istat
working
papers

N.13
2016

La procedura di editing selettivo nell'indagine su Struttura e Produzioni delle Aziende Agricole 2013

Orietta Luzi, Giovanni Seri, Roberta Varriale

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.13
2016

La procedura di editing selettivo nell'indagine su Struttura e Produzioni delle Aziende Agricole 2013

Orietta Luzi, Giovanni Seri, Roberta Varriale

Istat

Istat Working Papers

La procedura di editing selettivo nell'indagine su Struttura e Produzioni delle aziende agricole 2013

N. 13/2016

ISBN 978-88-458-1904-9

La procedura di *editing* selettivo nell'indagine su Struttura e Produzioni delle Aziende Agricole 2013[1]

Orietta Luzi, Giovanni Seri e Roberta Varriale

Sommario

L'indagine annuale su Struttura e Produzioni delle Aziende Agricole (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. In questo documento viene fornita la descrizione e la valutazione dei risultati della procedura di editing selettivo utilizzata ai fini dell'individuazione dei valori influenti per le principali variabili numeriche continue rilevate nell'indagine SPA (anno di riferimento 2013) relativamente alle sezioni su coltivazioni e allevamenti.

Parole chiave: editing selettivo, valori influenti, aziende agricole.

Abstract

The annual Survey on agricultural holdings structure and outputs collects information on areas cultivated with different crops, livestock dimensions, agricultural production, structure and amount of labour involved in the holding. The present work describes and evaluates the results of the selective editing strategy used to identify the influential values on the main numeric continuous variables for crops and livestock (reference year 2013).

Keywords: selective editing, influential values, agricultural holdings.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Indice

1. Introduzione

La rilevazione su Struttura e Produzioni delle Aziende Agricole (SPA nel seguito) raccoglie informazioni sulle aziende agricole italiane 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).

In occasione della prima edizione post-censuaria della rilevazione (anno di riferimento 2013), è stata disegnata una nuova procedura di controllo e correzione dei dati, in cui risultano integrati metodi e strumenti differenti per l'individuazione e il trattamento delle diverse tipologie di errore non campionario presenti nei dati.

In particolare, questo lavoro si concentra sulla descrizione della strategia di *editing* selettivo disegnata ai fini dell'individuazione degli *errori influenti* per le principali variabili numeriche continue rilevate dall'indagine (superfici, allevamenti e alcune produzioni).

Nell'ambito di ogni strategia di controllo e correzione di dati in indagini su imprese (in questo caso, aziende agricole) è infatti generalmente prevista una fase di individuazione degli errori aventi impatto rilevante (e quindi potenziale effetto distorsivo) sulle stime finali delle variabili principali oggetto di rilevazione (vedi ad esempio EDIMBUS, 2007, oppure MEMOBUST, 2014). Una classe di metodi particolarmente efficace in questo ambito è nota come *editing* selettivo (Latouche *et al.*, 1992), in cui le unità potenzialmente affette da errori influenti vengono selezionate in modo da garantire un controllo più stringente (generalmente di tipo manuale/interattivo) della loro natura/origine e una riduzione del loro potenziale effetto distorsivo sui risultati finali. Nel caso della rilevazione SPA, la metodologia di *editing* selettivo adottata per l'individuazione degli errori influenti è quella basata su modelli di contaminazione a classi latenti proposta da Di Zio e Guarnera (2013) e implementata nel pacchetto generalizzato Selemix (Guarnera *et al.,* 2013 e 2014).

Il lavoro è strutturato nel modo seguente. La sezione 2 contiene una descrizione delle principali caratteristiche della rilevazione SPA e della procedura complessiva di controllo e correzione per le variabili numeriche continue in essa rilevate. La sezione 3 contiene una breve descrizione della metodologia di *editing* selettivo adottata e del software Selemix in cui essa è implementata. Nella sezione 4 sono riportati gli esiti dell'applicazione del metodo ad alcune delle principali variabili dell'indagine SPA 2013. Nella sezione 5 sono valutati analiticamente i risultati e l'efficienza della procedura di *editing* selettivo a posteriori del processo di produzione delle stime delle variabili di superficie. Infine, la sezione 6 contiene alcune considerazioni conclusive e prospettive di lavoro.

2. La rilevazione SPA

L'indagine SPA 2013 risponde al Regolamento UE N. 1166/2008 del Parlamento Europeo e del Consiglio del 19 Novembre 2008 ed ha l'obiettivo di produrre in modo sistematico una serie di statistiche sulla struttura delle aziende agricole e sui metodi di produzione agricola. Per le variabili numeriche continue rilevate dall'indagine, i parametri oggetto di stima sono totali a livello di dettaglio Nazionale e regionale.

Il campione teorico della SPA 2013 comprende 42.723 aziende (D'Orazio, 2013[2]), per una popolazione di riferimento di 1.138.214 aziende attive sul territorio nazionale (circa il 70% delle 1.620.844 rilevate al 6° censimento agricoltura 2010).

Una caratteristica peculiare di questa edizione della rilevazione SPA è che si tratta della prima indagine effettuata dopo il 6° Censimento generale dell'agricoltura, cosa che ha reso possibile uti-

[2] I criteri di definizione della popolazione di interesse e degli standard di qualità che devono essere garantiti dal campione teorico sono fissati dal Regolamento UE N. 1166/2008. Al fine di garantire una ulteriore copertura in termini di *standard output* il campione è stato integrato con circa 2000 aziende non facenti parte della popolazione di riferimento e non rilevanti ai fini della procedura di *editing* selettivo. Il numero programmato di aziende da rilevare è stato, quindi, di 44.753.

lizzare il censimento come fonte ausiliaria, in modo da rendere più efficiente il processo di indagine e più accurati i risultati finali.

