CONTRIBUTI ISTAT

n. 13/2008

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

AA. VV.

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

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

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

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

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

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

Direttore responsabile della Rivista di Statistica Ufficiale: Patrizia Cacioli

Comitato di Redazione delle Collane Scientifiche dell'Istituto Nazionale di Statistica

Coordinatore: Giulio Barcaroli

Membri:

Corrado C. Abbate	Rossana Balestrino	Giovanni A. Barbieri
Giovanna Bellitti	Riccardo Carbini	Giuliana Coccia
Fabio Crescenzi	Carla De Angelis	Carlo M. De Gregorio
Gaetano Fazio	Saverio Gazzelloni	Antonio Lollobrigida
Susanna Mantegazza	Luisa Picozzi	Valerio Terra Abrami
Roberto Tomei	Leonello Tronti	Nereo Zamaro

Segreteria: Gabriella Centi, Carlo Deli e Antonio Trobia

Responsabili organizzativi per la *Rivista di Statistica Ufficiale*: Giovanni Seri e Carlo Deli

Responsabili organizzativi per i *Contributi ISTAT* e i *Documenti ISTAT*: Giovanni Seri e Antonio Trobia

ISTAT – Produzione libraria e centro stampa

CONTRIBUTI ISTAT

n. 13/2008

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

AA. VV.

(*) ISTAT - Direzione

I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto

Contributi e Documenti Istat 2008

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

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

Indice

Premessa

Orietta Luzi, Istat, Servizio Metodologie, Tecnologie e Software per la Produzione dell'Informazione Statistica (MTS/G)

Questo volume raccoglie i lavori presentati al seminario "Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali" (Istat, Roma, 15 Novembre 2008). Si tratta del secondo seminario[1] organizzato nell'ambito del progetto Europeo EDIMBUS (EDiting and IMputation in cross-sectional BUsiness Surveys) per la realizzazione di un manuale di *Pratiche Raccomandate per il controllo e la correzione dei dati per le indagini trasversali sulle imprese*. Entrambi i seminari sono nati dalla duplice esigenza di: 1) fare il punto della situazione dello stato dell'arte in Istat nell'area del controllo e correzione dei dati (CC nel seguito) da un punto di vista sia metodolologico sia operativo, e 2) condividere le esperienze maturate in questo ambito nei settori produttivi dell'Istat.

Il progetto EDIMBUS (edimbus.istat.it) ha rappresentato una occasione per avviare una riflessione sulla necessità di portare avanti a livello Europeo, e quindi negli Istituti di Statistica dei Paesi Membri, un processo di progressiva armonizzazione delle procedure di individuazione e trattamento degli errori non campionari. Questo sia termini di metodologie e pratiche, sia di approccio al disegno, valutazione e documentazione di procedure di CC. In effetti, nonostante sia ormai ampiamente riconosciuta la rilevanza della fase di CC in termini di impatto sulla qualità dei dati (tempestività, accuratezza, comparabilità, ecc.) e sui costi dei processi di produzione dell'informazione statistica, persiste sia nell'ambito del Sistema Statistico Europeo (SSE), sia internamente ai singoli Istituti Nazionali di Statistica, una forte eterogeneità nell'area del CC. Questa eterogeneità riguarda sia le definizioni e concetti di base, sia i metodi e le tecniche adottate, sia gli approcci al disegno, alla valutazione, e alla documentazione. Una maggiore armonizzazione consentirebbe maggiori livelli di accuratezza e tempestività dell'informazione prodotta, maggiore comparabilità delle statistiche a livello Nazionale e Europeo, minori costi.

Le pratiche raccomandate sviluppate nell'ambito del progetto EDIMBUS (Luzi et al., 2007) sono il primo tentativo di procedere in questa direzione a livello di SSE per la specifica area delle statistiche economiche[2]. Lo sviluppo di un manuale a livello europeo, tuttavia, pone un problema di successiva revisione, eventuale integrazione e test delle raccomandazioni nell'ambito di ogni Paese Membro sulla base delle specificità di ogni contesto Nazionale, tenendo conto delle peculiarità e dei vincoli di ogni Istituto di Statistica. Il punto di partenza non può che essere una ricognizione interna per la verifica dello stato dell'arte in ogni Istituto nel contesto delle indagini economiche.
Per quanto riguarda l'Istat, questa attività di ricognizione ha portato a far emergere alcune esperienze molto significative da un punto di vista sia metodologico sia operativo, mettendo in luce l'enorme lavoro correntemente svolto dai settori produttivi dell'Istituto, e gli avanzamenti metodologici in atto. Questa attività, che può essere considerato il punto di partenza di un continuo processo di condivisione delle informazioni e delle esperienze correnti in Istat, rimarrebbe però inefficace se non fosse affiancata da altre iniziative volte a garantire una maggiore armonizzazione del processo di CC dati. In particolare, andrebbero individuati gli strumenti più efficaci per la diffusione e la condivisione di una nuova concezione degli obiettivi e del "ciclo di vita" della fase di CC, visto come parte integrante e inscindibile del processo di indagine. E' noto infatti che la fase di CC dati è spesso considerata un semplice insieme

[1] Il primo seminario, dal titolo "Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali" si è tenuto all'Istat il 25 maggio 2007. Per gli interventi, vedi AA.VV. (2008).
[2] Alcuni singoli Istituti Nazionali di Statistica Europei, come Statistics Sweden e Statistics Finland, hanno sviluppato *Quality Guidelines* sul processo di indagine che includono anche linee guida sulla fase di CC dati.

di operazioni di elaborazione dati finalizzate alla eliminazione delle incongruenze (errori, valori mancanti, dati anomali) ne dati. Va invece sempre più affermandosi una nuova concezione, in cui: 1) la fase di CC è vista come un processo di *analisi statistica* dei dati finalizzato alla individuazione e risoluzione delle situazioni anomale o non accettabili attraverso l'utilizzo di strumenti anche complessi per la modellazione, la previsione, l'esplorazione dei dati; 2) un ruolo fondamentale della fase di CC è supportare lo statistico nell'individuazione delle cause di errore, al fine di attivare di un processo virtuoso che porti al miglioramento dell'intero processo di indagine attraverso la rimozione di tali cause; 3) la fase di CC ha un carattere di forte trasversalità nell'ambito del processo di indagine, dovuto ai suoi stretti legami con le altre fasi del processo stesso, per cui il disegno e l'implementazione dei flussi di dati e operazioni del processo di CC assumono livelli di complessità maggiori.
Relativamente al terzo aspetto, vanno inoltre considerati i seguenti elementi di complessità:

- il sempre più diffuso ricorso all'uso di questionari elettronici (CATI, CAPI, web) che consentono di anticipare alla fase di raccolta/registrazione dei dati parte dei controlli di qualità tradizionalmente effettuati esclusivamente in fase di CC;
- il sempre maggior ricorso all'integrazione di altre fonti nel processo di indagine, con conseguente necessità di integrare diversi processi di controllo di qualità delle diverse fonti di informazione, e di monitorare il diverso impatto delle diverse fonti sulla qualità complessiva dei dati finali;
- i legami esistenti fra la fase di CC e la fase di stima (necessità di valutazione dell'effetto del CC sull'accuratezza delle stime e sulle analisi dei dati successive).

I nove lavori raccolti in questo volume vanno letti alla luce di queste considerazioni. Questi lavori, che illustrano altrettante esperienze nell'area delle statistiche economiche congiunturali, evidenziano uno o più elementi di complessità metodologica e/o operativa fin qui richiamati. A chiusura del volume è riportata la sintesi dll'intervento al seminario del discussane, dott. Roberto Gismondi, della Direzione Centrale delle statistiche economiche congiunturali su imprese, servizi e occupazione.

La giornata seminariale è stata aperta dal Direttore della Direzione Centrale delle Statistiche Economiche Congiunturali su Imprese, Servizi e Occupazione, dott. Gian Paolo Oneto, il quale ha richiamato l'attenzione sulla importanza dell'occasione offerta dal seminario per portare alla luce attività spesso non sufficientemente valorizzate nell'ambito della produzione dell'informazione statistica, nonostante il loro impatto cruciale in termini di tempi, costi, accuratezza e trasparenza dell'informazione prodotta. Il Seminario è stato chiuso dal dott. Roberto Monducci, Direttore della Direzione Centrale delle Statistiche sui Prezzi e Commercio con l'Estero, il quale ha auspicato una sempre più intensa collaborazione fra settori produttivi dell'Istituto sui temi cruciali toccati nel corso delle presentazioni, anche nell'ambito di gruppi di lavoro o altre opportune forme di collaborazione.

Bibliografia

AA.VV. "Seminario: Metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali". *Contributi Istat*, n. 7/2008.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Rapporto tecnico del progetto Europeo EDIMBUS, 2007. (http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/RPM_EDIMBUS.PDF)

Controllo e correzione nell'indagine mensile sulle grandi imprese: metodi e prime evidenze da un'analisi retrospettiva sulla qualità

Fabiana Rocci, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
Laura Serbassi, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi

Sommario: Il presente lavoro descrive il processo di controllo e correzione adottato dall'indagine mensile "Lavoro e retribuzioni nelle grandi imprese". Le tecniche utilizzate tengono conto di alcuni aspetti caratteristici dell'indagine, in particolare: ogni unità di rilevazione, in quanto grande impresa, è considerata autorappresentativa e potenzialmente influente, la distanza tra la diffusione dei dati e il periodo di riferimento è molto contenuta. Infine, oltre alla produzione degli indicatori d'indagine, i microdati delle grandi imprese vengono integrati nei dati amministrativi (DM10 Inps) dell'indagine Oros (Occupazione, retribuzioni lorde e oneri sociali). Tutto ciò rende indispensabile effettuare dei controlli capillari per garantire un'elevata qualità anche a livello dei dati elementari. In generale, il processo di controllo e correzione si affianca in modo continuo a tutte le attività di raccolta, elaborazione e analisi dei dati. Nonostante il diffuso impiego di strumenti informatici, esso è basato principalmente sull'attività di revisori esperti che curano anche il contatto diretto con le singole imprese. Le principali fasi esaminate nel dettaglio sono il microediting interattivo e la metodologia seguita per l'imputazione delle mancate risposte totali (MRT). Nell'ultimo paragrafo vengono presentati i risultati di una prima ricognizione sulle tipologie degli errori non campionari più frequentemente registrati e una stima dell'impatto dell'imputazione delle MRT sugli indicatori prodotti dall'indagine.

Parole chiave: grandi imprese, microediting interattivo, mancate risposte totali, macroediting, errori non campionari, indicatori di qualità.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

Introduzione[3]

I fabbisogni conoscitivi ai quali la rilevazione mensile "Lavoro e retribuzioni nelle grandi imprese" intende corrispondere sono riferiti all'analisi di breve periodo dell'andamento congiunturale dell'occupazione, delle ore lavorate, delle retribuzioni e del costo del lavoro nelle imprese di grande dimensione (almeno 500 addetti nella media dell'anno base) appartenenti ai settori dell'industria e dei servizi distributivi e alle imprese.

I dati vengono diffusi mensilmente dopo circa 60 giorni dalla fine del mese di riferimento[4] e insieme a quelli derivanti dall'indagine sulle retribuzioni contrattuali e sui conflitti di lavoro, costituiscono l'unica fonte di informazione statistica ufficiale con tale cadenza sull'andamento del mercato del lavoro. Inoltre, i microdati raccolti dall'indagine sono integrati con quelli dell'archivio Inps dei modelli mensili DM10 per consentire all'Istat di produrre trimestralmente, attraverso la rilevazione Oros (Occupazione, retribuzioni lorde e oneri sociali), informazioni sulla dinamica occupazionale e retributiva di tutte le imprese con almeno un dipendente del settore privato non agricolo e per soddisfare due esigenze conoscitive di livello europeo: la costruzione dell'indice trimestrale europeo del costo del lavoro (Lci, Reg. Ce n. 450/2003) e la produzione degli indicatori sulle ore lavorate destinati a dare attuazione al Regolamento comunitario sulle statistiche economiche congiunturali (Sts, Reg. Ce n. 1165/98).

La rilevazione sulle grandi imprese si caratterizza per i tempi brevi di diffusione e per le unità di rilevazione, peculiari da un punto di vista statistico: infatti esse sono considerate tutte autorappresentative e potenzialmente influenti. Per questi motivi, la raccolta, la registrazione e la validazione dei microdati seguono un processo continuo e capillare, con una struttura ben definita in ogni sua parte. L'attenzione è posta sulla prevenzione degli errori non campionari, in tutte le loro forme, e sulla riduzione al minimo delle mancate risposte totali.

L'obiettivo di questo documento è di inquadrare il processo di controllo e correzione dell'indagine nello schema suggerito dalle raccomandazioni di Eurostat (Istat, CBS, SFSO, Eurostat, 2007). Dopo una breve descrizione delle principali fasi dell'indagine, illustreremo nel dettaglio quella relativa al microediting, la metodologia utilizzata per l'imputazione delle mancate risposte totali e la fase di macroediting, eseguita mensilmente per la validazione finale dei dati prima della diffusione degli indicatori. Infine, vengono presentati i risultati di una ricognizione sugli errori non campionari più frequentemente registrati e una stima dell'impatto delle imputazioni delle mancate risposte totali sugli indicatori prodotti, che forniscono un prima misura della qualità del processo di controllo e correzione implementato per tale indagine.

1. Aspetti generali dell'indagine Grandi imprese (GI)

Campo di osservazione

L'universo teorico dell'indagine è costituito dall'insieme delle imprese appartenenti al settore privato non agricolo ad esclusione dei servizi sociali e personali (settori di attività economica da C a K della classificazione Ateco 2002) che nell'archivio di riferimento (Asia) hanno, nell'anno relativo alla base di riferimento, un numero medio annuo di dipendenti di almeno 500 unità.

In accordo con le definizioni e le metodologie prevalenti a livello internazionale, l'unità di rilevazione dell'indagine è l'impresa, mentre l'unità di analisi è l'unità funzionale. Tutte le imprese con oltre 500 addetti sono tenute a fornire i dati richiesti mensilmente per ogni unità funzionale, cioè per ogni unità o insieme di unità locali contraddistinte da una specifica attività economica (KAU). Tuttavia,

[3] Il lavoro è il risultato delle riflessioni comuni degli autori; tuttavia, i paragrafi 2.3, 2.4 e 3.2 possono essere attribuiti a Fabiana Rocci, mentre i paragrafi 2.1, 2.2 e 3.1 a Laura Serbassi. Il paragrafo 1, l'introduzione e le conclusioni sono a cura di entrambi gli autori.

[4] La diffusione avviene tramite il comunicato stampa "Lavoro e retribuzioni nelle grandi imprese". Assieme al comunicato stampa gli indicatori vengono messi a disposizione nella banca dati congiunturale Con Istat in forma disaggregata per gruppo Ateco e qualifica professionale.

nella pratica dell'indagine si riscontrano diverse eccezioni a questa regola, riferite ai casi in cui le imprese non dispongano di una contabilità del personale e retributiva distinta per singola unità funzionale.

Panel di rilevazione

La rilevazione è basata su un panel di imprese definito nell'anno base, che comprende solo imprese che abbiano almeno 500 dipendenti in media annua. Per quanto riguarda il panel 2005, rispetto all'universo delle posizioni dipendenti nelle imprese con almeno 500 addetti il grado di copertura aggregato risulta pari al 90,6 per cento (93,8 nell'industria e 88,7 nei servizi).

Si tratta di un panel chiuso per il quale, durante il periodo di vigenza di ciascuna base, non viene considerata la componente demografica di entrata–uscita dalla soglia dimensionale. L'obiettivo è quindi rilevare ogni mese le stesse unità statistiche, anche nel caso queste siano scese sotto la soglia occupazionale o abbiano subito delle trasformazioni giuridiche, e gli indicatori prodotti si riferiscono al solo insieme delle imprese oggetto di rilevazione.

Le unità prese in considerazione sono considerate tutte autorappresentative, per cui nel caso di mancate risposte totali è prevista una procedura di ricostruzione dei microdati delle imprese non rispondenti, mentre il trattamento statistico dei dati raccolti non contempla alcuna procedura di riporto all'universo.

In altri termini, nell'attuale approccio l'universo di riferimento dell'indagine è costituito dalle grandi imprese presenti nell'anno base e l'aggiornamento del panel d'indagine viene effettuato ogni cinque anni con l'introduzione della nuova base di calcolo dei numeri indice.

Grado di copertura

Attualmente, con la base 2005, il numero delle imprese del panel d'indagine è di circa 1.100 unità, che corrisponde a circa due milioni di posizioni lavorative (Prospetto 1). In termini generali, con riferimento al totale delle posizioni lavorative dei settori C-K presenti nell'archivio delle imprese attive Asia 2005, le imprese rilevate rappresentano mediamente il 20 per cento delle posizioni lavorative dipendenti; la quota è pari al 15,5 per cento nell'industria e al 24,3 per cento nei servizi.

Dal punto di vista delle unità funzionali, alle 1.107 imprese del panel 2005 corrisponde un totale di 1.439 unità funzionali, rispettivamente 832 unità per l'industria e 559 per i servizi.

Prospetto 1 – **Grado di copertura del panel 2005 per sezione di attività economica**

Sezioni di attività economica Ateco 2002	**Numero imprese indagine GI**	**Dipendenti indagine GI** (media 2005)	**Dimensione media imprese indagine GI**	**Dipendenti archivio ASIA 2005**	**Grado di copertura** (valore %)
INDUSTRIA	**607**	**788.326**	**1.299**	**5.074.742**	**15,5**
Estrazione minerali	2	12.268	6.134	37.377	32,8
Attività manifatturiera	538	679.982	1.264	3.850.508	17,7
Energia, gas ed acqua	43	76.979	1.790	114.497	67,2
Costruzioni	24	19.097	796	1.072.360	1,8
SERVIZI	**500**	**1.308.947**	**2.618**	**5.394.171**	**24,3**
Commercio	101	244.934	2.425	1.738.054	14,1
Alberghi e ristoranti	29	82.012	2.828	644.115	12,7
Trasporti, magazzinaggio e comunicazioni	105	483.385	4.604	1.016.108	47,6
Intermediazione monetaria e finanziaria	123	323.932	2.634	486.234	66,6
Altre attività professionali e imprenditoriali (a)	142	174.684	1.230	1.292.201	13,5
TOTALE	**1.107**	**2.097.273**	**1.895**	**10.468.913**	**20,0**

(a) Dalle Altre attività professionali e imprenditoriali sono escluse le imprese di lavoro interinale.

Modello di rilevazione

Lo strumento di rilevazione utilizzato è costituito da una scheda anagrafica e da un modello mensile di rilevazione. Come precedentemente specificato, le unità di rilevazione sono le unità funzionali, ciò implica che per alcune imprese debbano essere contemplati più modelli di rilevazione. Per questo motivo, è stato costruito un sistema di codici interni di indagine, ognuno dei quali identifica in modo univoco un'unità funzionale, mentre il codice fiscale rimane l'identificativo dell'impresa (a cui possono far capo più unità funzionali, quindi più codici di indagine). La scheda anagrafica consente di catalogare nel modo più corretto le informazioni concernenti l'identificazione della singola unità funzionale e del referente d'indagine, che vengono costantemente aggiornate nel corso dell'anno.

Il modello di rilevazione mensile è strutturato in 4 sezioni: occupazione dipendente, volume di lavoro, ore di cassa integrazione guadagni e spese per il personale, per un totale di 71 variabili suddivise su due qualifiche professionali. Tutte le variabili contenute nel modello sono quantitative, ad eccezione di tre domande con risposta dicotomica poste in calce al modello. Esse sono volte a conoscere se, nel mese di riferimento, c'è stata l'applicazione di un contratto integrativo o se vi sono state variazioni contrattuali dell'orario di lavoro e/o delle retribuzioni. Tali informazioni non sono utilizzate direttamente per la produzione di indicatori, ma servono a fornire indicazioni fondamentali per l'interpretazione dei dati presenti nel modello agli operatori, durante la fase di controllo e correzione.

La redazione del questionario da parte delle imprese avviene tramite autocompilazione con un tempo medio stimato di circa un'ora e un quarto.

Attualmente la trasmissione dei modelli può avvenire con tre modalità: a) su supporto cartaceo tramite fax; b) su supporto informatico (file excel) tramite posta elettronica[5]; c) tramite web sul sito dell'Istituto per il *data capturing*; mentre l'utilizzo della posta ordinaria sta progressivamente scomparendo. Nel 2007 la modalità di invio più utilizzata è quella via fax che rappresenta circa il 43 per cento degli arrivi, seguita da quella via mail (circa il 35 per cento) e dalla compilazione on line (circa il 22 per cento). In considerazione dei vantaggi che comporta il *data capturing*, quali il rispetto della segretezza, i brevi tempi di acquisizione dei dati già in formato elettronico e la possibilità di inserire nel modello controlli interattivi durante la fase di compilazione, negli ultimi anni è stato fatto uno sforzo considerevole per indurre i rispondenti ad utilizzare tale supporto. Tra il 2005 e il 2007 si è avuto un incremento del 22 per cento degli utilizzatori web nonostante la resistenza delle imprese a modificare le abitudini di compilazione acquisite nel corso degli anni.

Variabili rilevate

Tutte le informazioni del modello vengono acquisite distintamente per le categorie professionali degli operai e apprendisti e degli impiegati e intermedi.

I dirigenti vengono rilevati solo come consistenza alla fine del mese di riferimento nella sezione dedicata all'occupazione, mentre vengono esclusi in tutte le altre sezioni.

Per quanto riguarda le variabili contenute nel questionario, la prima sezione è diretta alla rilevazione dei dati sulle posizioni lavorative dipendenti. In particolare viene richiesto di indicare lo stock di occupati presenti alla fine del mese precedente a quello in corso di rilevazione, i flussi in entrata e in uscita nel corso del mese e lo stock di occupati presenti alla fine del mese di riferimento. L'informazione sugli occupati alla fine del mese precedente rilevata nel mese corrente è particolarmente importante, in quanto garantisce la continuità con il modello acquisito il mese passato, e rappresenta il primo controllo di coerenza dei dati forniti dall'azienda. I flussi in entrata vengono articolati per tipologia contrattuale distinguendo tra: entrati a tempo indeterminato, entrati con contratto a termine, con contratto di apprendistato, con contratto di inserimento, stagionali e entrati a seguito di operazioni societarie (fusioni, incorporazioni, acquisizione di ramo d'azienda, eccetera). Per gli usciti viene richiesta la causa di cessazione del rapporto, distinguendo tra cessazioni spontanee, cessazioni incentivate, cessazioni per raggiunti limiti di età, scadenza termini, licenziamento e usciti a seguito di operazioni societarie (cessioni di ramo d'azienda, scorpori, eccetera).

[5] Per motivi di sicurezza l'acquisizione dei modelli via mail dovrà essere al più presto sostituita dall'invio tramite web. Tuttavia, per evitare una caduta nel tasso di risposta la sostituzione sta avvenendo in modo graduale.

La seconda sezione è dedicata alle ore di lavoro, articolate in ore ordinarie effettivamente lavorate, ore di straordinario, ore non lavorate ma retribuite dal datore di lavoro e ore non lavorate per sciopero. Nella terza sezione vengono rilevate le ore di cassa integrazione guadagni utilizzate nel mese, distinte in ordinarie e straordinarie. La quarta ed ultima sezione è diretta all'osservazione delle spese retributive e contributive sostenute per il personale nel mese di riferimento. Per ciascuna delle due tipologie è richiesta una disaggregazione in componenti sulla base di un elenco predefinito.

La retribuzione lorda (al lordo delle ritenute previdenziali e fiscali) viene disaggregata in sette componenti:

1. retribuzione continuativa (compensi corrisposti sistematicamente ad ogni periodo di paga);
2. retribuzione per prestazione straordinaria;
3. mensilità aggiuntive (eccedenti le dodici);
4. premi e gratifiche legati a parametri gestionali e/o di redditività aziendale;
5. premi e gratifiche non legati a parametri gestionali e/o di redditività aziendale;
6. arretrati e una tantum;
7. incentivi all'esodo.

Le voci contributive a carico del datore di lavoro sono distinte in:

1. TFR (Conferimenti ai fondi e Accantonamento per rivalutazione);
2. versamenti aggiuntivi ai fondi di previdenza integrativa effettuati dal datore di lavoro;
3. contributi sociali legali, contrattuali e volontari (esclusi i conferimenti del TFR al fondo Inps);
4. provvidenze al personale.

Le informazioni acquisite si riferiscono alle effettive erogazioni mensili effettuate dalle imprese, secondo un criterio di cassa e non di competenza.

Il ciclo di produzione

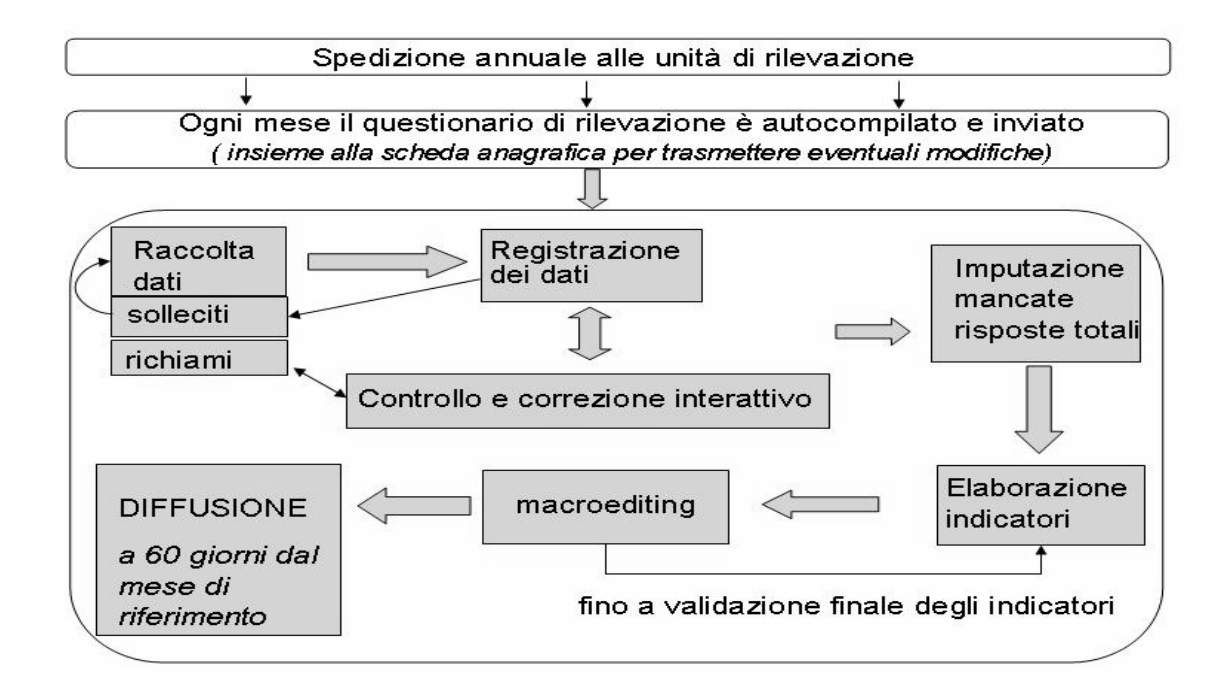
All'inizio di ogni nuovo anno viene inviato, a tutte le unità di rilevazione, un plico postale contenente il materiale da utilizzare durante tutto il ciclo di rilevazione annuale. Il plico contiene la circolare del Presidente, una scheda anagrafica, dodici modelli di rilevazione (uno per ciascun mese) e la guida alla compilazione. Sia la scheda anagrafica, sia i modelli di rilevazione vengono precompilati in alcune parti utilizzando i dati già presenti nell'archivio dell'indagine. In particolare nella scheda anagrafica vengono riportate tutte le informazioni concernenti l'identificazione dell'unità di rilevazione e del referente per l'indagine (che può essere lo stesso per diverse unità), mentre sui modelli di rilevazione viene prestampato il codice identificativo dell'unità, il codice di attività economica (Ateco 2002), l'anno e i mesi di riferimento dei dati.

Con l'acquisizione delle schede anagrafiche inizia la prima fase del ciclo produttivo. La scadenza per l'invio dei questionari mensili é il ventesimo giorno del mese successivo a quello di riferimento. Ogni mese, dopo la scadenza ufficiale, viene inviato alle imprese inadempienti un sollecito via fax o e-mail che definisce una ulteriore scadenza inderogabile (circa una settimana).

Decorso questo ultimo termine viene effettuata una ricognizione sull'intera lista delle imprese presenti nel panel, per individuare le mancate risposte totali che dovranno essere imputate con una procedura automatica. Successivamente viene eseguita la fase di elaborazione degli indicatori d'indagine, calcolati sui dati delle unità appartenenti al panel completo, costituito dai dati rilevati e stimati.

L'ultima fase del processo prevede la validazione degli indicatori finale attraverso una attività di macro editing e la diffusione dei dati.

Grafico 1 - Il ciclo di produzione mensile dell'indagine sulle grandi imprese



2. Il processo di controllo e correzione

2.1 Aspetti Generali del processo di controllo e correzione

Tenuto conto delle caratteristiche delle grandi imprese (tutte autorappresentative e potenzialmente influenti) e dei ristretti tempi di diffusione, la gestione dei dati richiede una trattazione attenta e puntuale. Il processo di controllo e correzione dell'indagine è fondato sull'utilizzo di un gruppo di sette revisori esperti, ciascuno dei quali gestisce in modo esclusivo un determinato insieme di imprese.

Le fasi dello schema di controllo e correzione possono essere così descritte :

1. controllo all'inizio dell'anno della correttezza della lista anagrafica e verifica dei referenti d'impresa;
2. registrazione mensile:
   - delle eventuali modifiche anagrafiche e longitudinali avvenute alle imprese;
   - dei modelli di rilevazione del mese di riferimento e di quelli dei mesi pregressi eventualmente arrivati in ritardo;
3. attività di sollecito;
4. imputazione delle mancate risposte totali;
5. validazione finale degli indicatori, attraverso una fase di controllo macro sui risultati ottenuti.

Tutto ciò che riguarda il punto 2. viene ripetuto ogni mese, ed è assimilabile ad un processo di microediting. I dati correnti sono registrati in un'apposita tabella sul database di indagine, dove ogni record è univocamente identificato con la data di riferimento (mese e anno) e il codice del modello.

Se i modelli arrivano via mail o fax, i dati vengono caricati dagli operatori direttamente nel bata base d'indagine attraverso una maschera elettronica di *data entry* (in Oracle forms) appositamente predisposta, che consente di eseguire la revisione dei dati contestualmente alla registrazione. Per i modelli arrivati via web si dispone direttamente della versione elettronica dei dati che vengono acquisiti in una tabella separata. Prima di essere inseriti nel data base d'indagine anche i modelli web vengono sottoposti dagli operatori ad una fase di microediting attraverso un'analoga maschera elettronica.

Il compito fondamentale degli operatori consiste, oltre alla registrazione dei modelli e delle variazioni anagrafiche, nella revisione dei dati per individuare ogni possibile errore nella compilazione dei questionari. A tal fine, entrambe le maschere elettroniche contengono numerosi controlli interattivi, definiti con riferimento alle possibili fonti di errore individuate sulla base dell'esperienza di rilevazione per ogni singola variabile. Tutti gli errori e le anomalie individuate, comprese le mancate risposte parziali, devono essere risolte dai revisori che contattano direttamente il referente interno all'azienda. Nel paragrafo successivo illustreremo nel dettaglio i controlli utilizzati e il funzionamento delle maschere.

Contestualmente alla registrazione dei dati, allo scadere della data prevista per l'invio, viene effettuato un sollecito mensile con l'obiettivo di accelerare l'acquisizione dei modelli mancanti. In diversi casi ad esso si accompagna il contatto telefonico diretto per garantire l'acquisizione dei modelli più importanti. In particolare per le imprese superiori a 15.000 dipendenti non viene utilizzata la procedura automatica di imputazione delle mancate risposte, di conseguenza nel caso di mancato invio viene fatta una sorta di intervista telefonica per acquisire le informazioni di base per la ricostruzione dei modelli. Si consideri che la prassi di effettuare un sollecito sistematico non solo ha sensibilmente migliorato il tasso di risposta, ma ha consentito di ridurre i tempi di diffusione dei dati che sono passati, tra il 2000 e il 2003, da 85 a 65 giorni dalla fine del mese di riferimento e dal 2003 al 2005 da 65 agli attuali 60. Oltre al sollecito mensile, due volte l'anno vengono inviati dei solleciti straordinari per integrare le serie storiche, in modo da garantire la completezza dell'archivio dei dati longitudinali, che è alla base dell'imputazione delle mancate risposte totali.

Dopo circa 50 giorni dalla fine del mese di riferimento è necessario chiudere le fasi relative alla raccolta e alla revisione dei dati per poter effettuare l'elaborazione e il controllo degli indicatori. Purtroppo, nonostante i solleciti e i ripetuti contatti, ogni mese una parte delle imprese del panel risulta non rispondente, in numerosi casi si tratta di modelli che arrivano in ritardo rispetto alle scadenza prevista per la diffusione dei dati. Infatti, la quota delle mancate risposte totali è di circa il 20 per cento al momento dell'elaborazione degli indicatori e scende di circa la metà a 80 giorni. Per i modelli non pervenuti in tempo utile viene utilizzata una procedura automatica di ricostruzione dei microdati in modo da ottenere ogni mese il panel completo (modelli arrivati + modelli stimati).

Una volta ricostruiti i dati di tutte le unità facenti parte del panel, si procede all'elaborazione degli indicatori. Infine, un ulteriore passaggio prima della diffusione degli indicatori prevede una attività di macroediting sulle principali variabili a livello aggregato di attività economica, al fine di trovare eventuali errori influenti sfuggiti alla fase di microediting.

Il trattamento dei dati anagrafici

L'invio delle informazioni anagrafiche viene effettuato dalle imprese e/o unità funzionali sulla base della scheda precompilata inviata con la spedizione annuale. Tra gennaio e febbraio le imprese devono restituire le schede con la conferma e/o modifica delle informazioni contenute. Successivamente, mese per mese, le eventuali modificazioni longitudinali vengono trasmesse via mail, telefono o web e registrate secondo una codifica standardizzata al fine di definire per ogni evento il trattamento statistico più adeguato.

Gli eventi di trasformazione giuridica delle imprese hanno effetto non solo sulla dimensione occupazionale delle unità di rilevazione, ma anche sulla loro collocazione in termini di attività economica. Nel corso del tempo, infatti, le imprese facenti parte del panel sono soggette a trasformazioni giuridiche di diversa natura (fusioni, cessazioni, scorpori, scissioni, eccetera), che possono comportare modifiche rilevanti nella composizione e nella struttura del panel e produrre negli indici variazioni spurie, che non derivano da effettive evoluzioni congiunturali dei fenomeni rilevati. In particolare, è apparso necessario distinguere tra gli eventi che coinvolgono imprese appartenenti al panel e quelli che avvengono tra imprese interne e esterne al panel. Nel primo caso, infatti, si deve operare in modo da minimizzare gli effetti spuri della trasformazione giuridica (in particolare sull'occupazione), mentre nel secondo gli effetti vengono trattati come derivanti da movimenti con effettiva valenza congiunturale.

Inoltre, la scelta strategica di integrare le diverse indagine (Gi, Oros e Vela) per la produzione dei dati sul mercato del lavoro si è basata sull'integrazione dei microdati. Attualmente, l'indagine Oros utilizza le informazioni sull'occupazione e le spese del personale, mentre l'indagine Vela sfrutta i dati sull'occupazione e sulle ore lavorate. Rispetto all'indagine Gi l'elemento identificativo delle unità che può essere utilizzato come raccordo dalle altre due indagini è il codice fiscale, che consente anche di individuare (e unificare) le diverse unità funzionali appartenenti alla medesima impresa. Un importante elemento da considerare è che il procedimento di *record linkage* effettuato de Oros e Vela avviene con un ritardo di due tre mesi rispetto all'indagine Gi. Ciò implica la necessità, da parte di Gi, di una gestione continua, puntuale e standardizzata delle informazioni identificative delle imprese (e delle unità funzionali), delle trasformazioni giuridiche e più in generale di tutti i cambiamenti nei codici fiscali, per minimizzare gli errori nella fase di integrazione degli archivi. Per superare i problemi legati allo sfasamento temporale è necessario ricostruire lo status di ogni impresa a ritroso nel tempo.

Nel database di indagine, sono predisposte due tabelle per la registrazione delle informazioni anagrafiche delle imprese. Una rappresenta la fotografia corrente, quanto più aggiornata, della situazione per le singole unità funzionali, mentre l'altra è stata costruita appositamente per mantenere le informazioni storiche, in modo da poter ricostruire in qualsiasi momento le vicende legate alle trasformazioni longitudinali.

In sintesi, sia l'adozione di un panel chiuso, sia l'utilizzo dei microdati da parte di altre indagini rendono di fondamentale importanza la corretta acquisizione e registrazione delle informazioni anagrafiche correnti e storiche.

2.2 Il microediting interattivo

Come accennato precedentemente, l'attività di microediting effettuata sui dati dell'indagine grandi imprese viene svolta dai revisori attraverso una maschera di controllo elettronica. Ogni modello di rilevazione, identificabile attraverso un codice impresa, viene visualizzato dall'operatore preposto e controllato in modo puntuale. Si sottolinea il fatto che non è prevista nessuna correzione automatica, tutti i problemi riscontrati (valori anomali, errori o semplici incertezze) vengono risolti tramite contatto diretto con le imprese (rete di referenti esterni – interni) e deve esserne appurata la causa.

Le singole variabili sono verificate sulla base di un piano di *check* predefinito (prospetto 2), nel caso in cui i vincoli imposti non siano rispettati la maschera mette in evidenzia le anomalie e visualizza i valori registrati dalla medesima variabile nei mesi precedenti.

I *check* immessi sulla maschera sono di diversa natura e hanno lo scopo di:

- individuare dati mancanti (mancate risposte parziali);
- individuare la presenza di errori sistematici, in particolare gli errori nell'unità di misura o quelli derivanti da una non comprensione della domanda;
- verificare che siano soddisfatte le coerenze algebriche fra le variabili che sono vincolate da totali;
- individuare errori attraverso il controllo della coerenza longitudinale dei dati storici dell'impresa stessa.

I casi di mancate risposte parziali sono molto limitati e solitamente dovuti a problemi occasionali dei rispondenti (modifiche di procedure informatiche, cambio di software, mancanza del referente aziendale, eccetera) che riescono comunque a rendere disponibili tutte le informazioni necessarie per effettuare una corretta imputazione della variabile mancante.

I *check* predisposti per l'individuazione degli errori sono definiti sulla base delle caratteristiche delle variabili di volta in volta sotto osservazione.

Nello specifico, il primo gruppo di controlli si basa sulle coerenze algebriche delle variabili rilevate, che riguardano:

i. la continuità tra modelli (relativi al mese *t* e *t-1)*: essa consiste nella verifica che, per ogni mese di rilevazione, il numero degli "occupati fine mese precedente" rilevati sul modello corrente coincida con gli occupati fine mese indicati sul modello del mese passato;

ii. nel singolo mese *t:*

ii1. che la somma algebrica degli occupati fine mese precedente e il numero dei dipendenti entrati ed usciti dia come risultato il valore dell'occupazione totale di fine mese;
ii2. che la retribuzione lorda sia uguale alla somma di quella continuativa e di tutte le parti accessorie riportate sul modello.

Successivamente, il secondo gruppo di controlli è basato sulla considerazione che alcune variabili hanno un andamento mensile caratterizzato da un profilo specifico di impresa. In base a tale ipotesi, un eventuale forte scostamento dai propri valori caratteristici deve essere analizzato, per valutare se esso sia un errore o un dato anomalo generato da uno specifico evento.

A tal fine, per ogni variabile sono state poste delle soglie per gli intervalli di variazione delle differenze tra i dati rilevati al mese *t* e i dati relativi ai mesi precedenti nella medesima unità. In particolare, il periodo utilizzato per il confronto è il mese precedente (variazione congiunturale) per le variabili caratterizzate da un comportamento legato alla congiuntura economica (come ad esempio la retribuzione continuativa), mentre per le variabili caratterizzate da un comportamento stagionale (ad esempio premi e gratifiche, mensilità aggiuntive) è lo stesso mese dell'anno precedente (variazione tendenziale), in diversi casi si utilizzano entrambi i riferimenti temporali. Quando i dati superano le soglie predefinite, essi vengono considerati anomali e quindi da verificare in modo puntuale. In questi casi l'operatore individua lo scostamento, ne accerta le cause, contattando il rispondente, e corregge l'errore o accetta il dato rilevato.

L'aspetto importante è che il confronto viene sempre effettuato sui dati storici dell'impresa stessa, siano essi valori pro capite, aliquote (come nel caso dei contributi sociali) o rapporti rispetto ad altre variabili contenute nel modello. Di conseguenza, l'efficienza e l'efficacia dei controlli dipende dalla completezza e dalla qualità dell'archivio storico che viene alimentato anche con i modelli arrivati in ritardo e con le stime.

Infine, alcune variabili particolari, che non hanno alcun tipo di regolarità (nemmeno specifiche di impresa), vengono validate rispetto a dei valori massimi uguali per tutte le imprese. Nel prospetto 2, sono riportati i principali controlli presenti nelle maschere

In conclusione l'attività di microediting dell'indagine si caratterizza per due aspetti. Il primo riguarda l'utilizzo dei dati longitudinali dell'impresa stessa, il secondo consiste nell'importanza attribuita all'attività dei revisori che devono conoscere a fondo ciascuna impresa, per interpretarne correttamente le informazioni e svolgere in modo efficace il controllo, ma soprattutto la correzione dei dati.Nel paragrafo 3 vengono illustrati i risultati di una ricognizione effettuata per verificare in cosa si sono concretizzati gli interventi dei revisori nel periodo 2003-2005 sui modelli acquisiti tramite web, per i quali si dispone sia dei dati grezzi che dei dati revisionati. Viceversa per tutte le altre modalità di acquisizione la revisione avviene contestualmente alla registrazione, di conseguenza i dati grezzi sono disponibili solo su supporto cartaceo rendendo di fatto non immediatamente documentabile l'attività dei revisori.

Prospetto 2 – Principali controlli presenti nelle maschere di registrazione

SEZIONE 1 - OCCUPAZIONE ALLE DIPENDENZE		TIPO CONTROLLO
Dipendenti alla fine del mese precedente	01	Uguale a dipendenti alla fine del mese del modello precedente (campo P16)
Entrati nel mese - **Totale**	02	Uguale alla somma degli entrati per contratto (cod. 03-08)
Usciti nel mese - **Totale**	09	Uguale alla somma di entrati per causa (cod. 10-15)
Dipendenti alla fine del mese	16	Uguale a fine mese precedente + entrati - usciti
Dirigenti	17	Minore di dipendenti fine mese (campo 16) Se valore = 0 e dirigenti mese precedente > 0 Visualizza valore mese precedente
Dipendenti a part-time	18	Minore di dipendenti fine mese (campo 16) Se valore = 0 e mese precedente > 0 Visualizza valore mese precedente
SEZ 2 - ORE DI LAVORO		
Ordinarie effettivamente lavorate	21	Somma variabili 21+23+31+32 maggiore o minore del 15% rispetto al pro capite dello stesso mese anno precedente della medesima impresa. Se non c'è il modello precedente il confronto viene fatto con un valore soglia predefinito.
Straordinarie	22	Il confronto viene fatto con un valore soglia predefinito.
Non lavorate ma retribuite dal datore di lavoro	23	Controllo congiunto con le ore ordinarie 21
Ore non lavorate a causa di scioperi	24	Il confronto viene fatto con un valore soglia predefinito.
SEZ 3 - CASSA INTEGRAZIONE GUADAGNI (Cig)		
Cig ordinaria: ore utilizzate	31	Controllo congiunto con sezione 2
Cig straordinaria: ore utilizzate	32	Controllo congiunto con sezione 2
SEZ 4 SPESE PER IL PERSONALE		
Retribuzione continuativa per prestazione ordinaria	41	Soglia +\- 8% rispetto al pro capite (lordo cig) del mese precedente e allo stesso mese dell'anno anno precedente Se non sono disponibili i dati storici visualizza valore
Retribuzione continuativa per prestazione straordinaria	42	Il confronto viene fatto con un valore soglia predefinito.
Mensilità aggiuntiva (eccedenti le dodici)	43	Soglia +\- 10% rispetto al pro capite stesso mese anno precedente (visualizza anche valore mese precedente)
Premi e gratifiche **legati** a parametri gestionali e/o di redditività aziendale	44	Somma 44+45 Se valore corrente o valore stesso mese anno prec. maggiore di soglia massima Visualizza pro capite mese corrente, mese precedente e stesso mese anno precedente
Premi e gratifiche **non legati** a parametri gestionali e/o di redditività aziendale	45	
Arretrati ed una-Tantum	46	Il confronto viene fatto con un valore soglia predefinito.
Incentivi all'esodo diversi dal TFR	47	Il confronto viene fatto con un valore soglia predefinito.
TOTALE	48	Uguale a somma voci P41-P47
Contributi sociali a carico del datore di lavoro		
TFR (Conferimento ai fondi e accantonamenti per rivalutazione)	49	Aliquota massima
Versamenti **aggiuntivi** ai fondi di previdenza integrativa effettuati dal datore di lavoro	50	Incidenza sulla retribuzione lorda al netto degli incentivi all'esodo (47) maggiore del 10%
Contributi sociali legali, contrattuali e volontari, esclusi (esclusi i conferimenti TFR al fondo Inps)	51	Incidenza sulla retribuzione lorda al netto degli incentivi all' esodo (47) +/- 10 % rispetto allo stesso mese anno precedente (con visualizzazione mese precedente)
Provvidenze al personale	52	

2.3 Il trattamento delle mancate risposte totali (MRT)

La procedura di imputazione delle mancate risposte ha l'obiettivo di stimare mensilmente le variabili relative ai modelli delle unità non rispondenti (mancate risposte totali). Le stime confluiscono nel calcolo degli indicatori finali, assicurando la completezza dei dati sul panel di imprese oggetto di indagine.

La procedura, nata come evoluzione di quelle precedenti[6], è stata introdotta con il passaggio alla base 2000 con l'obiettivo di ottimizzare l'utilizzo dei dati disponibili, sia longitudinali che trasversali. Il risultato ottenuto è la predisposizione di una metodologia deterministica, grazie alla quale viene sistematicamente ricostruito il microdato dell'unità statistica mancante. Ogni singolo dato imputato è il risultato di un'equazione definita sulla base delle caratteristiche economiche della variabile stessa, che sfrutta o le informazioni rilevate nelle unità rispondenti o quelle contenute nella storia dell'impresa non rispondente[7].

Le uniche variabili che non vengono imputate sono: le ore di sciopero (campo 24), gli incentivi all'esodo (campo 47), gli arretrati e una tantum (campo 48) e i contributi a fondi di previdenza integrativa (campo 50), che sono poste pari a zero.

Inoltre, la ricostruzione mensile delle mancate risposte garantisce la completezza delle serie storiche dei microdati di ogni singola impresa rispondendo anche al fabbisogno delle indagini utilizzatrici (Oros e Vela) dei dati sulle grandi imprese.

Equazioni di stima

L'occupazione fine-mese precedente (rilevata nel mese corrente) e il numero dei dirigenti sono le uniche variabili che subiscono un'operazione di *editing*. La prima variabile viene posta uguale all'occupazione fine-mese corrente rilevata nel mese precedente e la seconda viene posta uguale all'ammontare osservato nel mese precedente.

L'imputazione delle altre variabili è impostata su un modello che presuppone la soluzione di una serie di equazioni, risolte a cascata a partire dall'occupazione fine-mese corrente e dal valore pro capite che esse hanno assunto nei dati storici, riproporzionato con l'occupazione fine mese corrente stimata.

Le equazioni posso essere raggruppate secondo due criteri di base, uno secondo il quale si sfruttano le informazioni registrate sulle unità rispondenti (criterio A) e l'altro che sfrutta invece le informazioni tratte dal profilo longitudinale, che può essere di carattere congiunturale o tendenziale a seconda della variabile in questione, dell'unità sottoposta ad imputazione per mantenerne le specificità d'impresa (criterio B).

Introducendo i seguenti simboli:

j unità di rilevazione non rispondente;

S gruppo di attività economica a cui appartiene l'unità j-ma;

SRt insieme delle unità rispondenti nel mese t nel gruppo S.

Criterio A) nell'ipotesi che la dinamica delle variabili dell'occupazione e della retribuzione continuativa[8] sia analoga per tutte le imprese appartenenti alla medesima attività economica, le unità vengono aggregate nei gruppi Ateco 2002 (codici a tre cifre) di appartenenza[9]. In quest'ottica, per ogni mese di elaborazione e per ciascun gruppo, è possibile attribuire le variazioni registrate nelle unità rispondenti anche alle unità non rispondenti, nell'ipotesi appunto che l'andamento di tali variabili sia caratterizzato da un comportamento specifico di settore economico.

Il metodo utilizzato è assimilabile a quello denominato *ratio imputation*, molto usato nei modelli per statistiche economiche (cfr. Luzi et al., 2007, c.4.2), che è un caso particolare di regressione con un solo regressore. La stima dell'occupazione fine-mese corrente si ottiene moltiplicando il tasso di variazione

[6] Inizialmente (a partire dal 1972) l'obiettivo della procedura era la ricostruzione delle variabili relative all'occupazione, alle ore di lavoro e alle retribuzioni solo a livello aggregato.

[7] La procedura automatica di stima descritta nel presente paragrafo non viene utilizzata per l'imputazione delle mancate risposte totali delle imprese maggiori di 15.000 dipendenti e delle imprese soggette a trasformazioni giuridiche. In tali casi, la ricostruzione viene operata sulla base di informazioni puntuali riferite alla specifica situazione dell'impresa.

[8] La retribuzione continuativa comprende i compensi corrisposti sistematicamente ad ogni periodo di paga, quali: la paga base, l'indennità di contingenza, gli aumenti periodici di anzianità, i superminimi individuali e collettivi, gli aumenti di merito, ecc.; le maggiorazioni per lavoro notturno, festivo, in condizioni di disagio; le indennità di turno, le indennità di cassa, di maneggio valori e simili; le retribuzioni per ferie e festività; le indennità di alloggio e quelle di trasporto dal domicilio al posto di lavoro; l'incentivo al posticipo della pensione (superbonus) previsto dalla legge 243/2004; gli importi corrisposti in caso di malattia, maternità e infortuni sul lavoro; altro.

[9] Quando la numerosità delle unità presenti in un gruppo Ateco è inferiore alle 20 unità, i modelli vengono aggregati a livello superiore (di divisione, a 2 cifre).

occupazionale osservato sulle unità rispondenti appartenenti al gruppo dell'unità non rispondente, per il valore dell'occupazione fine-mese precedente dell'unità da stimare.

L'occupazione fine-mese viene così ottenuta:

$$\hat{Of}_{j,t} = \hat{Ofp}_{j,t}\, \Delta O_{SR_t}$$

$\hat{Ofp}_{j,t} = Of_{j,t-1}$ l'occupazione fine mese precedente nel mese *t* viene posta uguale a quella registrata alla fine del mese *t*-1;

$\Delta O_{SR_t} = \frac{Of_{SR_t}}{Ofp_{SR_t}}$ tasso di variazione occupazionale, calcolato come rapporto tra l'occupazione fine-mese corrente e l'occupazione fine-mese precedente registrate sulle unità rispondenti nel mese *t*.

Analogamente, la stima della retribuzione continuativa per una generica unità *j* non rispondente è ottenuta moltiplicando il tasso di variazione della retribuzione continuativa pro capite registrato dalle unità rispondenti del gruppo Ateco di appartenenza per il valore pro capite del mese precedente e per l'occupazione del mese corrente stimata dell'unità non rispondente:

$$\hat{R}_{j,t} = r_{j,t-1}\, \Delta r_{SR_t} \cdot \hat{O}_{j,t}$$

$r_{j,t-1}$ retribuzione continuativa pro capite della unità *j*-ma al tempo *t*-1;

Δr_{SR_t} tasso di variazione tra il mese *t* e il mese *t*-1 della retribuzione continuativa pro capite, registrato sulle unità rispondenti;

$\hat{O}_{j,t}$ media tra occupazione fine-mese corrente e fine-mese precedente ottenuti prima.

Criterio B) Le restanti variabili (quelle riferite al tempo di lavoro e tutte le voci retributive diverse dalla retribuzione continuativa) sono imputate in base al profilo longitudinale dell'unità non rispondente. L'ipotesi di base è che esse abbiano dei valori pro capite, o dei rapporti caratteristici (quote rispetto al totale) caratterizzati da comportamenti specifici d'impresa e presentino valori pressocchè costanti nel tempo. Quindi, viene calcolato il valore caratteristico di ogni variabile, in forma di rapporto rispetto a una variabile ausiliaria, al tempo in cui esso è giudicato significativo (congiunturale o tendenziale a seconda della variabile). Le variabili ausiliarie possono essere l'occupazione, la retribuzione continuativa o le ore ordinarie, precedentemente stimate o con il criterio A o in un passo precedente con il criterio B stesso. Questo evidenzia come anche le stime ottenute con questo criterio sono legate a quelle elaborate con il primo.

I livelli delle variabili sotto osservazione al tempo *t* vengono ottenuti, quindi, riproporzionando tali rapporti, registrati nell'unità *j*-ma nei dati storici, al livello stimato al tempo *t* della variabile ausiliaria. L'equazione per la stima della generica variabile *Z* si può scrivere nel seguente modo:

$$\hat{Z}_{j,t} = \frac{Z_{j,t-K}}{Y_{j,t-K}}\ \hat{Y}_{j,t}$$

dove la variabile *Y* e il *lag* temporale *K* sono scelte sulla base delle caratteristiche della variabile *Z*, nello specifico per ognuna delle singole variabili del modello si ha:

Z=Ore ordinarie effettivamente lavorate	⇒	*Y*=occupazione e *K*= 12
Z=Ore di cassa integrazione ordinaria e straordinaria	⇒	*Y*=occupazione e *K*= 1
Z=Numero di occupati in regime di part-time	⇒	*Y*=occupazione e *K*= 1
Z=Ore di straordinario,Ore non lavorate ma retribuite	⇒	*Y*=ore ordinarie e *K*= 12
Z=Mensilità aggiuntive e Premi e gratifiche	⇒	*Y*=retribuzione continuativa e K= 12
Z=Oneri sociali	⇒	*Y*=retribuzione lorda[10] e *K*= 12

Ogni modello stimato è registrato nel database dell'indagine cosicché le serie longitudinali (a partire da gennaio 2000) dei microdati delle unità appartenenti al panel di rilevazione sono sempre complete. Nel caso in cui, successivamente alla diffusione degli indicatori, i dati di un modello che era stato oggetto di stima vengano acquisiti, dopo aver superato i controlli standard, essi vengono sostituiti a quelli stimati in precedenza. In media, per circa il 40 per cento delle mancate risposte i dati reali arrivano con circa un mese di ritardo rispetto alla diffusione degli indicatori. Attualmente tali informazioni vengono utilizzate solo ai fini interni in quanto non è prevista nessuna revisione nei dati pubblicati[11].

Se la stima di un modello *j*-mo corrente richiede le informazioni storiche dell'unità, ove non siano disponibili dati reali (sia quelli utilizzati per l'elaborazione degli indicatori sia quelli arrivati in ritardo) vengono utilizzate le stime effettuate precedentemente.

Questo sistema assicura, da una parte la possibilità di stimare ogni mese tutti i modelli facenti parte del panel, dall'altra, un forte aggancio delle stime successive con i dati reali, particolarmente importante per le variabili imputate sulla base del dato congiunturale (occupazione e retribuzione continuativa).

2.4 Macroediting

Una volta acquisiti e registrati i dati di tutte delle unità facenti parte del panel (arrivati e stimati) è possibile elaborare gli indicatori dell'indagine. Prima del rilascio finale, tuttavia, viene eseguita un'ulteriore fase di controllo, al fine di individuare eventuali errori influenti sfuggiti alla fase di microediting e la presenza di *outlier* che devono essere evidenziati in fase di rilascio dei dati.

Tale controllo consiste in un'attività di macroediting, sulla quale si concentra tutto il gruppo di lavoro per circa una giornata, basata sull'analisi dei valori delle variazioni tendenziali dei principali indicatori prodotti per il totale dei dipendenti a livello di divisione Ateco. Gli indici presi in considerazione sono: l'indice dell'occupazione (netto e lordo c.i.g.), l'indice delle ore effettivamente lavorate pro capite, l'indice delle retribuzione continuativa pro capite, indice della retribuzione lorda pro capite e per ora lavorata e l'indice del costo del lavoro pro capite e per ora lavorata.

I livelli di riferimento delle variazioni considerati anomali non sono definiti a priori, bensì sulla base dell'esperienza degli analisti e delle informazioni derivanti dall'indagine sulle retribuzioni contrattuali, che rilascia i dati con trenta giorni di anticipo rispetto alla rilevazione sulle grandi imprese.

I casi che presentano variazioni anomale vengono analizzati nel dettaglio esaminando i microdati, al fine di individuare la causa che ha concorso a tale risultato. Generalmente l'attività di controllo accerta anomalie già controllate e validate dai revisori, in quanto derivanti da eventi specifici noti. Accade a volte che vengano individuate anomalie sfuggite al microediting, in questi casi si procede, a seconda della situazione, o alla correzione dell'errore o contattando l'impresa per ulteriori chiarimenti.

A questo punto gli indicatori vengono nuovamente elaborati e sottoposti a un ulteriore controllo fino alla validazione finale. Si consideri che, quando gli indicatori presentano delle variazioni anomale

[10] La retribuzione lorda considerata è al netto degli incentivi all'esodo (somma della retribuzione continuativa e delle varie componenti accessorie stimate nei passi precedenti).

[11] Attualmente l'indagine pubblica direttamente dati definitivi, la diffusione di indici provvisori con revisione a t+1 è stata abbandonata nel 1996.

particolarmente rilevanti, esse vengono messe in evidenza in sede di comunicato stampa con l'indicazione delle principali cause che le hanno generate.

3. Documentazione e indicatori[12]

L'attuale processo di controllo e correzione dell'indagine sulle grandi imprese consente alcune analisi sulle singole fasi. Sull'attività di microediting, attraverso il confronto tra dati grezzi e dati revisionati, e sull'attività di imputazione delle mancate risposte totali, attraverso il confronto tra i dati reali arrivati in ritardo rispetto ai tempi di diffusione dei dati e le stime effettuate al momento del calcolo degli indicatori.

3.1 Statistiche sull'attività di microediting

Per quanto riguarda l'attività di microediting, la possibilità di effettuare un'analisi sull'attività dei revisori è condizionata alla modalità di compilazione utilizzata dalle imprese. Secondo il tipo di acquisizione del modello si hanno due possibili situazioni:

- se il modello è stato inviato via web esso è disponibile direttamente su supporto informatico e si dispone della registrazione sia del dato grezzo, sia di quello validato;
- se invece il modello è stato inviato via fax o e-mail, non rimane traccia telematica del dato grezzo ma soltanto di quello validato. In tale caso, infatti, nella fase di registrazione non è contemplata la possibilità di salvare il dato originario ma soltanto la versione revisionata.

Sui modelli pervenuti tramite web nel periodo gennaio 2003 settembre 2005 (circa 5800 modelli, pari al 15 per cento del totale) è stata svolta un'analisi di qualità confrontando i dati grezzi con quelli finali validati. In tal modo si è ottenuta una misurazione dell'attività dei revisori sia in termini di entità, sia in termini di tipologia degli interventi di correzione effettuati.

Prospetto 3 - Distribuzione dei modelli web per stato e tipologia di variabile *(valori percentuali)*

	Totale variabili	Variabili occupazionali	Variabili orarie	Variabili retributive
Modelli che non hanno subito rettifiche	57,7	90,5	91,6	66,0
Modelli che hanno subito rettifiche	42,3	9,5	8,4	34,0
Totale	**100,0**	**100,0**	**100,0**	**100,0**
Numero medio di variabili rettificate per modello	5,9	3,1	1,8	6,0

La prima informazione ottenuta è che sul 57,7% dei modelli considerati i revisori non hanno effettuato nessun intervento di modifica, vi è quindi perfetta identità tra il modello originale e quello validato. Sul restante 42,3% dei modelli è stata rilevata almeno una differenza tra i dati grezzi e quelli definitivi. Tenendo conto del fatto che diverse variabili sono legate tra loro da vincoli di coerenza, il numero medio di variabili rettificate per modello è di 5,9. La distribuzione per tipologia di variabile evidenzia la maggiore problematicità di quelle retributive che risultano rettificate nel 34% dei modelli, rispetto a quelle occupazionali e orarie dove gli interventi hanno riguardato meno del 10% dei casi.

In considerazione del fatto che tra gli interventi di rettifica vi sono oltre alle correzioni vere e proprie, anche numerosi interventi che non influiscono sul valore (e quindi sulla qualità) del dato fornito (arrotondamenti, cancellazione di spazi, eccetera), sui modelli che hanno subito delle rettifiche durante il processo di microediting è stata effettuata per alcune variabili un'analisi puntuale al fine di individuare la tipologia più frequente di intervento. Nel prospetto che segue sono riportati i principali tipo di rettifica per tre variabili: retribuzione continuativa, mensilità aggiuntive e ore ordinarie effettivamente lavorate.

[12] Le analisi presentate in questo paragrafo utilizzano i dati del panel 2000.

Prospetto 4 - Variabili per tipologia di rettifica *(valori percentuali)*

Tipo di rettifica	Retribuzione continuativa		Mensilità aggiuntive		Ore ordinarie	
	Impiegati	Operai	Impiegati	Operai	Impiegati	Operai
Arrotondamento	22,6	21,6	9,6	7,5	-	-
Unità di misura	49,4	53,9	54,8	64,2	7,4	2,3
Rettifiche (*)	25,8	22,1	32,0	23,5	74,3	81,8
di cui: minore o uguale al 10%	*12,0*	*9,3*	*15,9*	*12,8*	*11,8*	*7,0*
compreso tra il 10 e il 30%	*8,0*	*6,4*	*6,5*	*5,3*	*29,7*	*18,1*
superiore al 30%	*5,8*	*6,4*	*9,6*	*5,3*	*32,8*	*56,7*
Altro	2,2	2,4	3,6	4,8	18,3	15,9
Totale	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

(*) si tratta di interventi di rettifica di errori rilevati dagli operatori. In questo caso sono stati distinti tre livelli di correzione basati sulla differenza percentuale tra il dato definitivo e quello grezzo (dato revisionato/dato grezzo *100-100).

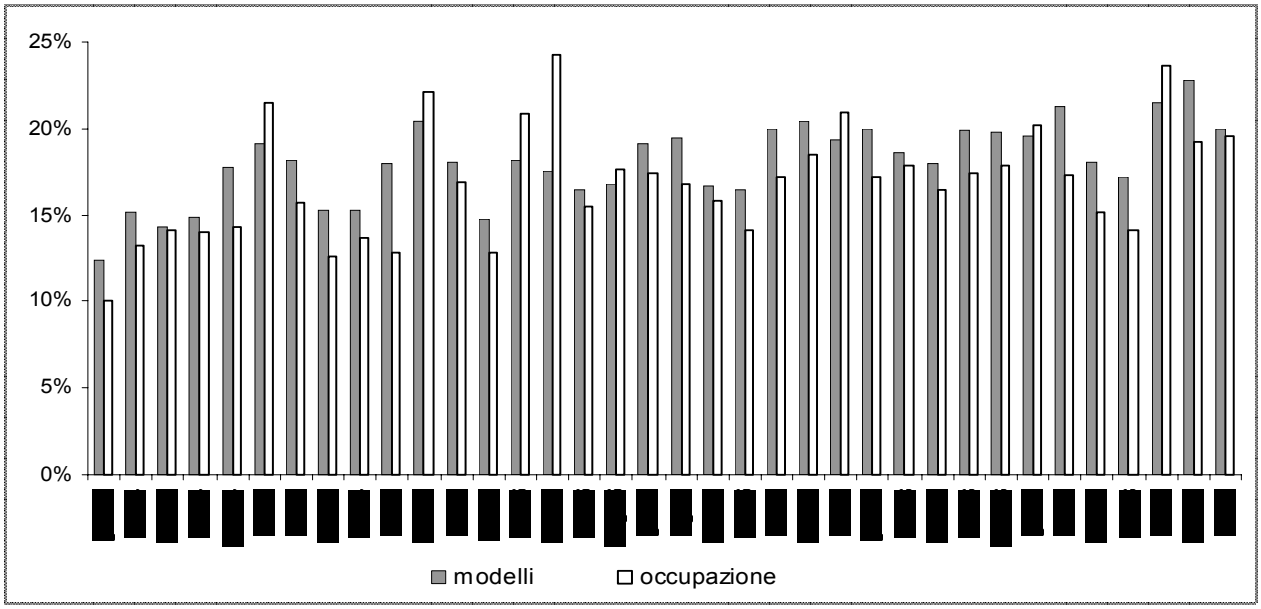
Per quanto riguarda le variabili economiche si noti che la rettifica dell'unità di misura risulta l'intervento più frequente, pari a circa la metà del totale. Nel caso degli arrotondamenti, ossia di rettifiche di piccolissima entità volte a modificare i decimali dopo la virgola generalmente al fine di garantire la coerenza della somma delle diverse voci retributive, la quota di interventi è circa il 22 per cento per la retribuzione continuativa, di circa il 10 per cento nelle mensilità aggiuntive degli impiegati e del 7,5 per cento negli operai. La diversa frequenza degli interventi tra le due variabili è riconducibile alla diversa natura delle due poste. Si consideri che mentre la retribuzione continuativa è una voce presente in tutti i periodi di retribuzione, le mensilità aggiuntive vengono erogate generalmente soltanto due volte l'anno (tredicesima e quattordicesima).

Infine, gli interventi di rettifica vera e propria presentano valori molto diversi tra le variabili considerate e variano dal 22,1 per cento della retribuzione continuativa operai all'81,8 per cento delle ore ordinarie operai. Si noti che per le due variabili economiche la quota di tali interventi costituisce circa un quarto del totale, con una netta prevalenza delle revisioni di bassa entità, mentre per le ore lavorate esse rappresentano una quota prossima o superiore ai tre quarti di tutte le rettifiche. In particolare, il fatto che esse si concentrino prevalentemente nella categoria delle rettifiche "superiore al 30%" lascia sottintendere più che un problema di revisione un problema di mancata risposta parziale.

3.2 Indicatori di qualità dell'imputazione delle mancate risposte totali (MRT)

La procedura di imputazione delle MRT viene tenuta costantemente sotto controllo, poiché, come evidenziato nel grafico 2 in alcuni mesi anche un tasso di mancata risposta inferiore al 20 per cento in termini di modelli, può comportare la stima di una quota più alta dell'occupazione (ad esempio marzo 2005), variabile pivot per la stima di tutte la altre.

Grafico 2 - Numero di modelli e occupazione stimata per mese di rilevazione. Anni 2004 - 2006 (*quote percentuali sul panel 2000)*



Nel prospetto 5 vengono riportate le medie annue, sia della quota dei modelli, sia della relativa occupazione stimata. La media mensile dei modelli stimati è cresciuta nel corso del biennio 2004-2006, sia per motivi legati al naturale attrito di cui soffrono le rilevazioni statistiche, in particolare quelle congiunturali, sia della diminuzione della distanza in giorni della data di rilascio dei dati rispetto al mese di riferimento degli stessi.

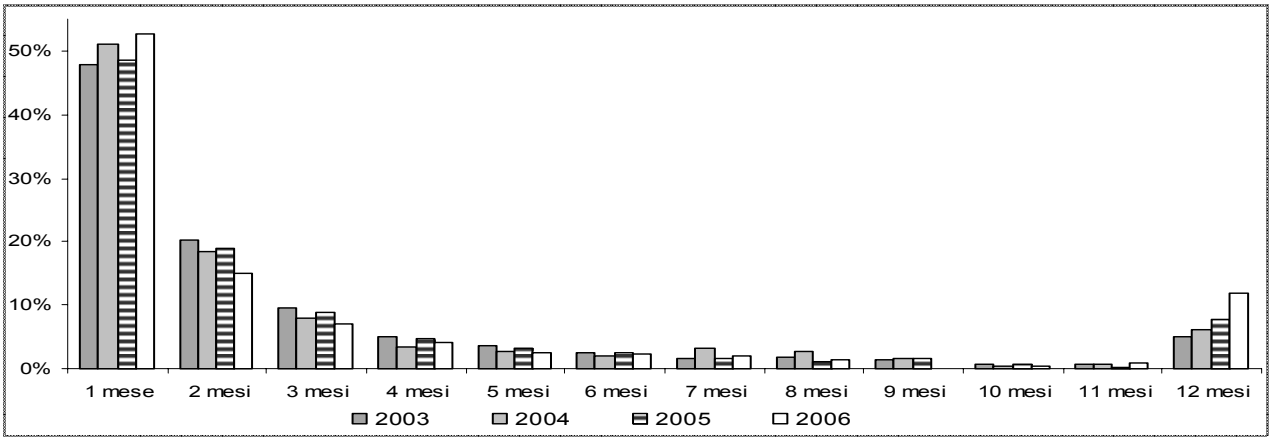
Prospetto 5 - Quota media annua di modelli e di occupazione stimata

Anno	Quota media annua di modelli stimati *(valore percentuale)*	Quota media annua di occupazione stimata *(valore percentuale)*	Giorni di distanza tra la fine del mese di riferimento e la diffusione dei dati
2004	16,6	15,1	63
2005	17,9	17,6	60
2006	19,7	18,0	60

Un aspetto molto interessante è conoscere il numero di mesi consecutivi che, nell'ambito del calcolo degli indicatori ufficiali, un modello relativo alla stessa unità funzionale viene stimato. Questo permette di monitorare l'efficacia della costante attività di ricognizione e dei solleciti sistematici, sia telefonici che per mail o fax.

Come evidenziato dal grafico 3, dal 2003 al 2006 è stato sempre possibile tenere la percentuale dei record stimati per un solo mese vicino al 50 per cento. Ciò significa che è alta la percentuale di imprese che, dopo un solo mese in cui sono state stimate, tornano a rispondere in tempo utile per l'elaborazione degli indicatori. Questo ci consente di affermare che il fenomeno di mancata risposta totale ha carattere per lo più accidentale, dovuto a un ritardo nella trasmissione dei dati, e non a comportamenti assimilabili a rifiuti nella risposta. Purtroppo, se da un parte l'aumento di tale percentuale può essere considerato un successo, si registra un acutizzarsi del fenomeno comunemente denominato attrito del panel; per cui alcune unità, anche se più volte sollecitate, tendono a non rispondere. Nel nostro caso, la percentuale dei modelli che non sono pervenuti per tutti i dodici mesi dell'anno è passata dal 5,0 per cento del 2003 al 11,9 del 2006.

Grafico 3 - Mesi consecutivi di ricostruzione dei modelli non rispondenti. Anni 2003-2006



Impatto dell'imputazione dei non rispondenti sugli indici finali

Come più volte ripetuto, per i modelli arrivati in ritardo si dispone sia della stima, che è stata utilizzata per il calcolo degli indicatori diffusi, sia del dato reale. In questo caso, quindi, è possibile analizzare la "bontà" delle stime effettuate e dare una misura dell'impatto delle stime sugli indicatori finali.

Considerando che dal 2004 al 2006 per le stime effettuate si dispone di circa il 40 per cento dei modelli rilevati (su un totale di 8397 stime, sono disponibili 3517 modelli arrivati in ritardo), presentiamo di seguito i risultati di una simulazione che sfrutta tali dati al fine di misurare l'impatto delle stime sugli indicatori. Utilizzando i modelli arrivati in ritardo è stato possibile ricalcolare gli indicatori principali sostituendo alla stima il dato reale.

Dati $I_{Y,t}$ e $I_{Y,t}^*$ rispettivamente gli indici pubblicati e quelli ricalcolati per la generica variabile Y al tempo t, ne vengono calcolate le differenze assolute come misura dell'impatto delle imputazioni delle mancate risposte totali.

$DF_{Y,t} = I_{Y,t} - I_{Y,t}^*$ differenza assoluta tra l'indice pubblicato e quello ricalcolato per la generica variabile Y per ogni t, dal 2004 al 2006.

Nei quattro grafici che seguono vengono mostrati i risultati ottenuti per gli indici dell'occupazione lorda, della retribuzione continuativa pro capite, delle ore lavorate pro capite e della retribuzione lorda pro capite.

Si noti, in particolare, l'entità più contenuta delle differenze tra gli indici dell'occupazione e della retribuzione continuativa pro capite (grafici 4 e 5) rispetto a quelle registrate sugli indici delle ore lavorate e della retribuzione lorda pro capite (grafici 6 e 7). Nel primo caso le stime hanno avuto un impatto minimo sugli indicatori, evidenziando la distanza contenuta tra dato stimato e dato reale Ciò conferma il fatto che l'occupazione e la retribuzione continuativa sono variabili che per natura hanno una variabilità molto bassa.

Grafico 4 - Indice dell'occupazione lordo c.i.g. nelle grandi imprese: differenze tra gli indici diffusi e quelli ricalcolati. Anni 2004-2006 *(differenze assolute tra numeri indice in base 2000=100)*

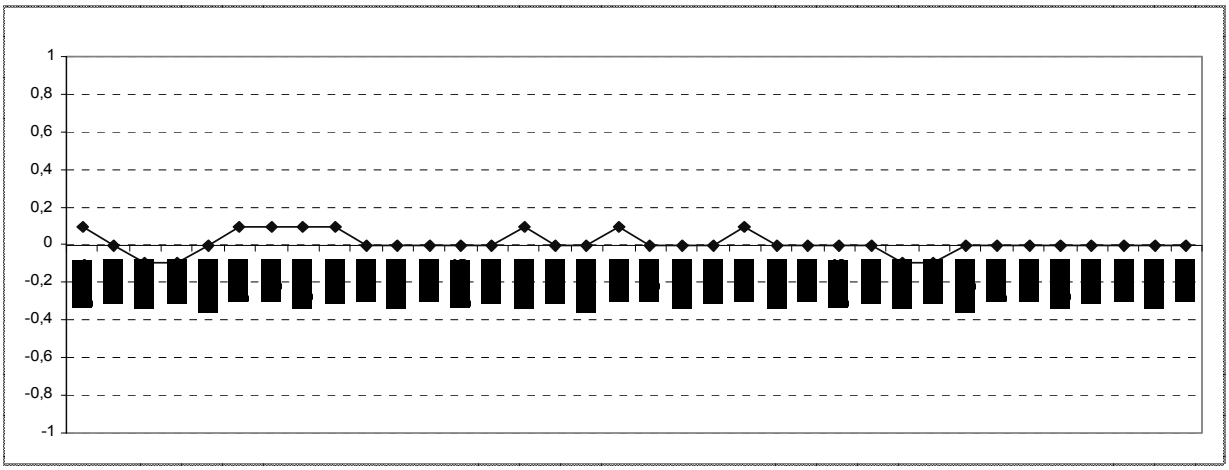


Grafico 5 - Indice della retribuzione continuativa pro capite nelle grandi imprese: differenze tra gli indici diffusi e quelli ricalcolati. Anni 2004-2006 *(differenze assolute tra numeri indice in base 2000=100)*

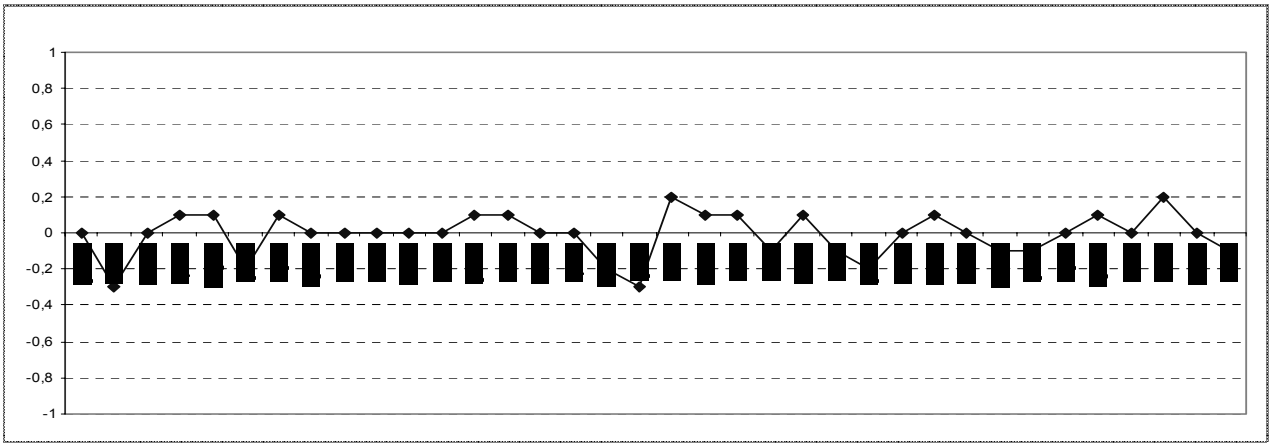


Grafico 6 - Indice delle ore lavorate per dipendente nelle grandi imprese: differenze tra gli indici diffusi e quelli ricalcolati. Anni 2004-2006 *(differenze assolute tra numeri indice in base 2000=100)*

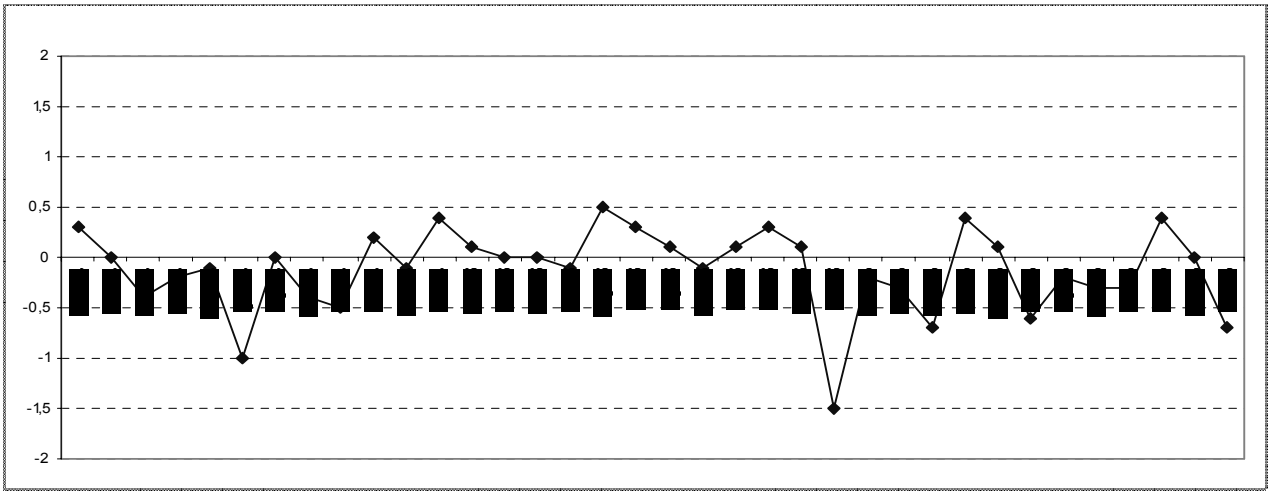
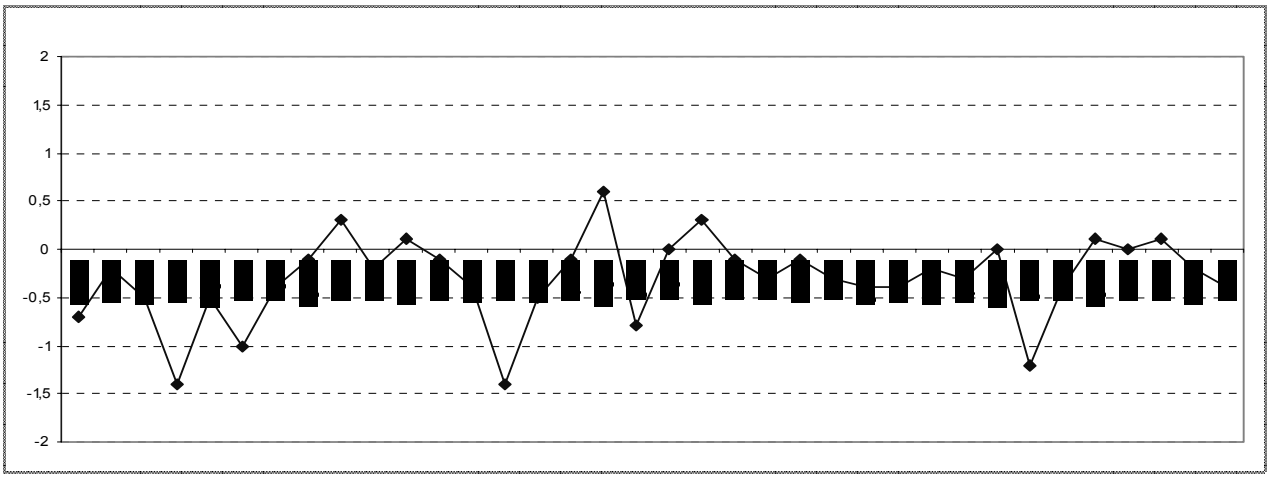


Grafico 7 - Indice della retribuzione lorda pro capite nelle grandi imprese: differenze tra gli indici diffusi e quelli ricalcolati. Anni 2004-2006 *(differenze assolute tra numeri indice in base 2000=100)*



Per consentire un'analisi più dettagliata, abbiamo calcolato ulteriori indicatori quali il valore massimo e il valore minimo, la media delle differenze e la media dei valori assoluti delle differenze (prospetto 6).

Prospetto 6 – Massimo, minimo e media delle differenze tra gli indici diffusi e quelli ricalcolati nel periodo gennaio 2004 dicembre 2006

Differenze assolute / Indicatori	Occupazione lordo c.i.g.	retribuzione continuativa pro capite	ore lavorate per dipendente	retribuzione lorda pro capite
Max.	0,1	0,2	0,5	0,6
Min.	-0,1	-0,3	-1,5	-1,4
Media	0,0	0,0	-0,1	-0,3
Media delle differenze in valore assoluto	0,0	0,1	0,3	0,4

In linea generale consideriamo i risultati ottenuti soddisfacenti, infatti tutte le variabili presentano una media degli errori molto vicina allo zero. Per la precisione, a conferma delle differenze mensili riportate nei grafici, l'occupazione e la retribuzione continuativa hanno una media uguale a zero; inoltre il range della prima variabile è estremamente contenuto (tra meno e più 0,1). Le variabili ore lavorate e retribuzione lorda hanno una media diversa da zero, ma comunque molto bassa. Tuttavia, si notano dei valori di massimo e di minimo più alti, in particolare per lo ore lavorate.

Successivamente, distinguendo le variabili sulla base del criterio di stima utilizzato, possiamo sottolineare che le prime due sono stimate sulla base delle informazioni raccolte sui rispondenti nello stesso mese (cirterio A), mentre le altre due sulla base del profilo dell'impresa stessa (criterio B).

Quindi, da una parte è si è avuta una conferma che l'ipotesi secondo cui la dinamica dell'occupazione e della retribuzione continuativa può essere rappresentata dalla quella registrata sulle unità rispondenti è valida. Mentre l'imputazione delle variabili, di per sé più irregolari, seguendo strettamente il profilo di impresa può portare a una maggiore variabilità nelle differenze tra dato reale e dato stimato. Ciò accade per le ore lavorate, i cui sbalzi possono dipendere anche, a parità di giorni lavorativi, dalle modalità con

cui i dipendenti stessi usufruiscono delle ferie e per la retribuzione lorda, formata oltre che dalla retribuzione continuativa anche dalle componenti saltuarie la cui erogazione può subire cambiamenti (diversa o mancata erogazione dei premi da un anno all'altro o anche slittamenti tra mesi successivi) non prevedibili.

Le differenze riscontrate nei risultati delle stime tra i due criteri, suggeriscono che per ottimizzare ulteriormente lo stimatore basato sul criterio B sia il caso di contemplare lo studio di un approccio più differenziato tra i due criteri, in particolare che sia più flessibile rispetto alle caratteristiche della variabile in questione.

Conclusioni e prospettive

Nel corso del tempo, l'indagine sulle grandi imprese ha subito diverse ristrutturazioni per uniformarsi ai mutamenti avvenuti nelle unità di rilevazione e per favorire i miglioramenti resi possibili dalle nuove tecnologie informatiche sia nella gestione e nell'archiviazione dei dati, sia nell'implementazione di nuove metodologie di calcolo. Una parte fondamentale in tale processo è stata svolta dagli operatori stessi, che per anni hanno lavorato al controllo dei microdati; infatti, come abbiamo descritto, tutti i controlli sono stati formulati tenendo conto delle loro considerazioni sulle possibili fonti di errore non campionario. Anche gli studi preliminari alla definizione della metodologia di imputazione delle mancate risposte sono stati condotti al fine di analizzare e validare le ipotesi fatte sulle singole variabili.

In questa ottica, dopo qualche anno dall'introduzione dell'informatizzazione e della standardizzazione nel processo di produzione dei dati, è molto interessante monitorare i risultati ottenuti in termini di capacità di prevenzione degli errori non campionari, di tipologia degli errori rilevati e delle revisioni effettuate, così come valutare la qualità delle imputazioni delle mancate risposte totali rispetto ai dati reali e l'impatto di queste sugli indicatori finali. Per quanto riguarda il primo aspetto, una valutazione è stata possibile sul sottoinsieme dei dati acquisiti via web. Ciò che è emerso dall'analisi è da un lato, la presenza di numerosi interventi che potrebbero essere oggetto di un trattamento più automatizzato e standardizzato, ma anche la criticità di alcune variabili per le quali, in fase di controllo, rimane fondamentale ricontattare il rispondente e quindi mantenere la presenza di operatori che alimentino un rapporto di collaborazione con tali imprese.

I confronti tra le stime delle MRT e i corrispondenti dati reali mostrano un ottimo risultato in termini di impatto sugli indicatori finali, a conferma della validità delle ipotesi fatte sulla base dell'esperienza degli operatori.

In conclusione, le fasi del processo di controllo e correzione che presentano margini di implementazione, con l'obiettivo di aumentare il livello di qualità o l'efficienza dell'intero processo , sono:

- l'attività di microediting attualmente consente di avere una buona qualità dei dati, tuttavia si tratta di un'attività molto costosa che impegna circa il 75% delle risorse disponibili, sarebbe pertanto molto utile definire dei controlli automatizzati in modo da ridurre la quantità di lavoro manuale necessario in tale fase e standardizzare gli interventi di rettifica. Un requisito importante per incrementare il livello di informatizzazione del processo di controllo e correzione è aumentare la quota dei modelli acquisiti attraverso il web (*data capturig*). In questo modo sarebbe possibile ridurre a priori la numerosità degli errori non campionari inserendo una serie di controlli interattivi in fase di compilazione, ma soprattutto disporre dei dati direttamente su supporto informatico che permetterebbe di monitorare il processo di controllo e correzione in modo efficiente e a costi in termini di documentazione davvero bassi;
- la fase del macroediting, ovvero valutare la possibilità di attivare una procedura standard e automatizzata, che guidi gli analisti nell'individuazione dei dati influenti e degli outlier.
- sviluppare una maggiore integrazione con le indagini Oros e Vela, che alimenti uno scambio bidirezionale delle informazioni. Ossia, oltre all'attuale fornitura dei microdati da parte della rilevazione

Gi, avviare un flusso inverso per sfruttare i dati relativi a tali indagini nell'imputazione delle MRT di imprese con un tasso di risposta molto basso;

- studiare in modo approfondito le caratteristiche degli errori commessi nelle stime delle MRT, al fine di ridurre ulteriormente lo scarto tra queste e i dati reali, soprattutto per quanto riguarda le variabili meno regolari. In questo contesto, impostare degli studi più ampi sulla natura dei dati e quindi delle variabili sotto osservazione, per individuare dei modelli teorici di riferimento che permetterebbero, in un ambito di inferenza statistica, maggiori strumenti di stima e di controllo ed eventuale correzione dell'errore di stima. In particolare, è stata impostata un'ipotesi di studio che verifichi le condizioni per cui il metodo utilizzato secondo il criterio A (ora costruito sulla base di un modello di *ratio imputation*) possa essere impostato come una regressione nei modelli di superpopolazione (Cicchitelli et al., 1992). In tale caso, infatti, sarebbe possibile sviluppare lo stimatore attualmente utilizzato per ottenerne uno che si dimostra essere ottimo (in termini di scarto quadratico medio minimo) nella classe di tutti i possibili stimatori.

Bibliografia

AA.VV. *Model quality Report in Business Statistics. Volume 1. Theory and Methods for Quality Estimation*. Eurostat, 2001.

Amato G., Gismondi R., Rocci F., Serbassi L. *L'effetto delle modificazioni longitudinali delle unità statistiche e di diverse metodologie di calcolo sugli indici dell'occupazione e delle retribuzioni nelle grandi imprese*. Documento interno Istat, Roma, 2007.

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

Cicchitelli G., Herzel A. e Montanari G. E. *Il campionamento statistico*. Il Mulino, Bologna, 1992.

Hidiroglou M.A., Berthelot J.M. "Statistical Editing and Imputation for Periodic Business Survey". *Survey Methodology*, 12, Statistics Canada, Ottawa, 1986.

Istat. *Metodi statistici per il record linkage*. Collana Metodi e Norme n. 16/2003.

Istat. *Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese*, Collana Metodi e Norme n. 29/2006. Disponibile sul sito www.istat.it/dati/catalogo/

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Rapporto tecnico del progetto Europeo EDIMBUS, 2007. (http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/RPM_EDIMBUS.PDF)

Oneto G. P. *La diffusione delle tecnologie di data capturing telematico nelle indagini congiunturali presso le imprese: risultati dell'analisi e proposte operative*. Documento interno Istat, Roma, 2006.

Il controllo e la correzione in una indagine congiunturale basata su dati amministrativi. Il caso della rilevazione Oros

Ciro Baldi, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
Francesca Ceccato, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
Eleonora Cimino, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
M. Carla Congia, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
Silvia Pacini, Fabio Rapiti, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi
Donatella Tuzi, Istat, Servizio Statistiche congiunturali sull'occupazione e sui redditi

Sommario: Il lavoro descrive le principali caratteristiche del processo di controllo e correzione della rilevazione trimestrale Oros che produce indicatori su retribuzioni di fatto e costo del lavoro utilizzando i dati amministrativi di fonte INPS integrati con i dati della Rilevazione mensile Istat sulle imprese di grandi dimensioni. Tale processo di controllo e correzione è particolarmente articolato e si contraddistingue da quello di altre rilevazioni congiunturali classiche per molteplici aspetti: la forte dipendenza dall'ente fornitore, l'enorme massa di dati amministrativi dettagliati e disaggregati, l'evoluzione continua del contenuto informativo della fonte amministrativa, l'integrazione con i dati d'indagine. La strategia adottata per garantire la qualità degli indicatori prodotti si basa principalmente su procedure specifiche per il controllo preliminare dei dati amministrativi e la loro traduzione in variabili statistiche, una sequenza di controlli sistematici per individuare gli errori influenti in tutte le fasi del processo, procedure altamente selettive e interattive e una documentazione sistematica che garantisce la riproducibilità e la ripetibilità del processo stesso.

Parole chiave: dati amministrativi, indicatori congiunturali, controllo e correzione, integrazione.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Introduzione[13]

Gli indicatori trimestrali Oros (Occupazione, Retribuzioni, Oneri Sociali) su retribuzioni di fatto e costo del lavoro per Unità di lavoro equivalenti a tempo pieno (Ula) vengono stimati integrando i dati amministrativi di fonte INPS con informazioni tratte dalla Rilevazione mensile dell'Istat su occupazione, orari di lavoro e retribuzioni nelle grandi imprese (da ora in poi, GI). La popolazione obiettivo è rappresentata dalle imprese attive con almeno un lavoratore dipendente, nei settori di attività economica dell'industria e dei servizi privati (sezioni da C a K dell'Ateco 2002).
Il processo di produzione dei dati Oros risulta estremamente complesso per varie ragioni, da attribuire principalmente: all'uso di dati amministrativi molto dettagliati e disaggregati (l'intera dichiarazione contributiva mensile DM10 dell'INPS), all'enorme massa di micro dati utilizzati (oltre 60 milioni di record ogni trimestre), alla necessità di integrare i dati amministrativi con quelli della Rilevazione GI, all'obbligo di produrre e diffondere stime con elevatissima tempestività come prescritto dai regolamenti europei, alla necessità di rivedere regolarmente le stime effettuando ogni trimestre, oltre alla stima provvisoria del trimestre corrente *t*, la stima definitiva del trimestre *t*-5. Il processo di controllo e correzione (C&C) è pertanto particolarmente articolato e si differenzia notevolmente da quello di altre rilevazioni congiunturali tradizionali per il grado di pervasività che si estende a tutte le fasi della rilevazione, dalla riaggregazione iniziale dei dati amministrativi, alla validazione finale dei macro dati, interessando persino i metadati.
Questo documento illustra la strategia, le caratteristiche e le diverse fasi del processo di controllo e correzione nella produzione corrente della rilevazione Oros. Il paragrafo 2 descrive brevemente le principali caratteristiche della rilevazione. Il paragrafo 3 affronta le peculiarità di una rilevazione basata su dati amministrativi e le conseguenze sul processo di C&C. Il paragrafo 4 presenta in sintesi le diverse fasi in cui si struttura il processo di C&C. Dal quinto al dodicesimo paragrafo vengono descritte in dettaglio le varie fasi del processo di C&C. Nel tredicesimo paragrafo vengono presentate alcune considerazioni conclusive.

2. I principali aspetti della rilevazione Oros

Per comprendere le peculiarità del processo di C&C Oros è necessario descrivere brevemente gli aspetti principali della rilevazione (Baldi et al. 2004).

- Le caratteristiche, il flusso e la tempistica dei dati amministrativi inviati dall'INPS all'Istat. La principale fonte dei dati è costituita dalle dichiarazioni dei contributi previdenziali e assistenziali (modello DM10) presentate mensilmente all'INPS dalle imprese con almeno un lavoratore dipendente[14]. La dichiarazione deve essere inviata entro 30 giorni dal mese di riferimento. L'impresa poteva presentare il DM10 su supporto cartaceo o magnetico o inviarlo telematicamente fino all'inizio del 2004 quando, per modifiche normative, l'invio telematico è divenuto l'unica modalità di presentazione accettata. Per la stima provvisoria vengono utilizzate le dichiarazioni "grezze" inviate all'INPS per via telematica o su supporto magnetico. Queste ultime, fino al 2004, sono state trattate come un campione "non casuale" mentre successivamente sono divenute un vero e proprio "universo provvisorio" che l'Istat acquisisce

[13] Sebbene il documento sia frutto del lavoro congiunto degli autori, i vari paragrafi possono essere come di seguito attribuiti. Il paragrafo 3 è stato curato da Fabio Rapiti, il paragrafo 5 da Eleonora Cimino, M. Carla Congia, Francesca Ceccato e Fabio Rapiti mentre il paragrafo 6 è stato curato da Eleonora Cimino e Fabio Rapiti. Il paragrafo 7 da Silvia Pacini e Donatella Tuzi mentre il paragrafo 8 è a cura di Ciro Baldi, Silvia Pacini e Donatella Tuzi. Il paragrafo 9 è stato curato da Ciro Baldi e Silvia Pacini, mentre il paragrafo 10 da Ciro Baldi, Silvia Pacini e Francesca Ceccato. Il paragrafo 11 è a cura di Silvia Pacini e Donatella Tuzi, mentre il paragrafo 12 di M. Carla Congia. I restanti paragrafi sono stati scritti interamente in collaborazione. Ringraziamo Rosa Sepe, Andrea Colace, Armando De Angelis che collaborano regolarmente all'elaborazione degli indici Oros.

[14] Ogni impresa con dipendenti per ottemperare agli obblighi contributivi deve aprire presso l'INPS una o più posizioni contributive alla quale è assegnato un identificativo (matricola). Quindi, la posizione contributiva rappresentata l'unità di rilevazione.

ogni trimestre sotto forma di tre archivi mensili con circa 35 giorni di ritardo rispetto al periodo di riferimento[15]. La stima definitiva viene effettuata utilizzando le dichiarazioni contributive dell'archivio dell'INPS che contiene l'"universo" dei DM10 che hanno subito un parziale trattamento di controllo da parte dell'Istituto di Previdenza e che vengono estratte dall'INPS e inviate all'Istat in tre archivi mensili scaricati a circa 12 mesi dal periodo di riferimento. A questi sei archivi va aggiunta l'anagrafica trimestrale INPS contenente le informazioni strutturali relative a ogni posizione contributiva. Ogni archivio mensile contiene oltre 1,3 milioni di modelli DM10 che vengono acquisiti nella forma integrale (circa 10 milioni di record al mese). Nel complesso ogni trimestre si acquisiscono e trattano oltre 60 milioni di record.

- La presenza di due processi di stima paralleli e distinti. La stima provvisoria viene rivista dopo 12-15 mesi quando sono disponibili dati amministrativi con una copertura completa. La stima definitiva, viene realizzata incorporando anche ulteriore e/o più aggiornata informazione resasi disponibile successivamente al rilascio della stima provvisoria[16]. L'archivio dei DM10 con cui si effettua la stima definitiva ha la stessa struttura disaggregata di quello utilizzato per la stima provvisoria ma contiene dati già controllati dall'INPS. Ciò nonostante è necessario comunque effettuare un processo di C&C anche se parzialmente differente da quello a cui sono sottoposti i dati utilizzati per la stima provvisoria. In definitiva ogni trimestre vengono effettuati due processi paralleli e distinti per produrre e diffondere la stima provvisoria relativa al trimestre corrente *t* e quella definitiva relativa al trimestre *t*-5 (insieme alla semidefinitiva di *t*-4[17]).
- La stima differenziata per quattro sottopopolazioni di imprese. Nella procedura di stima degli indicatori provvisori e definitivi, le unità presenti negli archivi INPS vengono distinte in quattro sottopopolazioni:
    1) le imprese di piccola e media dimensione (PMI);
    2) le imprese che vengono rilevate dall'Indagine GI;
    3) le imprese di grandi dimensioni non rilevate nell'indagine GI[18] (GI-INPS);
    4) le imprese interinali.

