Statistica Applicata Vol. 10, n. 2, 1998

LA CORREZIONE PROBABILISTICA DEI DATI: IL TRATTAMENTO CONGIUNTO DEGLI ERRORI DI RILEVAZIONE CASUALI E SISTEMATICI MEDIANTE L'APPLICAZIONE DEL TEOREMA DI BAYES ALLA METODOLOGIA FELLEGI–HOLT

Giulio Barcaroli

ISTAT, Servizio Studi Metodologici.

Riassunto

La metodologia Fellegi–Holt è anche nota come "approccio probabilistico" alla correzione dei dati statistici, contrapposto a quello "deterministico". Quest'ultimo prevede procedure di correzione che fissano a priori le azioni di imputazione da intraprendere per ogni particolare incompatibilità riscontrata in un dato record: non si pone cioè un problema di identificazione degli errori, di cui le incompatibilità attivate sono una conseguenza. Al contrario, l'approccio probabilistico considera fondamentale risolvere tale problema in modo ottimale per giungere ad una correzione degli errori, e dunque ad una eliminazione delle incompatibilità, che mediamente centri l'obiettivo di individuare le variabili che sono realmente errate, ripristinando i valori "veri" al posto di quelli "falsi". Fellegi e Holt, nel loro articolo di illustrazione della metodologia che ne ha preso il nome, non hanno mai esplicitato le basi teoriche grazie alle quali il "principio del minimo cambiamento", criterio alla base dell'algoritmo di identificazione degli errori, è in grado di garantire tale obiettivo. Applicando il teorema di Bayes è possibile verificare che l'applicazione della metodologia, sotto determinate ipotesi, permette effettivamente di determinare le variabili che più probabilmente sono errate e che determinano dunque l'attivazione delle incompatibilità in un dato record. La metodologia in questione, nella versione attuale, permette però di trattare adeguatamente i soli errori casuali, non quelli sistematici: per questi ultimi è stato finora utilizzato l'approccio deterministico. La formalizzazione ottenuta mediante l'applicazione del teorema di Bayes permette, mediante una modifica dell'algoritmo di identificazione degli errori, di inglobare il trattamento degli errori sistematici all'interno della metodologia.

1. INTRODUZIONE

La fase di controllo e correzione dei dati rilevati mediante un'indagine statistica riveste un'importanza fondamentale ai fini del contenimento della com-

ponente non campionaria dell'errore. Quando le attività in essa rientranti sono svolte in modo automatico mediante utilizzo di appositi programmi, si parla di approccio deterministico oppure probabilistico a seconda della metodologia incorporata nel software.

La differenza fondamentale è riscontrabile nella fase dell'identificazione degli errori, nel momento cioè in cui, essendo state riscontrate incompatibilità logico–matematiche o anomalie statistiche in un dato record, si procede a determinare, tra tutte le potenziali candidate, quelle variabili che, contenendo valori errati, ne sono responsabili. Nell'approccio deterministico, per ogni singola incompatibilità riscontrabile in un record, sono predeterminate le variabili errate che ne costituiscono la causa e, in quanto tali, da correggere. Ciò, per di più, avviene in modo indipendente dall'eventuale attivazione da parte dello stesso record di altre incompatibilità, il che porta a correzioni che possono rivelarsi contraddittorie con quelle già effettuate. Un tale approccio non solo non dà garanzie circa l'obiettivo di trovare e correggere quanti più errori possibili, ma al contrario è spesso fonte di introduzione di nuovi errori. Al contrario, l'approccio probabilistico, quale quello introdotto da Fellegi e Holt nel 1976, è così detto perché, obbedendo al principio del minimo cambiamento (l'identificazione degli errori avviene minimizzando il numero di variabili giudicate errate responsabili delle incompatibilità attivate), permette di scegliere la configurazione di errore più probabile tra quelle possibili.

Ciò avviene, però, solo se sono rispettate alcune condizioni: prima tra tutte, la natura casuale degli errori. Se al contrario, sono presenti componenti sistematiche dell'errore, la metodologia proposta da Fellegi e Holt non è più in grado di risolvere correttamente il problema dell'identificazione degli errori: paradossalmente, l'approccio deterministico si rivela più efficace (anche se non ottimale) nel trattamento di tale tipologia di errori.

Il presente lavoro intende proporre una estensione dell'approccio probabilistico al trattamento non solo degli errori casuali ma anche di quelli sistematici, mediante una modifica dell'algoritmo di Fellegi e Holt per l'identificazione degli errori, basata sulla considerazione esaustiva delle probabilità di ogni potenziale configurazione di errore, computate in accordo col teorema di Bayes, e sulla scelta di quella con probabilità più alta.

Nel secondo paragrafo vengono rapidamente descritte le caratteristiche degli approcci deterministici e probabilistici e, di quest'ultimo, viene riportato l'algoritmo per l'identificazione degli errori.

Nel terzo paragrafo, sono definiti gli errori casuali e quelli sistematici, assieme al differente trattamento richiesto dalle due tipologie.

Nel quarto viene proposta una formalizzazione dell'approccio probabilistico facente uso del teorema di Bayes.

Nel quinto, infine, viene proposto un algoritmo per l'identificazione degli errori analogo a quello di Fellegi ed Holt, ma basato sul principio della massimizzazione della probabilità della configurazione dell'errore, anziché su quello del minimo cambiamento.

