CONTRIBUTI ISTAT

n. 7/2008

Seminario: Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali

AA.VV.

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

Contributi e Documenti Istat 2008

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

CONTRIBUTI ISTAT

n. 7/2008

Seminario: Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali

AA.VV.

(*) ISTAT –

Indice

Premessa

Savatore Filiberti, *Istat, DCSS/SSI*
Orietta Luzi, *Istat*, *DCMT/MTS-G*

Lo studio di metodologie e strumenti validi e affidabili per poter trattare i dati raccolti nell'ambito delle indagini sulle imprese è indubbiamente un tema di grande importanza che merita di essere trattato con la dovuta attenzione. La realizzazione di una corretta metodologia di controllo e correzione delle informazioni rilevate contribuisce fortemente al miglioramento della qualità delle statistiche sulle imprese. In altre parole, l'implementazione delle fasi di controllo e correzione, se condotta seguendo i criteri suggeriti dalla metodologia e secondo l'esperienza professionale del responsabile d'indagine, permette di ottenere un incremento significativo della qualità dei risultati prodotti.

La fase di controllo e correzione dati (C&C nel seguito), particolarmente nelle indagini economiche sulle imprese, è caratterizzata da alcuni elementi che ne aumentano la complessità sia da un punto di vista metodologico, sia operativo. Innanzi tutto, la fase di C&C ha effetti su alcune importanti dimensioni della qualità (in particolare, accuratezza, tempestività, confrontabilità). Tale fase è legata inoltre a molti altri passi fasi del processo di indagine, dalla definizione del questionario, alla pianificazione della rilevazione sul campo per la prevenzione di errori e mancate risposte, alle modalità di registrazione dei dati, all'accuratezza delle stime.

Inoltre, benché riconosciuta come una delle più costose in termini di tempi e risorse impiegate, e nonostante i continui avanzamenti metodologici che la interessano, la fase di C&C è ancora caratterizzata da una elevata eterogeneità anche nelle definizioni e nei concetti a livello sia Europeo sia di singoli Istituti di Statistica, e risulta ancora poco sistematizzata da un punto di vista sia teorico, sia progettuale.

Le indagini economiche pongono inoltre problemi peculiari, legati alle presenza nei dati di forti asimmetrie delle distribuzioni osservate, dovute alla presenza di unità più influenti delle altre sui parametri oggetto di stima. Questo aspetto va anche considerato alla luce del fatto che le variabili economiche sono generalmente caratterizzate da forti relazioni statistico-matematiche, che devono essere tenute sotto controllo, e sfruttate, in fase di individuazione e trattamento degli errori.

Va ricordato, infine, che nelle indagini economiche la difficoltà ad ottenere informazioni corrette dai rispondenti è in genere maggiore rispetto a quelle sulle famiglie, ad esempio a causa della possibile indisponibilità delle informazioni richieste nei sistemi informativi delle imprese, oppure a causa di diverse definizioni dei fenomeni rilevati adottate dai rispondenti e dagli Istituti di Statistica. Ciò contribuisce anche ad aumentare il tasso di non risposta, con conseguenti elevati tassi di imputazione e quindi effetti distorsivi sulle stime dei parametri obiettivo. Un altro elemento da considerare è infine quello legato alla pressione sui rispondenti, che nelle indagini economiche è spesso accresciuta da ritorni non ottimizzati sui non rispondenti, per la correzione di errori influenti o la compensazione dei valori mancanti.

Per tutte queste ragioni, è stato ritenuto importante avviare, a livello Europeo e quindi anche in ISTAT, una riflessione sullo stato dell'arte nell'area del C&C per indagini economiche, ed è stato avviato un processo di ricerca comune sulle migliori metodi e pratiche per l'ottimizzazione della qualità dei dati e la riduzione dei costi di produzione dell'informazione statistica. L'avvio di questo processo si è concretizzato nel progetto Europeo EDIMBUS (EDiting and IMputation for Cross-Sectional BUsiness Surveys) (edimbus.istat.it) finalizzato alla produzione di un Manuale di Pratiche Raccomandate (PR) per il C&C nelle indagini trasversali sulle imprese. Nel manuale (Luzi et al., 2007) sono descritti "buoni" metodi per effettuare le diverse operazioni di una strategia di C&C, insieme con i rispettivi attributi. Più

in generale, il manuale rappresenta uno strumento di supporto di tipo operativo e non (solo) metodologico per il disegno e la gestione di una procedura di C&C in una indagine di tipo economico.

La necessità di avviare anche internamente all'Istat un processo di confronto sulle migliori pratiche e strategie per il C&C dati è stata rafforzata dall'indagine conoscitiva effettuata nel corso del 2006 nell'ambito del progetto EDIMBUS. L'indagine, che ha coinvolto tutti gli Istituti di Statistica Europei più alcuni Istituti leader nel campo del C&C, ha consentito di raccogliere informazioni sulle strategie correnti in termini non solo di metodi e pratiche per il C&C, ma più in generale su aspetti legati alla progettazione, alla realizzazione, alla documentazione e alla valutazione di processi di C&C, nonché alle strategie dei vari Istituti di Statistica per l'armonizzazione, la pianificazione e il monitoraggio dei processi di C&C.

Alla ricognizione effettuata in ISTAT (Luzi e Della Rocca, 2006) hanno partecipato 23 indagini, fra cui i due censimenti economici dell'Industria e dell'Agricoltura, 8 indagini di tipo strutturale e 13 indagini di tipo congiunturale. I risultati più significativi hanno riguardato gli aspetti relativi ai costi e alla valutazione/documentazione delle strategie di C&C adottate. Se il 64% delle indagini rispondenti ha dichiarato di impiegare più del 40% delle risorse disponibili nelle varie attività di C&C, solo il 36% ha dichiarato di effettuare test preliminari della propria procedura di C&C. Il 78% delle indagini ha comunque dichiarato di documentare la procedura e/o i suoi risultati (nell'88% anche con indicatori). I risultati dell'indagine vanno comunque valutati, anche negli elementi positivi come quello appena citato relativo alla documentazione, alla luce di alcune evidenze:

- esiste anche in Istat una significativa eterogeneità in termini di definizioni e concetti nell'area del C&C;
- si opera in assenza di standard/linee guida specifiche nell'area del C&C, per cui a livello di Istituto si ha un basso livello di standardizzazione e controllo dei processi di C&C.

Questi risultati, se da un lato hanno confermato la necessità di disporre di linee guida per tutte le "fasi di vita" di una procedura di C&C in indagini sulle imprese, dall'altro hanno stimolato l'avvio di un confronto a livello Istat sullo stato dell'arte e sulle prospettive future di sviluppo e standardizzazione dei processi di C&C.

Un primo passo in questo senso è stato appunto il seminario *Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali* (ISTAT, Roma, 25 Maggio 2007). Attraverso l'analisi di esperienze e soluzioni avanzate in diversi contesti d'indagine, il seminario ha voluto essere un primo momento di riflessione per individuare prospettive di lavoro e di ricerca future in Istituto, per una prima valutazione delle aree di maggior eterogeneità, per una prima riflessione sulla fattibilità di una progressiva armonizzazione in Istat delle attività di progettazione, valutazione, documentazione di strategie complesse di C&C nelle indagini sulle imprese, e per sottolineare la necessità di sempre maggiore integrazione delle competenze e delle esperienze in Istat nell'area del C&C.

Il presente Contributo Istat raccoglie i lavori presentati nel corso del seminario. La giornata seminariale è stata aperta dal Direttore della Direzione Centrale per le Tecnologie e il Supporto Metodologico, dott. Gerardo Giacummo, che ha richiamato l'attenzione sul problema del miglioramento della qualità dei dati attraverso il miglioramento continuo dei processi di produzione dell'informazione statistica. Al seminario sono intervenuti in qualità di discussant il dott. Piero Demetrio Falorsi, Dirigente del Servizio Progettazione e Supporto Metodologico nei Processi di Produzione Statistica, e il dott. Marco Ballin, Dirigente del Servizio Statistiche sull'Agricoltura. Entrambi hanno richiamato l'attenzione sulla rilevanza delle tematiche affrontate nel corso dei vari interventi, sulla necessità di proseguire verso una sempre più approfondita analisi metodologico/operativa dei processi di C&C dati, e sull'esigenza di un sempre maggiore confronto metodologico su questi temi internamente all'Istituto.

Le conclusioni sono state tratte dal Direttore della Direzione Centrale delle Statistiche Economiche Strutturali, dott. Giuseppe Antonio Certomà, il quale ha ribadito l'importanza di condurre all'interno dell'Istituto una più stretta collaborazione tra il reparto metodologico e gli uffici direttamente impegnati nella produzione delle statistiche sulle imprese. L'obiettivo di tale attività è quello di affrontare insieme e risolvere i problemi relativi al controllo e alla correzione dei dati sulle imprese, peculiari e altamente specifici nell'ambito di ogni singola indagine, e pertanto non facilmente generalizzabili.

Bibliografia

Luzi O., Della Rocca G. *Risultati dell'indagine sulle pratiche correnti in Istat per il controllo e la correzione dei dati*. Rapporto tecnico redatto nell'ambito del "Gruppo di lavoro avente il compito di predisporre un manuale di pratiche raccomandate per il controllo e la correzione dei dati nelle indagini trasversali sulle imprese collegato al progetto europeo EDIMBUS". ISTAT. (2006).

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Rapporto tecnico del progetto Europeo EDIMBUS. Anche disponibile sul sito edimbus.istat.it. (2007)

Un approccio armonizzato al controllo e correzione dei dati: il caso della Community Innovation Survey

Valeria Mastrostefano, *Istat, Direzione Centrale delle Statistiche Economiche Strutturali, SSI*

Sommario: Obiettivo del presente contributo è descrivere la metodologia di controllo e correzione (C&C) dei dati della *Community Innovation Survey* (Cis), un'indagine sviluppata congiuntamente dall'Eurostat e dagli Istituti statistici dei Paesi Ue allo scopo di raccogliere informazioni quali-quantitative sulle attività di innovazione delle imprese europee. La specificità del processo di C&C della Cis risiede nell'armonizzazione delle strategie e procedure adottate a livello europeo. Per la Cis, infatti, Eurostat raccomanda l'utilizzo di pratiche e tecniche standard di C&C per gli Stati membri Ue e gli altri paesi europei coinvolti nell'indagine.
Il presente lavoro è strutturato in quattro parti. Nella prima sono descritti sinteticamente il contenuto informativo e le caratteristiche metodologiche dell'indagine. Nella seconda sono illustrati gli obiettivi dell'approccio armonizzato del processo di C&C impiegato nell'ambito della Cis. Tra questi vanno principalmente sottolineati: 1) l'uso di procedure 'su misura' per garantire, oltre a un alto livello di affidabilità delle singole stime, anche il rispetto dei legami tra variabili e delle peculiarità settoriali e dimensionali delle imprese, dal momento che l'innovazione è un processo complesso, sistemico e fortemente differenziato rispetto alle caratteristiche strutturali dell'economia; 2) la standardizzazione delle metodologie adottate in ambito europeo in modo da assicurare una migliore comparabilità internazionale dei dati; 3) la trasparenza del processo di C&C ottenuta mediante una appropriata documentazione delle procedure implementate. La terza parte analizza nel dettaglio il funzionamento del meccanismo di C&C. In particolare, sono descritte la struttura e la sequenza delle diverse fasi in cui è articolato il processo, nonché le regole impiegate per la localizzazione degli errori, le modalità di trattamento degli *outlier* e le tecniche di imputazione. L'ultima sezione è dedicata alla documentazione prodotta al fine di monitorare il processo di C&C, esaminare l'entità del fenomeno e le principali componenti del profilo dell'errore, calcolare indicatori di qualità di processo finalizzati alla valutazione complessiva della qualità dei dati. E' importante, infine, precisare che parte degli indicatori di processo prodotti vanno a costituire la base informativa per la valutazione dell'accuratezza e dell'impatto degli errori non campionari del *Quality Report* della Cis che tutti i Paesi europei interessati alla rilevazione sono tenuti a redigere a conclusione dell'indagine.

Parole chiave: innovazione, Community Innovation Survey, localizzazione deterministica degli errori, valori anomali, imputazione logico-deduttiva, imputazione da regressione, imputazione da donatore.

1. Introduzione

La Community Innovation Survey (Cis), avviata a livello europeo all'inizio degli anni '90, è attualmente alla sua quarta edizione. Sviluppata congiuntamente dall'Eurostat e dagli Istituti statistici dei Paesi Ue (in collaborazione con la Commissione europea), è finalizzata a raccogliere informazioni sui processi di innovazione delle imprese europee dell'industria e dei servizi. In particolare, la Cis fornisce un set integrato di indicatori volti a quantificare il fenomeno (in termini di soggetti coinvolti e di impegno finanziario sostenuto) e a qualificare le attività innovative, nonché ad analizzare le strategie, i comportamenti e le performance innovative di imprese con caratteristiche strutturali differenti, i fattori di ostacolo e di supporto all'innovazione, le complesse interazioni sistemiche che si attivano tra gli attori del processo innovativo.

L'indagine è condotta sulla base di criteri definitori e metodologie di rilevazione comuni a tutti i Paesi dell'Unione europea ed è inserita nel quadro concettuale del Manuale di Oslo, che adotta un approccio di analisi 'soggettivo' per l'esplorazione dei fenomeni innovativi. L'impresa, l'attore fondamentale del cambiamento economico, diventa l'unità statistica di rilevazione e di analisi per la misurazione statistica dell'innovazione (Ocse, 1997).

La rilevazione Cis viene svolta con cadenza biennale (quadriennale fino al 2004) e con riferimento ad un periodo di osservazione di un triennio.

Dopo alcuni anni di preparazione, le attività di rilevazione legate alla Cis sono state inserite in un quadro normativo europeo (Regolamento Ce 1450/2004) che ne stabilisce l'obbligatorietà per gli stati membri.

Nel corso degli anni, l'indagine ha subito significativi miglioramenti in termini sia di arricchimento dei contenuti che di affinamento delle metodologie adottate. Rispetto a quest'ultimo punto, rilevanti avanzamenti sono stati raggiunti per garantire un sostanziale allineamento di tutti i paesi aderenti allo standard Cis e una buona armonizzazione metodologica, non più solo sotto il profilo definitorio-concettuale, ma anche rispetto alle modalità di raccolta delle informazioni, alle procedure di controllo e correzione (C&C) dei dati e alle tecniche di produzione delle stime finali.

Attualmente, l'indagine è arrivata a coprire tutti i 27 Paesi membri, i Paesi Candidati più altri Paesi Efta (Svizzera, Norvegia e Islanda). È significativo che il modello sviluppato dall'Unione sia stato adottato da vari altri paesi, tra cui Canada ed Australia, per valutare l'innovazione delle loro imprese.

L'ultima edizione (Cis4) è stata condotta tra il 2005-2006 con riferimento al triennio 2002-2004. In Italia, la rilevazione sull'innovazione, campionaria per le imprese da 10 a 249 addetti e censuaria per quelle con almeno 250 addetti, ha interessato circa 45.000 unità, rappresentative dell'universo delle imprese italiane con almeno 10 addetti attive nel 2004 nei settori dell'industria e nei servizi (Sezioni C-K dell'Ateco 2002). La rilevazione è stata condotta mediante auto-compilazione di un questionario cartaceo. La rilevazione si è conclusa con un tasso di risposta effettivo pari al 49 per cento delle imprese del campione iniziale (pari a circa 22.000 unità rispondenti finali). A conclusione della raccolta dati, secondo quanto richiesto da Eurostat, si è proceduto anche ad una rilevazione campionaria telefonica (Cati) delle imprese non rispondenti, alle quali è stato somministrato un questionario ridotto. La rilevazione Cati ha permesso di verificare le eventuali divergenze sistematiche nel comportamento delle imprese non rispondenti rispetto alle imprese rispondenti e di ridurre quella componente distorsiva della stima dovuta ad un'autoselezione dei rispondenti. Data la tipologia campionaria dell'indagine e la presenza di imprese non rispondenti, la stima dei totali delle variabili di interesse è stata calcolata attribuendo ad ogni unità rispondente un coefficiente di riporto (o peso), indicante il numero di unità della popolazione rappresentate dall'impresa, inclusa se stessa. La metodologia di riporto dei dati all'universo è basata sugli "stimatori di ponderazione vincolata" (Deville e Särndal, 1992). Il calcolo dei pesi finali ha tenuto inoltre conto dei risultati della rilevazione di controllo effettuata sul sottocampione di imprese non rispondenti (Istat, 2007).

2. La strategia del processo di controllo e correzione della CIS: caratteristiche e finalità

Per la Cis, Eurostat raccomanda l'utilizzo di procedure standard per il controllo e correzione (C&C) dei dati, definite dallo stesso istituto europeo, al fine di implementare una metodologia per il trattamento dei dati armonizzata a livello europeo. Nell'ultima edizione dell'indagine, la maggior parte dei paesi UE interessati alla Cis hanno implementato le pratiche di correzione e imputazione sviluppate da Eurostat (Eurostat, 2006).

Le scelte strategiche di C&C adottate nell'ambito della Cis europea sono guidate da criteri rigorosi (che derivano da un'ampia conoscenza degli aspetti teorico-concettuali e da una ricca esperienza nella misurazione statistica del fenomeno) e ottimali sia in termini di efficacia (l'idea è di garantire tanto l'affidabilità delle stime quanto una buona qualità dei processi di produzione, soprattutto riguardo agli aspetti di comparabilità europea dei dati) che di efficienza (razionalizzazione del processo di C&C e ottimizzazione dell'impiego delle risorse finanziarie e umane disponibili).

Gli obiettivi dell'approccio armonizzato al processo di C&C dei dati Cis possono essere riassunti nei seguenti punti principali:

1. Individuazione e implementazione di procedure calibrate sulle situazioni da sottoporre a controllo. L'innovazione è un processo complesso, fortemente differenziato rispetto alle caratteristiche strutturali e alle capacità tecnologiche delle imprese e in cui le componenti di 'sistema' giocano un ruolo fondamentale. In quest'ottica, l'attenzione rivolta a quelle relazioni tra le diverse variabili in grado di esplorare l'eterogeneità e la dimensione sistemica dell'innovazione è di fondamentale importanza non solo in fase di costruzione del questionario e di formulazione dei quesiti, ma anche nella definizione delle scelte metodologiche relative al trattamento dei dati, alla specificazione degli *edit*, alle modalità di imputazione. In fase di progettazione del processo di C&C è stato, quindi, necessario adottare procedure 'su misura' per garantire non solo un alto livello di affidabilità delle singole stime, ma anche il rispetto dei legami tra variabili e della forte caratterizzazione strutturale delle imprese in modo da soddisfare le esigenze conoscitive specifiche degli utilizzatori interessati alle dinamiche interne, alle relazioni causa-effetto, alle specificità settoriali e dimensionali dei processi innovativi.
2. Armonizzazione delle fasi e delle operazioni componenti il processo di C&C e standardizzazione delle metodologie adottate in modo da assicurare un trattamento più oggettivo ed omogeneo delle situazioni di errore e contribuire, quindi, ad una migliore comparabilità internazionale dei dati sull'innovazione.
3. Trasparenza del processo di C&C ottenuta mediante una appropriata documentazione delle procedure implementate. La trasparenza del sistema facilita, inoltre, il monitoraggio delle attività e delle prestazioni del C&C, consentendo, da un lato, ai responsabili di indagine di tenere sotto controllo il processo, dall'altro, agli utilizzatori (analisti e esperti di settore) di disporre di un sistema informativo sulla qualità della CIS tale da agevolare l'utilizzo e l'interpretazione dei dati.
4. Flessibilità nell'utilizzo del programma di C&C. Il software definito *ad hoc* per il C&C della CIS è costituito da un insieme di moduli, ciascuno dei quali implementa una particolare funzione delle fasi principali (*editing*, individuazione degli errori, imputazione, individuazione degli *outlier*). Questo consente ai singoli paesi di: optare solo per alcuni moduli qualora si ritenesse più appropriato o meno *time-consuming* ricorrere ad altre tecniche e procedure di imputazione; operare delle modifiche o degli adattamenti nel caso in cui il questionario contenesse variabili (o addirittura intere sezioni) aggiuntive rispetto allo schema standard europeo; 'rilassare' i vincoli e ritoccare i parametri prefissati, per adattarli a casi particolarmente problematici, specifici delle singole indagini nazionali.
5. Facilità d'uso del programma di C&C. L'intero processo di C&C dei dati viene effettuato utilizzando un software appositamente sviluppato da Eurostat per la CIS. Il programma, realizzato in linguaggio Sas Macro (Windows – versione 8.2) é dotato di interfaccia grafica. Alcune divergenze nel campo di osservazione e nel questionario dell'edizione italiana (rispetto a quella europea) hanno reso necessario un adattamento del programma Sas ai target di rilevazione italiani. La facilità d'uso del programma ha permesso di personalizzare il programma secondo le esigenze nazionali prima di procedere al C&C.

3. L'implementazione del processo di controllo e correzione della CIS

Il processo di C&C implementato per la Cis si articola in 5 fasi (Figura 1):

1. Controllo e correzione manuale dei dati. Si tratta di un'attività di revisione, svolta prima della registrazione dei questionari, che prevede una serie di operazioni di individuazione e di correzione degli errori eseguite manualmente da tecnici mediante revisione dei modelli, re-intervista, uso di informazioni ausiliarie e/o di conoscenze sul fenomeno investigato.
2. Localizzazione deterministica degli errori mediante procedure automatiche. Il processo automatico di C&C comincia con l'individuazione deterministica delle situazioni di errore e delle variabili da imputare, a partire dagli *edit* specificati da Eurostat sulla base delle regole interne del questionario e di conoscenze a priori del fenomeno oggetto di rilevazione. Di diversa natura sono gli errori non campionari che possono originarsi nella Cis (errori di dominio, mancate risposte parziali, valori anomali e incompatibilità fra risposte, errori di codifica e di percorso) e diverse sono le procedure utilizzate per il trattamento degli errori.
3. Controllo e correzione interattivi dei valori anomali (*outliers*) di tipo micro. L'insieme dei record considerati non corretti comprende anche quelle unità che presentano, per le variabili quantitative rilevate, valori che si discostano in modo significativo dai valori che le stesse variabili assumono nel resto delle unità rispondenti (valori anomali o 'sospetti') e che possono incidere pesantemente sulle stime finali e sull'imputazione delle altre variabili. Il programma individua diverse modalità di trattamento degli *outlier* applicate solo dopo una preliminare e scrupolosa analisi circa la loro natura e l'impatto sulle stime finali.
4. Imputazione dei valori mancanti o errati di tipo misto mediante l'attivazione di procedure automatiche di tipo sia deterministico che probabilistico. Questa fase prevede l'implementazione sequenziale di tecniche diverse individuate in funzione della tipologia di variabile (quantitativa o qualitativa) e dell'errore riscontrato (incoerenze o valori mancanti). In particolare, il processo di correzione si compone di tre passi:
   - l'esecuzione iniziale delle procedure di imputazione logico-deduttiva;
   - l'adozione di procedure di imputazione basate prevalentemente sullo 'stimatore rapporto' per le variabili quantitative;
   - l'applicazione di tecniche di imputazione del donatore per le variabili qualitative (dicotomiche e categorico-ordinali).
5. Validazione del processo di C&C e analisi delle stime delle variabili principali del questionario per dominio di impresa mediante un confronto dei dati aggregati corretti e opportunamente ponderati con informazioni storiche o ausiliarie al fine di evidenziare eventuali situazioni 'sospette'.

Al termine delle attività di C&C il programma produce una serie di indicatori di qualità del processo allo scopo di monitorare il processo di produzione e calcolare indicatori di qualità di processo finalizzati alla valutazione complessiva della qualità dei dati.

3.1 Revisione manuale dei modelli

I questionari compilati sono sottoposti ad una preliminare attività di revisione manuale che prevede una serie di operazioni di individuazione e di correzione degli errori dovuti prevalentemente a problemi di carattere sistemico emersi in fase di compilazione del modello (ad esempio, causati da una errata interpretazione da parte del rispondente dei quesiti e delle regole di compilazione). In questa fase sono verificate le seguenti condizioni:

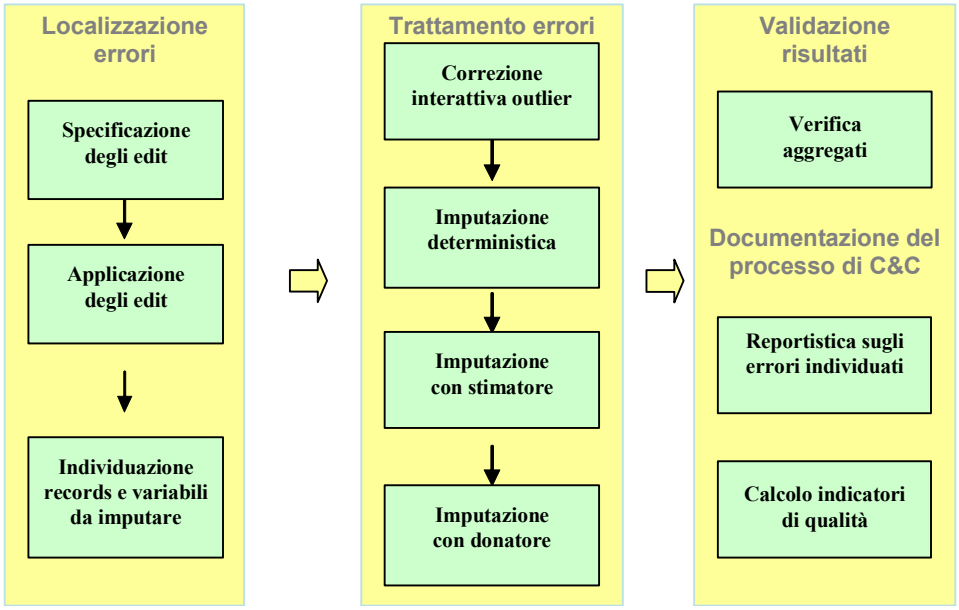
- l'appartenenza dei rispondenti al campo di osservazione dell'indagine per valutare l'eleggibilità dei rispondenti ed escludere, quindi, le unità non eleggibili contattate per errori di lista (unità non più esistenti, unità con variazioni di stato, unità fuori target);
- la corretta compilazione dei quesiti-chiave sull'innovazione e il grado di completezza dell'informazione fornita per distinguere, in relazione agli obiettivi conoscitivi dell'indagine, le mancate risposte totali dalle mancate risposte parziali e, quindi, decidere per l'inclusione o meno delle singole unità nel campione finale dei rispondenti;

- la presenza di errori di misura (spesso responsabili anche della presenza di valori anomali) e di altri errori sistematici (errori di esistenza, di dominio e di percorso), la cui origine è da attribuirsi a problemi dovuti a una non adeguata specificazione del problema, dei concetti e delle definizioni (difetti legati alla struttura del modello o alle norme di compilazione del questionario);
- la presenza di incongruenze logiche tra le variabili.

I casi dubbi riguardanti l'appartenenza delle unità al campo di osservazione della Cis (e, quindi, alla popolazione di riferimento dell'Indagine) e quelli relativi alla completezza dell'informazione rilevata (a partire dalla quale sono definite le unità del campione finale dei rispondenti) sono risolti mediante azioni di *follow-up* (re-intervista telefonica) nel caso di unità influenti (grandi imprese o unità attive in strati considerati 'critici' a causa della bassa numerosità delle unità presenti). Negli altri casi si ricorre, invece, all'uso di informazioni ausiliarie di fonte esterna: in particolare, all'Archivio Asia per informazioni di tipo strutturale; all'indagine sui Conti Economici per le variabili economiche (fatturato, ecc.); all'indagine R&S per i dati sulle spese innovative sostenute.

Al termine del controllo manuale i questionari sono registrati su supporto magnetico. L'acquisizione dei dati avviene mediante data entry sulla base di criteri armonizzati a livello europeo. Gli errori che non sono stati risolti nella fase precedente sono sottoposti al processo di C&C sviluppato da Eurostat.

Figura 1. *Le fasi principali del processo di controllo e correzione della Community Innovation Survey*



3.2 Localizzazione deterministica degli errori eseguita mediante procedure automatiche

L'individuazione dei valori non validi su variabili e record avviene nella fase di *editing* che identifica, sulla base di un insieme di regole e vincoli (non ridondanti e non contraddittori) i valori errati da correggere in modo tale che a livello di singolo record siano soddisfatti tutti gli *edit*. Un approccio di tipo deterministico è adottato da Eurostat nella costruzione delle procedure di *editing*.

I controlli derivano direttamente da regole interne al modello (relazioni logiche tra le variabili) come da relazioni di tipo statistico-matematico e da conoscenze specifiche a priori del fenomeno oggetto di rilevazione. I controlli sono finalizzati a verificare:

- il rispetto di certi requisiti per prefissate combinazioni di valori assunti da variabili rilevate in una stessa unità (controlli di coerenza);
- l'accettabilità o meno di valori mancanti per date variabili del modello condizionatamente alle risposte fornite ad uno o più particolari quesiti precedenti (controlli degli errori di percorso);

- la presenza di valori fuori dominio, ossia valori esterni all'intervallo di valori ammissibili per una data variabile (controlli degli errori di dominio);
- la coincidenza tra la somma dei dati parziali e il totale (controlli di quadratura);
- la presenza, in alcune unità statistiche, di valori al di fuori di prefissate soglie (controlli statistici).

Il controllo delle situazioni di errore è eseguito mediante applicazione di un programma di *Edit Fails*. Il sistema definisce una struttura gerarchica degli edit che, a fronte di più incoerenze interne, stabilisce un ordine sequenziale delle modifiche da effettuare su quelle variabili responsabili dell'attivazione degli *edit* (*edit fails*).

Questa fase si conclude con la elaborazione, da un lato, di un listato dei record errati per tipologia di *edit* violato e, dall'altro, di statistiche riassuntive che forniscono informazioni sulla frequenza (assoluta e relativa) delle regole di controllo predisposte e violate per record e variabile rilevata.

3.3 Individuazione e correzione dei valori anomali di tipo micro eseguite mediante procedure di tipo interattivo

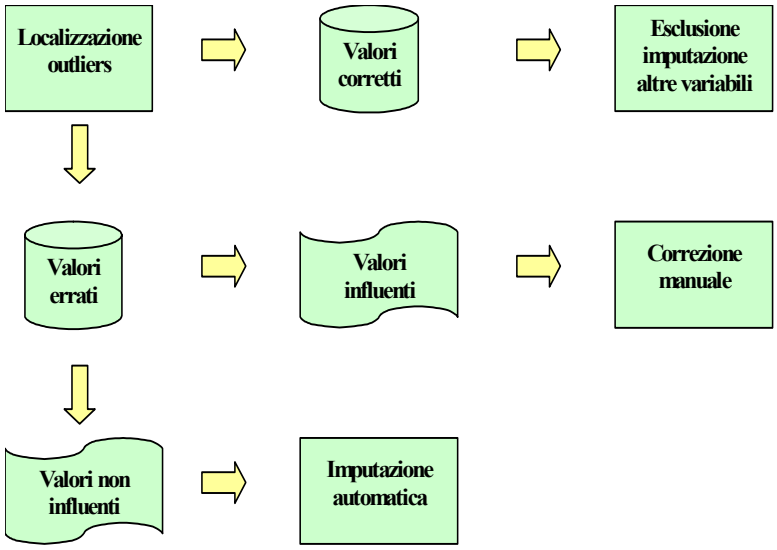
I valori 'sospetti' che sfuggono alla preliminare fase di revisione, o sono originati in fase di registrazione per l'errata trascrizione dei dati, sono sottoposti ad un'analisi micro per verificare se corrispondono a situazioni errate (sono cioè dovuti a errori di compilazione o di registrazione e quindi devono essere sottoposti a imputazione) oppure a casi reali (non sono errori, bensì valori estremi e, quindi, devono essere accettati come corretti e opportunamente considerati in fase di calcolo delle stime. L'analisi dei singoli *outliers* assume, quindi, un ruolo di fondamentale importanza per il loro trattamento in quanto da un lato, se i valori risultassero errati, in assenza di un'adeguata correzione o ponderazione si avrebbe un impatto considerevole sulle stime con inevitabili distorsioni sui risultati finali dell'indagine; dall'altro, invece, se gli stessi valori corrispondessero a situazioni reali, in presenza di un'errata imputazione si otterrebbe una forte perdita di informazione. Solo dopo averli analizzati, si decide come trattarli: se accettarli come valori validi, correggerli manualmente o includerli nel processo di imputazione probabilistica e automatica. E nel caso in cui siano accettati come corretti, se utilizzarli o meno in fase di imputazione delle altre variabili numeriche e qualitative.

Il controllo che si effettua è di tipo *inter-record*, ossia consiste nel confrontare i valori delle variabili di interesse in record diversi (controllo verticale) e l'individuazione degli *outliers* è effettuata su domini disgiunti dell'intera popolazione, all'interno dei quali si ritiene che il comportamento delle unità rispetto al carattere in esame sia omogeneo.

La localizzazione degli *outliers* avviene mediante determinazione di intervalli di accettazione al di fuori dei quali l'unità statistica è da considerare anomala. In particolare, i valori anomali per una certa variabile osservata sono individuati calcolando le distanze relative di ogni unità dal centro dei dati (considerato per domini disgiunti), e determinando i valori di soglia oltre i quali le unità sono da considerare sospette. L'intervallo di accettazione è ottenuto come funzione della distanza interquartilica della distribuzione dei dati. In particolare, calcolati Q1 e Q3 (primo e terzo quartile), i valori che cadono al di fuori l'intervallo individuato da (Q1 - 1,5*(Q3 -Q1)⌊—⌋ Q3 + 1,5*(Q3-Q1)) si considerano anomali e quelli che restano al di fuori di (Q1 - 3*(Q3 -Q1)⌊—⌋ Q3 + 3*(Q3-Q1)) sono *extreme outliers*. Si ricorre a quest'ultimo metodo per evitare l'esclusione di numerose osservazioni in presenza di una distribuzione fortemente asimmetrica.