La rilevazione è stata condotta mediante somministrazione di questionari elettronici via web. Molteplici fattori contribuiscono alla complessità della strategia di controllo e correzione 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 (ad esempio, duplicazioni, errori di 'salto', quadrature) vengono individuati già in fase di *data entry*, grazie alla nuova modalità di acquisizione dei dati mediante questionario elettronico in cui alcuni controlli di coerenza sono effettuati all'atto della immissione delle informazioni. Questa strategia garantisce una maggiore accuratezza dei dati rispetto ad un insieme di base di controlli, anche se la non esaustività dei controlli stessi fanno sì che nei dati registrati permangano situazioni di non accettabilità o di incoerenza, e quindi di errore.

Gli errori verificatisi in fase di raccolta o in fase di registrazione dei dati possono essere di natura sia sistematica sia casuale: questi errori sono identificabili se danno luogo a incoerenze logico/statistiche/matematiche, a valori anomali, influenti o meno sulle stime dei parametri obiettivo dell'indagine.

Ai fini dell'individuazione delle varie tipologie di errore potenzialmente presenti nelle variabili su superfici, allevamenti e produzioni, è stata disegnata una procedura complessa di controllo e correzione le cui fasi principali sono:

1. individuazione di errori di identificazione delle unità e/o di copertura (errori nei codici identificativi, fusioni/scorpori di aziende agricole rispetto al Censimento, appartenenza alla popolazione obiettivo, presenza di contenuti informativi minimali) - pre-*editing*;
2. individuazione degli errori di natura sistematica (in particolare errori di unità di misura) sulla base di un approccio di tipo deterministico, ai fini della loro correzione automatica;
3. individuazione degli errori influenti di natura presumibilmente casuale mediante un approccio di tipo selettivo, ai fini della loro revisione manuale/interattiva;
4. individuazione e correzione automatica degli errori non influenti di natura presumibilmente casuale sulla base di un approccio di tipo probabilistico. In particolare, è stata utilizzata la metodologia di tipo *data-driven* implementata nel software Diesis (Bruni et al., 2002).

Nel seguito del lavoro, l'attenzione sarà focalizzata sulla seconda e terza fase della procedura.

3. L'*editing* selettivo per l'individuazione degli errori influenti

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 attraverso una loro revisione manuale/interattiva. Al fine di ridurre i tempi e i costi di questo tipo di trattamento, è necessario limitare tali controlli alle unità maggiormente critiche, lasciando a procedure di tipo automatico la risoluzione delle situazioni di errore residue. L'approccio noto come *editing* selettivo (Latouche et al., 1992) è stato proposto per l'individuazione dei valori influenti sotto vincoli di costo: tale approccio presuppone un ordinamento delle unità potenzialmente errate in base al loro impatto potenziale sulle stime dei parametri obiettivo (nel caso della SPA, i totali delle variabili quantitative), e la selezione delle prime m unità, secondo questo ordinamento, dove il numero m è funzione di un prefissato livello di accuratezza. Il livello di accuratezza corrisponde generalmente ad un valore di soglia per l'errore residuo accettabile sulle stime calcolate sui dati non editati. Questo modo di procedere consente di ottenere una riduzione sia dei tempi e dei costi del controllo interattivo per il prefissato livello di accuratezza, sia del 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.

3.1 Modelli di contaminazione per l'*editing* selettivo: il pacchetto Selemix

Fra i metodi per l'*editing* selettivo sviluppati negli anni recenti, quello proposto da Di Zio e Guarnera (2013) è basato sulla modellizzazione esplicita dei dati 'veri' (cioè non contaminati) e del meccanismo di errore. In particolare, il ricorso ad un modello di contaminazione a classi latenti - ovvero ad un modello di regressione a classi latenti qualora siano presenti delle covariate - stimato sui dati osservati consente di catturare la natura 'intermittente' degli errori e quindi di attribuire ad ogni osservazione di indagine una probabilità di presenza o meno dell'errore. Per ogni valore osservato è quindi fornita una previsione del corrispondente valore 'vero' (e dell'errore, come scostamento tra valore osservato e valore 'vero') basata sulla appropriata distribuzione condizionata. Questa caratteristica consente di associare il numero di unità da revisionare all'accuratezza desiderata per le stime di interesse. Il metodo, basandosi sull'utilizzo di un modello a classi latenti, non richiede la disponibilità simultanea di dati contaminati e 'puliti' su cui stimare il modello d'errore.

È importante sottolineare che la stima dei modelli di contaminazione può avvenire nell'ambito di una opportuna stratificazione delle unità nella popolazione. Tale stratificazione non coincide necessariamente con quella che definisce i domini delle stime di interesse sui quali vengono individuati i casi influenti. Nella pratica, è frequente il caso in cui i domini di stima costituiscono una partizione più fine della stratificazione utilizzata per la stima dei modelli di contaminazione, per i quali è in generale necessario garantire robustezza statistica attraverso adeguate numerosità campionarie.