2. APPROCCIO DETERMINISTICO VS. APPROCCIO PROBABILISTICO

La fase di applicazione delle regole di dominio, di compilazione e di compatibilità ai dati grezzi non può che essere compiuta in modo deterministico: per ogni record, o per gruppi di record, vengono applicate tali regole che, se verificate, segnalano la presenza di errori. Ad esempio:

SE (età $15 \leq$ E stato civile $\neq$ *celibe*) ALLORA incompatibilità

Una regola di questo tipo non individua, di per sé, l'errore che ne causa l'attivazione: infatti, l'errore (inteso come *valore non vero,* cioè non rispondente alla modalità del carattere che l'unità effettivamente possiede) può celarsi in una o nell'altra delle variabili, o in entrambe.

È al momento dell'*identificazione degli errori* che diviene decisivo il tipo di approccio adottato. Nell'*approccio deterministico,* ad ogni situazione di incompatibilità segue, contestualmente, l'indicazione delle variabili che debbono essere considerate errate, e, in quanto tali, il cui valore deve essere modificato. Nell'esempio considerato avremo, per ipotesi:

SE (età ≤ 15 E stato civile $\neq$ *celibe*) ALLORA stato civile $\leftarrow$ *celibe*

il che significa che, se in un record è attivata la condizione di incompatibilità indicata nella parte "SE", la regola indica l'azione da effettuare per correggere l'errore, che consiste nell'imputare la modalità *celibe* alla variabile "stato civile". Una regola di questo tipo configura una sorta di gerarchia tra le variabili che vi compaiono, implicitamente determinata dal grado di affidabilità di ognuna di esse: la scelta di considerare errato lo stato civile, anziché l'età, indica che quest'ultima viene supposta più "sicura" rispetto alla prima.

Generalizzando, una volta verificate, mediante le regole di incompatibilità contenute nella parte "SE", una o più situazioni di incoerenza in un dato record, sono determinate a priori le azioni da intraprendere per riportare il medesimo record in una situazione di coerenza.

Le procedure deterministiche sono generalmente costituite da regole di imputazione deterministica (R.I.D.) del tipo:

SE [condizione di incompatibilità] ALLORA [azione di correzione]

La condizione di incompatibilità della regola esprime le relazioni intercorrenti tra le variabili implicate; l'azione di correzione riguarda delle variabili che possono essere o meno incluse nella parte "SE".

Un record, durante l'esecuzione della procedura di correzione, potrà causare l'attivazione di alcune di queste regole (quelle in corrispondenza delle quali è verificata la parte SE): in tal caso saranno modificate le variabili indicate nella parte ALLORA assegnando loro valori predefiniti o scelti in altro modo.

Al contrario di quello precedente, l'*approccio probabilistico* non prevede la possibilità (o la necessità) di definire a priori, per ogni possibile incoerenza o anomalia riscontrata, le azioni da intraprendere per eliminare gli errori che ne sono la causa: l'esperto statistico deve limitarsi a definire le situazioni di incompatibilità, demandando ad un prefissato algoritmo il compito di effettuare le imputazioni necessarie a riportare il record ad una situazione di coerenza.

L'approccio probabilistico ha il suo punto di riferimento nella cosiddetta metodologia Fellegi–Holt (Fellegi, Holt 1976). Su tale metodologia, nata per il trattamento delle variabili qualitative e successivamente ampliata e modificata per quelle quantitative, si basano numerosi sistemi software correntemente applicati in diversi Istituti nazionali di statistica: SCIA (Barcaroli *et al.* 1995) e DAISY (Barcaroli, Venturi 1997) in Italia, SPEER (Winkler, Draper 1997) e DISCRETE (Winkler, Petkunas 1997) nell'U.S. Bureau of the Census, GEIS (Kovar *et al.* 1991) in Canada, DIA (Garcia Rubio, Villan Criado 1988) in Spagna, CHERRYPI (De Waal 1996) in Olanda.

Una procedura probabilistica è composta da regole di incompatibilità che, seguendo la terminologia di Fellegi e Holt, vengono chiamate *edit in forma normale*. Un edit in forma normale è costituto dalla congiunzione di due o più condizioni sui valori di variabili del record. La parte SE di una R.I.D. (cioè quella che esprime la condizione di incompatibilità) può corrispondere a uno o più edit in forma normale; riconsiderando il precedente esempio di R.I.D. il corrispondente edit sarà:

$$(\text{età} \leq 15) \cap (\text{stato civile} \neq \textit{celibe})$$

Ogni record sottoposto a controllo può determinare l'attivazione di uno o più edit. L'algoritmo per l'identificazione degli errori ideato da Fellegi e Holt si basa sulla considerazione *simultanea* (non sequenziale come nel caso deterministico) di tutte le regole di incompatibilità attivate. Considerando tutte le k variabili coinvolte in queste ultime, il problema di determinare il sottoinsieme di quelle che contengono i valori effettivamente errati determinanti l'attivazione degli edit, è risolto individuando tra tutti i 2^k insiemi alternativi l'insieme di dimensione minima le cui variabili, una volta corrette, permettono di disattivare tutti gli edit "falliti". La

minimalità dell'insieme così determinato fa sì che, *sotto determinate condizioni,* la scelta effettuata, oltre a soddisfare un requisito di base (intervenire il meno possibile nei dati rilevati), sia mediamente quella con la maggiore probabilità di essere corretta (vedi par.4). Vedremo che la presenza di errori sistematici determina il venire meno di tali condizioni.

