Metodi per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di variazioni

Roberto Gismondi (*)

(*) ISTAT - Servizio SCO

ESTRATTO

Il documento propone alcune semplici procedure per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di un indice di variazione relativo ad una variabile quantitativa *Y*, che si supporrà generalmente riferita a dati d'impresa. Nonostante tematiche quali l'individuazione di eventuali errori di misura e di valori *outlier* in indagini quantitative siano ampiamente trattate in letteratura, non esistono molte valutazioni metodologiche ed applicazioni empiriche concepite *ad hoc* in presenza di misurazioni ripetute nel tempo sullo stesso insieme di unità. Ci si soffermerà soprattutto sulle possibili modalità d'intervento – direttamente sui microdati o sui pesi originari associati alle unità del campione disponibile – tese a migliorare la precisione della stima di una variazione supponendo che gli *outlier* non siano (del tutto) eliminabili dalla base dati disponibile (ad esempio, perché i valori anomali potrebbero essere associati ad unità autorappresentative non ricontattabili e che non possono essere escluse dai calcoli). D'altra parte, il problema preliminare legato alla individuazione dell'intervallo di accettazione sarà affrontato solo in parte, risultando particolarmente legato alle specificità dell'indagine e comunque già oggetto di diverse trattazioni specifiche. Dopo una premessa generale, nel secondo paragrafo verrà introdotta l'espressione generale dello stimatore di un indice di variazione, evidenziando il legame esistente tra scelta dello stimatore, disegno campionario e peso campionario associato ad ogni unità. Nel terzo paragrafo verranno commentate alcune strategie operative per il trattamento degli *outlier*. Il paragrafo 4 è dedicato ai metodi basati sulla correzione degli indici *outlier* senza alterazione dei pesi, mentre il paragrafo 5 ai metodi basati sulla modifica dei pesi senza correzione degli indici *outlier*. La trattazione si conclude con una applicazione empirica tesa a confrontare l'efficacia delle diverse metodologie.

1. Il trattamento dei dati anomali nella stima di un indice di variazione

Si supponga di operare nel contesto di un'indagine campionaria finalizzata al calcolo della variazione dell'ammontare di una certa variabile quantitativa *Y* - che per semplicità si supporrà non negativa - intercorso tra il tempo *t* ed un certo tempo (*t*-*k*) scelto come base di riferimento. Si farà preferibilmente riferimento ad indagini in cui la suddetta variabile esprime un indicatore economico riferito a dati d'impresa (ricavi, valore aggiunto, costi, investimenti, occupazione), sebbene molte delle considerazioni proposte nel prosieguo siano adattabili a contesti operativi più ampi. Dopo aver estratto un campione *S*, all'interno di un certo strato - che si supporrà d'ora in poi prefissato - non di rado vengono osservati alcuni valori relativi alle variazioni individuali di *Y* particolarmente "anomali", in quanto sensibilmente diversi dalla media (e/o dalla mediana) della distribuzione di frequenze empirica relativa allo strato stesso, in altri termini valori posizionati su una delle due code di tale distribuzione.

L'influenza di tali valori estremi sulla stima della variazione media riferita allo strato suddetto, soprattutto se associati ad unità caratterizzate da un ordine di grandezza di *Y* elevato, può risultare incontrollabile senza il ricorso ad un adeguato programma di controllo di validità dei dati elementari (*editing*). Si pone conseguentemente il problema, ampiamente dibattuto in letteratura[1], di come trattare le osservazioni effettivamente identificate come *outlier*.

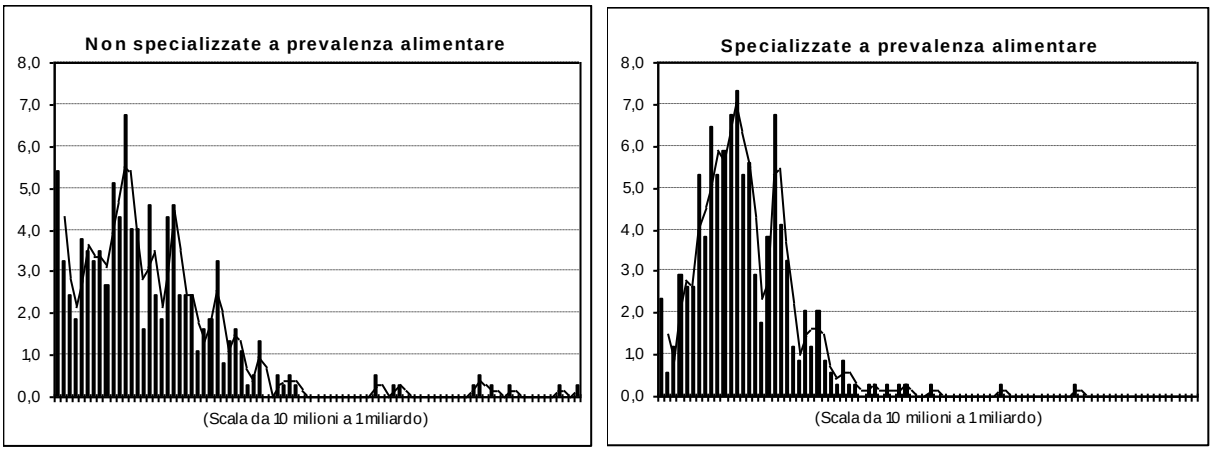
Come esempio, i due grafici seguenti riportano le distribuzioni di frequenze relative alle imprese fino a due addetti operanti nel commercio al dettaglio prevalentemente alimentare in sede fissa nel 1997. Sono stati considerati i dati campionari - disponibili sulla base dell'indagine mensile sulle vendite al dettaglio condotta correntemente dall'ISTAT - distintamente per le imprese non specializzate a prevalenza alimentare (in maggioranza minimercati, piccoli supermercati, *discount* ed altri grandi negozi a prevalenza alimentare) e per quelle specializzate a prevalenza alimentare (negozi tradizionali e di prossimità) sulla base, rispettivamente, di 371 e 468 imprese.

La scelta di tale esempio è dovuta essenzialmente al fatto che il comparto commerciale al dettaglio è notoriamente caratterizzato da una elevata eterogeneità imprenditoriale, anche all'interno di strati particolarmente circoscritti come quelli esaminati: esso risulta, quindi, particolarmente adatto ad evidenziare come anche una preventiva stratificazione piuttosto

[1] Si ricorda, tra tanti, il noto saggio di Fellegi e Holt (1976).

dettagliata (imprese fino a due addetti e classificazione ATECO a 3 e 4 cifre[2]) non comporti necessariamente l'assenza di *outlier* come quelli nella coda di destra delle due densità empiriche.

Grafico 1.1 - Distribuzione di frequenze del fatturato 1995 per alcune imprese commerciali al dettaglio fino a due addetti (ATECO 52.11 e 52.2)



Nota: elaborazioni su dati ISTAT.

Entrambe le distribuzioni sono chiaramente caratterizzate da asimmetria positiva, sebbene quella relativa alle imprese non specializzate (generalmente più omogenee appartenendo in maggioranza alla grande distribuzione organizzata) risulti più uniforme, almeno fino a alle frequenze corrispondenti ai 350 milioni circa. Sono evidenti le imprese posizionate sulla coda di destra di entrambe le distribuzioni, comunque più numerose tra le imprese non specializzate, il cui peso economico è generalmente assai superiore a quello delle imprese alimentari specializzate di uguale dimensione.

In questo caso l'anomalia non è dovuta alla presenza di errori di misurazione, bensì al fatto che alcune imprese dello strato presentano connotati economici tipici delle imprese appartenenti alle classi di addetti superiori.

L'operazione generalmente più immediata da compiere in presenza di valori anomali consiste proprio nel ricontattare il rispondente per accertarsi dell'esattezza del dato.

Se non è possibile ricontattare l'impresa, o se questa conferma come vero il dato apparentemente anomalo, la soluzione apparentemente più semplice del problema del trattamento degli *outlier*, piuttosto diffusa in pratica e consistente nel non considerare affatto le unità *outlier* nei calcoli (in altri termini di assegnarle un peso nullo) è spesso rischiosa.

[2] Per dettagli sulla classificazione delle attività economiche ATECO si rimanda a Gismondi (1998).

Infatti all'interno dello strato potrebbero ricadere poche unità campionarie (ed eventualmente poche unità nello stesso universo), alcune delle quali praticamente autorappresentative (non esistono nell'universo unità ad esse "simili", per cui queste devono appartenere al campione effettivo con probabilità uno), e quindi eliminarle dalla base dati disponibile per i calcoli successivi appare quanto mai sconsigliabile.

Una seconda procedura, consistente nel post-stratificare il campione sulla base della presenza di *outlier* in certi strati originari, richiederebbe comunque almeno una stima del numero di unità nell'universo dalle caratteristiche assimilabili a quelle delle unità *outlier* osservate nel campione, il che è spesso impossibile per carenza di informazioni al riguardo e/o la necessità di produrre le stime entro un intervallo temporale assai ristretto (ad esempio, ciò accade di frequente nel caso di indagini mensili). Le principali alternative alle soluzioni precedenti sono nel complesso classificabili in due gruppi di tecniche:

1. tecniche tese alla correzione del dato elementare ritenuto anomalo;
2. tecniche che non alterano il dato elementare ma correggono (generalmente diminuendolo) il peso con cui tale dato entra nella procedura di stima della variazione media complessiva dello strato di riferimento.

In generale, non esiste un approccio al problema che risulti sempre preferibile, dipendendo la scelta dal grado di conoscenza del fenomeno studiato, dall'ammontare degli interventi sui microdati e dalle stesse finalità dell'indagine (si veda in proposito il paragrafo 3).

In questo contesto si cercherà di illustrare alcune semplici metodologie finalizzate al miglioramento della precisione della stima campionaria di una variazione in presenza di *outlier* generalmente *non eliminabili* dalla base dati disponibile.

Più precisamente, seguendo Kovar e Winkler (1996), definiremo come *outlier* una unità che, con riferimento ad una variabile Y di interesse, presenta un valore situato alla coda di una distribuzione empirica di valori relativi ad unità ad essa teoricamente simili. Generalmente questo criterio di similitudine deriva da una preventiva suddivisione delle unità in strati in merito a cui si ipotizzano uguale media, omoschedasticità ed incorrelazione. Non sarà preso in esame il caso delle unità *inlier*, che cioè presentano un valore di Y invariante per più misurazioni successive, generalmente dovuto alla replicazione del valore dichiarato in un questionario precedente. Gli *outlier* possono essere distinti in:

1. *outlier non rappresentativi*: si tratta di valori anomali a causa di veri e propri errori in fase di compilazione del questionario[3]. Un caso classico è costituito dall'errore nell'unità di misura utilizzata per la risposta (ad esempio, lire invece di migliaia di lire, per cui i valori dichiarati dovrebbero essere divisi per mille). La loro non rappresentatività va intesa con riferimento alle unità della popolazione non incluse nel campione, perchè non contribuiscono alla variabilità campionaria fornendo informazioni su di esse. Si tratta di veri e propri errori che occorrerebbe individuare e correggere a monte;
2. *outlier rappresentativi*: si tratta di valori anomali non dovuti ad errori di misurazione, bensì ad eventi relativi all'unità di riferimento non (del tutto) valutabili sulla base delle informazioni disponibili su di essa. Si tratta comunque di osservazioni rappresentative di un certo numero di unità della popolazione non incluse nel campione, di cui generalmente non si conosce l'ammontare.

