DOCUMENTI ISTAT

n. 9/2009

La procedura automatica di controllo e correzione dell'indagine SPA 2007: aggiornamenti e integrazioni

G. Guarnera, O. Luzi e M. Greco

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

ISTAT – Servizio Editoria - Centro stampa

DOCUMENTI ISTAT

n. 9/2009

La procedura automatica di controllo e correzione dell'indagine SPA 2007: aggiornamenti e integrazioni

U. Guarnera(*), O. Luzi(**) e M. Greco(***)

(*) ISTAT - Direzione Centrale per le tecnologie ed il supporto metodologico
(**) ISTAT - Direzione Centrale delle statistiche economiche strutturali, sulle imprese, commercio con l'estero e prezzi al consumo
(***) ISTAT - Direzione Centrale dei censimenti generali

Contributi e Documenti Istat 2009

Istituto Nazionale di Statistica
Servizio Editoria – Centro stampa
Via Tuscolana, 1788 - 00173

La procedura automatica di controllo e correzione dell'indagine SPA 2007: aggiornamenti e integrazioni

Guarnera Ugo, Istat, DCMT/MTS
Luzi Orietta, Istat, DCSP/B
Greco Massimo, Istat, DCCG/SCE

Sommario: L'ottimizzazione delle prestazioni di procedure complesse di controllo e correzione per indagini su larga scala può essere facilitata nel caso in cui la procedura stessa sia stata progettata e documentata in modo accurato. In questi casi, l'integrazione nelle procedure pre-esistenti di nuovi strumenti e metodologie, e l'aggiornamento del processo di controllo e correzione per tener conto di cambiamenti strutturali dell'indagine (ad esempio, dei contenuti informativi del questionario e/o della loro struttura) hanno un impatto molto contenuto in termini di tempi e costi aggiuntivi. In questo lavoro viene descritto il processo di aggiornamento della procedura di controllo e correzione dei dati dell'indagine su Struttura e Produzioni delle Aziende Agricole – Anno 2007, evidenziando come il carattere modulare della procedura pre-esistente, completamente ridisegnata in occasione dell'indagine 2003, abbia semplificato il processo di integrazione di un nuovo modulo per l'imputazione delle mancate risposte parziali di una nuova sezione del questionario, nonché gli aggiornamenti delle altre fasi a seguito di variazioni nella struttura e nei contenuti del questionario.

Parole chiave: Editing, Mancate risposte, Imputazione

Indice

1. Introduzione

Le indagini strutturali sulle imprese sono caratterizzate da elementi di complessità legati non soltanto alle diverse tipologie di errore non campionario e di mancata risposta presenti nei dati, ma anche alla mole di dati rilevati (in termini sia di unità rilevate nella popolazione, sia di variabili osservate su ciascuna unità). Per questo tipo di indagini, la progettazione di procedure di controllo e correzione (C&C) il più possibile *modulari*, in cui cioè siano integrate diverse metodologie e tecniche per le diverse classi di variabili/errori, garantisce non solo una maggiore efficienza del processo di controllo e correzione e migliori livelli qualitativi dell'informazione statistica prodotta, ma anche la possibilità di procedere ad eventuali aggiornamenti ed integrazioni della procedura stessa con maggiore possibilità di controllo, minori tempi e costi. In effetti, da una edizione all'altra di una certa indagine è usuale che possano cambiare l'organizzazione dell'indagine stessa (ad esempio, la modalità di acquisizione o la modalità di codifica dei dati, e quindi la qualità dell'informazione in input alla procedura di C&C), la struttura del questionario, nonché i suoi contenuti informativi. Orientarsi verso procedure modulari è inoltre reso sempre più opportuno dalla maggiore complessità delle indagini strutturali derivante dalla crescente disponibilità di informazioni esterne (da fonte amministrativa oppure da altre indagini), che devono essere integrate nel flusso dei dati anche ai fini della localizzazione degli errori e dell'integrazione delle mancate risposte (totali e parziali). In questa prospettiva, è allora necessario procedere verso la progettazione di procedure di C&C sempre più flessibili, appunto modulari, che possano essere aggiornate e integrate senza bisogno di ridisegnare completamente l'impianto iniziale.

In questo documento viene esaminato il caso dell'indagine su *Struttura e Produzioni delle Aziende Agricole* (*SPA*). L'indagine SPA (ISTAT, 2008 e 2009) rileva annualmente circa 400 variabili sulle aziende agricole relativamente a caratteristiche strutturali e gestionali, utilizzazione dei terreni e informazioni collegate, tipologia e consistenza degli allevamenti, struttura e consistenza della manodopera familiare e non. L'unità di rilevazione è l'azienda agricola ovvero: *l'unità tecnico-economica costituita da terreni, anche in appezzamenti non contigui, ed eventualmente da impianti ed attrezzature varie, in cui si attua la produzione agraria, forestale o zootecnica ad opera di un conduttore, cioè, persona fisica, società od ente che ne sopporta il rischio aziendale sia da solo (conduttore coltivatore e conduttore con salariati e/o compartecipanti) sia in associazione ad un mezzadro o colono parziario* (Regolamento (CE) della Commissione n. 204/2006 del 6 febbraio 2006, che adegua il Regolamento (CEE) n.571/88 del Consiglio e modifica la decisione 2000/115/CE della Commissione) e Direttiva (CE) n. 109/2001 del 19 dicembre 2001).

Svolta dall'Istat di concerto con le Regioni e Province autonome competenti per territorio, l'indagine ha subito una vasta operazione di ridisegno nell'edizione del 2003. Di conseguenza, nella stessa occasione anche la procedura di C&C è stata completamente riprogettata (Ballin *et al.*, 2004; Guarnera e Luzi, 2004). La nuova procedura era caratterizzata da una struttura sostanzialmente modulare, in cui risultavano chiaramente identificabili i vari approcci in essa integrati (grafico, automatico, interattivo, deterministico, probabilistico) per il trattamento dei vari tipi di errori e mancate risposte (errori sistematici, errori influenti, valori anomali, incompatibilità non influenti). La procedura sviluppata nel 2003 aveva poi subito nelle edizioni successive alcune modifiche e integrazioni volte sia a tener conto di alcune modifiche nei contenuti e nella struttura del questionario, sia a migliorare la qualità del trattamento degli errori e delle mancate risposte di specifiche sezioni del modello (ad esempio, attraverso la riprogettazione della fase di imputazione delle mancate risposte della sezione *Lavoro* (Guarnera *et al.*, 2006).

