DOCUMENTI ISTAT

n. 11/2007

Strategie di correzione del questionario sulla qualità della vita dell'infanzia e dell'adolescenza. Indagine multiscopo sulle famiglie Aspetti della vita quotidiana 2005

D. Adamo, D. Cardoni, V. Greco, S. Montecolle, S. Orsini, A. Ortenzi e M. Savioli

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

DOCUMENTI ISTAT

n. 11/2007

Strategie di correzione del questionario sulla qualità della vita dell'infanzia e dell'adolescenza. Indagine multiscopo sulle famiglie Aspetti della vita quotidiana 2005

D. Adamo(*), D. Cardoni(*), V. Greco(*), S. Montecolle(*), S. Orsini(*), A. Ortenzi(*) e M. Savioli(*)

(*) ISTAT - Servizio Struttura e dinamica sociale

Contributi e Documenti Istat 2007

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

INDICE

Abstract

Nel 2005 l'Istat è tornato a rivolgere un'attenzione specifica al mondo dell'infanzia e dell'adolescenza, progettando e realizzando un'indagine sulla qualità della vita dei bambini e dei ragazzi di età compresa tra 0 e 17 anni. La visibilità statistica di questo segmento di popolazione rappresenta una acquisizione importante nell'ambito delle statistiche sociali e, sebbene garantita nei suoi elementi fondamentali dal sistema di indagini multiscopo, l'esigenza di un *focus* informativo specifico è sempre attuale.
Il questionario sulla qualità della vita dei bambini e degli adolescenti è stato inserito nell'ambito dell'indagine annuale sulle famiglie Aspetti della vita quotidiana.
Nel complesso l'indagine nel 2005 si è avvalsa di tre modelli di rilevazione: il questionario base dell'intervista, contenente i questionari individuali (uno per ogni componente della famiglia) e un questionario familiare, il modello autocompilato (uno per ogni componente della famiglia) e il questionario rivolto ai componenti della famiglia da 0 a 17 anni.
La peculiarità e la complessità dell'indagine e dei modelli di rilevazione hanno reso necessario sviluppare adeguate procedure di controllo e correzione dei dati per gestire l'impatto derivante dalla presenza di un questionario rivolto ad una specifica sottopopolazione.
Il documento descrive la filosofia di controllo e correzione e le procedure adottare per gestire questo modulo di indagine all'interno del processo già consolidato di controllo e correzione dell' indagine Aspetti della vita quotidiana.

1. Premessa[1]

Nel 2005 l'Istat è tornato a rivolgere un'attenzione specifica al mondo dell'infanzia e dell'adolescenza, progettando e realizzando un'indagine sulla qualità della vita dei bambini e dei ragazzi di età compresa tra 0 e 17 anni. La visibilità statistica di questo segmento di popolazione[2] rappresenta una acquisizione importante nell'ambito delle statistiche sociali e, sebbene garantita nei suoi elementi fondamentali dal sistema di indagini multiscopo, l'esigenza di un *focus* informativo specifico è sempre attuale. Questa rilevazione giunge, infatti, ad una certa distanza di tempo dalla precedente che data ormai al 1998, quando nell'ambito dell'indagine "Famiglie e soggetti sociali" veniva realizzato un analogo approfondimento.
La rilevanza del tema trattato e le sue potenzialità informative hanno spinto a ripetere la rilevazione e questa esperienza ha, stavolta, coinvolto più soggetti istituzionali: il Ministero del Lavoro e delle Politiche Sociali e l'Istituto degli Innocenti, che hanno contribuito anche alla progettazione del questionario con le loro specifiche competenze tematiche.
Attraverso questa indagine l'Istat ha voluto mettere in evidenza la vita quotidiana dei bambini e dei ragazzi – lo studio, il gioco, il tempo libero - per fornire un quadro sulla condizione dell'infanzia nel contesto sociale e familiare in continuità con la precedente indagine. Ma un'attenzione particolare è stata riservata a quella che sembra una novità rilevante e sempre più presente del mondo dei bambini e dei ragazzi: il loro rapporto con i media, soprattutto quelli "nuovi" quali il pc, Internet e il telefono cellulare elementi, questi ultimi, che sette anni fa erano totalmente assenti dal panorama di osservazione.
Il questionario sulla qualità della vita dei bambini e degli adolescenti è stato inserito, questa volta, nell'ambito dell'indagine annuale sulle famiglie Aspetti della vita quotidiana. Questa indagine fa parte del sistema di indagini multiscopo sulle famiglie e dal 1993 viene realizzata ogni anno per fornire informazioni su molteplici aspetti della vita quotidiana. Oggetto dell'indagine è, appunto, la vita quotidiana che è vista come ambito unitario in cui i ruoli e le attività dei soggetti sociali s'intersecano e si fondono in un "tutto" organico. Scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita e rapporto con i servizi sono indagati in un'ottica in cui *oggettività* dei comportamenti e *soggettività* delle aspettative, delle motivazioni, dei giudizi contribuiscono a definire l'informazione sociale. È possibile, in questo modo, cogliere importanti aspetti legati alla qualità della vita, non solo in base all'osservazione diretta dei comportamenti, ma anche alle indicazioni che provengono dalla dimensione percettiva e autovalutativa delle persone.
Come si è detto dal 1993 l'indagine viene realizzata tutti gli anni perseguendo l'obiettivo di garantire la continuità delle serie storiche e quindi continuità dei contenuti, allo stesso tempo però si presta all'introduzione di nuovi quesiti. I contenuti informativi dell'indagine possono, infatti, definirsi fissi, rotanti e modulari.
I contenuti fissi sono, ovviamente, rilevati ogni anno e sono standardizzati sia nella formulazione (testo della domanda e modalità di risposta) sia nella collocazione nell'ambito della sezione di appartenenza (sequenza interna alla sezione); quelli rotanti riguardano fenomeni il cui *trend* evolutivo richiede una rilevazione approfondita con cadenza pluriennale; i contenuti modulari, infine, sono approfondimenti su temi specifici ed includono l'accoglimento della domanda di nuova informazione statistica.
Il questionario sulla qualità della vita dell'infanzia e dell'adolescenza appartiene a questa terza tipologia, rappresentando però, rispetto agli approfondimenti modulari effettuati nelle precedenti occasioni di indagine[3], un'importante innovazione. Mentre, infatti, tutti i precedenti approfondimenti tematici erano

[1] Il lavoro è frutto dell'attività di ricerca congiunta degli autori. In ogni caso, ai soli fini dell'attribuzione, i paragrafi 4.1 e 4.6 sono da attribuirsi a Domenico Adamo; i paragrafi 4.2, 5.4.3, 5.4.4, 5.4.7 a Damiana Cardoni; il paragrafo 5.4.6 a Valeria Greco; i paragrafi 4.3, 4.4, 5.5, 5.6, 6.2 a Silvia Montecolle; i paragrafi 5.4.1 e 5.4.2 ad Alessandro Ortenzi; la premessa e i paragrafi 2.1, 2.2 e 4.5 a Sante Orsini; il capitolo 3 e i paragrafi 2.3, 5,1, 5.2, 5.3, 5.4.5, 6.1 a Miria Savioli.

[2] La frattura culturale rispetto ad una tradizione, anche statistica, che vede i bambini come semplici appedici degli adulti: figli, scolari o studenti, avviene in Istituto con il terzo ciclo di indagini multiscopo quando si assume per la prima volta un ottica di infanzia come gruppo ben definito di popolazione.

[3] Per un quadro completo sui moduli di approfondimento inseriti nell'indagine Aspetti della vita quotidiana si veda il Prospetto 2.5, pagg. 31,32 del volume *Il sistema di indagini multiscopo*, Istat, collana Metodi e Norme n. 31.

stati inseriti come nuove sezioni all'interno dei questionari d'indagine (dal 1993 vengono utilizzati due questionari: uno somministrato per intervista diretta tramite un rilevatore comunale e uno autocompilato da ciascun componente della famiglia), l'indagine sull'infanzia e l'adolescenza ha previsto l'utilizzo di un questionario aggiuntivo somministrato per intervista a tutti i componenti della famiglia in età compresa tra 0 e 17 anni.
Nel complesso, quindi, l'indagine nel 2005 si è avvalsa di tre modelli di rilevazione: il questionario base dell'intervista, contenente i questionari individuali (uno per ogni componente della famiglia) e un questionario familiare, il modello autocompilato (uno per ogni componente della famiglia) e il questionario rivolto ai componenti della famiglia da 0 a 17 anni.
Di seguito sono illustrate le principali caratteristiche dell'indagine Aspetti della vita quotidiana e il processo di controllo e correzione che ha riguardato il questionario sui minori. La peculiarità e la complessità dell'indagine e dei modelli di rilevazione hanno reso necessario sviluppare adeguate procedure di controllo e correzione dei dati per gestire l'impatto derivante dalla presenza di un questionario rivolto ad una specifica sottopopolazione.

2. Le principali caratteristiche dell'indagine

2.1 Periodo di rilevazione e tecnica di campionamento

Nel 2005 l'indagine Aspetti della vita quotidiana è stata realizzata nel mese di febbraio. È un'indagine che prevede un campionamento a due stadi con stratificazione delle unità primarie. Le unità primarie sono costituite dai comuni italiani (sono stati estratti 852 comuni su 8.100 in base all'ampiezza demografica), le unità di secondo stadio sono le famiglie estratte in modo casuale dalle liste anagrafiche di ogni comune campione. Non sono state ammesse sostituzioni delle famiglie non intervistate[4].
L'indagine ha come popolazione di riferimento la popolazione residente in Italia, al netto dei membri permanenti nelle convivenze. Nel 2005, sono state intervistate 18.944 famiglie per un totale di 49.288 individui, di cui 8.739 con un'età compresa tra 0 e 17 anni.

2.2 Unità di rilevazione e unità di analisi

L'unità di rilevazione è la famiglia, intesa come un "insieme di persone dimoranti abitualmente nella stessa abitazione e legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi".
Non sono stati considerati membri della famiglia gli ospiti, i domestici o le persone che condividevano l'abitazione per motivi economici (affittuari, pensionanti, ecc). Inoltre non sono state intervistate le persone che avevano lasciato definitivamente la famiglia, anche se non avevano ancora effettuato il cambio di residenza (ad esempio, il figlio che si è sposato ed è andato a vivere con la moglie in un altro appartamento, ma ha ancora la residenza a casa dei genitori). Viceversa, per le persone con dimora abituale nelle abitazioni, ma temporaneamente assenti, le informazioni sono state raccolte intervistando i familiari presenti (intervista *proxy*).
Si fa comunque presente che la composizione della famiglia di fatto può differire da quella della famiglia anagrafica estratta dalle liste anagrafiche dei comuni. Nel caso in cui la famiglia di fatto differisca dalla famiglia anagrafica come descritta nello stato di famiglia, è la famiglia di fatto che viene rilevata.
Le unità di analisi sono le famiglie di fatto e gli individui che le compongono. Nel caso particolare del questionario sui minori, l'unità di analisi è costituita dai bambini e ragazzi di età compresa tra 0 e 17 anni.

---

[4] Per un approfondimento si veda l'appendice "Strategia di campionamento e livello di precisione dei dati" del volume *La vita quotidiana nel 2005*. Roma: Istat, (Informazioni n. 4).

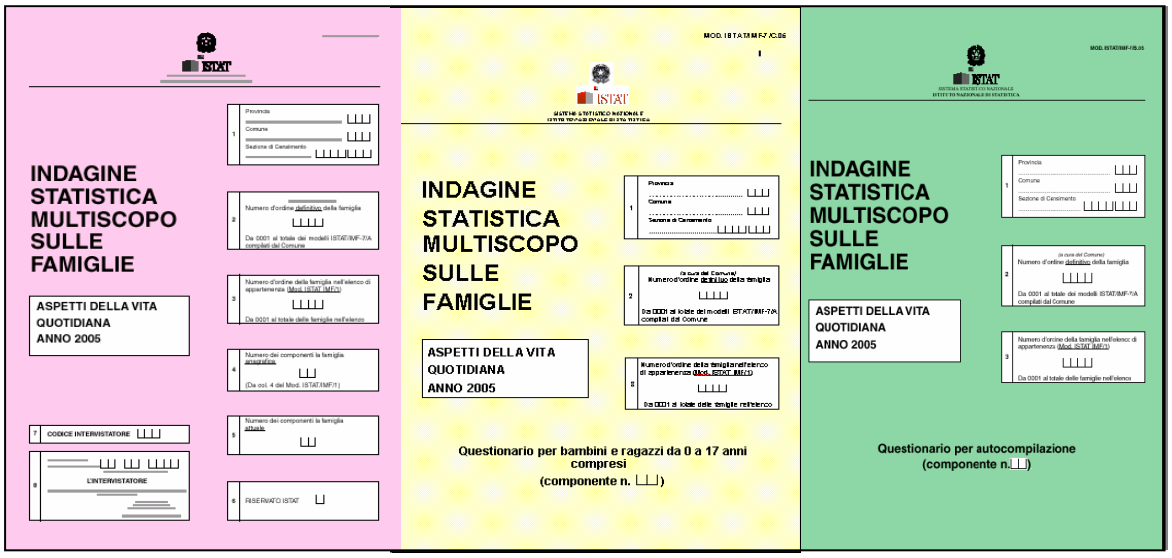
2.3 Modelli e tecnica di rilevazione

Nel 2005 per l'indagine Aspetti della vita quotidiana sono stati utilizzati tre questionari (Figura 2.1): il questionario base (MOD. ISTAT/IMF-7A.05 di colore rosa), che si compone di una scheda generale iniziale, dei questionari individuali e di un questionario familiare; un questionario per i bambini e ragazzi tra 0 e 17 anni (MOD. ISTAT/IMF-7C.05 di colore giallo) e, infine, un questionario autocompilato per tutti i componenti della famiglia (MOD. ISTAT/IMF-7B.05 di colore verde).
Sono state utilizzate due diverse tecniche di raccolta dei dati.

A) L'intervista diretta (*face to face*) rivolta a tutti i componenti della famiglia, condotta da un rilevatore comunale ed effettuata con tecnica PAPI (*Paper and Pencil interview*). Il rilevatore ha utilizzato il modello cartaceo (MOD. ISTAT/IMF-7A.05 di colore rosa) composto da:

- una scheda generale, dove sono state rilevate le informazioni sulle caratteristiche demografiche e socio-economiche di ogni componente la famiglia: relazione di parentela con l'intestatario dello stato di famiglia (persona di riferimento), sesso, data di nascita, titolo di studio, condizione e posizione nella professione, stato civile, anno del matrimonio, eccetera;
- un questionario individuale, in cui sono state rilevate per ciascun componente informazioni su diverse aree tematiche relative allo studio e alla formazione scolastica, agli spostamenti quotidiani, allo stato di salute, all'uso e soddisfazione per i servizi sanitari, socio assistenziali, ospedalieri e ad alcune attività del tempo libero tra cui le vacanze, l'attività fisica e sportiva e le relazioni amicali. Hanno risposto direttamente tutti i componenti di 14 anni e più. Per i bambini con meno di 14 anni ha risposto sempre un adulto, preferibilmente la madre;
- un questionario familiare, in cui sono state rilevate varie informazioni, tra cui l'uso e la soddisfazione per i servizi di elettricità e gas, la zona e l'abitazione in cui vive la famiglia, l'accessibilità ai servizi, il possesso di elettrodomestici, la situazione economica della famiglia.
  Al questionario ha risposto un solo componente la famiglia, preferibilmente la donna o un altro adulto della famiglia.

Figura 2.1: *Modelli di rilevazione. Indagine Aspetti della vita quotidiana – Anno 2005*



All'intervista diretta rivolta a tutti i componenti della famiglia, è seguita una seconda intervista *face to face* solo per i bambini e ragazzi da 0 a 17 anni, condotta dal rilevatore comunale. Il rilevatore ha utilizzato un modello cartaceo (MOD. ISTAT/IMF-7C.05 di colore giallo), dove sono state approfondite diverse aree tematiche sulla vita quotidiana dei bambini e ragazzi. In particolare l'affidamento del bambino, la scuola (utilizzo di attrezzature, partecipazione a corsi extra-scolastici, svolgimento dei compiti, eccetera), le attività svolte nel tempo libero, le relazioni amicali, il gioco, la fruizione della televisione e il possesso del telefono cellulare, l' autonomia, l'aiuto in casa e ai familiari.
I ragazzi tra i 14 e i 17 anni sono stati intervistati direttamente; per i bambini e ragazzi fino a 13 anni, invece, è stato intervistato un genitore, preferibilmente la madre, o un'altra persona adulta della famiglia.

B) L'autocompilazione di un questionario cartaceo (MOD. ISTAT/IMF-7B.05 di colore verde) dove sono stare rilevate informazioni sulla salute, gli stili alimentari, l'abitudine al fumo, l'utilizzo di vecchi e nuovi *media* (radio, tv, personal computer, internet), la fruizione di spettacoli fuori casa (cinema, teatro eccetera), l'abitudine alla lettura di libri e quotidiani, la soddisfazione per l'anno trascorso, l'uso e la soddisfazione di servizi di pubblica utilità (anagrafe, ASL, posta e banca, trasporti), la partecipazione politica e sociale.
Tutti i componenti di 14 anni e più hanno compilato personalmente il questionario. Per i bambini con meno di 14 anni ha provveduto alla compilazione un genitore o un altro adulto della famiglia.

3. Il processo di correzione

Alla fase della raccolta dei dati ha fatto seguito un articolato processo di correzione che si è basato su due operazioni sostanziali:

1. identificazione degli errori, attraverso l'esplorazione dei dati, basata su una reportistica (piani di *check*) che ne evidenzia anomalie e incoerenze;
2. correzione degli errori con procedure sia deterministiche che probabilistiche.

La prima operazione è fondamentale, poiché solo una puntuale identificazione di tutti i possibili errori ne permette la correzione. Allo stesso tempo la seconda operazione è estremamente delicata, poiché regole di correzione inadeguate possono inficiare ulteriormente la qualità dei dati.
Si è trattato di un processo di controllo, correzione e validazione dei dati, che ha avuto il duplice obiettivo di garantire un'elevata qualità delle stime prodotte e di produrre un archivio di dati elementari privo di incoerenze.
In questo processo i dati presenti nell'archivio sono stati analizzati non in maniera indipendente, ma in base a una precisa "gerarchia":

- Fase 1: correzione delle variabili demografiche e socio-economiche degli individui e delle variabili relative alla struttura della famiglia;
- Fase 2: correzione delle variabili tematiche rilevate nell'ambito del questionario sui minori.

La scelta di procedere gerarchicamente risiede nella necessità di stabilire l'esattezza delle informazioni fondamentali di famiglia/individuo, per poi assumerle come base per le successive informazioni, che ruotano attorno ad esse. Le variabili demografiche e socio-economiche relative agli individui e alle famiglie assumono tanta importanza da essere definite "pilastro", poiché sono utilizzate come filtri delle sezioni del questionario.
La Fase 1 è costituita da un insieme di procedure standardizzate. La raccolta ripetuta nel tempo di queste informazioni, infatti, ha favorito la realizzazione di procedure automatizzate modulari e facilmente mantenibili che perseguono nell'ordine i seguenti obiettivi:

1. controllo e correzione dei codici identificativi territoriali, familiari e individuali che consentono di identificare univocamente l'unità di rilevazione e di analisi;
2. costruzione di variabili derivate dal disegno campionario relative al contesto territoriale quali la ripartizione geografica e il tipo di comune;
3. controllo e correzione del sesso e dell'età degli individui;
4. controllo e correzione delle variabili socio-economiche (titolo di studio, condizione e posizione nella professione, attività economica);

5. ricostruzione della famiglia nella sua composizione interna, attraverso il controllo e la correzione di variabili quali la relazione di parentela, lo stato civile e l'anno di matrimonio.

La Fase 2 è caratterizzata da un percorso di correzione più lungo e articolato del precedente, nel quale è stato implementato un complesso piano di *check*, volto ad evidenziare anomalie e incoerenze nei dati. Gli errori riscontrati per queste variabili sono stati sanati tramite interventi di correzione, sia deterministici, sia probabilistici che hanno consentito di ottenere un archivio di dati elementari privo di anomalie e incoerenze.

Questo processo ha riguardato 319 variabili tematiche, rilevate attraverso il questionario sui minori, e 8.739 record, pari al numero di bambini e ragazzi di età compresa tra 0 e 17 anni presenti nel campione.

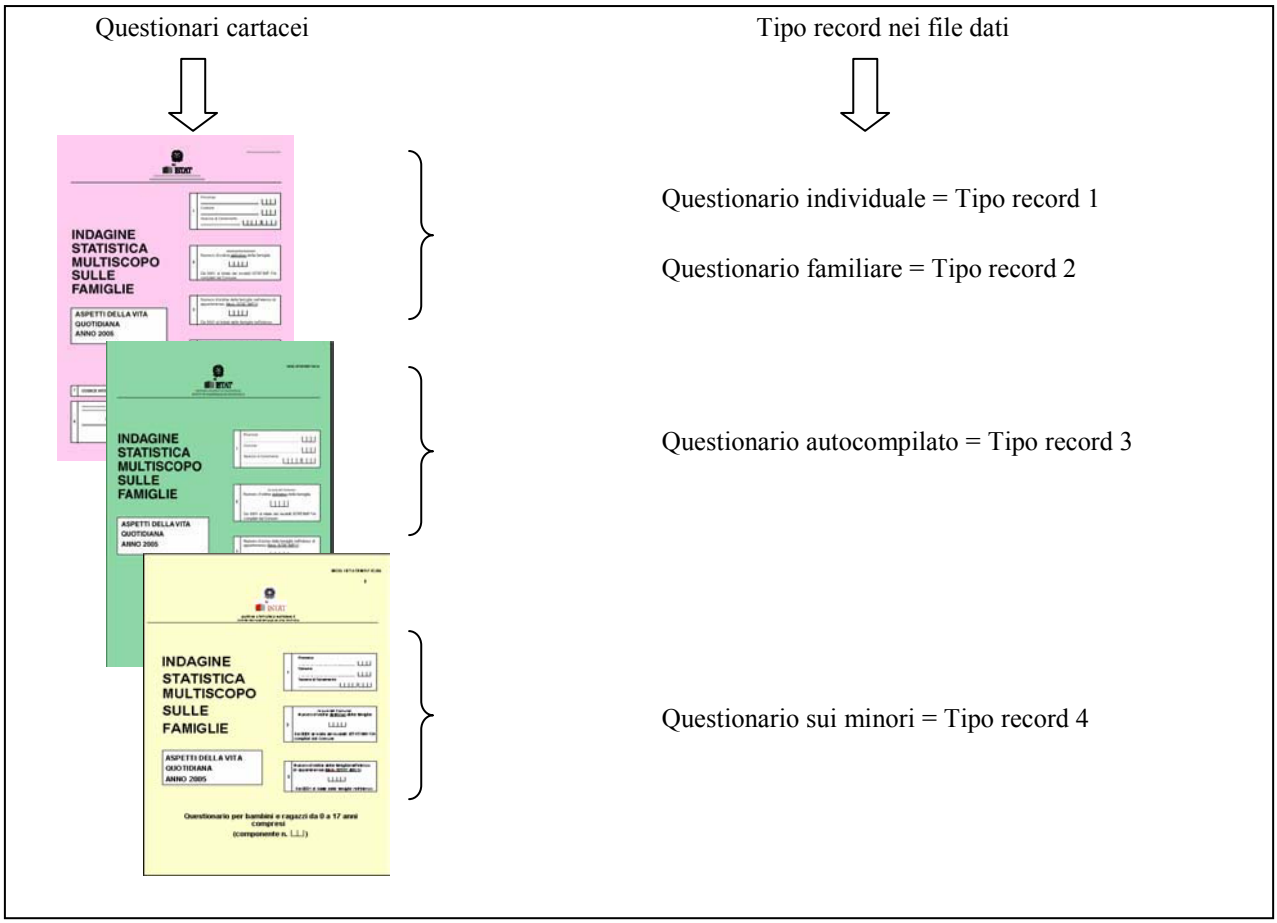
4. La correzione delle variabili demografiche e socio-economiche

4.1 Controllo e correzione dei codici identificativi territoriali, familiari ed individuali

Terminata la fase di rilevazione ogni comune ha inviato all'Istat i questionari compilati. Una volta arrivati, i modelli sono stati controllati da personale specializzato; si è trattato di una revisione quantitativa, consistente nel conteggio, per ciascun comune, delle famiglie intervistate e di quelle che non hanno partecipato all'indagine, nonché dei questionari compilati dai bambini di 0-17 anni.

I modelli revisionati sono stati inviati alla ditta di registrazione, sulla base di un calendario in cui sono stati stabiliti con precisione i tempi di revisione, di invio in registrazione, nonché di ritorno del materiale su supporto informatico. La consegna dei modelli alla società di registrazione è avvenuta con 9 invii.

