	16,18	18,19	21,31	23,82	16,18	19,51	21,29	16,51	18,88	**18,36**
Tipo rapporto di lavoro	28,44	27,98	31,12	33,16	29,80	34,33	29,87	24,09	27,35	**28,95**

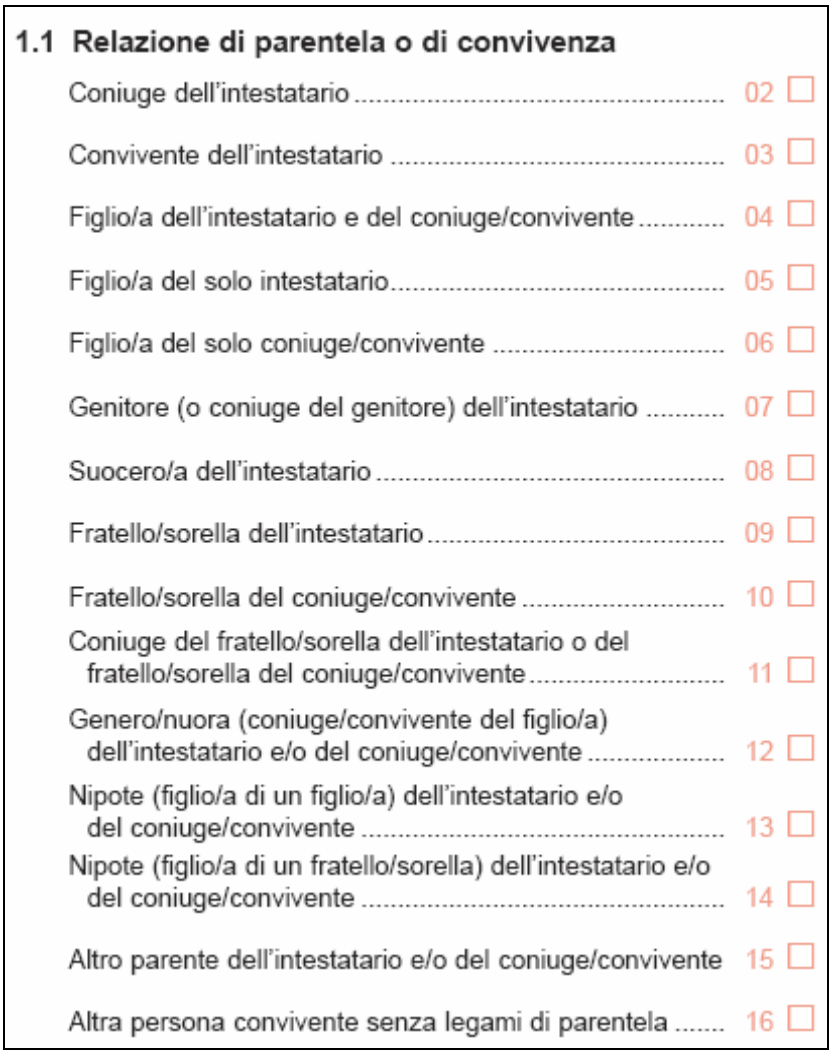
(a) Esclusi i comuni metropolitani.

3.3.2 - Le stime per i singoli quesiti del Foglio di famiglia

Il confronto tra le distribuzioni della variabile "Relazione di parentela o di convivenza" al Censimento e all'Idc, non evidenzia particolari differenze. La stima del Gdr per il totale Italia (riga "Aggregato" nella tavola 3.5) è di 5,26 per cento. A livello di domini territoriali, valori maggiori del Gdr si ottengono solo per la ripartizione territoriale Centro e per i comuni metropolitani. Analogo andamento si osserva per le stime aggregate dell'indice di inconsistenza, che comunque risultano sempre al di sotto della soglia critica del 20 per cento. Quando invece si osserva l'indice di inconsistenza a livello di singole modalità di risposta, con l'eccezione delle modalità "01", "02" e "04" (Figura 3.1), si ottengono stime elevate, che però non devono trarre

in inganno, dato che si tratta di modalità di risposta poco frequenti (la loro incidenza è sempre al di sotto del 5 per cento).

Figura 3.1 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sulle notizie anagrafiche



1.1 Relazione di parentela o di convivenza

- Coniuge dell'intestatario ..... 02 ☐
- Convivente dell'intestatario ..... 03 ☐
- Figlio/a dell'intestatario e del coniuge/convivente ..... 04 ☐
- Figlio/a del solo intestatario ..... 05 ☐
- Figlio/a del solo coniuge/convivente ..... 06 ☐
- Genitore (o coniuge del genitore) dell'intestatario ..... 07 ☐
- Suocero/a dell'intestatario ..... 08 ☐
- Fratello/sorella dell'intestatario ..... 09 ☐
- Fratello/sorella del coniuge/convivente ..... 10 ☐
- Coniuge del fratello/sorella dell'intestatario o del fratello/sorella del coniuge/convivente ..... 11 ☐
- Genero/nuora (coniuge/convivente del figlio/a) dell'intestatario e/o del coniuge/convivente ..... 12 ☐
- Nipote (figlio/a di un figlio/a) dell'intestatario e/o del coniuge/convivente ..... 13 ☐
- Nipote (figlio/a di un fratello/sorella) dell'intestatario e/o del coniuge/convivente ..... 14 ☐
- Altro parente dell'intestatario e/o del coniuge/convivente 15 ☐
- Altra persona convivente senza legami di parentela ..... 16 ☐

Tavola 3.5 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per la domanda relativa alla "Relazione di parentela o di convivenza" *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
01 (a)	39,15	39,21	-0,06	-0,13	0,00	1,79	1,73	1,86	3,76	3,62	3,90
02 (b)	23,29	23,42	-0,13	-0,19	-0,07	1,43	1,38	1,49	4,00	3,84	4,17
03	1,05	0,91	0,14	0,11	0,17	0,50	0,46	0,53	25,69	23,96	27,55
04	28,29	28,48	-0,18	-0,27	-0,10	2,67	2,59	2,75	6,58	6,38	6,78
05	3,85	3,71	0,14	0,07	0,22	2,29	2,22	2,37	31,52	30,51	32,56
06	0,40	0,33	0,07	0,04	0,10	0,36	0,33	0,39	49,04	45,18	53,23
07	0,82	0,78	0,04	0,02	0,07	0,29	0,27	0,32	18,31	16,72	20,06
08	0,41	0,40	0,01	0,00	0,03	0,09	0,07	0,10	10,90	9,24	12,86
09	0,73	0,71	0,01	-0,01	0,03	0,19	0,17	0,22	13,56	12,12	15,16
10	0,10	0,09	0,00	-0,01	0,02	0,06	0,04	0,07	29,59	24,00	36,48
11	0,04	0,04	0,00	-	-	-	-	-	-	-	-
12	0,45	0,46	-0,01	-0,03	0,01	0,14	0,12	0,16	15,28	13,39	17,42
13	0,92	0,92	-0,01	-0,03	0,02	0,27	0,25	0,30	14,95	13,61	16,43
14	0,11	0,11	0,00	-0,02	0,02	0,10	0,09	0,12	46,96	40,32	54,71
15	0,20	0,20	0,00	-0,02	0,02	0,14	0,12	0,16	35,38	31,01	40,37
16	0,19	0,21	-0,02	-0,04	0,00	0,15	0,13	0,17	35,83	31,50	40,75
Aggregato	**-**	**-**	**-**	**-**	**-**	**5,26**	**5,15**	**5,37**	**7,41**	**7,25**	**7,57**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	4,87	4,67	5,08	6,96	6,66	7,26
Nord-est	-	-	-	-	-	5,04	4,79	5,28	7,09	6,74	7,46
Centro	-	-	-	-	-	7,09	6,80	7,38	9,86	9,45	10,29
Sud	-	-	-	-	-	4,77	4,56	4,99	6,78	6,48	7,10
Isole	-	-	-	-	-	4,52	4,23	4,82	6,42	6,01	6,87
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	4,48	4,30	4,65	6,33	6,07	6,59
10.000-100.000	-	-	-	-	-	4,70	4,55	4,86	6,63	6,41	6,86
100.000 e più (c)	-	-	-	-	-	5,32	4,93	5,72	7,55	7,00	8,15
Comuni metropolitani	-	-	-	-	-	8,69	8,32	9,05	12,23	11,71	12,78

(a) Intestatario del Foglio di famiglia.
(b) Per la decodifica delle modalità di risposta si veda la Figura 3.1 sulle Relazioni di parentela.
(c) Esclusi i comuni metropolitani.

L'analisi del Ndr evidenzia che la distribuzione delle modalità di risposta nelle due occasioni di indagine non presenta sostanziali differenze e pertanto si può ritenere valida l'ipotesi che le due rilevazioni siano state condotte nelle medesime condizioni, almeno per quel che riguarda la domanda in questione.

La variabile Sesso, può essere considerata come un *baseline*, e fornire i valori di riferimento per l'errore di risposta (Figura 3.2). La differenza tra le stime al Censimento e all'Idc per le due modalità del sesso è dell'ordine dello 0,10 per cento. Il Gdr generale è pari a 0,81 per cento e l'indice di inconsistenza è pari a 1,62 per cento. Per questa variabile non sembrano emergere differenze rilevanti nelle stime per i diversi domini territoriali (Tavola 3.6).

Figura 3.2 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sul sesso dell'individuo

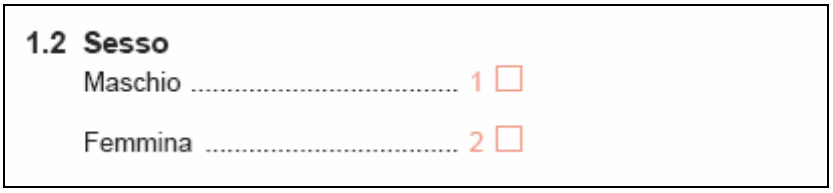


Tavola 3.6 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per la domanda "Sesso" *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 – Maschio	48,23	48,33	-0,10	-0,14	-0,06	0,81	0,77	0,85	1,62	1,53	1,71
2 – Femmina	51,77	51,67	0,10	0,06	0,14	0,81	0,77	0,85	1,62	1,53	1,71
Aggregato	**-**	**-**	**-**	**-**	**-**	**0,81**	**0,77**	**0,85**	**1,62**	**1,53**	**1,71**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	0,62	0,55	0,70	1,25	1,11	1,41
Nord-est	-	-	-	-	-	0,81	0,71	0,91	1,63	1,43	1,84
Centro	-	-	-	-	-	0,89	0,78	0,99	1,77	1,57	2,00
Sud	-	-	-	-	-	0,93	0,83	1,02	1,85	1,67	2,06
Isole	-	-	-	-	-	0,85	0,72	0,98	1,70	1,45	1,98
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	0,85	0,77	0,92	1,69	1,54	1,86
10.000-100.000	-	-	-	-	-	0,78	0,71	0,84	1,56	1,43	1,69
100.000 e più (a)	-	-	-	-	-	0,70	0,55	0,84	1,40	1,14	1,73
Comuni metropolitani	-	-	-	-	-	0,88	0,76	1,00	1,77	1,54	2,03

(a) Esclusi comuni metropolitani.

Per la data di nascita (Figura 3.3) in primo luogo è stata valutata la variabilità di risposta e le discrepanze tra le due occasioni di rilevazione attraverso il confronto sulla stringa "giorno/mese/anno". Successivamente, la stringa è stata utilizzata per calcolare l'età in anni compiuti al 21 ottobre 2001 (data di riferimento per il Cpa e per la Idc) e sono state considerate delle classi di età quinquennali.

Figura 3.3 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sulla data di nascita

Se si analizza la perfetta corrispondenza tra le date di nascita attraverso il confronto sulle stringhe, si osserva che la stima dell'errore variabile di risposta (Gdr) è pari a 6,71 per cento (Intervallo di confidenza al 95 per cento: 6,58-6,83) con un gradiente Nord-Sud, e cioè con errore più elevato nel Sud e nelle Isole, e con una maggiore tendenza all'errore nei grandi centri e nei comuni metropolitani (Tavola 3.7).

In questo caso non è possibile calcolare il Ndr e l'indice di inconsistenza perché la variabile stringa analizzata è una variabile categoriale che presenta un numero troppe grande di modalità di risposta e quindi risulterebbe troppo oneroso esplorare la distribuzione delle concordanze/discordanze per ciascuna modalità.

Come atteso, il livello di errore diminuisce notevolmente quando si considerano le classi di età in anni compiuti. Sull'età calcolata e raggruppata in classi, per il Gdr si ottiene il valore di 2,26 per cento contro il precedente 6,71 per cento, inoltre aumenta la confidenza nei risultati del Censimento essendo l'intervallo di confidenza più stretto (I.C. 95 per cento: 2,19-2,34). Si continua ad osservare l'andamento geografico e stime moderatamente più ampie per i comuni di maggiore ampiezza demografica.

Per l'età in classi (Tavola 3.8), la stima dell'indice di inconsistenza è molto bassa, attestandosi al 2,40 per cento, poco più alta di quella ottenuta per la variabile sesso (1,62 per cento).

Tavola 3.7 - Stime del *Gross difference rate* e relativi intervalli di confidenza al 95 per cento per la data di nascita espressa come stringa giorno/mese/anno *(valori percentuali)*

DATA DI NASCITA	Gross difference rate		
	Stima	Intervallo di confidenza	
		Lim. inferiore	Lim. superiore
Aggregato	**6,71**	**6,58**	**6,83**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE			
Nord-ovest	5,65	5,43	5,87
Nord-est	6,15	5,88	6,42
Centro	7,54	7,24	7,84
Sud	7,28	7,02	7,54
Isole	7,44	7,06	7,81
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI			
Meno 10.000	6,28	6,07	6,48
10.000-100.000	6,63	6,45	6,81
Oltre 100.000 (a)	7,23	6,78	7,68
Comuni metropolitani	7,64	7,30	7,98

(a) Esclusi i comuni metropolitani.

Tavola 3.8 - Stime del *Net difference rate*, *Gross difference rate*, Indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per classi di età in anni compiuti *(valori percentuali)*

CLASSI DI ETÀ CALCOLATA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. onf.	Lim. sup.		Lim. inf.	Lim. sup.
0-5	5,39	5,45	-0,06	-0,09	-0,04	0,31	0,29	0,34	3,07	2,81	3,35
6-10	4,65	4,65	-0,01	-0,03	0,02	0,25	0,22	0,27	2,78	2,52	3,06
11-14	4,02	4,05	-0,03	-0,05	-0,01	0,21	0,19	0,23	2,72	2,44	3,02
15-19	5,10	5,13	-0,03	-0,05	0,00	0,25	0,23	0,28	2,60	2,36	2,87
20-24	6,00	6,00	0,00	-0,03	0,02	0,31	0,28	0,34	2,72	2,49	2,97
25-29	7,51	7,54	-0,03	-0,06	0,00	0,32	0,29	0,35	2,28	2,09	2,48
30-34	7,70	7,76	-0,06	-0,09	-0,03	0,38	0,35	0,41	2,65	2,44	2,87
35-39	7,99	8,02	-0,02	-0,05	0,01	0,32	0,29	0,35	2,16	1,98	2,36
40-44	6,97	6,97	0,00	-0,03	0,03	0,33	0,30	0,36	2,55	2,34	2,77
45-49	6,60	6,61	-0,01	-0,03	0,02	0,27	0,25	0,30	2,20	2,00	2,42
50-54	6,94	6,94	0,00	-0,03	0,02	0,28	0,25	0,31	2,16	1,97	2,37
55-59	5,95	6,00	-0,05	-0,07	-0,02	0,23	0,21	0,25	2,03	1,83	2,25
60-64	6,20	6,21	-0,02	-0,04	0,01	0,24	0,22	0,27	2,09	1,89	2,31
65-69	5,67	5,65	0,02	0,00	0,04	0,17	0,15	0,19	1,62	1,44	1,82
70-74	4,97	4,95	0,02	-0,01	0,04	0,19	0,17	0,22	2,05	1,84	2,30
75 e oltre	8,28	8,06	0,22	0,19	0,25	0,38	0,35	0,41	2,55	2,35	2,76
Aggregato	-	-	-	-	-	**2,26**	**2,19**	**2,34**	**2,42**	**2,34**	**2,50**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	1,97	1,84	2,10	2,11	1,97	2,26
Nord-est	-	-	-	-	-	1,95	1,79	2,10	2,09	1,92	2,26
Centro	-	-	-	-	-	2,44	2,27	2,62	2,62	2,43	2,81
Sud	-	-	-	-	-	2,59	2,43	2,75	2,77	2,60	2,95
Isole	-	-	-	-	-	2,45	2,23	2,68	2,62	2,39	2,87
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	2,17	2,05	2,30	2,32	2,19	2,46
10.000-100.000	-	-	-	-	-	2,25	2,14	2,36	2,41	2,29	2,53
100.000 e più (b)	-	-	-	-	-	2,29	2,03	2,55	2,46	2,19	2,76
Comuni metropolitani	-	-	-	-	-	2,49	2,29	2,69	2,66	2,46	2,89

(a) L'età è calcolata a partire dalla stringa giorno/mese/anno.
(b) Esclusi i comuni metropolitani.

Il livello di errore stimato per il luogo di nascita (Figura 3.4), così come rilevato al Censimento, risulta piuttosto basso (Gdr pari a 1,88 per cento; indice di inconsistenza pari a 3,6 per cento). A livello territoriale il valore del Gdr per il Nord-ovest risulta un po' più basso che per le altre ripartizioni territoriali (Tavola 3.9).

Figura 3.4 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sul luogo di nascita

Tavola 3.9 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per il luogo di nascita *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 - In questo comune	42,96	42,33	0,62	0,56	0,69	1,79	1,72	1,85	3,65	3,52	3,79
2 - In altro comune italiano	53,58	54,18	-0,61	-0,67	-0,54	1,84	1,78	1,91	3,71	3,57	3,84
3 - All'estero	3,47	3,48	-0,02	-0,03	0,00	0,13	0,11	0,15	1,90	1,65	2,18
Aggregato	**-**	**-**	**-**	**-**	**-**	**1,88**	**1,81**	**1,95**	**3,57**	**3,44**	**3,70**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	1,37	1,26	1,48	2,81	2,59	3,05
Nord-est	-	-	-	-	-	1,85	1,70	2,01	3,57	3,29	3,88
Centro	-	-	-	-	-	2,11	1,94	2,27	3,93	3,63	4,25
Sud	-	-	-	-	-	2,42	2,26	2,57	4,64	4,35	4,95
Isole	-	-	-	-	-	1,58	1,40	1,76	3,19	2,84	3,57
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	**-**	**-**	**-**	**-**	**-**	1,52	1,42	1,63	3,29	3,06	3,52
10.000-100.000	**-**	**-**	**-**	**-**	**-**	2,04	1,94	2,15	3,91	3,71	4,12
100.000 e più (a)	**-**	**-**	**-**	**-**	**-**	2,20	1,95	2,46	4,28	3,80	4,81
Comuni metropolitani	**-**	**-**	**-**	**-**	**-**	2,00	1,81	2,18	4,10	3,74	4,50

(a) Esclusi i comuni metropolitani.

