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

Ugo Guarnera, Orietta Luzi, Stefano Salvi

ISTAT

SOMMARIO

1. INTRODUZIONE

In questo documento sono descritte le nuove metodologie per il controllo e la correzione automatica di dati quantitativi su superfici aziendali e consistenza degli allevamenti adottate nell'indagine annuale *Struttura e Produzioni delle Aziende Agricole* (SPA) 2003. La SPA raccoglie informazioni sulle aziende agricole relativamente a superfici per tipo di coltivazioni, tipo e quantità degli allevamenti, tipo di produzioni, struttura e ammontare della manodopera familiare e non (ISTAT, 2001).

Molteplici fattori contribuiscono alla complessità della strategia complessiva di controllo e correzione (C&C nel seguito) per questa indagine: l'elevato numero di variabili osservate, la loro tipologia (sia categoriche sia numeriche continue), le complesse relazioni sia di tipo strutturale sia di tipo statistico/matematico esistenti fra loro. Relativamente agli errori non campionari (Lessler, Kalsbeek, 1992), è necessario sottolineare innanzi tutto che alcuni di essi (in particolare, duplicazioni, errori di codifica, errori di "salto", quadrature) vengono preliminarmente individuati già in fase di *data entry*, attraverso l'uso dello strumento Blaise in cui alcuni vincoli (controlli) di coerenza sono verificati all'atto della registrazione delle informazioni (Ballin *et al*., 2004). Questa strategia garantisce una maggiore accuratezza dei dati rispetto ad un insieme di base di controlli, anche se la non esaustività dei controlli stessi e la possibilità di "forzare" un dato incoerente fanno sì che nei dati registrati permangano situazioni di non accettabilità e quindi di errore. Oltre a questo tipo di errori residuali, altri tipi di errore verificatisi in fase di raccolta o in fase di registrazione dei dati possono contaminare i dati *grezzi* dell'indagine, di natura sia sistematica sia casuale: questi errori possono dar luogo sia a incoerenze logico/statistiche/matematiche, sia a valori anomali, influenti o meno sulle stime dei parametri obiettivo dell'indagine.

Al fine di tenere sotto controllo i diversi elementi di complessità del contenuto informativo dell'indagine, è stata adottata una strategia mista e gerarchica di C&C dei dati, in cui diversi metodi e approcci sono integrati insieme in modo complementare al fine di trattare opportunamente le varie tipologie di errore e di mancata risposta (parziale) individuate nei dati (Guarnera e Luzi, 2004 e 2005). Le maggiori innovazioni sono state introdotte nel trattamento delle variabili relative a superfici aziendali e consistenza degli allevamenti, di tipo numerico continuo. In fase di localizzazione degli errori, il tradizionale approccio deterministico è stato adottato per l'individuazione degli errori di natura sistematica o generati da meccanismi di errore noti, mentre approcci innovativi sono stati adottati per l'individuazione dei valori anomali influenti, e degli errori casuali non influenti. Per quanto riguarda i valori influenti, la loro individuazione è stata basata sull'uso di una metodologia di *editing selettivo* (Latouche *et al*., 1992; Lawrence *et al.,* 2000) in combinazione con tecniche di tipo grafico; gli errori non influenti supposti di origine completamente casuale sono stati invece individuati mediante un approccio automatico di tipo probabilistico basato sulla metodologia Fellegi-Holt (Fellegi and Holt, 1976). Per quanto riguarda la fase di correzione/imputazione, diverse tecniche e approcci sono stati utilizzati a seconda della tipologia e della natura degli errori individuati: correzione automatica deterministica degli errori sistematici, correzione manuale/interattiva dei valori anomali influenti, imputazione automatica con soluzione analitica o con donatore di minima distanza stratificato per gli errori non influenti di natura casuale.

L'intera procedura di C&C è stata implementata in una nuova versione del software AGAIN (*Analisi e Gestione Automatica delle Indagini*), in cui è possibile effettuare l'intero trattamento dei dati in un ambiente completamente integrato, strutturato ed omogeneo. AGAIN è stato sviluppato in SAS dallo staff Istat responsabile delle indagini nel settore dell'agricoltura (Benedetti *et al.,* 2002).

La nuova versione di AGAIN utilizzata per l'indagine SPA consente di gestire in modo integrato tutti i processi di elaborazione dei dati effettuati a valle del processo di data entry. Rispetto alle precedenti versioni del software, le innovazioni introdotte sono relative alla possibilità di effettuare elaborazioni sui dati in parallelo, e la possibilità di aggiornare in modo molto semplice l'insieme dei controlli effettuati sui dati. L'innovazione metodologica principale consiste nell'inclusione in

AGAIN di alcune procedure (*PROC*) del SAS che implementano la metodologia Fellegi-Holt per la localizzazione probabilistica degli errori e le tecniche di imputazione con soluzione analitica o con donatore di minima distanza. Queste procedure sono parte del software *Banff* (Statistics Canada 2003 e 2005), sviluppato in linguaggio SAS da Statistics Canada per il controllo e la correzione di variabili quantitative. Da un punto di vista tecnico, *Banff* rappresenta la versione SAS del software GEIS (*Generalized Edit and Imputation System*), sviluppato però in linguaggio C (Kovar *et al.,* 1988; Cotton, 1991). GEIS utilizza database ORACLE in ambiente Unix ed implementa le medesime metodologie disponibili in *Banff*.

In questo lavoro vengono illustrate le componenti *automatiche* della procedura di C&C relative alla fase di individuazione probabilistica e di imputazione degli errori per le variabili quantitative rilevate dall'indagine SPA 2003. Un breve cenno è riservato al controllo degli errori influenti mediante l'approccio dell'*editing settivo*.

Il lavoro è strutturato nel modo seguente. La sezione 2 contiene una breve descrizione dell'approccio di *editing selettivo*. Nella sezione 3 è descritta brevemente la metodologia probabilistica proposta da Fellegi e Holt per l'individuazione degli errori in variabili di natura quantitativa, e le tecniche utilizzate per la correzione degli errori/l'integrazione delle mancate risposte parziali. Nella sezione 4 è illustrata la procedura complessiva di C&C disegnata per l'indagine SPA 2003. La sezione 5 contiene una descrizione dettagliata delle fasi automatiche della procedura (individuazione probabilistica e imputazione degli errori non influenti).