La correzione avviene generalmente mediante imputazione automatica; per unità influenti si ricorre, invece, a informazioni ausiliarie o storiche sui rispondenti. Infine, nel caso di *outlier* corrispondenti a valori estremi reali (e quindi non errati) non si procede ad una loro imputazione, sebbene siano esclusi dal set di dati da utilizzare per l'imputazione delle altre variabili (Figura 2).

Figura 2. *Flusso delle fasi di individuazione e correzione dei valori anomali*



3.4 Imputazione dei valori mancanti o errati di tipo misto eseguita mediante procedure automatiche

Le procedure di imputazione impiegate per le mancate risposte parziali e i valori errati, interamente automatizzate, sono di tipo micro, ossia prevedono il controllo di tutti i record presenti nel data set e la correzione di tutti quelli che hanno determinato l'attivazione di un qualsiasi *edit*.

L'imputazione segue una struttura gerarchica: le variabili sono corrette secondo un ordine specificato a priori e mediante l'impiego di tecniche definite in funzione della tipologia di variabile esaminata (quantitativa o qualitativa) e dell'errore riscontrato (incoerenze logiche o valori mancanti). In particolare, la strategia di imputazione si caratterizza per l'applicazione sequenziale dei seguenti metodi:

- in una prima fase, si procede con un'imputazione logico-deduttiva che prevede l'assegnazione alla variabile errata di quell'unico valore in grado di riportare i record nella regione di accettazione;
- il secondo passo è costituito dall'implementazione delle procedure di imputazione delle variabili numeriche. Per queste si utilizza un metodo di regressione basato sullo stimatore rapporto, che tiene conto di una serie di relazioni predefinite tra le variabili da imputare e alcune variabili ausiliarie ad esse altamente correlate;
- l'ultima fase è rappresentata dall'imputazione delle variabili qualitative (dicotomiche e ordinali) mediante la tecnica del donatore che consiste nell'individuare, per ogni record errato e rispetto a ciascuna tipologia di variabile, il record donatore più simile (rispetto alle caratteristiche strutturali e ai comportamenti innovativi delle unità/imprese) i cui valori, sostituiti ai valori errati, consentano al ricevente di soddisfare tutti gli *edit*.

La scelta dei tre metodi di imputazione nasce dall'esigenza di garantire non solo un buon livello di precisione delle singole stime finali ma di preservare anche le relazioni interne alle variabili del questionario, dal momento che lo studio dei legami tra le diverse informazioni rilevate dalla Cis è di cruciale importanza nell'analisi della multi-dimensionalità e eterogeneità dei processi innovativi nelle imprese.

Per ciascuna variabile (o set di variabili riconducibili alla stessa famiglia di indicatori), la presenza di valori imputati è segnalata da variabili-flag che indicano se e in che misura l'informazione originaria è stata modificata nel corso del processo di C&C in modo che gli analisti possano tenerne conto ed eventualmente valutare il potenziale effetto dell'imputazione sui risultati ottenuti.

3.4.1 Imputazione logico-deduttiva

Prima di passare all'imputazione dei valori mancanti con metodi più strettamente 'statistici', il programma procede ad un'imputazione di tipo logico-deduttivo per 'ripulire' il dataset dei possibili valori non validi dovuti ad un'errata compilazione dei modelli o a problemi generati in fase di registrazione. Il metodo logico-deduttivo permette di dedurre il valore da sostituire al dato mancante da una o più variabili ausiliarie, sfruttando le informazioni presenti nel dataset relative alle relazioni interne alle variabili del questionario. I valori da imputare per unità aventi le stesse caratteristiche sono, quindi, determinati in modo univoco in modo da garantire il rispetto dei legami tra le diverse variabili, precedentemente definiti da specifici modelli di comportamento del fenomeno investigato. Le procedure di correzione sono costituite da regole di imputazione deterministica del tipo: SE [condizione di errore] ALLORA [azione di correzione]. Il programma è impostato in modo tale che queste procedure di correzione non saranno eseguite finché nel campione dei rispondenti finali vi sia anche un solo valore mancante sulla variabile-filtro (Impresa con attività innovativa nel triennio di riferimento) ottenuta dalla somma logica di alcune variabili-chiave sull'innovazione rilevate dal questionario.

3.4.2 Imputazione delle variabili numeriche con lo stimatore-rapporto

Dopo aver rimosso dal dataset le incoerenze logiche e aver imputato, laddove possibile, i valori mancanti utilizzando le informazioni contenute nel resto dei dati o desunte da relazioni note, il programma esegue l'imputazione delle variabili quantitative continue, applicando un metodo di regressione che tiene conto dei rapporti tra la variabile da imputare ed alcune variabili ausiliarie ad essa altamente correlate (stimatore-rapporto) in modo da preservare le relazioni interne tra le variabili del questionario e garantire così una stima non distorta della covarianza tra le due variabili.

In particolare, due diversi stimatori-rapporto sono utilizzati per due set differenti di variabili: per le variabili economiche (addetti e fatturato totale), rilevate sia nel primo che nell'ultimo anno del triennio di riferimento della Cis, si impiega il 'rapporto di variazione', che permette di cogliere le variazioni temporali intervenute nelle unità rispondenti, ponendo a rapporto i valori assunti dalle variabili in corrispondenza dei due anni esaminati; per le variabili quantitative di innovazione (spese e fatturato) viene, invece, utilizzato il 'rapporto corrente' tra la variabile da imputare e una ausiliaria strettamente correlata alla prima rilevata nello stesso anno (il fatturato totale).

Il metodo prevede la suddivisione in classi delle unità in quanto le relazioni tra le variabili da imputare e le covariate possono cambiare molto da strato a strato. L'imputazione avviene, quindi, all'interno di singole celle di imputazione. Le celle (o classi) di imputazione sono ottenute operando una serie di opportune stratificazioni che risultano dalla concatenazione di due variabili di struttura (attività economica e dimensione aziendale) e che identificano sottoinsiemi omogenei di record/imprese con caratteristiche strutturali simili. Il numero delle celle è determinato in modo da assicurare la presenza di un numero minimo di rispondenti in ogni cella al fine di ottenere stime affidabili dei valori mancanti.

Eurostat, infine, prevede un ordine sequenziale nell'imputazione delle variabili numeriche che definisce prioritaria la correzione di quelle variabili che saranno utilizzate nell'imputazione delle altre variabili quantitative del questionario.

La procedura standard può essere così riassunta (Figura 3):

1. creazione delle celle di imputazione;
2. specificazione dei parametri che controllano il processo di individuazione degli *outliers* da escludere dalla stima dei valori mancanti delle variabili numeriche. In particolare sono definiti: il parametro che controlla la percentuale massima di records con valori anomali da rimuovere in fase di imputazione; i parametri della funzione che individua gli *outliers*. Per evitare l'esclusione di molte unità dal calcolo dei rapporti costruiti per l'imputazione, in presenza di una distribuzione della variabile fortemente

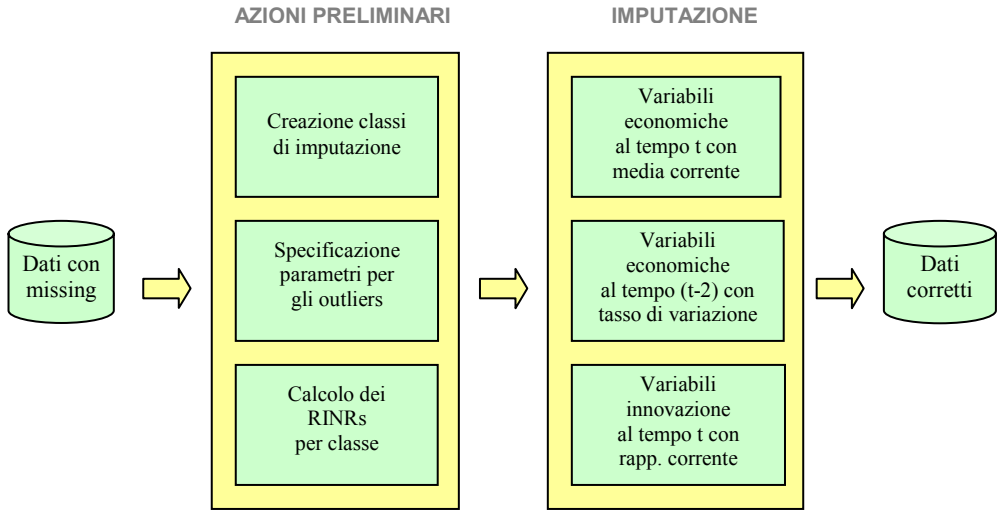
asimmetrica le procedure prevedono la possibilità di rilassare i vincoli per la definizione degli *outliers* in modo da individuare solo i valori anomali estremi;

3. calcolo dei tassi di mancata risposta parziale (RINRs) per strato. Se i RINRs sono inferiori al 50%, allora si procede all'imputazione. Altrimenti, il programma procede con l'aggregazione di più classi di imputazione fino a ridurre il RINR ad una percentuale inferiore al 50%;
4. imputazione delle variabili economiche. Questa azione prevede tre passi:
   - l'imputazione delle variabili economiche chiave del questionario (addetti e fatturato totale) nell'anno t (anno di riferimento dell'indagine) per i record/imprese che presentano un missing. Eurostat raccomanda di reperire i dati contattando le imprese interessate o ricorrendo a fonti statistiche ausiliarie (in particolare all'Archivio delle Imprese per il numero di addetti e all'Indagine Sci-Pmi per i dati sul fatturato). In caso di persistenza del fenomeno di mancata risposta sulle due variabili, il programma procede automaticamente ad un'imputazione con lo stimatore del valore medio che viene eseguita all'interno di ogni classe di imputazione;
   - il calcolo, all'interno di ciascuna cella di imputazione, del rapporto tra il totale del fenomeno nell'anno t-2 ed il suo totale nell'anno t per tutte le unità del campione rispondenti;
   - l'imputazione dei valori mancanti per l'anno t-2 mediante l'utilizzo dello stimatore rapporto calcolato come sopra;
5. imputazione delle variabili di innovazione (spese e quote di fatturato dovute ai prodotti innovativi) mediante lo stimatore 'rapporto corrente', ottenuto ponendo a confronto la variabile da imputare (numeratore) con la variabile 'fatturato totale'. Il processo termina con la verifica delle quadrature.

3.4.3 Imputazione delle variabili qualitative con il metodo del 'donatore hot deck'

Il passo successivo è la correzione delle variabili qualitative (dicotomiche e categorico-ordinali). Il metodo utilizzato è l'imputazione da donatore tramite hot deck gerarchico, che permette di individuare, per ogni record errato (ricevente), il record donatore 'più vicino', ovvero quello più simile rispetto alle caratteristiche strutturali (dimensione e settore di attività) e ai comportamenti innovativi delle unità/imprese ed i cui valori consentono al recipiente di soddisfare tutti gli edit. La procedura consente, per ogni variabile (o set di variabili riconducibili ad una specifica sottosezione del questionario), l'attribuzione all'unità con dato non valido (mancante o errato) del valore corretto dell'unità donatrice che immediatamente la precede secondo un dato ordinamento gerarchico. Le unità (rispondenti e non) sono, quindi, collegate secondo una base gerarchica e ordinate in modo che risultino vicino unità tra loro simili dati i valori assunti per un certo set di variabili.

Figura 3. *Flusso delle fasi di imputazione delle variabili quantitative*



In ambito europeo si è scelto questo metodo di imputazione perchè:

- consente, per ogni unità con dato mancante, di individuare unità con dati completi che presentano caratteristiche simili;
- il valore sostituito al posto del dato mancate è un valore 'reale';
- è in grado di garantire il rispetto delle distribuzioni congiunte fra le variabili e, quindi, di preservare le relazioni fra le variabili, soprattutto se uno stesso donatore è usato per imputare tutte le mancate risposte parziali di un record.

Nel metodo del 'record più vicino', il donatore è scelto in modo tale da minimizzare una misura della distanza multivariata tra esso ed il ricevente. Le procedure di imputazione delle variabili qualitative della Cis impiegano una funzione di distanza basata sull'Entropia relativa, che misura la disomogeneità nella distribuzione di una variabile con due o più modalità. Quando tutte le unità presentano la stessa modalità, l'entropia è minima. E' massima, invece, quando le unità sono massimamente equidistribuite tra le modalità (ogni modalità ha la stessa frequenza relativa).

Tale metodo è basato su alcuni passi principali (Figura 4):

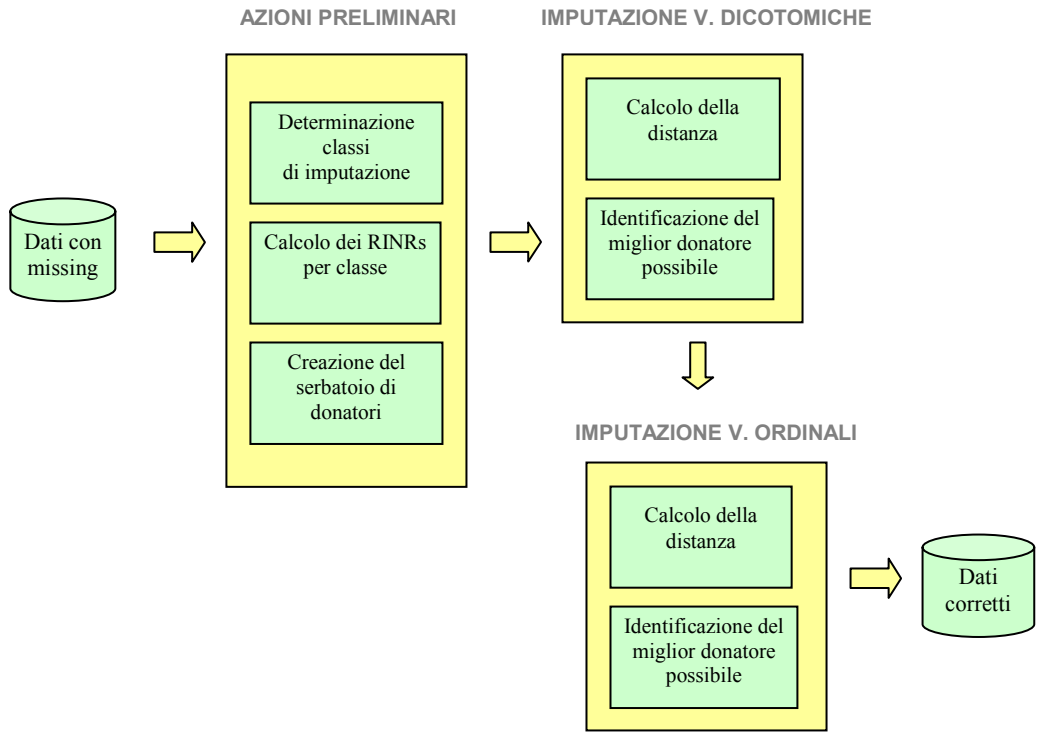
- la selezione delle variabili di stratificazione necessarie alla creazione delle classi di imputazione. Per costruire le classi sono utilizzate delle variabili ausiliarie che devono soddisfare i seguenti requisiti: non essere affette da errore (l'imputazione dei valori mancanti deve essere avvenuta preliminarmente); figurare negli edit violati dal recipiente; essere correlate con le variabili da imputare. Sono utilizzate variabili differenti per diversi set di imputazione. Per la creazione delle classi di imputazione si utilizzano anche due variabili strutturali: dimensione e attività economica;
- la conversione delle variabili di stratificazione continue in variabili ordinali. Le variabili di accoppiamento vengono trasformate in variabili ordinali in modo da essere ricondotte ad una scala comune. Queste nuove variabili possono assumere dalle 2 alle 5 modalità in funzione del numero di record presenti;
- la creazione delle classi di imputazione. Tutte le unità (rispondenti e non) sono, quindi, raggruppate in 'classi di imputazione', costruite sulla base di set dettagliati di variabili ausiliarie (strutturali e di indagine). In questo modo si dispone, per ciascun record con missing, di un serbatoio di donatori 'simili', ossia appartenenti alla stessa classe di imputazione. Nel caso in cui non sia possibile trovare un donatore adatto nell'iniziale classe di imputazione, le classi sono collassate per consentire il matching ad un livello più basso. Si utilizza un serbatoio di donatori unico per ciascun set di variabili 'simili' affette da errore;
- il calcolo dei tassi di mancata risposta parziale (RINRs) per strato. Se i RINRs sono inferiori al 50%, allora si procede alla identificazione del pool di donatori. Altrimenti, il programma procede con l'aggregazione di più classi di imputazione fino a ridurre il RINR ad una percentuale inferiore al 50%;
- il calcolo della distanza tra l'unità del campione con mancata risposta e tutte le altre unità con dati corretti all'interno di ciascuna classe di imputazione. Il miglior donatore possibile sarà scelto tra quelli che sono più 'vicini' (più simili) al 'ricevente' e con la più bassa entropia possibile. L'imputazione è di tipo congiunto per le variabili appartenenti ad uno stesso set di informazioni (dato un record con più valori mancanti, viene utilizzato un solo donatore per integrarne simultaneamente le mancate risposte), mentre è di tipo sequenziale per le variabili appartenenti a sottoinsiemi di informazione differenti (dato un record con più valori mancanti, viene utilizzato un donatore diverso per ogni mancata risposta). Per tenere sotto controllo l'uso multiplo dei donatori ed evitare che uno stesso donatore venga utilizzato troppe volte, è introdotta nella funzione di distanza una funzione di penalizzazione che tiene conto del numero di volte che un dato record è già stato utilizzato come donatore.

A conclusione dell'imputazione, viene svolto un ulteriore controllo dei dati imputati per verificare se tutti gli edit e i vincoli di uguaglianza originari risultano ancora soddisfatti.

3.4.4 Validazione delle procedure di controllo e correzione

La fase finale del processo di C&C prevede un'analisi comparativa dei dati aggregati corretti e opportunamente ponderati con i dati Cis della precedente edizione (per le variabili di innovazione) e con le informazioni ausiliarie provenienti dall'Indagine Sbs (per le variabili economiche, come fatturato e addetti), al fine di misurare e valutare le eventuali divergenze dei dati prodotti con quelli pubblicati, individuare eventuali valori sospetti a livello macro (per analizzarne le cause e correggerne le distorsioni) e procedere a una validazione finale dei dati aggregati. Per il calcolo delle stime vengono utilizzati pesi 'preliminari' (rispetto a quelli finali), ottenuti 'aggiustando' i pesi iniziali per tener conto, in questa fase, solo delle mancate risposte totali e non dei risultati della rilevazione di controllo sul sottocampione di non rispondenti.

Figura 4. *Flusso delle fasi di imputazione delle variabili qualitative*



Le informazioni sulla precedente edizione della Cis e quelle relative alla Sbs sono forniti dalla banca dati *New Cronos* dell'Eurostat. Per effettuare i controlli sono utilizzati due indicatori che misurano la variazione percentuale delle stime tra indagini diverse. L'analisi è condotta per sottoinsiemi aggregati secondo i domini di appartenenza (settore economico e classe di addetti). Il confronto finale tra stime si rende necessario per garantire sia un alto grado di comparabilità temporale dei dati di una stessa indagine, che un buon livello di coerenza con dati relativi allo stesso fenomeno ma provenienti da fonti statistiche differenti.

4. La documentazione del processo di controllo e correzione della Cis

Al termine delle singole fasi delle attività di C&C, come a conclusione dell'intero processo, il programma produce una serie di indicatori di qualità del processo allo scopo di:

- monitorare il processo di produzione;
- valutare le principali componenti del profilo dell'errore e documentare l'entità del fenomeno;
- individuare le cause (strutturali, organizzative, metodologiche, ecc.) delle componenti sistematiche degli errori;

- programmare interventi correttivi nel caso gli indicatori segnalino problemi in qualche fase del processo di produzione;
- calcolare indicatori di qualità di processo finalizzati alla valutazione complessiva della qualità dei dati.

L'analisi dei risultati, basata sulla produzione di una dettagliata reportistica, fornisce informazioni di tipo sia micro che macro relative all'incidenza degli errori riscontrati nel complesso e per le singole variabili controllate. In particolare, a livello micro sono prodotte informazioni su:

- il numero di edit attivati;
- il numero e il tipo di variabili responsabili dell'attivazione degli edit;
- il numero di inconsistenze, errori di dominio e di codifica;
- il numero di outlier individuati il tipo di trattamento cui sono stati sottoposti.

A livello aggregato e per sottoinsiemi di dati (per settore economico e classe di addetti), gli indicatori sintetici prodotti sono:

- il tasso di mancata risposta parziale per variabile, come risulta al termine delle fasi di *editing* e di imputazione logico-deduttiva;
- il tasso di imputazione totale delle singole variabili al termine della fase di imputazione;
- una tavola riassuntiva che riporta per le principali variabili della Cis un'analisi comparativa del numero di records/imprese calcolato sulla base del file iniziale di dati grezzi, del file di dati puliti e del file finale di dati corretti, al fine di misurare l'impatto complessivo del processo di C&C sui dati;
- alcuni indicatori relativi alla coerenza con le stime prodotte da fonti esterne e alla comparabilità temporale delle stime della Cis con i risultati della stessa indagine svolta con riferimento al triennio precedente.

La produzione di indicatori per sottoinsiemi di dati permette, infine, l'analisi della variabilità degli indicatori tra sottogruppi di unità statistiche aggregate secondo i domini di appartenenza.

Questi indicatori costituiscono la base informativa per la valutazione e la documentazione dell'accuratezza e dell'impatto degli errori non campionari sulle stime riportate nello *Standard Quality Report* della Cis (Eurostat, 2007) redatto da tutti i Paesi europei interessati alla rilevazione al fine di garantire agli utenti finali la massima trasparenza del processo di produzione e facilitare i confronti tra le diverse indagini europee.

Bibliografia

Deville J.-C., Särndal C. E. (1992) *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, N. 87, p. 376-382.

Eurostat (2006) *Community Innovation Survey. User Guide for Windows SAS application,* published at EUROSTAT WEB page: http://forum.europa.eu.int/Public/irc/dsis/Home/main - Section S&T and Innovation Statistics/ CIS4.

Eurostat (2007) *CIS 4. The 4th Community Innovation Survey, Quality Report for Country Italy*, published at EUROSTAT WEB page: http://forum.europa.eu.int/Public/irc/dsis/Home/main - Section S&T and Innovation Statistics/ CIS4/CIS4 Quality Reports.

Istat (2007) *Statistiche sull'Innovazione nelle Imprese. Anni 2002-2004*, Collana Informazioni, prossima pubblicazione.

Ocse (1997) *Proposed guidelines for collecting and interpreting technological innovation data – Oslo Manual*, Ocse, Paris.

Il metodo di controllo e correzione dei dati utilizzato per la Rilevazione sulla Struttura delle retribuzioni (anno 2002)

Roberto Sanzo, *Istat, Direzione Centrale delle Statistiche Economiche Strutturali, SSI*

Sommario: La strategia di controllo e correzione dei dati di indagine della Rilevazione sulla struttura delle retribuzioni (SES) per l'anno 2002 è stata caratterizzata dall'uso di tecniche di imputazione diverse a seconda della natura delle variabili rilevate. La particolarità del piano di campionamento (uno dei pochi casi di campione a doppio stadio nell'ambito delle rilevazioni economiche strutturali) e la complessità del questionario adottato hanno reso necessaria l'applicazione congiunta di procedure generalizzate e di procedure implementate *ad hoc*. In particolare, le prime sono state utilizzate essenzialmente per imputare, probabilisticamente o con tecnica del donatore, le variabili di tipo qualitativo, sia per le unità di primo stadio (imprese) che per quelle di secondo stadio (dipendenti); le seconde soprattutto per correggere le variabili di tipo quantitativo, retribuzioni lorde, costo del lavoro e ore lavorate su tutte. Per questa seconda tipologia di variabili, inoltre, la correzione è avvenuta in maniera gerarchica, correggendo prima gli aggregati economici ritenuti più importanti ai fini dell'indagine e poi agganciando a queste i risultati delle altre variabili. Il controllo è avvenuto attraverso l'utilizzo contemporaneo di *set* di indicatori caratteristici che ha permesso di individuare eventuali incongruenze tra le variabili costituenti tali indicatori; molto importante a tale scopo è stata la possibilità di utilizzo di informazioni provenienti da altre fonti, statistiche e amministrative, sia per la costruzioni di *range* di accettazione dei valori dei suddetti indicatori, sia per l'eventuale imputazione di valori mancanti o ritenuti errati attraverso l'applicazione di indicatori mediani di strato.

Parole chiave: struttura delle retribuzioni, imprese, dipendenti, retribuzioni lorde, costo del lavoro, ore lavorate, localizzazione errori, correzione probabilistica, correzione deterministica, fonti amministrative, indicatori caratteristici.

1. Introduzione

La rilevazione sulla struttura delle retribuzioni del 2002 (Structure of Earning Survey), prevista dal regolamento della U.E. n° 530/1999, è una rilevazione a cadenza quadriennale ed ha l'obiettivo di fornire importanti evidenze sui differenziali salariali esistenti nel mercato del lavoro subordinato riferiti alle principali caratteristiche personali e professionali dei lavoratori (sesso, età, titolo di studio, qualifica professionale). A tale scopo, vengono richieste, per i dipendenti, informazioni relative alle ore, alle retribuzioni e ai giorni lavorati, sia per l'intero anno 2002 che per il mese di ottobre (nel seguito indicate come informazioni mensili). In aggiunta, vengono rilevate anche alcune informazioni sulle caratteristiche delle imprese di cui i lavoratori rilevati sono dipendenti: in particolare, oltre a informazioni prettamente strutturali e relative alle caratteristiche dell'impresa (attività economica, localizzazione geografica, dimensione aziendale), sul questionario sono presenti alcuni quesiti relativi alle ore (ordinarie, straordinarie, non lavorate ma retribuite) e al costo del lavoro nonché informazioni circa le tipologie di contratto e le modalità di organizzazione del lavoro.
La rilevazione ha come universo di riferimento le imprese (con dipendenti) con almeno 10 addetti, appartenenti ai settori da C a K della Classificazione delle attività economiche Nace Rev.1 e si basa su un piano di campionamento a due stadi: il primo stadio è rappresentato dalle imprese, stratificate rispetto ai settori di attività economica (codici a due cifre), a 7 classi di dipendenti (10-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1000 e oltre) e 5 ripartizioni territoriali (Nord-Ovest, Nord-Est, Centro, Sud, Isole); il secondo stadio è rappresentato dalla selezione di un campione casuale di lavoratori dipendenti, il cui numero varia in ciascuna impresa selezionata nel primo stadio a seconda della sua dimensione (da un minimo di cinque ad un massimo di cento dipendenti per le imprese con oltre 1.000 addetti). Al fine di garantirne una selezione casuale delle unità di secondo stadio, le imprese rilevate sono state istruite relativamente alle modalità di selezione dei dipendenti stessi.
Allo scopo di diminuire gli effetti negativi delle mancate risposte totali, che nelle indagini sulle imprese risultano strutturalmente elevate, si è pianificato un incremento del 30% del campione base, pervenendo ad una lista finale di circa 22.300 imprese-campione per un totale di 188.000 "fogli-dipendenti" attesi.
La rilevazione è condotta mediante autocompilazione di questionari inviati per posta alle imprese selezionate, con solleciti telefonici e postali volti a garantire una buona copertura del campione base.
Il tasso di risposta è stato complessivamente pari al 43,7 % in termini di imprese, rappresentative del 46,2% dei dipendenti totali attesi.
Il campione finale rispondente è costituito da 9.771 imprese e da 86.753 dipendenti ed è risultato sufficiente ad assicurare la significatività statistica dei risultati, dato l'iniziale sovracampionamento del 30%, e a garantire la corrispondenza ai severi requisiti di "qualità statistica" richiesti dal Regolamento comunitario.
Il riporto dei dati all'universo è stato effettuato con l'ausilio della procedura Istat solitamente utilizzata per le indagini campionarie sulle imprese. Tale metodologia è basata sullo stimatore di ponderazione vincolata e si avvale delle informazioni strutturali ausiliarie fornite dall'archivio di riferimento ASIA 2002 per la stima dei pesi finali da assegnare alle unità rispondenti. Si assume quindi che le variabili oggetto di indagine siano strettamente correlate con quelle strutturali ausiliarie. Inoltre viene garantito il rispetto di uguaglianza tra i "totali noti" delle variabili ausiliarie e le stime delle stesse. Le variabili ausiliarie utilizzate sono costituite da: numero di imprese e numero di dipendenti.
I domini di stima sono stati definiti a livello di sottosezione di attività economica, classe di dipendenti (10-19; 20-49; 50-99; 100-249; 250-499; 500 e oltre) e 11 macroregioni della classificazione NUTS1 a due *digit*.

2. Fase di controllo e correzione

La fase di controllo e correzione delle informazioni rilevate sui rispondenti all'indagine in oggetto, è stata svolta, in maniera differenziata, su due livelli:

- sulle unità di rilevazione di primo stadio (le imprese);

- sulle unità di rilevazione di secondo stadio (i dipendenti).

Dopo una prima fase di controllo delle informazioni strutturali complessive delle imprese rispondenti (codice di attività economica e numero medio di addetti), in termini operativi si è preferito effettuare la fase di controllo e correzione delle informazioni desunte dal questionario in maniera gerarchica, correggendo prima le informazioni relative ai singoli dipendenti (unità di secondo stadio) e poi quelle relative al complesso dell'impresa (unità di primo stadio).
Inoltre, sia per le unità di primo che di secondo stadio, per le variabili di tipo quantitativo, è stato possibile svolgere una serie di controlli utilizzando diverse fonti esterne, tutte relative al 2002, sia statistiche che amministrative: la Rilevazione sul sistema dei conti delle imprese e la Rilevazione sulle piccole e medie imprese e sull'esercizio delle professioni (in seguito congiuntamente indicate con Sci-Pmi), l'Indagine sull'occupazione nelle Grandi Imprese (GI) e l'Archivio Oros-Inps. L'utilizzo di quest'ultima fonte, in particolare, ha permesso di effettuare non solo controlli a livello di impresa ma anche a livello di singolo dipendente: infatti le informazioni relative alle retribuzioni da essa desumibili hanno garantito, nella maggior parte dei casi, l'individuazione di *range* di variazione delle retribuzioni stesse per categoria professionale all'interno delle singole imprese, sia a livello annuale che mensile (relativi cioè a ottobre del 2002, come richiesto dal questionario). La correzione delle variabili di tipo quantitativo è avvenuta con metodi di tipo deterministico, basati essenzialmente sull'utilizzo di indicatori caratteristici, metodi che sono stati implementati in SAS con programmi *ad hoc*.
Le variabili qualitative sono state invece sottoposte principalmente a controllo e correzione di tipo probabilistico e tramite donatore utilizzando il software generalizzato Concord ed in particolare il modulo SCIA (per le unità di primo stadio) e il modulo RIDA (per le unità di secondo stadio). La scelta di RIDA in luogo di SCIA è derivata dall'esigenza di considerare nella procedura di correzione dei valori errati e di imputazione dei valori mancanti delle variabili qualitative dei *fogli dipendenti*, alcune variabili quantitative. Le variabili qualitative *multiresponse* e le variabili *filtro* del modello relativo alle informazioni sulle imprese sono state trattate con metodi deterministici.
Nel seguito, verrà presentata una sintetica descrizione delle diverse fasi attraverso le quali si sono esplicitate le attività di controllo e di correzione dei dati di indagine: essa seguirà le modalità con le quali queste ultime sono state praticamente messe in atto, a partire cioè dal *check* delle unità di secondo stadio, per passare poi a quelle di primo stadio.

2.1. Controllo e correzione delle unità di secondo stadio (fogli dipendenti)

Questa fase ha riguardato un totale di 86.753 "fogli dipendenti" pervenuti. La procedura ha inizialmente suddiviso le unità rilevate in due gruppi in base alla durata dell'orario di lavoro,

- dipendenti a tempo pieno
- dipendenti a tempo parziale.

Pur utilizzando per grandi linee la medesima procedura di controllo e correzione, si è preferito separare i due gruppi sia per favorire un controllo più puntuale, specialmente dei secondi, sia per evitare che i livelli degli indicatori calcolati su questi ultimi finissero per influire in maniera determinante su situazioni medie. La distinzione tra i due gruppi è stata effettuata in base alle domande dirette rivolte al singolo dipendente, dalle quali sono risultati:

- 78.847 dipendenti a tempo pieno;
- 7.906 dipendenti a tempo parziale.

Nel prosieguo ci si concentrerà sulla procedura utilizzata per i full-time: come detto, infatti, la procedura utilizzata per i part-time non presenta significative differenze rispetto alla precedente. La strategia che si è adottata è stata quella di ricondurre, in base al quesito sulla percentuale di orario a tempo parziale, tutti i valori di ore e retribuzioni rilevati sui dipendenti part-time al caso teorico di lavoro a tempo pieno. In tal modo i controlli che sono stati effettuati sono stati i medesimi del caso full-time[1], il che ha garantito la possibilità di utilizzo degli stessi programmi *ad hoc* già implementati.

[1] Per i dipendenti part-time, pur utilizzando la stessa metodologia, si sono considerate, laddove numericamente possibile, distribuzioni specifiche degli indicatori per strato.

