Università degli studi di Modena e Reggio Emilia
Dipartimento di Economia Politica

CAPP Centro di Analisi delle Politiche Pubbliche

\\ xxx \\

Il disegno dell'indagine sulle condizioni economiche e sociali delle famiglie nella Provincia di Modena

di

Michele Lalla

Materiali di discussione

Università degli Studi di Modena e Reggio Emilia
Dipartimento di Economia Politica
Via Jacopo Berengario 51
41100 Modena (Italia)
e-mail: lalla@unimo.it

Lavoro svolto nell'àmbito del progetto di ricerca

«*Costruzione di un'indagine sulle famiglie e di un modello di microsimulazione per l'analisi delle politiche sociali e fiscali a livello locale*»

cofinziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR).

Assegnazione: Anno 2001 – prot. 2001135524.

Coordinatore: Paolo Bosi

1. Introduzione

Negli anni recenti, accanto all'affinamento delle indagini nazionali (sui bilanci delle famiglie condotta dalla Banca d'Italia e sui consumi delle famiglie condotta dall'Istat), si è manifestato un interesse crescente per indagini di dimensioni minori, che diano una rappresentazione più precisa delle specifiche realtà locali, al fine di affinare le politiche di intervento o di esaminare con maggiore dettaglio gli effetti dei provvedimenti adottati dalle Pubbliche Amministrazioni. Le indagini "locali" possono aiutare a individuare sia una funzione di controllo e di valutazione delle strategie politiche (economiche e/o sociali) applicate, sia una verifica dell'attendibilità dei risultati derivati da indagini condotte a livello nazionale che hanno una precisione spesso assai insoddisfacente. Tali considerazioni costituiscono il presupposto della proposta di condurre un'indagine sui bilanci delle famiglie nella Provincia e nel Comune di Modena per la valutazione degli effetti delle politiche sociali e fiscali.

Gli obiettivi dell'indagine sono molteplici, ma ai fini del campionamento si possono limitare all'analisi della distribuzione del reddito e dei servizi sociali, e alla costruzione di un modello di microsimulazione (*tax-benefit model*) statico che consenta di valutare gli effetti di politiche fiscali e sociali redistributive nella Provincia e nel Comune di Modena. Le informazioni necessarie a tali fini sono di carattere economico, sociale, e demografico che si devono raccogliere tramite un'indagine (*survey*) specifica sulla situazione delle famiglie perché non esistono informazioni già disponibili per rispondere alle domande poste dagli obiettivi. Il campione da costituire deve essere, quindi, in grado: di rappresentare la distribuzione del reddito, del risparmio, e degli investimenti; di fornire il supporto informativo per la costruzione di un modello di microsimulazione; di accertare alcuni aspetti della domanda dei servizi pubblici offerti all'infanzia, agli anziani, e ai disabili; di rilevare le condizioni di salute e l'uso del tempo libero.

La rilevazione dei dati avviene (è avvenuta) tramite intervista diretta presso le famiglie incluse nel campione. Gli intervistatori utilizzano un questionario, che deve essere (è stato) realizzato in modo da consentire anche la costruzione di una base di dati, contenente informazioni sia inerenti alla famiglia, sia ai suoi singoli componenti. Le domande inserite nel questionario accertano, pertanto, tutte le informazioni relative al reddito, al patrimonio, alle caratteristiche sociali e demografiche; infatti, si deve essere in grado di ricostruire per l'individuo (e le famiglie) i carichi fiscali e i benefici derivanti da un insieme predefinito di istituti fiscali e di programmi di spesa nazionale e locale. Tra quelli nazionali si possono ricordare: l'IRPEF, gli assegni pensionistici di varia natura, gli assegni al nucleo famigliare, le imposte indirette nazionali. Tra gli istituti locali si ricordano: l'ICI; le tariffe relative alle forniture dei principali beni di utilità pubblica (luce, acqua, gas); la tassa per la raccolta dei rifiuti urbani; le tasse per la scuola materna, gli asili nido, le strutture protette per anziani e l'assistenza domiciliare, le forme di minimo vitale, e cosí via.

Sulla base dei dati raccolti si realizzerà la programmazione del modello di microsimulazione, che necessita di fonti informative costituite tramite indagini dirette a rilevare i dati individuali. Le indagini analoghe condotte a livello nazionale sono: l'indagine (quotidiana, riportata all'anno) sui consumi delle famiglie e l'indagine multiscopo sulle famiglie condotte dall'Istat (2002a,b,c); l'indagine biennale sui bilanci delle famiglie condotta dalla Banca d'Italia. L'indagine sui consumi condotta dall'Istat ha subíto diversi cambiamenti nel tempo per affrontare le varie difficoltà che si presenta-

vano nella rilevazione e nell'analisi (Falorsi, Russo, 1992; Filippucci, Marliani, 1992; De Vitiis, Falorsi, 2000) e è ancora oggetto di riflessioni (Barcherini, Calia, Filipucci, Grassi, 2002); inoltre, dal 1994 è stata introdotta una indagine longitudinale sulle famiglie estesa a tutti i paesi dell'Unione europea e coordinata dall'EUROSTAT, l'Ufficio di statistica dell'Unione europea (Istat, 2002d). L'indagine sui consumi delle famiglie rileva anche il reddito, ma non in forme dettagliate e accurate sicché il legame esistente tra reddito e consumo non può essere analizzato compiutamente. La Banca d'Italia conduce, invece, una indagine mirata a rilevare con precisione le varie tipologie di reddito, risparmio, e investimenti (Cannari, Gavosto, 1994; Brandolini, Cannari, 1994; Brandolini, 1999), ma il consumo rimane pressoché irrilevante. Queste fonti hanno permesso di effettuare importanti analisi del comportamento dei consumatori e delle famiglie e di verificare empiricamente teorie fondate sul comportamento del singolo agente; tuttavia, presentano una frattura concettuale perché l'indagine della Banca d'Italia rileva con più accuratezza il reddito e il patrimonio delle famiglie e non il consumo, mentre l'indagine dell'Istat rileva con più accuratezza il consumo delle famiglie e con molta approssimazione il reddito. I limiti menzionati non hanno limitato del tutto gli studi; infatti, i modelli di microsimulazione si sono rivelati utili per valutare, nelle applicazioni di politica economica, l'impatto distributivo di riforme fiscali e tariffarie (modificazioni delle imposte sul reddito e indirette, dei benefici dello stato sociale, e così via).

L'indagine dovrebbe acquisire informazioni utili sotto il profilo sia dell'analisi dei comportamenti di consumo e dell'offerta di lavoro, sia delle applicazioni di politica economica per la valutazione delle politiche sociali. Il *primo* profilo è di particolare interesse perché consentirebbe di disporre di una importante base di dati per la ricerca, utilizzabile anche da altri ricercatori di scienze sociali. Tuttavia, almeno nell'accezione più comune, l'indagine sui comportamenti di consumo, in un solo periodo dell'anno, è pressoché irrealizzabile perché richiede la rilevazione giornaliera delle spese: i costi sarebbero elevati e la strategia di rilevazione assai complessa. L'indagine può stimolare, sí, un alto interesse metodologico perché rileverebbe congiuntamente consumo e reddito, ma si può concludere a priori che il rapporto costo/prestazione non è accettabile. Il consumo sarà rilevato inevitabilmente con molta approssimazione e, per gli obiettivi fissati per l'indagine, si rileveranno dati che presenteranno le stesse limitazioni dell'indagine della Banca d'Italia: reddito e il patrimonio saranno accurati, ma la spesa per i consumi sarà sottostimata. Il *secondo* profilo è di interesse non solo per la ricerca, ma anche per il Comune, che potrebbe, partendo dai risultati conseguiti, impostare programmi di lavoro per la realizzazione di uno strumento assolutamente innovativo di monitoraggio e valutazione dei vantaggi e dei costi a livello famigliare e/o individuale delle politiche sociali, realizzando così un vero e proprio "laboratorio" delle politiche sociali locali. Si osserva, tuttavia, che l'azione delle politiche sociali potrebbe richiedere una valutazione molto più accurata per accertarne l'efficienza e l'efficacia; pertanto, una indagine un po' più generale non ha i requisiti della specificità. La metodologia che ne deriva, combinata con l'uso di dati amministrativi, potrebbe fornire una pregevole base di partenza per applicazioni concrete che abbiano effetto (di ritorno) sulla valutazione delle azioni intraprese e da intraprendere.

La costruzione di un campione per conseguire gli obiettivi dell'indagine richiede di possedere una buona lista (*frame*) della popolazione di riferimento o obiettivo (*target*), ossia priva di difetti vari relativi alle unità statistiche: incompletezza, sopracompletezza, ridondanza, inesistenza, inattualità, imprecisioni (Cicchitelli, Herzel, Montanari, 1997). Il piano di campionamento si potrebbe progettare con piú efficacia,

se fosse possibile avere informazioni relative alle unità statistiche della popolazione, utili anche per gli obiettivi dell'indagine. Le basi di dati di origine amministrativa sono utili per determinare la lista, anche se non sono esenti da problemi (Martini, Aimetti, 1989; Martini, 1990), specifici per ogni tipo ente che li produce e per ogni tipo di indagine (Abbate, Baldassarini, 1994; Cannari, Pellegrino, Sestito, 1996; Lucifora, 1995). L'accesso alla banca dati di origine fiscale sarebbe ideale per costruire un campione con l'obiettivo di indagare la distribuzione del reddito, del risparmio, e degli investimenti; tuttavia, per motivi di riservatezza è pressoché impossibile accedervi (v. *infra*) e occorre procedere senza informazioni specifiche sulle unità statistiche da selezionare.

La struttura del lavoro è la seguente. Nel paragrafo 2 si illustrano alcuni aspetti del piano di campionamento: dimensione campionaria e stratificazione. Nel paragrafo 3 si riportano alcune considerazioni sugli esiti della rilevazione. Nel paragrafo 4 si espongono i procedimenti adottati per determinare i fattori di riporto alla popolazione obiettivo e le varianze degli stimatori di interesse. Nel paragrafo 5 si riassumono le caratteristiche degli errori non campionari, in generale e in particolare per l'indagine corrente. Le conclusioni seguono, infine, nel paragrafo 6 con un breve cenno sulle eventuali repliche dell'indagine negli anni futuri, le quali potrebbero dare origine a un panel; l'utilità del panel per l'analisi di aspetti di dinamica sociale sarebbe elevata.

2. Piano di campionamento

Il piano di campionamento descritto valuta: il numero di unità statistiche (dimensione) da selezionare dalla popolazione di riferimento, che sia idoneo a soddisfare gli obiettivi dell'indagine (§2.1); la strategia di campionamento più efficace rispetto alla base campionaria disponibile e alle informazioni relative alla popolazione di riferimento, che si può utilizzare nella costruzione del campione (§2.2). In particolare, si è scelta una strategia a due stadi: le Unità di Primo Stadio (UPS) sono i Comuni della Provincia di Modena; le Unità di Secondo Stadio (USS) sono le famiglie, che costituiscono proprio l'oggetto dell'indagine e alle quali ci si riferirà con il termine «unità statistiche». Per il Comune di Modena (§2.3) si è previsto un campione con una dimensione piú elevata, rispetto agli altri e una strategia diversa. Le strategie alternative sono limitate (§2.4).

2.1. Dimensione campionaria

Si supponga che almeno una variabile da stimare, Y, sia nota; allora, Y rappresenta una caratteristica ideale per la stratificazione (Cochran, 1977: p.101); inoltre, la stratificazione cosí ottenuta permette di migliorare le stime dei parametri di tutte le altre grandezze che sono correlate con essa (Cochran, 1977; Cicchitelli, Herzel, Montanari, 1992). Nel caso in cui Y sia una variabile continua, la valutazione della dimensione del campione si ottiene dalla seguente relazione (Cochran, 1977)

$$n_e = \frac{\dfrac{z_{1-\alpha/2}^2 S^2}{r^2 \overline{Y}^2}}{1 + \dfrac{1}{N}\left(\dfrac{z_{1-\alpha/2}^2 S^2}{r^2 \overline{Y}^2} - 1\right)}, \qquad (1)$$

dove S^2 indica la varianza (non corretta) della Y, $\overline{Y}$ la media, N la dimensione della popolazione obiettivo, r l'errore relativo (percentuale) che si commette nella stima dei

parametri (media o totale) della Y, $z_{1-\alpha/2}$ l'ascissa della curva normale in cui la funzione di ripartizione vale $(1-\alpha/2)$ e α rappresenta il livello di significatività desiderato per le stime che si ottengono dal campione, n_e indica la dimensione del campione risultante dalla precisione desiderata delle stime. Qui e oltre, le grandezze indicate con le lettere maiuscole si riferiscono alla popolazione obiettivo, mentre le grandezze indicate con le lettere minuscole si riferiscono al campione selezionato e osservato; inoltre, il valore del livello di significatività α si può fissare pari al 5% per cui il valore di $z_{1-\alpha/2}$ è uguale a 1,96 e si può approssimare a 2. Infine, si noti che il denominatore esprime l'effetto della correzione per popolazioni finite; pertanto, occorre conoscere la dimensione della popolazione di riferimento.

Nel caso in oggetto, una caratteristica adeguata è il reddito delle famiglie o il risparmio o il patrimonio, ma non si conosce alcuna variabile rilevante da stimare. Si suppone allora di fissare l'errore sulla stima di una proporzione, P, della modalità di una data variabile qualitativa. La valutazione della dimensione del campione si ottiene dalla seguente relazione (Cochran, 1977)

$$n_e = \frac{\dfrac{z_{1-\alpha/2}^2 P(1-P)}{e^2}}{1+\dfrac{1}{N}\left(\dfrac{z_{1-\alpha/2}^2 P(1-P)}{e^2}-1\right)}, \tag{2}$$

dove e l'errore (assoluto) che si commette nella stima della proporzione P, $z_{1-\alpha/2}$ l'ascissa della curva normale in cui la funzione di ripartizione vale $(1-\alpha/2)$ e α rappresenta il livello di significatività desiderato per le stime che si ottengono dal campione.

Sia m la dimensione del campione ottenuto dall'indagine; a causa delle mancate risposte o partecipazioni m può risultare inferiore a n_e. I fallimenti nelle interviste sono sempre negativi e possono causare distorsioni anche rilevanti nelle stime. Nell'ipotesi che i dati mancanti si distribuiscano in modo casuale e siano incorrelati con le variabili oggetto di stima, si può rivalutare la precisione che fornisce il campione effettivo, ottenuto dalla rilevazione, calcolando: l'errore relativo r dalla relazione precedente

$$r = \frac{z_{1-\alpha/2} S}{\bar{Y}} \sqrt{\frac{1}{m}\left(\frac{N-m}{N-1}\right)}, \tag{3}$$

per la variabile continua Y; e l'errore (assoluto)

$$e = z_{1-\alpha/2} \sqrt{\frac{P(1-P)}{m}\left(\frac{N-m}{N-1}\right)}, \tag{4}$$

per le variabili dicotome.

La dimensione del campione dipende dalla precisione desiderata delle stime delle diverse variabili. Per ognuna di esse di ottiene un valore della dimensione, n_i, e la dimensione finale può essere data dal massimo tra le n_i, per $i = 1, \ldots, p$, dove p è il numero di caratteri considerati nella stima di n. Spesso la dimensione ottimale è in contrasto con le risorse finanziarie e umane disponibili e il valore si riduce per i vincoli di bilancio. Sia C l'ammontare delle risorse disponibili, sia C_0 il costo fisso da sostenere per condurre l'indagine, sia c_u il costo unitario di ogni intervista; allora il numero di unità statistiche che si possono includere nel campione, n_{costo}, è dato da

$$n_{\text{costo}} = \frac{C - C_0}{c_u} \le n_e . \tag{5}$$

La dimensione finale sarà data dal minimo delle due dimensioni ottenute:

$$n \equiv n_{\text{finale}} = \min(n_e, n_{\text{costo}}). \tag{6}$$

2.1.1. Valutazione della dimensione totale del campione

In assenza di qualunque informazione sulla popolazione di riferimento, com'è nel caso in oggetto, si può adottare l'espressione (2) per valutare la dimensione del campione perché, tramite essa, si fissa la precisione di una proporzione, *P*, relativa a una variabile dicotoma o a una modalità di una variabile qualitativa (rispetto alle altre modalità): la dimensione campionaria massima si ha con *P*=1/2. In base alle risorse disponibili, la dimensione n non dovrebbe superare 1500 unità statistiche (famiglie). La scelta ragionevole dei parametri che si possono, ora, considerare "fissi" è: *P*=0,5; e un livello di confidenza del 95% (che comporta un valore di $z_{1-a/2} \cong 2$). La dimensione risulta, allora, una funzione dell'errore desiderato: cosí con *e*=0,030 si ha n=1094 quando si assume che il numero di unità statistiche (le famiglie) della popolazione di riferimento sia *N*=252.968 al 31 dicembre 2000 (senza la correzione per popolazioni finite, n=1112); con *e*=0,025 si ha n=1590. Tali valori non sono alti per le risorse disponibili o per le precisioni desiderate: la dimensione effettiva può essere intermedia tra i due. La precisione delle stime per il Comune di Modena dovrebbe essere, inoltre, equiparabile a quella provinciale. Allora, a livello provinciale, si è fissato l'errore assoluto uguale al 3,1% (*e*=0,031), mantenendo costanti gli altri parametri dell'espressione (2), e si è ottenuta una dimensione campionaria pari a n=1040. A livello del Comune di Modena, invece, si è fissato l'errore assoluto uguale al 4% (*e*=0,040) che dà una dimensione campionaria pari a n_{MO}=620; un errore del 3,1% avrebbe condotto a un valore elevato per le risorse disponibili.

Si deve notare che l'errore assoluto è lo stesso per ogni valore *P* della popolazione di riferimento sicché la dimensione così ottenuta non garantisce la precisione adeguata per le proporzioni piccole; per esempio, inferiori al 10% (Cochran, 1977). Per migliorare la valutazione della dimensione del campione in base alla precisione desiderata delle stime, si considera che il carattere oggetto di stima è dicotomo e che si può rappresentare con una distribuzione bernoulliana. Allora, si adotta l'espressione (1) per valutare la dimensione adatta per stimare la proporzione di un carattere raro, ricordando che per la distribuzione bernoulliana: il valore atteso (media) è *P*, e la varianza è *P*(1–*P*). L'espressione diventa:

$$n_e = \frac{\dfrac{z_{1-\alpha/2}^2 (1-P)}{r^2 P}}{1 + \dfrac{1}{N}\left(\dfrac{z_{1-\alpha/2}^2 (1-P)}{r^2 P} - 1\right)}. \tag{7}$$

Diversamente, si fissa l'errore relativo sulla proporzione *P*; allora, l'errore assoluto è dato da $e = rP$ e, sostituendolo nella (2), si ottiene l'espressione (7). Si può mantenere, quindi, costante l'errore relativo rispetto a *P*; nel caso *P*=0,5 e un errore *e*=0,05 si ha un errore relativo del 10%; infatti, *r*=*e*/*P*. Ne consegue che l'errore relativo è uguale: a 0,04 per *P*=0,4; a 0,03 per *P*=0,3; a 0,02 per *P*=0,2; e cosí via. Analogamente varierà la dimensione del campione: *n*=600, *n*=933, *n*=1600; e cosí via — i valori sono stati ottenuti ignorando la correzione per popolazione finita, ossia considerando solo il numeratore della (7). Per una proporzione *P*=0,1 si ottiene una dimensione *n*=3600 e per *P*=0,05 si

ottiene una dimensione n=7600; si veda Fabbris (1989, pp. 61-64). Si noti che certi aspetti della povertà o delle politiche sociali potrebbero appartenere alla classe di percentuali inferiori al 10%; tuttavia, i costi pongono un limite massimo alla precisione desiderata delle stime. Per conoscere tali aspetti, con una precisione elevata o una conoscenza più dettagliata, si può ricorrere eventualmente a una indagine focalizzata.

2.2. Campionamento stratificato

La procedura di stratificazione realizza il raggruppamento delle unità statistiche, secondo strati che sono «omogenei» rispetto a certe caratteristiche; ciò consente di migliorare l'efficienza delle stime e la prestazione complessiva del campione. La scelta della stratificazione è condizionata, però, dalle informazioni relative alla popolazione di riferimento disponibili nella fase iniziale che, allo stato attuale, sono assai ridotte per l'indagine in oggetto. Il primo carattere considerato ha riguardato la suddivisione geografica del territorio secondo i distretti sociosanitari (§2.2.1).

