Individuazione e trattamento statistico di osservazioni *outlier* per la stima di una variazione in indagini longitudinali

Roberto Gismondi (*)

(*) ISTAT - Servizio SCO

1. Premessa[1]

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[2]. Si farà preferibilmente riferimento ad indagini in cui la suddetta variabile esprime un indicatore economico riferito a dati d'impresa (ricavi, valore aggiunto, 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.

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 pongono conseguentemente i problemi, ampiamente dibattuti in letteratura[3], di:

a) come identificare le unità anomale, o *outlier*;
b) come trattare le osservazioni effettivamente identificate come *outlier*.

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. Inoltre 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[4]. 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

[1] Le opinioni espresse in questo articolo non impegnano l'ISTAT e sono da attribuirsi esclusivamente all'autore, che è anche responsabile di eventuali errori od omissioni.
[2] Più precisamente, si supporrà 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.
[3] Si ricorda, tra tanti, il noto saggio di Fellegi e Holt (1976).
[4] In proposito si veda Weir (1997).

campionaria fornendo informazioni su di esse;

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, si utilizzerà il termine *outlier* nella sua massima generalità, anche perchè in pratica non è sempre possibile distinguere con certezza tra le due tipologie suddette; va peraltro notato come l'eventuale persistenza di *outlier* non informativi nella base dati potrebbe inficiare l'efficacia delle procedure di *editing* adottate per l'identificazione degli *outlier* informativi[5].

Con riferimento a tale aspetto, se è possibile ricontattare il rispondente per accertarsi dell'esattezza del dato, i problemi relativi alla prima situazione sono risolvibili immediatamente, mentre se ciò non è possibile – oppure se si ha la conferma di un *outlier* rappresentativo da parte del rispondente – le soluzioni più frequenti in pratica consistono:

a) nel non considerare affatto le unità *outlier* nei calcoli (in altri termini di assegnarle un peso nullo);
b) nel correggere il dato *outlier* sostituendovi un dato "corretto", ad esempio tramite donazione od imputazione della variazione media registrata per le unità non *outlier*;
c) nel non alterare il dato elementare ma nel correggere (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.

In questo contesto si cercherà di:

- valutare il criterio di identificazione degli *outlier* proposto da Hidiroglou e Berthelot (1986), proponendone una versione alternativa (paragrafi 2 e 3);
- proporre e confrontare alcuni criteri per l'identificazione delle soglie di accettazione, tra cui una famiglia di metodi che si basa sulla conoscenza della distribuzione teorica di y (paragrafo

4);

- proporre e confrontare alcuni criteri per ridurre il peso delle osservazioni *outlier* nel processo di stima (paragrafi 5 e 6).

Nel paragrafo 7 verrà inoltre proposta una applicazione ad un caso concreto, particolarmente idoneo per il confronto tra le varie procedure di identificazione e trattamento degli *outlier*.

2. La procedura di individuazione degli *outlier* proposta da Hidiroglou e Berthelot

Il metodo, proposto originariamente da Hidiroglou e Berthelot (1986) e ripreso successivamente da Lee (1995), si basa sul ricorso a soglie di accettazione derivate dai quartili della distribuzione empirica; in proposito, come fatto notare proprio dallo stesso Lee (p.506), il ricorso alternativo a soglie basate sulla media e la varianza empiriche potrebbe risultare molto rischioso in presenza di numerose osservazioni *outlier*, soprattutto se esse sono localizzate prevalentemente ad una coda della distribuzione.

Data una qualunque variabile di interesse *z*, ed indicati con $q_{0,25}$, $q_{0,50}$ e $q_{0,75}$ i primi tre quartili della distribuzione empirica di *z* relativa ad un campione di *n* osservazioni, si possono definire gli scarti interquartili inferiore e superiore, dati dalle relazioni:

$$d_{\inf} = q_{0,50} - q_{0,25} \tag{2.1}$$

$$d_{\sup} = q_{0,75} - q_{0,50} \tag{2.2}$$

e l'intervallo di accettazione di una generica osservazione sarà dato da:

$$A = (q_{0,50} - c_{\inf}\, d_{\inf}\,;\, q_{0,50} + c_{\sup}\, d_{\sup}) = (A_{\inf}\,;\, A_{\sup}) \tag{2.3}$$

dove $c_{\inf}$ e $c_{\sup}$ sono parametri arbitrari, eventualmente diversi; in particolare, se essi sono posti entrambi uguali a 1 l'intervallo di accettazione si riduce allo scarto interquartile[6].

---

[5] Potendo influenzare, ad esempio, il calcolo dei quantili e quindi delle soglie di accettazione, che rischierebbero di risultare esageratamente ampie.

[6] Che, si ricorda, è dato dalla differenza tra il terzo ed il primo quartile.

Poiché però, in pratica, $d_{\inf}$ e $d_{\sup}$ potrebbero risultare molto piccoli – il che comporterebbe un intervallo di accettazione troppo ristretto – viene proposta l'opzione alternativa data dalle relazioni seguenti:

$$d_{\inf} = \max\left(q_{0,50} - q_{0,25}, \left|B\, q_{0,50}\right|\right) \tag{2.4}$$

$$d_{\sup} = \max\left(q_{0,75} - q_{0,50}, \left|B\, q_{0,50}\right|\right) \tag{2.5}$$

dove *B* è un ulteriore parametro arbitrario compreso tra 0 e 1. Al riguardo, la scelta *B*=0,05 è risultata adeguata in molte applicazioni empiriche, come quella illustrata nel paragrafo 7.

In particolare, la metodologia basata sull'intervallo di accettazione definito dalle relazioni (2.3), (2.4) e (2.5), definibile come *metodo dei quartili*, può essere applicata direttamente alla variabile *z* data dal rapporto:

$$r_{ti} = \frac{y_{ti}}{y_{t-k,i}}\,. \tag{2.6}$$

dove y_{ti} indica il valore assunto da *y* sulla unità *i-ma* al tempo *t*. Nella pratica, tale approccio può rivelarsi rischioso nei casi, assai frequenti, in cui i valori r_{ti} siano caratterizzati da una distribuzione fortemente asimmetrica e risultino molto variabili per le unità con piccoli valori di *y*. In effetti ciò comporta che, se si applica il metodo dei quartili direttamente ai rapporti r_{ti}, i rapporti con valori piccoli di *y* saranno più facilmente identificabili come *outlier* sebbene, d'altra parte, siano proprio le unità con grandi valori di *y* a contribuire maggiormente al *trend* complessivo e dovrebbero essere quindi oggetto di una maggiore attenzione. Questo effetto è definito come *masking effect*. Per aggirare il problema, Hidiroglou e Berthelot hanno proposto la seguente trasformazione

$$s_{ti} = \begin{cases} 1 - \dfrac{q_{0,50}}{r_{ti}} & \text{se} \quad 0 < r_{ti} < q_{0,50} \\ \dfrac{r_{ti}}{q_{0,50}} - 1 & \text{se} \quad r_{ti} \geq q_{0,50} \end{cases} \tag{2.7}$$

e, successivamente, l'applicazione del metodo dei quartili alla funzione *z* data dalla funzione

effetto dell'unità *i-ma*:

$$E_{ti} = s_{ti}[\max(y_{ti}; y_{t-k,i})]^V \tag{2.8}$$

con V parametro arbitrario compreso tra 0 e 1. Se V è prossimo a 0, il *masking effect* non è rimosso, mentre al crescere di V l'ordine di grandezza di y assume un'importanza via via più rilevante.

La procedura proposta, per quanto assai utile e di non complessa applicabilità empirica, presenta alcuni inconvenienti, tra i quali:

- non prevede una metodologia per la scelta delle soglie $c_{\inf}$ e $c_{\sup}$, che resta legata a valutazioni sostanzialmente soggettive; analogamente, non sono avanzate ipotesi circa la distribuzione teorica di riferimento per la variabile y, il che non consente di identificare intervalli di accettazione su basi più obiettive e stabili[7];
- si basa su una funzione *effetto* che, essendo data dal prodotto di due componenti (livello e variazione), può assumere valori difficilmente interpretabili; in altri termini, potrebbe non risultare evidente il motivo per il quale una certa unità venga identificata come *outlier* (in funzione del livello, della variazione o di entrambe le componenti?);
- potrebbe comportare l'identificazione di un numero eccessivo di *outlier*, prevedendo la correzione anche in presenza di valori della funzione *effetto* molto piccoli, che potrebbero derivare da valori piccoli sia della variazione che del livello. In realtà, come suggerito nel paragrafo seguente, un criterio alternativo potrebbe essere basato sulla valutazione delle sole unità caratterizzate da un livello di y molto elevato;
- non consente un ricorso agevole a distribuzioni teoriche per i rapporti r_{ti} al fine della determinazione delle soglie di accettazione (cfr. il paragrafo seguente).

3. Una procedura alternativa

La procedura alternativa si basa su una funzione s^* sostanzialmente analoga a quella definita dalla relazione (2.1), a meno della costante pari a 1, per cui si avrà:

[7] Come anche ricordato dallo stesso Lee (*op. cit.*, p.509).

$$s_{ti}^{*} = \begin{cases} \dfrac{q_{0,50}}{r_{ti}} & \text{se} \quad 0 < r_{ti} < q_{0,50} \\ \dfrac{r_{ti}}{q_{0,50}} & \text{se} \quad r_{ti} \geq q_{0,50} \end{cases} \tag{3.1}$$

e tale espressione risulterà sempre non inferiore a 1[8].

I due indicatori, rispettivamente, del "rapporto" tra le due osservazioni e del "livello" associabile all'unità *i-ma* sono definibili tramite la relazione:

$$\begin{cases} E_{tri} = s_{ti*}^{U} \\ E_{tli} = l_{ti}^{V} = [\max(y_{ti}; y_{t-k,i})]^{V} \end{cases} \tag{3.2}$$

dove non si pongono limitazioni teoriche alla variabilità dei parametri *U* e *V*. Si noti come, ponendo *V*=0 e $y_{t-k,i} = 1$ ci si possa ricondurre al caso[9] in cui la variabile di interesse sia il livello y_{ti} piuttosto che il rapporto di variazione r_{ti}, e tale circostanza caratterizza anche il metodo originario di Hidiroglou e Berthelot.

Il criterio consiste nell'individuare come "*outlier* significativo" ogni osservazione tale che, se analizzati separatamente, *sia* E_{tri}, *sia* E_{tli} non superino il test dei quartili, dove con riferimento ad entrambi gli indicatori tale test sarà unidirezionale (ossia la zona di rifiuto sarà identificata dalla sola area a destra della distribuzione). In altri termini, una osservazione caratterizzata da un rapporto di variazione *r* molto basso o molto alto (e quindi di s_{ti}^{*} alto) potrebbe comunque non essere soggetta ad alterazioni nell'ambito della procedura di stima, qualora il suo livello dimensionale *l* non risultasse particolarmente elevato. Tale accorgimento dovrebbe comportare un minor numero di osservazioni identificate come *outlier* – e quindi di alterazioni dei microdati o dei relativi pesi - secondo una impostazione concettuale per certi versi simile a quella che ispira il macroediting (Barcaroli e Luzi, 1995).

In simboli, l'intervallo di accettazione sarà dato per entrambe le funzioni dalle relazioni:

$$A_{\text{sup}} = q_{0,50} + c_{\text{sup}}\, d_{\text{sup}} \tag{3.3}$$

[8] Ovviamente l'eliminazione dell'addendo unitario che compare nella (2.7) semplifica l'interpretazione della funzione, non altera la forma della sua distribuzione, comportandone solo una traslazione. Per un ulteriore, importante utilizzo della (3.1) si rimanda al paragrafo 6.

[9] L'ulteriore condizione *U*=1 è consigliabile, ma non indispensabile.

$$d_{\sup} = \max\left(q_{0,75} - q_{0,50}, \left|B\, q_{0,50}\right|\right). \tag{3.4}$$

Si noti poi come, più in generale, si possa porre:

$$s_{ti}^{*} = r_{ti}^{\alpha}\, q_{0,50}^{\beta} \tag{3.5}$$

da cui deriva la relazione (3.1) come caso particolare ponendo α=-1 e β=1 se $0 < r_{ti} < q_{0,50}$, e ponendo α=1 e β=-1 se $r_{ti} \geq q_{0,50}$.

L'utilità della formulazione (3.5) sta anche nel fatto che include come caso particolare la situazione in cui α=1 e β=0, ossia in cui la variabile di base per la definizione dell'indicatore del rapporto si riduce a r_{ti}, e quindi non si ricorre ad alcun accorgimento per tener conto della asimmetria strutturale della popolazione oggetto di studio.

Se ora si suppone una distribuzione esponenziale negativa[10] per le osservazioni di *y*, è possibile analizzare l'andamento della funzione (3.2) – e quindi derivare gli intervalli di accettazione – facendo riferimento ad alcune distribuzioni di probabilità teoriche note. In effetti alcuni casi utilizzabili in pratica sono quelli sintetizzati nel prospetto seguente, in cui sono anche indicate le distribuzioni teoriche con cui possono essere analizzate le funzioni E_{tri} e E_{tli}, dove per i riferimenti alle distribuzioni *F* di Fisher e Weibull si rimanda all'appendice[11].

Come già evidenziato, le precedenti relazioni sono utili ai fini dell'individuazione delle soglie di accettazione, così come indicato nel paragrafo seguente[12].

