Università degli Studi di Modena e Reggio Emilia
Dipartimento di Economia Politica
CAPP: Centro di Analisi delle Politiche Pubbliche

\\ 512 \\

Il disegno della seconda indagine sulle condizioni economiche e sociali delle famiglie nella Provincia di Modena

di

Michele Lalla

e-mail: lalla.michele@unimore.it

Materiali di discussione

Università degli Studi di Modena e Reggio Emilia
Dipartimento di Economia Politica
Via Jacopo Berengario 51
41100 Modena (Italia)

RINGRAZIAMENTI

Si ringraziano il dott. Giuliano Orlandi, dirigente del Servizio Statistica 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.

Lavoro svolto nell'àmbito del progetto di ricerca

«*Valutazione delle politiche fiscali e sociali locali con modelli di microsimulazione statici e dinamici*»

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

Assegnazione: Anno 2003 – prot. 2003139941.

Coordinatore: Paolo Bosi

1. Introduzione

L'indagine sulle condizioni economiche e sociali delle famiglie nella provincia di Modena è stata già condotta nell'anno 2002. Gli obiettivi, il piano di campionamento, i problemi, e i risultati dell'indagine sono descritti in Baldini, Bosi, e Silvestri (2004). Una indagine analoga deve essere realizzata nel 2006 perché gli esiti conseguiti sono stati di interesse per la comunità; infatti, rilevazioni simili sono svolte anche in altre località perché consentono di fotografare con una maggiore precisione lo stato della società rispetto al reddito, al patrimonio, agli investimenti, al lavoro, e alle condizioni di vita. Il piano di campionamento della seconda indagine sarà inevitabilmente simile a quello precedente, sebbene arricchito dall'esperienza sul campo; pertanto, nel séguito, si ripeterà, pur succintamente, molto di quanto è stato già stato esposto in Lalla (2003).

Gli obiettivi dell'indagine sono molteplici e conducono alla rilevazione di dati economici, sociali, e demografici delle famiglie. Il campione da costituire deve essere, quindi, in grado di rappresentare la distribuzione di alcuni caratteri fondamentali (il reddito, il risparmio, gli investimenti) e di fornire il supporto informativo per la costruzione di un modello di microsimulazione. Oggetto di rilevazione sono anche alcuni aspetti della domanda dei servizi pubblici (offerti all'infanzia, agli anziani, e ai disabili), delle condizioni di salute, e dell'uso del tempo libero. Si deve ricostruire per l'individuo (e le famiglie) i carichi fiscali e i beneficî derivanti da un insieme predefinito di istituti fiscali e di programmi di spesa nazionale e locale. Tra quelli nazionali si ricordano l'IRPEF, gli assegni pensionistici di varia natura, gli assegni al nucleo famigliare. Tra gli istituti locali si ricordano l'ICI, le tariffe per le forniture dei principali beni di utilità pubblica (luce, acqua, gas), le forme di minimo vitale, le strutture protette per anziani, l'assistenza domiciliare, le tasse per la raccolta dei rifiuti urbani, gli asili nido, e cosí via. I risultati attesi possono essere molto elevati, considerando le domande presenti nel questionario, sicché occorre súbito precisare che per problemi specifici, inerenti a segmenti ridotti di popolazione, si devono utilizzare indagini mirate e non generali, perché le proporzioni di popolazione interessate sono piccole e le variabili che determinano il fenomeno sono molteplici. Il numero di casi concernenti fenomeni particolari è spesso troppo piccolo nel campione: non si riesce a analizzare, perciò, le determinanti delle risposte in profondità e in modo esauriente (v. *infra*).

Le indagini analoghe condotte a livello nazionale sono molteplici; tra le altre, si ricordano l'indagine (quotidiana, riportata all'anno) sui consumi delle famiglie (Istat, 2002, 2004) e l'indagine biennale sui bilanci delle famiglie condotta dalla Banca d'Italia (2002, 2004). L'indagine sui consumi delle famiglie condotta dall'Istat è sempre oggetto di analisi e riflessioni (De Vitiis, Falorsi, 2000; Barcherini *et al*., 2002) e rileva anche il reddito, ma non in forme dettagliate e accurate sicché il legame esistente tra reddito e consumo non può essere analizzato compiutamente. L'indagine della Banca d'Italia, invece, rileva in dettaglio il reddito, il risparmio, e gli investimenti, ma le informazioni sul consumo sono pressoché irrilevanti (Brandolini, 1999, 2005). La rilevazione dei due dati, reddito e consumo, rimane tecnicamente problematica, anche se avere entrambe le informazioni è utile per modellare i comportamenti di un agente economico. L'accertamento del consumo, seppure riferito a 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. Si può concludere a priori che il rapporto costo/prestazione non è accettabile e le distorsioni per le mancate

collaborazioni e risposte sarebbero elevate. 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: il reddito e il patrimonio saranno accurati, ma la spesa per i consumi sarà trascurata, in generale. Si sono condotte, inoltre, anche diverse indagini a livello locale, piú o meno simili tra loro, che hanno operato seguendo una metodologia di rilevazione comparabile a quella utilizzata per l'indagine in oggetto (Baldi, Lemmi, Sciclone, 2005; Benassi, 2005; Betti *et al*., 2003; Plaseller, Vogliotti, Zeppa, 2005; Palamenghi, Riva, Trentini, 2005).

La struttura del lavoro è la seguente. Nel paragrafo 2 si illustrano gli aspetti del piano di campionamento concernenti la determinazione della dimensione campionaria, la stratificazione, e il criterio di selezione delle unità statistiche campionarie. Nel paragrafo 3 si espongono i procedimenti adottati per determinare i fattori di riporto alla popolazione obiettivo e le varianze degli stimatori di interesse. Nel paragrafo 4 si delineano alcune caratteristiche delle indagini per analizzare i fenomeni che evolvono nel tempo, in generale e in particolare per l'indagine corrente: strategie di campionamento, vantaggi e svantaggi, stimatori, e pesi. Nel paragrafo 5 si riassumono le tipologie di errori non campionarî. Nel paragrafo 6 seguono, infine, le conclusioni.

2. Piano di campionamento

La costruzione di un campione per conseguire gli obiettivi di una indagine richiede di possedere una lista (*frame*) della popolazione di riferimento o obiettivo (*target*), che sia priva di carenze informative sulle unità statistiche: incompletezza, sopracompletezza, ridondanza, inesistenza, inattualità, imprecisioni. Il piano di campionamento si potrebbe progettare con piú efficacia, se fosse possibile avere informazioni sulle 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, 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. Per motivi di riservatezza è, tuttavia, impossibile accedervi (Lalla, 2003); si procederà, quindi, senza informazioni specifiche sulle unità statistiche ricorrendo agli archivi anagrafici dei comuni, ai quali ci si riferirà brevemente con il temine «*lista anagrafica*».

Il piano di campionamento descritto valuta il numero di unità statistiche (dimensione) da selezionare dalla popolazione di riferimento, idoneo a soddisfare gli obiettivi dell'indagine (§2.1), e 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 anche solo con il termine «unità statistiche». Per il comune di Modena si è previsto un campione con una dimensione piú elevata, rispetto agli altri e una stratificazione per ampiezza della famiglia, classe di età, e genere del capofamiglia. L'estrazione delle famiglie si effettuerà con un campionamento sistematico dalla lista anagrafica (§2.3).

2.1. Dimensione campionaria

L'indagine sulle condizioni economiche e sociali delle famiglie, condotta nel 2002, fornisce alcune indicazioni sul reddito famigliare, $\mathcal{Y}$, che si possono utilizzare per valutare la dimensione del campione tramite la seguente relazione (Cochran, 1977):

$$n_e \frac{\frac{z_{1-\alpha/2}^2 S^2}{r^2 \bar{Y}^2}}{1+\frac{1}{N}\left(\frac{z_{1-\alpha/2}^2 S^2}{r^2 \bar{Y}^2}-1\right)}, \tag{1}$$

dove S^2 indica la varianza (non corretta) della $\mathcal{Y}$, $\bar{Y}$ la media, N la dimensione della popolazione obiettivo, r l'errore relativo (percentuale) che si commette nella stima dei parametri (media o totale) della $\mathcal{Y}$, $z_{1-\alpha/2}$ 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. Le grandezze indicate con le lettere maiuscole si riferiscono alla popolazione di riferimento, mentre le grandezze indicate con le lettere minuscole si riferiscono al campione selezionato e osservato. Il valore del livello di significatività α si può fissare pari al 5%, sicché il valore di $z_{1-\alpha/2}$ è uguale a 1,96 e si approssima a 2, per semplicità. Infine, si noti che il denominatore esprime l'effetto della correzione per popolazioni finite; pertanto, occorre conoscere N.

Se non si conosce alcuna variabile rilevante da stimare, si può fissare l'errore sulla stima di una proporzione, P, della modalità di una data variabile qualitativa. La dimensione del campione si ottiene, allora, dalla seguente relazione (Cochran, 1977):

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

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

La dimensione del campione ottenuto dall'indagine, m, può risultare inferiore a n_e per mancate risposte o partecipazioni. 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 (1) per la variabile continua $\mathcal{Y}$,

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

e l'errore (assoluto) dalla (2) per la variabile dicotoma,

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

La dimensione del campione dipende dalla precisione desiderata delle stime delle diverse variabili. Per ognuna di esse, si 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 che si deve 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

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

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

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

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 la (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 1600 unità statistiche (famiglie). Una 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 assoluto desiderato, che si voleva uguale per il comune di Modena e per la provincia. Per la provincia di Modena, esclusa Modena, al 31 dicembre del 2004, si ha N=193.276 e con e=0,0353 si ottiene n=800; per il comune di Modena si ha N=78.962 e con e=0,03518 si ottiene n_{MO}=800. Si può notare che la dimensione del campione è quasi insensibile alla variazione della estensione dell'area di studio e alla consistenza della popolazione; ossia, indagare la popolazione di una città, di una provincia, di una regione, o dell'intera nazione non altera la quantità di unità statistiche necessarie per avere un campione adeguato a fornire un determinato errore campionario sulle caratteristiche oggetto di stima (Barisione, Mannheimer, 1999) e per il quale si usa spesso, ma impropriamente, il termine *rappresentativo* (Calandi, 2003).

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 valutare correttamente la 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. Si adotta, quindi, l'espressione (1) per valutare la dimensione adeguata a 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 per il calcolo di n_e da una proporzione P, data dalla (2) per un fissato valore assoluto, e, diventa la seguente per un fissato errore relativo, r:

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

Diversamente, si fissa l'errore relativo sulla proporzione P; allora, l'errore assoluto è dato da $e = rP$ e, sostituendolo nella (2), si ottiene la (7). Si può mantenere, quindi, costante l'errore relativo rispetto a P. Nel caso P=0,5 e un errore assoluto e=0,05 si ha un errore relativo del 10%; infatti, $r=e/P$. L'errore relativo è uguale, allora, per esempio, a 0,04 per P=0,4 e a 0,03 per P=0,3. Analogamente varierà la dimensione del campione: n=600, n=933, 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 deve ricorrere a una indagine focalizzata o a gruppi opportunamente selezionati.

L'indagine condotta nel 2002 ci fornisce, però, alcune indicazioni per valutare sia la precisione conseguita sul reddito (errore relativo), sia la dimensione del campione utilizzando l'equazione (1) nella quale occorrono, tra le altre, la media e la deviazione standard del reddito. Nella Tabella 1 si può osservare che la precisione relativa del reddito è piú del 10% per i singoli distretti sociosanitarî, eccetto il comune di Modena per il quale è circa il 5%; per tutta la provincia l'errore relativo è del 3,46% equivalente a circa 1,263 € (dato in migliaia) sull'anno, che è assai simile a quello di indagini locali confrontabili: a Bolzano, per esempio, è 1,4 € pari a un errore relativo del 2,4% con media 29,9 € e deviazione standard 31,0 € (Plaseller, Vogliotti, Zeppa, 2005).

Tabella 1 – Numero di famiglie nell'indagine 2002 (n), reddito medio e deviazione standard (DS) in migliaia di euro, popolazione al 31/12/2000 (N), errore relativo osservato, dimensione del campione in base ai dati del reddito, e ripartizione proporzionale per distretto sociosanitario

Distretto Sociosanitario	**n-2002**	**Media $\mathscr{Y}$**	**DS di $\mathscr{Y}$**	**N-2000**	**Err. Rel. 2002**	**(*)n-2006**	**n-2006 proporz.**
Carpi	148	37,649	23,354	36534	0,1020	154	163
Mirandola	119	33,558	15,698	30160	0,0858	88	140
Modena	589	38,400	24,855	74675	0,0531	801	800
Sassuolo	167	35,619	18,987	42584	0,0824	114	169
Pavullo	39	33,529	20,413	15968	0,1960	147	129
Vignola	109	36,263	22,857	30613	0,1207	159	116
Castelfranco Emilia	64	36,003	23,840	22434	0,1654	175	109
Totale	**1235**	**36,457**	**22,194**	**252968**	**0,0346**	**1638**	**1626**

(*) n-2006 deriva da un errore relativo sul reddito del 10% per la provincia e del 4,55% per il comune di Modena

Si nota, infine, che nei distretti sociosanitarî la deviazione standard è abbastanza stabile, ma a Modena è un po' piú alta del valore provinciale e a Pavullo è un po' piú bassa, a Sassuolo è molto piú bassa, e a Mirandola è ancora piú bassa. Il valore piú alto a Modena può derivare da una certa polarizzazione dei redditi che induce una dilatazione della loro distribuzione, come pure il valore piú basso di Pavullo può

derivare da una minore ampiezza dei redditi che si ha nelle aree montane; meno evidenti sono le origini degli scarti osservati a Sassuolo e a Mirandola, ma potrebbero derivare proprio dall'autoselezione dei rispondenti (*selectivity bias*). Tale instabilità genera la differenza tra le dimensioni campionarie ottenute nei diversi distretti con l'equazione (1) e un errore relativo del 10% (penultima colonna della Tabella 1); mentre le dimensioni campionarie ottenute con la ripartizione proporzionale alla numerosità della popolazione di riferimento (ultima colonna della Tabella 1) variano solo in base all'entità della popolazione stessa nei distretti. Sia per semplicità e sia per le difficoltà che si incontrano a stimare la deviazione standard nelle indagini complesse (Cochran, 1977, pp. 78-81), come quella condotta in precedenza, che ha anche una bassa numerosità per strato, si è adottato ancora il criterio della ripartizione proporzionale.