Figura 4.1: *Modelli di rilevazione e tipo record. Indagine Aspetti della vita quotidiana – Anno 2005*



-

I file di ritorno dalla registrazione sono di tipo sequenziale (in formato Ascii) e contengono diverse tipologie di record (Figura 4.1):

- tipo record 1: è un record per ogni individuo, contiene informazioni raccolte con la scheda generale e il questionario individuale (modello rosa);
- tipo record 2: è un record per ogni famiglia, contiene informazioni raccolte con il questionario familiare (modello rosa);
- tipo record 3: è un record per ogni individuo, contiene informazioni raccolte con il questionario autocompilato (modello verde);
- tipo record 4: è un record per ogni individuo di età compresa tra 0 e 17 anni, contiene informazioni raccolte con il questionario sui minori (modello giallo).

Ad esempio per una famiglia di tre componenti composta da due adulti e un ragazzo con meno di 18 anni il file di ritorno dalla registrazione, nell'ipotesi di completa collaborazione all'intervista, conterrebbe: tre record di tipo 1, un record di tipo 2, tre record di tipo 3 e un record di tipo 4 per un totale di otto record.

Per ciascun file di ritorno dalla registrazione si è proceduto al controllo e alla correzione dei codici identificativi territoriali, familiari e individuali.

Il codice identificativo ha lo scopo fondamentale di distinguere una unità statistica dalle altre dello stesso tipo (in questa indagine una famiglia dalle altre famiglie e un individuo dagli altri individui). La sua univocità in un archivio di dati, oltre che la sua costruzione in modo gerarchico, consente l'assegnazione dell'unità statistica alla rispettiva unità di ordine superiore (l'individuo alla famiglia).

In particolare in questa fase di lavorazione (detta posizioni di verifica) sono state controllate e corrette le informazioni relative a:

- le chiavi territoriali, ovvero i codici provincia e comune dei Comuni appartenenti al campione;
- i codici identificativi della famiglia, ovvero all'interno di ciascun comune le variabili che identificano la famiglia (numero generale progressivo della famiglia e numero d'ordine della famiglia nell'elenco del comune di appartenenza);
- i codici identificativi degli individui, ovvero il numero d'ordine del componente all'interno della famiglia.

Inoltre in questa fase di lavorazione si è tenuto sotto controllo:

- il numero delle famiglie intervistate nell'ambito della provincia/comune, intesa come differenza tra numero di famiglie previste nel campione e numero di famiglie che non hanno partecipato all'indagine;
- le informazioni generiche sulla composizione della famiglia (numero di componenti della famiglia anagrafica, numero componenti della famiglia di fatto).

La presenza del tipo record 4 (relativo al questionario sui minori) ha comportato un trattamento specifico. In primo luogo la verifica della presenza del record di tipo 4 quando nella famiglia era presente un minorenne, individuato tramite un primo calcolo approssimativo dell'età. In secondo luogo è stato effettuato un controllo e una correzione preliminare delle variabili sesso e data di nascita dei bambini di 0-17 anni. Ciò ha facilitato la successiva fase di controllo e correzione delle variabili sesso ed età per i minorenni.

4.2 Creazione del file individuale

Dopo aver corretto i codici identificativi, il passo successivo è stato la creazione di un file unico dove le informazioni di ciascun individuo contenute nei tipo record 1, 3 e 4 (prima accodati uno di seguito all'altro) sono riportate su un unico record convenzionalmente chiamato tipo record 5, mentre le informazioni familiari (tipo record 2) sono ripetute alla fine di ciascun record individuale per ogni componente della stessa famiglia

In altre parole mentre nei file inviati dalla ditta di registrazione le informazioni relative ad un individuo si leggono su più righe, su ciascuna delle quali si ripetono i codici identificativi e i dati anagrafici, nel file finale le informazioni relative a ciascun individuo sono collocate su una stessa riga.

Figura 4.2: *Esempio di creazione del file individuale a partire dai file dopo la registrazione. Indagine Aspetti della vita quotidiana – Anno 2005*

FILE DOPO LA REGISTRAZIONE

	Codici identificativi		Variabili socio-demografiche		Variabili tematiche di ciascun tipo record	
	Tipo record	Numfam	Ordcom	Sesso	Età	P_1	P_2
riga 1	1	0001	1	1	40	1	2
riga 2	1	0001	2	2	38	2	1
riga 3	1	0001	3	1	13	2	1
riga 4	2	0001				2	2
riga 5	3	0001	1	1	40	80	180
riga 6	3	0001	2	2	38	54	168
riga 7	3	0001	3	1	13	69	170
riga 8	4	0001	3	1	13	1	2

FILE INDIVIDUALE (TIPO RECORD 5)

	Codici identificativi			Variabili socio-demografiche		Variabili tematiche di tutti i tipo record							
						Tipo record 1		Tipo record 3		Tipo record 4		Tipo record 2	
	Tipo record	Numfam	Ordcom	Sesso	Età	P_1 Spocon	P_2 Ricov	P_1 Peso	P_2 Stat	P_1 Pransc	P_2 Motpra	P_1 Telef	P_2 Elenc
riga 1	5	0001	1	1	40	1	2	80	180			2	2
riga 2	5	0001	2	2	38	2	1	54	168			2	2
riga 3	5	0001	3	1	13	2	1	69	170	1	2	2	2

Nell'esempio, nel file dopo la registrazione le prime 3 righe riportano le informazioni che ciascun componente ha fornito nel questionario individuale (tipo record 1):

1. il primo componente (*ordcom*=1) è un maschio di 40 anni (*sesso*=1; *eta*=40), non pratica sport con continuità (P_1=*spocon*=1) e negli ultimi 3 mesi è stato ricoverato (P_2=*ricov*=2);
2. il secondo componente (*ordcom*=2) è una femmina di 38 anni (*sesso*=2; *eta*=38), pratica sport con continuità (P_1=*spocon*=2) e negli ultimi 3 mesi non è stata ricoverata (P_2=*ricov*=1);
3. il terzo componente (*ordcom*=3) è un maschio di 13 anni (*sesso*=1; *eta*=13), pratica con continuità sport (P_1=*spocon*=2) e negli ultimi 3 mesi non è stato ricoverato (P_2=*ricov*=1).

La riga 4 riporta le informazioni raccolte con il questionario familiare (tipo record 2) che vengono inserite una sola volta indipendentemente dal numero dei componenti la famiglia:

4. la famiglia ha indicato che l'abitazione in cui vive dispone di telefono (P_1=*telef*=2) e che il numero telefonico dell'abitazione principale è riportato nell'elenco telefonico (P_2=*nelenc*=2).

Nelle righe 5,6 e 7 sono presenti le informazioni che ciascun componente della famiglia ha fornito riempiendo il questionario autocompilato (tipo record 3):

5. il primo componente pesa 80 Kg ed è alto 1 metro e 80 cm *(*P_1=*peso=80;* P_2=*stat=180)*;
6. il secondo componente pesa 54 kg ed è alto 1 metro e 68 cm *(*P_1=*peso=54;* P_2=*stat=168)*;
7. il terzo componente pesa 69 Kg ed è alto 1 metro e 70 cm *(*P_1=*peso=69;* P_2=*stat=170)*.

Nella riga 8 sono presenti le informazioni rilasciate dal componente con meno di 18 anni (tipo record 4):

8. il componente non consuma il pranzo a scuola (P_1=*pransc*=1) perché torna a casa prima di pranzo (P_2=*motpra*=2).

Il file dopo la creazione del file individuale (tipo record 5) presenta un numero di righe pari al totale dei componenti della famiglia (3 record) ed un numero di colonne pari al numero delle variabili rilevate (codici identificativi, variabili socio-demografiche, variabili tematiche presenti nei quattro questionari).

Nella Figura 4.2 si è considerato l'esempio di una famiglia composta da tre componenti, due adulti e un ragazzo con meno di 18 anni. Come si è detto, il file dopo la registrazione è composto per questa famiglia da otto record: tre record di tipo 1, un record di tipo 2, tre record di tipo 3 e, infine, un record di tipo 4 per il componente minorenne. Lo schema mostra sinteticamente il processo di creazione del file individuale e di come avviene il passaggio da otto record a tre record, uno per ciascun individuo.
Per semplicità nell'esempio vengono riportate soltanto alcune variabili: per i codici identificativi, il numero d'ordine della famiglia nell'elenco di appartenenza (*numfam*) e il numero d'ordine del componente familiare (*ordcom*); per le variabili socio-demografiche, il sesso (*sesso*) e l'età (*eta*)[5]; per le variabili tematiche, sono state selezionate, per ciascun tipo record, due variabili, che nei file inviati dalla ditta di registrazione si trovano nella stessa posizione, ma che ovviamente contengono informazioni diverse e seconda del tipo record a cui fanno riferimento (nella Figura 4.2 le posizioni di tali variabili tematiche sono genericamente indicate con P1, P2). Per il tipo record 1, in P1 e P2 ci sono le informazioni relative alla pratica sportiva continuativa (*spocon*) e agli eventuali ricoveri (*ricov*); per il tipo record 2, il possesso del telefono da parte della famiglia (*telef*) e la presenza del numero nell'elenco telefonico (*nelenc*); per il tipo record 3, il peso (*peso*) e la statura (*stat*); per il tipo record 4, il luogo abituale del pranzo (*pransc*) e se il ragazzo non pranza a scuola il motivo (*motpra*).
Alla fine del processo di creazione, il file individuale presenta un numero di righe pari al totale della popolazione intervistata ed un numero di colonne pari al numero delle variabili rilevate (codici identificativi, variabili socio-demografiche, variabili tematiche presenti nei quattro questionari).

4.3 Costruzione delle variabili territoriali

La strategia di campionamento adottata permette la produzione di stime valide a livello regionale, per grandi ripartizioni geografiche e per alcune tipologie comunali.
La referenziazione geografica delle interviste avviene associando le informazioni per provincia e comune di residenza dei rispondenti presenti sul file dati con le informazioni sulla tipologia del comune di residenza, che derivano dal disegno di campionamento. A partire dai codici di provincia e comune vengono create le variabili territoriali relative a:

- regione di residenza: 20 regioni, più le province autonome di Trento e Bolzano;
- ripartizione territoriale: Nord-ovest, Nord-est, Centro, Sud, Isole;
- ripartizione socio-demografica dei comuni: comuni centro dell'area metropolitana, comuni che gravitano intorno al centro dell' area metropolitana, comuni fino a 2.000 abitanti, comuni da 2.001 a 10.000 abitanti, comuni da 10.001 a 50.000 abitanti, comuni oltre 50.000 abitanti.

La costruzione delle variabili territoriali avviene preliminarmente alle procedure di controllo e correzione delle altre variabili perché fornisce una necessaria informazione di *breakdown* per le procedure stesse.

4.4 Controllo e correzione delle variabili sesso ed età

4.4.1 Le procedure standardizzate

La correzione delle incoerenze (informazioni errate o mancanti) relative al sesso e all'età degli intervistati assume un ruolo importante nel processo, in quanto tali informazioni sono determinanti per caratterizzare l'individuo, per definire con precisione la composizione della famiglia e per verificare i percorsi di compilazione dei questionari.
L'età dell'intervistato, in particolare, costituisce il criterio di ingresso a molte sezioni dei questionari, dove, di volta in volta, viene selezionata una specifica sottopopolazione di rispondenti. Inoltre, per l'indagine Aspetti della vita quotidiana del 2005, la presenza del questionario specifico per bambini e

[5] Lo schema riportato nella Figura 4.2 semplifica molto la costruzione del file individuale. Si tenga presente infatti che, nel file individuale, i valori delle variabili *sesso* ed *età* (o data di nascita) sono ripetuti tante volte quante sono le informazioni raccolte sui dati anagrafici nei diversi questionari (cfr § 4.4.1).

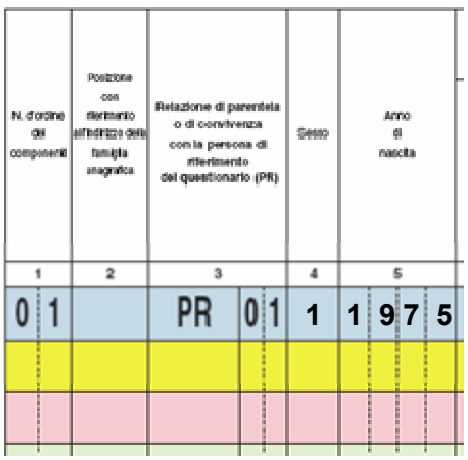
ragazzi di età compresa tra 0 e 17 anni ha reso particolarmente importante la corretta individuazione dei minorenni.
I dati anagrafici dell'intervistato sono informazioni richieste in diversi punti nei modelli di rilevazione dell'indagine (Figura 4.3): si rileva il sesso (*sesso1*) e l'anno di nascita (*anasc1*) nella scheda generale (modello rosa); il giorno (*gnasc2*), il mese (*mnasc2*), l'anno di nascita (*anasc2*), il sesso (*sesso2*) e l'età in anni compiuti (*eta*) nel questionario individuale (modello rosa) e il giorno (*gnasc3*), il mese (*mnasc3*), l'anno di nascita (*anasc3*) ed il sesso (*sesso3*) nel questionario autocompilato (modello verde).
Il sesso e la data di nascita (da cui si ricava l'età), oltre al numero d'ordine del componente familiare, sono variabili fondamentali per associare in modo esatto i diversi questionari ai corrispettivi componenti della famiglia e, di conseguenza, per costruire in modo corretto il file individuale a partire dai file con i tipo record 1, 2 e 3, ed è per questo motivo che è fondamentale rilevarli nei diversi questionari.
Di contro, l'esistenza di quesiti che rilevano lo stesso dato rende possibile la presenza su uno stesso record di indicazioni diverse che necessitano di un allineamento. Può accadere, quindi, a questo punto del processo, che su uno stesso record individuale siano presenti informazioni sul sesso e la data di nascita dell'individuo discordanti, dovute ad una errata compilazione del questionario o registrazione dei dati.
Per sanare questo tipo di discordanze dal 1993 sono state implementate in Sas delle procedure di controllo e correzione che hanno il fine di determinare univocamente il sesso e l'età di ogni individuo intervistato, allineando le informazioni raccolte nei vari punti dei questionari. Sono dette procedure standardizzate perché consentono di correggere ogni anno con gli stessi criteri variabili che vengono rilevate con le stesse modalità. Le procedure di correzione si basano fondamentalmente sul criterio della prevalenza: le informazioni su sesso e anno di nascita sono rilevate per tre volte, la più ricorrente è considerata quella valida.
Nella Figura 4.3 è riportato un esempio di incongruenza delle informazioni: la persona intervistata risulta essere nella scheda generale un maschio nato nel 1975 (la sua età è quindi 30 anni); nel questionario individuale una femmina, nata il 2 giugno del 1975 e di 30 anni (età dichiarata); nell'autocompilato un maschio nato il 2 giugno del 1955 e pertanto la sua età risulterebbe di 50 anni.
Il record risultante dalle risposte della Figura 4.3 è individuato come errato, sia dalla procedura di controllo della variabile sesso sia da quella della variabile età. Infatti, le variabili che si riferiscono al sesso contengono per due volte l'indicazione *maschio* (*sesso1*=1, *sesso3*=1) ed una volta quella di *femmina* (*sesso2*=2). La procedura di correzione, sulla base del criterio della prevalenza, modifica l'unico dato discordante allineandolo agli altri.
Allo stesso modo, la procedura per la correzione dell'età individua che per due volte è stato indicato l'anno di nascita uguale a 1975 (*anasc1*=1975, *anasc2*=1975) ed una volta anno di nascita pari a 1955 (*anasc3*=1955), per allineare i dati corregge l'anno di nascita che differisce dagli altri.
Per casi come questo riportato nell'esempio, le procedure agiscono automaticamente ed effettuano le correzioni, sostituendo il valore errato con quello esatto.
Quando i dati, però, sono tutti diversi tra loro o mancanti, il criterio della prevalenza non può essere utilizzato e le procedure non effettuano le correzioni in automatico. In queste situazioni dubbie (che riguardano un numero ridotto di casi) la correzione è demandata all'utente che, sul record errato, inserisce i valori che ritiene più opportuni basandosi sulle altre informazioni disponibili sull'individuo, che le procedure consentono di visualizzare in modo interattivo (informazioni aggiuntive[6]).
Al termine delle correzioni effettuate con le procedure standardizzate, nel file risultante (file di *output*) le tre variabili riferite al sesso e le tre riferite alle date di nascita assumono tutte lo stesso valore[7].

[6] Ci sono variabili che possono essere utilizzate come informazioni aggiuntive per la correzione del sesso o dell'età: ad esempio, per il sesso può essere indicativa la condizione professionale uguale a casalinga; per l'età variabili d'ausilio possono essere lo stato civile o la condizione professionale.

[7] L'età dichiarata è presa in considerazione dalla procedura per effettuare l'allineamento degli anni di nascita. Al termine delle correzioni assume lo stesso valore dell'età calcolata sulla base dei tre anni di nascita allineati.

Figura 4.3: *Quesiti su sesso e data di nascita, esempio di compilazione errata. Indagine Aspetti della vita quotidiana – Anno 2005*

Scheda generale (modello rosa)



Questionario individuale (modello rosa)

Questionario autocompilato (modello verde)

RECORD DOPO LA REGISTRAZIONE

			Scheda generale			Questionario individuale						Questionario autocompilato			
Tipo record	Numfam	Ordcom …	Sesso1	Anasc1	…	Gnasc2	Mnasc2	Anasc2	Sesso2	Eta	…	Gnasc3	Mnasc3	Anasc3	Sesso3
1	0001	1	1	1975		02	06	1975	2	30		.	.	.	.
3	0001	1	.	.		.	.	.	.	.		02	06	1955	1

RECORD NEL FILE INDIVIDUALE (TIPO RECORD 5)

Tipo record	Numfam	Ordcom …	Sesso1	Sesso2	Sesso3	…	Anasc1	Anasc2	Anasc3	Eta …
5	0001	1	1	2	1		1975	1975	1955	30

4.4.2 L'utilizzo delle informazioni aggiuntive per i minorenni

Nel 2005, solo per i ragazzi tra 0 e 17 anni, le informazioni sul sesso e l'anno di nascita sono state rilevate una quarta volta nel questionario sui minori. Ciò è stato dettato dall'esigenza di avere tutti i dati utili per collegare correttamente a ciascun componente minorenne della famiglia anche questo questionario.
La presenza di una quarta informazione, sia per il sesso che per l'anno di nascita, avrebbe comportato, qualora si fosse deciso di introdurla nelle procedure standardizzate di correzione, un riadattamento delle stesse alla nuova situazione. Le procedure utilizzate dal 1993 lavorano, infatti, allineando tre sessi e tre date di nascita, l'introduzione di un quarto dato avrebbe implicato la reingegnerizzazione dei *software* da attuare solo per l'indagine del 2005[8].

[8] La correzione del sesso ed in particolare dell'età per i minorenni è stata oggetto di maggiore attenzione già nella fase di posizione di verifica, dove, quando possibile, sono state apportate le prime correzioni di allineamento.

Per tale ragione si è deciso di non modificare le procedure standardizzate e di correggere i dati anagrafici di tutti gli individui intervistati nel modo usuale. Per i ragazzi di 0-17 anni il sesso e l'anno di nascita dichiarati nel questionario per i minori sono stati utilizzati nelle procedure standardizzate come informazione aggiuntiva, quindi, nelle situazioni in cui la procedura non ha attuato una correzione automatica e la decisione sul valore da inserire è stata demandata all'utente, la correzione dei dati anagrafici è stata fatta anche in base alle informazioni relative al questionario sui minori.
Ad esempio, nel caso del record di un individuo di 0-17 anni con i tre valori della variabile *sesso* tutti mancanti, la procedura di correzione non ha agito automaticamente e ha visualizzato il record all'utente, che ha effettuato la correzione, analizzando le informazioni aggiuntive, tra le quali anche il sesso indicato nel questionario sui minori.
L'utilizzo nel modo usuale delle procedure standardizzate ha però comportato una mancanza di controllo sul record dell'allineamento tra i valori inseriti automaticamente dalle procedure di correzione e quelli risultanti dalla registrazione del questionario dei minori. Per questo motivo, una volta terminata la correzione dei dati anagrafici con le procedure standardizzate è stato necessario un ulteriore passo di controllo per verificare l'allineamento tra il sesso e l'età presenti nel file dopo le procedure di correzione e quelli indicati sul questionario per i minori, utilizzando tabulati di controllo appositamente costruiti.
I casi discordanti sono stati analizzati e corretti. Per decidere come correggere si sono considerate le informazioni presenti sul file di input delle procedure di correzione (file grezzo), ovvero sono stati analizzati i valori presenti nel file prima di effettuare le correzioni automatiche di allineamento.
Per quanto riguarda la variabile *sesso*, per avere una visione sintetica delle tre informazioni presenti nel file grezzo, sono stati costruiti due indicatori. Questi hanno contato per ogni record quante volte le variabili che si riferivano al sesso nel file di input della procedura erano pari a femmina (*conta_f*) e quante volte a maschio (*conta_m*). Gli indicatori hanno assunto valori da 0 a 3 (vengono considerati i valori della variabile *sesso* indicati nella scheda generale, nel questionario individuale e nel questionario autocompilato). Nel caso di allineamento di tutte le informazioni si ha per i maschi *conta_f =0* e *conta_m =3*, per le femmine *conta_f =3* e *conta_m =0*[9].
Quando il valore della variabile *sesso* indicato nel questionario dei minori è risultato discordante con quello presente nel file dopo la procedura di correzione, la situazione è stata valutata tenendo presente i valori assunti da tali indicatori. Nella Tavola 4.1 sono riportati alcuni esempi di discordanza tra il valore della variabile *sesso* risultante dalla procedura e il dato relativo al questionario dei minori.

Tavola 4.1: *Esempi di record con informazioni discordanti sulla variabile sesso. Indagine Aspetti della vita quotidiana – Anno 2005*

RECORD	FILE DI INPUT DELLA PROCEDURA			INDICATORI COSTRUITI SUL FILE DI INPUT		FILE DI OUTPUT DELLA PROCEDURA	QUESTIONARIO MINORI
	Scheda generale Mod. rosa (*sesso1*)	Questionario individuale Mod. rosa (*sesso2*)	Questionario autocompilato Mod. verde (*sesso3*)	Conta_m	Conta_f	Valore dopo la procedura di correzione	Mod. giallo
Record_1	M	M	M	3	0	M	F
Record_2	M	F	M	2	1	M	F
Record_3	-	-	M	1	0	M	F

Nel record_1 c'è il caso di un minorenne con *conta_m=3*, *conta_f=0*, il valore della variabile *sesso* risultante dalla procedura è uguale a *maschio* (*sesso=M),* mentre il valore della variabile *sesso* del questionario sui minori è pari a *femmina (sesso=F).* Avere *conta_m=3* e *conta_f=0* significa che nel file grezzo le informazioni sul sesso erano già allineate e tutte uguali a *maschio* (la procedura di correzione ha considerato questo record come esatto). L'unico valore discordante, dunque, è quello del questionario dei minori; in questo caso la correzione ha riguardato quest'ultimo valore che è stato allineato agli altri.

[9] Se la somma dei due indicatori è minore di 3 vuol dire che almeno una delle tre informazioni è mancante nel file grezzo.

Il record_2 ha *conta_m=2*, *conta_f=1* e *sesso=F* dal questionario dei minori. La procedura di correzione per il criterio di prevalenza ha posto uguale a *maschio* il valore della variabile *sesso*, perchè le informazioni che aveva a disposizione indicavano due casi di *sesso=M* e un caso di *sesso=F*. La quarta informazione relativa alla registrazione del questionario dei minori, però, riportava di nuovo *femmina*. In realtà, quindi, le due alternative maschio o femmina erano parimenti probabili.
Analogamente, nel record_3 si ha *conta_m=1* e *conta_f=0*, la procedura, seguendo sempre il criterio della prevalenza, ha assegnato alla variabile *sesso* la modalità *maschio*, ma sul modello dei minori era presente *femmina*, quindi anche in questo caso le due modalità erano parimenti probabili.
Per il secondo e il terzo record, dunque, la considerazione del dato relativo al questionario dei minori ha fatto decadere la situazione di prevalenza in base alla quale la procedura ha eseguito la correzione. Per aiutare l'individuazione del sesso nei casi ambigui come questi sono state considerate altre variabili relative a quesiti presenti nel questionario dei minori potenzialmente discriminanti rispetto la variabile *sesso*, come ad esempio quelle sui giochi preferiti.
Nel caso delle variabili relative agli anni di nascita, la verifica e l'allineamento delle informazioni sono avvenuti in due *step*.
Come primo passo, nel file dati sono stati univocamente individuati i record da attribuire ai minorenni. Sono stati contati i record dei minorenni determinati in base all'età del file di output della procedura e quelli individuati sulla base dell'età indicata nel questionario dei minori. In caso di discordanza per le correzioni si è tenuto presente un indicatore (*compi*) costruito per segnalare se il questionario dei minori fosse stato compilato o meno: se *compi=pieno* si intende che è avvalorata almeno una delle variabili del questionario sui minori, se *compi=vuoto* non è presente nel record nessuna informazione relativa al questionario sui minori (Tavola 4.2).