3. ERRORI CASUALI ED ERRORI SISTEMATICI

Definiamo *errori casuali* quegli errori non campionari che non determinano distorsioni significative nelle distribuzioni. Infatti:

- nel caso delle variabili quantitative tali errori si distribuiscono con valore e segno variabili da unità a unità, in tal modo producendo un effetto nullo a livello di valori medi;
- per quanto riguarda le variabili qualitative, l'effetto nullo si produce a livello di distribuzione di frequenza delle modalità, dato che gli scambi tra modalità "vere" e "false" avvengono in modo tale da compensarsi per un numero di casi sufficientemente elevato.

Per tale tipologia di errori non è possibile individuare una causa precisa che non sia quella di una limitata accuratezza dei rispondenti nella valutazione dei valori delle variabili che sono richiesti nella rilevazione.

Al contrario, definiamo come *errori sistematici* gli errori non campionari dipendenti da elementi turbativi del buon andamento della rilevazione, tali da determinare nelle distribuzioni delle variabili interessate l'introduzione di distorsioni significative.

Le condizioni che danno luogo agli errori sistematici possono verificarsi essenzialmente nelle fasi di compilazione, di registrazione e di revisione dei modelli:

1. al momento della *compilazione,* i rispondenti rifiutano di fornire i valori "veri" (essenzialmente per motivi di privacy), oppure non comprendono quanto realmente richiesto nei quesiti o nelle norme di compilazione dei quesiti (per difetti insiti nel questionario, oppure per responsabilità dei rilevatori) e danno risposte tendenzialmente scorrette;
2. al momento della *registrazione,* gli operatori seguono un piano di registrazione con indicazioni errate, oppure ne ignorano sistematicamente le indicazioni;
3. al momento della *revisione,* gli impiegati seguono indicazioni tendenti, ad esempio, a eliminare dati che contrastino con informazioni della stessa indagine in occorrenze precedenti, o con informazioni provenienti da indagini diverse.

Normalmente, la presenza di errori sistematici determina nelle variabili interessate una probabilità di errore significativamente superiore rispetto a quella delle variabili affette da soli errori casuali.

Entrambi i tipi di errore, casuale e sistematico, sono identificabili solo se determinano l'attivazione di una o più condizioni di incompatibilità, altrimenti non esiste modo per individuarli se non quello di reintervistare le unità rispondenti e procedere a riconciliazione delle risposte divergenti.

Le condizioni di incompatibilità possono così essere classificate:

- valori fuori dominio;
- mancate risposte parziali;
- incoerenze tra i valori assunti da due o più variabili.

Le mancate risposte parziali sono scomponibili in:

1. mancate risposte parziali *certe,* in quanto la risposta ad una data variabile è sempre dovuta dal rispondente, indipendentemente dai valori assunti dalle altre variabili: questa componente è assimilabile a quella dei fuori dominio, in quanto la mancata risposta non è ammessa nel dominio della variabile;
2. mancate risposte parziali *potenziali,* in quanto il valore mancante registrato in una data variabile contrasta con i valori assunti dalle variabili "filtro" che rendono per il rispondente obbligatoria la risposta per quella variabile. Questa componente è assimilabile alle incoerenze, e come nel caso delle incoerenze pone un problema di *scelta* di quali variabili considerare errate: in altre parole, l'errore si trova nelle variabili–filtro, o in quelle la cui compilazione dipende dalle risposte fornite alle variabili–filtro?

Per quanto riguarda le incoerenze, esse possono essere di tipo *logico* quando coinvolgono variabili di tipo qualitativo, oppure *matematico–statistico* quando riguardano variabili quantitative.

Errori sistematici che danno luogo a valori fuori dominio

Consideriamo il seguente esempio(*):

Attività di qualificazione	**conclusa**	**interrotta**
volontariato	□ 01	□ 02
tirocinio	□ 03	□ 04
dottorato	□ 05	□ 06
borsa di studio	□ 07	□ 08
specializzazione	□ 09	□ 10
altre attività	□ 11	□ 12

(*) *Dal questionario dell'indagine ISTAT sull'Inserimento professionale dei Laureati del 1991 (semplificato).*

Il quesito dà luogo a 6 differenti variabili. Si è verificato nella realtà che, in sede di registrazione, ad ognuna di esse sono stati attribuiti i valori 01–02, anziché quelli previsti sul modello e sul piano di registrazione. Ne consegue che tutte le variabili, esclusa la prima, soffrono di errori sistematici immediatamente individuabili dalla condizione di fuori dominio, ed altrettanto immediatamente eliminabili mediante correzione deterministica (aggiunta di un valore costante).

Errori sistematici che danno luogo a mancate risposte parziali

Consideriamo il seguente esempio(*):

Reddito medio mensile della famiglia	
Fino a 600.000	□ 01
Da 600.001 a 1.000.000	□ 02
Da 1.000.001 a 1.500.000	□ 03
Da 1.500.001 a 2.000.000	□ 04
Da 2.000.001 a 3.000.000	□ 05
Da 3.000.001 a 4.000.000	□ 06
Oltre 4.000.000	□ 07

Per ragioni fiscali, i rispondenti sono portati a dichiarare classi di reddito inferiori a quelle reali, o a non dichiararle affatto. Solo in quest'ultimo caso l'errore è rilevabile (come mancata risposta parziale).