La specificità dell'indagine 2007 risiede nella circostanza che in questa occasione, oltre alle informazioni raccolte nelle edizioni dal 2003 in poi, sono state osservate un insieme di informazioni relative alle *Caratteristiche degli impianti ad alberi da frutto e agrumi* (*superficie coltivata, anno di impianto*, *numero di piante*) su un sottocampione di aziende, al fine di determinare il potenziale produttivo di alcune specie di alberi da frutto (*melo, pero, pesco, nettarina, albicocco, arancio, limone e agrumi a piccolo frutto*). La dimensione complessiva

del campione iniziale è di 63.922 aziende, di cui 20.744 aziende fruttifere. La raccolta dei dati è avvenuta tramite intervista diretta con un unico questionario cartaceo.

In questo documento è brevemente descritto il processo di aggiornamento e integrazione della procedura di C&C pre-esistente svolto in occasione dell'edizione 2007 dell'indagine. In particolare, sono descritte la nuova struttura della procedura e la strategia di imputazione delle mancate risposte parziali adottata per la nuova sezione del questionario sulle *Caratteristiche degli impianti ad alberi da frutto e agrumi*.

Il documento è articolato come segue. Nel paragrafo 2 è sintetizzata la struttura complessiva della procedura di C&C della SPA. Nel paragrafo 3 sono brevemente descritti i principali aggiornamenti apportati alla fase *automatica* della procedura, con riferimento sia agli algoritmi probabilistici pre-esistenti, sia all'imputazione delle mancate risposte della nuova sezione sulle *Caratteristiche degli impianti ad alberi da frutto e agrumi*. Relativamente a quest'ultima sezione, sono riportati alcuni risultati per la valutazione dell'impatto dell'imputazione su alcune statistiche univariate delle distribuzioni delle superfici delle specie di frutta rilevate.

2. Caratteristiche generali della procedura di controllo e correzione dell'indagine SPA

La procedura di C&C dell'indagine SPA (Greco *et al.*, 2008) consiste di diversi passi integrati, ciascuno dei quali destinati al trattamento di una particolare tipologia di errore e/o di variabile osservata. La procedura si articola nelle seguenti macro-fasi, schematizzate nella Figura 1:

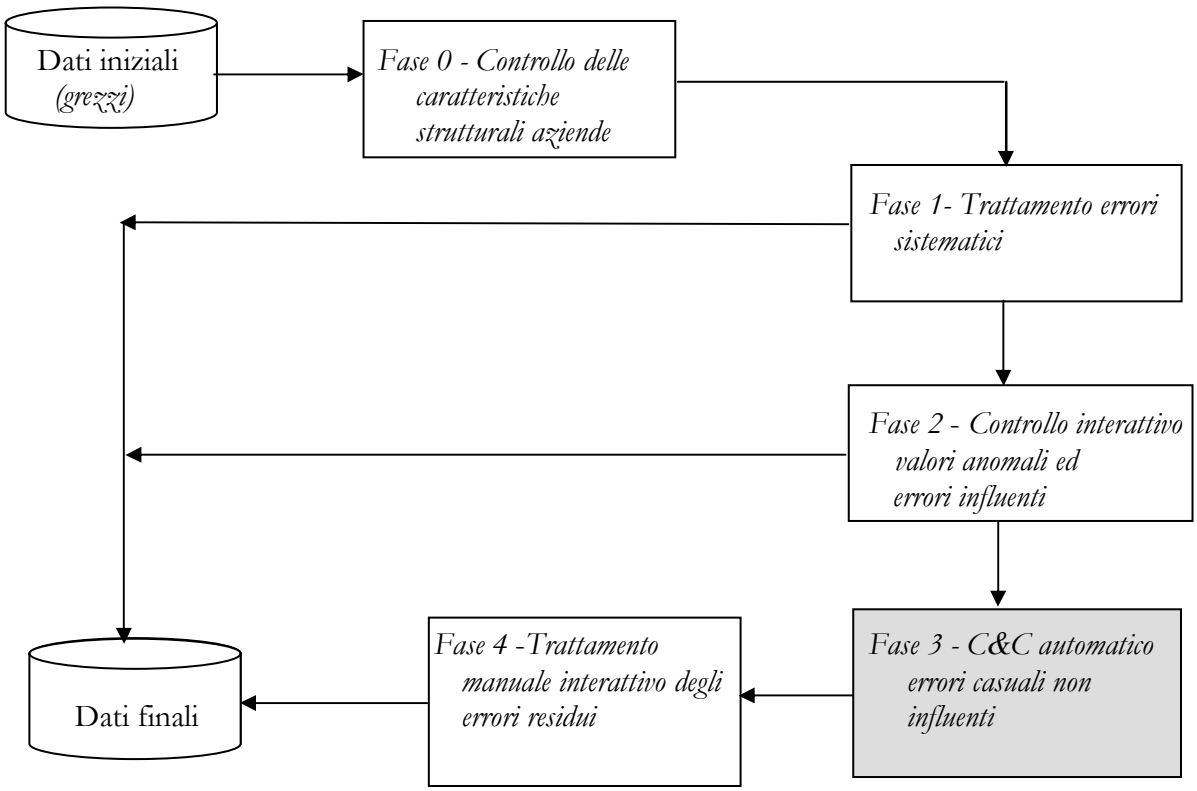
1. *Trattamento degli errori di tipo sistematico* mediante regole di compatibilità di tipo deterministico, per le varie sezioni del questionario.
2. *Controllo interattivo dei valori anomali e degli errori influenti* per le variabili principali del modello (ad es., *Totale seminativi, Totale Coltivazioni Legnose Agrarie, Superficie Agricola Utilizzata, Superficie Totale dell'Azienda,* ecc.). 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. I controlli interattivi sono effettuati in ambiente AGAIN (Benedetti et al., 2002), in cui è disponibile un supporto grafico per l'analisi e il trattamento dei dati individuati come influenti e/o anomali Per maggiori dettagli sulle metodologie adottate per queste tipologie di errore vedere Guarnera e Luzi (2005).
3. *C&C automatico degli errori casuali non influenti, inclusa l'imputazione delle mancate risposte parziali.* In questa fase, vengono adottate tecniche probabilistiche di individuazione degli errori di incompatibilità, e metodi non parametrici di imputazione delle mancate risposte parziali.
4. *Trattamento manuale-interattivo degli errori residui* delle precedenti fasi, e validazione finale dei dati.