Il primo passo è stato quello di individuare un gruppo all'interno dei full-time che soddisfacesse i seguenti criteri di correttezza:

1. presenza del valore delle retribuzioni lorde totali e delle ore totali lavorate, sia annuali che mensili;
2. soddisfacimento dell'appartenenza del seguente set di indicatori ad alcuni domini di accettazione:
   - retribuzioni orarie annuali;
   - retribuzioni orarie mensili;
   - rapporto tra retribuzioni annuali e retribuzioni mensili;
   - rapporto tra retribuzioni orarie annuali e retribuzioni orarie mensili.

I fogli dipendente che hanno soddisfatto tutte le condizioni sono stati circa 53 mila, cioè il 68 per cento dei dipendenti a tempo pieno rilevati. I circa 25 mila dipendenti rimanenti sono stati sottoposti alla procedura di controllo e correzione che verrà di seguito descritta. Tale procedura, inoltre, è stata differenziata a seconda del soddisfacimento o meno del primo criterio, cioè quello dell'esistenza di tutte e quattro le informazioni relative alle retribuzioni e alle ore, annuali e mensili.

2.1.1. Controllo e correzione dei fogli dipendenti con le informazioni relative alle ore e alle retribuzioni, annuali e mensili, tutte non mancanti

Questa fase del processo di controllo e correzione è stata quella che ha coinvolto il maggior numero di unità, circa 24 mila, cioè circa il 96 per cento delle unità sottoposte a procedura di controllo e correzione. Questo risultato sta ad evidenziare la particolare attenzione che si è avuta in fase di rilevazione e di revisione relativamente alle quattro variabili più rilevanti per gli obiettivi dell'indagine per le quali, quindi, si è registrato complessivamente soltanto il 4 per cento di unità con uno o più di quei quattro valori mancanti.

Sulle unità con presenza di valori relativamente alle retribuzioni e ore, annuali e mensili, la procedura qui utilizzata è andata prima di tutto a verificare la congruenza delle informazioni relative alle retribuzioni. Un primo segnale sulla bontà dei due valori (annuale e mensile) è stato possibile ottenerlo attraverso il rapporto tra le retribuzioni annuali e quelle mensili. In formule:

$$I_1 = \frac{R_a}{R_m} \quad \text{(rapporto tra le retribuzioni annuali e retribuzioni mensili)}$$

dove, con R_a si è indicato l'ammontare annuale delle retribuzioni lorde percepite dal dipendente in esame e con R_m l'ammontare mensile delle retribuzioni lorde relative allo stesso dipendente.

Il rapporto precedente, pur dando un importante contributo, non può essere preso a sé stante per discriminare le unità con valori accettabili da quelle con valori sospetti. Infatti l'indicatore in esame può essere influenzato da diversi fattori, tra i quali si ricordano:

- il rapporto può risultare accettabile anche in presenza di errori di unità di misura, della stessa entità, su entrambe le variabili;
- il rapporto potrebbe risultare non accettabile a causa della possibilità che il dipendente selezionato non abbia lavorato tutto l'anno o tutto il mese di ottobre 2002;
- il rapporto potrebbe risultare non accettabile anche a causa del fatto che le retribuzioni rilevate nel mese di ottobre non siano del tutto rappresentative della retribuzione media mensile a causa di particolari condizioni contrattuali delle imprese selezionate.

Per tener conto del fatto che le retribuzioni rilevate nel mese di ottobre possano non essere del tutto rappresentative della retribuzione media mensili, si è deciso di considerare valori accettabili di I_1 tutti quelli compresi tra 10 e 18.

Nel primo caso, invece, è stato possibile correggere l'unità di misura verificando il valore di altri due indicatori: le retribuzioni orarie annuali e le retribuzioni orarie mensili. In particolare si indicheranno le retribuzioni medie orarie annuali relative ad un dipendente con

$$I_2 = \frac{R_a}{H_a}$$ (retribuzioni orarie annuali)

dove R_a ha lo stesso significato visto in precedenza mentre H_a rappresenta il numero totale di ore nell'anno[2]. Il dominio di accettabilità di questo indicatore è stato individuato grazie all'utilizzo di fonti esterne ed in particolar modo della fonte Sci-Pmi che ha permesso di individuare per strato (divisioni di attività economica per quattro classi di dipendenti) la distribuzione del rapporto retribuzione oraria, sia complessiva che per categorie professionali. Altre fonti utilizzate per i confronti sono state Oros-Inps e GI.

Analogamente a quanto fatto per I_2, si indicheranno le retribuzioni medie orarie mensili con

$$I_3 = \frac{R_m}{H_m}$$ (retribuzioni orarie mensili)

del tutto analogo al precedente.

Nel caso in cui una certa correzione (dividendo o moltiplicando per multipli di 10) effettuata sul valore di una o di entrambe le due tipologie di retribuzione garantisse l'accettabilità di tutte e tre i rapporti finora specificati, tale correzione è stata applicata all'unità in esame. Come ulteriore indicatore utile ad individuare eventuali incongruenze tra i quattro ammontari considerati, è stato utilizzato il rapporto tra le retribuzioni orarie annuali e le retribuzioni orarie mensili, e cioè

$$I_4 = \frac{I_2}{I_3}$$ (rapporto tra retribuzioni orarie annuali e mensili)

con l'ipotesi che tale rapporto non debba allontanarsi troppo dal valore unitario, potendo supporre che nell'arco dell'anno, fatta salva la possibilità di un maggior o minor ricorso nel mese di ottobre alle ore di straordinario (pagate di più delle ore ordinarie), le retribuzioni orarie si mantengano abbastanza costanti. Conseguentemente, come dominio di accettabilità si è scelto l'intervallo [0,8;1,2].

Il confronto congiunto dei valori ottenuti dal calcolo di questi indicatori ha permesso di individuare non solo i record con valori delle retribuzioni e/o ore "sospette" (possibili errori o *outlier*), ma, nella maggior parte dei casi, anche di localizzare la possibile fonte dell'inaccettabilità di uno o più indicatori.

Inoltre, per evitare che la non accettabilità di I_1 dipendesse dal fatto che l'unità in esame avesse lavorato non per tutto l'anno (per esempio, a causa di un'assunzione avvenuta ad anno iniziato o a causa di dimissioni avvenute tra novembre e dicembre dell'anno di riferimento) e/o non completamente nel mese di ottobre 2002 (per esempio, assunzione a metà mese oppure casi di aspettativa non remunerata ecc.), si è corretto tale indicatore con un fattore che tenesse conto di questa eventualità. Per determinare questo fattore è stato necessario utilizzare le informazioni presenti sul questionario e relative ai quesiti indicanti il "numero di giorni a cui si riferisce la retribuzione lorda del mese di ottobre 2002" e il "numero di settimane a cui si riferisce la retribuzione lorda annuale".

In formule, si sono individuati, per ogni unità, i due coefficienti di correzione

$$c_a = \frac{MAX(S)}{NS}$$ e

$$c_m = \frac{MAX(G)}{NG}$$

con *NS* numero di settimane lavorate nell'anno, con *MAX(S)* numero massimo di settimane lavorabili nell'anno, con *NG* numero di giorni lavorati dall'unità in esame nel mese di ottobre 2002 e con *MAX(G)* numero massimo di giorni lavorabili nel mese di ottobre 2002. Inoltre, la determinazione dei due "valori massimi" ha dovuto tener conto delle diverse tipologie di contratto, mentre le informazioni ricavate dal questionario su *NS* e *NG* sono state oggetto di un preventivo passo di controllo e correzione; nei casi dubbi relativamente alla bontà di queste informazioni, si è comunque preferito non applicare alcuna correzione all'indicatore I_1.

[2] L'ammontare delle ore totali, sia annuali che mensili, è stato preliminarmente controllato anche in base ai valori indicati nelle diverse tipologie di ore (effettivamente lavorate, di straordinario, non lavorate ma retribuite) richieste dal questionario. A questo punto della procedura però non si sono effettuate correzioni, se non in casi estremamente evidenti: in pratica, si è verificata la quadratura tra voci e totale e i casi di errore sono stati segnalati per le successive fasi di correzione.

I due fattori correttivi così determinati hanno permesso di calcolare il seguente indicatore

$$I_1^{'} = \frac{R_a c_a}{R_m c_m}$$

che non è altro che l'indicatore I_1 calcolato nel caso in cui per l'unità considerata e per i periodi mancanti le sue retribuzioni fossero le stesse di quelle percepite nei periodi di presenza. È chiaro che nel caso in cui l'unità fosse costantemente presente nell'anno di riferimento, i due fattori correttivi c_a e c_m sono uguali entrambi all'unità e $I_1 = I_1^{'}$.

Le informazioni ricavate dal questionario su queste due variabili (numero di giorni lavorati in ottobre e numero di settimane lavorate nell'anno) sono state quindi anch'esse oggetto di un preventivo passo di controllo e correzione; data però la difficoltà di capire la bontà di queste informazioni, nei casi dubbi si è comunque preferito non applicare alcuna correzione all'indicatore I_1.

L'utilizzo di questo indicatore ha sicuramente favorito l'individuazione di unità corrette ma erroneamente ritenute sospette; in taluni casi, però, esso stesso può aver accentuato le possibili fonti di errore introducendo distorsioni dovute all'utilizzo di due variabili anch'esse affette da errori e spesso non facilmente controllabili. Per questo motivo il controllo effettuato in base a $I_1^{'}$ è stato affiancato a quello effettuato con gli altri indicatori ed è stato utilizzato soprattutto per rendere accettabili casi nei quali, in presenza di valori accettabili degli indicatori I_2, I_3 e I_4, l'indicatore I_1 risultasse esterno all'intervallo di accettabilità.

Il controllo attraverso questi indicatori, pur teoricamente effettuabile in maniera simultanea, in termini pratici ha seguito una metodologia di tipo sequenziale così da essere più facilmente implementabile e controllabile. Le unità soggette a controllo, quindi, sono passate attraverso le seguenti fasi successive:

1. verifica dell'accettabilità delle informazioni sulle retribuzioni annuali e mensili attraverso I_1;
2. calcolo degli indicatori orari e verifica dell'accettabilità di I_4;
3. verifica dell'accettabilità di I_3 e di I_2.

Mentre per il passo 1 è chiaro che l'accettabilità dell'indicatore deriva dalla congruenza della relazione che lega le due retribuzioni (annuale e mensile) e coinvolge solo esse, negli altri passi la procedura si complica perché vengono introdotte le informazioni relative alle ore. Nel secondo e nel terzo passo della procedura, infatti, la mancata accettabilità dell'indicatore pone il problema dell'individuazione di quale tra le due variabili (retribuzioni e ore) determina l'inaccettabilità; nel secondo, poi, la complicazione addirittura aumenta poiché bisogna verificare se è il dato annuale o quello mensile o entrambi a rendere l'indicatore non accettabile.

L'uso di fonti esterne, ancora una volta, può facilitare tale compito. Dalla fonte Oros-Inps è stato possibile individuare, sia a livello annuale che mensile, un valore medio a livello di impresa di retribuzione pro-capite, distinta anche per categoria professionale. Analogamente, la fonte Sci-Pmi ha fornito due tipi di informazioni molto utili:

- il link puntuale tra le imprese rispondenti alle due indagini ha permesso di calcolare un valore di retribuzione media pro-capite dell'impresa
- attraverso lo studio, all'interno di strati precostituiti e per categorie professionali, delle distribuzioni dei due indicatori "retribuzioni per dipendente" e "retribuzioni orarie", si è stati in grado di calcolare sia un valore di retribuzione media pro-capite che un valore di retribuzione media oraria relative allo strato di appartenenza.

Infine, la fonte GI, grazie anche in questo caso ad un link puntuale, ha permesso l'individuazione di indicatori medi per le singole grandi imprese rispondenti alle indagini.

Inoltre, sono state sfruttate anche le informazioni rilevate su quel 68 per cento di fogli dipendenti che hanno superato la fase preliminare di *check*: tenendo conto di tutte le precauzioni del caso (per esempio, la possibile esigua numerosità di dipendenti rilevati nelle singole imprese, soprattutto se si considera la distinzione per categoria professionale), tali informazioni sono state molto utili per determinare le eventuali specificità in termini retributivi o di regime orario di alcune imprese rispetto ad altre; è chiaro che in fase di controllo, una situazione sospetta rispetto a valori medi o mediani di strato per un

determinata unità (dipendente) potrebbe essere ritenuta accettabile se confrontata con una situazione similare rilevata nella stessa impresa.
Uno schema semplificato dei vari passi effettuati in questa fase è riportato nella Figura 1.
Di quel 32 percento di unità, rispetto al complesso di unità rilevate, aventi valori non nulli o non mancanti delle retribuzioni e delle ore, annuali e mensili (Figura 1), il 51 per cento (12.437 unità) ha mostrato valori accettabili per il rapporto I_1: di queste, più del 67 per cento ha evidenziato valori accettabili anche dell'indicatore I_4. Inoltre poco meno della metà delle unità con valori di I_1 non accettabili sono risultate compatibili con l'indicatore I_4 e di queste il 90 per cento ha evidenziato valori delle retribuzioni annuali e mensili congruenti se misurate attraverso l'indicatore I_1', applicando cioè la correzione per numero di giorni e di settimane lavorate.
Di seguito, le unità con valori non compatibili di ore e retribuzioni annuali e mensili sono state analizzate attraverso l'uso dei due rapporti I_2 e I_3: soltanto poco più del 23 per cento è risultato compatibile con uno dei rapporti precedenti[3]. I valori di retribuzioni o ore non compatibili di queste ultime sono stati corretti tenendo conto dei valori riscontrati da altre fonti, in particolare i valori medi di retribuzioni e ore per impresa calcolati sulle unità con tali valori corretti e la fonte Oros-Inps. Le correzioni effettuate sono state del tutto simili sia nel caso che I_2 fosse compatibile e I_3 non lo fosse, che nel caso opposto. Per esempio, se I_2 compatibile (valori annuali ritenuti accettabili) e I_3 non compatibile, si è proceduto a verificare l'accettabilità dei valori mensili e le correzioni sono state le seguenti:

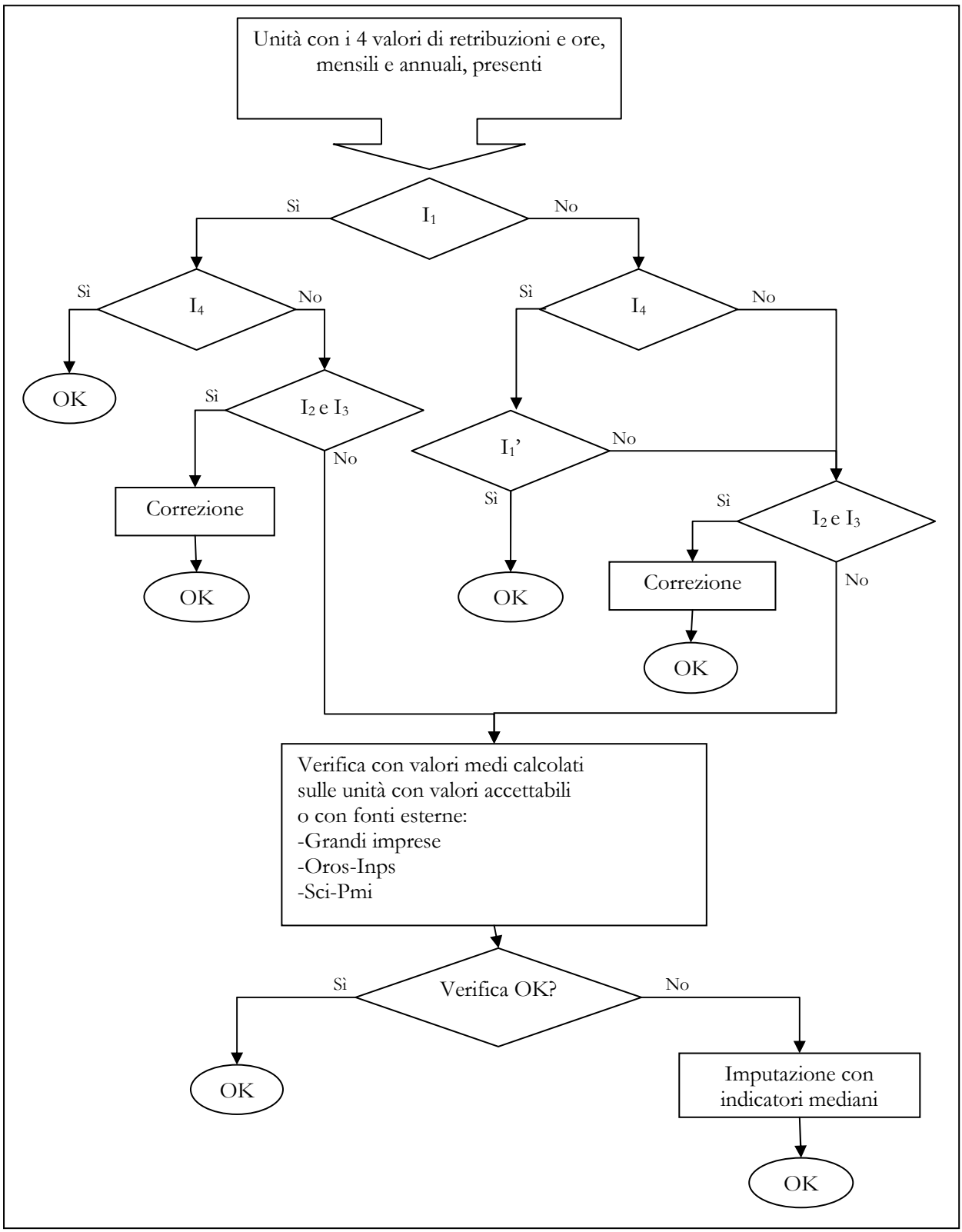
- verifica della presenza di errori di misura ed eventuale correzione interattiva dei valori sospetti;
- in caso di fallimento della prima fase, se le retribuzioni mensili non si discostavano in maniera evidente (più del 20 per cento) rispetto alla fonte considerata (unità con valori corretti o Oros-Inps), tale valore è stato accettato, mentre il valore delle ore è stato stimato in base al medesimo rapporto rilevato per impresa nel sottoinsieme di unità con valori corretti o in base al medesimo rapporto derivato dalla fonte Sci-Pmi, supponendo una uguaglianza tra retribuzioni medie annuali e quelle mensili;
- in caso di scostamento significativo del valore mensile si è preferito stimare il dato sulle retribuzioni in base a una delle due fonti (unità con valori corretti e poi eventualmente Oros-Inps) e verificare la coerenza con il dato sulle ore; in caso di compatibilità del nuovo valore con i *range* definiti per tutti gli indicatori considerati l'unità è stata considerata corretta altrimenti si è proceduto alla stima anche delle ore, in base al rapporto orario derivato da una delle fonti utilizzata per il confronto.

Come detto, nel caso di I_3 compatibile e I_2 non compatibile, la procedura è stata del tutto analoga, con la sola differenza che la verifica dei valori con le altre fonti ha riguardato il valore delle retribuzioni (e/o delle ore) annuali e non quello mensile.
Per le unità non compatibili nemmeno con gli indicatori I_2 e I_3, poiché non necessariamente ritenute aventi valori errati ma considerate come eventuali "spie" di comportamenti particolari messi in atto dalle imprese di appartenenza, i valori di retribuzioni e ore, annuali e mensili, sono stati sottoposti a controllo attraverso tutte le fonti disponibili. In ordine gerarchico di importanza, sono stati ritenuti corretti i valori delle retribuzioni annuali e/o mensili che non si discostavano troppo (meno del 20 per cento) dai valori medi per impresa derivati da:

- sottoinsieme di unità rispondenti con valori corretti; altrimenti
- Indagine sull'occupazione delle grandi imprese; altrimenti
- Oros-Inps; altrimenti
- valori annuali riscontrati in Sci-Pmi.

[3] Per le modalità di costruzione degli indicatori, un dato non compatibile con I_1 e I_4 non può essere compatibile con entrambi gli indicatori I_2 e I_3.

Figura 1: *Schema semplificato della procedura di controllo e correzione per le unità di secondo stadio (dipendenti) con i valori delle retribuzioni e delle ore, annuali e mensili, tutti presenti.*



Le unità che hanno manifestato valori coerenti con una delle fonti precedentemente indicate sono state circa l'85 per cento delle unità considerate; per garantire la coerenza del valore ritenuto accettabile con gli altri valori, si è proceduto ad una nuova fase di controllo in base ai cinque indicatori considerati. In caso di mancata coerenza con uno di essi, il valore coerente con una delle fonti è stato ritenuto esatto ed in base ad esso si sono stimati uno o più degli altri valori coinvolti.
Il restante 15 per cento è stato corretto utilizzando set di indicatori medi in base alla gerarchia di fonti sopra indicata; come ultima risorsa si sono utilizzate le mediane di strato calcolate in base a Sci-Pmi.

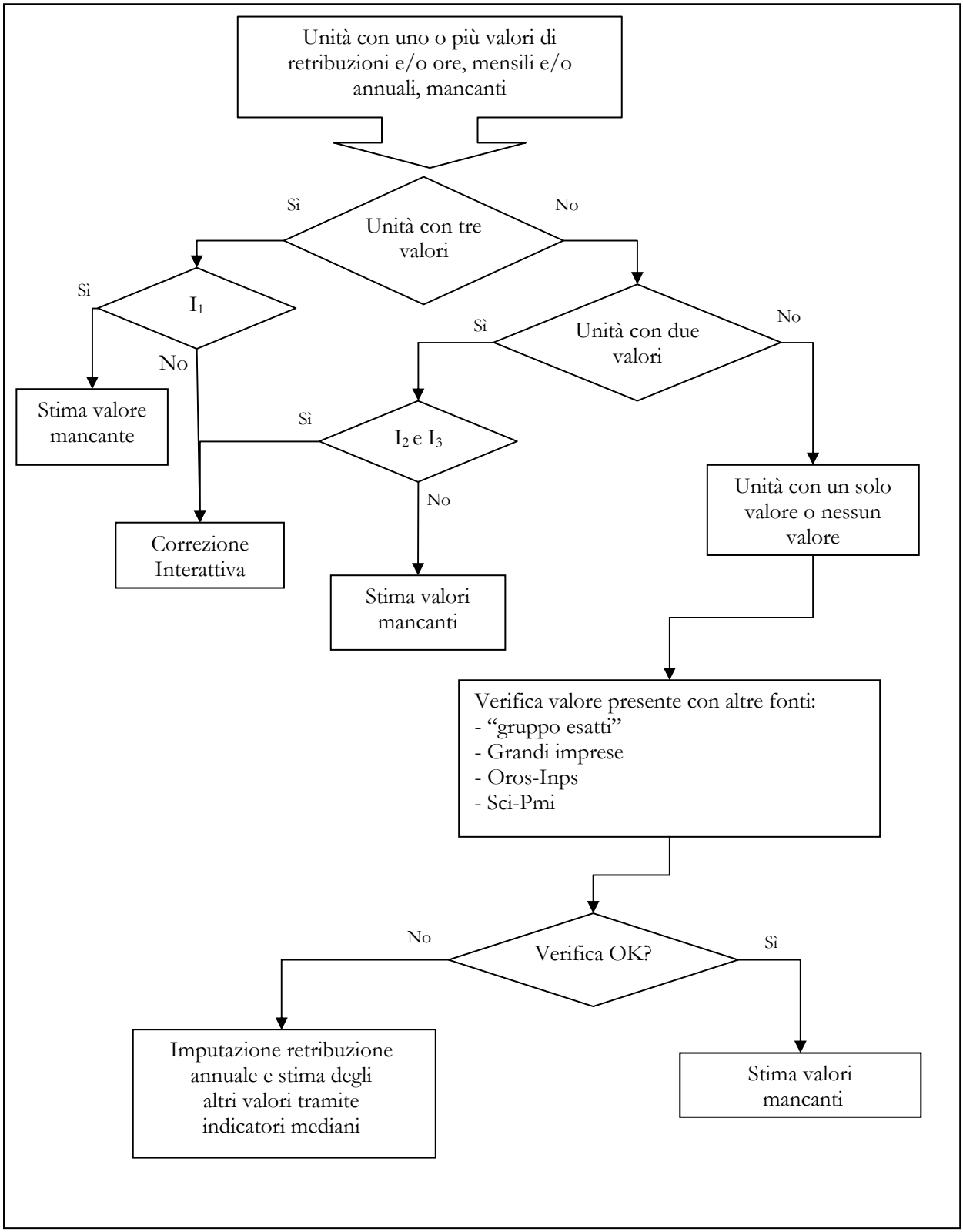
2.1.2. Controllo e correzione dei fogli dipendenti con mancante una o più informazioni relative alle ore e/o alle retribuzioni, annuali e/o mensili.

La procedura di controllo e correzione per questo gruppo di unità è sinteticamente riportato nella Figura 2.
Diversamente alla procedura descritta nel paragrafo precedente, non è possibile, per le unità con uno o più valori mancanti delle retribuzioni e delle ore (4 per cento del totale delle unità sottoposte a procedure di *check*), calcolare sempre e comunque gli indicatori prima considerati.
La procedura è stata quindi costruita per tenere conto sia del numero che del tipo di valore mancante. In pratica si è proceduto nel seguente modo:

1. sono state isolate tutte le unità con tre valori su quattro presenti; tra queste sono state individuate quelle con i valori delle retribuzioni annuali e mensili non mancanti, in modo da poter calcolare l'indicatore I_1. Nel caso di accettabilità di I_1, la variabile mancante sulle ore è stata stimata o usando una delle fonti esterne disponibili o usando il valore della retribuzione oraria disponibile (annuale o mensile) o usando indicatori mediani di strato. Nel caso di non accettabilità di I_1, così come nel caso di mancanza di uno dei due valori delle retribuzioni, data la loro scarsa numerosità (meno del 2% del totale delle unità coinvolte in questa fase del processo), i record sono stati corretti in maniera interattiva.
2. si sono considerate le unità con due valori su quattro presenti; tra queste si sono selezionate le unità per le quali era possibile calcolare o I_2 o I_3. In base all'accettabilità di uno dei due indicatori, si sono stimati i valori mancanti usando le medesime informazioni derivate da fonti esterne (ove disponibili) oppure indicatori mediani di strato. Nei casi di non accettabilità di I_2 o di I_3, i valori mancanti sono stati stimati come nel caso precedente oppure in maniera interattiva;
3. infine, si sono considerate tutte le restanti unità; su quelle con un unico valore presente, tale valore è stato confrontato con le altre fonti, i valori mancanti sono stati stimati attraverso le retribuzioni orarie (annuali o mensili, a seconda dei casi) calcolate su tali fonti; nei casi di totale assenza di valori significativi nelle quattro voci considerate, queste ultime sono state stimate a partire dalle retribuzioni annuali e stimando le rimanenti variabili in base agli indicatori calcolati su una delle fonti prese a riferimento: questa operazione ha coinvolto comunque solo meno del 2 per cento delle unità considerate in questa fase del processo di controllo e correzione.

Per quanto riguarda la verifica delle altre informazioni quantitative presenti sul "foglio dipendenti" (per es., disaggregazione delle ore in ordinarie, straordinarie e retribuite ma non lavorate, o la specificazione delle retribuzioni relative alle ore di straordinario o quelle derivanti da premi ecc.), si è tenuta presente la correzione già effettuata nei passi precedenti. Ovviamente ciò non ha impedito l'eventualità di *feedback* sui dati ritenuti già corretti qualora tali correzioni avessero generato situazioni assolutamente non conciliabili con le altre informazioni. Comunque, in generale, si è trattato di effettuare semplici controlli di quadratura con eventuali errori sanati attraverso operazioni di riproporzionamento.

Figura 2: *Schema semplificato della procedura di controllo e correzione per le unità di secondo stadio (dipendenti) con uno o più valori di retribuzioni e/o ore, annuali e/o mensili, mancanti.*



2.1.3 Controllo e correzione delle diverse tipologie di ore richieste dal questionario.

Una volta verificata la congruenza delle informazioni tra retribuzioni e ore, annuali e mensili, rilevate sui dipendenti, il passo successivo è stato quello di verificare su tutte le unità di secondo stadio la coerenza dell'informazione relativa alla disaggregazione nelle diverse tipologie di ore (ore ordinarie lavorate, dalle ore di straordinario e dalle ore non lavorate ma retribuite), annuali e mensili (al solito relative ad ottobre 2002). Tale procedura, che ha coinvolto circa il 6 per cento delle unità rilevate, si è svolta nel modo seguente:

- verifica dell'esistenza di valori mancanti per le diverse modalità di ore;
- verifica della quadratura tra ore totali e somma delle diverse modalità (annuali e mensili);
- verifica della congruenza tra le diverse tipologie delle ore annuali con le medesime modalità di ore mensili;
- verifica della congruenza dell'informazione relativa alle ore (mensili) di straordinario con quella relativa alle retribuzioni per straordinario;
- ulteriore verifica delle quadrature e delle congruenze tra le diverse tipologie di ore per le unità soggette a correzioni, automatiche e/o interattive;
- individuazione di indicatori mediani di strato calcolati sulle unità con valori ritenuti corretti;
- imputazione delle mancate risposte parziali.

Prima di tutto, quindi, si sono isolate le unità che presentavano valori mancanti per tutte le modalità di ore richieste, o annuali o mensili o entrambe. Questa situazione ha riguardato appena il 2 per cento delle unità analizzate. I valori del restante 98 per cento delle unità sono stati controllati innanzi tutto verificandone la quadratura con il dato complessivo, ritenuto corretto dopo la prima fase di *check*. I casi dubbi, cioè con differenze elevate tra totale e somma delle voci oppure con valori di alcune poste ritenuti "sospetti", sono stati soggetti a una verifica (ed eventuale correzione) di tipo interattivo, basata su un'analisi di coerenza tra le diverse quantità; in questa fase, in pratica, si è proceduto essenzialmente alla correzione delle poste non congruenti con il totale.

Le informazioni sulle ore, corrette e quadrate, annuali e mensili, sono state di seguito sottoposte ad un ulteriore controllo di congruenza tra le medesime voci relative ai due diversi periodi temporali.

Tale analisi di congruenza, data anche la specificità delle informazioni rilevate (il dato mensile può essere molto diverso da quello annuale per motivi contingenti quali per esempio malattia o ferie) si è basato essenzialmente sul controllo che il dato mensile fosse inferiore (o uguale) a quello annuale. In realtà si sono svolti anche dei controlli circa le distribuzioni ma la loro non decisività nell'individuazione di eventuali errori ha consigliato di utilizzare questa analisi solo per segnalare casi molto anomali da verificare interattivamente.

Inoltre, la presenza sul questionario di una informazione diretta relativa alla retribuzione per ore di straordinario relative al mese di ottobre, ha consentito di avere un altro segnale circa la bontà o meno delle informazioni relative alle ore mensili. Il controllo incrociato tra le due informazioni (ore di straordinario e relativa retribuzione) ha permesso di validare i valori ed eventualmente di correggere i casi "sospetti". Tali casi sono stati sanati verificando anche il rapporto esistente tra retribuzioni lorde e retribuzioni per straordinario: qualora tale rapporto risultasse superiore al 20 per cento, si è preferito correggere le retribuzioni per straordinario in base all'informazione relativa alle ore di straordinario; nel caso opposto, si sono stimate le ore di straordinario in base alle retribuzioni ad esse relative e si sono ricalcolate le altre informazioni sulle ore in modo da garantire ancora la quadratura.

Infine, tutte le correzioni effettuate nei passi precedenti hanno implicato un ulteriore passo di verifica della congruenza delle informazioni relative alle altre poste.

A questo punto, tutte le unità che hanno superato questa fase di verifica sono state poi utilizzate per derivare degli indicatori da utilizzare per il successivo passo di imputazione delle mancate risposte. Come indicatori si sono presi dei rapporti di composizione tra le diverse modalità ed il valore delle ore totali, calcolati per divisione di attività economica (o sue aggregazioni) e classi di dipendenti. Il processo di imputazione ha riguardato, comunque, solo l'1 per cento delle unità rilevate.