Errori sistematici che danno luogo ad incoerenze

In un questionario è dato il seguente flusso di quesiti(**):

1. POSIZIONE NELLA PROFESSIONE
   LAVORATORE DIPENDENTE □ 1
   LAVORATORE INDIPENDENTE □ 2

(*rispondere al seguente quesito solo se ha risposto 1*)
2. SVOLGE UN LAVORO NEL PUBBLICO O NEL PRIVATO?
   NEL PUBBLICO □ 1
   NEL PRIVATO □ 2

Nel corso dell'elaborazione dei dati dell'indagine, si è riscontrata la tendenza dei rispondenti ad ignorare l'istruzione di compilazione, ed a rispondere al secondo quesito indipendentemente dalla risposta fornita al primo. Ciò comporta la possi-

(*) *Dal modello dell'indagine ISTAT sui Consumi delle Famiglie (semplificato).*
(**) *Dal questionario dell'indagine ISTAT sull'Inserimento professionale dei Laureati del 1991 (semplificato).*

bilità di avere, in corrispondenza a valori della variabile "posizione nella professione" (POSPRO nel seguito) diverse da "1", risposte nella variabile "svolge un lavoro..." (LAVORA nel seguito) eguali a "1" e "2" anziché al valore nullo (*blank* nel seguito).

Poiché la distribuzione degli indipendenti è maggiormente accentrata nel settore privato, l'effetto finale dell'errore consisterà in una distorsione della variabile LAVORA verso la modalità "privato" a scapito della modalità "pubblico".

La presenza di tale errore è individuabile mediante l'applicazione della regola:

SE (POSPRO $\neq$ 1) E (LAVORA $\neq$ *blank*) ALLORA incompatibilità

L'incompatibilità individuata, però, non ci dice se la variabile da considerare errata e, tale dunque da determinarne l'attivazione, sia POSPRO oppure LAVORA.

L'identificazione degli errori e la correzione nel caso degli errori sistematici

Nel caso in cui l'errore sistematico dia luogo a valori fuori dominio o a mancate risposte parziali "sicure", analogamente al caso degli errori casuali non si pone il problema dell'identificazione dell'errore, di quale variabile, cioè, contenga effettivamente il valore errato e debba pertanto essere imputata.

Tale problema si pone invece nel caso in cui l'errore sistematico dia luogo ad incoerenze.

L'algoritmo probabilistico di identificazione degli errori, proprio della metodologia Fellegi–Holt, non è, in linea di principio, in grado di affrontare correttamente la identificazione di questo tipo di errore: anzi, la sua applicazione introduce ulteriori distorsioni. Infatti, consideriamo il seguente edit in forma normale definito per identificare l'incompatibilità dovuta al mancato rispetto della norma di compilazione:

$$(\text{POSPRO} \neq 1) \cap (\text{LAVORA} \neq \textit{blank})$$

Supponiamo di applicare un algoritmo di scelta che ottemperi unicamente al principio del minimo cambiamento posto da Fellegi e Holt. Nel caso in esame abbiamo tre insiemi candidati, due (uno contenente POSPRO, l'altro contenente LAVORA) con dimensione pari a uno ed il terzo di dimensione due (LAVORA e POSPRO). La scelta ovviamente ricadrà indifferentemente su uno dei primi due insiemi, e presupponendo un meccanismo di rotazione randomizzata di tale scelta, possiamo supporre che si determini un cambiamento del valore di POSPRO da 2 a 1 in circa il 50% dei casi, e di LAVORA da 1,2 a *blank* nel restante 50%: ma noi sappiamo che una grande percentuale delle volte in cui questo edit è attivato è dovuta al fatto che il rispondente ha ignorato la regola di compilazione della

variabile LAVORA, e che quindi l'errore è quasi sempre contenuto in questa variabile. Se ignoriamo questa considerazione, le imputazioni possono certamente ridurre la distorsione in LAVORA (anche se di una quantità lontana da quella necessaria), ma altrettanto certamente producono una nuova distorsione in POSPRO, che vede sistematicamente cambiato il proprio valore anche quando quest'ultimo è esatto.

D'altro canto, l'applicazione di una regola deterministica come la seguente:

SE (POSPRO ≠ 1) E (LAVORA ≠ *blank*) ALLORA LAVORA ← *blank*

ha come effetto quello di introdurre una distorsione in LAVORA di segno opposto a quella determinata dall'errore sistematico, in quanto in tal modo viene totalmente ignorata la componente casuale degli errori che fa sì che talvolta la variabile errata sia POSPRO e non LAVORA.

Riassumendo e generalizzando, se la identificazione degli errori avviene mediante l'approccio probabilistico, ciò determina l'introduzione di distorsioni nelle variabili collegate a quella/e affetta/e da errore sistematico. Se invece viene utilizzato l'approccio deterministico, vengono ignorati totalmente gli errori casuali che interessano sia la variabile affetta da errore sistematico che le variabili collegate. Sono state avanzate soluzioni (Garcia Rubio, Villan Criado 1988) per una integrazione del trattamento di errori sistematici e casuali. L'approccio proposto non supera la contrapposizione tra regole deterministiche per il trattamento degli errori sistematici ed edit in forma normale per quello degli errori casuali, ma ne propone un uso sequenziale e coordinato, con una analisi preventiva dei due gruppi di regole, al fine di evitare i conflitti che altrimenti possono verificarsi. Tutto ciò non impedisce però che l'applicazione del passo deterministico, effettuato separatamente e preventivamente a quello probabilistico, soffra degli inconvenienti citati in precedenza: le modalità di identificazione degli errori ignorano completamente la componente casuale, e per tale motivo vengono introdotte distorsioni nelle variabili indicate nella parte "SE" delle R.I.D..

