ISTITUTO NAZIONALE DI STATISTICA

METODI E SOFTWARE PER IL CONTROLLO E LA CORREZIONE DEI DATI

Giulio Barcaroli, Leandro D'Aurizio, Orietta Luzi, Antonia Manzari, Alessandro Pallara

Ottobre 1998

INDICE

Il lavoro è frutto della collaborazione degli autori. Tuttavia i paragrafi 6.2.1.1, 6.2.1.2, 6.2.1.3, 6.2.1.5, 6.2.2.2, 6.2.2.3, 8, 10 e 11 sono da attribuirsi a Giulio Barcaroli; i paragrafi 6.2.2.4 e 9 sono da attribuirsi a Leandro D'Aurizio; i paragrafi 2, 3, 4, 5, 6.1, 7.1 sono da attribuirsi ad Orietta Luzi; i paragrafi 6.2.2.1, 7.2 e l'Appendice sono da attribuirsi ad Antonia Manzari; i paragrafi 3.1, 6.2.1.4 e 6.2.1.6 sono da attribuirsi ad Alessandro Pallara.

1. Introduzione

La fase di *controllo e correzione dei dati* consiste nell'individuazione e nel trattamento degli *errori* (in senso generale) presenti nei dati raccolti mediante una certa indagine, allo scopo di garantire risultati finali con determinati livelli di qualità.

Gli *errori* presenti in un insieme di dati possono essere dovuti ad una qualunque delle fasi di acquisizione e messa a punto delle informazioni (raccolta, revisione, codifica, registrazione). Per questo motivo, mentre tradizionalmente il processo di controllo e correzione avveniva in un momento successivo alla fase di registrazione dei dati, la tendenza attuale è quella di spostare via via il controllo dei dati il più possibile vicino alla fase di raccolta delle informazioni presso le unità, in modo da rendere più agevole il reperimento di informazioni corrette laddove si verifichino situazioni non compatibili o anomale. Sono state quindi sviluppate tecnologie per l'integrazione del controllo e correzione dei dati con le fasi di intervista o di registrazione, in modo da eliminare o in ogni caso minimizzare la parte di errori attribuibile ad errori di compilazione o registrazione dei modelli (che rappresentano generalmente la parte più consistente del totale degli errori). Alcune tipologie di errori (di codifica, di percorso, di dominio, ecc.) vengono corretti contestualmente alla fase di intervista o di registrazione, producendo una migliore qualità finale dei dati ed un risparmio nei tempi e nei costi connessi alle fasi successive di controllo (interattivo o automatico) dei dati.

Il processo di verifica della qualità dei dati si è sviluppato originariamente come insieme di attività di tipo manuale-interattivo, in cui cioè entrambe le operazioni di individuazione e di correzione degli errori erano effettuate manualmente da esperti mediante revisione dei modelli, reintervista, uso di informazioni ausiliarie o di conoscenze soggettive sul fenomeno investigato.

Attualmente tale processo è basato, in tutto o in parte, sull'uso del calcolatore: l'utilizzo di procedure informatiche in fase di individuazione degli errori e/o in fase di correzione ha prodotto da un lato una notevole riduzione dei tempi e dei costi connessi con questa fase del processo d'indagine, dall'altro un generale aumento della qualità dei risultati finali.

Fra i principali problemi connessi all'attività di verifica di dati, infatti, vanno considerati i costi ed i tempi necessari per mettere a punto gli strumenti idonei e per portare a termine le operazioni connesse al controllo e correzione dei dati. Numerosi studi hanno dimostrato che i costi (in termini sia di risorse umane sia di budget impiegato) connessi con questa fase rappresentano percentuali molto elevate del costo totale dell'intero processo di produzione dell'informazione statistica e che, in generale, il tempo speso nell'effettuazione delle operazioni di controllo e correzione può assorbire una quota eccessiva del tempo totale disponibile (cosicché questa risorsa può risultare insufficiente per l'approfondimento di altri aspetti del processo di produzione dell'informazione statistica).

L'informatizzazione del processo di verifica della qualità dei dati ha portato contemporaneamente ad una maggiore trasparenza, ad un maggior controllo e generalmente ad un miglior livello di documentazione del processo di controllo e correzione, e ad un trattamento più oggettivo ed omogeneo delle situazioni di errore.

Tra i problemi connessi all'adozione di sistemi automatici, va sottolineata una maggiore rigidità dei controlli rispetto a certe tipologie di errori (ad esempio gli errori sistematici e le mancate risposte).

All'informatizzazione del processo di controllo e correzione può essere inoltre attribuito il fenomeno noto come *proliferazione dei controlli*: a causa della diminuzione dei tempi e dei

costi, gli operatori tendono a introdurre nel loro piano di verifica un gran numero di controlli, gran parte dei quali hanno in realtà un impatto trascurabile sulla qualità del risultato finale. A questo fenomeno è collegato il problema noto come *over-editing* (verifica eccessiva), che ha luogo quando vengono effettuate operazioni di verifica che, riguardando errori con un impatto trascurabile sulle stime finali, assorbono tempi e costi non giustificati da miglioramenti significativi nella qualità dei dati finali.

Per questi motivi, nelle attuali procedure di controllo e correzione dati si preferisce talvolta adottare un approccio di tipo *misto*, in cui le modalità di controllo e correzione interattiva ed automatica vengono impiegate in combinazione fra loro, in relazione alle caratteristiche del processo di produzione dell'informazione statistica, alla disponibilità di risorse (umane, di tempo, hardware e software), alle tipologie di errore ed alle esigenze qualitative. Anche in questa ottica, è necessario che le tecniche e le metodologie adottate nei due approcci (interattivo ed automatico) siano guidate da criteri rigorosi ed adottino metodologie ottimali in termini sia di efficacia (qualità dei risultati) che di efficienza (tempi, costi, carico sui rispondenti).

La situazione ottimale è quella in cui alla parte automatizzata del processo di controllo e correzione sono affidate le operazioni di individuazione degli errori e di correzione dei casi con minore impatto sulle stime finali dei dati, mentre la fase di controllo interattivo è utilizzata per la selezione e la correzione degli errori con maggior influenza sulle stime finali, al fine di eliminare da un lato le gravi distorsioni che tali errori possono provocare nei risultati, e di ridurre dall'altro i tempi e l'impiego di risorse connessi alla fase di controllo interattivo dei dati.

In conclusione, la tendenza attuale dei ricercatori e degli operatori nei vari enti produttori di statistiche è di lavorare su vari fronti, nell'ottica di una maggiore razionalizzazione del processo di controllo e correzione dai dati e di un aumento della qualità delle informazioni statistiche prodotte. Possiamo riassumere tale strategia nei seguenti punti principali:

- accuratezza nel disegno di procedure il più possibile calibrate sulla situazione da sottoporre a controllo;
- monitoraggio delle prestazioni delle procedure adottate, prevedendo attività di valutazione e documentazione delle stesse;
- adozione di strumenti per l'individuazione delle cause strutturali che, agendo sistematicamente sui dati, provocano percentuali significative di errori, mancate risposte, ecc.; intervento sull'organizzazione, sui rilevatori, sul modello ecc., al fine di rimuovere tali cause o ridurne il più possibile gli effetti;
- miglioramento delle metodologie e/o delle tecnologie utilizzate, per le parti sia manuale/interattiva, sia automatica della procedura complessiva di controllo e correzione;
- razionalizzazione del piano dei controlli di qualità, attraverso un'analisi approfondita dei risultati della loro applicazione; adozione di controlli mirati alla individuazione di specifiche tipologie di errori particolarmente frequenti o con significativo impatto sui risultati finali;
- ottimizzazione dell'impiego delle risorse disponibili, riducendo tempi, costi, carico sui rispondenti e sui rilevatori, soprattutto per la parte del controllo interattivo.

2. Errori non campionari: incompatibilità, valori anomali, mancate risposte parziali e mancate risposte totali

In generale, diciamo che una certa variabile rilevata in una data unità statistica è affetta da errore quando il suo valore non corrisponde al valore vero che essa presenta in quella unità. E' chiaro che la presenza di errori, di qualunque natura essi siano, può provocare distorsioni nelle distribuzioni delle variabili investigate, nelle stime finali dei dati (totali, medie, ecc.), e in tutte le analisi statistiche effettuate sui dati non corretti.

Gli errori da cui possono essere affette variabili rilevate statisticamente o per via amministrativa possono essere classificati secondo diversi criteri.

Distinguiamo innanzi tutto fra *distorsioni* ed *errori variabili* (Masselli, Panizon, Signore, 1992): nell'ipotesi che il processo di produzione dell'informazione statistica sia ripetibile nelle medesime condizioni, si assume che:

1. gli *errori variabili* siano casuali, e varino in ogni ripetizione del processo di produzione dell'informazione statistica;
2. le *distorsioni* sono il risultato di fattori sistematici, dipendono dalle condizioni in cui è effettuato il processo di produzione dell'informazione statistica, sono costanti in tutte le ripetizioni ed hanno segno specifico rispetto al valore vero.

Un'altra distinzione fra errori è basata sul livello (*microdati* o *macrodati*) a cui essi si verificano:

1. si dicono *non campionari* o *di misura* quegli errori che interessano direttamente i dati elementari: la differenza fra valore osservato y_i della variabile rilevata Y nella i-esima unità e valore vero Y_i è attribuibile a problemi nell'organizzazione del processo di produzione dell'informazione statistica (registri, preparazione dei rilevatori, controllo del territorio, ecc.), all'intervistato (che rifiuta di rispondere, fornisce un dato errato volontariamente o involontariamente, per errata comprensione della domanda, ecc.), all'intervistatore (carenza nell'addestramento, influenza dell'intervistatore, ecc.), alla tecnica di intervista (faccia a faccia, postale, telefonica, ecc.), alle caratteristiche del modello (lunghezza, complessità, terminologia, ecc.), a problemi nelle fasi di codifica e registrazione dati. Questi errori interessano le stime finali attraverso le operazioni di aggregazione dei dati elementari (totali, medie, frequenze relative ed assolute ecc.);
2. si dicono *campionari* quegli errori che dipendono sostanzialmente dalla circostanza che non tutta la popolazione, ma solo una porzione di essa (il campione) è soggetto a rilevazione (errori variabili di campionamento). Questi errori possono pertanto essere attribuiti esclusivamente all'effetto del caso, al disegno campionario, alla tecnica di campionamento o allo stimatore utilizzati in una data indagine statistica, e interessano solo le stime.

L'errore a livello di stima fra valore ottenuto mediante la rilevazione e valore vero è pertanto scomponibile in tre parti principali:

errore statistico=errore non campionario+

errore variabile campionario+

distorsione dello stimatore

Indicati con y_i e Y_i , rispettivamente, i valori osservati ed i valori veri della variabile Y, sia $f(y_1, y_2,..., y_n)$ lo stimatore del parametro di interesse $g(Y_1, Y_2,..., Y_N)$: l'errore che si

commette, cioè la discrepanza fra la stima ottenuta mediante la rilevazione ed il valore vero nella popolazione, può essere scomposto come segue:

$$f(y_1, y_2,..., y_n) - g(Y_1, Y_2,..., Y_N) = f(y_1, y_2,..., y_n) - f(Y_1, Y_2,..., Y_n) +$$
$$f(Y_1, Y_2,..., Y_n) - E[f(Y_1, Y_2,..., Y_n)] +$$
$$E[f(Y_1, Y_2,..., Y_n)] - g(Y_1, Y_2,..., Y_N)$$

Indicati con $y = f(y_1, y_2,..., y_n)$, $y^* = f(Y_1, Y_2,..., Y_n)$, $Y = g(Y_1, Y_2,..., Y_N)$, la relazione precedente può essere riscritta in modo più compatto come segue:

$$y - Y = [y-y^*] + [y^* - E(y^*)] + [E(y^*) - Y] =$$
$$= [v + b] + [y^* - E(y^*)] + [E(y^*) - Y]$$

dove v e b rappresentano, rispettivamente, le componenti dell'errore non campionario [y-y*] che si manifestano come errore *variabile* o come *distorsione*.

Pertanto l'errore totale di uno stimatore $y=f(y_1, y_2,..., y_n)$, calcolato generalmente mediante l'errore quadratico medio MSE, può essere scomposto nel modo seguente:

$$MSE = E(y - Y)^2 = E[v]^2 + [y^* - E(y^*)]^2 + [B + D]^2 + 2\,cov(v, y^*) =$$
$$= VNC + VC + (B+D)^2 + 2\,cov(v, y^*)$$

dove: VNC = varianza campionaria dell'errore variabile non campionario;
VC = varianza campionaria dell'errore variabile campionario;
B = distorsione non campionaria;
D = distorsione dello stimatore;
cov(v, y*) = covarianza tra l'errore variabile non campionario e la stima.

Si fa osservare che, mentre le distorsioni non campionarie B possono essere dovute sia ai rispondenti, (valori rilevati ma sistematicamente errati) sia ai non rispondenti (distorsioni indotte dalle mancate risposte parziali o totali), la distorsione D può essere dovuta all'uso di stimatori non corretti ma consistenti (annullate nel caso in cui la dimensione del campione è sufficientemente elevata o nel caso dei censimenti) oppure di stimatori non corretti e non consistenti (che permangono anche nel caso di indagini totali).

Mentre esiste ed è consolidata la teoria di valutazione e misura degli errori di tipo campionario, il trattamento della componente non campionaria dell'errore è reso più complesso dalle difficoltà connesse sia alla sua individuazione, sia alla determinazione e rimozione delle cause che l'hanno generata. In particolare, non esistono attualmente modelli esplicativi generali rigorosi dei fenomeni all'origine degli errori non campionari.

Il processo di controllo e correzione dei dati effettuato in fase di editing riguarda i soli errori *non campionari* presenti nei dati stessi. A questo tipo di errori faremo quindi riferimento nel seguito della trattazione.

Un secondo criterio di classificazione degli errori distingue fra errori *sistematici* ed errori *casuali* (o *stocastici* o *non sistematici*):

Si dicono *sistematici* quegli errori la cui origine è da attribuirsi a difetti strutturali o organizzativi del processo di produzione dell'informazione statistica, alla struttura del modello o al sistema di registrazione adottati, errori che si manifestano come deviazioni sistematiche dal valore vero di una o più variabili rilevate. La loro presenza può essere

segnalata da particolari frequenze di valori anomali, incongruenze o valori fuori dominio nelle variabili rilevate.

Si dicono *casuali* o *stocastici* quegli errori la cui origine è da attribuirsi a fattori aleatori: per questo tipo di errori è quindi ipotizzabile una distribuzione nella popolazione di riferimento di tipo normale a media nulla, nel caso di variabili quantitative, ed una distribuzione uniforme delle modalità errate nel caso di variabili qualitative.

Il trattamento delle due tipologie di errore appena descritte deve essere effettuato utilizzando procedure e tecniche sostanzialmente diverse (come vedremo meglio trattando le procedure di controllo e correzione), al fine di garantirne la corretta individuazione e la predisposizione degli opportuni interventi correttivi.

Gli errori da cui può essere affetto un insieme di dati possono essere ancora distinti in *mancate risposte totali* e *mancate risposte parziali*.

Si ha una *mancata risposta totale* (MRT nel seguito) quando una certa unità statistica inclusa nella rilevazione non fornisce risposta ad alcuno dei K quesiti previsti nel modello. La presenza di MRT può essere dovuta a varie cause: errore di lista, non reperibilità dell'unità statistica inclusa nella rilevazione, rifiuto di rispondere, incapacità di rispondere, ecc.

Le *mancate risposte parziali* (MRP nel seguito) si verificano quando per una certa unità statistica inclusa nella rilevazione non è disponibile l'informazione relativa ad un sottoinsieme k dei K quesiti previsti nel modello.

In generale, dato un vettore $(X_{i1}, X_{i2},...,X_{iK})$ contenente i valori assunti dalle K variabili (quesiti) rilevate in corrispondenza dell'i-esimo rispondente, si può parlare di mancata risposta parziale ogni qualvolta si verifica uno dei seguenti eventi:

- sulla base dei valori assunti dalle altre variabili nel vettore, è richiesta la presenza di un valore significativo in X_{ij} ed invece si riscontra un *valore mancante*;
- in X_{ij} il valore presente non corrisponde a quello assunto nella realtà dal carattere j nella i-esima unità.

La prima componente della MRP (*valori mancanti*) è dovuta prevalentemente a problemi in fase di *compilazione* del modello, consistenti o in una cattiva interpretazione dei quesiti o delle regole di compilazione da parte del rispondente e/o del rilevatore, oppure nel rifiuto da parte del rispondente. L'individuazione di questa componente richiede la verifica dell'accettabilità o meno di un valore mancante per una data variabile del modello condizionatamente alle risposte fornite ad uno o più particolari quesiti precedenti (detti di "svincolo" o di "controllo"), che possono essere a loro volta soggetti ad errore.

La seconda componente delle MRP (*valori errati*), oltre che ai problemi di compilazione già citati, risente in modo particolare di problemi in fase di *registrazione*. La presenza dei valori errati può dar luogo, e può essere segnalata da *valori fuori dominio*, *valori anomali*, *incompatibilità fra risposte* nello stesso modello:

- una certa variabile X_j presenta un valore *fuori dominio* quando essa presenta un valore esterno all'intervallo di valori per essa ammissibili. In questo caso, si può essere certi della presenza di un errore solo nel caso in cui X_{ij} assume un valore esterno al dominio di definizione della variabile X_j. Questo errore è tipico delle variabili qualitative;
- si ha un v*alore anomalo (outlier)* rispetto ad una certa variabile X_j quando una certa unità statistica inclusa nella rilevazione fornisce per tale variabile una risposta il cui valore si discosta in modo significativo dai valori che la stessa variabile assume nel resto delle unità campionarie. L'individuazione di questo tipo di errore implica un'analisi della distribuzione della variabile X_j nell'intera popolazione dei rispondenti. Questo errore è tipico delle variabili quantitative;

- si dice che in una unità rispondente sono presenti *incompatibilità* se i valori di una o più variabili in essa rilevate sono in contraddizione fra loro (in termini logici per le variabili qualitative, matematico-statistici per quelle quantitative), o con i corrispondenti valori in precedenti occasioni di rilevazione. Accertata la presenza di una o più incompatibilità in un record, è necessario individuare le variabili errate: ciò equivale a individuare un sottoinsieme di elementi del vettore $(X_{i1}, X_{i2},...,X_{iK})$ che, sulla base delle incompatibilità rilevate, sono *più probabilmente* affette da errore.

La distinzione fra MRT ed MRP è dovuta talvolta a considerazioni di tipo *soggettivo*, nel senso che dipende da una "soglia di accettabilità" per i modelli fissata volta per volta a seconda del tipo di rilevazione. Tale soglia viene generalmente stabilita sulla base del contenuto informativo dei modelli in rapporto agli obiettivi conoscitivi dell'indagine: per questo motivo, ad esempio, un modello in cui non siano riportati i valori di variabili strategiche per l'indagine può essere considerato una MRT.

Sia le MRT che le MRP possono essere di natura sistematica o stocastica: la comprensione delle cause alla loro origine è quindi di importanza fondamentale per l'adozione del metodo ottimale per il loro trattamento e, laddove possibile, per la loro prevenzione nelle successive occasioni di indagine.

Ciò che, in ultima analisi, distingue le MRP dalle MRT è che mentre le prime necessitano di una fase di individuazione, tale fase non è invece necessaria per le seconde. Per entrambe è invece necessaria una fase di analisi statistica allo scopo di:

1. valutare e documentare l'entità del fenomeno;
2. individuare le cause (strutturali, organizzative, psicologiche, metodologiche, ecc.) che lo hanno prodotto;
3. porre in atto le tecniche e le metodologie più appropriate per la prevenzione e/o il recupero delle situazioni di errore (miglioramento della preparazione degli intervistatori, del modello, dell'organizzazione, della tecnica di editing, ecc.).

Un aspetto importante del problema dell'individuazione e della correzione degli errori riguarda la necessità, sempre più sentita in termini non solo statistici ma anche e soprattutto di contenuto delle informazioni prodotte e rilasciate all'utente, di produrre dati non solo completi (cioè privi di mancate risposte e incongruenze interne), ma anche e soprattutto il più possibile corrispondenti al vero. In questo senso la ricerca e la correzione degli elementi errati vanno viste come operazioni attraverso cui, a fronte di una situazione di incertezza, vengono poste in atto tecniche di recupero e di ripristino dell'informazione "vera".

Nel paragrafo che segue vengono approfondite alcune problematiche connesse alle mancate risposte (parziali e totali) in dati statistici.

3. Le mancate risposte nelle indagini statistiche

L'obiettivo di un'indagine è di analizzare le unità campionarie rispetto a q variabili di studio y_1**,……,** y_j**,……,** y_q (q componenti di un questionario, ad es.). Sia y_{jk} il valore della variabile y_j per la unità k. Si denoti inoltre con n_s la dimensione di s, in cui s di solito costituisce un campione casuale ma può anche riferirsi all'universo U, nel caso di un'indagine censuaria.

In caso di risposta completa all'indagine si ha che, dopo la fase di raccolta e di controllo, i dati costituiscono, per ogni $k \in s$, un vettore q-dimensionale di valori osservati

$$\mathbf{y}_k = (y_{1k},....,y_{jk},....y_{qk});$$

questi dati formano una matrice di dimensione $n_s \times q$, con nessun valore mancante. Se la matrice non è piena, se cioè qualche y_{jk} ha un valore mancante, allora vi saranno mancate risposte (MR). Le MR di solito introducono un problema di distorsione nelle stime calcolate dai dati dell'indagine. Si denoti con r_j il j-esimo insieme di risposte, cioè il sottoinsieme di s per il quale siano state registrate risposte accettabili all'elemento j-esimo del questionario, vale a dire

$$r_j = \{k : k \in s \text{ e } y_{jk} \text{ è registrato}\}.$$

In una indagine vi sono q insiemi di risposte r_1, ..., r_q di solito non identici. Nel caso di un'indagine campionaria, l'insieme s è selezionato a partire da un disegno campionario noto e r_j è un sottoinsieme di s. La presenza di MR introduce una serie di difficoltà nel trattamento dei dati di un'indagine:

- il meccanismo di generazione di r_j, cioè $p_j(\cdot|s)$ non è noto, dove $p_j(\cdot|s)$ denota la probabilità di ogni insieme r_j di risposte, dato il campione s selezionato: la validità delle stime relative alla popolazione dipendono dalla validità delle assunzioni sul meccanismo di risposta;
- la stima $\hat{\theta}$ di un parametro θ e la stima della varianza dello stimatore $\hat{\sigma}^2(\hat{\theta})$ deve tenere conto della presenza di MR;
- poiché i q insiemi di risposte r_1, ..., r_q sono differenti, nella fase di stima risulta complesso il trattamento congiunto delle variabili di studio.

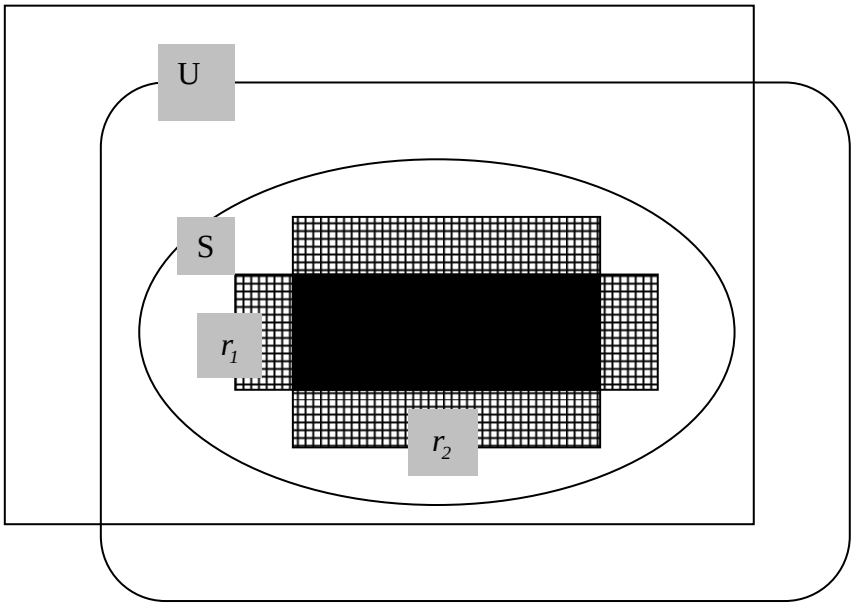
Non esiste un metodo che in assoluto consenta inferenze statistiche valide in presenza di MR. Tutti i metodi utilizzati in tale situazione costituiscono dei tentativi più o meno efficaci per ridurre la distorsione dovuta alle MR. E' difficile ottenere una misura precisa di tale distorsione, più facile è tentare di misurare l'estensione della non risposta. A questo riguardo si denoti con

$$r_u = r_1 \cup r_2 \cup ..... \cup r_q = \bigcup_{j=1}^{q} r_j$$

l'insieme di unità che hanno risposto ad almeno una componente del questionario. Il complemento di r_u, vale a dire s-r_u, costituisce l'insieme delle mancate risposte complete o totali (MRT). Si denoti poi con

$$r_c = r_1 \cap r_2 \cap ..... \cap r_q = \bigcap_{j=1}^{q} r_j$$

l'insieme di unità che hanno risposto a tutte le voci, per cui per ogni $k \in r_c$ il vettore $\boldsymbol{y}_k$ è completo. Allora r_u-r_c costituisce l'insieme delle unità che hanno risposto ad almeno una ma non a tutte le componenti, vale a dire le <u>mancate risposte parziali</u> (MRP). In particolare, r_u-r_j costituisce l'insieme delle MRP relative alla componente *j*. Un'illustrazione di quanto precede, per il caso *q*=2, è presentato nella figura che segue (cfr. Särndal, 1992).



Nel grafico r_1 ed r_2 rappresentano l'insieme di risposte alle variabili di studio y_1 e y_2, rispettivamente. L'area in nero costituisce l'insieme r_c mentre l'insieme r_u, è costituito da r_c e l'area ombreggiata dei due insiemi r_1 ed r_2.
<u>Caso speciale</u>: $r_1 = r_2 = \ldots\ldots = r_q = r$, cioè la non risposta influenza le variabili di studio alla stessa maniera; in tal caso, se $r \subset s$ vi sono MRT ma non MRP.

Si denoti con n_s,n_{ru},n_{rj} ed n_{rc} la numerosità di *s*, r_u, r_j e r_c, rispettivamente. Allora una misura semplice della risposta e della mancata risposta totale è rappresentata rispettivamente da:

$$p_{ru} = n_{ru} / n_s \qquad \text{e} \qquad 1 - p_{ru}$$

mentre una misura che tenga conto dei pesi campionari è costituita da:

$$\tilde{p}_{ru} = (\sum_{ru} 1/\pi_k) / (\sum_s 1/\pi_k) \qquad \text{e} \qquad 1 - \tilde{p}_{ru}$$

in cui π_k rappresenta la probabilità di inclusione per l'unità *k*. Similmente, una misura della risposta e della mancata risposta parziale è data da:

$$p_{rc} = n_{rc} / n_{ru} \qquad \text{e} \qquad 1 - p_{rc} \qquad \text{MRP totali}$$

$$p_{rj} = n_{rj} / n_{ru} \qquad \text{e} \qquad 1 - p_{rj} \qquad \text{componente } j\text{-esima.}$$

Una naturale estensione di queste espressioni consente poi di definire una misura della risposta e della mancata risposta parziale che incorpora i pesi campionari.

Il trattamento delle mancate risposte implica la messa a punto di un'insieme di strategie di intervento nelle diverse fasi di un'indagine, tendenti prima di tutto a ridurre il loro numero

oltre che ad ottenere stime che tengano conto dell'impatto che le MR possono avere sulla validità dei risultati. Tali interventi comprendono il ricorso a particolari tecniche di rilevazione e diverse forme di sollecito durante la fase di raccolta dei dati oppure l'adozione di specifiche tecniche campionarie per ottenere informazioni sui non rispondenti, ad es. procedendo ad una seconda fase di campionamento che restringa l'attenzione sulle unità campionarie che non hanno risposto alla prima fase dell'indagine, allo scopo di cercare di correggere le stime per la distorsione dovuta alle MR. In ogni caso è pressoché inevitabile che al termine della fase di raccolta e registrazione dei dati si sia in presenza di una certa percentuale di MR. A questo punto occorre introdurre esplicite assunzioni sul meccanismo di generazione delle risposte nella forma di un modello, utilizzato nella costruzione di stime che incorporino una componente legata alle MR. Nella maggior parte dei casi tali inferenze fanno uso di informazioni ausiliarie. I metodi per il trattamento delle MR si distinguono essenzialmente tra le tecniche di riponderazione, mediante le quali i pesi campionari associati alle unità rispondenti vengono modificati per tenere conto delle MR e l'imputazione dei valori mancanti mediante valori artificiali opportunamente selezionati. Di solito si ricorre alla riponderazione in presenza di MRT (vedi paragrafo 5) mentre per le MRP si preferisce procedere all'imputazione dei valori mancanti.

3.1. Un modello di risposta

Nel seguito, per semplicità si restringe l'attenzione al caso di una sola variabile di studio, per cui la distinzione fra MRP e MRT si perde. Si voglia stimare il totale della popolazione $t = \sum_U y_k$. A tale scopo viene selezionato un campione di dimensione n_s dalla popolazione (finita) $U = \{1,.....,k,.....N\}$ in base ad un disegno $p(s)$, con probabilità di inclusione π_k e π_{kl}, in cui con π_{kl} si denota la probabilità della contemporanea inclusione nel campione delle unità k ed l. Dopo la fase di raccolta, nel campione selezionato risultano MR: si denoti con r l'insieme di risposte, di dimensione $m_r<n_s$. Sul meccanismo di generazione di tali risposte si assume che:

$$Pr(k \in r|s) = \theta_k = \theta \qquad \text{probabilità di risposta costante}$$

$$Pr(k \cap l \in r|s) = \theta_k\theta_l = \theta^2 \qquad \text{indipendenza tra le risposte}$$

per ogni k ed $l \in s$ e per ogni s, con $\theta>0$ (modello di risposta per dati MAR - *missing at random*, Rubin, 1976). Una stima non distorta di t in presenza di risposta completa (cfr. Sarndal *et al.*, 1992) è costituita da:

$$\hat{t} = N\tilde{y}_s = N\frac{\sum_s y_k / \pi_k}{\sum_s 1/\pi_k}$$

in cui $\tilde{y}_s$ denota la stima della media della popolazione in cui si tenga conto dei pesi campionari. In presenza di MR una stima di t, con il modello di risposta ipotizzato, è data da:

$$\hat{t}_1 = N\tilde{y}_r = N\frac{\sum_r y_k / \pi_k\theta}{\sum_r 1/\pi_k\theta} = N\frac{\sum_r y_k / \pi_k}{\sum_r 1/\pi_k}$$

se il modello MAR è vero, $\hat{t}_1$ è una stima (sostanzialmente) non distorta di t. Come si vede, $\hat{t}_1$ differisce da t solo per il fatto che si basa sull'insieme di risposte r e non sull'intero

campione *s*. Il modello di risposta ipotizzato appare però poco realistico ed è utile valutare la distorsione di $\hat{t}_1$ con differenti distribuzioni delle risposte.

i. $Pr(k \in r|s) = \theta_k$ probabilità di risposta variabile
$Pr(k \cap l \in r|s) = \theta_k \theta_l$ indipendenza tra le risposte;

in questo caso si può mostrare (cfr. Sarndal *et al*., 1992, pag.577) che la distorsione (relativa) di $\hat{t}_1$ è tanto maggiore quanto maggiore è la correlazione nella popolazione tra la variabile di studio *y* e la probabilità di risposta θ, in altre parole quanto più ci si allontana da un meccanismo di (mancata) risposta ignorabile verso uno non ignorabile (Little e Rubin, 1987);

ii. distribuzione deterministica delle risposte - nella popolazione vi sono due gruppi, U_1 e U_2 di dimensione N_1 e N_2, rispettivamente: le unità in U_1 rispondono con probabilità 1, se selezionate, mentre per le unità in U_2 la probabilità di risposta è 0. Siano $\bar{y}_{U1}$ e $\bar{y}_{U2}$ le medie nei due gruppi. Si dimostra che la distorsione di $\hat{t}_1$ è

$$B(\hat{t}_1) = E(\hat{t}_1) - t \cong N_2(\bar{y}_{U1} - \bar{y}_{U2})$$

la distorsione aumenta all'aumentare della differenza tra le medie dei due strati e con la dimensione del gruppo dei non rispondenti.

iii. modello di non risposta MAR all'interno di sottogruppi omogenei: il campione *s* suddiviso in H_s gruppi omogenei s_h ($h=1, ..., H_s$). Per ogni *s* e per $h=1, ..., H_s$,

$$Pr(k \in r|s) = \pi_{k|s} = \theta_{hs} > 0 \qquad \text{per ogni } k \in s_h$$

$$Pr(k \cap l \in r|s) = \pi_{kl|s} = Pr(k \in r|s)\,Pr(l \in r|s) \qquad \text{per ogni } k \neq l \in s$$

dato *s*, tutte le unità in s_h hanno la stessa probabilità di risposta e differenti gruppi hanno diverse probabilità di risposta e le risposte sono indipendenti. Tale modello può portare ad una significativa riduzione della distorsione dovuta alla non risposta rispetto al modello di partenza che ipotizzava una probabilità di risposta costante su tutta la popolazione, soprattutto se la scelta dei gruppi è fatta in maniera efficace.
Si denoti con r_h il sottoinsieme di risposte nel gruppo s_h e con n_h e m_h la dimensione di s_h e r_h, rispettivamente. Un'espressione generale per uno stimatore rapporto basato su classi di aggiustamento per MR che utilizza variabili ausiliarie è definita da (cfr. Sarndal *et al*., 1992, pag.585):

$$\hat{t}_2 = \frac{N}{n}\left(\sum_s x_k\right)\frac{\sum_{h=1}^{H_s} n_h \bar{y}_{rh}}{\sum_{h=1}^{H_s} n_h \bar{x}_{rh}}$$

in cui $\bar{y}_{rh}$ e $\bar{x}_{rh}$ rappresentano le medie tra i rispondenti nella classe *h* della variabile di studio e della variabile ausiliaria, rispettivamente. Tale stimatore è ottenuto postulando l'esistenza di una relazione tra variabile di studio e variabile ausiliaria descritta da un modello di regressione e assumendo noti i valori x_k per rispondenti e non rispondenti. Sulla base del tipo di informazione ausiliaria utilizzata (o, più precisamente, disponibile) possono essere ottenute forme particolari dello stimatore $\hat{t}_2$ che costituiscono casi speciali noti:

a) la sola informazione ausiliaria utilizzata è l'appartenenza di ciascuna unità *k* ad uno degli *H* strati ($h=1, ..., H$) in cui è suddivisa la popolazione e di cui è nota la

numerosità (N_1, ..., N_H) e per la definizione dei gruppi omogenei in cui suddividere il campione *s* si utilizzano gli strati stessi della popolazione. In tal caso si può dimostrare che $\hat{t}_2$ diviene

$$\hat{t}_{2a} = \sum_{h=1}^{H} N_h \bar{y}_{rh}$$

che è lo stimatore poststratificato di *t*;

b) i gruppi sono definiti a partire dal campione realizzato, perciò possono variare da un campione all'altro e ogni unità può appartenere a gruppi differenti con due differenti campioni: in tal caso, dato *s*, l'aggiustamento per le MR viene realizzato sulla base dei gruppi costruiti e lo stimatore $\hat{t}_{2a}$ diviene

$$\hat{t}_{2b} = \frac{N}{n} \sum_{h=1}^{H_s} n_h \bar{y}_{rh}$$

c) una sintesi dei due stimatori precedenti può essere ottenuta combinando il modello di (non) risposta con una stratificazione del campione *s* in G_s sottogruppi s_1, ..., s_g, ..., s_{G_s}, tali che i valori y_k delle unità nel medesimo gruppo siano caratterizzati da una minima variazione intorno alla media del gruppo. I sottogruppi s_g possono essere visti come strati, ciascuno di numerosità n_g, in ognuno dei quali può essere adattato un modello ANOVA tale che

$$E(y_k) = \beta_g \qquad V(y_k) = \sigma^2$$

in cui $E(\cdot)$ e $V(\cdot)$ denotano, rispettivamente, valore atteso e varianza rispetto al modello, con parametri β_g e σ^2 . Con il modello di risposta ipotizzato, il campione *s* è caratterizzato dall'incrocio di $H_s G_s$ celle. Si denoti con s_{gh} il sottoinsieme di *s*, di dimensione n_{gh}, che ricade nella cella *gh* e con r_{gh} i rispondenti in s_{gh}, di numerosità m_{gh}. Il tasso di risposta nel gruppo *h* è definito da $f_h = m_{\cdot h}/n_{\cdot h}$, con $n_{\cdot h} = \sum_{g=1}^{G_s} n_{gh}$ e $m_{\cdot h} = \sum_{g=1}^{G_s} m_{gh}$. In questo caso lo stimatore di *t* è lo stimatore di regressione:

$$\hat{t}_{2c} = \frac{N}{n} \sum_{g=1}^{G_s} n_{g\cdot} \hat{\beta}_{gr}$$

in cui $n_{g\cdot} = \sum_{h=1}^{H_s} n_{gh}$ e

$$\hat{\beta}_{gr} = \left(\sum_{h=1}^{H_s} \frac{1}{f_h} \sum_{r_{gh}} y_k \right) \Bigg/ \sum_{h=1}^{H_s} \frac{m_{gh}}{f_h}$$

É abbastanza intuitivo ricondurre lo stimatore $\hat{t}_{2c}$ allo stimatore poststratificato $\hat{t}_{2a}$, se si considera $\hat{\beta}_{gr}$ come la stima della media del *g*-esimo strato, con correzione per le MR, e $\hat{N}_g = N n_{g\cdot}/n$ come stima della dimensione dello strato.

Le proprietà degli stimatori presentati in precedenza sono discusse diffusamente in Sarndal *et al*. (1992, cap.15). Rispetto ai semplici modelli di risposta ipotizzati in i. e ii., ciascuno degli stimatori discussi in iii. conduce ad una riduzione della varianza e risulta tanto più robusto di fronte ad errori di specificazione del modello quanto maggiore è l'informazione ausiliaria utilizzata e quanto più forte è la relazione tra variabili ausiliarie e variabile di studio. Un aspetto importante in tal senso può essere costituito dal criterio

per la costruzione dei gruppi omogenei, all'interno dei quali ipotizzare una probabilità di risposta costante. Proposte su questo specifico problema sono contenute in Rosenbaum e Rubin (1983), Little (1986). Occorre comunque sottolineare che tali stimatori sono stati ottenuti ipotizzando che la mancata risposta sia di tipo ignorabile. La situazione cambia quando le MR sono di tipo non ignorabile, quando cioè il meccanismo di risposta dipende dalla variabile di studio. In tal caso, gli stimatori discussi in precedenza possono condurre a risultati fortemente distorti. Il trattamento dei dati incompleti con MR non ignorabile è un'area di ricerca relativamente poco esplorata e che richiede una specifica attenzione per il fatto che risulta difficile in tali casi ottenere uno stimatore soddisfacente sulla base del disegno campionario. Alcune soluzioni sono presentate in Heckman (1976), Greenlees *et al*. (1982) e Little (1993).

4. Individuazione e correzione delle mancate risposte parziali (MRP)

Poiché varie e di diversa origine sono le componenti delle MRP, diverse saranno le tecniche utilizzate per la loro localizzazione e, eventualmente, per la loro correzione. In particolare, la fase di localizzazione non è richiesta per i *valori mancanti,* mentre può risultare molto complessa per le altre tipologie di errore (*incompatibilità* e *valori anomali*). Nel seguito analizzeremo separatamente le due fasi di *individuazione* e di *integrazione* delle MRP e, per ognuna di esse, considereremo le diverse tecniche utilizzabili separatamente per ciascuna tipologia di errore.

4.1. Individuazione

La localizzazione delle risposte errate in un certo insieme di dati statistici è basata su diversi tipi di controlli (o *regole* o *edit*), che possono essere classificati in tre categorie principali:

1. *controlli di consistenza:* verificano che prefissate combinazioni di valori assunti da variabili rilevate in una stessa unità soddisfino certi requisiti (*regole di incompatibilità*).
2. *controlli di validità o di range*: verificano che i valori assunti da una data variabile siano interni all'intervallo di definizione della variabile stessa.
3. *controlli statistici:* utilizzati al fine di isolare quelle unità statistiche che presentano, per alcune delle variabili in esse contenute, valori che si discostano in modo significativo dai valori che le stesse variabili assumono nel resto delle unità campionarie o rispetto ad una rilevazione precedente. Questi valori sono con alta probabilità errati, ma l'asserzione della loro non correttezza necessita di ulteriori e approfondite verifiche.

Vediamo come possono essere strutturati tali controlli per l'individuazione delle corrispondenti tipologie di errore.

Piani di incompatibilità

Gli *edit di consistenza* vengono utilizzati per la costruzione dei cosiddetti *piani di incompatibilità*. Più rigorosamente, si definisce p*iano di incompatibilità* un insieme di vincoli (edit) non ridondanti e non contraddittori che devono essere contemporaneamente soddisfatti da ogni unita' statistica affinché l'informazione corrispondente possa essere considerata corretta.

Il controllo effettuato sui dati mediante un piano di incompatibilità è di tipo *intra-record* se utilizza le sole informazioni fornite da ogni singola unità statistica, è di tipo *inter-record* quando i dati relativi ad una certa osservazione vengono confrontati con informazioni prodotte da altre osservazioni della stessa popolazione. Naturalmente, al variare dell'insieme di vincoli, e quindi della tipologia di errore, si otterranno diversi sottoinsiemi di record incorretti.

Gli edit componenti un piano di incompatibilità possono essere distinti in:

1. *regole formali*, che derivano dalla struttura del modello, cioè direttamente dalle norme di compilazione e dai "percorsi interni" (salti) del modello;
2. *regole sostanziali*, che derivano da considerazioni di tipo statistico-matematico, o da conoscenze specifiche a priori del fenomeno oggetto di rilevazione.

E' chiaro che la natura degli edit (sia formali che sostanziali) di un piano di incompatibilità è strettamente dipendente dal tipo di variabili (qualitative o quantitative) oggetto di verifica. Mentre nel caso di variabili qualitative, infatti, tali edit hanno la forma di relazioni *logiche* tra le variabili, nel caso di variabili quantitative le regole di incompatibilità sono espresse in forma di relazioni *statistico/matematiche* (equazioni o disequazioni lineari, rapporti, ecc.).

Una volta individuati i record i cui valori violano uno o più vincoli del piano di incompatibilità, il problema diventa la localizzazione delle variabili responsabili di tale violazione: sono solo queste, infatti, le variabili i cui valori devono essere considerati errati (cioè mancanti) e quindi corretti.

Sia il problema della localizzazione dei record errati, sia quello dell'individuazione delle variabili che, per ogni record errato, sono da considerarsi responsabili della violazione di una o più regole di incompatibilità, possono essere risolti adottando un approccio di tipo *interattivo* oppure *automatico*.

Nel caso di approccio *interattivo*, l'individuazione delle incompatibilità è basata sull'interazione tra esperto e dati, per cui il processo di verifica e correzione dipende strettamente da decisioni umane prese caso per caso.

Nel caso dell'editing *automatico*, si deve distinguere il caso in cui si utilizzi software *ad hoc* (cioè specificamente sviluppato per una data tipologia di rilevazioni) oppure *generalizzato* (cioè immediatamente adattabile a diverse tipologie di indagine).

Nell'ambito dell'editing di tipo *automatico* possiamo ulteriormente distinguere a seconda che per la costruzione della procedura di editing si adotti un approccio di tipo *deterministico* oppure *probabilistico*.

Procedure di localizzazione dei valori anomali

L'insieme degli *edit statistici* costituiscono la base per le cosiddette *procedure di localizzazione dei valori anomali e dei valori sospetti*.

Abbiamo già osservato che la presenza, per una data variabile, di valori anomali dovuti a risposte errate è spesso un efficace indicatore di presenza di errore sistematico per quella variabile. Inoltre, potendo tali valori avere un impatto considerevole sulle statistiche e sulle stime calcolate sui dati, se non opportunamente corretti o ponderati possono produrre delle notevoli distorsioni sui risultati finali dell'indagine.

In particolare, nel caso di indagini censuarie o esaustive gli outlier corrispondono generalmente a valori che si discostano in modo significativo dal resto delle unità della popolazione rispetto ad un certo fenomeno rilevato, e quindi hanno sempre un impatto sulle stime finali.

Nel caso di indagini campionarie il concetto di outlier assume un significato più complesso: essi possono infatti corrispondere non solo a valori estremi, ma anche a valori non estremi ma che hanno un impatto eccessivo sulle stime finali a causa del loro elevato peso campionario (*valori influenti*)[1]. Questo problema si acuisce nel caso di popolazioni con grosse concentrazioni, in cui è possibile che unità di grosse dimensioni abbiano un alto peso campionario: in questi casi, il campione deve essere disegnato in modo che grosse unità siano

[1] Outlier corrispondenti a valori estremi non sono necessariamente *valori influenti* nel caso in cui abbiano un piccolo peso campionario.

selezionate con probabilità 1 o con elevata probabilità, in modo da poter assegnare ad esse piccoli coefficienti di espansione.

L'origine degli outlier può essere dovuta a errori di misura commessi in una qualunque delle fasi della rilevazione, ad errata interpretazione del modello, ad errata trascrizione dei dati, ma anche alla variabilità intrinseca del fenomeno. E' pertanto importante verificare se tali valori corrispondono a risposte errate (devono quindi essere assimilati a mancate risposte e, di conseguenza, devono essere imputati) oppure a dati reali (e quindi devono essere accettati come corretti e opportunamente considerati in fase di calcolo delle stime).

La localizzazione degli outlier avviene mediante determinazione di *intervalli di accettazione* al di fuori dei quali una unità statistica è da considerare anomala e quindi da sottoporre a controllo ed, eventualmente, a correzione. L'individuazione degli outlier è generalmente più efficace se essa è effettuata su *domini* disgiunti dell'intera popolazione, determinati sulla base di una o più variabili presenti per tutte le unità rispondenti. Questo se all'interno di tali domini si ritiene che il comportamento delle unità rispetto al carattere in esame sia omogeneo.

La determinazione degli intervalli di accettazione può essere:

1. *empirica* se i limiti degli intervalli di accettabilità sono determinati dallo statistico sulla base della distribuzione della variabile stessa (o di una sua funzione) nella popolazione di riferimento;
2. a*utomatica* quando i limiti di accettabilità sono determinati sulla base di algoritmi implementati in programmi software.

Generalmente, i valori anomali per una certa variabile osservata sono individuati calcolando le distanze relative di ogni unità dal *centro* dei dati (considerati nel loro complesso o per domini), e determinando un valore di *soglia* oltre il quale le unità sono da considerare sospette. Il centro della distribuzione può essere determinato utilizzando media e varianza (che però sono a loro volta stimatori molto sensibili alla presenza di outlier), oppure utilizzando i quantili della distribuzione (in particolare, mediana, primo e terzo quartile). Tipicamente, i metodi basati sull'uso dei quantili non tengono conto, nel caso di indagini campionarie, dei pesi campionari delle unità: in questi casi vengono infatti localizzati solo i valori estremi non pesati.

La tecnica di determinazione degli intervalli è di tipo diverso a seconda che, per la variabile in esame, si disponga o meno di valori storici (cioè a seconda che l'indagine in esame sia o meno di tipo periodico). Nel caso di indagini periodiche, in cui cioè si disponga di informazioni storiche, i limiti di accettabilità vengono determinati utilizzando funzioni che considerano i valori delle variabili in ripetizioni precedenti dell'indagine. In questi casi, per ogni variabile da controllare, il criterio adottato è basato sul calcolo per ogni record rispondente di una particolare funzione (rapporto, differenza, ecc.) del suo trend storico: i limiti di accettabilità per la selezione delle osservazioni anomale vengono calcolati sulla base della distribuzione della stessa funzione nella popolazione dei rispondenti.

Per indagini non periodiche, o qualora lo statistico lo ritenga opportuno anche in presenza di dati storici, i limiti di accettabilità possono essere determinati mediante funzioni che utilizzano i soli valori correnti delle variabili di interesse: in questo caso, gli outlier per una data variabile possono essere individuati in due modi distinti:

1. confrontando i valori della variabile con limiti di accettabilità calcolati sulla base dei valori che la variabile assume in altri record nello stesso periodo di riferimento;

2. confrontando i valori di rapporti, calcolati fra variabili correlate fra loro, con limiti di accettazione calcolati sulla base della distribuzione degli stessi rapporti all'interno della stessa popolazione.

Fra i metodi più efficaci per la localizzazione degli outlier ricordiamo la procedura di Hidiroglou-Berthelot, che sarà descritta in dettaglio nel paragrafo relativo alle tecniche di correzione di tipo interattivo note come *macroediting univariato*. In realtà, tale metodo è stato utilizzato anche nell'ambito delle procedure automatiche generalizzate, e più precisamente nel software GEIS.

Le procedure per la localizzazione di valori con comportamento anomalo rispetto alla popolazione di riferimento e con impatto importante sulle stime finali sono, in generale, tutte le procedure afferenti al *macroediting* e all'*editing selettivo*.

Altre tecniche di localizzazione degli outlier presuppongono assunzioni distribuzionali sulla popolazione investigata: nel caso in cui tale popolazione abbia una specifica distribuzione parametrica, è infatti possibile sviluppare test statistici per individuare quei valori che non sono originati dal prefissato modello generale.

4.2. Correzione

A fronte della verifica della presenza di mancate risposte nei dati esistono varie alternative possibili per lo statistico: *ignorarne la presenza, integrarle, ponderarle*.

Nel primo caso, quando cioè si sceglie di utilizzare i soli dati completi, insiemi di dati con alte percentuali di valori mancanti o comunque con informazioni incomplete possono rendere non utilizzabili i risultati dell'indagine, o non significative o inattendibili le stime e le analisi statistiche prodotte. Questo in quanto è noto che raramente l'insieme dei rispondenti può essere considerato un sotto campione rappresentativo dell'intera popolazione, compresi i non rispondenti.

In alternativa, gli elementi che presentano valori mancanti o che appartengono al sottoinsieme delle variabili errate e le mancate risposte totali possono essere *integrati* o, come si dice con altra terminologia, le corrispondenti variabili possono essere *imputate*. La fase di integrazione delle mancate risposte, essendo in ogni caso un processo di *ricostruzione* di informazione, deve essere effettuato con grande cautela: infatti, esiste sempre il rischio che vengano distrutte informazioni per rendere possibile la creazione di dati consistenti rispetto a prefissati modelli. Per questo motivo, è innanzitutto opportuno che l'imputazione di dati avvenga sulla base di metodologie e tecniche che diano prefissate garanzie di qualità e di efficienza, come il mantenimento delle distribuzioni originali dei dati, l'oggettività e la riproducibilità. In ogni caso, è sempre necessario tenere traccia di quali e quante imputazioni sono state effettuate: il processo di imputazione deve essere attentamente monitorato e valutato, deve essere esaustivamente documentato. In queste condizioni, il risultato delle imputazioni può essere vantaggioso sia perché consente la disponibilità di insiemi di dati completi e coerenti, sia perché rende più agevole l'effettuazione di analisi statistiche sui dati.

L'ultima possibilità in presenza di mancate risposte è quella di operare a livello di *stima finale*, ricalcolando il sistema iniziale dei pesi. Questa soluzione è generalmente preferibile nel caso delle MRT, ma può essere adottata in teoria anche nel caso delle MRP.

Correzione degli errori

Una volta individuati i record contenenti valori errati, e quindi non accettabili, e le variabili responsabili di tale non correttezza, si pone il problema della loro modifica in modo da riportare il record nella condizione di accettabilità rispetto ai criteri (piano di incompatibilità o piano di localizzazione dei valori anomali) utilizzato.

Le procedure esistenti per tale operazione possono essere classificati secondo diversi punti di vista.

