UN CONFRONTO TRA METODI DI IDENTIFICAZIONE DI OSSERVAZIONI OUTLIER IN INDAGINI LONGITUDINALI FINALIZZATE ALLA STIMA DI UNA VARIAZIONE: PROPOSTE TEORICHE E VERIFICHE EMPIRICHE

Roberto Gismondi

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

---

[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).

situato alla coda di una distribuzione empirica di valori relativi ad unità ad essa teoricamente simili. 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, euro invece di migliaia di euro, per cui i valori dichiarati dovrebbero essere divisi per mille). La loro non rappresentatività va intesa con riferimento alle unità della popolazione non incluse nel campione, perchè non contribuiscono alla variabilità campionaria fornendo informazioni su di esse;
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 assegnare loro 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);

---

[4] In proposito si veda Weir (1997).

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

– proporre una applicazione ad un caso concreto particolarmente idoneo per il confronto tra le varie procedure di identificazione e trattamento degli *outlier* basate sulle soglie di accettazione (paragrafo 5).

Per una trattazione teorica più ampia circa le procedure di riponderazione di cui al precedente punto *c*) si rimanda a Rizzo, Kalton e Brick (1996), mentre per un confronto empirico completo tra le procedure *a*), *b*) e *c*) si può fare riferimento a Gismondi (2000a).

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 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:

$$d_{\inf} = q_{0,50} - q_{0,25} \tag{1}$$

$$d_{\sup} = q_{0,75} - q_{0,50} \tag{2}$$

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

$$A = (q_{0,50} - c_{\inf} d_{\inf} ; q_{0,50} + c_{\sup} d_{\sup}) = (A_{\inf} ; A_{\sup}) \tag{3}$$

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].

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:

$$d_{\inf} = \max\left(q_{0,50} - q_{0,25}, \left|B q_{0,50}\right|\right) \tag{4}$$

$$d_{\sup} = \max\left(q_{0,75} - q_{0,50}, \left|B q_{0,50}\right|\right) \tag{5}$$

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

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 5.

In particolare, la metodologia basata sull'intervallo di accettazione definito dalle relazioni (3), (4) e (5), definibile come *metodo dei quartili*, può essere applicata direttamente alla variabile z data dal rapporto:

$$r_{ti} = \frac{y_{ti}}{y_{t-k,i}}. \tag{6}$$

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

$$s_{ti} = \begin{cases} 1 - \dfrac{q_{0,50}}{r_{ti}} & \text{se} \quad 0 < r_{ti} < q_{0,50} \\ \dfrac{r_{ti}}{q_{0,50}} - 1 & \text{se} \quad r_{ti} \geq q_{0,50} \end{cases} \tag{7}$$

e, successivamente, l'applicazione del metodo dei quartili alla variabile z data dalla funzione effetto dell'unità i-esima:

$$E_{ti} = s_{ti}[\max(y_{ti}; y_{t-k,i})]^V \tag{8}$$

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[7].

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[8];

– si basa su una funzione effetto che, essendo data dal prodotto di due componenti (livello e variazione), può assumere valori difficilmente interpretabili; in altri

[7] Una procedura per determinate il valore ottimale da attribuire a V è descritta in Davila (1992).
[8] Come anche ricordato dallo stesso Lee (*op. cit.*, p.509).

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 (1), a meno della costante pari a 1, per cui si avrà:

$$s^*_{ti} = \begin{cases} \dfrac{q_{0,50}}{r_{ti}} & \text{se} \quad 0 < r_{ti} < q_{0,50} \\[2ex] \dfrac{r_{ti}}{q_{0,50}} & \text{se} \quad r_{ti} \geq q_{0,50} \end{cases} \tag{9}$$

e tale espressione risulterà sempre non inferiore a 1[9].

