CONTRIBUTI ISTAT

n. 1/2009

Valutazione dell'idoneità del software DIESIS all'individuazione dei valori errati in variabili quantitative

G. Bianchi, A. Manzari, A. Reale e S. Salvi

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 ISTAT

n. 1/2009

Valutazione dell'idoneità del software DIESIS all'individuazione dei valori errati in variabili quantitative

G. Bianchi(*), A. Manzari(**), A. Reale(*) e S. Salvi(***)

(*) ISTAT Servizio Metodi, tecniche e organizzazione dei censimenti
(**) ISTAT Servizio Metodologie, tecnologie e software per la produzione dell'informazione statistica
(***) ISTAT Servizio Statistiche sull'agricoltura

Contributi e Documenti Istat 2009

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

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

Sommario

Nell'ambito dei lavori del Comitato d'indirizzo per la progettazione di sistemi informatici e tecnologici per i censimenti del 2010 e 2011 è emerso l'interesse a valutare la possibilità di utilizzare il Sistema DIESIS per il 6° Censimento generale dell'agricoltura, caratterizzato dalla prevalenza di variabili quantitative. A tal fine, si è proceduto a realizzare un primo studio di valutazione dell'efficacia ed efficienza del Sistema DIESIS nella fase di localizzazione degli errori (casuali non influenti) in variabili quantitative. Lo studio è stato condotto nell'ambito delle attività svolte dal Gruppo di lavoro avente il compito di definire il piano di campionamento e l'impianto metodologico del sistema di controllo e correzione dei dati censuari relativamente al 6° Censimento generale dell'agricoltura.

Il presente documento descrive la sperimentazione effettuata sui dati dell'indagine su Struttura e Produzione delle Aziende Agricole (SPA) anno 2005 e riporta i risultati conseguiti.

INDICE

1 Introduzione

L'attuale politica informatica dell'Istituto persegue la riduzione della dipendenza dal sistema SAS (anche a seguito delle recenti richieste del Centro Nazionale per l'Informatica nella Pubblica Amministrazione – CNIPA) e il riuso del software disponibile (in modo tale da evitare duplicazioni di codice e quindi dispendio di risorse). Pertanto, è di primario interesse per l'Istituto poter disporre di versioni di software generalizzati (non SAS) da utilizzare nelle attività di produzione dei dati.

Relativamente al problema della localizzazione e correzione degli errori casuali non influenti in variabili quantitative, l'unico software generalizzato attualmente disponibile in Istat è il software Banff (Kovar *et al*., 1988; Cotton, 1991). Tale software è stato ideato e sviluppato da Statistics Canada in ambiente SAS ed è strutturato secondo la filosofia SAS delle procedure (*Proc*). Banff è uno dei software raccomandati ed utilizzati dalla comunità internazionale di data-editing per il trattamento degli errori casuali non influenti nelle variabili quantitative.

La localizzazione e correzione degli errori nelle variabili anagrafiche e di stato civile del Censimento della Popolazione Italiana 2001 è stata effettuata con il Sistema DIESIS (Data Imputation Editing System - Italian Software) (Bruni *et al*., 2001; Bianchi *et al.*, 2005b). DIESIS è stato ideato e sviluppato in linguaggio C++ nell'ambito di una collaborazione scientifica tra la Direzione dei Censimenti (Istat) e il Dipartimento di Informatica e Sistemistica dell'Università degli Studi di Roma "La Sapienza". Ad oggi DIESIS è stato testato (con ottimi risultati, cfr. Manzari and Reale, 2001) ed utilizzato esclusivamente per il trattamento delle variabili anagrafiche e di stato civile rilevate al Censimento della popolazione e delle abitazioni, ma gli algoritmi in esso implementati sono teoricamente idonei al trattamento singolo o congiunto di variabili qualitative e quantitative.

Il perseguimento degli obiettivi di politica informatica sopra citati suggerisce di valutare quanto prima l'idoneità del Sistema DIESIS al trattamento delle variabili quantitative. A tal fine, si è proceduto a realizzare un primo studio di valutazione dell'efficacia ed efficienza del Sistema DIESIS nella fase di localizzazione degli errori casuali non influenti in variabili quantitative. Lo studio è stato condotto utilizzando i dati grezzi relativi alle coltivazioni dell'indagine su *Struttura e Produzione delle Aziende Agricole (SPA)* anno 2005. Il presente documento descrive la sperimentazione effettuata e riporta i risultati conseguiti.

2 La sperimentazione DIESIS/Banff

2.1 Obiettivi

L'idoneità del Sistema DIESIS alla localizzazione degli errori casuali non influenti in variabili quantitative è stata valutata mediante uno studio comparativo: i processi di individuazione dei valori errati di DIESIS e di Banff sono stati applicati ad uno stesso insieme di dati quantitativi grezzi realmente osservati (e quindi contenenti errori); successivamente sono stati confrontati i risultati ottenuti dalle due applicazioni (indipendenti). In pratica, il software Banff, essendo riconosciuto dalla comunità internazionale come uno dei migliori software esistenti per la localizzazione degli errori casuali non influenti in variabili quantitative, è stato utilizzato come riferimento o "*gold standard*" per la valutazione del Sistema DIESIS.

L'obiettivo dello studio di valutazione descritto in questo documento è quello di misurare e porre a confronto alcune caratteristiche dei processi di localizzazione degli errori che determinano la capacità di soddisfare specifiche esigenze dell'Istituto (utente del processo).

Un'esigenza prioritaria dell'Istituto, connessa al requisito di *accuratezza* dei dati statistici, è quella di disporre di una procedura di localizzazione degli errori che sia in grado di rimuovere i valori errati salvaguardando il più possibile l'informazione raccolta. Un modo per soddisfare questa esigenza è quello di utilizzare una procedura basata sul principio del *minimo cambiamento* (Fellegi & Holt, 1976). Tale principio consiste nell'identificare, per ciascun record errato, il minimo numero di campi (variabili) da modificare (imputare) affinché il record risultante soddisfi tutte le regole (*soluzione di cardinalità minima*). In generale, il principio del minimo cambiamento è considerato un criterio ottimale per risolvere il problema della localizzazione degli errori di natura casuale. Infatti, sotto le ipotesi di indipendenza degli errori sulle singole variabili e di bassa probabilità di errore per ciascuna variabile, la modifica del minor numero di valori consente di massimizzare la probabilità di localizzare correttamente i valori errati. Il problema della localizzazione degli errori è quindi ricondotto alla risoluzione di un problema di programmazione lineare intera[1] con struttura di *Set Covering*, ovvero un problema di ricopertura di insiemi: si ricerca l'insieme di peso minimo di variabili il cui cambiamento permette di ripristinare la correttezza del record. Tale problema è risolto utilizzando tecniche di ottimizzazione.