2.1.4. Controllo e correzione di alcune voci relative alle retribuzioni.

La procedura di controllo e correzione utilizzata per le diverse tipologie di retribuzioni richieste dal modello è risultata alquanto differente rispetto a quella utilizzata per il controllo delle modalità delle ore. La differenza principale è consistita nel fatto che le voci relative alle retribuzioni, essendo solo alcune delle diverse componenti del totale, non necessariamente dovevano quadrare con il totale delle retribuzioni lorde (annuali e/o mensili). Anche in questo caso, il valore complessivo delle retribuzioni è stato considerato già controllato (ed eventualmente corretto) nelle diverse fasi precedenti ed è stato assunto, quindi, come valido elemento di confronto intorno al quale far ruotare tutti i controlli. Inoltre, poiché la variabile relativa alle retribuzioni per straordinario è entrata nella procedura per il controllo e la correzione delle ore di straordinario, anch'essa viene considerata già coerente con il resto delle informazioni e quindi non viene ulteriormente verificata.
Un'altra differenza è dovuta al fatto che i due dettagli delle retribuzioni mensili non corrispondono a quelli rilevati per le retribuzioni annuali: il controllo di queste tre voci[4] è stato effettuato essenzialmente basandosi sulla relazione che le lega, mediamente, al complesso delle retribuzioni mensili e/o annuali. Sono state, quindi, individuate per ognuna di queste voci delle soglie di accettazione, prima più ristrette e poi, dopo un'analisi delle diverse distribuzioni, più allargate in modo da permettere di distinguere casi di errore da correggere rispetto a valori anomali ammissibili. Le unità con valori ritenuti errati sono state quindi messe da parte mentre quelle ritenute corrette hanno contribuito all'individuazione di alcuni indicatori mediani (per lo più rapporti di composizione) calcolati a livello di strato (divisione di attività economica per classi di dipendenti). Tali indicatori, infine, sono stati utilizzati per imputare i dati che in prima istanza erano stati giudicati errati e per i quali non è stato possibile effettuare una idonea correzione di tipo interattivo.
Per la variabile "Retribuzioni per turni di lavoro notturno o festivo" inizialmente è stata scelta una soglia del 10 per cento rispetto al totale delle retribuzioni lorde. L'analisi della distribuzione ha consigliato di alzare tale soglia al 20 per cento: oltre detta soglia, il valore della variabile è stato sottoposto innanzitutto ad una correzione di tipo interattivo (per esempio, per errore di unità di misura) e poi eventualmente imputato. Quest'ultima operazione ha coinvolto comunque meno dell'1 per cento dei rispondenti. Anche per la variabile "Totale annuale dei premi" si è preferito alzare la soglia di ammissibilità dal 10 al 20 per cento rispetto al totale delle retribuzioni annuali: in questo caso, l'imputazione ha riguardato un numero più elevato di unità (circa l'8 per cento). Infine, per la variabile "Premi di risultato" si è effettuato un controllo (in termini di rapporto) con la variabile "Totale annuale dei premi": situazioni con valori dell'indicatore superiori all'unità sono state ovviamente ritenute errate. Esse sono state quindi sottoposte a correzione interattiva e, dove ciò non fosse possibile, si è proceduto ad imputare tali valori per mezzo degli indicatori mediani di strato prima accennati: tale operazione è stata utilizzata in meno dell'1 per cento dei casi.
Le restanti variabili qualitative, relative alle caratteristiche personali dei dipendenti (età, titolo di studio, ecc.) sono state controllate, corrette ed eventualmente imputate attraverso il software Concord, modulo RIDA, che implementa la tecnica probabilistica del donatore.

2.2 Fase di controllo e correzione delle unità di primo stadio (imprese)

La fase di controllo e correzione delle informazioni relative alle imprese (unità di primo stadio) alle quali appartengono le unità di secondo stadio si è basata su un'analisi preliminare sulle caratteristiche strutturali delle imprese. Il controllo di informazioni generali quali il settore di attività economica e il numero di addetti è stato svolto utilizzando l'Archivio Statistico delle Imprese Attive (Asia) per l'anno 2002. Soprattutto per il numero di addetti (e di dipendenti), differenze marcate dei dati rilevati con quelli di archivio hanno indotto un'ulteriore fase di verifica che è consistita, in primo luogo, nel controllo della correttezza del dato registrato con quello presente sul questionario e, in secondo luogo,

[4] Le retribuzioni mensili per straordinario sono, come detto, già state controllate e corrette nella fase precedente.

nel ricontattare le imprese (soprattutto quelle di dimensioni maggiori) al fine di ottenere nei casi dubbi il dato corretto o eventuali spiegazioni sui valori scritti sul questionario. La procedura di controllo e correzione della quale, nel seguito, si illustrano brevemente gli aspetti fondamentali, ha come base questa verifica preliminare: in altri termini, si suppone che le informazioni sul numero totale di addetti e di dipendenti siano corrette e tutto il resto deve essere congruente con queste informazioni.
La procedura di controllo e correzione ha riguardato 9.771 imprese ed è stata svolta attraverso le seguenti fasi:

1. verifica delle informazioni relative al numero medio di addetti e di dipendenti;
2. verifica del quadro relativo al numero di ore lavorate;
3. verifica del quadro relativo al costo del lavoro;
4. verifica delle informazioni di tipo qualitativo.

Le prime tre fasi hanno utilizzato esclusivamente criteri di correzione di tipo deterministico attuati attraverso programmi costruiti *ad hoc*; la quarta fase, invece, ha combinato criteri deterministici su alcune variabili con criteri di tipo probabilistico su altre, utilizzando sia programmi *ad hoc* che software generalizzati di controllo e correzione dei dati di indagine (nella fattispecie Concord, ed in particolare il modulo SCIA).
La procedura utilizzata è stata di tipo gerarchico: la verifica delle informazioni relative al costo del lavoro, per esempio, è stata effettuata supponendo esatte le informazioni su dipendenti e ore lavorate, già verificate nei due passi precedenti. In realtà, dato il forte legame esistente tra le variabili (dipendenti, ore, retribuzioni e costo del lavoro) le prime due fasi sono state eseguite simultaneamente e la terza ha comportato talvolta operazioni di *feedback* qualora l'informazione sul costo del lavoro o sulle retribuzioni risultasse corretta ma incompatibile con quella relativa a ore e dipendenti.
In particolare, in questa terza fase, i risultati sono stati vincolati all'accettabilità dei valori di alcuni indicatori, quali il costo del lavoro orario, le retribuzioni orarie, il rapporto retribuzioni su costo del lavoro e, ovviamente, le ore lavorate per dipendente. Tali indicatori hanno permesso nella maggior parte dei casi sia il trattamento dell'errore (localizzazione e correzione) che degli *outlier*, oltre a fornire gli indicatori mediani di strato utili all'imputazione delle mancate risposte parziali.
Per le prime tre fasi, inoltre, è stato possibile utilizzare per il controllo anche le informazioni provenienti da altre fonti: oltre alle già citate indagini statistiche (GI e Sci-Pmi) e alla fonte amministrativa (Oros-Inps) è stato possibile utilizzare anche le informazioni provenienti dall'elaborazione dei dati relativi alle unità di secondo stadio. In questo caso si è trattato essenzialmente di un controllo di coerenza tra le informazioni provenienti dai due diversi stadi del campione.

2.2.1. Controllo e correzione delle informazioni sugli addetti, sui dipendenti e sulle ore

Come detto, in questa fase il numero complessivo di addetti si ritiene corretto poiché già verificato attraverso il confronto con l'archivio Asia e l'eventuale ricontatto telefonico dell'impresa. In questa fase, relativamente agli addetti e ai dipendenti, la procedura ha svolto le seguenti attività:

- verifica della presenza di mancate risposte parziali alle voci relative al numero di imprenditori, al numero di dipendenti (esclusi gli apprendisti) a al numero di apprendisti. Il numero di imprenditori è stato imputato attraverso il dato esistente nell'archivio Asia, mentre le informazioni relative alle due altre voci sono state imputate utilizzando Asia (per il complesso) e/o informazioni provenienti da Sci-Pmi o attraverso un'analisi di tipo interattivo.
- Verifica della congruenza delle informazioni sul numero di addetti a tempo parziale e sul numero di addetti con contratto a tempo indeterminato rispetto al numero complessivo di addetti per categoria professionale. In questa fase il controllo è stato basato su semplici disuguaglianze e i casi di incoerenza sono stati sanati attraverso metodi di correzione di tipo interattivo.
- Verifica della quadratura delle informazioni per categoria professionale con il totale corrispondente. Anche in questo caso, i casi di incoerenza sono stati corretti interattivamente.

- Verifica della coerenza, in termini di semplice presenza, delle informazioni sugli addetti e delle corrispondenti informazioni sulle ore per categoria professionale. I casi di incoerenza sono stati segnalati per una successiva imputazione: in alcuni casi evidenti (per esempio, slittamento delle informazioni da un campo all'altro) i dati sono stati corretti in maniera interattiva.

Per il quadro relativo alle ore si è, prima di tutto, verificata la quadratura per colonna; questa verifica ha inoltre messo in evidenza anche casi di mancata risposta a una o più voci richieste dal questionario. La procedura di correzione dei dati di questo quadro si è svolta essenzialmente secondo le seguenti fasi:

- correzione interattiva, laddove l'errore e la relativa correzione fossero ben chiari (slittamento delle informazioni da una casella ad un'altra, errori di unità di misura ecc.);
- correzione o imputazione delle voci errate o mancanti relativamente alle ore ordinarie, a quelle straordinarie e quelle retribuite ma non lavorate, per categoria professionale, attraverso:
  - l'indicatore "ore ordinarie lavorate per dipendente" per categoria professionale, per l'impresa in esame, derivato dalle indagini Sci o Pmi, qualora l'impresa stessa fosse stata rispondente anche ad una di queste due;
  - l'indicatore mediano relativo alle "ore ordinarie lavorate per dipendente" per categoria professionale, per strato, derivato ancora dalle indagini Sci-Pmi;
  - l'indicatore mediano relativo alle "ore ordinarie, straordinarie e retribuite ma non lavorate per dipendente" per categoria professionale, derivato dall'elaborazione delle informazioni desunte dalle unità di secondo stadio (fogli dipendenti).

Si può notare che con questa procedura la correzione delle informazioni relative alle ore ordinarie lavorate può avvenire a diversi livelli e, grazie anche al ricorso alla fonte Sci-Pmi, in maniera abbastanza soddisfacente. Per le altre tipologie di ore, invece, la correzione è affidata esclusivamente alle informazioni parziali, perché rilevate su un numero di unità non sempre comprensivo di tutto il personale utilizzato dall'impresa, e quindi potenzialmente distorte, che derivano da elaborazioni a livello di singola impresa effettuate sui corrispondenti fogli dipendenti. Inoltre, una procedura di correzione del quadro relativo alle ore che non tenga conto delle informazioni provenienti dal quadro relativo agli addetti non può ritenersi del tutto soddisfacente: è chiaro che la correzione di variabili dei due quadri in presenza di un forte legame possa causare errori ancora più gravi se effettuata a prescindere da questo legame.

Per questi motivi le due fasi precedenti sono state svolte esclusivamente per verificare in prima analisi eventuali errori di quadratura. La verifica delle informazioni desunte dai due quadri, perciò, sia per quel che riguarda la coerenza interna delle informazioni, sia per quanto riguarda l'entità vera e propria dei valori rilevati, è stata effettuata facendo uso di alcuni indicatori pro-capite ed in particolar modo delle ore ordinarie, straordinarie e retribuite ma non lavorate per dipendente, per categoria professionale.

Lo studio delle distribuzioni per strato (divisioni di attività economica, o loro aggregazioni, per classe di dipendenti) di questi tre indicatori ha permesso di individuare casi incoerenti e/o casi anomali. In caso di errore si è proceduto nel modo seguente:

- verifica e correzione interattiva dei valori sospetti per le unità influenti;
- per le altre unità, verifica e correzione automatica degli errori di unità di misura;
- verifica della presenza di errori derivanti dalle due precedenti fasi di correzione ed eventuale ripristino dei valori originali in precedenza corretti ed intervento su altre variabili al fine di una maggiore coerenza complessiva delle informazioni rilevate per l'unità in esame;
- imputazione delle mancate risposte parziali e dei valori anomali non verificati e/o corretti nelle fasi precedenti attraverso:
  - dato proveniente dalle indagini Sci-Pmi o GI, qualora esso fosse presente;
  - dato proveniente dall'elaborazione delle informazioni derivate dalle unità di secondo stadio, qualora il numero di dipendenti rilevati fosse uguale al numero complessivo dei dipendenti presenti nell'impresa;
  - indicatore medio per impresa proveniente dall'elaborazione delle informazioni rilevate sulle unità di secondo stadio;
  - indicatore mediano di strato delle ore ordinarie pro-capite proveniente da Sci-Pmi;

- indicatore mediano di strato delle ore straordinarie e delle ore retribuite ma non lavorate pro-capite provenienti dalle imprese rilevate e ritenute corrette.

Una procedura analoga è stata infine utilizzata per la verifica delle informazioni relative al personale a tempo parziale utilizzato dall'impresa. In mancanza di informazioni più dettagliate, le correzioni sono avvenute ipotizzando un comportamento analogo degli indicatori pro-capite tra personale a tempo pieno e personale part-time.

2.2.2. Controllo e correzione delle informazioni sul costo del lavoro e sulle retribuzioni

Questa fase del processo di controllo e correzione si è basata sulla costruzione di tre indicatori e sulla verifica della loro accettabilità. Gli indicatori presi in considerazione sono stati i seguenti:
costo del lavoro orario:

$$I_1^* = \frac{CL}{H};$$

retribuzioni orarie:

$$I_2^* = \frac{R}{H};$$

rapporto tra retribuzioni e costo del lavoro:

$$I_3^* = \frac{R}{CL};$$

dove con *CL* si è indicato il costo del lavoro, con *H* il numero di ore e con *R* le retribuzioni lorde; inoltre, per garantire anche la coerenza con il quadro relativo ai dipendenti, si è considerato un quarto indicatore:
ore lavorate per dipendente:

$I_4^* = \frac{H}{DIP}$, con *DIP* ad indicare il numero complessivo di dipendenti.

I quattro indicatori sono stati calcolati per tutte le imprese del campione e sono stati usati per individuare le imprese con valori corretti: su 9.771 imprese, più dell'81 per cento hanno mostrato tutti e quattro gli indicatori negli intervalli prestabiliti.
La procedura di controllo ed eventuale correzione dei valori sospetti è stata svolta in maniera gerarchica, verificando prima la bontà del valore relativo al costo del lavoro e verificando poi via via i valori relativi alle retribuzioni lorde in denaro e alle altre voci del costo del lavoro (contributi sociali effettivi a carico del datore di lavoro e gli altri costi del personale); infine, in ultima analisi, si sono verificate le due voci retributive considerate nel questionario relative al Totale dei premi corrisposti e ai Premi di risultato.
Dato che il numero di ore lavorate sono state controllate ed eventualmente corrette nei passi precedenti, in base anche alla loro relazione con il numero di dipendenti, l'utilizzo dell'indicatore I_1^* ha permesso di individuare il 9 per cento circa di imprese con un valore sospetto del costo del lavoro totale. Per queste, in più dell'84 per cento dei casi, il valore del costo del lavoro stimato attraverso la somma delle sue componenti garantiva l'accettabilità dell'indicatore I_1^*: in questi casi, quindi, il totale è stato sostituito con detta somma.
Il restante 16 per cento di imprese con valori dell'indicatore fuori *range* è stato invece analizzato osservando anche i valori riscontrati per I_2^* e I_3^*: nei casi di accettabilità dei valori di tali indicatori con quelli di I_1^* prossimi agli estremi dell'intervallo di accettabilità[5] (54 per cento), il dato sul costo del lavoro viene ritenuto corretto e le sue diverse voci vengono stimate assegnando la differenza proporzionalmente alle singole voci. Per le rimanenti imprese, invece, si è confrontato il dato rilevato

[5] In questo contesto si intendono "prossimi" tutti quei valori che differiscono da uno dei due estremi della regione di accettabilità dell'indicatore meno del 10 per cento del valore dell'estremo stesso.

con quello presente nella fonte GI o in quella Oros-Inps, laddove presenti[6]. Nei casi di differenze degli indicatori del tipo I_1^* costruiti sui dati GI o Oros-Inps rispetto a I_1^* calcolato sui dati rilevati inferiori al 10%, il valore rilevato è stato ritenuto accettabile; in caso contrario, lo stesso è stato stimato attraverso la fonte GI o la fonte Oros-Inps, se presenti, o in maniera interattiva nei casi evidenti (per esempio, errori di unità di misura di una o più voci).

Anche per le imprese con un valore dell'indicatore nei parametri prefissati è stata verificata la quadratura delle diverse voci del costo del lavoro con il totale rilevato: nel 10 per cento dei casi, i due valori sono risultati diversi e si è agito correggendo in modo proporzionale al totale le diverse voci costituenti il costo del lavoro.

Infine, qualora il dato relativo al costo del lavoro non fosse risultato accettabile, in base all'indicatore I_1^* o in base al confronto con le altre fonti, esclusivamente per un errore o una mancanza nella voce "Altri costi", si è proceduto con la stima di tale informazione in base all'indicatore mediano per strato "Altri costi su Costo del lavoro" costruito sui rispondenti: ciò ha riguardato comunque soltanto 48 imprese che rappresentano circa lo 0,5 per cento del totale delle imprese rilevate.

Una volta corretto il costo del lavoro, il passo successivo è stato quello di verificare l'accettabilità del dato relativo alle retribuzioni lorde. L'indicatore preso come primo segnale di tale accettabilità è stato I_2^*: poiché le ore si ritengono già corrette, un eventuale scostamento dell'indicatore dai limiti prefissati si ritiene imputabile al valore delle retribuzioni. Quasi la totalità delle unità rilevate (99,4 per cento) ha mostrato valori dell'indicatore all'interno del *range* scelto in base alle conoscenze del fenomeno da parte del ricercatore. Le imprese residue sono state oggetto di ulteriore verifica effettuata in base all'indicatore I_3^*: ben l'88 per cento di queste ha mostrato valori di questo indicatore interni all'intervallo di accettabilità. Le poche unità con valori delle retribuzioni ancora non accettabili sono state corrette in maniera interattiva da revisori esperti.

Inoltre, lo stretto legame che esiste tra costo del lavoro e retribuzioni ha suggerito l'effettuazione di un ulteriore controllo della variabile retribuzione in base all'indicatore I_3^* anche per tutte le altre unità che mostravano valori accettabili dell'indicatore I_2^*. L'applicazione di I_3^* ai dati rilevati ha evidenziato l'8 per cento circa di unità con valori dell'indicatore fuori *range*. Per queste unità si è effettuato il confronto tra gli indicatori calcolati sui dati rilevati con i medesimi indicatori costruiti in base ai dati provenienti dalla fonte Oros-Inps; da tale confronto sono scaturite le seguenti tre situazioni:

- nei casi in cui tutti e due gli indicatori calcolati sulla fonte amministrativa fossero risultati interni al corrispondente intervallo di accettabilità e il valore del costo del lavoro calcolato in Oros-Inps risultasse differente da quello rilevato non più del 10% (52 per cento circa dei fuori *range*), si è deciso di ritenere corretto il dato di indagine relativo al costo del lavoro e di stimare il valore delle retribuzioni e dei contributi in base agli indicatori calcolati su base amministrativa;
- nei casi, invece, in cui tutti e due gli indicatori calcolati sulla fonte amministrativa fossero risultati interni al corrispondente intervallo di accettabilità ma con differenze di valori del costo del lavoro rispetto a quelli rilevati di entità superiori al 10% (30 per cento di unità con valori di I_3^* fuori *range*), si è preferito sostituire completamente i valori rilevati con quelli della fonte amministrativa;
- infine, nei casi in cui entrambi gli indicatori calcolati sulla fonte amministrativa non fossero risultati interni a limiti ragionevoli (18 per cento circa delle unità con valori sospetti), il dato di indagine è stato corretto in maniera interattiva tenendo conto dell'eventuale presenza di informazioni dalle due fonti GI e Sci-Pmi, o imputato in base all'indicatore I_3^* mediano di strato calcolato sulla fonte Sci-Pmi.

Al termine di questo processo, i valori relativi al costo del lavoro e alle retribuzioni sono ritenuti corretti e quindi verranno utilizzati nel seguito come punto di riferimento per il controllo e l'eventuale

[6] Per entrambe le fonti, non essendo disponibile il dato relativo agli "Altri costi del personale", il costo del lavoro è stato stimato sommando alle retribuzioni e ai contributi sociali desunti da tali fonti, il dato sugli "Altri costi" derivato dall'indagine stessa. Per questo motivo, gli indicatori costruiti utilizzando la variabile costo del lavoro per la fonte GI e per la fonte Oros-Inps, devono essere trattati con cautela.

correzione delle altre voci del costo del lavoro e per le voci delle retribuzioni rilevate dal questionario (premi corrisposti e premi di risultato).

2.2.3. Controllo e correzione delle altre voci del costo del lavoro e delle voci relative alle retribuzioni rilevate dal questionario

Il controllo della variabile "Contributi sociali effettivi a carico del datore di lavoro" è stato effettuato basandosi essenzialmente sull'utilizzo del seguente indicatore:

$$I_5^* = \frac{CS}{R}$$

dove con *CS* si intende la variabile in esame e con *R* si intende la variabile Retribuzioni lorde in denaro.
La definizione di un intervallo di accettabilità dell'indicatore, effettuata in base all'esperienza del ricercatore e attraverso lo studio delle distribuzioni per strato dell'indicatore stesso rilevato, ha permesso di evidenziare circa un 11 per cento di imprese con valori "sospetti" della variabile in oggetto e per le quali, quindi, si è resa necessario una verifica più puntuale. I casi di mancata risposta alla variabile Contributi sociali sono stati invece isolati: tale situazione ha richiesto un successivo passo di stima del valore della variabile.
Per le unità con valori di I_5^* fuori *range*, laddove possibile, è stato calcolato il medesimo indicatore in base alle informazioni desunte dalla fonte Oros-Inps; si sono osservate le seguenti situazioni:

- i casi di evidenti errori di unità di misura sono stati corretti in maniera interattiva;
- nei casi in cui l'indicatore calcolato su base amministrativa fosse compatibile con l'intervallo di accettabilità prescelto, il dato relativo ai Contributi sociali è stato stimato applicando questo indicatore al valore delle retribuzioni rilevato;
- i casi in cui si è potuta ragionevolmente supporre una inversione durante la fase di compilazione del questionario dei valori relativi ai Contributi sociali e agli Altri costi sono stati sanati riassegnando in maniera corretta i valori alle due variabili;
- nei casi in cui sia l'indicatore calcolato sui dati rilevati che quello derivato dalla fonte amministrativa, pur entrambi fuori *range*, comunque molto vicini tra loro (differenze inferiori al 10%), il dato relativo ai Contributi sociali è stato ritenuto corretto;
- nei casi in cui nessuna delle situazioni precedenti ha portato ad una correzione del valore della variabile in esame, tale valore è stato azzerato e successivamente imputato in base all'indicatore I_5^* mediano di strato, derivato dal gruppo di unità con valore dell'indicatore ritenuto corretto.

Inoltre, tutte quelle imprese per le quali non erano disponibili informazioni derivate dalla fonte Oros-Inps (4 per cento circa delle unità con valori di I_5^* non accettabili) ma con valori dell'indicatore prossimi agli estremi dell'intervallo (distanti cioè meno del 10%) sono state ritenute corrette, mentre per le rimanenti (7 unità) il dato sui Contributi sociali è stato imputato in base all'indicatore mediano di strato e al valore delle retribuzioni rilevato.
Infine, la voce relativa agli Altri costi è stata utilizzata soprattutto come voce residuale per cui il controllo su di essa è avvenuto semplicemente verificando che fosse garantita la quadratura delle diverse voci con il totale del costo del lavoro.
Il controllo dei valori rilevati per le variabili "Totale premi corrisposti" e "Premi di risultato" è stato effettuato verificando la relazione esistente tra esse e la variabili Retribuzioni. La correzione dei casi sospetti è avvenuta essenzialmente in maniera interattiva, valutando anche la congruenza delle informazioni a livello di impresa con quelle riscontrate sulle unità di secondo stadio (fogli dipendenti); nei casi dubbi si è preferito imputare il valore relativo al totale dei premi e/o ai premi di risultato utilizzando il rapporto di queste rispetto alle retribuzioni, calcolato in base alle informazioni provenienti dai fogli dipendenti relativi all'impresa in esame.

2.2.4. Controllo e correzione delle variabili di tipo qualitativo

In base alla definizione di un piano di compatibilità che tenesse conto delle relazioni interne alle variabili qualitative del questionario, il controllo e l'eventuale correzione sono avvenuti seguendo tre fasi:

1. Correzione deterministica su variabili *multiresponse* o variabili filtro;
2. Correzione e imputazione delle mancate risposte parziali con metodo probabilistico per tutte le altre variabili (escluse le percentuali);
3. Correzione o stima delle variabili esprimenti percentuali attraverso l'uso di rapporti di composizione.

La prima fase, come detto, ha coinvolto essenzialmente i quesiti sulle modalità di organizzazione del lavoro almeno per quel che concerneva l'utilizzo o meno di tali modalità e altri quesiti (per esempio quelli sulla contrattazione collettiva o sulle materie oggetto di contrattazione integrativa), limitatamente alle variabili filtro e *multiresponse*. La verifica delle prime è avvenuta tenendo conto dell'esistenza di informazioni significative presenti nei blocchi sottostanti ad esse.
Per le variabili *multiresponse* la correzione si basa sull'ipotesi che l'alta presenza di valori mancanti sia dovuta essenzialmente alla mancata indicazione delle risposte negative, quando la modalità in esame non interessava le specifiche aziendali. Quindi, nei casi di mancata risposta (ad una o più modalità) e in presenza di almeno una risposta positiva (ad una delle altre modalità), tale mancata risposta è stata imputata deterministicamente come risposta negativa.
Tutte le altre variabili, ad esclusione di quelle esprimenti percentuali, sono state successivamente sottoposte ad una fase di controllo e correzione di tipo probabilistico attraverso l'utilizzo del software generalizzato Concord ed in particolare del modulo SCIA basato sulla metodologia di Fellegi-Holt. In questa fase, le variabili filtro corrette deterministicamente nella fase precedente sono state considerate come punti di partenza per l'imputazione delle restanti variabili; per cui sono entrate nel processo di correzione in maniera, però, da risultare immodificabili.
Infine, l'imputazione di valori mancanti relativi a variabili esprimenti valori percentuali (di dipendenti sul totale) si è basata sull'utilizzo di rapporti mediani costruiti su un sottoinsieme di unità con valori ritenuti corretti e suddivise in strati omogenei in base al settore di attività economica (divisione) e alla dimensione aziendale (classe di dipendenti): il valore di questi rapporti è stato usato per imputare tout-court le percentuali mancanti qualora la modalità in esame prevedesse, a seguito di una risposta positiva, la specificazione di detta percentuale.

3. Alcuni indicatori relativi al processo di controllo e correzione

Per valutare l'entità dell'intervento sui dati del processo di correzione si possono osservare alcuni indicatori che vengono riportati nella Tavola 1. I valori di tali indicatori, (in termini percentuali) riferiti alle unità di primo e di secondo stadio, sono stati calcolati considerando solo otto e dieci, rispettivamente, delle variabili quantitative ritenute più importanti. In particolare per le imprese si sono considerate le variabili relative al totale dipendenti, al totale delle ore, ordinarie, di straordinario e non lavorate ma retribuite, nonché quelle relative al costo del lavoro e alle sue componenti. Per le unità di secondo stadio, invece, sono state utilizzate le ore (totali e relative alle sue diverse componenti) e le retribuzioni, sia annuali che mensili.
Si può notare che il tasso di imputazione è più basso nelle unità di secondo stadio rispetto a quello calcolato per le unità di primo stadio (9,0 per cento contro 14,1 per cento). Ciò è dovuto essenzialmente alla maggiore attenzione data, in fase di revisione e di pre-registrazione, alle sezioni del questionario relative alle informazioni sui singoli dipendenti. In entrambi i casi, comunque, si può notare la scarsa presenza di valori mancanti relativamente alle variabili fondamentali dell'indagine: il tasso di imputazione netta infatti non arriva all'1 per cento né per le unità del primo stadio né per quelle di secondo stadio. Inoltre l'elevata percentuale di modificazione (superiore al 90 per cento) mostra che la maggior parte delle imputazioni hanno riguardato modifiche di valori non mancanti e non la stima di

un dato mancante, il che dimostra il buon lavoro effettuato in fase di revisione e di pre-registrazione che ha permesso di ridurre al minimo l'esistenza dei valori mancanti.

Tavola 1: *Alcuni Indicatori relativi alla qualità del processo di correzione dei dati (valori percentuali) relativi a dieci variabili rilevate sulle unità di secondo stadio (dipendenti) e ad otto variabili rilevate sulle unità di primo stadio (imprese)***.**

Indicatori	Valori percentuali	
	Unità di primo stadio (imprese)	Unità di secondo stadio (dipendenti)
Tasso di imputazione	14,1	9,0
- Tasso di modificazione	13,2	8,1
- Tasso di imputazione netta	0,9	0,8
- Tasso di cancellazione	0,0	0,1
Tasso di Non Imputazione	85,9	91,0
- Tasso di valori missing non modificati	0,0	0,0
- Tasso dei valori non missing non modificati	85,9	91,0
Percentuale di modificazione	93,6	90,3
Percentuale di imputazione netta	6,1	8,9
Percentuale di cancellazione	0,2	0,8
Percentuale di valori missing non modificati	0,0	0,0
Percentuale di valori non missing non modificati	100,0	100,0
Primo Quartile del Tasso di imputazione (variable)	12,4	7,1
Terzo Quartile del Tasso di imputazione (variable)	16,2	10,5
Primo Quartile del Tasso di imputazione (record)	0,0	0,0
Terzo Quartile del tasso di imputazione (record)	25,0	10,0

Bibliografia

Barcaroli G., D'Aurizio L., Luzi O., Manzari A., Pallara A.. *Metodi e software per il controllo e la correzione dei dati*. Documenti ISTAT n. 1/1999.

Barcaroli G., Luzi O. (2004). *Situazione attuale e prospettive sui metodi e software per il controllo e la correzione dei dati*. Lucidi del seminario del 16/2/ 2004.

Grande E., Luzi O.. *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat*. Contributi ISTAT, n.6/2003.

Manzari A.. *Aspetti generali sulle procedure di controllo e correzione dei dati*. Dispense del corso Concord, 2006.

Manzari A.. *Cenni sulla valutazione delle procedure di controllo e correzione dei dati*. Dispense del corso Concord, 2006.

Riccini E.. *Seminario Concord*, Istat, 15 febbraio 2001.

Statistics Canada. *Quality guidelines*. Third edition, 1998.

Analisi dei metodi di controllo e correzione dei dati della rilevazione sul sistema dei conti delle imprese

Roberto Nardecchia, *Istat , Direzione Centrale delle Statistiche Economiche Strutturali, SSI*

Sommario: La procedura di controllo e correzione dei dati della "Rilevazione sul sistema dei conti delle imprese" è un insieme integrato di diverse metodologie. L'innovazione fondamentale che ha permesso notevoli guadagni in termini di tempestività ed accuratezza delle stime finali è sicuramente rappresentata dall'utilizzo dei dati dei bilanci civilistici. Essi costituiscono un valido riferimento sia per la fase di *editing* che per l'imputazione delle mancate risposte parziali e totali. Nel presente lavoro viene analizzato l'intero processo di controllo e correzione dell'indagine evidenziando in particolare i vantaggi derivanti dall'impiego delle informazioni di fonte amministrativa.

Parole chiave: Rilevazione sul sistema dei conti delle imprese, bilanci civilistici, imputazione.

1. Caratteristiche della rilevazione

La Rilevazione sul sistema dei conti delle imprese (SCI) ha come campo di osservazione l'insieme delle imprese italiane con almeno 100 addetti operanti nei settori industriali e dei servizi; sono escluse dalla popolazione di riferimento le imprese appartenenti ad alcune divisioni dell'attività di intermediazione monetaria e finanziaria, delle assicurazioni e dei servizi domestici. L'indagine è quindi totalitaria ed ha cadenza annuale.

L'archivio di riferimento dell'indagine è costituito dall'archivio statistico delle imprese attive (ASIA) realizzato dall'Istat sulla base dell'integrazione di varie fonti, sia di natura amministrativa che statistica.

I dati raccolti rilevano importanti variabili riguardanti gli aspetti economici, finanziari e patrimoniali dell'azienda. In particolare le voci di bilancio vengono richieste secondo lo schema stabilito dalla IV Direttiva CEE al fine di soddisfare il regolamento comunitario sulle statistiche strutturali N. 58/97 (SBS).

L' unità di rilevazione è costituita dall'impresa; le informazioni richieste riguardano sia l'azienda nel suo complesso (modello SCI) sia le eventuali unità funzionali (modello SCI-UF) al fine di disaggregare alcuni dei risultati per attività economiche omogenee. Complessivamente sono coinvolte dalla rilevazione circa 10.000 unità.

Le informazioni rilevate consentono di esaminare i principali aspetti della gestione aziendale ed a livello aggregato i flussi dei ricavi e dei costi sono utilizzati per il calcolo del valore aggiunto nell'ambito dei conti economici nazionali e della tavola intersettoriale dell'economia italiana.

Il questionario si compone di 8 sezioni. La prima contiene tutte le voci relative al conto economico (fatturato, valore della produzione costi della produzione), la seconda registra alcune voci dello stato patrimoniale, mentre la terza e la quarta si riferiscono agli aspetti occupazionali; in particolare, la terza sezione richiede informazioni circa l'occupazione totale e gli addetti suddivisi per qualifica professionale e sesso, la quarta sezione riguarda i costi sostenuti per il personale. La quinta sezione registra l'acquisizione dei capitali fissi effettuata nell'esercizio, mentre la sesta riguarda un insieme di altri dati. Nella settima sezione si richiedono alcuni dati disaggregati per regione. La sezione 8 è utilizzata come indagine multiscopo (esternalizzazioni, relazioni internazionali, comportamento sociale).

Il questionario contiene 352 variabili quantitative : le variabili economiche sono espresse in valore (migliaia di euro), l'occupazione in unità mentre le ore lavorate sono espresse in migliaia. Inoltre, il questionario personalizzato, contiene le variabili anagrafiche che identificano l'impresa e le variabili di classificazioni caratterizzanti l'attività produttiva (Ateco, localizzazione).

La tecnica di indagine, a partire dal 2003, ha subito una profonda innovazione; parallelamente al tradizionale invio del questionario cartaceo per via postale è stato introdotto il questionario elettronico. La modalità telematica ha apportato un significativo miglioramento della qualità delle informazioni raccolte ed una sensibile diminuzione dei tempi di risposta ed ha permesso di prevenire e monitorare alcuni degli errori non campionari caratteristici della fase di raccolta e registrazione dei dati. Con il questionario elettronico alcuni controlli sono stati avvicinati al rispondente (quadrature, unità di misura). Inoltre, le tecniche di acquisizione automatica controllata configurabili come un *pre-editing,* hanno consentito di individuare ed eliminare all'origine numerosi errori non campionari (verifica codice, doppioni, incompletezza nei dati, variabili doppie, addetti, stato dell'impresa, ateco).