Il Gdr per la cittadinanza (Figura 3.5) è molto basso (anche inferiore al risultato per il sesso); al contrario l'indice di inconsistenza stimato è pari al 4,93 per cento, di poco superiore a quello relativo al luogo di nascita (Tavola 3.10). In ogni caso bisogna avere molta cautela nel commentare le stime ottenute per l'indice di inconsistenza (a livello aggregato e per i vari domini territoriali) data la particolare distribuzione della variabile in esame, che risulta particolarmente concentrata sulla modalità di risposta "Italiana". Si noti che per la modalità di risposta "Apolide" non sono state riportate le stime di Gdr e *I* dato che la stessa si può ritenere una modalità rarissima.

Figura 3.5 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sulla cittadinanza

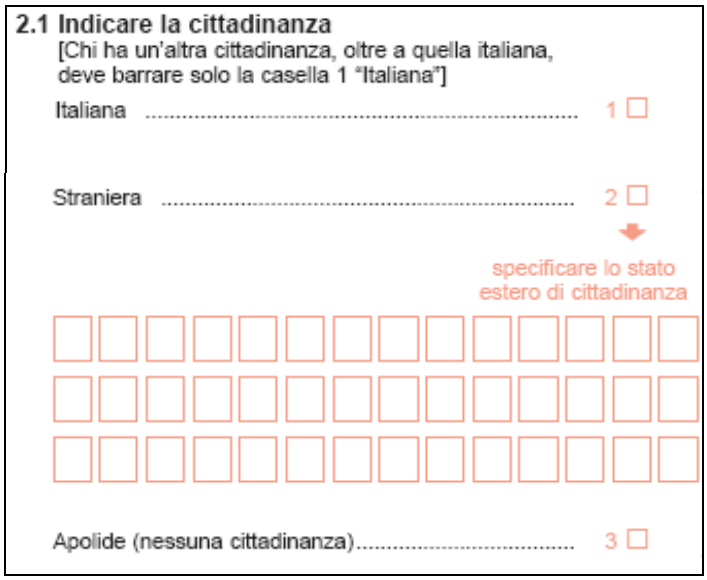


Tavola 3.10 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per la cittadinanza *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallodi confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 – Italiana	97,88	97,86	0,02	0,00	0,05	0,20	0,18	0,22	4,71	4,20	5,29
2 – Straniera	2,11	2,13	-0,02	-0,05	0,00	0,20	0,18	0,22	4,62	4,11	5,19
3 - Apolide	0,01	0,01	0,00	-	-	-	-	-	-	-	-
Aggregato	**-**	**-**	**-**	**-**	**-**	**0,21**	**0,18**	**0,23**	**4,93**	**4,40**	**5,52**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	0,27	0,21	0,32	4,99	4,12	6,04
Nord-est	-	-	-	-	-	0,21	0,16	0,26	3,64	2,84	4,67
Centro	-	-	-	-	-	0,27	0,21	0,33	5,88	4,69	7,36
Sud	-	-	-	-	-	0,14	0,10	0,18	7,90	5,93	10,52
Isole	-	-	-	-	-	0,07	0,03	0,11	3,02	1,76	5,18
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	0,22	0,18	0,26	4,26	3,53	5,13
10.000-100.000	-	-	-	-	-	0,14	0,11	0,17	4,71	3,83	5,79
100.000 e più (a)	-	-	-	-	-	0,15	0,08	0,22	3,68	2,33	5,82
Comuni metropolitani	-	-	-	-	-	0,40	0,32	0,49	7,41	5,98	9,17

(a) Esclusi i comuni metropolitani.

Anche per lo stato civile (Figura 3.6), la variabilità di risposta è piuttosto limitata (Gdr=1,81; I.C. 95 per cento: 1,75-11,88) e l'indice di inconsistenza aggregato, pari al 3,06 per cento, si può ritenere trascurabile (Tavola 3.11). I valori elevati stimati per l'indice di inconsistenza in corrispondenza delle modalità di risposta 3, 4 e 5 sono da ritenersi inattendibili dato che siamo in presenza di modalità poco frequenti (con frequenza inferiore al 5 per cento).

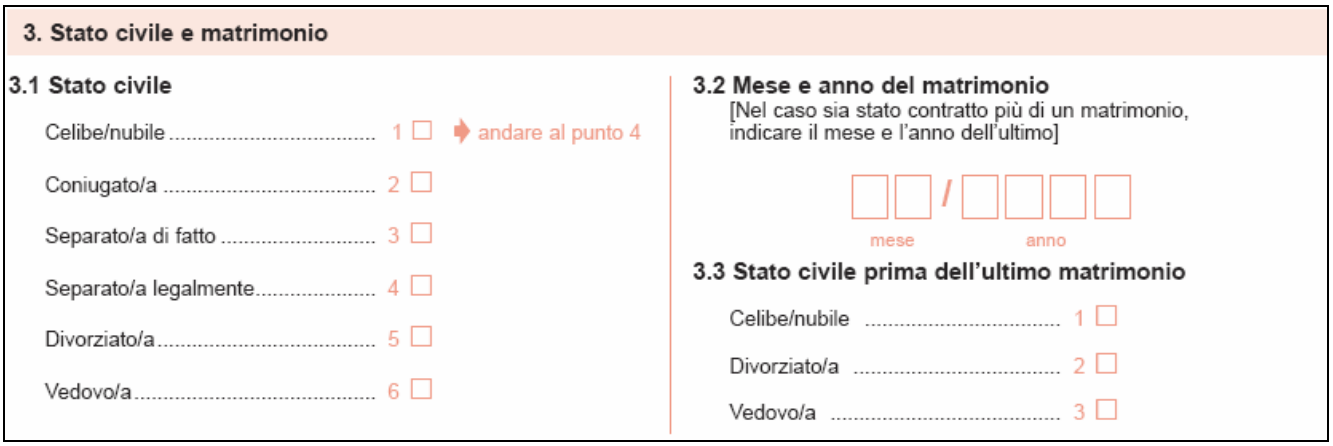
Figura 3.6 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sullo stato civile

Tavola 3.11 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per lo stato civile *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 - Celibe/nubile	39,08	39,11	-0,02	-0,07	0,03	1,01	0,96	1,06	2,12	2,02	2,22
2 – Coniugato/a	49,70	49,78	-0,08	-0,14	-0,03	1,22	1,17	1,28	2,44	2,33	2,55
3 - Separato/a di fatto	0,44	0,41	0,03	0,00	0,06	0,36	0,33	0,39	42,48	39,14	46,09
4 - Separato/a legalmente	1,55	1,46	0,09	0,05	0,12	0,47	0,43	0,50	15,73	14,64	16,90
5 - Divorziato/a	1,24	1,21	0,03	0,00	0,05	0,30	0,27	0,32	12,22	11,16	13,37
6 - Vedovo/a	7,98	8,01	-0,02	-0,05	0,00	0,26	0,24	0,29	1,78	1,62	1,96
Aggregato	**-**	**-**	**-**	**-**	**-**	**1,81**	**1,75**	**1,88**	**3,06**	**2,95**	**3,17**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	1,82	1,69	1,94	3,01	2,80	3,23
Nord-est	-	-	-	-	-	1,71	1,56	1,85	2,84	2,61	3,10
Centro	-	-	-	-	-	2,02	1,86	2,18	3,41	3,15	3,69
Sud	-	-	-	-	-	1,81	1,68	1,94	3,13	2,91	3,37
Isole	-	-	-	-	-	1,65	1,47	1,83	2,84	2,54	3,18
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	1,77	1,65	1,88	2,97	2,79	3,17
10.000-100.000	-	-	-	-	-	1,65	1,56	1,74	2,83	2,68	3,00
100.000 e più (a)	-	-	-	-	-	1,76	1,53	1,99	2,92	2,56	3,33
Comuni metropolitani	-	-	-	-	-	2,45	2,25	2,65	3,97	3,66	4,31

(a) Esclusi i comuni metropolitani.

Figura 3.7 - Foglio di famiglia del Censimento della popolazione 2001. Quesiti sullo stato civile e sullo stato civile prima del matrimonio



Alle domande 3.2 e 3.3 (Figura 3.7) dovevano rispondere solo quanti avessero dichiarato uno stato civile diverso dalla modalità "celibe/nubile". Pertanto, ai fini del calcolo del Gdr e delle altre stime, l'attenzione è stata limitata a quel sottoinsieme di individui *linked* che avessero dichiarato di essere in una condizione diversa da "celibe/nubile" e tali che la modalità di risposta fosse identica in entrambe le occasioni di indagine. In pratica sono stati eliminati dalle analisi le unità con valori discrepanti per lo stato civile o che ad entrambe le rilevazioni avessero dichiarato stato civile pari a "celibe/nubile". In tal modo, le stime per errore di risposta alle domande in questione sono da intendersi "al netto" degli errori di risposta commessi alla domanda filtro che le precede.

Maggiore errore di risposta variabile si osserva sulla data di matrimonio (Gdr pari a 9,02; I.C. al 95 per cento: 8,84-9,21), analizzata come stringa "mese/anno". Ciò significa che circa il 9 per cento delle date di matrimonio riportate al Censimento e alla Idc erano discrepanti in almeno una delle due componenti della data (mese o anno). Come per altre variabili, l'errore di risposta risulta maggiore al Centro Italia e nei grandi comuni e nei comuni metropolitani (Tavola 3.12).

Tavola 3.12 - Stime del *Gross difference rate* e relativi intervalli di confidenza al 95 per cento per la data di matrimonio (mese/anno) per individui che abbiano dichiarato di non essere "celibe/nubile" *(valori percentuali)*

DATA DI MATRIMONIO	Gross difference rate		
	Stima	Intervallo di confidenza	
		Lim. inferiore	Lim. superiore
Aggregato	**9,02**	**8,84**	**9,21**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE			
Nord-ovest	7,71	7,38	8,04
Nord-est	8,98	8,56	9,40
Centro	10,31	9,86	10,76
Sud	9,45	9,05	9,85
Isole	9,28	8,71	9,85
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI			
Fino a 10.000	7,89	7,59	8,19
10.000-100.000	8,36	8,09	8,63
100.000 e più (a)	10,34	9,65	11,04
Comuni metropolitani	13,02	12,44	13,61

(a) Esclusi i comuni metropolitani.

Lo stato civile prima dell'ultimo matrimonio presenta un livello di errore di risposta molto basso (Tavola 3.13). Come più volte detto, nell'interpretazione dell'indice di inconsistenza deve essere usata molta cautela, dato che la distribuzione delle risposte è concentrata quasi esclusivamente sulla prima modalità di risposta.

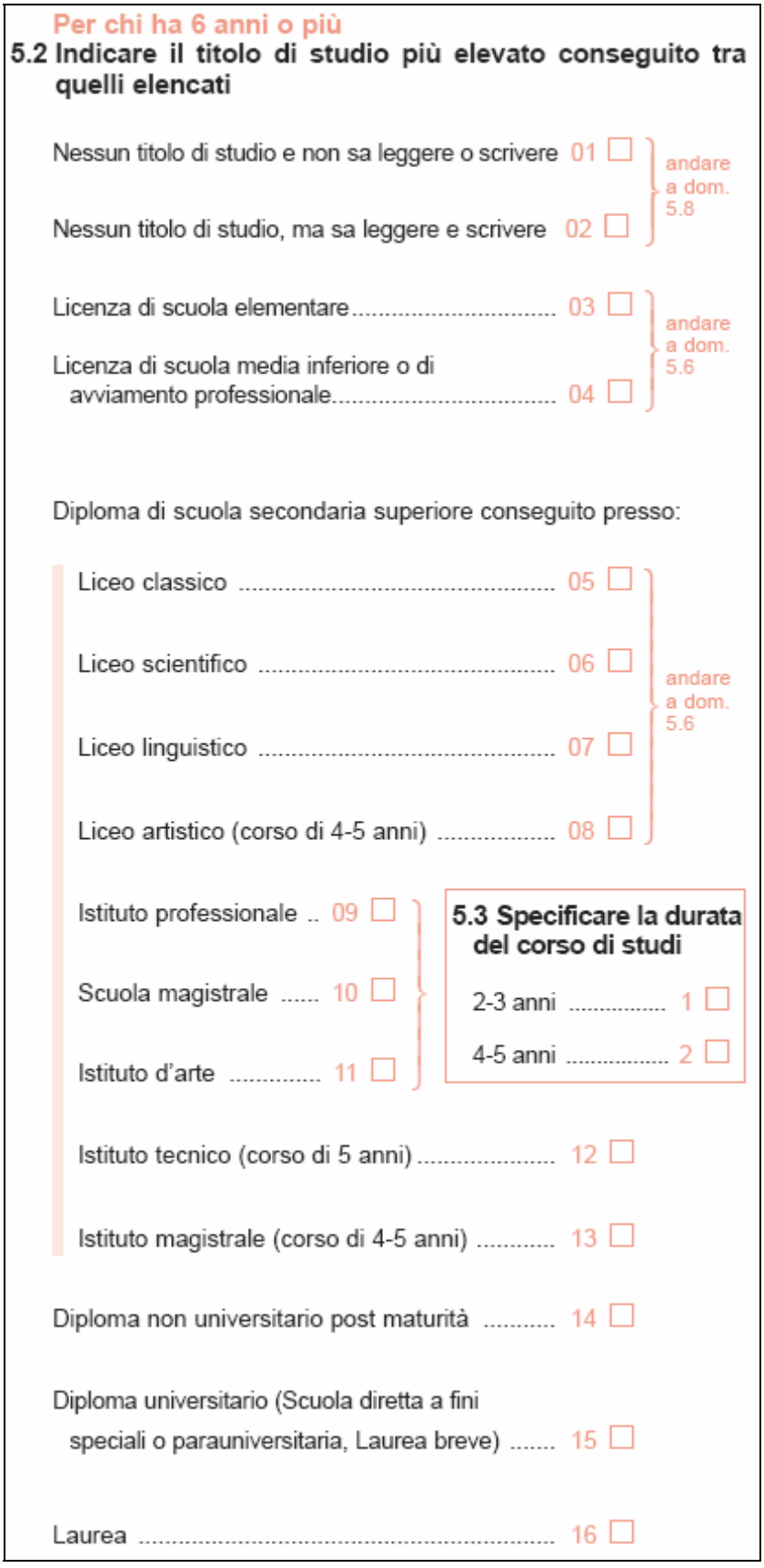
Tavola 3.13 - Stime *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per lo "stato civile prima dell'ultimo matrimonio" per individui che non si siano dichiarati "celibe/nubile" *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 - Celibe/nubile	97,87	98,28	-0,40	-0,47	-0,34	0,91	0,85	0,98	24,10	22,44	25,87
2 - Divorziato/a	1,33	1,16	0,17	0,12	0,22	0,48	0,44	0,53	19,55	17,73	21,55
3 - Vedovo/a	0,79	0,55	0,24	0,19	0,28	0,46	0,42	0,51	34,53	31,26	38,15
Aggregato	**-**	**-**	**-**	**-**	**-**	**0,93**	**0,86**	**0,99**	**24,51**	**22,84**	**26,29**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	1,10	0,97	1,23	23,96	21,23	27,04
Nord-est	-	-	-	-	-	0,85	0,71	0,99	21,69	18,43	25,53
Centro	-	-	-	-	-	0,93	0,79	1,08	23,99	20,45	28,14
Sud	-	-	-	-	-	0,92	0,78	1,06	32,07	27,50	37,41
Isole	-	-	-	-	-	0,64	0,48	0,80	21,00	16,27	27,12
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	0,80	0,69	0,90	22,74	19,92	25,95
10.000-100.000	-	-	-	-	-	0,84	0,74	0,93	26,79	23,99	29,92
100.000 e più (a)	-	-	-	-	-	1,12	0,87	1,37	22,65	18,15	28,27
Comuni metropolitani	-	-	-	-	-	1,42	1,21	1,64	24,15	20,75	28,12

(a) Esclusi i comuni metropolitani.

Le analisi condotte per il titolo di studio hanno preso in considerazione solo gli individui *linked* la cui età in anni compiuti, calcolata a partire dalla data di nascita, fosse maggiore di cinque anni. In primo luogo si sono considerate le 16 modalità di risposta originarie (Figura 3.8). Il Gdr complessivo è risultato pari al 12,7 per cento, ovvero un valore piuttosto elevato se confrontato con quelli registrati per le variabili anagrafiche o per quelle relative allo stato civile. Analogo discorso vale per l'indice di inconsistenza attestatosi al 15,7 per cento (Tavola 3.14). Se si scende a livello dei domini territoriali si nota che per i comuni metropolitani l'Indice di inconsistenza stimato è di poco superiore al 20 per cento, che rappresenta la soglia empirica al di sopra della quale l'impatto degli errori di risposta casuali sulle stime non può più ritenersi trascurabile.

Figura 3.8 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sul titolo di studio



Per chi ha 6 anni o più

5.2 Indicare il titolo di studio più elevato conseguito tra quelli elencati

Nessun titolo di studio e non sa leggere o scrivere 01 ☐
Nessun titolo di studio, ma sa leggere e scrivere 02 ☐
→ andare a dom. 5.8

Licenza di scuola elementare ..... 03 ☐
Licenza di scuola media inferiore o di avviamento professionale ..... 04 ☐
→ andare a dom. 5.6

Diploma di scuola secondaria superiore conseguito presso:

Liceo classico ..... 05 ☐
Liceo scientifico ..... 06 ☐
Liceo linguistico ..... 07 ☐
Liceo artistico (corso di 4-5 anni) ..... 08 ☐
→ andare a dom. 5.6

Istituto professionale .. 09 ☐
Scuola magistrale ..... 10 ☐
Istituto d'arte ..... 11 ☐

5.3 Specificare la durata del corso di studi
2-3 anni ..... 1 ☐
4-5 anni ..... 2 ☐

Istituto tecnico (corso di 5 anni) ..... 12 ☐

Istituto magistrale (corso di 4-5 anni) ..... 13 ☐

Diploma non universitario post maturità ..... 14 ☐

Diploma universitario (Scuola diretta a fini speciali o parauniversitaria, Laurea breve) ..... 15 ☐

Laurea ..... 16 ☐

Se si considerano le singole modalità di risposta, limitando l'attenzione a quelle maggiormente frequenti (con frequenza superiore al 5 per cento), ci si accorge che la stima per l'indice di inconsistenza in corrispondenza della modalità "02" ("Nessun titolo di studio, ma sa leggere e scrivere") risulta pari al 22,4 per cento, mentre per la categoria "09" ("Diploma di Istituto professionale") si attesta intorno al 27,5 per cento. Valori ben al di sopra della soglia del 20 per cento, che stanno ad indicare una certa difficoltà degli individui a classificarsi in tali categorie.