Tavola 4.2: *Correzioni effettuate sulla variabile età. Indagine Aspetti della vita quotidiana – Anno 2005*

ETA'		COMPILAZIONE DEL QUESTIONARIO SUI MINORI (*compi*)	CASI	CORREZIONE EFFETTUATA
Dopo la procedura	Questionario sui minori			
0-17 anni	mancante	VUOTO	200	età questionario minori = età dopo la procedura
0-17 anni	0-17 anni	VUOTO	20	OK
0-17 anni	18 anni e più	PIENO	3	età questionario minori = età dopo la procedura
18 anni e più	0-17 anni	PIENO	6	età dopo la procedura = età questionario minori
0-17 anni	0-17 anni	PIENO	8.510	OK

Come secondo passo si è verificata in modo puntuale l'uguaglianza tra l'età presente nel file dopo la procedura di correzione e quella del questionario sui minori. I casi non allineati sono stati estratti, analizzati e corretti. Per la correzione si è fatto ricorso alle informazioni del file di input ed è stato costruito un indicatore (*cfr_eta*) che ha contato in quanti casi l'età risultante dopo la procedura di correzione è uguale a quelle calcolate sul file di input.
Nel caso di *cfr_eta=3* c'è perfetto allineamento tra le tre età di input e quella del file di output: in questi casi la procedura di correzione non è intervenuta perché le tre età calcolate a partire dagli anni di nascita dei diversi questionari erano già allineate, di conseguenza l'età del file di output coincide con le tre età del file di input. Per questo i record con *cfr_eta=3* ed età rilevata nel questionario dei minori discordante da quella del file di output sono stati corretti deterministicamente, allineando l'età presente nel questionario sui minori con quella del file di output senza estrarre ed analizzare ulteriormente i record.
Un esempio è riportato nel record_1 della Tavola 4.3, in cui l'individuo risulta avere 10 anni sia in base agli anni di nascita del questionario rosa che in base a quello del verde, la procedura di correzione dell'età non ha avuto bisogno di agire (*cfr_eta=3)* perchè le età calcolate sul file di input erano già allineate, l'età riportata nel questionario sui minori è l'unica che differisce, quindi la correzione è stata effettuata allineando quest'ultima alle altre tre.

Tavola 4.3: *Esempi di record con informazioni discordanti sull'età. Indagine Aspetti della vita quotidiana – Anno 2005*

RECORD	FILE DI INPUT DELLA PROCEDURA			INDICATORE COSTRUITO SUL FILE DI INPUT	FILE DI OUTPUT DELLA PROCEDURA	QUESTIONARIO MINORI
	Scheda generale Mod. rosa (*anasc1*)	Questionario individuale Mod. rosa (*anasc2*)	Questionario autocompilato Mod. verde (*anasc3*)	Cfr_eta	Valore dopo la procedura di correzione	Mod. giallo
Record_1	10	10	10	3	10	15
Record_2	15	15	13	2	15	13
Record_3	15	-	-	1	15	13

Gli esempi successivi mostrano, invece, record in cui *cfr_eta<3*. Per questi casi la procedura di correzione ha agito in base al criterio di prevalenza, ma l'informazione del questionario minori può portare a riconsiderare alcune scelte. Nel record_2, l'età del questionario sui minori è allineata con l'età del questionario autocompilato che la procedura ha scartato e nel record_3, essa è diversa dall'unica età presente. Decade, dunque, il principio della prevalenza. Per questi casi specifici la correzione è stata verificata sulla base di ulteriori approfondimenti utilizzando variabili legate all'età.

4.5 Controllo e correzione delle variabili socio-economiche

Questa fase riguarda la correzione delle altre informazioni strutturali dell'individuo quali il titolo di studio, la condizione professionale, la posizione nella professione e l'attività economica.
Per queste variabili, così come per il sesso e l'età, è prevista oltre alla correzione delle incompatibilità, anche l'imputazione dei dati mancanti e la correzione è stata effettuata utilizzando le procedure di correzione probabilistica utilizzate in Istituto (Scia - Sistema Controllo e Imputazione Automatici).
La correzione delle variabili è avvenuta suddividendo la popolazione in due fasce di età, 0-14 anni e 15 anni e più, poiché i percorsi di compilazione della scheda generale (modello rosa), contenente le informazioni socio-anagrafiche sono differenti per le due categorie di rispondenti. In particolare i quesiti sulla condizione professionale, sulla posizione nella professione e sull'attività economica sono rivolti solo alle persone di 15 anni e più, mentre i bambini e ragazzi fino a 14 anni sono per definizione esclusi dalla popolazione attiva.
Per questo vengono effettuati due passi di correzione:

- nel primo vengono sanate le eventuali incompatibilità esistenti tra età, titolo di studio e frequenza scolastica degli individui con meno di 15 anni;
- nel secondo viene analizzata la popolazione di 15 anni e più, per la quale, insieme alle informazioni relative al titolo di studio ed alla frequenza scolastica, vengono sanate le eventuali incompatibilità relative alla condizione professionale, alla posizione nella professione e all'attività economica.

In termini operativi, sui record in cui sono stati riscontrati errori (incoerenze e valori mancanti), in corrispondenza delle variabili coinvolte sono stati imputati valori tra quelli ammissibili con probabilità condizionata alla distribuzione di frequenza relativa all'intero file di indagine.
Questi passi di correzione probabilistica hanno richiesto sia la definizione delle regole di incompatibilità e dei parametri (variabili di strato, grado di fissità, eccetera) richiesti dal *software* Scia per procedere all'imputazione, sia l'utilizzo di una complessa reportistica, che ha consentito di tenere sotto controllo la distribuzione di frequenza delle variabili coinvolte nel passo di correzione e le variazioni che esse subiscono nelle varie fasi di lavorazione.

4.6 Controllo e correzione delle variabili relative alla composizione familiare

In questa fase vengono controllate e corrette le informazioni relative alla composizione interna della famiglia. La famiglia è analizzata controllando l'esattezza delle variabili: relazione di parentela, stato civile, stato civile nell'anno precedente l'intervista, stato civile precedente al matrimonio e anno di matrimonio. Viene verificata, quindi, la congruenza delle informazioni di ogni componente in rapporto a quelle di tutti gli altri membri della famiglia.
L'obiettivo è quello di ricostruire la complessità delle relazioni intercorrenti tra i vari individui della famiglia non più solo sulla base delle informazioni anagrafiche, ma analizzando a posteriori e complessivamente tutti i membri della famiglia. Questo avviene in tre fasi principali.
In primo luogo la procedura controlla e corregge le informazioni anagrafiche (fase 1) e quindi la relazione di parentela con la persona di riferimento, lo stato civile e l'anno del matrimonio, ma non il sesso e l'età, già corrette e quindi fissate nelle precedenti fasi di correzione.
Partendo dalle informazioni validate nel passo precedente, la procedura ricostruisce per tutti i componenti della famiglia il ruolo che occupano rispetto al nucleo/i eventualmente presente/i attribuendo la tipologia al nucleo (fase 2) [10].
Infine, l'analisi congiunta, all'interno della famiglia, della relazione di parentela (a 17 modalità) di ciascun componente della famiglia con la persona di riferimento, del tipo nucleo, del numero di nucleo e dello stato civile permette alla procedura di individuare la specifica tipologia familiare (fase 3). Le tipologie familiari sono 41 e descrivono la struttura familiare in base alla presenza/assenza di un nucleo, se presenti al numero di nuclei, al tipo di nucleo, al sesso e allo stato civile dei genitori, alla presenza/assenza di figli, alla presenza/assenza di altri membri isolati (ovvero persone non facenti parte del nucleo ma della famiglia).
Sul piano operativo, le fasi sopra descritte non sono rigide e possono reiterarsi fino al raggiungimento di una situazione coerente.
Ad esempio, una volta conclusa la fase di creazione dei nuclei, l'analisi dei report può evidenziare che la procedura automatizzata non ha attribuito un minore al nucleo giusto, quello dei genitori, ma a quello degli zii. Questo inconveniente può avvenire in alcune famiglie che comprendono più nuclei per le quali è necessario rivedere la combinazione di relazioni di parentela (reiterando la fase 1), che mantenendo inalterata i rapporti tra i componenti familiari, facilita l'algoritmo per la costruzione dei nuclei presenti nella famiglia.

5. La correzione delle variabili tematiche

Una volta corrette le variabili relative alle caratteristiche demografiche e socio-economiche degli individui e alla struttura della famiglia (variabili strutturali) si è proceduto all'individuazione e alla correzione degli errori (incoerenze o valori mancanti) presenti sulle variabili tematiche d'indagine. Come già detto solo disponendo di variabili strutturali corrette è possibile procedere al controllo delle altre variabili specifiche di indagine, dal momento che le variabili strutturali individuano i percorsi di compilazione delle varie sezioni dei questionari.
Il processo di correzione ha considerato singolarmente ogni sezione del questionario sui minori. Per ciascuna sezione sono stati implementati i piani di *check* (ovvero procedure Sas di controllo volte ad evidenziare le incoerenze presenti nei dati) che hanno previsto l'utilizzo di macrovariabili (ovvero variabili di appoggio funzionali all'individuazione di errori nei dati), sono stati analizzati gli output risultanti dalle procedure di controllo e, infine, sono stati corretti gli eventuali errori riscontrati attraverso regole di correzione deterministiche del tipo *If-Then*. Gli errori non sanabili attraverso regole di correzione deterministiche sono stati corretti attraverso algoritmi di correzione probabilistica.

[10]I nuclei sono costituiti dai legami di coppia e da quelli genitori/figli e sono di 4 tipi: coppia con figli, coppia senza figli, monogenitore maschio e monogenitore femmina. Affinché un genitore formi un nucleo con il proprio figlio questo ultimo deve essere celibe/nubile. Pertanto formano un nucleo un genitore anziano che vive con un figlio adulto solo se questo ultimo è celibe/nubile. Se un genitore anziano vive con un figlio adulto che, dopo essersi separato, ritorna a vivere nella famiglia di origine, i due non formano un nucleo.

Prima di entrare nel dettaglio del processo delle correzioni deterministiche e probabilistiche verranno trattati brevemente i tipi di errore che si possono riscontrare nei dati e verrà presentata la strategia di individuazione degli errori, che ha previsto la realizzazione di piani di *check* e la costruzione di macrovariabili.

5.1 Tipi di errore

Gli errori presenti nel file dati (intesi in senso lato come incoerenze e come mancate risposte parziali) possono essere distinti in due tipologie principali: gli errori sistematici e gli errori casuali[11].
Si parla di errore sistematico quando si ipotizza l'esistenza di un meccanismo di condizionamento della risposta, scoperto il quale è possibile individuare il valore "corretto" da sostituire a quello errato.
Gli errori sistematici si identificano spesso per una incidenza percentuale elevata, a fronte di una bassa frequenza di errori di altra natura. Questi vengono generalmente corretti con procedure deterministiche, scritte in linguaggio Sas, del genere "*If* (condizione di errore), *Then* (correzione da effettuare)". Quindi se in un record è attivata la condizione di errore ricercata, la regola indica l'azione da effettuare per correggere l'errore.
Rientrano in questa tipologia:

- gli *errori di tipo formale*: ad esempio gli errori di *range*, ovvero la presenza di codici esterni al dominio di ciascuna variabile; gli errori di percorso, ovvero il mancato rispetto dei filtri di domanda o di sezione;
- gli *errori di tipo logico*, in genere rilevabili dall'incrocio di due o più informazioni presenti nel questionario: ad esempio, l'indicazione dello svolgimento di particolari attività non compatibili con le caratteristiche dell'intervistato.

Gli errori casuali, invece, sono dovuti a fattori aleatori e per la loro correzione si adotta un approccio probabilistico. Questo non prevede regole di correzione definite a priori: i record errati vengono sanati col ricorso all'algoritmo utilizzato dai *software* di imputazione. Per le variabili qualitative generalmente il *software* utilizzato è Scia (Sistema Controllo e Imputazione Automatici), la cui strategia di *editing* e correzione è basata sulla metodologia Fellegi-Holt. Per le variabili quantitative si utilizza invece, il *software* Rida (Ricostruzione delle informazioni con Donazione Automatica) che utilizza l'imputazione da donatore con distanza mista minima, rispetto ad alcune variabili di *matching* ritenute determinanti per l'individuazione dei donatori.
Per il questionario sui minori, la strategia di correzione utilizzata ha previsto per prima la correzione deterministica degli errori sistematici. Tutti gli errori non sanabili attraverso un programma deterministico (in particolare le mancate risposte parziali) sono stati, invece, risolti attraverso l'applicazione di metodi probabilistici. Questo perché l'applicazione corretta di un programma probabilistico richiede, preliminarmente, l'individuazione e l'eliminazione di tutti gli errori che si possono correggere con un metodo deterministico. Infatti, la mancata individuazione di errori sistematici e la conseguente "consegna" dei dati errati ad un processo di correzione probabilistica può avere effetti altamente distorcenti rispetto al fenomeno indagato, ed è per questo motivo che i piani di *check* che precedono gli eventuali interventi di tipo deterministico risultano di particolare importanza nel processo complessivo di correzione[12].

5.2 Piani di *check*

L'individuazione degli errori sia sistematici che casuali avviene tramite un articolato piano di controllo dei dati (distribuzioni di frequenze, verifica della coerenza dei percorsi di compilazione, eccetera), anche detto piano di *check*, che si compone di procedure Sas di vario livello di complessità.

---

[11] Istat. *Strategie di correzione del file dati relativo all'indagine "Tempo libero e cultura" Anno 1995*. Roma: Istat, 1998 (Documenti n.2), pag. 7.
[12] Istat. *Il sistema di indagini multiscopo.* Roma: Istat, 2006 (Metodi e Norme n. 31), pag. 202.

Nel processo di correzione del questionario sui minori il piano di *check* predisposto ha avuto l'obiettivo di individuare i seguenti errori sistematici:

1. valore fuori dominio: il valore rilevato è esterno all'intervallo dei valori ammissibili (errori di *range*);
2. valore fuori filtro sezione: rispondono alle domande di una sezione anche individui che per le loro caratteristiche (ad esempio l'età) non devono rispondere (errori di percorso);
3. valore anomalo o *outlier*: il valore rilevato si discosta in modo significativo dai valori assunti nelle restanti unità (tipico delle variabili quantitative);
4. valore mancante (*missing*) o mancata risposta parziale: il valore non è disponibile;
5. incoerenze tra valori relativi a domande appartenenti ad una stessa sezione del questionario: su una unità statistica il valore rilevato per una variabile è in contraddizione con il valore rilevato per almeno un'altra variabile della stessa sezione. Sono generalmente errori di percorso (incompatibilità vicine);
6. incoerenze tra valori relativi a domande appartenenti a sezioni di questionari diversi: su una unità statistica il valore rilevato per una variabile è in contraddizione con il valore rilevato per almeno un'altra variabile presente in un'altro questionario (incompatibilità lontane di domande poste su due diversi questionari);
7. incoerenze tra valori relativi a domande appartenenti a sezioni diverse dello stesso questionario: su una unità statistica il valore rilevato per una variabile è in contraddizione con il valore rilevato per almeno un'altra variabile presente in un'altra sezione dello stesso questionario (incompatibilità lontane di domande poste su due diverse sezioni dello stesso questionario).

Sul piano operativo i piani di *check* sono stati eseguiti in diversi momenti della lavorazione del file dati. In avvio di correzione hanno dato, infatti, indicazioni sulla dimensione degli errori e sulla tipologia degli stessi. In fase di correzione hanno consentito di tenere sotto controllo il progressivo "aggiustamento" del file dati, ovvero la bontà e l'efficacia degli interventi correttivi adottati, evidenziando le incompatibilità residue o le incompatibilità eventualmente generate da un erroneo intervento[13]. Alla fine del processo di correzione hanno consentito di verificare che tutte le incoerenze fossero risolte.

5.3 Macrovariabili

Nella realizzazione dei piani di *check* è stata utilizzata una metodologia di analisi, già sperimentata in altre occasioni di indagine, tanto semplice concettualmente quanto efficace.

Per ciascuna sezione del questionario sui minori sono state costruite nuove variabili chiamate "macrovariabili". Una macrovariabile è una variabile dicotomica (con modalità *ha risposto* e *non ha risposto*), la cui valorizzazione avviene considerando l'unione di tutte le domande sottostanti ad uno stesso filtro di domanda: in pratica quando è presente almeno una risposta nel relativo insieme di quesiti la macrovariabile assume valore *'1 – ha risposto'*, altrimenti assume valore *'0 – non ha risposto'*.

Per ogni sezione del questionario quindi sono state costruite un numero di macrovariabili corrispondente al numero di sottosezioni individuate dai filtri presenti, utilizzando sia i filtri di sezione che quelli di domanda.

Consideriamo, ad esempio, la domanda 2.6 con la quale si chiede a tutti i bambini e ragazzi iscritti alla scuola se hanno compiti da svolgere a casa. Le modalità di risposta previste sono: "No, mai", "Si, alcune volte", "Si, spesso o sempre".

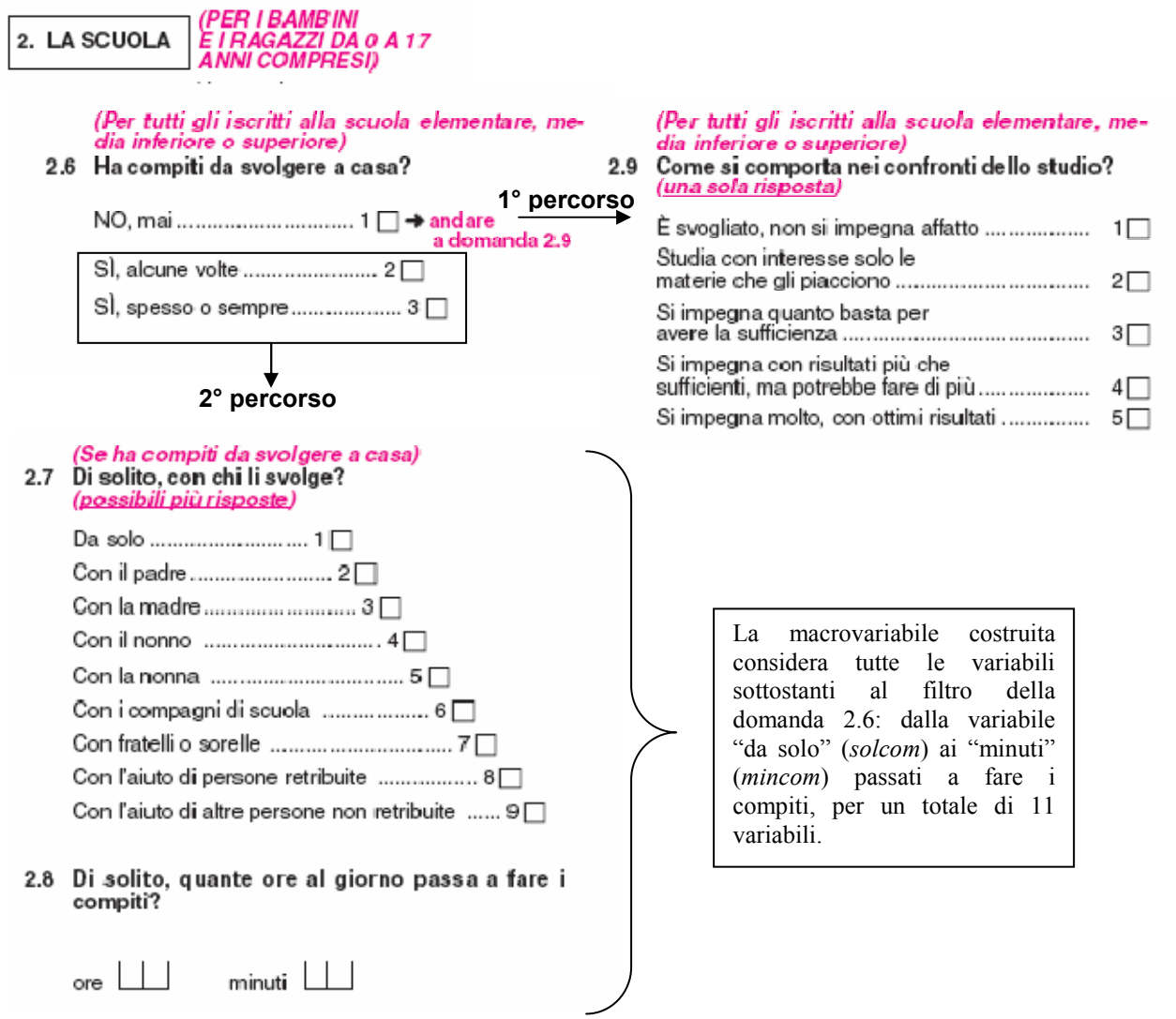
In base alla risposta data alla domanda 2.6 (domanda filtro) si individuano due diversi percorsi (Figura 5.1):

- coloro che non hanno compiti da svolgere a casa "saltano" le domande relative alle persone con cui svolgono i compiti e le ore trascorse a fare i compiti (domande 2.7 e 2.8) e rispondono direttamente alla domanda 2.9 sul comportamento verso lo studio (1° percorso);

[13] Istat. *Il sistema di indagini multiscopo.* Roma: Istat, 2006 (Metodi e Norme n. 31), pag. 204.

- coloro che dichiarano di avere compiti da svolgere a casa devono indicare con chi svolgono i compiti (domanda 2.7) e il numero di ore al giorno passate a fare i compiti (domanda 2.8) (2°percorso).

Figura 5.1: *Questionario sui minori: esempio di costruzione di una macrovariabile. Indagine Aspetti della vita quotidiana – Anno 2005*



A partire dalle variabili associate alle domande 2.7 e 2.8 si costruisce, con semplici istruzioni in Sas, una macrovariabile (Figura 5.2)[14].

Figura 5.2: *Esempio di costruzione di una macrovariabile in linguaggio Sas*

```
* COSTRUZIONE MACROVARIABILE DOMANDE 2.7-2.8 *;
array arr (11)$ solcom--mincom;
macro='0';
do i=1 to 11;
if arr11(i) ne ' ' then macro='1';
end;
```

[14] Si tenga presente che se il quesito prevede una sola risposta c'è una corrispondenza biunivoca tra quesito e variabile, se le risposte possibili sono più di una a ciascuna modalità di risposta corrisponde una variabile.

Se l'intervistato ha dato anche una sola risposta alle domande 2.7 e 2.8, la macrovariabile assume valore *'1 – ha risposto'*, altrimenti avrà valore *'0 – non ha risposto'*. A questo punto in base al piano di *check* implementato si incrocia la variabile corrispondente alla domanda filtro (2.6) con la macrovariabile costruita (cfr §*5.4.5*).

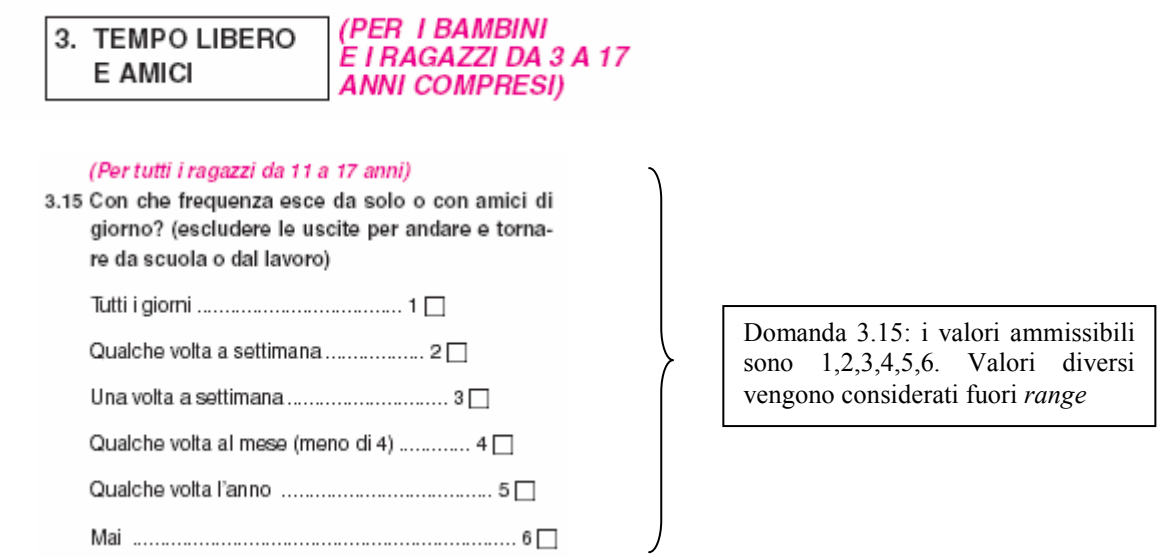
5.4 Correzioni deterministiche

Prima di iniziare il processo di controllo e correzione di ciascuna sezione del questionario, sono stati effettuati controlli e correzioni deterministici preliminari con l'obiettivo di sanare gli errori di *range* e di percorso.

