I lavori raccolti in questa collana hanno avuto origine nell'ambito del progetto di ricerca dell'ISTAT «Le statistiche dei consumi privati nel sistema statistico nazionale» e del progetto di ricerca MURST 40% «La misura dei consumi privati: uno studio sull'accuratezza, coerenza e qualità dei dati». Al progetto di ricerca hanno partecipato i ricercatori dell'ISTAT e dei seguenti Dipartimenti e Istituti universitari:

- Dipartimento di Scienze Statistiche, Bologna
- Dipartimento Statistico, Firenze
- Dipartimento di Scienze Statistiche, Perugia
- Dipartimento di Contabilità Nazionale, Roma
- Istituto di Statistica e Matematica, Istituto Universitario Navale, Napoli
- Istituto di Statistica, Messina.

IL TRATTAMENTO DELLA MANCATA RISPOSTA TOTALE MEDIANTE LA SOSTITUZIONE NELL'INDAGINE SUI CONSUMI DELLE FAMIGLIE

Silvia Pacei*

Rapporto di ricerca n. 7

CON PRI - La misura dei consumi privati

* Dipartimento di Scienze Statistiche "Paolo Fortunati", Università di Bologna

Dipartimento di Scienze Statistiche "Paolo Fortunati"
dell'Università degli Studi di Bologna
Gennaio 1995

Stampa: TECNOPRINT S.N.C., *Via del legatore 3, Bologna.*

INDICE

1. Introduzione

L'obiettivo principale di questo lavoro è stato quello di analizzare gli effetti della sostituzione delle famiglie partecipanti all'indagine ISTAT relativa ai consumi (BF) sulle stime ottenute. A tal proposito si è proceduto confrontando i risultati derivanti dalla sostituzione con quelli conseguiti ricorrendo sia all'utilizzo dell'insieme incompleto dei dati, sia ad una soluzione di trattamento *a posteriori* della non risposta.

Nell'ambito delle indagini campionarie, i provvedimenti volti a risolvere le problematiche indotte da eventuali defezioni delle unità svolgono un ruolo determinante nell'assicurare la rappresentatività del campione e la correttezza delle stime. E' noto, infatti, che le non risposte totali possono rivelarsi fonti di distorsione per le stime desumibili dalla rilevazione. Ciò è tanto più rilevante quanto più le caratteristiche di coloro che non collaborano differiscono da quelle di quanti, invece, partecipano all'indagine (Chapman, 1982).

Lo studio è stato compiuto sui dati rilevati dall'indagine BF negli ultimi sei mesi del 1990, periodo in cui un successivo sondaggio postale, condotto sulle famiglie non rispondenti, ha consentito di ottenere informazioni aggiuntive su di esse. Tale sondaggio è stato organizzato nell'ambito del progetto di ricerca CON.PRI., al fine di acquisire maggiori conoscenze sull'atteggiamento delle famiglie verso la rilevazione (De Simoni, Filippucci e Marliani, 1992). Le informazioni così ottenute, usualmente non disponibili per lo studio degli effetti delle non risposte, ci hanno permesso, da una parte, di valutare l'esistenza o meno di una effettiva differenza fra famiglie intervistate e non partecipanti all'indagine, dall'altra, di attuare un procedimento alternativo di stima basato sull'imputazione.

Il confronto fra i metodi sopra indicati è stato effettuato con riferimento alla variabile spesa familiare mensile, considerata nel suo complesso e suddivisa nelle sue principali categorie. In particolare, l'imputazione è stata condotta in modo diverso per beni non durevoli e per beni durevoli, in quanto la distribuzione delle spese per questi ultimi è caratterizzata dalla presenza di valori nulli.

Il lavoro risulta così articolato. La parte iniziale comprende una breve rassegna delle tematiche d'analisi affrontate nelle varie fasi della ricerca, come gli effetti della non risposta ed i metodi per risolvere il problema (par. 2). Vengono introdotti inoltre alcuni riferimenti alle caratteristiche salienti

dell'indagine sui consumi, con particolare attenzione alle modalità con cui hanno luogo la sostituzione delle famiglie ed il procedimento di stima delle spese (par. 3). Nella seconda parte, dopo una breve descrizione delle informazioni disponibili (par. 4.1), si passa all'analisi dell'utilizzo dei dati relativi ai soli intervistati (par. 4.2), e quindi alla illustrazione dei procedimenti di imputazione sperimentati (par. 4.3 e 4.4), riportando i risultati ottenuti nei vari casi. Infine, nel paragrafo 5 vengono presentati gli esiti dei confronti fra i differenti metodi utilizzati, dai quali emerge che la non risposta produce una sottovalutazione delle spese, e che il trattamento della non risposta tramite la sostituzione delle unità mancanti non rappresenta una soluzione efficace.

2. Il problema della mancata risposta totale

La non risposta totale è riconosciuta come uno dei principali problemi nelle indagini campionarie, ma l'estensione e gli effetti di tale fenomeno possono variare considerevolmente a seconda delle caratteristiche della rilevazione. L'effetto sulle stime che si ritiene maggiormente preoccupante è la distorsione sistematica, solitamente introdotta dall'impossibilità o incapacità di ottenere informazioni relative a particolari unità campionate (Sarndal, Swensson e Wretmann, 1992).

La distorsione aumenta, infatti, come evidenziato anche da Cochran (1977), al crescere della proporzione dei non rispondenti nella popolazione e della differenza fra la variabile di interesse nello strato dei rispondenti ed in quello dei non intervistati. Il "tasso di non risposta", dato dal rapporto fra il numero delle non risposte totali e quello delle unità campionate, può rappresentare una misura dell'incidenza della mancata risposta totale nella popolazione considerata, oltre ad un primo indicatore della qualità del lavoro di rilevazione svolto (raccolta dei dati, solleciti, ecc...). Per questo molti autori hanno sottolineato l'importanza di una sua riduzione (Stopher e Sheskin, 1981).

D'altra parte, il rilievo del trattamento delle non risposte totali dipende anche dal fatto che, qualora la non risposta sia non trascurabile, non esistono metodi ottimali e sempre validi per fare inferenza. Le molteplici proposte avanzate in letteratura per trattare il problema, si distinguono principalmente in procedimenti che si concentrano sull'analisi della

distorsione indotta, ed in metodi per prevenire le cadute campionarie e ridurne le conseguenze.

Il concentrarsi delle cadute campionarie in corrispondenza di particolari sottogruppi della popolazione, può produrre una distorsione nelle stime dovuta alla discrepanza fra campione realizzato e campione programmato (Lessler e Kalsbeek, 1992). In questo caso l'enfasi è posta sulla perdita di rappresentatività nel campione dovuta alle cadute campionarie. Di conseguenza, l'individuazione di una distorsione da esse indotta si effettua attraverso confronti fra rispondenti e non rispondenti, eseguiti rispetto a variabili di tipo demografico, sociale, economico e culturale ("studi di identificazione"). L'ipotesi è che differenze fra i due gruppi rispetto a queste variabili siano indicative di differenze significative anche fra le variabili di interesse stesse (Donald, 1960; Parten, 1966). Tuttavia, il semplice confronto può fornire solo qualche prima indicazione sulla potenziale entità e sulla direzione del contributo delle cadute campionarie alla distorsione, poiché difficilmente è noto il livello di correlazione fra le variabili ausiliarie e quelle di interesse.

Sull'altro versante, numerosi espedienti sono stati suggeriti in letteratura per incrementare i tassi di risposta e, soprattutto, per ridurre le conseguenze delle cadute campionarie. Chapman (1982) sostiene che sia possibile tentare di ridurre ad un livello minimo la distorsione intervenendo in più momenti dell'indagine: nella fase di rilevazione sul campo, includendo procedure di sollecito volte a fornire un basso tasso di non risposta totale; ad indagine completata, utilizzando uno dei numerosi metodi di imputazione e sostituzione per le unità mancanti suggeriti in letteratura (Rubin, 1976; Bailar, Bailey e Corby, 1978; Platek, Singh e Tremblay, 1978; Little e Rubin, 1987).

La sostituzione rappresenta un metodo per trattare la non risposta totale nella fase di lavoro sul campo, che si considera utile ai fini dell'indagine in particolar modo quando si ritiene che gli intervistati non siano in grado di rappresentare tutte le tipologie dei non partecipanti all'indagine. Quindi, selezionando altre unità per sostituirli si possono acquisire valori migliori di quelli ottenibili tramite procedimenti di imputazione alternativi. Benché tale tecnica sia spesso proposta come soluzione e offra la possibilità di aggiustare a proprio piacere la dimensione campionaria, non è detto che riduca la distorsione indotta dalle non risposte totali. Infatti, quando la sostituzione è realizzata secondo una scelta casuale e le unità sostitute sono

soggette allo stesso criterio di selezione del campione iniziale, è probabile che la distorsione aumenti, giacché i non intervistati verrebbero sostituiti con unità che assomigliano a quelle già presenti nel campione, mentre alcune sezioni della popolazione potrebbero rimanere non adeguatamente rappresentate (Kish, 1965). Per questo motivo i valori sostitutivi devono essere visti come un particolare tipo di valori imputati e non essere trattati come provenienti da intervistati. Se, invece, ad ogni unità che non collabora si abbina una unità sostitutiva sulla base di variabili ausiliarie note e correlate a quella da rilevare, cioè la scelta non avviene a caso, l'effetto distorsivo della non risposta tende a diminuire.

Condizione necessaria per la realizzazione di una tale scelta ragionata è pertanto la disponibilità di informazioni sulle unità della popolazione (Marbach, 1964). Inoltre, nei casi in cui la sostituzione porti ad un considerevole incremento dei costi della rilevazione, occorre valutare se i suoi effetti sulle stime (riduzione della distorsione e riduzione della varianza dovuta all'aumento della dimensione campionaria) sono tali da renderla tutto sommato conveniente.