Tavola 3.14 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per il "titolo di studio più elevato" per individui con età in anni compiuti superiore a 5 *(valori percentuali)*

MODALITÀ DI RISPOSTA (a)	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
01	1,60	1,47	0,13	0,08	0,18	0,93	0,88	0,98	30,76	29,16	32,45
02	8,55	8,78	-0,23	-0,33	-0,13	3,55	3,45	3,65	22,43	21,83	23,06
03	26,19	25,99	0,21	0,09	0,33	5,44	5,32	5,56	14,10	13,80	14,42
04	30,94	30,21	0,73	0,61	0,85	5,43	5,31	5,55	12,79	12,51	13,08
05	1,68	1,72	-0,03	-0,07	0,00	0,44	0,40	0,47	13,03	12,05	14,09
06	2,67	2,64	0,03	0,00	0,07	0,45	0,42	0,49	8,71	8,07	9,41
07	0,48	0,49	-0,02	-0,04	0,00	0,16	0,14	0,18	16,47	14,47	18,75
08	0,34	0,34	0,00	-0,02	0,02	0,12	0,10	0,14	18,05	15,56	20,92
09	6,08	6,57	-0,49	-0,59	-0,40	3,26	3,17	3,35	27,50	26,72	28,30
10	1,52	1,50	0,03	-0,03	0,08	1,16	1,10	1,21	38,85	37,03	40,76
11	0,42	0,40	0,03	0,00	0,05	0,16	0,14	0,19	19,88	17,50	22,59
12	10,50	10,70	-0,20	-0,28	-0,13	2,12	2,05	2,20	11,20	10,81	11,61
13	1,87	1,94	-0,07	-0,13	-0,02	1,14	1,08	1,19	30,35	28,91	31,85
14	0,37	0,37	0,01	-0,02	0,04	0,35	0,32	0,38	47,51	43,54	51,84
15	0,58	0,54	0,04	0,01	0,06	0,29	0,26	0,32	25,98	23,60	28,60
16	6,20	6,34	-0,15	-0,18	-0,11	0,45	0,42	0,49	3,85	3,57	4,16
Aggregato	**-**	**-**	**-**	**-**	**-**	**12,7**	**12,55**	**12,90**	**15,71**	**15,50**	**15,93**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	11,67	11,35	11,99	14,78	14,38	15,20
Nord-est	-	-	-	-	-	13,68	13,27	14,08	16,96	16,46	17,47
Centro	-	-	-	-	-	14,28	13,86	14,70	17,30	16,80	17,82
Sud	-	-	-	-	-	12,36	12,02	12,71	15,11	14,69	15,54
Isole	-	-	-	-	-	11,83	11,35	12,31	14,66	14,08	15,28
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	11,61	11,32	11,90	14,85	14,49	15,22
10.000-100.000	-	-	-	-	-	11,88	11,63	12,14	14,61	14,31	14,93
100.000 e più (b)	-	-	-	-	-	14,06	13,43	14,69	16,73	16,00	17,50
Comuni metropolitani	-	-	-	-	-	17,12	16,61	17,63	20,63	20,02	21,25

(a) Per la decodifica delle modalità di risposta si veda la Figura 3.8 sul titolo di studio più alto conseguito.
(b) Esclusi i comuni metropolitani.

Le stime sono state ricalcolate prendendo in considerazione un accorpamento delle 16 modalità di risposta originarie nelle seguenti categorie:
01 - Analfabeti
02 - Alfabeti privi di titolo di studio
03 - Licenza elementare

04 - Licenza media
05 - Diploma scolastico di qualifica (corso scolastico di 2-3 anni)
06 - Diploma di maturità (corso scolastico di 4-5 anni)
07 - Diploma terziario di tipo non universitario
08 - Diploma universitario
09 - Diploma di laurea

Per desumere l'appartenenza alle nuove categorie "05" e "06" si è tenuto conto della domanda relativa alla durata del corso di studi.

Come si può notare dalla tavola 3.15 questo accorpamento delle modalità di risposta porta ad una leggera diminuzione dei valori degli indicatori a livello aggregato complessivo (sia a livello Italia che a livello di ripartizioni territoriali). Se invece si considerano i valori degli indici per ciascuna nuova modalità di risposta si può osservare che permane la criticità per la modalità di risposta "02" (che non è stata variata rispetto al caso originario) mentre scompare il problema relativo alla modalità di risposta originaria "Diploma di Istituto professionale" che adesso confluisce nelle nuove modalità "05" o "06" a seconda della durata della corso di scuola secondaria superiore.

Tavola 3.15 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per il titolo di studio più elevato ricodificato in 9 modalità di risposta Individui con età in anni compiuti superiore a 5 *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
01	1,61	1,48	0,13	0,08	0,18	0,94	0,89	0,99	30,74	29,14	32,43
02	8,61	8,84	-0,23	-0,33	-0,13	3,57	3,48	3,67	22,44	21,83	23,06
03	26,37	26,16	0,21	0,09	0,33	5,47	5,35	5,59	14,11	13,80	14,43
04	31,13	30,37	0,77	0,65	0,89	5,40	5,28	5,52	12,67	12,39	12,96
05	3,61	3,97	-0,36	-0,44	-0,28	2,25	2,17	2,33	30,87	29,83	31,96
06	21,48	21,90	-0,42	-0,50	-0,33	2,69	2,61	2,78	7,92	7,68	8,18
07	0,37	0,36	0,01	-0,02	0,04	0,34	0,31	0,37	46,54	42,58	50,87
08	0,58	0,54	0,04	0,01	0,06	0,29	0,26	0,32	25,96	23,58	28,58
09	6,24	6,38	-0,14	-0,18	-0,11	0,45	0,42	0,49	3,80	3,52	4,11
Aggregato	**-**	**-**	**-**	**-**	**-**	**10,70**	**10,54**	**10,86**	**13,78**	**13,58**	**13,99**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	9,91	9,61	10,21	12,99	12,58	13,41
Nord-est	-	-	-	-	-	11,66	11,28	12,04	14,99	14,51	15,48
Centro	-	-	-	-	-	11,82	11,44	12,21	15,15	14,67	15,66
Sud	-	-	-	-	-	10,38	10,06	10,71	13,31	12,90	13,73
Isole	-	-	-	-	-	9,75	9,31	10,20	12,59	12,00	13,21
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	10,34	10,07	10,62	13,62	13,27	13,99
10.000-100.000	-	-	-	-	-	9,90	9,67	10,13	12,74	12,43	13,05
100.000 e più (a)	-	-	-	-	-	10,90	10,33	11,47	13,81	13,11	14,56
Comuni metropolitani	-	-	-	-	-	13,86	13,39	14,33	17,71	17,12	18,32

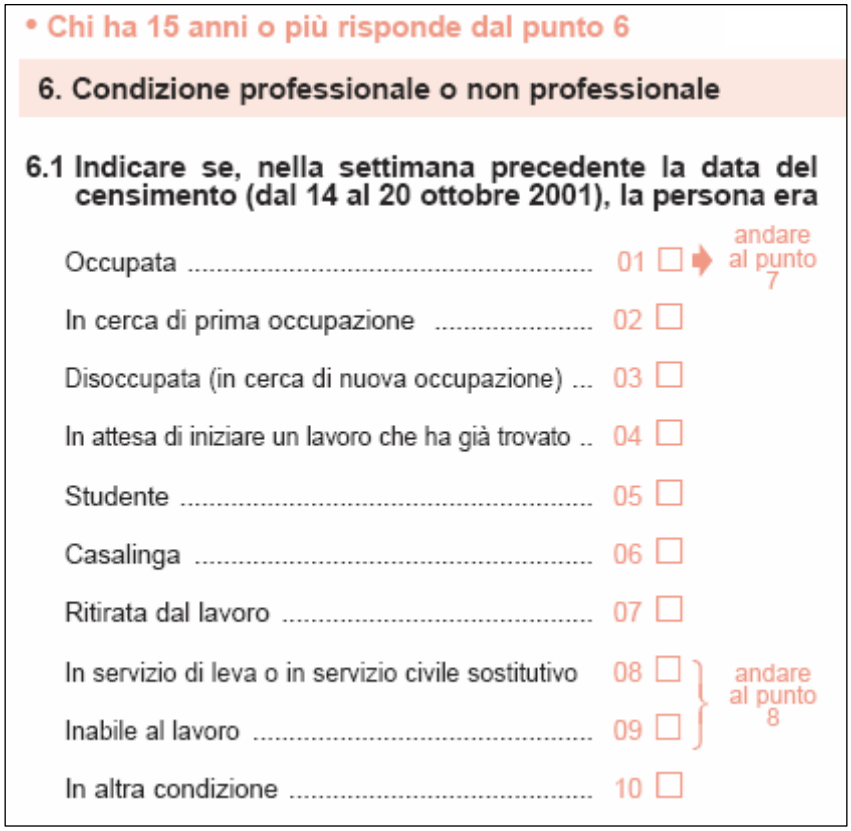
(a) Esclusi i comuni metropolitani.

Le analisi condotte per la condizione professionale (Figura 3.9) hanno preso in considerazione solo gli individui *linked* la cui età in anni compiuti, calcolata a partire dalla data di nascita, fosse maggiore di 14 anni. In generale, si osserva circa il 12 per cento degli individui come discordanti nelle risposte alle due occasioni di

rilevazione. L'indice di inconsistenza si attesta intorno al 16 per cento, quindi al di sotto della soglia critica (Tavola 3.16). Escludendo le modalità di risposta con frequenza inferiore al 5 per cento, non si osservano valori dell'indice di inconsistenza al di sopra della soglia empirica considerata critica. L'analisi per ripartizione evidenzia un livello di errore superiore per il Centro, il Sud e le Isole, mentre, come osservato per le altre variabili, sono i comuni metropolitani quelli dove si rileva la maggiore entità dell'errore.

Per questa variabile è stata condotta anche un'analisi sulle classi ricodificate in due sole categorie (Tavola 3.17). Come atteso, i livelli di errore scendono fortemente, confortando sui risultati diffusi a livello più aggregato.

Figura 3.9 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sulla condizione professionale o non professionale


• Chi ha 15 anni o più risponde dal punto 6

6. Condizione professionale o non professionale

6.1 Indicare se, nella settimana precedente la data del censimento (dal 14 al 20 ottobre 2001), la persona era

Occupata ..... 01 ☐ → andare al punto 7
In cerca di prima occupazione ..... 02 ☐
Disoccupata (in cerca di nuova occupazione) ... 03 ☐
In attesa di iniziare un lavoro che ha già trovato .. 04 ☐
Studente ..... 05 ☐
Casalinga ..... 06 ☐
Ritirata dal lavoro ..... 07 ☐
In servizio di leva o in servizio civile sostitutivo 08 ☐ } andare al punto 8
Inabile al lavoro ..... 09 ☐ }
In altra condizione ..... 10 ☐

Tavola 3.16 - Stime del *Net difference rate*, *Gross difference rate*, Indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per condizione professionale. Individui con età in anni compiuti superiore a 14 *(valori percentuali)*

MODALITÀ DI RISPOSTA (a)	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
01	41,99	42,34	-0,35	-0,44	-0,26	2,80	2,71	2,89	5,74	5,56	5,93
02	3,68	4,02	-0,34	-0,42	-0,26	2,16	2,08	2,24	29,18	28,14	30,26
03	4,70	3,99	0,70	0,62	0,79	2,75	2,66	2,84	33,06	32,02	34,15
04	0,38	0,24	0,14	0,11	0,17	0,32	0,30	0,36	52,14	47,47	57,26
05	7,74	7,82	-0,08	-0,13	-0,03	0,76	0,71	0,80	5,26	4,95	5,60
06	15,74	15,90	-0,16	-0,28	-0,05	4,71	4,60	4,83	17,70	17,27	18,14
07	19,82	21,01	-1,18	-1,31	-1,06	5,50	5,37	5,62	16,91	16,53	17,30
08	0,21	0,20	0,01	0,00	0,02	0,06	0,05	0,07	13,98	11,17	17,49
09	1,25	1,10	0,15	0,10	0,20	0,78	0,73	0,83	33,45	31,49	35,54
10	4,49	3,38	1,11	1,00	1,22	4,34	4,23	4,46	57,39	55,94	58,88
Aggregato	**-**	**-**	**-**	**-**	**-**	**12,09**	**11,92**	**12,26**	**16,24**	**16,01**	**16,48**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	9,54	9,24	9,85	13,71	13,26	14,18
Nord-est	-	-	-	-	-	9,65	9,29	10,01	14,48	13,92	15,06
Centro	-	-	-	-	-	13,47	13,05	13,89	18,60	18,03	19,19
Sud	-	-	-	-	-	14,11	13,72	14,49	17,44	16,97	17,92
Isole	-	-	-	-	-	15,70	15,13	16,27	19,48	18,79	20,20
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	10,96	10,67	11,25	14,95	14,56	15,35
10.000-100.000	-	-	-	-	-	11,74	11,48	12,00	15,66	15,32	16,01
100.000 e più (b)	-	-	-	-	-	11,96	11,35	12,58	16,58	15,76	17,45
Comuni metropolitani	-	-	-	-	-	15,76	15,26	16,27	20,72	20,07	21,40

(a) Per la decodifica delle modalità di risposta si veda la figura 3.8 sul titolo di studio più alto conseguito.
(b) Esclusi i comuni metropolitani.

Tavola 3.17 - Stime del *Net difference rate*, *Gross difference rate*, Indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per condizione professionale ricodificata in due categorie. Individui con età in anni compiuti superiore a 14 *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
"Occupata"	41,99	42,34	-0,35	-0,44	-0,26	2,80	2,71	2,89	5,74	5,56	5,93
Diversa da "Occupata"	58,01	57,66	0,35	0,26	0,44	2,80	2,71	2,89	5,74	5,56	5,93
Aggregato	**-**	**-**	**-**	**-**	**-**	**2,80**	**2,71**	**2,89**	**5,74**	**5,56**	**5,93**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	2,17	2,02	2,32	4,36	4,06	4,67
Nord-est	-	-	-	-	-	2,19	2,02	2,37	4,39	4,04	4,77
Centro	-	-	-	-	-	3,21	3,00	3,43	6,50	6,07	6,96
Sud	-	-	-	-	-	3,31	3,11	3,51	7,51	7,07	7,98
Isole	-	-	-	-	-	3,57	3,28	3,86	8,01	7,37	8,70
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	2,94	2,79	3,10	6,00	5,68	6,33
10.000-100.000	-	-	-	-	-	2,53	2,40	2,65	5,18	4,93	5,45
100.000 e più (a)	-	-	-	-	-	2,88	2,56	3,19	5,83	5,22	6,51
Comuni metropolitani	-	-	-	-	-	3,26	3,01	3,51	6,87	6,35	7,42

(a) Esclusi i comuni metropolitani.

Per quanto riguarda il quesito sul tipo di attività lavorativa (Figura 3.10), l'attenzione è limitata a quel sottoinsieme di individui *linked*, con età maggiore di 14 anni, che si sono dichiarati occupati a entrambe le occasioni di indagine. Il Gdr risulta piuttosto limitato, mentre l'indice di inconsistenza si attesta poco al di sopra della soglia di errore trascurabile (Tavola 3.18). Non si osservano differenze rilevanti per ripartizione geografica ed ampiezza demografica del comune.

Figura 3.10 - Foglio di famiglia del Censimento della popolazione 2001. Quesito sul tipo di attività lavorativa

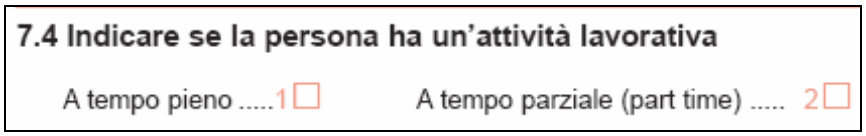
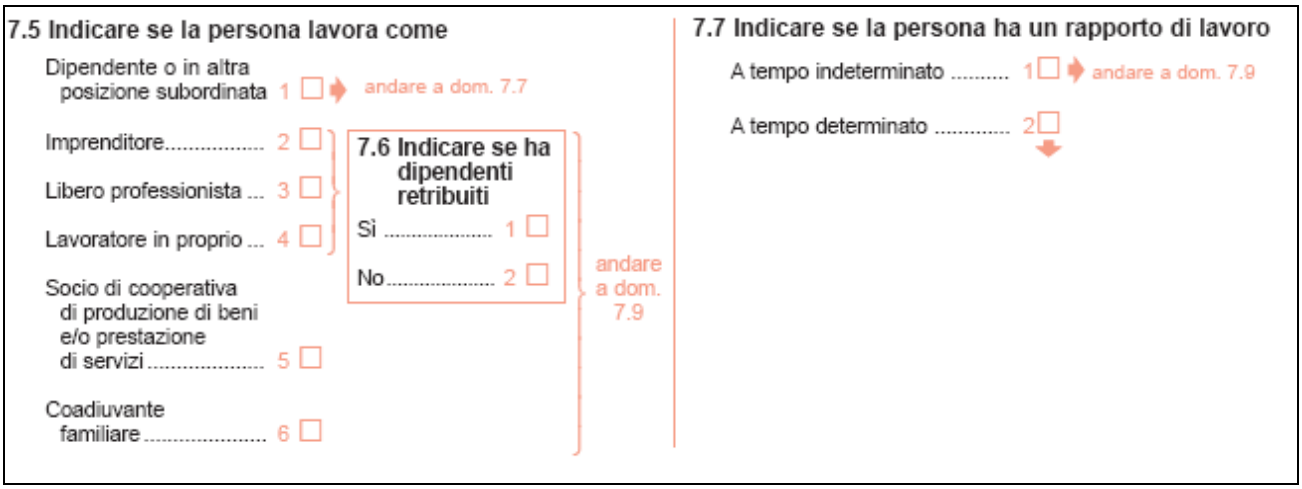


Tavola 3.18 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per tipo di attività lavorativa. Stime calcolate per gli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 - A tempo pieno	89,92	91,34	-1,42	-1,58	-1,26	3,57	3,41	3,73	20,98	20,06	21,96
2- A tempo parziale	10,08	8,66	1,42	1,26	1,58	3,57	3,41	3,73	20,98	20,06	21,96
Aggregato	**-**	**-**	**-**	**-**	**-**	**3,57**	**3,41**	**3,73**	**20,98**	**20,06**	**21,96**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	3,13	2,86	3,40	17,31	15,85	18,90
Nord-est	-	-	-	-	-	3,21	2,90	3,52	19,18	17,38	21,16
Centro	-	-	-	-	-	3,68	3,32	4,04	21,58	19,53	23,84
Sud	-	-	-	-	-	4,53	4,11	4,96	28,51	25,88	31,40
Isole	-	-	-	-	-	3,85	3,30	4,40	24,29	21,00	28,09
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	3,25	2,99	3,51	19,94	18,38	21,63
10.000-100.000	-	-	-	-	-	3,35	3,12	3,58	19,67	18,34	21,10
100.000 e più (a)	-	-	-	-	-	3,61	3,07	4,16	22,16	19,01	25,83
Comuni metropolitani	-	-	-	-	-	5,05	4,54	5,57	26,61	23,97	29,54

(a) Esclusi i comuni metropolitani.

