

PRESIDENZA DEL CONSIGLIO DEI MINISTRI

COMMISSIONE PER LA GARANZIA DELL'INFORMAZIONE STATISTICA

**ANALISI DELLE PROCEDURE DI CORREZIONE/IMPUTAZIONE
UTILIZZATE DALL'ISTAT
NELLE PRINCIPALI INDAGINI SULLE FAMIGLIE**

VOLUME I

Rapporto di Ricerca

00.02

Luglio 2000

a cura di:

Luigi Fabbris

Maria Elena Graziani

Cristina Panattoni

La Commissione per la Garanzia dell'Informazione Statistica (CGIS), istituita presso la Presidenza del Consiglio dei Ministri con il decreto legislativo n. 322 del 1989, art.12, è un organo collegiale indipendente chiamato a garantire il principio della imparzialità e della completezza dell'informazione statistica. A tal fine la CGIS vigila: (a) sulla imparzialità e completezza dell'informazione statistica; (b) sulla tutela della riservatezza delle informazioni fornite agli enti del Sistema Statistico Nazionale; (c) sulla qualità delle metodologie statistiche e delle tecniche informatiche impiegate nella raccolta, nella conservazione e nella diffusione dei dati; (d) sulla conformità delle rilevazioni alle direttive degli organismi internazionali e comunitari.

La serie "Rapporti di ricerca" raccoglie i risultati di attività di appositi gruppi di lavoro, promossi e coordinati dalla CGIS in relazione all'adempimento dei propri compiti.

La responsabilità del contenuto del rapporto è degli autori, e non coinvolge la Commissione.

PRESIDENZA DEL CONSIGLIO DEI MINISTRI

COMMISSIONE PER LA GARANZIA DELL'INFORMAZIONE STATISTICA

Ugo Trivellato, *Presidente*

Graziella Caselli

Pierluigi Ciocca

Bruno De Leo

Antonio Golini

Vittorio Grilli

Renato Guarini

Cesare Imbriani

Luisa Torchia

Commissione per la Garanzia dell'Informazione Statistica

Via Po, n.16/A

00198 Roma

tel. ++39-6-8598.2010/8598.2132

fax ++39-6-8598.2012

INDICE

Introduzione	4
1. L'errore nelle risposte	5
1.1 <i>Una classificazione degli errori di rilevazione</i>	8
1.2 <i>L'errore nell'indagine panel</i>	11
1.3 <i>Un modello di analisi dell'errore</i>	12
2. I metodi di rilevazione <i>computer assisted</i>	14
A. <i>CASI</i>	15
B. <i>CATI</i>	18
C. <i>CAPI</i>	19
2.1 <i>I vantaggi dell'intervista on line</i>	21
3. Il controllo di ammissibilità on line delle risposte	25
4. Controlli di coerenza delle risposte in indagini trasversali	28
5. Controlli di verosimiglianza delle risposte in indagini trasversali	29
6. Controlli di coerenza delle risposte in indagini <i>panel</i>	32
7. Controlli di verosimiglianza in indagini <i>panel</i>	33
8. La correzione "a freddo"	35
8.1 <i>La correzione nella fase di registrazione dati</i>	35
8.2 <i>L'editing</i>	35
8.2.1 <i>La correzione deterministica</i>	36
8.2.2 <i>Le procedure probabilistiche di individuazione e imputazione</i>	37
8.3 <i>I softwares per la correzione probabilistica più utilizzati presso l'ISTAT e tecniche implementate</i>	41
9. La possibilità di correggere gli errori <i>on line</i>	43
9.1 <i>Correzione di dati inammissibili in un'indagine trasversale</i>	44
9.2 <i>Correzione di dati inammissibili in un'indagine panel</i>	46
10. Il recupero di informazioni per aggiustare le stime	48
11. Indicazioni propositive	50
Riferimenti bibliografici	52

Introduzione

Il progetto di ricerca relativo alla “*Analisi delle procedure di Correzione/Imputazione utilizzate dall’ISTAT nelle principali Indagini sulle Famiglie*” è indirizzato all’esame delle metodologie adottate dall’ISTAT per il trattamento dei dati mancanti o inammissibili nelle principali indagini sulle famiglie.

In relazione alle suddette finalità, il gruppo di ricerca ha effettuato un accurato studio della letteratura esistente in materia, analizzando gli aspetti teorici relativi al problema del controllo e della correzione dei dati con particolare riguardo alle problematiche innovative del controllo *on line* ed esprimendo valutazioni comparative e suggerimenti all’ISTAT per migliorare le procedure e la documentazione sulle stesse.

Il presente rapporto documenta in sintesi il lavoro di valutazione critica delle informazioni acquisite, fornendo una introduzione ai temi del controllo e della correzione dei dati nonché uno schema di riferimento e una chiave di lettura per le indagini empiriche condotte dal gruppo di ricerca presso i Servizi dell’ISTAT.

In linea con gli obiettivi prefissati, nel rapporto contenente la seconda parte dello studio verranno presentate nel dettaglio le tre indagini campionarie condotte dall’ISTAT sulle famiglie (*Indagine Multiscopo, Indagine sulle Forze di Lavoro, Indagine sui Consumi delle Famiglie*), specificando, per ciascuna di esse, la logica dei controlli e delle correzioni effettuate sui dati errati o mancanti.

Verranno, inoltre, offerti spunti propositivi per il miglioramento del processo di controllo della qualità dei dati raccolti, sia in un’ottica trasversale che longitudinale.

1. L'errore nelle risposte

In questo primo volume del rapporto¹ presentato alla CGIS si tratta il problema del controllo della qualità dei dati ottenuti in indagini svolte mediante intervista sulla popolazione e delle possibilità date dalla metodologia statistica e informatica per correggere i dati affetti da errore durante la rilevazione dei dati stessi.

Il percorrere nuovamente i sentieri definitivi, metodologici e progettuali per la qualità del dato mira non solo a fare il punto e a sistematizzare ciò che già si sa, ma anche a creare i presupposti per una ridefinizione dei termini, della metodologia delle indagini statistiche nella versione tecnologicamente avanzata, delle tecniche (psicologiche, informatiche) di gestione dell'intervista finalizzate ad incrementare la qualità dei dati (O'Muircheartaigh, 1997).

Il ragionamento che stiamo per sviluppare poggia sui seguenti presupposti:

1. La qualità dei dati è argomento basilare in ogni indagine statistica diretta². Non esiste indagine, o dato rilevato in un'indagine, esente dal rischio di errore nei dati. Lasciando sullo sfondo gli errori di impostazione dell'indagine e del campione, nel seguito, si tratta il problema dell'errore statistico con in mente la rilevazione dei dati *computer assisted*, assistita da computer³. Le soluzioni specificamente inerenti a questo criterio di rilevazione si possono assumere a paradigmi metodologici per le rilevazioni basate su questionario.
2. La letteratura sull'argomento è vasta sia in relazione alla metodologia per l'identificazione degli errori e il trattamento dei dati rilevati, sia rispetto all'ampiezza del-

¹ Il presente rapporto è attribuibile a Luigi Fabbris tranne i paragrafi 2.1, 8.2, 8.2.1, 8.2.2, attribuibili a Maria Elena Graziani e i paragrafi 1.2, 8.1 e 8.3 attribuibili a Cristina Panattoni che ha redatto, inoltre, l'introduzione al rapporto e curato l'integrazione dei diversi elaborati.

² Per rilevazione diretta s'intende il metodo di reperimento dei dati statistici chiedendo, mediante un questionario, le informazioni direttamente alla persona che ne è depositaria. La rilevazione diretta si può svolgere mediante autocompilazione del questionario o con l'intermediazione di intervistatori.

³ La rilevazione assistita da computer è la gestione completamente automatizzata della composizione e somministrazione del questionario, della registrazione delle risposte e del controllo di qualità delle stesse. Il questionario è un programma informatico che, nella forma normale, si propone sullo schermo del computer e scorre via via che si registrano, in tempo reale, le risposte su apposite basi di dati. Con termini suggestivi, si denomina questo tipo di rilevazione *questionario elettronico*, o *informatizzato* (Fabbris, 2000).

le applicazioni. Nella letteratura citata continua, tuttavia, a vigere un sistema classificatorio degli errori mirato all'identificazione e alla misura dell'effetto degli errori dopo il completamento della raccolta dei dati. La collocazione del controllo e dell'eventuale correzione *on line* comporta un ripensamento metodologico e tecnico. Il sistema classificatorio di riferimento per le nostre analisi è esposto nei Paragrafi 1.1, 1.2 e 1.3.

3. I sistemi di rilevazione *computer assisted* rappresentano i metodi più vincolanti e, ad un tempo, più promettenti per la rilevazione di dati statistici. Sono vincolanti perché i passaggi logici per l'impostazione e l'informatizzazione del questionario e la gestione dell'intervista richiedono capacità metodologiche e tecniche specifiche. Sono più promettenti, da una parte, perché le opportunità di memoria e di calcolo offerte dai sistemi informatici permettono di accedere in tempo reale a risorse utilizzate nel passato solo "a freddo" (Paragrafo 8), alla fine del processo d'indagine e, d'altra parte, perché il rigore procedurale nell'impostazione e nella gestione dell'indagine imporrà la ricerca di una più fine metodologia. I metodi *computer assisted* permettono e richiedono un avanzamento metodologico e lo spostamento dei controlli dalla fase di trattamento pre-analisi al controllo in linea⁴, ossia realizzato mentre si sta dialogando con il rispondente. Le rilevazioni *computer assisted* sfruttano anche la tecnologia per la trasmissione a distanza dei quesiti e dei dati. Un sintetico excursus terminologico dei metodi di rilevazione *computer assisted* è oggetto del Paragrafo 2.

Rispetto alla rilevazione basata sul questionario cartaceo, la rilevazione *computer assisted* dà la possibilità di effettuare controlli sull'ammissibilità della risposta prima che sia posta la successiva domanda (Baker, Bradburn e Johnson, 1995). L'eventuale incompatibilità della risposta rispetto ad altri dati registrati nel sistema permette – nei limiti del ragionevole – di ottenere informazioni aggiuntive per risolvere l'incompatibilità subito o in un successivo momento. Il beneficio rispetto al questionario cartaceo, in cui il controllo può

⁴ Si dice in linea (*on line*) il controllo svolto sulle risposte quando sono in esercizio strumenti automatici di rilevazione dei dati. Se il controllo riguarda la risposta ottenuta ad un quesito, esso è effettuato prima della somministrazione del quesito successivo. Questa procedura, *mutatis mutandis*, si applica sia alla rilevazione di dati con sistemi mediati da rilevatori, sia a sistemi basati sull'autosomministrazione del questionario, nonché a sistemi che effettuano il reperimento dei dati da fonti non umane.

tutt'al più essere svolto dopo la registrazione dei dati, senza una reale possibilità di confronto con il rispondente, è notevole (Paragrafo 2.1).

Una delle possibilità offerte dal questionario elettronico è la possibilità di controllare non solo le eventuali incoerenze di una risposta ottenuta con regole o con informazioni prefissate, ma anche la sua verosimiglianza, ossia la probabilità che sia errata con riferimento ad una ipotizzata distribuzione statistica delle risposte⁵.

I fondamenti delle metodiche statistiche e informatiche per la valutazione dell'ammissibilità *on line* delle risposte sono discusse nel Paragrafo 3. Nel presentare le metodiche si distinguono quelle fondate su criteri deterministici di identificazione dell'errore da quelle che derivano da criteri probabilistici. I criteri di identificazione deterministici sono regole di coerenza di una risposta con dati o con metadati inseriti nella memoria del computer. Per questo sono spesso denominati "criteri di coerenza" o "di compatibilità". Si presentano nei Paragrafi 4 e 6 distinguendo i metodi basati sulla rilevazione trasversale dei dati da quelli raccolti in indagini ripetute nel tempo su campioni fissi di soggetti (*panel*).

I criteri stocastici sono regole per determinare la compatibilità di una risposta rispetto alla distribuzione statistica ipotizzata delle risposte della categoria di rispondenti cui appartiene la persona che sta rispondendo. Se la risposta è improbabile, viene sostanzialmente trattata come se fosse una risposta incompatibile, con alcuni distinguo. I criteri di valutazione della verosimiglianza delle risposte sono presentati nei Paragrafi 5 e 7, anche per questi distinguendo il caso della rilevazione trasversale da quella longitudinale basata su *panel*.

Quando, sulla base di un criterio tra quelli introdotti, si determini la presenza di un certo errore nella risposta appena ottenuta, si dovrà procedere alla rettifica della risposta o alla rettifica dei dati "di sistema" con cui questa è stata confrontata. La rettifica sarà diversa secondo i casi: per una mancata risposta, o per una risposta elusiva, si dovrà prima "girare al largo" per capire se e perché il quesito origina ritrosia nella persona interpellata e poi cercare un barlume d'informazione che permetta di arrivare nelle prossimità del dato "vero". Per una risposta incoerente con i dati di sistema, o inverosimile rispetto a standard distributivi ipotizzati, si dovrà accertare l'origine dell'errore e conciliare le situazioni messe a

⁵ Senza forzatura, il questionario che contenga un insieme coerente di regole per il controllo della qua-

confronto. Di ciò si parla nel Paragrafo 9.

Purtroppo, quantunque sulla base di un criterio di ammissibilità sia determinata l'inadeguatezza della risposta ottenuta, non è sempre possibile ottenere informazioni suppletive per eliminare o comprimere l'errore all'istante. Conviene allora recuperare informazioni a latere per correggere le stime. Infatti, i questionari in forma elettronica danno anche l'opportunità di semplificare il questionario "apparente", ossia visibile a chi lo somministra, per cogliere sperimentalmente - somministrando questionari alternativi a sottinsiemi casualmente determinati di rispondenti, senza che questi siano di ciò consapevoli - le differenze tra modi diversi di esprimere lo stesso concetto, tra sequenze diverse delle stesse domande, o delle stesse modalità di risposta offerte, e così via, che sono all'origine degli "errori di risposta" (Willemborg, 1987). Nel Paragrafo 10 si presentano alcune metodiche di somministrazione del questionario che possono annullare o ridurre al minimo, e nella peggiore delle ipotesi solo misurare, l'entità degli errori di rilevazione.

1.1 *Una classificazione degli errori di rilevazione*

In un'indagine statistica svolta con una struttura complessa di rilevazione, gli errori di rilevazione⁶ si possono distinguere *secondo la fonte* in:

1. ***Errori indotti dal rispondente.*** Questi possono essere originati da difetti nel ricordare eventi o situazioni, da compiacenza verso chi conduce l'indagine, da modifica della realtà per timore che la risposta non sia socialmente accettabile, da deliberata volontà di dare risposte errate.
2. ***Errori indotti dal rilevatore sul rispondente.*** Gli errori conseguono al modo di porre le domande, al modo di accettare le risposte ottenute, al modo di registrarle. Soprattutto il dialogo con il rispondente, il *feedback* emotivo dopo l'ottenimento della risposta e, quando il rispondente è visibile al rispondente, il suo modo di apparire sono causa di condizionamento delle risposte che può dare origine a serie distorsioni nei dati.

lità in linea delle risposte si può denominare "intelligente".

⁶ In questa nota si lasciano sullo sfondo le fonti di variabilità "endogena" dei fenomeni su cui s'indaga e si approfondiscono solo le fonti di variabilità dovute ad errori "esogeni", determinati dal sistema di rilevazione.

3. **Errori indotti dai supervisori sui rilevatori.** I supervisori sono le persone che effettuano azioni di controllo ed indirizzo dell'attività dei rilevatori. Hanno una grande importanza perché risolvono in modo autocratico le incertezze definitorie e procedurali della rilevazione. Possono generare errori anche di notevole entità se indirizzano soggettivamente l'operato dei rilevatori in una direzione diversa da quella intesa dal ricercatore.

Secondo le caratteristiche distributive, gli errori si distinguono in variabili e sistematici:

- sono *variabili*, o *casuali*, gli errori che si manifestano in entità e segno variabili da individuo a individuo. Tra gli errori variabili di rilevazione sono particolarmente temuti i cosiddetti "errori correlati", ossia quegli errori di risposta che hanno valori simili presso sottinsiemi di unità statistiche. Sono *casuali semplici* gli errori dei rispondenti in indagini basate sull'autocompilazione del questionario (tra le altre, nelle indagini postali), sono *casuali correlati* gli errori di risposta nelle indagini con rilevatore;
- sono *sistematici* gli errori costanti in ogni risposta della associata domanda *Y*. Gli errori sistematici si dicono anche *distorsioni*.