La tabella 3.1 seguente riporta un esempio da cui risulta evidente come il ricorso alla procedura alternativa definita dalle relazioni (3.1) e (3.2) possa ridurre drasticamente il numero di osservazioni identificate come *outlier*. Nell'esempio si è posto $c_{\inf} = c_{\sup} = U = V = 1$ e *B*=0,05; inoltre le unità anomale sono evidenziate da un numero 1 nelle colonne che iniziano con la sigla "Out". Ne deriva che la prima procedura (metodo di Hidiroglou e Berthelot, colonna "Out_1") identifica 6 unità anomale (ossia ben il 60% del totale), rispetto ad una sola unità (colonna "Out_4") identificata dalla procedura alternativa (l'ottava, risultata *outlier* anche con la prima procedura); ciò deriva dal fatto che delle 3 unità caratterizzate da rapporti di variazione *r* esterni all'intervallo di accettazione (colonna "Out_2"), solo una presenta un livello di *y*

[10] Tale proposta appare più generale rispetto a quanto suggerito originariamente da Fuller (1987), che ipotizza il ricorso alla distribuzione di Weibull.

[11] L'ipotesi implicita nel caso della funzione rapporto riferita alla *F* di Fisher è che le due variabili a rapporto (ossia i valori che *y* assume su ogni unità *i* nei tempi *t* e (*t*-*k*)) siano indipendenti.

particolarmente elevato (colonna "Out_3").

Prospetto 3.1 – Distribuzione teoriche di riferimento per le variabili E_{tri} e E_{tli}

U	V	E_{tri}	E_{tli}	*Distribuzioni teoriche di riferimento*	
				E_{tri}	E_{tli}
1	0	s_{ti}	1	F di Fisher	-
1	1	s_{ti}	y_{ti} se $y_{ti} \geq y_{t-k,i}$, altrimenti $y_{t-k,i}$	F di Fisher	Esponenziale negativa
1	2	s_{ti}	y_{ti}^2 se $y_{ti} \geq y_{t-k,i}$, altrimenti $y_{t-k,i}^2$	F di Fisher	Weibull

Tabella 3.1 – Un esempio di applicazione della procedura di Hidiroglou e Berthelot e della procedura alternativa.

Unità	y_{t-k}	y_t	r_t	s_t	E_t	**Out_1**	E_{tr}	E_{tl}	Out_2	Out_3	**Out_4**
1	10	12	1,200	0,121	1,458	0	1,121	12	0	0	0
2	10	11	1,100	0,028	0,308	0	1,028	11	0	0	0
3	15	25	1,667	0,558	13,941	1	1,558	25	1	0	0
4	20	19	0,950	-0,126	-2,526	0	1,126	20	0	0	0
5	20	27	1,350	0,262	7,065	1	1,262	27	1	0	0
6	25	22	0,880	-0,216	-5,398	1	1,216	25	0	0	0
7	25	26	1,040	-0,029	-0,750	0	1,029	26	0	0	0
8	25	36	1,440	0,346	12,449	1	1,346	36	1	1	1
9	40	37	0,925	-0,157	-6,270	1	1,157	40	0	1	0
10	60	55	0,917	-0,167	-10,036	1	1,167	60	0	1	0
$q_{0,25}$			0,931		-4,680		1,123	21,250			
$q_{0,50}$			1,070		-0,221		1,162	25,500			
$q_{0,75}$			1,313		5,664		1,250	33,750			
$d_{\inf}$					4,459		0,058	4,250			
$d_{\sup}$					5,884		0,088	8,250			
$c_{\inf}$					1						
$c_{\sup}$					1		1	1			
$A_{\inf}$					-4,680						
$A_{\sup}$					5,664		1,250	33,750			
Stima						**0,990**					**1,085**

Nota: la prima stima (Out_1) si riferisce al metodo di Hidiroglou e Berthelot, la seconda (Out_4) alla procedura alternativa.

L'aspetto problematico della procedura di Hidiroglou e Berthelot è dato dal fatto che risultano *outlier* unità caratterizzate da una variazione elevata ma da un livello non particolarmente elevato (ad esempio, ciò accade per la terza osservazione, il cui livello di *y*, pari a 25, è solo al sesto posto nella graduatoria) e, viceversa, unità caratterizzate da un livello elevato

[12] Il ricorso a distribuzioni teoriche è ampiamente documentato in Barnett (1978), che propone una rassegna di criteri finalizzati alla identificazione degli *outlier* tramite il ricorso a test d'ipotesi.

ma da una variazione non molto distante dalla mediana (ciò accade per la nona osservazione, il cui livello di *y*, pari a 37, è al secondo posto nella graduatoria ma il cui rapporto di variazione, pari a 0,925, non sembra particolarmente anomalo rispetto al valore mediano di 1,070). E' proprio tale caratteristica sostanzialmente indesiderata del metodo a suggerire quantomeno il confronto con tecniche alternative.

4. Individuazione delle soglie

Uno dei problemi basilari per l'identificazione delle osservazioni *outlier* è dato dalla scelta del criterio con cui determinare le soglie di accettazione per i rapporti r_{ti} e, quindi, sulla base della relazione (2.3), dei coefficienti $c_{\inf}$ e $c_{\sup}$. In generale, le soglie possono essere determinate:

a) ricorrendo a valutazioni soggettive, basate generalmente sull'esperienza acquisita relativamente alla distribuzione empirica dei rapporti r_{ti} con riferimento a periodi precedenti a *t* (ad esempio, lo stesso mese dell'anno precedente a quello di riferimento), o a fenomeni similari (Garcia e Peirats, 1994), come suggerito nel primo dei criteri di seguito proposti[13];
b) imponendo che le code della distribuzione empirica sottendano una quota prefissata delle frequenze osservate (ISTAT, 1998), come suggerito nel secondo dei criteri di seguito proposti;
c) ipotizzando una distribuzione teorica per la variabile *y* – e/o per i rapporti r_{ti} – anch'essa derivata da conoscenze acquisite sul fenomeno studiato (Granquist, 1995; Pizzi e Pellizzari, 1998), come illustrato nel terzo dei suddetti criteri;
d) supponendo di disporre della distribuzione delle variazioni tendenziali per l'intera popolazione con riferimento ad un tempo precedente a *t*, o ad una variabile ausiliaria *z* (nota) correlata con la variabile *y* oggetto di interesse e riferita la tempo *t*, come suggerito nel quarto dei suddetti criteri;
e) sulla base di variabili ausiliarie (note) ed il ricorso ad algoritmi complessi, come proposto da Drapler e Winkler (1997) e Thompson (1998) con riferimento al programma *SPEER*.

In particolare, viene proposta questa griglia di metodi, descritti con riferimento alla procedura alternativa del paragrafo precedente.

Criterio 1: ricorso a soglie prefissate

Ricordando la definizione dell'intervallo di accettazione data dalla (2.3), questo criterio consiste nel prefissare, nel caso delle variabili E_{tri} e E_{tli}, la soglia di accettazione $A_{\sup} = (q_{0,50} - c_{\sup} d_{\sup})$, da cui si ricava immediatamente $c_{\sup}$.

Criterio 2: ricorso alla distribuzione empirica

Sulla base di tale criterio, riguardo alle funzioni E_{tri} e E_{tli}, la soglia $c_{\sup}$ può essere identificata imponendo che alla coda della distribuzione empirica venga lasciata una quota di frequenze osservate pari a P.

Criterio 3: ricorso alla distribuzione teorica

Si può dimostrare[14] che se U=1, l'estremo superiore dell'intervallo di accettazione per la variabile E_{tri} sarà definito dalle relazioni:

$$q_{0,50} \frac{E(y_{t-k})}{E(y_t)} F_{\sup,P} \qquad \text{se} \qquad 0 < r_{ti} < q_{0,50} \tag{4.1}$$

$$\frac{E(y_t)}{q_{0,50} E(y_{t-k})} F_{\sup,P} \qquad \text{se} \qquad r_{ti} \geq q_{0,50} \tag{4.2}$$

dove il simbolo $E(\cdot)$ indica la speranza matematica, $F_{\sup,P}$ è il percentile della distribuzione F di Fisher (con entrambi i gradi di libertà pari a 1) che lascia alla sua destra una probabilità pari a P, dove si può porre P=0,05 o P=0,10. Tali estremi rappresentano automaticamente le soglie di accettazione per la variabile rapporto E_{tri}.

Per quanto riguarda la funzione E_{tli}, l'intervallo di accettazione sarà definito da tutti i valori non superiori alla soglia:

$$PERC_{\sup,P} \tag{4.3}$$

---

[13] Ricade in questa famiglia di procedure il caso in cui si prefissino delle soglie invarianti al variare del tempo *t*.
[14] Si rimanda al primo sottoparagrafo dell'appendice.

dove il percentile $PERC_{\sup,P}$ si riferirà alla distribuzione esponenziale negativa di parametro α=1/E(E_{tli}) se V=1 ed alla distribuzione di Weibull di parametri α=[1/E(E_{tli})]^2 e $\beta = 0{,}5$ se V=2[15].

In pratica, potrebbe essere necessario "adattare" preliminarmente l'intervallo di definizione della funzione F alla base dati disponibile, limitandone a priori il campo di variazione in funzione della variabilità intrinseca della variabile r osservata, al fine di non ricavare intervalli di accettazione irrealisticamente ampi. Ad esempio, sulla base dei dati relativi all'indagine mensile sulle vendite al dettaglio – e che saranno oggetto dell'applicazione del paragrafo 7 – il rapporto r tra i fatturati relativi allo stesso mese di due anni consecutivi mostra una certa concentrazione[16] nell'intervallo compreso tra l'estremo inferiore 0,603 e l'estremo superiore 1,626: troncando la curva F in tale intervallo si ricava, ponendo P=0,05, $F_{\sup}$=1,47. Riguardo alla soglia di accettazione per il livello, posto V=1, se il fatturato medio per impresa è pari a 11,7 milioni di lire sulla base del valore atteso della funzione esponenziale negativa, si ricava una soglia superiore pari a circa 32 milioni.

Va infine notato come possa essere facilmente individuato un legame formale tra i criteri 2 e 3: con riferimento alla funzione E_{tri}, ricordando la relazione (2.3) si può porre la seguente identità:

$$q_{0,50} + c_{\sup} d_{\sup} = q_{0,50} \frac{E(y_t)}{E(y_{t-k})} F_{\sup,P}$$

da cui si ricava il valore di $c_{\sup}$, dato dalla relazione:

$$c_{\sup} = \frac{q_{0,50}}{d_{\sup}} \left[\frac{E(y_t) F_{\sup,P/2} - E(y_{t-k})}{E(y_{t-k})} \right]. \qquad (4.4)$$

Riguardo invece alla funzione E_{tli} la soglia di destra può essere identificata ponendo l'identità:

[15] Si rimanda al secondo sottoparagrafo dell'appendice.
[16] Nell'intervallo citato cade infatti il 90% dei casi analizzati nel paragrafo 7. Se si fosse preso in considerazione il 98% dei casi, si sarebbe ricavato il valore 2,17 per il percentile della funzione F.

$$q_{0,50} + c_{\sup} d_{\sup} = PERC_{\sup,P} \qquad \text{da cui si ottiene:} \qquad c_{\sup} = \frac{PERC_{\sup,P} - q_{0,50}}{d_{\sup}}. \qquad (4.5)$$

Criterio 4: ricorso a dati elementari

In alcuni casi si verifica la situazione seguente: con riferimento ad un tempo t_0 antecedente al tempo *t* di riferimento, si dispone:

- di un sottoinsieme di osservazioni di una variabile *z*, misurate su unità appartenenti alla medesima popolazione studiata al tempo *t* ed ottenute in condizioni "simili" a quelle del tempo *t* (stesse definizioni, stesso disegno campionario, stessa tecnica di compilazione dei questionari e di invio dei dati, ecc.);
- del valore noto R_{z,t_0} della variazione della variabile *z* intercorsa tra i tempi t_0 e (t_0-*k*) per l'intera popolazione di riferimento.

Se si indica con $r_{z,ti}$ la variazione relativa alla variabile *z* intercorsa tra i tempi *t* e (*t*-*k*) e misurata sulla unità *i-ma* di un sottoinsieme *S*, l'intervallo di accettazione *A* – dove *A* è definito dalle condizioni (2.3) e (2.4) – potrà essere determinato imponendo la seguente condizione di minimo:

$$\left| \sum_{r_{z,t0i} \in A} (r_{z,t_0 i})\, D_{zi} - R_{z,t_0} \right| = \text{minimo} \qquad (4.6)$$

dove i termini D_{zi} indicano degli opportuni pesi campionari, definibili in modo analogo a quanto sarà illustrato nella successiva formula (6.1).

L'identificazione dell'intervallo *A* di accettazione – e quindi dei coefficienti $c_{\inf}$ e $c_{\sup}$ – per i rapporti di variazione tra i tempi t_0 e (t_0-*k*) di *z* consentirà di ricavare anche l'intervallo di accettazione per gli omologhi rapporti di variazione[17] r_{ti} relativi a *y* nei casi in cui:

- le due variabili possano ritenersi molto correlate (ad esempio, *y* e *z* possono riferirsi al fatturato delle imprese commerciali di due provincie limitrofe del Veneto);
- le due variabili siano caratterizzate da ordini di grandezza simili (ad esempio, valore aggiunto e fatturato).