[1] Sia la *funzione obiettivo* sia i *vincoli* sono funzioni lineari e le variabili oggetto di ottimizzazione sono vincolate ad assumere valori interi.

La procedura di localizzazione degli errori implementata nel software Banff (*proc Errorloc*) è basata sul principio del *minimo cambiamento*. Per la ricerca della soluzione ottima la *proc Errorloc* utilizza un algoritmo sviluppato da Chernikova (1964, 1965) e generalizzato da Rubin (1975). Tale procedura è *efficace*[2] rispetto al principio del *minimo cambiamento* in quanto è in grado di individuare il numero minimo di campi da modificare per soddisfare tutte le regole. Anche l'algoritmo di localizzazione degli errori "*first field then donor*" implementato nel Sistema DIESIS è basato sul principio del *minimo cambiamento*. Questo algoritmo utilizza un metodo di ricerca della soluzione ottima basato sulla tecnica del *Branch and Cut*[3] (Sassano, 1999).

L'***efficacia*** rispetto al principio del *minimo cambiamento* dell'algoritmo "*first field then donor"* del Sistema DIESIS è stata valutata confrontando, per ciascun record errato, il numero di campi da esso identificati con il numero di campi identificati dalla *proc Errorloc* del software Banff.

Un'altra esigenza fondamentale dell'Istituto, connessa al requisito di *tempestività* dei dati statistici, è quella di disporre di una procedura di localizzazione degli errori casuali non influenti che sia in grado di ottenere il prodotto atteso in tempi di esecuzione contenuti. Tale requisito è richiesto per tutte le indagini, ma è particolarmente rilevante per quelle censuarie, poiché il tempo di elaborazione aumenta all'aumentare del numero di record sottoposti a controllo (e del numero di regole di controllo definite). In generale, l'importanza dell'*efficienza* computazionale di una procedura destinata alla produzione è manifesta, se si considera che una procedura pur eccellente in termini di efficacia potrebbe risultare completamente inutilizzabile se non fosse applicabile al processo corrente di produzione dei dati a causa di tempi di esecuzione eccessivamente elevati. Si è ritenuto, pertanto, essenziale valutare anche l'*efficienza operativa* del sistema DIESIS.

---

[2] Per *efficacia* di un processo si intende la capacità del processo stesso di ottenere il prodotto desiderato (ovvero di raggiungere l'obiettivo).

[3] L'idea fondamentale di questo metodo è quella di effettuare una esplorazione implicita di tutta la regione ammissibile S (insieme delle soluzioni ammissibili del problema) alla ricerca della soluzione ottima. Per chiarire il senso del termine esplorazione implicita osserviamo innanzitutto che, essendo l'insieme ammissibile costituito da un numero finito di vettori, è sempre possibile, almeno in linea di principio, risolvere il problema di programmazione lineare intera calcolando il valore della funzione obiettivo in ciascuna soluzione appartenente ad S e scegliendo poi la migliore. Questo metodo può essere definito come "Enumerazione Totale di S". Se la cardinalità di S è piccola, allora l'enumerazione totale non solo è possibile ma è certamente il modo di risolvere il problema. Se, viceversa, come generalmente accade nei casi reali, la cardinalità dell'insieme delle soluzioni ammissibili è molto grande, l'enumerazione totale diviene non proponibile. Il Branch and Cut fornisce una metodologia di ricerca della soluzione che effettua una esplorazione parziale dell'insieme S. In particolare, la funzione obiettivo viene calcolata per un insieme piccolo di punti ammissibili con la proprietà di contenere almeno una soluzione ottima.

L'***efficienza operativa*** dell'algoritmo "*first field then donor"* del Sistema DIESIS è stata valutata confrontando i tempi di esecuzione da esso impiegati con quelli impiegati dalla *proc Errorloc* del software Banff[4].

2.2 I dati

La sperimentazione DIESIS/Banff è stata condotta utilizzando i dati grezzi relativi alle coltivazioni dell'indagine campionaria su ***Struttura e Produzioni delle Aziende Agricole (SPA)*** **anno 2005**. L'indagine *SPA*, svolta dall'Istat con periodicità biennale, ha l'obiettivo di monitorare la struttura aziendale e la sua evoluzione nel tempo attraverso la rilevazione di circa 400 caratteri riguardanti: superfici dedicate alle diverse coltivazioni, produzioni delle coltivazioni, consistenza degli allevamenti, forma organizzativa, rapporti dell'azienda con il mercato, lavoro, vendita, pratiche agronomiche, azienda agricola e ambiente. La raccolta dei dati è effettuata tramite intervista diretta con questionario cartaceo. L'unità di rilevazione è l'azienda agricola (con produzione agricola o zootecnica).

I dati *SPA* sono stati selezionati per la sperimentazione DIESIS/Banff in quanto:

- le variabili rilevate con l'indagine *SPA* sono molto simili a quelle che saranno rilevate con il Censimento dell'agricoltura dell'anno 2010. La procedura per il controllo e la correzione dei dati di tale censimento è ancora da definire, e potrebbe avvalersi del software DIESIS per la fase di localizzazione e correzione degli errori casuali non influenti.
- il trattamento degli errori casuali non influenti dei dati *SPA* (anni 2003 e 2005) è stato effettuato utilizzando il software Banff (Ballin *et al*., 2004; Guarnera *et al*., 2006) e quindi le regole di controllo utilizzate in produzione sono prontamente disponibili per la sperimentazione DIESIS/Banff.

[4] E' necessario far presente che nel software Banff è implementata esclusivamente la *proc Errorloc* per la localizzazione degli errori. Diversamente, nel software DIESIS oltre all'algoritmo "*first fields then donors*" è implementato anche un altro algoritmo, denominato "*first donors then fields"*, basato su un approccio *data-driven*, rivolto all'individuazione del *minimo cambiamento condizionato dai donatori disponibili*. Tale approccio, particolarmente idoneo al trattamento di variabili qualitative, non è oggetto della sperimentazione DIESIS/Banff descritta nel presente documento.