Secondo il modo in cui si manifestano, gli errori di rilevazione sono:

- *mancate risposte* quando non si ottiene la collaborazione dell'unità designata a rispondere. La mancata risposta può riguardare un singolo quesito o l'intero questionario. La mancata collaborazione all'intero questionario si dice anche mancata risposta totale, o mancata intervista, quella ad un singolo quesito si dice mancata risposta parziale, o semplicemente mancata risposta quando è implicito il significato. La mancata risposta è il fallimento totale nella richiesta di informazioni. La mancanza di risposta è, infatti, causata dal rifiuto di collaborare a causa di un sentimento negativo originato dal quesito, dai contenuti o dai criteri di gestione dell'indagine. La mancata risposta spesso provoca distorsioni nelle stime;
- *risposte elusive*, quando si ottengono risposte che mirano a schivare la richiesta dell'informazione cercata. Le tipiche risposte elusive sono "Non so, non ricordo", "Non sono in grado di dare un giudizio", "Altra risposta", "La via di mezzo" (in relazione ad

una valutazione), ecc. Le risposte elusive sono sostanzialmente delle mancate risposte e provocano, quindi, distorsioni nelle stime. Nei dati incompatibili prevalgono gli errori casuali, ma possono anche essere presenti delle distorsioni;

- *dato palesemente errato, o incompatibile*, quando è certo che la modalità rilevata è errata. Che una risposta sia manifestamente errata si capisce quasi esclusivamente confrontandola con altri dati o metadati noti a chi svolge la rilevazione. Un dato palesemente errato non è necessariamente un dato irrecuperabile, soprattutto nelle rilevazioni ripetute nel tempo e quando esiste un'ampia base informativa sul rispondente. Nei dati inverosimili prevalgono gli errori casuali;
- *dato probabilmente errato, o inverosimile*, quando esistono motivi per sospettare che la risposta sia affetta da un livello d'errore così ampio da considerarla un'eccezione rispetto alla norma. Un valore è probabilmente errato quando è esagerato per la categoria cui appartiene il rispondente. Si possono comunque avere dati inverosimili anche con riferimento a quesiti dicotomici, su scala ordinale, o politomici (nominali).

Si può osservare che le quattro categorie d'errore rappresentano altrettanti livelli di conoscenza sul dato. La risposta elusiva è un'offerta informativa un po' superiore alla mancata risposta, almeno sul piano della disponibilità a collaborare. Il dato probabilmente errato deve aver superato il controllo di compatibilità, ossia tutti i controlli formali. Il controllo della verosimiglianza di una risposta è il livello più fine di controllo tra quelli ipotizzati.

1.2 L'errore nell'indagine panel

Se l'indagine statistica prevede la reintervista in tempi successivi degli stessi individui secondo un disegno campionario di tipo *panel*, le possibilità di errore descritte nel precedente paragrafo si ampliano.

Nella categoria delle mancate risposte, oltre alle mancate risposte totali e parziali, si deve aggiungere il fenomeno delle mancate risposte *panel* (*wave nonresponse*; Duncan e Kalton, 1987) che si verifica quando il rispondente partecipa solo ad alcune delle occasioni di indagine previste.

Nel caso delle mancate risposte *panel* si possono verificare tre situazioni differenti (Ghellini e Pannuzi, 1996):

1. l'intervistato non risponde alla prima onda ma partecipa a tutte quelle successive (*initial nonresponse*);
2. l'intervistato partecipa saltuariamente alle varie occasioni di indagine;
3. l'intervistato, dopo aver risposto alle prime occasioni di indagine, esce dal *panel* senza più rientrarvi (fenomeno del logorio o *attrition*).

Per il trattamento di questi casi si può procedere nei seguenti modi:

1. per le mancate risposte iniziali si può optare per due differenti soluzioni:
 - l'unità viene esclusa dall'analisi longitudinale;
 - vengono ricostruite le informazioni relative alla prima occasione, ma solo per le principali variabili strutturali e di contesto;
2. le mancate risposte saltuarie possono essere considerate come mancate risposte parziali all'interno di un record longitudinale e le occasioni mancanti possono essere ricostruite tramite le informazioni acquisite nelle *waves* adiacenti;
3. il logorio può essere assimilato a una mancata risposta totale "ritardata" nel tempo e trattato secondo le tecniche del caso (ponderazione).

Inoltre, nel corso di una indagine *panel* si possono verificare altri tipi di errore non campionario riconducibili alle classificazioni precedentemente esposte:

Errori indotti dal rispondente

- condizionamento da *panel* (*panel conditioning*). Il rispondente è condizionato dal fatto di aver risposto in tempi precedenti alle stesse domande e fornisce risposte che siano in linea con quelle passate. Ridurre l'onere statistico sulle unità di rilevazione diventa, allora, un obiettivo cruciale, evitando duplicazioni e richieste ridondanti nelle occasioni di indagine successive alla prima;

Errori indotti dal rilevatore sul rispondente

- condizionamento da intervistatore. Il fatto che le interviste successive alla prima siano condotte dallo stesso intervistatore o da un altro intervistatore può indurre dei cambiamenti nel modo di rispondere;

Errori indotti dalle strategie di rilevazione

- cambiamento delle modalità e/o degli strumenti di indagine da un'onda all'altra per cui possono verificarsi incoerenze temporali nelle risposte;
- cambiamenti nelle codifiche da un'onda all'altra;
- errati abbinamenti dei record individuali (*linkage*) al fine di ricostruire l'informazione longitudinale relativa a ciascuna unità.

Il mancato abbinamento dei record individuali può derivare da:

- sostituzione delle unità di campionamento con unità estratte da liste aggiuntive;
- cambiamenti nei codici identificativi degli individui da un'onda all'altra;
- errata trascrizione dei codici identificativi degli individui;
- logorio, per cui una unità si perde nel corso delle varie occasioni di indagine.

Infine, l'errore di copertura - che si manifesta nell'impossibilità di contattare le stesse unità in tempi diversi o, comunque, di seguirle nel tempo - anche se afferisce più strettamente alla sfera degli errori campionari, pone problemi relativi sia all'adozione di definizioni dinamiche di popolazione sia alla specificazione di regole e/o di tecniche per mantenere il contatto e per seguire le unità *panel* mobili.

1.3 Un modello di analisi dell'errore

Si consideri il valore osservato in una indagine diretta alla quale partecipino rilevatori e

supervisor. L'errore di rilevazione si può scomporre secondo il seguente modello causale:

$$y_{jis} = \mathbf{m}_j + \mathbf{n}_j + \mathbf{a}_i + \mathbf{b}_s + \mathbf{g}_p, \quad (j=1, \dots, n; i=1, \dots, k; s=1, \dots, h) \quad (1)$$

dove:

y_{jis} è il valore rilevato presso il rispondente j ($j=1, \dots, n$), interpellato dal rilevatore i ($i=1, \dots, k$) controllato dal supervisore s ($s=1, \dots, h$)

\mathbf{m}_j è il "valore vero"⁷ del rispondente j per la variabile Y

\mathbf{n}_j è il termine d'errore attribuibile al rispondente j , senza tener conto dei possibili motivi (difetto di memoria, "filtro sociale" della risposta, volontà di rispondere in modo errato)

\mathbf{a}_i denota il termine d'errore del rilevatore i , anche questo come somma delle possibili cause d'errore attribuibili alla soggettività dell'esercizio del proprio ruolo da parte del rilevatore

\mathbf{b}_s denota il termine d'errore del supervisore s

\mathbf{g}_p è la cosiddetta distorsione del ricercatore, ossia il modo di impostare la ricerca, di fare le domande e di predisporre il sistema ad accettare risposte. Questa distorsione è sempre presente in una indagine diretta.

Il modello di scomposizione degli errori si semplifica se il disegno di rilevazione è meno complesso. Se non sono utilizzati supervisor, manca dall'equazione (1) il termine d'errore attribuibile ai supervisor, se la rilevazione si svolge senza l'intermediazione di rilevatori, rimangono solo i termini d'errore del rispondente e del ricercatore che sono sempre presenti in una rilevazione diretta. Per quanto riguarda la possibilità di stimare l'effetto sulle stime degli errori suddetti, si può affermare che (Bassi e Fabbris, 1994):

- L'effetto dell'errore del rispondente sulle stime si può stimare sulla base della variabilità della variabilità di y_{jis} attorno al suo valore atteso⁸. Il metodo basilare per la stima

⁷ Si denomina *valore vero* la modalità della variabile oggetto d'interesse per un dato rispondente. Più precisamente, si dovrebbe parlare di *modalità vera*, dato che valore è solo una modalità quantitativa. D'altronde, il modello di misura degli errori si applica nella forma presentata in questa nota a variabili quantitative o dicotomiche.

⁸ Il valore atteso di y_{jis} è \mathbf{m}_j solo se le osservazioni non sono uniformemente distorte. In genere, il

dell'effetto è la ripetizione della rilevazione (reintervista, o PES – *Post Enumeration Survey*) su un campione statistico di unità interpellate nella rilevazione principale. La logica della ripetizione della rilevazione mira esclusivamente a determinare l'attendibilità delle stime ed, eventualmente, a rettificare le stime di cui sia stata constatata la distorsione. La possibilità di rettificare i dati elementari è data solo dalla conciliazione tra le risposte date nell'indagine principale e in quella di controllo. La conciliazione è il processo di convergenza delle risposte date nelle due indagini con la mediazione del rilevatore⁹.

- L'effetto degli errori del rilevatore e del supervisore si possono stimare con il metodo della compenetrazione delle assegnazioni (Mahalanobis, 1946). Applicato ai rilevatori, il metodo consiste nel suddividere l'insieme da esaminare in tanti sub-campioni quanti sono i rilevatori e nell'assegnare i sub-campioni casualmente ai rilevatori. Questo metodo mira esclusivamente a determinare l'attendibilità delle stime (Fabbris, 1983). E', però, possibile misurare la propensione all'errore per singolo rilevatore ("distorsione del rilevatore").

2. I metodi di rilevazione computer assisted

L'acronimo CASIC (*Computer Assisted Survey Information Collection*) è il termine che indica l'insieme delle nuove tecniche di indagine per condurre rilevazioni statistiche assistite da computer e *software* specializzato¹⁰. Le principali tecniche di indagine *computer-assisted* sono (Figura 1):

- CATI (*Computer Assisted Telephone Interviewing*) per l'intervista telefonica;
- CAPI (*Computer Assisted Personal Interviewing*) per l'intervista faccia a faccia;

valore atteso di y_{jis} è Y_j e lo scarto tra Y_j e m_j è la distorsione uguale per tutte le unità statistiche.

⁹ In teoria, esiste la possibilità di conciliare coppie di risposte inerenti alla stessa unità statistica anche nelle rilevazioni con questionario autosomministrato. La tecnica della conciliazione è, tuttavia, scarsamente praticabile.

¹⁰ CASIC fa concorrenza ad altri termini nati indipendentemente, come CAI (*Computer Assisted Interviewing*) e CAQI (*Computer Assisted Questionnaire Interviewing*).

- CASI (*Computer Assisted Self-administered Interviewing*) per l'autocompilazione del questionario direttamente dal rispondente.

A. CASI

Nella categoria CASI rientrano tutte le tecniche che non prevedono il ricorso ad intervistatori, qualunque sia il mezzo utilizzato per raggiungere il rispondente (Capiluppi, 2000):

Figura 1: *Tecniche di indagine computer-assisted*

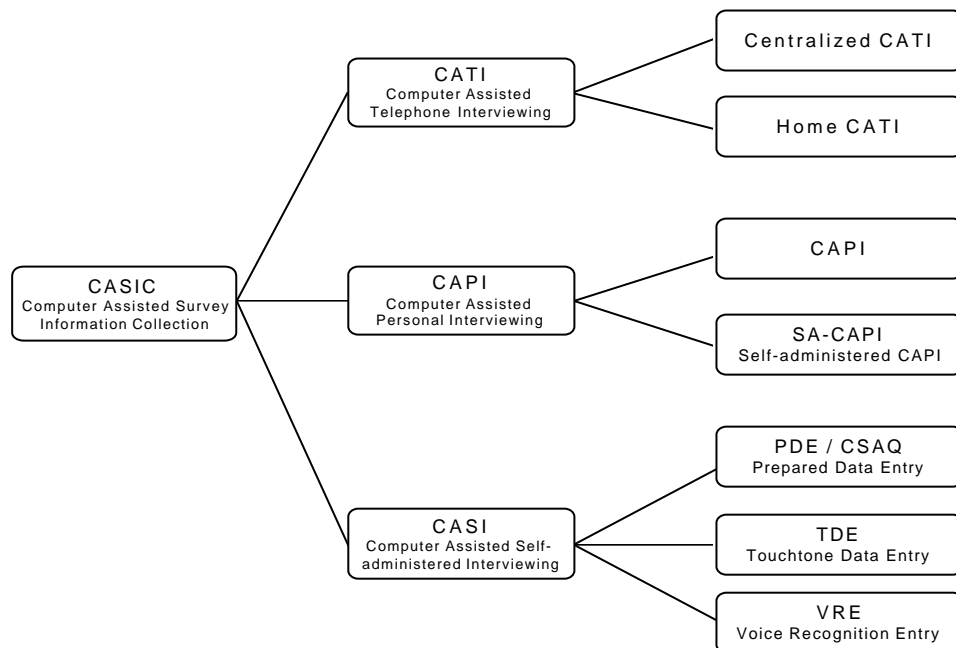
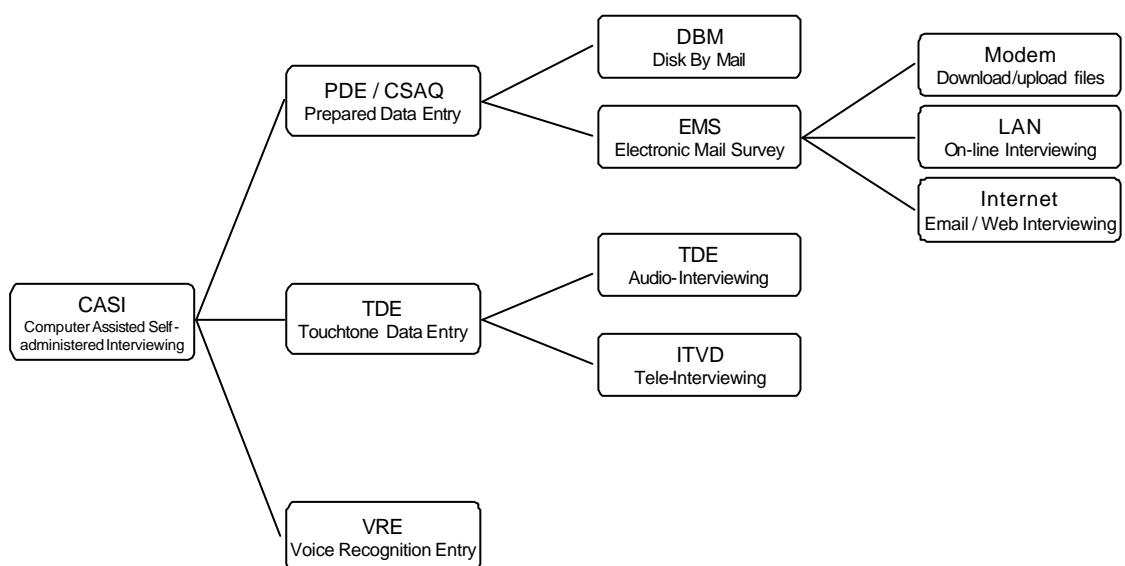


Figura 2: *Tecniche di indagine CASI*



- PDE (*Prepared Data Entry*): il rispondente utilizza personalmente il questionario elettronico sul proprio PC. Un tale tipo di indagine ha due requisiti fondamentali: la disponibilità di un PC da parte dell'intervistato (nel caso che ne sia sprovvisto l'organizzatore dell'indagine dovrà renderglielo disponibile ed eventualmente provvedere alla sua manutenzione) e una seppur limitata capacità di quest'ultimo nell'utilizzarlo¹¹ (McDonald, 1996; Saris e De Pijper, 1986; Nicholls *et al.*, 1997).
- TDE (*Touch-tone Data Entry*): il rispondente telefona a un numero (verde) gestito da un computer dedicato che lo “intervista” e risponde alle domande mediante la tastiera del telefono¹² (Clayton e Harrel, 1989; McKay e Robinson, 1994);
- VRE (*Voice Recognition Entry*): è simile alla precedente nell'attivazione del processo di somministrazione dei quesiti, ma le risposte sono fornite a voce e registrate mediante un sistema di riconoscimento vocale¹³. Le domande vengono poi ripetute all'intervistato che le conferma con un sì o un no (Appel e Cole, 1994; Blyth e Piper 1994; Blyth, 1997).

La tecnica PDE, detta anche CSAQ (*Computerised Self-Administered Questionnaire*), prevede diverse opzioni operative, in funzione della modalità di invio del programma e di raccolta dei file contenenti le risposte:

- DBM (*Disk By Mail*): i rispondenti ricevono il programma su *floppy disk* tramite posta, e sempre per posta rispediscono il dischetto con il file delle risposte;
- EMS (*Electronic Mail Survey*): il questionario informatizzato e/o il file delle risposte sono inviati mediante un mezzo telematico:
 - ❖ *via modem*: il programma viene scaricato (*download*) tramite una connessione telefonica diretta con un computer dell'ente che esegue l'indagine e, ad intervista ultimata, il file delle risposte viene trasferito per la stessa via;

¹¹ Il PDE offre molte possibilità: se il PC viene collegato ad un lettore ottico è possibile leggere codici a barre e, in prospettiva, anche testi. Poiché il PDE richiede la disponibilità di un computer e di utilizzatori con esperienza, è stato prevalentemente impiegato in indagini presso le imprese (Weeks, 1992).

¹² Per poter applicare il Televideo interattivo il rispondente deve disporre di un apparecchio telefonico multifrequenza, che genera toni decodificabili da parte del software TDE.

¹³ Il software VRE riconosce solo cifre e parole elementari.

- ❖ *via LAN*: i rispondenti hanno accesso al programma tramite una rete locale; le risposte possono essere registrate direttamente su un server centrale;
- ❖ *via Internet*: i rispondenti hanno accesso al questionario tramite Internet; si possono avere in questo caso diverse forme di utilizzo della rete: da semplice canale per la trasmissione di *files* a supporto per condurre interviste *online*.