L'uso appropriato delle informazioni utili, però, richiede: (a) una elaborazione per conoscere la struttura della popolazione di riferimento e per poter progettare la consistenza del campione per strato; (b) una elaborazione successiva per l'estrazione delle famiglie da includere nel campione. Allora, le difficoltà operative e di accesso possono risultare proibitive perché bisogna ricorrere ai dati dell'Anagrafe che sono protetti dalla legge sulla riservatezza dei dati (Legge n. 675/96, G.U. n. 5 dell'8 gennaio 1997). Sia le informazioni (aggregate) sulla struttura della popolazione rispetto a determinati caratteri per la fase (a), sia i dati individuali delle USS per la fase (b) si possono richiedere alle persone autorizzate all'acceso alle basi di dati. Si dipenderebbe, però, da altri per ottenere le informazioni desiderate e il tempo di esecuzione delle operazioni potrebbe diventare eccessivamente lungo perché si usufruirebbe del lavoro di personale adibito a altri incarichi. L'estrazione della lista delle famiglie campione con indirizzo, e tutti i possibili recapiti (anche telefonici), rimane un problema delicato, rispetto alla legge sulla riservatezza. Se l'Anagrafe non è automatizzata, allora l'operazione non si può eseguire in pratica; e la maggior parte dei Comuni piccoli non l'ha ancora. In ogni caso, è emerso che non era conveniente condurre tale stratificazione per tutti i Comuni, eccetto Modena per il quale le USS (le famiglie) sono state stratificate secondo la loro ampiezza (§2.2.2), l'età del capofamiglia (§2.2.3), e il genere del capofamiglia (§2.2.4). L'allocazione del campione tra gli strati è nel paragrafo 2.3.

Gli altri caratteri di interesse sono: la tipologia famigliare, l'accesso alla banca dati di origine fiscale, e il consumo di alcuni beni pubblici (§2.4). La loro importanza non coinvolge solo la realizzazione di strategie di campionamento migliori e piú efficienti; ma anche l'elaborazione dei dati: sia per una eventuale post-stratificazione, sia per i possibili riscontri sui risultati ottenuti. Il processo di allocazione adottato è stato proporzionale alla numerosità della popolazione di riferimento nelle celle per la scarsità di informazioni quantitative disponibili al momento della definizione del piano.

2.2.1. Stratificazione sui distretti sociosanitari

La strategia di campionamento suddivide la Provincia di Modena in aree geografiche (macrostrati) secondo i distretti socio-sanitari (Zoda, 1998; Benassi, Zoda 2002), la denominazione dei quali è data dalla città piú rappresentativa, come riportato nella Tabella 1 (a pagina 11). Il distretto N.1, di Carpi (D1), contiene anche i Comuni di Campogal-

liano, Novi di Modena, e Soliera. Il distretto N.2, di Mirandola (D2), contiene anche i Comuni di Camposanto, Cavezzo, Concordia sulla Secchia, Finale Emilia, Medolla, San Felice sul Panaro, San Possidonio, e San Prospero. Il distretto N.3, di Modena (D3), non contiene altri Comuni. Il distretto N.4, di Sassuolo (D4), contiene anche i Comuni di Fiorano Modenese, Formigine, Frassinoro, Maranello, Montefiorino, Palagano, e Prignano sulla Secchia. Il distretto N.5, di Pavullo nel Frignano (D5), contiene anche i Comuni di Fanano, Fiumalbo, Lama Mocogno, Montecreto, Pievepelago, Polinago, Riolunato, Serramazzoni, e Sestola. Il distretto N.6, di Vignola (D6), contiene anche i Comuni di Castelnuovo Rangone, Castelvetro, Guiglia, Marano sul Panaro, Montese, Savignano sul Panaro, Spilamberto, e Zocca. Il distretto N.7, di Castelfranco Emilia (D7), contiene anche i Comuni di Bastiglia, Bomporto, Nonantola, Ravarino, e San Cesario sul Panaro.

L'Unità statistica di Primo Stadio (UPS) è costituita, come si è detto, dal Comune; mentre l'Unità statistica di Secondo Stadio (USS) è costituita dalle famiglie, come in molte altre indagini condotte dall'Istat —sui consumi delle famiglie (Falorsi, Falorsi, Russo, 1992; De Vitiis, Falorsi, 2000) o sulle forze di lavoro (Di Pietro, 1993; Barcaroli, Di Pietro, Venturi, 1993) — e dalla Banca d'Italia (2000, 2002) sui bilanci delle famiglie. Le UPS sono state raggruppate in due categorie o strati: AutoRappresentative (AR), corrispondenti ai Comuni che *denominano* i distretti; e Non AutoRappresentative (NAR), tutti gli altri. I Comuni AR superano la soglia di 20000 residenti, con l'eccezione di Pavullo nel Frignano; mentre i Comuni NAR hanno un numero di residenti inferiori alla soglia, con l'eccezione di Formigine; si noti che tale soglia costituisce un estremo di classe anche nella stratificazione dei Comuni operata dalla Banca d'Italia (Brandolini, Cannari, 1994; Cannari, Gavosto, 1994). Nella Tabella 1 si mostra una ripartizione della dimensione campionaria provinciale, n=1040, proporzionale alla numerosità (frequenze) di USS per ogni UPS, n_{dc}, come se fossero tutte AR. I valori di n_{dc} sono stati arrotondati tutti per eccesso e ciò ha generato un lieve aumento della dimensione totale, che è passata da n=1040 a n=1062. Si è eseguita anche una ripartizione secondo la numerosità della popolazione residente, $\wp$, perché i fenomeni da indagare sono correlati anche a questa grandezza che, indirettamente, coglie la dimensione delle USS; ma, come si può osservare nella Tabella 1, le variazioni non sono rilevanti per l'omogeneità della struttura demografica delle famiglie nel territorio sicché si è proceduto nel séguito considerando prevalentemente la numerosità delle USS.

La determinazione del numero dei Comuni NAR per ogni distretto è stata eseguita in base al numero di USS per distretto, considerando la mediana della dimensione dei comuni AR arrotondata per eccesso a un numero pari: il Comune che corrisponde alla mediana è Castelfranco Emilia con 9903 USS. Allora, si è assegnato a ciascun distretto un Comune NAR ogni 10000 USS. In termini formali

$$n_{NAR;d} = \left\lfloor \left(\frac{1}{10000} \sum_{c=1}^{C_d^*} N_{dc} \right) + 1 \right\rfloor \tag{8}$$

dove $n_{NAR;d}$ è il numero di NAR da selezionare nel *d*-esimo distretto, C_d^* è il numero totale di NAR nel *d*-esimo distretto per il quale si ha $C_d^* = C_d - 1$ (dove C_d è il numero totale di Comuni), N_{dc} è il numero di USS del *c*-esimo NAR del *d*-esimo distretto, il simbolo $\lfloor \cdot \rfloor$ indica la parte intera dell'argomento. Le UPS da includere nei distretti sono state determinate con una generazione di numeri casuali proporzionali alla loro dimen-

sione N_{dc} (*Probability Proportional to Size* o PPS), ossia al numero di famiglie residenti, perché: fornisce una media campionaria non distorta, e non è soggetta all'inflazione della varianza (Hansen, Hurwitz, 1943; Cochran, 1977, p. 295). La dimensione campionaria provinciale, n=1040, nel primo passo, è stata ripartita proporzionalmente tra i vari distretti secondo la corrispondente numerosità di USS, N_d, ottenendo la dimensione campionaria per distretto, n_d. Nel secondo passo, la dimensione n_d è stata ripartita proporzionalmente tra le UPS campionarie del *d*-esimo strato, ottenendo le n_{dc}, per mantenere un certo equilibrio tra le numerosità delle UPS campionarie a livello distrettuale. I risultati della selezione dei Comuni sono esposti nella Tabella 2 (a pagina 12), dove l'approssimazione nel calcolo delle n_{dc} è stata eseguita sempre per eccesso e ciò ha generato una piccola differenza tra i totali parziali e il totale complessivo della colonna n_{dc}, rispetto a quelli della colonna n_d. I Comuni AR sono: Carpi, Mirandola, Modena, Sassuolo, Pavullo nel Frignano, Vignola, e Castelfranco Emilia. I Comuni NAR inclusi nel campione sono: Soliera (D1); Cavezzo, Finale Emilia (D2); Formigine, Maranello, Montefiorino (D4); Polinago (D5); Spilamberto, Guiglia (D6); Nonantola (D7).

Nel Comune di Modena, per il quale si voleva una precisione circa uguale a quella provinciale, occorreva un campione aggiuntivo di 313 USS per raggiungere la dimensione fissata a n_{MO}=620. Per migliorare ancora la precisione delle stime relative al Comune di Modena, si è operata una stratificazione delle USS in base a caratteri specifici delle USS: l'ampiezza della famiglia, l'età e il genere del capofamiglia.

2.2.2. Stratificazione sull'ampiezza delle famiglie

La stratificazione sull'ampiezza della famiglia è una scelta adeguata perché sembra correlata con gli obiettivi dello studio, distribuzione del reddito e valutazione degli effetti delle politiche fiscali e sociali; infatti, le famiglie più numerose potrebbero essere quelle più bisognose e soggette alle conseguenze delle politiche intraprese o la presenza di piú percettori di reddito influenza l'ammontare complessivo del reddito disponibile. Le famiglie con uno o due membri potrebbero costituire anche l'oggetto di interessi particolari dell'indagine quando sono anziani.

L'ampiezza delle famiglie è stata suddivisa in *I*=4 classi, come si può osservare nella distribuzione marginale (delle righe) della Tabella 3 (a pagina 12), cercando di conciliare il significato intrinseco di ogni classe con una frequenza dell'ordine di circa 1/*J*: famiglie con un solo membro, con due membri, con tre membri, con quattro o più membri. Si noti che la data di riferimento per il Comune di Modena, in Tabella 3 e seguenti (nelle pagine 12 e 13), è l'anno successivo a quello della provincia perché il campione è stato progettato alla fine del 2001; allora, erano disponibili i dati provinciali aggiornati al 31/12/2000. I dati comunali ci sono giunti all'inizio del 2002 e, quindi, erano aggiornati al 31/12/2001. Per la stabilità della popolazione nel tempo, non si alterano in modo sensibile i risultati relativi alle dimensioni campionarie e alle stime.

2.2.3. Stratificazione sull'età del capofamiglia

L'età del capofamiglia rappresenta un altro elemento discriminatorio tra gruppi diversi di famiglie; per esempio, i nuclei piccoli con capofamiglia anziano possono trovarsi in condizioni difficili e rappresentare gruppi caratterizzati rispetto alle aree oggetto di indagine. Si potrebbero fissare classi anche molto ampie del tipo: fino a 29 anni, da 30 a 49 anni, da 50 a 64 anni, da 65 a 74 anni, da 75 in avanti; ma si è optato per una suddi-

visione in cinque classi, $J = 5$, come si può osservare in Tabella 3 (a pagina 12): fino a 34 anni, da 35 a 49 anni, da 50 a 64 anni, da 65 a 74 anni, da 75 in avanti. Le classi sono state formate considerando sia i punti di suddivisione tradizionali (di cinque in cinque), sia la possibilità di avere classi con una numerosità circa uguale, sia l'opportunità di una aggregazione piú «fine» nell'età successiva al ritiro dal mondo del lavoro. La prima classe ha un numero di unità statistiche pari a circa la metà di quelle che sono nella seconda e nella terza classe, che hanno circa la stessa numerosità e quasi suddividono il periodo lavorativo di un soggetto in due; specie per quelli che entrano nel mondo del lavoro in ritardo. La quarta e la quinta classe suddividono in due parti il periodo di ritiro dal lavoro e presentano una numerosità pressoché comparabile tra loro, ma pari a poco piú della metà di quella della seconda e della terza classe di età.

2.2.4. Stratificazione sul genere del capofamiglia

La stratificazione sul genere del capofamiglia, K=2, è conveniente perché consente di migliorare la rappresentatività, nel campione, di segmenti di popolazione che possono avere problemi e comportamenti particolari; per esempio, i giovani che formano una famiglia con un solo componente (*single*) e gli anziani. Per questi caratteri si consegue, cosí, un controllo sulle distribuzioni marginali del campione rispetto a quelle della popolazione di riferimento, con un certo beneficio per le stime.

2.3. Dimensione campionaria per strato nel Comune di Modena

L'allocazione ottimale di Neyman (Cochran, 1977), vincolata a un totale prefissato, è la strategia piú idonea quando si dispongono delle grandezze quantitative per strato. In loro assenza, com'è in questo caso, si è applicata una allocazione proporzionale che definisce la dimensione del campione nello strato in proporzione alla dimensione della popolazione di riferimento nello stesso strato:

$$n_{MO;ijk} = \left\lfloor n_{MO} \left(\frac{N_{MO;ijk}}{N_{MO}} \right) + 1 \right\rfloor \tag{9}$$

dove $n_{MO;ijk}$ è il numero di USS da selezionare nello strato *ijk* (*i*-esimo numero di componenti la famiglia, *j*-esima classe di età del capofamiglia, *k*-esimo valore del genere) del Comune di Modena, n_{MO} è l'ampiezza della dimensione campionaria nel Comune di Modena (620 famiglie), $N_{MO;ijk}$, è il numero di famiglie nello strato *ijk*, N_{MO} è il numero totale di famiglie (75748 famiglie), e il simbolo $\lfloor \cdot \rfloor$ indica la parte intera dell'argomento.

La determinazione della dimensione campionaria per strato, $n_{MO;ijk}$, è stata eseguita arrotondando il valore decimale ottenuto in ogni dominio di studio, sicché la dimensione del campione è diventata n_{MO}=637, come riportato in Tabella 5 (pagina 13).

Si è eseguito un sopracampionamento per sopperire alle eventuali mancate risposte. Per stabilire l'ammontare delle USS in aggiunta alla dimensione programmata, si può considerare il tasso di mancate partecipazioni in altre indagini simili, date le difficoltà nella rilevazione di dati inerenti a fenomeni complessi, come il consumo e il reddito. Il tasso finale di non risposta è: dell'ordine del 15% nell'indagine sui consumi delle famiglie condotta dall'Istat, dopo avere sostituito le famiglie non disponibili a partecipare (Lucev, 1992); dell'ordine del 60% nell'indagine sui bilanci delle famiglie con-

dotta dalla Banca d'Italia (2002, p. 31). La notevole differenza tra i due dati deriva, oltre che dall'obbligatorietà della partecipazione alle indagini condotte dall'Istat, almeno da due motivi: la sostituzione delle mancate partecipazioni nel calcolo e la difficoltà intrinseca nel rilevare dati inerenti al reddito (Quintano, Lucev, 1990). Da ciò si può arguire che la dimensione ipotizzata ottimale deve essere triplicata. A causa delle supposte difficoltà di relazione con gli uffici dell'anagrafe dei comuni, il numero di USS estratte è stato pari al quadruplo della dimensione del campione sopra determinata.

2.4. Stratificazioni alternative: tipologia famigliare, reddito, consumi

La tipologia famigliare disponibile presso gli uffici anagrafici distingue tra: persone sole; coppie coniugate; coppie coniugate e figli; coppie coniugate, figli, e altre persone; genitori e figli; genitori, figli e altre persone; altro tipo di famiglia. Tale carattere è simile al numero di componenti la famiglia anche se, come si nota dalla Tabella 4 (a pagina 13), vi è una differenziazione interessante tra le USS; tuttavia, per evitare una eccessiva proliferazione di strati nei quali indagare, si è deciso di semplificare lo schema di campionamento tralasciando tale informazione.

La caratteristica ideale di stratificazione per l'indagine in oggetto è senza dubbio il reddito delle famiglie che, però, non è disponibile. Tale indisponibilità è motivata sulla base della riservatezza; infatti, il Ministero delle finanze (o l'Ufficio distrettuale delle imposte dirette) possiede informazioni sufficienti (*banca dati fiscale*) per costruire un buon campione. Tali dati potrebbero essere forniti in una forma anonima per motivi di ricerca, ma con un codice identificativo che consentirebbe di ottenere gli indirizzi dopo avere eseguito l'estrazione del campione; tuttavia, per i piccoli Comuni esistono oggettivamente ostacoli legali e burocratici perché vi possono essere informazioni che rendono identificabili gli individui. Occorrerebbe trovare una forma di impegno o responsabilità legale del richiedente perché l'elaborazione della banca dati fiscale non è da sottovalutare: le informazioni sulle fonti del reddito sono sia dettagliate, sia abbastanza inesplorate; la loro precisione potrebbe superare le attese con notevole sorpresa dei critici dell'attendibilità e validità dei dati fiscali. Si nota, tuttavia, che l'uso di questa base campionaria potrebbe essere non idonea per l'indagine sulla povertà e sugli effetti delle politiche sociali perché la rilevazione sarebbe eseguita con minore precisione: sotto un certo reddito, i percettori non sono obbligati a effettuare la dichiarazione.

Altre fonti informative di interesse sono agli archivi di alcune aziende che distribuiscono beni di utilità pubblica (energia elettrica, acqua, e gas). L'abbinamento dei dati contenuti in questi archivi condurrebbe a costituire una base informativa notevole che consentirebbe sia di estrarre il campione, sia di controllare le informazioni rilevate. La metodologia e l'alta qualità dei dati rappresentano gli aspetti più rilevanti di un piano di campionamento, senza i quali i margini di errori sono assai rilevanti, soprattutto per i problemi intrinseci delle indagini sul campo che soffrono diverse difficoltà connesse alle mancate risposte totali e parziali, all'autoselezione dei rispondenti, al carico di lavoro cui sono sottoposti gli intervistati in una indagine sul reddito e sui bilanci delle famiglie (Martini, Aimetti, 1989; Martini, 1990).

Tabella 1 – Numero di famiglie (USS), numero di famiglie cumulate (USSC), dimensione campionaria proporzionale al numero di famiglie (n_{dc}), numero totale per distretto (n_d), numero di abitanti (P_d), numero di abitanti cumulati ($P_d C$), dimensione campionaria proporzionale al numero di abitanti ($n_{P;dc}$), e totali ($n_{P;d}$) per i Comuni della Provincia di Modena suddivisi per distretto sociosanitario(*) al 31/12/2000

	Comune	USS	USSC	n_{dc}	n_d	P_d	$P_d C$	$n_{P;dc}$	$n_{P;d}$
D1	**Carpi**	24674	36534	102	**150**	61631	92562	102	**152**
	Campogalliano	2992	2992	13		7671	7671	13	
	Novi di Modena	3940	6932	17		10358	18029	17	
	Soliera	4928	11860	21	**=153]**	12902	30931	22	**=154]**
D2	**Mirandola**	8711	30160	36	**124**	22077	78607	37	**129**
	Camposanto	1145	1145	5		3031	3031	5	
	Cavezzo	2549	3694	11		6716	9747	12	
	Concordia sulla Secchia	3164	6858	13		8342	18089	14	
	Finale Emilia	6122	12980	26		15129	33218	25	
	Medolla	2091	15071	9		5504	38722	10	
	San Felice sul Panaro	3586	18657	15		9821	48543	17	
	San Possidonio	1322	19979	6		3497	52040	6	
	San Prospero	1470	21449	7	**=128]**	4490	56530	8	**=134]**
D3	**Modena**	74675	74675	307	**307**	176965	176965	291	**291**
D4	**Sassuolo**	15685	42584	65	**175**	40872	113073	67	**186**
	Fiorano Modenese	5687	5687	24		16046	16046	27	
	Formigine	10953	16640	45		29827	45873	49	
	Frassinoro	1041	17681	5		2218	48091	4	
	Maranello	5678	23359	24		15819	63910	26	
	Montefiorino	1052	24411	5		2337	66247	4	
	Palagano	1112	25523	5		2488	68735	5	
	Prignano sulla Secchia	1376	26899	6	**=179]**	3466	72201	6	**=188]**
D5	**Pavullo nel Frignano**	5997	15968	25	**66**	14851		25	**62**
	Fanano	1350	1350	6		2905	2905	5	
	Fiumalbo	615	1965	3		1389	4294	3	
	Lama Mocogno	1389	3354	6		3040	7334	5	
	Montecreto	424	3778	2		934	8268	2	
	Pievepelago	883	4661	4		2150	10418	4	
	Polinago	906	5567	4		1870	12288	4	
	Riolunato	333	5900	2		749	13037	2	
	Serramazzoni	2832	8732	12		6710	19747	11	
	Sestola	1239	9971	6	**=70]**	2696	22443	5	**=66]**
D6	**Vignola**	8553	30613	36	**126**	20954	76200	35	**125**
	Castelnuovo Rangone	4535	4535	19		11759	11759	20	
	Castelvetro	3476	8011	15		9388	21147	16	
	Guiglia	1497	9508	7		3635	24782	6	
	Marano sul Panaro	1447	10955	6		3640	28422	6	
	Contese	1421	12376	6		3183	31605	6	
	Svignano sul Panaro	3221	15597	14		8323	39928	14	
	Spilamberto	4293	19890	18		10725	50653	18	
	Zocca	2170	22060	9	**=130]**	4593	55246	8	**=129]**
D7	**Castelfranco Emilia**	9903	22434	41	**92**	24518	57924	41	**95**
	Pastiglia	1248	1248	6		3236	3236	6	
	Comporto	2714	3962	12		7398	10634	13	
	Nonantola	4717	8679	20		12318	22952	21	
	Ravarino	1944	10623	8		5185	28137	9	
	San Cesario sul Panaro	1908	12531	8	**=95]**	5269	33406	9	**=99]**
	Totale Provincia/NAR	**252968**	**104770**	**1062**			**270757**	**1061**	**1040**

(a) Il totale di colonna n_d ($n_{P;d}$) è inferiore al totale di colonna n_{dc} ($n_{P;dc}$) per arrotondamenti eseguiti sempre per eccesso

Tabella 2 – Numero di famiglie (USS), dimensione campionaria proporzionale in base al numero di famiglie (n_{dc}), numero totale per distretto (n_d) per i Comuni inclusi (selezionati) nel campione della Provincia di Modena suddivisi per distretto sociosanitario[(a)] al 31/12/2000[(b)]

	Comune	USS	n_{dc}	n_d		Comune	USS	n_{dc}	n_d
D1	**Carpi**	**24674**	**126**	**150**	**D5**	**Pavullo nel Frignano**	**5997**	**58**	**66**
	Soliera	4928	25			Polinago	906	9	
	Totale D1		**151**			**Totale D5**		**67**	
D2	**Mirandola**	**8711**	**63**	**124**	**D6**	**Vignola**	**8553**	**74**	**126**
	Cavezzo	2549	19			Spilamberto[(d)]	4293	40	
	Finale Emilia	6122	44			Guiglia	1497	13	
	Totale D2		**126**			**Totale D6**		**127**	
D3	**Modena**	74675	**307**	**307**					
D4	**Sassuolo**	**15685**	**83**	**175**	**D7**	**Castelfranco Emilia**	**9903**	**63**	**92**
	Formigine	10953	58			Nonantola	4717	30	
	Maranello[(c)]	5678	30			**Totale D7**		**93**	
	Montefiorino	1052	6						
	Totale D4		**177**						
						Totale Provincia	**190076**	**1048**	**1040**

[(a)] Il totale di colonna n_d è inferiore al totale di colonna n_{dc} per gli arrotondamenti eseguiti sempre per eccesso.