Il metodo è implementato nello strumento generalizzato SeleMix, sviluppato in R internamente all'Istituto Nazionale di Statistica (Guarnera e Buglielli, 2013 e 2014). Il package SeleMix è costituito da tre funzioni principali che si occupano rispettivamente di stimare il modello di contaminazione (*ml.est*), di calcolare le previsioni dei valori 'veri' (*pred.y*), e di selezionare le unità contenenti errori potenzialmente influenti sulle stime obiettivo, in base ad una soglia di accuratezza specificata dallo statistico (*sel.edit*). Ai fini di selezionare le unità con errori influenti, qualora l'analisi venga svolta su dati di indagine, la procedura di stima utilizza i pesi campionari (eventualmente corretti per mancata risposta totale).

4. La procedura di *editing* selettivo per l'indagine SPA

Nel caso dell'indagine SPA 2013, l'approccio dell'*editing* selettivo implementato nel pacchetto Selemix è stato utilizzato per l'individuazione degli errori influenti per alcune variabili principali sulle superficie dell'azienda - superficie agricola totale, superficie agricola utilizzata, superficie irrigabile, superficie effettivamente irrigata - e per le variabili più significative relative agli allevamenti - totale bovini, totale bufalini, totale equini, totale ovini, totale caprini, totale suini, totale conigli - per le quali si dispone, tra l'altro, di informazioni censuarie affidabili raccolte in occasione del 6° Censimento generale dell'agricoltura. Le altre variabili sottoposte a *editing* selettivo per le quali non sono disponibili informazioni censuarie sono le variabili relative al latte munto da vacche da latte e altre vacche, da bufale, da pecore e da capre.

La fase di *editing* selettivo ha riguardato le 38.330 aziende agricole, sulle 42.723 appartenenti alla popolazione obiettivo della rilevazione SPA[3], correttamente identificate nella fase di pre-*editing*. Il modello di contaminazione, applicato a livello regionale, il metodo di calcolo delle previsioni dei valori 'veri' (valori predetti dal modello in alternativa all'utilizzo dei valori delle corri-

[3] La procedura di *editing* selettivo ha riguardato le osservazioni relative alle aziende classificate come rispondenti e, per quelle generate da una scissione, ricomposte al fine di ottenere unità comparabili a quelle rilevate al Censimento.

spondenti variabili censuarie) e la soglia di accuratezza delle stime obiettivo sono stati definiti mediante procedure statistiche esplorative e sotto la supervisione di esperti del settore. Più in particolare, la scelta di quale modello di contaminazione stimare per singola variabile, e quindi le informazioni ausiliarie da utilizzare, è stata basata, tra le altre cose, sulla consistenza numerica e sulle caratteristiche dei dati (ad esempio, eccessiva presenza di zeri nelle distribuzioni).

Nel prosieguo della trattazione viene descritta la procedura di *editing* selettivo separatamente per ogni singola variabile.

4.1 Superficie agricola utilizzata e superficie agricola totale

Le variabili contenenti l'informazione sulla superficie agricola utilizzata (SAU) e sulla superficie agricola totale (SAT) vengono rilevate nell'indagine in due modi differenti. Nel primo caso (SEZIONE II – AGGIORNAMENTO NOTIZIE STRUTTURALI), le variabili SAU1 e SAT1 vengono rilevate come sommatoria della quota di superficie (utilizzata e totale) di *Proprietà, usufrutto,* ecc., *in Affitto* e *ad Uso gratuito*. Nel secondo caso (SEZIONE III – UTILIZZAZIONE DEI TERRENI), la superficie agricola utilizzata (SAU2 nel seguito) viene calcolata come sommatoria delle variabili: *Totale seminativi, Totale coltivazioni legnose agrarie, Orti familiari* (per autoconsumo), *Totale prati permanenti e pascoli utilizzati*, *Prati permanenti e pascoli non più destinati alla produzione ed ammessi a beneficiare di aiuti finanziari*; mentre la superficie agricola totale (SAT2 nel seguito) viene calcolata come sommatoria delle variabili: *Superficie agricola utilizzata, Totale arboricoltura da legno, Totale boschi, Superficie agricola non utilizzata, Altra superficie*. E' importante notare che già in fase di compilazione del questionario elettronico sono presenti controlli sull'uguaglianza delle variabili SAU e SAT rilevate nei due modi differenti.

Sulla base di analisi esplorative dei dati, dopo aver trattato in fase preliminare alcuni casi affetti da errore di unità di misura, la strategia identificata come 'migliore' è consistita nel trattare le due variabili SAU2 e SAT2 separatamente, piuttosto che congiuntamente. Inoltre, avendo a disposizione le corrispondenti variabili al censimento (c_SAU2 e c_SAT2 nel seguito), il potere predittivo aggiuntivo in fase di stima del modello di contaminazione (modello di regressione a classi latenti) di qualsiasi altra variabile rilevata in sede di indagine è risultato essere molto basso.

Per le variabili SAU2 e SAT2 è stata utilizzata una procedura analoga.

Relativamente alla variabile SAU2, è stata adottata una procedura in cui:

- per le osservazioni tali che SAU2≠ c_SAU2, con SAU2 e c_SAU2 non nulle in entrambe le rilevazioni, è stato adottato, separatamente per ciascuna regione, un modello di contaminazione con variabile risposta SAU2 e covariata c_SAU2. Per tali osservazioni, le previsioni dei valori 'veri' sono state ottenute mediante la funzione *pred.y* di SeleMix;
- per le unità in cui SAU2 = c_SAU2 e quelle in cui una delle due variabili risultava nulla, la previsione del valore 'vero' è stata simulata ponendo SAU2 pari al corrispondente valore della variabile osservato al censimento.