Internet è tuttora poco diffuso nell'ambito delle famiglie italiane, ma si tratta di un canale di comunicazione interessante in una prospettiva temporale di medio termine. *Internet* può essere utilizzato come un canale per inoltrare il questionario, via posta elettronica o via FTP, facilitato rispetto alla tradizionale connessione via modem. Le possibilità offerte da Internet vanno però ben oltre questo tipo di utilizzo, e le nuove tecnologie di programmazione per sviluppare applicazioni su *Internet* permettono di concepire la realizzazione di vere e proprie interviste *on-line* su *Web*, con interfacce grafiche di facile utilizzo e controlli di coerenza in linea che assistono la compilazione del questionario come in un vero e proprio programma PDE (Hardie e Neou, 1994; Quarterman, 1994; Pitkow, 1995).

La tecnica TDE può essere realizzata mediante il supporto del Televideo televisivo, dando origine al cosiddetto Televideo Interattivo (ITVD). In pratica, grazie al supporto di un *provider* Televideo, una pagina TVD "dinamica" viene utilizzata per visualizzare, una alla volta, le domande del questionario con le relative risposte, in alternativa alla somministrazione audio telefonica tipica di TDE. Il video dell'apparecchio TV viene quindi utilizzato come terminale del computer remoto che gestisce l'intervista, mentre la tastiera telefonica continua ad essere utilizzata per digitare le risposte (Lindström, 1995)

Il questionario può essere definito una volta per tutte ma è anche possibile, grazie al collegamento telematico del computer con l'organizzatore dell'indagine, che il questionario venga inviato di volta in volta e che la stessa unità statistica sia chiamata a collaborare a più indagini (Jenkins e Dillman, 1997)

Anche la registrazione dell'informazione può subire cambiamenti di rilievo. Il computer può essere collegato ad uno *scanner*, oppure le domande potrebbero essere trasmesse col segnale televisivo, proposte tramite il televisore e le risposte fornite mediante un tele-

comando. Se poi il questionario è trasmesso via Internet appare possibile anche un'interazione tra intervistato ed intervistatore (Fisher *et al.*, 1994; Shing e Chu, 1996).

B. CATI

Il metodo CATI è basato su un sistema di intervista interattiva mediata dall'uso del telefono. Il computer, grazie ad un apposito *software*, effettua le chiamate, il rilevatore legge le domande, che appaiono sullo schermo, e le registra direttamente nella memoria elettronica (House, 1985; Chiaro, 1996).

E' stato il primo metodo proposto per la rilevazione automatica di dati statistici (Nicholls, 1978; Miller e Cannell, 1982; Nicholls e Groves, 1986; House e Nicholls, 1988). Attualmente, la rilevazione CATI è il modo comune di rilevare dati mediante intervista telefonica: nelle rilevazioni di tipo commerciale, nelle quali la rapidità d'esecuzione è connotata con gli obiettivi della ricerca e si desiderano stime per grandi domini di studio, l'impiego di sistemi CATI è il metodo più comune in assoluto per rilevare dati; negli U.S.A., le rilevazioni con sistemi CATI rappresentano oltre l'85% del totale delle indagini svolte sulla popolazione e sulle imprese.

Le modalità organizzative della rilevazione CATI sono due: (i) un sistema centralizzato, formato da un gran numero di postazioni interconnesse e sottoponibili a controllo e supervisione¹⁴, (ii) un sistema decentrato, composto da un certo numero di postazioni atomistiche collegate via modem con il centro (Shanks e Tortora, 1985; Tortora, 1985; Chiaro, 1996; Batcher e Scheuren, 1997).

I sistemi CATI permettono una più efficiente gestione del lavoro sul campo tramite la programmazione (*scheduling*) delle telefonate e dei nuovi tentativi di chiamata quando il primo è andato a vuoto. La programmazione delle chiamate ha raggiunto alti livelli di sofisticazione, definendo algoritmi che comprendono il giorno e l'ora più opportuni per il contatto, nonché l'ordinamento delle chiamate per orario, in funzione della probabilità di contattare l'unità designata in un dato istante e di ottenere risposta dalle unità contattate (Gro-

¹⁴ Il controllo del supervisore si attua nel collegamento con le postazioni dei rilevatori e nel seguire le interviste senza che i rilevatori siano consapevoli di essere controllati. In alcuni Paesi, tra cui l'Italia, il controllo non può avvenire senza che l'operatore ne sia consapevole.

ves e Kahn, 1979; Kerssemakers *et al.*, 1987; Kulka e Weeks, 1988; Fabbris e Martini, 1998; 1999; Pratesi, 2000).

Un altro risultato delle rilevazioni CATI è la mitigazione dell'effetto intervistatore, ossia delle distorsioni generate dalla soggettività che i singoli rilevatori pongono nello svolgimento del proprio compito e che si traduce in una perdita di stabilità delle stime aggiuntiva rispetto a quella di tipo campionario e agli errori di memoria, agli sbagli che commettono inevitabilmente coloro che collaborano ad un'intervista¹⁵.

Lo stato del SISTAN e i vincoli del lavoro nella pubblica amministrazione deporrebbero a favore dell'appalto esterno delle rilevazioni CATI. La pianificazione delle rilevazioni correnti e di quelle di supporto alle indagini strutturali, essendo inserita nel Programma statistico nazionale, permette di modulare le esigenze anche nei tempi lunghi e giocare su ambedue i piani, in pratica, organizzando un sistema CATI centralizzato e ricorrendo ad appalti per le attività eccezionali (Fabbris, 1999).

C. CAPI

I sistemi CAPI consistono nella rilevazione faccia a faccia mediante computer di minime dimensioni¹⁶ utilizzati dal rilevatore per leggervi le domande e memorizzarvi all'istante le risposte¹⁷. E' un sistema completamente decentrato, essendo i computer consegnati ai rilevatori, talvolta per svolgervi una pluralità d'indagini¹⁸.

¹⁵ Il lettore interessato a studi sull'errore dell'intervistatore può consultare, tra gli altri, i lavori di Hanson e Marks (1958), Kish (1962), Fellegi (1964, 1974), Bailar *et al.* (1977), Tucker (1983), Groves e Magilavy (1986), Pannekoek (1987), Bassi e Fabbris (1994), Allard *et al.* (1996)

¹⁶ Le dimensioni dei computer utilizzati per le rilevazioni CAPI hanno un ingombro minimo. Nella rilevazione di dati ufficiali, con l'intervistatore di fronte all'intervistato, lo strumento di rilevazione è tipicamente un computer portatile (*notebook*) a batteria e si usa su un tavolo. Nella rilevazione svolta in piedi (per esempio, per strada, su un mezzo di trasporto, ecc.), la dimensione del computer è minima, tanto da poter stare sul palmo della mano del rilevatore ("computer palmare", cfr. Brakenhoff *et al.*, 1987).

¹⁷ Si possono applicare anche sistemi di rilevazione *Audio-CAPI*, nei quali l'intervistatore fornisce una cuffia all'intervistato che ascolta le domande da un sintetizzatore vocale e risponde utilizzando una tastiera o un altro mezzo di registrazione informatica dei dati (Filippucci, 2000).

¹⁸ Un particolare sviluppo del CAPI è il cosiddetto "*audio CAPI*". In questo caso l'intervistatore fornisce una cuffia all'intervistato che ascolta le domande da un sintetizzatore vocale e risponde usando una tastiera o un altro mezzo per la registrazione informatica dei dati. Dopo che l'ultima risposta è stata fornita, l'intervistato può spegnere il computer.

I metodi CAPI hanno cominciato a diffondersi verso la fine degli anni Ottanta quando lo sviluppo dei computer portatili ha reso economicamente e praticamente pensabile la sostituzione del questionario tradizionale con quello elettronico installato su di un computer¹⁹. Il CAPI ha proprie caratteristiche che lo distinguono dagli altri metodi CASIC: i problemi principali da considerare in questo caso riguardano: l'accettazione dello strumento da parte dell'intervistato e dell'intervistatore, la realizzazione di un *software* adeguato e la formazione dell'intervistatore²⁰.

Un sistema CAPI richiede l'installazione di un modem presso i gangli periferici dai quali i rilevatori possono inviare al centro periodicamente le informazioni raccolte (Metz, 1987). Non è necessario che i calcolatori impiegati nella raccolta dei dati siano "intelligenti", è sufficiente che abbiano memorie RAM e di massa e capacità di elaborazione mirate. I centri dai quali sono incanalate le informazioni raccolte sono spesso le abitazioni dei rilevatori.

I sistemi CAPI sono stati introdotti primariamente per reperire dati sulle forze di lavoro e sulle opinioni della gente in vari Paesi europei e americani²¹. Il riscontro è stato positivo, sia rispetto all'obiettivo di ridurre le mancate risposte (Van Bastelaer e Sikkel, 1987), sia riguardo all'interazione con i rispondenti (Van Bastelaer *et al.*, 1987; Couper *et al.*, 1997; Santi *et al.*, 1997), sia in funzione della qualità dei dati (Benelmans-Spork e Sikkel, 1985).

L'ISTAT ha svolto rilevazioni sulle forze di lavoro, con computer portatili forniti dall'EUROSTAT, e sui bilanci delle famiglie italiane. Le esperienze hanno dato risultati complessivamente positivi (Filippucci *et al.*, 2000), anche se l'impiego episodico dei com-

¹⁹ Ricordiamo alcuni lavori che testimoniano del vasto arco di indagini che sono interessate dalle tecniche CAPI: Lyberg (1985), Van Bastelaer e Sikkel (1987), Foxon (1987), Bernard (1989), Keller *et al* (1990), Weeks (1992), Martin *et al* (1993), Matchett *et al* (1994), Baker *et al.* (1995), Couper (1996). Fra le esperienze si segnala quella italiana condotta dall'ISTAT nel 1996 relativamente all'indagine sui consumi (Filippucci, Drudi e Ferrante, 2000).

²⁰ I campi di ricerca in cui queste tecniche vengono utilizzate è ormai vasto: dalle statistiche del mercato del lavoro (US Bureau of Labor Statistics, *Current Employment Statistics*), a quelle relative ai consumi (*US Consumer expenditure panel*), alle ricerche di mercato, fino ai censimenti (censimento USA del 2000).

²¹ Nell'intento di diffondere l'uso del computer anche nelle interviste faccia a faccia, l'EUROSTAT ha favorito l'acquisizione, da parte degli istituti nazionali di statistica dei paesi europei, di insiemi limitati di computer da utilizzare a fini sperimentali nelle rilevazioni sulle forze di lavoro. Questi dati sono utilizzati, assieme ad altri rilevati in modo convenzionale, dall'EUROSTAT per produrre periodicamente statistiche sulle forze di lavoro in Europa.

puter nelle rilevazioni faccia a faccia, rispetto alla prassi dell'Istituto Nazionale di Statistica di reclutare *ad hoc* intervistatori, ha generato problemi di tipo logistico, problemi che diverrebbero insormontabili se l'uso del computer sul campo fosse esteso alle migliaia di rilevatori reclutati *in loco* per le indagini correnti sulle famiglie.

La rilevazione di dati mediante intervista CAPI presso le famiglie in una prospettiva longitudinale, la quale implichi il contatto ripetuto delle stesse unità a cadenze predeterminate, comporta la definizione di una rete di rilevatori addestrati su scala nazionale, la consegna di un computer e l'instaurazione di una rete di assistenza tecnica completamente innovative per l'ISTAT.

2.1 *I vantaggi dell'intervista on line*

L'uso del computer rappresenta un motore importantissimo della rivoluzione che sta, in questi anni, attraversando il terreno della misura statistica dei fenomeni per i seguenti due motivi:

- ✓ *riduce enormemente i tempi e i costi delle indagini*: il programma informatico, infatti, mostrando sullo schermo del computer le domande e permettendo di digitare direttamente le risposte degli intervistati, unifica in un solo passo la stampa dei questionari, la spedizione, la codifica e l'immissione dei dati.
- ✓ *riduce notevolmente il tempo e ottimizza il funzionamento della convalida e della correzione degli errori o mancate risposte, che avviene in tempo reale*. Il programma, infatti, può segnalare eventuali valori fuori campo, valori contraddittori o mancanti, permettendo così a chi intervista di ripetere la domanda per verificare la correttezza della risposta. Così facendo, si può pensare che, alla fine dell'intervista, gli errori siano molto pochi.
- ✓ *conduce al miglioramento della qualità dei dati nelle indagini e nelle ricerche panel*.

Perché, però, sia conveniente sostituire il questionario cartaceo con un'intervista assistita dal computer, telefonica, 'faccia a faccia' o autogestita, essa lo deve al meno riprodurre

e, quindi, rimpiazzare. Con maggiore ambizione si può dire che essa dovrebbe anche sostituire gli intervistatori o, al meno, ridurre notevolmente il loro ruolo, lasciando solamente il compito di motivare il rispondente e spiegare le domande (Saris, 1991).

Si espongono ora i requisiti che deve avere un questionario *computer assisted* nel caso in cui

- a) sostituisca solamente il questionario cartaceo;
 - b) sostituisca il questionario cartaceo e, parzialmente, l'intervistatore;
 - c) sia migliore di una intervista normale.
- a) Un programma che sostituisce il questionario cartaceo è caratterizzato principalmente da tre requisiti di base. Anzitutto la presentazione delle informazioni, delle domande, delle categorie, delle risposte e delle istruzioni; poi la registrazione delle risposte e infine il passaggio alla domanda successiva.

Per essere in grado di far ciò, il programma di intervista deve avere informazioni su: il tipo di domanda, la lunghezza della risposta, la tipologia del dato da inserire (numerico o alfanumerico), l'inizio del testo, la fine del testo.

b) In generale, i requisiti richiesti a chi intervista sono:

- ⇒ ottenere collaborazione;
- ⇒ motivare il rispondente a rispondere;
- ⇒ costruire un grado di fiducia sufficiente a ottenere risposte oneste;
- ⇒ interrompere il rispondente quando racconta storie o avvenimenti irrilevanti;

Inoltre, l'intervistatore deve:

- ⇒ controllare se la risposta è appropriata;
- ⇒ aiutare il rispondente se la domanda non è compresa;
- ⇒ codificare le risposte in categorie precostituite;
- ⇒ scrivere le risposte.

Chiaramente, un programma informatico non è in grado di svolgere tutte le funzioni appena elencate, soprattutto quelle di tipo più 'sociale'. Quindi, i requisiti auspicati dal questionario *on line* affinché sostituisca almeno parzialmente l'intervistatore sono solo

- quelli ‘tecnici’, cioè gli ultimi quattro:
- ⇒ controllare che la risposta sia appropriata → il programma riesce, non solo a verificare se il dato digitato non appartiene all’intervallo di variazione (cosa che potrebbe fare anche un intervistatore), ma anche confrontarlo con ogni altra informazione che si voglia inserire per il controllo della correttezza (cfr. Paragrafo 3);
 - ⇒ aiutare il rispondente se la domanda non è compresa → il programma, se ben costruito, può fornire delle opzioni cosiddette di *help* che possono essere richiamate all’occorrenza;
 - ⇒ codificare le risposte in categorie precostituite → la codifica della risposta normalmente fa perdere molto tempo all’intervistatore ed è un’operazione piuttosto complessa; le procedure informatiche possono, quindi, essere di grande aiuto;
 - ⇒ scrivere le risposte → non è difficile fornire al programma un *editor* semplice che permetta di inserire le risposte con sigle brevi che, cioè, sintetizzino parole intere e, talvolta, anche frasi.
- c) I requisiti di base che caratterizzano un programma che sia migliore di una intervista normale sono:
- ⇒ fare calcoli utili per la verifica della correttezza delle risposte → per esempio verificare se la spesa totale per un prodotto e la quantità consumata siano coerenti con il prezzo medio del prodotto considerato;
 - ⇒ modificare in un secondo tempo le risposte date;
 - ⇒ rendere casuale l’ordine delle domande in batteria e le categorie di risposta → spesso il rispondente sceglie la prima o l’ultima delle risposte elencate a causa della difficoltà di memorizzare tutte le risposte dell’elenco o di ricordare la domanda stessa; l’elencazione casuale delle domande o delle risposte possibili può minimizzare questo rischio di errore;
 - ⇒ fare ‘salti’ complessi tra una domanda e l’altra (*branching*) facilmente e in breve tempo → se l’operazione dura più di pochi secondi, il rispondente può annoiarsi e ridurre la volontà di collaborare al minimo;
 - ⇒ eseguire codifiche elaborate in breve tempo;

⇒ formulare in maniera corretta le domande e le informazioni necessarie per rispondere;

⇒ fornire informazioni ausiliarie in ogni momento;

⇒ convalidare le risposte → in genere l'intervistatore non ha tempo o capacità di convalidare le risposte nel corso dell'intervista. Con un programma informatico, grazie alle caratteristiche appena analizzate, è il programma stesso ad evidenziare eventuali discrepanze consentendo, così, la verifica ed eventualmente la correzione del valore inserito: questa operazione può essere fatta in tempo reale.

Per effettuare le correzioni, le procedure sono diverse, ma tutte partono dallo stesso principio: il programma confronta le risposte a differenti domande e evidenzia le incoerenze o le combinazioni di risposte troppo contraddittorie; chiaramente, ciò che non può fare il programma è segnalare quale tra quelle evidenziate sia la risposta sbagliata.