Le fasi 1, 2 e 4, effettuate in ambiente AGAIN, sono caratterizzate da una forte componente interattiva, in cui cioè è richiesto un intervento umano diretto per l'individuazione delle situazioni (potenzialmente) errate e per la loro correzione.

Nella fase 3, al contrario, vengono utilizzate procedure completamente automatiche di individuazione e correzione degli errori, in particolare quelle disponibili nel software Banff (Statistics Canada, 2003 e 2005; Kovar *et al.*, 1988). Questa parte della procedura di C&C ha subito una serie di aggiornamenti, rispetto all'edizione 2005, in seguito alle modifiche e integrazioni dei contenuti informativi

dell'indagine. Nel seguito di questo documento, ci si limiterà alla descrizione delle innovazioni apportate a tale fase del processo.

Figura 1: *Fasi della procedura di C&C dell'indagine SPA 2007*



3. La fase di controllo e correzione automatica dell'indagine SPA 2007

La nuova fase di C&C *automatico* (*Fase 3* della Figura 1) è schematizzata nella Figura 2. Le principali innovazioni rispetto alla procedura pre-esistente sono le seguenti:

- aggiornamento e integrazione della procedura automatica di C&C e dei relativi parametri (vincoli di coerenza, pesi delle variabili, cardinalità delle soluzioni di minimo cambiamento, ecc.) in seguito alle modifiche apportate alla struttura e al contenuto informativo della sezione *Utilizzazione dei terreni* (Sezione II) e alla Sezione IV: *Pratiche agronomiche e altre notizie*, nonchè alla introduzione della nuova Sezione III: *Caratteristiche degli impianti ad alberi da frutto e agrumi*);
- sviluppo di un modulo di imputazione delle mancate risposte per le variabili della nuova Sezione III: *Caratteristiche degli impianti ad alberi da frutto e agrumi.*

La nuova procedura automatica è gerarchicamente articolata nei seguenti passi:

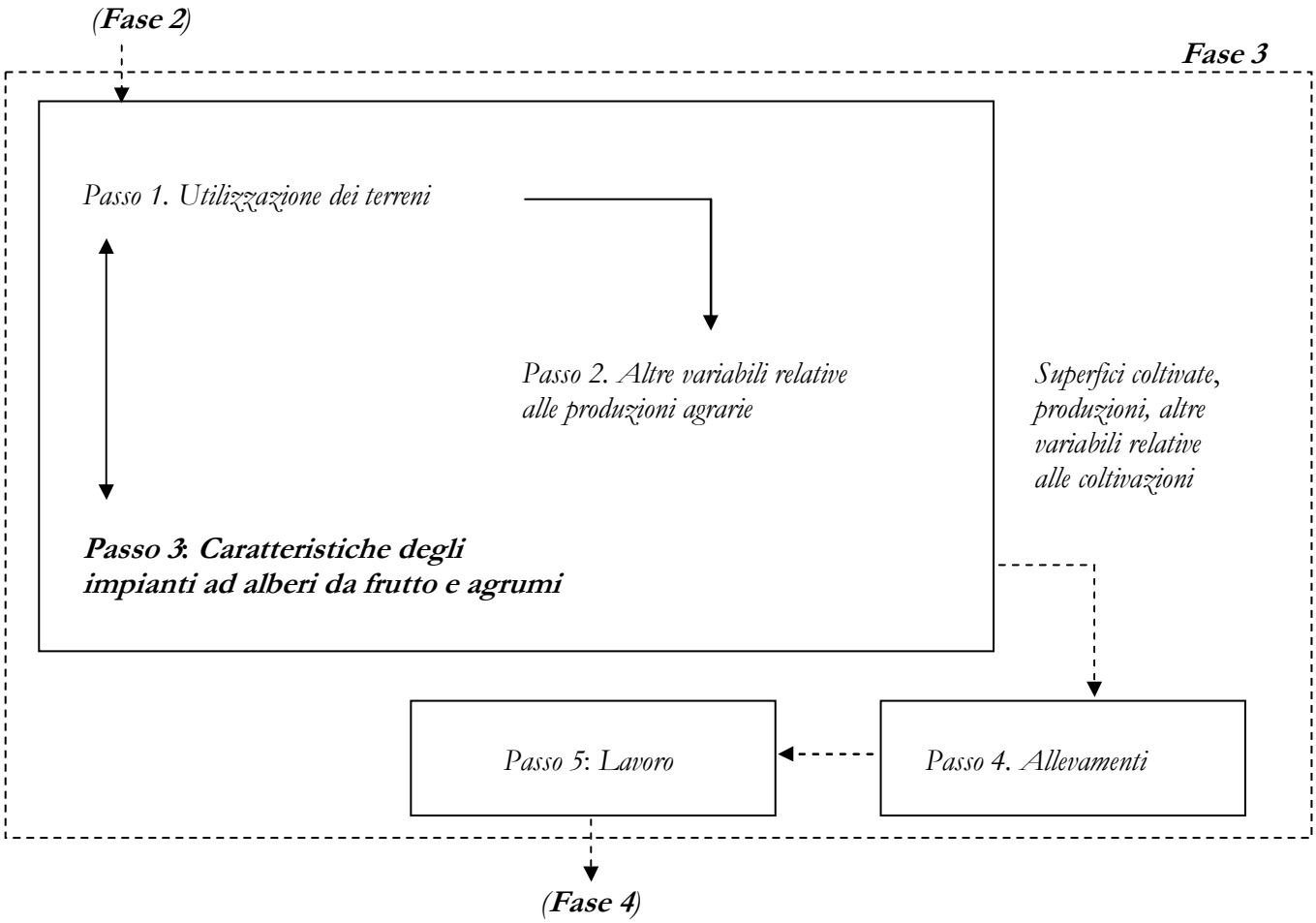
Passo 1: controllo delle relazioni fra le variabili contenute nella Sezione II (*Utilizzazione dei terreni*) e le corrispondenti variabili (totali delle superfici coltivate a frutta) della Sezione III (*Caratteristiche degli impianti ad alberi da frutto e agrumi*). Questi controlli sono evidentemente adottati solo per il sottocampione di aziende agricole produttrici di frutta.