4. FORMALIZZAZIONE DELL'APPROCCIO PROBABILISTICO MEDIANTE IL TEOREMA DI BAYES

Richiamiamo brevemente il teorema di Bayes.

Se $q_1, q_2, \ldots, q_n$ sono un insieme di eventi mutuamente esclusivi ed esaustivi, allora la probabilità di q_r, condizionata ad un determinato ulteriore evento p, è data dalla seguente espressione:

$$\Pr(q_r \mid p) = \Pr(q_r)\Pr(p \mid q_r) / \sum_{i=1}^{n} \Pr(q_i)\Pr(p \mid q_i) \qquad (1)$$

Nelle principali applicazioni, p è un evento osservato, le q_r ne sono le possibili ipotesi esplicative. La Pr $(q_r|p)$ è detta *probabilità a posteriori,* la Pr (q_r) è la *probabilità a priori,* mentre la Pr $(p|q_r)$ costituisce la *verosimiglianza.* In sostanza, il teorema permette di determinare le probabilità delle ipotesi esplicative, una volta verificatosi un dato evento.

L'identificazione degli errori nella metodologia Fellegi–Holt

Per l'illustrazione completa della metodologia rimandiamo al testo fondamentale (Fellegi, Holt, 1976). Qui riportiamo il classico esempio di funzionamento dell'algoritmo di identificazione degli errori, basato sul principio del minimo cambiamento.

Sia dato il seguente piano di incompatibilità, stilato nella notazione propria della metodologia Fellegi–Holt:

e_1 : (ETA ∈ 0–14) ∩ (STATO_CIVILE ≠ *celibe*)

e_2 : (STATO_CIVILE = celibe) ∩ (RELAZ_CF(*) = coniuge)

e_3 : (ETA ∈ 0–14) ∩ (RELAZ_CF = coniuge)

Si consideri ora il seguente record:

[(ETA = 9) ∩ (STATO_CIVILE = coniugato) ∩ (RELAZ_CF = coniuge)]

che attiva le incompatibilità e_1 ed e_3. L'algoritmo di identificazione degli errori determina l'insieme minimo di variabili che, una volta opportunamente modificate, possano disattivare tutti gli edit attivati senza al contempo attivarne degli altri. Gli insiemi candidati sono:

1. { ETA }
2. { STATO_CIVILE }
3. { RELAZ_CF }
4. { ETA, STATO_CIVILE }
5. { ETA, RELAZ_CF }
6. { STATO_CIVILE, RELAZ_CF }
7. { ETA, STATO_CIVILE, RELAZ_CF }

Tutti gli insiemi della lista, ad eccezione del secondo e del terzo, rispondono al requisito di disattivazione potenziale degli edit attivati: viene scelto il primo, che è di dimensione minima. La correzione del record avviene dunque assegnando alla variabile ETA un valore interno all'intervallo complementare a quello indicato negli edit e_1 ed e_3.

(*) *RELAZ_CF: relazione col capofamiglia, o con la persona di riferimento.*

L'ottimalità delle scelte dell'algoritmo di Fellegi–Holt sotto certe condizioni, la sua capacità cioè di localizzare gli errori più probabili è stata affermata in più occasioni (Barcaroli, D'Angiolini, 1994). Il ricorso al teorema di Bayes permette di formalizzare tale ottimalità e di individuare in modo rigoroso le condizioni che la assicurano.

Formalizzazione della identificazione degli errori mediante Bayes

Utilizzando l'esempio testé introdotto, definiamo come evento osservato l'evento

p : attivazione degli edit e_1 ed e_3.

Definiamo quindi l'insieme delle possibili ipotesi esplicative di p nel seguente modo:

	ETA=9	STATO_CIVILE = coniugato	RELAZ_CF = coniuge
q_1	falso	vero	vero
q_2	vero	falso	vero
q_3	vero	vero	falso
q_4	falso	falso	vero
q_5	falso	vero	falso
q_6	vero	falso	falso
q_7	falso	falso	falso

(Si noti che le q_r corrispondono esattamente ai sette insiemi "candidati" introdotti in precedenza).

Le q_r sono ipotesi che si escludono a vicenda, ed esaustive: siamo quindi nelle condizioni di applicabilità del teorema di Bayes.

Dobbiamo ora indicare come valutare le verosimiglianze e le probabilità a priori.

Le verosimiglianze $\Pr(p|q_r)$ possono assumere solo valori 1 e 0, a seconda che la configurazione di errore q_r possa effettivamente causare o meno p, cioè l'attivazione congiunta degli edit e_1 ed e_3.