Riguardo all'anno di riferimento della rilevazone, i dati *SPA* 2005 sono stati preferiti a quelli dell'anno 2003 sia perchè la struttura del questionario dell'anno 2005 è più simile a quella del questionario del Censimento dell'agricoltura 2010, sia perchè i dati *SPA* 2005 non contengono errori sistematici, a differenza di quelli dell'anno 2003.

E' opportuno osservare che la fase di trattamento degli errori casuali non influenti dei dati *SPA* (2003 e 2005) è stata preceduta da una fase interattiva, eseguita con il software AGAIN (*Analisi e Gestione Automatica delle Indagini*) (Benedetti *et al*., 2002), avente l'obiettivo di individuare e correggere i valori anomali e gli errori influenti. I dati dell'anno 2005 sono stati corretti quasi completamente nel corso della fase interattiva, e quindi l'insieme dei dati elaborato con il software Banff conteneva un numero esiguo di valori errati. Per questa ragione l'insieme dei dati effettivamente elaborato con Banff non è stato considerato idoneo ad essere utilizzato per il confronto dei risultati DIESIS *vs* Banff e si è deciso di condurre la sperimentazione sui **dati grezzi** dell'anno 2005.

Le variabili selezionate per la sperimentazione sono quelle relative alle **superfici aziendali** (o *coltivazioni*). Tali variabili sono state preferite a quelle relative alla *consistenza degli allevamenti* in quanto più numerose e con un maggior numero di informazioni disponibili. Inoltre, le variabili relative alle coltivazioni sono connesse da un numero maggiore di vincoli, utilizzabili per la definizione delle regole di controllo, rispetto alle variabili relative alla consistenza degli allevamenti.

L'insieme di unità considerate nello studio DIESIS/Banff è costituito da 49014 aziende agricole (dati Italia). Le variabili relative alle coltivazioni controllate con il software Banff sono 209[5]. L'unità di misura (numeri interi positivi) è l'*ara* (1 *ara*=100 *mq*) ad eccezione delle superfici relative alle coltivazioni di funghi (variabile C110) e serre (variabile C111) che sono registrate in *mq*.

[5] In produzione, il trattamento delle variabili relative alle superfici aziendali dei dati *SPA* 2005 è stato effettuato mediante due sottoprocessi distinti. L'insieme iniziale di regole è stato suddiviso in due sottoinsiemi e i dati sono stati elaborati (localizzazione e imputazione) utilizzando separatamente i due sottoinsiemi. Poiché alcune variabili (94) erano coinvolte in entrambi i sottoinsiemi di regole, è stato necessario definire un ordine prefissato di esecuzione dei sottoprocessi e tenere fissi, durante l'esecuzione del secondo sottoprocesso, i valori delle variabili in comune.

2.3 Le regole

La presenza di errori nei dati quantitativi è rilevata per mezzo di un insieme di regole (edit) matematiche. In generale, le regole possono essere espresse come condizioni di correttezza dei valori osservati nell'unità (regole PASS) o condizioni di errore (regole FAIL). Un record è definito **esatto** quando tutte le regole PASS sono *vere* per quel record e tutte le regole FAIL sono *false*. Un record è **errato** quando almeno una regola PASS è *falsa* o almeno una regola FAIL è *vera*.

Nei software Banff e DIESIS le regole sono specificate in forma di uguaglianze o disuguaglianze lineari. Banff accetta regole specificate sia in forma FAIL sia in forma PASS e provvede (mediante una funzione interna) a convertire automaticamente la forma FAIL in forma PASS. L'attuale versione di DIESIS richiede che le regole siano espresse esclusivamente in forma PASS[6]. Nel seguito ci riferiremo esclusivamente a regole (o edit) PASS.

Un insieme di edit che coinvolge n variabili definisce una regione di accettazione[7] nello spazio n-dimensionale R^n. I valori dei record che soddisfano tutti gli edit (record esatti) appartengono alla regione di accettazione. Di contro, i valori dei record errati cadono fuori dalla regione di accettazione. Secondo il paradigma del *minimo cambiamento*, l'obiettivo di una procedura di localizzazione degli errori consiste nell'identificare il minor numero di campi (o l'insieme di peso minimo di campi) da modificare in modo da riportare il record errato nella regione di accettazione. Come accennato nella sezione 2.1, questo problema è ricondotto alla risoluzione di un problema di programmazione lineare intera (i campi da imputare sono individuati minimizzando una funzione obiettivo soggetta ad una serie di vincoli) dove gli edit rappresentano i vincoli del problema e definiscono l'insieme delle soluzioni ammissibili.

Nelle elaborazioni eseguite dai due software è stato utilizzato lo stesso insieme di regole di controllo. Tale insieme di regole è costituito dal *minimal set di edit*[8] determinato da Banff[9] a partire

---

[6] Non esiste però alcuna preclusione per l'utilizzo futuro in DIESIS di regole espresse dall'utente in forma FAIL in quanto è attualmente in corso lo sviluppo di un modulo che automaticamente consente la traduzione di regole espresse in forma FAIL (regole di incompatibilità) in regole in forma PASS (regole di compatibilità).

[7] La regione di accettazione è convessa e contiene la frontiera.

[8] Il *minimal set di edit* contiene il minimo numero di regole necessario per definire la regione di accettazione relativa all'insieme iniziale di regole. Il *minimal set* di edit è ottenuto verificando che le regole siano consistenti tra loro (ossia che la regione di accettazione non sia vuota) ed eliminando le regole ridondanti (regole che non contribuiscono a definire la regione di accettazione), le uguaglianze nascoste (ossia uguaglianze implicite) e le variabili determinate (ossia variabili che possono assumere un solo valore).

[9] Mediante la *proc Verifyedits*.

dalle regole iniziali specificate dal servizio SAG per il trattamento in produzione delle variabili relative alle superfici aziendali dei dati *SPA* 2005. L'insieme di regole è composto da 262 edit di cui 154 esprimenti vincoli di positività per i valori di singole variabili (*non-negativity edits*) ossia regole del tipo "*nome_var* ≥ 0". I restanti 108 edit sono riportati nella Tabella 1 dell'Appendice. Tra questi, 13 (ID: 93, 94, 96, 99-108) sono vincoli di quadratura (*balance edit*) e 3 (ID: 95, 97, 98) sono vincoli di uguaglianza tra i valori di due variabili.

Sulla base delle regole utilizzate sono stati individuati 8017 record errati (16.36%).

2.4 Elaborazione con Banff

E' stata utilizzata la versione Banff 2.02.002 su Windows per SAS v9.