Nel prosieguo, pur facendo maggiormente riferimento alla seconda tipologia - il che escluderebbe il ricorso a correzioni dei dati sospetti - si utilizzerà il termine *outlier* nella sua massima generalità, sia perchè in pratica non è sempre possibile distinguere con certezza tra le due tipologie suddette, sia perchè possono essere individuate tecniche di stima di un indice di variazione in cui il trattamento delle unità anomale rimane valido per entrambe le situazioni. Si escluderà comunque il caso dei veri e propri errori riconducibili all'unità di misura con cui è stato espresso il fenomeno, che si supporranno risolvibili con ovvie correzioni dei microdati antecedenti alle valutazioni di cui in seguito.

Inoltre non verrà affrontato esplicitamente il problema di *come* individuare le unità *outlier*, sebbene tale aspetto venga implicitamente sfiorato in diverse circostanze (in particolare nel paragrafo 4.2); per dettagli su tale problema si rimanda a Hidiroglou e Berthelot (1986), Hennig (1998), Pizzi e Pellizzari (1998), Thompson e Sigman (1998). L'attenzione verrà piuttosto concentrata sul trattamento statistico delle unità *outlier* in sede di stima di un indice di variazione, senza alcuna ipotesi circa la persistenza nella base dati di possibili *outlier* non rappresentativi e supponendo di aver già rimosso gli eventuali errori di misura, sebbene tale fase non sia indispensabile ai fini della validità di gran parte delle procedure proposte in seguito.

Infine si supporrà, per semplicità, di disporre di una base dati longitudinale basata

sullo stesso insieme di unità intervistate al tempo *t*, quindi di un *panel* senza rotazioni, ipotesi non irrealistica se si fa riferimento a rilevazioni esaustive di tipo annuale o mensili svolte in un arco temporale di riferimento non superiore ai 12-18 mesi.

Si ricorda, infine, che in ISTAT risulta attualmente disponibile il software GEIS (Barcaroli e Luzi, 1995), elaborato da Statistics Canada, finalizzato alla individuazione delle osservazioni *outlier* ed alla loro sostituzione con stime prodotte con una varietà di metodi piuttosto comuni in pratica. Sono però scarse le applicazioni finora condotte con riferimento ad indagini longitudinali sulle imprese. E' poi in fase di valutazione l'efficacia del software SPEER, prodotto dal *Bureau of Census* statunitense, che si differenzia da GEIS per la diversa formulazione algebrica dei vincoli imposti per determinare le soglie di accettazione[4].

Nel paragrafo seguente verrà introdotta l'espressione generale dello stimatore di un indice di variazione, evidenziando il legame esistente tra scelta dello stimatore, disegno campionario e peso campionario associato ad ogni unità. Nel paragrafo 3 verranno commentate alcune strategie operative per il trattamento degli *outlier*. Il paragrafo 4 è dedicato ai metodi basati sulla correzione degli indici *outlier* senza alterazione dei pesi, mentre il paragrafo 5 ai metodi basati sulla modifica dei pesi senza correzione degli indici *outlier*. La trattazione si conclude con una applicazione empirica tesa a confrontare l'efficacia delle diverse metodologie.

2. Definizione generale dello stimatore di un indice di variazione

Siano date queste definizioni:

Y_{ti} : valore assunto dalla variabile quantitativa Y al tempo t sull'unità *i-ma* della popolazione;

Y_t : ammontare complessivo, riferito all'intera popolazione, della variabile quantitativa Y al tempo t;

Y_{Sti} : valore assunto dalla variabile quantitativa Y al tempo t sull'unità *i-ma* appartenente al campione S;

Y_{St} : ammontare complessivo riferito al campione S della variabile

[3] Per una rassegna si veda Fuller (1987).
[4] Mentre GEIS si basa su vincoli espressi in forma di equazioni e/o disequazioni lineari, SPEER si basa su rapporti.

quantitativa Y al tempo t;

$I_{Sti/k} = Y_{Sti} / Y_{S,t-k,i}$: rapporto tra i valori di Y riferiti ai tempi t e (t-k) calcolato sull'unità *i-ma* del campione S, anche detto "variazione tendenziale individuale";

N_t : numero di unità della popolazione appartenenti all'universo al tempo t;

n_{St} : numero di unità del campione S riferito al tempo t;

π_{Sti} : probabilità di inclusione della unità *i-ma* nel campione S estratto al tempo t;

p_{Sti} : probabilità individuale di estrazione della *i-ma* unità

$I_t = Y_t / Y_{t-k}$: rapporto tendenziale, riferito all'intera popolazione, tra i valori di Y riferiti ai tempi t e (t-k), *incognito ed oggetto di stima*.

$I_{St} = Y_{St} / Y_{S,t-k}$: rapporto tendenziale, riferito al campione S, tra i valori di Y riferiti ai tempi t e (t-k).

Si può poi supporre quanto segue:

1. di conoscere l'ammontare Y_{t-k} : tale ipotesi è plausibile supponendo che (t-k) si riferisca ad un certo anno base preso come riferimento per il calcolo della variazione[5];
2. di conoscere $Y_{S,t-k,i}$ per ogni unità i del campione S estratto al tempo t: ciò è plausibile supponendo che nei due tempi il campione resti lo stesso, oppure che ad ogni unità intervistata al tempo t si chieda anche l'ammontare di Y riferito al tempo (t-k).

Se si indica allora con T_{St} uno stimatore del suddetto rapporto incognito, sulla base della espressione generale dello stimatore di Horvitz e Thompson esso assumerà varianza minima nell'insieme degli stimatori lineari e corretti se sarà genericamente scrivibile in questa forma:

$$T_{St} = \frac{\sum_{i=1}^{n_{St}} \frac{Y_{Sti}}{\pi_{Sti}}}{\sum_{i=1}^{N_{t-k}} Y_{t-k,i}} = \sum_{i=1}^{n_{St}} \left(\frac{Y_{Sti}}{Y_{S,t-k,i}} \right) \left(\frac{1}{\pi_{Sti}} \right) \frac{Y_{S,t-k,i}}{Y_{t-k}} = \sum_{i=1}^{n_{St}} \left(I_{Sti/k} \right) \frac{Y_{S,t-k,i}}{\pi_{Sti} Y_{t-k}} . \qquad [2.1]$$

Come caso particolare della espressione [2.1], se si optasse al tempo t per un disegno campionario casuale semplice (*PPS*), dove la variabile dimensionale è posta uguale a $Y_{t-k,i}$ si avrebbero le due relazioni seguenti:

$$\begin{cases} p_{Sti} = \dfrac{Y_{t-k,i}}{Y_{t-k}} \\ \pi_{Sti} = n_{St}\, p_{Sti} \end{cases} \qquad [2.2]$$

e di conseguenza lo stimatore assumerebbe questa forma semplificata:

$$T_{St} = \sum_{i=1}^{n_{St}} \left(\frac{I_{Sti/k}}{n_{St}} \right); \qquad [2.3]$$

si tratterebbe, quindi, alla semplice media aritmetica non ponderata degli n_{St} rapporti tendenziali individuali calcolabili sul campione al tempo t, e quindi lo stimatore del rapporto tra gli ammontari di Y riferiti ai tempi t e (t-k) in questo caso può essere ricondotto alla media dei rapporti tendenziali individuali tra gli ammontari di Y nei due tempi riferiti a ciascuna delle n_{St} unità inserite nel campione. Per l'effettiva implementazione del disegno *PPS* al tempo t, qualora non si conoscessero i valori di Y relativi ad ogni singola unità della popolazione al tempo (t-k), si può utilizzare una variabile Z correlata a Y secondo la relazione $Y_{ti} \cong \rho_t Z_{ti} + \varepsilon_{ti}$, ponendo:

$$\begin{cases} p^*_{Sti} = \dfrac{Z_{t-k,i}}{Z_{t-k}} \\ \pi^*_{Sti} = n_{St}\, p^*_{Sti} \end{cases} \qquad [2.4]$$

dove Z_{t-k} indica l'ammontare complessivo di Z nella popolazione al tempo (t-k). Se Y indica una variabile di *output* (fatturato, valore della produzione, valore aggiunto), Z può essere ragionevolmente data dal numero degli addetti relativi all'unità di analisi.

Un secondo caso particolare della [2.1] è ottenibile ponendo:

[5] In alternativa è generalmente possibile stimare le grandezze riferite all'anno base utilizzando le informazioni raccolte con alcune tra le principali indagini strutturali sulle imprese condotte dall'ISTAT, quali le indagini PMI (imprese fino a 19 addetti, campionaria) e SCI (imprese con almeno 20 addetti, totalitaria).

$$\begin{cases} p_{Sti} = 1/N_t \\ \pi_{Sti} = n_{St}\, p_{Sti} \end{cases} \qquad [2.5]$$

ossia utilizzando un disegno casuale semplice (*SRS*), e di conseguenza lo stimatore assumerebbe questa ulteriore forma semplificata:

$$T_{St} = \frac{\left(\frac{N_t}{n_{St}}\right)\sum_{i=1}^{n_{St}} Y_{Sti}}{Y_{t-k}}; \qquad [2.6]$$

si tratterebbe, quindi, del semplice rapporto tra la stima dell'ammontare complessivo di *Y* al tempo *t* - ottenuta moltiplicando per il relativo coefficiente di espansione l'ammontare campionario - e l'ammontare complessivo di *Y* al tempo (*t-k*). Vale la pena di notare che, qualora non fosse noto l'ammontare vero Y_{t-k}, al denominatore della espressione che definisce T_{St} basterebbe operare la sostituzione seguente:

$$\sum_{i=1}^{n_{S,t-k}} \frac{Y_{S,t-k,i}}{\pi_{S,t-k,i}} \qquad \text{in luogo di} \qquad Y_{t-k},$$

dove potrebbe ragionevolmente valere la semplificazione $\pi_{S,t-k,i} = \pi_{Sti}$.