Sui dati così ottenuti, la funzione *sel.edit* ha permesso quindi di selezionare le unità contenenti errori potenzialmente influenti sulla base di una soglia di accuratezza specificata pari al 5% sulle stime dei totali calcolate per regione e classe OTE (Orientamento Tecnico Economico, 9 modalità).

Analogo procedimento ha riguardato la variabile SAT2.

Come risultato, sono state identificate 183 osservazioni influenti per la variabile SAU2 e 129 per la variabile SAT2 (rispettivamente circa lo 0.5% e lo 0.3% del totale delle aziende agricole campionarie considerate) corrispondenti a meno di 250 codici azienda da controllare in quanto circa 80 aziende risultano contenere valori influenti per entrambe le variabili. Il numero di osservazioni influenti, per regione, è riportato nella Tavola 1.

Un elemento che si evidenzia è la concentrazione di casi segnalati come influenti in alcune regioni 'piccole' almeno dal punto di vista demografico, come la Valle d'Aosta, il Molise e l'Umbria. Concentrazioni consistenti di casi influenti in alcune regioni si verificano anche per altre variabili (vedi sezioni 4.2, 4.3 e 4.4) anche se in questo caso una spiegazione almeno parziale può derivare dal diverso impatto che una soglia uguale per tutti i domini di stima considerati può avere su regioni con caratteristiche diverse dal punto di vista dell'economia delle aziende agricole.

Tavola 1 – Numero di osservazioni influenti, per regione e tipo di prodotto

Regione	SAU2	SAT2	BOVINI	EQUINI	BUFALINI	OVINI	CAPRINI	SUINI	CONIGLI	LATTE DI MUCCA	LATTE DI BUFALA	LATTE DI PECORA	LATTE DI CA-PRA	IRR1	IRR2	Totale (a)
Piemonte	12	4	9	4	2	17	10	**80**	4	7	0	0	0	0	0	144
Valle D'Aosta	17	12	1	1	0	3	0	0	0	0	0	1	3	5	5	32
Lombardia	13	4	0	5	0	6	0	**68**	5	5	0	1	0	0	0	104
Veneto	5	4	3	5	1	4	11	28	8	2	1	0	0	**240**	**431**	586
Friuli-Venezia Giulia	4	3	8	1	1	3	8	12	2	0	0	3	0	0	**77**	114
Liguria	13	4	5	0	0	2	11	0	0	0	0	1	1	**115**	**90**	164
Emilia-Romagna	3	3	0	1	0	2	8	32	1	2	0	1	0	0	44	94
Toscana	5	9	**44**	9	3	**67**	11	10	4	4	0	**25**	1	0	0	171
Umbria	18	9	19	2	1	7	1	10	3	3	1	1	0	0	0	61
Marche	12	5	18	1	0	30	6	1	5	2	0	3	1	54	50	162
Lazio	10	5	**88**	**18**	8	29	6	5	2	5	3	17	0	0	0	178
Abruzzi	14	13	8	3	0	3	7	8	1	0	0	6	0	1	32	80
Molise	13	17	26	3	3	31	4	11	0	4	0	2	0	39	23	130
Campania	3	2	11	1	16	20	8	8	3	8	4	0	1	0	1	78
Puglia	13	12	11	4	1	29	14	6	2	0	0	7	1	**202**	72	274
Basilicata	9	6	0	6	3	**57**	**38**	8	2	0	0	**32**	3	31	7	171
Calabria	4	4	**69**	2	3	**110**	**96**	21	1	0	1	9	0	0	0	280
Sicilia	5	4	**170**	**24**	2	**58**	**71**	2	3	4	0	**60**	**12**	34	0	406
Sardegna	3	1	3	3	1	0	0	37	1	0	0	1	0	81	0	129
Trentino-Alto Adige-Trento	1	1	0	0	0	4	1	0	0	0	0	0	0	0	0	6
Trentino-Alto Adige-Bolzano	6	7	1	0	0	6	3	2	1	0	0	0	1	1	0	22
Totale	183	129	494	93	45	488	314	349	48	46	10	170	24	803	832	3386

(a) Il Totale non rappresenta la somma di riga ma il numero di aziende agricole distinte per cui è stato osservato almeno un valore influente.

4.2 Allevamenti: bovini, bufalini, equini, ovini, caprini, suini, conigli

Le variabili principali rilevate dall'indagine SPA nella SEZIONE V - CONSISTENZA DEGLI ALLEVAMENTI AL 1° DICEMBRE 2013 sono: *totale bovini, totale bufalini, totale equini, totale ovini, totale caprini, totale suini, totale allevamenti avicoli, totale conigli, totale struzzi, altri allevamenti*. Ogni tipo di allevamento è stato trattato con procedure di *editing* selettivo in maniera indipendente e la procedura seguita è stata impostata in maniera analoga per tutti i tipi di allevamento. Le variabili relative al *totale allevamenti avicoli, totale struzzi* e *altri allevamenti* sono state escluse dalle procedure di *editing* selettivo per la loro peculiarità. Infatti, mentre per il *totale allevamenti avicoli* la consistenza degli allevamenti è legata in maniera molto forte alla data di rilevazione, per le altre due variabili la specificità del tipo di allevamento ha fatto sì che le imprese con presenza di tali allevamenti fossero poche in termini numerici.