[17] Con le opportune trasformazioni la procedura può essere applicata anche alle funzioni (2.7) e (3.1).

Un esempio concreto in tal senso si riferisce al caso in cui si ha proprio $z=y$. L'attuale rilevazione mensile sul movimento turistico svolta correntemente dall'ISTAT è di tipo esaustivo e consente la diffusione dei dati definitivi sugli arrivi e le presenze nelle strutture ricettive con un ritardo di circa 9 mesi dalla fine dell'ultimo mese di riferimento. In precedenza, vengono diffuse stime provvisorie a 90 giorni, 120 giorni, e così via secondo aggiornamenti mensili che si avvalgono dei dati via via pervenuti. Uno dei controlli di qualità fondamentali per la revisione dei questionari si basa sul confronto tra i dati relativi allo stesso mese di due anni successivi (k=12), ossia sui rapporti $r_{t_0 i} = (y_{t_0 i} / y_{t_0 - 12, i})$. Se dopo circa 9 mesi è noto il valore relativo alla variazione complessiva (e definitiva) R_{t_0}, le soglie $c_{\inf}$ e $c_{\sup}$ possono essere determinate a posteriori imponendo che la stima di R_{t_0} effettuata in un qualunque periodo compreso tra i 3 mesi ed i 9 mesi successivi a t_0 sia il più possibile simile al suddetto valore noto, secondo quanto espresso nella formula (4.6). Ovviamente, tale valutazione sarà effettuata con riferimento ad un anno A precedente a quello effettivo di interesse, di modo che l'intera procedura assuma un significato operativo concreto. Così, ad esempio, le soglie per il rapporto tra le presenze nelle strutture ricettive a gennaio 1999 ed a gennaio 1998 possono essere poste uguali alle soglie riferite al rapporto tra le presenze nelle strutture ricettive a gennaio 1998 ed a gennaio 1997, individuate a posteriori una volta noto il rapporto di variazione definitivo, disponibile alla fine di settembre del 1998.

5. Metodi di stima

Va sottolineato come sussistano almeno due caratteri peculiari delle procedure di trattamento degli *outlier* basate sul ricorso alle soglie di accettazione:

- nella maggioranza dei casi (tra cui quelli contemplati nei precedenti paragrafi 2 e 3), se tali procedure definiscono le soglie in funzione delle osservazioni campionarie essi identificheranno *sempre* almeno una osservazione *outlier*;
- la presenza anche di un solo *outlier* caratterizzato da un livello molto elevato può influenzare notevolmente l'identificazione delle soglie di accettazione.

Entrambe le peculiarità potrebbero non essere desiderabili. Nel primo caso, se si opera in strati con pochi rispondenti, l'identificazione delle soglie basata sui dati campionari potrebbe risultare problematica (ad esempio, per la difficoltà di stimare correttamente i quantili), ed inoltre l'identificazione di troppi *outlier* finirebbe con: a) il depauperare ulteriormente una base dati già

esigua se si decidesse si assegnare un peso nullo a tali osservazioni in fase di stima; b) comportare problemi di stima qualora si decidesse di correggere a posteriori il valore *outlier* sulla base delle poche osservazioni disponibili dello strato non risultate *outlier*. Nel secondo caso la natura del problema risulta evidente, e potrebbe complicarsi ulteriormente in presenza di strati di piccola dimensione.

Alla luce di queste considerazioni, assume un certo rilievo il confronto tra le tecniche basate sull'identificazione preliminare e l'eventuale correzione degli *outlier* e le tecniche di riponderazione delle osservazioni disponibili.

Se si decide di modificare l'osservazione y_{ti} ritenuta anomala, moltiplicando tale valore per il fattore w_{1i}, si utilizzerà il nuovo valore:

$$y_{ti}\, w_{1i} \qquad \text{in luogo del valore originario} \qquad y_{ti}\,. \tag{5.1}$$

D'altra parte, se in fase di stima si assegna alla *i-ma* unità *outlier* un peso diverso dal peso originario p_i (dunque senza alterare il dato di base), ricorrendo al fattore w_{2i}, si utilizzerà il nuovo peso:

$$p_i\, w_{2i} \qquad \text{in luogo del peso originario} \qquad p_i\,. \tag{5.2}$$

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 alla prima tipologia 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[18].

Va comunque notato che, in pratica, ad ogni procedura di correzione dei dati anomali di tipo (5.1) corrisponde implicitamente una riponderazione di tipo (5.2): infatti, se dopo la correzione (5.1) il nuovo microdato è dato da $y_{ti} w_{1i}$ e dopo la riponderazione (5.2) il nuovo peso associato alla unità *i-ma* è dato da $p_i w_{2i}$ - per cui il prodotto tra il valore originario di *y* associato a tale unità ed il nuovo peso è dato da $y_i p_i w_{2i}$ - la corrispondenza tra le procedure (5.1) e (5.2) è stabilita dalla relazione:

$$w_{1i} = p_i w_{2i}\,, \tag{5.3}$$

ossia il criterio della riponderazione tramite il fattore w_{2i} equivale al criterio della modifica del valore *outlier* sulla base del fattore dato dal membro di destra della (5.3).

La disponibilità di basi di dati longitudinali anche non particolarmente lunghe comporta spesso il ricorso al generico stimatore così definito:

$$\hat{y}_{ti} = y_{t-k,i}\left(\frac{\bar{y}_t}{\bar{y}_{t-k}}\right)\left(\frac{x_{ti}}{x_{t-k,i}}\right)(1+\varepsilon_{ti}) \tag{5.4}$$

dove la quantità a destra del segno di uguale è data dal prodotto tra:

- il rapporto (prima parentesi) tra il valore medio campionario di *y* misurato *sulle sole unità non outlier* ai tempi *t* e (*t-k*);
- il rapporto (seconda parentesi) tra il valore di una variabile ausiliaria *x* misurato sulla *i-ma* unità *outlier* ai tempi *t* e (*t-k*);
- una componente (terza parentesi) basata su una ulteriore variabile ε[19].

La formula precedente può assumere diverse forme – alcune delle quali saranno applicate nell'esempio del paragrafo 7 – tra le quali si menzionano le seguenti:

1. se x=1 e ε=0 per ogni unità *i*, si sostituisce l'osservazione y_{ti} *outlier* con la media

[18] 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.

[19] L'introduzione della variabile ε serve a ridurre il rischio di un eccessivo appiattimento della distribuzione disponibile dopo le correzioni.

campionaria calcolata sulle sole unità "buone"[20]; qualora ε rappresentasse una variabile casuale a media nulla e varianza costante al variare di *i*, si correggerebbe tale stima con un effetto casuale;

2. se *x* varia al variare di *i*, 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, se *y* rappresenta il fatturato, *x* dato dal 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* (Gismondi, *op. cit.*, 1996);
3. se si pone $x_{ti} = (A_{\sup} / \bar{y}_t)$ e $x_{t-k,i} = (1/\bar{y}_{t-k})$ nel caso in cui $r_{ti} > A_{\sup}$, e si pone invece $x_{ti} = (A_{\inf} / \bar{y}_t)$ ed ancora $x_{t-k,i} = (1/\bar{y}_{t-k})$ nel caso in cui $r_{ti} < A_{\inf}$ si ricorre, in pratica, ad un troncamento (Searls, 1966), tramite il quale si sostituisce il valore anomalo con uno degli estremi dell'intervallo di accettazione; tale procedura può essere iterata aggiungendo l'effetto indotto dalla componente casuale ε;
4. se si pone $x_{ti} = t_{td(i)} / \bar{y}_t$ e $x_{t-k,i} = (t_{t-k,d(i)} / \bar{y}_{t-k})$, dove il pedice indica l'unità *d* scelta come donatrice rispetto alla *i-ma* unità *outlier*, si sostituisce l'osservazione *outlier* con il valore di una unità donatrice, scelta con un qualche criterio di distanza minima (Gismondi, 1999).

6. Criteri basati sulla modifica dei pesi campionari

Come accennato nella premessa, in presenza di *outlier* rappresentativi potrebbe essere auspicabile non introdurre alterazioni *ad hoc* su tali osservazioni che, per quanto anomale, risultano "vere". D'altra parte, per evitare di introdurre distorsioni nella procedura di stima, il peso associato a tali unità dovrebbe essere cautelativamente modificato (generalmente viene diminuito) sulla base di procedure come quella descritta in questo paragrafo.

Preliminarmente va ricordato che una soluzione non infrequente in pratica consiste nel "congelare" le osservazioni *outlier*, ossia non considerarle per la stima assegnando loro un peso nullo[21]. Tale procedura potrebbe però rivelarsi dannosa se si opera in strati poco numerosi, in cui la strategia più efficiente potrebbe consistere nell'utilizzare comunque tutte le unità disponibili ai fini della stima della variazione.

[20] Cfr. Hidiroglou e Berthelot (*op. cit.*, p.79).

In tale ottica, il criterio proposto nel prosieguo aggira il problema di dover predefinire un intervallo di accettazione, come commentato nei paragrafi precedenti. In altri termini, in questo contesto si supporrà di non trattare separatamente i due sottoinsiemi degli indici *outlier* e degli indici "buoni", bensì di procedere alla rideterminazione dei pesi di tutte le unità secondo la semplice metodologia descritta. 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).

Preliminarmente è essenziale far notare come, se è disponibile un campione *S* di *n* osservazioni relative alla variabile *y* (dove ogni unità è caratterizzata da una probabilità di inclusione pari a π_i), un generico stimatore del rapporto tra due ammontari relativi alla medesima variabile misurata ai tempi *t* e (*t*-*k*) sia scrivibile in questa forma:

$$\hat{R}_t = \frac{\sum_{i\in S} \frac{y_{ti}}{\pi_i}}{\sum_{i\in S} \frac{y_{t-k,i}}{\pi_i}} = \sum_{i\in S} \left(\frac{y_{ti}}{y_{t-k,i}} \right) \left(\frac{\frac{y_{t-k,i}}{\pi_i}}{\sum_{i\in S} \frac{y_{t-k,i}}{\pi_i}} \right) = \sum_{i\in S} (r_{ti})\, D_{ti} \qquad (6.1)$$

dove per la determinazione dei valori D_{ti} si può supporre che $y_{t-k,i}$ sia noto, oppure che tale variabile sia approssimabile con la variabile ausiliaria $x_{t-k,i}$ (anche in questo contesto può valere l'esempio di fatturato ed addetti). Dunque tale stimatore è scrivibile nella forma di una media aritmetica ponderata[22] delle variazioni tendenziali individuali osservate sulle unità del campione, con pesi D_{ti} direttamente proporzionali alla dimensione di *y* sulla unità *i-ma* al tempo (*t*-*k*) – supposta nota o stimabile con una variabile *x* ad essa correlata – ed inversamente proporzionali alla probabilità di inclusione. In particolare, nel caso in cui le probabilità di inclusione siano costanti al variare di *i* formalmente si è in presenza di un disegno campionario di tipo *PPS*, basato sulla variabile $y_{t-k,i}$. Il peso D_{ti} è dunque il peso originario associato alla unità *i-ma* e derivato dal disegno campionario originale[23].

---

[21] In tale circostanza può risultare conveniente normalizzare i pesi associati alle rimanenti unità non *outlier*, imponendo che la loro somma sia uguale alla somma dei pesi originari (ossia calcolata includendo anche gli *outlier*), e a tale fine è sufficiente un semplice riproporzionamento.

[22] La somma dei pesi sarà generalmente diversa da uno.

[23] Per approfondimenti si rimanda a Tremblay (1986).

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 *i-ma* dovrebbe essere assegnato un peso decrescente tanto più il valore assunto da tale unità si discosta dalla mediana calcolata sulle unità rispondenti dello strato[24]. Se tutte le unità avessero approssimativamente lo stesso valore di *y*, non sarebbe necessaria alcuna modifica sostanziale dei pesi originari.

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

$$W_{ti} = \frac{D_{ti}}{s^*_{ti}}. \qquad (6.2)$$

Tale peso sarà sempre non inferiore a 1 ed assumerà valori crescenti al crescere dello scostamento tra il valore osservato sulla unità *i-ma* ed il valore mediano, mentre *W* tenderà a coincidere con *D* per valori della variazione *r* prossimi alla mediana. Il criterio (6.2) implica comunque una modifica di tutti i pesi originari, anche quelli delle unità non identificate come *outlier*, sebbene le modifiche dovrebbero risultare assai ridotte per valori di *r* prossimi alla mediana.

Si noti, infine, come con la formula (6.2) si finisca, in pratica, con il calcolare la media ponderata con pesi sempre pari a D_{ti} dei nuovi rapporti dati da $(r_{ti}^2 / q_{0,50})$ se $0 < r_{ti} < q_{0,50}$ e semplicemente da $q_{0,50}$ altrimenti.

I nuovi pesi W_{ti} determinati sulla base della formula (6.2) potrebbero essere inseriti direttamente nella (6.1) al posto dei pesi originari D_{ti}.

Alternativamente, potrebbe essere conveniente normalizzare i nuovi pesi imponendo, ad esempio, che la loro somma W_t^* sia pari alla somma dei pesi originari D_{ti}, a sua volta pari a 1 sulla base della loro stessa definizione.