[(b)] La data di riferimento è antecedente (di un anno) alle date di riferimento delle Tabelle 3, 4, e 5 relative al Comune di Modena perché al momento della realizzazione del piano di campionamento non erano ancora disponibili i dati provinciali della popolazione.

[(c)] Il Comune di Maranello ha rifiutato di partecipare all'indagine e, perciò, è stato sostituito con Fiorano Modenese.

[(d)] Il Comune di Spilamberto ha rifiutato di partecipare all'indagine e, perciò, è stato sostituito con Castelnuovo Rangone.

Tabella 3 – Numero di famiglie (USS, $N_{MO;ijk}$) per numero di componenti la famiglia, per classi di età e per genere del capofamiglia, nel Comune di Modena al 31/12/2001

Numero di Componenti	**Genere**	**Classi di età del capofamiglia**						
		<=34 anni	**35-49 anni**	**50-64 anni**	**65-74 anni**	**>=75 anni**	**Totale**	**Totale**
1 componente	M	2722	2736	1629	915	1166	**9168**	
	F	1797	1901	1942	2926	5316	**13882**	**23050**
2 componenti	M	1872	2076	4264	4589	3883	**16684**	
	F	970	1574	1446	894	1055	**5939**	**22623**
3 componenti	M	1458	4543	5326	1908	718	**13953**	
	F	527	1142	638	256	331	**2894**	**16847**
4 componenti e piú	M	899	5752	3879	712	298	**11540**	
	F	322	746	260	161	199	**1688**	**13228**
Totale	M	6951	15107	15098	8124	6065	**51345**	
Totale	F	3616	5363	4286	4237	6901	**24403**	
Totale	M+F	**10567**	**20470**	**19384**	**12361**	**12966**	**75748**	**75748**

Tabella 4 – Numero di famiglie (USS, $N_{MO;ijk}$) per tipologia famigliare e per numero di componenti la famiglia nel Comune di Modena al 31/12/2001

Tipologia famigliare	Numero di componenti						
	1	2	3	4	5	6 e +	Totale
Persone sole	23050						**23050**
Coppie coniugate		14626					**14626**
Coppie coniugate e figli			12778	7822	1186	267	**22053**
Coppie coniugate e altre persone			872	110	23	13	**1018**
Coppie coniugate, figli, e altre persone				968	737	448	**2153**
Genitori e figli		4577	1313	175	31	4	**6100**
Genitori, figli, e altre persone			1252	702	280	121	**2355**
Altro tipo di famiglia		3420	632	189	93	59	**4393**
Totale	**23050**	**22623**	**16847**	**9966**	**2350**	**912**	**75748**

Tabella 5 – Numero di famiglie nel campione (USS, $n_{MO;ijk}$) per numero di componenti la famiglia, per classi di età del capofamiglia, e per genere nel Comune di Modena al 31/12/2001

Numero di Componenti	Genere	Classi di età del capofamiglia						
		<=34 anni	35-49 anni	50-64 anni	65-74 anni	>=75 anni	Totale	Totale
1 componente	M	24	24	14	8	10	**80**	
	F	16	16	17	24	45	**118**	**198**
2 componenti	M	15	17	36	39	33	**140**	
	F	8	14	12	7	9	**50**	**190**
3 componenti	M	12	37	44	16	6	**115**	
	F	4	10	5	2	3	**24**	**139**
4 componenti e piú	M	7	48	32	6	3	**96**	
	F	3	6	2	1	2	**14**	**110**
Totale	M	58	126	126	69	52	**431**	
Totale	F	31	46	36	34	59	**206**	
Totale	M+F	**89**	**172**	**162**	**103**	**111**	**637**	**637**

3. Gli esiti della rilevazione campionaria

I Comuni coinvolti nell'indagine differiscono da quelli riportati in Tabella 2 perché due di essi non hanno collaborato. Il responsabile dell'Anagrafe di Maranello ha sostenuto di essere tempestato e oberato da richieste di interviste e, pertanto, ha negato la sua collaborazione; data la peculiarità del Comune, può essere plausibile. In sua sostituzione si è selezionato il Comune di Fiorano Modenese che è molto vicino a Maranello e ha circa la stessa numerosità di famiglie. Il responsabile dell'Anagrafe di Spilamberto ha sostenuto di versare in una grave carenza temporanea di personale e non è stato possibile ottenere l'estrazione casuale del campione di famiglie; pertanto, è stato sostituito con il Comune di Castelnuovo Rangone, che si è dimostrato piú cooperativo. Questa sostituzione è stata eseguita, come la precedente, con la selezione di un Comune assai simile per dimensione e caratteristiche ambientali, economiche, e sociali.

Il campione è, come già detto, a due stadi (i Comuni della Provincia di Modena e le famiglie residenti), con stratificazione delle UPS secondo il distretto sociosanitario e la loro dimensione, e con una selezione PPS per distretto, relativamente ai Comuni NAR. Le USS sono state selezionate dalla lista anagrafica di ciascun Comune con il metodo del campionamento sistematico circolare; ossia, senza reimmissione e con probabilità uguali (Särndal, Swensson, Wretman, 1992). Alle Anagrafi è stato fornito il passo a valore intero, $a_{dc} = \lfloor N_{dc} / n_{dc} \rfloor$, e il punto di partenza, ρ, determinato generando un numero casuale con distribuzione uniforme discreta in $[1, N_{dc}]$. Si sono selezionate le famiglie che nella lista anagrafica occupavano le posizioni generate dall'espressione:

$$\rho + (j-1)\ a_{dc} - N \cdot 1_{[N_{dc}+1,\infty)}\left[\rho + (j-1)\ a_{dc}\right] \qquad \text{per} \quad j = 1, \cdots, n_{dc};$$

dove $1_{[\bullet)}[\cdot]$ è la funzione indicatrice che vale 1, se l'argomento appartiene all'insieme specificato nell'indice, 0 altrimenti. L'estrazione iniziava, quindi, dal punto di partenza casuale fornito e proseguiva «lungo» la lista, ricominciando all'inizio dopo la fine della lista. La selezione delle famiglie dalla base di dati anagrafica dei Comuni è stata eseguita da un dipendente. Tutti i membri delle famiglie sono stati inclusi nel campione.

Per sopperire all'eventuale insuccesso degli intervistatori si è estratta la lista «suppletiva», che contiene le USS (dette anche, per brevità, «riserve») tra le quali selezionare le sostitutive di quelle che non si riescono a intervistare sia per il rifiuto di rispondere o di entrare in contatto con l'intervistatore, sia per l'irreperibilità (indirizzo sbagliato, trasferimento, assenza perdurante da casa). L'entità della lista di riserva è stata fissata a circa il quadruplo della dimensione obiettivo (cfr. §5). La lista di riserva è stata estratta assieme alle unità campionarie in tutti i comuni, eccetto Modena; pertanto, il passo è stato determinato riportando il quintuplo di n_{dc} al denominatore della frazione per il calcolo del passo e con un successivo campionamento sistematico si è determinato l'elenco base degli intestatari delle schede di famiglia: i nominativi del campione obiettivo. A Modena, invece, si sono estratti cinque campioni indipendenti: il primo costituisce la lista base, e gli altri rappresentano la lista suppletiva.

La rilevazione è iniziata a giugno 2002 e è proseguita fino a dicembre dello stesso anno. Le difficoltà incontrate sono fisiologiche nelle indagini di questa natura; in particolare, si sono avute mancate partecipazioni («rifiuti») e non si sono rintracciate alcune famiglie («irreperibili»). Nella Tabella 6 sono riportati alcuni dati essenziali del processo: il numero delle famiglie della lista base, il numero di famiglie partecipanti all'indagine, il numero di mancate interviste rispetto all'obiettivo, la copertura del campione, il numero di rifiuti, e il numero di irreperibili. Nella Tabella 7 si sono riportate le informazioni analoghe relative al Comune di Modena.

Le interviste realizzate sono state complessivamente 1235; pertanto, si è ottenuto il 10,4% in meno di famiglie rispetto al campione obiettivo. I comuni con lo scarto piú elevato tra interviste obiettivo e interviste realizzate sono sei: tre dell'area montana (Pavullo, Montefiorino, Polinago), uno dell'area pedemontana (Guiglia), e due dell'area della «pianura centrale» (Castelfranco Emilia e Nonantola).

Il mancato conseguimento della dimensione obiettivo nei Comuni del campione dipende da varie ragioni (Bigarelli, Fregni, Silvestri, 2003). Nell'area montana e pedemontana le famiglie si sono rivelate meno disponibili a rilasciare l'intervista, e i tassi di rifiuto sono stati superiori alla media provinciale (dal 65% all'85%, contro una media del 56%). L'insistenza operata sulle famiglie è stata anche un po' piú bassa per ragioni logistiche: le distanze inducono un aumento di costi e di tempi per gli spostamenti, con

conseguenti visite in orari non sempre favorevoli. Si è riscontrata, poi, una rilevante presenza di famiglie residenti in case sparse. Nel caso di impossibilità a stabilire un contatto telefonico con queste famiglie, gli intervistatori vi hanno potuto effettuare una sola visita diretta, a causa delle notevoli distanze che le separavano dai centri abitati. Nella maggior parte di questi casi, gli intervistatori hanno avuto l'impressione che l'abitazione corrispondesse a una seconda casa, ma non è stato possibile accertarlo (Bigarelli, Fregni, Silvestri, 2003). Nella «pianura centrale» il mancato raggiungimento dell'obiettivo campionario è dipeso da difficoltà inerenti agli intervistatori. Nei Comuni dove non è stata raggiunta la dimensione prevista, le liste delle famiglie (campione piú riserve) non sono state sempre esaurite. La decisione di fermarsi al numero di interviste raccolte e di sospendere la rilevazione è stata presa tenendo conto dei vincoli di tempo stabiliti e della disponibilità degli intervistatori.

I risultati conseguiti sono complessivamente soddisfacenti; anche se il confronto con alcuni indici dell'indagine della Banca d'Italia (2002, pp. 31-32), calcolati per la quota non *panel*, che è quella confrontabile, mostra valori inferiori: le interviste completate nel Comune di Modena sono il 33,4% (nella Provincia il 31,3%) contro il 38,3%; le famiglie indisponibili nel Comune di Modena sono il 39,9% (nella Provincia il 40,8%) contro il 57,2%; le famiglie irreperibili nel Comune di Modena sono il 26,6% (nella Provincia il 27,9%) contro il 15,2%. Si noti che per le famiglie irreperibili non si è potuto accertare se erano ineleggibili: famiglie non esistenti all'indirizzo anagrafico per errori, decessi, o trasferimenti; ciò avrebbe migliorato l'«efficienza» della rilevazione.

Tabella 6 – Numero di famiglie o interviste obiettivo (n_{dc}), numero di famiglie partecipanti ($n_{p;dc}$), numero complessivo di famiglie estratte dalla lista (campione piú «riserve», n^0_{dc}), differenza tra il numero delle interviste obiettivo e il numero di famiglie partecipanti (Δn_{dc}), copertura del campione (Cop. %=$100\, n_{p;dc} / n_{dc}$), numero di famiglie che rifiutano di partecipare ($n_{r;dc}$), e numero di famiglie irreperibili ($n_{*;dc}$) per i Comuni nel campione della Provincia di Modena secondo il distretto sociosanitario

	Comune	n_{dc}	$n_{p;dc}$	n^0_{dc}	Δn_{dc}	**Cop. %**	$n_{r;dc}$	$n_{*;dc}$
D1	**Carpi**	**126**	**123**	**511**	**–3**	**97.6**	**99**	**92**
	Soliera	25	25	100	0	100,0	22	18
D2	**Mirandola**	**63**	**56**	**252**	**–7**	**88.9**	**86**	**35**
	Cavezzo	19	19	76	0	100,0	27	14
	Finale Emilia	44	44	176	0	100,0	17	9
D3	**Modena**	**637**	**589**	**2549**	**–48**	**92.5**	**704**	**472**
D4	**Sassuolo**	**83**	**76**	**332**	**–7**	**91.6**	**87**	**65**
	Fiorano Modenese(*)	30	30	121	**0**	100,0	46	35
	Formigine	58	58	232	0	100,0	69	43
	Montefiorino	6	3	24	–3	50,0	16	2
D5	**Pavullo nel Frignano**	**58**	**33**	**232**	**–25**	**56,9**	**95**	**42**
	Polinago	9	6	36	–3	66,7	11	6
D6	**Vignola**	**74**	**67**	**318**	**–7**	**90,5**	**130**	**120**
	Castelnuovo Rangone(**)	40	40	165	0	100,0	55	19
	Guiglia	13	2	56	–11	15,4	11	11
D7	**Castelfranco Emilia**	**63**	**44**	**254**	**–19**	**69,8**	**48**	**27**
	Nonantola	30	20	120	–10	66,7	26	40
	Totale Provincia	**1378**	**1235**	**5554**	**–143**	**–10.4**	**1549**	**1050**

(*) Il Comune di Fiorano sostituisce il Comune di Maranello perché l'Ufficio dell'anagrafe ha rifiutato di collaborare.

(**) Il Comune di Castelnuovo Rangone sostituisce il Comune di Spilamberto, *idem*.

Tabella 7 – Numero di famiglie o interviste obiettivo (n_{dc}), numero di famiglie partecipanti ($n_{p;dc}$), numero complessivo di famiglie estratte dalla lista (campione piú «riserve», n^0_{dc}), differenza tra il numero delle interviste obiettivo e il numero di famiglie partecipanti (Δn_{dc}), copertura del campione (Cop. %=$100\, n_{p;dc} / n_{dc}$), numero di famiglie che rifiutano di partecipare ($n_{r;dc}$), e numero di famiglie irreperibili ($n_{*;dc}$) per il Comune di Modena secondo il numero di componenti la famiglia, le classi di età e il genere del capofamiglia

NCF	Età	Genere	n_{dc}	$n_{p;dc}$	n^0_{dc}	Δn_{dc}	Cop. %	$n_{r;dc}$	$n_{*;dc}$
1 comp.	**<=34 a.**	**M**	24	14	96	10	58,3	8	43
		F	16	16	64	0	100,0	12	1
	35-49 a.	**M**	24	24	96	0	100,0	17	25
		F	16	16	64	0	100,0	12	7
	50-64 a.	**M**	14	9	56	5	64,3	8	28
		F	17	15	68	2	88,2	22	30
	65-74 a.	**M**	8	8	32	0	100,0	8	15
		F	24	24	96	0	100,0	51	20
	>=75 a.	**M**	10	10	40	0	100,0	10	4
		F	45	27	180	18	60,0	89	63
2 comp.	**<=34 a.**	**M**	15	13	60	2	86,7	17	26
		F	8	8	32	0	100,0	5	10
	35-49 a.	**M**	17	17	68	0	100,0	9	2
		F	14	14	56	0	100,0	5	7
	50-64 a.	**M**	36	34	144	2	94,4	30	33
		F	12	12	48	0	100,0	12	1
	65-74 a.	**M**	39	38	156	1	97,4	99	16
		F	7	7	28	0	100,0	7	12
	>=75 a.	**M**	33	32	132	1	97,0	63	5
		F	9	10	36	–1	111,1	9	3
3 comp.	**<=34 a.**	**M**	12	12	48	0	100,0	13	10
		F	4	4	16	0	100,0	3	0
	35-49 a.	**M**	37	37	149	0	100,0	13	4
		F	10	10	40	0	100,0	2	1
	50-64 a.	**M**	44	43	176	1	97,7	28	16
		F	5	5	20	0	100,0	4	1
	65-74 a.	**M**	16	16	64	0	100,0	32	0
		F	2	1	8	1	50,0	6	0
	>=75 a.	**M**	6	6	24	0	100,0	4	3
		F	3	3	12	0	100,0	8	0
>=4comp.	**<=34 a.**	**M**	7	7	28	0	100,0	3	2
		F	3	3	12	0	100,0	1	1
	35-49 a.	**M**	48	42	192	6	87,5	49	79
		F	6	6	24	0	100,0	6	0
	50-64 a.	**M**	32	32	128	0	100,0	18	1
		F	2	2	8	0	100,0	1	0
	65-74 a.	**M**	6	6	24	0	100,0	10	0
		F	1	1	4	0	100,0	3	0
	>=75 a.	**M**	3	3	12	0	100,0	3	3
		F	2	2	8	0	100,0	4	0
Totale			**637**	**589**	**2549**	**48**	**92,5**	**704**	**472**

4. I fattori di riporto alla popolazione obiettivo

In una popolazione $\wp$ di N unità, sia Y il carattere oggetto di stima (per esempio, il reddito totale delle famiglie) con una distribuzione statistica incognita e valori $(Y_1, Y_2 \cdots, Y_N)$. Si voglia stimare il totale della Y in $\wp$, dato da $Y = \sum_{i=1}^{N} Y_i$, in base al campione osservato $(y_1, y_2 \cdots, y_n)$, con l'eventuale uso di variabili ausiliarie, dove y_1 indica il valore osservato di Y nell'unità ottenuta dalla prima estrazione, y_2 indica il valore osservato di Y nell'unità ottenuta dalla seconda estrazione, e cosí via fino all'*n*-esima estrazione. Gli stimatori che si considerano, in genere, sono lineari del tipo

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i \tag{10}$$

dove le quantità w_i, dette *pesi*, non dipendono dal numero d'ordine delle osservazioni, ma possono dipendere dal tipo di campionamento adottato e dall'etichetta che individua l'unità statistica selezionata (Cicchitelli, Herzel, Montanari, 1997).

Si consideri, ora, la Provincia di Modena, stratificata per distretto sociosanitario. Il totale della caratteristica, Y, è dato dalla somma estesa a tutte le unità statistiche della Provincia. Sia Y_{dci} il valore di Y per l'*i*-esima famiglia nel *c*-esimo Comune del *d*-esimo strato. Il totale della Y, che nell'esempio è il reddito delle famiglie, sarà dato da

$$Y = \sum_{d=1}^{D} \sum_{c=1}^{C_d} \sum_{i=1}^{N_{dc}} Y_{dci} \tag{11}$$

dove D è il numero di distretti, C_d è il numero di Comuni nel *d*-esimo distretto, N_{dc} è il numero di USS nel *c*-esimo Comune del *d*-esimo distretto.