2. L'APPROCCIO DELL'EDITING SELETTIVO PER L'INDIVIDUAZIONE DEGLI ERRORI INFLUENTI

In generale, data una variabile X ed un parametro θ oggetto di stima sulla distribuzione di X, definiamo influenti rispetto a θ quei valori di X che sono affetti da errore e che hanno un potenziale effetto distorsivo sulla stima di θ. A causa della loro rilevanza statistica, tali valori errati necessitano di essere controllati in modo più accurato rispetto agli altri, cioè manualmente. Al fine di ridurre i tempi e i costi della revisione manuale, è necessario limitare tali controlli alle unità maggiormente critiche. L'approccio noto come *editing selettivo* (Latouche *et al.*, 1992; Di Zio *et al.,* 2002) è stato proposto per l'individuazione dei valori influenti sotto vincoli di costo: tale approccio presuppone l'ordinamento gerarchico delle unità errate (cioè incoerenti rispetto a un prefissato insieme di vincoli) in base al loro impatto potenziale sulle stime dei parametri obiettivo (nel caso della *SPA*, i totali), e la selezione delle prime m unità, secondo questo ordinamento, dove il numero m è funzione di un prefissato livello di accuratezza. Fissato quindi un certo livello di errore residuo accettabile sulle stime, le operazioni di controllo interattivo vengono effettuate solo sul sottoinsieme di unità errate che *spiegano* la corrispondente percentuale di errore totale. In questo modo si ottiene sia di minimizzare i tempi e i costi del controllo interattivo per il prefissato livello di accuratezza, sia di ridurre il fenomeno dell'*over-editing* (risorse spese per il controllo interattivo di errori con effetto trascurabile sulle stime). E' chiaro che in questo approccio il ruolo fondamentale è svolto dal criterio utilizzato per "stimare" l'impatto sulle stime obiettivo delle unità affette da errore. In generale, tale criterio viene espresso da una *funzione punteggio* che assegna ad ogni unità un valore che tiene conto, tra gli altri elementi, dell'entità dell'errore (o degli errori) di cui è responsabile l'unità stessa e del peso campionario. Naturalmente, poiché in genere si è interessati a stime su diverse variabili, funzioni punteggio globali possono essere ottenute come sintesi di più funzioni punteggio relative a singole variabili.

3. L'APPROCCIO PROBABILISTICO ALL'INDIVIDUAZIONE E ALLA CORREZIONE DEGLI ERRORI CASUALI NON INFLUENTI IN DATI QUANTITATIVI

La metodologia Fellegi-Holt (*FH* nel seguito) e la tecnica *hot deck* nota come *donatore di distanza minima* (*nearest-neighbor donor, NND* nel seguito) descritte in questa sezione sono quelle disponibili nel software Geis/Banff (Kovar *et al.*, 1988; Cotton, 1991) per il controllo e la correzione automatica di variabili numeriche continue. Questi metodi sono particolarmente adatti al trattamento di errori di origine completamente casuale non influenti in termini di impatto sulle stime dei parametri obiettivo.

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

Dal momento che gli errori sono individuati sulla base di un prefissato insieme di edit, quando si adotta la metodologia FH è necessaria la massima cautela nella definizione delle regole di controllo. Infatti, se viene utilizzato un insieme con regole poco numerose ma soprattutto poco connesse (in termini di variabili coinvolte in esse), l'algoritmo potrebbe determinare una scelta sostanzialmente casuale delle variabili da modificare. In tali casi, un approccio di tipo deterministico potrebbe risultare preferibile. D'altra parte, troppi vincoli di coerenza potrebbero dare origine ad eccessiva complessità del problema di localizzazione dell'errore, con conseguente impossibilità per l'algoritmo a determinare una soluzione. Per questo motivo, in alcune applicazioni particolarmente complesse si ricorre alla suddivisione degli edit in due o più sottoinsiemi, ed alla loro applicazione sequenziale ai dati in una prefissata gerarchia. Solo se nessuna variabile è coinvolta in edit appartenenti ai diversi sottogruppi, l'applicazione dei diversi sottoinsiemi di regole (e quindi dell'algoritmo probabilistico) produce risultati che non dipendono dalla sequenza con cui essi sono verificati sui dati. In caso contrario, se cioè una o più variabili compaiono in diversi sottoinsiemi di vincoli, la sequenza di applicazione influenza il risultato, e le variabili trattate nell'ambito di un sottoinsieme di edit devono essere rese non modificabili (attraverso l'uso dei pesi) nei passi successivi di controllo.

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

Per quanto riguarda la tecnica del *donatore di minima distanza* (*nearest-neighbor donor, NND*), per ogni unità *i* che viola almeno un *edit* (unità *ricevente*), i valori mancanti o classificati come errati per una o più variabili sono sostituiti congiuntamente con i valori delle stesse variabili osservati nell'unità più vicina d_i (*donatore*). Il donatore d_i è selezionato da un *serbatoio* di unità complete (cioè prive di valori mancanti) e coerenti (cioè che soddisfano tutti gli *edits*) in base alla funzione di distanza *minmax* calcolata rispetto a un insieme di *variabili di accoppiamento* (o *matching*) trasformate mediante la funzione *rango*. Una importante caratteristica del metodo adottato consiste nel fatto che l'imputazione avviene nel rispetto dei vincoli: per ogni unità *ricevente*, il donatore d_i viene utilizzato per l'imputazione solo se l'unità risultante diviene coerente

rispetto a tutti gli *edits*. Nella pratica, nel metodo NND disponibile in Banff il donatore d_i viene utilizzato per l'imputazione solo se l'unità risultante diviene coerente rispetto a tutti gli *edits di post-imputazione*: questi *edits* (detti *post-edits*) corrispondono a vincoli originali opportunamente resi meno stringenti (ad esempio, vincoli di quadratura rilassati in modo da ammettere come validi valori in un intorno del valore esatto di confronto - il totale). L'uso dei *post-edits* ha l'obiettivo di allargare la rosa dei potenziali donatori. Infatti, dato un record *i* che fallisce una certa uguaglianza, può accadere che il sistema:

- non riesca a trovare un donatore che fornisca ad *i* un valore tale da riportarlo nella condizione di soddisfare l'uguaglianza;
- scarti record donatori "vicini" ad *i*, ma che non gli garantiscono il rispetto dell'uguaglianza, e seleziona un donatore più "distante" da *i*, ma che gli fornisce il valore richiesto.

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

In generale, le tecniche di imputazione (incluso il *NND*) risultano più efficienti se esse vengono applicate all'interno di *celle di imputazione*: si tratta di sotto-popolazioni definite sulla base di variabili (*covariate*) statisticamente associate alle variabili oggetto di correzione/imputazione, e quindi ritenute maggiormente omogenee in termini del/dei fenomeno/i oggetto di imputazione.

I metodi da donatore (o *hot deck*), sono riconosciuti fra i più adatti alla ricostruzione delle mancate risposte in termini di preservazione della variabilità delle distribuzioni semplici delle variabili imputate e delle associazioni fra esse, se di tali associazioni si tiene conto nel modello di imputazione (in fase di stratificazione e/o calcolo della distanza e/o adottando modalità di imputazione congiunta delle variabili con uno stesso donatore) (Kalton e Kasprzyk, 1982; Chen e Shao, 2000)

4. LA PROCEDURA COMPLESSIVA DI CONTROLLO E CORREZIONE PER L'INDAGINE SPA 2003

Nella nuova procedura di C&C predisposta per l'indagine SPA2003 possiamo distinguere le seguenti fasi principali:

1) Controllo quantitativo dei modelli pervenuti
2) Controlli qualitativi preliminari sulle variabili identificative dei modelli e sulle caratteristiche principali delle unità campione
3) Controllo qualitativo delle variabili quantitative principali del questionario
4) Controllo qualitativo delle altre variabili quantitative del questionario
5) Controllo qualitativo delle altre variabili qualitative del questionario
6) Controllo qualitativo della sezione LAVORO

Ciascuna fase è stata effettuata (come generalmente accade) con approcci e metodi di tipo diverso: alcune fasi conterranno controlli di tipo esclusivamente deterministico (ad esempio, i passi 1, 2), in altre fasi sarà necessario l'uso combinato di rappresentazioni grafiche e funzioni selettive (ad esempio le fasi 3 e 4), in altre fasi verrà utilizzato anche l'approccio probabilistico (ad esempio nelle fasi 3,4 e 6).