Passo 2: controllo delle relazioni fra le variabili contenute nella sezione IV (*Pratiche agronomiche e altre notizie)* e le relazioni fra queste e le variabili della sezione II (*Utilizzazione dei terreni*). E' evidente che, in questa fase, le variabili della Sezione II vengono rese non modificabili attraverso l'assegnazione di pesi opportuni.

Passo 3: imputazione delle mancate risposte della sezione *Caratteristiche degli impianti ad alberi da frutto e agrumi.*

Passo 4: verifica delle relazioni fra le variabili contenute nella sezione V (*Allevamenti e consistenza al 1° Dicembre 2007*). Tale sezione non ha legami con altre sezioni del questionario e non ha subito variazioni rispetto alle dizioni precedenti dell'indagine, per cui i parametri della procedura di C&C (vincoli, variabili, pesi) sono rimasti invariati rispetto alla procedura pre-esistente.

Figura 2: *I passi della procedura automatica di C&C dell'indagine SPA 2007*



Passo 5: trattamento delle variabili della sezione VI: *Lavoro.* 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 (Greco et al., 2008). Tale sezione non ha legami con altre sezioni del questionario e non ha subito variazioni né di contenuto né di struttura. Le metodologie di individuazione degli errori e di imputazione delle mancate risposte, non hanno quindi subito modifiche rispetto alla precedente edizione dell'indagine.

Nei paragrafi che seguono gli aggiornamenti effettuati vengono sinteticamente illustrati separatamente per le sezioni:

- *Sezione II: Utilizzazione dei terreni* e *Sezione IV: Pratiche agronomiche e altre notizie*
- *Sezione III: Caratteristiche degli impianti ad alberi da frutto e agrumi*

3.1. Aggiornamento del trattamento delle sezioni II e IV

Per quanto riguarda le Sezioni II e IV, relative alle superfici utilizzate per tipo di coltivazione, e a tutte le informazioni ad esse collegate (*Pratiche colturali, Lavorazione del terreno, Irrigazione, Criteri di intervento fitosanitario,* ecc.), per l'individuazione degli errori è utilizzata la metodologia probabilistica di Fellegi e Holt (Fellegi, Holt, 1976), mentre per l'imputazione di errori e mancate risposte parziali vengono utilizzati i metodi deduttivo (*Deterministic Imputation*) e del donatore di distanza minima (*Nearest-Neighbour Donor, NND*) (per ulteriori dettagli vedere Greco *et al.*, 2008).

Gli aggiornamenti effettuati su queste sezioni tengono conto delle variazioni strutturali e di contenuto apportate al questionario nell'edizione 2007 rispetto all'edizione precedente dell'indagine, ed hanno perciò hanno riguardato:

- l'integrazione di nuove variabili nel piano dei controlli e l'eliminazione di variabili non più rilevate;
- l'integrazione e la messa a punto dei nuovi controlli (*edit*) e dei corrispondenti *post-edit* per fa fase di imputazione mediante NND;
- l'aggiornamento dei pesi delle variabili per l'individuazione probabilistica degli errori.

I controlli aggiuntivi sulla Sezione II hanno riguardato le relazioni con le variabili sulle produzioni di frutta osservate nella nuova Sezione III: *Caratteristiche degli impianti ad alberi da frutto e agrumi*. Inoltre, sono stati necessari aggiornamenti dovuti alla eliminazione delle informazioni relative alle Superfici Forestali per forma di gestione.

Per quanto riguarda invece la Sezione IV, gli aggiornamenti di vincoli, variabili e parametri sono state principalmente conseguenza della eliminazione di alcune variabili (ad es. quelle relative alla utilizzazione dei mezzi meccanici) e all'introduzione di nuove informazioni (ad es. quelle relative all'utilizzo di prodotti fitosanitari nell'ambito delle *Pratiche Colturali*).

In seguito alle modifiche apportate, è stato necessario:

- procedere all'ottimizzazione dei nuovi insiemi di vincoli (analisi di incongruenze e/o ridondanze;
- verificare l'eventuale presenza di nuove tipologie di errore sistematico attraverso l'analisi delle frequenze di attivazione dell'insieme ottimizzato di vincoli. A fronte dei nuovi errori sistematici individuati, sono state predisposte le opportune regole deterministiche di correzione a monte del trattamento probabilistico.

Per una descrizione dettagliata degli effetti della nuova procedura automatica di controllo e correzione, vedere ISTAT (2008), pp. 29-43.

3.2. Imputazione della Sezione III - Caratteristiche degli impianti ad alberi da frutto e agrumi

Come accennato nell'introduzione, la sezione Caratteristiche degli impianti ad alberi da frutto e agrumi è osservata solo sul sottocampione di aziende produttrici di frutta. In questa sezione vengono rilevate

informazioni su *Anno di impianto, Superficie investita, Numero di piante* per tutte le *varietà* coltivate di 8 tipologie di frutta (*specie*): *Melo, Pero, Pesco, Nettarina, Albicocco, Arancio, Limone, Agrumi e piccoli frutti.*

Per il controllo dei dati di questa sezione, è stata messa a punto una procedura mista in cui, oltre ai controlli di coerenza interni, effettuati e risolti o deterministicamente, oppure in modo interattivo in ambiente AGAIN, è stato previsto un passo di controllo e di imputazione automatica.

Per quanto riguarda il controllo, è stato sfruttato il legame fra le superfici investite per le varie *specie* di frutta (somma delle superfici delle relative *varietà*) risultanti dalla Sezione III, e le corrispondenti superfici coltivate a quel tipo di frutta indicate nella sezione relativa all'utilizzazione dei terreni.

Per quanto riguarda invece *l'imputazione delle mancate risposte parziali*, i dati osservati presentano un solo *pattern* di mancata risposta, quello relativo al caso di aziende in cui, pur essendo riportati i totali di alcune specie di frutta, risultano mancanti tutte le informazioni relative alle corrispondenti varietà (in particolare, *superficie, anni di impianto, numero di piante*). In totale, sono 144 le aziende affette da questo tipo di mancata risposta (1%). Per l'imputazione di questa casistica, è stato predisposto un metodo con donatore di distanza minima in cui, una volta individuata opportunamente un'azienda donatrice, i totali osservati nell'azienda ricevente vengono ri-proporzionati sulla base delle varietà e delle relative proporzioni osservate nel donatore. Il metodo è descritto più in dettaglio nel paragrafo seguente.