Considerata la numerosità e l'instabilità nel tempo (rispetto alle superfici) delle consistenze animali, nonché i conseguenti risultati forniti dalla stima del modello di contaminazione con variabile risposta *Totale capi dell'allevamento* rilevata nella SPA e con covariata *Totale capi dell'allevamento* rilevata al Censimento generale dell'agricoltura, è stato deciso di non effettuare la previsione da modello, ma di 'imputare' come valore 'vero' il valore osservato al censimento. Le unità influenti sono state quindi determinate per regione (senza distinzione per OTE) selezionando quelle in cui lo scarto tra i valori osservati alla SPA e al censimento fosse superiore a soglie s prefissate suggerite dagli esperti di settore (per bovini, equini, ovini e caprini s= 20 capi, per i bufalini s=10, per i suini s=50, per i conigli s=100 unità).

I risultati della procedura di *editing* selettivo sono riportati, suddivisi per regione, nella Tavola 1. Ciò che appare evidente, è che vi sono concentrazioni del numero di osservazioni influenti in alcune regioni per tipologia di allevamento (valori in grassetto nella tavola). Questo dovrebbe portare, preliminarmente alla fase di controllo manuale dei valori influenti, ad un'analisi del processo di rilevazione e acquisizione dei dati ai fini di accertare la presenza di eventuali cause sistematiche di errore.

4.3 Produzione del latte munto in azienda: latte di mucca, bufala, pecora, capra

La SEZIONE VI del questionario SPA è dedicata alla rilevazione di variabili relative alla - PRODUZIONE ED IMPIEGO DEL LATTE MUNTO IN AZIENDA, in particolare del latte di mucca, bufala, pecora e capra. Per queste variabili non è disponibile un corrispondente valore censuario per cui nella procedura di *editing* selettivo sono state utilizzate come variabili ausiliarie alcune variabili interne alla rilevazione SPA. Le quattro variabili relative alla produzione di latte sono state trattate in maniera indipendente, ciascuna con riferimento alle corrispondenti variabili che individuano le consistenze degli animali produttori. Quindi, per la produzione di latte di mucca, bufala, pecora e capra sono state utilizzate come informazioni ausiliarie, rispettivamente, la consistenza di mucche da latte, bufale da latte, pecore e capre. Un modello di contaminazione è stato quindi stimato utilizzando come variabile risposta il latte prodotto e come covariata la consistenza di animali da latte. In particolare, il modello è stato stimato per le osservazioni in cui la covariata è osservata maggiore di zero. Le osservazioni 'vere' sono state quindi determinate attraverso la funzione *pred.y* di SeleMix. Per le osservazioni in cui la produzione di latte è presente (maggiore di zero) mentre la covariata è assente (nulla o *missing*), è stato assunto come 'vero' il valore osservato dato che le due variabili non hanno una gerarchia di rilevazione e/o di accuratezza.

La procedura utilizzata è analoga per le quattro variabili con la differenza che nel caso del latte di mucca la numerosità più consistente di osservazioni con valori maggiori di zero ha permesso la stima tramite modello di contaminazione (*ml.est*) per regione mentre per le altre variabili si è ritenuto opportuno mantenersi a livello nazionale. I valori influenti sono stati determinati per regione.

Si noti che i casi in cui a una produzione di latte positiva non corrisponde una consistenza animale che lo produce, sono segnalati come anomali a priori per uno specifico controllo.

I risultati della procedura di *editing* selettivo sono riportati, suddivisi per regione, nella Tavola 1. Come per gli allevamenti, vi sono concentrazioni del numero di osservazioni influenti in alcune

regioni per tipologia di prodotto. Anche in questo caso, quindi, è necessaria un'analisi, preliminare alla fase di controllo manuale degli errori influenti, del processo di rilevazione e acquisizione dei dati ai fini di accertare la presenza o meno di problemi specifici.

4.4 Superficie irrigabile e superficie irrigata

La SEZIONE IV – LAVORAZIONE DELTERRENO, IRRIGAZIONE E PRODUZIONE DI ENERGIA RINNOVABILE del questionario raccoglie, tra le altre cose, informazioni sulla *Superficie irrigabile* e la *Superficie effettivamente irrigata* (IRR2).

In considerazione della limitata bontà di adattamento ai dati del modello di contaminazione basato sul dato censuario assunto come covariata, è stato scelto di determinare le osservazioni teoriche 'vere' direttamente con i valori censuari. Data l'elevata variabilità del fenomeno in esame (anche rispetto al censimento) è stata adottata una soglia di accuratezza delle stime dei totali per regione pari a 0.1 (invece di 0.05).

I risultati della procedura di *editing* selettivo sono riportati, suddivisi per regione, nella Tavola 1. Il numero delle osservazioni influenti risulta essere piuttosto alto rispetto alle altre variabili trattate. Data la specificità delle quantità in analisi, dovranno essere condotte da parte degli esperti ulteriori analisi dei dati, preliminarmente alla fase di controllo manuale degli errori influenti, al fine di individuare eventuali criticità in fase di rilevazione e/o di registrazione .

5. Valutazione della procedura di *editing* selettivo

In questo paragrafo è presentata la valutazione della procedura di *editing* selettivo relativamente alle variabili SAU2 e SAT2. L'analisi è limitata alle due variabili che rilevano la superficie delle aziende agricole in quanto sono quelle che presentano il maggior grado di stabilità nel tempo delle informazioni rilevate rispetto a quelle ausiliarie.