Probabilmente è solo attraverso ricerche empiriche che l'effetto delle tecniche di sostituzione può essere valutato. Poche indagini sono state condotte in passato per valutare le conseguenze delle sostituzioni sulle stime (vedi ad esempio Durbin e Stuart, 1954; Williams e Folsom, 1977; Chapman e Roman, 1985). Benché nessuno di questi studi fosse effettuato in condizioni ideali, tutti sembrano indicare che i metodi di sostituzione non eliminano completamente gli effetti della distorsione indotta dalle non risposte totali. Ma questa critica, come sostenuto anche da Chapman (1982), può essere diretta anche ai metodi alternativi di trattamento della non risposta.

Passando ai criteri di compensazione applicabili una volta raccolti i dati, questi sono molteplici e si possono distinguere sulla base delle assunzioni che implicitamente ammettono sulle differenze fra rispondenti e non rispondenti, ovvero sul meccanismo che governa i dati mancanti. Secondo la definizione di Little e Rubin (1987), il meccanismo di non risposta si può considerare "ignorabile" quando la probabilità di rispondere è indipendente sia dalla variabile soggetta a non risposta, Y, sia dalle altre variabili note per ogni unità, i cui valori immaginiamo siano contenuti nella matrice X. Tale situazione è tuttavia piuttosto improbabile, poiché i valori osservati devono potersi considerare alla stregua di un campione casuale di

quelli campionati. Se invece la probabilità di risposta dipende solo dalle X e non dalla Y, il meccanismo si considera ignorabile entro le sottoclassi definite dalle variabili ausiliarie. Infine, lo si ritiene "non ignorabile", se questa probabilità dipende dalla variabile di interesse Y.

I metodi più frequentemente proposti in letteratura sono così raggruppabili: a) il sottocampionamento dei non rispondenti; b) l'aggiustamento dei pesi; c) l'imputazione esplicita; d) l'inferenza basata sul modello.

L'idea di estrarre un sottocampione dei non intervistati al primo tentativo fu suggerita da Hansen e Hurwitz (1946) e prevede che le informazioni così ottenute siano utilizzate congiuntamente a quelle relative agli intervistati per ottenere le stime finali. Ovviamente, uno dei presupposti del metodo è che tutti i non intervistati campionati forniscano delle risposte. La proposta iniziale, che si inseriva nell'ambito dell'approccio classico all'inferenza statistica, ha dato luogo successivamente ad estensioni che si avvalgono di metodi di stima basati sul modello (Erickson, 1967).

Le procedure di aggiustamento dei pesi, prevedendo una modificazione dei pesi da attribuire a quanti hanno partecipato all'indagine in modo da tenere conto delle cadute campionarie, possono essere visti come una forma implicita di imputazione, poiché si risolvono nell'attribuire a questi ultimi i dati ottenuti dagli intervistati. Le unità che hanno partecipato all'indagine vengono sostanzialmente pesate con l'inverso della probabilità di selezione e di risposta. Una variante più efficiente è costituita dall'aggiustamento realizzato all'interno di celle mutuamente esclusive, esaustive ed intenzionalmente omogenee, in cui si suddivide il campione originario. In questo caso si suppone che la risposta, che possiamo rappresentare con una variabile indicatrice R, sia indipendente dalla Y e dalla probabilità di inclusione nel campione (inclusione che indichiamo con la variabile dicotomica I), entro i gruppi individuati (C)

$$R \perp (Y, I) \mid C.$$

Con i metodi di imputazione espliciti, invece, ci si propone di determinare un valore sostitutivo che sia il più vicino possibile a quello mancante. La vicinanza viene stabilita sulla base dei valori assunti da un insieme di variabili ausiliarie, che devono essere correlate alla variabile da imputare e disponibili per entrambi gli intervistati ed i non intervistati.

Queste tecniche vengono prevalentemente sfruttate per trattare la non risposta parziale, benché siano in teoria utilizzabili anche in presenza di non risposta totale.

Nell'ambito di queste tecniche, le informazioni ausiliarie possono essere impiegate con differenti scopi: definire celle di imputazione entro cui individuare il valore da sostituire, definire un modello di regressione che fornisca i valori mancanti, od anche quantificare il grado di vicinanza fra le unità, così da identificare la più simile a quella che non ha collaborato (Lessler e Kalsbeek, 1992).

Entro le celle di imputazione si può scegliere di imputare il valore medio, con lo svantaggio però di distorcere la distribuzione della variabile di interesse e di minimizzarne la varianza, o di imputare un valore selezionato a caso. Quest'ultima procedura costituisce una tecnica Hot deck, denominata da Kalton e Kasprzyk (1982) "Hot deck imputazione casuale entro le classi". Gli altri metodi di tipo Hot deck proposti in letteratura sono quello tradizionale, quello gerarchico e quello sequenziale. In generale queste tecniche prevedono che ciascun valore mancante sia rimpiazzato dall'osservazione relativa ad una unità intervistata simile.

Per quanto riguarda l'utilizzo di variabili ausiliarie nell'adattamento di modelli di regressione, questi sono quasi sempre lineari, di tipo deterministico o stocastico. In questo caso non si attribuisce più alle non risposte totali una osservazione relativa ai partecipanti all'indagine, bensì un valore predittivo ottenuto dal modello. L'ipotesi è che la relazione fra le variabili esplicative e quella dipendente sia la stessa per chi ha partecipato all'indagine e chi non l'ha fatto. Il maggior vantaggio di questa tecnica è quello di consentire sempre di trovare un valore sostitutivo per i non intervistati, mentre con i metodi Hot deck, ad esempio, talvolta può essere difficile individuare un donatore appropriato.

Lo stesso inconveniente si può verificare, quando l'abbinamento fra gli intervistati ed i non intervistati viene realizzato minimizzando una funzione di distanza calcolata sui valori assunti dalle variabili esplicative scelte. Trattandosi di una funzione di distanza queste variabili devono essere continue. Perciò tale metodo può essere visto come una versione numerica dell'Hot deck, che si può definire invece categorico in quanto crea classi di imputazione.

Le tecniche sin qui menzionate, producendo un unico valore per ogni osservazione mancante, possono condurre ad una distorsione della

distribuzione naturale della variabile considerata. A questo proposito Rubin (1978) propose l'utilizzo dell'imputazione multipla, in cui l'imputazione singola viene ripetuta con lo stesso criterio più volte, supponiamo m (con m almeno uguale a 2), in modo da ottenere differenti completi *data-set*. La variabilità fra di essi e fra le stime da essi ottenibili, mette in evidenza l'incertezza che caratterizza la scelta del valore da imputare sotto uno stesso modello per le non risposte, ossia quello assunto implicitamente tramite la regola di imputazione scelta. In seguito, si possono sintetizzare le diverse stime facendone la media, al fine di produrre un unica stima della variabile di interesse. La varianza dello stimatore finale si ottiene sommando la media delle varianze delle stime ottenute con l'imputazione singola e la varianza fra di esse. Quest'ultima componente, ignorata dalle comuni tecniche di imputazione, riflettendo la variabilità fra gli m *data-set* ottenuti, esprime la perdita di precisione dovuta alla mancanza dei dati (Herzog e Rubin, 1983). Il principale svantaggio di questo metodo è la mole di lavoro che la sua realizzazione richiede.

L'approccio all'inferenza basato sul modello si distingue da quello classico poiché in esso i valori di una popolazione finita non sono trattati come fissati, bensì come realizzazioni campionarie di variabili casuali che si distribuiscono sulla base di un modello postulato per la superpopolazione (Smith, 1983). L'estensione di tale approccio al problema delle non risposte prevede l'assunzione di un particolare meccanismo stocastico per esse, ovviamente di tipo non ignorabile, la cui forma viene utilizzata al fine di ottenere stimatori che compensino la mancanza di osservazioni. Si possono distinguere attualmente due approcci all'argomento: quello di Royall, in cui l'inferenza è fondata sulle proprietà delle stime per le quantità della popolazione in campionamenti ripetuti dalla distribuzione della superpopolazione, e quello bayesiano di Ericson, dove le distribuzioni a priori dei parametri della superpopolazione sono specificate e l'inferenza si basa sulla distribuzione a posteriori delle quantità della popolazione condizionata ai dati ottenuti (Little, 1982).

Modelli appropriati per la non risposta, che tengano conto sia del meccanismo di campionamento che di quello di risposta e che offrano garanzie contro errori di specificazione, sono attualmente poco sviluppati in letteratura. Uno dei principali svantaggi di questo metodo è rappresentato proprio dalla considerevole sensibilità delle stime ad errori di specificazione nel modello. Tuttavia, i suoi promotori sostengono che

stimatori desunti da un modello appropriato soddisfino proprietà più desiderabili di quelli sviluppati dai metodo classici. Infatti, se si è assunto un modello idoneo, le stime ottenute sono corrette rispetto al modello e la distorsione dovuta al disegno campionario assume in questo caso un ruolo molto meno rilevante che nell'inferenza basata sul disegno (Cassel, Sarndal e Wretman, 1983).