2.2. Stratificazione

La procedura di stratificazione realizza il raggruppamento delle unità statistiche, per 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 sulla popolazione di riferimento disponibili nella fase iniziale che, allo stato attuale, sono assai ridotte per l'indagine in oggetto. La base di dati utile per costruire un buon campione è, infatti, l'archivio del Ministero delle finanze (banca dati fiscali). L'accesso è, difatti, impossibile perché riservata e, pertanto, richiederebbe una collaborazione molto attiva del personale che è, in genere, carente. L'uso di caratteri individuali per realizzare la stratificazione, infatti, comporta: (a) l'elaborazione per conoscere la struttura della popolazione di riferimento e determinare la consistenza del campione per strato; (b) l'estrazione successiva delle famiglie da includere nel campione. Le difficoltà operative e di accesso sono quasi proibitive; perciò, in alternativa alla base di dati fiscali, si può ricorrere agli archivi anagrafici della popolazione residente, ugualmente protetti dalla legge sulla riservatezza dei dati (dalla Legge n. 675 del 31 dicembre 1996 al Decreto Legislativo n. 196 del 30 giugno 2003), ma piú facilmente trattabili con l'aiuto degli addetti che già attuano, spesso per conto dell'Istat, l'estrazione di unità statistiche da includere nelle sue varie indagini. I dati sulla struttura della popolazione rispetto a determinati caratteri per la fase (a) e i dati individuali delle famiglie per la fase (b) si possono richiedere, infatti, alle persone autorizzate all'acceso alle basi di dati e già con esperienze di estrazioni di nominativi per indagini campionarie.

Si è eseguita soltanto una stratificazione del territorio, date le difficoltà, ma si è deciso di modificare, rispetto alla indagine precedente, la suddivisione dell'area provinciale al fine di migliorare l'efficacia della rilevazione e aumentare l'omogeneità territoriale delle UPS (§2.2.1). Le famiglie sono state stratificate solo per il comune di Modena, secondo la loro ampiezza, l'età del capofamiglia, e il genere del capofamiglia perché il «Servizio Statistica» è stato disponibile a cooperare (§2.2.2). Il processo di allocazione adottato è stato proporzionale alla numerosità della popolazione negli strati rispetto alle suddivisioni operate, sia territoriale e sia per caratteri della famiglia nel comune di Modena, anche se per quest'ultimo si poteva pensare all'uso di dati già raccolti, simili a quelli della Tabella 1, e illustrati nella Tabella 4.

2.2.1. Stratificazione territoriale

Una tipica suddivisione della provincia di Modena è costituita dalle aree geografiche (macrostrati) individuate dai distretti sociosanitarî (Benassi, Zoda, 2002), la denominazione dei quali è data dalla città piú rappresentativa. Essi sono stati confrontati con i sistemi locali del lavoro e con l'organizzazione amministrativa delle Comunità montane al fine di incrementare l'omogeneità delle aree. L'aggregazione dei comuni è diventata, quindi, la seguente. Il distretto N.1, di Carpi (D1), contiene anche i comuni di Campogalliano, Novi di Modena, e Soliera: coincide con quello sociosanitario. 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: coincide con quello sociosanitario. 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, e Maranello. Il distretto N.5, di Pavullo nel Frignano (D5), contiene tutti i comuni della montagna: Fanano, Fiumalbo, Lama Mocogno, Montecreto, Pievepelago, Polinago, Riolunato, Serramazzoni, e Sestola, che sono quelli del distretto sociosanitario; piú quelli montani dei distretti sociosanitarî di Sassuolo (Frassinoro, Montefiorino, Palagano, e Prignano sulla Secchia) e di Vignola (Guiglia, Marano sul Panaro, Montese, e Zocca). Il distretto N.6, di Vignola (D6), contiene anche i comuni di Castelnuovo Rangone, Castelvetro, Savignano sul Panaro, Spilamberto. Il distretto N.7, di Castelfranco Emilia (D7), contiene anche i comuni di Bastiglia, Bomporto, Nonantola, Ravarino, e San Cesario sul Panaro: coincide con quello sociosanitario. Aumenta cosí l'omogeneità delle aree di Sassuolo e Vignola e si definisce un'area montana unica che, seppure eterogenea, possiede una peculiare fisionomia amministrativa e geografica, ossia una intrinseca omogeneità territoriale. Ci si riferirà a tali aree solo con il termine *distretti* e solo in casi di ambiguità si aggiungerà la specificazione *d'area*.

In ciascun distretto d'area, le UPS sono state raggruppate in due categorie o strati: AutoRappresentative (AR), corrispondenti ai comuni che *denominano* i distretti o superano la soglia fissata di 10000 USS; e Non AutoRappresentative (NAR), tutti gli altri. Solo quattro comuni AR, su sette che denominano i distretti, superano la soglia di 10000 USS e il comune mediano è Castelfranco Emilia. Tra i comuni che non denominano i distretti, solo Formigine supera la soglia e diventa AR. Si noti che la soglia di 10000 USS è simile a quella utilizzata dalla Banca d'Italia nella stratificazione dei comuni (Brandolini, Cannari, 1994; Cannari, Gavosto, 1994). Anche l'Istat opera una analoga stratificazione del territorio nell'indagine sui consumi delle famiglie (Falorsi, Falorsi, Russo, 1992; De Vitiis, Falorsi, 2000) e sulle forze di lavoro (Di Pietro, 1993; Barcaroli, Di Pietro, Venturi, 1993). Nella Tabella 2 si mostra una ripartizione della dimensione campionaria provinciale, n=1600, proporzionale alla numerosità (frequenze) di USS per ogni UPS, n_{dc}, come se fossero tutte AR. L'indice d di n_{dc} denota il distretto e l'indice c denota il comune. I valori di n_{dc} sono stati arrotondati tutti per eccesso e ciò ha generato un lieve aumento della dimensione totale, che è passata a $n = 1626$. Non si è eseguita la ripartizione secondo la numerosità della popolazione residente perché le USS sono correlate a essa e, dunque, le variazioni non sono rilevanti per l'omogeneità della struttura demografica delle famiglie nel territorio.

Tabella 2 – Numero di famiglie (N_{dc}) e dimensione campionaria proporzionale (n_{dc}) per tutti i comuni della provincia di Modena suddivisi per distretto d'area al 31/12/2004

D	COMUNE	N_{dc}	n_{dc}	D	COMUNE	USS	n_{dc}
D1	**CARPI**	26019	108		Riolunato	337	2
	Campogalliano	3167	14		Serramazzoni	3230	14
	Novi di Modena	4165	18		Sestola	1264	6
	Soliera	5454	23		Frassinoro (S)	1017	5
	Totale D1	**38805**	**163**		Montefiorino (S)	1050	5
D2	**MIRANDOLA**	9211	36		Palagano (S)	1118	5
	Camposanto	1184	5		Prignano sulla Secchia (S)	1423	6
	Cavezzo	2775	11		Guiglia (V)	1737	8
	Concordia sulla Secchia	3366	13		Marano sul Panaro (V)	1540	7
	Finale Emilia	6298	26		Montese (V)	1532	7
	Medolla	2382	9		Zocca (V)	2236	10
	San Felice sul Panaro	4172	15		**Totale D5**	**28758**	**129**
	San Possidonio	1448	6	**D6**	**VIGNOLA**	9251	39
	San Prospero	1994	7		Castelnuovo Rangone	4980	21
	Totale D2	**32830**	**128**		Castelvetro	3950	17
D3	**MODENA**	**78962**	**800**		Savignano sul Panaro	3403	15
D4	**SASSUOLO**	16463	69		Spilamberto	5577	24
	Fiorano Modenese	6011	25		**Totale D6**	**27161**	**116**
	Formigine	11666	49	**D7**	**CASTELFRANCO E.**	11162	47
	Maranello	6042	26		Bastiglia	1424	6
	Totale D4	**40182**	**169**		Bomporto	3093	13
D5	**PAVULLO nel Frignano**	6573	28		Nonantola	5364	23
	Fanano	1437	6		Ravarino	2264	10
	Fiumalbo	585	3		San Cesario sul Panaro	2233	10
	Lama Mocogno	1405	6		Totale D7	**25540**	**109**
	Montecreto	436	2		**Totale comuni AR**	157641	1130
	Pievepelago	966	5		**Totale comuni NAR**	114597	496
	Polinago	872	4		**Totale Provincia**	**272238**	**1626**

La determinazione del numero dei comuni NAR per ogni distretto è stata eseguita in base al numero di USS per distretto, N_d, considerando la mediana della dimensione dei comuni AR, approssimata a 10000 per comodità. Allora, si è assegnato a ciascun distretto un Comune NAR ogni 10000 USS. In termini formali

$$c_d^{NAR} = \left\lfloor \left(\frac{1}{10000} \sum_{c=1}^{C_d^{NAR}} N_{dc} \right) + 1 \right\rfloor . \qquad (8)$$

dove c_d^{NAR} è il numero di NAR da selezionare nel *d*-esimo distretto, C_d^{NAR} è il numero totale di NAR nel *d*-esimo distretto per il quale si ha, in genere, $C_d^{NAR} = C_d - 1$ (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 $\lfloor . \rfloor$ indica la parte intera dell'argomento. Le UPS da includere nei distretti sono state determinate con una generazione di numeri casuali proporzionali alla loro dimensione 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=1600, nel primo passo, è stata ripartita in parti uguali tra Modena e il resto della provincia. Nel secondo passo, la dimensione n=800 è stata ripartita proporzionalmente tra i vari distretti secondo la corrispondente numerosità di USS, N_d, ottenendo la dimensione campionaria per distretto, $n_d = n N_d / N$. Nel terzo passo, la dimensione n_d è stata ripartita proporzionalmente tra le UPS campionarie del *d*-esimo strato, per mantenere un certo equilibrio tra le numerosità delle UPS campionarie a livello distrettuale. Si è ottenuto, cosí, $n_{dc} = n_d \, N_{dc} \Big/ \sum_{c=1}^{c_d} N_{dc}$, dove c_d indica il numero di comuni nel campione del *d*-esimo distretto. I risultati della selezione dei comuni sono esposti nella Tabella 3, dove l'approssimazione nel calcolo delle n_{dc} è stata eseguita sempre per eccesso e ciò ha generato un aumento di 37 unità in piú delle 1600 previste. I comuni AR che denominano il distretto sono Carpi, Mirandola, Modena, Sassuolo, Pavullo nel Frignano, Vignola, e Castelfranco Emilia; mentre Formigine è AR e appartiene al distretto di Sassuolo. I comuni NAR inclusi nel campione sono Novi di Modena (D1), Concordia sulla Secchia (D2), Finale Emilia (D2), San Prospero (D2), Fiorano modenese (D4), Serramazzoni (D5), Palagano (D5), Prignano sulla Secchia (D5), Montese (D5), Savignano sul Panaro (D6), Spilamberto (D6), Bomporto (D7), Nonantola (D7).

Tabella 3 – Numero di famiglie (N_{dc}), dimensione campionaria proporzionale (n_{dc}), e numero totale per distretto (n_d) per i comuni inclusi (selezionati) nel campione della provincia di Modena suddivisi per distretto d'area al 31/12/2004[(a)]

D	COMUNE	N_{dc}	n_{dc}	n_d	D	COMUNE	N_{dc}	n_{dc}	n_d
D1	**Carpi**	26019	141	164	**D2**	**Mirandola**	9211	62	142
	Novi di Modena	4165	23			Concordia sulla Secchia	3366	23	
						Finale Emilia	6298	43	
						San Prospero	1994	14	
D3	**Modena**	78962	800	800					
D4	**Sassuolo**	16463	82/164	400	**D5**	**Pavullo nel Frignano**	6573	62	133
	Fiorano modenese	6011	30/ 60			Serramazzoni	3230	31	
	Formigine	11666	58/116			Palagano (S)	1118	11	
	Maranello[(*)]	6042	60			Prignano s. Secchia (S)	1423	14	
						Montese (V)	1532	15	
D6	**Vignola**	9251	59/137	400	**D7**	**Castelfranco Emilia**	11162	63	111
	Castelnuovo Rangone[(*)]	4980	73			Bomporto	3093	18	
	Castelvetro[(*)]	3950	58			Nonantola	5364	30	
	Savignano sul Panaro	3403	22/ 50						
	Spilamberto	5577	36/ 82						
						Totale Provincia	**272238**	**[(o)]1637**	**[(o)]2150**

[(a)] La data di riferimento è antecedente (circa un anno) alle date di riferimento delle tabelle relative al comune di Modena perché al momento della realizzazione del piano di campionamento non erano ancora disponibili i dati provinciali della popolazione.
[(*)] Il comune è stato aggiunto a causa dell'espansione del campione ordinario del distretto d'area. Le dimensioni del campione, senza espansione, sono riportate prima del simbolo "/" per fornire un'idea dell'aumento che ne è conseguito.
[(o)] Il primo totale (1637) è riferito al campione senza espansioni, il secondo totale (2150) è riferito al campione con espansioni.

Le dimensioni del campione, nei distretti di Sassuolo e Vignola, sono state aumentate perché le comunità locali hanno fornito un contributo per condurre le interviste mancanti a raggiungere una numerosità sufficientemente alta, 400 unità, per ottenere una precisione adeguata di alcune informazioni. Per esempio, fornisce un errore assoluto del 5% per le stime delle proporzioni della popolazione. Tutti i comuni del distretto sono stati inclusi, allora, nel campione, con probabilità di inclusione pari a 1, come per i comuni AR, al fine di garantire una maggiore *rappresentatività* territoriale.