1) *Controllo quantitativo dei modelli pervenuti*
Questi controlli hanno lo scopo principale di verificare quanto materiale è stato raccolto rispetto a quello pianificato soprattutto per strati, tipicamente territoriali (regioni, provincia, comuni), ma anche ad esempio per rilevatore, per lotto di registrazione ecc.

2) *Controlli qualitativi preliminari sulle variabili identificative dei modelli e sulle caratteristiche principali delle unità campione*

Un tipico obiettivo di questi controlli è accertarsi che il modello corrisponda ad una ed una sola unità statistica e sia correttamente identificato.
Altri controlli riguardano la coerenza fra loro delle caratteristiche strutturali dell'azienda. Tipicamente, si marcano opportunamente tutti i pattern di incoerenza per la costruzione di tabelle di frequenza sia per la verifica della presenza di errori sistematici sia per l'impostazione delle azioni di correzione.

3) *Controllo e correzione delle variabili quantitative del questionario*

Questa fase consiste a sua volta di più sottoprocessi di C&C dei dati.

a) <u>Controllo degli errori sistematici, anomali, influenti</u>

Questo controllo ha l'obiettivo da un lato di fissare punti di riferimento importanti per il controllo delle altre variabili quantitative, dall'altro di trattare gerarchicamente prima variabili cruciali del questionario. Nel caso dell'indagine in oggetto, in questa fase si sottopongono a controllo le principali variabili di superficie dell'azienda e quelle più significative relative agli allevamenti, per le quali si dispone tra l'altro di informazioni censuarie affidabili. Tipici controlli sono:

- controllo degli errori sistematici, per i quali sono state predisposte procedure automatiche di correzione deterministica;
- controllo dei valori anomali e degli errori influenti, per i quali sono state predisposte opportune procedure di individuazione, e il controllo manuale interattivo per l'eventuale correzione.

b) <u>Controllo e correzione automatica degli errori casuali *non* influenti</u>

Obiettivo di questa fase è risolvere in modo automatico (quindi poco costoso in termini sia di tempo che di risorse) tutte le incoerenze logico-matematiche presenti nei dati ma che sono considerate poco influenti in quanto hanno superato tutti i controlli precedenti. L'algoritmo utilizzato è quello probabilistico di FH implementato in Banff. Le regole di controllo utilizzate sono quadrature di tabelle, esistenze incrociate fra variabili collegate fra loro nel modello, uguaglianze fra valori osservati o ricavabili in più punti del modello, disuguaglianze.

Il trattamento è stato effettuato separatamente per gruppi di variabili distinte: superfici da un lato, allevamenti dall'altro. Per quanto riguarda le superfici e le relative coltivazioni/produzioni, data la quantità delle variabili presenti e la complessità delle relazioni fra esse, si propende per un trattamento gerarchico: prima le *variabili principali* del modello (sezioni I e II e parte della III), poi altre variabili quantitative di altre sezioni del questionario relative a superfici e produzioni (*variabili secondarie*) per le quali è possibile utilizzare l'approccio probabilistico. Le variabili quantitative definite *secondarie* sono quelle relative alle sezioni:

- *Successioni colturali*
- *Superfici soggette a regime aiuto*
- *Contoterzismo*
- *Altre notizie sulla floricoltura*
- *Altre notizie sulle ortive*
- *Produzione di qualità*

- *Pratiche colturali*
- *Quesiti aggiuntivi di interesse regionale*

Dal momento che le variabili secondarie sono legate da relazioni con le variabili principali, una volta controllate e corrette le variabili principali, queste ultime sono state tenute fisse nei passi di localizzazione degli errori.
La fase di controllo automatico probabilistico prevede evidentemente che tutte le relazioni fra le variabili del questionario siano individuate ed esplicitate. Si tratta di un lavoro impegnativo data la complessità del questionario, che è stato effettuato sulla base dei legami fra variabili derivanti sia dall'analisi del modello, sia dalla conoscenza dei fenomeni coinvolti.

In questa fase del processo, tutte le coerenze vengono ripristinate mediante tecniche di correzione o di imputazione vera e propria (correzione deterministica, con soluzione analitica, imputazione con donatore di minima distanza, ecc.) a seconda delle tipologie degli errori (sistematici o casuali), del tipo di vincoli violati e dei pattern di errore riscontrati nei dati.

Per la fase di imputazione è stato inoltre necessario individuare criteri di similitudine (ad esempio, la ripartizione territoriale delle zone in cui si trovano le aziende agricole, l'essere state censite o campionate, ecc.) per la costruzione di serbatoi di donatori.

4) *Controllo delle altre variabili quantitative del questionario*

Obiettivo di questa fase è completare il controllo degli errori sulle altre variabili quantitative del modello per cui non è possibile adottare l'approccio probabilistico in quanto per esse non è possibile stabilire relazioni di coerenza con altre informazioni presenti nel modello. Per queste variabili sono state predisposti controlli di tipo deterministico (quindi di tipo gerarchico) per cui, ad una certa tipologia di errore, corrisponde un certo insieme di variabili errate e la corrispondente azione di correzione.
Le variabili quantitative trattate secondo questo approccio sono quelle relative alle sezioni:

- *Commercializzazione*
- *Degrado del territorio*
- *Lavorazione del terreno*
- *Vendita dei prodotti dell'azienda*
- *Sezione VII – Attività connesse all'agricoltura*

5) *Controllo delle altre variabili qualitative del questionario*

I quesiti di tipo qualitativo presenti nel questionario dell'indagine non si prestano ad un trattamento di tipo probabilistico utilizzando le tecniche disponibili in Istat. In tale situazione, per il controllo e la correzione sono stati predisposti criteri di controllo e di correzione di tipo deterministico studiati *ad hoc*.

6) *Controllo della sezione LAVORO*

In questa sezione del questionario vengono rilevate variabili di tipo sia categorico (caratteristiche demografiche delle persone impiegate nell'azienda), sia di tipo numerico continuo (principalmente, giornate lavorate in azienda). Pertanto, per il controllo di coerenza di questa sezione sono necessari sia i classici controlli fra caratteristiche individuali (ad es. coerenza fra sesso dei coniugi, date di nascita ecc.), sia controlli di coerenza sulle giornate lavorate, complessivamente e per tipologia di manodopera impiegata, tenendo anche conto ad esempio della grandezza dell'azienda in termini di

superficie. Anche in questo caso,è stato adottato un approccio di tipo deterministico per il controllo degli errori, l'individuazione delle variabili errate e la loro correzione.

5. LA PROCEDURA AUTOMATICA DI CONTROLLO E CORREZIONE DELLE VARIABILI QUANTITATIVE SU SUPERFICI, ALLEVAMENTI E RELATIVE PRODUZIONI

Obiettivo di questa sezione è illustrare le fasi automatiche della procedura di C&C per le variabili quantitative su superfici aziendali e consistenza degli allevamenti (in particolare, individuazione probabilistica e correzione/imputazione degli errori casuali non influenti).

In generale, la procedura di C&C per le variabili quantitative dell'indagine SPA è strutturata in più passi principali, relativo ciascuno ad un sottoinsieme di variabili quantitative. Ciascun passo è a sua volta composto da sottofasi relative ciascuna ad una tipologia di errori ed a uno specifico tipo di trattamento (individuazione o correzione degli errori, interattivo o automatico).