Le Tavole 2 e 3 riportano per Regione - rispettivamente per le variabili SAU2 e SAT2 - il numero di: dati complessivamente rilevati (A), dati effettivamente corretti (*imputati*) sia attraverso procedure di *editing* automatiche che interattive (B), dati segnalati come influenti (C) e, tra questi, dati corretti mediante *editing* interattivo (D). Inoltre, nelle Tavole sono riportate le seguenti percentuali:

- *Tasso di imputazione (B/A)*: dati complessivamente corretti sul numero di dati rilevati,
- *Tasso di errori influenti (C/A)*: dati segnalati come influenti sul numero di dati rilevati,
- *Tasso di imputazione per errori influenti (D/B)*: dati segnalati come influenti e corretti interattivamente sul numero di dati complessivamente corretti
- *Hit Rate (D/C)*: dati segnalati come influenti e corretti mediante editing interattivo sul numero di dati segnalati come influenti.

Dalla Tavola 2 si evince che delle 44.552 osservazioni della variabile SAU2 nel dataset finale il 97.18% non subisce alcuna correzione. Complementarmente, il tasso di imputazione è pari al 2.82% mentre la quota dei casi segnalati come 'influenti' è dello 0.41% che in termini assoluti equivalgono a 183 casi. Di questi solo 20 (*hit rate* =10,93%) sono stati classificati come errori e corretti. Per la variabile SAT2 (si veda Tavola 3), analogamente, il 96.53% delle osservazioni non subisce alcuna correzione. La quota dei casi segnalati come 'influenti' è dello 0.29% che in termini assoluti equivalgono a 129 casi. Di questi solo 24 (*hit rate*=18.60%) sono stati classificati come errori e corretti. Le differenze tra le due variabili SAU2 e SAT2 riguardano sostanzialmente il numero complessivo di correzioni subite, più alto per la seconda variabile, rispetto al numero di casi influenti corretti, al contrario più alto per SAU2. I valori 10.93% e 18.60%, rispettivamente per le variabili SAU2 e SAT2, dell'hit rate possono essere interpretate come misura di 'efficienza' della procedura adottata nell'individuare, tra i casi influenti, gli errori effettivi. I valori generalmente bassi dell'indice possono essere dovuti al fatto che nei modelli utilizzati per queste variabili è stata considerata come variabile ausiliaria un'informazione (rilevata al censimento) di tre anni precedente all'indagine, quindi con ridotto potere predittivo, e che tale informazione potrebbe a sua volta manifestare difetti di qualità.

Tavola 2 – Numero di osservazioni totali, influenti e corrette, per regione: SAU2

Regione	Totale (A)	Totale Corretti (B)	Influenti (C)	Influenti Corretti (D)	B/A (%)	C/A (%)	D/B (%)	D/C (%)
Piemonte	2236	82	12	1	3.67	0.54	1.22	8.33
Vale D'Aosta	230	21	17	4	9.13	7.39	19.05	23.53
Lombardia	2090	96	13	1	4.59	0.62	1.04	7.69
Veneto	2877	39	5	0	1.36	0.17	0.00	0.00
Friuli-Venezia Giulia	1152	66	4	1	5.73	0.35	1.52	25.00
Liguria	649	39	13	2	6.01	2.00	5.13	15.38
Emilia-Romagna	2698	82	3	1	3.04	0.11	1.22	33.33
Toscana	2400	14	5	0	0.58	0.21	0.00	0.00
Umbria	1117	32	18	1	2.86	1.61	3.13	5.56
Marche	1716	91	12	1	5.30	0.70	1.10	8.33
Lazio	3231	116	10	1	3.59	0.31	0.86	10.00
Abruzzi	2342	13	14	0	0.56	0.60	0.00	0.00
Molise	974	26	13	2	2.67	1.33	7.69	15.38
Campania	2847	38	3	0	1.33	0.11	0.00	0.00
Puglia	3081	49	13	1	1.59	0.42	2.04	7.69
Basilicata	1749	73	9	1	4.17	0.51	1.37	11.11
Calabria	4181	49	4	0	1.17	0.10	0.00	0.00
Sicilia	5432	240	5	2	4.42	0.09	0.83	40.00
Sardegna	2218	9	3	0	0.41	0.14	0.00	0.00
Trentino-Alto Adige-Trento	524	52	1	1	9.92	0.19	1.92	100.00
Trentino-Alto Adige-Bolzano	808	29	6	0	3.59	0.74	0.00	0.00
Totale	44552	1256	183	20	2.82	0.41	1.59	10.93

Tavola 3 – Numero di osservazioni totali, influenti e corrette, per regione: SAT2