Figura 3.11 - Foglio di famiglia del Censimento della popolazione 2001. Quesiti sulla posizione nella professione e sulla durata del rapporto di lavoro



Anche per questi ultimi quesiti (Figura 3.11) le stime sono calcolate sul sottoinsieme di individui *linked*, di età maggiore di 14 anni, e che si siano dichiarati occupati.

Complessivamente, per la variabile posizione nella professione a sei categorie, il livello di errore non è particolarmente allarmante, anche se bisogna sottolineare che, in questi quesiti filtrati, il sottoinsieme di popolazione che si va a studiare è già in qualche modo selezionato, essendo le stime calcolate sul sottoinsieme di individui in accordo sul quesito di filtro.

Anche per questo quesito la stima dell'errore è leggermente maggiore per i grandi comuni e i comuni metropolitani e per il Centro e il Sud (Tavola 3.19).

Tavola 3.19 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per posizione nella professione. Stime calcolate per gli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
1 - Dipendente	73,98	75,00	-1,02	-1,17	-0,87	3,17	3,02	3,32	8,34	7,95	8,75
2 – Imprenditore	5,37	4,62	0,75	0,60	0,90	3,09	2,95	3,25	32,62	31,08	34,23
3 - Libero professionista	5,69	5,66	0,03	-0,10	0,16	2,32	2,20	2,45	21,68	20,50	22,92
4 - Lavoratore in proprio	11,45	11,97	-0,53	-0,71	-0,34	4,91	4,72	5,10	23,73	22,83	24,65
5 - Socio di cooperativa	1,71	1,18	0,53	0,44	0,62	1,16	1,08	1,26	40,80	37,72	44,14
6 – Coadiuvante familiare	1,81	1,55	0,25	0,17	0,34	0,95	0,87	1,03	28,67	26,28	31,28
Aggregato	**-**	**-**	**-**	**-**	**-**	**7,81**	**7,58**	**8,04**	**18,36**	**17,81**	**18,93**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	6,42	6,04	6,80	16,18	15,22	17,20
Nord-est	-	-	-	-	-	8,38	7,89	8,86	19,51	18,37	20,73
Centro	-	-	-	-	-	9,68	9,12	10,24	21,29	20,03	22,62
Sud	-	-	-	-	-	6,91	6,40	7,43	16,51	15,27	17,84
Isole	-	-	-	-	-	8,52	7,73	9,31	18,88	17,14	20,79
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	7,22	6,85	7,60	16,18	15,33	17,08
10.000-100.000	-	-	-	-	-	7,62	7,28	7,96	18,19	17,37	19,05
100.000 e più (a)	-	-	-	-	-	9,16	8,32	9,99	21,31	19,37	23,45
Comuni metropolitani	-	-	-	-	-	9,07	8,40	9,74	23,82	22,05	25,74

(a) Esclusi i comuni metropolitani.

Come atteso, il livello di errore diventa ancor più trascurabile se la variabile posizione nella professione viene ricodificata in due classi: dipendente e diverso da dipendente (Tavola 3.20).

Tavola 3.20 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per posizione nella professione ricodificata in due categorie. Stime calcolate per gli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
Dipendente	73,98	75,00	-1,02	-1,17	-0,87	3,17	3,02	3,32	8,34	7,95	8,75
Diversa da "Dipendente"	26,02	24,98	1,02	0,87	1,17	3,17	3,02	3,32	8,34	7,95	8,75
Aggregato	**-**	**-**	**-**	**-**	**-**	**3,17**	**3,02**	**3,32**	**8,34**	**7,95**	**8,75**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	2,90	2,64	3,16	8,09	7,39	8,86
Nord-est	-	-	-	-	-	2,95	2,65	3,24	7,71	6,96	8,53
Centro	-	-	-	-	-	3,35	3,01	3,69	8,35	7,53	9,26
Sud	-	-	-	-	-	3,55	3,17	3,92	9,43	8,46	10,51
Isole	-	-	-	-	-	3,56	3,04	4,08	8,86	7,63	10,28
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	2,81	2,57	3,05	7,06	6,47	7,70
10.000-100.000	-	-	-	-	-	3,13	2,90	3,35	8,33	7,75	8,96
100.000 e più (a)	-	-	-	-	-	3,44	2,91	3,97	8,99	7,69	10,50
Comuni metropolitani	-	-	-	-	-	4,08	3,62	4,54	11,79	10,51	13,23

(a) Esclusi i comuni metropolitani.

Tavola 3.21 - Stime del *Net difference rate*, *Gross difference rate*, indice di inconsistenza e relativi intervalli di confidenza al 95 per cento per rapporto di lavoro. Stime calcolate per gli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati in posizione di dipendente *(valori percentuali)*

MODALITÀ DI RISPOSTA	Stima Cpa	Stima Idc	Net difference rate			Gross difference rate			Indice di inconsistenza		
			Stima	Intervallo di confidenza		Stima	Intervallo di confidenza		Stima	Intervallo di confidenza	
				Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.		Lim. inf.	Lim. sup.
Tempo indeterminato	87,68	89,95	-2,27	-2,51	-2,03	5,76	5,52	6,00	28,95	27,76	30,18
Tempo determinato	12,32	10,05	2,27	2,03	2,51	5,76	5,52	6,00	28,95	27,76	30,18
Aggregato	**-**	**-**	**-**	**-**	**-**	**5,76**	**5,52**	**6,00**	**28,95**	**27,76**	**30,18**
STIMA AGGREGATA PER RIPARTIZIONE TERRITORIALE											
Nord-ovest	-	-	-	-	-	4,96	4,57	5,35	29,80	27,49	32,30
Nord-est	-	-	-	-	-	5,77	5,29	6,25	34,33	31,51	37,40
Centro	-	-	-	-	-	5,96	5,42	6,50	29,87	27,21	32,78
Sud	-	-	-	-	-	6,43	5,83	7,02	24,09	21,88	26,51
Isole	-	-	-	-	-	6,83	5,97	7,69	27,35	24,01	31,15
STIMA AGGREGATA PER AMPIEZZA DEMOGRAFICA DEI COMUNI											
Fino a 10.000	-	-	-	-	-	5,95	5,54	6,36	28,44	26,48	30,54
10.000-100.000	-	-	-	-	-	5,76	5,41	6,11	27,98	26,28	29,79
100.000 e più (a)	-	-	-	-	-	5,27	4,50	6,03	31,12	26,81	36,12
Comuni metropolitani	-	-	-	-	-	5,61	4,99	6,23	33,16	29,58	37,17

(a) Esclusi i comuni metropolitani.

Infine, per la variabile sul tipo di rapporto di lavoro (Tavola 3.21), se a tempo determinato o indeterminato, complessivamente si ha una percentuale di discrepanza tra le due rilevazioni inferiore al 6 per cento, però se si considera l'indice di inconsistenza, si ha un livello di errore del 28,95 per cento, considerato un errore di entità media.

3.4 - Conclusioni

I risultati delle analisi sulla stima della variabilità di risposta per il Censimento del 2001 sembrano indicare che la formulazione dei quesiti ha funzionato bene, tuttavia c'è una qualche difficoltà per quesiti con elevato numero di modalità di risposta (es. titolo di studio) e per quelli relativi alla condizione professionale e attività lavorativa. Si osserva una minore accuratezza delle risposte per in alcuni domini di stima (Centro e Sud, Comuni Metropolitani). Non vi sono dati per stabilire quali siano i fattori che causano questa minore accuratezza. Tendenzialmente, non esistono motivi per ritenere che la variabilità delle risposta possa variare in funzione delle aree geografiche o del tipo di comune di residenza. Si possono fare alcune ipotesi. Da una parte può esserci un fattore dovuto alla maggior fretta o minore attenzione da parte dei rispondenti nella compilazione del questionario nei grandi comuni. Dall'altra, ci può essere l'ipotesi di una minore qualità del lavoro dei rilevatori. Tale ipotesi è però difficilmente plausibile dato che nel caso del Censimento i rilevatori si dovrebbero limitare a consegnare e ritirare i questionari e solo in pochi casi a supportare i rispondenti nella compilazione.

Dal punto di vista generale, si può sicuramente affermare che l'esperienza di valutare l'errore di risposta utilizzando i dati dell'Idc, e quindi di unificare le due indagine post-censuarie (quella di qualità e quella di copertura), sia stata una scelta di successo. Non bisogna tuttavia dimenticare che tale la possibilità è fortemente legata alle performance delle procedure di *record linkage*, infatti, una procedura di *record linkage* poco efficiente ed accurata (elevate probabilità di falsi *link* ed elevato numero di unità non *linked*) rende poco attendibile il confronto delle risposte alle due occasioni d'indagine e, conseguentemente distorte le stime della variabilità dovuta agli errori di risposta.

Bibliografia

Abbate, C.C., M. Masselli e M. Signore. 1993. “A combined post-enumeration survey of the 1991 population and industrial census”, *Proceedings of Isi*, 2, 16.3.

Australian Bureau Of Statistics. 2003. *Information paper: census of population and housing, data quality-undercount, 2001*. Australia**:** Australian Bureau of Statistics.

Balestrino, R., Baiocchi, F. e A. Reale. 1999. ”Census forms optical reading: taking the opportunity”, *Proceedings of European workshop on the preparation of the census fieldwork, Joint Istat-Ece-Eurostat Meeting*, Roma: 12-14 Aprile 1999

Biemer, P.P. 2004. “The twelfth Morris Hansen lecture simple response variance: then and now”. *Journal of Official Statistics*, 20, n. 3: 417-439.

Biemer, P. P. e G. Forsman. 1992. “On the quality of reinterview data with application to the current population survey”. *Journal of the American Statistical Association*, 87, n.420, 915-923.

Brancato, G., M. D’Orazio e M. Fortini. 2004. “Response error estimation in presence of record linkage Errors: the case of the Italian population census”. In *Proceedings of the European Conference on quality and methodology in official statistics (Q2004)*. Mainz, 24-26 May 2004. Wiesbaden: Statistisches Bundesamt.

Brown, L. D., T. Toni Cai e A. DasGupta. 2001. “Interval estimation for a binomial proportion”. *Statistical Science*, 16, 2: 101-133.

Carbonetti, G., M. Fortini, N. Mignolli e A. Nuccitelli. 2005. “L’indagine sul grado di copertura del 14° censimento della popolazione: considerazioni di carattere metodologico e primi risultati”. Poster presentato alle *Giornate di studio sulla popolazione*, Padova: 16 – 18 febbraio 2005.

Cariani, G. “I controlli ED del censimento demografico”. 1983. In *Atti del Convegno intermedio della Società italiana di statistica: La qualità dei dati statistici. Trieste, 21-23 aprile.*

Casale, D., A. Chieppa e F. Panizon. 2002. “Progettazione, integrazione e utilizzo delle informazioni di controllo per il miglioramento della qualità dei dati del Censimento della popolazione 2001”. Paper presentato al Workshop *Strategie e modelli per il controllo della qualità dei dati*, Bologna: 22 aprile 2002.

Chan, K. W. 2003. “Chinese Census 2000: new opportunities and challenges” In *The China Review*, 3, 2. 1-12.

Chieppa, A. e F. Panizon. 2001. “Data quality control system for the 2001 Italian population census”. Paper presentato all’International Conference on quality in official statistics**,** Q2001, Stoccolma: 14-15 maggio.

Chieppa, A. e F. Panizon. 2002. “Design of ‘quality archives’ for controls on population and house census data”. In *Atti della XLI Riunione scientifica della Società italiana di statistica. Milano, 5-7 giugno.*

Cortese, A. 1983. “Indagine sul confronto censimento-anagrafe: scopi, modalità di esecuzione, principali risultati”. In *Atti del Convegno intermedio della Società italiana di statistica: La qualità dei dati statistici. Trieste, 21-23 aprile.*

Cortese, A. e M. Greco. 1993. “Il grado di copertura del censimento demografico 1991: considerazioni sulla base del confronto con le risultanze anagrafiche”. Paper presentato alle Giornate di studio sulla popolazione, Bologna, 6-7 dicembre.

Crescenzi, F. 1999. “New challenges and opportunities in redesign and updating territorial database”. In *Proceedings of European Workshop on the preparation of the Census fieldwork. Joint Istat-Ece-Eurostat meeting. Roma, 12-14 aprile.*

Depoutot R. 1998. “Quality of international statistics: comparability and coherence”. Paper presentato alla Conference on methodological issues in official statistics. Stockholm.

De Santis, G., S. Salvini e A. Santini. 1995. “La qualità dei dati dell’ultimo Censimento generale della popolazione e delle abitazioni (Data quality of the last population and housing Census)”. Relazione presentata alla Commissione per la garanzia dell’informazione statistica. Presidenza del Consiglio dei ministri, novembre.

Deville, J. C. e C.-E. Särndal. 1992. “Calibration estimators in survey sampling” . *Journal of the American statistical association* 87, 418: 376-382.

Di Consiglio, L. e S. Falorsi. 2003. *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento generale della popolazione 2001*. Roma: Istat. (Documenti, n. 11). http://www.istat.it/dati/pubbsci/documenti/Doc_anno2003.htm

Egidi, V. 1999. “Le strategie dell’Istat per i Censimenti del 2000”. Paper presentato alla Conferenza Verso i censimenti del 2000, Udine, 7-9 giugno.

Engstrom, P. e L. Granquist. 1999. “Improving quality by modern editing”. Paper presentato alla Unece work session on statistical data editing, Roma, 2-4 giugno.

Eurostat. 2004. *Documentation of the 2000 round of population and housing censuses in the Ee, Efta and candidate countries.* In *Population and social condition 3/2004/F/n. 01*. Lussemburgo: Eurostat.

Filippucci, C. 2000. “Qualità delle statistiche e controllo del processo di misura”. Paper presentato al Seminario Sieds-Istat. La qualità dell'informazione statistica, Roma, 6-7 aprile.

Fortini, M. 1994. “Un’applicazione del modello a classi latenti per l’analisi dell’errore di copertura del XIII censimento della popolazione”. In *Atti della XXXVII Riunione scientifica della Società italiana di statistica,* San Remo, 6-8 aprile.

Fortini, M., B. Liseo, A. Nuccitelli e M. Scanu. 2001.“On bayesian record linkage”. *Research in official statistics*. 4. 185-198.

Fortini, M., M. Scanu e M. Signore. 1999. “Measuring and analysing the data editing activity in Istat information system for survey documentation”. Paper presentato alla Unece Work session on statistical data editing, Roma, 2-4 giugno.

Hansen, M. H., W.N. Hurwitz e L. Pritzker. 1964. “The estimation and interpretation of gross differences and simple response variance”. In *Contributions to statistics*, C. R. Rao, 111-136. Calcutta: Statistical publishing society.

Istat. 1989. *Il sistema di controllo della qualità dei dati - Manuale di tecniche di indagine vol. 6*. Roma: Istat. (Note e relazioni, n. 1).

Istat. 1993. *La progettazione dei Censimenti 1991: basi territoriali, organizzazione, campagna di informazione, piano dei controlli - 13° Censimento generale della popolazione e delle abitazioni*, Roma: Istat.

Istat. 1997. *I controlli di qualità: l'elaborazione dei dati - 13° Censimento generale della popolazione e delle abitazioni*, Roma: Istat.

Istat. 2001. *Disposizioni per gli organi periferici e istruzioni per il rilevatore*. Roma: Istituto poligrafico e zecca dello Stato.

Istat. 2004. *Concord v. 1.0 - controllo e correzione dei dati - Manuale utente e aspetti metodologici*. Roma: Istat. (Tecniche e strumenti, n. 1)

Istat. 2005. *14° Censimento della popolazione e delle abitazioni 2001, struttura demografica e familiare della popolazione residente – Italia*. Roma: Istat.

Istat. 2006°. *Il piano di rilevazione e il sistema di produzione - 14° Censimento generale della popolazione e delle abitazioni*. Roma: Istat.

Istat. 2006b. *I documenti - 14° Censimento generale della popolazione e delle abitazioni*. Roma: Istat.
Laihonen, A. 2000. "2001 Round population censuses". In Insee-Eurostat Seminar on censuses after 2001, Paris: novembre.

Linacre, S. J. 1991. "Approaches to quality assurance in the Australian bureau of statistics business surveys". In *Bulletin of the international statistical institute, Proceedings of the 48th Session.* 297-321. Il Cairo.

Masselli, M. 1983. "Risultati dell'indagine di controllo sulla qualità dei dati del censimento del 1981". In *Atti del Convegno intermedio della Società italiana di statistica: La qualità dei dati statistici*. Trieste, 21-23 aprile.

Massimini, G., e P. Valente 1998. "Processing the Italian population and housing census data". Paper presentato all'International seminar on census methodology. Portsmouth (UK), 29 aprile-1 maggio.

Morganstein, D., e D. A. Marker. 1997. "Continuous quality improvement in statistical agencies". In *Survey measurement and process quality*. Lyberg L.E. et al. 475-500. New York: Wiley.

Nuccitelli, A. 2001. *Integrazione di dati mediante tecniche di abbinamento esatto: una rassegna critica ed una proposta in ambito bayesiano*. Tesi di dottorato in Metodi statistici per l'economia e l'impresa. Roma: Università degli studi di Roma Tre.

Nuccitelli, A. 2005. "La strategia di abbinamento dei dati del 14° Censimento della popolazione con i dati dell'indagine di copertura". In *L'integrazione di dati di fonti diverse - Tecniche e applicazioni del record linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*. A cura di P. D. Falorsi, A. Pallara, A. Russo. 61-91. Milano: Franco Angeli.

Nuccitelli, A., F. Bosio, e L. Fioriti. 2004. *L'applicazione RECLINK per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*. Roma: Istat. (Documenti, n.10).