I passi principali e le relative sottofasi sono illustrati di seguito.

FASE 1 - Controllo e correzione delle superfici e delle coltivazioni principali del questionario

- controllo e correzione manuale-interattiva di valori anomali e degli errori influenti
- controllo automatico probabilistico degli errori residui
- correzione automatica con soluzione analitica
- imputazione automatica con donatore di minima distanza
- correzione manuale-interattiva dei casi non risolti dal probabilistico o non imputati con successo

FASE 2 - Controllo e correzione delle superfici e delle coltivazioni secondarie del questionario
(fissate le variabili trattate al passo precedente)

- controllo automatico probabilistico degli errori
- correzione automatica con soluzione analitica
- imputazione automatica con donatore di minima distanza
- correzione manuale-interattiva dei casi non risolti dal probabilistico o non imputati con successo

FASE 3 - Controllo e correzione delle variabili su allevamenti e relative produzioni

- controllo e correzione manuale-interattiva di valori anomali ed errori influenti
- controllo automatico probabilistico degli errori
- correzione automatica con soluzione analitica
- imputazione automatica con donatore di minima distanza
- correzione manuale-interattiva dei casi non risolti dal probabilistico o non imputati con successo

Nelle sezioni 5.1, 5.2 e 5.3, le caratteristiche delle componenti automatiche delle fasi 1, 3 verranno analizzate in dettaglio.

5.1. Controllo e correzione automatica degli errori casuali sulle variabili principali relative a superfici e produzioni (Sezione I)

La procedura automatica per l'individuazione probabilistica degli errori casuali e la loro correzione/imputazione relativamente alle 274 variabili quantitative definite come *principali* del

questionario SPA2003 è schematizzata nella figura 1 dell'Allegato 1. I dati sono stati elaborati separatamente per regione.

Per ogni regione R_k, i dati sottoposti a controllo automatico sono stati preliminarmente depurati dagli eventuali valori anomali e dagli errori influenti su alcune variabili critiche (SAU e SAT) con una procedura basata su un approccio di tipo selettivo (vedi paragrafo 2).

Successivamente, mediante correzioni di tipo deterministico, i dati sono stati depurati da alcuni errori sistematici individuati attraverso l'analisi delle frequenze di violazione dei vincoli definiti per le variabili oggetto del controllo.

Sui dati così pre-elaborati, la prima fase di localizzazione probabilistica degli errori è stata effettuata utilizzando un opportuno insieme di vincoli lineari o linearizzati (quadrature, esistenze, inclusioni, uguaglianze) del tipo illustrato nell'Allegato 2. Si tratta di 133 vincoli, di cui 12 quadrature e 121 disuguaglianze, che coinvolgono le 274 variabili trattate in questa fase. Le variabili hanno tutte uguale peso (1 per *default*), tranne le variabili della *sezione III – Irrigazione* trattate in questa fase[1], aventi tutte peso pari a 0,5: questo equivale a dire che, se un ogni record errato viola vincoli in cui sono coinvolte le variabili sull'irrigazione, non tutte le variabili hanno la stessa probabilità di essere incluse nella soluzione di minimo cambiamento, le variabili sull'irrigazione hanno probabilità maggiore di essere selezionate come errate.

I parametri utilizzati in questa fase, oltre ovviamente ai vincoli di coerenza, sono:

- la massima cardinalità C ammessa per la soluzione di minimo cambiamento, cioè il massimo numero di variabili selezionabili come errate nella soluzione di minimo cambiamento dell'algoritmo;
- il massimo tempo disponibile T, per ognuno dei record errati, perché l'algoritmo individui una possibile soluzione (entro il limite di massima cardinalità assegnato).

Nella prima fase di localizzazione con metodologia FH sono stati posti C=8, T=45.
Generalmente, a causa del limite imposto su C o su T, accade che per un sottoinsieme di unità errate (cioè che violano almeno un vincolo), nessuna soluzione sia individuata. Un secondo passo di localizzazione è stato pertanto effettuato su questo sottoinsieme di dati con nuovi parametri C=8, T=100[2]. Le unità non risolte in questo secondo passo (*Unità non risolte 1*) rappresentano il primo gruppo di aziende da sottoporre a revisione manuale per questo gruppo di variabili.

Tutte le unità risolte al primo o al secondo passo di localizzazione degli errori sono state sottoposte a un passo di correzione con soluzione analitica: per ogni unità errata, sulla base dei vincoli definiti e delle variabili etichettate come errate dall'algoritmo di localizzazione degli errori, un algoritmo verifica se per qualche variabile errata esiste uno ed un solo valore possibile. Al termine di tale fase alcune unità saranno state completamente risolte, cioé non presenteranno valori errati residui, mentre per altre osservazioni sarà necessario procedere all'imputazione del sottoinsieme residuo di valori errati.
Questo passo di imputazione è stato effettuato utilizzando la tecnica del donatore di distanza minima disponibile in *Banff* (vedi sezione 3). Imputazioni distinte sono state fatte all'interno di strati (*celle di imputazione*) definiti, per ogni regione, dal tipo di selezione (unità "censite" o unità "campionate") e dalla forma di conduzione. Per ogni strato, la tecnica di imputazione è stata applicata secondo la seguente strategia ciclica (vedi figura 1 dell'Allegato 1):

1. imputazione con donatore utilizzando: un insieme di edit uguale a quello iniziale, ma con i vincoli di quadratura rilassati del 5%[3]; un insieme di *post-edit* (vedi sezione 3) corrispondente a quello iniziale, ma senza le quadrature. Questo per evitare che la necessità di imputare valori che garantiscano il rispetto delle quadrature possa impedire del tutto l'individuazione di

[1] ir0, ir1, ir2, ir3, ir4, ir5, ir6, ir7, ir8, ir9, ir10, ir11, ir12, ir13, ir14, ir15, ir16

[2] La cardinalità C non è stata modificata perché unità con più di 8 variabili da modificare sono state considerate particolarmente critiche, quindi da rivedere in modo interattivo.

[3] Ogni quadratura del tipo x+y=z è stata sostituita dalle due disuguaglianze x+y<=1.5*z e x+y>=0.5*z, che definiscono un intorno di Z di ampiezza ±0.5.

donatori adatti. Le unità per le quali risulta comunque impossibile individuare un donatore costituiscono il secondo gruppo di aziende (*Unità non risolte 2*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.

2. Per le unità con donatore individuato (*Unità risolte*), nuovo passo di localizzazione rispetto agli edit iniziali per garantire la coerenza dei dati rispetto ai vincoli nella loro forma originaria (in particolare, le quadrature).
3. Correzione con soluzione analitica utilizzando l'insieme iniziale di vincoli. Le unità non completamente risolte costituiscono il terzo gruppo di aziende (*Unità non risolte 3*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.

Al termine di questa procedura, tutte le unità coerenti rispetto ai vincoli sulle variabili principali (*Unità Risolte 1* e *Unità Risolte 2* nella Figura 1 dell'Allegato1) sono riunite in un unico data set (*Dati output Fase 1*) per il controllo successivo rispetto agli altri vincoli riguardanti superfici, coltivazioni, produzioni. Tutte le aziende per le quali invece non è stato possibile completare il processo di controllo (*Unità non risolte 1*) o di imputazione (*Unità non risolte 2* e *Unità non risolte 3*) costituiscono il primo insieme di osservazioni, per la regione R_k, su cui è necessario procedere al controllo e alla correzione manuale/interattivo.