5.4.1 Errori di range

Vengono definiti errori di *range* (coerenza del campo di variazione) quelle situazioni individuate da valorizzazioni della variabile esterne all'intervallo dei valori per essa ammissibili (espliciti o impliciti). L'azione derivante dall'individuazione di questi valori è la loro cancellazione.
Ad esempio, nella sezione 3 se la variabile associata alla domanda 3.15 "Con che frequenza esce da solo o con amici di giorno?", che presenta 6 modalità, ha un valore non compreso tra 1 e 6, tale valore viene cancellato (Figura 5.3).

Figura 5.3: *Questionario sui minori: domanda 3.15. Indagine Aspetti della vita quotidiana – Anno 2005*



Nel caso di variabili quantitative il *range* è implicito, ad esempio per il quesito in cui si è chiesto l'ora di rientro a casa il *range* varia tra '00' e '24' per le ore e tra '00' e '59' per i minuti, per cui i valori esterni a questi sono stati cancellati.
Si è individuata, inoltre, una tipologia particolare di fuori *range* che ha caratterizzato le domande con più modalità di risposta (domande *multiresponse*) ed è stata definita "slineamento". Per questo tipo di domande ad ogni modalità di risposta è associata una variabile. Per slineamento si intende il caso in cui una di queste variabili ha presentato un valore che è fuori il suo *range*, ma è ammissibile per la variabile che la precede o la segue. In questi casi si è ipotizzato che vi fosse stato un errore in fase di registrazione dei dati e quindi si è proceduto alla cancellazione del valore errato e contemporaneamente alla sua attribuzione alla variabile per la quale esso è risultato ammissibile.
Ad esempio se nella domanda 2.7, relativa alle persone con cui i bambini svolgono i compiti (Figura 5.4), la variabile associata alla modalità di risposta "Con il padre", che dovrebbe assumere valore '2',

presenta il valore '1' significativo per la variabile "Da solo" che la precede, si procede alla cancellazione del valore '1' riferito alla variabile "Con il padre" e all'attribuzione dello stesso alla variabile "Da solo".

Figura 5.4: *Questionario sui minori: domanda 2.7. Indagine Aspetti della vita quotidiana – Anno 2005*

Domanda 2.7: ogni modalità di risposta è una variabile. La variabile "Con il padre" può assumere solo il valore '2'. Se assume valore '1' significativo per la variabile "Da solo" che la precede, si procede alla cancellazione del valore '1' riferito alla variabile "Con il padre" e all'attribuzione dello stesso alla variabile "Da solo".

Altre situazioni specifiche hanno riguardato la correzione di variabili quantitative a due o più *byte*, per le quali, anziché alla cancellazione, si è proceduto alla traslazione a destra del valore e, successivamente, all'inserimento di zeri a sinistra. Ad esempio, nell'ipotesi in cui la variabile associata al quesito "Quante ore al giorno passa a fare i compiti" presenti il valore 4 seguito da uno spazio ('4 '), la correzione ha previsto la trasformazione del valore in '04'.

5.4.2 Errori di fuori filtro sezione

Si definiscono errori di fuori filtro sezione le risposte fornite da soggetti con età o caratteristiche non corrispondenti a quelle indicate nei filtri di ciascuna sezione.
Ad esempio, chi ha un'età maggiore di 13 anni non deve rispondere ai questi della sezione 5 relativa al gioco e riservata ai soli bambini di età compresa tra i 3 e i 13 anni. Il problema delle informazioni ridondanti, ovvero non dovute, è molto frequente ed è generalmente un tipo di errore che si commette in fase di intervista a causa di una lettura affrettata e superficiale dei filtri di accesso alle sezioni. Per correggere questo tipo di errore si è proceduto per fasi.

Fase 1: Costruzione degli indicatori di rispetto della condizione di filtro.
Sono state dapprima create delle variabili utilizzate come indicatori di rispetto delle condizioni di filtro: ad esempio per la sezione 3 relativa al tempo libero e agli amici, dove dovevano rispondere solo i bambini dai 3 ai 17 anni, il filtro sezione è stato costruito come riportato nella Figura 5.5.

Figura 5.5: *Esempio di costruzione di un indicatore relativo al filtro sezione in linguaggio Sas*

```
* FASE 1: Costruzione indicatori di "rispetto delle condizioni di filtro" *;
  ETA3_17 = 0
  IF 3<=ETA=<17  THEN  ETA3_17 = 1
```

→ Tale indicatore assume valore "1" nel caso in cui l'intervistato ha un'età compresa tra i 3 e i 17 anni, mentre assume valore "0" in tutti gli altri casi.

Fase 2: Definizione di un nuovo tracciato record.
Il tracciato del file dei dati individua posizione e lunghezza di tutte le variabili associate alle domande presenti nel questionario. In questa fase, per ciascuna sezione del questionario è stato definito un nuovo tracciato nel quale, piuttosto che definire la posizione di ogni singola variabile nel file dati, vengono lette intere stringhe di *byte* riferite alle singole sezioni.
Ad esempio la sezione 3 sul tempo libero e gli amici viene letta nel tracciato come un'unica stringa con il nome *temlibe*. In pratica ogni sezione viene trattata come se fosse un'unica variabile. La variabile *temlibe* è vuota quando l'intervistato non ha dato nessuna risposta alle domande della sezione 3 e valorizzata quando è presente almeno una risposta.

Fase 3: *Check* dei dati.
L'utilizzo di un tracciato specifico ha consentito di incrociare direttamente le singole stringhe di sezione con gli indicatori di rispetto della condizione di filtro per prendere visione degli errori di percorso e successivamente di intervenire con la cancellazione dei campi errati. Ogni sezione letta come un'unica variabile è stata incrociata con il filtro di sezione relativo. Di seguito si riporta l'incrocio tra la nuova variabile *temlibe* relativa alla sezione Tempo libero e amici e l'indicatore di rispetto della condizione di filtro eta3_17 (Figura 5.6).

Figura 5.6: *Esempio di incrocio dei record individuali con il filtro di sezione*

```
* FASE  3: Output della fase 3 - Check dei dati *;
* Esempio di incrocio dei record individuali con il filtro di sezione;
```

temlibe(3.TEMPO LIBERO E AMICI)

	eta3_17		
Frequenza Percentuale	fuori filtro	filtro ok	Totale
VUOTA	1268 14.51	217 2.48	1485 16.99
VALORIZZATA	73 0.84	7181 82.17	7254 83.01
Totale	1341 15.35	7398 84.65	8739 100.00

Tra i minorenni (in totale 8.739) ci sono 73 bambini da 0 a 2 anni (*eta3_17=0*) che hanno erroneamente risposto alle domande della sezione 3. Le informazioni da essi rilasciate sono state successivamente cancellate. Gli individui dai 3 ai 17 anni sono 7.398 e di questi 7.181 hanno risposto alla sezione.

Fase 4: Correzioni deterministiche.
L'uso del nuovo tracciato record costruito nella Fase 2 consente di considerare una sezione come se fosse un'unica variabile, con conseguente guadagno in termini di numero di istruzioni necessarie nella fase di correzione. Ad esempio, per la sezione 3 relativa al tempo libero e agli amici, contenente 61 variabili, si sarebbe dovuto scrivere (usando il tracciato standard) la seguente serie di istruzioni:

```
IF  ETA3_17  NE  1  THEN DO ;
      CONCOR = ' ' ;
      SIPACO = ' '  ;
      ..... = ' '  ;
END;
```

e così via fino alla 61esima variabile.
Utilizzando il nuovo tracciato, con una sola istruzione si è ottenuto lo stesso risultato:

```
IF  ETA3_17  NE  1  THEN TEMLIBE = ' ' ;
```

Quindi, essendo l'intero questionario costituito da 8 sezioni, sono bastate 8 istruzioni per eseguire la cancellazione delle risposte fuori filtro di tutte le sezioni.

```
IF  ETA0_17  NE  1  THEN NONNI   = ' ' ;
IF  ETA0_17  NE  1  THEN SCUOLA  = ' ' ;
IF  ETA3_17  NE  1  THEN TEMLIBE = ' ' ;
IF  ETA3_17  NE  1  THEN TELEV   = ' ' ;
IF  ETA3_13  NE  1  THEN GIOCO   = ' ' ;
IF  ETA6_17  NE  1  THEN TELECE  = ' ' ;
IF  ETA6_17  NE  1  THEN CHIAVI  = ' ' ;
IF  ETA6_17  NE  1  THEN AIUTO   = ' ' ;
```

Tavola 5.1: *Questionario sui minori: numero record con campi errati per ciascuna sezione. Indagine Aspetti della vita quotidiana – Anno 2005*

Sezione	Filtro sezione	Indicatore	Variabile sezione	Record errati
1. Nonni e affidamento del bambino	0-17 anni	eta0_17	nonni	0
2. La scuola	0-17 anni	eta0_17	scuola	0
3. Tempo libero e amici	3-17 anni	eta3_17	temlibe	73
4. La televisione	3-17 anni	eta3_17	telev	60
5. Il gioco	3-13 anni	eta3_13	gioco	232
6. Telefono cellulare	6-17 anni	eta6_17	telece	49
7. Chiavi di casa e autonomia	6-17 anni	eta6_17	chiavi	36
8. Lavoretti in casa e aiuto familiari	6-17 anni	eta6_17	aiuto	21

La fase delle correzioni deterministiche preliminari ha consentito di correggere in un solo *step* gli errori di accesso a tutte le sezioni del questionario, lasciando alla fase successiva la correzione di eventuali errori interni a ciascuna sezione.

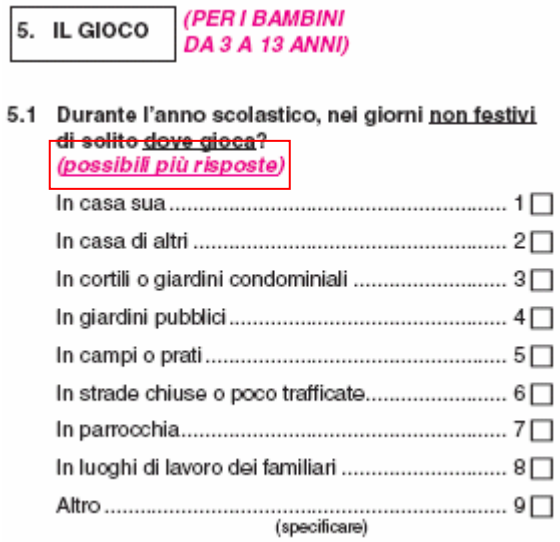
5.4.3 L'analisi delle mancate risposte parziali

Prima di iniziare le correzioni deterministiche su ogni singola sezione del questionario sui minori è stata effettuata una prima fase di controllo delle variabili tematiche attraverso la realizzazione di un'ampia reportistica di distribuzioni di frequenza riferita, per ciascuna variabile, solo alla specifica sottopopolazione di potenziali rispondenti (frequenze filtrate) allo scopo di evidenziare la quota di mancate risposte parziali (*missing*) di ciascuna variabile presente nel questionario. Infatti una quota di *missing* superiore alla soglia di volta in volta considerata critica è un segnale di mal funzionamento della domanda, dovuto, ad esempio, alla sua formulazione, alla sua posizione nella sequenza delle domande o all'organizzazione dei filtri che guidano la compilazione del questionario[15].
Le distribuzioni di frequenza filtrate, però, consentono di valutare la quota di *missing* solo per le domande che prevedono una sola risposta, nel caso di domande *multiresponse* invece è necessario costruire una macrovariabile per valutare il livello complessivo di *missing* della domanda.
Ad esempio con la domanda 5.1 si chiede ai bambini da 3 a 13 anni in quali luoghi di solito giocano durante l'anno scolastico nei giorni non festivi e gli viene proposto un elenco di 9 modalità di risposta e la possibilità di indicarne più di una (Figura 5.7).
La frequenza di ogni singola modalità di risposta indica quanti intervistati hanno indicato quel luogo di gioco, mentre la quota di mancata risposta di ciascuna modalità singolarmente considerata dà una duplice informazione non scindibile: è, infatti, la somma dei bambini che non giocano in quel luogo e di quelli che non hanno risposto (Tavola 5.2). Per valutare la quota "reale" dei *missing* è necessario costruire una macrovariabile che consenta di valutare la mancata risposta complessiva. Questa quota, infatti, è data da quei bambini che non hanno indicato nessuna delle nove modalità di risposta previste.

[15] In generale le domande con un'elevata percentuale di mancata risposta parziale non entrano nel processo di correzione.

La macrovariabile è costruita in modo tale da assumere valore '*1 - ha risposto*' se l'intervistato ha dato almeno una risposta alla domanda 5.1 e valore '*0 - non ha risposto*' nel caso in cui l'intervistato non ha dato nessuna risposta. La frequenza dei valori '*0 - non ha risposto*' fornisce la quota reale dei *missing*. Nel caso della domanda 5.1 la quota reale dei *missing* (macrovariabile= 0) è risultata pari al 3,3%.

Figura 5.7: *Questionario sui minori: domanda 5.1. Indagine Aspetti della vita quotidiana – Anno 2005*


5. IL GIOCO (PER I BAMBINI DA 3 A 13 ANNI)

5.1 Durante l'anno scolastico, nei giorni non festivi di solito dove gioca? (possibili più risposte)

In casa sua 1
In casa di altri 2
In cortili o giardini condominiali 3
In giardini pubblici 4
In campi o prati 5
In strade chiuse o poco trafficate 6
In parrocchia 7
In luoghi di lavoro dei familiari 8
Altro 9
(specificare)

Tavola 5.2: *Bambini di 3-13 anni per luogo dove giocano nei giorni non festivi. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 bambini di 3-13 anni)*

LUOGO DI GIOCO NEI GIORNI NON FESTIVI	Si	No, Non indicato	Totale
In casa sua	90,1	9,9	100,0
In casa di altri	37,6	62,4	100,0
In cortili e giardini condominiale	25,9	74,1	100,0
In giardini pubblici	22,8	77,2	100,0
In campi o prati	13,1	86,9	100,0
In strade chiuse o poco trafficate	9,9	90,1	100,0
In parrocchia	10,7	89,3	100,0
In luoghi di lavoro dei familiari	2,6	97,4	100,0
Altro	2,6	97,4	100,0

Pertanto sono state costruite tante macrovariabili quante sono le domande *multiresponse* presenti nel questionario e per ognuna di esse la quota di *missing* è stata valutata attraverso la frequenza della macrovariabile stessa.
Di seguito si fornisce un esempio della procedura utilizzata, scritta in linguaggio Sas (Figura 5.8).

Figura 5.8: *Questionario sui minori: esempio di creazione di una macrovariabile in linguaggio Sas*

```
* 1 - COSTRUZIONE MACROVARIABILE DOMANDA 5.1 *;
   array arr23a (9)$ ovegio1--ovegio9;
   macro23a='0';
   do i=1 to 9;
   if arr23a(i) ne ' ' then macro23a='1';
   end;

 * 2 - CHECK: FREQUENZA MACROVARIABILE DOMANDA 5.1 *;
   proc freq;
   where '003'<=eta<='013';     *FILTRO SEZIONE;
   table macro23a/ missing; run;
```

5.4.3 Il trattamento della modalità "altro specificare"

Nel questionario dei minori sono presenti 18 domande semiaperte nelle quali è prevista la modalità di risposta altro ed accanto uno spazio per specificare cosa si intende per altro (Figura 5.9). L'analisi delle modalità di risposta "altro specificare" è stata effettuata attraverso lo studio delle distribuzioni di frequenza filtrate, ovvero distribuzioni di frequenza riferite, per ciascuna variabile, solo alla specifica sottopopolazione di potenziali rispondenti. Quando si è ritenuto che la distribuzione di frequenza della modalità "altro" fosse troppo elevata, si è proceduto ad una analisi più accurata del contenuto dell'"altro specificare", che ha fatto emergere due situazioni rilevanti:

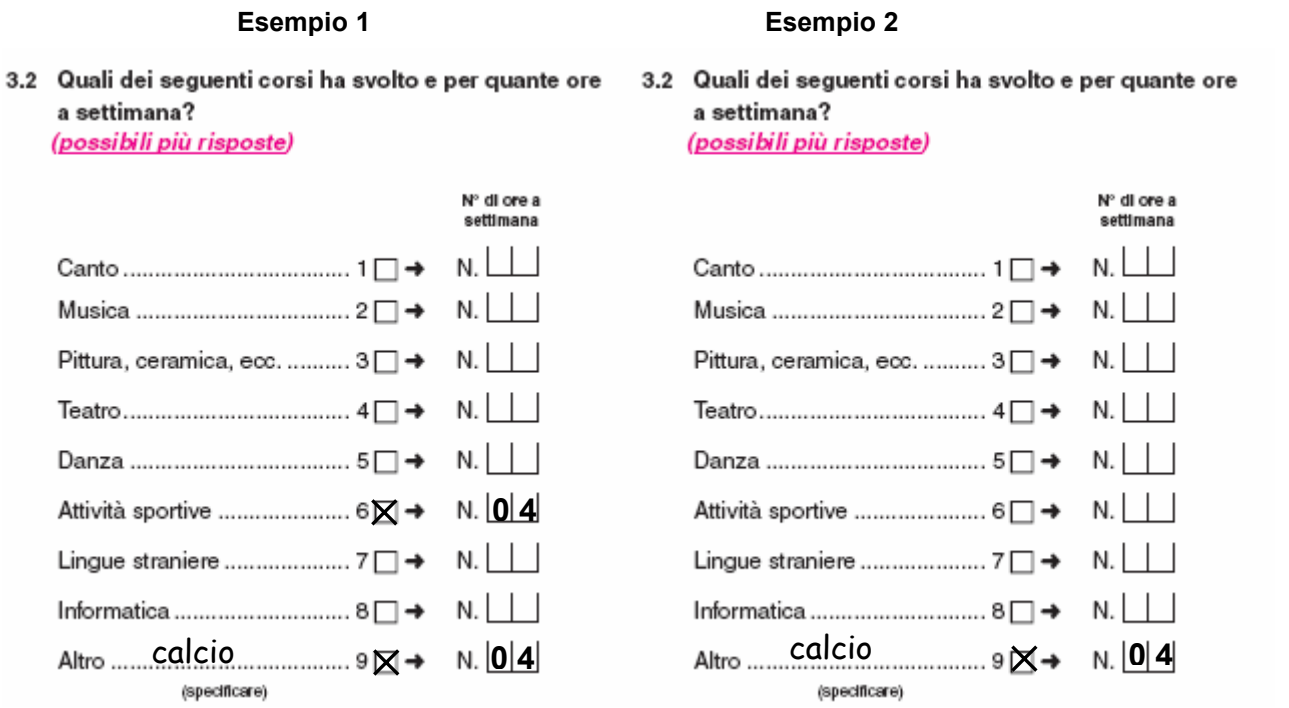
1. l'esplicitazione nella voce "altro specificare" di modalità di risposta già presenti nell'elenco fornito;
2. la presenza nell'"altro specificare" di nuove modalità di risposta per una quota non trascurabile di casi.

Con riferimento alla prima situazione questa può prefigurare sia una ridondanza di informazione, quando la stessa risposta viene ripetuta due volte per eccesso di dettaglio, sia una "distrazione" dell'intervistatore che non ha correttamente individuato la risposta nell'elenco fornito. In entrambi i casi la correzione ha previsto la cancellazione della modalità altro, ma nel secondo (Figura 5.9, esempio 2) anche l'imputazione del valore alla variabile corrispondente alla modalità di risposta già presente nell'elenco.

Un esempio è dato dalla domanda 3.2 dove ai bambini e ragazzi da 3 a 17 anni sono stati chiesti i corsi svolti. Nel caso in cui un ragazzo avesse svolto corsi non compresi tra le modalità di risposta, l'intervistatore doveva barrare la voce altro, specificando il tipo di corso svolto nella casella per la risposta aperta.

Così come riportato nell'esempio (Figura 5.9), l'errore più frequente è risultato l'indicazione di tipi di sport come calcio, nuoto, basket e così via nonostante fosse presente, nell'elenco dei corsi, la modalità "attività sportive". Si tratta, probabilmente, di un errore dovuto ad un eccesso di precisione da parte dell'intervistato e/o dell'intervistatore che, trovando tra le modalità di risposta il nome generico "attività sportive", ha preferito specificare il nome dello sport praticato.

Figura 5.9: *Questionario sui minori: esempi di compilazione dell' "altro specificare". Indagine Aspetti della vita quotidiana – Anno 2005*



Nella Figura 5.10 si riportano le istruzioni Sas attraverso le quali è stata corretta la situazione dell'esempio 2. Preventivamente sulle risposte fornite è stata effettuata una analisi testuale che ha consentito di individuare parti di stringhe comuni a più risposte che sono state poi utilizzate nel programma Sas per ricodificare l'altro specificare.
Così, ad esempio, per ricodificare tutte le risposte in cui è stato specificato il termine 'calcio' è stata utilizzata la stringa '*cal*' (Figura 5.10).

Figura 5.10: *Esempio di ricodifica dell' "altro specificare" in linguaggio Sas*

```
* REGOLA DI CORREZIONE DETERMINISTICA: DOM 3.2 - RICODIFICA ALTRO SPECIFICARE;
  str=upcase(alcosp);
  a=index(str,'BASK') or index(str,'CAL') or index(str,'SPORT') or
  index(str,'EQUI') or index(str,'GIN') or index(str,'J')
  or index(str,'KA') or index(str,'NUO') or index(str,'PAL')
  or index(str,'PAT') or index(str,'PISC') or index(str,'RUG')
  or index(str,'SCI') or index(str,'TEN');

* QUANDO NELLA VARIABILE "ALTRO SPECIFICARE" E' STATO INDICATO IL NOME DI UNO
  SPORT, ASSEGNO IL VALORE 6 E LE RELATIVE ORE ALLA VARIABILE "ATTIVITA'
  SPORTIVE" (sport, nsport ) E CANCELLO LA MODALITA' ALTRO TIPO DI CORSO (alco)
  E LE RELATIVE ORE (nalco);

  if a>0 then do;
  sport='6';
  nsport=nalco;
  alco=' ';
  nalco=' ';
  end;
```

Rispetto invece alla seconda situazione, un esempio è dato dalla domanda 3.7 relativa al luogo delle feste. La distribuzione di frequenza filtrata ha evidenziato una percentuale molto alta della modalità di risposta "in altro luogo" (12,1%) (Figura 5.11).
Tale risultato ha suggerito una analisi dettagliata della voce "altro specificare", da cui è emerso che ci sono altri due luoghi dove è abitudine organizzare feste, oltre a quelli riportati tra le modalità di risposta della domanda 3.7, che hanno una ricorrenza significativa: la pizzeria e la parrocchia.

Figura 5.11: *Questionario sui minori: domande 3.6 e 3.7. Indagine Aspetti della vita quotidiana – Anno 2005*

Il passo successivo è stato quello di scrivere una procedura Sas che ha ricodificato queste risposte in due nuove modalità di risposta: Pizzeria/ristorante/birreria/pub (*festel7*=7) e Oratorio/Parrocchia (*festel8*=8). In questo modo la percentuale dell'altro specificare si è notevolmente ridotta arrivando al 3,9% (Figura 5.12).

Figura 5.12: *Questionario sui minori: esempio di creazione di una nuova variabile, a partire dall'altro specificare, in linguaggio Sas*

```
* DOM 3.7 - CREAZIONE DI UNA NUOVA VARIABILE A PARTIRE DALL'"ALTRO SPECIFICARE"

  str=upcase(festesp);
  c=index(str,'PIZ')   or index(str,'CINE')or index(str,'RIST')
  or index(str,'PLAY') or index(str,'PUB') or index(str,'BIRR')
  or index(str,'BAR') or index(str,'DISC');
  if c>0 then festel7='7';

  str=upcase(festesp);
  d=index(str,'PARR')      or index(str, 'CHIE')      or index(str,'CONV')
  or index(str,'ORATORIO') or index(str,'ORATORIALE') or index(str,'ACR')
  or index(str,'GIOVANI');
  if d>0 then festel8='8';
```

Nell'esempio appena effettuato, una semplice distribuzione di frequenza è stata sufficiente a mettere in rilievo la mancanza di una esaustiva gamma di modalità di risposta per l'intera popolazione, tuttavia a volte tale problematica può interessare anche solo una popolazione specifica, ad esempio soltanto le femmine o solo gli individui in una specifica classe di età.
In questi casi la necessità di una ricodifica dell' "altro specificare" emerge soltanto con una più attenta analisi, mediante tavole per sesso e classe di età o per zona. E' stato il caso della domanda 7.4 con la quale è stato chiesto ai bambini e i ragazzi da 6 a 17 anni, che ricevono regolarmente una somma di denaro dai genitori, come spendono questa somma. Le modalità di risposta sono 21 compresa la modalità "altro specificare". La lettura della distribuzione di frequenza delle modalità di spesa della paghetta ha messo in rilievo una percentuale pari al 5,9% della modalità "altro specificare", che è stata giudicata accettabile. Successivamente però la distribuzione per sesso ed età della domanda ha evidenziato come la percentuale di "altro specificare" salisse al 10,1% tra i bambini di 6-10 anni. Per questo si è ritenuto necessario analizzare, sempre attraverso distribuzioni di frequenza filtrate, cosa i bambini di 6-10 anni avessero indicato nel campo aperto "altro specificare". Molti dei bambini di questa classe di età hanno scritto nel campo aperto "non spende paghetta".
A questo punto è stata creata una nuova modalità di risposta *denge22*="Non la spende". In questo modo la modalità di risposta "altro" è passata dal 5,9 al 3,6 sul totale dei bambini di 6-17 anni e dal 10,1% al 5,8% per i bambini di 6-10 anni.