La procedura di localizzazione degli errori del software Banff (*Proc Errorloc*) consente la specificazione di alcuni parametri (opzionali). Poiché in Banff il tempo di elaborazione aumenta considerevolmente all'aumentare del numero di edit e di record sottoposti a controllo, il *massimo tempo disponibile per record in secondi* (*timeperobs*) è un parametro solitamente utilizzato per evitare che il costo di elaborazione si protragga oltre un limite non accettabile. Generalmente la fase di localizzazione degli errori è effettuata in più passi. Nel primo passo si definisce un valore per il parametro *timeperobs* che consenta di individuare una soluzione di minimo cambiamento (tra quelle possibili) per la maggioranza dei record. Successivamente, i record residui, ossia quelli per i quali il *timeperobs* non ha consentito di individuare una soluzione, vengono sottoposti ad una nuovo passo di localizzazione utilizzando un valore maggiore per il parametro *timeperobs*[10].

Nella sperimentazione in oggetto il *timeperobs* è stato l'unico parametro opzionale utilizzato nella *proc Errorloc* di Banff[11]. Poiché si voleva disporre di una soluzione per ciascun record errato

[10] In produzione, il primo sottoprocesso (localizzazione e imputazione) delle variabili relative alle superfici aziendali dei dati SPA 2005 ha utilizzato due passi di localizzazione degli errori aventi, rispettivamente, *timeperobs*=45 e *timeperobs*=100.

[11] Altri parametri spesso usati nelle procedure di localizzazione sono i "pesi" di affidabilità assegnati alle variabili. In questi casi l'algoritmo non identifica l'insieme minimo di variabili da modificare bensì l'insieme di peso minimo. Nonostante nel processo di produzione dei dati SPA siano stati utilizzati dei pesi per alcune variabili, nella sperimentazione DIESIS/Banff si è preferito utilizzare pesi unitari per tutte le variabili.

(da usare per il confronto con la soluzione individuata da DIESIS) sono stati eseguiti diversi passi di localizzazione aumentando di volta in volta il valore del parametro *timeperobs*[12].

2.5 Elaborazione con DIESIS

La versione attuale del sistema DIESIS è di tipo multi-piattaforma (UNIX, LINUX e WINDOWS). Tale versione è interfacciata con il software commerciale Cplex della ILOG, che è riconosciuto, nell'ambito della Ricerca Operativa, come il migliore solutore di problemi di programmazione lineare intera al momento in commercio.

La procedura di localizzazione degli errori del sistema DIESIS richiede l'assegnazione di alcuni parametri. E' necessario indicare al sistema di attivare solo l'approccio "*first fields then donors*" (minimo cambiamento assoluto). E' possibile assegnare pesi di affidabilità a ciascuna delle variabili, e in particolare per ciascuna variabile è possibile assegnare anche pesi diversi a sottoinsiemi di valori ammissibili definiti dall'utente.

Oltre ai file dei record errati e delle regole, DIESIS richiede la predisposizione di due file per la descrizione dettagliata delle variabili e dei loro intervalli di variazione.

I passi principali dell'elaborazione sono i seguenti:

1. lettura delle variabili

Inizialmente il sistema legge tutte le informazioni riguardanti la tipologia delle variabili e le modalità di utilizzo delle stesse. Ad esempio, se la variabile è imputabile, se è una variabile di strato, se è una variabile di coppia, se ha le tabelle di distanza, se è presente nel problema di ottimizzazione, ed infine il numero d'ordine delle variabili nel problema della localizzazione dell'errore.

2. lettura delle regole

Le regole sono distinte per tipologia secondo la seguente classificazione, ottenuta in funzione del tipo di regola (logica, matematica, oppure logico matematica) e del numero di individui implicati (un solo individuo, due o tre individui):

- logiche individuali generiche;
- logiche individuali specifiche;

[12] Trattandosi di una sperimentazione, il valore di *timeperobs* è stato "stressato" ad un valore molto più elevato di quello consentito dalle normali esigenze di produzione.

- logiche interpersonali 2 individui;
- logiche interpersonali 3 individui;
- logiche matematici individuali;
- logiche matematici interpersonali 2 individui;
- logiche matematici interpersonali 3 individui;
- matematiche individuali;
- matematiche interpersonali 2 individui;
- matematiche interpersonali 3 individui.

Ogni regola viene analizzata e decomposta nei suoi elementi base individuando il campo e l'individuo a cui si riferisce la variabile, le operazioni matematiche eventualmente presenti, il numero di disgiunzioni logiche e matematiche, i versi e il termine noto delle eventuali disequazioni.

Dalle regole vengono estratti i *break-point* delle variabili, valori costanti che sono determinati dalle regole.

Un *break-point* rappresenta essenzialmente un punto di demarcazione sul dominio, cioè un confine che permette di partizionare il dominio in tanti sotto-insiemi disgiunti

3. lettura dei record errati

I campi del record errato sono codificati utilizzando i codici definiti nel file delle variabili. In particolare per ogni campo si utilizza il codice della variabile associata a quel campo, il valore della modalità della variabile e la codifica della modalità.

4. localizzazione dell'errore

Il cuore del programma è chiaramente in questo blocco, che contiene tutte le funzionalità richieste per modellare e risolvere il problema della localizzazione dell'errore.

Il sistema DIESIS è interfacciato con il suddetto solutore commerciale Cplex della ILOG. Attraverso un insieme di procedure dedicate sia alla creazione dei vincoli, secondo la sintassi imposta da Cplex, sia alla modellazione della funzione obiettivo, si giunge ad istanziare un modello matematico che rappresenta efficacemente il problema di ottimizzazione.

Per risolvere tale problema il sistema utilizza il metodo del *Branch and Bound*.

5. generazione output

Dopo aver risolto il modello matematico il sistema fornisce in uscita le informazioni riguardanti: il codice della regola violata, l'identificativo del record che la viola, il numero e le variabili da cambiare per ripristinare la correttezza del record rispetto all'insieme di regole in esame.

2.6 Risultati

Tutte le elaborazioni (sia Banff sia DIESIS) sono state effettuate su un PC AMD Athlon(tm) XP 2600+ con 2.14GHz e 512 MB di RAM.

Come descritto nella sezione 2.4, la procedura di localizzazione degli errori del software Banff è stata eseguita in più passi aumentando di volta in volta il valore del parametro *timeperobs (massimo tempo disponibile per record in secondi)*. La Tabella 1 riporta, per ciascun passo di localizzazione di Banff, il valore del parametro *timeperobs,* il numero di record sottoposti ad elaborazione, quelli con soluzione e quelli senza soluzione, nonché il tempo (ore:minuti:secondi) reale complessivo di esecuzione.