2. La procedura di controllo e correzione.

2.1 Il piano di compatibilità: individuazione e trattamento degli errori

Il processo di controllo e correzione dell'indagine è di tipo misto; le modalità, automatiche ed interattive, vengono utilizzate in combinazione al fine di individuare e trattare gli errori non campionari presenti nei dati raccolti.

In particolare, alle procedure automatizzate sono affidate le operazioni di individuazione e di correzione degli errori che hanno minor impatto sulle stime finali mentre gli errori influenti sono risolti con controllo interattivo al fine di ottenere un insieme di dati coerente e completo.

L'*editing* automatico utilizza un software specificatamente sviluppato ed è effettuato seguendo un approccio di tipo deterministico; esso risolve tutti gli errori compresi in un *range* di tolleranza.

L'*editing* interattivo è invece di tipo *micro,* prevede cioè che tutte le unità rilevate siano sottoposte a controllo e correzione. Ad ogni situazione di incompatibilità il check fa seguire contestualmente l'indicazione delle variabili che non rispettano gli edit attraverso una 'descrizione dell'errore' che i revisori analizzano per il processo di imputazione e correzione.

Al fine di limitare l' *over editing* le azioni che sono intraprese dal piano di controllo sono classificabili in tre tipologie:

- ✓ **Forzature -** quando un valore viene sostituito con un altro in presenza di differenze comprese nel *range* di tolleranza
- ✓ **Accertamenti -** quando una determinata situazione può far sorgere il sospetto di presenza di valori errati e va quindi verificata specificatamente
- ✓ **Errori** – quando si rileva una incongruenza grave fra le informazioni relative alla stessa unità o tra unità collegate (UF).

La procedura è di tipo gerarchico, i vincoli sono individuati sulla base delle relazioni esistenti tra le variabili rilevate; in particolare, per ogni sezione del questionario, i totali sono considerati valori più accurati dei parziali, sulla base dell'ipotesi che l'impresa ha maggiore informazione per essi rispetto ai singoli *item* che li compongono.

Gli *edit* di tipo lineare controllano se certe variabili (o somme algebriche) assumono valori uguali, minori o maggiori di altre variabili.

Gli *edit* di tipo rapporto esprimono, invece, valori medi e rapporti caratteristici.

La procedura di controllo e correzione esegue:

1 → Controllo dell'unità di misura
2 → Controllo valori negativi
3 → Ricostruzione dei valori totali mancanti in presenza degli addendi
4 → Controllo consistenza (addetti<100)
5 → Ricostruzione addendi in caso di sola presenza dei valori totali
6 → Controllo quadratura
7 → Controlli incrociati tra sezioni
8 → Controllo congruenza dati.

Gli errori di coerenza o di incompatibilità che si possono verificare sono sia di tipo *intra-record* quando riguardano variabili appartenenti alla stessa impresa, sia *inter-record* quando le informazioni disaggregate per unità funzionali contraddicono quelle che si riferiscono all' impresa ad esse collegate.

Durante il processo di *editing* un utile confronto per i valori anomali è sicuramente costituito dalla disponibilità di dati relativi a precedenti edizioni della rilevazione che consentono di effettuare controlli ed analisi longitudinali permettendo al revisore di avere a disposizione, *on line* durante il *chek*, i dati della stessa impresa relativi all'anno precedente (se rispondente).

Le correzioni automatiche (forzature) che risolvono le situazioni attivate dai controlli di quadratura e di ricostruzione dei totali e degli addendi si basano sulle regole di *riproporzionamento* e di *arrotondamento*.

Ad esempio, se il valore dei ricavi (voce 11100 del questionario) differisce dalla somma dei suoi addendi (11101, 11102, 11103, 11104, 11105, 11106, 11107) e tale differenza è in valore assoluto inferiore a 50 mila euro (tolleranza), gli addendi sono riproporzionati con un coefficiente ottenuto rapportando il valore del totale (11100) alla somma errata degli addendi; gli eventuali arrotondamenti sono effettuati sull'addendo maggiore.

I controlli incrociati tra sezione consentono di verificare e rendere coerente l'informazione relativa alla stessa unità. Ad esempio il valore "dell'utile o perdita d'esercizio" presente nello Stato Patrimoniale

(25400) deve coincidere con quello riportato nel Conto Economico (19000); allo stesso modo il valore totale dei Costi del personale riportato nella sezione 4 (44000) deve essere uguale ai costi per personale (12400) riportati nella sezione 1 – Conto Economico.

Le incongruenze nei dati si manifestano, invece, quando i valori delle voci '*di cui*' sono maggiori dei totali o quando gli addetti rilevati si discostano significativamente da quelli registrati da ASIA.

Ad ogni impresa durante il controllo e correzione, viene assegnato un codice di qualità che migliora man mano che i campi sono corretti; il processo termina solo quando l'impresa ha raggiunto il livello di qualità stabilito.

2.2 Imputazione delle variabili errate o mancanti

Nella strategia di correzione, le tecniche di imputazione dei valori anomali e quelle di integrazione delle mancate risposte parziali vengono utilizzate congiuntamente: l'accuratezza e la coerenza dei risultati finali dell'indagine è infatti garantita solo se l'insieme dei dati elementari non contiene record errati o incompleti.

Generalmente, per ciascuna impresa, l'imputazione delle variabili segue un approccio di tipo deterministico e viene effettuata in base all'insieme delle restanti risposte valide o attraverso funzioni analitiche che esprimono le relazioni di bilancio esistenti tra le variabili del questionario.

Inoltre, la disponibilità dei dati amministrativi consente di sostituire alcuni dei valori mancanti con l'informazione proveniente dai bilanci civilistici che l'imprese di capitale depositano presso le camere di commercio.

Tale metodologia, come vedremo in seguito, ha assunto negli ultimi anni un ruolo fondamentale anche per l'integrazione delle mancate risposte totali ai fini della tempestività e della riduzione dell'onere statistico sulle imprese.

Frequentemente i dati sono imputati tramite relazioni lineari in cui la variabile ausiliare è costituita dagli addetti ed i coefficienti sono valori medi. Alcuni rapporti hanno un *range* di variazione pressoché costante nel tempo, come ad esempio le ore per dipendente; altre volte i coefficienti sono stimati con le informazioni relative alla stessa impresa contenute nell'indagine effettuata nell'anno precedente.

Il *microediting* interattivo si avvale anche della disponibilità dei valori medi per ateco relativi alle imprese rispondenti e validate che sono calcolati *ad hoc* dal programma di controllo e correzione.

In alcuni casi, inoltre, si ricorre anche alla imputazione logica sulla base dell'attività economica svolta (Ateco); ad esempio, se nel questionario è presente il dato sui ricavi (voce 11100) ma non è indicato quale elemento /i dei ricavi contribuisce a tale valore si attribuisce l'importo dei rivavi :

- Alle "Vendite di prodotto fabbricati dall'impresa" (voce 11101) se trattasi di un'impresa industriale;
- alle "Attività di intermediazione (commissioni, provvigioni, ecc.)" (voce 11105) se trattasi di impresa commerciale ;
- agli "Introiti lordi del traffico" (voce11106) per le imprese di trasporti **;**
- ai ricavi per "prestazione di servizi a terzi" (voce 11107) per i restanti tipi di imprese.

I questionari delle imprese che hanno al proprio interno numerose variabili mancanti, non ricostruibili da altre fonti, vengono ricontattate telefonicamente ed invitate a fornire le informazioni necessarie. Tale pratica, tuttavia, è utilizzata solo per quelle imprese che hanno un peso rilevante dal punto di vista dimensionale.

2.3 Individuazione dei valori anomali residui

Dopo i controlli di consistenza, la base dei microdati corretti relativi alle imprese rispondenti viene sottoposta ad ulteriori controlli automatici attraverso il confronto con fonti esterne.

In particolare il *benchmark* per le variabili economiche è rappresentato dai bilanci civilistici, mentre per gli addetti è costituito da ASIA.

L'obiettivo perseguito è quello della coerenza e della confrontabilità dei dati prodotti unitamente alla accuratezza dell'intero processo di indagine.

La corrispondenza dei dati provenienti dall'archivio dei bilanci e i dati rilevati dall'indagine viene condotto con riferimento a variabili importanti ed ampiamente confrontabili dal punto di vista definitorio come ad esempio: fatturato, costi, spese per il personale, immobilizzazioni. L'analisi statistica si basa sui quantili della distribuzione delle differenze percentuali calcolate a livello micro per ogni impresa presente nelle due basi di dati come differenza tra il dato amministrativo e quello rilevato dall'indagine.

L'indicatore di coerenza utilizzato è del tipo :

$$C = \frac{Y_b - Y_i}{Y_i} * 100$$

dove Y esprime la variabile confrontata (ad es. fatturato), i il dato di indagine mentre b il corrispondente presente nel *database* dei bilanci.

Generalmente si riscontra una forte convergenza informativa tra le due fonti come risulta dai prospetti seguenti; le imprese che invece presentano differenze significative nei valori vengono controllate nuovamente ed analizzate (*editing selettivo*) accertando il valore vero.

Tavola 1 - *Analisi della distribuzione delle differenze percentuali del fatturato tra la fonte bilanci e la fonte Sci (rilevato) per impresa - Anno 2003*

Numero imprese	3.722
Quinto percentile	-0,4
Venticinquesimo percentile	0,0
Mediana	0,0
Settantacinquesimo percentile	0,0
Novantacinquesimo percentile	0,04

Tavola 2 - *Analisi della distribuzione delle differenze percentuali del costo del lavoro tra la fonte bilanci e la fonte Sci (rilevato) per impresa - Anno 2003*

Numero imprese	3.722
Quinto percentile	-0,5
Venticinquesimo percentile	0,0
Mediana	0,0
Settantacinquesimo percentile	0,0
Novantacinquesimo percentile	0,5

Inoltre i dati per impresa dell'indagine dell'anno t sono confrontati con quelli della rilevazione relativi all'anno t-1. Le imprese che presentano forti variazioni dei valori non dovuti al trend economico o a trasformazione societarie vengono esaminati attentamente.

Sono altresì calcolati i valori medi e i rapporti caratteristici per impresa (ad esempio valore aggiunto per addetto, costo del personale per dipendente, ore per dipendente, valore aggiunto su fatturato) che consentono di individuare le unità con valori anomali che si collocano al di fuori di determinate soglie di accettazione o alle code della distribuzione.

La maggiore accuratezza a livello micro è propedeutica e necessaria alla integrazione delle mancate risposte totali che nella prima fase utilizza la tecnica del donatore (hot deck); un impresa con valori errati o anomali, se donatrice, potrebbe inficiare i valori finali a livello aggregato nell'ambito della classe di attività economica di appartenenza.

2.4 La procedura di integrazione delle mancate risposte totali (MRT)

Le mancate risposte totali nell'indagine SCI rappresentano circa il 50% della popolazione di riferimento; conseguentemente il loro trattamento è piuttosto rilevante allo scopo di prevenire distorsioni sulle stime finali.

La ricostruzione delle informazioni delle imprese non rispondenti all'indagine è effettuata sulla base di tre fonti:

- ✓ Archivio statistico sulle imprese (ASIA), da cui sono desumibili le caratteristiche strutturali (popolazione di riferimento, attività economica, addetti, localizzazione)
- ✓ L'insieme dei dati corretti relativi alle imprese rispondenti all'indagine
- ✓ La base dati dei bilanci civilistici.

In una prima fase la stima delle variabili economiche avviene attraverso una strategia di imputazione tramite donatore. Tale metodo, garantisce la correttezza finale del record ed il rispetto delle distribuzioni semplici e congiunte fra le variabili del questionario; essa consiste nell'individuare, per ogni impresa non rispondente, il donatore più vicino all'interno di insiemi di imprese caratterizzati da analoga dimensione aziendale, regione di appartenenza e settore di attività economica. Per evitare una riduzione della variabilità, insita nel metodo, si pone un vincolo al processo di estrazione del donatore ponendo un tetto al numero di volte che uno stesso *record* può essere utilizzato come donatore.

Nella seconda fase i dati stimati delle imprese non rispondenti sono sostituiti, con i valori provenienti dai bilanci civilistici.

Tramite la fonte bilanci sono disponibili 22 variabili relative al conto economico, 4 variabili relative allo stato patrimoniale e 6 relative al costo del lavoro; si tratta di 7 variabili di primo livello (capoconti che nel questionario sono indicati tramite lettera, più il costo del personale) e 25 variabili di secondo livello (capoconti individuati da lettera e cifra, più alcuni aggregati relativi al costo del lavoro). In particolare sono disponibili dal file dei bilanci civilistici le seguenti variabili:

Conto economico

- valore della produzione (A), e le 5 componenti (A1-A5);
- costi della produzione (B), e le 9 componenti (B6-B14);
- proventi e oneri finanziari (C), e le 3 componenti (C15-C17);
- imposte sugli utili lordi (T);
- utili netti (U).

Stato Patrimoniale

- immobilizzazioni (B), e le 3 componenti (B1-B3).

Costo del lavoro

- totale del costo del lavoro;
- salari e stipendi;
- oneri sociali;
- quiescenza;
- altri costi;
- trattamento di fine rapporto

Per lo Stato Patrimoniale, oltre alle variabili elencate sopra, sono disponibili altre 16 voci che rappresentano variabili di terzo livello e che vengono utilizzate (se presenti) nell'integrazione delle mancate risposte totali.

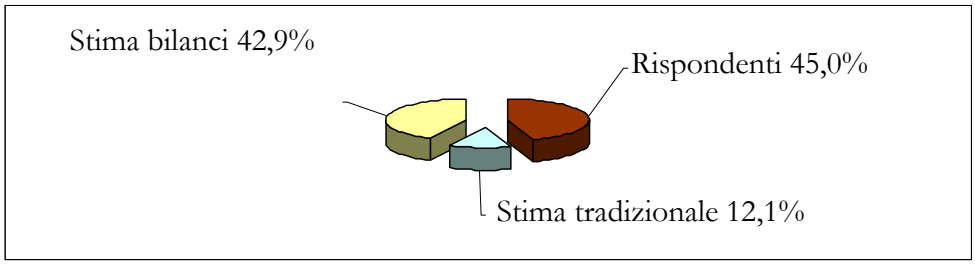
La procedura prevede innanzitutto l'integrazione diretta o tramite modello delle variabili di primo e secondo livello del questionario che dal punto di vista definitorio corrispondono a quelle contenute nella fonte dei bilanci. Quindi dal momento che nella rilevazione Sci le variabili richieste presentano un

livello di dettaglio superiore a quello riportato nel conto economico della fonte bilanci, le sottovoci vengono stimate utilizzando come pesi i valori ricavati dalla procedura di integrazione tramite donatore.

In tal modo, la fonte amministrativa consente di integrare 120 variabili (comprese quelle riponderate) della rilevazione con un notevole guadagno in termini di qualità per le stime finali. Per le restanti variabili si lasciano i valori ottenuti tramite l'usuale procedura di integrazione; infine tutte le variabili stimate vengono successivamente quadrate con una procedura gerarchica.

La figura 1 mostra l'analisi dei dati distinti per fonte relativamente al 2003.

Figura 1 – *Imprese per modalità di risposta e di integrazione – Anno2003*



Il confronto tra la procedura di integrazione delle mancate risposte totali tramite donatore e la procedura di integrazione tramite bilanci sulle stime finali, sintetizzato nella tavola 3, mostra differenze contenute soprattutto per le grandi imprese; tuttavia la prevalenza dei segni negativi indica che il metodo del donatore determina stime tendenzialmente superiori rispetto a quelle ottenute con il metodo dei bilanci. Tale risultanza è dovuta alla dimensione media delle imprese rispondenti che è generalmente più elevata rispetto a quella delle imprese non rispondenti.

Tavola 3 – *Confronto tra le procedure di integrazione delle mancate risposte totali tramite bilanci e tramite donatore nell'indagine SCI – Anno 2003.*

Classi di addetti	Totale imprese	Fatturato	Costo del lavoro	Valore della produzione	Costi della produzione
100 - 249	6.880	- 1,27	-2,48	-2,63	-0,65
250 - 499	1.861	- 2,42	-2,57	-4,15	-2,43
500 - 999	783	0,91	-1,33	0,32	1,16
1000 +	503	- 0,39	-0,43	-0,32	0,31
Totale	10.027	- 0.77	-1,43	-1,42	-0,29

2.5 Controlli a livello aggregato

Completata la fase di integrazione delle MRT, prima di effettuare le stime finali, i dati dell'indagine SCI vengono preventivamente controllati analizzando il loro comportamento in forma aggregata (*macroediting*). La disponibilità dei dati storici relativi a precedenti edizioni della rilevazione consente di effettuare confronti longitudinali fra i valori correnti delle stime e quelli che si riferiscono agli anni precedenti.

Le variabili sottoposte a verifica sono scelte per controllare gli aggregati rilevanti del conto economico, dell'occupazione, della spesa del personale e della spesa per investimenti.

Il ritorno selettivo sulle imprese da controllare viene effettuato da revisori esperti.

3. Conclusioni

La procedura di controllo e correzione dei dati dell'indagine sul sistema dei conti delle imprese si presenta quindi come un insieme integrato di diverse metodologie.

Le varie operazioni sono tra loro complementari ed interagiscono al fine di migliorare l'intero processo per produrre dati completi e coerenti ed un flusso informativo aderente alla reale situazione delle unità osservate.

L'innovazione fondamentale che ha permesso notevoli guadagni in termini di tempestività ed accuratezza delle stime finali è sicuramente rappresentato dall'utilizzo dei dati dei bilanci civilistici.

Essi rappresentano un valido *benchmark* sia per le procedure di controllo e correzione che per l'imputazione delle mancate risposte parziali e totali, consentendo apprezzabili vantaggi in termini di qualità dei dati e riduzione dell'onere statistico sulle imprese.

Figura 2 – *Schema della procedura di controllo e correzione dell'indagine " Rilevazione sul sistema dei conti delle imprese".*

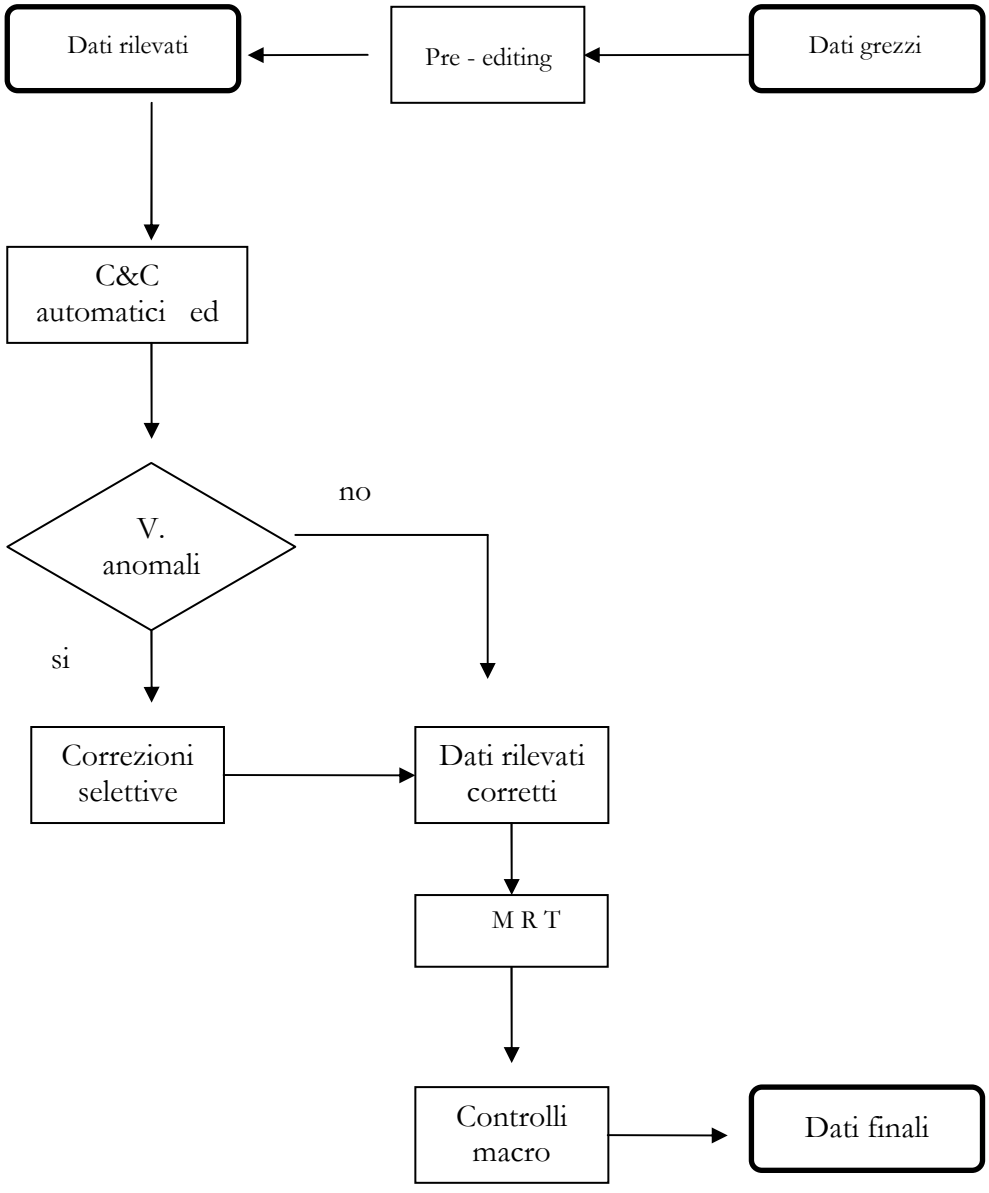


Tavola 4 – *Risultati della procedura di C & C per modalità di risposta e tipologia di controlllo nell'indagine SCI 2005 (Dati provvisori)* .

Valori assoluti

	Arrivi	Accertamenti	Correzioni	Forzature
Posta	1.424	1.271	6.750	6.964
Web	3.139	2.660	9.045	10.640

Valori medi

	Arrivi	Accertamenti	Correzioni	Forzature
Posta	1.424	0.89	4,74	4,89
Web	3.139	0.85	2,88	3,39

Bibliografia

Barcaroli G., D'Aurizio L., Luzi O., Manzari A.., Pallara A. *Metodi e software per il controllo e la correzione dei dati,* Quaderni di ricerca ISTAT n. 1/1999.

Dabbicco G., De Gregorio C. *L'utilizzo dei dati dei bilanci civilistici per l'integrazione delle mancate risposte totali alla rilevazione sul sistema dei conti delle imprese (SCI)* Risultati del gruppo di lavoro UDA*S* (2002)

ISTAT *Conti economici delle imprese- Anno 2000* Roma : ISTAT, 2005 (*Informazioni n.6)*

ISTAT *Conti economici delle imprese- Anno 2003* Roma : ISTAT, 2007 (*Informazioni n.8)*

ISTAT, CBS, SFSO. *Recommended practices for editing and imputation in cross – sectional business surveys* draft version 1.4 , ISTAT, CBS, SFSO, 2007.

Le nuove procedure di controllo e correzione delle indagini sull'agricoltura SPA e RICA-REA

Massimo Greco, *Istat, Direzione Centrale delle Statistiche Economiche Strutturali, SSI*
Ugo Guarnera, *Istat, Direzione Centrale per le Tecnologie e il Supporto Metodologico, MTS*
Orietta Luzi, *Istat, Direzione Centrale per le Tecnologie e il Supporto Metodologico, MTS*

Sommario: Nel presente lavoro sono descritte le nuove procedure di controllo e correzione dati realizzate per due indagini economiche del settore dell'agricoltura, l'indagine su Struttura e Produzione delle Aziende Agricole (SPA) e sulla Rete Contabile Agricola/Risultati Economici delle aziende agricole (RICA-REA). In entrambe le procedure sono stati implementati approcci avanzati per la risoluzione delle diverse tipologie di errore non campionario (incluse le mancate risposte parziali) per l'ottimizzazione dell'efficienza del trattamento dei dati in termini sia di tempi e costi delle elaborazioni, sia di riduzione dell'intervento interattivo e del peso sui rispondenti dovuto a follow-up, sia di maggiore accuratezza dei risultati con particolare riferimento al controllo degli errori influenti e alla stima delle mancate risposte.

Parole chiave: Errore non campionario, mancata risposta parziale, imputazione.

1. Introduzione

L'attuale sistema delle statistiche agricole in Italia è molto articolato e complesso in quanto risponde all'ampia richiesta informativa proveniente da utenti, pubblici e privati, sia nazionali che internazionali. In particolare la Politica Agricola Comune, per la cui gestione la Commissione richiede la disponibilità di dati agricoli tempestivi ed affidabili in maniera regolare da parte degli Stati Membri, influenza fortemente la struttura del sistema stesso.

Un'ulteriore peculiarità delle indagini di questo settore è la presenza di una pluralità di soggetti titolari dei progetti all'interno del SISTAN (ISTAT, MIPAAF, ISMEA. INEA, IREPA, Corpo Forestale dello Stato, ecc.). Per questo motivo negli ultimi anni l'ISTAT sta tentando di razionalizzare il sistema delle statistiche agricole, attraverso i circoli di qualità e vari protocolli d'intesa, con l'obiettivo di eliminare le duplicazioni esistenti, uniformare le differenti metodologie ed introdurre l'utilizzo di dati amministrativi disponibili per altre finalità.

All'interno dell'attuale complesso sistema delle statistiche agricole un ruolo rilevante occupano le rilevazioni sulla Struttura e Produzione delle Aziende agricole (SPA) e sulla Rete Contabile Agricola/Risultati Economici delle aziende agricole (RICA-REA) perché svolgono un ruolo di riferimento, se non in alcuni casi di *benchmark*, per le altre indagini del sistema.

La SPA (Ballin e Greco, 2006) è svolta ogni due anni con la finalità di aggiornare i dati raccolti dal Censimento agricolo rilevando, quindi, informazioni di tipo strutturale e principalmente quantitative sulle coltivazioni, gli allevamenti, il lavoro, i metodi di produzione e gli aspetti agro-ambientali (ISTAT, 2007a).

La RICA-REA, condotta annualmente, stima le principali variabili economiche aziendali secondo schemi concettuali analoghi a quelli adottati per l'analisi dei risultati economici delle imprese operanti nei settori dell'industria e dei servizi (ISTAT, 2007b). Questa seconda rilevazione è, in realtà, il frutto dell'integrazione di due distinte indagini condotte dall'INEA (RICA) e dall'ISTAT (REA) con modalità di raccolta dei dati diverse.

Per entrambe le rilevazioni, di tipo campionario, la popolazione di riferimento è rappresentata dalle aziende agricole rilevate al Censimento al di sopra di una determinata soglia fisica e/o economica che identifica l'universo UE e rende paragonabile i risultati a livello comunitario. Inoltre la raccolta dei dati avviene attraverso intervista diretta del conduttore di azienda da parte della rete di tecnici rilevatori delle Regioni e Province autonome.

La dimensione del campione è di circa 55.000 unità per la SPA e di 20.000 unità per la RICA-REA mentre il numero delle variabili presenti nel questionario è di circa 300 per la SPA e 200 per la RICA-REA.

Per ridurre gli errori dovuti alla mancata risposta (totale e parziale) da parte dei rispondenti, in fase di progettazione dell'indagine SPA vengono esaminate, assieme ai responsabili degli uffici di statistica e degli assessorati all'agricoltura delle Regioni e Province autonome coinvolti nella rilevazione, le problematiche che caratterizzano normalmente una rilevazione di questo tipo e le misure necessarie (specifica formazione dei rilevatori e sensibilizzazione nei riguardi dei conduttori di azienda agricola) per aumentare il grado di collaborazione dei rispondenti. Inoltre, prima dell'avvio della rilevazione i conduttori dell'azienda agricola vengono informati dell'indagine mediante l'invio di una lettera di preavviso.

Allo scopo di migliorare la qualità del dato raccolto, la collaborazione tra Regioni, Province autonome ed Istat inizia fin dalla fase di definizione dei contenuti, di scelte delle metodologie connesse alla rappresentatività dei risultati e si amplia in fase di sviluppo del questionario d'azienda e di stesura del libretto di istruzione.

Nel corso della rilevazione, a garanzia della qualità dei dati, è fornita una costante ed adeguata assistenza alla rete di rilevazione sia per la raccolta sia per la revisione dei dati, risolvendo molti casi di non corretta interpretazione dei quesiti e stimolando alla collaborazione anche i conduttori che presentano scarso interesse per l'indagine (fornendo chiarimenti sulla utilità dell'indagine e sul ruolo dell'azienda agricola per l'economia locale).

In fase di trattamento dei dati raccolti, per l'indagine SPA, si cerca di minimizzare le incongruenze formali, ovvero la presenza di codici di unità o di variabili non ammissibili, ricorrendo alla registrazione

mediante l'utilizzo del software BLAISE, sviluppato da Statistic Netherlands (CBS, 2005). Il software garantisce conformità tra le informazioni memorizzate sul questionario informatizzato e quello cartaceo rendendo nullo l'errore. Nella realizzazione del questionario informatizzato, prodotto all'interno dell'Istat, sono state introdotte alcune regole di compatibilità con lo scopo di evidenziare ed eliminare già in fase di registrazione dei dati particolari tipi di incongruenze.

Per l'indagine RICA-REA, a partire dal 2005 e solo per il campione REA gestito dall'ISTAT (circa 5.000 unità sulle 20.000 totali), è utilizzato un questionario elettronico in formato excel contenente dei controlli di compatibilità sulle variabili.

La registrazione dei dati (tramite software BLAISE per la SPA e questionario excel per la REA) avviene presso le Regioni e Province autonome a carico degli uffici preposti alla raccolta dei dati. Questa soluzione favorisce un tempestivo intervento per i casi di mancata risposta sia totale che parziale, consentendo a livello locale il ripristino dell'informazione mancante e/o alla rettifica di quella errata.

Nel corso delle ultime occasioni di indagine, le procedure di controllo e correzione di entrambe le rilevazioni sono state completamente riprogettate, al fine di rendere più efficiente (soprattutto in termini di tempi e risorse impiegate) il processo di individuazione e trattamento degli errori e delle mancate risposte parziali. Obiettivo di questo lavoro è descrivere le principali innovazioni metodologiche introdotte nella procedura di controllo e correzione delle due indagini (SPA e campione REA), con particolare riferimento al trattamento automatico degli errori. Nel paragrafo 2 è illustrata la struttura generale delle due procedure di controllo e correzione (C&C), disegnata tenendo conto non solo delle varie tipologie di errore presenti nei dati, ma anche delle caratteristiche dei fenomeni trattati e della rilevanza delle unità sottoposte a controllo. Nei paragrafi 3 e 4 sono approfonditi alcuni aspetti relativi alle metodologie adottate per il trattamento automatico degli errori e delle mancate risposte per le diverse (sotto)sezioni dei questionari adottati: in particolare, il paragrafo 4 si concentra sulle sezioni relative al Lavoro in azienda. Il paragrafo 5 contiene alcune considerazioni conclusive.

2. Le nuove procedure di controllo e correzione

In entrambe le indagini considerate in questo lavoro, la procedura di C&C delle due indagini è strutturata in più passi, integrati fra loro secondo una gerarchia predefinita al fine di razionalizzare e semplificare il trattamento delle diverse tipologie di errore nel consistente numero di variabili osservate: in ciascun passo, infatti, vengono trattati sottoinsiemi di variabili contenute in sezioni "connesse" del questionario. Tali sottoinsiemi sono individuati sulla base delle caratteristiche delle variabili stesse e delle relazioni fra esse esistenti.

Per ogni passo della procedura, cioè per ogni gruppo di sezioni, la localizzazione e il trattamento degli errori sono stati progettati secondo una strategia sostanzialmente comune, in cui il flusso delle operazioni è basato sulla **tipologia** e sulla **rilevanza** degli errori che si assume siano presenti nei dati. In tale strategia diverse metodologie e tecniche sono integrate gerarchicamente tenendo conto della diversa rilevanza delle varie tipologie di errore in termini di potenziale effetto distorsivo sulle stime. Coerentemente con quanto raccomandato nelle *Pratiche Raccomandate per il Controllo e la Correzione dei dati nelle Indagini Trasversali sulle Imprese* (Luzi *et al.*, 2007), la procedura di C&C nelle due indagini prevede la seguente gerarchia: errori sistematici, valori anomali ed errori influenti, errori casuali non influenti. Le mancate risposte parziali, ricostruite mediante diversi modelli di imputazione, sono trattate al termine della fase di controllo degli errori, al fine sia di evitare la propagazione degli errori stessi (ad esempio in caso di imputazione con donatore di distanza minima), sia di ottenere previsioni non distorte dei dati mancanti (ad esempio a causa dei possibili effetti distorsivi sui parametri dei modelli di imputazione dovuti alla presenza di valori anomali).

I vari passi della procedura di C&C si articolano a loro volta nelle macro-fasi descritte di seguito (Guarnera e Luzi, 2004; Ballin *et al.*, 2004):

1. **Trattamento degli errori di tipo sistematico** mediante regole di compatibilità di tipo deterministico.