  La stima sulla popolazione sub 2, viene ottenuta utilizzando i dati provenienti dalla Rilevazione GI[19]. Tale stima non subisce per definizione revisioni mentre le stime provvisorie derivanti dai dati delle dichiarazioni DM10, relative alle altre tre sottopopolazioni, sono soggette a revisione.
- L'elevata tempestività della diffusione della stima provvisoria. I dati della rilevazione Oros vengono utilizzati non solo per produrre le stime per Ula diffuse trimestralmente a circa 70 giorni dal trimestre di riferimento attraverso regolari comunicati stampa, ma anche per soddisfare due regolamenti comunitari. Il primo è relativo alle statistiche congiunturali sulle imprese (STS) e prevede l'invio all'Eurostat di due indicatori, uno relativo all'occupazione entro 60 giorni dal trimestre di riferimento e uno relativo alle retribuzioni lorde entro 90 giorni. Il secondo regolamento è riferito all'indice del costo del lavoro trimestrale (Labour cost index - LCI) e prevede la diffusione di indici su retribuzioni lorde, oneri sociali e, quale sintesi dei due precedenti, del costo del lavoro per ora lavorata, che devono essere inviati all'Eurostat entro 70 giorni dal trimestre di riferimento.

---

[15] La metodologia di stima provvisoria si è evoluta nel corso del tempo a seguito dei cambiamenti nell'insieme informativo. Fino al 2004 si è basata su un modello predittivo stimato per sottogruppi della popolazione, che utilizzava informazioni correnti e ausiliarie. A partire dalla stima del secondo trimestre 2004, rilasciata nel mese di settembre 2004, vista la disponibilità di un universo provvisorio, non è stato più necessario ricorrere al modello predittivo.

[16] Alcune variabili che utilizzano altre fonti vengono calcolate/assegnate in base alla versione più aggiornata delle informazioni: per esempio, l'attività economica proviene da versioni successive dell'Archivio Statistico delle Imprese Attive (ASIA).

[17] In realtà ogni trimestre viene effettuata una revisione intermedia, la stima "semidefinitiva" del trimestre t-4, che si differenzia da quella "definitiva" del trimestre t-5 solo per l'utilizzo di minore informazione longitudinale nel processo di imputazione (si veda il cap. 8).

[18] La Rilevazione GI si basa su un panel chiuso di grandi imprese che viene aggiornato ogni 5 anni.

[19] Le ragioni della scelta di utilizzare i dati della Rilevazione GI al posto di quelli INPS vengono illustrate nel par. 9.

3. La strategia del processo di controllo e correzione della rilevazione Oros: peculiarità e caratteristiche

Il C&C in una rilevazione basata su dati amministrativi si pone in termini molto diversi rispetto alle rilevazioni tradizionali. In queste ultime, infatti, è possibile progettare e disegnare la rilevazione in modo tale da prevenire e ridurre al minimo i possibili errori non campionari in tutte le fasi della rilevazione (progettazione, stesura del questionario, wording, data entry, ecc.). Utilizzando dati amministrativi, invece, è impossibile intervenire ex ante sul modello in cui sono presenti i dati o sul processo di gestione e produzione degli stessi. Tale processo è fortemente influenzato da fattori amministrativi dettati da leggi, regolamenti, circolari ed è, quindi, totalmente indipendente dal controllo degli statistici che utilizzano successivamente i dati. Inoltre, in genere, i concetti, le definizioni e le classificazioni adottate nella raccolta dei dati amministrativi coincidono solo parzialmente con quelli delle indagini statistiche e gli statistici possono intervenire soltanto ex post.
La realizzazione del processo di C&C Oros non ha potuto basarsi su esperienze precedenti nell'uso di dati amministrativi a livello congiunturale rinunciando, inoltre, a priori ad una fase fondamentale presente in tutte le rilevazioni tradizionali sulle imprese vale a dire la revisione dei modelli effettuata da personale esperto e l'eventuale contatto diretto con le imprese che hanno fornito dati sospetti o errati.
Nel caso della rilevazione Oros c'è anche un ulteriore aspetto legato alla struttura fortemente disaggregata dei dati amministrativi elementari che vengono acquisiti dall'Istat nella forma originale (l'intera dichiarazione con le informazioni registrate su più record, in media circa 8 per DM10)[20]. Con due importanti conseguenze sul processo di produzione: da un lato un forte appesantimento perché si sono rese necessarie un'estesa e complessa fase di controllo preliminare dei dati amministrativi elementari e la ricostruzione delle variabili statistiche, dall'altro un aumento consistente della quantità di dati da trasferire dall'INPS all'Istat (non solo il numero dei record ma il numero delle variabili amministrative elementari) creando, soprattutto nei primi anni di implementazione della rilevazione, gravi problemi di disponibilità di spazio disco informatico alla struttura Istat che ha gestito l'intero processo. Quest'ultimo aspetto ha rappresentato un forte vincolo di cui si è dovuto tenere conto nello sviluppo delle procedure nella fase di impianto della rilevazione.
D'altro canto il vincolo di dover catturare il DM10 integrale nella sua forma originale ha consentito di ottenere due indubbi vantaggi: la disponibilità di molta più informazione utilizzabile anche da altri settori dell'Istat e un più stretto controllo sulla qualità dei dati.
Si è trattato, quindi, di impostare una strategia che consentisse di affrontare globalmente tutte le difficoltà derivanti da precondizioni ineliminabili. Di conseguenza i principi su cui si basa l'organizzazione del processo di controllo e correzione sono:

- pervasività dei controlli. In tutte le fasi della rilevazione sono presenti controlli su possibili errori non campionari; il C&C è stato esteso ai metadati ed è stato indispensabile avere come supporto una Banca Dati Normativa aggiornata trimestralmente;
- selettività e interattività. Sebbene sia stato necessario sviluppare processi estremamente selettivi per la localizzazione dei possibili errori, è sembrato ottimale correggere i valori anomali in modo interattivo utilizzando idonee "maschere di controllo", appositamente implementate, contenenti numerosi indicatori;
- documentazione, standardizzazione, trasparenza e condivisione. Ogni fase del C&C ha determinati indicatori di output per valutarne la correttezza e l'efficacia, indicatori che vengono registrati ogni trimestre e costituiscono serie storiche utili alla valutazione del C&C stesso. Poiché, inoltre, modificandosi la normativa è necessario cambiare frequentemente la struttura dei controlli, è stato indispensabile catalogare in modo standardizzato le diverse versioni trimestrali delle procedure informatiche utilizzate (*versioning*) in modo da mantenere una memoria storica facilmente ed immediatamente accessibile. In definitiva è l'intero processo

20 Nella fase iniziale di progettazione e sperimentazione della rilevazione, è risultato impossibile prevedere che l'INPS riaggregasse direttamente i dati nelle modalità necessarie agli obiettivi della rilevazione nei tempi ridotti previsti per il loro rilascio.

trimestrale di C&C ad essere completamente standardizzato e documentato in maniera sistematica e continua. Tutto ciò permette, tra l'altro, una completa condivisione dell'intero processo e di conseguenza un'elevata interscambiabilità dei ruoli tra il personale preposto alla rilevazione[21].

Le scelte strategiche iniziali di C&C sono state sperimentate e implementate progressivamente per un certo periodo prima dell'entrata a regime della rilevazione. Nel corso del tempo sono state razionalizzate ed adattate sulla base dell'esperienza e delle modifiche delle caratteristiche tecniche e di contenuto informativo delle basi di dati amministrative.

Il C&C Oros viene effettuato utilizzando delle procedure totalmente originali sviluppate appositamente dal personale responsabile della rilevazione prevalentemente in linguaggio Sas[22]. Nel corso del tempo tali procedure sono risultate affidabili in termini di efficacia (qualità dell'intero processo), in termini di efficienza (durata limitata e utilizzo ridotto di risorse umane e finanziarie).

4. Le diverse macro fasi del C&C nel processo di produzione Oros

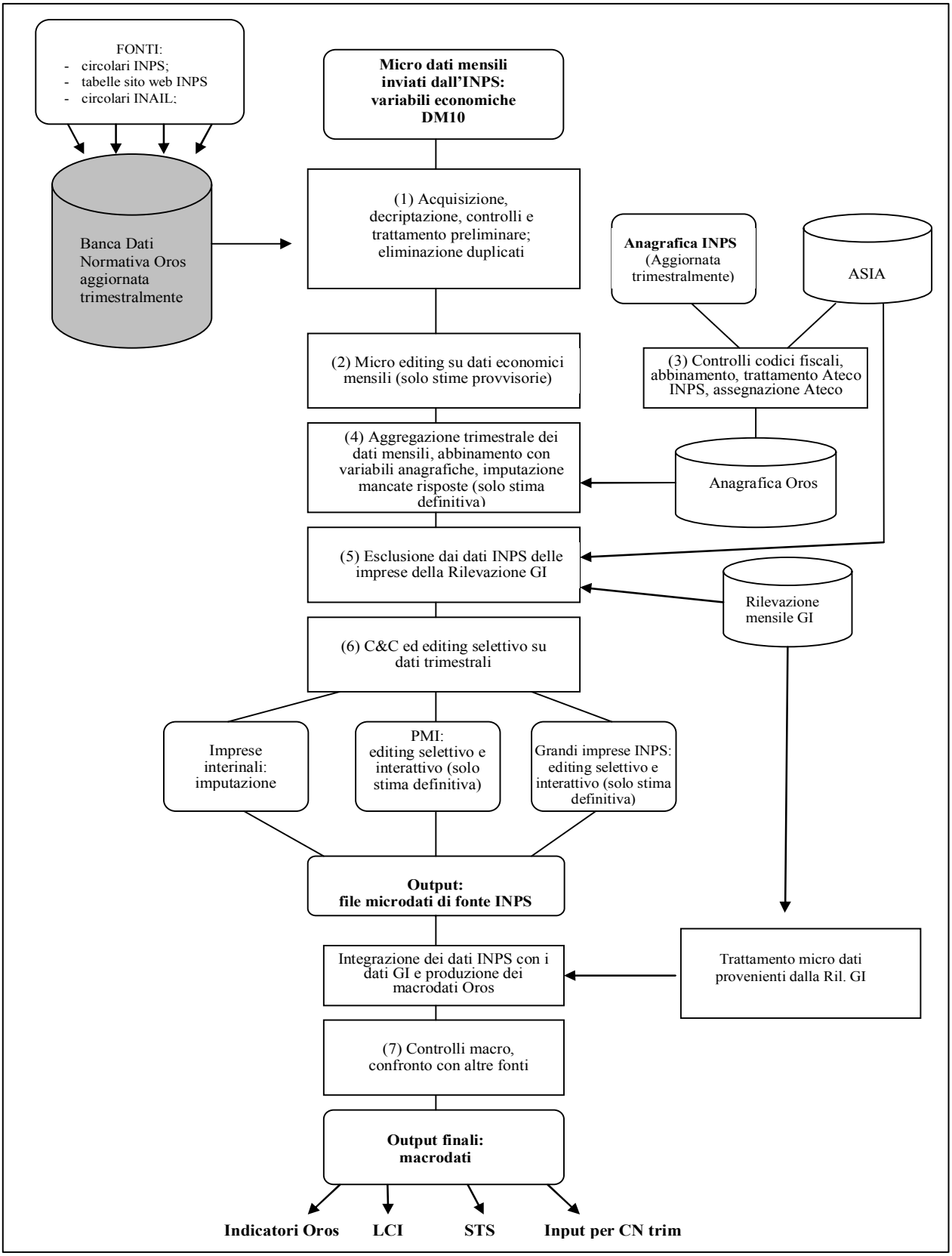
Il processo di C&C Oros si articola essenzialmente in 7 fasi principali con alcune diversità fra la stima provvisoria e quella definitiva (Figura 1).

1. Le procedure di controllo preliminare e di trasformazione in variabili statistiche dei dati amministrativi utilizzando la Banca Dati Normativa. In ogni singola dichiarazione DM10 le informazioni sono classificate in base a numerosi codici amministrativi. Per riuscire ad utilizzare a fini statistici le informazioni associate a tali codici è necessario interpretarne correttamente il significato amministrativo, verificare la correttezza formale e, in relazione ad altre informazioni presenti sulla stessa dichiarazione, riaggregare i dati in variabili statistiche. Tutto ciò può essere effettuato soltanto con l'ausilio di una Banca Dati Normativa appositamente sviluppata e aggiornata trimestralmente.
2. Il micro editing interattivo dei dati economici mensili dell'universo provvisorio. Sulle variabili economiche mensili viene effettuato un processo di controllo e correzione interattivo dei valori anomali per tutte le unità appartenenti alle due sottopopolazioni delle GI-INPS e delle imprese interinali. Viene effettuato, inoltre, un microediting selettivo per le PMI attraverso una localizzazione deterministica dei possibili errori ed un controllo interattivo dei valori anomali. In entrambi i casi il microediting viene effettuato utilizzando una maschera di controllo contenente indicatori trasversali e longitudinali.
3. Controlli sulle variabili anagrafiche. Nella fase di costruzione dell'anagrafica Oros e in quella di abbinamento fra informazioni economiche ed anagrafiche vengono effettuati diversi controlli su alcune variabili (codice fiscale, Codice Statistico Contributivo, Ateco, etc.) e viene attribuito il codice di attività economica.
4. Il trattamento delle mancate risposte nella stima definitiva. Poiché la stima definitiva deve garantire la copertura certa della intera popolazione, e ciò non era completamente garantito dagli archivi "definitivi" INPS in cui mancavano unità ritardatarie, è stata prevista una procedura di imputazione delle mancate risposte. A causa dei problemi di copertura dell'anagrafe INPS, nella rilevazione Oros non esiste una lista teorica di unità attive che va pertanto predetta per poter in seguito individuare le mancate risposte. L'imputazione delle variabili di interesse sulle singole unità si basa su modelli deterministici che sfruttano la notevole ricchezza di informazioni longitudinali disponibili.

[21] Le peculiarità della rilevazione hanno implicato che il personale sviluppasse professionalità e competenze trasversali in modo da conoscere e seguire la normativa contributiva, saper programmare in SAS in modo rapido ed efficiente, conoscere ed interpretare le procedure statistiche, gli indicatori ed i risultati economici aggregati delle diverse variabili obiettivo.

[22] Utilizzando principalmente i moduli Base, Macro, SQL, IML, ETS, INSIGHT.

Figura 1: *Le diverse macro fasi del processo di produzione Oros*



5. Integrazione e trattamento dei dati GI. L'integrazione tra le due fonti (INPS e GI) implica la costruzione di due liste complementari di imprese al fine di evitare da un lato duplicazioni delle stesse imprese, dall'altro la mancata assegnazione di una impresa ad una delle due liste.
6. Il processo di C&C delle variabili economiche trimestrali prodotte a partire dall'"universo". Per la sottopopolazione delle PMI viene effettuata una procedura di editing selettivo individuando prima le unità influenti e successivamente effettuando un controllo interattivo basato su una maschera di controllo contenente indicatori trasversali e longitudinali. L'individuazione delle unità influenti si basa su un criterio di sensibilità della stima alla singola unità, misurando l'effetto che l'esclusione dell'unità considerata ha sulla stima del parametro di interesse. Le sottopopolazioni delle GI-INPS e delle imprese Interinali subiscono un processo ad hoc in considerazione delle loro caratteristiche peculiari.
7. Controlli macro sui domini e sottodomini di stima e di diffusione. Prima di essere diffusi i dati vengono controllati analizzando il loro andamento in forma aggregata in confronto ad altre fonti disponibili e alla serie storica. Nel caso emergano anomalie si effettuano ulteriori controlli con procedure Sas sviluppate ad hoc fino ad individuare eventuali outliers o problemi legati a particolari modifiche normative non correttamente incorporate nell'aggiornamento dei controlli.

5. Le procedure di controllo preliminare e di trasformazione in variabili statistiche dei dati amministrativi

5.1 La struttura delle informazioni all'interno del modello DM10

L'acquisizione dei dati del modello DM10 nella loro forma integrale e grezza, cioè senza interventi o aggregazioni di sorta da parte dell'INPS, ha obbligato l'Istat a sottoporre i microdati amministrativi ad un trattamento preliminare del tutto peculiare, propedeutico alle fasi successive, che permetta di ricostruire correttamente il microdato statistico per posizione contributiva.
Prima di illustrare tale trattamento è necessario descrivere brevemente la struttura del modello DM10 e le caratteristiche dei dati in esso riportati. La dichiarazione mensile risulta suddivisa in quattro sezioni denominate "quadri". In particolare, nel quadro A sono riportate alcune caratteristiche anagrafiche relative alla posizione contributiva (la matricola assegnata dall'INPS, la forma giuridica, ecc). Nel quadro B-C, sono riportate le informazioni relative al numero dei dipendenti, alle giornate retribuite, al monte retributivo ed ai contributi a debito complessivi, a carico del datore di lavoro e del lavoratore. Nel quadro D vengono riportati gli importi a credito derivanti da riduzioni contributive o da indennità anticipate dal datore di lavoro (malattia, maternità, assegni familiari, ecc.). Il saldo fra il totale dei contributi a debito del quadro B-C e degli importi a credito del quadro D rappresenta quanto il datore di lavoro deve versare nel mese di riferimento all'INPS.
Tutte le informazioni (dipendenti, giornate retribuite, monti retributivi e contributivi) riportate nei quadri B, C e D del modello sono disaggregate in base ad una particolare "variabile amministrativa" che identifica la tipologia occupazionale e/o contributiva dei lavoratori. Tale "variabile amministrativa" di classificazione è composta da 4 caratteri (numerici o alfanumerici) e presenta un numero di modalità, i cosiddetti codici, molto elevato e crescente nel tempo. L'evoluzione della normativa sugli adempimenti contributivi comporta delle continue modifiche di tali codici: ogni trimestre ne vengono inseriti nuovi, altri vengono annullati, alcuni assumono un nuovo significato.
I codici possono assumere significati molto diversi. Vista la complessità del contenuto informativo del modello, per il suo sfruttamento a fini statistici è necessario classificare tutti i codici per individuare quelli da selezionare per la corretta aggregazione delle variabili amministrative.

5.2 La banca dati normativa Oros

Nel caso della rilevazione Oros, senza dei metadati completi e aggiornati continuamente per interpretare il contenuto informativo della dichiarazione, i micro dati amministrativi non potrebbero essere utilizzati in modo corretto. Pertanto, si è reso necessario rintracciare, raccogliere, archiviare in una forma standardizzata e facilmente accessibile non soltanto i metadati relativi alla dichiarazione contributiva ma anche quelli indispensabili per la stima di alcune componenti del costo del lavoro non rilevate nel modello DM10. A tale scopo è stata progettata e realizzata una Banca Dati Normativa (BDN) per organizzare in modo sistematico e aggiornare trimestralmente i riferimenti normativi, i metodi e le procedure utilizzati (Cimino et al., 2003).

In sintesi, la procedura della BDN si sviluppa ogni trimestre nelle seguenti fasi:

- Costruzione della lista dei codici validi, cioè delle modalità della "variabile amministrativa" di classificazione ammissibili nel trimestre di riferimento.
- Classificazione dei codici in tre tipologie:
    1. i "codici occupazione" che indicano gruppi omogenei di lavoratori a fini contributivi (operai, impiegati, dirigenti, apprendisti, lavoratori assunti con CFL, ecc.) e individuano il numero dei dipendenti, le rispettive giornate retribuite, retribuzioni imponibili e contribuzioni di base;
    2. i "codici contribuzione" rappresentano, invece, delle particolarità contributive aggiuntive rispetto ai contributi di base già registrati con i codici occupazione;
    3. gli "altri codici" che hanno significati diversi. Questi devono essere esclusi dalla ricostruzione delle variabili statistiche obiettivo della rilevazione Oros.

  I "codici occupazione" sono presenti solo nel quadro B-C e aggregando opportunamente le variabili amministrative ad essi associate si possono calcolare correttamente il numero dei dipendenti e il monte retributivo. Per ricostruire i contributivi complessivi versati dal datore di lavoro devono essere correttamente aggregati i contributi di base associati ai "codici occupazione", quelli aggiuntivi indicati con i "codici contribuzione" del quadro B-C e all'importo ottenuto va detratta la somma delle riduzioni contributive associate ai "codici contribuzione" del quadro D[23]. Tale classificazione permette di ricostruire per ogni dichiarazione DM10 le variabili posizioni lavorative, giornate retribuite e monti retributivi per otto categorie – quattro qualifiche (operai, impiegati, apprendisti, dirigenti) per tempo di lavoro (tempo pieno e tempo parziale). A causa dell'assenza di informazioni specifiche, non è possibile, invece, ricostruire con questo livello di dettaglio la variabile oneri sociali.
- Identificazione e aggiornamento delle aliquote contributive a carico del lavoratore. L'importo complessivo degli oneri sociali riportato sul modello DM10, comprende sia la parte a carico del datore di lavoro sia quella a carico del dipendente. Per calcolare correttamente il costo del lavoro è necessario scorporare dagli oneri sociali la quota a carico del lavoratore in quanto già compresa nella retribuzione imponibile. Per individuare la specifica aliquota da applicare a carico del dipendente, vengono utilizzati i metadati derivanti dallo studio di una numerosa serie di tabelle sulle aliquote contributive presenti sul sito web dell'INPS.
- Aggiornamento delle aliquote INAIL. Il costo dell'assicurazione per gli infortuni sul lavoro INAIL rappresenta un'importante componente degli oneri sociali, ma l'informazione sull'importo del premio versato dal datore di lavoro per ogni dipendente non è presente nella dichiarazione DM10. Pertanto è necessario ricorrere ad altre fonti, in particolare le aliquote medie utilizzate per il calcolo del costo del lavoro contrattuale[24] e le aliquote medie per gruppo di attività economica pubblicate sul sito internet dell'INAIL.

[23] Escludendo gli importi relativi ai crediti del datore di lavoro derivanti da prestazioni anticipate per conto dell'INPS presenti nel quadro D che non costituiscono componenti del costo del lavoro.

[24] Stimato nel quadro dell'Indagine Costo del lavoro e retribuzioni nette su base contrattuale, effettuata soltanto relativamente all'anno 1995 dalla struttura Istat che si occupa della produzione degli Indici mensili sulle retribuzioni contrattuali.

- Monitoraggio della normativa sul costo del lavoro. Per la stima di eventuali ulteriori componenti del costo del lavoro non rilevate dal DM10 viene regolarmente monitorata l'evoluzione della normativa di riferimento e analizzate fonti alternative (Es: riduzioni del costo del lavoro attraverso il credito di imposta o la riduzione dell'Irap).

La gestione trimestrale della BDN è particolarmente onerosa, in quanto è soltanto parzialmente automatizzabile e richiede allo stesso tempo competenze giuridiche, informatiche e statistiche. L'aggiornamento della BDN è un'attività assai delicata che deve essere esaustiva e precisa nell'individuazione delle componenti da includere nella ricostruzione delle variabili statistiche obiettivo, in modo da non provocare distorsioni nelle stime degli indicatori pubblicati (cfr. par. 12).

5.3 Le diverse fasi del trattamento preliminare

I metadati prodotti nella BDN vengono utilizzati per la corretta trasformazione dei micro dati amministrativi nei micro dati statistici, che solo successivamente vengono sottoposti alle tradizionali procedure di controllo e correzione. La fase di trattamento preliminare, pertanto, si caratterizza per originalità e complessità, richiedendo un notevole sforzo sia di progettazione sia di implementazione e aggiornamento trimestrale. Le principali fasi (Figura 2) in cui si articola il trattamento preliminare sono:

- Controlli quantitativi sul numero e i legami tra i record ai fini della valutazione del grado di popolamento dell'archivio e della necessità di richiedere all'INPS eventuali scarichi supplementari:
  - numerosità totale dei record (circa 10 milioni);
  - numerosità totale dei modelli DM10 (circa 1,3 milioni).
- Controllo sulla presenza di errori formali nei codici:
  - controllo della compatibilità dei codici con la lista aggiornata di tutti quelli ammissibili nel trimestre (identificazione di 19 tipologie di errore);
  - normalizzazione della stringa dei codici attraverso la correzione di 3-4 tipologie di errore in base ad un set di regole impostate con l'ausilio dei metadati della BDN (es: la trasformazione dello 0 al primo digit in O, ecc).
- Controllo di compatibilità delle variabili quantitative associate ai singoli codici:
  - controllo sul numero delle variabili quantitative associate a ciascun codice (numero dipendenti, numero giornate retribuite, monti retributivi, monti contributivi) e segnalazione dei casi di possibili mancate risposte parziali;
  - identificazione dei codici ripetuti all'interno del singolo modello e controllo dei valori delle variabili quantitative ad esso riferite, distinguendo i casi in cui le variabili associate ai codici ripetuti sono tutte uguali (ripetuti "identici") oppure le variabili assumono valori diversi (ripetuti "diversi");
  - correzione dei record con i codici ripetuti attraverso la conservazione del primo dei record ripetuti "identici" oppure aggregando i record ripetuti "diversi" in un unico record sommando i valori delle variabili quantitative.
- Controllo qualitativo/quantitativo sulle relazioni tra le variabili associate ai singoli codici all'interno dello stesso DM10:
  - controllo dei valori assunti dalle variabili quantitative in corrispondenza di codici che identificano tipologie contrattuali e/o contributive di particolare complessità espositiva, attraverso il check delle relazioni formali tra questi codici e le informazioni ad essi associate (CFL, apprendisti trasformati, ecc.).
- Ricostruzione delle variabili statistiche all'interno del singolo DM10:
  - aggregazione dei record relativi ai codici selezionati, sulla base dei metadati della BDN, per il calcolo dei dipendenti, delle giornate retribuite, delle retribuzioni lorde e dei contributi per qualifica. Ciò consente di sintetizzare in un solo record i dati relativi a ciascun modello DM10;

- scorporo dei contributi a carico del lavoratore dal totale dei contributi complessivi utilizzando le aliquote a carico del lavoratore stimate nella BDN;
- stima delle componenti degli oneri sociali non registrate nel modello DM10 (premi Inail, accantonamento TFR) sulla base dei metadati della BDN.

- Trattamento dei modelli DM10 ripetuti:
  - identificazione di più modelli DM10 riferiti alla stessa posizione contributiva;
  - trattamento dei duplicati sulla base dei valori assunti da un set di variabili statistiche rilevanti:
    - duplicazioni "identiche": si conserva solo il primo modello in quanto gli altri sono considerati dei modelli inviati per errore;
    - duplicazioni "diverse": si aggregano i modelli multipli in un unico record sommando i valori di tutte le variabili perché tali modelli sono considerati integrativi o complementari.

Le procedure di controllo e di correzione preliminari illustrate vengono applicate sia ai dati dell'universo sia a quelli dell'universo provvisorio dei DM10. I primi presentano una qualità più elevata in quanto sottoposti ai controlli formali e di merito da parte dell'INPS. Al contrario, i modelli DM10 acquisiti a 35 giorni dalla fine del trimestre di riferimento utilizzati per la produzione delle stime provvisorie, vengono messi a disposizione dell'Istat senza subire alcun trattamento da parte dell'INPS. E' quindi sui dati dell'universo provvisorio che le procedure di trattamento preliminare individuano un maggior numero di anomalie che, tuttavia, risulta essere di bassa entità in rapporto alla numerosità dei record e delle variabili trattatati.

6. Il micro editing sui dati mensili nella stima preliminare

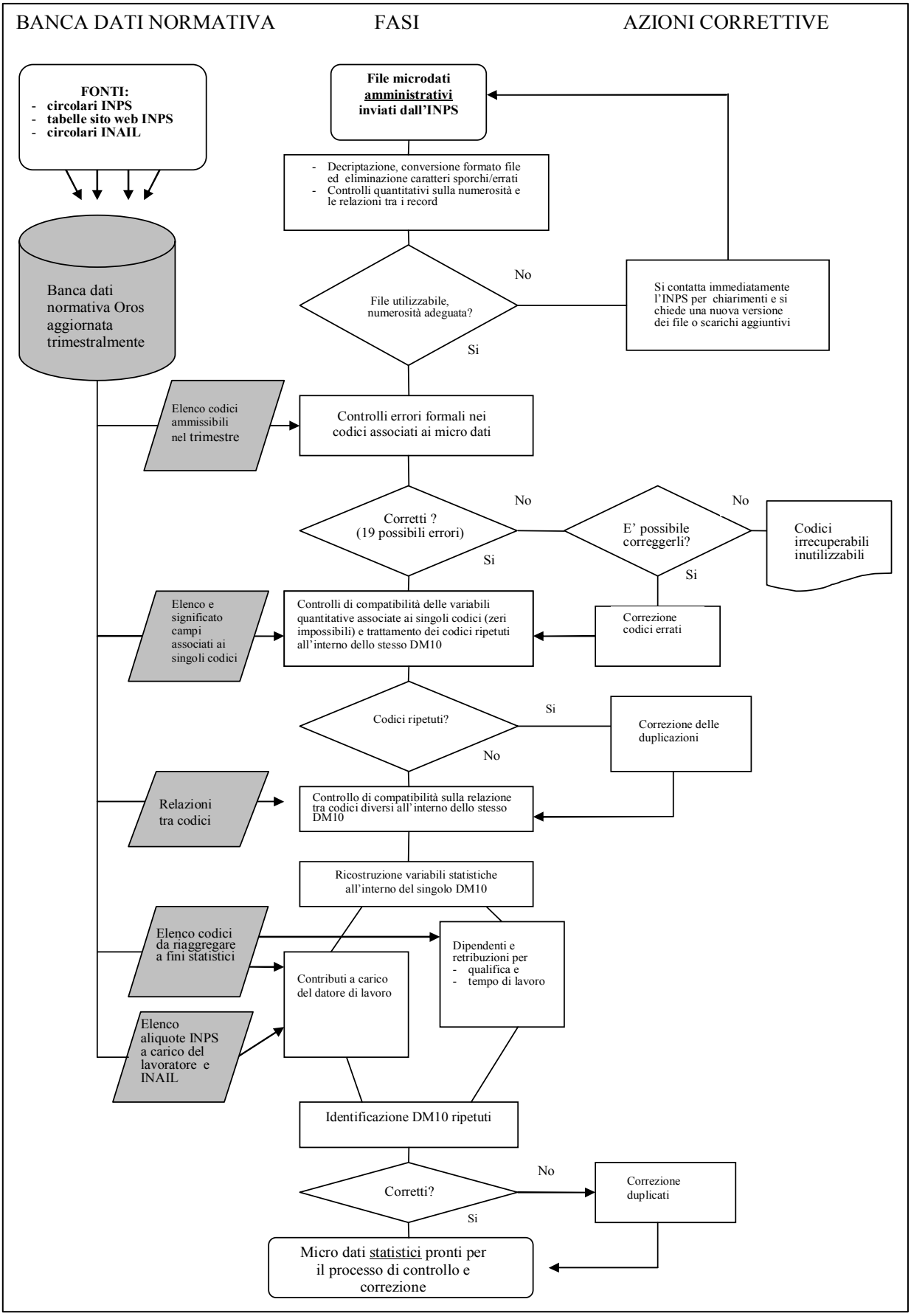
Una volta ricostruite le variabili utilizzabili a fini statistici si deve procedere al controllo e correzione dei micro dati mensili con l'obiettivo di individuare ed eventualmente correggere gli errori di misura, in particolare i valori estremi. Il livello di aggregazione delle variabili su cui viene effettuato il controllo è mirato esclusivamente alle variabili obiettivo. Come è stato accennato nel paragrafo precedente, poiché gli indicatori congiunturali Oros si riferiscono all'insieme dei dipendenti con l'esclusione dei dirigenti si è ritenuto opportuno sottoporre alla procedura di micro editing tutte le variabili di interesse distinte per soli due aggregati: il totale dei dipendenti e il "di cui" dei dirigenti in modo da ottenere per semplice differenza l'insieme del totale dei dipendenti esclusi i dirigenti.

I controlli si basano sia su rapporti caratteristici trasversali a livello di singola osservazione sia su informazioni longitudinali riferite alla stessa unità nel mese precedente. Alcune sperimentazioni hanno consentito di escludere la possibilità di fare riferimento al mese corrispondente dell'anno precedente. I benefici di cogliere la stagionalità sono inferiori ai costi in termini di minore compresenza delle unità nei due mesi[25] e di maggiore complessità della procedura. Quindi, il controllo viene effettuato sul mese in esame (m) con l'ausilio delle informazioni del mese precedente (m-1) già eventualmente corrette. Questa procedura viene eseguita a catena sui tre mesi del trimestre oggetto di controllo.

I. Per le sottopopolazioni delle GI-INPS e delle imprese interinali il controllo è esaustivo. Per le PMI, vista la numerosità (oltre 1,2 milioni di unità), il controllo non può che essere di tipo selettivo. Le soglie di selezione sono state scelte in modo altamente prudenziale cercando di minimizzare la probabilità di controllare valori corretti.

[25] Il forte e continuo turnover demografico in particolare di piccole e piccolissime imprese che iniziano attività con dipendenti o che la cessano comporta che fra il mese corrente m e quello relativo all'anno precedente m-12 vi sia una così elevata quota di unità presenti soltanto in uno due mesi che viene meno l'utilità del controllo longitudinale.

Figura 2: *Le procedure del trattamento preliminare*



Per effettuare il micro editing mensile è stata progettata e realizzata una "maschera di controllo" sviluppata ad hoc con la procedura FSEDIT di Sas, che riporta un insieme di variabili e indicatori che letti in modo combinato consentono all'operatore di discriminare, in modo rapido ed efficace se si è in presenza di osservazioni corrette o di errori. Le variabili riportate dalla "maschera di controllo" sono numerose e possono essere raggruppate in quattro sottoinsiemi:

II. informazioni identificative dell'unità: numero di matricola, attività economica nel mese *m* e nel mese *m*-1;
III. variabili di interesse:
    a. il numero totale dei dipendenti, distinguendo ulteriormente quelli a tempo pieno e quelli a tempo parziale[26], il totale delle giornate retribuite, il monte retributivo e il monte oneri;
    b. alcuni rapporti caratteristici calcolati combinando le suddette variabili: retribuzione media pro capite mensile, retribuzione media giornaliera, oneri medi pro capite mensili, rapporto percentuale fra oneri e retribuzione;
IV. variabili ausiliarie, rappresentate dalle variabili di interesse e dai rapporti caratteristici citati sopra riferite al mese precedente (*m*-1) e da altri indicatori come: la variazione del numero dei dipendenti, la variazione del rapporto percentuale tra gli oneri e le retribuzioni, le aliquote contributive percentuali a carico del lavoratore nei due mesi analizzati;

informazioni sul numero delle unità e il numero di occupati totali nella divisione in cui è classificata l'osservazione, derivati dall'universo dei dati INPS dell'anno precedente.

Per il controllo di un generico mese *m* la procedura può essere suddivisa in 5 passi (Figura 3).

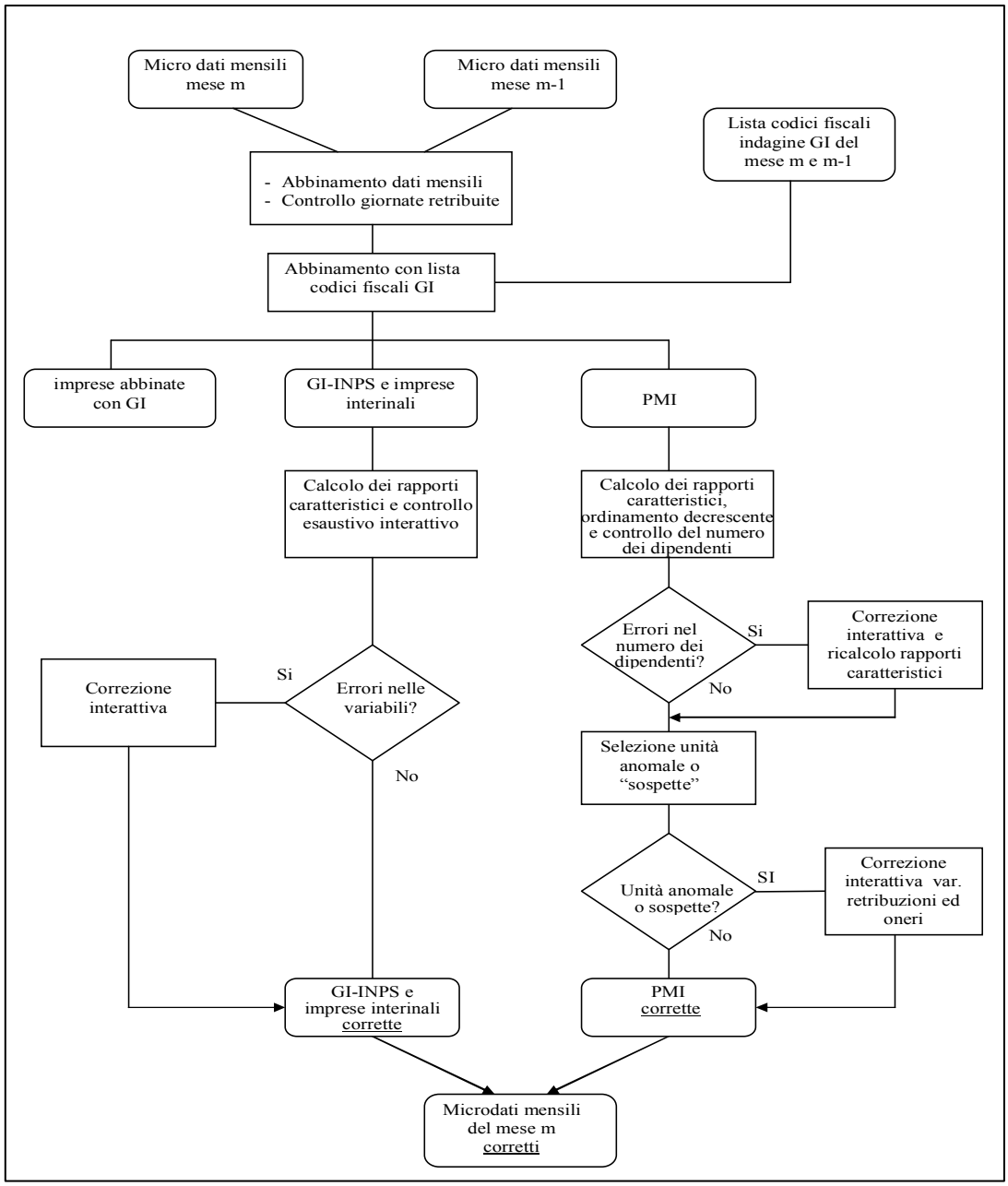
1. Abbinamento delle informazioni economiche relative al mese in esame *m* con le stesse informazioni relative al mese precedente *m*-1 a livello di ogni singola unità.
2. Abbinamento con la lista GI che consente di individuare le unità abbinate che vanno escluse (poiché esse verranno stimate in modo censuario utilizzando la rilevazione sulle GI) e le tre sottopopolazioni da controllare con diversa procedura: le unità GI-INPS, le imprese Interinali e le PMI.
3. Controllo e correzione della variabile "giornate retribuite" totali partendo dall'analisi delle distribuzioni mensili delle giornate retribuite pro capite, aggiustate per i lavoratori part-time e gli apprendisti[27]. Il controllo è necessario per l'individuazione di valori impossibili nelle giornate retribuite, variabile che in seguito contribuisce all'analisi interattiva degli errori.
4. Controllo esaustivo ed eventuale correzione interattiva delle variabili delle unità appartenenti alle sottopopolazioni GI-INPS e delle imprese Interinali. Le correzioni, effettuate attraverso la maschera di controllo interattiva, vengono segnalate con un flag e quelle sugli occupati vengono riportate anche sulle variabili utili alla costruzione delle Ula (part-time, full-time, ecc.).
5. Nella sottopopolazione PMI si procede con le seguenti fasi:
    - controllo longitudinale della variabile totale occupazione ed eventuale correzione interattiva. A tale scopo, si utilizza la distribuzione delle variazioni mensili del numero di dipendenti. Ordinando in modo decrescente le osservazioni secondo la stessa variabile si limita il controllo interattivo alle imprese con variazioni (positive o negative) superiori a 80 dipendenti. Sulla base dell'esperienza si è osservato che eventuali errori compaiono ai due estremi della distribuzione: basta scorrere poche decine di unità che hanno subito forti variazioni, per individuare eventuali valori estremi errati. Se necessario si effettuano le correzioni attraverso la maschera interattiva e, anche in questo caso, quelle sugli occupati vengono riportate anche sulle variabili utili alla costruzione delle Ula;
    - individuazione di un insieme "sospetto" di unità che rispetto a diversi rapporti caratteristici (retribuzioni medie pro capite e oneri medi pro capite) superano dei valori soglia prestabiliti. Inizialmente modulando opportunamente i valori soglia, erano stati individuati due insiemi: il primo che conteneva le poche osservazioni caratterizzate da situazioni molto anomale (probabili errori casuali) o sistematicamente distaccati dal corpo delle distribuzioni (errori

[26] Informazioni indispensabili per il calcolo delle Ula.

[27] Il valore originale era espresso in settimane nel caso del tempo parziale e in ore quello degli apprendisti.

nell'unità di misura) su cui era possibile effettuare correzioni automatiche; il secondo contenente altri valori anomali da valutare interattivamente. Dopo un certo periodo di esperienza si è preferito far rientrare in un unico insieme tutte le unità anomale e non procedere più a correzioni automatiche ma controllare interattivamente tutte le unità;

Figura 3: *Le fasi del micro editing interattivo sui dati mensili*



- ordinamento decrescente delle unità dell'insieme "sospetto" in base ad un indicatore di sintesi della dimensione della unità (media fra il totale delle retribuzioni e il totale degli oneri del mese *m*) in modo da controllare per prime le osservazioni che dovrebbero avere un impatto maggiore sulle stime. Successivamente si procede al controllo interattivo ed all'eventuale correzione delle variabili retributive e contributive scorrendo le osservazioni fino a quando si valuta che la dimensione decrescente delle unità rende inutile continuare l'esame. Nel caso si identifichino degli errori, le singole variabili vengono corrette sostituendo valori più plausibili basati sulle

variabili ausiliarie del mese precedente o, in caso di errori nell'unità di misura, dividendo o moltiplicando per multipli di 10.

6. Una volta effettuate tutte le eventuali correzioni si procede alla riunificazione dei diversi file contenenti le unità appartenenti alle sottopopolazione delle GI-INPS e delle imprese Interinali, con quelle della sottopopolazione PMI ottenendo il file dei micro dati corretti.

Va segnalato che la particolarità dei dati, legati alla provenienza amministrativa, ha un notevole impatto anche sulla distribuzione delle variabili retribuzioni e oneri sociali e sulla identificazione di valori soglia che riescano a discriminare fra valori accettabili ed errori. La distribuzione della retribuzione pro capite media mensile delle PMI come prevedibile presenta una certa asimmetria con una lunga coda a destra di retribuzioni pro capite molto elevate. Ma presenta anche una coda a sinistra con un considerevole numero di osservazioni con retribuzioni estremamente ridotte che declinano lentamente fino allo zero. In altri contesti queste osservazioni sarebbero considerate affette da errori ma in questo caso si tratta di fenomeni reali dovuti all'erogazione di retribuzioni ridotte o integrazioni salariali alle indennità erogate dall'INPS per maternità, malattia, CIG, in imprese con un numero di dipendenti estremamente ridotto. Il caso degli oneri sociali è ancora più particolare perché questa variabile a livello mensile può assumere addirittura valori negativi. Le imprese, infatti, a causa di modifiche normative relative, ad esempio, a determinati gruppi di lavoratori, frequentemente hanno diritto a sgravi contributivi o al recupero di contributi pagati in eccesso nei mesi precedenti che possono compensare riducendo i versamenti nel mese corrente. Questo fenomeno interessa maggiormente le imprese con pochi dipendenti perché, in genere, in quelle grandi i recuperi contributivi vengono più che compensati da altri versamenti. Nell'editing delle PMI, pertanto, si sono dovute fissare soglie con valori negativi accettabili.

Il numero delle correzioni effettuate complessivamente sulle imprese grandi, sulle interinali e sulle PMI risulta estremamente ridotto, relativamente al totale dei record sottoposti al micro editing. Tuttavia, l'entità di queste correzioni di frequente può essere molto rilevante. A titolo di esempio un caso estremo di errore nelle posizioni lavorative di una singola impresa che non venisse eliminato correttamente dalla procedura di micro editing, nel terzo trimestre del 2007 comporterebbe un impatto nella variazione tendenziale della sezione G di 2,2 punti percentuali (che corretta è pari allo 0,8 per cento, in luogo del 3,0 per cento affetto da errore).

7. Il controllo delle variabili anagrafiche

Ai dati economici presenti nel DM10 devono essere abbinate le informazioni anagrafiche, in particolare l'identificativo d'impresa (codice fiscale) e la classificazione economica dell'attività presenti nell'anagrafica trimestrale Oros (Baldi et al., 2001), costruita dall'integrazione dell'archivio anagrafico INPS con l'archivio statistico annuale delle imprese attive ASIA (Figura 4).

La fonte principale, il registro anagrafico delle posizioni contributive fornito trimestralmente dall'INPS, è utilizzabile a fini statistici solo in seguito ad alcune operazioni preliminari di controllo, trattamento ed integrazione di dati. Tra le informazioni da monitorare vi è il codice fiscale che, avendo poca rilevanza per i fini amministrativi, potrebbe non essere riportato correttamente nel registro INPS. Allo scopo di identificare e corregge i codici fiscali mancanti o errati è stata sviluppata una procedura che sfrutta in modo sequenziale alcune informazioni ausiliarie (Partita Iva, matricola di riferimento[28], altre informazioni longitudinali). Nel corso del tempo la qualità di questa variabile è andata progressivamente migliorando e la quota di codici fiscali formalmente[29] corretti in origine è ormai quasi totalitaria.

Disporre di un codice fiscale di buona qualità è anche il presupposto per la corretta assegnazione del codice di attività economica. Infatti, sebbene il registro INPS contenga almeno due fonti informative

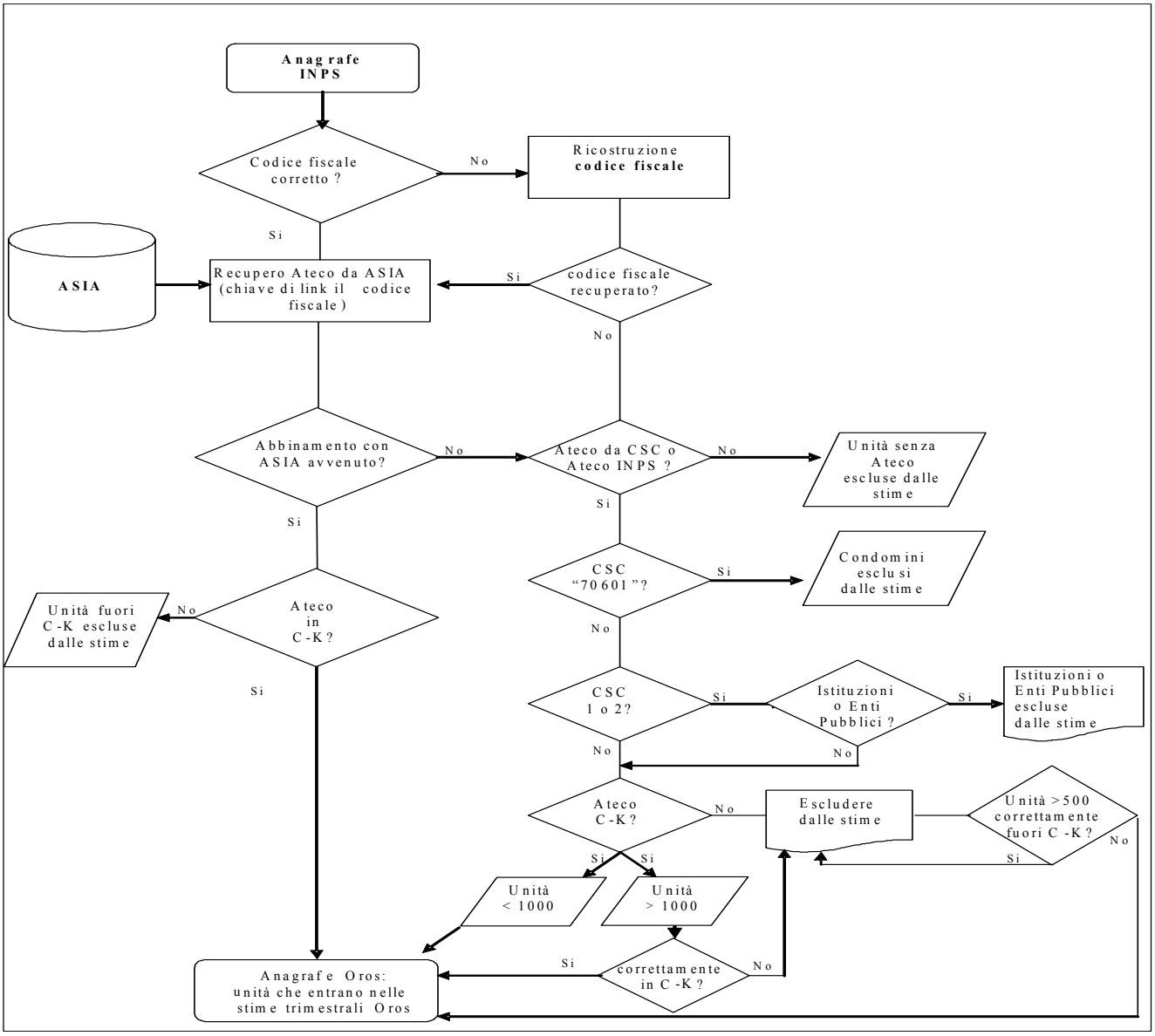
[28] La "matricola di riferimento" è un identificativo che lega, a fini amministrativi, diverse posizioni contributive della stessa impresa.

[29] E' stata riscontrata la presenza di un numero non trascurabile di unità con codice fiscale formalmente corretto, ma ripetutamente assegnato d'ufficio dagli operatori INPS ad unità con identificativo d'impresa non comunicato. Per evitare che informazioni assegnate mediante codice fiscale risultino scorrettamente utilizzate, è stata creata una lista di questi codici fiscali "anomali", periodicamente aggiornata, per controllare le unità interessate.

che consentono di individuare il settore di attività, si è stabilito di assegnare il codice Ateco ufficiale d'impresa attribuito dall'Istat, nel registro delle imprese ASIA che, tuttavia, ha un ritardo di aggiornamento di 15-22 mesi rispetto all'anagrafe INPS[30].
Il *link* dell'archivio INPS con ASIA mediante codice fiscale individua un insieme di unità che non si abbinano. In prevalenza si tratta di unità di recente costituzione che per definizione non sono presenti nel registro delle imprese che è relativo ad un periodo precedente. Vi sono anche casi in cui il mancato abbinamento dipende da una diversa tempistica di aggiornamento del codice fiscale tra i due archivi. Alle posizioni a cui non è possibile assegnare un Ateco ASIA si attribuisce l'Ateco di fonte INPS. L'assegnazione di questa variabile da parte dell'INPS avviene d'ufficio, mediante una ricerca di corrispondenza tra descrizione dell'attività svolta da parte dell'impresa e i codici Ateco riportati in un manuale operativo predisposto appositamente. La qualità di questa variabile, pertanto, può essere fortemente influenzata dal giudizio discrezionale dell'operatore INPS[31] anche se, in generale, viene ritenuta soddisfacente[32]. Nei pochi casi residui si assegna un Ateco dedotto dal codice statistico contributivo (CSC), una variabile di contenuto amministrativo che, tuttavia, non ha un'immediata ed univoca corrispondenza con i codici convenzionali della classificazione dell'attività economica e va quindi trascodificata.
In definitiva, la procedura assegna ad oltre il 90% delle unità un Ateco di fonte ASIA, alle restanti (circa centomila unità), viene assegnato l'Ateco INPS e solo ad un centinaio di unità circa viene attribuito un codice di attività economica attraverso la trascodifica del CSC.
Le unità a cui è stato assegnato un Ateco di fonte INPS vengono sottoposte anche a controlli per verificare se rientrano nel campo di osservazione della rilevazione. Considerando ASIA come *benchmark* ufficiale di riferimento, vengono in particolare controllate le unità che, pur risultando classificate secondo INPS nel dominio C-K, non si trovano nel Registro delle imprese. Escludendo le nate o le riattivate nel periodo successivo all'istante di aggiornamento di Asia, alcune di queste unità sono errori di sovracopertura della popolazione obiettivo. Tra queste, circa 400 posizioni contributive configurate come enti e amministrazioni pubbliche che, pur svolgendo un'attività economica classificata nelle sezioni C-K, per la loro natura giuridica non rientrano nel campo di osservazione Oros. Queste posizioni vengono identificate mediante il CSC ed escluse dalle stime. Allo stesso modo, vengono escluse circa 20 mila unità classificate come condomini. Tra le unità con Ateco di fonte INPS, in realtà, vi è un insieme di posizioni che, per definizione, dovrebbero rientrare anche in ASIA. Si tratta di aziende mediamente di piccole dimensioni e che operano in settori particolari (come il commercio al minuto e l'edilizia) e, come tali, caratterizzate da frequenti modifiche della configurazione giuridica. E' possibile che il codice fiscale di queste posizioni, più suscettibili di cambiamenti, non venga aggiornato nell'archivio INPS contrariamente a quanto avviene in ASIA. Queste posizioni non devono quindi essere escluse dal dominio d'interesse ma vanno monitorate (in particolare, nei *check* ordinari si analizzano quelle con maggiore peso occupazionale). Un ulteriore controllo riguarda le unità con oltre 500 dipendenti, assenti in ASIA e classificate da INPS fuori C-K, al fine di verificare la corretta classificazione al di fuori del campo di osservazione.

[30] In genere a maggio dell'anno *a* viene rilasciato il registro ASIA relativo all'anno *a*-2.
[31] Un esempio per comprendere il problema è il codice Ateco "99999", formalmente inesistente, ma convenzionalmente assegnato dagli operatori INPS ad unità che svolgono attività alle dipendenze di condomini. In Oros, questo codice viene corretto in Ateco "70320" ossia "Amministrazione e gestione di beni immobili per conto terzi".
[32] E' stato verificato, infatti, che oltre il 50% delle unità compresenti in ASIA e in INPS hanno un Ateco a cinque cifre uguale nei due archivi, percentuale che supera l'80% nella corrispondenza a due cifre.

Figura 4: *Il processo di C&C delle variabili anagrafiche*



8. Il trattamento delle mancate risposte totali nella stima definitiva

La rilevazione Oros è stata progettata e realizzata prevedendo una revisione semi-definitiva e una definitiva rispettivamente dopo 4 e 5 trimestri dal rilascio della stima preliminare. La possibilità di rilasciare una stima migliore basata sull'universo nasce dal fatto che l'INPS, con un certo ritardo, mette a disposizione l'intera popolazione delle dichiarazioni DM10. Questo archivio si riempie gradualmente e diviene effettivamente completo solo dopo alcuni anni. Nella fase di primo impianto della rilevazione, sulla base di una valutazione del trade-off tra completezza dell'archivio e tempestività del rilascio della stima definitiva, si è deciso di far scaricare gli archivi dei DM10 a 12 mesi in quanto dopo tale data gli arrivi residui sono molto dilazionati nel tempo. Fino al 2004 l'incidenza dei DM10 non pervenuti entro 12 mesi non ha mai superato il 3% ed ha interessato circa l'1% dell'occupazione totale. Le unità ritardatarie, ai fini della rilevazione, vengono ritenute mancate risposte totali e vengono imputate per vari motivi. Anzitutto potrebbero essere assenti unità di grandi dimensioni, con effetti rilevanti sulle stime di tutte le variabili. D'altra parte, non risposte relative a imprese di piccole dimensioni se possono

risultare pressoché ininfluenti sulle stime delle variabili retributive pro capite non lo sono per la stima dell'occupazione.

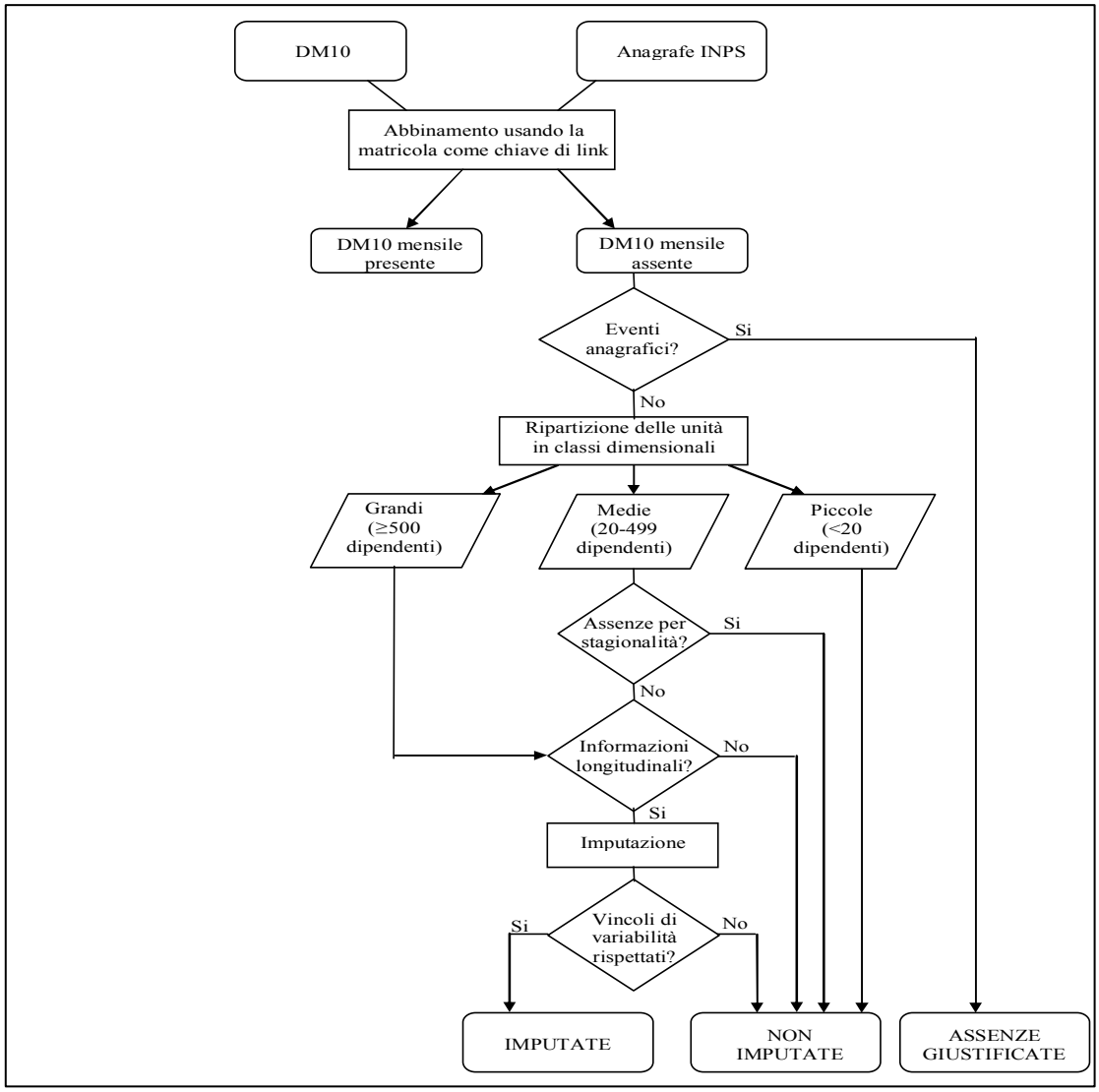
8.1 L'identificazione delle mancate risposte

La procedura di imputazione delle mancate risposte si pone l'obiettivo di ricostruire in valore assoluto le variabili sfruttando le serie storiche di micro dati. Il primo aspetto da affrontare riguarda l'individuazione stessa delle mancate risposte (Figura 5). Diversamente dalle indagini tradizionali, in cui la lista dei non rispondenti è nota, nella rilevazione Oros non esiste una lista teorica di unità attive a causa dei problemi di copertura che caratterizzano l'archivio anagrafico INPS e va, quindi, predetta. Ipotizzando che i trimestri contigui a quello di stima possano essere informativi sullo stato di attività a t, per effetto della bassa persistenza nei ritardi empiricamente constatata nei dati Oros, la lista di stima viene definita come l'insieme delle unità che hanno inviato almeno un DM10 nei 4 trimestri precedenti a quello di stima o nel trimestre di stima e almeno un DM10 in quello successivo. Infatti la scelta di effettuare prima una stima semi-definitiva riferita al trimestre t-4 e di posticipare di un trimestre la pubblicazione del dato definitivo a t-5 risponde proprio alla necessità di utilizzare anche le informazioni del trimestre successivo a quello di stima per una migliore individuazione della lista di unità attive. Dall'imputazione vengono escluse le unità sospese per stagionalità, individuate secondo ipotesi che valutano la sistematicità delle assenze in un periodo prefissato, e le unità con meno di 20 addetti, per le quali si ipotizzano naturali brevi e/o frequenti periodi di sospensione dell'attività. Sono, invece, sempre considerate mancate risposte le assenze che riguardano posizioni di medie e grandi dimensioni, per le quali si suppongono poco probabili momenti di completa inattività per stagionalità o brevi periodi di sospensione. Applicando queste ipotesi, le mancate risposte stimate nella situazione informativa fino al 2004 sono circa 20 mila con una tendenza a ridursi nel tempo. Il loro numero scende drasticamente nell'attuale situazione informativa, in cui si stimano meno di 10 mila mancate risposte. Interessante sottolineare che nei trimestri in cui si sono registrati cali nell'invio dei modelli per cause amministrative, il numero di mancate risposte individuate è risultato visibilmente più elevato.

8.2 I criteri d'imputazione

Per il trattamento delle mancate risposte in casi analoghi a quello in esame la letterature suggerisce due possibili strade: il ricorso a unità donatrici e/o l'adozione di modelli statistico - matematici (Istat, 1989). Entrambe le soluzioni sono subordinate alla disponibilità di informazioni ausiliarie affidabili e alla conoscenza delle caratteristiche strutturali delle unità indagate.
Disponendo di informazioni longitudinali a livello di micro dato, l'imputazione delle mancate risposte è stata pertanto impostata sulla base del solo secondo criterio, adottando un approccio deterministico, in cui si assume una dipendenza di tipo lineare fra il sottoinsieme di variabili di interesse ed un insieme di variabili esplicative. Queste ultime sono principalmente le stesse variabili considerate in istanti temporali differenti da quello corrente. Le funzioni che esprimono tale dipendenza sono a loro volta dipendenti da un insieme di parametri, che vengono stimati sulla base delle informazioni fornite dalle unità rispondenti o dalle stesse unità in istanti temporali precedenti. Naturalmente, il ricorso a questo approccio richiede la verifica del rapporto tra le variabili ausiliarie ed il modello di risposta e che sia appurata l'indipendenza del meccanismo aleatorio di riposta dal livello delle variabili ausiliarie utilizzate. Di fatto, analizzando la distribuzione delle mancate riposte, si è notato come non vi sia una significativa dipendenza tra dati mancanti e livello medio dell'occupazione delle unità ritardatarie.

Figura 5: *Imputazione delle mancate risposte nella stima definitiva*



L'imputazione prevede la ricostruzione dei monti trimestrali delle variabili d'interesse. Noto il numero di mesi di mancata risposta nel trimestre, condizione necessaria per la correzione dei totali delle variabili retributive è la ricostruzione del monte dipendenti. Quando è possibile correggere questa variabile, si garantisce sempre la ricostruzione delle retribuzioni e quindi degli oneri sociali.
La diversa caratterizzazione economica delle variabili d'interesse comporta il ricorso a regole differenziate per la ricostruzione dei dati, orientate alla conservazione dei modelli evolutivi delle variabili osservate.
L'imputazione dell'occupazione avviene sulla base dei valori della stessa variabile nel trimestre in corso, se l'unità ha presentato almeno un DM10 o, quando non disponibile, nel trimestre precedente, nell'ipotesi che la *proxy* più attendibile sia la stessa variabile selezionata dai mesi contigui. Se l'unità è in ritardo da almeno due trimestri viene definita non eleggibile ad imputazione. Genericamente, se nel trimestre *t* di stima definitiva è presente almeno un DM10 con dipendenti, la ricostruzione[33] si basa sulla seguente formula:

[33] L'imputazione, inizialmente, viene effettuata sul totale dei dipendenti ma vengono stimate anche le quote rappresentate dai dirigenti e fatta la distinzione tra *part-time* e *full-time*.

$$\hat{D}_{it} = D_{it} + \frac{D_{it}}{n_{it}}(\hat{n}_{it} - n_{it})$$
[1]

in cui, per l'unità *i*, D_{it} e $\hat{D}_{it}$ rappresentano il monte trimestrale dei dipendenti pre e post imputazione, mentre n_{it} e $\hat{n}_{it}$ sono, rispettivamente, il numero trimestrale dei DM10 pre e post imputazione (e la loro differenza, il numero di mancate risposte nel trimestre).
In alternativa, se l'unità è assente completamente in *t*, ma presenta almeno un DM10 con dipendenti nel trimestre *t*-1, allora si applica la formula:

$$\hat{D}_{it} = \frac{D_{it-1}}{n_{it-1}}\hat{n}_{it}$$
[2]

Per evitare che eventuali anomalie presenti nei dati selezionati per la ricostruzione e/o che *pattern* stagionali non rilevati in fase di definizione dello stato di attività (paragrafo 8.1) si riflettano sul dato ricostruito, si impone un vincolo al valore del dato imputato che viene ritenuto accettabile (mancata risposta imputabile) se la variabilità dell'indicatore imputato, misurata in un numero prefissato di trimestri che includono il dato ricostruito, non viene alterata oltre certi limiti[34].
La natura di spiccata periodicità delle retribuzioni richiede una particolare attenzione nella scelta della base di riferimento da cui effettuare la correzione del dato mancante. Per effetto della presenza di poste erogate periodicamente come la tredicesima e la quattordicesima, l'uso esclusivo dell'informazione corrente, anche se più aggiornata, potrebbe indurre distorsioni nel dato medio mensile ricostruito. In generale, le retribuzioni riferite allo stesso trimestre del precedente anno sono ritenute un riferimento fondamentale per l'imputazione della variabile.
Dato il contesto informativo in cui una generica mancata risposta si colloca, anche la ricostruzione del monte retribuzioni avviene secondo diverse regole d'imputazione[35]. Una prima situazione è quella in cui nel trimestre in corso tutte le dichiarazioni risultano mancanti. In questo caso, l'imputazione fa riferimento allo stesso trimestre dell'anno precedente, quando tale informazione è disponibile. In formule:

$$\hat{R}_{it} = r_{it-4}\left(\frac{\Delta r}{r}\right)_{gt}\hat{D}_{it}$$
[3]

In cui $\hat{R}_{it}$ è il monte trimestrale delle retribuzioni post imputazione, basato sulla retribuzione pro capite $r_{it-4} = \frac{R_{it-4}}{D_{it-4}}$ selezionata da *t*-4 in riferimento all'unità *i*, aggiornata con un tasso di variazione $\left(\frac{\Delta r}{r}\right)_{gt}$ stimato a livello di gruppo *g* (per Ateco e classe dimensionale).
Quando anche in *t*-4 non è presente alcun DM10, le retribuzioni correnti vengono imputate ricorrendo ad un valore mediano della variabile (*Me(r)*), calcolato al tempo corrente a livello di gruppo di stima *g*.
Un secondo contesto informativo è quello in cui nel trimestre di stima almeno un DM10 è presente. In questo caso l'imputazione viene effettuata con un metodo misto, basato su una media ponderata dell'informazione corrente sulla retribuzione pro capite e dell'informazione ausiliaria che, come sopra precisato, può essere, a seconda del contesto informativo, la retribuzione riferita alla stessa posizione contributiva a *t*-4, o una retribuzione mediana calcolata in *t*. La ponderazione si basa su un peso prefissato[36].

---

[34] La valutazione viene effettuata sulla base del coefficiente di variazione calcolato sulla nuova serie di dati relativa agli ultimi cinque trimestri (incluso quello di riferimento con dato imputato). Esso non deve superare di un valore stabilito lo stesso coefficiente calcolato sulla serie di stessa lunghezza, ma con il dato corrente non imputato.
[35] Le variabili obiettivo della rilevazione non devono includere i dirigenti. Quindi, dalla ricostruzione delle retribuzioni totali occorre distinguere e detrarre le retribuzioni dei dirigenti. A tal fine, si ricorre a dei coefficienti di rapporto dedotti dalle risposte fornite dalle stesse unità in altri istanti temporali, o applicando coefficienti stimati a livello di celle aggregate.
[36] Attualmente all'informazione di *t*-4 viene conferito un peso pari all'80% e il rimanente all'informazione corrente.

La ricostruzione del monte oneri sociali, infine, si basa sull'attribuzione dell'aliquota contributiva media (rapporto tra monte oneri e monte retribuzioni) rilevata nel trimestre corrente o in quello più prossimo al trimestre di riferimento, andando indietro fino ad un massimo di un anno. Questa scelta si basa sulla considerazione che i trimestri più vicini a quello di stima colgono più tempestivamente eventuali variazioni legislative. Per le posizioni per cui non si riesce ad effettuare una ricostruzione si attribuisce un'aliquota determinata ex-ante[37]. In formule, la ricostruzione del monte oneri avviene secondo la seguente relazione:

$$\hat{O}_{it} = c_{(t-j)i}\hat{R}_{it} \qquad j=0,\ldots,4 \qquad [4]$$

in cui $\hat{O}_{it}$ è il monte trimestrale degli oneri post-imputazione e c è l'aliquota contributiva media. Nella relazione [4] la scelta di j, ossia del trimestre ausiliario, ricade sul primo valore che rappresenta il trimestre in cui è garantito il verificarsi dell'espressione $c_{(t-j)}>0$[38].
L'impatto complessivo dell'imputazione, si è, notevolmente ridimensionato in seguito al consolidarsi del nuovo contesto informativo. Non solo il numero delle unità individuate come mancate risposte si è ridotto ma negli ultimi trimestri le mancate risposte imputate sono ormai poche centinaia e il loro peso rispetto alle imputabili è sceso a meno del 5%, per la mancanza di informazioni utili alla ricostruzione del dato dovuta alla maggiore persistenza dei ritardi. Tuttavia, l'imputazione viene ritenuta uno strumento ancora utile per far fronte a eventuali cali amministrativi che l'Istat non può prevedere né come tempistica, né come entità.

9. Il trattamento delle imprese di grandi dimensioni e l'integrazione dei dati INPS con la Rilevazione GI

Un aspetto peculiare della stima degli indicatori Oros riguarda il trattamento delle variabili di interesse sul sottoinsieme di imprese di maggiori dimensioni, ossia con più di 500 dipendenti.
In origine, i problemi connessi a tale stima erano prevalentemente riconducibili alle caratteristiche del campione INPS che, pur costituito da un elevato numero di unità, non presentava caratteristiche di casualità, e in particolare risultavano sottorappresentate le imprese di grandi dimensioni. Ciò ha indotto a integrare i dati di fonte INPS con quelli della "Rilevazione sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese dell'Industria e dei Servizi" (GI) (Istat, 2006), che rileva imprese che impiegano circa il 20 per cento del totale dei lavoratori dipendenti nei settori privati extra agricoli dell'economia italiana. L'evoluzione del campione verso l'universo provvisorio ha cambiato le condizioni iniziali; tuttavia, per una serie di motivazioni fondamentali, per le imprese grandi si continuano a utilizzare i dati della Rilevazione GI. Infatti, la presenza di esperti Istat che hanno contatti diretti con queste imprese, che per loro natura sono caratterizzate da continui e significativi cambiamenti, garantisce una migliore qualità dei dati e una più rapida e aggiornata gestione dei fenomeni che le possono interessare. Inoltre, la constatazione che il processo di controllo e correzione delle imprese grandi, specialmente per quanto concerne le loro frequenti trasformazioni giuridiche, viene effettuato nell'ambito della Rilevazione GI, ha indotto a continuare ad integrare le due fonti. Questo implica che, per la quota di popolazione di imprese coperta dalla Rilevazione GI, sia per la stima preliminare sia per quella definitiva, vengono usati i dati di quest'ultima piuttosto che i dati INPS.
L'integrazione dei micro dati della Rilevazione GI implica sostanzialmente due processi fondamentali.
Il primo è rappresentato dalla produzione, a partire dai micro dati economici mensili GI, di variabili che siano il più possibile allineate in termini definitori con le variabili di interesse Oros. Tra le operazioni da effettuare necessariamente ci sono l'aggregazione trimestrale dei dati mensili e il calcolo delle Ula[39].

[37] Attualmente il valore attribuito all'aliquota è pari al 40%.
[38] Per la stima della quota di oneri dei dirigenti, all'aliquota totale si applica una percentuale di differenza pari a quella rilevata nel trimestre selezionato come base per l'imputazione oppure, se il dato non è disponibile, moltiplicando per una quota fissa.

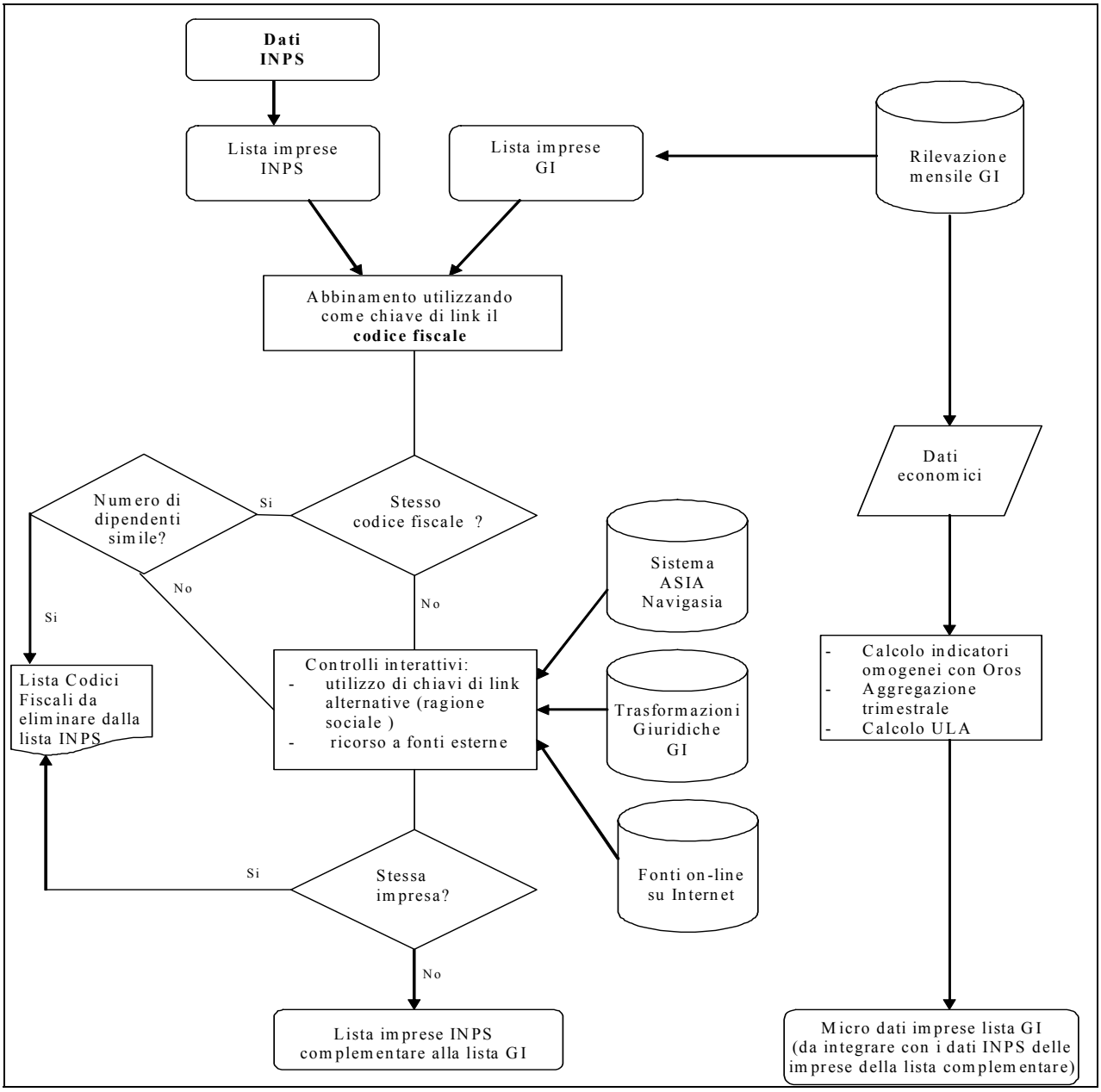
Il secondo, invece, è rappresentato dalla procedura di controllo e correzione volta a garantire la corretta individuazione nell'archivio amministrativo delle imprese contenute nella lista GI. L'integrazione tra le due fonti, infatti, implica la costruzione di due liste complementari di imprese al fine di evitare da un lato duplicazioni delle stesse, dall'altro la mancata assegnazione di una impresa ad una delle due liste. Le difficoltà sono principalmente legate alla constatazione che il codice fiscale, chiave primaria di aggancio tra i due archivi, non è sempre sufficiente per individuare le stesse imprese. Oltre a possibili errori formali nel codice fiscale, infatti, occorre tenere in considerazione che le trasformazioni giuridiche, da cui le grandi imprese sono frequentemente interessate, vengono registrate con tempi e con regole di aggiornamento (implicite o esplicite) diversi negli archivi utilizzati.
Come ci si può attendere, la maggior parte delle imprese della Rilevazione GI sono presenti in INPS con lo stesso codice fiscale; alcune di queste, tuttavia, possono mostrare uno scostamento significativo nel numero di dipendenti rilevati dalle due fonti suggerendo che lo stesso identificativo molto probabilmente fa riferimento a due entità differenti. Questa problematica si presenta, ad esempio, quando una società si divide in due o più società, ma i due archivi registrano l'evento in tempi diversi. Complessivamente queste ultime imprese, unitamente a quelle che nella Rilevazione GI hanno un codice fiscale non presente in INPS, occupano il 12 per cento circa del totale dei dipendenti rilevati dall'indagine GI.
Ogni trimestre, quindi, occorre prestare particolare attenzione a "migrazioni" che possono avvenire da una lista all'altra a seguito di trasformazioni giuridiche. A tale proposito occorre ricordare che la popolazione teorica a cui si riferisce la Rilevazione GI è la popolazione di imprese, nel campo di osservazione compreso tra le sezioni C-K dell'Ateco 2002, con almeno 500 dipendenti in media nell'anno di definizione della base secondo l'archivio di riferimento (ASIA). A causa delle mancate risposte sistematiche, tuttavia, l'indagine si basa su un panel che non include tutte le imprese comprese nell'universo teorico[40]. La logica panel, inoltre, implica che non vengano inserite le nuove grandi imprese entranti nell'universo teorico, né eliminate quelle che scendono sotto la soglia dei 500 dipendenti. Per tenere conto, invece, delle evoluzioni che le imprese presenti nel panel subiscono sono state stabilite precise regole per il trattamento delle trasformazioni giuridiche che dovrebbero consentire di seguire quelle stesse imprese nel tempo, almeno fino al successivo cambio della base (Istat, 2006).
I cambiamenti giuridici, quindi, determinano la nascita e/o la cessazione di alcune società che devono essere necessariamente individuate per evitare distorsioni nelle stime a causa di possibili e frequenti disallineamenti temporali nella registrazione dell'evento nelle due fonti. Per poter correttamente distinguere due liste complementari, quindi, è necessario ricorrere a chiavi di aggancio e informazioni ulteriori rispetto al codice fiscale, come ad esempio la ragione sociale. L'utilizzo di questa informazione, tuttavia, è reso particolarmente complesso dal fatto che tale variabile non è standardizzata nei due archivi. Il processo di normalizzazione, che consentirebbe di fare un *record linkage* probabilistico tra le due fonti, non è ancora stato posto in essere e pertanto questa attività è solo in parte automatizzata.
Altre informazioni per agganciare i due archivi vengono ricavate da fonti esterne: in particolare viene utilizzato, in modo parzialmente interattivo, il database delle trasformazioni giuridiche elaborato nel contesto della Rilevazione GI nel quale vengono registrate informazioni relative ad eventi quali scorpori, fusioni, etc. che consente di ricostruire l'evoluzione delle imprese oggetto di indagine mantenendo memoria storica delle informazioni anagrafiche e dei loro cambiamenti nel tempo. Allo stesso modo, vengono utilizzate le informazioni presenti nel sistema ASIA, visualizzabile attraverso il software Navigasia, che contiene tutte le evoluzioni delle singole imprese (cambi di attività economica, connessioni con altre imprese, partecipazioni a gruppi di impresa, etc.), aggiornato sostanzialmente in tempo reale.

---

[39] Una considerazione importante riguarda l'unità di analisi che è differente nei due casi: per GI, pur con qualche eccezione, è l'unità funzionale laddove l'obiettivo di Oros è, almeno per quanto riguarda il settore di attività economica, l'impresa. Su tale aspetto, è stata scelta una soluzione flessibile permettendo unità di analisi diverse, anche alla luce della considerazione che sono soprattutto le grandi imprese ad avere attività economiche differenziate al loro interno.

[40] Il grado di copertura del panel 2005 rispetto alle posizioni dipendenti delle imprese con almeno 500 addetti risultante all'Archivio Asia 2005 è del 90,6%.

Figura 6: *Il processo di integrazione dei dati INPS con i dati della Rilevazione GI*



Dal 2000, quindi, nella produzione trimestrale a regime degli indicatori Oros si realizza un'integrazione dei dati INPS con la totalità dei microdati GI, sia per la stima definitiva sia per quella anticipata. Le grandi imprese non presenti nel panel della rilevazione GI (denominate GI-INPS), invece, vengono stimate utilizzando i dati INPS. Il trattamento di controllo e correzione delle GI-INPS viene illustrato nel paragrafo 10 che affronta il tema dell'editing selettivo e interattivo: i metodi usati per le imprese grandi sono sostanzialmente identici a quelli impiegati per le imprese di piccole e medie dimensioni sebbene le prime vengano sottoposte ad analisi preliminari in virtù delle differenze che le caratterizzano e che devono essere necessariamente tenute in considerazione.

10. L'editing selettivo e interattivo sui dati trimestrali

10.1 La procedura per le PMI: individuazione delle unità influenti e controlli interattivi

I dati usati per la produzione delle stime definitive vengono sottoposti, a fini amministrativi, a procedure di controllo e correzione da parte dell'INPS. Considerato che queste procedure non sono del tutto note e trasparenti e che sono principalmente orientate alla correzione di alcuni errori di misura, si è ritenuto opportuno sottoporre anche i dati dell'universo ad una fase di editing selettivo che, tuttavia, ha delle caratteristiche diverse da quelle del micro editing effettuato sui dati mensili dell'universo provvisorio. In particolare, per i dati dell'universo l'editing agisce sui dati già aggregati trimestralmente ed eventualmente imputati. Esso, inoltre, è impostato per evidenziare anche eventuali errori su alcune caratteristiche strutturali delle unità, che possono avere effetti rilevanti sulle stime, in particolare errori di classificazione per settore di attività economica e per sottopopolazione di stima.
La procedura di editing selettivo si compone di una fase di individuazione delle unità influenti e di una fase di controllo interattivo di queste ultime e di correzione di eventuali errori.
L'individuazione delle unità influenti si basa su un criterio di sensibilità che misura l'effetto che l'esclusione di ogni singola unità ha sulla stima del parametro di interesse (Hidiroglou e Berthelot, 1986). Questo metodo era stato implementato in passato, per la verifica dei dati per la stima preliminare, quando la numerosità del campione era bassa e la stima era prodotta per ponderazione (Baldi et al. 2004). In quel contesto informativo, il dato di una singola unità poteva diventare influente a causa del suo peso nella procedura di espansione alla popolazione. La correzione poteva riguardare sia il valore della specifica variabile, se palesemente errato, oppure il fattore di espansione se si riteneva che il valore della variabile, seppur anomalo, non rappresentasse un errore, ma risultasse ingiustificato espandere l'anomalia ad un numero alto di unità. Da quando il campione ha praticamente raggiunto le dimensioni dell'universo, e infatti viene definito "universo provvisorio" la stima per ponderazione non è più opportuna, poiché ogni unità rappresenta solo se stessa.
Nelle formulazioni che vengono di seguito presentate, si fa riferimento alla versione generalizzata, che fa uso dei fattori di espansione, in quanto può essere applicata anche ad altre indagini campionarie ed è facilmente adattabile alla stima definitiva ponendo i fattori pari ad uno.
La base del metodo consiste nello scomporre la stima della variazione congiunturale delle variabili di interesse in modo da evidenziare il contributo che ciascuna unità statistica apporta alla variazione. Infatti, anche se i livelli delle variabili sono sicuramente un parametro di interesse, essendo Oros una rilevazione longitudinale, ha come obiettivo primario quello di misurare correttamente le variazioni.
Le variabili per le quali si studia l'influenza delle singole unità sono le retribuzioni lorde e gli oneri sociali rapportati alle Ula. La variabile occupazione, espressa in termini di Ula, entra indirettamente nelle funzioni di influenza, attraverso la variazione tra i trimestri della struttura occupazionale ma al momento, dato che il suo rilascio non è imminente, non si studia l'influenza delle singole unità sull'occupazione. In quanto segue, si concentra l'attenzione sulle retribuzioni, ma le espressioni possono essere generalizzate anche agli oneri sociali. La variabile retribuzioni per Ula può essere scritta come:

$$r_{dt} = \frac{\sum_{i \in C_{dt}} k_{it} R_{it}}{\sum_{i \in C_{dt}} k_{it} U_{it}}$$

[5]

in cui k_{it} è il fattore di espansione, R_{it} è il monte retributivo e U_{it} il numero di Ula per la generica unità i nel dominio d appartenente al campione C_{dt} al tempo t. Data la [5], la stima della variazione assoluta delle retribuzioni per Ula assume la seguente forma:

$$\Delta r_{dt} \equiv r_{dt} - r_{dt-1} = \frac{\sum_{i \in C_{dt}} k_{it} R_{it}}{\sum_{i \in C_{dt}} k_{it} U_{it}} - \frac{\sum_{i \in C_{dt-1}} k_{it-1} R_{it-1}}{\sum_{i \in C_{dt-1}} k_{it-1} U_{it-1}}$$

[6]

Indicando con E_{dt} l'insieme delle imprese entranti tra *t*-1 e *t* nel dominio *d* e con U_d l'insieme delle imprese uscenti tra *t*-1 e *t* dal dominio *d*, si ha che:

$$E_{dt} = C_{dt} - (C_{dt} \cap C_{dt-1})$$
$$U_{dt-1} = C_{dt-1} - (C_{dt} \cap C_{dt-1})$$

[7]

E' utile ricordare che le imprese entranti nel dominio *d* al tempo *t* sono:
1. entrate per motivi demografici, nuove nate e riattivate;
2. entrate per inclusione nel campione, escluse in *t*-1 e incluse in *t*[41];
3. entrate per cambio di attività economica, prima appartenenti ad un dominio *d'* diverso da *d*[42];
4. entrate spurie per trasformazione giuridica[43].

Analogamente le imprese uscenti dal dominio *d* al tempo *t*-1 sono:
1. uscite per motivi demografici, cessate e sospese;
2. uscite per esclusione dal campione, incluse in *t*-1 ed escluse in *t*;
3. uscite per cambio di attività economica, ora appartenenti ad un dominio *d'* diverso da *d*;
4. uscite spurie per trasformazione giuridica.

La [6] può essere scritta come:

$$\Delta r_{dt} = \sum_{i \in C_{dt} \cap C_{dt-1}} \left(g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right) + \sum_{i \in E_{dt}} g_{it} r_{it} - \sum_{i \in U_{dt-1}} g_{it-1} r_{it-1}$$

[8]

dove:

$g_{it} = \frac{k_{it} U_{it}}{\sum_{i \in C_{dt}} k_{it} U_{it}}$ e $g_{i,t-1} = \frac{k_{it-1} U_{i,t-1}}{\sum_{i \in C_{dt-1}} k_{it-1} U_{i,t-1}}$ sono rispettivamente, il peso al tempo *t* e *t-1* in termini di Ula, opportunamente ponderate, relativo all'unità *i*-esima sul totale del dominio di stima *d*.

I tre termini della [8] rappresentano rispettivamente il contributo alla variazione delle retribuzioni per Ula del generico dominio *d* delle unità panel, delle unità entranti, delle unità uscenti.

Considerando che per le entranti $r_{it-1} = g_{it-1} = 0$ e per le uscenti $r_{it} = g_{it} = 0$, la [8] si può scrivere come:

---

41 In questo caso non si usa il termine rispondente, bensì il termine incluse, per indicare genericamente le unità appartenenti al campione INPS. Dato che esso è formato dalle unità che in ciascun periodo inviano il DM10 in formato elettronico, non si dispone a priori di una lista dei possibili rispondenti, così come avviene nelle indagini classiche in cui un campione è definito dall'estrazione dall'archivio di riferimento. Qui, invece, le unità appartenenti al campione si verificano ex-post e l'inclusione implica la risposta.

42 Questa eventualità si può presentare sia per effetto di cambiamenti di attività economica che avvengono nell'archivio Asia, sia per modifiche negli archivi INPS. In ogni caso, in alcune situazioni si può rivelare necessario, al fine di evitare effetti rilevanti sulle stime, di ripristinare i vecchi codici Ateco.

43 In particolare ci si riferisce a quelle unità che per effetto di trasformazioni giuridiche e contestuale cambiamento del codice fiscale confluiscono dal gruppo di stima delle grandi imprese della Rilevazione GI alla lista complementare di imprese INPS. Dati i criteri di conservazione del panel della Rilevazione GI e il modo in cui l'indagine tratta la specifica trasformazione giuridica (cfr. par. 9), al fine di evitare duplicazioni può essere opportuno mantenere le imprese interessate dal cambiamento nel gruppo delle grandi della Rilevazione GI.

$$\Delta r_{dt} = \sum_{i \in C_{dt} \cup C_{dt-1}} \left(g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right)$$

[9]

In termini semplici, la [9] mostra come le singole unità contribuiscono a formare la variazione assoluta delle retribuzione per Ula del dominio *d*.
Data la [9], come indicatore dell'impatto della unità *i*-esima sulla variazione retributiva viene utilizzata la seguente funzione di influenza:

$$f^{r}{}_{i} = \frac{\left| g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right|}{\sum_{i \in C_{dt} \cup C_{dt-1}} \left| g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right|}$$

[10]

dove l'utilizzo del valore assoluto si rende necessario per individuare le unità influenti ai due estremi della distribuzione. La [10] può essere definita *local function*, ovvero funzione che misura l'impatto delle unità su una singola variabile di interesse.
Al fine di disporre di uno strumento unico per individuare sia le unità influenti in termini di retribuzione, le cui relazioni sono appena state esposte, che di oneri sociali, per i quali la *local funcion* ha un'espressione equivalente alla [10], si rende utile la costruzione di una sola funzione di influenza, ovvero il passaggio ad una *global function*. Quella proposta può essere scritta come:

$$f_i = \frac{\left| g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right| + \left| g_{it} \Delta o_{it} + o_{i,t-1} \Delta g_{it} \right|}{\sum_{i \in C_{dt} \cup C_{dt-1}} \left(\left| g_{it} \Delta r_{it} + r_{i,t-1} \Delta g_{it} \right| + \left| g_{it} \Delta o_{it} + o_{i,t-1} \Delta g_{it} \right| \right)}$$

[11]

Le funzioni così definite vengono calcolate per ogni dominio, individuato per divisione di attività economica della classificazione Ateco 2002[44].
Le unità esaminate interattivamente sono quelle che hanno i valori maggiori di f_i. I valori soglia, o di *cut-off*, sul campo di variazione di f_i, vengono scelti in base alla numerosità della popolazione del dominio, in modo da generare un numero accettabile di casi da controllare in maniera interattiva. Tale numero, infatti, deve essere contenuto perché la fase di controllo e correzione avvenga in tempi molto brevi, dati i pressanti vincoli di tempo nel rilascio dei dati congiunturali. D'altra parte esso deve essere sufficientemente ampio per contenere tutte le unità rilevanti. Nel corso del tempo i valori soglia sono stati più volte ritoccati, sulla base dell'esperienza pratica. Nell'editing sull'universo, le soglie attuali sono tre ed equivalgono al 5 per cento quando le unità nel dominio sono numericamente inferiori a 1.000. Se tale numero è compreso tra 1.000 e 49.999 la soglia è dell'1 per cento e scende allo 0,5 per cento quando ci sono 50.000 o più unità nel dominio.
Una volta selezionate le unità influenti, segue la fase di controllo interattivo per individuare eventuali errori. Per agevolare i controlli è stata sviluppata *ad hoc* una maschera di controllo, avvalendosi della procedura Sas FSEDIT, in cui le unità sono presentate, all'interno di ogni dominio, in ordine decrescente di influenza. Le variabili riportate nella maschera possono essere raggruppate in tre sottoinsiemi:

1. informazioni identificative e strutturali della unità (matricola, codice fiscale, ragione sociale, ecc.), sia per il trimestre corrente che per il precedente, per mettere in luce eventuali errori di collocazione dell'unità nel settore di attività economica o nella sottopopolazione di stima;

[44] La divisione rappresenta un livello di disaggregazione maggiore rispetto al dominio di pubblicazione attuale di Oros, ossia la sezione, ma è in linea con il dettaglio necessario per le stime richieste dal Regolamento STS.

2. informazioni sull'influenza dell'unità. Oltre al valore della funzione globale di influenza f_i, vengono riportati i valori delle funzioni locali f_i^r e f_i^o per dare un segnale immediato di incoerenza tra le variabili relative a retribuzioni lorde ed oneri sociali. Normalmente, infatti, le due funzioni locali hanno valori molto simili; quindi uno scostamento tra le due indica che una delle variabili è affetta da qualche anomalia. L'informazione della funzione globale è utile invece per valutare l'eventualità di proseguire nell'analisi: se l'influenza diventa molto bassa, può rivelarsi opportuno passare al dominio successivo;
3. variabili di interesse e variabili ausiliarie. Le informazioni ausiliarie sono rappresentate dalle variabili di interesse, riferite ai trimestri *t*-1, *t*-4 e *t*-5. In particolare la presenza dei valori a *t*-4 aiuta a capire se una anomalia identificata nel confronto tra i trimestri *t* e *t*-1, non sia in realtà spiegabile attraverso variazioni stagionali. Altra variabile rilevante di controllo è il rapporto tra oneri sociali e retribuzioni, che rappresenta l'aliquota contributiva media effettiva dell'impresa. Costituisce un rapporto caratteristico che ha un campo di variazione, ad eccezione di alcuni settori e altri casi particolari, compreso tra il 30 e il 40 per cento pertanto scostamenti da questi valori tipici sono segnale di errore di misura.

Nel corso del tempo, gli interventi mirati a correggere gli errori di misura segnalati dall'editing selettivo si sono notevolmente ridotti. Le correzioni attuate sui dati economici hanno riguardato sia le retribuzioni che gli oneri sociali e sono state generalmente risolte mediante divisioni o moltiplicazioni per multipli di 10. Attualmente vengono principalmente corretti errori di classificazione per attività economica e per sottopopolazione di stima.

10.2 La procedura per le imprese INPS di grandi dimensioni

La procedura di editing selettivo descritta per il controllo dei dati trimestrali delle PMI viene applicata anche alle imprese di grandi dimensioni che non fanno parte della Rilevazione GI (vedi paragrafo 9), per la stima delle quali viene utilizzato il dato amministrativo[45]. Considerata la peculiarità delle imprese grandi, tuttavia, per esse sono previsti anche dei controlli preliminari ulteriori.

Il processo di costituzione della lista di imprese INPS, complementare alla lista GI, dovrebbe garantire di evitare duplicazioni di stime per le stesse imprese tra le due diverse fonti o l'esclusione totale di alcune imprese da entrambe le stime; tuttavia, un ulteriore verifica è comunque utile e necessaria.

Un primo controllo consiste nel fornire agli esperti della Rilevazione GI la lista delle imprese più grandi non presenti nella rilevazione ma presenti nei dati amministrativi: facendo prevalentemente riferimento alla ragione sociale, si deve stabilire se nella lista sono erroneamente presenti imprese che vengono rilevate dall'indagine GI.

Inoltre, considerato che le GI-INPS sono tutte potenzialmente influenti in virtù delle loro dimensioni, è importante valutare e individuare a priori eventuali cambiamenti che, tra un trimestre e l'altro, hanno interessato queste unità per comprendere e quantificare correttamente il loro impatto sulle stime degli indicatori di interesse. L'uscita o l'entrata, in un trimestre, di un'impresa di grandi dimensioni, nella maggior parte dei casi non è attribuibile a un fenomeno vero e proprio di nati-mortalità d'impresa, ma è l'effetto di fenomeni spuri, come trasformazioni giuridiche e cambiamenti dell'attività economica svolta, che determinano imprese solo "apparentemente" nuove. Ad esempio, l'uscita di un'impresa da un settore potrebbe essere compensata dall'entrata di un'impresa simile, frutto di una trasformazione giuridica della prima, in un altro settore.

Come potenziale segnalatore dell'esistenza di un fenomeno che ha interessato l'unità di analisi, si utilizza la variazione occupazionale congiunturale: una variazione superiore a una soglia di *cut-off* stabilita potrebbe essere rivelatrice di un possibile evento che ha interessato l'impresa.

Per calcolare tale indicatore, tuttavia, è necessaria un'operazione preliminare di ricostruzione dell'impresa. Si ricorda, infatti, che l'unità di rilevazione amministrativa è la posizione contributiva e, spesso, le imprese di dimensioni maggiori attivano presso l'INPS più posizioni contributive. Per

[45] Si tratta delle GI-INPS che, mediamente, occupano poco più dell'1 per cento dei dipendenti totali delle imprese nelle sezioni da C a K dell'Ateco 2002.

valutare l'impresa nel suo complesso, quindi, è necessario ricostruirla e a tale scopo viene usato come identificativo il codice fiscale. Qualora il codice fiscale non sia presente o sia stato definito non formalmente corretto a seguito della procedura descritta nel paragrafo 7, ogni singola posizione contributiva viene considerata un'impresa.
Le imprese che presentano una variazione dei dipendenti totali rispetto al trimestre precedente superiore ad un valore prestabilito vengono selezionate e abbinate, utilizzando come chiave di *link* il codice fiscale, con il Sistema informativo ASIA dell'Istat, in particolare con il database storico delle imprese italiane in cui vengono registrate le informazioni più aggiornate sulle stesse. Tale sistema prevede, tra le altre cose, anche la gestione di tutti gli eventi di connessione, quali fusioni, scorpori, cessioni, cessazioni, ecc.. Le informazioni storiche sul codice di attività economica prevalente, presenti nello stesso sistema informativo, vengono invece utilizzate per giustificare eventuali passaggi da un'attività ad un'altra. L'occupazione d'impresa viene riportata nella maschera di controllo, introducendo un ulteriore elemento di novità rispetto all'editing sulle PMI.
Questi controlli preliminari sulle GI-INPS, vengono effettuati sia nel processo di stima preliminare sia in quello di stima finale, mentre l'editing selettivo per quella stessa sottopopolazione viene condotto solo sui dati utilizzati per la stima finale.