Gli stimatori [2.3] e [2.6] sono comunemente definiti, rispettivamente, "media di rapporti" e "rapporto di ammontari", e le considerazioni precedenti hanno evidenziato sotto quali disegni campionari essi risultano coincidenti con lo stimatore di Horvitz e Thompson e, dunque, in quali contesti operativi ne può risultare raccomandabile l'utilizzo. In pratica può però accadere che, indipendentemente dal disegno campionario adottato, la "media di rapporti" si riveli comunque più efficiente del "rapporto di ammontari", pur essendo affetta da distorsione. Nell'esempio seguente[6] tale evidenza viene verificata ricorrendo ad un approccio modellistico, supponendo cioè che ad ognuna delle dieci unità considerate sia associata una superpopolazione di cui, tramite le cinque osservazioni disponibili e riferite ad altrettanti anni consecutivi, si possono stimare le medie μ(i) e le varianze σ^2(i)[7]. Dalle relazioni [2.3] e [2.6]

[6] Si tratta di dieci imprese operanti nella grande distribuzione alimentare, i cui ricavi annui riferiti al periodo 1991-1995 sono stati espressi in decine di miliardi.
[7] Per una trattazione *model-based* esaustiva si veda il paragrafo 5.2.

si ricavano facilmente queste formule per le medie e le varianze:

$$E(T_{t(2.3)}) \cong \overline{Y}_5 \sum_{i=1}^{10} \frac{1}{10 Y_{0i}} \quad E(T_{t(2.6)}) \cong \frac{\overline{Y}_5}{\overline{Y}_0} \quad V(T_{t(2.3)}) \cong \sum_{i=1}^{10} \frac{V(Y_{5i})}{10^2 (Y_{0i})^2} \quad V(T_{t(2.6)}) \cong \sum_{i=1}^{10} \frac{V(Y_{5i})}{10^2 (\overline{Y}_0)^2}.$$

Tabella 2.1 - Un confronto tra gli stimatori [2.3] e [2.6]

i	Y_{0i}	Y_{1i}	Y_{2i}	Y_{3i}	Y_{4i}	Y_{4i}/ Y_{0i}	V(i)*
1	9	12	13	17	20	2,222	15,0
2	8	11	12	15	19	2,375	14,0
3	8	7	13	17	25	3,125	43,2
4	12	19	17	19	23	1,917	12,8
5	10	15	20	24	28	2,800	40,6
6	9	12	14	18	24	2,667	27,0
7	12	20	26	29	35	2,917	61,8
8	15	21	15	22	32	2,133	38,8
9	12	26	31	35	39	3,250	87,4
10	12	24	25	30	37	3,083	67,4
Media	10,7	16,7	18,6	22,6	28,2		

	Media di rapporti (2,649)	**Rapporto di ammontari (2,636)**
VAR	0,010	0,036
$BIAS^2$	0,011	0,000
MQE	0,022	0,036

(*) V=varianza.

Questo caso concreto presenta alcune peculiarità, peraltro frequenti qualora si studino variabili economiche riferite ad imprese:

- si verifica una forte crescita per tutte le unità passando dal tempo 0 al tempo 4;
- la varianza "longitudinale" delle unità cresce al crescere della dimensione di *Y*;
- le variazioni individuali più elevate si verificano prevalentemente in corrispondenza delle unità con i valori di *Y* più elevati.

Nonostante i due stimatori conducano a stime assai simili (2,649 per lo stimatore [2.3] basato sulla media di rapporti e 2,636 per lo stimatore [2.6] basato sul rapporto di ammontari), l'errore quadratico medio del primo è sensibilmente più basso rispetto al secondo (0,022 contro 0,036), sebbene la media di rapporti sia affetta da una distorsione quadratica pari a 0,011.

In realtà, le posizioni [2.3] e [2.6] costituiscono solo due casi particolari della più ampia gamma di possibili disegni campionari implementabili in sede di stima di un rapporto. Un disegno più generale è dato da:

$$\begin{cases} p_{Sti} = \dfrac{X_{t-h,i}}{X_{t-h}} \\ \pi_{Sti} = n_{St}\, p_{Sti} \end{cases} \qquad [2.7]$$

dove X_{t-k} indica l'ammontare complessivo di X nella popolazione al tempo (t-k); la variabile X dovrebbe risultare correlata positivamente con Y, si tornerà nel paragrafo seguente e h è un intero non negativo minore od uguale a k. Ad esempio, se Y indica i ricavi totali d'impresa, t=1998 e (t-k)=1991, si può supporre che in corrispondenza di quest'ultimo anno sia disponibile una stima per l'intero universo delle imprese in esame sulla basa dell'ultimo Censimento dell'Industria e dei Servizi. La variabile X potrebbe essere costituita dalle dichiarazioni fiscali delle imprese, generalmente disponibile con un certo ritardo rispetto al periodo di riferimento, per cui si potrebbe porre (t-h)=1997. Si noti come in questo caso, essendo $h<k$, è preferibile utilizzare una variabile ausiliaria X diversa da Y, perchè riferita ad un tempo assai più vicino rispetto al 1998 di quanto non sia il 1991[8]. Partendo dalla ultima relazione della [2.1] si avrà dunque che:

$$T_{St} = \sum_{i=1}^{n_{St}} (I_{Sti/k}) \frac{Y_{S,t-k,i}}{\pi_{Sti} Y_{t-k}} = \sum_{i=1}^{n_{St}} \left(\frac{I_{Sti/k}}{n_{St}} \right) \frac{\dfrac{Y_{S,t-k,i}}{Y_{t-k}}}{\dfrac{X_{S,t-h,i}}{X_{t-h}}} = \left(\frac{N_{t-h}}{N_{t-k}} \right) \sum_{i=1}^{n_{St}} \frac{I_{Sti/k}}{n_{St}} \left[\frac{\dfrac{Y_{S,t-k,i}}{\left(\dfrac{Y_{t-k}}{N_{t-k}} \right)}}{\dfrac{X_{S,t-h,i}}{\left(\dfrac{X_{t-h}}{N_{t-h}} \right)}} \right].$$

Quindi in generale lo stimatore di un rapporto riferito alla variabile Y ed al tempo t è scrivibile come prodotto tra l'indice di variazione del tempo (t-h) rispetto al tempo (t-k) del numero di unità appartenenti allo strato esaminato nella popolazione (che è chiaramente pari ad uno se k=h) e la media aritmetica semplice di n_{St} addendi, ognuno dei quali è uguale al prodotto tra l'indice di variazione individuale ed un fattore esprimibile come rapporto tra:

- l'intensità relativa di $Y_{t-k,i}$ rispetto all'intensità media di Y in (t-k) nella popolazione;
- l'intensità di rispetto all'intensità media di $X_{t-h,i}$ nella popolazione. Si noti in proposito

[8] Sempre che i ricavi riferiti al 1991 non risultino comunque più correlati ai ricavi del 1998 rispetto alle dichiarazioni fiscali 1997.

come, in pratica, i valori della variabile *X* al tempo (*t-h*) potrebbero non essere noti per tutte le unità della popolazione, nel qual caso si può utilizzare una stima campionaria di tale intensità media, senza alterare il significato intrinseco della precedente formulazione, sebbene tale procedura renda più imprecisa l'implementazione del disegno *PPS*.

Alla luce di quanto visto, è infine immediato semplificare ulteriormente la formulazione generale dello stimatore di un rapporto tra ammontari tramite la posizione:

$$T_{St} = \sum_{i=1}^{n_{St}} (I_{Sti/k})\, D_{Sti} \,. \qquad [2.8]$$

dove in generale si avrà che $D_{Sti} = f(N_{t-k}; N_{t-h}; n_{St}; Y_{S,t-k,i}; X_{S,t-h,i})$, e la somma di tali pesi rispetto al totale delle unità campionarie non è necessariamente vincolata ad uno.

Mentre in questo paragrafo i pesi *D* sono derivabili in funzione del disegno campionario utilizzato, nel paragrafo 5 si determineranno altre possibili scelte di pesi tali da migliorare la precisione delle stime in presenza di *outlier*, sulla base di valutazioni in tutto od in parte svincolate dal disegno. In effetti, la ponderazione degli indici individuali con pesi che non sono necessariamente coincidenti con quelli ottimali di Horvitz e Thompson - che risultano ottimali *in media al variare del campione estratto* - rappresenta una soluzione operativa prettamente mirata ad ottimizzare la precisione della stima con riferimento *allo specifico campione disponibile al tempo t*, soprattutto qualora fosse affetto da alcune osservazioni chiaramente anomale, il cui effetto distorcente potrebbe risultare assai rilevante se si ricorresse a stimatori basati sul disegno campionario del tipo [2.8].

3. Possibili strategie per il trattamento dei microdati in presenza di *outlier*

Prima di procedere, è necessario richiamare all'attenzione la coesistenza di due metodologie di base con cui è possibile trattare il problema delle unità *outlier* nell'ambito di indagini volte alla stima di un indice di variazione secondo la formula generale [2.8].

Una prima famiglia di tecniche ha come fine la correzione del dato di base Y_{Sti}. Una alterazione del dato di base equivale ad una trasformazione di tale dato secondo il "fattore"

w_{1i}, in modo che si utilizzerà:

$$Y_{Sti}\, w_{1i} \qquad \text{in luogo del valore originario} \qquad Y_{Sti}\,.$$

Una seconda famiglia di tecniche prevede invece, senza una alterazione del dato di base, una correzione del peso originario con cui una unità anomala dovrebbe entrare nel meccanismo di stima di una variazione, secondo quanto riportato nella formula [2.8], tramite il "fattore" w_{2i}, comporterà l'utilizzo del nuovo peso complessivo:

$$D_{Sti}\, w_{2i} \qquad \text{in luogo del valore originario} \qquad D_{Sti}\,.$$

A priori non è in genere possibile stabilire quale delle due tipologie di procedure sia preferibile: va però sottolineato come il ricorso alla prima tipologia implichi necessariamente la possibilità di identificare le unità effettivamente *outlier* e risulti consigliabile se tra le finalità dell'indagine c'è anche quella di fornire una base dati coerente per gli utenti finali.

In generale, rispetto al secondo gruppo di tecniche - certamente più cautelativo riguardo al delicato aspetto del trattamento dei microdati - il ricorso al primo gruppo comporta la massima fiducia nella procedura di controllo predisposta per assegnare o meno ad un'unità l'attributo di *outlier*: in altri termini, si suppone soprattutto che la sua *potenza* sia molto elevata (ossia sia molto elevata la probabilità di definire *outlier* una unità quando questa è effettivamente affetta da errore), e che l'errore di prima specie sia comunque contenuto (ossia sia bassa la probabilità di considerare erroneamente una unità come *outlier*). Questa impostazione concettuale è coerente con la necessità, assai frequente nei contesti operativi d'indagine, di doversi cautelare soprattutto dal rischio di persistenza di valori anomali nella base dati predisposta per i calcoli, prima ancora che dal rischio di correggere dati viceversa non affetti da errore[9].