Distinguiamo innanzi tutto fra tecniche di correzione di tipo *micro* e di tipo *macro*. Le prime prevedono il controllo di tutti i record presenti nel data set e la correzione di tutti quelli che hanno determinato l'attivazione di un qualsiasi edit. L'approccio macro, invece, prevede la verifica e l'eventuale correzione delle sole unità che incidono maggiormente sulle stime finali dei dati. Nell'ambito dei metodi di tipo macro distinguiamo fra tecniche del *macroediting* e tecniche di tipo *selettivo*. Entrambe sono di tipo *interattivo*, cioè prevedono che i record errati o con alta probabilità di esserlo vengano corretti sulla base dell'intervento diretto dell'operatore, il quale provvede a rimuovere l'errore mediante verifica del modello cartaceo o, laddove possibile, mediante reintervista.

Le tecniche di correzione di tipo *micro*, invece, possono essere di tipo sia *interattivo* sia *automatico*. I metodi rientranti nel primo tipo possono essere utilizzati in contesti sia interamente interattivi (in cui cioè anche la determinazione degli errori avviene attraverso l'interazione fra dati ed esperto), sia in ambiti parzialmente automatici (in cui cioè l'individuazione delle componenti errate nei record avviene attraverso l'utilizzo di software automatico in cui sono implementate le regole di controllo). In quest'ultimo caso si parla di procedure di controllo e correzione di tipo *misto*.

Le tecniche di tipo automatico sono invece generalmente implementate in procedure interamente automatizzate. In questo caso, viene sviluppato software (ad hoc o generalizzato) in cui, una volta individuate (automaticamente) le eventuali incompatibilità nel record, i valori da imputare sono determinati, sempre automaticamente, attraverso metodi *deterministici* o *stocastici*.

I metodi *deterministici* prevedono che, dato un insieme di rispondenti, i valori da imputare siano determinati in modo univoco. Nei metodi *stocastici*, invece, i valori da imputare sono soggetti a maggiore o minore grado di casualità. Nella maggior parte dei casi questi metodi sono ottenuti mediante introduzione di una componente stocastica nei modelli di tipo deterministico.

Un'altra importante distinzione nell'ambito dei metodi di tipo micro può essere fatta fra metodi in cui il valore da imputare per una data variabile è ottenuto esclusivamente sulla base dei valori delle restanti variabili o in base all'utilizzo di informazioni ausiliarie, e metodi in cui si ricorre ai valori osservati in altre unità rispondenti, ad esempio sostituendo direttamente il valore errato con il corrispondente valore presente in una unità *donatrice*, oppure ricorrendo all'uso di modelli (rapporti, regressione, ecc.).

Correzione dei valori anomali

Una volta individuate le unità in cui una o più variabili presentano valori anomali rispetto all'insieme residuo dei rispondenti, esistono due possibili alternative:

1. escludere i valori anomali dalle elaborazioni successive e dal calcolo delle stime finali;

2. verificare se gli outlier individuati corrispondono o meno a situazioni errate, sono cioè dovuti a errori di compilazione o di registrazione o se invece corrispondono alla situazione reale del rispondente rispetto al carattere rilevato. Questo tipo di analisi può essere solo di tipo *interattivo*, e può consistere nella revisione dei modelli cartacei (laddove disponibili) o dei record corrispondenti, oppure, laddove praticabile, nella reintervista del rispondente. Nel caso in cui i valori anomali corrispondano alla reale situazione dell'unità rispondente, trattandosi non errori, ma di *valori estremi*, è necessario verificare se essi corrispondono o meno ad *unità influenti*, cioè se la loro inclusione o esclusione ha o meno un impatto importante sulle stime.

Nel caso 1 (esclusione totale degli outlier) possono essere introdotte gravi distorsioni nei risultati finali del processo di produzione dell'informazione statistica dal momento che, se gli outlier corrispondono a valori reali, si rinuncia a informazioni in ogni caso corrette, che rappresentano modalità possibili dell'evolversi del fenomeno in oggetto. Questa soluzione è accettabile solo nel caso in cui gli outlier corrispondano ad osservazioni errate e non influenti (cioè con trascurabile impatto sulle stime).

Nel secondo caso, al controllo interattivo possono seguire le seguenti operazioni:

1. in fase di editing, imputazione dei valori anomali corrispondenti a risposte errate;
2. trattamento dei valori anomali dovuti al reale evolversi del fenomeno (cioè degli outlier corrispondenti a valori corretti) a livello di stima.

La prima operazione può avvenire in due modi distinti, a seconda del tipo di verifica effettuata sui valori anomali:

1. se il controllo interattivo avviene mediante reintervista oppure se l'outlier è dovuto ad un errore di registrazione, la correzione del dato avviene contestualmente a questa fase;
2. se l'outlier è dovuto ad un errore di compilazione e non è possibile ricontattare il rispondente, analogamente a qualunque altro tipo di errore, tali valori possono essere considerati errori a tutti gli effetti e, quindi, sottoposti a imputazione mediante uno qualunque dei metodi esistenti per la correzione degli errori (interattiva, automatica deterministica o probabilistica).

La seconda operazione, che prevede il trattamento degli outlier a livello di calcolo della stima finale, introduce normalmente distorsioni negli stimatori utilizzati (cioè una deviazione del valore atteso dello stimatore rispetto al parametro vero). Per questo motivo, tutte le tecniche utilizzate a livello di stima hanno l'obiettivo di ridurre la varianza di stimatori non corretti. Esistono sostanzialmente tre approcci al trattamento degli outlier in fase di stima:

1. modifica dei valori degli outlier;
2. determinazione per gli outlier di nuovi pesi che tengano opportunamente conto dell'impatto che le unità anomale hanno sul fenomeno nel suo complesso;
3. utilizzo di tecniche di stima *robuste*, cioè poco sensibili alla presenza nei dati di valori anomali.

La prima alternativa, nota come *tecnica di Winsorizzazione*, richiede l'ordinamento dei valori delle variabili di interesse nella popolazione rilevata, ed è applicabile solo nel caso di indagini campionarie in cui il disegno sia casuale semplice. Tale tecnica consiste nel sostituire i valori della variabile Y di interesse nelle k osservazioni anomale poste all'estremo della lista ordinata di valori con k valori precedenti opportunamente determinati.

Alla modifica dei valori dei k outlier individuati viene però generalmente preferita la riduzione dei loro pesi: in questa ottica sono stati proposti degli stimatori, detti *stimatori riponderati*, in cui il peso delle osservazioni anomale influenti viene modificato in modo tale

da soddisfare prefissate condizioni[2]. Nell'ambito di questo approccio, gli studi sono attualmente concentrati nella ricerca di pesi correttivi *ottimi* nel senso che minimizzino la varianza degli *stimatori riponderati*, e quindi nella ricerca dello stimatore più efficiente per ogni prefissato parametro di interesse (totali, medie, percentuali, ecc.).

Il terzo approccio al trattamento degli outlier in fase di stima, di più recente sviluppo, è basato sull'uso di tecniche di stima robuste, come lo stimatore di massima verosimiglianza (*stimatore M* nel seguito). In realtà, la maggior parte degli stimatori robusti finora proposti (rapporto, di regressione, ecc.) sono ottenuti, a partire dai corrispondenti stimatori non robusti, mediante applicazione dello stimatore M. Gli stimatori robusti sono generalmente non corretti e non lineari, e la loro efficienza dipende dal tipo di informazioni disponibili sul fenomeno investigato (tali informazioni possono essere infatti tradotte in ipotesi distribuzionali ed utilizzate in fase di costruzione dello stimatore M).

Sia l'individuazione che l'imputazione dei valori anomali risultano più efficaci se vengono utilizzati metodi che tengono conto in qualche misura di informazioni storiche o ausiliarie sui rispondenti, naturalmente laddove queste siano disponibili. La maggior parte dei più sofisticati metodi interattivi per la localizzazione dei valori anomali, infatti, sono pensate per essere applicate a indagini periodiche o per le quali in ogni caso si disponga di informazioni ausiliarie provenienti anche da fonti esterne. Tipico esempio sono i metodi afferenti al macroediting ed all'editing selettivo, come vedremo meglio nel seguito.

5. Trattamento delle mancate risposte totali (MRT)

La presenza di MRT nei dati è un problema comune a tutte le indagini, sia campionarie sia censuarie: tutti gli strumenti adottabili per la prevenzione di tale fenomeno possono solo ridurne l'intensità, ma non riescono in ogni caso ad eliminarne del tutto la presenza.

Le mancate risposte totali hanno due effetti sui risultati finali:

- riducono la quantità di informazione disponibile: nel caso di indagini campionarie, attraverso la riduzione della numerosità campionaria, viene prodotto un incremento del relativo errore di campionamento;
- introducono distorsioni nelle stime quando il meccanismo che le genera è non casuale (come generalmente accade).

Il trattamento delle MRT ha lo scopo di prevenire le distorsioni che la loro presenza può provocare sui risultati finali del processo di produzione dell'informazione statistica. Questo trattamento può avvenire a tre livelli: in fase di rilevazione, in fase di editing oppure in fase di stima finale. Nel primo caso si cerca di ridurre il fenomeno della MRT prevedendo delle sostituzioni per le unità eventualmente non rispondenti. Nel secondo caso, le MRT vengono sottoposte a integrazione analogamente a quanto avviene per le MRP. Nel terzo caso, il problema consiste nell'eliminazione o nella riduzione della distorsione prodotta dalla presenza di MRT nelle stime finali attraverso l'utilizzo di opportuni pesi correttivi.

[2] Ad esempio, nel caso della stima di un totale, i pesi correttivi vengono determinati in modo tale che la somma dei pesi finali rimanga pari a N (dimensione della popolazione).

Rilevazione

Nel caso di indagini campionarie, le unità non rispondenti possono essere sostituite direttamente in fase di rilevazione con altre unità precedentemente selezionate casualmente dalla stessa lista. Questo metodo presenta il vantaggio di ripristinare la numerosità campionaria iniziale, ma possono non essere eliminati gli effetti distorsivi sulle stime finali se la sub-popolazione dei rispondenti rappresentata dalle unità sostitutive hanno caratteristiche sistematicamente diverse da quelle dei non rispondenti.

Sempre nel caso di indagini campionarie, un metodo di correzione degli effetti della presenza di MRT sulle stime finali consiste nell'estrarre un sub-campione casuale semplice dalla popolazione dei non rispondenti, e di procedere alla reintervista, mediante ritorni successivi, delle unità selezionate. In tal modo, ottenuta la stima relativa ai non rispondenti, è possibile ridurre la distorsione della stima finale (ad esempio ottenuta mediante combinazione lineare delle due stime parziali, quella dei rispondenti e quella del campione dei non rispondenti)[3]. Questa tecnica è però raramente praticabile nel caso di indagini di tipo amministrativo, ed in ogni caso risulta essere piuttosto costosa e in termini sia economici che organizzativi.

Imputazione

Se fra le esigenze dell'indagine c'è la costruzione di un archivio *completo* di informazioni le MRT possono essere sottoposte a imputazione analogamente alle MRP. Ciò è possibile nel caso in cui siano disponibili le caratteristiche strutturali della popolazione investigata e informazioni ausiliarie affidabili. I metodi utilizzabili a questo scopo possono essere basati sull'uso di *donatori* oppure sull'adozione di *modelli statistico-matematici* di varia natura.

Nel primo caso, le informazioni relative ad ogni unità totalmente non rispondente vengono ottenute mediante duplicazione di una unità rispondente *donatrice*, scelta secondo un prefissato criterio casuale fra un insieme di unità donatrici candidate. Le unità donatrici candidate sono generalmente ottenute classificando tutti i possibili donatori sulla base di variabili ausiliarie, note per tutte le unità rispondenti, che si suppone discriminino fra diversi modelli di risposta. E' evidente che il rapporto fra tali variabili ausiliarie ed il modello di risposta vanno verificate, così come va verificata l'indipendenza del meccanismo aleatorio di riposta dal livello delle variabili ausiliarie utilizzate.

Nel caso di imputazione mediante modelli vengono utilizzati generalmente modelli *deterministici* in cui si assume una dipendenza di tipo lineare fra un sottoinsieme di variabili di interesse ed un insieme di variabili esplicative. Le funzioni che esprimono tale dipendenza sono generalmente a loro volta dipendenti da un insieme di parametri, che devono essere stimati sulla base delle informazioni fornite dalle unità rispondenti.

Riponderazione

[3] La stime finale è non distorta solo nel caso in cui si proceda alla reintervista di tutti i non rispondenti.

E' chiaro che, quando le informazioni relative ad alcune unità statistiche risultano completamente mancanti e non è possibile o non si ritiene opportuno procedere alla loro integrazione, è necessario tenere conto di questa assenza di informazione a livello di stima finale: ciò può essere fatto incrementando il valore dei pesi campionari di unità rispondenti considerate rappresentative di quelle non rispondenti. E' chiaro che l'assunzione alla base di questo approccio è piuttosto critica, in quanto si assume una omogeneità di probabilità di risposta fra rispondenti e non rispondenti non sempre accettabile, e che dovrebbe essere in ogni caso sempre accuratamente verificata.

Fra le tecniche di riponderazione più diffuse ricordiamo il *metodo geografico* e il *metodo della ponderazione vincolata*.

Il primo metodo consiste nel far rappresentare le unità non rispondenti da unità appartenenti a classi territoriali contigue, e viene spesso usato in combinazione col *criterio dell'aggregazione degli strati*, consistente appunto nell'integrazione fra strati in cui si verifica un completa caduta delle unità campione e strati contigui che ne diventano così rappresentativi. Vantaggio di tale metodo è il fatto che la somma dei pesi modificati coincide col totale delle unità della popolazione. Il principale svantaggio è legato alla non correttezza generale delle stime finali: tali stime risultano infatti non distorte solo nel caso in cui il fattore correttivo applicato ai pesi iniziali sia il reciproco della probabilità di risposta delle unità rispondenti.

Il *metodo della ponderazione vincolata* (Falorsi P.D., Falorsi S. 1995) può essere adottato per tutte quelle indagini per le quali si dispone di totali noti sulla popolazione oggetto di indagine, ottenuti o da fonti esterne oppure sulla base dell'archivio da cui il campione di unità statistiche è stato selezionato. Questo metodo consiste nel calcolare i fattori correttivi per i pesi campionari in modo tale che siano rispettati i vincoli di uguaglianza fra i totali noti e le rispettive stime campionarie. Gli stimatori utilizzati per il calcolo di tali stime, detti *stimatori di ponderazione vincolata*[4], consentono in generale di attenuare gli effetti distorsivi dovuti alla presenza di MRT, e, sotto particolari condizioni sui modelli probabilistici che generano la mancata risposta, si dimostra che tali stimatori sono non distorti.

Il metodo della ponderazione vincolata e, in generale, tutti i metodi di riponderazione in presenza di MRT, presuppongono la specificazione di *modelli probabilistici di interpretazione* della mancata risposta totale, o *modelli di mancata risposta*. Questi modelli vengono utilizzati, in presenza di MRT, per la stima delle probabilità di risposta delle unità campionarie, qualora tali probabilità siano incognite. Attraverso queste stime vengono calcolati fattori correttivi ai pesi iniziali tali che gli stimatori delle quantità finali siano non distorti.

La maggior parte dei modelli di mancata risposta fanno uso, oltre che delle informazioni fornite dai rispondenti, di informazioni ausiliarie, che possono consistere in dati sui non-rispondenti ottenibili o dalle liste di selezione del campione, oppure da fonti esterne all'indagine che descrivono alcune caratteristiche dei non rispondenti.

[4] Ad esempio, lo *stimatore di regressione generalizzato* e lo *stimatore ratio raking*.

6. Procedure di controllo e correzione

Una volta definito il piano di incompatibilità, l'insieme cioè delle regole che permettono di individuare, ed eventualmente correggere, gli errori all'interno dei dati, è necessario definire le modalità di applicazione di tali regole ai dati stessi.

Distinguiamo anzitutto tra le seguenti fasi:

1. individuazione delle situazioni di errore, mediante verifica delle situazioni di *fuori dominio*, *mancate risposte parziali* e *incompatibilità*;
2. localizzazione degli errori che causano le incompatibilità;
3. correzione degli errori mediante attribuzione di nuovi valori alle variabili errate.

Ognuna delle fasi citate può essere effettuata in modo manuale, interattivo, automatico o misto. Le modalità *manuale* e *interattiva* presuppongono l'intervento umano per ogni tipo di decisione, quella *automatica* prevede la totale delega al computer di tali decisioni, mentre quella *mista* fa ricorso sia all'intervento umano che a quello della macchina. La differenza tra manuale ed interattiva è data dalla diversa modalità di utilizzo del computer da parte dell'operatore umano: nel primo caso il processo decisionale è totalmente indipendente dall'elaboratore, mentre nel secondo caso si determina attraverso una continua interazione tra uomo e macchina.

Normalmente, la fase di individuazione delle situazioni di errore è sempre compiuta in modo automatico, in quanto non vi sono particolari decisioni da prendere: si tratta solo di verificare se un record presenta mancate risposte parziali, valori fuori dominio o dà luogo o meno a incompatibilità. Per quanto riguarda invece le altre fasi, le decisioni da prendere sono estremamente delicate, in quanto, se non eseguite correttamente, possono portare non già alla correzione degli errori presenti, ma addirittura all'introduzione di nuovi, e, in ultima analisi, allo stravolgimento della distribuzione originale. In merito all'adozione della modalità interattiva od automatica giocano considerazioni relative alle conseguenze sulla qualità finale dei dati e sui costi in termini di risorse e di tempi necessari. Mentre per quanto riguarda quest'ultimo elemento è innegabile che la soluzione automatica risulta essere sempre vantaggiosa, non altrettanto si può dire riguardo la qualità: sotto questo aspetto, è decisiva la valutazione delle tecniche e degli algoritmi utilizzati nell'uno e nell'altro caso. Ad esempio, un conto è che le procedure automatiche di localizzazione e correzione degli errori si basino su algoritmi di tipo deterministico, che spesso non danno garanzie di qualità, un altro è che facciano riferimento a metodi di tipo stocastico, che massimizzano la probabilità di trovare l'errore "vero" e di ripristinare il valore reale nelle variabili errate. Analogamente, le tecniche interattive, se non guidate da metodologie corrette, dipendono in modo drammatico dalle capacità dei singoli operatori.

Le differenti combinazioni delle citate modalità nelle varie fasi danno luogo ad alcune caratteristiche tipologie organizzative del processo di controllo e correzione dei dati.

L'**organizzazione**, che potremmo definire "**tradizionale"**, della fase di controllo e correzione dei dati, è caratterizzata dai seguenti elementi (Granquist, 1995):

1. l'acquisizione dei dati avviene mediante *data entry senza controlli ("heads down")*: in altre parole, si provvede alla trascrizione dei dati su supporto magnetico senza controllare contestualmente se tali dati sono tra loro compatibili;

2. il *controllo* delle situazioni di errore è eseguito mediante applicazione di un programma contenente le regole di compatibilità (*machine editing*): il risultato è un listato dei record errati, riportante accanto ad ognuno di essi le incompatibilità riscontrate;
3. la *localizzazione degli error*i e la loro *correzione* vengono effettuate manualmente dagli operatori umani che provvedono a modificare i valori delle variabili sul listato dei record errati: si procede quindi ad una nuova registrazione dei record corretti in tal modo, e si ritorna al passo 2, iterando il processo finché non si riscontrano più situazioni di errore o finché queste non sono giudicate trascurabili. Una variante, in caso di indisponibilità di risorse umane sufficienti, consiste nell'applicazione di *procedure deterministiche per la correzione automatica*: anche in tal caso può essere necessario iterare il processo, dato che tali procedure non garantiscono l'eliminazione immediata delle incompatibilità riscontrate.

Tale organizzazione, la prima ad essere utilizzata negli enti produttori di statistiche, negli anni '60 e '70, quando l'ambiente elaborativo era caratterizzato da mainframe operanti in modo batch, soffre di limiti evidenti, di cui il più vistoso è senz'altro quello costituto dalla necessità di iterare le operazioni di registrazione dei record corretti in modo manuale su supporto cartaceo, in caso di correzioni manuali, oppure dagli scarsi livelli qualitativi ottenuti mediante l'applicazione delle procedure deterministiche.

La prima importante modifica a tale impostazione riguarda l'adozione di **modalità interattive di lavoro**, con un rilevante impatto sia nella fase di acquisizione dei dati, che in quella della correzione degli errori:
1. l'acquisizione dei dati avviene con dei controlli contestuali di compatibilità (*data entry "heads up"*): vengono immessi i dati relativi ad ogni singolo modello, il programma segnala le eventuali situazioni di errore all'operatore, che provvede a controllare la corrispondenza tra le risposte sul questionario e quanto digitato, ed in caso di discordanza procede alla correzione; se si verificano ancora delle incompatibilità, queste sono dovute in linea di massima ad errori di compilazione del modello, e non ad errori di registrazione;
2. come prima, il controllo è effettuato in modo automatico, ma il risultato è una suddivisione del file dei dati registrati in due sottoinsiemi, quello degli esatti e quello degli errati;
3. il file degli errati viene sottoposto a correzione interattiva: ogni record è visualizzato assieme alle segnalazioni di errore, l'operatore provvede ad eseguire le correzioni fino ad eliminare le incompatibilità.

Questo tipo di organizzazione è soggetto a un limite rilevante: la *soggettività* degli operatori cui compete la totalità delle decisioni, sia quelle riguardanti la scelta, per ogni record errato, di quali variabili correggere, sia quelle relative ai valori da assegnare ad esse. Da un punto di vista della qualità dei dati, questa scelta si giustifica solo nel caso di *follow-up*: i rispondenti vengono ricontattati e, assieme a loro, si verificano le risposte fornite e registrate nel record. Se non si procede al follow-up, la correttezza delle correzioni apportate può essere garantita solo dall'esperienza e dalla sensibilità dell'operatore, qualità che non sempre sono assicurate. Si parla in tal caso di *editing creativo* e non in senso positivo: le correzioni vengono apportate non tanto per ripristinare i valori "veri" all'interno del record, ma per assicurare la plausibilità di tali valori e per rimuovere le inconsistenze.

Per tale motivo, un ulteriore modifica riguarda le modalità di localizzazione degli errori e della loro correzione: non più di competenza esclusiva dell'operatore, ma eseguita in modo **automatico** da un opportuno algoritmo che, per ogni record errato, sulla base delle

incompatibilità attivate oppure valutando la distribuzione complessiva dei dati, ricerca le variabili che più probabilmente sono quelle errate. All'operatore competono ancora la conferma delle indicazioni fornite dal computer, e la scelta tra diverse possibili modalità di correzione (deterministiche o probabilistiche).

Un'organizzazione di questo tipo, pur assicurando alti livelli qualitativi, è però anch'essa affetta da un limite: le risorse ed i tempi necessari per attuarla. Infatti, dovendo sottoporre a trattamento tutti i record errati, il numero di operatori da impiegare ed il tempo per eseguire tutti i controlli e le correzioni in modo interattivo possono risultare inaccettabili rispetto a vincoli esterni.

La risposta a questo problema può essere l'adozione di **procedure totalmente automatiche**, che permettano l'esecuzione del passo 3 in modo rapido e indipendente da risorse umane: se gli algoritmi di localizzazione dell'errore e di imputazione sono basati su metodologie rigorose, oltre alla minimizzazione dei costi è assicurata anche la massimizzazione della qualità.

Una soluzione alternativa è quella costituita dall'adozione di **tecniche selettive** (in particolare, le più note: *macroediting* e *editing selettivo*): viene controllato e corretto interattivamente, con particolare cura, solo quel sottoinsieme di record che ha un maggiore impatto sulle stime finali di interesse, mentre la restante parte dei record non viene sottoposta ad editing, oppure lo è in modo automatico. Un approccio di questo tipo è possibile quando la distribuzione delle variabili di interesse mostra picchi di concentrazione in sottoinsiemi di unità di rilevazione, ed è conveniente in quanto, minimizzando l'impiego di risorse umane, lo concentra sulle unità più importanti garantendo affidabilità delle stime.

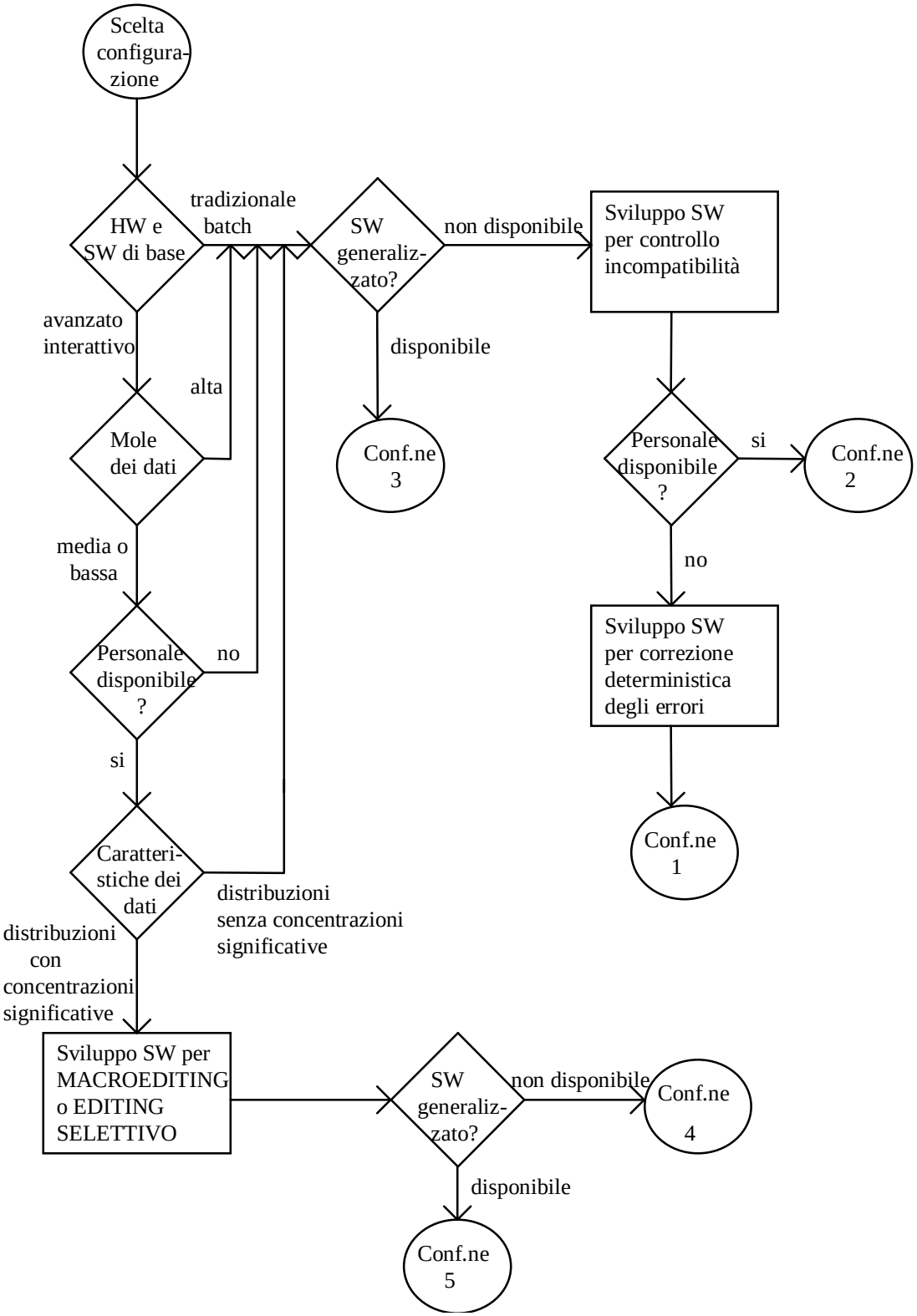
Abbiamo visto come la fase di controllo e correzione può essere configurata secondo numerose organizzazioni alternative, determinate dalle scelte relative all'impiego di procedure automatiche o interattive. I parametri che guidano nella scelta della specifica configurazione sono essenzialmente i livelli di qualità auspicati, le risorse e gli strumenti disponibili ed i vincoli sui tempi di esecuzione.

Avendo posto come obiettivo la massimizzazione dei livelli di qualità, nel processo decisionale relativo alla scelta della configurazione vanno valutati i seguenti elementi:

1. **caratteristiche dei dati**: variabili qualitative o quantitative, distribuzioni con o senza concentrazioni rilevanti, disponibilità di serie storiche, numerosità (bassa, media, alta);
2. **disponibilità hardware e software di base**: ambiente tradizionale utilizzabile essenzialmente in modo batch, oppure configurato con stazioni di lavoro adatte alle modalità interattive (ambiente database e interfacce avanzate);
3. **disponibilità di software generalizzato** che incorpori algoritmi ottimizzati per la localizzazione degli errori e/o per l'imputazione delle variabili errate;
4. **disponibilità di personale** motivato e con elevata esperienza.

Nel seguito è riportato un possibile schema decisionale che lega la valutazione degli elementi ora citati alla scelta della particolare organizzazione. In esso ogni configurazione possibile è caratterizzata dal tipo di risposte che si danno alle domande contenute nei simboli di decisione.

SCHEMA PER LA SCELTA DELLA CONFIGURAZIONE DELLE PROCEDURE DI CONTROLLO E CORREZIONE DEI DATI



La **prima configurazione** è quella cui si perviene in caso di risposte totalmente negative: l'ambiente, di tipo tradizionale, non consente trattamenti interattivi di tipo avanzato, ma solo quelli batch. Non è disponibile software generalizzato che incorpori algoritmi ottimizzati, è dunque necessario sviluppare il software ad hoc per il controllo delle incompatibilità presenti nei dati. Infine, non sono disponibili adeguate risorse di elevata esperienza in grado di localizzare e correggere gli errori sulla base delle segnalazioni ed è dunque necessario ricorrere allo sviluppo di programmi deterministici che effettuino tali operazioni.

La **seconda configurazione** si distingue dalla precedente in quanto la disponibilità di personale di buona professionalità consente di attuare la fase di localizzazione degli errori e loro correzione senza essere costretti a ricorrere all'approccio automatico deterministico.

La **terza configurazione** è quella caratterizzata dall'approccio automatico non deterministico (o, come vedremo, non prevalentemente deterministico): non è necessario, o è molto limitato, lo sviluppo di software ad hoc, è sufficiente fornire le indicazioni dei domini ammissibili per ogni variabile, delle regole di compilazione del questionario e delle incompatibilità che possono manifestarsi tra i valori delle diverse variabili nel record: è il sistema che provvede al controllo dei dati, alla localizzazione degli errori ed alla loro correzione. Questo tipo di scelta non è condizionata solo dal tipo di ambiente elaborativo, ma anche da altre situazioni, quali la scarsità di personale, un'elevata mole di dati, caratteristiche dei dati che non consentono l'utilizzo di particolari tecniche alternative. E' una scelta che può essere preferita anche indipendentemente da ogni considerazione, vista l'elevata qualità assicurata e la garanzia di tempi rapidi di esecuzione.

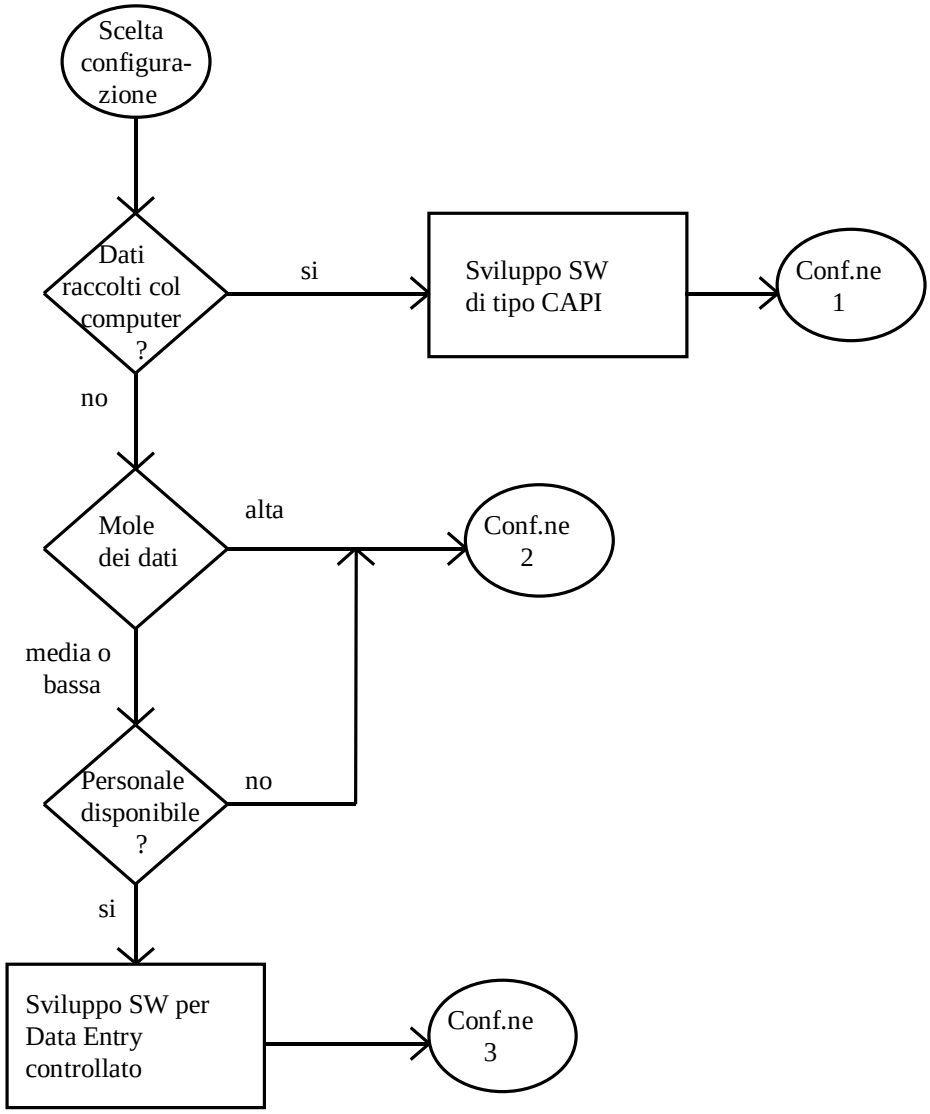
La **quarta configurazione** è invece quella che prevede un trattamento interattivo dei dati, basato su tecniche di tipo avanzato quali il macroediting o l'editing selettivo. Ciò è possibile quando ricorrono le condizioni di un ambiente elaborativo avanzato, di sufficienti risorse umane di elevata professionalità, e di specifiche caratteristiche dei dati. Questa scelta comporta la non esaustività dei controlli: solo un sottoinsieme dei dati, quelli con maggiore impatto sulle stime dei dati, è sottoposto a controllo e correzione.
Questa soluzione può però essere abbinata all'applicazione di software generalizzato per la restante parte dei dati, dando luogo alla **quinta configurazione**, che da un punto di vista qualitativo risulta essere la più avanzata.

Giova sottolineare che, laddove l'ambiente elaborativo permetta lavorazioni di tipo interattivo, per quanto riguarda la fase di **acquisizione dei dati** è di enorme importanza il tipo di data entry che si effettua, se con controlli o meno. L'esperienza ha dimostrato che la quantità di errori introdotti nella fase di registrazione dei modelli spesso non è inferiore a quella dovuta alla loro compilazione: la minimizzazione di tale componente attraverso l'utilizzo di programmi contenenti controlli sul dominio delle variabili e regole di compilazione e di compatibilità è dunque estremamente rilevante. Infatti, un conto è procedere all'individuazione ed alla rimozione degli errori, con il ripristino dei valori "veri" nelle variabili controllate, con una percentuale di successo talvolta drammaticamente bassa, un altro è operare per fare in modo che tali errori abbiano meno possibilità di verificarsi. Una situazione ottimale si ha quando si può fare a meno del questionario cartaceo, e le domande vengono poste direttamente all'interessato utilizzando il computer: in tal caso, se il programma che presiede alla registrazione delle risposte effettua controlli, è possibile, in caso di incompatibilità tra le risposte, dirimere le controversie direttamente col rispondente, minimizzando in tal modo anche gli errori di compilazione. Queste tecniche, rientranti nel

filone del CASIC (Computer Assisted Survey Information Collection), iniziano ad essere utilizzate da diversi enti produttori di statistiche, ufficiali o privati, per le operazioni di raccolta dati: realisticamente, un ente amministrativo non può pensare di introdurre tali tecniche solo per esigenze di tipo statistico. Se però nel processo amministrativo di raccolta dati è già previsto l'uso diretto dei computer, lo statistico può intervenire per far introdurre nei programmi tutti i possibili controlli, simulando la tecnica nota come CAPI (Computer Assisted Personal Interviewing).

Nello schema che segue riportiamo le valutazioni da compiere per giungere ad una determinata configurazione della fase di acquisizione dati.

SCHEMA PER LA SCELTA DELLA CONFIGURAZIONE DELLA FASE DI ACQUISIZIONE DEI DATI



La **prima configurazione** è quella più avanzata, di tipo CAPI: essa è realisticamente possibile per l'ente amministrativo senza costi aggiuntivi, se già le operazioni di raccolta dati vengono effettuate utilizzando computer con sistemi on-line. E' la situazione ottimale: i controlli permettono di contenere al massimo gli errori alla fonte.

La **seconda configurazione** è quella di tipo tradizionale, *data entry "heads down"*, la peggiore da un punto di vista qualitativo, cui si è costretti a ricorrere in caso di grande mole di dati e/o limitata disponibilità di personale.

La **terza** soluzione è invece intermedia rispetto alle precedenti: non ha influenza sugli errori che si verificano al momento della compilazione del modello, ma è in grado di contenere quelli dovuti alla registrazione.

Conclusa la parentesi relativa ai controlli da effettuare al momento della raccolta dei dati, nei paragrafi successivi entreremo in dettaglio relativamente alle caratteristiche delle procedure di tipo interattivo o automatico per il controllo e la correzione degli errori nella fase successiva all'acquisizione delle informazioni.

Per quanto riguarda le prime, descriveremo i principali metodi utilizzati, tra cui, oltre i già citati filoni del macroediting e dell'editing selettivo, considereremo anche le tecniche di tipo grafico.

Per quanto concerne invece le seconde, tratteremo della differenza tra l'approccio deterministico e quello non deterministico (o probabilistico), illustrando i principali software generalizzati che fanno riferimento al secondo.

6.1. Procedure interattive di controllo e correzione

Le procedure interattive possono essere utilizzate nelle seguenti fasi del processo produttivo dell'informazione statistica o amministrativa:

1. *raccolta dei dati*;
2. *registrazione dei dati*;
3. *editing dei dati*.

Nel primo caso le tecniche interattive di controllo sono generalmente integrate in sistemi software di acquisizione controllata dei dati (CATI o CAPI), per cui la verifica della correttezza dei dati avviene contestualmente all'intervista delle unità.

Nel secondo caso, le procedure di controllo sono incorporate nei programmi di registrazione dei dati, e sono quindi eseguite contestualmente a tale fase.

In fase di editing, l'applicazione delle procedure interattive può essere rivolta:

1. alla localizzazione ed alla correzione degli errori che si configurano come fuori dominio, incompatibilità, mancate risposte totali e parziali;
2. alla localizzazione ed alla correzione degli outlier.

Le procedure interattive possono essere:

1. *interamente interattive*, in cui sia la fase di localizzazione, sia la fase di correzione degli errori è basata sull'intervento umano;
2. *miste*, in cui la fase di localizzazione degli errori è di tipo automatico, e la fase di correzione è (parzialmente o interamente) interattiva.

Possiamo inoltre distinguere fra procedure interattive:

1. di tipo *micro* (in cui cioè tutte le osservazioni in errore vengono sottoposte a controllo interattivo);
2. di tipo *macro* o *selettivo* (in cui solo le unità errate con impatto significativo sulle stime finali vengono sottoposte a verifica interattiva);
3. di tipo *misto* (in cui le unità errate con impatto significativo sulle stime finali vengono sottoposte a verifica interattiva e le rimanenti unità sono corrette in modo automatico).

Le procedure di tipo *macro* e di tipo *selettivo* sono normalmente applicate solo nel caso di variabili quantitative. Le tecniche di tipo *macro* possono essere:

- *univariate*, quando la localizzazione degli errori avviene considerando una variabile per volta;
- *multivariate*, quando le osservazioni errate vengono individuate tenendo conto congiuntamente di un insieme di variabili rilevate e verificando le relazioni esistenti fra esse.

Nell'ambito delle procedure *macro* vanno ancora distinte le procedure di tipo *grafico*, in cui l'individuazione delle osservazioni errate avviene mediante visualizzazione su diagrammi, grafici o tabelle dei valori delle variabili rilevate sulle unità.

Le procedure interattive possono riguardare *variabili qualitative* oppure *quantitative*, possono essere *sviluppate ad hoc*, oppure possono essere implementate in *software generalizzato*.

6.1.1. Il microediting interattivo

L'editing interattivo tradizionale è di tipo *micro*, prevede cioè che tutte le unità elementari oggetto di indagine siano sottoposte a controllo e correzione interattiva. Ciò implica che tutti gli errori presenti nei dati sono considerati aventi la stessa importanza ed a tutti viene riservato lo stesso tipo di trattamento, indipendentemente dall'impatto che l'eventuale correzione del dato errato avrà sulle stime finali. E' questa un'altra forma in cui si può manifestare il fenomeno che nell'introduzione abbiamo chiamato *over-editing*, cioè un eccessivo numero di correzioni sui dati non giustificato da un apprezzabile miglioramento della qualità dei risultati.

Mentre in origine le procedure interattive erano interamente di tipo manuale, cioè prevedevano che entrambe le operazioni di localizzazione e di correzione degli errori fossero il risultato dell'interazione fra operatori e dati, prevalgono attualmente procedure interattive di tipo *misto*, automatiche per la localizzazione degli errori e interattive per la loro correzione.

In generale, una procedura di controllo interattiva può essere utilizzata nelle seguenti attività, in cui è presente l'intervento dell'operatore:

1. controllo manuale dei dati al momento della revisione dei modelli;
2. controllo e correzione interattiva dei dati all'atto della registrazione;
3. editing in cui l'integrazione dei dati mancanti e la correzione degli errori segnalati da un opportuno piano di incompatibilità avviene interattivamente mediante reintervista o verifica dei modelli cartacei.

I principali inconvenienti che possono derivare dall'utilizzo di procedure di correzione siffatte possono essere riassunti nei punti seguenti:

- soggettività delle correzioni laddove non sia possibile ricorrere alla reintervista o non siano reperibili i modelli cartacei (rischio di *editing creativo*);
- costi elevati in termini di risorse umane e notevole carico sui rispondenti;

- tempi elevati per grosse moli di dati;
- non riproducibilità del processo;

Il microediting interattivo è garanzia di ottima qualità dei risultati finali e di risparmio in termini di tempi e costi laddove esso venga effettuato in fase di intervista (CATI o CAPI) oppure in fase di registrazione dei dati: in questi casi, infatti, esso assicura l'eliminazione della maggior parte degli errori di codifica, di registrazione, di dominio, di percorso contestualmente all'intervista oppure quando le informazioni sono reperibili senza bisogno di attività di ricerca dei modelli cartacei.

Qualora adottato in fase di controllo e correzione dei dati, l'approccio del microediting interattivo è in generale efficace nelle situazioni in cui:

- le risorse ed i tempi lo rendano praticabile in rapporto alla mole di dati da trattare;
- si disponga di software (ad hoc o generalizzato) con prefissate proprietà statistico-matematiche;
- sia possibile il reperimento dell'informazione corretta (mediante reintervista o recupero dei modelli);
- sia possibile l'utilizzo di informazioni ausiliarie o storiche.

6.1.1.1. Procedure generalizzate per il microediting

Le procedure di controllo e correzione di tipo interattivo sono generalmente implementate in software ad hoc: per ogni particolare processo di produzione dell'informazione statistica e per ogni approccio prescelto (microediting, macroediting, editing selettivo) vengono messe a punto procedure automatiche che implementano il particolare algoritmo di localizzazione dei record errati adottato, record che vengono quindi sottoposti interattivamente a verifica ed eventuale correzione.

Sono stati però sviluppati alcuni sistemi generalizzati per il trattamento interattivo dei dati, nei quali, fissata la metodologia di localizzazione dei dati errati o sospetti, si ha una standardizzazione delle fasi e delle operazioni componenti il processo di controllo e correzione. Questi sistemi, se da un lato implicano grossi costi iniziali per la loro realizzazione, dall'altro producono sia l'eliminazione di duplicazioni e ridondanze di software, sia una omogeneità di trattamento per le indagini cui vengono applicati. Anche se devono essere in ogni caso adattati alla particolare indagine oggetto di verifica, in essi definizioni, operazioni e parametri sono fissati e standardizzati.

In alcuni casi questi sistemi consentono tra l'altro l'integrazione del processo di editing interattivo nelle fasi di raccolta o di registrazione dei dati (ad esempio, Blaise). Infatti, qualora i costi e le risorse umane lo consentano, tali sistemi possono essere utilizzati sia per la conduzione di interviste (dirette o telefoniche) assistite da computer, sia come strumenti per l'acquisizione controllata dei dati, con tutti i vantaggi in termini di efficienza e di qualità già illustrati.

Nel seguito vengono illustrate le caratteristiche e le potenzialità di alcuni sistemi generalizzati per l'editing interattivo dei dati, Blaise per il trattamento di variabili qualitative, SPEER e ARIES per l'analisi di variabili quantitative.

BLAISE

Il sistema BLAISE, sviluppato dall'Istituto nazionale di statistica olandese (Statistics Netherlands), è un classico strumento che permette l'applicazione delle tecniche CAPI e CATI, ma che può anche essere utilizzato per l'acquisizione controllata dei dati, e per la correzione interattiva di questi.

La sintassi di BLAISE permette, una volta definita la struttura del questionario ed i quesiti che ne fanno parte, di associarvi delle regole di controllo, ognuna delle quali viene applicata ogni volta che il cursore giunge all'ultima delle variabili interessate. Ad esempio:

(ANNO_RILEVAZIONE - ANNO_NASCITA ≥ 23) AND TITOLO_STUDIO = 1

"La persona ha conseguito il dottorato e deve avere almeno 23 anni"

In questa regola, che stabilisce una condizione di *correttezza*, vengono coinvolte tre variabili: l'anno cui si riferiscono i dati, l'anno di nascita della persona, ed il suo titolo di studio. Se le variabili compaiono in questa sequenza nel questionario, ogni volta che si giunge ad inserire un valore in TITOLO_STUDIO, o comunque se il cursore supera il campo in questione, il sistema valuta se la regola è rispettata, oppure no. In questo ultimo caso, viene proposto all'operatore il messaggio racchiuso tra le virgolette, unitamente all'elenco delle tre variabili e dei valori che assumono nel questionario. L'operatore può scegliere quale/i variabile/i correggere, e quale/i valore/i assegnare ad essa/e, oppure può decidere di ignorare la segnalazione di errore, proseguendo nell'attività di memorizzazione dati, o di controllo interattivo.

BLAISE opera nell'ambiente DOS dei personal computer, anche collegati in LAN, e gode di una larga diffusione, soprattutto all'interno degli Istituti nazionali di statistica. Uno dei motivi di tale diffusione risiede nella grande facilità di sviluppo di programmi utilizzabili, come già detto, sia per le applicazioni CAPI/CATI, che per il data entry e per il controllo interattivo: essendo un sistema nato a tal fine, può essere utilizzato anche da non esperti programmatori, in quanto tutto ciò che si richiede è la descrizione del questionario attraverso l'indicazione dei quesiti coi relativi domini associati, e con le regole di compilazione e di compatibilità.

Un'applicazione BLAISE utilizzabile per il controllo interattivo comporta la disponibilità dei dati memorizzati secondo varie possibilità, la conversione del file originale in un file leggibile da BLAISE, la suddivisione dei record in esatti e errati mediante l'applicazione di particolari utilities, ed un'attività interattiva che vede il singolo operatore intervenire sui singoli modelli errati seguendo le indicazioni fornite dal programma BLAISE contenente la descrizione del modello con le varie regole di compatibilità.

SPEER

Nel paragrafo 5.2.2.3 sono descritte la struttura e le funzionalità del sistema SPEER, nell'ambito dei metodi automatici generalizzati per il controllo e la correzione dei dati. In realtà, il sistema è utilizzabile anche interattivamente: in questo caso, la correzione dei record che non soddisfano i controlli del prefissato piano di incompatibilità non è affidata al programma, il quale si limita, per ogni variabile da imputare, a suggerire all'operatore i possibili valori da assegnare, valori prodotti dai vari metodi previsti nel sistema. La scelta di quali valori assegnare è lasciata alle valutazioni soggettive dell'esperto, il quale procede alle correzioni in modo interattivo.

L'utilizzo in senso interattivo del sistema SPEER produce, rispetto alla modalità di utilizzo interamente automatica, da un lato un maggiore utilizzo di risorse umane per il controllo interattivo dei record, dall'altro una notevole riduzione o addirittura l'eliminazione del processo ciclico generato dalla versione batch del sistema.

E' evidente che l'utilizzo in modalità interattive del programma risulta più efficiente rispetto alla soluzione di tipo batch solo nel caso in cui si disponga dei modelli cartacei, oppure qualora sia possibile ricontattare le unità rispondenti.

SPEER può essere utilizzato in modalità interattiva non solo nella fase di controllo e correzione dei dati, ma anche in fase di registrazione, avendo a disposizione tempi e risorse umane sufficienti: in questo caso, i record verrebbero controllati e corretti direttamente all'atto della registrazione.

6.1.2. Il macroediting, l'editing grafico e l'editing selettivo

Il problema del controllo e della correzione di variabili di tipo quantitativo, oltre che con tecniche di tipo micro, in cui tutte le osservazioni rilevate vengono sottoposte a controllo e correzione, può essere affrontato utilizzando metodologie che consentono di concentrare le operazioni di verifica sulle sole unità statistiche con impatto consistente sulle stime finali. Questi approcci, noti come *macroediting* e *editing selettivo*, producono al contempo:

- una riduzione dei costi e del "carico" sui rispondenti (essendo gli eventuali ritorni limitati a poche unità);
- un aumento dell'efficienza del processo di produzione in termini di tempestività;
- una diminuzione non significativa dell'efficacia in termini di qualità delle stime prodotte;
- una riduzione del fenomeno dell'*over-editing*, cioè delle operazioni di verifica che, riguardando dati con un impatto trascurabile sulle stime finali, producono un eccessivo e spesso superfluo impiego di risorse (umane e di tempo).

In termini generali, le strategie di editing alla base del *macroediting* e dell'*editing selettivo* prevedono un ritorno selettivo sulle unità con mancate risposte o con incongruenze rispetto alle regole di controllo utilizzate nel piano di incompatibilità: tali unità vengono selezionate sulla base di *funzioni*, *punteggi* o *indici* che misurano sostanzialmente l'impatto delle unità sulle quantità stimate.

Queste tecniche sono tutte di tipo *interattivo*: il controllo e l'eventuale correzione dei dati riconosciuti come anomali sono basate sulla revisione dei modelli cartacei, su tecniche di *reintervista*, su valutazioni soggettive di esperti.

Tipicamente, gli approcci del *macroediting* e dell'*editing selettivo* sono stati sviluppati per il trattamento di dati rilevati in indagini di tipo *periodico* (censuarie o campionarie, meglio se di tipo panel) *con elevata frequenza di ripetizione*: è per queste indagini, infatti, che risultano più importanti gli aspetti della *tempestività* dei risultati finali e della ottimizzazione dei *costi* e dei *tempi* di revisione e correzione dei dati. Queste tecniche si sono dimostrate tanto più efficaci quanto minore è la periodicità dell'indagine per due motivi principali:

1. il numero di verifiche e di correzioni interattive è generalmente molto basso rispetto ai tradizionali approcci di tipo *micro*, siano esse interattive o automatiche;
2. poiché tali tecniche prevedono in generale l'uso di informazioni storiche, la disponibilità di dati storici "recenti" rende più affidabili i risultati della loro applicazione, essendo più probabile che forti variazioni verificatesi in una certa variabile in un breve intervallo di tempo siano dovute a errore piuttosto che al naturale evolversi del fenomeno stesso.