Molte altre proposte sono state avanzate per fare inferenza in presenza di dati mancanti, come, ad esempio, l'applicazione di algoritmi di *data augmentation* (Tanner, 1993) o di tecniche di ricampionamento frequentiste, *bootstrap* e *jack-knife*, congiuntamente alle procedure di imputazione già illustrate (Efron, 1994; Rao e Shao, 1992). In definitiva, la vasta letteratura in merito alle molteplici strategie attuabili rende non banale la decisione su quale di esse sia la più appropriata in ogni particolare situazione, anche perché ciascuno degli approcci utilizzabili si fonda su un insieme di assunzioni che spesso non è possibile verificare. Per questo motivo i giudizi espressi su di essi sono talvolta poco incoraggianti, come quello di Stopher e Sheskin (1981): "*...all traditional methods for dealing with nonresponse bias have been shown to have significant disadvantages*". D'altra parte la necessità di ridurre la distorsione dei risultati finali induce a scegliere di attuare alcuni provvedimenti in grado di risolvere, seppure parzialmente, il problema.

In generale, la scelta deve essere innanzi tutto guidata da considerazioni di tipo statistico e dagli obiettivi dell'indagine, da mettere poi a confronto con i vantaggi e gli svantaggi connessi alle varie tecniche. Innanzi tutti si cercherà il metodo che consente di minimizzare gli effetti della non risposta sulle stime. Poi, se l'obiettivo fosse, ad esempio, studiare una relazione fra variabili, come nella regressione multipla, allora si cercherà un metodo di compensazione che alteri il meno possibile tale relazione, piuttosto che fare riferimento all'errore quadratico medio delle statistiche univariate.

Inoltre, molti metodi di compensazione richiedono la disponibilità di informazioni ausiliarie: per aggiustare i dati o assegnare valori da imputare sono necessarie informazioni su tutte le unità che compongono il campione, mentre alcuni metodi basati sul modello richiedono che le informazioni siano disponibili per l'intera popolazione. Altre considerazioni possono poi fondarsi sui costi da sostenere e sulla complessità di realizzazione dei vari metodi. Ad esempio, la sostituzione ed il sottocampionamento dei rispondenti possono incrementare i costi della rilevazione, mentre approcci

sofisticati come l'imputazione multipla applicata al metodo Hot-deck ed alcuni metodi basati sul modello possono risultare non agevoli e richiedere lunghi tempi di attuazione.

3. Caratteristiche dell'indagine ISTAT

Com'è noto, l'indagine BF si avvale di un disegno campionario in due stadi con stratificazione a priori delle unità di primo stadio, i comuni, e stratificazione a posteriori delle unità di secondo stadio, le famiglie, estratte in modo sistematico dalle liste anagrafiche comunali. La rilevazione è trimestrale e si propone principalmente di raccogliere informazioni sull'ammontare di beni e servizi acquistati o autoconsumati dalle famiglie per soddisfare i bisogni dei singoli individui. Per una descrizione più dettagliata del disegno campionario e degli strumenti di rilevazione si rimanda a Innocenzi (1992).

Le famiglie-campione, che vengono estratte con un tasso di campionamento proporzionale al peso di ciascuno strato di comuni, ammontano a circa 3.250 ogni mese. La loro partecipazione è obbligatoria alla luce delle norme vigenti, ma, nonostante ciò, le famiglie estratte possono rifiutarsi di collaborare, risultare momentaneamente assenti, oppure si possono verificare errori di lista derivanti dal non tempestivo aggiornamento delle informazioni anagrafiche.

Questi sono i motivi che principalmente conducono alle mancate risposte totali, problema a cui si è rivolto il nostro interesse, mentre non abbiamo considerato gli errori che possono derivare dalla compilazione parziale dei libretti di spesa (non risposta parziale), per i quali l'ISTAT procede colmando le informazioni mancanti e verificando che siano compatibili con quelle fornite.

Quando risulta impossibile eseguire la rilevazione su una famiglia, gli intervistatori, che sono scelti dai comuni, hanno il compito di sostituirla con un'altra selezionata con scelta ragionata, ossia possibilmente caratterizzata dallo stesso "numero dei componenti" e dallo stesso tipo di "ubicazione dell'abitazione", che peraltro rappresentano, con il "sesso del capofamiglia" ed il "comune di residenza", le uniche informazioni disponibili sulla famiglia prima della rilevazione.

Le stime delle spese per l'universo si ottengono applicando ai dati ottenuti dalle famiglie dei coefficienti di riporto, che riproporzionano il campione nei vari strati. A tal proposito va precisato, che il riporto all'universo nel procedimento di stima per la popolazione è stato da noi effettuato così come proposto dall'ISTAT, senza affrontare un'analisi critica della definizione dei coefficienti e degli strati, per la mancanza delle informazioni necessarie sulla popolazione. Comunque, è già stato suggerito un nuovo disegno campionario, che consenta di ottenere classi più omogenee rispetto al fenomeno oggetto di indagine (Filippucci e Marliani, 1992).

Poiché si è dovuto procedere ricalcolando i coefficienti, per applicarli alle spese familiari campionarie ottenute con tecniche alternative alla sostituzione, passiamo ad illustrare nel dettaglio come essi vengono costruiti dall'ISTAT.

Facendo riferimento ad un trimestre e ad una data regione geografica, per ciascun strato di comuni in essa individuabile le famiglie osservate vengono distribuite per numero dei componenti. Indichiamo con f_{sh} il numero delle famiglie nel campione di dimensione s appartenenti allo strato *h.mo*. Moltiplicando tali valori per il coefficiente dato dal rapporto fra la popolazione nello strato (P_h) ed il totale dei componenti-campione dello stesso (p_h), si ottiene una prima stima del numero delle famiglie universo per ampiezza familiare e per strato

$$\hat{F}_{sh} = f_{sh}\frac{P_h}{p_h}.$$

Tali stime risulteranno facilmente distorte soprattutto rispetto alla distribuzione per ampiezza, che può essere stimata in modo soddisfacente solo a livello regionale (Falorsi, Falorsi e Russo, 1992). Pertanto, al fine di correggere rispetto alla dimensione familiare, al primo coefficiente $\frac{P_h}{p_h}$ se ne aggiunge un secondo contenente, al numeratore, una stima delle famiglie di ampiezza s nella regione calcolata sulla base dei dati censuari aggiornati con informazioni desunte dall'indagine sulle forze di lavoro, ed al

denominatore, la stima della stessa quantità ottenuta come somma per strato delle stime $\hat{F}_{sh}$

$$\frac{F_s'}{\hat{F}_s} \qquad \text{dove} \qquad \hat{F}_s = \sum_{h=1}^{n}\hat{F}_{hs} \qquad (n\text{=numero di strati nella regione}).$$

La necessità di aggiornare F_s' si deve al modificarsi negli anni della struttura per ampiezza delle famiglie con uno spostamento dalle ampiezze maggiori a quelle minori.

In definitiva il coefficiente di riporto familiare complessivo risulta dal prodotto fra i due coefficienti descritti

$$C_{hs} = \frac{P_h}{p_h}\cdot\frac{F_s'}{\hat{F}_s} = \frac{P_h}{p_h}\cdot\frac{F_s'}{\sum_h f_{hs}\frac{P_h}{p_h}}.$$

4. Utilizzo di procedimenti alternativi alla sostituzione

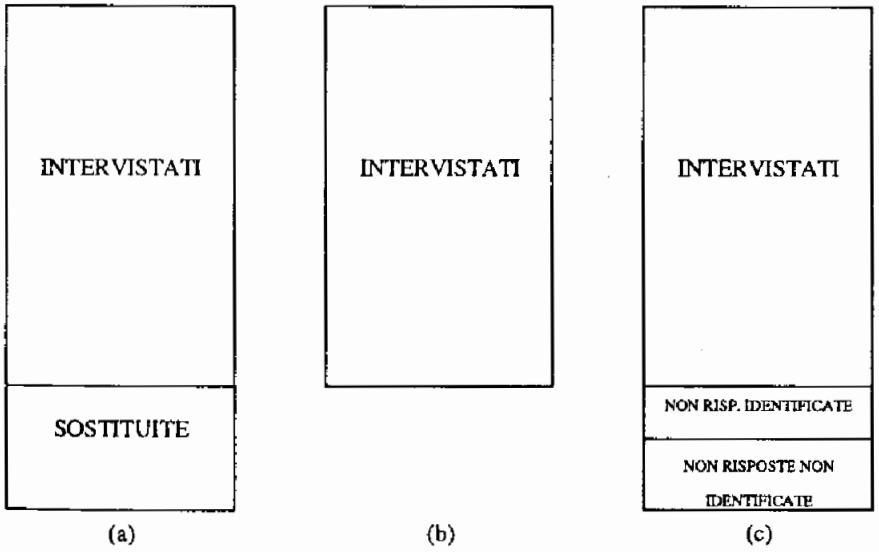
4.1. Informazioni disponibili

La scarsezza che caratterizza, come già detto, le informazioni usualmente disponibili sulle famiglie sostituite non consente di affrontare un'analisi approfondita della non risposta totale e del meccanismo di sostituzione. Pertanto, al fine di acquisire maggiori conoscenze sulle caratteristiche socio-economiche di tali famiglie, sul loro atteggiamento verso l'indagine e sul rispetto delle procedure di rilevazione da parte dei comuni, fu proposto di inviare un breve questionario postale a tutte le famiglie selezionate nel campione che per un qualsiasi motivo non avevano partecipato all'indagine (De Simoni, Filippucci e Marliani, 1992). Tale indagine supplettiva ha avuto luogo negli ultimi due trimestri del 1990, tramite un questionario denominato "scheda di ispezione SMI" (fig. 1 in appendice), che conteneva quesiti riguardanti principalmente le caratteristiche della famiglia e dell'abitazione in cui viveva, il modo in cui era stata contattata ed i motivi della sua mancata collaborazione.