Si consideri, poi, il piano di campionamento probabilistico a due stadi che genera un campione di *n* unità estratte senza ripetizione (reimmissione), come nel caso in oggetto, in cui sia le UPS e sia le USS vengano estratte con probabilità variabili. Siano $(y_1, y_2, \cdots, y_n)$ le osservazioni campionarie; siano $(\pi_{d1}, \pi_{d2}, \cdots, \pi_{dc_d})$ le probabilità di inclusione delle UPS, dove l'indice c_d indica il numero di Comuni nel campione del *d*-esimo distretto; siano $(\pi_{dc1}, \pi_{dc2}, \cdots, \pi_{dcn_{dc}})$ le probabilità di inclusione delle USS, una volta che sia stata estratta la *c*-esima UPS, dove n_{dc} indica il numero di famiglie nel campione del *d*-esimo distretto del *c*-esimo Comune; allora, lo stimatore corretto del totale, $\hat{Y}$, è

$$\hat{Y} = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \sum_{i=1}^{n_{dc}} \frac{y_{dci}}{\pi_{dc} \pi_{dci}} = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \frac{\hat{Y}_{dc}}{\pi_{dc}}, \tag{12}$$

che è uno stimatore di Horvitz-Thompson (Horvitz, Thompson, 1952), ottenuto dalla combinazione lineare delle osservazioni campionarie nei $D = 7$ distretti con pesi pari a $1/(\pi_{dc} \pi_{dci})$, dove $(c = 1, \cdots, c_d)$ e $(i = 1, \cdots, n_{dc})$, dipendenti dalle etichette delle unità cui si riferiscono le osservazioni, ossia dal piano di campionamento adottato. La quantità $\hat{Y}_{dc}$ è lo stimatore di secondo stadio del totale dell'UPS *c* del *d*-esimo distretto e le probabilità di selezione delle UPS sono uguali all'unità, $\pi_{dc} = 1$, per i Comuni AR. I pesi delle combinazioni lineari degli stimatori sono dati, dunque, dall'espressione inversa delle probabilità di selezione delle unità statistiche nel campione.

Gli stimatori associati al campionamento a piú stadi sono complessi e, pertanto, anche le varianze degli stimatori assumono espressioni complicate. In generale, la va-

rianza dello stimatore del totale, $\hat{Y}$, assume la forma seguente (Cicchitelli, Herzel, Montanari, 1997, p. 194)

$$V\left(\hat{Y}\right)=V_1\left(\sum_{d=1}^{D}\sum_{c=1}^{c_d}\frac{\hat{Y}_{HT;dc}}{\pi_{dc}}\right)+\sum_{d=1}^{D}\sum_{c=1}^{c_d}\frac{V_2\left(\hat{Y}_{dc}\right)}{\pi_{dc}} \tag{13}$$

dove il primo termine a secondo membro è la varianza di primo stadio dello stimatore di Horvitz-Thompson del totale di $\wp$ nel campionamento a grappoli a un solo stadio e $V_2\left(\hat{Y}_{dc}\right)$ è la varianza di secondo stadio dello stimatore $\hat{Y}_{dc}$ del totale del grappolo *c* del campione nel distretto *d*. L'espressione finale della varianza si ottiene partendo dalla (13) e adattandola alla specifica strategia.

Le probabilità di inclusione derivano dall'entità della popolazione di riferimento, $\wp$, al momento del campionamento. Nell'espressione di uno stimatore, come indicato nella (10), il peso di una unità *i*, w_i, è il reciproco della probabilità di inclusione, detto *peso base*. Il peso deve essere spesso aggiustato per sopperire a varie difficoltà; ma, da un lato, l'aggiustamento migliora la rappresentatività del campione, dall'altro lato, introduce una non linearità negli stimatori. Si perviene al peso finale, pertanto, con una serie di correzioni. Nel caso in oggetto si possono avere, poi, almeno due diversi tipi di pesi perché $\wp$ può essere: sia le famiglie residenti, N; sia la popolazione residente, P. Se l'unità di analisi è la famiglia, allora si usano i pesi determinati secondo espressioni che contengono N, che indica il numero di famiglie. Se l'unità di analisi è l'individuo, allora è sufficiente sostituire nelle espressioni il simbolo *N* con il simbolo P, che indica il numero di individui, anche se cosí, in effetti, si attua una «post-stratificazione».

Il tempo al quale «ancorare» la popolazione di riferimento deve essere fissato, dato che subisce una evoluzione nel tempo e l'indagine è stata svolta in un lasso di tempo che coincide, pressappoco, circa con il secondo semestre del 2002; pertanto, riferirsi a una data precisa non è strettamente necessario. Una possibilità consiste nell'usare come popolazione di riferimento la media dei dati disponibili al 31/12/2001 e al 31/12/2002, $\wp=\left(\wp_{01}+\wp_{02}\right)/2$; tuttavia, l'attuale indisponibilità dei dati del 2002 induce a utilizzare la popolazione al 31/12/2001. Tale scelta altera, però, le probabilità di inclusione e si configura come una specie di «post-stratificazione».

4.1. I fattori di riporto alla popolazione obiettivo per il Comune di Modena

I pesi sono già predeterminati al momento della progettazione dell'indagine perché le probabilità di selezione delle UPS e USS sono note, ma le mancate partecipazioni introducono un fattore di disturbo di cui tenere conto; quindi, per il Comune di Modena, si devono presumibilmente usare pesi diversi per ciascun dominio di studio (o strato), anche se si tratta di un campione autoponderante, per correggere le mancate collaborazioni. L'espressione per stimare il totale del carattere *Y* si ottiene adattando l'equazione precedente al piano di campionamento adottato nel distretto di Modena (*d*=3), che è stratificato: per classe di ampiezza della famiglia, i, dove $i=1,\cdots,I(=4)$; per classe di età del capofamiglia, j, dove $j=1,\cdots,J(=5)$; per genere del capofamiglia, k, dove $k=1,2(=K)$:

$$\hat{Y}_{d=3}=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{n_{ijk|d=3}}\frac{1}{\pi_{ijk|d=3}}y_{ijkl|d=3}\,. \tag{14}$$

Tale statistica è detta anche *stimatore per espansione* perché nel caso di un campionamento casuale semplice o autoponderante, come è nel Comune di Modena, diventa semplicemente il prodotto della corrispondente grandezza campionaria moltiplicata per l'inverso della frazione di campionamento: $\hat{Y}_{d=3} = (N/n)\sum_{ijk} y_{ijk|d=3} = N\ \bar{y}_{d=3}$. La frazione di campionamento è, in totale, n/N; quindi, si trattano i dati come se ogni unità del campione rappresentasse N/n unità della popolazione e, perciò, il fattore N/n è detto anche *coefficiente di espansione*. Nel caso in oggetto, all'interno di ogni strato (o dominio di studio) si ha un peso che corrisponde proprio a questa rappresentazione, dato dall'inverso della probabilità di selezione del primo ordine $1/\pi_{ijk|d=3}$. Per semplificare le espressioni, si ometterà nel séguito l'indice *d*=3 quando è chiaro l'universo di riferimento; in particolare, quando i titoli dei sottoparagrafi specificano che ci si riferisce al Comune di Modena. All'interno di ogni strato il fattore di espansione o peso è dato da

$$w_{ijk} = \frac{1}{\pi_{ijk}} = \frac{N_{ijk}}{n_{ijk}}. \tag{15}$$

Per la determinazione dei pesi, che riportano la popolazione alla data di riferimento della lista, occorre considerare: la non appartenenza alla popolazione di riferimento $\wp$; l'emigrazione o l'uscita da $\wp$; e la non rintracciabilità che può includere sia l'emigrazione, sia l'uscita, sia gli errori di registrazione negli archivi. Il trattamento delle unità che rientrano in tali categorie può seguire diverse strategie alternative nel calcolo dei pesi.

(a) Si ignorano e si trattano come non rispondenti, ma ne consegue una possibile sovrastima della popolazione di riferimento.
(b) Si assume che nella lista vi siano unità estranee alla popolazione di riferimento.

L'entità degli errori può essere trascurabile e la dimensione campionaria per strato piccola; allora la stima risente della rarità degli eventi e della scarsa numerosità per strato del campione, sicché si rischia di enfatizzare l'effetto nelle stime e ottenere una considerevole sottostima dei valori della popolazione. Nel calcolo dei pesi per strato si utilizzerà la popolazione di riferimento originaria, N_{ijk}. Per il calcolo dei pesi finali occorre considerare la probabilità di rintracciare una unità e la probabilità di ottenere la sua partecipazione:

$$w_{ijk} = \frac{1}{\pi_{ijk}}\ \frac{1}{\pi_{r;ijk}}\ \frac{1}{\pi_{p;ijk}} = \frac{N_{ijk}}{n_{ijk}}\ \frac{n_{ijk}}{n_{c;ijk}}\ \frac{n_{c;ijk}}{n_{p;ijk}} \tag{16}$$

dove $\pi_{r;ijk}$ è la probabilità che l'unità sia rintracciata, $\pi_{p;ijk}$ è la probabilità che l'unità partecipi all'indagine, n_{ijk} è il numero di unità selezionate nello strato *ijk*, $n_{c;ijk}$ è il numero di unità contattate, e $n_{p;ijk}$ denota il numero di unità che partecipano all'indagine e rispondono alle domande del questionario. Il peso finale per strato, *ijk*, corrisponde, banalmente, al rapporto tra il numero di unità della popolazione nello strato *ijk* e il numero di unità partecipanti all'indagine

$$w_{ijk} = \frac{N_{ijk}}{n_{p;ijk}} = \frac{1}{\pi^*_{p;ijk}} \tag{17}$$

dove $1/\pi^*_{p;ijk}$ può interpretarsi come una «pseudo-probabilità» di selezione o probabilità di rilevare effettivamente i dati dell'unità statistica perché deriva dalla probabilità di inclusione modificata o corretta per le difficoltà incontrate e che sarà utile in questa

forma solo per determinare l'espressione di normalizzazione a uno dei pesi (v. *infra*); infatti, è in questa forma espressiva che si utilizzerà per ricavarli.

Nella Tabella 8 sono riportati i pesi, w_{ijk}, calcolati secondo l'espressione (17) dove si può notare che sono tutti circa dello stesso ordine di grandezza. In realtà, dovrebbero essere tutti uguali; ma già in fase di estrazione i passi per strato differivano tra loro sia perché vi erano stati arrotondamenti per eccesso nella determinazione di n_{ijk}, sia perché l'arrotondamento dei passi a un numero intero induceva variazioni per dominio, sia perché la popolazione al momento della selezione (il 29 maggio 2002) aveva già subíto modifiche rispetto alla data di riferimento (31/12/2001). Le mancate risposte per dominio comportano una ulteriore modifica dei pesi; infatti, nei domini dove non c'è stata una copertura completa si osserva un aumento del peso piú alto del suo valore «medio». Nella Tabella 9 sono riportati, invece, i pesi $w_{P;ijk}$ riferiti alla popolazione (persone) residente nel Comune di Modena.

Tabella 8 – Pesi, w_{ijk}, per riportare al numero di famiglie del Comune di Modena, $N_{MO;ijk}$, in data 31/12/2001, per numero di componenti la famiglia, per classi di età e per genere del capofamiglia

Numero di Componenti	**Genere**	**Classi di età del capofamiglia**				
		<=34 anni	**35-49 anni**	**50-64 anni**	**65-74 anni**	**>=75 anni**
1 componente	M	194,4286	114,0000	181,0000	114,3750	116,6000
	F	112,3125	118,8125	129,4667	121,9167	196,8889
2 componenti	M	144,0000	122,1176	125,4118	120,7632	121,3438
	F	121,2500	112,4286	120,5000	127,7143	105,5000
3 componenti	M	121,5000	122,7838	123,8605	119,2500	119,6667
	F	131,7500	114,2000	127,6000	256,0000	110,3333
4 componenti e piú	M	128,4286	136,9524	121,2188	118,6667	99,3333
	F	107,3333	124,3333	130,0000	161,0000	99,5000

Tabella 9 – Pesi, $w_{P;ijk}$, per riportare al numero di soggetti residenti nel Comune di Modena, $P_{MO;ijk}$, in data 31/12/2001, per numero di componenti la famiglia, per classi di età e per genere del capofamiglia

Numero di Componenti	**Genere**	**Classi di età del capofamiglia**				
		<=34 anni	**35-49 anni**	**50-64 anni**	**65-74 anni**	**>=75 anni**
1 componente	M	143,2632	76,0000	162,9000	114,3750	116,6000
	F	81,6818	100,0526	114,2353	121,9167	196,8889
2 componenti	M	144,0000	115,3333	125,4118	119,1948	121,3438
	F	114,1176	112,4286	125,7391	137,5385	111,0526
3 componenti	M	118,2162	123,9000	123,8605	127,2000	134,6250
	F	121,6154	110,5161	127,6000	384,0000	110,3333
4 componenti e piú	M	138,5862	146,1588	133,8560	127,6000	114,9167
	F	118,7500	141,3913	130,4444	182,0000	97,3333

La soluzione adottata è la piú semplice per compensare le stime dalle difficoltà delle indagini e dalle non risposte; altre strategie, piú sofisticate e complesse, che non si possono spesso applicare alle indagini su larga scala, si trovano in Little e Rubin (1987) e Rubin (1988). Gli stimatori diventano, però, non lineari e le loro varianze aumentano (Kish, 1990, 1992); inoltre, le correzioni apportate non sono correlate con le variabilità negli strati e tendono a incrementare la varianza (Bethlehem, Keller, 1987; Potter, 1990); infatti, il peso dei rispondenti è incrementato perché devono rappresentare, in un certo senso, anche le unità che rifiutano di partecipare o che sono irreperibili.

4.2. I fattori di riporto alla popolazione obiettivo per la Provincia di Modena

Il «peso» di ogni USS che partecipa all'indagine «rappresenta», in un certo senso, il numero di UPS del Comune e del distretto di appartenenza. Per semplificare le espressioni si indica con $c=1$ il Comune AR del *d*-esimo strato e con i valori successivi gli altri Comuni di $\wp$ o del campione. Nel calcolo dei pesi occorre distinguere: (a) i Comuni AR dove $\pi_{dc}=1$, (b) i distretti con una o piú UPS tipo NAR. Naturalmente, si possono considerare strategie diverse a seconda della numerosità delle UPS selezionate: una, due, o piú. Per semplicità, ci si è limitati ai primi due casi, nei quali i pesi si ottengono come segue:

$$w_{d,c=1}=\frac{1}{\pi_{dc}}\frac{1}{\pi_{dci}}=\frac{N_{d1}}{n_{d1}}, \tag{18}$$

$$w_{d,c>1}=\frac{1}{\pi_{dc}}\frac{1}{\pi_{dci}}\equiv\frac{N_d^*}{c_d N_{dc}}\frac{N_{dc}}{n_{dc}}=\frac{1}{c_d}\frac{N_d^*}{n_{dc}}, \tag{19}$$

dove, relativamente al *d*-esimo distretto, $N_d^*=N_d-N_{d1}$ è il totale delle famiglie nello strato NAR, N_{d1} è il numero di famiglie del Comune AR, c_d è il numero di UPS di tipo NAR estratte nel campione, n_{d1} e n_{dc} sono le dimensioni dei campioni nel Comune AR e nei Comuni NAR estratti, rispettivamente. Dalla precedente espressione, si ha che la probabilità di selezione del *c*-esimo Comune del *d*-esimo distretto è pari a $c_d n_{dc}/N_d^*$. Nel séguito, i pesi saranno indicati solo con w_{dc} per semplificare le espressioni.

Nella Tabella 10 sono esposti i pesi w_{dc} e $w_{P;dc}$, riferiti alle famiglie e alla popolazione residente, rispettivamente, e calcolati secondo la (18) per i Comuni AR e secondo la (19) per i Comuni NAR. Nella stessa sono riportati anche la popolazione di famiglie nel distretto (N_d) e nei Comuni campione (N_{dc}), il numero di famiglie nel campione (n_{dc}), il numero di soggetti nel distretto (P_d) e nei Comuni campione (P_{dc}), il numero di soggetti nel campione (p_{dc}). I valori dei pesi sono molto diversi tra loro: non solo per compensare le mancate risposte; ma, soprattutto, perché ogni UPS stima una parte della popolazione dello strato data dal reciproco del numero di UPS estratte, ossia di c_d. Si hanno cosí valori assai elevati nei Comuni piccoli e con poche unità rilevate. Valori sorprendenti e pressoché inaccettabili si hanno nei comuni di Guiglia, Montefiorino, e Polinago. Tale risultato è un indicatore dei problemi riscontrati nella raccolta dei dati tra le famiglie residenti nei Comuni di montagna; in un certo senso, esse sono sottorappresentate; ma si era deciso di non stratificare secondo l'altezza sul livello del mare per convenienze organizzative. I pesi relativi ai soggetti non sono molto diversi da quelli relativi alle famiglie, se non nei Comuni montani già citati.

Il peso base, eventualmente già aggiustato per le mancate partecipazioni, può essere aggiustato con il metodo della post-stratificazione, che diventa piú efficace quando si conosce la distribuzione congiunta di due o piú caratteri della popolazione di riferimento; per esempio, il numero di componenti la famiglia, la classe di età e il genere del capofamiglia. Si assumono come post-strati i domíni definiti dagli incroci (celle) dei caratteri noti e il peso base è moltiplicato per il quoziente tra il numero di unità della popolazione appartenenti al post-strato e la somma dei pesi delle unità campionarie che appartengono al post-strato stesso.

Tabella 10 – Numero di famiglie di $\wp$ nello strato (N_d), numero di famiglie di $\wp$ nei Comuni campione (N_{dc}), numero di famiglie nel campione (n_{dc}), pesi relativi alle famiglie (w_{dc}), numero di soggetti nello strato (P_d), numero di soggetti nei Comuni campione (P_{dc}), numero di soggetti nel campione (p_{dc}), pesi relativi ai soggetti ($w_{P;dc}$), per i Comuni campione della Provincia di Modena al 31/12/2001

	Comune	N_d	N_{dc}	n_{dc}	w_{dc}	P_d	P_{dc}	p_{dc}	$w_{P;dc}$
D1	**Carpi**	25020	25020	123	203,4146	62288	62288	304	204,8947
	Soliera	(*)12120	5068	25	484,8000	(*)31462	13238	62	507,4516
D2	**Mirandola**	8763	8763	56	156,4821	22115	22115	130	170,1154
	Cavezzo	(*)21876	2618	19	575,6842	(*)56918	6775	50	569,1800
	Finale Emilia		6098	44	248,5909		15212	118	241,1780
D3	**Modena**	75748	75748	589	128,6044	178013	178013	1388	128,2514
D4	**Sassuolo**	15854	15854	76	208,6053	41003	41003	195	210,2718
	Fiorano Modenese	(*)27343	5778	30	303,8111	(*)72736	16106	94	257,9291
	Formigine		11204	58	157,1437		30252	173	140,1464
	Montefiorino		1052	3	3038,1111		2332	6	4040,8889
D5	**Pavullo nel Frignano**	6105	6105	33	185,0000	15126	15126	92	164,4130
	Polinago	(*)10073	875	6	1678,8333	(*)22674	1888	12	1889,5000
D6	**Vignola**	8717	8717	67	130,1045	21276	21276	170	125,1529
	Castelnuovo Rangone	(*)22562	4679	40	282,0250	(*)56227	12081	102	275,6225
	Guiglia		1536	2	5640,5000		3709	4	7028,3750
D7	**Castelfranco Emilia**	10283	10283	23	447,0870	25359	25359	103	246,2039
	Nonantola	(*)12968	4863	29	447,1724	(*)34118	12562	59	578,2712
	Totale Provincia	**257432**	**179797**	**1223**	**210,4922**	**639315**	**441976**	**3062**	**208,7900**

(*) Totale di USS nello strato NAR dal quale sono estratti i Comuni elencati nella riga o nelle righe corrispondenti.

4.3. Normalizzazione dei pesi all'unità

Per eseguire test statistici e/o stimare i parametri di modelli rappresentativi della realtà indagata non si può pesare con w_{ijk}, dato dalla precedente equazione perché esso altera la numerosità campionaria e, quindi, le probabilità di significatività relative alle ipotesi da sottoporre a verifica. In pratica, quindi, per rimediare a tali inconvenienti è utile «scalare» i pesi in modo che la loro somma sia uguale all'unità, anche se i totali non sono cosí riportati alla popolazione di riferimento (Verma, 1995). Per incorporare la struttura del campione nella determinazione degli stimatori e non alterare la numerosità campionaria, si può utilizzare un insieme di pesi che mantengano inalterate le caratteristiche del campione. Le stime si eseguono, separatamente, a due livelli: uno per il Co-

mune di Modena, e l'altro per la Provincia di Modena senza il Comune di Modena.