2.2.2. Stratificazione nel comune di Modena

Nel comune di Modena, le USS si sono stratificate secondo l'ampiezza della famiglia, l'età, e il genere del capofamiglia perché si ha la collaborazione piena del personale degli uffici competenti, una maggiore esperienza e efficienza nell'elaborazione dei dati. La data di riferimento della popolazione obiettivo è, quindi, piú recente rispetto a quella dei restanti comuni della provincia perché si opera in diretto contatto con gli uffici e si ricevono i dati un po' prima della selezione. Per la stabilità della popolazione nel tempo, tuttavia, non si alterano in modo sensibile i risultati delle dimensioni campionarie e delle stime. Si è proceduto, quindi, secondo lo schema seguíto nella precedente indagine (Lalla, 2003), brevemente descritto di séguito.

La stratificazione sull'ampiezza della famiglia è utile perché si suppone correlata con la distribuzione del reddito e è stata suddivisa in I=4 classi, come si può osservare nella distribuzione marginale (delle righe) della Tabella 4: famiglie con un solo membro, con due membri, con tre membri, con quattro o più membri.

L'età del capofamiglia è un altro carattere distintivo tra le famiglie e si è optato per una suddivisione in cinque classi, $J = 5$, per motivi di uniformità: 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 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, ma l'efficienza della stratificazione dipende dalla possibilità di costruire strati con una variabilità minore di quella totale della $\mathscr{Y}$: dai dati della Tabella 4, la scelta sembra piú di ordine logico.

Le informazioni raccolte sul reddito, tramite l'indagine condotta nel 2002, si possono utilizzare per determinare la dimensione del campione, infatti, come già mostrato per i distretti. I dati dell'indagine condotta nel 2002, rilevanti a tal fine, sono riportati nella Tabella 4, con le precisioni relative e le dimensioni del campione per l'indagine da realizzare, calcolate utilizzando la media, la deviazione standard delle celle (strati), e un errore relativo pari al 15,9% e uguale in tutte le celle: si è eliminata la suddivisione per genere al fine di aumentare la numerosità nella cella e ottenere una maggiore stabilità nei dati. Il confronto tra le dimensioni dell'indagine precedente e

quella da eseguire suggerisce che, negli strati determinati dai capifamiglia con piú di 64 anni e una dimensione famigliare maggiore di tre membri, si deve estrarre un numero di USS maggiore del numero ottenuto con l'allocazione proporzionale o quanto meno in quegli strati si devono effettuare piú sforzi per non avere mancate risposte.

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:

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

dove $n_{MO;ijk}$ è il numero di famiglie 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 la dimensione campionaria nel comune di Modena (pari a 800 famiglie), $N_{MO;ijk}$, è il numero di famiglie nello strato *ijk* riportato in Tabella 5, N_{MO} è il numero totale di famiglie, e il simbolo $\lfloor \cdot \rfloor$ indica la parte intera dell'argomento: l'arrotondamento è eseguito, quindi, per eccesso dato il «+1» nella (9).

Tabella 4 – Numero di famiglie rilevate nell'indagine precedente (n-2002), reddito medio e deviazione standard (DS) in migliaia di euro, errore relativo osservato, e numero di famiglie da rilevare nell'indagine corrente (n-2006) ottenute da un errore relativo pari al 15,9% per numero di componenti la famiglia e per classi di età del capofamiglia

Numero componenti	Dati cella	Classi di età del capofamiglia					Totale
		18-34	35-49	50-64	65-74	75 +	
1 componente	**n-2002**	30	40	24	32	37	**163**
	Media	26,52	27,54	25,12	18,91	19,17	23,17
	DS	11,94	12,40	14,75	10,63	10,72	12,42
	Err. Rel.	0,164	0,142	0,239	0,198	0,185	**0,0837**
	n-2006	**32**	**32**	**54**	**50**	**50**	**218**
2 componenti	**n-2002**	21	31	46	45	42	**185**
	Media	31,63	39,43	46,42	40,59	33,07	39,11
	DS	8,03	22,56	29,47	26,76	19,96	24,26
	Err. Rel.	0,112	0,208	0,187	0,197	0,187	**0,0910**
	n-2006	**11**	**52**	**64**	**68**	**57**	**252**
3 componenti	**n-2002**	16	47	48	17	9	**137**
	Media	37,52	47,45	56,77	45,99	48,18	49,44
	DS	10,20	24,04	36,29	20,20	19,45	27,81
	Err. Rel.	0,138	0,148	0,184	0,218	0,275	**0,0958**
	n-2006	**12**	**41**	**64**	**31**	**26**	**174**
4 componenti e più	**n-2002**	10	48	34	7	5	**104**
	Media	45,09	44,75	57,81	51,61	53,87	49,66
	DS	15,17	19,60	31,67	28,84	15,79	24,58
	Err. Rel.	0,219	0,126	0,189	0,428	0,285	**0,0967**
	n-2006	**18**	**31**	**47**	**47**	**14**	**157**
Totale	**n-2002**	**77**	**166**	**152**	**101**	**93**	**589**
	Media	32,11	40,65	48,11	35,58	28,14	38,40
	DS	12,57	21,41	32,28	24,55	18,71	24,85
	Err. Rel.	**0,0891**	**0,0815**	**0,1085**	**0,1372**	**0,1377**	**0,0531**
	n-2006	**73**	**156**	**229**	**196**	**147**	**801**

L'arrotondamento per eccesso della dimensione del campione per strato, $n_{MO;ijk}$, ha generato un aumento di 20 unità: n_{MO}=820, come risulta dalla Tabella 6.

Tabella 5 – 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 20/12/2005

Numero componenti	Genere	Classi di età del capofamiglia					Totale
		18-34	35-49	50-64	65-74	75 +	
1 componente	Uomo	3503	3652	1895	1006	1345	**11401**
	Donna	2279	2591	2474	2769	5914	**16027**
2 componenti	Uomo	1409	2051	4118	4930	4248	**16756**
	Donna	1015	1824	1572	889	1222	**6522**
3 componenti	Uomo	1089	3935	4734	1819	796	**12373**
	Donna	567	1470	740	226	321	**3324**
4 componenti e più	Uomo	706	5731	3581	723	335	**11076**
	Donna	343	1176	296	175	210	**2200**
Totale	Uomo	**6707**	**15369**	**14328**	**8478**	**6724**	**51606**
	Donna	**4204**	**7061**	**5082**	**4059**	**7667**	**28073**

Tabella 6 – 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 20/12/2005

Numero componenti	Genere	Classi di età del capofamiglia					Totale
		18-34	35-49	50-64	65-74	75 +	
1 componente	Uomo	36	37	20	11	14	**118**
	Donna	23	27	25	28	60	**163**
2 componenti	Uomo	15	21	42	50	43	**171**
	Donna	11	19	16	9	13	**68**
3 componenti	Uomo	11	40	48	19	8	**126**
	Donna	6	15	8	3	4	**36**
4 componenti e più	Uomo	8	58	36	8	4	**114**
	Donna	4	12	3	2	3	**24**
Totale	Uomo	**70**	**156**	**146**	**88**	**69**	**529**
	Donna	**44**	**73**	**52**	**42**	**80**	**291**
Totale	**U+D**	**114**	**229**	**198**	**130**	**149**	**820**

2.3. Selezione delle unità statistiche campionarie

Si deve eseguire 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 nell'indagine precedente e in altre indagini simili, date le difficoltà nella rilevazione di informazioni 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). Il tasso finale di non risposta è dell'ordine del 65% nell'indagine sui bilanci delle famiglie condotta dalla Banca d'Italia (2004, pp. 35-39): piú bassa per la componente longitudinale o *panel* (25,5%), e molto piú alta per la componente trasversale o non *panel* (76,2%). 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). Si evince, quindi, che la dimensione ipotizzata ottimale deve essere almeno triplicata per ottenere il numero desiderato di unità statistiche realmente rilevate; ossia, ogni unità campionaria dovrebbe avere due unità aggiuntive con funzione di *riserva*, se tutte le unità estratte fossero contattate. Per aumentare la probabilità di intervistare una unità statistica prima di esaurire le sue corrispondenti riserve, se ne predispongono tre per ogni unità del campione; pertanto, il numero di USS estratte sarà pari al quadruplo della dimensione del campione sopra determinata.

Nelle indagini complesse, la difficoltà piú rilevante è l'indisponibilità o «rifiuti»: il 60,7% nelle indagini della Banca d'Italia (2004, pp. 35-39) e il 56% nell'indagine precedente, che presenta anche una notevole variabilità territoriale. Le altre mancate partecipazioni derivano dall'impossibilità di contattare la famiglia per telefono o di trovare qualcuno a casa quando ci si reca presso l'abitazione («irreperibili»): il 5% nelle indagini della Banca d'Italia (2004, pp. 35-39) con quattro punti percentuali di scarto tra le due componenti (*panel*, non *panel*) e il 27% circa nell'indagine precedente. La Banca d'Italia le distingue dalle ineleggibili —famiglie non esistenti all'indirizzo anagrafico per errori, decessi, o trasferimenti— che sono circa il 2%. Nell'indagine precedente non è stata eseguita tale distinzione o ricerca delle cause di irreperibilità perché, da un lato, migliora l'«efficienza» della rilevazione o la conoscenza dei movimenti delle unità statistiche, dall'altro lato, comporta un aumento di costi e di tempi spesi nei rapporti con gli uffici anagrafici dei comuni. I dati della Banca d'Italia (2004) sembrano mostrare anche una lieve flessione rispetto al passato (2002). Nella quota non *panel*, le interviste completate sono il 34,3% contro il 38,3%; le famiglie indisponibili sono il 60,7% contro il 57,2%: le differenze potrebbero derivare dal caso, da una minore accuratezza degli operatori, da una maggiore diffusione del diritto alla riservatezza dei dati personali.

Nel secondo stadio del campionamento si selezionano, quindi, le famiglie, utilizzando la lista anagrafica di ciascun comune e il metodo del campionamento sistematico. Tale metodo fornisce stime non distorte, quando il passo di campionamento, a_{dc}, è un numero intero; ossia, il rapporto $a_{dc} = N_{dc}/n_{dc}$ ha resto uguale a zero. Si ha, allora, un campionamento casuale semplice senza reimmissione e con probabilità uguali (Särndal, Swensson, Wretman, 1992). Se il resto è diverso da zero, si può ricorrere al campionamento sistematico circolare: dato 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 selezionano le famiglie che nella lista anagrafica occupano le posizioni date da:

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

dove $1_{[\bullet)}[\cdot]$ è la funzione indicatrice che vale 1, se l'argomento appartiene all'insieme specificato nell'indice, 0 altrimenti. L'estrazione inizia, quindi, dal punto di partenza casuale ρ e prosegue «lungo» la lista, ricominciando all'inizio dopo avere raggiunto la fine della lista.

La selezione delle famiglie dalla lista anagrafica dei comuni è eseguita da un dipendente, pertanto, il sistema circolare di estrazione può generare diverse difficoltà. Pare piú conveniente, quindi, fornire un punto di partenza casuale, ρ, che sia all'inizio della lista. Al momento dell'estrazione si chiederà all'addetto quanto è la consistenza

della lista, ossia N_{dc}. Si genera un numero casuale con distribuzione uniforme discreta in $[1, N_{dc}]$. Sia r_{dc}. Il punto di partenza casuale, ρ, sarà dato da $\rho = r_{dc}/n_{dc} - \lfloor r_{dc}/n_{dc} \rfloor$, ossia dal resto della divisione r_{dc}/n_{dc}, se questo è maggiore di zero; se il resto è uguale a zero, allora $\rho = a_{dc}$. Tale metodo ha il vantaggio sia di partire sempre dall'inizio della lista, sia di generare stime non distorte del totale, della media, e della proporzione (Levy, Lemeshow, 1991, pp. 82-84). Tutti i membri delle famiglie, che convivono a qualsiasi titolo nello stesso nucleo, sono inclusi nel campione.

Per sopperire all'eventuale insuccesso degli intervistatori si estrae la cosiddetta 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, come detto, uguale al triplo della dimensione obiettivo.

L'estrazione sarà effettuata ordinando la lista per strada e numero civico. La prima USS sarà la famiglia che si trova nell'ordine corrispondente al punto di partenza casuale ρ. Le tre USS che si trovano nelle tre posizioni successive ($\rho+1$, $\rho+2$, $\rho+3$) vanno a costituire la lista suppletiva o lista di riserva, che è stata estratta, quindi, assieme alle unità campionarie.

Nel comune di Modena si è deciso di creare una componente longitudinale (v. *infra*). Per semplicità, la quota longitudinale (di individui che hanno partecipato nel 2002 e partecipano ancora all'indagine corrente) è stata fissata pari al 50% del totale, ossia 400 USS. Nell'indagine della Banca d'Italia (2004), la componente longitudinale effettiva è, infatti, il 45% del totale. L'indagine svolta in precedenza non aveva previsto, tuttavia, la possibilità di una ripetizione, dato i costi; pertanto, non si è chiesto alle famiglie selezionate se erano disponibili a ripetere in futuro l'intervista. Allo stato attuale non è facile prevedere quante saranno le unità effettivamente intervistate; in base a altre esperienze si può prevedere che piú del 25% non parteciperà e circa il 3% non sarà reperibile. Si suggerisce di procedere, allora, come specificato di séguito.

Il primo passo consiste nella verifica della disponibilità delle 589 famiglie, intervistate nel 2002, a essere intervistate di nuovo nel 2006. In ogni cella di Tabella 6 si deve intervistare la metà del numero di famiglie ivi riportato. Se il numero di famiglie disponibili per cella è superiore, allora si eliminano con una selezione sistematica quelle in eccesso e fungeranno da riserve. Se una famiglia decidesse di non partecipare piú all'indagine durante l'intervista, allora si potrebbe scegliere una sostituta dall'elenco delle eccedenti. Si deve procedere, quindi, súbito alla rilevazione delle disponibili per conoscere con certezza il numero di USS della componente longitudinale, $n_{MO;L;ijk}$.

Si devono includere nel *panel* tutte le famiglie che si sono formate dalle unità originarie, cosí opera anche la Banca d'Italia (Banca d'Italia, 2004; Kasprzyk *et al.*, 1989); ma diventa difficile intervistare tutte le famiglie che si sono trasferite in un altro comune, specialmente se non è un comune nel campione o è fuori provincia.

