Uno stimatore ottimale in presenza di non risposte

Roberto Gismondi (*)

(*) ISTAT - Servizio SCO

RIASSUNTO

Nell'ambito di un'indagine campionaria, finalizzata alla stima di un ammontare medio od assoluto di una variabile quantitativa, il verificarsi di mancate risposte, oltre a ridurre la dimensione del campione effettivo, può comportare anche significative distorsioni nel processo di stima. I problemi che conseguentemente si pongono sono i seguenti: a) decidere se stimare o meno le mancate risposte; b) nel caso si optasse per la stima delle mancate risposte, decidere quale tecnica utilizzare. In questo lavoro si propone una tecnica di stima in due fasi: nella prima fase si stimano le mancate risposte individuali sulla base di un qualunque metodo di stima probabilistico. Nella seconda si determinano i pesi ottimali da attribuire a ciascuna delle unità oggetto di stima, generalmente diversi dai pesi originari derivati dal disegno campionario o dal particolare modello di superpopolazione generatore dei dati. Oltre alla tecnica di determinazione dei pesi "ottimale" – basata sulla minimizzazione del quadrato della differenza tra ammontare stimatore ed ammontare vero con riferimento alle unità non rispondenti – si propone anche una tecnica "sub-ottimale", che peraltro nell'applicazione empirica proposta – relativa all'indagine mensile sulle vendite al dettaglio condotta correntemente dall'ISTAT – si è rivelata più precisa delle tecnica ottimale.

ABSTRACT

An Optimal Estimator when Dealing with Non-response

In the frame of a sample survey, aimed at estimating a mean or an amount concerning a quantitative variable *y*, the presence of non-responses reduces the effective sample size and can cause relevant biases in the estimation process. Consequent problems are: a) to decide if missing answers must be imputed or not; b) if missing answers must be imputed, to decide which estimation technique has to be used. In this paper we propose a two-steps technique: in the first non-responses are estimated on the basis of a whatever estimation technique. In the second we obtain optimal weights to be assigned to each unit object of estimation, generally different from the original ones derived from the sampling design or from the superpopulation model underlying data. In addition to the "optimal" technique – based on the minimisation of the squared difference between true and estimated weighted amounts related to non-respondents – we propose a "sub-optimal" technique as well. In particular, in the empirical attempt, concerned with the monthly retail trade survey currently carried out by ISTAT, this sub-optimal technique performed better than the optimal one.

1. Premessa[1]

Si supponga di operare nel contesto di un'indagine campionaria finalizzata al calcolo dell'ammontare medio di una certa variabile quantitativa *y* – che per semplicità si supporrà non negativa – e di aver registrato n_0 non risposte. In pratica, i problemi che si presentano sono i seguenti:

a) decidere se stimare o meno le non risposte;
b) se si opta per la stima delle non risposte, decidere quale tecnica di stima utilizzare;
c) verificare se l'utilizzo degli n_0 valori imputati congiuntamente a quello degli $(n-n_0)$ valori noti possa o meno avvenire utilizzando i pesi originari attribuiti alle unità non rispondenti – derivanti dal disegno campionario o dal modello di superpopolazione generatore dei dati – oppure utilizzando pesi modificati, al fine di attenuare la variabilità addizionale indotta dal meccanismo d'imputazione.

Con riferimento al primo punto, come ampiamente documentato in Quintano e Castellano (2001), le principali indagini su famiglie ed imprese condotte dall'ISTAT possono basarsi su tecniche di stima basate sulla ponderazione vincolata[2] – che non richiede la stima di mancate risposte – oppure su procedure alternative che presuppongono quasi sempre tale stima.

Riguardo al secondo aspetto, nella pratica si ricorre ad una vasta gamma di metodologie, derivate dal particolare contesto d'indagine.

Più in generale, seguendo Lucev (1997) si può notare come tutte le procedure d'imputazione si basino su una funzione di regressione del tipo $y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon_i$, in quanto il problema dell'imputazione presenta chiari punti di contatto con quello della previsione e dell'interpolazione statistica, consistendo nello stimare il valore della caratteristica *y* relativa all'unità *i-ma* a partire dalla conoscenza del vettore **x** contenente *k* informazioni ausiliarie x_{hi}. Sotto l'ipotesi di normalità, $\hat{y}_i = f(x_{1i}, x_{2i}, ..., x_{ki})$ rappresenta il miglior previsore non distorto di *y*, condizionatamente all'informazione contenuta in **x**.

In relazione al valore assegnato al residuo ε si distinguono i metodi deterministici da quelli stocastici. I primi attribuiscono un residuo nullo ad ogni non rispondente, facendo derivare il valore imputato direttamente dalla relazione sistematica tra *y* e **x**; i secondi assegnano residui non nulli ed incorporano un elemento d'incertezza, dato dal termine erratico che cerca di compensare le discrepanze tra modello e realtà.

In realtà, come si vedrà nel prosieguo, l'uso di un meccanismo d'imputazione deterministico rappresenta un caso particolare di quello probabilistico, sebbene questa ipotesi conduca a soluzioni banali, ossia alla conferma del ricorso ai pesi originari anche per le unità oggetto d'imputazione.

Riguardo al terzo punto, supponendo di aver stimato le mancate risposte con un meccanismo probabilistico, si pone il problema dei pesi da assegnare alle unità oggetto d'imputazione. L'idea di fondo, che sarà sviluppata tecnicamente nei prossimi paragrafi, è che è possibile determinare nuovi pesi, diversi da quelli originari, il cui utilizzo potrebbe condurre a stime finali più efficienti di quelle ottenibili continuando ad utilizzare i pesi originari. In particolare, con riferimento alla valutazione dell'efficienza si ricorda che in un contesto in cui si decide di stimare le non risposte con un metodo probabilistico si possa fare riferimento a tre tipi di aleatorietà:

---

[1] Le opinioni espresse nell'articolo non impegnano l'ISTAT, ma vanno attribuite all'autore, che è anche l'unico responsabile di eventuali errori ed omissioni.

[2] Si veda in proposito Rizzo, Kalton e Brick (1996).

1. quella derivata dal meccanismo di imputazione;
2. quella relativa al disegno campionario utilizzato;
3. quella relativa al modello di superpopolazione sottostante i dati.

In pratica, gli sviluppi analitici faranno riferimento solo al primo tipo di struttura stocastica, per cui ulteriori valutazioni di efficienza dovranno basarsi sul ricorso ai consueti strumenti di stima della varianza campionaria e della distorsione indotta dal disegno campionario.

Il lavoro è così articolato: nel paragrafo seguente si definiranno la struttura formale dello stimatore e del meccanismo di stima delle mancate risposte. Nel paragrafo 3 sarà ricavato lo stimatore ottimale, basato sulla minimizzazione del quadrato della differenza tra l'ammontare medio vero e stimato della variabile *y* oggetto d'interesse con riferimento alle sole unità non rispondenti. Nel paragrafo 4 sarà ricavato uno stimatore alternativo sub-ottimale, mentre il paragrafo 5 è dedicato ad alcuni aspetti operativi. Infine, il paragrafo 6 contiene un'applicazione all'indagine sulle vendite al dettaglio condotta correntemente dall'ISTAT, in cui sono state utilizzate due tecniche di imputazione stocastica: mediante regressione casuale e mediante scelta casuale di un donatore all'interno di una data classe d'imputazione.

2. La struttura formale della strategia di stima delle mancate risposte

In generale, lo stimatore di una ammontare medio incognito $\overline{Y}$ basato su un campione teorico di *n* osservazioni si può scrivere nella forma seguente:

$$T = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n_0} w_i y_i + \sum_{i=n_0+1}^{n} w_i y_i = T_0 + T_1 \tag{2.1}$$

dove si è supposto di aver registrato n_0 mancate risposte, posizionate – per semplicità e senza perdita di generalità – in corrispondenza delle prime n_0 unità campionarie e dove $w_i = 1/N\pi_i$ è il peso campionario della unità *i-ma*, con π_i pari alla probabilità di inclusione di-tale unità nel campione.