Regione	Totale (A)	Totale Corretti (B)	Influenti (C)	Influenti Corretti (D)	B/A (%)	C/A (%)	D/B (%)	D/C (%)
Piemonte	2236	82	4	0	3.67	0.18	0.00	0.00
Valle D'Aosta	230	21	12	4	9.13	5.22	19.05	33.33
Lombardia	2090	95	4	1	4.55	0.19	1.05	25.00
Veneto	2877	38	4	0	1.32	0.14	0.00	0.00
Friuli-Venezia Giulia	1152	62	3	0	5.38	0.26	0.00	0.00
Liguria	649	32	4	2	4.93	0.62	6.25	50.00
Emilia-Romagna	2698	91	3	0	3.37	0.11	0.00	0.00
Toscana	2400	24	9	1	1.00	0.38	4.17	11.11
Umbria	1117	50	9	2	4.48	0.81	4.00	22.22
Marche	1716	97	5	1	5.65	0.29	1.03	20.00
Lazio	3231	145	5	1	4.49	0.15	0.69	20.00
Abruzzi	2342	22	13	1	0.94	0.56	4.55	7.69
Molise	974	48	17	5	4.93	1.75	10.42	29.41
Campania	2847	50	2	0	1.76	0.07	0.00	0.00
Puglia	3081	65	12	3	2.11	0.39	4.62	25.00
Basilicata	1749	92	6	2	5.26	0.34	2.17	33.33
Calabria	4181	160	4	0	3.83	0.10	0.00	0.00
Sicilia	5432	270	4	0	4.97	0.07	0.00	0.00
Sardegna	2218	18	1	0	0.81	0.05	0.00	0.00
Trentino-Alto Adige-Trento	524	58	1	0	11.07	0.19	0.00	0.00
Trentino-Alto Adige-Bolzano	808	27	7	1	3.34	0.87	3.70	14.29
Totale	44552	1547	129	24	3.47	0.29	1.55	18.60

I valori 10.93% e 18.60%, rispettivamente per le variabili SAU2 e SAT2, dell'*hit rate* possono essere interpretate come misura di 'efficienza' della procedura adottata nell'individuare, tra i casi influenti, gli errori effettivi. I valori generalmente bassi dell'indice possono essere dovuti al fatto che nei modelli utilizzati per queste variabili è stata considerata come variabile ausiliaria un'informazione (rilevata al censimento) di tre anni precedente all'indagine, quindi con ridotto potere predittivo, e che tale informazione potrebbe a sua volta manifestare difetti di qualità.

Il t*asso di imputazione per errori influenti* supera il 10% solo in Valle d'Aosta e Molise, evidentemente risentendo del basso numero di casi totali, mentre a Trento risulta poco significativo in quanto vi è solo un caso segnalato come possibile valore anomalo influente.

Come ulteriore elemento di valutazione, nella Tavola 4 si riportano, rispettivamente per le variabili SAU2 e SAT2, le differenze percentuali tra le stime calcolate sui seguenti insiemi di dati: prima della fase di controllo e correzione (*Grezzi*), dopo la fase di controllo interattivo degli errori influenti (*Semi finali*), e a valle della procedura complessiva di controllo e correzione (*Finali*). Le stime sono calcolate utilizzando i pesi campionari aggiustati per mancata risposta totale. In altre parole i dati riportati nella Tavola esprimono, in percentuale, l'impatto sulle stime ottenute sui dati *Grezzi* rispettivamente delle correzioni effettuate sui soli dati influenti e l'impatto totale delle correzioni. L'impatto è misurato sulle stime totali e sono pertanto possibili compensazioni per correzioni di diverso segno sui dati. Sul totale nazionale le correzioni nel loro complesso comportano una variazione inferiore all'uno percento della stima rispetto ai dati grezzi.

Tavola 4 – Differenze % rispetto alle stime di SAU2 e SAT2 calcolate sui dati Grezzi di quelle che si ottengono dopo i controlli interattivi (Semi finali) e come esito della rilevazione (Finali)

Regione	(Semi finali-Grezzi)/Grezzi (%)	(Finali-Grezzi)/Grezzi (%)
Totale SAU2	0.03	0.98
Totale SAT2	-0.04	0.09

Per meglio interpretare l'impatto contenuto delle correzioni sui dati segnalati come influenti vengono presentate in Tavola 5 la somma e la media delle differenze assolute tra i dati segnalati come influenti e corretti in quanto errori e il rispettivo dato grezzo in raffronto con le stesse misure per le differenze assolute dei dati corretti ma non segnalati come influenti e i rispettivi dati grezzi. Quel che si osserva è che le correzioni sui dati influenti sono in media più consistenti di quelle effettuate sui dati non influenti: circa 1.23 volte per la variabile SAU2 e 2.8 volte per la variabile SAT2. Questo avvalora la capacità della procedura di *editing* selettivo di identificare gli errori maggiormente influenti sulle stime finali.

Tavola 5 – Somma e media delle differenze in valore assoluto tra dati Influenti corretti e dati Grezzi e tra dati Corretti e Grezzi per le variabili SAU2 e SAT2

Variabile	Somma	N	Media	A Media/B Media % C Media/D Media %
A. SAU2 |Influenti corretti – Grezzi|	402342	20	20117	
B. SAU2 |Non Influenti corretti – Grezzi|	20179311	1236	16326	123
C. SAT2 |Influenti corretti – Grezzi|	2104088	24	87670	
D. SAT2 |Non Influenti corretti – Grezzi|	47493823	1523	31184	281

6. Considerazioni finali

L'integrazione della procedura di controllo e correzione dei dati della Rilevazione annuale sulla Struttura e Produzioni delle Aziende Agricole 2013 con una procedura di *editing* selettivo ha avuto come obiettivo principale quello di individuare le osservazioni errate di maggiore influenza sui ri-

sultati finali per sottoporle ad un controllo manuale interattivo da parte delle strutture preposte alla conduzione della rilevazione. Dato che questo tipo di controlli è particolarmente oneroso, l'obiettivo di una tale procedura è quello di circoscrivere i controlli ad un numero limitato di casi confidando che tra di essi siano concentrati gli errori più rilevanti in termini di potenziale effetto distorsivo sulle stime obiettivo. L'efficacia della procedura è tanto maggiore quanto migliore è la qualità delle informazioni ausiliarie eventualmente disponibili.