Terminata la rilevazione della componente longitudinale, $n_{MO;L;ijk}$, si determina il numero di USS rimanenti per strato, componente trasversale o $n_{MO;T;ijk}$, e ci si può rivolgere all'anagrafe del comune di Modena per l'estrazione del campione di famiglie, nei varî strati con ampiezze pari a $(n_{MO;ijk} - n_{MO;L;ijk})$, dove $n_{MO;ijk}$ è in Tabella 6.

3. I fattori di riporto alla popolazione obiettivo

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

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

dove w_i è il *peso*, che non dipende dal numero d'ordine delle osservazioni, ma può dipendere dal tipo di campionamento adottato e dall'etichetta che individua l'unità statistica selezionata.

Si consideri, ora, la provincia di Modena, stratificata per distretto d'area, *d*, e per comune, *c*. Sia Y_{dci} il valore di $\mathscr{Y}$ per l'*i*-esima famiglia nel *c*-esimo comune del *d*-esimo strato. Il totale di $\mathscr{Y}$ è dato dalla somma estesa a tutte le unità statistiche di $\wp$:

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

dove D è il numero di distretti, C_d è il numero di comuni nel *d*-esimo distretto, N_{dc} è il numero di famiglie nel *c*-esimo comune del *d*-esimo distretto.

Un campionamento probabilistico a due stadi generi un campione di *n* unità, estratte senza ripetizione, in cui le UPS e 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, dopo l'estrazione della *c*-esima UPS, dove n_{dc} indica il numero di famiglie nel campione del *d*-esimo distretto del *c*-esimo comune. Lo stimatore 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.

Gli stimatori associati al campionamento a piú stadi sono complessi e le varianze degli stimatori assumono espressioni complicate. In generale, la varianza dello stimatore del totale, $\hat{Y}$, è data da (Cicchitelli, Herzel, Montanari, 1997, p. 194):

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

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.

3.1. I fattori di riporto alla popolazione obiettivo per la provincia di Modena

Il «peso» di ogni unità campionaria che partecipa all'indagine indica, in un certo senso, il numero di USS del comune e/o del distretto di appartenenza «rappresentate» dall'unità stessa. Nelle espressioni per il calcolo dei pesi, si distinguerà tra i comuni AR, dove $\pi_{dc}=1$, e i distretti con una o piú UPS tipo NAR:

$$w_{dci}^{AR}=\frac{1}{\pi_{dc}}\frac{1}{\pi_{dci}}=\frac{N_{dc}^{AR}}{n_{dc}^{AR}}, \tag{14}$$

$$w_{dci}^{NAR}=\frac{1}{\pi_{dc}}\frac{1}{\pi_{dci}}\equiv\frac{N_{d}^{NAR}}{c_{d}^{NAR}N_{dc}^{NAR}}\frac{N_{dc}^{NAR}}{n_{dc}^{NAR}}=\frac{1}{c_{d}^{NAR}}\frac{N_{d}^{NAR}}{n_{dc}^{NAR}}, \tag{15}$$

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

Nella Tabella 7 sono riportati i pesi w_{dc} per le famiglie-campione, calcolati secondo la (14) per i comuni AR e secondo la (15) per i comuni NAR. I valori dei pesi sono molto diversi tra loro perché ogni UPS stima una parte della popolazione dello strato data dal reciproco del numero di UPS estratte, ossia di c_d^{NAR}. Si hanno cosí valori assai elevati nei comuni piccoli e con poche unità incluse nel campione.

Le mancate partecipazioni introducono un fattore di disturbo sicché si dovranno usare pesi diversi dal peso base al fine di correggere per le mancate collaborazioni. In generale, si considerano la non appartenenza alla popolazione di riferimento, l'emigrazione o l'estinzione, e la non rintracciabilità dell'unità che può includere sia l'emigrazione, sia l'estinzione, sia gli errori di registrazione negli archivi. Si ignorano, per semplicità, tali distinzioni e si trattano tutti come non rispondenti, anche se ne potrebbe conseguire una sovrastima della popolazione di riferimento. Per il calcolo dei pesi finali, se si considerasse la probabilità di rintracciare una unità e la probabilità di ottenere la sua partecipazione, si otterrebbe comunque la semplice espressione seguente:

$$w_{dc} = \frac{1}{\pi_{dc}} \cdot \frac{1}{\pi_{r;dc}} \cdot \frac{1}{\pi_{p;dc}} = \frac{N_{dc}^{q}}{c_d^q \, n_{dc}} \cdot \frac{n_{dc}}{n_{r;dc}} \cdot \frac{n_{r;dc}}{n_{p;dc}} = \frac{N_{dc}^{q}}{c_d^q \, n_{p;dc}} = \frac{1}{\pi_{p;dc}^{*}}, \qquad (16)$$

dove $\pi_{r;dc}$ è la probabilità che l'unità sia rintracciata, $\pi_{p;dc}$ è la probabilità che l'unità partecipi all'indagine, n_{dc} è il numero di unità selezionate nel comune *c* del distretto *d*, $n_{r;dc}$ è il numero di unità rintracciate, e $n_{p;dc}$ denota il numero di unità che partecipano all'indagine e rispondono alle domande del questionario. Il numeratore, N_{dc}^{q}, indica la popolazione di riferimento: per un comune autorappresentativo, $q = AR$ e $c_d^{AR} = 1$, per un comune non autorappresentativo, $q = NAR$ e $c_d^{NAR} \geq 1$. Il peso finale è dato da $1/\pi_{p;dc}^{*}$, dove $\pi_{p;dc}^{*}$ 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.

Tabella 7 – Numero di famiglie nei comuni campione (N_{dc}) e nel campione (n_{dc}), e pesi relativi alle famiglie (w_{dc}) per i comuni campione della provincia di Modena al 31/12/2004

D	COMUNE	N_d	n_{dc}	w_{dc}	D	COMUNE	N_d	n_{dc}	w_{dc}
D1	**Carpi**	26019	141	184,5319	**D2**	**Mirandola**	9211	62	148,5645
	Novi di Modena	4165	23	237,1304		Concordia sulla Secchia	3366	23	342,3043
						Finale Emilia	6298	43	183,0930
						San Prospero	1994	14	562,3571
	Totale senza AR	***12786***				***Totale senza AR***	***23619***		
D3	**Modena**	78962	800						
D4	**Sassuolo**	16463	82/164	100,3841	**D5**	**Pavullo nel Frignano**	6573	62	106,0161
	Fiorano modenese	6011	30/ 60	100,1833		Serramazzoni	3230	31	178,9113
	Formigine	11666	58/116	100,5690		Palagano (S)	1118	11	504,2045
	Maranello(*)	6042	60	100,7000		Prignano s. Secchia (S)	1423	14	396,1607
						Montese (V)	1532	15	369,7500
	Totale senza AR	***23719***				***Totale senza AR***	***22185***		
D6	**Vignola**	9251	59/137	67,5255	**D7**	**Castelfranco Emilia**	11162	63	177,1746
	Castelnuovo Rangone(*)	4980	73	68,2192		Bomporto	3093	18	399,3889
	Castelvetro(*)	3950	58	68,1034		Nonantola	5364	30	239,6333
	Savignano sul Panaro	3403	22/ 50	68,0600					
	Spilamberto	5577	36/ 82	68,0122					
	Totale senza AR	***17910***				***Totale senza AR***	***14378***		

La soluzione adottata è la piú semplice per compensare le stime dalle mancate partecipazioni; altre strategie, piú sofisticate e complesse, sono difficili da applicare alle indagini su larga scala (Little, Rubin, 1987; Rubin, 1988). Gli stimatori diventano, però, non lineari e le loro varianze aumentano (Kish, 1990, 1992), specialmente se le correzioni apportate non sono correlate con le variabilità negli strati (Bethlehem, Keller,

1987; Potter, 1990); infatti, il peso dei rispondenti incrementa perché «devono rappresentare» anche le unità che rifiutano di partecipare o che sono irreperibili.

3.2. I fattori di riporto alla popolazione obiettivo per il comune di Modena

Nel comune di Modena (d=3) si è adottata l'allocazione proporzionale, che è autoponderante, tra gli strati determinati dalla classe di ampiezza della famiglia, i (dove $i=1,\cdots,I\,[=4]$), dalla classe di età del capofamiglia, j (dove $j=1,\cdots,J\,[=5]$), dal genere del capofamiglia, k (dove $k=1,2\,[=K]$); pertanto, la stima del totale del carattere $\mathscr{Y}$ è:

$$\hat{Y}_{d=3}=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{n_{ijk|d=3}}\frac{1}{\pi_{ijk|d=3}}y_{ijkl|d=3}\,. \qquad (17)$$

Tale statistica è detta anche *stimatore per espansione* perché nel caso di un campionamento casuale semplice o autoponderante diventa semplicemente il prodotto della corrispondente grandezza campionaria moltiplicata per l'inverso della frazione di campionamento: $\hat{Y}_{d=3}=\left(N_{d=3}/n_{d=3}\right)\sum_{ijk}y_{ijk|d=3}=N_{d=3}\bar{y}_{d=3}$. Ogni unità del campione rappresenta, quindi, $N_{d=3}/n_{d=3}$ unità della popolazione; perciò, il fattore $N_{d=3}/n_{d=3}$ è detto anche *coefficiente di espansione*. In altre parole, all'interno di ogni strato si ha un peso che corrisponde proprio all'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 che l'universo di riferimento è il comune di Modena, nel quale, all'interno di ogni strato il fattore di espansione o peso è dato da

$$w_{ijk}=\frac{1}{\pi_{ijk}}=\frac{N_{ijk}}{n_{ijk}}\,. \qquad (18)$$

Nella Tabella 8 sono riportati i pesi, w_{ijk}, calcolati secondo la (18), che dovrebbero essere esattamente uguali. Le differenze derivano dagli arrotondamenti per eccesso della dimensione n_{ijk} e dal troncamento della parte decimale dei passi.

Tabella 8 – Pesi base, w_{ijk}, per numero di componenti la famiglia, per classi di età, e per genere del capofamiglia, nel comune di Modena al 31/12/2005

N. componenti	Genere	Classi di età del capofamiglia				
		<=34 anni	35-49 anni	50-64 anni	65-74 anni	>=75 anni
1 componente	M	97,31	98,70	94,75	91,45	96,07
	F	99,09	95,96	98,96	98,89	98,57
2 componenti	M	93,93	97,67	98,05	98,60	98,79
	F	92,27	96,00	98,25	98,78	94,00
3 componenti	M	99,00	98,38	98,63	95,74	99,50
	F	94,50	98,00	92,50	75,33	80,25
4 componenti e piú	M	88,25	98,81	99,47	90,38	83,75
	F	85,75	98,00	98,67	87,50	70,00

Le mancate partecipazioni introducono un fattore di disturbo, che anche per il comune di Modena è variabile per strato, sicché i pesi differiranno per strato ancora di piú, anche se si tratta di un campione autoponderante; infatti, nei domini dove non ci

sarà una copertura completa si osserverà un aumento del peso piú alto del suo valore «medio». Alla fine della rilevazione, l'espressione per il calcolo del peso finale è simile alla (16), basta sostituire all'indice di distretto-comune, *dc*, l'indice di strato, *ijk*: $w_{ijk} = N_{ijk} / n_{p;ijk}$. Le altre osservazioni (§3.1) restano immutate. I valori finali dei pesi saranno descritti, pertanto, in fase di consuntivo dell'indagine.

3.3. Normalizzazione dei pesi all'unità

Per eseguire test statistici e/o stimare i parametri di modelli rappresentativi della realtà indagata non si possono usare i pesi dati dalla (16) e (18) perché alterano la numerosità campionaria e, quindi, le probabilità di significatività relative alle ipotesi da verificare. Per rimediare a tali inconvenienti si possono «scalare» i pesi in modo che la loro somma sia uguale all'unità (Verma, 1995). Si incorpora, cosí, la struttura del campione nella determinazione degli stimatori e non si altera la numerosità campionaria. Anche qui, si specifica solo il procedimento e si rinvia il calcolo definitivo a fine rilevazione.

3.3.1. Normalizzazione nella provincia di Modena

Per compensare la non proporzionalità nella scelta delle unità statistiche campionarie e le mancate partecipazioni, si può utilizzare un insieme di pesi, w^*_{dc}, che mantengano inalterate le caratteristiche del campione, ossia soddisfacciano il vincolo:

$$\sum_{d=1}^{D}\sum_{c=1}^{c_d} w^*_{dc} n_{dc} = n\,.$$

Il peso dato dal rapporto tra i pesi «originari», $1/\pi^*_{p;dc}$, e un peso medio, $1/\bar{\pi}^*_p$, può soddisfare la condizione data. Le grandezze figurano al denominatore, sicché si può calcolare la media usando come aggregazione la funzione somma delle quantità inverse perché tutte positive (sono «pseudo-probabilità»). Si definisce, quindi, la funzione $f(\cdot)$ come somma degli inversi dei valori osservati, $f(y_1,\cdots,y_n) = \sum_{i=1}^{n} \frac{1}{y_i}$, da cui si ottiene la sequenza di relazioni:

$$\sum_{d=1}^{D}\sum_{c=1}^{c_d}\sum_{i=1}^{n_{dc}} \frac{1}{\pi^*_{p;dc}} = \sum_{d=1}^{D}\sum_{c=1}^{c_d}\sum_{i=1}^{n_{dc}} \frac{1}{\bar{\pi}^*_p} = \sum_{d=1}^{D}\sum_{c=1}^{c_d} \frac{n_{dc}}{\bar{\pi}^*_p} \quad \Leftrightarrow \quad \bar{\pi}^*_p = \frac{\sum_{d=1}^{D}\sum_{c=1}^{c_d} n_{dc}}{\sum_{d=1}^{D}\sum_{c=1}^{c_d} \frac{n_{dc}}{\pi^*_{p;dc}}},$$

dove $\bar{\pi}^*_p$ è la media armonica delle probabilità di selezione per i vari comuni, *dc*, nel campione. Il peso normalizzato a uno, per ogni comune campione sarà dato dal rapporto tra i pesi effettivi finali $\pi^*_{p;dc}$ e il peso medio dato dall'inverso della media armonica, $1/\bar{\pi}^*_p$. Allora, il peso normalizzato a uno, w^*_{dc}, che rispetta il vincolo (Lalla, 2003) è

$$w^*_{dc} = \frac{\bar{\pi}^*_p}{\pi^*_{p;dc}} = \frac{N^q_{dc}}{c_d\, n_{p;dc}} \times \frac{n}{N}\,. \qquad (19)$$

3.3.2. 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, $1/\bar{\pi}^*_p$, in modo da soddisfare le condizioni (a) e (b). Come per la provincia, si otterrà, adattando i simboli agli strati *ijk*:

$$\bar{\pi}^*_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} \Bigg/ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{ijk}}{\pi^*_{p;ijk}},$$

dove $\bar{\pi}^*_p$ 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, $1/\bar{\pi}^*_p$:

$$w^*_{ijk} = \frac{\bar{\pi}^*_p}{\pi^*_{p;ijk}} = \frac{N_{ijk}}{n_{p;ijk}} \times \frac{n}{N}\,. \tag{20}$$

Ossia, 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: $w^*_{ijk} = W_{ijk}/w_{ijk} = \left(N_{ijk}/N\right):\left(n_{p;ijk}/n\right)$. Questi pesi w^*_{ijk} alterano la struttura delle dimensioni campionarie per strato rispetto al campione effettivo.