Office for National Statistics. 2003.*Census strategic development review. Alternative to a census: review of international approaches.* Gran Bretagna*:* ONS. ottobre.

Orasi, A. 2002. "Contenuti informativi e operazioni sul campo dei censimenti generali della popolazione". Relazione presentata alla Sesta Conferenza nazionale di statistica, Roma, 6-8 novembre.

Pagliuca, D. 2005. *Genesess v.3.0, funzione di stime ed errori. Manuale utente ed aspetti metodologici.* Roma: Istat. (Tecniche strumenti Istat, n.3)

Parson, N., e G. Jones. 1999. "A quality assurance strategy for processing the UK 2001 population Census". Paper presentato alla Conferenza Verso i censimenti del 2000, Udine, 7-9 giugno.

Petersen, C. G. J. 1896. "The yearly immigration of young plaice into Limfjord from the German Sea." *Report of the Danish biological station*.6:1Ð48.

Pollock, K. H. et al. 1990. *Statistical inference for capture-recapture experiments*. Bethesda. (Wildlife Monographs, 107)

Särndal, C.-E., B. Swensson e J. Wretman. 1992. *Model assisted survey sampling*. New York: Springer Verlag.

Sekar, C. C., e W. E. Deming. 1949. "On a method of estimating birth and death rates and the extent of registration". *Journal of the American statistical association*, 44, 245. 101-115.

Skinner, T., J. Struik e M. Butterfield. 1998. "Managing the census process at the Australian Bureau of statistics". Paper presentato all'International seminar on census methodology, Portsmouth (UK), 29 aprile-1 maggio.

Statistics Canada. 2004. *Coverage, 2001 Census technical report.* Statistics Canada.

Statistics New Zealand. 2002. *A report on the post-enumeration survey 2001*. Statistics New Zealand.

Statistical Office of Estonia. 2005. *General information of 2000 population and housing census in Estonia.* Statistical Office of Estonia.

Statistics South Africa. 2004. "Census 2001: post-enumeration survey: results and methodology". *Report No. 03-02-17*.

Terra Abrami, V. e M. Masselli. 1983. "L'indagine di controllo di copertura del censimento della popolazione". In *Atti del Convegno intermedio della Società italiana di statistica: La qualità dei dati statistici. Trieste, 21-23 aprile.*

US Census Bureau. 1985. "Evaluating censuses of population and housing". *Statistical training document.* ISP-TR-5.

US Census Bureau. 2003. "Census 2000 content reinterview survey: accuracy of data for selected population and housing characteristics as measured by reinterview". *Census 2000 evaluation, Final report B*.5: 9.

US Census Bureau. 2004. "Coverage measurement from the perspective of March 2001 accuracy and coverage evaluation" In *Census 2000 topic report No. 4.* US Census Bureau.

US Department of Education - National Center for Education Statistics. 1997. "Reinterview results for the school safety & discipline and school readiness components". *Technical report*, NCES 97-339.

Winkler, E. 1999. "Draft glossary of terms used in data editing". Paper presentato alla Unece work session on statistical data editing, Roma: 2-4 giugno.

Winkler W. E. 2001. "Record linkage software and methods for merging administrative lists". In *Statistical research report series RR2001/03*. Washington: U.S. Bureau of the Census.

Wolter, K. M. 1986. “Some coverage error models for census data”, *Journal of the American statistical association*, 81, 394. 338-346.

Woodruff, R. S. 1971.**”**A simple method for approximating the variance of a complicated estimate**”,** *Journal of the American statistical association*, 66, 334. 411-414.

APPENDICE

Appendice A
Denominazione delle variabili utilizzate nel questionario del 14° Censimento generale della popolazione e delle abitazioni e loro significato

RESIDENTI IN FAMIGLIA (Modello Istat CP.1. Sezione II)

AMAT	anno di matrimonio
ANAS	anno di nascita
ANNTER	anno ritiro dal lavoro
ANNTES	anni studio all'estero
ANNTRA	anno di trasferimento in Italia (per gli stranierei nati all'estero)
CERCAT	ricerca di lavoro
CITTAD	cittadinanza
CONDIZ	condizione professionale o non professionale
CONTIN	proseguimento altra dimora
CORFOR	frequenza corso di formazione/aggiornamento professionale
DADOVE	alloggio di uscita
DIMNAP	dimora un anno fa
DUESET	disponibilità al lavoro
ESTNAS	codice stato estero di nascita
FREQUE	frequenza asilo nido o scuola materna
GNAS	giorno di nascita
HADIP	presenza dipendenti
HASVOL	lavoro nel corso della vita
ISCRIZ	iscrizione ad un corso regolare di studi
ITANAS	origine della cittadinanza italiana
LUONAS	luogo di nascita (Italia o estero)
LUOSL	luogo di studio o lavoro
MEZZOT	mezzo di trasporto utilizzato
MMAT	mese di matrimonio
MNAS	mese di nascita
MOTIVO	motivo altra dimora
MOTNES	motivo eventuale mancanza ore
MOTTRA	motivo di trasferimento in Italia (per gli stranieri nati all'estero)
NOVANT	numero giorni altra dimora
OHOW	(*one hour one week*) effettuazione ore lavoro nella settimana precedente
ORELAV	numero di ore di lavoro effettuate
ORESP	domanda filtro sulle ore di lavoro
POSIZ	posizione nella professione
PRESEN	presenza alla data di censimento
RAPPOR	durata rapporto di lavoro
RELPAR	relazione di parentela
RIENTR	alloggio di rientro
SCIVUM	stato civile precedente al matrimonio
SESSO	sesso

SETATT	settore di attività economica
SIRECA	studio o lavoro fuori casa
SITUAT	ubicazione altra dimora
SPECIA	specializzazione post-laurea e/o dottorato di ricerca
STAC	codice stato estero di cittadinanza (per gli stranieri)
STACIV	stato civile
STAP	codice stato estero di cittadinanza precedente (per gli italiani acquisiti)
TEMIMP	tempo impiegato per recarsi al luogo di studio o lavoro
TEMPP	durata attività
TIPATT	attività lavorativa svolta
TIPCOR	tipo di corso formazione/aggiornamento professionale
TIPRAP	tipo rapporto di lavoro
TITEST	studio all'estero
TITSTU	titolo di studio
VISSUT	altra dimora ultimo anno

NON RESIDENTI IN FAMIGLIA (Modello Istat CP.1. Sezione III)

ANAS	anno di nascita
APRES	anno arrivo in Italia
CITTAD	cittadinanza
CONDSP	condizione professionale o non professionale
DADOVE	alloggio di uscita
DIMANC	dimora abituale alla data di censimento
GNAS	giorno di nascita
LUOSL	luogo di studio o lavoro
MEZZOT	mezzo di trasporto utilizzato
MNAS	mese di nascita
MOTAL	motivo utilizzo alloggio
MOTPRE	motivo di presenza in Italia (solo stranieri residenti all'estero)
MPRES	mese arrivo in Italia
PRESEN	presenza alla data di censimento
RIENTR	alloggio di rientro
SESSO	sesso
SETATP	settore di attività economica
SIRECA	studio o lavoro fuori casa
STAC	codice stato estero di cittadinanza (per gli stranieri)
STACIV	stato civile
TEMIMP	tempo impiegato per recarsi al luogo di studio o lavoro
TMPVIS	numero giorni dimora in questo alloggio

RESIDENTI IN CONVIVENZA (Modello Istat CP.2. Sezione I)

ADAC	anno arrivo nella convivenza
ANAS	anno di nascita
ANNTER	anno ritiro dal lavoro
ANNTES	anni studio all'estero
ANNTRA	anno di trasferimento in Italia (per gli stranieri nati all'estero)
CERCAT	ricerca di lavoro
CITTAD	cittadinanza
CONDIZ	condizione professionale o non professionale

CONTIN	proseguimento altra dimora
CORFOR	frequenza corso di formazione/aggiornamento professionale
DIMAPC	dimora un anno fa
DIMCOP	tipo di dimora un anno fa
DUESET	disponibilità al lavoro
ESTNAS	codice stato estero di nascita
FREQUE	frequenza asilo nido o scuola materna
GNAS	giorno di nascita
HADIP	presenza dipendenti
HASVOL	lavoro nel corso della vita
ISCRIZ	iscrizione ad un corso regolare di studi
ITANAS	origine della cittadinanza italiana
LUONAS	luogo di nascita (Italia o estero)
MDAC	mese arrivo nella convivenza
MNAS	mese di nascita
MOPERC	motivo di permanenza nella convivenza
MOTIVO	motivo altra dimora
MOTNES	motivo eventuale mancanza ore
MOTTRA	motivo di trasferimento in Italia (per gli stranieri nati all'estero)
NOVANT	numero giorni altra dimora
OHOW	(*one hour one week*) effettuazione ore lavoro nella settimana precedente
ORELAV	numero di ore di lavoro effettuate
ORESP	domanda filtro sulle ore di lavoro
POSIZ	posizione nella professione
PRESEN	presenza alla data di censimento
RAPPOR	durata rapporto di lavoro
SESSO	sesso
SETATT	settore di attività economica
SITUAT	ubicazione altra dimora
SPECIA	specializzazione post-laurea e/o dottorato di ricerca
STAC	codice stato estero di cittadinanza (per gli stranieri)
STACIV	stato civile
STAP	codice stato estero di cittadinanza precedente (per gli italiani acquisiti)
TEMPP	durata attività
TIPATT	attività lavorativa svolta
TIPCOR	tipo di corso formazione/aggiornamento professionale
TIPRAP	tipo rapporto di lavoro
TITEST	studio all'estero
TITSTU	titolo di studio
VISSUT	altra dimora ultimo anno

NON RESIDENTI IN CONVIVENZA (Modello Istat CP.2. Sezione I)

ANAS	anno di nascita
APRES	anno arrivo in Italia (mese; anno)
CITTAD	cittadinanza
CONDSP	condizione professionale o non professionale
DIMANC	dimora abituale alla data di censimento
GNAS	giorno di nascita
MNAS	mese di nascita
MOPERC	motivo di permanenza nella convivenza
MOTPRE	motivo di presenza in Italia (solo stranieri residenti all'estero)

MPRES	mese arrivo in Italia
PRESEN	presenza alla data di censimento
SESSO	sesso
SETATP	settore di attività economica
STAC	codice stato estero di cittadinanza (per gli stranieri)
STACIV	stato civile
TMPVIS	numero giorni dimora in questo alloggio

EDIFICI (Modello Istat CP.ED)

ASCENS	ascensore
CONTIG	contiguità
EDUTIL	utilizzazione edificio
EPOCOS	epoca costruzione
NPIAFT	numero piani fuori terra
NSCALE	numero scale
PIAINT	presenza piani interrati
STCONS	stato conservazione
TIPEDI	tipo costruzione
TIPMAT	materiale costruzione
TIPUSO	tipo edificio
TOTINT	totale interni edificio

ABITAZIONI (Modello Istat CP.1, Sezione I)

ACQCAL	disponibilità acqua calda
ANGCOT	presenza angolo cottura
BOXPRI	disponibilità box
CORTIL	disponibilità posto auto cortile
CUCINI	presenza cucinino
CUCSTA	numero cucine
ENRIAC	produzione acqua calda
FONTAC	disponibilità acqua potabile
GABIN, GASUP	numero di gabinetti
GARAGE	disponibilità posto auto garage
IMPRID	tipo impianto riscaldamento
NPIANI	numero piani
NSTAB	numero stanze
OPEIMP	ristrutturazione impianti negli ultimi dieci anni
OPENOS	ristrutturazione interni negli ultimi dieci anni
OPESTR	interventi strutturali negli ultimi dieci anni
PROPR	proprietà
RISACQ	riscaldamento comune per acqua calda
STANUF, STUSUP	numero stanze uso ufficio
SUPERF	superficie in metri quadri
TELFIS	disponibilità telefono fisso
TICOMB	combustibile riscaldamento
TITGOD	titolo godimento abitazione
VASDOC, VADOSU	numero di docce e vasche

Appendice B
Tabelle di confronto tra le risposte al Censimento e le risposte all'Indagine di copertura

Tavola B.1 - Confronto tra le risposte al Censimento e all'Indagine di copertura per "Relazione di parentela o di convivenza" *(stime ottenute con dati pesati)*

RISPOSTA A IDC (a)	Risposte al Cpa								
	1	2	3	4	5	6	7	8	9
1	20.141.589	247.769	29.416	47.177	35.210	1.382	56.665	12.916	24.898
2	215.881	11.910.507	106.238	31.372	13.288	1.782	16.070	3.426	2.729
3	34.455	20.804	382.931	17.376	4.514	475	508	-	499
4	38.835	39.206	7.948	14.228.273	548.859	63.161	3.906	904	8.709
5	39.957	6.376	7.445	457.401	1.384.160	41.148	1.972	123	5.875
6	832	489	1.676	47.777	22.074	99.097	-	-	-
7	43.100	7.244	313	3.723	180	102	344.725	7.538	182
8	7.017	1.641	-	268	-	-	4.441	190.856	1.305
9	22.850	2.587	1.147	4.813	4.329	-	1.058	-	327.751
10	4.577	-	518	-	-	-	-	-	3.994
11	8.029	456	-	-	-	-	-	-	273
12	15.237	12.141	357	1.878	517	805	617	1.255	-
13	3.366	-	391	34.616	8.330	1.751	-	-	-
14	1.743	143	-	4.022	730	-	-	-	335
15	6.440	1.915	474	1.846	541	1.327	2.546	728	2.999
16	10.967	183	11.541	2.202	1.113	1.613	-	274	2.282
VNA (b)	1.115	1.294	-	-	-	-	-	-	-
VM (c)	93.483	57.921	1.792	113.119	11.897	348	2.247	1.852	2.676
Totale	**20.689.473**	**12.310.676**	**552.187**	**14.995.863**	**2.035.742**	**212.991**	**434.755**	**219.872**	**384.507**

RISPOSTA A IDC (a)	Risposte al Cpa									Totale
	10	11	12	13	14	15	16	Vna (b)	Vm (c)	
1	2.649	1.445	7.460	4.058	1.142	4.343	10.836	-	48.341	20.677.296
2	-	960	14.316	800	245	4.298	678	-	284.832	12.607.422
3	-	-	883	1.876	-	826	11.637	-	10.660	487.444
4	618	-	1.759	32.923	2.892	870	1.189	-	299.134	15.279.186
5	99	-	537	3.393	664	260	-	-	64.192	2.013.602
6	-	-	-	-	-	369	2.619	-	5.336	180.269
7	-	-	194	-	-	2.067	622	-	7.672	417.662
8	-	-	1.748	-	-	2.426	474	-	4.617	214.793
9	6.303	1.481	-	999	499	1.639	155	-	8.938	384.549
10	34.500	497	-	-	-	3.810	-	-	3.969	51.865
11	510	12.012	-	168	-	980	893	-	1.224	24.545
12	689	-	204.467	2.066	-	1.825	1.429	-	6.212	249.495
13	158	-	2.997	411.485	19.120	2.075	397	-	18.846	503.532
14	-	-	-	18.474	30.873	1.912	-	-	1.682	59.914
15	4.538	3.462	1.399	4.646	2.084	66.929	2.056	-	3.782	107.712
16	-	790	3.218	540	557	8.371	68.855	-	4.654	117.160
VNA (b)	-	-	-	-	-	-	-	-	-	2.409
VM (c)	666	-	2.792	5.044	937	1.954	658	-	10.326	307.712
Totale	**50.730**	**20.647**	**241.770**	**486.472**	**59.013**	**104.954**	**102.498**	**-**	**784.417**	**53.686.567**

(a) Legenda:
1 Intestatario del Foglio di famiglia.
2 - Coniuge dell'intestatario.
3 - Convivente dell'intestatario.
4 - Figlio/a dell'intestatario e del coniuge/convivente.
5 - Figlio/a del solo intestatario.
6 - Figlio/a del solo coniuge/convivente.
7 - Genitore (o coniuge del genitore) dell'intestatario.
8 - Suocero/a dell'intestatario.
9 - Fratello/sorella dell'intestatario.
10 - Fratello/sorella del coniuge/convivente.
11 - Coniuge del fratello/sorella dell'intestatario o del fratello/sorella del coniuge/convivente.
12 - Genero/nuora (coniuge/convivente del figlio/a) dell'intestatario e/o del coniuge convivente.
13 - Nipote (figlio/a di un figlio/a) dell'intestatario e/o del coniuge convivente.
14 - Nipote (figlio/a di un fratello/sorella) dell'intestatario e/o del coniuge convivente.
15 - Altro parente dell'intestatario e/o del coniuge/convivente.
16 - Altra persona convivente senza legami di parentela.

(b) Vna – Valore non ammissibile.
(c) Vm – Valore mancante.