5.4.5 Incompatibilità tra domande di una stessa sezione del questionario

Effettuate le correzioni preliminari su tutto il questionario dell'infanzia (fuori *range*, fuori filtro sezione, altro specificare) sono stati effettuati i *check* e le correzioni deterministiche su ogni singola sezione. Per ciascuna sezione sono state costruite le macrovariabili e implementati i relativi *check*.
Un valido esempio di come la strategia di *check* utilizzata consente con un numero molto ridotto di incroci di tenere sotto controllo i diversi percorsi di compilazione interni a ciascuna sezione è dato dalla sezione 2 relativa alla scuola, dove con la domanda 2.4 si chiede a tutti i bambini e ragazzi iscritti alla scuola elementare, media inferiore o superiore se nel corso dell'anno scolastico hanno partecipato a corsi di musica, sport, lingue, informatica, ecc. organizzati dalla scuola al di fuori dell'orario scolastico. Le modalità di risposta previste sono: "No", "Si" (Figura 5.13).

I bambini che dichiarano di non seguire corsi saltano la domanda 2.5 relativa al tipo di corsi effettuati e al numero delle ore settimanali e rispondono direttamente alla domanda 2.6 relativa ai compiti (1° percorso); i bambini e ragazzi invece che dichiarano di seguire corsi devono rispondere alla domanda 2.5, indicando il tipo di corsi effettuati e il numero delle ore settimanali (2° percorso).
La macrovariabile costruita considera tutte le risposte date alla domande 2.5. Se l'intervistato ha dato anche una sola risposta la macrovariabile assume valore '*1 – ha risposto*', altrimenti avrà valore '*0 – non ha risposto*' (Fase 1). A questo punto si incrocia la variabile che si riferisce alla domanda filtro (2.4) con la macrovariabile costruita (Fase 2).
Questo sistema consente di tenere sotto controllo tutte le risposte date alla domanda 2.5 utilizzando un solo *check* relativo alla valorizzazione della macrovariabile anzichè di ogni singola variabile che la compone.
L'esempio di programma scritto in linguaggio Sas (Figura 5.14) mette in evidenza la semplicità e la sinteticità di ogni singolo *check*, ma anche la praticità e l'efficacia che questi possono avere in un processo di correzione.

Figura 5.13: *Questionario sui minori: esempio di costruzione di una macrovariabile. Indagine Aspetti della vita quotidiana – Anno 2005*

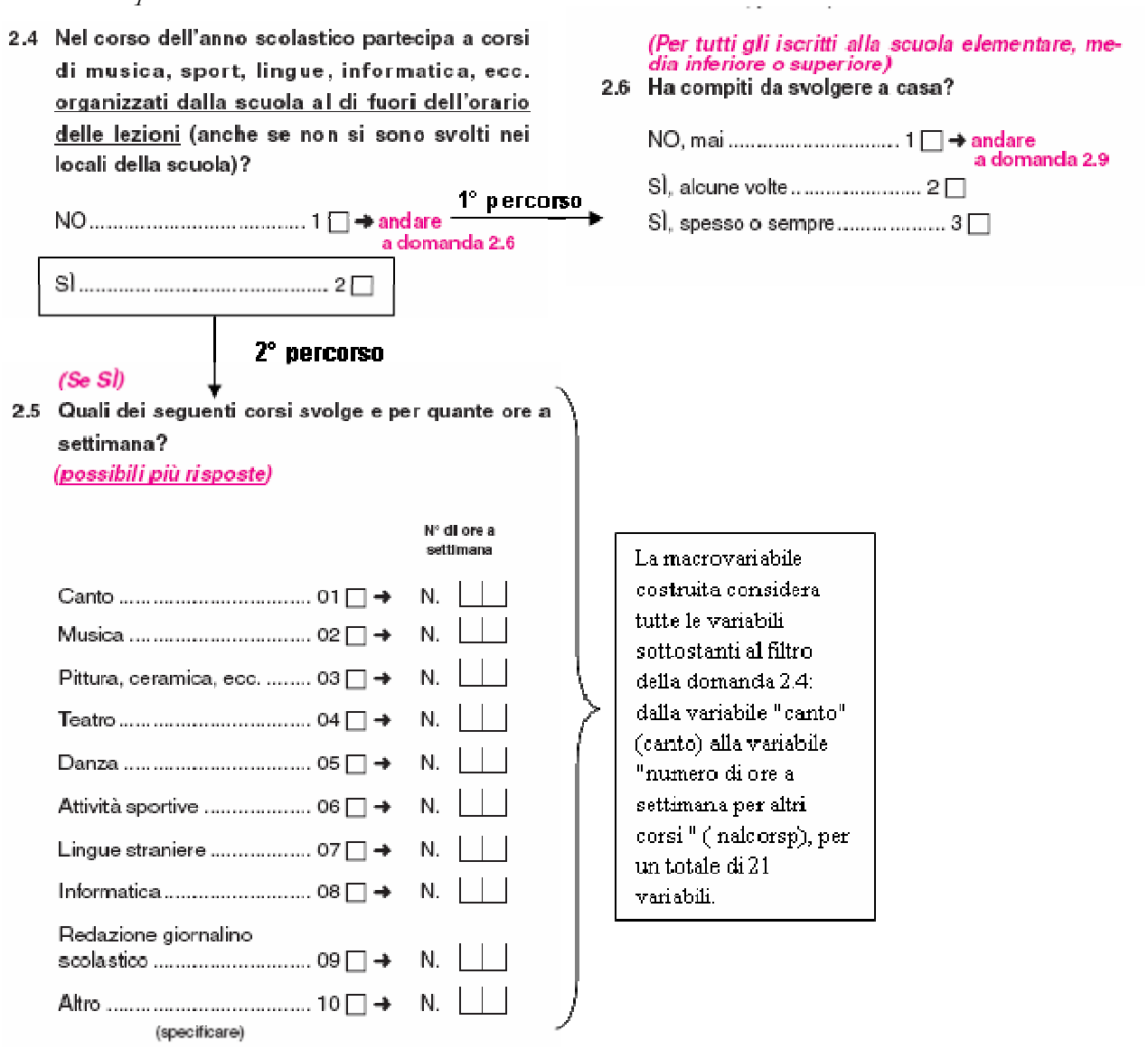


Figura 5.14: *Questionario sui minori: esempio di creazione di una macrovariabile in linguaggio Sas*

```
* FASE 1 - COSTRUZIONE MACROVARIABILE DOMANDE 2.4-2.5 *;
  array arr (21)$ canto--nalcor;
  macro='0';
  do i=1 to 21;
  if arr(i) ne ' ' then macro='1';
  end;

* FASE 2 – CHECK: INCROCIO TRA VARIABILE FILTRO (corsiex) E MACROVARIABILE *;
  proc freq;
  where eta<'018' and '05'<=frsc<='13';     *FILTRO SEZIONE;
  table corsiex*macro/missing;
  run;
```

Analizzando ogni singola cella dell'incrocio tra le variabili associate alle domande filtro e le macrovariabili (Fase 3) si sono progressivamente individuate le situazioni coerenti con la struttura del questionario, quelle che indicavano chiaramente la presenza di un errore sistematico e quelle che dovevano essere passate alle fasi di correzione probabilistica.
Il primo passo di *check* relativo all'incrocio tra variabile filtro e macrovariabile non fornisce in prima battuta tutte le informazioni necessarie alla strategia di correzione, ma rappresenta il punto di partenza di analisi sempre più mirate. L'incrocio tra variabile filtro e macrovariabile, infatti, consente solo di individuare le celle con errore sistematico, ovvero le celle critiche. Il passo successivo è quello di isolare ogni singola cella critica e di stampare tutti i casi che vi appartengono andando a vedere nel dettaglio cosa gli intervistati hanno risposto ad ogni singola domanda che compone la macrovariabile (Figura 5.15).

Figura 5.15: *Esempio di incrocio tra variabile filtro (domanda 2.4) e macrovariabile*

```
* OUTPUT DELLA FASE 2 – CHECK: INCROCIO TRA VARIABILE FILTRO E MACROVARIABILE *

Nell'anno scolastico partecipa a corsi (CORSIEX) per macrovariabile domanda 2.5

CORSIEX              macrovariabile
```

Frequenza Percentuale Pct riga Pct col	0	1	Totale
	291	8	299
	4.98	0.14	5.11
	97.32	2.68	
	6.68	0.54	
1	4049	48	4097
	69.24	0.82	70.06
	98.83	1.17	
	92.91	3.22	
2	18	1434	1452
	0.31	24.52	24.83
	1.24	98.76	
	0.41	96.24	
Totale	4358	1490	5848
	74.52	25.48	100.00

1. 8 bambini non rispondono alla domanda sui corsi seguiti a scuola (*corsiex=non indicato*), ma poi indicano il tipo di corso seguito e/o il numero di ore a settimana. La macrovariabile infatti assume valore '1' – *Ha risposto*.

2. 48 bambini hanno indicato che non seguono corsi a scuola (*corsiex=1),* ma poi rispondono alla domanda sul tipo di corso seguito e/o il numero di ore a settimana. La macrovariabile infatti assume valore '1' – *Ha risposto*.

3. 18 bambini dichiarano che seguono corsi a scuola (*corsiex=2),* ma poi non rispondono alla domanda sul tipo di corso seguito e/o il numero di ore a settimana.. La macrovariabile infatti assume valore '0' – *Non ha risposto*.

Consideriamo, ad esempio, gli otto bambini che non hanno risposto alla domanda sui corsi seguiti a scuola, ma poi hanno la macrovariabile valorizzata. Analizzando nel dettaglio questi otto record errati emerge come si tratti di bambini che nella quasi totalità dei casi hanno sempre indicato di praticare sport e il numero di ore a settimana, oltre ad altri corsi (Figura 5.16). Si tratta quindi di un errore sistematico che è possibile correggere deterministicamente attribuendo il valore '2' ("Si") alla variabile *corsiex*.

Figura 5.16: *Esempio di stampa dei record errati*

```
* FASE 3 - CHECK: STAMPA DEI RECORD ERRATI *;
  proc freq;
  where eta<'018' and '05'<=frsc<='13'
  and corsiex=' ' and macro='1';         * FILTRO SEZIONE;
  table canto--nalcor /norow nocol missing list;
  run;


* OUTPUT FASE 3 - CHECK: STAMPA DEI RECORD ERRATI *;
 -·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-
                           M                   N       N   I
       N   M   N       N   T   N       N   S   S   L   L   N   N       N       N
   C   C   U   M   P   P   E   T   D   D   P   P   I   I   F   I   G   G   A   A
   A   A   S   U   I   I   A   E   A   A   O   O   N   N   O   N   I   I   L   L
   N   N   I   S   T   T   T   A   N   N   R   R   G   G   R   F   O   O   C   C
   T   T   C   I   T   T   R   T   Z   Z   T   T   S   S   M   O   R   R   O   O
   O   O   A   C   U   U   O   R   A   A   I   I   T   T   A   R   N   N   R   R Freq.  Freq.
                                                                                         cum.
 --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -----  -----
                                           6   2                                     1      1
                                           6  10           8   2                     1      2
           2   1                                   7   2   8   1                     1      3
           2   2   3   2                   6   2   7   5   8   2                     1      4
                                           6   8                                     1      5
                                           6   3                                     1      6
                                           6   4                                     1      7
                                           6   1   7   2                  10   2     1      8
 -·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-·-
```

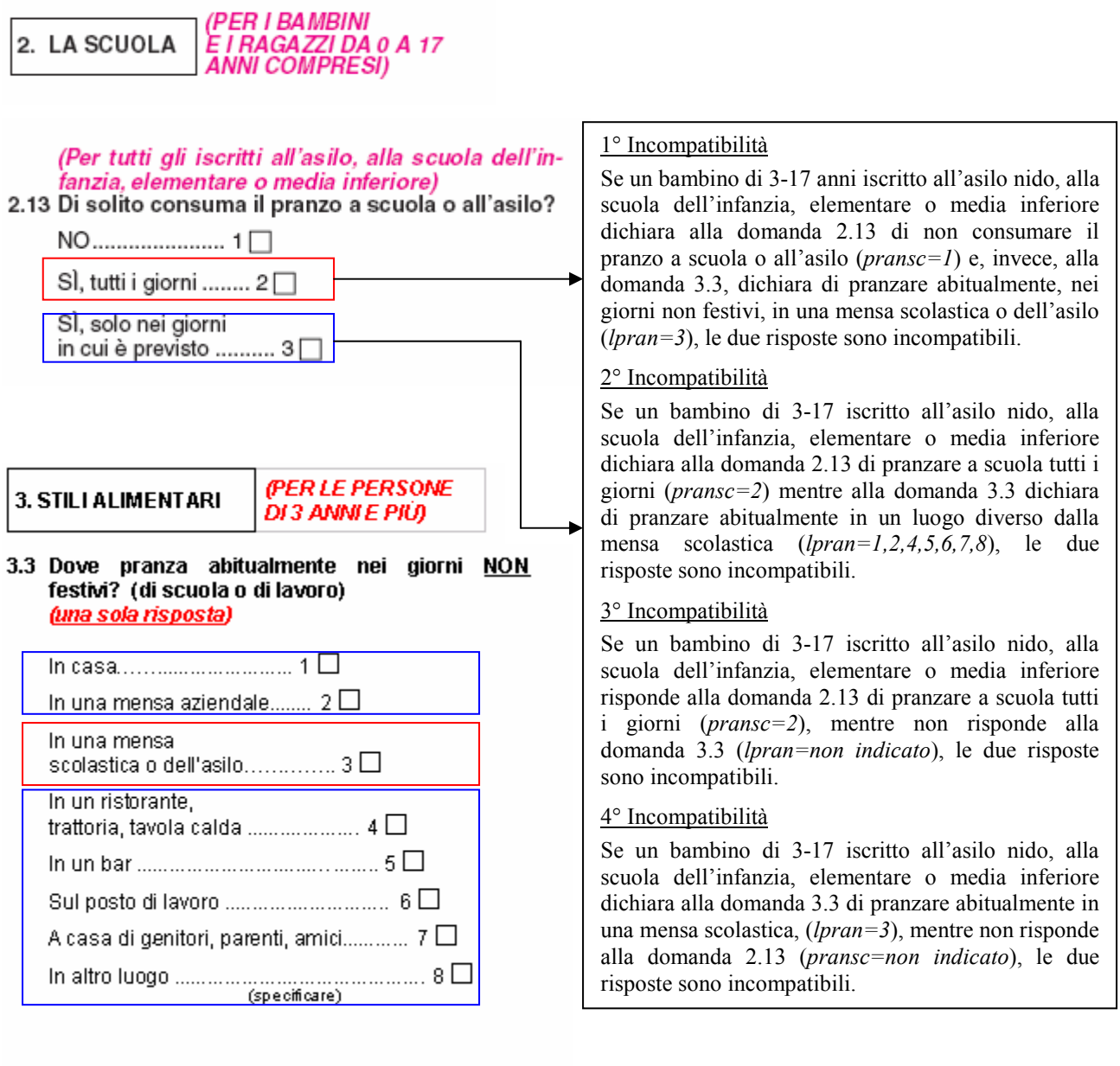
Le due fasi di *check*: incrocio tra variabile filtro e macrovariabile (fase 2) e stampa dei record errati (fase 3) devono essere sequenziali, poiché è molto difficile prevedere a priori la stampa di tutte le combinazioni errate. Le due fasi quindi si alternano continuamente in un ciclo di approfondimenti successivi che permettono di evidenziare via via le celle critiche da stampare.

5.4.6 Incompatibilità tra domande di questionari diversi

Le relazioni tra le informazioni sottoposte a controllo hanno riguardato non solo quelle dello specifico questionario sui minori, ma anche quelle presenti sugli altri questionari di indagine. Sono state considerate cioè, tutte le variabili del questionario dei minori in relazione logica con variabili presenti negli altri questionari utilizzati nell'indagine. A tale scopo è stato implementato uno specifico piano di *check* volto a verificare l'esistenza o meno di incoerenze nelle risposte fornite; una volta, evidenziati gli errori, sono state decise le regole deterministiche di correzione.
Sono state individuate 38 possibili situazioni di incompatibilità logica tra informazioni del questionario minori e quelle degli altri questionari, di queste 10 con il questionario familiare, 9 con quello individuale e 19 con l'autocompilato.

Per illustrare come le incompatibilità logiche sono state analizzate, consideriamo come esempio l'incompatibilità individuata tra una domanda presente nel questionario sui minori (modello giallo) e una dell'autocompilato (modello verde). Per il modello dei minori è stata considerata la domanda 2.13 della sezione 2 (*pransc*), con la quale si chiede ai bambini e ragazzi iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore, se consumano il pranzo a scuola o all'asilo e nel caso lo consumino, con quale frequenza lo fanno, se tutti i giorni o solo nei giorni in cui è previsto. Le risposte date a questa domanda potrebbero essere logicamente incompatibili con quelle fornite alla domanda 3.3 presente sul questionario autocompilato. La domanda 3.3, infatti, è rivolta a tutte le persone di 3 anni e più e con essa si chiede il luogo abituale del pranzo (*lpran*). Le modalità di risposta sono: nella propria casa, in una mensa aziendale, in una mensa scolastica, in un ristorante, in un bar, a casa di genitori o parenti.

Figura 5.17: *Questionario sui minori: esempio di incompatibilità tra la domanda 2.13 e la domanda 3.3 del questionario autocompilato. Indagine Aspetti della vita quotidiana – Anno 2005*



La domanda 2.13 inserita nel modello dei minori, quindi, è parzialmente sovrapposta con quella presente sul questionario autocompilato. In questa edizione di indagine, nonostante la parziale ridondanza di informazione rilevata, è stato deciso comunque di inserire nei rispettivi modelli entrambe le domande. Questo perché la domanda 2.13 garantisce la confrontabilità con la stessa informazione

rilevata nell'indagine Famiglia e soggetti sociali effettuata nel 1998[16] ed, inoltre, attraverso di essa viene rilevata la frequenza con cui i bambini mangiano a scuola, il concetto di abitudine quindi è indotto dalle modalità di risposta e non è esplicitato nel testo della domanda.
D'altra parte, la domanda 3.3 è presente nel questionario autocompilato dell'indagine Aspetti della vita quotidiana fin dal 1993 e viene utilizzata per realizzare la serie storica sul luogo abituale del pranzo di tutta la popolazione di 3 anni e più e con questa non si chiede la frequenza ma si fa riferimento ad un comportamento abituale.
Per analizzare l'incompatibiltà tra i due quesiti si è analizzata solo la popolazione che risponde ad entrambi. Pertanto poiché la domanda 2.13 è rivolta a tutti i ragazzi da 0 a 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore, mentre la domanda 3.3 è rivolta a tutta la popolazione di 3 anni e più, le incompatibilità rilevabili riguardano solo la sottopopolazione dei ragazzi tra 3 e 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore (5.476 casi).
Le possibili situazioni di incompatibilità logica tra queste due domande, che si possono risolvere deterministicamente, si hanno quando (Figura 5.17):

1. un bambino dichiara alla domanda 2.13 di non consumare il pranzo a scuola, mentre invece alla domanda 3.3 del questionario autocompilato, dichiara di pranzare abitualmente, nei giorni non festivi, in una mensa scolastica;
2. un bambino dichiara alla domanda 2.13 di pranzare a scuola tutti i giorni, mentre alla domanda 3.3 del questionario autocompilato, dichiara di pranzare abitualmente in un luogo diverso dalla mensa scolastica;
3. un bambino dichiara alla domanda 2.13 di pranzare a scuola tutti i giorni, mentre non risponde alla domanda 3.3 del questionario autocompilato;
4. un bambino dichiara alla domanda 3.3 del questionario autocompilato di pranzare abitualmente in una mensa scolastica, mentre non risponde alla domanda 2.13.

Si è effettuato quindi un *check* in cui si sono incrociate le due variabili considerando tutti i bambini e ragazzi da 3 a 17 anni iscritti all'asilo, alla scuola dell'infanzia, elementare o media inferiore (5.476).

Figura 5.18: *Questionario sui minori: esempio di check tra la domanda 2.13 e la domanda 3.3 del questionario autocompilato*

```
* CHECK DOMANDA 2.13 *;
  proc freq ;
  where '003'<=eta<='017' and frsc in ('12','13','14','15'); *FILTRO SEZIONE;
  format pransc $pransc. lpran $lpran.;
  table lpran*pransc/missing;
  run;
```

Il *check* evidenzia la presenza di casi per ognuna dei 4 tipi di incompatibilità possibili:

1. sono 35 i record errati riscontrati attraverso la 1° incompatibilità e rappresentano lo 0,6% dei bambini tra i 3 e i 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore;
2. sono 282 (269+9+4) i record errati riscontrati attraverso la 2° incompatibilità e rappresentano il 5,1% dei bambini tra i 3 e i 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore;
3. sono 138 i record errati riscontrati attraverso la 3° incompatibilità e rappresentano il 2,5% dei bambini tra i 3 e i 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore;
4. sono 55 i record errati riscontrati attraverso la 4° incompatibilità e rappresentano l'1% dei bambini tra i 3 e i 17 anni iscritti all'asilo nido, alla scuola dell'infanzia, elementare o media inferiore).

[16] Istat. *La vita quotidiana di bambini e ragazzi*, Roma: Istat, 2000 (Informazioni n. 23).

Non è stata considerata incompatibilità logica, invece, la situazione in cui i bambini hanno indicato alla domanda 2.13 del questionario dei minori che consumano il pranzo a scuola solo nei giorni in cui è previsto, mentre alla domanda 3.3 del questionario autocompilato hanno indicato che pranzano in un luogo diverso da una mensa scolastica (694 casi in totale). Il motivo di tale scelta è dovuto al fatto che nella domanda 3.3 del questionario autocompilato si chiede il consumo abituale del pranzo, quindi la risposta ha una componente soggettiva: è possibile infatti che un bambino pranzi abitualmente a casa, ma nei giorni in cui è previsto pranzi a scuola.

Figura 5.19: *Output generato dal programma di check tra la domanda 2.13 e la domanda 3.3 del questionario autocompilato*

* OUTPUT GENERATO DAL PROGRAMMA DI CHECK *

Dove pranza abitualmente nei giorni non festivi (LPRAN)	Consuma il pranzo a scuola? (PRANSC) NO	SI, tutti i giorni	Si, solo nei giorni in cui è previsto	Totale
	27 0.49 11.84 11.07	43 0.79 18.86 1.64	138 2.52 60.53 8.42	20 0.37 8.77 2.05	228 4.16
In casa	155 2.83 4.45 63.52	2420 44.19 69.44 92.44	269 4.91 7.72 16.41	641 11.71 18.39 65.74	3485 63.64
In una mensa scolastica o all'asilo	55 1.00 3.50 22.54	35 0.64 2.23 1.34	1219 22.26 77.64 74.37	261 4.77 16.62 26.77	1570 28.67
In un ristorante	0 0.00 0.00 0.00	2 0.04 100.00 0.08	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.04
In un bar	1 0.02 20.00 0.41	3 0.05 60.00 0.11	0 0.00 0.00 0.00	1 0.02 20.00 0.10	5 0.09
A casa di genitori, parenti, amici	5 0.09 2.94 2.05	109 1.99 64.12 4.16	9 0.16 5.29 0.55	47 0.86 27.65 4.82	170 3.10
In altro luogo	1 0.02 6.25 0.41	6 0.11 37.50 0.23	4 0.07 25.00 0.24	5 0.09 31.25 0.51	16 0.29
Totale	244 4.46	2618 47.81	1639 29.93	975 17.80	5476 100.00

138 record errati riscontrati con la 3° incompatibilità.

282 (269+9+4) record errati riscontrati con la 2° incompatibilità.

35 record errati riscontrati con la 1° incompatibilità.

55 record errati riscontrati con la 4° incompatibilità.