5.2. Controllo e correzione automatica degli errori casuali sulle variabili secondarie relative a superfici e produzioni (Sezione II)

La procedura automatica di individuazione probabilistica degli errori e di correzione/imputazione predisposta per le variabili quantitative definite come *secondarie* del questionario SPA2003 è schematizzata nella figura 2 dell'Allegato 1.

Sempre procedendo per regione, i dati sottoposti a controllo automatico sono quelli provenienti dalla prima fase del controllo automatico (*Dati output Fase 1*), in cui le variabili principali del questionario sono coerenti e complete rispetto ai vincoli utilizzati nel primo passo della procedura.

Mediante alcune correzioni deterministiche, i dati in input vengono preliminarmente depurati da alcuni errori sistematici sulle variabili secondarie oggetto del controllo individuati attraverso l'analisi delle frequenze di violazione dei vincoli definiti per le variabili oggetto del controllo.

Sui dati così pre-trattati, la prima fase di localizzazione probabilistica degli errori viene effettuata utilizzando un insieme di 63 vincoli (di cui 1 quadratura), che coinvolgono 235 variabili, fra cui 81 nuove variabili rispetto al passo relativo alle variabili principali su superfici e coltivazioni. Le 154 variabili principali coinvolte nei vincoli (che hanno forma del tipo descritto nell'Allegato 2), ma trattate nella prima fase della procedura, sono rese non modificabili attraverso l'uso dei pesi (vedi sezione 3): ad esse è assegnato un peso molto elevato che ne impedisce la selezione come variabili errate in questa fase. Pertanto per ogni record errato solo le nuove 81 variabili possono essere incluse, con uguale probabilità, nella soluzione di minimo cambiamento qualora coinvolte in edit violati.

I parametri utilizzati nella fase di localizzazione probabilistica di questo gruppo di variabili, oltre ovviamente ai vincoli di coerenza, sono stati:

- massima cardinalità C ammessa per la soluzione di minimo cambiamento C=12;
- il massimo tempo disponibile T =100.

Anche in questo caso, a causa del limite imposto su C o su T, può accadere che nessuna soluzione venga individuata per un sottoinsieme di unità errate della regione. Le unità non risolte in

questo passo (*Unità non risolte 1*) rappresentano il primo gruppo di aziende da sottoporre a revisione manuale per questo gruppo di variabili.

Tutte le unità con errori individuati sono state invece sottoposte a un passo di correzione con soluzione analitica. Al termine di tale fase alcune unità vengono completamente risolte (*Unità Risolte 1*), cioè non presentano valori errati residui, mentre per altre osservazioni è necessario procedere all'imputazione del sottoinsieme residuo di valori errati.

Come per le superfici principali, anche in questo caso le imputazioni mediante la tecnica del donatore di distanza minima sono state effettuate all'interno di strati distinti definiti in modo analogo al caso delle variabili principali. Per ogni strato, la tecnica di imputazione è stata applicata secondo la seguente strategia (vedi Figura 2 dell'Allegato 1):

1. imputazione con donatore utilizzando: un insieme di edit uguale a quello iniziale con associato un insieme di *post-edit* corrispondente a quello iniziale, ma con i vincoli di quadratura rilassati del 5%. Le unità per le quali risulta impossibile individuare un donatore costituiscono il secondo gruppo di aziende (*Unità non risolte 2*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.
2. Per le unità con donatore individuato (*Unità risolte*), nuovo passo di localizzazione rispetto agli edit iniziali per garantire la coerenza dei dati rispetto ai vincoli nella loro forma originaria (in particolare, le quadrature).
3. Correzione con soluzione analitica utilizzando l'insieme iniziale di vincoli. Le unità non completamente risolte costituiscono il terzo gruppo di aziende (*Unità non risolte 3*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.

Al termine di questa procedura, tutte le unità coerenti rispetto ai vincoli sulle variabili secondarie (*Unità Risolte 1* e *Unità Risolte 2* nella Figura 2 dell'Allegato1) sono riunite in un unico data set (*Dati output Fase 2*) per il controllo successivo rispetto agli altri vincoli riguardanti gli allevamenti e le relative produzioni.

Tutte le aziende per le quali invece non è stato possibile completare il processo di controllo (*Unità non risolte 1*) o di imputazione (*Unità non risolte 2* e *Unità non risolte 3*) costituiscono il secondo insieme di osservazioni, per la regione R_k, su cui è necessario procedere al controllo e alla correzione manuale/interattive.

Al termine di questa fase di C&C automatica, tutte le variabili relative a superfici, coltivazioni e produzioni trattabili con approccio probabilistico sono coerenti e complete rispetto ai vincoli utilizzati.

5.3. Controllo e correzione automatica degli errori sulle variabili relative ad allevamenti e relative produzioni (Sezione III)

La procedura automatica di individuazione probabilistica degli errori e di correzione/imputazione predisposta per le variabili quantitative relative agli allevamenti e relative produzioni nel questionario SPA 2003 è schematizzata nella Figura 3 dell'Allegato 1.

Sempre procedendo per regione, i dati sottoposti a controllo automatico sono quelli provenienti dalla seconda fase del controllo automatico (*Dati output Fase 2*), in cui le variabili del questionario relative a superfici, relative coltivazioni e produzioni sono coerenti e complete rispetto ai vincoli utilizzati.

Anche in questo passo della procedura sono state effettuate alcune correzioni di tipo deterministico per depurare i dati in input da alcuni errori sistematici sulle variabili oggetto del controllo individuati attraverso l'analisi delle frequenze di violazione dei vincoli definiti per le variabili oggetto del controllo.

Sui dati così pre-trattati, la prima fase di localizzazione probabilistica degli errori viene effettuata utilizzando opportuni vincoli lineari o linearizzati (equazioni e disequazioni della forma generale descritta nell'Allegato 2). Si tratta di 53 vincoli (di cui 23 quadrature o uguaglianze), che coinvolgono 133 variabili. In questo caso, tutte le variabili hanno la medesima probabilità di essere incluse nella soluzione di minimo cambiamento, in quanto tutte hanno peso (di *default*) pari a 1.

I parametri utilizzati nella fase di localizzazione probabilistica per questo gruppo di variabili sono:

- massima cardinalità C ammessa per la soluzione di minimo cambiamento C=10;
- il massimo tempo disponibile T =20.

Anche in questo caso, a causa del limite imposto su C o su T, può accadere che nessuna soluzione venga individuata per un sottoinsieme di unità errate della regione. Le unità non risolte in questo passo (*Unità non risolte 1*) rappresentano il primo gruppo di aziende da sottoporre a revisione manuale per questo gruppo di variabili.

Tutte le unità con errori individuati sono state invece sottoposte a un passo di correzione con soluzione analitica. Al termine di tale fase alcune unità vengono completamente risolte (*Unità Risolte 1*), cioè non presentano valori errati residui, mentre per altre osservazioni è necessario procedere all'imputazione del sottoinsieme residuo di valori errati.