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

$$\begin{cases} E_{tri} = s^{U}_{ti*} \\ E_{tli} = l^{V}_{ti} = [\max(y_{ti}; y_{t-k,i})]^{V} \end{cases} \tag{10}$$

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[10] 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 da s^*_{ti} alto) potrebbe comunque non essere soggetta ad alterazioni nell'ambito della procedura di stima, qualora il suo livello

[9] Ovviamente l'eliminazione dell'addendo unitario che compare nella (7) semplifica l'interpretazione della funzione, non altera la forma della sua distribuzione, comportandone solo una traslazione.

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

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:

$$A_{\sup} = q_{0,50} + c_{\sup} d_{\sup} \tag{11}$$

$$d_{\sup} = \max\left(q_{0,75} - q_{0,50}, \left|B q_{0,50}\right|\right). \tag{12}$$

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

$$s_{ti}^{*} = r_{ti}^{\alpha} q_{0,50}^{\beta} \tag{13}$$

da cui deriva la relazione (9) 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 (13) 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[11] per le osservazioni di Y, è possibile analizzare l'andamento della funzione (10) – 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[12].

TABELLA 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}	Distribuzioni teoriche di riferimento 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

Come già evidenziato, le precedenti relazioni sono utili ai fini dell'individuazione delle soglie di accettazione, così come indicato nel paragrafo seguente[13].

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

[12] 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.

[13] 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.

La tabella 2 seguente riporta un esempio da cui risulta evidente come il ricorso alla procedura alternativa definita dalle relazioni (9) e (10) 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 particolarmente elevato (colonna "Out_3").

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 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). È proprio tale caratteristica sostanzialmente indesiderata del metodo a suggerire quantomeno il confronto con tecniche alternative.

TABELLA 2

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.

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 (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[14];

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} ed E_{tli}, la soglia di accettazione $A_{\text{sup}} = (q_{0,50} - c_{\text{sup}} d_{\text{sup}})$, da cui si ricava immediatamente c_{sup}.

Criterio 2: *ricorso alla distribuzione empirica*

Sulla base di tale criterio, riguardo alle funzioni E_{tri} ed 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.

---

[14] Ricade in questa famiglia di procedure il caso in cui si prefissino delle soglie invarianti al variare del tempo *t*.

Criterio 3: *ricorso alla distribuzione teorica*

Si può dimostrare[15] che se U=1, l'estremo superiore dell'intervallo di accettazione per la variabile E_{tri} sarà definito dalle relazioni:

$$q_{0,50}\frac{\mathrm{E}(y_{t\text{-}k})}{\mathrm{E}(y_t)}F_{\sup,p} \qquad \text{se} \qquad 0 < r_{ti} < q_{0,50} \tag{14}$$

$$\frac{\mathrm{E}(y_t)}{q_{0,50}\mathrm{E}(y_{t\text{-}k})}F_{\sup,p} \qquad \text{se} \qquad r_{ti} \geq q_{0,50} \tag{15}$$

dove il simbolo E(·) 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:

$$PERC_{\sup,p} \tag{16}$$

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[16].

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 5 – il rapporto r tra i fatturati relativi allo stesso mese di due anni consecutivi mostra una certa concentrazione[17] 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 (3) si può porre la seguente identità:

---

[15] Si rimanda al primo sottoparagrafo dell'appendice.

[16] Si rimanda al secondo sottoparagrafo dell'appendice.

[17] Nell'intervallo citato cade infatti il 90% dei casi analizzati nel paragrafo 5. Se si fosse preso in considerazione il 98% dei casi, si sarebbe ricavato il valore 2,17 per il percentile della funzione F.

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

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

$$c_{\sup} = \frac{q_{0,50}}{d_{\sup}} \left[\frac{\mathrm{E}(y_t) F_{\sup,p/2} - \mathrm{E}(y_{t-k})}{\mathrm{E}(y_{t-k})} \right]. \tag{17}$$

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

$$q_{0,50} + c_{\sup} d_{\sup} = PERC_{\sup,p} \quad \text{da cui si ottiene:} \quad c_{\sup} = \frac{PERC_{\sup,p} - q_{0,50}}{d_{\sup}}. \tag{18}$$

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, relative ad 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 $(t_0\text{-}k)$ 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 $(t\text{-}k)$ e misurata sulla unità i-esima di un sottoinsieme S, l'intervallo di accettazione A – dove A è definito dalle condizioni (3) e (4) – potrà essere determinato imponendo la seguente condizione di minimo:

$$\left| \sum_{r_{z,t_0 i} \in A} (r_{z,t_0 i}) D_{zi} - R_{z,t_0} \right| = \text{minimo} \tag{19}$$

dove i termini D_{zi} indicano degli opportuni pesi campionari.