Tabella 1 – Elaborazione con il software Banff

Passo	*Timeperobs* (secondi)	Record elaborati	Record con soluzione	Record senza soluzione	Tempo di esecuzione (ore:minuti:secondi)
1	45	8017	7694	323	8:57:51
2	200	323	80	243	15:45:36
3	850	243	35	208	53:25:24
4	1200	208	9	199	69:18:50
1-4		**8017**	**7818**	**199**	**147:25:41**

Come si evince dalla Tabella 1, il primo passo di localizzazione è stato eseguito ponendo *timeperobs* = 45 secondi; il tempo di esecuzione complessivo è stato di circa 8 ore e 57 minuti; il software Banff ha trovato una soluzione di cardinalità minima per 7694 record su 8017 record elaborati; 323 record errati sono rimasti senza soluzione a causa di vincoli temporali e sono stati sottoposti nuovamente alla procedura di localizzazione degli errori (secondo passo). Nel secondo passo di localizzazione è stato posto *timeperobs* = 200; il tempo di esecuzione complessivo è stato di circa 15 ore e 45 minuti; il software Banff ha trovato una soluzione di cardinalità minima per 80 record su 323 elaborati mentre 243 record errati sono rimasti senza soluzione a causa di vincoli temporali; tali record sono stati sottoposti al terzo passo di localizzazione. Al terzo passo di localizzazione è stato posto *timeperobs* = 850; il tempo di esecuzione complessivo è stato di circa 53 ore e 25 minuti; il software Banff ha trovato una soluzione di cardinalità minima per 35 record

su 243 elaborati; i 208 record errati rimasti senza soluzione a causa di vincoli temporali sono stati sottoposti al quarto passo di localizzazione. Al quarto passo di localizzazione è stato posto *timeperobs* = 1200 corrispondente a 20 minuti per record, un tempo eccessivamente elevato rispetto a quello consentito dalle normali esigenze di produzione; il tempo di esecuzione complessivo è stato di circa 69 ore e 18 minuti; i 20 minuti a disposizione per ciascun record non sono risultati sufficienti all'algoritmo di localizzazione per individuare una soluzione di cardinalità minima per 199 record errati che, pertanto, sono rimasti senza soluzione (record *rejected* da Banff). Non è stato ritenuto opportuno aumentare il tempo disponibile per record (valore del parametro *timeperobs*) e quindi non sono stati eseguiti ulteriori passi di localizzazione con Banff.

La procedura di localizzazione degli errori mediante l'algoritmo "*first field then donor"* del software DIESIS è stata eseguita in un solo passo. Nella Tabella 2 sono riportati il numero di record sottoposti ad elaborazione, quelli con soluzione e quelli senza soluzione, nonché il tempo (ore:minuti:secondi) reale complessivo di esecuzione.

Tabella 2 – Elaborazione con il sistema DIESIS

Passo	*Timeperobs* (secondi)	Record elaborati	Record con soluzione	Record senza soluzione	Tempo di esecuzione (minuti)
1	_*	8017	8017	0	30

*DIESIS non necessita di parametri vincolanti il tempo massimo di elaborazione per singolo record.

Come si evince dalla Tabella 2, il software DIESIS ha individuato una soluzione di cardinalità minima (tra quelle possibili) per tutti gli 8017 record errati nel corso di una elaborazione durata circa 30 minuti. Il tempo medio di elaborazione per singolo record è stato pertanto di circa 0.22 secondi. Tale tempo di esecuzione può essere ulteriormente ridotto (di circa il 20%) se si inibisce la stampa di alcuni messaggi (*log*) utilizzati per il controllo del processo e assegnando in modo ottimale i valori dei parametri utilizzati dal sistema DIESIS.

Poiché il software Banff è stato in grado di trovare una soluzione solo per 7818 record errati (nei tempi a disposizione), solo per questi è possibile confrontare il numero dei campi identificati dai due software ossia, la cardinalità delle soluzioni trovate dai due software. Per ciascuno dei 7818 record errati per i quali entrambi i software hanno trovato una soluzione è stato identificato lo **stesso numero di campi da modificare**. La distribuzione di frequenza della cardinalità della soluzione per i record errati risolti da entrambi i software è riportata nella Tabella 3.

Tabella 3 – Banff e DIESIS – Distribuzione della cardinalità della soluzione (record risolti da entrambi i software)

Cardinalità	Frequenza	*Percentuale*
1	4901	*62.69*
2	2074	*26.53*
3	518	*6.63*
4	239	*3.06*
5	68	*0.87*
6	17	*0.22*
7	1	*0.01*
Totale	7818	*100.00*

Dalla Tabella 3 si evince che per ripristinare la situazione di correttezza, rispetto alle regole specificate, per il 62.69% dei record errati occorre modificare il valore di una sola variabile, mentre per il 26.53% dei record errati è necessario modificare il valore di due variabili, e così via.

Quando la soluzione di minimo cambiamento non è unica (sono identificati due o più insiemi di variabili da modificare di uguale cardinalità minima) Banff e DIESIS selezionano una soluzione in modo casuale. Questo elemento di casualità fa sì che per un dato record errato i due software possano identificare differenti insiemi di variabili (a parità di cardinalità). Né Banff né DIESIS forniscono l'indicazione delle soluzioni multiple trovate. Per verificare l'esistenza di soluzioni multiple, nella Tabella 4, per ciascuna cardinalità della soluzione, è riportata la distribuzione del numero di campi diversi nelle soluzioni individuate da Banff e DIESIS (la percentuale è calcolata rispetto al numero di soluzioni con la cardinalità in questione, ossia le frequenze della Tabella 3).