La rilevanza di questo aspetto è ulteriormente sottolineata dalla necessità, attualmente in fase di rapida diffusione, di finalizzare una indagine statistica non solo al calcolo di alcuni aggregati sintetici fondamentali (indifferentemente ammontari od indici di variazione), ma anche di predisporre una base di dati individuali d'impresa utilizzabili per altre finalità. Ciò implica una particolare attenzione all'uso delle metodologie di correzione, che dovrebbero

garantire la coerenza trasversale e longitudinale delle informazioni sia a livello *macro* che a livello *micro*, e questa duplice finalità potrebbe implicare il ricorso a procedure di riponderazione dei dati differenziate. In realtà, come si vedrà la distinzione tra le due famiglie di tecniche non è sempre così netta; inoltre possono esistere diverse scelte dei fattori w_{1i} e w_{2i}, di cui un possibile riepilogo è fornito nella seguente tabella 3.1.

Tabella 3.1 - Possibili strategie per l'archiviazione dei microdati ed il calcolo di una variazione in presenza di unità *outlier*

TIPOLOGIA DI FATTORE CORRETTIVO	a	b	c	d	e	f
1) Fattore assegnato all'unità nella base dati finale	0	1	1	w_1	1	w_1
2) Fattore assegnato all'unità in fase di stima (formula 2.8)	0	0	1	1	w_2	w_2

a) In tale situazione l'unità anomala viene eliminata sia dalla base dati sia dalla procedura di calcolo della variazione. Viene così del tutto assimilata ad una mancata risposta. Questa soluzione, particolarmente drastica, sembra plausibile solo nei casi in cui l'unità in questione non sia autorappresentativa, o comunque non presenti dimensioni troppo rilevanti con riferimento alla variabile *Y*, oppure qualora non si disponga del tempo e/o delle metodologie idonee per una correzione.

b) In questo caso l'unità anomala viene inserita nella base dati ma non è considerata per i calcoli. Si tratta di una soluzione possibile qualora il dato anomalo sia in realtà veritiero - benchè particolarmente al di fuori del normale intervallo di accettazione predisposto per l'indagine - e non si disponga di metodologie affidabili per una sua opportuna riponderazione in sede di stima di una variazione.

c) In tale circostanza l'unità anomala viene conservata senza alcuna alterazione tanto nella base dati quanto in fase di calcolo di una variazione. Rispetto alla situazione di cui al punto precedente si suppone di disporre di informazioni addizionali circa la vera dinamica della variabile *Y* nello strato in esame, in base a cui non si ritiene distorcente la permanenza della unità anomala, senza alcuna alterazione del suo peso campionario, nell'insieme delle unità utilizzate per il calcolo della variazione[10].

d) L'unità anomala viene inserita nella base dati con una alterazione basata sul fattore w_1, ma

[9] In tale ottica, Weir (1997) suggerisce come la valutazione del numero di errori commessi in una procedura di *editing*, distinguendo le erronee correzioni dei dati buoni dalle mancate correzioni dei dati errati, rappresenti una utile metodologia per verificare l'efficacia di una procedura di correzione.

in sede di stima di una variazione le viene assegnato un peso uguale a quello originario (il precedente caso c) è un caso particolare di d) per w_1=1).

e) L'unità anomala viene inserita nella base dati senza modifiche, ma in sede di stima di una variazione le viene assegnato un peso diverso da quello originario, alterato secondo il fattore w_2 (i precedenti casi b) e c) sono due soluzioni particolari per w_2=0 e w_2=1).

f) L'unità anomala viene inserita nella base dati con una alterazione basata sul fattore w_1 ed in sede di stima di una variazione le viene assegnato un peso diverso dal peso originario, alterato secondo il fattore w_2. Può verificarsi $w_1 = w_2$, come ad esempio nel precedente caso c); ovviamente d) ed e) sono casi particolari di f).

Dei primi tre casi elencati, certamente più immediati, a) sembra di scarsa utilità, c) sembra alquanto rischioso, mentre b) è di uso non infrequente in pratica, trattandosi di un meccanismo alquanto cautelativo.

D'altra parte, il ricorso al metodo d) equivale in pratica ad una correzione dei dati anomali (e quindi degli indici anomali), mentre il metodo e) equivale ad una riallocazione dei pesi originari che dovrebbe assegnare, *coeteris paribus*, un peso più basso alle unità anomale. Nel paragrafo 4 si tratteranno alcune metodologie tese ad implementare la soluzione d), mentre nel paragrafo 5 verranno illustrate tecniche di tipo e). Tecniche "miste" di tipo f) sono meno frequenti, sebbene risultino ancora più cautelative delle precedenti, intervenendo sia sui microdati che sui pesi.

Va infine notato che, in effetti, ad ogni procedura di correzione dei dati anomali di tipo d) corrisponde implicitamente una riponderazione di tipo e): infatti se dopo la correzione il nuovo microdato è dato da $Y_{Sti}\, w_{1i}$, e quindi il nuovo indice da $I_{Sti/k}\, w_{1i}$, ricordando la [2.8] tale trasformazione equivale, ai fini del calcolo dello stimatore T_{St}, al ricorso agli indici originari $I_{Sti/k}$ con i nuovi pesi $D_{Sti}\, w_{1i}$, per cui in generale la corrispondenza tra le procedure d) ed e) è stabilita dalla relazione:

$$w_{2i} = D_{Sti}\, w_{1i}\,. \qquad [3.1]$$

Il reciproco di tale connessione, per quanto immediato dal punto di vista algebrico, sembra meno realistico da un punto di vista prettamente interpretativo: infatti se è vero che il ricorso

[10] Si noti come sia stata omesso, in quanto irrealistico, il caso in cui una unità anomala venga esclusa dalla base dati ma venga utilizzata senza alterazione del suo peso originario per il calcolo di una variazione.

ad una tecnica e) equivale ad una invarianza dei pesi originari con una alterazione degli indici originari $I_{Sti/k}$, il significato economico di un indice pari a $I_{Sti/k} D_{Sti}$ potrebbe non risultare ammissibile, così come l'introduzione nella base dati finali del valore "corretto" $Y_{Sti} D_{Sti}$.

4. Correzione degli indici *outlier* senza alterazione dei pesi

Si tratta generalmente di metodi preferibili nei contesti in cui si è sufficientemente sicuri della natura di *outlier* di alcuni valori e si ritiene che l'anomalia derivi da un errore nella dichiarazione e/o nella registrazione del microdato di base. A differenza dei criteri basati sulla modifica dei pesi originari assegnati ad ogni unità campionaria, correggendo gli *outlier* si può controllare meglio il numero di interventi che saranno effettuati sui dati elementari, intervenendo eventualmente sugli estremi dell'intervallo di accettazione. Una raccomandazione preliminare consiste nel "simmetrizzare" la distribuzione originaria dei rapporti tendenziali individuali per evitare che risulti problematica l'individuazione dei valori anomali e/o che pochi valori particolarmente anomali possano influire eccessivamente sull'intera procedura di correzione, specie se associati alle unità più significative nei termini dell'ordine di grandezza di *Y*.

4.1 Metodo della modifica minima rispetto agli indici originari

Supponendo di poter suddividere l'insieme degli n_{St} indici individuali disponibili al tempo *t* nei due sottoinsiemi E_{St} (indici *outlier*) e $\overline{\mathrm{E}}_{St}$ (indici *non outlier*, più brevemente indicabili con la terminologia *buoni*), una semplice metodologia per la correzioni degli indici *outlier* che non altera i pesi campionari *D* introdotti nel paragrafo 2 è la seguente. Se $I_{Sti/k}$ è il generico indice *outlier* appartenente a E_{St}, l'obiettivo di tale procedura consiste nel sostituire a tale indice il nuovo indice $I^{*}_{Sti/k}$, in modo che i nuovi indici non siano troppo diversi dagli originari ritenuti *outlier*, nell'ottica di alterare il meno possibile i microdati disponibili. Tale opzione metodologica deriva anche dalla ricorrente situazione di incertezza circa la vera natura di *outlier* di alcuni indici, già menzionata nel paragrafo precedente, che sconsiglia il ricorso ad alterazioni troppo severe della base dati disponibile. Per tale motivo questo metodo può essere definito come quello della *modifica minima*. In simboli occorre determinare i nuovi indici $I^{*}_{Sti/k}$ tali che risulti minima la funzione:

$$\Phi^2 = \sum_{i \in \mathrm{E}_{St}} (I_{Sti/k} - I^*_{Sti/k})^2 . \qquad [4.1.1]$$

Nella procedura di minimizzazione è utile introdurre il vincolo seguente:

$$\sum_{i \in \mathrm{E}_{St}} I^*_{Sti/k} D_{Sti} = f(\mathrm{E}_{St}) \qquad [4.1.2]$$

dove sono possibili diverse scelte della funzione *f*, comunque dipendente dall'insieme degli indici *outlier* al tempo *t*. Ad esempio, con riferimento ad un periodo antecedente (*t*-*h*) si potrebbe calcolare l'indice medio $I(\mathrm{E}_{St}, t-h)$ ottenibile per media aritmetica ponderata degli indici $I_{S,t-h,i/k}$ relativi alle unità risultate *outlier* al tempo *t*, secondo la relazione:

$$\sum_{i \in \mathrm{E}_{St}} I_{S,t-h,i/k} D_{Sti} = I(\mathrm{E}_{St}, t-h),$$

mentre l'indice medio $I(\overline{\mathrm{E}}_{St}, t-h)$ ottenibile per media aritmetica ponderata degli indici $I_{S,t-h,i/k}$ relativi alle unità risultate *buone* al tempo *t* sarà dato da:

$$\sum_{i \in \overline{\mathrm{E}}_{St}} I_{S,t-h,i/k} D_{Sti} = I(\overline{\mathrm{E}}_{St}, t-h).$$

Se in luogo della [4.1.2] valesse la seguente identità:

$$\sum_{i \in \mathrm{E}_{St}} I_{Sti/k} D_{Sti} = I(\mathrm{E}_{St}, t),$$

si può ragionevolmente porre:

$$f(\mathrm{E}_{St}) = I(\mathrm{E}_{St}, t-h)\left[\frac{I(\overline{\mathrm{E}}_{St}, t)}{I(\overline{\mathrm{E}}_{St}, t-h)}\right] \qquad [4.1.3]$$

e quindi *f* rappresenta una stima dell'indice relativo alle sole unità *outlier* al tempo *t*, ottenuta moltiplicando il corrispondente indice riferito al tempo (*t*-*h*) per la variazione dell'indice

relativo alle sole unità *buone* intercorsa tra i tempi (*t*-*h*) e *t*.

Ovviamente l'efficacia di questa procedura dipende strettamente dalla scelta del periodo (*t*-*h*): in indagini annuali si dovrebbe porre *h*=1, in indagini congiunturali *h*=4 (trimestrali) o *h*=12 (mensili)[11], nonchè dall'ipotesi che nell'intervallo di durata *h* i pesi campionari restino inalterati. E' inoltre implementabile solo se si dispone di osservazioni retrospettive per il medesimo campione di imprese rispondenti al tempo *t*. In caso contrario si potrebbe porre:

$$f(\mathrm{E}_{St}) = I(\mathrm{E}_{St}, X)\left[\frac{I(\overline{\mathrm{E}}_{St})}{I(\overline{\mathrm{E}}_{St}, X)}\right] \qquad [4.1.4]$$

dove $I(\mathrm{E}_{St}, X)$ indica l'indice relativo al tempo *t* calcolato con riferimento ad una variabile *X* correlata positivamente con *Y* ed il cui ammontare al tempo *t* è noto per l'intero universo: potrebbe trattarsi, ad esempio, del numero degli addetti, generalmente disponibili dall'archivio di estrazione del campione.