In corrispondenza della *i-ma* mancata risposta si suppone, inoltre, di aver stimato il valore ignoto y_i con il valore $\hat{y}_i$, generato da una procedura di imputazione probabilistica caratterizzata da questa struttura stocastica:

$$\hat{y}_i = y_i + \varepsilon_i \tag{2.2}$$

dove si può ragionevolmente supporre che valgano queste relazioni:

$$\begin{cases} E(\varepsilon_i) = 0 \\ VAR(\varepsilon_i) = \sigma_i^2 \\ COV(\varepsilon_i;\varepsilon_j) = 0 \quad \text{se} \quad i \neq j \end{cases} \tag{2.3}$$

Si noti come nelle espressioni della (2.2) i simboli di speranza matematica, varianza e covarianza si riferiscano al *meccanismo di stima delle mancate risposte*, e non al disegno campionario utilizzato o al modello di superpopolazione[3] generatore delle realizzazioni di *y*. Si suppone,

[3] Per ulteriori dettagli si rimanda a Brewer (1998).

inoltre, di stimare l'ammontare incognito $\sum_{i=1}^{n_0} w_i y_i$ con l'espressione:

$$\hat{T}_0 = \sum_{i=1}^{n_0} \hat{w}_i \hat{y}_i \tag{2.4}$$

basata sui valori stimati e su dei nuovi pesi $\hat{w}_i$ da determinare secondo un criterio di ottimalità che sarà discusso nel paragrafo 3. In tale ottica, una condizione che sarà imposta è che valga la condizione:

$$\sum_{i=1}^{n_0} \hat{w}_i = \sum_{i=1}^{n_0} w_i = \mathrm{W} \tag{2.5}$$

ossia che la somma dei nuovi pesi assegnati alle unità non rispondenti sia uguale alla somma dei pesi iniziali[4]. Ad esempio, nel caso di un disegno casuale semplice varrebbe la condizione $\mathrm{W} = n_0/n$. In generale lo stimatore $\hat{T} = \hat{T}_0 + T_1$ è distorto sia rispetto al disegno campionario, sia rispetto ad un generico modello di superpopolazione. Il valore atteso di $\hat{T}$ rispetto alla tecnica di imputazione è dato da:

$$E(\hat{T}) = \sum_{i=1}^{n} w_i y_i + \sum_{i=1}^{n_0} (\hat{w}_i - w_i) y_i \,. \tag{2.6}$$

Di conseguenza la media rispetto al disegno, indicabile con E_d, sarà data, ricordando la (2.1) e la (2.2), dalla relazione:

$$E_d[E(\hat{T})] = \overline{Y} + \sum_{i=1}^{n_0} E_d(\hat{w}_i - w_i) y_i \tag{2.7}$$

dove il secondo addendo indica la distorsione dello stimatore $E(\hat{T})$. Se invece si supponesse un modello di superpopolazione tale che:

$$E_m(y_i) = \mu \quad \text{per ogni } i \tag{2.8}$$

lo stimatore $E(\hat{T})$ sarebbe corretto rispetto al modello, dato che:

$$E_m[E(\hat{T})] = E_m(\overline{Y}) + \sum_{i=1}^{n_0} (\hat{w}_i - w_i) E_m(y_i) = \mu + \mu \sum_{i=1}^{n_0} (\hat{w}_i - w_i) = \mu \tag{2.9}$$

dove si è tenuto conto della (2.5), con l'ulteriore ipotesi che i pesi $\hat{w}_i$ non dipendano dalle realizzazioni (vere o stimate) di y. Quest'ultima condizione potrebbe peraltro non essere soddisfatta, come si vedrà nei paragrafi seguenti, in cui saranno ricavate due soluzioni ottimali entrambe dipendenti dalle osservazioni incognite y_i

[4] Tale condizione consente di imporre un vincolo logico che consenta di legare i nuovi pesi a quelli iniziali. Secondo una logica almeno in parte analoga Ballin, Falorsi e Russo (2001) suggeriscono la ricerca di nuovi pesi basata sulla minimizzazione della loro distanza media rispetto ai pesi iniziali, con il vincolo, maggiormente restrittivo, che le stime finali basate sui nuovi pesi soddisfino vincoli di coerenza relativi ad una o più variabili ausiliarie misurabili nel campione.

3. La selezione ottimale dei pesi individuali

L'errore quadratico medio dello stimatore (2.4), sulla base delle ipotesi (2.2), sarà dato da[5]:

$$\Psi(\hat{w}_i) = E\left[\left(\sum_{i=1}^{n_0} \hat{w}_i \hat{y}_i - \sum_{i=1}^{n_0} w_i y_i\right)^2\right] = E\left[\left(\sum_{i=1}^{n_0} \hat{w}_i \varepsilon_i + \sum_{i=1}^{n_0} (\hat{w}_i - w_i) y_i\right)^2\right] =$$

$$= \sum_{i=1}^{n_0} \hat{w}_i^2 \sigma_i^2 + \left[\sum_{i=1}^{n_0} (\hat{w}_i - w_i) y_i\right]^2. \qquad (3.1)$$

Dalla (3.1) si deduce come l'uso dei pesi iniziali w_i renderebbe nullo il secondo addendo, ma comporterebbe un valore tanto più elevato del primo addendo quanto più i pesi più elevati si associano alle unità non rispondenti caratterizzate dalle varianze più alte, come si verifica non di rado nella pratica corrente. Ad esempio, una situazione del genere si verifica quando le non risposte riguardano unità con diversi livelli di magnitudine della variabile *y*, la cui varianza (e quindi, ragionevolmente, anche la varianza delle relative stime) cresce più che proporzionalmente al crescere della dimensione

Se, dunque, si impone il vincolo che la somma dei pesi stimati riproduca la somma dei pesi iniziali, occorre determinare gli n_0 valori ottimali di $\hat{w}_i$ ed il valore di λ in modo che sia minimizzata la seguente funzione di Lagrange:

$$\Psi(\hat{w}_i) + \lambda\left(\sum_{i=1}^{n_0} \hat{w}_i - \mathrm{W}\right). \qquad (3.2)$$

Derivando la funzione (3.2) rispetto a n_0 e λ si ottiene, dopo alcuni passaggi, la soluzione ottimale[6]:

$$\hat{w}_i^* = \left(\frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n_0} \frac{1}{\sigma_i^2}}\right)\left[\mathrm{W} - \mathrm{A}\sum_{j=1}^{n_0} \frac{(y_i - y_j)}{\sigma_j^2}\right] \qquad (3.3)$$

dove si ha:

---

[5] L'ulteriore valutazione dell'errore quadratico medio dello stimatore complessivo $\hat{T}$ risulterebbe più laboriosa, anche qualora si considerasse il valore atteso – rispetto al disegno o al modello – dello scarto $\left[E(\hat{T}) - \bar{Y}\right]^2$ anziché di $\left[\hat{T} - \bar{Y}\right]^2$.