A cura di Marcello D'Orazio

Tavola B.2 - Confronto tra le risposte al Censimento e all'Indagine di copertura per sesso *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa				
	Maschio	Femmina	Vna (a)	Vm (b)	Totale
Maschio	25.186.860	238.877	-	293.664	25.719.401
Femmina	186.379	26.993.210	-	280.488	27.460.077
VNA (a)	-	-	-	-	-
VM (b)	236.842	262.905	-	7.342	507.089
Totale	**25.610.081**	**27.494.992**	**-**	**581.494**	**53.686.567**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.3 - Confronto tra le risposte al Censimento e all'Indagine di copertura per classi di età in anni compiuti *(stime ottenute con dati pesati)*

ETÀ' A IDC	Età calcolata al Cpa								
	--5	6-10	11-14	15-19	20-24	25-29	30-34	35-39	40-44
--5	2.789.382	19.053	8.084	2.296	2.013	3.510	2.137	3.272	1.201
6-10	20.939	2.399.526	9.833	8.977	2.396	1.931	2.004	2.504	1.944
11-14	15.200	8.244	2.084.186	20.662	10.880	-	1.269	600	1.958
15-19	3.840	14.731	15.689	2.643.388	16.085	4.367	658	3.118	2.851
20-24	3.676	3.565	5.276	9.918	3.099.814	29.547	12.216	347	2.534
25-29	2.453	2.092	-	5.984	25.790	3.904.571	19.494	8.504	865
30-34	3.703	1.268	1.954	1.073	6.181	17.504	4.000.551	31.812	18.686
35-39	3.993	3.957	364	3.453	959	7.126	21.850	4.160.064	15.720
40-44	2.695	1.303	655	2.908	2.587	857	12.122	12.139	3.610.239
45-49	483	2.688	1.388	1.866	3.205	953	62	8.317	25.331
50-54	635	1.617	521	733	3.417	2.318	3.625	-	7.248
55-59	1.093	683	-	344	1.160	3.291	1.004	2.888	395
60-64	2.908	-	1.619	167	496	1.819	3.174	1.258	5.117
65-69	972	2.658	-	547	159	518	674	1.812	1.216
70-74	1.325	379	1.658	-	979	268	878	899	2.139
75 e oltre	1.478	1.671	827	501	4.089	1.192	2.118	1.085	1.030
Vna (a)	970	-	-	-	-	154	271	-	-
Vm (b)	28.248	17.315	14.934	23.716	29.026	33.964	37.276	31.334	28.768
Totale	**2.883.993**	**2.480.750**	**2.146.988**	**2.726.533**	**3.209.236**	**4.013.890**	**4.121.383**	**4.269.953**	**3.727.242**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.3 segue **- Confronto tra le risposte al Censimento e all'Indagine di copertura per classi di età in anni compiuti** *(stime ottenute con dati pesati)*

ETÀ A IDC	Età calcolata al Cpa									
	45-49	50-54	55-59	60-64	65-69	70-74	75 e oltre	Vna (a)	Vm (b)	Totale
--5	431	1.372	1.849	2.409	-	1.783	48.452	2.388	13.363	2.902.995
6-10	1.662	840	836	675	2.384	126	7.588	1.957	8.947	2.475.069
11-14	-	100	-	382	-	1.503	1.155	1.750	5.002	2.152.891
15-19	3.675	584	782	584	2.140	-	3.195	1.966	6.350	2.724.003
20-24	2.330	3.179	389	1.022	-	2.817	3.848	1.791	8.995	3.191.264
25-29	3.912	1.697	2.306	1.859	1.075	435	13.541	2.472	10.669	4.007.719
30-34	366	4.471	2.018	3.968	859	1.850	12.220	8.875	14.850	4.132.209
35-39	15.301	675	2.490	938	2.505	304	8.482	2.002	15.395	4.265.578
40-44	22.889	14.103	989	4.538	1.926	1.398	3.730	2.225	18.664	3.715.967
45-49	3.432.551	14.219	3.816	513	4.799	1.130	4.205	1.587	10.689	3.517.802
50-54	9.792	3.606.496	20.217	7.689	463	5.026	7.430	3.609	14.818	3.695.654
55-59	4.230	25.558	3.107.788	13.680	2.840	886	13.085	2.103	12.303	3.193.331
60-64	1.504	4.004	10.354	3.225.229	19.538	9.103	7.125	795	10.585	3.304.795
65-69	1.552	463	334	13.708	2.953.593	10.627	4.729	1.101	13.387	3.008.050
70-74	432	841	-	6.912	8.628	2.578.083	20.804	975	10.635	2.635.835
75 e oltre	1.521	1.209	1.280	1.195	3.188	18.562	4.229.789	1.775	28.975	4.301.485
Vna (a)	-	380	-	-	-	-	472	1.886	-	4.133
Vm (b)	26.579	31.857	23.435	30.987	26.192	23.679	44.116	408	5.953	457.787
Totale	**3.528.727**	**3.712.048**	**3.178.883**	**3.316.288**	**3.030.130**	**2.657.312**	**4.433.966**	**39.665**	**209.580**	**53.686.567**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.4 - Confronto tra le risposte al Censimento e all'Indagine di copertura per luogo di nascita *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa					Totale
	In questo comune	In altro comune italiano	All'estero	Vna (a)	Vm (b)	
In questo comune	21.639.320	293.971	7.302	-	319.665	22.260.258
In altro comune italiano	613.020	27.446.844	21.404	-	949.092	29.030.360
All'estero	10.852	26.034	1.767.777	-	108.284	1.912.947
Vna (a)	-	240	400	-	-	640
Vm (b)	84.063	354.746	15.505	-	28.048	482.362
Totale	**22.347.255**	**28.121.835**	**1.812.388**	**-**	**1.405.089**	**53.686.567**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.5 - Confronto tra le risposte al Censimento e all'Indagine di copertura per cittadinanza *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa					Totale
	Italiana	Straniera	Apolide	Vna (a)	Vm (b)	
Italiana	47.525.991	38.994	2.863	-	2.688.810	50.256.658
Straniera	49.796	983.090	2.318	-	51.272	1.086.476
Apolide	2.255	1.842	393	-	181	4.671
Vna (a)	1.665	182	-	-	454	2.301
Vm (b)	2.059.539	16.792	-	-	260.130	2.336.461
Totale	**49.639.246**	**1.040.900**	**5.574**	**-**	**3.000.847**	**53.686.567**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.6 - Confronto tra le risposte al Censimento e all'Indagine di copertura per stato civile *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa								Totale
	Celibe/nubile	Coniugato/a	Separato/a di fatto	Separato/a legalmente	Divorziato/a	Vedovo/a	Vna (a)	Vm (b)	
Celibe/nubile	20.381.836	203.414	8.704	14.369	28.429	17.207	-	226.586	20.880.545
Coniugato/a	219.268	25.949.531	43.661	41.475	5.895	33.787	-	150.465	26.444.082
Separato/a di fatto	2.796	27.375	130.126	51.722	5.242	1.277	-	2.604	221.142
Separato/a legalmente	5.399	12.870	44.148	672.988	35.942	1.771	-	8.406	781.524
Divorziato/a	15.800	7.017	4.453	34.827	568.700	8.785	-	5.093	644.675
Vedovo/a	17.032	43.334	1.159	4.248	9.672	4.154.199	-	25.554	4.255.198
Vna (a)	-	6.261	215	-	-	-	-	-	6.476
Vm (b)	243.506	165.098	2.326	5.120	2.429	27.520	-	6.926	452.925
Totale	**20.885.637**	**26.414.900**	**234.792**	**824.749**	**656.309**	**4.244.546**	**-**	**425.634**	**53.686.567**

(a) Vna – Valore non ammissibile.
(b) Vm – Valore mancante.

Tavola B.7 - Confronto tra le risposte al Censimento e all'Indagine di copertura per stato civile prima dell'ultimo matrimonio. Stime limitate agli individui che hanno dichiarato lo stato civile diverso da "celibe/nubile" al quesito sullo stato civile *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa					Totale
	Celibe/nubile	Divorziato/a	Vedovo/a	Vna (a)	Vm (b)	
Celibe/nubile	27.070.124	88.015	93.966	-	1.964.097	29.216.202
Divorziato/a	40.212	279.331	3.036	-	8.813	331.392
Vedovo/a	28.648	2.416	122.609	-	15.141	168.814
Vna (a)	1.359	-	-	-	-	1.359
Vm (b)	1.450.458	14.437	10.509	-	282.374	1.757.778
Totale	**28.590.801**	**384.199**	**230.120**	**-**	**2.270.425**	**31.475.545**

(a) Vma - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.8 - Confronto tra le risposte al Censimento e all'Indagine di copertura per titolo di studio. Stime limitate agli individui con età in anni compiuti superiore a 5 *(stime ottenute con dati pesati)*

RISPOSTA A IDC (a)	Risposte al Cpa							
	1	2	3	4	5	6	7	8
1	512.256	135.280	43.252	6.804	708	-	790	-
2	178.282	3.293.190	664.180	49.724	700	764	-	178
3	58.549	580.763	11.171.259	569.812	3.156	2.464	774	107
4	10.065	59.159	583.363	13.316.520	23.782	20.883	4.198	6.271
5	-	1.155	5.877	37.533	708.524	21.471	3.629	1.036
6	1.301	601	5.437	25.151	13.942	1.161.667	6.210	-
7	-	1.019	785	8.366	2.279	7.793	193.398	-
8	-	371	328	5.769	1.458	1.262	402	132.497
9	218	4.585	22.642	530.994	7.268	6.665	3.888	1.177
10	475	1.531	4.604	24.306	3.887	1.111	2.494	464
11	-	-	678	9.341	178	62	-	9.669
12	3.151	4.614	9.335	159.523	10.582	19.785	5.368	1.087
13	499	882	3.631	17.846	3.337	3.719	1.450	313
14	-	760	910	7.869	3.631	4.927	795	4.793
15	369	211	383	4.080	2.014	2.215	681	1.270
16	596	2.904	3.631	15.997	19.267	22.314	3.039	2.118
Vna (b)	-	215	99	-	370	-	370	-
Vm (c)	34.509	200.434	170.567	237.038	12.541	14.298	3.942	326

(a) Legenda:
1 - Nessun titolo di studio e non sa leggere o scrivere.
2 - Nessun titolo di studio, ma sa leggere e scrivere.
3 - Licenza di scuola elementare.
4 - Licenza di scuola media inferiore o di avviamento professionale.
5 - Liceo classico.
6 - Liceo scientifico.
7 - Liceo linguistico.
8 - Liceo artistico (corso di 4-5 anni).
9 - Istituto professionale.
10 - Scuola magistrale.
11 - Istituto d'arte.
12 - Istituto tecnico (corso di 5 anni).
13 - Istituto magistrale (corso di 4-5 anni).
14 - Diploma non universitario post maturità (ad esempio Accademia di belle arti, Scuola di archivistica, Istituto di musica pareggiato, eccetera).
15 - Diploma universitario (Scuola diretta a fini speciali o parauniversitaria, Laurea breve).
16 - Laurea.

(b) Vna – Valore non ammissibile.
(c) Vm – Valore mancante.

Tavola B.8 segue **- Confronto tra le risposte al Censimento e all'Indagine di copertura per titolo di studio. Stime limitate agli individui con età in anni compiuti superiore a 5** *(stime ottenute con dati pesati)*

RISPOSTA A IDC (a)	Risposte al Cpa										Totale
	9	10	11	12	13	14	15	16	Vna (a)	Vm (b)	
1	1.841	376	-	1.247	371	673	-	245	-	13.452	717.295
2	3.864	1.731	-	2.514	364	-	169	806	-	119.443	4.315.909
3	16.159	4.505	369	9.684	2.688	659	474	790	-	194.616	12.616.828
4	277.604	17.700	8.133	79.980	14.294	3.119	3.535	11.203	-	345.245	14.785.054
5	8.840	3.635	1.100	11.086	1.881	1.784	3.250	9.324	-	25.762	845.887
6	7.174	1.368	922	16.907	3.458	5.592	2.934	8.878	-	29.464	1.291.006
7	4.900	2.301	-	10.139	-	3.027	692	434	-	13.015	248.148
8	1.308	759	12.229	628	892	2.328	1.113	781	-	5.197	167.322
9	2.244.747	10.295	5.884	269.340	8.985	15.736	7.152	2.101	-	161.140	3.302.817
10	13.924	445.607	1.613	8.771	194.201	6.919	2.061	3.094	-	39.517	754.579
11	6.629	369	157.819	1.433	238	2.396	1.943	-	-	10.992	201.747
12	287.689	5.874	3.337	4.560.906	10.471	13.560	8.084	13.435	-	130.376	5.247.177
13	6.194	223.857	687	11.961	640.413	4.961	5.630	3.845	-	55.153	984.378
14	12.644	3.725	8.426	7.127	3.998	93.031	16.989	5.360	-	33.880	208.865
15	5.790	1.233	364	6.862	3.821	18.168	197.776	12.776	-	25.520	283.533
16	6.407	5.219	1.919	21.750	8.385	6.366	23.513	2.888.374	-	173.714	3.205.513
Vna	-	-	-	-	-	-	-	-	-	-	1.054
Vm	37.602	13.270	4.260	44.463	9.785	5.786	7.483	38.322	-	53.562	888.188
Totale	**2.943.316**	**741.824**	**207.062**	**5.064.798**	**904.245**	**184.105**	**282.798**	**2.999.768**	**-**	**1.430.048**	**50.065.300**

(a) Legenda:
1 - Nessun titolo di studio e non sa leggere o scrivere .
2 - Nessun titolo di studio, ma sa leggere e scrivere.
3 - Licenza di scuola elementare.
4 - Licenza di scuola media inferiore o di avviamento professionale.
5 - Liceo classico.
6 - Liceo scientifico.
7 - Liceo linguistico.
8 - Liceo artistico (corso di 4-5 anni).
9 - Istituto professionale.
10 - Scuola magistrale.
11 - Istituto d'arte.
12 - Istituto tecnico (corso di 5 anni).
13 - Istituto magistrale (corso di 4-5 anni).
14 - Diploma non universitario post maturità (ad esempio Accademia di belle arti, Scuola di archivistica, Istituto di musica pareggiato, eccetera).
15 - Diploma universitario (Scuola diretta a fini speciali o parauniversitaria, Laurea breve).
16 - Laurea.

(b) Vna – Valore non ammissibile.
(c) Vm – Valore mancante.

Tavola B.9 - Confronto tra le risposte al Censimento e all'Indagine di copertura per titolo di studio ricodificato in 9 categorie. Stime limitate agli individui con età in anni compiuti superiore a 5 *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa											Totale
	1	2	3	4	5	6	7	8	9	Vna (a)	Vm (b)	
1 - Analfabeti	512.256	135.280	43.252	6.804	1.440	3.892	673	-	245	-	13.453	717.295
2 - Alfabeti privi di titolo di studio	178.282	3.293.190	664.180	49.724	2.582	7.146	-	169	806	-	119.830	4.315.909
3 - Licenza elementare	58.549	580.763	11.171.259	569.812	8.902	27.990	659	474	790	-	197.630	12.616.828
4 - Licenza media	10.065	59.159	583.363	13.316.520	208.903	219.263	3.119	3.535	11.203	-	369.924	14.785.054
5 – Diploma scolastico di qualifica (corso di 2-3 anni)	-	2.202	15.046	452.624	1.263.711	137.250	8.681	2.969	664	-	163.867	2.047.014
6 – Diploma di maturità (corso di 4-5 anni)	5.169	12.108	35.907	356.191	212.570	9.657.389	45.599	29.889	40.923	-	486.118	10.881.863
7 – Diploma terziario di tipo non universitario	-	760	910	7.869	9.143	36.587	93.031	16.989	5.360	-	38.216	208.865
8 – Diploma universitario	369	211	383	4.080	3.683	20.454	18.168	197.776	12.776	-	25.633	283.533
9 – Diploma di laurea	596	2.904	3.631	15.997	1.524	86.007	6.366	23.513	2.888.374	-	176.601	3.205.513
Vna (a)	-	215	99	-	-	740	-	-	-	-	-	1.054
Vm (b)	34.984	200.882	172.931	247.052	45.661	163.818	7.809	7.484	38.627	-	83.124	1.002.372
Totale	**800.270**	**4.287.674**	**12.690.961**	**15.026.673**	**1.758.119**	**10.360.536**	**184.105**	**282.798**	**2.999.768**	**-**	**1.674.396**	**50.065.300**

(a) Vna – Valore non ammissibile.
(b) Vm – Valore mancante.

Tavola B.10 - Confronto tra le risposte al Censimento e all'Indagine di copertura per condizione professionale o non professionale. Stime limitate agli individui con età in anni compiuti superiore a 14 *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa					
	1	2	3	4	5	6
1 - Occupata	18.181.135	84.526	219.959	53.061	40.350	109.672
2 - In cerca di prima occupazione	74.671	1.234.650	292.560	12.442	65.854	75.495
3 - Disoccupata (in cerca di nuova occupazione)	118.544	130.032	1.325.368	26.690	16.492	99.820
4 - In attesa di iniziare un lavoro che aveva già trovato	16.283	4.198	10.785	66.854	2.925	2.790
5 - Studente	54.612	69.436	27.090	2.785	3.302.008	14.138
6 - Casalinga	91.520	96.161	160.272	5.787	11.495	6.004.268
7 - Ritirata dal lavoro	116.987	3.059	19.737	-	2.662	499.959
8 - In servizio di leva o in servizio civile sostitutivo	2.629	2.946	1.781	261	1.445	-
9 - Inabile al lavoro	7.441	3.241	10.203	395	1.225	20.175
10 - In altra condizione	62.441	12.835	28.012	2.185	8.095	192.847
Vna (a)	395	-	-	-	-	357
Vm (b)	112.974	7.269	12.868	7.794	22.609	41.058
Totale	**18.839.632**	**1.648.353**	**2.108.635**	**178.254**	**3.475.160**	**7.060.579**

RISPOSTA A IDC	Risposte al Cpa						Totale
	7	8	9	10	Vna (a)	Vm (b)	
1 - Occupata	79.124	3.608	10.913	101.708	-	140.206	19.024.262
2 - In cerca di prima occupazione	1.807	4.560	7.931	21.464	-	32.761	1.824.195
3 - Disoccupata (in cerca di nuova occupazione)	19.777	795	8.628	35.491	-	26.774	1.808.411
4 - In attesa di iniziare un lavoro che aveva già trovato	1.468	-	-	2.703	-	1.580	109.586
5 - Studente	1.024	4.550	1.616	11.135	-	51.171	3.539.565
6 - Casalinga	437.907	496	35.623	247.901	-	94.934	7.186.364
7 - Ritirata dal lavoro	7.879.848	512	91.715	754.884	-	133.161	9.502.524
8 - In servizio di leva o in servizio civile sostitutivo	684	77.676	-	136	-	1.186	88.744
9 - Inabile al lavoro	56.724	-	351.412	41.045	-	9.594	501.455
10 - In altra condizione	364.040	782	50.216	787.125	-	40.302	1.548.880
Vna (a)	-	-	-	-	-	-	752
Vm (b)	51.760	1.992	4.276	14.722	-	9.304	286.626
Totale	**8.894.163**	**94.971**	**562.330**	**2.018.314**	**-**	**540.973**	**45.421.364**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.11 - Confronto tra le risposte al Censimento e all'Indagine di copertura per condizione professionale ricodificata in 2 classi. Stime limitate agli individui con età in anni compiuti superiore a 14 (*stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa				Totale
	Occupata	Diversa da "Occupata"	Vna (a)	Vm (b)	
Occupata	18.181.135	702.921	-	140.206	19.024.262
Diversa da "Occupata"	545.128	25.173.133	-	391.463	26.109.724
Vna (a)	395	357	-	-	752
Vm (b)	112.974	164.348	-	9.304	286.626
Totale	**18.839.632**	**26.040.759**	**-**	**540.973**	**45.421.364**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.12 - Confronto tra le risposte al Censimento e all'Indagine di copertura per tipo di attività lavorativa. Stime limitate agli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al CPA				Totale
	A tempo pieno	A tempo parziale	Vna (a)	Vm (b)	
A tempo pieno	15.480.836	433.977	-	574.613	16.489.426
A tempo parziale	186.740	1.321.459	-	47.023	1.555.222
Vna (a)	430	-	-	-	430
Vm (b)	112.331	12.603	-	11.123	136.057
Totale	**15.780.337**	**1.768.039**	**-**	**632.759**	**18.181.135**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.13 - Confronto tra le risposte al Censimento e all'Indagine di copertura per posizione nella professione. Stime limitate agli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa								Totale
	1	2	3	4	5	6	Vna (a)	Vm (b)	
1 - Dipendente	12.908.220	57.664	71.807	96.586	105.449	39.641	-	202.560	13.481.927
2 - Imprenditore	21.399	609.776	24.763	145.267	6.029	10.151	-	9.639	827.024
3 - Libero professionista	44.002	42.026	799.682	106.756	7.182	2.729	-	20.113	1.022.490
4 - Lavoratore in proprio	66.666	226.747	105.563	1.638.837	28.930	52.932	-	28.602	2.148.277
5 - Socio di cooperativa	31.943	5.534	1.287	16.515	153.055	847	-	3.935	213.116
6 - Coadiuvante familiare	23.507	8.296	4.948	22.347	2.345	213.676	-	5.245	280.364
Vna (a)	2.613	-	-	304	-	-	-	371	3.288
Vm (b)	157.148	5.369	6.537	14.665	3.987	2.004	-	14.939	204.649
Totale	**13.255.498**	**955.412**	**1.014.587**	**2.041.277**	**306.977**	**321.980**	**-**	**285.404**	**18.181.135**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.14 - Confronto tra le risposte al Censimento e all'Indagine di copertura per posizione nella professione ricodificata in due classi. Stime limitate agli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa				Totale
	Dipendente	Diversa da "Dipendente"	Vna (a)	Vm (b)	
Dipendente	12.908.220	371.147	-	202.560	13.481.927
Diversa da "Dipendente"	187.517	4.236.220	-	67.534	4.491.271
Vna	2.613	304	-	371	3.288
Vm	157.148	32.562	-	14.939	204.649
Totale	**13.255.498**	**4.640.233**	**-**	**285.404**	**18.181.135**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.15 - Confronto tra le risposte al Censimento e all'Indagine di copertura per tipo di rapporto di lavoro. Stime limitate agli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa				Totale
	Tempo indeterminato	Tempo determinato	Vna (a)	Vm (b)	
Tempo indeterminato	10.885.839	508.607	-	58.110	11.452.556
Tempo determinato	220.878	1.052.069	-	12.419	1.285.366
Vna (a)	-	-	-	-	-
Vm (b)	146.551	18.965	-	4.782	170.298
Totale	**11.253.268**	**1.579.641**	**-**	**75.311**	**12.908.220**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Tavola B.16 - Confronto tra le risposte al Censimento e all'Indagine di copertura per tipo di rapporto di lavoro. Stime limitate agli individui con età in anni compiuti superiore a 14 che si siano dichiarati occupati *(stime ottenute con dati pesati)*