Ad esempio, il programma può comunicare:

<u>Domanda</u>	<u>Risposta</u>
Età padre	35
Età figlio	36

Impossibile!

Il programma evidenzia sullo schermo le domande le cui risposte sono in contraddizione. Supponendo che l'errore sia in una di esse, l'intervistatore deve:

- * comunicare che qualcosa è sbagliato (con frasi del tipo "lei ha fatto un errore ...", "mi spiace, c'è una difficoltà ...", "mi spiace, il computer mi segnala che ...");
- * fare domande per chiarire l'errore (quante più domande si pongono, tanto minore è il rischio di correggere una risposta che in realtà non è sbagliata) → il programma dovrebbe fornire le possibili domande da porre e non lasciare all'intervistatore il difficile compito di trovare le parole giuste da dire;
- * tornare indietro alla domanda a cui è stata data la risposta sbagliata e correggerla;
- * se vi sono altre domande che dipendono da questa risposta, vanno ricercate per verificare la correttezza;

* infine, il programma ritorna al punto in cui era stata evidenziata l'incoerenza.

Il problema sorge se non sono le domande evidenziate ad essere errate, ma altre che non appaiono direttamente: il nome della variabile non compare sullo schermo e l'intervistatore è costretto a ricercarla per correggerla.

Questa operazione richiede del tempo, a meno che il programma non sia stato elaborato in modo tale da fornire sullo schermo i nomi di tutte le domande coinvolte (anche indirettamente) nella incoerenza.

Ci sono, chiaramente, anche degli svantaggi collegati all'utilizzo dell'assistenza del computer:

- ◇ lo schermo del computer generalmente è più piccolo del foglio del questionario e può contenere meno informazioni → ciò può provocare la perdita a metà dell'intervista del suo scopo;
- ◇ è più semplice eseguire le correzioni su un foglio piuttosto che andare avanti e indietro tra le schermate del programma;
- ◇ si perde la coordinazione tra mano e occhio.

3. Il controllo di ammissibilità on line delle risposte

Il controllo di ammissibilità, o plausibilità, di una risposta riguarda la coerenza della stessa con le informazioni di cui dispone il ricercatore. Nei sistemi *computer assisted*, le informazioni possono essere trasferite nella memoria del calcolatore come regole di elaborazione *on line* del dato. La plausibilità della risposta determina il proseguimento dell'intervista, l'inammissibilità della risposta attiva un processo che comprende il tentativo di recupero di un'informazione plausibile e di eventuali informazioni supplementari finalizzate all'aggiustamento delle stime e al miglioramento del questionario per indagini dello stesso tipo.

Le informazioni per il controllo di ammissibilità possono essere:

1. la definizione del campo di variazione X della variabile Y , ossia l'insieme delle risposte plausibili per il quesito sottoposto;

2. le caratteristiche della distribuzione di frequenze $f(y)$ della variabile Y associata al quesito sottoposto;
3. le caratteristiche del campo di variazione X della variabile Y condizionatamente ad una caratteristica nota Z del rispondente. Se si conosce la caratteristica ascrittiva Z del rispondente j , il campo di variazione $X|Z=z_j$ della risposta y_j può essere più limitato di quello dell'insieme delle unità statistiche su cui è svolta la rilevazione e la plausibilità della risposta può essere, pertanto, prevedibile per l'unità j con maggiore precisione che sull'intero campione di unità investigate. La variabile Z può essere una combinazione di modalità di variabili note al ricercatore;
4. le caratteristiche della distribuzione di frequenze $f(y|Z=z_i)$ della variabile Y condizionatamente ad una caratteristica nota Z del rispondente. Se si conosce la caratteristica ascrittiva Z (o combinazione di caratteristiche) del rispondente j , la distribuzione di frequenze $Y|Z=z_j$ della categoria di persone h cui appartiene l'unità i ($h \in i$) può essere più circoscritta di quella dell'intero campione su cui è svolta la rilevazione e l'ammissibilità della risposta può essere, quindi, più facilmente determinabile.

I controlli richiedono che, per ciascuna variabile, si specifichino sia il campo di variazione incondizionato, sia quelli condizionati alle caratteristiche eventualmente note prima della somministrazione del quesito Y , sia le caratteristiche distributive della variabile note (o ipotizzate) prima dell'inizio della rilevazione, sia le caratteristiche distributive di Y condizionate alle caratteristiche delle classi di rispondenti note/ipotizzate prima della somministrazione del quesito.

I controlli in linea della qualità delle risposte fornite in una rilevazione automatica dei dati sono possibili se si stabilisce una gerarchia tra le variabili. La gerarchia riguarda sia le informazioni raccolte in un'indagine trasversale, sia quelle di un'indagine longitudinale.

In un'indagine trasversale, le informazioni possono essere in relazione gerarchica secondo (Bassi e Fabbris, 2000):

1. *il contenuto*, se si riferiscono a fenomeni legati da una relazione di dipendenza incrociata. Ad esempio, la dichiarazione di aver compiuto un viaggio come conducente di auto privata, non è logicamente compatibile con un'età inferiore ai 18 anni;

2. *lo spazio*, se fanno riferimento a popolazioni contenute una nell'altra. Per esempio, si esaminano i viaggi compiuti nella settimana di riferimento da ciascuna unità di rilevazione: se il rispondente dichiara di avere compiuto n viaggi occasionali nella settimana, il *file* dovrà contenere n distinti *record* che descrivono analiticamente questi viaggi;
3. *il tempo*, se fanno riferimento a caratteristiche strutturali che non mutano nel tempo (per esempio, il sesso dei rispondenti) o che sono soggette a variazioni determinabili in funzione del tempo (per esempio, l'età).

In un'indagine longitudinale, la dipendenza tra le informazioni deriva, analogamente, da:

1. *la stabilità temporale, o la variazione secondo andamenti noti*, di variabili ascrivibili (come sesso, l'età e le caratteristiche fisiche delle persone, la superficie dell'alloggio, ma anche, entro limiti accertabili, il titolo di studio, il possesso della patente, il numero di componenti della famiglia, ecc.);
2. *il condizionamento temporale delle risposte inerenti ad uno stesso fenomeno osservato ripetutamente nel tempo* presso la stessa unità o a fenomeni dipendenti rilevati in tempi diversi presso la stessa unità (per esempio, le caratteristiche culturali e politiche dell'individuo, le sue abitudini sanitarie, alimentari, ricreative, ecc.);
3. la dinamica della sottopopolazione cui l'unità statistica appartiene (per esempio, le modifiche nel reddito e nei consumi della categoria sociale cui appartiene la famiglia; le variazioni di voto degli appartenenti ad un gruppo politico, ecc.).

Nel seguito, condiscendendo ad una terminologia dominante, si denominano “di compatibilità” i controlli di ammissibilità inerenti al campo di variazione (condizionato e incondizionato) e “di verosimiglianza” quelli inerenti alle caratteristiche distributive.

La logica delle due classi di controlli di ammissibilità è piuttosto simile, e così pure sono simili le conseguenze sul piano pratico. La sequenzialità logica dei criteri di controllo è già stata argomento delle presentazioni sopra riportate; i sistemi di controllo di ammissibilità sono praticamente composte di regole di controllo prefissate, sia che si tratti di campi di variazione, sia che si tratti di caratteristiche distributive della variabile associata al quesito. Le differenze sono solo nella teoria di riferimento, che nel caso del campo di variazione è la struttura formale, deterministica (si potrebbe dire “informatica”), dei dati. Nel caso della

distribuzione di frequenze, il riferimento concettuale è la struttura stocastica, o probabilistica (per simmetria, si può dire “statistica”), dei dati.

4. Controlli di coerenza delle risposte in indagini trasversali

I controlli di coerenza, o compatibilità, delle risposte hanno come scopo quello di individuare e, là dove possibile, correggere eventuali incoerenze tra le informazioni presenti nel sistema di rilevazione. Queste possono essere regole predefinite o dati forniti dal rispondente a domande diverse di uno stesso questionario o da altri rispondenti legati a quello di cui si controlla la risposta.

I controlli basati su informazioni presenti nel sistema riguardano il campo di variazione, quelli basati su risposte date dal rispondente a quesiti antecedenti connessi a quello in esame riguardano il campo di variazione condizionato. Le informazioni contenute nel sistema si possono assumere “ancillari”, o “a priori”, del rispondente j ($j=1, \dots, n$).

I controlli si basano sul confronto tra la modalità ammissibile X_j del campo di variazione X che dovrebbe contenere la risposta (modalità osservata) y_j della variabile Y : X_j è il campo di variazione Y per l’unità statistica j .

Se il campo di variazione è condizionato, la combinazione di modalità che determinano X_j si denomina $Z_{h \subset j}$.

In un’indagine trasversale, la combinazione delle risposte ammissibili, X_j , e di quella ottenuta dall’unità j , y_j , si dice *incompatibile* se ha probabilità nulla di verificarsi nella realtà che si esamina:

$$P(y_j; X_j) = 0 \quad (j=1, \dots, n), \quad (2)$$

mentre la “modalità vera”, Y_j , rilevabile presso l’unità j in assenza d’errore, ha forzatamente probabilità 1 di essere compresa nel campo di variazione X_j :

$$P(Y_j; X_j) = 1 \quad (j=1, \dots, n), \quad (3)$$

Se si ottiene una risposta incompatibile, il sistema di raccolta automatica dei dati darà un messaggio di errore. Se il ricercatore non dispone di elementi esterni per decidere sulla verosimiglianza dell'osservazione y_j o del campo di variazione condizionato $X=x_j$ (a causa della misura con errore di $Z_{h \subset j}$), o di ambedue, i dati si considerano entrambi errati.

Se, invece, X_j è certo, o comunque più attendibile di y_j , si considererà errato y_j . Se si stabilisce una gerarchia tra informazioni in termini di attendibilità è, quindi, possibile determinare quale tra le modalità incompatibili va considerata errata.

Controlli di compatibilità particolarmente efficaci sono quelli che sfruttano la dipendenza gerarchica delle informazioni derivante dall'essere le unità di rilevazione suddivise in sottopopolazioni. Ovviamente, a questo scopo, è necessario definire le categorie di una qualche variabile filtro che identifica, senza errore, i diversi gruppi di unità, le informazioni fornite dai quali saranno sottoposte a controlli di coerenza *ad hoc*.

5. Controlli di verosimiglianza delle risposte in indagini trasversali

I controlli di verosimiglianza si basano sul principio che, se le frequenze di una certa variabile nella popolazione di riferimento seguono una distribuzione nota, allora è determinabile la probabilità di osservare qualunque insieme di valori. Ciò consente di individuare i valori osservati poco verosimili, ovvero poco probabili.

In una distribuzione regolare, unimodale, con la massa delle frequenze concentrata verso il centro della distribuzione, si può determinare la probabilità di una classe di valori in base alla media e alla varianza. Un valore osservato che si allontani dalla media, in più o in meno, per più di tre scarti quadratici medi si può considerare anomalo (*outlier*), nel senso che la sua probabilità di verificarsi è quasi nulla. In una distribuzione di frequenze empiriche, tra i valori anomali, possono coesistere valori autenticamente molto grandi o molto piccoli e dati errati a causa di errori nelle risposte. I primi sono, naturalmente, accettati e costituiscono anzi materia di analisi particolare, mentre i secondi sono gravi errori di rilevazione che possono compromettere l'esito delle analisi.

Un dato, quindi, è posto a verifica se risulta poco verosimile secondo il criterio delineato sopra. Esso, ad esempio, è sottoposto a riscontro se assume valori poco probabili rispetto alla categoria cui appartiene l'unità di rilevazione, come nel caso di un rispondente che dichiara di essere operaio, ma presenti lo stesso reddito di un dirigente d'azienda.

I controlli di verosimiglianza diventano più efficaci se sono impostati per tenere conto delle relazioni causali esistenti tra le informazioni raccolte: la verosimiglianza di un'informazione può allora essere valutata non solo facendo riferimento alla distribuzione di frequenze univariata della variabile in questione, ma a distribuzioni multivariate di variabili concatenate.

Si denoti con y_j la risposta fornita dall'unità i per il fenomeno osservato Y e sia $X_{h \subset j} = f(y_j | Z_{h \subset j})$ la distribuzione di frequenze della variabile Y attesa per il rispondente j nell'ipotesi che la modalità condizionante Z_j (o la combinazione di modalità condizionanti $Z_{h \subset j}$) sia osservata senza errore. Si dice che la risposta y_j è poco verosimile se

$$P(y_j | X_{h \subset j}) < \mathbf{p} \quad (j=1, \dots, n; h=1, \dots, H) \quad (4)$$

dove \mathbf{p} è un limite di probabilità determinato. La disuguaglianza (4) definisce un intervallo di variabilità di Y_j .

Se la distribuzione di frequenze è condizionata, è determinabile la probabilità che l'unità statistica j risponda y_j al quesito Y se appartiene alla categoria $Z_{h \subset i}$ ($i = 1, \dots, n$), dove h ($h=1, \dots, H$) è la sottopopolazione definita dalla "modalità filtro" $Z_{h \subset i}$.

Le differenze tra gli H valori attesi della variabile Y nelle H categorie sono determinate da un'analisi della varianza nelle categorie²². Le modalità condizionanti la distribuzione di frequenze della Y faranno prevalentemente riferimento a caratteristiche ascrittive delle unità o, comunque, a caratteristiche di natura strutturale delle stesse. Se la Y è una variabile sociale degli adulti, una caratteristica ascrittiva condizionante può essere la professione, op-

²² Un'analisi statistica appropriata per la massimizzazione delle differenze tra medie della variabile Y nelle categorie è il metodo della segmentazione binaria. Applicando questo metodo, si possono determinare le modalità e le eventuali interazioni tra modalità delle variabili condizionanti.

pure la classe di reddito. Per certi tipi di atteggiamenti o comportamenti, può essere importante l'essere proprietari dell'abitazione in cui la famiglia vive, oppure l'essere proprietari di almeno un'abitazione per le vacanze. Per prevedere i comportamenti di viaggio (Fabbris e Bassi, 1997), è importante sapere se l'individuo è un pendolare giornaliero, o settimanale, o se è un "grande viaggiatore", o "un viaggiatore occasionale", o una persona impedita a viaggiare.

Per essere gestibili, le categorie condizionanti dovrebbero:

- massimizzare la differenza tra le medie della variabile Y nelle categorie formate;
- avere varianza interna non superiore a quella dell'intero campione,
- essere in numero ridotto.

I controlli di verosimiglianza possono essere definiti nel sistema di controllo

- *come regole deterministiche*, ossia come controlli del superamento di una soglia di valore prefissata. Per esempio, conoscendo la distribuzione dei redditi degli operai, si definisce come improbabile il reddito dell'unità osservata j , di cui sia stata accertata la posizione nella professione $Z_j = \text{"operaio"}$, che sia molto lontano dalla media $Z = z_h \subset j$ degli operai;
- *come regole stocastiche*, ossia come controlli del superamento della soglia di probabilità calcolata in funzione di parametri determinati per categorie di rispondenti variamente determinate.

Le informazioni di base per la determinazione delle caratteristiche della distribuzione della variabile Y possono essere ipotizzate sulla base di informazioni a priori, oppure in funzione di dati acquisiti nel corso della rilevazione. Questo secondo caso configura il sistema come "esperto", ossia in grado di modificare le informazioni di base, uguali per tutte le unità del campione, con informazioni acquisite via via che si realizzano le interviste del campione²³.

Analogamente al controllo di coerenza, se si ottiene una risposta inverosimile, il sistema di rilevazione dei dati darà un messaggio. Nel caso dell'analisi della verosimiglianza, il ri-

²³ In questa nota si trascura l'analisi della possibilità di acquisire informazioni sulle caratteristiche della distribuzione della variabile nel corso della rilevazione. Questa logica è virtualmente interessante per i sistemi *computer assisted*, ma al momento attuale poco praticabile.

cercatore dovrà disporre di modalità condizionanti certe per l'unità j (si dovrà, cioè, assumere che $Z_{h \subset j}$ sia stata misurata senza errore). In questo modo, muovendo dall'assunto che nessun dato è privo d'errore, si stabilisce una gerarchia tra le informazioni a priori e quelle osservate in termini di attendibilità, con le prime, naturalmente, più credibili delle seconde.

6. Controlli di coerenza delle risposte in indagini panel

I controlli di coerenza delle risposte ottenute in un'indagine longitudinale possono essere regole predefinite o dati forniti da un rispondente a domande poste alla stessa persona e/o a persone a questa legate in occasioni d'indagine precedenti.

I controlli riguardano il campo di variazione della variabile Y_t ($t=1, \dots, T$) in funzione delle risposte date dal rispondente nelle occasioni ($1, \dots, t-1$). Le informazioni presenti nel sistema si possono assumere "ancillari" della risposta del rispondente j ($j=1, \dots, n$).

In un'indagine longitudinale, X_{jt} è il campo di variazione di Y_t per l'unità statistica j nell'occasione di rilevazione t ($t=1, \dots, T$). La variabile Z condizionante il campo di variazione di Y_t può rappresentare lo stesso fenomeno descritto da Y misurato in tempi diversi presso l'unità j , o un fenomeno concatenato a Y_t misurato in tempi diversi.