4.3.1. Normalizzazione nel Comune di Modena

L'allocazione proporzionale, che è autoponderante, non comporta la necessità di normalizzare all'unità i pesi durante l'elaborazione dei dati; ma, per compensare le mancate partecipazioni, si può utilizzare un insieme di pesi che, partendo da w_{ijk}, mantengano inalterate le caratteristiche del campione, ossia soddisfacciano due vincoli:

$$\text{(a)} \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w^*_{ijk} = IJK \qquad\qquad \text{(b)} \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w^*_{ijk} n_{ijk} = n\,.$$

Per soddisfare entrambi i criteri si può utilizzare un peso dato dal rapporto tra i pesi «originari», $1/\pi^*_{p;ijk}$, e un peso medio, $1/\overline{\pi}^*_p$, in modo da soddisfare le condizioni (a) e (b). Le grandezze figurano al denominatore, sicché si può calcolare la media secondo il criterio del Chisini (1929), usando come aggregazione la funzione somma delle quantità inverse perché tutte positive (sono «pseudo-probabilità»). La media, secondo Chisini, di un insieme di n osservazioni di una variabile Y è quel valore intermedio $\overline{Y}$ (compreso tra il minimo, $y_{(1)}$, e il massimo, $y_{(n)}$) che, sostituito a ciascuna osservazione, lascia invariato il valore una funzione sintetica delle osservazioni:

$$f(y_1, y_2, \cdots, y_n) = f(\overline{Y}, \overline{Y}, \cdots, \overline{Y}).$$

La definizione comporta la trasferibilità del carattere Y perché il valore $\overline{Y}$ uguagli la funzione $f(\cdot)$ quando si sostituiscono le osservazioni con il valore costante $\overline{Y}$. Si richiede, pertanto, di specificare la $f(\cdot)$ in base alla natura del carattere (additiva, moltiplicativa, inversa, e cosí via) e alla sua trasferibilità (Piccolo, 1998, pp. 78-92). Nel caso in oggetto, si definisce la funzione $f(\cdot)$ come somma degli inversi dei valori osservati $f(y_1, y_2, \cdots, y_n) = \sum_{i=1}^{n} \frac{1}{y_i}$ da cui si ottiene, adattando i simboli agli strati *ijk*:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{n_{ijk}} \frac{1}{\pi^*_{p;ijk}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{n_{ijk}} \frac{1}{\overline{\pi}^*_p} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{ijk}}{\overline{\pi}^*_p} \quad \Leftrightarrow \quad \overline{\pi}^*_p = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{ijk}}{\pi^*_{p;ijk}}}$$

dove $\overline{\pi}^*_p$ è la media armonica delle probabilità di selezione per i vari strati *ijk*. Il peso normalizzato a uno per ogni strato *ijk* sarà dato dal rapporto tra i pesi effettivi finali $\pi^*_{p;ijk}$ e il peso medio dato dall'inverso della media armonica, $1/\overline{\pi}^*_p$. Allora, il peso normalizzato a uno, w^*_{ijk}, che rispetta entrambi i vincoli (a) e (b) diventa

$$w^*_{ijk} = \frac{\overline{\pi}^*_p}{\pi^*_{p;ijk}} = \frac{1}{\pi^*_{p;ijk}} \; \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{ijk}}{\pi^*_{p;ijk}}}. \tag{20}$$

Si può mostrare che i pesi w^*_{ijk} sono dati dal rapporto tra i pesi degli strati rispetto alla popolazione totale di riferimento e i pesi degli strati nel campione rispetto alla dimensione totale del campione: $w^*_{ijk} = W_{ijk}/w_{ijk} = (N_{ijk}/N):(n_{p;ijk}/n)$; infatti,

$$w^*_{ijk} = \frac{N_{ijk}}{n_{p;ijk}} \frac{n}{N} = \frac{1}{\pi^*_{p;ijk}} \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{p;ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} N_{ijk} \frac{n_{p;ijk}}{n_{p;ijk}}} = \frac{1}{\pi^*_{p;ijk}} \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{p;ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{p;ijk}}{\pi^*_{p;ijk}}}.$$

Si noti che i pesi w^*_{ijk} possono alterare completamente la struttura delle dimensioni campionarie per strato rispetto al campione effettivo. Nella Tabella 11 sono riportati i pesi normalizzati all'unità per il Comune di Modena, dove si può notare che i pesi sono vicini all'unità perché si tratta di una allocazione proporzionale. Nella Tabella 12 sono riportati i pesi normalizzati all'unità relativi alla popolazione dei soggetti residenti.

Tabella 11 – Pesi relativi al numero di famiglie ($N_{MO;ijk}$) del Comune di Modena, normalizzati all'unità (w^*_{ijk}) e riferiti al 31/12/2001, per numero di componenti la famiglia, per classi di età e per genere del capofamiglia

Numero di Componenti	**Genere**	**Classi di età del capofamiglia**				
		<=34 anni	**35-49 anni**	**50-64 anni**	**65-74 anni**	**>=75 anni**
1 componente	M	1,5118	0,8864	1,4074	0,8894	0,9067
	F	0,8733	0,9239	1,0067	0,9480	1,5310
2 componenti	M	1,1197	0,9496	0,9752	0,9390	0,9435
	F	0,9428	0,8742	0,9370	0,9931	0,8203
3 componenti	M	0,9448	0,9547	0,9631	0,9273	0,9305
	F	1,0245	0,8880	0,9922	1,9906	0,8579
4 componenti e piú	M	0,9986	1,0649	0,9426	0,9227	0,7724
	F	0,8346	0,9668	1,0109	1,2519	0,7737

Tabella 12 – Pesi relativi al numero di soggetti residenti ($P_{MO;ijk}$) del Comune di Modena, normalizzati all'unità ($w^*_{P;ijk}$) e riferiti al 31/12/2001, per numero di componenti la famiglia, per classi di età e per genere del capofamiglia

Numero di Componenti	**Genere**	**Classi di età del capofamiglia**				
		<=34 anni	**35-49 anni**	**50-64 anni**	**65-74 anni**	**>=75 anni**
1 componente	M	1,1269	0,5978	1,2813	0,8997	0,9172
	F	0,6425	0,7870	0,8986	0,9590	1,5487
2 componenti	M	1,1327	0,9072	0,9865	0,9376	0,9545
	F	0,8976	0,8843	0,9890	1,0819	0,8735
3 componenti	M	0,9299	0,9746	0,9743	1,0005	1,0589
	F	0,9566	0,8693	1,0037	3,0205	0,8679
4 componenti e piú	M	1,0901	1,1497	1,0529	1,0037	0,9039
	F	0,9341	1,1122	1,0261	1,4316	0,7656

4.3.2. Normalizzazione nella Provincia di Modena

La normalizzazione dei pesi all'unità, nella Provincia di Modena, si esegue con un procedimento analogo al precedente; ossia, i pesi w^*_{dc} (o $w^*_{P;dc}$) sono dati dal rapporto tra i pesi degli strati rispetto alla popolazione totale di riferimento e i pesi degli strati nel

campione rispetto alla dimensione totale del campione:

$$w^*_{dc} = \frac{N^*_{dc}}{c_d n_{dc}} \quad \frac{n}{N} = \frac{\bar{\pi}}{\pi^*_{dc}} \tag{21}$$

e sostituendo N con P si ottengono i pesi $w^*_{P;dc}$ da usare quando si trattano gli individui e non le famiglie. Nella Tabella 13 sono riportati i pesi w^*_{dc} e $w^*_{P;dc}$. Si può osservare che valori molto grandi, rispetto a uno, si ottengono proprio per quei Comuni della zona montana che hanno presentato piú problemi in fase di rilevazione.

Tabella 13 – Pesi (w^*_{dc}) relativi al numero di famiglie (N_{dc}) della Provincia di Modena, normalizzati all'unità e pesi ($w^*_{P;dc}$) relativi al numero di soggetti residenti (P_{dc}) nella Provincia di Modena, riferiti al 31/12/2001, per i Comuni inclusi nel campione della Provincia suddivisi per distretto sociosanitario

	Comune	w^*_{dc}	$w^*_{P;dc}$		Comune	w^*_{dc}	$w^*_{P;dc}$
D1	**Carpi**	0,9759	0,9813	**D5**	**Pavullo nel Frignano**	0,8875	0,7875
	Soliera	2,3258	2,4304		Polinago	8,0540	9,0498
D2	**Mirandola**	0,7507	0,8148	**D6**	**Vignola**	0,6242	0,5994
	Cavezzo	2,7618	2,7261		Castelnuovo Rangone	1,3530	1,3201
	Finale Emilia	1,1926	1,1551		Guiglia	27,0596	33,6624
D3	**Modena**	0,6170	0,6143				
D4	**Sassuolo**	1,0008	1,0071	**D7**	**Castelfranco Emilia**	1,1212	1,1792
	Fiorano Modenese	1,4575	1,2354		Nonantola	3,1106	2,7696
	Formigine	0,7539	0,6712				
	Montefiorino	14,5750	19,3538				

4.4. Varianza della stima del reddito totale

Il totale della caratteristica Y è espresso nella (11); il suo stimatore derivato dalla (12) è

$$\hat{Y} = \sum_{d=1}^{D} \hat{Y}_d = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \sum_{i=1}^{n_{dc}} w_{dc} y_{dci} \tag{22}$$

dove y_{dci} è il reddito dell'i-esima unità campionaria, nel c-esimo Comune del d-esimo distretto. Ogni metodo di stima campionaria assume il principio che le unità incluse nel campione rappresentano anche le altre unità della popolazione che non sono state selezionate; ossia, nel caso in oggetto, ogni unità nel campione rappresenta le $(w_{dc} - 1)$ unità della popolazione che non sono state selezionate. Nello stadio iniziale del processo di elaborazione dei dati si useranno i pesi già calcolati. Per valutare la varianza dello stimatore del totale si distinguono i seguenti casi.

Negli strati AR di un disegno di campionamento a grappoli, dove le USS (famiglie) sono selezionate senza reimmissione e con probabilità uguali, si ottengono stime della varianza campionaria che risultano corrette e lo stimatore $\hat{Y}$ è dato da $\hat{Y}_{AR} = \sum_{d=1}^{D} N_{d1} \bar{y}_{d1}$ con varianza pari a

$$V\left(\hat{Y}_{AR}\right) = \sum_{d=1}^{D} N_{d1}^2 \frac{s^2_{2;d1}}{n_{d1}} \left(1 - f_{2;d1}\right), \tag{23}$$

dove $s^2_{2;d1}$ è la varianza campionaria del reddito e $f_{2;d1} = n_{d1}/N_{d1}$ è la frazione di unità nel campione del d-esimo distretto.

Negli strati NAR con un solo Comune selezionato si ha $c_d = 1$ e, quindi, $c = 2$. Nel metodo di Hansen e Hurwitz (1943), adottato per la selezione, lo stimatore del totale si può ottenere dallo stimatore della media campionaria (Cochran, 1977, p. 295)

$$\hat{Y}_d^* = N_d^* \hat{\bar{y}}_d^* = N_d^* \bar{y}_{d2}, \tag{24}$$

dove, per i Comuni del *d*-esimo distretto, l'asterisco (*) indica sempre il riferimento allo strato NAR, $\hat{\bar{y}}_d^*$ è lo stimatore della media nello strato NAR, $\bar{y}_{d2}$ è la media osservata nell'unico campione del Comune campione nello strato NAR. La sua varianza è data da

$$V\left(\hat{Y}_d^*\right) = N_d^* \left[\sum_{c=1}^{C_d^*} \left(N_{dc} - n_{dc}\right) \frac{S_{2;dc}^2}{n_{dc}} + \sum_{c=1}^{C_d^*} N_{dc} \left(\bar{Y}_{dc} - \bar{\bar{Y}}_d^*\right)^2 \right] \tag{25}$$

dove $S_{2;dc}^2$ è la varianza e $\bar{Y}_{dc}$ è la media della popolazione del *c*-esimo Comune NAR del *d*-esimo distretto, mentre $\bar{\bar{Y}}_d$ la media totale dello strato NAR del *d*-esimo distretto. Senza informazioni sulla popolazione delle UPS non è possibile calcolare tale espressione.

Negli strati NAR con due o piú Comuni campione, lo stimatore del totale è dato dalla (22), con una varianza

$$V\left(\hat{Y}_d^*\right) = \sum_{c=1}^{C_d^*} \sum_{c' \neq c}^{C_d^*} \left(\frac{\pi_{dc}\pi_{dc'}}{\pi_{dcc'}} - 1 \right) \left(\frac{\hat{Y}_{dc}}{\pi_{dc}} - \frac{\hat{Y}_{dc'}}{\pi_{dc'}} \right)^2 + \sum_{c=1}^{C_d^*} \frac{N_{dc}^2}{\pi_{dc}} \frac{S_{2;dc}^2}{n_{dc}} \left(1 - f_{2;dc}\right), \tag{26}$$

dove $\pi_{dcc'}$ è la probabilità di inclusione di secondo ordine, $\hat{Y}_{dc}$ è sempre lo stimatore di Horvitz-Thompson del totale. La sua stima campionaria si ottiene facilmente.

4.5. Post-stratificazione

La popolazione può essere suddivisa in base a alcuni caratteri; per esempio, nella fase iniziale dell'analisi, i dati facilmente accessibili sono la classe di età $(j = 1, \ldots, J)$ e il genere $(k = 1,\ 2\ (= K))$ degli individui; si usano gli indici già definiti in precedenza. La loro conoscenza consente di costruire $J \times K$ post-strati. In ogni distretto *d* e in ogni strato di Comuni (AR o NAR), si può costruire uno stimatore della popolazione che ricade in ogni post-strato *jk* sicché si ottiene

$$\hat{Y}_d = \sum_{j=1}^{J} \sum_{k=1}^{K} N_{d1jk} \bar{y}_{d1jk} + \sum_{j=1}^{J} \sum_{k=1}^{K} N_{djk}^* \bar{y}_{djk}^*. \tag{27}$$

I soggetti inclusi nel campione avranno, in questa procedura, dei nuovi pesi che si ottengono immediatamente dall'espressione precedente in una forma simile alla (17):

$$w_{PS;d1jk} = \frac{N_{d1jk}}{n_{d1jk}} \qquad \text{e} \qquad w_{PS;djk}^{NAR} = \frac{N_{djk}^*}{n_{djk}^*}. \tag{28}$$

dove n_{djk}^* è la dimensione del campione nel post-strato *jk* del *d*-esimo distretto, relativamente ai Comuni campione NAR. Nei piani di campionamento complessi, la varianza degli stimatori post-stratificati presenta una espressione abbastanza complicata (Cochran, 1977; Cicchitelli, Herzel, Montanari, 1997): sia per gli strati AR, stimati con il primo termine di secondo membro della (27); sia per gli strati NAR, stimati con il secondo termine della (27). Per semplificare, non si riportano per esteso, ma per una applicazione nelle indagini complesse si vedano: Falorsi, Falorsi, e Russo (1992); Falorsi e Russo (1992); Little (1993); Zhang (2000).

4.6. Stimatori di ponderazione vincolata

Si consideri sempre la stima del totale del reddito espressa, analogamente all'equazione precedente, come

$$\hat{Y} = \sum_{d=1}^{D} \hat{Y}_d = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \sum_{i=1}^{n_{dc}} W_{dci} y_{dci} \tag{29}$$

dove W_{dci} è il *peso finale* da attribuire a tutti i componenti della *i*-esima famiglia del *c*-esimo Comune del *d*-esimo distretto; in breve, «famiglia dci». La determinazione di W_{dci} dovrebbe conseguire gli obiettivi: (1) di ottenere stime coerenti per famiglie e individui, attribuendo a ciascuna famiglia dci e a tutti i suoi componenti lo stesso peso finale W_{dci}; (2) di correggere la distorsione per le mancate risposte; (3)di produrre stime campionarie di totali di alcune importanti variabili ausiliarie coincidenti con i loro valori noti nella popolazione di riferimento.

L'accesso ai dati del 14° Censimento generale della popolazione del 21 ottobre 2001, che è prossima alla data adottata in precedenza, consentirà di ottenere una distribuzione della popolazione simile a quella operata nel Comune di Modena: per numero di componenti la famiglia, per classe di età, e per genere. Il numero di strati è assai elevato; quindi, ci si può limitare alla classe di età e al genere per ogni strato AR e NAR di ciascun distretto, come nella procedura adottata dall'Istat (2002b,c). In tal caso, si possono utilizzare anche altre fonti, come le statistiche sulla popolazione della Provincia di Modena (Benassi, Zoda, 2002). Sia ${}_lX$, con $(l = 1, \ldots, L)$, il totale noto della *l*-esima variabile ausiliaria, allora dovrà risultare

$${}_lX_d = {}_l\hat{X}_d = \sum_{c=1}^{c_d} \sum_{i=1}^{n_c} {}_lx_{dci} \tag{30}$$

dove ${}_lx_{dci}$ è il valore che la *l*-esima variabile ausiliaria assume nella famiglia *dci*. Per ogni distretto si avrà una stima per il Comune AR e una stima per lo strato NAR.

I pesi base già ottenuti, tramite il prodotto dei pesi diretti con i fattori correttivi per mancata risposta totale, devono essere, quindi, ancora corretti per soddisfare le condizioni di uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime campionarie, a livello di distretto. Il fattore di correzione è determinato, in genere, risolvendo un problema di minimo vincolato: si minimizza una funzione della distanza tra i pesi finali e i pesi di base in modo che siano soddisfatte le uguaglianze tra i valori dei totali noti della popolazione e le corrispondenti stime campionarie. Gli stimatori sono detti di ponderazione vincolata (*calibration estimators*) e costituiscono una classe generale (Falorsi, Falorsi, 1995). Si è dimostrato (Deville, Särndal, 1992), però, che tutti gli stimatori di ponderazione vincolata convergono allo stimatore di regressione generalizzata, che si ottiene quando si adotta una funzione di distanza euclidea.

4.6.1. Stimatore di regressione generalizzata

Lo stimatore di ponderazione vincolata è definito sulla base di una funzione di distanza; per esempio, la funzione adottata dall'Istat (2002b,c) è di tipo logaritmico troncato. Lo stimatore $\hat{V}(\hat{Y}_d)$ della varianza $V(\hat{Y}_d)$ non è una funzione lineare dei dati campionari, ma si può ottenere un'espressione lineare approssimata con il metodo proposto da Woodruff (1971), che usa uno sviluppo in serie di Taylor, e ricavare da quella la varianza (Cicchitelli, Herzel, Montanari, 1997, pp. 234-242). Si possono utilizzare gli stimatori di regressione generalizzata perché tutti gli stimatori di ponderazione vincolata conver-

gono a essi, quando adottano una funzione di distanza euclidea. L'espressione lineare dell'addendo dello stimatore è data da

$$\hat{Y} \cong \hat{Z} = \sum_{d=1}^{D} \hat{Z}_d = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \hat{Z}_{dc} = \sum_{d=1}^{D} \sum_{c=1}^{c_d} \sum_{i=1}^{n_{dc}} W_{dci} Z_{dci} \,, \tag{30}$$

dove Z_{dci} è la variabile linearizzata espressa dalla relazione

$$Z_{dci} = Y_{dci} - X'_{dci}\beta \,, \tag{31}$$

nella quale $X'_{dci} = \left({}_1 x_{dci}, \ldots, {}_K x_{dci}\right)'$ è il vettore contenente i valori delle K variabili ausiliarie relativi alla generica famiglia dci e β è il vettore dei coefficienti di regressione della variabile di interesse Y sulle K variabili ausiliarie X. Lo stimatore della varianza dello stimatore del totale, $\hat{Y}$, sarà dato da

$$\hat{V}\left(\hat{Y}\right) = \hat{V}\left(\hat{Z}\right) = \sum_{d=1}^{D} \hat{V}\left(\hat{Z}_d\right) = \sum_{d=1}^{D} \hat{V}\left(\hat{Z}_{d1}\right) + \sum_{d=1}^{D} \sum_{c=1}^{c_d^*} \hat{V}\left(\hat{Z}_{dc}\right). \tag{32}$$

La stima della varianza della stima $\hat{Y}$ risulta espressa dalla somma di due quantità: la somma delle stime delle varianze delle stime negli strati AR e la somma delle stime delle varianze delle stime negli strati NAR. Nella fase iniziale della elaborazione dei dati si useranno i pesi calcolati in precedenza. Una specificazione piú dettagliata del procedimento di calcolo degli stimatori di ponderazione vincolata sarà l'oggetto di una nota successiva.