Tabella 4 – Banff e DIESIS – Distribuzione del numero di campi diversi nelle soluzioni Banff e DIESIS (record risolti da entrambi i software)

Cardinalità	N.ro di campi diversi	Frequenza	*Percentuale*
1	0	3697	*75.43*
	1	1204	*24.57*
2	0	1058	*51.01*
	1	576	*27.77*
	2	440	*21.22*
3	0	175	*33.78*
	1	155	*29.92*
	2	154	*29.73*
	3	34	*6.56*
4	0	70	*29.29*
	1	50	*20.92*
	2	92	*38.49*
	3	20	*8.37*
	4	7	*2.93*
5	0	15	*22.06*
	1	13	*19.12*
	2	23	*33.82*
	3	13	*19.12*
	4	3	*4.41*
	5	1	*1.47*
6	0	1	*5.88*
	1	2	*11.76*
	2	6	*35.29*
	3	4	*23.53*
	4	1	*5.88*
	5	2	*11.76*
	6	1	*5.88*
7	2	1	*100.00*

Per ciascuna cardinalità della soluzione, si osservano soluzioni Banff e DIESIS composte da campi diversi, a riprova dell'esistenza di soluzioni multiple al problema di localizzazione degli errori.

Per i 199 record per i quali il software Banff non è riuscito a trovare una soluzione di minimo cambiamento a causa di vincoli temporali, pur avendo a disposizione un *massimo tempo disponibile per record* di 20 minuti, è possibile conoscere solo la cardinalità della soluzione trovata dal sistema DIESIS. Nella Tabella 5 è riportata la distribuzione di frequenza della cardinalità della soluzione DIESIS relativa al suddetto sottoinsieme di record.

Tabella 5 – DIESIS – Distribuzione della cardinalità della soluzione (record risolti solo da DIESIS)

Cardinalità	Frequenza	*Percentuale*
1	1	*0.50*
2	49	*24.62*
3	81	*40.70*
4	28	*14.07*
5	21	*10.55*
6	17	*8.54*
7	2	*1.01*
Totale	199	*100.00*

Come si evince dal confronto tra i valori nelle Tabelle 3 e 5, la distribuzione di frequenza della cardinalità della soluzione per i 199 record risolti solo da DIESIS differisce da quella osservata per i record risolti da entrambi i software. Per i record risolti da entrambi i software la distribuzione risulta essere fortemente asimmetrica (valore modale=1). Tale asimmetria si riduce notevolmente per i record risolti solo da DIESIS (valore modale=3).

3 Conclusioni

L'obiettivo dello studio descritto nel presente documento è quello di valutare alcune caratteristiche di *efficacia* (rispetto al principio del *minimo cambiamento*) e di *efficienza operativa* dell'algoritmo di localizzazione "*first fields then donors*" implementato nel sistema DIESIS nell'ambito della localizzazione degli errori in variabili quantitative. Le suddette caratteristiche sono state considerate congiuntamente in quanto sono requisiti necessari per l'applicabilità del software nelle procedure di controllo e correzione dei dati in produzione. La strategia di valutazione adottata è di tipo comparativo, ed il software Banff, in quanto raccomandato ed utilizzato a livello

internazionale, è stato scelto come riferimento. Gli indicatori utilizzati per la valutazione sono costituiti dalla cardinalità della soluzione (per la valutazione dell'*efficacia*) e dal tempo di esecuzione (per la valutazione dell'*efficienza operativa*).

Dai risultati ottenuti si evince che il Sistema DIESIS è efficace quanto il software Banff (per i record risolti da entrambi i software, la cardinalità della soluzione DIESIS è uguale alla cardinalità della soluzione Banff) ma è di gran lunga più efficiente (DIESIS ha elaborato con successo tutti gli 8017 record errati in un tempo complessivo di circa 30 minuti, mentre il software Banff non è stato in grado di trovare una soluzione per 199 record a fronte di un tempo complessivo di elaborazione dei quattro passi di localizzazione applicati superiore a 147 ore). La minore efficienza (in termini di tempo di esecuzione) del software Banff non ha consentito di pervenire ad una soluzione per una quota di record errati che devono essere, pertanto elaborati in modo alternativo. Nel trattamento in produzione dei dati *SPA* anno 2005, i record errati senza soluzione Banff sono stati sottoposti a revisione manuale (Guarnera *et al*., 2006). La revisione manuale dei record errati privi di soluzione automatica difficilmente si concilia con le esigenze di produzione dei dati censuari.

Tutto ciò suggerisce indubbiamente di proseguire lo sviluppo del Sistema DIESIS.

Lo sviluppo di tale Sistema, pur avendo raggiunto livelli avanzati, necessita di ulteriori ampliamenti che consentirebbero:

- un utilizzo agevole anche da parte di altri utenti (attualmente sono in grado di utilizzare DIESIS solo coloro che lo hanno progettato e sviluppato[13]);
- l'impiego del Sistema per il trattamento dei dati di altre indagini condotte dall'Istituto (attualmente DIESIS è stato utilizzato per il controllo e la correzione delle variabili anagrafiche e di stato civile rilevate al Censimento della Popolazione del 2001);
- la diffusione del Sistema ad altri Istituti nazionali di statistica, favorendo in tal modo la collaborazione scientifica internazionale.

Tali ampliamenti riguardano:

1. la creazione di un'interfaccia (grafica o no) di facile utilizzo per gli utenti, che renda agevole la dichiarazione dei parametri e delle regole;
2. la scelta e l'adattamento a DIESIS di un solutore *open source* che possa sostituire il solutore commerciale Cplex della Ilog, con conseguenti benefici in termini economici;
3. l'inserimento di ulteriori moduli per l'imputazione delle variabili quantitative. Ad esempio, l'imputazione con metodi da modello (attualmente DIESIS esegue solo l'imputazione di tipo

[13] Alessandra Reale e Gianpiero Bianchi.

sequenziale attingendo il valore dal donatore più vicino o generandolo dalla distribuzione marginale della variabile);

4. lo sviluppo di un insieme più ampio di funzioni per il calcolo della distanza tra record donatori e record errati;
5. la realizzazione di manuali di metodo e d'uso per l'utente.

Tali ampliamenti consentirebbero di disporre di un sistema efficiente ed altamente versatile da utilizzare per il controllo e la correzione degli errori casuali non influenti in variabili quantitative e/o qualitative, e quindi dei dati provenienti da indagini di carattere demografico, sociale o economico.

In particolare, la sostituzione di Cplex con un solutore open source (punto 2.) è condizione necessaria sia per far conseguire risparmi all'Istituto sia per consentire la completa libertà gestionale dello strumento, e quindi la sua fruibilità non onerosa in ogni Direzione di produzione e, in prospettiva, anche in ambito SISTAN e internazionale.

E' importante, inoltre, tener presente che DIESIS consente anche il trattamento di dati a struttura gerarchica (ad esempio, dati familiari per i quali devono essere considerate regole intra- ed inter-componente) e l'uso della metodologia *data driven* (in modalità isolata o congiunta a quella di minimo cambiamento) (Manzari and Reale, 2001; Bianchi *et al.*, 2005a).