3.2.1 Imputazione delle mancate risposte

I dati della Sezione III: *Caratteristiche degli impianti ad alberi da frutto e agrumi* sono organizzati in modo tale che ciascun record è univocamente determinato da un codice azienda, un codice varietà e l'anno di impianto della varietà, oppure da un codice azienda e da un codice di "specie". Nel primo caso le variabili riportate sono, oltre a quelle già menzionate, la superficie e il numero di piante relative all'impianto della varietà nell'anno riportato; nel secondo caso, la sola superficie totale dedicata alla coltivazione della specie (somma delle superfici relative alle singole varietà).

In un certo numero di casi (144 aziende) del sottocampione dei fruttiferi sono risultate disponibili solo le *superfici totali* di specie, non essendo state riportate le informazioni relative alle singole varietà (*codice varietà, anno di impianto, superficie, numero piante*). Per queste aziende si è posto dunque il problema di ricostruire, per ogni specie presente nell'azienda, sia i codici delle diverse varietà della specie, sia tutte le caratteristiche di impianto. L'approccio adottato a questo scopo si è basato sull'utilizzo del metodo del donatore stratificato. Più in dettaglio, nel caso in cui, per una certa azienda (azienda *ricevente*) fosse riportato solo la superficie totale di una determinata specie *S*, si è proceduto con i seguenti passi:

1. Individuazione delle aziende in cui è praticata la coltivazione della specie in questione e sono disponibili le informazioni relative alle diverse varietà

2. Selezione, all'interno dell'insieme di aziende di cui al punto 1), del sottoinsieme di aziende (serbatoio di *donatori*) appartenenti alla stessa Regione della azienda ricevente e alla stessa classe dimensionale "relativamente alla specie S". L'ultima caratteristica è stata determinata mediante i quartili della distribuzione della superficie coltivata con la specie S (considerando esclusivamente le aziende in cui tale superficie è risultata maggiore di zero). Per ogni specie dunque, le aziende sono state divise in quattro classi dimensionali.

3. Selezione casuale di un'azienda donatrice all'interno del serbatoio definito ai punti 1) e 2).

4. Attribuzione all'azienda ricevente delle seguenti caratteristiche dell'azienda donatrice:
   - codici di varietà
   - anni di impianto

- per ciascun anno di impianto, rapporti tra le superfici relative a ciascuna varietà e superficie totale coltivata con la specie in questione
- per ciascun anno di impianto, numero di piante per unità di superficie

5. Ricostruzione delle superfici mediante moltiplicazioni dei rapporti imputati per il totale della superficie di specie.

6. Ricostruzione del numero piante mediante moltiplicazione del numero di piante per unità di superficie per la superficie ricostruita.

Da quanto esposto sopra risulta chiaro che per una stessa azienda possono essere usati, in generale, diversi donatori. Infatti le procedure di imputazione relative a diverse specie sono svolte indipendentemente. E' opportuno notare che, se da una parte questo approccio penalizza la preservazione di eventuali relazioni di associazioni tra le variabili che si riferiscono a varietà di diverse specie, dall'altra garantisce la disponibilità di serbatoi di donatori di ampiezza adeguata.

3.2.2 Effetto dell'imputazione delle mancate risposte per la sezione III: alcuni risultati

In questo paragrafo riportiamo alcuni risultati relativi all'imputazione delle varietà relative alle specie mancanti nelle 144 aziende identificate come produttrici di frutta ma non rispondenti alla Sezione III.

Nella tabella 1 è indicato il numero e la frequenza di imputazioni per specie di frutta imputata. Come si può vedere, il numero massimo di imputazioni (45 aziende) ha riguardato la specie '699' = *Arancio*, mentre la specie che ha subito il minor numero di imputazioni è la '499' = *Nettarina* (6 aziende). In totale, sono stato imputati 173 nuovi valori (totali di specie).

Tabella 1: *Numero e frequenza di aziende imputate per tipo di specie imputata*

Specie	Frequenza	%	Freq. Cumul .	% cumulata
Melo	13	7.51	13	7.51
Pero	27	15.61	40	23.12
Pesco	29	16.76	69	39.88
Nettarina	6	3.47	75	43.35
Albicocco	19	10.98	94	54.34
Arancio	45	26.01	139	80.35
Limone	14	8.09	153	88.44
Agrumi a *piccoli frutti*	20	11.56	173	100.00

Al fine di valutare l'impatto delle imputazioni effettuate sulle principali statistiche univariate delle distribuzioni delle varie specie, sono state calcolate le variazioni fra valori di totale (*var_T*), media (*var_M*), mediana (*var_Med*), e deviazione standard (*var_STD*) calcolate prima e dopo l'imputazione nelle 144 aziende dei 173 valori mancanti. Tali variazioni sono ottenute mediante i seguenti indicatori:

$$var\_T(specie) = \frac{T_p(specie) - T_a(specie)}{T_p(specie)} \times 100 \qquad [1]$$

dove:

- $T_a(specie) = \sum_{i=1}^{nspecie(ante)} superficie_i(ante)$

- $T_p(specie) = \sum_{i=1}^{nspecie(post)} superficie_i(post)$

- ***nspecie(ante)*** e ***nspecie(post)*** sono rispettivamente il numero di casi per una certa specie di frutta disponibili prima e dopo il passo di imputazione.

Le variazioni relative alle altre statistiche considerate sono state calcolate in modo del tutto analogo:

$$var\_M(specie) = \frac{Media_p(specie) - Media_a(specie)}{Media_p(specie)} \times 100 \quad [2]$$

$$var\_STD(specie) = \frac{DevStandard_p(specie) - DevStandard_a(specie)}{DevStandard_p(specie)} \times 100 \quad [3]$$

$$var\_Med(specie) = \frac{Mediana_p(specie) - Mediana_a(specie)}{Mediana_p(specie)} \times 100 \quad [4]$$

Nella tabella 2 sono indicati valori degli indicatori precedenti ottenuti confrontando le distribuzioni osservata e imputata per ciascuna specie di frutta.