RISPOSTA A IDC	Risposte al Cpa				Totale
	Tempo indeterminato	Tempo determinato	Vna (a)	Vm (b)	
Tempo indeterminato	10.885.839	508.607	-	58.110	11.452.556
Tempo determinato	220.878	1.052.069	-	12.419	1.285.366
Vna (a)	-	-	-	-	-
Vm (b)	146.551	18.965	-	4.782	170.298
Totale	**11.253.268**	**1.579.641**	**-**	**75.311**	**12.908.220**

(a) Vna - Valore non ammissibile.
(b) Vm - Valore mancante.

Appendice C
Metodologie per la stima dell'errore di risposta

C.1 - Modelli e stima dell'errore di risposta

Questo paragrafo presenta sinteticamente la teoria alla base di stimatori utilizzati per stimare la varianza di risposta e le altre quantità riportate nelle tabelle precedentemente introdotte. Il modello matematico di riferimento è quello introdotto Hansen, Hurwitz e Pritzker (1964 e successivamente ripreso da diversi autori tra i quali Biemer e Forsman, 1992, Särndal et al, 1992, Cap. 16 e Biemer, 2004).

Si consideri una popolazione finita *U* composta da *N* unità. Sia μ_k il valore vero della variabile di interesse per l'unità u_k della popolazione ($k = 1,2,\ldots,N$). Si supponga che per ogni unità della popolazione siano disponibili *T* misurazioni, y_{tk}, di tale variabile. Per tali misurazioni si assume che valga il seguente modello:

$$y_{tk} = \mu_k + b_{tk} + \varepsilon_{tk}, \quad t = 1,2,\ldots,T$$

dove:

$$E_m(\varepsilon_{tk}) = 0$$
$$Var_m(\varepsilon_{tk}) = \sigma^2_{tk}$$
$$Cov_m(\varepsilon_{tk}, \varepsilon_{tl}) = \sigma_{tkl}, \qquad t = 1,2,\ldots,T, \qquad k \neq l, \qquad k,l = 1,2,\ldots,N$$
$$Cov_m(\varepsilon_{tk}, \varepsilon_{t'k}) = \sigma_{tt'k} \qquad t \neq t' \qquad t,t' = 1,2,\ldots,T, \qquad k = 1,2,\ldots,N$$

La *m* al pedice sta a indicare che valore atteso varianza e covarianza sono riferite al modello di misurazione.

In pratica, questo modello assume che la misurazione di un fenomeno su una unità dia origine ad una quantità che è il frutto della somma del valore vero del fenomeno, μ_k, con due componenti di errore, una, b_{tk}, di natura sistematica e l'altra, ε_{tk}, di natura puramente casuale. In tal modo, nell'ipotesi di condurre sulla stessa unità una serie elevata di misurazioni (tutte nelle medesime condizioni) gli errori di natura casuale tenderanno a compensarsi mentre gli errori di natura sistematica, essendo tutti nella medesima direzione, tenderanno a sommarsi:

$$E_m(y_{tk}) = \mu_k + b_{tk} = \psi_{tk}.$$

Limitando l'attenzione al caso di sue sole misurazioni ($T = 2$) si può affermare che la coppia di misurazioni (y_{1k}, y_{2k}) è una variabile casuale bivariata con media (ψ_{1k}, ψ_{2k}) e matrice di varianza e covarianza:

$$\begin{pmatrix} \sigma^2_{1k} & \sigma_{12k} \\ \sigma_{21k} & \sigma^2_{2k} \end{pmatrix}$$

Si assuma adesso che l'obiettivo di inferenza sia la media del fenomeno nella popolazione. A tal fine si consideri di estrarre da *U* un campione casuale semplice *s* ($s \subseteq U$) di *n* unità, e di procedere alla doppia misurazione del fenomeno indagato per ciascuna unità campione. In tal caso, la media campionaria, $\bar{y} = (1/n)\sum_{k=1}^{n} y_k$, è uno stimatore non distorto, rispetto al disegno di campionamento, della media nella popolazione. In virtù del modello di misurazione, considerando la media campionaria delle osservazioni alla *t*-esima occasione d'indagine, $\bar{y}_t = (1/n)\sum_{k=1}^{n} y_{tk}$, accade che il valore atteso complessivo dello stimatore $\bar{y}_t$ (sia rispetto al disegno di campionamento *p* che al modello di misurazione *m*) è uguale alla somma della media dei veri valori nella popolazione, $\bar{\mu}_U$ (il parametro che si vuole stimare), con una quantità, $\bar{b}_{U\,t}$, che in quanto funzione degli errori di misurazione di natura sistematica rappresenta la *distorsione da misurazione*:

A cura di Giovanna Brancato

$$E(\bar{y}_t) = E_{pm}(\bar{y}_t) = E_p\left[E_m(\bar{y}_t|s)\right] = \frac{1}{N}\sum_{k=1}^{N}(\mu_k + b_{tk})$$

$$= \bar{\mu}_U + \bar{b}_{Ut}$$

Per la varianza complessiva dello stimatore in questione, vale il seguente risultato:

$$V_{pm}(\bar{y}_t) = E_p\left[V_m(\bar{y}_t|s)\right] + V_p\left[E_m(\bar{y}_t|s)\right]$$

$$= \left[\frac{n-1}{n}\mathrm{CRV}_t + \frac{\mathrm{SRV}_t}{n}\right] + \left(1 - \frac{n}{N}\right)\frac{1}{n}\mathrm{SV}_t$$

dove il primo termine (quantità nelle parentesi quadre) costituisce la *varianza dovuta agli errori di misurazione* mentre il secondo è la *varianza dovuta al campionamento*.

L'espressione della varianza campionaria segue direttamente dal disegno di campionamento prescelto che nel nostro caso è di tipo casuale semplice senza reinserimento. In essa:

$$\mathrm{SV}_t = \frac{1}{N-1}\sum_{k=1}^{N}(\psi_{tk} - \bar{\psi}_t)^2$$

Per quel che riguarda la varianza dovuta agli errori di risposta, la quantità SRV_t rappresenta la varianza media delle risposte dalla stessa unità alla stessa domanda in indagini ripetute:

$$\mathrm{SRV}_t = \frac{1}{N}\sum_{k=1}^{N}\sigma_{tk}^2$$

Essa è comunemente denotata come *varianza di risposta semplice* (Srv: *Simple Response Variance*). La componente della varianza dovuta agli errori di misurazione è denotata come componente correlata della varianza di risposta (Crv: *Correlated Response Variance*):

$$\mathrm{CRV}_t = \frac{2}{N(N-1)}\sum_{k=1}^{N}\sum_{l\neq k}\sigma_{tkl}$$

Essa è determinata da quei fattori che causano somiglianza tra gli errori che si manifestano nella osservazione di unità distinte. Uno dei fattori che maggiormente contribuisce a questo termine è l'intervistatore. Altri fattori, probabilmente meno rilevanti, potrebbero essere rappresentati da supervisori, codificatori, eccetera. Nella nostra trattazione questo termine viene considerato trascurabile e posto a 0, in quanto il contesto applicativo non implica intervistatori, trattandosi del censimento condotto con questionario auto-somministrato. In tal modo la varianza complessiva dello stimatore si riduce a due termini:

$$V_{pm}(\bar{y}_t) = \frac{1}{n}\mathrm{SRV}_t + \left(1 - \frac{n}{N}\right)\frac{1}{n}\mathrm{SV}_t$$

$$\cong \frac{1}{n}\mathrm{SRV}_t + \frac{1}{n}\mathrm{SV}_t$$

Quest'ultima approssimazione si ritiene valida quando la frazione di campionamento $f = n/N$ è trascurabile.

Per valutare l'impatto della varianza di risposta sulla varianza complessiva conviene calcolare l'*indice di inconsistenza*:

$$I_t = \frac{\mathrm{SRV}_t}{\mathrm{SRV}_t + \mathrm{SV}_t}$$

che ha il vantaggio di variare tra 0 e 1.

Per l'indice di inconsistenza vale la seguante regola empirica:[1]

$I \leq 0{,}20$	Varianza di risposta bassa
$0{,}20 < I < 0{,}50$	Varianza di risposta moderata
$I \geq 0{,}50$	Varianza di risposta elevata

C.1.1 - Stima della varianza semplice di risposta per variabili continue

Lo stimatore usuale per la varianza semplice di risposta al tempo $t = 1$ (nel nostro caso il censimento) è $\hat{\mathrm{SRV}}_1 = g/2$ dove

$$g = \frac{1}{n}\sum_{k=1}^{n}\left(y_{1k} - y_{2k}\right)^2$$

tale statistica è nota come *gross difference rate* (GDR). Si noti che il valore atteso di questo stimatore è (Biemer e Forsman, 1992):

$$E(g/2) = \frac{1}{2}\left[\mathrm{SRV}_1 + \mathrm{SRV}_2 - \frac{2}{N}\sum_{k=1}^{N}\sigma_{12k} + \frac{1}{N}\sum_{k=1}^{N}\left(\psi_{1k} - \psi_{2k}\right)^2\right]$$

Se ne deduce che lo stimatore è non distorto per SRV_1 solo se valgono le seguanti ipotesi:

a.1) $E_m\left(y_{1k}|s\right) = E_m\left(y_{2k}|s\right)$ per tutte le unità della popolazione, ossia se l'errore di natura sistematica commesso nelle due misurazione di una stessa unità è costante ($b_{1k} = b_{2k} = b_k$; $k = 1,2,\ldots,N$).

a.2) $\mathrm{SRV}_1 = \mathrm{SRV}_2$ ($\sigma_{1k} = \sigma_{2k} = \sigma_k$, $k = 1,2,\ldots,N$), ossia identica varianza di risposta semplice alle due misurazioni;

a.3) $\sigma_{12k} = 0$ per tutte le unità; ossia indipendenza per i valori osservati sulla stessa unità nelle due misurazioni

Queste tre condizioni equivalgono ad avere una seconda misurazione che si possa configurare come una replicazione indipendente e identicamente distribuita della prima misurazione. Si noti che se dovessero valere le prime due assunzioni ma non la terza, avendosi correlazione positiva tra le successive misurazioni sulla stessa unità ($\sigma_{12k} > 0$), allora $g/2$ fornirebbe una sottostima della varianza semplice risposta.

C.1.2 - Stima della varianza semplice di risposta per variabili categoriali

Per i dati qualitativi è necessario introdurre una nuova parametrizzazione del modello di misurazione differente da quella per le variabili continue. Per semplicità conviene trattare il caso di variabili dicotomiche (anche nel caso di più di due modalità di risposta ci si può ricondurre al caso dicotomico considerando ciascuna modalità in relazione a tutto il resto).[2]

Per dati dicotomici il valore vero, μ_k, è pari a 0 o 1 ed analogo discorso vale per i valori osservati alla *t*-esima misurazione, y_{tk}. Pertanto, si possono definire due *probabilità di errata classificazione*:

$\phi_{tk} = Pr\left(y_{tk} = 1|\mu_k = 0\right)$: probabilità di osservare 1 quando il vero valore è 0;

$\theta_{tk} = Pr\left(y_{tk} = 0|\mu_k = 1\right)$: probabilità di osservare 0 quando il vero valore è 1

Ai fini della stima della frequenza relativa di una certa caratteristica nella popolazione, $P = (1/N)\sum_{k=1}^{N}\mu_k$, in presenza di un campione casuale semplice senza reinserimento usualmente si fa ricorso alla frequenza relativa

[1] US Census Bureau. "Evaluating Censuses of population and Housing". *Statistical Training Document* ISP-TR-5 (1985).

[2] Hansen, M. H., W.N. Hurwitz e L. Pritzker. "The Estimation and Interpretation of Gross Differences and Simple Response Variance". In *Contributions to Statistics*, C. R. Rao, 111-136. Calcutta: Statistical Publishing Society, 1964.

campionaria, $p_t = (1/n)\sum_{k=1}^{n} y_{tk}$. Tale stimatore è corretto rispetto al disegno di campionamento ma, per via degli errori di misurazione, il suo valore atteso complessivo è (Biemer e Forsman, 1992):

$$E(p_t) = E_p\left[E_m\left(p_t|s\right)\right] = (1/N)\sum_{k=1}^{N}\mu_k + \left(-P\bar{\theta}_t + Q\bar{\phi}_t\right)$$
$$= P + B_t$$

in cui $Q = 1 - P$, $\bar{\theta}_t = (1/N_1)\sum_{k=1}^{N_1}\theta_{tk}$ è la probabilità di un errore di "falso negativo" alla *t*-esima misurazione (la sommatoria è estesa alle sole $N_1 = PN$ unità per le quali $\mu_k = 1$) e, infine, $\bar{\phi}_t = (1/N_0)\sum_{k=1}^{N_0}\phi_{tk}$ è la probabilità di un errore di "falso positivo" alla *t*-esima misurazione (la sommatoria è estesa alle sole $N_0 = N - N_1$ unità per le quali $\mu_k = 0$). Quindi p_t è uno stimatore distorto per *P* a meno che non valga l'uguaglianza $P\bar{\theta}_t = Q\bar{\phi}_t$. Tale uguaglianza si verifica raramente nella pratica, con l'eccezione dei casi in cui $\bar{\phi}_t$ e $\bar{\theta}_t$ non siano piuttosto piccoli.

In tale contesto, la varianza semplice di risposta è

$$\mathrm{SRV}_t = P\left[\bar{\theta}_t\left(1-\bar{\theta}_t\right) - \sigma^2_{\theta t}\right] + Q\left[\bar{\phi}_t\left(1-\bar{\phi}_t\right) - \sigma^2_{\phi t}\right]$$

in cui

$$\sigma^2_{\theta t} = \frac{1}{N_1}\sum_{k=1}^{N_1}(\theta_{tk} - \bar{\theta}_t)^2 \qquad \text{e} \qquad \sigma^2_{\phi t} = \frac{1}{N_0}\sum_{k=1}^{N_0}(\phi_{tk} - \bar{\phi}_t)^2$$

Ossia, la varianza semplice di risposta è una funzione del parametro da stimare, *P*, e delle probabilità di errata classificazione alla misurazione *t*-esima.

Anche in questo caso, la stima di Srv_1 è data dal metà del Gdr, $\hat{\mathrm{SRV}}_1 = g/2$, che, nel caso in questione, può essere espresso nel seguente modo.[3]

$$g = \frac{b+c}{n}$$

dove le quantità coinvolte sono le frequenze della seguente tabella di contingenza:

	Indagine principale		
Reintervista	$y_{1k} = 1$	$y_{1k} = 0$	Tot.
$y_{2k} = 1$	*a*	*b*	n_1
$y_{2k} = 0$	*c*	*d*	n_0
			n

In pratica, *g* non è altro che la frazione di unità che hanno risposto diversamente alla domanda nelle due misurazioni.

Il valore atteso del Gdr è dato da:

$$E(g) = P\left[\bar{\theta}_1 + \bar{\theta}_2 - 2\bar{\theta}_{12}\right] + Q\left[\bar{\phi}_1 + \bar{\phi}_2 - 2\bar{\phi}_{12}\right]$$

in cui

$$\bar{\theta}_{12} = \frac{1}{N_1}\sum_{k=1}^{N_1}\theta_{12k} \quad \text{e} \quad \bar{\phi}_{12} = \frac{1}{N_0}\sum_{k=1}^{N_0}\phi_{12k}$$

essendo θ_{12k} la probabilità la *k*-esima unità sia incorrettamente classificata in entrambe le misurazioni quando μ_k è pari a 1, e analogamente ϕ_{12k} la probabilità che l'unità sia incorrettamente nelle due misurazioni quando

[3] US Census Bureau. "Evaluating Censuses of population and Housing".

$\mu_k = 0$. Pertanto, $g/2$ fornisce una stima corretta di SRV_1 se (b.1) $\theta_{1k} = \theta_{2k}$; (b.2) $\phi_{1k} = \phi_{2k}$; (b.3) $\theta_{12k} = \theta_{1k}^2$ e (b.4) $\phi_{12k} = \phi_{1k}^2$, ossia se la reintervista rappresenta una replicazione identica della prima misurazione, essendo indipendente da quest'ultima. Se la replicazione dell'indagine non è indipendenti dalla intervista originale, perché, come spesso accade, gli individui ricordano le risposte fornite alla intervista originaria e le replicano alla reintervista allora $g/2$ fornisce una sottostima della reale varianza semplice di risposta.