Benché pensati per il trattamento di indagini di tipo periodico, tali metodi sono comunque generalizzabili anche a situazioni in cui non si disponga di informazioni storiche o queste non siano affidabili.

Oltre che la produzione di stime di qualità, obiettivo finale delle indagini è generalmente garantire il rilascio di dati corretti a livello elementare: in questi casi, è necessario che all'applicazione delle tecniche di tipo *macro* o di tipo *selettivo* faccia seguito l'uso di procedure automatiche di tipo *micro* per l'editing e la correzione delle rimanenti unità elementari[5].

Nell'ottica generale di ottimizzazione dei tempi, dei costi e delle risorse che abbiamo già sottolineato, la strategia di editing da adottare nel caso di variabili quantitative dovrebbe prevedere:

1. il follow-up completo nei casi di mancata risposta totale;
2. l'utilizzo di tecniche di macroediting o di editing selettivo per l'individuazione, il controllo e l'eventuale correzione interattiva delle unità con impatto importante sulle stime finali delle variabili di interesse;
3. la correzione automatica dei casi residui tramite un sistema automatico generalizzato probabilistico (del tipo di GEIS o SPEER) al fine di garantire file coerenti a livello di dati elementari.

6.1.2.1. Il macroediting

Per l'identificazione dei dati elementari sospetti con impatto significativo sulle stime finali, il macroediting parte dall'idea di considerare preventivamente particolari sotto insiemi (aggregati) dei dati stessi e, all'interno di quelli con comportamento sospetto in base ad un prefissato criterio, procedere alla localizzazione dei record da sottoporre a controllo interattivo. A questo punto, si procede alla verifica dei dati elementari sospetti a partire dai più importanti, fermandosi quando ulteriori correzioni non producono più alcun effetto significativo sulle stime finali.

Tutti i metodi afferenti al macroediting sono basati su particolari distribuzioni dei dati grezzi, quindi non possono essere utilizzati in fase di registrazione: ciò non ha effetti sulla tempestività, in quanto il pregio di questi metodi è che pochi dati sono generalmente segnalati per la verifica manuale.

Sperimentazioni e implementazioni di questi metodi hanno dimostrato che il risparmio operativo (in termini di riduzione del numero di correzioni effettuate) che si consegue rispetto alle tradizionali procedure di tipo micro va dal 30% all'80%, senza effetti rilevanti sulla qualità delle stime finali.

Le caratteristiche principali delle tecniche di tipo *macro*, oltre all'essere di tipo interattivo e di poter essere applicate alle sole variabili quantitative, possono essere così sintetizzate:

. analisi guidata dagli aggregati;
. analisi guidata dall'importanza delle unità sulle stime finali;
. uso di dati storici;

Nell'ambito del macroediting possiamo distinguere due diverse tipologie di metodi:

1. . *tecniche di tipo univariato*;

[5] In questo ambito i risultati migliori sono garantiti dall'uso di procedure di tipo probabilistico.

2. . *tecniche di tipo multivariato*.

Nel primo caso, dato un insieme di variabili rilevate di interesse, l'editing dei dati è basato sul controllo di una variabile per volta: il controllo della correttezza dei dati consiste nella localizzazione delle unità elementari *più importanti*, quelle cioè che presentano, rispetto alle altre unità rispondenti, una concentrazione significativa del carattere stesso e che, quindi, hanno un impatto maggiore sulle stime finali (totali, medie, ecc.).

Nel caso multivariato, invece, sono considerate unità anomale quelle che non solo hanno un impatto significativo sulle stime finali delle variabili di interesse, ma presentano relazioni sospette fra tali variabili: l'editing dei dati avviene in questo caso congiuntamente rispetto alle variabili da sottoporre a controllo. E' ovvio che, nel caso di outlier definiti in ambito multivariato, sia la loro individuazione, sia la loro interpretazione risultano più complesse rispetto al caso univariato.

Inoltre, mentre nel caso univariato gli outlier risiedono nelle code della distribuzione delle funzioni di controllo, nel caso multivariato i valori sospetti possono essere localizzati ovunque nella nube di punti delle osservazioni.

Un'altra differenza fra l'approccio univariato e quello multivariato al problema della localizzazione degli outlier risiede nella tecnica di individuazione delle osservazioni sospette: mentre le tecniche di tipo univariato ricorrono sostanzialmente all'uso delle cosiddette *funzioni di controllo* (rapporti, differenze, altre funzioni statistico-matematiche più o meno complesse), nei metodi multivariati le osservazioni sospette vengono individuate sulla base di *metodi di analisi multivariata* generalmente combinati con metodologie proprie della *teoria di verifica di test di ipotesi statistiche*.

6.1.2.1.1. Metodi univariati

In generale, nell'approccio *univariato* la selezione delle unità anomale avviene in tre fasi:

1. definizione dei domini di interesse (classi individuate sulla base di opportune variabili di classificazione);
2. aggregazione dei valori elementari correnti all'interno dei domini;
3. calcolo di opportune *funzioni di controllo* per il confronto fra tali valori ed opportuni valori di riferimento;
4. definizione di una o più regioni di accettazione sulla base delle funzioni di controllo per l'individuazione delle unità da sottoporre a verifica.

Pur essendo stato sviluppato stato sviluppato originariamente per il trattamento di dati rilevati in indagini di tipo periodico, nel caso si utilizzino funzioni di controllo che sfruttano le informazioni relative al solo periodo di rilevazione l'approccio del *macroediting* può essere applicato anche ad indagini di tipo non periodico.

Pertanto, una distinzione importante fra i metodi *univariati* afferenti al *macroediting* va collegata alla disponibilità o meno di *dati storici* per il fenomeno in esame. Per ogni variabile di interesse, quindi, i valori di riferimento di cui al punto 3 possono essere rappresentati in alternativa:

- dai corrispondenti valori relativi ad una precedente realizzazione dell'indagine (indagini censuarie o campionarie di tipo *panel*, cioè indagini in cui *uno stesso insieme di unità* è sottoposto periodicamente a rilevazione);

- dai valori presenti nelle altre unità rilevate o loro sintesi, oppure con dati provenienti da fonti esterne.

E' importante sottolineare che, nel caso si applichino tecniche *univariate* ad indagini di tipo campionario (panel o non panel), poiché le unità campione possiedono diverse probabilità di inclusione, è necessario ponderare la/le funzioni di controllo in accordo col disegno di campionamento, al fine di tenere conto del diverso impatto di ogni unità statistica sul/sui fenomeni di interesse.

Di seguito verranno brevemente illustrati i principali metodi di tipo *macro* attualmente esistenti, basati su approcci statistico/matematici e/o grafici per la determinazione dei record *outlier*.

Il metodo Aggregato

L'idea di base di questo metodo è quella di controllare i dati a livello aggregato, e di procedere poi al controllo ed alla correzione dei soli dati elementari che danno luogo ad aggregati *sospetti*.

Gli aggregati vengono calcolati all'interno di domini determinati sulla base delle modalità di una o più variabili di classificazione (ad esempio, per un'indagine che rilevi informazioni sulle imprese, *attività economica*, *classe di addetti*, *ripartizione territoriale*, ecc.): il metodo, quindi, procede all'interno di ogni variabile di interesse evidenziando a quali particolari aggregati è da attribuirsi l'eventuale anomalia della variabile stessa.

L'individuazione degli elementi sospetti avviene mediante il calcolo di prefissate *funzioni di controllo.*

A livello aggregato, per ogni variabile di interesse viene calcolata la funzione di controllo (fc1) sui valori aggregati: tale funzione restituisce una lista ordinata di valori sulla base della quale viene definita empiricamente la regione di accettazione[6]. Gli aggregati cui sono associati valori della fc1 esterni alla regione di accettazione sono considerati sospetti.

La funzione di controllo a livello elementare (fc2) viene calcolata di volta in volta sui record elementari che contribuiscono alla formazione dell'aggregato sospetto. Ordinata la lista dei valori, viene definita empiricamente la regione di accettazione, e gli elementi corrispondenti a valori della fc2 esterni a tale regione vengono controllati ed eventualmente corretti. Il processo termina appena la correzione di un record consente all'aggregato ad esso associato di rientrare nella regione di accettazione.

Le fasi critiche nella predisposizione di una strategia di controllo e correzione dati basata sul metodo *Aggregato* possono allora essere riassunte nei punti seguenti:

1. scelta della/e variabili di classificazione e dei domini di interesse;
2. scelta della/e funzioni di controllo a livello di valori aggregati e di microdati;
3. scelta della/e soglie per la delimitazione della regione di accettazione a livello *aggregato* (ampiezza delle code della distribuzione di fc_1);
4. scelta della/e soglie per la delimitazione della regione di accettazione a livello di microdati (ampiezza delle code della distribuzione di fc_2).

[6] Generalmente la *regione di rifiuto*, a livello sia aggregato che elementare, è costituita dalle code della distribuzione della funzione di controllo.

Un esempio di funzioni di controllo, rispettivamente a livello aggregato ed elementare, è dato dalle seguenti espressioni[7]:

$$fc_{1kj} = \frac{\left| A_{kj}(t) - A_{kj}(t-1) \right|}{\left| \mathrm{Min}\left[A_{kj}(t), A_{kj}(t-1) \right] \right|}, \quad j=1,2,...,m \qquad [1]$$

dove $A_{kj}(t)$ e $A_{kj}(t-1)$ sono i valori aggregati per il dominio k e la variabile j rispettivamente ai tempi t e (t-1).

$$c_{2ijr} = \left| V_{kji}(t) - V_{kji}(t-1) \right|, \quad i=1,2,...,n \; ; \; j= 1,2,...,m \qquad [2]$$

dove $V_{kji}(t)$ e $V_{kji}(t-1)$ sono, rispettivamente, i valori corrente e storico della variabile j nell'unità i del dominio k.

Il metodo Top – Down

A differenza del *metodo Aggregato*, il metodo *Top-Down* (Granquist 1990b, 1992c) non prevede alcun controllo a livello *aggregato*: con tale tecnica si procede direttamente al controllo dei dati elementari all'interno dei domini di interesse. L'individuazione delle unità statistiche *sospette*, fra quelle che hanno maggior impatto sulla determinazione delle stime, avviene mediante l'uso di una o più funzioni di controllo che agiscono direttamente a livello di *microdati*.

Come per il *metodo Aggregato*, la localizzazione degli outlier avviene separatamente per variabili e per domini:

1. definizione dell'insieme di funzioni di controllo che si intende utilizzare (eventualmente ponderate in accordo col disegno campionario);
2. per una data variabile, calcolo dei valori della funzione di controllo prescelta sui dati elementari dei domini di interesse;
3. ordinamento dei valori del codominio di tale funzione e formazione di una lista;
4. controllo interattivo dall'inizio o dalla fine della lista;
5. ad ogni correzione effettuata, ricalcolo delle stime: il processo di correzione ha termine appena si ottiene un miglioramento trascurabile nella determinazione della stima.

Il procedimento va ripetuto per ogni variabile e per ogni dominio di interesse.

In una sperimentazione effettuata dall'Istituto di Statistica svedese Statistics Sweden, il metodo è stato applicato all'indagine su Fatturato e Ordinativi utilizzando funzioni di controllo che calcolano, per ogni record e per ognuna delle variabili di interesse:

. le 15 maggiori variazioni positive;
. le 15 maggiori variazioni negative;
. i 15 maggiori contributi alla stima dell'aggregato.

In ognuno di questi casi, l'operatore deve controllare ed eventualmente correggere interattivamente 15 record: per ogni correzione effettuata, il sistema riapplica la funzione di controllo e ridetermina il nuovo valore della stima. Il record corretto esce dalla lista dei

[7] Queste funzioni sono state utilizzate in un'applicazione del metodo aggregato effettuata in ISTAT sui dati dell'indagine sul Sistema dei Conti delle Imprese (Barcaroli, Ceccarelli, Luzi, 1995).

sospetti, ed un altro record vi entra: il processo di editing ha termine quando ulteriori correzioni hanno un effetto trascurabile sulle stime.

La funzione di Hidiroglou-Berthelot

La procedura ideata da Hidiroglou-Berthelot (H&B nel seguito) appartiene a quella classe di metodi di editing basati sull'uso di controlli in forma di rapporti (*ratio methods*).

In generale, il metodo consente di isolare quelle incompatibilità nei dati che si manifestano come divergenza dei valori di alcune variabili, rilevate in certe unità in una data indagine in un dato periodo, rispetto ai valori che le stesse variabili presentavano nelle stesse unità e nella stessa rilevazione effettuata però in un periodo precedente[8].

In realtà, il metodo H&B può essere utilizzato anche in assenza di dati storici, allo scopo di individuare quelle unità che presentano valori inconsistenti nell'ambito di una stessa rilevazione (valori cioè che si discostano in modo significativo dai valori che le stesse variabili assumono nel resto delle unità rilevate). In particolare, il metodo può essere utilizzato per l'editing di indagini non periodiche in due modi distinti:

1. confrontando i valori dei rapporti calcolati fra due variabili correlate fra loro all'interno di una data popolazione;
2. confrontando i valori di una data variabile con limiti di accettabilità calcolati sulla base dei valori che la variabile assume in altri record in uno stesso periodo di riferimento (Cotton, 1991);

La procedura H&B adotta una funzione di controllo piuttosto complessa, che può essere applicata indifferentemente sia nel microediting (Kovar, MacMillian, Whitridge, 1988) sia nel macroediting (Hidiroglou, Berthelot, 1986): in quest'ultimo caso, essa può essere utilizzata come una particolare funzione di controllo nell'ambito di altre metodologie (ad esempio, nei metodi Aggregato e Top-Down), sia come metodo a sé stante.

La funzione H&B utilizza tre parametri: l'opportuna "calibrazione" di tali parametri permette di controllare l'ampiezza della regione di accettazione e, quindi, di decidere variabile per variabile quale tipologia di variazioni considerare anomale rispetto all'entità del fenomeno oggetto di studio (controllando anche l'importanza da associare alle unità in base alla loro "grandezza", eventualmente dando più importanza a relativamente piccole variazioni in grandi unità che a relativamente grandi variazioni in piccole unità).

In sintesi, l'algoritmo di localizzazione delle osservazioni anomale rispetto ad una data variabile di interesse X prevede le seguenti fasi:

1. per ogni unità i calcolo del *tasso di variazione*

$$r_i = x_i(t+1) / x_i(t)$$

dove $x_i(t+1)$ e $x_i(t)$ sono rispettivamente i valori corrente e storico della variabile X nell'unità i.

[8] In realtà, la procedura generale di editing proposta da H&B prevede anche controlli di incompatibilità, cioè controlli che verifichino che combinazioni lineari fra variabili rilevate in una stessa unità campionaria in un dato momento soddisfino certi requisiti (Hidiroglou-Berthelot, 1986).

2. Per trattare in modo equilibrato sia le variazioni positive che quelle negative, trasformazione degli r_i in nuovi valori S_i :

$$S_i = \begin{cases} 1 - r_{mediana} / r_i & \text{se } 0 < r_i < r_{mediana} \\ r_i / r_{mediana} - 1 & \text{se } r_i >= r_{mediana} \end{cases}$$

dove $r_{mediana}$ è la mediana dei rapporti: in tal modo metà dei valori di s_i risultano positivi e l'altra metà negativi.

3. Per tener conto della diversa "grandezza" delle osservazioni, trasformazione dei valori S_i nei nuovi valori E_i :

$$E_i = s_i \times \{ MAX [x_i(t), x_i(t+1)] \}^U$$

dove U è un parametro compreso tra 0 e 1. E_i è detto *effetto* associato all'i-esima unità, e l'esponente U consente di controllare l'importanza da associare alla "grandezza" dell'unità stessa. In pratica, questa trasformazione consente di dare più importanza a relativamente piccole variazioni in grandi valori rispetto a grandi variazioni in piccoli valori: quando U=0 non si dà alcuna importanza alla dimensione dei valori, se U=1 si dà ad essi la massima importanza. Gli E_i sono distribuiti attorno allo zero (in modo tanto più disperso quanto più alto è il valore di U): gli E_i troppo piccoli o troppo alti sono considerati come possibili outlier.

4. Definizione dei limiti della regione di accettazione per i valori E_i :

4.1. calcolo delle deviazioni :

$$d_{Q1} = MAX \{ E_{mediana} - E_{Q1} , A*E_{mediana} \}$$

$$d_{Q3} = MAX \{ E_{Q3} - E_{mediana} , A*E_{mediana} \}$$

dove E_{Q1}, $E_{mediana}$, E_{Q3} sono, rispettivamente, il primo quartile, la mediana ed il terzo quartile della distribuzione degli E_i . Il termine $A*E_{mediana}$ è una protezione contro la possibilità di trovare troppi outlier quando gli E_i sono concentrati attorno ad un singolo valore con troppe poche deviazioni, cioè quando gli intervalli $[E_{mediana}-D_{Q1}]$, $[D_{Q3}-E_{mediana}]$ sono troppo piccoli (viene normalmente utilizzato il valore A=0.05). Per la costruzione degli intervalli di accettazione il metodo utilizza i quartili al posto delle deviazioni standard per impedire che gli outlier abbiano un peso troppo importante sul calcolo degli intervalli stessi.

4.2. Definizione dei limiti inferiore (L) e superiore (U) della regione di accettazione come segue:

$$L = E_{mediana} - C \times D_{Q1}$$

$$U = E_{mediana} + C \times D_{Q3}$$

in cui la costante C permette di ampliare o restringere l'ampiezza dell'intervallo di accettazione. Tutte le unità in cui i corrispondenti effetti E_i assumono valori esterni all'intervallo [L, U]sono da considerare outlier.

La stima dei parametri della funzione H&B, e, quindi, delle soglie di delimitazione della regione di accettazione, viene effettuata, variabile per variabile, per tentativi, osservando i risultati ottenuti per le diverse combinazioni dei valori dei parametri (n° di outlier correttamente ed erroneamente identificati, impatto degli outlier identificati sulle stime finali).

Un procedimento sperimentato in Svezia da Statistics Sweden nell'indagine su "Fatturato e Ordinativi" (Davila, 1992) è basato sul calcolo, per ogni combinazione dei parametri, della cosiddetta probabilità empirica di corretta classificazione degli outlier: in questo caso è stato imposto che il processo di correzione degli outlier proseguisse finché le stime finali calcolate sui dati non fossero risultate stabili (cioè fino a quando ulteriori correzioni non avessero più alcun impatto su tali stime).

In termini di applicabilità, vanno ricordate alcune circostanze che devono essere tenute presenti prima di procedere all'editing di dati mediante il metodo H&B:

- il metodo prevede l'elaborazione di una variabile per volta, per cui ne è consigliabile l'uso o su indagini che rilevino poche variabili, oppure nel caso si intenda sottoporre a controllo un numero limitato di variabili rilevate (ad esempio, le sole variabili strategiche o quelle di maggior interesse);
- poiché il metodo è basato sull'uso di rapporti, esso risente in modo consistente della presenza di zeri e di valori mancanti: è quindi opportuno, prima di procedere alla sua applicazione, verificare l'entità di valori nulli e di missing nei dati.

6.1.2.1.2. Metodi multivariati

Le metodologie sviluppate nell'ambito del *macroediting multivariato* presentano la caratteristica comune di prevedere l'individuazione, la verifica e l'eventuale correzione dei record del data-set che, oltre ad avere un *impatto significativo sulle stime finali*, presentano anche *relazioni sospette fra un prefissato insieme di variabili*. In questo caso, quindi, l'editing dei dati viene effettuato *congiuntamente* per l'insieme prefissato di variabili di interesse.

E' ovvio che, nel caso di outlier definiti in ambito multivariato, sia la loro individuazione, sia la loro interpretazione risultano più complesse rispetto al caso univariato.

Inoltre, mentre nel caso univariato gli outlier risiedono nelle code della distribuzione delle funzioni di controllo, nel caso multivariato i valori sospetti possono essere localizzati ovunque nella nube di punti delle osservazioni.

In questo approccio ogni unità statistica, in quanto portatrice di informazioni su ognuna delle n variabili rilevate, può essere rappresentata nello spazio m dimensionale dal vettore $X_i=(x_{i1}, x_{i2}, ..., x_{im})$, dove i valori , x_{i1},,x_{i2}, ..., x_{im} rappresentano i valori delle variabili 1,2,...,m.

In generale, le procedure di controllo di tipo multivariato prevedono l'effettuazione dei seguenti passi:

1) identificazione di *domini* di interesse all'interno dei quali effettuare analisi separate al fine di controllare la variabilità delle relazioni fra le variabili rispetto a fattori noti;
2) definizione e calcolo di un particolare *indice di distanza* fra i vettori di valori elementari e opportuni valori di riferimento;
3) definizione del *limite della regione di accettazione a livello di unità* (in base alla distribuzione dell'indice di distanza utilizzato) per l'individuazione delle osservazioni da considerare sospette;

4) definizione del *limite della regione di accettazione a livello di variabile* per l'individuazione delle variabili da sottoporre a verifica all'interno di ogni record outlier.

Analogamente al caso *univariato*, i metodi *multivariati* possono essere applicati ad indagini di tipo sia periodico (censuarie o campionarie di tipo panel) sia non periodico.

Nel caso si disponga di *dati storici* per il/i fenomeno/i in esame, tali informazioni possono essere utilizzate per "ponderare" i valori correnti nel calcolo delle distanze.

Nel caso si applichino tecniche *multivariate* ad indagini di tipo campionario (panel o non panel), poiché le unità campione possiedono diverse probabilità di inclusione, è necessario ponderare la/le funzioni di distanza in accordo col disegno di campionamento, al fine di tenere conto del diverso impatto di ogni unità statistica sul/sui fenomeni di interesse.

Per il tipo di controllo che prevedono sui dati, le tecniche multivariate sono particolarmente adatte in quei casi in cui è importante verificare particolari relazioni fra alcune variabili, come nel caso delle indagini che rilevano bilanci.

Di seguito verrà illustrato un metodo di tipo *multivariato* basato sull'uso congiunto dell'analisi in componenti principali e dell'indice T^2 di Hotelling.

Metodo dell'ACP

Nel quadro delle tipologie di controllo della qualità dei dati di tipo multivariato, è stato proposto un metodo (Jackson, 1959) in cui le unità outlier vengono determinate utilizzando le proprietà delle componenti principali combinate all'utilizzo dell'indice T^2 di Hotelling.

Nel contesto del problema della correzione di dati statistici, l'ACP può essere utilizzata, infatti, come base per la costruzione di test grafici che rendano possibile la localizzazione di osservazioni anomale in un contesto multivariato. Questi test consistono nella costruzione di *carte di controllo multivariate* (in cui cioè più variabili sono poste simultaneamente sotto controllo) che rendono possibile, a prefissati livelli di significatività, l'individuazione le cosiddette *osservazioni fuori controllo* da sottoporre a verifica.

In particolare, carte di controllo di tipo parametrico[9] possono essere costruite utilizzando l'indice T^2 di Hotelling calcolato sulla base dei valori delle variabili originali trasformati mediante la tecnica ACP: questa tecnica consente di ridurre la dimensionalità del fenomeno osservato (in termini di numero di variabili da trattare) trasformando le m variabili originali $x_1, x_2,..,x_m$ (correlate fra loro) in m' variabili $z_1, z_2,..,z_m$ (m'<m) incorrelate fra loro.

In particolare, sulla carta di controllo T^2 (largamente applicate nelle fasi di controllo di processi produttivi) vengono riportate le grandezze:

$$Z_i = V' V \qquad [3]$$

dove V è un'opportuna trasformazione lineare delle osservazioni X:

$$V = W (X - \mu_0) \qquad [4]$$

La matrice *W* rappresenta i coefficienti delle variabili ottenuti mediante le componenti principali:

$$W = \Lambda^{-1/2} A \qquad [5]$$

dove Λ è la matrice diagonale costruita a partire dagli autovettori della matrice X dei dati e della sua matrice Σ di varianza e covarianza, mentre *A* è la matrice degli autovettori.

[9] L'ipotesi di base è quella di multinormalità delle variabili.

Il metodo dell'analisi in componenti principali (ACP nel seguito) si articola attraverso le fasi seguenti:

- calcolo delle componenti principali dalla matrice di varianza e covarianza S;
- trasformazione delle variabili di ciascuna unità osservata sulla base degli autovalori ed autovettori calcolati secondo le formule [3] e [4] precedenti;
- per ogni osservazione, calcolo dell'indice T^2 come somma dei valori Z_i e creazione della corrispondente carta di controllo per la visualizzazione delle unità sospette;
- determinazione della soglia di accettazione ad un prefissato livello di confidenza α sulla distribuzione *F* legata alla distribuzione T^2 dalla relazione

$$T^2 = F_{\alpha, m, n-m} \frac{(n-1)m}{(n-m)} \qquad [6]$$

dove m è il numero di variabili ed n è il numero di unità osservate.

- (passo necessario solo nel caso che gli outlier debbano essere corretti e non eliminati) individuazione, per ogni record sospetto, della variabile responsabile (o maggiormente responsabile) di tale anomalia. Un possibile metodo consiste nello scomporre il valore del T^2 relativo all'unità sospetta (la distanza totale dal baricentro della nuvola dei punti in R^n) negli n *contributi* attribuibili alle singole variabili: i contributi maggiori di una certa soglia (ad esempio del contributo medio) sono considerate sospette, e quindi devono essere verificate.

L'applicazione del metodo ACP presenta alcuni punti critici:

- *quali* variabili analizzare, tenuto conto che diversi set di variabili possono condurre a risultati diversi;
- *quante* variabili considerare (un numero troppo basso di variabili può condurre ad una trattazione poco dettagliata delle relazioni fra variabili, un numero troppo elevato può implicare una eccessiva complessità dell'analisi);
- *scelta delle variabili di classificazione* (il metodo risulta tanto più efficiente quanto maggiore è l'omogeneità dei dati nei domini, cioè quanto minore è al loro interno la variabilità delle relazioni fra variabili);
- *scelta dei limiti delle regioni di accettazione* a livello di unità e di variabili sospette.

6.1.2.2. L'editing grafico

Nell'ambito del macroediting sono stati sviluppati una serie di prodotti software di tipo grafico per facilitare ed accelerare le operazioni di controllo e correzione interattiva di dati quantitativi.

Gli algoritmi alla base delle procedure di editing di tipo grafico sono sostanzialmente delle generalizzazioni grafiche dei metodi di tipo macro non grafici: ad esempio, il metodo Box- Plot descritto nel seguito utilizza lo stesso algoritmo di editing del metodo Top-Down.

I vantaggi derivanti dall'uso di metodi di tipo grafico in fase di editing possono essere riassunti nei punti seguenti:

1. possibilità di avere una visione globale della distribuzione dei dati rispetto agli aggregati di interesse e, quindi, visualizzazione immediata delle unità sospette;
2. basso costo di sviluppo nel caso si utilizzino pacchetti software specificamente pensati per applicazioni di tipo grafico;

3. nel caso siano sviluppati per *moduli*, i programmi che implementano metodi grafici possono essere facilmente integrati e riutilizzati per l'editing di indagini con caratteristiche simili.

In generale, tali metodi sono tanto più efficienti quanto minore è la numerosità delle osservazioni sottoposte ad editing, dal momento che la visualizzazione su diagrammi di un numero eccessivamente elevato di punti può rendere difficoltosa l'individuazione delle unità anomale. In presenza di indagini di grandi dimensioni, pertanto, resta valido il suggerimento di effettuare una stratificazione (il più omogenea possibile) delle unità rispondenti e di effettuare applicazioni separate dei metodi agli strati così ottenuti.

Accanto al software sviluppato *ad hoc* per rispondere alle esigenze di indagini particolari, è stato messo a punto un prodotto generalizzato, di nome ARIES, descritto nel relativo paragrafo.

Il metodo Box-Plot

Si tratta di un metodo grafico che, analogamente al metodo *Top-Down*, prevede il calcolo di una funzione di controllo sui dati elementari all'interno dei domini di interesse per ogni variabile da sottoporre a controllo, funzione che tiene conto eventualmente dei pesi associati a tali unità (Granquist, 1992a).

I valori della funzione di controllo vengono rappresentati graficamente in modo tale che la definizione degli intervalli di accettazione e, quindi, la localizzazione delle unità sospette avvenga interattivamente agendo direttamente sul diagramma. La scelta dei record *sospetti*, pertanto, è lasciata all'esperienza del ricercatore che, in base alle proprie conoscenze dirette sulla materia, decide quali unità statistiche sottoporre a controllo interattivo. Si intuisce, altresì, che quante e quali unità statistiche possano essere controllate dipende da criteri empirici e non dall'applicazione di metodi statistici e/o matematici.

Il nome del metodo deriva dal fatto che i record sospetti vengono evidenziati disegnando un quadratino (*box*) intorno all'immagine del record elementare proiettata (*plot*) nel grafico dalla funzione di controllo utilizzata.

L'Australian Bureau of Statistics ha applicato questo metodo all'indagine sui "Guadagni Settimanali" ottenendo, a fronte di una perdita non significativa della precisione delle stime, un risparmio operativo quantificabile intorno al 75%.

Metodo dell'Editing Grafico

Il metodo noto come *Editing Grafico* (Engstrom, Angsved, 1994), fornisce una soluzione di tipo grafico al problema particolare dell'editing di *variabili quantitative in indagini campionarie di tipo panel*, pur essendo generalizzabile anche al caso di indagini campionarie non periodiche o periodiche non panel.

Il metodo è stato implementato in applicazioni realizzate con procedure personalizzate *object oriented* in modo da sfruttare tutte le potenzialità offerte da linguaggi come il Visual Basic o lo stesso SAS in ambiente UNIX.

Nelle sue applicazioni, l'*Editing Grafico* risulta integrato nelle fasi di data-entry e di editing del sistema di produzione del dato statistico. In particolare, esso si può pensare costituito da due "moduli" principali:

1. *localizzazione degli outlier*;
2. *editing grafico delle unità sospette*.

In generale, la localizzazione degli outlier è basata sul calcolo, per ogni variabile, dominio di interesse e unità elementare, di due indicatori[10]:

1. il *contributo* dell'unità alla stima dell'aggregato;
2. la *variazione relativa* fra due periodi di tempo successivi[11].

Un valore è considerato outlier quando i valori degli indicatori 1 e 2 sono esterni a prefissate *soglie*. Le soglie di accettazione vengono definite manualmente sulla base del contenuto di un diagramma *scatter*, in cui viene visualizzata la distribuzione delle unità rilevate rispetto alla variabile ed al dominio di interesse, ed in cui gli elementi anomali della distribuzione appaiono colorati in rosso. Variando i valori di soglia è possibile osservare, per ogni scelta, quante e quali osservazioni sono classificate come *sospette*.

Una volta scelto il valore ottimale, ogni unità sospetta viene visualizzata nella sua interezza per le operazioni di verifica.

In termini di applicabilità, l'*Editing Grafico* è per lo più adatto al trattamento di piccole indagini periodiche (panel se di tipo campionario). Nel caso di indagini di dimensioni più grandi, date le difficoltà di chiara visualizzazione di nuvole di punti troppo fitte, è consigliabile "partizionare" il data-set originale (ad esempio utilizzando opportune variabili di classificazione) ed applicare il metodo ai sottoinsiemi così ottenuti.

6.1.2.2.1 ARIES

Sviluppato dall'U.S. Bureau of the Census per la revisione interattiva di dati, il sistema ARIES (Automated Review of Industry Employment Statistics) adotta un approccio di tipo macro al problema del controllo e della correzione di dati di tipo quantitativo. In particolare, esso utilizza un metodo del tipo *top-down* per la localizzazione dei valori outlier, in quanto la ricerca dei valori sospetti è guidata dalla localizzazione preliminare di valori aggregati sospetti.

Per la visualizzazione degli elementi anomali ai vari livelli di aggregazione ARIES adotta rappresentazioni di tipo grafico e/o tabellare: a partire dai livelli di aggregazione più elevati, per ogni variabile l'operatore è in grado di visualizzare le stime che deviano significativamente dai corrispondenti trend storici, e può quindi restringere la sua ricerca di valori anomali a tale sottoinsieme di stime. Individuati gli outlier nell'ambito degli aggregati sospetti, essi vengono corretti (nel caso in cui corrispondano a valori errati) oppure vengono opportunamente riponderati: in entrambi i casi il sistema procede al ricalcolo automatico delle stime ai vari livelli di aggregazione.

Il processo di editing grafico interattivo implementato in ARIES consiste quindi nelle seguenti fasi principali:

1. per ogni variabile di interesse, l'operatore utilizza una rappresentazione grafica di tutti gli aggregati di sua competenza per l'identificazione di quelli con stime sospette.

[10] Se l'indagine è di tipo campionario, entrambi gli indici vanno ponderati in accordo col disegno di campionamento.

[11] Questo indice e' calcolabile solo per indagini periodiche censuarie o campionarie di tipo panel. Nel caso di indagini campionarie periodiche non panel, il metodo deve essere applicato rivedendo la definizione di outlier (ad esempio utilizzando solo lo stimatore 1).

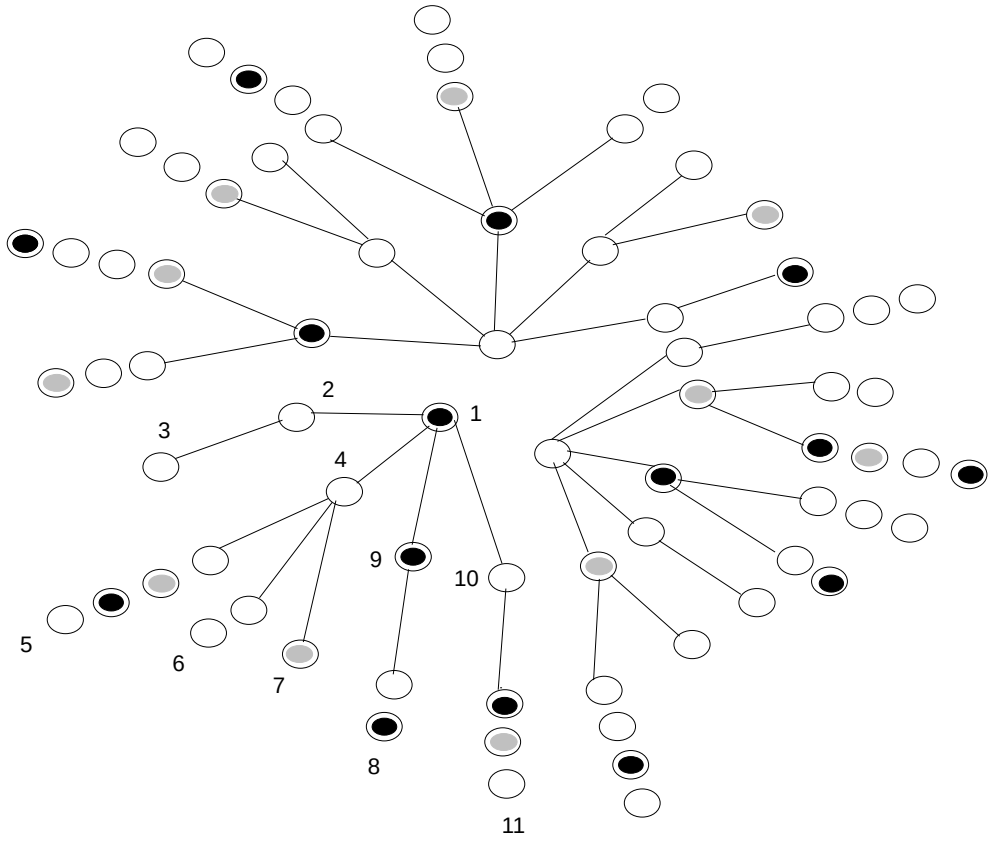
L'identificazione è basata sul calcolo delle variazioni subìte dalle stime stesse in un prefissato intervallo di tempo;

2. per ogni stima (aggregato) sospetta, l'operatore procede all'individuazione delle osservazioni elementari (ad esempio, le imprese) che possono aver contribuito a determinare la stima sospetta. A tale scopo, l'operatore può utilizzare tre metodi diversi: due metodi di tipo grafico oppure un metodo basato sull'uso di interrogazioni di controllo;
3. le unità sospette così identificate vengono corrette interattivamente (nel caso corrispondano a valori errati) oppure l'operatore assegna ad esse un peso appropriato. Le stime ai vari livelli vengono quindi automaticamente ricalcolate sulla base del nuovo sistema di pesi.

La rappresentazione grafica dei dati di cui alla fase 1 consiste nella costruzione delle cosiddette *mappe di anomalia* (vedi Figura 2), costituite da alberi di celle poste su circonferenze concentriche in cui viene rappresentata la struttura dei dati a vari livelli di aggregazione. In tali mappe, le celle più esterne corrispondono alle aggregazioni più fini, mentre le celle disposte su circonferenze via via più interne corrispondono alle successive aggregazioni delle celle stesse, fino alla circonferenza centrale, che rappresenta il livello di massima aggregazione dei dati. Le celle sono collegate fra loro da linee che, a partire dal centro, si dirigono a raggiera verso le celle periferiche. Le celle perimetrali corrispondono agli *strati per i quali vengono prodotte le stime finali dell'indagine*: le stime ai livelli di aggregazione superiori verso il centro dell'albero sono ottenute sommando le stime ai livelli immediatamente inferiori.

Una diversa colorazione viene utilizzata per evidenziare i nodi le cui stime sono esterne ai limiti di accettazione, fissati sulla base delle variazioni fra i dati correnti e i dati storici, variazioni che vengono calcolate per tutti i livelli di aggregazione presenti. Ad ogni colore corrisponde un diverso grado di sospettosità, cioè una diversa distanza fra la stima ed i prefissati limiti di accettazione. Anche le linee di collegamento dell'albero sono colorate qualora uniscano celle con stesso grado di sospettosità. Selezionando col mouse un nodo corrispondente ad una cella sospetta, il sistema visualizza le serie storiche relative a quell'aggregato (a prefissati intervalli di tempo, a seconda della periodicità dell'indagine) ed alla variabile di interesse, rendendo quindi disponibile l'informazione generale su trend e stagionalità del fenomeno in oggetto.

Figura 1 - Esempio di mappa di anomalia



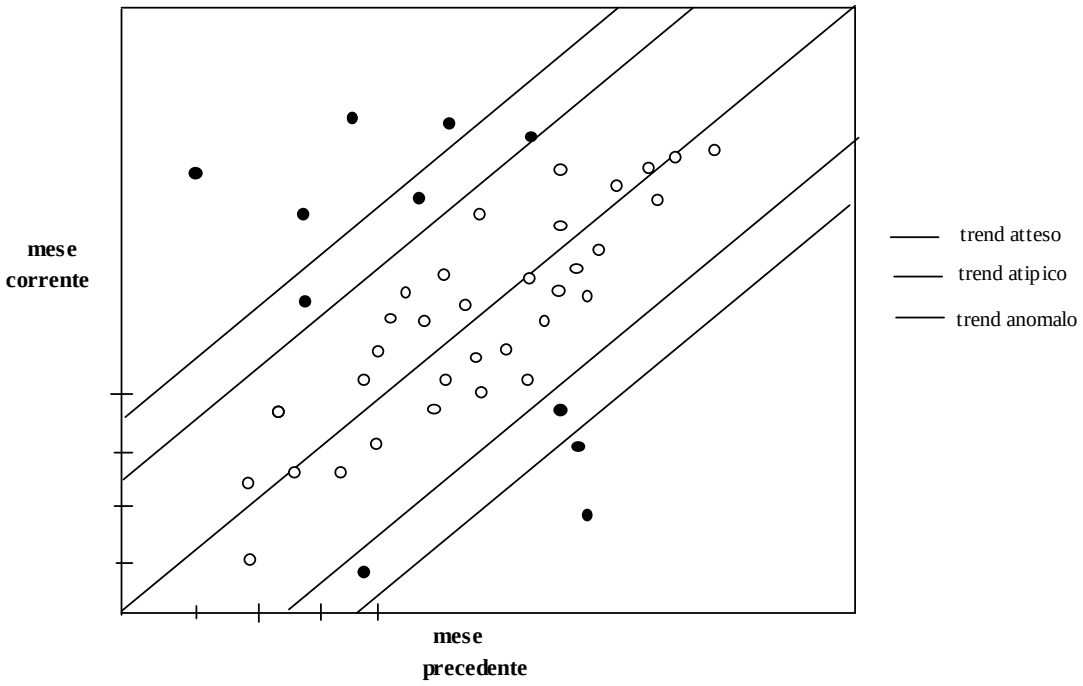
Due sono i vantaggi connessi all'utilizzo delle mappe di anomalia:

1. l'operatore ha un'immagine immediata dell'andamento delle variazioni intervenute nelle stime degli aggregati di sua competenza;
2. l'operatore è in grado di risalire rapidamente in modo selettivo alle unità elementari che con maggior probabilità sono responsabili dell'anomalia a livello di aggregato.

Per l'individuazione delle osservazioni elementari sospette in un aggregato con comportamento anomalo ARIES prevede l'utilizzo alternativo di tre metodi:

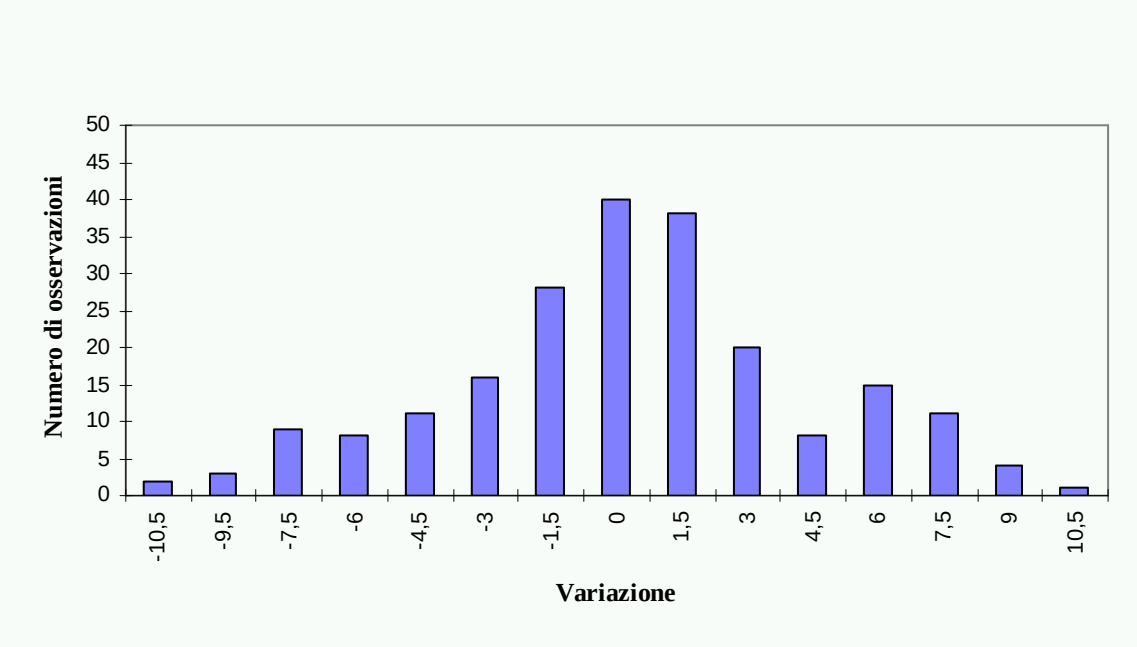
1. costruzione di un *diagramma di tipo scatter* in cui, per ogni unità elementare appartenente all'aggregato sospetto, vengono rappresentati i valori corrente e precedente (vedi Figura 2). Le osservazioni corrispondenti a valori esterni a prefissate regioni di accettazione (una più esterna ed una più interna, a seconda dell'entità delle variazioni ammesse) vengono sottoposte a revisione interattiva. Poiché la situazione ottimale si ha quando tutti i punti giacciono sulla diagonale principale, le regioni di accettazione vengono costruite sulla base di prefissati multipli della deviazione standard da tale diagonale. Selezionando col mouse un punto sospetto del diagramma, il sistema visualizza tutte le informazioni correnti e storiche ad esso relative mediante tabelle e grafici, con la visualizzazione delle serie storiche relative a tutte le variabili di interesse.

Figura 2 - Esempio di diagramma scatter per un singolo aggregato



2. rappresentazione grafica della *distribuzione delle variazioni* dal periodo corrente al precedente nelle osservazioni elementari appartenenti all'aggregato sospetto (vedi Figura 3). Le unità outlier corrispondono ai valori appartenenti alle code della distribuzione: l'ampiezza delle regioni di rifiuto (code) può essere fissata interattivamente dall'operatore muovendo delle barre lungo la distribuzione stessa.

Figura 3 - Esempio di distribuzione delle variazioni per un singolo aggregato



3. utilizzo di prefissate *interrogazioni di controllo*: sono considerate sospette tutte quelle osservazioni che soddisfano le condizioni stabilite attraverso tali interrogazioni. Questo metodo è particolarmente adatto al caso in cui si vogliano localizzare osservazioni con specifiche caratteristiche oppure con particolari relazioni fra un prefissato insieme di variabili.

Fra i vantaggi garantiti dall'utilizzo del sistema grafico interattivo generalizzato ARIES i principali sono la semplicità di utilizzo, la flessibilità nel trattamento dei dati grazie alla varietà di metodi utilizzabili, l'efficienza, in termini di costi e di tempi necessari per la revisione interattiva dei dati elementari, dovuta al tipo di metodo (il top-down) implementato per l'individuazione degli outlier.

In termini di applicabilità, il sistema automatico ARIES è utilizzabile per il trattamento di tutte le indagini periodiche che rilevano variabili quantitative, siano esse totali, campionarie o esaustive. Il sistema risulta in ogni caso tanto più efficace quanto maggiore è la frequenza di ripetizione dell'indagine, e quanto più omogeneo è il comportamento delle unità rilevate all'interno degli strati e degli aggregati di interesse.

6.1.2.3. L'Editing Selettivo

Le metodologie sviluppate nell'ambito dell'*editing selettivo* presentano la caratteristica comune di prevedere l'individuazione, la verifica interattiva e l'eventuale correzione dei record del data-set che:

- hanno un *impatto significativo sulle stime finali*;
- presentano un *consistente errore globale rispetto ad un prefissato insieme di variabili*.

A differenza del macroediting, i metodi di tipo selettivo propriamente detti necessitano dell'informazione preliminare sui record errati o sospetti, vengono applicati cioè sul sottoinsieme di dati grezzi segnalati in errore (certo o probabile) da un dato piano di incompatibilità.

Pertanto, il processo tipico di editing nell'ottica dell'editing selettivo è caratterizzato dalle fasi seguenti (Latouche, Berthelot, 1992):

1. individuazione preliminare delle unità errate mediante un piano di incompatibilità automatico (deterministico o probabilistico);
2. follow-up completo nei casi di mancata risposta totale, di unità che potrebbero non essere eleggibili (cioè non appartenenti alla popolazione di interesse), di record errati appartenenti a domini con numerosità molto bassa (per i quali è necessario effettuare una verifica più accurata);
3. individuazione delle altre unità da sottoporre a controllo interattivo o a follow-up sulla base di funzioni, dette *funzioni punteggio* ("*score functions"*), opportunamente costruite;
4. correzione delle restanti unità sospette o in errore mediante un piano di incompatibilità automatico probabilistico (del tipo di GEIS o SPEER).

Le funzioni di cui al punto 3 assegnano ai record riconosciuti come errati dalla procedura di controllo automatica un *punteggio* globale, calcolato considerando *contemporaneamente* tutte le variabili rilevate (o un prefissato sottoinsieme di esse). Tali funzioni tengono generalmente conto sia della *grandezza* di ogni unità rispondente rispetto ad ogni variabile rilevata (in termini di impatto sulle stime finali), sia dell'importanza degli errori presenti in ogni record che di quella delle variabili.

Altri fattori di cui si dovrebbe tenere conto nella fase di costruzione di funzioni punteggio sono:

- i pesi campionari, informazioni storiche, informazioni ausiliarie, ecc.;
- il tasso di non-risposta (un alto tasso di non-risposta dovrebbe indurre un incremento di ritorni per migliorare la qualità dei dati);
- la facilità di implementazione;
- la flessibilità in termini di adattabilità a varie indagini;
- l'indipendenza dal flusso delle risposte (cioè funzioni basate su parametri predefiniti sulla base ad esempio di precedenti ripetizioni dell'indagine);
- per una data variabile, la funzione punteggio dovrebbe produrre valori con distribuzione simile ai diversi livelli di aggregazione.

I metodi di tipo selettivo possono essere visti come un caso particolare del macroediting multivariato, dal momento che entrambi gli approcci prevedono che l'individuazione delle osservazioni sospette tenga conto simultaneamente delle variabili rilevate. Nel caso dell'editing selettivo, però, l'individuazione dei record da sottoporre a controllo interattivo avviene non sulla base di tecniche di analisi multivariata dei dati, ma sulla base di sintesi più o meno complesse (le funzioni punteggio) di edit costruiti per le variabili di interesse.

L'applicazione di metodi di tipo selettivo è indipendente sia dal numero di variabili quantitative rilevate (in quanto tutte le variabili vengono considerate contemporaneamente nella funzione punteggio), sia dalla dimensione dell'indagine (in quanto solo un numero limitato di unità vengono selezionate per la revisione interattiva).

In generale, l'adozione di tecniche di tipo selettivo si colloca in una strategia complessiva in cui gli obiettivi principali sono:

. la riduzione dei costi (risorse umane e economiche) legati alla fase di controllo interattivo dei dati (follow-up e/o recupero dei modelli cartacei);

. il miglioramento della qualità delle stime finali attraverso il controllo più accurato delle unità con maggiore influenza sulle stime stesse;
. il contenimento dei tempi della fase di editing e, quindi, dell'intero processo di produzione.

Di seguito sono illustrate alcune *funzioni punteggio* per la determinazione dei record da sottoporre a controllo interattivo ed il particolare metodo selettivo proposto da Van de Pol.

Funzioni punteggio

Una definizione più rigorosa e complessa di funzione punteggio è contenuto in Latouche M., Berthelot J.M. (1995), in cui gli autori propongono una strategia complessiva di editing per dati provenienti da indagini economiche basata sull'uso congiunto di tecniche di tipo selettivo e del tradizionale microediting (automatico probabilistico). Questa strategia prevede infatti il ricorso ai tradizionali piani di incompatibilità (meglio se automatici probabilistici) sia per l'individuazione del sottoinsieme di dati in errore su cui effettuare l'individuazione selettiva delle osservazioni anomale, sia per la correzione delle restanti unità in errore.

Nel loro lavoro, Latouche e Berthelot definiscono *funzione punteggio* una funzione che include quattro informazioni principali:

1. la *grandezza* di ogni unità rispondente rispetto ad ogni variabile rilevata (in termini di impatto sulle stime finali);
2. la *grandezza* dell'errore nelle variabili del modello risultate sospette in seguito al controllo preliminare degli errori sui dati grezzi;
3. il numero delle variabili del modello risultate sospette in seguito al controllo preliminare degli errori sui dati grezzi;
4. l'importanza relativa delle variabili (valutata soggettivamente da esperti o metodologi).

Una misura dell'elemento di cui al punto 1 può essere ottenuta dai dati correnti, dai dati storici o da fonti amministrative.

La grandezza dell'errore commesso per ogni variabile e per ogni record può essere approssimato usando il rapporto o la differenza fra i valori correnti e storici.

Per quanto riguarda il numero degli errori in ogni record, è ovvio che la funzione punteggio deve tenerne conto, in quanto priorità dovrà essere data a ritorni che ci consentano di verificare il maggior numero di valori possibile.

L'importanza relativa delle variabili, invece, è un elemento che deve essere valutato soggettivamente da esperti o metodologi: ci sono variabili strategiche da sottoporre a controllo più accurati, variabili che non possono essere imputate efficientemente in modo automatico, ecc.