11. L'imputazione delle agenzie di lavoro interinale

Nel settore della produzione dei servizi alle imprese (sezione K dell'Ateco 2002) un peso molto rilevante in termini di occupazione è rivestito dalle agenzie di fornitura di lavoro interinale individuate con il codice Ateco 74.50.2[46]. Ad oggi, queste unità rappresentano poco meno del 3% dell'occupazione totale della rilevazione e ricoprono circa il 20% dell'occupazione della sezione K. La loro dimensione media supera 1500 dipendenti e, tra di esse, poche unità assumono oltre il 50% del peso del gruppo: l'assenza nei dati anche di una sola di queste unità può avere un effetto rilevante sulle stime, rendendo opportuno un monitoraggio specifico. A tale scopo è stata predisposta una procedura *ad hoc* per l'imputazione delle imprese interinali ai fini della produzione delle stime preliminari, mentre per le stime finali ad esse viene applicata la procedura di imputazione generale (descritta al par. 8). Il controllo di queste unità e l'individuazione di assenze realmente attribuibili a mancate risposte è reso estremamente complicato dalla loro forte dinamicità che si riflette nella gestione delle posizioni, che vengono aperte e chiuse, sospese e riattivate con estrema velocità e con i relativi dipendenti che transitano tra posizioni anche afferenti ad imprese diverse. Inoltre, la diffusione nell'utilizzo di lavoro interinale osservata negli ultimi anni si è riflettuta in una considerevole crescita nel numero di posizioni contributive aperte presso l'INPS. Il problema è in parte attenuato dal numero limitato di unità che rientrano in questo gruppo: attualmente la lista di stima è costituita da circa 150 posizioni contributive, attive secondo le informazioni anagrafiche correnti (non cessate, né sospese) e in base ad alcune ipotesi sui loro pattern di presenza in termini di DM10 inviati. Nel trimestre t, infatti, le uniche indicazioni con le quali predire lo stato di attività derivano dalle variabili anagrafiche (data di costituzione o riattivazione, di cessazione o sospensione) il cui uso esclusivo, tuttavia, può determinare problemi di sovracopertura a causa dei ritardi di aggiornamento degli eventi da parte dell'INPS. Per contenere questo problema, la lista di stima delle interinali viene limitata all'insieme delle unità attive secondo le informazioni sui DM10 pervenuti nel corso dell'ultimo anno a cui vengono aggiunte le unità nate nello stesso intervallo di tempo, anche se non hanno presentato alcuna dichiarazione contributiva. Da questo insieme vengono escluse le unità non presenti a t per sospensione periodica dell'attività, secondo alcune ipotesi di persistenza nelle assenze. Le residue assenze vengono definite mancate risposte.
Data la tipicità delle imprese interinali, prima di procedere all'operazione d'imputazione delle mancate risposte così definite, vi è una fase preliminare di analisi, che ha l'obiettivo di verificare l'effettiva assenza di alcune unità definite mancate risposte secondo le ipotesi postulate. In particolare, vengono monitorate longitudinalmente le unità di più grandi dimensioni e, qualora vi siano state variazioni

[46] Le informazioni riferite ai lavoratori interinali nei DM10 sono rilevate solo dal lato delle società fornitrici e non dal lato di quelle utilizzatrici e vengono, pertanto, incluse tutte nella divisione 74 dei servizi alle imprese.

significative nel numero dei dipendenti, viene effettuato un monitoraggio interattivo per verificare se tali variazioni siano dovute ad assenze per eventi demografici o per trasformazioni giuridiche o sono attribuibili a mancate risposte. Nei primi due casi queste unità vengono sottratte dalla procedura automatica d'imputazione.
Per le unità interessate da mancate risposte, la ricostruzione delle variabili d'interesse avviene secondo un approccio deterministico che sfrutta le informazioni longitudinali sulle stesse variabili, disponibili a livello micro nell'arco del periodo su cui si definisce la lista di stima (t, t-4). In particolare, per dipendenti e retribuzioni si seleziona il primo dato medio mensile disponibile a partire da t-1[47], che viene attualizzato con un tasso mediano calcolato a partire dai tassi di variazione delle rispettive variabili, stimati sulle unità rispondenti (r) nel trimestre corrente (t) e nel trimestre da cui si selezionano le informazioni per l'imputazione (t-j). In formule, la ricostruzione di occupazione e retribuzioni pro-capite (Y) per la generica unità i può essere espressa come:

$$\hat{Y}_{it} = Y_{it-j}(1 + \dot{y}_{t,t-j}) \qquad j=1,\ldots,4$$
[12]

in cui $\dot{y}_{t,t-j} = Me(Y_{rt} / Y_{rt-j} - 1)$.
Per quanto riguarda gli oneri sociali, partendo dal presupposto che il loro livello sia fortemente influenzato dalla legislazione sugli obblighi contributivi, si ritiene che il trimestre corrente sia l'istante più adeguato da cui trarre informazione. A partire dal gruppo dei rispondenti (r) in t, si calcola una misura dell'aliquota effettiva vigente ($\hat{o}$) come mediana del rapporto tra monte oneri (O) e monte retribuzioni (R), in seguito applicata al monte retributivo ($\hat{R}$) ricostruito come prodotto tra le retribuzioni pro capite e i rispettivi dipendenti stimati. In formule:

$$\hat{O}_{it} = \hat{R}_{it}\hat{o}_t$$
[13]

dove $\hat{o}_t = Me(O_{rt} / R_{rt})$.
Data la modalità di definizione della lista di stima (le unità con mancata risposta hanno almeno un DM10 tra t-4 e t), la ricostruzione dei dati mancanti è sempre garantita, ad esclusione dei casi in cui vi siano soltanto dichiarazioni a t e le unità non sono neo nate nel trimestre. Se le unità sono neo-nate, infatti, la ricostruzione dell'occupazione si basa sull'informazione anagrafica, non sempre disponibile, relativa al numero di dipendenti all'iscrizione (per le retribuzioni si fa riferimento ad un dato mediano sui rispondenti). Nei casi in cui, invece, si ha solo informazione a t affetta da mancata risposta, non si imputa perchè i dati presenti potrebbe essere caratterizzati da variazioni occupazionali periodiche e/o da poste straordinarie sulle retribuzioni che andrebbero a riflettersi erroneamente sui dati da ricostruire. In generale, il dato corrente affetto da mancata risposta, tuttavia, viene utilizzato per monitorare la congruità dei valori ricostruiti sulle unità imputate.
L'imputazione delle mancate risposte ha avuto un impatto rilevante sulla stima preliminare delle imprese interinali nella situazione informativa precedente l'obbligo di invio telematico delle dichiarazioni contributive. Negli ultimi due anni, l'effetto si è notevolmente ridimensionato mostrando, tuttavia, una certa variabilità tra i diversi trimestri. In termini di occupazione, nella nuova situazione informativa l'imputazione ha comportato un aumento medio superiore all'1% mentre l'impatto è stato di minor rilievo su retribuzioni e oneri.
In generale, il metodo di imputazione delle interinali ha dato buoni risultati. Gli errori di revisione sulle variabili d'interesse, che vengono trimestralmente monitorati, si sono ridotti nel tempo, anche grazie ad una maggiore conoscenza del fenomeno che ha consentito, in alcune occasioni, di discriminare sull'eventualità di imputare o meno in casi di incertezza. Considerato che nelle stime definitive le

[47] Poiché le unità da ricostruire potrebbero essere state già interessate da mancate risposte nel periodo di riferimento per la ricostruzione del dato mancante, al fine di selezionare informazioni corrette, vengono utilizzati i dati eventualmente imputati.

imprese interinali vengono sottoposte al processo generalizzato di imputazione, applicato a tutte le unità (cfr. par. 8), gli errori di revisione vengono utilizzati anche per monitorare la correttezza dei dati ricostruiti a *t*-4 e *t*-5 e per apportare eventualmente delle modifiche al metodo di imputazione applicato.

12. I controlli per la validazione dei dati macro

L'ultima fase di C&C della rilevazione Oros è finalizzata alla validazione dei macro dati a livello di dominio di pubblicazione. Si tratta di una fase particolarmente delicata, in cui si deve essere in grado di discernere tra andamenti irregolari dovuti a fattori di natura economica o giuridica, quindi accettabili, e veri e propri errori residui e influenti sulle stime. In quest'ultimo caso può essere necessario anche ritornare sul micro dato, con vincoli temporali molto stretti (circa 2/3 giorni), e lo sviluppo di nuovi programmi ad hoc. Ogni trimestre devono essere validate le nuove stime: quella provvisoria riferita al trimestre *t* e quella definitiva relativa al trimestre *t-5*.
I controlli sviluppati al fine di individuare la presenza di eventuali anomalie sono molteplici e possono essere raggruppati in:

1. analisi in serie storica per domini e sottodomini di stima;
2. confronti con i dati di altre fonti;
3. analisi delle relazioni tra variabili;
4. analisi della revisione della stima provvisoria.

Mentre i primi tre tipi di controllo sono finalizzati a validare prevalentemente la stima provvisoria, l'ultimo riguarda esclusivamente la validazione della stima finale. Più nel dettaglio, di seguito vengono illustrati i metodi e le tecniche utilizzate nell'ambito di ciascuno dei quattro gruppi di controlli macro.
1. Il macroediting basato sull'analisi in serie storica viene effettuato su ogni singolo indicatore Oros e per ciascun aggregato di pubblicazione attraverso un'analisi di coerenza tra l'ultima stima provvisoria prodotta e l'intera serie storica. Se questa prima analisi evidenzia degli aggregati sospetti, identificati sulla base di soglie di accettazione stabilite, si approfondisce l'ispezione al loro interno con un controllo ad un livello di dettaglio maggiore, in particolare per divisione. Tra i check in serie storica è sempre prevista anche l'analisi degli indicatori per le quattro sotto-popolazioni di stima[48], molto utile per circoscrivere l'insieme delle unità sospette sulle quali proseguire i controlli a livello micro. Oltre ai metodi analitici basati sull'utilizzo di tabelle e grafici, prodotti in modo automatizzato in linguaggio Sas, è stata implementata una procedura automatica (Tuzi, 2008) basata sull'algoritmo di check degli outliers di TERROR ("Tramo for errors"), un'applicazione del software per la destagionalizzazione TRAMO-SEATS (Caporello e Maravall, 2002). Questo metodo basa la valutazione di casi "sospetti" sulla comparazione statistica tra la predizione ottimale del dato, ottenuta con un modello di serie storiche, e il dato effettivo. Oltre a consentire di testare in tempi molto brevi un numero ampio di serie storiche, esso permette di valutare l'effetto di possibili componenti di ciclo-trend e di stagionalità della serie stessa su eventuali dati anomali.
2. I controlli basati sul confronto con altre fonti vengono effettuati ogni trimestre con i dati della Rilevazione GI, i dati trimestrali della Contabilità Nazionale e quelli dell'Indice sulle Retribuzioni Contrattuali (IRC), attraverso un'analisi grafica completamente automatizzata. Per quanto riguarda la prima fonte, il confronto è possibile, tenendo conto di alcune piccole differenze definitorie[49], con le variabili retribuzioni lorde e costo del lavoro. Per i confronti con i dati trimestrali della Contabilità Nazionale, nonostante la comparabilità dei dati non sia totale[50], vengono utilizzati i monti del reddito da lavoro dipendente e delle retribuzioni lorde e, per differenza tra i due quello degli oneri sociali, che

[48] Cfr. paragrafo 2.
[49] In particolare, le retribuzioni delle grandi imprese comprendono anche i compensi in natura, mentre Oros rileva soltanto i compensi in denaro, e il costo del lavoro comprende oltre ai contributi obbligatori, anche quelli volontari e figurativi assenti invece nell'indicatore Oros.
[50] Oltre alle differenze definitorie su retribuzioni lorde e costo del lavoro simili a quelle con la Rilevazione GI, ricordiamo che i dati trimestrali stimati dalla Contabilità nazionale includono i dirigenti e i lavoratori irregolari e hanno come popolazione obiettivo anche le unità del settore pubblico non rilevate da Oros.

rapportati alle Ula medie permettono il calcolo degli indici di valore e delle variazioni tendenziali confrontabili con gli indicatori Oros. Infine, l'andamento tendenziale delle retribuzioni lorde per Ula viene confrontato con quello dell'indice sulle retribuzioni contrattuali (IRC), utile per spiegare variazioni legate alla sola componente tabellare.
3. Il terzo gruppo di controlli completa i precedenti attraverso l'analisi di alcune relazioni tra le variabili. Particolarmente importante è il controllo della stabilità della relazione tra gli oneri sociali e le retribuzioni lorde, effettuato calcolando il rapporto percentuale tra i monti trimestrali delle due variabili per sezione di attività economica e confrontandolo graficamente con lo stesso rapporto nei dati di Contabilità Nazionale. Un altro indicatore utilizzato per verificare la relazione attesa tra le due variabili è lo scostamento tra le variazioni tendenziali delle retribuzioni lorde per Ula e quelle degli oneri sociali per Ula.
4. Il set di controlli trimestrali a regime si conclude con l'analisi della revisione della stima provvisoria per la validazione della stima finale che viene prodotta e rilasciata con un ritardo di circa un anno. Le differenze tra la stima provvisoria e quella definitiva non riguardano soltanto il numero delle unità utilizzate, ma ricordiamo che la stima definitiva viene realizzata proprio per incorporare tutta l'informazione resasi disponibile successivamente al rilascio della stima provvisoria, anche al costo di revisioni non irrilevanti (Congia e Rapiti, 2007). Pertanto, l'analisi della revisione è uno strumento utile per controllare i diversi fattori che determinano la revisione della stima provvisoria e, quindi, per individuare eventuali anomalie non giustificate. A questo scopo viene analizzata la revisione delle variazioni tendenziali dei tre indicatori per dominio di pubblicazione e per sotto-popolazione di stima.
Se le analisi a livello macro sopra illustrate forniscono dei segnali che inducono a sospettare la presenza di anomalie che possono nascondere errori residui, si passa a una serie di controlli più approfonditi sui micro dati (*drill-down*). Questi ulteriori controlli micro sono stati sviluppati nel corso del tempo, inizialmente *ad hoc* sulla base delle evidenze macro che sono emerse di volta in volta ciascun trimestre e successivamente sono stati implementati e estesi, andando a costituire un'insieme di procedure consolidate che vengono utilizzate a regime qualora si presentino gli stessi o simili problemi.
Le anomalie possono derivare da outliers nei micro dati (errori o valori anomali), ma anche da modifiche nei metadati che non sono state adeguatamente recepite, vale a dire da cambiamenti normativi che non sono stati interpretati e inseriti in modo appropriato e completo nella fase di trattamento preliminare[51] (cfr. par.5). Poiché sono piuttosto frequenti modifiche della normativa che hanno un impatto visibile sui macro dati, è necessario distinguere i casi in cui l'impatto è corretto da quelli in cui le anomalie sono da attribuire a errori nel recepimento delle modifiche dei metadati[52]. Questo tipo di controlli generalmente porta a confermare che l'impatto delle modifiche della disciplina previdenziale sui macro dati è corretto, raramente ha evidenziato errori di interpretazione della normativa, mentre altre volte ha messo in luce errori introdotti nei programmi Sas in occasione del loro aggiornamento per recepire proprio le continue modifiche della normativa stessa.
Uno dei controlli più frequenti viene effettuato a livello settoriale analizzando l'impatto delle dinamiche demografiche sulle variazioni tendenziali di retribuzioni e oneri sociali, attraverso l'esame delle unità entranti-uscenti-compresenti. Infine, a livello micro è possibile verificare se alcuni aggregati anomali sono determinati da sovracorrezioni effettuate, o da outliers sfuggiti, nelle fasi di editing precedenti. L'impatto di ouliers non corretti in queste ultime fasi può essere notevole sui macro dati (cfr. par. 6).
La correzione degli eventuali errori individuati viene effettuata, ove possibile, a livello micro in modo da preservare la coerenza tra macro e micro dati.

[51] In particolare, un'errata interpretazione della normativa può portare alla non corretta selezione dei codici da includere nella ricostruzione delle variabili statistiche e ciò può avere un effetto di distorsione delle stime non trascurabile.
[52] Tra questo controlli ricordiamo, a titolo esemplificativo, quelli effettuati su nuove e particolari tipologie contrattuali, come il lavoro a chiamata che ha avuto un forte impatto sulle retribuzioni pro capite in quanto caratterizzato da un input di lavoro molto ridotto, oppure quelli per la verifica di andamenti tendenziali anomali degli oneri sociali determinati dalla concessione o dalla conferma di sgravi contributivi in alcuni settori, come quello edile in cui uno sgravio specifico viene reiterato di anno in anno ma in periodi diversi.

13. Conclusioni

In questo documento è stato descritto il processo di controllo e correzione della rilevazione Oros caratterizzata dall'utilizzo estensivo di dati amministrativi, integrati con i dati della Rilevazione GI, per la produzione di statistiche congiunturali da rilasciare con notevole tempestività.
Non potendosi basare su analoghe esperienze precedenti anche a livello internazionale, le procedure di C&C sono state sviluppate progressivamente attraverso un graduale processo di apprendimento, tenendo conto dei numerosi vincoli strutturali e temporali e dell'evoluzione continua del contenuto informativo della fonte amministrativa. Infatti, mentre in un'indagine tradizionale è possibile progettare e disegnare la rilevazione in modo da prevenire e ridurre al minimo i possibili errori non campionari, in una rilevazione che utilizza estensivamente i dati amministrativi, i controlli possono essere effettuati solo ex post. L'uso di dati amministrativi in questo contesto implica una forte dipendenza dall'ente fornitore e un maggior rischio di incorrere in problemi, anche di natura informatica e tecnica, che possono influire sulla qualità dei dati. La strategia adottata per ridurre questi rischi si è basata su: relazioni più strette e continue con l'ente fornitore dei dati; una sequenza di controlli sistematici e pervasivi che consentano di intercettare gli errori influenti; procedure altamente selettive e interattive; un sistema di controlli flessibile e modulare che apprende dall'esperienza e utilizza la sistematica documentazione per adattarsi di continuo alle mutevoli esigenze produttive.
Nel complesso le procedure di C&C sono risultate affidabili in termini di efficacia ed efficienza, anche se sarebbero possibili ulteriori miglioramenti. Nel trattamento preliminare, ad esempio, sarebbe utile semplificare, accorpare e automatizzare ulteriormente alcune fasi per gestire in modo più efficiente le modifiche della normativa e delle caratteristiche dei dati amministrativi. Nel microediting dei dati mensili, invece, potrebbero essere riviste alcune soglie per controllare in modo selettivo anche le unità di dimensioni maggiori, mentre l'imputazione delle mancate risposte nelle stime definitive potrebbe essere estesa anche alle unità di minori dimensioni sfruttando meglio le informazioni disponibili.
In ogni caso, il processo di C&C e l'impianto stesso di una rilevazione fortemente dipendente dai dati amministrativi è destinato, comunque, ad evoluzioni più frequenti rispetto alle indagini tradizionali. Pertanto, possibili futuri sviluppi del processo di C&C andrebbero collocati nel contesto di una revisione generale della rilevazione Oros. L'impianto della rilevazione, infatti, può essere semplificato grazie all'evoluzione dell'assetto informativo avvenuta a partire dalla seconda metà del 2004, quando è stato introdotto l'obbligo dell'invio telematico della dichiarazione contributiva.

Bibliografia

AA.VV. *Model Quality Report in Business Statistics. Theory and Methods for Quality Evaluation.* Volume 1. Lussemburgo: Eurostat, 2001.

Baldi C., F. Rapiti. "Wages and employment official statistics using INPS data: a preliminary proposal and some methodological and quality problems". *Contributi Istat*, n.16, 1999.

Baldi C., E. Cimino, F. Rapiti, P. Minicucci, R. Succi, D. Tuzi. "L'utilizzo dei dati INPS per la stima trimestrale del numero dei dipendenti, le retribuzioni, il costo del lavoro e le ore lavorate". *Documenti Istat*, n.14, 2001.

Baldi C., F. Ceccato, E. Cimino, M.C. Congia, S. Pacini, F. Rapiti, D. Tuzi. "Use of Administrative Data to produce Short Term Statistics on Employment, Wages and Labour Cost". *Essays*, 15. Roma: Istat.

Caporello G., A. Maravall. *A tool for quality control of time series data. Program TERROR.* Madrid: Bank of Spain, 2002.

Cimino E., M.C. Congia, P. Minicucci, F. Rapiti. *Banca dati normativa Oros sul lavoro dipendente e contribuzione. Procedure, metodi, riferimenti normativi.* Roma: Documento interno Istat, 2003.

Cochran, W.G. *Sampling techniques*, third edition. New York: Wiley, 1977.

Congia M.C., F. Rapiti. "Gli indicatori di revisione nella rilevazione trimestrale Oros sulle retribuzioni di fatto, gli oneri sociali e il costo del lavoro". In *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI – 1a giornata.* Roma. Anche disponibile in *Contributi Istat*, n. 6, 2007.

Hidiroglou M.A., J.M. Berthelot. "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology*, 12, Statistics Canada, Ottawa, 1986.

Istat. *Il sistema di controllo della qualità dei dati*. Note e relazioni, Volume 6, n. 1. Roma: Istat, 1989.

Istat. *Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese*. Metodi e Norme, n. 29, Roma: Istat, 2006 ().

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Rapporto tecnico del progetto Europeo EDIMBUS, 2007 (http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/RPM_EDIMBUS.PDF).

Tuzi D. *Una procedura automatica di controllo di dati anomali sulle serie storiche di Oros mediante TERROR*. Documento interno. Roma: Istat, 2008.

Prevenzione degli errori, integrazione dei dati e metodi statistici nel processo di controllo e correzione dell'Indagine trimestrale sui posti vacanti e le ore lavorate

Ciro Baldi, Istat, Occ/A
Diego Bellisai, Istat, Occ/A
Stefania Fivizzani, Istat, Occ/A
Marina Sorrentino, Istat, Occ/A

Sommario: Il presente lavoro documenta i metodi di controllo e correzione sulle posizioni lavorative e sui posti vacanti dell'Indagine Trimestrale sui Posti Vacanti e le Ore Lavorate sia riguardo alle operazione di prevenzione degli errori nella fase di rilevazione dei dati, sia riguardo la fase più strettamente detta di Editing and Imputation che opera sui dati raccolti. Il processo di controllo nella fase di rilevazione del dato, che avviene con modalità mista CATI e WEB/E-mail, è pervasivo e va dalla formazione degli operatori CATI, al monitoraggio del numero delle interviste, al folllow-up su imprese non rispondenti di grandi dimensioni. Questo insieme di operazioni si dimostrano essenziali per massimizzare il rendimento della rilevazione sia in termini di quantità che di qualità. Le operazioni di Editing and Imputation prevedono l'integrazione dell'indagine con altre due fonti di dati: l'Indagine Mensile su Occupazione, Orari di lavoro, Retribuzioni e Costo del Lavoro nelle Grandi Imprese (GI) e l'Indagine Trimestrale su Occupazione, Retribuzioni e Oneri Sociali (OROS). L'integrazione dei dati, da un lato consente di ottenere stime coerenti con il sistema OROS-GI, dall'altro migliora la qualità dei dati e delle stime attraverso le operazioni di controllo e di imputazione. Infine i metodi di controllo e correzione (editing selettivo, imputazione) dei posti vacanti sono stati sviluppati per tenere in debito conto le caratteristiche della variabile posti vacanti che si presenta fortemente asimmetrica e concentrata sullo zero. Tra essi l' imputazione delle mancate risposte mediante donatore di minima distanza si è dimostrato uno strumento flessibile per ottenere stime con le proprietà desiderate.

Parole chiave: Prevenzione degli errori, monitoraggio della rilevazione, integrazione dei dati, donatore di minima distanza, editing selettivo, imputazione.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Introduzione[53]

Nel quadro delle statistiche congiunturali sul mercato del lavoro, è emersa in anni recenti la necessità di avere informazioni infra-annuali sui posti di lavoro vacanti, ovvero sulle ricerche di personale da parte del sistema delle imprese. Indicatori su questa variabile possono aiutare a comprendere ed a prevedere il ciclo occupazionale dell'economia. Per capire l'importanza che la statistica europea assegna agli indicatori sui posti vacanti basti pensare che il *Job Vacancy Rate* è uno dei quattro *Principal European Economic Indicators* (*PEEIs*)[54] sul mercato del lavoro. È per questo motivo che il sistema statistico europeo, prima con un *Gentlemen's Agreement* fra i Paesi dell'Unione ed Eurostat, poi con un regolamento in corso di approvazione si è apprestato a regolamentare la produzione e la fornitura di questo tipo di informazioni.
La produzione di tali indicatori è realizzata in Italia, a partire dal terzo trimestre 2003, dall'Indagine trimestrale Istat sui posti vacanti e le ore lavorate (nel seguito, indicata anche con l'acronimo VELA). La rilevazione delle ore lavorate nella stessa indagine è chiaramente motivata dall'utilità derivante dal fatto che sia un'unica rilevazione a raccogliere un insieme di variabili atte a misurare la dinamica dell'input di lavoro. D'altro canto, la rilevazione delle ore lavorate soddisfa le esigenze informative che sono rappresentate da ben tre regolamenti comunitari: il calcolo del numero di ore lavorate secondo il regolamento europeo sulle statistiche congiunturali sulle imprese (STS, Reg. CE n. 1165/98); il calcolo dei conti nazionali trimestrali sull'occupazione espressa in ore lavorate (come richiesto dalla modifica del Reg. CE n. 2223/96); il calcolo del costo del lavoro per ora lavorata (come richiesto dal regolamento sull'indice del costo del lavoro trimestrale, LCI, Reg. CE n. 450/2003).
Nello specifico, l'indagine trimestrale sui posti vacanti e le ore lavorate raccoglie informazioni su tre famiglie di variabili, che caratterizzano le sezioni del questionario: le posizioni lavorative occupate, i posti vacanti e le ore lavorate e retribuite nell'impresa. Le posizioni occupate sono richieste in due istanti del tempo: la fine del trimestre precedente a quello di rilevazione e la fine del trimestre corrente. Inoltre, l'indagine richiede la quantità di posizioni occupate create nel trimestre (entrate) e di quelle distrutte (uscite). Per quanto riguarda i posti vacanti, l'indagine ne richiede il numero alla fine del trimestre di riferimento. Nella sezione relativa alle ore le principali variabili raccolte sono il numero di ore lavorate ordinarie, quelle straordinarie e le ore non lavorate ma retribuite dal datore di lavoro. Tutte le variabili sono raccolte distintamente per impiegati e quadri da un lato ed operai ed apprendisti dall'altro.
La popolazione di riferimento dell'indagine è costituita dall'insieme di imprese con almeno 10 dipendenti appartenenti all'industria e ai servizi distributivi e alle imprese (sezioni C-K della classificazione ATECO 2002).
Il campione teorico è costituito da circa 11.000 imprese, estratto secondo un disegno campionario stratificato dove gli strati sono definiti da attività economica, classe dimensionale e ripartizione geografica. Gli strati delle imprese con almeno 500 dipendenti sono campionati in maniera esaustiva (*take-all strata*). Il campione delle imprese non autorappresentative subisce una rotazione di circa un terzo ogni quarto trimestre dell'anno.
I dati sono raccolti attraverso due principali modalità di rilevazione, CATI e Web. A questi si aggiungono i questionari che sono ricevuti via Fax o posta ordinaria e che sono inseriti, prevalentemente, dai rilevatori CATI . La composizione della raccolta varia da trimestre a trimestre principalmente in relazione alla rotazione del campione: nel primo trimestre dopo la rotazione le

[53] L'ideazione e la stesura del presente documento è frutto del lavoro congiunto di tutti gli autori. Tuttavia è possibile attribuire il paragrafo 2 a M. Sorrentino e C. Baldi, il paragrafo 3 a D. Bellisai e S. Fivizzani, il paragrafo 4 a S. Fivizzani, i paragrafi 5.1 e 5.4 a D. Bellisai, il 5.2 e 5.5 a C. Baldi, il 5.3 e 5.6 a M. Sorrentino.
Il documento si basa sul lavoro di analisi, progettazione e programmazione informatica, oltre che degli autori, di Annalisa Lucarelli. Roberto Gismondi ha fornito una consulenza pressoché continua nella fase di definizione delle procedure di C&C ed a lui si devono puntuali suggerimenti e commenti su una precedente bozza. Il lavoro si è avvalso anche delle discussioni con Gian Paolo Oneto e Leonello Tronti. La cura editoriale della pubblicazione è stata prestata da Antonella Pietrantoni. A tutti loro va il sentito ringraziamento degli autori.
[54] I *PEEI* sono un insieme di indicatori, identificati da istituzioni ed altri utenti chiave, di primaria importanza per la conduzione della politica monetaria ed economica europea.

imprese appena entrate nel campione sono chiamate a rispondere tramite CATI ed è quindi massima la quota di raccolta CATI. Nei trimestri successivi le imprese possono scegliere di modificare la modalità di risposta e conseguentemente cresce la quota raccolta tramite Web. Per dare una misura della composizione della raccolta, nel terzo trimestre 2006, quindi nell'ultimo prima di una nuova rotazione, circa il 74% delle risposte è stato raccolto tramite CATI ed il 22% tramite Web.
Attualmente, la diffusione dei dati dell'indagine è limitata alla trasmissione trimestrale ad Eurostat delle stime sul numero di posti vacanti e sulle posizioni occupate per sezione di attività economica.
Lo scopo del presente lavoro è quello di documentare da un lato le procedure messe in atto per massimizzare i tassi di risposta e prevenire errori nella compilazione del questionario e, dall'altro, le procedure di controllo e correzione implementate sui dati raccolti. Dato l'attuale stadio di sviluppo dell'indagine, il lavoro si concentra principalmente sui posti vacanti e sulle variabili relative alle posizioni occupate. Alcuni dei metodi sono ancora provvisori e la loro integrazione nel processo regolare di produzione è attualmente in fase di test. La struttura del documento è la seguente. Nel paragrafo 2 è illustrata la strategia complessiva del processo di controllo e correzione. Nel paragrafo 3, dopo una disamina dei parametri obiettivo e delle caratteristiche della variabile sui posti vacanti, si discute delle loro implicazioni per il processo di controllo e correzione di questa variabile. Nel paragrafo 4 si affrontano i metodi adottati nella fase di raccolta dell'indagine per massimizzare la risposta e prevenire gli errori. Infine, nel paragrafo 5 si analizzano in dettaglio tutte le fasi di controllo e correzione sui dati raccolti. Nel paragrafo 6 si tracciano le conclusioni.

2. La strategia del processo di controllo e correzione dell'indagine

Il processo di controllo e correzione di VELA, che coinvolge tutte le fasi della rilevazione, si fonda logicamente su tre pilastri: l'organizzazione della raccolta dati, l'integrazione di altre fonti statistiche, e il processo di controllo e correzione sui dati raccolti.
Fin dall'organizzazione e dallo svolgimento della raccolta di dati sono implementate una serie di misure e procedure volte a prevenire errori e mancate risposte. La stessa scelta della principale modalità di rilevazione dati è ricaduta sulla CATI, in quanto tale modalità fornisce una serie di garanzie sulla massimizzazione della risposta e sulla prevenzione degli errori. In particolare, la CATI consente di minimizzare gli errori connessi:

- al rispondente: il contatto diretto con l'impresa permette di identificare correttamente uno o più referenti all'interno dell'impresa per la raccolta e la comunicazione dei dati;
- alla misurazione dei dati quantitativi, attraverso i controlli di coerenza implementati nel software CATI;
- alla corretta interpretazione delle variabili, soprattutto nel caso del posto vacante che può non essere di comune conoscenza e i cui dati non sono sempre presenti nei sistemi informativi delle imprese.

Per consentire che la rilevazione funzioni al meglio e la CATI sviluppi le sue potenzialità, una serie di procedure dirigono ed accompagnano tutta la durata della rilevazione. Tali procedure, spiegate in dettaglio nel paragrafo 4, sinteticamente consistono:

1) nel *data cleaning*, prima di ogni occasione di indagine, delle informazioni anagrafiche contenute nel database Oracle della rilevazione;
2) nella preparazione della rilevazione CATI (inclusa la formazione degli intervistatori e la creazione di gruppi specifici di rilevatori dedicati alle imprese più grandi o potenzialmente complicate);
3) nel monitoraggio della rilevazione;
4) nei solleciti e *follow-up* delle imprese più grandi.

Il secondo pilastro della strategia di controllo e correzione consiste nella integrazione dei dati di VELA con quelli di altre due rilevazioni ISTAT: quella mensile su occupazione, orari di lavoro, retribuzioni e costo del lavoro nelle grandi imprese (nel seguito indicata con la sigla GI) e OROS (Occupazione, Retribuzioni e Oneri Sociali).

I principi fondanti di questa integrazione, e le conseguenti modalità operative, consistono, da un lato, nel massimizzare la coerenza degli indicatori prodotti dalle tre rilevazioni e, dall'altro, nello sfruttare appieno le qualità delle singole indagini in un'ottica di specializzazione e divisione del lavoro. Per comprendere meglio la logica di questa operazione è utile descrivere brevemente le due indagini con i cui dati vengono integrati quelli di VELA.

GI rileva mensilmente presso un panel di circa 1.100 imprese (che avevano almeno 500 dipendenti in media nell'anno base 2005 e sono classificate nelle sezioni C-K dell'Ateco 2002) il numero delle posizioni occupate a fine mese, gli entrati e gli usciti, gli orari di lavoro (con le stesse definizioni di VELA[55]), e variabili relative a retribuzioni e costo del lavoro.

La rilevazione OROS, invece, si basa sulla quasi totalità dei modelli DM10 che mensilmente le imprese sono tenute a compilare e trasmettere all'INPS, per la dichiarazione dei contributi obbligatori. Integrando le informazioni contenute in questi modelli con quelle raccolte da GI per le imprese del panel, OROS produce indicatori trimestrali sulle retribuzioni, il costo del lavoro e le posizioni lavorative occupate per le sezioni C-K dell'Ateco 2002.

La variabile misurata da OROS sulle posizioni lavorative occupate è la media, sui tre mesi del trimestre, del numero di dipendenti a cui in ogni mese è stata retribuita almeno un'ora di lavoro.

La definizione delle posizioni occupate di OROS differisce dunque da quelle di VELA e GI, che misurano il numero di posizioni occupate all'inizio ed alla fine del trimestre o del mese. Tuttavia, una serie di analisi hanno mostrato che, in generale, il dato calcolabile come media sul trimestre di quelli rilevati da VELA per inizio e fine del trimestre non differisce sostanzialmente da quello calcolato come appena indicato da OROS.

Le informazioni raccolte da GI e OROS sono utilizzate in molte fasi del trattamento dei dati di VELA, e in particolare:

- per le imprese coinvolte anche nella rilevazione GI, le informazioni di questa rilevazione sono utilizzate per:
    - attribuire l'attività economica prevalente e il codice Ateco 2002
    - controllare i dati sulle posizioni occupate
    - imputare dati mancanti sulle posizioni occupate
- per le imprese non coinvolte nella rilevazione GI, le informazioni di OROS sono utilizzate:
    - per definire se un'impresa sia o meno attiva
    - per attribuire l'attività economica prevalente e il codice Ateco 2002
    - per controllare i dati sulle posizioni occupate
    - per imputare dati mancanti sulle posizioni occupate
    - come universo di riferimento.

In questo modo, si garantisce che:

- le stime di VELA dei totali per sezione delle posizioni lavorative occupate a fine trimestre riproducano i dati di OROS relativi alle imprese con almeno 10 dipendenti in ogni trimestre, al netto degli effetti delle differenze nelle definizioni delle variabili[56].
- se si considerano le sole imprese coinvolte in GI, le stime dei totali per sezione delle posizioni lavorative occupate a fine trimestre su queste imprese sulla base di VELA coincidano con quelle producibili da GI per la fine dell'ultimo mese del trimestre.

Perciò, le stime prodotte da VELA che utilizzano le posizioni occupate (ossia il tasso di posti vacanti e, in futuro, le ore lavorate pro-capite) sono basate su un denominatore coerente con le stime prodotte da GI e OROS.

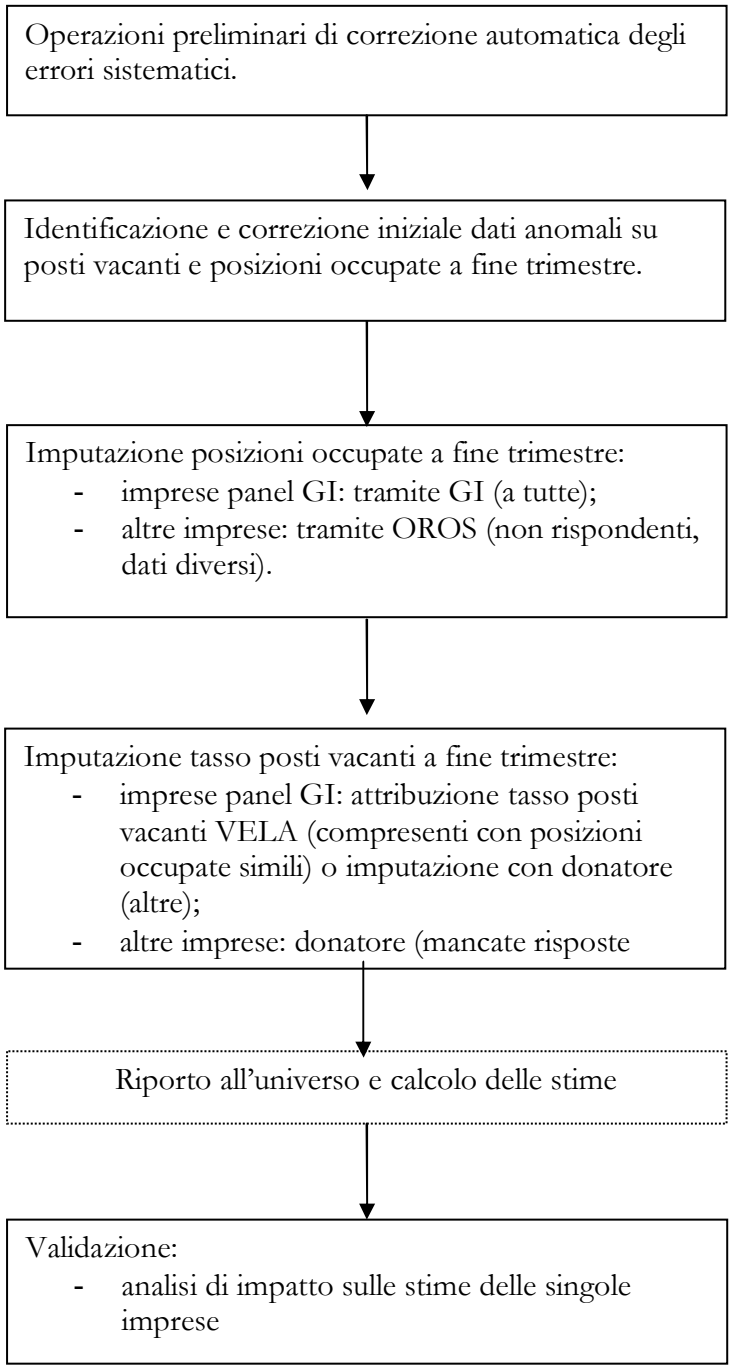
Il terzo pilastro logico del processo è la fase di controllo e correzione sui dati raccolti, il cui fulcro si colloca nell'imputazione delle posizioni occupate e dei posti vacanti. Come si è appena accennato, il

[55] Al fine di contenere il carico statistico sulle unità campionarie e i costi e le risorse necessari alla raccolta e al trattamento dei dati, le imprese coinvolte sia in VELA che in GI possono non fornire a VELA i dati sugli orari di lavoro.

[56] È da notare, a questo proposito, che l'utilizzo di OROS come universo di riferimento invece che la popolazione definita dall'ultima versione disponibile dell'archivio ASIA, assicura la proprietà che le stime di VELA siano ottenute mediante riporto ad una popolazione che si riferisce al medesimo trimestre delle stime.

controllo e correzione delle posizioni occupate e delle variabili strettamente connesse, avviene in gran parte integrando i dati di VELA con GI ed OROS. In questa operazione, infatti, i dati di GI sulle posizioni occupate, che sono già validati ed integrati per le mancate risposte nel processo di controllo e correzione dell'indagine, sono usati al posto dei dati raccolti da VELA. Questa scelta ha

Figura 1: *La struttura del processo di controllo e correzione sui dati raccolti nell'indagine trimestrale sui posti vacanti e le ore lavorate*



almeno quattro vantaggi: in primo luogo permette di avere un numero di rispondenti effettivi sulle grandi imprese maggiore di quello assicurato da VELA; in secondo luogo le *wave non response* di GI sono imputate con procedure ormai testate e stabili, con l'implicazione che non è stato necessario sviluppare procedure ridondanti sui dati di VELA riguardo alle posizioni occupate; in terzo luogo le singole imprese sono seguite nel tempo da revisori esperti che diventano profondi conoscitori delle caratteristiche delle unità di cui si occupano e svolgono un'attenta attività di *recall* per verificare i dati trasmessi che sembrano anomali rispetto alle serie storiche delle imprese. Come conseguenza, nell'ambito di GI è possibile identificare presto anche le trasformazioni societarie cui le imprese di grandi dimensioni sono soggette di frequente e considerarle in modo adeguato, un risultato molto rilevante per l'accuratezza delle stime dei parametri d'interesse.
Infine, questo trattamento delle imprese del panel di GI è conforme a quello realizzato nell'ambito della rilevazione OROS. Dunque è necessario per soddisfare l'obiettivo di produrre tramite VELA delle stime delle posizioni lavorative occupate (da usare al denominatore nel tasso di posti vacanti e nelle ore lavorate pro capite) coerenti con quelle di OROS per le sole imprese con almeno 10 dipendenti.
Per quanto riguarda i dati delle imprese appartenenti al campione di VELA ma non alla rilevazione GI, essi vengono integrati con le informazioni desunte da OROS. In questo caso la definizione della variabile posizioni occupate è differente da quella di VELA sulle posizioni a fine trimestre, ma empiricamente molto simile alla variabile posizioni occupate medie nel trimestre. Questa caratteristica ha consentito, da un lato, la correzione dei dati di VELA sulle stesse imprese, dall'altro, l'imputazione delle mancate risposte. Il vantaggio evidente di questa scelta risiede nel fatto che, data la copertura censuaria di OROS, è possibile ricostruire le posizioni occupate di tutte le mancate risposte di VELA.
Al termine delle procedure di integrazione dei dati sulle posizioni occupate, il campione di VELA risente solo delle mancate risposte parziali sui posti vacanti. Le scelte sull'imputazione dei posti vacanti, e più in generale sul controllo e correzione di questa variabile, si basano su una serie di caratteristiche dei parametri da stimare e della variabile rilevata ed è per tale ragione che a questo tema è dedicato tutto il paragrafo 3. L'imputazione delle posizioni occupate e dei posti vacanti rappresenta la fase centrale del processo di controllo e correzione sui dati raccolti. A completare il quadro vi sono una serie di procedure per il controllo degli errori sistematici, per l'individuazione e correzione dei dati outlier e influenti. L'intero processo di controllo e correzione sui dati raccolti è rappresentato sinteticamente nella Figura 1.

3. Definizioni, parametri obiettivo e caratteristiche delle variabili sui posti vacanti

Due tra i principali parametri obiettivo dell'indagine sono il numero di posti vacanti e il tasso di posti vacanti. Un posto vacante, nella definizione adottata dall'indagine, ed armonizzata a livello internazionale, è ogni posto di lavoro (di nuova creazione, non occupato o che sta per divenire non occupato) per cui il datore di lavoro ha compiuto azioni concrete di ricerca rivolte all'esterno dell'impresa miranti all'assunzione di un candidato idoneo, ed è pronto a compierne altre se necessario. Il tasso di posti vacanti è il rapporto percentuale fra posti vacanti e la somma fra questi e le posizioni lavorative occupate. In termini economici, il tasso di posti vacanti è una misura della domanda di lavoro insoddisfatta specularmente al tasso di disoccupazione che, definito in maniera simmetrica, è una misura dell'offerta di lavoro insoddisfatta. Accanto al numero ed al tasso di posti vacanti, un altro parametro obiettivo, la cui utilità è emersa nell'esperienza concreta della rilevazione, è la quota di imprese con almeno un posto vacante. Tale indicatore fornisce informazioni sulla diffusione della ricerca di personale che aiutano ad interpretare la dinamica della domanda di lavoro. Si pensi, come caso estremo, ad un trimestre in cui una sola impresa, di grandi dimensioni, aumenti notevolmente la ricerca del personale, mentre le altre si comportino esattamente come il trimestre precedente. Risulta chiaro che la stabilità della quota di imprese con posti vacanti qualifica l'aumento del tasso come un fenomeno scarsamente diffuso nel sistema delle imprese.
La variabile posti vacanti presenta alcune caratteristiche che la differenziano da molte altre variabili rilevate in indagini presso le imprese. Innanzitutto può essere una variabile difficile da rilevare, in

quanto si riferisce ad un fenomeno che spesso non viene registrato nei sistemi informativi delle imprese (a differenza di quanto accade per le variabili occupazionali e orarie che vengono registrate anche per fini fiscali e contributivi), o viene registrato in modo meno formalizzato. Si pensi, ad esempio, ai casi in cui la ricerca di personale avviene esclusivamente, o prevalentemente, attraverso passaparola fra datori di lavoro. La limitata formalizzazione della registrazione implica che possa essere difficile ricostruire anche la data in cui la ricerca ha effettivamente avuto inizio, ossia quando si è verificata la prima azione concreta rivolta all'esterno dell'impresa. Inoltre, minore è la formalizzazione della registrazione maggiori possono essere le difficoltà di rilevazione al crescere della distanza fra il momento a cui si riferisce la variabile che si intende rilevare e il momento in cui è rilevata.
Oltre a ciò, il concetto di posto vacante può essere di non semplice interpretazione e comprensione da parte delle imprese. Evidenza di queste difficoltà si ha, ad esempio, nei casi (non frequenti) in cui il numero di posti vacanti a fine trimestre fornito è uguale al numero di dipendenti entrati nel trimestre, o quando l'impresa dichiara di non effettuare ricerche di personale all'esterno di essa sebbene stia esaminando i curriculum ricevuti, che ora risiedono in banche dati interne[57].
Un'altra caratteristica deriva dal fatto che la durata dei posti vacanti, ovvero il tempo in cui i posti rimangono vacanti prima di essere riempiti o cancellati, può essere estremamente variabile in quanto dipende da un elevato numero di fattori quali, ad esempio, il tipo di professione a cui si riferiscono, le politiche dell'impresa nel reclutamento di nuovi lavoratori e la situazione congiunturale del mercato del lavoro,. Questa osservazione implica che alcuni posti vacanti non vengano rilevati dall'indagine perché nascono e terminano all'interno di un medesimo trimestre, prima dell'ultimo giorno del trimestre a cui si riferisce il dato misurato.
Dal punto di vista empirico, la distribuzione della variabile posti vacanti è fortemente concentrata sul valore zero, con una coda destra molto breve. La figura 2, che rappresenta la distribuzione dei posti vacanti per il terzo trimestre 2006, mostra che la percentuale di imprese che rispondono di non avere posti vacanti è di poco inferiore all'85% mentre la percentuale di imprese con più di 10 posti vacanti è solo del 2% circa.
Un'altra caratteristica importante, e problematica al tempo stesso, della variabile posti vacanti misurata sulla singola impresa è data dalle sue deboli relazioni a livello empirico con altre variabili rilevate dall'indagine o ricavabili da altre fonti. Fanno, in parte, eccezione le relazioni tra variabili correlate ai posti vacanti e la dimensione di impresa (vedi tavola 1). Ad esempio, si osserva empiricamente in tutti i trimestri che il tasso di posti vacanti medio nelle imprese con posti vacanti decresce all'aumentare della dimensione di impresa[58].
Le caratteristiche dei parametri obiettivo da stimare insieme a quelle definitorie, di misurazione ed empiriche della variabile posti vacanti hanno condizionato ed informato le scelte sui metodi di controllo e correzione.
Sul fronte dell'analisi dei dati anomali, l'addensarsi della distribuzione di questa variabile sullo zero rende praticamente impossibile l'identificazione di effettive osservazioni outlier tra le imprese che dichiarano zero posti vacanti. Ma la presenza di una relazione, seppur debole, tra posti vacanti ed occupazione ha permesso di elaborare metodi robusti per l'individuazione di outlier nella coda destra della distribuzione condizionale alla dimensione di impresa.

57 Un ultimo fenomeno che rende i posti vacanti difficili da rilevare è legato al fatto che, in alcuni casi, le imprese sono riluttanti a fornire informazioni sulla ricerca di personale, principalmente per una non chiara comprensione della definizione e delle norme sulla protezione dei dati personali che si applicano ai dati da loro forniti, il che le induce a temere che la trasmissione all'Istat di queste informazioni possa causare difficoltà con le organizzazioni sindacali a cui fanno capo i lavoratori. Ciò accade soprattutto in alcune imprese molto grandi. Per i succitati motivi, una grande cura è stata ed è, quindi, riposta nella formazione degli intervistatori particolarmente riguardo alla comprensione del concetto di posto vacante.

58 La relazione non dipende semplicemente dalla presenza del numero di posizioni occupate al denominatore del tasso di posti vacanti.

Figura 2. *Distribuzione dei posti vacanti nel terzo trimestre 2006*

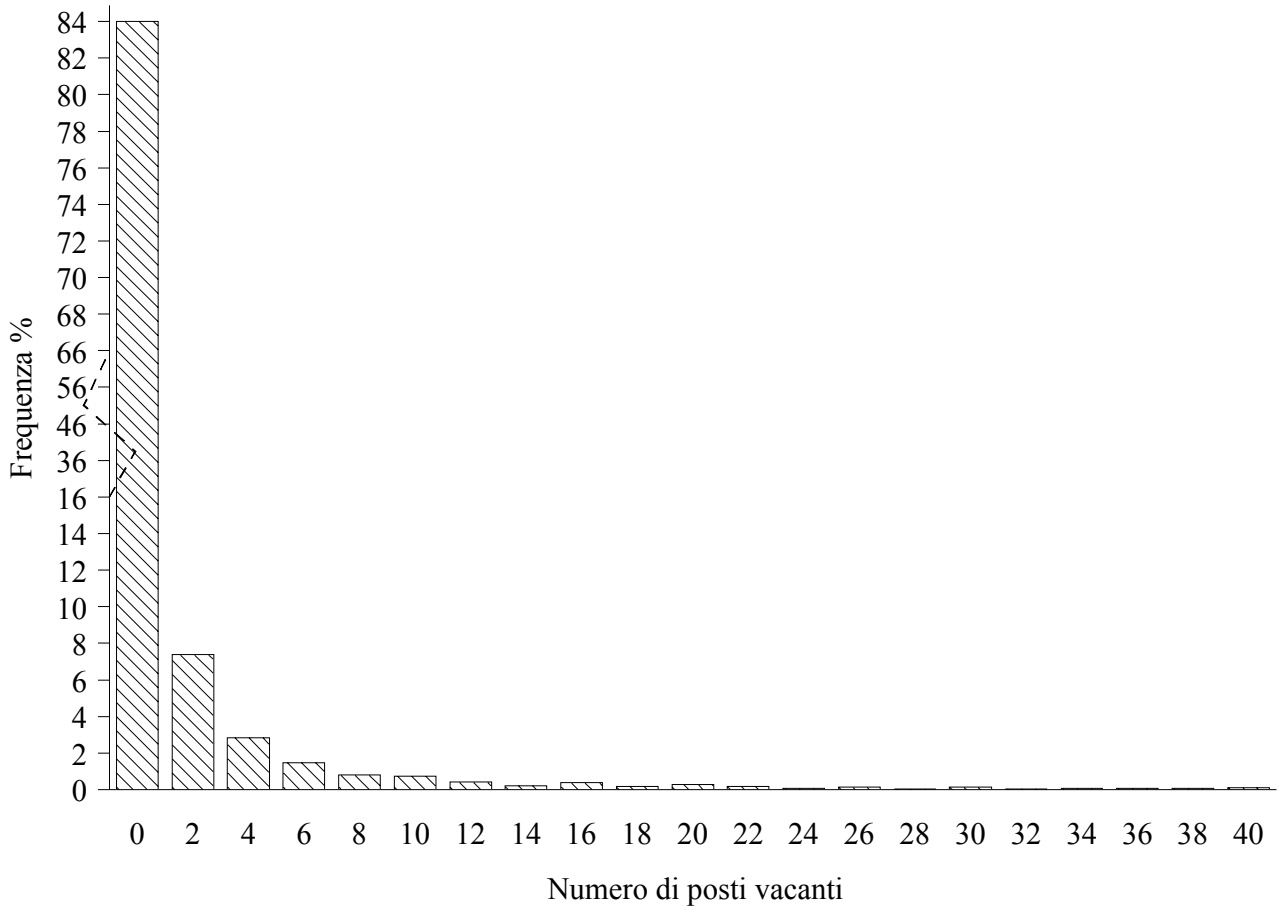


Tavola 1: *Relazione tra posti vacanti e dimensione di impresa (in termini di numero di dipendenti, esclusi i dirigenti) – terzo trimestre 2006*

Dimensione di Impresa (n. dipendenti)	Numero medio di posti vacanti per impresa	Quota di imprese con posti vacanti %	Tasso di posti vacanti (TPV) %	TPV nelle imprese con posti vacanti %
Fino a 20	0,1	7,1	0,95	13,3
20-100	0,3	12,0	0,74	6,1
100-500	2,1	26,7	1,16	4,4
Oltre 500	17,4	45,8	0,90	2,0
Totale	**2,4**	**16,5**	**0,87**	**5,3**

Sul fronte dell'imputazione dei dati mancanti, invece, è interessante notare come la necessità di produrre una stima affidabile della quota di imprese con posti vacanti implichi che non tutti i metodi di imputazione dei posti vacanti siano accettabili. In particolare, dall'insieme dei possibili metodi è stato necessario escludere quelli che imputano medie (o medie condizionali) che finirebbero per distorcere verso l'altro la stima della quota delle imprese con posti vacanti.

D'altro canto, anche le difficoltà nello specificare modelli di regressione parametrici ha reso non percorribile la strada di imputare le mancate risposte sui posti vacanti con un metodo a due stadi, in cui nel primo stadio si usi un modello logistico per stimare la presenza di posti vacanti e nel secondo stadio si usi un modello lineare per stimare il numero di posti vacanti per le sole imprese a cui è stata attribuita la presenza di posti vacanti al primo stadio.
Infine, la concentrazione della variabile sullo zero e la durata variabile dei posti vacanti fanno sì che, a livello d'impresa, ed escludendo quei settori caratterizzati da forte stagionalità, sia molto difficile specificare modelli autoregressivi che permettano di distinguere gli "zeri" dai valori positivi.
L'analisi condotta, e riassunta brevemente qui sopra, ha portato a scegliere come metodo di imputazione uno basato sulla donazione che consente di non distorcere la distribuzione dei posti vacanti con la conseguenza di poter stimare parametri non medi come la quota di imprese con posti vacanti. Inoltre, il metodo del donatore, essendo non parametrico, impone meno restrizioni sulla forma delle relazioni tra variabili. Infine, per tenere conto di una possibile persistenza dei posti vacanti, tra le variabili di matching della donazione, vengono usate, ove possibile, informazioni sui posti vacanti dei trimestri precedenti.

4. Gli strumenti primari per garantire la qualità dell'indagine: il processo di rilevazione dei dati ed il suo monitoraggio

4.1 I controlli e le operazioni che precedono la raccolta

Per far sì che lo strumento CATI, e più in generale la strategia di rilevazione, svolga al meglio la sua funzione di prevenzione della mancata risposta e dell'errore, tutto il processo di rilevazione viene organizzato a partire dalla fase di controllo che precede la raccolta fino alla selezione degli intervistatori, alla formazione teorica e pratica, al monitoraggio e alla parte tecnico-informatica.
Prima di ogni rilevazione vengono svolte una serie di operazioni di controllo e di pulizia delle informazioni anagrafiche che vengono aggiornate utilizzando i dati dell'ultimo trimestre. Questa serie di operazioni ha due obiettivi. In primo luogo, il controllo delle informazioni anagrafiche (soprattutto su eventuali modifiche della ragione sociale e dell'attività economica segnalate dalle imprese) è diretto a verificare l'appartenenza dei rispondenti al campo di osservazione e eventualmente ad escludere unità non più eleggibili perché cessate, fuori target e così via. Questo lavoro è supportato da un lato dalle note esplicative inserite dagli intervistatori in tutti i casi in cui l'intervista non viene conclusa o il contatto ha un esito negativo definitivo perché l'impresa risulta essere fuori target, cessata, in crisi, in liquidazione; e dall'altro utilizzando le informazioni dell'archivio ASIA per verificare ulteriormente quanto dichiarato dalle imprese. In secondo luogo, il controllo dei dati forniti per il trimestre precedente, consente di ottimizzare la gestione dei contatti e delle risposte. A questo fine l'individuazione e/o l'aggiornamento dei dati sui referenti dell'indagine presso le imprese ed i relativi recapiti permette di migliorare l'efficacia delle spedizioni utilizzando le informazioni più aggiornate. Inoltre la lista delle imprese da contattare viene suddivisa per modalità di contatto e di risposta (fax o e-mail, CATI o Web), utilizzando, in particolare l'ultima domanda del questionario che chiede espressamente all'impresa la modalità di contatto e di trasmissione del questionario preferita.
Nell'ambito delle operazioni che precedono la raccolta CATI, in sede di formazione gli intervistatori vengono istruiti su tutti i controlli di coerenza implementati nel software e, attraverso simulazioni vengono evidenziate le problematiche che potrebbero sorgere durante il contatto con le imprese.

4.2 I controlli effettuati durante la raccolta

Nel corso della rilevazione, viene selezionato un sottogruppo di intervistatori che seguono personalmente le imprese più importanti (grandi imprese e società di fornitura di lavoro temporaneo). Il monitoraggio periodico dei rilevatori sul campo nel corso della rilevazione permette di risolvere

eventuali difficoltà che dovessero insorgere e di intervenire subito (anche modificando il software CATI se necessario).
Nel software CATI sono implementati dei controlli di coerenza finalizzati all'individuazione e la correzione dei dati errati comunicati durante la rilevazione, con l'obiettivo di diminuire la necessità di interventi nella fase successiva di controllo e correzione sui dati raccolti.
È importante sottolineare, tuttavia, che i controlli del software CATI non sono bloccanti per cui l'intervistatore può forzarli e proseguire nell'intervista. Questo significa che spesso è demandata alla sua esperienza la decisione se passare alle domande successive del questionario, oppure insistere nel chiedere di verificare ulteriormente il dato che viola un controllo.
Si tratta di controlli progettati e implementati a livello micro (il singolo questionario) finalizzati a risolvere le incongruenze logiche tra i dati (che possono essere originate sia dal rispondente che dall'intervistatore) e a verificare l'attendibilità di valori ai margini della distribuzione. Le procedure di individuazione e trattamento dell'errore a questo livello vengono implementate attraverso *edit rules* che, nel caso del software CATI, servono ad identificare i record potenzialmente in errore. Tali controlli sono essenzialmente di tre tipi:

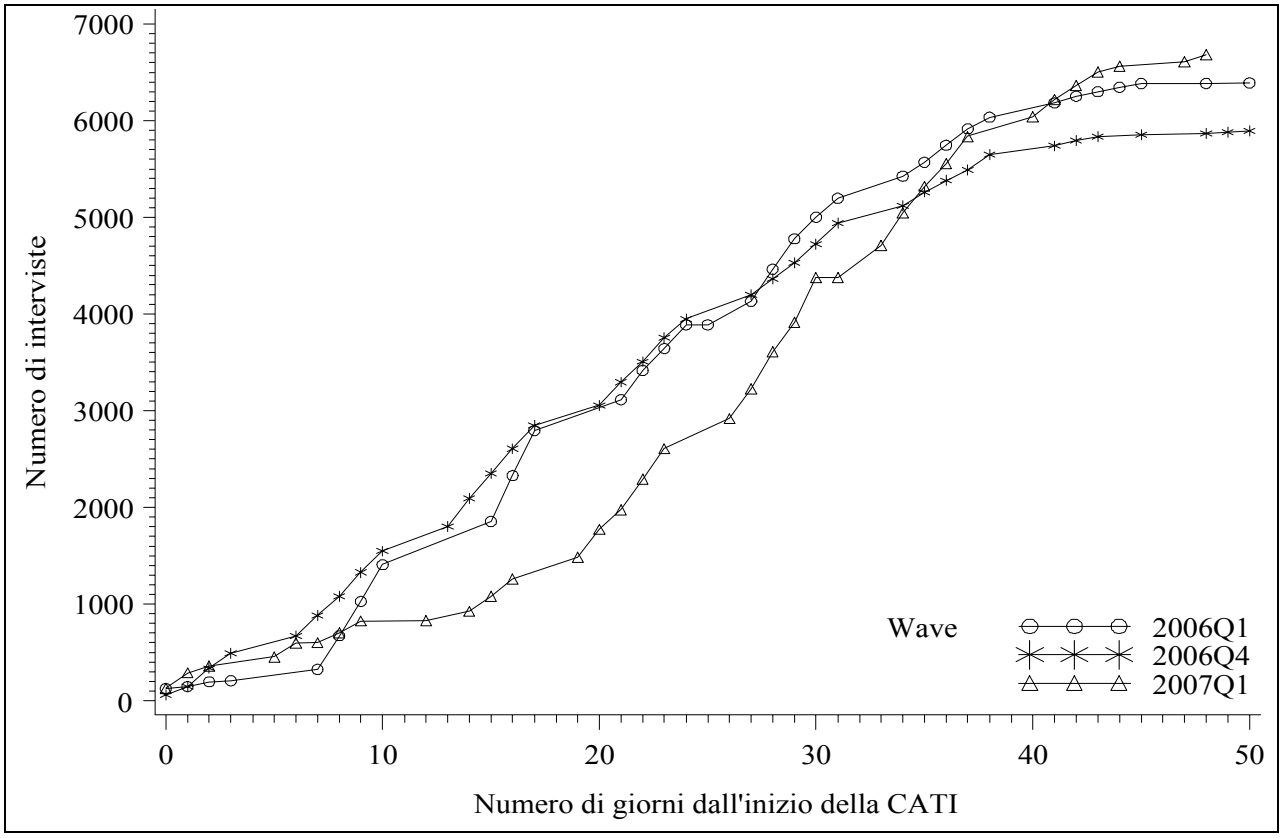
1. controlli di quadratura, principalmente per verificare la coerenza tra i dati parziali ed il totale;
2. controlli di *range*, sull'appartenenza dei valori assunti dalle variabili (in particolare sulle ore) all'intervallo di definizione prefissato per le variabili stesse (ad esempio, il numero di ore lavorate procapite in un trimestre non può superare una certa soglia);
3. controlli di tipo qualitativo sulla variabile posti di lavoro vacanti effettuati sulla base di alcune conoscenze empiriche del fenomeno oggetto d'indagine (ad esempio, se l'impresa intervistata è una società di fornitura di lavoro temporaneo che dichiara di non avere posti vacanti, ma ha flussi in entrata e uscita nel trimestre rilevanti rispetto al numero di posizioni occupate di fine trimestre, le viene chiesto di verificare il dato).

Mentre i controlli di quadratura, seguendo la terminologia utilizzata in Luzi e altri (2007), appartengono alla categoria dei *fatal* (o *hard*) *edits*, nel senso che la violazione di queste regole individua con certezza un errore, i controlli di *range* e qualitativi sono classificabili come *query* (o *soft*) *edits* in quanto costruiti su assunzioni basate sulle conoscenze sulle variabili in oggetto e sul tipo di imprese contattate.
La numerosità delle violazioni dei controlli implementati nel software CATI è monitorata attraverso un programma SAS a mano a mano che i dati sono acquisiti nel database dell'indagine.
Questo controllo assiduo ha permesso in varie occasioni di intervenire durante la rilevazione per risolvere problemi che avrebbero potuto creare distorsioni nei dati. In particolare, durante la rilevazione del primo trimestre 2007 ci si è accorti che il numero di mancate risposte alla domanda sugli entrati e gli usciti nel trimestre era troppo elevato, soprattutto nei dati delle imprese che avevano inviato il questionario via fax. Alcune delle imprese che compilano ed inviano il questionario via fax lasciano, infatti, il campo vuoto o sbarrato ad indicare la risposta "zero". Per questo motivo è stato diffuso un piccolo promemoria per tutti gli intervistatori in cui si raccomanda di inserire "zero" al posto di campi vuoti o sbarrati in un fax e, solo se viene violato qualche controllo su quei campi, contattare l'impresa per verificare se si tratti effettivamente di uno zero o di una mancata risposta. In questo modo si minimizza il costo di ulteriori contatti telefonici e la molestia statistica verso imprese che hanno già trasmesso all'Istat il questionario compilato.
Un'ulteriore procedura di controllo eseguita durante la raccolta dati consiste nel monitoraggio quotidiano delle numerosità dei questionari pervenuti sia attraverso la modalità CATI che quella Web e degli esiti dei contatti CATI. La procedura, che confronta le numerosità cumulate dall'inizio della rilevazione al giorno del controllo con le numerosità analoghe del trimestre precedente e dello stesso trimestre dell'anno precedente, ha lo scopo di seguire l'andamento della rilevazione nel tempo e di capire se eventuali problematiche sorte siano ricorrenti oppure specifiche della singola rilevazione in modo da attuare eventuali interventi correttivi volti a massimizzare la risposta.
Le principali statistiche monitorate sono:

- le frequenze cumulate di risposte CATI e Web, in modo da valutare l'andamento della raccolta;
- il tasso di risposta per strato di appartenenza delle imprese, in modo da poter controllare eventuali strati sottorappresentati e sovrarappresentati;

- le frequenze degli esiti dei contatti telefonici, in modo da misurare l'incidenza delle cause di non risposta delle imprese contattate CATI. La non risposta, infatti, può essere dovuta a rifiuto, irreperibilità dell'impresa, numero di telefono errato, oppure a non eleggibilità (perché le imprese sono cessate, inattive, fuori target, ecc.), ma può anche essere temporanea, nel caso di imprese che abbiano dichiarato di volere trasmettere diversamente i dati, o con cui ci sia un appuntamento per la comunicazione dei dati via CATI nei giorni successivi. Questi ultimi casi devono essere attentamente controllati soprattutto a ridosso della chiusura della rilevazione, affinché non diventino esiti definitivi negativi;
- il tasso di non risposta delle imprese a cui si è chiesto di rispondere via Web per modalità di risposta prevista nel trimestre precedente, in modo da comprendere se il cambiamento della modalità di contatto con le imprese (da CATI a Web, peraltro richiesto dalle imprese stesse e segnalato a tutte prima dell'inizio della rilevazione) produca un aumento delle mancate risposte totali.

Figura 3. *Frequenze cumulate di interviste CATI*



L'utilizzo periodico di questo programma ha permesso di intervenire in alcune situazioni critiche, come ad esempio nella rilevazione del primo trimestre 2007, in cui il numero delle imprese intervistate nei primi 10 giorni della rilevazione tramite CATI sembrava troppo basso rispetto ai trimestri precedenti (si veda Figura 3). Il dato è stato discusso con i responsabili della rilevazione telefonica. È così emerso un errore nel programma che gestisce i contatti telefonici, che non era predisposto in modo da contattare per la prima volta tutte le imprese durante la prima settimana di rilevazione. Un ulteriore ostacolo al raggiungimento di questo risultato era dato da un numero di intervistatori presenti nelle sessioni pomeridiane insufficiente a smaltire tutti i contatti previsti. Si è quindi chiesta la modifica del software di gestione dei contatti telefonici, e si sono formati intervistatori aggiuntivi, risolvendo così il problema.

4.3 Il monitoraggio e il follow-up delle imprese di grandissima dimensione

Al fine di prevenire le mancate risposte totali e parziali ripetute in imprese autorappresentative, che inciderebbero in maniera rilevante sulle stime, ogni trimestre viene ricontattato un certo numero di queste imprese da parte di un ricercatore Istat esperto dell'indagine.
Al momento, VELA dispone delle risorse umane per ricontattare solo le imprese non rispondenti (o parzialmente rispondenti) con più di 3.000 dipendenti. Per il *recall*, che avviene circa una settimana prima della conclusione stabilita per la rilevazione, viene utilizzata una maschera Oracle Forms appositamente creata che permette di richiamare dal database dell'indagine le informazioni anagrafiche sull'impresa, il numero delle posizioni lavorative totali, il nome, la funzione e i recapiti del referente, i trimestri e le modalità di risposta, l'esito dell'ultimo contatto e, inoltre, permette di inserire delle note esplicative. Le informazioni che si acquisiscono nei *recall* vengono registrate in una tabella Oracle, che dunque contiene traccia di tutti i contatti con le grandi imprese.
Se l'impresa non risponde da un certo numero di trimestri vengono incrociati i dati anche con le informazioni della rilevazione GI.
Il *follow-up* alle grandi imprese, cominciato nel quarto trimestre 2006, ha prodotto buoni risultati sia sul tasso di risposta che ai fini della comprensione delle difficoltà che queste imprese incontrano nel rispondere. Sono emerse in particolare alcune situazioni specifiche. Vi sono delle imprese che non hanno risposto perché è cambiata la modalità di contatto, oppure è cambiato il referente, oppure ancora perché è difficile reperire un referente all'interno di un'impresa di grandi dimensioni. In alcuni casi la non risposta è dovuta all'impossibilità da parte delle imprese di inviare il questionario nei tempi prestabiliti. La maggior parte di questi casi è facilmente risolvibile.
Vi sono, invece, casi di imprese che non rispondono da molti trimestri e risultano difficilmente contattabili. A queste per la prima volta è stato inviato per il secondo trimestre 2007 anche un sollecito postale indirizzato al responsabile del personale.
Un diverso tipo di *recall* viene effettuato sulle imprese molto grandi che rispondono all'indagine ma non forniscono dati sui posti vacanti. Sulla base dell'esperienza di rilevazione è plausibile pensare che un'impresa di grandi dimensioni cerchi personale da assumere in ogni trimestre. A queste imprese viene, quindi, chiesto di verificare se effettivamente in quel trimestre non stiano effettuando ricerche di personale all'esterno oppure se abbiano problemi a misurare il dato, siano restie a fornirlo, e così via.

5. Le procedure di controllo e correzione

5.1 Le operazioni preliminari di correzione: gli errori sistematici

L'obiettivo di questa prima fase è quello di ripulire i dati ricavati dalle risposte alle interviste CATI e dalla compilazione telematica da alcuni errori dovuti ad errata comprensione delle domande o a semplici dimenticanze da parte della persona incaricata della compilazione e/o dell'intervistatore.
La presenza di questi errori nei questionari CATi dipende essenzialmente dal fatto che i controlli di coerenza implementati nel software utilizzato dagli intervistatori non sono bloccanti. Inoltre nei questionari Web i controlli di coerenza sono praticamente assenti. La prima e più semplice operazione di controllo riguarda esclusivamente i questionari compilati per via telematica. Nel caso di ricezione di un questionario completamente vuoto tramite questa modalità, entro pochi minuti la procedura provvede automaticamente a sollecitare, tramite posta elettronica, il referente dell'impresa ad effettuare una nuova compilazione. Le altre operazioni consistono invece nel correggere alcuni errori, sulla base di semplici ipotesi.
I tipi di errore più frequenti si possono dividere in tre classi in cui:

- il valore dichiarato per una variabile "di cui" è maggiore del valore della variabile totale (ad esempio, il numero di posizioni occupate con regime orario part-time è maggiore del totale delle

posizioni occupate). In questo caso, si ipotizza che l'intervistato abbia inserito, al posto del valore della variabile totale, il valore della variabile totale al netto della variabile "di cui" (ad esempio le sole posizioni occupate full-time invece delle posizioni occupate totali). La correzione consiste nel sostituire il valore dichiarato per la variabile totale con lo stesso valore aumentato di quello dichiarato per la variabile "di cui". In particolare, questo criterio viene applicato alle posizioni occupate totali con regime part-time, a termine con regime part-time, e alle ore lavorate dagli occupati con regime part-time;

- vi sono mancate risposte parziali che possono essere interpretate come zeri (ad esempio nelle ore lavorate e nei dipendenti entrati ed usciti)
  Nel caso delle variabili relative alle ore, se le posizioni occupate a inizio e fine trimestre a cui si riferiscono le ore sono nulle, si assume che l'impresa o l'intervistatore abbiano volutamente tralasciato di compilare i campi relativi alle ore lavorate sottintendendo che il valore di ogni campo fosse zero. Nel caso di mancate risposte sui dipendenti entrati ed usciti, si assume che il loro valore sia pari a zero se il numero di posti occupati a inizio e fine trimestre è identico;
- Manca la risposta alla domanda sulla ricerca attiva o meno di candidati idonei all'assunzione. Nel questionario, la domanda sul numero di posti vacanti è preceduta da due domande filtro, che chiedono la prima se l'impresa stesse cercando personale da assumere all'ultimo giorno del trimestre di riferimento, e la seconda se alla stessa data l'impresa avesse già compiuto azioni concrete di ricerca. Nel caso di mancata risposta alla seconda domanda filtro, si assume che l'impresa abbia tralasciato di compilare il campo relativo alla ricerca attiva o meno di candidati idonei all'assunzione, giudicando sufficiente di avere dato la risposta positiva alla domanda filtro precedente. Il valore viene imputato in base alla presenza o meno di un valore dichiarato per i posti vacanti alla domanda successiva.

Nella tavola 2 viene riportata la consistenza numerica di tali tipi di violazioni per un trimestre tipo, il terzo del 2006.
Come si può vedere, l'entità delle correzioni è abbastanza elevata, soprattutto per quanto riguarda le ore lavorate, dipendenti entrati e usciti lasciati erroneamente missing, e i flag di ricerca attiva di candidati idonei all'assunzione. Questo è da imputare in larghissima parte alle imprese che rispondono con la modalità Web, in quanto i controlli di coerenza sono volutamente ridottissimi, soprattutto per quanto riguarda l'obbligatorietà di inserire valori zero, che potrebbe, per la sua frequenza, infastidire l'impresa durante la compilazione e causare l'abbandono di quest'ultima. Per quanto riguarda invece le correzioni sulle variabili "di cui", queste sono molto ridotte in quanto il controllo relativo è stato implementato sia nel software CATI che nella compilazione telematica.

Tavola 2: *Frequenza e percentuale correzioni automatiche errori pre-imputazione - terzo trimestre 2006*

Tipologia errore	Numero	% sulle risposte
Ore nulle invece che zero	1.317	17,1
Flag non valorizzati	1.336	17,4
"Di cui" maggiori dei totali	60	0,8
Flussi nulli invece che zero	675	8,8
Totale risposte	**7698**	

5.2 Il controllo dei dati outlier

Il processo di individuazione e correzione di dati outlier ed influenti si svolge in due fasi di editing selettivo. Al fine di comprendere questa scelta, è utile richiamare la distinzione tra osservazioni outlier ed osservazioni influenti, come illustrata da Luzi *e altri* (2007). Un outlier è una osservazione che è

predetta in maniera insoddisfacente da un modello statistico, laddove un'osservazione è detta influente se ha un impatto rilevante su una statistica, rilevante o di pubblicazione, dell'indagine. Sebbene sia probabile che un outlier sia anche influente, ciò non è strettamente necessario. Un esempio è quello di un'osservazione che ha un valore di una variabile lontano dal centro della distribuzione della variabile stessa, ma che, a causa di un peso campionario basso, non ha un effetto considerevole sulla statistica di interesse. Per un motivo opposto, ci possono essere osservazioni influenti che non sono considerate outlier rispetto al modello dei dati. Nella prima fase vengono identificate possibili osservazioni outlier. Nel controllo interattivo che segue, qualora si giudichi che il valore di qualche variabile sia errato, quest'ultimo viene corretto immediatamente, oppure il suo valore viene posto a missing e passato alla fase successiva di imputazione. Nella seconda fase, che si colloca dopo le procedure di integrazione con le fonti ausiliarie, imputazione delle mancate risposte e riponderazione dei dati campionari, si effettua il controllo delle unità influenti, che quindi tiene in considerazione l'effetto attribuibile al peso campionario.

Qui di seguito si descrive la procedura di identificazione e correzione dei dati outlier, mentre al paragrafo 5.5 si rimanda la spiegazione della procedura per i dati influenti.

L'individuazione dei valori outlier è basata su un metodo di regressione robusta. Come è stato osservato nella sezione 2 la probabilità che un'impresa abbia posti vacanti e il numero medio di posti vacanti per impresa crescono con la dimensione di impresa. Da ciò discende l'utilità di cercare gli outlier nella distribuzione condizionale dei posti vacanti rispetto al numero delle posizioni occupate dell'impresa. Inoltre, la distribuzione dei posti vacanti è molto variabile da settore a settore. Ad esempio, la sezione di attività economica Alberghi e Ristoranti ha un numero medio di posti vacanti molto alto nei trimestri precedenti a quelli di picco dell'attività stagionale. La scelta di uno stimatore robusto per la regressione mira invece ad evitare effetti di mascheramento di dati anomali che si possono produrre con stimatori non robusti come i minimi quadrati ordinari.

Il modello è specificato nel logaritmo del tasso di posti vacanti, previa esclusione delle osservazioni con zero posti vacanti[59]. L'esclusione di queste osservazioni, se da un lato è necessaria per stimare il modello in logaritmi, dall'altro non impone restrizioni alla procedura in quanto non c'è modo di distinguere errori tra gli "zeri" della distribuzione. Il modello, stimato separatamente per operai ed impiegati, contiene come variabili esplicative il logaritmo delle posizioni occupate alla fine del trimestre, e *dummies* per le sezioni di attività economica. Lo stimatore robusto scelto è lo stimatore MM di Yohai (1987)[60].

La procedura SAS fornisce un modo semplice per identificare gli outlier come quelle osservazioni che hanno un residuo (in modulo) troppo alto. Specificatamente, un'osservazione è definita outlier se $|r_i| > k$ dove $|r_i|$ è il valore assoluto del residuo associato all'i-esima osservazione, σ è una misura robusta di scala (che definisce l'errore di regressione), e k è un valore soglia che può essere definito dall'utente. Nella presente applicazione k è stato posto pari a 3,5, a seguito di un'analisi di sensibilità volta a minimizzare il numero di osservazioni da controllare interattivamente, pur evitando di escludere dal controllo interattivo valori che è possibile siano errati[61].

Circa 30 osservazioni vengono controllate interattivamente ogni trimestre, tramite una maschera Oracle Forms che mostra un insieme di variabili anagrafiche e strutturali dell'impresa (come la ragione sociale e

[59] Data la distribuzione dei posti vacanti mostrata precedentemente sarebbe stato preferibile usare un modello adatto a dati count come un modello di regressione Poissoniana o negativa binomiale, possibilmente tenendo anche conto della forte concentrazione della distribuzione sullo zero (zero inflation). Non è ancora disponibile una procedura nel sistema SAS che stimi modelli di dati count in maniera robusta. Quindi si è scelto di usare una regressione lineare robusta, calcolata attraverso la procedura Robustreg, stimata sul logaritmo del tasso di posti vacanti, previa esclusione delle osservazioni con zero posti vacanti.

[60] Tale stimatore garantisce stime efficienti con punto di *breakdown* elevato. Il punto o valore di *breakdown* si può definire, in maniera non formale, come la proporzione di osservazioni il cui valore deve essere "spostato ad infinito" affinché lo stimatore vada ad infinito. Si tratta quindi di una misura di resistenza agli outlier. Inoltre questo tipo di stimatori è resistente anche a osservazioni con forte effetto di leva. Con effetto di leva si indica comunemente il potenziale effetto che un'osservazione outlier sulle y ha sulla retta di regressione. Tale effetto è tanto maggiore quanto più l'osservazione è lontana dal centro della distribuzione dei dati nello spazio delle covariate. Una osservazione di questo tipo è definita un punto di leva o osservazione con forte effetto di leva.

[61] La procedura è stata analizzata rispetto ad una basata sul metodo di Hidiroglou-Berthelot (1986).

la sezione di attività economica), le informazioni sulle variabili di interesse nel trimestre corrente e nei tre trimestri precedenti, e le note dell'intervistatore e quelle rilasciate dalla persona intervistata. L'analisi interattiva consente di decidere se il dato è errato oppure, sebbene anomalo, possa essere ritenuto corretto. Per i dati affetti da errore, a meno che le informazioni mostrate dalla maschera non suggeriscano una correzione semplice ed univoca che viene applicata direttamente, il valore dei campi è posto pari a missing per essere riassegnato nella successiva fase di imputazione.
Il numero di campi corretto manualmente ogni trimestre è di circa 2-3, altrettanti sono quelli posti pari a missing. Le restanti osservazioni vengono invece giudicate corrette.

5.3 Il processo di integrazione ed imputazione delle posizioni lavorative occupate

L'obiettivo della procedura di controllo e correzione sui dati raccolti è quello di produrre un dataset di microdati validati che includa tutte le imprese contattate nel blocco di rotazione cui appartiene il trimestre di riferimento, anche tramite l'utilizzo di informazioni ausiliarie delle rilevazioni GI e OROS.
Il modo in cui questo obiettivo viene perseguito è diverso a seconda che l'impresa contattata dall'indagine VELA faccia parte o meno del panel della rilevazione GI.
Infatti, per le imprese che vengono contattate da entrambe le rilevazioni, la misura delle posizioni lavorative occupate di VELA alla fine e all'inizio del trimestre di riferimento viene sostituita con quella di GI. Le ragioni di questa scelta sono state esposte in dettaglio nel paragrafo 2.
Per le imprese contattate da VELA che non appartengono al panel di GI, invece, la procedura di controllo e correzione è più complessa.
Innanzitutto, vengono identificate fra le imprese contattate da VELA quelle che possono essere considerate attive e appartenenti alla popolazione di riferimento nel trimestre corrente, tramite abbinamento per codice fiscale con la lista di quelle rispondenti a OROS per lo stesso trimestre, con almeno 10 dipendenti e con attività classificata nelle sezioni da C a K dell'Ateco 2002. I record delle imprese che vengono ritenute non attive e/o non appartenenti alla popolazione di riferimento non vengono considerati nei passi successivi e nella produzione delle stime dei parametri di interesse.
Vengono in seguito trattati i record affetti da mancata risposta totale (la mancata risposta parziale sulle posizioni lavorative occupate è un fenomeno che in VELA non si verifica praticamente mai).
Prima di spiegare in dettaglio il tipo di trattamento utilizzato, occorre chiarire l'uso di alcuni termini. Sebbene VELA rilevi, in ogni occasione di indagine, le posizioni occupate riferite all'ultimo giorno del trimestre precedente, qui di seguito, per semplificare l'esposizione, ci riferiremo a questa variabile come "posizioni occupate all'inizio del trimestre di riferimento". Riserveremo, invece l'espressione "posizioni occupate alla fine del trimestre precedente" alla variabile raccolta nella precedente occasione di indagine e riferita alla fine del trimestre. L'imputazione delle posizioni lavorative occupate avviene facendo ricorso, ove possibile, ai dati raccolti sia da VELA per la stessa impresa per il trimestre precedente a quello di riferimento che da OROS, e in mancanza di quelli di VELA, ai soli dati raccolti da OROS. Nel primo caso, le posizioni occupate all'inizio del trimestre di riferimento vengono imputate con le posizioni occupate alla fine del trimestre precedente. Quelle per la fine del trimestre di riferimento, invece, vengono imputate applicando la variazione congiunturale delle posizioni occupate medie secondo OROS fra trimestre precedente e trimestre di riferimento alle posizioni occupate secondo VELA all'inizio del trimestre di riferimento. Ossia, in formule, rispettivamente:

$$\hat{oi}_{t,i} = of_{t-1,i} \qquad \text{e} \qquad \hat{of}_{t,i} = \hat{oi}_{t,i}\left(1+\frac{o_{t,i}^{OROS}-o_{t-1,i}^{OROS}}{o_{t-1,i}^{OROS}}\right).$$

Dove *oi* ed *of* rappresentano rispettivamente le posizioni occupate all'inizio ed alla fine del trimestre, o^{OROS} rappresenta le posizioni occupate rilevate da OROS, che, come si è detto, sono una media trimestrale di valori mensili, e il sottoscritto pedice *t* si riferisce al trimestre di rilevazione del dato Infine la notazione $\hat{x}$ indica che la variabile x è imputata.

Qualora, invece, l'impresa non abbia risposto a VELA nemmeno nel trimestre precedente, le posizioni occupate sia all'inizio che alla fine del trimestre di riferimento vengono imputate tramite le posizioni occupate medie secondo OROS. Ossia:

$$\hat{oi}_{t,i} = o_{t,i}^{OROS} \qquad \text{e} \qquad \hat{of}_{t,i} = o_{t,i}^{OROS} .$$

I dati di OROS sono usati anche per imputare le posizioni occupate nel caso in cui l'impresa sia rispondente a VELA nel trimestre di riferimento, ma abbia fornito informazioni su questa variabile sostanzialmente diverse da quanto misurato in OROS.
Per identificare questi casi si usa il metodo di Hidiroglou-Berthelot (1986), come implementato nella procedura outlier del software generalizzato BANFF sviluppato da Statistics Canada (si veda, in particolare, Statistics Canada, 2003).
Nel modo in cui è stata utilizzata per verificare i dati di VELA sulle posizioni lavorative occupate, questa procedura calcola innanzitutto il rapporto r_i, fra le posizioni occupate in media nel trimestre secondo VELA, calcolate come media tra le posizioni occupate all'inizio ed alla fine del trimestre, e le posizioni occupate medie secondo OROS nell'impresa *i*-esima, rispettivamente $\overline{o}_i$ e o_i^{OROS}.
Viene in seguito definita la variabile s_i, trasformata del rapporto r_i,

$$s_i = \begin{cases} 1 - \dfrac{r_M}{r_i} & 0 < r_i < r_M \\ \dfrac{r_i}{r_M} - 1 & r_i \geq r_M \end{cases}$$

dove r_M è la mediana della distribuzione degli r_i.
La variabile s_i ha distribuzione simmetrica e centrata sullo zero, e quindi è tale che una variazione di r_i di una certa percentuale pesa allo stesso modo su s_i indipendentemente dal suo segno.
Per ogni impresa viene quindi calcolata una statistica test e_i, che chiameremo effetto, nel seguente modo:

$$e_i = s_i \left[\max\left(\overline{o}_i, o_i^{OROS}\right) \right]^{\alpha}, \quad 0 \leq \alpha \leq 1$$

Nel presente lavoro α è stato posto uguale a 1, in modo da attribuire maggiore importanza a piccole deviazioni sulle imprese più grandi.
La distanza fra $\overline{o}_i$ e o_i^{OROS} è ritenuta troppo grande quando:

$$e_i < M - C_E d_{Q1} \quad \text{o} \quad e_i > M + C_E d_{Q3} ,$$

dove M è la mediana della distribuzione degli effetti e_i, C_E è un parametro definito dall'utente, che nel caso in questione è stato posto uguale a 30 per le imprese con non più di 100 dipendenti (secondo OROS), e a 10 per le imprese che superano questa soglia, d_{Q1} e d_{Q3} sono pari a

$$d_{Q1} = \max(M - Q1, |A \cdot M|) \qquad \text{e} \qquad d_{Q3} = \max(Q3 - M, |A \cdot M|),$$

con $Q1$ e $Q3$ rispettivamente primo e terzo quartile della distribuzione degli e_i, ed A parametro lasciato al valore di default del software di 0,05.

Questo metodo viene applicato separatamente in quattro classi dimensionali. I valori del parametro C_E sono fissati sulla base di un'analisi di sensibilità che ha tenuto conto che è probabile che parte della discrepanza fra i valori delle variabili rilevate dalle due indagini sia spiegata dalle diverse definizioni.
Per le imprese le cui posizioni lavorative occupate vengono identificate come non affidabili sulla base di questa procedura, i dati per l'inizio e la fine del trimestre di riferimento vengono calcolati riproporzionando quelli originariamente raccolti da VELA su posizioni occupate a inizio e fine trimestre con il rapporto fra le posizioni occupate medie secondo OROS e secondo VELA. Ossia:

$$\hat{oi}_i = oi_i \frac{o_i^{OROS}}{\overline{o}_i} \qquad \text{e} \qquad \hat{of}_i = of_i \frac{o_i^{OROS}}{\overline{o}_i} \quad .$$

Si può mostrare analiticamente (si veda Bellisai, 2006) che, a causa della differenza nelle definizioni, possono esistere sia situazioni in cui, per una certa impresa e in un dato trimestre, le posizioni occupate medie secondo OROS sono maggiori della media di quelle rilevate a inizio e fine trimestre da VELA, sia situazioni in cui la relazione vale nel verso opposto. Tuttavia, analiticamente non risulta semplice definire insiemi di condizioni associati a ciascuna di queste due situazioni.

Tavola 3: *Imprese da contattare da VELA e rispondenti a OROS, rispondenti e non rispondenti a VELA, con posizioni lavorative non imputate e imputate – terzo trimestre 2006*

	Numero Imprese	**Composizione percentuale**
Imprese da contattare e rispondenti a OROS	8.756	100,0
di cui rispondenti a VELA	6.413	73,2
di cui con posizioni lavorative imputate tramite OROS (per differenza sostanziale fra i dati delle due rilevazioni)	446	5,1
di cui non rispondenti a VELA	2.343	26,8
di cui con posizioni lavorative imputate tramite la risposta a VELA nel trimestre precedente e tramite OROS	649	7,4
di cui con posizioni lavorative imputate solo tramite OROS	1.694	19,3

NOTA: la tavola si riferisce alle sole imprese coinvolte nella rilevazione VELA, ma non in GI.

La scelta di usare le informazioni rilevate da OROS per integrare le posizioni occupate in VELA è comunque supportata da una serie di studi da cui risulta che in generale la differenza fra le posizioni occupate medie secondo OROS e la media di quelle rilevate a inizio e fine trimestre da VELA è vicina a zero (si vedano Bellisai, Pacini, Pennucci, 2005a e 2005b).
Come si può vedere dalla tavola 3 che descrive la situazione in un trimestre rappresentativo (il terzo del 2006) per le sole imprese coinvolte nella rilevazione VELA ma non in GI, il 73% delle imprese contattate ha risposto sia a VELA che a OROS, mentre il restante 27% non ha risposto a VELA pur potendo essere ritenuto attivo (in quanto rispondente a OROS). A poco più di un quarto delle imprese affette da mancata risposta totale sono state imputate posizioni lavorative occupate sia sulla base della risposta fornita dalle medesime imprese a VELA nel trimestre precedente che dei dati di OROS, mentre per le restanti si sono utilizzati i soli dati di OROS. Infine, solo nel 7% delle imprese rispondenti a entrambe le rilevazioni si sono riscontrati dati sostanzialmente diversi presso le due fonti.

5.4 L'imputazione del tasso di posti vacanti

Il passo successivo all'imputazione delle posizioni occupate a fine trimestre consiste nell'identificazione delle imprese alle quali deve essere imputato il tasso di posti vacanti (definito, in questo paragrafo, come il numero di posti vacanti a fine trimestre diviso per il numero di posizioni occupate a fine trimestre) e nella loro successiva imputazione.
La fase di imputazione del tasso di posti vacanti procede separatamente per le imprese della rilevazione che fanno parte anche del panel della rilevazione GI e per quelle che non ne fanno parte (da ora in poi chiamate PMI). Questo a causa delle diverse caratteristiche intrinseche e della disponibilità di differenti variabili ausiliarie per le due tipologie di imprese.
Le PMI alle quali va imputato il tasso di posti vacanti si suddividono in tre tipologie:

1. le imprese rispondenti alla rilevazione che non hanno risposto alla domanda sulla presenza o meno di posti vacanti (mancata risposta parziale)
2. le imprese rispondenti alla rilevazione che hanno fornito una risposta sul tasso di posti vacanti identificata come outlier nella fase iniziale e, per tale ragione, posta a missing
3. le imprese non rispondenti alle quali sono state precedentemente imputate le posizioni occupate

Per quanto riguarda le PMI, l'obiettivo della fase di imputazione è quello di stimare il tasso di posti vacanti per tutte le imprese contattate nei trimestri di un blocco di rotazione (dal quarto trimestre di un anno al terzo del successivo). Per quanto riguarda invece le GI, l'obiettivo è quello di stimare il tasso di posti vacanti per tutte le imprese della rilevazione GI che sono state contattate, indipendentemente dal blocco di rotazione, in quanto queste imprese fanno sempre parte del campione della rilevazione.
Le GI alle quali va imputato il tasso di posti vacanti si suddividono in cinque tipologie, di cui tre analoghe a quelle identificate per le PMI. I due casi aggiuntivi corrispondono a:

1. le imprese rispondenti che hanno fornito una risposta sul tasso di posti vacanti, ma la cui risposta sulle posizioni occupate a fine trimestre è ritenuta non compatibile con la stessa risposta alla rilevazione GI. Si ritiene in questo caso che la risposta a VELA non sia rappresentativa dell'impresa nel suo complesso e si pone a missing il valore dei posti vacanti dichiarato. In breve, il dato sul tasso dei posti vacanti di VELA è ritenuto accettabile se, in termini di posizioni occupate, l'impresa VELA è più grande di quella GI o se, pur essendo più piccola, la differenza è contenuta. In caso contrario l'impresa VELA non è ritenuta rappresentativa dell'impresa GI e il dato sul tasso dei posti vacanti non è ritenuto affidabile. Nel primo caso, il numero dei posti vacanti è ottenuto moltiplicando il tasso di posti vacanti rilevati da VELA per le posizioni occupate secondo la rilevazione GI.
   Riassumendo in formule, nel primo caso:

$$\text{se } (of_i > of_i^{GI})$$

oppure

$$0 \le \frac{(of_i^{GI} - of_i)}{of_i} \le 0,5$$

   dove con of_i^{GI} si intende il numero di posizioni occupate rilevate da GI alla fine del terzo mese del trimestre di riferimento, allora

$$\hat{pv}_i = tpv_i \cdot of_i^{GI}$$

2. le imprese rispondenti con almeno 10.000 posizioni occupate a fine trimestre che dichiarino di non avere posti vacanti e che non forniscano, mediante una nota, una spiegazione per questo dato. Per quanto detto sulla relazione tra il numero di posti vacanti e la dimensione dell'impresa, è improbabile che un'impresa di grandissime dimensioni (che ha quindi un

turnover pressoché continuo) non stia cercando personale all'esterno (si veda la sezione 2)[62]. In questo caso si pone a missing il valore dei posti vacanti dichiarato per una successiva imputazione.

Anche la strategia di imputazione, analogamente all'identificazione dei record da imputare, procede separatamente per le PMI e per le GI, differenziandosi ulteriormente nell'ambito delle GI stesse.
Si suddividono innanzitutto le imprese rispondenti alla rilevazione GI in due sottoinsiemi:

- le imprese che almeno in un trimestre dei primi nove della rilevazione avevano più di 10.000 posizioni occupate a fine trimestre (definite "grandissime")
- le imprese che in nessuno dei primi nove trimestri della rilevazione avevano più di 10.000 posizioni occupate a fine trimestre (definite "medio-grandi")

Per quanto riguarda le PMI e le imprese medio-grandi, l'imputazione del tasso di posti vacanti viene fatta mediante donatore *hot-deck* di minima distanza. Viene utilizzata a tale scopo la procedura DONORIMPUTATION del software generalizzato BANFF sviluppato da Statistics Canada (Statistics Canada, 2003).
Questa procedura utilizza la tecnica del donatore di minima distanza per trovare, per ogni record che necessita di essere imputato, il record valido che abbia le caratteristiche più simili ad esso e che permetta al record imputato di soddisfare eventuali regole di *edit* imposte dall'utente. L'imputazione viene effettuata se un tale record viene trovato. Uno dei maggiori vantaggi di questa tecnica di imputazione risiede nel fatto che tutti i campi che devono essere imputati in un record provengono dallo stesso record donatore e quindi le relazioni tra le variabili imputate vengono mantenute. Questo rende il metodo particolarmente utile in vista dell'estensione della metodologia di imputazione per ulteriori variabili rilevate da VELA, quali le ore lavorate.
La distanza tra il record *i*-esimo (ricevente) con valori $\{x_{ik}\}_{k=1}^{N}$ delle N variabili di matching (che sono le variabili sulle quali definire la distanza scelte dall'utente) e un altro record (possibile donatore) con valori $\{x_{jk}\}_{k=1}^{N}$ delle variabili di matching è data da[63]

$$D(x_i, x_j) = \min_{x_j \in X} \left[\max_k (| x_{ik} - x_{jk} |) \right]$$

dove X è il dominio dei possibili donatori. Il donatore viene quindi identificato come quel record, tra tutti i record del dominio dei possibili donatori, che minimizza la distanza tra le variabili di matching del donatore e del ricevente. La distanza così definita viene anche chiamata "distanza minimax" in quanto il donatore prescelto è quello che ha la minore differenza massima assoluta tra i valori delle variabili di matching e del ricevente, avendo dato pari peso a tutte le variabili di matching.
Per evitare che la donazione sia sistematicamente dominata dalle variabili di matching che assumono valori in intervalli di grandi dimensioni, la distanza viene calcolata dopo avere compiuto una trasformazione che induce una normalizzazione della scala sulle variabili di matching.
Nel caso dell'imputazione del tasso di posti vacanti, i possibili donatori vengono scelti all'interno di classi determinate da posizioni occupate e sezione di attività economica, e vengono utilizzati diversi insiemi di variabili di matching (a seconda della disponibilità delle variabili per le imprese riceventi e potenziali donatrici).
In particolare, per le PMI le classi di donazione risultano dall'intersezione di tre partizioni, definite da:

[62] E' stato osservato, tra l'altro, che molte imprese di grandi dimensioni tendono a rispondere che non hanno posti vacanti aperti, puntualizzando però in nota che è impossibile o molto difficile quantificare la ricerca di personale. È questa una conseguenza della difficoltà di misurazione per le imprese della variabile posti vacanti, menzionata nel paragrafo 2.

[63] La distanza è definita dalla norma L^{∞}, sullo spazio definito dall'intersezione dei domini di definizione delle variabili di matching.