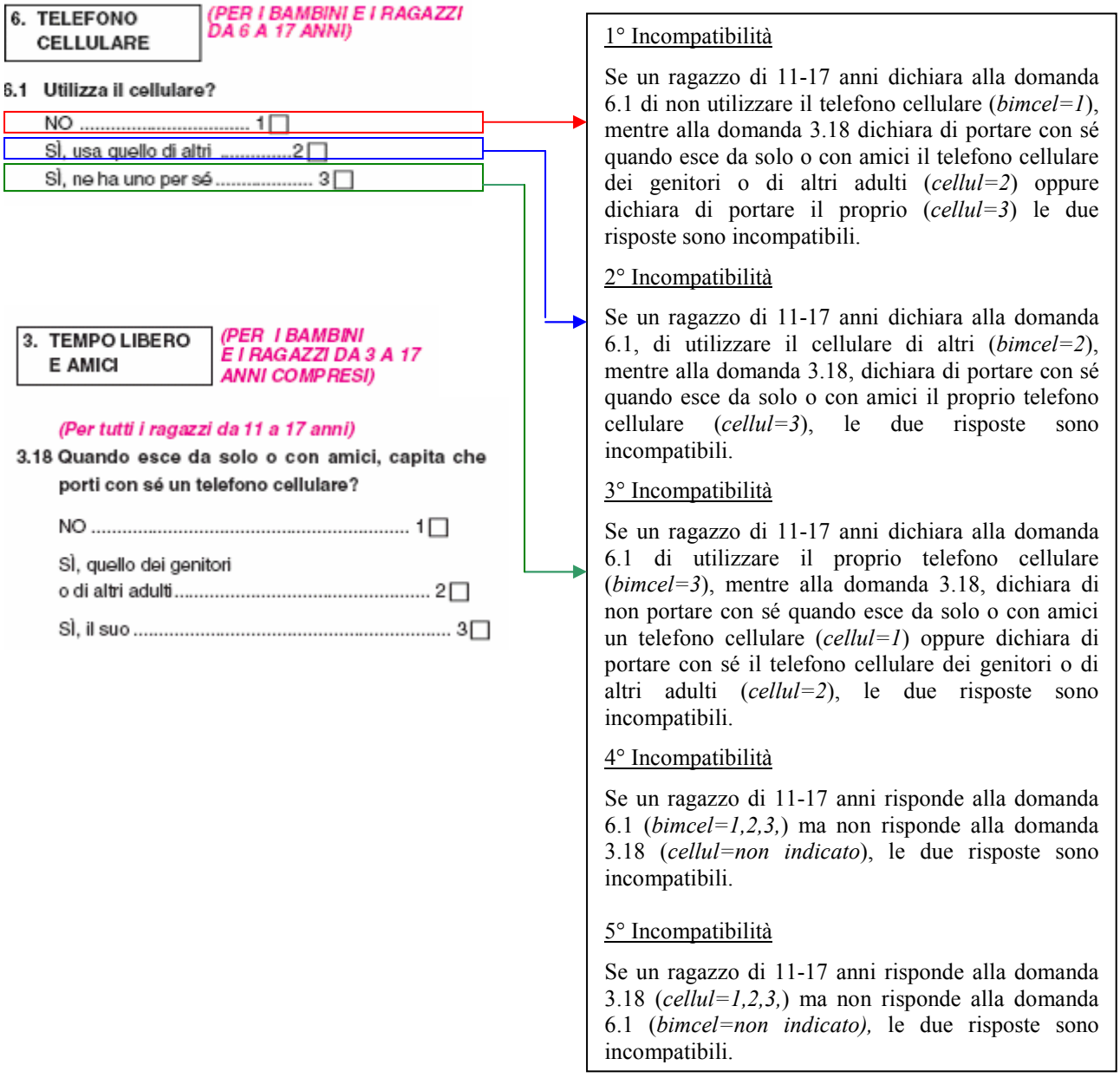
Le quattro incompatibilità logiche riscontrate sono state risolte applicando regole di correzione deterministica.
Il processo di correzione si considera concluso per i quesiti presenti nel questionario sui minori con l'imputazione delle mancate risposte come, ad esempio per i 27 bambini che non hanno risposto né alla domanda 2.13 del questionario dei minori, né alla domanda 3.3 del questionario autocompilato. In questo caso è stato applicato un passo di correzione successivo di tipo probabilistico che ha permesso l'imputazione di tutte le mancate risposte.

5.4.7 Incompatibilità tra domande di diverse sezioni dello stesso questionario

Una volta effettuate le correzioni deterministiche all'interno di ciascuna singola sezione del modello dei minori e corrette le incoerenze tra le risposte presenti in questo questionario e quelle riportate sugli altri questionari (rispettivamente incompatibilità interne alle sezioni e tra questionari), il processo è proseguito con l'individuazione e la correzione delle eventuali incompatibilità tra le diverse sezioni del questionario sui minori (modello giallo).
Analizzando il questionario per bambini e ragazzi da 0 a 17 anni sono state dapprima individuate quelle domande presenti in sezioni differenti che potevano presentare delle risposte logicamente legate tra loro, su queste sono stati realizzati i *check* che hanno consentito di verificare l'esistenza o meno di incoerenze nelle risposte fornite; in ultima fase, individuati gli errori, sono state generate le regole deterministiche di correzione.
All'interno del questionario dei minori sono state così individuate 19 possibili incompatibilità tra domande inserite in sezioni diverse del questionario. Un esempio di questo tipo di incompatibilità è rappresentato da due quesiti relativi al telefono cellulare che sono riportati nelle sezioni 3 e 6.
Con la domanda 3.18 appartenente alla Sezione 3 sul tempo libero e gli amici viene chiesto a tutti i ragazzi da 11 a 17 anni se, quando escono da soli o con amici, portano con loro un telefono cellulare. La domanda prevede tre possibili modalità di risposta: "No", "Si quello dei genitori o di altri adulti", "Si il suo".
Con la domanda 6.1, relativa alla Sezione 6 sul telefono cellulare si richiede a tutti i bambini e ragazzi da 6 a 17 anni se utilizzano il cellulare. Sono previste tre possibili risposte: "No", "Si usa quello di altri", "Si ne ha uno per sé".
Poiché la domanda 3.18 è rivolta a tutti i ragazzi da 11 a 17 anni, mentre la domanda 6.1 è rivolta ai bambini e ragazzi da 6 a 17 anni, le incompatibilità rilevabili riguardano la sottopopolazione dei ragazzi tra gli 11 e i 17 anni (3.633 casi).
Le possibili situazioni di incompatibilità logica tra queste due domande, che si possono risolvere deterministicamente, si hanno quando (Figura 5.20):

1. un ragazzo di 11-17 anni dichiara alla domanda 6.1 di non utilizzare il telefono cellulare, mentre alla domanda 3.18 dichiara di portare con sé, quando esce da solo o con amici, il telefono cellulare dei genitori o di altri adulti oppure dichiara di portare il proprio;
2. un ragazzo di 11-17 anni dichiara alla domanda 6.1 di utilizzare il cellulare di altri, mentre alla domanda 3.18 dichiara di portare con sé quando esce da solo o con amici il proprio telefono cellulare;
3. un ragazzo di 11-17 anni dichiara alla domanda 6.1 di utilizzare il proprio telefono cellulare, mentre alla domanda 3.18 dichiara di non portare con sé quando esce da solo o con amici un telefono cellulare oppure dichiara di portare con sé il telefono cellulare dei genitori o di altri adulti;
4. un ragazzo di 11-17 anni risponde alla domanda 6.1, ma non risponde alla domanda 3.18;
5. un ragazzo di 11-17 anni risponde alla domanda 3.18, ma non risponde alla domanda 6.1.

Figura 5.20: *Questionario sui minori: esempio di incompatibilità tra le domande 6.1 e 3.18. Indagine Aspetti della vita quotidiana – Anno 2005*



Non è stata considerata un'incompatibilità logica il caso in cui un ragazzo di 11-17 anni dichiara alla domanda 6.1 di utilizzare il telefono cellulare di altri e alla domanda 3.18 dichiara di non portare con sé quando esce da solo o con amici un telefono cellulare, in quanto si è ritenuto sia una situazione plausibile.

Il *check* delle incompatibilità tra le risposte date alla domanda 6.1 e alla domanda 3.18 è stato attuato soltanto dopo aver effettuato i *check* e le correzioni deterministiche all'interno della sezione 6. In seguito la domanda 6.1 è stata utilizzata come guida per la correzione della domanda 3.18.

Il controllo è stato effettuato incrociando le due variabili considerando tutti i bambini tra gli 11 e i 17 anni (3.633 casi). La procedura Sas applicata è riportata nella Figura 5.21.

La lettura dell'output prodotto dal *check* ha evidenziato cinque tipi di errori (Figura 5.22):

1. sono 39 (23+16) i record errati rilevati attraverso la 1° incompatibilità e rappresentano l'1% dei ragazzi di 11-17 anni;

2. sono 34 i record errati rilevati attraverso la 2° incompatibilità e rappresentano lo 0,9% dei ragazzi di 11-17 anni.
3. sono 65 (49+16) i record errati rilevati attraverso la 3° incompatibilità e rappresentano l'1,8% dei ragazzi di 11-17 anni.
4. sono 78 (26+3+39) i record errati rilevati attraverso la 4° incompatibilità e rappresentano l'1,9% dei ragazzi di 11-17 anni.
5. sono 24 (7+1+16) i record errati rilevati attraverso la 4° incompatibilità e rappresentano lo 0,6% dei ragazzi di 11-17 anni.

I record errati, individuati con i cinque tipi di incompatibilità sopra descritti, sono stati corretti attraverso l'applicazione di regole di correzione deterministica.

Figura 5.21: *Questionario sui minori: esempio di check tra le domande 6.1 e 3.18*

```
* CHECK DOMANDA 3.18 - 6.1;
  proc freq;
  where '011'<=eta<='017';    *FILTRO SEZIONE;
  format bimcel $bimcel. cellul $cellul.;
  table bimcel*cellul/missing;
  run;
```

Figura 5.22: *Output generato dal programma di check tra le domande 6.1 e 3.18*

```
* OUTPUT GENERATO DAL PROGRAMMA DI CHECK *
```

Utilizza il Cellulare? (BIMCELL)		Quando esce porta con sé un cellulare? (CELLUL) No	Quello dei genitori o di altri adulti	Si, il suo	Totale
	154	7	1	16	178
	4.24	0.19	0.03	0.44	4.90
	86.52	3.93	0.56	8.99	
	69.37	1.40	0.40	0.60	
No	26	366	23	16	431
	0.72	10.07	0.63	0.44	11.86
	6.03	84.92	5.34	3.71	
	11.71	73.05	9.09	0.60	
Si, usa quello di altri	3	79	213	34	329
	0.08	2.17	5.86	0.94	9.06
	0.91	24.01	64.74	10.33	
	1.35	15.77	84.19	1.28	
Si, ne ha uno per sé	39	49	16	2591	2695
	1.07	1.35	0.44	71.32	74.18
	1.45	1.82	0.59	96.14	
	17.57	9.78	6.32	97.52	
Totale	222	501	253	2657	3633
	6.11	13.79	6.96	73.14	100.00

78 (26+3+39) record errati riscontrati con la 4° incompatibilià

24 (7+1+16) record errati riscontrati con la 5° incompatibilià.

39 (23+16) record errati riscontrati con la 1° incompatibilià.

34 record errati riscontrati con la 2° incompatibilià.

65 (49+16) record errati riscontrati con la 3° incompatibilià.

Un analogo procedimento di *check* e correzione è stato realizzato per le restanti incompatibilità logiche individuate tra altre variabili presenti in sezioni diverse del questionario dei minori.
Tutti i record errati non sanabili con un metodo di correzione deterministica, ad esempio i 154 record corrispondenti a ragazzi di 11-17 anni che non hanno risposto né alla domanda 6.1 né alla domanda 3.18 sono stati corretti con un metodo di correzione probabilistico che ha permesso di imputare le mancate risposte.

5.5 Correzioni probabilistiche

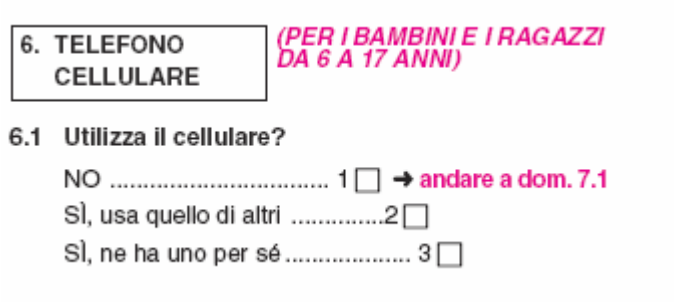
Dopo le correzioni deterministiche, il processo ha previsto l'applicazione di un passo di correzione probabilistico che ha consentito da un lato di sanare tutte le situazioni di incompatibilità non risolvibili con metodi deterministici e dall'altro di imputare i valori mancanti per tutte le variabili presenti nel questionario dei minori, sia qualitative che quantitative.
Al contrario delle correzioni deterministiche, per effettuare quelle probabilistiche non è necessario definire regole di correzione del tipo *If-then*. L'azione preliminare è la definizione delle situazioni di errore, nella fase successiva vengono individuati i record errati che un algoritmo di correzione, adeguatamente scelto, provvederà a sanare.
Per imputare le mancate risposte parziali presenti nel record, si è ricorso a *software* utilizzati in Istituto per l'imputazione dei dati: Scia (Sistema Controllo e Imputazione Automatici) e Rida (Ricostruzione delle Informazioni con Donazione Automatica). In generale Scia è stato usato per le variabili qualitative e Rida per le variabili quantitative.
Requisito fondamentale del file dati su cui si effettuano le imputazioni in modo probabilistico sono l'assenza di errori sistematici e la correttezza di tutte le variabili ausiliari che occorrono per l'esecuzione delle procedure di correzione probabilistica. Sia Scia che Rida lavorano distinguendo tra record esatti (donatori) e record errati (riceventi) e il valore da inserire (o sostituire) nel record errato è preso dal record esatto più somigliante al record da correggere. Alcune variabili sono dunque utilizzate per stabilire la similitudine tra record donatori e record riceventi.
In genere, le variabili utilizzate per questo scopo sono quelle relative alle caratteristiche socio-economiche dell'individuo (sesso, età, ecc.) e quelle che si ritengono in qualche modo correlate alla variabile da imputare. A questo punto del processo, le variabili delle principali caratteristiche dell'individuo sono tutte corrette, mentre va prestata particolare attenzione alle altre variabili che verranno utilizzate sia per la definizione delle incompatibilità sia per la distinzione tra record esatti ed errati. Questo implica una gerarchia nella scelta di quali variabili correggere per prime: per i quesiti preceduti da una domanda filtro, in genere, prima viene sanata la variabile filtro e successivamente le variabili che da essa dipendono.

5.5.1 Imputazione dei dati mancanti con Scia

Scia si basa sulla metodologia Fellegi-Holt. Questa prevede la definizione di un piano di incompatibilità tra le variabili, in cui ciascuna situazione di errore è definita dai valori delle variabili che non possono verificarsi contemporaneamente. Sulla base del piano di incompatibilità, l'algoritmo individua i record errati ed i record esatti. Ogni record è esaminato e se viola le regole di incompatibilità è considerato errato; la sua correzione avviene prendendo i valori corretti da un serbatoio di record donatori (record esatti).
La correzione viene effettuata rispettando i vincoli di correttezza e di minimalità, ovvero decidendo per ogni record e per ogni situazione di errore quali variabili modificare eliminando tutti gli errori individuati senza introdurne di nuovi (correttezza) e allo stesso tempo minimizzando il numero di variabili da modificare (minimalità). La scelta del record donatore avviene tentando di prendere sempre il record più somigliante al record errato, qualora ciò non fosse possibile la procedura corregge le variabili scegliendo tra i valori possibili della variabile quello che rende esatto il record.
Scia prevede che ogni variabile utilizzata per il piano di incompatibilità debba essere definita in una lista di variabili indicante per ognuna la posizione, la lunghezza e i valori che può assumere.
Alle variabili definite può essere associato un parametro crescente, detto grado di fissità, che va da 1 a 9. L'attribuzione del parametro alle variabili è facoltativo, se non è esplicitato la procedura associa lo

stesso livello di fissità a tutte le variabili, che sono ugualmente passibili di correzione, se però ci sono motivi per cui si vuole che una variabile sia modificata meno facilmente rispetto ad un'altra allora le si dà una fissità più alta delle altre.
Le variabili a cui si attribuisce fissità 9 sono considerate non modificabili ed entrano nella procedura di correzione solo come elementi delle regole di incompatibilità. La fissità pari a 9 è utilizzata per evitare ad esempio che la procedura corregga il record errato modificando valori a variabili già corrette in precedenza.
Ad alcune variabili, si può attribuire anche la proprietà di essere "chiave". Queste devono avere sempre fissità 9 e sono utilizzate come variabili di stratificazione per il cambio del serbatoio dei record dei donatori, al fine anche di evitare eccessivi riutilizzi di una stessa unità donatrice. Se una variabile è definita chiave si presuppone che i record esatti ed errati siano ordinati rispetto a questa e il rinnovo del serbatoio di record esatti avviene al cambio del valore della variabile chiave sui record errati[17].
Data la dimensione del file dati, si è deciso di lavorare separatamente ogni sezione del questionario e per ciascuna di esse è stato implementato un piano di incompatibilità. Nel caso di sezioni con quesiti logicamente legati tra loro, la correzione è stata effettuata anche tenendo in considerazione le incompatibilità logiche tra variabili appartenenti a sezioni diverse dello stesso questionario o di questionari differenti.
Di seguito, è mostrato in modo esemplificativo come è stata applicata la correzione alla variabile associata al primo quesito della sezione 6, in cui si chiede ai bambini e ai ragazzi da 6 a 17 anni se utilizzano il cellulare (Figura 5.23). Il quesito 6.1 è una domanda filtro a cui segue un'altra domanda sulla frequenza d'uso. La variabile associata alla domanda 6.1 (*bimcel*) dopo le correzioni deterministiche presentava una quota di *missing* pari al 4% dei bambini tra 6 e 17 anni.

Figura 5.23: *Questionario sui minori: domanda 6.1. Indagine Aspetti della vita quotidiana – Anno 2005*


6. TELEFONO CELLULARE

(PER I BAMBINI E I RAGAZZI DA 6 A 17 ANNI)

6.1 Utilizza il cellulare?

NO ..... 1 ☐ ➜ andare a dom. 7.1

Sì, usa quello di altri .....2 ☐

Sì, ne ha uno per sé ..... 3 ☐

Figura 5.24: *Esempio di definizione delle regole di incompatibilità in Scia*

1. ETA (6-17) BIMCEL( )	SE l'età è compresa tra 6 e 17 anni ALLORA la variabile bimcel non può essere missing.
2. ETA (6-17) TELCEL(1) BIMCEL(3)	SE l'età è compresa tra 6 e 17 anni e la famiglia non possiede telefoni cellulari (telcel=1) ALLORA la variabile bimcel non può essere uguale a 3 (Sì, ne ha uno per sé).

Le regole di incompatibilità utilizzate per questa domanda sono state due (Figura 5.24). La prima regola garantisce l'imputazione dei valori mancanti a tutti i bambini e ragazzi da 6 a 17 anni. La seconda regola invece prende in considerazione una variabile del questionario familiare che si riferisce al possesso da parte della famiglia di uno o più telefoni cellulari ed ha un duplice scopo:

1. correggere, qualora presenti, le incompatibilità di questo tipo,
2. impedire che la procedura di imputazione dei dati mancanti generi incompatibilità di questo tipo.

[17] Riccini, E., *CONCORD v. 1.0. Manuale Utente e aspetti metodologici*, Roma: Istat, 2002.

Sulla base delle regole definite, Scia individua tutti i record che presentano queste incompatibilità e li tratta come record errati. Si è inoltre ritenuto opportuno condizionare la correzione delle variabili con la definizione delle variabili fisse e delle variabili chiavi:

```
RIP   9K
CLASET9K
SESSO 9K
ETA   9
TELCEL9.
```

La ripartizione (*rip*), la classe di età (*claset*) e il sesso (*sesso*) sono state definite come variabili chiavi. Questo significa che i record dei donatori e i record da correggere sono stati ordinati rispetto ai valori assunti da queste variabili e il valore da inserire nel record errato per sanare le incompatibilità è scelto analizzando i record esatti che presentano le stesse caratteristiche.
L'età (*eta*) ed il possesso del telefono cellulare da parte della famiglia (*telcel*) sono state poste come variabili fisse, quindi non modificabili. Questo impedisce all'algoritmo di correzione di modificare il record errato, ad esempio, cambiando l'età, variabile ormai considerata "pilastro" per le correzioni, o modificando la variabile del questionario familiare che si ritiene corretta e non passibile di correzione. Per correggere i record e renderli compatibili rispetto alle regole definite l'unica variabile che l'algoritmo può quindi modificare è *bimcel*.
Al termine delle correzioni probabilistiche, le variabili trattate con Scia non presentano più casi di valori mancanti e il file dati dei corretti non ha nuovi casi di incompatibilità con altre variabili logicamente collegate e presenti nei questionari.

5.5.2 Imputazione dei dati mancanti con Rida

Rida utilizza l'imputazione da donatore con distanza mista minima, rispetto ad alcune variabili, dette di *matching*, ritenute determinanti per l'individuazione dei donatori. La funzione di distanza utilizzata per definire la somiglianza tra record ricevente e donatore è di tipo misto per consentire il trattamento simultaneo di variabili di diversa natura, per questo Rida è una procedura applicabile sia alle variabili qualitative che quantitative. Alle variabili di *matching* possono essere attribuiti pesi diversi nella funzione, qualora si ritenga che una variabile sia più rilevante di un'altra nell'individuare la somiglianza tra i record[18].
Oltre alle variabili di *matching*, è possibile definire delle variabili di strato. Le variabili di strato si utilizzano per limitare la ricerca del donatore all'interno di sottoinsiemi di record che presentano per queste variabili valori uguali al record ricevente.
A differenza di Scia, con questa procedura la distinzione tra record esatti (donatori) ed errati (riceventi) non avviene automaticamente, ma i record esatti devono essere preventivamente separati dai record errati con un'apposita procedura:

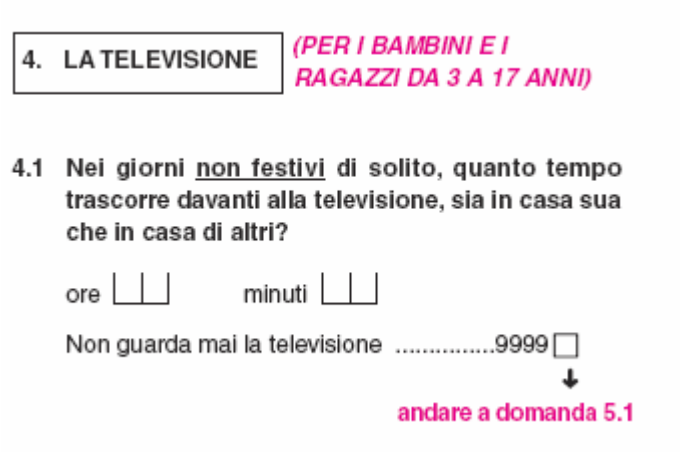
- il record è esatto quando non presenta valori mancanti o errati sia nelle variabili da imputare, che nelle variabili di *matching*;
- il record è errato quando ha valori mancati nella variabile da imputare e deve essere corretto rispetto alle variabili di strato e di *matching*. I record errati devono essere marcati con un carattere di errore che sostituisce il dato mancante per tutta la lunghezza della variabile da correggere.

Oltre ai file degli esatti e degli errati, Rida richiede la creazione di un file in cui sono indicati i parametri necessari per l'algoritmo di correzione. In esso si definiscono le variabili da utilizzare come strato, le variabili di *matching* e le variabili da imputare. Per ognuna di esse va segnalata la posizione, la lunghezza e il tipo (ad esempio se qualitativa o quantitativa); per le variabili di *matching* si può indicare il peso da attribuire loro nella funzione di distanza; per la variabile da imputare si indica il valore scelto come carattere di errore. In questo file, inoltre, è possibile definire dei parametri che regolamentano l'utilizzo dei record donatori (numero massimo di volte che un record può essere utilizzato come donatore, massima distanza ammissibile per considerare un record come donatore, fattore moltiplicativo da applicare alla funzione di distanza per penalizzare l'uso ripetuto dello stesso donatore).

[18] Abbate, C. , *La completezza delle informazioni e l'imputazione da donatore con distanza mista minima: il prodotto Rida (Ricostruzione delle Informazioni con Donazione Automatica),* Roma: Istat, 1993 (Documento interno).