3.4. Varianza della stima del reddito totale

Il totale della caratteristica $\mathscr{Y}$ è dato dalla (11) e il suo stimatore derivato dalla (12) è

$$\hat{Y} = \sum_{d=1}^{D} \hat{Y}_d = \sum_{d=1}^{D}\sum_{c=1}^{c_d}\sum_{i=1}^{n_{dc}} w_{dc}\, y_{dci}\,, \tag{21}$$

dove y_{dci} è il reddito dell'*i*-esima unità campionaria, nel *c*-esimo comune del *d*-esimo distretto. Con ciò si assume, come già detto, che le unità incluse nel campione rappresentano anche le altre $(w_{dc} - 1)$ unità della popolazione che non sono state selezionate. All'inizio del processo di elaborazione dei dati si usano 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 famiglie sono selezionate senza reimmissione e con probabilità uguali, lo stimatore $\hat{Y}$, del totale di $\mathscr{Y}$, è dato da $\hat{Y}_{AR} = \sum_{d=1}^{D}\sum_{c=1}^{C_d^{AR}} N_{dc}^{AR}\, \bar{y}_{dc}^{AR}$ perché nei distretti vi sono piú AR. Le stime della varianza campionaria risultano corrette e fornite dall'espressione seguente:

$$V\left(\hat{Y}_{AR}\right) = \sum_{d=1}^{D}\sum_{c=1}^{C_d^{AR}} \left(N_{dc}^{AR}\right)^2 \frac{{}_{AR}s^2_{2;dc}}{n_{dc}^{AR}} \left(1 - f_{2;dc}^{AR}\right), \tag{22}$$

dove ${}_{AR}s^2_{2;\,dc}$ è la varianza campionaria della $\mathscr{Y}$ e $f^{AR}_{2;\,dc} = n^{AR}_{dc} \big/ N^{AR}_{dc}$ è la frazione di unità nel campione del *c*-esimo comune AR, del *d*-esimo distretto di area.

Negli strati NAR con un solo comune selezionato si ha $c^{NAR}_d = 1$. 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):

$$\hat{Y}^{NAR}_d = N^{NAR}_d \; \hat{\bar{y}}^{NAR}_d = N^{NAR}_d \; \bar{y}^{NAR}_d , \tag{23}$$

dove $\hat{\bar{y}}^{NAR}_d$ è lo stimatore della media nello strato NAR, $\bar{y}^{NAR}_d$ è la media osservata nell'unico campione del comune campione nello strato NAR. La sua varianza è data da

$$V\left(\hat{Y}^{NAR}_d\right) = N^{NAR}_d \left[\sum_{c=1}^{C^{NAR}_d} \left(N^{NAR}_{dc} - n^{NAR}_{dc}\right) \frac{{}_{NAR}S^2_{2;dc}}{n^{NAR}_{dc}} + \sum_{c=1}^{C^{NAR}_d} N^{NAR}_{dc} \left(\bar{Y}^{NAR}_{dc} - \bar{\bar{Y}}^{NAR}_d\right)^2 \right], \tag{24}$$

dove ${}_{NAR}S^2_{2;\,dc}$ è la varianza e $\bar{Y}^{NAR}_{dc}$ è la media della popolazione del *c*-esimo comune NAR del *d*-esimo distretto, mentre $\bar{\bar{Y}}^{NAR}_d$ è la media totale dello strato NAR del *d*-esimo distretto. Senza dati 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 sempre dalla (22), con una varianza

$$V\left(\hat{Y}^{NAR}_d\right) = \sum_{c=1}^{C^{NAR}_d} \sum_{c'\neq c}^{C^{NAR}_d} \left(\frac{\pi_{dc}\pi_{dc'}}{\pi_{dcc'}} - 1 \right) \left(\frac{\hat{Y}_{dc}}{\pi_{dc}} - \frac{\hat{Y}_{dc'}}{\pi_{dc'}} \right)^2 + \sum_{c=1}^{C^{NAR}_d} \frac{(N^{NAR}_{dc})^2}{\pi_{dc}} \frac{{}_{NAR}S^2_{2;dc}}{n^{NAR}_{dc}} \left(1 - f^{NAR}_{2;dc}\right), \tag{25}$$

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 è un po' laboriosa.

3.5. Post-stratificazione

Per alcuni caratteri non presenti nella lista, possono essere disponibili dati in forma di tabelle sia nel campione e sia nella popolazione; per esempio, si conosce la distribuzione per classe di età $(j = 1, \ldots, J)$ e per genere $(k = 1,\ 2\ (= K))$ degli individui. La loro conoscenza consente di costruire $J \times K$ post-strati. In ogni distretto *d* e in ogni comune (AR, NAR), si può costruire uno stimatore che ricade in ogni post-strato *jk*:

$$\hat{Y}_d = \sum_{j=1}^{J} \sum_{k=1}^{K} N^{AR}_{djk} \; \bar{y}^{AR}_{djk} + \sum_{j=1}^{J} \sum_{k=1}^{K} N^{NAR}_{djk} \; \bar{y}^{NAR}_{djk} . \tag{26}$$

I soggetti inclusi nel campione avranno, in questa procedura, dei nuovi pesi che si ottengono immediatamente dall'espressione precedente in una forma simile alla (16):

$$w^{AR}_{PS;\,djk} = \frac{N^{AR}_{djk}}{n^{AR}_{djk}}; \qquad\qquad w^{NAR}_{PS;djk} = \frac{N^{NAR}_{djk}}{n^{NAR}_{djk}}; \tag{27}$$

dove n^{AR}_{djk} e n^{NAR}_{djk} sono, rispettivamente, le dimensioni dei campioni dei comuni AR e NAR nel post-strato *jk* del *d*-esimo distretto. L'uso di tali pesi generano gli *stimatori post-stratificati semplici*, ma nelle indagini complesse le probabilità di selezione variano in ciascun post-strato per effetto del disegno di campionamento o per l'aggregazione. Si può ottenere un miglioramento delle stime, quindi, con il cosiddetto

stimatore di Hajek:

$$\hat{Y}_{djk} = N_{djk}\left(\frac{\tilde{Y}_{djk}}{\tilde{N}_{djk}}\right) = \tilde{R}_{djk}\,\tilde{Y}_{djk} = \tilde{R}_{djk}\sum_{l\in\varsigma_{djk}} w_{djkl}\,y_{djkl}\,, \tag{28}$$

dove $\tilde{Y}_{djk}$ è la stima del totale e $\tilde{N}_{djk}$ 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* (Smith, 1991; Zhang, 2000). 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, allora, si possono cosí esprimere:

$$w_{PH;djk} = \sum_{l\in\varsigma_{djk}} \tilde{R}_{djk}\,w_{djkl}\,. \tag{29}$$

Nei piani di campionamento complessi, in generale, 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 del secondo membro della (26); sia per gli strati NAR, stimati con il secondo termine della (26). 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

La determinazione del peso dovrebbe conseguire gli obiettivi seguenti: (1) ottenere stime coerenti per famiglie e individui, attribuendo a ciascuna famiglia e a tutti i suoi componenti lo stesso peso finale; (2) correggere la distorsione per le mancate risposte; (3) produrre stime campionarie di totali di alcune importanti variabili ausiliarie coincidenti con i loro valori noti nella popolazione, $\wp$ (Falorsi, Falorsi, 1995). Tali obiettivi si possono conseguire con gli stimatori di ponderazione vincolata (*calibaration estimators*), che per il totale della $\mathcal{Y}$, ha una espressione analoga alla (21):

$$\hat{Y}_{PV} = \sum_{d=1}^{D}\sum_{c=1}^{c_d}\sum_{i=1}^{n_{dc}} d_{dci}\,\gamma_{dci}\,y_{dci} \equiv \sum_{k=1}^{n} d_k\,\gamma_k\,y_k = \sum_{k=1}^{n} w_k\,y_k\,, \tag{30}$$

dove con d_k si sono indicati i pesi iniziali, w_{dc} o w_{ijk}, detti *pesi diretti*, per potere indicare i *pesi finali* ancora con $w_k = d_k\,\gamma_k$; il fattore γ_k è il correttore dei pesi iniziali. Per semplificare le espressioni successive si usa un solo indice, *k*. I pesi iniziali dipendono dal piano di campionamento e dagli esiti della rilevazione, mentre i pesi finali dipendono dai totali noti delle L variabili ausiliarie, **X**, in $\wp$ e dai valori assunti dalle variabili ausiliarie del campione estratto. Per determinare i pesi finali occorre definire una funzione, *G*, che misura la distanza tra i pesi diretti d_k (noti) e i pesi finali w_k (incogniti). I pesi finali derivano dalla soluzione del minimo delle distanze

$$\min\left\{\sum_{k=1}^{n} G_k\left(w_k; d_k\right)\right\}, \tag{31}$$

soggetto al vincolo che i pesi finali soddisfacciano i totali noti delle variabili ausiliarie

$$\sum_{k=1}^{n} w_k \, \mathbf{x}_k = \mathbf{X} \, . \tag{32}$$

La funzione G deve soddisfare alcune condizioni di regolarità affinché il problema di minimo vincolato ammetta soluzioni (Deville, Särndal, 1992), che garantiscono l'esistenza di una funzione inversa, $g_k^{-1}(\cdot)$, con la quale si ottiene $w_k = g_k^{-1}\left[g_k(w_k; d_k)\right]$. Con il metodo dei moltiplicatori di Lagrange si ottiene il seguente sistema omogeneo:

$$\begin{cases} \dfrac{\partial L(\mathbf{w}, \boldsymbol{\lambda})}{\partial w_k} = g_k\left(w_k; d_k\right) - \mathbf{x}_k' \, \boldsymbol{\lambda} = 0 & \text{per} \quad k = 1, \ldots, n \\ \dfrac{\partial L(\mathbf{w}, \boldsymbol{\lambda})}{\partial \lambda_l} = \sum_{k=1}^{n} w_k \, x_{kl} - X_l = 0 & \text{per} \quad l = 1, \ldots, L \end{cases} \tag{33}$$

di $(n + L)$ equazioni nelle $(n + L)$ incognite $(\mathbf{w}, \boldsymbol{\lambda})$ in cui $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l, \ldots, \lambda_L)$ è il vettore dei moltiplicatori di Lagrange e $L(\mathbf{w}, \boldsymbol{\lambda})$ è la funzione di Lagrange. Dalle prime n equazioni si ottengono le soluzioni

$$w_k = g_k^{-1}\left(\mathbf{x}_k' \boldsymbol{\lambda}\right) = d_k \, \frac{1}{d_k} \, g_k^{-1}\left(\mathbf{x}_k' \boldsymbol{\lambda}\right) = d_k \, F_k\left(\mathbf{x}_k' \boldsymbol{\lambda}\right) = d_k \, \gamma_k \, , \tag{34}$$

dove la funzione $F_k\left(\mathbf{x}_k' \boldsymbol{\lambda}\right)$ corrisponde al correttore, γ_k, dei pesi di base, d_k. Si sostituisce, quindi, nelle ultime L equazioni della (33),

$$\sum_{k=1}^{n} d_k \, F_k\left(\mathbf{x}_k' \boldsymbol{\lambda}\right) \mathbf{x}_k = \mathbf{X} \, , \tag{35}$$

e si risolve il sistema risultante nel vettore delle L incognite $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l, \ldots, \lambda_L)$. Il vettore delle soluzioni, $\boldsymbol{\lambda}^*$, si sostituisce nella (34). Per ottenere il valore dei pesi finali occorre esplicitare e/o definire la funzione di distanza perché ve ne sono molte (Deville, Särndal, 1992; Singh, Mohl, 1996). La funzione di distanza piú comune è quella euclidea, $G(w_k; d_k) = \left(w_k - d_k\right)^2 / d_k$, dalla quale si ottiene (Falorsi, Rinaldelli, 1998):

$${}_{PV} w_k = d_k \left\{ 1 + \frac{1}{2} \, \mathbf{x}_k' \left(\sum_{k=1}^{n} d_k \, \mathbf{x}_k' \, \mathbf{x}_k \right)^{-1} \left(\mathbf{X} - \sum_{k=1}^{n} d_k \, \mathbf{x}_k \right) \right\} . \tag{36}$$

Solo ora i pesi finali sono stati indicati con ${}_{PV} w_k$ sia per non appesantire le formule precedenti e sia per distinguerli dagli altri tipi di pesi. Altre funzioni di distanza, come la logaritmica e la logaritmica troncata che sono utilizzate anche dall'Istat (Falorsi, Rinaldelli, 1998), generano stimatori con proprietà non note; tuttavia, gli stimatori di ponderazione vincolata convergono allo stimatore di regressione generalizzata ottenuto, quando si adotta una funzione di distanza euclidea (Deville, Särndal, 1992).