- il numero di posizioni occupate a fine trimestre, sulla base delle seguenti quattro classi: meno di 50, 50-100,100-500 e oltre 500
- un flag che caratterizza se un'impresa è una società di fornitura di lavoro temporaneo o meno (tutte le imprese di questo tipo sono trattate tra le PMI, indipendentemente dalla dimensione, in quanto non fanno parte del panel della rilevazione GI)
- l'attività economica prevalente sulla base delle seguenti classi: le sezioni C, E, F, G, H, I, J, K della classificazione Ateco 2002 e quattro sottoinsiemi della sezione D.

Le variabili di matching sono:

- il numero di posizioni lavorative occupate a fine trimestre (sempre disponibile)
- il tasso di posti vacanti medio negli eventuali trimestri precedenti quello corrente e appartenenti allo stesso blocco di imputazione (ove disponibile)
- il tasso di posti vacanti medio nei trimestri del precedente blocco di rotazione (ove disponibile)
- l'età dell'impresa in anni, sulla base della data di inizio attività fornita dall'archivio OROS (sempre disponibile).

Per le imprese medio-grandi invece le classi di donazione risultano dall'intersezione di due partizioni definite da:

- il numero di posizioni occupate a fine trimestre, sulla base delle seguenti due classi: meno di 1000, 1000-10000
- aggregazioni delle sezioni di attività economica prevalente (C-E-F aggregate, D, G-H aggregate, I, J, K)

In questo caso, le variabili di matching sono:

- il numero di posizioni occupate a fine trimestre (sempre disponibile)
- il tasso di posti vacanti medio negli ultimi 3 trimestri (ove disponibile)
- il rango del tasso di crescita dell'occupazione nell'ultimo trimestre (sempre disponibile)
- il tasso di entrata medio negli ultimi 3 trimestri (sempre disponibile solo per le GI, in quanto la rilevazione OROS non fornisce questa informazione).

Per quanto riguarda invece ciascuna impresa grandissima, avendo queste caratteristiche uniche e quindi non simulabili mediante altre, l'imputazione dei valori mancanti viene effettuata con un modello basato sulla serie storica delle risposte dell'impresa stessa, nel caso in cui si abbiano a disposizione sufficienti informazioni storiche su alcune variabili dell'impresa quali tasso di posti vacanti e tasso di entrata.
Per le imprese medio-grandi per cui non sia possibile utilizzare il donatore di minima distanza, e per quelle grandissime per cui non sia applicabile il modello di serie storica, l'imputazione del tasso di posti vacanti viene effettuata assegnando la media sulla sezione di attività economica del tasso di posti vacanti calcolata sulle grandi imprese rispondenti o imputate nel trimestre.
La tavola 4 mostra, separatamente per le PMI e le GI e per il totale delle imprese, il tasso medio di posti vacanti (moltiplicato per 100) prima (*pre*) e dopo (*post*) la fase di imputazione.
Come si può notare, per le PMI, come pure per il totale delle imprese, la differenza tra il tasso di posti vacanti calcolato sulle sole rispondenti e su tutte le imprese è trascurabile. Per quanto riguarda le GI la differenza (tra le imputate e le non imputate), pur essendo di maggiore entità, non è molto rilevante.
L'ultimo passo di questa fase consiste nell'imputazione del numero di posti vacanti, il quale è ottenuto semplicemente moltiplicando il tasso di posti vacanti (effettivo o imputato) per i posti occupati a fine trimestre quali risultano dalla fase di imputazione di cui alla sezione 5.3.

Tavola 4: *Tasso medio di posti vacanti, imprese imputate e non imputate - terzo trimestre 2006*

	PMI			GI			Tutte		
	numero	Tpv post	Tpv pre	numero	Tpv post	Tpv pre	numero	Tpv post	Tpv pre
Imputate	2.381	0,873	-	343	0,990	-	2.724	0.888	-
Non imputate	6.375	0,899	0,899	762	0,897	0,897	7.137	0.898	0.898
Tutte	**8.756**	**0,891**	**0,899**	**1.105**	**0,926**	**0,897**	**9.861**	**0,896**	**0,898**

5.5 Il controllo dei dati influenti

L'ultima fase di controllo e correzione, prima della compilazione dei dati aggregati, consiste in un editing selettivo delle osservazioni influenti. A questo punto del processo tutti i dati sono stati imputati e i pesi sono stati assegnati dalla procedura di riponderazione. L'inserimento in questo punto del processo di una procedura di controllo di dati influenti rispecchia proprio la necessità di tenere conto di questi due elementi: da un lato controllare che l'insieme delle procedure di imputazione non abbia generato valori anomali e dall'altro che eventuali anomalie rimaste o lasciate nei dati non abbiano ricevuto un peso campionario troppo elevato.
L'influenza dell'*i*-esima osservazione è calcolata semplicemente come la sensibilità della stima del parametro di interesse a quella osservazione, ovvero come la differenza tra la stima che include tale osservazione e quella che la esclude. Dato che in questa fase di sviluppo dell'indagine si è interessati principalmente al tasso di posti vacanti, l'influenza è calcolata come:

$$g_i = 100 \cdot \left| \frac{\sum_{j \in S} k_j pv_j}{\sum_{j \in S} k_j of_j} - \frac{\sum_{j \in S_{(i)}} k_j pv_j}{\sum_{j \in S_{(i)}} k_j of_j} \right|$$

dove l'indice *j* indica l'osservazione *j*-esima, k_j è il suo peso di riporto all'universo, pv_j è il numero di posti vacanti, of_j è il numero di posti occupati a fine trimestre. Infine S rappresenta il campione complessivo nel trimestre, mentre $S_{(i)}$ rappresenta il campione dal quale è stata esclusa l'osservazione *i*-esima.
Un indicatore alternativo, la differenza dovuta alla osservazione *i*-esima rispetto al tasso di posti vacanti complessivo, è rappresentabile come:

$$g_i^1 = \frac{g_i}{100 \cdot \frac{\sum_{j \in S} k_j pv_j}{\sum_{j \in S} k_j of_j}}$$

Gli indicatori g_i e g_i^1 sono calcolati per ogni dominio di pubblicazione, le sezioni di attività economica.
Le osservazioni influenti sono individuate come quelle osservazioni per cui $g_i \geq 0,2$ o $g_i^1 \geq 8$. La scelta delle soglie si è basata anche in questo caso su un'analisi tesa a determinare livelli che minimizzassero il numero di osservazioni da studiare e trattare interattivamente, senza tuttavia escludere quelle che si ritiene opportuno verificare.
Questo metodo conduce ad analizzare circa 20 osservazioni al trimestre. Anche in questo caso, come per il controllo degli outlier, le osservazioni sono controllate interattivamente, basandosi anche sui

valori storici dell'impresa. Inoltre, in questa fase un'attenzione particolare è posta alle imprese che hanno subito imputazione per le quali sono controllati anche i dati precedenti a tale trattamento.
Grazie a questi controlli si è potuto individuare un possibile punto critico della procedura di imputazione. Talvolta infatti assegnare il tasso di posti vacanti rilevato dall'indagine ad una osservazione le cui posizioni lavorative occupate sono state imputate tramite OROS conduce ad un valore troppo influente sulle stime. Ciò avviene quando le posizioni occupate di OROS sono molto maggiori rispetto a quelle di VELA e il tasso di posti vacanti rilevato da VELA è elevato. Questa correzione delle posizioni occupate non comporta, infatti, che il tasso di posti vacanti rilevato da VELA venga modificato. Tuttavia, riferendosi dopo l'imputazione delle posizioni occupate ad un'impresa più grande, può diventare non più accettabile. In questo e negli altri casi in cui si riscontra un errore nei dati dovuto all'imputazione oppure un valore anomalo cui viene attribuito un peso campionario molto elevato la procedura è quella di porre a missing il dato sui posti vacanti che verrà poi re-imputato tramite la media, nella cella di imputazione, del tasso di posti vacanti di tutte le altre imprese rispondenti ed imputate.

5.6 La validazione dei dati aggregati

L'ultima fase di controllo e correzione consiste nell'analisi e nella validazione dei dati aggregati. Questa fase è, al momento in cui si scrive, in via di sperimentazione e sviluppo e qui di seguito si descrivono brevemente gli esercizi esplorativi in corso.
Gli scopi di questa fase sono di due tipi. Un primo, che classicamente si persegue in tutte le indagini congiunturali, è l'identificazione di eventuali valori anomali nei dati aggregati. Un secondo scopo consiste nell'analizzare le relazioni sia tra le dinamiche degli indicatori prodotti dall'indagine, che fra queste ultime e gli andamenti di indicatori derivanti da altre fonti. Questo secondo tipo di analisi è di particolare importanza per gli indicatori sui posti vacanti che sono ritenuti fondamentali nell'analisi del ciclo economico: in particolare si ritiene che siano indicatori anticipatori, specialmente di variabili relative al mercato del lavoro. A questo proposito è in corso di sviluppo un'analisi grafica mirante a confrontare il tasso di posti vacanti con i tassi di crescita delle posizioni occupate e, per il solo settore manifatturiero, con la produzione industriale. Il secondo tipo di confronto mira a verificare la congruenza ciclica dei posti vacanti con quella del prodotto delle imprese: in particolare, si può pensare che i posti vacanti possano crescere dopo un incremento del livello della produzione che sia giudicato come persistente dall'imprenditore. Sulla base di questa ipotesi si è cominciato a verificare se il tasso di posti vacanti nel settore manifatturiero abbia un profilo ciclico simile, ma ritardato rispetto a quello dell'indice mensile della produzione industriale. Simili confronti possono essere fatti anche fra il tasso di posti vacanti nei servizi e indicatori (deflazionati) del fatturato nei medesimi settori.
Le relazioni tra posti vacanti e tassi di crescita delle posizioni occupate (o i tassi di entrata), entrambi di fonte VELA, studiate per singola sezione di attività economica e per gli aggregati di livello superiore, sono invece di immediato interesse per comprendere se il tasso di posti vacanti possa essere considerato un *leading indicator* dell'occupazione. È ragionevole ipotizzare, infatti, che un incremento dei posti vacanti possa essere un segnale anticipatore di un successivo aumento nelle posizioni occupate.
In futuro verrà sviluppato uno studio analogo sulle eventuali relazioni tra il tasso di posti vacanti e il tasso di disoccupazione.
Un diverso tipo di analisi è, invece, svolto per confrontare gli indicatori italiani sul tasso di posti vacanti con quelli di altri Paesi europei e dell'intera Unione, al fine di capire se l'ordine di grandezza dei livelli e le dinamiche congiunturali siano o meno simili. Si deve però sottolineare che eventuali discrepanze, sia nei livelli che nel profilo ciclico, non sono di per sé immediatamente interpretabili come segnali di limitata affidabilità dei dati italiani, perché potrebbero invece riflettere differenze nella struttura e/o nel ciclo delle economie dei diversi Paesi.

6. Conclusioni

Nel presente documento sono state descritte le procedure di controllo e correzione adoperate dall'Indagine Trimestrale Istat sui Posti Vacanti e le Ore Lavorate. Come si è detto, i tre pilastri sui quali poggia la strategia del processo sono: l'organizzazione della raccolta dati, curata in dettaglio a partire dalla formazione dei rilevatori CATI fino al monitoraggio durante la rilevazione e finalizzata a massimizzare le risposte e a prevenire errori; l'integrazione con altre fonti statistiche, che consente di ridurre la molestia statistica nei confronti delle imprese, di razionalizzare il sistema dei processi produttivi e produrre indicatori coerenti con le indagini GI e OROS: il trattamento delle Grandi Imprese, in particolare, va nella direzione frequentemente discussa e auspicata di una raccolta il più possibile centralizzata dei dati e di trattamento omogeneo tra le indagini; il processo di controllo e correzione in senso stretto che, con particolare riferimento ai posti vacanti, è stato elaborato dopo un'attenta analisi dei parametri da stimare e delle caratteristiche della variabile.
Per quanto riguarda le variabili sulle posizioni occupate e sui posti vacanti lo stadio di sviluppo e verifica delle procedure è quasi completato. In un prossimo futuro, queste ultime saranno estese per includere le variabili relative alle ore lavorate. Nei piani di sviluppo dell'indagine è prevista, inoltre, una fase di analisi mirante a valutare da un lato l'impatto delle singole fasi del processo di controllo e correzione sulle stime e dall'altro l'efficacia delle procedure in circostanze di raccolta dati differenti che si verificano di *wave* in *wave*.

Bibliografia

Bellisai D. *Confronti tra occupazione trimestrale OROS e occupazione VELA media trimestrale*. Mimeo, 2006.

Bellisai D., Pacini S. e M.A. Pennucci. *Analisi preliminare sui dati OROS e VELA – 1ª parte*. Mimeo, 2005a.

Bellisai D., Pacini S. e M.A. Pennucci. *Analisi preliminare sui dati OROS e VELA – 2ª parte*. Mimeo, 2005b.

Chen C. *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*. Proceedings of the Twenty-Seventh Annual SAS Users Group, 2002.

Hidiroglou M.A. e J.M. Berthelot. "Statistical editing and imputation for periodic business surveys". *Survey Methodology,* 12, (1986): 73-83.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B. e D. Kilchman. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Lussemburgo: Technical report of the European EDIMBUS project, http://edimbus.istat.it, 2007.

Statistics Canada. *Banff – Functional Description of the Banff System for Edit and Imputation – Version 1.02*. Statistics Canada December 2003.

Yohai V.J. "High breakdown point and high efficiency robust estimates for regression". *Annals of statistics*, 15, (1987): 642-656.

Le indagini sul fatturato degli altri servizi: metodi di controllo e correzione

Alfredo Cirianni, Istat, Servizio Statistiche sull'attività dei servizi
Salvatore Coppola, Istat, Servizio Statistiche sull'attività dei servizi
Fernanda Panizon, Istat, Servizio Statistiche sull'attività dei servizi

Sommario: La strategia per il controllo e la correzione degli errori non campionari, nelle rilevazioni trimestrali di fatturato degli altri servizi è omogenea per quanto riguarda la correzione degli errori di lista, di misura e di processo, mentre è più specifica per ciascuna indagine per le fasi di imputazione delle mancate risposte e per l'individuazione ed il trattamento degli *outlier*.
L'analisi illustrata mostra come l'impatto del processo di controllo e correzione sia significativo sulle stime definitive delle variazioni tendenziali. L'imputazione delle mancate risposte con criteri prudenti sembra comprimere la stima della variazione tendenziale degli indici di fatturato. La fase dell'esclusione degli *outlier* influenti è particolarmente critica: anche poche osservazioni anomale possono avere notevole impatto sulla stima delle variazioni, per cui è necessario valutare attentamente caso per caso, evidenziando e giustificando i motivi che suggeriscono un comportamento eterogeneo dell'impresa all'interno dello strato di appartenenza e quindi determinano la sua esclusione. Altri problemi di impatto sulle stime sono legati alla soglia di accettazione dei valori anomali, la cui componente soggettiva non sembra, allo stato attuale delle conoscenze, evitabile.

Parole chiave: Indagini ripetute nel tempo, errori non campionari, imputazione, editing, outlier

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Caratteristiche generali delle indagini trimestrali del fatturato nel settore degli altri servizi[(*)(**)]

Le rilevazioni trimestrali di fatturato degli "altri servizi" (così definiti per distinguerli dal comparto delle vendite al dettaglio) producono stime delle variazioni tendenziali degli indici di fatturato per alcuni settori del terziario, secondo uno schema comune a tutti i paesi dell'Ue. L'obiettivo è quello di fornire informazioni congiunturali sull'andamento del settore.

Caratteristiche comuni delle rilevazioni:

Parametro obiettivo: il parametro obiettivo è la stima della variazione tendenziale trimestrale di fatturato per i domini di stima richiesti dal Regolamento Comunitario sulle statistiche congiunturali (n. 1165/98 e successivi emendamenti). Tale parametro è calcolato come differenza percentuale tra due numeri indice a base fissa 2000=100, espressi al tempo corrente ed allo stesso trimestre dell'anno precedente. L'indice viene calcolato col metodo del concatenamento. Il cambiamento della base, secondo Regolamento, avviene ogni cinque anni;

Popolazione obiettivo: la popolazione obiettivo è costituita dalle imprese attive nell'anno di rilevazione, che non è nota in tempo reale. L'Archivio Statistico delle Imprese Attive (ASIA) aggiornato annualmente e disponibile a circa 18 mesi dalla fine dell'anno di riferimento, è la fonte primaria per disporre delle informazioni anagrafiche, dei dati sugli addetti e sul volume d'affari indispensabili per il disegno campionario.

Variabili richieste nel questionario: le variabili richieste sono il fatturato (al netto di IVA) ed il numero di addetti nel trimestre. La prima è la variabile obiettivo della rilevazione, mentre la seconda è una variabile ausiliaria, impiegata anche nell'aggiornamento annuale della stratificazione. Viene anche richiesto all'impresa di aggiornare l'attività principale e/o secondaria e di fornire spiegazioni su eventuali variazioni molto ampie del fatturato. Questa informazione è utile per il trattamento dei valori anomali.

Modalità di raccolta dati: le imprese vengono contattate con una lettera informativa, che specifica gli obiettivi dell'indagine; alla lettera sono allegati il questionario e le istruzioni alla compilazione. Le imprese rispondono per posta, via fax o, dal 2006, via web. I dati pervenuti su carta vengono acquisiti con registrazione interna, mentre i dati arrivati via internet sono caricati direttamente nel data base di produzione.

Disegno campionario: Per tutte le indagini è selezionato un campione di imprese *panel*, che vengono cioè intervistate per diversi anni nei quattro trimestri. Periodicamente i *panel* vengono aggiornati per tenere conto dei fenomeni demografici (cessazioni e nascite) e dell'*attrition,* (imprese mai rispondenti, imprese risultate non eleggibili, imprese che avendo partecipato per più anni all'indagine chiedono di essere esonerate e ove possibile vengono sostituite).

Diffusione: il rilascio dei dati prevede diversi momenti di calcolo degli indici. A 60 giorni dal trimestre di riferimento le stime vengono fornite in forma confidenziale all'Eurostat. A 90 giorni sono pubblicate mediante comunicato stampa trimestrale come dati provvisori e a 180 giorni, sempre sul comunicato stampa, si diffondono le stime definitive. Per ogni occasione di rilascio l'insieme di dati pervenuti fino a quel momento viene sottoposto all'intero processo di controllo e correzione. Per massimizzare l'insieme delle osservazioni disponibili, la cui consistenza aumenta di giorno in giorno, si cerca di spingere il momento del calcolo degli indici il più a ridosso possibile della data di diffusione. Ovviamente le successive revisioni risentono dell'effetto delle imprese "ritardatarie". I dati pervenuti entro i 180 giorni sono utilizzati nel calcolo delle stime definitive, mentre i dati arrivati oltre tale data sono archiviati ed usati l'anno seguente, per il calcolo del denominatore degli indicatori prodotti.

Specificità delle rilevazioni

(*) Le elaborazioni statistiche sono a cura di Salvatore Coppola per il commercio all'ingrosso e di Barbara Iaconelli per manutenzione e riparazione di autoveicoli.

(**) Gli esperti di indagine dei settori oligopolistici sono Roberto Braca e Fabio Mocavini.

Le caratteristiche dei campioni (disegno, dimensione, rotazione del panel) sono diversi a seconda del settore economico considerato (Tabella 1), e derivano da studi di settore effettuati in fase di progettazione.

Per le rilevazioni del **commercio all'ingrosso** e dell'**informatica**, si è adottato un disegno campionario casuale semplice stratificato La dimensione campionaria è scelta in base all'errore di stima desiderato sulla variabile fatturato (10%) ed in base ai vincoli di risorse professionali assegnate alla rilevazioni, legate quindi al numero di imprese campione che ciascun addetto alla revisione deve seguire. La dimensione campionaria è pari a circa 8.000 unità per ingrosso ed a circa 1.800 imprese per servizi informatici [64]. Per tali rilevazioni, viene effettuato un aggiornamento annuale del campione.

Per l'indagine sulla **manutenzione e riparazione di autoveicoli**, il campione, pari a circa 2.700 imprese, è di tipo bilanciato e ragionato (approccio *model based*[65]) di tipo stratificato[66]. Le imprese con oltre 50 addetti vengono incluse sempre nel campione. La selezione della restante parte avviene a partire da circa 3000 sub-strati in cui è stato partizionato (con metodo *cluster*) l'universo delle imprese di ASIA (2001). Da ogni sub-strato viene estratta l'impresa più rappresentativa della media (del fatturato) del sub-strato stesso. Le imprese che escono dal campione vengono sostituite periodicamente a livello di sub-strato, scegliendo di volta in volta l'impresa più vicina alla media. La rotazione è trimestrale per le imprese non eleggibili (cessate o fuori campo di osservazione) ed annuale per le imprese non rispondenti per quattro trimestri nell'anno. Un *refreshment* (mediante l'estrazione di un nuovo campione con riferimento ad ASIA 2005) del panel è previsto in occasione del cambio base.

Per i settori **oligopolistici** (trasporti aerei e marittimi, telecomunicazioni, servizi postali), caratterizzati da poche imprese dominanti con elevato fatturato, è stato scelto il disegno campionario di tipo *cut off*, ossia sono state selezionate le unità più grandi fino a raggiungere la copertura dell'80-90 per cento del volume d'affari dell'intera popolazione. Il campione è composto da circa 300 imprese per trasporti marittimi, circa 80 imprese per trasporti aerei, circa 140 imprese per servizi postali, e circa 200 unità per telecomunicazioni. Il panel viene aggiornato annualmente, con la disponibilità del nuovo archivio ASIA, per tener conto del fenomeno demografico e di quello dell'*attrition* che erode la dimensione campionare originaria.

Tabella 1 - Dimensione campionaria delle indagini trimestrali sul fatturato degli altri servizi

CODICE ATECO	SETTORE ATTIVITA' ECONOMICA	IMPRESE (ASIA 2004)	CAMPIONE 2007	Tasso di campionamento
51	INGROSSO E INTERMEDIARI COMMERCIO	416.317	7.878	1,89
72	INFORMATICA	89.755	1.829	2,04
50.2	RIPARAZIONE AUTOVEICOLI	91.898	2.734	2,98
61.1 - 61.2	TRASPORTI MARITTIMI E FLUVIALI	1.489	295	19,81
62.1	TRASPORTI AEREI	281	78	27,76
64.1	SERVIZI POSTALI	1.788	141	7,89
64.2	TELECOMUNICAZIONI	1.878	204	10,86

[64] La stratificazione del commercio all'ingrosso è per classe dimensionale (1-5, 6-19, oltre 20), per attività economica (9 modalità corrispondenti ai gruppi o combinazioni di classi di attività economica) e per ripartizione territoriale (Nord, Centro e Sud-Isole). La stratificazione dell'informatica è per classe di addetti (1-19, 20-99, oltre 100) e per 6 gruppi 6 di attività economica. Le imprese con oltre 100 addetti sono censite in quanto coprono quasi la metà del fatturato dell'intera popolazione.

[65] Gismondi R. (2002) *Model-based sample selection using balanced sampling,* Rivista di statistica ufficiale, n. 3/2002, pag. 81-111

[66] La stratificazione è per categoria economica (5 modalità) e per classe di addetti (1-2, 3-5, 6-9, 10-19, oltre 20).

2. Errori non campionari delle indagini (panel) sul fatturato degli altri servizi

I dati acquisiti sono sottoposti a diverse fasi di controllo e correzione per rimuovere gli errori non campionari, sia quelli certi, identificati come mancate risposte, sia quelli probabili, individuati come valori anomali in quanto posti alle estremità della distribuzione. I metodi di controllo e correzione adottati influiscono sui risultati finali ottenuti.
Per alcune tipologie di errore[67] le indagini si avvalgono di tecniche simili per ripristinare, ove possibile, situazioni di correttezza e coerenza; di seguito si fornisce una sintetica descrizione di tali tipologie.
Errori di lista dell'archivio: la popolazione di riferimento, dalla quale vengono estratte le imprese per formare i campioni si basa sull'Archivio Statistico delle Imprese Attive (ASIA), che è disponibile a 18-24 mesi dal periodo di riferimento. Questo ritardo comporta che al momento dell'aggiornamento dei *panel* (per rotazione o *refreshment)*, per le imprese nuove entrate si possano osservare errori di lista, per esempio a causa di indirizzi errati, di cambiamenti dell'attività economica o perché nel frattempo l'impresa è cessata. Nel caso di variazione di attività economica una impresa appartenente al campione può non risultare più eleggibile e, quindi viene esclusa e sostituita appena possibile. In caso di indirizzi errati, che si constatano quando la lettera non è stata consegnata e viene restituita all'Istat, si cerca invece di recuperare gli indirizzi esatti, mediante l'uso di strumenti quali le visure camerali o i siti web delle "pagine bianche" o "pagine gialle", in modo da provare a contattare le unità nelle successive occasioni di indagine. Le imprese non eleggibili o con indirizzi errati non recuperati, insieme alle unità che non hanno mai risposto nel corso dell'anno e che non appartengono a strati censuari[68], vengono ruotate annualmente.
Errori di misura: gli errori di misura si riferiscono soprattutto alla compilazione del questionario da parte dell'impresa. Ad esempio nel passaggio da lira a euro, numerosi sono stati i casi di imprese che fornivano i dati nella vecchia unità monetaria. Oppure, in caso di livelli di fatturato elevati (milioni dei euro) le imprese, per risparmiare spazio, tendono a compilare il questionario indicando il fatturato in migliaia di euro. Altri casi di errore derivano dal mancato rispetto da parte dei rispondenti del criterio richiesto della "competenza", che le imprese sostituiscono col criterio di "cassa": questo implica che i ricavi derivanti per esempio da commesse importanti risultano erroneamente concentrati in certi trimestri (quelli dell'incasso), pur riferendosi ad attività svolte lungo un arco di tempo più esteso. Le istruzioni alla compilazione chiedono chiaramente di indicare il fatturato del trimestre, e sul modello di rilevazione sono specificati esplicitamente i tre mesi di riferimento ma alcuni rispondenti forniscono come dato trimestrale il risultato cumulato di più trimestri. Tali errori possono essere intercettati nella fase di *microediting* (controllo sulle singole osservazioni): se un potenziale errore di misura comporta una variazione tendenziale di fatturato elevata, tale cioè da superare le soglie di accettazione previste nella fase di revisione preliminare[69], si provvede a ricontattare telefonicamente l'impresa per conferme o rettifiche;
Errori di processo: sono gli errori che si possono verificare nelle diverse fasi del processo di produzione. Ad esempio con l'acquisizione dei dati via *web* si è verificato un problema sui centesimi di euro dopo la virgola decimale, che non venendo riconosciuta correttamente dal programma informatico portava il dato di fatturato e di numero di addetti ad essere moltiplicato per cento. Fax illeggibili o errori di registrazione possono produrre valori sospetti, che richiedono un accertamento presso i rispondenti.

Questo documento focalizza l'attenzione sul trattamento degli errori non campionari (integrazione delle mancate risposte e trattamento degli outlier) nelle diverse rilevazioni, con metodi adeguati al disegno campionario e alla differente dimensione campionaria.

[67] Bergdahl, et al. (1997) *Model quality report in business statistics*, pag. 82-160, pubblicato da EUROSTAT
[68] Per la loro importanza e data l'impossibilità di trovare imprese sostitutive le imprese di questo strato vengono mantenute nella lista del campione.
[69] Usualmente le imprese che hanno, in termini assoluti, una variazione tendenziale superiore al 50 per cento sono oggetto di verifica..

Secondo la definizione di Eurostat[70], nel caso di rilevazioni disegnate con campioni stratificati, l'imputazione delle mancate risposte consiste nell'attribuzione di valori stimati alle unità non rispondenti, sulla base dei valori effettivi delle unità rispondenti non anomale, appartenenti allo stesso strato. Il vantaggio dell'imputazione consiste nel fatto che viene ricostruito il campione teorico
L'esclusione dei valori anomali nell'insieme di dati osservati che concorrono all'imputazione è motivata dall'esigenza di assegnare un valore "normale" all'unità non osservata del campione teorico, un valore che non dipenda dagli estremi della distribuzione e che sia desunto da unità omogenee (per classe dimensionale e attività economica)[71]. Questo significa che è necessaria una esplorazione preliminare dei dati per identificare ed escludere i valori estremi dal processo di imputazione delle mancate risposte. Questi *outlier preliminari* vengono successivamente inclusi nell'analisi e nei calcoli.
Per determinare quali sono i veri e propri *outlier*, da escludere dal calcolo degli indici si adottano invece specifiche metodologie[72].

3. Strategia complessiva della procedura di qualità dei dati (controllo e correzione)

Per le rilevazioni trimestrali di fatturato nel settore degli altri servizi, che sono state disegnate e implementate in periodi differenti, sono stati adottati metodi diversi per il controllo e correzione dei dati, sia per la fase di imputazione delle mancate risposte, che per il trattamento degli *outlier*.
Va premesso che alla fase di imputazione delle mancate risposte si ricorre dopo aver adottato strategie e soluzioni operative per ottenere la risposta alla fonte. La politica di prevenzione della mancata risposta, si svolge, per tutte le indagini, attraverso una serie di solleciti alle imprese che non hanno ancora risposto. Nel questionario, inviato per posta alla fine del trimestre di riferimento, si chiede all'impresa di rispettare una prefissata data di scadenza per la risposta; pochi giorni dopo quella data viene effettuato un sollecito postale. Le imprese che hanno fornito in precedenza il proprio indirizzo e-mail ricevono un ulteriore eventuale sollecito per posta elettronica. Nei giorni che precedono il calcolo dell'indice si effettuano solleciti telefonici mirati alle imprese di maggiore dimensione e/o alle imprese che appartengono a strati con più elevati tassi di mancata risposta o con basse coperture.

Tabella 2 - Schema sintetico dei metodi di imputazione e di trattamento degli *outlier* nelle diverse indagini

SETTORE ATTIVITA' ECONOMICA	Pre-selezione outlier	Imputazione Mancate risposte	INFLU-ENTI strati	INFLU-ENTI imprese	OUTLIER
INGROSSO E INTERMEDIARI COMMERCIO	NO	NO	FUNZ. PUNT. MACRO,	f. p micro (box plot)	H-B
INFORMATICA	SOGLIA 50%	QUO-ZIENTE	FUNZ. PUNT. MACRO	f. p. micro (box plot)	SOGLIA 50%
RIPARAZIONE MOTOVEICOLI	REGRESSIONE + RESIDUI STUDENTIZZATI		FUNZ. PUNT. MACRO,	f.p micro (box plot)	SOGLIA 50%
OLIGOPOLISTICI: Trasporti Marittimi e Fluviali Trasporti Aerei Servizi Postali Telecomunicazioni	Solo per il gruppo delle "piccole imprese" (peso < 1%) METODO DEL QUOZIENTE				Selezione dall' ordinamento decrescente delle variazioni tendenziali

[70] EUROSTAT (2005) *Methodology of short term business statistics*, dicembre 2005
[71] *Model quality report in business statistics*, pag. 133-137, pubblicato da EUROSTAT
[72] Lee H. (1995) *Outliers in business surveys*, in Cox, Binder, Chinappa, Christianson, Colledge, Kott *Business survey methods*, John Wiley & Sons, New York, pag. 503-523

La politica di prevenzione della mancata risposta assume particolare rilevanza per le imprese dominanti appartenenti ai settori oligopolistici, per le imprese "censuarie" appartenenti al settore dell'informatica e per le imprese che hanno un "peso" rilevante nello strato di appartenenza, per cui si cerca di garantite una buona copertura (in termini di fatturato) per tutti gli strati del campione.
Anche l'identificazione degli *outlier* viene curata nel corso della raccolta dei dati, e le imprese che forniscono valori ritenuti anomali (con variazione tendenziale in valore assoluto superiore ad un soglia prefissata) vengono identificate dal software di controllo (maschere inserimento dati) e contattate per conferma ed accettazione dei dati, o per correggere eventuali errori di misura o di processo.
Le diverse indagini affrontano il problema dell'imputazione e del trattamento degli *outlier* con gli strumenti statistici ritenuti più pertinenti al tipo di indagine.

4. Metodi comuni per l'imputazione delle mancate risposte e per l'individuazione e il trattamento degli outlier

Nei paragrafi successivi, verranno analizzati i metodi utilizzati mettendo in evidenza, anche ricorrendo a qualche caso concreto a carattere esemplificativo, quali siano i "problemi aperti" legati a specifiche scelte operative.

Metodi di imputazione delle mancate risposte

I metodi di imputazione delle mancate risposte si prefiggono l'obiettivo di ricostruire il campione teorico su cui basare i calcoli degli indici trimestrali. Per le indagini che adottano campioni stratificati, i metodi di imputazione correntemente utilizzati si basano sull'ipotesi che le imprese non rispondenti si comportino allo stesso modo delle imprese rispondenti, omogenee per caratteristiche strutturali (stessa classe dimensionale e stessa attività economica), che non ricadano negli estremi della distribuzione (ipotesi MAR – *Missing At Random*).
Per le imprese dominanti appartenenti ai settori oligopolistici invece tale omogeneità può essere ritenuta valida solo per il sottoinsieme delle imprese più piccole e non per le imprese dominanti, per le quali non esiste una impresa omogenea rispondente dalla quale desumere eventuali valori mancanti da imputare.
I metodi di imputazione utilizzati per le indagini sono il "metodo del quoziente" e il "modello di regressione" semplice (senza intercetta): entrambi assegnano alle osservazioni mancanti una variazione tendenziale che corrisponde a quella media dello strato di appartenenza.

$$F_t = \frac{\bar{F}_t}{\bar{F}_{t-4}} F_{t-4} \qquad (1)$$

Se chiamiamo F_t il fatturato del trimestre corrente ed F_{t-4} il fatturato dello stesso trimestre dell'anno precedente possiamo calcolare la variazione tendenziale media come rapporto, calcolato sulle sole imprese rispondenti nelle due occasioni *t* e *t-4*, fra il fatturato medio $\bar{F}_t$ e il fatturato medio $\bar{F}_{t-4}$.
Applicando questa variazione tendenziale media alle imprese non rispondenti al tempo *t*, di cui si conosce il fatturato a *t-4* si ottiene la stima di F_t.
La variazione tendenziale media calcolata con la (1) corrisponde alla stima di β nel modello di regressione (sui rispondenti) fra il fatturato a tempo *t* e al tempo *t-4,* che viene usato in alternativa.

$$F_t = \beta F_{t-4} + \varepsilon \qquad (2)$$

Metodo di individuazione e trattamento degli outlier influenti

Nelle rilevazioni caratterizzate da campioni stratificati (commercio all'ingrosso, manutenzione e riparazione di autoveicoli e informatica) per l'individuazione ed il trattamento degli *outlier* influenti, la cui inclusione o esclusione nel calcolo degli indicatori ha un impatto molto significativo sulle stime delle variazioni tendenziali, si utilizza un metodo aggregato univariato (approccio *macroediting*). L'obiettivo del *macroediting* è l'individuazione preliminare delle unità influenti, ossia quelle che contribuiscono maggiormente alla stima delle variazioni tendenziali di fatturato e per le quali è necessario verificare se siano o meno *outlier*.
La ricerca delle osservazioni influenti avviene in due passi, basandosi sulle variazioni tendenziali calcolate dai dati "grezzi" (dopo aver effettuato l'eventuale imputazione delle mancate risposte). Il primo passo consiste nella identificazione degli strati influenti attraverso la funzione di punteggio macro FC1.

$$FC1 = VTF_S * W_S \qquad (3)$$

dove FC1 indica la funzione punteggio macro, VTF_s indica la variazione tendenziale del fatturato dello strato *s* e W_s rappresenta il peso, riferito all'anno base, dello strato nel dominio di pubblicazione.
Gli strati ritenuti influenti sono quelli che presentano il maggior valore assoluto della funzione punteggio macro. Vengono selezionate le imprese con più alto punteggio macro, in ordine decrescente fino a spiegare almeno l'80 per cento della variazione tendenziale dell'indice del dominio di pubblicazione.
Il secondo passo consiste nell'individuazione, all'interno degli strati influenti appena selezionati, delle imprese influenti. Si calcola pertanto la funzione punteggio micro FC2:

$$FC2 = VTF_i * W_{i,(t-4)} \qquad (4)$$

data dalla variazione tendenziale del fatturato della singola impresa (VTF_i) moltiplicata per il peso della singola unità nello strato al tempo *t-4* ($W_{i,\ t-4}$). Si dimostra che la sommatoria della funzione punteggio micro è esattamente uguale alla variazione tendenziale di fatturato dello strato: il metodo del *macroediting* quindi permette di concentrare l'attenzione sulle imprese più influenti, ossia su quelle che hanno il maggiore impatto sulla stima dell'indice di strato.
Una volta calcolate le FC2, si analizza la loro distribuzione. Le imprese che, in base al metodo grafico del *box plot*, ovvero in base alla funzione interquartile corrispondente, cadono agli estremi della distribuzione (e che per costruzione sono imprese che appartengono agli strati influenti) sono definite influenti. Solo per queste imprese una variazione tendenziale di fatturato elevata può determinare un *outlier* da escludere.

5. Il settore dell'ingrosso

Per l'indagine sull'ingrosso non viene effettuata alcuna procedura si imputazione delle mancate risposte. Questa scelta implica che alle unità non rispondenti viene applicata implicitamente la media campionaria (della variazione tendenziale di fatturato) delle imprese rispondenti identificate come non anomale e appartenenti allo stesso strato.
La procedura del *macroediting* viene applicata per determinare quali degli 81 strati siano influenti e quali imprese siano influenti
La ricerca degli *outlier* si avvale del metodo di *Hidiroglou-Berthelot*, che viene inserito in una procedura automatica, particolarmente vantaggiosa per la rapidità di esecuzione.

La soglia di accettazione di *Hidiroglou-Berthelot*, al di fuori della quale cadono le unità anomale, viene definita a partire da una trasformazione lineare, che rende simmetrica la "variazione tendenziale di fatturato":

$$A = (q_{0,5} - c_{\inf} * (q_{0,5} - q_{0,25}); q_{0,5} + c_{\sup} * (q_{0,75} - q_{0,5})) = (A_{\inf}; A_{\sup}) \quad (5)$$

dove $q_{0,5}$ esprime il valore mediano, $q_{0,25}$ il primo quartile e $q_{0,75}$ il terzo quartile della variabile obiettivo; "c" è un parametro arbitrario (che teoricamente potrebbe anche essere assegnato in modo differente per l'estremo inferiore e superiore). La scelta del valore del parametro "c" nell'algoritmo di *Hidiroglou-Berthelot* è lasciata alla discrezionalità dell'esperto dell'indagine e riveste una notevole importanza nella determinazione delle quota di *outlier* identificati (Panizon e Cirianni, 2007; Cirianni e Gismondi, 2006).
Nella procedura correntemente utilizzata per il commercio all'ingrosso, il parametro "c" è posto uguale a 1,5 e con tale soglia risultano caratterizzati da valori anomali influenti, esclusi dalla stima, circa l'8 per cento delle imprese rispondenti.
Dopo l'esecuzione della procedura automatica e prima di escludere gli *outlier* individuati viene comunque effettuata una attività di *feed-back* per verificare, mediante l'analisi grafica (*box plot* e *scatter plot*) e il ricontatto telefonico, la correttezza dei dati delle unità influenti anomale. Se vengono trovati errori di misura o di processo, questi vengono corretti e la procedura automatica viene eseguita nuovamente.
Per il settore del commercio all'ingrosso e degli intermediari del commercio (tabella 2), nel periodo considerato (2006-2007), la procedura di *macroediting* automatizzato ha un impatto significativo sulle stime definitive, determinando sempre una riduzione delle stesse tranne, che nel caso del quarto trimestre 2006.
A livello di gruppo di attività economica, nella maggior parte dei casi, l'effetto del *macroediting* automatizzato è quello di ridurre le stime definitive rispetto a quelle grezze, ma esistono casi in cui viene amplificata la variazione tendenziale. In ogni caso, si riscontrano pochissimi cambiamenti di segno nella stima delle variazioni tendenziali.
In conclusione, rispetto alle stime grezze la procedura automatica di *Hidiroglou-Berthelot*, pur essendo adatta alla dimensione campionaria e consentendo un risparmio di risorse rispetto alle procedure interattive, sembra avere un effetto sistematico di riduzione dell'ampiezza della variazione tendenziale stimata.
Per comprendere l'importanza della scelta del parametro "c" è stata effettuata una simulazione con diversi valori del parametro; ciò permette un confronto sul numero di *outlier* influenti esclusi e sull'impatto sulle variazioni tendenziali di fatturato nel periodo 2003-2007[73].
Nella simulazione il parametro 'c'viene posto uguale a 1,5 (strategia attuale), a 2 (strategia B) ed a 2,5 (strategia A) e si confrontano le stime delle variazioni tendenziali nel periodo 2003-2007.
Il profilo delle variazioni tendenziali (Grafico 1), dopo il trattamento degli *outlier* con la procedura automatica non subisce sostanziali modifiche rispetto alle stime grezze per tutte le soglie del parametro "c". E', invece, netto lo scostamento tra le stime delle variazioni tendenziali prime e dopo l'esecuzione della procedura automatica e la soglia di accettazione sembra influire in misura limitata su questo risultato: l'effetto del trattamento degli *outlier* influenti è, quindi, quello di ridurre sistematicamente la variabilità dei tassi di variazione degli indici .

[73] Si considera solo il primo trimestre 2007, dato per il quale è disponibile l'ultima stima definitiva a 180 giorni.

Tabella 3 – Impatto del macroediting automatizzato sulle stime definitive delle variazioni tendenziali di fatturato del commercio all'ingrosso – Periodo primo trimestre 2006-primo trimestre 2007

ATECO 2002	FASE	PERIODO				
		2006				2007
		I	II	III	IV	I
51 - Commercio all'ingrosso e intermediari del commercio	Stime Grezze	7,6	5,7	6,6	4,6	5
	Stime Finali	6,3	4,4	4,5	5,1	4,6
	Differenza	*-1,3*	*-1,3*	*-2,1*	*0,5*	*-0,4*
511 - Intermediari del commercio	Stime Grezze	4,9	1,6	5,1	6,1	4,5
	Stime Finali	2,4	2,7	1,6	3,3	6,8
	Differenza	*-2,5*	*1,1*	*-3,5*	*-2,8*	*2,3*
512 - Commercio all'ingrosso di materie prime agricole e animali vivi	Stime Grezze	4,7	2,7	0,4	-2,9	12,5
	Stime Finali	4,2	0,3	2,1	5,2	8,1
	Differenza	*-0,5*	*-2,4*	*1,7*	*8,1*	*-4,4*
513 - Commercio all'ingrosso di prodotti alimentari, bevande e tabacco	Stime Grezze	0	4,9	5,5	4,8	7,5
	Stime Finali	-0,3	4,2	3,7	3	4,4
	Differenza	*-0,3*	*-0,7*	*-1,8*	*-1,8*	*-3,1*
514 - Commercio all'ingrosso di altri beni di consumo finale	Stime Grezze	8,7	2,8	3,5	0,6	-0,4
	Stime Finali	6,6	2,4	1,9	3,5	2,4
	Differenza	*-2,1*	*-0,4*	*-1,6*	*2,9*	*2,8*
515 - Commercio all'ingrosso di prodotti intermedi	Stime Grezze	10,2	9,2	9,4	6,4	8,2
	Stime Finali	10,2	7,7	7,8	7,2	6
	Differenza	*0*	*-1,5*	*-1,6*	*0,8*	*-2,2*
518 - Commercio all'ingrosso di macchinari e attrezzature	Stime Grezze	13,5	8,7	10	10,9	2,7
	Stime Finali	9,8	2,8	4,8	8,9	2,9
	Differenza	*-3,7*	*-5,9*	*-5,2*	*-2*	*0,2*
519 - Commercio all'ingrosso di altri prodotti	Stime Grezze	9,9	5,4	11,8	4,9	1,9
	Stime Finali	7	4,5	8,4	3,2	4,5
	Differenza	*-2,9*	*-0,9*	*-3,4*	*-1,7*	*2,6*

Grafico 1 – Confronto tra le stime definitive delle variazioni tendenziali di fatturato grezze e per varie soglie del parametro 'c' per il commercio all'ingrosso – Periodo 2003-2007

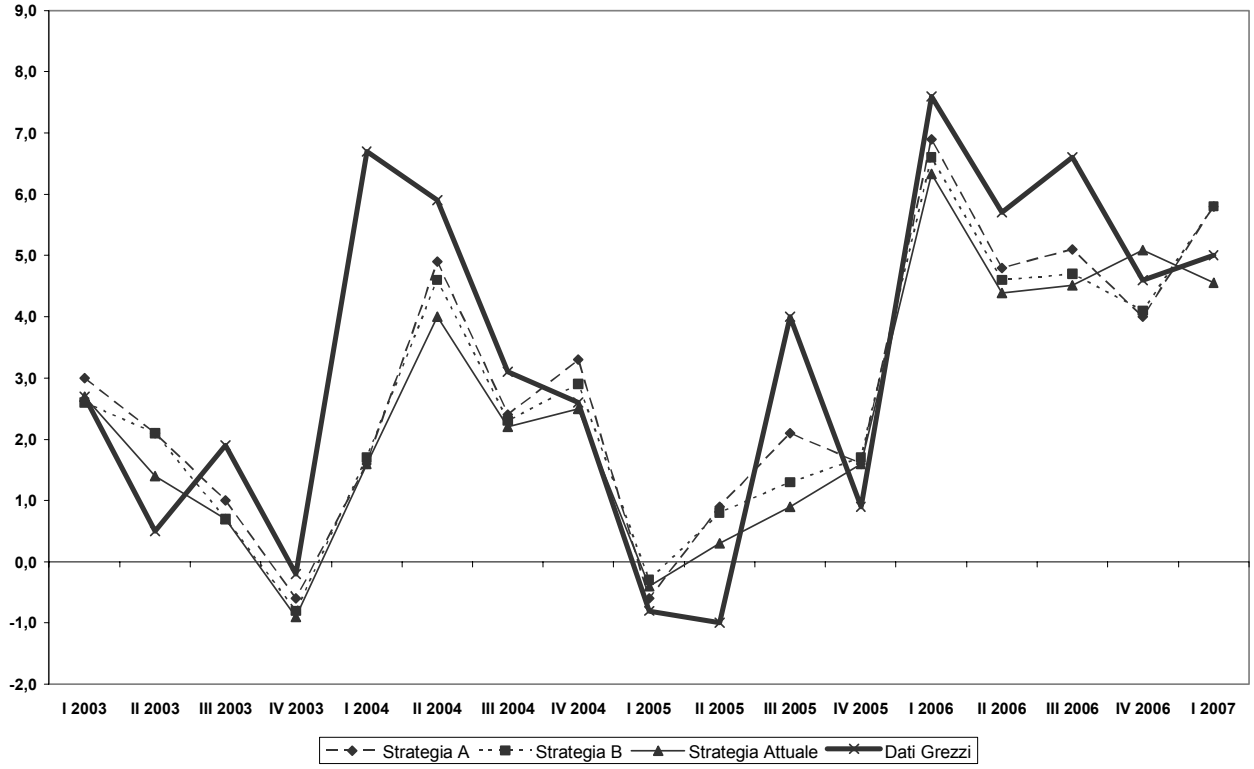
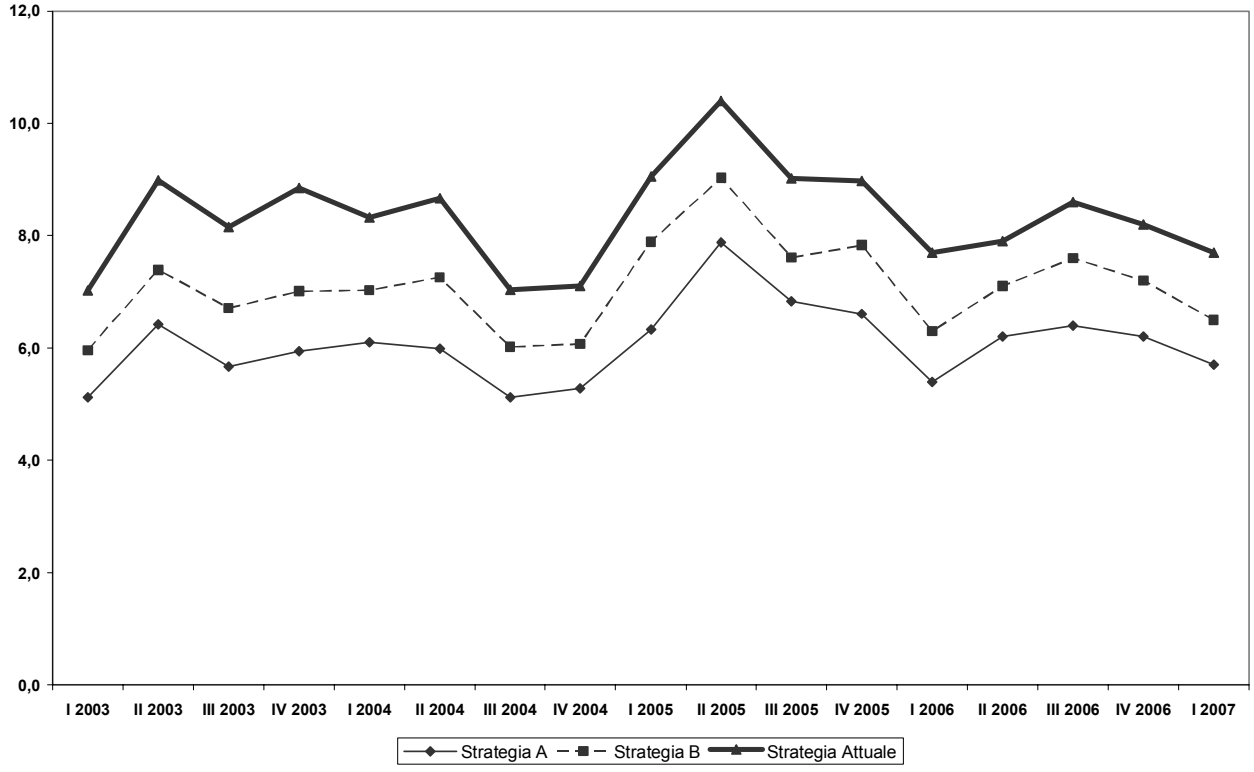


Grafico 2 – Frequenza percentuale degli *outlier* influenti esclusi dalla stima nelle diverse occasioni di indagine e con le diverse strategie – Periodo 2003-2007



La quota degli *outlier* influenti, esclusi dalla stima (Grafico 2), si abbassa rilassando la soglia di accettazione di *Hidiroglou-Berthelot:* la percentuale scende dall'attuale media dell'8 per cento con "c" =

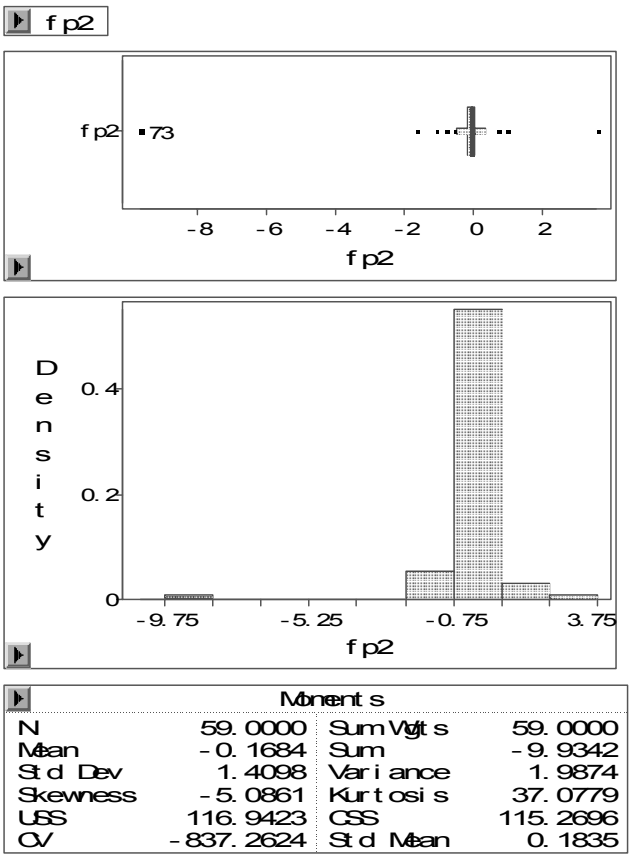
1,5 a circa il 5 per cento passando dal parametro con "c" = 2,5. L'analisi grafica sembra suggerire che rilassando le soglie si hanno nella maggioranza dei casi effetti poco significativi sulle stime, ma si riduce la perdita di informazioni dovuta all'esclusione di molte osservazioni anomale.

6. Il settore di servizi informatici

Per il settore dell'informatica l'imputazione delle mancate risposte utilizza il metodo del quoziente. La variazione media tendenziale dello strato viene calcolata sulle imprese rispondenti ritenute, in fase preliminare, non anomale. La selezione delle imprese (temporaneamente anomale) da escludere dai calcoli dello stimatore quoziente viene effettuata considerando una soglia della variazione tendenziale di fatturato pari al 50 per cento in valore assoluto. Se per una impresa non è disponibile il dato di fatturato relativo allo stesso trimestre dell'anno precedente, non si effettua alcuna imputazione.
Una volte effettuata l'imputazione, la fase di *macroediting* evidenzia le imprese influenti. Per definire se una impresa è o meno *outlier* si valuta se la sua variazione tendenziale, in valore assoluto, supera il 50%. Le imprese influenti e anomale vengono ricontattate telefonicamente per verificare l'eventuale presenza di errori di misura o di processo e per giustificare l'andamento della serie storica. Solo in alcuni rari casi, se si ritiene che la variazione tendenziale rilevata e controllata non sia rappresentativa della parte non osservata del campione e della popolazione, si procede all'esclusione dell'unità dal calcolo dell'indice per i servizi informatici. Gli eventi rari sono attribuiti, ad esempio, a cambiamenti di strato, fatturazione atipica[74], etc..
Gli *outlier* non rappresentativi esclusi del calcolo dell'indice riguardano una bassa percentuale di osservazioni (che non supera quasi mai l'1 per cento del campione teorico).
Un caso particolare riguarda le imprese appartenenti agli strati "censuari" dell'informatica in quanto in tale situazione, rilevando l'intera sottopopolazione e non solo un campione, anche l'unità posta agli estremi della distribuzione è sempre considerata rappresentativa, quindi sempre inclusa nei calcoli, anche se presenta una forte variazione tendenziale ed è influente.
Il metodo adottato nella fase di individuazione e trattamento degli *outlier,* se pure massimizza l'informazione raccolta, presenta lo svantaggio di richiedere più tempo e risorse rispetto a procedure automatiche.
Un caso di studio per la rilevazione dell'informatica riguarda la stima provvisoria a 90 giorni del secondo trimestre 2007 per lo strato delle piccole imprese appartenenti al gruppo (72.1) della "consulenza per l'installazione di sistemi informatici". Nel Grafico 3, l'osservazione n. 73, posizionata sulla coda negativa della funzione punteggio *micro*, è molto influente e, analizzando la serie storica del fatturato e degli addetti dei dati trimestrali 2006-2007, si vede che tale impresa ha aumentato notevolmente il numero di addetti, passando dalla classe 1-19 a quella 20-99. Il cambio di strato la rende eterogenea ed anomala rispetto alle altre unità dello strato originario. Siccome, l'aggiornamento delle variabili di stratificazione, attraverso informazioni desunte dall'indagine, avviene annualmente, si è ritenuto opportuno escludere dal calcolo dell'indice tale unità.

[74] In questo caso, si tratta di veri e propri errori di compilazione in quanto l'impresa non ha soddisfatto il principio di competenza ma quello di cassa nell'attribuzione del fatturato trimestrale, così come richiesto dalla definizione della variabile obiettivo. In particolare, tale problema si verifica nel settore dell'informatica in quanto l'erogazione del servizio è spesso superiore al trimestre di riferimento e l'impresa, a volte, ha difficoltà ad attribuire il fatturato di competenza in base allo stato avanzamento dei lavori eseguiti per l'espletamento del servizio.

Grafico 3 – Box plot della funzione punteggio micro (F P2) nel caso delle piccole imprese operanti nel gruppo 721 (consulenza per l'installazione di sistemi informatici) dell'informatica relativo alla stima provvisoria del secondo trimestre 2007



Proseguendo con l'esempio si vede dal *box plot* che un'altre impresa influente risulta posizionata sulla estrema coda positiva, anche se con più contenuta funzione punteggio: avendo confermato telefonicamente i dati e la crescita del fatturato tale impresa influente è stata considerata rappresentativa ed inclusa nel calcolo dell'indice.

Tabella 4 – Impatto del *macroediting* interattivo sulle stime definitive delle variazioni tendenziali di fatturato dell'informatica – Periodo primo trimestre 2006-primo trimestre 2007

Periodo		**Fasi del processo di controllo e correzione**		
Anni	**Trimestre**	**Stime grezze (solo rispondenti)**	**Stime con imputazione mancate risposte**	**Stime definitive (dopo *macroediting* interattivo)**
2006	I	6,8	6,2	2,6
	II	0,1	-0,3	1,6
	III	0,2	-0,6	1,9
	IV	-7,9	-7,2	0,7
2007	I	-0,5	-0,4	3,6

Nel settore dell'informatica, lo scostamento tra stima grezza (solo rispondenti) e quella successiva all'imputazione delle mancate risposte, che incorpora l'effetto dell'esclusione dal processo di

imputazione delle osservazioni anomale mette in evidenza differenze limitate e apparentemente non sistematiche .
Lo scostamento tra la stima con integrazione delle mancate risposte e definitiva (con il *macroediting* interattivo) è, invece, significativa e, tranne che nel caso del primo trimestre 2006, le stime, dopo la fase del *macroediting*, danno luogo a variazioni tendenziali positive (a fronte di valori negativi delle stime grezze). Ciò è dovuto al fatto che, pur escludendo o correggendo poche unità (al massimo l'1% del campione, vedi Tabella 5), queste ultime risultano molto influenti sulla stima delle variazioni tendenziali (alto punteggio micro) e, nella maggior parte dei casi, si posizionano nella coda negativa. Tali unità sono state considerate non rappresentative se appartenenti a strati campionari, mentre per gli strati "censuari" vengono corretti gli eventuali errori di misura e di processo, ma le osservazioni sono incluse nel calcolo dell'indice.
Il motivo dell'esclusione, oltre che ad ampie variazioni tendenziali di fatturato, è anche dovuto al peso di alcune imprese all'interno dello strato di appartenenza: queste imprese presentano un livello del fatturato trimestrale troppo elevato e nettamente estraneo rispetto alle altre unità omogenee per classe di addetti e attività economica. Pertanto si è deciso di escludere tali unità come se si trattasse di un errore di classificazione. Anche questo esempio dimostra come la fase di *macroediting* interattivo sia delicatissima nel processo di controllo e correzione.

Tabella 5 – Frequenza degli outlier non rappresentativi influenti sulla stima degli indici dell'informatica – Periodo primo trimestre 2006-primo trimestre 2007

Anni	Trimestri	Numero	Percentuale
2006	I	9	0,7
	II	6	0,5
	III	6	0,5
	IV	11	1,1
2007	I	11	0,7

7. Manutenzione e riparazione di autoveicoli

Per la rilevazione della manutenzione e riparazione di autoveicoli la fase di imputazione delle mancate risposte adotta un modello di regressione semplice (senza intercetta) del valore corrente con quello corrispondente dell'anno precedente. Qualora non sia disponibile l'osservazione sul fatturato al tempo (t-4), si usa come variabile ausiliaria il volume di affari di ASIA, (opportunamente trimestralizzato), quindi l'intero campione teorico viene sempre ricostruito. Per la stima del parametro β, effettuata coi minimi quadrati ordinari a livello di strato, vengono esclusi dall'insieme delle unità rispondenti i valori definiti anomali, individuati con la tecnica dei residui "studentizzati". In particolare, vengono eliminate le osservazioni per cui il valore assoluto del residuo studentizzato è maggiore di 2. Quindi le osservazioni che si discostano "troppo" dalla retta di regressione stimata, vengono esclusi nel processo di stima di β, secondo una procedura automatica. Usualmente l'effetto del processo di imputazione sulle stime degli indici è quello di ridurre la variazione tendenziale: la riduzione più marcata è stata registrata per le stime definitive del quarto trimestre 2005 (vedi Tabella 6).
Il *macroediting* per la rilevazioni della manutenzione e riparazione di autoveicoli determina le imprese influenti, mentre la soglia di accettazione per l'individuazione di *outlier* è definita da una variazione tendenziale di fatturato superiore al 50 per cento. Le unità influenti ed anomale vengono ricontattate telefonicamente per eliminare possibili errori e consentire rettifiche. Generalmente una osservazione *outlier* non viene esclusa dai calcoli dell'indice se si ritiene che essa sia rappresentativa della parte non osservata del campione e della popolazione. Se al contrario essa corrisponde ad un *outlier* non rappresentativo viene ristimata come mancata risposta (1% dei casi); questo è il caso di imprese che nel periodo considerato hanno visto una forte modifica del peso di attività normalmente di tipo secondario (ad esempio la vendita di veicoli, affiancata alla riparazione).

Il settore della manutenzione di autoveicoli è caratterizzato da una moltitudine di microimprese, in termini di fatturato, e quindi è molto raro, rispetto al settore dell'informatica, che vengano raggiunte elevate funzioni punteggio micro; ciononostante l'effetto del *macroediting* nella manutenzione, per quanto più contenuto è comunque piuttosto significativo e dà luogo, in questo caso, a una tendenza sistematica alla compressione delle variazioni tendenziali (positive) degli indici.

Tabella 6 – Impatto del processo di controllo e correzione sulle stime definitive delle variazioni tendenziali di fatturato della manutenzione e riparazione di autoveicoli – Periodo: primo trimestre 2005-Primo trimestre 2007

Periodo		Fasi del processo di controllo e correzione		
Anno	**Trimestre**	**Stime grezze (solo rispondenti)**	**Stime con imputazione delle mancate risposte**	**Stime definitive (dopo *macroediting* interattivo)**
2005	I	1,0	0,5	0,2
	II	3,3	2,9	2,0
	III	4,4	3,0	0,2
	IV	6,5	4,9	4,2
2006	I	8,0	7,6	6,4
	II	3,6	3,1	2,0
	III	2,7	2,1	2,0
	IV	3,4	2,9	2,8
2007	I	6,3	5,9	5,2

8. Settori oligopolistici

Nel caso delle rilevazioni dei settori oligopolistici (dei trasporti marittimi e aerei, dei servizi postali e attività di corriere, delle telecomunicazioni), si ricorre all'imputazione delle mancate risposte con il metodo del quoziente. Per le imprese dominanti si cerca sempre di ottenere il dato effettivo. In rari casi si ricorre a valutazione approssimate (soprattutto in sede di stime anticipata se il dato non è ancora disponibile) ricorrendo a riponderazioni sulle serie storiche, o riproporzionamenti su variabili altamente correlate. Ad esempio i dati tendenziali su merci e passeggeri trasportati vengono utilizzati nel processo di stima in caso di mancata risposta nel settore dei trasporti.

Le unità di minor dimensione, che presentano ampie variazioni tendenziali di fatturato, collocandosi alle estremità della distribuzione, si ricontattano telefonicamente per correggere eventuali errori. Se la variazione tendenziale viene confermata e si ritiene che l'impresa rappresenti un caso isolato viene temporaneamente esclusa del calcolo dell'indice, mentre le imprese che presentano scostamenti sistematici e persistenti di variazione tendenziale di fatturato vengono seguite nel tempo e solitamente escluse in modo permanente dal calcolo degli indicatori

Le imprese individuate come *outlier* vengono ricontattate telefonicamente per verificare l'eventuale presenza di errori, preservando la distribuzione dei dati osservati soprattutto per le imprese dominanti. Per queste ultime è possibile talvolta effettuare riscontri sull'andamento economico confrontando i bilanci semestrali ed annuali per la voce relativa ai ricavi, o studiando la relazione economica 'impresa per interpretare la dinamica del settore. Anche per le piccole imprese, che hanno comunque un peso marginale nella determinazione dell'indice, vengono individuate eventuali anomalie da verificare tramite contato telefonico.

Un caso particolare che riguarda le grandi imprese, particolarmente complesso da risolvere, è quello relativo ai *break* strutturali, dovuti ad esempio ad operazioni societarie di fusioni, scissioni, compravendite di aziende o di ramo di aziende, trasformazioni societarie (Buffelli e Sistoli, 2004) .

9. Conclusioni e sviluppi futuri

Le strategie, per il metodo di controllo e correzione degli errori non campionari adottato nelle rilevazioni trimestrali di fatturato degli altri servizi, presentano aspetti comuni a tutte le indagini (correzione degli errori di lista, di misura e di processo) e aspetti divergenti in base al disegno ed alla dimensione campionaria per l'imputazione delle mancate risposte ed il trattamento degli *outlier*.
Con il cambio base e con l'introduzione della nuova classificazione delle attività economiche previste per il 2008, si cercherà di integrare alcune metodologie. Per esempio è prevista una simulazione per verificare l'applicazione del metodo automatizzato di *Hidiroglou-Berthelot* alla rilevazione della manutenzione e riparazione di autoveicoli. Un altro aspetto che si sta valutando è legato al valore da assegnare al parametro "c" dell'algoritmo, al fine di ottenere una contenuta percentuale di *outlier* influenti e risultati coerenti e al fine di ridurre la differenza sistematica tra stime grezze e definitive.
Per le rilevazioni dei settori oligopolistici e per l'informatica un cambio di strategia sembra più difficile, considerando soprattutto l'importanza delle imprese dominanti appartenenti ai settori oligopolistici e delle imprese operanti in strati caratterizzati da un elevato peso nel settore dell'informatica, che richiedono procedure di revisione "manuali"/"artigianali", in quanto poche unità influenti anomale possono determinare significativi scostamenti tra stime grezze e definitive.
Resta comunque aperto il problema degli *outlier,* comune a tutte le rilevazioni campionarie sulle imprese: le regole per l'inclusione o l'esclusione di un dato anomalo dalle stime dell'indice congiunturale non sono precise e spesso è l'esperto ad analizzare il singolo caso e a definire, con un qualche grado di soggettività, la soluzione opportuna.
L'attenzione degli studiosi, in letteratura, si è sempre concentrata sugli *outlier* rappresentativi, ossia quelli non affetti da errori di misura o di processo, ma posti all'estremità della distribuzione e che devono pertanto essere inclusi nel calcolo degli indicatori.
Se gli *outlier* non sono omogenei all'interno dello strato di riferimento può porsi un problema di errata classificazione dell'impresa nello strato e in questo caso l'approccio più ricorrente nella prassi operativa è quello di escluderli dalla stima, ma non è noto se tale metodo possa provocare distorsione.
Talvolta invece di escludere un *outlier* si può considerarlo autorappresentativo, supponendo cioè che sia unico nella parte osservata e non osservata della popolazione.
Purtroppo non esiste ad oggi un metodo consolidato a livello teorico per il trattamento degli *outlier*. La regola di modificare il peso dei valori anomali in modo che il loro impatto sulla stima campionaria sia tenuto basso lascia aperto il problema della arbitrarietà del peso stesso. [75].
In tutti i casi, le caratteristiche peculiari dei singoli settori del comparto dei servizi determinano le scelte del disegno campionario e, di conseguenza, orientano anche le strategie di controllo e correzione degli errori non campionari.

[75] BERGDAHL,et altri *Model quality report in business statistics*, pag. 54-60, pubblicato da EUROSTAT

Bibliografia

Barcaroli G., Luzi O., Ceccarelli C. "Il macroediting: tecniche di correzione interattiva di variabili quantitative guidata dall'analisi degli aggregati. Il caso del sistema dei conti delle imprese". *Quaderni di ricerca ISTAT*, n. 1/1998.

Bergdahl, Black, Bowater, Chambers, Davies, Draper, Elvers, Full, Holmes, Lundqvist, Lundstrom, Nordberg, Perry, Pont, Prestwood, Richardson, Skinner, Smith, Underwood, Williams. *Model quality report in business statistics*. EUROSTAT, pp. 82-160 e pp. 54-60, 1997.

Buffelli G., Sirtoli M. *Le operazioni straordinarie delle società*. Edizione Giuffré, 2004.

Cirianni A., Gismondi R. "Identificazione e trattamento statistico delle osservazioni anomale nel settore degli altri servizi". In direttiva ISTAT TRAC16: "*Sperimentazione di stime anticipate per specifici indicatori congiunturali, finalizzata al rilascio in produzione delle relative metodologie*", paragrafo 4.4, 2006.

Cirianni A. *Refreshment e rotazione annuale del panel dell'informatica 2007*. Nota tecnica, 15 Marzo 2007.

Cirianni A., Panizon F. "The problem of outliers in short term surveys on turnover of other services". *Atti del Convegno Intermedio SIS 2007*, Venezia, 6-8 Giugno, 2007.

EUROSTAT. *Methodology of short term business statistics*. Dicembre, 2005.

Gismondi R. "Model-based sample selection using balanced sampling". *Rivista di statistica ufficiale*, n. 3/2002, pp. 81-111.

Hidiroglou M.A., Berthelot J.M. "Statistical editing and imputation for periodic business survey". *Survey methodology*, 12, pp. 73-84, 1986.

ISTAT. *Nota informativa del comunicato stampa relativo agli indici trimestrali di fatturato per alcune attività dei servizi*. 2007.

Lee H. *Outliers in business surveys*. In Cox, Binder, Chianappa, Christianson, Colledge, Kott *Business survey methods*. John Wiley & Sons, New York, 1995, pp. 503-523.

La rilevazione mensile sulle vendite al dettaglio: metodi per il controllo e la correzione dei dati

Annarita Giorgi, Istat, Servizio Statistiche sull'attività dei servizi
Tiziana Pichiorri, Istat, Servizio Statistiche sull'attività dei servizi

Sommario: nel presente lavoro vengono illustrati i principali metodi adottati per il controllo e la correzione dei dati della rilevazione mensile sulle vendite al dettaglio. Dopo una sintesi delle principali caratteristiche della rilevazione e dei punti critici che riguardano il processo produttivo ad essa sottostante vengono descritti i suddetti metodi. Essi si concretizzano, da un lato, nelle soluzioni attuate per ridurre il numero di mancate risposte e di dati anomali, dall'altro nella metodologia applicata per la stima e il trattamento degli errori.

Parole chiave: statistiche congiunturali, commercio al dettaglio, tasso di risposta, controllo dei dati.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Caratteristiche generali della rilevazione mensile sulle vendite al dettaglio

La rilevazione mensile sulle vendite al dettaglio nella sua struttura attuale ha avuto inizio dal 1996, anno in cui sono state introdotte sostanziali innovazioni che hanno riguardato la metodologia generale dell'indagine, la tecnica di calcolo nei numeri indici mensili, le modalità di diffusione delle informazioni e, più in generale, l'intero processo produttivo.

L'indagine ha come dominio di riferimento l'universo delle imprese la cui attività prevalente è il commercio al dettaglio realizzato mediante punti di vendita al dettaglio in sede fissa; la vendita riguarda prodotti nuovi, con esclusione delle rivendite di tabacchi, generi di monopolio, autoveicoli e combustibili. Tutte le imprese considerate sono classificate nella divisione 52 dell'ATECO 2002[76].

La produzione degli indicatori che risulta dall'elaborazione dei dati raccolti è disciplinata Regolamento STS dell'Unione europea che riguarda le Statistiche congiunturali sulle imprese e i servizi, nelle sue versioni originale (CE n. 1165/98) ed emendata (CE 1158/2005). In particolare l'allegato C dei suddetti Regolamenti prevede nel dettaglio i contenuti (elenco delle variabili e livello di dettaglio), la natura degli indicatori (grezzi o sottoposti a correzione per gli effetti dovuti al numero di giorni di calendario o alla stagionalità) e i tempi relativi alla trasmissione delle statistiche prodotte.

Sulla base di quanto stabilito dal suddetto Regolamento, vengono elaborati e diffusi indici mensili del valore delle vendite con base 2000=100[77] con un ritardo medio di 54 giorni calcolato a partire dalla fine del periodo di riferimento.

A partire dal 2003, a seguito dei risultati prodotti dai lavori di una task force coordinata da Eurostat e riguardante l'individuazione di una metodologia che consentisse l'elaborazione e la diffusione di indicatori a 30 giorni relativi al volume delle vendite a livello UE, è stata avviata la produzione di indici "anticipati" per il totale delle vendite, per le vendite di alimentari e non alimentari, trasmessi ad Eurostat stesso entro 30 giorni dalla fine del periodo di riferimento, in forma confidenziale. Tali indicatori sono stati utilizzati unicamente per il calcolo dei relativi indici aggregati riferiti all'Unione europea nel complesso, mentre non sono stati diffusi a livello nazionale.

Successivamente, il Regolamento STS emendato del 2005 ha stabilito l'obbligo di trasmissione di stime a 30 giorni relative ai domini più aggregati (gli stessi già individuati per gli indici anticipati). In altre parole il Regolamento emendato richiede la fornitura ad EUROSTAT di indici delle vendite a 30 giorni dalla fine del mese di riferimento; questi rappresentano stime anticipate degli indici definitivi, diffusi in Italia tra 50 e 55 giorni.

Le statistiche mensili relative all'andamento delle vendite al dettaglio fanno parte dei Principal European Economic Indicators (PEEIs), ovvero di un set di indicatori di rilevanza strategica per la politica economica a livello europeo. Di conseguenza negli anni più recenti l'attenzione rivolta agli aspetti qualitativi che riguardano sia i metodi di raccolta dei dati sia l'elaborazione degli stessi è fortemente cresciuta.

Per quanto riguarda gli aspetti più strettamente operativi, l'indagine prevede che i questionari predisposti per la rilevazione dei dati vengano inviati mensilmente per posta alle unità del campione, alle quali è chiesto di restituirli all'Istat, dopo averli compilati, entro 10 giorni dalla ricezione. Per la trasmissione dei modelli compilati all'Istat sono rese disponibili le seguenti modalità: fax, posta ordinaria, web; il fax e la posta ordinaria costituiscono di gran lunga i mezzi preferiti dalle imprese. Dal punto di vista della gestione delle operazioni di raccolta si cerca di incoraggiare i rispondenti all'uso del fax, anche perché la disponibilità di utenze fax-server dedicate facilita indubbiamente l'acquisizione. I ritorni tramite posta ordinaria (a carico dell'Istat) sono in buona parte connessi alle caratteristiche dei rispondenti (in larga parte piccole imprese, non sempre dotate di apparecchi fax e in generale poco informatizzate), tuttavia è constante lo sforzo dedicato all'abbattimento di questa componente inerziale che ha effetti negativi sulla tempestività. Allo stato attuale la percentuale di questionari che vengono restituiti a mezzo fax è superiore rispetto a quella relativa ai ritorni via posta ordinaria, che tuttavia non è ancora sufficientemente trascurabile.

Con riferimento alla modalità web, questa è stata introdotta in via sperimentale soltanto per le imprese appartenenti al cosiddetto "campione rapido", cioè un sottoinsieme del campione totale composto da

[76] Restano, fuori dal campo di osservazione i punti di vendita di beni usati, gli ambulanti e i mercati.

[77] Il passaggio alla base di riferimento 2005=100 è prevista per il 2009, con un ritardo rispetto ai tempi consueti di adeguamento della base dovuto all'adozione della classificazione ATECO che avviene contemporaneamente.

unità seguite con particolare attenzione affinché le informazioni che le riguardano siano disponibili per l'elaborazione degli indici anticipati. Attualmente la percentuale di rispondenti che utilizzano tale modalità è ancora molto ridotta, anche se è evidente la tendenza ad un incremento costante; un obiettivo da perseguire nell'immediato futuro è quello di fornire a tutte le imprese del campione la possibilità di compilare on-line il questionario mensile.
Nella tabella seguente sono riportate, con riferimento al periodo gennaio 2006-luglio 2007 le percentuali di risposta distinte in base alle modalità utilizzate; le percentuali sono state calcolate sia per i questionari relativi agli indici anticipati sia per quelli relativi agli indici diffusi a 54 giorni dalla fine del periodo di riferimento.
I dati riportati confermano che la modalità "posta" ha ancora un peso rilevante che in alcuni mesi supera addirittura quello del fax. Naturalmente ciò vale soprattutto nel caso degli indici a 54 giorni dal momento che per l'elaborazione degli indici a 30 giorni si considerano le risposte più tempestive e nella determinazione della tempestività, oltre all'impegno del singolo rispondente, gioca un ruolo rilevante anche il mezzo di trasmissione del questionario compilato.

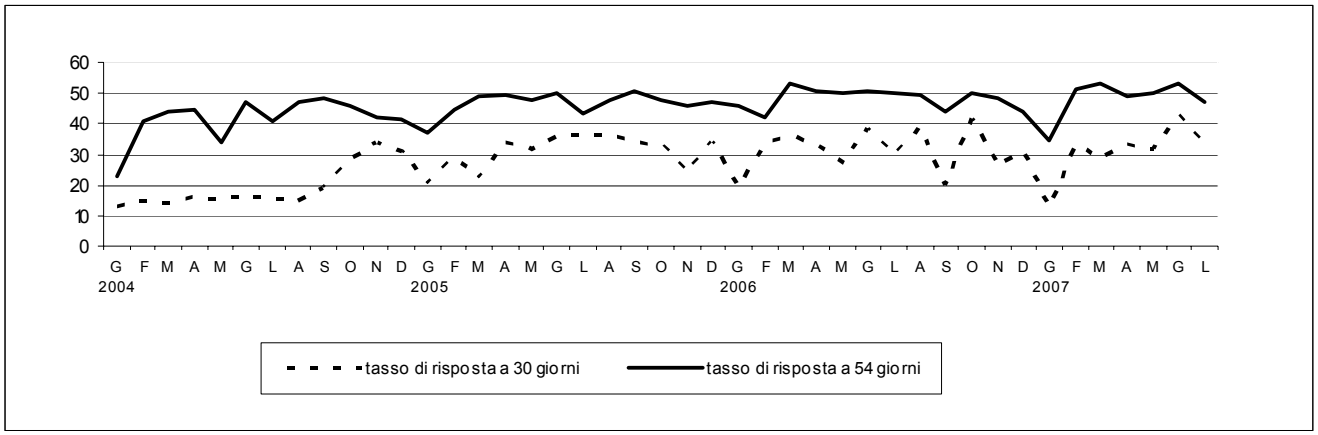
Tavola 1: percentuale di risposte per modalità e tipo di indice *(gennaio 2006-luglio 2007)*

Periodo		Indici a 30 giorni			Indici a 54 giorni		
Modalità		**POSTA**	**FAX**	**WEB**	**POSTA**	**FAX**	**WEB**
2006	Gennaio (*)	28,0	72,0	0,0	43,9	56,1	0,0
	Febbraio	44,4	53,4	2,2	54,0	43,8	2,2
	Marzo	31,8	65,0	3,1	38,7	58,7	2,7
	Aprile	23,5	72,3	4,2	42,2	54,7	3,1
	Maggio	49,0	46,0	5,0	40,5	56,7	2,9
	Giugno	31,0	65,5	3,5	35,1	61,9	3,0
	Luglio	19,6	75,8	4,6	41,2	55,5	3,3
	Agosto	39,3	56,5	4,1	39,3	57,1	3,5
	Settembre	21,9	68,7	9,3	41,9	53,7	4,4
	Ottobre	41,5	55,4	3,1	41,2	56,2	2,6
	Novembre	37,9	57,1	5,0	42,0	55,1	2,9
	Dicembre	37,2	59,2	3,6	45,0	52,2	2,8
2007	Gennaio	1,0	95,5	3,4	28,4	68,3	3,3
	Febbraio	20,7	74,7	4,6	39,1	57,3	3,6
	Marzo	12,9	81,5	5,5	41,9	54,1	4,0
	Aprile	43,7	50,1	6,2	51,4	42,9	5,7
	Maggio	49,1	44,2	6,7	42,7	52,4	4,9
	Giugno	41,8	53,5	4,7	44,3	51,1	4,6
	Luglio	48,2	46,2	5,6	42,9	51,9	5,2

(*) L'assenza di risposte via web per questo mese è da attribuire a problemi di natura tecnica che hanno riguardato il sistema di acquisizione dati

Il campione dei rispondenti effettivi, su cui si basa il calcolo degli indici mensili supera le 4.000 imprese. Al momento dell'elaborazione delle stime che verranno diffuse mediante il comunicato stampa il tasso di risposta si aggira intorno al 50%. Per ciascun periodo di riferimento (mese), successivamente alla diffusione del comunicato continuano ad arrivare altri questionari le cui informazioni vengono registrate ma non utilizzate nelle fasi di calcolo degli indici. Il numero di risposte pervenute in ritardo varia molto di mese in mese e può, in alcuni casi, essere influenzato da fattori stagionali (es. periodi di chiusura per ferie delle imprese). In generale va comunque sottolineato che l'incremento del tasso di risposta mensile costituisce uno degli obiettivi costantemente perseguiti nell'ottica del miglioramento della qualità delle statistiche prodotte. Compatibilmente con le regole di contenimento della pressione statistica, infatti, si cerca di incoraggiare i rispondenti mettendo a loro disposizione, come detto in precedenza, diverse opportunità per la restituzione dei questionari compilati ed offrendo tutto il supporto richiesto in termini di chiarimenti o aiuto alla compilazione.
Nella figura che segue sono rappresentati i tassi di risposta mensili rispettivamente per gli indici a 30 e a 54 giorni per il periodo che va da gennaio 2004 a luglio 2007.

Figura 1: tassi di risposta mensili della rilevazione sulle vendite al dettaglio relativi all'elaborazione degli indici a 30 e a 54 giorni *(gennaio 2004-luglio 2007)*



A parte l'anno 2004, per il quale i tassi di risposta sono più contenuti, soprattutto a causa della fase di rodaggio del processo relativo agli indicatori anticipati, dalla figura si nota come nell'ultimo periodo il tasso di risposta a 54 giorni sembra essersi stabilizzato intorno al 50%, mentre il tasso di risposta relativo agli indici a 30 giorni mostra un lieve trend crescente[78].

Tra le attività intraprese per limitare l'effetto delle mancate risposte è il ricorso a solleciti telefonici mirati che riguardano soprattutto le imprese più grandi e quelle appartenenti agli strati per i quali è necessario disporre di un numero maggiore di risposte. Con l'avvio del calcolo delle stime anticipate è stato sperimentato un sistema automatico di solleciti realizzati sfruttando il servizio fax-server dell'Istat. Tale sistema tuttavia si è rivelato troppo oneroso per le imprese che comunque sono chiamate a rispondere mensilmente e con scadenze piuttosto rigide. Un sistema di tal genere potrebbe funzionare perfettamente se, date le scadenze, fosse possibile monitorare in tempo reale i ritorni dei questionari, anche di quelli restituiti via posta.

Come si è detto la popolazione di riferimento per la rilevazione è costituita da un insieme di imprese classificate nella divisione 52 dell'ATECO 2002. Tale insieme è composto da oltre 500mila unità che occupano più di un milione e 500mila addetti. Va comunque osservato che si tratta in larga parte di imprese di piccole dimensioni, spesso addirittura individuali. Ciò produce degli effetti sia sulla definizione della metodologia della rilevazione sia sulle fasi operative.

L'unità di rilevazione, che coincide con l'unità di analisi (come richiesto dal Regolamento STS), è dunque l'impresa definita come sopra.

La lista di riferimento dalla quale viene estratto il campione è costituita dall'archivio ASIA[79].

Le informazioni rilevate sono quasi tutte di natura quantitativa e le corrispondenti variabili sono: valore delle vendite relativo al mese di riferimento, valore delle vendite relativo allo stesso mese dell'anno precedente e valore delle vendite realizzato nel corso del mese di riferimento attraverso la vendita di beni appartenenti a 15 gruppi di prodotti. Tali variabili vengono rilevate in corrispondenza di ciascuna forma di vendita[80] attraverso la quale l'impresa esercita la propria attività. Inoltre, per ciascuna forma di vendita, viene rilevato il numero complessivo degli addetti (indipendenti e dipendenti), il numero di punti di vendita, la superficie di vendita complessiva e il numero medio di giorni in cui i punti di vendita dell'impresa sono rimasti aperti al pubblico. Il valore delle vendite viene richiesto, in generale, a lordo dell'IVA[81]

[78] I valori più contenuti relativi ai mesi di gennaio degli anni rappresentati sono dovuti ad un ritardo nella raccolta dei dati che è conseguenza delle fasi di aggiornamento annuale del campione.

[79] La lista derivata dall'archivio ASIA presenta un ritardo nell'aggiornamento di oltre 20 mesi rispetto alla rilevazione.

[80] Per forma di vendita si intende il "modo" attraverso il quale l'impresa effettua l'attività di vendita, ovvero il tipo di organizzazione della distribuzione.

[81] Il regolamento STS prevede che il valore delle vendite sia rilevato al netto dell'IVA. Tuttavia per facilitare i rispondenti e in considerazione del fatto che l'obiettivo fondamentale è la stima della variazione tendenziale del valore delle vendite si è ritenuto finora di poter considerare dati che comprendessero anche l'IVA.

L'indagine è di tipo campionario e il relativo disegno prevede un campione casuale all'interno di strati individuati incrociando l'attività economica (a livello di classe ATECO), cinque classi dimensionali determinate sulla base del numero di addetti delle imprese (1-2, 3-5, 6-9, 10-19 e >19) e da due forme di vendita definite come "Grande distribuzione" e "imprese operanti su piccole superfici". I punti di vendita che fanno capo a tali imprese operano su tutto il territorio nazionale.

All'inizio di ciascun anno il campione viene aggiornato, ovvero parte delle unità vengono sostituite, soprattutto imprese di piccola dimensione. L'aggiornamento si rende necessario per tener conto delle trasformazioni strutturali delle unità e della natimortalità delle imprese oltre che per contenere il carico statistico. Per il 2007 il campione è composto da circa 8.000 imprese.

Obiettivo dell'indagine è la stima della variazione del valore mensile delle vendite riferito alle imprese nel complesso e classificate secondo opportuni domini di riferimento identificati dalla forma di vendita, dal settore merceologico o dalla dimensione determinata sulla base del numero di addetti. Tali domini sono di seguito elencati nel dettaglio:

- Alimentari e non alimentari (settore merceologico)
- Grande distribuzione e imprese operanti su piccole superfici (forma di vendita)
- Forme di vendita della Grande distribuzione
- Classi di addetti delle imprese

Vengono inoltre elaborati indici relativi a quattro ripartizioni geografiche (Nord-ovest, Nord-est, Centro, Sud e isole) e ai principali settori merceologici.

Gli indici per i domini sopra elencati sono ottenuti come media aritmetica ponderata di un opportuno insieme di indici elementari di strato; l'aggregazione viene effettuata utilizzando pesi proporzionali al fatturato stimato per ciascuno strato con riferimento all'anno 2000 (anno base).

Gli indici di strato si ottengono moltiplicando il corrispondente indice relativo allo stesso mese dell'anno precedente[82] per il rapporto tra le medie campionarie dello strato stesso – riferite rispettivamente ai due mesi messi a confronto - che rappresenta il rapporto tendenziale rilevato per ciascuno strato nel mese di riferimento.

2. Problematiche principali

Gli aspetti più problematici della rilevazione sono connessi essenzialmente a due fattori: la tempestività e le caratteristiche strutturali della popolazione di riferimento.

Per quanto riguarda la tempestività, la riduzione dei tempi di rilascio dei dati è un obiettivo costante che si è rafforzato nel corso degli ultimi anni fino all'emendamento del Regolamento STS che stabilisce che le stime relative agli aggregati principali vengano trasmesse entro 30 giorni rispetto al mese di riferimento.

L'avvio del calcolo delle stime anticipate di cui si è parlato in precedenza, finalizzato alla produzione da parte di Eurostat di indicatori sulle vendite al dettaglio per l'insieme degli aggregati europei (Ue ed Uem) a poco più un mese dalla fine del periodo di riferimento, ha comportato profondi cambiamenti nel processo che fa capo alla rilevazione. Grazie ad essi è possibile rispettare la tempistica fissata e, allo stesso tempo, elaborare indici anticipati la cui qualità – misurata in termini di revisione per la stima finale – ha raggiunto livelli soddisfacenti.

L'ulteriore impegno richiesto dal Regolamento STS emendato comporta lo sforzo di curare anche la diffusione delle stime anticipate che, nel momento in cui non saranno più considerate "confidenziali" da parte di Eurostat, saranno effettivamente disponibili agli utilizzatori. La naturale conseguenza di tutto ciò sarà la pubblicazione degli indicatori anticipati anche a livello nazionale e la necessità di curare le informazioni relative al rilascio delle informazioni anticipate e definitive, al fine di favorire una corretta interpretazione degli indicatori stessi.

[82] In realtà questa è soltanto una delle dimensioni della qualità delle stime prodotte, ma è quella che ne giustifica l'interesse da parte degli utilizzatori (primo fra tutti la Banca Centrale Europea) che, all'interno dell'UEM si trovano a decidere in merito alle strategie di politica monetaria.

Le caratteristiche strutturali della popolazione di riferimento costituiscono un ulteriore elemento critico. Innanzitutto, come si è detto, la compresenza di un insieme relativamente poco numeroso di imprese di grandi dimensioni e di un insieme ben più ampio di imprese piccole, spesso individuali e/o a conduzione familiare rende necessaria, nelle varie fasi del processo, l'adozione di opportune strategie che consentano di gestire tali elementi.

In primo luogo il campione mensile non permette di garantire, per l'insieme delle imprese piccole, una copertura elevata in termini di numero di unità. Al contrario si cerca di raccogliere informazioni per un numero di imprese grandi che, in alcuni strati, si avvicina molto alla numerosità della popolazione di riferimento. D'altra parte, poiché esse rappresentano le unità che hanno un peso maggiore in termini di fatturato e valore delle vendite, è fondamentale disporre tempestivamente delle informazioni che le riguardano. A questo proposito va osservato che le imprese più grandi sono quelle più organizzate e informatizzate, ovvero quelle per cui, almeno in linea teorica, la pressione statistica dovrebbe avere un peso meno rilevante. Viceversa sono molto numerosi i casi in cui gli esercizi commerciali di dimensioni più ridotte debbano ricorrere all'assistenza di un commercialista per la compilazione del questionario; da ciò deriva un costo per le imprese e un ritardo nella disponibilità dei dati determinati da tale intermediazione. Il ricorso al web per la compilazione e la trasmissione dei dati mensili all'Istat potrebbe costituire una soluzione per questo tipo di problema ma, come già osservato, la diffusione di questa modalità di fornitura dei dati resta limitata.

Un'altra caratteristica della popolazione di riferimento che incide fortemente sulle informazioni raccolte è l'elevato dinamismo di un settore nel quale sia la nascita di nuovi esercizi commerciali sia la cessazione di attività sono fenomeni particolarmente frequenti. Accanto a ciò va tenuta in considerazione una serie di eventi relativi alle imprese e che si manifestano in quantità rilevante, quali fusioni, scorpori, cessioni di rami d'azienda ecc… Ne derivano mutamenti significativi che riguardano struttura delle unità e molto spesso sanciscono la cessazione di una specifica attività. Tali eventi sono di frequente rilevati direttamente e in anticipo rispetto alla lista di estrazione, che di fatto viene aggiornata ogni anno. Tuttavia è indispensabile tener conto di essi nelle fasi di controllo dei valori anomali, dal momento che l'obiettivo fondamentale è l'elaborazione di indicatori di tipo longitudinale basati sul confronto dei risultati relativi ad una stessa unità osservata in due tempi diversi.

Accanto alle criticità di cui si è appena parlato vanno segnalati ulteriori aspetti che possono essere fonte di problemi. Un esempio è costituito dalla superficie di vendita; tale variabile viene utilizzata convenzionalmente anche per la definizione delle diverse forme distributive, ma può accadere che non vi sia esatta corrispondenza tra la definizione teorica basata sulla superficie di vendita e l'attività svolta di fatto. In conseguenza di ciò in molti casi si rende necessario effettuare degli accertamenti al fine di evitare errori nell'attribuzione dell'attività prevalente effettivamente svolta.

Questo tipo di problemi ha effetto anche sull'identificazione dell'insieme delle imprese che fanno parte della Grande distribuzione[83] e riguarda principalmente gli esercizi non specializzati, ovvero quelli per i quali non esiste una tipologia prevalente di prodotti venduti (si pensi ad esempio agli ipermercati o ai grandi magazzini).

3. Strategie di controllo e correzione

La rilevazione mensile sulle vendite al dettaglio prevede un processo di controllo e correzione dei dati in parte automatizzato, in parte interattivo. Tale processo coinvolge la rilevazione nel complesso, dalla raccolta dei dati, alla registrazione fino alle fasi successive e ha l'obiettivo di ridurre o eliminare gli errori non campionari laddove possono avere origine.

L'identificazione e la correzione di errori non influenti e l'accertamento del rispetto dei principali vincoli di compatibilità vengono effettuati tramite un controllo di tipo automatizzato. Le regole di compatibilità possono essere schematizzate come segue:

[83] Tale insieme non è automaticamente identificabile attraverso la classificazione ATECO poiché i criteri che permettono di identificare un'impresa come appartenente alla Grande distribuzione non tengono conto solo dell'attività economica prevalente ma anche di un'altra serie di caratteristiche, prima fra tutte il tipo di organizzazione attraverso il quale la suddetta attività si realizza.

- condizione di validità di un questionario: deve contenere almeno un valore per le vendite mensili maggiore di zero
- controllo delle varie sezioni del questionario (dati strutturali, valori mensili, sezione rimanente)
- controllo sull'attività economica dell'impresa e sulla forma di vendita
- controllo sul numero degli addetti.

Il controllo interattivo, al contrario, viene effettuato per l'individuazione e la correzione di errori influenti e nei casi in cui si ritiene necessario condurre degli accertamenti preliminari.

In linea di principio la regola sottostante l'intero processo di controllo e correzione riguarda la necessità di ridurre al minimo gli interventi sui dati e, allo stesso tempo, eliminare il maggior numero di errori.

I controlli effettuati riguardano principalmente i microdati (microediting) e sono finalizzati sia a garantire una buona qualità degli indicatori prodotti, sia alla costruzione di una base dati che fosse utilizzabile anche per i controlli di tipo longitudinale.

La predisposizione di una base di dati da utilizzare per il calcolo degli indici mensili delle vendite al dettaglio richiede che, a partire dall'insieme dei dati grezzi, vengano completati due processi di controllo, l'editing e l'imputazione delle mancate risposte. Tali processi hanno una doppia dimensione, trasversale e longitudinale. A questo proposito va osservato come essi possono, in alcuni casi, andare oltre il momento di rilascio delle stime. Ove infatti vengano accertati degli errori oppure nei casi in cui le imprese stesse comunichino una revisione dei dati si effettuano comunque le opportune correzioni sul database dei microdati

Le tipologie di errori da cui sono più frequentemente affetti i dati raccolti riguardano:

- le caratteristiche strutturali dell'impresa commerciale inserita nel campione, ovvero derivare da errori della lista di estrazione;
- le informazioni mensili delle singole unità.

In entrambi i casi occorre distinguere il caso in cui si tratti di errori influenti, ovvero tali da avere un effetto diretto sulle stime finali, dal caso in cui gli errori siano non influenti. Le unità influenti che presentano anomalie nei dati sono spesso ricontattate per accertare il tipo di problema ed eventualmente apportare una correzione. Le imprese della Grande distribuzione richiedono dei controlli particolari in quanto sono quelle che presentano una complessità maggiore e hanno un peso rilevante all'interno dei domini di cui fanno parte.

E' importante osservare che in alcuni casi l'individuazione di errori trova un ostacolo nella coerenza delle informazioni: può accadere, infatti, che a livello di singola impresa (ovvero di questionario) le informazioni fornite nel complesso siano coerenti, ma contengano ugualmente degli errori.

I controlli vengono effettuati sia nella fase di raccolta, soprattutto web con i vincoli nella maschera. Ciò permette di eliminare già nella fase di compilazione alcuni tipi di errore.

Condizione fondamentale affinché un questionario sia utilizzabile è che il valore delle vendite del mese di riferimento sia maggiore di zero. Raramente si presentano casi in cui un valore delle vendite uguale a zero sia di fatto corretto; tali casi vanno approfonditi singolarmente in quanto possono riferirsi ad imprese con attività stagionale (ma in tal caso sono i controlli longitudinali che prevalgono) oppure essere determinati da altri fattori occasionali. In generale i valori anomali per effetto della stagionalità non producono variazioni errate (si ricorda che la stima della variazione tendenziale è l'obiettivo principale). Tuttavia va precisato che un valore delle vendite pari a zero viene comunque considerato anomalo e pertanto imputato.

Di seguito si cercherà di illustrare in modo più dettagliato l'insieme delle operazioni eseguite per l'individuazione di errori nei dati e la correzione degli stessi.

Per quanto riguarda i dati anagrafici dell'impresa, i controlli vengono svolti senza soluzione di continuità trasversalmente al processo produttivo. In altre parole ogni occasione di contatto con l'impresa (e non solo le occasioni d'indagine) può fornire elementi utili per la verifica di tali informazioni. Nella maggior parte dei casi sono le imprese a fornire volontariamente indicazioni circa le

variazioni anagrafiche ma in alcuni casi, a fronte di segnali significativi, si provvede a contattare le unità per la verifica delle informazioni.

Diverso è il caso delle verifiche sull'attività prevalente dell'impresa. Errori su questo tipo di informazione pregiudicano non soltanto il contatto con l'unità rispondente ma anche tutte le informazioni fornite nel complesso. Gli errori relativi alla classificazione ATECO vengono individuati prima di tutto verificando che ci sia coerenza tra: valore mensile delle vendite, gruppo di prodotti venduti, forma di vendita dell'impresa. Qualora venga riscontrato un errore di questo tipo si procede con un accertamento presso l'impresa.

A proposito degli errori relativi alle informazioni strutturali si ricorda che il campione della rilevazione viene aggiornato all'inizio di ciascun anno. Affinché l'assegnazione delle imprese ai diversi strati venga effettuata nel modo più corretto si cerca di effettuare il maggior numero di verifiche a priori (in fase di estrazione dell'unità). Il progressivo incremento della qualità dell'archivio ASIA ha portato alla riduzione di errori di tal genere.

La tabella a pagina seguente illustra, con riferimento ai mesi compresi nel periodo gennaio 2006-luglio 2007, la composizione delle unità campionarie in relazione al loro stato.

Si osservi come in corrispondenza dei dati di gennaio i valori riportati sono più bassi rispetto agli altri mesi a causa degli effetti derivanti dalla revisione del campione dell'indagine. In particolare possono essere considerate "risolte" con certezza tutte e sole le unità rispondenti e le stesse sono automaticamente considerate "eleggibili".