Come per le altre variabili del questionario, anche in questo caso le imputazioni sono state effettuate mediante la tecnica del donatore di distanza minima, all'interno di strati distinti definiti in modo analogo al caso delle variabili su superfici, coltivazioni e relative produzioni. Per ogni strato, la tecnica di imputazione è stata applicata secondo la seguente strategia (vedi Figura 2 dell'Allegato 1):

1. imputazione con donatore utilizzando: un insieme di edit uguale a quello iniziale, ma con i vincoli di quadratura rilassati del 5%; un insieme di *post-edit* (vedi sezione 3) corrispondente a quello iniziale, ma senza le quadrature. Questo per evitare che la necessità di imputare valori che garantiscano il rispetto delle quadrature possa impedire del tutto l'individuazione di donatori adatti. Le unità per le quali risulta comunque impossibile individuare un donatore costituiscono il secondo gruppo di aziende (*Unità non risolte 2*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.
2. Per le unità con donatore individuato (*Unità risolte*), nuovo passo di localizzazione rispetto agli edit iniziali per garantire la coerenza dei dati rispetto ai vincoli nella loro forma originaria (in particolare, le quadrature).
3. Correzione con soluzione analitica utilizzando l'insieme iniziale di vincoli. Le unità non completamente risolte costituiscono il terzo gruppo di aziende (*Unità non risolte 3*) su cui effettuare una revisione di tipo manuale/interattivo per questo gruppo di variabili.

Al termine di questa procedura, tutte le unità coerenti rispetto ai vincoli sulle variabili secondarie (*Unità Risolte 1* e *Unità Risolte 2* nella Figura 3 dell'Allegato 1) sono riunite in un unico data set (*Dati output Fase 3*) contenente i dati coerenti e completi rispetto a tutti i vincoli utilizzati su tutte la variabili su superfici e allevamenti e relative produzioni trattate con approccio probabilistico.

Tutte le aziende per le quali invece non è stato possibile completare il processo di controllo (*Unità non risolte 1*) o di imputazione (*Unità non risolte 2* e *Unità non risolte 3*) costituiscono il terzo insieme di osservazioni, per la regione R_k, su cui è necessario procedere al controllo e alla correzione manuale/interattiva.

6. CONCLUSIONI

La strategia di C&C per la nuova indagine SPA può essere considerata un esempio di processo tipo per tutte le indagini che rilevano dati di natura quantitativa. In essa si riconosce una struttura "standard" in cui il flusso delle operazioni di trattamento dei dati è organizzato in modo tale da consentire un bilanciamento ottimale fra tempi di elaborazione, qualità attesa del risultato raggiunto, costi (in termini di risorse spese nel corso del processo e di peso sui rispondenti). In questa struttura, le unità anomale oppure affette da errori rilevanti (in termini di impatto sulle stime di interesse sulle variabili obiettivo) vengono individuate e trattate per prime: su di esse vengono concentrate le attività di revisione interattiva e/o di re-intervista, in modo da ridurre tempi, costi e peso sui rispondenti connessi alla fase di trattamento manuale/interattivo dei dati. Gli errori di natura sistematica, per i quali cioè sia possibile individuare la causa (o meccanismo) che li genera possono essere trattati automaticamente in modo ottimale mediante procedure di tipo deterministico (in cui cioè a fronte di un certo errore si specifichi direttamente l'azione di correzione) poco costose da implementare, manutenere e aggiornare. Per l'individuazione degli altri errori, cioè degli errori di natura casuale non influenti sui parametri oggetto di stima, un approccio automatico di tipo probabilistico garantisce qualità del risultato finale (in termini di probabilità di corretta individuazione e minimalità di valori modificati), e tempi ridotti, nonché semplicità di aggiornamento e manutenzione nelle successive occasioni di indagine. La strategia di imputazione automatica degli errori casuali e delle mancate risposte (effettuata scegliendo i metodi più opportuni, di natura parametrica o non parametrica) deve infine fornire garanzie di preservazione delle proprietà statistiche dei dati di interesse per la particolare indagine.

Pertanto, la realizzazione di una procedura di C&C di tipo misto, in cui da un lato il flusso dei controlli è basato sulla natura degli errori e sulle caratteristiche delle variabili oggetto del trattamento, dall'altro sono integrati strumenti di natura diversa, tra cui strumenti generalizzati, implica una revisione complessiva e profonda del processo di trattamento degli errori non campionari sui dati osservati (Di Zio e Luzi, 2002b), realizzabile in caso di processi in via di radicale ristrutturazione (come nel caso dell'indagine SPA) o di prima implementazione.

E' opportuno inoltre evidenziare alcuni valori aggiunti legati all'utilizzo dello strumento generalizzato Banff.

Innanzi tutto, questo strumento consente di effettuare un'analisi complessiva di coerenza e non contraddittorietà delle regole utilizzate nel piano di C&C dei dati, consentendo di individuare controlli superflui, oppure contraddittori, oppure mal specificati. Questo consente di tenere sotto controllo non solo la struttura dei controlli applicati ai dati, ma anche il fenomeno dell'*over-editing*, che consiste nel verificare la coerenza dei dati rispetto a vincoli che producono modifiche non rilevanti nei dati senza però contribuire ad aumentare il livello complessivo di qualità dei dati.

Inoltre, mediante l'analisi dei report prodotti è possibile ottenere utili informazioni sulle caratteristiche degli errori presenti nei dati e sulla loro possibile origine (sistematica o casuale). In particolare, dallo studio delle frequenze di attivazione delle regole di controllo utilizzate nel piano di compatibilità è possibile evidenziare una serie di problemi connessi all'adeguatezza ed alla formulazione degli edit ed alla presenza nei dati di eventuali errori di origine sistematica. Alte frequenze di violazione possono infatti segnalare la presenza di controlli non appropriati o mal formulati, oppure di errori di origine non casuale che richiedono interventi sull'organizzazione dell'indagine (istruzione ai rilevatori, piani di registrazione).

Infine, l'uso di strumenti generalizzati implica che i costi di disegno e implementazione siano concentrati nella prima occasione di indagine, abbattendo però drasticamente i costi di aggiornamento e manutenzione nelle successive occasioni di indagine (Di Zio e Luzi, 2002a).

RIFERIMENTI BIBLIOGRAFICI

Ballin M., Guarnera U., Luzi O., Salvi S. (2004), 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.

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

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

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

Di Zio M., Luzi O. (2002a), Comparing a purely deterministic and a semi-probabilistic approach for editing and imputation of Agricultural Census data, *Scritti di Statistica Economica*, **9**, Settembre 2002.

Di Zio M., Luzi O. (2002b), Combining Methodologies in a Data Editing Procedure: an Experiment on the Survey of Balance Sheets of Agricultural Firms, *Statistica Applicata*, **14**, 1, pp. 59-80.

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

Guarnera U., Luzi O. (2004), 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.

Guarnera U., Luzi O. (2005), Valutazione del trattamento degli errori di misura e di risposta nell'indagine SPA. *Convegno AGR@STAT: Verso un nuovo sistema di statistiche agricole*, Università degli Studi di Firenze, 30-31 Maggio.

ISTAT (2001), *Struttura e Produzioni delle Aziende Agricole – Anno 1999 - Italia*. Collana Informazioni ISTAT.

Latouche M., Berthelot J.M. (1992), Use of Score Functions to Prioritise and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics*, 8, 3, Part II.