Gli autori propongono le tre funzioni seguenti:

1. *Funzione RATIO*

Questa funzione è basata principalmente sui rapporti tra i valori correnti e quelli storici (valori finali della ripetizione precedente dell'indagine) delle variabili di interesse. Essa associa ad ogni record k un valore ottenuto attraverso la seguente formula :

$$RATIO_{k,.,t} = \sum_{J=1,k} (f_{k,j,t}\ z_{k,j,t}\ v_{.,j,t})$$

dove:

- f_{kJt} = funzione del peso dell'unità k-esima al tempo t rispetto alla variabile J-esima, e del rapporto tra i valori della variabile J-esima assunti al tempo t ed al tempo t-1 (contributo alla variazione relativa della stima dell'aggregato);

- $z_{k,J,t} = \begin{cases} 0 & \text{se la i-esima variabile e' accettata dagli edit} \\ 1 & \text{altrimenti} \end{cases}$ (*variabile indicatrice di errore*)

- $v_{.,J,t}$ = importanza della variabile J-esima

In termini generali, tale funzione può essere considerata un'estensione della funzione di Hidiroglou-Berthelot, essendo basata su analoghe trasformazioni dei rapporti fra valori correnti e precedenti e sull'uso di parametri per la calibrazione dell'importanza associata alla dimensione delle unità. In termini di applicabilità e di efficienza, per questa funzione valgono le stesse considerazioni fatte per la funzione di Hidiroglou-Berthelot.

2. *Funzione FLAG*

Questa funzione è stata sviluppata con l'obiettivo di dare massima importanza ad una data variabile, soggettivamente scelta da un esperto. Se J è l'indice di tale variabile, la funzione FLAG è definita alla seguente espressione:

$$FLAG_{k,.,t} = w_{k,J,t} \sqrt{MAX(y_{k,J,t}, y_{k,J,t-1})} \sum_{i=1}^{I} (z_{k,i,t}\, v_{.,i,t})$$

in cui:
- $w_{k,J,t}$ è il peso campionario;
- $y_{k,J,t}$, $y_{k,J,t-1}$ sono i valori forniti dal rispondente k al quesito i rispettivamente ai tempi corrente e precedente;
- $z_{k,J,t}$ e $v_{.,J,t}$ hanno lo stesso significato che avevano nella funzione RATIO.

3. *Funzione DIFF*

Questa funzione enfatizza, per ogni variabile, le differenze assolute tra i valori correnti e quelli finali del ciclo precedente. Ogni differenza viene ponderata utilizzando il totale $Y_{.,J,t-1}$ che ogni variabile aveva nel periodo precedente ad un prefissato livello.

Il punteggio globale di ogni record è dato dalla somma seguente:

$$DIFF_{k,.,t} = \sum_{i=1}^{I} \frac{w_{k,i,t} \left| y_{k,i,t} - y_{k,i,t-1} \right| z_{k,i,t}\, v.,i,t}{Y_{.,i,t-1}}$$

Le esperienze mostrano che le soglie critiche che comportano basse percentuali di follow-up (20-30%) garantiscono il contenimento entro limiti accettabili della distorsione delle stime.

L'applicabilità delle funzioni punteggio fin qui illustrate è indipendente dalla dimensione dell'indagine sia in termini di numero di unità statistiche rilevate (in quanto esse vengono applicate ad un sottoinsieme delle sole osservazioni errate), sia in termini di numero di variabili quantitative di interesse (in quanto esse tengono conto contemporaneamente di tali variabili).

E' chiaro che l'efficienza di una strategia di editing basata sull'uso di tali funzioni in fase di controllo interattivo dei dati è tanto maggiore quanto maggiori sono la disponibilità di informazioni (ausiliarie, storiche, ecc.) e la loro affidabilità. E' anche evidente che, dando un'importanza strategica alle informazioni storiche, tali funzioni sono tanto più efficienti quanto maggiore è la frequenza di ripetizione dell'indagine: in questo caso, infatti, variazioni consistenti nell'entità delle variabili di interesse saranno più probabilmente dovute a errore che non al naturale evolversi dei fenomeni investigati.

Il Metodo di Van de Pol

Partendo dall'idea della procedura di Hidiroglou-Berthelot, il metodo selettivo proposto da Van de Pol (Van de Pol F., 1994) è basato sul presupposto che, per un controllo più efficiente dei dati, è necessario assegnare diverse priorità agli edit definiti fra le variabili di interesse. Tale metodo è basato sull'ulteriore considerazione che, ai fini dell'individuazione dei record errati, sia più efficiente accentrare l'attenzione non sulle singole variabili, ma sui record nel loro complesso: in altri termini, si procede alla verifica e correzione dei modelli nel loro complesso piuttosto che, volta per volta, delle singole variabili. Il metodo consente quindi il controllo della correttezza delle unità statistiche non solo con riferimento alla loro "importanza" (in termini di impatto sulle stime finali), ma anche alla correttezza globale, al loro interno, delle relazioni prefissate tra le variabili di interesse.

In questa ottica, il metodo prevede che il controllo della correttezza dei dati sia effettuato sulla base dei valori assunti dalla cosiddetta *variabile di punteggio sintetica*: tale variabile rappresenta un indicatore globale della correttezza di un dato record rispetto ad un insieme di edit.

Gli edit, detti *variabili messaggio di errore*, sono costruiti come *rapporti caratteristici* fra le variabili obiettivo dell'indagine. Sia

$$r_{vf} = \frac{y_{vf}}{x_{vf}}$$

il v-esimo edit fra la variabile (dipendente) y da sottoporre a controllo e la variabile ausiliaria (indipendente) x nell'unità f. Nella situazione teorica campionamento casuale, i rapporti così costruiti dovrebbero avere priorità dipendente dall'entità dell'errore verificatosi in y_{vf} , cioè della quantità $e_{vf} = y_{vf} - Y_{vf}$ dove Y_{vf} è il valore vero (o esatto) della variabile Y nell'unità f.

Per ogni record, la *variabile di punteggio sintetica* è ottenuta mediante una media (ponderata) delle *variabili di punteggio* associate ad ogni rapporto caratteristico: per ogni

record e per ogni rapporto, la *variabile di punteggio* è inversamente proporzionale alla probabilità di inclusione dell'unità campionaria, e direttamente proporzionale all'*importanza* dell'errore in quel rapporto, cioè

$$s^{*}_{vf} \approx \frac{e_{vf}}{\pi_f} \qquad [7]$$

dove π_f è la probabilità di inclusione nel campione dell'unità f[12].

L'errore e_{vf} viene stimato utilizzando, analogamente a quanto avviene nel metodo Hidiroglou-Berthelot, la *devianza* d_r fra il valore osservato di ciascun rapporto e la mediana dei rapporti calcolata sull'intero campione:

$$d_{vf} = \text{abs}\left[\, r_{vf} - \text{median}_f(r_{vf}) \,\right]$$

In corrispondenza di un valore grande di d_{vf} è allora probabile che si sia verificato un errore, con entità dello stesso ordine di grandezza della devianza.

Nell'ipotesi in cui il denominatore x_{vf} è corretto[13], l'errore e_{vf} può essere stimato mediante la quantità $e_{vf} = d_{vf} \times x_{vf}$, per cui la *variabile punteggio* [7] per l'edit r_{vf} diventa

$$s_{vf} = \frac{d_{vf}\, x_{vf}}{\pi_f} \qquad [8]$$

Per tenere conto dell'eventualità in cui si ritenga che alcuni rapporti siano più importanti, la *variabile punteggio sintetica* viene definita, in termini generali, mediante la seguente media ponderata:

$$s_{\cdot f} = \frac{1}{\sum_{v=1}^{V} w_v} \sum_{v=1}^{V} w_f\, s^{*}_{vf} \qquad [9]$$

dove V è il numero di rapporti considerati, e le singole variabili punteggio sono state preventivamente standardizzate. Nel caso in cui tutti gli edit abbiano stesso peso, $w_v = 1$ la [9] si semplifica nella

$$s_{\cdot f} = \frac{1}{n} \sum_{v=1}^{V} s^{*}_{vf} \qquad [10]$$

Una volta costruita la *variabile di punteggio sintetica*, il problema diventa determinare un *valore soglia* per essa, al di sotto del quale non è necessario effettuare alcuna operazione di editing. La determinazione di tale soglia avviene sulla base dell'analisi dell'*effetto cumulativo* sulle stime finali degli ammontari di editing corrispondenti a vari livelli di soglia.

Supponendo che obiettivo dell'indagine sia la stima dei totali delle variabili rilevate, cioè delle quantità

---

12 Naturalmente, per piccole unità π_f è piccola.

13 L'ipotesi di correttezza delle variabili utilizzate a denominatore deve essere testata ed assicurata introducendo rapporti caratteristici fra tali variabili ed altre variabili ausiliarie.

$$\hat{Y} = \sum_{f=1}^{N} Y_{vf} \qquad [11]$$

dove N è la numerosità della popolazione, e che come stimatore di tale paramtero si utilizza la quantità

$$\hat{Y} = \sum_{f=1}^{n} \frac{y_{vf}}{\pi_f} \qquad [12]$$

dove n è la dimensione del campione, l'analisi dell'effetto cumulativo può essere effettuata mediante l'uso di un grafico nel quale vengono rappresentate le stime

$$\hat{Y} = \sum_{f=1}^{n} \left[\delta_f Y_{vf} + (1-\delta_f) y_{vf} \right] \frac{\left(\sum_{f'} \pi_{f'}\right)}{\pi_{f'}} \qquad [13]$$

per valori decrescenti della variabile punteggio $s_{.f}$, dove con f' sono indicate le unità di piccole dimensioni. Per un dato valore soglia $s_{.}\mid$, tutte le osservazioni f con $s_{.f} > s_{.}\mid$ ($\delta_f = 1$) hanno il valore y_{vf} , che è sospetto, sostituito con il valore editato Y_{vf} (valore vero), mentre per le osservazioni con $s_{.f} < s_{.}\mid$ ($\delta_f = 0$) viene utilizzato il valore non editato y_{vf} . E' chiaro che per valori molto grandi di $s_{.}\mid$ si ha assenza di editing, mentre a valori molto piccoli corrisponde la situazione di editing completo dei dati.

Come valore soglia ottimale viene scelto quello che fornisce stime il più vicine possibile a quelle prodotte dall'editing completo, a fronte di un numero di verifiche minore possibile.

Per la costruzione dei grafici di cui sopra, quindi, è necessario conoscere i valori delle stime ottenute nel caso di editing completo. Di conseguenza, il *valore soglia ottimo* per il metodo di Van de Pol può essere determinato:

- se si conoscono i valori delle stime relativi all'editing completo (Sc), utilizzando il metodo grafico; in tal caso, il valore di soglia è quello che fornisce stime il più vicine possibile alle corrispondenti stime Sc e minimizza il numero di record da sottoporre a verifica;
- se non si può utilizzare il metodo grafico, il valore di soglia può essere determinato come previsto nel metodo Hidiroglou-Berthelot.

6.2. Procedure automatiche di controllo e correzione

Le procedure automatiche possono essere utilizzate nei seguenti passi:

1. applicazione delle regole di *dominio*, di *compilazione* e di *compatibilità* ai dati (**controllo** dei record);
2. **localizzazione degli errori** (individuazione delle variabili errate sulla base delle regole violate);
3. correzione degli errori (**imputazione** delle variabili errate).

Le procedure automatiche possono essere di tipo *deterministico, probabilistico* o misto; possono essere *sviluppate ad hoc*, oppure possono essere il risultato dell'applicazione di *software generalizzato*; infine, possono riguardare *variabili qualitative* oppure *quantitative*.

6.2.1 Approccio deterministico vs. approccio probabilistico

6.2.1.1. Fase di controllo dei record

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 sicuramente la presenza di errori.

Ad esempio:

SE (sesso = *maschio* E professione = *casalinga*) ALLORA sussiste incompatibilità x

Una regola di questo tipo non individua, di per sè, 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.

6.2.1.2. Fase di localizzazione degli errori

E' in questa fase che diviene decisivo il tipo di approccio adottato. Nell'***approccio deterministico***, ad ogni situazione di incompatibilità segue, contestualmente, l'indicazione delle variabili che debbono considerarsi errate, e, in quanto tali, da imputare. Nell'esempio considerato avremo, per ipotesi:

SE (sesso = *maschio* E professione = *casalinga*) ALLORA sesso ← *femmina*

il che significa che, se in un record è attivata la condizione di incompatibilità "maschio/casalinga", la regola indica l'azione da effettuare per correggere l'errore, che consiste nell'imputare la modalità *femmina* alla variabile sesso.

Generalizzando, una volta attivate, mediante le regole di compatibilità, una o più condizioni di errore in un dato record, sono determinate a priori le azioni da intraprendere per riportare il medesimo record in una situazione di correttezza.

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

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

La condizione di errore 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 situazione di errore, l'elenco delle azioni da intraprendere per eliminare gli errori dai dati: l'esperto statistico deve limitarsi a definire le situazioni di errore, demandando ad un prefissato algoritmo il compito di riportare il record ad una situazione di correttezza.

L'approccio probabilistico ha il suo punto di riferimento nella cosiddetta *metodologia Fellegi-Holt*, esposta nell'articolo "A systematic approach to automatic edit and imputation" di I.Fellegi e D.Holt, pubblicato nel 1976 sul Journal of the American Statistical Association, di cui riportiamo una sintesi in appendice.

Un piano probabilistico è composto, quindi, 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: l'edit è attivato da un dato record quando sono verificate simultaneamente tutte le condizioni in esso definite. La parte SE di una R.I.D. (cioè quella che esprime la situazione di errore) può corrispondere a uno o più edit in forma normale.

L'algoritmo che elimina gli errori provvede a determinare, per ogni record e per ogni situazione di errore, quali variabili modificare in modo da avere la certezza di eliminare gli errori individuati e, soprattutto, di non introdurne altri nel record, minimizzando nel contempo il numero di variabili modificate.

Gli edit in forma normale definiti dall'esperto, gli edit espliciti, sono sufficienti ad individuare la presenza di errori all'interno dei record di un file, ma non a garantire una imputazione di valori corretta ed ottimale. Infatti, la scelta di quali variabili modificare e di quali nuovi valori assegnare, è condizionata dai vincoli di correttezza (non introdurre nuovi errori nel record) e di minimalità (modificare il minor numero possibile di variabili). A tal fine, occorre considerare anche i cosiddetti edit impliciti, derivabili da quelli espliciti ed individuare così l'insieme minimo e completo degli edit.

La metodologia di Fellegi-Holt prevede che, una volta definiti gli edit espliciti, questi siano analizzati sia per scoprire la presenza di contraddizioni e/o ridondanze che per derivare tutti gli edit impliciti in essi contenuti.

La fase dell'analisi e della derivazione degli edit, produce un insieme di regole che ha le seguenti caratteristiche:

1. è minimale, privo cioè di edit ridondanti;
2. è corretto, privo di edit tra loro contraddittori;
3. è completo, in quanto contiene esplicitamente tutti gli edit implicitamente definiti all'interno di quelli iniziali.

La derivazione degli edit impliciti nell'ambito della metodologia Fellegi-Holt, rappresenta un'operazione altamente critica: infatti la generazione degli edit impliciti richiede un numero di operazioni che è esponenziale rispetto al numero di edit espliciti. Spesso, la derivazione degli edit impliciti risulta impossibile; in questo caso si ricorre ad euristiche che permettono di limitare a priori il numero delle operazioni necessarie e alla partizione dell'insieme iniziale di edit suddividendo la fase di correzione in tante sottofasi quanti sono i sottoinsiemi di edit cosi definiti.

Nel caso delle variabili quantitative, non si tratta solo di individuare le variabili errate, cioè quelle che determinano l'attivazione di incompatibilità, ma anche di determinare, per ogni variabile, dei limiti al di là dei quali i valori riscontrati possono essere considerati come

outlier, cioè valori che contraddicono la tendenza generale, a livello trasversale (relativamente all'insieme delle unità rispondente in una stessa ripetizione dell'indagine) oppure a livello longitudinale (relativamente alle risposte fornite dalle unità in ripetizioni differenti della stessa indagine). I vari metodi utilizzati (cfr. *procedura di Hidiroglou-Berthelot*) possono essere considerati metodi deterministici: sulla base dell'andamento effettivo di una data variabile, viene determinato l'intervallo di accettazione: se in un record il valore della variabile cade al di fuori di tale intervallo, la variabile è considerata errata, e candidata all'imputazione.

6.2.1.3. Fase di correzione degli errori (imputazione delle variabili errate)

Una volta individuate le variabili contenenti gli errori che hanno causato l'attivazione delle incompatibilità, oppure i cui valori sono stati giudicati outlier, occorre procedere alla fase di imputazione di tali variabili, onde rimuovere gli errori, cercando di ripristinare i valori veri.

Anche in questo caso, i metodi per l'imputazione possono far riferimento ad uno dei due opposti approcci: in un caso parleremo di metodi deterministici, nell'altro di metodi stocastici (Kovar, Whitridge, 1995).

Un metodo di imputazione è deterministico quando il nuovo valore di una variabile è stabilito con certezza sulla base di un'indicazione diretta di tale valore, oppure di vincoli logici, o mediante calcolo. Tra i vari **metodi deterministici** citiamo:

- *imputazione da valore prefissato*: nell'esempio di R.I.D. citato in precedenza, nella parte ALLORA della regola non solo si definiva la variabile "sesso" come la variabile errata da correggere, ma veniva anche indicato il valore da assegnare a tale variabile, cioè "femmina";
- *imputazione "logica"* (da vincoli logici): l'imputazione, in tal caso, è determinata da vincoli tali da restringere ad un valore unico quello da assegnare alla variabile errata. L'esempio precedente è anche un caso di imputazione logica: il valore "femmina" è l'unico in grado di disattivare la condizione della parte SE. Un ulteriore esempio, per le variabili quantitative: se una regola di compatibilità stabilisce che le spese devono essere inferiori al reddito, e se quest'ultimo è pari a zero, allora l'unico valore imputabile alla variabile "spese" è proprio zero;
- *imputazione da serie storica*: per variabili che tendono ad essere stabili nel tempo, in caso di imputazione viene riproposto il valore disponibile nel periodo immediatamente precedente. Come variante, tale valore viene "aggiustato" per tener conto del trend della serie storica relativa alla variabile;
- *imputazione del valor medio*: alla variabile viene imputato il valor medio calcolato sui dati disponibili, o in un opportuno strato di questi (è un metodo che può essere utilizzato solo per le variabili quantitative). Lo svantaggio è che in tal modo viene introdotta una seria distorsione nella distribuzione della variabile, creando un picco artificiale in corrispondenza del suo valor medio;
- *imputazione sequenziale da donatore "hot deck"*: in una data variabile il valore errato viene sostituito dal valore corrispondente della ultima unità rispondente. Con questo metodo è estremamente importante l'ordinamento cui è sottoposto il file oggetto della correzione: le variabili di ordinamento sono quelle rispetto alle quali è assicurata la minima distanza tra il record *ricevente* e quello *donatore*. Un possibile aspetto negativo di tale metodo è nel fatto

che uno stesso donatore può essere utilizzato più volte, tante quanto la dimensione di un insieme di record adiacenti che necessitano di correzione: ciò può creare picchi artificiali nei valori della variabile;

- *imputazione dal più vicino donatore*: la differenza col metodo precedente consiste nel fatto che il donatore è scelto in modo tale da una qualche misura della distanza tra esso ed il ricevente è minimizzata. In genere, la distanza scelta non è di tipo spaziale, ma una misura multivariata basata sui dati disponibili: per tale ragione, il metodo è più appropriato per le variabili quantitative. Tra i vantaggi, citiamo quello relativo al mantenimento ottimale delle distribuzioni multivariate originali. Lo svantaggio è comune al metodo precedente: uno stesso donatore può essere utilizzato più volte; esiste però la possibilità di limitare questo svantaggio, ponendo un tetto al numero di volte che uno stesso record può essere utilizzato come donatore, oppure introducendo nella funzione di distanza una funzione di penalizzazione che tiene conto del numero di volte che un dato record è già stato utilizzato come donatore;
- *imputazione da regressione*: per l'imputazione di una data variabile viene utilizzato il valore fornito da una funzione di regressione che fa uso di una o più variabili ausiliarie. La variabile da imputare deve essere quantitativa, mentre le variabili indipendenti possono essere continue o discrete. Il metodo assicura buoni risultati sotto due condizioni: (i) alta correlazione tra variabile da imputare e variabili ausiliarie e (ii) disponibilità di valori corretti delle variabili ausiliarie per tutti i (o per gran parte dei) record. Un caso particolare è dato dall'*imputazione da rapporto (ratio)* in cui viene considerata la relazione tra la variabile da imputare ed una variabile ausiliaria con essa altamente correlata: in tal caso entrambe devono essere di tipo continuo. Questo metodo si rivela adeguato nei casi in cui la variabile da imputare è affetta da un errore di tipo stocastico, oppure sistematico ma il cui andamento è legato alla variabile ausiliaria.

I limiti fondamentali di cui soffrono i metodi deterministici risiedono nel fatto che essi spesso riducono la variabilità della variabile imputata, e talvolta introducono distorsioni. Per queste ragioni sono state introdotte delle tecniche di imputazione stocastica, molte delle quali rappresentano varianti dei metodi deterministici, ideate per mantenere le distribuzioni e la variabilità dei dati.

6.2.1.4. Un modello generale di imputazione

Molti dei metodi di imputazione possono essere visti come casi particolari della stima di un modello di regressione:

$$\hat{y}_{mk} = \hat{\beta}_{r0} + \sum_j \hat{\beta}_{rj} x_{mjk} + \hat{e}_{mk}$$

in cui $\hat{y}_{mk}$ rappresenta il valore imputato per la *k*-esima unità con un valore mancante, x_{mjk} è il valore delle variabili ausiliarie $\hat{\beta}_{r0}$ e $\hat{\beta}_{rj}$ sono i coefficienti della regressione di *y* su *x* per i rispondenti, mentre $\hat{e}_{mk}$ costituisce un residuo corrispondente ad uno schema probabilistico associato al particolare metodo di imputazione prescelto. Alcuni casi particolari:

i. $\hat{e}_{mk} = 0$; in questo caso $\hat{y}_{mk}$ costituisce il valore stimato con modello di regressione;
ii. se $\hat{e}_{mk} = 0$ e x_j è una variabile *dummy* che denota la classe allora l'equazione equivale all'imputazione con media della classe, ossia $\hat{y}_{mk} = \bar{y}_{rh}$, di cui l'imputazione mediante

media globale può essere vista come un caso particolare in cui non si utilizzano informazioni ausiliarie;

iii. se alla media della classe in ii. si aggiunge una componente casuale individuale e_{rhk} si è ricondotti all'imputazione stocastica all'interno di classi, che equivale all'adattamento ai dati di un modello ANOVA con effetti casuali, in cui il residuo è costituito dallo scarto per ciascun rispondente dalla media della classe, ossia $\hat{y}_{mk} = \bar{y}_{rh} + e_{rhk} = y_{rhk}$. Le imputazioni con metodo *hot-deck* (sequenziale o gerarchico, all'interno di classi) possono essere rappresentate come casi particolari di questo tipo di imputazione.

La distinzione essenziale tra metodi deterministici e metodi stocastici di imputazione dipende dall'avere posto $\hat{e}_{mk} = 0$ oppure no. La scelta tra un metodo di imputazione deterministica ed uno stocastico può essere fatta sulla base degli obiettivi che l'analisi dei dati dell'indagine si prefigge. Così per la stima della media della popolazione sulla base di valori osservati e valori imputati è preferibile utilizzare un metodo di imputazione deterministica in quanto, pur potendo effettuare una scelta controllata della componente casuale in una imputazione stocastica, ciononostante ne consegue una certa perdita di precisione delle stime.

Per contro, ai fini della stima della variabilità e della distribuzione della variabile di studio un'imputazione deterministica può condurre a risultati di modesta qualità. Un semplice esempio è rappresentato dall'imputazione mediante valore medio. La sostituzione in tutte le MR del valore medio dei rispondenti (eventualmente all'interno di classi di imputazione) crea picchi artificiali nella distribuzione delle risposte in corrispondenza del valore medio delle classi, riducendo la variabilità della variabile di studio, soprattutto per la parte di variabilità *all'interno* delle classi. In tali casi, l'uso di un metodo di imputazione stocastica, di tipo *hot-deck* ad es., consegue migliori risultati. Esistono poi particolari proposte metodologiche (imputazione multipla, ad es.) che cercano, oltre che ricostruire le MR garantendo la variabilità dei valori dei rispondenti, di ottenere una stima di una componente aggiuntiva della variabilità totale, legata al processo stesso di ricostruzione.

Qualora si opti per una imputazione stocastica si pone il problema della scelta di una opportuna distribuzione da cui estrarre la componente stocastica. Una scelta naturale con una imputazione mediante modello (di regressione) è quella di una distribuzione dei residui normale, con media zero e varianza uguale alla varianza residua della regressione sui rispondenti. Possibili alternative sono rappresentate dalla scelta casuale dalla distribuzione empirica dei residui dei rispondenti o la scelta di un residuo a partire da unità rispondenti considerate "vicine" all'unità con valore mancante sulla base di variabili ausiliarie. Ciò è, ad esempio, quello che si verifica con una imputazione con donatore (*hot-deck* o *nearest-neighbour*), in cui all'unità con MR è assegnato un valore da un sottoinsieme di unità rispondenti considerate "vicine".

6.2.1.5. Vantaggi e svantaggi degli approcci deterministico e probabilistico

Quali sono i vantaggi e gli svantaggi dei due diversi approcci? Molto schematicamente, possiamo ascrivere ai vantaggi del metodo deterministico:

- la completa applicabilità: una procedura deterministica è sempre applicabile ai dati una volta tradotte le regole di imputazione deterministica in istruzioni di un programma;
- l'efficienza elaborativa: il tempo necessario per eseguire il programma che traduce la procedura deterministica è lineare rispetto al numero di regole e al numero di record;

- l'orientabilità degli effetti: lo statistico può orientare i risultati dell'applicazione della procedura deterministica definendo opportunamente la parte imputazione di ogni regole, e la sequenza di queste nel piano.

Quest'ultimo elemento è di una certa importanza: ad esempio, sulla base della fiducia che lo statistico nutre rispetto alla correttezza delle variabili, egli può implicitamente stabilire una gerarchia tra queste, orientando la modifica verso quelle che egli ritiene meno affidabili. In realtà, questo è un risultato che si può ottenere, utilizzando opportuni pesi, anche nel caso delle procedure non deterministiche.

Tra gli svantaggi ed i limiti del deterministico citiamo:

- la mancata garanzia di correttezza dei record alla fine della fase di correzione e la conseguente necessità di cicli di controllo e correzione;
- la mancata garanzia di minimizzazione dei cambiamenti della distribuzione originale, ovvero non è assicurato il risultato che in ogni record errato il numero di variabili modificate per riportarlo ad una situazione di correttezza sia il minimo possibile;
- l'introduzione di distorsioni nelle distribuzioni e la perdita di variabilità.

I vantaggi dell'approccio probabilistico sono speculari ai limiti di quello deterministico:

- la correttezza finale dei record sottoposti a correzione,
- la minimalità del cambiamento (assicurata dallo stesso algoritmo che provvede a trovare il numero minimo di variabili, l'insieme minimale, da modificare),
- il maggior rispetto della distribuzione congiunta delle variabili, presente nell'insieme dei dati rilevati.

In caso di errori sistematici, l'approccio deterministico si rivela, nella maggior parte dei casi, il più adatto, soprattutto nel passo di localizzazione degli errori. L'applicazione del probabilistico, al contrario, rischia di introdurre nuove distorsioni nei dati, qualora non si pesino opportunamente le variabili per tener conto della sistematicità di tali errori.

6.2.1.6. Effetto dell'imputazione sulle relazioni bivariate

Anche se normalmente nel valutare vantaggi e svantaggi di una particolare tecnica di imputazione si concentra l'attenzione sul suo impatto sulla stima di statistiche e distribuzioni univariate il processo di imputazione può avere forti effetti sui legami fra due o più variabili, spesso con il risultato di attenuare le relazioni di associazione. Allo scopo di esaminare tali effetti si consideri il caso di una variabile di studio *y* con dati incompleti caratterizzati da un meccanismo di mancata risposta MAR mentre un'altra variabile *x* sia osservata con risposta completa. Si voglia ottenere una stima $\hat{\sigma}_{yx}$ della covarianza σ_{yx} tra *y* ed *x* sulla base dei valori veri dei rispondenti e dei valori imputati per i non rispondenti. Si può dimostrare (Kalton e Kasprzyk, 1986) che, dal momento che il valore atteso di $\hat{y}_{mk}$ ottenuto con imputazione stocastica — sotto l'assunzione che $E(\hat{e}_{mk}) = 0$ — è uguale al valore $\hat{y}_{mk}$ risultante dalla corrispondente imputazione deterministica, il valore atteso di $\hat{\sigma}_{yx}$ risultante da un metodo di imputazione deterministico è lo stesso di quello ottenuto con il corrispondente metodo stocastico. Dunque, indipendentemente dall'aver scelto un'imputazione deterministica o stocastica si dimostra (Santos, 1981) che la distorsione relativa di $\hat{\sigma}_{yx}$ dipende dal metodo di imputazione utilizzato. In particolare:

- imputazione media: in questo caso non vi è nessuna relazione tra il valore imputato della *y* e il valore della *x* e la distorsione relativa di $\hat{\sigma}_{xy} \cong -\overline{M}$, con $\overline{M}$ tasso di non risposta;

dunque quanto più elevato è il numero di MR tanto più tendente a zero sarà la stima della covarianza;

- imputazione (media o casuale) all'interno di classi: in tal caso nel processo di imputazione si ricorre all'uso di variabili ausiliarie, con il risultato di attenuare la distorsione relativa di $\hat{\sigma}_{xy} \cong -\overline{M}(\sigma_{xy.z} / \sigma_{yx})$, in cui $\sigma_{xy.z} = \sum W_h \sigma_{xyh}$ valor medio della covarianza *all'interno* delle classi ottenute sulla base della variabile ausiliaria z, σ_{xyh} è la covarianza all'interno della classe *h*-esima e W_h è la proporzione della popolazione all'interno della classe *h*. Un caso particolare si ha quando *x* = *z*, cioè quando la variabile *x* è utilizzata come variabile ausiliaria: in tal caso si avrà che $\sigma_{xy.z} = 0$, con la conseguenza che $\hat{\sigma}_{xy}$ diviene una stima non distorta della covarianza della popolazione;
- imputazione (deterministica o stocastica) con modello di regressione semplice di *y* su *x*: la distorsione relativa di $\hat{\sigma}_{xy} \cong -\overline{M}\left[1 - (\rho_{xz}\rho_{yz} / \rho_{yx})\right]$ con ρ_{uv} correlazione tra *u* e *v*. Anche in questo caso, se *x* = *z*, $\hat{\sigma}_{xy}$ diventa una stima non distorta della covarianza.

Se dunque lo studio della relazione tra *x* ed *y* rappresenta una componente importante dell'analisi dei dati di un'indagine vale la pena di utilizzare *x* come variabile ausiliaria nel processo di imputazione dei valori mancanti di *y*. Se poi *x* e *y* contengono entrambe MR, la covarianza può risultare attenuata per effetto della imputazione sulle due variabili. Un caso particolare si ha quando *x* ed *y* contengono MR in corrispondenza della *stessa* unità: in tal caso, imputando congiuntamente, vale a dire utilizzando la stessa unità rispondente per l'imputazione sia di *x* che di *y*, la struttura della covarianza viene preservata.

6.2.2 Il software generalizzato

Le procedure automatiche di controllo e correzione possono essere implementate ad hoc, oppure mediante il ricorso a sistemi di tipo generalizzato. Tali sistemi offrono vantaggi notevoli:

- in essi sono generalmente incorporate metodologie sofisticate per il trattamento degli errori, che difficilmente possono essere introdotte in una singola applicazione;
- le funzionalità dei sistemi sono tali da favorire una corretta ed ottimale applicazione di tali metodologie;
- non occorre essere esperti di programmazione di computer: tali sistemi possono essere utilizzati direttamente dai responsabili dell'indagine e/o da metodologi;
- lo sforzo di implementazione delle procedure è minimo se paragonato a quello richiesto dallo sviluppo di procedure ad hoc.

Il software generalizzato per il controllo e la correzione automatica, prodotto e reso disponibile da alcuni Istituti Nazionali di Statistica, è generalmente improntato ai criteri della metodologia Fellegi-Holt, permette cioè lo sviluppo e l'applicazione di procedure prevalentemente probabilistiche, soprattutto per quanto riguarda le modalità di localizzazione degli errori. Nel seguito presenteremo le caratteristiche fondamentali di tre distinti sistemi, SCIA per il trattamento delle variabili qualitative e GEIS e SPEER per quello delle variabili qualitative.

6.2.2.1. SCIA (Sistema Controllo e Imputazione Automatici)

SCIA è un sistema per l'editing e l'imputazione automatica di variabili qualitative realizzato autonomamente dall'Istituto Nazionale di Statistica a partire dal 1990 nell'ambito di un gruppo di progetto[14] diretto alla costruzione di un *ambiente* di strumenti che coprano le diverse situazioni che si presentano in fase di editing e imputazione dei dati.

Attualmente esistono due versioni di SCIA, una in ambiente mainframe IBM compatibile, l'altra in ambiente UNIX; è prevista una terza versione in ambiente personal computer.

La versione attuale di SCIA opera esclusivamente su variabili di tipo qualitativo. Esso consente l'individuazione e la correzione automatica degli errori di tipo sia casuale sia, parzialmente, sistematico presenti nei dati.

Più precisamente, mentre per il trattamento degli errori stocastici SCIA adotta un approccio probabilistico interamente basato sulla metodologia di Fellegi-Holt, per quanto riguarda gli errori sistematici l'attuale versione del sistema prevede la possibilità di verificarne la presenza attraverso opportune analisi delle imputazioni probabilistiche effettuate sui dati, consentendo così di definire le appropriate regole deterministiche per la loro eliminazione. Una volta definite tali regole, SCIA provvede a generare automaticamente programmi COBOL eseguibili indipendentemente dal sistema.

Nel caso degli errori di tipo stocastico, la strategia di editing e correzione dei dati è interamente basata sulla metodologia di Fellegi-Holt.

Per eseguire i processi di controllo e di imputazione dei valori delle variabili devono essere specificate le elaborazioni da effettuare sui dati e le modalità con cui esse devono essere realizzate. La definizione di tali processi prevede 4 passi:

1. *specificazione dei parametri*, mediante i quali vengono definite modalità generali per la fase di imputazione. SCIA prevede che siano specificati:
   - il grado di fissità delle variabili;
   - le variabili da sottoporre a imputazione forzata, quando viene usato il metodo di imputazione basato sulle distribuzioni marginali delle variabili stesse ;
   - i pesi da assegnare alle modalità delle variabili in caso di imputazione mediante distribuzioni marginali;
   - le variabili chiave sulle quali viene ordinato il file di input quando viene usato il metodo di imputazione da donatore: quando questo parametro è specificato, un record errato viene corretto con un donatore avente le stesse chiavi del record errato;
   - criteri per la gestione dell'insieme di record da cui vengono scelti i donatori, e cioè: la dimensione per il *serbatoio* dei record candidati come donatori e il numero massimo di volte che può essere usato uno stesso record donatore;
2. *definizione delle regole formali (o strutturali)*, cioè di quelle regole che derivano direttamente dalla struttura del questionario (in particolare, dalle istruzioni di compilazione del questionario stesso). Esse esprimono condizioni di incompatibilità fra variabili, specificano cioè situazioni di non correttezza dei record. In particolare, esse indicano quando la presenza o l'assenza di risposta per una variabile o una lista di variabili risulta incompatibile con i valori assunti da variabili precedenti. Esse vengono inserite in SCIA direttamente in forma normale;
3. *definizione delle regole sostanziali*, cioè di quelle regole che derivano dalle conoscenze a priori sulle relazioni esistenti fra le variabili rilevate. Come le regole formali anch'esse

[14] Del gruppo hanno fatto parte: D.Sabatini, E.Riccini, F.Silvestri, M.Masselli, G.Barcaroli, A.Manzari

esprimono condizioni di incompatibilità fra variabili, specificano cioè situazioni di non correttezza dei record e vengono inserite in SCIA direttamente in forma normale;

4. *creazione dell'insieme completo*: questo passo è eseguito dopo aver inserito le regole originali. E' dalle caratteristiche di questo insieme (completezza, non contraddittorietà, ecc.) che dipende in massima parte la qualità dei risultati finali.

Una volta inserito l'insieme delle regole formali e sostanziali (insieme *iniziale* delle regole) per l'esecuzione della fase di editing ed i parametri utilizzati nella fase di imputazione, il sistema prevede che vengano effettuate alcune elaborazioni distinte.

Sono previsti tre diversi tipi di elaborazioni, non tutte necessarie o praticabili:

1. Generazione dell'insieme minimale di edit

La generazione dell'insieme minimale viene effettuata al fine di :

a) eliminare eventuali regole *ridondanti*, ossia regole che esprimono condizioni già implicate in altre regole;
b) segnalare regole direttamente *contraddittorie*;
c) aggregare regole che si possono combinare fra loro.

L'insieme minimale è quindi di dimensioni non superiori a quelle dell'insieme originale ed è sufficiente per il controllo dei dati. Il suo utilizzo ai fini dell'imputazione, però, non garantisce nè la correttezza dei risultati finali, nè la minimalità nel numero di correzioni che verranno effettuate sui dati: solo l'insieme completo fornisce questo tipo di garanzie. Inoltre, i tempi di elaborazione per l'imputazione di ogni record possono essere molto elevati, in quanto il sistema può dover effettuare un gran numero di tentativi prima di individuare la soluzione per la correzione del record. Maggiore è il rapporto fra il numero di edit impliciti ed il numero di edit originali, maggiore è la frequenza con cui tali inconvenienti si verificano. Nel caso in cui tale rapporto è molto basso, l'applicazione ai dati dell'insieme minimale dà risultati soddisfacenti: per questo motivo si può decidere di generare ed applicare ai dati solo l'insieme minimale, senza porsi il problema della generabilità dell'insieme completo ed, eventualmente, della suddivisione dell'insieme originale degli edit.

Le necessità di cui ai punti a, b e c fanno sì che l'insieme minimale venga sempre generato, anche nel caso in cui sia possibile generare l'insieme completo. Il ricorso ad esso per la localizzazione degli errori o la loro imputazione è perseguito nel caso in cui risulti impossibile o eccessivamente onerosa la generazione dell'insieme completo.

2. Generazione dell'insieme completo di edit

Con la generazione dell'insieme completo di edit il sistema individua tutti gli edit implicitamente contenuti nell'insieme iniziale di regole (edit *espliciti*), combinando fra loro gli edit originali secondo la metodologia di Fellegi-Holt.

Il procedimento di generazione consiste nel tentare di combinare gli edit assumendo come campo generatore via via tutte le variabili coinvolte: se questo procedimento produce nuovi edit, esso va ripetuto anche combinando gli edit nuovi con quelli preesistenti, e così via fino a quando nessun nuovo edit viene prodotto.

La generazione dell'insieme completo garantisce la creazione di un insieme di edit non contraddittorio, ed il suo utilizzo garantisce la correttezza dei dati rispetto a tali regole.

3. Suddivisione dell'insieme originale di regole

Nel caso in cui non sia stato possibile generare l'insieme completo di edit a causa dell'eccessiva complessità ed onerosità dell'operazione, occorre procedere a ridurre tale complessità suddividendo l'insieme iniziale di regole in due o più sottogruppi: tali sottogruppi saranno poi sottoposti, separatamente al processo di generazione dell'insieme completo.

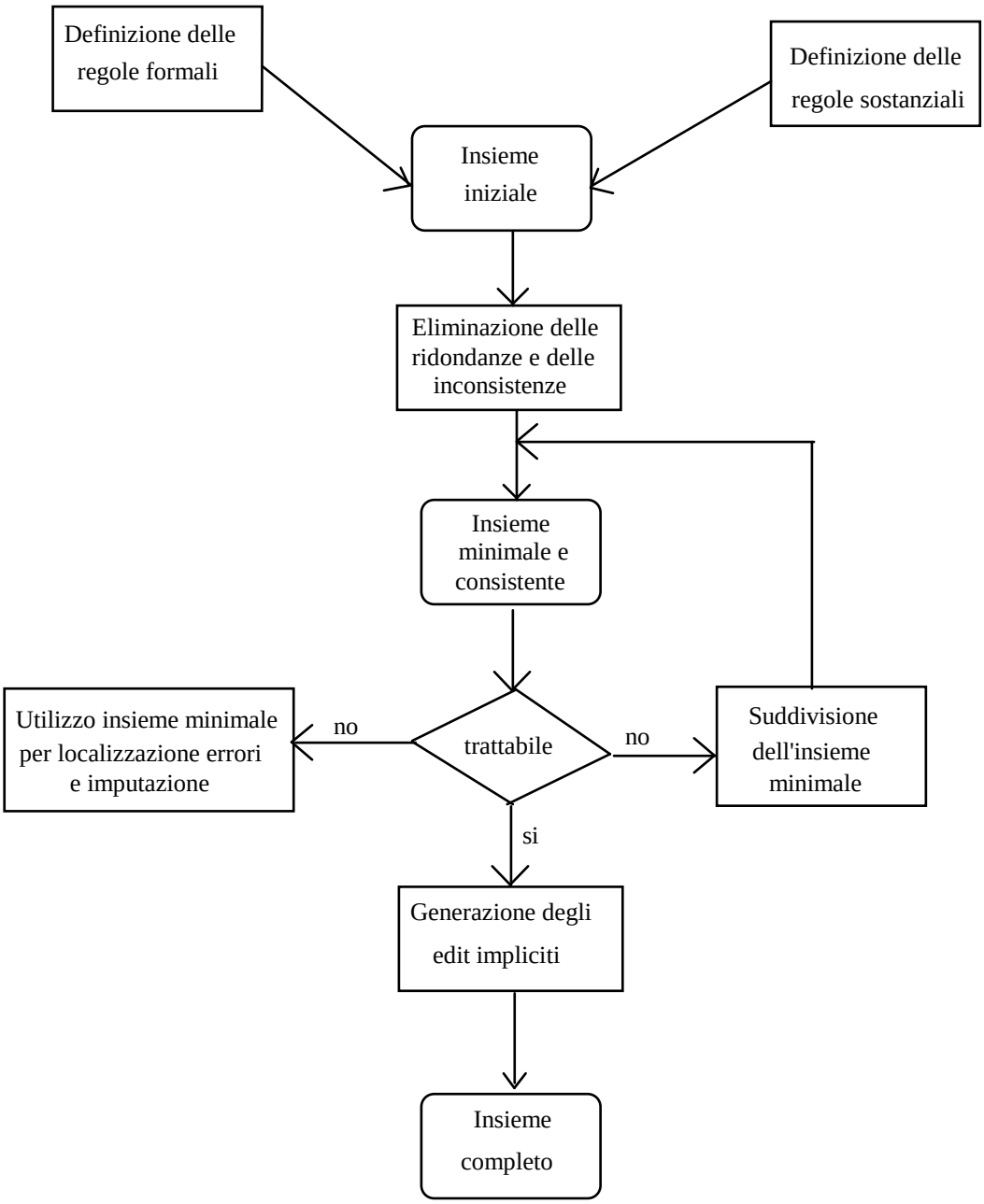
Poichè, lo ricordiamo, ad ogni insieme di regole corrisponde un processo di elaborazione dei dati (editing e imputazione), l'operazione di suddivisione dell'insieme originale di edit comporta la generazione di due o più distinti processi di elaborazione.

Se gli insiemi di regole generati sono disgiunti, l'ordine di esecuzione è ininfluente. Se invece gli insiemi di regole generati contengono variabili comuni, durante l'esecuzione di uno dei processi di elaborazione sarà necessario tenere fisse tutte le variabili imputate dal/dai processi precedentemente eseguiti.

Naturalmente la suddivisione sarà tanto migliore quanto minore è il numero di variabili comuni: il risultato ottimale è quello in cui i vari sottoinsiemi di regole risultano completamente disgiunti.

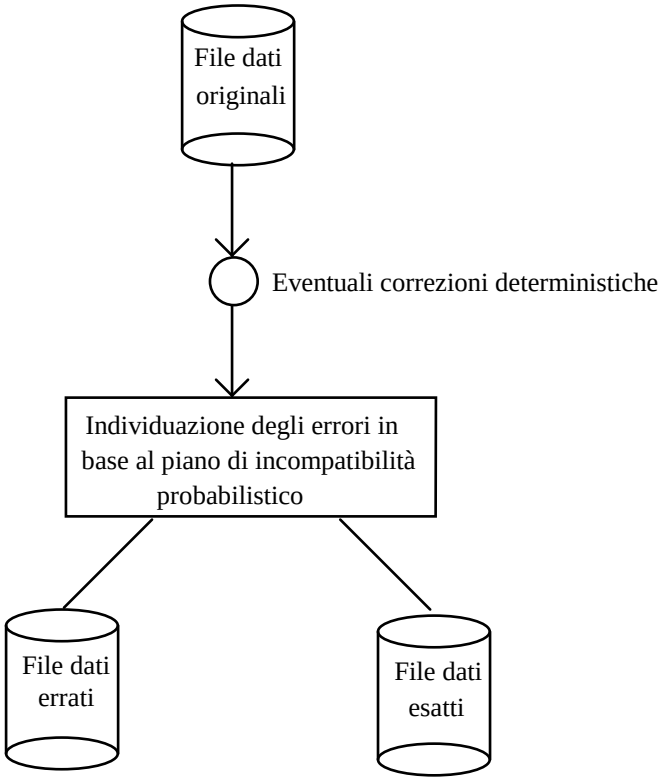
L'intero processo di definizione e messa a punto del piano di incompatibilità per una data applicazione può essere riassunto nello schema della seguente Figura 3:

Figura 3 - Flusso di definizione di un piano di incompatibilità probabilistico



Il problema dell'individuazione dei record errati consiste nel localizzare le unità statistiche in cui le variabili rilevate assumono valori tali da attivare uno o più edit del piano di incompatibilità.

Il risultato dell'operazione di individuazione degli errori nei dati rispetto ad un dato piano di incompatibilità è schematizzato nella figura seguente:



Identificati i record che violano uno o più edit, per ognuno di essi il sistema deve individuare l'insieme di variabili da modificare e l'insieme dei valori da assegnare ad esse in modo tale che siano garantite le seguenti proprietà:

i) il numero di correzioni per ogni record sia minimo;
ii) restino invariate le distribuzioni originali dei dati;
iii)il record risultante soddisfi tutti gli edit.

Il problema di cui al punto i) viene risolto da SCIA implementando l'algoritmo proposto da Fellegi-Holt (vedi relativa appendice): l'insieme minimo di variabili da imputare viene determinato attraverso l'identificazione di quelle variabili che "coprono" tutti gli edit attivati dal record errato. L'utente può comunque impedire o rendere meno probabile l'inserimento di una o più variabili nell'insieme minimale, assegnando a ciascuna di esse un grado di fissità (da 1 a 9) dipendente dalla probabilità di errore pevista per tali variabili[15].

Per quanto riguarda i problemi di cui ai punti ii) e iii), essi trovano soluzione all'interno degli algoritmi implementati in SCIA per l'imputazione dei dati.

In particolare, SCIA offre tre possibili strategie di correzione:

1. *imputazione congiunta*;
2. *imputazione sequenziale* ;
3. *imputazione basata sulle distribuzioni marginali* o *imputazione forzata*.

Le prime due sono strategie di correzione del tipo "da donatore", mentre la terza tecnica è basata sull'analisi e sull'utilizzo delle distribuzioni marginali semplici rilevate nell'indagine per le variabili dell'insieme minimale.

[15] Valore di fissità pari a 9 deve sempre essere assegnato a quelle variabili che sono state già editate e corrette in passi precedenti della procedura e che, pertanto, non sono più modificabili

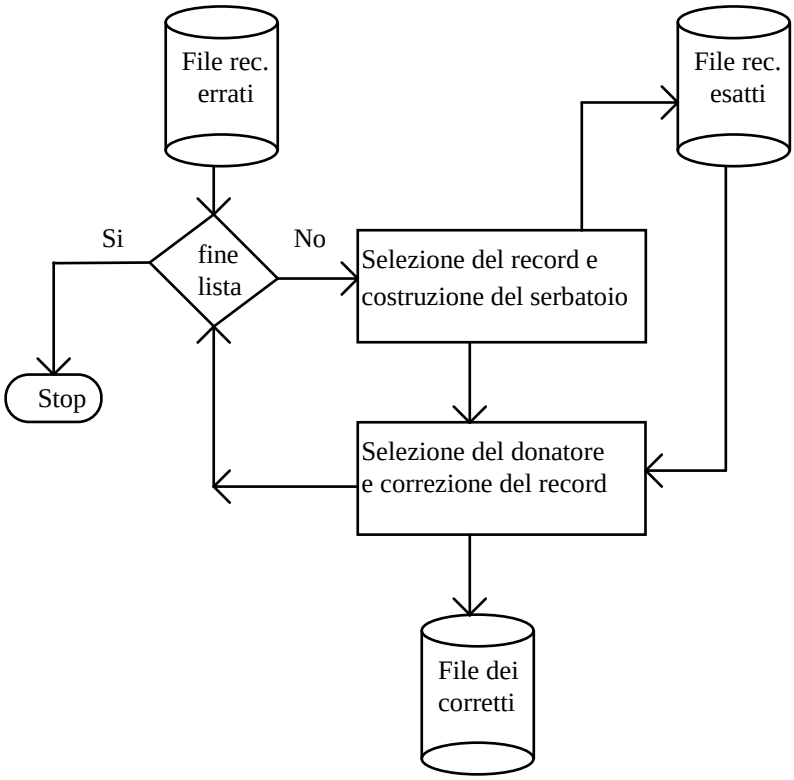
Per quanto riguarda la correzione dei dati, in SCIA è presente una procedura generale di imputazione che, implementando al suo interno (sequenzialmente) le suddette strategie di correzione, ha una struttura indipendente dalle caratteristiche delle strategie stesse. Le caratteristiche delle diverse tecniche di imputazione agiscono infatti solo all'interno di alcune delle fasi componenti la procedura di correzione stessa.

Tale procedura generale è costuita da due fasi principali:

- selezione del record errato e costruzione di un "*serbatoio*" di record donatori, costruzione dipendente dalle specifiche assegnate dall'utente tramite i *parametri*;
- scelta del donatore e correzione dei record, le cui modalità dipendono dal tipo di algoritmo di correzione utilizzato.

Si tenga presente che, mentre il meccanismo di costruzione del serbatoio può essere controllato dall'utente (appunto attraverso i parametri), la particolare strategia di imputazione che il sistema adotterà per l'effettiva correzione di un certo record dipende quasi esclusivamente da criteri ed elaborazioni interni al sistema stesso[16].

La struttura di base del sistema è schematizzata nella seguente figura:



Come già detto, in SCIA sono implementate due metodologie di imputazione da donatore:

- imputazione congiunta;
- imputazione sequenziale.

La prima tecnica, in particolare, prevede le due seguenti versioni:

[16] L'unica opzione che consente all'utente di modificare la sequenza di applicazione delle diverse tecniche di correzione è rappresentata dalla specificazione di distribuzioni marginali per date variabili.

1. *imputazione congiunta ristretta*, in cui, dato un certo record errato, vengono selezionati come possibili donatori quei record che possiedono, per le *variabili di accoppiamento*[17], valori identici a quelli contenuti nel record errato;
2. *imputazione congiunta allargata*, in cui, dato un certo record errato, vengono selezionati come possibili donatori quei record che possiedono, per le *variabili di accoppiamento*, valori contenuti nei corrispondenti intervalli (*range*) opportunamente determinati.

Nel caso di *imputazione sequenziale* si procede ad imputare una variabile alla volta: per ciascuna variabile appartenente all'insieme minimo viene calcolato il *range* dei valori ammissibili; per ciascuna di esse viene quindi cercato nel serbatoio e, se esiste, selezionato un record donatore con valore compreso nel corrispondente *range*.