Negli stessi mesi invece non si hanno informazioni relative al numero di unità fuori campo di osservazione, cessate oppure che sono state interessate da variazioni di stato.

In generale occorre attendere almeno l'occasione d'indagine successiva, ovvero quella di febbraio per iniziare a raccogliere informazioni che riguardano lo stato delle unità.

Sia la percentuale di unità risolte sia la percentuale di unità eleggibili aumenta di mese in mese anche se alla fine di ciascun anno il numero di unità non risolte e non eleggibili risulta ancora non trascurabile.

Tavola 2: composizione del campione in relazione allo stato delle unità *(gennaio 2006-luglio 2007).*

Periodo		UNITA' DEL CAMPIONE						
		Totale unità	Unità risolte (%)	Unità eleggibili (%)	Unità Fuori Target (%)	Unità non più esistenti (%)	Variazioni di stato (%)	Tasso di risposta
2006	gennaio	7.936	45,58	45,58	-	-	-	45,58
	febbraio	7.936	55,24	54,32	0,89	0,46	0,32	42,5
	marzo	7.936	64,82	61,90	1,85	1,46	1,21	55,06
	aprile	7.936	68,50	64,76	2,23	1,82	1,42	52,61
	maggio	7.936	70,63	66,31	2,43	2,07	1,62	52,32
	giugno	7.936	71,88	67,55	2,38	2,03	1,60	53,1
	luglio	7.936	74,09	68,60	2,62	2,72	2,07	53,24
	agosto	7.936	75,40	69,12	2,69	3,13	2,52	52,74
	settembre	7.936	76,39	69,38	2,77	3,51	2,89	47,11
	ottobre	7.936	76,99	69,66	2,82	3,76	2,95	54,03
	novembre	7.936	77,71	69,80	2,87	4,05	3,26	52,12
	dicembre	7.936	78,18	70,00	2,93	4,19	3,34	48,04
2007	gennaio	7.990	34,94	34,94	-	-	-	34,94
	febbraio	7.990	55,98	55,31	0,30	0,18	0,20	51,29
	marzo	7.990	65,87	62,30	1,09	1,39	1,09	53,05
	aprile	7.990	70,24	64,56	1,43	2,55	1,70	49,10
	maggio	7.990	72,98	65,91	1,78	3,05	2,24	50,10
	giugno	7.990	75,32	67,51	1,90	3,35	2,55	53,09
	luglio	7.990	76,57	68,09	2,07	3,64	2,78	47,12

Per quanto riguarda le informazioni sulla forma di vendita dell'impresa si possono presentare errori qualora non vi sia coerenza tra i caratteri strutturali (es. superficie di vendita), l'attività economica e il numero e/o il tipo di prodotti venduti.

Questo carattere, tuttavia, può presentare errori anche a seguito della non univocità delle definizioni sottostanti. Ad esempio la superficie di vendita costituisce un carattere discriminante per l'individuazione delle diverse forme distributive, sia quelle più tradizionali sia quelle della Grande distribuzione. Di conseguenza non sono rari i casi in cui un'impresa si riconosce appartenente ad una data forma di vendita anche se dal punto di vista della definizione rigorosa ad essa corrispondente non lo sarebbe.

Il numero di addetti (dipendenti più indipendenti) dell'impresa commerciale, oltre ad essere una variabile di stratificazione, riveste particolare importanza nella fase di editing dal momento che ad essa si fa riferimento sia nel controllo della struttura dell'impresa, sia nel controllo del valore mensile delle vendite, sia nella stima dei dati mancanti (attraverso il rapporto fatturato per addetto sia a livello di singola unità sia a livello di strato). Nel caso in cui non si abbiano informazioni sul numero degli addetti (non reperibili neanche attraverso il contatto telefonico) si imputa il dato derivato dalla lista di estrazione. Valori apparentemente anomali per questa variabile in un dato mese potrebbero indicare un mutamento nella struttura dell'impresa stessa.

La variabile relativa al numero di punti di vendita dell'impresa non viene utilizzata per la stratificazione delle unità della popolazione di riferimento né rientra direttamente nel calcolo delle stime mensili. Tuttavia le informazioni ad essa relative sono particolarmente rilevanti per il controllo sulla coerenza dei dati a livello di singola unità.

Per ciò che riguarda più strettamente l'individuazione degli errori, con riferimento ai dati rilevati mensilmente, i metodi adottati si basano su regole deterministiche. In particolare per le verifiche che riguardano numero di addetti e del numero di punti di vendita si richiede che il rapporto tendenziale (ovvero il rapporto tra il dato corrente e quello relativo al corrispondente periodo dell'anno precedente) non superi (in valore assoluto) più del 10% il valore relativo al mese di riferimento. Per i dati relativi al valore delle vendite, invece, la regola deterministica di localizzazione degli errori porta comunque alla definizione di un intervallo di accettazione, i cui valori soglia sono calcolati con il metodo dei quartili (proposto da Hidiroglou e Berthelot 1986).

Una volta che un dato viene identificato come outlier viene corretto. In particolare gli errori relativi al numero di addetti e al numero dei punti di vendita vengono corretti sostituendo il valore errato con quello relativo allo stesso mese dell'anno precedente.

La variabile relativa al valore mensile delle vendite viene corretta e imputata come se fosse una mancata risposta.

L'imputazione delle mancate risposte relative alle variabili numero di addetti e numero di punti di vendita segue la medesima regola, ovvero il dato errato viene sostituito con quello relativo allo stesso mese dell'anno precedente. I dati mancanti relativi al valore mensile delle vendite sono invece imputati come segue: per l'insieme delle imprese rispondenti, il cui valore delle vendite sia compreso nell'intervallo di accettazione, si calcolano, per ciascuno strato, la media del valore delle vendite del mese corrente e la media del valore delle vendite del corrispondente mese dell'anno precedente. In questo modo, sempre per ciascuno strato, è possibile determinare il rapporto tendenziale come rapporto tra le due medie di cui sopra. Il dato errato viene dunque ristimato applicando la variazione data dal rapporto tendenziale medio di strato al rapporto fatturato per addetto relativo all'impresa moltiplicato per il numero degli addetti del mese. [84]

La tabella 3 riporta la percentuale di dati anomali rilevati per i diversi mesi del periodo gennaio 2006-luglio 2007 e l'indicazione di quanti di questi sono stati corretti interattivamente.

Se si effettua un confronto tra i mesi comuni nei due anni (ovvero gennaio-luglio) si può notare come la percentuale di dati anomali risulti nel 2007 leggermente più contenuta.

[84] L'ipotesi che questa regola sottintende è che il rapporto fatturato per addetto dell'impresa abbia subito, in termini tendenziali, la stessa variazione rilevata per lo strato di cui essa fa parte.

Tavola 3: percentuale di dati anomali individuati *(gennaio 2006-luglio 2007)*

	2006		2007	
Mese	**Percentuale di dati anomali**	**Di cui corretti con controllo interattivo**	**Percentuale di dati anomali**	**Di cui corretti con controllo interattivo**
gennaio	13,95	3,34	10,68	4,34
febbraio	13,87	3,63	16,48	3,04
marzo	17,01	3,63	17,06	3,60
aprile	16,23	4,11	14,47	4,58
maggio	15,31	3,79	15,36	3,75
giugno	15,71	3,61	15,79	3,57
luglio	15,49	3,74	14,09	4,09
agosto	15,94	3,95	-	-
settembre	13,62	4,35	-	-
ottobre	15,76	4,40	-	-
novembre	15,31	2,39	-	-
dicembre	13,23	3,52	-	-

4. Conclusioni

L'introduzione di processi automatizzati per il controllo e la correzione dei dati costituisce un elemento importante per la qualità dei risultati prodotti, pertanto sarebbe auspicabile in futuro ricorrere ad essi sempre più frequentemente. Tuttavia nel caso della rilevazione mensile sulle vendite al dettaglio i controlli interattivi hanno un ruolo fondamentale a causa delle peculiarità delle unità presso le quali si effettua l'indagine. Il miglioramento della tecnologia (soprattutto informatica) potrà garantire una sempre maggiore efficienza del processo favorendo l'integrazione delle varie fasi ed ottimizzando le operazioni di tipo interattivo.

Bibliografia

ISTAT. *La nuova indagine sulle vendite al dettaglio: aspetti metodologici e contenuti innovativi*, Metodi e norme, 3, Istat, Roma, 1998.

Brancato G., Carbini R. *Controllo di qualità e documentazione standard dei processi produttivi con SIDI*. Documento tecnico interno, Istat, Roma, 2006.

EUROSTAT. *Short-term Statistics Manual*, Eurostat, Lussemburgo, 2005*a*.

EUROSTAT. *Council Regulation No 1158/05 amending Council regulation No 1165/98 concerning short-term statistics*, Eurostat, Lussemburgo, 2005*b*.

EUROSTAT. *PEEIs in focus – Retail trade*, Eurostat, Lussemburgo, 2006.

ISTAT. *Manuale di tecniche d'indagine,* Vol. 4-5, Istat, Roma, 1989.

Hidiroglou M.A., Berthelot J.M. "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology*, 12, 73-84, Statistics Canada, Ottawa, 1986.

Indagine su fatturato e ordinativi: verso un sistema integrato per il controllo e correzione dati

Fabio Bacchini, Istat, Servizio Statistiche Congiunturali dell'Industria e delle Costruzioni
Roberto Iannaccone, Istat, Servizio Statistiche Congiunturali dell'Industria e delle Costruzioni
Enzo Salvatori, Istat, Servizio Statistiche Congiunturali dell'Industria e delle Costruzioni

Sommario: L'indagine mensile su fatturato ed ordinativi permette l'elaborazione ed il rilascio, tramite comunicato stampa, degli indicatori di fatturato ed ordinativi previsti dal regolamento europeo sulle statistiche congiunturali. In particolare, l'indice degli ordinativi costituisce uno degli indicatori economici principali (PEEIs) tra quelli monitorati a livello europeo per le decisioni di politica monetaria. L'indagine è tra quelle tradizionali nel dominio delle statistiche congiunturali. Le serie storiche sono disponibili, a partire dal gennaio 1990 (*con.istat.it*), fino al livello di gruppo di attività economica (3 digit). Per i rilascio dei dati è prevista, all'interno del processo di produzione, sia una fase di editing sia una fase di imputazione per le mancate risposte.
Accanto ai metodi tradizionali sviluppati nel corso del tempo e basati essenzialmente su un macroediting, negli ultimi mesi sono stati implementati dei controlli sistematici a livello micro per l'identificazione dei valori anomali e per la valutazione degli effetti del processo di imputazione. Tuttavia queste innovazioni vengono gestite con dei programmi creati ad hoc, esterni alla procedura informatica tradizionale. E' quindi necessaria una ulteriore fase di approfondimento che permetta l'implementazione del sistema di editing e imputazione all'interno del processo di produzione.)

Parole chiave: fatturato e ordinativi, editing, imputazione.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Introduzione

La rilevazione mensile sul fatturato e gli ordinativi permette l'elaborazione ed il rilascio, tramite comunicato stampa, degli indicatori di fatturato ed ordinativi previsti dal regolamento sulle statistiche congiunturali. La diffusione del comunicato stampa avviene solitamente a circa 50 giorni dal mese di riferimento (ad esempio gli indici per il mese di agosto 2007 sono stati pubblicati lo scorso 19 ottobre (figura 1[85])).
L'indice degli ordinativi elaborato dall'Italia concorre al rilascio del valore aggregato a livello europeo, uno degli indicatori principali a livello europeo (PEEIs) utilizzato per le analisi di politica economica (ad esempio il 23 ottobre è stato rilasciato il dato europeo per il mese di agosto (figura 2)).
L'intensità (11 comunicati l'anno, i mesi di giugno e luglio vengono pubblicati congiuntamente nel mese di settembre), la tempestività nel rilascio degli indicatori (50 giorni dal periodo di riferimento) e il numero di indici grezzi calcolati (1.989 indici articolati per gruppo ed aggregazioni superiori) richiedono sia un consistente impiego di risorse umane sia un sistema informativo in grado di governare, integrandole tra di loro, le diverse fasi del processo produttivo.
L'attuale procedura informatica definita ai tempi del porting per l'anno 2000, prevede una procedura gestita attraverso il modulo del programma SAS (denominato AF) che gestisce le maschere di dialogo collegate direttamente alle tabelle Oracle. Attraverso il menu principale delle procedura sono disponibili i moduli principali del processo produttivo: gestione dell'anagrafica, inserimento dei dati, stima delle mancate risposte, calcolo dei macrodati, calcolo degli indici.
Fino a poco tempo fa, l'attività di editing ed imputazione (d'ora in avanti E&I) era quasi esclusivamente legata a questi moduli sviluppati secondo lo schema descritto in figura 3. In particolare, la fase di editing era essenzialmente concentrata sull'analisi in serie storica dei macrodati (ultimi 13 mesi) per gruppo di attività economica congiuntamente ai valori degli indici calcolati successivamente all'imputazione delle mancate risposte.
Negli ultimi mesi il processo di editing è stato esteso anche ai microdati, sia nella fase precedente la stima delle mancate risposte, per valutare l'esistenza di errori di imputazione e/o valori anomali, sia dopo l'imputazione per verificarne l'effetto. Le informazioni acquisite in questa fase, permettono di integrare ed in molti casi agevolare la lettura dei controlli macro.
Questi miglioramenti sono stati accompagnati da significative innovazioni informatiche che da un lato permettono l'immediata lettura dei dati per singola impresa, realizzando il link con il modello originario (fax, mail, web), dall'altro rendono automatico il caricamento delle informazioni acquisite in forma telematica.
Accanto alle attività descritte, è stata sviluppata una attività di editing mirata al miglioramento della qualità dei dati di fatturato e ordinativi riferiti alla zona euro. I dati necessari al calcolo dei rispettivi indicatori, sono stati inizialmente acquisiti in via sperimentale e successivamente validati attraverso i microdati di impresa dell'indagine sul commercio estero (il progetto è stato in parte finanziato da Eurostat). I risultati positivi ottenuti (si veda Bacchini, 2006), costituiscono il presupposto per un suo futuro tentativo di standardizzazione con immediati vantaggi sia in termini di qualità dei dati che di riduzione dell'onere statistico.
L'insieme delle attività presentate hanno in comune l'obiettivo di migliorare, nel complesso, la gestione del processo di produzione. In questa fase, la riflessione sugli strumenti utilizzati per le fasi di E&I, costituisce un passo importante. Una loro sintesi in un quadro unitario, stabilmente inserito all'interno del processo di produzione e descritto da una appropriata documentazione è il prossimo obiettivo.
Il lavoro è articolato come segue. Nel paragrafo due si descrive la rilevazione; nel tre il precedente sistema di E&I; nel quattro le recenti innovazioni ed infine il paragrafo 5 è dedicato ai problemi aperti e alle prospettive di lavoro.

[85] Tutte le figure sono riportate in coda al lavoro

2. Descrizione della rilevazione

Il campo di osservazione per la rilevazione sul fatturato comprende le attività economiche delle industrie estrattive e manifatturiere. Sono escluse le industrie dell'energia elettrica, gas ed acqua. Solo per un sottoinsieme di queste attività vengono richieste anche le informazioni sugli ordinativi. L'identificazione delle attività è armonizzata a livello europeo ed è definita all'interno del regolamento sulle statistiche congiunturali (per il dettaglio si veda la tabella 14 delle Note metodologiche allegate al comunicato stampa mensile).
Per entrambe le variabili, l'unità di rilevazione è l'impresa. Tuttavia nel caso di imprese il cui fatturato/ordinativo si riferisce a differenti attività economiche (a livello di 3 digit della classificazione Ateco), è richiesto il dettaglio dei dati per singola "unità funzionale".
Il panel delle imprese selezionate per l'indagine è estratto in modo ragionato dall'universo delle imprese con più di 20 addetti. La scelta del campione di imprese è realizzata a livello di gruppo di attività economica selezionando un numero di imprese tali che il loro fatturato totale rappresenti almeno il 70% del fatturato totale del gruppo. La numerosità del campione all'anno base (il 2000) era pari a circa 6.000 unità. La progressiva erosione dovuta sia alla persistenza di comportamenti di mancata risposta sia alla cessazione di attività ha portato la numerosità nel 2005 a 5.802 unità[86].
Per fatturato si intende l'ammontare del valore risultante da tutte le fatture emesse nel mese per vendite sul mercato interno, su quello estero e su quello della zona euro, al netto dell'Iva fatturata ai clienti e degli abbuoni e sconti esposti in fattura e al lordo delle spese (trasporti, imballaggi, ecc.) e delle altre imposte (per es. imposte di fabbricazione) addebitate.
Per gli ordinativi è rilevato l'ammontare di quelli nuovi assunti ed accettati definitivamente nel corso del mese. Nel caso che alcuni ordinativi siano stati commissionati soltanto in termini di quantità (es. tonnellate di filati, numero di pezzi, ecc.) viene richiesta la quantificazione in termini di valore in base ai prezzi medi correnti di vendita. I dati vanno distinti a seconda che gli ordini provengano da clienti nazionali o da clienti esteri. Per il mercato estero, va riportato sia il valore totale che quello parziale relativo ai clienti dei Paesi appartenenti alla Zona Euro.
Tutte le variabili vengono richieste alle imprese attraverso il modello di rilevazione riportato in figura 4 (per semplicità riferito ai soli mesi di gennaio, febbraio, marzo e aprile, nel caso di impresa operante anche su ordinativi). Nel dettaglio ad inizio anno ad ogni impresa appartenente al campione viene inviata la circolare, le istruzioni ed il modello di rilevazione riferito ai 12 mesi dell'anno, richiedendo l'aggiornamento mensile dei dati entro il giorno 10 del mese successivo a quello richiesto.
La trasmissione dei dati dalle imprese all'Istat avviene attraverso tre diversi canali. Il tradizionale invio tramite fax a dei numeri gestiti dal servizio fax-server (circa il 66% dei casi), quello tramite e-mail (circa il 21% dei casi) e, a partire dal 2007, quello attraverso l'uploading del modello in formato excel sul sito *indata.istat.it* (circa il 13%).
L'attività di sollecito ai non rispondenti avviene tramite il servizio Pt-fax con un invio a circa 35 giorni dal mese di riferimento. Infine a circa 47 giorni viene chiuso il comunicato stampa.

3. L'attuale sistema di editing e imputazione

La procedura per il calcolo degli indici di fatturato e ordinativi è gestita attraverso un'interfaccia realizzato in SAS/AF che dialoga con tabelle Oracle. Il menu principale (figura 5) prevede, tra l'altro, le operazioni per la gestione dei questionari, per le stime ed il controllo, per il calcolo dei macrodati e degli indici e la produzione dei report.
E' importante sottolineare che per elaborare gli indici mensili, l'attuale procedura richiede i valori delle variabili oggetto di studio, per ciascuna delle imprese appartenenti al panel. Questo implica la necessità di stimare le mancate risposte al momento della chiusura dell'indice (blocco 2 nella figura 3).

[86] Per ridurre il fenomeno dell'attrition e per migliorare la rappresentatività del campione selezionato, a partire dal 2007 ha avuto inizio una attività annuale di verifica dell'allineamento del panel rispetto all'Archivio statistico delle imprese attive (Asia).

3.1 La stima delle mancate risposte

La procedura prevede due tipi di stima, automatica oppure manuale. Il primo caso, che riguarda quasi la totalità delle mancate risposte, definisce il valore della stima pari alla variazione tendenziale del valore del fatturato per le imprese rilevate appartenenti allo stesso gruppo di attività economica. Indicando con i_G la generica impresa da stimare, con G il gruppo di attività economica, con G_R il sottoinsieme di imprese rispondenti al tempo t si ha:

$$fatt_{i_G,t} = \frac{\sum_{i \in G_r} fatt_{i,t}}{\sum_{i \in G_r} fatt_{i,t-12}} fatt_{i_G,t-12}$$

La stima automatica viene gestita dalla procedura attraverso un apposito bottone all'interno del modulo "stime/controllo". La stima manuale riguarda esclusivamente i casi in cui, per l'impresa da stimare, non si dispone dell'informazione al tempo t-12. Questo capita in presenza di nuove imprese. In questi casi si acquisiscono direttamente dalla impresa le informazioni utili alla stima.

3.2 L'editing dei macro dati e degli indici

Stimate le mancate risposte è possibile procedere al calcolo dei macrodati e degli indici per gruppo di attività economica (figure 6 e 7 per l'ateco 272) e per le aggregazioni successive. La procedura offre la possibilità di visualizzare o stampare gli aggregati.

Le tabelle dei macrodati analizzate congiuntamente a quelle degli indici permettono di valutare il peso dell'imputazione, ovvero la presenza di indicatori fuori range. Alla fine della figura 6 è riportato il valore della collaborazione (copertura) espressa come rapporto tra il valore rilevato ed il totale (rilevato + stimato) del fatturato nazionale, nell'esempio pari a 96.5%. Dalla figura emerge anche un'altra particolarità dell'attuale processo di imputazione: per alcuni settori, i dati degli ordinativi sono stimati come uguali al valore del fatturato.

Per quanto riguarda gli indici, per ciascuna delle variabili viene evidenziata con un asterisco una variazione tendenziale maggiore del 10% (figura 7).

4 Le recenti innovazioni

Le tradizionali attività di controllo basate sui macrodati e sull'analisi delle variazioni tendenziali per singolo indicatore e gruppo di attività economica sono state affiancate da un processo di microediting rivolto alla valutazione sia degli errori di inserimento dati che degli effetti del processo di imputazione. Il microediting è ormai parte integrante del processo di elaborazione degli indici di fatturato e ordinativi anche se, allo stato attuale, viene gestito esternamente alla procedura con programma ad hoc (paragrafo 4.1).

Inoltre, con riferimento ai valori del fatturato e degli ordinativi per la zona Euro, raccolti sperimentalmente a partire dal 2002 ed utilizzati per il calcolo e l'invio, in forma confidenziale, dei corrispondenti indici ad Eurostat, è stato realizzato un linkage a livello di impresa con i dati raccolti dall'indagine mensile sul commercio estero. Il linkage ha permesso il controllo della qualità e la validazione dei microdati (paragafo 4.2).

Infine, particolare attenzione è stata dedicata all'implementazione dei programmi disponibili per il supporto alla attività di microediting e all'acquisizione automatica dei dati acquisiti per via telematica (paragrafo 4.3).

4.1 Il microediting

La gestione della procedura secondo le linee tracciate nel diagramma presentato in figura 3 è stata parzialmente rivista con un duplice obiettivo: dedicare maggiore attenzione all'analisi micro prima della stima delle mancate risposte; effettuare una valutazione del processo di stima automatica.

Nella settimana precedente alla chiusura dell'indice, viene effettuata un'analisi delle mancate risposte con lo scopo di realizzare dei solleciti telefonici mirati. Vengono individuate le imprese più rilevanti che, nei mesi immediatamente precedenti, avevano collaborato alla rilevazione. Allo stesso tempo vengono effettuate delle ricerche per individuare le imprese che negli ultimi mesi presentano un break nella collaborazione. Ad esempio per il mese di agosto 2007 sono state effettuati circa 50 solleciti mirati.
Parallelamente viene effettuate un controllo sui dati pervenuti. Il controllo riguarda tutte le imprese in cui il valore di una delle variabili è superiore ai 5 milioni di euro[87]. Vengono, quindi, individuate le imprese con una variazione tendenziale maggiore del 30% per le quali si procede all'accertamento dei dati. Per il mese di agosto 2007, questa procedura ha portato a 112 controlli, che hanno permesso di evidenziare 5 errori di imputazione, individuati attraverso il ritorno al modello, e tre errori significativi di compilazione del questionario da parte di altrettante imprese, accertati mediante il ricontatto dell'impresa. Per circa altri 10 casi è stato possibile acquisire informazioni complementari utili alla interpretazione di valori delle variazioni tendenziali particolarmente rilevanti.
Dopo questi controlli si è proceduto alla stima delle mancate risposte. Per i microdati imputati si è realizzato un controllo a livello micro speculare a quello realizzato per i dati rilevati. Per il mese di agosto 2007, questa procedura ha portato ad ulteriori 14 controlli che hanno determinato una correzione per gli ordinativi di una impresa, per un settore in cui il valore era posto uguale a quello del fatturato.

4.2 Il linkage con i dati del commercio estero

Nel periodo 2005-06 Eurostat ha concesso un finanziamento all'Istat con l'obiettivo di procedere al rilascio degli indicatori di fatturato ed ordinativi esteri separatamente per la zona euro e non euro. Sono stati acquisiti i microdati sulle esportazioni nell'ambito della rilevazione mensile sul commercio estero.
Il lavoro è proceduto in due fasi. Le imprese sono state collegate sulla base del codice Asia. Successivamente si è proceduto ad una riaggregazione delle informazioni del commercio estero secondo il codice di attività economica dell'impresa e sono state comparate le variazioni tendenziali.
I risultati ottenuti hanno permesso di effettuare alcune correzioni per i valori riferiti alla zona euro (prevalentemente situazioni in cui l'impresa comunicava un valore pari a zero per un determinato periodo).
Nel dettaglio l'operazione a livello micro ha permesso, per il periodo 2002-2005, di collegare in media l'82.9% delle imprese (figura 8). Anche i risultati del confronto delle variazioni tendenziali per sottosezione sono stati positivi (figura 9). Per la sottosezione 21 il confronto riportato ha permesso di individuare alcuni errori a livello di impresa la cui correzione determina un significativo miglioramento dell'accostamento tra le due serie.

4.3 I nuovi strumenti informatici

4.3.1 La riaggregazione per unità delle informazioni disponibili

L'intensificazione dei controlli per microdato ha reso necessario poter disporre di un pannello di controllo in grado di facilitare la consultazione di tutte le informazioni riferite ad una singola ditta. Per soddisfare questa richiesta è stato sviluppato un modulo in linguaggio php, in linea quindi con lo sviluppo in istituto di applicazioni open-source. Una volta identificata l'impresa oggetto di indagine (figura 10, le chiavi di ricerca sono il codice asia, il codice nai e la denominazione) è possibile accedere (figura 11, bottoni in alto) alle informazioni sui dati (serie storica di tutte le variabili a partire dal 2000, differenziando il colore nel caso di dato imputato), agli attributi anagrafici, all'identificativo del referente per la rilevazione, alle caratteristiche della spedizione della circolare, alle credenziali necessarie per accedere ai servizi telematici.
Il pannello di controllo permette anche il collegamento, per ciascuna impresa, con il modello pervenuto, sia esso acquisito via fax o via e-mail/web (figura 12). In questo modo per ogni impresa si ha, in tempo reale, un quadro su tutti gli aspetti necessari alla intepretazione del dato osservato.

[87] Un analogo controllo viene realizzato a cadenza trimestrale per le imprese al di sotto della soglia specificata.

4.3.2 Il caricamento automatico dei dati pervenuti per via telematica

L'acquisizione telematica dei dati attraverso la funzione di uploading del modello in formato excel sul sito *indata*, ha portato alla realizzazione di un software in grado di archiviare direttamente i dati e proporli all'operatore per la convalida. Il programma, anche esso sviluppato in php[88] è stato esteso anche ai modelli che pervenuti attraverso e-mail[89].

Il programma prevede l'interpretazione delle informazioni acquisite per via telematica, raggruppandole in 3 sottoinsiemi: notizie di carattere anagrafico, dati sulle variabili della rilevazione, eventuali messaggi. Le informazioni vengono ripartite tra gli operatori secondo il codice di gruppo di attività economica. Ogni operatore può quindi collegarsi per analizzare i dati pervenuti (figura 13), selezionare i microdati per mese e ditta e valutare il suo inserimento (figura 14). Al momento dell'inserimento è anche possibile accedere a tutte le informazioni inviate nel testo dell'e-mail ed eventualmente aprire direttamente il modello in formato excel.

La procedura è in fase di test presso gli operatori ed è previsto il suo rilascio nei primi mesi del 2008. Il risultato atteso è duplice: riduzione degli errori di inserimento e diminuzione dei tempi necessari all'acquisizione dei dati che, al momento, vengono inseriti tramite procedura direttamente dall'operatore.

5 Problemi aperti e prospettive a breve

L'introduzione di una analisi sistematica a livello micro per l'identificazione e la gestione di eventuali valori anomali nella rilevazione sul fatturato e ordinativi, ha portato considerevoli vantaggi rispetto alla tradizionale valutazione basata sui macrodati e gli indici.

Queste implementazioni sono state supportate da novità sia nel processo di acquisizione dei dati, con l'introduzione della trasmissione via web e l'acquisizione dei dati pervenuti per via telematica sia nella riaggregazione per ciascuna impresa di tutte le informazioni disponibili.

Rimangono ancora diversi punti da approfondire per giungere ad un sistema integrato, come ad esempio:

- introduzione del microediting e della gestione dei solleciti a livello di singolo operatore, mantenendo segnali a livello centralizzato;
- valutazione dell'impatto della stima delle mancate risposte;
- dell'attuale procedura di stima per le mancate risposte degli ordinativi;
- sperimentazione di un link sistematico con la rilevazione mensile del commercio estero con immediati vantaggi in termini di qualità dei dati;
- necessità di una documentazione della fase di *E&I* in linea con gli standard qualitativi richiesti.

Ringraziamenti

Il lavoro, originariamente preparato per il seminario 'Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali', non sarebbe stata scritta senza la disponibilità e la pazienza di Orietta Luzi, Marco Di Zio, Ugo Guarnera e Antonia Manzari che hanno dato vita al gruppo di lavoro e organizzato il seminario. Ringraziamo inoltre Giulio Barcaroli e l'anonimo referee per i commenti ricevuti.

L'elaborazione degli indici del fatturato e degli ordinativi è possibile grazie al contributo di tutte le persone che compongono l'unità operativa SCI/A: Daniela Birzò, Antonella Catapano, Paola Chialastri, Lucia Curradi, Angela Cuzari, Patrizia De Horatis, Stefania Ducci, Maria Cristina Mastrolillo, Giancarlo Nuccilli, Maria Vittoria Quintili e Ornella Trabalza e di Stefania Nappi (unità operativa SCI/D).

[88] Il programma è stato sviluppato in collaborazione con Lamberto Franci

[89] In linea con le strategie dell'Istituto in termini di privacy, è obiettivo dell'unità operativa dismettere quanto prima questo tipo di acquisizione dei dati

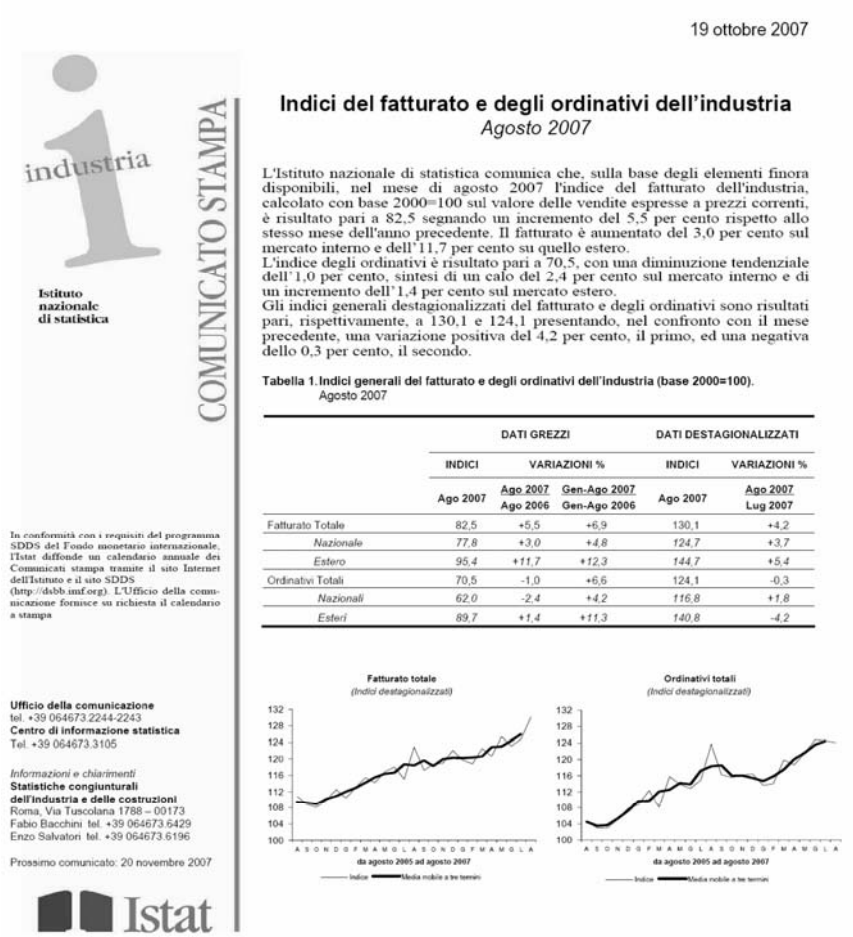
Bibliografia

Bacchini F. *Progetto Eurozone/non Eurozone breakdown for industrial turnover and new orders. Rapporto finale di ricerca*. Istat, 2006.

Barcaroli G., O. Luzi e C. Ceccarelli. *Il macroediting: tecniche di correzione interattiva di variabili quantitative guidata dall'analisi degli aggregati. Il caso del sistema dei conti delle imprese*. Quaderni di ricerca Istat n. 1/1998 Roma, Istat, 1998.

Istat, CBF, SFSO. *Recommended practices for editing and imputation in cross sectional business surveys. Rapporto finale di ricerca.* Roma: Istat, 2007.

Figura 1: *Comunicato stampa fatturato e ordinativi mese di agosto 2007*



19 ottobre 2007

industria

COMUNICATO STAMPA

Istituto nazionale di statistica

Indici del fatturato e degli ordinativi dell'industria

Agosto 2007

L'Istituto nazionale di statistica comunica che, sulla base degli elementi finora disponibili, nel mese di agosto 2007 l'indice del fatturato dell'industria, calcolato con base 2000=100 sul valore delle vendite espresse a prezzi correnti, è risultato pari a 82,5 segnando un incremento del 5,5 per cento rispetto allo stesso mese dell'anno precedente. Il fatturato è aumentato del 3,0 per cento sul mercato interno e dell'11,7 per cento su quello estero.
L'indice degli ordinativi è risultato pari a 70,5, con una diminuzione tendenziale dell'1,0 per cento, sintesi di un calo del 2,4 per cento sul mercato interno e di un incremento dell'1,4 per cento sul mercato estero.
Gli indici generali destagionalizzati del fatturato e degli ordinativi sono risultati pari, rispettivamente, a 130,1 e 124,1 presentando, nel confronto con il mese precedente, una variazione positiva del 4,2 per cento, il primo, ed una negativa dello 0,3 per cento, il secondo.

Tabella 1. Indici generali del fatturato e degli ordinativi dell'industria (base 2000=100). Agosto 2007

	DATI GREZZI			DATI DESTAGIONALIZZATI	
	INDICI	VARIAZIONI %		INDICI	VARIAZIONI %
	Ago 2007	Ago 2007 / Ago 2006	Gen-Ago 2007 / Gen-Ago 2006	Ago 2007	Ago 2007 / Lug 2007
Fatturato Totale	82,5	+5,5	+6,9	130,1	+4,2
Nazionale	*77,8*	*+3,0*	*+4,8*	*124,7*	*+3,7*
Estero	*95,4*	*+11,7*	*+12,3*	*144,7*	*+5,4*
Ordinativi Totali	70,5	-1,0	+6,6	124,1	-0,3
Nazionali	*62,0*	*-2,4*	*+4,2*	*116,8*	*+1,8*
Esteri	*89,7*	*+1,4*	*+11,3*	*140,8*	*-4,2*

Fatturato totale (Indici destagionalizzati)

Ordinativi totali (Indici destagionalizzati)

da agosto 2005 ad agosto 2007

Indice — Media mobile a tre termini

In conformità con i requisiti del programma SDDS del Fondo monetario internazionale, l'Istat diffonde un calendario annuale dei Comunicati stampa tramite il sito Internet dell'Istituto e il sito SDDS (http://dsbb.imf.org). L'Ufficio della comunicazione fornisce su richiesta il calendario a stampa

Ufficio della comunicazione
tel. +39 064673.2244-2243
Centro di informazione statistica
Tel. +39 064673.3105

Informazioni e chiarimenti
Statistiche congiunturali dell'industria e delle costruzioni
Roma, Via Tuscolana 1788 – 00173
Fabio Bacchini tel. +39 064673.6429
Enzo Salvatori tel. +39 064673.6196

Prossimo comunicato: 20 novembre 2007

Istat

Figura 2: *Euroindicators - nuovi ordinativi mese di agosto 2007*

eurostat EURO-INDICATORS news release

143/2007 - 23 October 2007

August 2007 compared to July 2007

Industrial new orders up by 0.3% in euro area

Up by 1.0% in EU27

The **euro area**[1] (EA13) industrial new orders index[2] rose by 0.3% in August 2007 compared with July 2007. The index fell by 2.6% in July[3]. **EU27**[1] new orders increased by 1.0% in August 2007, after a decrease of 3.5% in July[3]. Excluding ships, railway and aerospace equipment[4] industrial new orders gained 0.3% in the **euro area** and 0.9% in the **EU27** in August 2007.

In August 2007 compared with August 2006, industrial new orders increased by 5.1% in the **euro area** and by 8.2% in the **EU27**. Total industry excluding ships, railway and aerospace equipment grew by 6.3% in the **euro area** and by 9.8% in the **EU27**.

These estimates are released by **Eurostat, the Statistical Office of the European Communities**.

Figura 3: *Indagine F&O schema processo di produzione*

Figura 4: *Indagine F&O modello di rilevazione*

ISTITUTO NAZIONALE DI STATISTICA
Mod. ISTAT/SCI/FO

INDAGINE MENSILE SUL FATTURATO E GLI ORDINATIVI DELL'INDUSTRIA

CODICE DITTA	/0 ATTIVITÀ ECONOMICA

Anno 2007 — Valori in **migliaia** di euro

VOCI	GENNAIO	FEBBRAIO	MARZO	APRILE
Fatturato nazionale				
Fatturato estero				
di cui a clienti **Zona Euro**				
Ordini nazionali				
Ordini esteri				
di cui da clienti **Zona Euro**				

Figura 5: *Indagine F&O menu principale del sistema informativo*

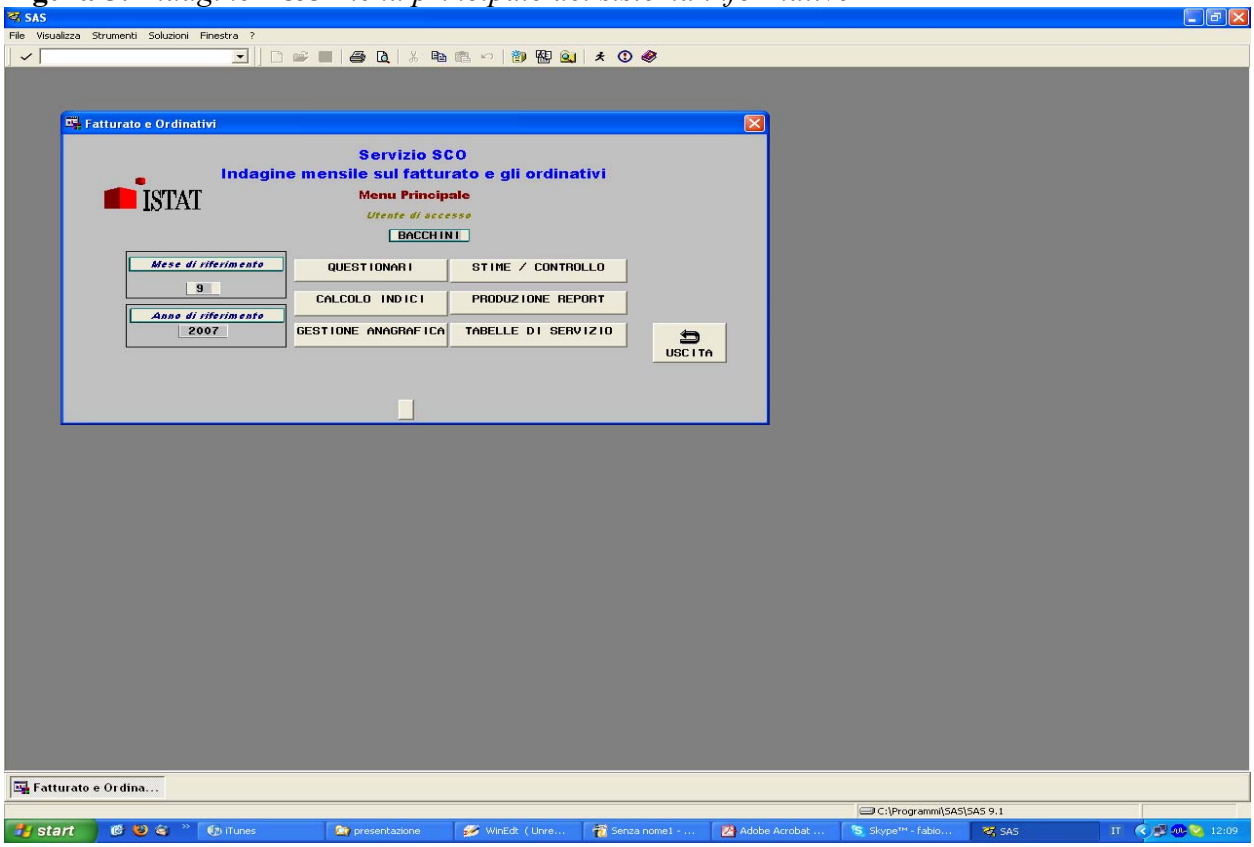


Figura 6: *Indagine F&O macrodati chiusura preliminare agosto 2007 - ateco 272*

CODICE ATECO 272													
TOTALI													
fn	127641	271370	280059	278502	185092	264888	292099	325255	258679	315639	281050	271542	86311
fe	104473	250461	251573	262380	188995	316551	300646	300986	258027	320831	309008	279537	145158
fc	86270	172122	175638	183753	94127	189038	182699	186697	175141	192003	182153	174214	83529
onz	131645	270476	277038	271839	185729	266522	293079	320416	256587	316827	273285	267299	87610
oe	101588	248037	243986	264831	190997	317117	299090	309066	249986	316574	306396	273195	140326
oc	83340	167355	168217	181822	95914	187646	180352	189148	168371	187907	179865	169320	81372
cn	141513	141719	138698	132035	132672	134136	134960	130168	128417	129605	120682	116821	118460
ce	103621	103243	98033	102651	105223	107508	109296	117421	109387	106424	99833	93491	91524
cc	77623	76559	70995	73731	75733	78288	78145	82747	78502	76852	69916	65022	63325
STIMATE													
fn	58668	4996	3253	6197	11263	26824	24460	22601	12491	18501	8661	14870	3035
fe	18796	5350	1602	7900	10365	27236	19377	15878	11355	15106	21144	19822	3105
fc	9507	3962	693	4064	9502	14666	12397	13434	9405	8794	12912	15597	2170
onz	116483	229298	237791	234716	164840	232376	266940	287769	228338	259322	238098	230181	75848
oe	88414	219568	220140	233794	175724	287982	274728	272290	228684	212426	282369	253093	128245
oc	71534	143672	146939	157922	82391	163111	157186	160920	148432	149793	158302	150899	71226
cn	141513	141719	138698	132035	132672	134136	134960	130168	128417	106012	120292	116821	118460
ce	103621	103243	98033	102651	105223	107508	109296	117421	109387	85009	99801	93491	91524
cc	77623	76559	70995	73731	75733	78288	78145	82747	78502	74565	69884	65022	63325
COPERTURA	96.5												

Figura 7: *Indagine F&O macrodati chiusura preliminare agosto 2007 - ateco 272*

	Aug-06	Sep-06	Oct-06	Nov-06	Dec-06	Jan-07	Feb-07	Mar-07	Apr-07	May-07	Jun-07	Jul-07	Aug-07	
fatt naz	69.2	147.2	151.9	151.1	100.4	143.7	158.5	176.5	140.3	171.2	152.5	147.3	46.8	
- var. tend.	48.2	14.6	29.3	28.1	10.2	32	26.8	28.4	27.2	16.7	-4.4	-13.6	-32.4	*
fatt est	99.6	238.9	239.9	250.2	180.2	301.9	286.7	287	246.1	306	294.7	266.6	138.4	
- var. tend.	36.1	44.4	30.7	37.2	40.2	74.2	43.9	36.1	49.4	45.7	54.1	39.7	39	*
fatt euro	79.5	178.3	181.8	184.7	127.5	197.4	202	214	176.2	217	200.8	187.8	77.9	
- var. tend.	42.7	26.5	29.9	32.1	22.8	50.9	34.5	31.8	36.8	29	17.8	5.9	-2	
ord naz	71	146	149.5	146.7	100.2	143.8	158.2	172.9	138.5	171	147.5	144.2	47.3	
- var. tend.	49.5	13.4	29.5	23.6	8.8	31.6	27.3	26.9	24.3	18.7	-9.1	-14.7	-33.4	*
ord est	95.8	233.9	230.1	249.7	180.1	299	282	291.5	235.7	298.5	288.9	257.6	132.3	
- var. tend.	33.1	42.6	28.5	38.6	40.7	74.3	42.6	38.2	43.5	44.1	53.5	39.7	38.1	*
ord euro	79.5	176	177	181.8	127.5	196.7	200.4	213.3	171.7	214.5	195.7	182.9	76.3	
- var. tend.	42.2	25.1	29.1	30.2	22.2	50.7	34.2	31.9	32.7	29.5	14.4	4.9	-4	

Figura 8: *Indagine F&O e commercio estero percentuale imprese collegate per sottosezione*

codice ateco	2002	2003	2004	2005	media
14	43.9	44.7	40.4	36.8	41.4
15	78.2	80.5	76.7	74.3	77.4
16	52.9	11.8	11.8	11.8	22.1
17	84.7	85.7	84.4	81.4	84.0
18	93.7	92.7	88.3	86.8	90.4
19	90.7	89.6	88.5	86.0	88.7
20	76.1	75.7	66.7	65.8	71.1
21	88.7	88.0	81.0	81.0	84.7
22	67.9	74.3	58.8	54.0	63.8
23	81.6	78.9	78.9	71.1	77.6
24	90.6	92.7	90.6	90.3	91.1
25	91.8	92.2	90.5	90.8	91.3
26	71.1	70.8	66.8	66.1	68.7
27	90.6	92.8	88.5	86.4	89.6
28	83.6	85.3	80.8	79.7	82.3
29	93.2	93.6	92.5	91.1	92.6
30	77.8	88.9	88.9	88.9	86.1
31	90.7	91.8	90.7	88.5	90.4
32	91.7	88.9	86.1	81.9	87.2
33	92.9	95.3	93.7	90.6	93.1
34	91.2	91.2	89.6	89.6	90.4
35	79.3	80.5	79.3	72.4	77.9
36	91.6	92.6	89.4	85.2	89.7
37	33.3	38.1	31.7	34.9	34.5
total	84.4	85.3	81.9	80.0	82.9

Figura 9: *Indagine F&O e commercio estero variazioni tendenziali zona euro per sottosezione*

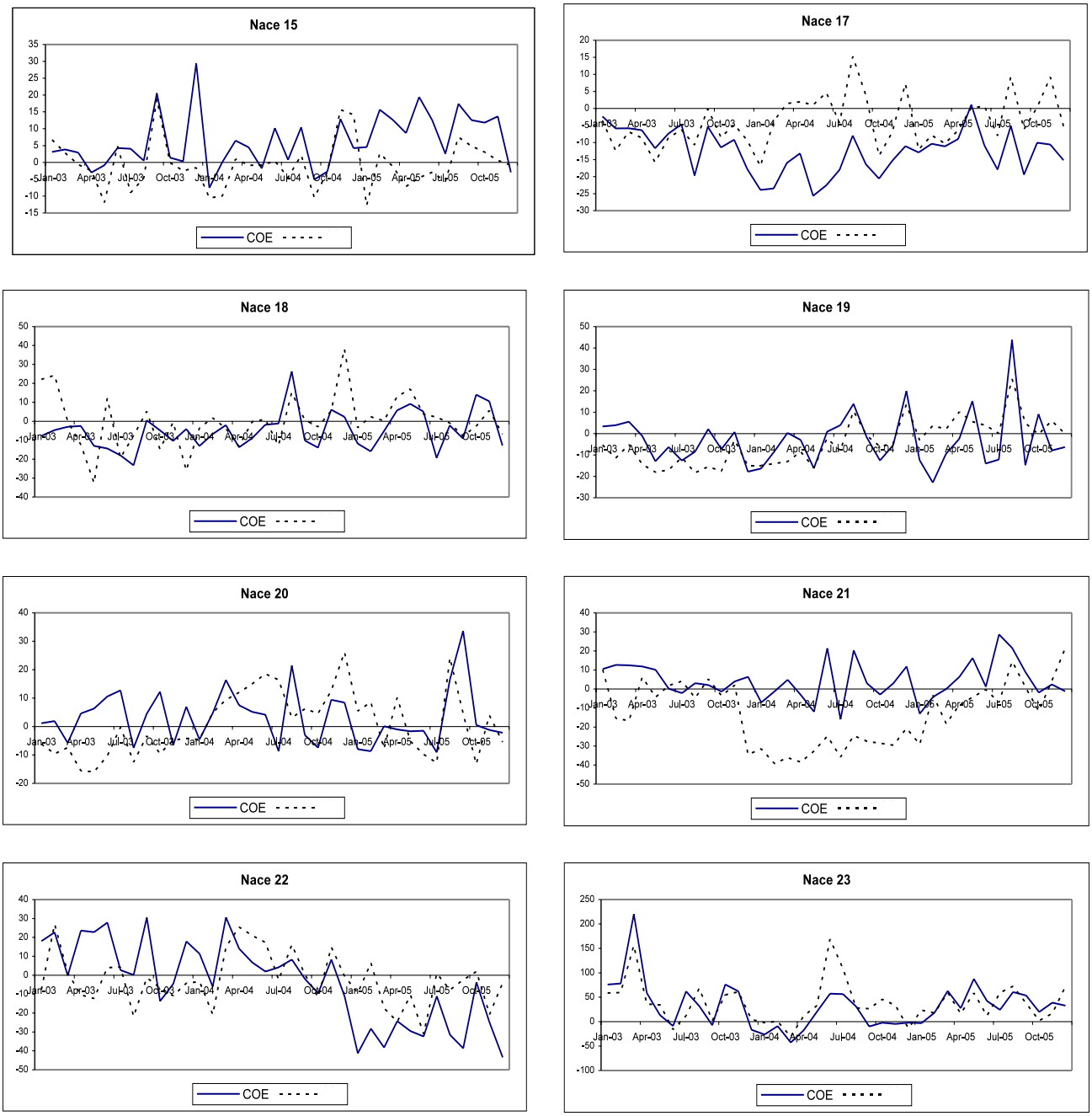


Figura 10: *Indagine F&O gestione delle informazioni per la singola impresa*

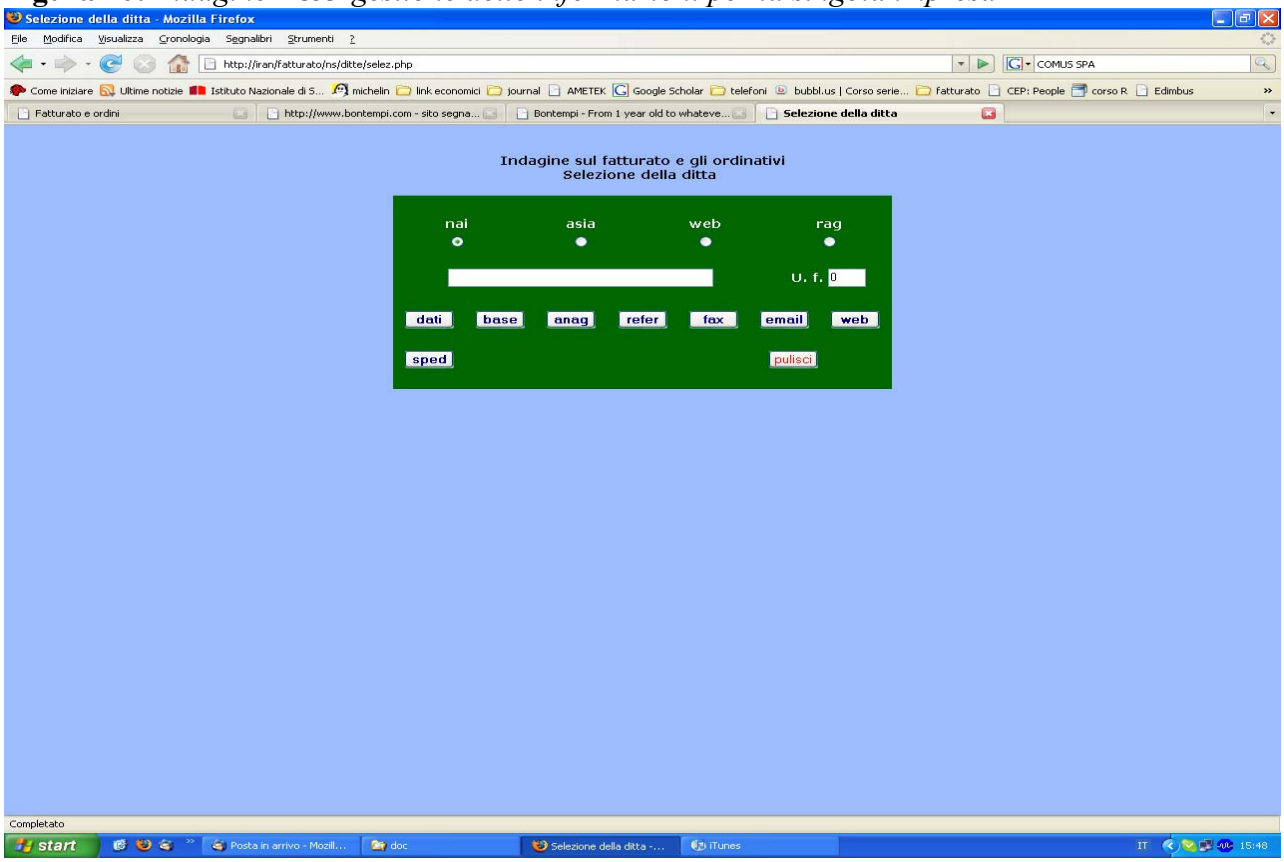


Figura 11: *Indagine F&O collegamento ai fax pervenuti*

Figura 12: *Indagine F&O collegamento ai modelli arrivati via mail o web*

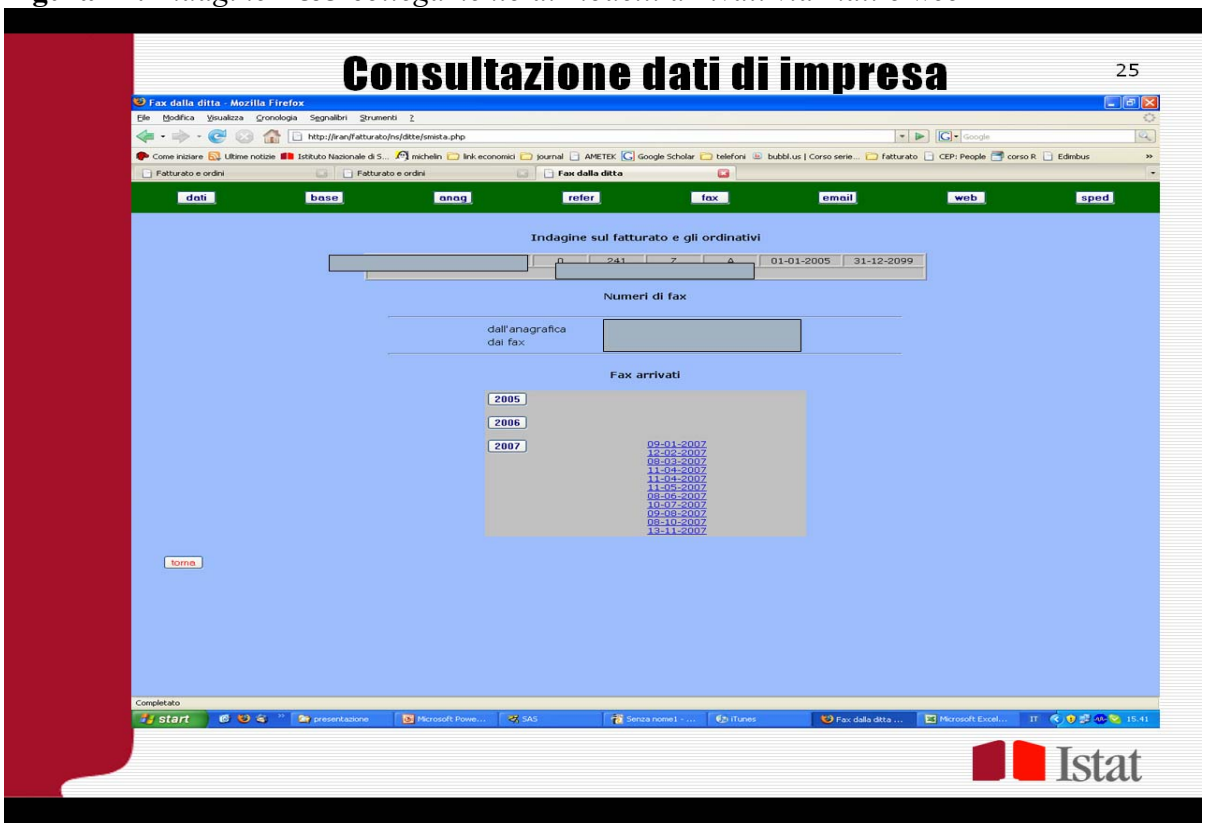


Figura 13: *Indagine Indagine F&O lista dati inviati per mail o web*

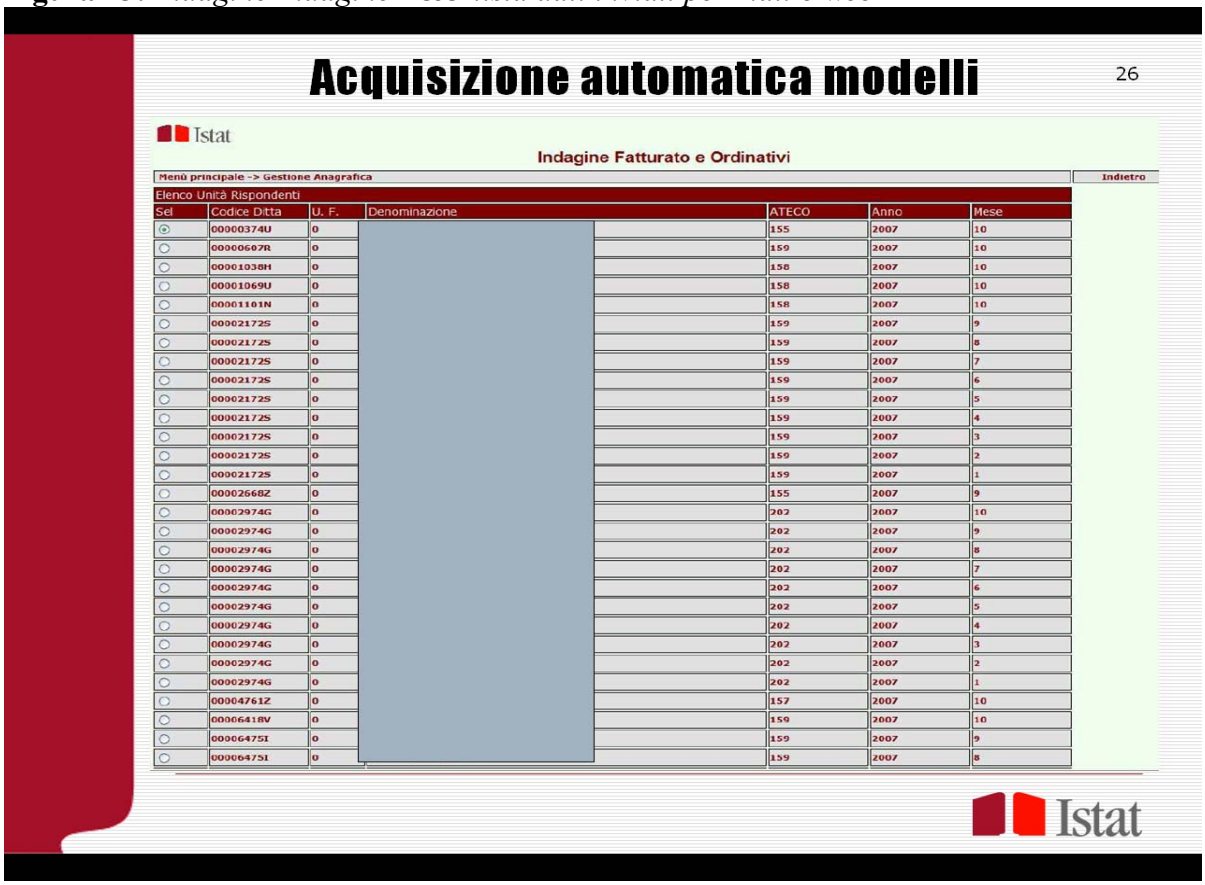
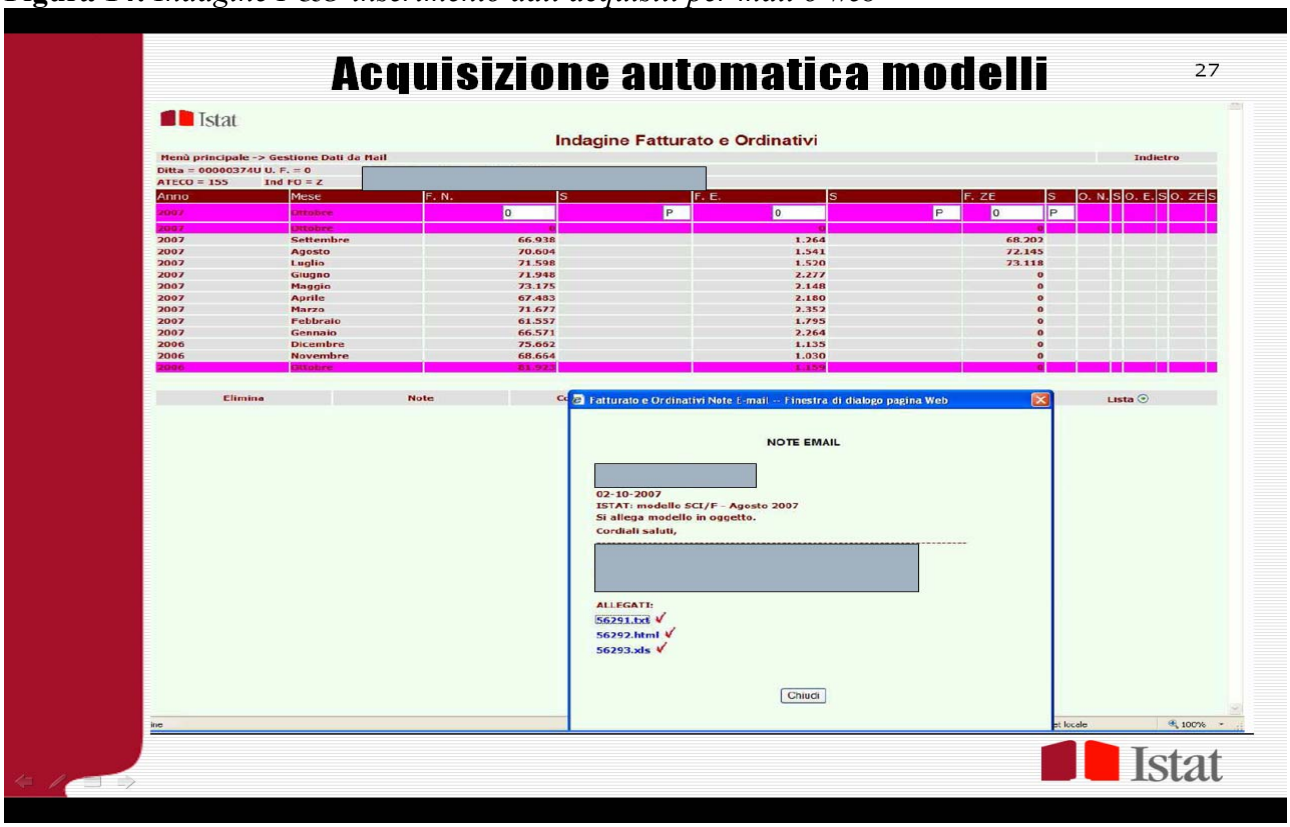


Figura 14: *Indagine F&O inserimento dati acquisiti per mail o web*



Metodi di controllo e correzione dei dati nell'indagine mensile sulla produzione industriale: stato attuale e possibili sviluppi

Annarita Mancini, Istat, Servizio Statistiche congiunturali dell'industria e delle costruzioni
Teresa Gambuti[90], Istat, Servizio Statistiche congiunturali dell'industria e delle costruzioni

Sommario: Il presente lavoro descrive i metodi di controllo e correzione utilizzati nell'indagine sulla produzione industriale mensile con particolare attenzione alle innovazioni di processo conseguite con il passaggio al nuovo sistema informativo la cui fase di sviluppo non è ancora terminata. Le funzionalità descritte sono, pertanto, quelle disponibili al momento della presentazione al seminario "Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali" tenutosi a Roma il 15 novembre 2007.

Parole chiave: Indagine mensile sulla produzione industriale, sistema informativo, metodi di controllo e correzione.

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

[90] I paragrafi da 1 a 4 sono a cura di Annarita Mancini, il paragrafo 5 è a cura di Teresa Gambuti, l'appendice a cura di Maria Bruno.

1. L'indagine mensile sulla produzione industriale

1.1 Caratteristiche generali dell'indagine

L'indice della produzione industriale (IPI) misura la variazione nel tempo del volume fisico della produzione effettuata dall'industria in senso stretto (ovvero dell'industria con esclusione delle costruzioni). Esso si basa sui risultati di una rilevazione statistica longitudinale condotta presso un panel di imprese comprendente quelle più rappresentative dell'industria manifatturiera (sezione C, D ed E della NACE Rev1.1 ad esclusione della NACE 41 e 40.3 della sezione E). In conformità a quanto stabilito dal Regolamento STS n. 1165/98 del Consiglio dell'Unione europea relativo alle statistiche congiunturali, tale rilevazione misura la variazione nel tempo del volume di produzione dei beni inclusi in un paniere rappresentativo di prodotti. Ciò consente di calcolare i numeri indici per ciascuna voce di prodotto inserita nel paniere (indici elementari) che, a loro volta, sono sintetizzati per attività economica secondo la formula di Laspeyres.

Più in dettaglio, l'indagine mensile sulla produzione industriale viene effettuata direttamente presso un *panel longitudinale* di circa 6.000 unità rispondenti[91], che comunicano i volumi di produzione mensili relativi ad oltre 1.100 prodotti[92], definiti generalmente in termini di quantità fisiche[93]. Ad integrazione di tali dati, per la stima degli andamenti produttivi di specifici settori industriali, come ad esempio quello energetico e estrattivo, vengono utilizzate altre fonti che ne curano la collezione statistica in maniera aggregata e totalitaria.

Di tutti i prodotti rilevati, 848 sono utilizzati[94] per calcolare i 548 indici elementari grezzi, con uno sfasamento di 40 giorni, secondo quanto stabilito dal Regolamento STS Emendato (Regolamento STS n. 1158/2005).

Negli ultimi anni è fortemente cresciuta, in campo nazionale ed internazionale, l'esigenza di produrre un'informazione sempre più completa e tempestiva. Ciò ha comportato, oltre ad un progressivo e maggior carico di informazioni da rilevare ed elaborare mensilmente, anche una progressiva riduzione del tempo necessario a diffondere le stime richieste, tanto che il comunicato stampa viene diffuso attualmente a poco più di 40 giorni dal mese di riferimento (anziché dopo 45 giorni come richiesto dal vecchio Regolamento STS). I dati devono venire comunque inviati ad EUROSTAT tassativamente non oltre il 10 del secondo mese successivo a quello di rilevazione, eventualmente sotto embargo, come richiesto nella nuova versione del suddetto Regolamento STS, entrata in vigore l'11 agosto 2005.

La spinta alla tempestività conduce quindi a diffondere dati con maggior grado di provvisorietà, ovvero stimati sulla base delle risposte pervenute entro un tempo utile rispetto alla data prefissata. Al contempo, si richiede che le stime provvisorie siano affidabili e che si discostino il meno possibile da quelle definitive, con il conseguente obiettivo di ridurre al minimo le revisioni delle stime precedentemente diffuse in via provvisoria. Diventa imperativa, quindi, l'esigenza di affinare le tecniche di Controllo e Correzione (indicate con C&C di seguito). In particolare, per le indagini congiunturali, per le quali la tempestività è un requisito fondamentale, le mancate risposte e/o gli errori, soprattutto quelli influenti, impattano in maniera importante sulla qualità delle stime provvisorie dato che è spesso

---

[91] Circa 5.100 nel panel corrente di indagine e oltre 1.000 in un panel sostitutivo.

[92] Uno o più prodotti elementari *j* formano il macroprodotto *k* (o serie elementare) inserito nel paniere rappresentativo della base di riferimento dell'indice. L'indice della produzione industriale, infatti, è un indice a base fissa, aggiornata ogni 5 anni in termini di *panel* di imprese, ponderazione e paniere. Questi microprodotti sviluppano oltre 12000 microdati nel computo dell'indice mensile.

[93] A seconda delle specificità intrinseche del settore/prodotto oggetto di rilevazione si può far ricorso a dati di input (ore lavoro) oppure a dati di output (quantità prodotte, valore della produzione al netto delle variazioni dei prezzi).

[94] In realtà in precedenza il numero dei prodotti utilizzati nel calcolo dell'indice era maggiore (circa 950), in seguito tutta la nomenclatura dei prodotti IPI ha subito una riorganizzazione ai fini della spedizione relativa all'anno 2005 (coincidente con l'inizio della nuova base 2005) che ha comportato, in dettaglio, l'eliminazione di alcune voci di prodotto ormai obsolete, definizioni più fini e conformi alle classificazioni europee, l'inserimento di nuove voci di prodotto per le classi ATECO poco rappresentate, la modificazione, quando necessario, dell'intero questionario, l'introduzione e/o la modificazione delle avvertenze di compilazione relative a determinati prodotti e, infine l'aggiunta di unità di misura più adatte a osservare il prodotto rilevato oltre a quelle attualmente richieste (si veda Mancini A., 2005).

problematico il ricorso a tecniche di imputazione o di riponderazione efficienti, non applicabili nel breve arco temporale intercorrente tra l'acquisizione dei microdati e la data di diffusione delle stime a causa della mancanza di tempo e/o della non disponibilità di variabili ausiliarie.

1.2 Il modello di rilevazione

A partire dalla spedizione relativa all'anno 2007, al fine di rendere più agevole il compito delle imprese, riducendone quindi il carico statistico, a ciascuna unità rispondente viene inviato un questionario mensile personalizzato[95] in termini di lista delle voci di prodotto[96] inserite nel modello stesso. Per ognuna delle voci inserite nel questionario vengono richieste le quantità prodotte[97] nel mese corrente e nel mese precedente; è richiesto, inoltre, alla stessa unità rispondente *i-ma* di segnalare il numero di giorni lavorati nel mese corrente, inteso come il numero di giorni in cui l'unità ha svolto attività produttiva, a prescindere dal reale volume di lavoro impiegato in ciascuna linea produttiva. Per il solo mese di dicembre, inoltre, viene richiesto il numero medio di addetti dell'anno appena trascorso. Per alcuni prodotti, infine, è prevista un'unità di misura accessoria. Questa, in particolare, ha il duplice scopo di unità di misura di controllo e unità di misura "pilota"[98].

Tuttavia, i giorni lavorati, il numero medio di addetti e l'eventuale volume di produzione espresso nella seconda unità di misura, non sono utilizzati ai fini della costruzione dell'indice mensile della produzione industriale, per cui non sempre sono oggetto di controlli capillari e, inoltre, qualora mancanti, non determinano necessariamente la mancata risposta parziale dell'unità rispondente che, quindi, non viene sollecitata in caso abbia risposto alla variabile principale.

Sono altresì richieste nel modello informazioni relative ad eventuali cambiamenti anagrafici e il nome del compilatore (se diverso da quello del mese precedente). Eventuali definizioni o avvertenze di compilazione sono esplicitate nel modello stesso.

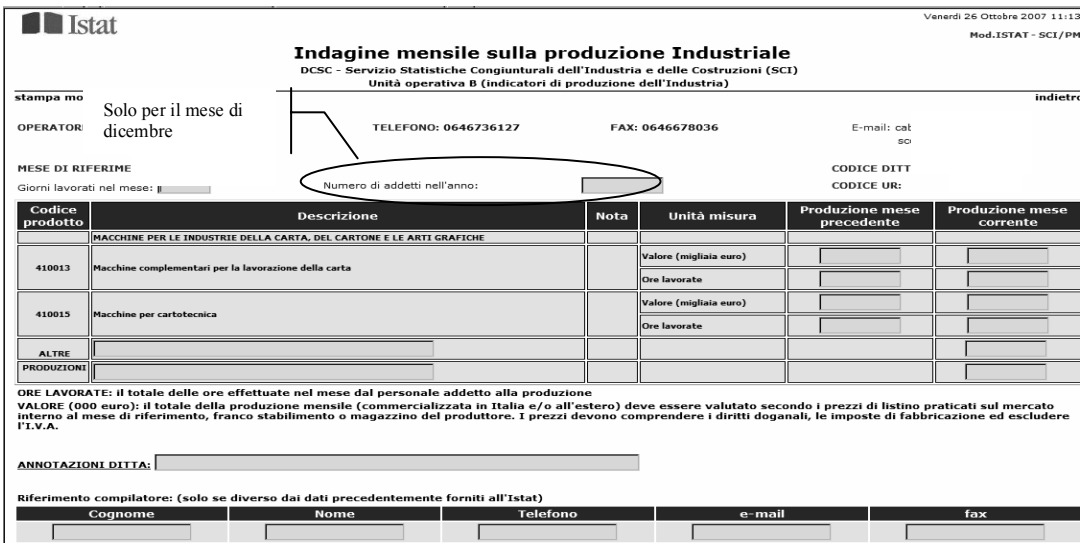
[95] Fino al 2006 venivano spediti 44 questionari disegnati in funzione dell'omogeneità di settore (ovvero in ciascun modello venivano elencate tutte le voci di prodotto della nomenclatura d'indagine appartenenti ad una o più classi ATECO). Alle imprese, quindi, potevano venire inviati più modelli nel caso in cui i prodotti manufatti da queste erano eterogenei in termini di classe ATECO. Ad esempio, alcune unità rispondenti operanti nel settore chimico ricevevano addirittura fino a 5 questionari, essendo tale settore molto complesso, in termini di prodotti rilevati, al punto di dover utilizzare 9 modelli di rilevazione. A volte accadeva, inoltre, che le imprese ricevevano più modelli ma il numero di prodotti, specifici della produzione della stessa impresa, in ognuno di essi poteva essere esiguo. Alla luce di ciò, si è disegnato un modello "personalizzato", definito a partire dalle unità rispondenti anziché dai prodotti inseriti nel campo di osservazione dell'indagine. Così anziché avere 44 modelli di rilevazione definiti sulla base delle produzioni omogenee di ciascun settore, attualmente si definiscono circa 6000 modelli, tanti quante sono le unità rispondenti dell'indagine.

[96] In fase di definizione del modello di rilevazione per la spedizione annuale (che attualmente avviene ad inizio dell'anno con l'invio dei 12 modelli mensili) vengono agganciate a ciascuna unità rispondente i codici prodotti *j* della nomenclatura IPI per cui ci sia stata almeno una risposta negli ultimi 24 mesi rilevati.

[97] Il volume di produzione può essere misurato in diverse unità di misura. Queste sono predefinite nella nomenclatura IPI e specifiche per ogni tipologia di manufatto rilevato. Si veda MANCINI A.et altri (2005 bis), per approfondimenti.

[98] Si inserisce un'altra unità di misura al fine di avere una misurazione alternativa del volume di produzione. Se si osserva che tale unità è più affidabile a misurare i volumi produttivi di uno specifico prodotto *j* si valuterà la possibilità di utilizzare tale unità per il computo dell'indice di produzione nella base successiva a quella di riferimento.

Figura 1.1 *Modello di rilevazione ISTAT–SCI/PM*



Istat

Venerdì 26 Ottobre 2007 11:13

Mod.ISTAT - SCI/PM

Indagine mensile sulla produzione Industriale

DCSC - Servizio Statistiche Congiunturali dell'Industria e delle Costruzioni (SCI)
Unità operativa B (indicatori di produzione dell'Industria)

stampa mo | indietr

OPERATOR: | TELEFONO: 0646736127 | FAX: 0646678036 | E-mail: cat sci

Solo per il mese di dicembre

MESE DI RIFERIME | CODICE DITT

Giorni lavorati nel mese: | Numero di addetti nell'anno: | CODICE UR:

Codice prodotto	Descrizione	Nota	Unità misura	Produzione mese precedente	Produzione mese corrente
	MACCHINE PER LE INDUSTRIE DELLA CARTA, DEL CARTONE E LE ARTI GRAFICHE				
410013	Macchine complementari per la lavorazione della carta		Valore (migliaia euro)		
			Ore lavorate		
410015	Macchine per cartotecnica		Valore (migliaia euro)		
			Ore lavorate		
ALTRE PRODUZIONI					

ORE LAVORATE: il totale delle ore effettuate nel mese dal personale addetto alla produzione

VALORE (000 euro): il totale della produzione mensile (commercializzata in Italia e/o all'estero) deve essere valutato secondo i prezzi di listino praticati sul mercato interno al mese di riferimento, franco stabilimento o magazzino del produttore. I prezzi devono comprendere i diritti doganali, le imposte di fabbricazione ed escludere l'I.V.A.

ANNOTAZIONI DITTA:

Riferimento compilatore: (solo se diverso dai dati precedentemente forniti all'Istat)

Cognome	Nome	Telefono	e-mail	fax

La struttura del modello rimane, quindi, sostanzialmente snella e semplice nella compilazione e ulteriormente alleggerita dal fatto che vengono elencati esclusivamente i prodotti che l'impresa produce, sulla base delle informazioni caricate nel database.

Una novità del modello "personalizzato" è la sezione dedicata ad altre produzioni, dove l'unità rispondente potrà annotare sia le produzioni che non rientrano nel campo di osservazione dell'indagine, sia quelle che, pur facendone parte, non risultavano essere prodotte dall'unità. In tal modo si possono avere informazioni aggiuntive per l'aggiornamento della base degli indici, rispettivamente sull'ampliamento del paniere con l'inserimento di nuovi prodotti e sull'allargamento del panel di imprese rispondenti al paniere dei prodotti già rilevati con l'attuale base.

1.3 Modalità di acquisizione dati

Per la raccolta dei dati si inviano a ciascuna unità rispondente 12 modelli ISTAT/SCI/PM (uno per ogni mese). Ciascun questionario mensile compilato deve essere inviato all'ISTAT, non oltre il giorno 10 del mese successivo a quello di riferimento per via telematica (da gennaio 2007), tramite fax o tramite posta. In aggiunta i dati possono essere acquisiti telefonicamente, soprattutto per effetto dei solleciti.

Mentre i dati trasmessi via web sono acquisiti direttamente e, quindi, sottoposti a microediting automatico, tutte le altre modalità di trasmissione prevedono una revisione preliminare interattiva del questionario prima della registrazione interna sul database, dove viene effettuato il microediting automatizzato.

Di seguito si riporta la percentuale di risposte, per tipologia di trasmissione, acquisite in media per il 2007:

- Questionario cartaceo autocompilato (via fax) 79%
- Questionario elettronico autocompilato 15%
- Dati inviati per e-mail[99] 4%
- Dati acquisiti per telefono in fase sollecito 2%

La modalità di trasmissione per posta, anche se prevista, non è idonea per un'indagine tempestiva come quella della produzione industriale, per cui resta con un'incidenza marginale, per lo più già comunicati per telefono.

[99] Tale modalità in realtà a partire dal mese di agosto è stata inibita ed è probabile che le risposte confluiranno nella trasmissione via web.

1.4 La revisione periodica dei dati grezzi

Gli indici della produzione industriale sono soggetti a due revisioni che si effettuano per motivi differenti.

Una prima revisione viene effettuata nel mese successivo al rilascio della prima stima, sulla base di informazioni aggiuntive che pervengono dalle imprese.

Una seconda revisione, che è stata introdotta a partire dal comunicato relativo ai dati di agosto 2004, avviene a cadenza semestrale e riguarda le serie storiche degli indici. Tale revisione ha lo scopo di incorporare negli indici tre tipologie di informazioni che si rendono disponibili successivamente alla pubblicazione della prima rettifica. Nello specifico, gli elementi considerati nel processo di revisione sono:

1. le risposte pervenute dalle imprese dopo la chiusura degli indici rettificati (che avviene di regola intorno a 60 giorni dalla fine del periodo di riferimento); si tratta di una quota di risposte molto limitata, ma che può determinare rettifiche di un qualche rilievo sugli indici disaggregati.
2. le correzioni a posteriori di informazioni già pervenute dalle imprese e che sulla base di successive verifiche sono risultate affette da imprecisioni nella misurazione del fenomeno. Si tratta di modifiche che hanno, in media, un effetto contenuto sugli indici aggregati ma che, occasionalmente, possono causare revisioni significative per specifici settori.
3. l'aggiornamento e la periodica revisione (relativa all'ultimo triennio), delle stime di contabilità nazionale degli aggregati di valore aggiunto e sulle ore effettivamente lavorate su cui si basano i coefficienti annuali di produttività utilizzati per i prodotti rilevati tramite i flussi mensili di ore lavorate[100]. Ne deriva che l'effetto della revisione dei coefficienti può risultare sensibile per quegli specifici settori.

Queste revisioni avvengono in occasione della diffusione degli indici relativi al mese di febbraio e di agosto. Nella prima sono incorporate sia le nuove stime annuali di contabilità nazionale per i tre anni precedenti, sia le rettifiche basate sulle risposte giunte con ritardo e sulle correzioni di informazioni già pervenute. Nella seconda si tiene conto della sola componente dovuta a informazioni supplementari e rettifiche, operando una revisione a partire dal gennaio dell'anno corrente; questa, inoltre, ha lo scopo di verificare la significatività di dati anomali riscontrati in corso di rilevazione [101].

2. Problematiche principali

2.1 Individuazione degli outlier

Nell'indagine mensile sulla produzione industriale i dati anomali sono individuati confrontando i volumi di produzione mensile relativi all'unità *i* per il prodotto *j* con il *range* del volume di produzione osservato nella serie storica del microdato:

$$ {}^{j}y_{mi} \notin \left[{}^{j}y_{\min i} ; {}^{j}y_{\max i} \right] $$

Vengono altresì considerati anomali quei valori che determinano una variazione tendenziale fuori i limiti di accettazione, definiti per ciascun prodotto *j*, i cui estremi sono in genere individuati dal decimo percentile (${}^{j}p_{10i}$) e dal 90-mo percentile (${}^{j}p_{90i}$) della distribuzione di frequenza delle variazioni tendenziali calcolate negli ultimi 4 anni[102] e simmetrizzati rispetto a 100. Per i domini più importanti (determinati dal peso attribuito nell'anno base) i limiti di accettazione sono più rigorosi[103].

---

[100] Tali prodotti pesano, nel complesso, per il 6% dell'indice generale sebbene risultano concentrati in alcuni settori (in particolare, macchine e apparecchi meccanici, apparecchi elettrici e di precisione, mezzi di trasporto).

[101] Per un'analisi approfondita delle revisioni operate negli ultimi anni sugli indici della produzione industriale si rimanda al documento "Revisione degli indici della produzione industriale" nell'area *download* del sito Istat.

[102] Con l'esclusione del mese di agosto che presenta di norma andamenti discontinui legati al periodo di chiusura estiva che in genere, per alcuni settori e per alcune imprese, è correlata all'andamento del proprio ciclo economico.

[103] In effetti, vista l'estrema variabilità di alcuni domini, si è deciso comunque di mantenere intervalli di accettazione più rigorosi.

Per formalizzare:

${}^{j}W_b$ è il peso del prodotto *j* nell'anno base b,

${}^{[j]}W_b$ è il peso percentuale cumulato, in ordine decrescente di peso, relativo al prodotto localizzato al *j*-mo posto,

${}^{j}\Delta_{mi} = (\frac{{}^{j}y_{mi}}{{}^{j}y_{(m-12)i}}) \times 100$ è il rapporto tendenziale percentuale dell'unità i per il prodotto *j*,

${}^{j}Inf_i = 100 * \frac{{}^{j}p_{10i}}{\sqrt{{}^{j}p_{10i}\,{}^{j}p_{90i}}}$ è il decimo percentile simmetrizzato[104] rispetto a 100 (caso di invarianza del dato),

${}^{j}Sup_i = 100 * \frac{{}^{j}p_{90i}}{\sqrt{{}^{j}p_{10i}\,{}^{j}p_{90i}}}$ è il 90-mo percentile simmetrizzato[15] rispetto a 100.

${}^{j}\Delta_{mi}$ sarà considerata anomala se valgono le seguenti condizioni:

1) $\begin{cases} \forall [j]: {}^{[j]}W_b \le 25 \quad \text{e} \quad \begin{cases} {}^{j}Inf_i < 75 \\ {}^{j}Sup_i > 135 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [75;135] \\ \textit{oppure} \\ \forall [j]: {}^{[j]}W_b \le 25 \quad \text{e} \quad \begin{cases} {}^{j}Inf_i \ge 75 \\ {}^{j}Sup_i \le 135 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [{}^{j}Inf_i; {}^{j}Sup] \end{cases}$

2) $\begin{cases} \forall [j]: {}^{[j]}W_b \in (25;50] \quad \text{e} \quad \begin{cases} {}^{j}Inf_i < 50 \\ {}^{j}Sup_i > 200 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [50;200] \\ \textit{oppure} \\ \forall [j]: {}^{[j]}W_b \in (25;50] \quad \text{e} \quad \begin{cases} {}^{j}Inf_i \ge 50 \\ {}^{j}Sup_i \le 200 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [{}^{j}Inf_i; {}^{j}Sup] \end{cases}$

3) $\begin{cases} \forall [j]: {}^{[j]}W_b \in (50;75] \quad \text{e} \quad \begin{cases} {}^{j}Inf_i < 33 \\ {}^{j}Sup_i > 300 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [33;300] \\ \textit{oppure} \\ \forall [j]: {}^{[j]}W_b \in (50;75] \quad \text{e} \quad \begin{cases} {}^{j}Inf_i \ge 33 \\ {}^{j}Sup_i \le 300 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [{}^{j}Inf_i; {}^{j}Sup] \end{cases}$

4) $\begin{cases} \forall [j]: {}^{[j]}W_b > 75 \quad \text{e} \quad \begin{cases} {}^{j}Inf_i < 25 \\ {}^{j}Sup_i > 400 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [25;400] \\ \textit{oppure} \\ \forall [j]: {}^{[j]}W_b > 75 \quad \text{e} \quad \begin{cases} {}^{j}Inf_i \ge 25 \\ {}^{j}Sup_i \le 400 \end{cases} \Rightarrow \text{anomalo se } {}^{j}\Delta_{mi} \notin [{}^{j}Inf_i; {}^{j}Sup] \end{cases}$

[104] Questa operazione vuole considerare come significativi gli scostamenti in entrambe le direzioni dello stesso ordine.

Tali valori vengono considerati come "potenziali" errori e/o "potenziali" *outlier* da verificare tramite *follow-up*.

Nel caso in cui non si riesca a contattare il rispondente in un tempo utile per la validazione dell'indice provvisorio, può accadere che tali valori vengano trattati come potenziali outlier e le eventuali correzioni, in seguito a *follow-up*, verranno inserite con le successive revisioni.

Invece, nel caso di unità influenti, o comunque di variazioni influenti per una singola serie elementare, il dato viene imputato in maniera interattiva in base alla serie storica osservata e caricato nel database dei microdati con flag "forzato". Queste forzature dovranno essere risolte con successivi *follow-up* e hanno la priorità rispetto agli altri potenziali errori o *outlier*. I valori anomali influenti vengono pertanto "congelati" fino al momento in cui si è risolto il *follow-up*: se non si ha una conferma motivata da parte dell'unità rispondente il dato anomalo viene registrato ma non utilizzato ai fini del calcolo dell'indice. Al suo posto viene imputato un valore tenendo conto soprattutto

- della serie storica del microdato in questione;
- dell'informazione aggiuntiva che ha determinato il dato anomalo;
- dell'andamento delle altre imprese rispondenti nel dominio di appartenenza dell'unità influente per cui si è rilevato un valore anomalo "non risolto" e nei domini il cui processo produttivo è verticalmente integrato con quello di interesse.

I casi i valori anomali, sia per casi influenti che non, se confermati dall'unità rispondente sono registrati e utilizzati nel calcolo dell'indice.

Tutti i microdati sono corredati di un campo nota per appuntare i motivi addotti come causa del valore anomalo dichiarati dall'unità rispondente.

2.2 Individuazione e trattamento degli errori

2.2.1 Errori Sistematici

Nell'indagine mensile sulla produzione industriale possono essere riscontrati, sebbene non frequentemente, errori ripetuti nel tempo; questi vengono individuati solo a posteriori e, quindi, corretti in occasione delle revisioni.

Un tipico errore classificato come tale è la diversa dichiarazione di dominio j_1 al mese m che può verificarsi principalmente per un cambiamento della figura del compilatore che interpreta in maniera diversa la definizione del prodotto j_1 indagato, rispetto a quanto veniva fatto in precedenza: il nuovo compilatore, dunque, inizia a redigere una nuova voce di prodotto j_2, nella sezione altre produzioni del questionario, lasciando in bianco (o con produzione nulla) una o più voci di prodotti della lista attribuita all'unità in base alle risposte precedenti. Nel caso che si verifichi l'evento, il livello del dominio j_1 presente fino al mese $(m-1)$ viene mantenuto applicando ad esso gli andamenti congiunturali che si osserveranno a partire dal tempo m per l'output del nuovo dominio j_2. Quest'ultimo, invece, viene tenuto in osservazione. L'adeguamento dei valori registrati in $(m-1)$ e m effettuato per il dominio j_1, in pratica, si basa solo sui valori osservati dell'unità rispondente coinvolta e non su una media dei valori calcolata sulle unità, appartenenti allo stesso dominio, non soggette a modifiche di strato. Tale imputazione, a partire dal mese m, viene impostata automaticamente e registrata sul database con il flag "calcolato".

Altro errore di tipo sistematico dovuto al rispondente, è quello di misurazione: i dati possono essere forniti in un'unità di misura non appropriata anche per un lungo periodo di osservazione e si ha coscienza dell'errore solo a posteriore per effetto di un cambiamento del compilatore o anche per effetto del *benchmarking* che si effettua periodicamente con i dati dell'indagine annuale sulla produzione industriale (PRODCOM). Il trattamento di questi errori prevede di mantenere il livello pregresso, applicando quindi la variazione effettiva. Analogamente a quanto descritto sopra l'imputazione viene impostata in maniera automatica sul database.

Infine, tra gli errori sistematici dovuti principalmente ai rispondenti devono essere annoverati quelli causati da un evento di demografia di impresa. Nel caso in cui avvenga, per esempio, una fusione tra un'unità presente nel panel e un'altra non inserita inizialmente nella base, l'aumento degli output produttivi senza dichiarazione dell'evento demografico comporterebbe un incremento errato dei livelli

produttivi dell'unità originaria, rispetto alla base di riferimento, e, conseguentemente, dei prodotti da questa manufatti; il panel, quindi, perderebbe la rappresentatività longitudinale. Specularmente si ha lo stesso effetto per gli scorpori non segnalati dall'impresa rispondente. Di solito questi errori vengono individuati rapidamente, anche se non sempre si riesce a effettuare il *follow-up* in tempo utile per la validazione dell'indice provvisorio, per cui tali valori vengono trattati come gli outlier descritti nel precedente paragrafo. Successivamente, se in seguito al *follow-up* l'impresa dichiara l'evento di trasformazione dell'impresa, ma non riesce a mantenere la continuità delle vecchie serie (ad esempio attribuire la produzione afferente alla vecchia impresa e alla nuova) si utilizzano tecniche di imputazione simili a quelli delle mancate risposte totali (MRT) descritta più avanti[105].

2.2.2 Errori casuali

Può accadere che un errore avvenga isolatamente per un singolo mese. Di norma questi sono più frequenti di quelli sistematici e possono riguardare semplici errori di digitazione (sia da parte del compilatore sia da parte del revisore ISTAT nel caso di acquisizione non via web), errori di misurazione (produzione cumulata di più mesi) o di errata interpretazione della nomenclatura di indagine (esclusione e/o inclusione di produzioni escluse dal campo di osservazione, ecc.). Tale errori vengono di norma risolti con la verifica presso le imprese. Nel caso non si riesca a contattare l'unità rispondente in tempo utile per la validazione dell'indice provvisorio può accadere che tali valori vengono trattati con tecniche di imputazione simili a quelli per la MRT descritta più avanti.

2.3 Mancate risposte: totali e parziali

Le mancate risposte sono il principale fattore che può influenzare la qualità delle statistiche prodotte dall'indagine mensile sulla produzione industriale. È da rilevare, inoltre, che la maggior parte delle indagini congiunturali sono caratterizzate da una caduta di risposta direttamente proporzionale alla lunghezza del periodo di osservazione. Infatti, l'aumento di mancate risposte con l'allontanarsi dall'anno base di riferimento, è dovuto all'effetto dell'attrito ma anche del deterioramento in termini di rappresentatività sia del panel che del paniere oggetto di indagine.

In aggiunta a questi fattori "fisiologici", la tendenza alla diffusione di dati provvisori sempre più tempestivi, oltre alla richiesta di stime provvisorie il meno possibile diverse da quelle definitive, ha comportato negli ultimi anni l'investimento, da parte dell'unità operativa, di risorse nell'affinare le tecniche di imputazione delle mancate risposte[106]. Ed è per questo motivo che a tali tecniche è dedicato il paragrafo 3.5.

Sebbene il questionario rilevi più variabili quantitative, solo una è utilizzata ai fini del calcolo dell'indice mensile, ovvero il volume di produzione per il prodotto j espressa nell'unità di misura prescelta ($^{j}y_{mi}$) in fase di definizione del paniere di riferimento. Pertanto come mancata risposta viene considerato il dato mancante relativamente a tale variabile, anche quando l'unità rispondente abbia fornito altre variabili. Conseguentemente le tecniche di C&C vengono applicate solo per $^{j}y_{mi}$.

In genere, i motivi della mancata risposta totale (MRT) sono da attribuirsi ai rispondenti e sono conseguenti a attrito, a non reperibilità del compilatore (per ferie, malattia, messa in cassa integrazione o altro), a motivi contingenti all'unità rispondente (come per esempio per un maggior carico di lavoro dovuta alla chiusura bilanci, inventario magazzino, ecc.) per lo più caratteristici di specifici mesi, a eventi accidentali e del tutto esogeni alle decisioni dell'impresa (scioperi, alluvioni, terremoti, incendi, ecc) e, infine, a variazioni demografiche (cessazione attività, cessioni/acquisizioni di rami d'azienda) o trasferimenti territoriali (oltre il confine nazionale).

[105] Per ora il flag è quello della stima di mancate risposte, in futuro si metterà un flag apposito. È in atto, infatti, uno studio per la classificazione delle mancate risposte tipiche dell'indagine mensile sulla produzione industriale.
[106] Si veda Gismondi (2006) e Gismondi (2004)

Di solito la MRT coinvolge tutti i prodotti attribuiti all'unità stessa e su tutte le variabili indagate. Raramente è da attribuirsi a smarrimento del questionario compilato[107].

Le mancate risposte parziali (MRP), invece, sono relative soprattutto alle variabili quantitative ausiliarie e sono dovute sia ai rispondenti sia alla raccolta. In particolare le MRP derivano in principal modo da problemi di misurazione (per esempio, i rispondenti non riescono ad esprimere i volumi produttivi di uno specifico prodotto nell'unità di misura previsti dal questionario), ma possono derivare anche dal metodo di raccolta: infatti, quando viene effettuato il sollecito telefonico, per motivi di tempo, si richiede soprattutto la variabile ${}^{j}y_{mi}$ trascurando quelle ausiliarie.

Naturalmente le MRP possono riguardare anche la variabile quantitativa ${}^{j}y_{mi}$, ma i motivi sono attribuibili sia ai rispondenti sia alla raccolta. Questi vengono trattati con tecniche di imputazione particolari, sfruttando le variabili ausiliarie se esistenti (unità di misura secondaria) o seguendo le stesse tecniche di imputazione utilizzate per la MRT.

3. Il processo di C&C nel processo di diffusione degli indici grezzi provvisori

3.1 Schema generale: processo diffusione degli indici grezzi provvisori

Il processo di controllo e correzione dei dati per la diffusione degli indici grezzi provvisori adottato nell'indagine mensile sulla produzione industriale è abbastanza articolato, sebbene si svolga in un lasso di tempo relativamente breve. Ricordiamo, infatti, che le imprese devono inviare i dati di produzione entro il 10 del mese successivo a quello di riferimento e le stime provvisorie dell'indice mensile vengono diffuse in media a 40 giorni.

Pertanto le operazioni relative al processo di elaborazione delle stime preliminari sono compresse in un mese di calendario e sono, in pratica, basate sulle informazione pervenute fino a quel momento. Successivamente si provvederà a calcolare un indice rettificato, poco prima del comunicato successivo, e ad effettuare revisioni periodiche.