Per garantire tale condizione si può ricorrere ad un semplice riproporzionamento, sulla base della formula:

[24] 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.

$$W_{ti}^{*} = \left(\frac{W_{ti}}{W_t} \right) \tag{6.3}$$

Alternativamente si può ricorrere ad un criterio ottimale, che consiste nel determinare, a partire dai pesi W_{ti}, i pesi finali W_{ti}^{*} che risultino il più possibile simili, in media, ai corrispondenti pesi W_{ti} e tali che la somma dei pesi W_t^{*} sia ancora pari a 1. Tale criterio si basa dunque sull'ipotesi che una buona rideterminazione dei pesi dovrebbe comunque alterare il meno possibile i pesi campionari originari[25], e la scelta di una funzione di perdita data dalla somma dei quadrati degli scarti tra pesi originari e nuovi pesi comporta, definite le seguenti grandezze:

$$\sum_{i \in S} D_{ti} = 1 \qquad \text{e} \qquad \sum_{i \in S} W_{ti} = W_t$$

che la funzione di Lagrange da minimizzare è data da:

$$\Phi^2 = \sum_{i \in S} (W_{ti}^{*} - W_{ti})^2 + \lambda \left(\sum_{i \in S} W_{ti}^{*} - 1 \right). \tag{6.4}$$

Derivando la funzione (6.4) rispetto a W_{ti}^{*} ed uguagliando a zero si ottiene la relazione:

$$W_{ti}^{*} = W_{ti} + \frac{\lambda}{2} \qquad \rightarrow \qquad \lambda = \frac{2(1 - W_t)}{n}$$

dalle quali consegue la relazione finale:

$$W_{ti}^{*} = W_{ti} + \frac{(1 - W_t)}{n}. \tag{6.5}$$

I nuovi pesi W_{ti}^{*} saranno tutti non negativi, perché il numeratore del secondo addendo della (6.5) sarà sempre non negativo, dato che $s_{ti}^{*} \geq 1$, per cui dalla (6.2) si avrà che $W_{ti} \leq D_{ti}$ e quindi, sommando rispetto a tutte le unità del campione, $\sum_{i \in S} W_{ti} = W_t \leq 1$.

[25] Peraltro tale ipotesi andrà accuratamente verificata in pratica, perché potrebbe condurre – in modo solo apparentemente paradossale – a risultati peggiori rispetto a quelli ottenibili con il criterio definito dalla (6.3). In proposito si rimanda al paragrafo 7.

Infine, un criterio che comporta una generalizzazione della formula (6.5)[26] si basa sulla funzione di Lagrange data dalla relazione:

$$\Phi^2 = \sum_{i \in S} (W_{ti}^* - W_{ti})^2 + \lambda \left(\sum_{i \in S} W_{ti}^* z_{ti} - z_t^* \right). \tag{6.6}$$

dove *z* è una variabile ausiliaria correlata con la variabile rapporto *r* e di cui è noto l'ammontare complessivo z_t^* nell'intera popolazione di riferimento al tempo *t*, laddove si ha che $\sum_{i \in S} W_{ti} z_{ti} = z_t$. La soluzione ottimale è data dalla relazione:

$$W_{ti}^* = W_{ti} + z_{ti} \left[\frac{(z_t^* - z_t)}{\sum_{i \in S} z_{ti}^2} \right] \tag{6.7}$$

che si riduce nuovamente alla (6.5) per $z_{ti} = 1$.

Nella tabella 6.1 sono riportati i risultati salienti di un esempio concreto relativo all'uso di criteri di riponderazione. Si è supposto un disegno campionario casuale semplice basato su una popolazione di dimensione *N*=100 e un campione di dimensione *n*=30, per cui per ciascuna unità si avrà $\pi_i = 0{,}333$. Si supporrà inoltre che il rapporto di variazione "vero" sia assimilabile a quello ottenibile considerando solo le prime 24 unità (ossia le unità "buone"), per semplicità tutte caratterizzate da ammontari di *y* pari a 50000 nel periodo (*t-k*) ed a 55000 nel periodo *t*, per cui il rapporto *r* "vero" sarà pari a 1,100. Sono state quindi ipotizzate 4 situazioni relativamente alla natura degli *outlier*, ossia delle unità da 25 a 30:

1. 3 rapporti di variazione *r* sono molto bassi (0,200) e 3 sono molto alti (5,000), secondo intensità simili, dato che si è in presenza di 3 diminuzioni del 500% e di 3 incrementi del 500%;
2. 3 rapporti *r* molto bassi e 3 molto alti, con una intensità media dei rapporti alti più elevata di quella dei rapporti bassi. Un caso del genere può presentarsi a causa di *outlier* non informativi (derivati ad esempio da errori nell'unità di misura utilizzata dal compilatore o da errori di registrazione);
3. 6 rapporti *r* molto elevati, anch'essi dovuti ad *outlier* non informativi, oppure a crescite anormalmente alte registrate da alcune unità;

[26] Vale la medesima osservazione della nota precedente.

4. 6 rapporti del tutto “normali”, ma derivati da coppie di dichiarazioni errate, dovute anche in questo caso ad errori di compilazione o di registrazione.

Su queste 4 basi di dati sono stati confrontati:

a) lo stimatore senza correttivi per gli *outlier* (6.1);
b) lo stimatore (6.2) basato sui nuovi pesi *W*;
c) lo stimatore (6.3) basato sui pesi *W* riproporzionati;
d) lo stimatore (6.4) basato sui pesi *W* ottimali.

Come indicatore di qualità è stata calcolata la differenza (in valore assoluto) tra la variazione vera (pari come già detto a 1,100) e quella ottenuta riponderando le 6 unità *outlier* con i vari criteri.

In tutte le circostanze, come prevedibile, lo stimatore meno preciso è risultato il primo, sebbene con l’eccezione del quarto caso, in cui la maggiore imprecisione spetta al secondo criterio: ciò deriva dal fatto che la presenza di 6 ammontari *outlier* che *non generano* rapporti anomali induce una riponderazione sostanzialmente inutile, a meno che non venga corretta sulla base dei criteri c) e d).

Il criterio d) non presenta *performance* particolarmente buone (è al terzo posto nei primi 3 casi), con la sola suddetta eccezione del quarto caso, che dovrebbe quindi risultare l’unico (o uno dei pochi) in cui, in pratica, può essere consigliabile il suo utilizzo.

Il criterio che nel complesso è caratterizzato dalla affidabilità più elevata è il terzo, basato sul riproporzionamento dei nuovi pesi *W*. Tale criterio è risultato il migliore nel primo caso e si è posizionato al secondo posto negli altri casi. Va comunque notato che, in effetti, se si esclude il quarto caso (in cui peraltro un confronto significativo tra i vari livelli di errore è possibile solo a partire dalla quinta cifra decimale), il criterio migliore risulterebbe il secondo: infatti il criterio b) si posiziona al primo posto nel secondo e nel terzo caso, ossia in presenza di rapporti fortemente anomali, laddove evidentemente il riproporzionamento rispetto alla somma dei pesi originari indotto dal criterio c) finisce con il peggiorare la qualità della stima.

In conclusione, il ricorso alla correzione dei pesi originari basata sui valori s^* sembra complessivamente utile, come sarà anche confermato dall’applicazione ad un caso concreto illustrata nel paragrafo seguente.

Tabella 6.1 – Un confronto empirico tra alcuni criteri di modifica dei pesi campionari in presenza di osservazioni *outlier*.

Unità	y_{t-k}	y_t	r_t	D_t	$W_t = D_t / s_t^*$	W_t^*	W_t^* (6.5)
Da 1 a 24	50000	55000	1,100	0,03226	0,03226	0,03816	0,03741
25	50000	10000	0,200	0,03226	0,00587	0,00694	0,01102
26	50000	10000	0,200	0,03226	0,00587	0,00694	0,01102
27	50000	10000	0,200	0,03226	0,00587	0,00694	0,01102
28	50000	250000	5,000	0,03226	0,00710	0,00840	0,01225
29	50000	250000	5,000	0,03226	0,00710	0,00840	0,01225
30	50000	250000	5,000	0,03226	0,00710	0,00840	0,01225
Scarti				**0,29032**	**0,10293**	**0,07949**	**0,11927**
Da 1 a 24	50000	55000	1,100	0,03221	0,03221	0,03995	0,03867
25	50000	54300000	1086,000	0,03214	0,00003	0,00004	0,00655
26	52360	50448000	963,475	0,03366	0,00004	0,00005	0,00655
27	49754	52610000	1057,413	0,03198	0,00003	0,00004	0,00655
28	51519	527	0,010	0,03312	0,00031	0,00038	0,00682
29	49902	529	0,011	0,03208	0,00031	0,00038	0,00682
30	52100	520	0,010	0,03349	0,00030	0,00038	0,00682
Scarti				**100,93840**	**0,10855**	**0,13230**	**20,31164**
Da 1 a 24	50000	55000	1,100	0,03999	0,03999	0,03999	0,03999
25	50	54678	1093,560	0,00004	0,00000	0,00000	0,00001
26	51	53478	1048,588	0,00004	0,00000	0,00000	0,00001
27	43	53267	1238,767	0,00003	0,00000	0,00000	0,00001
28	47	52315	1113,085	0,00004	0,00000	0,00000	0,00001
29	39	45324	1162,154	0,00003	0,00000	0,00000	0,00001
30	48	54267	1130,563	0,00004	0,00000	0,00000	0,00001
Scarti				**0,25036**	**0,00000**	**0,00024**	**0,05046**
Da 1 a 24	50000	55000	1,100	0,04000	0,04000	0,04000	0,04000
25	50	54	1,080	0,00004	0,00004	0,00004	0,00004
26	46	45	0,986	0,00004	0,00003	0,00003	0,00003
27	52	52	0,992	0,00004	0,00004	0,00004	0,00004
28	49	50	1,010	0,00004	0,00004	0,00004	0,00004
29	47	46	0,986	0,00004	0,00003	0,00003	0,00003
30	52	51	0,978	0,00004	0,00004	0,00004	0,00004
Scarti				**0,0000223**	**0,0000424**	**0,0000201**	**0,0000198**

7. Una applicazione

L'ISTAT rileva ogni mese l'ammontare delle vendite al dettaglio presso un campione di circa 8.000 imprese commerciali, e sulla base delle informazioni raccolte elabora e diffonde una serie di indici che esprimono, per ogni strato, la variazione delle vendite intercorsa tra un certo mese e la media mensile dell'anno base 1995.

In tale contesto, l'informazione congiunturale di maggiore rilievo è però data dalla variazione relativa intercorrente tra l'indice di un certo mese *m* e l'indice dello stesso mese *m* riferito all'anno precedente (la cosiddetta *variazione tendenziale*). Inoltre, a causa della forte

stagionalità delle vendite la verifica qualitativa dei dati raccolti in un dato mese si basa proprio sul calcolo, per ogni impresa, del suddetto rapporto tendenziale, la cui variabilità intrinseca varierà, in genere, in funzione sia dello strato di appartenenza, sia del mese di riferimento.

Dopo aver eliminato gli eventuali *outlier* non informativi, il problema consiste nel definire una strategia per l'identificazione dei rapporti tendenziali anomali e, successivamente, per il loro trattamento in sede di stima di una variazione. Tale variazione è espressa come rapporto tra il valore medio delle vendite relative al mese *m* dell'anno *A* ed il valore medio delle vendite del mese *m* dell'anno (*A*-1): l'indice a base 1995=100 sarà successivamente calcolabile tramite una procedura concatenata, ossia moltiplicando il suddetto rapporto per l'indice a base 1995=100 relativo al mese (*m*-1)[27].

In questa applicazione è stata considerata la base dati disponibile con riferimento al mese di giugno 1999, in quanto nell'arco di tale anno è proprio in questo mese che si è registrato il più alto numero di risposte a cui è stato possibile associare anche l'informazione sulle vendite realizzate nel medesimo mese del 1998[28]. Inoltre il mese di giugno è, in genere, meno affetto da fattori esogeni che possono incidere sulla comparabilità delle vendite registrate nello stesso mese di due anni successivi, come ad esempio gli effetti di calendario (che riguardano i mesi di marzo ed aprile, qualora la Pasqua cadesse un anno in un mese e l'anno successivo nell'altro) o gli effetti stagionali (che riguardano soprattutto agosto e dicembre). Tale mese dovrebbe dunque risultare piuttosto "stabile", per cui i risultati ottenuti, pur non essendo immediatamente generalizzabili, possono comunque ritenersi fortemente indicativi.

Sebbene la stratificazione originaria adottata per l'indagine sulle vendite preveda circa 170 domini per i quali ogni mese viene calcolato un indice distinto, in questo contesto è stata adottata una stratificazione semplificata, basata su 20 domini, definiti nel prospetto 7.1 seguente.

Prospetto 7.1 – Descrizione dei 20 strati utilizzati per l'applicazione al caso delle vendite al dettaglio

Attività prevalente dell'impresa	**Classi di addetti**				
	1-2	*3-5*	*6-9*	*10-19*	*>19*
Impresa specializzata alimentare	1 (315)	2 (63)	3 (22)	4 (31)	5 (16)
Impresa specializzata non alimentare	6 (1587)	7(391)	8 (212)	9 (226)	10 (135)

[27] Per ulteriori dettagli si rimanda a ISTAT (1998).

[28] L'indagine sulle vendite si caratterizza per una elevato tasso di non risposta – dovuto tanto alla natura estremamente polverizzata del dominio osservato quanto alla necessità di dover diffondere gli indici delle vendite a distanza di appena un mese l'uno dall'altro – e prevede ogni anno una rotazione parziale delle imprese (il tasso di rotazione è pari a circa 1/3). Di conseguenza, anche qualora un'impresa rispondesse in un dato mese *m*, potrebbe non essere disponibile l'informazione relativa al mese *m* dell'anno precedente.