In un'indagine longitudinale, la combinazione delle risposte ammissibili, X_{jt} , e della risposta ottenuta dall'unità j nell'occasione d'indagine t , y_{jt} , si dice *incompatibile* se ha probabilità nulla di verificarsi nella realtà che si esamina:

$$P(y_{jt}; X_{jt}) = 0 \quad (j=1, \dots, n; t=1, \dots, T) \quad (5)$$

mentre la "modalità vera", Y_{jt} , rilevabile presso l'unità j al tempo t in assenza d'errore, ha forzatamente probabilità 1 di essere compresa nel campo di variazione X_j :

$$P(Y_{jt}; X_{jt}) = 1 \quad (j=1, \dots, n; t=1, \dots, T). \quad (6)$$

Se si ottiene una risposta incompatibile, il sistema di raccolta automatica dei dati darà

un messaggio di errore. Se il ricercatore non dispone di elementi esterni per decidere sulla verosimiglianza dell'osservazione y_{jt} , o del campo di variazione condizionato $X_t=x_{jt}$ (a causa della misura con errore delle informazioni concomitanti $Z_{ht \subset jt}$), o di ambedue, si considerano corretti i dati coerenti con la modalità prevalente o con la razionalità implicita nella sequenza storica dei dati posti a confronto.

In presenza di dati certi, si può, naturalmente, ancorarsi a questi per stabilire una gerarchia di attendibilità nei controlli. La dipendenza gerarchica si può stabilire sia con riferimento ad informazioni trasversali, sia con riferimento a categorie definite longitudinalmente. Ciò significa che, mentre l'attendibilità di una variabile "trasversale" Z può essere valutata in base alla coerenza tra dati raccolti in nella stessa indagine in cui si osserva Y , o rispetto a dati "ufficiali" ottenuti esternamente all'indagine, l'attendibilità "longitudinale" Z_t si determina in base alla coerenza tra informazioni dello stesso tipo di Y_t , oppure di variabili correlate a Z_t .

7. Controlli di verosimiglianza in indagini panel

I controlli di verosimiglianza in un'indagine longitudinale si basano sul principio che presso l'unità statistica j , già interpellata in $t-1$ occasioni, la risposta al quesito Y_t è prevedibile se lo stesso quesito è stato posto in occasioni precedenti e Y_t è correlato temporalmente con Y_g ($g=1, \dots, t-1$) ed esistono, eventualmente, altre variabili Z correlate con Y_t . Le osservazioni poco verosimili sono quelle che manifestano uno scarto dal valore stimato superiore a un determinato valore, oppure una probabilità di verificarsi inferiore a una determinata soglia. Si considera inverosimile l'osservazione y_{jt} che differisce dal suo valore atteso \hat{y}_{jt} con probabilità superiore ad una soglia prefissata:

$$P(y_{jt} - \hat{y}_{jt}) > \mathbf{p} \quad (j=1, \dots, n; t=1, \dots, T) \quad (7)$$

dove:

$$\hat{y}_{jt} = f(Y_{j1}, \dots, Y_{jt-1}; Z_{j1}, \dots, Z_{jt}) \quad (j=1, \dots, n; t=1, \dots, T) \quad (8)$$

I controlli riguardano il valore atteso della variabile Y_t ($t=1, \dots, T$) in funzione delle risposte date dal rispondente nelle occasioni ($1, \dots, t-1$) e di eventuali altre informazioni acquisite esternamente all'indagine. Le informazioni presenti nel sistema si possono assumere "ancillari" della risposta del rispondente j ($j=1, \dots, n$).

Y può essere una variabile su qualunque scala. Se la variabile Y è quantitativa, la funzione (8) è un'equazione di regressione; se è dicotomica, la funzione sarà di regressione logistica. La logica della probabilità del possesso di un attributo si può anche applicare a variabili qualitative. Per queste, il metodo multivariato da applicare è l'analisi della funzione discriminante.

In un'indagine longitudinale, il sistema di rilevazione accerta la verosimiglianza della risposta:

1. "a freddo", in sede di trattamento dei dati, evidenziando (graficamente o numericamente) la successione delle risposte, eventualmente suddivise in segmenti temporali, e determinando le eccezioni al comportamento standard del rispondente. I valori eccezionali sono determinabili abbastanza facilmente a posteriori, più complesso è l'evidenziarli in linea²⁴;
2. sia in linea, sia a posteriori, valutando le risposte ottenute dall'interpellato, eventualmente classificato in una categoria omogenea di rispondenti. Se il gruppo in cui l'individuo è stato classificato ha subito, col passare del tempo, delle variazioni nei comportamenti standard, ci si aspetta che analoghe variazioni si osservino anche presso le singole unità che fanno parte del gruppo. Se non vale questo assunto, la previsione si può determinare in funzione dei soli valori espressi dal rispondente nelle occasioni di rilevazione precedenti.

²⁴ I controlli di verosimiglianza che evidenzino le risposte ottenute in precedenti occasioni di rilevazione sono gestibili ragionevolmente in indagini svolte mediante rilevatori. Nelle rilevazioni di tipo CASI, i controlli e i relativi interventi debbono essere interni al sistema di rilevazione.

8. *La correzione “a freddo”*

E' la strategia più comunemente utilizzata per la correzione dei dati nella statistica ufficiale.

Comprende tutte quelle correzioni che vengono apportate in una fase diversa da quella dell'acquisizione nelle indagini con questionario cartaceo.

8.1 *La correzione nella fase di registrazione dati*

Nella fase di registrazione, appositamente progettata e indipendente dall'intervista, possono essere attivati controlli volti a identificare errori di codifica e trascrizione dei dati.

A questo scopo, vengono sviluppate forme di controllo automatico che segnalano la presenza di valori fuori dominio o valori anomali, tramite *warning* a video.

La correzione avviene solitamente attingendo dal questionario cartaceo e verificando il dato immesso.

Contattare nuovamente l'unità di rilevazione è una pratica non frequente poiché di difficile attuazione e dispendiosa sia in termini economici che di tempo.

Nel caso di rilevazioni *computer-assisted*, si elimina completamente il problema della errata trascrizione dei dati in quanto l'immissione in archivio avviene contestualmente all'intervista, con la possibilità di correggere “a caldo” le informazioni eventualmente errate per trascrizione.

8.2 *L'editing*

Nel progettare un'indagine statistica, si prevede una fase *ad hoc* per la correzione degli errori nei diversi passi, denominata *editing*.

Per svolgere un'operazione di *editing*, si svolgono due attività, l'individuazione e la correzione dell'errore.

Per l'individuazione, vanno formulate delle regole, dette di incompatibilità, che permettano di far emergere se un record risulta affetto da errore.

Per la correzione, va individuato il nuovo dato da imputare al posto di quello errato, correggendo il file grezzo.

La formulazione delle regole e l'assegnazione dei nuovi valori può avvenire mediante:

- ✓ procedure deterministiche, in genere per gli errori sistematici;
- ✓ procedure probabilistiche o stocastiche, in genere per gli errori casuali.

8.2.1 La correzione deterministica

Applicando il criterio deterministico, l'individuazione dell'errore e la scelta della modalità da forzare avvengono nello stesso momento; infatti le tecniche di correzione (i piani di correzione) sono costituiti da Regole di Imputazione Deterministica (R.I.D.) del tipo SE ... ALLORA (IF ... THEN) definite a priori; cioè regole costituite da (Riccini *et al*, 1995): una 'parte condizione' = **IF ((cond1) AND/OR (cond2) ...)** che esprime la condizione per cui si verifica l'errore nel record; una 'parte azione' = **THEN ((azione1) AND/OR (azione2) ...)** che stabilisce a priori il tipo di correzione da effettuare sul record per eliminare l'errore.

Per esempio:

SE ((età=13) E (stato civile = coniugato)) ALLORA (stato civile = celibe).

Quindi, quando un record, durante l'esecuzione della procedura di correzione, attiva alcune di queste regole, vengono automaticamente modificate le variabili indicate nella 'parte azione' assegnando i valori attribuiti alla destra del segno di attribuzione.

I piani deterministici operano le correzioni sulla base di regole fondate su dati disponibili ritenuti 'veri', scelti a priori dal ricercatore (metodo *if-then*).

I vantaggi dell'approccio deterministico sono (Barcaroli, Di Pietro, Venturi, 1993):

- 1) la completa applicabilità: un piano deterministico è, infatti, sempre applicabile ai dati una volta tradotte le R.I.D. in istruzioni di programma;
- 2) l'efficienza elaborativa: il tempo necessario per eseguire il programma che traduce il piano deterministico è proporzionale al numero di R.I.D. e al numero di record;
- 3) l'orientabilità degli effetti: quando vi sono delle incompatibilità, è possibile orientare i ri-

sultati dell'applicazione del piano deterministico, definendo opportunamente la 'parte condizione' di ogni R.I.D. In tal modo, sulla base della fiducia che si nutre rispetto alla correttezza delle variabili, si può stabilire una gerarchia tra di esse, modificando quelle che si ritengono meno affidabili.

I limiti di questo approccio sono (Barcaroli, Di Pietro, Venturi, 1993: 9-10):

- a) assenza di garanzia di correttezza finale dei risultati;
- b) il risultato finale delle correzioni dipende dalla particolare sequenza con cui le regole vengono applicate ai dati;
- c) possono avere origine *loop* (cioè 'circuiti' di correzioni) e può, quindi, risultare impossibile correggere un dato record;
- d) possono esserci R.I.D. ridondanti;
- e) possono essere introdotte distorsioni sulle distribuzioni originali delle variabili.

8.2.2 *Le procedure probabilistiche di individuazione e imputazione*

L'approccio probabilistico al problema della correzione dei dati, invece, non prevede la possibilità/necessità di definire a priori l'elenco di azioni da intraprendere per eliminare gli errori dei dati. Si basa, quindi, su due momenti differenti: individuazione degli errori e correzione degli stessi.

1) L'individuazione

Per individuare gli errori occorre definire un insieme di regole di incompatibilità, espresse attraverso una sequenza di relazioni logiche tra le possibili risposte alle domande e derivanti, in genere, da conoscenze a priori. Questo insieme è chiamato 'insieme completo degli *edit*' (Fellegi e Holt, 1976), che determinano la non ammissibilità di modalità in una variabile in presenza di determinate modalità in altre variabili (Fellegi, Holt, 1976).

Per essere completo, l'insieme degli edit deve essere:

- 1) 'minimale', cioè privo di edit ridondanti; ciò assicura che:
 - una volta modificati i valori delle variabili di cui è costituito l'edit, il record può considerarsi corretto rispetto agli errori;

- le variabili considerate dagli edit dell'insieme sono quelle che con maggior probabilità sono effettivamente affette da errori;
- 2) 'corretto', cioè privo di edit tra loro contraddittori;
 - 3) 'completo', cioè che contenga in forma esplicita tutti gli edit implicitamente definiti.

La metodologia che consente di definire l'insieme completo prevede una serie di passaggi.

- 1) Il primo consiste nel definire le regole di incompatibilità che stabiliscono le condizioni di errore all'interno di un record, permettono cioè di individuare la presenza di errori all'interno dei dati. Questo insieme di regole si chiama 'originale' ed è costituito da 'edit in forma normale'.

Le regole di incompatibilità prendono il nome di 'edit in forma normale' quando sono date dall'intersezione di sottodomini di almeno due variabili presenti nel record, ad esempio:

ETA' (0, 1, 2, ..., 14) AND STATO_CIVILE (coniugato, divorziato, separato, vedovo)

E' sempre possibile tradurre un edit qualsiasi in uno o più edit in forma normale.

- 2) Nella seconda fase si deve modificare l'insieme originale per ottenere l'insieme 'minimale'.

Ciò si ottiene provvedendo a:

- eliminare le ridondanze;
 - aggregare le regole uguali.
- 3) L'insieme minimale, pur essendo internamente coerente, non è corretto perché non sono ancora state eliminate le regole contraddittorie, le quali si considerano 'regole degeneri'.

Per eliminare le regole contraddittorie, devono essere generate le regole implicite a partire da quelle esplicite trattate fino ad adesso: questo permette di generare tutte le regole che guidano nel processo di correzione e permettono di assegnare dei valori che sicuramente riconducono il record ad una situazione di correttezza, assicurando il minimo cambiamento dei dati originari.

Questo passo, decisamente critico nell'applicazione della metodologia di Fellegi-Holt (a causa dell'enorme numero di edit impliciti che è possibile generare), conduce all'insieme completo.

Una volta creato l'insieme di regole di incompatibilità, si deve procedere all'individuazione degli errori: la procedura indica la presenza di un dato errato quando un edit viene 'attivato' da un dato record, cioè se questo assume valori delle variabili presenti nell'edit che siano interni al campo di variazione stabilito.

Limiti (Barcaroli, Venturi, 1997):

- i) la derivazione degli edit impliciti rappresenta una operazione critica: quando il questionario è complesso possono essere definiti moltissimi edit espliciti e un numero enorme di edit impliciti (da un punto di vista teorico, infatti, dati m edit espliciti iniziali, il numero di edit impliciti che possono essere costruiti è 2^m);
- ii) la metodologia di Fellegi-Holt localizza in maniera ottimale gli errori solo se la loro natura è stocastica. Se essi sono, invece, sistematici (del tutto o una parte) si corre il rischio di mantenere errori nelle variabili o correggere variabili corrette.

2) La correzione

La correzione delle informazioni errate (cioè l'imputazione dei valori) avviene in una seconda fase, dopo aver localizzato gli errori nei record.

Si possono adottare due strategie di correzione (Abbate, 1997: 72-74):

- 1) il valore può essere fornito applicando un modello di regressione i cui parametri sono stimati sulla base delle risposte valide, ipotizzando, quindi, una relazione probabilistica tra la variabile carente di risposte e determinate variabili predittive. Questa strategia, denominata 'imputazione secondo modello', garantisce il rispetto della distribuzione semplice della variabile e delle distribuzioni congiunte. Il suo limite risiede nella sua non immediata applicabilità, perché la ricerca del modello richiede una adeguata analisi del fenomeno e l'individuazione del modello più efficace;
- 2) la correzione può essere fatta forzando il valore valido di una unità simile (detta 'donatore'), con risposte complete e corrette (cfr. Paragrafo 8.3). La correzione di un record errato secondo questa tecnica è un processo di definizione e scelta:

- definizione dell'insieme di record alternativi e candidati alla sostituzione;
- scelta, tra questi, di un record sicuramente corretto. Nel processo di scelta si può porre come criterio di ottimalità che il record scelto per la sostituzione sia il meno distante possibile da quello da sostituire.

La similitudine tra unità donatrice e unità ricevente è determinata sulla base di alcune variabili scelte sulla base della loro correlazione con la variabile da imputare.

Nella prima strategia, per le variabili esplicative (sulla base delle quali si determina la funzione di probabilità delle risposte da imputare) e, nella seconda, per le variabili di matching (utilizzate nella determinazione della distanza tra unità) è necessario assicurarsi la correttezza. Si può anche applicare una procedura di correzione gerarchica delle variabili se le più importanti non contengono mancate risposte parziali.

Caratteristiche importanti (Barcaroli, 1993):

- a) la coerenza del record con l'insieme di regole e, dunque, la correttezza finale della correzione;
- b) che ogni record, per essere dichiarato corretto, soddisfi la regola del 'minimo cambiamento', cioè garantisca la minimalità delle modifiche apportate;
- c) che le distribuzioni di frequenza dei dati originali rimangano il più possibile invariate;
- d) che le regole di imputazione derivino dalle stesse regole di controllo senza necessità di una esplicita specificazione.

La tecnica Fellegi-Holt è sicuramente in grado di mantenere la distribuzione marginale di ogni variabile imputata, ma non sempre assicura la distribuzione multivariata, cosa invece indispensabile da assicurare nella fase della correzione.

8.3 *I softwares per la correzione probabilistica più utilizzati presso l'ISTAT e tecniche implementate*

Nel tempo sono stati sviluppati una serie di *softwares* generalizzati (applicabili a situazioni diverse) che implementano l'algoritmo di Fellegi-Holt e che ne permettono l'applicazione in modo automatico (Barcaroli *et al*, 1999).

Presso l'ISTAT i più utilizzati sono:

- SCIA (*Sistema Controllo e Imputazione Automatici*) per il trattamento di variabili qualitative, sviluppato all'interno dell'Istituto stesso;
- GEIS (*Generalised Edit and Imputation System*) per il trattamento di variabili quantitative, sviluppato da Statistics Canada.

SCIA, programma largamente utilizzato per la correzione dei dati nelle indagini sulle famiglie, procede all'imputazione dei valori errati secondo due tecniche "da donatore", entrambe di tipo *hot deck*: imputazione congiunta (ristretta e allargata) e imputazione sequenziale.

Se non esistono le condizioni per l'applicazione di queste metodologie, SCIA procede all'imputazione tramite forzatura sulle variabili errate tenendo conto dei valori ammissibili e delle distribuzioni di frequenza marginali delle variabili stesse.

Nel dettaglio, la tecnica di imputazione "da donatore" prevede che i valori errati all'interno di un record vengano sostituiti con i corrispondenti valori assunti dalla stessa variabile in una unità "donatrice", cioè che non ha attivato nessuna regola del piano di incompatibilità.

L'identificazione del donatore può avvenire secondo due differenti metodologie (Masselli e Barcaroli, 1994):

- *cold deck* ("a parte");
- *hot deck* ("in corso d'opera").