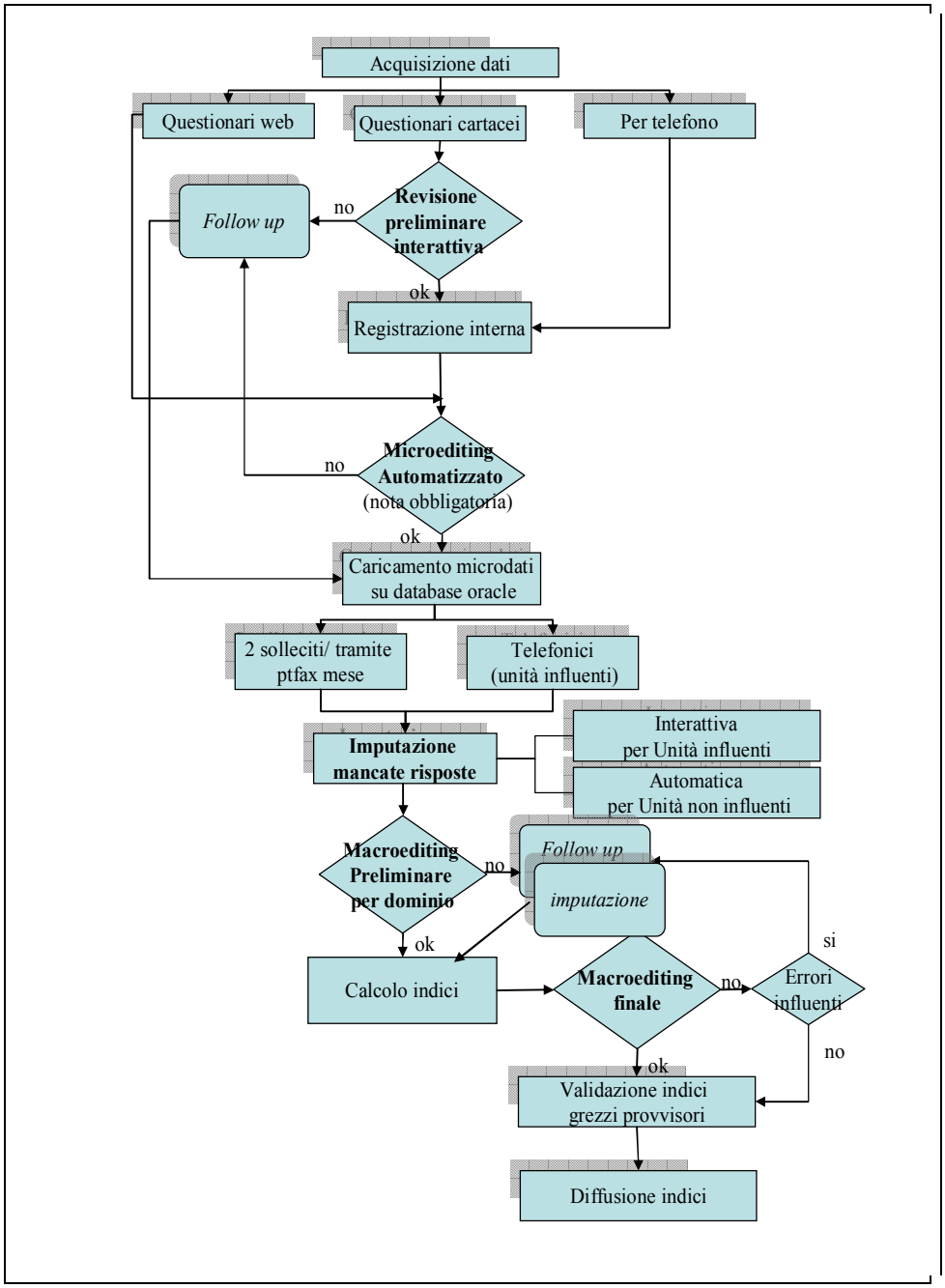
Le fasi operative dell'intero processo di diffusione degli indici provvisori possono schematizzarsi come segue:

1. acquisizione dei dati
1.1. gestione dei questionari
1.2. inserimento e revisione dei microdati
1.3. solleciti
2. elaborazione dei dati
2.1 stima dei dati mancanti
2.2 definizione degli indici elementari provvisori
2.3 calcolo degli indici per aggregati economici
2.4 calcolo degli indici derivati (corretti per giorni lavorativi e stagionalità)
3. diffusione dei dati per l'indice provvisorio (comunicato stampa)
3.1 comunicato stampa (indici provvisori del mese di riferimento e indici rettificati del mese precedente)
3.2 CON.ISTAT (indici provvisori del mese di riferimento e indici rettificati del mese precedente
3.3 Enti Internazionali
3.4 Altri utenti

Sembra opportuno, in aggiunta, rappresentare il processo di diffusione delle stime preliminari con un diagramma di flusso, come riportato nella successiva figura.

[107] La maggior parte dei questionari, come descritto nel paragrafo 1.3, perviene via fax, prevalentemente sul fax server dell'Istat, o è acquisito direttamente via web.

Figura 3.1 *Schema generale del processo di diffusione degli indici grezzi provvisori*



3.2 L'individuazione delle unità rilevanti

Una delle caratteristiche dei microdati relativi all'indagine mensile sulla produzione industriale è la loro grande eterogeneità, persino all'interno dello stesso raggruppamento di prodotti (dominio). Questa eterogeneità spesso comporta una grande variabilità nei dati.

In particolare, è importante individuare le unità "rilevanti", allo scopo di poter privilegiare il trattamento statistico dei dati ad esse relativi. Queste normalmente vengono trattate in modo interattivo, per cui una mancata risposta non è oggetto di stima tramite procedure automatiche di imputazione.

La revisione della procedura di identificazione delle unità rilevanti è stata recentemente oggetto di studio, ed una nuova proposta operativa[108] è stata predisposta ed ampiamente documentata (Gismondi et altri, 2006, Gismondi et altri, 2007). In sintesi, il nuovo criterio di identificazione cerca di tenere conto simultaneamente – ed in modo più sistematico rispetto a quanto fatto in precedenza – dei seguenti aspetti: a) il peso relativo dell'unità in termini di produzione; b) la variabilità longitudinale dei dati (ossia la propensione dell'impresa a generare forti variazioni congiunturali e/o tendenziali; c) il livello di concentrazione della produzione nello strato di appartenenza dell'impresa:
Nello specifico l'individuazione delle unità rilevanti si basa su tre condizioni:

1) per un singolo prodotto *j* le n unità influenti individuate garantiscono una copertura di almeno il 50% (in termini di produzione)
2) ciò implica che ogni strato *j* abbia almeno una unità influente che copre il 20% della produzione dello strato
3) Ogni macroprodotto *k* deve essere caratterizzato da almeno una unità rilevante.

In generale, facendo riferimento all'attuale struttura delle unità rilevanti si può affermare che il loro peso (in termini produzione) nei vari settori è fortemente variabile: è piuttosto basso nei settori meno concentrati e senza un rilevante peso sull'intera economia (DA: alimentare; DB: tessile), mentre cresce fortemente nelle attività più concentrate e con un peso rilevante, come DJ (produzione di metallo e prodotti in metallo) e DK (produzione di macchine e apparecchi meccanici).

Si ricorda che per le unità rilevanti non viene applicata la procedura di stima automatica. L'ipotesi di base è che queste imprese siano talmente importanti da poter influenzare l'andamento dei comparti ai quali appartengono e il dato di produzione da esse fornito, per cui dovrebbero risultare sempre rispondenti al momento del calcolo dell'indice.

Quando, tuttavia, per motivi contingenti queste imprese non sono in grado di rispettare le scadenze richieste l'imputazione del dato mancante viene fatta ricorrendo sopratutto all'osservazione della serie storica. Le informazioni riguardanti la dinamica della singola impresa vengono, inoltre, adeguate sulla base delle informazioni ausiliarie disponibili al momento dell'elaborazione dell'indice, ad esempio le informazioni riguardanti gli scioperi, i blocchi della produzione, le condizioni climatiche e così via. Stesso discorso vale per valori anomali pervenuti, come descritto nel paragrafo 2.1.

3.3 Revisione preliminare interattiva

La prima attività svolta sui questionari cartacei (inviati, quindi, per fax) è la revisione critica del modello in termini di risposte per:

- codici prodotto: il revisore[109] controlla se tutte le voci di prodotto sono state compilate,
- volumi di produzione: i modelli cartacei[110]prevedono una colonna relativa ai volumi di produzione del mese precedente a quello di rilevazione, per cui un primo controllo è vedere se ci siano "salti di livello" o dichiarazioni di produzione nulla.

In questa prima fase il revisore è coadiuvato da tre maschere disponibili nel nuovo sistema informativo[111]:

1) La *maschera di visualizzazione M1;* questa permette di avere per ciascuna impresa (codice ASIA) l'elenco delle unità rispondenti[112] (UR) con l'indicazione delle produzioni attive e di quelle cessate.

---

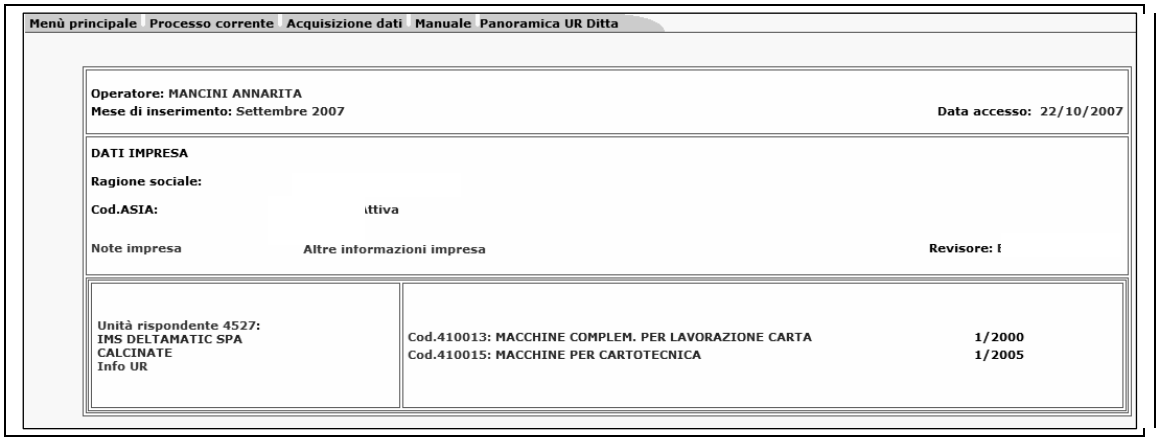
[108] In precedenza le unità rilevanti erano individuate combinando la predominanza dell'impresa, il peso dello strato e la concentrazione dello strato in termini di produzione. La procedura non era però sviluppata in maniera rigorosa e la lista veniva aggiornata solo in fase di cambiamento di base.

[109] Ciascuno dei 16 revisori in forza nell'unità operativa è considerato esperto di un particolare settore. Ad ognuno, infatti, sono assegnate più classi ATECO di cui sono "responsabili" per le quali curano i dati della maggior parte delle unità rispondenti. Di solito si cerca di mantenere fissi i settori assegnati ad ogni revisore in modo che le professionalità acquisite nel tempo per quel settore specifico non vengano disperse.

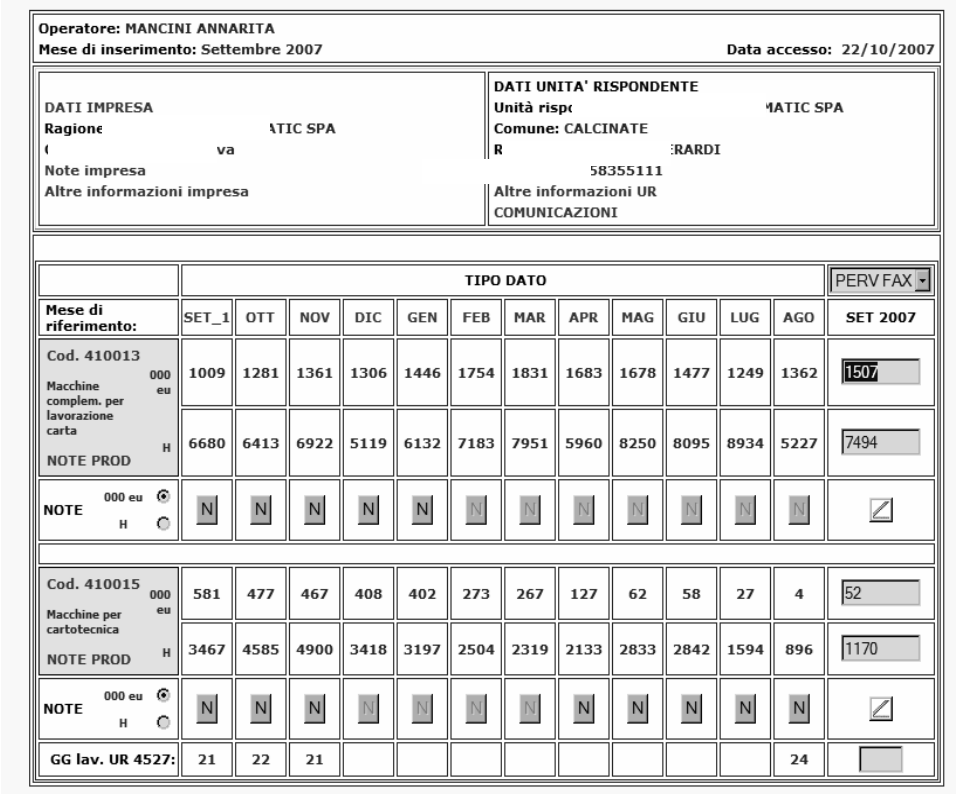
[110] Si veda come esempio la Figura 1.1 del paragrafo 1.3.

[111] Tutti i precedenti applicativi sono stati riprogrammati sulla nuova base dati oracle prevedendo l'utilizzo del linguaggio PHP per lo sviluppo delle applicazioni grafiche interattive disponibili nel nuovo sistema informativo della produzione industriale; i modelli di interfaccia sono stati impostati in modo "user friendly".

[112] Che possono coincidere con le unità locali dell'impresa.

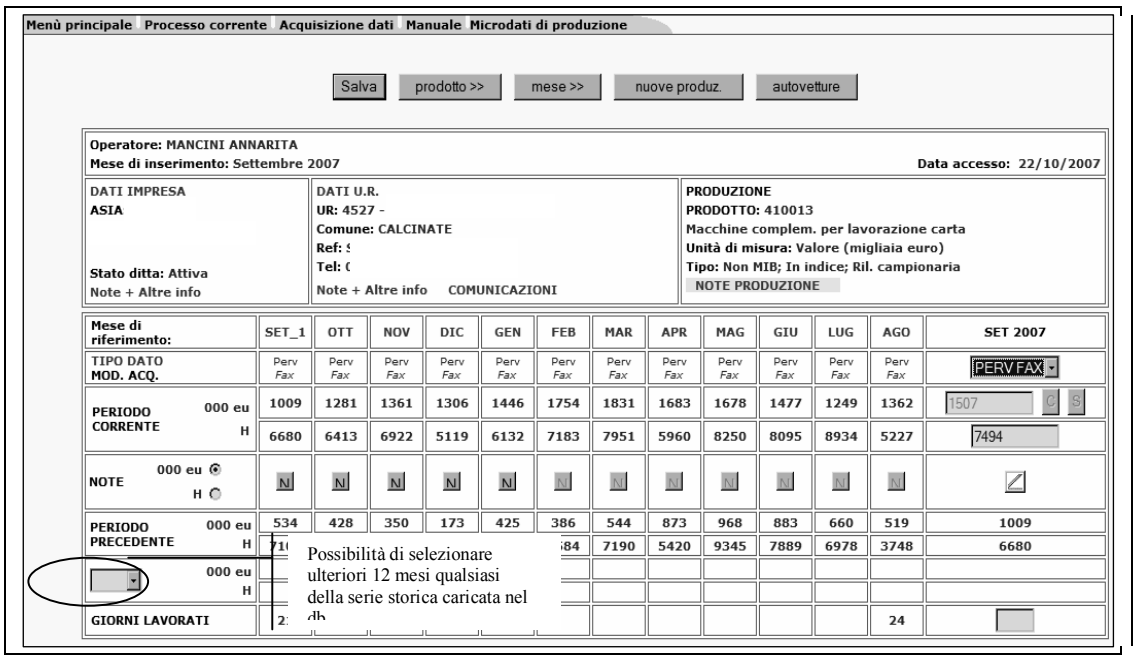


2) La *maschera di visualizzazione M2*; la quale fornisce la lista di tutte le produzioni di ciascuna unità rispondente; questa permette di effettuare una prima verifica della compilazione di tutti i codici prodotti attribuiti all'unità rispondente, l'eventuale spostamento di produzione tra codici di prodotto, gli eventuali errori di misurazione (per esempio inversione tra unità di misura nel caso di prodotti che prevedono la rilevazione in due dimensioni), oppure le possibili mancate risposte parziali. In tale maschera sono già visibili eventuali note ai microdati rilevati inserite nei mesi precedenti. Queste sono informazioni molto utili perché specifiche di ciascuna produzione dichiarata in un certo mese *m* e, quindi, danno la possibilità di annotare in particolar modo gli eventi contingenti relativi ad uno specifico periodo evitando spesso di effettuare inutili *follow-up* presso i rispondenti: ad esempio, se nel mese precedente l'unità rispondente ha segnalato una commessa particolarmente rilevante per un prodotto e che interessa più mesi, è possibile accettare un valore pervenuto particolarmente alto caricando il dato direttamente nel database senza ulteriore *follow-up*. È da rilevare che la nota, digitabile nella successiva maschera, è specifica per ciascun prodotto *j* dell'unità *i* selezionata per il mese *m* e per ciascuna unità di misura in cui il prodotto *j* è espresso.



3) *Maschera di imputazione M3*: questa prevede la visualizzazione in serie storica dei dati per ciascuna

produzione (ultimi 25 mesi con possibilità di selezionare ulteriori 2 mesi), in modo da poter valutare interattivamente se il dato inviato sia plausibile o meno in termini longitudinali (molto utile soprattutto per i dati acquisiti telefonicamente).



Più in dettaglio, nella maschera M3 è riportata la nota di produzione specifica per il prodotto *j* dell'unità *i*. Si pensi, per esempio, al caso in cui per una specifica linea produttiva (coincidente con una produzione i*j) il personale sia messo in cassa integrazione; se nella nota di produzione il revisore ha segnalato questo stato per il prodotto *j*, sebbene temporaneo, volumi di produzione bassi possono essere accettati e, quindi, inseriti senza ulteriori riscontri presso il rispondente.

Tutte e tre le maschere hanno, inoltre, un corredo informativo relativo all'impresa (codice Ateco, eventi demografici, note immesse nel sistema) e all'unità rispondente analizzata (riferimenti telefonici, indirizzi e-mail, note e tipologia di non risposta[113]).

Nel caso di incertezza su valori o sui codici compilati dall'unità rispondente, il revisore può decidere se eseguire direttamente il *follow-up* presso l'unità rispondente o provare comunque a caricare i dati direttamente sul database, il quale provvederà ad effettuare automaticamente il microediting come descritto nel successivo paragrafo.

3.4 Microediting automatizzato

Il microediting automatizzato è eseguito sia per i dati caricati manualmente[114], registrati tramite la maschera M3 descritta nel precedente paragrafo, sia per i dati inviati per via telematica. Tale controllo è determinato da un software sviluppato ad hoc su piattaforma Oracle.

In dettaglio i controlli effettuati in questa fase sono:

- Segnalazioni di missing su codici prodotti assegnati alla UR (solo per dati acquisiti via web)
- Segnalazioni, non vincolanti, di zeri
- Controlli di quadratura (solo sui microdati inviati dalle raffinerie di petrolio essendo dati di bilancio)
- Relazioni logiche non valide (esempio volumi di produzione padre < volumi di produzione prodotti figli)

[113] In realtà tale variabile non è ancora efficientemente definita, ma si sta completando la classificazione delle singole UR secondo classi di non risposta analogamente a quanto richiesto per la diffusione di indicatori SIDI (sistema di documentazione delle indagini dell'Istat),: distinzione tra unità eleggibili e non eleggibili e, per quelle eleggibili, il motivo della non risposta.

[114] Quindi dati pervenuti via fax (questionari cartacei), acquisiti telefonicamente oppure inviati per e-mail.

- Segnalazioni di fuori range dei valori assoluti[115] (range calcolato sulla serie storica degli ultimi 4 anni e per cui ci siano almeno 12 mesi inseriti nel database[116])
- Segnalazioni di fuori range dei rapporti tendenziali rispetto a intervalli costruiti utilizzando le distanze interquantiliche della distribuzione dei rapporti tendenziali[24117].

Il microediting automatizzato, insieme a quello interattivo, permettono di raggiungere obiettivi di accuratezza, per mezzo di strumenti quali:

- la disponibilità della lista prodotti completa per unità rispondente: questa favorisce il controllo della validità del codice prodotto segnalato dall'impresa
- la disponibilità della lista prodotti completa dell'indagine: questa favorisce l'introduzione di nuovi prodotti in osservazione
- la segnalazione di "potenziali" errori e/o outliers

Inoltre, permette di effettuare in tempo reale l'eventuale *follow-up* dei dati segnalati anomali, rendendo possibile un'immediata separazione tra errori effettivi (risoluzione immediata con l'unita rispondente) e semplici outlier (in questo caso il sistema richiede necessariamente di compilare il campo "nota microdato" in cui viene inserita la motivazione dichiarata dal rispondente).

3.5 Le tecniche di riduzione delle mancate risposte e la procedura corrente di integrazione delle mancate risposte

3.5.1 Tecniche di riduzione delle mancate risposte

Allo scopo di mantenere un tasso di risposta soddisfacente sono messe in atto campagne di sensibilizzazione dei rispondenti o altre misure di carattere preventivo dell'errore, come il coinvolgimento di associazioni di categoria, l'invio di lettere di preavviso e di *gadgets* e informazioni sull'utilizzo dei dati forniti. Mentre il coinvolgimento delle associazioni di categoria viene attuato *una tantum* e di solito in fase di definizione del nuovo paniere e di selezione del panel delle imprese, l'invio di eventuali *gadgets*[118] e di informazioni inerenti l'utilizzo e la diffusione dei dati forniti è condotto annualmente insieme alla lettera di spedizione dei questionari. L'invio di lettere di preavviso, invece, viene effettuato mensilmente ed in maniera mirata soprattutto per via telematica.

Ad ogni cambiamento di base, inoltre, si effettua, nei domini di analisi per i quali ciò è possibile, l'ampliamento del campione e la sostituzione dei non rispondenti al fine di ridurre la varianza dell'errore di stima.

Particolarmente curato, infine, è il ricorso ad operazioni di sollecito e di contatto ripetuto, principalmente per telefono, dei non rispondenti. In particolare, si effettuano due solleciti automatici tramite *pt-fax*: il primo sollecito è fissato di solito nella terza settimana dopo la fine del mese di riferimento; il secondo sollecito, invece, viene di norma fissato nella quarta settimana e, comunque, almeno dieci giorni prima della diffusione del comunicato stampa e coinvolge tutte le imprese, anche quelle non inserite nel panel oggetto di calcolo dell'indice mensile.

Con riferimento alle unità più rilevanti si effettuano solleciti mirati sia in termini di tempistica che di tipologia. Inoltre, vi sono contatti telefonici diretti con il personale dell'impresa, che a volte prevedono accordi preventivi per esonerare le imprese stesse da spiacevoli solleciti automatici.

Solo dopo aver attuato tutte queste misure cautelative, si procede all'imputazione dei dati relativi alle unità non rispondenti mediante l'applicazione di procedure automatiche e/o interattive, in quest'ultimo caso a cura degli esperti di settore.

3.5.2 Mancate risposte totali

Si è scelto di sviluppare le tecniche di stima a livello di microdato, e non ad un livello superiore di aggregazione dei dati, per poter disporre anche di serie storiche *complete* di microdati validati (pervenuti o stimati).

---

115 Si veda per maggiori dettagli il paragrafo 2.1 per individuazione valori anomali

116 Altrimenti per nuove produzioni scatterebbe il controllo automatico sempre.

117 Si veda paragrafo 2.1.

118 In realtà questi sono stati utilizzati in via sperimentale per la spedizione dell'anno 2005 e 2006.

Per poter gestire operativamente e nel rispetto di chiari criteri metodologici la procedura di integrazione delle mancate risposte totali, le unità rispondenti sono suddivise in due tipologie:

1. unità rilevanti (ovvero particolarmente influenti)
2. unità poco rilevanti come descritto nel precedente paragrafo 3.2.

In ogni caso, prima di lanciare a automatica, per ogni dominio *j* viene calcolata la stima dell'indice di copertura $\sum_{i=1}^{n_{Rmj}} {}^{j}y_{(m-1)i}$ nel mese (*m*-1),

$${}^{j}\tilde{C}op_{m} = \frac{\sum_{i=1}^{n_{Rmj}} {}^{j}y_{(m-1)i}}{\sum_{i=1}^{n_{j}} {}^{j}y_{(m-1)i}}.$$

dove, ${}_{n}R_{mj}$ il numero di unità del panel riferite al prodotto *j* che effettivamente rispondono nel mese *m*.
In definitiva, le tecniche di imputazione per MRT si dividono in

- <u>automatiche</u> per le unità poco rilevanti e appartenenti a domini per i quali $\tilde{C}op_{m}^{j} \geq s$. La soglia *s*, di norma, è fissata a 0,7, ma dipende dal dominio *j* (ovvero dal peso del prodotto *j* rispetto all'indice aggregato di ordine superiore) e soprattutto dal mese di rilevazione (il tasso di risposta può risultare particolarmente basso ad agosto – dato che la maggior parte delle imprese rimangono chiuse anche per tre settimane – o nei mesi di chiusura dei bilanci aziendali, in cui le imprese sono impegnate nelle scritture contabili e quindi meno propense a rispondere tempestivamente).

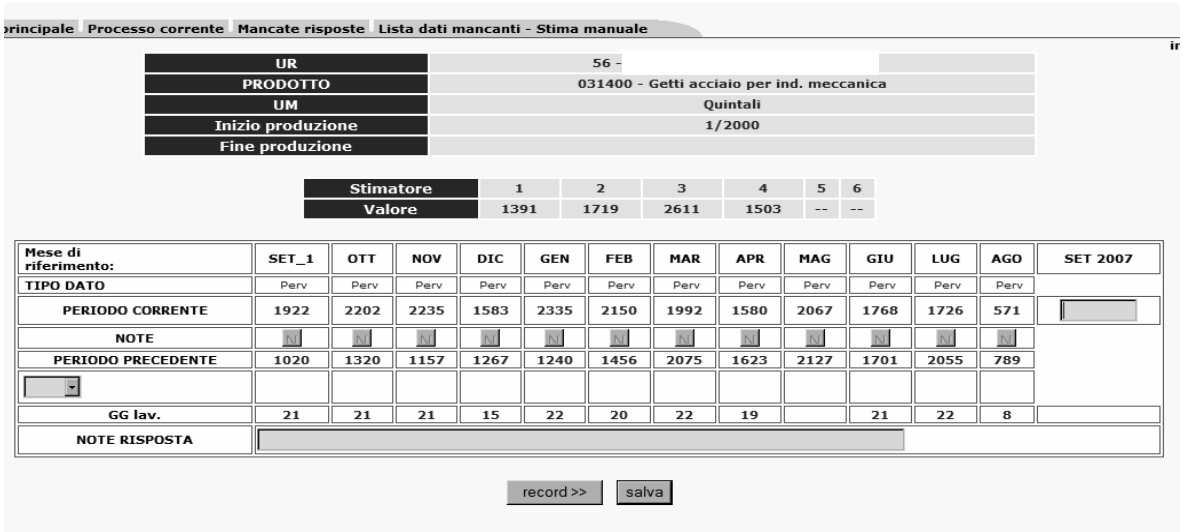
In questo caso viene applicata la stima automatica secondo il metodo proporzionale su base congiunturale (come suggerito, d'altra parte, dal manuale metodologico[119]): la stima di una risposta mancante si ottiene moltiplicando il dato del mese precedente (*m*-1) per il tasso di variazione congiunturale medio delle imprese rispondenti nel dominio *j*[120].

$${}^{j}\hat{y}_{Ami} = {}^{j}y_{A(m-1)i}\left(\frac{{}^{j}\bar{y}_{Am\Re}}{{}^{j}\bar{y}_{A(m-1)\Re}}\right)$$

In questo modo si assume implicitamente l'ipotesi, spesso confermata dal riscontro empirico, che nel breve periodo il "peso economico" delle imprese rimanga mediamente costante all'interno di ciascun dominio

- <u>interattive</u> per le unità rilevanti o dove la copertura dello strato è troppo bassa. La stima è basata soprattutto sull'osservazione della serie storica del microdato in questione o da deduzioni logiche (ad esempio, nel caso di prodotti padre per cui ci sia risposta per prodotti figlio).

Nella maschera per l'effettuazione della stima manuale il sistema, comunque, suggerisce 6 valori, calcolati con 6 stimatori diversi[121]:



Questi sei stimatori sono attualmente definiti come segue:

[119] EUROSTAT (2000).
[120] Si tratta, in pratica, dello stimatore 5 suggerito dalla procedura per la stima interattiva, come illustrato di seguito.
[121] Si veda GISMONDI et altri (2006).

1) Variazione congiunturale anno precedente stessa unità

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{A(m-1)i}\left(\frac{{}^{j}y_{(A-1)mi}}{{}^{j}y_{(A-1)(m-1)i}}\right) $$

2) Variazione tendenziale trimestrale stessa unità

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{Ami}\left(\frac{\mathrm{E}\left({}^{j}y_{A(m-3)i}, {}^{j}y_{A(m-2)i}, {}^{j}y_{A(m-1)i}\right)}{\mathrm{E}\left({}^{j}y_{(A-1)(m-3)i}, {}^{j}y_{(A-1)(m-2)i}, {}^{j}y_{(A-1)(m-1)i}\right)}\right) $$

3) Media variazioni congiunturali dei 2 anni precedenti stessa unità

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{Ami}\mathrm{E}\left(\frac{{}^{j}y_{(A-1)mi}}{{}^{j}y_{(A-1)(m-1)i}}, \frac{{}^{j}y_{(A-2)mi}}{{}^{j}y_{(A-2)(m-1)i}}\right) $$

4) Variazione tendenziale del bimestre precedente stessa unità

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{Ami}\left(\frac{\mathrm{E}\left({}^{j}y_{A(m-2)i}, {}^{j}y_{A(m-1)i}\right)}{\mathrm{E}\left({}^{j}y_{(A-1)(m-2)i}, {}^{j}y_{(A-1)(m-1)i}\right)}\right) $$

5) Variazione congiunturale valutata su unità rispondenti in m e (m-1) nello strato (se copertura adeguata)

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{A(m-1)i}\left(\frac{{}^{j}\bar{y}_{Am\Re}}{{}^{j}\bar{y}_{A(m-1)\Re}}\right) $$

Questo stimatore è utilizzato nella procedura automatica di stima dell'indagine.

6) Variazione tendenziale valutata su unità rispondenti in m e (m-12) nello strato (se copertura adeguata)

$$ {}^{j}\hat{y}_{Ami} = {}^{j}y_{(A-1)mi}\left(\frac{{}^{j}\bar{y}_{Am\Re}}{{}^{j}\bar{y}_{(A-1)m\Re}}\right) $$

Il revisore, eventualmente, può scegliere la stima più realistica, delle sei possibili suggerite dal sistema, come valore da imputare interattivamente al dato mancante influente.

I dati ottenuti per imputazione automatica o interattiva sono opportunamente individuati con un flag di stima che specifica il tipo di tecnica di imputazione utilizzata. Questi, infine, vengono sottoposti a revisione critica e nei casi in cui il coefficiente di variazione risulti anomalo, si procede ad una stima sostitutiva effettuata tenendo presente diversi fattori, quali:

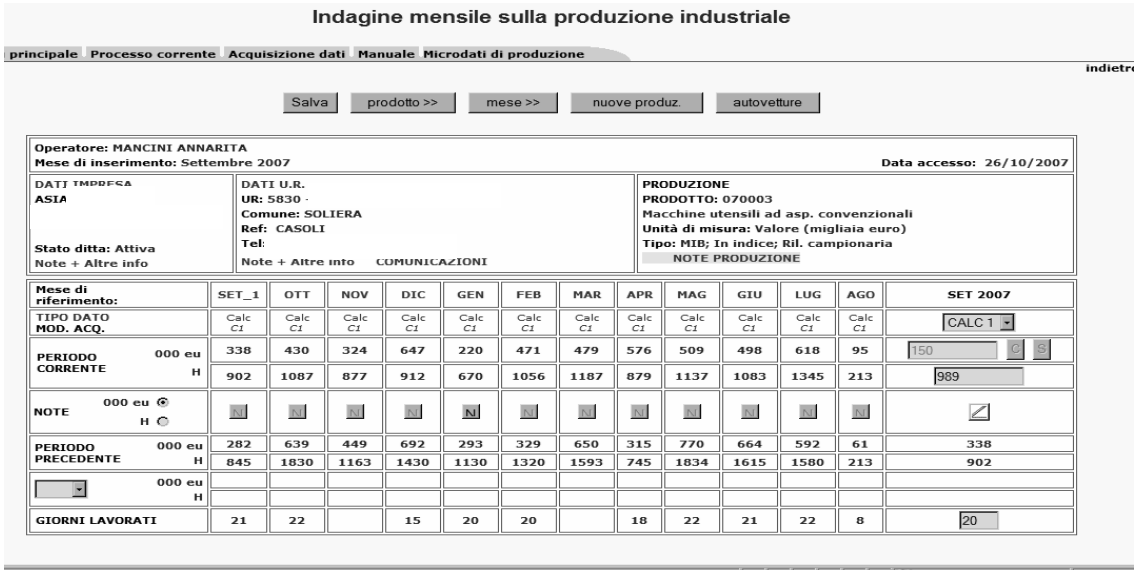
- la natura del bene,
- il livello produttivo dei mesi antecedenti quello della stima ed in particolare nel mese *m* dell'anno precedente,
- il numero dei giorni lavorati nel mese *m*,
- l'eventuale presenza di una componente stagionale,
- la rilevanza economica.

3.5.3 Mancate risposte parziali

Nel caso la mancata risposta riguardi solo la variabile ${}^{j}y_{mi}$ espressa nell'unità di misura di interesse per il calcolo dell'indice mensile e sia stata, invece, fornita l'altra grandezza (possibilità che riguarda solo i domini j[122] che prevedono la rilevazione con due unità di misura), si utilizza l'informazione longitudinale pervenuta per la misura ausiliaria stimando sulla base della sua variazione il livello di quella oggetto del computo mensile, apponendo un opportuno flag ("calcolato"). Tali

[122] Attualmente il 20% dei prodotti viene rilevato con 2 unità di misura.

microdati, comunque, non vengono considerati nel calcolo del tasso di mancata risposta in quanto sono considerati pervenuti e non *missing value*.



Un esempio tipico di MRP si osserva per i prodotti per i quali è richiesto il valore di produzione mensile espresso in migliaia di euro. Spesso le imprese hanno difficoltà a valorizzare tale grandezza e indicano il fatturato; quest'ultimo non è una *proxy* ideale dei quantitativi prodotti in quanto, tra l'altro, non considera il mese in cui effettivamente il prodotto è stato realizzato. Inoltre, non essendo contemplata nel modello una definizione specifica per questa variabile, spesso essa è fornita comprensiva dell'IVA nonché degli altri ricavi e proventi ordinari dell'azienda. Per questo motivo, consapevoli della difficoltà che un'unità rispondente può avere nel misurare tale grandezza, il modello viene corredato, per i prodotti rilevati in valore, di un'unità di misura ausiliaria che di solito è costituita dalle ore lavorate (dati di input) o alcune volte da "numeri di pezzi" prodotti o "peso" del prodotto a seconda della specificità del prodotto stesso.

Una volta scelta la tipologia di imputazione (in altre parole il tipo di informazione longitudinale che si vuole utilizzare nel calcolo della varabile di interesse) l'operazione viene impostata direttamente dal sistema informativo. Il revisore in aggiunta alla nota di produzione che descrive il metodo di imputazione scelto, nei mesi successivi avrà tale indicazione direttamente sulla maschera: l'opzione di calcolo sarà biffata in maniera automatica dal sistema. In questo modo si ha la certezza di utilizzare per lo stesso microdato un'identica tecnica di imputazione.

Di seguito si riporta l'interfaccia a disposizione del revisore per l'imputazione automatica di un dato calcolato. Ovviamente la maschera viene utilizzata anche per imputare dati pervenuti affetti da errori, per lo più sistematici (si veda sottoparagrafo 2.2.1).

Le prime tre opzioni fornite dal sistema si riferiscono principalmente alle correzioni di errori sistematici, mentre le ultime tre sono specifiche per le MRP. In dettaglio, le possibili scelte fornite dal sistema sono:

- congiunturale: il valore errato viene corretto applicando la variazione congiunturale osservata sul dato pervenuto,
- tendenziale: il valore errato viene corretto applicando la variazione tendenziale osservata sul dato pervenuto,
- quoziente il valore errato viene corretto applicando un coefficiente al dato pervenuto non corretto; un esempio tipico è quando il dato viene inviato in una unità di misura diversa da quella richiesta nel questionario per la quale, di conseguenza, occorre moltiplicare per un coefficiente di correzione (un esempio rilevante è nel settore 2411 - Fabbricazione di gas industriali, per il quale i prodotti rilevati dall'indagine hanno una tabella di conversione specifica per ciascuna unità di misura idonea a misurarli, si veda appendice),
- congiunturale II unità di misura: il valore mancante viene imputato applicando la variazione congiunturale osservata sulla variabile ausiliaria,
- tendenziale II unità di misura il valore mancante viene imputato applicando la variazione tendenziale osservata sulla variabile ausiliaria,
- quoziente II unità di misura il valore mancante viene imputato applicando un coefficiente al dato pervenuto per la variabile ausiliaria.

3.6 Macroediting preliminare

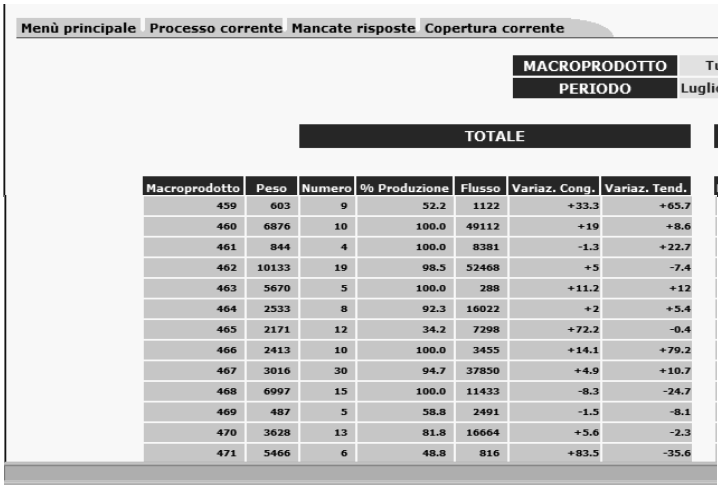
3.6.1 Verifica Copertura corrente

Una volta inseriti tutti i dati pervenuti e dopo aver effettuato tutte le imputazioni necessarie, ciascun revisore procede a effettuare controlli aggregati per macroprodotto[123] *k*, relativamente a quelli di cui è responsabile[124].

Per fare ciò il sistema informativo mette a disposizione una tabella denominata "*verifica copertura corrente*" che evidenzia contemporaneamente una serie di informazioni aggregate per tutti i microdati utilizzati per costruire l'indice mensile, escludendo, quindi, quelli in osservazione.

[123] Ricordiamo che un macroprodotto *k* può essere il risultato dell'aggregazione di uno o più prodotti *j*, alcuni dei quali possono essere realizzati anche da un'impresa diversa da quelle che il revisore segue personalmente.

[124] Sebbene un macroprodotto k o un dominio j possa essere visualizzato anche da più revisori la "responsabilità" di tutti i dati inclusi in quel macroprodotto è attribuita a uno solo.



Macroprodotto	Peso	Numero	% Produzione	Flusso	Variaz. Cong.	Variaz. Tend.
459	603	9	52.2	1122	+33.3	+65.7
460	6876	10	100.0	49112	+19	+8.6
461	844	4	100.0	8381	-1.3	+22.7
462	10133	19	98.5	52468	+5	-7.4
463	5670	5	100.0	288	+11.2	+12
464	2533	8	92.3	16022	+2	+5.4
465	2171	12	34.2	7298	+72.2	-0.4
466	2413	10	100.0	3455	+14.1	+79.2
467	3016	30	94.7	37850	+4.9	+10.7
468	6997	15	100.0	11433	-8.3	-24.7
469	487	5	58.8	2491	-1.5	-8.1
470	3628	13	81.8	16664	+5.6	-2.3
471	5466	6	48.8	816	+83.5	-35.6

In dettaglio in questo tabulato sono riportati:

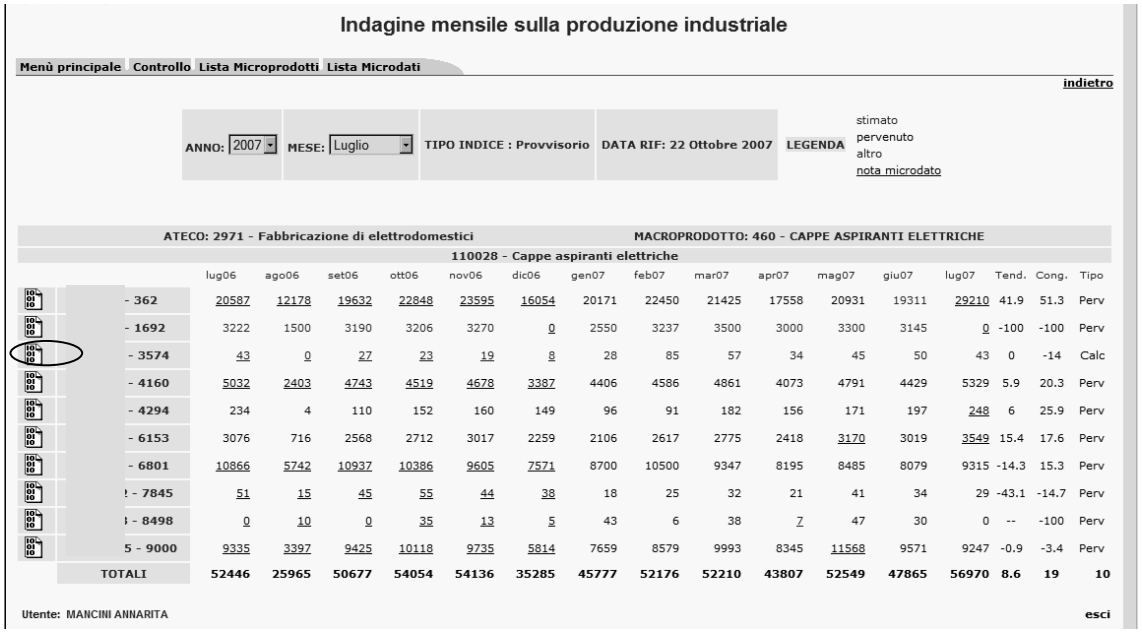
- l'identificazione del macroprodotto e/o dei macroprodotti selezionati,
- il peso del macroprodotto nella ponderazione dell'anno base
- il numero delle imprese che, in totale, entrano nel calcolo dell'indice elementare relativo a quel macroprodotto,
- la copertura, in termini di produzione, del macroprodotto calcolata su tutte le imprese disponibili[125] fino a quel momento,
- il flusso di produzione per ciascun macroprodotto, ovvero la somma dei dati assoluti relativi alle unità rispondenti[126] per quanto riguarda i prodotti che definiscono quel macroprodotto
- Le variazioni congiunturali e tendenziali di ogni macroprodotto.

Questo controllo, in effetti, può essere effettuato in qualunque momento del processo di rilevazione dei dati, selezionando a piacere anche uno o più domini *j* invece di scegliere l'aggregazione *k*, per verificare sia la copertura corrente, in quanto non sono presenti le stime dei dati mancanti, sia eventuali dati anomali che si evidenziano in qualsiasi fase dell'acquisizione in relazione al dominio *j* o alla serie *k*, punto di partenza per l'elaborazione degli indici elementari.

3.6.2 Maschera controllo microdati per dominio

Unitamente alla precedente tabella è possibile stampare la *lista microdati* che fornisce per ciascun prodotto *j*, la serie storica di tutte le unità rispondenti per quel prodotto evidenziando a video fino a m-13 mesi in modo da poter confrontare sia i dati tendenziali sia quelli congiunturali. Anche questo tabulato è costruito solo sulle informazioni utili al calcolo dell'indice mensile.

[125] Per disponibili si intendono tutti i dati immessi nel sistema e, quindi, pervenuti, forzati o calcolati.
[126] Relativamente alle sole unità inserite nel panel utilizzato per il calcolo dell'indice.



Per ogni unità rispondente sono evidenziati le variazioni congiunturali e tendenziali, l'inserimento di eventuali note di microdato (segnalate con la sottolineatura dello stesso valore e visibili a mezzo di *tooltips* evidenziabili posizionandosi sopra con il mouse) e la serie storica dei dati, graficamente diversificati a seconda che essi siano stimati, pervenuti, forzati o calcolati con possibilità di visualizzare anche la serie storica di ciascun microdato per tipo dato (pervenuto, calcolato, rettificato, forzato):

Questa maschera è utile soprattutto nel caso in cui si è fatto ricorso ad un'imputazione con flag "calcolato".

La lista microdati, di solito viene selezionata quando si vuole approfondire una particolare variazione relativa ad un dominio *j* o una sua aggregazione *k*, individuata nel precedente step di macroediting preliminare, per accertare quale unità rispondente abbia influito sull'andamento aggregato analizzato.

Tale lista può essere comunque selezionata dal revisore in qualunque momento della fase di rilevazione per controllare lo stato di ciascun dominio di sua competenza.

3.6.3 Serie storica microdato

Una volta individuata, dalla tabella precedente, una variazione o comunque un dato anomalo sul quale si ritiene opportuno indagare, prima di effettuare il *follow-up* o procedere a eventuale imputazione, può essere utile effettuare il controllo in serie storica per il microdato "incriminato"[127]. Si richiede quindi al sistema di visualizzarne la serie storica degli ultimi 6 anni, o comunque per il periodo per cui ci sia continuità di serie, riportando anche la serie delle variazioni tendenziali e congiunturali, con possibilità di effettuare grafici, sia sui livelli sia sulle variazioni, ed eventualmente segnalare gli scostamenti, su mese e/o su anno, superiori a soglie scelte al momento.

DITTA	9812	INIZIO PRODUZIONE	Gennaio 2000
UNITA' RISP	362 - [redacted]	FINE PRODUZIONE	
PRODOTTO	110028 - Cappe aspiranti elettriche	FINE PERIODO	Luglio 2007
ATECO	2971 - Fabbricazione di elettrodomestici	MACROPRODOTTO	460 - CAPPE ASPIRANTI ELETTRICHE - 6876

ricarica

DATI ASSOLUTI

	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic	TOTALE
2002	11802	13616	13942	12289	14035	13055	16380	4940	15235	17118	14577	11365	158354
2003	14213	14340	13401	12777	14988	14448	16387	6604	16600	20158	17916	13007	174839
2004	14844	20530	21596	19047	18870	18447	20437	7428	18394	17735	18945	13022	209295
2005	12335	18445	20829	19052	21631	18251	20550	10695	21893	19906	21420	13205	218212
2006	19500	20106	16150	16652	21076	19445	20587	12178	19632	22848	23595	16054	227823
2007	20171	22450	21425	17558	20931	19311	29210						151056

VARIAZIONI TENDENZIALI PERCENTUALI SOGLIA 30

	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic	MEDIA
2002	--	--	--	--	--	--	--	--	--	--	--	--	--
2003	20.4	5.3	-3.9	4	6.8	10.7	0	33.7	9	17.8	22.9	14.4	11.8
2004	4.4	43.2	61.2	49.1	25.9	27.7	24.7	12.5	10.8	-12	5.7	0.1	21.1
2005	-16.9	-10.2	-3.6	0	14.6	-1.1	0.6	44	19	12.2	13.1	1.4	6.1
2006	58.1	9	-22.5	-12.6	-2.6	6.5	0.2	13.9	-10.3	14.8	10.2	21.6	7.2
2007	3.4	11.7	32.7	5.4	-0.7	-0.7	41.9	--	--	--	--	--	13.4

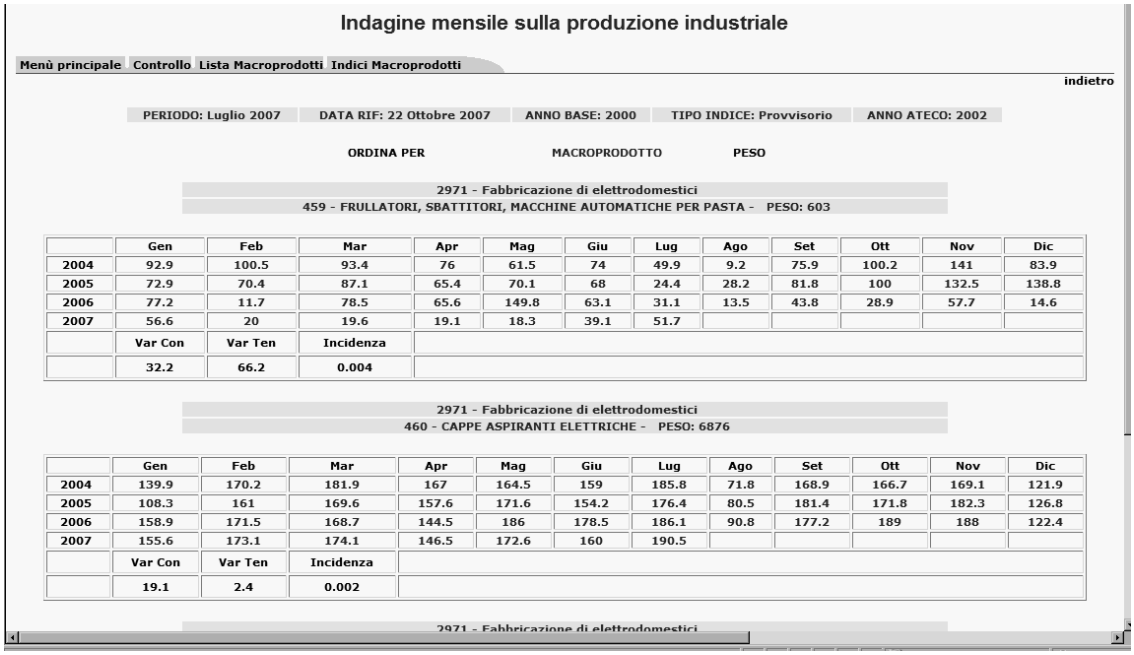
VARIAZIONI CONGIUNTURALI PERCENTUALI SOGLIA 50

	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic	MEDIA
2002	--	15.4	2.4	-11.9	14.2	-7	25.5	-69.8	208.4	12.4	-14.8	-22	13.9
2003	25.1	0.9	-6.5	-4.7	17.3	-3.6	13.4	-59.7	151.4	21.4	-11.1	-27.4	9.7
2004	14.1	38.3	5.2	-11.8	-0.9	-2.2	10.8	-63.7	147.6	-3.6	6.8	-31.3	9.1
2005	-5.3	49.5	12.9	-8.5	13.5	-15.6	12.6	-48	104.7	-9.1	7.6	-38.4	6.3
2006	47.7	3.1	-19.7	3.1	26.6	-7.7	5.9	-40.8	61.2	16.4	3.3	-32	5.6
2007	25.6	11.3	-4.6	-18	19.2	-7.7	51.3	--	--	--	--	--	11

3.6.4 Analisi serie storica macrodato

Si può richiedere al sistema di visualizzare la serie storica degli ultimi 4 anni degli indici elementari "ufficiali". Nello specifico, vengono definiti indici "ufficiali" quelli diffusi al più recente comunicato stampa riferito al mese in esame (si veda paragrafo 1.4 per la politica di diffusione relativa all'indagine sulla produzione industriale). Infatti, dopo la validazione di un indice mensile relativo al mese *m*, sia esso provvisorio, rettificato o revisionato, può accadere che uno o più valori relativi a quel mese subiscano una rettifica per effetto delle risposte tardive e/o correzioni di valori già validati, per cui i dati vengono immessi nel sistema come rettificati. Questi dati, quindi, sono visibili nella serie storica del microdato, ma non sono ancora utilizzati per il calcolo dell'indice del mese *m*: essi verranno utilizzati per la successiva revisione.

[127] L'unità rispondente che nell'ambito del macroprodotto e successivamente del microprodotto ha determinato la variazione o il dato sospetto.



Nel tabella sono riportate, oltre alle variazioni dell'indice, anche l'incidenza assoluta percentuale[128] rispetto all'indice generale.

Si può scegliere di visualizzare un singolo macroprodotto *k* o un insieme di questi, appartenenti per esempio ad una stessa ATECO, potendo ordinare le serie in ordine decrescente di peso, incidenza[129] o di codice identificativo di macroprodotto.

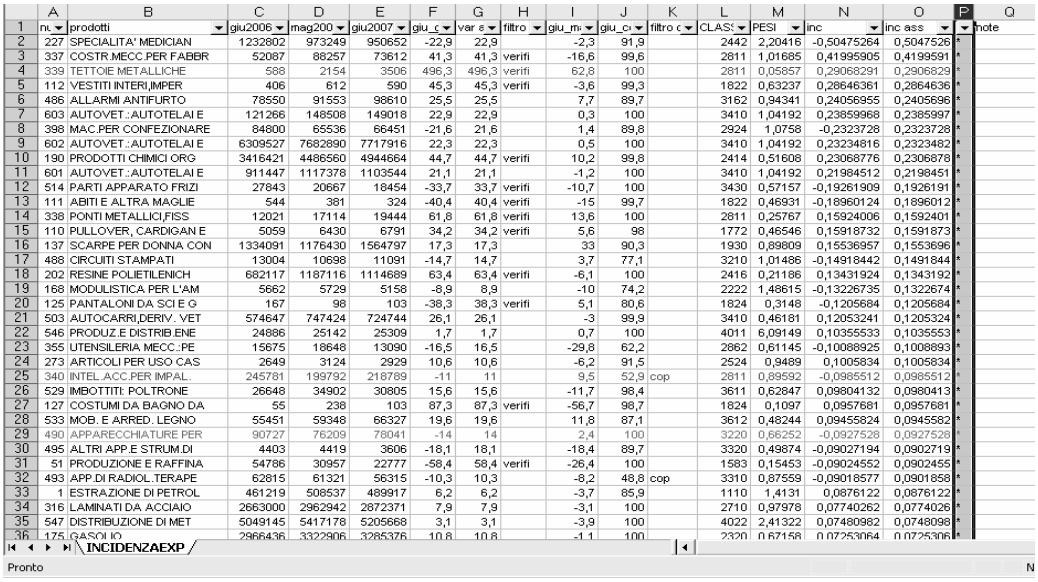
3.7 Macroediting finale

Circa tre/quattro giorni[130] prima della data di diffusione degli indici mensili a mezzo del comunicato stampa, viene effettuato dal responsabile di indagine il macroedinting finale. Attualmente le procedure sono ancora esterne al sistema informativo, per cui i dati caricati nel database Oracle vengono dapprima esportati in dataset SAS e poi un programma ad hoc sviluppato nello stesso ambiente permette di ottenere un listato in formato excel, editabile e idoneo a documentare i controlli effettuati in questa fase.

[128] Si veda paragrafo successivo per la definizione.

[129] Quest'ultima opzione è ancora in fase di sviluppo.

[130] In realtà il momento in cui viene effettuato il macroediting finale definitivo dipende dalla copertura ottenuta. In altre parole se un'unità particolarmente influente risulta ancora mancante, si cercherà di mirare le ultime telefonate di sollecito per ottenere tali dati prima della validazione dei dati grezzi provvisori.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	nr	prodotti	giu2006	mag200	giu2007	giu_c	var e	filtro	giu_m	giu_c	filtro c	CLAS	PESI	inc	inc ass	note	
2	227	SPECIALITA' MEDICIAN	1232802	973249	950652	-22,9	22,9		-2,3	91,9		2442	2,20416	-0,50475264	0,5047526	*	
3	337	COSTR.MECC.PER FABBR	52087	88257	73612	41,3	41,3	verifi	-16,6	99,6		2811	1,01685	0,41995905	0,4199591	*	
4	339	TETTOIE METALLICHE	588	2154	3506	496,3	496,3	verifi	62,8	100		2811	0,05857	0,29068291	0,2906829	*	
5	112	VESTITI INTERI,IMPER	406	612	590	45,3	45,3	verifi	-3,6	99,3		1822	0,63237	0,28646361	0,2864636	*	
6	486	ALLARMI ANTIFURTO	78550	91553	98610	25,5	25,5		7,7	89,7		3162	0,94341	0,24056955	0,2405696	*	
7	603	AUTOVET.:AUTOTELAI E	121266	148508	149018	22,9	22,9		0,3	100		3410	1,04192	0,23859868	0,2385997	*	
8	398	MAC.PER CONFEZIONARE	84800	65536	66451	-21,6	21,6		1,4	89,8		2924	1,0758	-0,2323728	0,2323728	*	
9	602	AUTOVET.:AUTOTELAI E	6309527	7682890	7717916	22,3	22,3		0,5	100		3410	1,04192	0,23234816	0,2323482	*	
10	190	PRODOTTI CHIMICI ORG	3416421	4486560	4944664	44,7	44,7	verifi	10,2	99,8		2414	0,51608	0,23068776	0,2306878	*	
11	601	AUTOVET.:AUTOTELAI E	911447	1117378	1103544	21,1	21,1		-1,2	100		3410	1,04192	0,21984512	0,2198451	*	
12	514	PARTI APPARATO FRIZI	27843	20667	18454	-33,7	33,7	verifi	-10,7	100		3430	0,57157	-0,19261909	0,1926191	*	
13	111	ABITI E ALTRA MAGLIE	544	381	324	-40,4	40,4	verifi	-15	99,7		1822	0,46931	-0,18960124	0,1896012	*	
14	338	PONTI METALLICI,FISS	12021	17114	19444	61,8	61,8	verifi	13,6	100		2811	0,25767	0,15924006	0,1592401	*	
15	110	PULLOVER, CARDIGAN E	5059	6430	6791	34,2	34,2	verifi	5,6	98		1772	0,46546	0,15918732	0,1591873	*	
16	137	SCARPE PER DONNA CON	1334091	1176430	1564797	17,3	17,3		33	90,3		1930	0,89809	0,15536957	0,1553696	*	
17	488	CIRCUITI STAMPATI	13004	10698	11091	-14,7	14,7		3,7	77,1		3210	1,01486	-0,14918442	0,1491844	*	
18	202	RESINE POLIETILENICH	682117	1187116	1114689	63,4	63,4	verifi	-6,1	100		2416	0,21186	0,13431924	0,1343192	*	
19	168	MODULISTICA PER L'AM	5662	5729	5158	-8,9	8,9		-10	74,2		2222	1,48615	-0,13226735	0,1322674	*	
20	125	PANTALONI DA SCI E G	167	98	103	-38,3	38,3	verifi	5,1	80,6		1824	0,3148	-0,1205684	0,1205684	*	
21	503	AUTOCARRI,DERIV. VET	574647	747424	724744	26,1	26,1		-3	99,9		3410	0,46181	0,12053241	0,1205324	*	
22	546	PRODUZ.E DISTRIB.ENE	24886	25142	25309	1,7	1,7		0,7	100		4011	6,09149	0,10355533	0,1035553	*	
23	355	UTENSILERIA MECC.:PE	15675	18648	13090	-16,5	16,5		-29,8	62,2		2862	0,61145	-0,10088925	0,1008893	*	
24	273	ARTICOLI PER USO CAS	2649	3124	2929	10,6	10,6		-6,2	91,5		2524	0,9489	0,1005834	0,1005834	*	
25	340	INTEL.ACC.PER IMPAL.	245781	199792	218789	-11	11		9,5	52,9	cop	2811	0,89592	-0,0985512	0,0985512	*	
26	529	IMBOTTITI: POLTRONE	26648	34902	30805	15,6	15,6		-11,7	98,4		3611	0,62847	0,09804132	0,0980413	*	
27	127	COSTUMI DA BAGNO DA	55	238	103	87,3	87,3	verifi	-56,7	98,7		1824	0,1097	0,0957681	0,0957681	*	
28	533	MOB. E ARRED. LEGNO	55451	59348	66327	19,6	19,6		11,8	87,1		3612	0,48244	0,09455824	0,0945582	*	
29	490	APPARECCHIATURE PER	90727	76209	78041	-14	14		2,4	100		3220	0,66252	-0,0927528	0,0927528	*	
30	495	ALTRI APP.E STRUM.DI	4403	4419	3606	-18,1	18,1		-18,4	89,7		3320	0,49874	-0,09027194	0,0902719	*	
31	51	PRODUZIONE E RAFFINA	54786	30957	22777	-58,4	58,4	verifi	-26,4	100		1583	0,15453	-0,09024552	0,0902455	*	
32	493	APP.DI RADIOL.TERAPE	62815	61321	56315	-10,3	10,3		-8,2	48,8	cop	3310	0,87559	-0,09018577	0,0901858	*	
33	1	ESTRAZIONE DI PETROL	461219	508537	489917	6,2	6,2		-3,7	85,9		1110	1,4131	0,0876122	0,0876122	*	
34	316	LAMINATI DA ACCIAIO	2663000	2962942	2872371	7,9	7,9		-3,1	100		2710	0,97978	0,07740262	0,0774026	*	
35	547	DISTRIBUZIONE DI MET	5049145	5417178	5205668	3,1	3,1		-3,9	100		4022	2,41322	0,07480982	0,0748098	*	
36	175	GASOLIO	2966436	3322906	3285376	10,8	10,8		-1,1	100		2320	0,67158	0,07253064	0,0725306	*	

INCIDENZAEXP

Pronto

In questo listato, sono riportate tutte le serie elementari (k) ordinate secondo valori decrescenti di incidenza assoluta, corredate da molte informazioni:

- flusso di produzione "ufficiale", ovvero diffuso, per lo stesso mese dell'anno precedente
- flusso di produzione "ufficiale", ovvero diffuso, per il mese precedente
- flusso di produzione corrente
- la variazione tendenziale percentuale (sul cui valore assoluto è fatto un filtro per evidenziare valori maggiori di 30%):

$$\left| \left(\frac{{}^{k}y_{Ami}}{{}^{k}y_{(A-1)mi}} - 1 \right) \times 100 \right| > 30\%$$

- la variazione congiunturale percentuale
- la copertura finale per ciascun macroprodotto (filtro per coperture inferiori al 60%)

$$\left(\frac{{}^{k}y_{Am\Re}}{{}^{k}y_{Am\Re} \bigcup {}^{k}y_{Am\overline{\Re}}} \times 100 \right) < 60\%$$

- la classe ATECO
- il peso
- contributo alla variazione dell'indice (filtro per incidenze superiori a 0.05%)

$$\left| \left(\frac{{}^{k}y_{Ami}}{{}^{k}y_{(A-1)mi}} - 1 \right) \times 100 \times {}^{k}W_{b} \right| > 0.05\% .$$

- Note (è una colonna vuota che verrà riempita con le annotazioni utili a interpretare i valori filtrati e per annotare eventuali *follow-up* o reimputazioni da fare).

Per ognuna delle serie k, per cui uno dei tre filtri imposti abbia dato esito positivo, viene effettuato un controllo presso i revisori e/o tramite l'esame diretto della serie storica per macroprodotto. Dopo tali verifiche, eventualmente, si effettuano *follow-up* o imputazioni per valori giudicati anomali.

Sono, inoltre, controllati in questa fase i valori calcolati e stimati più influenti i quali, eventualmente, vengono reimputati interattivamente[131].

[131] Eventuali reimputazioni di stime o di valori calcolati sono individuati opportunamente sul database.

Una volta validati i dati grezzi correnti si procede al calcolo dell'indice provvisorio grezzo, corretto e destagionalizzato, procedendo ad ulteriori controlli interattivi (variazioni aggregate su dati grezzi e corretti).

4. Il processo di C&C nel processo di diffusione degli indici grezzi rettificati

4.1 Schema generale: processo diffusione degli indici grezzi rettificati

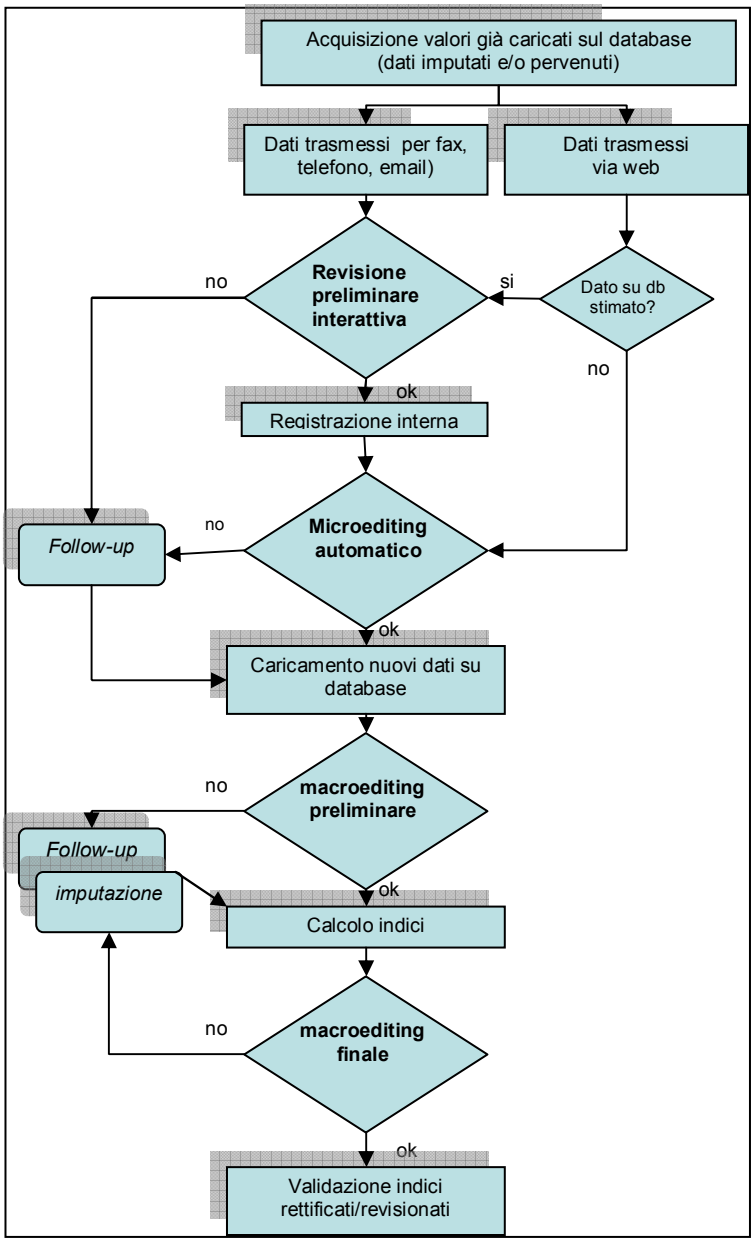
Analogamente a quanto fatto per il processo di elaborazione degli indici mensili provvisori si riporta nella Figura 4.1 il diagramma di flusso che sintetizza le fasi operative necessarie per diffondere gli indici rettificati.

Si ricorda che tali indici vengono diffusi al comunicato stampa successivo, quindi dopo ulteriori 30 giorni rispetto alla diffusione delle stime preliminari[132], e si basano soprattutto sull'aggiunta di risposte tardive (pervenuto su mancata risposta) anche se in questa fase possono essere inserite correzioni di valori già registrati sul database conseguenti a verifiche successive alla diffusione delle stime preliminari, effettuate sia dalle imprese sia dai revisori. Queste correzioni vengono caricate nel database come rettifiche che, in dettaglio, sostituiscono valori già registrati con etichetta pervenuto o calcolato o forzato.

La struttura del processo di elaborazione degli indici revisionati periodicamente è analoga a quella descritta in questo paragrafo.

[132] Tali dati vengono, quindi, lavorati in contemporanea con quelli relativi al mese t+1.

Figura 4.1 Schema generale del processo di diffusione degli indici grezzi rettificati



4.2 Revisione preliminare interattiva

Successivamente alla validazione dei dati grezzi provvisori relativi ad un mese *m*, si possono ricevere informazioni per lo stesso mese *m* sia per i dati che in precedenza erano stati oggetto di imputazione per mancata risposta sia per i dati già inseriti come pervenuti e che, quindi debbono subire una rettifica. Questi ultimi sono registrati sul database con un'etichetta specifica. Il primo controllo che viene svolto è quello di verificare l'etichetta del mese compilato, soprattutto se sul database ci sono valori registrati, per lo stesso mese *m*, come pervenuti. Questo perché il compilatore può aver sbagliato a digitare il mese di riferimento: nel senso che il rispondente ha compilato, per il questionario relativo al mese t, valori di pertinenza invece al mese successivo *m+1* .Di conseguenza, si potrebbe caricare erroneamente il dato come una rettifica di dato già pervenuto, per il mese *m*.

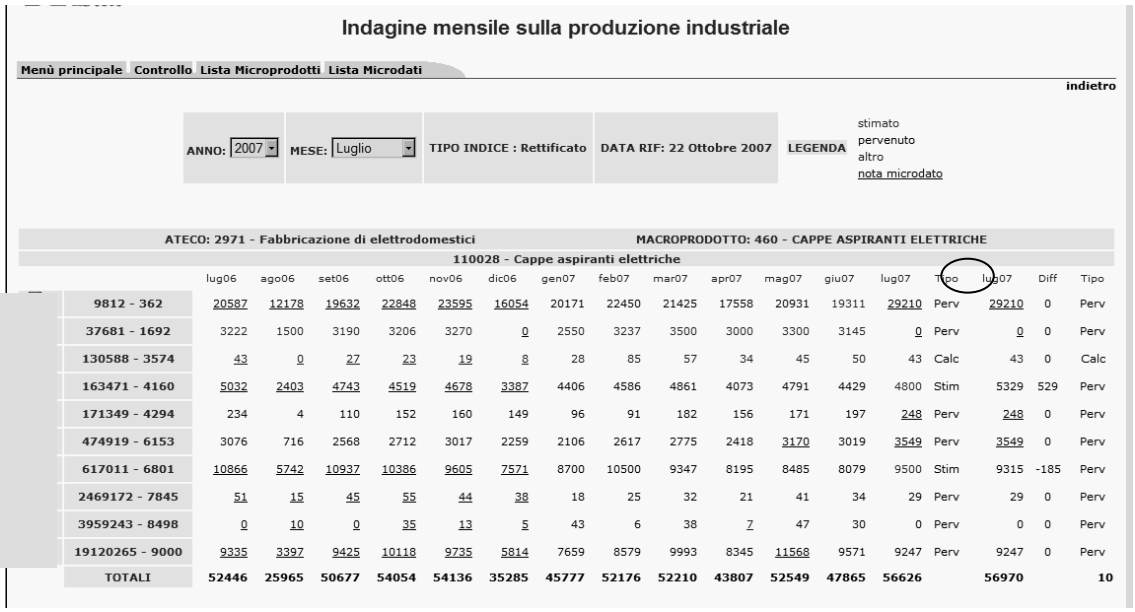
Successivamente, coadiuvati dalle maschere descritte nel paragrafo 3.3, si procede a confrontare i valori già validati, sia stimati che pervenuti, con quelli inseriti *ex novo* controllando principalmente i livelli.

Una volta caricati i dati, il database effettua, comunque, il *microediting* automatizzato così come descritto nel paragrafo 3.4.

Per i dati acquisiti via web, invece, l'inserimento automatico avviene solo nei casi in cui il dato era stato stimato: in questo caso si ha un controllo automatizzato identico a quello descritto nel paragrafo 3.4. Se invece il dato risulta già pervenuto non è catturato automaticamente dal sistema ma deve essere inserito a mano.

4.3 Macroeding preliminare per dominio

In questa fase si utilizza prettamente la lista microdati[133] per dominio che, anziché le variazioni di microdato, visualizza le differenze (revisioni) nei livelli tra i nuovi valori di produzione misurati il mese *m* che si sta sottoponendo a rettifica e quelli precedentemente inseriti, unitamente all'indicazione della tipologia di dato (stimato, pervenuto, ecc.).



Indagine mensile sulla produzione industriale

Menù principale | Controllo | Lista Microprodotti | Lista Microdati

indietro

ANNO: 2007 MESE: Luglio | TIPO INDICE : Rettificato | DATA RIF: 22 Ottobre 2007 | LEGENDA: stimato, pervenuto, altro, nota microdato

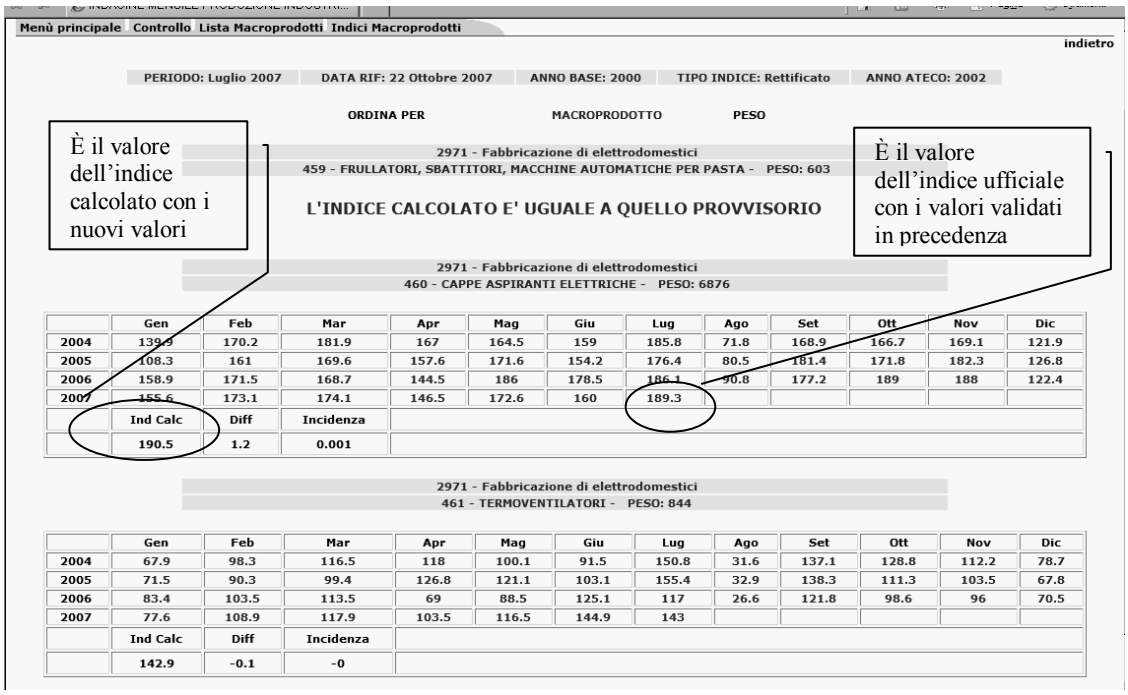
ATECO: 2971 - Fabbricazione di elettrodomestici — MACROPRODOTTO: 460 - CAPPE ASPIRANTI ELETTRICHE

110028 - Cappe aspiranti elettriche

	lug06	ago06	set06	ott06	nov06	dic06	gen07	feb07	mar07	apr07	mag07	giu07	lug07	Tipo	lug07	Diff	Tipo
9812 - 362	20587	12178	19632	22848	23595	16054	20171	22450	21425	17558	20931	19311	29210	Perv	29210	0	Perv
37681 - 1692	3222	1500	3190	3206	3270	0	2550	3237	3500	3000	3300	3145	0	Perv	0	0	Perv
130588 - 3574	43	0	27	23	19	8	28	85	57	34	45	50	43	Calc	43	0	Calc
163471 - 4160	5032	2403	4743	4519	4678	3387	4406	4586	4861	4073	4791	4429	4800	Stim	5329	529	Perv
171349 - 4294	234	4	110	152	160	149	96	91	182	156	171	197	248	Perv	248	0	Perv
474919 - 6153	3076	716	2568	2712	3017	2259	2106	2617	2775	2418	3170	3019	3549	Perv	3549	0	Perv
617011 - 6801	10866	5742	10937	10386	9605	7571	8700	10500	9347	8195	8485	8079	9500	Stim	9315	-185	Perv
2469172 - 7845	51	15	45	55	44	38	18	25	32	21	41	34	29	Perv	29	0	Perv
3959243 - 8498	0	10	0	35	13	5	43	6	38	7	47	30	0	Perv	0	0	Perv
19120265 - 9000	9335	3397	9425	10118	9735	5814	7659	8579	9993	8345	11568	9571	9247	Perv	9247	0	Perv
TOTALI	52446	25965	50677	54054	54136	35285	45777	52176	52210	43807	52549	47865	56626		56970		10

[133] Analogamente a quanto descritto per il macroediting preliminare del processo di validazione dei dati provvisori (paragrafo 3.6)

Unitamente a questa maschera è possibile visualizzare la serie storica ufficiale dell'indice elementare *k* e vedere quanto la revisione incide sull'indice generale (la formula è riportata nel successivo paragrafo):



	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic
2004	139.9	170.2	181.9	167	164.5	159	185.8	71.8	168.9	166.7	169.1	121.9
2005	108.3	161	169.6	157.6	171.6	154.2	176.4	80.5	181.4	171.8	182.3	126.8
2006	158.9	171.5	168.7	144.5	186	178.5	186.1	90.8	177.2	189	188	122.4
2007	155.6	173.1	174.1	146.5	172.6	160	189.3					

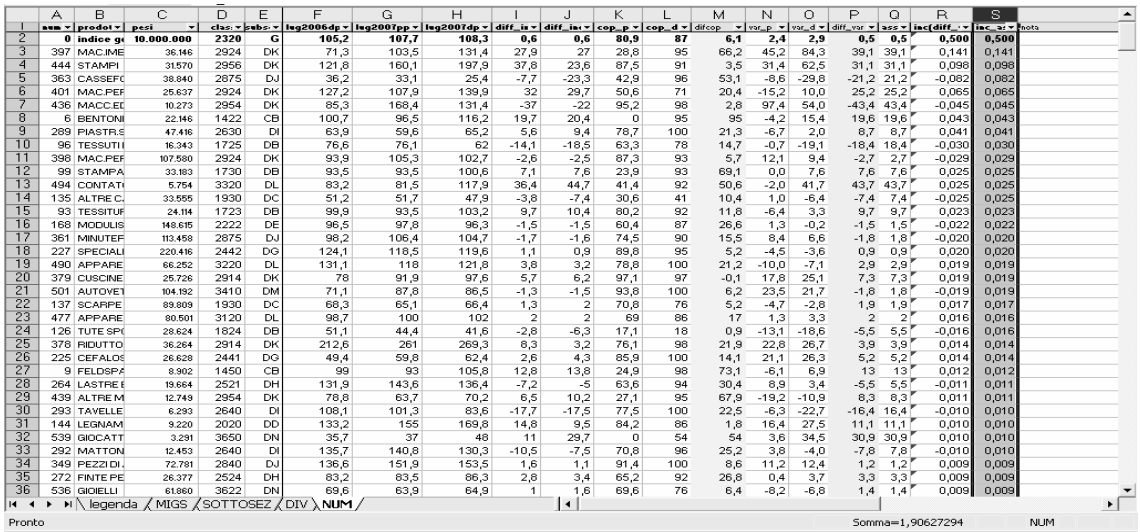
Ind Calc	Diff	Incidenza
190.5	1.2	0.001

	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic
2004	67.9	98.3	116.5	118	100.1	91.5	150.8	31.6	137.1	128.8	112.2	78.7
2005	71.5	90.3	99.4	126.8	121.1	103.1	155.4	32.9	138.3	111.3	103.5	67.8
2006	83.4	103.5	113.5	69	88.5	125.1	117	26.6	121.8	98.6	96	70.5
2007	77.6	108.9	117.9	103.5	116.5	144.9	143					

Ind Calc	Diff	Incidenza
142.9	-0.1	-0

4.4 Macroeding finale

Anche per questa fase di C&C, relativa alla fase di revisione degli indici, le procedure sono ancora esterne al sistema informativo per cui ci si avvale di un programma ad hoc sviluppato in SAS per ottenere un listato in formato excel con tutte le informazioni utili ai controlli da fare in questa fase.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	sum	prodot	pesi	clas	subc	leg2006dp	leg2007pp	leg2007dp	diff_is	diff_isc	cop_p	cop_d	difcop	var_p	var_d	diff_var	ass	isc(diff_	isc_st
2	0	indice g	10.000.000	2320	G	105,2	107,7	108,3	0,6	0,6	80,9	87	6,1	2,4	2,9	0,5	0,5	0,500	0,500
3	397	MAC.IME	36.146	2924	DK	71,3	103,5	131,4	27,9	27	28,8	95	66,2	45,2	84,3	39,1	39,1	0,141	0,141
4	444	STAMPI	31.570	2956	DK	121,8	160,1	197,9	37,8	23,6	87,5	91	3,5	31,4	62,5	31,1	31,1	0,098	0,098
5	363	CASSEFO	38.840	2875	DJ	36,2	33,1	25,4	-7,7	-23,3	42,9	96	53,1	-8,6	-29,8	-21,2	21,2	-0,082	0,062
6	401	MAC.PEF	25.637	2924	DK	127,2	107,9	139,9	32	29,7	50,6	71	20,4	-15,2	10,0	25,2	25,2	0,065	0,065
7	436	MACC.ED	10.273	2954	DK	85,3	168,4	131,4	-37	-22	95,2	98	2,8	97,4	54,0	-43,4	43,4	-0,045	0,045
8	6	BENTONI	22.146	1422	CB	100,7	96,5	116,2	19,7	20,4	0	95	95	-4,2	15,4	19,6	19,6	0,043	0,043
9	289	PIASTR.S	47.416	2630	DI	63,9	59,6	65,2	5,6	9,4	78,7	100	21,3	-6,7	2,0	8,7	8,7	0,041	0,041
10	96	TESSUTI	16.343	1725	DB	76,6	76,1	62	-14,1	-18,5	63,3	78	14,7	-0,7	-19,1	-18,4	18,4	-0,030	0,030
11	398	MAC.PEF	107.580	2924	DK	93,9	105,3	102,7	-2,6	-2,5	87,3	93	5,7	12,1	9,4	-2,7	2,7	-0,029	0,029
12	99	STAMPA	33.183	1730	DB	93,5	93,5	100,6	7,1	7,6	23,9	93	69,1	0,0	7,6	7,6	7,6	0,025	0,025
13	494	CONTATI	5.754	3320	DL	83,2	81,5	117,9	36,4	44,7	41,4	92	50,6	-2,0	41,7	43,7	43,7	0,025	0,025
14	135	ALTRE C.	33.555	1930	DC	51,2	51,7	47,9	-3,8	-7,4	30,6	41	10,4	1,0	-6,4	-7,4	7,4	-0,025	0,025
15	93	TESSITUF	24.114	1723	DB	99,9	93,5	103,2	9,7	10,4	80,2	92	11,8	-6,4	3,3	9,7	9,7	0,023	0,023
16	168	MODULIS	148.615	2222	DE	96,5	97,8	96,3	-1,5	-1,5	60,4	87	26,6	1,3	-0,2	-1,5	1,5	-0,022	0,022
17	361	MINUTEF	113.458	2875	DJ	98,2	106,4	104,7	-1,7	-1,6	74,5	90	15,5	8,4	6,6	-1,8	1,8	-0,020	0,020
18	227	SPECIALI	220.416	2442	DG	124,1	118,5	119,6	1,1	0,9	89,8	95	5,2	-4,5	-3,6	0,9	0,9	0,020	0,020
19	490	APPARE	66.252	3220	DL	131,1	118	121,8	3,8	3,2	78,8	100	21,2	-10,0	-7,1	2,9	2,9	0,019	0,019
20	379	CUSCINE	25.726	2914	DK	78	91,9	97,6	5,7	6,2	97,1	97	-0,1	17,8	25,1	7,3	7,3	0,019	0,019
21	501	AUTOVET	104.192	3410	DM	71,1	87,8	86,5	-1,3	-1,5	93,8	100	6,2	23,5	21,7	-1,8	1,8	-0,019	0,019
22	137	SCARPE	89.809	1930	DC	68,3	65,1	66,4	1,3	2	70,8	76	5,2	-4,7	-2,8	1,9	1,9	0,017	0,017
23	477	APPARE	80.501	3120	DL	98,7	100	102	2	2	69	86	17	1,3	3,3	2	2	0,016	0,016
24	126	TUTE SP(	28.624	1824	DB	51,1	44,4	41,6	-2,8	-6,3	17,1	18	0,9	-13,1	-18,6	-5,5	5,5	-0,016	0,016
25	378	RIDUTTO	36.264	2914	DK	212,6	261	269,3	8,3	3,2	76,1	98	21,9	22,8	26,7	3,9	3,9	0,014	0,014
26	225	CEFALOS	26.628	2441	DG	49,4	59,8	62,4	2,6	4,3	85,9	100	14,1	21,1	26,3	5,2	5,2	0,014	0,014
27	9	FELDSPA	8.902	1450	CB	99	93	105,8	12,8	13,8	24,9	98	73,1	-6,1	6,9	13	13	0,012	0,012
28	264	LASTRE E	19.664	2521	DH	131,9	143,6	136,4	-7,2	-5	63,6	94	30,4	8,9	3,4	-5,5	5,5	-0,011	0,011
29	439	ALTRE M	12.749	2954	DK	78,8	63,7	70,2	6,5	10,2	27,1	95	67,9	-19,2	-10,9	8,3	8,3	0,011	0,011
30	293	TAVELLE	6.293	2640	DI	108,1	101,3	83,6	-17,7	-17,5	77,5	100	22,5	-6,3	-22,7	-16,4	16,4	-0,010	0,010
31	144	LEGNAM	9.220	2020	DD	133,2	155	169,8	14,8	9,5	84,2	86	1,8	16,4	27,5	11,1	11,1	0,010	0,010
32	539	GIOCATT	3.291	3650	DN	35,7	37	48	11	29,7	0	54	54	3,6	34,5	30,9	30,9	0,010	0,010
33	292	MATTON	12.453	2640	DI	135,7	140,8	130,3	-10,5	-7,5	70,8	96	25,2	3,8	-4,0	-7,8	7,8	-0,010	0,010
34	349	PEZZI DI.	72.781	2840	DJ	136,6	151,9	153,5	1,6	1,1	91,4	100	8,6	11,2	12,4	1,2	1,2	0,009	0,009
35	272	FINTE PE	26.377	2524	DH	83,2	83,5	86,3	2,8	3,4	65,2	92	26,8	0,4	3,7	3,3	3,3	0,009	0,009
36	536	GIOIELLI	61.860	3622	DN	69,6	63,9	64,9	1	1,6	69,6	76	6,4	-8,2	-6,8	1,4	1,4	0,009	0,009

In questo listato, in dettaglio, sono riportate tutte le serie elementari (*k*) ordinate secondo valori decrescenti di incidenza percentuale assoluta calcolata, in questo caso, sulla differenza tra le variazioni tendenziali calcolate con i valori degli indici ${}^{k}I_{Am}^{R}$ ottenuti con i nuovi dati giunti successivamente

rispetto alla validazione dei dati provvisori e quelle calcolate con gli indici diffusi con l'ultimo comunicato stampa [134] relativamente allo stesso mese *m* in analisi():

$$ {}^{k}inc^{R} = \left| \left(\frac{{}^{k}I^{R}_{Am}}{{}^{k}I_{(A-1)m}} - \frac{{}^{k}I^{P}_{Am}}{{}^{k}I_{(A-1)m}} \right) \times 100 \times {}^{k}W_{b} \right| $$

Il listato, inoltre, fornisce per ogni macroprodotto k:

- la classe ATECO
- il peso
- Il livello dell'indice elementare "ufficiale" ${}^{k}I^{P}_{(A-1)m}$ per lo stesso mese *m* dell'anno precedente
- Il livello dell'indice elementare "ufficiale" diffuso per lo stesso mese *m* con il precedente comunicato stampa
- Il livello dell'indice corrente sempre per lo stesso mese *m*
- La differenza tra i livelli degli indici per il mese m ${}^{k}I^{R}_{Am} - {}^{k}I^{P}_{Am}$ e la stessa espressa in termini percentuali $(\frac{{}^{k}I^{R}_{Am} - {}^{k}I^{P}_{Am}}{{}^{k}I^{P}_{Am}} \times 100)$
- La copertura finale relativamente al mese *m* nei due momenti di diffusione e la loro differenza
- la variazione tendenziale percentuale calcolata sempre relativamente al mese *m* nei due momenti di diffusione
- l'incidenza ${}^{k}inc^{R}$
- Note, è una colonna vuota che verrà riempita con le annotazioni utili a interpretare i valori cambiati rispetto a quelli precedentemente validati e per annotare eventuali *follow-up* o reimputazioni da fare: in questo caso le reimputazioni riguardano soprattutto il ripristino dei dati già validati in precedenza se non si riesce a effettuare *follow-up* in tempo utile.

In aggiunta al precedente listato, un altro, programma sviluppato sempre in SAS permette di estrarre, solo per i macroprodotti k per i quali ${}^{k}inc^{R}$ sia superiore a 0,001%, la serie storica completa di tutti microdati dello strato k con la segnalazione dei microdati che sono cambiati. Per questi, più in dettaglio, viene riportata la serie storica completa con le due colonne relative ai valori caricati sul database, per lo stesso mese *m*, nei due momenti diversi (similmente a quanto descritto nel paragrafo 4.3).

5. Gli sviluppi futuri

5.1 Premessa

La recente migrazione su piattaforma Oracle dei dati della Rilevazione mensile sulla produzione industriale ha consentito di recepire e di implementare alcune innovazioni riguardanti diversi aspetti dell'indagine come già ampiamente illustrato in precedenza: modello di rilevazione personalizzato, acquisizione dati via web, metodo dei quartili per l'individuazione degli outlier, criteri aggiornati per la definizione delle imprese rilevanti e degli stimatori ottimali per l'imputazione delle mancate risposte. Tali innovazioni sono il frutto di un ampio lavoro di ristrutturazione dell'indagine intrapreso alla fine del 2004 che ha visto la collaborazione del responsabile e degli esperti di indagine nonché dei metodologi e degli informatici della Direzione Centrale delle Statistiche congiunturali su imprese, servizi e costruzioni. I primi risultati sono stati presentati nel corso del seminario dal titolo "La

[134] In effetti ${}^{k}I^{P}_{Am}$ si riferisce, comunque, all'ultimo valore dell'indice diffuso a mezzo di comunicati stampa e a seconda se si stanno diffondendo dati rettificati o revisionati (si veda paragrafo 1.4 per la politica delle revisioni) indicherà indici provvisori o rettificati/revisionati.

rilevazione mensile della produzione industriale: aggiornamento metodologico e disegno del nuovo sistema informativo" che si è svolto nel mese di marzo 2006.

In quel momento il sistema informativo della produzione industriale (di seguito indicato con SIPI) era ancora nella fase di progettazione e test; il lavoro è attualmente ancora in corso nell'ottica di perfezionare le funzionalità base e definirne delle nuove più specializzate. Il progetto, sicuramente molto ambizioso, prevede che il complesso insieme di procedure utilizzate in tutte le fasi dell'indagine siano integrate in un'unica procedura per individuare con certezza i tempi e i responsabili di ogni attività[135].

Il prodotto finale sarà corredato da "un'area simulazioni" che permetterà di valutare gli effetti sull'indice generale delle variazioni di alcuni parametri quali le modificazioni nei dati, nel panel delle imprese rispondenti, nel paniere dei prodotti, nei pesi utilizzati, nell'anno base di riferimento e così via.

Nel seguito l'attenzione sarà limitata a ciò che ancora resta da completare rispetto alle procedure C&C e si definiranno gli interventi pianificati nel breve periodo rispetto alle varie fasi individuate dallo schema generale di C&C di pag 11.

5.2 Il microediting

Per quanto riguarda il microediting alcuni interventi volti a migliorare e ad integrare le attuali procedure, sono stati definiti sulla base delle evidenze emerse nel corso dei primi mesi di utilizzo del sistema SIPI. Qui di seguito sarà presentato un breve elenco di tali attività.

1. Per quanto riguarda l'acquisizione dei dati via web si ritiene utile fornire alle imprese in fase di compilazione dei dati on-line, l'accesso al repertorio dei prodotti rientrati nel campo di osservazione dell'indagine, limitato inizialmente alla classe ATECO di appartenenza dell'impresa ma estendibile a tutti i domini rilevati. In questo modo si faciliterà il lavoro del compilatore del questionario nel caso in cui si trovi a dover dichiarare una nuova produzione. L'ideale sarebbe poter disporre di un elenco all'interno del quale navigare tramite funzioni di ricerca più o meno complesse. Altra esigenza manifestata dalle imprese è quella di ricevere una conferma dell'avvenuta acquisizione dei dati che possa essere archiviata come documentazione interna. A tal proposito si sta valutando la possibilità di inserire, nella pagina per la compilazione on-line del questionario, uno schema che elenchi i questionari inviati negli ultimi 12 mesi e le relative date di invio.
2. In merito all'individuazione degli outlier si potrebbe considerare la possibilità di utilizzare dei limiti di accettazione derivanti dalle variazioni congiunturali[136] in combinazione con quelli derivanti dalle variazioni tendenziali. Primi tentativi condotti in tal senso hanno evidenziato che gli intervalli tendono ad essere molto restrittivi, generando di conseguenza over-editing. Sembra comunque interessante, indagare sulla possibilità di utilizzare tutte le informazioni disponibili al fine di individuare gli outlier. Alla fine del prossimo anno, inoltre, quando i dati registrati con il nuovo sistema SIPI copriranno un orizzonte temporale più vasto, si potrà valutare l'opportunità di condurre uno studio sulle giustificazioni fornite dalle imprese per i valori individuati come potenziali outlier ma rivelatisi corretti in seguito al follow-up. In questo modo si potrebbero predisporre alcune voci predefinite che permetterebbero di velocizzare le attività di microediting e di quantificare gli eventi che determinano andamenti particolari per alcuni settori e/o mesi specifici.
3. Per quanto riguarda il trattamento delle mancate risposte totali, attualmente si procede ad una imputazione automatica per le unità non rilevanti e ad una imputazione interattiva per le unità rilevanti, come già descritto in precedenza. E' tuttavia, in corso uno studio che ha l'obiettivo di individuare uno stimatore ottimale che possa variare in base al dominio di stima e al mese di riferimento (Gismondi, Carone, 2006). Il sistema SIPI è stato predisposto per accogliere sei stimatori ma sono necessarie ulteriori simulazioni volte a valutare il grado di accuratezza che essi forniscono. Questa ultima innovazione resta in fase di sperimentazione poiché solo con il passaggio

[135] Il sistema permette di risalire alla data e all'ora di tutte le modifiche e all'operatore che le ha realizzate. L'idea alla base dell'attuale struttura del database SIPI è che dati relativi alla singola impresa ed al singolo prodotto non vadano a sovrascrivere i precedenti già validati ma si aggiungano ad essi determinando un aggiornamento del flag dell'inserimento.

[136] escludendo i mesi di luglio, di settembre, di novembre e di gennaio poiché si confronterebbero con i mesi di agosto e dicembre il cui andamento dipende molto dai periodi di chiusura aziendali.

alla base 2005=100 potrà essere introdotta nel processo corrente, al fine di non generare break nelle serie storiche attualmente diffuse agli utilizzatori. I sei stimatori proposti dal sistema vengono, dunque, utilizzati esclusivamente dai revisori per la stima interattiva delle unità influenti (non sono utilizzati cioè dalla procedura di stima automatica).

4. Con riferimento all'individuazione delle unità rilevanti si sta intraprendendo un ulteriore studio il cui obiettivo è individuare quelle influenti per ciascun microprodotto. Allo stato attuale, infatti, un'unità che realizza più produzioni e che risulta influente per una di queste viene considerata influente per tutti i prodotti che essa fornisce. Questo tipo di classificazione permette di definire delle strategie di *follow-up* efficienti ma limita l'utilizzo delle stima automatica che è utilizzata solo per le imprese non rilevanti. Si tratta di un ulteriore sviluppo del lavoro presentato recentemente nel corso del seminario sulle stime anticipate (Gismondi- Carone, 2007). Il lavoro originale prevedeva, tra le diverse opzioni presentate, l'individuazione delle imprese influenti attraverso la definizione di funzioni di rischio basate sull'indice generale calcolato escludendo di volta in volta una singola impresa. Lo sviluppo metodologico prevede che le simulazioni siano fatte escludendo dal calcolo dell'indice di volta in volta le singole microproduzioni.
5. Infine, un'altra attività che è in corso riguarda il trattamento delle mancate risposte totali: si tratta della codifica dei motivi della non risposta. Fino ad oggi, infatti, tali informazioni sono archiviate come testo libero in una nota che accompagna tutte le unità rispondenti rendendo poi di fatto estremamente difficoltoso quantificare l'attrito, l'impatto degli eventi demografici, gli errori di classificazione delle imprese e della lista di spedizione. La codifica di questi eventi, permetterà di calcolare alcuni indicatori sul processo corrente come ad esempio, il numero di unità divenute non eleggibili nelle diverse occasioni di indagine a causa di variazioni di stato o di cessazione dell'attività.

5.3 Il macroediting

Le attività di controllo dei dati a livello macro allo stato attuale non sono ancora del tutto integrate nel SIPI, anche se lo saranno nel brevissimo termine, in quanto restano da completare le funzionalità del macroediting finale che sono svolte direttamente dai responsabili di indagine utilizzando delle apposite procedure SAS come già illustrato nel paragrafo 3.7.

In fase di realizzazione del sistema informativo, infatti, si è data la precedenza a tutte le attività svolte dai revisori (inserimento e verifica dati, solleciti imprese, stime mancate risposte, elaborazione indici elementari), in quanto a monte dell'intero processo produttivo e non più realizzabili all'esterno del database in seguito alla migrazione nel nuovo sistema. All'inizio del 2007, inoltre, la necessità di passare all'utilizzo del modello personalizzato e sopratutto all'acquisizione dei dati via web sono apparsi non più procastinabili portando alla decisione di passare al nuovo sistema SIPI sebbene non fossero state completate tutte le funzionalità previste in fase di progettazione.

I prossimi sviluppi riguardanti il macroediting finale sono indicati qui di seguito:

1. Gli output generati dai programmi SAS saranno integrati nel SIPI e avranno il grande vantaggio di permettere una navigazione di tipo ipertestuale che partendo dal macrodato consenta di arrivare con pochi e semplici passaggi ai microdati che lo hanno generato. La strategia di controllo dei dati, infatti, anche a livello macro è svolta dal basso verso l'alto poiché si verificano prima gli indici elementari e poi le aggregazioni successive (classi, gruppi, divisioni, sottosezioni e sezioni Ateco) nell'ottica di ottenere serie elementari validate.
2. Al termine di questa fase si potrà cercare di aggiungere delle funzioni che aiutino ulteriormente il ricercatore nella valutazione degli andamenti ottenuti. Ad esempio, il risultato del benchmarking con i dati provenienti da fonti esterne potrà essere archiviato così come le eventuali attività di controllo che da esso sono scaturite. Un esempio ulteriore potrebbe essere il confronto tra gli andamenti mensili dei settori integrati verticalmente[137] anche se questa informazione dovrà poi essere necessariamente integrata con l'andamento degli ordinativi e dell'import/export. Lo studio necessario per il prossimo cambiamento dell'anno base volto a recepire le modifiche della classificazione delle attività economiche, potrà rappresentare l'occasione per ipotizzare, definire e codificare le eventuali relazioni fra i diversi processi produttivi.