Riassumiamo quindi le varie fasi di cui si compone il processo di correzione dei dati con tecniche da donatore:

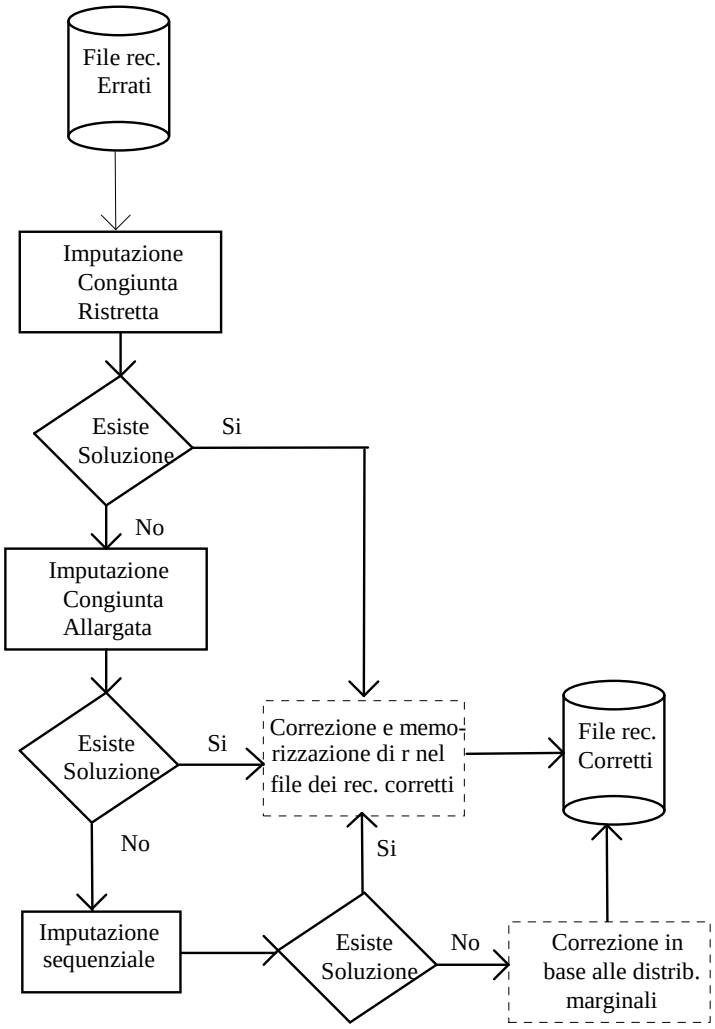
I. *Inviduazione dell'insieme minimale.* In questa fase, dato il record errato r, viene determinato il minimo numero di variabili da correggere tra quelle presenti in tutti gli edit attivati da r.
II. *Controllo delle variabili marginali*. Prima di procedere alla ricerca del donatore, il sistema verifica la presenza di qualcuna delle variabili dell'insieme minimale all'interno della lista di variabili specificata come marginali. In caso positivo, tali variabili vengono corrette direttamente col metodo dell'imputazione forzata, e si procede alla ricerca del donatore per la correzione delle variabili residue dell'insieme minimale.
III. *Selezione del donatore*. In questa fase, a seconda della strategia di imputazione adottata, viene selezionato dal serbatoio il record donatore d[18].

La sequenza in cui vengono applicate le tre metodologie di imputazione da donatore è schematizzata nella seguente Figura 4 (la fase relativa all'imputazione basata sulle distribuzioni marginali è stata tratteggiata per sottolineare il fatto che essa viene applicata, eventualmente, al termine del ciclo di correzione da donatore).

[17] Ricordiamo che le *variabili di accoppiamento* sono le variabili che compaiono negli edit attivati o attivabili, ma non appartengono all'insieme minimo

[18] Nel caso in cui non sia stato possibile individuare un donatore adatto con nessuna delle tecniche da donatore previste, il record viene corretto mediante imputazione forzata

Figura 4 - Ciclo di applicazione delle strategie di imputazione da donatore



Nel caso in cui, per un certo record errato r, non sia stato possibile individuare un donatore adatto con nessuna delle tecniche da donatore disponibili in SCIA, il sistema corregge automaticamente le variabili dell'insieme minimale utilizzando le corrispondenti distribuzioni marginali semplici (*correzione forzata*). La correzione di tali variabili avviene sequenzialmente.

Per la correzione di una o più variabili l'utente può anche decidere di non tentare affatto la correzione basata sulle tecniche da donatore, può cioè richiedere al sistema di sottoporre direttamente tali variabili al metodo dell'imputazione forzata specificandole come marginali.

La tecnica di correzione forzata è basata su un algoritmo random di estrazione del valore da assegnare alla variabile errata (selezionato tra i valori ammissibili), estrazione guidata da una funzione di probabilità definita sulla base della distribuzione di frequenze che la variabile stessa assume nel file dei dati originari.

Gli eventuali pesi da assegnare alle modalità delle variabili da correggere mediante imputazione forzata devono essere specificati.

L'uso di regole di tipo deterministico è previsto solo nel caso si debbano correggere errori di tipo sistematico: questo tipo di errori derivano, generalmente, da problemi strutturali nel questionario, nell'organizzazione della rilevazione, nella registrazione dei dati.

La presenza di errori sistematici nei dati viene generalmente verificata attraverso un'analisi delle imputazioni probabilistiche effettuate dal sistema, analisi condotta possibilmente in fase di test del piano di incompatibilità.

Questa analisi viene condotta sulla base dei report che SCIA produce automaticamente al termine del processo di correzione.

6.2.2.2. GEIS (Generalised Edit and Imputation System)

GEIS (Generalized Edit and Imputation System) è un prodotto sviluppato da Statistics Canada: esso implementa una metodologia di controllo ed imputazione per variabili quantitative basata sulla metodologia di Fellegi-Holt (1976), e adotta un approccio, proposto da Sande (1978 e 1979), che utilizza metodi propri della programmazione lineare per risolvere problemi di ottimizzazione nell'ambito di variabili rilevate in indagini statistiche.

In particolare, per risolvere alcuni specifici problemi di ottimo, GEIS utilizza un algoritmo sviluppato da Chernikova (1964 e 1965) e generalizzato da Rubin (1975).

GEIS è realizzato in linguaggio C e richiede la presenza di un DBMS ORACLE per la costruzione del data base nel quale immagazzinare i dati da sottoporre a controllo e correzione. Di conseguenza, l'utente di GEIS deve essere anche un utente ORACLE, e deve essere in grado di utilizzare il linguaggio SQL.

Attualmente, GEIS è disponibile sulle tre piattaforme più diffuse, e cioè in ambiente mainframe IBM, in ambiente UNIX ed in quello DOS per personal computer. Può essere utilizzato nelle applicazioni in cui :

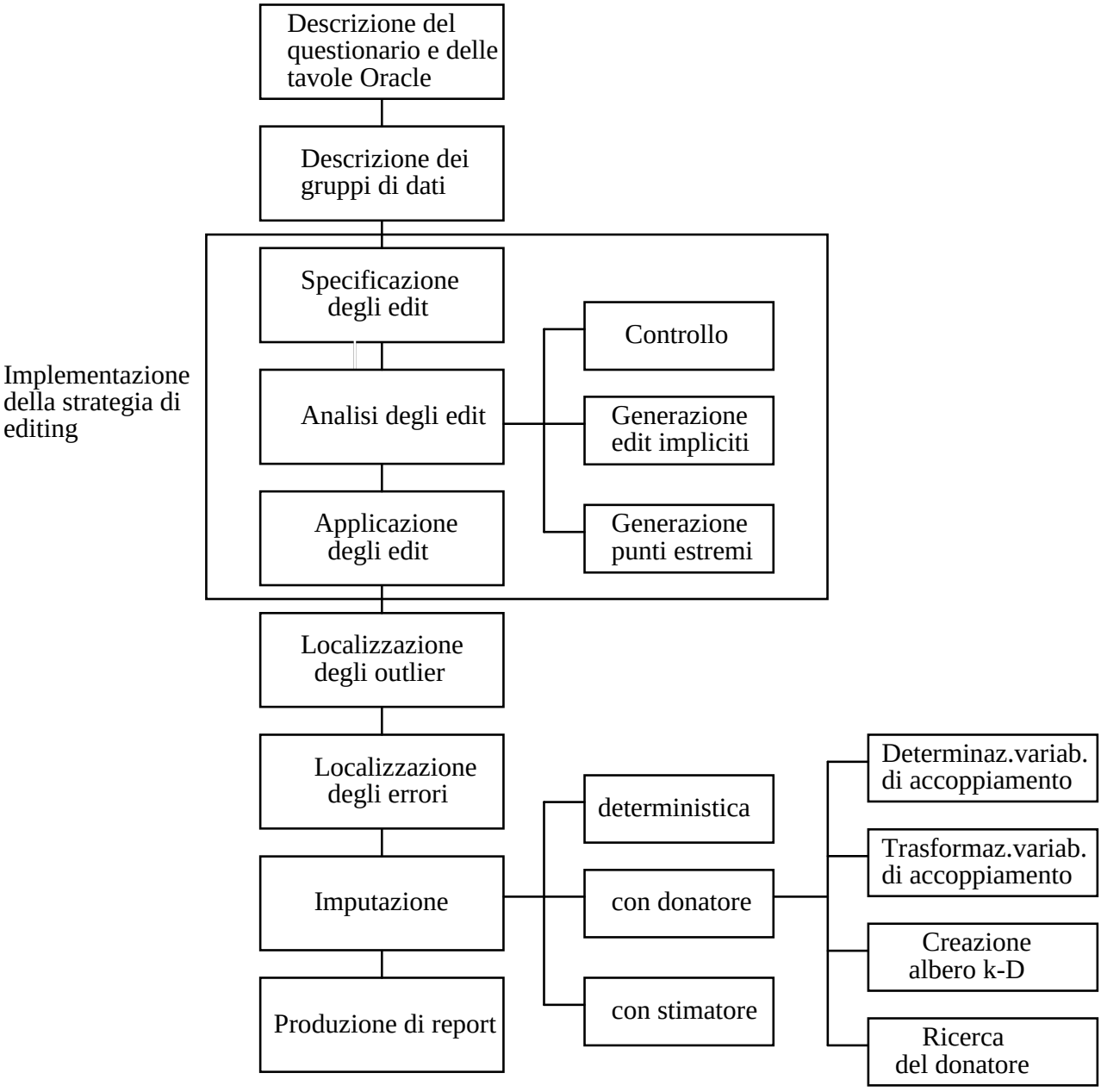
- tutte le variabili sono *numeriche*, *continue* e *non negative;*
- i vincoli possono essere espressi sotto forma di disequazioni *lineari*.

L'utilizzo di GEIS presuppone che parte delle correzioni e dei controlli sui dati (in particolare, quelli relativi agli errori di tipo sistematico) siano stati effettuati in una fase preliminare di editing, e che, quindi, solo i casi residui vengano sottoposti ad imputazione automatica.

Un aspetto di GEIS cui prestare particolare attenzione è l'efficienza: il costo dell'elaborazione aumenta considerevolmente all'aumentare del numero di edit e di record da sottoporre a controllo.

GEIS è costituito da un insieme di *moduli*, ciascuno dei quali implementa una particolare sottofunzione di una delle funzioni principali (editing, individuazione degli errori, imputazione, individuazione degli outlier) in esso implementate.

La struttura di GEIS può essere schematizzata come segue:



Nel passo di descrizione del questionario e delle tavole Oracle vengono effettuate le operazioni di descrizione della tavola ORACLE contenente il file di dati da sottoporre a imputazione, e della eventuale tavola ORACLE contenente il file di dati "storici" da utilizzare nelle fasi di individuazione degli outlier o di imputazione mediante stimatori.

La descrizione dei gruppi di dati consiste nella definizione (mediante espressioni SQL) dei gruppi di dati sui quali effettuare eventualmente imputazioni distinte. La necessità di separare l'insieme globale dei dati in sottogruppi distinti può derivare dall'esigenza di ridurre il costo dell'elaborazione e/o dall'opportunità di sottoporre a controlli ad hoc sottoinsiemi particolari di dati.

La fase di implementazione della strategia di editing prevede la strutturazione ed il controllo di validità dell'insieme di edit specifico per l'indagine che si vuole sottoporre a controllo. A questo scopo sono state predisposte le funzioni descritte di seguito.

Specificazione degli edit

Attraverso gli edit, l'utente fornisce la descrizione del record corretto: gli edit esprimono infatti le regole che ogni record deve soddisfare per essere considerato corretto.
Nessuna azione viene specificata a fronte dell'eventuale violazione di un dato edit, quindi gli errori hanno tutti lo stesso "peso": è il sistema stesso che, a fronte di una violazione, identifica i singoli campi da modificare.

Nel loro insieme, gli n edit definiscono la *regione di accettazione* in R^n, convessa e contenente i confini.
In GEIS gli edit sono espressi come condizioni di correttezza o di errore (PASS/FAIL), e sono rappresentati attraverso uguaglianze o disuguaglianze lineari della forma:

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i \quad \text{oppure:} \quad \sum_{j=1}^{n} a_{ij}x_j = b_i \qquad i=1,...,m$$

dove x_j sono le n variabili risposta rilevate su una unità campionaria, m è il numero di edit, ed i valori a_{ij} , b_i (j=1,...,n ; i=1,..,m) sono costanti specificate dall'utente.

All'insieme degli edit così definiti, GEIS aggiunge automaticamente i vincoli di positività sulle variabili.

Nel caso di edit non lineari, è necessario verificare:

- che tra le variabili risposta esista effettivamente una relazione non lineare;
- la possibilità di linearizzare tale relazione;
- l'opportunità di inserire nell'analisi tale relazione.

Ad esempio, l'edit

$$x_1 * x_2 = x_3$$

può essere sostituito dall'edit

$$y_1 + y_2 = y_3$$

dove:

$$y_1 = \log x_1$$
$$y_2 = \log x_2$$
$$y_3 = \log x_3$$

(si noti che tale operazione è possibile solo se x_1, x_2 ed x_3 non compaiono in altri edit).

Poichè la fase di individuazione degli errori prevede un limite nel numero di variabili che possono essere trattate contemporaneamente, nel caso di indagini di grandi dimensioni è necessario suddividere le variabili in due o più gruppi di edit, possibilmente indipendenti fra loro. Questi insiemi devono essere disgiunti, cioè non possedere variabili in comune: in caso contrario, le variabili modificate nel corso dell'applicazione di uno di essi devono essere tenute fisse nel corso dell'applicazione di tutti i gruppi di edit successivi.

Analisi degli edit

L'analisi degli edit viene condotta esaminando la regione di accettazione da essi delimitata, che deve essere convessa e contenere la frontiera.Viene in primo luogo determinato l'*insieme consistente minimale* degli edit, che definisce la regione di accettazione e che verrà utilizzato nella fase di individuazione dei record errati. Viene verificato che:

- l'insieme di edit sia *consistente*, cioè la regione di accettazione non sia vuota: l'eliminazione delle inconsistenze avviene per tentativi, sottraendo edit dall'insieme originale e controllando se la nuova regione di accettazione è vuota o meno;
- non ci siano *ridondanze*, cioè non esistano edit che non contribuiscono a definire la regione di accettazione. Gli edit ridondanti possono essere direttamente eliminati;
- non ci siano *uguaglianze nascoste*, cioè uguaglianze implicite nell'insieme di edit;
- la presenza di *variabili determinate*, variabili cioè che possono assumere un solo valore.

Ciascuna delle suddette verifiche viene effettuata risolvendo in successione particolari problemi di programmazione lineare del tipo:

$$\max (\min)\ S(x) : C'x$$

in cui la funzione S(x) è lineare nelle variabili x ed è soggetta ai vincoli lineari:

$$Ax \leq b$$

ed ai vincoli di positività:

$$x \geq 0.$$

Nelle espressioni precedenti si è indicato con C la matrice dei coefficienti e dei termini noti, con x la matrice delle variabili rilevate, e con A la matrice dei coefficienti delle combinazioni lineari che definiscono gli edit.

Vengono quindi calcolati i cosiddetti *punti estremi* cioè i vertici della regione di accettabilità descritta da un gruppo di edit (essi possono essere visti come i record che sono ai limiti dell'accettabilità). Geometricamente, i punti estremi corrispondono alle intersezioni fra n edit nello spazio n-dimensionale, e vengono individuati per rendere più comprensibile la struttura della regione di accettabilità. Essi vengono determinati utilizzando l'algoritmo di Chernikova.

Si passa quindi a generare gli *edit impliciti*, che rappresentano relazioni implicitamente contenute in un dato gruppo di edit. Nel caso di variabili quantitative, tali edit sono sempre ridondanti e non vengono mai aggiunti all'insieme minimale di edit: in questo contesto essi sono utilizzati solo per consentire una migliore esplicitazione, a livello concettuale, degli edit originali. Formalmente, un edit implicito è ottenuto da una combinazione lineare di k edit in cui almeno (k-1) variabili siano state eliminate. Essi vengono generati utilizzando, analogamente al caso dei punti estremi, una particolare versione dell'algoritmo di Chernikova.

A questo punto, gli edit così perfezionati vengon applicati ai record da trattare, producendo delle tabelle di statistiche riassuntive che forniscono informazioni, ad esempio, su quanti record violano le regole di controllo predisposte, quali di queste sono violate e quante volte, ecc.

Individuazione degli outlier

Questa funzione è notevolmente diversa dalle altre funzioni di controllo implementate in GEIS, in quanto effettua un tipo di analisi *inter-record*, e non *intra-record* come nel caso degli edit sopra descritti. In pratica, il controllo consiste nel confrontare i valori delle variabili

di interesse in record diversi (controllo verticale), invece che nel verificare la coerenza delle variabili all'interno di uno stesso record (controllo orizzontale).

In realtà, questo modulo può essere utilizzato anche da solo, cioè senza alcun legame con le successive fasi di elaborazione, solo allo scopo di individuare gli eventuali outlier: solo dopo averli analizzati si può decidere il loro trattamento, se controllarli manualmente o se integrarli nel processo di imputazione (cioè se sottoporli a imputazione automatica, se includerli o meno nella popolazione dei possibili donatori, se escluderli o meno dal calcolo degli eventuali totali per il ricorso all'imputazione con stimatori).
Pertanto, l'insieme dei record considerati non corretti comprende, oltre ai record che violano un qualche edit, anche tutti quei record che presentano, per qualcuna delle variabili rilevate, valori al di fuori di prefissate soglie (valori outlier).

Il metodo utilizzato è dovuto a Hidiroglou-Berthelot (1986), ed individua i valori outlier relativi ad una certa variabile confrontando i valori di tale variabile all'interno di un insieme di record: fissati i limiti (superiore e inferiore) considerati accettabilitali per ogni variabile da sottoporre a controllo, vengono così identificati quelli fra loro che fuoriescono da tali limiti.

I valori riconosciuti come outlier possono essere o sottoposti a imputazione (valori ODI), analogamente ai record localizzati in fase di individuazione degli errori, oppure accettati come corretti ma non utilizzati come possibili donatori in fase di imputazione (valori ODE).

Per l'individuazione degli outlier sono previsti due metodi:

- *metodo dei dati correnti:* i valori della variabile selezionata vengono confrontati con limiti di accettabilità calcolati sulla base dei valori che essa assume in altri record in uno stesso periodo di riferimento;
- *metodo del trend storico*: in presenza di dati "storici", per la variabile selezionata viene calcolata in ogni record una funzione del suo trend storico: i limiti di accettabilità vengono determinati in questo caso sulla base della stessa funzione calcolata su tutti i record del file.

Localizzazione degli errori

Il legame tra la fase di editing e quella di imputazione è costituito dalla funzione di individuazione degli errori, che identifica il minimo numero di campi da correggere (positivamente o negativamente) affinché ogni record errato sia riportato nella condizione di soddisfare tutti gli edit.

Un record è considerato non corretto se cade al di fuori della regione di accettazione associata all'insieme minimo di edit.
In generale, per riportare il record (vettore x) nella regione di accettabilità, è necessario applicare ad x una correzione positiva (vettore y) o negativa (vettore z) in modo tale che:

1. il record (x+y-z) sia corretto;
2. ogni volta che $y_i>0$ allora $z_i<0$;
3. la cardinalità f del vettore (y-z) sia minima.

Il problema dell'individuazione dei campi da imputare all'interno dei record errati viene dunque formulato in GEIS come un problema di programmazione lineare con il vincolo di minima cardinalità della soluzione, seguendo la metodologia di Sande, e viene risolto utilizzando l'algoritmo di Chernikova, così come generalizzato da Rubin.

Un ruolo importante gioca, in questa fase, il costo in termini di tempo impiegato per risolvere il problema dell'individuazione dei record errati e della soluzione ottimale per ciascuno di essi: si può verificare, infatti, che per alcuni record il sistema non riesca a trovare una soluzione accettabile a causa del limite di tempo concesso. In questi casi è necessario effettuare un'elaborazione ad hoc concedendo un limite tempo maggiore.

Correzione degli errori

Identificati i campi da modificare in un dato record, il sistema trova quei valori che, sostituiti ai valori originali, garantiscono che il record risultante soddisfi tutti gli edit, e che resti invariata la struttura originale dei dati.

Per l'imputazione GEIS offre tre possibili strategie:

1. imputazione deterministica;
2. imputazione con donatore;
3. imputazione con stimatori.

In fase di *correzione deterministica* viene analizzata ogni variabile fra quelle imputabili, e si verifica se esiste uno ed un solo valore che, una volta assegnato, riporti il record nella regione di accettazione. Se tale valore esiste, esso viene direttamente assegnato alla variabile. La procedura prevede i seguenti passi:

1. vengono presi in considerazione i soli edit *attivi* nel record x (ricordiamo che un edit è attivo per un dato record se è violato dal record stesso);
2. in questi edit si procede ad assegnare, alle variabili che non devono essere imputate, il loro valore nel record: il record (vettore) x può dunque essere scomposto come $x=(x_i, x_r)$, dove x_i è il sub-vettore delle variabili da imputare, ed x_r è il sub-vettore delle altre variabili;
3. viene definito il *sistema ridotto* di vincoli
$$A_{1i}\, x_i \leq b_{1i}$$
$$A_{2i}\, x_i = b_{1i}$$
$$x_i \geq 0$$
4. per ogni variabile x_i da imputare vengono calcolati i valori massimo e minimo risolvendo il sistema ridotto con funzione obiettivo, rispettivamente, $Max(x_i)$ e $Min(x_i)$;
5. se $Max(x_i) = Min(x_i) = \bar{x}_i$,allora il valore $\bar{x}_i$ viene assegnato alla variabile x_i.

L'*imputazione con donatore* consiste nell'individuare, per ogni record errato (*ricevente*), il record *donatore* "più vicino" ed i cui valori consentono al recipiente di soddisfare tutti gli edit.

Nel metodo del "record più vicino" la distanza fra ricevente e donatore viene calcolata sulla base delle *variabili di accoppiamento*. Tale metodo garantisce il rispetto delle distribuzioni semplici e congiunte fra le variabili ed è basata su tre passi principali:

1. selezione delle variabili di accoppiamento;
2. trasformazione dei valori e calcolo della distanza;
3. creazione dell'albero k-D, ricerca del record donatore e imputazione.

Nel passo di *selezione delle variabili di accoppiamento* si procede ad individuare un insieme di variabili (dette di accoppiamento) da utilizzare nel calcolo della distanza tra record ricevente e record donatore.Queste variabili devono :

1. non richiedere imputazione;

2. figurare negli edit violati dal recipiente;
3. possibilmente, essere correlate con le variabili da imputare.

A questo insieme di variabili l'utente può aggiungere altre variabili di accoppiamento (ad esempio nel caso in cui esistano variabili molto correlate con quelle da imputare, ma che non compaiono esplicitamente nel gruppo di edit sottoposto a verifica).

Nel passo di *trasformazione dei valori e calcolo della distanza*, le variabili di accoppiamento vengono trasformate in modo da essere ricondotte ad una scala comune, al fine di evitare effetti di scala nel calcolo delle distanze (peso eccessivo sulla distanza di variabili con range molto ampio). Per il calcolo della distanza, GEIS utilizza la metrica in L^{∞} : la distanza fra due record x ed y è definita come il valore

$$D(x,y) = \max(|x_1-y_1|, |x_2-y_2|,..., |x_n-y_n|)$$

Tale distanza è nota come *distanza minimax*, in quanto fra i record donatori viene scelto quello con la più piccola differenza assoluta massima tra i valori trasformati delle variabili di accoppiamento nel record donatore stesso e nel record recipiente.

La creazione dell'albero k-D è propedeutica all'individuazione, per un dato record errato, del donatore più vicino i cui valori riportino il recipiente nella regione di accettazione.

Al fine di rendere più efficiente tale ricerca, GEIS implementa un metodo basato sulla costruzione di un albero di ricerca binario, l'albero k-D. Per ogni record errato x, la ricerca viene allora effettuata come segue:

1. viene esplorato l'albero per identificare gli n_1 donatori più vicini (cioè quelli con D(x,y) minima) fra gli n donatori possibili;
2. cominciando dal donatore più vicino, si verifica se esiste qualche donatore i cui valori, se imputati, consentono al ricevente x di soddisfare gli edit di post-imputazione: se tale donatore viene trovato, l'imputazione viene effettuata;
3. altrimenti, l'albero viene attraversato una seconda volta e viene selezionato un gruppo di n_2 donatori più vicini fra gli $(n-n_1)$ donatori residui;
4. si ripete la ricerca come al passo 2: se il donatore viene trovato, l'imputazione viene effettuata, altrimenti il modulo avverte che non è stato individuato alcun donatore e che, quindi, l'imputazione con tale metodo è impossibile.

Nel caso in cui il ricevente non possiede variabili di accoppiamento, la scelta del donatore è di tipo completamente casuale.

Nella ricerca del record donatore gioca un ruolo decisivo la predisposizione e l'utilizzo degli *edit di post-imputazione*, generalmente ottenuti "rilassando" edit originali. Questo tipo di operazione consente, in generale, di allargare la rosa dei potenziali donatori per un record che violi un certo edit originale.

Il caso più frequente di rilassamento consiste nel trasformare una uguaglianza in due disuguaglianze, definendo così un limite inferiore e uno superiore per il valore esatto di confronto.

Dato un record x che fallisce una certa uguaglianza può accadere, infatti, che il sistema:

1. non riesca a trovare un donatore che fornisca ad x un valore tale da riportarlo nella condizione di soddisfare l'uguaglianza;
2. scarti potenziali record donatori molto "vicini" ad x, ma che non gli garantiscono il rispetto dell'uguaglianza stretta, e seleziona un donatore che non è "vicino" ad x, ma gli fornisce il valore richiesto.

L'unico vincolo imposto da GEIS è che il gruppo di edit di post-imputazione e il gruppo di edit originali corrispondente contengano le stesse variabili.

Se da un lato l'uso degli edit di post-imputazione facilita l'individuazione del record donatore, dall'altro esso rende necessaria una successiva verifica dei dati imputati per controllare se qualcuno di essi, pur soddisfacendo gli edit di post-imputazione, violi i vincoli di uguaglianza originali. Generalmente, i casi di violazione con questa origine vengono risolti col metodo di imputazione deterministica.

Imputazione con stimatori

In questo modulo è prevista l'imputazione di una variabile alla volta utilizzando vari tipi di stimatori: rapporti, medie (correnti e storiche), serie storiche (con e senza aggiustamento del trend).

E' anche possibile limitare il calcolo delle eventuali medie solo sui valori accettabili della variabile da correggere o su particolari sottoinsiemi di record.

Va sottolineato che, poiché gli stimatori vengono applicati indipendentemente per ogni variabile, i record così imputati possono non soddisfare l'insieme originale di edit: occorrerà quindi sottoporre nuovamente questi record alla fase di localizzazione degli errori, sfruttando le opzioni previste da GEIS per la rielaborazione dei dati.

Siano: X la variabile da imputare, t il periodo di riferimento, (t-1) il periodo precedente, x_{it} il valore assunto da X nel record i al tempo t, Y la variabile ausiliaria (correlata con X), $\overline{X}$ il valore medio di X (al tempo t o t-1) ed $\overline{Y}$ il valore medio di Y (al tempo t o t-1). Gli stimatori previsti in GEIS sono i seguenti:

1. VALORE PRECEDENTE: $x_{it} = x_{i(t-1)}$
2. MEDIA PRECEDENTE: $x_{it} = \overline{X}_{(t-1)}$
3. MEDIA CORRENTE: $x_{it} = \overline{X}_t$
4. RAPPORTO CORRENTE: $x_{it} = \overline{X}_t \frac{y_{it}}{\overline{Y}_t}$
5. TREND AUSILIARIO: $x_{it} = x_{i(t-1)} \frac{y_{it}}{y_{i(t-1)}}$
6. TREND DIFFERENZA: $x_{it} = x_{i(t-1)} \frac{\overline{X}_t}{\overline{X}_{(t-1)}}$

Produzione di report

In questa fase viene data all'utente la possibilità di generare una serie di report relativi all'applicazione effettuata (ad esempio, frequena degli operatori relazionali utilizzati o delle variabili presenti negli edit, lista egli edit che non appartengono a nessun gruppo di edit, lista degli stimatori utilizzati per l'imputazione ecc.).

Rielaborazione dei dati

Se qualche record viola ancora qualche edit, si può ripetere la fase di localizzazione degli errori limitandosi a considerare le sole variabili già imputate.

Un record imputato può non risultare corretto:

1. perché la fase di imputazione da donatore è stata eseguita con edit di "post-imputazione" rilassati rispetto a quelli iniziali.
2. perché è stato utilizzato il metodo di imputazione con stimatori.

In questi casi GEIS prevede varie opzioni per ripetere l'operazione di ricerca degli errori solo per sottoinsiemi di record (ad esempio, per i soli record con almeno una variabile imputata con donatore, per i soli record con almeno una variabile imputata con stimatori, ecc.)

E' anche possibile rielaborare tutti quei record per i quali nella fase iniziale di localizzazione dell'errore non sia stata trovata una soluzione a causa del limite di tempo prefissato.

6.2.2.3. SPEER (Structured Programs for Economic Editing and Referrals)

Lo SPEER è un sistema costituito da un insieme di programmi che permettono l'applicazione dei principi di base della metodologia Fellegi-Holt a indagini che rilevano variabili di tipo quantitativo. Di tale metodologia sono rispettati in particolare i seguenti elementi:

1. garanzia di correttezza finale;
2. minimo cambiamento apportato ai dati.

(Non è invece sempre garantito il rispetto della distribuzione multivariata originale dei dati, dipendendo dalla particolare modalità di imputazione dei dati).

Così come nel caso delle variabili qualitative, dagli edit inizialmente definiti vengono generati tutti quelli in essi implicitamente contenuti, fino ad ottenere l'insieme completo. Sulla base di tale insieme, ogni record viene analizzato per stabilire se e quali edit vengono violati, e la determinazione delle variabili da modificare e degli intervalli di valori da assegnare ad esse è fatta in modo tale da disattivare gli edit violati senza attivarne nel contempo altri. Ogni variabile da modificare riceve il valore, interno all'intervallo di accettazione predefinito, secondo procedure personalizzabili a seconda delle situazioni concrete.

L'unica forma ammissibile degli edit è quella del rapporto tra variabili. Sia $(x_1, \ldots, x_n)$ un vettore rappresentante un record con variabili quantitative. Un edit-rapporto (*ratio edit*) esprime il vincolo (di compatibilità) che il quoziente tra i valori assunti nel record da due variabili x_i e x_j giaccia in un intervallo delimitato da un limite inferiore L_{ij} e un limite superiore U_{ij} :

$$L_{ij} \leq x_i / x_j \leq U_{ij}$$

Lo statistico può definire k edit di questo tipo. Tipicamente, la definizione dei rapporti di interesse rispetta i seguenti passi:

1. tra gli n (n-1) / 2 possibili edit di questo tipo che è possibile stabilire tra le n variabili, vengono definiti quelli tra le coppie di variabili maggiormente correlate: a tal fine si

costruisce la matrice dei coefficienti di correlazione, e per ogni variabile x_i si definiscono edit che la pongono in relazione con le variabili maggiormente correlate, con limiti inferiori L_{ij} e superiori U_{ij} ancora indefiniti;

2. per determinare i valori dei limiti inferiori e superiori, si ricorre ad uno strumento per la determinazione della regione di accettazione dei quozienti che massimizza la probabilità di escludere i soli outlier.

Lo strumento di cui al punto 2 è il *Distance Measurement Algorithm for Selection of Outliers* (D-MASO). Il requisito per l'applicazione di questo metodo è che la distribuzione del quoziente da circoscrivere sia di tipo normale. L'assunto alla base dell'algoritmo è il seguente: anzichè scegliere un intervallo di ampiezza tale da escludere una percentuale fissa di casi nelle code di sinistra e di destra della distribuzione ("blind cutoff"), il che può portare o a includere valori che sono outlier o ad escluderne di ammissibili, è preferibile procedere all'individuazione di quelli che sono i primi probabili outlier procedendo dal centro della distribuzione verso sinistra e destra. L'individuazione si basa sulla considerazione degli *intervalli* tra i valori dei quozienti: l'ipotesi è che *se c'è un intervallo tra due valori prossimi ad una estremità della distribuzione che è significativamente più ampio della maggior parte degli altri intervalli, il più estremo dei due valori è probabilmente un outlier*. Il meno estremo dei due valori è allora assunto come limite (inferiore o superiore, a seconda che ci si trovi nella coda sinistra o in quella destra) dell'intervallo di accettazione.

L'algoritmo di D-MASO è composto dai seguenti passi:

a) data la distribuzione dei valori dei quozienti, se ne calcolano le distanze: queste possono essere di tipo *additivo* (differenze) o *proporzionale* (rapporti). L'esperienza ha dimostrato che è più conveniente considerare le distanze di tipo "rapporto";
b) i valori dei rapporti sono ordinati in senso ascendente, ed accanto sono listati i rapporti tra il quoziente corrente e quello successivo;
c) viene determinato un *cutoff* tale che tutte le distanze maggiori di questo devono essere analizzate come possibili identificatrici di outlier. Tale cutoff è dato dalla distanza mediana (che di per sè individua metà dei rapporti) moltiplicato per un opportuno fattore moltiplicativo che permetta di restringere l'arco delle possibilità: nella pratica, questo fattore è posto a 1.2, il che permette di restringere l'analisi ad una percentuale di rapporti che va dal 25% al 5%;
d) viene inoltre scelto un α che è la massima percentuale di casi in una delle due code esclusi dai limiti dell'intervallo di accettazione da scegliere;
e) a questo punto, viene scelto come limite inferiore (superiore) dell'intervallo di accettazione quel valore che, essendo più vicino al centro della distribuzione, risponde ai seguenti requisiti: ha una distanza dal precedente (successivo) maggiore del cutoff e taglia non più dell'α% dei casi nella coda sinistra (destra).

Una volta definiti gli edit come intervalli di accettazione per i quozienti di variabili quantitative, analogamente a quanto prescritto dalla metodologia Fellegi-Holt per gli edit in forma normale relativi alle variabili qualitative, si procede alla generazione dell'insieme completo di edit, composto dagli edit iniziali depurati da eventuali ridondanze e dagli edit *impliciti*, quegli edit cioè implicitamente contenuti negli iniziali.

Ad esempio, definiti i seguenti due edit iniziali:

$$L_{12} \leq x_1/x_2 \leq U_{12}$$

e

$$L_{23} \leq x_2/x_3 \leq U_{23}$$

è possibile derivare il seguente edit implicito:

$$L_{12}\, L_{23} \;\leq\; x_1/x_3 \;\leq\; U_{12}U_{23}$$

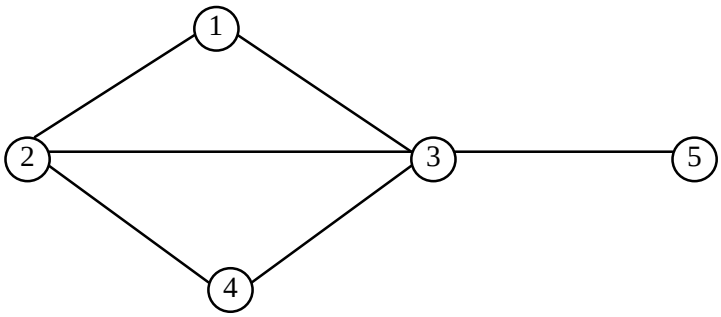
Questo passo è anche utile per evidenziare inconsistenze all'interno dell'insieme iniziale di edit definiti dallo statistico: qualora infatti vengano generati edit degeneri (in cui, cioè, il limite inferiore sia maggiore o uguale a quello superiore), ciò può derivare solo dalla contraddittorietà tra due o più edit iniziali, che dovranno quindi essere analizzati per individuare ed eliminare l'inconsistenza.

Ogni record viene confrontato con l'insieme di edit (è sufficiente l'insieme iniziale di edit). Se il record non viola alcun edit, può essere ritenuto corretto, altrimenti viene considerato errato e deve essere sottoposto a trattamento di correzione, che si compone di due passi: la localizzazione degli errori, e l'imputazione delle variabili errate.

Il passo di localizzazione degli errori consiste nel determinare il minimo insieme di variabili da modificare in modo da disattivare tutti gli edit violati, senza violarne alcun altro. Le variabili possono essere *pesate* in modo da esprimere la fiducia nella loro correttezza: in pratica, due variabili con peso 1 hanno la stessa probabilità di essere inserite nell'insieme minimo quanto una variabile con peso 2.

L'algoritmo di localizzazione degli errori si basa sulla costruzione di un grafo i cui archi rappresentano gli edit violati e i nodi collegati dagli archi le variabili dichiarate negli edit. Si consideri l'esempio di Figura 5:

Figura 5 - Grafo degli edit violati



Il grafo indica che la variabile 1 viola gli edit in cui è presente assieme alle variabili 2 e 3, la variabile 2 viola gli edit con le variabili 1 e 4, la 4 viola gli edit con le 2 e 3, e così via.

L'obiettivo dell'algoritmo è quello di *cancellare un sottoinsieme minimo di nodi in modo che non ci siano più archi.* Nell'esempio in questione è sufficiente cancellare i nodi 2 e 3, il che equivale a inserire le variabili 2 e 3 nell'insieme di variabili da imputare.

Il problema di scelta delle variabili è trattato come un classico Set Covering Problem. L'obiettivo successivo è quello di determinare, per ogni variabile da modificare, l'insieme di valori ammissibili che garantisca sia il rispetto dei vincoli precedentemente violati, sia che nessun nuovo vincolo sia violato dopo l'imputazione.

A tal fine si supponga, riordinando le variabili, che le variabili x_k con $k < n$, *non* siano da imputare, in quanto non sono state scelte dall'algoritmo precedente: ciò significa che esse sono tra di loro mutuamente consistenti e che, per ogni $i,j \leq k$, gli edit

$$L_{ij} \leq x_i / x_j \leq U_{ij}$$

sono sempre soddisfatti.

Consideriamo ora la prima variabile da imputare, cioè x_{k+1} : per tutti gli $j \leq k$ sono definiti gli edit:

$$L_{k+1,j} \leq x_{k+1} / x_j \leq U_{k+1,j}$$

Moltiplicando ogni membro per x_j otteniamo:

$$x_j L_{k+1,j} \leq x_{k+1} \leq x_j U_{k+1,j}$$

dove x_j, $L_{k+1,j}$ e $U_{k+1,j}$ sono costanti note per tutti gli $j \leq k$. Dunque ogni j=1,...,k determina un intervallo all'interno del quale il valore di x_{k+1} deve risiedere per assicurare il rispetto degli edit. L'intersezione di tutti gli intervalli così ottenuti è l'*intervallo dei valori imputabili* alla variabile x_{k+1} . Un teorema (Garfinkel, Kunnathur, Liepins, 1986) assicura che se l'insieme degli edit iniziali è consistente e se la ricerca dell'intervallo è effettuata rispetto all'insieme completo, l'intersezione è non nulla.

Una volta imputata la variabile x_{k+1} si passa a calcolare l'intervallo dei valori imputabili per x_{k+2} e così via, fino a coprire sequenzialmente l'intero insieme di variabili da correggere.

Una volta determinato, per ogni variabile da modificare, l'intervallo dei valori imputabili, si procede ad assegnare alla variabile in questione un valore interno a tale intervallo. A tal fine, viene richiamato un modulo di imputazione che non è definito una volta per tutte, ma è legato alle particolari caratteristiche dell'indagine trattata. Normalmente, tale modulo è caratterizzato da una sequenza di metodi di imputazione, applicati a cascata: la prima regola di imputazione determina un valore che, se risulta interno all'intervallo, viene assegnato; altrimenti si procede con la seconda regola e così via. I metodi correntemente utilizzati sono:

- riferimenti a valori in data set di controllo (ad esempio, ripetizioni precedenti della stessa indagine, o di altre indagini, o dati amministrativi);
- modelli di regressione;
- medie;
- valori forniti dallo statistico.

L'applicazione sequenziale delle regole fino a trovare un valore imputabile avviene nella versione *batch* di SPEER. Nella versione *online*, invece, per ogni variabile da imputare viene proposta una schermata con i valori possibili di imputazione così come calcolati dai diversi metodi, lasciando al revisore la scelta tra questi.

6.2.2.4. La metodologia del donatore implementata in RIDA: una descrizione formale.

Il sistema RIDA (Ricostruzione delle Informazioni con Donazione Automatica) è stato realizzato interamente in ISTAT, ad opera di G. Massimini e C. Runci. Esso realizza la correzione di un file di dati di qualsiasi tipo tramite la tecnica del donatore. Verranno di seguito descritti i principi su cui la tecnica si basa, nonché brevemente i passi che l'utente deve eseguire per rendere operativo il sistema.

Rappresentazione dei dati.

Sia data una matrice di dati X, formata da n unità e k variabili di tipo qualsiasi. Le unità rappresentano i vettori-riga, le variabili i vettori-colonna. Le variabili sono di tipo qualsiasi.

Dal punto di vista della archiviazione elettronica della informazione, la matrice dei dati X è contenuta in un file, costituito da un insieme di record, ognuno rappresentante una unità, e contenente un numero di campi pari al numero di variabili (da ora in poi useremo il termine record o unità come sinonimi).. Un insieme di campi (al limite anche uno solo) consente di identificare in modo univoco il record-unità ed è detto **chiave** o **identificativo** del record.

Dividiamo in due gruppi le variabili:

1) variabili affette da errore (in numero di h<k);
2) variabili esatte (in numero di k-h).

Supponiamo di sottoporre ad un processo di controllo ogni record, in modo che ognuno degli h campi corrispondenti alle variabili affette da errore contenga o un flag di errore o un valore esatto. Il file risulta diviso in due:

- insieme dei record totalmente esatti;
- insieme dei record che presentano almeno un flag di errore.

Costruzione della metrica delle distanze.

Proponiamoci ora di misurare la distanza tra due unità, rispetto alle variabili esatte. A questo scopo è necessario introdurre una metrica per ogni tipologia di variabile (si veda Abbate, 1996 a questo proposito).Sia quindi d la distanza tra due unità, misurata rispetto ad una variabile :

a) **Variabile qualitativa sconnessa.** Si pone d=0 se le unità presentano la stessa modalità, d=1 se la modalità è diversa.
Formalmente: $X_1 = X_2 \Rightarrow d=0$, $X_1 \neq X_2 \Rightarrow d=1$

b) **Variabile ordinata con m modalità.** Si pone d=0 se sulle due unità è stata rilevata la stessa modalità, d=1 se le modalità sono adiacenti, d=2 se tra di esse ce n'è una sola, e così via fino a d=m-1, se le due modalità sono agli estremi opposti. Per rendere d variabile tra 0 ed 1, essa viene divisa per il suo massimo m-1.

Formalmente: $X_1 = X_2 \Rightarrow d=0$, $X_1 = r$, $X_2 = s$ $(r \neq s) \Rightarrow d = \frac{|r-s|}{m-1}$

c) **Variabile qualitativa telescopica.** Tali variabili sono rappresentabili tramite un insieme di gruppi primari di livello 1, contenenti ognuno più sottogruppi di livello 2. Ogni sottogruppo di livello 2 contiene più sottogruppi di livello 3 e così via fino ad un sottogruppo di livello j, contenente modalità non ulteriormente scomponibili in sottogruppi, che sono al livello più basso j+1. Una modalità siffatta può essere codificata con g gruppi di bit, ognuno dei quali è dimensionato in modo da poter rappresentare tutti i sottogruppi relativi a quel livello. Poniamo d=0 se le due unità presentano stessa modalità, d=1 se le due modalità diverse sono nello stesso sottogruppo di livello j, d=2 se esse sono in gruppi differenti di livello j, ma nello stesso sottogruppo di livello j-1, d=3 se sono in gruppi differenti di livello j-1, ma nello stesso sottogruppo di livello j-2 e così via fino ad un massimo di d=j+1 se le due modalità sono in gruppi primari diversi di livello 1. Rendiamo la distanza variabile tra 0 ed 1 dividendola per il suo massimo pari a j+1.
Sia r il livello più alto a partire dal quale si riscontra una differenza tra X_1 ed X_2, r assume quindi valori tra 1 e j+1.

Formalmente: $X_1 = X_2 \Rightarrow d=0$, $X_1 \neq X_2 \Rightarrow d=\dfrac{|j+2-r|}{j+1}$

In RIDA questo tipo di distanza è utilizzato nel caso particolare che sia sufficiente una sola cifra per rappresentare ogni livello. Date quindi due generiche modalità di una variabile di tipo telescopico, esse distano 0 se tutte le cifre sono uguali, 1 se solo l'ultima è diversa, 2 se sono diverse soltanto l'ultima e la penultima e così via;

d) **Variabile quantitativa.** Sia X_1 il valore assunto dalla variabile X nella prima unità, X_2 nella seconda. Poniamo $d=|X_1 - X_2|$. La distanza può essere resa variabile tra 0 e 1 dividendola per il suo massimo, pari alla differenza tra i valori massimo (Xmax) e minimo (X_{min}) della variabile X presenti nel file.

Formalmente: $X_1 = X_2 \Rightarrow d=0$, $X_1 \neq X_2 \Rightarrow d=\dfrac{|X1 - X2|}{X\max - X\min}$

Nella versione di RIDA su CMS, il valore assoluto della differenza tra X_1 e X_2 è diviso per X_1+1, misurando uno scostamento relativo rispetto ad X_1 (la scelta di X_1+1 serve per evitare un denominatore degenere, nel caso che sia $X_1=0$). E' evidente che la scelta di una distanza siffatta privilegia l'importanza della variabile quantitativa, in particolare se il valore di X_2 risultasse molto distante da quello di X_1.

Formalizzazione della funzione di distanza mista ponderata.

Assegnata una matrice di dati, presentante k-h variabili non affette da errore, definiamo distanza mista ponderata D tra due generiche unità una espressione del tipo:

$$D = \sum_{i=1}^{r} W_i D_i ,$$

dove D_i è la distanza tra le due unità rispetto alla variabile i, misurata con una delle espressioni di cui sopra e W_i è un numero reale positivo che rappresenta l'importanza assegnata alla variabile i nel calcolo della distanza. Le r variabili sono scelte tra le k-h quelle non affette da errore. L'attuale versione di RIDA accetta solo numeri naturali per W_i.
Chiamiamo **variabili di accoppiamento o di matching** le r variabili scelte per il calcolo della distanza.

Scelta dell'unità donatrice.

Data un'unità affetta da errore nella variabile k si vuole trovare l'unità esatta posta alla distanza minima. Essa è detta **unità donatrice**, perché il valore della variabile k relativo ad essa è "donato" all'unità affetta da errore. L'insieme della unità tra le quali è scelta l'unità donatrice è detto **serbatoio dei donatori.** Il serbatoio dei donatori può essere costruito in due modi:

1) selezionando le unità esatte rispetto alla sola variabile k;
2) selezionando le unità esatte rispetto a tutte le variabili.

Nel primo caso si usa un diverso serbatoio per ogni variabile da errata, nel secondo caso si utilizza un serbatoio unico per tutte le variabili affette da errore. La prima procedura è utile quando si desidera disporre di serbatoi di donatori relativamente numerosi per ogni variabile da correggere.

Questa scelta deve essere effettuata e realizzata prima di utilizzare RIDA.

La scelta dell'unità donatrice è ulteriormente affinabile scegliendo, nell'insieme delle variabili non affette da errore e non usate come variabili di accoppiamento, delle variabili dette di **strato.** Dopo aver formato il serbatoio dei donatori in uno dei due modi di cui sopra, si seleziona l'unità donatrice tra quelle che inoltre, rispetto alle variabili di strato, presentano le stesse modalità dell'unità affetta da errore. L'uso di variabili di strato implica l'accettazione della possibilità di non avere donatori idonei per quell'unità.

Funzione di distanza mista ponderata corretta.

Possiamo introdurre un perfezionamento alla distanza mista ponderata sopra introdotta, per penalizzare l'unità del serbatoio che è già stata utilizzata nella donazione. Ridefiniamo la distanza D come:

$$D = \sum_{i=1}^{r} W_i D_i + kp,$$

dove k è il numero di volte per cui l'unità è stata precedentemente utilizzata, p è un fattore di penalità. Questa espressione più completa è adottata da RIDA, che richiede che p sia un numero intero.

Ponderazione delle variabili di matching.

Sono molte le tecniche possibili di ponderazione delle variabili di matching. Le applicazioni finora realizzate nell'interno dell'istituto hanno utilizzato il criterio del χ^2 (si veda [1]). Esso si applica nel seguente modo:

1) si misura la connessione tra la variabile affetta da errore e quelle esatte tramite l'indice χ^2. Il valore dell'indice dipende dal numero di celle della tabella di contingenza. Poiché bisogna confrontare il valore dei χ^2 ottenuti, per renderli confrontabili occorre o riclassificare in modo opportuno almeno la variabile da correggere, se di tipo quantitativo, in modo da ottenere tabelle di contingenza di dimensioni omogenee, oppure dividere

direttamente il valore del χ^2 per il numero di gradi di libertà, che è pari al prodotto tra il numero delle righe e delle colonne della tabella di contingenza diminuiti entrambi di uno;

2) l'utilizzatore del metodo deve esaminare criticamente i valori di χ^2 così ottenuti, eventualmente divisi per il numero dei gradi di libertà: le variabili non affette da errore che presentano il valore più alto sono le migliori candidate ad essere variabili di strato, quelle con valore immediatamente inferiore possono diventare variabili di matching. L'utilizzatore del metodo deve usare i valori come supporto a una decisione che tiene anche conto della sua conoscenza dell'indagine.

La scelta delle variabili di strato deve tener conto anche del fatto che all'aumentare del loro numero, aumenterà la selettività nell'ambito del serbatoio dei donatori, ma aumenterà anche la probabilità di non trovare il donatore. Nelle applicazioni finora realizzate è stata sempre impiegata una sola variabile di strato.

Modalità di utilizzo del sistema RIDA.

L'utente del prodotto deve preparare 4 file.

A) File contenente i record errati. In esso le variabili affette da errore debbono contenere un carattere di errore ripetuto per tutta la lunghezza del campo.

B) File contenente i record esatti, costituenti il serbatoio dei potenziali donatori.

C) File dei parametri. Esso contiene le variabili di strato e di matching. Per ogni variabile occorre specificare la posizione iniziale, la lunghezza, il tipo (obbligatorio per le variabili di matching, al fine della scelta della funzione di distanza da adottare) e il peso. Le variabili quantitative possono essere riclassificate, specificando l'estremo superiore di ogni classe. Si possono poi inserire i parametri U,R,L. Essi sono, rispettivamente, il numero massimo di volte che la stessa unità può essere utilizzata come donatrice, il fattore moltiplicativo che penalizza l'uso ripetuto dello stesso donatore e la massima distanza a cui può essere considerato un donatore. Vale la stessa avvertenza formulata a proposito dell'uso degli strati: l'uso dei parametri U e L implica la possibilità di non riuscire a trovare il donatore. I parametri U, R, L possono anche non essere utilizzati: questo implica che non si pone alcun limite alla possibilità di riutilizzare lo stesso donatore ed esso può essere scelto anche molto distante rispetto all'unità donatrice. L'ultimo parametro da inserire è il carattere di errore: esso è particolarmente importante perché solo i campi in cui esso è presente sono soggetti a correzione.