Se al vincolo [4.1.2] si associa il moltiplicatore di Lagrange λ e la relativa sommatoria di sinistra è indicata con Ψ, occorre dunque minimizzare la funzione:

$$\Omega = \Phi^2 + \lambda(\Psi - f).$$

Uguagliando a zero la generica derivata prima della funzione Ω rispetto al generico indice incognito *i-mo* si avrà:

$$2(I_{Sti/k} - I^*_{Sti/k}) + \lambda D_{Sti} = 0$$

da cui, moltiplicando l'equazione per D_{Sti} e sommando per $i \in \mathrm{E}_{St}$, dopo alcuni passaggi si può esplicitare il parametro λ ottenendo:

$$\lambda = \frac{2[f - I(\mathrm{E}_{St}, t)]}{\sum_{i \in \mathrm{E}_{St}} (D_{Sti})^2}.$$

Risolvendo la precedente equazione rispetto a $I^*_{Sti/k}$ si ottiene infine la soluzione ottimale:

$$I^{*}_{Sti/k} = I_{Sti/k} + \frac{D_{Sti}[f - I(\mathrm{E}_{St}, t)]}{\sum_{i \in E_{St}} (D_{Sti})^2} \qquad [4.1.5]$$

Esistono due limiti principali in sede di implementazione empirica della [4.1.5]:

1. il significato economico delle correzioni potrebbe essere ambiguo, trattandosi di una procedura meccanicistica che viene applicata simultaneamente per tutti gli indici *outlier*;
2. in particolare, non è escluso che alcuni indici corretti possano risultare negativi qualora $I(\mathrm{E}_{St}, t) > f$.

Infine, si noti come sostituire ad ogni indice *outlier* un indice corretto secondo la [4.1.5] equivale a non alterare gli indici *outlier,* variandone però i pesi campionari secondo la relazione:

$$W_{Sti} = D_{Sti} + \frac{(D_{Sti})^2 [I(\mathrm{E}_{St}, t) - f]}{I_{Sti/k} \sum_{i \in E_{St}} (D_{Sti})^2} \; .$$

4.2 Metodi basati sul troncamento

In questo paragrafo si introdurrà una semplice procedura di "troncamento" di un dato *outlier*, cercando di evidenziarne l'estrema generalità formale per poi derivarne alcuni casi particolari. assai diffusi. Dato il forte parallelismo con la più ampia tematica della stima di mancate risposte, per una rassegna al riguardo si rimanda a Lucev (1997).

Nella sua versione originaria tale tecnica di correzione di un *outlier* consisteva semplicemente nel sostituire il valore originario con uno degli estremi dell'intervallo di accettazione, e la sua efficienza dipende strettamente dai criteri di definizione di tale intervallo (Searls, 1966). Il criterio proposto in questo paragrafo è di tipo "misto", contemplando una soluzione generalmente più efficiente in presenza di informazioni ausiliarie sulle unità affette da valori anomali.

Come già nel primo paragrafo, si ribadisce che non verrà affrontato in dettaglio il problema della scelta dell'intervallo di accettazione, la cui determinazione appare

[11] In effetti è generalmente preferibile fare riferimento allo stesso periodo dell'anno precedente (confronto tendenziale) piuttosto che al periodo precedente (confronto congiunturale).

sensibilmente legata alle specificità del contesto in cui si opera. Non è comunque superfluo ricordare che in genere si è in presenza di distribuzioni caratterizzate da asimmetria positiva, con riferimento alle quali *è difficile individuare gli outlier*: quelli *per eccesso* perchè spesso la coda di destra è molto lunga e piatta[12], quelli per difetto perchè il ricorso ai consueti intervalli di confidenza (μ-rσ) condurrebbe rapidamente (e spesso quasi esclusivamente) a zone di rifiuto negative, non ammissibili se si suppone la non negatività di *Y*. In questo caso è consigliabile il ricorso ad una preventiva simmetrizzazione della distribuzione - come già ricordato - ad esempio con delle trasformazioni logaritmiche come proposto in Thompson e Sigman (op.cit.).

Ricordando le simbologie proposte nella tabella 3.1, si tratta di trasformare l'indice originario identificato come *outlier* secondo la formula:

$$I^{*}_{Sti/k} = I_{Sti/k}\, w_{1Sti}\,. \qquad [4.2.1]$$

Si supponga di aver individuato un sottoinsieme di unità *outlier*, caratterizzate cioè da indici individuali esterni all'intervallo [*L*,*U*] i cui estremi, per semplicità, saranno riportati senza i pedici relativi al campione *S* ed al periodo *t* di riferimento. Il vettore di osservazioni buone è quindi così definito:

$$\mathbf{Y}_{LU} = \left\{ Y_{Sti} : L \le Y_{Sti} \le U \right\},$$

la cui media sarà indicata senza ambiguità con $\overline{\mathrm{Y}}_{St/LU}$, ed in generale con riferimento ad una variabile ausiliaria *X* si supporrà disponibile anche il vettore:

$$\mathbf{X}_{LU} = \left\{ X_{Sti} : L \le Y_{Sti} \le U \right\},$$

la cui media sarà indicata senza ambiguità con $\overline{\mathrm{X}}_{St/LU}$. Se ε_{Sti} rappresenta una variabile casuale normale a media nulla definita nell'intervallo [-α,+α] piccolo a piacere, si possono poi introdurre le tre funzioni seguenti:

[12] In proposito si rimanda a Draper e Winkler (1997): tale inconveniente implica l'impossibilità di determinare le soglie di accettazione (quindi nessuna unità risulta essere un *outlier*), o il ricorso ad una ulteriore stratificazione dell'insieme originario di unità analizzate.

$$A_{Sti} = I_{Sti/k}\left(\frac{X_{Sti}}{Y_{Sti}}\right) g(\mathbf{Y}_{LU}, \mathbf{X}_{LU})(1+\varepsilon_{Sti}) \qquad B_{Sti} = U\left(1-\frac{|\varepsilon_{Sti}|}{2}\right) \qquad C_{Sti} = L\left(1+\frac{|\varepsilon_{Sti}|}{2}\right).$$

In corrispondenza di una unità *outlier* si possono quindi verificare due possibilità:

Caso I: $Y_{Sti} > U$.

Si può allora porre

$$I^*_{Sti/k} = \begin{cases} m_{AB} = \min[A_{Sti}; B_{Sti}] & se \quad L \le m_{AB} \le U \\ B_{Sti} & altrimenti \end{cases} \qquad [4.2.2]$$

Caso II: $Y_{Sti} < L$.

Si può allora porre:

$$I^*_{Sti/k} = \begin{cases} m_{AC} = \max[A_{Sti}; C_{Sti}] & se \quad L \le m_{AC} \le U \\ C_{Sti} & altrimenti. \end{cases} \qquad [4.2.3]$$

Con tale procedura si sostituiranno ai valori anomali dei nuovi valori certamente interni all'intervallo di accettazione ed il più possibile lontani dagli estremi dello stesso; l'introduzione della variabile ε serve a ridurre il rischio di un eccessivo appiattimento della distribuzione, successivo alle correzioni, verso gli estremi *L* e *U*.

Ricordando la [4.2.1] è infine immediato verificare che le opzioni A_{Sti}, B_{Sti} e C_{Sti} sopra introdotte equivalgono, rispettivamente, alle trasformazioni:

$$w_{1Sti} = \left(\frac{X_{Sti}}{Y_{Sti}}\right) g(\mathbf{Y}_{LU}, \mathbf{X}_{LU})(1+\varepsilon_{Sti}) \qquad w_{1Sti} = \frac{U}{I_{Sti/k}}\left(1-\frac{|\varepsilon_{Sti}|}{2}\right) \qquad w_{1Sti} = \frac{U}{I_{Sti/k}}\left(1+\frac{|\varepsilon_{Sti}|}{2}\right).$$

E' dalle varie forme che può assumere l'espressione A_{Sti}, dipendenti dalle scelte di *X* e *g*, che derivano alcune delle soluzioni operative più frequenti in pratica[13]:

[13] Per semplicità si trascurerà l'effetto della componente causale ε.

1. se X =1 per ogni unità i e $g = \overline{Y}_{St/LU}$, si sostituisce l'osservazione Y_{Sti} *outlier* con la media campionaria calcolata sulle sole unità buone;
2. se X varia al variare di i e $g = \overline{Y}_{St/LU} / \overline{X}_{St/LU}$, si ha un raffinamento del metodo precedente in cui si sfrutta la correlazione positiva tra *X* e *Y*, con *X* nota su tutte le unità campionarie. E' frequente il caso in cui *X* rappresenta il numero degli addetti, oppure il valore di *Y* associato all'unità *i* ritardato di *h* periodi e supposto noto: se tale valore fosse disponibile per tutte le unità del campione dovrebbe essere quasi certamente preferito ad una generica variabile ausiliaria *X* (cfr. Gismondi, 1997);
3. se X =1 per ogni unità i e $g = Y_{Std(i)}$, dove il pedice indica l'unità *d* donatrice rispetto all'*outlier i*, si sostituisce l'osservazione Y_{Sti} *outlier* con il valore di una unità donatrice, scelta con un criterio di distanza minima;
4. se X varia al variare di i e $g = Y_{Std(i)} / X_{Std(i)}$, dove la quantità a denominatore indica il valore di *X* associato all'unità donatrice *d*, si sostituisce l'osservazione *outlier* con il valore di una unità donatrice, corretta sulla base del rapporto tra i valori della variabile *X* della unità ricevente e donatrice (tale metodo è proposto ed ulteriormente dettagliato in Istat, 1998);
5. se X varia al variare di i e $g = \hat{\beta}_{St}$, si sostituisce l'osservazione Y_{Sti} *outlier* con una stima basata sul metodo della regressione, dove il coefficiente di regressione è stato stimato sulla base delle sole unità buone.

Si ribadisce come, in pratica, la disponibilità di basi di dati longitudinali anche non particolarmente lunghe comporti spesso la posizione $X_{Sti} = Y_{S,t-h,i}$ per cui ad esempio lo stimatore relativo al precedente caso 1 comporta che:

$$A_{Sti} = I_{Sti/k}\left(\frac{Y_{S,t-h,i}}{Y_{Sti}}\right)\left(\frac{\overline{Y}_{St}}{\overline{Y}_{S,t-h}}\right)(1+\varepsilon_{Sti}) = \left(\frac{\overline{Y}_{St}}{Y_{S,t-k,i}}\right)\left(\frac{Y_{S,t-h,i}}{\overline{Y}_{S,t-h}}\right)(1+\varepsilon_{Sti}) \qquad [4.2.4]$$

dove la quantità a destra è data dal prodotto tra:

- il rapporto tra il valore medio campionario di *Y* al tempo *t* ed il valore di *Y* relativo all'unità *i-ma* al tempo base (prima parentesi);
- il rapporto tra il valore di *Y* relativo all'unità *i-ma* al tempo (*t-h*) ed il valore medio

campionario di *Y* al tempo (*t*-*h*) (seconda parentesi);

- una componente casuale.