Con il metodo *cold deck* il file dei dati viene idealmente scisso in due parti : l'una, con-

tenente i record esatti - tutti possibili donatori - l'altra, contenente i record da correggere. La procedura di correzione agisce esclusivamente sul serbatoio delle unità errate e sceglie un donatore nel serbatoio delle unità esatte in base alla minore distanza tra le due (unità errata/esatta) rispetto a un insieme scelto di variabili da non imputare.

Con il metodo *hot deck*, la procedura opera su un unico file procedendo sequenzialmente all'analisi dei record; quando incontra un record che non attiva nessuna regola lo mantiene in memoria, nel serbatoio delle unità esatte (di dimensioni fisse, per cui a ciascuna entrata corrisponde una uscita); quando incontra un record errato, lo corregge attingendo il valore corretto dall'ultimo donatore incontrato più vicino (sempre in termini di minima distanza rispetto all'insieme di variabili da non imputare).

Secondo l'imputazione "congiunta ristretta" il donatore deve presentare per tutte le variabili da non imputare gli stessi valori dell'individuo errato. In questo caso si attribuiscono in blocco alle variabili errate del ricevente i valori delle corrispondenti variabili del donatore.

In base all'imputazione congiunta allargata il donatore deve assumere per le variabili da non imputare valori ammissibili, cioè all'interno del campo di variazione delle stesse, che non entrino in conflitto con nessuna regola. Come nel caso precedente, si procede alla donazione in blocco verso il ricevente.

L'imputazione sequenziale consiste, infine, nello scegliere come donatore, per ciascuna variabile da imputare del record errato, il record che presenta per quella variabile valori compresi nel rispettivo campo di variazione, cioè ammissibili.

9. La possibilità di correggere gli errori on line

Allo scopo di decidere l'ammissibilità di una risposta tramite controlli di compatibilità e verosimiglianza, si utilizzano, in genere, le procedure seguenti:

1. *la conciliazione tra le risposte incoerenti*. Si tenta di stabilire, insieme al rispondente, quale tra le informazioni poste a confronto, oppure utilizzate per determinare la verosimiglianza dell'ultima risposta, è attendibile, rendendo evidente al rispondente che ha in tempi diversi ha dato informazioni incoerenti, oppure che la risposta data non è coerente con informazioni acquisite esternamente. L'interpellato, posto in contraddizione, dà una risposta che concilia l'incoerenza registrata. La modalità conciliante può essere l'ultima, una delle precedenti, o una nuova modalità. Il sistema poi si incarica di forzare la modalità di conciliazione al posto di quella/e errata/e, ristabilendo la coerenza interna al sistema (Bates, 1996);
2. *lo scandaglio (probing)* della posizione del rispondente rispetto al fenomeno sul quale si è realizzata una incoerenza ovvero è emersa una inverosimiglianza. Lo scandaglio si attua con una o più domande che "prendono alla larga" l'argomento, cercando di aiutare il rispondente a formarsi un'idea più precisa di ciò che è chiesto nell'indagine e dei modi più efficienti per cercare la risposta nella propria memoria, allo scopo di ottenere informazioni corrette. La procedura dello scandaglio può essere utile anche per eliminare mancate risposte o risposte elusive.

I criteri di applicazione delle due procedure sono esplicitati nel seguito con riguardo ad una indagine trasversale, per la quale si disponga tutt'al più dei dati raccolti nel corso dell'intervista presso lo stesso soggetto, o presso altre fonti, sia ufficiali sia familiari (Paragrafo 9.1), e ad una indagine longitudinale, per la quale al tempo t si disponga in linea delle informazioni fornite dallo stesso soggetto in $t-1$ occasioni di rilevazione precedenti, oltre che di informazioni da fonti esterne, come in qualsiasi rilevazione trasversale (Paragrafo 9.2).

9.1 *Correzione di dati inammissibili in un'indagine trasversale*

In termini molto sintetici, una buona procedura per il controllo e la correzione di dati inammissibili deve rispondere esaurientemente ai due quesiti seguenti:

1. “*Quali, tre le informazioni disponibili, sono valide a fini di controllo della qualità delle singole risposte?*”
2. “*Quale criterio è applicabile per correggere la risposta y_j , di cui si è constatata l'inammissibilità, o per ottenere on line informazioni suppletive sul probabile valore di Y_j ?*”

Ambedue i quesiti richiedono approfondimenti di natura statistica. La identificazione *on line* degli errori con criteri di compatibilità tra risposte, grazie alla velocità e alla capacità di calcolo dei moderni computer, rimane un criterio importante, ma banale. Qualunque programmatore, dotato di sufficienti capacità logiche, sarebbe in grado di sviluppare i controlli adeguati all'indagine per cui sta operando. Comunque sia, i moderni programmi per lo svolgimento di indagini *computer assisted* (BLAISE, prodotto da Statistics Netherlands, cfr. Bethlehem, 1987, 1992; Bethlehem *et al.*, 1987; Bethlehem e Keller, 1992; Martin, 1993; Schuerhoff, 1993, Stol, 1993; Manners e Diamond, 1994; Kent 2000; CASES, prodotto da University of California at Berkeley; CAPTOR, cfr. Capiluppi, 2000) sono predisposti per realizzare controlli di qualità di tipo deterministico *on line*.

Il sistema di controllo e correzione delle risposte che si è prefigurato nei paragrafi precedenti necessita di analisi statistiche sulla prevedibilità delle risposte in funzione di caratteristiche note prima di porre la domanda. Il sistema di controllo va sviluppato prima che sia dato avvio alla rilevazione. Ciò implica l'anticipo delle analisi per la verifica della qualità dei dati dalla fase che precede la stima alla fase di predisposizione del questionario.

Questo modo di procedere non inficia la possibilità di svolgere controlli a posteriori sui singoli dati o sulle stime aggregate. Per esempio, i controlli che applica “a freddo” l'ISTAT (1989) su dati rilevati con questionari cartacei sono ugualmente applicabili dopo aver svolto una rilevazione assistita da computer.

Siccome la maggior parte delle relazioni tra variabili non sono calcolabili prima di raccogliere i dati, per valutare l'ammissibilità delle risposte si può:

- A. *definire il campo di variazione di ciascuna variabile;*
- B. *cercare di ottenere informazioni "ufficiali", o molto attendibili, da fonti esterne all'indagine. In base a tali dati si definiranno categorie "gerarchicamente superiori" a cui ascrivere le unità statistiche da osservare,*
- C. *predisporre un piano di controllo delle compatibilità tra modalità delle variabili rilevate nell'ambito della stessa indagine. Il piano dovrà anche predeterminare il grado di attendibilità a priori delle modalità incompatibili;*
- D. *predisporre un piano di controllo delle incompatibilità tra le modalità di risposta nell'indagine e quelle derivanti da fonti esterne. Il piano dovrà anche stabilire se le informazioni acquisite dall'esterno sono certe o hanno attendibilità superiore alle informazioni sottoposte ai controlli;*
- E. *ipotizzare caratteristiche delle distribuzioni delle variabili che si vogliono sottoporre a controllo di verosimiglianza. Le caratteristiche faranno convenientemente riferimento a distribuzioni condizionate da attributi certi delle unità campionarie.*

Se le variabili possono essere gerarchizzate in termini di attendibilità e le informazioni disponibili sono strettamente consequenziali, si possono applicare:

- regole di correzione deterministiche di tipo "*IF <evento> THEN <correzione>*" (Paragrafo 8.2.1);
- regole di correzione probabilistiche (Paragrafo 8.2.2).

I metodi comunemente applicati nell'ambito delle statistiche ufficiali sono:

- *il metodo "del donatore" (Paragrafo 8.3);*
- *per le variabili quantitative, il metodo della regressione tra la variabile Y di cui si vogliono stimare i valori (per le n^* unità con dati mancanti o palesemente errati) e una o più variabili predittive:*

$$\hat{y}_j = \hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_{1j} + \dots + \hat{\mathbf{b}}_p x_{pj} \quad (j=1, \dots, n^*) \quad (9)$$

I parametri della funzione si stimano con il metodo dei minimi quadrati sulla base degli $(n-n^*)$ valori validamente rilevati;

- per le variabili dicotomiche, il metodo della regressione logistica tra la probabilità di Y (per le n^* unità con dato mancante) e una o più variabili predittive:

$$\log \text{it}(p(y)) = \hat{b}_0 + \hat{b}_1 x_{1j} + \dots + \hat{b}_p x_{pj} \quad (j=1, \dots, n^*) \quad (10)$$

dove $\text{logit}(p)$ indica $\log(p/(1-p))$.

9.2 Correzione di dati inammissibili in un'indagine panel

In un'indagine longitudinale, una procedura per il controllo e la correzione di dati inammissibili deve rispondere esaurientemente ai tre quesiti seguenti:

1. *“In quale misura le informazioni raccolte sulla stessa unità statistica sono valide a fini di controllo della qualità delle singole risposte²⁵?”* Il quesito si può porre anche in un altro modo: *“Sono intervenuti dei fatti rilevanti che non permettono l'estrapolazione a fini previsivi delle informazioni raccolte presso l'unità statistica j nelle occasioni di rilevazione precedenti?”*
2. *“In quale misura le variazioni intervenute, tra le occasioni t-1 e t, nella variabile Y presso la categoria cui appartiene l'unità statistica j si applicano alla stessa unità?”*. Naturalmente, prima di porsi il quesito, è necessario classificare le unità in categorie consolidate a fini previsionali.
3. *“Come distinguere, tra i valori che superano determinate soglie individuati da un controllo di verosimiglianza, gli outliers e i dati affetti da errori di rilevazione?”*
4. *“Quale criterio è applicabile per correggere la risposta y_j , di cui si è constatata l'inammissibilità, o per ottenere on line informazioni suppletive sul probabile va-*

²⁵ Va ricordato che, nelle indagini longitudinali, si assiste allo strano fenomeno della “distorsione del gruppo di rotazione”, ossia alla propensione a rispondere con diversa accuratezza in ragione del numero di occasioni in cui il rispondente è interpellato. Bailar (1975) trova che la proporzione di disoccupati nell'indagine sulle forze di lavoro statunitense alla prima occasione di rilevazione è superiore a quella rilevabile alla seconda e, gradatamente, alle occasioni terza e quarta. Il fenomeno si rinnova, in proporzione diversa, dopo ogni rientro nel campione. In altri contesti, non è stato constatato lo stesso fenomeno.

lore di Y_j ?”. Il quesito si può porre anche nella forma seguente: “*Quale modello statistico è adeguato per la previsione, essendo in grado di bilanciare la necessità di descrivere il più dettagliatamente possibile le relazioni causali tra le variabili e mantenere una struttura semplice*²⁶?”

Siccome la maggior parte delle relazioni tra variabili non sono calcolabili prima di raccogliere i dati, per valutare l’ammissibilità delle risposte si può:

- A. *ipotizzare, per ciascuna variabile, il campo di variazione (per variabili su qualsiasi scala) o il valore probabile (per variabili quantitative o dicotomiche) consequenziale alle osservazioni raccolte nelle precedenti $t-1$ occasioni presso l’unità statistica j ;*
- B. *ipotizzare, per ciascuna variabile, il sistema di regole di ammissibilità della modalità y_{jt} . Le regole saranno sia di coerenza rispetto ad altre caratteristiche rilevate nella stessa occasione t d’indagine, sia di coerenza logica rispetto alla sequenza di modalità osservate nell’insieme delle t occasioni, sia di verosimiglianza rispetto alle precedenti occasioni di rilevazione e ad altre informazioni correlate.*

Se la risposta fornita nell’occasione di rilevazione t , y_{jt} , è poco verosimile, data la serie delle $t-1$ risposte ottenute precedentemente, Y_{jt} ($t = 1, \dots, t-1$), la procedura della conciliazione forza la modalità \hat{y}_{jt} che rende la sequenza di osservazioni nuovamente verosimile, ovvero con probabilità di osservazione congiunta sulle t occasioni superiore alla soglia prefissata.

La procedura dello scandaglio si attua, invece, ponendo al rispondente domande che mirano a capire i motivi della mancata verosimiglianza della sequenza di osservazioni: gli si ricorda, ad esempio, che ha già collaborato in altre occasioni e gli si chiede se siano intervenute variazioni per un certo fenomeno dall’ultimo contatto, ed, eventualmente, quali.

Le informazioni raccolte in più occasioni di rilevazione possono essere utilizzate per at-

meno (Corradi *et al.*, 1991)

²⁶ Nel lavoro di Bassi e Fabbris (2000) che hanno applicato il controllo di verosimiglianza ad una indagine sui viaggi degli italiani, il quesito si pone nel modo seguente: “Quali valori dei parametri utilizzare nell’impiego in linea del modello? Con riferimento all’applicazione (...), bisogna valutare se è più opportuno utilizzare i coefficienti del modello di regressione completo o quelli del modello in cui si sono tagliate le code.

tuare la procedura dello scandaglio con i seguenti strumenti (Saris, 1991, Dibbs *et al.*, 1995):

1. gli “schermi dinamici”, ossia schermate di dati alcuni dei quali sono modificabili dal rispondente. Le schermate sono presentabili convenientemente in forma di finestra, entro riquadri posti in posizione marginale se sono informazioni ausiliarie, o a pieno schermo se sono dati che si chiede alla persona che risponde di confermare o modificare opportunamente;
2. i “calendari dinamici”, ossia calendari entro i quali l’interpellato apporrà i propri dati modificati nelle intestazioni in funzione delle caratteristiche spaziali e temporali dell’indagine e delle caratteristiche del panelista e/o della sua famiglia. Il calendario può contenere una riga in cui si riportano eventi che possono essere utili per aiutare il rispondente nella collocazione temporale delle notizie richieste e nel sollecitare la memoria. Tra gli eventi che si possono segnalare ci sono, ad esempio, i compleanni dei membri della famiglia, ed altri che dipenderanno anche dal tipo di informazioni che si stanno raccogliendo e dalla cadenza temporale del fenomeno cui si fa riferimento.

10. Il recupero di informazioni per aggiustare le stime

Nel seguito, si descrivono le informazioni qualitative e quantitative utilizzabili non per la correzione in tempo reale della singola risposta, bensì per aggiustare le stime a posteriori e per svolgere eventuali analisi degli errori al fine di guidare le future indagini dello stesso tipo.

Il rischio più grave per le stime è costituito dalla distorsione, ossia dalla differenza sistematica tra il valore atteso dei dati rilevati e il valore vero della stima. Il rischio di distorsione nasce soprattutto dalle mancate risposte, dato che coloro che non rispondono sono, per molte variabili cruciali, differenti da coloro che collaborano all’indagine pienamente. Per questo, in relazione ai mancati rispondenti totali, è opportuno:

- raccogliere dati ufficiali di tipo ascrittivo al fine di determinare le frequenze o le medie delle categorie di mancati rispondenti e correggere con la tecnica della stima basata su

quozienti (tra le frequenze dei rispondenti e quelle del campione completo) per le categorie ascrittive individuate la stima che interessa;

- rilevare, là dove è possibile, la dimensione dell'unità, caratteristica che permette di stimare con discreta accuratezza molte grandezze correlate alla dimensione, sia per gruppi di unità, sia per la stima di frequenze o grandezze inerenti all'intero campione.

Inoltre, si sa che l'errore del rilevatore può, tutt'al più, essere contenuto e che i contenuti e il disegno della rilevazione interagiscono con le caratteristiche del rilevatore nel determinare le variabili su cui si manifesta e l'entità con cui si manifesta l'errore del rilevatore. La costruzione di una banca dati sui rilevatori collegabile a quella dei dati dell'indagine principale, da utilizzare per stabilire le caratteristiche maggiormente correlate alle caratteristiche del rilevatore, può dare indicazioni sia per la stessa indagine²⁷, sia per indagini future su argomenti analoghi.

In ogni indagine vanno, inoltre, rilevate informazioni sul contesto in cui si è svolta la rilevazione, sulle caratteristiche del rispondente, sul grado di accettazione di singoli quesiti o di insiemi di quesiti, nonché dell'intera indagine da parte di ciascun rispondente. Queste variabili si sono dimostrate, in vari contesti di ricerca, importanti per prevedere mancate risposte ed errori di risposta.

I dati qualitativi da recuperare presso i rispondenti per progettare altri questionari e altre rilevazioni riguardano:

- informazioni a latere sulle risposte molto improbabili o elusive sulle quali sia stato realizzabile solo lo scandaglio e non la conciliazione,
- suggerimenti e altre indicazioni qualitative sul modo di porre i quesiti, sul contenuto della rilevazione, sulle caratteristiche tecniche dell'indagine.

²⁷ Si può azzardare un utilizzo dei dati sui rilevatori anche per correggere le stime campionarie con tecniche di stima basate sulla regressione.

11. Indicazioni propositive

La distinzione tra controlli di compatibilità e controlli di verosimiglianza delineata in questo lavoro ha lo scopo di evidenziare le potenzialità specifiche pertinenti a ciascuna delle due logiche di costruzione di controlli in linea.

I controlli di compatibilità risultano di indubbia utilità per l'eliminazione di errori evidenti nei dati raccolti (incoerenze tra informazioni, valori esterni all'intervallo di ammissibilità), causati principalmente da disattenzione da parte del rispondente o da errori di digitazione degli addetti.

I controlli di verosimiglianza dimostrano grandi potenzialità per migliorare la qualità dei dati in quanto permettono di individuare errori meno palesi, ma non meno gravi di quelli di incoerenza, se portano a distorsioni delle stime. La natura probabilistica di questi controlli, le assunzioni in termini di distribuzione delle variabili e, spesso, anche di comportamento dei rispondenti, richiedono che essi siano messi in opera solo dopo accurata riflessione.