[6] Questa soluzione rappresenta una generalizzazione di un risultato ottenuto da Lettau e Loewenstein (2000) con riferimento alla stima campionaria di numeri indici. Per la dimostrazione si rimanda all'appendice.

$$A = \frac{W\sum_{i=1}^{n_0}\frac{y_i}{\sigma_i^2} - \sum_{i=1}^{n_0} w_i y_i \sum_{i=1}^{n_0}\frac{1}{\sigma_i^2}}{\left(\sum_{i=1}^{n_0}\frac{1}{\sigma_i^2}\right)\left[1+\sum_{i=1}^{n_0}\left(\frac{y_i}{\sigma_i}\right)^2\right] - \sum_{i=1}^{n_0}\left(\frac{y_i}{\sigma_i}\right)^2}. \qquad (3.4)$$

L'equazione (3.3) indica che il grado di ottimalità implicito nella sovrastima (o sottostima) dei pesi di strato dipende in modo cruciale dalle differenze $(y_i - y_j)$ e dalle varianze σ_j^2. In generale, più piccola è la differenza tra il valore y_i da stimare ed i rimanenti $(n_0 - 1)$ valori[7], più il nuovo peso tenderà alla quantità:

$$\hat{w}_i^* = \left(\frac{\frac{W}{\sigma_i^2}}{\sum_{i=1}^{n_0}\frac{1}{\sigma_i^2}} \right) \qquad (3.5)$$

per cui il peso di ogni strato risulterebbe inversamente proporzionale alla varianza di strato. A tale soluzione semplificata si perviene, più in generale, se si verifica il caso limite in cui tutte i valori y_i fossero uguali. Più in generale, quanto più y_i è elevato (in confronto alle restanti $(n_0 - 1)$ osservazioni), tanto più il peso ottimale decresce, e viceversa.

Qualora le varianze fossero tutte uguali, non si perverrebbe ad una soluzione ottimale particolarmente semplificata, a meno che non si supponga un disegno casuale semplice, con w_i=1/n - ipotesi peraltro ragionevole nel caso di varianze uguali - nel qual caso si avrebbe nuovamente la relazione (3.5), che equivarrebbe all'uguaglianza $\hat{w}_i = w_i = 1/n$.

La formula della media quadratica dell'errore imputabile al meccanismo di imputazione è ottenibile sostituendo la (3.3) nella (3.1). Per pervenire ad una stima di tale media si possono utilizzare le stime $\hat{y}_i$ al posto dei valori incogniti y_i.

Questo approccio utilizza pesi ottimali - rispetto alla procedura di stima delle mancate risposte utilizzata - determinati sulla base della conoscenza preliminare di segnali di intensità per le singole unità non rispondenti, sintetizzati nelle rispettive varianze. Come tale, potrebbe però risultare parzialmente tautologico, e produrre nuovi pesi anche molto diversi da quelli iniziali, conseguenza poco desiderabile in un contesto operativo dove la principale valutazione della precisione è effettuata in funzione del disegno campionario, o di un modello di superpopolazione sulla cui base lo stimatore corretto della media incognita implicherebbe il ricorso ai pesi originari w_i. Ciò nonostante, il principio di base che lo ispira è il miglior sfruttamento possibile delle informazioni disponibili sulle unità non rispondenti, come sarà meglio descritto nel paragrafo 5.

In ogni caso, sia la soluzione ottimale determinata sopra, sia quella descritta nel paragrafo seguente propongono il ricorso a stimatori in cui si opta per una stima delle mancate risposte, la cui efficienza dovrà sempre essere confrontata con quella di altri stimatori più tradizionali e, in particolare, con lo stimatore basato sulla semplice media dei rispondenti[8]. In particolare, si ribadisce come l'ottimalità dello stimatore è stata valutata in funzione del solo meccanismo di stima delle mancate risposte e non, ad esempio, sulla base della aleatorietà campionaria, per cui solo il riscontro empirico può fornire indicazioni più precise circa la convenienza ad utilizzare

[7] Questa situazione si verifica quando la *i-ma* osservazione da stimare è approssimativamente pari alla media delle $(n_0 - 1)$ osservazioni rimanenti.

[8] Si noti come tale stimatore sia ottenibile dalla (2.1) ponendo pari a zero il peso di ogni unità non rispondente e pari a $1/(n - n_0)$ quello di ogni unità rispondente.

una simile tipologia di stimatori.

4. Una procedura alternativa

Nella (3.1) si è imposto di selezionare i pesi in modo da rendere minimo il valore atteso del quadrato dello scarto tra T_0 e la sua stima $\hat{T}_0$. Una condizione meno restrittiva consiste nel ricercare le stime dei pesi che rendono minimo il valore atteso della somma dei quadrati degli n_0 scarti tra ogni valore $w_i y_i$ e la sua stima $\hat{w}_i \hat{y}_i$. In simboli, la funzione di Lagrange da minimizzare è data da:

$$\Phi(\hat{w}_i, \lambda) = \sum_{i=1}^{n_0} E(\hat{w}_i \hat{y}_i - w_i y_i)^2 + \lambda \sum_{i=1}^{n_0} (\hat{w}_i - \mathrm{W}) \tag{4.1}$$

che, dopo alcuni passaggi, si riduce alla funzione:

$$\Phi(\hat{w}_i, \lambda) = \sum_{i=1}^{n_0} \hat{w}_i^2 \sigma_i^2 + \sum_{i=1}^{n_0} (\hat{w}_i - w_i)^2 y_i^2 + \lambda \sum_{i=1}^{n_0} (\hat{w}_i - \mathrm{W}) .$$

Derivando rispetto a $\hat{w}_i$ si ottiene la soluzione ottimale:

$$\hat{w}_i^{**} = w_i a_i + b_i \left(\frac{\mathrm{W} - \sum_{i=1}^{n_0} w_i a_i}{\sum_{i=1}^{n_0} b_i} \right) \tag{4.2}$$

dove si ha:

$$a_i = \frac{y_i^2}{y_i^2 + \sigma_i^2} \qquad b_i = \frac{1}{y_i^2 + \sigma_i^2} \tag{4.3}$$

La soluzione ottimale (4.2) gode di queste proprietà.

1. Garantisce, per costruzione, che sia minimizzata la media degli scarti tra ciascuna coppia di valori ponderati y_i veri e stimati. Tale aspetto è particolarmente rilevante quando l'indagine in oggetto è finalizzata, oltre alla produzione, anche al rilascio di microdati, sebbene parzialmente stimati a causa delle non risposte, di cui è ovviamente richiesto un accettabile livello di qualità.
2. E' più semplice, da un punto di vista computazionale, della soluzione ottimale (3.3).
3. Rende minimo il primo addendo della media quadratica dell'errore definita nel paragrafo precedente. In effetti, tale media si può scrivere come:

$$E\left[\left(\sum_{i=1}^{n_0} (\hat{w}_i \hat{y}_i - w_i y_i) \right)^2 \right] = E\left[\sum_{i=1}^{n_0} (\hat{w}_i \hat{y}_i - w_i y_i)^2 \right] + E\left[\sum_{i=1}^{n_0} \sum_{j \neq i=1}^{n_0} (\hat{w}_i \hat{y}_i - w_i y_i)(\hat{w}_j \hat{y}_j - w_j y_j) \right]$$

ed il primo addendo del membro di destra della precedente identità corrisponde al primo

addendo della funzione di Lagrange (4.1). Se l'evidenza empirica dimostrasse che tale addendo assume un peso preponderante nell'errore quadratico medio complessivo, la soluzione sub-ottimale (4.2) potrebbe essere utilizzata in alternativa (o parallelamente) alla soluzione ottimale (3.3).