È facile verificare quanto riportato nella seguente tabella:

	e_1 attivato	e_3 attivato	$\Pr(p\|q_r)$
q_1	si	si	1
q_2	si	no	0
q_3	no	si	0
q_4	si	si	1
q_5	si	si	1
q_6	si	si	1
q_7	si	si	1

A titolo esplicativo, consideriamo le ipotesi q_1 e q_2. La prima, sulla base dei valori assunti dal record, e dell'assunzione di falsità del valore contenuto nella variabile ETA, determina necessariamente l'attivazione sia di $\hat{e}_1$ che di e_3: in tal caso, Pr $(p \mid q_1) = 1$. La seconda è in grado di spiegare solo l'attivazione di e_1, ma non di e_3: in quest'ultimo edit è infatti assente la variabile che si suppone errata (STATO_CIVILE), e ciò fa sì che Pr $(p|q_2) = 0$.

Ipotizzato tale comportamento delle verosimiglianze, la (1) diventa:

$$\Pr(q_r \mid p) = \begin{cases} \Pr(q_r) / \sum_{i=1}^{n'} \Pr(q_i) & (\forall i : \Pr(p \mid q_i) = 1) \\ 0 & \text{altrimenti} \end{cases} \qquad (2)$$

Il termine a denominatore è costante, dunque la probabilità a posteriori dipende solo dalla probabilità a priori della configurazione di errore q_r.

Condizioni per l'ottimalità dell'algoritmo Fellegi–Holt per la identificazione degli errori

Per quanto detto nel precedente paragrafo, *sotto determinate condizioni il principio del minimo cambiamento alla base dell'algoritmo di Fellegi–Holt per la identificazione degli errori, assicura la scelta della configurazione di errore più probabile.*

Tali ipotesi sono:

1. indipendenza degli errori sulle singole variabili;
2. uniformità delle probabilità di errore sulle singole variabili;
3. probabilità di errore sulle singole variabili strettamente inferiori a 0.5.

Infatti, nel caso di indipendenza degli errori abbiamo:

$$\Pr(q_r) = \prod_{i=1}^{m} \Pr(A_i^r)$$

dove A_i^r è l'asserzione di verità/falsità sulla i–esima variabile sotto l'ipotesi q_r. Ne deriva che la (2) diviene:

$$\Pr(q_r \mid p) = \prod_{i=1}^{m} \Pr(A_i^r) / \text{costante} \qquad (\forall r : \Pr(p \mid q_r) = 1) \qquad (3)$$

Sostenere l'uniformità delle probabilità di errore, equivale a dire:

Pr (valore di v_i: falso) = Pr (valore di v_j: falso) $\quad \forall$i,j

o, identicamente:

Pr (valore di v_i: vero) = Pr (valore di v_j: vero) $\quad \forall$i,j

Se le probabilità di errore sono al di sotto del valore 0.5 (o se le probabilità di esattezza sono al di sopra di 0.5), ne consegue che *ogni configurazione di errore* q_i *che abbia meno asserzioni di falsità di una configurazione di errore* q_j, *ha una probabilità a posteriori più alta di quest'ultima.*

Nell'esempio svolto in precedenza, supponiamo errori indipendenti sulle tre variabili, e probabilità di errore equidistribuita e pari a 0.1. Si consideri la situazione riportata nella tabella seguente:

	ETA=9	STATO_CIVILE = coniugato	RELAZ_CF = coniuge	$\prod_1^m i \ \Pr(A_i^r)$	$\Pr(q_r \mid p)$
q_1	falso (0.1)	vero (0.9)	vero (0.9)	0.081	0.743
q_4	falso (0.1)	falso (0.1)	vero (0.9)	0.009	0.082
q_5	falso (0.1)	vero (0.9)	falso (0.1)	0.009	0.082
q_6	vero (0.9)	falso (0.1)	falso (0.1)	0.009	0.082
q_7	falso (0.1)	falso (0.1)	falso (0.1)	0.001	0.009
					1.000

Sotto le condizioni ipotizzate, la scelta della variabile ETA da correggere, scelta effettuata dall'algoritmo di Fellegi–Holt in base al principio del minimo cambiamento e riportata nel paragrafo 2, è quindi equivalsa a localizzare la configurazione di errore più probabile tra quelle possibili, in base ai valori del record ed alle incompatibilità da questo attivate.

5. UNA MODIFICA DELL'ALGORITMO FELLEGI–HOLT DI IDENTIFICAZIONE DEGLI ERRORI PER IL TRATTAMENTO CONGIUNTO DI ERRORI CASUALI E SISTEMATICI

L'algoritmo di identificazione degli errori proposto da Fellegi e Holt sceglie le variabili da modificare in base al principio del minimo cambiamento, cioè in pratica viene scelta la configurazione di errore che risponde a due requisiti:

- "spiega" gli edit attivati;
- minimizza le modifiche necessarie per disattivare tali edit.

Abbiamo visto come sotto determinate ipotesi ciò equivale a scegliere la configurazione di errore più probabile: *ma se queste ipotesi non sono rispettate, come nel caso in cui sono presenti errori sistematici, l'algoritmo non è in grado di comportarsi correttamente.*

In questo lavoro si propone una semplice modifica dell'algoritmo di Fellegi–Holt per la identificazione degli errori che permette di trattare correttamente e

congiuntamente sia gli errori casuali che quelli sistematici: *l'algoritmo deve scegliere non già l'insieme minimale di variabili da modificare, ma quello corrispondente alla configurazione di errore con la probabilità a posteriori più alta, calcolata mediante le singole probabilità di errore relative alle diverse variabili.*

Tutto ciò che si richiede è la stima di tali probabilità, cioè delle Pr (A_i^r) presenti nella (3).