L'imputazione tramite Rida non consente di tenere sotto controllo le incompatibilità con altre variabili. I record donatori sono esatti al loro interno, ma possono donare valori che contrastano con le modalità di altre variabili presenti nel record ricevente. Le azioni di imputazione avvengono indipendentemente dalla presenza di vincoli di coerenza fra le variabili. Un modo per affrontare il problema è quello di considerare le variabili correlate a quella da imputare come variabili di strato e costringere quindi l'algoritmo di correzione a scegliere il record donatore tra quelli che presentano una determinata caratteristica che è compatibile con quella che si va ad imputare. Questa soluzione, tuttavia, diventa impraticabile quando le variabili che possono creare un'incompatibilità sono tante o con tante modalità. In questi casi, la costruzione degli strati ottenuta incrociando molte modalità riduce drasticamente il serbatoio dei donatori per alcune tipologie di record riceventi e l'algoritmo non è in grado di trovare un donatore adeguato.
Anche per l'imputazione con Rida, ogni sezione del questionario minori è stata lavorata separatamente, generando ogni volta un serbatoio dei donatori corretti rispetto alle variabili considerate.
Rida è stata utilizzato, come già detto, per imputare le variabili quantitative presenti nel questionario dei minori. Un esempio è il quesito sul tempo trascorso davanti alla televisione dai ragazzi di 3-17 anni nei giorni non festivi.
Il primo passo è stato creare i file degli esatti e degli errati. Si sono considerati esatti tutti i record di bambini e ragazzi tra i 3 e i 17 anni che hanno indicato le ore e i minuti che hanno trascorso davanti la tv ed errati tutti gli altri relativi a chi non ha risposto al quesito.
Le ore e i minuti sono state trattate come un'unica variabile e i campi da imputare sono stati marcati con un carattere di errore. La percentuale dei record errati dopo le correzioni deterministiche era il 4,6% dei ragazzi tra 3 e 17 anni.

Figura 5.25: *Questionario sui minori: domanda 4.1. Indagine Aspetti della vita quotidiana – Anno 2005*

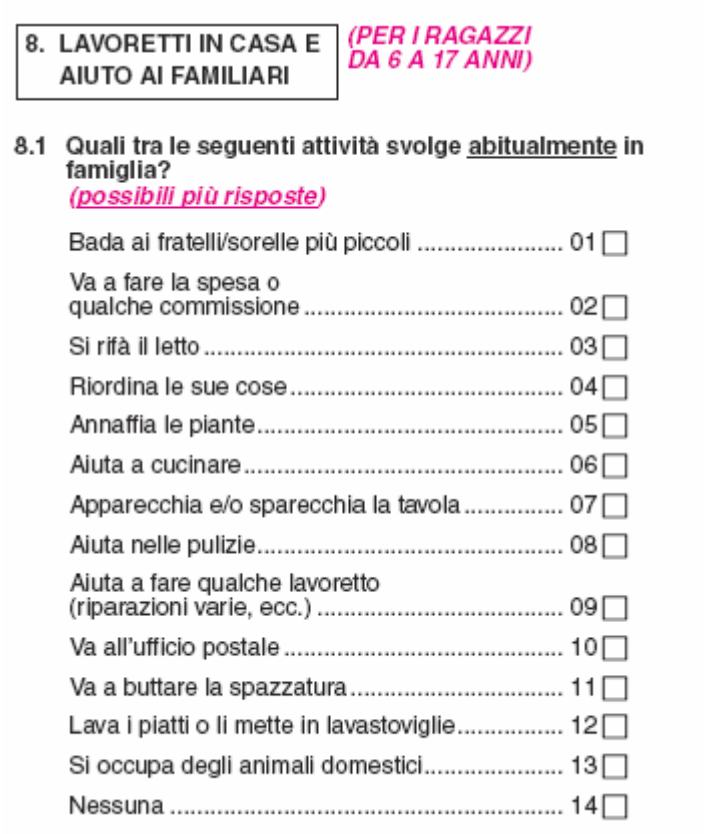


Al termine della procedura di imputazione il quesito non presentava più valori nulli, ma il file dati non poteva essere considerato ancora definitivamente corretto. L'imputazione con Rida non ha infatti tenuto conto delle possibili relazioni tra la variabile da imputare e le altre variabili che sono presenti nel questionario, come ad esempio i quesiti 5.8 e 5.11 del questionario dei minori in cui si chiede ai bambini se guardano la TV insieme al padre e/o alla madre. Quindi ad un bambino che ha dichiarato di guardare la televisione tutti i giorni con la madre (*mvedtv='1'*) e che non ha indicato alla domanda 4.1 le ore, il *software* Rida può aver attribuito il valore '9999'= non guarda la televisione, generando una nuova incompatibilità logica non presente nell'archivio originario dei dati.
Per tener conto di questa incompatibilità logica, le variabili sulla frequenza con cui guarda la tv con il padre (*pvedtv*) o con la madre (*mvedtv*) possono essere utilizzate come strato. In questo caso sono considerati record esatti quelli in cui chi guarda la televisione tutti i giorni con uno dei genitori ha indicato anche quante ore trascorre davanti la tv nei giorni non feriali: se il record errato ricevente ha modalità "tutti giorni" in una delle due variabili (*pvedtv* o *mvedtv*), allora la modalità che Rida imputa alle ore e i minuti non può essere '9999'= non guarda la televisione. Il risultato è un archivio di dati in cui non sono generate nuove incompatibilità di questo tipo.

5.5.3 Il caso particolare delle domande a risposta multipla: un confronto tra Scia e Rida

I quesiti che prevedevano la possibilità di più risposte sono stati oggetto di una particolare attenzione durante la fase di correzione probabilistica. Gli algoritmi di correzione utilizzati da Scia e Rida e applicati a questo tipo di quesiti conservano pressoché inalterate le distribuzioni semplici di ciascuna modalità di risposta, diverso, invece, è il comportamento dei due algoritmi di correzione rispetto al mantenimento delle distribuzioni di tutte le diverse combinazioni delle modalità di risposta del quesito considerato (distribuzioni congiunte).
Come esempio dell'impatto dei due metodi di correzione sulle distribuzioni di frequenza di quesiti che prevedono più risposte si è scelto il quesito 8.1 relativo ai lavoretti svolti in casa dai ragazzi tra i 6 e i 17 anni (Figura 5.26). La domanda prevede la possibilità di indicare più di una modalità di risposta. Il record è considerato errato quando non è presente nessuna delle modalità indicate, complessivamente la percentuale dei ragazzi che non ha indicato nulla è stata il 4,6%.
I valori mancanti del quesito sono stati imputati prima con Scia e poi con Rida e i risultati sono stati posti a confronto.

Figura 5.26: *Questionario sui minori: domanda 8.1. Indagine Aspetti della vita quotidiana – Anno 2005*



8. LAVORETTI IN CASA E AIUTO AI FAMILIARI *(PER I RAGAZZI DA 6 A 17 ANNI)*

8.1 Quali tra le seguenti attività svolge abitualmente in famiglia?
(possibili più risposte)

- Bada ai fratelli/sorelle più piccoli ..... 01 ☐
- Va a fare la spesa o qualche commissione ..... 02 ☐
- Si rifà il letto ..... 03 ☐
- Riordina le sue cose ..... 04 ☐
- Annaffia le piante ..... 05 ☐
- Aiuta a cucinare ..... 06 ☐
- Apparecchia e/o sparecchia la tavola ..... 07 ☐
- Aiuta nelle pulizie ..... 08 ☐
- Aiuta a fare qualche lavoretto (riparazioni varie, ecc.) ..... 09 ☐
- Va all'ufficio postale ..... 10 ☐
- Va a buttare la spazzatura ..... 11 ☐
- Lava i piatti o li mette in lavastoviglie ..... 12 ☐
- Si occupa degli animali domestici ..... 13 ☐
- Nessuna ..... 14 ☐

L'imputazione con Scia è stata fatta utilizzando una funzione del programma che permette di definire delle liste, utili ad agevolare la scrittura delle regole. In pratica una lista consente di considerare un insieme di variabili contemporaneamente in una regola. Queste variabili possono essere legate da un operatore logico "AND" o "OR". Nel primo caso la condizione deve essere verificata da tutte le variabili, nel secondo da almeno una. Tutte le 13 variabili relative ai lavoretti svolti più quella relativa a "nessuna attività" sono definite in una lista legate dall' "AND".
Sono state definite due condizioni di incompatibilità:

- la prima definisce un record errato se l'età è compresa tra i 6 e i 17 anni e la lista delle varabili sono tutte contemporaneamente *missing*;
- la seconda rende incompatibile la presenza di "nessuna attività" con una qualsiasi delle variabili relative ai lavoretti svolti.

Per applicare Rida, invece, sono stati considerati esatti i record dei bambini tra 6 e 17 anni in cui è stato dichiarato almeno un lavoretto e quelli in cui è stata indicata solo la modalità "nessuna attività". Nei record errati l'insieme delle modalità di risposta relative al quesito 8.1 è stato trattato come un'unica macrovariabile ed è stato marcato tutto con uno stesso carattere di errore. In questo modo nel record donatore l'insieme delle modalità da donare è stato considerato come se fosse un'unica variabile.
La Tavola 5.3 riporta le percentuali dei ragazzi di 6-17 anni che hanno dichiarato di aver svolto abitualmente le attività indicate nel quesito 8.1 calcolate per i rispondenti prima dell'imputazione e dopo l'imputazione effettuata con Scia e con Rida.

Tavola 5.3: *Ragazzi di 6-17 anni che hanno svolto diverse attività in famiglia. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 ragazzi di 6-17 anni)*

Tipo di attività svolte	PRIMA delle correzioni probabilistiche	DOPO le correzioni probabilistiche effettuate con:	
		Scia	Rida
Bada ai fratelli/sorelle più piccoli	23,2	22,5	22,9
Va a fare la spesa o qualche commissione	32,4	31,2	32,3
Si rifa il letto	30,4	29,3	30,3
Riordina le sue cose	58,6	56,3	58,8
Annaffia le piante	12,6	12,3	12,1
Aiuta a cucinare	20,8	20,1	20,5
Apparecchia e/o sparecchia la tavola	54,4	52,2	54,6
Aiuta nelle pulizia	24,8	24,1	24,7
Aiuta a fare qualche lavoretto (riparazioni varie, ecc.)	15,3	15,0	15,2
Va all'ufficio postale	5,5	5,6	5,4
Va a buttare la spazzatura	34,5	33,2	34,6
Lava i piatti o li mette nella lavastoviglie	18,9	18,4	19,1
Si occupa degli animali domestici	19,4	18,8	19,0
Nessuna	10,9	10,5	10,7

La quota dei ragazzi che svolge ciascuna attività rimane pressoché inalterata prima e dopo l'applicazione dei due metodi di correzione: le distribuzioni semplici delle variabili sono state preservate, come previsto.
Andando però ad analizzare la distribuzione dei rispondenti rispetto al numero di attività svolte per controllare l'effetto sulle distribuzioni congiunte, si sono evidenziate delle differenze. I risultati sono riportati nella Tavola 5.4, in essa è evidente che l'imputazione effettuata con Rida, trattando come un'unica variabile la lista delle modalità di risposta, ha consentito di mantenere inalterata anche la distribuzione congiunta.
L'imputazione con Scia, poiché avviene rendendo esatto il record secondo il principio del minimo cambiamento, ha determinato, invece, un incremento dei record in cui è indicata una sola attività, con conseguente alterazione delle distribuzioni congiunte. L'inserimento di una sola attività, infatti, è sufficiente a risolvere l'incompatibilità con il minimo cambiamento. La quota di bambini che hanno svolto una sola attività era l'11,9% prima delle correzione probabilistiche ed è rimasta 11,9% dopo le correzioni apportate con Rida, mentre sale al 15,5% dopo le correzioni effettuate con Scia.
Con Scia, diversamente che con Rida, non è possibile trattare l'insieme delle variabili come se fosse una variabile unica, perché è prevista per ciascuna variabile la definizione dei valori ammissibili. In questo caso i valori ammissibili sarebbero tutte le possibili combinazioni delle modalità non definibili con un *range*.

Tavola 5.4: *Ragazzi di 6-17 anni per numero di attività svolte in famiglia Indagine. Aspetti della vita quotidiana – Anno 2005 (per 100 ragazzi di 6-17 anni)*

Numero di attività svolte	PRIMA delle correzioni probabilistiche	DOPO le correzioni probabilistiche effettuate con:	
		Scia	Rida
0	10,6	10,5	10,7
1	11,9	15,5	11,9
2	16,9	16,2	17,1
3	16,9	16,1	16,9
4	13,5	12,9	13,5
5	9,4	9,0	9,5
6	7,5	7,1	7,4
7	5,1	4,9	5,1
8	3,6	3,5	3,5
9	2,5	2,4	2,5
10	1,1	1,0	1,1
11	0,5	0,5	0,5
12	0,3	0,3	0,3
13	0,1	0,1	0,0

In pratica si è sfruttata la maggiore duttilità di Rida rispetto a Scia, che la rende applicabile anche al trattamento di variabili qualitative non ordinabili. Le modalità di risposta della domanda 8.1 sono, infatti, state trattate come un'unica variabile qualitativa sconnessa.
In sintesi, le caratteristiche di Scia non hanno permesso di operare preservando adeguatamente la distribuzione congiunta delle variabili, quando ciò sarebbe opportuno, come nel caso delle domande a risposta multipla. Questo ha portato alla scelta di imputare i valori dei quesiti che prevedevano più di una modalità di risposta con Rida.

5.6 Controlli e correzioni deterministiche finali

Terminata la fase di correzione probabilistica, è stato necessario effettuare un ulteriore fase di controllo e correzione deterministica dei dati. La fase di imputazione dei valori mancanti può infatti aver generato *ex-novo* incompatibilità logiche analizzate e risolte deterministicamente in precedenza.
Scia prevede la localizzazione automatica dell'errore tramite delle regole di incompatibilità e l'algoritmo di correzione sceglie i valori da inserire nel record errato in modo da non violare queste regole. È quindi possibile operare in modo da tenere sotto controllo le relazioni logiche tra le diverse variabili. L'imputazione con Rida invece, come già detto, non consente di tenere sempre sotto controllo le incompatibilità logiche.
In questa fase del processo di correzione è allora opportuno far passare di nuovo il piano dei *check* utilizzato a monte del processo di correzioni deterministiche per il controllo delle incompatibilità logiche tra domande appartenenti a sezioni diverse (dello stesso questionario o di questionari diversi).
Un esempio è dato dal quesito 7.4, in cui si chiede ai ragazzi tra i 6 e i 17 anni in che modo spendono la paghetta. Le procedure di controllo hanno individuato come incompatibile la modalità di risposta del quesito 7.4 spende la paghetta in "ricariche del telefono" con la modalità "non utilizza il cellulare" della domanda 6.1 e l'incompatibilità è stata corretta in modo deterministico.
Il quesito su come spende la paghetta prevede la possibilità di indicare più modalità di risposta e per questa ragione i valori mancanti sono stati imputati con Rida. Dopo aver imputato è stato allora necessario ripetere le procedure di controllo e, se presente l'incompatibilità, sanarla nei nuovi record errati con la correzione deterministica.

6. Indicatori di qualità

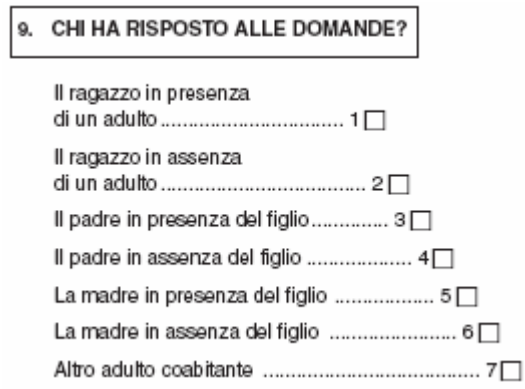
6.1 Chi ha risposto alle domande

Nella progettazione del questionario dei bambini e ragazzi di 0-17 anni un'attenzione particolare è stata dedicata alle modalità di compilazione del questionario al fine di tenere sotto controllo le risposte *proxy*. Con questo termine si indicano le risposte date da un componente della famiglia per conto di un altro (ad esempio, la madre per il figlio).
Esistono difficoltà oggettive che suggeriscono l'utilizzo di questa tecnica (ad esempio assenza del figlio per tutto il periodo della rilevazione), che però va valutato con oculatezza, sia in relazione all'oggetto dell'indagine sia in relazione all'età dell'intervistato. Ad esempio supponiamo che la madre risponda per il figlio di 14 anni ad un quesito sul tipo di malattie infettive che il figlio ha avuto durante l'infanzia, probabilmente la madre può rispondere addirittura meglio del figlio. Ma se consideriamo il quesito relativo a quali funzioni del cellulare il figlio utilizza, probabilmente il figlio è maggiormente in grado di rispondere e il risultato in termini di attendibilità dell'informazione raccolta sarà sicuramente migliore.
Oltre al tipo di fenomeno rilevato un altro elemento importante da considerare, quando in fase di progettazione dell'indagine si decide di accettare la risposta *proxy*, è l'età dell'intervistato. Per bambini molto piccoli, infatti, le riposte *proxy* non solo sono consigliate, ma addirittura necessarie (si pensi, in particolare, ai bambini nella fascia 0-5 anni).
In tutte le indagini multiscopo è stata sempre seguita la linea metodologica di ridurre la molestia statistica sui bambini fino a 13 anni, per questa fascia d'età quindi tutte le informazioni vengono chieste ai genitori (preferibilmente la madre).
In linea con le scelte metodologiche già effettuate in altre indagini multiscopo, anche nella progettazione del questionario sui minori, si è deciso di accettare interviste *proxy* per i bambini e ragazzi fino a 13 anni, pertanto:

- per i bambini e ragazzi fino a 13 anni ha risposto alle domande un genitore (preferibilmente la madre) o un'altra persona adulta della famiglia (intervista *proxy*);
- i ragazzi tra i 14 e i 17 anni hanno risposto personalmente.

Per analizzare chi ha fornito le risposte è stato inserito alla fine del questionario un quesito (da compilare a cura del rilevatore) in cui sono state chieste informazioni sulla compilazione del modello stesso. Il quesito mira a rilevare se l'intervistato ha risposto direttamente alle domande (da solo o in presenza di un adulto), se invece un genitore (distinguendo tra padre e madre) ha risposto per lui (da solo o in presenza del figlio) (Figura 6.1).

Figura 6.1: *Questionario sui minori: domanda 9.1. Indagine Aspetti della vita quotidiana – Anno 2005*


9. CHI HA RISPOSTO ALLE DOMANDE?

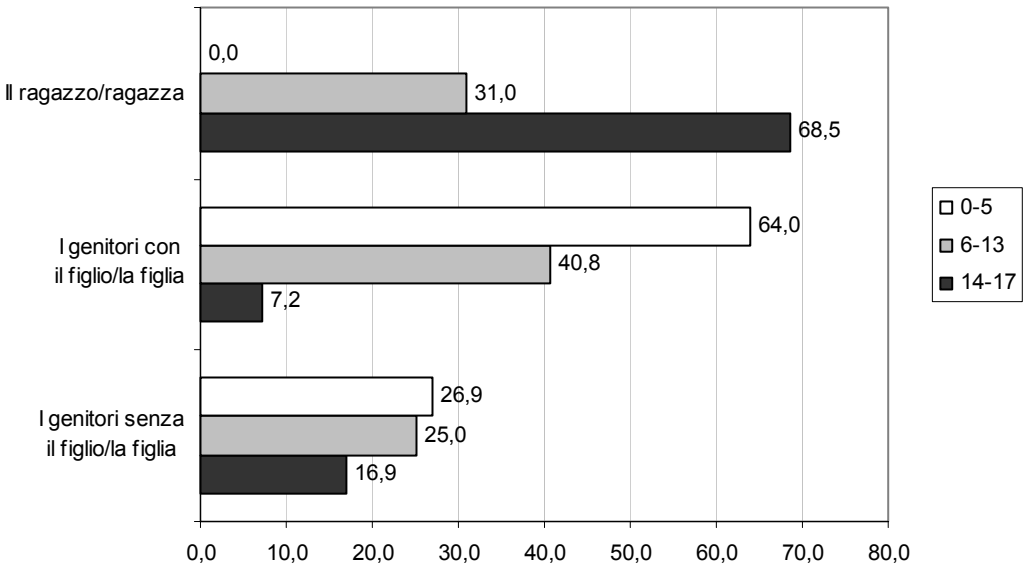
Il ragazzo in presenza di un adulto ..... 1 ☐
Il ragazzo in assenza di un adulto ..... 2 ☐
Il padre in presenza del figlio ..... 3 ☐
Il padre in assenza del figlio ..... 4 ☐
La madre in presenza del figlio ..... 5 ☐
La madre in assenza del figlio ..... 6 ☐
Altro adulto coabitante ..... 7 ☐

Come si è detto, in generale, si considerano *proxy* le risposte date da un componente per conto di un altro membro della famiglia, queste sono esplicitamente richieste per i bambini in età 0-13 (indipendentemente dalla presenza o meno del bambino), nella fascia tra i 14 e i 17 anni invece possono essere effettuate solo in casi eccezionali e in generale sono sconsigliate. Tuttavia, se si considerano tre fasce d'età e l'eventuale presenza/assenza del bambino si possono verificare le seguenti situazioni:

- per i bambini in età 0-5 anni le risposte sono, ovviamente, tutte *proxy* e la presenza o meno del bambino può non avere un effetto sull'informazione raccolta;
- passando alla fascia d'età 6-13, la presenza del figlio al momento dell'intervista può influenzare la qualità delle informazioni raccolte in *proxy* se si ipotizza, ad esempio, che eventuali informazioni non pertinenti fornite dal genitore possano essere corrette dal figlio presente (si pensi ad informazioni relative ai servizi o attrezzature utilizzate a scuola o al numero di ore settimanali di svolgimento di corsi scolastici, di cui non sempre il genitore può essere a conoscenza), in ogni caso le informazioni raccolte possono essere frutto di un'interazione;
- per i ragazzi di 14-17 anni (per i quali, lo ricordiamo, non erano previste le risposte *proxy* se non in casi eccezionali), nel caso abbia risposto un genitore si possono fare le stesse considerazioni metodologiche della fascia 6-13.

In base a quanto detto, nella presentazione e nell'analisi dei risultati si sono tenute distinte queste situazioni. Ovviamente nei casi in cui l'informazione è stata raccolta in assenza dell'intervistato l'effetto *proxy* è più netto.

Figura 6.2: *Bambini e ragazzi di 0-17 anni per modalità di compilazione del questionario e classe di età. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 bambini e ragazzi di 0-17 anni della stessa classe di età)*



Per quanto riguarda i bambini di 6-13 anni, per i quali era richiesta la risposta *proxy*, per il 75,8% di essi ha risposto un genitore (nel 40,8% dei casi era presente anche il figlio), mentre il 31% ha risposto direttamente alle domande poste dal rilevatore (Tavola 6.1).

Considerando i ragazzi tra i 14 e i 17 anni per i quali non era richiesta la risposta *proxy*, di questi il 68,5% ha risposto direttamente alle domande e il 7,2% pur non avendo risposto direttamente era però presente all'intervista, mentre le interviste totalmente *proxy* sono state il 16,9%. In questa fascia di età, la quota di coloro che hanno risposto direttamente è maggiore tra le ragazze (il 72,5% rispetto al 64,6% dei ragazzi) (Tavola 6.2).

Inoltre considerando il territorio la quota di coloro che hanno risposto direttamente risulta più alta nel Nord dove si colloca sul 70,4%, mentre nel Centro-sud scende sotto il 68% (Tavola 6.3).

Tavola 6.1: *Bambini e ragazzi di 0-17 anni per modalità di compilazione del questionario e classe di età. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 bambini e ragazzi di 0-17 anni della stessa classe di età)*

CLASSI DI ETA'	Modalità di compilazione del questionario					Totale
	Il ragazzo/la ragazza	Il padre/la madre		Altro adulto coabitante	Non indicato	
		con il figlio	senza il figlio			
0-5	0,0	64,0	26,9	1,3	7,9	100,0
6-13	31,0	40,8	25,0	1,1	2,1	100,0
14-17	68,5	7,2	16,9	1,1	6,3	100,0
Totale	**30,4**	**39,9**	**23,6**	**1,2**	**4,9**	**100,0**
Valori assoluti	*2.658*	*3.486*	*2.063*	*101*	*431*	*8.739*

Tavola 6.2: *Bambini e ragazzi di 0-17 anni per modalità di compilazione del questionario, sesso e classe di età. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 bambini e ragazzi di 0-17 anni dello stesso sesso e classe di età)*

SESSO E CLASSI DI ETA'	Modalità di compilazione del questionario					Totale
	Il ragazzo/la ragazza	Il padre/la madre		Altro adulto coabitante	Non indicato	
		con il figlio	senza il figlio			
			MASCHI			
0-5	0,0	65,4	25,1	1,3	8,2	100,0
6-13	29,1	41,7	25,9	1,2	2,1	100,0
14-17	64,6	8,4	19,9	1,4	5,7	100,0
Totale	**28,3**	**41,3**	**24,2**	**1,3**	**4,9**	**100,0**
Valori assoluti	*224*	*1.299*	*1.894*	*1.110*	*58*	*4.585*
			FEMMINE			
0-5	0,0	62,3	28,9	1,2	7,6	100,0
6-13	33,2	39,8	24,0	1,0	2,1	100,1
14-17	72,5	6,0	13,9	0,9	6,8	100,1
Totale	**32,7**	**38,3**	**22,9**	**1,0**	**5,0**	**99,9**
Valori assoluti	*1.359*	*1.592*	*953*	*43*	*207*	*4.154*

Tavola 6.3: *Bambini e ragazzi di 0-17 anni per modalità di compilazione del questionario, ripartizione geografica e classe di età. Indagine Aspetti della vita quotidiana – Anno 2005 (per 100 bambini e ragazzi di 0-17 anni della stessa ripartizione geografica e classe di età)*

RIPARTIZIONE GREOGRAFICA E CLASSI DI ETÀ	Modalità di compilazione del questionario					Totale
	Il ragazzo/la ragazza	Il padre/la madre		Altro adulto coabitante	Non indicato	
		con il figlio	senza il figlio			
			NORD			
0-5	0,0	62,2	29,7	1,1	7,0	100,0
6-13	31,1	38,3	27,1	1,2	2,3	100,0
14-17	70,4	5,8	17,5	1,0	5,3	100,0
Totale	**29,2**	**39,2**	**25,9**	**1,1**	**4,6**	100,0
Dati in migliaia	*946*	*1.269*	*837*	*36*	*149*	*3.237*
			CENTRO			
0-5	0,0	60,2	28,1	2,0	9,7	100,0
6-13	33,8	37,1	26,3	1,4	1,4	100,0
14-17	66,7	6,3	18,4	1,4	7,2	100,0
Totale	**31,2**	**37,0**	**25,0**	**1,6**	**5,2**	100,0
Dati in migliaia	*465*	*552*	*374*	*24*	*79*	*1.494*
			SUD			
0-5	0,0	67,1	23,7	1,2	8,0	100,0
6-13	29,9	44,4	22,7	0,8	2,2	100,0
14-17	67,7	8,5	16,1	1,1	6,6	100,0
Totale	**31,1**	**41,5**	**21,3**	**1,0**	**5,1**	100,0
Dati in migliaia	*1.247*	*1.665*	*852*	*41*	*203*	*4.008*

6.2 Indicatori di valutazione globale degli effetti del processo di correzione

Al termine delle procedure di controllo, correzione e imputazione si è ottenuta una matrice dei dati "pulita" da tutte le incompatibilità individuabili tra i quesiti delle diverse sezioni e priva di mancate risposte parziali.