4.6.2. Livello di precisione delle stime

La valutazione della variabilità campionaria delle stime prodotte da una indagine si possono esprimere sia con l'errore assoluto, sia con l'errore relativo o coefficiente di variazione. L'errore assoluto si può valutare con la deviazione standard della stima. Sia $\hat{Y}_d$ lo stimatore con $V\left(\hat{Y}_d\right)$, allora l'entità dell'accuratezza della stima si può ottenere da

$$\hat{\sigma}\left(\hat{Y}_d\right) = \sqrt{V\left(\hat{Y}_d\right)} \,; \tag{33}$$

mentre la stima del corrispondente errore relativo è data da

$$\hat{\varepsilon}\left(\hat{Y}_d\right) = \frac{\hat{\sigma}\left(\hat{Y}_d\right)}{\hat{Y}_d} \,. \tag{34}$$

La valutazione degli errori di campionamento espressi dalla (33) o dalla (34), commessi nell'indagine in oggetto, si possono ottenere dalle espressioni precedenti per gli strati AR; inoltre, il disegno di campionamento adottato consente di ottenere stime della varianza campionaria che sono corrette. Negli strati NAR, si possono ottenere stime corrette della varianza degli stimatori, se si seguono procedimenti che semplificano il loro calcolo (Fabbris, 1989; Särndal, Swensson, Wretman, 1992); per esempio, in ogni strato: (1) vi sono due o piú UPS (Comuni); (2) le UPS sono scelte con reimmissione. La prima condizione non è sempre soddisfatta perché vi sono strati con un solo Comune campione (distretti: D1-Carpi, D5-Pavullo nel Frignano, e D7-Castelfranco Emilia); si può rimediare con la tecnica di *collassamento degli strati*, ma la limitata entità territoriale del campione non consente di applicarla in modo totalmente appropriato e, inoltre, comporta una inflazione della varianza di campionamento effettiva. La seconda non è ugualmente soddisfatta perché le selezioni delle UPS sono avvenute senza reimmissione e ne consegue ancora una sovrastima della varianza che diminuisce con il diminuire della frazione di campionamento di ciascun strato NAR fino a diventare trascurabile per

frazioni molto piccole. Anche le valutazioni degli errori campionari saranno l'oggetto della nota successiva.

4.7. Stime del reddito individuale per alcuni dominî di studio

Il reddito è la variabile di maggiore interesse nell'indagine sulle condizioni economiche e sociali delle famiglie nella provincia di Modena. La sua stima è eseguita con l'uso dei pesi derivati in precedenza; ma i pesi riportati nella Tabella 10 possono destare qualche perplessità perché in alcuni distretti sociosanitari vi sono Comuni, e quindi famiglie, che presentano pesi con valori assai elevati e un numero basso di unità statistiche. Si è deciso di adottare, pertanto, una post-stratificazione che, in base ai dati disponibili e alle dimensioni del campione per strato, è stata eseguita: per i sette distretti sociosanitari ($d = 1, \cdots, D$); per otto classi di età ($j = 1, \cdots, J$) di dieci anni l'una, definite in modo da non avere celle (domíni) con frequenze nulla $(0-9,\ 10-19, \cdots, \geq 70)$; per due valori del genere dell'individuo, $k = 1,2 (= K)$.

Negli strati determinati dalla post-stratificazione i pesi possono essere calcolati come indicato nel paragrafo precedente (§4.5); allora, si hanno gli *stimatori post-stratificati semplici*. Nei piani di campionamento complessi, le probabilità di selezione delle unità statistiche variano in ciascun post-strato sia per il piano di stratificazione originario, sia per raggruppamenti di unità. In ogni post-strato djk, lo stimatore del totale, Y_{djk}, è dato dallo stimatore di Horvitz-Thompson; tuttavia, si suggerisce di usare il cosiddetto stimatore di Hajek (Zhang, 2000):

$$\hat{Y}_{djk} = N_{djk}\left(\frac{\widetilde{Y}_{djk}}{\widetilde{N}_{djk}}\right) = \widetilde{R}_{djk}\widetilde{Y}_{djk} = \widetilde{R}_{djk}\sum_{l \in \varsigma_{djk}} w_{djkl}\, y_{djkl}$$

dove $\widetilde{Y}_{djk}$ è la stima del totale e $\widetilde{N}_{djk}$ è la stima della popolazione nel post-strato djk (entrambe ottenute con i pesi derivati dalle probabilità di selezione e aggiustati), ς_{djk} indica l'insieme di unità statistiche del post-strato djk; si applica, in definitiva, uno stimatore di rapporto all'interno di ciascun post-strato. Alcune giustificazioni per tale procedura sono esposte in Särndal, Swensson, e Wretman (1992, §5.7). I pesi per gli stimatori di Hajek, $w_{PH;djk}$, si possono determinare con la relazione seguente:

$$w_{PH;djk} = \sum_{l \in \varsigma_{djk}} \widetilde{R}_{djk}\, w_{djkl}\ . \qquad (35)$$

Nella Tabella 14 sono riportate, per gli strati campionarî del Comune di Modena, le stime del reddito medio individuale ottenute senza e con diversi pesi. Si può notare che le medie non pesate sono uguali alle medie pesate con $w_{P;ijk}$; ciò perché si tratta di un disegno autoponderante. Non si sono riportate le stime ottenute pesando con w_{ijk} perché sono ancora coincidenti con le medie non pesate. La maggiore differenza tra le stime si riscontra con i pesi della post-stratificazione semplice, $w_{PS;djk}$, perché questi non incorporano le probabilità di selezione, mentre $w_{PH;djk}$ ne tiene conto e riduce, perciò, l'entità degli scarti. Le differenze sono, però, trascurabili statisticamente perché l'errore relativo, calcolato *ex post* con la (3), varia da strato a strato assumendo valori piú elevati del 10-15%. Solo nel campione complessivo è dell'ordine del 5%.

Tabella 14 – Dimensione della popolazione e del campione, errore standard (ES) della media campionaria, e reddito medio individuale non pesato, con i pesi individuali $w_{P;ijk}$, con pesi da post-stratificazione semplice $w_{PS;djk}$, e con lo stimatore di Hajek $w_{PH;djk}$ per numero di componenti la famiglia, per classi di età e genere del capofamiglia a Modena

Classi di età	Comp. Fam.	N	n	ES della Media	No pesato Media	$w_{P;ijk}$ Media	$w_{PS;djk}$ Media	$w_{PH;djk}$ Media
Uomo								
18-34 a.	**1 C.**	2722	19	3751,776	38227,057	38227,057	39456,605	39533,158
	2 C.	3744	26	3093,509	29192,356	29192,356	29239,472	29305,243
	3 C.	4374	37	3987,034	23331,554	23331,554	24708,050	25232,107
	≥4 C.	4019	29	4961,585	21722,435	21722,435	23718,715	24358,396
35-49 a.	**1 C.**	2736	36	4651,640	39216,315	39216,315	38091,675	38367,621
	2 C.	4152	36	6454,588	42899,658	42899,658	43544,906	43654,256
	3 C.	13629	110	3465,812	32926,084	32926,084	32488,322	33086,524
	≥4 C.	24847	170	2345,954	20663,546	20663,546	20330,983	20969,804
50-64 a.	**1 C.**	1629	10	9552,472	49052,563	49052,563	49137,780	49069,796
	2 C.	8528	68	3991,501	41474,047	41474,047	41056,826	41041,419
	3 C.	15978	129	3695,061	37261,899	37261,899	35305,990	35279,260
	≥4 C.	16732	125	3476,009	29667,524	29667,524	28031,107	27963,654
64-74 a.	**1 C.**	915	8	3142,942	27630,438	27630,438	27709,272	27730,804
	2 C.	9178	77	4303,816	40249,877	40249,877	39849,192	40185,130
	3 C.	5724	45	3452,699	33332,546	33332,546	32423,510	32813,307
	≥4 C.	3190	25	5978,856	25048,368	25048,368	24614,992	24878,881
>=75 a.	**1 C.**	1166	10	9713,021	49550,400	49550,400	49550,400	49550,400
	2 C.	7766	64	3977,968	32983,806	32983,806	31930,147	33009,297
	3 C.	2154	16	6368,327	33255,107	33255,107	32674,559	33513,951
	≥4 C.	1379	12	3705,960	20700,635	20700,635	20248,655	20495,942
Donna								
18-34 a.	**1 C.**	1797	22	3954,172	36711,183	36711,183	35898,494	35998,611
	2 C.	1940	17	3315,222	28806,785	28806,785	29226,180	29297,178
	3 C.	1581	13	5150,534	22947,495	22947,495	23520,681	24102,598
	≥4 C.	1425	12	7236,705	19450,522	19450,522	19725,718	20146,401
35-49 a.	**1 C.**	1901	19	2838,682	36921,015	36921,015	36792,420	36787,875
	2 C.	3148	28	4769,777	28496,081	28496,081	27063,800	27530,502
	3 C.	3426	31	3989,061	21799,384	21799,384	21090,005	21582,364
	≥4 C.	3252	23	6683,039	24649,265	24649,265	23611,910	24281,905
50-64 a.	**1 C.**	1942	17	4670,368	38347,146	38347,146	38457,501	38422,838
	2 C.	2892	23	9448,793	57031,617	57031,617	55637,821	55818,821
	3 C.	1914	15	4451,775	29711,784	29711,784	28431,448	28505,103
	≥4 C.	1174	9	3686,570	9148,167	9148,167	8754,797	8760,495
64-74 a.	**1 C.**	2926	24	4555,080	39405,619	39405,619	39966,488	39664,366
	2 C.	1788	13	4447,504	34142,346	34142,346	33691,346	33900,087
	3 C.	768	2	9608,915	24578,885	24578,885	24972,024	24884,173
	≥4 C.	728	4	6232,744	18577,750	18577,750	18322,238	18416,149
>=75 a.	**1 C.**	5316	27	3296,804	34881,050	34881,050	34881,050	34881,050
	2 C.	2110	19	3146,642	28131,960	28131,960	28032,416	27925,707
	3 C.	993	9	7980,302	34987,523	34987,523	32679,992	33173,024
	≥4 C.	876	9	9183,733	31825,496	31825,496	32047,668	31763,418
Totale		**176459**	**1388**	**861,336**	**31975,703**	**31615,965**	**31399,808**	**31399,230**

Tabella 15 – Dimensione della popolazione e del campione, errore standard (ES) della media campionaria, e reddito medio individuale non pesato, con i pesi famigliari $w_{\bullet}$, con i pesi individuali $w_{P;\bullet}$, con pesi da post-stratificazione semplice $w_{PS;jk}$, e con lo stimatore di Hajek $w_{PH;jk}$ per classi di età, per genere, e per distretto sociosanitario

DS	Genere	Classi di età	N	n	SE(M)	No peso Media	$w_{\bullet}$ Media	$w_{P;\bullet}$ Media	$w_{PS;djk}$ Media	$w_{PH;djk}$ Media
D1	**Uomo**	**0-9 a.**	4079	11	0	0	0	0	0	0
		10-19 a.	3742	16	1671,27	3574,96	3574,96	3574,96	3574,96	3574,96
		20-29 a.	6129	17	4304,92	20021,77	22688,27	22830,99	20021,77	22830,99
		30-39 a.	8366	28	4443,29	44256,67	45535,70	45606,96	44256,67	45606,96
		40-49 a.	6614	20	17165,84	70169,03	63228,00	62885,94	70169,03	62885,94
		50-59 a.	6169	34	4775,54	54146,93	55090,00	55142,28	54146,93	55142,28
		60-69 a.	5270	24	7920,11	48339,83	46046,51	45922,33	48339,83	45922,33
		≥70 a.	5043	23	4255,09	36575,32	35548,74	35500,95	36575,32	35500,95
	Donna	**0-9 a.**	3894	10	0	0	0	0	0	0
		10-19 a.	3616	16	575,00	575,00	401,46	393,44	575,00	393,44
		20-29 a.	5881	21	2392,95	12559,88	13484,88	13539,59	12559,88	13539,59
		30-39 a.	7779	28	3327,41	25949,68	25128,66	25082,91	25949,68	25082,91
		40-49 a.	6566	23	4383,81	26484,65	24122,94	23996,15	26484,65	23996,15
		50-59 a.	6597	34	3179,85	24648,23	23779,65	23731,5	24648,23	23731,50
		60-69 a.	5838	25	3171,27	21404,11	20430,91	20375,13	21404,11	20375,13
		≥70 a.	8167	36	3225,98	30313,44	28496,20	28403,88	30313,44	28403,88
D2	**Uomo**	**0-9 a.**	3484	12	0	0	0	0	0	0
		10-19 a.	3347	16	755,67	755,67	901,87	867,50	755,67	867,50
		20-29 a.	5079	12	3795,13	28122,18	26939,52	27031,04	28122,18	27031,04
		30-39 a.	6602	23	3731,04	39181,71	35305,58	35645,86	39181,71	35645,86
		40-49 a.	5604	26	4480,32	49208,65	50876,39	50940,55	49208,65	50940,55
		50-59 a.	4904	15	5454,55	52107,18	49718,11	50073,67	52107,18	50073,67
		60-69 a.	4508	19	4702,74	36852,72	36751,32	36962,12	36852,72	36962,12
		≥70 a.	4925	22	3897,77	36065,91	36241,68	36274,77	36065,91	36274,77
	Donna	**0-9 a.**	3117	9	0	0	0	0	0	0
		10-19 a.	3147	13	1397,07	1443,11	795,16	849,33	1443,11	849,33
		20-29 a.	4950	17	3439,46	15053,71	13901,74	14049,14	15053,71	14049,14
		30-39 a.	6209	21	4569,10	26025,97	24176,45	24511,44	26025,97	24511,44
		40-49 a.	5373	29	2393,36	21431,13	21491,78	21478,64	21431,13	21478,64
		50-59 a.	5067	20	3580,81	23916,05	18042,62	18154,37	23916,05	18154,37
		60-69 a.	4914	22	3161,82	21826,85	23392,16	23405,84	21826,85	23405,84
		≥70 a.	7803	22	2578,37	18663,43	20336,57	20375,65	18663,43	20375,65
D3	**Uomo**	**0-9 a.**	7556	70	0	0	0	0	0	0
		10-19 a.	7015	50	318,81	384,16	413,44	424,43	384,16	424,43
		20-29 a.	11004	64	2122,83	18316,99	19757,06	18766,12	18316,99	18766,12
		30-39 a.	15507	112	3170,99	49368,07	49504,58	49760,30	49368,07	49760,30
		40-49 a.	12472	104	2640,76	52083,06	52222,13	52158,73	52083,06	52158,73
		50-59 a.	11467	94	4586,71	67180,75	66782,94	66866,42	67180,75	66866,42
		60-69 a.	10038	86	4149,17	56680,84	56558,40	56650,75	56680,84	56650,75
		≥70 a.	10643	83	3789,01	48271,79	48390,71	48272,53	48271,79	48272,53

(*continua*)

Tabella 15 – Dimensione della popolazione e del campione, errore standard (ES) della media campionaria, e reddito medio individuale non pesato, con i pesi famigliari $w_{\bullet}$, con i pesi individuali $w_{P;\bullet}$, con pesi da post-stratificazione semplice $w_{PS;djk}$, e con lo stimatore di Hajek $w_{PH;djk}$, per classi di età, per genere, e per distretto sociosanitario

(*continua*)

DS	Genere	Classi di età	N	n	SE(M)	No peso Media	$w_{\bullet}$ Media	$w_{P;\bullet}$ Media	$w_{PS;djk}$ Media	$w_{PH;djk}$ Media
D3	**Donna**	**0-9 a.**	7215	56	22,32	34,17	30,69	31,01	34,17	31,01
		10-19 a.	6336	46	215,69	460,85	433,07	394,44	460,85	394,44
		20-29 a.	10666	72	1554,49	13223,30	13266,47	12880,49	13223,30	12880,49
		30-39 a.	14403	129	1846,34	29439,97	29029,79	28572,95	29439,97	28572,95
		40-49 a.	12451	106	1608,82	30313,04	29868,92	29597,86	30313,04	29597,86
		50-59 a.	12289	104	2789,42	33083,69	33008,43	32861,92	33083,69	32861,92
		60-69 a.	11633	96	2169,34	27578,79	27703,58	27801,10	27578,79	27801,10
		≥70 a.	17318	116	1786,72	25641,56	26728,86	26634,73	25641,56	26634,73
D4	**Uomo**	**0-9 a.**	5586	20	0	0	0	0	0	0
		10-19 a.	5621	32	656,83	843,60	660,43	642,03	843,60	642,03
		20-29 a.	8082	34	2455,25	18306,89	18455,17	18300,41	18306,89	18300,41
		30-39 a.	10099	30	4477,00	47822,73	48316,74	47912,54	47822,73	47912,54
		40-49 a.	8436	47	2726,84	52894,32	44609,78	42508,19	52894,32	42508,19
		50-59 a.	7165	35	3429,31	52955,22	53106,10	52992,83	52955,22	52992,83
		60-69 a.	6006	14	4052,35	54071,52	54600,77	54198,42	54071,52	54198,42
		≥70 a.	5472	14	3559,60	43521,44	44462,47	43921,04	43521,44	43921,04
	Donna	**0-9 a.**	5404	16	0	0	0	0	0	0
		10-19 a.	5349	31	33,17	40,92	40,25	42,20	40,92	42,20
		20-29 a.	7644	35	2218,77	11737,38	14701,12	15453,47	11737,38	15453,47
		30-39 a.	9354	34	2336,42	27433,15	28319,80	28395,12	27433,15	28395,12
		40-49 a.	8305	49	3111,34	28833,72	25944,87	25104,21	28833,72	25104,21
		50-59 a.	6949	39	4258,45	24825,83	25556,41	25148,77	24825,83	25148,77
		60-69 a.	6154	16	2884,22	29091,14	28929,09	29255,07	29091,14	29255,07
		≥70 a.	8113	22	2790,56	25711,3	18896,53	17984,51	25711,3	17984,51
D5	**Uomo**	**0-9 a.**	1690	8	0	0	0	0	0	0
		10-19 a.	1551	5	0	0	0	0	0	0
		20-29 a.	2295	8	5506,85	21139,73	21139,73	21139,73	21139,73	21139,73
		30-39 a.	3106	8	7763,44	66881,91	66881,91	66881,91	66881,91	66881,91
		40-49 a.	2684	9	13329,65	64665,75	66397,89	66637,34	64665,75	66637,34
		50-59 a.	2379	8	5154,60	41164,85	35649,59	35193,82	41164,85	35193,82
		60-69 a.	2158	5	32622,46	71368,49	44298,90	42739,38	71368,49	42739,38
		≥70 a.	2913	4	5463,88	43887,88	34651,45	33888,17	43887,88	33888,17
	Donna	**0-9 a.**	1533	6	0	0	0	0	0	0
		10-19 a.	1449	7	1678,23	1678,23	7071,87	7718,10	1678,23	7718,10
		20-29 a.	2231	4	2591,82	9011,72	10162,55	10217,72	9011,72	10217,72
		30-39 a.	2882	13	3885,39	28953,40	28953,40	28953,40	28953,40	28953,40
		40-49 a.	2444	10	4453,02	20273,55	21548,55	21671,76	20273,55	21671,76
		50-59 a.	2039	3	21131,30	22363,21	22363,21	22363,21	22363,21	22363,21
		60-69 a.	2240	4	6065,59	21283,71	20842,79	20806,36	21283,71	20806,36
		≥70 a.	4206	2	4355,75	17074,75	17074,75	17074,75	17074,75	17074,75

(*continua*)

Tabella 15 – Dimensione della popolazione e del campione, errore standard (ES) della media campionaria, e reddito medio individuale non pesato, con i pesi famigliari $w_{\bullet}$, con i pesi individuali $w_{P;\bullet}$, con pesi da post-stratificazione semplice $w_{PS;djk}$, e con lo stimatore di Hajek $w_{PH;djk}$, per classi di età, per genere, e per distretto sociosanitario

(*continua*)