4. Da un punto di vista concettuale, i pesi sub-ottimali (4.2) dovrebbero sempre risultare mediamente più simili ai corrispondenti pesi iniziali, come evidenzia il primo addendo della (4.2). I pesi ottimali (3.3) non dipendono da tali pesi iniziali, se non per il fatto che essi stessi devono riprodurre, se sommati, il valore W. Come già ricordato, tale aspetto potrebbe sconsigliare il ricorso ai pesi ottimali in quanto, sebbene risultanti da una ottimizzazione matematica, potrebbero condurre a stime non sempre realistiche a causa della possibile sovrastima dei pesi assegnati ad unità caratterizzate da una forte variabilità individuale nei valori di *y*.
5. Si può presentare, in pratica, un caso di particolare interesse nel momento in cui si avesse $\sigma_i^2 \cong cv^2 y_i^2$, ossia in cui il coefficiente di variazione *cv* fosse approssimativamente costante al variare delle unità. Se si pone $c = 1/(1+cv^2)$ è facile verificare che la relazione (4.2) si riduce, dopo alcuni passaggi, alla:

$$\hat{w}_i^{**} = c\, w_i + (1-c)\left(\frac{\frac{1}{y_i^2}}{\sum_{i=1}^{n_0} \frac{1}{y_i^2}} \right) \qquad (4.3)$$

ossia alla media aritmetica ponderata tra il peso originario ed un termine inversamente proporzionale all'intensità dell'ammontare del carattere sull'unità *i-ma*, dove il coefficiente di ponderazione *c* assegnato al peso originario è a sua volta inversamente proporzionale alla variabilità relativa di *y*. Questa semplificazione non si verificherebbe nel caso della formula ottimale (3.3). Peraltro la soluzione sub-ottimale (3.2) non si riduce a forme particolarmente semplici nemmeno nei casi in cui i valori y_i o le relative varianze, fossero costanti al variare delle unità, circostanza peraltro piuttosto improbabile in pratica.

Una considerazione di carattere empirico deriva dall'osservare come, in presenza di molte unità non rispondenti e/o in una situazione operativa in cui non sia possibile discostarsi eccessivamente dai pesi originari, o nel caso in cui informazioni dettagliate su medie e varianze siano disponibili solo con riferimento a poche unità, il valore della media quadratica dell'errore potrebbe essere fortemente influenzato solo da pochi addendi, ossia dalle stime relative a poche unità non rispondenti.

Dalla espressione della media quadratica dell'errore riportata nella (3.1) il contributo di ogni unità oggetto di stima all'errore quadratico complessivo è dato dalla espressione (approssimata per quanto riguarda il secondo addendo):

$$\hat{w}_i^2 \sigma_i^2 + (\hat{w}_i - w_i)^2 y_i^2 \qquad (4.4)$$

in cui il primo addendo esprime la quota della MQE dovuta alla variabilità della stima ed il secondo addendo dipende invece dalla differenza tra pesi ottimali e pesi iniziali. Di conseguenza, una strategia operativa può consistere nei seguenti passi.

1. Raccolta delle risposte relative ad un'indagine basata su un dato disegno campionario e, quindi, su un dato stimatore del tipo (2.1).

2. Selezione ed implementazione di *k* tecniche probabilistiche per la stima di mancate risposte.
3. Calcolo degli stimatori (3.3) e/o (4.2) per ciascuna delle procedure di imputazione adottate.
4. Calcolo della relazione (4.4) per ogni unità non rispondente e ciascuno dei casi sviluppati in funzione del precedente punto 3.
5. Per ciascuna delle procedure del punto 3, ordinamento decrescente dei valori della (4.4).
6. Per ciascuna delle procedure del punto 3, si calcola la media quadratica dell'errore considerando solo la prima unità oggetto d'imputazione nell'ordinamento di cui al precedente punto 5, poi solo le prime due unità, e così via.
7. La strategia migliore, ossia quella che identifica la tecnica d'imputazione e, al contempo, il numero di unità campionarie non rispondenti da considerare per la stima, è quella a cui corrisponde il valore minimo osservato al passo descritto nel punto 6.

In pratica, l'importanza della valutazione dei pesi ottimali o sub-ottimali sta anche nella possibilità di identificare le unità in corrispondenza delle quali occorre supportare gli sforzi maggiori per ottenere le risposte, in quanto ad esse sono generalmente associate le varianze di stima delle mancate risposte più elevate.

5. L'implementazione empirica degli stimatori ed il ricorso a dati storici

Il principale problema insito nella implementazione degli stimatori introdotti nei paragrafi 3 e 4 riguarda, ovviamente, la non conoscenza dei valori veri y_i e delle varianze individuali oggetto di stima. In merito, le soluzioni operative più semplici sono le seguenti:

1. Sostituire ai valori incogniti le rispettive stime $\hat{y}_i$. Una strategia analoga può essere seguita per le varianze, per la cui stima è necessario disporre di una serie storica di osservazioni di *y* per ogni unità non rispondente, o per unità rispondenti "simili" ad essa (ad esempio, appartenenti allo stesso strato). Di conseguenza, le condizioni di ottimalità suddette si devono continuare a ritenere valide in chiave teorica, ma non effettiva, dovendosi tenere conto della struttura aleatoria di tali stime. Ovviamente la qualità delle stime rimarrà elevata nei casi in cui il criterio di stima delle mancate risposte risulti sufficientemente preciso.
2. Sostituire ai suddetti valori incogniti i rispettivi valori, supposti noti, relativi ad un periodo di rilevazione precedente a quello di interesse.
3. Sostituire ai suddetti valori incogniti i rispettivi valori, supposti noti e riferiti al medesimo tempo di riferimento, relativi ad una variabile nota per tutte le unità non rispondenti e sufficientemente correlata a *y*.

Ad esempio, nelle indagini ripetute nel tempo condotte su imprese, non è raro disporre di dati su variabili quantitative per più periodi di osservazione anche molto ravvicinati (a volte relativi a mesi consecutivi): è il caso delle indagini mensili sulla produzione industriale e sulle vendite al dettaglio, condotte correntemente dall'ISTAT, che rilevano il fatturato mensile su panels di imprese. La *wave non-response* che caratterizza la seconda delle indagini suddette non consente di disporre per ogni mese di tutte le risposte, sebbene per circa l'80% delle imprese sia sempre disponibile il valore del fatturato relativo allo stesso mese dell'anno precedente, che consentirebbe il ricorso alla soluzione operativa 2.

D'altra parte, per tutte le imprese italiane risulta anche disponibile il fatturato annuo[9], particolarmente utile per implementare la soluzione operativa 3 qualora la variabile *y* sia il

[9] Il fatturato annuo d'impresa è disponibile nell'archivio ASIA dell'ISTAT (ASIA è l'acronimo per "Archivio Statistico delle Imprese Attive"). Tale informazione, attualmente aggiornata all'anno 1999, deriva dalle dichiarazioni fiscali messe a disposizione dal Ministero delle Finanze.

fatturato annuale od infraannuale.

La disponibilità di dati storici è in realtà sfruttabile anche con un criterio di ottimalità lievemente diverso da quello descritto nel paragrafo 3.

Se risultasse noto il valore vero relativo ad un periodo "0" precedente al periodo *t* oggetto d'indagine, potrebbe essere richiesto che i pesi da utilizzare al tempo *t* siano tali da rendere minima la somma dei quadrati degli scarti tra T_0 e la sua stima $\hat{T}_0$, dove entrambe le quantità si riferiscono al tempo 0, con un vincolo analogo alla (2.5).

Si tratta di un'impostazione piuttosto simile a quella vista nel paragrafo 2, sebbene in questo caso non si coinvolga nessuna ipotesi stocastica e ci si basi, in sostanza, sulla ricerca di una soluzione ottimale di tipo deterministico che si suppone trasferibile al tempo *t* d'interesse.