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 $(t_0\text{-}k)$ di Z consentirà di ricavare anche l'intervallo di accettazione per gli omologhi rapporti di variazione[18] r_{ti} relativi a Y nei casi in cui:

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

– le due variabili possano ritenersi molto correlate (ad esempio, Y e Z possono riferirsi al fatturato delle imprese commerciali di due provincie limitrofe di una data regione);

– le due variabili siano caratterizzate da ordini di grandezza simili (ad esempio, valore aggiunto e fatturato).

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 6 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 6 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 6 mesi successivi a t_0 sia il più possibile simile al suddetto valore noto, secondo quanto espresso nella formula (19). 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.

Prima di proporre un confronto empirico tra diverse tecniche di identificazione degli outlier basate sulle soglie di accettazione, va sottolineato come a priori non sia in genere possibile stabilire se tale famiglia di procedure sia o meno preferibile rispetto alle tecniche basate sulla modifica dei pesi campionari.

Va però sottolineato come il ricorso alla prima famiglia di tecniche implichi necessariamente la possibilità di identificare ciascuna delle 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.

Sussistono, inoltre, 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.

Si potrebbe comunque verificare facilmente come, in pratica, ad ogni procedura di correzione dei dati anomali basata su soglie di accettazione e sulla successiva nuova stima delle osservazioni outlier corrisponda, implicitamente, una procedura di modifica dei pesi campionari (Gismondi, 2000).

5. UNA APPLICAZIONE

L'ISTAT rileva ogni mese l'ammontare delle vendite al dettaglio presso un campione di circa 7.200 imprese commerciali, e sulla base delle informazioni raccolte elabora e diffonde una serie di indici che esprimono la variazione delle vendite intercorsa tra un certo mese e la media mensile dell'anno base 1995.

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)[19].

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[20]. 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

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

[20] 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.

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 tabella 3 seguente.

TABELLA 3

Descrizione dei 20 domini 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)
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)

Sostanzialmente sono state considerate due tipologie di imprese commerciali al dettaglio: le imprese specializzate (ossia quelle che vendono esclusivamente o in prevalenza una sola tipologia di prodotti) e quelle non specializzate, a loro volta distinte in base al fatto che la tipologia dei prodotti venduti esclusivamente o in prevalenza sia di tipo alimentare o non alimentare. L'ulteriore elemento di stratificazione è dato dalla dimensione aziendale, misurata sulla base delle classi di addetti 1-2, 3-5, 6-9, 10-19 e da 20 in poi. 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 5.1. In tale tabella è riportato, in forma sintetica, il numero di unità identificate come outlier, separatamente per ciascuna delle 4 tipologie di imprese: specializzate alimentari (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 per le 5 classi di addetti: (i domini 1, 6, 11, 16 identificano la classe di addetti 1-2; 2, 7, 12, 17 la classe 3-5; 3, 8, 13, 18 la classe 6-9; 4, 9, 14, 19 la classe 10-19; 5, 10, 15, 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 $P\approx 0{,}05$. Si noti come il criterio standard operi direttamente sui rapporti *r* senza comportarne alcuna trasformazione ulteriore.

– Il metodo dei quartili definito dalle relazioni (3), (4) e (5), anch'esso applicato direttamente ai rapporti *r*; i parametri c_{inf} e $c_{\sup}$ sono stati posti entrambi uguali alla stessa costante *c*, con *c*=1 e *c*=3[21].