Come è intuibile dalla uniformità della simbologia adottata nella presentazione, i controlli di compatibilità e di verosimiglianza sono specificazioni di un'unica azione di ammissibilità dei dati rilevati. La rilevazione *computer assisted* rende possibile l'anticipazione del controllo di ammissibilità del dato al momento del colloquio che si svolge tra l'interpellato e il computer o un altro tramite visivo o sonoro (questionario autosomministrato) o tra l'interpellato e un intervistatore (rilevazione tramite intervistatori).

Alcune cause d'errore non sono eliminabili neppure con un sistema assistito da computer. In particolar modo, le ragioni che inducono gli interpellati a non collaborare del tutto o in parte all'indagine, o a reagire in modo negativo a quesiti imbarazzanti o ansiogeni, non sono eliminabili se non si trovano criteri adeguati di approccio al rispondente, di proposizione dei quesiti e delle modalità di risposta. Il ripensare i modi di svolgimento dell'indagine statistica nell'ottica della rilevazione *computer assisted* è un'area di studio che si propone da quando l'avvento del computer e dei sistemi telematici ha cambiato in modo irreversibile la rilevazione.

I controlli a posteriori restano importanti, anche perché gli errori sono così sfuggenti, così diversi nelle varie sottopopolazioni in relazione a tipi di quesiti, e così mutevoli nel tempo in funzione delle modificazioni culturali, che l'agire per ottimizzare la rilevazione deve sempre accompagnarsi alla verifica a posteriori, soprattutto per scoprire eventuali distorsioni nei dati rilevati.

È in uno spirito positivo, di sollecitazione di una continua ricerca delle vie più adeguate per definire la metodologia dell'indagine statistica, che concludiamo questa prima parte del rapporto alla Commissione per la Garanzia dell'Informazione Statistica.

Riferimenti bibliografici

- Abbate C. (1997) La completezza delle informazioni e l'imputazione da donatore con distanza minima, *Quaderni di Ricerca*, ISTAT, n. 4.
- Allard B., Brisebois F., Dufour J., Simard M. (1996) How do interviewers do their job? A look at new data quality measures for the Canadian labour force survey, *International Conference on "Computer-assisted survey information collection"*, San Antonio, Texas.
- Appel M.V., Cole R. (1994) Spoken language recognition for the year 2000 census questionnaire, *American Statistical Association for Public Opinion Research Annual Conference*, Danvers, MA.
- Baker, R.P. (1987) Information systems in survey research, *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, Washington: 166-177.
- Baker R.P., Bradburn N.M., Johnson R.A. (1995) Computer-assisted personal interviewing: an experimental evaluation of data quality and cost, *Journal of Official Statistics*, **11**, 413-431.
- Bailar B.A. (1975) The effect of rotation group bias on estimates from panel surveys, *Journal of the American Statistical Association*, **70**: 3-30.
- Bailar B.A. (1975) The effect of rotation group bias on estimates from panel surveys, *Journal of the American Statistical Association*, **70**: 3-30.
- Bailar B.A., Baily L., Stevens J. (1977) Measures of interviewer bias and variance, *Journal of Marketing Research*, **14**: 337-343.
- Balbi S., Verde R. (1999) Una strategia di imputazione per mancate risposte in questionari strutturati come oggetti simbolici. *Intervento al convegno della SIS dal titolo "Ingegnierizzazione del processo di produzione dei dati statistici"*, Firenze 9-4-1999.
- Balbi S., Verde R. (2000) Una struttura simbolica per il controllo della coerenza del questionario. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 191-206.
- Baldassarri E. Guida all'utilizzo della procedura di editing ed imputazione dei dati dell'indagine sulle Forze di Lavoro. *Servizio FIL/C. DOCUMENTO INTERNO*, ISTAT, Roma.
- Barcaroli G. (1993) Un approccio logico formale al problema del controllo e della correzione dei dati statistici, *Quaderni di Ricerca*, ISTAT, n. 9.
- Barcaroli G., D'Aurizio L., Luzi O., Manzari A., Pallara A. (1999) Metodi e software per il controllo e la correzione dei dati. *Documenti ISTAT*, n.1, ISTAT, Roma.
- Barcaroli G., Di Pietro E., Venturi M. (1993) L'applicazione della metodologia Fellegi-Holt per il controllo e la correzione dei dati relativi alla nuova indagine trimestrale sulle

- forze di lavoro, *Quaderni di Ricerca, ISTAT*, n. 9.
- Barcaroli G., Venturi M. (1997) The probabilistic approach to automatic edit and imputation: improvements of the Fellegi-Holt methodology, *Quaderni di Ricerca, ISTAT*, n. 4.
- Bassi F., Fabbris L. (1994) L'errore statistico nella produzione di microdati e macrodati. In: Colombo B., Cortese A., Fabbris L. (a cura di) *La produzione di statistiche ufficiali*, CLEUP, Padova: 247-266.
- Bassi F., Fabbris L. (2000) Controlli di verosimiglianza in linea in una rilevazione CASI sui viaggi degli italiani. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 171-190.
- Batcher M., Scheuren F. (1997) CATI site management in a survey of service quality. In: Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwartz N., Trewin D. (a cura di) *Survey Measurement and Process Quality*, Wiley, New York: 573-588.
- Bates N. (1996) Reinterviews and reconciliation using CAPI: the integrated coverage measurement (ICM) interview, *International conference on "Computer-assisted survey information collection"*, San Antonio, Texas.
- Benelmans-Spork M.E.J., Sikkel D. (1985) Data collection with hand-held computers, *Proceedings of the 45th Session*, International Statistical Institute, Book III, topic 18.3.
- Bernard C. (1989) *Survey data collection using labtop computers*, INSEE Report n. 01/C520, Paris.
- Bethlehem J.G. (1987) The Data Editing Research Project. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 55-66.
- Bethlehem, J.G. (1992) A new approach to statistical information processing, *CBS-report*, Central Bureau of Statistics, Voorburg, The Netherlands.
- Bethlehem J.G., Denteneer D., Hundepool A.J., Keller W.J. (1987) *The BLAISE Reference Manual*, Voorburg, The Netherlands (internal CBS Report).
- Bethlehem, J.G., Keller, W.J. (1992) The BLAISE system for integrated survey processing, *Survey Methodology*, **17**: 43-56.
- Blom, E. (1994) Building Integrated Systems of CASIC Technologies at Statistics Sweden, *Proceedings of the Annual Research Conference and CASIC Technologies Interchanges*, U.S. Bureau of the Census: 623-634.
- Blyth B. (1997) Developing a speech recognition application for survey research. In: Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwarz N., Trewin D. (eds), *Survey measurement and process quality*, New York, J. Wiley & Sons: 249-266.
- Blyth W.G., Piper H. (1994) Speech recognition - A new dimension in survey research,

- Journal of the Market Research Society*, **36**: 183-204.
- Brakenhoff W.J., Remmerswaal P.W.M., Sikkel D. (1987) A test of the Netherlands Continuous Labour Force Survey with hand-held computers: Interviewer behaviour and data quality. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 13-26.
- Capiluppi, C. (2000). Il sistema CAPTOR. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 225-240.
- Chiaro M. (1996) *I sondaggi telefonici*, Roma, CISU.
- Clayton R.L., Harrel L.J. (1989) Developing a cost model for alternative data collection methods: mail, CATI, and TDE, *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 264-269.
- Corradi F., Fabbris L., Sanetti I., Zuliani A. (1991) Proposte in tema di stime tempestive dei disoccupati. In: Trivellato U. (a cura di) *Forze di lavoro: disegno dell'indagine e analisi strutturali*, Annali di Statistica, Anno 120, serie IX, Vol. 11, ISTAT, Roma: 83-98.
- Couper M.P. (1996) Changes in interview setting under CAPI, *Journal of Official Statistics*, **12**.
- Couper M.P., Hansen S.E., Sadosky A.A. (1997) Evaluating interviewer use of CAPI technology. In: Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwartz N., Trewin D. (a cura di) *Survey Measurement and Process Quality*, Wiley, New York: 267-285.
- Dibbs, R., Hale, A., Loverock, R., Michaud S. (1995). Some effects of computer-assisted interviewing on the data quality of the survey of labor and income dynamics, *Proceedings of the Conference on Survey Measurement and Process Quality (Bristol) - Contributed Papers*: 174-177.
- Duncan G.J. (1992) Household panel studies: prospect and problems, Survey Research Center, University of Michigan, Ann Arbor, Michigan.
- Duncan G.J., Kalton G. (1987) Issues of design and analysis of surveys across time. *International Statistical Review*, **55** (1): 97-117.
- Fabbris L. (1983) Una esperienza di stima dell'errore non campionario mediante reintervista e compenetrazione dell'assegnazione degli intervistatori. In: *Atti del Convegno 1983, Società Italiana di Statistica*, Vol. I, Litografia Ricci, Trieste: 515-531.
- Fabbris L. (1989) *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*, La Nuova Italia Scientifica, Roma.
- Fabbris L. (1994) *L'error profile* di un'indagine statistica, In: Colombo B., Cortese A., Fabbris L. (a cura di) *La produzione di statistiche ufficiali*, CLEUP, Padova: 287-290.

- Fabbris L. (1999) Rilevazione di dati assistita da computer per via telematica nelle indagini longitudinali prospettiche su famiglie e imprese (documento interno, ISTAT, Roma).
- Fabbris L. (a cura di) (2000) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova.
- Fabbris L., Bassi F. (1997) On-line Likelihood Controls in Computer-assisted Interviewing. In: *Proceedings Invited Papers (Istanbul, August 18-26 1997)*, International Statistical Institute: 38-47 (also published in a reduced version in: *Bulletin of the International Statistical Institute*, **LVII (1)**: 515-518.
- Fabbris L., Martini M.C. (1999) Programmazione efficiente delle chiamate in indagini CATI sulle famiglie italiane. In: Società Italiana di Statistica, *Atti del convegno "Verso i Censimenti del 2000, Raccolta degli abstract e delle comunicazioni spontanee (Udine, 7-9 giugno 1999)*: 55-57.
- Fellegi I.P. (1964) Response variance and its estimation, *Journal of the American Statistical Association*, **59**: 1016-1041.
- Fellegi I.P. (1974) An improved method of estimating the correlated response variance, *Journal of the American Statistical Association*, **69**: 496-501.
- Fellegi I.P., Holt D. (1976) A systematic approach to automatic editing & imputation, *Journal of the American Statistical Association*, **71**: 17-35.
- Fellegi I.P., Sunter A.B. (1969) A theory for record linkage, *Journal of the American Statistical Association*, **64**.
- Filippucci C. (2000). La rilevazione di dati mediante computer presso le famiglie: prospettive e problemi sulla base di alcune esperienze italiane. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 1-32.
- Filippucci C., Drudi I., Ferrante R. (2000) Le tecniche assistite da computer nelle indagini sui consumi in Italia. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 69-94.
- Fisher B., Margolis M., Resnick D., Bishop G. (1995) Survey research in cyberspace: breaking ground on the virtual frontier, *Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol.
- Foxon J. (1987) Field trials in the Labour Force Survey Lab-top Computer Project, *Survey Methodology*, **12**: 12-25.
- Ghellini G., Pannuzi N. (1996) Non risposta ed errori di risposta in indagini panel: strategie per il controllo di qualità, la ponderazione e l'imputazione. In: Società Italiana di Statistica - *Atti della XXXVIII riunione scientifica (Rimini, 9-13 aprile 1996)*, **1**: 219-229.

- Green T.M. (1996) An investigations of response effects for responders and refusers in an on-line organizational survey, *International conference on "Computer- assisted survey information collection"*, San Antonio, Texas.
- Groves R.M., Kahn R. (1979) *Surveys by Telephone: A National Comparison with Personal Interviews*, Academic Press, New York.
- Groves R.M., Magilavy L.J. (1986) Measuring and explaining interviewer effects in centralized telephone surveys, *Public Opinion Quarterly*, **55**: 251-266.
- Groves, R.M., Nicholls, W.L.II (1986). The status of computer-assisted telephone interviewing: Part II. Data quality issues, *Journal of Official Statistics*, **2**, 117-134.
- Groves R.M., Tortora R. (1996) Integrating CASIC into existing designs and organizations: a survey of the field, *International conference on "Computer- assisted survey information collection"*, San Antonio, Texas.
- Hanson R.H., Marks E.S. (1958) Influence of the interviewer on the accuracy of survey results, *Journal of the American Statistical Association*, **53**: 635-655.
- Hardie E.T.L., Neou V. (a cura di) (1994) *Internet Mailing Lists, 1994 Edition*, Englewood Cliffs, Prentice-Hall.
- House C.H. (1985) Questionnaire design with computer assisted telephone interviewing, *Journal of Official Statistics*, **1**: 209-220.
- House, C.C., Nicholls, W.L.II (1988). Questionnaire design for CATI: design objectives and methods. In: R. Groves *et al* (eds.) *Telephone Survey Methodology*, Wiley, New York, 421-436.
- ISTAT (1989) *Manuale di tecniche d'indagine. Volume 6: Il sistema di controllo della qualità dei dati*, Note e relazioni, n.1, ISTAT, Roma.
- Jenkins C.R., Dillman D.A. (1997) Towards a theory of self-administered questionnaire design. In: Lyberg L, Biemer P., Collins M., De Leeuw E., Dippo C., Schwarz N., Trewin D (eds) *Survey Measurement and Process Quality*, J. Wiley & Sons, New York. 165-196.
- Kalton G. (1986) Handling wave nonresponse in panel survey. *Journal of Official Statistics*, **2** (3): 303-314.
- Keller W., Metz K.J., Bethlehem J.G. (1990) The impact of microcomputers on survey processing at the Netherlands Central Bureau of Statistics, *Proceedings of the Annual Research Conference, US Bureau of the Census*: 637-675.
- Kent J.-P. (2000) BLAISE: experiences, recent improvements, possible developments. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 55-68.
- Kerssemakers F.A.M., De Mast F.A.C., Remmerswaal P.W.M. (1987) Computer assisted telephone interviewing, some response findings. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The

- Netherlands: 119-131.
- Kish L. (1962) Studies of interviewer variance for attitudinal variables, *Journal of the American Statistical Association*, **57**: 92-115.
- Kulka R.A., Weeks M.F. (1988) Toward the development of optimal calling protocols for telephone surveys: a conditional probability approach, *Journal of Official Statistics*, **4**: 319-332.
- Lindström H.L. (1995) The touch-tone data entry experiment in the Swedish producer price index survey 1993, *Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol.
- Lyberg L. (1985) Plans for computer assisted data collection at Statistics Sweden, *Bulletin of the International Statistical Institute*, Book 3.
- Mahalanobis I.P. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute, *Journal of the Royal Statistical Society*, **109**: 325-370.
- Manners T., Diamond A. (1994) Integrated field and office editing in BLAISE: OPCS's Experience of complex financial Surveys, *Proceedings of the Annual Research Conference and CASIC Technologies Interchange, U.S. Bureau of the Census*: 723-737.
- Martin J. (1993) From PAPI to CAPI: the OPCS experience, *Essays on BLAISE, Proceedings of the Second International BLAISE Users Conference*: 96-117.
- Martin J., O'Muircheartaigh C., Curtice J. (1993) The use of CAPI for attitude surveys: an experimental comparison with traditional methods, *Journal of Official Statistics*, **9**: 641-661.
- Matchett S.D., Creighton K.P., Landman C.R. (1994) Building integrated system of CASIC technology at the US Bureau of the Census, *Proceedings of the Annual Research Conference, US Bureau of the Census*: 573-595.
- McDonald J.L. (1996) The Bureau of Census, economic directorate's use of computerized self-administered questionnaires, *International Conference on "Computer-assisted survey information collection"*, San Antonio, Texas.
- McKay R.B., Robinson E.L. (1994) Touch-tone data entry for CPS sample expansion, *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 509-511.
- Metz K.J. (1987) Decentralization in statistical data processing: experiences with mini- and microcomputers. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 133-144.
- Miller V.P., Cannell C.F. (1982) A study of experimental techniques for telephone interviewing, *Public Opinion Quarterly*, **46**: 250-269.
- Nicholls W.L.II (1978) Experiences with CATI in a large-scale survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 9-17.