Se si desidera rendere minimo il quadrato dello scarto tra lo stimatore e l'ammontare incognito oggetto di stima, si può definire la funzione di Lagrange:

$$\Phi(\hat{w}_i, \lambda) = \left[\sum_{i=1}^{n_0} (\hat{w}_i \hat{y}_{i0} - w_i y_{i0}) \right]^2 + \lambda \sum_{i=1}^{n_0} (\hat{w}_i - \mathrm{W}) \tag{5.1}$$

da cui è possibile ricavare la soluzione ottimale data da:

$$\hat{w}_i^* = \left(\frac{\frac{\mathrm{W}}{\hat{y}_{i0}^2}}{\sum_{i=1}^{n_0} \frac{1}{\hat{y}_{0i}^2}} \right) \tag{5.2}$$

Anche in questo caso si potrebbe optare per una soluzione sub-ottimale basata sulla minimizzazione della somma degli scarti al quadrato tra ogni stima ponderata ed il rispettivo valore vero ponderato, considerando quindi la funzione:

$$\Phi(\hat{w}_i, \lambda) = \sum_{i=1}^{n_0} (\hat{w}_i \hat{y}_{i0} - w_i y_{i0})^2 + \lambda \sum_{i=1}^{n_0} (\hat{w}_i - \mathrm{W}) \tag{5.3}$$

con la conseguente soluzione sub-ottimale:

$$\hat{w}_i^{**} = \frac{y_{i0} w_i}{\hat{y}_{i0}} + \frac{1}{\hat{y}_{i0}^2} \left(\frac{\mathrm{W} - \sum_{i=1}^{n_0} \frac{y_{i0} w_i}{\hat{y}_{i0}}}{\sum_{i=1}^{n_0} \frac{1}{\hat{y}_{i0}^2}} \right). \tag{5.4}$$

Tale soluzione equivale alla soluzione ottimale (4.2) ponendo tutte e varianze pari a zero e sostituendo ai valori di *y* al denominatore le rispettive stime.

6. Una applicazione

L'ISTAT rileva ogni mese l'ammontare delle vendite al dettaglio presso un campione di circa 8.000 imprese commerciali al dettaglio in sede fissa, e sulla base di tali informazioni elabora e diffonde una serie di indici che esprimono la variazione delle vendite intercorsa tra un

certo mese e la media mensile dell'anno base 1995[10].

I dati che ogni mese si rendono via via disponibili sono utilizzabili anche per la stima del fatturato medio del comparto commerciale al dettaglio, su base mensile od annuale.

Al termine della fase di raccolta dei dati, che normalmente avviene dopo circa 3 mesi dalla fine del mese di riferimento, il tasso medio di non risposta è di circa il 30%.

In questa applicazione è stata considerata la base dati disponibile con riferimento agli anni 1999 e 1998. Sebbene la stratificazione originaria adottata per l'indagine sulle vendite preveda circa 170 domini per i quali ogni mese è calcolato un indice distinto, in questo contesto è stata adottata una stratificazione semplificata, basata sui 20 domini definiti nella tabella 6.1

Sostanzialmente sono state considerate due tipologie di imprese commerciali al dettaglio: le imprese specializzate (ossia quelle che vendono esclusivamente o in prevalenza una sola tipologia di prodotti) e quelle non specializzate, a loro volta distinte in base al fatto che la tipologia dei prodotti venduti esclusivamente o in prevalenza sia di tipo alimentare o non alimentare[11]. L'ulteriore elemento di stratificazione è dato dalla dimensione aziendale, misurata sulla base delle classi di addetti 1-2, 3-5, 6-9, 10-19 e da 20 in poi.

Tabella 6.1 – Numero di unità del panel, valori medi veri del fatturato annuale e coefficienti di variazione del fatturato (valori in milioni di lire)

Attività prevalente	Addetti	Numero di unità		Fatturato medio		C.V.
		Rispondenti	Non rispondenti	Rispondenti	Non rispondenti	
Totale commercio al dettaglio		5.505	2.363	1.019	1.044	2,85
Specializzati alimentari		772	333	421	449	2,89
Specializzati non alimentari		3.697	1.585	566	588	2,74
Despecializzati alimentari		804	345	3.007	3.083	3,49
Despecializzati non alimentari		232	100	3.349	3.226	2,32
	1-2	2.764	1.185	47	51	3,09
	3-5	941	405	179	194	1,82
	6-9	491	210	809	756	3,10
	10-19	693	297	2.107	2.231	2,92
	>19	616	266	5.610	5.667	3,11
Specializzati alimentari	1-2	443	190	49	43	2,95
	3-5	153	66	218	281	3,98
	6-9	64	28	842	696	1,98
	10-19	62	27	895	915	1,07
	>19	50	22	3.211	3.567	2,37
Specializzati non alimentari	1-2	2.078	890	38	44	3,25
	3-5	701	301	146	151	1,00
	6-9	269	115	459	483	3,03
	10-19	380	163	704	633	2,93
	>19	269	116	5.646	5.937	2,82
Despecializzati alimentari	1-2	190	82	75	81	2,03
	3-5	62	27	256	225	5,97
	6-9	120	51	1.340	1.215	3,74
	10-19	206	88	4.345	4.986	3,98
	>19	226	97	5.891	5.673	3,45
Despecializzati non alimentari	1-2	53	23	265	280	1,84
	3-5	25	11	682	776	1,29
	6-9	38	16	1.554	1.360	3,40
	10-19	45	19	5.377	5.045	0,47
	>19	71	31	6.266	6.130	3,64

Tra le 7.868 imprese prese in esame, di cui si conosce l'ammontare annuo delle vendite per gli anni 1999 e 1998, sono state scelte a caso 2.363 imprese, di cui sono state azzerate le risposte e che nella simulazione hanno giocato il ruolo di unità non rispondenti. La scelta di una

[10] I dettagli metodologici dell'indagine sono disponibili in ISTAT (1998), mentre la rassegna dei dati diffusi più recente è data da ISTAT (2000).

[11] Esercizi despecializzati a prevalenza alimentare sono i supermercati, gli ipermercati ed i discount; a prevalenza non alimentare sono i grandi magazzini e le altre grandi superfici non alimentari.

quota di unità non rispondenti pari a circa il 30% del campione all'interno di ogni strato è in sintonia con la suddetta quota finale di non rispondenti dopo 90 giorni dalla fine del mese di riferimento dei dati[12].

Tale scelta casuale è stata replicata per 100 volte; per ciascuna delle scelte sono state iterate le procedure di stima delle mancate risposte descritte nel prosieguo, dove la variabile oggetto di stima è stata rappresentata dal fatturato annuale del 1999.

Dalla tabella 6.1 emerge come circa la metà delle imprese considerate nella simulazione abbiano non più di 2 addetti e circa i due terzi svolgano un'attività di vendita specializzata di prodotti non alimentari. La tipologia di imprese meno frequente è quella degli esercizi despecializzati non alimentari; tra questi, quelli con addetti tra 3 e 5 sono solo 11 tra i non rispondenti e 25 tra i rispondenti, sulla cui base sono state stimate le non risposte.

La variabilità media del fatturato tra impresa ed impresa è piuttosto elevata, essendo il coefficiente di variazione medio – calcolato sulle sole unità rispondenti – pari a 2,85. La variabilità più forte si verifica per gli esercizi despecializzati a prevalenza alimentare (coefficiente pari a 3,49) ed è rilevante verificare nel prosieguo se tale circostanza si traduca o meno in una maggiore imprecisione nel processo di stima delle risposte mancanti.

Per stimare le mancate risposte sono stati utilizzati due criteri probabilistici: la stima mediante regressione stocastica e mediante donatore scelto casualmente[13], che si descrivono brevemente.

Imputazione mediante regressione casuale (IR)

Per la *i-ma* unità non rispondente si utilizza la stima:

$$\hat{y}_i = \hat{\beta}_{R0} + \sum_{h=1}^{k} \hat{\beta}_{Rh} x_{ih} + \hat{\varepsilon}_i \qquad (6.1)$$

dove i (k+1) coefficienti $\hat{\beta}_{Rh}$ sono stati stimati sulla base dei dati disponibili per le unità rispondenti (da cui deriva il pedice *R*), x_{ih} è il valore assunto dalla *h-ma* variabile ausiliaria *x* sulla *i-ma* unità non rispondente ed $\hat{\varepsilon}_i$ è il residuo della regressione. I residui possono essere scelti in base a diversi criteri:

a) Si può ipotizzare una distribuzione teorica con media zero e varianza uguale alla varianza residua della regressione e da essa campionare i residui per i non rispondenti.
b) Sotto l'ipotesi MAR (risposte *Missing At Random*) i residui dei non rispondenti potrebbero distribuirsi secondo una distribuzione analoga a quella dei rispondenti e, quindi, possono essere campionati casualmente da essi.
c) Si può scegliere il residuo del rispondente che presenta valori delle variabili ausiliarie più vicini ai valori del non rispondente, cosicché se il rispondente "donatore" del residuo presenta la stessa serie di valori per le variabili *x*, di fatto, si duplica il suo valore.