Si consideri lo stimatore di regressione, approssimato al primo termine con lo sviluppo in serie di Taylor,

$$\hat{Y}_{\text{Regr}} \cong \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \boldsymbol{\beta} = \mathbf{X}' \boldsymbol{\beta} + \sum_{k=1}^{n} d_k \, \gamma_k \, Z_k = \mathbf{X}' \boldsymbol{\beta} + \sum_{k=1}^{n} {}_{PV} w_k \, Z_k \, , \tag{37}$$

dove γ_k è il correttore dei pesi iniziali d_k ottenuto con la funzione di distanza euclidea, $\mathbf{X} = ({}_1\mathbf{X}_1, \ldots, {}_L\mathbf{X}_n)'$ è il vettore contenente i valori delle L variabili ausiliarie, $\boldsymbol{\beta}$ è il

vettore dei coefficienti di regressione del modello lineare che mette in relazione la variabile Y con le L variabili ausiliarie ${}_l X$ per $l = (1,\ldots,L)$, e $Z_k = Y_k - \mathbf{X}_\mathbf{k}^{'}\,\beta$.

Lo stimatore $\hat{V}\left(\hat{Y}_{\text{Regr}}\right)$ della varianza $V\left(\hat{Y}_{\text{Regr}}\right)$ non è, quindi, una funzione lineare dei dati campionarî, 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). L'espressione lineare dello stimatore di Z è data da

$$\hat{Z} = \sum_{d=1}^{D} \hat{Z}_d = \sum_{d=1}^{D}\sum_{c=1}^{c_d} \hat{Z}_{dc} = \sum_{d=1}^{D}\sum_{c=1}^{c_d}\sum_{i=1}^{n_{dc}} {}_{PV}\, w_{dci} Z_{dci}\,. \tag{38}$$

Lo stimatore della varianza dello stimatore del totale, $\hat{Y}$, in ciascun dominio territoriale può essere espresso come somma degli stimatori delle varianze dei singoli strati AR e NAR, appartenenti al dominio d, che sono differenti tra loro e risulterà:

$$\hat{V}\left(\hat{Y}\right) \cong \sum_{d=1}^{D} \hat{V}\left(\hat{Z}_d\right) = \sum_{d=1}^{D}\sum_{c=1}^{c_d^{AR}} \hat{V}\left(\hat{Z}_{dc}\right) + \sum_{d=1}^{D}\sum_{c=1}^{c_d^{NAR}} \hat{V}\left(\hat{Z}_{dc}\right), \tag{39}$$

dove c_d^{AR} e c_d^{NAR} indicano il numero di comuni AR e NAR, rispettivamente, nello strato d. Nei distretti di Sassuolo e Vignola i comuni sono tutti AR per l'espansione del campione e negli altri c'è un solo comune AR. Il disegno di campionamento adottato consente di ottenere stime corrette della varianza campionaria. Negli strati AR, l'espressione per il primo termine, del membro a destra del segno di uguaglianza nella (39), è data dalla (22), dove ${}_{AR} s^2_{2;\,dc}$ è riferita alla Z_{dc}. 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 il distretto di Carpi ha un solo comune campione, ma si può rimediare con la tecnica di *collassamento degli strati*, associando lo strato NAR di Carpi con quello di Mirandola; tuttavia, 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. In generale, dopo il collassamento, siano D^* il numero di distretti con comuni NAR e C_d^* il numero di comuni NAR nel distretto d, con $C_d^* \geq 2$. Allora, l'ultimo termine della (39) diventa

$$\sum_{d=1}^{D^*} \hat{V}\left(\hat{Z}_d\right) = \sum_{d=1}^{D^*} \frac{C_d^*}{C_d^* - 1} \sum_{c=1}^{C_d^*} \left(\hat{Z}_{dc} - \frac{\hat{Z}_d}{C_d^*} \right)^2, \tag{40}$$

dove $\hat{Z}_{dc} = \sum_{i=1}^{n_{dc}} {}_{PV}\, w_{dci}\, Z_{dci}$ è il totale dello strato NAR dc, $\hat{Z}_d = \sum_{c=1}^{C_d^*}\sum_{i=1}^{n_{dc}} {}_{PV}\, w_{dci}\, Z_{dci}$ è il totale del d-esimo strato NAR. Si può cosí determinare completamente $\hat{V}\left(\hat{Y}\right)$.

4. L'indagine longitudinale sul reddito delle famiglie

Il reddito è una grandezza che evolve nel tempo e per analizzarlo è necessario, quindi, introdurre la dimensione temporale nello schema di campionamento perché anche le caratteristiche e la composizione della popolazione di riferimento cambiano, sicché la comparabilità dei dati rilevati nelle varie fasi può essere compromessa (Kish, 1986).

La stima dei parametri della popolazione al variare del tempo, tramite campione, fu analizzata in principio da Jessen (1942), Yeats (1949), e Patterson (1950). In séguito, furono condotti diversi studi sugli schemi di rotazione da Hansen e altri (1955), Eckler (1955), Rao e Graham (1964), Gurney e Daly (1965). Gli obiettivi di stima delle indagini sui fenomeni nel tempo sono (Duncan, Kalton, 1987; Kalton, Citro, 1993):
(a) stima dei parametri della popolazione relativi a tempi distinti,
(b) stima dei valori medi nel tempo dei parametri della popolazione,
(c) stima della variazione netta (*net changes*),
(d) stima dei cambiamenti individuali (*gross changes*),
(e) cumulazione dei dati individuali nel tempo,
(f) cumulazione dei campioni nel tempo,
(g) osservazione di eventi che accadono in un periodo fissato.

I metodi di indagine dei fenomeni nel tempo sono principalmente l'indagine ripetuta, l'indagine longitudinale o *panel*, l'indagine rotata, e l'indagine *split panel*.

L'indagine ripetuta o periodica consiste nell'effettuare osservazioni o misure simili su campioni distinti estratti, in momenti diversi del tempo, da una popolazione equivalente che è univocamente definita. Ogni campione corrisponde a una indagine trasversale (*cross-section*), detta anche per contemporanei, che misura i caratteri delle unità statistiche alla data di riferimento, ossia tra loro contemporanee. L'indagine ripetuta è la piú semplice da realizzare e permette di conseguire gli obiettivi: (a), (b), (c), (f), (g); ma non gli obiettivi (d), (e).

L'indagine *panel*[1], proposta da Lazarsfeld e Fiske (1938), richiede la rilevazione delle stesse misure sullo stesso campione a momenti diversi del tempo, ovvero è un'indagine ripetuta nel tempo sulle stesse unità (soggetti, famiglie, imprese) selezionate al momento della costituzione iniziale del campione. Con l'indagine *panel* si conseguono gli obiettivi (c), (d), (e), (g), ma la sua realizzazione è assai problematica.

L'indagine rotata o *panel* rotante osserva gli stessi elementi solo per un periodo limitato: a ogni fase, una quota di elementi, $Q = 1 - P$, lascia il campione, e una quota di nuovi elementi, pari a Q, viene aggiunta. Si conseguono gli obiettivi: (a), (b), (c); mentre (d), (e), (f), (g) sono penalizzati dalla breve permanenza degli elementi nel campione. I piani di campionamento si distinguono per l'entità di P e per il tipo di rotazione tra i periodi: per P=1 si ha una indagine *panel*, mentre per P=0 si ha una indagine ripetuta.

---

[1] La terminologia non è consolidata: un sinonimo è *indagine longitudinale*, che è preferibile usarlo per una indagine trasversale (*cross-section survey*) nella quale i dati sono acquisiti retrospettivamente, come suggeriscono Duncan e Kalton (1987), ma per questo tipo di rilevazione sarebbe meglio usare il sintagma *indagine retrospettiva* (*retrospective survey*). Altri termini, come *indagine prospettica* (*prospective survey*) o *indagine multifase* (*multi-round survey*), sono ugualmente utilizzati, benché siano meno diffusi dei precedenti. In epidemiologia è privilegiato piú frequentemente il termine *follow-up*. Le indagini concernenti sottogruppi di popolazione che hanno sperimentato lo stesso evento nello stesso periodo di tempo, come il numero di coppie che hanno celebrato il matrimonio o il numero di laureati in un dato anno, sono chiamate *studi di coorti* o *analisi per generazioni*.

L'indagine *split panel* utilizza un campione suddiviso in due indagini distinte e "indipendenti": una *panel* e l'altra ripetuta o rotata (Kish, 1983, 1986); infatti, il termine *split* significa «diviso o spaccato o scisso». L'indagine *panel* consente di ottenere gli obiettivi (c), (d), (e), (g). L'indagine ripetuta permette di conseguire gli obiettivi (a), (b), (c), (f); inoltre, consente di accertare gli ingressi e le uscite che avvengono nella popolazione o di controllare la distorsione che può avvenire nell'indagine panel, data dal condizionamento, dalla perdita dei soggetti, e dal logoramento del campione. La sua realizzazione è, però, piú dispendiosa.

L'indagine retrospettiva o trasversale con domande retrospettive si caratterizza per la raccolta, su ogni unità, di dati riferiti sia al momento in cui si realizza la misura, sia a momenti precedenti. Si basa sul ricordo dei soggetti intervistati per ricostruire il flusso di informazioni nel tempo. I dati possono presentare notevoli errori di tipo non campionario perché gli intervistati dimenticano a volte un evento (errore di memoria) o la data dell'avvenimento o la sua entità (errore telescopico). A questi problemi, si deve aggiungere la distorsione per la perdita delle unità che sono uscite dalla popolazione prima della data dell'indagine (Sudman, Brandburn, 1973; Moss, Goldstein, 1979; Cannell, Miller, Oksenberg, 1981; Mathiowetz, Duncan, 1984). Malgrado questi limiti, si possono conseguire gli obiettivi (d), (e), (g), e, con difficoltà, (a), (b), (c), (f).

4.1. Confronto tra alcuni metodi di indagine

L'indagine retrospettiva è stata richiamata per completezza, ma è difficile utilizzarla nel caso della rilevazione del reddito delle famiglie per le sue caratteristiche e per la natura delle informazioni da rilevare. Le indagini *panel* e le indagini ripetute usano, invece, criteri antitetici per includere i soggetti nelle fasi successive; pertanto, le difficoltà e le facilitazioni, che ne conseguono dalla loro applicazione, costituiscono una guida nella scelta del metodo di indagine (Kish, 1965; Sudman, 1976; Sudman, Ferber, 1979).

I problemi e gli svantaggi che derivano dall'uso delle indagini longitudinali sono: (1) l'*auto-selezione* iniziale dei soggetti inclusi nel campione; (2) l'*attrito* o perdita di soggetti per stanchezza o noia o apatia o irreperibilità; (3) la *non riposta temporanea*, dovuta o alla non presenza a casa o al rifiuto di rispondere che può essere piú alta dell'attrito, 3-6% contro l'1-2%; (4) la *reattività alla reintervista*, perché l'esperienza di quella passata e l'anticipazione di un'altra da realizzare in futuro possono cambiare il comportamento, le attitudini, e le opinioni dell'intervistato — il fenomeno è denominato anche condizionamento, contaminazione, sensibilizzazione, apprendimento, e distorsione del *panel*; (5) la *disattenzione nella reintervista* sia da parte dell'intervistato, sia da parte dell'intervistatore per accidia, affaticamento, monotonia; (6) la *mobilità* dei soggetti che rende la loro reperibilità difficile; (7) i *cambiamenti* subíti nel tempo dalle unità che complicano l'elaborazione dei dati, come le separazioni, i divorzi, i matrimoni dei figli, eccetera; (8) la *mortalità e emigrazione* che riducono l'ampiezza del campione; (9) la *natalità e immigrazioni* che sono ignorate nelle indagini panel (Sobol, 1959), mentre le indagini ripetute le incorporano automaticamente; (10) le *verifiche e i controlli* necessari sia per evitare distorsioni e perdita dei soggetti nel campione, sia per ottenere una elevata qualità dei dati, in termini di completezza, consistenza, e coerenza; (11) i *limiti nell'analisi dei dati*, perché il condizionamento dei soggetti e l'onere continuo nel tempo comportano errori di misura e/o omissione di variabili; (12) la *scelta del periodo di reintervista*, perché la reattività dell'intervistato potrebbe indicare periodi ottimali che sono in conflitto con le esigenze

conoscitive del fenomeno. Tra le difficoltà citate sono state incluse, per completezza, anche quelle che sono tipiche di ogni indagine campionaria: (7), (8), (9), (10), (11), e (12). Cosí, per esempio, le verifiche e i controlli, (10), per ottenere dati accurati e non distorti o i limiti nell'analisi dei dati, (11), che emergono da errori di misura e/o omissioni di variabili, riguardano un po' tutte le indagini.

Il vantaggio principale delle indagini longitudinali rispetto alle indagini ripetute è (a) la *potenzialità di analisi*, perché i «dati panel» sono piú numerosi e piú variabili sicché diminuisce la collinearità tra i caratteri esplicativi e aumentano i gradi di libertà e l'efficienza delle stime; inoltre, i dati individuali consentono di studiare la natura del cambiamento e del comportamento. La varianza nelle osservazioni ha una componente inter-individuale, poi, che è dominante sul resto e assicura una maggiore robustezza a alcune stime (Dormont, 1989). Tra gli altri vantaggi si citano: (b) la *rimozione e riduzione dell'errore*, perché si possono introdurre procedure di controllo della coerenza e consistenza dei dati; (c) la *famigliarità* tra gli intervistati e gli intervistatori, perché spesso facilita il flusso di informazioni dai primi ai secondi in contrasto con l'attrito, la reattività, e la disattenzione; (d) l'*effetto organizzazione*, perché la necessità di disporre di una struttura che funzioni continuamente nel tempo, comporta uno sviluppo di conoscenze e risorse che contribuiscono a migliorare la realizzazione delle diverse fasi dell'indagine; (e) i *costi minori*, perché l'acquisizione dei dati di base, o che non mutano nel tempo, non si deve ripetere nelle fasi successive alla prima. La questione dei costi è, tuttavia, controversa perché la conoscenza dei soggetti facilita il contatto in termini di spesa e di tempo, ma la convenienza non sembra cosí scontata (Kish, 1986, 1989).