Come si può osservare, l'imputazione delle mancate risposte ha avuto un impatto molto contenuto su tutte le statistiche considerate per tutte le varietà soggette a imputazione. Per quanto riguarda il *Totale di superficie coltivata*, le variazioni (percentuali) maggiori in valore assoluto si sono verificate per le specie *Agrumi e piccoli frutti* (+1,76%) e *Arancio* (+1,39%); le stesse due specie hanno subito le variazioni (percentuali) maggiori in valore assoluto anche in termini di *Superficie Media* coltivata; e di *Deviazione standard*; in termini di *Mediana* delle distribuzioni, solo le specie *Pero* e *Pesco* hanno subito variazioni diverse da zero.

Tabella 2: *Variazioni fra i valori di totale, media, deviazione standard e mediana nelle distribuzioni non ponderate delle specie di frutta prima e dopo il passo di imputazione*

Specie	var_T	var_M	var_STD	var_Med
Melo	+0,058	+0,058	-0,007	0
Pero	+0,79	+0,79	+0,16	+1,84
Pesco	+0,55	+0,55	+0,15	+1,43
Nettarina	+0,10	+0,11	-0,014	0
Albicocco	+0,14	+0,14	-0,029	0
Arancio	+1,39	+1,39	+0,53	0
Limone	+0,68	+0,68	+0,05	0
Agrumi a piccoli frutti	+1,76	+1,76	+0,69	0

Nella tabella 3 sono invece riportate le variazioni ante- e post-imputazione per le stesse statistiche univariate calcolate però sulle distribuzioni ponderate ante e post imputazione delle varie specie di frutta.

Anche in questo caso, le variazioni assolute non superano mai l'1,5% per nessuna delle specie trattate su nessuna delle statistiche univariate considerate. Tuttavia, si può notare aumento dell'entità dell'impatto del processo di imputazione sulle distribuzioni ponderate di *Pero* e *Pesco*, con riferimento alle superfici sia *Totale* sia *Media*, nonché alla *Deviazione standard* (passata quest'ultima da +0,16% a +0,69% per il *Pero*, e da +0,5% a +0,89% per il *Pesco*). Un incremento di impatto è osservabile anche sulla variazione della *Deviazione Standard* sulla distribuzione ponderata del *Limone*, passata da +0,05% sulla distribuzione non ponderata a +0,168% su quella ponderata.

Tabella 3: *Variazioni fra i valori di totale, media, deviazione standard e mediana nelle distribuzioni ponderate delle specie di frutta prima e dopo il passo di imputazione*

Specie	var_T	var_M	var_STD	var_Med
Melo	+0,047	+0,047	-0,010	0
Pero	+1,093	+1,093	+0,699	0
Pesco	+1,158	+1,158	+0,889	+1,562
Nettarina	+0,151	+0,151	-0,004	0
Albicocco	+0,295	+0,295	-0,041	0
Arancio	+0,778	+0,778	+0,441	+1,538
Limone	+0,606	+0,606	+0,168	0
Agrumi a piccoli frutti	+1,348	+1,348	+0,733	0

Complessivamente, anche a causa della bassa frequenza percentuale di valori imputati, la ricostruzione dei valori mancanti della sezione *Caratteristiche degli impianti ad alberi da frutto e agrumi* non sembra aver sostanzialmente alterato il contenuto informativo statistico dei dati a livello aggregato, né per quanto riguarda le distribuzioni non ponderate, né quelle ponderate delle superfici coltivate a frutta.

4. Conclusioni

Il processo di controllo e correzione dei dati di un'indagine statistica economica su larga scala è caratterizzato da elementi di complessità derivanti non solo dal tipo degli errori non campionari presenti nei dati, ma anche dalla mole di informazione da trattare (in termini sia di unità campionate, sia di variabili osservate). Per quanto riguarda la fase di controllo e correzione dei dati, la necessità di tener conto in modo integrato di questi diversi elementi, di disegnare una procedura efficiente in termini di tempi e costi impiegati, e di garantire risultati qualitativamente soddisfacenti, obbliga a rivedere, aggiornare, monitorare e ottimizzare in modo continuo le procedure pre-esistenti di individuazione e di trattamento degli errori e quelle di ricostruzione dei valori mancanti, sfruttando le metodologie e le tecnologie più appropriate. Questo processo di aggiornamento e ottimizzazione è largamente facilitato in presenza di procedure aventi struttura modulare, in cui cioè i vari passi del processo di controllo e

correzione relativi alle diverse tipologie di errore e/o di variabili e/o alle diverse parti del questionario siano implementati in procedure separate strutturate in un flusso integrato di processi e dati.

La procedura di controllo e correzione dell'indagine su *Struttura e Produzioni delle Aziende Agricole* è un ottimo esempio di procedura complessa per indagini economiche strutturali su larga scala in cui, per le diverse sezioni del questionario, sono integrate modularmente metodi e tecniche per il trattamento delle diverse tipologie di errore e di mancate risposte raccomandati a livello internazionale (Luzi *et al.*, 2008). L'indagine è inoltre caratterizzata da revisioni periodiche e miglioramenti nella struttura e nei contenuti informativi del questionario, che rendono necessaria la revisione e l'aggiornamento della procedura di controllo e correzione dei dati ad ogni ciclo di indagine.

D'altra parte, nel caso dell'indagine SPA, si ha un esempio avanzato anche dal punto di vista dell'integrazione del processo di controllo e correzione con le altre fasi dell'indagine, a partire dall'acquisizione controllata delle informazioni (che garantisce una elevata qualità dei dati in input), per finire con il processo di stima, che è agganciato ai risultati del processo di controllo e correzione attraverso le fasi di *editing selettivo* e di *macroediting* rispettivamente a monte e a valle del trattamento degli errori e delle mancate risposte. L'accuratezza e completezza della documentazione e valutazione delle caratteristiche della procedura e dei suoi effetti sui dati dell'indagine (ISTAT, 2008) sono ulteriori elementi di qualità che caratterizzano l'indagine.

Alla luce di queste considerazioni, si può concludere che la procedura finale dell'indagine 2007, con la inclusione della sezione sulle *Caratteristiche degli impianti ad alberi da frutto e agrumi*, può essere considerato uno degli esempi più avanzati di procedure modulari di controllo e correzione in Istat, da un punto di vista sia tecnologico, sia metodologico, sia operativo.