Impresa non specializzata alimentare	11 (79)	12 (37)	13 (58)	14 (81)	15 (97)
Impresa non specializzata non alimentare	16 (9)	17 (5)	18 (11)	19 (4)	20 (25)

Sono state considerate 2 tipologie di imprese commerciali al dettaglio: le imprese specializzate (ossia quelle che vendono esclusivamente o in prevalenza una sola tipologia di prodotti) e quelle non specializzate (che equivalgono con buona apprtossimazione alle imprese operanti con punti di vendita della grande distribuzione), a loro volta distinte in base al fatto che la tipologia dei prodotti venduti esclusivamente o in prevalenza sia di tipo alimentare o non alimentare. L'ulteriore elemento di stratificazione è dato dalla dimensione aziendale, misurata sulla base delle 5 classi di addetti 1-2, 3-5, 6-9, 10-19 e da 20 in poi. I 20 strati così definiti sono indicati con i numeri da 1 a 20 contenuti nel prospetto: alla destra di ogni numero è riportata, tra parentesi, la numerosità dello strato con riferimento al campione di imprese rispondenti a giugno 1999 e di cui fosse noto il valore delle vendite anche a giugno 1998.

I 3.404 rapporti tendenziali $r_t=y_t/y_{t-12}$, stratificati secondo i 20 domini appena descritti, sono stati sottoposti a diversi criteri di *editing*, elencati nella testata della tabella 7.1. In tale tabella è riportato il numero di unità identificate come *outlier*, separatamente per ciascuno dei 20 domini di base, informazione ulteriormente sintetizzata secondo le 4 tipologie di imprese (specializzate alimentari - ossia i domini da 1 a 5, specializzate non alimentari – domini da 6 a 10, non specializzate alimentari – domini da 11 a 15, non specializzate non alimentari – domini da 16 a 20) e le 5 classi di addetti (i domini 1, 6, 11 e 16 identificano la classe di addetti 1-2; 2, 7, 12 e 17 la classe 3-5; 3, 8, 13 e 18 la classe 6-9; 4, 9, 14 e 19 la classe 10-19; 5, 10, 15 e 20 la classe da 20 in poi). I criteri di *editing* posti a confronto sono i seguenti:

- il criterio 1 del paragrafo 4, basato sul ricorso a soglie prefissate, poste pari rispettivamente a 0,2 e 5; tale criterio è definito come "standard", perché è quello effettivamente utilizzato nell'ambito dell'indagine sulle vendite. In realtà esso è stato derivato sull'osservazione ripetuta della distribuzione empirica dei rapporti *r*, ed è quindi interpretabile anche come un caso particolare del criterio 2 del paragrafo 4, con P≈0,05. Si noti come il criterio standard operi direttamente sui rapporti *r* senza comportarne alcuna trasformazione ulteriore.
- Il metodo dei quartili definito dalle relazioni (2.3), (2.4) e (2.5), anch'esso applicato direttamente ai rapporti *r*; i parametri c_{inf} e c_{sup} sono stati posti uguali ad un'unica costante *c*, e si è posto *c*=1,2,3[29].
- Il metodo di Hidiroglou e Berthelot descritto nel paragrafo 2, sperimentato utilizzando i medesimi valori per *c*.

[29] Dato l'elevato numero di strati e il numero spesso esiguo di unità disponibili in ogni strato, i criteri 3 e 4 per la stima delle soglie di accettazione descritti nel paragrafo 4 non sono stati sperimentati.

- La procedura alternativa descritta nel paragrafo 2, utilizzando sempre i medesimi valori per *c*.

Dall'esame della tabella 7.1 emerge uno dei risultati salienti: il criterio standard identifica il numero più contenuto di *outlier* (25), concentrati nello strato più numeroso (17), che comprende le imprese specializzate non alimentari. La maggioranza degli altri criteri identifica un numero molto più elevato di rapporti anomali, e per qualunque scelta di *c* è il criterio di Hidiroglou e Berthelot a porsi al primo posto in tal senso, con l'ammontare più elevato di *outlier* in corrispondenza di *c*=1, nel quale caso la metà dei rapporti verrebbero considerati anomali. La sola procedura alternativa consente di identificare come anomali un numero ridotto di rapporti, nei casi in cui *c*=2 (77 rapporti) e, in particolare, *c*=3 (38).

Nel complesso, la quota relativa di rapporti identificati come anomali sul totale dei rapporti dello strato si mantiene piuttosto stabile al variare sia della tipologia di imprese, sia della classe di addetti; in particolare, dopo la procedura standard, la procedura alternativa identifica sempre il numero più contenuto di *outlier* per ogni scelta di *c* e/o dello strato considerato.

La tabella 7.2 riporta, sulla base di una stratificazione analoga a quella della tabella 7.1, i rapporti di variazione medi di strato ottenuti utilizzando i suddetti criteri di *editing* e trattando in 3 modi diversi le osservazioni identificate come anomale:

1. assegnando loro un peso nullo, ossia escludendole dall'analisi;
2. ristimando l'ammontare delle vendite di giugno 1999 moltiplicando l'ammontare delle vendite di giugno 1998 per il rapporto di variazione medio registrato nello strato calcolato sulle sole unità "buone" (formula (5.4) con x=1 e ε variabile casuale compresa tra –0,05 e +0,05);
3. ristimando l'ammontare delle vendite di giugno 1999 applicando anche in questo caso la formula (5.4), ma con x=addetti e ε variabile casuale compresa tra –0,05 e +0,05.

Inoltre sono riportati anche i risultati ottenibili senza il ricorso a nessuna procedura di *editing*, bensì ricorrendo ai criteri di riponderazione (6.2), (6.3) e (6.5). Nel complesso sono stati quindi confrontati 31 criteri di stima: il metodo standard, i 3 suddetti criteri di riponderazione e le procedure dei quartili sui rapporti *r*, di Hidiroglou e Berthelot e alternativa per *c*=1,2,3 e con le 3 suddette opzioni per l'imputazione dei valori anomali.

Per verificare a posteriori la qualità di tali stime – o quantomeno il loro livello medio di discordanza – dato che il rapporto di variazione delle vendite "vero" è comunque ignoto[30], si supporrà di approssimarlo con la media delle stime ottenute con i vari criteri posti a confronto,

per ciascuno degli strati considerati.

Considerando il totale delle imprese, solo la procedura basata sulla riponderazione espressa dalla (6.2) – ossia senza alcun ricondizionamento al vincolo di somma unitaria per i nuovi pesi – ha comportato una stima piuttosto diversa dalla stima media: lo scarto è risultato infatti pari a –0,12, ottenuto per differenza tra la stima della variazione ottenuta con tale procedura (pari a 0,88) e la stima ottenuta come media delle 31 procedure, pari a 1,01. Tutte le altre procedure hanno condotto a stime della variazione media diverse dalla stima media generale per 1 o 2 centesimi, ad eccezione del metodo di Hidiroglou e Berthelot per c=1 e c=2 (lo scarto è di 4 centesimi).

Più indicativo è il confronto tra le medie degli scarti dalla stima media riportati nell'ultima colonna della tabella 7.2, che consente di individuare le tipologie di imprese per le quali l'uso dell'una o dell'altra procedura ha condotto a risultati significativamente diversi. In effetti, la stratificazione per classi di addetti comporta stime mediamente più diverse tra loro rispetto alla stratificazione per attività prevalente delle imprese: nel primo caso si oscilla dalle eterogeneità più elevate relative alle classi di addetti estreme (lo scarto medio dalla stima media è pari a 0,07 per la classe di addetti 1-2 ed a 0,06 per la classe da 20 addetti in poi) a quelle più contenute delle classi 3-5 (0,02) e 10-19 (0,01); nel secondo caso le eterogeneità maggiori caratterizzano le imprese alimentari (0,05 sia per le specializzate che per le non specializzate) rispetto a quelle non alimentari (0,04 per le specializzate e 0,03 per le non specializzate). Va peraltro notato come le metodologie più anomale rispetto alle altre, e quindi sostanzialmente "rischiose", siano risultate la suddetta riponderazione (6.2) – per ogni strato considerato – ed il metodo di Hidiroglou e Berthelot qualora si ricorra all'imputazione corretta con gli addetti (si sono verificati scarti dalla stima media superiore ad un punto decimale per le imprese non specializzate alimentari se c=1 o c=2, per le imprese fino a 2 addetti se c=3 e per le imprese con almeno 20 addetti).

In sintesi, il risultato saliente è che la comparazione tra metodi di *editing* e di trattamento delle unità anomale sembra più sensibile rispetto alla dimensione aziendale che alla tipologia imprenditoriale, sebbene l'identificazione di strati particolarmente adatti per confrontare criteri diversi debba preferibilmente basarsi sull'incrocio tra i due caratteri. In effetti, gli strati più idonei in tal senso includono:

a) le imprese specializzate a prevalenza alimentare fino a 2 addetti (lo scarto medio dalla stima media della variazione è pari a 0,13);
b) le imprese specializzate a prevalenza non alimentare con addetti tra 6 e 9 (scarto pari a 0,11);

[30] L'universo di riferimento è composto da circa 560mila imprese, ed è quindi pressochè impossibile disporre di una

c) le imprese non specializzate a prevalenza alimentare con almeno 20 addetti (scarto pari a 0,15).

E' possibile una interpretazione economica di tale evidenza empirica: le unità di tipo a) rappresentano forse la tipologia di imprese al dettaglio maggiormente colpita dallo sviluppo della grande distribuzione e sono oggetto di un processo evolutivo piuttosto turbolento, che ne esalta le eterogeneità gestionali e può almeno in parte spiegare la minore convergenza dei metodi confrontati. Le unità di tipo b) costituiscono una delle poche tipologie di imprese specializzate che si sono ricollocate sul mercato secondo una impostazione manageriale che può risultare tanto tradizionale quanto moderna, e la cui eterogeneità è accentuata dalla vastità del comparto *non food*. Infine, le imprese di tipo c) identificano la componente più dinamica e numerosa della grande distribuzione (supermercati, ipermercati, discount, minimercati), e si caratterizzano per dimensioni aziendali e strategie di mercato molto differenziate.

Per quanto riguarda, infine, le soglie di accettazione per i rapporti *r*, nella tabella 7.3 è riportato un confronto tra il metodo dei quartili applicato direttamente a tali rapporti e la procedura alternativa. Va notato come la maggiore convergenza tra le 2 procedure si verifichi quando con riferimento alla procedura alternativa si considerano i rapporti *r* inferiori alla mediana: infatti, con riferimento ai rapporti *r* maggiori od uguali rispetto alla mediana il metodo dei quartili comporta soglie inferiori e superiori sensibilmente più basse rispetto alle corrispondenti soglie identificate dalla procedura alternativa, ed implica dunque – *coeteris paribus* – una minore tolleranza rispetto ai rapporti di variazione più elevati presenti nella base dati.

Tabella 7.1 – Numero di *outlier* identificati sulla base di diversi criteri e tre possibili scelte del parametro *c*

Strato	Metodo standard	Quartili su *r*			Hidiroglou-Berthelot			Alternativa			Totale unità
		c=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	
					Valori assoluti						
1	0	158	75	40	158	97	66	40	3	2	315
2	0	32	13	10	32	22	16	7	2	1	63

certezza assoluta circa la dinamica del fatturato commerciale per il complesso delle imprese attive.