In corrispondenza ad ogni record che attivi delle incompatibilità, le Pr (A_i^r) vanno calcolate nel seguente modo:

$$\text{Pr(valore di } v_i : \text{ falso)} = \begin{cases} P'_j & \text{se ricorrono le condizioni per un dato errore sistematico} \\ P'' & \text{altrimenti} \end{cases} \qquad (4)$$

Pr (valore di v_i: vero) = 1 – Pr (valore di v_i: falso)

dove P'_j è una stima della probabilità di errore sistematico sulla v_i qualora se ne verifichino le condizioni, mentre P'' è una stima della probabilità dell'errore casuale che si suppone eguale per tutte le variabili.

Se si pone $P'_j = 1$, ciò equivale a imporre la regola deterministica che se il record verifica le condizioni per l'occorrenza dell'errore sistematico, allora la v_i deve essere senz'altro imputata: infatti, la probabilità a posteriori di tutte le configurazioni di errore in cui sia (valore di v_i: vero) risultano nulle, e qualunque scelta effettuata dall'algoritmo di identificazione degli errori prevederà l'inserimento della v_i tra le variabili da modificare.

Se invece $P'_j < 1$, allora non è detto che la v_i debba essere imputata: ciò dipenderà dal valore delle probabilità a posteriori delle configurazioni q_r in cui risulti (valore di v_i: vero): se almeno una di esse risulta maggiore di tutte quelle in cui (valore di v_i: falso), allora la v_i non rientrerà nell'insieme di variabili da imputare.

Riprendiamo in considerazione l'esempio già citato.

Supponiamo che il nostro piano di incompatibilità sia costituito dal solo edit

e_1: (POSPRO ≠ 1) ∩ (LAVORA ≠ *blank*)

e si consideri il record:

[(POSPRO = 2) (LAVORA = 2)]

che attiva tale edit.

Le possibili ipotesi esplicative di tale attivazione sono mostrate nella seguente tabella:

	POSPRO = 2	LAVORA = 2
q_1	falso	vero
q_2	vero	falso
q_3	falso	falso

Sotto l'ipotesi di equidistribuzione delle probabilità di errore, la scelta sarebbe indifferente tra q_1 e q_2:

	POSPRO = 2	LAVORO = 2	$\prod_1^m i \ \Pr(A_i^r)$	$\Pr(q_r \mid p)$
q_1	falso (0.1)	vero (0.9)	0.09	0.47
q_2	vero (0.9)	falso (0.1)	0.09	0.47
q_3	falso (0.1)	falso (0.1)	0.01	0.06
				1.00

Se invece, sulla base di una stima che fissa ad esempio a 0.5 la probabilità che un rispondente, pur dichiarandosi lavoratore indipendente, risponda al quesito successivo, applicando la (4) abbiamo:

$$\Pr(\text{LAVORA} = 2 : \text{ falso}) = \begin{cases} 0.5 & \text{se POSPRO} = 2 \\ 0.1 & \text{altrimenti} \end{cases}$$

e conseguentemente:

	POSPRO = 2	LAVORO = 2	$\prod_1^m i \ \Pr(A_i^r)$	$\Pr(q_r \mid p)$
q_1	falso (0.1)	vero (0.5)	0.05	0.09
q_2	vero (0.9)	falso (0.5)	0.45	0.82
q_3	falso (0.1)	falso (0.5)	0.05	0.09
				1.00

il che porta a scegliere, come la più probabile, la configurazione di errore q_2.

Generalizzando, l'integrazione degli errori sistematici nell'approccio probabilistico avviene nel seguente modo:

1. per ogni errore sistematico, viene fornita la sua "descrizione", e cioè la condizione di occorrenza dell'errore, secondo le stesse modalità con cui viene definita, nell'approccio deterministico, una regola di imputazione deterministica;
2. ad ogni descrizione viene associata una "probabilità di accadimento dell'errore sistematico", stimata mediante analisi dei dati o indagini di controllo (ad esempio, reinterviste con conciliazione). Ad esempio:

   SE (POSPRO ≠ 1) E (LAVORA ≠ *blank*) ALLORA *errore sistematico su* LAVORA *con probabilità* 0.5;

3. per ogni record errato, al momento della identificazione degli errori vengono calcolate le probabilità di errore relative ad ogni variabile, che in alcuni casi saranno date dalla probabilità dell'errore casuale, comune a tutte, in altri dalla probabilità dell'errore sistematico, se le condizioni per la sua occorrenza sono attivate dal record stesso. Sulla base di tali probabilità di errore relative alle singole variabili, viene calcolata la probabilità a posteriori di ogni possibile configurazione di errore secondo la formula di Bayes: l'algoritmo sceglierà la configurazione che ha probabilità massima.

È doveroso rimarcare che si pone un problema di praticabilità della soluzione proposta, legato alla sua *complessità esponenziale* rispetto al numero di variabili potenzialmente errate determinato, in corrispondenza ad ogni unità rilevata, dal numero di edit attivati e dalla quantità di variabili in essi contenuti: in altre parole, se k è il numero di variabili differenti contenute negli edit attivati da un dato record (variabili "sospette"), esistono 2^k diverse configurazioni di errore di cui occorre calcolare la probabilità a posteriori. Ovviamente, per valori elevati di k il lavoro di calcolo e valutazione può risultare proibitivo: la valutazione dell'effettiva applicabilità dell'approccio proposto dipende:

- dalla potenza di calcolo a disposizione;
- dalla distribuzione dei record per numero di variabili "sospette".