2. **Controllo dei valori anomali e degli errori influenti**. Per l'individuazione degli errori influenti viene adottato l'approccio dell'editing selettivo (Latouche *et al.*, 1992), mentre l'individuazione dei valori anomali avviene principalmente attraverso l'analisi di grafici e matrici di transizione in grado di evidenziare le variazioni (rispetto a quanto osservato con il censimento generale dell'agricoltura) delle principali caratteristiche strutturali e produttive di ciascuna unità e la loro influenza sul livello delle stime finali. Per maggiori dettagli sulle metodologie adottate per queste tipologie di errore vedere Guarnera e Luzi, 2005 per la SPA; e Guarnera *et al.*, 2006b per la REA.
3. **Individuazione ed eliminazione degli errori casuali non influenti ed integrazione delle mancate risposte parziali** (escluse le sezioni sulla *Manodopera Aziendale*). In questa fase, l'individuazione di errori è effettuata mediante la metodologia probabilistica nota come *algoritmo di Fellegi e Holt* (Fellegi e Holt, 1976). Per la ricostruzione delle informazioni mancanti o incoerenti vengono adottate due tecniche di imputazione: con soluzione analitica, ed una versione della tecnica del donatore di distanza minima (*Nearest-Neighbour Donor*, *NND*) in cui si tiene conto dei vincoli di coerenza fra le variabili (paragrafo 3.1).
4. **Individuazione ed eliminazione degli errori casuali non influenti ed integrazione delle mancate risposte parziali per le sezioni relative al Lavoro**. In entrambe le indagini, gli errori e le mancate risposte parziali sono individuati mediante regole di compatibilità di tipo deterministico. Il loro trattamento (imputazione) è stato effettuato utilizzando la tecnica NND non vincolato e, per l'indagine REA, anche la tecnica di imputazione con regressione multivariata mediante algoritmo *EM* (Dempster *et al.*, 1977) (paragrafo 4.1).
5. **Trattamento manuale-interattivo degli errori residui** delle precedenti fasi, e validazione finale dei dati.

Le fasi 1, 2 e 5 sono state sviluppate all'interno del Servizio Agricoltura in ambiente AGAIN (modulo SAS), per l'indagine SPA, e con procedure SAS, per la rilevazione REA.
La fase 3, invece, è stata implementata, per entrambe le indagini, attraverso l'uso del software generalizzato *Banff* (paragrafo 3).

L'imputazione delle mancate risposte nella fase 4 è stata effettuata utilizzando il software *QUIS* (paragrafo 4).

Al fine di illustrare schematicamente il flusso appena descritto, nella figura 1 si riporta come esempio la strategia generale adottata per l'indagine REA (per maggiori dettagli, vedere Guarnera *et al.*, 2006b). In questo caso, la procedura consiste di tre passi principali (variabili rilevate nella Sezione sulla **Struttura economica**, variabili rilevate alle sottosezioni **Manodopera Familiare** e **Altra Manodopera** della sezione **Lavoro**). Per ogni passo, gli errori sono trattati gerarchicamente come indicato in figura.

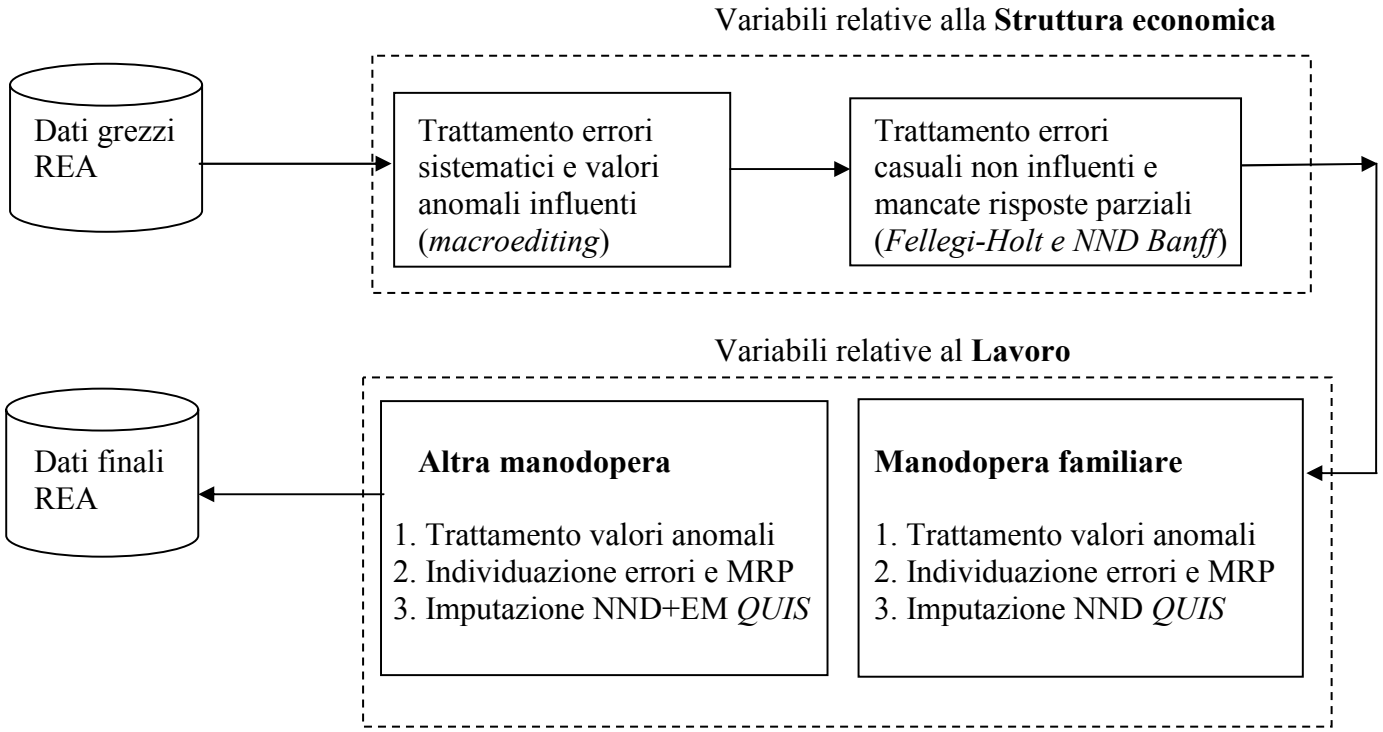
Per quanto riguarda il problema dell'individuazione dei valori anomali (e degli errori influenti) nelle due sottosezioni della sezione **Lavoro**, sono in corso sperimentazioni per ottimizzare la procedura REA utilizzando tecniche alternative di tipo selettivo combinate con metodi multivariati di individuazione dei valori anomali (Di Zio *et al.*, 2007).

3. Il trattamento degli errori e delle mancate risposte mediante Banff

In questa fase vengono trattate tutte le variabili quantitative del questionario che sono interconnesse da vincoli logici e aritmetici, quali quadrature (i valori di voci parziali di una grandezza devono sommare al valore totale della stessa grandezza), disuguaglianze (relazioni di inclusione tra quantità), esistenze (la positività di una o più variabili implica la positività di altre variabili). Per l'indagine SPA, si tratta per lo più delle variabili relative a superfici coltivate per tipo di coltivazione, produzione corrispondente, capi di allevamento, ecc. Per l'indagine REA le principali variabili trattate in questa fase si riferiscono alla struttura dei costi e dei ricavi (*Giacenze e scorte, Reimpieghi, Ricavi e autoconsumo*, ecc.).

Per queste variabili, la localizzazione e la correzione degli errori, e l'imputazione delle mancate risposte parziali è stata effettuata utilizzando le metodologie disponibili nel software generalizzato *Banff*, che rappresenta la versione in ambiente SAS del software generalizzato GEIS (Kovar *et al.*, 1988).

Figura 1: *Strategia complessiva di controllo e correzione dell'indagine REA*



3.1. Metodologie per il controllo e la correzione dei dati disponibili in Banff

Il controllo e la correzione automatica di variabili numeriche continue avviene nel software *Banff* sulla base della metodologia Fellegi-Holt (FH nel seguito) e della tecnica hot deck nota come donatore di distanza minima (NND). Questi metodi sono particolarmente adatti al trattamento di errori originati da meccanismi di tipo completamente casuale che siano non influenti in termini di impatto sulle stime dei parametri obiettivo.

L'algoritmo probabilistico FH può essere utilizzato per identificare errori casuali in dati che devono risultare coerenti, a livello micro, rispetto a prefissati vincoli o relazioni di compatibilità (*edit*) fra variabili osservate. Per ogni unità *i* che viola almeno un edit, l'algoritmo identifica il minimo numero di valori (variabili) da modificare in modo da poter riportare il record alla situazione di coerenza (soluzione di minimo cambiamento). In altri termini, l'algoritmo assegna ad ogni variabile che compare in almeno un edit una probabilità di essere errata proporzionale al numero di edit violati che coinvolgono tale variabile. La selezione delle variabili classificate come errate, e quindi da correggere, può essere influenzata dallo statistico attraverso l'uso di pesi, che possono essere associati alle variabili e che ne misurano il grado di affidabilità (tanto maggiore è il peso di una variabile, tanto minore è la sua probabilità di essere inclusa nella soluzione di minimo cambiamento). Nel caso di utilizzo dei pesi, la soluzione di minimo cambiamento per un record errato è quella che coinvolge il sottoinsieme di variabili aventi la minima somma dei pesi.

Una volta localizzati, gli errori ed i valori originariamente mancanti (mancate risposte parziali) devono essere sostituiti (imputati) con valori ammissibili. A questo fine, diverse tecniche sono disponibili in *Banff*: un metodo di imputazione con soluzione analitica (o deduttiva) basato sulla ricerca dell'unico valore (se esiste) che sostituito ai valori errati soddisfa tutti i vincoli di compatibilità; una

tecnica di imputazione con NND sotto vincoli; diverse tecniche da modello esplicito (inclusi modelli di regressione).

Per quanto riguarda la tecnica del NND, per ogni unità *i* che viola almeno un edit (*unità ricevente*), i valori mancanti o classificati come errati per una o più variabili sono sostituiti congiuntamente con i valori delle stesse variabili osservati nell'unità più vicina di (*donatore*). Il donatore di è selezionato da un serbatoio di unità complete (cioè prive di valori mancanti) e coerenti (cioè che soddisfano tutti gli edit) in base alla funzione di distanza *minmax* calcolata rispetto a un insieme di variabili di accoppiamento (o *matching*) trasformate mediante la funzione rango. Una importante caratteristica del metodo adottato consiste nel fatto che l'imputazione avviene nel rispetto dei vincoli: per ogni unità ricevente, il donatore di viene utilizzato per l'imputazione solo se l'unità risultante diviene coerente rispetto a tutti gli edit. Nella pratica, nel metodo NND disponibile in *Banff* il donatore viene utilizzato effettivamente solo se l'unità ricevente, una volta imputata, risulta coerente rispetto a tutti gli edit di post-imputazione: questi edit corrispondono a vincoli iniziali opportunamente resi meno stringenti (ad esempio, vincoli di quadratura rilassati in modo da ammettere come validi valori in un intorno del valore esatto di confronto - il totale). L'uso degli edit di post-imputazione ha l'obiettivo di allargare la rosa dei potenziali donatori per ogni unità errata. Infatti, dato un record *i* che fallisce un certo vincolo di uguaglianza, può accadere che il sistema non riesca a trovare un donatore che fornisca ad *i* un valore tale da riportarlo nella condizione di soddisfare l'uguaglianza, oppure che scarti record donatori "vicini" ad *i*, ma che non gli garantiscono il rispetto dell'uguaglianza, selezionando un donatore più "distante" da *i*, ma che gli fornisce il valore richiesto.

E' evidente che se da un lato l'uso degli edit di post-imputazione facilita l'individuazione di un donatore appropriato, dall'altro esso rende necessaria una successiva verifica dei dati imputati per verificare se qualcuno di essi viola i vincoli di uguaglianza iniziali.

In generale, le tecniche di imputazione (incluso il NND) risultano più efficienti se vengono applicate all'interno di *celle di imputazione*: si tratta di sotto-popolazioni definite sulla base di covariate statisticamente associate alle variabili oggetto di correzione/imputazione, e quindi ritenute maggiormente omogenee in termini del/dei fenomeno/i oggetto di imputazione. Tutte le tecniche di imputazione disponibili in *Banff* prevedono la possibilità di effettuare la ricostruzione dei casi errati o mancanti all'interno di celle definite sulla base di variabili definite dall'esperto.

3.2. Strategia di localizzazione e correzione automatica nelle indagini SPA e REA

In questa fase vengono risolte in modo automatico (quindi poco costoso in termini di tempo e risorse impiegate) tutte le incoerenze logico-matematiche presenti nei dati, ma che sono considerate poco influenti ai fini della stima avendo superato la fase precedente di individuazione di valori anomali ed errori influenti.

Dal momento che gli errori sono individuati sulla base di un prefissato insieme di edit, quando si adotta la metodologia FH è necessaria la massima cautela nella definizione delle regole di controllo. Infatti, se viene utilizzato un insieme con regole poco numerose ma soprattutto poco connesse (in termini di variabili coinvolte in esse), l'algoritmo potrebbe determinare una scelta sostanzialmente casuale delle variabili da modificare. In tali casi, un approccio di tipo deterministico potrebbe risultare preferibile. D'altra parte, troppi vincoli di coerenza potrebbero dare origine ad eccessiva complessità del problema di localizzazione dell'errore, con conseguente impossibilità per l'algoritmo a determinare una soluzione. Per questo motivo, in alcune applicazioni particolarmente complesse si ricorre alla suddivisione degli edit in due o più sottoinsiemi, ed alla loro applicazione sequenziale ai dati in una prefissata gerarchia. E' stato appunto questa la strategia utilizzata per l'indagine SPA. In questa indagine infatti, la procedura di controllo e correzione automatica si è articolata in tre fasi. Nella prima sono state trattate le variabili principali relative alle *coltivazioni*. Nella seconda fase sono state corrette altre variabili relative alle coltivazioni, mantenendo fissi i valori del primo gruppo di variabili. Nella terza fase sono state trattate le variabili relative agli *allevamenti* e alle relative produzioni. Per maggiori dettagli sulla struttura della procedura di controllo degli errori mediante *Banff* nell'indagine SPA, vedere Guarnera *et al.*, 2006b.

Nell'indagine REA le variabili relative alla **Struttura Economica** sono state trattate tutte simultaneamente. Per maggiori dettagli sulla struttura della procedura di controllo degli errori nell'indagine REA, vedere Guarnera *et al.*, 2006b.

In entrambe le indagini, per ogni fase o gruppo di variabili, la procedura di individuazione e correzione/imputazione di errori e mancate risposte è ciclica, e prevede la ripetizione per un numero prefissato di volte dei seguenti passi (vedere l'Appendice per lo schema relativo ad una delle fasi dell'indagine SPA, estratto da Guarnera *et al.*, 2006):

1. *Localizzazione errori con edit iniziali*
   - individuazione degli errori e delle mancate risposte con un set iniziale di parametri (numero massimo di valori classificabili come errati, tempo massimo di elaborazione per record)
   - ri-elaborazione dei casi non risolti con nuovi parametri
2. *Imputazione degli errori e delle mancate risposte*
   - imputazione con soluzione analitica
   - sui casi residui, imputazione con metodo NND con edit di post-imputazione
3. *Nuovo passo di localizzazione errori su record imputati con edit iniziali*
4. *Imputazione degli errori con soluzione analitica*

Gli errori residui ai vari passi (in particolare, 1, 2 e 4) vengono sottoposti a revisione manuale interattiva: si tratta comunque di un numero limitato di osservazioni per entrambe le indagini, per cui tale fase non comporta un significativo incremento di tempi e risorse impiegate.

La procedura automatica può quindi essere considerata rispondente all'esigenza di ottimizzazione di tempi e costi del trattamento degli errori casuali non influenti. D'altra parte, relativamente all'impatto delle correzioni effettuate automaticamente sulle variabili obiettivo, in entrambe le indagini si sono osservati effetti contenuti. Ad esempio, nella Tabella 1 sono mostrati gli effetti su valor medio e deviazione standard delle principali variabili rilevate nelle due indagini (per maggiori dettagli, vedi Guarnera e Luzi, 2005 per la SPA; Guarnera *et al.*, 2006a per la REA).

Tabella 1: *Effetto della procedura automatica di C&C per indagine e per variabile*

		Indice	
Indagine	**Variabile**	**VRM**	**VRSTD**
RICA-REA	Tot spese coltivazioni	0,00	0,00
	Tot spese allevamenti	19,26	18,43
	Tot spese meccanizzazione	-0,75	-1,01
	Tot spese generali e varie	0,00	0,00
	Tot costi	-0,71	-4,49
	Tot ricavi	0,00	0,00
SPA[7]	Superficie Aziendale Totale (SAT)	0,71	1,17
	Superficie Aziendale Utilizzata (SAU)	0,21	3,69
	Totale Seminativi	0,32	0,18
	Totale Coltivazioni Legnose	0,07	0,63
	Totale Superficie Irrigata	4,90	12,05

[7] In questo caso, i valori degli indicatori VRM e VRSTD sono valori medi calcolati su 50 applicazioni iterative della procedura automatica di C&C (vedi Guarnera e Luzi, 2005).

Gli indicatori sono definiti come segue: siano X_i i valori *grezzi* e con X_i^* i valori *finali* (controllati e corretti) che una variabile X assume sulle unità i (i=1,...,n). Le misure VRM e VRSTD sono definite come segue:

- *Variazione Relativa della Media:* $VRM = \frac{\overline{X}_w^* - \overline{X}_w}{\overline{X}_w} \times 100$, dove $\overline{X}_w = \frac{1}{\sum_{1=1}^{n1} w_i} \sum_{i=1}^{n1} w_i X_i$, e

  $\overline{X}_w^* = \frac{1}{\sum_{i=1}^{n2} w_i} \sum_{i=1}^{n2} w_i X_i^*$ sono le medie ponderate ottenute rispettivamente dai valori grezzi e *puliti* di X calcolati sui corrispondenti sottoinsiemi di n_1 e n_2 unità ($n_1 \leq n$, $n_2 \leq n$) tali che $X_i \neq 0$ e $X_i^* \neq 0$. Questo indice fornisce una indicazione dell'effetto dell'individuazione e trattamento degli errori e delle mancate risposte sul valor medio della variabile X.

- *Variazione Relativa della Deviazione Standard*: $VRSTD = \frac{STD_w - STD_w^*}{STD_w} \times 100$, dove STD_w e STD^*_w sono rispettivamente le deviazioni standard ponderate sui valori grezzi e puliti di X calcolati sui corrispondenti sottoinsiemi di n_1 e n_2 unità ($n_1 \leq n$, $n_2 \leq n$) tali che $X_i \neq 0$ e $X_i^* \neq 0$. L'indice fornisce una indicazione dell'effetto dell'individuazione e trattamento degli errori e delle mancate risposte sulla variabilità di X.

4. Il trattamento delle variabili della Sezione Lavoro mediante QUIS

In questa fase vengono trattate tutte le variabili osservate nelle sezioni dei questionari di indagine che rilevano informazioni sulla manodopera (familiare e non) impiegata nelle aziende Agricole, e i relativi costi.

In effetti, la sezione relativa al **Lavoro** (occupazione e redditi) presenta caratteristiche specifiche che richiedono un diverso trattamento delle variabili in essa riportate. In particolare, la sostanziale assenza di vincoli tra le variabili rilevate rende inutilizzabile la metodologia FH per la localizzazione degli errori. Per questo motivo, in entrambe le indagini la localizzazione degli errori è stata effettuata utilizzando un approccio di tipo deterministico, rilevando sostanzialmente i valori "fuori range" (ad es. numero di giornate lavorate pro-capite annuale maggiore di 365), e determinando i casi in cui non sono accettabili valori nulli delle variabili analizzate (ad esempio, in presenza di retribuzioni devono esistere giornate lavorate).

Per quanto riguarda la correzione degli errori e la ricostruzione delle mancate risposte parziali, invece, i diversi tipi di relazioni esistenti fra sottogruppi di variabili osservate in tale sezione hanno reso necessario differenziare i modelli di imputazione utilizzati. Tutte le procedure di imputazione usate per tale sezione si basano su routine contenute nel software *QUIS*.

4.1. Metodologie per l'imputazione dei dati disponibili in QUIS

Il software QUIS, interamente sviluppato in ambiente SAS, è uno strumento di semplice utilizzo per l'applicazione di alcune delle principali tecniche di trattamento dei valori mancanti nel caso di dati quantitativi. Nel software sono stati inclusi i seguenti metodi:

1) imputazione per regressione basata sull'algoritmo EM
2) donatore di minima distanza
3) predictive mean matching multivariato
4) imputazione multipla

Tutti i metodi implementati sono applicabili in presenza di pattern arbitrari di mancata risposta e presuppongono che i valori errati siano stati preliminarmente identificati ed eliminati dal dataset di analisi. Tra i diversi metodi, l'imputazione basata su EM (Dempster *et al.*, 1977) e l'imputazione multipla si basano fortemente sull'ipotesi di normalità dei dati e possono essere definiti parametrici. Più precisamente, il primo metodo utilizza l'algoritmo EM per stimare i parametri di un modello normale multivariato da cui vengono derivate le distribuzioni condizionate corrispondenti ai vari pattern di mancata risposta utilizzate per l'imputazione. L'imputazione può essere effettuata con o senza l'aggiunta di un residuo casuale (normale) a seconda che si sia interessati o meno alla preservazione delle caratteristiche distribuzionali dei dati. L'imputazione multipla (Rubin, 1987) è una tecnica che consente di stimare la componente della variabilità delle stime dovuta all'incompletezza dell'informazione, mediante un'opportuna replicazione del processo di imputazione con la conseguente produzione di un certo numero di dataset completi. Nella versione implementata in QUIS si assume un modello normale multivariato per i dati.

Ancora basato sull'assunzione di normalità, ma probabilmente più robusto rispetto ad eventuali allontanamenti dall'ipotesi, è il Predictive Mean Matching (Little,1988). Anche questo metodo stima un modello normale multivariato mediante algoritmo EM, ma utilizza il modello solo per definire un'opportuna funzione di distanza utilizzata per associare ad un dato record incompleto un opportuno "donatore".

Infine in QUIS è incluso un classico metodo di imputazione basato su donatore di minima distanza, che consente di selezionare il donatore più vicino utilizzando una delle tre possibili funzioni di distanza tra Euclidea, Minimax e Manhattan. A differenza del donatore di minima distanza implementato in Banff tuttavia, il metodo non consente di tener conto di eventuali vincoli di compatibilità tra le variabili.

4.2. Strategia di localizzazione e correzione automatica nelle indagini SPA e REA

In questo paragrafo vengono descritte più in dettaglio le strategie di imputazione adottate nelle indagini REA e SPA per la ricostruzione degli errori e delle mancate risposte parziali per le sezioni relative alla manodopera impiegata nelle aziende Agricole, e ai relativi costi.

Nel caso dell'indagine REA (vedere Guarnera *et al.*, 2006b), l'imputazione degli errori e delle mancate risposte parziali è stata effettuata con procedure diverse a seconda che le variabili trattate si riferissero: 1) alla *manodopera familiare*, 2) all'*altra manodopera aziendale a tempo determinato*, 3) all'*altra manodopera a tempo indeterminato*. Nel primo caso i record sono a livello di individuo, e l'imputazione delle variabili di interesse (*giornate lavorate*, *ore medie lavorate giornaliere*) è stata effettuata con il metodo del donatore di minima distanza all'interno di opportune celle di imputazione definite sulla base della relazione di parentela con il conduttore e di alcune variabili strutturali dell'azienda (UDE, *altimetria*, ecc.). Il donatore di minima distanza è stato utilizzato anche nel caso della manodopera aziendale extra familiare. In questo caso però, ogni unità si identifica con un'azienda anziché con un individuo e le variabili analizzate (*giornate lavorate, retribuzioni, contributi, TFR*) si riferiscono alla totalità degli individui che lavorano nell'azienda. Nel caso della manodopera extra-familiare a tempo indeterminato infine, un'analisi esplorativa dei dati ha evidenziato relazioni approssimativamente lineari, in scala logaritmica, tra variabili di interesse (le stesse del caso precedente). Ciò ha suggerito di adottare un approccio basato su modello. Più precisamente, i dati sono stati modellati, in scala logaritmica, mediante una distribuzione normale quadrivariata. I parametri del modello sono stati stimati mediante algoritmo EM e utilizzati per stimare le distribuzioni condizionate corrispondenti ai diversi pattern di valori mancanti. Per ogni pattern infine, l'imputazione è stata effettuata generando da queste distribuzioni, anziché calcolando le medie condizionate, al fine di preservare i momenti superiori al primo delle distribuzioni marginali e congiunte.

Nel caso dell'indagine SPA, la fase di imputazione delle variabili osservate nelle sezioni relative alla manodopera aziendale è stata effettuata utilizzando la tecnica NND non vincolato disponibile in

QUIS, con parametri diversi a seconda che le variabili trattate si riferissero a: 1) *Famiglia e parenti del conduttore,* 2) *Altri lavoratori dell'azienda in forma continuativa* (a tempo determinato o indeterminato), 3) *Altri lavoratori dell'azienda in forma saltuaria*. Per le tre le tipologie di variabili l'imputazione è avvenuta all'interno di celle di imputazione definite utilizzando diversi sottoinsiemi di covariate, e specificando diversi insiemi di variabili di *matching* per il calcolo della distanza.

Per le variabili di cui ai punti 1 e 2, in cui ogni osservazione corrisponde a un singolo individuo, l'imputazione delle variabili di interesse (*numero di giornate lavorate*, *ore medie lavorate giornaliere*) è stata effettuata in celle definite sulla base del *tipo di lavoratore* (tipo di parentela col conduttore e tipo di contratto di lavoro), della *forma giuridica* dell'azienda e della *regione* di appartenenza. Come variabili di *matching* sono state utilizzate l'*altimetria,* la SAT (*Superficie Agricola Totale*) e la classe UBA (*Unità Bestiame Adulto*) dell'azienda.

Nel caso degli *Altri lavoratori in forma saltuaria*, in cui ogni record corrisponde al totale di lavoratori di questo tipo presenti in azienda (separatamente per maschi e femmine), l'imputazione delle variabili di interesse (*totale persone impiegate* e *totale giornate lavorate*), è stata effettuata in celle di imputazione definite sulla base delle variabili *sesso*, *forma giuridica* e *regione*. In questo caso, come variabili di *matching* sono utilizzate, oltre ad *altimetria, SAT* e *UBA*, anche le stesse variabili *Totale persone impiegate* e *Totale giornate lavorate*, al fine di garantire la selezione di un'azienda donatrice simile rispetto a tali caratteristiche. In questo caso, per ottenere valori imputati più stabili rispetto alla relazione fra il numero complessivo di persone impiegate e il numero totale di giornate lavorate in azienda, la ricostruzione di valori errati o mancanti delle variabili di interesse è stata effettuata passando per il valore *pro-capite* di giornate lavorate in azienda, definito come:

Tot Giorni Pro = Totale giornate lavorate/Totale persone impiegate

Per ogni azienda con valore da imputare di una delle due variabili obiettivo, i valori finali (indicati con asterisco) sono stati quindi ottenuti, dopo il passo di imputazione del valore *pro-capite* mediante il NND di QUIS, attraverso le semplici relazioni:

$Totale\ giornate\ lavorate^{*} = Tot\ Giorni\ Pro \times Totale\ persone\ impiegate$)
$Totale\ persone\ impiegate^{*} = Totale\ giornate\ lavorate\ /\ Tot\ Giorni\ Pro$).

5. Conclusioni

La strutturazione di una procedura di controllo in fasi il più specializzate possibile rispetto alle tipologie di errori non campionari presenti nei dati di indagine consente sia una maggior razionalizzazione e un maggior controllo del flusso dei dati, sia una maggiore possibilità di aggiornare/monitorare il processo di controllo e correzione.

In questo contesto, l'adozione in ogni fase delle tecniche più appropriate per la tipologia di errore e di variabile oggetto di controllo e correzione in quella fase consente un trattamento ottimale sia dal punto di vista della qualità attesa del risultato, sia dal punto di vista dei tempi e dei costi.

Le procedure di controllo e correzione realizzate per le indagini SPA e REA costituiscono un esempio di procedura integrata per indagini economiche. In esse, diversi approcci sono utilizzati a seconda non solo della tipologia di errore (errori influenti, valori anomali, errori di compatibilità casuali, mancate risposte parziali), ma anche della tipologia di fenomeni (variabili) oggetto del controllo/imputazione e delle relazioni esistenti fra essi.

Specifici vantaggi dell'utilizzo di software generalizzato sono la riduzione dei costi di sviluppo di nuovo codice, lo sfruttamento di algoritmi complessi altrimenti non utilizzabili, la facilità di documentazione del processo (grazie alla reportistica prodotta automaticamente dal software, e la semplicità di aggiornamento del processo di controllo e correzione dovuta ad eventuali cambiamenti strutturali dell'indagine (stratificazioni, aggiunta/eliminazione di variabili, modifica della struttura interna del questionario, ecc.).

Bibliografia

Ballin M. e Greco M. *Indagine sulla struttura e produzione delle aziende agricole. Anno 2005. Rapporto sulla qualità,* Direttiva DCSS1, Produzione di rapporti di qualità delle rilevazioni su imprese, aziende agricole, istituzioni pubbliche e private, Roma, Dicembre 2006.

Ballin M., Guarnera U., Luzi O. e Salvi S. "Nuove metodologie e strumenti per il trattamento degli errori non campionari nell'indagine su Struttura e Produzione delle Aziende Agricole2. *Atti del Convegno Metodi d'Indagine e di Analisi per le Politiche Agricole - MIAPA 2004*, Università di Pisa, 21-22 Ottobre 2004.

CBS. *Blaise for Windows 4.5 Developer's Guide.* http://www.cbs.nl/en-GB/menu/informatie/onderzoekers/blaise-software/blaise-voor-windows/manuals/default.htm 2005.

Dempster, A.P., Laird, N.M. and Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38, 1977.

Di Zio M., Guarnera U., Luzi O. e Tommasi I. "Detection of potentially Influential Errors in Statistical Survey Data. *Convegno intermedio SIS 2007*. Venezia, 6-8 Giugno 2007.

Fellegi, I.P. e Holt, T.D. "A Systematic Approach to Edit and Imputation". *Journal of the American Statistical Association*, 71, 17-35, 1976.

Guarnera U. "Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi. Il software QUIS". *Contributi di Ricerca ISTAT*, N. 5/2004.

Guarnera U., Luzi O. "Editing and Imputation Methods in the ISTAT Survey on Structure and Production of Agricultural Firms". *Atti del Convegno "L'Informazione Statistica e le Politiche Agricole" - ISPA 2004*, Università di Cassino, 6 Maggio 2004.

Guarnera U. e Luzi O. "Valutazione del trattamento degli errori di misura e di risposta nell'indagine SPA". *Atti del Convegno AGRI@Istat - Verso un nuovo sistema di Statistiche Agricole*, Firenze, 30-31 Maggio 2005.

Guarnera U., Luzi O. e Salvi S. "La nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti". *Documenti Istat*, N. 8/2006.

Guarnera U., Luzi O. e Tommasi I. "Metodi Parametrici e non parametrici per la ricostruzione dei valori mancanti nell'indagine RICA-REA". *Atti del Convegno "Le statistiche agricole verso il Censimento del 2010: valutazioni e prospettive"*, Università di Cassino, 26-27 Ottobre, 2007a.

Guarnera U., Luzi O. e Tommasi I. "La nuova procedura di controllo e correzione degli errori e delle mancate risposte parziali nell'Indagine sui Risultati Economici delle Aziende Agricole". *Documenti Istat*, N. 3/2007, 2007b.

Kovar, J.G., MacMillan, J. e Whitridge, P. *Overview and Strategy for the Generalized Edit and Imputation System*, Statistics Canada, Methodology Branch Working Paper No, BSMD-88-007E/F, Ottawa, 1988.

ISTAT *Indagine sulla struttura e sulle produzioni delle aziende agricole*, http://www.istat.it/strumenti/rispondenti/indagini/spa2007/spa2007.html, 2007a.