Nell'ambito della rilevazione sono state considerate le variabili relative alle coltivazioni, rappresentate dalle variabili che esprimono le superfici totali delle imprese agricole, e agli allevamenti. Le informazioni ausiliarie utilizzate nella procedura di *editing* selettivo sono le analoghe informazioni rilevate in occasione del 6° Censimento generale dell'agricoltura del 2010.

Gli esiti della procedura sono risultati diversi nei due ambiti di analisi. Le variabili di superficie, come atteso, si sono dimostrate più stabili nel tempo oltre a presentare naturalmente una numerosità di informazioni rilevate più elevata. Ciò ha consentito di impostare la procedura di *editing* selettivo su domini di stima definiti come combinazione delle variabili ausiliarie Regione e OTE adottando una prefissata soglia *standard* di accuratezza delle stime (comune a tutti i domini) per ottenere un insieme di osservazioni 'influenti' da sottoporre a controllo interattivo di numerosità accettabile.

Per le consistenze animali, invece, dal momento che sul dato relativo alla presenza dei capi nelle aziende intervistate può risultare rilevante anche il giorno in cui esso viene rilevato - questo è particolarmente vero per gli avicoli ma riscontrabile anche per la altre tipologie animali - l'informazione ausiliaria utilizzata per la procedura di *editing* selettivo è risultata meno efficiente. Ciò ha portato probabilmente a identificare un numero di casi 'influenti' più ampio di quello atteso e/o auspicato ai fini del contenimento del controllo interattivo, nonostante un'attenta definizione di soglie diversificate per le diverse tipologie animali. Anche le informazioni sulla produzione di latte, dipendenti naturalmente dalla consistenza degli animali, ha risentito di una tale instabilità. Inoltre, la concentrazione di casi influenti in alcune regioni per alcune variabili ha portato ad ipotizzare la presenza di fattori sistematici incidenti sui risultati. Ulteriori approfondimenti sono dunque necessari al fine di verificare da un lato l'efficacia dei modelli e delle covariate utilizzati per questa tipologia di variabili, dall'altro l'esistenza di ulteriori, più aggiornate, fonti di informazione ausiliaria da usare come predittori.

La valutazione degli effetti della procedura di *editing* selettivo proposta è stata concentrata sulle variabili di superficie SAU2 e SAT2. La quota di casi segnalati come influenti ed effettivamente corrispondenti ad errori si è rivelata piuttosto bassa. Ciononostante, la consistenza media degli errori identificati è risultata sensibilmente più alta per i casi segnalati come influenti rispetto alla media generale delle osservazioni corrette. Le numerosità degli insiemi di osservazioni sottoposti a controllo interattivo è sembrato adeguato rispetto alle risorse preventivate per questo tipo di attività.

Nel complesso si può ritenere che la procedura di *editing* selettivo ha prodotto risultati incoraggianti/soddisfacenti, ma che nel caso specifico la sua efficacia è risultata inferiore alle attese sia per la natura dei dati che per l'assenza di informazioni ausiliarie di qualità rispondente alle esigenze di analisi.

Per rendere più efficiente la procedura nelle edizioni successive della rilevazione occorrerà valutare, sulla base anche di questa esperienza, quali variabili investigare in funzione delle informazioni ausiliarie che saranno disponibili sia da altre fonti/archivi amministrativi, sia longitudinalmente (sfruttando cioè il dato 2013). In questo senso, potrebbe essere rivista anche il tipo di modellazione passando ad una strategia di tipo multivariato. Inoltre, può essere utile ripensare la strategia di definizione delle soglie di accuratezza delle stime diversificandole sui singoli domini in funzione della dimensione degli stessi per uniformare i criteri di identificazione delle osservazioni influenti anche in termini assoluti oltre che relativi. Infine, vanno effettuati approfondimenti sulle varie fasi del processo di rilevazione (ad esempio in fase di registrazione dei dati) per identificare e prevenire eventuali cause sistematiche di errore sulle variabili principali del questionario di indagine.

Riferimenti bibliografici

Bruni R., Reale A., Torelli R. 2002. DIESIS: a New Software System for Editing and Imputation. *Proceedings SIS2002*. Milano.

D'Orazio M. 2013. *Il campione per l'indagine sulla struttura delle aziende agricole*. (Documento interno).

Di Zio, M. and U. Guarnera. 2013. A Contamination Model for Selective Editing. *Journal of Official Statistics*, 29, 4: 539-555.

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

Guarnera, U. e M.T. Buglielli. 2013. SeleMix: an R Package for Selective Editing. http://cran.r-project.org/web/packages/SeleMix/vignettes/SeleMix-vignette.pdf .

Guarnera, U. e M.T. Buglielli. 2014. Package SeleMix. Reference manual http://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf

ISTAT. 2001. *Struttura e Produzioni delle Aziende Agricole – Anno 1999 - Italia*. Roma: Istat.

Latouche M. and J.M. Berthelot. 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 R. McKenzie. 2000. The General Application of Significance Editing. *Journal of Official Statistics*, 16: 243-253.

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