Il campione estratto nel terzo trimestre era costituito da 8.201 famiglie, quello del quarto da 8.301 per un totale di 16.502 unità estratte nell'intervallo di tempo considerato. Di queste, 1.424 erano state sostituite (765 nel terzo trimestre e 659 nel quarto), dando luogo ad un tasso di caduta campionario pari rispettivamente al 9,3% ed al 7,9% nei due periodi. Delle 1.424 schede SMI complessivamente inviate, 550 sono pervenute non in bianco. Per le restanti famiglie non rispondenti le uniche informazioni note rimanevano pertanto quelle anagrafiche.

Lo schema riportato in figura 4.1 riassume le situazioni che ci siamo proposti di mettere a confronto. Il caso (a) rappresenta il *data-set* analizzato dall'ISTAT, nel caso (b) si considera solo la parte relativa agli intervistati, mentre nell'ultimo, caso (c), la disponibilità di informazioni ausiliarie sulle famiglie cadute, differenziata a seconda che queste abbiano o meno partecipato all'indagine postale (rispettivamente "non risposte identificate" e "non risposte non identificate"), ci ha consentito di sperimentare tecniche alternative.

Fig. 4.1: Quadro dei casi messi a confronto.



Le categorie di spese in cui, come già detto, i consumi totali sono stati considerati suddivisi sono: spese per beni alimentari; spese per articoli e servizi correnti (tabacchi, giornali, ecc...); spese per articoli, servizi e beni semidurevoli (abbigliamento, mobili, ecc...) e spese per beni durevoli.

4.2. Analisi dell'insieme incompleto dei dati

Un primo confronto è stato realizzato con le spese stimate per il paese considerando l'insieme incompleto dei dati, ossia escludendo dal campione osservato dall'ISTAT le unità sostitute. Benché questo modo di procedere venga solitamente evitato a causa della scarsa informatività che, come sostiene anche Binder (1991), generalmente caratterizza i dati ottenuti, nell'indagine BF non appare una procedura azzardata, poiché il tasso di non risposta totale non si attesta generalmente su valori molto elevati (10% circa), al contrario di quanto avviene nelle indagini sui bilanci di famiglia di altri paesi (Innocenzi, 1992).

Ovviamente, al fine di ottenere tali stime, si sono ricalcolati i coefficienti di riporto, che, come descritto in precedenza, dipendono dal numero delle famiglie e dal numero di individui osservati.

A questo punto si poneva però il problema dell'impossibilità di ricalcolare con precisione tali rapporti, a causa della non disponibilità delle informazioni sulla popolazione e del numero troppo elevato di C_{hs} da determinare per l'intero territorio nazionale. Pertanto, come peraltro consigliato dallo stesso istituto nazionale di statistica, ci si è limitati a modificare il primo coefficiente, sostituendo a p_h, il numero $\dot{p}_h$ di componenti nello strato h nel campione degli intervistati. In questo modo, la stima della popolazione di strato ottenibile applicando i nuovi coefficienti al campione rimane pressoché inalterata.

Dall'analisi della tavola 4.1, che mostra le stime ottenute considerando solo i partecipanti alla rilevazione e le confronta con quelle conseguite con la sostituzione, emerge il primo risultato rilevante: le differenze sono trascurabili per ogni categoria di spesa considerata e per ogni trimestre (la spesa media familiare è risultata superiore dello 0,13% nel terzo trimestre e dello 0,47% nel quarto). Stando a ciò la sostituzione, che costituisce una soluzione spesso dispendiosa, può essere evitata senza comportare

significative modifiche nei risultati finali. L'unico vantaggio che potrebbe continuare ad offrire sarebbe quello di consentire una riduzione della varianza delle stime dovuta al mantenimento della numerosità campionaria programmata. Tuttavia tali aspetti della non risposta non verranno da noi approfonditi, poiché, come già detto, si ritiene molto più preoccupante la distorsione che ne deriva rispetto alla crescita della varianza.

Tav. 4.1: Spese familiari mensili in lire 1990 stimate con la sostituzione e senza. La differenza percentuale rappresenta la variazione nelle stime ottenuta senza sostituzione.

	SOSTITUZIONE	NIENTE	DIFF. %
	3° trimestre		
Spesa totale	2.540.824	2.544.059	+0,13
Alimentari	737.990	739.081	+0,15
Correnti	748.779	752.231	+0,46
Semidurevoli	926.032	925.362	-0,07
Durevoli	128.021	127.865	-0,12
	4° trimestre		
Spesa totale	2.699.686	2.712.397	+0,47
Alimentari	743.590	745.183	+0,21
Correnti	791.543	796.397	+0,60
Semidurevoli	1.052.749	1.059.193	+0,61
Durevoli	111.804	111.624	-0,16

4.3. Una tecnica di imputazione

Un primo confronto realizzato fra le caratteristiche delle famiglie campione e delle famiglie sostituite, ha messo in evidenza differenze rilevanti rispetto ad un certo numero di variabili che, come risultava anche da precedenti studi (Drudi, 1992), potevano essere considerate legate al fenomeno del consumo. Le variabili considerate sono state:

1 - tipo di comune, AR e NAR[1] (TIPCOM);
2 - regione (REG);

[1] Con AR si intendono i comuni autorappresentativi, ossia i capoluoghi di provincia e quelli con oltre 50.000 abitanti. I restanti comuni sono denominati non autorappresentativi o NAR.

3 - ripartizione territoriale (RIPTER);
4 - mese (MESE);
5 - numero dei componenti la famiglia (COMP);
6 - titolo di studio del capofamiglia (ISTRUZ)*;
7 - condizione del capofamiglia (COND)*;
8 - titolo di godimento dell'abitazione (TITAB)*;
9 - età del capofamiglia (ETA)*;
10 - sesso (SESSO)*.

Quelle denotate con l'asterisco rappresentano informazioni acquisite attraverso il questionario SMI e pertanto non disponibili per tutte le famiglie non intervistate.

Dal confronto, realizzato distintamente per i due trimestri in esame sia perché la rilevazione è trimestrale, sia al fine di evidenziare un eventuale legame fra cadute campionarie e periodo dell'anno, sono emerse alcune tendenze a collocarsi, da parta delle famiglie sostituite, in corrispondenza di particolari modalità dei caratteri. Sono prevalentemente le famiglie con capofamiglia laureato, di età superiore ai 65 anni, di sesso femminile, pensionato od occupato in qualità di dirigente, impiegato, imprenditore e libero professionista a sottrarsi all'indagine. Differenze si riscontravano anche nei confronti realizzati con le variabili note per tutti i non rispondenti, soprattutto per quanto riguarda l'ampiezza della famiglia, (tendono a non partecipare all'indagine i nuclei unipersonali mentre prevalgono quelli composti da 4 o 5 persone), il periodo dell'anno (la partecipazione cala nel mese di agosto[2]) e le caratteristiche geografiche (il tasso di caduta di risposta appare più alto nelle regioni del nord d'Italia e nei comuni AR) (tavv. 1, 2, 3, 4, 5, 6, 7, 8, 9, e 10 in appendice).

Stando alla definizione già illustrata di ignorabilità della non risposta, tali osservazioni inducono a ritenere il meccanismo che governa le caduta campionarie "non ignorabile" in questo caso, se non altro con riferimento al legame fra le esplicative sopra citate e la spesa per consumo.

Questa considerazione ci porta ad escludere alcune soluzioni per la non risposta più semplici, inducendo a vagliare quelle che, tenendo conto di questi legami fra le variabili, li sfruttino nel procedimento di compensazione. Pertanto la scelta si è orientata verso un criterio di

[2] Va ricordato che, poiché l'indagine è trimestrale, la numerosità campionaria può variare nei mesi, sia nel suo complesso che a livello regionale. Ciò nonostante, il confronto fra tassi di non risposta mensili non perde di significato.

imputazione casuale entro celle opportunamente individuate. Inoltre, anziché limitarci ad una singola attribuzione dei valori mancanti, si è deciso di adottare l'imputazione multipla, per i vantaggi che, come già argomentato, essa offre in termini di distorsione nella distribuzione della variabile oggetto di studio.

Per quanto riguarda la determinazione delle celle, questa può risultare: i) da una classificazione incrociata di variabili ausiliarie, che dovrebbero essere correlate con quella di interesse ma non fra loro; ii) determinando direttamente gruppi di unità, omogenei rispetto alla *Y* e sufficientemente numerosi, attraverso tecniche multivariate di raggruppamento o sulla base della propensità a rispondere stimata attraverso un modello Logit o Probit (David e altri, 1983).

Per la scelta delle variabili ausiliarie da considerare per la classificazione incrociata, la regressione multipla, nella versione *stepwise*, consente di selezionare il gruppo di regressori in grado di interpretare al meglio la variabilità della dipendente. La natura esplorativa del processo di selezione delle variabili rende questa procedura una adeguata soluzione al problema. Infatti, trasformando le variabili esplicative, spesso classificate con scale nominali, in variabili *dummy* indicanti l'appartenenza o meno di un'unità alla classe, si può ottenere un sistema di potenziali variabili predittive, attraverso il quale identificare gruppi omogenei (Fabbris, 1983). In pratica, si vogliono identificare strati all'interno dei quali le medie previste siano approssimativamente costanti. Tale metodo è noto infatti come *predicted means stratification.*

Così è stata condotta un'analisi di regressione *stepwise* sull'insieme degli intervistati, con variabile dipendente pari alla spesa familiare mensile e come potenziali esplicative le variabili: numero dei componenti; età in classi del capofamiglia; titolo di studio del capofamiglia; ripartizione territoriale; trimestre; tipo di comune; sesso del capofamiglia e condizione del capofamiglia. In particolare, vale la pena di osservare come la variabile "condizione del capofamiglia" sia stata costruita in modo da riassumere sia la condizione lavorativa che l'eventuale posizione occupata nella professione dal capofamiglia, formando le seguenti classi

1 - pensionato;
2 - dirigente, impiegato o quadro intermedio;
3 - operaio o altro tipo di dipendente;
4 - imprenditore, libero professionista;

5 - lavoratore in proprio (artigiano, commerciante o coltivatore diretto);
6 - disoccupato;
7 - altro.