E' questa la funzione utilizzata nell'applicazione del paragrafo 6.

5. Modifica dei pesi campionari senza alterazione degli indici *outlier*

A parità di condizioni, si tratta di metodi preferibili nei contesti in cui si desideri comunque cautelarsi dalla eventuale presenza di *outlier* - indipendentemente dal fatto che se ne siano effettivamente accertati alcuni - oppure nel caso di *outlier* rappresentativi di cui non si vuole alterare il microdato di base.

5.1 Metodo della modifica minima rispetto ai pesi originari

In questo contesto, a differenza di quanto vista nel paragrafo 4, si supporrà di non trattare separatamente i due sottoinsiemi degli indici *outlier* e degli indici buoni, bensì di accertare la presenza di unità *outlier* e, successivamente, di procedere alla rideterminazione dei pesi campionari di tutte le unità secondo la semplice metodologia di cui in seguito. L'idea di fondo è che in alcuni casi la correzione degli indici *outlier* potrebbe non essere possibile (ad esempio, per mancanza di informazioni ausiliarie sufficienti per garantire una buona qualità del processo di correzione) o consigliabile (ad esempio, perchè i valori assunti da tali indici, per quanto anomali rispetto al recente *trend*, potrebbero non essere necessariamente errati, e quindi una correzione dei relativi microdati risulterebbe inopportuna).

Se nella [2.8] si sostituisce al peso campionario *D* un nuovo peso *W*, da determinare secondo un criterio di ottimalità, il generico stimatore di un indici di variazione sarà allora scrivibile nella forma:

$$T_{St} = \sum_{i=1}^{n_{St}} (I_{Sti/k}) W_{Sti} \,. \qquad [5.1.1]$$

Un criterio ragionevole consiste nell'imporre che una buona rideterminazione dei pesi dovrebbe alterare il meno possibile i pesi campionari relativi a tutte le unità, per cui risulta fondamentale la scelta del vincolo da introdurre nel processo di minimizzazione della somma dei quadrati degli scarti tra pesi originari e nuovi pesi. Occorre quindi minimizzare la

funzione:

$$\Phi^2 = \sum_{i=1}^{n_{St}} (D_{Sti} - W_{Sti})^2 , \qquad [5.1.2]$$

con il vincolo seguente:

$$\sum_{i=1}^{n_{St}} (I_{X,Sti}) W_{Sti} = I_{X,t} , \qquad [5.1.3]$$

dove $I_{X,t}$ è l'indice di variazione rispetto al periodo (*t*-*h*), supposto noto per l'intero universo di riferimento, di una variabile *X* correlata positivamente con *Y*, di cui sono note le modalità assunte sulle unità del campione *S*[14]. In tale ottica si suppone di disporre di un *totale noto* relativo all'intero strato da cui è stato estratto il campione *S* e di vincolare la determinazione dei nuovi pesi al soddisfacimento del vincolo di uguaglianza rispetto a tale totale, indipendentemente dal fatto che nel campione si siano verificate unità *outlier*. Omettendo lo sviluppo algebrico, formalmente analogo a quello visto nel paragrafo 4.1, si ottiene la relazione:

$$W_{Sti} = D_{Sti} + \left[\frac{I_{X,t} - \sum_{i=1}^{n_{St}} (I_{X,Sti}) D_{Sti}}{\sum_{i=1}^{n_{St}} (I_{X,Sti})^2} \right] . \qquad [5.1.4]$$

Con riferimento alla formula precedente valgono le stesse limitazioni riportate al termine del paragrafo 4.1. Inoltre è immediato verificare che la modifica dei pesi secondo la [5.1.4] equivale ad utilizzare per la stima dell'indice I_{St} ancora i pesi campionari originari, modificando però gli indici originari, moltiplicandoli per il rapporto W_{Sti} / D_{Sti}.

5.2 Una soluzione model-based

In questo contesto verrà adottato un approccio svincolato dal disegno campionario, privilegiando una impostazione metodologica *model-based*. In altri termini si supporrà che la variabilità delle misurazioni non derivi dalla aleatorietà del campione, bensì dalla variabilità

[14] Anche in questo caso potrebbe trattarsi del numero di addetti, o dei valori di *Y* ritardati di *h* periodi rispetto a quello di riferimento.

intrinseca di tutte le possibili realizzazioni comunque fissato un campione di riferimento[15]. Questa impostazione può rivelarsi utile soprattutto nei casi in cui risulti complessa la valutazione della precisione delle stime secondo un approccio basato sul disegno campionario, e si ripone la massima fiducia nella verosimiglianza dell'unico campione effettivamente disponibile. Tra le motivazioni più ricorrenti si citano le seguenti:

- l'implementazione del disegno campionario previsto in origine potrebbe risultare affetta da numerose distorsioni operative; in particolare, in un'ottica longitudinale il campione di rispondenti è spesso assimilabile più ad un *panel* autoselezionatosi naturalmente piuttosto che al risultato di un preciso disegno campionario, e questo implica che nessuna tecnica tradizionale di campionamento sottostà all'insieme dei dati disponibili;
- all'interno di strati campionari in cui si è supposta uguale media ed omoschedasticità per tutte le unità che vi fanno parte tali ipotesi potrebbero risultare irrealistiche, come è implicitamente confermato dal verificarsi di unità *outlier*;
- in diversi casi un approccio basato su un modello piuttosto che sul disegno campionario può risultare metodologicamente più semplice, nonchè più adeguato, potendo basarsi su una modellizzazione più realistica dei dati in esame, soprattutto con riferimento alla evidente eteroschedasticità di molte unità appartenenti al medesimo strato.

Si supponga che all'interno dello strato esaminato valgano le seguenti ipotesi:

a) una varianza (rispetto al modello) di ogni rapporto tendenziale individuale generalmente diversa da unità ad unità, ossia $VAR(I_{Sti/k}) = \sigma^2_{ti/k}$;
b) incorrelazione (rispetto al modello) tra le variazioni individuali relative ad unità diverse.

Riprendendo lo stimatore [5.1.1] - dove i pesi *W* dovrebbero in generale essere soggetti ad un vincolo di normalizzazione, ossia la loro somma estesa alle unità campionarie dovrebbe risultare pari a uno[16]- si avranno quindi le due relazioni seguenti, in cui il simbolo *E* indica il valore atteso rispetto al modello:

$$E(T_{St}) = \sum_{i=1}^{n_{St}} E(I_{Sti/k}) W_{Sti} \qquad VAR(T_{St}) = \sum_{i=1}^{n_{St}} VAR(I_{Sti/k}) W^2_{Sti} = \sum_{i=1}^{n_{St}} \sigma^2_{ti/k} W^2_{Sti}\,, \qquad [5.2.1]$$

[15] Un approccio simile è riportato in Gismondi (1997).
[16] In effetti se si adottano pesi svincolati dal disegno campionario e liberi da ulteriori vincoli lo stimatore in questione potrebbe assumere valori ampiamente fuori dominio. In questo caso si potrebbe verificare come l'introduzione esplicita di un vincolo di somma unitaria per i

e la prima delle due relazioni precedenti evidenzia come non sia garantita la correttezza (rispetto al modello) dello stimatore T_{St}. D'altra parte, se si suppone noto l'indice $I_{S,t-h/k}$ relativo ad un periodo (*t-h*) antecedente a *t* - ad esempio perchè noto sulla base di informazioni non campionarie o perchè precedentemente stimato tramite uno stimatore basato sul disegno campionario del tipo [2.8] - una condizione che equivale alla richiesta di correttezza empirica dello stimatore *T* riferita quantomeno al tempo (*t-h*) è la seguente:

$$\sum_{i=1}^{n_{St}} (I_{S,t-h,i/k}) W_{Sti} = I_{S,t-h/k} \qquad [5.2.2]$$

nell'ipotesi che $n_{S,t-h} = n_{St}$, sicuramente realistica supponendo una rilevazione di tipo *panel*, o comunque (*t-h*) prossimo a *t*. Ovviamente il ricorso alla condizione [5.2.2] implica l'assenza di osservazioni anomale al tempo (*t-h*). Si può così definire la seguente funzione di Lagrange da minimizzare:

$$\Phi^2 = \sum_{i=1}^{n_{St}} \sigma_{ti/k}^2 (W_{Sti})^2 + \lambda \left(\sum_{i=1}^{n_{St}} I_{S,t-h,i/k} W_{Sti} - I_{S,t-h/k} \right). \qquad [5.2.3]$$

Uguagliando a zero la derivata prima della [5.2.2] rispetto a W_{Sti} si ricava:

$$\frac{\partial \Phi^2}{\partial W_{Sti}} = 2\sigma_{ti/k}^2 W_{Sti} + \lambda I_{S,t-h,i/k} = 0 \qquad \text{ossia:} \qquad W_{Sti} = -\frac{\lambda I_{S,t-h,i/k}}{2\sigma_{ti/k}^2};$$

moltiplicando ambo i membri dell'ultima uguaglianza per $I_{S,t-h,i/k}$ e sommando rispetto a *i* si ottiene:

$$\lambda = -\frac{2 I_{S,t-h/k}}{\sum_{i=1}^{n_{St}} \frac{(I_{S,t-h,i/k})^2}{\sigma_{ti/k}^2}}$$

ed è quindi immediato ricavare la soluzione:

---

pesi nella funzione di Lagrange renderebbe assai più complessa la soluzione finale e non garantirebbe la non negatività dei pesi. In tal senso la normalizzazione può essere effettuata a posteriori.

$$W_{Sti} = \frac{\left[\frac{(I_{S,t-h,i/k})(I_{S,t-h/k})}{\sigma^2_{ti/k}}\right]}{\sum_{i=1}^{n_{St}} \frac{(I_{S,t-h,i/k})^2}{\sigma^2_{ti/k}}}. \qquad [5.2.4]$$

Risulta evidente come tale metodo, più che alla effettiva identificazione delle unità *outlier* e quindi alla valutazione del meccanismo generatore dei valori anomali, è teso piuttosto all'impostazione di una procedura di calcolo dell'indice che risulti in qualche modo cautelativa rispetto alla possibile persistenza di valori anomali nella base dati utilizzata. Si noti come la somma dei pesi al variare di *i* per tutte le unità del campione non sia pari ad uno, sebbene tale vincolo, peraltro non richiesto in origine, possa risultare approssimativamente verificato in pratica; inoltre i pesi sono tutti uguali ad una costante se il rapporto $I_{S,t-h,i/k} / \sigma^2_{ti/k}$ è costante al variare di *i*.