Non trascurabile è, infine, la considerazione che la disponibilità di un Sistema sviluppato internamente all'Istituto, consente di poter effettuare aggiornamenti ed eventuali riadattamenti per fronteggiare situazioni *ad hoc* in maniera efficace e con un contenimento dei costi.

Riferimenti bibliografici

Ballin M., Guarnera U., Luzi O., Salvi S. (2004) New methodologies and tools for dealing with non-sampling errors the Istat survey on structure and Production of Agricutural firms. *Atti del Convegno Metodi d'Indagine e di Analisi per le Politiche Agricole-MIAPA 2004*- Università di Pisa 21-22 Ottobre.

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

Bianchi G., Manzari A., Pezone A., Reale A., Saporito G. (2005a) New procedures for editing and imputation of demographic variables. *Contributed paper presented at UN-ECE Work Session on Statistical Data Editing, Canada (Ottawa),* May 16-18, 2005.

Bianchi G., Manzari A., Reale A. (2005b) Relazione sull'attività di ricerca Istat-DIS (Dipartimento di Informatica e Sistemistica dell'Università 'La Sapienza' di Roma - Facoltà di Ingegneria) sui temi: Modelli e algoritmi per problemi di edit ed imputation e Modelli e metodi per problemi di linkage e clustering di dati. *Istat-Documento Interno* Ottobre 2005.

Bruni R., Reale A., Torelli R. (2001) Optimization Techniques for Edit Validation and Data Imputation, *presented at the Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" XVIIIth International Symposium on Methodological Issues*.

Chernikova N. V. (1964) Algorithm for Finding a General Formula for the Non-negative solutions of a System of Linear Equations, *USSR Computational Mathematics and Mathematical Physics*, 4, 151-158.

Chernikova N. V. (19645 Algorithm for Finding a General Formula for the Non-negative solutions of a System of Linear Inequalities, *USSR Computational Mathematics and Mathematical Physics*, 5, 228-233.

Cotton C. (1991) Functional description of the generalized edit and imputation system. *Business Survey Methods Division* - July 25 Statistics Canada.

Fellegi, I.P., Holt D. (1976) A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* 71, 17-35.

Guarnera U., Luzi O., Salvi S. (2006) 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.

Kovar J.G., MacMillian J.H., and Whitridge P. (1988) Overview and strategy for the generalized edit and imputation system. *Report, Methodology Branch - April 1988 (updated February 1991) Statistics Canada.*

Manzari A., Reale A. (2001) Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, *In Proc. 53rd Session of The International Statistical Institute*, August 22-29, 2001, pp. 634-655. Sydney: International Statistical Institute.

Rubin D. S. (1975) Vertex Generation and Cardinality Constrained Linear Programs. *Operations Research*, 23, 555-565.

Sassano A. (1999) Modelli ed Algoritmi della Ricerca Operativa. *Ed. Franco Angeli*.

Appendice

L'insieme di regole di controllo utilizzate nella sperimentazione DIESIS/Banff è composto da 262 edit aritmetici specificati in forma di uguaglianze o disuguaglianze lineari. Ben 154 edit esprimono vincoli di positività per i valori di singole variabili (*non-negativity edits*) ossia regole del tipo "*nome_var* ≥ 0"; tali edit non sono riportati nel presente documento. I restanti 108 edit sono riportati nella Tabella 1 che segue. Tra questi, 13 (ID 93, 94, 96, 99-108) rappresentano vincoli di quadratura (*balance edit*) e 3 rappresentano vincoli di uguaglianza tra i valori di due variabili (ID 95, 97, 98).

Tabella1 – Regole di controllo utilizzate nella sperimentazione (esclusi i vincoli di positività)

ID	Regola
0	- SAT1 + SAU1 <= 0
1	- SAT2 + SAU2 <= 0
2	- SAT3 + SAU3 <= 0
3	- SAT4 + SAU4 <= 0
4	- SAT5 + SAU5 <= 0
5	- SAT6 + SAU6 <= 0
6	- SAT7 + SAU7 <= 0
7	- C57 + S57 <= 0
8	- C58 + S58 <= 0
9	- C59 + S59 <= 0
10	- C61 + S61 <= 0
11	- C62 + S62 <= 0
12	- C63 + S63 <= 0
13	- C64 + S64 <= 0
14	- C65 + S65 <= 0
15	- C66 + S66 <= 0
16	- C67 + S67 <= 0
17	- C68 + S68 <= 0
18	- C69 + S69 <= 0
19	- C70 + S70 <= 0
20	- C71 + S71 <= 0
21	- C72 + S72 <= 0
22	- C73 + S73 <= 0
23	- C74 + S74 <= 0
24	- C75 + S75 <= 0
25	- C76 + S76 <= 0
26	- C77 + S77 <= 0
27	- C78 + S78 <= 0
28	- C79 + S79 <= 0

29	- C80 + S80 <= 0
30	- C81 + S81 <= 0
31	- C82 + S82 <= 0
32	- C86 + S86 <= 0
33	- C87 + S87 <= 0
34	- 100 * C100 + C110 <= 0
35	- 100 * C101 + C111 <= 0
36	- FOR1 + FOR11 + FOR5 + FOR8 <= 0
37	FOR12 - FOR2 + FOR6 + FOR9 <= 0
38	FOR10 + FOR13 - FOR3 + FOR7 <= 0
39	- C94 - C97 + IR0 <= 0
40	- IR0 + IR16 <= 0
41	C89 <= 30
43	- IR16 + IR20 <= 0
44	- IR16 + IR21 <= 0
45	- IR16 + IR22 <= 0
46	- IR16 + IR23 <= 0
47	- IR16 + IR25 <= 0
48	- IR23 + IR24 <= 0
49	C104 + C105 + C38 + C40 + C41 + C43 + C44 + C89 - C94 - C97 + IR20 <= 0
50	C104 + C105 + C38 + C40 + C41 + C43 + C44 + C89 - C94 - C97 + IR21 <= 0
51	C38 + C40 + C41 + C43 + C44 + C89 - C94 - C97 + IR22 <= 0
52	C104 + C105 + C38 + C40 + C41 + C43 + C44 + C89 - C94 - C97 + IR23 <= 0
53	C104 + C105 + C38 + C40 + C41 + C43 + C44 + C89 - C94 - C97 + IR25 <= 0
54	- C94 + COLT1 <= 0
55	- C94 + COLT2 <= 0
56	- C94 + COLT3 <= 0
57	- C56 - C88 + COLT4 <= 0
58	- C93 + COLT5 <= 0
59	- C97 - C98 + COLT6 <= 0
60	- C55 + COLT7 <= 0
61	- C100 - C99 + COLT8 <= 0
62	- C101 + COLT9 <= 0
63	- C101 + COLT10 <= 0
64	- C94 - C97 + SUP1 <= 0
65	- C94 - C97 + SUP2 <= 0
66	- C94 - C97 + SUP3 <= 0
67	- C94 - C97 + SUP4 <= 0
68	- C94 - C97 + SUP5 <= 0
69	- C94 - C97 + SUP6 <= 0