Le *dummy* selezionate utilizzando come criterio d'arresto del processo quello del quadrato del coefficiente di correlazione parziale (la quota della varianza spiegata dalle variabili predittive già entrate nell'equazione, spiegata dall'ultima variabile entrata deve essere maggiore al 5 per mille), suddividevano le variabili esplicative in questo modo

età del capofamiglia	1) meno di 65 anni, 2) oltre i 65 anni;
numero dei componenti	1) un componente, 2) due componenti, 3) tre componenti, 4) quattro componenti e oltre;
livello di istruzione	1) da analfabeta a scuola media inferiore, 2) scuola media superiore, 3) laurea;
trimestre	1) terzo, 2) quarto;
ripartizione territoriale	1) nord, 2) centro e sud.

L'R^2 corretto non è risultato molto elevato (0,3), come peraltro ci si aspettava, ma la corretta specificazione del modello non assume una grande importanza in questo caso, poiché la regressione non viene utilizzata direttamente per l'imputazione, bensì al fine di ordinare parzialmente il campione (Little, 1986).

D'altra parte, poiché, come si è detto, le variabili indicanti i gruppi dovrebbero essere poco correlate fra loro, è stata svolta anche un'analisi della multicollinearità, seguendo l'approccio suggerito da Belsley, Kuh e Welsch (1980) basato sulla proporzione della varianza delle stime dei coefficienti di regressione $\hat{\beta}_j$ associata a ciascun autovalore della matrice $X'X$. I risultati ottenuti hanno messo in evidenza un livello di multicollinearità trascurabile (*condition indices* inferiori a 30).

A tal punto si è proceduto con l'imputazione multipla delle spese all'interno delle celle così determinate per ciascuna categoria di beni, escluso i durevoli, per i quali l'imputazione è stata preceduta da un'analisi della probabilità di acquisto come verrà descritto nel par. 2.3.1.

Il procedimento sin qui esposto è stato realizzato per le famiglie sostituite che avevano partecipato all'indagine postale condotta su di esse.

Tuttavia, il nostro obiettivo era quello di compensare le cadute campionarie nel loro complesso, in modo da ottenere stime che tenessero conto dei non intervistati, senza escludere coloro per i quali si disponeva di un numero inferiore di informazioni. Di conseguenza, supponendo che quanti avevano rispedito i questionari SMI compilati potessero rappresentare l'insieme completo delle unità cadute, si è proceduto aggiustando i pesi delle prime all'interno degli strati di appartenenza e della dimensione familiare, in modo da tener conto anche delle altre famiglie cadute. La scelta delle celle era questa volta vincolata alla necessità di utilizzare gli stessi coefficienti di riporto alla popolazione previsti dall'indagine ISTAT, che vengono elaborati proprio per strato e per numero dei componenti.

L'idea di ripesare i dati ottenuti dal gruppo degli intervistati al secondo tentativo trattandolo come rappresentativo del complesso delle non risposte totali alla prima indagine, fu suggerita per la prima volta da Bartholomew (1961). Ovviamente la riduzione della distorsione, con questo approccio, è direttamente correlata alla somiglianza fra gli intervistati alla seconda chiamata ed i non intervistati ad entrambe le chiamate. Nel nostro caso l'ipotesi che il meccanismo di autoselezione che governa la risposta alla scheda SMI sia ignorabile, e quindi il campione di non risposte ottenuto sia trattabile alla stregua di un campione casuale semplice, può apparire un po' forzata. Ma ammettendo una relazione inversa fra spese e propensione a partecipare all'indagine, come spesso ipotizzato per il reddito (Greenless, Reece e Zieschang, 1982), l'*hard core* di famiglie non partecipanti all'indagine, che avevano in prevalenza rifiutato di collaborare alla rilevazione BF, dovrebbe essere caratterizzato da consumi ancor più elevati. Pertanto i risultati che otteniamo con la nostra assunzione potrebbero al limite sottostimare la spesa per consumo e non sovrastimarla.

Le stime per la popolazione sono state infine ottenute correggendo ancora una volta i coefficienti di riporto, così come descritto per l'analisi dell'insieme incompleto dei dati, in modo da considerare l'insieme delle famiglie campione selezionato originariamente.

Come si evince dalla tavola 4.2, in cui sono riportati i risultati ottenuti, la spesa media risulta complessivamente più elevata nel quarto trimestre con entrambi i metodi. In particolar modo sono gli acquisti di beni correnti e semidurevoli a subire un incremento.

Per quanto riguarda le due tecniche utilizzate, quella da noi sperimentata ha fornito una stima della spesa totale familiare superiore, se

confrontata con quella ottenuta con la sostituzione, rispettivamente dell'1,2% e dell'1,5% nei due trimestri in esame. Poiché la spesa totale rappresenta la somma dei consumi delle diverse categorie di acquisti considerate, si può notare come tale variazione si debba soprattutto ad un incremento delle stime ottenute per i beni semidurevoli (+1,7% nel terzo trimestre e +1,9% nel quarto con l'imputazione) e per i beni durevoli (per i risultati relativi a questi ultimi si rimanda al prossimo paragrafo, dopo la spiegazione del metodo utilizzato). Passando alle spese per beni alimentari, l'incremento nelle stime risulta più contenuto (+0,7% e +0,8% nei due periodi), mentre per i correnti si può notare una maggior crescita nel quarto trimestre rispetto a quello risultante per il terzo.

Tav. 4.2: Spese familiari mensili in lire 1990 stimate con la sostituzione e con l'imputazione. La differenza percentuale rappresenta la variazione nelle stime ottenuta con l'imputazione.

	SOSTITUZIONE	IMPUTAZIONE	DIFF. %
	3° trimestre		
Spesa totale	2.540.824	2.572.692	1,2%
Alimentari	737.990	743.270	0,7%
Correnti	748.779	754.347	0,7%
Semidurevoli	926.032	942.202	1,7%
	4° trimestre		
Spesa totale	2.699.686	2.741.767	1,5%
Alimentari	743.590	749.760	0,8%
Correnti	791.543	803.988	1,5%
Semidurevoli	1.052.749	1.072.862	1,9%

Poiché i coefficienti di riporto alla popolazione sono stati aggiustati in modo da mantenere pressoché inalterato il numero di famiglie e di individui complessivamente stimati per l'Italia con essi, le spese considerate a livello *pro-capite* mostrano approssimativamente le stesse variazoni osservate per i consumi familiari.

Tali aumenti registrati nelle stime familiari vanno comunque valutati tenendo conto dell'ordine di grandezza dei loro errori relativi percentuali. Poiché il calcolo di tali errori, data la complessità del disegno campionario, appariva piuttosto complicato, si è fatto riferimento ai valori per essi

stimati da Falorsi, Falorsi e Russo (1992) con riferimento ad alcuni singoli capitoli di spesa per il 1990. Queste stime risultano generalmente al di sotto dell'1% per le spese alimentari (come pane, pasta, ecc...) ed un po' più elevati, ossia dall'1 al 5% circa per i generi non alimentari (come abiti, calzature, tabacchi, ecc...). Benché questi errori dovrebbero subire una riduzione se calcolati per le più ampie categorie di spese da noi considerate, gli incrementi ottenuti con l'imputazione non possono essere giudicati considerevoli. Comunque, il fatto che si siano sistematicamente ottenute delle stime più elevate, per ogni categoria di acquisti e per ogni trimestre, fornisce una evidenza della sottostima probabilmente indotta dal processo di sostituzione.

La diversa efficacia delle due tecniche di trattamento della non risposta totale risulta invece più evidente, quando ci si limita ad osservare le famiglie sostitute e le famiglie sostituite a cui sono state imputate le spese, ossia la parte di campione relativa alla non risposta, che ne rappresenta circa l'8-9%. Infatti, le differenze fra i consumi medi familiari stimati da questi due gruppi di unità trattate in modo differente appaiono decisamente più rilevanti, ovvero più elevati con l'imputazione di circa 7-12 punti percentuali (tav. 4.3).

Tav. 4.3: Spese familiari stimate nel sottoinsieme delle famiglie sostitute ed in quello delle famiglie non intervistate trattate con l'imputazione. La differenza percentuale rappresenta la variazione nelle stime ottenuta con l'imputazione.

	SOSTITUTE	FAM. TRATTATE CON L'IMPUTAZIONE	DIFF. %
	3° trimestre		
Spesa totale	2.485.556	2.683.286	7,4%
Alimentari	716.442	728.496	1,7%
Correnti	725.564	741.026	2,1%
Semidurevoli	929.042	1.049.108	11,4%
	4° trimestre		
Spesa totale	2.444.587	2.781.195	12,1%
Alimentari	678.030	713.353	4,9%
Correnti	678.767	771.775	12,0%
Semidurevoli	967.568	1.137.805	14,9%

In generale, la differenza riscontrata nelle spese si può considerare frutto della diversa composizione dei sottoinsiemi costituiti dalle famiglie sostitute e dalle famiglie sostituite considerate per l'imputazione, oltre che della particolare metodologia utilizzata. Pertanto, la si può considerare un indicatore della diversità dei comportamenti nei due gruppi: i non intervistati appaiono caratterizzati da una più elevata propensione al consumo. Se poi la spesa complessiva si può considerare *proxy* del reddito, si spiega come mai tale maggiore propensione al consumo dei non rispondenti sia particolarmente accentuata verso i beni non alimentari, portandoci ai risultati osservati per i semidurevoli.