Lawrence, D., and McKenzie, R. (2000), The General Application of Significance Editing. *Journal of Official Statistics*, **16**, 243-253.

Lessler J.T., Kalsbeek W.D. (1992), *Non Sampling Errors in Surveys*, New York: Wiley.

Statistics Canada (2003) *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 (2005) *Banff Users Guide, Version 1.04*, Generalized Systems Methods Section, Business Survey Methods Division.

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

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

ALLEGATO 1 – FLUSSO DELLA PROCEDURA DI CONTROLLO E CORREZIONE INDAGINE SPA 2003

Figura 1. Flusso della procedura – Superfici e Coltivazioni Principali

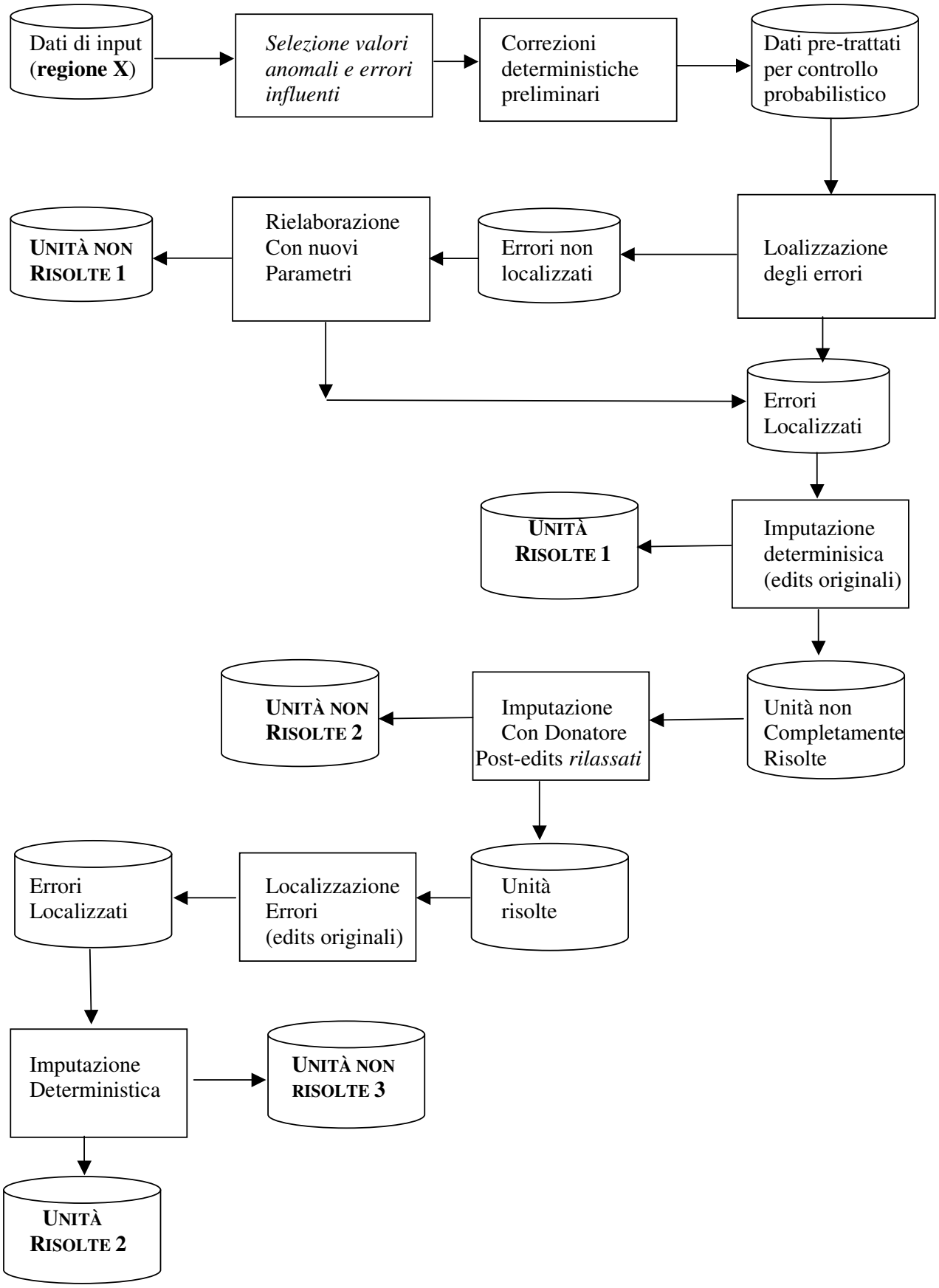


Figura 2. Flusso della procedura – Superfici e Coltivazioni Secondarie

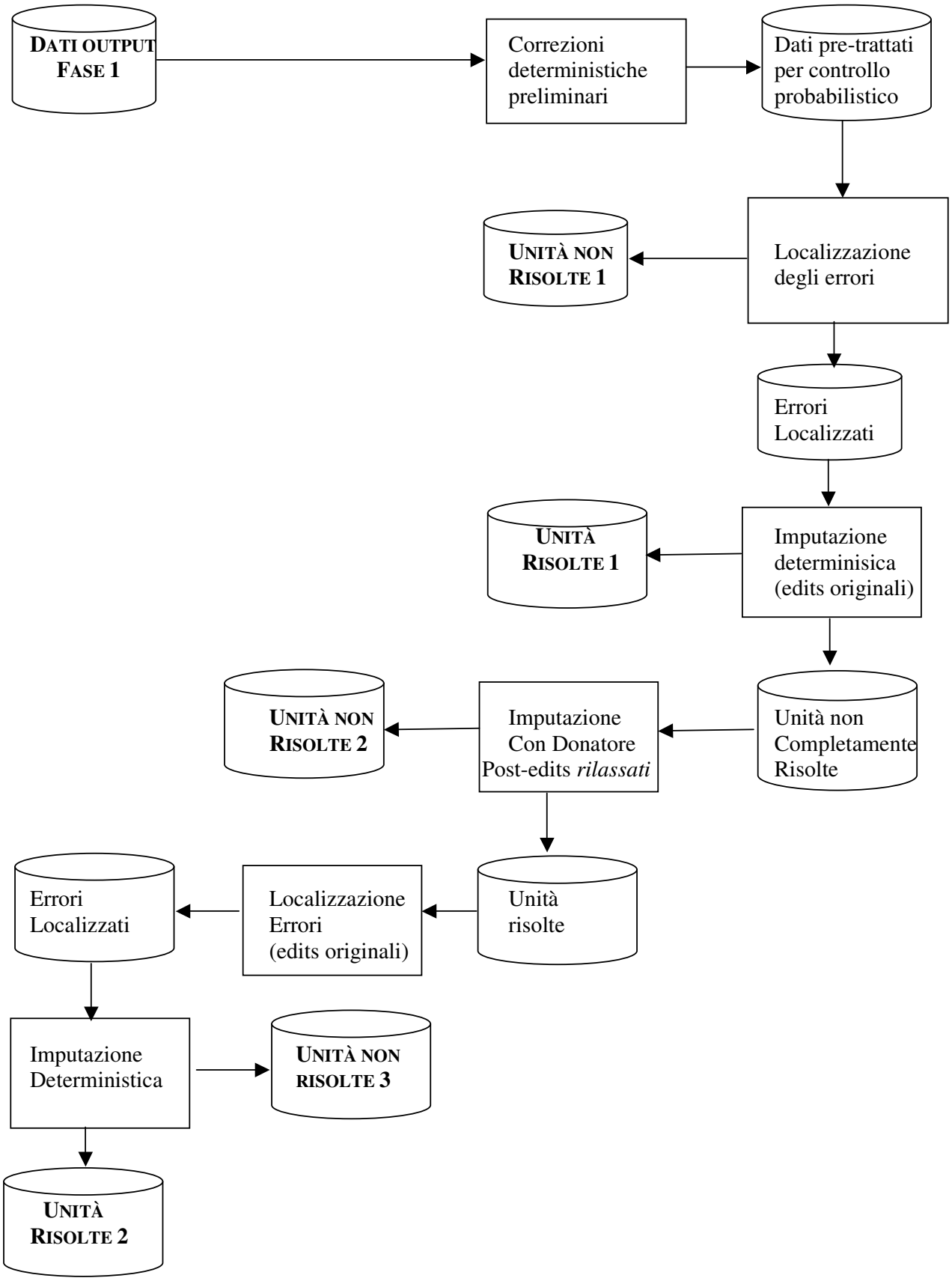
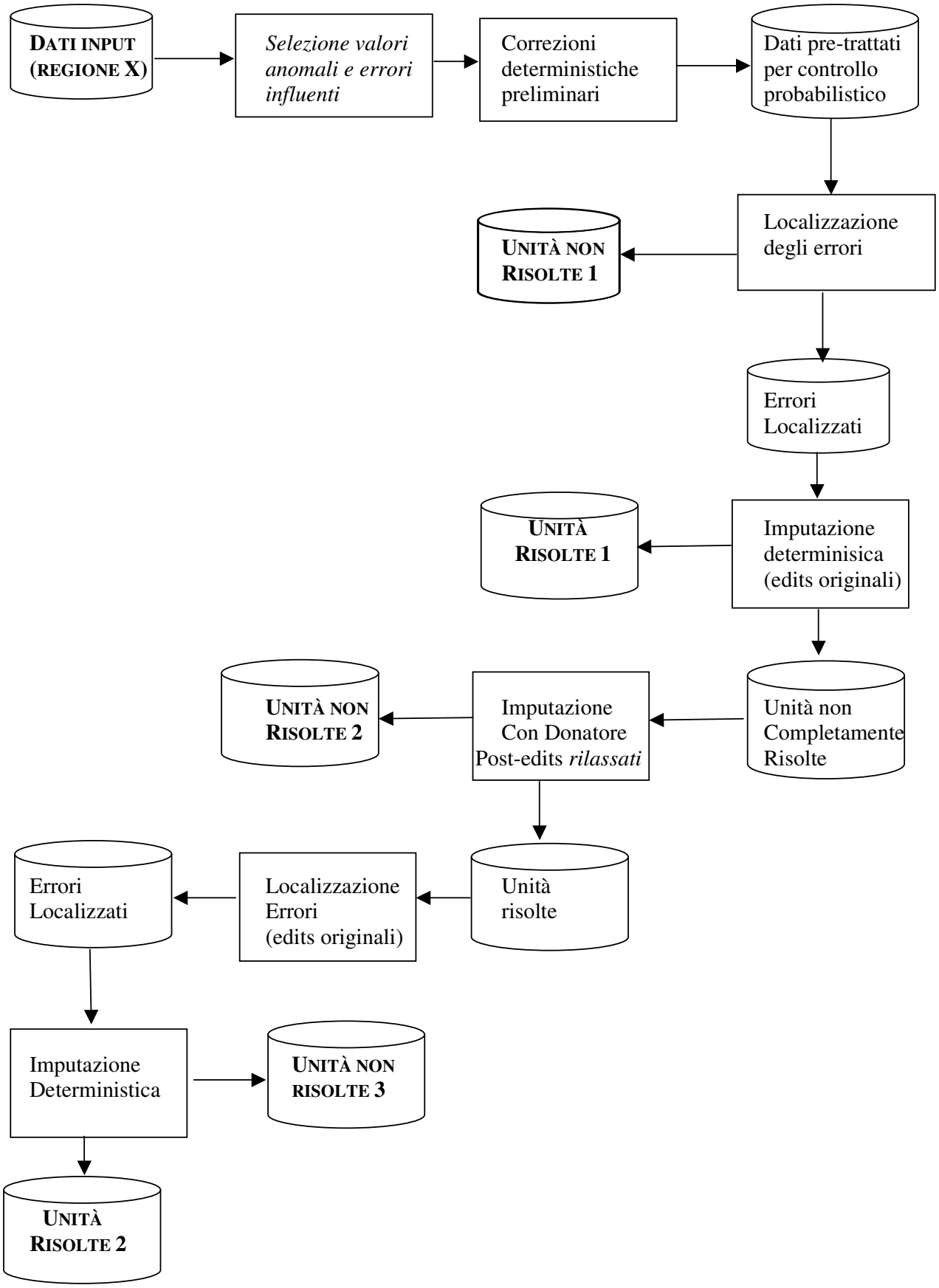


Figura 3 Flusso della procedura –Allevamenti



ALLEGATO 2 – TIPOLOGIE DI CONTROLLI UTILIZZATI NELL'INDAGINE

Nelle procedure di individuazione probabilistica e di imputazione degli errori realizzate mediante il software Banff, la verifica della coerenza dei dati viene effettuata rispetto a insiemi di vincoli (o *edit*) esprimenti le condizioni di accettabilità delle informazioni fornite da ogni unità in forma di equazioni o disequazioni lineari.

Di seguito sono elencate le tipologie di vincoli utilizzati nel caso dell'indagine SPA, per il controllo delle variabili su superfici, allevamenti e produzioni.