In generale, verranno assegnati pesi più bassi alle unità caratterizzate da una elevata varianza e più alti alle unità caratterizzate da livelli elevati dell'indice relativo al periodo (*t*-*h*). Quindi in sede di calcolo dell'indice sintetico $I_{St/k}$ le unità *outlier* avranno un peso tanto più basso quanto più elevato risulterà il rapporto tra la loro variabilità ed il relativo indice di variazione del periodo (*t*-*h*), in confronto con l'andamento medio di tale rapporto nel campione osservato. Generalmente la somma dei pesi espressi dalla [5.2.3] potrebbe risultare di per sè molto vicina ad uno, per cui la loro normalizzazione potrebbe risultare poco influente.

E' evidente, da quanto sopra esposto, che la stima della varianza individuale riveste un ruolo fondamentale ai fini della affidabilità della procedura di riponderazione tesa a ridurre l'influenza di eventuali indici *outlier*. Se si limita l'orizzonte temporale al solo periodo *t* di riferimento si potrebbe stimare la varianza individuale *i-ma* con la formula:

$$\left(I_{Sti/k} - \sum_{i=1}^{n_{St}} \frac{I_{Sti/k}}{n_{St}} \right)^2 .$$

Altrimenti una stima più precisa dovrebbe essere ottenuta, ipotizzando la disponibilità degli indici individuali fino ad un periodo (*t*-*h*) precedente a *t*, tramite la relazione:

$$\sum_{r=t-h}^{t} \frac{(I_{Sri/k} - \bar{I}_{Shi/k})^2}{h+1}, \qquad \text{dove} \qquad \bar{I}_{Shi/k} = \sum_{r=t-h}^{t} \frac{I_{Sri/k}}{h+1}.$$

In tale ottica potrebbe risultare necessaria una preventiva normalizzazione degli indici relativi ai periodi precedenti a *t*, per renderli effettivamente comparabili (ad esempio, una destagionalizzazione nel caso di indicatori mensili, un aggiustamento dei livelli medi nel caso di indici annui caratterizzati da una forte componente tendenziale).

5.3 Determinazione "empirica" dei pesi W

La semplice idea di base per la determinazione di pesi che tengano conto della distribuzione delle frequenze osservate di *Y* è che, a parità di condizioni, ad ogni unità dovrebbe essere assegnato un peso crescente al crescere del suo peso economico (in termini di *Y*) e decrescente tanto più la sua variazione osservata (in termini dell'indice *I*) risulta troppo piccola o troppo grande rispetto alla variazione media osservata nello strato[17]. Se tutte le unità avessero approssimativamente lo stesso valore di *Y* e di *I*, dovrebbero avere approssimativamente lo stesso peso.

Con riferimento alle distribuzioni empiriche verificatesi in corrispondenza dell'estrazione campionaria relativa al tempo *t* ed allo strato *S* si possono definire per l'unità *i-ma* queste due funzioni:

$$\eta_{Sti} = \left[\frac{1 + F(Y_{Sti})}{1 + F(\bar{Y}_{St})} \right] \qquad [5.3.1]$$

$$\gamma_{Sti} = \frac{1 + \max[G(I_{Sti/k}); G(\bar{I}_{St/k})]}{1 + \min[G(I_{Sti/k}); G(\bar{I}_{St/k})]} \qquad [5.3.2]$$

basate, rispettivamente, sulle funzioni di ripartizione campionarie *F* e *G* dei valori Y_{Sti} e degli indici $I_{Sti/k}$, dove i simboli $\bar{Y}_{St}$ e $\bar{I}_{St/k}$ indicano le corrispondenti medie calcolate *su tutte* le unità campionarie[18]. Si verifica immediatamente che:

[17] Implicitamente si suppone, quindi, di riporre una elevata fiducia nel grado di precisione della stratificazione adottata, per cui la presenza di valori molto diversi in media viene attribuita al caso piuttosto che ad una possibile erroneità nella definizione dello strato stesso.

[18] Nel paragrafo 4.2 erano state introdotte medie formalmente analoghe, ma basate sulle sole unità campionarie appartenenti all'intervallo di accettazione.

$$\eta_{Sti} = 1 \qquad \text{se} \qquad Y_{Sti} = \overline{Y}_{St}$$
$$\eta_{Sti} \to 2 \qquad \text{se} \qquad F(Y_{Sti}) \to 1 \qquad \text{e} \qquad F(\overline{Y}_{St}) \to 0$$
$$\eta_{Sti} \to 0{,}5 \qquad \text{se} \qquad F(Y_{Sti}) \to 0 \qquad \text{e} \qquad F(\overline{Y}_{St}) \to 1$$

mentre per quanto riguarda I_{Sti} si avrà:

$$\gamma_{Sti} = 1 \qquad \text{se} \qquad I_{Sti/k} = \bar{I}_{St/k}$$
$$\gamma_{Sti} \to 2 \qquad \text{se} \qquad G(I_{Sti/k}) \to 1 \qquad \text{e} \qquad G(\bar{I}_{St/k}) \to 0$$
$$\gamma_{Sti} \to 2 \qquad \text{se} \qquad G(I_{Sti/k}) \to 0 \qquad \text{e} \qquad G(\bar{I}_{St/k}) \to 1.$$

Una possibile determinazione del peso da assegnare ad ogni unità campionaria sarà data allora dalla relazione:

$$W_{Sti} = \frac{\eta_{Sti}}{\gamma_{Sti}}. \qquad [5.3.3]$$

Tale peso varia tra 0,25 e 2, a parità del valore dell'indice di variazione cresce al crescere del peso dell'unità *i-ma* in termini dell'ordine di grandezza di *Y*, a parità di valori di *Y* decresce ad uno sia per valori dell'indice di variazione molto grandi che per valori molto piccoli (rispetto alla variazione media). Ovviamente si può pervenire facilmente a pesi variabili tra zero e uno sottraendo alla [5.3.3] il suo valore minimo e dividendo per la differenza tra il valore massimo ed il minimo. Dato che la somma dei pesi espressi dalla [5.3.3] potrebbe risultare spesso assai diversa da uno, si raccomanda la loro normalizzazione.

La [5.3.3] presenta il vantaggio di tenere conto sia dei valori medi di *Y* e di *I* nel campione, sia della forma delle relative funzioni di ripartizione. Una scelta più semplice dei pesi che risponde alla stessa logica della precedente ma non tiene conto della forma di tali funzioni di ripartizione è ottenibile utilizzando ancora la [5.3.3] ma ponendo:

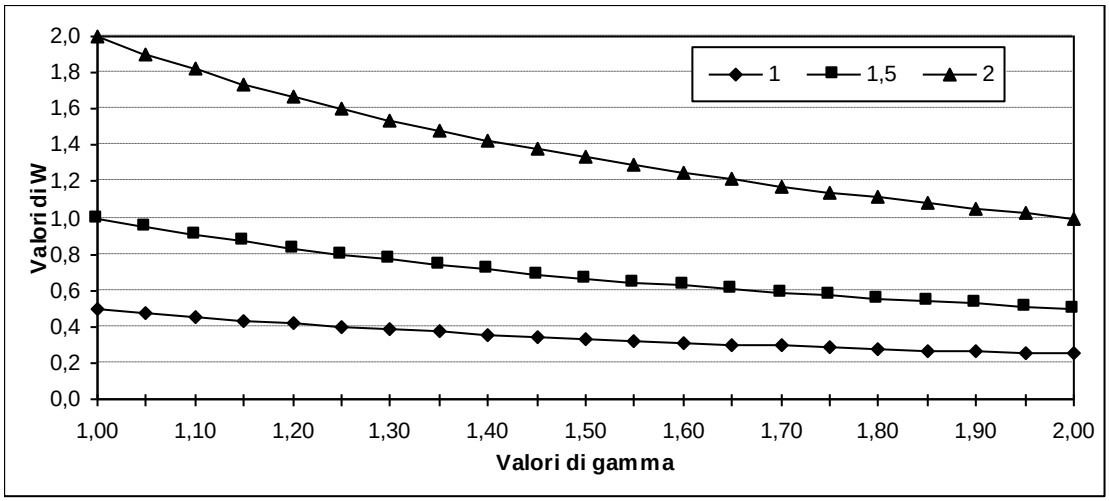
$$\eta'_{Sti} = \left(\frac{1 + Y_{Sti}}{1 + \overline{Y}_{St}} \right) \qquad [5.3.4]$$

$$\gamma'_{Sti} = \frac{1 + \max(I_{Sti/k}; \bar{I}_{St/k})}{1 + \min(I_{Sti/k}; \bar{I}_{St/k})} \qquad [5.3.5]$$

sebbene in questo caso la variabilità dei pesi risulti assai più elevata e generalmente incontrollabile a priori, con il rischio conseguente di appiattire od esaltare eccessivamente la ponderazione assegnata ad alcune unità.

Nel grafico 5.3.1 seguente è riportato l'andamento del peso *W* definito dalla [5.3.3] in funzione di γ per tre diversi livelli di η: 1, 1,5 e 2. Al crescere di γ, il peso *W* decresce secondo una pendenza sostanzialmente costante per γ superiore a 1,5. Il tasso di decrescita è più elevato per livelli elevati di η.

Grafico 5.3.1 - Valori del peso *W* in funzione di γ per tre livelli di η (1, 1,5 e 2)



6. Una applicazione

E' stato ripreso l'esempio relativo al grafico 1.1, supponendo di considerare il campione effettivamente disponibile - distintamente per le imprese classificate nelle già citate attività economiche codificate con le sigle ATECO 52.11 e 52.2 - come l'intero universo e di riestrarre da tale universo virtuale un altro campione tale da garantire una frazione sondata pari a circa il 50% (tabelle 6.1). In tal modo si è supposto di disporre di un campione di 186 unità (su 371) per le imprese specializzate a prevalenza alimentare e di 234 unità (su 468) per quelle specializzate a prevalenza alimentare.

L'estrazione è stata condotta con un disegno PPS quale quello descritto dalla [2.4], dove la variabile ausiliaria è stata rappresentata dalla media mensile del valore delle vendite

nel 1995, anno scelto come base, ed è stata reiterata per 100 volte.

La variabile *Y* è stata data dal valore delle vendite nel mese di maggio 1998, scelto tra i primi otto mesi di tale anno - ossia i mesi disponibili al momento della sperimentazione - perchè caratterizzato dalla più elevata variabilità degli indici individuali, misurata tramite il coefficiente di variazione. Ogni indice di variazione individuale è dato dal rapporto tra il valore delle vendite di maggio ed il corrispondente valore medio mensile relativo all'anno base 1995.