3	0	11	4	2	12	6	4	2	1	1	22
4	0	16	10	6	16	6	3	2	1	1	31
5	0	7	1	0	8	4	2	1	0	0	16
6	17	794	402	214	794	530	392	205	34	21	1587
7	3	196	82	45	196	118	82	48	9	2	391
8	1	106	58	33	106	60	42	25	5	2	212
9	2	114	62	36	114	73	52	30	5	1	226
10	0	68	35	16	68	39	31	13	2	0	135
11	1	40	25	17	40	20	16	10	3	2	79
12	1	18	9	4	18	12	7	4	0	0	37
13	0	30	14	7	30	21	14	7	1	0	58
14	0	35	17	8	40	28	10	5	3	1	81
15	0	41	22	14	48	34	28	9	5	2	97
16	0	4	3	0	4	4	4	1	0	0	9
17	0	2	2	1	2	2	1	0	0	0	5
18	0	6	3	2	6	2	2	1	0	0	11
19	0	2	2	0	2	2	0	0	0	0	4
20	0	12	8	6	12	6	6	4	3	2	25
Da 1 a 5	0	224	103	58	226	135	91	52	7	5	447
Da 6 a 10	23	1278	639	344	1278	820	599	321	55	26	2551
Da 11 a 15	2	164	87	50	176	115	75	35	12	5	352
Da 16 a 20	0	26	18	9	26	16	13	6	3	2	54
1-6-11-16	18	996	505	271	996	651	478	256	40	25	1990
2-7-12-17	4	248	106	60	248	154	106	59	11	3	496
3-8-13-18	1	153	79	44	154	89	62	35	7	3	303
4-9-14-19	2	167	91	50	172	109	65	37	9	3	342
5-10-15-20	0	128	66	36	136	83	67	27	10	4	273
Totale	25	1692	847	461	1706	1086	778	414	77	38	3404
Valori percentuali											
1	0,0	50,2	23,8	12,7	50,2	30,8	21,0	12,7	1,0	0,6	100,0
2	0,0	50,8	20,6	15,9	50,8	34,9	25,4	11,1	3,2	1,6	100,0
3	0,0	50,0	18,2	9,1	54,5	27,3	18,2	9,1	4,5	4,5	100,0
4	0,0	51,6	32,3	19,4	51,6	19,4	9,7	6,5	3,2	3,2	100,0
5	0,0	43,8	6,3	0,0	50,0	25,0	12,5	6,3	0,0	0,0	100,0
6	1,1	50,0	25,3	13,5	50,0	33,4	24,7	12,9	2,1	1,3	100,0
7	0,8	50,1	21,0	11,5	50,1	30,2	21,0	12,3	2,3	0,5	100,0
8	0,5	50,0	27,4	15,6	50,0	28,3	19,8	11,8	2,4	0,9	100,0
9	0,9	50,4	27,4	15,9	50,4	32,3	23,0	13,3	2,2	0,4	100,0
10	0,0	50,4	25,9	11,9	50,4	28,9	23,0	9,6	1,5	0,0	100,0
11	1,3	50,6	31,6	21,5	50,6	25,3	20,3	12,7	3,8	2,5	100,0
12	2,7	48,6	24,3	10,8	48,6	32,4	18,9	10,8	0,0	0,0	100,0
13	0,0	51,7	24,1	12,1	51,7	36,2	24,1	12,1	1,7	0,0	100,0
14	0,0	43,2	21,0	9,9	49,4	34,6	12,3	6,2	3,7	1,2	100,0
15	0,0	42,3	22,7	14,4	49,5	35,1	28,9	9,3	5,2	2,1	100,0
16	0,0	44,4	33,3	0,0	44,4	44,4	44,4	11,1	0,0	0,0	100,0
17	0,0	40,0	40,0	20,0	40,0	40,0	20,0	0,0	0,0	0,0	100,0
18	0,0	54,5	27,3	18,2	54,5	18,2	18,2	9,1	0,0	0,0	100,0
19	0,0	50,0	50,0	0,0	50,0	50,0	0,0	0,0	0,0	0,0	100,0
20	0,0	48,0	32,0	24,0	48,0	24,0	24,0	16,0	12,0	8,0	100,0
1-5	0,0	50,1	23,0	13,0	50,6	30,2	20,4	11,6	1,6	1,1	100,0
6-10	0,9	50,1	25,0	13,5	50,1	32,1	23,5	12,6	2,2	1,0	100,0
11-15	0,6	46,6	24,7	14,2	50,0	32,7	21,3	9,9	3,4	1,4	100,0
16-20	0,0	48,1	33,3	16,7	48,1	29,6	24,1	11,1	5,6	3,7	100,0
1,6,11,16	0,9	50,1	25,4	13,6	50,1	32,7	24,0	12,9	2,0	1,3	100,0
2,7,12,17	0,8	50,0	21,4	12,1	50,0	31,0	21,4	11,9	2,2	0,6	100,0
3,8,13,18	0,3	50,5	26,1	14,5	50,8	29,4	20,5	11,6	2,3	1,0	100,0
4,9,14,19	0,6	48,8	26,6	14,6	50,3	31,9	19,0	10,8	2,6	0,9	100,0
5,10,15,20	0,0	46,9	24,2	13,2	49,8	30,4	24,5	9,9	3,7	1,5	100,0
Totale	0,7	49,7	24,9	13,5	50,1	31,9	22,9	12,2	2,3	1,1	100,0

Tabella 7.2 – Confronto tra la procedura standard, il metodo dei quartili su *r*, il metodo di Hidiroglou e Berthelot e la procedura alternativa

	PESO NULLO PER GLI *OUTLIER*										IMPUTAZIONE SEMPLICE						
		Quartili su *r*			Hidiroglou-Berthelot			Alternativa				Quartili su *r*			Hidiroglou-Berthelot		
Strato	Stan-dard	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	Stan-dard	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3
Variazioni medie																	
1	0,66	0,98	0,91	0,91	0,98	0,99	0,99	0,90	0,92	0,92	0,66	0,98	0,91	0,66	0,98	0,99	0,66
2	1,00	1,00	1,00	1,00	0,96	0,97	0,97	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,96	0,97	1,00
3	0,81	0,98	0,99	0,99	0,98	0,99	0,99	0,99	0,95	0,95	0,81	0,98	0,99	0,81	0,98	0,99	0,81

4	1,03	1,02	0,99	0,99	1,00	1,01	1,00	0,98	1,00	1,00	1,03	1,02	0,99	1,03	1,00	1,01	1,03
5	1,00	1,04	1,00	1,00	1,03	1,03	1,03	0,99	1,00	1,00	1,00	1,04	1,00	1,00	1,03	1,03	1,00
6	0,97	1,00	0,92	0,95	1,00	1,00	1,00	0,85	0,96	0,96	0,97	1,00	0,92	0,97	1,00	1,00	0,97
7	0,97	1,04	0,97	0,97	1,03	1,03	1,04	0,88	0,97	0,97	0,97	1,04	0,97	0,97	1,03	1,03	0,97
8	1,21	1,00	0,99	0,99	1,00	1,00	1,00	0,98	1,21	1,21	1,21	1,00	0,99	1,21	1,00	1,00	1,21
9	1,00	1,00	1,00	1,00	1,02	1,02	1,02	1,01	1,00	1,00	1,00	1,00	1,00	1,00	1,02	1,02	1,00
10	1,05	1,05	1,04	1,04	1,05	1,05	1,05	1,04	1,05	1,05	1,05	1,05	1,04	1,05	1,05	1,05	1,05
11	0,95	0,99	0,99	0,98	0,97	0,98	0,98	0,99	0,99	0,98	0,95	0,99	0,99	0,92	0,97	0,98	0,92
12	0,99	0,99	0,99	0,99	0,98	0,97	0,97	1,01	0,99	0,99	0,99	0,99	0,99	0,99	0,98	0,97	0,99
13	0,99	1,00	0,99	0,98	1,00	1,01	0,99	1,01	1,00	0,99	0,99	1,00	0,99	0,99	1,00	1,01	0,99
14	0,97	1,00	1,00	0,98	0,99	0,99	0,99	0,99	0,98	0,97	0,97	1,00	1,00	0,97	0,99	0,99	0,97
15	0,88	1,04	1,04	1,04	1,02	1,03	1,03	1,04	1,04	1,04	0,88	1,04	1,04	0,88	1,02	1,03	0,88
16	1,10	1,11	1,11	1,10	1,00	1,00	1,00	1,02	1,10	1,10	1,10	1,11	1,11	1,10	1,00	1,00	1,10
17	1,15	1,11	1,11	1,05	1,11	1,11	1,23	1,15	1,15	1,15	1,15	1,11	1,11	1,15	1,11	1,11	1,15
18	1,02	1,01	1,01	1,00	0,99	1,00	1,00	1,02	1,02	1,02	1,02	1,01	1,01	1,02	0,99	1,00	1,02
19	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14
20	1,14	1,05	0,93	0,94	1,05	1,11	1,11	0,94	0,94	1,00	1,14	1,05	0,93	1,14	1,05	1,11	1,14
1-5	0,90	1,00	0,98	0,98	0,99	1,00	0,99	0,97	0,97	0,97	0,90	1,00	0,98	0,90	0,99	1,00	0,90
6-10	1,04	1,02	0,98	0,99	1,02	1,02	1,02	0,95	1,04	1,04	1,04	1,02	0,98	1,04	1,02	1,02	1,04
11-15	0,95	1,00	1,00	0,99	0,99	0,99	0,99	1,01	1,00	0,99	0,95	1,00	1,00	0,95	0,99	0,99	0,95
16-20	1,11	1,08	1,06	1,05	1,06	1,07	1,10	1,05	1,07	1,08	1,11	1,08	1,06	1,11	1,06	1,07	1,11
1,6,11,16	0,92	1,02	0,98	0,99	0,99	0,99	0,99	0,94	0,99	0,99	0,92	1,02	0,98	0,91	0,99	0,99	0,91
2,7,12,17	1,03	1,04	1,02	1,00	1,02	1,02	1,05	1,01	1,03	1,03	1,03	1,04	1,02	1,03	1,02	1,02	1,03
3,8,13,18	1,01	1,00	0,99	0,99	1,00	1,00	1,00	1,00	1,04	1,04	1,01	1,00	0,99	1,01	1,00	1,00	1,01
4,9,14,19	1,03	1,04	1,03	1,03	1,04	1,04	1,04	1,03	1,03	1,03	1,03	1,04	1,03	1,03	1,04	1,04	1,03
5,10,15,20	1,02	1,05	1,00	1,00	1,04	1,05	1,06	1,00	1,01	1,02	1,02	1,05	1,00	1,02	1,04	1,05	1,02
Totale	1,00	1,03	1,00	1,00	1,02	1,02	1,03	1,00	1,02	1,02	1,00	1,03	1,00	1,00	1,02	1,02	1,00
								Scarti dalla stima media									
1	-0,18	0,14	0,06	0,07	0,14	0,14	0,15	0,06	0,07	0,07	-0,18	0,14	0,06	-0,18	0,14	0,14	-0,18
2	0,01	0,01	0,01	0,01	-0,03	-0,03	-0,03	0,01	0,01	0,01	0,01	0,01	0,01	0,01	-0,03	-0,03	0,01
3	-0,10	0,07	0,07	0,08	0,07	0,08	0,08	0,08	0,04	0,04	-0,10	0,07	0,07	-0,10	0,07	0,08	-0,10
4	0,02	0,01	-0,01	-0,01	0,00	0,00	0,00	-0,02	-0,01	-0,01	0,02	0,01	-0,01	0,02	0,00	0,00	0,02
5	-0,01	0,03	-0,01	-0,01	0,02	0,03	0,03	-0,01	-0,01	-0,01	-0,01	0,03	-0,01	-0,01	0,02	0,03	-0,01
6	0,01	0,04	-0,04	0,00	0,05	0,05	0,05	-0,11	0,01	0,01	0,01	0,04	-0,04	0,01	0,05	0,05	0,01
7	-0,01	0,06	-0,01	-0,01	0,05	0,05	0,06	-0,10	-0,01	-0,01	-0,01	0,06	-0,01	-0,01	0,05	0,05	-0,01
8	0,12	-0,09	-0,10	-0,10	-0,09	-0,09	-0,09	-0,11	0,12	0,12	0,12	-0,09	-0,10	0,12	-0,09	-0,09	0,12
9	0,00	0,00	-0,01	-0,01	0,01	0,01	0,02	0,01	0,00	0,00	0,00	0,00	-0,01	-0,01	0,01	0,01	-0,01
10	0,00	0,00	-0,01	-0,01	0,00	0,01	0,00	-0,01	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,01	0,00
11	0,02	0,06	0,06	0,06	0,04	0,05	0,05	0,07	0,06	0,05	0,02	0,06	0,06	-0,01	0,04	0,05	-0,01
12	0,00	0,00	0,00	0,00	-0,01	-0,01	-0,01	0,03	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00
13	-0,01	0,01	-0,01	-0,01	0,01	0,02	0,00	0,02	0,01	-0,01	-0,01	0,01	-0,01	-0,01	0,01	0,02	-0,01
14	-0,01	0,02	0,02	0,00	0,01	0,01	0,01	0,01	0,00	-0,01	-0,01	0,02	0,02	-0,01	0,01	0,01	-0,01
15	-0,15	0,01	0,01	0,01	-0,01	0,00	0,00	0,02	0,01	0,01	-0,15	0,01	0,01	-0,15	-0,01	0,00	-0,15
16	0,03	0,04	0,04	0,03	-0,06	-0,06	-0,06	-0,05	0,03	0,03	0,03	0,04	0,04	0,03	-0,06	-0,06	0,03
17	0,03	-0,02	-0,02	-0,07	-0,02	-0,02	0,11	0,03	0,03	0,03	0,03	-0,02	-0,02	0,03	-0,02	-0,02	0,03
18	0,01	0,00	0,00	-0,01	-0,02	-0,01	-0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,01	-0,02	-0,01	0,01
19	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
20	0,10	0,01	-0,11	-0,11	0,01	0,06	0,06	-0,11	-0,11	-0,05	0,10	0,01	-0,11	0,10	0,01	0,06	0,10
1-5	-0,05	0,05	0,02	0,03	0,04	0,05	0,04	0,02	0,02	0,02	-0,05	0,05	0,02	-0,05	0,04	0,05	-0,05
6-10	0,03	0,00	-0,03	-0,03	0,00	0,00	0,01	-0,06	0,02	0,02	0,03	0,00	-0,03	0,03	0,00	0,00	0,03
11-15	-0,03	0,02	0,02	0,01	0,01	0,01	0,01	0,03	0,02	0,01	-0,03	0,02	0,02	-0,04	0,01	0,01	-0,04
16-20	0,03	0,01	-0,02	-0,03	-0,02	0,00	0,02	-0,02	-0,01	0,00	0,03	0,01	-0,02	0,03	-0,02	0,00	0,03
1,6,11,16	-0,03	0,07	0,03	0,04	0,04	0,04	0,05	-0,01	0,04	0,04	-0,03	0,07	0,03	-0,04	0,04	0,04	-0,04
2,7,12,17	0,01	0,01	0,00	-0,02	0,00	0,00	0,03	-0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,00	0,01
3,8,13,18	0,01	0,00	-0,01	-0,01	-0,01	0,00	-0,01	0,00	0,04	0,04	0,01	0,00	-0,01	0,01	-0,01	0,00	0,01
4,9,14,19	0,00	0,01	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,01	0,01	0,00
5,10,15,20	-0,01	0,01	-0,03	-0,03	0,00	0,02	0,02	-0,03	-0,03	-0,01	-0,01	0,01	-0,03	-0,01	0,00	0,02	-0,01
Totale	-0,01	0,02	0,00	0,00	0,01	0,01	0,02	-0,01	0,01	0,02	-0,01	0,02	0,00	-0,01	0,01	0,01	-0,01