- Nicholls W.L.II, Appel M.A. (1994) New CASIC technologies at the Bureau of the Census, *Proceedings of the Section on Survey Methods, American Statistical Association*: 757-762.
- Nicholls W.L.II, Baker R.P., Martin J. (1997) The effects of new data collection technologies on survey data quality. In: Lyberg L, Biemer P., Collins M., De Leeuw E., Dippo C., Schwarz N., Trewin D. (eds) *Survey Measurement and Process Quality*, J. Wiley & Sons, New York: 221-248.
- Nicholls II W.L., Groves, R.M. (1986) The status of computer-assisted telephone interviewing. Part I: Introduction and impact on cost and timeliness of survey data, *Journal of Official Statistics*, **2**: 93-115.
- Nicholls W.L.II, Kindel K.K. (1993) Case management and communications for CAPI, *Journal of Official Statistics*, **9**, 623-639.
- O'Muirheartaigh C. (1989) Sources of nonsampling error: discussion, In: Kasprzyk D., Duncan G., Kalton G., Singh M.P. (a cura di) *Panel Survey*, Wiley, New York: 271-287.
- O'Muirheartaigh C.A. (1997) Measurement error in surveys: a historical perspective. In: Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwartz N., Trewin D. (a cura di) *Survey Measurement and Process Quality*, Wiley, New York: 1-28.
- O'Reilly, J., Hubbard, M., Lassler, J., Biemer, P. & Turner, C. (1994). Audio and video computer assisted self-interviewing: preliminary test and new technologies for data collection, *Journal of Official Statistics*, **10**, 197-214.
- Pannekoek J. (1987) Interviewer variance in a telephone survey. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 107-117.
- Piazza T. (1997) New methodological possibilities offered by computer assisted interviewing, *Bulletin of the International Statistical Institute - 51st Session - Proceedings*, **1**, 511-514.
- Pitkow J. (1995) Using the Web as a survey tool: result from the second WWW user survey, <http://www/cc/gatech/edu/cogsci/faculty/recker.html>.
- Pratesi M. (2000) Gestione automatica delle interviste e modelli per non risposta nelle indagini telefoniche. In: Fabbris L. (a cura di) *Il questionario elettronico. Metodi e tecniche per le rilevazioni assistite da computer*, CLEUP editore, Padova: 147-158.
- Quarterman J. S. (1994) Preliminary partial results of the second TIC/MIDS internet demographic survey, *Matrix News*, **4**, n.12.
- Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995) La metodologia di editing e imputazione per variabili qualitative implementate in SCIA (documento interno), ISTAT, Roma.
- Santi E., Melcarne M.D., Fabbris L., Bosisio P. (1997) L'effetto dell'intervistatore in una

- rilevazione diretta della qualità di servizi comunali. In: Corsi M., Fabbris L., Franci A. (a cura di) *La valutazione della qualità dei servizi socio-assistenziali*, CLEUP, Padova: 163-170.
- Saris, W.E. (1991). *Computer-assisted Interviewing*, Sage, Newbury Park.
- Saris W.E., De Pijper W.M. (1986) Computer assisted interviewing using home computers, *European Research*, **14**: 144-150.
- Schuerhoff, M.H. (1993) BLAISE as a statistical control centre, *Bulletin of the International Statistical Institute*, **2**: 273-282.
- Shanks J.M., Tortora R. (1985) Beyond CATI: generalized and distributed systems for computer assisted surveys, *Proceedings of the First Annual Research Conference of the Bureau of the Census*, Bureau of the Census, Washington D.C.: 358-377.
- Shing C.C., Chu S.C. (1996) A web-based intelligent survey tool, *International Conference on "Computer-assisted survey information collection"*, San Antonio, Texas.
- Stol, H.R. (1993) An architecture for EDI in business surveys based on the use of BLAISE, *Essays on BLAISE, Proceedings of the Second International BLAISE Users Meeting*: 143-153.
- Tortora R. (1985) CATI in an agricultural statistical agency, *Journal of Official Statistics*, **1**: 301-314.
- Tucker C. (1983) Interviewer effects in telephone surveys, *Public Opinion Quarterly*, **47**: 84-95.
- Turner, M.J. (1994) General survey processing software: an architectural model, *Proceedings of the Annual Research conference and CASIC Technologies Interchange*, U.S. Bureau of the Census: 596-606.
- Van Bastelaer A.M.L., Sikkel D. (1987) From Three to Three Hundred Hand-held Computers. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 27-36.
- Van Bastelaer A.M.L., Kerssemakers F.A.M., Sikkel D. (1987) A Test of the Netherlands Continuous Labour Force Survey with Hand-held Computers: Interviewer Behaviour and Data Quality. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4, Voorburg, The Netherlands: 37-54.
- Weeks M.F. (1992) Computer-assisted survey information collection: A review of CASIC methods and their implication on survey operations, *Journal of Official Statistics*, **8**: 445-465.
- Werking G.S., Clayton R.L. (1990) Estimating the quality of time critical estimates through the use of mixed mode CATI/CASI collection, *Measurement and Improvement of Data Quality*, Proceedings of Statistics Canada Symposium 93, October.
- Willemborg L.C.R.J. (1987) The routing structure of the questionnaire. In: Central Bureau Voor de Statistiek (a cura di) *Automation in Survey Processing*, CBS Select, 4,

Voorburg, The Netherlands: 97-106.

Winter D.L.S., Clayton R.L. (1990) Speech data entry: results of the first test of voice recognition for data collection, *Annual Conference of the American Statistical Association*, Arnheim.

SERIE “RAPPORTI DI RICERCA”

- 93.01 Valutazioni di procedure di oscuramento delle informazioni individuali e di canoni di pubblicazione di informazioni a minimo rischio di individuazione, (*M. Angrisani*)
- 93.02 Gli investimenti pubblici: problemi di contabilità pubblica e di contabilità nazionale, (*G. Trupiano*)
- 93.03 Investimenti pubblici lordi e netti: problemi analitici, (*V. Selan*)
- 93.04 L'indice dei prezzi al consumo in Italia, (*F. Franceschini, G. Marliani, M. Martini*)
- 94.01 Privatizzazione e sistema statistico nazionale, (*G. Di Gaspare*)
- 94.02 Stato delle statistiche sociali in Italia, (*G.B. Sgritta*)
- 94.03 Statistica sociale e Statistiche sociali, (*L. Bernardi*)
- 94.04 Prospettive preliminari per possibili analisi longitudinali nella statistica ufficiale italiana, (*U. Trivellato, G. Ghellini, C. Martelli, A. Regoli*)
- 94.05 Analisi di alcune caratteristiche del Programma Statistico Nazionale 1995-1997, (*D. Cotzia, S. D'Andrea, E. Mastantuoni*)
- 94.06 Verifica dei ritardi rispetto alle previsioni di stampa delle pubblicazioni ISTAT negli anni 1993 e 1994, (*D. Cotzia*)
- 94.07 Analisi sulla tempestività della Produzione di informazione statistica (Esame di alcune rilevazioni ed elaborazioni dell'Istat), (*D. Cotzia*)
- 94.08 La suddivisione territoriale della spesa pubblica per investimenti, (*G. Trupiano*)
- 94.09 Il consolidamento della spesa pubblica per investimenti, (*G. Trupiano*)
- 94.10 Investimenti netti, ammortamenti e spese di manutenzione. Stock di capitale: un'ipotesi censuaria, (*V. Selan*)
- 94.11 Le spese per investimenti nelle statistiche Eurostat sui conti delle amministrazioni pubbliche, (*M. Colazingari*)
- 94.12 Gli investimenti pubblici del Comune di Roma, (*P. Palmarini*)
- 94.13 La revisione del Sistema dei Conti Nazionali: problemi e prospettive per

- l'Italia, (B. Bracalente, G. Carbonaro, M. Carlucci, M. Di Palma, L. Esposito, G. Ferrari, R. Zelli)*
- 94.14 La tutela della riservatezza e l'identificazione dei rispondenti alle rilevazioni statistiche svolte nell'ambito del Sistan: rapporto preliminare, *(M. Angrisani, L. Buzzigoli, A. Giusti, L. Grassini, G. Marliani)*
- 94.15 I dati statistici produttivi di effetti giuridici determinati e la loro sindacabilità, *(G. Manto)*
- 94.16 Ufficialità del dato e Programma Statistico Nazionale, *(G. D'Alessio)*
- 94.17 Valutazioni preliminari sulla qualità dei dati dell'ultimo censimento generale della popolazione e delle abitazioni, *(G. De Santis, A. Bonaguidi, A. Santini)*
- 94.18 La revisione del Sistema dei Conti Nazionali: problemi e prospettive per l'Italia - rapporto finale, *(B. Bracalente, G. Carbonaro, M. Carlucci, M. Di Palma, L. Esposito, G. Ferrari, R. Zelli)*
- 95.01 Classificazione delle province italiane in clusters e determinazione delle province outliers in riferimento alle correzioni degli errori di coerenza e di range del censimento dell'agricoltura 1991, *(S. D'Andrea)*
- 95.02 La qualità dei dati dell'ultimo censimento generale della popolazione e delle abitazioni, *(G. De Santis, S. Salvini, A. Santini)*
- 95.03 Stato delle Statistiche sociali in Italia - Sintesi del rapporto, *(G. B. Sgritta)*
- 95.04 Lo Stato dell'informazione statistica nei comuni e negli altri enti territoriali intermedi del Sistan: le province di Ferrara e Siena, *(A. Buzzi Donato, I. Drudi, M.R. Ferrante, C. Filippucci, G. Gesano, G. Ghellini, T. Giovani, A. Lemmi)*
- 95.05 Analisi delle funzioni del Sistema di Informazione Geografica-GISCO della Commissione delle Comunità Europee, *(E. Mastantuoni)*
- 95.06 Stato ed evoluzione delle statistiche ambientali in Italia, *(L. Fabbris, M. Lo Cascio)*
- 95.07 Rapporto sugli aspetti statistici nella Legislazione Ambientale - I. Aria, *(S. Bordignon, A. C.S. Capelo, G. Lovison, G. Masarotto)*
- 95.08 Il Sistema Statistico delle Imprese in Italia: rapporto preliminare, *(S. Biffignandi, M. Pratesi, T. Proietti, L. Schionato)*

- 95.09 Prospettive per possibili analisi longitudinali nella statistica ufficiale italiana, (*U. Trivellato, G. Ghellini, C. Martelli, A. Regoli*)
- 95.10 Per una estensione dei compiti della Commissione per la Garanzia dell'informazione statistica, (*G. Calvi, M.T. Crisci, S. Draghi, L. Ferrari, A. Rizzi*)
- 95.11 Rapporto sugli aspetti statistici nella legislazione ambientale - II. Rumore, (*S. Bordignon, A. C.S. Capelo, G. Lovison, G. Masarotto*)
- 95.12 Innovazioni integrazioni nel sistema dei conti nazionali: Problemi aperti e soluzioni possibili - Sintesi e suggerimenti -, (*B. Bracalente, G. Carbonaro, M. Carlucci, M. Di Palma, L. Esposito, G. Ferrari, R. Zelli*)
- 95.13 Disaggregazione spaziale e temporale delle statistiche ufficiali sulla qualità dell'aria, (*L. Fabbris*)
- 95.14 Disaggregazione spaziale e temporale delle statistiche ufficiali sulla qualità delle acque, (*L. Fabbris*)
- 95.15 L'esercizio della funzione statistica a livello locale: lo stato degli uffici di statistica comunali dopo il d.lgs. n.322/89, (*G. Manto*)
- 95.16 Gli uffici di statistica dei Ministeri, (*C. Gallucci*)
- 95.17 Le statistiche comunitarie e le statistiche nazionali: evoluzione, coordinamento, integrazione e processi di uniformazione, (*G. Di Gaspare*)
- 95.18 Organizzazione ed attività statistica delle regioni nel contesto del Sistan, (*G. D'Alessio*)
- 96.01 Rapporto sullo stato dell'informazione statistica nei comuni della provincia di Bari, (*C. Cecchi, V. Nicolardi, A. Pollice, N. Ribecco*)
- 96.02 Sistemi Nazionali di statistica: loro organizzazione e funzionamento in alcuni paesi dell'unione europea, (*B. Carelli*)
- 96.03 L'attività delle amministrazioni centrali dello Stato per il programma statistico nazionale del triennio 1996-98, (*G. Filacchione*)
- 96.04 Rapporto sugli aspetti statistici nella legislazione ambientale - III. Dati mancanti -, (*S. Bordignon, A.C.S. Capelo, G. Lovison, G. Masarotto*)
- 96.05 Osservatorio Statistico Locale: Studio di un modello per il Sistan, (*P. Bellini, S. Campostrini, T. Di Fonzo, M.P. Bellini*)

- 96.06 La tutela della riservatezza e l'identificazione dei rispondenti alle rilevazioni statistiche svolte nell'ambito del Sistan - rapporto finale, (*M. Angrisani, L. Buzzigoli, A. Giommi, A. Giusti, L. Grassini, G. Marliani*)
- 96.07 Analisi dell'organizzazione e delle iniziative del Sistan - Esame delle pubblicazioni presenti nel Catalogo Sistan 1994, (*A. De Nardo, S. Sagramora*)
- 96.08 Sistema Statistico delle Imprese, (*S. Biffignandi, M. Pratesi, T. Proietti, L. Schionato*)
- 96.09 Monitoraggio della diffusione dei dati riguardanti alcuni indicatori dell'Istat su prezzi, lavoro e commercio con l'estero, (*A. De Nardo, E. Mastantuoni, M. Notarnicola, S. Sagramora*)
- 96.10 Monitoraggio della qualità e tempestività dell'indice della produzione industriale, (*V. Napoli, F. Tagliafierro*)
- 96.11 La qualità dei dati del VII censimento dell'industria e dei servizi: alcune valutazioni dal punto di vista dell'utilizzatore, (*R. Guarini, R. Zelli*)
- 96.12 Analisi del processo di revisione corrente delle stime provvisorie dei dati del Commercio con l'Estero, (*E. Mastantuoni, S. Sagramora*)
- 96.13 Prime indagini sull'accesso ai dati statistici individuali nell'ambito del Sistan, (*L. Buzzigoli, C. Martelli, N. Torelli*)
- 97.01 Interconnessione di basi di dati: problemi di sfruttamento statistico, (*A. Cortese*)
- 97.02 La formazione statistica nelle amministrazioni dello Stato: profili comparativi ed elementi propositivi, (*F. Covino*)
- 97.03 Rapporto sull'autonomia degli uffici di statistica nelle amministrazioni centrali dello Stato, (*F. Covino*)
- 97.04 Rapporto sulle regioni e le province autonome nel sistema statistico nazionale, (*N. Belvedere*)
- 97.05 Il sistema statistico europeo. Stato attuale e possibile riforma, (*I. Savi*)
- 97.06 Rapporto preliminare sulla statistica in Francia e nel Regno Unito, (*E. Marotta*)
- 97.07 Verifica della programmazione nell'attività del Sistan e dell'attività di vigilanza, (*F. Bigazzi*)

- 97.08 Indagine sulle statistiche della Sanità, (*P. Golini*)
- 98.01 Evoluzione e prospettive della statistica comunitaria: un aggiornamento, (*I. Savi*)
- 98.02 L'incidenza sul SISTAN delle leggi di riforma amministrativa e della disciplina in materia di privacy, (*N. Belvedere, I Savi*)
- 98.03 Analisi sullo stato di attuazione degli uffici di statistica dei comuni. Analisi preliminari e progetto di rilevazione, (*A. De Nardo, M. Notarnicola*)
- 98.04 Documentazione statistica su fenomeni di emarginazione sociale: offerta e fabbisogni: Tossicodipendenze, (*B. Colombo, G. Filacchione*)
- 98.05 Analisi delle caratteristiche dei non rispondenti con riferimento alle principali indagini campionarie sulle famiglie condotte dall'ISTAT, (*E. Mastantuoni, S. Sagramora, F. Tagliafierro*)
- 98.06 La razionalizzazione della statistica giudiziaria, (*F. Giusti, S. Andreano, M. Fabri, V. Napoli, R. Santoro*)
- 99.01 Validità e qualità degli indici dei prezzi al consumo. *Atti del Seminario, Roma, 12 dicembre 1997*
- 99.02 Analisi della disponibilità delle statistiche di genere, (*M.E. Graziani*)
- 99.03 La razionalizzazione della statistica giudiziaria - Rapporto finale, (*F. Giusti, S. Andreano, M. Fabri, V. Napoli, R. Santoro*)
- 99.04 Le procedure di destagionalizzazione di serie storiche economiche: esperienze internazionali e pratica nell'ambito dell'Istat, (*T. Di Fonzo, B. Fischer, T. Proietti*)
- 99.05 Lo stato dell'informazione statistica sul lavoro, con particolare riguardo alla partecipazione al lavoro ed a retribuzioni e costo del lavoro, (*G. Faustini, E. Rettore, P. Sestito*)
- 99.06 Analisi delle caratteristiche dei non rispondenti con riferimento alle principali indagini campionarie sulle famiglie condotte dall'Istat, (*E. Mastantuoni, S. Sagramora*)
- 99.07 Statistiche dei rifiuti, (*L. Fabbri, G. Nebbia*)
- 99.08 Problemi di adeguamento della legislazione italiana alla normativa comuni-

taria e internazionale sulla tutela della riservatezza di dati personali utilizzati per finalità statistiche, (*N. Belvedere, I. Savi, F. Tufarelli*)

- 99.09 Stato di attuazione degli uffici di statistica dei comuni, (*A. De Nardo, M. Notarnicola*)
- 99.10 Il confronto tra censimento ed anagrafe: per un maggior grado di coerenza tra le due fonti, (*L. Ciucci, G. De Santis, M. Natale, M. Ventisette*)
- 99.11 Censimenti economici e schedari delle imprese, (*R. Castellano, C. Quintano, G. Screpis, F. Tassinari*)
- 99.12 Accesso ai dati statistici individuali: l'esperienza di altri paesi, (*L. Buzzigoli, C. Martelli, N. Torelli*)
- 00.01 Analisi della qualità delle operazioni sul campo con riferimento alle principali indagini campionarie dell'Istat sulle famiglie, (*C. Filippucci, B. Buldo, V. Napoli, R. Bernardini Papalia*)
- 00.02 Analisi delle procedure di correzione/imputazione utilizzate dall'Istat nelle principali indagini sulle famiglie: volume I, (*L. Fabbris, C. Panattoni, M. Graziani*)

Il presente rapporto di ricerca è stato riprodotto nel mese di luglio 2000