In corrispondenza di ognuna delle estrazioni si è simulata una incidenza relativa di valori *outlier* (sia per eccesso che per difetto) pari a circa il 30% delle unità campionarie, quota assai simile a quella massima riscontrabile in pratica nell'indagine sulle vendite al dettaglio; gli indici individuali *outlier* sono stati generati tramite l'introduzione di perturbazioni casuali degli indici individuali aventi media pari a 0,5, e tale procedura è stata reiterata per 100 volte, per cui l'intera sperimentazione si è avvalsa in ognuno dei due strati analizzati di 10.000 replicazioni (100 campioni di uguale dimensione per 100 generazioni casuali di 3 indici individuali *outlier* ogni 10). In tal senso le stime riportate nella tabella 1 vanno intese come medie calcolate su 10.000 osservazioni.

Gli indici di variazione "veri" sono risultati pari a 1,0031 per la classe ATECO 52.11 ed a 0,9994 per il gruppo ATECO 52.2., per cui la forte eterogeneità degli indici di variazione individuali riscontrata a maggio non ha prodotto indici di variazione medi di strato particolarmente significativi, risultando entrambi molto vicini all'unità.

Per ogni campione estratto nella simulazione la stima della variazione media per l'intero "universo" è stata effettuata ricorrendo alla formula [2.3], ottenibile dalla più generale [2.8] ponendo $D_{Sti} = 1/n_{St}$, nei seguenti casi: metodo [4.1] della modifica minima rispetto agli indici originari e metodo [4.2] basato sul troncamento, ossia i metodi finalizzati alla correzione degli indici *outlier*. D'altra parte, si è ricorsi allo stimatore [5.1.1] nel caso dei metodi finalizzati al ricorso a pesi alternativi a quelli originari derivati dal disegno campionario, ossia i metodi [5.1] della modifica minima, [5.2] di tipo *model-based* e [5.3] basato su una procedura empirica.

Nel metodo [4.2] basato sul troncamento le soglie di accettazione inferiore e superiore sono state poste pari, rispettivamente, a 0,5 e 2, sulla base dell'esame preliminare della distribuzione di frequenze empirica dei rapporti individuali, mentre lo stimatore A_{Sti} è stato definito dalla [4.2.4], in cui (*t*-*h*) si riferisce al mese di aprile 1998, così come l'indice di

variazione $I_{X,t}$ utilizzato nel vincolo [5.1.3] e quindi per implementare il metodo [5.1]. Per la stima delle varianze individuali, necessaria ai fini dell'implementazione del metodo [5.2], si è utilizzata la seconda delle due formule proposte alla fine del paragrafo 5.2, dove *h*=4 e quindi sono stati considerati, in un'ottica longitudinale, i mesi da gennaio a maggio 1998. Infine per l'implementazione del metodo [5.3] è stata utilizzata la formula [5.3.3].

Va ricordato che il contenuto della tabella 1 si riferisce ai risultati medi ottenuti replicando per 100 volte l'estrazione campionaria fittizia (a parità di dimensione del campione) e per altrettante volte la generazione degli *outlier*.

Come evidenziato dalla tabella, l'effetto della presenza di *outlier* è più rilevante per le imprese non specializzate (l'indice calcolato in presenza di *outlier* è pari a 1,0818 rispetto a 1,0031 vero) che per quelle specializzate (1,0293 contro 0,9994), probabilmente perchè i valori anomali sono risultati spesso associati ad unità con peso economico rilevante, fenomeno più significativo tra le imprese non specializzate, caratterizzate da livelli medi delle vendite più elevati di quelle specializzate della stessa dimensione. I risultati salienti emersi dalla sperimentazione sono i seguenti:

- il metodo migliore, almeno limitatamente a questa applicazione, è risultato il [5.1], basato sulla modifica minima rispetto ai pesi campionari originari: l'errore percentuale di stima (in valore assoluto) scende dal 7,84% al 3,29% per le imprese non specializzate e dal 3,00% al 2,46% per quelle specializzate;
- tra i metodi rimanenti, l'unico in grado di ridurre l'errore di stima è il [4.2] basato sul troncamento, con un guadagno di precisione relativamente alle imprese non specializzate quasi simile a quello del metodo precedente (l'errore è del 3,32%) e lievemente peggiore per le imprese specializzate (2,90%);
- i rimanenti tre metodi realizzano guadagni di precisione solo con riferimento ad uno dei due strati analizzati: in particolare il metodo [5.3] basato sulla determinazione empirica dei pesi *W* è il meno efficiente, dato che peggiora la precisione della stima con riferimento alle imprese non specializzate e la migliora pochissimo per quelle specializzate (l'errore è del 2,95%). Ciò dipende, probabilmente, dalla difficoltà di definire con una sufficiente precisione la forma corretta delle funzioni di ripartizione potendosi avvalere di un campione di unità di così ridotte dimensioni;
- mentre il metodo [4.1] migliora sensibilmente la precisione della stima nel caso delle imprese non specializzate e la peggiora lievemente per quelle specializzate (errore del 3,78%), il metodo *model-based* [5.2] risulta in media assai preciso con riferimento alle

prime (l'errore è pari ad appena lo 0,06%) e molto più impreciso con riferimento alle seconde (errore medio del 5.53%), per la maggiore difficoltà incontrata in quest'ultimo caso per stimare correttamente le varianze individuali.

Pur ribadendo che una sola applicazione non consente valutazioni troppo generali sulla qualità delle varie metodologie proposte, il suggerimento emergente con chiarezza è che, a parità di condizioni e di informazioni addizionali disponibili, sembrano preferibili le tecniche che conservano il più possibile i riferimenti di ponderazione individuale derivati dal disegno campionario originario, quali risultano essere i metodi [4.2] e [5.1].

Tabella 6.1 - Risultati di una sperimentazione comparativa tra i 5 metodi introdotti

Variabile	Imprese commerciali al dettaglio non specializzate a prevalenza alimentare (ATECO 52.11)	Imprese commerciali al dettaglio specializzate a prevalenza alimentare (ATECO 52.2)
Principali parametri della sperimentazione		
Numerosità dello strato	371	468
Numerosità del campione	186	234
Quota sondata	50,1	50,0
Numero valori outlier	56	70
Quota outlier nel campione	30,1	29,9
Variazioni vere e stimate (maggio 1998 su media 1995) (*)		
Variazione vero nello strato	**1,0031**	**0,9994**
Variazione stimata con outlier	1,0818	1,0293
Stima con metodo [4.1]	1,0424	1,0372
Stima con metodo [4.2]	1,0365	1,0284
Stima con metodo [5.1]	1,0361	1,0240
Stima con metodo [5.2]	1,0025	1,0547
Stima con metodo [5.3]	1,0932	1,0289
Scarti percentuali (in valore assoluto) tra variazioni stimate e vere (*)		
Variazione stimata con outlier	0,0784	0,0300
Stima con metodo [4.1]	0,0391	0,0378
Stima con metodo [4.2]	0,0332	0,0290
Stima con metodo [5.1]	0,0329	0,0246
Stima con metodo [5.2]	0,0006	0,0553
Stima con metodo [5.3]	0,0898	0,0295

(*) Elaborazioni su dati ISTAT. Si tratta dei risultati medi su 10.000 replicazioni.

Riferimenti bibliografici

BARCAROLI G.-LUZI O. (1995), "Sistema generalizzato per l'editing e l'imputazione di variabili quantitative (GEIS)", *Quaderni di ricerca*, 1, 1-83, Istat, Roma.

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

COCHRAN W.G. (1977), *Sampling Techniques - 3th edition*, John Wiley & Sons, New York.

DIGGLE P.J.-YEE LIANG K.-ZEGER S.L. (1994), *Analysis of Longitudinal Data*, Oxford Statistical Science Series, 13, Oxford Science Publications.

DRAPER L.R.-WINKLER W.E. (1997), "Balancing and Ratio Editing with the New SPEER System", paper presented at the *Work Session on Statistical Data Editing*, 14-17 Ottobre, Praga.

EDWARDS W.S.-CANTOR D. (1991), *Towards a Response Model in Establishment Surveys*, in P.P.Biemer-R.M.Groves-L.E.Lyberg-N.A.Mathiowetz-S.Sudman (editors) "Measurement Errors in Surveys", 211-236, John Wiley & Sons, New York.

FELLEGI I.P.-HOLT D. (1976), "A Sistematic Approach to Authomatic Editing and Imputation, *Journal of the American Statistical Association*, 71, 17-35.

FULLER W.A. (1987), *Measurement Error Models*, John Wiley & Sons, New York.

GARCIA E.-PEIRATS V. (1994), "Evaluation of Data Editing Procedures: Results of a Simulation Approach", *Statistical Data Editing Methods and Techniques Vol.I*, Conference of European Statisticians and Studies, 44, 52-68.

GISMONDI R. (1997), "Domain Change in Panel Surveys: a Model Based Strategy to Estimate Net Changes in Uncertainty Conditions", paper presented at the *IASS/IAOS Satellite Meeting on Longitudinal Studies*, Session 1.5, August 1997, Jerusalem.

GISMONDI R. (1998), *Definizione e classificazione delle unità di rilevazione*, dattiloscritto non pubblicato, Istat, Roma.

GRANQUIST L. (1995), "Improving the Traditional Editing Process", *Business Survey Methods*, 381-385, John Wiley & Sons, New York.

HENNIG C. (1998), "Clustering and Outlier Identification: Fixed Point Cluster Analysis", in Rizzi A.-Vichi M.-Bock H.H., *Advances in Data Science and Classification*, Springer.

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

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

ISTAT (1998), "La nuova indagine sulle vendite al dettaglio: aspetti metodologici e contenuti

innovativi", *Metodi e norme*, 3, Istat, Roma.

KALTON G.-KASPRZYK D.-MCMILLEN D. (1989), "Nonsampling Errors in Panel Surveys", *Panel Surveys*, 249-270, John Wiley & Sons, New York.

KOVAR J.G.-WINKLER W.E. (1996), "Editing Economic Data", *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 81-87.

LATOUCHE M.-MICHAUD S. (1997), "Impact of Different Weighting Schemes in the Measurement of Flows in a Longitudinal Survey", paper presented at the *IASS/IAOS Satellite Meeting on Longitudinal Studies*, Session 1.4, August 1997, Jerusalem.

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

PIZZI C.-PELLIZZARI P. (1998), "Detecting Outliers in Time Series", in Rizzi A.-Vichi M.-Bock H.H., *Advances in Data Science and Classification*, Springer.

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

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

SEARLS D. (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large True Observation", *Journal of the American Statistical Association*, 4, 1200-1204.

SMITH P. (1997), "Winsorisation: an Update", paper presented at the *Work Session on Statistical Data Editing*, 14-17 Ottobre, Praga.

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

THOMPSON K.J.-SIGMAN R.S. (1998), "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data", unpublished paper submitted to the *Journal of Official Statistics*, Bureau of the Census, Washington.

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

WEIR P. (1997), "Data Editing and Performance Measures", paper presented at the *Work Session on Statistical Data Editing*, 14-17 Ottobre, Praga.