Tabella 7.2 (*segue*) – Confronto tra la procedura standard, il metodo dei quartili su *r*, il metodo di Hidiroglou e Berthelot e la procedura alternativa

	IMP.SEMPLICE			IMPUTAZIONE CORRETTA CON GLI ADDETTI												
	Alternativa				Quartili su *r*			Hidiroglou-Berthelot			Alternativa			Riponderazione			
Strato	*c*=1	*c*=2	*c*=3	Stan-dard	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	*c*=1	*c*=2	*c*=3	(6.2)	(6.3)	(6.5)	Media
								Variazioni medie									
1	0,90	0,92	0,66	0,66	0,98	0,91	0,66	0,98	0,99	0,66	0,90	0,92	0,66	0,46	0,70	0,81	0,84
2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,96	0,97	1,00	1,00	1,00	1,00	0,96	1,00	1,00	0,99
3	0,99	0,95	0,81	0,81	0,94	0,93	0,79	0,96	0,96	0,79	0,93	0,90	0,79	0,71	0,91	0,92	0,91
4	0,98	1,00	1,03	1,03	1,01	0,98	1,01	0,99	1,00	1,01	0,98	0,99	1,02	0,92	1,02	1,02	1,00
5	0,99	1,00	1,00	1,00	1,04	1,00	1,00	1,03	1,04	1,00	1,00	1,00	1,00	0,95	1,00	1,00	1,01

6	0,85	0,96	0,97	0,97	0,99	0,92	0,97	0,99	0,99	0,96	0,84	0,96	0,97	0,83	0,97	0,99	0,96
7	0,88	0,97	0,97	0,97	1,04	0,97	0,97	1,03	1,03	0,97	0,88	0,97	0,97	0,88	0,98	1,00	0,98
8	0,98	1,21	1,21	1,21	1,00	0,99	1,20	1,00	1,00	1,21	0,98	1,21	1,21	0,95	1,20	1,16	1,09
9	1,01	1,00	1,00	1,00	1,00	1,00	1,00	1,02	1,02	1,00	1,01	1,00	1,00	0,97	1,00	1,00	1,00
10	1,04	1,05	1,05	1,05	1,06	1,04	1,05	1,06	1,06	1,06	1,04	1,05	1,05	0,96	1,05	1,05	1,05
11	0,99	0,99	0,92	0,95	0,95	0,95	0,89	0,65	0,65	0,63	0,95	0,94	0,91	0,84	0,97	0,96	0,93
12	1,01	0,99	0,99	0,99	0,99	0,99	0,99	0,98	0,97	0,99	1,01	0,99	0,99	0,98	0,99	0,99	0,99
13	1,01	1,00	0,99	0,99	1,00	0,99	1,00	1,01	1,01	0,99	1,00	1,00	0,99	0,90	0,99	1,00	0,99
14	0,99	0,98	0,97	0,97	1,00	1,00	0,97	0,99	0,99	0,97	0,99	0,98	0,97	0,90	0,97	0,97	0,98
15	1,04	1,04	0,88	0,88	0,87	0,87	0,79	1,89	1,90	1,61	0,87	0,86	0,80	0,75	0,96	0,98	1,03
16	1,02	1,10	1,10	1,10	1,11	1,11	1,10	1,00	1,00	1,10	1,02	1,10	1,10	0,91	1,07	1,08	1,06
17	1,15	1,15	1,15	1,15	1,11	1,11	1,15	1,11	1,11	1,15	1,15	1,15	1,15	0,93	1,07	1,13	1,13
18	1,02	1,02	1,02	1,02	1,01	1,01	1,02	0,99	1,00	1,02	1,02	1,02	1,02	0,95	1,01	1,01	1,01
19	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,14	1,06	1,15	1,15	1,14
20	0,94	0,94	1,14	1,14	1,08	0,95	1,18	1,07	1,13	1,16	0,95	0,96	1,14	0,88	1,07	1,07	1,05
1-5	0,97	0,97	0,90	0,90	0,99	0,96	0,89	0,98	0,99	0,89	0,96	0,96	0,90	0,80	0,93	0,95	0,95
6-10	0,95	1,04	1,04	1,04	1,02	0,98	1,04	1,02	1,02	1,04	0,95	1,04	1,04	0,92	1,04	1,04	1,02
11-15	1,01	1,00	0,95	0,95	0,96	0,96	0,93	1,10	1,10	1,04	0,96	0,95	0,93	0,87	0,98	0,98	0,98
16-20	1,05	1,07	1,11	1,11	1,09	1,06	1,12	1,06	1,08	1,11	1,06	1,07	1,11	0,95	1,07	1,09	1,08
1,6,11,16	0,94	0,99	0,91	0,92	1,01	0,97	0,90	0,91	0,91	0,84	0,93	0,98	0,91	0,76	0,93	0,96	0,95
2,7,12,17	1,01	1,03	1,03	1,03	1,04	1,02	1,03	1,02	1,02	1,03	1,01	1,03	1,03	0,93	1,01	1,03	1,02
3,8,13,18	1,00	1,04	1,01	1,01	0,99	0,98	1,00	0,99	0,99	1,00	0,98	1,03	1,00	0,88	1,03	1,02	1,00
4,9,14,19	1,03	1,03	1,03	1,03	1,04	1,03	1,03	1,04	1,04	1,03	1,03	1,03	1,03	0,96	1,04	1,04	1,03
5,10,15,20	1,00	1,01	1,02	1,02	1,01	0,96	1,00	1,26	1,28	1,21	0,97	0,97	1,00	0,89	1,02	1,03	1,03
Totale	1,00	1,02	1,00	1,00	1,02	0,99	0,99	1,04	1,05	1,02	0,98	1,01	0,99	0,88	1,00	1,01	1,01
								Scarti dalla stima media									
1	0,06	0,07	-0,18	-0,18	0,14	0,06	-0,18	0,14	0,14	-0,18	0,06	0,07	-0,18	-0,38	-0,14	-0,03	0,13
2	0,01	0,01	0,01	0,01	0,01	0,01	0,01	-0,03	-0,03	0,01	0,01	0,01	0,01	-0,04	0,01	0,01	0,01
3	0,08	0,04	-0,10	-0,10	0,03	0,02	-0,12	0,05	0,05	-0,13	0,02	-0,01	-0,12	-0,20	0,00	0,01	0,07
4	-0,02	-0,01	0,02	0,02	0,01	-0,03	0,01	-0,01	-0,01	0,01	-0,03	-0,01	0,02	-0,08	0,02	0,02	0,02
5	-0,01	-0,01	-0,01	-0,01	0,03	-0,01	-0,01	0,02	0,03	-0,01	-0,01	-0,01	-0,01	-0,06	-0,01	-0,01	0,02
6	-0,11	0,01	0,01	0,01	0,04	-0,04	0,01	0,03	0,03	0,01	-0,12	0,01	0,01	-0,13	0,01	0,03	0,04
7	-0,10	-0,01	-0,01	-0,01	0,06	-0,01	-0,01	0,05	0,05	-0,01	-0,10	-0,01	-0,01	-0,10	0,00	0,02	0,03
8	-0,11	0,12	0,12	0,12	-0,09	-0,10	0,12	-0,09	-0,09	0,12	-0,11	0,12	0,12	-0,14	0,11	0,08	0,11
9	0,01	0,00	-0,01	0,00	0,00	-0,01	-0,01	0,01	0,01	-0,01	0,01	0,00	-0,01	-0,04	0,00	0,00	0,01
10	-0,01	0,00	0,00	0,00	0,01	-0,01	0,01	0,02	0,01	0,01	0,00	0,00	0,00	-0,09	0,00	0,01	0,01
11	0,07	0,06	-0,01	0,02	0,02	0,02	-0,04	-0,28	-0,28	-0,30	0,02	0,02	-0,02	-0,08	0,04	0,04	0,06
12	0,03	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,03	0,00	0,00	-0,01	0,00	0,00	0,01
13	0,02	0,01	-0,01	-0,01	0,01	-0,01	0,00	0,01	0,01	-0,01	0,01	0,01	-0,01	-0,10	0,00	0,00	0,01
14	0,01	0,00	-0,01	-0,01	0,02	0,02	-0,01	0,01	0,01	-0,01	0,01	0,00	-0,01	-0,08	-0,01	-0,01	0,01
15	0,02	0,01	-0,15	-0,15	-0,16	-0,16	-0,24	0,86	0,87	0,58	-0,16	-0,17	-0,23	-0,28	-0,07	-0,05	0,15
16	-0,05	0,03	0,03	0,03	0,04	0,04	0,03	-0,06	-0,06	0,03	-0,05	0,03	0,03	-0,15	0,00	0,01	0,04
17	0,03	0,03	0,03	0,03	-0,02	-0,02	0,03	-0,02	-0,02	0,03	0,03	0,03	0,03	-0,20	-0,06	0,01	0,03
18	0,01	0,01	0,01	0,01	0,00	0,00	0,01	-0,02	-0,01	0,01	0,01	0,01	0,01	-0,07	0,00	0,00	0,01
19	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,08	0,01	0,00	0,01
20	-0,11	-0,11	0,10	0,10	0,03	-0,10	0,13	0,03	0,09	0,12	-0,09	-0,09	0,00	-0,17	0,03	0,03	0,08
1-5	0,02	0,02	-0,05	-0,05	0,04	0,01	-0,06	0,03	0,04	-0,06	0,01	0,01	-0,06	-0,15	-0,02	0,00	0,05
6-10	-0,06	0,02	0,03	0,03	0,00	-0,03	0,02	0,00	0,00	0,02	-0,07	0,02	0,03	-0,10	0,03	0,03	0,04
11-15	0,03	0,02	-0,04	-0,03	-0,02	-0,03	-0,06	0,12	0,12	0,05	-0,02	-0,03	-0,05	-0,11	-0,01	0,00	0,05
16-20	-0,02	-0,01	0,03	0,03	0,01	-0,01	0,04	-0,01	0,00	0,04	-0,02	0,00	0,01	-0,13	0,00	0,01	0,03
1,6,11,16	-0,01	0,04	-0,04	-0,03	0,06	0,02	-0,04	-0,04	-0,04	-0,11	-0,02	0,03	-0,04	-0,18	-0,02	0,01	0,07
2,7,12,17	-0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,00	0,01	-0,01	0,01	0,01	-0,09	-0,01	0,01	0,02
3,8,13,18	0,00	0,04	0,01	0,01	-0,01	-0,02	0,00	-0,01	-0,01	0,00	-0,02	0,03	0,00	-0,12	0,03	0,02	0,05
4,9,14,19	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,07	0,00	0,00	0,01
5,10,15,20	-0,03	-0,03	-0,01	-0,01	-0,02	-0,07	-0,03	0,23	0,25	0,18	-0,07	-0,07	-0,06	-0,15	-0,01	0,00	0,06
Totale	-0,01	0,01	-0,01	-0,01	0,01	-0,02	-0,01	0,04	0,04	0,01	-0,02	0,00	-0,02	-0,12	0,00	0,01	0,04

Tabella 7.3 – Soglie di accettazione inferiore (A_{inf}) e superiore (A_{sup}). Confronto tra il metodo dei quartili applicato ai rapporti *r* e la procedura alternativa[31]

	Quartili su *r*						Procedura alternativa per $r \geq$ mediana						Procedura alternativa per $r <$ mediana					
	c=1		*c*=2		*c*=3		*c*=1		*c*=2		*c*=3		*c*=1		*c*=2		*c*=3	
Strato	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}	A_{inf}	A_{sup}

[31] La distinzione tra i casi in cui *r* sia maggiore od inferiore rispetto alla mediana deriva dalla formula (3.1), da cui sono facilmente derivabili gli intervalli di accettazione per *r*. Si riportano anche gli estremi superiori relativi al livello, ossia al valore delle vendite per impresa espresso in milioni di lire, rispettivamente per gli strati da 1 a 20 e per c=1,2,3. **1**:27,37,46; **2**:94,133,173; **3**:215,285,356; **4**:361,462,563; **5**:924,1283,1642; **6**:24,36,47; **7**:110,158,205; **8**:235,299,363; **9**:474,629,783; **10**:1994,2979,3964; **11**:39,57,74; **12**:125,166,207; **13**:306,375,444; **14**:573,677,781; **15**:3348,5088,6828; **16**:21,28,35; **17**:96,104,112; **18**:211,243,275; **19**:557,649,740; **20**:2251,2867,3484.