In questo caso, per le stime delle spese *pro-capite* le differenze fra i metodi si accentuano a causa delle diverse distribuzioni per numero di componenti delle famiglie che compongono i due gruppi. La presenza di un numero più folto di nuclei unipersonali fra i non rispondenti porta a stimare un numero maggiore di famiglie e minore di individui rispetto a quanto non avvenga con il sottocampione dei sostituti. Così le differenze osservate per le stime delle spese familiari aumentano ulteriormente di circa 2-3 punti percentuali a livello *pro-capite*.

Ovviamente, per come sono stati costruiti i coefficienti di riporto applicati a tutto il campione, si verifica il contrario esaminando le spese desunte dal sottogruppo più numeroso composto esclusivamente dalle famiglie che hanno partecipato all'indagine.

4.4. Trattamento delle spese per beni durevoli

Gli acquisti di beni durevoli[3], benché siano rilevati dall'ISTAT con riferimento ad un intervallo trimestrale, sono caratterizzati da una frequenza talmente scarsa da causare il prevalere degli zeri di spesa. Una distribuzione così "troncata", con un picco in corrispondenza del valore zero, si presta ad un tipo di imputazione in due fasi (Ford, Kleweno e

[3] Nella categoria beni durevoli sono comprese le seguenti merci: automobile; moto, scooter e motorino; roulotte, rimorchio e camper; canotto, gommone, barca, motoscafo e wind-surf; televisore; videoregistratore e telecamera; personal computer e periferiche; registratore, giradischi, lettore di compact disc, alta fedeltà (piastra, amplificatore, ecc...); radio, autoradio e radio portatili; lavatrice; frigorifero, congelatore, surgelatore e combinati; lavastoviglie; lucidatrice, aspirapolvere e battitappeto; condizionatore e umidificatore d'aria; cucina, forno a microonde, stufa e scaldabagno; macchina da scrivere.

Tortora, 1981). Nella prima fase si stabilisce a quali unità, fra quelle non rispondenti, assegnare una spesa per durevoli nulla ed a quali una spesa non nulla, per poi attribuire, nella seconda fase, un valore preciso unicamente a queste ultime.

Le famiglie cadute a cui imputare una spesa maggiore di zero sono state identificate con l'ausilio di un modello di regressione logistica, che consente di stimare la probabilità di acquisto di ciascuna famiglia condizionatamente alle sue caratteristiche. La variabile dipendente contemplata dal modello era una variabile dicotomica distinguente fra acquirenti e non acquirenti di beni durevoli nel trimestre.

Il modello, adattato dapprima nell'insieme dei rispondenti, ha messo in evidenza come gran parte delle variabili esplicative note influiscano sull'acquisto di tali beni. Infatti, la versione che meglio si è adattata sulla base del rapporto di verosimiglianza e del criterio AIC di Akaike, conteneva come esplicative le variabili: numero dei componenti la famiglia, età (x), livello di istruzione e condizione del capofamiglia, mese in cui la rilevazione aveva avuto luogo e regione di residenza della famiglia

$$\log\left(\frac{m_{ijkhs1}}{m_{ijkhs0}}\right) = a + \tau_i^{comp} + \tau_j^{istruz} + \tau_k^{cond} + \tau_h^{mese} + \tau_s^{reg} + \beta x.$$

0 e 1 indicano, come già detto, l'acquisto o meno di beni durevoli. L'intercetta α, che costituisce la media dei logit, ed i coefficienti τ, che esprimono l'influenza esercitata da ogni singola modalità sulla variabile dipendente, sono stati stimati con il metodo della massima verosimiglianza.

Dai test effettuati sui parametri stimati (test di Wald basato sulla matrice delle informazioni) e dai valori assunti dalle misure della bontà dell'adattamento, riportati nella tavola 4.4, si evince che tutte le variabili sopra elencate sono altamente significative e che l'adattamento raggiunge un buon livello (la probabilità associata al rapporto di verosimiglianza si approssima a uno).

Tav. 4.4: Risultati dei test effettuati sul modello logit.

	Gradi di libertà	Wald test	Prob.
INTERCETTA	1	136,30	0,0000
COMP	5	193,02	0,0000
ETA	1	10,42	0,0012
ISTRUZ	4	13,01	0,0112
COND	6	19,77	0,0030
MESE	5	38,90	0,0000
REG	19	217,16	0,0000
Rapp. di verosimiglianza	15474	10487,23	1,0000

AIC = - 2log(L^*) + 2p = 11.038,587 (con L^* = funz. di verosim. e p = n° dei parametri)

In particolare il modello stimato mette in evidenza che la probabilità di acquistare beni durevoli nel trimestre risulta particolarmente alta quando si incrociano le seguenti modalità:
- le regioni Trentino Alto Adige, Veneto, Fiuli Venezia Giulia, Emilia Romagna e Sardegna;
- i mesi di dicembre e di luglio;
- i livelli di istruzione del capofamiglia più alti (si vede che tale probabilità cresce al crescere di del titolo di studio);
- i capofamiglia occupati con le qualifiche di dirigenti e impiegati, imprenditori e liberi professionisti, lavoratori in proprio, mentre è molto bassa per i disoccupati;
- per le più elevate dimensioni familiari (cresce con il crescere di tali dimensioni);
- per le età intermedie del capofamiglia (35-55 anni), mentre è molto bassa per le famiglie con capofamiglia ultra-sassantacinquenne.

Lo stesso modello è stato poi applicato alle famiglie non intervistate che avevano risposto alla SMI. Per quelle a cui corrispondeva una probabilità di acquisto stimata superiore ad una soglia, pari a circa la quota dei non acquirenti nel campione, si è proceduto con l'imputazione multipla.

Le famiglie che non hanno collaborato, come si evince anche dalle tavole in appendice, risultano concentrarsi maggiormente in corrispondenza delle modalità associate ad una più elevata spesa per durevoli rispetto alle famiglie-campione, soprattutto per quanto riguarda le variabili regione, mese, titolo di studio e condizione del capofamiglia. Infatti, come ci si

attendeva, la spesa famigliare per durevoli stimata per la popolazione con lo stesso procedimento descritto per le altre categorie di acquisti è risultata più elevata (tavola 4.5), soprattutto nel terzo trimestre (+3,7%), in cui peraltro il tasso di non risposta totale superava di circa 1,5 punti percentuali quello registratosi nel quarto trimestre.

Tav. 4.5: Spese familiari mensili per beni durevoli in lire 1990 stimate con la sostituzione e l'imputazione. La differenza percentuale rappresenta la variazione ottenuta nelle stime con l'imputazione.

	SOSTITUZIONE	IMPUTAZIONE	DIFF. %
	3° trimestre		
Durevoli	128.021	132.873	3,7%
	4° trimestre		
Durevoli	111.804	115.156	2,9%

Se ci si limita a considerare, come già fatto per le altre categorie di beni, le spese stimate dalle famiglie sostitute e dalle famiglie che non collaborano trattate con l'imputazione, le differenze fra i risultati medi familiari si amplificano anche in questo caso, attestandosi rispettivamente sul 30,5 e 24,0% nei due trimestri (tav. 4.6).

Tav. 4.6: Spese familiari per beni durevoli stimate nel sottoinsieme delle famiglie sostitute ed in quello delle famiglie non intervistate trattate con l'imputazione. La differenza percentuale rappresenta la variazione ottenuta nelle stime con l'imputazione.

	SOSTITUTE	FAM. TRATTATE CON L'IMPUTAZIONE	DIFF. %
	3° trimestre		
Durevoli	114.508	164.657	30,5%
	4° trimestre		
Durevoli	120.221	158.262	24,0%

Stando alle stime degli errori relativi percentuali familiari calcolate da Falorsi, Falorsi e Russo (1992), che per i vari tipi di beni durevoli risultano mediamente pari a al 6-7%, tale differenza si rivela decisamente significativa, benché poi, come si è già osservato, essa si riduca

notevolmente nella stima complessiva realizzata riportando alla popolazione le spese desunte dal campione considerato nella sua interezza.

5. Considerazioni conclusive

Un primo risultato che emerge dall'analisi condotta riguarda il distribuirsi in modo non casuale delle cadute campionarie, che, al contrario, si concentrano in corrispondenza di particolari segmenti della popolazione (monocomponenti, laureati, pensionati o occupati come dirigenti, liberi professionisti, ecc...). Pertanto, il trattamento della non risposta appare necessario per evitare l'insorgere di un problema di distorsione nelle stime.

I confronti realizzati in seguito fra le spese stimate avvalendosi di due strategie di trattamento della non risposta alternative alla sostituzione hanno fornito interessanti evidenze in merito agli effetti di quest'ultima sulle stime.

Per entrambi i trimestri in esame con il primo metodo utilizzato, che consisteva nel trascurare le famiglie sostituite e stimare così le spese solo attraverso l'insieme degli intervistati, le stime si avvicinano moltissimo a quelle ottenute dall'ISTAT mediante la sostituzione. Quest'ultima appare quindi un provvedimento che è possibile evitare, senza che ciò comporti significativi peggioramenti nelle stime. L'unico vantaggio che questo metodo può continuare ad offrire rispetto all'analisi dell'insieme incompleto dei dati è la riduzione della varianza delle stime dovuta al mantenimento della numerosità campionaria programmata.