– Il metodo di Hidiroglou e Berthelot descritto nel paragrafo 2, sperimentato utilizzando i medesimi valori per *c*.

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

Dall'esame della tabella risulta evidente come il criterio standard identifichi il numero più contenuto di outlier (25), concentrati nello strato più numeroso (23), 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 nel caso in cui *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 5 riporta, sulla base di una stratificazione analoga a quella della tabella 4, i rapporti di variazione medi di strato ottenuti utilizzando i suddetti criteri di editing e trattando in due 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 tramite un criterio di "imputazione semplice", ossia moltiplicando l'ammontare delle vendite di giugno 1998 per il rapporto di variazione medio registrato nello strato calcolato sulle sole unità "buone".

Nel complesso sono stati quindi confrontati 14 criteri di stima: il metodo standard, le procedure dei quartili sui rapporti *r*, di Hidiroglou e Berthelot e alternativa per *c*=1 e *c*=3 e con le due suddette opzioni per il trattamento 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[22], 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 (penultima riga della tabella 5.2), tutte le procedure hanno condotto a stime della variazione media diverse dalla stima

---

[21] 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.

[22] L'universo di riferimento è composto da circa 560mila imprese, ed è quindi pressochè impossibile disporre di una certezza assoluta circa la dinamica del fatturato commerciale per il complesso delle imprese attive nel comparto.

media generale per 1 o 2 centesimi, ad eccezione del metodo dei quartili per c=3 e con un peso nullo per gli outlier, nel qual caso lo scarto è inferiore a 1 centesimo.

Se però si considera la media degli scarti presi in valore assoluto (ultima riga della tabella), questa risulta sempre pari a 0,02, a meno che non si usi la procedura di Hidiroglou e Berthelot per c=1, nel qual caso lo scarto medio è pari a 0,01.

Più indicativo è il confronto tra le medie degli scarti dalla stima media riportati nell'ultima colonna della tabella 5.2, 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 attività prevalente delle imprese comporta stime mediamente più diverse tra loro rispetto alla stratificazione per classi di addetti: nel primo caso si passa dallo scarto più elevato per le imprese specializzate alimentari (0,04) agli scarti delle tre tipologie restanti, tutti pari a 0,02; nel secondo caso, a fronte dello scarto più elevato relativo alle imprese fino a 2 addetti – pari anch'esso a 0,04 – gli scarti delle altre classi sono sempre più contenuti, risultando pari a 0,01, ad eccezione della classe 10-19 per la quale lo scarto medio è sostanzialmente nullo.

Va poi notato come le procedure che hanno generato casi con scarti in valore assoluto dalla stima media pari ad almeno 0,05 sono:

– la procedura standard per le imprese specializzate a prevalenza alimentare;

– la procedura dei quartili per le medesime imprese del punto precedente con c=1 e c=3; per la classe di addetti 1-2 con c=1;

– la procedura di Hidiroglou e Berthelot per le imprese specializzate a prevalenza alimentare e la classe di addetti 1-2 con c=3;

– la procedura alternativa per le imprese specializzate a prevalenza non alimentare con c=1 e c=3; per le imprese specializzate a prevalenza alimentare con c=3.

In sintesi, le indicazioni più rilevanti emerse dalla comparazione tra diverse procedure di editing e di trattamento delle unità anomale sembrano le seguenti:

1. la procedura alternativa comporta una riduzione molto significativa del numero di unità identificate come anomale rispetto ai metodi dei quartili e di Hidiroglou e Berthelot.

2. Nel complesso, le stime finali ottenibili dopo aver applicato le diverse procedure di controllo sono piuttosto simili tra loro, sebbene il metodo di Hidiroglou e Berthelot garantisca una stabilità delle stime, al variare dei singoli domini oggetto d'interesse, lievemente superiore rispetto alle altre procedure.