Nell'ambito della simulazione si è utilizzata una sola variabile ausiliaria *x*, data dal fatturato annuo del 1998, la cui correlazione con la variabile oggetto di stima (il fatturato del 1999) è risultata superiore a 0,90 in ogni strato. I residui sono stati scelti sulla base del criterio c).

[12] I fatturati annui 1998 e 1999 delle unità che hanno risposto solo in alcuni mesi di tali anni sono stati ricavati dal suddetto archivio ASIA dell'ISTAT.
[13] Per una rassegna più ampia si rimanda a Lucev (op. cit., 1997).

Imputazione casuale (IC)

Per la *i-ma* unità non rispondente si utilizza la stima:

$$\hat{y}_{si} = y_{sj} \tag{6.2}$$

dove y_{sj} è il valore di *y* associato alla generica unità *j-ma* rispondente appartenente alla cella di stratificazione *s-ma*.

Tale metodo assegna ad ogni non rispondente *i* il valore di un donatore selezionato a caso nella classe d'imputazione a cui appartiene il non rispondente, dove le classi d'imputazione sono costruite sulla base delle variabili ausiliarie disponibili. Esso rappresenta una caso particolare del metodo IR precedente, in cui le variabili esplicative sono *dummy* connesse con le celle di classificazione[14].

Nell'ambito della simulazione le classi d'imputazione sono state determinate sulla base di cinque classi dimensionali della variabile ausiliaria *x* data dal fatturato annuo del 1998[15]. A causa dell'ampiezza delle classi, la tecnica di stima utilizzata si è basata sulla formula:

$$\hat{y}_{si} = y_{sj}\left(\frac{x_{si}}{x_{sj}}\right) \tag{6.3}$$

dove x_{si} è il numero di addetti dell'impresa *i-ma*.

La tecnica di stima utilizzata è in due fasi. Nella prima fase si stimano le mancate risposte individuali sulla base di ciascuno due metodi probabilistici appena descritti. Nella seconda si determinano i pesi ottimali da attribuire a ciascuna delle unità oggetto di stima, generalmente diversi dai pesi originari derivati dal disegno campionario o dal particolare modello di superpopolazione generatore dei dati, basati rispettivamente sulle formule 3.3 e 4.2.

Nella tabella 6.2 le stime ottenute con le due tecniche suddette sono indicate con le sigle IR e IC e saranno definite come stime "dirette", per distinguerle dalle stime "ottimali" IR* e IC* (basate sulla formula 3.3) e dalle stime "sub-ottimali" IR** e IC** (basate sulla formula 4.2).

Per ognuna delle unità non rispondenti il peso originario *w* è costante all'interno dello strato ed è dato dal reciproco del numero di unità dello strato, sia rispondenti sia non rispondenti. Per quanto riguarda la stima della varianza individuale di ogni unità non rispondente, essa è stata effettuata:

1) sulla base delle risposte fornite negli anni dal 1995 al 1999 per le unità appartenenti al panel *stabilmente* nel quinquennio suddetto (circa il 30% del campione);
2) sulla base della stima $\hat{\sigma}^2_{si} = \sigma^2_s y^2_{si}$ per le unità restanti, dove y_{si} è il fatturato del 1999 e σ^2_s è un parametro medio di strato stimato sulla base delle unità del precedente punto 1.

Nella tabella 6.3 sono riportate le medie degli scarti percentuali, in valore assoluto, tra i valori stimati ed i valori veri. Sono stati evidenziati in grassetto i casi corrispondenti allo scarto più basso per ogni dominio d'interesse, mentre i *second-best*[16] sono sottolineati. Per un'interpretazione più chiara, si ricorda che i risultati si riferiscono agli scarti percentuali medi tra valori veri e stimati calcolati come media di 100 replicazioni con scelta casuale dei non

[14] Una tecnica del donatore alternativa descritta in Abbate (1998).
[15] Le classi di fatturato sono: da 1 a 200 milioni, da 201 a 500, da 501 a 1.000, da 1.000 a 5.000, oltre 5.000.
[16] Acronimo utilizzato per indicare le seconde migliori prestazioni nell'ambito di valutazioni qualitative.

rispondenti, coerente con l'ipotesi di una struttura MAR per le non risposte.

Tabella 6.2 – Numero di unità non rispondenti, valori medi veri del fatturato annuale e valori medi stimati con diversi criteri (valori in milioni di lire)

Attività prevalente	Addetti	n	Valori veri	Valori stimati con vari criteri					
				IR	IC	IR*	IC*	IR**	IC**
Totale commercio al dettaglio		2.363	957	945	956	944	954	946	956
Specializzati alimentari		333	449	461	447	449	436	462	447
Specializzati non alimentari		1.585	588	595	592	588	586	596	593
Despecializzati alimentari		345	2.521	2.373	2.501	2.442	2.570	2.374	2.501
Despecializzati non alimentari		100	3.106	3.190	3.075	3.061	2.940	3.191	3.076
	1-2	1.185	51	57	51	56	50	57	51
	3-5	405	194	195	192	193	191	195	192
	6-9	210	756	755	749	741	735	756	749
	10-19	297	2.295	2.290	2.295	2.343	2.347	2.292	2.296
	>19	266	4.821	4.695	4.817	4.637	4.762	4.700	4.822
Specializzati alimentari	1-2	190	43	47	43	47	43	47	43
	3-5	66	281	281	279	297	295	281	279
	6-9	28	696	713	687	709	683	713	687
	10-19	27	915	922	919	769	766	927	923
	>19	22	3.567	3.693	3.550	3.655	3.529	3.696	3.555
Specializzati non alimentari	1-2	890	44	51	43	48	41	51	44
	3-5	301	151	152	150	147	144	152	150
	6-9	115	483	484	486	472	473	485	486
	10-19	163	633	649	646	676	672	650	647
	>19	116	5.937	5.954	5.984	5.863	5.904	5.964	5.994
Despecializzati alimentari	1-2	82	81	87	82	103	99	87	82
	3-5	27	225	231	215	239	225	231	215
	6-9	51	1.215	1.199	1.196	1.182	1.179	1.200	1.197
	10-19	88	4.986	4.963	4.962	5.184	5.189	4.965	4.963
	>19	97	3.673	3.169	3.635	3.208	3.668	3.169	3.635
Despecializzati non alimentari	1-2	23	280	277	275	280	279	277	276
	3-5	11	776	761	771	744	753	761	771
	6-9	16	1.360	1.364	1.323	1.331	1.291	1.364	1.324
	10-19	19	6.045	5.935	6.046	5.719	5.803	5.941	6.052
	>19	31	5.130	5.472	5.051	5.212	4.786	5.472	5.052

Nel complesso, l'imputazione casuale IC comporta stime più precise rispetto all'imputazione per regressione stocastica IR: l'errore medio è dello 0,15% nel primo caso e dell'1,20% nel secondo. Il ricorso al metodo sub-ottimale comporta buoni risultati, in quanto riduce l'errore medio di stima allo 0,06% con IC ed all'1,12% con IR. Di contro, il metodo ottimale conduce a stime mediamente più imprecise, ed il peggioramento nella qualità media della stima è particolarmente rilevante nel caso si utilizzi la tecnica IC (l'errore medio raddoppia, arrivando allo 0,32%). Tale risultato può essere attribuito a diverse cause: a) alla necessità, implicita nel metodo ottimale, di stimare un numero maggiore di parametri disponendo spesso di poche osservazioni; b) al fatto già citato che i pesi ottimali sono meno vincolati ai pesi originari rispetto ai pesi sub-ottimali; c) alla possibile mancanza di validità, nel riscontro empirico, della terza relazione della (2.3), ossia dell'ipotesi di non correlazione tra gli errori di stima tra mancate risposte diverse che, d'altra parte, non entra direttamente in gioco negli sviluppi analitici che hanno condotto alla soluzione sub-ottimale.