DS	Genere	Classi di età	N	n	SE(M)	No peso Media	$w_{\bullet}$ Media	$w_{P;\bullet}$ Media	$w_{PS;djk}$ Media	$w_{PH;djk}$ Media
D6	**Uomo**	**0-9 a.**	3487	3	0	0	0	0	0	0
		10-19 a.	3104	11	2126,95	2130,52	1615,916	1604,43	2130,52	1604,43
		20-29 a.	4959	14	4570,09	20254,76	20560,90	20567,65	20254,76	20567,65
		30-39 a.	6861	20	5706,53	42947,34	42898,32	42897,37	42947,34	42897,37
		40-49 a.	5607	23	5290,62	51085,84	35546,21	34046,61	51085,84	34046,61
		50-59 a.	5024	26	7153,02	65672,23	69583,99	69656,40	65672,23	69656,40
		60-69 a.	4525	20	11402,05	60709,07	51423,99	50565,87	60709,07	50565,87
		≥70 a.	4906	10	4317,28	42708,73	42464,58	42459,27	42708,73	42459,27
D6	**Donna**	**0-9 a.**	3154	10	9,00	13,48	16,16	16,22	13,48	16,22
		10-19 a.	2913	15	1770,24	1822,22	2968,56	2994,28	1822,22	2994,28
		20-29 a.	4737	15	3979,83	16103,71	14955,33	14934,60	16103,71	14934,60
		30-39 a.	6375	21	2427,92	27009,34	12942,84	11196,01	27009,34	11196,01
		40-49 a.	5315	32	4299,31	34685,58	37079,24	37128,13	34685,58	37128,13
		50-59 a.	4955	28	4192,86	22933,57	25096,20	25138,50	22933,57	25138,50
		60-69 a.	4620	19	1937,05	14112,26	13713,57	13713,26	14112,26	13713,26
		≥70 a.	6961	9	4461,08	22570,19	23405,99	23423,67	22570,19	23423,67
D7	**Uomo**	**0-9 a.**	2836	15	1,35	1,35	0,85	0,93	1,35	0,93
		10-19 a.	2461	6	4155,64	6310,60	3965,24	4353,39	6310,60	4353,39
		20-29 a.	4032	11	5499,78	20328,82	17687,59	18080,40	20328,82	18080,40
		30-39 a.	5613	16	5612,55	42913,27	42174,42	42306,97	42913,27	42306,97
		40-49 a.	4445	8	6052,45	56530,98	54654,53	54953,49	56530,98	54953,49
		50-59 a.	3645	12	11239,92	61153,77	73254,48	71395,94	61153,77	71395,94
		60-69 a.	3250	5	6153,98	38253,51	44072,63	43165,83	38253,51	43165,83
		≥70 a.	3288	9	2669,04	32648,21	33683,03	33511,77	32648,21	33511,77
	Donna	**0-9 a.**	2629	7	0	0	0	0	0	0
		10-19 a.	2306	6	42,68	67,43	99,57	93,28	67,43	93,28
		20-29 a.	3842	5	8296,73	13574,46	8306,911	9127,76	13574,46	9127,76
		30-39 a.	5173	18	3410,11	20871,92	19603,01	19813,02	20871,92	19813,02
		40-49 a.	4155	15	8165,86	35379,83	31083,15	31794,24	35379,83	31794,24
		50-59 a.	3549	8	28854,83	51337,81	45186,92	46166,89	51337,81	46166,89
		60-69 a.	3453	13	4889,96	26046,90	23594,78	23957,35	26046,90	23957,35
		≥70 a.	4800	8	5060,34	13741,03	14767,55	14620,47	13741,03	14620,47
Totale			639315	3062	545,44	29936,53	28696,68	28500,38	29086,89	28183,74

Nella Tabella 15 sono riportate, per i diversi post-strati djk della Provincia di Modena, le stime del reddito medio individuale ottenute senza e con diversi pesi. L'Errore Standard (ES) per la media campionaria è dato da $s_{2;djk} \big/ \sqrt{n_{djk}}$ e una forma analoga vale per la Tabella 14, nella quale cambiano gli indici. Si può notare, ora, che le medie non pesate non sono piú uguali alle medie pesate sia con $w_{\bullet}$, sia con $w_{P;\bullet}$ perché il disegno non è piú autoponderante; pertanto, è nel distretto di Modena che si osservano le minori differenze. Si noti che si è omessa la specificazione degli indici perché

sono diversi per il Comune di Modena e per il resto della Provincia. Le differenze piú elevate tra le stime si riscontrano nei distretti di Pavullo nel Frignano (D5) e Vignola (D6) sia perché hanno diversi Comuni montani, dove si sono avute maggiori difficoltà di rilevazione, sia perché le dimensioni del campione sono piuttosto piccole. La classe di età 40-49 anni è quella che presenta quasi sempre le differenze piú elevate; ma anche nelle due classi adiacenti si hanno scarti di rilievo, specie nei distretti D5, D6, e D7. I pesi della post-stratificazione semplice, $w_{PS;djk}$, producono stime uguali a quelle non pesate, mentre tutti gli altri pesi tendono a produrre stime pressoché uguali tra loro e, a volte, diverse da quelle della post-stratificazione semplice. Le differenze sono, in realtà, trascurabili statisticamente perché l'errore relativo, calcolato *ex post* con la (3), varia da strato a strato assumendo valori piú elevati del 10-15%. Nel campione complessivo, l'errore relativo resta sempre dell'ordine del 5%.

Tali risultati sono coerenti con le attese, per la strategia di campionamento adottata, specie nel Comune di Modena. Nel commento ai dati occorre tenere presente l'entità degli errori campionari per valutare correttamente le differenze osservate. Le uguaglianze osservate tra le stime esposte nelle Tabelle 14 e 15 sembrano indicare, in prima approssimazione, che si possa anche trascurare di pesare con i pesi normalizzati all'unità per stimare i parametri dei modelli; tuttavia, se i pesi ottenuti con la post-stratificazione si riportano all'unità con il solito procedimento, allora si ottengono valori del reddito medio individuale che sono statisticamente diversi da quelli ottenuti senza pesare.

5. Errori non campionari

Le indagini dirette all'accertamento del reddito, del patrimonio, del risparmio, e degli investimenti risultano sempre estremamente complicate. Si possono utilizzare diversi accorgimenti per migliorare la rilevazione (Quintano, Lucev, 1990), ma le capacità degli intervistatori sono fondamentali sia per la qualità dei dati raccolti, sia per ottenere la partecipazione delle unità statistiche (Bigarelli, Fregni, Silvestri, 2003; Couper, Groves, 1992; Hox, de Leeuw, 2002). Il processo di raccolta dei dati è suscettibile di miglioramento; purtroppo, i vincoli temporali, logistici, e di risorse umane e finanziarie pongono limiti insormontabili ai possibili miglioramenti del piano di campionamento e di raccolta dei dati; e in ciò emergono i tratti fisiologici di ogni piano concreto. L'esperienza maturata nelle indagini condotte dalla Banca d'Italia (2002, p. 33) ha mostrato che l'attendibilità dei dati è migliore per le famiglie nelle quali il capofamiglia è giovane, ha un elevato titolo di studio, è un lavoratore dipendente. Per esempio, nel Comune di Modena si sarebbe potuto stratificare ulteriormente sia sul titolo di studio (o scolarità), sia sulla posizione professionale; ma l'accesso ai dati anagrafici non è agevole, la posizione professionale non è attendibile, e il numero dei domíni di studio sarebbe diventato molto elevato: già la proposta attuale conta di $4 \times 5 \times 2$ (numero di componenti $\times$ classe di età $\times$ genere)=40 domíni. Il miglioramento del piano, in tal senso, avrebbe comportato un aggravio di costi e un allungamento dei tempi, non sostenibili; forse, non avrebbe prodotto un sostanziale guadagno nelle stime e, soprattutto, non sarebbe stato praticabile negli altri Comuni.

Nelle indagini campionarie occorre prestare attenzione all'insieme complessivo delle operazioni che si devono eseguire nella raccolta dei dati, che si articola in varie

fasi e coinvolgono molteplici persone, come gli intervistatori e gli intervistati. Le cause di errore sono, pertanto, molteplici e non sempre controllabili; i loro effetti sono denominati *errori non campionari*. Si sono indirizzati tutti gli sforzi nel ridurli perché possono diventare anche preponderanti, rispetto agli errori campionari. Le caratteristiche ideali del processo di indagine sono: (a)assenza di errori nella lista di $\wp$, ossia a ogni nominativo della lista corrisponde una e una sola unità di $\wp$ e viceversa, senza gli altri tipi di errori già menzionati; (b)la selezione delle unità è coerente con il piano di campionamento, ossia sono definite le probabilità di inclusione del primo e del secondo ordine; (c)le variabili sono rilevate senza errore per tutte le unità campionarie; (d)la codifica e la trascrizione su supporto magnetico è esente da errore (Cicchitelli, Herzel, Montanari, 1997). Non esiste ancora una teoria completa degli errori non campionari; pertanto, ogni indagine è un caso a sé e presenta un proprio *profilo dell'errore*. L'individuazione di tali errori richiedono una analisi dettagliata sul campo in cui si opera che descriva in modo completo e circoscritto tutte le operazioni necessarie e le relative (potenziali) fonti di errore e, possibilmente, anche il loro effetto sull'errore complessivo (Bailar, 1983; Bigarelli, Fregni, Silvestri, 2003).

Gli errori non campionari sono distinti generalmente in tre tipologie (Lessler, Kalsbeek, 1992): (*i*) errori nella lista o errori di copertura; (*ii*) errori da mancata risposta, derivanti sia dall'impossibilità di procedere alla rilevazione per non reperibilità o assenza di alcune unità statistiche incluse nel campione (Kish, 1965), sia dalla non partecipazione all'indagine delle unità statistiche selezionate e rintracciate —*rifiuto totale*—, sia dall'assenza di cooperazione su una particolare domanda del questionario —*rifiuto parziale*—; (*iii*) errori di misurazione, generati da numerosi fattori che alterano il valore da osservare introducendo una differenza con il valore reale.

Gli errori della lista sono i piú perniciosi perché è quasi impossibile porvi rimedio. Gli archivi anagrafici dei Comuni, utilizzati nell'indagine, costituiscono una lista ben aggiornata (attuale), con un ottimo grado di copertura di $\wp$ (completezza), senza duplicazioni di unità (ridondanza), senza grappoli di unità corrispondenti a uno stesso nominativo (molteplicità), include poche unità senza un reale corrispondente empirico o estranee a $\wp$ (inesistenza, sopracompletezza), è quasi esente da errori di imputazione: nei nomi e negli indirizzi. La scelta della lista, come la selezione delle unità dalla stessa, ha tenuto conto delle esigenze del committente, degli obiettivi dell'indagine, e delle risorse disponibili: istanze a volte in conflitto tra loro e con la dimensione del campione perché al suo aumento, cala l'errore campionario, ma tende a aumentare anche l'errore non campionario.

La riduzione degli errori da mancata risposta è il primo compito da perseguire in una indagine perché migliorare la *qualità dei dati* raccolti è lo sforzo da compiere per ottenere risultati piú affidabili e fedeli alla realtà (Liepins, Uppuluri, 1990). Proprio gli strumenti che si usano nell'indagine costituiscono una fonte primaria di errore. L'esperienza può aiutare a progettare strategie efficienti, ma nelle realtà complesse, non si riesce a sfuggire alle difficoltà tipiche di ogni rilevazione campionaria. Una vasta letteratura fornisce utili raccomandazioni sui procedimenti da seguire, ma spesso in pratica ci si trova alla corda. Ecco un elenco di princípi che si possono rintracciare in un qualunque manuale di metodologia per la ricerca sociale (Bailey, 1994): (a)cominciare l'intervista presentandosi e descrivendo concisamente obiettivi e oggetto dell'indagine; (b)sottolineare l'importanza della collaborazione degli intervistati perché consentirà, nel caso specifico, di conoscere la realtà sociale e migliorare i possibili interventi;

(c)rassicurare che i dati saranno assolutamente segreti, non ceduti a altri enti, e non usati per scopi diversi da quelli dell'indagine; (d)le risposte devono essere completamente libere e senza vincoli di sorta; (e)disporre alla fine le domande delicate, potenzialmente imbarazzanti o compromettenti —per esempio, le domande sul salario percepito e sul voto di diploma o di laurea—; (f)prestare attenzione alle batterie di domande, specie quelle con la stessa scala o con le stesse opzioni di scelta, per evitare risposte seriali —*response set*— (g)controllare il flusso informativo generato dalla serie di domande affinché sia coerente e efficace, evitando eccessivi salti logici e strutturali; (h)usare un linguaggio chiaro e semplice, evitando espressioni gergali o dialettali o tecniche; (i)predisporre un questionario il piú breve possibile, evitando l'uso di domande lunghe e relative combinazioni di esse, ma per la rilevazione in oggetto è impossibile rispettare tale indicazione; (j)aiutare il rispondente nel ricordare eventi inerenti al passato; (k)minimizzare l'introduzione di aspetti sensibili, ma reddito, patrimonio, e risparmio sono di per sé sensibili; (l)verificare all'inizio, e revisionare successivamente, gli strumenti adottati (pre-test, test).

Gli errori di misurazione si sovrappongono, in parte, a quelli da mancata risposta perché l'assenza di una risposta potrebbe dipendere proprio da una formulazione ambigua o inadeguata. La misurazione comporta, in generale, che il processo applicato goda di alcune proprietà fondamentali (Torgerson, 1962; Zeller, Carmines, 1980): la *validità*, quando rileva effettivamente l'intensità o la proprietà del concetto in esame, ossia, consegue gli obiettivi fissati; l'*attendibilità*, quando applicato piú volte agli stessi fenomeni, nelle stesse condizioni, riproduce (entro certi limiti) gli stessi risultati; la *precisione*, quando c'è la possibilità di valutare i sottomultipli dell'unità di misura. La terminologia non è ancora consolidata sicché, in alcuni contesti si usano termini piú suggestivi: sinonimi di attendibilità sono i sintagmi *stabilità* (della misura) o *fedeltà* (dello strumento); sinonimo di precisione è il termine *accuratezza* (Nunnally, Bernstein, 1994). Una distinzione tipica degli errori di misurazione è basata sulla causa che li ha prodotti: (1) errori di *strumenti*, in genere, riconducibili al questionario per domande formulate in modo ambiguo, ordinate in modo inadeguato, o batterie dei test non tarati bene, e cosí via; (2) errori di *tecniche*, in genere, legati al tipo di tecnica utilizzata, come il questionario postale, l'intervista auto-somministrata, l'intervista telefonica, il CATI —*Computer Assisted Telephone Interviewing*— oppure il CAPI —*Computer Assisted Personal Interviewing*—, la batteria di test; (3) errori dell'*intervistatore*, derivanti dalla influenza che esercita sull'intervistato sia nell'incentivare o disincentivare la sua partecipazione, sia nel fornire o non fornire una data risposta; (4) errori dell'*intervistato*, connessi alla capacità di comprensione dell'intervistato o di ricordare gli eventi accaduti, alla sua idoneità e volontà di fornire risposte veritiere. L'indagine deve rilevare, in particolare, i periodi di occupazione e disoccupazione o ricchezza famigliare e consumi o redditi o servizi; pertanto, si va incontro a diversi tipi di errori (Neter e Waksberg, 1964): l'errore *telescopico*, il rispondente ricorda l'evento, ma lo colloca in un momento errato del tempo tendendo a avvicinarlo al presente o a allontanarlo, rispetto alla data vera; l'errore di *condizionamento*, si ha in una intervista ripetuta nel tempo, quando si ha un decremento del numero di eventi riportati rispetto a quello reale; l'errore di *richiamo* (*recall loss effect*), quando si verifica una perdita di informazione dovuta o all'incapacità del rispondente di ricordare, o alla numerosità eccessiva di eventi da riportare (*report loading effect*).

La qualità dei dati raccolti si migliora soprattutto con l'accortezza nel reclutamento e addestramento dei rilevatori e si è lavorato molto in questa direzione, ma i risultati sono spesso affetti ugualmente da variabilità imponderabili (Bigarelli, Fregni, Silvestri, 2003). In questa indagine ci si è rivolto alla ditta R&I di Carpi che ha istruito e coordinato gli intervistatori, con la supervisione di Paolo Silvestri. Nonostante un impegno notevole, non si è sfuggiti alla riluttanza delle unità selezionate a collaborare all'indagine, a causa di una intervista cosí onerosa e delicata. Nel caso specifico, il tasso di rifiuto si può stimare preventivamente intorno al 60%, in base alla affidabilità della lista, alle esperienze condotte in precedenza, e anche alla letteratura esistente (Goyder, 1987; Groves, 1989; Groves *et al*., 2002). Si è notato che le difficoltà a ottenere le interviste crescono con il crescere del reddito, della ricchezza, del titolo di studio del capofamiglia (Banca d'Italia, 2002, p. 32); ma qui si sono riscontrati inconvenienti anche con un capofamiglia che ha uno stato civile libero (*single*), con gli anziani perché non aprono facilmente agli sconosciuti, con le dimensioni dei Comuni. Le relazioni sono un po' diverse da quelle riscontrate dalla Banca d'Italia, data la differente scala delle indagini: maggiori ostacoli si incontrano con Comuni piccoli e/o in montagna, con un ridotto numero di componenti, con un capofamiglia pensionato.

Forme ulteriori di errori non campionari possono emergere in altre fasi del processo di indagine: durante la codifica, la revisione, la registrazione, e l'elaborazione dei dati. Questi sono non meno rilevanti dei precedenti, ma non coinvolgono rispondenti e intervistatori (Cicchitelli, Herzel, Montanari, 1997), bensí il personale addetto di R&I.

Il trattamento degli errori non campionari richiede assunti sulle caratteristiche di $\wp$, sulla natura, e sulla distribuzione degli errori. Tali assunti non hanno sempre un corrispondente empirico e, pertanto, occorre sempre operare con la maggiore coerenza possibile rispetto alle condizioni ideali di svolgimento dell'indagine. Solo cosí si ottengono dati validi, attendibili, e precisi; ma l'ideale non corrisponde al reale, sicché occorre anche accettare l'imprecisione, fissando eventualmente un limite massimo oltre il quale ricorrere a interventi migliorativi, seppur costosi. D'altronde, anche gli istituti specializzati, come l'Istat, o con ampie risorse umane e finanziarie, come la Banca d'Italia, che sono piú accreditati presso la popolazione e supportati dalla legge —gli intervistati sono «obbligati» a partecipare all'indagine—, non riescono a ottenere il successo prescritto dalle condizioni ideali.

5.1. Misure relative alle mancate risposte

Si possono definire alcune percentuali (o indicatori) che esprimono la qualità del processo di indagine. Il Tasso percentuale di Efficienza dell'Intervistatore (TEI) esprime la percentuale di volte che un intervistatore ottiene le interviste agli indirizzi campionari contattati:

$$\text{TEI} = 100 \ \frac{\text{Numero di interviste}}{\text{Numero di contatti}},$$

ma la specificazione del numeratore e del denominatore possono evidenziare aspetti diversi del processo di intervista. Si noti che i tassi sono, in genere, rapporti «unitari» e il termine percentuale indica espressamente come è effettivamente espresso; per brevità, si ometterà di specificare «percentuale», eccetto nella definizione. Tali tipi di rapporti sono denominati anche Tassi di Completamento (TC, *completion rate*); riguardano il successo delle interviste e possono definirsi variando relativamente numeratore e deno-

minatore per ottenere indicazioni diverse sul processo di indagine. L'indice TEI, cosí definito, esprime la potenzialità di partecipazione o di successo (rispondenti eleggibili) di quella indagine; oppure il grado medio di successo degli intervistatori nell'ottenere la cooperazione degli elementi di $\wp$. Un indicatore diverso, e un po' piú preciso di TEI, può definirsi come Tasso percentuale di Interviste Completate (TIC)

$$\text{TIC} = 100 \; \frac{\text{Numero di interviste completate}}{\text{Numero di unità campionarie eleggibili}}$$

dove le unità campionarie eleggibili si riferiscono, in alcuni testi, a quelle unità che potenzialmente possono essere intervistate; ossia, l'insieme delle unità intervistate completamente, piú quelle intervistate parzialmente, piú quelle che rifiutano di partecipare, piú quelle che presentano uno stato di appartenenza non determinato, piú quelle mai rintracciate. L'aggiunta o l'eliminazione, di queste ultime, consentono di ottenere tassi di risposta diversi che consentono di cogliere aspetti diversi del processo. Le combinazioni sono diverse e tante, qui ci si limiterà soltanto a alcune variazioni per definire i principali indicatori di processo delle interviste:

$$\text{TEIC} = 100 \; \frac{\text{Numero di Interviste Completate}}{\text{Numero di Contatti}}$$

esprime il Tasso percentuale di Efficienza degli Intervistatori nei Contatti (TEIC) avuti con gli intervistati;

$$\text{PPPI} = 100 \; \frac{\text{Numero di Rispondenti a tutte le domande}}{\text{Numero di Rispondenti Cominciato Intervista}};$$

esprime la Propensione (in percentuale) degli elementi della Popolazione a Partecipare all'Indagine (PPPI), diversamente è interpretabile come l'efficienza o abilità degli intervistatori a ottenere la collaborazione degli intervistati;

$$\text{TUR} = 100 \; \frac{\text{Numero di Interviste Completate}}{\text{Numero di Unità nel Campione (Eleggibili + Ineleggibili)}}$$

dove «eleggibile», qui, deve essere inteso come il numero di unità appartenenti alla popolazione, viceversa per «ineleggibile», e esprime il Tasso percentuale di Unità statistiche Rilevate (TUR);

$$\text{TUU} = 100 \; \frac{\text{Numero di Interviste Completate Eleggibili}}{\text{Numero di Unità nel Campione (Eleggibili + Ineleggibili)}}$$

esprime il Tasso percentuale di Unità statistiche Utili ai fini della stima dei parametri di $\wp$. Nelle indagini complesse, come quella in oggetto, è difficoltoso ricorrere alle interviste per telefono; tuttavia, si possono definire analogamente alcuni indici. Per esempio, è interessante considerare il rapporto tra il numero di contatti avuti e il numero complessivo di tentativi eseguiti per accertare il peso di lavoro compiuto dagli intervistatori, definibile Tasso percentuale di Successo nei Contatti Telefonici (TSCT)

$$\text{TSCT} = 100 \; \frac{\text{Numero di Successi nei Contatti Telefonici}}{\text{Numero Totale di Tentativi di Contatti Telefonici}};$$

con piccole variazioni di numeratore e denominatore, si possono ottenere indicatori un po' diversi e interessanti a seconda degli obiettivi che si vogliono conseguire.