Bibliografia

Ballin M., Guarnera U., Luzi O., Salvi S.. New Methodologies and tools for Dealing with Non-Sampling Errors in the Istat Survey on Structure and Production of Agricultural Firms. *Atti del Convegno Metodi d'Indagine e di Analisi per le Politiche Agricole - MIAPA 2004*, Università di Pisa, 21-22 Ottobre 2004.

Benedetti R., Espa G., Piersimoni F. Available methods, techniques and software for survey data editing, *Conference on Agricultural and Enviromental Statistical applications in Rome*, Roma, Giugno 2001. 3, 631-644, 2002.

Chen J., Shao J. Nearest Neighbour Imputation for Survey Data, *Journal of Official Statistics*, 16, 113-131, 2000.

Fellegi I.P., 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 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 Nazionale l'Informazione Statistica e le Politiche Agricole - ISPA 2004*, Università di Cassino, 6 Maggio 2004.

Guarnera U., Luzi O. Indagine Struttura e Produzioni delle Aziende Agricole: la nuova procedura di Controllo e Correzione Automatica per le Variabili su Superfici Aziendali e Consistenza degli Allevamenti, *Documenti Istat*,n.8/2006.

Greco M., Guarnera U., Luzi O. Le nuove procedure di controllo e correzione delle indagini agricoltura SPA e RICA-REA, *Contributi Istat*, n.7/2008.

ISTAT. *Rapporto sulla Qualità - Indagine sulla struttura e produzione delle aziende agricole. Anno 2007*, 2008.

ISTAT. *Struttura e produzioni delle aziende agricole*. http://www.istat.it/dati/dataset/20090120_01/. 2009.

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.

Statistics Canada. *Banff - Functional Description of the Banff System for Editing and Imputation*, Version 1.02, Generalized Systems Methods Section, Business Survey Methods Division, December 2003.

Statistics Canada. *Banff Users Guide*, Version 1.04, Generalized Systems Methods Section, Business Survey Methods Division, 2005.

Kalton G., Kasprzyk D. Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 22-31, 1982.

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

Documenti ISTAT(*)

1/2005 – Francesco Cuccia, Simone De Angelis, Antonio Laureti Palma, Stefania Macchia, Simona Mastroluca e Domenico Perrone – *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*

2/2005 – Marina Peci – *La statistica per i Comuni: sviluppo e prospettive del progetto Sisco.T (Servizio Informativo Statistico Comunale. Tavole)*

3/2005 – Massimiliano Renzetti e Annamaria Urbano – *Sistema Informativo sulla Giustizia: strumenti di gestione e manutenzione*

4/2005 – Marco Broccoli, Roberto Di Giuseppe e Daniela Pagliuca – *Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo Microstrat di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat*

5/2005 – Mauro Albani e Francesca Pagliara – *La ristrutturazione della rilevazione Istat sulla criminalità minorile*

6/2005 – Francesco Altarocca e Gaetano Sberno – *Progettazione e sviluppo di un "Catalogo dei File Grezzi con meta-dati di base" (CFG) in tecnologia Web*

7/2005 – Salvatore F. Allegra e Barbara Baldazzi – *Data editing and quality of daily diaries in the Italian Time Use Survey*

8/2005 – Alessandra Capobianchi – *Alcune esperienze in ambito internazionale per l'accesso ai dati elementari*

9/2005 – Francesco Rizzo, Laura Vignola, Dario Camol e Mauro Bianchi – *Il progetto "banca dati della diffusione congiunturale"*

10/2005 – Ennio Fortunato e Nadia Mignolli – *I sistemi informativi Istat per la diffusione via web*

11/2005 – Ennio Fortunato e Nadia Mignolli – *Sistemi di indicatori per l'attività di governo: l'offerta informativa dell'Istat*

12/2005 – Carlo De Gregorio e Stefania Fatello – *L'indice dei prezzi al consumo dei testi scolastici nel 2004*

13/2005 – Francesco Rizzo e Laura Vignola – *RSS: uno standard per diffondere informazioni*

14/2005 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Launching and implementing the job vacancy statistics*

15/2005 – Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni e Leonardo Tininini – *La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT*

16/2005 – Ennio Fortunato e Nadia Mignolli – *Verso il Sistema di Indicatori Territoriali: rilevazione e analisi della produzione Istat*

17/2005 – Raffaella Cianchetta e Daniela Pagliuca – *Soluzioni Open Source per il software generalizzato in Istat: il caso di PHPSurveyor*

18/2005 – Gianluca Giuliani e Barbara Boschetto – *Gli indicatori di qualità dell'Indagine continua sulle Forze di Lavoro dell'Istat*

19/2005 – Rossana Balestrino, Franco Garritano, Carlo Cipriano e Luciano Fanfoni – *Metodi e aspetti tecnologici di raccolta dei dati sulle imprese*

1/2006 – Roberta Roncati – *www.istat.it (versione 3.0) Il nuovo piano di navigazione*

2/2006 – Maura Seri e Annamaria Urbano – *Sistema Informativo Territoriale sulla Giustizia: la sezione sui confronti internazionali*

3/2006 – Giovanna Brancato, Riccardo Carbini e Concetta Pellegrini – *SIQual: il sistema informativo sulla qualità per gli utenti esterni*

4/2006 – Concetta Pellegrini – *Soluzioni tecnologiche a supporto dello sviluppo di sistemi informativi sulla qualità: l'esperienza SIDI*

5/2006 – Maurizio Lucarelli – *Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica*

6/2006 – Natale Renato Fazio – *La ricostruzione storica delle statistiche del commercio con l'estero per gli anni 1970-1990*

7/2006 – Emilia D'Acunto – *L'evoluzione delle statistiche ufficiali sugli indici dei prezzi al consumo*

8/2006 – Ugo Guarnera, Orietta Luzi e Stefano Salvi – *Indagine struttura e produzioni delle aziende agricole: la nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti*

9/2006 – Maurizio Lucarelli – *La regionalizzazione del Laboratorio ADELE: un'ipotesi di sistema distribuito per l'accesso ai dati elementari*

10/2006 – Alessandra Bugio, Claudia De Vitiis, Stefano Falorsi, Lidia Gargiulo, Emilio Gianicolo e Alessandro Pallara – *La stima di indicatori per domini sub-regionali con i dati dell'indagine: condizioni di salute e ricorso ai servizi sanitari*