Considerando l'insieme dei 30 domini per i quali sono stati riportati gli errori medi di stima (10 dei quali derivati dai 20 di base, che stratificano le quattro tipologia di vendita per cinque classi di addetti), in 12 casi il metodo migliore è risultato RC**, in 7 casi RC, in 4 casi RR* e RR**, in 3 casi RR e in 2 casi RC*.

Tabella 6.3 – Numero di unità non rispondenti, valori medi veri del fatturato annuale e medie degli scarti percentuali (in valore assoluto) tra valori veri e stimati

Attività prevalente	Addetti	n	Valori veri	Scarti % tra valori veri e stimati (in valore assoluto)					
				IR	IC	IR*	IC*	IR**	IC**
Totale commercio al dettaglio		2.363	957	1,20	0,15	1,41	0,32	1,12	**0,06**
Specializzati alimentari		333	449	2,85	0,46	**0,14**	2,89	3,00	0,29
Specializzati non alimentari		1.585	588	1,21	0,79	**0,03**	0,35	1,36	0,96
Despecializzati alimentari		345	2.521	5,87	0,80	3,13	1,96	5,85	**0,78**
Despecializzati non alimentari		100	3.106	2,68	1,03	1,45	5,37	2,72	**0,98**
	1-2	1.185	51	12,45	0,32	10,62	1,94	12,54	**0,17**
	3-5	405	194	0,48	0,96	**0,23**	1,69	0,54	0,89
	6-9	210	756	0,15	0,98	1,96	2,84	**0,06**	0,89
	10-19	297	2.295	0,20	**0,00**	2,09	2,28	0,13	0,07
	>19	266	4.821	2,61	0,08	3,83	1,23	2,51	**0,03**
Specializzati alimentari	1-2	190	43	9,96	0,27	9,82	0,60	9,97	**0,23**
	3-5	66	281	0,12	0,71	5,67	5,10	**0,22**	0,57
	6-9	28	696	2,34	1,31	1,79	1,90	2,35	**1,30**
	10-19	27	915	0,82	**0,42**	15,90	16,21	1,31	0,89
	>19	22	3.567	3,51	0,48	2,46	1,06	3,62	**0,34**
Specializzati non alimentari	1-2	890	44	16,19	0,21	9,75	6,63	16,28	**0,05**
	3-5	301	151	0,79	0,68	2,80	4,46	0,84	**0,62**
	6-9	115	483	**0,11**	0,46	2,32	2,15	0,26	0,64
	10-19	163	633	2,47	**1,96**	6,74	6,14	2,58	2,09
	>19	116	5.937	**0,29**	0,79	1,23	0,56	0,45	0,97
Despecializzati alimentari	1-2	82	81	6,99	**0,44**	26,91	22,22	7,13	0,73
	3-5	27	225	2,77	4,22	6,47	**0,12**	2,78	4,21
	6-9	51	1.215	1,31	1,56	2,74	2,97	**1,24**	1,50
	10-19	88	4.986	0,47	0,50	3,96	4,05	**0,44**	0,47
	>19	97	3.673	13,70	1,02	12,64	**0,11**	13,70	1,02
Despecializzati non alimentari	1-2	23	280	1,19	1,84	**0,02**	0,40	1,06	1,67
	3-5	11	776	1,96	**0,66**	4,17	3,01	1,96	**0,66**
	6-9	16	1.360	**0,26**	2,70	2,17	5,08	0,28	2,69
	10-19	19	6.045	1,81	**0,03**	5,38	4,01	1,72	0,12
	>19	31	5.130	6,66	**1,54**	1,58	6,72	6,66	**1,54**

Considerando nella valutazione complessiva anche i *second-best*, i metodi migliori sono risultati RC** e RC con 19 primi o secondi posti, seguiti da RR con 11, RR** con 8, RR* con 5 e RC* con 3. Dunque, in sintesi si può affermare che:

- in questa particolare applicazione l'uso dei criteri ottimale e sub-ottimale non comporta miglioramenti particolarmente significativi nella precisione delle stime, poiché le stime dirette risultano esse stesse già piuttosto precise.
- L'uso dell'imputazione casuale comporta comunque risultati mediamente preferibili a quelli ottenuti utilizzando l'imputazione per regressione casuale.
- In ogni caso, il criterio sub-ottimale risulta piuttosto utile soprattutto nel caso in cui si utilizzi l'imputazione casuale, che comporta riduzioni significative dell'errore medio di stima per l'insieme di tutte le imprese e per le imprese fino a 2 e con oltre 19 addetti; d'altra parte, qualora si ricorra all'imputazione casuale l'uso del metodo ottimale risulta di scarsa utilità.
- Il metodo ottimale è invece utile qualora si utilizzi l'imputazione per regressione stocastica, con una riduzione dell'errore medio di stima molto significativa per gli esercizi specializzati che, nel campione come nell'intero universo delle imprese commerciali al dettaglio, rappresenta la quota più consistente delle imprese[17].

[17] Alla fine del 1999 il comparto commerciale al dettaglio era composto da circa 690.000 imprese, di cui il 67% specializzate.

Nella tabella 6.4 sono riportate le stime delle medie quadratiche dell'errore relative alle sei famiglie di stime disponibili.

E' evidente la sostanziale discordanza tra gli errori medi attesi e quelli effettivamente realizzatisi appena commentati: come era lecito attendersi, nel complesso il metodo ottimale comporta il maggior numero di casi con la più bassa stima dell'errore quadratico medio (16 primi posti e 6 *second-best* con RR*, 6 primi posti e 8 *second-best* con RC*). Il metodo sub-ottimale comporta invece 9 primi posti e 4 *second-best* con RR**, 2 primi posti e 3 *second-best* con RC**.

In realtà i miglioramenti teorici nella precisione delle stime hanno corrisposto solo in parte, come appena visto, a miglioramenti effettivi, il che indica una parziale irrealisticità di alcune delle ipotesi fatte nel paragrafo 2, come ad esempio la già citata in correlazione tra i residui degli errori di stima della (2.3).

Sembra pertanto necessario, prima di ricorrere a procedure di stima più complesse di quelle dirette, verificare la validità del modello tramite analisi preliminari, il che conferma anche come l'applicabilità di tali metodi risulti realistica solo in contesti in cui si disponga di buone conoscenze preliminari sul fenomeno oggetto di studio e di una base di dati storici sufficientemente lunga.