3. Gli effetti del trattamento delle unità anomale sembrano più sensibili rispetto alla tipologia imprenditoriale piuttosto che alla dimensione aziendale. Ciò può derivare dal fatto che la sola dimensione aziendale, espressa dal numero di addetti, non consente sempre di discriminare nettamente tra gruppi di imprese con strutture organizzative interne e performance economiche significativamente differenziate, mentre l'attività prevalente svolta dall'impresa al dettaglio influenza molto più significativamente, in generale, la dinamica congiunturale del fatturato.

Per quanto riguarda, infine, le soglie di accettazione per i rapporti r, nella tabella 5.3 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 due procedure si verifichi quando con riferimento alla procedura alternativa si considerano i rapporti *r* inferiori alla mediana[23]: 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 4

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

Strato		Metodo standard	Quartili su r c=1	Quartili su r c=3	Hidiroglou-Berthelot c=1	Hidiroglou-Berthelot c=3	Alternativa c=1	Alternativa c=3	Totale unità
Valori assoluti									
1-5	(specializzate alimentari)	0	224	58	226	91	52	5	447
6-10	(specializzate non alimentari)	23	1278	344	1278	599	321	26	2551
11-15	(despecializzate alimentari)	2	164	50	176	75	35	5	352
16-20	(despecializzate non alimentari)	0	26	9	26	13	6	2	54
1,6,11,16	(fino a 2 addetti)	18	996	271	996	478	256	25	1990
2,7,12,17	(da 3 a 5 addetti)	4	248	60	248	106	59	3	496
3,8,13,18	(da 6 a 9 addetti)	1	153	44	154	62	35	3	303
4,9,14,19	(da 10 a 19 addetti)	2	167	50	172	65	37	3	342
5,10,15,20	(da 20 addetti in poi)	0	128	36	136	67	27	4	273
Totale		25	1692	461	1706	778	414	38	3404
Valori percentuali									
1-5	(specializzate alimentari)	0,0	50,1	13,0	50,6	20,4	11,6	1,1	100,0
6-10	(specializzate non alimentari)	0,9	50,1	13,5	50,1	23,5	12,6	1,0	100,0
11-15	(despecializzate alimentari)	0,6	46,6	14,2	50,0	21,3	9,9	1,4	100,0
16-20	(despecializzate non alimentari)	0,0	48,1	16,7	48,1	24,1	11,1	3,7	100,0
1,6,11,16	(fino a 2 addetti)	0,9	50,1	13,6	50,1	24,0	12,9	1,3	100,0
2,7,12,17	(da 3 a 5 addetti)	0,8	50,0	12,1	50,0	21,4	11,9	0,6	100,0
3,8,13,18	(da 6 a 9 addetti)	0,3	50,5	14,5	50,8	20,5	11,6	1,0	100,0
4,9,14,19	(da 10 a 19 addetti)	0,6	48,8	14,6	50,3	19,0	10,8	0,9	100,0
5,10,15,20	(da 20 addetti in poi)	0,0	46,9	13,2	49,8	24,5	9,9	1,5	100,0
Totale		0,7	49,7	13,5	50,1	22,9	12,2	1,1	100,0

[23] La distinzione tra i casi in cui *r* sia maggiore od inferiore rispetto alla mediana deriva dalla formula (9), da cui si possono ricavare 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 i domini da 1 a 20 del prospetto 5.1 e per c=1,3. **1**:27,46; **2**:94,173; **3**:215,356; **4**:361,563; **5**:924,1642; **6**:24,47; 7:110,205; **8**:235,363; **9**:474,783; **10**:1994,3964; **11**:39,74; **12**:125,207; **13**:306,444; **14**:573,781; **15**:3348,6828; **16**:21,35; **17**:96,112; **18**:211,275; **19**:557,740; **20**:2251,3484.