Trattando invece le cadute campionarie con una tecnica di imputazione multipla, abbinata ad un aggiustamento dei pesi entro gli strati, le stime delle spese familiari mensili sono risultate più elevate. A livello complessivo l'incremento si è attestato sull'1,2% e l'1,5% nei due trimestri in esame, e si osserva che è più contenuto per i beni alimentari, mentre le spese per i durevoli risentono di un maggiore incremento (3,7% e 2,9% nei due periodi). Tuttavia, se si valutano tali aumenti alla luce degli errori relativi percentuali delle stime, il rilievo delle differenze non appare particolarmente apprezzabile.

Ciò non toglie che, se ci si limita ad analizzare i sottogruppi composti dalle unità cadute e da quelle utilizzate per la sostituzione (che costituiscono circa l'8-9% del campione complessivo) le stime delle spese

ottenute con l'imputazione si rivelano decisamente più elevate per ogni categoria di beni, raggiungendo per i durevoli circa il 30%. La tecnica di imputazione appare quindi molto più adatta a ridurre la componente di sottostima che notoriamente influenza le stime dei consumi per beni non alimentari.

Naturalmente, considerata la scarsa incidenza della non risposta totale in questa indagine, le differenze osservate fra i metodi si attenuano considerevolmente a livello complessivo, risultando meno rilevanti. I livelli di consumo sempre superiori ottenuti con le tecniche di imputazione, per ogni categoria di beni e per ogni trimestre, mettono comunque in evidenza la presenza di una distorsione nelle stime finali desunte dall'indagine, conseguente alla non risposta ed al metodo di sostituzione utilizzato.

In sintesi, dal momento che il ricorso alla sostituzione, peraltro non facilmente controllabile e relativamente costosa, non produce particolari effetti sulla qualità delle stime, vale la pena di esplorare le possibilità offerte dai procedimenti di imputazione per il miglioramento della valutazione delle spese.

Riferimenti bibliografici

BAILAR B. A., L. BAILEY, C. CORBY (1978), 'A Comparison of some Adjustment and Weighting Procedures for Survey Data', *Survey Sampling and Measurement*, N. K. Namboodiri, Academic Press, New York.

BARTHOLOMEW D. J. (1961), 'A Method of Allowing for "Not-at-home" Bias in Sample Surveys', *Applied Statistics*, vol. 10.

BELSLEY D. A., E. KUH, R. E. WELSCH (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

BINDER D. A. (1991), 'A Framework for Analysing Categorical Survey Data with Nonresponse', *Journal of Official Statistics*, vol. 7, n. 4.

CASSEL C. M., C. E. SARNDAL, J. H. WRETMAN (1983), 'Some Uses of Statiatical Models in Connection with the Nonresponse Problem', in W. G. Madow e I. Olkin eds., *Incomplete Date in Sample Survey*, vol. 3, *Proccedings of the Symposium*, Academic Press, New York.

CHAPMAN D. W. (1982), 'Substitution for Missing Units', *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

CHAPMAN D. W., A. ROMAN (1985), 'An Investigation of Substitution for an RDD Survey', *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

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

DAVID M., R. J. A. LITTLE, M. SAMUHEL, R. TRIEST (1983), 'Nonrandom Nonresponse Model Based on the Propensity to Respond', *Proc. Bus. Econ. Statist. Sec., American Statistical Association.*

DE SIMONI S., C. FILIPPUCCI, G. MARLIANI (1992), *Un progetto di ricerca sulla misura dei consumi privati in Italia*, CON.PRI. Rapporto di ricerca n. 1, Bologna, Dipartimento di Scienze Statistiche.

DONALD M. N. (1960), 'Implication of Nonresponse for the Interpretation of Mail Questionnarie Data', *Public Opinion Quarterly*, vol. 24.

DRUDI I. (1992), *Analisi delle frequenze di acquisto nell'indagine sui consumi delle famiglie,* CON.PRI. Rapporto di ricerca n. 8, Bologna, Dipartimento di Scienze Statistiche.

DURBIN J., A. STUART (1954), 'Callback and Clustering in Sample Survey: An Experimental Study', *The Journal of the Royal Statistical Society*, Serie A, vol. 117, part. 4.

EFRON B. (1994), 'Missing Data, Imputation and the Bootstrap', *Journal of the American Statistical Association*, vol. 89, n. 426.

ERICKSON W. A. (1967), 'Optimal Sample Design with Nonresponse', *Journal of the American Statistical Association*, vol. 62.

FABBRIS L. (1983), *Analisi Esplorativa dei Dati Multidimensionali*, CLUEP, Padova.

FALORSI P. D., S. FALORSI, A. RUSSO (1992), *Indagine campionaria sui consumi delle famiglie: strategia di campionamento e precisione delle stime*, CON.PRI. Rapporto di ricerca n. 3, Bologna, Dipartimento di Scienze Statistiche.

FILIPPUCCI C., G. MARLIANI (1992), *La misura dei consumi delle famiglie: una riflessione a partire dall'esperienza italiana*, CON.PRI. Rapporto di ricerca n. 6, Bologna, Dipartimento di Scienze Statistiche.

FORD B. L., G. K. DOUGLAS, R. D. TORTORA (1981), 'The Effects of Procedures which Impute for Missing Items: a Simulation Study Using an Agricultural Survey', *Current Topics in Survey Sampling*, Academic Press, New York.

GREENLESS W. S., J. S. REECE, K. D. ZIESCHANG (1982), 'Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputated', *Journal of the American Statistical Association*, vol. 77.

HANSEN M. H., W. N. HURWITZ (1946), 'The Problem of Nonresponse in Sample Survey', *Journal of the American Statistical Association*, vol. 41.

HERZOG T. N., D. B. RUBIN (1983), 'Using Multiple Imputation to Handle Nonresponse in Sample Surveys', *Incomplete Data in Sample Surveys*, vol. 2, *Panel on Incomplete Data*, Academic Press, New York.

INNOCENZI G. (1992), *Principali Aspetti dell'indagine ISTAT sui Consumi delle Famiglie*, CON.PRI. Rapporto di ricerca n. 2, Bologna, Dipartimento di Scienze Statistiche.

KALTON G., D. KASPRZYK (1982), 'Imputation for Missing Survey Response', *Proceedinds of the Section on Survey Research Methods, American Statistical Association.*

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

LESSLER J. T., W. D. KALSBEEK (1992), *Nonsapling Error in Survey*, John Wiley & Sons, New York.

LITTLE R. J. A. (1982), 'Models for Nonresponse in Sample Surveys', *Journal of the American Statistical Association*, vol. 77.

LITTLE R. J. A. (1986), 'Survey Nonresponse Adjustments for Estimates of Means', *International Statistical Review*, vol. 54, n. 2.

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

MARBACH G. (1964), 'Sui Campioni di Composizione Non Proporzionale', *Atti della 24ª Riunione Scientifica della Società Italiana di Statistica.*

PARTEN M. B. (1966), *Survey, Polls and Samples Practical Procedures*, Cooper Square, New York.

PLATEK R., M. P. SINGH, V. TREMBLAY (1978), 'Adjustament for Nonresponse in Survey', *Survey Sampling and Measurement*, N. K. Namboodiri, Academic Press.

RAO J. N. K., J. SHAO (1992), 'Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation', *Biometrika*, vol. 79.

RUBIN D. B. (1976), 'Inference and Missing Data', *Biometrika*, vol. 63, n. 3.

RUBIN D. B. (1978), 'Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Non Response', *Imputation and Editing of Faulty or Missing Data*, U. S. Department of Commerce, Social Security Administration, Washington D. C. (anche in *Proceeding American Statistical Association Section Survey Research Methods*, 1978).

SARNDAL C. E., B. SWENSSON, J. WRETMANN (1992), *Model Assisted Survey Sampling*, New York, Springer.

SMITH T. M. F. (1983), 'On the Validity of Inference from Non-random Samples', *Journal of the Royal Statistical Society*, Serie A, n. 146, part 4.

STOPHER P., I. SHESKIN (1981), 'A Method for Determining and Reducing Nonresponse Bias', *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

WILLIAMS S., R. E. FOLSOM JR (1977), *Bias Resultig from School Nonresponse: Methodology and Findings*, National Centre for Educational Statistics, New York.

Appendice

Figura 1: Scheda SMI

istat

INDAGINE SUI BILANCI DI FAMIGLIA-RICERCA SULLA QUALITA' DEI DATI
QUESTIONARIO SULLE MODALITA' DI ESECUZIONE DELL'INDAGINE

1. Può fornirci le seguenti informazioni relative alla sua famiglia e all'abitazione:

a) Numero di componenti la famiglia |__|__|

b) Numero di componenti la famiglia che lavorano |__|__|

c) L'abitazione nella quale vive la famiglia è:
- di proprietà ..... ❑1
- in affitto ..... ❑2
- utilizzata ad altro titolo ..... ❑3

d) Età del capofamiglia |__|__|

e) Sesso del capofamiglia (M=1, F=2) |__|

f) Titolo di studio del capofamiglia:
- nessun titolo ..... ❑1
- licenza elementare ..... ❑2
- diploma media inferiore ..... ❑3
- diploma media superiore ..... ❑4
- laurea ..... ❑5

g) Condizione del capofamiglia:
- Occupato ..... ❑1
- In cerca di prima occupazione ..... ❑2
- In cerca di nuova occupazione ..... ❑3
- Pensionato ..... ❑4
- Casalinga ..... ❑5
- Altra condizione ..... ❑6

h) Se il capofamiglia è occupato, indicare se
- Alle dipendenze, come:
  - dirigente o figura assimilata ..... ❑1
  - impiegato ..... ❑2
  - quadro intermedio ..... ❑3
  - operaio ..... ❑4
  - altro dipendente ..... ❑5
- In conto proprio, come:
  - imprenditore ..... ❑6
  - libero professionista ..... ❑7
  - lavoratore in proprio (artigiano, commerciante, coltivatore diretto) .... ❑8
  - coadiuvante ..... ❑9

2. La sua famiglia ha ricevuto dal Comune una lettera nella quale si comunicava che avrebbe dovuto partecipare all'indagine Istat sui consumi? SI ❑1 NO ❑2

3. Un rilevatore incaricato dal Comune si è mai messo in contatto con lei o con altro membro della famiglia (personalmente o per telefono) per richiedere la vostra collaborazione all'indagine?