[137] Come ad esempio la produzione di carta da giornale e la stampa dei quotidiani.

3. Nel controllo dei macrodati finali le attività di verifica si concentrano sui flussi mensili e sugli indici elementari che essi permettono di calcolare. Occorrerebbe prevedere un flag nel database che permetta di distinguere se un indice elementare è stato verificato e quale ne sia stato il motivo: le coperture basse, le variazioni tendenziali o le incidenze elevate come già avviene nella procedura esterna descritta in precedenza.
4. Resta, infine, da integrare nel SIPI la procedura che fornisce i grafici delle serie storiche degli indici ai diversi livelli di aggregazione. Saranno disponibili dunque, i 548 grafici degli indici elementari e quelli per le aggregazioni superiori con la possibilità di visualizzare un grafico per volta oppure dei gruppi di grafici. Le serie mostreranno l'andamento degli ultimi quattro anni disponibili con la possibilità di scegliere periodi più ampi o più brevi.

5.4 Gli indicatori prodotti per il sistema SIDI

In SIDI "Il sistema di documentazione delle indagini dell'Istat", sono definiti i seguenti gruppi di indicatori: Copertura e Mancata Risposta, Controllo e Correzione, Tempestività e Puntualità, Costi, Controllo e Correzione per Variabile, Codifica, Coerenza con Fonti Esterne, Coerenza tra Dati Provvisori e Dati Definitivi, Confrontabilità e Revisione dei dati finali.

Attualmente per l'indagine mensile sulla produzione industriale il sistema risulta popolato per gli indicatori di copertura e mancata risposta, di tempestività e puntualità, per gli indicatori sui costi e sulla coerenza tra dati provvisori e dati definitivi.

Il prossimo piano di immissione degli indicatori non prevede il calcolo degli indicatori su controllo e correzione poiché richiederebbe delle modifiche al processo produttivo molto costose.

Se, infatti, un dato è stato indicato in maniera non corretta da un'impresa, segnalato dal sistema e verificato tramite contatto diretto, il revisore attualmente inserisce solo il dato finale già revisionato. Tenere traccia di queste attività, per quanto possibile apportando delle modifiche all'attuale database, comporterebbe una modifica del processo produttivo particolarmente onerosa in termini di risorse e di tempo impiegato che mal si concilia con le richieste di sempre maggiore tempestività avanzate in sede comunitaria.

Gli indicatori previsti dal SIDI sono:

1. **Tasso di Imputazione:** $\frac{\text{Valori Imputati}}{\text{Totale record}} * 100$
2. **Tasso di Modificazione:** $\frac{\text{Valori Modificati da Codice a Codice diverso}}{\text{Totale record}} * 100$
3. **Tasso di Imputazione Netta:** $\frac{\text{Valori Modificati da Blank a Codice}}{\text{Totale record}} * 100$
4. **Tasso di Cancellazione:** $\frac{\text{Valori Modificati da Codice a Blank}}{\text{Totale record}} * 100$
5. **Tasso di Non Imputazione:** $\frac{\text{Valori Non Imputati}}{\text{Totale record}} * 100$
6. **Tasso di Valori Non Blank Immutati:** $\frac{\text{Valori Non Blank Non Imputati}}{\text{Totale record}} * 100$
7. **Tasso di Valori Blank Immutati:** $\frac{\text{Valori Blank Non Imputati}}{\text{Totale record}} * 100$

Si tratta di indicatori sicuramente più adatti per le indagini strutturali che si basano su questionari complessi e utilizzano piani di C&C automatizzati.

Esaminando gli indicatori proposti si potrebbe pensare, comunque, di calcolare il tasso di valori non blank immutati (6) e il tasso di modificazione (2) derivante dalle attività di macro e di microediting. I "valori modificati da codice a codice diverso" possono essere, infatti, assimilati ai valori che vengono modificati e codificati con il flag calcolato o forzato. Non bisogna, però, dimenticare che si tratterebbe di una misura approssimata poiché mancano le correzioni derivanti dal contatto diretto con le imprese.

Il tasso di imputazione (1) e il relativo tasso di non imputazione (5), invece non possono essere calcolati poiché i dati non sono sottoposti a piani di C&C completamente automatici e quindi non è definito il numero di valori sui quali hanno agito le regole d'imputazione e volendo ricorrere ad un'approssimazione finirebbero per coincidere con il tasso di modificazione. Nemmeno il tasso di imputazione netta (3) può essere calcolato poiché "il cambiamento da blank a nuovo valore non blank" coincide con la stima delle mancate risposte totali. I rimanenti tassi di cancellazione (4) e di valori blank immutati (7) non possono essere, infine, calcolati in quanto nessun dato che concorre al calcolo dell'indice può essere cancellato per diventare un blank e nessun blank può rimanere tale trattandosi di una mancata risposta totale.

5.5 La documentazione interna sul processo di controllo e correzione

Come già evidenziato più volte in precedenza, se moltissimo è stato fatto in fase di progettazione e realizzazione del sistema informativo SIPI per avere traccia di ogni attività realizzata sui microdati, restano ancora da formalizzare gli strumenti idonei a fornire una documentazione completa sul processo di C&C.

Secondo il manuale di pratiche raccomandate affinché un processo di C&C si possa ritenere documentato devono essere prodotti:

- un manuale metodologico,
- un sistema di indicatori di qualità
- ed un sistema di indicatori per l'archiviazione dei dati.

Il manuale metodologico è in fase di preparazione e conterrà molte delle indicazioni sul processo di controllo e correzione riportate in questo documento, nonché indicazioni di carattere pratico volte a ridurre il rischio di editing creativo da parte del personale in forza all'unità operativa. Allo stato attuale tale rischio è considerato piuttosto basso poiché il personale ha maturato mediamente più di 15 anni di esperienza nel complesso meccanismo che parte dall'acquisizione dei dati per arrivare all'elaborazione degli indici elementari e il controllo finale, realizzato dai ricercatori in fase di macroediting, permette di ridurre ulteriormente tale rischio. La realizzazione del manuale può essere considerata, dunque, frutto della necessità di documentare le attività svolte e di formare eventuale nuovo personale.

Nella realizzazione del manuale sono stati coinvolti anche i revisori che hanno partecipato alla fase di progettazione iniziale del sistema informativo.

Per quanto riguarda il sistema degli indicatori di qualità, il cui obiettivo è informare gli utilizzatori sui principali aspetti del processo di C&C e sulle modifiche che questi determinano nella qualità dei dati finali, verrà valutata la possibilità di elaborare degli indicatori, oltre a quelli richiesti da SIDI, pesati e non pesati, sui tassi di risposta e di imputazione così come indicato dal manuale di pratiche raccomandate per il controllo e la correzione dei dati. Tali indicatori potrebbero essere estesi anche alle variabili ausiliare permettendo di produrre indicatori del tipo "item non response".

Occorre ricordare, come già ampiamente illustrato nel paragrafo 3, che attualmente il SIPI permette di registrare tutte le informazioni riguardanti le imprese e i prodotti che esse forniscono, mentre le procedure esterne al sistema permettono di tener traccia dei controlli (e dei relativi esiti) effettuati in fase di macroediting finale.

In fase di microediting, infatti, si dispone attualmente di tre differenti tipologie di note in cui è possibile inserire del testo libero:

1. la nota che riguarda il microprodotto che è associata a ciascun mese e a ciascuna tipologia di dato inserito. Ad esempio, è possibile inserire una nota per il dato stimato, una per il dato successivamente pervenuto ed una per l'eventuale dato rettificato e questo per tutti i mesi dell'anno. Tale nota è compilabile anche in riferimento alla seconda unità di misura quando è prevista dal questionario di rilevazione.
2. la nota che riguarda le unità rispondenti (che come già detto possono corrispondere o meno alle unità locali) che è unica ma può essere aggiornata nel corso dell'anno. Qui possono essere segnalati, ad esempio, i motivi della non collaborazione oppure dei periodi di manutenzione impianti o di cassa integrazione o ancora le informazioni riguardanti i dati inviati mensilmente (tipicamente errori nell'unità di misura fornita).
3. la nota che riguarda l'impresa e che viene utilizzata per mostrare le informazioni sulle variazioni anagrafiche. Il SIPI dispone, infatti, di una funzionalità che permette di definire e trattare gli eventi demografici ai fini di conservare il più possibile la continuità delle serie storiche elementari; la nota compilata in occasione della formalizzazione degli eventi viene mostrata nella maschera di inserimento M3.

In fase di macroediting finale, infine, i file excel utilizzati sia per i controlli effettuati all'elaborazione della prima stima dell'indice generale che alle successive revisioni consentono di annotare i controlli effettuati ed i relativi esiti.

Per quanto riguarda, infine, le attività di archiviazione occorre sottolineare che sebbene ogni cambiamento nei dati allo stato attuale venga registrato, non sono formalizzati tutti i flag necessari a codificare ogni aspetto delle attività svolte: *follow-up* e suo/suoi esiti nel caso occorra ritornare più di una volta sulla stessa unità rispondente, controllo di coerenza con le variabili ausiliarie fornite, con fonti esterne, con la serie storica disponibile e così via.

Inoltre il brevissimo intervallo temporale durante il quale vengono effettuati tutti i controlli dei dati a livello macro (mediamente due giorni con la possibilità di arrivare a tre in situazioni che necessitano particolari approfondimenti) limita la possibilità di produrre mensilmente un set di indicatori molto particolareggiato.

L'inserimento non automatico di un flag per ogni attività svolta rischia di rallentare il processo produttivo come già evidenziato. Occorrerà, quindi, effettuare delle analisi di tipo costi-benefici prima di pianificare l'inserimento di miglioramenti volti a produrre indicatori dettagliati sul processo di controllo e correzione.

Bibliografia

EUROSTAT. *Council Regulation N° 1165/98 Amended by the Regulation N° 1158/2005 of the European Parliament and of the Council* , 2005.

EUROSTAT. *Short-term Statistics Manual*, Eurostat, Luxembourg, 2000.

EUROSTAT. *Council Regulation N° 588/2001*, 2001.

EUROSTAT. *Council Regulation N° 586/2001*, 2001.

EUROSTAT (1998), *Council Regulation N° 1165/98*, 1998.

Gismondi R., CARONE A. "Statistical Criteria to Manage Non-respondents' Intensive Follow Up in Surveys Repeated along Time". Seminario: *Stima anticipata di indicatori congiunturali: teoria e applicazioni*, Istat, Roma, 24 Ottobre, 2007.

Gismondi R., Carone A., Iannaccone R. "L'individuazione delle unità statistiche "influenti" nell'indagine mensile sulla produzione industriale". Seminario: *La rilevazione mensile della produzione industriale: aggiornamento metodologico e disegno del nuovo sistema informativo*, Istat, Roma, 14 Marzo 2006.

Gismondi R., et altri. "La stima delle mancate risposte nell'indagine mensile sulla produzione industriale". Seminario: *La rilevazione mensile della produzione industriale: aggiornamento metodologico e disegno del nuovo sistema informativo*, Istat, Roma, 14 Marzo, 2006.

Gismondi R., Carone A., Ciammola A., Gambuti T., Iannaccone R., Mancini A., Moschetta M. "Non Response Treatment in the Italian Monthly Survey on Industrial Production". Meeting *Implementation of the Council Regulation No. 1165/98 on short term statistics*. Eurostat, Luxembourg, 15 Ottobre, 2004.

Gomez V., Maravall A. *Programs TRAMO and SEATS, Instructions for the User*. Bank of Spain. Working paper n.9628, Madrid, 1996.

Hidiroglou M.A., Berthelot J.M. "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology*, 12, 73-84, Statistics Canada, Ottawa, 1986.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Rapporto tecnico del progetto Europeo EDIMBUS, 2007. (http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/RPM_EDIMBUS.PDF)

ISTAT. "Le modificazioni longitudinali delle imprese: definizioni e trattamento statistico nel contesto della stima di un indice di variazione". Documento del *Gruppo di lavoro per l'aggiornamento metodologico ed il ridisegno informativo della rilevazione mensile sulla produzione industriale*, Istat, Roma, 2005.

ISTAT . *Manuali di tecniche d'indagine*, Istat, Roma, 1989.

Lucev D. *Tipologie e controllo dell'errore di non risposta per la qualità dei dati economici*, Rocco Curto Editore, Napoli, 1997.

Lundstrom S., Särrndal C.E. "Calibration as a Standard Method for Treatment of Non-response". *Journal of Official Statistics*, Vol.15, 2, 305-327, 1999.

Mancini A. "La rilevazione mensile della produzione industriale". Seminario: *La rilevazione mensile della produzione industriale: aggiornamento metodologico e disegno del nuovo sistema informativo*, Istat, Roma, 14 Marzo, 2006.

Mancini A. "La scelta del paniere dei prodotti per l'indagine mensile della produzione industriale: il legame con l'indagine annuale sulla produzione industriale". Documento del *Gruppo di lavoro per l'aggiornamento metodologico ed il ridisegno informativo della rilevazione mensile sulla produzione industriale*, Istat, Roma, 2005.

Mancini A., Bruno M., Genovese C. Moschetta M.G., Rocchetti M. "Report del sottogruppo sull'adeguamento delle classificazioni agli standard Istat utilizzati". Documento del *Gruppo di lavoro per l'aggiornamento metodologico ed il ridisegno informativo della rilevazione mensile sulla produzione industriale*, Istat, Roma, 2005bis.

Mancini A. "Proposta per aggiustamento per giorni lavorativi dell'indice della produzione industriale: metodo di calcolo basato sulla variabile endogena rilevata dai modelli ISTAT/SCO/PM". Documento di lavoro del *gruppo di lavoro destagionalizzazione*, Istat, Roma, 2002.

Quintano C., Castellano R., Romano A.A. "L'imputazione delle mancate risposte nelle indagini con parte panel: il caso dei redditi familiari della Banca d'Italia". *Quaderni di discussione*, 10, Istituto di Statistica e Matematica, Istituto Universitario Navale, Rocco Curto Editore, Napoli, 1996.

Rao J.N.K., Srinath K.P., Quenneville B. "Estimation of Level and Change Using Current Preliminary Data", In: Kasprzyk D., Duncan G., Kalton G., Singh MP. (*eds*), *Panel Surveys*, J.Wiley & Sons, New York, , 457-485, 1989.

Royall R.M. "Robustness and Optimal Design Under Prediction Models for Finite Populations". *Survey Methodology*, 18**,** 79-185, 1992.

Särrndal C.E., Swensson B., Wretman J. *Model Assisted Survey Sampling*, Springer Verlag, 1993.

Tam S.M. "Analysis of Repeated Surveys Using a Dynamic Linear Model". *International Statistical Review*, 55, 1, 63-73, International Statistical Institute, Great Britain, 1987.

APPENDICE – TABELLA DI CONVERSIONE DEI GAS

ii.GAS		m³ gas tecnici (*)	m³ gas normali (**)	kg
ARGON	m³ tecnici	1	0,9174	1,6364
	m³ normali	1,0900	1	1,7837
	m³ standard	1,0333	0,9479	1,6908
	litri	0,8511	0,7808	1,3928
	chilogrammi	0,6111	0,5606	1
AZOTO	m³ tecnici	1	0,9174	1,1473
	m³ normali	1,0900	1	1,2506
	m³ standard	1,0333	0,9479	1,1855
	litri	0,7048	0,6466	0,8086
	chilogrammi	0,8716	0,7996	1
ELIO	m³ tecnici	1	0,9174	0,1638
	m³ normali	1,0900	1	0,1785
	m³ standard	1,0333	0,9479	0,1692
	litri	0,7631	0,7003	0,1250
	chilogrammi	6,1050	5,6022	1
IDROGENO	m³ tecnici	1	0,9174	0,0825
	m³ normali	1,0900	1	0,0899
	m³ standard	1,0333	0,9479	0,0852
	litri	0,8602	0,7898	0,0710
	chilogrammi	12,1212	11,1235	1
OSSIGENO	m³ tecnici	1	0,9174	1,3108
	m³ normali	1,0900	1	1,4288
	m³ standard	1,0333	0,9479	1,3544
	litri	0,8705	0,7986	1,1410
	chilogrammi	0,7629	0,6999	1

(*) a 15°C e 98067 Pa (735.5 mm Hg); (**) a 0°C e 101325 Pa (760 mm Hg)

La rilevazione dei permessi di costruire: il controllo e la correzione dei dati

Silvana Garozzo, Istat, Servizio Statistiche congiunturali dell'industria e delle costruzioni
Giuliano Rallo, Istat, Servizio Statistiche congiunturali dell'industria e delle costruzioni

Sommario: Nella trattazione che segue si farà riferimento a due indagini mensili.
La prima è la Rilevazione statistica dei permessi di costruire, di tipo censuario e finalizzata alla produzione di statistiche strutturali sul settore dell'attività edilizia.
La seconda, introdotta nel 2003, è la Rilevazione statistica "rapida" dei permessi di costruire di tipo campionario e finalizzata alla produzione di indicatori trimestrali sull'edilizia entro 90 giorni dal periodo di riferimento, come richiesto dal regolamento comunitario n. 1165/98.
I due processi, che hanno come oggetto di rilevazione lo stesso fenomeno, procedono in parallelo con la differenza che una parte delle informazioni provenienti dalle unità campionarie viene sottoposta ad un trattamento più rapido, per consentire la produzione degli indicatori congiunturali nei tempi previsti.
Dopo aver descritto le caratteristiche delle due rilevazioni e aver definito i concetti di MRP e MRT, verrà esaminata la fase di controllo e correzione dati e le tecniche utilizzate, nell'ambito dell'una e dell'altra rilevazione, sia per il micro-editing sia per il macro-editing.

Parole chiave: permesso di costruire, fabbricato, censuaria, rapida, Mancate risposte parziali, Mancate risposte totali

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminare di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Le due rilevazioni

1.1. Caratteristiche della Rilevazione dei permessi di costruire (censuaria)

La rilevazione dei permessi di costruire è di tipo censuario a cadenza mensile. Attraverso la rete degli 8100 comuni italiani, vengono rilevate le principali caratteristiche dei nuovi fabbricati distinti in residenziali e non residenziali (anche se demoliti e interamente ricostruiti). Vengono, inoltre, rilevati gli ampliamenti dei fabbricati preesistenti, la cui costruzione sia stata autorizzata dal competente ufficio comunale. Le trasformazioni e le ristrutturazioni di fabbricati già esistenti, che non comportano variazioni di volumi degli stessi, non rientrano nel campo di rilevazione.
La realizzazione di un'opera edilizia nuova può essere autorizzata attraverso la presentazione del cosiddetto "Permesso di costruire" o della "Denuncia di inizio attività" (DIA), come previsto dalle leggi vigenti in materia.[138]
Lo strumento di rilevazione è costituito da un questionario cartaceo (modello Istat/AE), per il quale vi è l'obbligo di compilazione a cura del richiedente il titolo abilitativo.
Due o più opere, relative allo stesso titolo abilitativo, costituiscono due o più unità di rilevazione per le quali vengono, quindi, compilati altrettanti modelli di rilevazione.

Gli uffici comunali, mensilmente, hanno il compito di raccogliere tutti i questionari relativi alle opere edilizie per le quali siano stati ritirati i "Permessi di costruire", o decorra la validità delle "Denunce di inizio attività", controllare l'esattezza delle informazioni che vi sono riportate, compilarne il riquadro riservato al Comune ed inviarli, tramite posta, entro il quinto giorno del mese successivo. Quindi, il mese di riferimento della rilevazione coincide con il mese in cui avviene il ritiro del generico permesso di costruire o decorre la validità della denuncia di inizio attività. In caso di assenza di permessi di costruire ritirati e di DIA in corso di validità, il Comune deve inviare una segnalazione di attività edilizia nulla, attraverso la quale si rileva l'assenza del fenomeno nel mese di riferimento. Se il Comune non invia alcun questionario o segnalazione di attività edilizia nulla, relativi al mese di riferimento, è considerato non rispondente.

Il modello di rilevazione, oltre al riquadro contenente i dati identificativi del comune, è composto da tre parti: la prima raccoglie le notizie generali sull'opera (tempi previsti per la realizzazione, ubicazione, natura dell'opera, destinazione d'uso, concessionario, finanziamento, regime dei suoli, impianto termico, struttura portante); la seconda relativa ai soli fabbricati residenziali, contiene informazioni sui piani, sul volume, sulla superficie, sul numero di abitazioni e la ripartizione delle abitazioni secondo il numero di stanze per abitazione e secondo le classi di superficie utile abitabile; la terza comprende le notizie relative ai soli fabbricati non residenziali e indaga sulla dimensione del fabbricato, sulla parte ad uso abitativo, sulla destinazione economica e sulla tipologia dell'opera. La prima parte del questionario va sempre compilata. Mentre, a seconda che si tratti di fabbricato residenziale o non residenziale, verrà compilata rispettivamente la seconda o la terza parte.

1.2 Caratteristiche della Rilevazione "rapida" dei permessi di costruire

Il Regolamento del Consiglio Europeo sulle statistiche congiunturali (CE n. 1165/98) impone agli Stati Membri l'obbligo di produrre e trasmettere alla Commissione alcuni indicatori congiunturali sull'attività edilizia entro 90 gg dal periodo di riferimento (cfr. Tavola 1).
Pur raccogliendo le informazioni necessarie alla produzione degli indicatori richiesti, la rilevazione totalitaria dei permessi di costruire non consente di rispettare le scadenze comunitarie. Si è resa, pertanto, necessaria l'introduzione di alcune innovazioni per ridurre i tempi di produzione di tali indicatori.

[138] D.P.R. 6 giugno 2001 n. 380 – Testo unico delle disposizioni legislative e regolamenti in materia edilizia – G.U. n. 245 del 20 ottobre 2001 – s.o. n. 239

Le innovazioni sono consistite nella creazione della Rilevazione statistica "rapida" dei permessi di costruire e, nell'ambito della stessa, nella costruzione di un disegno di campionamento per la selezione di un campione rappresentativo di comuni e nella messa a punto di nuove metodologie per la stima degli indicatori congiunturali.

Tavola 1: *Indicatori richiesti dal Regolamento STS*

Unitá di misura	Variabile	Descrizione	Disaggre-gazioni	Descrizione
Abitazioni (numero)	4610	Numero Totale Abitazioni in nuovi fabbricati residenziali	4611	Abitazioni in nuovi fabbricati residenziali con 1 abitazione
			4612	Abitazioni in nuovi fabbricati residenziali con 2 abitazioni ed oltre
Superfici (mq)	4610	Superficie dei fabbricati residenziali = Superficie Utile Abitabile delle abitazioni in nuovi fabbricati residenziali + Superficie totale dei fabbricati destinati a collettività	4610	Superficie Utile Abitabile delle abitazioni in nuovi fabbricati residenziali con 1 abitazione
			4612	Superficie Utile Abitabile delle abitazioni in nuovi fabbricati residenziali con 2 abitazioni ed oltre
			4613	Superficie totale dei Fabbricati per collettività
	4620	Superficie totale Fabbricati non residenziali	4621	Superficie totale Fabbricati per Uffici
			4629	Superficie totale Altri Fabbricati non residenziali

La rilevazione rapida è, in effetti, un sottoprocesso della tradizionale Rilevazione dei permessi di costruire che rileva mensilmente, su un campione ristretto di comuni, soltanto le informazioni necessarie al calcolo degli indicatori congiunturali richiesti dal regolamento, utilizzando gli stessi questionari che, successivamente, vengono lavorati nel processo della rilevazione tradizionale, insieme a quelli pervenuti da tutti gli altri comuni non campionari.
Il campo di rilevazione dell'indagine rapida è costituito, quindi, da un campione di 814 comuni, di cui 160 comuni capoluogo e non capoluogo con più di 50.000 abitanti che costituiscono uno strato autorappresentativo, mentre i restanti 654 sono stratificati per ripartizione geografica e classi di popolazione in 20 gruppi (cfr. Tavola 2).
Dalla prima parte del questionario di rilevazione, vengono acquisite soltanto due variabili: la natura dell'opera (nuovo o ampliamento), e la destinazione d'uso (residenziale o non residenziale). Vengono, inoltre, acquisiti il numero dei piani, il volume, e la superficie totale. Ed, infine, il numero di abitazioni e la superficie utile abitabile, se si tratta di opera residenziale, la destinazione economica e la tipologia dell'opera, se si tratta di opera non residenziale.

Tavola 2: *Numero di comuni campione per ripartizione e classi demografiche**

Classi di popolazione	Centro	Nord-Est	Nord-Ovest	Sud-Isole	Totale
Fino a 3000	15	28	79	29	**151**
Da 3001 a 7000	20	45	57	42	**164**
Da 7001 a 13000	22	54	37	38	**151**
Da 13001 a 25000	14	28	32	32	**106**
Da 25000 a 50000	20	13	21	28	**82**
Totale	**91**	**168**	**226**	**169**	**654**

* *esclusi i 160 comuni autorappresentativi*

1.3. La raccolta dei dati

I questionari autocompilati vengono raccolti, mensilmente, sia presso i comuni campione della rilevazione rapida sia presso i comuni non campione. Ma, mentre i primi li inviano direttamente all'Istat di Roma, tutti i restanti comuni li inviano alle relative camere di commercio. Queste, in qualità di enti intermedi di rilevazione, raccolgono tutti i modelli pervenuti nel mese dai comuni non campione della provincia di propria competenza, e li inviano, a loro volta, all'Istat di Roma.
Questa duplice procedura permette, da un lato di velocizzare la raccolta dei questionari dei comuni campione e dall'altro di avvalersi della preziosa collaborazione degli enti provinciali per la raccolta e la gestione dei modelli dei restanti numerosi comuni non campione.
Anche se il questionario è lo stesso, come si è detto, il numero di variabili rilevate è differente tra le due rilevazioni. Infatti, per quanto riguarda la censuaria vengono raccolte tutte le variabili contenute nel modello di rilevazione, mentre per l'indagine rapida si rilevano soltanto una minima parte delle informazioni.
Nella tavola 3 è riportato il numero di variabili rilevate dalle due rilevazioni.

Tavola 3: *Numero delle variabili rilevate*

	Rilevazione	
Fabbricato	Censuaria	Rapida
Residenziale	56	12
Non residenziale	47	12

2. Controllo e correzione dei dati (errori non campionari)

2.1. Definizione di MRP e MRT

Prima di descrivere il processo di controllo e correzione per le due rilevazioni, è opportuno definire i concetti di mancata risposta parziale (MRP) e mancata risposta totale (MRT).
A tal proposito, occorre precisare che la fase di controllo e correzione dati prevede interventi su due diversi livelli: prima sui microdati poi sui dati aggregati per comune.
Pertanto i concetti di mancata risposta vengono definiti a seconda che si riferiscano alle correzioni di primo o di secondo livello.

2.1.1. Correzione di primo livello (sul singolo questionario/record)

Come si è detto, l'unità di rilevazione è costituita dalla singola opera edilizia da realizzare, relativa a un nuovo fabbricato o all'ampliamento di un fabbricato preesistente, per la quale viene compilato il questionario che riporta le principali caratteristiche dell'opera stessa.
Quindi, a livello di microdato, si definisce MRP il caso in cui un questionario viene compilato solo in parte. Tali MRP vengono opportunamente localizzate e corrette per riportare il questionario in una situazione di completezza.
Mentre, si definisce MRT il caso di mancata compilazione del questionario, che non è altrettanto facile da individuare.
Bisogna ricordare, infatti, che la compilazione è a cura del singolo cittadino, in quanto titolare del titolo che abilita alla realizzazione dell'opera, mentre la raccolta e l'invio dei questionari è a cura del competente ufficio comunale.
Ma, poiché non tutte le amministrazioni comunali collaborano effettivamente alla rilevazione, non è possibile risalire a tutti quei casi, oggetto di rilevazione, per i quali non è pervenuto il questionario.

2.1.2. Correzione di secondo livello (sui dati comunali)

Sulla base della mancata collaborazione dei comuni si identificano, nella seconda fase di controllo e correzione, le MRP e le MRT. Infatti, la MR non si riferisce più alla mancata compilazione del questionario, relativo ad una singola unità di rilevazione, bensì alla non risposta o non collaborazione da parte del Comune che non invia alcuna comunicazione nel mese di riferimento. Pertanto, con riferimento ad un intervallo temporale di un anno, si definisce MRP il caso in cui un comune ha risposto almeno un mese su 12, mentre si ha MRT quando un comune è totalmente assente nei 12 mesi considerati.
Nella fase di controllo e correzione dei macrodati aggregati per comune, si procede alla localizzazione dei comuni non rispondenti per mese e alla loro imputazione.
In generale, la fase di controllo e correzione avviene secondo due diverse procedure, una relativa alla rilevazione dei permessi di costruire censuaria e una alla rilevazione rapida campionaria, sebbene i presupposti di base siano comuni ad entrambe.

2.2. Rilevazione sui permessi di costruire – censuaria

2.2.1. Microediting di tipo misto interattivo e automatico (primo livello)

In questa fase vengono localizzati e corretti i valori fuori dominio e le risposte parziali, le incompatibilità e i valori anomali. I questionari cartacei vengono integralmente revisionati manualmente, secondo specifiche norme. Alcuni criteri di revisione manuale sono deterministici, mentre altri possono avere un margine di soggettivà da parte del revisore, in quanto si potrebbero presentare più soluzioni possibili e la scelta dipende da una valutazione complessiva delle informazioni presenti sul questionario. Sono, quindi, previsti una serie di editing relativi a controlli di incompatibilità o di consistenza intra-record, controlli di range di accettazione e controlli statistici per la localizzazione di valori anomali, attraverso intervalli di accettazione determinati da algoritmi, sulla base di distribuzioni di rapporti caratteristici.

Revisione manuale
Per ogni questionario, che può essere relativo ad un nuovo fabbricato o ad un ampliamento di un fabbricato preesistente, viene controllato sempre il riquadro con i dati identificativi del comune e della provincia, per garantire una giusta successiva registrazione dei codici. Viene poi verificata la coerenza tra quanto indicato in "natura dell'opera" e "destinazione d'uso" e le successive compilazioni della parte

seconda o parte terza (residenziale o non residenziale). Viene poi messa a confronto la sequenza temporale delle date di richiesta e di ritiro del Permesso di costruire, per verificarne la coerenza.
Questi primi controlli sono di fondamentale importanza per la corretta interpretazione di tutte le notizie successive riguardanti il fabbricato in quanto collocano territorialmente e temporalmente le notizie successive del questionario.
Il successivo controllo è di completezza; viene verificata la compilazione di tutti i quesiti relativi al tipo di opera. Nel caso di MRP si procede all'imputazione dei valori mancanti sulla base delle altre informazioni presenti sul questionario.
Per quanto riguarda, invece, la verifica delle informazioni presenti sul modello si effettuano sia controlli di range di accettazione sui valori, sia controlli di incompatibilità tra valori sulla base di relazioni tra variabili.
Con riferimento a quest'ultimi, ad esempio, in caso di nuovo fabbricato residenziale, è necessario che sia stata indicata la superficie totale in mq del fabbricato e che il suo valore sia compatibile con il corrispondente volume in metricubi, e cioè che il rapporto tra il volume e la superficie totale sia compreso tra 1,5 e 6. Se il rapporto non rientra in tale intervallo, in generale non viene modificato il volume mentre viene corretta la superficie mettendola in relazione con altre variabili.
Per quanto riguarda l'individuazione di eventuali valori anomali, si può fare il seguente esempio, in cui il dato anomalo è identificato mettendo in relazione la superficie utile abitabile con il numero di abitazioni, in base alla regola che il rapporto tra la superficie e le abitazioni deve essere compreso tra 15 e 400. Se tale rapporto è minore di 15, occorre modificare opportunamente la superficie; se è maggiore di 400, siamo in presenza di un valore anomalo da accertare o attraverso un'analisi complessiva del questionario o contattando il comune per chiedere dei riscontri con i progetti originali.

Registrazione e correzioni interattive
La tavola 4 contiene la numerosità per anno dei questionari (corrispondenti ai record) registrati, successivamente sottoposti a controllo presso l'UO ed archiviati nel database dell'indagine censuaria. Il numero si attesta intorno ad un livello pari a circa 120.000 unità. La quota dei record relativi ai fabbricati residenziali costituisce la quota prevalente, in quanto costituita da poco più del 50% del totale. I record di tipo non residenziale, invece, costituiscono appena il 20% del totale, mentre quelli relativi alle segnalazioni di attività negativa del comune nel mese pesano quasi il 30%.

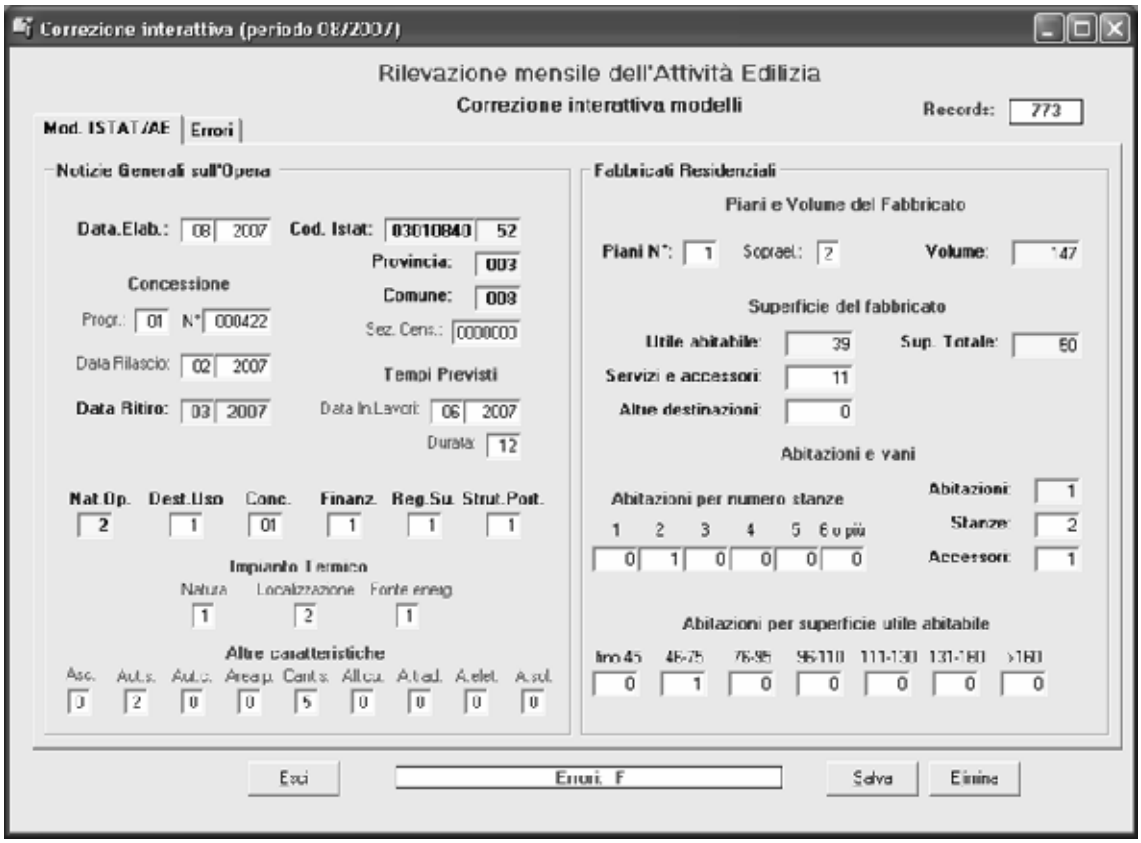
Tavola 4: *Numero di record registrati per anno*

	Rilevazione censuaria					
Tipo record	**2003**	**2004**	**2005**	**2003**	**2004**	**2005**
Residenziale	62.815	65.651	64.825	51%	54%	55%
Non Residenziale	24.601	23.580	21.768	20%	20%	18%
Negativi	35.008	31.581	31.918	29%	26%	27%
Totale	**122.424**	**120.812**	**118.511**	**100%**	**100%**	**100%**

Detto materiale viene inviato mensilmente, dopo i controlli illustrati nel precedente paragrafo, alla fase di registrazione che avviene in service esterno e produce il file dei microdati. Questo file viene sottoposto, in primo luogo, ad una procedura di correzione automatica che avviene con approccio deterministico per correggere alcuni tipi errori sfuggiti alla revisione manuale oppure dovuti alla registrazione. Il software delle correzioni produce un tabulato che riporta per comune e per singolo record gli errori e gli accertamenti. I primi dovuti alla non completa revisione del questionario o ad errori di digitazione, i secondi relativi alla localizzazione di valori ritenuti anomali.

La correzioni interattive successive richiedono il frequente ritorno sui questionari contenenti gli errori da correggere e l'accertamento da fare sui valori segnalati come anomali. L'interfaccia grafica del programma che viene utilizzato per le correzioni interattive è riportato in Figura 1. Esistono tre tipologie di schermate in corrispondenza dei tre tipi record (residenziale, non residenziale, negativo).

Figura 1: *Maschera per le correzioni interattive dei record*



In alto a destra della maschera è riportato il numero di record con uno o più errori. Per ciascun record tutti i campi compaiono sulla singola schermata. Nella parte inferiore della maschera sono evidenziati i tipi di errore indicati in lettere, la cui decodifica è possibile tramite l'utilizzo della bandiera "errori". Questi derivano dal piano generale del check e sono riportati in una lista che ha le caratteristiche del fac-simile riportato nella Tavola 5.

Tavola 5: *Esempio di alcune descrizioni di errori, estratte dalla lista generale*

Codice errore	Descrizione
N	Il numero piani residenziale è uguale a zero (nuovo fabbricato)
M	Il rapporto tra Volume / (Sup.Utile+Sup.serv.accessori) non è compreso fra 1.5 e 5
M	Il rapporto tra Volume / (Sup.Utile+Sup.serv.accessori+Sup altre dest) non è compreso fra 1.5 e 6
N	Per 1 abitazione è ammesso un massimo di 6 piani
N	Per 2 abitazioni è ammesso un massimo di 8 piani
N	Per più di 2 abitazioni il rapporto tra piani ed abitazioni deve essere inferiore a 4

Tutti i record corretti e convalidati vengono infine immessi sul database generale dei permessi di costruire, che viene alimentato mensilmente e che costituisce la base informativa dei dati rilevati, generalmente parziali a causa di mancata collaborazione di alcuni comuni.

2.2.2. Validazione dei risultati delle elaborazioni

Prima di procedere all'imputazione della MRT ai fini della pubblicazione annuale dei risultati strutturali, si effettuano dei confronti con i dati degli anni precedenti sul livello di copertura territoriale raggiunto, sui totali delle principali variabili rilevate e sui valori medi caratteristici per verificare che non ci siano valori anomali che possano provocare distorsioni nelle stime finali. Nel caso in cui si riscontri la presenza di valori anomali, si approfondisce la ricerca restringendo l'analisi a livello territoriale e poi si ritorna sui microdati al fine di stabilire se si tratta di errore e, quindi, procedere alla correzione, o se si è in presenza di risultati da imputarsi alla variabilità del fenomeno rilevato.

2.2.3. Macroediting sui dati comunali (secondo livello)

In questa fase vengono imputate le mancate risposte dei comuni, o di secondo livello, in base alla definizione data in precedenza.
L'imputazione delle MR dovute alla mancata collaborazione dei comuni, nella rilevazione censuaria dei permessi di costruire, costituisce una importante innovazione rispetto ad alcuni anni fa, quando la diffusione dei dati era il risultato dell'aggregazione dei record relativi soltanto ai Comuni rispondenti alla rilevazione (ISTAT, 2005).
Per l'imputazione delle MR, che avviene su dati mensili relativi ad un periodo di un anno, si distinguono due insiemi di comuni: il primo comprende tutti i comuni che siano capoluogo di provincia o che abbiano più di 50.000 abitanti (160 comuni); il secondo comprende i restanti 7940 comuni non capoluogo e con popolazione non superiore a 50.000.
Dalla tavola 6 si evince il diverso comportamento in termini di collaborazione tra i due gruppi di comuni. Infatti, ben il 72% dei comuni appartenenti al primo insieme rispondono 12 mesi e soltanto il 3% non ha mai collaborato nell'anno considerato. Mentre per il secondo insieme si ha una minore percentuale di comuni che hanno risposto tutti e 12 mesi (40%) e un tasso di non risposta totale (zero mesi nell'anno) molto più elevato, pari al 25%.

Tavola 6: *Percentuale di comuni per numero mesi di risposta - anno 2005*

Mesi di risposta	Comuni capoluogo e comuni con più di 50.000 abitanti (160 unità)	Comuni con popolazione non superiore a 50.000 abitanti (7940 unità)	Totale comuni (8100 unità)
0	3,1	25,0	24,6
1	0,6	2,2	2,2
2	0,6	1,3	1,3
3	0,0	0,9	0,9
4	0,6	0,9	0,9
5	0,0	1,0	1,0
6	1,9	1,2	1,2
7	0,0	1,5	1,5
8	1,3	2,0	2,0
9	3,1	3,5	3,5
10	3,8	6,6	6,5
11	13,1	14,4	14,4
12	71,9	39,5	40,1

Totale	**100**	**100**	**100**

**a tutto periodo (18 mesi)*

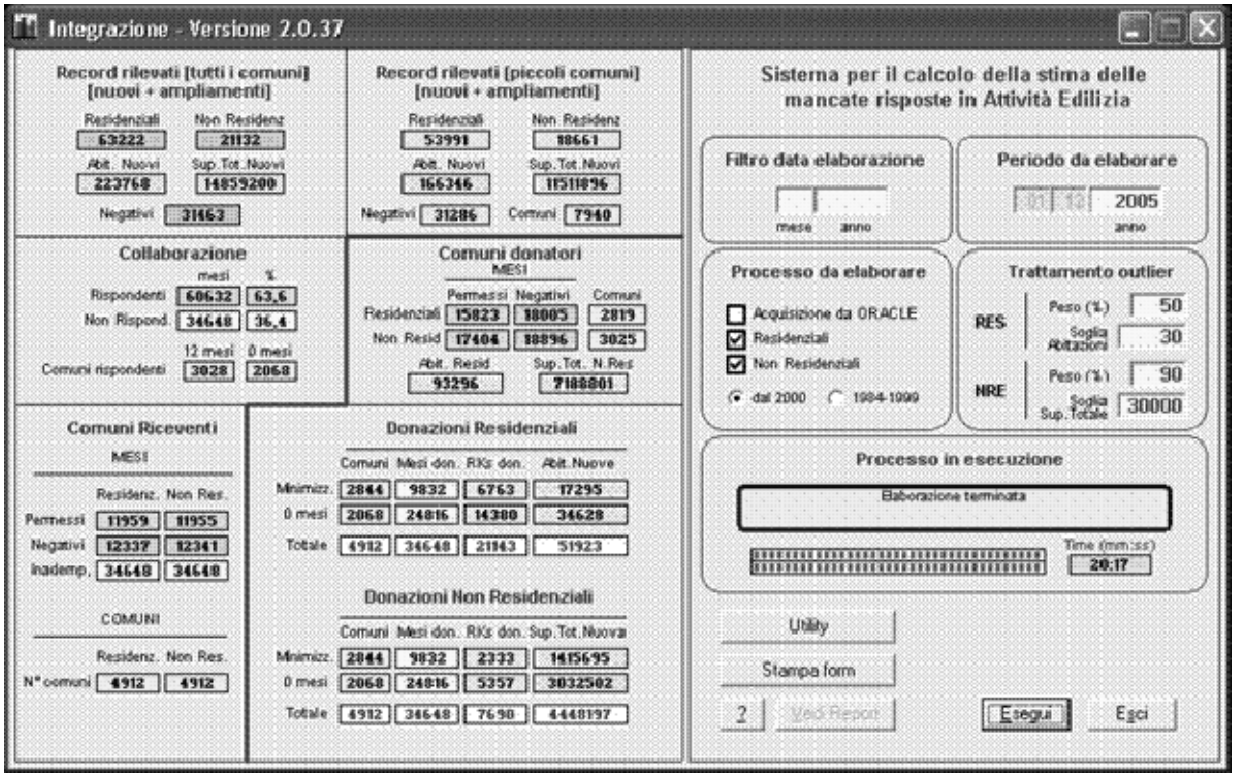
Il metodo utilizzato per i comuni del primo insieme, costituito da 160 comuni, tiene conto dell'importanza che essi assumono in termini di peso nella rilevazione e della loro collaborazione complessiva. L'integrazione dei dati mensili si basa su un'analisi delle informazioni elementari dei comuni, rispondenti nei 12 mesi considerati, che conduce a individuare l'insieme di record da utilizzare per l'imputazione delle mancate risposte.
In sintesi, in caso di MRP, i dati mensili mancanti sono imputati sulla base di quelli forniti dal medesimo comune per i mesi contigui o, in caso di ulteriori mancate risposte, nel medesimo mese di anni contigui. Nei casi di MRT, cioè assenza totale di risposta in tutti i mesi dell'anno, l'imputazione avviene tramite donatore, ovvero attribuendo al comune non rispondente i record relativi ad un altro comune con caratteristiche simili (per dimensione demografica, regione di appartenenza e zona altimetrica) scelto come donatore.

Per l'imputazione delle MR nell'insieme dei restanti 7940 comuni si utilizza il metodo del donatore scelto mediante una funzione che minimizza la distanza, basata su variabili territoriali e demografiche (Bacchini, Iannaccone e Otranto, 2005).
In generale, in caso di MRP, e cioè per i comuni rispondenti in almeno in uno dei mesi del generico anno, all'interno di ciascuno strato definito dalle variabili ausiliare (ripartizione geografica e popolazione), il donatore viene individuato minimizzando, per ciascun comune *j*, la seguente funzione di distanza:

$$\min_{1\le k\le r_h} \sum_{m\in M} | x_k^m - x_j^m |$$

dove M indica l'insieme dei mesi in cui l'unità *j* ha risposto nel corso dell'anno e r_h il numero dei rispondenti 12 mesi nello strato *h* cui appartiene il comune *j*.
Nel caso di MRT, in cui il comune non abbia risposto in nessuno dei 12 mesi dell'anno, la selezione del donatore avviene estraendo casualmente un comune dall'insieme dei comuni rispondenti 12 mesi nello strato.
In entrambe le situazioni il donatore individuato viene utilizzato per imputare congiuntamente tutti i mesi mancanti al fine di preservare il profilo temporale del fenomeno.
Il pannello che segue (Figura 2) è l'interfaccia grafica del programma che opera l'imputazione dei comuni (i 7940 inferiori a 50.000 abitanti e non capoluogo) non rispondenti nel singolo mese.

Figura 2: *Maschera per l'imputazione delle MR sui comuni non rispondenti*



In questo, riferito all'anno 2005, nel riquadro "Collaborazione" sono riportati i mesi di risposta per comune: pari al 63,6%; e quelli di mancata risposta: 36,4%. Nello stesso riquadro è riportato l'importante numero dei comuni *sempre* rispondenti (3.028 comuni), che costituisce il serbatoio dei potenziali donatori. Nel riquadro "Comuni donatori" compaiono le numerosità distinte per residenziale (2819 comuni) e non residenziale (3025), derivanti dai filtri imposti per trattare gli "outlier".
In basso a sinistra della maschera compaiono, infine, due zone destinate all'informazione sulla donazione con l'evidenziazione di alcune caratteristiche del "ricevuto" in termini di tipo record e del "donato".

2.3. Rilevazione rapida dei permessi di costruire - campionaria

Come si è detto, gli 814 comuni campione inviano mensilmente i questionari compilati direttamente all'Istat di Roma, presso un'apposita casella postale dedicata alla rilevazione rapida, in modo da ridurre i tempi di raccolta.

2.3.1. Microediting sui questionari/record (primo livello)

I questionari, pervenuti dagli 814 comuni campione, vengono sottoposti ad una lavorazione rapida che prevede la revisione manuale e la registrazione, all'interno dell'unità operativa, delle sole variabili utili al calcolo degli indicatori congiunturali.
Ciascun questionario viene, quindi, sottoposto ad una parziale revisione manuale, relativa soltanto alla parte contenente le informazioni da digitare.
I criteri di revisione sono gli stessi di quelli utilizzati nella rilevazione censuaria e quindi, anche in questo caso, vengono corretti gli errori riguardanti le mancate risposte parziali, i valori fuori dominio, le incompatibilità e i valori anomali.

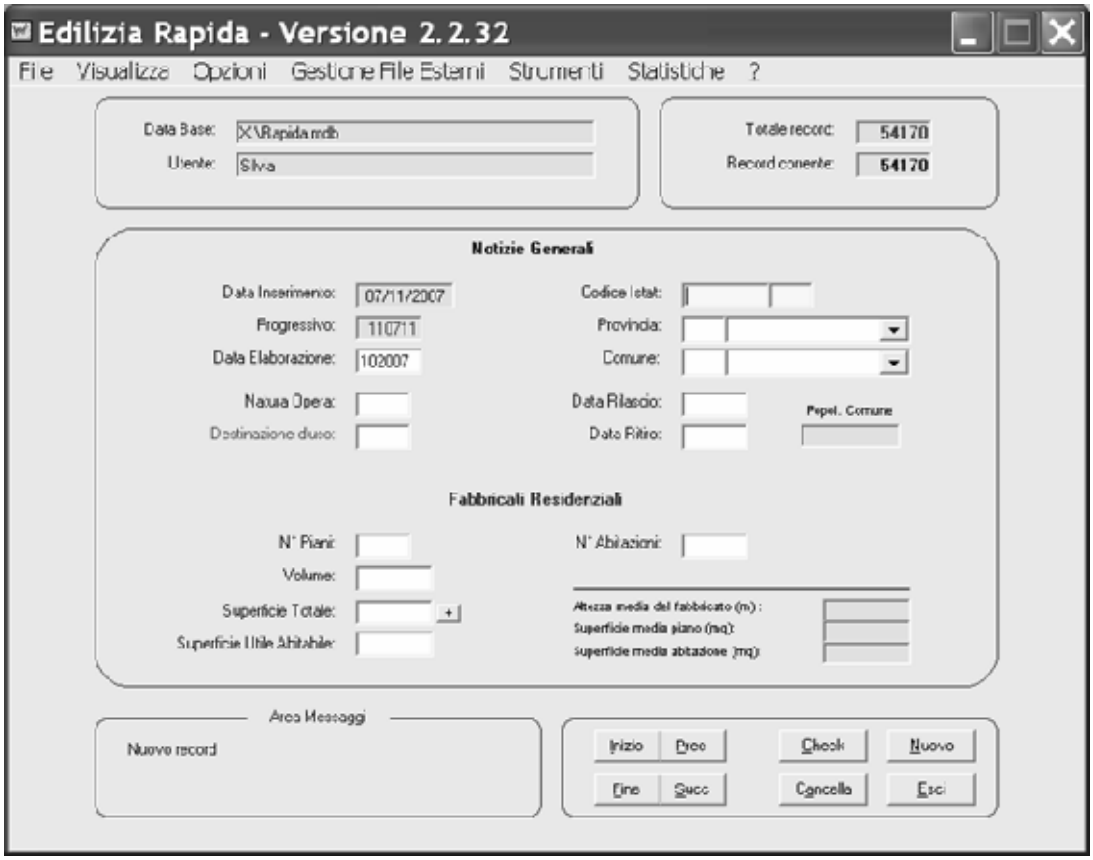
La tavola 7 contiene il numero di record in assoluto e in percentuale sottoposti a parziale controllo nell'ambito dell'indagine rapida. In totale i questionari o record relativi ai soli comuni campione sono circa 27.000 all'anno, che costituiscono il 22% circa rispetto al totale comuni.
In percentuale, a differenza di quanto visto per l'indagine censuaria, si ha una netta predominanza di record relativi a fabbricati residenziali, con oltre il 70% di questionari sul totale. Per contro la percentuale di record negativi si presenta, in questo caso, piuttosto esigua aggirandosi intorno all'8%. Ciò si spiega per il fatto che, in questo caso, è maggiore il peso dei 160 comuni capoluogo o con più di 50.000 abitanti, per i quali si può ipotizzare una concentrazione più alta del fenomeno residenziale e nello stesso tempo una scarsa propensione all'assenza del fenomeno. Infine, i record relativi a fabbricati non residenziali rappresentano, come per l'indagine censuaria, il 20% circa dei questionari lavorati.

Tavola 7: *Numero di record registrati per anno*

	Rilevazione Rapida					
Tipo record	**2003**	**2004**	**2005**	**2003**	**2004**	**2005**
Residenziale	18.388	20.188	19.079	69%	74%	72%
Non Residenziale	6.295	5.196	5.382	23%	19%	20%
Negativi	2.161	1.935	2.072	8%	7%	8%
Totale	**26.844**	**27.319**	**26.533**	**100%**	**100%**	**100%**

La registrazione avviene mediante un software ad hoc che prevede l'immissione autocontrollata dei dati di ciascun record con controlli, in minima parte di tipo automatico con approccio deterministico e in parte di tipo interattivo con l'intervento dell'operatore. I controlli automatici prevedono l'immissione, per ciascun record, di alcune informazioni come il nome dell'operatore, la data di inserimento e di un numero progressivo. I controlli di tipo interattivo sono contestuali all'immissione dei dati e finalizzati alla correzione degli errori non risolti in fase di revisione e degli errori di digitazione.
Dopo questa serie di operazioni il record viene, quindi, validato e memorizzato nell'archivio dell'indagine rapida.
La Maschera che segue è quella utilizzata per la digitazione autocontrollata di ogni singolo record relativo a un fabbricato di tipo residenziale. Analogamente si ha un pannello per la registrazione dei record di tipo non residenziale ed inoltre uno per la registrazione dei record negativi
In alto a sinistra appare il nome del database in cui vengono archiviati, una volta validati, tutti record residenziali e non residenziali e il nome dell'utente che immette i dati. In alto a destra viene indicato il numero di record totale contenuti in archivio e il corrispondente numero record oggetto di nuova immissione o di modifica. Mentre la parte centrale della maschera contiene le informazioni di carattere generale e le variabili specifiche che vengono immesse per ciascun record di tipo residenziale. Infine, in basso si hanno i pulsanti di comando che consentono di operare l'immissione, la modifica o la validazione di un record.

Figura 3: *Maschera per la registrazione controllata dei questionari dei comuni campione*



2.3.2. Macroediting sui dati comunali (secondo livello)

Per quanto riguarda la correzione delle MR dovute alla mancata collaborazione dei comuni, anche qui si distingue l'insieme dei 160 comuni autorappresentativi (quelli capoluogo e quelli con più di 50.000 abitanti) da quello costituito dai restanti comuni, come nel caso della rilevazione censuaria, ma a differenza di quest'ultima, non si imputano i microdati, bensì i dati aggregati per comune.
Tale diversità è dovuta al fatto che la rilevazione censuaria permette di produrre numerose tabelle che prevedono incroci non solo tra variabili di tipo quantitativo ma anche di tipo qualitativo, la cui aggregazione non renderebbe possibile una serie di analisi strutturali.
Invece, nel caso della rilevazione rapida, l'obiettivo specifico è quello di stimare gli indicatori richiesti dal regolamento STS, che riguardano esclusivamente variabili di tipo quantitativo, e il cui obiettivo è principalmente quello di descrivere la dinamica del fenomeno.
Come si dirà nel seguito, sono in corso delle sperimentazioni alternative ai metodi attualmente in uso per l'imputazione delle MR per verificare la possibilità di ottenere dei miglioramenti della qualità delle stime prodotte, sia per i comuni autorappresentativi sia per i restanti comuni.

Per quanto riguarda il sottoinsieme dei 160 comuni autorappresentativi, l'imputazione mensile in caso di MRP, sui dati aggregati a livello comunale, avviene attraverso il valore medio che le variabili (abitazioni e superfici) assumono nello stesso comune nei 12 mesi precedenti in cui ha risposto, compreso il mese che si sta elaborando.
Una sperimentazione condotta per valutare la capacità di tale tecnica d'imputazione ha portato alla conclusione che il metodo usato è migliore rispetto ad altri metodi alternativi.
In particolare, all'interno dell'Unità operativa "Metodi per il trattamento degli errori non campionari" (MTS/G), sono state effettuate delle analisi basate su simulazioni tipo Monte Carlo di MRP e

sull'applicazione comparativa del metodo attuale della media longitudinale e di diverse tecniche da donatore di distanza minima longitudinale per strati.
Il metodo del donatore mostra una buona capacità di ricostruzione del valore totale mensile delle variabili da stimare, ma il metodo attuale, sfruttando in modo diretto l'informazione longitudinale della stessa unità oggetto d'imputazione, risulta leggermente preferibile, in parte anche a causa dell'esiguità del numero di unità (160 comuni) e dalla loro eterogeneità, che non costituiscono dei buoni presupposti per l'imputazione tramite donatore (Di Zio, Guarnera e Luzi , 2007).

In caso di MRT, quando cioè il comune non ha mai risposto nei 12 mesi precedenti, non viene imputato nessun valore, con una conseguente leggera sottostima del fenomeno.
Questo problema è stato di recente oggetto di ulteriori sperimentazioni, svolte nell'ambito dell'Unità operativa "Metodi per il trattamento degli errori non campionari" (MTS/G), per trovare una soluzione ed adottare un opportuno metodo d'imputazione anche per le MRT.
Tenuto conto degli obiettivi dell'indagine e che il fenomeno delle MRT nell'insieme dei comuni autorappresentativi è piuttosto esiguo, sono stati messi a confronto due diversi metodi, uno basato sulla media trasversale per classi e uno basato sul donatore trasversale di minima distanza, sempre per classi.
La sperimentazione ha mostrato che i due metodi hanno prestazioni pressoché analoghe. Il metodo della media sembra da preferire in relazione alle finalità dell'indagine (stima di totali), alla semplicità dell'approccio e all'omogeneità di trattamento rispetto alle MRP. Mentre il metodo del donatore è da preferire per omogeneità di trattamento delle MRT rispetto al metodo usato nella rilevazione censuaria (Di Zio, 2007).
In ogni caso, data la sostanziale equivalenza delle prestazioni, si valuterà all'interno della struttura responsabile della rilevazione l'opportunità di scegliere l'uno o l'altro dei due metodi, anche in base a considerazioni di tipo operativo.

Per quanto riguarda le MR, nell'insieme dei restanti 654 comuni campione stratificati si applica una diversa metodologia che, anziché utilizzare l'imputazione, si basa sulla riponderazione delle osservazioni rilevate sui comuni rispondenti in modo da rappresentare i restanti comuni.
Inoltre, tenendo conto del fatto che, al momento del calcolo delle stime, la copertura del campione non è completa, mentre sono disponibili informazioni su alcune unità non incluse nel campione, si è ritenuto opportuno sfruttare non solo le informazioni pervenute dai 654 comuni campione ma anche quelle pervenute, entro i termini utili per l'invio delle variabili ad Eurostat, dagli altri 7286 comuni non campionari.
A tale scopo, quindi, è stata messa a punto una metodologia d'imputazione che, a partire dai dati aggregati per comune, stima per mese il numero totale delle abitazioni in nuovi fabbricati residenziali e la superficie totale non residenziale per i 7940 comuni (di cui 654 campionari e 7286 non campionari), attribuendo un peso al valore assunto dalle variabili d'interesse nei comuni rispondenti nel periodo considerato.
Lo stimatore attualmente utilizzato, denominato stimatore FABI, è il seguente:

$$\widetilde{Y}_t = \sum_{k \in r_t} y_{t,k}\, a_{t,k}$$

Dove $y_{t,k}$ è la variabile d'interesse al tempo t nella k-esima unità e $a_{t,k}$ è il peso attribuito alla stessa unità determinato in base ai parametri stimati da una regressione logistica (vedi Falorsi, Alleva, Bacchini, Iannaccone, 2005).
Tale stimatore ha spesso evidenziato una sistematica sovrastima degli indicatori calcolati a 90 giorni dal trimestre di riferimento rispetto a quelli rivisti a distanza di tre mesi.
Pertanto, nell'ottica di migliorare la qualità delle stime preliminari a 90 giorni, si è introdotto un altro stimatore che attualmente è in sperimentazione.
Il nuovo stimatore, di tipo Shrinkage, anziché combinare insieme le informazioni campionarie e non campionarie, calcola separatamente le due stime indipendenti, una basata sui comuni campione e l'altra

sui comuni non campione, combinandole in modo ottimale al fine di rendere minima la varianza complessiva (Bacchini, Falorsi e Iannaccone, 2005).
Dopo avere stimato il numero delle abitazioni in nuovi fabbricati, si moltiplica tale valore per la superficie utile abitabile media, riscontrata sui dati rilevati, ottenendo così il corrispondente valore della Superficie Utile abitabile.

Dalla somma dei valori stimati precedentemente per il gruppo dei 160 comuni autorappresentativi e per quello dei restanti 7940 comuni, si ottiene il numero totale delle abitazioni nei nuovi fabbricati residenziali, la relativa superficie utile abitabile e la superficie totale non residenziale (compresa la superficie per collettività) per l'universo degli 8100 comuni italiani.

Nella tavola 8 vengono riportati i tassi di risposta per mese dei vari gruppi di comuni che entrano a far parte della metodologia di calcolo delle stime.
Tali tassi si riferiscono al momento in cui avviene la revisione delle stime congiunturali che diventano, quindi, definitive e che vengono elaborate a 180 dal periodo di riferimento.
I tassi più elevati si registrano per lo strato dei 160 comuni autorappresentativi, per i quali in media si ha ben oltre l'80% di rispondenti per mese. I restanti 654 comuni campione raggiungono mediamente un tasso di risposta del 70% ed, infine, gli oltre 7000 comuni non campione rispondenti per mese superano in media il 40%.
In generale, per tutti e tre i gruppi di comuni considerati si nota una leggera tendenza alla diminuzione del tasso di risposta man mano che ci si sposta dai primi mesi dell'anno fino a dicembre.
Al fine di mantenere un contatto con i comuni, finalizzato a mantenere oltre che migliorare il livello dei tassi di risposta, si effettuano periodicamente dei solleciti telefonici ai comuni campione e, tramite le Camere di commercio, ai comuni non campione.

Tavola 8: *Tassi di risposta mensile - anno 2005 **

mese	Comuni autorappresentativi (160 unità)	Restanti comuni campione (654 unità)	Totale comuni campione (814 unità)	Comuni non campione (7286 unità)
gen	83,1	69,7	72,4	46,4
feb	84,4	69,0	72,0	44,5
mar	83,8	69,9	72,6	38,2
apr	87,5	71,3	74,4	42,1
mag	81,9	69,7	72,1	37,1
giu	78,1	65,0	67,6	34,6
lug	82,5	69,0	71,6	44,2
ago	80,0	67,7	70,1	44,9
set	79,4	66,7	69,2	43,1
ott	84,4	70,5	73,2	49,0
nov	83,1	67,6	70,6	46,3
dic	83,1	67,0	70,1	42,0

** A 180 giorni dal periodo di riferimento degli indicatori trimestrali*

3. Sviluppi futuri

Tra le prossime innovazioni, da realizzare al fine di aumentare l'affidabilità delle stime, vengono proposte delle modifiche sia di metodo che di processo; tra le prime, cui abbiamo accennato anche in precedenza, citiamo:

a) l'introduzione del trattamento delle MRT nello strato dei comuni autorappresentativi dell'indagine campionaria che, per omogeneità, verrà adottato sugli stessi comuni anche nell'indagine censuaria;
b) la sostituzione dell'attuale stimatore (FABI), usato per il calcolo degli indicatori congiunturali dei comuni non autorappresentativi, con uno stimatore di tipo Shrinkage, di cui si stanno valutando i risultati.

Un'altra importante modifica riguarda la riorganizzazione di alcune fasi dei processi operativi delle due rilevazioni.

Abbiamo visto che l'indagine censuaria e quella campionaria hanno in comune il questionario e parte della rete di rilevazione. Le successive fasi di controllo e correzione, di registrazione e validazione dei dati, rilevati sui comuni campione, vengono gestite separatamente, prima nell'ambito della rilevazione campionaria, per elaborare gli indicatori nei tempi previsti e poi nell'ambito della rilevazione censuaria. Pertanto, due lavorazioni separate comportano sia la duplicazione di attività del tutto analoghe sia la coesistenza di due diversi archivi di dati, uno relativo alla rilevazione censuaria (totale comuni) e uno alla campionaria, che contiene soltanto le principali variabili rilevate sui comuni campione. Tra l'altro, i due archivi, pur avendo in comune una parte di informazioni relative ai comuni campione, possono divergere tra loro, a causa di aspetti legati al doppio trattamento dei dati.

Quindi, per rendere più efficiente il processo produttivo e, nel contempo, eliminare le difformità tra i database sembra opportuno unificare il più possibile il trattamento dei dati relativi ai comuni campione, limitandosi a trattare separatamente soltanto le informazioni pervenute oltre il tempo limite che consente di farle confluire nel processo censuario per la produzione degli indicatori alle scadenze prefissate.

Avendo, infatti, verificato che è possibile trattare con le procedure previste dalla rilevazione censuaria tutti i questionari, dei comuni campione e non, pervenuti e revisionati entro i primi 50 giorni circa di ogni trimestre, soltanto i modelli che perverranno nei successivi 25 giorni verranno trattati a parte per rilevare le informazioni utili alla stima degli indicatori, che altrimenti non entrerebbero a far parte del calcolo.

Bibliografia

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

Di Zio M., Guarnera U., Luzi O. *Rilevazione rapida dei permessi di costruire (grandi comuni): Risultati della sperimentazione delle tecniche di imputazione con donatore longitudinale e con media longitudinale per le MRP*. Documento interno Istat, 2007.

Di Zio M. *Sperimentazione sul trattamento delle mancate risposte totali per la stima anticipata della rilevazione rapida dei permessi di costruire*. Documento interno Istat, 2007.

Istat (a cura di G. Rallo). *Statistiche sui permessi di costruire*. Collana informazioni n. 32/2005.

Garozzo S. *Progetto di indagine rapida dell'attività edilizia*. Documento interno Istat, 2003.

Bacchini F., Iannaccone R., Otranto E. "L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione italiani". *Contributi Istat,* n. 4/2005.

Falorsi P.D., Alleva G., Bacchini F., Iannaccone R. "Estimated based on preliminary data from a specific subsample and from respondents not included in the subsample". *Statistical Methods and Applications*, vol. 14, n.1, 83-99, 2005.

Bacchini F., Falorsi P.D., Iannaccone R. *Il miglioramento della stima dei permessi si costruzione a 90 giorni*". Documento interno Istat, 2005.

Prime innovazioni nel processo di controllo e correzione dei dati della rilevazione Extrastat

Mariagloria Narilli, Istat, Servizio Statistiche sul Commercio con l'Estero
Alessandra Nuccitelli, Istat, Direzione Centrale delle Statistiche sui Prezzi e il Commercio con l'Estero

Sommario: Fino alla fine del 2006, i parametri di segnalazione del prezzo medio nella rilevazione del commercio con l'estero Extrastat erano stabiliti in modo ragionato. Tali parametri richiedevano un aggiornamento periodico - svolto annualmente o, al più, con cadenza biennale - con notevole dispendio in termini di tempo e di risorse umane. Inoltre, l'aggiornamento dei parametri si rivelava spesso non abbastanza tempestivo da tenere conto delle variazioni dei prezzi sui mercati internazionali, provocando in tal modo un numero eccessivamente elevato di "false" segnalazioni.
A partire da gennaio 2007, è stata introdotta una procedura automatica per la determinazione dei parametri. Secondo tale procedura, i parametri di segnalazione sono calcolati mensilmente sulla base della distribuzione dei dati "grezzi" relativi agli ultimi 24 mesi. Tra i vantaggi connessi all'adozione della procedura, vanno evidenziati sia una maggiore accuratezza ed oggettività nell'individuazione delle situazioni di errore rispetto al passato, sia una riduzione dei tempi e dei costi (in termini di risorse umane) legati alla revisione dei dati e all'aggiornamento dei parametri di segnalazione.

Parole chiave: Nomenclatura Combinata, intervallo di accettazione, analisi della varianza, editing selettivo

Le collane esistenti presso l'ISTAT - Contributi e Documenti - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell'ISTAT e del Sistan, o da studiosi esterni.
I lavori pubblicati nella collana Contributi Istat vengono fatti circolare allo scopo di suscitare la discussione attorno ai risultati preliminari di ricerca in corso.
I Documenti Istat hanno lo scopo di fornire indicazioni circa le linee, i progressi ed i miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.
I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

1. Introduzione[139]

La rilevazione del commercio con l'estero Extrastat fornisce informazioni sullo scambio di merci dell'Italia con i Paesi non appartenenti all'Unione Europea. Si tratta di un'indagine di tipo totale: i dati provengono dalle dichiarazioni (bollette doganali o DAU[140]) rese dagli operatori economici presso la dogana attraverso cui le merci transitano. I dati riportati sulle bollette doganali sono registrati e inviati per via telematica al Centro Elaborazioni Dati dell'Agenzia delle Dogane che li trasmette all'Istat con cadenza mensile (circa 2,5 milioni di records ogni mese).

Le informazioni contenute nella bolletta doganale sono rilevate con finalità sia fiscale sia statistica: valore e quantità della merce scambiata, origine, provenienza, destinazione, ecc.. La compilazione della bolletta risulta più affidabile per quanto riguarda le informazioni che hanno rilevanza fiscale, alle quali sia gli operatori che gli uffici doganali prestano maggiore attenzione; le informazioni raccolte soltanto a fini statistici risultano indicate in maniera più approssimativa[141].

Le merci scambiate possono essere classificate a vari livelli di dettaglio: la classificazione merceologica più articolata è rappresentata dalla Nomenclatura Combinata (*NC8*) che prevede quasi 10.000 raggruppamenti di merci in posizioni ad otto cifre[142].

Il massimo livello di dettaglio per la diffusione dei dati di valore e quantità delle merci scambiate ogni mese è dato dall'incrocio delle modalità relative alle variabili *codice merceologico NC8*, *tipo di movimento* (importazione o esportazione), *Paese statistico*[143]. Dal momento che i dati sono diffusi ad un livello così poco aggregato, il processo di controllo e correzione deve garantire un elevato grado di accuratezza delle informazioni rilasciate.

La segnalazione delle unità statistiche - le *transazioni commerciali* - con dati potenzialmente errati avviene principalmente sulla base di intervalli di accettazione definiti per:

- il *prezzo medio*, dato dal rapporto tra il valore (in euro) e la quantità (in chilogrammi) della merce scambiata;
- il *peso medio*, dato dal rapporto tra la quantità espressa in chilogrammi e la quantità espressa in unità supplementari (ad esempio, litri, numero di pezzi, metri quadrati, ecc.) della merce scambiata.

Tali intervalli sono individuati per ogni combinazione di tipo di movimento e di merce a livello di Nomenclatura Combinata. Le osservazioni i cui valori cadono al di fuori dell'intervallo di accettazione sono considerate "segnalate" e quindi da sottoporre a revisione manuale. In fase di controllo e correzione delle segnalazioni è possibile anche un ritorno alla fonte dei dati (operatori commerciali o dogane).

Fino alla fine del 2006, le soglie delimitanti gli intervalli di accettazione (chiamate, d'ora in poi, *parametri di segnalazione*) erano stabilite in modo ragionato sulla base dell'esperienza acquisita dai revisori dei dati della rilevazione Extrastat. Tali parametri richiedevano un aggiornamento periodico - svolto annualmente o, al più, con cadenza biennale - con notevole dispendio in termini di tempo e di risorse

---

([139]) Il lavoro è frutto della collaborazione delle autrici. Il paragrafo 2 è stato redatto da Mariagloria Narilli; i paragrafi 4, 5, 6, 7 e l'Appendice sono stati redatti da Alessandra Nuccitelli; i paragrafi 1 e 3 sono stati redatti congiuntamente da Mariagloria Narilli e Alessandra Nuccitelli. Un ringraziamento va a Ersilia Di Pietro, Natale Renato Fazio, Pasquale Mazza e alle persone impegnate nella revisione dei dati Extrastat per i preziosi suggerimenti forniti.

([140]) Documento Amministrativo Unico.

([141]) Per ulteriori dettagli sulle caratteristiche della rilevazione Extrastat si rimanda a Di Pietro (2006).

([142]) La Nomenclatura Combinata rappresenta una disaggregazione del Sistema Armonizzato che è costituito da raggruppamenti di merci in posizioni a 6 cifre (*SH6*); ad un livello più aggregato si considerano i sottocapitoli, ovvero i raggruppamenti in posizioni a 4 cifre (*SH4*), e i capitoli, ovvero i raggruppamenti in posizioni a 2 cifre (*SH2*).

([143]) Per le esportazioni, il *Paese statistico* coincide con il Paese di destinazione della merce. Per le importazioni, il *Paese statistico* è il Paese di origine della merce; tuttavia, in alcuni casi - ad esempio, per le merci la cui origine non è nota, per le opere d'arte, ecc. - al *Paese statistico* viene attribuito il Paese di provenienza.

umane dedicate. Inoltre, l'aggiornamento dei parametri relativi al *prezzo medio* si rivelava spesso non abbastanza tempestivo da tenere conto delle variazioni dei prezzi sui mercati internazionali, provocando in tal modo un numero eccessivamente elevato di "false" segnalazioni.
A partire da gennaio[144] 2007, contestualmente all'adeguamento del processo di produzione dei dati alle novità normative legate all'ingresso della Romania e della Bulgaria nell'Unione Europea e all'aggiornamento annuale della Nomenclatura Combinata, sono stati introdotti criteri innovativi per la segnalazione dei valori anomali.
Al momento, le innovazioni introdotte nel processo di lavorazione dei dati sono state dettate soprattutto dall'esigenza di migliorare la qualità delle stime finali senza dover apportare modifiche sostanziali al sistema informativo già esistente e alle applicazioni correntemente utilizzate per il trattamento interattivo dei dati, dal momento che tali modifiche avrebbero comportato tempi e costi di attuazione non sostenibili.
Questo documento è incentrato sull'elemento innovativo principale, rappresentato dall'introduzione di una procedura automatica per la determinazione degli intervalli di accettazione relativi al *prezzo medio*. Secondo tale procedura, i parametri di segnalazione sono calcolati mensilmente sulla base della distribuzione dei dati "grezzi" (cioè, non lavorati) relativi agli ultimi 24 mesi. Tra i vantaggi connessi all'adozione della procedura, va sottolineata una maggiore accuratezza ed oggettività nell'individuazione delle situazioni di errore rispetto al passato.
Il documento è strutturato come segue. Nel paragrafo 2 è illustrato il processo di lavorazione dei dati della rilevazione Extrastat; inoltre, nel paragrafo 3 viene fornito un quadro degli interventi migliorativi introdotti gradualmente in tale processo a partire da aprile 2006. La procedura automatica per la determinazione degli intervalli di accettazione relativi al prezzo medio è descritta nel paragrafo 4; nel paragrafo 5 sono riportati i risultati principali delle sperimentazioni effettuate, volte a confrontare i parametri determinati automaticamente con quelli individuati in modo ragionato dai revisori. Nel paragrafo 6 viene valutata l'opportunità di utilizzare intervalli di accettazione specifici, definiti per *Paese statistico* o per *operatore commerciale*, oltre che per tipo di movimento e tipologia merceologica. Infine (paragrafo 7), sono evidenziate alcune criticità non risolte e si accenna ai futuri sviluppi della procedura proposta al fine di migliorare ulteriormente la qualità delle informazioni finali.

2. Il processo di lavorazione dei dati

2.1 Un quadro d'insieme

Con cadenza mensile l'Istat riceve dall'Agenzia delle Dogane circa 2,5 milioni di records - di cui 1,5 milioni di interesse statistico - contenenti:

- i microdati grezzi relativi alle transazioni commerciali registrate nel mese precedente (mese di riferimento o di rilevazione dei dati[145]);
- i microdati relativi a rettifiche e annullamenti[146] che si riferiscono a transazioni registrate nel mese precedente o anche nei mesi passati.

I dati di interesse statistico, oggetto principale di revisione, riguardano il *valore*, espresso in euro, e la *quantità*, espressa in chilogrammi o in unità supplementari, relativi ad ogni transazione commerciale dell'Italia con i Paesi extra UE.

(144) Dove non altrimenti specificato, il *mese* indicherà d'ora in poi il mese di riferimento dei dati.
(145) Per maggiori precisazioni sul periodo di riferimento dei dati si rimanda a Di Pietro (2006).
(146) Una "rettifica" consiste in una dichiarazione che modifica una bolletta precedente in una qualsiasi delle informazioni riportate; per "annullamento" si intende la cancellazione di una dichiarazione presentata in precedenza.

Facendo il rapporto tra il valore e la quantità espressa in chilogrammi relativi ad una certa transazione, si ottiene il *prezzo medio* della merce scambiata. Nei casi in cui il volume della merce scambiata è misurato utilizzando una unità supplementare, è possibile definire anche il *prezzo unitario*, espresso in euro per unità supplementare, e il *peso medio*, ovvero il rapporto tra la quantità espressa in chilogrammi e la quantità espressa in unità supplementari.

I dati elementari pervenuti da parte dell'Agenzia delle Dogane vengono sottoposti ad un complesso processo di lavorazione che può essere schematicamente suddiviso in tre fasi:

1. dall'acquisizione dei dati dall'Agenzia delle Dogane alla produzione dei dati ("provvisori") per il comunicato stampa;
2. dal comunicato stampa alla produzione dei dati (detti "definitivi" o "revisionati") per l'aggiornamento mensile di Coeweb, la banca dati sulle statistiche del commercio estero consultabile all'indirizzo www.coeweb.istat.it;
3. dall'aggiornamento mensile di Coeweb alla produzione dei cosiddetti dati "storici" - anch'essi diffusi tramite Coeweb - risultanti dall'ultima revisione che i dati "definitivi" subiscono entro la fine dell'anno successivo a quello di riferimento.

Con riferimento all'anno 2005, nella tavola 1 sono riportate le variazioni percentuali del valore mensile risultanti alla fine delle varie fasi del processo di lavorazione dei dati, per tipo di movimento. La tavola 2, riferendosi all'anno 2006, riporta le variazioni relative alle sole prime due fasi. Come si può vedere, le maggiori variazioni avvengono nella fase di produzione dei dati provvisori; nelle fasi successive sono apportate correzioni di minore entità.

La prima fase di lavorazione prevede controlli di carattere generale e determina la validazione dei dati in forma provvisoria. I dati prodotti al termine di questa fase sono utilizzati esclusivamente per la predisposizione del comunicato stampa che viene diffuso circa 25 giorni dopo la fine del mese di riferimento. Il comunicato stampa contiene dati aggregati sulle importazioni e sulle esportazioni e il saldo della bilancia commerciale per settore di attività economica (individuato dalle prime due cifre della classificazione ATECO 2002) e per Paese statistico e area geoeconomica[147]. Il comunicato stampa rappresenta una scadenza di particolare importanza, vincolata alle date di rilascio stabilite dall'Istat alla fine dell'anno precedente e all'acquisizione dei dati dall'Agenzia delle Dogane. In questa fase la revisione, concentrata in pochi giorni lavorativi (circa cinque), riguarda soltanto le transazioni con maggiore impatto sulla bilancia commerciale, chiamate "alti valori", ovvero le transazioni caratterizzate da valori superiori a 500 mila euro, da quantità superiori a un milione di chilogrammi oppure da quantità superiori a 10 milioni di unità supplementari. La revisione consiste nel controllo accurato e nella eventuale correzione dei dati di valore e/o quantità utilizzando tutte le informazioni disponibili, in alcuni casi ottenute anche con un ritorno alla fonte (operatori commerciali o dogane).

Rispetto alla precedente, la seconda fase di lavorazione, di durata maggiore (10-15 giorni lavorativi), prevede controlli e verifiche puntuali sulle transazioni potenzialmente errate (segnalate) ed un'eventuale rettifica dei dati provvisori diffusi mediante il comunicato stampa. I dati prodotti alla fine di questa fase sono diffusi circa due mesi e mezzo dopo la fine del mese di riferimento tramite Coeweb. Questa banca dati contiene informazioni sui flussi commerciali dell'Italia con gli altri Paesi ad un livello molto più dettagliato di quello del comunicato stampa: infatti, accedendo al data warehouse è possibile effettuare un'ampia gamma di interrogazioni sui dati di valore e quantità (in chilogrammi e in unità supplementari) fino al livello di dettaglio dato dall'incrocio delle modalità relative alle variabili *codice merceologico NC8*, *tipo di movimento*, *Paese statistico*.

Terminata la seconda fase di lavorazione, i dati sono conservati in una banca dati di produzione dove possono subire altre modifiche, chiamate "correzioni fuori mese", apportate dai revisori sulla base di

(147) I Paesi statistici e le aree geoeconomiche considerate per il comunicato stampa sono: EFTA, Russia, Altri Paesi europei, Turchia, OPEC, USA, Mercosur, Cina, Giappone, EDA, Altri Paesi.

informazioni giunte in ritardo o a seguito di ulteriori verifiche resesi necessarie per rilievi da parte di utenti. I dati possono essere modificati fino alla fine del mese di ottobre dell'anno successivo a quello di riferimento; a questo punto sono considerati "storici" ed estratti dalla banca dati di produzione per l'aggiornamento annuale di Coeweb.

***Tavola 1*: Valore mensile (in migliaia di euro) dei dati iniziali, provvisori, definitivi e storici con variazioni percentuali, per tipo di movimento**

Dati Extrastat (anno 2005)

mese di riferimento	dati iniziali	dati provvisori	variazione tra dati provvisori e dati iniziali	dati definitivi	variazione tra dati definitivi e dati provvisori	dati storici	variazione tra dati storici e dati definitivi
				importazione			
Gennaio	9.435.481	9.439.004	0,04	9.439.252	0,00	9.464.315	0,27
Febbraio	10.148.487	9.674.708	-4,67	9.679.572	0,05	9.562.471	-1,21
Marzo	11.523.282	11.182.727	-2,96	11.187.545	0,04	11.103.985	-0,75
Aprile	11.066.861	10.953.945	-1,02	10.865.700	-0,81	10.867.017	0,01
Maggio	11.744.311	11.133.878	-5,20	11.138.866	0,04	11.172.862	0,31
Giugno	10.797.502	10.793.928	-0,03	10.795.100	0,01	10.886.813	0,85
Luglio	11.156.203	11.171.105	0,13	11.171.461	0,00	11.238.943	0,60
Agosto	9.467.448	9.482.065	0,15	9.484.193	0,02	9.539.077	0,58
Settembre	12.284.608	12.053.794	-1,88	12.054.548	0,01	12.057.810	0,03
Ottobre	11.824.812	11.449.994	-3,17	11.460.656	0,09	11.444.374	-0,14
Novembre	11.999.850	12.084.002	0,70	12.067.574	-0,14	12.063.309	-0,04
Dicembre	11.469.215	11.352.230	-1,02	11.347.133	-0,04	11.345.809	-0,01
totale	**132.918.060**	**130.771.380**	**-1,62**	**130.691.600**	**-0,06**	**130.746.785**	**0,04**
				esportazione			
Gennaio	8.841.825	7.479.825	-15,40	7.480.429	0,01	7.474.247	-0,08
Febbraio	8.959.060	8.546.834	-4,60	8.540.538	-0,07	8.546.123	0,07
Marzo	10.503.131	10.537.556	0,33	10.532.949	-0,04	10.194.613	-3,21
Aprile	12.285.230	9.877.245	-19,60	9.879.700	0,02	9.879.950	0,00
Maggio	14.794.069	10.647.276	-28,03	10.591.949	-0,52	10.602.853	0,10
Giugno	11.663.179	10.644.004	-8,74	10.645.371	0,01	10.653.555	0,08
Luglio	13.065.954	11.673.243	-10,66	11.700.269	0,23	11.695.335	-0,04
Agosto	9.724.515	9.109.704	-6,32	9.111.295	0,02	9.135.002	0,26
Settembre	10.233.301	9.952.093	-2,75	9.967.829	0,16	9.999.832	0,32
Ottobre	11.366.928	11.218.360	-1,31	11.225.957	0,07	11.260.975	0,31
Novembre	11.921.114	11.262.425	-5,53	11.203.524	-0,52	11.222.392	0,17
Dicembre	12.049.682	11.480.773	-4,72	11.488.796	0,07	11.504.943	0,14
totale	**135.407.988**	**122.429.338**	**-9,58**	**122.368.606**	**-0,05**	**122.169.820**	**-0,16**

***Tavola 2*: Valore mensile (in migliaia di euro) dei dati iniziali, provvisori e definitivi con variazioni percentuali, per tipo di movimento**

Dati Extrastat (anno 2006)

mese di riferimento	dati iniziali	dati provvisori	variazione tra dati provvisori e dati iniziali	dati definitivi	variazione tra dati definitivi e dati provvisori
			importazione		
Gennaio	12.581.670	12.513.964	-0,54	12.523.907	0,08
Febbraio	12.687.345	12.632.670	-0,43	12.640.045	0,06
Marzo	13.950.184	13.658.157	-2,09	13.659.121	0,01
Aprile	12.083.099	11.952.516	-1,08	11.955.456	0,02
Maggio	14.087.855	13.981.423	-0,76	13.983.095	0,01
Giugno	13.020.420	12.988.494	-0,25	13.012.327	0,18
Luglio	13.494.885	13.442.097	-0,39	13.446.218	0,03
Agosto	11.633.777	11.661.089	0,23	11.665.911	0,04
Settembre	13.778.197	13.419.079	-2,61	13.477.460	0,44
Ottobre	13.823.095	13.641.133	-1,32	13.637.428	-0,03
Novembre	13.383.905	13.352.789	-0,23	13.355.341	0,02
Dicembre	12.474.358	12.425.302	-0,39	12.427.959	0,02
totale	**156.998.790**	**155.668.713**	**-0,85**	**155.784.268**	**0,07**
			esportazione		
Gennaio	9.194.895	8.927.940	-2,90	8.905.277	-0,25
Febbraio	11.608.701	10.041.547	-13,50	10.066.110	0,24
Marzo	13.538.394	12.222.926	-9,72	12.245.455	0,18
Aprile	10.362.257	10.159.203	-1,96	10.192.371	0,33
Maggio	14.966.025	12.373.077	-17,33	12.378.858	0,05
Giugno	14.255.643	12.211.768	-14,34	12.214.908	0,03
Luglio	12.335.791	11.941.252	-3,20	11.985.332	0,37
Agosto	10.079.874	9.731.466	-3,46	9.744.418	0,13
Settembre	12.265.430	10.663.155	-13,06	10.750.348	0,82
Ottobre	12.915.540	12.574.223	-2,64	12.593.541	0,15
Novembre	13.089.825	12.642.438	-3,42	12.651.964	0,08
Dicembre	13.270.129	12.919.888	-2,64	12.937.515	0,14
totale	**147.882.504**	**136.408.883**	**-7,76**	**136.666.097**	**0,19**

Le tecniche di controllo e correzione utilizzate nel processo di lavorazione dei dati sono di due tipi:

- *verifica interattiva*: avviene in tutte le tre fasi. I revisori richiamano il record relativo alla singola transazione in esame tramite un'applicazione client-server disponibile sul loro personal computer, controllano ed eventualmente correggono i dati in modo da assicurare la coerenza delle informazioni. L'applicazione client-server, sviluppata in Oracle Forms Developer, presenta molte funzionalità; ad esempio, permette di effettuare:
  - l'aggiornamento e la consultazione del registro delle imprese;
  - la consultazione dei metadati (nomenclature, classificazioni geografiche, elenco degli uffici doganali);
  - la consultazione, la correzione, l'inserimento e la cancellazione di dati elementari;
  - la consultazione degli intervalli di accettazione relativi a prezzo medio e a peso medio per ogni combinazione di tipo di movimento e di merce a livello di Nomenclatura Combinata (la consultazione degli intervalli di accettazione viene effettuata soprattutto durante la correzione del dato, in quanto fornisce al revisore utili indicazioni sul valore o sulla quantità da imputare);
  - il confronto dei valori e delle quantità (in chilogrammi e in unità supplementari) rispetto allo stesso mese dell'anno precedente secondo diversi criteri di aggregazione (operatore commerciale, Nomenclatura Combinata, Paese statistico, ecc.);
  - il calcolo e la consultazione di bilance[148] mensili (o cumulate sui mesi precedenti) secondo diversi criteri di aggregazione (Nomenclatura Combinata, dogana, Paese statistico, ecc.);
- *correzioni automatiche*: avvengono soltanto nelle fasi 1 e 2 e sono discusse più approfonditamente nei sottoparagrafi 2.2 e 2.3.

Nei sottoparagrafi seguenti si illustrano più in dettaglio le prime due fasi del processo di lavorazione dei dati.

2.2 Fase 1: dai dati iniziali ai dati provvisori

Nella prima fase, dopo opportuni controlli formali e quantitativi, il file acquisito dall'Agenzia delle Dogane viene caricato nell'ambiente di produzione Oracle dove subisce una serie di trasformazioni che portano i dati dal formato originario al formato desiderato per i successivi trattamenti.
Fondamentalmente, tali trasformazioni sono finalizzate a:

- l'eliminazione dei campi relativi ad informazioni che non sono di interesse statistico;
- il calcolo, a partire dai valori delle variabili originarie, dei valori relativi alle variabili di tipo derivato (ad esempio, *Paese statistico*, *prezzo medio*, *peso medio*, ecc.);
- l'esclusione di records non rilevanti ai fini delle statistiche del commercio con l'estero (ad esempio, i records relativi ai transiti);
- l'individuazione di particolari tipologie di records, come le rettifiche e gli annullamenti - che sono messi da parte per un utilizzo successivo - e gli EU da DAU[149] - che vengono passati alla rilevazione Intrastat, riguardante gli scambi di merci dell'Italia con i Paesi appartenenti all'Unione Europea;
- l'esclusione dei records relativi ai "bassi valori"[150];

(148) Per "bilance" si intendono le aggregazioni dei dati elementari di valore e quantità (in chilogrammi e in unità supplementari) effettuate per entrambi i tipi di movimento secondo varie classificazioni (merceologiche, economiche, territoriali, ecc.).
(149) Con la dicitura "EU da DAU" sono indicati i records relativi a scambi con zone dell'Unione Europea soggette a particolari trattamenti fiscali, per i quali è stato emesso il DAU anziché il documento di cessione o di acquisto.
(150) Il Regolamento comunitario n° 1669/2001 stabilisce che siano oggetto delle statistiche del commercio con l'estero le transazioni con valore o massa netta superiore rispettivamente a 1.000 euro o a 1.000 chilogrammi. L'applicazione di tali limiti è facoltativa, purché ogni Stato membro comunichi la soglia e i metodi di adeguamento prescelti.

- l'individuazione dei records *formalmente incompatibili*, ovvero dei records che presentano un errore di classificazione (merceologica, territoriale, ecc.) o mancano delle informazioni sul mezzo di trasporto attivo[151] o sulla quantità espressa in unità supplementari; se prima non si effettua la correzione manuale o automatica che sana l'incompatibilità, non si può procedere con la lavorazione di tali records e tenerne conto nel calcolo delle bilance.

Alla fine di questa fase i records dei microdati grezzi mensili sono approssimativamente 800.000; circa un terzo di questi riguarda le importazioni, mentre i restanti due terzi si riferiscono alle esportazioni. I records formalmente incompatibili sono circa 10.000 e riguardano soprattutto le esportazioni.

Terminato il caricamento dei dati nell'ambiente di produzione Oracle, vengono calcolate le bilance[152] iniziali, visualizzabili da parte dei revisori tramite l'applicazione client-server, secondo diversi criteri di aggregazione dei dati. A questo stadio della lavorazione, la verifica interattiva dei dati riguarda solo gli "alti valori". Tali transazioni, pur rappresentando meno dell'1 per cento del totale delle transazioni, hanno un forte impatto sulla bilancia commerciale.

I revisori controllano ed eventualmente correggono i dati di valore e/o quantità utilizzando tutte le informazioni disponibili, in alcuni casi ottenute anche con un ritorno alla fonte. Particolare attenzione viene posta nella revisione dei records relativi alle voci che incidono maggiormente sulla bilancia commerciale italiana:

- mezzi di trasporto (aerei, navi e automobili);
- prodotti energetici (metano, energia elettrica e petrolio grezzo).

Per quanto riguarda i records formalmente incompatibili, viene data la precedenza a quelli relativi agli "alti valori", in modo da poterne tenere conto nell'elaborazione dei dati provvisori per il comunicato stampa. In questa prima fase sono recuperati circa 150 records.

Oltre alla verifica interattiva, vengono effettuate alcune correzioni di tipo automatico.

Un primo tipo di correzione automatica riguarda i records formalmente incompatibili per mancanza di informazioni sul mezzo di trasporto attivo (1.000-1.500 records). La correzione avviene attribuendo il mezzo di trasporto più appropriato in base alla posizione geografica e alle caratteristiche della dogana coinvolta nello scambio commerciale. Si assegna il mezzo di trasporto navale in caso di dogana marittima e quello aereo in caso di dogana aeroportuale; in caso di dogana terrestre, viene data la preferenza al trasporto su strada, in quanto più frequente del trasporto su rotaia.

Un secondo tipo di correzione automatica riguarda le provviste di bordo[153] portuali in esportazione (circa 1.000 records). Ai fini delle statistiche del commercio estero per questi prodotti va considerata la nazionalità del mezzo di trasporto. Nel caso di provviste destinate a navi extra UE, i dati provenienti dall'Agenzia delle Dogane risultano sistematicamente inferiori al dato reale in quanto, nella compilazione del DAU, nel campo relativo al Paese di destinazione viene indicata la nazionalità della bettolina (italiana) che serve per il trasbordo delle provviste dal porto alla nave, anziché la nazionalità della nave (extra UE) a cui sono destinati i prodotti. Pertanto, tramite una procedura automatica di correzione, il valore delle provviste di bordo viene attribuito ai Paesi extra UE.

Fino a dicembre 2006 l'Istat ha applicato le seguenti soglie, relative al valore della transazione:

- € 516 per i capitoli 1-24 (animali vivi e prodotti del regno animale, prodotti del regno vegetale, prodotti alimentari);
- € 620 per gli altri capitoli (25-99).

Dal 1° gennaio 2007, invece, sono applicati i seguenti limiti:

- € 700 per i capitoli 1-14 (animali vivi e prodotti del regno animale, prodotti del regno vegetale);
- € 1.000 per gli altri capitoli (15-99).

Le transazioni il cui valore è inferiore alle soglie prefissate sono definite "bassi valori".

(151) Il "mezzo di trasporto attivo" indica il tipo di trasporto con cui si presume che le merci entrino o escano dal territorio doganale della Comunità.

(152) Le stesse bilance saranno poi ricalcolate in fase di elaborazione dei dati provvisori per il comunicato stampa e dei dati definitivi mensili (fase 2).

(153) Per "provviste di bordo" si intendono le merci consumate a bordo di navi o aerei dalle persone oppure le merci che servono per il funzionamento dei motori.

Un terzo tipo di correzione automatica riguarda le transazioni relative ad acque e vini, per i quali è prevista l'indicazione della quantità in litri, oltre che in chilogrammi. La procedura automatica - che interessa circa 3.000 records - viene applicata per correggere la quantità (in chilogrammi o in unità supplementari), quando erroneamente riportata, sulla base della corrispondenza esistente tra chilogrammo e litro per questa tipologia di prodotti.
Infine, un altro tipo di correzione automatica riguarda le transazioni relative a prodotti di abbigliamento e calzature, per i quali è prevista l'indicazione del numero di pezzi. La procedura, che imputa automaticamente le unità supplementari, viene eseguita per quei records (circa 1.000) per cui le unità supplementari sono pari a 1, il prezzo medio è interno e il peso medio è esterno ai rispettivi intervalli di accettazione stabiliti per tipo di movimento e di merce a livello di Nomenclatura Combinata.
Ogni mese le correzioni automatiche suddette interessano complessivamente circa 6.000-6.500 records.
Terminata la revisione degli "alti valori" ed effettuate le correzioni automatiche, vengono estratti i dati provvisori per il comunicato stampa e calcolate le bilance per settore di attività economica e per Paese statistico e area geoeconomica.

2.3. Fase 2: dai dati provvisori ai dati definitivi mensili

Diffusi i dati provvisori, l'individuazione delle transazioni con dati potenzialmente errati avviene sulla base di intervalli di accettazione per il *prezzo medio* e per il *peso medio*, definiti per ogni combinazione di tipo di movimento e di merce a livello di Nomenclatura Combinata. Sono considerate segnalate:

- le transazioni per cui il prezzo medio e/o il peso medio cade al di fuori del corrispondente intervallo di accettazione;
- le transazioni con quantità espressa in chilogrammi uguale a zero;
- le transazioni relative a prodotti che incidono in modo rilevante sulla bilancia commerciale, che presentano delle specificità da tenere sotto controllo o sui quali sono frequenti i rilievi da parte degli utenti (caffè [solo importazioni], petrolio greggio, energia elettrica, diamanti, perle, aerei, navi, opere d'arte).

Le transazioni segnalate, superiori a particolari soglie[154] di valore e di quantità - definite per tipo di movimento e per capitolo - sono sottoposte a revisione manuale; sono controllati accuratamente ed eventualmente corretti i dati di *valore*, *quantità*, *codice merceologico NC8*, *Paese statistico* utilizzando tutte le informazioni disponibili, in alcuni casi ottenute anche con un ritorno alla fonte (operatori commerciali o dogane).
Parallelamente si prosegue nella correzione dei records formalmente incompatibili; alla fine di questa fase vengono recuperati circa 6.000 records. Tuttavia, allo stato attuale il processo di lavorazione non prevede che i valori di tali records, una volta risolta l'incompatibilità, siano sottoposti al confronto con gli intervalli di accettazione; pertanto, i dati di interesse statistico (*valore* e *quantità*) relativi ai records recuperati potrebbero ancora contenere errori non accuratamente revisionati.
Le correzioni di tipo automatico effettuate in questa fase riguardano le transazioni relative a una serie di prodotti per i quali è prevista l'indicazione della quantità in unità supplementari: legno e prodotti in legno, porte e finestre, abbigliamento, calzature, prodotti ceramici, perle e metalli preziosi, cinture di sicurezza, armi e cartucce. Tali correzioni avvengono dopo la revisione e l'eventuale imputazione dei dati di *valore*, *quantità* (in chilogrammi), *codice merceologico NC8*, *Paese statistico* e sono applicate alle sole transazioni il cui peso medio cade al di fuori dell'intervallo di accettazione corrispondente; in questi casi la correzione della quantità espressa in unità supplementari avviene sulla base della corrispondenza tipica osservata tra chilogrammo e unità supplementare specifica di ciascuna tipologia di prodotto.