1	0,90	1,08	0,81	1,19	0,73	1,30	1,15	1,33	1,09	1,44	1,02	1,55	0,92	1,06	0,85	1,12	0,79	1,19
2	0,89	1,04	0,82	1,12	0,75	1,20	1,11	1,25	1,05	1,32	1,00	1,39	0,94	1,05	0,89	1,11	0,84	1,18
3	0,93	1,05	0,88	1,12	0,83	1,19	1,09	1,21	1,03	1,26	0,98	1,32	0,95	1,05	0,91	1,11	0,87	1,18
4	0,94	1,07	0,86	1,12	0,78	1,17	1,12	1,29	1,06	1,41	1,00	1,52	0,91	1,05	0,84	1,11	0,77	1,18
5	0,97	1,12	0,92	1,21	0,87	1,31	1,09	1,21	1,04	1,27	0,98	1,32	0,95	1,05	0,91	1,11	0,87	1,18
6	0,85	1,20	0,69	1,41	0,54	1,61	1,29	1,71	1,15	1,99	1,01	2,27	0,84	1,11	0,72	1,24	0,63	1,41
7	0,87	1,20	0,73	1,37	0,58	1,55	1,26	1,55	1,14	1,73	1,03	1,91	0,89	1,09	0,79	1,20	0,72	1,34
8	0,89	1,10	0,78	1,20	0,68	1,31	1,17	1,38	1,10	1,53	1,04	1,68	0,89	1,06	0,81	1,12	0,74	1,19
9	0,90	1,12	0,78	1,23	0,66	1,34	1,16	1,43	1,07	1,60	0,97	1,78	0,88	1,08	0,78	1,18	0,71	1,29
10	0,97	1,13	0,88	1,21	0,80	1,28	1,10	1,28	1,04	1,39	0,99	1,51	0,91	1,05	0,83	1,11	0,77	1,18
11	0,91	1,03	0,85	1,10	0,79	1,16	1,09	1,28	1,03	1,41	0,97	1,54	0,90	1,05	0,81	1,11	0,74	1,18
12	0,92	1,04	0,87	1,10	0,82	1,16	1,07	1,19	1,02	1,24	0,96	1,30	0,95	1,05	0,91	1,11	0,87	1,18
13	0,92	1,05	0,84	1,10	0,76	1,16	1,08	1,23	1,02	1,33	0,96	1,43	0,92	1,05	0,85	1,11	0,79	1,18
14	0,94	1,04	0,88	1,09	0,82	1,14	1,03	1,14	0,98	1,20	0,93	1,25	0,95	1,05	0,91	1,11	0,87	1,18
15	0,96	1,09	0,91	1,16	0,86	1,23	1,06	1,18	1,00	1,25	0,95	1,31	0,94	1,05	0,90	1,11	0,85	1,18
16	0,94	1,26	0,89	1,52	0,83	1,78	1,16	1,44	1,01	1,57	0,86	1,70	0,91	1,13	0,84	1,30	0,77	1,53
17	0,93	1,26	0,88	1,54	0,83	1,82	1,13	1,44	1,01	1,63	0,89	1,82	0,87	1,10	0,76	1,23	0,68	1,40
18	0,95	1,08	0,90	1,15	0,85	1,23	1,08	1,22	1,02	1,30	0,96	1,39	0,93	1,05	0,87	1,11	0,81	1,18
19	1,05	1,19	0,97	1,25	0,89	1,31	1,08	1,22	1,02	1,30	0,97	1,38	0,93	1,05	0,88	1,11	0,83	1,18
20	0,93	1,13	0,82	1,22	0,72	1,31	1,15	1,38	1,05	1,52	0,95	1,66	0,90	1,09	0,82	1,19	0,75	1,31
1-5	0,92	1,07	0,86	1,15	0,79	1,23	1,11	1,26	1,05	1,34	0,99	1,42	0,93	1,05	0,88	1,11	0,83	1,18
6-10	0,89	1,15	0,77	1,28	0,65	1,42	1,20	1,47	1,10	1,65	1,01	1,83	0,88	1,08	0,79	1,17	0,71	1,28
11-15	0,93	1,05	0,87	1,11	0,81	1,17	1,07	1,20	1,01	1,29	0,95	1,37	0,93	1,05	0,87	1,11	0,82	1,18
16-20	0,96	1,18	0,89	1,34	0,82	1,49	1,12	1,34	1,02	1,46	0,93	1,59	0,91	1,09	0,83	1,19	0,77	1,32
1,6,11,16	0,90	1,14	0,81	1,30	0,72	1,46	1,17	1,44	1,07	1,60	0,97	1,76	0,89	1,09	0,80	1,19	0,73	1,33
2,7,12,17	0,90	1,13	0,82	1,28	0,74	1,43	1,14	1,35	1,06	1,48	0,97	1,61	0,91	1,08	0,84	1,16	0,78	1,27
3,8,13,18	0,92	1,07	0,85	1,15	0,78	1,22	1,10	1,26	1,04	1,36	0,98	1,45	0,92	1,05	0,86	1,11	0,80	1,18
4,9,14,19	0,96	1,11	0,87	1,17	0,79	1,24	1,10	1,27	1,03	1,38	0,97	1,48	0,92	1,06	0,85	1,13	0,79	1,20
5,10,15,20	0,96	1,11	0,88	1,20	0,81	1,28	1,10	1,26	1,03	1,36	0,96	1,45	0,93	1,06	0,86	1,13	0,81	1,21
Totale	0,93	1,11	0,85	1,22	0,77	1,33	1,12	1,32	1,05	1,43	0,97	1,55	0,91	1,07	0,84	1,15	0,78	1,24

APPENDICE

Distribuzione del rapporto tra due variabili aleatorie esponenziali negative

Il risultato vale, con maggiore generalità, anche nel caso di variabili aleatorie di tipo Gamma. Come noto, l'espressione generale della funzione di densità di una variabile aleatoria x di tipo Gamma, con parametri α e β è data da:

$$GA(\alpha,\beta)=\frac{\alpha^{\beta}}{\Gamma(\beta)}x^{\beta-1}\exp(-\alpha x).$$

Se si hanno due variabili aleatorie X_1 e X_2 distribuite secondo le due distribuzioni Gamma $GA(\alpha_1,\beta_1)$ e $GA(\alpha_2,\beta_2)$, supponendo l'indipendenza tra tali distribuzioni si può dimostrare (Stuart e Ord, 1992) che la nuova variabile $\eta=\left(\frac{\alpha_1\beta_2}{\alpha_2\beta_1}\right)\frac{X_1}{X_2}=\left(\frac{\alpha_1\beta_2}{\alpha_2\beta_1}\right)I_{12}$ si distribuisce come una variabile *F* di Fisher con β_1 e β_2 gradi di libertà. Se dunque si desidera individuare il percentile $I_{12,\mathrm{P}}$ relativo alla variabile I_{12} tale che l'area sottesa dalla coda di destra sia pari a P, basta imporre la condizione:

$$(1-\mathrm{P})=PROB(\eta\le F_{\mathrm{sup,P}})$$

dove $F_{\mathrm{sup,P}}$ è il percentile della distribuzione *F* di Fisher con β_1 e β_2 gradi di libertà tali che l'area alla destra dello stesso sia pari a P, e dalla definizione di η segue facilmente che:

$$I_{12,\mathrm{sup,P}}=\left(\frac{\alpha_2\beta_1}{\alpha_1\beta_2}\right)F_{\mathrm{sup,P}}$$

Ricordando poi che la distribuzione di una variabile aleatoria di tipo esponenziale negativa di parametro α equivale alla distribuzione di una variabile Gamma di parametri α e 1, ossia $EN(\alpha)=GA(\alpha,1)$, in generale si avrà che:

$$I_{12,\mathrm{P}}=\left(\frac{\alpha_2}{\alpha_1}\right)F_{\mathrm{P}}=\left(\frac{\frac{1}{\alpha_1}}{\frac{1}{\alpha_2}}\right)F_{\mathrm{P}}=\frac{E(X_1)}{E(X_2)}F_{\mathrm{P}},$$

dove il simbolo *E*(*X*) indica il valore atteso di *X* e la funzione *F* ha entrambi i gradi di libertà pari a 1. Se poi la variabile I_{12} è moltiplicata per la costante *a*, segue ovviamente che: $a\,I_{12,\mathrm{P}}=a\frac{E(X_1)}{E(X_2)}F_{\mathrm{P}}$. Ponendo prima $a=q_{0,5}$ e $I_{12}=1/r_{\mathrm{i}}$, e poi $a=1/q_{0,5}$ e $I_{12}=r_{\mathrm{i}}$ si perviene,

infine, alle relazioni (4.1) e (4.2).

Distribuzione del quadrato di una variabile aleatoria esponenziale negativa

Sulla base di quanto appena ricordato, l'espressione generale della funzione di densità di una variabile aleatoria *x* di tipo Esponenziale negativa, con parametro α, è data da:

$$f_x(x) = EN(\alpha) = \alpha \exp(-\alpha x).$$

Se si effettua la trasformazione $y = g(x) = x^2$, da cui consegue $x = g^{-1}(y) = y^{0,5}$, la funzione di densità della nuova variabile *y* sarà ricavabile tramite la formula:

$$f_y(y) = \frac{f_x[g^{-1}(y)]}{\left|g'[g^{-1}(y)]\right|}.$$

Consegue che $f_y(y) = \frac{\alpha}{2\,y^{0,5}} \exp(-\alpha\, y^{0,5}) = \frac{\alpha}{2}\, y^{-0,5} \exp(-\alpha\, y^{0,5}) = \frac{0,5}{\theta^{0,5}}\, y^{-0,5} \exp\left[-\left(\frac{y}{\theta}\right)^{0,5}\right]$, dove nell'ultimo passaggio si è posto $\alpha = \theta^{-0,5}$. Poiché l'espressione generale della funzione di densità di una variabile aleatoria di Weibull di parametri θ e β è data da:

$$WE(\theta, \beta) = \frac{\beta}{\theta^{\beta}}\, y^{\beta-1} \exp\left[-\left(\frac{y}{\theta}\right)^{\beta}\right],$$

consegue che $f_y(y) = WE(\theta;0,5) = WE(\alpha^2;0,5)$, dove α è il reciproco del valore atteso di *x*.

Bibliografia

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

BARNETT V. (1978), *Outliers in Statistical Data*, 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 Biemer, Groves, Lyberg, Mathiowetz, Sudman (eds) "Measurement Errors in Surveys", John Wiley & Sons, New York, pp.211-236.

FELLEGI I.P., HOLT D. (1976), *A Sistematic Approach to Authomatic Editing and Imputation*, "Journal of the American Statistical Association", 71, pp.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, pp.52-68.

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

GISMONDI R. (1998), *Metodi per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di variazioni*, documento preparato per la "Commissione campioni", Istat, Roma.

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

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

HAWORTH M.F. (1996), *Re-engineering Data Production and Measuring Quality in the UK Retail Prices Index*, paper presented at the "Annual Research Conference and Technology Interchange", March 1999, Arlington, USA.

HENNIG C. (1998), *Clustering and Outlier Identification: Fixed Point Cluster Analysis*, in Rizzi, Vichi, Bock (eds) "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, Statistics Canada, Ottawa, pp.73-84.

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

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

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

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

LEE H. (1995), *Outliers in Business Surveys*, in Cox, Binder, Chinnappa, Christianson, Colledge, Kott (eds), "Business Survey Methods", John Wiley & Sons, New York, pp.503-523.

PIZZI C., PELLIZZARI P. (1998), *Detecting Outliers in Time Series*, in Rizzi,Vichi, Bock (eds.) "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, pp.43-53.

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

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

STUART A., ORD K. (1992), *Advanced Theory of Statistics*, vol.I, Edward Arnold, London.

THOMPSON K.J. (1998), *Generalised SAS Ratio Edit Parameter Program*, internal document, Boureau of the Census, Washington, USA.

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

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

RIASSUNTO

Nell'ambito di una indagine campionaria finalizzata alla stima di una variazione di una variabile quantitativa, la presenza di osservazioni particolarmente anomale (*outlier*) può comportare significative distorsioni nel processo di stima. I problemi che conseguentemente si pongono sono i seguenti: a) come identificare le unità anomale; b) come trattarle da un punto di vista statistico. In questo lavoro si è cercato di valutare criticamente il criterio di identificazione degli *outlier* proposto originariamente da Hidiroglou e Berthelot, proponendone una versione alternativa, che in genere consente di ridurre sensibilmente il numero di unità identificate come anomale e,

quindi, il numero di interventi sui microdati. Sono poi stati proposti e confrontati alcuni criteri per l'identificazione delle soglie di accettazione e per ridurre il peso delle osservazioni *outlier* nel processo di stima. E' stata infine illustrata una applicazione ad un caso concreto, in cui sono state poste a confronto 31 modalità di trattamento statistico degli *outlier*.

ABSTRACT

Outliers' Detection and Treatment when Estimating Change in Longitudinal Surveys

If the main purpose of a sampling survey consists in the estimation of a change concerning a certain quantitative variable, the presence of outliers could lead to significant biases in the estimation process. As a consequence, problems occurring are: a) how identifying outliers; b) how treating them from a statistical point of view. In this paper we tried to evaluate the technique for identifying outliers originaly proposed by Hidiroglou and Berthelot, presenting a new technique as well, that generally leads both to a significantly lower amount of units identified as outliers and of microdata alterations. Moreover, we proposed and compared some criteria for defining the tolerance intervals and for reducing the weight of outliers in the estimation process. Finally, we carried out an empirical study, in which we compared 31 different ways for tretaing outliers.