D) File eseguibile. Esso contiene il nome del file dei parametri, nonché dei file degli esatti e degli errati. Inoltre l'utente deve specificare il nome di due file che saranno creati nel corso dell'esecuzione del programma, uno contenente i record che sono stati corretti, l'altro quelli che la procedura non è riuscita a correggere (file degli incorretti). Le ultime due righe contengono rispettivamente il nome del disco di sistema su cui risiedono i file e saranno generati gli output della procedura e due variabili di tipo Y/N, per indicare se si desiderano o meno delle statistiche di correzione Il file deve avere il suffisso DON01. L'istruzione EXEC DON01, seguita dal nome del file privo di suffisso lancia la procedura.

Dopo la corretta esecuzione di RIDA, l'utente dovrà provvedere a fondere in un file unico il file dei record esatti, quello dei corretti e quello eventuale degli incorretti.

Se è stato prodotto il file dei record incorretti, nel file unico creato dall'utente i record non corretti conterranno ancora il carattere di errore: essi debbono essere corretti con una tecnica alternativa.

7. Disegno ed implementazione

La fase di disegno ed implementazione delle procedure di controllo e correzione è fortemente influenzata dagli obiettivi dell'elaborazione.

In particolare, se il fine principale è quello di disporre di stime di aggregati che siano nel contempo affidabili e tempestive, senza che sia necessario disporre di dati che siano il più possibile liberi da errori, allora si ricorrerà alle tecniche proprie del macroediting e dell'editing selettivo. Se, al contrario, non solo gli aggregati, ma anche i dati elementari devono essere il più possibile esenti dagli errori, allora la procedura di controllo e correzione deve essere di tipo esaustivo, dovendo coinvolgere la totalità dei dati, e verranno applicate tecniche automatiche di correzione, oppure verranno implementate procedure di tipo misto (con sottofasi automatiche e sottofasi interattive).

Nei due casi, diversi saranno i criteri cui improntare sia il disegno, che lo sviluppo e la messa a punto delle procedure.

7.1. Disegno ed implementazione delle procedure interattive

In generale, le strategie di controllo e correzione di tipo interattivo sono applicabili a tutte le principali tipologie di indagine (campionarie, censuarie o esaustive), siano esse di tipo statistico o amministrativo. L'intervento interattivo in una o più fasi del processo di produzione di dati statistici o di altro tipo è spesso necessario per garantire una migliore qualità delle informazioni finali prodotte: la collocazione e l'entità di tale intervento dipendono, nella pratica, dai fattori illustrati nei paragrafi relativi alla scelta della configurazione delle procedure di controllo e correzione.

In generale, nell'ambito delle possibili strategie di tipo interattivo, la scelta dell'una o dell'altra procedura deve essere basata su considerazioni strettamente legate alle caratteristiche dell'indagine da sottoporre a controllo ed all'organizzazione dell'indagine stessa, alle risorse ed alle tecnologie disponibili. Poiché il trattamento dei dati avviene un record alla volta, importanza primaria hanno ad esempio la dimensione dell'indagine, il livello di correzione dei dati richiesto, la disponibilità di personale, i tempi ed i costi.

Nella fase di disegno ed implementazione di una procedura interattiva di controllo e correzione è necessario innanzi tutto distinguere fra i seguenti due casi:

1. procedure interattive integrate nelle fasi di raccolta o di registrazione dei dati;
2. procedure interattive integrate nella fase di editing.

Abbiamo già illustrato quali sono gli approcci, i metodi ed i prodotti utilizzabili in ciascuna delle due situazioni. Di seguito sono elencati quelli che possono essere considerati i principali fattori discriminanti fra le possibili scelte, raggruppati per tipologia.

a. Vincoli dovuti alle caratteristiche dell'indagine:

1. tipo di variabili rilevate(qualitative o quantitative);
2. dimensione dell'indagine (numero unità rilevate);
3. livello di correzione dei dati richiesto (aggregato, elementare);
4. periodicità dell'indagine;
5. presenza ed omogeneità di domini disgiunti (definiti sulla base di prefissate classificazioni);

6. esistenza, tipologia e misurabilità di relazioni statistico/matematiche fra variabili rilevate;
7. disponibilità di informazioni storiche;
8. possibilità di utilizzo di informazioni ausiliarie;
9. tasso di non risposta parziale e totale dell'indagine;
10. tipologia dei controlli richiesti (correzione delle incompatibilità, trattamento degli outlier e/o delle mancate risposte, ecc.).

b. Vincoli dovuti alle risorse disponibili:
1. disponibilità di risorse umane e budget;
2. tempi da rispettare;
3. costi (in particolare per l'implementazione della procedura di controllo, per la conduzione delle attività di re-intervista, ecc.);
4. disponibilità di analisti e di programmatori specializzati.

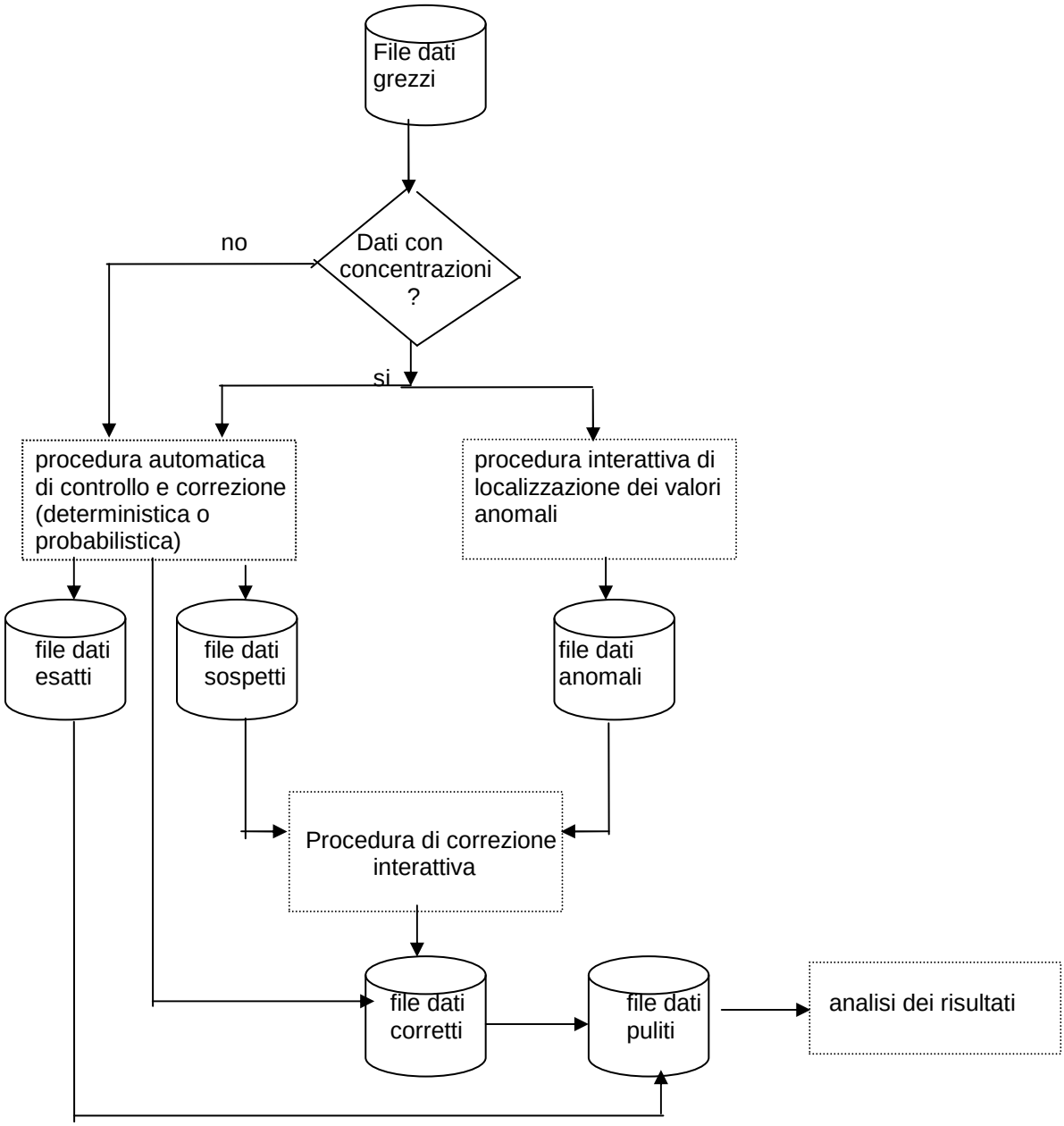
c. Vincoli dovuti alla tecnologia:
1. piattaforma hardware (mainframe, PC, rete) disponibile;
2. sistema operativo esistente;
3. disponibilità di software generalizzato o necessità di sviluppo di programmi ad hoc;
4. disponibilità di software grafico di qualità.

d. Vincoli dovuti all'organizzazione :
1. tecnica di intervista o di raccolta delle informazioni;
2. qualità della formazione degli intervistatori;
3. modalità di registrazione dei dati;
4. possibilità di re-intervista;
5. disponibilità dei modelli cartacei.

Il disegno di una procedura di editing prevede più alternative possibili:
1. sviluppo di una procedura interamente interattiva di tipo micro;
2. sviluppo di una procedura interattiva di tipo macro o selettivo;
3. sviluppo di una procedura interattiva micro di tipo misto (in parte interattiva, in parte automatica);
4. sviluppo di una procedura interattiva macro di tipo misto (in parte interattiva, in parte automatica).

Le alternative 1 e 3 sono state in realtà ampiamente discusse nell'ambito delle possibili configurazioni vincolanti l'adozione di una procedura di controllo e correzione dati. Nello schema della Figura 1 è descritto il disegno di una tipica procedura di controllo e correzione micro di tipo misto correntemente utilizzata in ISTAT.

Figura 1 - Schema di procedura di editing micro di tipo misto



La tecnica di localizzazione degli outlier può essere affidata a procedure generalizzate del tipo di SPEER o ARIES o può richiedere lo sviluppo di software ad hoc. In questo caso, è possibile implementare uno dei metodi di tipo macro o di tipo selettivo.

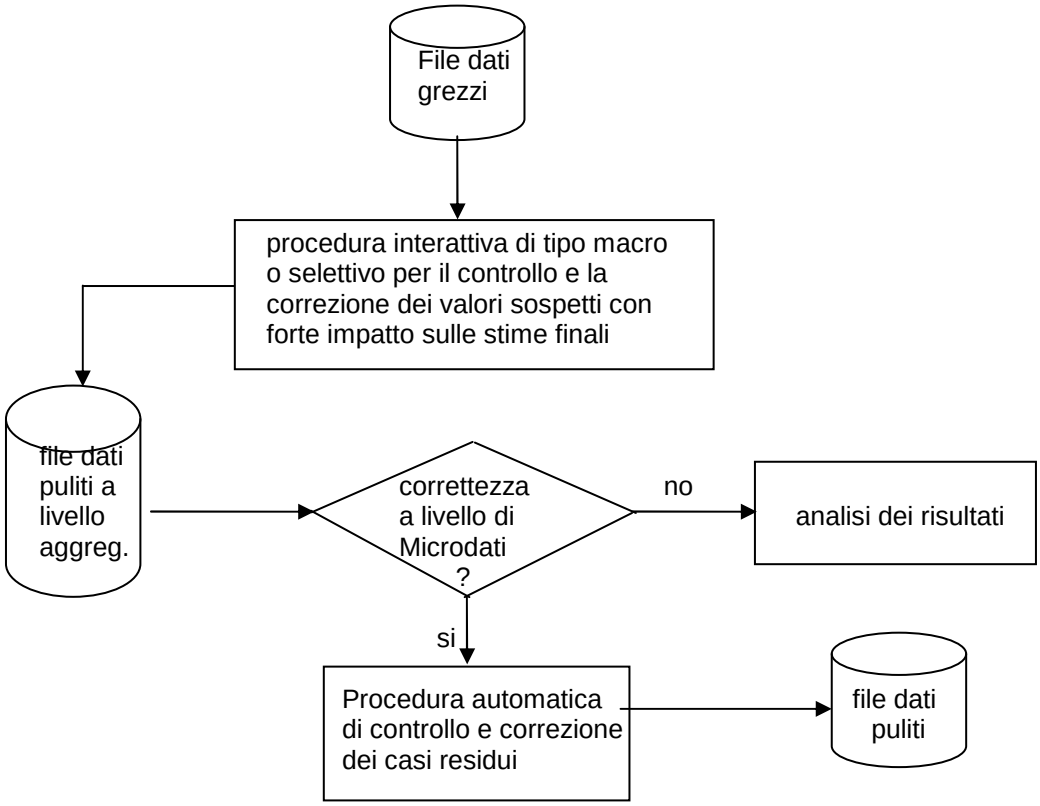
Le alternative 2 e 4 sono possibili solo nel caso in cui i dati presentino concentrazioni significative, in particolare laddove abbia senso parlare di osservazioni anomale con impatto significativo sulle stime finali prodotte dall'indagine (tipicamente, indagini economiche o amministrative che rilevano ammontari o frequenze).

La scelta dell'una o dell'altra strategia di editing dipende sostanzialmente dalla necessità o meno di produrre dati corretti a livello elementare: nel caso sia sufficiente per gli scopi

dell'indagine avere stime corrette a livello di aggregati, è infatti possibile e consigliabile l'adozione delle sole tecniche di tipo macro o di tipo selettivo. Tali tecniche, lo ricordiamo, uniscono un elevato livello di efficacia, in termini di qualità dei risultati prodotti, ad un altrettanto elevato grado di efficienza, in termini di risparmio operativo prodotto (tempi e numero di verifiche interattive necessarie).

Il tipico disegno di una procedura interattiva macro di tipo misto è schematizzata nella Figura 2.

Figura 2 - Schema di procedura di editing macro di tipo misto



La procedura interattiva dei valori sospetti o dei valori anomali (di tipo micro, macro univariato o multivariato, grafico, selettivo) può essere caratterizzata o meno dai seguenti fattori principali:

- utilizzo di informazioni storiche;
- utilizzo di informazioni da fonti esterne;-
- utilizzo di variabili di classificazione per la defiinizione di domini omogenei;

Quanto maggiori sono le informazioni di cui ci si avvale, tanto migliori sono le prestazioni della procedura stessa.

La procedura di correzione interattiva presente in entrambi gli schemi delle Figure 1 e 2 può consistere:

- nella re-intervista dei rispondenti cui corrispondono i valori sospetti;
- nel controllo manuale dei questionari;
- nelle valutazioni soggettive di un esperto caso per caso.

Le tre operazioni precedenti, che ovviamente non si escludono fra loro, non sono però sempre praticabili, sia a causa dell'organizzazione del processo produttivo (che ad esempio non prevede il mantenimento dei modelli cartacei, oppure presenta l'impossibilità di effettuare reinterviste) sia a causa dei tempi e dei costi necessari alla loro effettuazione.

Negli schemi delle figure 1 e 2 abbiamo lasciato tratteggiato il blocco relativo alla fase di analisi dei risultati e delle prestazioni della procedura di editing. Questa fase riveste un'importanza cruciale nell'ambito del processo di progettazione, disegno e messa a punto di una procedura di controllo, poiché su di essa si basano le fasi di:

1. validazione e documentazione della procedura di editing;
2. revisione della procedura di editing;
3. revisione dell'organizzazione dell'indagine;
4. revisione del questionario;

Per quanto riguarda il processo di revisione della procedura interattiva di editing, questo consiste essenzialmente nella verifica delle sue prestazioni in termini di:

- rispondenza alle caratteristiche dell'indagine e dei fenomeni rilevati;
- qualità delle stime finali prodotte;
- tempi di implementazione e di effettuazione;
- costi di implementazione e di effettuazione.

I primi due aspetti possono essere investigati a partire dall'analisi dei risultati generati dalla procedura. Questa analisi permetterà, attraverso la verifica della presenza nei dati di errori sistematici, di individuare eventuali inconsistenze o difetti nella procedura, e di mettere in atto le misure correttive necessarie alla rimozione di tali errori.

Gli elementi fondamentali per la conduzione dell'analisi dei risultati sono rappresentati dalle seguenti informazioni:

- frequenza di attivazione degli edit;
- numero di correzioni effettuate per variabile e per edit;
- matrice o grafico di transizione per ogni variabile;
- entità e tipologia dei valori anomali localizzati;
- impatto delle correzioni apportate sulle stime finali.

Dall'analisi di questi dati è possibile verificare la natura stocastica o sistematica degli errori rilevati e corretti: laddove si riscontri una sistematicità di tali errori, è necessario stabilire la loro origine più probabile, intervenendo di conseguenza.

Nel caso l'origine dell'errore sistematico non sia attribuibile alla procedura di controllo e correzione interattiva, ma piuttosto a difetti o imperfezioni nell'organizzazione dell'indagine o nella struttura interna del questionario, vanno attivati (laddove possibile) i processi di revisione di cui ai punti 3 e 4 precedenti per la localizzazione e la modifica dei fattori di disturbo.

Laddove invece si verifichino carenze nella struttura o nelle caratteristiche della procedura stessa, vanno analizzate:

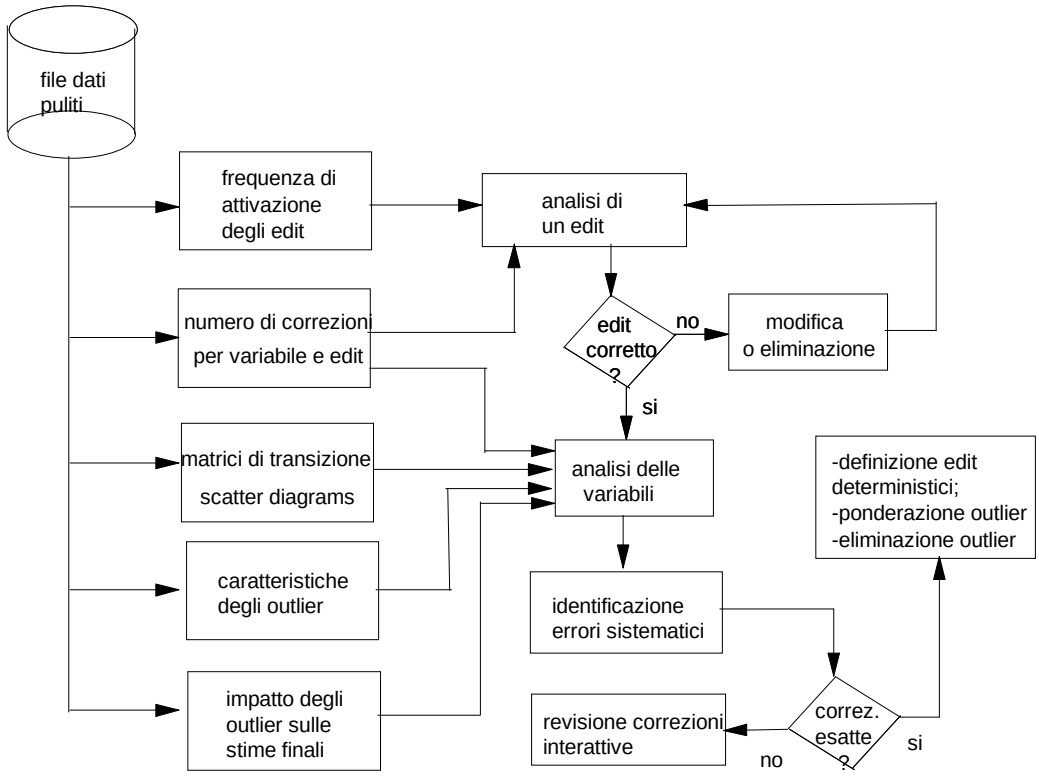
1. la tipologia degli edit utilizzati (edit logici, regole statistico-matematiche come rapporti o differenze, altre funzioni fra variabili) e la loro capacità di cogliere l'errore;
2. la completezza e l'esaustività dell'insieme di edit utilizzato;
3. il corretto uso delle informazioni storiche o ausiliarie eventualmente utilizzate;
4. l'idoneità del metodo di localizzazione degli outlier utilizzato rispetto alle caratteristiche dell'indagine e dei dati;

5. la correttezza delle operazioni di correzione interattiva effettuate (fenomeno dell'over-editing, rischio di editing creativo).

Alla luce delle precedenti considerazioni, lo schema generale della fase di analisi dei risultati della procedura di editing interattiva è rappresentato in Figura 3

In tale schema si è ipotizzata una procedura mista, cioè in parte automatica, in parte interattiva (di tipo micro, macro oppure misto), ma il meccanismo di verifica delle prestazioni della procedura stessa non cambia sostanzialmente nei casi di procedure interamente interattive. Nel caso di procedure esclusivamente interattive di tipo macro o selettivo, gli edit sono soltanto quelli utilizzati per la localizzazione delle unità sospette e degli outlier (rapporti, differenze, altre funzioni statistico-matematiche).

Figura 3 - Analisi dei risultati prodotti dalla procedura di controllo e correzione interattiva



Dall'analisi delle frequenze di attivazione degli edit, del numero di correzioni apportate e delle caratteristiche delle osservazioni anomale è possibile verificare la correttezza o meno degli edit utilizzati in termini di:

1. idoneità a cogliere l'errore;
2. corretta formulazione;

L'aspetto 1 è relativo all'eventualità che, in fase di localizzazione degli outlier, siano state utilizzate regole poco adatte a cogliere l'errore: può quindi essere necessario utilizzare

rapporti invece che differenze, fare un diverso uso delle eventuali informazioni storiche o ausiliarie a disposizione, definire nuovi domini ecc.

La corretta formulazione di un edit consiste nella sua effettiva corrispondenza o meno a situazioni di incompatibilità fra variabili.

Se la frequenza di attivazione di un edit è troppo alta, vanno quindi innanzi tutto verificate, per quell'edit, le condizioni 1 e 2 precedenti.

Al contrario, se la frequenza di attivazione di un edit è troppo bassa oppure l'impatto delle correzioni derivanti dalla sua attivazione è trascurabile, è opportuno prendere in considerazione l'eventualità di eliminare tale edit, onde ridurre il fenomeno dell'over-editing.

Se la causa di frequenze di attivazione anomale degli edit non è attribuibile agli edit stessi, è necessario procedere alla verifica della presenza nei dati di errori sistematici.

Nel caso di variabili qualitative, tale analisi può essere condotta sulla base delle matrici di transizione delle variabili, mentre nel caso delle quantitative si può ricorrere a scatter diagrams.

Gli errori sistematici, in ogni caso, possono essere causati anche da una errata interpretazione delle risposte a certi quesiti da parte degli esperti addetti alla registrazione o alla revisione interattiva (*editing creativo*): è quindi opportuno verificare la correttezza delle modifiche apportate ai dati (alla luce della tecnica adottata per effettuare tale operazione).

Laddove si sia proceduto correttamente, è necessario eliminare l'errore sistematico ricorrendo all'uso di regole deterministiche nel caso di errori di incompatibilità, a eliminazione o ponderazione dei dati nel caso di valori anomali.

Qui di seguito sono elencate alcune proprietà generali di cui una procedura di controllo e correzione interattiva dovrebbe godere, nell'ottica della funzionalità, della riproducibilità e della generalizzabilità:

- *documentabilità del processo di correzione*: al termine del processo di controllo e correzione devono essere note le informazioni relative a:
  - quali e quante variabili sono state corrette e per quante volte;
  - quali edit sono stati attivati e quante volte;
  - quanti outlier sono stati individuati ed il tipo di trattamento ad essi riservato (eliminazione, imputazione, ponderazione);
  - approccio utilizzato per la correzione interattiva (re-intervista, controllo modelli, valutazione di esperti.

- *trasparenza della procedura e sua semplicità di utilizzo*: le caratteristiche della procedura devono essere chiare in termini di:
  - piano di incompatibilità utilizzato (deterministico, probabilistico, generalizzato o ad hoc) e il tipo di regole utilizzate;
  - approccio adottato per la localizzazione degli outlier (micro, macro, selettivo, generalizzato oppure sviluppato ad hoc) ed ipotesi eventualmente fissate;

- *possibilità di apportare miglioramenti, aggiornamenti*: la procedura deve essere tale che qualunque modifica nei dati, nella struttura del questionario o nell'organizzazione dell'indagine sia facilmente trasferibile all'interno della procedura stessa (*flessibilità*). Questo risultato è generalmente conseguibile con maggior semplicità nei casi in cui la procedura abbia una struttura *modulare*, cioè quando le varie fasi del processo di controllo

e correzione sono implementate in porzioni indipendenti di software, modificabili senza compromettere le altre funzionalità della procedura di controllo e correzione stessa.

- *possibilità di generalizzare il software e la procedura*: soprattutto nel caso di software sviluppato ad hoc, è consigliabile che in fase di disegno e implementazione dell'algoritmo di controllo e correzione non siano fissati, laddove possibile, parametri o vincoli troppo specifici dell'indagine di interesse. Date le caratteristiche dell'indagine, sarebbe opportuno prevedere una struttura dell'algoritmo il più generale possibile, all'interno della quale si possano specificare di volta in volta i parametri relativi all'indagine di interesse. Questo consentirebbe l'utilizzo della procedura per il trattamento di indagini con caratteristiche generali simili fra loro, con un conseguente risparmio (nel lungo periodo) in termini di tempo e di costi.

7.2 Disegno ed implementazione delle procedure automatiche

La definizione, lo sviluppo e la messa a punto hanno come scopo finale la creazione di una procedura automatica per l'editing e la correzione dei dati che :

- localizzi ed elimini il maggior numero di errori possibile;
- non introduca distorsioni nei dati.

Tra i due approcci descritti in precedenza, è quello probabilistico l'unico in grado di assicurare questo tipo di risultato, almeno in una situazione di tipo "ideale", tale cioè che la tipologia degli errori presenti nei dati sia di carattere stocastico, o quantomeno che la componente sistematica negli errori sia trascurabile. Se ciò non avviene, se cioè gli errori sistematici sono presenti in quantità tale da non poter essere considerati trascurabili, deve essere introdotta una specifica componente deterministica nella procedura, dato che è dimostrato che l'approccio probabilistico non è adatto al trattamento di tali errori, ma anzi è suscettibile di introdurre ulteriori distorsioni nei dati.

Nel breve periodo, quindi, in fase di disegno della procedura complessiva occorre:

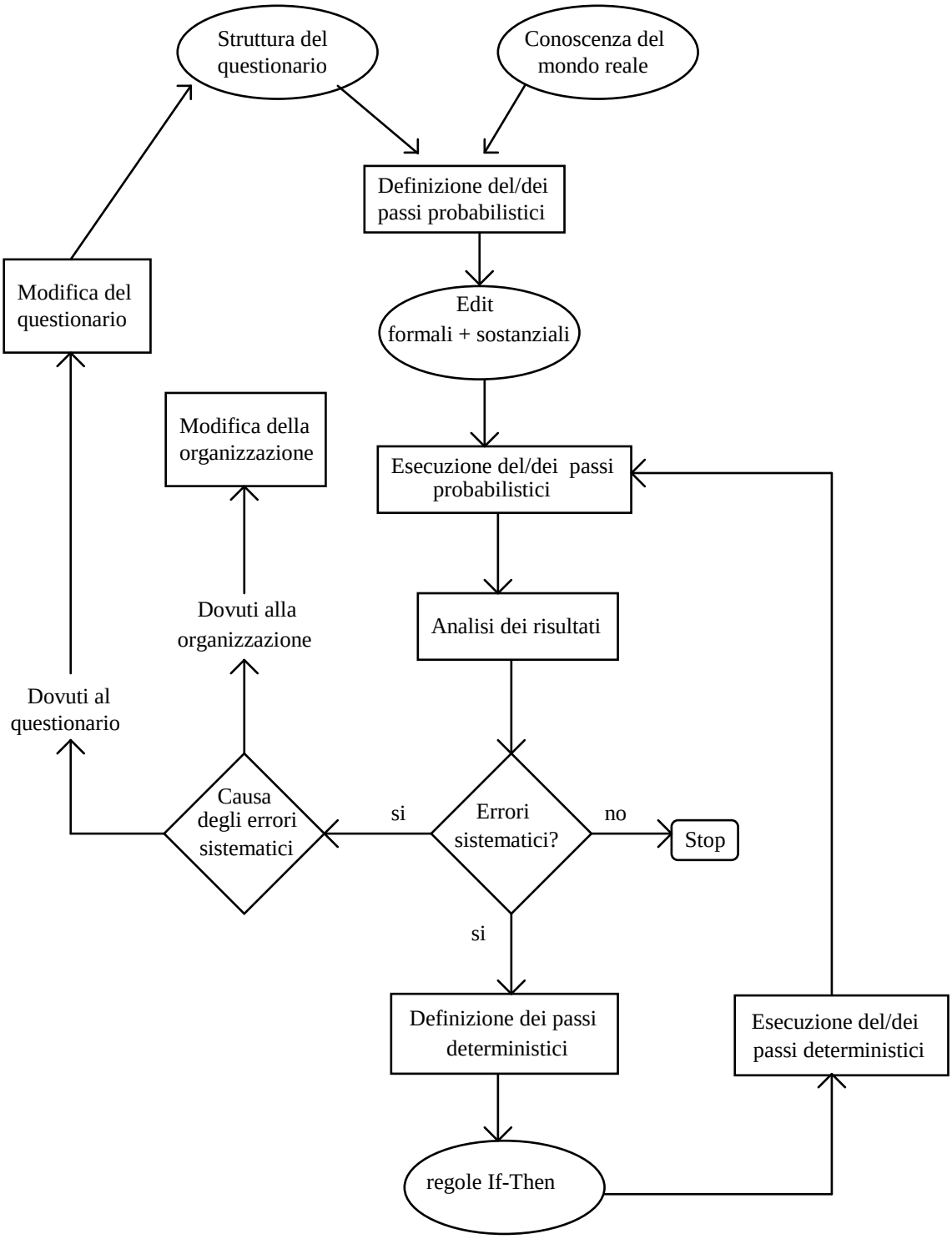
a) prevedere comunque la massimizzazione del ricorso all'approccio probabilistico, disegnando un piano di compatibilità che ricalchi i principi della metodologia Fellegi-Holt;
b) individuare le componenti sistematiche dell'errore e prevedere, come eccezione, l'applicazione di procedure deterministiche per la loro rimozione.

Nel lungo periodo, invece, qualora si possa intervenire sul processo di raccolta e registrazione dei dati, e si abbia quindi la possibilità di rimuovere le cause che producono gli errori sistematici, occorre intervenire in tal senso, al fine di minimizzare e, al limite, eliminare il ricorso a passi di tipo deterministico (che sono comunque suscettibili di introdurre distorsioni addizionali nei dati).

Tutto ciò implica che la fase di messa a punto delle procedure non è finalizzata solo ad una ottimizzazione della procedura probabilistica ideata nella fase di disegno (verifica della completezza e correttezza del piano di compatibilità), ma anche all'individuazione della componente sistematica degli errori (per lo sviluppo di passi deterministici), ed alla identificazione delle cause di tali errori (per la loro rimozione dal processo produttivo).

Lo schema della metodologia proposta è contenuto nella seguente Figura 4.

Figura 4 - La metodologia per la messa a punto della procedura di edit ed imputazione



Il punto di partenza è dato dal modello di rilevazione dei dati. Esso contiene, generalmente, delle regole di compilazione: ognuna di esse dà luogo a uno o più edit in forma normale, che sono generalmente del tipo:

(variabile-filtro ≠ valori discriminanti) ∩ (variabile-dipendente = valori significativi) [14]
oppure:
(variabile-filtro = valori discriminanti) ∩ (variabile-dipendente ≠ valori significativi) [15]

Ad esempio, supponiamo che in un modello la risposta alle variabili "condizione professionale" (CONPRO) e "posizione nella professione" (POSPRO) sia condizionata dalla risposta alla variabile-filtro ETA': devono cioè rispondere solo gli intervistati con almeno 15 anni. Sarà quindi necessario definire i seguenti edit in forma normale:

(ETA' < 15) ∩ (CONPRO ≠ *blank*)
(ETA' < 15) ∩ (POSPRO ≠ *blank*)

che corrispondono agli [14]. Mediante tali edit si stabilisce che è una condizione di errore la presenza di valori significativi dei domini di CONPRO e POSPRO in corrispondenza a rispondenti con età inferiore al minimo previsto. Se per gli intervistati con età superiore a tale minimo, la risposta è obbligatoria, occorre introdurre anche gli edit corrispondenti a [15]:

(ETA' > 14) ∩ (CONPRO = *blank*)
(ETA' > 14) ∩ (POSPRO = *blank*)

che stabiliscono che è un errore la non-risposta in corrispondenza di individui con più di 14 anni. Questi edit sono anche detti *formali*, in quanto dipendono solo dalle regole formali di compilazione del questionario. I sistemi generalizzati come SCIA sono utili anche per verificare la consistenza interna formale di un modello di rilevazione: se infatti ci limitiamo a definire tutti e solo gli edit formali, sottoponendoli poi a generazione dell'insieme completo, qualora venga prodotto un edit cosiddetto *degenere*[19] possiamo essere assolutamente certi che le regole di compilazione sono tra di loro contraddittorie. In tal caso, dobbiamo sottoporre il modello di rilevazione a revisione onde individuare e rimuovere tale contraddittorietà.

Assieme agli edit formali devono essere definiti quelli *sostanziali*, quelli cioè che dipendono dalla conoscenza del mondo reale sottoposto ad indagine. Ad esempio, il fatto che una donna non possa svolgere la professione di carabiniere, o che un dipendente di un'impresa non possa eccedere una certa retribuzione, nè possa lavorare gratis, dà luogo ad altrettanti edit sostanziali.

L'unione degli edit formali e di quelli sostanziali forma l'*insieme iniziale* degli edit. Tale insieme deve rispondere ad alcuni importanti criteri:

- in primo luogo deve essere il più *ricco* possibile, deve cioè contenere tutta la conoscenza che possa permetterci di individuare il maggior numero possibile di errori: è fondamentale, a questo riguardo, la disponibilità di informazioni complete sulle relazioni ed i vincoli che sussistono nella porzione di mondo oggetto di indagine;
- in secondo luogo deve essere *corretto*, non deve cioè contenere al proprio interno asserzioni che si contraddicono: per tale motivo occorre analizzare gli edit per individuare e risolvere le eventuali inconsistenze.

Una volta garantita la consistenza, si procede a generare l'*insieme completo*, quello necessario ad una corretta localizzazione e correzione degli errori. Questi compiti sono generalmente svolti in modo automatico o assistito dai sistemi generalizzati, sia quelli che

[19] Vedi Appendice sulla metodologia Fellegi-Holt

trattano variabili qualitative (SCIA) che quelli relativi alle variabili quantitative (SPEER, GEIS[20]).

La generazione dell'insieme completo di edit non sempre è possibile: l'algoritmo che espleta questa funzione è di complessità esponenziale rispetto al numero iniziale di edit espliciti, il che significa che quando tale numero eccede una certa soglia (nei casi pratici, 150-200 edit), i tempi di esecuzione e/o la memoria necessaria diventano inaccettabili. La soluzione che generalmente si adotta in questi casi è quella di suddividere l'insieme iniziale di edit in due o più sottoinsiemi, ognuno dei quali sia "trattabile" (sia cioè possibile, per esso, generare il corrispondente insieme completo). In tal caso, non si ha più l'esecuzione di unico passo probabilistico, con l'applicazione contemporanea di tutte le regole di compatibilità, bensì un'applicazione *sequenziale* di più passi probabilistici: in ognuno di essi, non devono più essere imputate le variabili che lo siano già state in passi precedenti. Si rinuncia, in tal modo, alla ottimalità delle operazioni di localizzazione degli errori, ma si garantisce in ogni caso la correttezza finale dei dati.

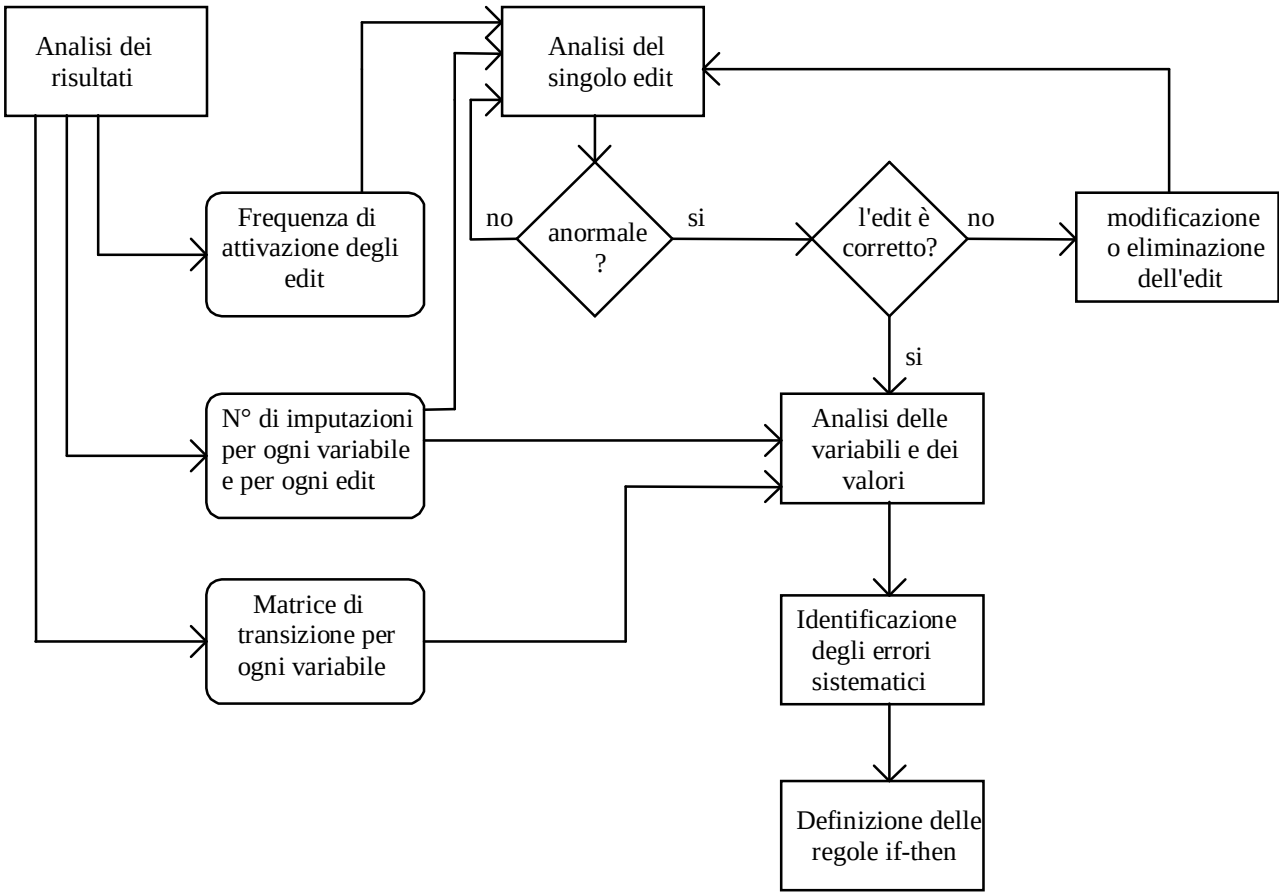
Una volta ottenuto l'insieme completo di edit (o la sequenza di insiemi completi, nel caso di suddivisione dell'insieme iniziale), si procede applicandolo ai dati grezzi a disposizione. L'analisi delle statistiche relative agli edit attivati ed alle variabili imputate è fondamentale per:

1. verificare la correttezza delle regole di incompatibilità dichiarate;
2. individuare componenti sistematiche degli errori.

Vediamo più in dettaglio, come questo può essere possibile.

[20] Nel caso di GEIS, la generazione degli edit impliciti non costituisce un passo importante ai fini delle operazioni di localizzazione e correzione degli errori, mentre vengono comunque eseguite le operazioni di individuazione e rimozione di inconsistenze e ridondanze tra gli edit

Figura 5 - Analisi dei risultati dell'imputazione



Come evidenziato nello schema della Figura 5, l'analisi dei risultati viene condotta principalmente considerando:

- le frequenze di attivazione dei singoli edit;
- le frequenze delle imputazioni effettuate su ogni variabile, a seconda dell'edit che ha causato l'imputazione;
- le matrici o i grafici[21] di transizione, che evidenziano la quantità e la qualità delle trasformazioni subite da ogni variabile in seguito all'imputazione.

La presenza di frequenze di attivazione anormali nei dati possono essere dovute:

- alla non corretta formulazione dell'edit responsabile di tale attivazione (ad esempio, l'edit può non rappresentare un'effettiva incoerenza fra variabili, ma una condizione possibile nella realtà), da cui la necessità di modificare o rimuovere l'edit stesso;
- l'errore evidenziato dall'attivazione dell'edit è di tipo sistematico e non stocastico, per cui la sua eliminazione richiede il ricorso a regole di tipo deterministico.

[21] Le **matrici di transizione** sono utilizzate per le variabile qualitative, e riportano, in riga, le modalità della variabile *prima* della correzione e, in colonna, le modalità della stessa variabile *dopo* la correzione: le frequenze indicate in una casella (i,j) indicano il numero di casi (record) in cui la variabile è stata corretta transitando dalla modalità i-esima a quella j-esima. I **grafici di transizione** sono invece utilizzati per le variabili quantitative (o per le qualitative con elevato numero di modalità), e riportano in ascissa i valori della variabile *prima* della correzione, ed in ordinata i valori della stessa variabile *dopo* la correzione.

Se, attraverso l'analisi delle informazioni di cui ai punti precedenti, viene accertata la non correttezza dell'edit in esame, si deve procedere alla sua modifica o, se del caso, alla sua eliminazione.

In caso contrario, se l'anomalia nella frequenza di attivazione dell'edit non risulta attribuibile ad una errata definizione dell'edit, è necessario procedere all'analisi delle matrici o dei grafici di transizione delle variabili della cui imputazione l'edit è responsabile. La presenza di una concentrazione di valori fuori della diagonale principale delle matrici consente, attraverso l'analisi delle coppie di modalità che corrispondono a tali frequenze, di identificare sia la presenza, sia la causa dell'errore sistematico e, quindi di definire l'opportuna regola deterministica per la sua eliminazione. Analogamente, nel caso dei grafici, l'addensamento di punti in zone esterne alla bisettrice è indizio della presenza di elementi sistematici di turbamento.

L'insieme delle regole deterministiche prodotte al termine di questa analisi dovrà essere applicato ai dati *prima* dell'applicazione ad essi dei passi probabilistici.

Una volta ripetuto il processo di imputazione, l'analisi dei risultati verrà iterata in modo analogo per verificare eventuali ulteriori anomalie nei dati.

Pertanto, come già accennato all'inizio del capitolo, nel breve periodo la procedura risulta essere di tipo *misto*, cioè composta da passi sia probabilistici che deterministici.

Nel lungo periodo, invece, la parte deterministica del processo è destinata ad essere soppressa, attraverso un accurato processo di eliminazione delle cause strutturali che determinano gli errori sistematici.

Infatti, l'analisi di tali cause può rivelare imperfezioni sia nella struttura del questionario, sia nell'organizzazione della rilevazione, sia nella memorizzazione dei dati. Una volta eliminate queste imperfezioni, l'errore sistematico dovrebbe diminuire: la verifica di tale riduzione può essere condotta sulla base delle frequenze di attivazione delle regole deterministiche.

Nel caso di utilizzo di dati amministrativi è probabile che vi siano rigidità tali nel processo di raccolta e registrazione dei dati da non consentire interventi migliorativi sostanziali, tesi ad eliminare le carenze che generano gli errori sistematici: in tal caso, l'obiettivo di rendere totalmente probabilistica la procedura è difficilmente raggiungibile, ed anche a regime la componente deterministica sarà presente al suo interno.

8. Validazione delle procedure di controllo e correzione

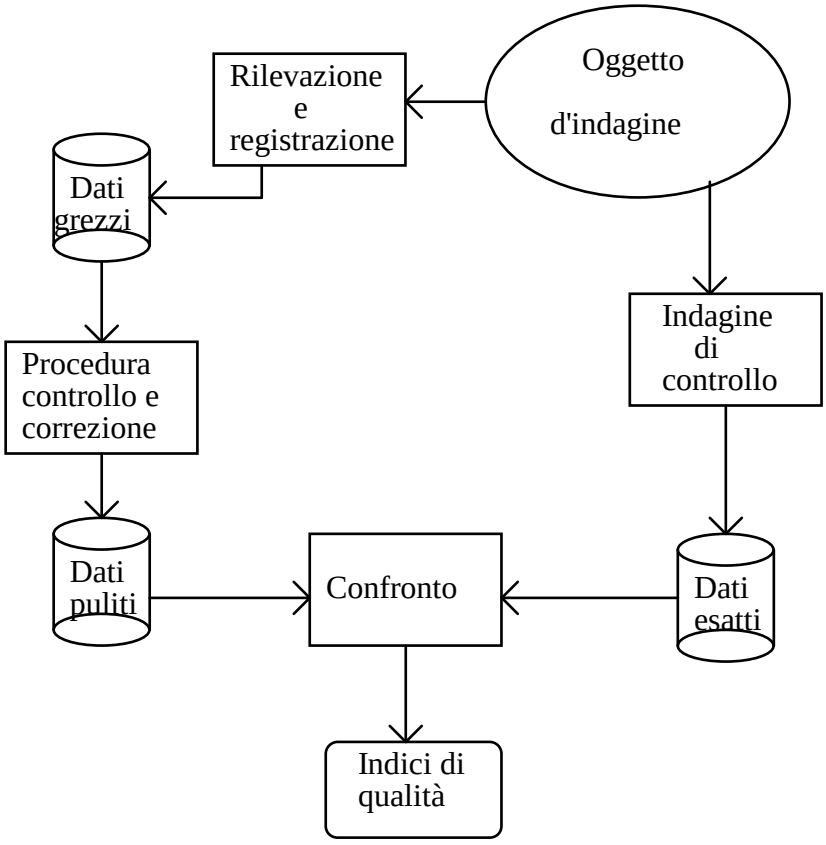
Una procedura di controllo dei dati e di correzione degli errori, sia di tipo interattivo, che automatico, che mista, deve essere sottoposta a *validazione*, ne deve cioè essere verificata la capacità di:

1. individuare effettivamente gli errori presenti nei dati;
2. correggere gli errori individuati, ripristando i valori veri al posto di quelli errati.

In una situazione ottimale, una procedura di validazione dovrebbe poter disporre dei seguenti file:

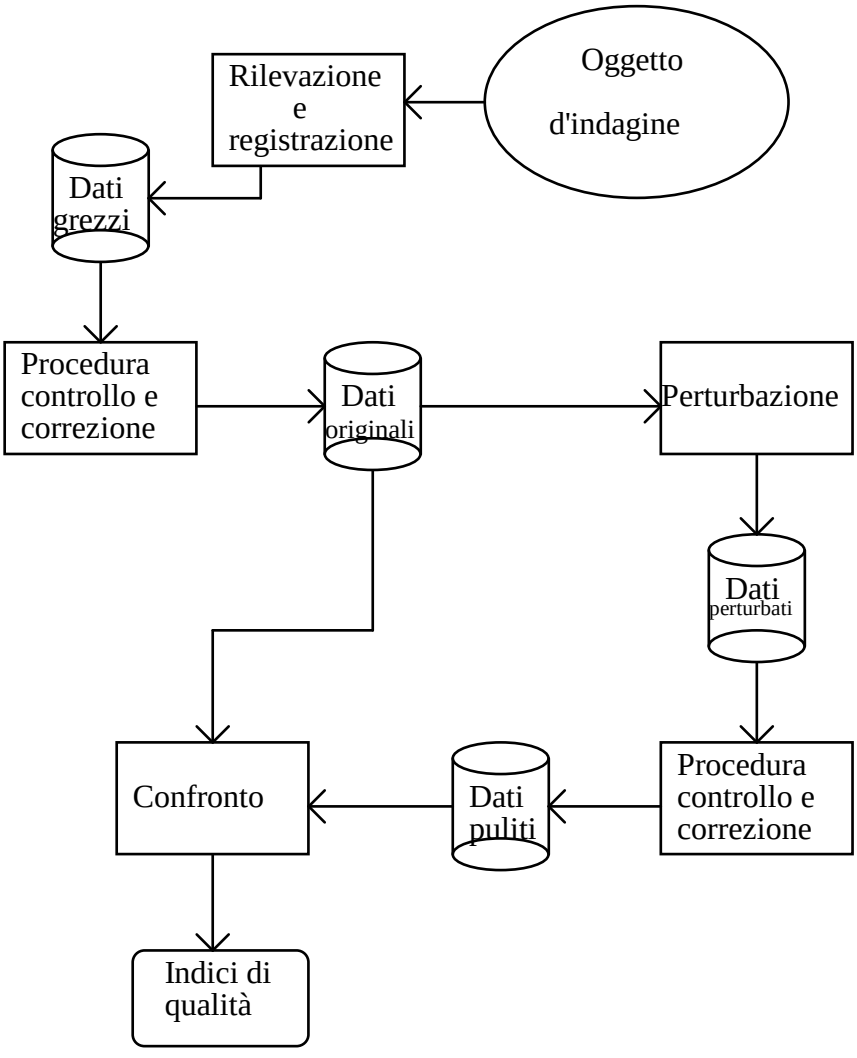
- dati *esatti*: ogni variabile di ogni record del file rivela una piena corrispondenza tra valori registrati e quelli veri relativi alle unità rilevate nel mondo oggetto di indagine;
- dati *grezzi*, i dati effettivamente a disposizione una volta rilevati e memorizzati su supporto magnetico mediante le procedure correnti;
- dati *puliti*, ottenuti dall'applicazione di una data procedura di controllo e correzione ai dati grezzi.

Ognuno dei file deve contenere i record relativi alle stesse unità rilevate nella stessa ripetizione dell'indagine, e deve essere possibile mettere in corrispondenza esatta ogni record dei tre file. Soddisfatte queste condizioni, onde valutare la bontà della procedura, si pongono a confronto i dati esatti con quelli puliti, verificando la "distanza" tra questi e quelli.



In realtà, è estremamente difficile poter disporre di dati "esatti", completamente liberi da errori. Solo nel caso in cui si effettui un'indagine campionaria di controllo, in cui si ponga

estrema cura in tutte le fasi di rilevazione e memorizzazione dei dati (per esempio attraverso tecniche avanzate di acquisizione dei dati: CAPI, CATI, CASI), è possibile considerare trascurabili gli errori contenuti nei dati. Se questo non è possibile, occorre rinunciare al concetto di "dati esatti" e considerare, al suo posto, quello di "dati originali", o "dati di confronto". Generalmente, viene preso un file di dati che risulti almeno "corretto" rispetto ad un dato piano di compatibilità: tale file può essere il risultato dell'applicazione della stessa procedura di controllo e correzione oggetto di test (ed in tal caso coincide col file dei dati "puliti"), o di altre procedure più accurate, i cui controlli costituiscano un sovrainsieme di quelli della procedura da validare. A questo punto, il file in questione viene detto "originale" (o "di riferimento"), e viene sottoposto a perturbazione controllata: all'interno delle variabili, cioè, vengono introdotti in una porzione più o meno grande dei record, valori diversi da quelli originali. Questa perturbazione può essere di tipo puramente stocastico (generazione di valori casuali eseguita in modo indipendente per ogni variabile), ed in tal caso produce errori di tipo casuale, oppure può essere effettuata in modo da simulare i difetti riscontrati nei processi di acquisizione e memorizzazione dei dati, dando luogo in tal modo alla componente sistematica degli errori. In ogni caso, possiamo assimilare il risultato della perturbazione dei dati originali ai dati "grezzi" contenenti errori. A questi viene applicata la procedura da validare, ottenendo i dati "puliti", che vengono messi a confronto con quelli originali.