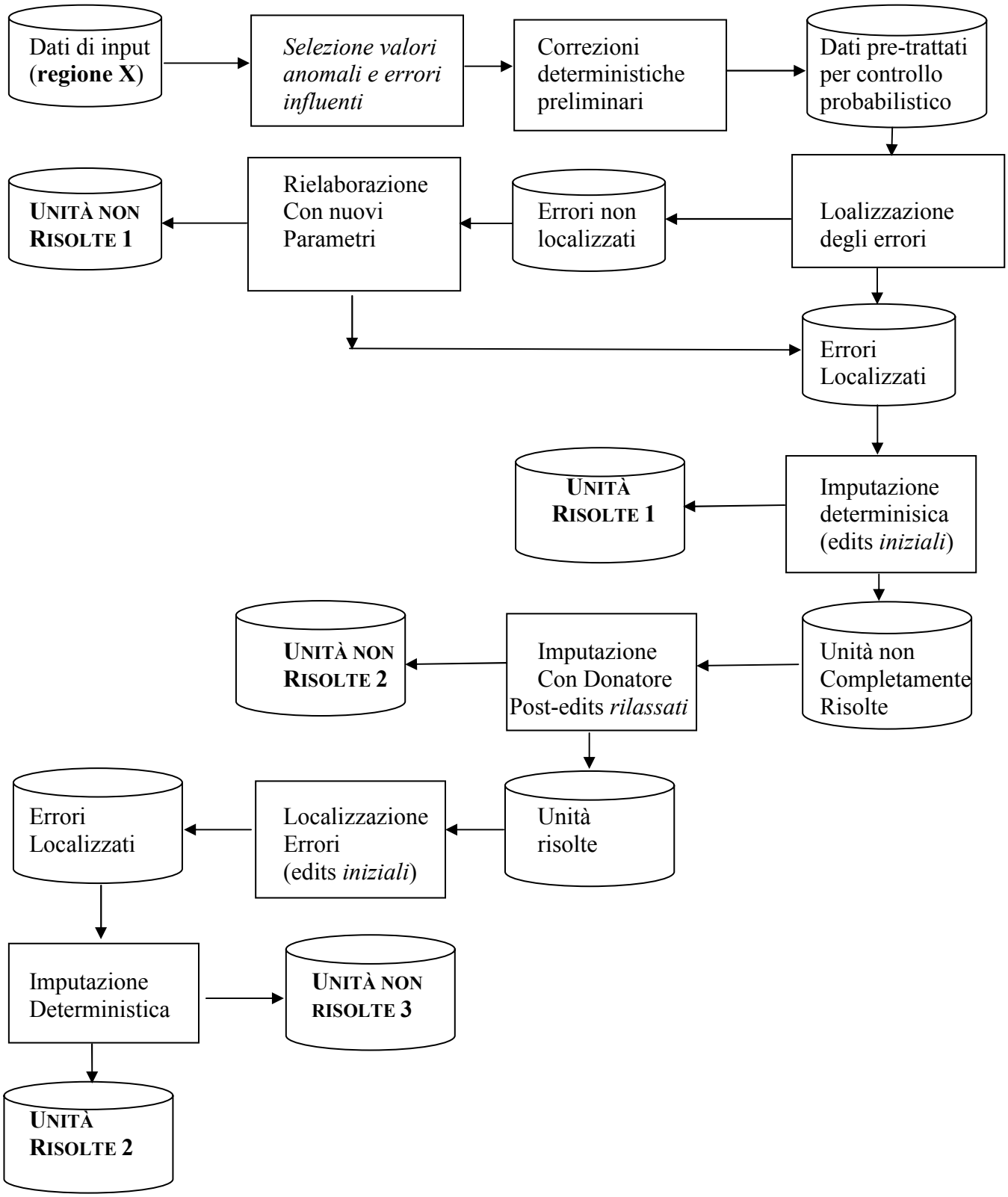
ISTAT. *Indagine REA – Rilevazione sui Risultati Economici delle Aziende Agricole.* http://www.istat.it/strumenti/rispondenti/indagini/rea/indice_rea.html, 2007b.

Little, R.J.A. "Missing Data Adjustments in Large Survey". *Journal of Business & Economic Statistics, 6, 287*-295, 1988.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Tempelman C., Hulliger B. e Kilchmann D. *Recommended Practices for Editing and Imputation for Cross-Sectional Business Surveys*, disponibile sul sito edimbus.istat.it, 2007.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*..Wiley & Sons, New York, 1987.

APPENDICE - Schema generale della procedura automatica di controllo e correzione delle variabili su superfici e coltivazioni principali (fase 1) dell'indagine SPA



Il nuovo sistema di controllo e correzione dei dati nella rilevazione annuale della produzione industriale

Carlo Ferrante, *Istat, Direzione Centrale delle Statistiche Economiche Strutturali*

Sommario: obiettivo del presente documento è la descrizione della nuova procedura di controllo e correzione utilizzata nella rilevazione annuale della produzione industriale. Una breve presentazione delle principali caratteristiche dell'indagine precede la definizione dei criteri che hanno ispirato la strategia complessiva della procedura di check. Successivamente vengono illustrate le problematiche principali che presenta l'indagine e uno schema generale che descrive il flusso delle operazioni di trattamento dei dati. I paragrafi seguenti descrivono le fasi più significative e maggiormente innovative che caratterizzano la procedura: i controlli implementati nel questionario elettronico, la fase di microediting e le tecniche di imputazione sul singolo questionario di unità locale, la gestione delle unità influenti, la fase di microediting finale applicata sui dati di impresa, le tecniche di macroediting e la fase di valutazione e documentazione tecnica e metodologica. Il documento presenta, inoltre, una serie di tavole che riportano informazioni quantitative elaborate dalla procedura sui dati di indagine per l'anno 2005.

Parole chiave: prodcom (produzione industriale), questionario elettronico, unità locale, unità influenti, microediting, macroediting

1. Caratteristiche generali dell'indagine

L'indagine annuale della produzione industriale rileva informazioni armonizzate a livello europeo sulle tipologie di prodotti industriali realizzati in Italia e sui livelli produttivi conseguiti, espressi in quantità e valore, dettagliati per ciascuna delle circa 4500 voci presenti nella classificazione di prodotti definita lista Prodcom. La predisposizione della classificazione è stata curata da Eurostat in collaborazione con i paesi e con le rappresentanze delle associazioni industriali europee. L'elenco si riferisce principalmente ai beni materiali, ma sono compresi anche alcuni servizi industriali (perfezionamento, riparazione, manutenzione e installazione). La descrizione dei singoli prodotti è accompagnata da un codice identificativo a otto cifre concordato a livello comunitario (codice Prodcom), una definizione, un'unità di misura per la rilevazione delle quantità prodotte e un riferimento alla Nomenclatura combinata, utilizzata per le statistiche di interscambio con l'estero. In linea generale, le prime sei cifre corrispondono alla classificazione Cpa (Nomenclatura comunitaria dei prodotti per attività[8]), mentre le ultime due cifre stabiliscono un riferimento alla Nomenclatura combinata.[9] Ogni voce, inoltre, è collegata a una divisione della classificazione europea delle attività economiche Nace Rev.1.1. Le divisioni incluse nell'elenco Prodcom sono quelle da 13 a 22 e da 24 a 36.

Le informazioni statistiche sono elaborate secondo metodi, concetti, definizioni e classificazioni coerenti con le disposizioni del regolamento (Cee) del Consiglio n. 3924 del 19 dicembre 1991, relativo a un'indagine comunitaria sulla produzione industriale, integrato dal regolamento Ce della Commissione n. 912/2004, riguardante le modalità di applicazione del regolamento Ce n. 3924/91 del Consiglio.

La rilevazione è condotta sulla totalità delle unità locali produttive delle imprese industriali con almeno 20 addetti e su un campione rappresentativo delle imprese industriali aventi numero di addetti compreso fra 3 e 19. Le unità di osservazione (unità locali produttive) investigate sono state rispettivamente 68.738 e 63.028 per gli anni di rilevazione 2005 e 2006, mentre le unità di rilevazione (imprese industriali e non) ammontano per i due anni citati a 47.257 e 47.132.

L'universo di riferimento della rilevazione è quello delle unità locali produttive che effettuano attività di trasformazione industriale e che rientrano nelle seguenti divisioni della classificazione delle attività economiche Nace Rev.1.1:

13 - Estrazione di minerali metalliferi;
14 - Altre industrie estrattive;
15 - Industrie alimentari e delle bevande;
16 - Industria del tabacco;
17 - Industrie tessili;
18 - Confezione di articoli di abbigliamento; preparazione e tintura di pellicce;
19 - Preparazione e concia del cuoio; fabbricazione di articoli da viaggio, borse, marocchineria, selleria e calzature;
20 - Industria del legno e dei prodotti in legno e sughero, esclusi i mobili; fabbricazione di articoli di paglia e materiali da intreccio;
21 - Fabbricazione della pasta-carta, della carta e dei prodotti di carta;
22 - Editoria, stampa e riproduzione di supporti registrati;
24 - Fabbricazione di prodotti chimici e di fibre sintetiche e artificiali;
25 - Fabbricazione di articoli in gomma e materie plastiche;
26 - Fabbricazione di prodotti della lavorazione di minerali non metalliferi;
27 - Metallurgia;
28 - Fabbricazione e lavorazione dei prodotti in metallo, escluse macchine e impianti;
29 - Fabbricazione di macchine e apparecchi meccanici;
30 - Fabbricazione di macchine per ufficio, di elaboratori e sistemi informatici;
31 - Fabbricazione di macchine e apparecchi elettrici n.c.a.;

[8] Regolamento (Ce) n.204/2002 della Commissione, del 19 dicembre 2001, che modifica il regolamento (Cee) n. 3696/93 del Consiglio relativo alla classificazione statistica dei prodotti associati all'attività economica nella Comunità economica europea.

[9] In taluni casi le ultime due cifre possono assumere altro significato, specificato in apposite note allegate al regolamento Prodcom.

32 - Fabbricazione di apparecchi radiotelevisivi e di apparecchiature per le comunicazioni;
33 - Fabbricazione di apparecchi medicali, di apparecchi di precisione, di strumenti ottici e di orologi;
34 - Fabbricazione di autoveicoli, rimorchi e semirimorchi;
35 - Fabbricazione di altri mezzi di trasporto;
36 - Fabbricazione di mobili; altre industrie manifatturiere;

Sono pertanto escluse dall'osservazione le voci di prodotto appartenenti alle seguenti divisioni:
10 - Estrazione di carbon fossile e lignite; estrazione di torba;
11 - Estrazione di petrolio greggio e di gas naturale; servizi connessi all'estrazione di petrolio e di gas naturale, esclusa la prospezione;
12 - Estrazione di minerali di uranio e di torio;
23 - Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari;
37 – Recupero e preparazione per il riciclaggio;
40 - Produzione e distribuzione di energia elettrica, di gas e di calore.

Per ciascuna voce di prodotto dell'elenco Prodcom la rilevazione osserva le seguenti variabili:

- la quantità prodotta in conto proprio o per conto terzi nell'unità locale durante l'anno di riferimento;
- la quantità prodotta nell'unità locale, anche anteriormente l'anno di riferimento, reimpiegata nel processo produttivo nel corso dell'anno di riferimento per la produzione di altri prodotti;
- la quantità prodotta nell'unità locale, anche anteriormente l'anno di riferimento, trasferita ad altre unità locali dell'impresa per una successiva lavorazione e/o trasformazione;
- la quantità prodotta per conto terzi in Italia durante l'anno di riferimento, con la precisazione che secondo il regolamento Prodcom, effettua produzione per conto terzi l'impresa (commissionario) che riceve le materie prime da un'altra impresa (committente) senza fattura, le trasforma e rende al committente il prodotto di tale processo: committente e commissionario debbono essere due imprese diverse e non stabilimenti della stessa impresa;
- il compenso corrisposto dalle imprese committenti, al netto dell'Iva, per la produzione effettuata in Italia per conto terzi nel corso dell'anno di riferimento;
- la produzione venduta durante l'anno di riferimento indipendentemente dall'epoca in cui è stata realizzata: la variabile non comprende né la produzione effettuata per conto terzi né la produzione acquistata da terzi e rivenduta nel medesimo stato, mentre comprende la produzione fatta realizzare a terzi in Italia, dietro fornitura di materie prime senza fattura, venduta nell'anno di riferimento;
- la produzione fatta realizzare a terzi in Italia dietro fornitura di materie prime senza fattura, venduta nell'anno di riferimento;
- il valore della produzione venduta durante l'anno di riferimento.

Oltre alle informazioni relative ai prodotti realizzati, la rilevazione osserva alcune variabili ausiliarie, quali gli acquisti di prodotti energetici, in quantità e valore, e la media degli occupati nel periodo di riferimento. In totale il questionario di indagine presenta 15 variabili tutte di tipo quantitativo.

Il disegno campionario adottato prevede un campione di imprese con 3-19 addetti a uno stadio stratificato. In particolare, si tratta di un piano di campionamento equiprobabilistico all'interno di ciascuno strato (classe di attività economica e ripartizione geografica) con selezione delle unità senza reimmissione. La selezione delle unità da includere nella rilevazione è avvenuta ricorrendo all'archivio Asia. Il disegno di campionamento è stato definito nel quadro della strategia di coordinamento dei campioni per le indagini strutturali sulle imprese, utilizzata dall'Istat per minimizzare l'onere statistico sulle unità produttive. La selezione delle unità dall'archivio Asia è avvenuta secondo criteri che hanno assicurato la casualità del campione.

La metodologia utilizzata per il calcolo dei pesi finali è quella degli stimatori di ponderazione vincolata. Tale tecnica consente di modificare i pesi iniziali, ovvero quelli che descrivono il piano di campionamento, di ciascuna unità rispondente in pesi finali che, sotto certe ipotesi, attenuano l'effetto distorsivo delle stime dovuto sia alle mancate risposte totali, sia alla sottocopertura della lista da cui è selezionato il campione. Questi stimatori garantiscono inoltre l'uguaglianza tra alcuni parametri noti

della popolazione e le corrispondenti stime campionarie: quanto più le variabili ausiliarie sono correlate alle variabili oggetto d'indagine, tanto più efficienti risultano essere gli stimatori.

A partire dall'anno di riferimento 2004 accanto al tradizionale questionario cartaceo autocompilato, le imprese possono scegliere di compilare il questionario elettronico disponibile sul sito INDATA dell'ISTAT.

2. Strategia complessiva della nuova procedura di controllo e correzione: criteri ispiratori

Nel corso del 2006 e nei primi mesi del 2007 è stata effettuata una revisione radicale della procedura di controllo e correzione dell'indagine Prodcom, applicata per la prima volta ai dati rilevati per l'anno di indagine 2005. La procedura di check è stata aggiornata in modo sostanziale sia negli aspetti metodologici che in quelli informatici.

La procedura di controllo precedente risaliva all'anno di riferimento 1996; essa recepiva per la prima volta in Italia le disposizioni previste nel regolamento Prodcom ed era implementata in diversi ambienti di sviluppo (Cobol, Sas, Oracle). Il primo obiettivo, pertanto, è stato quello di aggiornare una procedura ormai datata e di unificare gli ambienti di sviluppo decidendo di implementare la procedura in un'unica piattaforma informatica (Oracle).

Inoltre, la nuova procedura di controllo si è posta la finalità di ottimizzare gli interventi interattivi da parte dei revisori. La riduzione di risorse umane verificatasi negli ultimi anni ha reso necessario un incremento degli automatismi e una riduzione del numero dei ritorni sullo stesso questionario per essere controllato dai singoli revisori, con l'obiettivo finale di revisionare i dati in un'unica occasione. Questa ottimizzazione è stata conseguita mediante l'utilizzo di un approccio di tipo selettivo sia in fase di editing sia in fase di follow-up, verificando solo i questionari delle unità influenti e limitando, di conseguenza, il fenomeno dell'over-editing. Ulteriori funzionalità introdotte nella nuova procedura riguardano l'implementazione di controlli ad hoc nella fase di verifica e correzione dei microdati, sempre per contenere il numero dei ritorni successivi al calcolo delle stime (macrodati). Per aumentare la qualità degli interventi manuali sono state uniformate le modalità di intervento dei revisori mediante l'incremento della disponibilità di fonti ausiliarie, la documentazione con indicazioni precise per le tipologie di errore più frequenti, la sostituzione di alcuni interventi interattivi con delle forzature automatiche, la formazione continua del personale.

La progettazione della nuova procedura di controllo e correzione è in completa sintonia con le potenzialità offerte dall'introduzione del questionario elettronico. In particolare sono stati introdotti una serie di controlli direttamente nella fase di raccolta dati, al momento della compilazione del questionario elettronico, utilizzato da un numero crescente di imprese. Inoltre, a partire dall'anno di riferimento 2004, entrambe le tipologie di questionari utilizzati (cartacei e elettronici) sono precodificati con le informazioni sui codici di prodotto dichiarati dalle imprese negli anni precedenti, incrementando così la qualità e la precisione delle informazioni raccolte.

Le novità apportate alla procedura di check hanno permesso di colmare alcune lacune presenti nella vecchia procedura. Ad esempio sono tenute maggiormente in considerazione le relazioni fra unità locali appartenenti alla stessa impresa: ai controlli intra-questionario sono stati aggiunti quelli fra questionari (ad esempio sulla destinazione dei trasferimenti interni e sulla ricostruzione dei livelli di fatturato complessivi dell'impresa per evidenziare eventuali unità locali mancanti). Sono stati estesi i controlli riguardanti la sezione degli acquisti di prodotti energetici data l'importanza che i dati in questione hanno acquisito ultimamente, originando una pubblicazione autonoma. Particolare attenzione è stata rivolta all'applicazione più rigorosa e puntuale di controlli basati sulle relazioni esistenti fra le sezioni del questionario, così come all'individuazione di fonti di errore di natura sistematica (ad esempio produzione estera, reimpieghi, trasferimenti interni e servizi industriali) fondamentale per intervenire nella fase di progettazione dell'indagine dell'anno successivo. Inoltre la procedura è stata predisposta per fornire stime affidabili a livello aggregato congiuntamente ad un file di microdati corretto a livello elementare e per incrementare la coerenza e la confrontabilità con fonti esterne a livello nazionale ed europeo.

La revisione della procedura di controllo si è accompagnata ad un'attività, purtroppo carente in passato, di documentazione completa sia della procedura in sé sia dei risultati dell'applicazione di tale procedura ai dati di indagine. Tale attività di documentazione è stata sviluppata con l'obiettivo di soddisfare diverse esigenze di rilevanza strategica: calcolare indicatori di qualità sui dati raccolti richiesti dal sistema SIDI; adeguamento alle direttive interne (es. TRAD03) e comunitarie; fornire statistiche utili per il perfezionamento della stessa procedura di controllo; monitorare lo stato di lavorazione in ciascun settore produttivo; fornire informazioni per il miglioramento dell'intero processo produttivo di indagine.

Una scelta che ha influito sulla progettazione e predisposizione della procedura è stata quella di gerarchizzare gli errori in tre macro classi. Gli errori definiti "bloccanti", che devono obbligatoriamente essere corretti dai revisori, e riguardano principalmente errori delle unità influenti, inesattezze sulle unità di misura associate ai prodotti o relazioni errate tra variabili (ad esempio la produzione realizzata conto terzi non può mai essere maggiore della produzione totale). La seconda tipologia definita "accertamento" riguarda invece i potenziali errori che possono essere mantenuti nei microdati (ad esempio il prezzo medio di un prodotto dichiarato da un'impresa non in linea con il prezzo medio di riferimento nazionale). La terza tipologia, infine, è costituita dalle "forzature automatiche", ovvero gli errori individuati e corretti direttamente dal sistema di controllo che riguardano principalmente le unità non influenti oppure il ripristino di alcune relazioni esatte tra variabili (ad esempio sono stati dichiarati i 'di cui' di una certa variabile quantitativa senza i dati di livello superiore nella sezione acquisti energetici; in questo caso la procedura imputa automaticamente i dati mancanti).

3. Problematiche principali dell'indagine

Un'attenta analisi delle problematiche principali della rilevazione Prodcom ha preceduto e successivamente indirizzato la strategia complessiva di riprogettazione della procedura di controllo e correzione.

Oltre alle mancate risposte totali, presenti nella maggior parte delle indagini strutturali sulle imprese, e dovute soprattutto ai rispondenti o ad errori di lista, questa analisi preliminare ha evidenziato la presenza di mancate risposte parziali dovute principalmente alla presenza nei dati di unità di misura per le quantità, associate al prodotto, non coerenti con quelle in uso presso le imprese. Inoltre alcune fra le variabili richieste nel questionario non sempre sono disponibili nella contabilità di impresa o nel controllo di gestione.

Le principali cause che generano errori di misura dovuti al rispondente sono:

la non corretta individuazione dei prodotti fra quelli disponibili nell'elenco Prodcom;

i dati in quantità e/o valore forniti con unità di misura non appropriata;

la non corretta interpretazione di alcune delle variabili richieste (in particolare produzione conto terzi, reimpieghi, trasferimenti interni e servizi industriali);

la non corretta attribuzione territoriale di produzione realizzata all'estero. In particolare le industrie tessili, calzaturiere e della produzione di macchine, dove è in atto una forte delocalizzazione dell'attività produttiva all'estero, a volte tendono a dichiarare anche la produzione che non viene realizzata sul territorio italiano;

per le imprese plurilocalizzate, esiste una eterogeneità nella gestione del 'grappolo' di unità locali produttive, sia nel numero dei questionari compilati, in quanto molte imprese compilano uno o parte dei questionari di rilevazione, sia nel soggetto che effettua la compilazione, a livello centralizzato o di singolo stabilimento produttivo. Questa mancanza di uniformità può generare dati non armonizzati, mancanti (ad esempio trasferimenti fra stabilimenti produttivi) o incoerenti (ad esempio incongruenze tra la produzione dichiarata a livello di impresa e gli addetti che si riferiscono al singolo stabilimento).

Alcuni errori di misura sono dovuti agli strumenti di raccolta delle informazioni. Ad esempio, l'elenco dei prodotti Prodcom non sempre è conforme con le classificazioni in uso presso le imprese.

Infine alcuni errori di elaborazione sono generati dalla duplice modalità di raccolta delle informazioni (cartacea ed elettronica).

4. Il processo di controllo e correzione: schema generale

Il seguente diagramma di flusso descrive il processo di controllo e correzione dalla fase di ricezione dei questionari fino alla produzione delle stime finali. In seguito saranno descritte solo alcune delle fasi, in particolare quelle più significative e maggiormente innovative del processo stesso: i controlli implementati nel questionario elettronico, la fase di microedeting e le tecniche di imputazione applicate sul singolo questionario di unità locale, la gestione delle unità influenti, la fase di microediting finale a livello di impresa, la fase di macroediting.

Diagramma1 - Il processo di controllo e correzione: schema generale

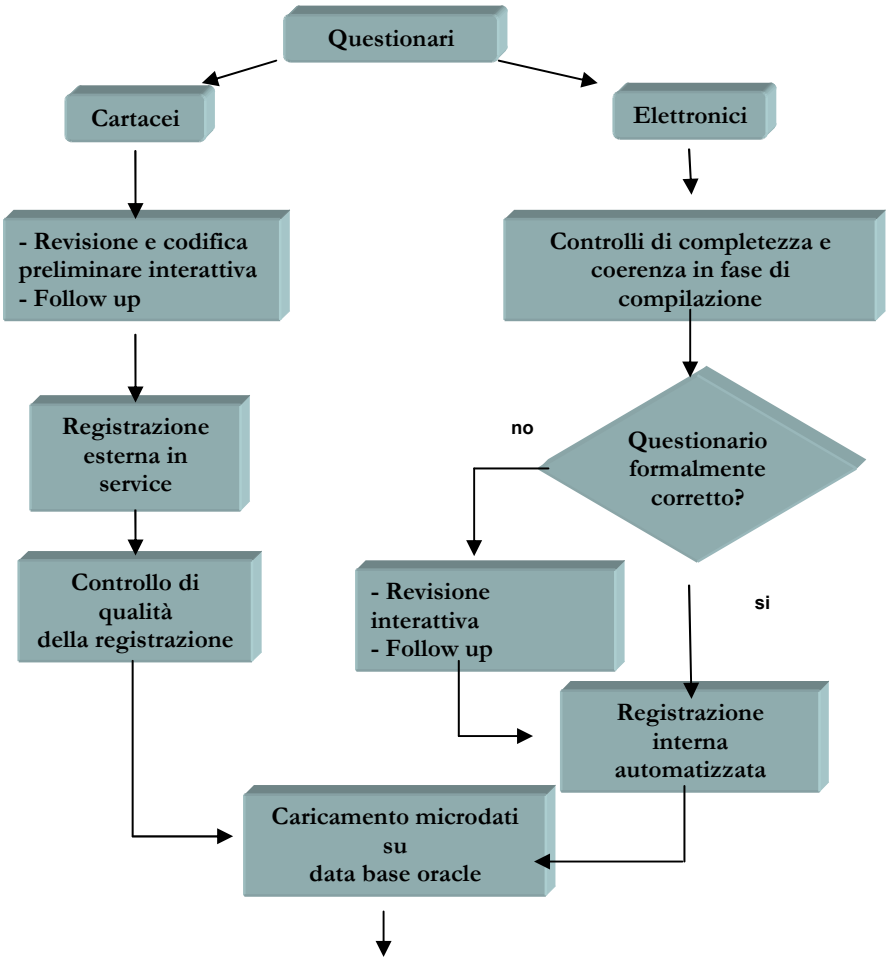
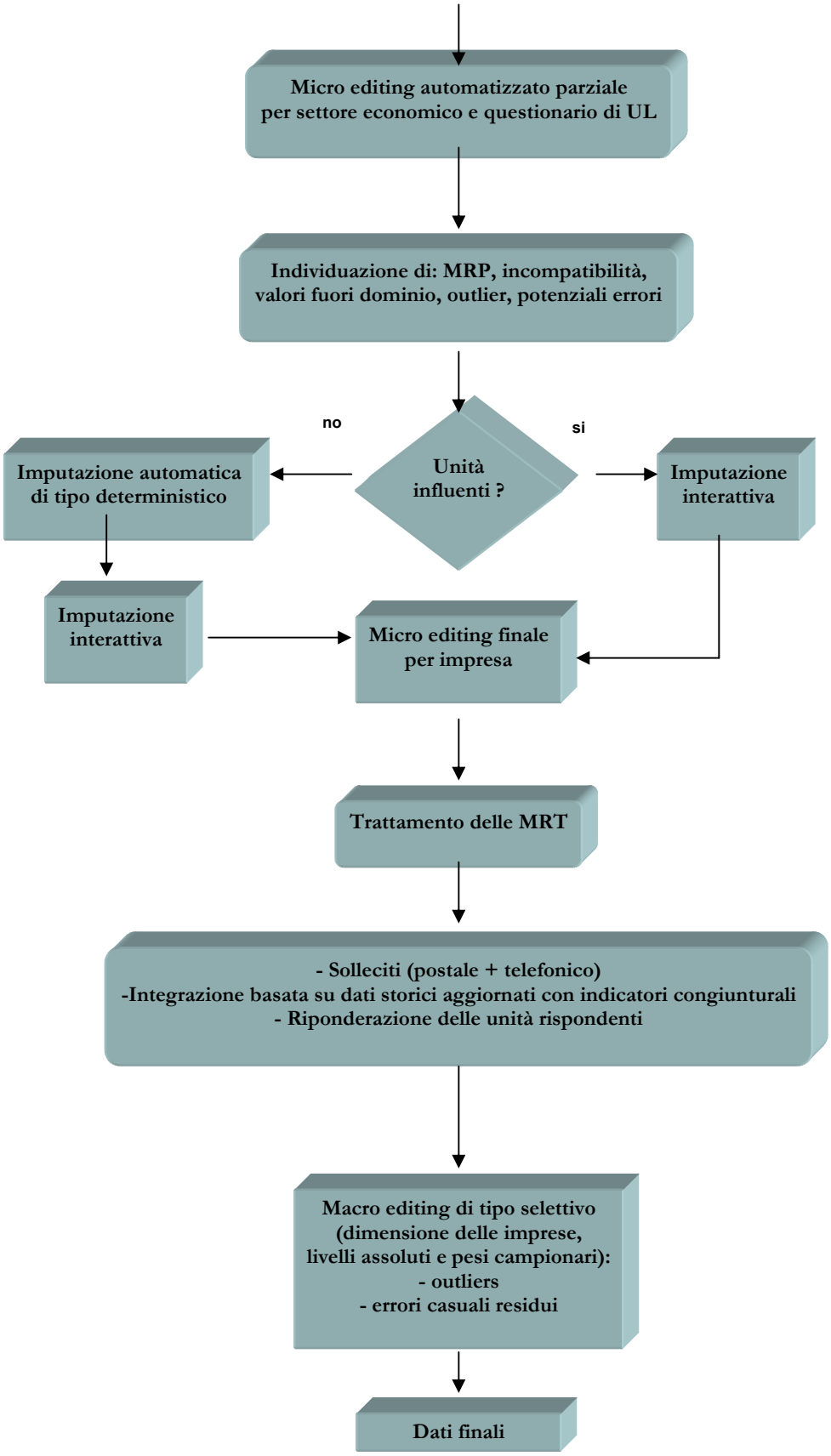


Diagramma1 segue - Il processo di controllo e correzione: schema generale



5. Controlli di completezza, coerenza e accuratezza nella fase di compilazione del questionario elettronico

Come già accennato, dall'anno di riferimento 2004 i dati della rilevazione Prodcom vengono rilevati con una duplice modalità di raccolta: oltre al tradizionale questionario cartaceo, le imprese possono scegliere di compilare il questionario elettronico. Questo, implementato su fogli di lavoro Excel, può essere prelevato direttamente dal compilatore accedendo al sito Indata dell'Istituto, salvato in locale su un proprio pc, compilato e successivamente trasmesso accedendo nuovamente al sito Indata. Come evidenziato dal diagramma di flusso, i questionari cartacei e i questionari elettronici seguono due metodi di lavorazione sensibilmente diversi prima che le informazioni di entrambi convergano nello stesso data base Oracle, ove sono sottoposti alla procedura di check.

I questionari cartacei subiscono inizialmente un controllo preliminare interattivo in cui si verifica la presenza delle informazioni obbligatorie, vengono in parte sanate le mancate risposte parziali (presenza di dati solo in valore o solo in quantità) e si codificano le unità di misura dichiarate dalle imprese e i prodotti che le stesse hanno fornito solo mediante la descrizione senza il relativo codice prodotto (sezione B del questionario).

A differenza dei questionari cartacei, che vengono tutti revisionati preliminarmente in modo manuale e il cui impatto in termini numerici tende notevolmente a ridursi negli anni, i questionari elettronici sono sottoposti ad un processo di lavorazione iniziale completamente diverso che tende a sfruttare le potenzialità offerte dallo strumento informatico in cui sono implementati, con evidenti ricadute positive in termini di riduzione delle risorse impiegate per la loro lavorazione. Infatti, molti dei questionari elettronici, correttamente compilati già in fase di data entry attraverso la verifica delle informazioni dichiarate all'atto della compilazione, vengono direttamente registrati senza richiedere un controllo preliminare interattivo; di questi, una buona percentuale risulterà corretta anche alla successiva fase di controllo e correzione. L'utilizzo di alcuni vincoli di completezza, coerenza e accuratezza applicati all'atto della registrazione delle informazioni garantisce una maggiore accuratezza dei dati rispetto ad un insieme di controlli di base, anche se la non esaustività dei controlli stessi (lo strumento Excel non permette l'applicazione di alcuni vincoli di coerenza) e la possibilità di mantenere attraverso delle forzature un dato incoerente, fanno sì che nei microdati registrati possano ancora permanere situazioni di errore.

I controlli di completezza presenti hanno comunque permesso la riduzione di alcune mancate risposte parziali (es. addetti, i 'di cui' nella sezione dei prodotti energetici).

La precodifica dei codici prodotto dichiarati dalle imprese negli anni precedenti permette di avere dati storici più coerenti. I controlli di coerenza implementati si basano sulla presenza di segnalazioni non vincolanti di incoerenze tra rapporti di variabili strategiche (ad esempio valori e compensi medi unitari calcolati sui dati dichiarati che risultano "fuori range"), oppure sulla segnalazione di alcune relazioni logiche non valide (ad esempio produzione conto terzi maggiore della produzione totale; la produzione venduta ricevuta da terzi maggiore della produzione venduta; i 'di cui' dei prodotti energetici maggiori dei valori dei totali di riferimento). I controlli di incoerenza verificano anche le relazioni esistenti tra sezioni del questionario (ad esempio il valore totale degli acquisti di prodotti energetici e il valore totale della produzione venduta).

L'accuratezza delle informazioni dichiarate è garantita attraverso una serie di controlli ad hoc. I codici prodotto dichiarati sono sottoposti ad un controllo di validità mediante il confronto con la lista Prodcom. Il questionario elettronico dispone, a differenza del questionario cartaceo, della lista completa dei prodotti ed inoltre, facilita la selezione del codice prodotto attraverso dei criteri gerarchici di ricerca a livello di classe di attività economica o di classificazione CPA (prime sei cifre del codice Prodcom). Le unità di misura sono codificate automaticamente, così come gli addetti dichiarati sono arrotondati automaticamente all'unità più prossima. Vengono segnalati i casi di dati non validi (i valori dichiarati devono essere tutti interi e maggiori di zero). Infine, per i questionari pervenuti che non risultano essere formalmente corretti, la procedura di trasmissione telematica permette al singolo revisore la possibilità di effettuare un ritorno presso le imprese in tempo quasi reale.

6. Microediting parziale per questionario di unità locale

I microdati presenti in ciascun questionario, dopo essere stati registrati in un data base Oracle, sono sottoposti alla procedura di controllo e correzione. Questa è suddivisa in due macro fasi. La prima fase è indirizzata esclusivamente alla verifica delle informazioni del singolo questionario ed ha inizio dal momento in cui cominciano a tornare i questionari compilati trasmessi dalle imprese. La seconda fase, che inizia al termine della fase di raccolta dei questionari, ha invece l'obiettivo di validare le informazioni a livello di impresa. Infatti, per le imprese plurilocalizzate, i singoli questionari di unità locale, specialmente se per la compilazione sono "distribuiti" dall'impresa ai singoli stabilimenti produttivi, possono arrivare in momenti diversi, per cui la ricostruzione esatta della situazione di impresa ed il relativo controllo può avvenire solo al termine dell'indagine.

Il microediting parziale per questionario di unità locale è stato sviluppato mediante un software ad hoc su piattaforma Oracle. Pertanto esso non fa uso, come il resto della procedura, di software dedicati generalizzati. Si basa su di un controllo intra-record ed è organizzato sia per singole sezioni del questionario (informazioni anagrafiche, prodotti, servizi industriali, lavoro, acquisti energetici) sia tra sezioni.