Sulla base del primo elemento, è possibile stimare i tempi richiesti per la soluzione del problema di identificazione degli errori in corrispondenza alle diverse numerosità degli insiemi di variabili da valutare. La stima dei tempi permette di definire un numero massimo accettabile di variabili "sospette": se il numero così determinato permette di trattare la totalità o almeno la gran parte dei record, allora l'approccio è praticabile.

6. CONCLUSIONI

La fase di controllo e correzione dei dati rilevati è, in molte indagini statistiche, estremamente importante ai fini del contenimento della componente non campionaria dell'errore. In caso di utilizzo di procedure automatiche, risulta decisivo il tipo di algoritmo utilizzato per la identificazione degli errori in corrispondenza di ogni record per il quale vengano riscontrate delle incompatibilità: adottando l'approccio deterministico, non vi sono garanzie che gli errori presenti nei dati vengano effettivamente rimossi, e che non vengano introdotti nuovi errori. Ricorrendo all'approccio probabilistico nella forma definita dalla metodologia di Fellegi e Holt, in una situazione ideale di assenza di errori sistematici la procedura automatica raggiunge l'obiettivo di una corretta identifi-

cazione ed eliminazione degli errori. Se al contrario sono presenti anche errori di tipo sistematico, l'algoritmo proposto da Fellegi ed Holt non è grado di assicurare la correttezza delle scelte effettuate, tant'è che nella pratica si ricorre ad un preventivo trattamento deterministico di tali errori, con risultati non ottimali.

Il calcolo delle probabilità a posteriori di ogni possibile ipotesi esplicativa delle incompatibilità attivate da un dato record permette di localizzare correttamente gli errori sia che questi siano di natura casuale che di tipo sistematico. Decisive risultano essere, a tal fine, le stime delle probabilità di errore relative ad ogni variabile, in particolare delle probabilità di occorrenza degli errori sistematici. Se tali stime risultano corrette, l'approccio probabilistico è in grado di trattare congiuntamente entrambe le tipologie di errore, ed il dualismo (errore casuale/ approccio probabilistico) vs. (errore sistematico/approccio deterministico) può dirsi superato.

RIFERIMENTI

BARCAROLI G., CECCARELLI C., LUZI O., MANZARI A., RICCINI E., SILVESTRI F. (1995), "The methodology of editing and imputation of qualitative variables implemented in SCIA", *Documento interno ISTAT.*

BARCAROLI G., VENTURI M. (1997), "DAISY (Design, Analysis and Imputation SYstem): structure, methodology and first applications" in *Statistical Data Editing Methods and Techniques – Vol.2 (Statistical Standards and Studies N.48)* - United Nations Conference of European Statisticians.

BARCAROLI G., D'ANGIOLINI G. (1994), "Controllo e correzione inter-record", *Documento interno ISTAT.*

DE WAAL TON (1996), "CHERRYPI: A computer program for automatic edit and imputation", *Research Paper n.9635* – Centraal Bureau voor de Statistiek Netherlands.

FELLEGI I.P., HOLT D. (1976) "A systematic approach to edit and imputation", *Journal of the American Statistical Association,* vol.71, pp.17–35.

GARCIA RUBIO E., VILLAN CRIADO I. (1988) "The DIA System (An Automatic Edit and Imputation System)", Volume I, DIA System Description, INE Madrid.

KOVAR J.G., MAC MILLIAN J.H., WHITRIDGE P. (1991) "Overview and strategy for the Generalised Edit and Imputation System", *Working Paper BSMD–88–007E* Methodology Branch Statistics Canada.

WINKLER W.E., DRAPER L.R. (1997) "The SPEER edit system" in *Statistical Data Editing Methods and Techniques – Vol.2 (Statistical Standards and Studies N.48)* – United Nations Conference of European Statisticians.

WINKLER W.E., PETKUNAS T.F. (1997) "The DISCRETE edit system" in *Statistical Data Editing Methods and Techniques – Vol.2 (Statistical Standards and Studies N.48)* – United Nations Conference of European Statisticians.

THE PROBABILISTIC DATA EDITING: THE JOINT TREATMENT OF SYSTEMATIC AND STOCHASTIC ERRORS BY MEANS OF THE APPLICATION OF BAYES THEOREM TO THE FELLEGI–HOLT METHODOLOGY

Summary

Fellegi–Holt methodology adopts a "probabilistic approach" to the correction of statistical data, in contrast to the "deterministic" : the latter is based on modes of correction that fix a priori the actions of imputation for each different editing rule failed in a given record, while the former considers fundamental to resolve the problem of the error localisation in such a way to achieve the result to modify the variables that most probably are wrong, restoring the "real" values instead of those "false". Fellegi and Holt did not make clear the theoretic bases for which the "principle of minimum change", basic criterion of the algorithm for the error localisation, can guarantee this result. Applying the theorem of Bayes, it is possible to verify that the methodology, under definite hypothesis, allows to determine the variables that are most likely to be wrong. The methodology in question, in its actual version, allows to treat adequately only stochastic errors, not the systematic ones: until now deterministic approach has been used to remove the latter. The formalisation obtained by means of the application of the theorem of Bayes allows, with a simple modification of the algorithm of error localisation, to treat jointly both stochastic and systematic errors.

Keywords: non sampling errors, statistical data editing, Bayes theorem.