TABELLA 5

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		Alternativa		
Strato	Standard	*c*=1	*c*=3	*c*=1	*c*=3	*c*=1	*c*=3	*c*=1	*c*=3	*c*=1	*c*=3	*c*=1	*c*=3	*Media (*)*
							Variazioni medie							
1-5	0,90	1,00	0,98	0,99	0,99	0,97	0,97	1,00	0,90	0,99	0,90	0,97	0,90	0,95
6-10	1,04	1,02	0,99	1,02	1,02	0,95	1,04	1,02	1,04	1,02	1,04	0,95	1,04	1,02
11-15	0,95	1,00	0,99	0,99	0,99	1,01	0,99	1,00	0,95	0,99	0,95	1,01	0,95	0,98
16-20	1,11	1,08	1,05	1,06	1,10	1,05	1,08	1,08	1,11	1,06	1,11	1,05	1,11	1,08
1,6,11,16	0,92	1,02	0,99	0,99	0,99	0,94	0,99	1,02	0,91	0,99	0,91	0,94	0,91	0,96
2,7,12,17	1,03	1,04	1,00	1,02	1,05	1,01	1,03	1,04	1,03	1,02	1,03	1,01	1,03	1,03
3,8,13,18	1,01	1,00	0,99	1,00	1,00	1,00	1,04	1,00	1,01	1,00	1,01	1,00	1,01	1,01
4,9,14,19	1,03	1,04	1,03	1,04	1,04	1,03	1,03	1,04	1,03	1,04	1,03	1,03	1,03	1,03
5,10,15,20	1,02	1,05	1,00	1,04	1,06	1,00	1,02	1,05	1,02	1,04	1,02	1,00	1,02	1,03
Totale	1,00	1,03	1,00	1,02	1,03	1,00	1,02	1,03	1,00	1,02	1,00	1,00	1,00	1,01
							Scarti dalla stima media							
1-5	-0,05	0,05	0,03	0,04	0,04	0,02	0,02	0,05	-0,05	0,04	-0,05	0,02	-0,05	0,04
6-10	0,03	0,00	-0,03	0,00	0,01	-0,06	0,02	0,00	0,03	0,00	0,03	-0,06	0,03	0,02
11-15	-0,03	0,02	0,01	0,01	0,01	0,03	0,01	0,02	-0,04	0,01	-0,04	0,03	-0,04	0,02
16-20	0,03	0,01	-0,03	-0,02	0,02	-0,02	0,00	0,01	0,03	-0,02	0,03	-0,02	0,03	0,02
1,6,11,16	-0,03	0,07	0,04	0,04	0,05	-0,01	0,04	0,07	-0,04	0,04	-0,04	-0,01	-0,04	0,04
2,7,12,17	0,01	0,01	-0,02	0,00	0,03	-0,01	0,01	0,01	0,01	0,00	0,01	-0,01	0,01	0,01
3,8,13,18	0,01	0,00	-0,01	-0,01	-0,01	0,00	0,04	0,00	0,01	-0,01	0,01	0,00	0,01	0,01
4,9,14,19	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,00
5,10,15,20	-0,01	0,01	-0,03	0,00	0,02	-0,03	-0,01	0,01	-0,01	0,00	-0,01	-0,03	-0,01	0,01
Totale	-0,01	0,02	0,00	0,01	0,02	-0,01	0,02	0,02	-0,01	0,01	-0,01	-0,01	-0,01	0,01
Media (*)	0,02	0,02	0,02	0,01	0,02	0,02	0,02	0,02	0,02	0,01	0,02	0,02	0,02	0,02

(*) Media degli scarti presi in valore assoluto.