SI ❑1 → 4.1. Se SI, la sua famiglia ha collaborato all'indagine?
- SI ❑1
- NO: ho rifiutato fin dall'inizio ❑2
- NO: ho rifiutato durante l'indagine ❑3

4.1.1. Se non ha collaborato, può dirci per quale motivo? (*Barrare una sola risposta*)
- si trattava di un lavoro troppo gravoso ❑1
- non avevo abbastanza tempo da dedicarvi ❑2
- la famiglia doveva assentarsi in quel periodo ❑3
- sono contrario ai sondaggi ❑4
- avevo dubbi sull'anonimato dell'indagine ❑5
- altro motivo ❑6

NO ❑2 → 4.2. Se NO, la sua famiglia sarebbe disposta a partecipare ad un'indagine sui consumi che prevede la trascrizione giornaliera su un apposito libretto di tutte le spese sostenute per un periodo di 10 giorni? SI ❑1 NO ❑2

Tav. 1a: Distribuzione per regione delle famiglie-campione e delle famiglie non rispondenti (terzo trimestre).

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Piemonte	585	7,1	112	14,6
Valle d'Aosta	246	3,0	32	4,2
Lombardia	861	10,5	123	16,1
Trentino-A. A.	451	5,5	57	7,5
Veneto	441	5,4	46	6,0
Friuli-V. G.	236	2,9	14	1,8
Liguria	388	4,7	12	1,6
Emilia-Romagna	488	6,0	75	9,8
Toscana	566	6,9	49	6,4
Umbria	254	3,1	22	2,9
Marche	362	4,4	18	2,4
Lazio	569	6,9	49	6,4
Abruzzo	253	3,1	7	0,9
Molise	262	3,2	17	2,2
Campania	486	5,9	22	2,9
Puglia	447	5,5	20	2,6
Basilicata	230	2,8	13	1,7
Calabria	297	3,6	17	2,2
Sicilia	519	6,3	36	4,7
Sardegna	260	3,2	24	3,1
Totale	8.201	100,0	765	100,0

Tav. 1b: Distribuzione per regione delle famiglie-campione e delle famiglie non rispondenti (quarto trimestre).

	Campione		Non rispondenti	
4° trimestre 1990	N	%	N	%
Piemonte	606	7,3	80	12,1
Valle d'Aosta	233	2,8	27	4,1
Lombardia	938	11,3	122	18,5
Trentino-A. A.	467	5,6	37	5,6
Veneto	429	5,2	36	5,5
Friuli-V. G.	230	2,8	19	2,9
Liguria	409	4,9	25	3,8
Emilia-Romagna	499	6,0	68	10,3
Toscana	573	6,9	26	3,9
Umbria	274	3,3	24	3,6
Marche	365	4,4	10	1,5
Lazio	600	7,2	39	5,9
Abruzzo	257	3,1	10	1,5
Molise	233	2,8	0	0,0
Campania	474	5,7	36	5,5
Puglia	440	5,3	28	4,2
Basilicata	198	2,4	6	0,9
Calabria	317	3,8	13	2,0
Sicilia	522	6,3	38	5,8
Sardegna	237	2,9	15	2,3
Totale	8.301	100,0	659	100,0

Tav. 2: Distribuzione per tipo di comune di residenza delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
AR	4.073	49,7	522	68,2
NAR	4.128	50,3	243	31,8
Totale	8.201	100,0	765	100,0
4° trimestre 1990	N	%	N	%
AR	4.163	50,2	450	68,3
NAR	4.138	49,8	209	31,7
Totale	8.301	100,0	659	100,0

Tav. 3: Distribuzione per ripartizione territoriale delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Nord	3696	45,1	471	61,6
Centro	1751	21,3	138	18,0
Sud	2754	33,6	156	20,4
Totale	8201	100,0	765	100,0
4° trimestre 1990	N	%	N	%
Nord	3811	45,9	414	62,8
Centro	1812	21,8	99	15,0
Sud	2678	32,3	146	22,2
Totale	8301	100,0	659	100,0

Tav. 4: Distribuzione per numero dei componenti delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
1	1.447	17,6	221	28,9
2	2.026	24,7	187	24,4
3	2.007	24,5	168	22,0
4	1.864	22,7	142	18,6
5	606	7,4	34	4,4
6	167	2,0	8	1,0
oltre 6	84	0,9	5	0,7
Totale	8.201	100,0	765	100,0
4° trimestre 1990	N	%	N	%
1	1.404	16,9	228	34,6
2	2.021	24,3	164	24,9
3	2.013	24,3	134	20,3
4	1.997	24,1	100	15,2
5	621	7,5	25	3,8
6	172	2,1	7	1,1
oltre 6	73	0,8	1	0,2
Totale	8.301	100,0	659	100,0

Tav. 5: Distribuzione per titolo di studio del capofamiglia delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Nessuno	862	10,5	20	6,2
Licenza elementare	3.090	37,7	92	28,4
Diploma media inf.	2.213	27,0	77	23,8
Diploma media sup.	1.589	19,4	85	26,2
Laurea	447	5,5	50	15,4
Totale	8.201	100,0	324	100,0
4° trimestre 1990	N	%	N	%
Nessuno	817	9,8	15	7,0
Licenza elementare	3.048	36,7	51	24,1
Diploma media inf.	2.254	27,2	51	24,1
Diploma media sup.	1.702	20,5	53	25,0
Laurea	480	5,8	42	19,8
Totale	8.301	100,0	212	100,0

Tav. 6: Distribuzione per condizione occupazionale del capofamiglia delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Pensionati	2.355	28,7	121	37,3
Dirigenti e impiegati	1.699	20,7	86	26,5
Operai e assimilati	1.948	23,8	47	14,5
Imprenditori e liberi professionisti	290	3,5	19	5,9
Lavoratori in proprio	1.097	13,4	27	8,3
Disoccupati	88	1,1	6	1,9
Altra condizione	724	8,8	18	5,6
Totale	8.201	100,0	324	100,0
4° trimestre 1990	N	%	N	%
Pensionati	2.416	29,1	99	46,7
Dirigenti e impiegati	1.823	22,0	50	23,6
Operai e assimilati	1.800	21,7	29	13,7
Imprenditori e liberi professionisti	313	3,8	11	5,2
Lavoratori in proprio	1.162	14,0	10	4,7
Disoccupati	121	1,5	5	2,4
Altra condizione	666	8,0	8	3,8
Totale	8.301	100,0	212	100,0

Tav. 7: Distribuzione per titolo di godimento dell'abitazione delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Di proprietà	5.749	70,1	200	61,7
In affitto	2.067	25,2	94	29,0
Altro titolo	385	4,7	30	9,3
Totale	8.201	100,0	324	100,0
4° trimestre 1990	N	%	N	%
Di proprietà	5.653	68,1	130	61,3
In affitto	2.233	26,9	71	33.5
Altro titolo	415	5,0	11	5,2
Totale	8.301	100,0	212	100,0

Tav. 8: Distribuzione per classi d'età del capofamiglia delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Fino a 25	94	1,1	2	0,6
da 26 a 30	508	6,2	25	7,7
da 31 a 35	752	9,2	27	8,3
da 36 a 40	818	10,0	33	10,2
da 41 a 45	782	9,5	24	7,4
da 46 a 50	857	10,4	37	11,4
da 51 a 55	857	10,4	24	7,4
da 56 a 60	845	10,3	26	8,0
da 61 a 65	802	9,8	34	10,5
oltre 65	1.886	23,0	92	28,4
totale	8.201	100,0	324	100,0
4° trimestre 1990	Campione		Non rispondenti	
Fino a 25	148	1,8	4	1,9
da 26 a 30	508	6,1	15	7,1
da 31 a 35	771	9,3	16	7,5
da 36 a 40	818	9,9	24	11,3
da 41 a 45	874	10,5	15	7,1
da 46 a 50	834	10,0	19	9,0
da 51 a 55	827	10,0	17	8,0
da 56 a 60	822	9,9	24	11,3
da 61 a 65	803	9,7	20	9,4
oltre 65	1.896	22,8	58	27,4
totale	8.301	100,0	212	100,0

Tav. 9: Distribuzione per sesso del capofamiglia delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Maschio	6576	80,2	256	79,0
Femmina	1625	19,8	68	21,0
Totale	8201	100,0	324	100,0
4° trimestre 1990	N	%	N	%
Maschio	6686	80,5	158	74,5
Femmina	1615	19,5	54	25,5
Totale	8301	100,0	212	100,0

Tav. 10: Distribuzione per mese delle famiglie-campione e delle famiglie non rispondenti

	Campione		Non rispondenti	
3° trimestre 1990	N	%	N	%
Luglio	2.599	31,7	238	31,1
Agosto	2.639	32,2	289	37,8
Settembre	2.963	36,1	238	31,1
Totale	8.201	100,0	765	100,0
4° trimestre 1990	N	%	N	%
Ottobre	2.492	30,0	205	31,1
Novembre	2.555	30,8	241	36,6
Dicembre	3.254	39,2	213	32,3
Totale	8.301	100,0	659	100,0