I vantaggi delle indagini longitudinali sono notevoli, nonostante le numerose difficoltà, perché consentono di approfondire alcuni aspetti dei fenomeni che, altrimenti resterebbero inesplorati (Dormont, 1989). Per esempio, solo con esse si può esaminare la natura o la struttura del cambiamento individuale o aggregare i dati degli individui nel tempo o costruire modelli di comportamento e stimare i relativi parametri; inoltre, ci si può limitare a raccogliere solo i dati che mutano nel tempo aumentando le informazioni disponibili e si possono controllare anche gli errori telescopici, che si verificano quando gli eventi vengono riferiti a date sbagliate.

L'indagine longitudinale è piú efficiente nella stima della variazione netta. Siano $\bar{y}_1$ e $\bar{y}_2$ le medie della variabile in esame ai tempi $t=1$ e $t=2$; sia $\Delta y = \bar{y}_2 - \bar{y}_1$ lo stimatore della variazione netta. Allora, la varianza della stima è data

$$V(\bar{y}_2 - \bar{y}_1) = V(\bar{y}_1) + V(\bar{y}_2) - 2\,\rho\,\sqrt{V(\bar{y}_1)\,V(\bar{y}_2)}$$

dove ρ è il coefficiente di correlazione tra le $\mathscr{Y}$ rilevate nei due tempi. In una indagine ripetuta i due campioni sono indipendenti e $\rho = 0$; in una indagine panel, invece, i valori individuali della $\mathscr{Y}$ dovrebbero essere positivamente correlati nel tempo e produrre stime piú precise della variazione netta. Si noti, poi, che la misura della variazione netta conseguita con un'indagine ripetuta nel tempo riflette sia il cambiamento dei valori della variabile, sia il cambiamento della struttura della popolazione; mentre con una indagine longitudinale si può tenere conto di entrambi i cambiamenti.

Gli svantaggi delle indagini longitudinali sono limitati con diversi accorgimenti. Per limitare la distorsione introdotta dal logoramento del campione e dalle risposte non date (*selectivity bias*) si interpolano i dati mancanti, si aggiustano i pesi (Kish, 1990), si modellano le mancate risposte (Little, Rubin, 1987). Per arginare la perdita di

rappresentatività del campione si applicano varianti dell'indagine *panel* che prevedono l'aggiunta di campioni di nuovi entranti in ciascuna fase: campioni per *coorti di nascite* che prevedono l'aggiunta di elementi selezionati dalle coorti di nascita man mano che passa il tempo; le *coorti multiple* che includono in ogni periodo un campione delle unità entranti. L'inefficienza statistica che ne deriva è compensata dalla maggiore quantità di informazione disponibile poiché delle nuove unità si conosce l'origine.

Nell'indagine rotata la permanenza di famiglie nel campione è limitata a un certo periodo di tempo; si perdono i vantaggi del panel, mentre si riduce la perdita di rappresentatività del campione perché una parte delle famiglie intervistate è rinnovata. Gli svantaggi potrebbero essere compensati da una rilevazione piú accurata, ma rimane faticoso per l'intervistato sicché la percentuale delle non risposte può diventare alta e il rischio di distorsione da partecipazione diventa sempre piú elevato. Una panoramica sui molteplici aspetti delle indagini longitudinali è riportata in Kasprzyk e altri (1989).

4.2. Stimatori e pesi per le indagini longitudinali

La scelta di un'indagine longitudinale per la rilevazione dei bilanci delle famiglie richiede ovviamente di definire la popolazione innanzitutto e quali unità statistiche rilevare (le famiglie o gli individui o le abitazioni), di stabilire con quale periodicità eseguire la rilevazione, biennale o annuale o semestrale, e di fissare per quante fasi si ripete l'intervista alle unità del campione. Questi aspetti devono essere definiti in anticipo perché connessi agli obiettivi e all'organizzazione dell'indagine. Cosí, per esempio, se si vuole verificare come muta il comportamento di risparmio, il contatto con le famiglie deve avvenire una volta ogni periodo di tempo dato; ma per la sua scelta non esistono criteri univoci e dipende sia dai fattori che determinano il fenomeno oggetto di studio, sia dalla tecnica di misurazione: domande retrospettive o compilazione di moduli lasciati presso l'intervistato.

Lo stimatore di minima varianza della media al tempo t è dato dalla seguente espressione (Fabbris, 1989):

$$\bar{y}_t = \alpha\, \bar{y}_t^T + (1-\alpha)\bar{y}_t^L + (1-\alpha)\rho\left(\bar{y}_{t-1} - \bar{y}_{t-1}^L\right), \tag{41}$$

dove $\bar{y}_t$ e $\bar{y}_{t-1}$ sono le medie di $\mathscr{Y}$ al tempo t e $(t-1)$, rispettivamente, $\bar{y}_t^L$ e $\bar{y}_{t-1}^L$ sono le medie di $\mathscr{Y}$ al tempo t e $(t-1)$ per la componente longitudinale, $\bar{y}_t^T$ è la media al tempo t per la componente trasversale, ρ è il coefficiente di correlazione tra $\bar{y}_t$ e $\bar{y}_{t-1}$, α è il coefficiente della combinazione lineare pari a

$$\alpha = \frac{Q\left(1-\rho^2 Q\right)}{1-\rho^2 Q^2}, \tag{42}$$

dove Q è la quota di famiglie trasversali o non *panel*. Lo stimatore $\bar{y}_t$ non è una media ponderata dei valori rilevati al tempo t e $(t-1)$ perché tiene conto sia del coefficiente di correlazione e sia della parte trasversale e longitudinale. Se si può ragionevolmente assumere che $\bar{y}_{t-1} \cong \bar{y}_{t-1}^L$, allora l'ultimo termine del secondo membro della (41) è trascurabile e lo stimatore della media presenta la seguente semplificazione:

$$\bar{y}_t^\circ = \tilde{\alpha}\, \bar{y}_t^T + (1-\tilde{\alpha})\bar{y}_t^L, \tag{43}$$

dove $\bar{y}_t^\circ$ è lo stimatore approssimato della $\bar{y}_t$ nella (41) e corrisponde alla media

ponderata della componente trasversale e longitudinale, $\tilde{\alpha}$ è il valore campionario di α espresso nella (42), dove ρ è stimato dai dati del campione.

Lo stimatore $\bar{y}_t^{\circ}$ assegna alla parte longitudinale del campione un peso relativo maggiore perché si basa sulla correlazione esistente tra le variabili rilevate sulle stesse famiglie in tempi successivi; pertanto, si riduce il peso della componente trasversale. La stima si può ottenere come media dei dati rilevati al tempo *t*, ponderata con pesi pari a:

$$\begin{cases} w_{dci}^{T} = w_{dci}\,\dfrac{\tilde{\alpha}}{Q} \\ w_{dci}^{L} = w_{dci}\,\dfrac{1-\tilde{\alpha}}{1-Q} \end{cases} \tag{44}$$

dove w_{dci}^{T} e w_{dci}^{L} sono, rispettivamente, i pesi da applicare alla componente trasversale e longitudinale. La determinazione dei pesi per le indagini longitudinale è un argomento ampio e complesso (Kalton, Brick, 1995; Lavallée, 1995; Rizzo, Kalton, Brick, 1996), che ora si omette per brevità.

5. Errori non campionarî

Le indagini che accertano il reddito, il patrimonio, il risparmio, e gli investimenti risultano sempre complicate e non bastano gli accorgimenti a migliorare la rilevazione (Quintano, Lucev, 1990), ma occorrono intervistatori capaci sia per la qualità dei dati raccolti, sia per ottenere la partecipazione delle unità statistiche (Baldini *et al*., 2004; Couper, Groves, 1992; Hox, de Leeuw, 2002). Anche se il processo di raccolta dei dati può essere migliorato, i vincoli temporali, logistici, e di risorse umane e finanziarie pongono limiti decisivi. L'esperienza maturata nelle indagini condotte dalla Banca d'Italia (2004) mostra 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.

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; (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. Gli ultimi due punti riguardano il processo di raccolta dei dati, che si articola in varie fasi e coinvolge molteplici persone, come gli intervistatori e gli intervistati, che non sono sempre controllabili. Si possono generare, quindi, degli errori, detti *non campionarî*, che bisogna cercare di ridurre con tutti i mezzi disponibili perché possono diventare anche preponderanti, rispetto agli errori campionarî. Non esiste ancora una teoria completa degli errori non campionarî; pertanto, ogni indagine è un caso a sé e presenta un proprio *profilo dell'errore*. L'individuazione di tali errori richiede 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).

Gli errori non campionarî sono classificati in tre categorie (Lessler, Kalsbeek, 1992): (*i*) errori di lista o errori di copertura, (*ii*) errori da mancata risposta, (*iii*) errori

di misurazione, generati da numerosi fattori che alterano il valore osservato introducendo una differenza con il valore reale.

Gli errori della lista (*i*) sono i peggiori 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. Nell'impossibilità di usare la banca dati fiscale e tributaria dei contribuenti, la lista anagrafica è un buon compromesso per gli obiettivi dell'indagine.

La riduzione degli errori da mancata risposta (*ii*) è il compito primario da perseguire in una indagine perché migliora la *qualità dell'indagine in sé*, riducendo tutte le difficoltà menzionate in precedenza. La mancata risposta può derivare sia dalla 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*). Sul campo si è accumulata molta esperienza, ma non è sempre possibile applicarla a causa dei costi. Le persone esperte in interviste sono rare e il loro costo è elevato. Si ricorre, pertanto, a intervistatori che, seppure addestrati, sono spesso alla loro prima esperienza e ciò può non bastare a migliorare il tasso di partecipazione (Groves, 1989).

Gli errori di misurazione (*iii*) si sovrappongono, in parte, a quelli da mancata risposta perché l'assenza di una risposta potrebbe dipendere proprio da una formulazione ambigua o inadeguata. Una rilevazione accurata migliora la *qualità dei dati* e, quindi, la precisione dei risultati che diventano piú affidabili e fedeli alla realtà: rappresenta, perciò, un obiettivo essenziale (Liepins, Uppuluri, 1990). Gli strumenti che si usano nell'indagine possono costituire una fonte di errore e l'esperienza può aiutare a progettare strategie efficienti, ma nelle realtà complesse, le difficoltà non sono eludibili. Una distinzione tipica degli errori di misurazione è basata sulla causa che li ha prodotti: (1) errori di *strumenti*, riconducibili al questionario per domande formulate in modo ambiguo o disposte in un ordine inadeguato, per batterie di test non tarati bene, e cosí via; (2) errori di *tecniche*, dipendenti dal tipo di procedura o tecnica utilizzata, come il questionario postale, l'intervista auto-somministrata, l'intervista telefonica, la batteria di test; (3) errori dell'*intervistatore*, derivanti dall'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 o di ricordare gli eventi accaduti, alla sua idoneità e volontà di fornire risposte veritiere.

L'esperienza ci fornisce già l'ordine di grandezza delle mancate partecipazioni, come già indicato in precedenza (§2.3). Il tasso di rifiuto si può stimare preventivamente intorno al 60%, in base all'affidabilità della lista, alle esperienze condotte in precedenza, e anche alla letteratura esistente (Goyder, 1987; 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, 2004); ma nella indagine condotta nel 2002 si sono riscontrati inconvenienti anche con un capofamiglia che aveva uno stato civile libero (*single*), con gli anziani perché non aprono facilmente agli sconosciuti, con le dimensioni dei comuni (Baldini *et al*., 2004).

Le relazioni sono un po' diverse da quelle riscontrate dalla Banca d'Italia, data la differente scala delle due indagini: maggiori ostacoli si incontrano con comuni piccoli e/o in montagna, con famiglie aventi un ridotto numero di componenti, con un capofamiglia pensionato.

Forme ulteriori di errori non campionarî 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, bensí il personale addetto alla rilevazione e immagazzinamento dei dati.

Il trattamento degli errori non campionarî 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

La raccolta dei dati può essere sintetizzata con alcuni indicatori, in genere, definiti dal rapporto di due quantità, numeratore e denominatore. In base alla specificazione delle due quantità, si evidenziano aspetti differenti del processo di intervista. Tali rapporti sono denominati «tassi di completamento» (*completion rate*) perché riguardano il successo delle interviste. Per semplicità, si usa il termine «tasso», nonostante esso indichi, in genere, rapporti «unitari»; per brevità, si ometterà di specificare «percentuale». Gli indicatori piú frequenti sono definiti nella Tabella 9.

Il Tasso di Efficienza dell'Intervistatore (TEI) esprime la percentuale di volte che un intervistatore ottiene le interviste agli indirizzi campionarî contattati. Un indicatore diverso, e un po' piú preciso di TEI, è il Tasso di Interviste Completate (TIC) sul numero di unità campionarie eleggibili, che 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 differenti che colgono aspetti diversi del processo. Le combinazioni sono diverse e tante, ma nella Tabella 9 ci si è limitati soltanto al Tasso di Efficienza degli Intervistatori nei Contatti (TEIC) avuti con gli intervistati, alla Propensione degli elementi della Popolazione a Partecipare all'Indagine (PPPI) che è anche interpretabile come l'efficienza o abilità degli intervistatori a ottenere la collaborazione degli intervistati, al Tasso di Unità statistiche Rilevate (TUR), al Tasso di Unità statistiche Utili (TUU) ai fini della stima dei parametri di $\wp$. Al loro denominatore, «eleggibile» deve essere inteso come il numero di unità appartenenti alla popolazione, viceversa per «ineleggibile».

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 di Successo nei Contatti Telefonici (TSCT). 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 di Non Risposta (TNR). Si possono determinare diversi tassi variando il denominatore, ossia, distinguendo tra tutti i potenziali rispondenti o tra tutti i rispondenti con numero di telefono.