In termini complementari, si possono calcolare i tassi di rifiuto; per esempio, il Tasso percentuale di Non Risposta, TNR, può essere così definito

$$\mathrm{TNR} = 100 \; \frac{\text{Numero di Rifiuti}}{\text{Numero di Interviste} + \text{Numero di Rifiuti}} .$$

Si possono determinare diversi tassi variando il denominatore, ossia, distinguendo tra tutti i potenziali rispondenti o tra tutti i rispondenti con numero di telefono:

$$\mathrm{TNR}_1 = 100 \; \frac{\text{Numero di Rifiuti}}{\text{Numero di Potenziali Rispondenti}} ,$$

$$\mathrm{TNR}_2 = 100 \; \frac{\text{Numero di Rifiuti}}{\text{Numero di Potenziali Rispondenti con Numero di Telefono}} .$$

La rilevazione dei dati è stata eseguita con questionari cartacei, che sono stati trasferiti successivamente su documento elettronico dalla società R&I. Le interviste potevano essere effettuate anche con la metodologia CAPI (*Computer-Assisted Personal Interview*) perché consente sia di memorizzare direttamente le informazioni su supporto elettronico, sia di effettuare una serie di controlli sui dati immessi alla presenza dell'intervistato (la famiglia) e di correggere le eventuali incoerenze che in essi si riscontrano. Tale metodologia è usata dalla Banca d'Italia (2002, p. 30) nel 67% dei casi nel 2002, ma l'ampiezza e la portata dell'indagine in oggetto ha condotto all'esclusione della metodologia CAPI perché i costi iniziali sono assai elevati; si è preferito ricorrere, perciò, solo al tradizionale questionario cartaceo.

6. Conclusioni

Gli aspetti piú critici derivano dalle difficoltà di realizzazione che sono già note e riscontrate da ogni esperienza sul campo: non si riusciranno a rilevare tutte le unità statistiche programmate nei diversi strati; ma occorre accontentarsi perché sarebbe arduo migliorare l'entità dei dati rilevati senza aumentare i costi oltre le quote accettabili. Per quanto concerne la sostituzione delle unità che si rifiutano di collaborare, si raccomanda sempre di non ricorrere a tale pratica perché se da un lato si migliora la precisione delle stime, dall'altro si consegue un aumento della distorsione, perché le unità piú disponibili a collaborare potrebbero avere caratteristiche distintive che inficiano o distorcono le stime dei parametri della popolazione. Per esempio, è noto che le difficoltà a ottenere le interviste crescono con il crescere del reddito, della ricchezza, del titolo di studio del capofamiglia (Banca d'Italia, 2002, p. 32), con lo stato civile indipendente e giovane (*single*), con gli anziani perché non aprono facilmente a sconosciuti, con la dimensione o «natura» dei Comuni. La sostituzione delle UPS è di per sé ancora piú rilevante, ma non si è riusciti a ottenere la collaborazione delle Anagrafi e altre strade non erano praticabili. L'impegno si è concentrato, pertanto, in un'accurata attività di controllo sull'operato degli intervistatori per verificare se il lavoro svolto fosse stato puntuale, accorto, e onesto. Il risultato conseguito sembra complessivamente soddisfacente.

Una indagine che persegue obiettivi plurimi adottando una complessa strategia di campionamento non riesce a assicurare prefissati livelli di precisione di tutte le stime

prodotte. La complicazione aumenta quando, oltre alle stime di statistiche comuni, si devono stimare i parametri di alcuni modelli statistici. La soluzione di usare i pesi w^*_{ijk} o w^*_{dc}, nelle elaborazioni dei dati che coinvolgono verifiche di ipotesi, non risolve il problema perché si consegue, in genere, una sottostima dell'errore. Nel Comune di Modena si è ottenuto un campione che può essere considerato alla stessa stregua di un campionamento casuale semplice; pertanto, i pesi possono essere anche ignorati nella stima di medie e parametri di modelli statistici. Nella Provincia di Modena si sono ottenuti, invece, risultati che non consentono di ignorare il piano di campionamento, specie a causa delle mancate partecipazioni, che non si sono distribuite uniformemente tra gli strati.

Il costo dell'indagine è assai elevato; quindi, è difficile che venga ancora attuata. Nel caso si debba attuare di nuovo, tra non molto tempo, sarebbe interessante includere una parte delle famiglie già intervistate per ottenere un campione longitudinale e stimare meglio i cambiamenti e le modifiche che avvengono nella situazione economica e sociale delle famiglie (Banca d'Italia, 2002; Duncan, Kalton, 1987; Kalton, Citro, 1993).

Bibliografia

Abbate C., Baldassarini A. (1994). Contenuto informativo degli archivi INPS e confronto con altre fonti sul mercato del lavoro, *Economia & Lavoro*, **XXVIII**, n. 2, pp. 115–133.

Bailar B. A. (1983). Error profiles: uses and abuses, in Wright T., *Statistical Methodology Improvement Data Quality*, Academic Press, New York, pp. 117-130.

Bailey K. D. (1994). *Methods of Social Research*, 4th edition, The Free Press, New York. Tr. it. (1995) *Metodi della ricerca sociale*, il Mulino, Bologna.

Banca d'Italia (2000). *I bilanci delle famiglie italiane nell'anno 1998*, a cura di D'Alessio G., Faiella I., Supplementi al bollettino statistico (nuova serie), **anno X**, n. 22, Banca d'Italia, Roma.

Banca d'Italia (2002). *I bilanci delle famiglie italiane nell'anno 2000*, a cura di D'Alessio G., Faiella I., Supplementi al bollettino statistico (nuova serie), **anno XII**, n. 6, Banca d'Italia, Roma.

Barcaroli G., Di Pietro E., Venturi M. (1993). La nuova indagine trimestrale sulle forze di lavoro: aspetti metodologici e analisi dell'impatto delle innovazioni introdotte sulla stima degli aggregati, *Politiche del lavoro*, **22-23**, pp. 35-49.

Barcherini S., Calia P., Filippucci C., Grassi D. (2002). Qualità nel processo di produzione nell'indagine sui consumi dell'Istat, in Filippucci C. (a cura di) (2002), *Strategie e modelli per il controllo della qualità dei dati*, Franco Angeli, Milano.

Benassi P., Zoda G. (2002). *La popolazione modenese 2002. Analisi sulla struttura, per sesso e per classi di età, della popolazione residente nei comuni e nelle aree della programmazione sovracomunale al 31 dicembre 2001*, SISTAN, Provincia di Modena.

Bethlehem J. G., Keller W. J. (1987). Linear weighting of sample survey data, *Journal of Official Statistics*, **3**, pp. 141–153.

Bigarelli D., Fregni C., Silvestri P. (2003). Rilevazione dei dati e attendibilità delle risposte nell'indagine sulle condizione economiche delle famiglie nella Provincia di Modena, mimeo.

Brandolini A. (1999). The distribution of personal income in post-war Italy: source description, data quality, and the time pattern of income inequality, *Giornale degli Economisti e Annali di Economia*, **58**, n. 2, pp. 183-239.

Brandolini A., Cannari L. (1994). Methodological Appendix: the Bank of Italy's Survey of Households Income and Wealth, in Ando A., Guiso L., Visco I. (eds.), *Saving and the Accumulation of Wealth*, Cambridge University Press, Cambridge, pp. 369-386.

Cannari L., Gavosto A. (1994). L'indagine della Banca d'Italia sui bilanci delle famiglie: una descrizione dei dati sul mercato del lavoro, *Economia & Lavoro*, **XXVIII**, n. 1, pp. 63-79.

Cannari L., Pellegrini G., Sestito P. (1996). *L'utilizzo di microdati d'impresa per l'analisi economica: alcune indicazioni metodologiche alla luce delle esperienze in Banca d'Italia*, Temi di discussione, Numero 286, Banca d'Italia, Roma, pp. 1-49.

Chisini O. (1929). Sul concetto di media, *Periodico di matematiche*, **9** (4).

Cicchitelli G., Herzel A., Montanari G. E. (1997). *Il campionamento statistico*, II edizione, il Mulino, Bologna.

Cochran W. G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Couper M. P., Groves R. M. (1992). The Role of Interviewer in Survey Participation, *Survey Methodology*, **18**, pp. 263-278.

De Vitiis C., Falorsi S. (2000). *Analisi dell'impatto della nuova strategia di campionamento dell'indagine Istat sui consumi delle famiglie*, Documenti ISTAT, n. 5, ISTAT, Roma.

Deville J. C., Särndal K. E. (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, pp. 376–282.

Di Pietro E. (1993). La nuova indagine Istat sulle forze di lavoro, *Economia & Lavoro*, **XXVII**, n. 1, pp. 57-64.

Duncan G. J., Kalton G. (1987). Issue of design and analysis of surveys across time, *International Statistic Review*, **55**, 97-117.

Fabbris L. (1989). *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*, La Nuova Italia Scientifica, Roma, 1989.

Falorsi P. D., Falorsi S., Russo A. (1992). *Indagine campionaria sui consumi delle famiglie: strategia di campionamento e precisione delle stime*, Rapporto di ricerca N. 3, CONPRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, Bologna.

Falorsi P. D., Russo A. (1992). *La mancata risposta totale nei campioni complessi: un'applicazione all'indagine campionaria sui consumi delle famiglie*, Rapporto di ricerca N. 23, CONPRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, Bologna.

Falorsi P. D., Falorsi S. (1995). *Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese*, Rapporto di ricerca N. 13, CONPRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, Bologna.

Filippucci C., Marliani G. (1992). *La misura dei consumi delle famiglie: una riflessione a partire dall'esperienza italiana*, Rapporto di ricerca N. 6, CONPRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, Bologna.

Goyder J. (1987). *The Silent Minority*, Basil Blackwell, Oxford.

Groves R. M. (1989). *Survey Errors and Survey Costs*, Wiley & Sons, New York.

Groves R. M., Dillman D. A., Eltinge J. L., Little R. J. A. (2002). *Survey Nonresponse*, Wiley & Sons, New York.

Hansen M. H., Hurwitz W. N. (1943), On the theory of sampling from finite populations, *The Annals of Mathematical and Statistics*, **14**, pp. 333-362.

Horvitz D. G., Thompson D. J. (1952). A Generalization of Sampling Without Replacement from a finite Universe, *Journal of the American Statistical Association*, **47**, pp. 663–685.

Hox J., de Leeuw E. (2002). The Influence of Interviewers' Attitude and Behavior on Household Survey Nonresponse: An International Comparison, in Groves R. M., Dillman D. A., Eltinge J. L., Little R.J.A. (2002). *Survey Nonresponse*, Wiley & Sons, New York, pp. 103-120.

ISTAT (2002a). *I consumi delle famiglie. Anno 2000*, a cura di Pannuzi N., Annuario, n. 7, Istat, Roma.

ISTAT (2002b). *Stili di vita e condizioni di salute. Indagine multiscopo sulle famiglie: «Aspetti della vita quotidiana». Anno 2000*, a cura di Orsini S., Informazioni, n. 3, Roma, Appendice C, pp. 63-75.

ISTAT, (2000c). *Le condizioni di salute della popolazione. Indagine multiscopo sulle famiglie «Condizioni di salute e ricorso ai servizi sanitari». Anni 1999-2000*, a cura di Gargiulo L., Sebastiani G., Informazioni, n. 12, Roma, Nota metodologica, pp. 109-127.

ISTAT (2002d). *Panel europeo sulle famiglie*, a cura di Pauselli C., Metodi e Norme, nuova serie, n. 15, Roma.

Kalton G., Citro C. F. (1993). Panel Surveys: Adding the Fourth Dimension, *Survey Methodology*, **19**, pp. 205-215.

Kish L. (1965). *Survey Sampling*, John Wiley & Sons, New York.

Kish L. (1990). Weighting: why, when, and how, *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 121–130.

Kish L. (1992). Weighting for unequal P_i, *Journal of Official Statistics*, **8**, 2, pp. 121–130.

Lessler J. T., Kalsbeek W. D. (1992). *Nonsampling Errors in Surveys*, Wiley & Sons, New York.

Liepins G. E., Uppuluri V. R. R. (1990). *Data Quality Control. Theory and Pragmatics*, Marcel Dekker, New York.

Little R. J. A., Rubin D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Little R. J. A. (1993). Post-Stratification: A Modeler's Perspective, *Journal of the American Statistical Association*, **88**, pp. 1001-1012.

Lucev D. (1992). *Le mancate risposte totali nell'indagine sui consumi delle famiglie*, Rapporto di ricerca N. 14, CONPRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, Bologna.

Lucifora C. (1995). L'analisi del mercato del lavoro con micro-dati: l'utilizzo degli archivi amministrativi INPS, *Economia & Lavoro*, **XXIX**, n. 3, pp. 3–20.

Martini M. (1990). I dati amministrativi come fonte di informazione statistica sulle imprese, *Economia & Lavoro*, **XXIV**, n. 2, pp. 45–58.

Martini M., Aimetti P. (1989). *Un archivio delle imprese per l'analisi economica. Fonti, metodi e risultati*, Union-camere e Regione Lombardia, Milano.

Neter J., Waksberg J. (1964). A Study of Response Errors in Expenditures Data form Household Survey, *Journal of the American Statistical Association*, **59**, pp. 18-55.

Nunnally J. C., Bernstein I. H. (1994). *Psychometric Theory*, McGraw-Hill, New York.

Piccolo D. (1998). *Statistica*, il Mulino, Bologna.

Potter F. J. (1990). A study of procedures to identify and trim extreme sampling weights, *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 121–130.

Quintano C., Lucev D. (1990). Le mancate risposte in esperienze di indagini reddituali, *Quaderni sardi di economia*, **20**, n. 3, pp. 253-278.

Rubin D. B. (1988). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Särndal C. E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, Berlin.

Torgerson W. S. (1962). *Theory and Methods of Scaling*, Wiley & Sons, New York.

Verma V. (1995). *Weighting for Wave 1*, Working Group "European Community Household Panel", Doc. PAN 36/95, Statistical Office of the European Communities, Luxembourg.

Woodruff R. S. (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, **66**, pp. 411–414.

Zeller R. A., Carmines E. G. (1980). Measurement in the Social Sciences: the Link between Theory and Data, Cambridge University Press, Cambridge.

Zhang L.-C. (2000). Post-Stratification and Calibration — A Synthesis, *The American Statistician*, **54**, n. 3, pp. 178-184.

Zoda G. (1998). *La popolazione modenese 1997. Analisi sulla struttura, per sesso e per classi di età, della popolazione residente nei comuni e nelle aree della programmazione sovracomunale al 31 dicembre 1997*, SISTAN, Provincia di Modena.

Elenco dei simboli piú frequenti

D Numero di distretti sociosanitari o strati, D=7.

C_d Numero di Comuni nel distretto d.

c_d Numero di Comuni inclusi nel campione del distretto d.

C_d^* Numero di Comuni NAR nel distretto d, risulta $C_d^* = C_d - 1$.

c_d^* Numero di Comuni NAR inclusi nel campione del distretto d, vale $c_d^* = c_d - 1$.

c_{d1} Comune AR incluso nel campione del distretto d.

N_d Numero di USS (famiglie) nel distretto d.

N_d^* Numero di USS (famiglie) nel distretto d, strato di Comuni NAR.

n_d^* Numero di USS (famiglie) nel distretto d, strato di Comuni NAR, nel campione.

N_{d1} Numero di USS (famiglie) nel distretto d, Comune AR.

n_{d1} Numero di USS (famiglie) nel distretto d, Comune AR, nel campione.

N_{dc} Numero di USS (famiglie) nel distretto d, Comune c (per c>1).

n_{dc} Numero di USS (famiglie) nel distretto d, Comune c (per c>1), nel campione.

w_{dc} peso delle USS (famiglie) nel distretto d, e nel Comune c; per c=1 il peso si riferisce al Comune AR, per c>1 al Comune NAR della Provincia senza Modena.

$w_{P;dc}$ peso degli individui residenti nel distretto d, e nel Comune c.

w_{dc}^* peso normalizzato a uno delle USS (famiglie) nel distretto d, e nel Comune c.

$w_{P;dc}^*$ peso normalizzato a uno degli individui residenti nel distretto d, e nel Comune c.

w_{ijk} peso delle USS (famiglie) nel Comune di Modena.

$w_{P;ijk}$ peso degli individui residenti nel Comune di Modena.

w_{ijk}^* peso normalizzato a uno delle USS (famiglie) nel Comune di Modena.

$w_{P;ijk}^*$ peso normalizzato a uno degli individui residenti nel Comune di Modena.

$\lfloor \bullet \rfloor$ parte intera dell'argomento; ossia, arrotondamento per difetto.

Y se sta per carattere indica il reddito, se sta per parametro indica il totale in $\wp$.

$\hat{Y}_d$ stimatore del totale (il reddito) a livello di distretto («distrettuale») in $\wp$.

$s_{2;dc}^2$ varianza campionaria a livello comunale o di secondo stadio.

f $f = n/N$, frazione di campionamento totale o provinciale.

$f_{1;d}$ $f_{1;d} = n_d / N_d$, frazione di campionamento «distrettuale».

$f_{2;dc}$ $f_{2;dc} = n_{dc} / N_{dc}$, frazione di campionamento comunale o di secondo stadio.

$1_{[\bullet]}[\bullet]$ funzione indicatrice che vale 1, se l'argomento appartiene all'insieme specificato nell'indice, vale 0 altrimenti; per esempio, $1_{[a,b)}[x]$ è uguale a 1 se $x \in [a,b)$, è uguale a 0 se $x \notin [a,b)$. Si noti che la parentesi quadra indica che il valore estremo adiacente è incluso nell'insieme, mentre la parentesi tonda indica che il valore estremo adiacente non è incluso nell'insieme.

$\widetilde{Y}_{djk}$ stimatore di Horvitz-Thompson del totale di Y nel post-strato jk del distretto d.

$\widetilde{N}_{djk}$ stimatore del totale dei soggetti nel post-strato jk del distretto d.

RINGRAZIAMENTI

Per svolgere una ricerca complessa è necessaria la collaborazione di numerose persone che lavorano gratuitamente e con dedizione. Non è né in questa nota che si possa esprimere la nostra gratitudine a tutti, indicando esplicitamente anche i loro nomi, né compito dell'autore perché spetta al coordinatore della ricerca manifestarla nella sede opportuna. Si coglie l'occasione, tuttavia, di ringraziare già nel presente lavoro i dirigenti dell'Anagrafe dei comuni selezionati nel campione che hanno cooperato e coloro che hanno estratto le unità campionarie in ogni Comune campione; in particolare, si ringraziano il dott. Giuliano Orlandi, dirigente del Servizio Statistica e Mercati del Comune di Modena, che si è mostrato sempre cortese e disponibile e il dott. Giovanni Bigi che ci ha fornito, ogni volta, tutti i dati richiesti con competenza e sollecitudine. Si esprimono, infine, ringraziamenti a Cinzia Mortarino che, con una lettura attenta e meticolosa di una precedente versione del testo e con le sue preziose osservazioni, ha contribuito a renderlo piú chiaro e preciso. Vale il solito *caveat*: responsabile di errori e eventuali omissioni è, naturalmente, solo l'autore.

Lavoro svolto nell'àmbito del progetto di ricerca

«*Costruzione di un'indagine sulle famiglie e di un modello di microsimulazione per l'analisi delle politiche sociali e fiscali a livello locale*»

cofinziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR).

Assegnazione: Anno 2001 – prot. 2001135524.

Coordinatore: Paolo Bosi