Per quanto riguarda l'indice di inconsistenza, una sua stima è fornita da:[4]

$$\hat{I} = \frac{g}{p_1 q_1 + p_2 q_2}$$

Inoltre, il denominatore di questo indice può essere stimato in modo corretto sia da $p_1 q_1$ che da $p_2 q_2$. Ciò è valido anche in presenza di errori correlati. Altri possibili stimatori per *I* possono essere ottenuti usando $2 p_1 q_1$ oppure $p_1 q_2 + p_2 q_1$ al denominatore (US Census Bureau, 1985; Biemer, 2004)

C.1.3 - Stima della varianza semplice di risposta in presenza di disegni di campionamento complessi

La teoria sin qui illustrata si dimostra valida nel caso in cui la reintervista venga condotta su un campione casuale semplice dei rispondenti alla indagine originaria. Nel caso la reintervista preveda un campionamento complesso che comporti diverse probabilità di inclusione delle unità allora ai fini della stima della varianza di risposta semplice è necessario introdurre i pesi nel calcolo del GDR. In particolare nel caso di variabili continue converrà utilizzare la seguente formula:

$$g_w = \frac{1}{N} \sum_{k \in s} w_k (y_{1i} - y_{21})^2$$

o, ancora meglio[5]:

$$g_w = \frac{1}{\hat{N}} \sum_{k \in s} w_k (y_{1i} - y_{21})^2$$

dove $\hat{N} = \sum_{k \in s} w_k$. Le due espressioni coincidono nel caso che i pesi finali delle unità campione siano calibrati in modo che $\hat{N} = N$.

Nel caso di variabili categoriali si può far riferimento alla seguente formula

$$g_w = \frac{b_w + c_w}{\hat{N}}$$

in cui b_w e c_w sono le frequenze della tabella di contingenza ottenuta incrociando i dati della reintervista con quelli della intervista originaria tenendo conto dei pesi campionari.

C.2 - La stima degli indicatori di riferimento per la varianza semplice di risposta al Censimento

Le variabili su cui si è deciso di indagare sono tutte di natura categoriale prevedendo nella maggior parte dei casi più di due modalità di risposta. Nel seguito si riportano in dettaglio i passi seguiti per il calcolo dei vari indicatori presi in considerazione per valutare l'impatto degli errori di risposta sulle stime prodotte attraverso i dati censuari.

Innanzitutto è necessario partire dalla seguente tabella generica costruita per una certa domanda in comune tra le due indagini. In essa si incrociano le risposte ottenute alla Idc con quelle ottenute al Cen. Si considera il caso generale di una domanda che prevede due o più ($C \geq 2$) modalità di risposta.

[4] US Census Bureau. "Evaluating Censuses of population and Housing".
[5] US Department of Education - National Center for Education Statistics. "Reinterview Results for the School Safety & Discipline and School Readiness Components". *Technical Report*, NCES 97-339 (1997).

Idc	Risposte al Cen								
	Tot. Individui Linked	Valori Mancanti	**Tot.**	1	…	i	…	C	VNA
Tot. Individui linked	m_L	$m_{\text{VM,CEN}}$							
Valori Mancanti	$m_{\text{VM,IDC}}$	$m_{\text{VM,VM}}$							
Totale			n	$n_{\bullet 1}$	…	$n_{\bullet i}$	…	$n_{\bullet C}$	$n_{\bullet \text{VNA}}$
Categorie Risposta:									
1			$n_{1\bullet}$	n_{11}	…	n_{1i}	…	n_{1C}	$n_{1,\text{VNA}}$
…			…	…	…	…	…	…	…
i			$n_{i\bullet}$	n_{i1}	…	n_{ii}	…	n_{iC}	$n_{i,\text{VNA}}$
…			…	…	…	…	…	…	…
C			$n_{C\bullet}$	n_{C1}	…	n_{Ci}	…	n_{CC}	$n_{C,\text{VNA}}$
Valore Non Ammissibile (VNA)			$n_{\text{VNA}\bullet}$	$n_{\text{VNA},1}$	…	$n_{\text{VNA},i}$	…	$n_{\text{VNA},C}$	$n_{\text{VNA,VNA}}$

In essa:

- m_L è la somma dei pesi campionari associati agli individui che la procedura di *record linkage* ha accoppiato;
- $m_{\text{VM,CEN}}$ è la somma dei pesi campionari associati agli individui *linked* per i quali è stata ottenuta una risposta mancante al Cpa ma non alla Idc.
- $m_{\text{VM,IDC}}$ è la somma dei pesi campionari associati agli individui *linked* per i quali è stata ottenuta una risposta mancante al Idc ma non al Cen.
- $m_{\text{VM,VM}}$ è la somma dei pesi campionari associati agli individui *linked* per i quali è stata ottenuta una risposta mancante sia al Cpa che alla Idc.
- n è la somma dei pesi associati agli individui *linked* per i quali è stata ottenuta una risposta sia al Cpa che alla Idc. Tale quantità è data dalla seguente espressione:

$$n = m_L - m_{\text{VM,CEN}} - m_{\text{VM,IDC}} + m_{\text{VM,VM}}$$

Si noti che, i dati presi in considerazione non sono stati sottoposti ad *editing* e pertanto si considerano come risposte (valori non mancanti) anche valori osservati che però risultano al di fuori del campo di definizione $\{1,\ldots,C\}$ della variabile (denotati come valori non ammissibili).

C.2.1 - Stima del Net difference rate e del relativo intervallo di confidenza

Il *Net difference rate* (Ndr) è la differenza tra la frazione di individui che al Cpa hanno optato per una certa categoria di risposta e la medesima frazione calcolata per la Idc:[6]

$$d_i = \left(p_{\bullet i} - p_{i\bullet}\right) = \left(\frac{n_{\bullet i}}{n} - \frac{n_{i\bullet}}{n}\right), \qquad i = 1,\ldots,C$$

dove $p_{\bullet i} = n_{\bullet i}/n$ è la frequenza relativa della *i*-esima categoria al censimento (prima misurazione) mentre $p_{i\bullet} = n_{i\bullet}/n$ è la frequenza relativa della *i*-esima categoria alla indagine di copertura (seconda misurazione).

Solitamente tale indice viene espresso in termini percentuali ($d_i \times 100$). Esso misura come la variazione della categoria di risposta tra una occasione di indagine e la successiva vada a modificare la distribuzione finale delle risposte nelle diverse occasioni di indagine. In questo contesto, è particolarmente utile per stabilire se le assunzioni fatte ai fini della stima della varianza di risposta possano considerarsi attendibili. In particolare, se le seconda rilevazione può considerarsi una replicazione della prima, condotta cioè nelle medesime condizioni. In

[6] US Census Bureau. "Evaluating Censuses of population and Housing".

pratica, laddove l'Ndr risulti significativamente diverso da zero risulta difficile poter assumere che le due rilevazioni siano state condotte nelle medesime condizioni.

Talvolta il Ndr viene espresso in termini relativi:

$$d_{Ri} = \frac{|d_i|}{p_{\bullet i}} = \frac{|n_{\bullet i} - n_{i\bullet}|/n}{n_{\bullet i}/n} = \frac{|n_{\bullet i} - n_{i\bullet}|}{n_{\bullet i}}$$

e si considera la seguente regola empirica[7]

$d_{Ri} < 0{,}01$	Differenza trascurabile tra stima proporzioni in diverse indagini
$0{,}01 \le d_{Ri} \le 0{,}05$	Differenza moderata tra stima proporzioni in diverse indagini
$d_{Ri} > 0{,}05$	Differenza elevata tra stima proporzioni in diverse indagini

Ai fini del calcolo degli intervalli di confidenza delle stima del Ndr è necessario stimare la corrispondente varianza campionaria. In questo caso, tale operazione può risultare piuttosto difficoltosa dato che lo stimatore dell'Ndr si configura come un rapporto di variabili aleatorie e il disegno di campionamento è stratificato a più stadi. Per tale motivo conviene far riferimento a formule basate sulla approssimazione alla normale e derivate assumendo che il campionamento sia assimilabile ad un campionamento casuale semplice con frazione di campionamento trascurabile.[8] In genere per l'intervallo di confidenza al 95 per cento vale la seguente approssimazione:[9]

$$d_i \pm \left[\frac{2}{n}\sqrt{n_{\bullet i} + n_{i\bullet} - 2n_{ii} + 1}\right], \qquad i = 1, \ldots, C$$

Con l'eccezione dei seguenti casi particolari:

1) se $n_{\bullet i} \neq n_{ii}$ e $n_{i\bullet} = n_{ii}$

Limite di confidenza inferiore: $d_i - \frac{2}{n}\sqrt{n_{\bullet i} + n_{i\bullet} - 2n_{ii} + 1}$

Limite di confidenza superiore: $d_i + \left(\frac{2}{n} + \frac{2}{n}\sqrt{n_{\bullet i} + n_{i\bullet} - 2n_{ii} + 1}\right)$

2) se $n_{\bullet i} = n_{ii}$ e $n_{i\bullet} \neq n_{ii}$

Limite di confidenza inferiore: $d_i - \left(\frac{2}{n} + \frac{2}{n}\sqrt{n_{\bullet i} + n_{i\bullet} - 2n_{ii} + 1}\right)$

Limite di confidenza superiore: $d_i + \left(\frac{2}{n}\sqrt{n_{\bullet i} + n_{i\bullet} - 2n_{ii} + 1}\right)$

3) se $n_{\bullet i} = n_{ii}$ e $n_{i\bullet} = n_{ii}$

Limite di confidenza inferiore: $-4/n$

Limite di confidenza superiore: $+4/n$

[7] US Census Bureau. "Evaluating Censuses of population and Housing".
[8] Idem.
[9] Idem.

C.2.2 - Stima del Gross difference rate e del relativo intervallo di confidenza

Il *Gross difference rate* (Gdr) misura le discrepanze tra le risposte al censimento e quella alla reintervista. Esso può essere calcolato sia per ciascuna categoria di risposta che a livello aggregato. In quest'ultimo caso non è altro che la frazione di discrepanze osservate tra le due rilevazioni. Come anticipato precedentemente la sua metà può essere considerata una stima della varianza di risposta semplice.

Il Gdr per una data categoria è calcolato nel seguente modo:[10]

$$g_i = \frac{n_{\bullet i} + n_{i\bullet} - 2n_{ii}}{n}, \qquad i = 1, \ldots, C .$$

Solitamente il Gdr viene espresso in termini percentuali moltiplicando g_i per 100.

Per quel che riguarda gli intervalli di confidenza al 95 per cento si distinguono due casi:[11]

1) Se $g_i \leq 0{,}1$:

$$\left(g_i + \frac{2}{n} \right) \pm \frac{2}{n} \sqrt{n g_i + 1} .$$

2) se invece $g_i > 0{,}1$:

$$\left(g_i + \frac{2}{n} \right) \pm \frac{2}{n} \sqrt{n g_i (1 - g_i)}$$

Il Gdr a livello aggregato è dato dalla percentuale di discrepanze nella tabella di contingenza[12]

$$g = \frac{n - \sum_{i=1}^{C} n_{ii}}{n}$$

e anch'esso è solitamente espresso in termini percentuali ($g \times 100$).

Per quel che riguarda i relativi intervalli di confidenza al 95 per cento vale la seguente espressione approssimata:[13]

$$g \pm \frac{2}{n} \sqrt{n g (1 - g)}$$

C.2.3 - Stima dell'Indice di inconsistenza e del relativo intervallo di confidenza

L'indice di inconsistenza permette di ricavare un'idea della incidenza della varianza di risposta sulla varianza complessiva:

$$I = \frac{\mathrm{SRV}}{\mathrm{SRV} + \mathrm{SV}}$$

Anch'esso può essere calcolato per ciascuna categoria della variabile in esame:

$$I_i = \frac{\mathrm{SRV}_i}{\mathrm{SRV}_i + \mathrm{SV}_i}, \qquad i = 1, \ldots, C$$

Nell'ipotesi che la reintervista possa essere considerata una perfetta replicazione della prima misurazione, la stima di I_i è data da[14]

[10] US Census Bureau. "Census 2000 Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by reinterview". *Census 2000 Evaluation, Final Report B*.5 (2003): 9.
[11] Idem.
[12] Idem.
[13] Idem.
[14] US Census Bureau. "Evaluating Censuses of population and Housing".

$$\hat{I}_i = \frac{g_i}{p_{\bullet i}(1-p_{i\bullet}) + p_{i\bullet}(1-p_{\bullet i})}, \qquad i = 1, \ldots, C$$

Generalmente si preferisce riportare la stima dell'indice di inconsistenza in termini percentuali ($\hat{I}_i \times 100$).

Per quel che riguarda il calcolo degli intervalli di confidenza al 95 per cento (approssimati) si segue la stessa regola illustrata per il Gdr:

1) Se $g_i \leq 0{,}1$:

$$\frac{(g_i + 2/n) \pm (2/n)\sqrt{ng_i + 1}}{p_{\bullet i}(1-p_{i\bullet}) + p_{i\bullet}(1-p_{\bullet i})}$$

2) se invece $g_i > 0{,}1$:

$$\frac{(g_i + 2/n) \pm (2/n)\sqrt{ng_i(1-g_i)}}{p_{\bullet i}(1-p_{i\bullet}) + p_{i\bullet}(1-p_{\bullet i})}$$

Una stima dell'indice di inconsistenza aggregato è data da:

$$\hat{I} = \frac{g}{1 - \sum_{i=1}^{C} p_{\bullet i} p_{i\bullet}}$$

Tale stima si può ritenere attendibile se vale quanto già detto a proposito della stima dell'indice di inconsistenza per ciascuna categoria della variabile considerata.

La determinazione degli intervalli di confidenza al 95 per cento per l'indice di inconsistenza segue le stesse regole viste precedentemente:

1) Se $g \leq 0{,}1$:

$$\frac{(g + 2/n) \pm (2/n)\sqrt{ng + 1}}{1 - \sum_{i=1}^{C} p_{\bullet i} p_{i\bullet}}$$

2) se invece $g > 0{,}1$:

$$\frac{(g + 2/n) \pm (2/n)\sqrt{ng(1-g)}}{1 - \sum_{i=1}^{C} p_{\bullet i} p_{i\bullet}}$$

C.2.4 - Alcune considerazioni relativamente Intervalli di confidenza degli stimatori

Gli intervalli di confidenza presentati per Ndr, Gdr ed *I* sono tratti dal volume del US Census Bureau (1985). Tali intervalli di confidenza sono il frutto di approssimazioni dovute al fatto che la derivazione delle varianza campionaria degli stimatori in questione (Ndr, Gdr, *I*) risulta piuttosto difficoltosa. Infatti, tali stimatori sono degli stimatori per rapporto ossia funzioni non lineari delle osservazioni campionarie. La derivazione della loro varianza campionaria dovrebbe quindi passare attraverso una linearizzazione degli stessi; sfortunatamente le difficoltà aumentano laddove il disegno di campionamento risulti di tipo complesso (stratificato a più stadi). Per questi motivi, nella definizione degli intervalli di confidenza delle statistiche in questione US Census Bureau (1985) suggerisce di rifarsi alla teoria dell'inferenza classica utilizzando l'intervallo di confidenza per proporzioni $\hat{p} = X/n$ derivato con il metodo *score* da Wilson:[15]

$$\frac{X + z^2/2}{n + z^2} \pm \frac{z}{n + z^2}\sqrt{X - \frac{X^2}{n} + \frac{z^2}{4}}$$

[15] Brown, L. D., T. Toni Cai e A. DasGupta. "Interval Estimation for a Binomial Proportion". *Statistical Science*, 16, 2 (2001): 101-133.

in cui z è il percentile della distribuzione normale standardizzata. In tale caso si assume che $\hat{p} = X/n$ segua la distribuzione binomiale. Si noti che nel caso di $\alpha = 0{,}05$, si ha $z = 1{,}96 \cong 2$ e pertanto l'intervallo di confidenza si semplifica in:

$$\frac{X+2}{n+4} \pm \frac{z}{n+4}\sqrt{X - \frac{X^2}{n} + 1}$$

In presenza di fenomeni rari è preferibile utilizzare lo stesso metodo ma ipotizzando che $\hat{p} = X/n$ segua la distribuzione di Poisson:

$$\frac{X + z^2/2}{n} \pm \frac{z}{n}\sqrt{X + \frac{z^2}{4}}.$$

Nel caso che $\hat{p} = X/n$ si distribuisca secondo una binomiale Brown *et al.* (2001) suggeriscono di utilizzare il metodo di Wilson in presenza di campioni piccoli ($n \leq 40$) mentre negli altri casi si suggerisce di utilizzare la modifica dell'intervallo di Wilson proposta da Agresti-Coull:[16]

$$\tilde{p} \pm z\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

in cui $\tilde{p} = \tilde{X}/\tilde{n}$, essendo $\tilde{X} = X + z^2/2$ e $\tilde{n} = n + z^2$. E' facile osservare che i due intervalli sono entrambi centrati intorno a $\tilde{p} = \tilde{X}/\tilde{n}$, di contro l'intervallo di Wilson non supera mai in ampiezza quello di Agresti-Coull.

[16] Brown, L. D., T. Toni Cai e A. DasGupta. "Interval Estimation for a Binomial Proportion".

La qualità dei dati

Il volume fornisce gli elementi utili per valutare la qualità dei dati del 14° Censimento generale della popolazione e delle abitazioni del 21 ottobre 2001.
I tre capitoli di cui è composta la pubblicazione prendono in esame le diverse dimensioni delle qualità: nel primo si considerano gli aspetti legati al processo di produzione dei dati, nelle sue diverse fasi di acquisizione, di controllo e correzione; nel secondo si descrive la realizzazione dell'Indagine di copertura, condotta per verificare la corretta enumerazione degli individui, e se ne illustrano i principali risultati; nel terzo, infine, si presentano ulteriori analisi della qualità, quali le stime della varianza di risposta - calcolate utilizzando i dati dell'Indagine di copertura - necessarie per poter effettuare una valutazione della accuratezza dei dati raccolti.

The Quality of Data

The publication presents useful information to evaluate data quality for the 14th General Population and Housing Census, which was held on 21 October 2001.
The three chapters examine several dimensions of quality: the first deals with the aspects connected to data production process, considered in the phases of data recording, check, editing and imputation; the second chapter describes how the Post Enumeration Survey was implemented - which was carried out to check whether people were correctly enumerated - and it also presents the main results; the third chapter gives space to more detailed analysis of quality, such as estimates of response variance - calculated using data of the Post Enumeration Survey - which are necessary to correctly evaluate the accuracy of the collected data.

1C01200900220000

ISBN 978-88-458-1624-6

€ 20,00