Tabella 6.4 – Numero di unità non rispondenti, valori medi veri del fatturato annuale e stime della media quadratica dell'errore relativa a vari metodi di stima

Attività prevalente	Addetti	n	Valori veri	Stima coefficienti di variazione %					
				IR	IC	IR*	IC*	IR**	IC**
Totale commercio al dettaglio		2.363	957	1,77	1,75	**1,56**	1,58	1,69	1,66
Specializzati alimentari		333	449	2,91	3,01	**2,63**	2,72	2,83	2,88
Specializzati non alimentari		1.585	588	2,26	2,27	**1,93**	2,03	2,08	2,08
Despecializzati alimentari		345	2.521	1,21	1,15	1,10	**1,05**	1,20	1,14
Despecializzati non alimentari		100	3.106	1,22	1,27	**1,15**	1,20	1,21	1,25
	1-2	1.185	51	4,71	5,31	4,74	5,35	**4,69**	5,27
	3-5	405	194	3,53	3,58	**3,52**	3,57	**3,52**	3,57
	6-9	210	756	2,02	2,04	2,01	2,03	**2,00**	2,01
	10-19	297	2.295	1,75	1,75	**1,54**	**1,54**	1,73	1,73
	>19	266	4.821	1,48	1,44	**1,21**	1,25	1,35	1,30
Specializzati alimentari	1-2	190	43	**4,14**	4,57	4,22	4,66	**4,14**	4,56
	3-5	66	281	4,74	4,78	**4,48**	4,51	4,72	4,75
	6-9	28	696	1,92	1,99	**1,86**	1,93	1,92	1,99
	10-19	27	915	3,19	3,20	2,70	**2,67**	2,84	2,88
	>19	22	3.567	2,51	2,61	**2,18**	2,28	2,47	2,49
Specializzati non alimentari	1-2	890	44	4,19	4,88	4,39	5,16	**4,17**	4,84
	3-5	301	151	3,61	3,66	3,68	3,75	**3,60**	3,65
	6-9	115	483	2,64	2,64	2,61	2,60	2,58	**2,56**
	10-19	163	633	2,61	2,63	**2,44**	2,45	2,58	2,59
	>19	116	5.937	1,96	1,95	**1,53**	1,64	1,73	1,70
Despecializzati alimentari	1-2	82	81	8,71	9,28	**7,28**	7,56	8,67	9,18
	3-5	27	225	2,96	3,18	**2,84**	3,02	2,96	3,18
	6-9	51	1.215	1,65	1,66	1,67	1,67	**1,64**	1,65
	10-19	88	4.986	1,48	1,48	**1,30**	**1,30**	1,48	1,48
	>19	97	3.673	0,52	0,45	0,50	**0,44**	0,52	0,45
Despecializzati non alimentari	1-2	23	280	4,65	4,68	**4,39**	4,48	4,62	4,64
	3-5	11	776	0,85	**0,84**	0,86	0,85	0,85	**0,84**
	6-9	16	1.360	1,58	1,62	1,60	1,66	**1,57**	1,62
	10-19	19	6.045	1,69	1,65	1,43	**1,41**	1,65	1,62
	>19	31	5.130	**0,76**	0,82	0,79	0,86	**0,76**	0,82

7. Appendice: dimostrazione della relazione (3.3)

La funzione di Lagrange da minimizzare è data da:

$$\Phi(\hat{w}_i, \lambda) = \sum_{i=1}^{n_0} \hat{w}_i^2 \sigma_i^2 + \left[\sum_{i=1}^{n_0} (\hat{w}_i - w_i) y_i \right]^2 + \lambda \left(\sum_{i=1}^{n_0} \hat{w}_i - \mathrm{W} \right).$$

Derivando rispetto all'*i-mo* peso incognito si ottiene la relazione:

$$\frac{\partial \Phi}{\partial \hat{w}_i} = 2\,\hat{w}_i \sigma_i^2 + 2\,y_i \left[\sum_{i=1}^{n_0} (\hat{w}_i - w_i) y_i \right] + \lambda = 0$$

che, ponendo $(\hat{w}_i - w_i) = x_i$ e $\sum_{i=1}^{n_0} x_i y_i = X$ diventa:

$$(x_i + w_i) + \frac{y_i}{\sigma_i^2} X + \frac{\lambda}{2\sigma_i^2} = 0\,. \qquad (7.1)$$

Sommando la (7.1) rispetto a *i* si ottiene la relazione:

$$X \sum_{i=1}^{n_0} \frac{y_i}{\sigma_i^2} + W + \frac{\lambda}{2} \sum_{i=1}^{n_0} \frac{1}{\sigma_i^2} = 0 \qquad (7.2)$$

mentre moltiplicando la (7.1) per y_i, sommando rispetto a *i* e ricordando che $\sum_{i=1}^{n_0} x_i = 0$ si ottiene la relazione:

$$\left(1 + \sum_{i=1}^{n_0} \frac{y_i^2}{\sigma_i^2} \right) X + W + \frac{\lambda}{2} \sum_{i=1}^{n_0} \frac{y_i}{\sigma_i^2} = 0\,. \qquad (7.3)$$

Risolvendo la (7.2) e la (7.3) rispetto a X e λ e sostituendo le soluzioni nella (7.1) si ottiene infine la relazione (3.3).

Bibliografia

ABBATE C. (1998), "La completezza delle informazioni e l'imputazione da donatore con distanza mista minima", *Quaderni di ricerca*, 4, pagg.67-102, Istat, Roma.

BALLIN M., FALORSI P.D., RUSSO A. (2000), "Condizioni di coerenza e metodi di stima per le indagini campionarie sulle imprese", *Quaderni di ricerca*, 2, pagg.31-52, Franco Angeli, Milano.

BARROSO L.P., BUSSAB W.O., KNOTT M. (1996), "Imputation in Panel Data Using the Mixed Model", *Relatorio Tecnico*, 9629, Universidade de São Paulo, Brasil.

BREWER K.R.W. (1995), "Combining Design-Based and Model-Based Inference", *Business Survey Methods*, pagg.589-606, John Wiley & Sons, New York.

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

EDWARDS W.S., CANTOR D. (1991), "Towards a Response Model in Establishment Surveys", in *Statistical Data Editing Methods and Techniques - Conference of European Statisticians and Studies*, Vol. 1, 44, pagg.52-68.

GISMONDI R. (1996), "Gli effetti delle non risposte nell'indagine sulle vendite al dettaglio delle piccole imprese", *Quaderni di ricerca*, 4, pagg.199-236, Istat, Roma.

GISMONDI R. (1999), "Un criterio generalizzato per l'imputazione di dati mancanti in indagini congiunturali", *Statistica*, 1, pagg.83-100, Clueb, Bologna.

GISMONDI R. (2000), "Confronti tra misure di accuratezza per la stima di mancate risposte in indagini longitudinali", *Statistica*, 3, pagg.557-575, Clueb, Bologna, Bologna.

GISMONDI R. (2001), "Pesi reali e pesi ottimali nella stima di numeri indici", mimeo, Istat, Roma.

HIDIROGLOU M.A., BERTHELOT J.M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology*, 12, pagg.73-84, Statistics Canada, Ottawa.

HIDIROGLOU M.A., SRINATH K.P. (1993), "Problems Associated with Designing Subannual Business Surveys", *Journal of Business & Economic Statistics*, Vol.11, 4, pagg.397-405.

ISTAT (1998), *La nuova indagine sulle vendite al dettaglio: aspetti metodologici e contenuti innovativi*, "Metodi e norme", 3, Istat, Roma.

ISTAT (2000), *Gli indici delle vendite al dettaglio nel 2000*, "Informazioni", 48, Istat, Roma.

LUCEV D. (1997), *Tipologie e controllo dell'errore di non risposta per la qualità dei dati economici*, Rocco Curto Editore, Napoli.

LETTAU M.K., LOEWENSTEIN M.A. (2000), "Optimal Weighting of Index Components: An Application to the Employment Cost Index", *Journal of Official S*tatistics, Vol.16, 1, pagg.39-52.

QUINTANO C., CASTELLANO R. (2001), *Strategies for Dealing with Non-responses for Quality in some Istat Surveys*, "Essays", Istat, Roma.

RIZZO L., KALTON G., BRICK J.M. (1996), "A Comparison of Some Weighting Adjustment Methods for Panel Non-response", *Survey Methodology*, 22, pagg.43-53.

TREMBLAY V. (1986), "Practical Criteria for Definition of Weighting Classes", *Survey Methodology*, Vol.12, 1, pagg.85-98.