TABELLA 6

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

	Quartili su *r*				Procedura alternativa (*r*≥mediana)				Procedura alternativa (*r*<mediana)			
	c=1		*c*=3		*c*=1		*c*=3		*c*=1		*c*=3	
Strato	Ainf	Asup	Ainf	Asup	Ainf	Asup	Ainf	Asup	Ainf	Asup	Ainf	Asup
1-5	0,92	1,07	0,79	1,23	1,11	1,26	0,99	1,42	0,93	1,05	0,83	1,18
6-10	0,89	1,15	0,65	1,42	1,20	1,47	1,01	1,83	0,88	1,08	0,71	1,28
11-15	0,93	1,05	0,81	1,17	1,07	1,20	0,95	1,37	0,93	1,05	0,82	1,18
16-20	0,96	1,18	0,82	1,49	1,12	1,34	0,93	1,59	0,91	1,09	0,77	1,32
1,6,11,16	0,90	1,14	0,72	1,46	1,17	1,44	0,97	1,76	0,89	1,09	0,73	1,33
2,7,12,17	0,90	1,13	0,74	1,43	1,14	1,35	0,97	1,61	0,91	1,08	0,78	1,27
3,8,13,18	0,92	1,07	0,78	1,22	1,10	1,26	0,98	1,45	0,92	1,05	0,80	1,18
4,9,14,19	0,96	1,11	0,79	1,24	1,10	1,27	0,97	1,48	0,92	1,06	0,79	1,20
5,10,15,20	0,96	1,11	0,81	1,28	1,10	1,26	0,96	1,45	0,93	1,06	0,81	1,21
Totale	0,93	1,11	0,77	1,33	1,12	1,32	0,97	1,55	0,91	1,07	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 $I_{12,p}$ relativo alla variabile I_{12} tale che l'area sottesa dalla coda di destra sia pari a *p*, basta imporre la condizione:

$$(1-p)=\text{Prob}(\eta\le F_{\sup,p})$$

dove $F_{\sup,\text{P}}$ 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:

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

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:

$$I_{12,p}=\left(\frac{\alpha_2}{\alpha_1}\right)F_p=\left(\frac{\frac{1}{\alpha_1}}{\frac{1}{\alpha_2}}\right)F_p=\frac{\text{E}(X_1)}{\text{E}(X_2)}F_p,$$

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 la variabile I_{12} è moltiplicata per la costante *a*, segue ovviamente che: $a I_{12,p} = a \dfrac{E(X_1)}{E(X_2)} F_p$. Ponendo prima a=$q_{0,5}$ e $I_{12} = 1/r_i$, e poi a=$1/q_{0,5}$ e $I_{12} = r_i$ si perviene, infine, alle relazioni (14) e (15).

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:

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

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.

RIFERIMENTI BIBLIOGRAFICI

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

V. BARNETT (1978), *Outliers in Statistical Data*, John Wiley & Sons, New York.

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

H.E. DAVILA (1992), *The Hidiroglou-Berthelot Method*, "Statistical Data Editing Methods and Techniques", United Nations.

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

W.S. EDWARDS, D. CANTOR (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.

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

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

E. GARCIA, V. PEIRATS (1994*), Evaluation of Data Editing Procedures: Results of a Simulation Approach*, in "Statistical Data Editing Methods and Techniques - Conference of European Statisticians and Studies", Vol. 1, 44, pp. 52-68.

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

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

R. GISMONDI (2000a), *Metodi per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di variazioni*, "Contributi Istat", 8.

R. GISMONDI (2000b), *Metodi per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di variazioni*, "Rivista di statistica ufficiale", 2, pp. 97-130, Franco Angeli, Milano.

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

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

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

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

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

J.G. KOVAR, W.E. WINKLER (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.

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

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

O. LUZI (1998), "The Outlier Localisation Based on the Use of the Hidiroglou and Berthelot Function", *Italian Journal of Applied Statistics*, vol.10, 2, pp. 197-216.

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

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

D. SEARLS (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.

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

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

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

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

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

RIASSUNTO

Un confronto tra metodi di identificazione di osservazioni outlier in indagini longitudinali finalizzate alla stima di una variazione: proposte teoriche e verifiche empiriche

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' stata infine illustrata una applicazione ad un caso concreto, in cui sono state poste a confronto 14 modalità di trattamento statistico degli outlier.

ABSTRACT

A comparison among methods for detecting outliers in longitudinal surveys: theoretical proposals and empirical attempts

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, originally 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 micro-data alterations. Moreover, we proposed and compared some criteria for defining the tolerance intervals. Finally, we carried out an empirical study, in which we compared 14 different ways for dealing with outliers.