Nell'effettuare la perturbazione dei dati, va per prima cosa definito il *livello di errore*, cioè la percentuale α di valori da perturbare per ogni variabile. Tipicamente, si lavora sotto diverse ipotesi: di livello basso, medio e alto (ad esempio, rispettivamente, α = 5%, 10%, 20%). L'inverso di α rappresenterà il passo di campionamento, indicherà cioè ogni quanti record occorrerà procedere alla perturbazione della singola variabile.

Se si è interessati a valutare solamente la capacità della procedura di *ricostruire* i valori originari, i valori attribuiti al momento della perturbazione devono essere tali da richiedere comunque l'imputazione: si procede perciò ad attribuire dei valori che risultino fuori del dominio di definizione della variabile da perturbare.

Se al contrario, si intende valutare l'adeguatezza complessiva della procedura, quindi anche la sua potenza in termini di capacità di *individuazione* degli errori, i valori attribuiti dovrebbero rispecchiare la distribuzione effettiva della variabile nei dati a disposizione. In altre parole, la generazione dei valori dovrebbe essere governata dalla particolare funzione di densità della variabile da perturbare, funzione di densità che può essere approssimata dalla distribuzione delle frequenze dei valori assunti da tale variabile nel file dei dati originali. Consideriamo, ad esempio, la seguente distribuzione relativa alla variabile SESSO, desunta dai dati a disposizione:

Valori	distribuz. % delle frequenze	% cumulata
fuori dominio	5	5
maschi	47	52
femmine	48	100

Generare una modalità in modo da tener conto di questa distribuzione è estremamente semplice: è sufficiente generare un numero r compreso tra 0 e 100. Se r è minore di 6, viene attribuito un valore fuori dominio alla variabile SESSO, se è compreso tra 6 e 52, il valore "maschio", se infine r risulta tra 53 e 100 viene attribuito il valore "femmina". La variabile nell'esempio è di tipo qualitativo: per quelle di tipo quantitativo si può invece procedere come segue:

- se la forma della funzione di densità è nota, se ne stimano i parametri mediante i dati a disposizione, ed il nuovo valore viene generato sulla base di tale funzione;
- altrimenti, si procede come nel caso delle variabili qualitative, provvedendo a suddividere la variabile quantitativa in classi di ampiezza, e generando, in successione, dapprima un numero casuale che individui la classe (secondo le stesse modalità viste nell'esempio), e, successivamente, un numero compreso tra gli intervalli della classe assumendo distribuzione uniforme al suo interno.

Si può o meno porre il vincolo che il nuovo valore sia diverso da quello corrente: supponiamo di assumere tale vincolo.

Effettuata la perturbazione, si possono dare le seguenti possibilità:

1. il valore attribuito è fuori dominio: la variabile verrà sicuramente definita come errata dalla procedura di controllo e, in quanto tale, imputata;
2. il valore attribuito è nel dominio dei valori ammissibili: tale valore può o meno essere in contraddizione coi valori assunti da altre variabili nello stesso record o in altri record. Nel

secondo caso il valore errato non attiva alcuna delle incompatibilità, il record risulta pertanto formalmente esatto, e la variabile errata non verrà mai riconosciuta come tale;

3. nel caso invece di attivazione di incompatibilità, la procedura può o meno essere in grado di stabilire, al momento della localizzazione degli errori, che tale attivazione è dovuta in tutto o in parte al valore della variabile perturbata, e può quindi procedere alla sua imputazione. Altrimenti, la procedura può stabilire che l'attivazione delle incompatibilità è dovuta ai valori di altre variabili, e procedere all'imputazione di queste ultime, anche se esatte, introducendo in tal modo nuovi errori;
4. in caso di localizzazione corretta, la procedura procede all'imputazione della variabile perturbata. Si danno due casi: viene ripristinato il valore originario della variabile, oppure un altro valore che, pur determinando il rientro del record in una condizione di correttezza (in quanto determina la disattivazione delle incompatibilità), è diverso da quello "vero".

La procedura da validare è qualitativamente accettabile quanto più riesce ad individuare le variabili perturbate (e dunque concettualmente "errate"), e a ripristinarne il valore originario (e dunque "vero"), mantenendo nel contempo entro limiti ragionevoli l'introduzione di nuovi errori nei dati (attribuzione di valori "falsi" a variabili non perturbate).

Le statistiche da valutare nel processo di validazione sono quindi le seguenti:

- nel complesso del file, il rapporto tra il numero di record che contenengono errori (variabili perturbate) e che sono riconosciuti come errati dalla procedura, ed il totale dei record perturbati (indice di ***adeguatezza del controllo***):

$$C = \frac{R_{err}^{ind}}{R_{err}} \times 100$$

- per ogni variabile, il rapporto tra il numero di volte che la j-esima variabile, essendo stata perturbata, è stata correttamente individuata come errata sul totale di volte che è stata perturbata; ed inoltre il rapporto tra il numero di volte che la variabile, non perturbata, è stata giudicata errata sul totale di volte che non è stata perturbata (indici di ***qualità della localizzazione degli errori***):

$$L_j^1 = \frac{p_j^{err}}{p_j} \times 100 \qquad \text{e} \qquad L_j^2 = \frac{\bar{p}_j^{err}}{\bar{p}_j} \times 100$$[22]

- per ogni variabile sottoposta ad imputazione, la distanza tra valori originali e valori imputati (***qualità dell'imputazione***). Qui proponiamo due indicatori per le variabili quantitative: l'*errore sistematico relativo* $ESR_j = \frac{\sum_{i=1}^{p_j}(O_{ji} - I_{ji})}{\sum_{i=1}^{p_j} I_{ji}} \times 100$ che evidenzia quanto, ed in quale verso, i valori imputati si discostano, in media, da quelli originali[23], ed

[22] Il termine p_j indica il numero di volte che la variabile j-esima è stata perturbata, mentre $\bar{p}_j$ indica le volte che *non* è stata perturbata. L'apice "err" indica le volte che la variabile è stata giudicata errata dalla procedura, e come tale suscettibile di imputazione

[23] Si intende con O_{ji} il valore originale, e con I_{ji} il valore imputato

il *coefficiente di disuguaglianza* $CDI_j = \sqrt{\frac{\sum_{i=1}^{p_j}(O_{ji} - I_{ji})^2}{\sum_{i=1}^{p_j} I_{ji}^2}}$ che fornisce una misura euclidea della distanza esistente tra valori originali e valori imputati della variabile j (errore quadratico), indipendente dall'unità di misura della variabile stessa. Per quanto riguarda invece le variabili qualitative, proponiamo il più semplice *indice di ripristino*

$$IR_j = \frac{\sum_{i=1}^{p_j} y_{ji}}{p_j} \times 100 \qquad y_{ji} = \begin{cases} 1 \text{ se } O_{ji} = I_{ji} \\ 0 \text{ altrimenti} \end{cases}$$

Tutti gli indici proposti variano tra 0 e 100. Alti valori di C, L_j^1 e IR_j, unitamente a bassi valori di L_j^2, ESR_j e CDI_j indicano una buona qualità della procedura di controllo e correzione degli errori.

Questi problemi saranno ulteriormente approfonditi nel capitolo 8 del manuale, relativo alla valutazione delle prestazioni e dell'impatto di un piano di editing sui dati dell'indagine.

9. La Valutazione delle prestazioni di un piano di editing: un quadro complessivo.

I maggiori istituti di statistica ricorrono sempre più all'uso di tecniche di data editing per il controllo e la correzione dei dati delle indagini effettuate. Tali tecniche sono applicate con l'aiuto di strumenti informatici che incorporano metodologie statistiche molto sofisticate sia per la localizzazione dell'errore che per la correzione del dato errato. La correzione è spesso effettuata con la tecnica del donatore o con metodi ausiliari se non è reperibile una adeguata unità donatrice. In questo filone rientrano i sistemi SCIA e GEIS, il primo concepito e realizzato in ambito ISTAT per la correzione delle variabili qualitative tramite la metodologia Fellegi-Holt, il secondo realizzato da Statistics Canada (Kovar, MacMillan, Whitridge, 1991) per la correzione di variabili di tipo quantitativo.

L'impiego di questi metodi automatici non si è però sempre affiancato ad un tentativo di misurare il loro impatto sui dati. Non riteniamo sufficiente la sola conoscenza accurata del metodo e delle ipotesi statistiche su cui si basa la tecnica di data editing, all'atto della sua applicazione. Essa infatti deve accompagnarsi ad una misura della modifica subita dai dati di partenza. Questa misurazione aiuta a valutare criticamente il comportamento della tecnica di data editing e a raffinarla per un suo riutilizzo successivo sugli stessi dati.

Si vuole sottolineare che sempre i dati di una indagine "puliti" da una tecnica di editing sono il frutto di un processo iterativo, in cui le prime applicazioni sono provvisorie e successivamente raffinate sulla base di molte considerazioni, tra cui la misura dell'impatto sui dati iniziali.

Per questa esigenza di misurazione si sono enucleate dalla letteratura statistica una serie di tecniche e indici atti a confrontare due distribuzioni secondo uno stesso carattere.

Un'altra necessità dell'utilizzatore delle tecniche di data editing è quella di una valutazione della loro capacità correttiva, anche al fine di valutare l'efficienza comparativa di più tecniche di editing. Una misura dell'efficienza si effettua con la valutazione del ripristino di dati veri che sono stati alterati con un processo di perturbazione. Si presenta a questo scopo un semplice modello formale di analisi delle varie fasi operative del piano di editing, dalla localizzazione dell'errore alla sua correzione, in cui si introducono indici di misura della capacità correttiva del piano di editing, nonché delle sue fonti di errore.

Ponendosi sempre nell'ottica dell'utilizzatore del piano di editing, spesso si desidera conoscere quali siano le unità statistiche o le variabili maggiormente modificate dal piano. Nel caso delle unità, la loro conoscenza è necessaria o per predisporre una eventuale reintervista delle unità maggiormente modificate, oppure se si desidera studiare le caratteristiche di queste unità ai fini di una migliore successiva applicazione del piano. Per quel che concerne le variabili, se si stanno correggendo più variabili quantitative sintetizzate da una variabile di somma, interessa conoscere quali siano le più importanti per l'azione correttiva del piano. In quest'ottica sono presentati alcuni indicatori.

9.1. Definizione generale del problema.

Si suppone di disporre dei dati di una indagine organizzati in un file. Definiamo **piano di editing** un insieme di passi logici, tramite i quali si può definire una singola osservazione esatta o errata e, nel caso risulti errata, procedere alla sua correzione. Dalla definizione, segue che le principali fonti di errore del piano sono:

1) una osservazione esatta può essere giudicata errata;

2) una osservazione errata può essere giudicata esatta,
3) non sempre la correzione ricostruisce esattamente il valore esatto.

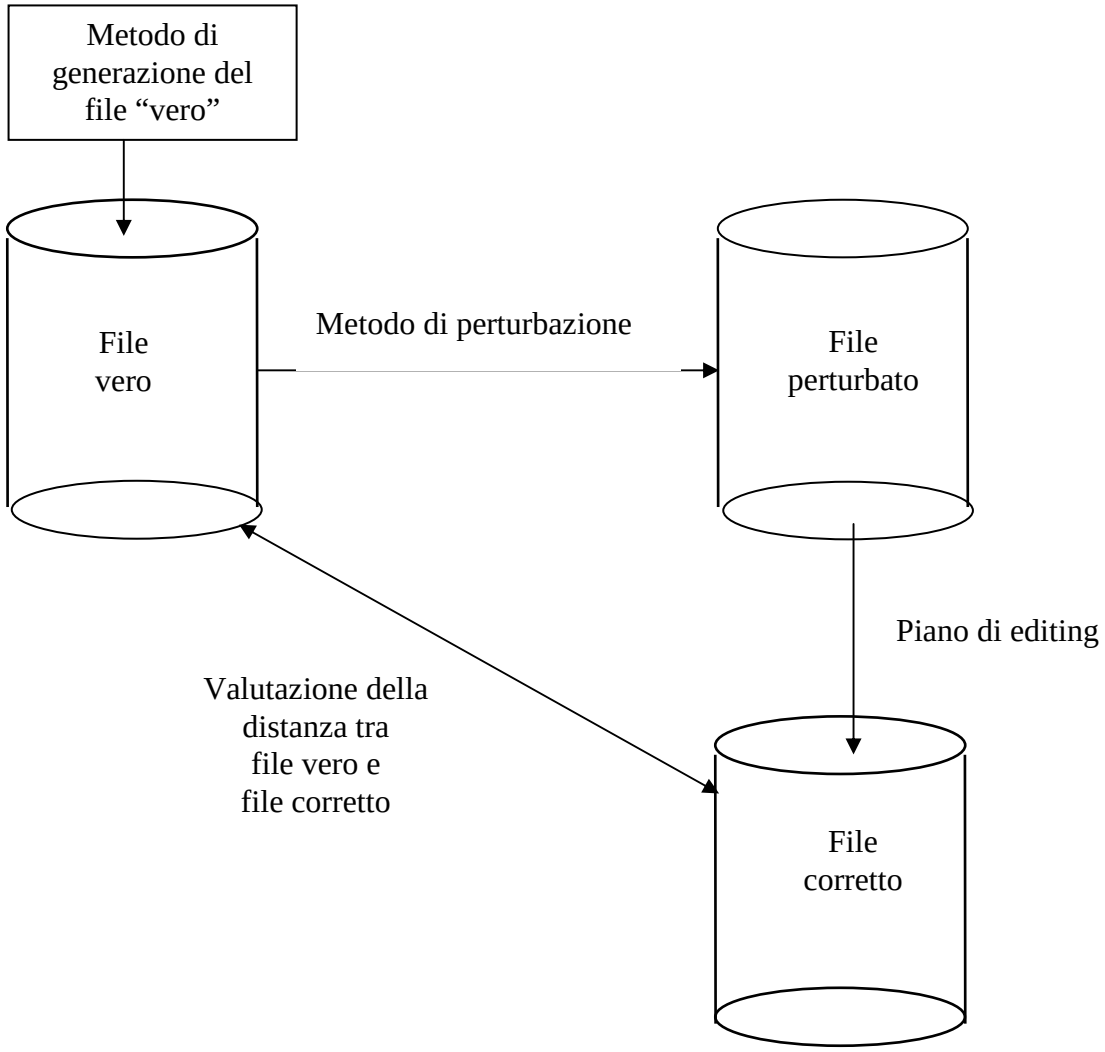
L'obiettivo è di valutare come il piano di editing utilizzato agisce sui dati. Distinguiamo tre possibili contesti operativi.
1) Si dispone di un file di dati grezzi da correggere, risultante dalla registrazione dei risultati dell'indagine, che differisce in qualche misura dal file dei dati "vero". Ci si può limitare a confrontare i dati grezzi e quelli puliti dal piano e valutare la loro distanza. La valutazione della variazione apportata ai dati può essere di ausilio per raffinare il piano di editing per una sua successiva riapplicazione.
2) Si dispone di un file di dati che si ritiene "vero". In questo caso si può introdurre un meccanismo artificiale di generazione di errori per ottenere dati alterati. E' possibile valutare il piano di editing applicato su questi ultimi dati, verificando se esso riesce a ripristinare i dati veri di partenza.
3) Si dispone sia di un file di dati "vero", sia di un file di dati grezzi. Nella realtà il primo file può essere disponibile per un sottoinsieme delle unità formanti il file dei dati grezzi. In questo caso si potrà verificare se il piano di editing applicato ai dati grezzi ricostruisce dati vicini a quelli veri.

9.2. Metodi basati sulla perturbazione di un file "vero".

Approfondiamo il secondo contesto operativo, tra quelli introdotti nel paragrafo 2. Si sottopone a perturbazione un file di dati "vero", modificando un sottoinsieme dei dati, in modo da ottenere un file di dati perturbato. Ad essi viene applicato il piano di editing, per ottenere un file dei dati corretto. Occorrerà quindi definire:
A) il metodo di generazione del file "vero", laddove esso non sia già disponibile.
B) il metodo di perturbazione;
C) una metrica per valutare la distanza del file corretto rispetto a quello "vero" che si cerca di ripristinare.

Possiamo rappresentare il processo nel seguente modo grafico (vedi Figura 1):

Figura 1 - Schema della valutazione del piano di editing tramite perturbazione di un file di dati veri



I metodi di generazione del file "vero" sono essenzialmente di due tipi.

1) Metodi basati sull'integrazione e correzione dei dati di una indagine per cui sono disponibili dati grezzi. Dopo aver revisionato il file dei dati grezzi, ottenendo in questo modo un file di dati "veri", si prescinde dal file dei dati grezzi di partenza.

2) Metodi di generazione automatica, di tipo casuale. Essi consistono nell'applicazione di un generatore automatico di numeri casuali secondo una certa distribuzione per creare delle osservazioni artificiali, con alcune operazioni preliminari:

- si crea una tabella contenente l'insieme completo delle modalità rilevate sulle unità ;
- si predispongono i vincoli di compatibilità tra le varie combinazioni delle variabili;
- si ipotizzano delle distribuzioni di frequenza per le modalità delle variabili e si predetermina il numero totale di unità da generare;
- si generano le singole unità, a partire dalle distribuzioni definite a priori, rispettando i vincoli di compatibilità. Un'applicazione di questo metodo si trova in Nordbotten (1995). Per una

trattazione generale delle applicazioni della simulazione si veda il testo di Klejinen, e Van Groedendaal (1992).

La perturbazione del file "vero" può avvenire secondo:
a) modelli di generazione puramente casuali dell'errore,
b) modelli non puramente casuali di generazione dell'errore.

Gli errori di registrazione delle indagini reali sono ben simulabili con il metodo a), mentre il metodo b) si adatta meglio a modellizzare gli errori di compilazione, che possono nascondere un meccanismo non casuale di generazione dell'errore.

Barcaroli e Luzi (1995) presentano un esempio di piano di editing applicato all'indagine mensile ISTAT su "Occupazione, Retribuzioni e Orario di lavoro nelle grandi imprese", per i mesi di Gennaio e Febbraio 1992. Trattandosi di caratteri quantitativi, per perturbare i dati sono state utilizzate due funzioni generatrici di valori casuali: la RANUNI, che genera un valore di una distribuzione uniforme e la RANNORM che genera un valore estratto da una distribuzione normale standardizzata. In base al primo valore prodotto dalla RANUNI, si decide se modificare o meno la variabile per quel particolare record-unità; in caso di modifica, è generato un secondo valore mediante la stessa funzione, in base al quale si decide se cancellare il valore in questione o perturbarlo. In caso di perturbazione, il valore che va a sostituire l'originale è generato casualmente da una RANNORM con media e varianza rilevate sulla variabile oggetto di editing. Questa applicazione è un esempio di metodo di tipo a) di generazione di errore puramente casuale.

Pallara e Abbate (1997), relativamente al carattere qualitativo volume di affari dell'archivio ASIA della provincia di Cagliari, perturbano i dati dapprima con un meccanismo puramente casuale (viene oscurato il 10% delle osservazioni), in seguito oscurano circa il 40% dei dati delle divisioni 45 (Costruzioni) e 81 (Sanità e altri servizi sociali) di ATECO, relativi alle imprese con il minor volume di affari. Quest'ultimo metodo è del tipo b) non casuale, perché, come gli autori spiegano, si basa:
1) sul fatto che per quelle divisioni di ATECO si rileva una maggiore frequenza di non risposte;
2) per problemi organizzativi le imprese a basso volume di affari hanno maggiori probabilità di non risposta.

9.3. Disponibilità di un file dei dati grezzi e di un file dei dati "vero": metodi basati su indagini di controllo.

Approfondiamo il terzo contesto operativo, tra quelli introdotti nel paragrafo 2. Un metodo per ottenere un file "vero", in modo da avere un termine di paragone affidabile sull'impatto di un piano di editing applicato ad un file di dati grezzi consiste nel condurre una indagine ripetuta sulle stesse unità. Nel corso di essa, le unità possono essere sottoposte ad intervista: o si tratta di una intervista ripetuta, oppure di una prima intervista in assoluto, se esse hanno in precedenza compilato il questionario di risposta senza assistenza dell'intervistatore.

Questo metodo ha lo svantaggio di essere costoso e poco tempestivo, inoltre non tutte le unità che sono entrate a far parte dell'indagine sono propense a farsi reintervistare. Ne consegue che le unità che possono essere ricontattate sono una frazione esigua di quelle originarie, quindi il file "vero" rappresenta un sottoinsieme ristretto di quello grezzo di partenza. Per un interessante esempio applicativo si rimanda ad un lavoro di Granquist (1995).

9.4. Tecniche di confronto tra il file dei dati grezzi e quello ripristinato da un piano di editing.

9.4.1. Presentazione del problema.

L'obiettivo di questa sezione è presentare una serie di tecniche statistiche per confrontare due differenti rilevazioni della stessa variabile sullo stesso insieme di unità. Queste tecniche sono utilizzabili sia per il confronto tra dati grezzi e dati corretti da un piano di editing, sia per il confronto tra dati veri e dati ripristinati dal piano in seguito ad una perturbazione.

I metodi che si presenteranno servono quindi per confrontare due distribuzioni empiriche di una variabile e misurare la loro distanza. Nel caso si disponga soltanto di un file di dati grezzi da correggere e non si ha il file "vero", queste tecniche sono utili per misurare di quanto il file dei dati corretti dal piano si differenzi rispetto al file grezzo originario. Se si utilizza un file di dati "veri", sottoposto a perturbazione e corretto, le tecniche sono di ausilio per verificare se il piano ha ripristinato una distribuzione di dati prossima a quella "vera".
Tutti gli indici presentati sono desunti dal testo di G. Leti (1983).

9.4.2. Confronti tra due distribuzioni semplici secondo un carattere qualsiasi.

Una tecnica elementare di confronto è costituita dall'esame visivo della differenza tra gli istogrammi di frequenza che rappresentano le distribuzioni del carattere nelle due distribuzioni.

Se si vuole misurare la dissomiglianza con un indice, si può procedere nel seguente modo. Sia assegnato un carattere qualsiasi avente k modalità. Siano f_{Ai} per i=1....k le frequenze relative delle modalità nel primo gruppo di unità, f_{Bi} per i=1....k quelle del secondo gruppo. Si noti che non si presuppone - e non lo si farà neanche in seguito - che i due gruppi abbiano la stessa numerosità, cosa che nell'applicazione di un piano di editing normalmente si presume perché si mettono a confronto due distribuzioni relative allo stesso insieme di unità. Un indice di dissomiglianza variabile tra 0 e 1 tra i due gruppi, indicato con z_1 , si costruisce sommando le differenze assolute tra le frequenze corrispondenti alle stesse modalità e dividendo la somma per il suo valore massimo pari a 2, che si ottiene nel caso di massima dissomiglianza. Esso si presenta quando le unità dei due gruppi assumono un'unica modalità, che però è differente tra i due gruppi. Si ha quindi:

$$z_1 = \frac{1}{2}\sum_{i=1}^{k} \left| f_{Ai} - f_B \right|$$

9.4.3. Considerazioni sull'applicabilità dell'indice z_1

Si pone il problema di quali modalità considerare nell'applicazione della formula: spesso nel file dei dati grezzi un insieme di unità presenta la modalità 'missing' (dato

mancante), oppure un insieme di modalità non ammissibili diverse dal 'missing'. Il piano di editing sostituirà a tali valori degli altri, compresi nell'insieme dei valori ammissibili. Quando ci si trova di fronte a questi casi due sono le possibili strategie.

1) La prima consiste nel considerare per il file dei dati grezzi soltanto le unità-record che presentano valori ammissibili. Occorre in questo caso ricalcolare le f_{Ai} , escludendo quelle riferite ai valori non accettabili e facendo in modo che la somma di quelle restanti sia pari all'unità. In questo modo si porranno a confronto due insiemi di dati di diversa numerosità, ma aventi le stesse modalità. Questa scelta è accettabile in particolare quando i valori anomali sono molto numerosi e la loro inclusione genererebbe un valore dell'indice molto elevato soprattutto a causa della loro presenza. Si sceglie inoltre questa strategia se si vuole confrontare la distribuzione di partenza dei valori ammissibili con quella di uscita dal piano di editing, in termini di sole frequenze relative, senza considerare i valori anomali.
2) La seconda strategia consiste nel non eliminare i valori anomali, perché si è interessati anche a studiare la loro redistribuzione ad opera del piano di editing all'interno delle modalità ammissibili. Ovviamente si perde questa importante informazione adottando la prima strategia.

La scelta migliore sembra quella di calcolare l'indice z_1 in entrambi i casi e di esaminare la relativa distribuzione di frequenza dei dati grezzi confrontata con quella finale. Dalle due analisi si ricaveranno diversi tipi di informazioni. Illustriamo queste considerazioni tramite un semplice esempio numerico.

Esempio

Supponiamo di aver rilevato su n=4000 unità un carattere qualitativo non ordinabile con 6 modalità A, B, C, D, E, F e di avere la seguente distribuzione dei dati grezzi.

Tabella 1 - Distribuzione globale dei dati grezzi

MODALITA'	**NUMEROSITA'**	**FREQ. RELATIVA(%)**
A	400	10
B	600	15
C	200	5
D	900	22,5
E	700	17,5
F	900	22,5
Dato Mancante	300	7,5
Totale	4000	100

Supponiamo inoltre che la modalità 'Dato Mancante' non sia ammissibile: essa verrà quindi sostituita dal piano di editing con una delle sei modalità ammissibili. Siamo quindi interessati a verificare come la distribuzione dei valori grezzi ammissibili sia modificata. Essa si presenta nel seguente modo:

Tabella 2 - Distribuzione dei dati grezzi ammissibili

MODALITA'	NUMEROSITA'	FREQ. RELATIVA(%)
A	400	10,8
B	600	16,3
C	200	5,4
D	900	24,3
E	700	18,9
F	900	24,3
Totale	3700	100

Dopo l'applicazione del piano di editing, supponiamo di avere la seguente situazione.

Tabella 3 - Distribuzione dei dati dopo l'applicazione del piano di editing

MODALITA'	NUMEROSITA'	FREQ. RELATIVA(%)
A	420	10,5
B	640	16
C	200	5
D	930	23,3
E	750	18,7
F	1060	26,5
Totale	4000	100

Confrontando la tabella 3 e la tabella 1, si può studiare la redistribuzione dei valori anomali nell'insieme di quelli ammissibili.

L'indice z_1, calcolato tra la distribuzione finale e quella dei grezzi globale, vale:

$$\frac{1}{(2*100)}(|10,5-10|+|16-15|+|5-5|+|23,3-22,5|+|18,7-17,5|+|26,5-22,5|+|0-7,5|)=0,07$$

L'indice z_1, calcolato tra la distribuzione finale di tabella 3 e quella dei grezzi ammissibile di tabella 2, vale invece:

$$\frac{1}{(2*100)}(|10,5-10,8|+|16-16,3|+|5-5,4|+|23,3-24,3|+|18,7-18,9|+|26,5-24,3|)=0,02$$

La differenza tra l'indice calcolato nel primo e nel secondo modo consente di misurare l'impatto della redistribuzione del valore non ammissibile 'Dato Mancante' all'interno dei valori ammissibili.

9.4.4. Confronti tra due distribuzioni semplici secondo un carattere ordinato.

Anche per questo caso, se non ci si vuole limitare al solo esame visivo delle due distribuzioni di frequenza, è possibile costruire un indice di dissomiglianza. Sia dato un carattere ordinato avente k modalità. Siano F_{Ai} per i=1....k le frequenze relative cumulate delle modalità nel primo gruppo di unità, F_{Bi} per i=1....k quelle del secondo gruppo. Un indice di dissomiglianza tra i due gruppi variabile tra 0 e 1, indicato con z_2 , si costruisce sommando le differenze assolute tra le frequenze cumulate corrispondenti alle stesse modalità e dividendo la somma per il suo valore massimo pari a k-1, che si ottiene nel caso di massima dissomiglianza. Quest'ultima si ottiene quando in un gruppo tutte le unità presentano la modalità ad un estremo, mentre nell'altro tutte presentano la modalità all'estremo opposto. Si ha perciò:

$$z_2 = \frac{1}{k-1}\sum_{i=1}^{k}|F_{Ai} - F_{Bi}|$$

Circa l'applicabilità di questo indice, valgono le stesse considerazioni fate per z_1. Occorre però, trattandosi di un carattere ordinato, decidere in quale posizione considerare gli eventuali valori anomali presenti nell'insieme dei dati grezzi: la decisione può essere presa solo conoscendo le caratteristiche dell'indagine. La scelta più frequente è di collocarlo ad uno degli estremi dell'insieme ordinato delle modalità del carattere.

9.4.5. Confronti tra due distribuzioni semplici secondo un carattere quantitativo.

Il confronto tra due distribuzioni semplici secondo un carattere quantitativo si può effettuare confrontando il valore assunto in esse da indici quali media, deviazione standard, coefficiente di variazione, mediana, quantili, etc.

Si possono anche costruire degli indici di dissomiglianza: una trattazione esaustiva della costruzione di essi per questo caso si trova nel già citato libro di G. Leti. Di essa esponiamo il seguente caso semplificato: sia dato un carattere quantitativo rilevato su due gruppi di unità aventi stessa numerosità n. Si costruisce un indice di dissomiglianza z_3, variabile tra 0 e 1, nel seguente modo: si ordinano in sequenza crescente i valori nei due gruppi, ottenendo due insiemi di dati così indicati:

$x_{(1)}, x_{(2)}, x_{(3)},....., x_{(n)}$ nel primo gruppo di unità;
$y_{(1)}, y_{(2)}, y_{(3)},....., y_{(n)}$ nel secondo gruppo.

Siamo M ed m rispettivamente i valori massimo e minimo rilevati nei due gruppi considerati nel loro complesso. z_3 si esprime nel seguente modo:

$$z_3 = \frac{1}{M-m}\left(\frac{1}{n}\sum_{i=1}^{n}\left|x_{(i)} - y_{(i)}\right|\right),$$

dove M-m è il massimo dell'espressione contenuta in parentesi. Tale valore è ottenuto quando in un gruppo tutte le unità assumono il valore massimo M, nell'altro assumono il valore minimo m.

9.4.6. Confronti aggregati tra due distribuzioni semplici secondo un carattere quantitativo.

In alcuni casi, soprattutto quando sono in gioco grandezze di tipo monetario, può essere interessante misurare lo scostamento relativo tra il totale della variabile nel file sottoposto al piano di editing e il totale della stessa variabile in un file precedente, che può essere quello grezzo o quello 'vero' sottoposto a perturbazione. Convenendo al solito che il record contiene l'informazione relativa ad un'unica unità statistica, indichiamo con Ii il valore della variabile nel record i-esimo in seguito al piano di editing e con Oi il valore corrispondente nel file precedente. Sia n il numero totale di record presenti in entrambi i file. Lo scostamento può essere misurato con il seguente indice:

$$I = \frac{\sum_{i=1}^{n}\left(I_i - O_i\right)}{\sum_{i=1}^{n} O_i} 100$$

L'indice misura la variazione percentuale che subisce il valore totale della variabile aggregata, rapportato al totale originario.

Un indice analogo, che però misura la variazione soltanto relativamente ai record che sono stati effettivamente modificati è il seguente (indichiamo con p il numero di record modificati dal piano di editing):

$$\bar{I} = \frac{\sum_{i=1}^{p}\left(I_i - O_i\right)}{\sum_{i=1}^{p} O_i} 100$$

Quest'ultimo indice è anche detto **errore sistematico relativo**, introdotto anche nel capitolo 7 del manuale. I due indici debbono essere entrambi utilizzati, perché il primo serve per valutare la variazione del totale della variabile relativamente a tutti i record presenti nel file, il secondo soltanto per quelli che sono stati modificati.

9.4.7. Caratteri quantitativi: tecniche di confronto basate sulla verifica di ipotesi.

Nell'ambito dell'inferenza statistica sono stati sviluppati una serie di test che prescindono da qualsiasi ipotesi sulla forma della distribuzione, definiti test non parametrici. Consideriamo il caso di carattere quantitativo perché più frequente nelle applicazioni. Tra questi test, sono di particolare interesse quelli che permettono di confrontare due insiemi di

unità e verificare l'ipotesi nulla che esse non siano diverse rispetto a certi parametri o certe caratteristiche funzionali. Nel ambito del data editing i due insiemi di unità possono essere:
1) quello su cui sono stati rilevati i dati grezzi e quello sottoposto al piano di editing;
2) quello su cui sono stati rilevati i valori "veri" e quello sottoposto al piano di editing.

Nel primo caso ha interesse misurare l'impatto del piano sulla distribuzione di partenza, nel secondo è importante verificare se la distribuzione di partenza sia stata rispettata. Nel caso di popolazioni provenienti da distribuzioni continue, può essere utilizzato il test di Smirnov-Kolmogorov (Vitale, 1993). Nel caso di una distribuzione qualsiasi, si utilizza il test del χ^2. Questo test è di grande interesse per il nostro particolare problema, perché nella presentazione e nella sintesi dei risultati di molte indagini statistiche, le variabili quantitative vengono riclassificate e rese discrete, ottenendo una distribuzione di frequenza rappresentata tramite istogramma. E' perciò importante verificare se la distribuzione discretizzata di partenza sia stata modificata.

Sia data quindi una distribuzione così definita:
su n unità è rilevato un carattere quantitativo riclassificato in k classi $(-\infty,a_0]$, $(a_0, a_1]$, $(a_1, a_2]$, $(a_2, a_3]$,....., $(a_{k-1}, +\infty)$. Si indica con f_{Ai} la frequenza relativa delle unità nella classe i-esima per il primo insieme di unità (costituito dai dati grezzi o da quelli "veri"), con f_{Bi} la stessa frequenza rilevata sui dati corretti dal piano di editing. Interessa verificare se le frequenze relative restino stabili. In termini di teoria della prova di ipotesi, si vuole sottoporre a test l'ipotesi nulla:
$H_0 : f_{Ai} = f_{Bi}$ per ogni i=1....k, contro l'ipotesi alternativa $H_1 : f_{Ai} \neq f_{Bi}$ per qualche i=1....k. Si utilizza per il test la statistica:

$$\chi^2 = \sum_{i=1}^{k} \frac{(nf_{Bi} - nf_{Ai})^2}{nf_{Ai}}$$

che sotto l'ipotesi nulla si distribuisce come un χ^2 con k-1 gradi di libertà. Per la validità dei risultati del test, si raccomanda in genere di avere $nf_{Ai} \geq 5$. Se per alcune classi questo non si verifica, si raccomanda di fonderle con alcune classi vicine, per soddisfare alla condizione con un numero ridotto di classi. Di recente, (si veda sempre Vitale) risultati analitici e numerici hanno portato a definire questa regola: se il numero k di classi è non inferiore a 3, indicato con m il numero di classi per le quali si ha: $nf_{Ai} \leq 5$, ne segue che, indicando con nf_{Ai*} la frequenza minima, per poter applicare il test, si deve verificare: $nf_{Ai*} \geq 5\left(\frac{m}{k}\right)$.

9.4.8. Tecniche di confronto per più caratteri quantitativi basate sulla costruzione di una variabile derivata.

Spesso non interessa come risultato dell'indagine il valore delle variabili rilevate direttamente, ma il valore di variabili che sono una loro sintesi. Si confrontano perciò i due gruppi rispetto a queste variabili sintetiche, utilizzando qualcuna delle tecniche finora descritte. Se una variabile di sintesi presenta notevoli variazioni nei due gruppi, ciò implica che in essi esistono notevoli differenze per quel che riguarda le singole variabili originarie, sulle quali occorrerà effettuare analisi dettagliate.

9.5. Valutazione della capacità di un piano di editing.

Si vogliono valutare le prestazioni di un piano di editing per tutte le variabili osservate su un insieme di unità. Il metodo che si descrive di seguito deve essere iterato per tutte le variabili in gioco, che possono essere di tipo qualsiasi. E' disponibile un file di dati veri, indicato con F1, contenente un insieme di V record, ciascuno corrispondente ad un'unità. Da ora in poi indichiamo con $|V|$ il numero di elementi formanti l'insieme V.

Si usa un meccanismo di perturbazione per ottenere un file di dati perturbati, indicato con F2: esso è composto dagli insiemi V_1 e V_2, contenenti rispettivamente i record non perturbati e quelli perturbati.

Definiamo come **tasso di perturbazione** il seguente rapporto:

$$p = \frac{|V_2|}{|V|}$$

E' utile usare diversi valori di p per testare il piano di editing in situazioni di crescente difficoltà operativa dal punto di vista dell'incidenza degli errori, per considerare un insieme completo di condizioni operative.

Abbiamo già dato la definizione di piano di editing. Vogliamo ora valutare se esso agisca correttamente. Assumiamo che esso agisca in due fasi: nella prima sono localizzati i record errati e come risultato si produce un file con alcuni record contrassegnati come errati tramite un flag di errore. Questo file è indicato con F3 e risulta composto dai seguenti insiemi:
-V_3 , insieme dei record individuati come errati;
-V_4 , insieme dei record individuati come esatti.
I due insiemi possono essere così scomposti:

$$V_3 = V_5 \cup V_6 \ (V_5 \cap V_6 = \phi), \text{ dove } V_5 \subseteq V_1 \text{ e } V_6 \subseteq V_2$$
$$V_4 = V_7 \cup V_8 \ (V_7 \cap V_8 = \phi), \text{ dove } V_7 \subseteq V_2 \text{ e } V_8 \subseteq V_1$$

Si tenga altresì conto che:

$$V_1 = V_5 \cup V_8 \text{ e } V_2 = V_6 \cup V_7$$

Quindi V_5 è l'insieme dei record giudicati incorrettamente errati, mentre V_7 contiene i record erroneamente giudicati corretti. Questi due insiemi sono le prime fonti di errore del piano di editing. Possiamo ora definire i seguenti tre indici.

La **capacità identificativa dei valori errati:**

$$CI = \frac{|V6|}{|V2|} \in [0,1]$$

è l'indice che quantifica la capacità del piano di identificare correttamente gli errori nei dati grezzi.

L'**errore di identificazione dei valori errati:**

$$EI = \frac{|V5|}{|V1|} \in [0,1]$$

è l'indice che quantifica gli errori commessi identificando valori esatti come errati (può assimilarsi ad un *erro*re *di primo tipo*).
Infine, l'**errore di identificazione dei valori corretti:**

$$EII = \frac{|V7|}{|V2|} \in [0,1]$$

quantifica gli errori commessi nell'identificare dati errati come veri (è assimilabile ad un *errore di secondo tipo*).

I tre termini non sono indipendenti, come si mostra facilmente:

$$E_{II} = \frac{|V7|}{|V2|} = \frac{|V2| - |V6|}{|V2|} = 1 - C_I$$

L'indice C_I, insieme agli errori E_I ed E_{II}, è utile per la valutazione delle prestazioni del piano in fase di individuazione degli errori.

Nella seconda fase il piano effettua la correzione di una parte dei record contrassegnati come errati, generando un file pulito indicato come F4. I 4 file finora definiti contengono lo stesso insieme di record-unità. Il valore dell'unica variabile in studio può cambiare solo in uno dei modi seguenti:

- può essere alterato dal meccanismo di perturbazione, nel passaggio da F1 a F2;
- il valore della variable può essere contrassegnato come errato, nel passaggio da F2 a F3;
- un valore della variabile contrassegnato come errato può essere modificato, nel passaggio da F3 a F4.

Il file F4 risulta composto da due insiemi di record:
-V_9 , insieme dei record sottoposti a correzione;
-V_{10} , insieme dei record non sottoposti a correzione.

Imponiamo un vincolo di *coerenza* al piano di editing, intendendo con ciò che esso può correggere soltanto valori precedentemente contrassegnati come errati. Il vincolo si esprime algebricamente come: $V_9 \subseteq V_3$.

Possiamo anche scomporre gli insiemi V_9 e V_{10} :

$$-V_9 = V_{11} \cup V_{12} \quad (V_{11} \cap V_{12} = \phi), \text{ dove } V_{11} \subseteq V_5 \subseteq V_1 \text{ e } V_{12} \subseteq V_6 \subseteq V_2$$

$$-V_{10} = V_{13} \cup V_{14} \quad (V_{13} \cap V_{14} = \phi), \text{ dove } V_{13} \subseteq V_3 \text{ e } V_{14} \equiv V_4 = V_7 \cup V_8$$

V_{11} rappresenta così l'insieme dei record veri, che però sono sottoposti a correzione come conseguenza diretta del fatto che E_I nella realtà è superiore a zero.

Introduciamo ora l'**indice di efficienza operativa** del piano di editing:

$$IEFF = \frac{|V9|}{|V3|} \in [0,1]$$

Esso quantifica la capacità del piano di correggere la frazione più alta possibile dei record contrassegnati come errati. Ne segue che V_{10} è formato sia da record che non dovrebbero essere corretti (V_4) che da record da correggere, ma che non lo sono (V_{13}) a causa di problemi interni di efficienza operativa del piano, esprimibili tramite la condizione $I_{EFF}<1$.

Supponiamo che I_{EFF} sia uniforme su tutti i sottoinsiemi di V_3 e V_9. Tale uniformità può esprimersi come:

$$\forall\, V_3^1 \subset V_3, \text{ se } V_9^1 = V_3^1 \cap V_9 \Rightarrow \frac{|V_9^1|}{|V_3^1|} = \frac{|V_9|}{|V_3|} = I_{EFF}$$

Si tratta di una assunzione ragionevole: si richiede in pratica che il comportamento del piano, per quanto riguarda la frazione di record corretti, non dipenda dal particolare sottoinsieme di record contrassegnati come errati.

F4 può in alternativa considerarsi formato da:

- record in realtà errati (V_{ES}),
- record in realtà esatti (V_{ERR}).

Approfondiamo la struttura dei due insiemi. Con questo obiettivo, decomponiamo ulteriormente V_9, scrivendo:

$$V_{12} = V_{15} \cup V_{16} ,$$

dove V_{15} e V_{16} sono nell'ordine:

1) l'insieme dei record corretti in modo appropriato, fra tutti quelli correttamente giudicati errati,

2) l'insieme dei record non adeguatamente corretti, fra tutti quelli correttamente giudicati errati.

In questo modo si crea un'altra fonte di incoerenza per il piano di editing: una significativa percentuale di record errati, contrassegnati come tali, non corretti in modo appropriato.

E' possibile effettuare una ulteriore decomposizione di V_{10}, scrivendo:

$$V_{13} = V_{17} \cup V_{18}, \text{ dove } V_{17} \subseteq V_6 \subseteq V_2 \text{ e } V_{18} \subseteq V_5 \subseteq V_1$$

L'insieme V_{17} è una fonte di errore derivante direttamente dal fatto che $I_{EFF} <1$, mentre V_{18} è lo strano risultato del fatto che sia E_I che I_{EFF} sono diversi dal loro valore ideale. Poiché V_{18} è formato da record esatti, erroneamente giudicati affetti da errore, ma non corretti, non costituisce una fonte di errore.

Possiamo poi scrivere:

$$(V_{20} \equiv V_8^{(*)}), \qquad V_{ES} = V_{20} \cup V_{15} \cup V_{18},$$

dove dunque V_{18} è una componente non desiderabile dell'insieme.

Essendo necessariamente: $V_{19} \equiv V_7 \subset V_{ERR}$, ne segue che:

$$V_{ERR} = V_{11} \cup V_{16} \cup V_{17} \cup V_{19}$$

(*) il simbolo ≡ significa "uguale per definizione a"

Introduciamo due ulteriori indici.

La **capacità correttiva:**

$$C_{II} = \frac{|V_{15}|}{|V_2|} \in [0,1]$$

esprime la capacità del piano di assegnare il valore corretto al record errato in fase di correzione.

D'altro canto l'**errore di correzione:**

$$E_{III} = \frac{|V_{16}|}{|V_2|} \in [0,1]$$

quantifica la frazione di record errati che sono stati corretti con un valore non appropriato.
E_{III} può derivarsi da grandezze introdotte in precedenza:

$$E_{III} = \frac{|V_{16}|}{|V_2|} = \frac{I_{EFF}|V_6| - |V_{15}|}{|V_2|} = I_{EFF}\, C_I - C_{II},$$

avendo assunto l'uniformità di I_{EFF}. Assegnati $|V_1|$ and $|V_2|$, con semplici passaggi algebrici si ottiene la numerosità di tutti gli insiemi che compongono V_{ERR} e V_{ES} in funzione esclusivamente di C_I, C_{II}, E_I e I_{EFF}, come mostrato nelle due successive tabelle.

Tabella 4 - Numerosità degli insiemi affetti da errore

V_{ERR}	$0<I_{EFF}<1$	$I_{EFF}=1$
V_{11}	$I_{EFF}\, E_I\, \|V_1\|$	$E_I\, \|V_1\|$
V_{16}	$(I_{EFF}\, C_I - C_{II})\, \|V_2\|$	$(C_I - C_{II})\, \|V_2\|$
V_{17}	$(1-I_{EFF})C_I\, \|V_2\|$	0
V_{19}	$(1-C_I)\, \|V_2\|$	$(1-C_I)\, \|V_2\|$

Tabella 5 - Numerosità degli insiemi esatti

V_{ES}	$0<I_{EFF}<1$	$I_{EFF}=1$
V_{15}	$C_{II}\|V_2\|$	$C_{II}\|V_2\|$
V_{20}	$(1-E_I)\|V_1\|$	$(1-E_I)\|V_1\|$
V_{18}	$(1-I_{EFF})E_I\|V_1\|$	0

Si mostrano i risultati anche nel caso particolare di $I_{EFF}=1$ perché, nelle applicazioni reali, un metodo ausiliario di correzione è spesso aggiunto al piano automatico di editing, che presenta $I_{EFF}<1$, allo scopo di ottenere un piano complessivo con $I_{EFF}=1$. Si noti che per derivare le espressioni contenenti I_{EFF}, si è spesso sfruttata la proprietà della sua uniformità.

I risultati ottenuti nel caso $I_{EFF}<1$ sono particolarmente utili per valutare il nucleo di un piano di editing, in genere realizzato con una procedura informatica. Molti sistemi di editing sono infatti composti da un metodo principale automatico, concepito per eliminare la maggior parte degli errori, lasciando la parte restante di errori a metodi ausiliari, espressamente pensati per il caso di mancato funzionamento del metodo principale. Si noti che i valori nelle tabelle potevano scriversi in funzione di $|V|$, anziché di $|V_1|$ e $|V_2|$, tramite le relazioni:

$$|V_1|=(1-p)|V|,\quad |V_2|=p|V|$$

Ricaviamo le numerosità degli insiemi presentate nelle Tabelle 4 e 5.

Per la Tabella 4 si ha:

$$|V_{11}|=I_{EFF}|V_5|=I_{EFF}E_I|V_1| \qquad (I_{EFF}=1\Rightarrow |V_{11}|=E_I|V_1|)$$

$$|V_{16}|=|V_{12}|-|V_{15}|=I_{EFF}|V_6|-|V_{15}|=I_{EFF}C_I|V_2|-C_{II}|V_2|=(I_{EFF}C_I-C_{II})|V_2|$$
$$(I_{EFF}=1\Rightarrow |V_{16}|=(C_I-C_{II})|V_2|)$$

$$|V_{17}|=|V_6|-|V_{12}|=(1-I_{EFF})|V_6|=(1-I_{EFF})C_I|V_2|$$
$$(I_{EFF}=1\Rightarrow |V_{17}|=0)$$

$$|V_{19}|\equiv|V_7|=E_{II}|V_2|=(1-C_I)|V_2|$$

Per la Tabella 5 si ha:

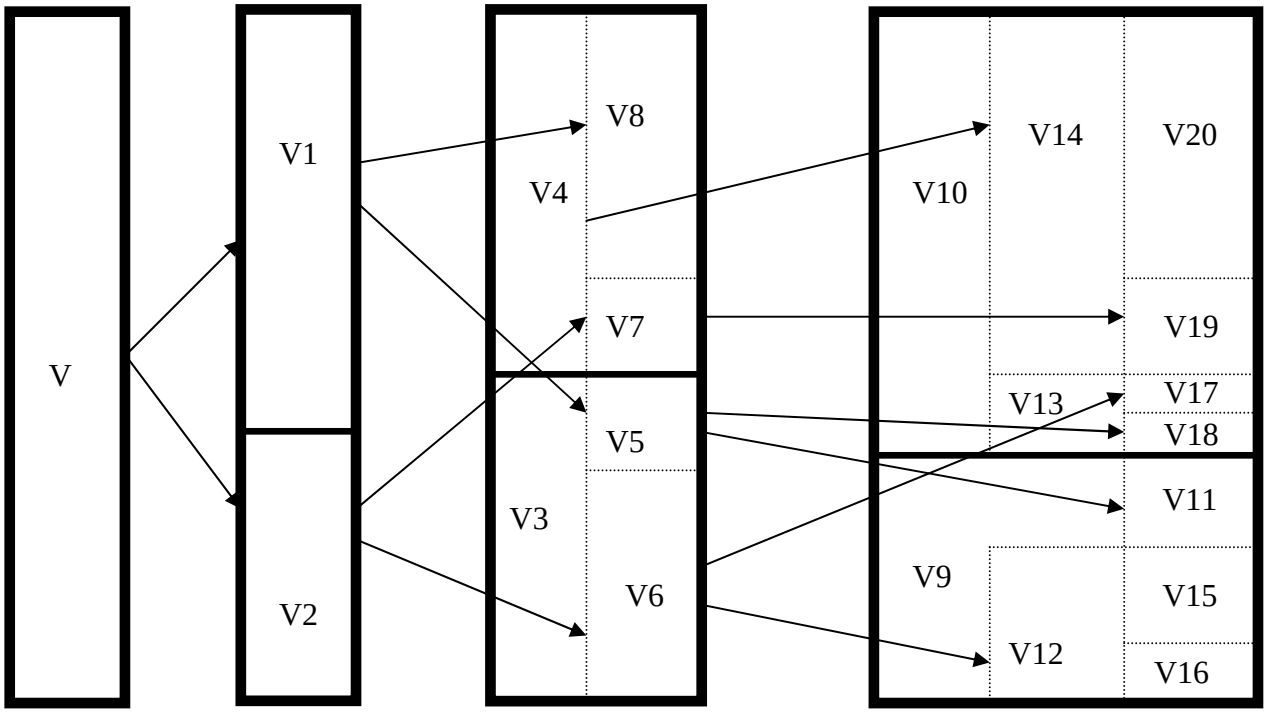
$$|V_{15}| = C_{II}|V_2|$$

$$|V_{20}| \equiv |V_8| = |V_1| - |V_5| = |V_1| - E_I|V_1| = (1 - E_I)|V_1|$$

$$|V_{18}| = |V_5| - |V_{11}| = (1 - I_{EFF})|V_5| = (1 - I_{EFF})E_I|V_1|$$
$$(I_{EFF} = 1 \Rightarrow |V_{18}| = 0)$$

Cerchiamo ora di ottenere una sintesi visiva del modello, che evidenzi le relazioni tra i vari insiemi di record introdotti. Ciò può effettuarsi con lo schema in Figura 2:

Figura 2 -Schema completo di passaggio dai dati veri a quelli corretti dal piano di editing



F1: dati veri F2: dati grezzi F3: dati controllati F4: dati corretti

Legenda:

V : dati veri
V_1: dati non perturbati
V_2: dati perturbati
V_3: dati contrassegnati con il carattere di errore per la correzione
V_4: dati non contrassegnati con il carattere di errore

V_5: dati veri contrassegnati con il carattere di errore per la correzione
V_6: dati perturbati contrassegnati con il carattere di errore per la correzione
V_7: dati perturbati non contrassegnati con il carattere di errore per la correzione
V_8: dati veri non contrassegnati con il carattere di errore per la correzione
V_9: dati sottoposti a correzione
V_{10}: dati non sottoposti a correzione
V_{11}: dati veri contrassegnati con il carattere di errore e sottoposti a correzione
V_{12}: dati perturbati contrassegnati con il carattere di errore e sottoposti a correzione
V_{13}: dati contrassegnati con il carattere di errore e non sottoposti a correzione
V_{14}: dati non contrassegnati con il carattere di errore e non sottoposti a correzione
V_{15}: dati perturbati contrassegnati con il carattere di errore e sottoposti ad appropriata correzione
V_{16}: dati perturbati contrassegnati con il carattere di errore e sottoposti a non appropriata correzione
V_{17}: dati perturbati contrassegnati con il carattere di errore e non sottoposti a correzione
V_{18}: dati veri contrassegnati con il carattere di errore e non sottoposti a correzione
$V_{19} \equiv V_7$
$V_{20} \equiv V_8$

Nel grafico di Figura 2, i sottoinsiemi di tonalità grigia corrispondono agli errori contenuti nei vari file, ad iniziare da F2 per finire a F4. In quest'ultimo, i sottoinsiemi di tonalità grigia contengono valori errati non corretti, o valori veri erroneamente corretti, o valori errati non adeguatamente corretti. La definizione di questi sottoinsiemi e delle loro interrelazioni ci ha permesso la costruzione degli indicatori di valutazione della qualità del piano di editing.