Per valutare l'impatto del processo di correzione e imputazione sulla matrice dei dati originali, si è pensato di calcolare degli indicatori attraverso uno strumento disponibile in Istituto chiamato IDEA (*Indices for Data Editing Assesment*). Questo applicativo fornisce un insieme di indicatori utili per la valutazione globale degli effetti del processo di correzione. Tali indicatori sono calcolati ponendo a confronto la matrice dei dati grezzi con la matrice dei dati ottenuta dopo il processo di correzione[19].

La valutazione del processo è stata fatta sugli 8.739 record di bambini e ragazzi tra 0-17 anni e le 319 variabili rilevate con il questionario sui minori, tutte soggette a controllo e passibili di correzione ed imputazione: la matrice risultante ha una dimensione pari a 8.739 x 319 = 2.787.741 campi.

Il processo di correzione e imputazione può modificare i valori della matrice in tre modi:

1. da codice a codice diverso;
2. da *missing* a codice;
3. da codice a *missing*.

Per ciascun tipo di modifica è calcolato un indicatore che dà una valutazione in percentuale dei campi modificati, rapportando i valori modificati su tutti quelli passibili di correzione.

[19] Manuale IDEA, *Indices for Data Editing Assessment*, Indicatori per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI, Istat, Metodologie di base per la produzione statistica.

Considerando i diversi tipi di correzione, si hanno:

1. tasso di modificazione (per le correzioni da codice a codice diverso);
2. tasso di imputazione netta (per le correzioni da *missing* a codice);
3. tasso di cancellazione (per le correzioni da codice a *missing*).

La Tavola 6.4 mostra che in totale il 2,3% dei dati presenti nella matrice ha subito una correzione, in particolare l'1,4% di esse ha riguardato l'attribuzione di un codice alla mancata risposta parziale, mentre solo per lo 0,2% dei casi il dato originale è stato modificato in un altro codice e nello 0,7% si è proceduto alla cancellazione di valori errati.

Tavola 6.4: *Indicatori di valutazione globale degli effetti del processo di correzione sulla matrice dei dati del Questionario sui minori. Indagine Aspetti della vita quotidiana – Anno 2005*

Numero osservazioni	**8.739**
Variabili soggette a correzione	**319**
	%
Tasso di modificazione	0,2
Tasso di imputazione netta	1,4
Tasso di cancellazione	0,7
Totale	**2,3**

La Tavola 6.5 riporta per i campi che hanno subito una correzione le percentuali del tipo di errore riscontrati per età del rispondente. In oltre il 60% dei casi dei ragazzi con più di 6 anni, le correzioni hanno riguardato l'imputazione di valori mancanti, mentre per i più piccoli, la maggior parte delle correzioni è stata la cancellazione di valori.

Tavola 6.5: *Indicatori di valutazione globale degli effetti del processo di correzione sulla matrice dei dati del Questionario sui minori per classe di età. Indagine Aspetti della vita quotidiana – Anno 2005*

CLASSE DI ETA'	% modificazione	% imputazione netta	% cancellazione	Totale
0-5	9,4	38,2	52,4	100,0
6-10	13,1	66,0	20,9	100,0
11-17	9,1	64,8	26,1	100,0
Totale	**10,2**	**60,0**	**29,8**	**100,0**

Bibliografia

Abbate, C. , *La completezza delle informazioni e l'imputazione da donatore con distanza mista minima: il prodotto Rida (Ricostruzione delle Informazioni con Donazione Automatica),* Roma: Istat, 1993 (Documento interno).

Istat. *Il sistema di indagini multiscopo*. Roma: Istat, 2006 (Metodi e Norme n. 31).

Istat. *La vita quotidiana di bambini e ragazzi*, Roma: Istat, 2000 (Informazioni n. 23).

Istat. *Strategie di correzione del file dati relativo all'indagine "Tempo libero e cultura" Anno 1995*. Roma: Istat, 1998 (Documenti n.2).

Istat. *Il mondo dei bambini*. Roma: Istat, 1994.

Riccini, E., *CONCORD v. 1.0. Manuale Utente e aspetti metodologici*, Roma: Istat, 2002.

Appendice

Aree tematiche, filtri di sezione e indicatori presenti nel Questionario sui minori

Area tematiche	Filtro di sezione	Indicatori
Nonni e affidamento del bambino	Per i bambini e ragazzi da 0 a 17 anni	Frequenza con cui vede i nonni Persone adulte a cui è abitualmente affidato il bambino quando non è con i genitori o a scuola Persone adulte a cui è abitualmente affidato il bambino quando non è con i genitori o a scuola
La scuola	Per i bambini e ragazzi da 0 a 17 anni	Iscrizione Servizi o attrezzature scolastiche utilizzate Partecipazione a corsi di recupero Partecipazione a corsi organizzati dalla scuola al di fuori dell'orario scolastico Tipo di corsi svolti e numero di ore settimanali Svolgimento compiti a casa Persone con cui svolge i compiti a casa Numero ore impiegate per fare i compiti Comportamento verso lo studio Promozione nell'anno appena trascorso Presenza in classe di compagni stranieri e numero Incontro compagni stranieri al di fuori dell'orario scolastico Consumo del pranzo a scuola Motivi per cui non consuma il pranzo a scuola Frequenza in passato dell'asilo o della scuola dell'infanzia
Tempo libero e amici	Per i bambini e ragazzi da 3 a 17 anni	Partecipazione a corsi non organizzati dalla scuola Tipo di corsi svolti e numero di ore settimanali Incontro con i coetanei nel tempo libero e numero di coetanei Frequenza con cui vede i coetanei Frequenta più maschi o più femmine Partecipazione a feste Luogo in cui erano organizzate le feste In quale stanza della casa dorme Possesso di una tv personale Possesso del pc a casa Atti di prepotenza da pare dei coetanei Partecipazione ad attività di associazioni ricreative, culturali, ambientali, boy-scouts Frequenza con cui partecipa ad attività di associazioni ricreative, culturali, ambientali, boy-scouts Frequenza con cui si è recato in sala giochi, fast food, strada piazza, oratorio, parrocchia, luoghi di lavoro di familiari, spazi condominiali, cortili Frequenza con cui esce da solo o con amici di giorno Frequenza con cui esce da solo o con amici di sera Ora del rientro quando esce la sera Frequenza con cui si è recato in: bar, birreria, pub, pizzeria, discoteca, stadio
La televisione	Per i bambini e ragazzi da 3 a 17 anni	Tempo trascorso davanti alla televisione nei giorni non festivi Momenti della giornata in cui guarda la televisione e con chi Tramissioni segutite alla tv I genitori fanno attenzione ai programmi e/o videocassette viste dai bambini
Il gioco	Per i bambini e ragazzi da 3 a 13 anni	Dove giova il bambino Con chi gioca nei giorni non festivi Con chi gioca nei giorni festivi Freqnenza con cui va a giardini/parchi attrezzati, giardini/parchi non attrezzati, luna park/giostre, sale giochi Quali sono i giochi preferiti Frequenza con cui gioca con il padre Giochi svolti insiema al padre Frequenza con cui il padre svolge alcune attività Frequenza con cui gioca con la madre Giochi svolti insieme alla madre Frequenza con cui la madre svolge alcune attività Frequenza con cui capita che il bambino si annoi
Telefono cellulare	Per i bambini e ragazzi da 6 a 17 anni	Utilizzo del telefono celluare Frequenza di utilizzo del telefono cellulare Persone con cui comunica con il cellulare Funzioni del cellulare utilizzate
Chiavi di casa a utonomia	Per i bambini e ragazzi da 6 a 17 anni	Disponibilità delle chiavi di casa Riceve regolarmente una somma di denaro dai genitori Quota settimanale ricevuta Come spende questa somma di denaro Abitudine a risparmiare
Lavoretti in casa e aiuto ai familiari	Per i bambini e ragazzi da 6 a 17 anni	Attività svolte abitualmente in famiglia

Contributi ISTAT(*)

1/2002 - Francesca Biancani, Andrea Carone, Rita Pistacchio e Giuseppina Ruocco - *Analisi delle imprese individuali*

2/2002 - Massimiliano Borgese .-.*Proposte metodologiche per un progetto d'indagine sul trasporto aereo alla luce della recente normativa comunitaria sul settore*

3/2002 - Nadia Di Veroli e Roberta Rizzi - *Proposta di classificazione dei rapporti di lavoro subordinato e delle attività di lavoro autonomo: analisi del quadro normativo*

4/2002 – Roberto Gismondi – *Uno stimatore ottimale in presenza di non risposte*

5/2002 – Maria Anna Pennucci – *Le strategie europee per l'occupazione dal Libro bianco di Delors al Consiglio Europeo di Cardiff*

1/2003 – Giovanni Maria Merola – *Safety Rules in Statistical Disclosure Control for Tabular Data*

2/2003 – Fabio Bacchini, Pietro Gennari e Roberto Iannaccone – *A new index of production for the construction sector based on input data*

3/2003 – Fulvia Cerroni e Enrica Morganti – *La metodologia e il potenziale informativo dell'archivio sui gruppi di impresa: primi risultati*

4/2003 – Sara Mastrovita e Isabella Siciliani – *Effetti dei trasferimenti sociali sulla distribuzione del reddito nei Paesi dell'Unione europea: un'analisi dal Panel europeo sule famiglie*

5/2003 – Patrizia Cella, Giuseppe Garofalo, Adriano Paggiaro, Nicola Torelli e Caterina Viviano – *Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità*

6/2003 – Enrico Grande e Orietta Luzi – *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat*

7/2003 – Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Indagine sperimentale sui posti di lavoro vacanti*

8/2003 – Mario Adua – *L'agricoltura di montagna: le aziende delle donne, caratteristiche agricole e socio-rurali*

9/2003 – Franco Mostacci e Roberto Sabbatini – *L'euro ha creato inflazione? Changeover e arrotondamenti dei prezzi al consumo in Italia nel 2002*

10/2003 – Leonello Tronti – *Problemi e prospettive di riforma del sistema pensionistico*

11/2003 – Roberto Gismondi – *Tecniche di stima e condizioni di coerenza per indagini infraannuali ripetute nel tempo*

12/2003 – Antonio Frenda – *Analisi delle legislazioni e delle prassi contabili relative ai gruppi di imprese nei paesi dell'Unione Europea*

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys):*

un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi **–** *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo **–** *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni **–** *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 **–** Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo **–** *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

(*) ultimi cinque anni

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

5/2007 – Giulio Barcaroli, Tiziana Pellicciotti – *Strumenti per la documentazione e diffusione dei microdati d'indagine: il Microdata Management Toolkit*

Documenti ISTAT(*)

1/2002 – Paolo Consolini e Rita De Carli - *Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati*

2/2002 – Stefania Macchia - *Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività economiche*

3/2002 – Maria De Lucia - *Applicabilità della disciplina in materia di festività nel pubblico impiego*

4/2002 – Roberto Gismondi, Massimo Marciani e Mauro Giorgetti - *The italian contribution towards the implementantion of an european transport information system: main results of the MESUDEMO project*

5/2002 – Olimpio Cianfarani e Sauro Angeletti - *Misure di risultato e indicatori di processo: l'esperienza progettuale dell'Istat*

6/2002 – Riccardo Carbini e Valerio De Santis – *Programma statistico nazionale: specifiche e note metodologiche per la compilazione delle schede identificative dei progetti*

7/2002 – Maria De Lucia – *Il CCNL del personale dirigente dell'area 1 e la valutazione delle prestazioni dei dirigenti*

8/2002 – Giuseppe Garofalo e Enrica Morganti – *Gruppo di lavoro per la progettazione di un archivio statistico sui gruppi d'impresa*

1/2003 – Francesca Ceccato, Massimiliano Tancioni e Donatella Tuzi – *MODSIM-P:Il nuovo modello dinamico di previsione della spesa pensionistica*

2/2003 – Anna Pia Mirto – *Definizioni e classificazioni delle strutture ricettive nelle rilevazioni statistiche ufficiali sull'offerta turistica*

3/2003 – Simona Spirito – *Le prestazioni assistenziali monetarie non pensionistiche*

4/2003 – Maria De Lucia – *Approfondimenti di alcune tematiche inerenti la gestione del personale*

5/2003 – Rosalia Coniglio, Marialuisa Cugno, Maria Filmeno e Alberto Vitalini – *Mappatura della criminalità nel distretto di Milano*

6/2003 – Maria Letizia D'Autilia – *I provvedimenti di riforma della pubblica amministrazione per l'identificazione delle "Amministrazioni pubbliche" secondo il Sec95: analisi istituzionale e organizzativa per l'anno 2000*

7/2003 – Francesca Gallo, Pierpaolo Massoli, Sara Mastrovita, Roberto Merluzzi, Claudio Pauselli, Isabella Siciliani e Alessandra Sorrentino – *La procedura di controllo e correzione dei dati Panel Europeo sulle famiglie*

8/2003 – Cinzia Castagnaro, Martina Lo Conte, Stefania Macchia e Manuela Murgia – *Una soluzione in-house per le indagini CATI: il caso della Indagine Campionaria sulle Nascite*

9/2003 – Anna Pia Maria Mirto e Norina Salamone – *La classificazione delle strutture ricettive turistiche nella normativa delle regioni italiane*

10/2003 – Roberto Gismondi e Anna Pia Maria Mirto – *Le fonti statistiche per l'analisi della congiuntura turistica: il mosaico italiano*

11/2003 – Loredana Di Consiglio e Stefano Falorsi – *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento Generale della Popolazione 2001*

12/2003 – Roberto Gismondi e Anna Rita Giorgi – *Struttura e dinamica evolutiva del comparto commerciale al dettaglio: le tendenze recenti e gli effetti della riforma "Bersani"*

13/2003 – Donatella Cangialosi e Rosario Milazzo – *Fabbisogni formativi degli Uffici comunali di statistica: indagine rapida in Sicilia*

14/2003 – Agostino Buratti e Giovanni Salzano – *Il sistema automatizzato integrato per la gestione delle rilevazioni dei documenti di bilancio degli enti locali*

1/2004 – Giovanna Brancato e Giorgia Simeoni – *Tesauri del Sistema Informativo di Documentazione delle Indagini (SIDI)*

2/2004 – Corrado Peperoni – *Indagine sui bilanci consuntivi degli Enti previdenziali: rilevazione, gestione e procedure di controllo dei dati*

3/2004 – Marzia Angelucci, Giovanna Brancato, Dario Camol, Alessio Cardacino, Sandra Maresca e Concetta Pellegrini – *Il sistema ASIMET per la gestione delle Note Metodologiche dell'Annuario Statistico Italiano*

4/2004 – Francesca Gallo, Sara Mastrovita, Isabella Siciliani e Giovanni Battista Arcieri – *Il processo di produzione dell'Indagine ECHP*

5/2004 – Natale Renato Fazio e Carmela Pascucci – *Gli operatori non identificati nelle statistiche del commercio con l'estero: metodologia di identificazione nelle spedizioni "groupage" e miglioramento nella qualità dei dati*

6/2004 – Diego Moretti e Claudia Rinaldelli – *Una valutazione dettagliata dell'errore campionario della spesa media mensile familiare*

7/2004 – Franco Mostacci – *Aspetti Teorico-pratici per la Costruzione di Indici dei Prezzi al Consumo*

8/2004 – Maria Frustaci – *Glossario economico-statistico multilingua*

9/2004 – Giovanni Seri e Maurizio Lucarelli – *"Il Laboratorio per l'analisi dei dati elementari (ADELE): monitoraggio dell'attività dal 1999 al 2004"*

10/2004 – Alessandra Nuccitelli, Francesco Bosio e Luciano Fioriti – *L'applicazione RECLINK per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*

1/2005 – Francesco Cuccia, Simone De Angelis, Antonio Laureti Palma, Stefania Macchia, Simona Mastroluca e Domenico Perrone – *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*

2/2005 – Marina Peci – *La statistica per i Comuni: sviluppo e prospettive del progetto Sisco.T (Servizio Informativo Statistico Comunale. Tavole)*

3/2005 – Massimiliano Renzetti e Annamaria Urbano – *Sistema Informativo sulla Giustizia: strumenti di gestione e manutenzione*

4/2005 – Marco Broccoli, Roberto Di Giuseppe e Daniela Pagliuca – *Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo Microstrat di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat*

5/2005 – Mauro Albani e Francesca Pagliara – *La ristrutturazione della rilevazione Istat sulla criminalità minorile*

6/2005 – Francesco Altarocca e Gaetano Sberno – *Progettazione e sviluppo di un "Catalogo dei File Grezzi con meta-dati di base" (CFG) in tecnologia Web*

* ultimi cinque anni

7/2005 – Salvatore F. Allegra e Barbara Baldazzi – *Data editing and quality of daily diaries in the Italian Time Use Survey*

8/2005 – Alessandra Capobianchi – *Alcune esperienze in ambito internazionale per l'accesso ai dati elementari*

9/2005 – Francesco Rizzo, Laura Vignola, Dario Camol e Mauro Bianchi – *Il progetto "banca dati della diffusione congiunturale"*

10/2005 – Ennio Fortunato e Nadia Mignolli – *I sistemi informativi Istat per la diffusione via web*

11/2005 – Ennio Fortunato e Nadia Mignolli – *Sistemi di indicatori per l'attività di governo: l'offerta informativa dell'Istat*

12/2005 – Carlo De Gregorio e Stefania Fatello – *L'indice dei prezzi al consumo dei testi scolastici nel 2004*

13/2005 – Francesco Rizzo e Laura Vignola – *RSS: uno standard per diffondere informazioni*

14/2005 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Launching and implementing the job vacancy statistics*

15/2005 – Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni e Leonardo Tininini – *La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT*

16/2005 – Ennio Fortunato e Nadia Mignolli – *Verso il Sistema di Indicatori Territoriali: rilevazione e analisi della produzione Istat*

17/2005 – Raffaella Cianchetta e Daniela Pagliuca – *Soluzioni Open Source per il software generalizzato in Istat: il caso di PHPSurveyor*

18/2005 – Gianluca Giuliani e Barbara Boschetto – *Gli indicatori di qualità dell'Indagine continua sulle Forze di Lavoro dell'Istat*

19/2005 – Rossana Balestrino, Franco Garritano, Carlo Cipriano e Luciano Fanfoni – *Metodi e aspetti tecnologici di raccolta dei dati sulle imprese*

1/2006 – Roberta Roncati – *www.istat.it (versione 3.0) Il nuovo piano di navigazione*

2/2006 – Maura Seri e Annamaria Urbano – *Sistema Informativo Territoriale sulla Giustizia: la sezione sui confronti internazionali*

3/2006 – Giovanna Brancato, Riccardo Carbini e Concetta Pellegrini – *SIQual: il sistema informativo sulla qualità per gli utenti esterni*

4/2006 – Concetta Pellegrini – *Soluzioni tecnologiche a supporto dello sviluppo di sistemi informativi sulla qualità: l'esperienza SIDI*

5/2006 – Maurizio Lucarelli – *Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica*

6/2006 – Natale Renato Fazio – *La ricostruzione storica delle statistiche del commercio con l'estero per gli anni 1970-1990*

7/2006 – Emilia D'Acunto – *L'evoluzione delle statistiche ufficiali sugli indici dei prezzi al consumo*

8/2006 – Ugo Guarnera, Orietta Luzi e Stefano Salvi – *Indagine struttura e produzioni delle aziende agricole: la nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti*

9/2006 – Maurizio Lucarelli – *La regionalizzazione del Laboratorio ADELE: un'ipotesi di sistema distribuito per l'accesso ai dati elementari*

10/2006 – Alessandra Bugio, Claudia De Vitiis, Stefano Falorsi, Lidia Gargiulo, Emilio Gianicolo e Alessandro Pallara – *La stima di indicatori per domini sub-regionali con i dati dell'indagine: condizioni di salute e ricorso ai servizi sanitari*

11/2006 – Sonia Vittozzi, Paola Giacchè, Achille Zuchegna, Piero Crivelli, Patrizia Collesi, Valerio Tiberi, Alexia Sasso, Maurizio Bonsignori, Giuseppe Stassi e Giovanni A. Barbieri – *Progetto di articolazione della produzione editoriale in collane e settori*

12/2006 – Alessandra Coli, Francesca Tartamella, Giuseppe Sacco, Ivan Faiella, Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Isabella Siciliani, Sara Colombini e Alessandra Masi – *La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*

13/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Intrastat*

14/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Extrastat*

15/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: comparazione tra rilevazione Intrastat ed Extrastat*

16/2006 – Fabio M. Rapiti – *Short term statistics quality Reporting: the LCI National Quality Report 2004*

17/2006 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zeli – *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI*

18/2006 – Mauro Albani – *La nuova procedura per il trattamento dei dati dell'indagine Istat sulla criminalità*

19/2006 – Alessandra Capobianchi – *Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa*

20/2006 – Francesco Altarocca – *Gli strumenti informatici nella raccolta dei dati di indagini statistiche: il caso della Rilevazione sperimentale delle tecnologie informatiche e della comunicazione nelle Pubbliche Amministrazioni locali*

1/2007 – Giuseppe Stassi – *La politica editoriale dell'Istat nel periodo 1996-2004: collane, settori, modalità di diffusione*

2/2007 – Daniela Ichim – *Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment*

3/2007 – Ugo Guarnera, Orietta Luzi e Irene Tommasi – *La nuova procedura di controllo e correzione degli errori e delle mancate risposte parziali nell'indagine sui Risultati Economici delle Aziende Agricole (REA)*

4/2007 – Vincenzo Spinelli – *Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770*

5/2007 – Anna Di Carlo, Maria Picci, Laura Posta, Michaela Raffone, Giuseppe Stassi e Fiorella Tortora – *La progettazione dei Censimenti generali 2010-2011: 1 - Analisi, valutazione e proposte in merito ad atti di normazione e finanziamento*

6/2007 – Silvia Bruzzone, Antonia Manzari, Marilena Pappagallo e Alessandra Reale – *Indagine sulle Cause di Morte: Nuova procedura automatica per il controllo e la correzione delle variabili demo-sociali*

7/2007 – Maura Giacummo, Carlo Vaccari e Monica Scannapieco – *Indagine sulle Scelte Tecnologiche degli Istituti Nazionali di Statistica*

8/2007 – Lamberto Pizzicannella – *Sviluppo del processo di acquisizione e trattamento informatico degli archivi relativi al modello di dichiarazione 770. Anni 2004 – 2005*

9/2007 – Damiano Abbatini, Lorenzo Cassata, Fabrizio Martire, Alessandra Reale, Giuseppina Ruocco e Donatella Zindato – *La progettazione dei Censimenti generali 2010-2011 2 - Analisi comparativa di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani*

10/2007 – Marco Fortini, Gerardo Gallo, Evelina Paluzzi, Alessandra Reale e Angela Silvestrini – *La progettazione dei censimenti generali 2010 – 2011 3 – Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento*

11/2007 – Domenico Adamo, Damiana Cardoni, Valeria Greco, Silvia Montecolle, Sante Orsini, Alessandro Ortenzi e Miria Savioli – *Strategie di correzione del questionario sulla qualità della vita dell'infanzia e dell'adolescenza. Indagine multiscopo sulle famiglie. Aspetti della vita quotidiana 2005*