11/2006 – Sonia Vittozzi, Paola Giacchè, Achille Zuchegna, Piero Crivelli, Patrizia Collesi, Valerio Tiberi, Alexia Sasso, Maurizio Bonsignori, Giuseppe Stassi e Giovanni A. Barbieri – *Progetto di articolazione della produzione editoriale in collane e settori*

12/2006 – Alessandra Coli, Francesca Tartamella, Giuseppe Sacco, Ivan Faiella, Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Isabella Siciliani, Sara Colombini e Alessandra Masi – *La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*

13/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Intrastat*

14/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Extrastat*

15/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: comparazione tra rilevazione Intrastat ed Extrastat*

16/2006 – Fabio M. Rapiti – *Short term statistics quality Reporting: the LCI National Quality Report 2004*

17/2006 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zeli – *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI*

18/2006 – Mauro Albani – *La nuova procedura per il trattamento dei dati dell'indagine Istat sulla criminalità*

19/2006 – Alessandra Capobianchi – *Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa*

20/2006 – Francesco Altarocca – *Gli strumenti informatici nella raccolta dei dati di indagini statistiche: il caso della Rilevazione sperimentale delle tecnologie informatiche e della comunicazione nelle Pubbliche Amministrazioni locali*

1/2007 – Giuseppe Stassi – *La politica editoriale dell'Istat nel periodo 1996-2004: collane, settori, modalità di diffusione*

2/2007 – Daniela Ichim – *Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment*

3/2007 – Ugo Guarnera, Orietta Luzi e Irene Tommasi – *La nuova procedura di controllo e correzione degli errori e delle mancate risposte parziali nell'indagine sui Risultati Economici delle Aziende Agricole (REA)*

* ultimi cinque anni

4/2007 – Vincenzo Spinelli – *Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770*

5/2007 – Anna Di Carlo, Maria Picci, Laura Posta, Michaela Raffone, Giuseppe Stassi e Fiorella Tortora – *La progettazione dei Censimenti generali 2010-2011: 1 - Analisi, valutazione e proposte in merito ad atti di normazione e finanziamento*

6/2007 – Silvia Bruzzone, Atonia Manzari, Marilena Pappagallo e Alessandra Reale – *Indagine sulle Cause di Morte: Nuova procedura automatica per il controllo e la correzione delle variabili demo-sociali*

7/2007 – Maura Giacummo, Carlo Vaccari e Monica Scannapieco – *Indagine sulle Scelte Tecnologiche degli Istituti Nazionali di Statistica*

8/2007 – Lamberto Pizzicannella – *Sviluppo del processo di acquisizione e trattamento informatico degli archivi relativi al modello di dichiarazione 770. Anni 2004 – 2005*

9/2007 – Damiano Abbatini, Lorenzo Cassata, Fabrizio Martire, Alessandra Reale, Giuseppina Ruocco e Donatella Zindato – *La progettazione dei Censimenti generali 2010-2011 2 - Analisi comparativa di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani*

10/2007 – Marco Fortini, Gerardo Gallo, Evelina Paluzzi, Alessandra Reale e Angela Silvestrini – *La progettazione dei censimenti generali 2010 – 2011 3 – Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento*

11/2007 – Domenico Adamo, Damiana Cardoni, Valeria Greco, Silvia Montecolle, Sante Orsini, Alessandro Ortensi e Miria Savioli – *Strategie di correzione del questionario sulla qualità della vita dell'infanzia e dell'adolescenza. Indagine multiscopo sulle famiglie. Aspetti della vita quotidiana 2005*

12/2007 – Carlo Nappi – *Manuale per la preparazione di originali "ready to print"*

1/2008 – Franco Lorenzini – *Indagine sulle unità locali delle imprese: la flessibilità organizzativa e il ruolo degli uffici regionali come strategia per la riduzione del disturbo statistico e il raggiungimento di elevati tassi di risposta*

2/2008 – Elisa Berntsen, Simone De Angelis, Simona Mastroluca – *La progettazione dei Censimenti generali 2010-2011 4-L'uso dei dati censuari del 2000-2001: alcune evidenze empiriche*

3/2008 – Marina Peci – *Progetto SCQ -Scuola Conoscenza Qualità-Statistica e Studenti*

4/2008 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zel – *Messa a regime dell'uso dei dati fiscali (Modelli UNICO) per l'integrazione delle mancate risposte e la riduzione del numero delle unità campione della rilevazione PMI*

5/2008 – Giovanni Seri e Maurizio Lucarelli – *A.D.ELE. Il laboratorio per l'Analisi dei Dati ELEmentari. Monitoraggio dell'attività Anni 2004-2007*

6/2008 – Francesco Altarocca – *Strumenti informatici innovativi nella conduzione di indagini statistiche*

1/2009 – Silvia Dardanelli, Simona Mastroluca, Alessandro Sasso e Mariangela Verrascina – *La progettazione dei censimenti generali 2010 – 2011 5 - Novità di regolamentazione internazionale per il 15° Censimento generale della popolazione e delle abitazioni*

2/2009 – Rossana Balestrino e Alberto Gaucci – *Tecniche di cattura dati nei processi di produzione statistica*

3/2009 – Barbara Fiocco – *Le "misure" dell'Italia nell'Annuario Statistico Italiano*

4/2009 – Daniela Pagliuca, Raffaella Cianchetta, Marco Broccoli, Teresa Buglielli, Roberto Di Giuseppe e Diego Zardetto – *L'Osservatorio Tecnologico per i Software generalizzati (OTS) nel 2008*

5/2009 – Silvia Losco – *Il riuso informatico nelle Pubbliche Amministrazioni: normativa e prime esperienze in Istat*

6/2009 – Fabio Crescenzi Marco Fortini, Gerardo Gallo e Andrea Mancini – *La progettazione dei censimento generali 2010 – 2011 6 - Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione*

7/2009 – Silvia Losco – *Gli standard informatici dell'Istat*

8/2009 – Alfredo Roncaccia e Roberto Iannaccone – *L'indagine sulle Opere Pubbliche dalla costituzione dell'Istituto Centrale di Statistica ai giorni nostri*

9/2009 – Ugo Guarnera, Orietta Luzi e Massimo Greco – *La procedura automatica di controllo e correzione dell'indagine SPA 2007: aggiornamenti e integrazioni*