9.6. Razionalizzazione del piano di editing: individuazione delle unità o variabili maggiormente modificate.

9.6.1. Correzione di una sola variabile di tipo quantitativo.

Supponiamo di disporre di una sola variabile quantitativa Y. Sia Y_I il totale della variabile rilevata sul file di ingresso, Y_{II} il suo totale calcolato sul file di uscita del piano di editing. Definiamo **variazione totale** della variabile Y e la indichiamo con Δ_Y la seguente espressione:

$$\Delta_Y = \left| Y_{II} - Y_I \right|$$

Sia n il totale delle unità; su ognuna si rileva Y_{Ii} , contenuto nel file di ingresso. Nel file di uscita si ha per la stessa unità Y_{IIi} , per cui: $Y_{Ii} = Y_{IIi}$, se non è stato alterato il valore della variabile per quell'unità, altrimenti si ha: $Y_{Ii} \neq Y_{IIi}$. Supponiamo di riordinare le n unità in modo che si abbia:

$\left| Y_{II(1)} - Y_{I(1)} \right| \geq \left| Y_{II(2)} - Y_{I(2)} \right| \geq \ldots\ldots\ldots\ldots\ldots \geq \left| Y_{II(n)} - Y_{I(n)} \right|$
Si può quindi scrivere:

$$\Delta_Y = \sum_{i=1}^{n} \left| Y_{II(i)} - Y_{I(i)} \right|$$

Per un generico i≤n il rapporto:

$$C^I{}_i = \frac{\sum_{j=1}^{i} \left| Y_{II(j)} - Y_{I(j)} \right|}{\Delta_Y} 100$$

rappresenta il *contributo percentuale apportato alla variazione totale della variabile Y dalle prime i unità più modificate dal piano di editing*.

Questo indice può avere diversi significati. Ad esempio, nel caso in cui il file di ingresso sia quello dei valori grezzi, esso indica le unità più pesantemente modificate: su di esse si può decidere ad esempio di intervenire di nuovo, o con tecniche di correzione alternative o tramite reintervista. Un semplice esame grafico dell'andamento di C^I_i rispetto al numero di unità darebbe un'idea della concentrazione dell'azione correttiva.

9.6.2. Correzione di più variabili di tipo quantitativo.

Si suppone di disporre di k variabili di tipo quantitativo Y_1 ..... Y_k, rilevate su n unità. Facciamo l'ipotesi che sia di interesse per l'indagine la **variabile aggregata** Y, rappresentata dal totale delle k variabili calcolato sia su ogni unità che sul loro complesso.

Siano Y_{TI1}..... Y_{TIk} i totali delle k variabili nel file di ingresso. Analogamente Y_{TII1},.....,Y_{TIIk} sono i totali delle k variabili nel file di uscita. Indichiamo rispettivamente con Y_{Iij} e Y_{IIij} il valore della variabile j rilevato nell'unità i nel file di ingresso e in quello di uscita. Possiamo quindi scrivere:

$$Y_{TIj} = \sum_{i=1}^{n} Y_{Iij} \quad \text{e} \quad Y_{IIj} = \sum_{i=1}^{n} Y_{IIij} \text{ per ogni j=1.....k.}$$

Indichiamo poi con Y_{TI} e Y_{TII} i totali della variabile aggregata Y nei due file di ingresso e di uscita. Essi si scrivono come:

$$Y_{TI} = \sum_{j=1}^{k} Y_{TIj} = \sum_{j=1}^{k} \sum_{i=1}^{n} Y_{Iij} \quad \text{e} \quad Y_{TII} = \sum_{j=1}^{k} Y_{TIIj} = \sum_{j=1}^{k} \sum_{i=1}^{n} Y_{IIij}$$

Riordiniamo le differenze tra i totali delle variabili calcolati nei due file, in modo che risulti:

$$\left| Y_{TII(1)} - Y_{TI(1)} \right| \geq \left| Y_{TII(2)} - Y_{TI(2)} \right| \geq \ldots\ldots\ldots\ldots\ldots\ldots \geq \left| Y_{TII(k)} - Y_{TI(k)} \right|$$

Per la variabile aggregata Y indichiamo con Δ_{TY} l'espressione $\Delta_{TY} = \left| Y_{TII} - Y_{TI} \right|$. Si può quindi scrivere:

$$\Delta_{TY} = \sum_{j=1}^{k} \left| Y_{TII(j)} - Y_{TI(j)} \right|$$

Per un generico j≤k il rapporto:

$$C^{II}{}_j = \frac{\sum_{i=1}^{j} \left| Y_{TII(i)} - Y_{TI(i)} \right|}{\Delta_{TY}} 100$$

rappresenta il *contributo percentuale apportato alla variazione totale della variabile aggregata Y dalle prime j variabili più modificate dal piano di editing***.**

L'indice serve ovviamente per determinare quali siano le variabili più importanti per il processo di correzione dei dati. Notiamo poi che si può scrivere:

$$\Delta TY = \sum_{j=1}^{k}\sum_{i=1}^{n} \left|Y_{IIij} - Y_{Iij}\right| = \sum_{i=1}^{n}\sum_{j=1}^{k} \left|Y_{IIij} - Y_{Iij}\right|$$

Ordiniamo in modo crescente le somme delle differenze tra i valori assunti dalle k variabili nella stessa unità, in modo da avere:

$$\sum_{j=1}^{k} \left|Y_{II(1)j} - Y_{I(1)j}\right| \geq \sum_{j=1}^{k} \left|Y_{II(2)j} - Y_{I(2)j}\right| \geq \ldots\ldots \geq \sum_{j=1}^{k} \left|Y_{II(n)j} - Y_{I(n)j}\right|$$

Per un generico i≤n il rapporto:

$$C^{III}i = \frac{\sum_{h=1}^{i} \sum_{j=1}^{k} \left|Y_{II(h)j} - Y_{I(h)j}\right|}{\Delta_{TY}} 100$$

rappresenta il ***contributo percentuale apportato alla variazione totale della variabile aggregata Y dalle prime i unità più modificate dal piano di editing*.**

L'indice quantifica il contributo delle i unità più rilevanti per il piano di editing. Su queste unità si deve accentrare l'attenzione di chi conduce l'indagine, ai fini di una loro eventuale reintervista.

9.6.3. Una semplice applicazione dell'indice C^{I}_{i} all'editing selettivo.

Spesso è di interesse verificare che lo scostamento relativo tra il totale di una variabile aggregata prima e dopo l'applicazione del piano non superi una certa soglia p, compresa tra 0 e 1. In questo caso si decide che ci si trova in un caso sospetto e si desidera sottoporre almeno a un riesame i risultati. Usando le stesse quantità introdotte nel sotto-paragrafo 7.1, si vuole verificare questo vincolo:

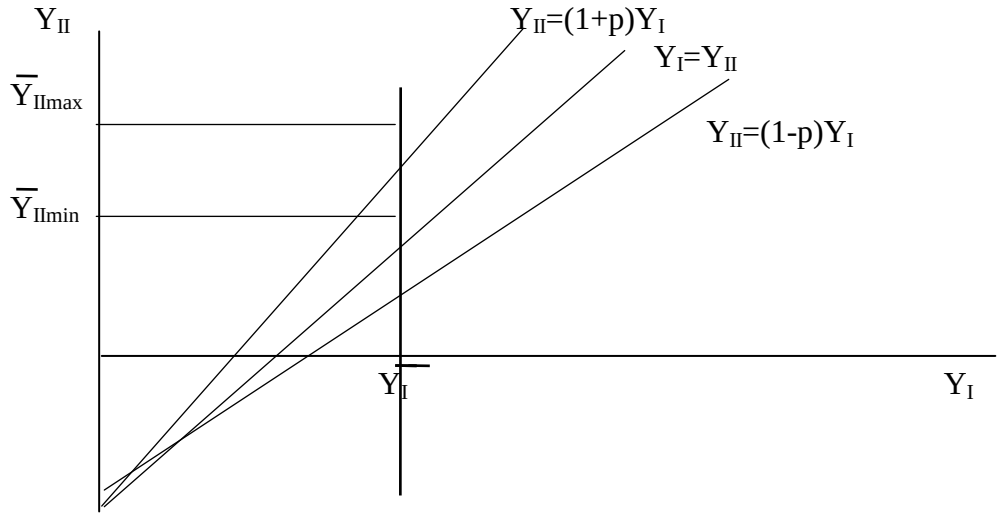
$$\frac{\Delta Y}{Y_I} \leq p ,$$

che può essere diviso nella coppia di disequazioni:

$$Y_{II} \leq (1+p)Y_I , \; Y_{II} \geq (1-p)Y_I$$

La situazione può essere visualizzata su un grafico cartesiano (vedi Figura 3):

Figura 3 - Grafico di controllo dell'andamento di Y_{II} in funzione di Y_I



Con Y_I si indica il valore effettivo di Y_I. Se Y_{II} risulta esterno all'intervallo $[Y_{Imin}, Y_{IImax}]$, l'indice C_1^I segnala l'unità maggiormente responsabile della variazione del totale della Y. Si procede quindi a un riesame del valore $Y_{II(1)}$ relativo a quell'unità e lo si sottopone a revisione, eventualmente ricontattando l'unità. Se il valore $Y_{II(1)}$ è stato variato in seguito alla revisione, si procede al ricalcolo di Y_{II} e si verifica se rientra nell'intervallo di accettazione. Il processo può anche proseguire riesaminando le unità relative ai valori $Y_{II(2)}$, $Y_{II(3)}$ e così via.

10. Documentazione

Dobbiamo distinguere tra la documentazione delle procedure in sè, e la documentazione dei risultati dell'applicazione di tali procedure ai dati a disposizione.

10.1. Documentazione delle procedure

Una procedura di controllo e correzione è normalmente composta da una serie di *passi*, ognuno dei quali è un'elaborazione, interattiva o automatica, contraddistinta da uno o più file di dati in input, ed uno o più file di dati in output[24].

Una procedura si intende documentata quando:

- è descritto il suo *flusso interno*, cioè la particolare concatenazione dei passi che ne fanno parte;
- sono fornite le *specifiche di ogni passo* di elaborazione.

La descrizione complessiva è adeguatamente fornita da un *diagramma di flusso*, da una rappresentazione grafica, cioè, della sequenza delle elaborazioni che intercorrono tra la disponibilità dei dati grezzi e l'ottenimento di quelli puliti.

Per quanto riguarda invece la documentazione del singolo passo, occorre distinguere rispetto alla sua natura, tra

1. passi di predisposizione dei dati;
2. passi propri di controllo/correzione.

I primi sono costituiti essenzialmente da operazioni di ordinamento, unione, suddivisione, ecc. dei dati, e derivazione/calcolo di nuove variabili a partire da quelle disponibili: per loro natura, necessitano di una documentazione limitata. I secondi, invece, costituiscono il cuore della procedura, e va posta particolare cura nella scrittura delle specifiche che li caratterizzano.

La documentazione minima richiesta è data dall'insieme di regole di controllo e correzione che vengono utilizzate nel singolo passo: nel caso di un passo deterministico, si tratterà dell'*insieme dell'insieme di regole di imputazione deterministica* (utilizzate sia per il controllo che per la correzione), in cui sarà importante anche la sequenza di applicazione; nel caso invece di un passo probabilistico, deve essere fornito l'*insieme iniziale di edit in forma normale* (utilizzato nella fase di controllo), già depurato da eventuali inconsistenze e ridondanze, e l'*insieme completo* (utilizzato nella fase di correzione).

Poichè le regole sono riportate secondo il particolare linguaggio richiesto dal sistema elaborativo, è bene accompagnare ogni regola con una sua descrizione in chiaro, laddove il linguaggio non sia di comprensione immediata.

Nel caso di procedure non automatiche, è bene riportare, per ogni attività di tipo interattivo, le norme fornite agli addetti alle operazioni di controllo e correzione dei dati.

[24] Nel caso di procedure di tipo interattivo, il file di input può coincidere col file di output, nel senso che le lavorazioni accedono in lettura/scrittura a record dello stesso file

10.2. Documentazione delle singole applicazioni

E' fondamentale mantenere traccia di ogni singola applicazione di una procedura di controllo e correzione. Relativamente ad ogni data elaborazione, ciò è necessario per valutare le prestazioni della procedura e il suo *impatto* della procedura sui dati e sulle stime degli aggregati ricavabili da tali dati. Considerando invece il complesso delle applicazioni, organizzate secondo una serie storica, è possibile studiare l'andamento sia delle prestazioni che dell'impatto: una procedura giudicata soddisfacente al momento della sua implementazione, può, con l'andar del tempo, non risultare più adeguata a causa del cambiamento delle modalità e/o dell'oggetto della rilevazione. Oppure, al contrario, una procedura che in fase iniziale soffre di limiti derivanti da inadeguatezze presenti nei processi dell'indagine, può superare tali limiti grazie alla rimozione delle carenze individuate.

Per quanto riguarda le procedure di tipo automatico, per ogni esecuzione è utile produrre e mantenere almeno la seguente documentazione:

1. la lista delle *regole di controllo* (edit in forma normale nel caso di procedure probabilistiche, parte "condizione" delle regole di imputazione nel caso deterministico) attivate, *per frequenza di attivazione*;
2. la lista delle *variabili imputate per frequenza di imputazione*, con l'indicazione delle regole che ne hanno causato l'imputazione.

Si suggerisce anche, per le variabili più importanti, di mantenere le *matrici* o i *grafici di transizione*, che mostrano il tipo di impatto non solo quantitativo, ma anche qualitativo, della procedura sulla distribuzione di tali variabili.

Le stesse statistiche possono essere prodotte anche nel caso di procedure di tipo interattivo, o misto: anzichè essere il risultato contestuale all'esecuzione delle procedure automatiche, potranno essere ottenute dall'esecuzione di programmi ad hoc che mettono a confronto i dati grezzi con quelli puliti.

Molto importante, ai fini della documentazione, è il calcolo del cosiddetto *pseudo-bias* relativo alle stime degli aggregati più importanti prodotti dall'indagine:

$$PB_i = \frac{| Y_i - Y_i^{'} |}{Y_i^{'}} \times 100$$

dove PB_i esprime la "distorsione" apportata dall'applicazione della procedura di correzione dei dati alla stima dell'i-esimo aggregato: Y_i è la stima dell'aggregato calcolata sui dati grezzi, mentre $Y_i^{'}$ è la stessa stima calcolata sui dati puliti. Un pseudo-bias alto può indicare due cose:

1. la *procedura di correzione non è adeguata*, ed è anzi tale da allontanare la stima calcolata sui dati puliti dal valore "vero", fornito da quella calcolata sui dati grezzi;
2. *nei dati grezzi sono presenti errori sistematici*: la distorsione è già presente nella stima dell'aggregato calcolata sui dati grezzi, la procedura di correzione apporta una distorsione di segno contrario eliminando l'effetto degli errori sistematici.

La verifica dell'andamento degli pseudo-bias per gli aggregati più importanti è di estremo interesse, in quanto permette di tenere sotto controllo la presenza o l'insorgenza di errori di carattere sistematico, i più pericolosi ai fini della correttezza delle stime.

11. Peculiarità delle indagini sui dati amministrativi

Nei precedenti paragrafi di questo capitolo abbiamo riportato i metodi e le tecniche più importanti utilizzabili in fase di controllo dei dati e di correzione degli errori, unitamente al software che ne permette l'applicazione. Sia le metodologie che il software corrispondente sono stati ideati e sviluppati presso gli enti produttori di statistiche ufficiali, in particolare dagli Istituti nazionali di statistica, in situazioni, cioè, in cui l'attività prevalente è la rilevazione e l'elaborazione dei dati *statistici*, più che l'utilizzo dei dati *amministrativi*. Sorge immediato il quesito se e quanto tali tecniche e software siano utilizzabili non all'interno del classico ciclo produttivo dell'indagine statistica, bensì in una situazione in cui l'obiettivo è quello di utilizzare a fini statistici dati rilevati per altri scopi. Infatti, mentre nel caso dell'indagine statistica gli unici vincoli all'applicazione ottimale delle metodologie e degli strumenti riportati nel presente capitolo sono costituiti dalle risorse (elaborative ed umane) a disposizione e dai tempi da rispettare, al contrario nel caso in cui i dati sono acquisiti all'interno di una procedura amministrativa in una forma e con modalità prefissate occorre tener conto di scelte effettuate senza possibilità di influenza da parte dello statistico.

La differenza fondamentale tra le due situazioni è sintetizzabile nel fatto che mentre nel primo caso lo statistico ha il controllo *completo* del ciclo di produzione dell'informazione statistica, dal momento del disegno dell'indagine a quello della diffusione dei dati, nel secondo, al contrario, esso interviene quando già diverse decisioni sono già state prese.

Un elemento estremamente importante, all'interno di questo quadro, è il fatto che il supporto per l'acquisizione dei dati (il modello) risponde, nella maggioranza dei casi, a finalità di tipo puramente amministrativo anzichè statistico, con rilevanti conseguenze ai fini del controllo. Ad esempio, mentre è naturale, ai fini statistici, inserire nel modello di rilevazione dei *quesiti di controllo* (che costituiscono informazione ridondante, utile per verificare la correttezza dei dati e guidare nelle operazioni di correzione), ciò raramente viene fatto dai responsabili di procedure amministrative.

Ancora, mentre è normale, in molte indagini statistiche, pensare di ricontattare i rispondenti in caso di incongruenze delle risposte, ciò non è pensabile nelle situazioni in cui lo statistico non ha alcun titolo per far questo, non essendo il responsabile delle procedure amministrative che danno luogo ai dati da controllare.

Molto importante è poi la questione del *feedback* che l'analisi del processo di controllo e correzione dovrebbe avere sulle fasi precedenti: tale analisi consente di rilevare le carenze strutturali della rilevazione che danno luogo, ad esempio, agli errori di tipo sistematico. Mentre in un'indagine statistica ciò ha come conseguenza l'intervento su tali carenze ed il loro superamento (modifiche del questionario, formazione dei rilevatori, ecc.), ciò non è normalmente possibile in situazioni in cui le modalità delle procedure amministrative risultano essere sostanzialmente immutabili.

In sostanza, è possibile riassumere i limiti posti dalla particolare natura dei dati amministrativi al momento della fase di controllo, nel fatto che non si ha la possibilità di intervenire *a monte* del processo, in particolare nelle fasi di ideazione del modello e di raccolta dei dati. Questo non è vero in assoluto: nulla vieta che i responsabili amministrativi accolgano i suggerimenti degli statistici, se ciò non comporta sensibili aggravi nella gestione delle procedure. Ciò è tanto più possibile se si sottolinea che la ricaduta sui livelli di qualità dei dati raccolti garantisce ovviamente benefici al servizio cui le procedure in questione sono destinate. Ad esempio, se i dati amministrativi vengono creati da transazioni già informatizzate, l'adozione di tecniche rientranti nel filone CAPI ha un impatto minimo sui

costi e garantisce migliore qualità: pensiamo ad esempio a quegli istituti di cura che registrano già in archivi informatizzati le notizie relative ai ricoveri ed ai dimessi.

Detto questo, quanto esposto in precedenza è applicabile a molte delle situazioni concrete. E' senz'altro possibile applicare procedure automatiche di correzione; anche le tecniche interattive di controllo e correzione citate (macroediting, editing selettivo, grafico ecc.) sono applicabili, con il limite già citato della pratica impossibilità di ritorno al rispondente (follow up), limite che peraltro si riscontra anche in molte indagini statistiche. Estremamente importante, oltre che possibile, è eseguire le operazioni di validazione e documentazione riportate nei relativi paragrafi, che danno indicazioni estremamente preziose in termini di livello di qualità per dati che, ricordiamolo, vengono acquisiti per altri scopi, e la cui possibilità di utilizzo statistico non deve essere data per scontata.

Infine, vogliamo qui citare una tecnica che più che rientrare nella vera e propria fase di controllo e correzione, fa uso dei metodi in essa rientranti, ed è peculiare dei dati amministrativi: tale tecnica è nota come imputazione di massa.

Imputazione di massa

Alcuni Istituti di statistica, tra i quali in particolare Statistics Canada, hanno elaborato una particolare tecnica, che consiste di due fasi:

1. da una collezione di dati amministrativi viene estratto un campione di unità, rispetto al quale viene rilevata informazione addizionale mediante una indagine apposita;
2. anzichè calcolare le stime campionarie, attribuendo pesi alle unità selezionate, si ricostruisce l'intera popolazione imputando le parti mancanti delle unità non selezionate.

Il tasso di campionamento della prima fase si colloca normalmente tra il 10 ed il 30%, cosicchè la seconda fase richiede l'imputazione di una quota oscillante tra il 70 ed il 90% della popolazione (da cui il termine "imputazione di massa"). La differenza rispetto alla classica stima campionaria basata su attribuzione di pesi si ha in quanto mediante determinate tecnche di imputazione viene sfruttata la conoscenza che si ha relativamente a tutte le unità grazie alla disponibilità dei dati amministrativi, cosa che non avviene col semplice riporto all'universo: ciò è tanto più vantaggioso quanto maggiore è la correlazione tra le variabili rilevate con procedure amministrative e quelle rilevate mediante indagine statistica, e da imputare per le unità non selezionate nel campione.

Esiste una corrispondenza diretta tra metodo di aggiustamento del peso e metodo di imputazione: nel caso di campionamento sistematico semplice, imputare attribuendo il valor medio dei rispondenti equivale ad utilizzare un coefficiente di espansione diretto (pari all'inverso della probabilità di inclusione, o, in altri termini, pari all'intervallo di campionamento). L'imputazione da donatore ("nearest neighbor") è invece equivalente ad adottare coefficienti di espansione variabili, ognuno dei quali pari al numero delle volte che il record è stato utilizzato come donatore per una data variabile.

Poichè da un relativamente piccolo campione viene ricostruito l'universo mediante numerose imputazioni di dati, occorre usare le seguenti accortezze:

1. rendere trasparente il meccanismo, fornendo agli utenti il file ottenuto con l'indicazione dei record imputati, e del meccanismo di imputazione utilizzato;
2. valutare e documentare dettagliatamente l'impatto del processo di imputazione, mantendendo sempre separati i due archivi, quello dei dati originali a disposizione, e

quello dei dati imputati ottenuto, utilizzando la metodologia di validazione e documentazione.

Appendice - La metodologia Fellegi-Holt per il controllo e la correzione delle variabili qualitative

Tre sono i criteri fondamentali per l'imputazione delle variabili qualitative alla base della metodologia proposta da Fellegi e Holt[25]:

1. in ogni record i dati devono soddisfare tutte le regole di validità e compatibilità, cambiando il meno possibile il valore dei campi;
2. le regole di imputazione devono essere derivate dalle regole di controllo, senza esplicita specificazione;
3. le distribuzioni di frequenza marginali e congiunte devono essere mantenute il più possibile.

Edit in forma normale

Distinguiamo gli edit logici, riguardanti le variabili qualitative, dagli edit aritmetici, riguardanti le variabili quantitative.

DEFINIZIONE: un ***edit logico*** esprime una condizione di inaccettabilità su una data combinazione di valori di due o più variabili

Un edit può essere formalizzato come l'applicazione di una funzione f a sottoinsiemi dei domini di n variabili:

$$f(A_1^0, A_2^0, \ldots, A_n^0)$$

dove:

A_i^0 : sottoinsieme del dominio della variabile i-esima

f : funzione logica che connette i vari A_i^0 mediante gli operatori logici di intersezione ($\cap$) e unione ($\cup$)

Un record $\underline{a}$ è errato se:

$$\underline{a} \in f(A_1^0, A_2^0, \ldots, A_n^0)$$

Applicando ripetutamente alla f la legge distributiva otteniamo:

$$f(A_1^0, A_2^0, \ldots, A_n^0) =$$
$$(A_{i_1}^1 \cap A_{i_2}^1 \cap \ldots \cap A_{m_i}^1) \cup (A_{j_1}^2 \cap A_{j_2}^2 \cap \ldots \cap A_{m_j}^2) \cup \ldots \cup (A_{k_1}^r \cap A_{k_2}^r \cap \ldots \cap A_{m_k}^r)$$

Possiamo dire che un record è errato se appartiene ad almeno uno dei termini a secondo membro. Definiamo come "edit in forma normale" ognuno di tali termini.

DEFINIZIONE: un ***edit in forma normale*** è un edit logico in cui l'unico operatore ammesso è quello di intersezione

In simboli:

$$\bigcap_{i \in s} A_i^*$$

Ogni edit logico, di qualsiasi forma, può sempre essere tradotto in una serie di edit in forma normale. Consideriamo, ad esempio, la seguente regola (di compatibilità):

"Se una persona ha età inferiore a 16 anni, oppure frequenta una scuola elementare, allora non può essere capo-famiglia, ed il suo stato civile deve essere celibe o nubile"

[25] Quanto segue è una sintesi dell'articolo "A Systematic Approach to Automatic Edit and Imputation" di I.Fellegi e D.Holt pubblicato sul Journal of the American Statistical Association (marzo 1976)

Questa regola può essere convertita in una serie di edit in forma normale attraverso i seguenti passi[26]:

1. *formalizzazione*:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})] \rightarrow [\ (\neg\ \text{Capo-famiglia}) \cap (\text{Celibe/Nubile})\ ]$$

2. *traduzione in regola di incompatibilità*:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})]\ \cap \neg\ [\ (\neg\ \text{Capo-famiglia}) \cap (\text{Celibe/Nubile})\ ] = \text{errore}$$

3. *semplificazione*:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})]\ \cap [\ (\text{Capo-famiglia}) \cup (\neg\ \text{Celibe/Nubile})\ ] = \text{errore}$$

4. *applicazione della legge distributiva*:

$$[\ (\text{Età} < 16) \cap (\text{Capo-famiglia})] \cup$$
$$[\ (\text{Età} < 16) \cap (\neg\ \text{Celibe/Nubile})] \cup$$
$$[\ (\text{Scuola Elementare}) \cap (\text{Capo-famiglia})] \cup$$
$$[\ (\text{Scuola Elementare}) \cap (\neg\ \text{Celibe/Nubile})] = \text{errore}$$

I quattro termini nell'ultima espressione sono altrettanti edit in forma normale.

L'insieme completo degli edit

DEFINIZIONE: gli edit in forma normale specificati direttamente dallo statistico sono detti ***edit espliciti***.

Un record che non attiva alcun edit esplicito si dice corretto, e non necessita di alcuna modifica. Al contrario, un edit che attiva almeno un edit esplicito si dice errato, e necessita della modifica di almeno una variabile.
Mentre gli edit espliciti sono necessari e sufficienti per determinare la correttezza di un record, essi non sono sufficienti per una sua ottimale correzione.

DEFINIZIONE: chiamiamo ***edit implicito*** un edit logicamente contenuto negli edit espliciti.

La funzione degli edit impliciti, considerati congiuntamente con gli edit espliciti, è quella di permettere l'imputazione ottimale di un record errato.

DEFINIZIONE: l'***insieme completo*** degli edit è dato dall'unione degli edit espliciti e di quelli impliciti.

Per eseguire in modo ottimale il passo di scelta delle variabili da imputare, e di determinazione del range di valori imputabili, è necessario preventivamente generare l'insieme completo di edit.

Consideriamo il seguente esempio.
Supponiamo che un record contenga tre variabili, di cui siano definiti i seguenti domini:

[26] Oltre agli operatori di intersezione e unione, facciamo uso anche di quelli di negazione ($\neg$) e implicazione ($\rightarrow$)

VARIABILI	*DOMINI*
ETA'	0-14, 15-99
STATO CIVILE (STACIV)	celibe, coniugato, separato,divorziato,vedovo
RELAZIONE COL CAPO FAMIGLIA (RELCF)	capofamiglia, coniuge, altro

Siano stati definiti i seguenti edit in forma normale espliciti, esprimenti condizioni di incompatibilità:

I. (ETA' = 0-14) $\cap$ (STACIV = coniugato, separato,divorziato,vedovo)
II. (STACIV = celibe, separato, divorziato, vedovo) $\cap$ (RELCF = coniuge)

Possiamo riscriverli come condizioni di compatibilità nel seguente modo:

(ETA' = 0-14) $\rightarrow$ (STACIV = celibe)
(STACIV = celibe, separato, divorziato, vedovo) $\rightarrow$ (RELCF $\neq$ coniuge)

Poichè la conseguenza della prima implicazione è contenuta nella premessa della seconda, possiamo derivare che

(ETA' = 0-14) $\rightarrow$ (RELCF $\neq$ coniuge)

relazione che, opportunamente ritradotta in forma normale, diventa:

III. (ETA' = 0-14) $\cap$ (RELCF = coniuge)

Questo terzo edit era implicitamente contenuto nei primi due.

Supponiamo ora di considerare il seguente record:

(ETA' = 0-14) $\cap$ (STACIV = coniugato) $\cap$ (RELCF = coniuge)

Questo record attiva gli edit I e III.
Per correggere il record, ricerchiamo l'insieme minimo di variabili che *copra* tutti gli edit attivati dal record in questione. Nel nostro caso verifichiamo che la variabile ETA' è presente sia nel primo che nel terzo edit attivato. Per disattivare tali edit è sufficiente assegnare a ETA' un valore interno all'*intersezione dei complementi* dei valori che compaiono negli edit attivati o attivabili:

$$(\neg\ 0\text{-}14) \cap\ (\neg 0\text{-}14) = 15\text{-}99$$

Assegnando il valore 15-99 alla variabile ETA', il record può dirsi corretto, in quanto non attiva alcun edit: nel far ciò abbiamo tenuto conto del principio del minimo cambiamento, in quanto abbiamo modificato una sola variabile.

Se in questo processo di ricerca dell'insieme minimale di variabili da imputare non avessimo tenuto conto dell'edit implicito, avremmo considerato il solo edit I: per disattivarlo, avremmo potuto scegliere di imputare sia ETA' che STACIV. Se avessimo scelto STACIV, che compare anche nell'edit II, avremmo constatato che l'intersezione del complemento dei relativi valori è l'insieme vuoto ∅:

$$\neg \text{ (coniugato, separato,divorziato,vedovo)} \cap \neg \text{ (celibe, separato, divorziato, vedovo)} = \\ = \text{celibe} \cap \text{coniugato} = \varnothing$$

L'impossibilità di trovare dei valori imputabili a STACIV tali da correggere il record deriva dal fatto che STACIV non è contenuto nell'edit III, implicito, attivato dai valori delle variabili ETA' e RELCF. La conseguenza di carattere generale è che *la non considerazione degli edit impliciti non permette di definire sempre insiemi minimi di variabili da imputare che siano in grado di riportare il record in una situazione di correttezza.*

LEMMA: dati s edit $e_r e_r$ e n variabili, per ogni arbitraria variabile i, un edit $e_i : \bigcap_{j=1}^{n} A_j^*$ si dice generato dagli s edit se e solo se

$$\begin{cases} A_j^* = \bigcap_{r \in s} A_j^r & j = 1,2,...,n \quad i \neq j \\ A_i^* = \bigcup_{r \in s} A_i^r \end{cases}$$

In altri termini, fissata una variabile i (detta *generante*), il corrispondente A_i^* sarà ottenuto come *unione* degli A_i^r, mentre ogni altro A_j^* sarà ottenuto come *intersezione* degli A_j^r.

DEFINIZIONE: Un edit generato si dice ***edit implicito essenzialmente nuovo*** se e solo se:

1. A_i^* coincide col dominio della variabile i;
2. ogni A_i^r è non vuoto ed è un sottoinsieme proprio del dominio della variabile i;

Consideriamo il seguente esempio. Siano dati gli edit:

I. (ETA' = 0-14) ∩ (RELCF = qulasiasi) ∩ (STACIV ≠ celibe)
II. (ETA'=qualsiasi) ∩ (RELCF = coniuge) ∩ (STACIV = celibe, separato, divorziato, vedovo)

Se fissiamo ETA' come variabile generante otteniamo:
(ETA'=qualsiasi) ∩ (RELCF = coniuge) ∩ (STACIV = separato, divorziato, vedovo)
che è ridondante rispetto al secondo edit.

Fissando invece RELCF otteniamo:
(ETA'=0-14) ∩ (RELCF = qualsiasi) ∩ (STACIV = separato, divorziato, vedovo)
che è ridondante rispetto al primo edit.

Infine, scegliendo STACIV come variabile generante:

(ETA'=0-14) ∩ (RELCF = coniuge) ∩ (STACIV = qualsiasi)

che è un edit implicito essenzialmente nuovo.

DEFINIZIONE : Un edit generato da due o più edit tra loro contraddittori (inconsistenti) è detto ***edit degenere***

Consideriamo il seguente esempio:

I. (ETA' = 0-14) ∩ (STACIV ≠ celibe)

II. (ETA' = 15-99) ∩ (STACIV ≠ celibe)

Assumendo ETA' come campo generante, otteniamo l'edit esplicito

III. (ETA' = qualsiasi valore) ∩ (STACIV ≠ celibe) = (STACIV ≠ celibe)

che ci dice che sono errati tutti i valori di STACIV diversi da celibe, il che chiaramente contraddice la definizione del dominio della variabile STACIV. L'edit III è un edit degenere, ed in quanto tale può essere generato solo da edit tra loro contraddittori.

I seguenti teoremi e corollari assicurano che, *avendo a disposizione l'insieme completo di edit, un qualsiasi record errato è sempre correggibile, e lo è in modo ottimale*.

Sia $\mathbf{\Omega}$ l'insieme completo di edit, e sia $\mathbf{\Omega}_k$ un sottoinsieme tale da coinvolgere le prime k variabili (con l'esclusione, quindi, di tutti gli edit in cui compaiano le variabili k+1, k+2, ... , n).

TEOREMA 1: se gli a_i^0 sono possibili valori per le prime k-1 variabili, e se questi valori soddisfano tutti gli edit in $\mathbf{\Omega}_{k-1}$, allora esiste un qualche valore a_k^0 tale da soddisfare tutti gli edit in $\mathbf{\Omega}_k$.

La ripetuta applicazione del teorema 1 permette di conseguire il seguente

COROLLARIO 1: se un record ha n variabili, di cui le prime k-1 hanno valori a_i^0 (i=1,2,...,k-1) tali che tutti gli edit in $\mathbf{\Omega}_{k-1}$ sono soddisfatti, allora esistono valori a_i^0 (i=k,k+1,...,n) tali da soddisfare tutti gli edit in $\mathbf{\Omega}$.

Ed inoltre:

COROLLARIO 2: se un record ha n variabili, di cui un sottoinsieme s ha la proprietà che almeno uno dei valori a_i ($i \in s$) compare in ogni edit attivato dal record, allora esistono dei valori a_i^0 ($i \in s$) tali che, assieme agli a_i ($i \notin s$) fanno si che il record soddisfi tutti gli edit.

Metodi di imputazione

La metodologia prevede, per ogni record errato:

1. l'identificazione dell'*insieme minimo di variabili da modificare*;
2. per ogni variabile rientrante nell'insieme minimo, la *determinazione dell'insieme di valori attribuibili*, e *imputazione* di uno tra questi.

Per quanto riguarda il punto 1, ricordiamo che l'insieme minimo di variabili da imputare è costituito da quell'insieme di variabili che "coprono" tutti gli edit attivati dal record e che risulta essere di dimensione minima.

Per quanto concerne il punto 2, sono proposti due metodi, entrambi di tipo *hot deck*, consistenti nell'imputare in una variabile del record corrente (ricevente) il valore della stessa variabile in un record (donatore) scelto tra quelli esatti. I metodi in questione sono:

- metodo dell'imputazione sequenziale;
- metodo dell'imputazione congiunta.

METODO 1: IMPUTAZIONE SEQUENZIALE

Consideriamo un record errato di cui sia già stato identificato un insieme minimo di k variabili da imputare. Il metodo consiste nell'imputare dapprima la k-esima variabile, e poi, sequenzialmente, le variabili k-1,k-2,...,1.
Consideriamo tutti gli M edit in cui

- è presente la variabile k;
- non sono presenti le variabili 1,2,...,k-1.

Tra questi, consideriamo solo gli M' edit in cui non sono presenti gli edit sicuramente disattivati dai valori correnti delle variabili k+1, k+2, ... , n: gli M' edit sono quelli che possono essere attivati o meno in funzione dei valori della sola variabile k. Se vogliamo che il record soddisfi tali edit, il valore da assegnare alla variabile k deve soddisfare la condizione:

$$a_k^0 \in \bigcap_{r=1}^{M'} \overline{A_r^k}$$

cioè deve appartenere all'insieme intersezione dei complementi dei valori indicati per la variabile k in tutti gli M' edit: tale insieme non è mai vuoto per il teorema 1.
Lo stesso procedimento viene iterato per le variabili k-1, k-2, ...1, fino all'esaurimento dell'insieme minimo di variabili da imputare.
Consideriamo il seguente esempio, con 5 variabili:

VARIABILI	*DOMINI*
SESSO	maschio, femmina
ETA	0-14,15-16,17-99
STATO CIVILE (STACIV)	celibe, coniugato, separato, divorziato, vedovo
RELAZIONE COL CAPOFAMIGLIA (RELCF)	moglie, marito, figlio, altro
LIVELLO D'ISTRUZIONE (ISTRUZ)	nessuno,elementare, secondario, post-secondario

L'insieme (completo) degli edit è il seguente:

e_1 : (SESSO=maschio) $\cap$ (RELCF=moglie)
e_2 : (ETA'=0-14) $\cap$ (STACIV$\neq$celibe)
e_3 : (STACIV$\neq$coniugato) $\cap$ (RELCF=moglie,marito)
e_4 : (ETA'=0-14) $\cap$ (RELCF=moglie,marito)
e_5 : (ETA'=0-16) $\cap$ (ISTRUZ=post-secondaria)

Sia dato il seguente record:

VARIABILE	*VALORE*
SESSO	maschio
ETA	12
STACIV	coniugato
RELCF	moglie
ISTRUZ	elementare

Il record attiva gli edit e_1, e_2, e_4. Nessuna singola variabile "copre" i tre edit. Tre coppie di variabili coprono gli edit attivati: (SESSO, ETA'), (ETA', RELCF) e (STACIV, RELCF). Supponiamo di scegliere la coppia (SESSO, ETA'): la dimensione s dell'insieme è pari a 2.
Sia ETA' la variabile k-esima (k=2). Consideriamo tutti gli edit che contengono ETA' ma non SESSO (la variabile k-1=1):
e_2 : (ETA'=0-14) $\cap$ (STACIV$\neq$celibe)
e_4 : (ETA'=0-14) $\cap$ (RELCF=moglie,marito)
e_5 : (ETA'=0-16) $\cap$ (ISTRUZ=post-secondaria)

L'edit e_5 è sempre soddisfatto per qualsiasi valore di ETA' dal momento che nel record il valore di ISTRUZ è "elementare". Per calcolare i valori imputabili ad ETA' dobbiamo quindi considerare solo A_2^2 e A_2^4:

$$a_2^* \in \overline{A_2^2} \cap \overline{A_2^4} \equiv \overline{(0\text{-}14)} \cap \overline{(0\text{-}14)} = (15\text{-}99)$$

cercheremo quindi un record donatore con un valore di ETA' compreso tra 15 e 99: supponiamo 22.
Passiamo ora variabile SESSO (k-1=1). Solo l'edit e_1 la contiene, quindi:

$$a_1^* \in \overline{A_1^1} \equiv \overline{\text{maschio}} = \text{femmina}$$

Essendo unico, il valore "femmina" è direttamente imputato alla variabile SESSO. Il record corretto sarà quindi il seguente:

VARIABILE	*VALORE*
SESSO	femmina
ETA	22
STACIV	coniugato
RELCF	moglie
ISTRUZ	elementare

METODO 2: IMPUTAZIONE CONGIUNTA

Per un dato record errato siano state definite le k variabili da imputare. Si considerino gli M" edit con le k variabili

$$e_r : \bigcap_{i=1}^{n} A_i^r \quad (r=1,2,...,M")$$

dove $a_i^0 \in A_i^r$ (i=k+1,k+2,...,n). Sono gli edit in cui sono presenti le k variabili, e dove le variabili k+1, k+2, ..., n hanno nel record valori interni agli A_i^r: sono cioè gli edit attivabili o meno in funzione dei valori che si danno alle k variabili.
Si considerino gli insiemi

$$A_i^* = \bigcap_{r=1}^{M''} A_i^r \quad (i=k+1, k+2, \ldots ,n)$$

Se scegliamo un qualsiasi record, tra quelli esatti, i cui valori delle variabili k+1, k+2, ... ,n siano interni agli insiemi così definiti, i valori di tale record nelle variabili 1,2,...,k sono attribuibili in blocco al record errato corrente, in quanto costituiscono una combinazione che sicuramente garantisce che tutti gli M'' edit siano soddisfatti (cioè disattivati). Per tale motivo non c'è alcun bisogno di calcolare l'insieme dei valori attribuibili alle k variabili dell'insieme minimo.
Rprendiamo in considerazione l'esempio visto per l'imputazione sequenziale: siano ancora SESSO ed ETA' le variabili dell'insieme minimo: queste due variabili sono presenti negli edit e_1, e_2, e_4 ed e_5. Quest'ultimo è soddisfatto comunque per il valore di ISTRUZ. Restano:

e_1 : (SESSO=maschio) $\cap$ (RELCF=moglie)
e_2 : (ETA'=0-14) $\cap$ (STACIV$\neq$celibe)
e_4 : (ETA'=0-14) $\cap$ (RELCF=moglie,marito)

E' questo l'insieme M'' di edit. Si determinano gli insiemi di valori per le variabili k+1, k+2, ..., n, cioè per STACIV (3), RELCF (4) e ISTRUZ (5):
A_3^* = coniugato, separato, divorziato, vedovo
A_4^* = moglie $\cap$ (moglie, marito) = moglie
A_5^* = qualsiasi valore

A questo punto, tra i record esatti viene ricercato un donatore che abbia i valori di STACIV e RELCF interni agli insiemi così determinati, ed i relativi valori di SESSO ed ETA' vengono attribuiti al record errato corrente.

Bibliografia

ABBATE C., BOVE G., CRESCENZI F. (1992) "Metodi statistici multivariati per la ricostruzione dell'informazione mancante", in *Avanzamenti metodologici e statistiche ufficiali, Atti delle prime giornate di studio SIS-ISTAT*, Roma 13-14 dicembre 1992.

ABBATE C., GIOMMI A. (1993) "Metodi di ponderazione e di correzione di dati elementari", *Atti del Convegno "La qualità dell'informazione statistica e la qualità industriale"*, SIS-ISTAT-AICQ, Roma 10 maggio 1991.

ABBATE C., SCHIEVANO R. (1993) "Efficacia dell'imputazione da donatore con distanza minima", *Atti del Convegno SIS*, Sanremo 1993.

ABBATE C. (1993) "La completezza delle informazioni e l'imputazione da donatore con distanza mista minima: il prodotto RIDA (Ricostruzione delle Informazioni con Donazione Automatica)",*Documento interno Istat*.

ALBONI, F. (1994) "Il controllo e la revisione dei dati" in *L'utilizzazione della rete di contabilità agraria in Emilia-Romagna* (Cap.7), a cura di F.Alvisi e C.Filippucci, Calderini Editore.

BARCAROLI G. (1992) "An integrated system for edit and imputation of data in the Italian Statistical Insitute", *Survey and Statistical Computing*, pp.167-177.

BARCAROLI G. (1993) "Un approccio logico formale al problema del controllo e della correzione dei dati statistici", *Quaderni di ricerca ISTAT* n.9/1993

BARCAROLI G., CECCARELLI C., LUZI O. (1995) "An edit and imputation system of quantitative variables based on macroediting techniques", *Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol (UK), 1-4 Aprile 1995, pp.12-17

BARCAROLI G., CECCARELLI C., LUZI O., MANZARI A., RICCINI E., SILVESTRI F. (1995) "The Methodology of Editing and Imputation by Qualitative Variables implemented in SCIA", *Documento interno ISTAT*.

BARCAROLI G., LUZI O. (1995) "Sistema generalizzato per l'editing e l'imputazione di variabili quantitative (GEIS)", *Quaderni di ricerca ISTAT* n.1/1995 (nuova serie)

COTTON C. (1991) "Functional description of the generalized edit and imputation system", Statistics Canada, Business Survey Methods Division, July 25.

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

ENGSTROM P., ANGSVED C. (1994) "A description of a geographical Macro-editing application", *Statistical Commission and Economic Commission for Europe-Conference of European Statisticians*, Cork, Ireland, 17-20 October 1994.

ESPOSITO R., LIN D., TIDEMANN K.(1993) "The ARIES System in the BLS Current Employment Statistics Program", *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp. 425-429.

FALORSI P.D., FALORSI S. (1995) "Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese", *Rapposrto di ricerca n.13 Progetto CON.PRI*, Università di Bologna

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

FORD B.L. (1983) "An overvew of Hot-deck procedures" in *Incomplete data in sample survey*, vol.1, pg. 191, Academic Press, New York.

GARCIA RUBIO E., VILLAN CRIADO I. (1988) "Sistema DIA, Sistema de deteccion e imputacion automatica de errores para datos cualitativos", Istituto Nacional de Estadistica, Madrid, 1988.

GRANQUIST L. (1992a) "A Review of methods for rationalizing the editing of survey data", in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992.

(1992b) "The Aggregate Method", in *Statistical Data Editing Methods and Techniques*, United Nations, Vol. I, February, 1992.

(1992c) "The Top-Down Method", in *Statistical Data Editing Methods and Techniques*, United Nations, Vol. I, February, 1992.

(1992d) "On the need for generalized numeric and imputation system", in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992.

GRANQUIST L. (1995) "Improving the traditional editing process", in *Business Survey Methods*, John Wiley and sons

GRANQUIST L. (1995) "An overview of methods of evaluating editing processes", *Conference of European Statisticians* (Athens, Greece, 6-9 November), Working Paper n. 3.

GRENLESS J.S., REECE W.S., ZIESCHANG K.D. (1982): "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed", *Journal of the American Statistical Association*, 77, pp 251-261.

HAWKINS D. M. (1974) "The Detection of Errors in Multivariate Data Using Principal Components", *Journal of the American Statistical Association*, Vol. 69. No 346.

HECKMAN, J. (1976) "The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models", *Annals of Economic and Social Measurement*, 5:475-492.

HIDIROGLOU M.A., BERTHELOT J.M.(1986) "Statistical Editing and Imputation for Periodic Business Surveys", Survey Methodology, June 1986, vol.12, N.1, pp.73-83.

JACKSON J.E. (1959) "Quality control methods for several related variables", *Technometrics*, vol. 1, n. 1.

KALTON G. and KASPRZIK D. (1986) "The treatment of missing survey data", in "Survey methodology", 12, 1, Statistics Canada.

KLEJINEN J., VAN GROEDENDAAL W. (1992) - "*Simulation. A Statistical Perspective*", John Wiley, New York.

KOVAR J.G., MacMILLIAN J.H., WHITRIDGE P. (1988) "Overview and strategy for the generalized edit and imputation system", Statistics Canada, Methodology Branch, April 1988(updated February 1991).

KOVAR J.G., WHITRIDGE P. (1995) " Imputation of business survey data", in *Business Survey Methods*, John Wiley and sons

LATOUCHE M., BERTHELOT J.M. (1992) "Use of Score Function to Prioritize and Limit Recontacts in Editing Business Surveys", *Journal of Official Statistics*, Vol.8, No.3, Part II.

LEE H., GHANGURDE P.D., MACH L., YUNG W. (1992) - "Outliers in sample survey", *Statistics Canada Methodology Branch.*

LETI G. (1983)- "*Statistica descrittiva"*, Il Mulino, Bologna.

LINDELL K. (1994) "Evaluation of the editing process of the salary statistics for employees in country councils", *Statistical Commission and Economic Commission for Europe-Conference of European Statisticians*, Cork, Ireland, 17-20 October 1994.

LITTLE, R.J.A. (1986): "Survey nonresponse adjustments for estimates of means, *International Statistical Review*, 54: 139-157.

LITTLE R.J.A. (1993): "Pattern-mixture models for multivariate incomplete data", *Journal of the American Statistical Association*, 88: 125-134.

LITTLE R.J.A., RUBIN D.B. (1987): *Statistical Analysis with Missing Data,* John Wiley & Sons, Inc, New York.

LITTLE R. J. A., SMITH J. (1983) "Multivariate Edit and Imputation for Economic Data", *American Statistical Association, Proceedings of the Survey Research Methods Section.*

LITTLE R. J. A., SMITH J. (1987) "Editing and Imputation for Quantitative Survey Data", *Journal of the American Statistical Association,*Vol. 82, N. 397, Apoplications Section.

LINDSTROM K. (1992) "A macroediting application developed in PC-SAS", in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992.

LUZI O., CECCARELLI C. (1997) "Le componenti principali nello studio dell'editing multivariato", *Atti della XXXV Riunione Scientifica della Società Italiana di Economia, Demografia e Statistica*.

LUZI O. (1996), "Applicabilità ed impatto potenziale dei metodi per l'editing di dati quantitativi basati sugli approcci del Macroediting e dell'Editing Selettivo", *Contributi ISTAT*.

LUZI O. (1998), "L'Editing Selettivo come strumento per la razionalizzazione del processo di editing: un primo studio su Occupazione, Rteribuzioni e Orario di lavoro nelle grandi imprese ", *Quaderni di Ricerca ISTAT*, n.3 1998.

MAGAGNOLI, U. (1978) "L'assicurazione della qualità nelle interrelazioni azienda-mercato-ambiente", *X Convegno nazionale della Associazione Italiana per il Controllo di Qualità*, Torino, Memorie vol.2.

MASSELLI M., SIGNORE M., PANIZON F. (1992) "Il sistema di controllo della qualità dei dati" in *Manuale di tecniche di indagine*, Vol.6 ISTAT.

NORDBOTTEN S.(1995) "Editing statistical records by neural networks", *Conference of European Statisticians*, Athens, Greece, 6-9 November, Working Paper n. 40.

PALLARA A., ABBATE C. (1997) "L'uso dei metodi di partizione ricursiva nel processo di imputazione dei dati delle indagini sulle imprese", *Relazione presentata al Convegno SIS "La statistica per le imprese"*, (Torino 2-4 Aprile).

ROSENBAUM P.R., RUBIN D.B. (1983): "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70: 41-55.

RUBIN D.S. (1975) "Vertex generation and cardinality constrained linear programs", *Operations Research*, vol. 23, pp. 555-565.

RUBIN D.B. (1976): "Inference and missing data", *Biometrika*, 63: 581-592.

RUBIN D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc, New York.

SANTOS R.L. (1981): "Effects of imputation on regression coefficients", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.

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

VAN DE POL F. (1994) "Selective editing in the Netherlands annual construction survey", *Statistical Commission and Economic Commission for Europe-Conference of European Statisticians*, Cork, Ireland, 17-20 October 1994.

VITALE O. (1993) "*Statistica per le scienze applicate*", Cacucci, Bari.

WINKLER W.E. (1994) "SPEER Edit System", computer system and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., USA.

ZANELLA A. (1983) "Sulle procedure di classificazione simultanea", *Statistica*, Anno XXXIII, n.1.