Tabella 9 – Tassi di completamento delle interviste

Denominazione	**Definizione**
Tasso di **E**fficienza dell'**I**ntervistatore	$TEI = 100 \dfrac{\text{Numero di interviste}}{\text{Numero di contatti}}$
Tasso di **I**nterviste **C**ompletate	$TIC = 100 \dfrac{\text{Numero di interviste completate}}{\text{Numero di unità campionarie eleggibili}}$
Tasso di **E**fficienza dell'**I**ntervistatore nei **C**ontatti	$TEIC = 100 \dfrac{\text{Numero di Interviste Completate}}{\text{Numero di Contatti}}$
Propensione **P**opolazione a **P**artecipare all'**I**ndagine	$PPPI = 100 \dfrac{\text{Numero di Rispondenti a tutte le domande}}{\text{Numero di Rispondenti Cominciato Intervista}}$
Tasso di **U**nità (statistiche) **R**ilevate	$TUR = 100 \dfrac{\text{Numero di Interviste Completate}}{\text{Numero di Unità nel Campione (Eleggibili + Ineleggibili)}}$
Tasso di **U**nità (statistiche) **U**tili	$TUU = 100 \dfrac{\text{Numero di Interviste Completate Eleggibili}}{\text{Numero di Unità nel Campione (Eleggibili + Ineleggibili)}}$
Tasso di **S**uccesso nei **C**ontatti **T**elefonici	$TSCT = 100 \dfrac{\text{Numero di Successi nei Contatti Telefonici}}{\text{Numero Totale di Tentativi di Contatti Telefonici}}$
Tasso di **N**on **R**isposta	$TNR = 100 \dfrac{\text{Numero di Rifiuti}}{\text{Numero di Interviste + Numero di Rifiuti}}$
Tasso di **N**on **R**isposta 1	$TNR_1 = 100 \dfrac{\text{Numero di Rifiuti}}{\text{Numero di Potenziali Rispondenti}}$
Tasso di **N**on **R**isposta 2	$TNR_2 = 100 \dfrac{\text{Numero di Rifiuti}}{\text{Numero di Potenziali Rispondenti con Numero di Telefono}}$

6. Conclusioni

La proposta di campionamento attuale ha migliorato la stratificazione territoriale, rispetto all'indagine condotta nel 2002. L'interesse degli amministratori dei distretti di Sassuolo e Vignola, ai risultati prodotti dall'indagine, ha comportato un aumento della dimensione campionaria di distretto e l'inclusione di tutti i comuni del distretto nel

campione, che sono diventati autorappresentativi. Si ottengono, cosí, sia beneficî per le stime e sia una semplificazione nel calcolo della varianza delle stime per distretto.

Nel comune di Modena si è introdotta una componente longitudinale che offre la possibilità di valutare e modellare le variazioni dei redditi individuali e i comportamenti di risparmio, di migliorare l'efficienza della stima delle variazioni nette e di apprezzare meglio l'evoluzione temporale del fenomeno.

Gli aspetti piú critici derivano dalle difficoltà di realizzazione che sono già note e riscontrate da ogni esperienza sul campo: presumibilmente, non si riusciranno a rilevare tutte le unità statistiche programmate nei diversi strati, ma nei comuni piccoli, in particolare, bisogna effettuare tutti gli sforzi possibili per rilevare il numero di unità programmate e, analogamente, nel comune di Modena si devono moltiplicare gli sforzi per intervistare tutte le famiglie appartenenti all'ultima classe di età del capofamiglia e all'ultima classe di dimensione della famiglia. Il limite al miglioramento dell'entità e della qualità dei dati rilevati è costituito dai costi, che possono crescere tanto da bloccare il processo di ricerca della collaborazione delle famiglie «resistenti».

La sostituzione delle unità che rifiutano di partecipare non è, in generale, una buona 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, 2004), con lo stato civile indipendente e giovane (*single*), con gli anziani perché non aprono facilmente a sconosciuti, con la dimensione o «natura» dei comuni. Se non si fa ricorso alla sostituzione delle mancate partecipazioni, però, la dimensione campionaria diventa insufficiente per gli scopi dell'indagine. Si nota, poi, che la sostituzione di un comune nel suo complesso è di per sé ancora piú rilevante, ma potrebbe accadere di non riuscire a ottenere la collaborazione delle amministrazioni comunali e/o dei responsabili degli uffici dell'anagrafe. Altre strategie, come il campionamento a grappoli e/o per aree a livello comunale, non sono praticabili perché si perderebbe, poi, l'omogeneità con gli altri dati.

Si rileva, infine, che 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 ordinarie, 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 il piano di campionamento 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, invece, non si può ignorare il piano di campionamento, specie se si considerano le mancate partecipazioni che si osserveranno e che non si distribuiranno uniformemente tra gli strati.

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.

Baldi P., Lemmi A., Sciclone N. (a cura di) (2005). *Ricchezza e povertà. Condizioni di vita e politiche pubbliche in Toscana*, Franco Angeli, Milano.

Baldini M., Bosi P., Silvestri P. (a cura di) (2004). *La ricchezza dell'equità*¸ il Mulino, Bologna.

Baldini M., Bigarelli D., Colombini S., Fregni C., Silvestri P. (2004). Nota metodologica sull'indagine, in Baldini M., Bosi P., Silvestri P. (a cura di), *La ricchezza dell'equità*¸ il Mulino, Bologna, pp. 309–321.

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.

Banca d'Italia (2004). *I bilanci delle famiglie italiane nell'anno 2002*, a cura di D'Alessio G., Faiella I., Supplementi al bollettino statistico (nuova serie), **anno XIV**, n. 12, 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), *Strategie e modelli per il controllo della qualità dei dati*, Franco Angeli, Milano, pp. 135–161.

Barisione M., Mannheimer R. (1999). *I sondaggi*, il Mulino, Bologna.

Benassi D. (a cura di) (2005). *La povertà come condizione e come percezione*, 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.

Betti G., Lemmi A., Maltinti G., Sciclone N. (2003). *Indagine sulle condizioni di vita delle famiglie toscane. Primi risultati*, Irpet/Cridire, Firenze.

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. (2005). La disuguaglianza di reddito in Italia nell'ultimo decennio, *Stato e mercato*, **74**, n. 2, pp. 207–229.

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.

Calandi S. (2003). Il campionamento: analisi del concetto di rappresentatività, *Sociologia e ricerca sociale*, **70**, pp. 70–95.

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.

Cannell C. F., Miller P. V., Oksenberg L. (1981). Research on interviewing techniques, in Leinhardt S. (ed.), *Sociological Methodology*, Jossey-Bass, San Francisco, pp. 389–437.
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.
Dormont B. (1989). Petite apologie des données de panel, *Economie et Prevision*, **87**, pp. 19–32.
Duncan G. J., Kalton G. (1987). Issue of design and analysis of surveys across time, *International Statistic Review*, **55**, pp. 97–117.
Eckler A. R. (1955). Rotation sampling, *Annals of Mathematical Statistics*, **26**, pp. 664–685.
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.
Falorsi S., Rinaldelli C. (1998). Un software generalizzato per il calcolo delle stime e degli errori di campionamento, *Statistica Applicata*, **10** (2), pp. 217–233.
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.
Gurney M., Daly J. F. (1965). A multivariate approach to estimation in periodic sample survey, in *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 242–257.
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.
Hansen M. H., Hurwitz W. N., Nisselson H., Steinberg J. (1955). The redesign of the census current population survey, *Journal of the American Statistical Association*, **50**, pp. 701–719.
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., *Survey Nonresponse*, Wiley & Sons, New York, pp. 103–120.
ISTAT (2002). *Panel europeo sulle famiglie*, a cura di Pauselli C., Metodi e Norme, nuova serie, n. 15, Roma.
ISTAT (2004). *I consumi delle famiglie. Anno 2002*, a cura di Barcherini S., Marrone P., Annuario, n. 9, Istat, Roma.
Jessen R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts, *Iowa Agricultural Experimental Research Bulletin*, **304**, pp. 54–59.
Kalton G., Citro C. F. (1993). Panel Surveys: Adding the Fourth Dimension, *Survey Methodology*, **19**, pp. 205–215.
Kalton G., Brick M. (1995). Weighting Schemes for Household Panel survey, *Survey Methodology*, **21**, pp. 33–44.
Kasprzyk D., Duncan G. J., Kalton G., Singh M. P. (1989). *Panel Surveys*, John Wiley & Sons, New York.
Kish L. (1965). *Survey Sampling*, John Wiley & Sons, New York.
Kish L. (1983). Data collection for details over space and time, in Wright T. (ed.), *Statistical Methods and the Improvement of Data Quality*, Academic Press, New York, pp. 73–84.
Kish L. (1986). Timing of surveys for public policy, *The Australian Journal of Statistics*, **28**, pp. 1–12.
Kish L. (1989). *Sampling Methods for Agricultural Surveys*, FAO Statistical Development Series, N. 3, Roma.
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.
Lalla M. (2003). Il disegno dell'indagine sulle condizioni economiche e sociali delle famiglie nella Provincia di Modena, *Materiali di discussione*, N. **431**, Dipartimento di Economia Politica, Università di Modena e Reggio Emilia, pp. 1–45.
Lavallée P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, *Survey Methodology*, **21**, pp. 25–32.
Lazarsfeld P. F., Fiske M. (1938). The panel as a new tool for measuring opinion, *Public Opinion Quarterly*, **2**, pp. 596–612.
Lessler J. T., Kalsbeek W. D. (1992). *Nonsampling Errors in Surveys*, Wiley & Sons, New York.
Levy P. S., Lemeshow S. (1991). *Sampling of Populations: Methods and Applications*, Jhon 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. (1993). Post-Stratification: A Modeler's Perspective, *Journal of the American Statistical Association*, **88**, pp. 1001–1012.
Little R. J. A., Rubin D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
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.
Mathiowetz N., Duncan G. (1984). Temporal patterns of response errors in retrospective reports of unemployment and occupation, in *Proceedings of the Survey Research Methodology Section*, American Statistical Association, pp. 652–657.

Moss L., Goldstein H. (1979). *The Recall Method in Social Surveys*, Institute of Education, University of London, London.

Palamenghi M., Riva L., Trentini M. (2005). *Criteri e metodi di stima del reddito delle famiglie bresciane*, Rapporti di ricerca del Dipartimento di Metodi Quantitativi, Quaderno n. 247, Università degli Studi di Brescia.

Patterson H. D. (1950). Sampling on successive occasions with partial replacement of units, *Journal of the Royal Statistical Society*, **B**, **12**, pp. 241–255.

Plaseller C., Vogliotti S., Zeppa A. (2005). Situazione reddituale e patrimoniale delle famiglia in provincia di Bolzano – 2003-2004, Provincia Autonoma di Bolzano-Alto Adige e Istituto Provinciale di Statistica – ASTAT, n. **117**, Bolzano.

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.

Rao J. N. K., Graham J. E. (1964). Rotation designs for sampling on repeated occasions, *Journal of the American Statistical Association*, **59**, pp. 492–509.

Rizzo L., Kalton G., Brick M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse, *Survey Methodology*, **22**, pp. 43–53.

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.

Sing A. C., Mohl C. A. (1996). Understanding Calibration Estimators in Survey Sampling, *Survey Methodology*¸ **22** (2), pp. 107–115.

Smith T. M. F. (1991). Post-Stratification, *The Statistician*, **40**, pp. 315–323.

Sobol M. G. (1959). Panel mortality and panel bias, *Journal of the American Statistical Association*, **54**, pp. 52–68.

Sudman S. (1976). *Applied Sampling*, Academic Press, New York.

Sudman S., Brandburn N. M. (1973). Effects of time and memory factors on response in surveys, *Journal of the American Statistical Association*, **68**, pp. 805–815.

Sudman S., Ferber R. (1979). *Consumer Panels*, American Marketing Association, Chicago.

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.

Yates F. (1949). *Sampling Methods for Census and Surveys*, Charles Griffin and Co. Ltd., London.

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

Elenco dei simboli piú frequenti

D Numero di distretti di area 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^{AR} Numero di comuni AR inclusi nel campione del distretto d; in genere, $c_d^{AR} = 1$.

C_d^{NAR} Numero di comuni NAR nel distretto d; in genere, risulta $C_d^{NAR} = C_d - 1$.

c_d^{NAR} Numero di comuni NAR inclusi nel campione del distretto d, vale $c_d^{NAR} = c_d - 1$.

N_d Numero di USS (famiglie) nel distretto d.

N_d^{AR} Numero di USS (famiglie) nel distretto d, strato di comuni AR.

N_d^{NAR} Numero di USS (famiglie) nel distretto d, strato di comuni NAR.

n_d^{AR} Numero di USS (famiglie) nel distretto d, strato di comuni AR, nel campione.

n_d^{NAR} Numero di USS (famiglie) nel distretto d, strato di comuni NAR, nel campione.

N_{dc} Numero di USS (famiglie) nel distretto d e nel comune c.

n_{dc} Numero di USS (famiglie) nel distretto d e nel comune c del campione.

w_{dc} Peso delle USS (famiglie) nel distretto d e nel comune c del campione.

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

w_{ijk} Peso delle USS (famiglie) nel comune di Modena.

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

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

$\mathscr{Y}$ Variabile casuale quantitativa o carattere, come il reddito e il risparmio.

Y Totale in $\wp$, se è un parametro oggetto di stima.

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

$_{AR}s_{2;dc}^2$ varianza campionaria nel distretto d e comune c che è AR o di secondo stadio.

$_{NAR}s_{2;dc}^2$ varianza campionaria nel distretto d e comune c che è NAR 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».

$f_{2;dc}^{AR}$ $f_{2;dc}^{AR} = n_{dc}^{AR} / N_{dc}^{AR}$, frazione di campionamento di secondo stadio nel comune AR.

$f_{2;dc}^{NAR}$ $f_{2;dc}^{NAR} = n_{dc}^{NAR} / N_{dc}^{NAR}$, frazione di campionamento di secondo stadio nel comune NAR.

$1_{[\bullet]}[\bullet]$ 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.

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

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