1) Vincoli di quadratura
Verificano che la somma di k variabili $X_1,..,X_k$ sia pari al valore della variabile X_T in cui viene osservato il corrispondente totale.

Esempio: per una data azienda, la superficie agricola utilizzata dell'azienda (SAU4) deve essere pari alla somma della superficie agricola utilizzata in proprietà (SAU1), in affitto (SAU2), e in uso gratuito (SAU3) dell'azienda:

```
sau1+sau2+sau3 = sau4
```

2) Vincoli di inclusione
Verificano che una variabile X sia inferiore o al massimo uguale ad una variabile Y.

Esempio: la superficie agricola utilizzata dell'azienda (SAU4) deve essere inferiore o al più uguale della superficie agricola totale dell'azienda (SAT4):

```
sau4 <= sat1
```

3) Vincoli di esistenza
Verificano che se una variabile X>0, allora la variabile Y ad essa legata assuma valori non nulli.

Esempio: Se la produzione di *frumento tenero e spelta* (p1) è diversa da zero, allora la superficie coltivata (principale o secondaria - s1+c1) a *frumento tenero e spelta* deve essere diversa da zero:

```
s1+c1>=0.0001*p1;
```

4) Vincoli di uguaglianza
Verificano che i valori di uno stesso fenomeno rilevato in modo diverso nel questionario siano uguali fra loro.

Esempio: la superficie agricola utilizzata dell'azienda rilevata nella sezione I del questionario (SAU4) deve essere uguale alla superficie agricola utilizzata dell'azienda rilevata nella sezione II del questionario (c94):

```
sau4=c94 ;
sat4=c101
```