70	- C94 - C97 + SUP7 <= 0
71	- C94 - C97 + SUP8 <= 0
72	- C98 + SUP9 <= 0
73	- C98 + SUP10 <= 0
74	- C98 + SUP11 <= 0
75	- C98 + SUP12 <= 0
76	- C98 + SUP13 <= 0
77	- C98 + SUP14 <= 0
78	BIO1 - C1 - C10 - C2 - C3 - C4 - C5 - C6 - C7 - C8 - C9 <= 0
79	BIO2 - C102 - C103 - C104 - C105 - C32 - C33 - C34 - C35 - C37 - C38 - C40 - C41 <= 0
80	BIO3 - C57 - C58 - C59 - C60 <= 0
81	BIO4 - C61 - C62 <= 0
82	BIO5 - C63 - C64 - C65 - C66 - C67 <= 0
83	BIO6 - C68 - C69 - C70 - C71 - C72 - C73 - C74 - C75 - C76 - C77 - C78 - C79 - C80 - C81 - C82 <= 0
84	BIO7 - C93 <= 0
85	BIO8 - C11 - C13 - C14 - C16 - C17 - C18 - C19 - C20 - C21 - C22 - C23 - C24 - C25 - C26 - C27 - C28 - C29 - C30 - C31 - C42 - C43 - C44 - C45 - C46 - C47 - C48 - C50 - C52 - C53 - C54 - C55 - C83 - C84 - C85 - C86 - C87 - C89 <= 0
86	BIO10 - C94 <= 0
87	- C94 + PRA1 <= 0
88	- C94 + PRA2 <= 0
89	- C94 + PRA3 <= 0
90	- C94 + PRA4 <= 0
91	- C94 + PRA5 <= 0
92	- C94 + PRA6 <= 0
93	- BIO1 - BIO2 - BIO3 - BIO4 - BIO5 - BIO6 - BIO7 - BIO8 + BIO9 = 0
94	- COLT4 - COLT5 - COLT6 - COLT7 - COLT8 + COLT9 = 0
95	- C98 + FOR4 = 0
96	FOR1 + FOR2 + FOR3 - FOR4 = 0
97	- C101 + SAT8 = 0
98	- C94 + SAU8 = 0
99	C100 - C101 + C94 + C97 + C98 + C99 = 0
100	C56 + C88 + C89 + C93 - C94 = 0
101	C106 + C107 + C108 - C98 = 0
102	C95 + C96 - C97 = 0
103	C90 + C91 + C92 - C93 = 0
104	S57 + S58 + S59 + S61 + S62 + S63 + S64 + S65 + S66 + S67 + S68 + S69 + S70 + S71 + S72 + S73 + S74 + S75 + S76 + S77 + S78 + S79 + S80 + S81 + S82 + S86 + S87 - S88 = 0
105	C57 + C58 + C59 + C60 + C61 + C62 + C63 + C64 + C65 + C66 + C67 + C68 + C69 + C70 + C71 + C72 + C73 + C74 + C75 + C76 + C77 + C78 + C79 + C80 + C81 + C82 + C83 + C84 + C85 + C86 + C87 - C88 = 0

106	C1 + C10 + C102 + C103 + C104 + C105 + C11 + C13 + C14 + C16 + C17 + C18 + C19 + C2 + C20 + C21 + C22 + C23 + C24 + C25 + C26 + C27 + C28 + C29 + C3 + C30 + C31 + C32 + C33 + C34 + C35 + C37 + C38 + C4 + C40 + C41 + C42 + C43 + C44 + C45 + C46 + C47 + C48 + C5 + C50 + C52 + C53 + C54 + C55 - C56 + C6 + C7 + C8 + C9 = 0
107	SAU1 + SAU2 + SAU3 + SAU4 + SAU5 + SAU6 + SAU7 - SAU8 = 0
108	SAT1 + SAT2 + SAT3 + SAT4 + SAT5 + SAT6 + SAT7 - SAT8 = 0

Contributi ISTAT(*)

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*

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 Wînci/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*

(*) ultimi cinque anni

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 – Carlo Vaccari e Daniele Frongia – *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*

8/2008 – Monica Montella – *La nuova matrice dei margini di trasporto*

9/2008 – Antonia Boggia, Marco Fortini, Matteo Mazziotta, Alessandro Pallara, Antonio Pavone, Federico Polidoro, Rosabel Ricci, Anna Maria Sgamba e Angela Seeber – *L'indagine conoscitiva della rete di rilevazione dei prezzi al consumo*

10/2008 – Marco Ballin e Giulio Barcaroli – *Optimal stratification of sampling frames in a multivariate and multidomain sample design*

11/2008 – Grazia Di Bella e Stefania Macchia – *Experimenting Data Capturing Techniques for Water Statistics*

12/2008 – Piero Demetrio Falorsi e Paolo Righi – *A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation*

13/2008 – AA.VV. – *Seminario: Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali*

14/2008 – Francesco Chini, Marco Fortini, Tiziana Tuoto, Sara Farchi, Paolo Giorgi Rossi, Raffaella Amato e Piero Borgia – *Probabilistic Record Linkage for the Integrated Surveillance of Road Traffic Injuries when Personal Identifiers are Lacking*

15/2008 – Sonia Vittozzi – *L'attività editoriale e le sue regole: una ricognizione e qualche proposta per l'Istat editore*

16/2008 – Giulio Barcaroli, Stefania Bergamasco, Michelle Jouvenal, Guido Pieraccini e Leonardo Tininini – *Generalised software for statistical cooperation*

1/2009 – Gianpiero Bianchi, Antonia Manzari, Alessandra Reale e Stefano Salvi – *Valutazione dell'idoneità del software DIESIS all'individuazione dei valori errati in variabili quantitative*