L'approccio scelto si basa sulla localizzazione degli errori di tipo automatico che individua mancate risposte parziali, incompatibilità, valori fuori dominio, valori anomali e dati potenzialmente in errore. La tecnica di imputazione è di tipo deterministico. Questa scelta è stata guidata da una serie di analisi che hanno permesso di identificare abbastanza facilmente la fonte dell'errore per la maggior parte degli *edit* violati, essendo gli stessi costruiti su un insieme di variabili le cui relazioni non sono affatto complesse. Inoltre, la stessa analisi ha evidenziato la maggiore affidabilità delle variabili dichiarate in valore rispetto a quelle dichiarate in quantità e, per la maggior parte delle tipologie di errore, ha fornito indicazioni su quali variabili intervenire e quali valori assegnare alle variabili in errore. Un ulteriore elemento che ha condotto a questa scelta è la constatazione che la percentuale di record che presentano più di due errori è normalmente bassa.

Per quanto riguarda l'individuazione dei valori anomali, la localizzazione si basa su intervalli di accettazione calcolati in modo empirico su domini disgiunti (ad esempio divisione di attività economica, singolo prodotto industriale o energetico) dell'intera popolazione. Le tecniche di determinazione degli intervalli sono diverse a seconda della variabile in esame e si basano principalmente sull'utilizzo di funzioni (in forma di *edit* lineari o di rapporti), informazioni storiche e analisi congiunta delle distribuzioni a livello trasversale e longitudinale (ad esempio per i valori e i compensi medi unitari). L'analisi dei valori anomali individuati è di tipo interattivo. Se il dato risulta corretto, solo per quei valori che hanno un impatto notevole sulle stime (unità influenti) si interviene, se possibile, sui pesi campionari nella successiva fase di macroediting lasciando inalterato comunque il microdato.

In questa fase, una volta determinati i record con errori o potenzialmente in errore, si procede con l'applicazione di tecniche di *editing* selettivo utilizzato per individuare le unità influenti: viene controllato e corretto interattivamente, con particolare cura, solo quel sottoinsieme di record che ha un impatto maggiore sulle stime finali, mentre la restante parte dei record è sottoposta prevalentemente ad imputazione automatica di tipo deterministico. Questa scelta è seguita ad un'analisi delle distribuzioni delle principali variabili di stima (produzione totale e produzione venduta) che risultano essere fortemente asimmetriche: in questo modo si concentra l'impiego di risorse umane, decrescente nella disponibilità, sulle unità più importanti garantendo l'affidabilità delle stime.

7. Tecniche di imputazione applicate sul singolo questionario di unità locale

Una volta individuate le variabili contenenti gli errori che hanno causato l'attivazione delle incompatibilità, oppure i cui valori sono stati giudicati *outlier*, occorre procedere alla fase di imputazione di tali variabili, al fine di rimuovere gli errori cercando di ripristinare i valori corretti. I metodi di imputazione utilizzati sono di tipo automatico o interattivo a seconda della tipologia di errore e dell'importanza dell'unità di rilevazione (unità influente o meno).

L'imputazione automatica utilizzata è di tipo deterministico, principalmente perché l'indagine deve fornire esclusivamente stime di totali e, inoltre, questo metodo presenta il vantaggio dell'orientabilità degli effetti. Essendo stata individuata una gerarchia circa la correttezza delle variabili, il metodo permette anche di orientare le modifiche verso quelle variabili ritenute meno affidabili. In genere le variabili dichiarate in quantità sono meno affidabili di quelle dichiarate in valore e, inoltre, i concetti di produzione conto terzi, reimpieghi, trasferimenti interni e servizi industriali sono meno familiari di quelli di produzione totale o di produzione venduta.

Le tipologie di errore corrette con i metodi deterministici sono principalmente le mancate risposte parziali, le incompatibilità logico/matematiche tra variabili e alcuni errori sistematici. I metodi utilizzati si basano essenzialmente sui dati disponibili per le unità rispondenti congiuntamente ad altre informazioni ausiliarie usate per definire le celle di imputazione e basate su variabili di stratificazione (attività economica, classe di addetti e ripartizione geografica) e sui domini di stima (classe di attività economica). Le tecniche utilizzate, a seconda delle variabili coinvolte, consistono nell'imputazione da valore prefissato, imputazione logica, imputazione basata su modelli (ad esempio utilizzo dei valori mediani unitari per le mancate risposte parziali).

Per la maggior parte degli errori riscontrati sulle unità influenti e per gli *outlier* in generale si utilizzano tecniche di imputazione interattiva. Il sistema fornisce un elenco di errori suddiviso per settore economico (divisione e/o gruppo di attività economica) e tipologia di errore che può assumere la forma di accertamento o errore bloccante e che comprendono anche gli *outlier*. Il revisore ha la possibilità di mantenere anche gli errori bloccanti previa analisi approfondita e idonea documentazione giustificativa da indicare nel *check* e da utilizzare nella fase di monitoraggio finale della procedura di controllo. I principali strumenti utilizzati dai revisori in questa fase consistono nell'analisi del modello di rilevazione, *follow-up*, dati storici di impresa, valore mediano unitario di prodotto, relazioni logiche tra variabili e tra sezioni del questionario, utilizzo di donatori (ad esempio consumo energetico per addetto), valori mediani unitari di fonte Eurostat per prodotti nuovi o mai dichiarati, utilizzo di manuali per il trattamento di errori particolari, altre fonti interne ed esterne.

8. Gestione unità influenti

La procedura di controllo e correzione è differenziata a seconda dell'appartenenza o meno del *record* ad un'unità influente.

Le unità influenti sono definite in base alla dimensione (addetti e livello di fatturato), all'importanza nella determinazione delle stime finali di prodotto per le principali variabili (produzione totale e produzione venduta in quantità e valore), al peso campionario e alla gravità dell'errore. Ad esse, come già accennato, si applicano prevalentemente tecniche di correzione interattiva. Le stesse unità sono, inoltre, sottoposte ad un monitoraggio continuo per settore economico al fine di assicurarne la collaborazione all'indagine. Per le unità che presentano più di dieci unità locali è anche previsto il contatto diretto da parte dei revisori in fase di spedizione, per concordare l'invio del materiale di rilevazione. Infine, nella fase di gestione dei solleciti telefonici le unità influenti sono contattate direttamente da personale interno ISTAT, a differenza delle altre unità di rilevazione contattate da personale di apposite società specializzate.

Alle unità definite non influenti si applicano delle tecniche di correzione di tipo automatico deterministico. Le correzioni effettuate sono comunque rese disponibili agli operatori per valutare come la procedura ha modificato i dati grezzi con *edit* falliti. Le imputazioni automatiche, in passato di tipo interattivo e con elevate frequenze di attivazione, hanno permesso di liberare risorse da destinare a controlli più mirati e qualificanti, di ridurre il carico statistico sulle imprese limitando i ritorni sulle stesse per eventuali chiarimenti, di incrementare la tempestività di diffusione dei risultati di indagine verso gli utenti finali e, inoltre, di ridurre alcuni errori per le elaborazioni più complesse (ad esempio il riproporzionamento delle quantità dichiarate dai rispondenti utilizzando unità di misura diversa da quella prevista) o nei casi di *editing* creativo (ad esempio nelle dichiarazioni di reimpieghi e/o trasferimenti interni).

9. Microediting finale per impresa

Una volta terminata la prima delle due macrofasi di *check*, i microdati dei singoli questionari di unità locale risultano corretti e coerenti al loro interno. Ultimata la fase di ritorno dei questionari, inizia la seconda macrofase di controllo il cui obiettivo principale è quello di garantire la coerenza fra record di unità locali appartenenti alla stessa impresa (controllo inter-record) con l'eventuale possibilità di individuare unità locali non rispondenti da integrare o unità locali duplicate. Inoltre, nell'ottica di ottimizzare gli interventi interattivi da parte dei revisori, sono stati anticipati, in questa fase, alcuni controlli per limitare successivamente quelli a livello macro seguenti al calcolo dei pesi campionari.

I controlli implementati in questa fase sono tutti impostati come accertamenti e analizzati interattivamente. Essi mirano innanzitutto a verificare che i prodotti industriali e quelli energetici, dichiarati a livello di impresa nel complesso vista come insieme di questionari di unità locali, siano coerenti con quelli dichiarati dalla stessa impresa nell'anno precedente. Questo controllo, strategico per la qualità delle stime finali, è stato impostato sulla base di un'analisi specifica condotta sulle caratteristiche produttive delle imprese rispondenti alla rilevazione Prodcom, che ha permesso di stabilire che circa l'80% di esse sono monoprodotto. Un secondo controllo, anch'esso di importanza fondamentale per individuare unità locali completamente non rispondenti o duplicate, consiste nel confrontare con l'anno precedente, per l'intera impresa, il livello della variabile chiave valore della produzione.

Un ulteriore controllo specifico è finalizzato a verificare l'esatta interpretazione del concetto di trasferimento interno all'impresa. Se in un questionario è stato indicato dal rispondente un trasferimento di prodotto, deve necessariamente esistere almeno un altro questionario di unità locale della stessa impresa con un codice prodotto diverso da quello trasferito. Spesso le imprese dichiarano trasferimenti dello stesso prodotto fra unità locali a fini di vendita e non produttivi come dovrebbe essere correttamente.

In questa fase i revisori, per validare i dati di impresa potenzialmente in errore, oltre al consueto *follow up* o all'analisi dei dati storici, possono consultare *on-line* altre fonti rese disponibili direttamente nella schermata della procedura di controllo. Queste fonti comprendono dati di bilancio depositati presso le Camere di Commercio e dati di fonte Istat relativi alle rilevazioni SCI-PMI (sistema dei conti delle imprese), all'archivio ASIA e alla rilevazione IULGI (indagine unità locali delle grandi imprese).

10. Macroediting

Terminata la fase di microediting, il passo successivo riguarda la gestione delle mancate risposte totali. Le imprese non rispondenti all'indagine e ai solleciti previsti (postale e telefonico) durante la fase di raccolta dati, sono trattate in due modi distinti. Per le imprese che presentano informazioni riferite alla precedente occasione d'indagine vengono applicate tecniche di integrazione basate su dati storici opportunamente aggiornati attraverso idonei indicatori congiunturali integrabili con le statistiche Prodcom. Per le imprese per le quali non sono disponibili informazioni storiche, o non è possibile o non si ritiene opportuno procedere alla loro integrazione, si interviene sui pesi campionari di unità rispondenti considerate rappresentative di quelle non rispondenti.

Terminato il trattamento delle mancate risposte totali si procede alla stima provvisoria delle variabili investigate e quindi alla fase di *macroediting* il cui obiettivo è quello di individuare errori casuali residui, *outliers* o eventuali errori introdotti nelle precedenti fasi di controllo e imputazione. Questo controllo è applicato sulle stime provvisorie a livello di prodotto delle sole variabili chiave (produzione totale in quantità e produzione venduta, in quantità e valore) mediante tecniche di tipo univariato basate sul confronto con dati storici di prodotto. Per verificare eventuali stime di prodotto non in linea negli anni si effettua un controllo di coerenza con altre fonti interne ed esterne all'Istituto, in particolare: risultati della rilevazione mensile della produzione industriale, statistiche di fonte SBS (Structural Business Statistics), statistiche di interscambio con l'estero, dati prodotti da associazioni di categoria. Anche in questa fase la localizzazione dei record da sottoporre a controllo interattivo utilizza tecniche di tipo

selettivo nell'individuazione delle unità influenti, definite principalmente dai livelli assoluti delle variabili chiave e dai pesi campionari associati.

Da notare che indicazioni utili in questa fase sono fornite anche da Eurostat, che effettua analisi sulla completezza e coerenza delle statistiche Prodcom a livello europeo, basate sul confronto dei valori mediani di prodotto elaborati sui dati trasmessi dai diversi Paesi, sui consumi apparenti e su informazioni storiche.

11. Valutazione e documentazione

Nella nuova procedura di controllo e correzione sono previste una serie di informazioni che ne documentano sia le prestazioni sia le specifiche metodologiche.

La *performance* della procedura nel suo complesso è stata validata attraverso una simulazione effettuata sul file di microdati corretto dell'indagine riferita all'anno 2004, al quale sono state apportate opportune "perturbazioni".

Il sistema è stato progettato per garantire un monitoraggio continuo in corso d'opera, mediante un controllo puntuale distinto per fase (*microediting* parziale per questionario di unità locale o finale per impresa), tipo di questionario (cartaceo o elettronico), settore economico (divisione o gruppo di attività economica) e tipologia di errore.

La procedura è stata anche predisposta per calcolare i principali indicatori circa l'impatto del *check* e la qualità dei dati (tasso di imputazione e non imputazione e relative componenti).

La documentazione metodologica attualmente disponibile può essere consultata nella nota metodologica allegata alla pubblicazione, nei quality reports prodotti per Eurostat e nelle specifiche tecniche ad uso interno contenenti l'insieme degli *edit* e le linee guida per gli interventi interattivi.

12. Informazioni elaborate dalla procedura

Nell'anno di indagine 2005 sono pervenuti 24.327 questionari di unità locale, di cui 19.576 utilizzabili (vedi tavola 1), equamente ripartiti tra le due modalità di compilazione, con una leggera preferenza, da parte delle imprese, per il tradizionale questionario cartaceo. I questionari che presentano errori, o potenziali errori, sono 12.832. Va sottolineato che i questionari cartacei sono tutti sottoposti ad una revisione preliminare da parte dei revisori prima di essere registrati e di questa attività non è possibile avere informazioni. Per cui nell'analisi delle tabelle presentate, un eventuale confronto tra le due modalità di somministrazione dei questionari deve essere fatta tenendo conto del diverso percorso seguito dai questionari prima della fase di controllo e correzione. I questionari elettronici, dal canto loro, permettono di sfruttare i vantaggi di una maggiore tempestività nella disponibilità del microdato, con innegabili benefici che si ripercuotono su tutto il successivo percorso di validazione fino alla produzione delle stime.

Tavola 1 - Questionari utilizzabili pervenuti per tipologia - Rilevazione annuale della produzione industriale - Anno 2005

Tipologia	Questionari pervenuti			
			Di cui errati	
	Numero	%	Numero	%
Cartacea	10.502	54	6.791	53
Elettronica	9.074	46	6.041	47
Totale	19576	100	12.832	100

Durante la fase di editing sono individuati i questionari da sottoporre a controllo, di cui più della metà presentano al massimo due errori. Percentuale che, per quanto puntualizzato sopra, sale addirittura al 66% per i questionari cartacei (vedi tavola 2).

Tavola 2 - Questionari errati per numero di errori e tipologia - Rilevazione annuale della produzione industriale - Anno 2005

Numero errori	Tipologia				Totale	
	Elettronici		Cartacei			
	Frequenza	%	Frequenza	%	Frequenza	%
1	1.482	25	2.797	41	4.279	33
2	1.178	20	1.716	25	2.894	23
> 2	3.381	55	2.278	34	5.659	44
Totale	6.041	100	6.791	100	12.832	100

I microdati registrati dai 19.576 questionari pervenuti sono 95.720 e di questi, con riferimento ai 12.832 questionari individuati come errati dalla procedura, 42.264 sono sottoposti a controllo (vedi tavola 3).

Tavola 3 – Microdati registrati - Rilevazione annuale della produzione industriale - Anno 2005

Corretti	Errati	Totale
53456	42264	95720

La tipologia di imputazione maggiormente applicata dalla procedura è quella interattiva con possibilità di accertamento da parte dei revisori (vedi tavola 4).

Tavola 4 – Microdati errati per tipologia di imputazione e tipo questionario - Rilevazione annuale della produzione industriale - Anno 2005

Tipologia imputazione	Tipo questionario				Totale	
	Elettronico		Cartaceo			
	Frequenza	%	Frequenza	%	Frequenza	%
Automatica	3.983	16	2.849	16	6.832	16
Interattiva con accertamento	10.766	44	8.855	50	19.621	47
Interattiva bloccante	9.934	40	5.877	34	15.811	37
Totale	24.683	100	17.581	100	42.264	100

Dopo un'attenta analisi da parte del personale, circa un terzo dei microdati vengono lasciati inalterati avendo la procedura, in questo caso, attivato dei controlli su dei potenziali errori (vedi tavola 5).

Tavola 5 – Errori interattivi con accertamento non imputati, per tipo questionario - Rilevazione annuale della produzione industriale - Anno 2005

Tipo questionario	Frequenza	%
Elettronico	2.934	27
Cartaceo	3.205	36

Inoltre, una piccolissima percentuale di errori di tipo interattivo bloccante (117 su un totale di 15.811) è stata mantenuta invariata dopo un'approfondita verifica effettuata mediante un ritorno presso le imprese e utilizzando informazioni ausiliarie, per validare dati definiti come *outlier* dalla procedura di *check*. La maggior parte hanno riguardato le quantità e/o i valori dichiarati per l'acquisto di prodotti energetici poco diffusi.

Nel caso di controlli implementati nel questionario elettronico i relativi *edit* sono attivati con una minore frequenza rispetto al questionario cartaceo. Questo si verifica, ad esempio, quando il compilatore sceglie un'unità di misura associata al prodotto diversa da quella prevista, oppure quando il rapporto tra il valore e la quantità dichiarata della produzione venduta si discosta da un prezzo medio di riferimento (vedi tavola 6). Al contrario quando invece l'impresa è impossibilitata a fornire i dati richiesti, in particolare i dati in quantità della produzione realizzata e venduta o degli acquisti di prodotti energetici, nel questionario cartaceo il revisore tende a sanare questa situazione di mancate risposte parziali e i relativi *edit* sono attivati con una minore frequenza rispetto al questionario elettronico.

Tavola 6 – Principali edit attivati per tipo questionario - Rilevazione annuale della produzione industriale - Anno 2005

Edit attivati	Tipo questionario				Totale	
	Elettronico		Cartaceo			
	Frequenza	%	Frequenza	%	Frequenza	%
Valore unitario fuori range prodotti energetici	3.252	13	2.589	15	5.841	14
Valore unitario fuori range prodotti industriali	2.297	9	2.221	13	4.518	11
Produzione totale +/- 40% componenti	1.831	7	861	5	2.692	6
Unità misura diversa da ufficiale	593	2	1.050	6	1.643	4
Quantità mancante prodotti energetici	1.132	5	179	1	1.311	3
Quantità mancante produzione venduta	1.056	4	209	1	1.265	3

13. Conclusioni

Nonostante lo sforzo profuso nella ristrutturazione della procedura di controllo e correzione esistente, un ulteriore perfezionamento può essere raggiunto intervenendo in modo mirato su alcuni aspetti critici.

In primo luogo occorre ottimizzare la gestione dei tempi e delle risorse impiegate nella fase di controllo interattivo, nell'ottica di indirizzarle verso attività più prossime alla fase di raccolta delle informazioni: attualmente, circa il 32% degli errori accertati dal sistema vengono lasciati invariati dai revisori. Un'attenta riflessione dovrà riguardare anche l'opportunità di continuare o meno ad effettuare la revisione preliminare dei modelli cartacei prime del loro invio in registrazione.

E' necessario rivedere alcuni concetti e definizioni per meglio adattarli ai sistemi contabili di impresa per eliminare alcune cause di errori sistematici. Una riflessione è attualmente in atto anche presso Eurostat per meglio definire nel regolamento Prodcom alcuni principi causa di errate interpretazioni: definizione di produzione conto terzi, esatta attribuzione territoriale della produzione realizzata dalle imprese multinazionali.

Un aumento della qualità dei microdati raccolti è possibile adottando un questionario on-line in modo da estendere ulteriormente i controlli al momento della compilazione del questionario e cercando, nel contempo, di non accrescere il carico statistico sui rispondenti. Attualmente, nel questionario telematico, in formato excel, alcuni controlli non sono stati implementati proprio per i limiti tecnici che l'ambiente di sviluppo scelto presenta.

Indicazioni su come ottimizzare ulteriormente la fase di check possono essere ottenute mediante la realizzazione, al termine di ogni ciclo produttivo di indagine, di analisi *ad hoc* sulla procedura di controllo. Questa fase di analisi, da effettuare in modo costante e ripetitivo ed opportunamente documentata, permette anche di migliorare la conoscenza del processo produttivo di indagine nel suo complesso, e quindi di intervenire su fasi che presentano aree di criticità.

Infine sarebbe opportuno sperimentare l'utilizzo di un software generalizzato, adatto alle caratteristiche peculiari che l'indagine presenta, per sfruttarne i vantaggi in termini di flessibilità e di efficienza elaborativa.

Bibliografia

Barcaroli G.,D'Aurizio L.,Luzi O.,Manzari A.,Pallara A.. *Metodi e software per il controllo e la correzione dei dati*. Roma: Istat, 1999. (Quaderni di Ricerca, n.1).

Eurostat. *Prodcom Quality Report – Anno 2004*. Lussemburgo: Eurostat, 2006. (Documento interno).

ISTAT. *Statistica annuale della produzione industriale – Anno 2003*. Roma: Istat, 2006. (Collana Informazioni n. 5).

Statistics Canada. *Quality Guidelines Statistics Canada - Fourth Edition.* Ottawa: Statistics Canada, 2003 (Catalogue no. 12-539-XIE).

Discussione: Alcune riflessioni sui metodi per il controllo e la correzione dei dati nelle indagini strutturali sulle imprese

Piero Demetrio Falorsi, *Istat, DCMT/ PSN*
Stefano Falorsi, *Istat, DCMT/ PSN*

1. Contesto di riferimento

I lavori riportati in questo volume sono stati presentati nel seminario *Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali*, tenutosi all'Istat il 25 Maggio 2007 e rivolto a illustrare alcune specifiche esperienze di controllo e correzione dei dati nell'ambito delle indagini strutturali sulle imprese condotte dall'Istat. Il seminario ha avuto la finalità di discutere e porre a confronto esperienze diverse utili alla definizione di linee guida che tengono conto di quanto effettivamente realizzato nelle indagini dell'Istituto. Le principali caratteristiche del contesto caratterizzante il processo di controllo e correzione delle indagini in parola è di seguito brevemente riassunto.

- Tutte le indagini sono sottoposte a specifici regolamenti comunitari che definiscono anche i metadati necessari a documentare la qualità del processo di trattamento dei dati. Per alcune indagini sono anche indicati i metodi e i software da utilizzare.
- I dati elementari di una data indagine possono essere validati, controllati e corretti utilizzando anche l'informazione esterna all'indagine stessa, derivante da una grande disponibilità di fonti amministrative.
- I dati da trattare sono costituiti sia variabili qualitative sia da variabili quantitative in genere fortemente asimmetriche.
- Nel corso del tempo, il risultato congiunto di diversi processi relativi al reclutamento, alla formazione e al ritiro del lavoro del personale dell'Istat ha condotto a una sostanziale riduzione del numero di persone specificatamente dedicate alle fasi di controllo e dei dati. Emerge quindi l'esigenza di utilizzare metodi che privilegiano l'intervento automatico superando i metodi tradizionali che richiedevano alta intensità di lavoro, cercando di ottimizzare le eventuali fasi di intervento manuale sui dati.

Nel seminario sono state presentate una pluralità di soluzioni diverse caratterizzate da una sostanziale omogeneità nell'articolazione delle fasi in cui è suddiviso il processo di trattamento dei dati. L'omogeneità deriva sia dal contesto comune delle indagini sia da un lavoro di ricerca condotto con continuità dagli anni 90 che ha consentito lo sviluppo di metodi e software flessibili e applicabili alle diverse realtà di indagine. Nel corso del tempo, inoltre, i ricercatori dell'Istituto, che progettano e implementano le procedure di controllo e correzione dei dati, hanno sviluppato approcci condivisi e una cultura comune sull'argomento.

La generalità delle procedure descritte prevedono un approccio complessivo tendente a minimizzare gli interventi di controllo e correzione. Tale approccio si articola in complessi processi di trattamento dei dati che in generale prevedono:

- l'uso di dati amministrativi come fonte di variabili ausiliarie per la ricostruzione di dati mancanti o errati;
- l'individuazione di outlier e di dati influenti e loro trattamento specifico, spesso fondato su interventi di esperti o sul ricontatto delle unità;
- l'imputazione massiva per i dati non influenti in genere basata (*i*) su regole di imputazione definite nell'ambito di *celle* individuate dalla concatenazione di variabili ausiliarie e (*ii*) sull'utilizzo della tecnica del donatore per le variabili qualitative;

– una specifica fase di documentazione della procedura di controllo e correzione, la cui implementazione è, sovente guidata dai dettami previsti dai *Quality Report* definiti dai Regolamenti Comunitari.

Le differenze tra le varie esperienze riguardano il modo in cui gli specifici passi sono collegati tra loro e le scelte tecniche implementate nell'ambito di ciascun passo.

Si nota infine che l'individuazione di appropriate celle di imputazione è un compito che comporta un raffinato lavoro di modellizzazione statistica che utilizza in modo integrato e unificato le metodologie di stima (fondate o assistite da modello) con quelle di controllo e correzione. L'utilizzo di celle prevede modelli differenziati per cella, ipotizzando quindi che i dati siano generati da modelli in cui gli effetti incrociati delle variabili marginali la cui concatenazione individua le celle di imputazione siano significativi. Tale approccio si scontra frequentemente con il problema della esiguità campionaria in ambito di specifiche celle. Per superare questo problema, potrebbe essere analizzata la possibilità di fondarsi su modelli che utilizzano solo gli effetti marginali e non gli effetti incrociati.

2. Alcune riflessioni di carattere generale

La progettazione e l'implementazione di procedure di controllo e correzione dei dati è un compito complesso che implica il coinvolgimento integrato di ricercatori con capacità e cognizioni di natura diversa. Una procedura ben fatta è in genere il risultato di un lavoro in cui i ricercatori con competenze tematiche interagiscono positivamente sia con ricercatori con competenze metodologiche sia con i ricercatori dell'area informatica che progettano i Sistemi informativi delle singole indagini.
L'uso di procedure di controllo e correzione complesse e variamente articolate evidenzia un problema di trasparenza e la necessità di un'attività continua di verifica e monitoraggio delle scelte effettuate.

Trasparenza

Il problema di trasparenza è in parte affrontato mediante la costruzione di indicatori di qualità del processo previsti dai Quality Report definiti dai Regolamenti Comunitari. Un'ulteriore fonte di documentazione è reperibile dalla collezione degli indicatori di qualità di processo raccolti dal Sistema informativo sulla qualità dell'Istat, comunemente noto con l'acronimo SIDI. Gli indicatori in parola, tuttavia, risolvono solo in parte il problema di trasparenza. Evidenziano una possibile caduta della qualità, ma non permettono di verificare se il trattamento dei dati introduca o meno fattori distorsivi sulle stime diffuse. Ad esempio, un alto valore di un indicatore di qualità, definito come rapporto del numero dei dati imputati sul totale dei dati raccolti, serve a segnalare una caduta della qualità della rilevazione connesso all'elevata incidenza di dati ricostruiti tramite imputazione; esso, però, non indica se i dati ricostruiti possano causare delle sovrastime o delle sottostime dei fenomeni indagati. Vi è quindi la necessità di prevedere una specifica attività volta a controllare l'effetto dei processi di trattamento dei dati. Tale attività potrebbe esplicarsi con modalità differenti; ad esempio, nell'ambito delle celle di imputazione, o di loro opportune aggregazione, è utile confrontare i valori caratteristici (media e varianza) delle distribuzioni dei dati modificati dal processo di controllo e correzione con quelli della distribuzione dei dati corretti, ossia il sottoinsieme dei dati non modificati dal processo suddetto. Alternativamente, si possono condurre indagini di controllo basate sul confronto dei valore imputati con i corrispondenti valori desunti da fonti amministrative.

Uno specifico problema di trasparenza è connesso alle usuali misure di variabilità campionaria, espressi in termini di errore standard o di coefficienti di variazione, che accompagnano le stime diffuse. Tali misure devono tenere conto che il campione osservato su una data variabile è costituito unicamente dalle unità rispondenti corrette su cui non sono stati effettuati interventi di modifica o imputazione del valore. Le misure di accuratezza tradizionalmente diffuse tendono quindi a sottostimare la variabilità dei dati. A tal fine si possono utilizzare tecniche di calcolo della varianza basate su metodi di replicazione, o una strategia di imputazione multipla. E' importante, quindi, calcolare e diffondere l'incremento di varianza dovuto al processo di imputazione.

Si nota, inoltre, che alcune delle tecniche di imputazione proposte - come quella in cui un dato errato viene sostituito dal valore mediano di una data classe di imputazione – diminuiscono artificialmente ed in modo errato la variabilità dei dati.

Monitoraggio

Il controllo e correzione è una fase del processo di un'indagine: se l'indagine è progettata male, anche la procedura più sofisticata di controllo e correzione non può fare miracoli. Vi è la necessità quindi di monitorare le procedure nel tempo al fine di realizzare uno sviluppo evolutivo della fase di progettazione delle indagini. I risultati, opportunamente esaminati del processo di controllo e correzione possono, infatti, fornire indicazioni di eventuali carenze nella progettazione dell'indagine e consentono di individuare dove concentrare le risorse per realizzare un miglioramento dell'indagine: ad esempio, un alto tasso di imputazione di una data variabile può segnalare che vi sono problemi nel questionario che deve essere opportunamente modificato.

3. Alcune riflessioni sui singoli lavori

Qui di seguito si riportano alcune considerazioni sui singoli lavori presentati.

Indagine Community Innovation Survey

Il lavoro illustra un esempio molto rilevante di indagine in cui il processo di controllo e correzione è regolamentato in ambito comunitario, anche mediante la fornitura di specifici software.
Ottima la reportistica di documentazione del processo che potrebbe permettere di migliorare notevolmente una futura ripetizione dell'indagine. Anche in questa esperienza, la variabilità potrebbe essere artificialmente ridotta: le imputazioni sono basate su un modello rapporto, senza aggiunta di una componente stocastica.

Indagine sulla Struttura delle retribuzioni

Il lavoro presenta un importante esempio di utilizzo di quasi tutte le fonti amministrative esistenti nel mondo delle imprese.
Tuttavia, la struttura gerarchica e iterativa del processo di controllo e correzione rende complessa l'identificazione e la validazione dei modelli statistici utilizzati per l'imputazione dei dati.

Indagine Strutturale sui Conti delle Imprese

Si nota il grande lavoro e i notevoli risultati raggiunti nell'integrare i dati di indagine con i dati dei Bilanci Civilistici. L'uso dei modelli INPS e INAIL (già acquisiti dall'Istat) potrebbero fornire informazioni sulla parte relativa al costo del lavoro costo. La variabilità potrebbe essere artificialmente ridotta in quanto le imputazioni da modello non prevedono una componente stocastica.

Indagine Rica Rea

Il lavoro presenta l'esempio più raffinato di procedura di controllo e correzione. Il processo descritto è ben strutturato e articolato al proprio interno. Tutti i passi implementati si basano su alcune fra le metodologie più avanzate sviluppate negli ultimi anni.

Indagine Prodcom

La procedura di controllo e correzione è molto ben strutturata e documentata. E' interessante la scelta di diversificare le procedure in base alla modalità di acquisizione dei dati. Dai dati forniti non risulta un miglioramento della qualità derivante dall'uso del questionario elettronico; questo risultato è interessante in quanto smentisce la credenza diffusa che l'uso di strumenti elettronici di raccolta dati producano sempre un miglioramento dell'accuratezza del dato raccolto.

Contributi ISTAT(*)

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys): un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat*

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi – *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo – *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni – *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 – Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo – *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

(*) ultimi cinque anni

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

5/2007 – Giulio Barcaroli e Tiziana Pellicciotti – *Strumenti per la documentazione e diffusione dei microdati d'indagine: il Microdata Management Toolkit*

6/2007 – AA.VV. – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 1ª giornata*

7/2007 – Raffaella Cianchetta, Carlo De Gregorio, Giovanni Seri e Giulio Barcaroli – *Rilevazione sulle Pubblicazioni Scientifiche Istat*

8/2007 – Emilia Arcaleni, e Barbara Baldazzi – *Vivere non insieme: approcci conoscitivi al Living Apart Together*

9/2007 – Corrado Peperoni e Francesca Tuzi – *Trattamenti monetari non pensionistici metodologia sperimentale per la stima degli assegni al nucleo familiare*

10/2007 – AA.VV – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 2ª giornata*

11/2007 – Leonello Tronti – *Il prototipo (numero 0) dell'Annuario di statistiche del Mercato del Lavoro (AML)*

12/2007 – Daniele Frongia, Raffaello Martinelli, Fernanda Panizon, Bruno Querini e Andrea Stanco – *Il nuovo Sistema informatico Altri Servizi. Progetto di reingegnerizzazione dei processi produttivi delle indagini trimestrali di fatturato degli altri servizi*

1/2008 – Carlo De Gregorio, Stefania Fatello, Rosanna Lo Conte, Stefano Mosca, Francesca Rossetti – *Sampling design and treatment of products in Istat centralised CPI surveys*

2/2008 – Mario Albisinni, Elisa Marzilli e Federica Pintaldi – *Test cognitivo e utilizzo del questionario tradotto: sperimentazioni dell'indagine sulle forze di lavoro*

3/2008 – Franco Mostacci – *Gli aggiustamenti di qualità negli indici dei prezzi al consumo in Italia: metodi, casi di studio e indicatori impliciti*

4/2008 – Daniele Frongia e Carlo Vaccari – *Introduzione al Web 2.0 per la Statistica*

5/2008 – Antonio Cortese – *La conta degli stranieri: una bella sfida per il censimento demografico del 2011*

6/2008 – Carlo De Gregorio, Carmina Munzi e Paola Zavagnini – *Problemi di stima, effetti stagionali e politiche di prezzo in alcuni servizi di alloggio complementari: alcune evidenze dalle rilevazioni centralizzate dei prezzi al consumo*

7/2008 – AA.VV. – *Seminario: metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali*