Tecniche di stima e condizioni di coerenza per indagini infraannuali ripetute nel tempo

Roberto Gismondi (*)

(*) ISTAT – Servizio SCO

Riassunto

La qualità delle indagini infraannuali è spesso messa a serio rischio dalla vasta gamma di possibili fonti di errore che possono influire sulla precisione delle stime finali. Tra queste, assumono particolare rilievo la cadenza ravvicinata delle rilevazioni e la necessità di operare con strutture campionarie longitudinali di tipo panel, che possono comportare sia distorsioni, sia perdita di efficienza rispetto a strategie alternative. Con questa premessa, nel prosieguo vengono illustrate possibili opzioni metodologiche che, sotto certe condizioni, possono comportare un miglioramento nella qualità delle stime. In particolare, si è cercato di valutare: la distribuzione ottimale delle unità campionarie nei vari sottoperiodi oggetto d'interesse tale da rendere minimo il *gap* qualitativo nei confronti delle indagini strutturali; il possibile miglioramento delle stime ottenibile tramite l'introduzione di particolari vincoli o tramite l'utilizzo di serie storiche; l'utilizzo di stime basate sulla sintesi di due stime derivate da fonti diverse.

1. Premessa[1]

Nella produzione di statistiche ufficiali diviene sempre più pressante l'esigenza di costruire un sistema di indagini "integrato", in cui le stime prodotte dalle diverse rilevazioni siano facilmente e correttamente utilizzabili in modo congiunto. Questo richiede il graduale ripensamento dell'insieme delle statistiche prodotte secondo un'ottica sistemica; in chiave più tecnica, ciò implicherà un crescente ricorso a strategie campionarie che utilizzano al meglio, soprattutto nella fase di stima, informazioni ausiliarie.

Tale aspetto diviene ancora più rilevante nel caso delle indagini infraannuali (definite anche congiunturali), per le quali al problema operativo di dover gestire più rilevazioni a cadenze spesso molto ravvicinate (ad esempio mensili) - di cui occorre garantire un sufficiente livello qualitativo - si aggiunge quello della coerenza dei risultati riferiti all'intero anno, ottenuti per somma delle stime infraannuali, con la stima annua del medesimo aggregato d'interesse ottenuta con indagini strutturali *ad hoc*. Ad esempio, le indagini sui conti delle imprese condotte ogni anno dall'ISTAT producono stime dei livelli medi annui per diverse variabili economiche, tra cui il fatturato complessivo, per tutte le principale branche di attività economica. Altre indagini mensili misurano a loro volta la dinamica infraannuale del fatturato, e un'ovvia condizione di coerenza che tutti gli utilizzatori di statistiche vorrebbero soddisfatta è la coincidenza, almeno approssimata, tra le stime dei ricavi medi annui ottenibili con le due fonti. Due esempi sono dati dall'indagine mensile sul fatturato del commercio al dettaglio e dall'indagine mensile sul fatturato della grande industria. In realtà va osservato che:

1. il concetto di coerenza sviluppato nell'ambito della recente teoria dei campioni (cfr. il successivo paragrafo 6) non si riferisce direttamente alla variabile oggetto di interesse, ma ad una o più variabili ausiliarie correlate ad essa;
2. poiché le indagini congiunturali producono stime assai in anticipo rispetto alle indagini strutturali, non possono esistere tecniche di stima che possano garantire la coincidenza tra le stime annuali ottenute per somma di stime infraannuali o derivate direttamente da un'unica indagine strutturale, per cui ci si trova a fronteggiare un problema sostanzialmente tautologico.

Ciò nonostante, raccomandazioni a favore di una verifica di coerenza del tipo di quella descritta nel precedente punto 2 sono incluse nello stesso Regolamento Congiunturale sulle Imprese (COMMISSIONE EUROPEA, 1998) ed hanno già costituito un importante argomento di discussione in ambito internazionale (Gismondi, 1998*b*).

Resta il fatto che è utile verificare, da un lato, se e sotto quali condizioni la precisione di una stima annuale ottenuta da un'unica indagine strutturale sia effettivamente più elevata rispetto alla precisione di una stima annuale ottenuta per somma di k stime infraannuali; dall'altro, se ed in che modo è possibile utilizzare informazioni addizionali derivate da altre fonti per migliorare la qualità delle stime infraannuali.

Nel prosieguo, le condizioni legate al disegno di campionamento si suppongono date e si privilegerà un approccio di tipo *model-based*, come anche suggerito in altri contesti (Drudi e Filippucci, 1996; Binder e Dick, 1989). Tale impostazione è peraltro realistica nel caso in cui si ricorra a panel longitudinali selezionati con tecniche almeno parzialmente ragionate[2].

---

[1] Questo lavoro riassume in forma più sistematica l'intervento svolto dall'autore nell'ambito del workshop *Alcune problematiche teorico-applicative connesse alle metodologie e alla qualità delle indagini longitudinali*, svoltosi in ISTAT il 6 novembre 2001.

[2] Ad esempio, assai frequente in pratica è il caso di panel che includono solo le unità più grandi.

Dopo avere introdotto la struttura formale del modello di riferimento che si suppone generi le osservazioni campionarie nel vari sottoperiodi dell'anno (paragrafo 2), vengono sviluppati i punti seguenti:

- valutazione della distribuzione ottimale delle unità campionarie nei vari sottoperiodi oggetto d'interesse, distinguendo il caso di campioni infraannuali del tutto indipendenti (paragrafo 3) dal caso di campioni perfettamente dipendenti in quanto basati su una struttura panel (paragrafo 4).
- Uso di una serie di rilevazioni infraannuali abbinato ad un'unica rilevazione strutturale di fine anno (paragrafo 5).
- Possibile miglioramento delle stime ottenibile tramite l'introduzione di particolari vincoli (paragrafo 6).
- Possibile miglioramento delle stime ottenibile tramite l'utilizzo di serie storiche (paragrafo 7).
- Utilizzo di stime basate sulla sintesi di due stime derivate da fonti diverse (paragrafo 8).

Va infine notato come, sebbene la trattazione sia sviluppata supponendo un'ottica longitudinale ed un contesto applicativo maggiormente orientato a favore delle indagini sulle imprese, molte delle considerazioni proposte possono ritenersi ancora valide, *mutatis mutandis*, nel caso in cui si faccia riferimento ad indagini su individui e/o qualora si sostituiscano ai *k* periodi temporali presi a riferimento altrettanti domini di interesse (classi di attività economica, aree geografiche, ecc.).

2. Struttura formale delle indagini infraannuali e confronti con le indagini strutturali

Si suppone un disegno d'indagine di tipo campionario tramite il quale, con riferimento al sottoperiodo *t* dell'anno *A*, vengono osservate le realizzazioni campionarie della variabile *Y* di cui è oggetto di stima la media $\bar{Y}_{At}$. La generica realizzazione campionaria di tale variabile osservabile sull'unità *i-ma* è indicabile con:

y_{Ati} = valore di *y* osservato sull'unità *i-ma* nel sottoperiodo *t* dell'anno *A*.

Seguendo un approccio *model-based*, se μ_{At} indica il vero livello medio *rispetto al modello di superpopolazione* relativo alle unità appartenenti all'universo nel periodo *t* dell'anno *A*, si può ragionevolmente supporre che valga la relazione[3]:

$$y_{Ati} = \mu_{At} + \varepsilon_{Ati} \tag{2.1}$$

dove l'errore ε_{Ati} sarà caratterizzato da questa struttura stocastica per medie, varianze e covarianze:

$$\begin{cases} E(\varepsilon_{Ati}) = 0 & \forall i \\ VAR(\varepsilon_{Ati}) = \sigma^2_{At} & \forall i \\ COV(\varepsilon_{Ati}; \varepsilon_{Arj}) = 0 \quad \text{se} \quad i \neq j \end{cases} \tag{2.2}$$

[3] In realtà il modello più realistico è della forma $y = \beta x + \varepsilon$

dove i pedici *t* e *r* utilizzati nella formula della covarianza indicano due occasioni d'indagine relative all'anno *A* e dove si può avere *t*=*r*. Si possono poi introdurre le ulteriori definizioni seguenti:

n_{At} = numero di osservazioni campionarie disponibili con riferimento al sottoperiodo *t* dell'anno *A*;
$\bar{y}_{At}$ = media campionaria dei valori di *y* riferiti al periodo t dell'anno *A*

Si avrà quindi, con notazioni ovvie, la media campionaria:

$$\bar{y}_{At} = \sum_{i=1}^{n_{At}} \frac{y_{Ati}}{n_{At}} \tag{2.3}$$

che, in base alla prima delle relazioni (2.2), sarà uno stimatore corretto di μ_{At}, ossia si avrà che:

$$E(\bar{y}_{At} - \bar{Y}_{At}) = 0\,. \tag{2.4}$$

Sulla base della formula della covarianza nella (2.2) si avrà la varianza:

$$VAR(\bar{y}_{At}) = \frac{\sigma^2_{At}}{n_{At}}\,. \tag{2.5}$$

Si supponga poi di suddividere l'anno in *k* sottoperiodi, tali che la somma delle loro durate riproduca esattamente l'intero anno (si farà quindi riferimento a mesi, trimestri, quadrimestri, ecc.). Con riferimento a ciascun sottoperiodo si suppone disponibile una stima campionaria del valore medio $\bar{Y}_{At}$ della forma (2.3): di conseguenza, uno stimatore non distorto del valore medio $\bar{Y}_A$ relativo all'intero anno *A* sarà definibile nel modo seguente:

$\bar{y}_{A(k)}$ = stima campionaria dei valori di *y* riferiti all'anno *A*, basata sulla somma delle stime relative a *k* periodi infraannuali.

In simboli si avrà quindi:

$$\bar{y}_{A(k)} = \sum_{t=1}^{k} \bar{y}_{At}\,. \tag{2.6}$$

La numerosità campionaria complessiva relativa allo stimatore (2.6) è data semplicemente dalla somma delle numerosità campionarie riferite a ciascuno dei sottoperiodi, ossia in simboli:

$$n_{A(k)} = \sum_{t=1}^{k} n_{At}\,. \tag{2.7}$$

Nel prosieguo, si valuteranno la correttezza e, qualora possibile, l'efficienza di diversi stimatori di una media con riferimento al modello definito dalle relazioni (2.2), prescindendo dal particolare disegno campionario utilizzato. Si tratteranno separatamente le situazioni di indipendenza e di dipendenza (ricorso ad un panel) tra osservazioni successive, escludendo il

caso della rotazione parziale delle unità campionarie, rimandando per le conseguenze teoriche generali sulle tecniche di stima a Cicchitelli (et al., 1998) e per alcune applicazioni al caso della stima di variazioni a Gismondi (1998).

3. Ricorso a campioni infraannuali indipendenti

Per quanto riguarda la varianza dello stimatore (2.6), se si suppone che le *k* rilevazioni infraannuali si basino su campioni indipendenti (ossia se si suppone di utilizzare un campione completamente rinnovato per ogni rilevazione), si avrà:

$$COV(\varepsilon_{Ati};\varepsilon_{Ari}) = 0 \qquad \text{per ogni coppia } (t,r) \text{ con } t \neq r \tag{3.1}$$

e, di conseguenza, la varianza dello stimatore (2.6) sarà data da:

$$VAR(\bar{y}_{A(k)}) = \sum_{t=1}^{k} \frac{\sigma_{At}^2}{n_{At}}. \tag{3.2}$$

Si noti come la precedente espressione equivalga anche alla somma delle varianze dei *k* stimatori infraannuali, ossia a $\sum_{t=1}^{k} VAR(\bar{y}_{At})$ - che esprime il livello medio di imprecisione delle *k* stime - indifferentemente nei casi in cui tali campioni siano tra di loro dipendenti od indipendenti. Di conseguenza, le condizioni di ottimalità determinate nel prosieguo saranno riferibili anche al caso in cui si voglia rendere minima la somma delle varianze infraannuali anziché la varianza dello stimatore strutturale ottenuto per somma delle *k* stime infraannuali.

Con queste premesse, un particolare aspetto metodologico che si connette da vicino a quanto esposto nella premessa consiste nel verificare se ed in quali casi la stima di un ammontare medio annuo ottenibile tramite la (2.6) – ossia come somma di *k* stime infraannuali, in questo caso indipendenti – possa risultare più (o almeno altrettanto) precisa rispetto alla stima della media annua basata su un'unica rilevazione annuale, che nel prosieguo sarà definita come stima "strutturale".

Il problema può essere affrontato determinando, in via preliminare, la distribuzione ottimale delle numerosità campionarie in ciascuno dei sottoperiodi oggetto d'indagine, fissato il numero complessivo di unità intervistabili nel corso dell'intero anno, posto pari $n_{A(k)}$. La determinazione della distribuzione di tale numerosità nei vari sottoperiodi che rende minima la varianza di (2.6) è ottenibile impostando questa semplice funzione di Lagrange, da minimizzare uguagliandone a zero la derivata prima:

$$\Phi(n_{At},\lambda) = \sum_{t=1}^{k}\left(\frac{\sigma_{At}^2}{n_{At}}\right) + \lambda\left(\sum_{t=1}^{k} n_{At} - n_{A(k)}\right). \tag{3.3}$$

Derivando la (3.3) rispetto alle numerosità n_{At} e a λ si ricava facilmente la soluzione ottimale:

$$n_{At}^{*} = \frac{\sigma_{At}^2}{\sum_{t=1}^{k}\sigma_{At}^2} n_{A(k)} \tag{3.4}$$

sulla cui base la distribuzione ottimale delle numerosità campionarie tra i vari sottoperiodi risulta proporzionale al livello delle relative varianze. Sostituendo la soluzione ottimale (3.4) nella (3.2) si ricava la varianza minima data da:

$$VAR(\bar{y}^{*}_{A(k)}) = \frac{k\sum_{t=1}^{k}\sigma^{2}_{At}}{n_{A(k)}}. \tag{3.5}$$

Si noti come nel caso di un'unica rilevazione annuale si abbia k=1, sebbene sia preferibile indicare la stima annuale basata su una sola rilevazione strutturale semplicemente con il simbolo $\bar{y}_A$. Inoltre, è immediato verificare come la varianza dello stimatore $\bar{y}_{A(k)}$ risulterebbe ancora uguale alla (3.5) anche se si utilizzasse un'allocazione costante nei vari sottoperiodi, ossia se $n_{At} = n_{A(k)}/k$. Sulla base dell'ipotesi (3.1) e della formula generale (3.2), la varianza dello stimatore strutturale sarà data semplicemente da:

$$VAR(\bar{y}_A) = \frac{\sigma^{2}_{A}}{n_A}. \tag{3.6}$$

Si vuole dunque valutare se è possibile che si verifichi questa condizione:

$$VAR(\bar{y}^{*}_{A(k)}) \leq VAR(\bar{y}_A) \tag{3.7}$$

ossia che la varianza dello stimatore basato sulla somma di k stime infraannuali e sull'allocazione ottimale (3.4) sia non superiore alla varianza dello stimatore strutturale. Si noti come, sulla base delle relazioni (2.1) e (2.2), si abbia:

$$\sigma^{2}_{A} = VAR(y_{Ai}) = VAR\left(\sum_{t=1}^{k} y_{Ati}\right) = \sum_{t=1}^{k}\sigma^{2}_{At} + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k}\sigma_{Atr} = \sum_{t=1}^{k}\sigma^{2}_{At} + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k}\rho_{Atr}\,\sigma_{At}\,\sigma_{Ar} \tag{3.8}$$

Di conseguenza, poiché la condizione (3.7) equivale alla condizione:

$$\frac{k\sum_{t=1}^{k}\sigma^{2}_{At}}{n_{A(k)}} \leq \frac{\sigma^{2}_{A}}{n_A} \tag{3.9}$$

si dovrà avere:

$$n_{A(k)} \geq \left(\frac{k\sum_{t=1}^{k}\sigma^{2}_{At}}{\sum_{t=1}^{k}\sigma^{2}_{At} + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k}\rho_{Atr}\,\sigma_{At}\,\sigma_{Ar}}\right) n_A = (R)\,n_A. \tag{3.10}$$

Dunque, la (3.7) si potrà verificare solo se la somma delle numerosità campionarie associate alle k rilevazioni infraannuali $n_{A(k)}$ risulta sufficientemente più grande della numerosità campionaria n_A utilizzata nell'unica indagine annuale, dato che il rapporto R in parentesi tonde è sempre

maggiore od uguale a uno[4]. Di conseguenza, se anche le numerosità campionarie utilizzate in ciascuna rilevazione infraannuale fossero assegnate in modo ottimale, se la loro somma fosse esattamente uguale al numero di unità campionarie utilizzate nell'unica rilevazione strutturale (ossia se $n_{A(k)} = n_A$) lo stimatore $\bar{y}^*_{A(k)}$ basato sulla somma delle *k* stime infraannuali non sarà mai più efficiente dello stimatore strutturale $\bar{y}_A$. In alternativa, si dovrebbero utilizzare numerosità campionarie per le indagini infraannuali tali che la loro somma risulti almeno uguale alla parte intera del secondo membro della (3.9).

Se si decidesse di utilizzare campioni infraannuali tutti caratterizzati dalla stessa numerosità $n_{At} = n_{A(k)}/k$, poiché la varianza di $\bar{y}_{A(k)}$ risulterebbe in questo caso pari alla varianza di $\bar{y}^*_{A(k)}$, varrebbero considerazioni analoghe a quelle appena esposte.

Il caso in cui si ricorra a campioni del tutto indipendenti per ciascuno dei periodi infraannuali osservati non è particolarmente frequente. Ciò nonostante, è possibile immaginare un contesto operativo basato, ad esempio, su quattro rilevazioni trimestrali finalizzate a stimare le relative medie di trimestre, ciascuna basata su un campione di unità che saranno reintervistate solo nel corrispondente trimestre dell'anno successivo, al fine di contenere l'onere di risposta. Tale procedura potrebbe rivelarsi assolutamente realistica soprattutto nel caso in cui risultasse di particolare interesse la stima della variazione della media intercorsa tra gli stessi trimestri di due anni successivi (variazioni tendenziali).

4. Ricorso ad un panel

Questa situazione sarà sviluppata supponendo, in particolare, che siano intervistate sempre le stesse unità in ognuno dei *k* periodi di osservazione. Per semplicità si supporrà che $n_{At} = n_{A(k)}/k$, per cui si avrà:

$$VAR(\bar{y}_{At}) = k\frac{\sigma^2_{At}}{n_{A(k)}}. \tag{4.1}$$

Varrà poi la seguente relazione:

$$COV(\varepsilon_{Ati};\varepsilon_{Ari}) = \sigma_{Atr}. \tag{4.2}$$

Sulla base delle relazioni (2.5), (2.6) e (4.2), si avrà:

$$COV(\bar{y}_{At};\bar{y}_{Ar}) = COV\left(\sum_{i=1}^{n_{At}}\frac{y_{Ati}}{n_{At}};\sum_{i=1}^{n_{At}}\frac{y_{Ari}}{n_{At}}\right) = \sum_{i=1}^{n_{At}}\frac{\sigma_{Atr}}{n^2_{At}} = \frac{\sigma_{Atr}}{n_{At}} = k\frac{\sigma_{Atr}}{n_{A(k)}} \tag{4.3}$$

e quindi:

---

[4] Dalla disuguaglianza di Cauchy-Schwarz si ha che $\left(\sum_{t=1}^{k} a_t b_t\right)^2 \le \left(\sum_{t=1}^{k} a_t^2\right)\left(\sum_{t=1}^{k} b_t^2\right)$, che ponendo $b_t = 1$ si riduce a $\left(\sum_{t=1}^{k} a_t\right)^2 \le k\sum_{t=1}^{k} a_t^2$. Se anche si supponessero tutte le correlazioni longitudinali ρ_{Atr} pari ad uno, e ponendo $a_t = \sigma_{At}$, basta notare che il primo ed il secondo termine dell'ultima disuguaglianza risultano pari, rispettivamente, al denominatore ed al numeratore di *R*.

$$VAR(\bar{y}_{A(k)}) = \sum_{t=1}^{k} VAR(\bar{y}_{At}) + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k} COV(\bar{y}_{At};\bar{y}_{Ar}) = \frac{k}{n_{A(k)}}\left(\sum_{t=1}^{k}\sigma_{At}^2 + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k}\sigma_{Atr}\right). \quad (4.4)$$

Ricordando la (3.5) consegue che, a parità di condizioni, la varianza dello stimatore basato sulla somma di *k* stime infraannuali sarà più elevata nel caso di campioni dipendenti, dato che è ragionevole supporre $\sigma_{Atr} \geq 0$. Inoltre, ricordando la (3.8), il termine in parentesi tonde della (4.2) è pari a σ_A^2 e, sulla base della (3.6), si avrà infine:

$$VAR(\bar{y}_{A(k)}) = k\left(\frac{n_A}{n_{A(k)}}\right)VAR(\bar{y}_A). \quad (4.5)$$

Quindi, se *k*>1 la varianza dello stimatore della media annuale ottenuto per somma di *k* stimatori infraannuali ottenuti sulla base di altrettante rilevazioni basate sulle medesime unità sarà sempre più elevata della varianza dello stimatore basato su un'unica rilevazione strutturale fintanto che $n_{A(k)} < k\, n_A$; in particolare, varrà il segno di uguaglianza solo se in ogni sottoperiodo *t* si utilizza un numero di unità campionarie pari al numero di unità che si utilizzano nell'unica indagine annuale. Tale condizione, almeno in linea teorica, appare più restrittiva della condizione (3.10) ottenuta con riferimento al caso di campioni infraannuali indipendenti.

Come primo esempio, si può supporre il caso di due rilevazioni semestrali (*k*=2), con varianze uguali e covarianza pari a 0,5; consegue che nella (3.10) *R*=4/3, per cui la strategia basata sulla somma di due stime semestrali è almeno altrettanto efficace di quella basata su un'unica stima annuale fintanto che $n_{A(2)} \geq 4/3\, n_A$ nel caso in cui i campioni semestrali fossero indipendenti, mentre nel caso di campioni panel sulla base della (4.5) occorrerebbe che $n_{A(2)} \geq 2\, n_A$.

Come secondo esempio, si possono supporre 12 rilevazioni mensili (*k*=12), con varianze uguali e correlazioni pari a 0,8; si ha *R*=1,22, per cui la somma di 12 stime mensili è almeno altrettanto efficace di un'unica stima annuale fintanto che $(n_{A(12)}/12) \geq 0{,}102\, n_A$ nel caso in cui i campioni fossero indipendenti, mentre nel caso di campioni panel basterebbe che $(n_{A(12)}/12) \geq n_A$.

Questi risultati derivano dal fatto che se non si rinnova il campione da periodo a periodo si ottengono stime infraannuali mediamente più imprecise rispetto al caso di perfetta rotazione, perché le unità appartenenti ad un unico panel annuale, osservato in più sottoperiodi dell'anno potrebbero essere molto rappresentative per qualche sottoperiodo e molto poco rappresentative per altri.

Una formulazione più complessa si otterrebbe se invece della stima di un livello medio si desiderasse la stima di una variazione[5].

5. Altre strategie operative

Nel paragrafo precedente è stata evidenziata la sostanziale inefficienza delle stime infraannuali qualora esse siano utilizzate per stimare un ammontare medio riferito all'intero anno,

[5] Gismondi (op. cit., 1998).

in confronto alla qualità di un'unica stima strutturale e a parità di dimensione campionaria annua complessiva.

Per questo motivo è lecito immaginare strategie di rilevazione alternative che, a parità di numerosità campionarie totali, possano condurre a stimatori del livello medio annuo più efficienti.

In particolare, in questo contesto si suppone di aver prefissato il numero $n_{A(k)}$ di interviste realizzabili nell'arco di un intero anno e di valutare l'efficienza della strategia basata su (k-1) rilevazioni infraannuali, relative a ciascuno dei primi (k-1) sottoperiodi dell'anno e sulla cui base si calcolano gli stimatori $\bar{y}_{At}$ per t=1,…,(k-1), e su una *k-ma* rilevazione strutturale, in cui si chiede ai rispondenti di indicare il proprio ammontare annuo complessivo di *Y*, utile per stimare la media annua $\bar{y}_A$. Si suppone di intervistare n_{Ak} unità nell'ultima rilevazione e, quindi, $n_{At} = (n_{A(k)} - n_{Ak})/(k-1)$ per ciascuna delle precedenti (k-1).

L'utilità di una simile strategia potrebbe consistere nel risparmio di unità campionarie per le prime (k-1) rilevazioni infraannuali, bilanciata dall'utilizzo di molte unità campionarie nella rilevazione strutturale, ad esempio in contesti in cui esiste un serio vincolo al numero di unità intervistabili nell'arco dell'anno e si riscontra una forte variabilità dei dati in corrispondenza dell'ultimo sottoperiodo. In pratica, possono presentarsi i due casi operativi seguenti.

5.1 Somma di (k-1) varianze infraannuali e di una varianza annuale

Per $n_{A(k)}$ fissato, si vuole determinare la numerosità n^*_{Ak} da assegnare all'ultima rilevazione strutturale in modo da minimizzare la somma delle varianze delle (k-1) stime trimestrali e della varianza dell'ultima stima annuale. Per tenere conto del diverso ordine di grandezza delle varianze infraannuali e strutturale, si possono introdurre due pesi opportuni γ e λ e si avrà:

$$\Phi(n_{Ak}) = \gamma \sum_{t=1}^{k-1} VAR(\bar{y}_{At}) + \lambda VAR(\bar{y}_{Ak}) = \gamma(k-1) \sum_{t=1}^{k-1} \frac{\sigma^2_{At}}{n_{A(k)} - n_{Ak}} + \lambda \frac{\sigma^2_A}{n_{Ak}} \tag{5.1}$$

Dopo alcuni passaggi si ricava la soluzione ottimale:

$$n^*_{Ak} = \frac{\sqrt{\eta}\,\sigma_A}{\left(\sqrt{\gamma(k-1) \sum_{t=1}^{k-1} \sigma^2_{At}} + \sqrt{\eta}\,\sigma_A \right)} n_{A(k)} \tag{5.2}$$

che, nel caso di due sole rilevazioni con pesi uguali, si riduce a:

$$n^*_{Ak} = \left(\frac{\sigma_A}{\sigma^2_{A1} + \sigma_A} \right) n_{A(k)} . \tag{5.3}$$

Si noti come la soluzione ottimale resta inalterata sia nel caso di una struttura panel, sia in quello di campioni del tutto indipendenti.

5.1 Somma di k varianze infraannuali

In questo caso si suppone di voler determinare la numerosità n^{*}_{Ak} in modo da rendere minima la somma delle *k* stime infraannuali, di cui la *k-ma* - ossia la stima della media della media del *k-mo* sottoperiodo - sarà data dalla differenza $\bar{y}_A - \sum_{t=1}^{k-1} \bar{y}_{At}$. Lo sviluppo analitico è diverso a seconda che si supponga dipendenza od indipendenza tra le osservazioni infraannuali.

Se i campioni sono tutti indipendenti – ossia se si utilizzano sempre unità diverse – e se in ciascuna delle *k* occasioni d'indagine si intervistano n_{At} unità, si può facilmente verificare che la somma delle varianze dei *k* stimatori infraannuali è data da:

$$2\sum_{t=1}^{k-1} VAR(\bar{y}_{At}) + VAR(\bar{y}_A) = 2\sum_{t=1}^{k-1} \frac{\sigma^2_{At}}{n_{At}} + \frac{\sigma^2_A}{n_{Ak}} \tag{5.4}$$

che equivale alla (5.1) per γ=2, λ=1 e se n_{At} è costante nei primi (*k*-1) periodi. Qualora $k\, n_{At} = n_{A(k)}$ per ogni *t* – ossia se si intervista lo stesso numero di unità in ognuno dei *k* sottoperiodi – la (5.4) si riduce alla relazione:

$$\frac{k}{n_{A(k)}}\left(2\sum_{t=1}^{k-1}\sigma^2_{At} + \sigma^2_A\right) \tag{5.5}$$

Può essere utile derivare la relazione (5.4) rispetto a n_{Ak}, per ricavare la numerosità da assegnare all'ultima rilevazione e, quindi, alle prime (*k*-1) – supponendo una numerosità costante per ogni sottoperiodo – in modo da rendere minima la somma delle *k* varianze supponendo di aver fissato la numerosità annua complessiva $n_{A(k)}$. Si può verificare come è sufficiente risolvere rispetto a n_{Ak} la seguente equazione di secondo grado, previa verifica della non negatività[6] di almeno una delle due soluzioni:

$$\left[2(k-1)\sum_{t=1}^{k-1}\sigma^2_{At} - \sigma^2_A\right] n^2_{Ak} + 2\left[\sigma^2_A\, n_{A(k)}\right] n_{Ak} - \left[\sigma^2_A\, n^2_{A(k)}\right] = 0\,. \tag{5.6}$$

La trattazione è più complessa nel caso, peraltro più realistico, di *k* campioni dipendenti. Per semplicità si supporrà di intervistare sempre le stesse unità nei primi (*k*-1) periodi, mentre nel *k-mo* periodo si suppone di intervistare ancora tutte le unità intervistate nei periodi precedenti, più un contingente di nuove unità. Se in tutto si intervistano $n_{A(k)}$ unità, di cui n_{Ak} nell'ultimo periodo, consegue che in ciascuno dei precedenti (*k*-1) periodi si intervistano $n_{At} = (n_{A(k)} - n_{Ak})/(k-1)$ unità.

Si può dimostrare (cfr. appendice 2) che in questo caso la somma delle *k* varianze infraannuali è data dalla relazione:

[6] Nonché, ovviamente, della loro appartenenza al campo dei numeri reali.

$$(k-1)\sum_{t=1}^{k-1}\frac{\sigma_{At}^2}{n_{A(k)}-n_{Ak}}+\left\{\frac{1}{n_{Ak}}\left[\sigma_{Ak}^2-\frac{n_{A(k)}-k\,n_{Ak}}{n_{A(k)}-n_{Ak}}\left(\sum_{t=1}^{k-1}\sigma_{At}^2+\sum_{t=1}^{k-1}\sum_{t\neq r=1}^{k-1}\sigma_{Atr}\right)\right]\right\} \tag{5.7}$$

dove il termine in parentesi graffe è la varianza del *k-mo* stimatore infraannuale $\left(\bar{y}_A-\sum_{t=1}^{k-1}\bar{y}_{At}\right)$.

Va notato come se fosse $k\,n_{Ak}=n_{A(k)}$, ossia se anche nella *k-ma* occasione si intervistassero sempre e solo le stesse unità intervistate nei periodi precedenti, la precedente espressione si ridurrebbe alla (3.2), ossia a:

$$k\sum_{t=1}^{k}\frac{\sigma_{At}^2}{n_{A(k)}} \tag{5.8}$$

che è la somma delle varianze relative a *k* campioni infraannuali *tutti* di uguale numerosità $n_{A(k)}/k$. Dunque, usare una *k-ma* rilevazione annuale ha senso, ai fini di una possibile riduzione della somma delle varianze trimestrali, solo se per tale rilevazione si decide di intervistare un numero sufficientemente più alto di unità e se la varianza associata al *k-mo* sottoperiodo è sufficientemente più elevata rispetto alle precedenti (*k*-1).

Ad esempio, se *k*=4, le varianze dei primi tre periodi sono pari a 10, quella del quarto periodo è pari a 180 ed in tutto si possono intervistare 40 unità, tramite la (5.8) si avrebbe una varianza complessiva pari a 21. Se però la varianza annuale fosse pari a 200 e si optasse per la strategia basata sulla varianza (5.4), assegnando alle prime tre rilevazioni 5 unità campionarie ed alla quarta 25 unità, si avrebbe una varianza inferiore, pari a 20.

Anche in questo caso, si può derivare la relazione (5.7) rispetto a n_{Ak}, per ricavare la numerosità da assegnare all'ultima rilevazione in modo da rendere minima la somma delle *k* varianze, supponendo di aver fissato $n_{A(k)}$. Si può verificare come, ponendo:

$$\Omega=\left(\sum_{t=1}^{k-1}\sigma_{At}^2+\sum_{t=1}^{k-1}\sum_{t\neq r=1}^{k-1}\sigma_{Atr}\right) \tag{5.9}$$

è sufficiente risolvere rispetto a n_{Ak} la seguente equazione di secondo grado, con le avvertenze di cui alla nota 7:

$$\left[(k-1)\sum_{t=1}^{k-1}\sigma_{At}^2+k\Omega-\sigma_{Ak}^2\right]n_{Ak}^2-2\left[\left(\Omega-\sigma_{Ak}^2\right)n_{A(k)}\right]n_{Ak}+\left[\left(\Omega-\sigma_{Ak}^2\right)n_{A(k)}^2\right]=0\,. \tag{5.10}$$

Infine, va notato come un'ulteriore criterio utile per determinare la numerosità campionaria da assegnare all'ultima rilevazione si potrebbe basare sulla minimizzazione della somma – eventualmente ponderata – delle *k* varianze infraannuali e della stima strutturale, sulla base di un'espressione del tipo (5.1) qualora si aggiungesse la varianza della *k-ma* stima infraannuale.

Rimane evidentemente da verificare l'ammissibilità teorica di (*k*-1) stime infraannuali potenzialmente caratterizzate da bassa qualità (essendo basate su poche unità campionarie), a vantaggio di un'unica stima strutturale dotata di elevata qualità.

Nel prospetto seguente sono stati riepilogati gli stimatori e le relative varianze introdotti nei due paragrafi precedenti.

Prospetto 1 – Riepilogo di stimatori e varianze

Parametro	Varianza	
Stima strutturale: $\bar{y}_A$	$\frac{\sigma_A^2}{n_A}$	
Stima infraannuale: $\bar{y}_{At}$	$\frac{\sigma_{At}^2}{n_{At}} \rightarrow k\frac{\sigma_{Ak}^2}{n_{Ak}}$ se $n_{At} = \frac{n_{A(k)}}{k}$	
	Dipendenza	**Indipendenza**
Somma stime infraannuali: $\bar{y}_{A(k)} = \sum_{t=1}^{k} \bar{y}_{At}$	$\sum_{t=1}^{k} \frac{\sigma_{At}^2}{n_{At}}$	$\sum_{t=1}^{k} \frac{\sigma_{At}^2}{n_{At}}$
Somma varianze infraannuali: $\sum_{t=1}^{k} VAR(\bar{y}_{At})$	$k\frac{\sigma_A^2}{n_{A(k)}}$	$\sum_{t=1}^{k} \frac{\sigma_{At}^2}{n_{At}}$
$VAR\left(\sum_{t=1}^{k} \bar{y}_{At}\right) \leq VAR(\bar{y}_A)$ se:	$\frac{n_{A(k)}}{k} \geq n_A$	$\frac{n_{A(k)}}{k} \geq \frac{\sum_{t=1}^{k} \sigma_{At}^2}{\sigma_A^2} n_A$
Somma varianze con (k-1) infraannuali e una strutturale:	$(k-1)\sum_{t=1}^{k-1} \frac{\sigma_{At}^2}{n_{A(k)} - n_{Ak}} + \left\{ \frac{1}{n_{Ak}} \left[\sigma_{Ak}^2 - \frac{n_{A(k)} - k\, n_{Ak}}{n_{A(k)} - n_{Ak}} \left(\sum_{t=1}^{k-1} \sigma_{At}^2 + \sum_{t=1}^{k-1} \sum_{t \neq r=1}^{k-1} \sigma_{Atr} \right) \right] \right\}$	$2\sum_{t=1}^{k-1} \frac{\sigma_{At}^2}{n_{At}} + \frac{\sigma_A^2}{n_{Ak}}$

6. Possibili vincoli di coerenza

In un recente lavoro Ballin, Falorsi e Russo (2000) introducono e sviluppano formalmente alcuni concetti di coerenza relativamente ad indagini campionarie da cui si ricavano stime di cui si cerca di garantire la concordanza rispetto ad ammontari noti (od essi stessi stimati) derivati da altre fonti[7]. Rimandando a tale riferimento per ulteriori dettagli, in questo contesto si utilizzerà il concetto di *coerenza trasversale esterna*, dove la trasversalità dipende dal fatto che si focalizza l'attenzione sulla stima di un parametro di livello[8] (un ammontare, medio od assoluto), mentre la coerenza è esterna perché la (le) variabile (variabili) di vincolo non sono stimate tramite l'indagine stessa ma, più realisticamente, derivano da fonti esterne.

Si supponga di utilizzare l'indagine campionaria per stimare una funzione $f(\mathbf{y}_u)$, dove $\mathbf{y}_u$ è un vettore di N osservazioni relative all'universo u e f indica una generica funzione, e di osservare a livello campionario i vettori $\mathbf{y}_c$ e $\mathbf{x}_c$. La coerenza esterna si verifica se $f(\mathbf{x}_c) = f(\mathbf{x}_u)$, nell'ipotesi che il vettore sia noto da una fonte non campionaria o sia esso stesso una stima derivante però da un altro campione. Così, ad esempio, se y indica il fatturato e x il numero di addetti, in un'indagine campionaria finalizzata a stimare il fatturato medio $f(\mathbf{y}_u) = \sum_{i=1}^{N} y_i / N$ si può imporre il vincolo che la media campionaria $f(\mathbf{x}_c) = \sum_{i=1}^{n} x_i / n$ del numero di addetti sia pari

[7] Peraltro l'introduzione di vincoli comporta quasi certamente un peggioramento della qualità delle stime in termini di media quadratica dell'errore (Cicchitelli et al., op. cit.).

[8] A differenza del caso longitudinale in cui si vuole stimare un parametro di variazione tra due periodi.

alla media degli addetti nell'universo $f(\mathbf{x}_U) = \sum_{i=1}^{N} x_i / N$, oppure ad una stima sufficientemente affidabile derivante da un'altra indagine.

Se si pone $w_i = d_i \gamma_i$ il generico stimatore utilizzabile in questo contesto sarà dato dalla relazione:

$$T = \sum_{i=1}^{n} w_i y_i \tag{6.1}$$

dove d_i è il peso campionario diretto mentre γ_i è un suo correttore. Ad esempio, nel caso del disegno SRSWR si ha $d_i = 1/n$, così come in un contesto, come quello definito dalla (2.2), in cui si privilegia un'impostazione *model-based*. Se si suppone di utilizzare come vincolo un ammontare medio noto per l'intero universo la condizione di coerenza è esprimibile con le due relazioni:

$$\begin{cases} \underset{\gamma_i}{Min}\left\{ \sum_{i=1}^{n} p_i D(d_i \gamma_i, d_i) \right\} \\ \sum_{i=1}^{n} d_i \gamma_i x_i = \mu_x \end{cases} \tag{6.2}$$

dove *D* è una funzione di distanza, p_i è una costante nota - generalmente connessa alla dimensione dell'unità *i-ma* - e μ_x è il valore medio noto di una variabile ausiliaria *X* fortemente correlata a *Y* . In tale ottica, un caso classico è rappresentato dalla variabile ausiliaria addetti, generalmente molto correlata alla variabile fatturato oggetto di stima[9].

L'utilizzo effettivo di condizioni di coerenza nell'ambito di indagini infraannuali è particolarmente problematico, soprattutto a causa della carenza di ammontari medi noti da utilizzare come vincolo.

A titolo di esempio, la disponibilità su base mensile e trimestrale della base dati INPS sui contributi previdenziali versati ai lavoratori dipendenti potrebbe consentire di utilizzare come variabile di vincolo per la stima dei ricavi totali (ed eventualmente di altre variabili di conto economico) il numero di addetti dipendenti, monitorato da tale fonte informativa a cadenza trimestrale per il complesso delle imprese con lavoratori dipendenti regolarmente registrati.

Un caso particolare è poi rappresentato dall'indagine "rapida" sugli alberghi italiani[10], condotta dall'ISTAT in tre periodi dell'anno (Pasqua, Ferragosto e Natale) ed in cui la variabile di vincolo per la stima del numero dei pernottamenti è data dal numero dei posti letto.

Un'ulteriore soluzione operativa, come discusso anche in seguito, può consistere nel rilevare al tempo *t* l'ammontare di una variabile ausiliaria *X* riferito ad un periodo precedente "0" in relazione al quale si conosce l'ammontare vero di tale variabile, utilizzabile come vincolo nella seconda delle relazioni (6.2).

Nel contesto specifico la condizione di coerenza esterna comporta il ricorso allo stimatore:

$$T = \sum_{i=1}^{n} \frac{y_i \gamma_i}{n} \tag{6.3}$$

[9] Nella pratica, la correlazione si aggira sempre attorno a 0,8.
[10] Cfr. ISTAT (2001).

e si può definire la funzione di Lagrange:

$$\Phi(\gamma_i, \lambda) = \sum_{i=1}^{n} p_i \left(\frac{\gamma_i}{n} - \frac{1}{n} \right)^2 + \lambda \left(\sum_{i=1}^{n} \frac{x_i \gamma_i}{n} - \mu_x \right) \tag{6.4}$$

da cui si ricava la soluzione ottimale:

$$\gamma_i^* = 1 + \frac{n x_i}{p_i} \frac{(\mu_x - \bar{x})}{\left(\sum_{i=1}^{m} \frac{x_i^2}{p_i} \right)} \tag{6.5}$$

che per essere implementata in pratica non richiede la conoscenza di nessun indicatore di variabilità. In particolare, se i pesi *p* sono tutti uguali la soluzione si riduce al rapporto tra la media del modello e la media campionaria, per cui l'introduzione dei pesi (6.5) diviene ininfluente. Il limite principale di tale soluzione è quello di poter condurre a pesi negativi, il che è logicamente inammissibile[11].

Va notato come una possibile variante della procedura (6.2) consista nell'imporre questa condizione di vincolo:

$$\sum_{i=1}^{n} d_i \gamma_i x_{0i} = \bar{x}_0^* \tag{6.6}$$

in cui $\bar{x}_0^*$ rappresenta una stima sufficientemente affidabile del valore medio di *X* relativo ad un periodo di riferimento "0" precedente a quello d'interesse *t*, mentre x_{0i} è il valore assunto dalla variabile *X* al tempo 0 in corrispondenza della *i-ma* unità campionaria osservata nel campione al tempo *t*. In pratica, si può procedere in questo modo:

- si stima $\bar{x}_0^*$ sulla base di un campione di ampie dimensioni, oppure, come suggerito da Renssen e Nieuwenbroek (1997), tramite l'integrazione tra due fonti statistiche che consente di pervenire ad una stima più precisa (cfr. il paragrafo 8). Il periodo "0" potrebbe essere, ad esempio, il primo trimestre di un certo anno.
- Si determinano i pesi γ_{ti} sulla base del vincolo (6.6) seguendo una procedura di ottimizzazione analoga alla (2.2). A tale fine, occorre che le *n* unità inserite nel campione al tempo *t* appartengano tutte anche al campione utilizzato al tempo 0 (in modo da poterne conoscere le realizzazioni x_{0i}), oppure è necessario rilevare nell'indagine condotta al tempo *t* i valori di *X* relativi al periodo 0 (ad esempio, basterebbe chiedere nel questionario d'indagine relativo al secondo trimestre di un dato anno il numero di addetti del primo trimestre, scelto come periodo "0" e per il quale si suppone disponibile una stima molto affidabile del numero medio di addetti nell'universo).

In alternativa alla metodologia (6.2), si può utilizzare una funzione obiettivo diversa. Si fa ancora riferimento ad un generico disegno campionario basato su *n* osservazioni e caratterizzato da un modello di superpopolazione analogo a quello definito dalle relazioni (2.2), laddove si

[11] L'imposizione dell'ulteriore vincolo di non negatività di tutti i nuovi pesi è implicita nella soluzione ottenibile con algoritmi iterativi descritti, ad esempio, in Deville e Särndal (1992).

eliminino, per semplicità formale, i pedici riferiti all'anno (*A*) ed al periodo (*t*) e si supponga di riferirsi ad un unico periodo infraannaule dato (ad esempio un trimestre).

Si desidera determinare i nuovi pesi individuali tali che la media quadratica dell'errore di stima della media incognita μ del carattere *Y* sia minima, con il vincolo che la stima campionaria della media del carattere *X* riproduca esattamente il totale noto μ_x. In questo caso ci si basa, quindi, su questa condizione di coerenza:

$$\begin{cases} \underset{\gamma_i}{Min}\left\{ E\left[D\left(\sum_{i=1}^{n} d_i \gamma_i y_i, \mu \right) \right] \right\} \\ \sum_{i=1}^{n} d_i \gamma_i x_i = \mu_x \end{cases} \tag{6.6}$$

La funzione di Lagrange da minimizzare è data da:

$$\Phi(\gamma_i, \lambda) = E\left(\sum_{i=1}^{n} \frac{y_i \gamma_i}{n} - \mu \right)^2 + \lambda \left(\sum_{i=1}^{n} \frac{x_i \gamma_i}{n} - \mu_x \right) \tag{6.7}$$

e, in base al modello (2.2), la precedente espressione potrà essere scritta come:

$$\Phi(\gamma_i, \lambda) = \left[\mu^2 \left(\sum_{i=1}^{n} \frac{\gamma_i}{n} - 1 \right)^2 + \sigma^2 \sum_{i=1}^{n} \frac{\gamma_i^2}{n^2} \right] + \lambda \left(\sum_{i=1}^{n} \frac{x_i \gamma_i}{n} - \mu_x \right) \tag{6.8}$$

dove il termine in parentesi quadrate è la media quadratica dell'errore rispetto al modello. Si può dimostrare (cfr. l'appendice 2) che la soluzione ottimale è data dalla relazione:

$$\gamma_i^{**} = a + b\, x_i \tag{6.9}$$

dove si è posto:

$$a = -\frac{\mu^2}{\sigma^2} \Pi \qquad b = \frac{\mu^2 (\Pi - n) \sum_{i=1}^{n} x_i + n \sigma^2 \mu_x}{\sigma^2 \sum_{i=1}^{n} x_i^2} \qquad \Pi = \frac{n\mu^2 \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) + n\, \mu_x \sigma^2 \sum_{i=1}^{n} x_i}{n\mu^2 \left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) + \sigma^2 \sum_{i=1}^{n} x_i^2}.$$

Ovviamente, l'implementazione di tale procedura richiede la stima di un maggior numero di parametri rispetto alla soluzione (6.5), il che è possibile solo supponendo di avere già a disposizione una serie storica piuttosto lunga di osservazioni periodiche che possano consentire di stimare opportunamente medie e varianze del modello.

7. Utilizzo di serie storiche

La disponibilità di dati storici relativi anche ad un solo periodo precedente a quello di interesse consente di utilizzare stimatori parzialmente euristici, ma che potrebbero rivelarsi utili

in pratica, soprattutto nel breve periodo, supponendo che la struttura del modello (2.2) si mantenga stabile.

Nel dettaglio, si suppone di disporre di *k* stime infraannuali $\bar{y}_{Bt}$ relative ad un anno *B* precedente ad *A*, con riferimento al quale è anche noto il vero valore medio $\overline{Y}_B$ realizzatori nella popolazione oggetto di interesse.

Si desidera ricavare l'espressione dello stimatore infraannuale riferito a ciascun sottoperiodo *t* dell'anno *B* tale che la somma di tali stime infraannuali al variare di *t*, qualora utilizzabili nel corso di tale anno, avrebbe riprodotto il vero valore noto $\overline{Y}_B$. La ricerca può essere impostata imponendo la minimizzazione della somma degli scarti al quadrato tra ogni stimatore infraannuale "ottimale" ed il corrispondente stimatore infraannuale originario, dato dalla media campionaria. In simboli si ha la funzione di Lagrange:

$$\Phi(T_{Bt},\lambda)=\sum_{t=1}^{k}(T_{Bt}-\bar{y}_{Bt})^2+\lambda\left(\sum_{t=1}^{k}T_{Bt}-\overline{Y}_B\right) \tag{7.1}$$

da cui si ottiene facilmente il nuovo stimatore, corretto rispetto al modello (2.2), dato da:

$$T^{*}_{Bt}=\bar{y}_{Bt}+\frac{\left(\overline{Y}_B-\sum_{t=1}^{k}\bar{y}_{Bt}\right)}{k}. \tag{7.2}$$

Naturalmente, il precedente stimatore infraannuale non è praticamente utilizzabile con riferimento allo stesso anno *B*, essendo necessario conoscere, per tale anno, l'ammontare medio nell'universo della variabile d'interesse, ovviamente non disponibile in sede di stima infraannuale. E' però possibile traslare all'anno *A* il risultato ottimale (7.2) utilizzando la formula seguente:

$$T^{*}_{At}=\bar{y}_{At}+\frac{\left(\overline{Y}_B-\sum_{t=1}^{k}\bar{y}_{Bt}\right)}{k}\,f(\mathbf{y}_{At},\mathbf{y}_{Bt}) \tag{7.3}$$

dove la coppia di vettori $(\mathbf{y}_{At},\mathbf{y}_{Bt})$ contiene le osservazioni campionarie osservate fino al sottoperiodo *t* relative agli anni *A* e *B*. Una possibile scelta della funzione *f* è data dalla relazione:

$$f(\mathbf{y}_{At},\mathbf{y}_{Bt})=\frac{\sum_{r=1}^{t}\gamma_r\,\bar{y}_{Ar}}{\sum_{r=1}^{t}\gamma_r\,\bar{y}_{Br}} \tag{7.4}$$

che si basa, dunque, su un fattore correttivo di scala plausibile con il contesto di riferimento, dato dal rapporto tra la somma ponderata con i pesi γ delle medie infraannuali cumulate fino al sottoperiodo *t* d'interesse[12] e relative, rispettivamente, agli anni *A* e *B*.

Il precedente stimatore non è però corretto e la sua precisione potrebbe risultare molto variabile al variare del sottoperiodo considerato.

[12] Se tutti i coefficienti γ sono pari ad uno, si tratta del rapporto tra gli ammontari campionari degli anni *A* e *B* cumulati fino al sottoperiodo *t*.

Una soluzione alternativa consiste nello sfruttare la conoscenza del valore medio vero $\overline{Y}_B$ relativo al tempo *B* al fine di determinare dei pesi correttivi α_t, generalmente diversi per ogni sottoperiodo, tali che ciascuna delle nuove stime infraannuali $\alpha_t \bar{y}_{At}$ sia il più possibile simile alla stima non modificata $\bar{y}_{At}$, con il vincolo che la somma dei *k* stimatori infraannuali del tempo *B* modificati con i medesimi pesi riproduca esattamente $\overline{Y}_B$. In simboli:

$$\Phi(\alpha_t, \lambda) = \sum_{t=1}^{k} (\alpha_t \bar{y}_{At} - \bar{y}_{At})^2 + \lambda \left(\sum_{t=1}^{k} \alpha_t \bar{y}_{Bt} - \overline{Y}_B \right) \tag{7.5}$$

da cui si ricava la soluzione ottimale:

$$\alpha_t^* = 1 + \frac{\bar{y}_{Bt}}{\bar{y}_{At}^2} \left[\frac{\overline{Y}_B - \sum_{t=1}^{k} \bar{y}_{Bt}}{\sum_{t=1}^{k} \left(\frac{\bar{y}_{Bt}}{\bar{y}_{At}} \right)^2} \right] \tag{7.6}$$

e, quindi, il nuovo generico stimatore infraannuale sarà dato da:

$$T_{At}^* = \bar{y}_{At} + \left(\frac{\bar{y}_{Bt}}{\bar{y}_{At}} \right) \left[\frac{\overline{Y}_B - \sum_{t=1}^{k} \bar{y}_{Bt}}{\sum_{t=1}^{k} \left(\frac{\bar{y}_{Bt}}{\bar{y}_{At}} \right)^2} \right] \tag{7.7}$$

che rappresenta un caso particolare di (7.3) qualora si ponesse $f(\mathbf{y}_{At}, \mathbf{y}_{Bt}) = k \left(\frac{\bar{y}_{Bt}}{\bar{y}_{At}} \right) \Big/ \sum_{t=1}^{k} \left(\frac{\bar{y}_{Bt}}{\bar{y}_{At}} \right)^2$.
Un limite all'utilizzo della (7.6) consiste nel fatto che, come già visto nel paragrafo 6, alcuni dei pesi possono risultare negativi.

Si noti, infine, come nel caso in cui nella (7.5) si supponesse l'utilizzo di pesi tutti uguali tra loro, si otterrebbe la soluzione semplificata:

$$\alpha^* = \frac{\overline{Y}_B}{\sum_{t=1}^{k} \bar{y}_{Bt}} \tag{7.8}$$

Il limite principale delle procedure (7.6) e (7.8) riguarda l'entità della modifica indotta sugli stimatori diretti infraannuali a cui si applicano i pesi ottimali, che potrebbe appiattire eccessivamente l'informazione diretta derivata dalle misurazioni campionarie. Va però notato come l'effetto di eventuali errori nella stima di tali pesi risulterebbe più circoscritto se l'oggetto di interesse principale dell'indagine fosse la variazione del livello medio di *Y* intercorsa tra due periodi di riferimento: l'effetto dei pesi si annulla sempre sia che si focalizzi l'attenzione su variazioni tendenziali o congiunturali qualora si optasse per la (7.8), mentre il ricorso ai pesi

diversi (7.6) è ininfluente nel caso delle sole variazioni tendenziali, in cui si confrontano stime entrambe riferite al tempo t[13].

8. Uso congiunto di due stimatori

Sulla base della letteratura riguardante la stima per piccole aree (Rao e Choudry, 1995), la forma generale con cui si può esprimere uno stimatore ottenuto come sintesi lineare di due stimatori distinti dello stesso parametro incognito è data dalla relazione:

$$T_{At} = \alpha \bar{y}_{At} + (1-\alpha) S_{At} \qquad (8.1)$$

dove $\bar{y}_{At}$ esprime la stima campionaria diretta (non distorta) introdotta in precedenza, mentre S_{At} è un secondo stimatore della media infraannuale incognita, non necessariamente corretto.

Lo stimatore indiretto può derivare da diverse fonti, variabili in funzione del tipo di variabile oggetto di interesse:

a) fonti di tipo amministrativo. Tra questo assumono particolare rilevanza, in Italia, le già citate rilevazioni fiscali sui versamenti IVA infraannuali. Tali fonti possono essere supposte *indipendenti* dallo stimatore diretto.
b) Stime derivate da variabili rilevate nell'ambito della medesima indagine tramite cui si raccolgono i dati utilizzati per la stima diretta. Ad esempio, se *y* è il fatturato e *x* è il numero di addetti, è possibile ricavare una seconda stima indiretta del fatturato sulla base di un modello $y=f(x)$, soprattutto se il numero di risposte utili per la stima diretta del fatturato è molto più basso del numero di risposte pervenute con riferimento agli addetti. In questo caso le stima indiretta sarà necessariamente correlata con quella diretta.
c) Stime derivate dalla medesima indagine, ma riferite ad un dominio almeno in parte non coincidente con quello d'interesse. Ad esempio, si può essere interessati alla stima del fatturato medio per le attività di vendita di prodotti alimentari disponendo della stima riferita alla vendita di alimentari nei soli esercizi specializzati (complementare alla vendita di alimentari negli esercizi despecializzati[14]). Anche in questo caso le due stime risulteranno dipendenti. Se le stima indiretta deriva da una fonte diversa le stime saranno indipendenti.
d) Stime derivate da fonti addizionali ma riferite esattamente allo stesso dominio di interesse. Ad esempio, nel caso della stima dei flussi turistici interni al territorio nazionale l'ISTAT diffonde due stime, una derivata dall'indagine sull'offerta ed una dall'indagine sulla domanda[15]. Anche in questo caso le fonti si possono ragionevolmente supporre indipendenti.

Come noto, lo stimatore (8.1) è, in generale, uno stimatore distorto di μ_{At} - potendo risultare S_{At} distorto - sebbene potrebbe essere caratterizzato da una media quadratica dell'errore inferiore rispetto a quella dello stimatore diretto $\bar{y}_{At}$. Indicando con $E(S_{At})$ il valore atteso del secondo stimatore, l'errore quadratico medio dello stimatore composto sarà (Gismondi, 2001):

$$MQE(T_{At}) = VAR(T_{At}) + [BIAS(T_{At})]^2 =$$

[13] Si noti come l'uso della trasformazione $a+b\alpha_t$ sarebbe invece influente anche nel calcolo delle variazioni tendenziali.
[14] Gli esercizi despecializzati coincidono con buona approssimazione con gli esercizi della grande distribuzione commerciale (ipermercati, supermercati, grandi magazzini).
[15] Un estratto di tali informazioni è disponibile in ISTAT (2001).

$$= \alpha^2 VAR(\bar{y}_{At}) + (1-\alpha)^2 VAR(S_{At}) - 2\alpha(1-\alpha)COV(\bar{y}_{At}; S_{At}) + [E(S_{At}) - \bar{Y}_{At}]^2 . \quad (8.2)$$

La scelta del peso da assegnare allo stimatore diretto che rende minima la (8.2) è data da:

$$\alpha^* = \frac{VAR(S_{At}) - COV(\bar{y}_{At}; S_{At}) + [E(S_{At}) - \bar{Y}_{At}]^2}{VAR(\bar{y}_{At}) + VAR(S_{At}) - 2COV(\bar{y}_{At}; S_{At}) + [E(S_{At}) - \bar{Y}_{At}]^2} \quad (8.3)$$

da cui consegue che se i due stimatori sono linearmente indipendenti e la distorsione dello stimatore indiretto è nulla, il peso dipende esclusivamente dal rapporto tra le varianze dei due stimatori e crescerà al decrescere della varianza dello stimatore diretto.

9. Conclusioni prospettiche

Senza ombra di dubbio, la valutazione della qualità complessiva di un'indagine longitudinale – soprattutto se finalizzata alla stima di parametri con cadenza infraannuale – presenta una complessità superiore rispetto al caso, più tradizionale ed oggetto di studio in letteratura, di un'indagine "trasversale" (o *cross-section*) finalizzata alla stima di parametri riferiti ad uno specifico momento temporale. Il motivo principale sta nel fatto che nel caso delle indagini longitudinali occorre valutare due tipi di errore: il primo si riferisce alla stima di variazioni valutate su coppie di periodi definiti nell'ambito annuale (settimane, mesi, trimestri…) e generalmente catalizza gli sforzi maggiori del ricercatore per la finalità stessa delle indagini ripetute nel tempo; il secondo riguarda la necessità *logica* di una coerenza tra la stima di variazioni annuali desumibili cumulando le varie misurazioni infraannuali e la medesima stima ricavabile da un'indagine strutturale trasversale.

In proposito, si è visto come, almeno in linea teorica, le indagini infraannuali siano svantaggiate rispetto a quelle strutturali, nel senso che a parità di condizioni l'errore medio di stima di un ammontare annuale stimato cumulando stime infraannuali non sarà superiore all'errore medio di un'unica stima strutturale solo se in ognuno dei periodi infraannuali di osservazione si intervista lo stesso numero di unità intervistate nell'indagine strutturale. Ad esempio, nel caso di un'indagine annuale basata su 5.000 unità l'equivalenza qualitativa rispetto ad un'indagine mensile si avrà solo se in ciascuno dei 12 mesi si intervistano altrettante unità, per un totale di 60.000 interviste annue, che rappresentano un costo ben superiore tanto per chi realizza l'indagine quanto, ovviamente, per chi è chiamato a rispondere.

Una strada alternativa consiste nell'utilizzare, nell'ambito delle indagini infraannuali, tecniche di stima che incorporano, sotto varie forme, l'imposizione di uno o più vincoli di coerenza, sebbene con riferimento a tale possibilità vadano sottolineati due aspetti di fondamentale importanza:

- il concetto di coerenza deve preferibilmente riferirsi a variabili *diverse* da quella oggetto di stima, per quanto ad essa correlate;
- l'introduzione di tali vincoli può peggiorare la qualità delle stime infraannuali, che restano il principale obiettivo delle indagini congiunturali.

Dunque, la comprensibile necessità di molti utilizzatori – tra i più privilegiati dei quali si ricordano EUROSTAT e Banca Centrale Europea – di non rilevare discordanze significative tra dati congiunturali e strutturali sembra scontrarsi con l'attuale scarsità di strumenti metodologici mirati allo scopo. Inoltre tale richiesta, che sulla base di quanto appena ricordato rappresenta un'interpretazione almeno in parte impropria del concetto campionario di coerenza, costituisce

anche una *tautologia temporale*, dato che le indagini congiunturali, per loro natura, sono chiamate a fornire informazioni molto in anticipo rispetto a quelle strutturali, che d'altra parte vengono comprensibilmente considerate come il definitivo *benchmark* di riferimento.

Poiché, inoltre, l'ipotesi di intervistare sia nelle indagini strutturali, sia in quelle infraannuali il medesimo sottoinsieme di imprese è, almeno in Italia, del tutto irrealistica[16], la principale azione da intraprendere per ridurre le divergenze informative tra dati annuali ed infraannuali consiste in una attenta analisi circa la possibile distorsione degli stimatori utilizzati e le sue cause principali, da ricercarsi primariamente nell'autoselezione dei rispondenti e nell'impatto delle mancate risposte.

10. Appendice

10.1 Dimostrazione della (5.6)

Se si indica con $n_1 = (n_{A(k)} - n_{Ak})$ il numero di unità intervistate nel complesso delle prime (*k*-1) rilevazioni, lo stimatore del *k-mo* sottoperiodo sarà scrivibile come:

$$T = \bar{y}_A - \sum_{t=1}^{k-1} \bar{y}_{At} = \sum_{t=1}^{k} \sum_{i=1}^{n_{A(k)}} \frac{y_{Ati}}{n_1 + n_{Ak}} - \sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{y_{Ati}}{n_1} = \left[\sum_{i=1}^{n_1} \frac{y_{Aki}}{n_1 + n_{Ak}} - \sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{n_{Ak}\, y_{Ati}}{n_1 (n_1 + n_{Ak})} \right] +$$

$$+ \left[\sum_{t=1}^{k} \sum_{i=n_1+1}^{n_{Ak}} \frac{y_{Ati}}{n_1 + n_{Ak}} \right] = [T_1] + [T_2]$$

dove T_1 e T_2 sono indipendenti e $VAR(T\ ) = VAR(T_1) + VAR(T_2)$. Si ha poi:

$$VAR(T_1) = VAR\left[\sum_{i=1}^{n_1} \frac{y_{Aki}}{n_1 + n_{Ak}} \right] + VAR\left[\sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{n_{Ak}\, y_{Ati}}{n_1 (n_1 + n_{Ak})} \right] - 2COV\left[\sum_{i=1}^{n_1} \frac{y_{Aki}}{n_1 + n_{Ak}}; \sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{n_{Ak}\, y_{Ati}}{n_1 (n_1 + n_{Ak})} \right]$$

e si può verificare che valgono queste relazioni:

$$VAR\left[\sum_{i=1}^{n_1} \frac{y_{Aki}}{n_1 + n_{Ak}} \right] = \frac{n_1 \sigma_{Ak}^2}{(n_1 + n_{Ak})^2}$$

$$VAR\left[\sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{n_{Ak}\, y_{Ati}}{n_1 (n_1 + n_{Ak})} \right] = \frac{n_{Ak}^2}{n_1^2 (n_1 + n_{Ak})^2} \sum_{i=1}^{n_1} \left[\sum_{t=1}^{k-1} \sigma_{At}^2 + \sum_{t=1}^{k-1} \sum_{r \neq t=1}^{k-1} \sigma_{Atr} \right]$$

$$-2COV\left[\sum_{i=1}^{n_1} \frac{y_{Aki}}{n_1 + n_{Ak}}; \sum_{t=1}^{k-1} \sum_{i=1}^{n_1} \frac{n_{Ak}\, y_{Ati}}{n_1 (n_1 + n_{Ak})} \right] = -\frac{2\, n_{Ak}}{(n_1 + n_{Ak})^2} \sum_{t=1}^{k-1} \sigma_{Akt}$$

[16] Intervistare le medesime imprese dovrebbe comportare implicitamente la convergenza tra stime infraannuali ed annuali. Attualmente l'ISTAT sta cercando di operare in modo opposto, al fine di ridurre al minimo il numero di indagini in cui la medesima impresa è chiamata a collaborare.

$$VAR(T_2) = \sum_{i=n_1+1}^{n_{Ak}} VAR\left(\sum_{t=1}^{k} \frac{y_{Ati}}{n_1 + n_{Ak}}\right) = \frac{n_{Ak}}{(n_1 + n_{Ak})^2}\left[\sum_{t=1}^{k} \sigma_{At}^2 + \sum_{t=1}^{k}\sum_{r\neq t=1}^{k} \sigma_{Atr}\right].$$

Sommando le quattro precedenti identità si ottiene, dopo alcuni passaggi, la formula (5.6).

10.1 Dimostrazione della (6.9)

Uguagliando a zero la derivata prima della (6.8) si ricava facilmente questa espressione per il peso individuale γ_i:

$$\gamma_i = \frac{-\mu^2 n\left(\sum_{i=1}^{n} \frac{\gamma_i}{n} - 1\right)}{\sigma^2} - \frac{\lambda}{2}\left(\frac{n x_i}{\sigma^2}\right). \tag{10.1}$$

Sommando la (10.1) rispetto al pedice *i* si ricava poi:

$$-\frac{\lambda}{2} = \frac{\sum_{i=1}^{n} \gamma_i \left(1 + \frac{n\mu^2}{\sigma^2}\right) - \frac{n^2\mu^2}{\sigma^2}}{\frac{n}{\sigma^2}\sum_{i=1}^{n} x_i}. \tag{10.2}$$

Moltiplicando la (10.1) per x_i / n, sommando rispetto a *i*, ricordando il vincolo ed esplicitando rispetto al moltiplicatore λ si ottiene l'ulteriore identità:

$$-\frac{\lambda}{2} = \frac{\mu_x + \frac{\mu^2}{\sigma^2}\sum_{i=1}^{n} x_i\left(\sum_{i=1}^{n} \frac{\gamma_i}{n} - 1\right)}{\sum_{i=1}^{n} \frac{x_i^2}{\sigma^2}} \tag{10.3}$$

e, quindi, uguagliando la (10.2) alla (10.3) si ottiene, dopo alcuni passaggi, la relazione (6.9).

Bibliografia

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, 31-52, Franco Angeli, Milano.

BINDER D.A., DICK J.P. (1989), "Modelling and Estimation for Repeated Surveys", *Survey Methodology*, 15, 29-45.

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

CICCHITELLI G., HERZEL A., MONTANARI G.E. (1998), *Il campionamento statistico*, Il Mulino, Bologna.

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

COMMISSIONE EUROPEA (1998), "Regolamento n°1165/98 del Consiglio relativo alle statistiche congiunturali", *Gazzetta ufficiale delle Comunità europee*, 5 giugno, Bruxelles.

DEVILLE J.C., SARNDAL C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, Vol. 87, 376-382.

DRUDI I., FILIPPUCCI C. (1996), "Inference from Longitudinal Non-random Surveys: a Case Study", *Rivista di Statistica Applicata*, 1, Rocco Curto Editore, Milano.

EUROSTAT (2000), *Implementation of Council Regulation N°1165/98 Concerning Short-term Statistics – Definition of Variables (Version 2.2)*, Eurostat, Luxembourg.

GISMONDI R. (1998*a*), "Strategie ottimale per la stima di un rapporto in indagini panel", *Rivista di Statistica Applicata*, Vol.10, 459-477, Rocco Curto Editore, Milano.

GISMONDI R. (1998*b*), "The Impact of the Short-term Business Statistics Regulation", paper presentato al *13° Voorburg Group Meeting*, Roma.

GISMONDI R. (2001), "Integration Among Statistical Sources: Some Methodological Proposals", *Contributi*, 10, Istat, Roma.

HIDIROGLOU M.A., CHOUDHRY G.H., LAVALLEE P. (1991), "A Sampling and Estimation Methodology for Sub-Annual Business Surveys, *Survey Methodology*, Vol.17, 2, 195-210, Statistics Canada.

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

ISTAT (1989), *Manuali di tecniche d'indagine, vol. 4-5*, Istat, Roma.

ISTAT (2001), *Annuario statistico italiano*, Istat, Roma.

JONES R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys", *Journal of the Royal Statistical Society*, Series B, 42, 221-226.

KROESE A.H., RENSSEN R.H. (2000), "New Applications of Old Weighting Techniques; Constructing a Consistent Set of Estimates Based on Data from Different Sources", paper presentato alla *International Conference on Establishment Surveys*, Buffalo, 17-21 giugno, USA.

RAO J.N.K., CHOUDRY G.H. (1995), "Small Area Estimation: Overview and Empirical Study", in *Business Survey Methods* (Cox B.G. et al., Eds.), John Wiley & Sons, New York.

RENSSEN R.H., NIEUWENBROEK J.N. (1997), "Aligning Estimates for Common Variables in Two or More Sample Surveys", *Journal of the American Statistical Association*, vol.92, 437, 368-374.

SÄRNDAL C.E., SWENSSON B., WRETMAN J.H. (1992), *Model Assisted Survey Sampling*, Sprinter-Verlag, New York.

SINGH A.C., MOHL C.A. (1996), "Understanding Calibration Estimators in Survey Sampling", Survey Methodology, Vol.22, 2, 107-115.

SRINATH K.P., CARPENTER R.M. (1995), "Sampling Methods for Repeated Business Surveys", *Business Survey Methods*, 171-183, John Wiley & Sons, New York.

YANSANEH I.S., FULLER A.W. (1998), "Optimal Recursive Estimation for Repeated Surveys", *Survey Methodology*, vol.24, 1, 31-40, Statistics Canada.