(154) Tali soglie, definite sulla base dell'esperienza acquisita dai revisori in modo da concentrare le attività di controllo e correzione sulle osservazioni caratterizzate da valori o quantità elevate, sono state introdotte a partire da marzo 2006 per far fronte ad una drastica riduzione del personale addetto alla revisione dei dati.

3. Principali innovazioni nel processo di lavorazione dei dati a partire da aprile 2006

Vincoli di diversa natura condizionano l'introduzione di innovazioni nel processo di lavorazione dei dati della rilevazione Extrastat; i più rilevanti sono i seguenti:

- *vincoli temporali*: il controllo e la correzione dei dati devono essere effettuati in pochi giorni lavorativi per garantire il rispetto delle scadenze di rilascio delle informazioni;
- *vincoli tecnologici*: modifiche sostanziali al sistema informativo e alle applicazioni esistenti comporterebbero tempi e costi non sostenibili con le risorse attualmente a disposizione;
- *vincoli normativi*: la produzione delle statistiche sul commercio con l'estero è regolamentata da Eurostat e numerosi sono i cambiamenti introdotti ogni anno nella normativa vigente;
- *vincoli di diffusione*: i dati sono diffusi ad un livello molto dettagliato.

L'introduzione di innovazioni radicali di processo in un contesto così complesso quale quello della rilevazione Extrastat comporta tempi e costi notevoli che vanno pianificati accuratamente, in modo da gravare il meno possibile sull'attività di produzione corrente. Alla luce di queste considerazioni e tenendo conto delle risorse disponibili, si è preferito intervenire sul processo di lavorazione dei dati in maniera molto graduale.

Di seguito sono illustrate le innovazioni apportate a partire da aprile 2006; tali innovazioni sono state introdotte con l'obiettivo di ridurre la quantità dei controlli da effettuare da parte dei revisori, mantenendo inalterate le modalità di lavorazione e garantendo almeno lo stesso livello di qualità dei dati prodotti. Specificamente, le innovazioni riguardano:

1) il trattamento automatico delle rettifiche trasmesse mensilmente all'Istat dall'Agenzia delle Dogane (a partire da aprile 2006);
2) l'introduzione di una procedura automatica per la determinazione degli intervalli di accettazione per il prezzo medio (a partire da gennaio 2007).

Con la prima innovazione, è stata implementata una diversa utilizzazione delle rettifiche acquisite dall'Agenzia delle Dogane (circa 17.000 records al mese). Si ricorda che le rettifiche possono riguardare sia transazioni registrate nel mese precedente, sia transazioni registrate nei mesi passati. Fino a marzo 2006, le rettifiche pervenute mensilmente nel corso dell'anno venivano utilizzate a distanza di tempo per la definizione dei dati "storici", correndo il rischio di sovrascrivere records già revisionati e corretti. A partire da aprile 2006, le rettifiche sono lavorate subito dopo il caricamento dei dati nell'ambiente di produzione Oracle e, rappresentando una correzione alla fonte, il loro utilizzo consente una riduzione del numero di segnalazioni nella fase 2. Anche se finora, per motivi tecnici, non è stato possibile quantificare tale riduzione, l'inserimento immediato delle rettifiche nel processo di lavorazione ha rappresentato il primo passo verso l'ottimizzazione del processo di lavorazione e il miglioramento della qualità dei dati prodotti.

L'altra innovazione è rappresentata dall'introduzione di una procedura automatica per la determinazione degli intervalli di accettazione per il *prezzo medio*. Tale procedura viene discussa approfonditamente nei paragrafi successivi. Come si capirà meglio in seguito, l'applicazione della stessa procedura alla variabile *peso medio* non ha prodotto risultati soddisfacenti. Pertanto, gli intervalli di accettazione per il *peso medio* sono definiti dai limiti di peso contenuti nella descrizione merceologica associata al codice *NC8*, quando presenti, oppure in modo ragionato dai revisori.

Ulteriori interventi innovativi, attualmente in corso di implementazione, riguardano:

I. il trattamento dei records formalmente incompatibili;
II. il trattamento delle riparazioni[155], che al momento, per errori di compilazione del DAU, possono essere confuse con transazioni temporanee diverse dalle riparazioni;

([155]) Secondo il Regolamento comunitario n° 1949/2005, le riparazioni non sono oggetto delle statistiche del commercio con l'estero, in quanto non modificano la natura della merce scambiata. Le riparazioni sono caratterizzate dalla temporaneità

Per quanto riguarda il trattamento dei records formalmente incompatibili si sta agendo in più direzioni:

- correzione automatica di alcune incompatibilità di tipo sistematico, che attualmente vengono risolte manualmente;
- introduzione di nuove regole di incompatibilità;
- completamento della revisione dei records incompatibili prima dell'inizio della fase 2, in modo che i valori dei records recuperati possano essere sottoposti al confronto con gli intervalli di accettazione.

Per quanto riguarda il trattamento delle riparazioni, occorre tener conto del fatto che nel DAU il campo relativo al *regime doganale* risulta compilato più accuratamente del campo relativo alla *natura della transazione*. Pertanto, si sta procedendo in modo da:

- correggere automaticamente la natura della transazione, se la transazione è di tipo definitivo;
- lasciare invariata la lavorazione delle transazioni temporanee di valore superiore a 500 mila euro relative ai capitoli 84-90 (per cui i revisori già effettuano controlli puntuali);
- correggere automaticamente la natura della transazione a seconda delle caratteristiche merceologiche del bene scambiato per le transazioni temporanee residue di valore inferiore a 500 mila euro o relative a capitoli diversi da 84-90.

Infine, in futuro interventi ulteriori dovranno riguardare:

i. l'esclusione dei "bassi valori" solo dopo che questi siano stati sottoposti a revisione nella fase 2; infatti, allo stato attuale potrebbe accadere che alcuni records siano erroneamente esclusi dalla lavorazione dei dati, in quanto relativi a falsi "bassi valori";
ii. la sostituzione delle soglie di valore e di quantità - introdotte per contenere il numero di transazioni da sottoporre a revisione nella fase 2 (si veda la nota 16) - con un criterio di tipo selettivo basato sull'utilizzo di una misura di impatto dell'errore potenziale di ciascuna transazione sui valori oggetto di diffusione. Rispetto al sistema attuale di selezione delle segnalazioni, che sottopone a revisione solo le segnalazioni caratterizzate da valori eccessivamente alti, tale criterio consentirebbe di controllare anche quelle caratterizzate da valori eccessivamente bassi.

4. Determinazione automatica dei parametri di segnalazione

4.1 Concetti introduttivi

In questo paragrafo viene descritto il procedimento introdotto a partire da gennaio 2007 per la determinazione automatica dei parametri di segnalazione relativi al prezzo medio.

Gli intervalli di accettazione per la variabile *prezzo medio* possono essere definiti sulla base di una *funzione sospetto*, il cui calcolo richiede la determinazione di indici di posizione (quartili) della distribuzione dei valori assunti dalla stessa variabile; le transazioni i cui valori cadono al di fuori dell'intervallo di accettazione sono considerate "segnalate" e quindi da sottoporre a revisione manuale (Fescina *et al.*, 2004; Garcia *et al.*, 2006).

Per quanto riguarda le variabili legate alle unità supplementari (*peso medio* e *prezzo unitario*), la determinazione di intervalli di accettazione mediante criteri statistici risulta, invece, particolarmente problematica, a causa della bassa qualità dei dati grezzi relativi alla quantità espressa in unità supplementari.

---

dello scambio (indicata dal *regime doganale*) e dalla motivazione per cui la merce viene movimentata (indicata dalla *natura della transazione*). Al momento, per ragioni di efficienza e tempestività, i revisori effettuano controlli solo sulle transazioni temporanee di valore superiore a 500 mila euro relative a quei capitoli le cui merci possono essere più frequentemente soggette a riparazione (capitoli 84-90, relativi a macchine ed apparecchi meccanici, elettrici, ottici e a mezzi di trasporto).

Il calcolo degli indici di posizione della distribuzione dei valori del *prezzo medio* è effettuato sui microdati ad un certo stadio della loro lavorazione, ovvero subito dopo il loro caricamento nell'ambiente di produzione Oracle. Tali microdati possono essere considerati "grezzi" dal momento che, pur essendo stati sottoposti ad alcuni trattamenti preliminari (ad esempio, esclusione di records non rilevanti o non desiderati, esclusione di records incompatibili), non hanno ancora subìto correzioni da parte dei revisori.
Occorre osservare che, in generale, l'individuazione delle unità statistiche "anomale" è tanto più efficace quanto più il calcolo della funzione sospetto viene effettuato su un dominio della popolazione per cui il comportamento delle unità rispetto alla variabile oggetto di revisione sia omogeneo. Al momento, come già anticipato, i domini di riferimento utilizzati per il calcolo della funzione sospetto sono individuati dall'incrocio delle modalità relative alle variabili *tipo di movimento* e *codice merceologico NC8*.
Inoltre, una decisione fondamentale - in conflitto con quella relativa all'individuazione di gruppi di unità dal comportamento omogeneo - riguarda il numero di osservazioni da prendere in considerazione per il calcolo della funzione sospetto su un dato dominio. Ciascuna tipologia di merce è caratterizzata da un numero di transazioni commerciali molto variabile da mese a mese. L'uso dei soli microdati del mese corrente di lavorazione non è in generale sufficiente a garantire stime accurate dei parametri di segnalazione, anche utilizzando domini di riferimento più ampi, definiti, ad esempio, ad un livello di classificazione merceologica più aggregato di quello della Nomenclatura Combinata. Questo problema rende necessario il ricorso a dati riferiti ai mesi precedenti.
In base alle sperimentazioni effettuate, si ritiene che l'utilizzazione degli ultimi 2 anni di osservazioni mensili possa garantire stime sufficientemente accurate dei parametri di segnalazione per gran parte dei domini di riferimento. Tuttavia, poiché la memorizzazione dei microdati grezzi è stata avviata soltanto a partire da gennaio 2006, fino a novembre 2007 il calcolo della funzione sospetto verrà effettuato su un insieme di dati incompleto (nel momento in cui si scrive si dispone soltanto di 16 mesi di osservazioni).
Nel sottoparagrafo seguente è descritto l'algoritmo di calcolo dei parametri di segnalazione per un generico dominio di riferimento.

4.2 Calcolo dei parametri di segnalazione per un dominio di riferimento

Il procedimento di segnalazione qui descritto si basa sull'assunzione che la maggior parte dei valori della variabile oggetto di revisione non presenti errori. Questa assunzione consente di definire, sulla base della distribuzione di tutti i valori assunti dalla variabile per un certo dominio di riferimento, un intervallo al di fuori del quale ci si aspetta che cadano i valori anomali relativi ad osservazioni appartenenti a quel dominio.
Va premesso che sono esclusi dal calcolo automatico dei parametri di segnalazione per il prezzo medio, alcuni domini di riferimento che si riferiscono a tipologie merceologiche le cui caratteristiche richiedono una revisione accurata di tutte le osservazioni[156]; tali domini sono individuati dall'incrocio delle seguenti modalità relative alle variabili *tipo di movimento* e *codice merceologico NC8*:

([156]) Come anticipato nel paragrafo 2.3, si tratta di prodotti che hanno un peso rilevante sulla bilancia commerciale, che presentano delle specificità da tenere sotto controllo o sui quali sono frequenti i rilievi da parte degli utenti.

tipo di merce	*tipo di movimento*	*codice merceologico NC8*
caffè	importazione	09011100
petrolio greggio	importazione, esportazione	27090090
metano	importazione, esportazione	27112100
energia elettrica	importazione, esportazione	27160000
diamanti	importazione, esportazione	71022900, 71023900
perle	importazione, esportazione	71011000, 71012100, 71012200
aerei	importazione, esportazione	tutti quelli afferenti al capitolo 88
navi	importazione, esportazione	tutti quelli afferenti al capitolo 89
opere d'arte	importazione, esportazione	tutti quelli afferenti al capitolo 97

Sia PM_i il *prezzo medio* - espresso in euro per chilogrammo - relativo alla transazione *i* nel mese corrente di riferimento dei dati:

$$PM_i = \frac{valore_i}{quantità_i} \qquad quantità_i \neq 0;$$

siano, inoltre, $PM_{Q1}(i)$ e $PM_{Q3}(i)$ rispettivamente il primo ed il terzo quartile della distribuzione dei valori del *prezzo medio* nel dominio a cui appartiene la transazione *i*; tale distribuzione è calcolata a partire dai microdati grezzi del mese corrente di riferimento e degli ultimi 23 mesi immediatamente precedenti.
Una migliore individuazione dell'intervallo di accettazione si ottiene prendendo il logaritmo naturale dei valori del *prezzo medio*; questa trasformazione consente di rendere simmetrica la loro distribuzione, tipicamente caratterizzata da asimmetria positiva all'interno di ciascun dominio.
Con riferimento ad una generica transazione *i* (la cui quantità espressa in chilogrammi è diversa da zero), il valore della funzione sospetto è dato da:

$$sospetto_i = \begin{cases} \dfrac{log(PM_{Q1}(i)) - log(PM_i)}{log(PM_{Q3}(i)) - log(PM_{Q1}(i))} & \text{se} \quad PM_i < PM_{Q1}(i) \\ 0 & \text{se} \quad PM_{Q1}(i) \leq PM_i \leq PM_{Q3}(i) \\ \dfrac{log(PM_i) - log(PM_{Q3}(i))}{log(PM_{Q3}(i)) - log(PM_{Q1}(i))} & \text{se} \quad PM_i > PM_{Q3}(i). \end{cases} \qquad [1]$$

Le osservazioni del dominio per cui il valore della funzione sospetto è superiore ad una certa costante *C* (maggiore di 0) sono segnalate per essere sottoposte a revisione manuale.
In altre parole, una transazione commerciale è considerata "sospetta" se la distanza (in termini logaritmici) tra il *prezzo medio* osservato e il più vicino tra il primo e il terzo quartile è superiore a *C* volte la distanza interquartilica.
Oltre alle osservazioni "sospette", sono sottoposte a revisione manuale tutte le transazioni per cui la quantità espressa in chilogrammi risulta uguale a zero.
I parametri di segnalazione $PM_{min}(i)$ e $PM_{max}(i)$ ($PM_{min}(i) \leq PM_{max}(i)$), ovvero i valori di *prezzo medio* delimitanti l'intervallo al di fuori del quale le osservazioni del dominio sono considerate segnalate, possono essere facilmente ricavati in funzione della costante *C*:

$$PM_{min}(i) = exp\left(-C \times log\left(PM_{Q3}(i)\right) + (1+C) \times log\left(PM_{Q1}(i)\right)\right)$$
$$PM_{max}(i) = exp\left(-C \times log\left(PM_{Q1}(i)\right) + (1+C) \times log\left(PM_{Q3}(i)\right)\right). \quad [2]$$

Occorre osservare che, pur utilizzando gli ultimi 2 anni di osservazioni, non è detto che sia possibile garantire stime sufficientemente accurate dei parametri di segnalazione per tutti i domini di riferimento[157]. Pertanto, sono esclusi dal calcolo automatico dei parametri tutti i domini caratterizzati da meno di 20 transazioni commerciali nell'arco temporale degli ultimi 24 mesi[158]. Per tali domini i parametri sono stabiliti in modo ragionato dai revisori dei dati Extrastat ed aggiornati periodicamente.

5. Risultati principali delle sperimentazioni effettuate

In questo paragrafo sono riportati i risultati principali delle sperimentazioni effettuate, finalizzate alla valutazione dell'efficacia dei parametri determinati automaticamente. Alcuni accorgimenti adottati per superare i problemi indotti dall'aggiornamento annuale della Nomenclatura Combinata sono discussi in Appendice.

In una situazione ottimale, per poter valutare la capacità dei parametri di segnalare effettivamente i valori anomali bisognerebbe disporre di un insieme di dati "esatti", ovvero completamente esenti da errori. A causa della complessità della rilevazione Extrastat, è molto difficile disporre di un tale insieme di dati, anche ricorrendo ad un'indagine campionaria di controllo, in cui si ponga estrema cura in tutte le fasi di lavorazione dei dati; d'altra parte, i dati "puliti" - ottenuti alla fine del processo di controllo e correzione - non possono essere ritenuti completamente esatti, per la possibile presenza di errori non segnalati oppure ritenuti accettabili dai revisori.

I parametri di segnalazione utilizzati fino a dicembre 2006 erano stati stabiliti in modo ragionato circa due anni prima e, all'inizio del 2007, avrebbero richiesto un aggiornamento per gran parte dei domini di riferimento, con notevole dispendio in termini di tempo e di risorse umane. A partire da gennaio 2007, contestualmente all'adeguamento del processo di produzione dei dati alle novità legate all'ingresso della Romania e della Bulgaria nell'Unione Europea e all'aggiornamento annuale della Nomenclatura Combinata, è stata introdotta la procedura per il calcolo dei parametri descritta nel paragrafo 4.

Purtroppo, l'elevato ammontare delle transazioni lavorate ogni mese (circa 800.000), congiunto all'esiguo numero di persone impiegate nella revisione dei dati (10 revisori), non hanno consentito la sperimentazione in parallelo dei due criteri di segnalazione nel processo di lavorazione dei dati. Pertanto, i risultati qui presentati si limitano ad alcune considerazioni basate sul "confronto" tra gli intervalli di accettazione "vecchi" - individuati in modo ragionato e utilizzati fino a dicembre 2006 - e quelli "nuovi" - determinati in base alla distribuzione dei microdati grezzi degli ultimi mesi e utilizzati a partire da gennaio 2007. Ovviamente, tale confronto è possibile per i soli domini i cui intervalli di accettazione sono calcolabili statisticamente, vale a dire per i domini caratterizzati da almeno 20 osservazioni nel periodo di riferimento considerato (gennaio 2006 - aprile 2007).

([157]) D'altra parte, l'utilizzo di ulteriori dati ancora più lontani nel tempo dal mese corrente di riferimento, può avere anche degli inconvenienti; ad esempio, in caso di aumento dei prezzi di una certa merce nel mese corrente, i parametri di segnalazione potrebbero rivelarsi più facilmente inadeguati, conducendo ad un numero eccessivamente elevato di false segnalazioni.

([158]) Dal momento che la memorizzazione dei microdati grezzi è stata avviata soltanto a partire da gennaio 2006, fino a novembre 2007 la condizione relativa alle 20 transazioni commerciali sarà verificata su un arco temporale minore di 24 mesi. Inoltre, poiché a gennaio 2007 la Romania e la Bulgaria sono entrate a far parte dell'Unione Europea, per il 2007 viene adottato il criterio di escludere le osservazioni del 2006 riferite a questi due Paesi da tutti i domini caratterizzati da almeno 20 transazioni rispetto ad altri Paesi.

Per quanto riguarda le importazioni, questi domini sono 6.291, approssimativamente il 64 per cento del numero complessivo[159] (9.827); con riferimento alle esportazioni, i domini i cui parametri di segnalazione sono stimabili in modo sufficientemente accurato rappresentano circa il 69 per cento del totale.
Si ha ragione di credere che, utilizzando 24 mesi di osservazioni, sia possibile calcolare gli intervalli in modo automatico per la quasi totalità dei domini di riferimento.
Nei sottoparagrafi successivi le due tipologie di intervallo di interesse sono confrontate in termini di numero di segnalazioni (sottoparagrafo 5.1) e in termini di posizione reciproca e lunghezza (sottoparagrafo 5.2).

5.1 Confronto tra gli intervalli di accettazione in termini di numero di segnalazioni

Per un dato dominio, le due tipologie di intervallo di interesse possono essere confrontate in termini di numero di segnalazioni.
Il numero di segnalazioni costituisce un'importante misura del carico di lavoro dei revisori e può essere preso utilmente in considerazione come criterio per la calibrazione dei parametri determinati statisticamente. Nella scelta della costante *C* in [2], l'interesse è rivolto essenzialmente a ridurre l'incidenza degli intervalli dalle caratteristiche poco desiderabili, cioè gli intervalli troppo stretti - che danno luogo ad un numero eccessivamente elevato di segnalazioni - e gli intervalli troppo ampi - che comportano l'accettazione di tutte le osservazioni.
Tuttavia bisogna tenere conto che, all'aumentare del valore di *C*, diminuisce la frequenza dei casi caratterizzati da intervalli stretti, ma aumenta la frequenza delle situazioni senza alcuna segnalazione.
I risultati riportati in questo sottoparagrafo sono ottenuti facendo riferimento al mese di aprile 2007 che può essere ritenuto rappresentativo del numero di segnalazioni negli altri mesi dell'anno.
Le tavole 3 e 4 riportano la distribuzione di frequenza dei codici *NC8* secondo la percentuale di segnalazioni nel mese di aprile 2007, in base all'intervallo stabilito in modo ragionato e in base all'intervallo determinato statisticamente.
In particolare, la tavola 3 si riferisce alle importazioni, la tavola 4 alle esportazioni. La percentuale di segnalazioni per un certo dominio è calcolata rispetto al totale delle transazioni del dominio stesso nel mese di aprile 2007. Le segnalazioni in base all'intervallo determinato statisticamente sono ottenute utilizzando tre differenti valori di *C*: 0,4, 0,5 e 0,6.
Il valore "ottimale" di *C* è quello che, rispetto alle segnalazioni ottenute in base ai "vecchi" intervalli, consente di:

- ridurre la frequenza dei codici *NC8* caratterizzati da un numero di segnalazioni troppo elevato (superiore al 50 per cento delle osservazioni: "*50% ┤ 100%*") o abbastanza elevato (superiore al 25 e inferiore o uguale al 50 per cento delle osservazioni: "*25% ┤ 50%*");
- aumentare la frequenza dei codici *NC8* caratterizzati da un numero di segnalazioni considerato "ideale" (superiore a 0 e inferiore o uguale al 25 per cento delle osservazioni: "*0% ┤ 25%*");
- contenere l'aumento della frequenza dei codici *NC8* senza alcuna segnalazione ("*0%*").

Con riferimento alla tavola 3, in base ai "vecchi" intervalli, ben 1.696 codici *NC8* presentano un numero di segnalazioni superiore al 50 per cento delle osservazioni.

([159]) Con riferimento ad un solo *tipo di movimento* (importazione o esportazione), il numero complessivo di domini è dato dal numero dei raggruppamenti merceologici in posizioni ad otto cifre previsti dalla Nomenclatura Combinata. Nel 2007 la Nomenclatura Combinata prevede 9.827 codici *NC8*. I codici *NC8* relativi ai domini esclusi dal calcolo automatico dei parametri di segnalazione - in quanto si riferiscono a merci le cui caratteristiche richiedono una revisione di tutte le osservazioni - costituiscono meno dell'1 per cento del totale.

Per ciascun valore di *C*, i "nuovi" intervalli consentono una riduzione notevole di questa frequenza - 398, 312 e 262 per *C* = 0,4, 0,5 e 0,6 - e un aumento dell'incidenza dei casi con un numero di segnalazioni considerato "ideale" ("*0% ┤ 25%*").
Tuttavia, la riduzione della frequenza dei codici *NC8* con un numero di segnalazioni compreso tra il 25 e il 50 per cento delle osservazioni si ottiene solo per *C* pari a 0,5 e 0,6 (rispettivamente, 1.210 e 930 codici *NC8* contro i 1.341 risultanti in base ai "vecchi" intervalli). Inoltre, poiché per *C* = 0,6 i casi per cui non si verificano segnalazioni aumentano in misura non trascurabile, il valore 0,5 può essere considerato "ottimale" e viene utilizzato per il calcolo dei parametri $PM_{min}(i)$ e $PM_{max}(i)$.
Considerazioni analoghe valgono con riferimento alle esportazioni (tavola 4).

***Tavola 3*: Distribuzione di frequenza dei codici merceologici *NC8* (con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007) secondo la percentuale di segnalazioni del mese di aprile 2007, in base all'intervallo stabilito in modo ragionato e in base all'intervallo determinato statisticamente ($C = 0{,}4$, $C = 0{,}5$, $C = 0{,}6$)**
Valori assoluti - Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Importazioni

percentuale di segnalazioni in base all'intervallo stabilito in modo ragionato	codici merceologici *NC8*: percentuale di segnalazioni in base all'intervallo determinato statisticamente				totale
	0%	*0% ┤ 25%*	*25% ┤ 50%*	*50% ┤ 100%*	
		C = 0,4 *C = 0,5* *C = 0,6*			
0%	754 826 885	316 316 305	243 207 174	134 98 83	**1.447**
0% ┤ 25%	96 130 172	851 956 1.019	404 273 169	31 23 22	**1.382**
25% ┤ 50%	195 248 296	589 670 704	506 385 311	51 38 30	**1.341**
50% ┤ 100%	392 457 521	667 741 772	455 345 276	182 153 127	**1.696**
totale	**1.437** **1.661** **1.874**	**2.423** **2.683** **2.800**	**1.608** **1.210** **930**	**398** **312** **262**	**5.866** (*)

(*) 5.866 non coincide con il numero totale di codici *NC8* con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007 (6.291), in quanto per 425 di tali codici non si sono verificate importazioni nel mese di aprile 2007.

***Tavola 4*: Distribuzione di frequenza dei codici merceologici *NC8* (con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007) secondo la percentuale di segnalazioni del mese di aprile 2007, in base all'intervallo stabilito in modo ragionato e in base all'intervallo determinato statisticamente ($C = 0{,}4$, $C = 0{,}5$, $C = 0{,}6$)**

Valori assoluti - Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Esportazioni

percentuale di segnalazioni in base all'intervallo stabilito in modo ragionato	codici merceologici *NC8*: percentuale di segnalazioni in base all'intervallo determinato statisticamente				totale
	0%	*0% ┤ 25%*	*25% ┤ 50%*	*50% ┤ 100%*	
		C = 0,4 *C = 0,5* *C = 0,6*			
0%	527 591 619	206 187 197	141 115 98	105 86 65	**979**
0% ┤ 25%	90 122 157	1.031 1.172 1.256	505 346 236	42 28 19	**1.668**
25% ┤ 50%	179 230 281	722 855 940	684 527 408	74 47 30	**1.659**
50% ┤ 100%	430 496 572	814 965 1.028	637 454 338	233 199 176	**2.114**
totale	**1.226** **1.439** **1.629**	**2.773** **3.179** **3.421**	**1.967** **1.442** **1.080**	**454** **360** **290**	**6.420** (*)

(*) 6.420 non coincide con il numero totale di codici *NC8* con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007 (6.811), in quanto per 391 di tali codici non si sono verificate esportazioni nel mese di aprile 2007.

La tavola 5 riporta il numero di segnalazioni (in termini assoluti e percentuali) nel mese di aprile 2007 secondo il *tipo di intervallo* e il *tipo di movimento*, avendo posto in [2] la costante *C* pari a 0,5.

La percentuale di segnalazioni è calcolata rispetto al totale delle osservazioni nel mese di aprile 2007, per *tipo di movimento*. Per quanto riguarda la tipologia di intervallo, oltre al "vecchio" (stabilito in modo ragionato) e al "nuovo" (determinato statisticamente), viene presa in considerazione anche l'intersezione dei due.

Come si può vedere, per entrambi i tipi di movimento, l'adozione del criterio statistico consente di ridurre di quasi la metà il numero di segnalazioni che si ottengono in base al criterio ragionato. Inoltre, quasi due terzi delle segnalazioni in base al criterio statistico sono in comune con il criterio ragionato.

***Tavola 5*: Segnalazioni nel mese di aprile 2007 secondo la tipologia di intervallo, per tipo di movimento ($C = 0,5$)**

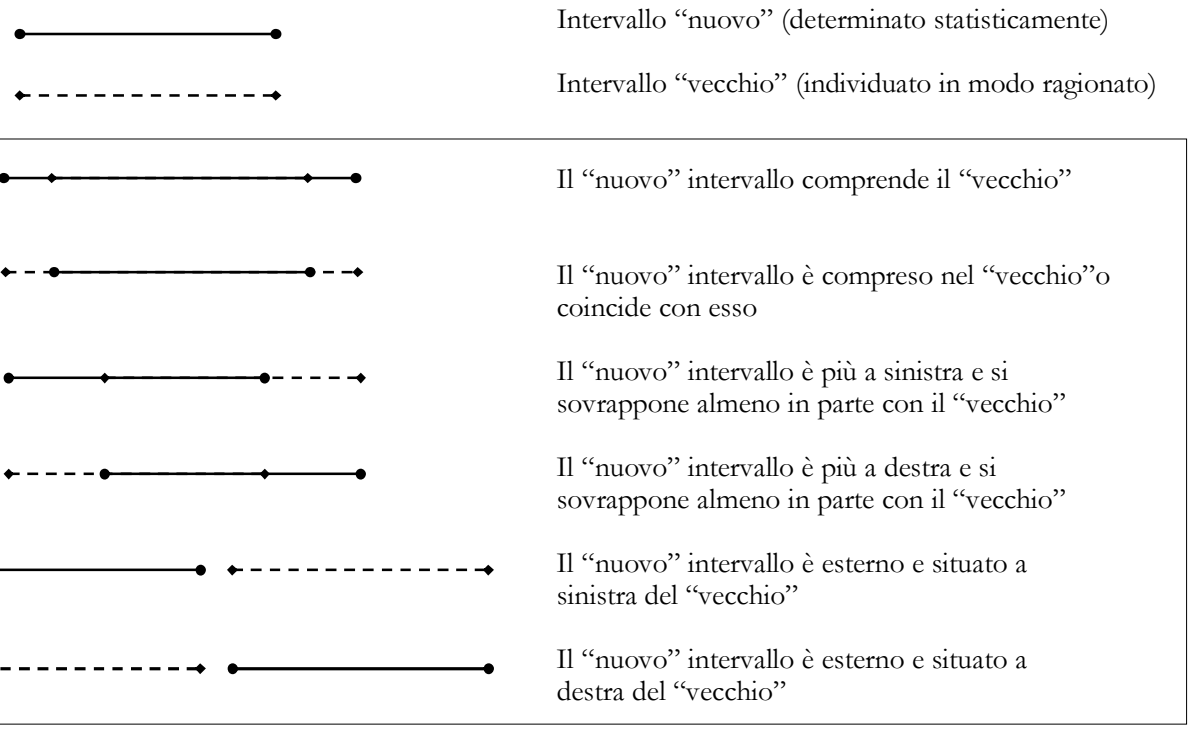
Valori assoluti e percentuali - Dati grezzi Extrastat (gennaio 2006 - aprile 2007)

tipo di intervallo	segnalazioni			
	importazione		esportazione	
	valori assoluti	%	valori assoluti	%
intervallo stabilito in modo ragionato	72.136	32,8	150.873	34,8
intervallo determinato statisticamente	41.023	18,6	82.498	19,0
intersezione tra l'intervallo stabilito in modo ragionato e l'intervallo determinato statisticamente	25.867	11,7	49.918	11,5

5.2 Confronto tra gli intervalli di accettazione in termini di posizione reciproca e lunghezza

Oltre al numero di segnalazioni, per un dato dominio le due tipologie di intervallo di interesse possono essere confrontate in termini di posizione reciproca e lunghezza.
Per quanto riguarda la posizione, l'intervallo determinato statisticamente si situa rispetto a quello individuato in modo ragionato secondo una delle modalità illustrate in figura 1.

***Figura 1*: Posizioni dell'intervallo determinato statisticamente rispetto a quello individuato in modo ragionato, per un dato dominio**



Nella tavola 6 è riportata la distribuzione di frequenza (assoluta e percentuale), per *tipo di movimento*, dei codici merceologici *NC8* secondo il *tipo di posizione del "nuovo" intervallo rispetto al "vecchio"*, avendo posto in [2] la costante *C* pari a 0,5.
Come si può osservare, i casi in cui l'intervallo "nuovo" è compreso nel "vecchio" - 30,0 per cento per le importazioni e 22,2 per cento per le esportazioni - superano quelli in cui il "nuovo" comprende il "vecchio" (21,5 e 13,0 per cento).
Inoltre, i parametri di segnalazione stabiliti dai revisori risultano spesso poco aggiornati in senso positivo o comunque sottostimati, e ciò sembra valere soprattutto per le esportazioni. Infatti, i codici *NC8* per cui l'intervallo determinato statisticamente è situato più a destra rispetto all'intervallo stabilito in modo ragionato (*tipo di posizione "4"* e *"6"*) costituiscono il 35,8 e il 58,8 per cento del totale (rispettivamente, per le importazioni e per le esportazioni); in particolare, nel 2,2 e nel 3,3 per cento dei casi, il "nuovo" intervallo, oltre a essere situato a destra, risulta addirittura esterno al "vecchio".
Scendendo nel dettaglio dei capitoli[160], con riferimento alle importazioni, il posizionamento a destra del "nuovo" intervallo rispetto al "vecchio" si verifica per numerosi codici *NC8* afferenti ai capitoli 84 *"Reattori nucleari, caldaie, macchine, apparecchi e congegni meccanici; parti di queste macchine o apparecchi"* e 85 *"Macchine, apparecchi e materiale elettrico e loro parti; apparecchi per la registrazione o la riproduzione del suono, apparecchi per la registrazione o la riproduzione delle immagini e del suono per la televisione, e parti ed accessori di questi apparecchi"*. Con riferimento alle esportazioni, il posizionamento a destra si osserva per molti codici *NC8* afferenti, oltre che ai capitoli 84 e 85, anche ai capitoli 61 *"Indumenti ed accessori di abbigliamento, a maglia"*, 62 *"Indumenti ed accessori di abbigliamento, diversi da quelli a maglia"*, 73 *"Lavori di ghisa, ferro o acciaio"* e 90 *"Strumenti ed apparecchi di ottica, per fotografia e per cinematografia, di misura, di controllo o di precisione; strumenti ed apparecchi medico-chirurgici; parti ed accessori di questi strumenti o apparecchi"*.

***Tavola 6*: Distribuzione di frequenza dei codici merceologici *NC8* (con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007) secondo il tipo di posizione del "nuovo" intervallo rispetto al "vecchio", per tipo di movimento (*C* = 0,5)**
Valori assoluti e percentuali - Dati grezzi Extrastat (gennaio 2006 - aprile 2007)

tipo di posizione del "nuovo" intervallo rispetto al "vecchio"	codici merceologici *NC8*			
	importazione		esportazione	
	valori assoluti	%	valori assoluti	%
1	1.354	21,5	886	13,0
2	1.885	30,0	1.510	22,2
3	625	9,9	300	4,4
4	2.112	33,6	3.778	55,5
5	41	0,6	47	0,7
6	138	2,2	228	3,3
n.d. (*)	136	2,2	62	0,9
totale	**6.291**	**100,0**	**6.811**	**100,0**

(*) "n.d." sta per "non definito" e indica i casi in cui, non essendo stato stabilito dai revisori un intervallo in modo ragionato, non esiste il termine di paragone per l'intervallo determinato statisticamente.

(160) Per brevità, non viene riportata la distribuzione doppia di frequenza dei codici *NC8* secondo il *tipo di posizione del "nuovo" intervallo rispetto al "vecchio"* e il *capitolo* del Sistema Armonizzato a 2 cifre.

Per un dato dominio, accanto alla posizione può essere preso in considerazione il rapporto $R(i)$ tra la lunghezza dell'intervallo determinato statisticamente e la lunghezza dell'intervallo individuato in modo ragionato:

$$R(i) = \frac{PM_{max}(i) - PM_{min}(i)}{PM^{r}_{max} - PM^{r}_{min}},$$

avendo indicato con PM^{r}_{max} e PM^{r}_{min} ($PM^{r}_{min} \leq PM^{r}_{max}$) i valori di *prezzo medio* delimitanti l'intervallo stabiliti dai revisori.

Le tavole 7 e 8 riportano la distribuzione di frequenza dei codici *NC8* secondo il rapporto $R(i)$ e il *tipo di posizione del "nuovo" intervallo rispetto al "vecchio"*, avendo posto in [2] la costante *C* pari a 0,5. In particolare, la tavola 7 si riferisce alle importazioni, la tavola 8 alle esportazioni.

La modalità VI (colonna 7) indica la classe dei valori di $R(i)$ per cui le due tipologie di intervallo differiscono in misura minore, in quanto un intervallo è al più doppio dell'altro. Le modalità I-V (VII-XI) indicano, invece, le classi dei valori di $R(i)$ per cui l'intervallo determinato statisticamente è più che dimezzato (raddoppiato) rispetto all'intervallo stabilito in modo ragionato.

Guardando l'ultima riga della tavola 7, complessivamente i "nuovi" intervalli risultano più spesso di dimensione maggiore rispetto ai "vecchi" (le frequenze corrispondenti alle modalità VII-XI sono più elevate delle frequenze corrispondenti alle modalità I-V). Tale fenomeno risulta più accentuato per le esportazioni (tavola 8).

Più precisamente, in quasi la metà dei casi per cui si verifica il posizionamento a sinistra del "nuovo" intervallo rispetto al "vecchio" (*tipo di posizione* "*3*" e "*5*"), l'intervallo determinato statisticamente risulta più che dimezzato rispetto a quello stabilito in modo ragionato.

***Tavola 7*: Distribuzione di frequenza dei codici merceologici *NC8* (con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007) secondo il rapporto *R(i)* e il tipo di posizione del "nuovo" intervallo rispetto al "vecchio" ($C = 0{,}5$)**

Valori assoluti - Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Importazioni

tipo di posizione del "nuovo" intervallo rispetto al "vecchio"	codici merceologici *NC8* — *R(i)*														
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**= 0** (*)	**< 0** (**)	**v. m.** (***)	**totale**
1	0	0	0	0	0	331	651	235	55	47	22	0	8	5	**1.354**
2	0	10	15	288	1.069	484	0	0	0	0	0	19	0	0	**1.885**
3	1	1	1	24	261	335	2	0	0	0	0	0	0	0	**625**
4	0	0	0	0	43	854	890	262	34	23	6	0	0	0	**2.112**
5	0	4	4	17	7	3	0	0	0	0	0	1	4	1	**41**
6	0	0	0	0	23	28	31	31	3	10	5	0	5	2	**138**
n. d.	0	0	0	0	0	0	0	0	0	0	0	0	0	136	**136**
totale	**1**	**15**	**20**	**329**	**1.403**	**2.035**	**1.574**	**528**	**92**	**80**	**33**	**20**	**17**	**144**	**6.291**

(*) *R(i) = 0* quando il numeratore è nullo, ovvero quando i valori di *prezzo medio* delimitanti l'intervallo determinato statisticamente coincidono.

(**) *R(i) < 0* quando il denominatore è negativo, ovvero quando per errore[161] $PM^{r}_{min} > PM^{r}_{max}$.

(***) "v. m." sta per "valore mancante"; ciò si verifica quando il "vecchio" intervallo non è definito oppure quando il denominatore di *R(i)* è nullo.

Classi di valori di R(i):

I. $R(i) < \frac{1}{500}$

II. $\frac{1}{500} \leq R(i) < \frac{1}{100}$

III. $\frac{1}{100} \leq R(i) < \frac{1}{50}$

IV. $\frac{1}{50} \leq R(i) < \frac{1}{10}$

V. $\frac{1}{10} \leq R(i) < \frac{1}{2}$

VI. $\frac{1}{2} \leq R(i) \leq 2$

VII $2 < R(i) \leq 10$

VIII $10 < R(i) \leq 50$

IX. $50 < R(i) \leq 100$

X. $100 < R(i) \leq 500$

XI. $R(i) > 500$

([161]) Fino a dicembre 2006, i parametri di segnalazione, essendo stabiliti in modo ragionato, venivano aggiornati manualmente, con la possibilità di commettere errori in fase di digitazione.

***Tavola 8*: Distribuzione di frequenza dei codici merceologici NC8 (con almeno 20 osservazioni nel periodo gennaio 2006 - aprile 2007) secondo il rapporto *R(i)* e il tipo di posizione del "nuovo" intervallo rispetto al "vecchio" ($C = 0{,}5$)**

Valori assoluti - Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Esportazioni

tipo di posizione del "nuovo" intervallo rispetto al "vecchio"	codici merceologici *NC8*														
	R(i)														
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**= 0** (*)	**< 0** (**)	**v. m.** (***)	**totale**
1	0	0	0	0	0	185	462	168	27	26	12	0	2	4	**886**
2	2	5	16	103	822	557	0	0	0	0	0	5	0	0	**1.510**
3	1	0	2	25	116	153	2	0	0	0	0	0	0	1	**300**
4	0	0	0	0	18	1.402	1.871	393	45	37	12	0	0	0	**3.778**
5	0	2	4	12	9	6	0	0	0	0	0	3	8	3	**47**
6	0	0	1	1	10	28	54	59	12	16	24	1	7	15	**228**
n. d.	0	0	0	0	0	0	0	0	0	0	0	0	0	62	**62**
totale	**3**	**7**	**23**	**141**	**975**	**2.331**	**2.389**	**620**	**84**	**79**	**48**	**9**	**17**	**85**	**6.811**

(*) *R(i)* = 0 quando il numeratore è nullo, ovvero quando i valori di *prezzo medio* delimitanti l'intervallo determinato statisticamente coincidono.

(**) *R(i)* < *0* quando il denominatore è negativo, ovvero quando per errore[162] $PM^{r}_{min} > PM^{r}_{max}$.

(***) "v. m." sta per "valore mancante"; ciò si verifica quando il "vecchio" intervallo non è definito oppure quando il denominatore di *R(i)* è nullo.

Classi di valori di *R(i)*:

I. $R(i) < \frac{1}{500}$

II. $\frac{1}{500} \leq R(i) < \frac{1}{100}$

III. $\frac{1}{100} \leq R(i) < \frac{1}{50}$

IV. $\frac{1}{50} \leq R(i) < \frac{1}{10}$

V. $\frac{1}{10} \leq R(i) < \frac{1}{2}$

VI. $\frac{1}{2} \leq R(i) \leq 2$

VII $2 < R(i) \leq 10$

VIII $10 < R(i) \leq 50$

IX. $50 < R(i) \leq 100$

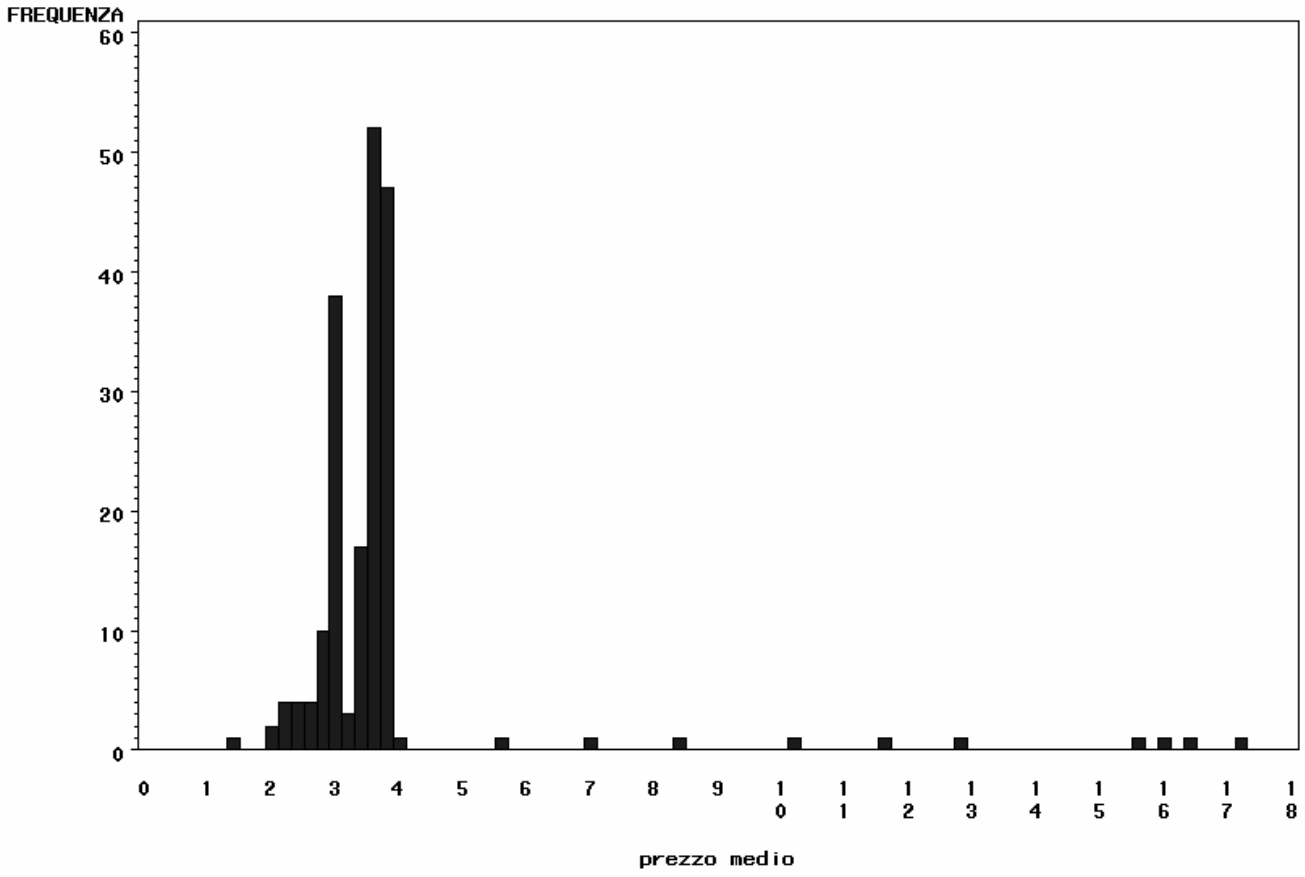
X. $100 < R(i) \leq 500$

XI. $R(i) > 500$

([162]) Fino a dicembre 2006, i parametri di segnalazione, essendo stabiliti in modo ragionato, venivano aggiornati manualmente, con la possibilità di commettere errori in fase di digitazione.

In generale, si tratta di domini al cui interno le osservazioni sono caratterizzate da prezzi medi vicini tra loro; in questi casi la distribuzione delle transazioni secondo il *prezzo medio* presenta una dispersione centrale molto minore di quella espressa dai revisori mediante l'intervallo stabilito in modo ragionato.
Il caso rappresentato in figura 2, relativo alle importazioni di "*Mobili congelatori-conservatori, tipo armadio, di capacità inferiore o uguale a 250 litri*", è molto esplicativo in tal senso. Il rapporto $R(i)$ vale 0,01; l'intervallo di accettazione indicato dai revisori, pari a 16,00-148,00 euro/kg, oltre ad essere non centrato rispetto ai valori assunti dal *prezzo medio* nel periodo gennaio 2006 - aprile 2007 (che vanno da 1,50 a 17,14 euro/kg), risulta palesemente troppo ampio. Più adeguato, invece, è l'intervallo 2,69-4,25 euro/kg, ottenuto in base alla procedura descritta nel paragrafo 4 (avendo posto in [2] la costante *C* pari a 0,5).

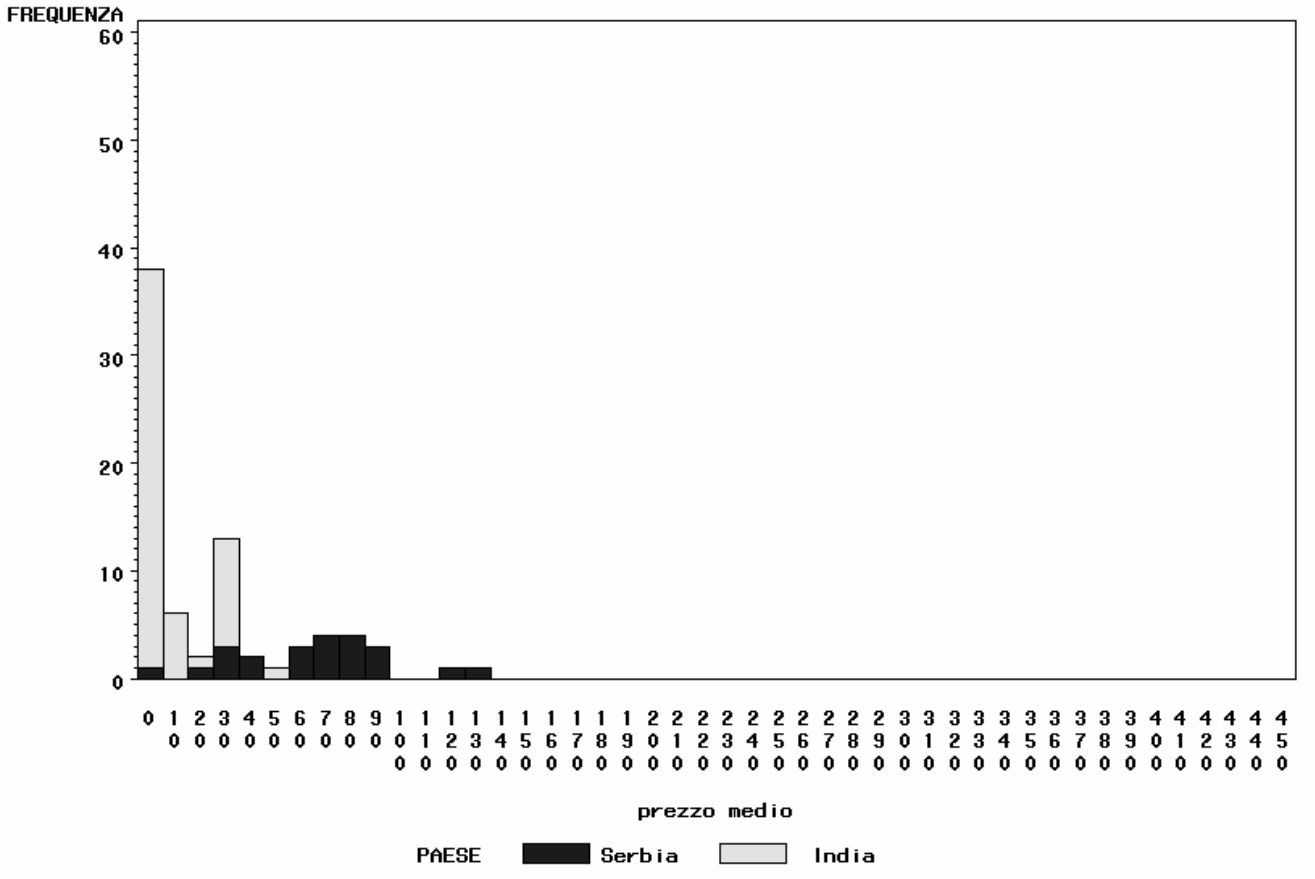
***Figura 2*: Transazioni relative a "Mobili congelatori-conservatori, tipo armadio, di capacità inferiore o uguale a 250 litri" (corrispondente al codice *NC8* "84184020"), secondo il prezzo medio**
Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Importazioni



Al contrario di ciò che avviene in occasione del posizionamento a sinistra, per oltre la metà dei casi in cui il "nuovo" intervallo comprende il "vecchio" (*tipo di posizione* "*1*") oppure si verifica il posizionamento a destra (*tipo di posizione* "*4*" e "*6*"), l'intervallo determinato statisticamente risulta più che raddoppiato rispetto a quello stabilito in modo ragionato; in particolare, in quasi il 2 per cento dei casi, la lunghezza del "nuovo" intervallo è addirittura almeno 100 volte quella del "vecchio" intervallo (modalità X e XI).
In generale, questi casi estremi corrispondono a domini poco omogenei, al cui interno le transazioni si riferiscono a merci caratterizzate da prezzi medi molto diversi. Spesso, tra i fattori osservabili

maggiormente esplicativi della disomogeneità all'interno dei domini rientrano il *Paese statistico* o l'*operatore* (a tale proposito si veda il paragrafo 6).
In questi casi, la distribuzione di frequenza per il *prezzo medio* su un singolo dominio può essere vista come la risultante di una mistura di sotto-distribuzioni, diversamente centrate a seconda del *Paese statistico* o dell'*operatore*. Inoltre, occorre osservare che quanto più i centri di tali sotto-distribuzioni sono distanti, tanto più gli intervalli determinati statisticamente in base alla distribuzione dei dati sull'intero dominio possono risultare meno accurati nella segnalazione delle situazioni di errore.
A titolo esemplificativo, si riporta in figura 3 l'istogramma del *prezzo medio* all'importazione di "*Cuoi e pelli, verniciati o laccati; cuoi e pelli metallizzati*". Nel periodo gennaio 2006 - aprile 2007, questo tipo di merce risulta provenire dall'India o dalla Serbia. Il rapporto $R(i)$ è pari a 105,77. Come si può vedere dalla figura, i prezzi medi praticati nelle transazioni dall'India sono mediamente inferiori a quelli praticati nelle transazioni dalla Serbia. Tuttavia, alla luce dei valori assunti dal *prezzo medio*, l'intervallo indicato dai revisori, pari a 4,00-5,00 euro/kg, è troppo stretto e, comunque, poco adeguato per le importazioni dalla Serbia. L'intervallo determinato in base alla distribuzione dei dati sull'intero dominio risulta senz'altro più appropriato (2,08-107,85 euro/kg); tuttavia, un maggior grado di accuratezza nella segnalazione delle situazioni di errore potrebbe essere raggiunto utilizzando intervalli determinati in modo distinto per i due Paesi (1,53-25,26 euro/kg, per l'India, e 27,68-123,00 euro/kg, per la Serbia).

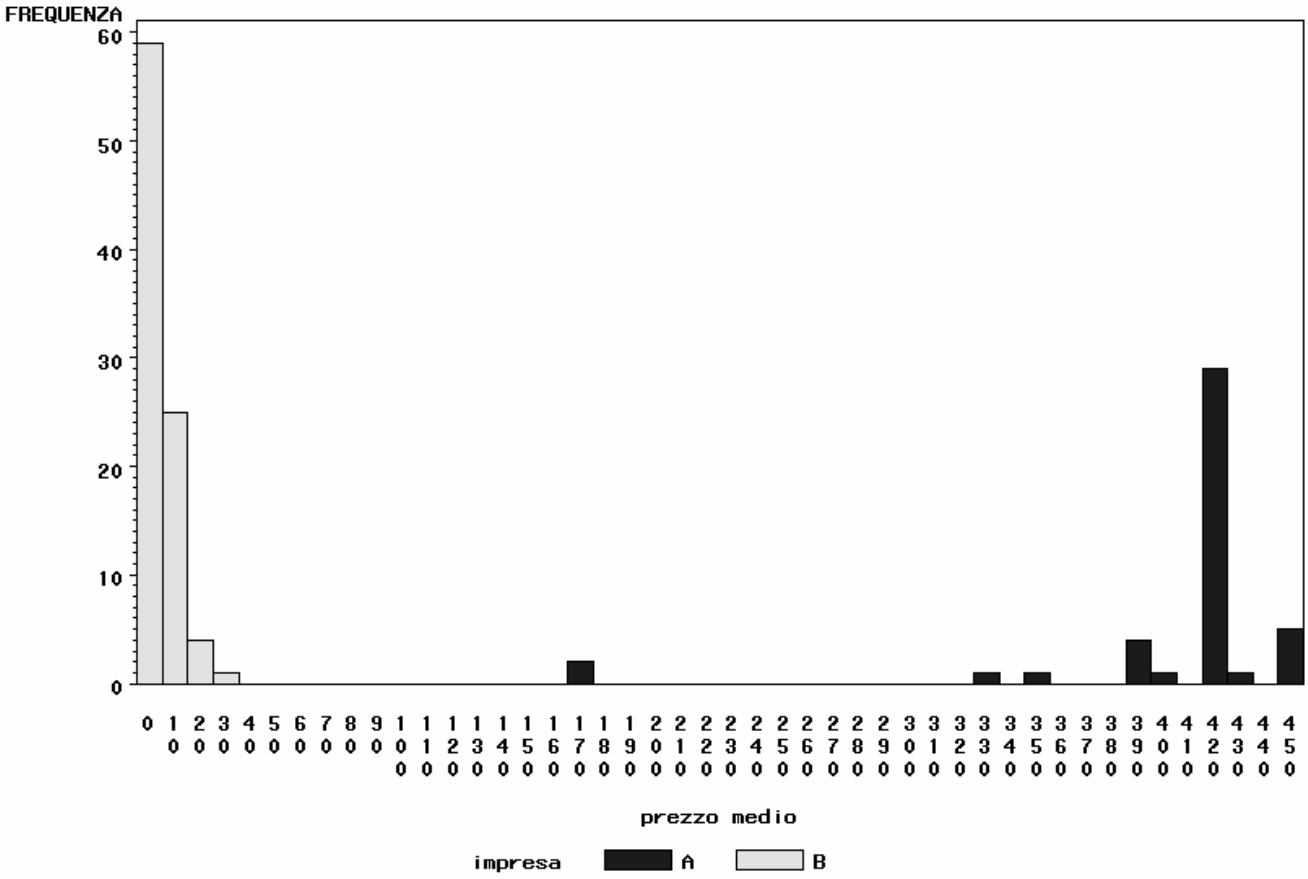
***Figura 3*: Transazioni relative a "Cuoi e pelli, verniciati o laccati; cuoi e pelli metallizzati" (corrispondente al codice *NC8* "41142000"), secondo il prezzo medio**
Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Importazioni



Al fine di illustrare gli inconvenienti causati da forte eterogeneità dei prezzi medi all'interno di uno stesso dominio, viene riportato il caso (fortunatamente abbastanza isolato) relativo alle esportazioni di una merce afferente al capitolo 29 "*Prodotti chimici organici*" (figura 4). Per vincoli di riservatezza, non è possibile rendere noto il codice *NC8* corrispondente. Come si può vedere, nel periodo gennaio 2006 - aprile 2007 le transazioni relative a questa merce sono state effettuate da due soli operatori; i prezzi medi praticati dall'operatore B risultano nettamente inferiori a quelli praticati dall'operatore A. Il rapporto $R(i)$ vale 1.540,02. L'intervallo di accettazione stabilito in modo ragionato dai revisori, pari a 0,20-6,00 euro/kg, oltre ad essere troppo stretto per i prezzi praticati dall'operatore B, comporta la segnalazione di tutte le transazioni relative all'operatore A; d'altro canto, anche l'intervallo determinato statisticamente si rivela del tutto inadeguato, in quanto, a causa della notevole differenza tra i prezzi medi praticati dai due operatori, risulta esageratamente ampio (0,04-8.932,16 euro/kg).
In questo caso, gli intervalli determinati in maniera distinta per i due operatori - rispettivamente, pari a 0,30-17,01 euro/kg e a 419,29-420,72 euro/kg - consentirebbero un guadagno notevole in termini di accuratezza nella segnalazione delle situazioni di errore.

***Figura 4*: Transazioni relative ad una merce del capitolo 29 "Prodotti chimici organici", secondo il prezzo medio**

Dati grezzi Extrastat (gennaio 2006 - aprile 2007) - Esportazioni



6. Individuazione di domini di riferimento specifici per il calcolo dei parametri di segnalazione

In questo paragrafo, viene valutata l'opportunità di utilizzare domini di riferimento specifici per il calcolo dei parametri di segnalazione.
A tal fine può essere utilizzata la tecnica nota con il nome di *analisi della varianza*[163].
A partire dalla partizione delle transazioni commerciali determinata dalla Nomenclatura Combinata per ciascun *tipo di movimento* (importazione o esportazione), si ricercano i domini di riferimento caratterizzati al loro interno da una maggiore omogeneità del fenomeno oggetto di studio (*variabile risposta*).
La variabile risposta presa in esame è costituita dal *logaritmo naturale del prezzo medio*[164].
Le variabili esplicative (o *fattori*) considerate per l'individuazione di domini di riferimento specifici sono:

- il *Paese statistico*;
- l'*operatore commerciale*, identificato dalla partita IVA.

Dal momento che il numero di osservazioni disponibili rappresenta un elemento critico per l'individuazione di domini di riferimento specifici, è prevista una fase preliminare all'analisi della varianza in cui, per ogni tipologia merceologica della Nomenclatura Combinata, vengono determinati:

a) il numero di modalità (o *livelli*) di ciascun fattore;
b) il numero di transazioni commerciali relative ad ogni livello di ciascun fattore.

Sulla base dei conteggi di tipo a) e b), per ogni fattore considerato sono poi selezionate le tipologie merceologiche della Nomenclatura Combinata per cui il numero di livelli e di osservazioni per livello è sufficiente per effettuare l'analisi della varianza[165]; i risultati qui presentati sono ottenuti selezionando, per ogni fattore, i codici merceologici *NC8* per i quali il numero dei livelli con almeno 20 transazioni commerciali è maggiore o uguale a 2.
L'analisi della varianza viene effettuata utilizzando la procedura GLM del SAS[166].
In base ai risultati dell'analisi della varianza, si riesce a stabilire - per ogni combinazione di *tipo di movimento* e di *codice merceologico NC8* per cui il numero di livelli con almeno 20 osservazioni è maggiore o uguale a 2 - se esistono almeno due medie significativamente diverse tra loro fra quelle determinate dai livelli del fattore preso in esame. Successivamente, nei casi in cui i livelli considerati siano più di due, può essere applicato il test di Tukey-Kramer per stabilire a quali livelli si riferiscono le medie significativamente diverse (Tukey, 1949; Kramer, 1956).
Nella tavola 9 sono riportati, per ciascuno dei fattori considerati, il numero di codici merceologici *NC8* per cui:

- il numero dei livelli con almeno 20 transazioni commerciali è maggiore o uguale a 2 (colonna 2);
- è stabilita una differenza significativa tra le medie determinate dai livelli del fattore e il fattore risulta spiegare almeno il 45 per cento della variabilità del *logaritmo naturale del prezzo medio* (colonna 3).

---

(163) Con questa terminologia viene indicato un insieme di procedure di tipo inferenziale che permettono di verificare l'ipotesi di uguaglianza tra le medie di un fenomeno quantitativo (*variabile risposta*) osservato in due o più gruppi di unità statistiche. Tali medie sono determinate dalle diverse modalità (o *livelli*) di una o più variabili di tipo nominale, dette *fattori*. L'ipotesi di uguaglianza tra le medie viene verificata mediante il confronto tra due diverse stime della variabilità del fenomeno (*variabilità tra i gruppi* e *variabilità interna ai gruppi*).
(164) La distribuzione dei valori del *prezzo medio* è generalmente asimmetrica e la trasformazione logaritmica la rende più assimilabile alle realizzazioni di una variabile aleatoria di tipo Normale (la normalità degli errori rientra tra le assunzioni fondamentali per l'applicazione delle tecniche di analisi della varianza; al riguardo si veda, ad esempio, la monografia di Scheffé, 1959).
(165) Dal momento che il numero di osservazioni per dominio rappresenta un elemento critico, viene effettuata l'analisi della varianza considerando un solo fattore alla volta (*analisi della varianza ad una via*).
(166) La procedura ANOVA non è indicata in questo contesto (SAS Institute Inc., 1989), dal momento che generalmente i dati da elaborare non sono bilanciati (ovvero il numero di osservazioni per livello non è costante).

Nella lettura della tavola occorre tenere presente che, nei casi in cui sulla variabile risposta sia verificato un effetto significativo sia del *Paese statistico* sia dell'*operatore*, viene preso in considerazione solo il fattore che spiega una quota maggiore della variabilità del fenomeno.
Come si può vedere dai dati riportati nella tavola (colonna 3), per circa un quarto dei quasi 10.000 codici merceologici *NC8* viene verificato un effetto significativo di almeno uno dei due fattori sulla variabile risposta.
Inoltre, nonostante le tipologie merceologiche sottoposte all'analisi della varianza siano più numerose per il fattore *Paese statistico*, il fattore *operatore* risulta molto più spesso determinante nella spiegazione della variabilità del *logaritmo naturale del prezzo medio*. Ciò vale in modo particolare per le esportazioni.

***Tavola 9*. Codici merceologici *NC8* sottoposti all'analisi della varianza e codici merceologici *NC8* per cui il fattore spiega almeno il 45 per cento della variabilità del logaritmo naturale del prezzo medio, per fattore e tipo di movimento**
Valori assoluti - Dati grezzi Extrastat (gennaio 2006 - aprile 2007)

fattore	codici *NC8* sottoposti all'analisi della varianza	codici *NC8* per cui il fattore spiega almeno il 45% della variabilità del logaritmo del prezzo medio
		importazione
Paese statistico	3.786	552
operatore	3.195	1.810
		esportazione
Paese statistico	4.275	231
operatore	3.792	2.133

7. Considerazioni conclusive

In base ai risultati delle sperimentazioni effettuate, la procedura proposta consente di ridurre di quasi la metà il numero di segnalazioni che si ottengono utilizzando il criterio ragionato (tavola 5), con una conseguente riduzione dei tempi e dei costi (in termini di risorse umane) legati alla revisione. Le osservazioni segnalate, anche se in numero minore, richiedono comunque, come in passato, un lavoro molto accurato di verifica e di eventuale correzione da parte dei revisori.
Un esame approfondito dei motivi di differenza tra i "nuovi" intervalli di accettazione e quelli utilizzati precedentemente, rivela che spesso questi ultimi sono obsoleti o comunque tali da non rispecchiare in modo adeguato la dispersione centrale della distribuzione dei valori della variabile oggetto di revisione.
Tra i vantaggi più rilevanti connessi all'adozione del criterio statistico vanno quindi evidenziati:

- un maggior grado di trasparenza, oggettività ed accuratezza nell'individuazione delle situazioni di errore;
- una riduzione dei tempi e dei costi legati all'aggiornamento dei parametri di segnalazione.

Per quanto riguarda gli aspetti critici della procedura proposta, va sottolineato che il criterio statistico di segnalazione si basa sull'assunzione che la maggior parte dei valori oggetto di revisione non presenti errori.

A tale proposito si ricorda che la stessa procedura, applicata alla variabile *peso medio*, non ha prodotto risultati soddisfacenti, proprio a causa della bassa qualità dei dati grezzi relativi alla quantità espressa in unità supplementari. Perciò, i parametri di segnalazione attualmente utilizzati per il *peso medio* sono definiti dai limiti di peso contenuti nella descrizione merceologica associata al codice *NC8*, quando presenti, oppure sono stabiliti in modo ragionato dai revisori.
Al fine di valutare l'opportunità di utilizzare domini di riferimento specifici per il calcolo dei parametri di segnalazione è stata utilizzata la tecnica di analisi della varianza. I risultati ottenuti evidenziano un effetto significativo del *Paese statistico* o dell'*operatore* sul *logaritmo del prezzo medio* per circa un quarto delle quasi 10.000 tipologie merceologiche previste dalla Nomenclatura Combinata. Inoltre, rispetto al *Paese statistico*, il fattore *operatore* risulta molto più spesso determinante nella spiegazione della variabilità del fenomeno oggetto di studio, soprattutto con riferimento alle esportazioni.
Pertanto, al fine di migliorare l'efficacia del criterio statistico di segnalazione è senz'altro auspicabile l'introduzione di domini definiti per *Paese statistico* o per *operatore*, oltre che per tipo di movimento e tipologia merceologica.
Finora, l'attuazione di questo ulteriore sviluppo del procedimento di segnalazione non è stata possibile essenzialmente per due ordini di problemi.
Innanzitutto, l'individuazione e l'utilizzazione di domini caratterizzati da una maggiore omogeneità interna sono subordinate alla disponibilità di un numero di osservazioni tale da garantire stime sufficientemente accurate dei parametri di segnalazione. Questo elemento è tanto più critico, quanto più i domini adottati per il calcolo dei parametri sono specifici e presentano una forte discontinuità temporale, come quella causata dagli eventi di nascita e cessazione di operatori (o meglio, di partite IVA). Basti pensare che al momento, utilizzando 16 mesi di osservazioni, è possibile fornire stime accurate dei parametri di segnalazione per domini specifici soltanto per circa un terzo (o poco più) delle tipologie merceologiche previste dalla Nomenclatura Combinata (colonna 2 di tavola 9). Si ha ragione di credere che in futuro, con 24 mesi di osservazioni a disposizione, l'impiego di domini specifici potrà essere meno limitato.
In secondo luogo, l'introduzione di domini di riferimento specifici richiede modifiche importanti, con tempi e costi finora non sostenibili, al sistema informativo esistente.
Infine, come ulteriore sviluppo della procedura proposta, di più semplice attuazione, si accenna alla possibilità di concentrare le operazioni di revisione manuale soltanto sulle osservazioni che incidono maggiormente sulle stime finali. Ciò può essere ottenuto introducendo, accanto al grado di sospettosità misurato dalla [1], un criterio selettivo basato sull'utilizzo di una misura di impatto potenziale di ciascuna transazione sui valori totali oggetto di diffusione[167]. Rispetto all'attuale sistema di selezione delle segnalazioni - fondato sull'utilizzo di soglie di valore e di quantità (sottoparagrafo 2.3) - che sottopone a revisione solo le segnalazioni caratterizzate da valori eccessivamente alti, tale criterio

---

([167]) Ad esempio, con riferimento ad una generica transazione *i*, l'impatto potenziale può essere definito nel modo seguente (Jäder e Norberg, 2005):

$$impatto\_potenziale_i = \frac{\left| valore_i - quantità_i \times PM_{Q2}(i) \right|}{\sum_{i \in I} valore_i},$$

avendo indicato con $PM_{Q2}(i)$, la mediana della distribuzione dei valori del *prezzo medio* nel dominio a cui appartiene la transazione *i* (tale distribuzione può essere calcolata a partire dai microdati grezzi degli ultimi 24 mesi); la sommatoria al denominatore si intende estesa a tutte le transazioni del dominio a cui appartiene *i* negli ultimi 24 mesi.
È importante notare che, in base a tale funzione, sia un errore nel valore osservato ($valore_i$) sia un errore nella quantità osservata ($quantità_i$) contribuiscono alla determinazione dell'impatto potenziale.

consentirebbe di controllare anche quelle caratterizzate da valori eccessivamente bassi, migliorando la qualità delle stime finali.

Bibliografia

Di Pietro E. "Le statistiche del Commercio estero dell'Istat. Rilevazione Extrastat". *Documenti ISTAT*, 14, Istituto Nazionale di Statistica, Roma, 2006.

Fescina R., Jennings A., Wroblewski M. "Automated production of foreign trade data parameters using resistant fences". *Proceedings of the Joint Statistical Meetings*, *Section on Survey Research Methods, American Statistical Association*, Toronto, Canada, August 8-12, 2004.

Garcia M. M., Gajcowski A., Jennings A. "Selective editing strategies for the U. S. Census Bureau Foreign Trade Statistics Programs". *UN/ECE Work Session on Statistical Data Editing*, Bonn, Germany, September 25-27, 2006.

Jäder A., Norberg A. "A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics". *UN/ECE Work Session on Statistical Data Editing*, Ottawa, Canada, May 16-18, 2005.

Hidiroglou M. A., Berthelot J. M.. "Statistical editing and imputation for periodic business surveys". *Survey Methodology*, vol. 12, n. 1, pp. 73-83, 1986.

Kramer, C. Y. "Extension of multiple range tests to group means with unequal numbers of replications". *Biometrics*, 12, pp. 307-310, 1956.

SAS Institute Inc. *SAS/STAT User's guide*, version 6, vol. 1, SAS Institute Inc., Cary, NC, 1989.

Scheffé H. *The analysis of variance*, J. Wiley & Sons, New York, 1959.

Tukey, J. W. "Comparing individual means in the analysis of variance". *Biometrics*, 5, pp. 99-114, 1949.

Appendice

Determinazione dei parametri di segnalazione del prezzo medio: problemi indotti dall'aggiornamento annuale della Nomenclatura Combinata

La Nomenclatura Combinata, in vigore dal 1° gennaio 1988, viene aggiornata ogni anno mediante la nascita, la soppressione o la modifica dei codici merceologici *NC8* e delle descrizioni ad essi relative.
A gennaio 2007, l'aggiornamento della Nomenclatura Combinata ha comportato la soppressione di 1.202 codici *NC8* del 2006; in sostituzione di questi codici sono stati introdotti 1.080 nuovi codici nella Nomenclatura del 2007. La corrispondenza tra ognuno dei nuovi codici e uno o più dei 1.202 codici *NC8* del 2006 è descritta nella cosiddetta tabella delle trasposizioni[168].
Va osservato che per i codici introdotti nel 2007 non sono disponibili gli intervalli definiti in modo ragionato dai revisori. Quindi, il confronto (i cui risultati sono riportati nel paragrafo 5) tra i parametri determinati automaticamente e quelli individuati dai revisori può essere effettuato anche per i domini individuati dai nuovi codici *NC8*, solo adottando opportuni accorgimenti.
Per quanto riguarda i parametri stabiliti in modo ragionato, questi possono essere definiti anche per i domini individuati dai nuovi codici *NC8* operando nel modo seguente (*procedimento A*):

- se uno o più dei codici *NC8* introdotti nel 2007 corrispondono ad un solo codice *NC8* del 2006, i nuovi codici ereditano i parametri di segnalazione stabiliti in precedenza dai revisori per il codice *NC8* corrispondente del 2006;
- se un codice *NC8* introdotto nel 2007 deriva dalla fusione di più codici *NC8* del 2006, il nuovo codice eredita, come parametri di segnalazione, il minimo e il massimo dei parametri relativi ai codici *NC8* corrispondenti del 2006.

Invece, per quanto riguarda i parametri determinati automaticamente relativi ai domini individuati dai nuovi codici *NC8*, si procede come segue (*procedimento B*):

- se uno o più codici *NC8* introdotti nel 2007 corrispondono ad un solo codice *NC8* del 2006, i nuovi codici ereditano le osservazioni relative al codice *NC8* corrispondente del 2006;
- se un codice *NC8* introdotto nel 2007 deriva dalla fusione di più codici *NC8* nel 2006, il nuovo codice eredita le osservazioni relative a tutti i codici *NC8* corrispondenti del 2006.

In altre parole, nel procedimento *B* le "vecchie" osservazioni partecipano, insieme alle "nuove", al calcolo dei parametri secondo la procedura esposta nel paragrafo 4, tenendo conto della corrispondenza stabilita tra i codici merceologici da un anno all'altro. Inoltre, con il passare dei mesi, le osservazioni relative ai nuovi codici sono sempre più numerose rispetto alle osservazioni relative ai vecchi codici e hanno quindi un peso sempre più rilevante nella determinazione dei parametri. I parametri di segnalazione ottenuti con il procedimento *B* presentano il vantaggio di poter essere calcolati (e aggiornati) automaticamente in modo accurato, senza richiedere l'intervento dei revisori, anche in occasione dei cambiamenti introdotti annualmente nella Nomenclatura Combinata.

([168]) disponibile su web:
http://www.agenziadogane.gov.it/wps/wcm/connect/ed/Servizi/Intrastat/Software+Intrastat/Software+Intrastat+anno+2007/

Discussione
Standardizzazione e personalizzazione delle fasi di controllo e correzione nell'ambito delle indagini congiunturali sulle imprese

Roberto Gismondi, Istat, Direzione Centrale delle Statistiche Economiche Congiunturali su Imprese,Servizi e Occupazione, Ufficio del direttore e attività di supporto metodologico e alla diffusione (DCSC/U)

1. Premessa

Nell'ambito delle attività che di norma caratterizzano gli istituti nazionali di statistica, la produzione di macrodati per la diffusione rappresenta, tradizionalmente, la fase maggiormente apprezzata dagli utilizzatori finali. Il recente consolidamento del ruolo sempre più rilevante assunto dai cosiddetti "utilizzatori privilegiati" (si pensi agli organismi internazionali, in particolare EUROSTAT e la Banca Centrale Europea) ha contribuito ad un ulteriore ampliamento della quantità di indicatori finali che è necessario produrre, accrescendone spesso anche la tempestività. Quest'ultimo aspetto è poi particolarmente rilevante nel contesto degli indicatori congiunturali, finalizzati a fornire segnali quanto mai tempestivi circa l'andamento di vari comparti del sistema economico nazionale.
In questo contesto, potrebbe rischiare di cadere erroneamente in secondo piano l'insieme delle fasi che precedono la diffusione degli indicatori finali. In particolare, la fase del trattamento dei micro-dati antecedente al calcolo delle stime e dei relativi errori riveste un ruolo di primaria rilevanza, sebbene essa non risulti sempre del tutto trasparente in sede di documentazione del processo produttivo, né tantomeno siano sempre chiare le implicazioni delle singole categorie di intervento sui micro-dati sulle stime finali.
Uno dei meriti del seminario di cui questa rassegna ha illustrato i contributi è stato proprio quello di far *emergere il problema*: ossia, perché è importante documentare le fasi di controllo e correzione antecedenti alla fase di stima, e con quale varietà di strumenti metodologici un istituto nazionale di statistica cerca di migliorare la qualità delle stime finali intervenendo sui dati di base, ossia confrontando l'insieme dei dati *osservati* con uno schema teorico ritenuto adeguato per modellizzare il comportamento *atteso* dei dati.
La rassegna dei documenti oggetto di presentazione nel corso del seminario è ampia, approfondita e ricca di spunti di riflessione. Dunque, un primo aspetto da evidenziare è proprio il constatare la vastità del patrimonio informativo disponibile, che sarebbe stato difficile far emergere del tutto senza l'opportunità derivata dal progetto EDIMBUS. In tale ottica, questo documento ha l'obiettivo di lasciare una traccia tangibile del lavoro svolto, che possa servire anche per permetterne un aggiornamento periodico ed un eventuale ampliamento con riferimento ad indagini statistiche attualmente non documentate (o documentate solo in parte).
Le rilevazioni congiunturali oggetto di discussione sono estremamente eterogenee in quanto a campo di osservazione e metodologia d'indagine. Indubbiamente tale aspetto giustifica la necessità di utilizzare strumenti di controllo e correzione dei micro-dati altrettanto diversi – come risulta evidente da una lettura attenta dei contributi – ma, d'altra parte, riporta l'attenzione verso il dualismo tra necessità di personalizzare le varie categorie di intervento e l'uso di strumenti il più possibile omogenei, o comunque riconducibili a canoni standardizzati.
Un ulteriore dualismo riguarda la coesistenza, in molto processi produttivi, di fasi di controllo altamente o completamente automatizzate ed interventi interattivi. Forse non si tratta di due approcci in reale contrasto: la loro coesistenza è possibile e per certi versi resa necessaria dalla non possibilità di individuare dei modelli comportamentali di riferimento per tutte le tipologie di dati e, soprattutto, che siano valide per tutte le dimensioni d'impresa. Normalmente le imprese di grandi dimensioni presentano profili molto specifici e difficilmente replicabili, per cui l'adozione di schemi di intervento troppo automatizzati finirebbe con l'essere rischiosa. E' altresì chiaro che la presenza di interventi sui

micro-dati di tipo interattivo comporta l'inevitabile impossibilità di valutare e quantificare tutte le possibili componenti di errore non campionario; inoltre, talvolta implica la non disponibilità di file di micro-dati ante e post correzione, dato che l'esito delle fasi di controllo e correzione può consistere nel sovrapporre un file di dati *corretti* al file originario dei cosiddetti dati *grezzi*.
Un aspetto che forse dovrebbe essere oggetto di un'attenzione maggiore – o essere comunque documentato con maggiore dovizia di dettagli – riguarda il legame tra le diverse fasi di controllo e correzione e la qualità delle stime finali, ossia l'effetto che le varie categorie di intervento hanno sull'errore di stima. Talvolta si è avuta l'impressione che determinate scelte operative – per quanto quasi sempre frutto di studi approfonditi della base dati disponibile – non siano state intraprese nella piena consapevolezza del legame esistente tra il pre-trattamento dei dati grezzi e la fase di stima, laddove invece sarebbe necessario considerare l'intera strategia di stima come un corpus unico, includendovi anche l'identificazione e correzione degli errori ed il trattamento delle mancate risposte.
In sostanza, una domanda che è ancora necessario porsi è la seguente: le stime finali ottenute dopo aver applicato una serie di interventi preliminari sui micro-dati sono *certamente* caratterizzate da una qualità superiore alle stime che si sarebbero potute ottenere senza intervenire affatto, o intervenendo in altro modo, o comunque in modo meno *intenso* sui dati di base?
Va da se che la possibilità di rispondere a questa domanda implica di non lasciare che i contenuti del seminario restino fine a se stessi, ma rappresentino la fase di avvio di ulteriori analisi, che possano condurre ad una risposta esauriente.

2. Il contesto internazionale

L'ampliamento dell'Unione Europea e l'introduzione della moneta unica hanno accelerato il processo di convergenza dei sistemi statistici nazionali in ambito UE verso uno standard produttivo e qualitativo il più possibile comune e condiviso. Peraltro, non sempre la necessità di ampliare il panorama degli indicatori congiunturali sulle imprese disponibili a livello UE comporta necessariamente una crescita del grado di soddisfazione degli utilizzatori, sia perché la tempestività richiesta per tali indicatori potrebbe non essere immediatamente garantita, sia perché l'impianto di nuove indagini finalizzate a produrre stime per le quali la tempestività finisce con il rappresentare il fattore qualitativo ritenuto più importante potrebbe non essere associato al rispetto di livelli qualitativi minimali, oltre a risultare spesso molto costoso per l'intera collettività. In questo contesto, il ricorso a fonti amministrative – se possibile – comporterebbe una scelta strategica da privilegiare, come in parte già fatto anche in Italia. In sintesi, occorre sottolineare il rischio dovuto al noto *trade/off* tra qualità e tempestività degli indicatori statistici.
EUROSTAT sta cercando da anni di uniformare la produzione statistica congiunturale degli stati dell'UE, sebbene privilegiando la quantità degli indicatori e la loro tempestività rispetto all'effettiva utilizzabilità economica immediata (molte serie storiche sono brevi e dunque non destagionalizzabili) ed alla affidabilità statistica degli stessi. In tal senso, i principali strumenti di indirizzo e controllo sono i seguenti: 1) adozione di Regolamenti e Direttive; il ruolo fondamentale è giocato dal Regolamento sulle statistiche congiunturali (*Short Term Statistics*, *STS*), che ha identificato la lista di indicatori da produrre, regolamentandone il livello di dettaglio, la tempestività ed alcuni requisiti minimi di qualità. 2) Aggiornamento quinquennale del manuale metodologico europeo di riferimento per gli indicatori congiunturali (l'ultima edizione si riferisce al 2007). 3) Raccolta ed aggiornamento annuale delle principali meto-informazioni relative ai singoli stati UE con riferimento al calcolo degli indicatori congiunturali inerenti al regolamento STS (*STS Sources*). 4) Un approfondimento metodologico relativo ai Principal European Economic Indicators (*PEEIs in focus*), con cui ogni anno si richiede agli stati UE di fornire informazioni - più dettagliate rispetto a *STS Sources* - per uno degli indicatori congiunturali di maggiore rilevanza (dal 2005 al 2008 sono stati analizzati l'indice della produzione industriale, l'indice delle vendite al dettaglio, i prezzi alla produzione dei prodotti industriali e l'indice di produzione delle costruzioni). Tali basi informative rappresentano una rilevante novità rispetto al passato recente, e consentono di poter comparare metodologie d'indagine e scelte operative connesse al trattamento dei micro-dati tra stati diversi. In particolare, i *PEEIs in focus* approfondiscono tematiche quali il trattamento delle modificazioni longitudinali delle imprese, l'analisi ed il trattamento della mancata

risposta (tassi di non risposta, azioni intraprese per ridurre l'effetto delle non risposte, stima della distorsione da non risposta), la misurazione dei possibili errori (processo di editing dei dati, errori non campionari).
La graduale convergenza verso standard europei comuni relativamente alla valutazione degli aspetti qualitativi connessi al processo di data editing non può che stimolare ed ampliare l'attività di documentazione dei processi, che indipendentemente dall'effettivo livello di precisione delle stime finali relative ad un dato indicatore rappresenta un fattore estremamente positivo ed aiuta a migliorare i processi stessi.
Tuttavia, è doveroso ricordare come sia sempre estremamente rischioso – e per certi versi non del tutto corretto – confrontare le caratteristiche metodologiche di processi inerenti lo stesso indicatore, ma sviluppati in contesti operativi spesso assai diversi, per loro natura non paragonabili tra loro. Ad esempio, si consideri la tabella seguente: il fatto che i livelli di tempestività nella diffusione dei dati nei vari stati siano piuttosto eterogenei implica la non confrontabilità dei tassi di risposta e dell'effetto degli eventuali solleciti. Inoltre, gli stessi tassi di risposta – nonché l'efficacia delle procedure messe in atto per stimare le non risposte – sono essi stessi fortemente interconnessi con la tipologia di indagine ed il tasso di campionamento adottato: in merito, si noti come in Italia si sia costretti ad operare con un tasso di sondaggio particolarmente basso (inferiore al 2%), data la vastità dell'universo oggetto di interesse (oltre mezzo milione di imprese commerciali al dettaglio).
Sebbene, negli ultimi anni, EUROSTAT abbia chiaramente investito più sulla quantità che sulla qualità degli indicatori congiunturali, va comunque sottolineato l'effetto indiretto indotto da tale attività sulla revisione ed il possibile miglioramento dei processi esistenti, talvolta in conflitto con la necessità di avviarne di nuovi.

Tabella 1: *Alcuni aspetti metodologici relativi all'indice delle vendite al dettaglio in alcuni stati europei*

VARIABILE	Italia	Germania	UK	Francia	Paesi Bassi	Spagna
TEMPESTIVITA'	30 giorni provvisorio, 51 definitivo	42 giorni	19 giorni	30 giorni provvisorio, 60 definitivo	35 giorni provvisorio (definitivo 4 mesi dopo)	30 giorni al provvisorio, definitivo 3 mesi dopo
RACCOLTA DATI	Soprattutto fax, poi postale, web in crescita	Soprattutto postale e fax, anche telefono, internet e e-mail	Postale	Dichiarazioni VAT e indagine per imprese non soggette a versamenti	60% postale, 40% e-mail	Postale e telefonico
TIPO DI INDAGINE	Campione casuale stratificato (attività economica, forma distributiva e dimensione)	Campione stratificato (regione, attività economica e numero di addetti)	Campione stratificato (attività economica e 4 classi di addetti)	Campione stratificato (attività economica)	Campione stratificato (censuaria oltre 50 addetti), basato su attività economica e numero di addetti	Campione stratificato (censuaria oltre 50 addetti), basato su attività economica e numero di addetti e regione
TASSO DI CAMPIONAMENTO	Inferiore al 2%	5%	5.000 imprese (censuaria oltre 99 addetti, 1% 1-9, 5% 10-19, 20% 20-99)	21.000 imprese, equivalenti a più del 95% del fatturato	10.000 imprese (70% della popolazione)	1,25% 1-2 5% 3-9 18% 10-49
TASSO DI RISPOSTA	Intorno al 50%	100°%	60% al provvisorio e 90% un mese dopo	90% al provvisorio, 100% al definitivo	80% al provvisorio	90% a provvisorio, 95% al definitivo
SOLLECITI	Solleciti telefonici mirati per le imprese più grandi	Postali e telefonici	Postali e telefonici	-	Telefonici	Telefonici e sanzioni

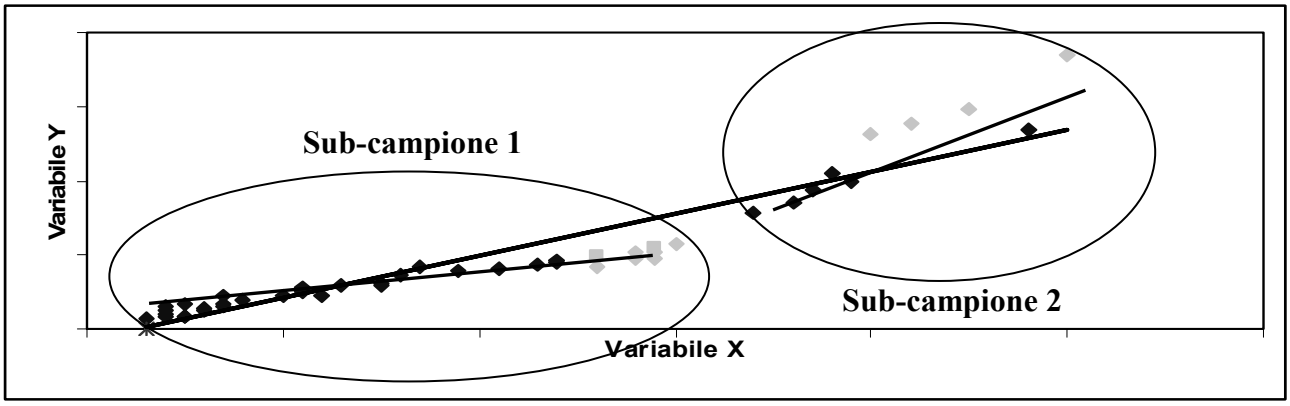
STIMA NON RISPOSTE	Diverse tecniche in base a dati disponibili (di norma: trend medio dei rispondenti)	Stimatore rapporto (relazione tra 3 mesi precedenti e stesso mese anno precedente	Stesso trend medio dei rispondenti	-	Stesso trend medio dei rispondenti o stessa media dei rispondenti	Stesso trend medio dei rispondenti

3. Alcuni aspetti specifici

Nel complesso, il problema della individuazione ed eventuale correzione degli "errori di misura" sembra rappresentare, nel processo di valutazione delle possibili fonti di errore, una componente piuttosto evoluta, nel senso che di norma gli esperti d'indagine sono in grado di fronteggiare con gli strumenti adeguati – spesso automatizzati – la possibile presenza di tali componenti erronee.
D'altra parte, due aspetti di particolare rilevanza, su cui persiste un'elevata varietà di tecniche d'intervento, riguardano le osservazioni anomale (*outlier*) ed il trattamento della mancata risposta.

L'individuazione delle osservazioni anomale è storicamente un tema tanto dibattuto quanto sfumato proprio in relazione alle sue connotazioni più specifiche, ossia: 1) quali siano le unità anomale e 2) quale sia il trattamento più adeguato da applicare a tali unità. Indirettamente, l'essere o meno "anomalo" implica il riferimento ad un modello di comportamento "corretto", ossia al dare massima fiducia al fatto che ad ogni unità statistica siano associati dei valori "attesi", o comunque più probabili di altri, in merito alle variabili oggetto di interesse. Al crescere della discrepanza tra valori osservati ed attesi, dovrebbe crescere la probabilità che una certa osservazione sia "anomala". Va da se che la ragionevolezza di tale impostazione implica che possa essere identificato un unico modello di riferimento, valido per tutte le osservazioni oggetto di interesse.
In realtà non di rado può accadere che siano osservati due sub-campioni nell'ambito dello stesso dominio campionario (figura 1). Al fine di identificare possibili osservazioni anomale, è corretto valutare la distanza tra tutte le osservazioni da un'unica retta di regressione, e non piuttosto ritenere più adeguata l'esistenza di due sub-popolazioni – da cui sono stati osservati due sub-campioni 1 e 2 – e quindi anche di due rette di regressione rispetto alle quali valutare separatamente la distanza dei singoli punti?

Figura 1: *Possibili outlier in un'indagine campionaria (punti in grigio)*



In sostanza, la questione di fondo consiste nel decidere se analizzare l'intero campione osservato in un'unica soluzione – ammettendo implicitamente che tutte le osservazioni disponibili derivino dal medesimo modello teorico generatore dei dati – oppure se condurre delle valutazioni separate in 2 (o eventualmente più di 2) sottocampioni del tipo di quelli esemplificati nella figura precedente. E' chiaro che, nel secondo caso, si deve affrontare il problema di come identificare tali sub-campioni, sulla base, possibilmente, di algoritmi oggettivi. Va notato come una possibile soluzione a tale problema potrebbe derivare da una semplice post-stratificazione delle unità campionarie all'interno dello strato originario oggetto d'interesse, che richiede però la disponibilità di almeno una variabile ausiliaria, preferibilmente

quantitativa, utile per identificare diversi modelli di riferimento che possano spiegare i vari profili comportamentali dei dati.
Più in generale, si dovrebbe sempre tenere conto del fatto che la peculiarità di essere "outlier" non va valutata tanto (o solo) in relazione alle sole osservazioni campionarie disponibili, ma anche (e soprattutto) in relazione alla distribuzione del carattere d'interesse nell'intera popolazione, ossia cercando di inferire quante possano essere, nella popolazione, le unità non osservate tipologicamente simili ad ogni unità campionaria sospettata di essere outlier (si veda, ad esempio, al lavoro di Welsch e Ronchetti, 1998). In altri termini, si dovrebbe cercare di confrontare la distribuzione di frequenze campionaria con quella della popolazione: ad esempio, se nel campione osservato la frequenza relativa delle unità con un valore di *y* superiore ad una data soglia fosse solo dell'1%, mentre l'analoga quota nella popolazione fosse del 3%, considerare come anomale tutte le unità campionarie che determinano il suddetto 1% sarebbe altamente deleterio, perché a fronte di un campione osservato che già di per se non sarebbe pienamente rappresentativo della coda di destra della distribuzione universo (1% rispetto a 3%), si andrebbe ulteriormente a smussare il peso (in certi casi addirittura azzerandolo) delle unità estreme, aumentando quasi certamente la distorsione per difetto della stima finale così producibile.
Quindi, in sintesi: se la distribuzione di frequenze campionaria è molto simile a quella dell'intera popolazione, ridurre il peso campionario delle unità identificate come anomale può peggiorare la precisione della stima finale. Per ridurre tale rischio, si dovrebbe valutare la probabilità che un'unità sia anomala condizionatamente al campione osservato *ed all'intera popolazione*, mentre in pratica si finisce spesso con il valutare tale probabilità condizionatamente al solo campione disponibile. E' chiaro che tale opzione è spesso l'unica applicabile, perché sarebbe irrealistico supporre di poter conoscere con certezza la distribuzione di *y* nella popolazione. Tuttavia, è necessario approfondire la ricerca in tale campo tramite il ricorso a variabili ausiliarie note per l'intera popolazione, o a valutazioni *ex post* circa il grado di precisione di stime relative a periodi precedenti prodotte ricorrendo a diverse procedure di identificazione e trattamento dei valori anomali (tra le quali, possibilmente, dovrebbe trovare posto anche l'opzione consistente nel non alterare il peso di nessun valore potenzialmente anomalo).

Riguardo al trattamento delle mancate risposte, il problema di fondo riguarda la poca chiarezza circa gli strumenti da applicare per poter valutare la presenza di un'eventuale componente distorsiva della stima dovuta alle mancate risposte. A volte si ha l'impressione che tale valutazione non sia gestita con strumenti non del tutto adeguati, e talvolta viene un po' ignorata. Ciò dipende anche dal fatto che la stessa letteratura più o meno recente sul tema ha speso molte energie nel campo della valutazione teorica degli effetti di diversi profili di non risposta (MCAR, MAR, NMAR), ma molte meno nel proporre un set adeguato di tali strumenti di valutazione, in funzione della tipologia di indagine e della disponibilità o meno di informazioni ausiliarie (in primo luogo, serie storiche). La principale conseguenza è che si rischia di impegnare tempo e risorse per tentare di rettificare le stime per tenere conto della mancata risposta, senza sapere esattamente se tale azione – spesso molto complessa e talvolta troppo personalizzata in funzione dell'indagine considerata – porti a qualche effettivo miglioramento nella qualità della stima finale, o conduca comunque a miglioramenti solo di lieve entità.
Ci sentiamo di aggiungere che, sebbene la riponderazione (particolarmente enfatizzata dalla letteratura recente) e l'imputazione della mancata risposta (approccio più tradizionale) abbiano origine da logiche diverse, il contrasto tra le due opzioni è in realtà spesso sfumato (si potrebbe facilmente verificare che ad ogni operazione di riponderazione corrisponde una certa tecnica di imputazione e viceversa). In realtà il problema di fondo consiste nel capire quali siano le assunzioni metodologiche sottostanti l'uso di una certa tecnica di riponderazione o di imputazione. Senza la disponibilità di almeno una variabile ausiliaria molto correlata con quella (principale) oggetto di stima, e senza l'uso corretto di tale informazione nel contesto dell'impianto metodologico più adeguato, si rischia, come già accennato, di implementare processi di correzione della mancata risposta di estrema complessità, ma dagli effetti incerti.
Senza entrare nel merito delle singole indagini trattate, in tutti i casi si è constatata l'estrema attenzione riposta verso il problema della non risposta, nonché la predisposizione di opzioni correttive quasi sempre capillari e razionali, tese a sfruttare nel modo migliore tutte le informazioni ausiliarie disponibili sulle singole unità non rispondenti (dando priorità alla disponibilità di dati della stessa unità relativi a

periodi precedenti). Si potrebbe comunque suggerire di approfondire (o documentare maggiormente) l'analisi a monte del possibile processo di generazione della non risposta e della presenza di fattori che potrebbero discriminare tra diversi *pattern* di non risposta. Si ritiene che, a tale fine, sia doveroso ricorrere a modelli utili per spiegare il verificarsi dei dati osservati, la cui disponibilità potrebbe essere utile anche per rivedere, nel tempo e con gradualità, le peculiarità dell'intero processo di stima. In tale ottica, è di fondamentale importanza poter confrontare i rispondenti con i non rispondenti, possibilmente recuperando i dati di almeno una parte dei non rispondenti anche successivamente alla data di diffusione delle stime preliminari e/o finali.
Ulteriori accorgimenti per la riduzione della non risposta sono particolarmente raccomandati e si è constatato come siano spesso già ampiamente implementati nei vari processi d'indagine: ci si riferisce, in particolare, alla gestione razionale dei solleciti (possibilmente *mirati*, privilegiando le unità più influenti, si veda anche Gismondi, 2007), alla registrazione della data di arrivo dei questionari e alla reintervista ex post per alcune unità sulla cui base poter valutare l'entità della possibile distorsione da non risposta.

4. Conclusioni e prospettive

Nel contesto delle rilevazioni congiunturali sulle imprese, negli ultimi anni la necessità di rinnovare indagini preesistenti o di impiantare ex novo nuove indagini ha comportato tanto l'ampliamento della base informativa a disposizione per gli utenti finali (primo tra tutti EUROSTAT), quanto la proliferazione di metodologie per il trattamento dei micro-dati quanto mai eterogenee e "personalizzate" in funzione della particolare tipologia dei dati trattati.
Di per sé l'eterogeneità degli approcci per il controllo ed il trattamento dei micro-dati non rappresenta necessariamente un fattore di rischio, dato che, come ampiamente documentato, le rilevazioni congiunturali esaminate si basano su disegni d'indagine altrettanto eterogenei, in quanto concepiti in momenti storici diversi e riferiti a domini di riferimento alquanto diversificati. D'altra parte, sarebbe forse utile approfondire le reali motivazioni che hanno condotto alla scelta di determinate strategie di data editing, ossia documentando nel miglior modo possibile la reale necessità di una diversificazione tra gli approcci, ossia in che misura tale scelta dipenda da un'effettiva diversità dei dati raccolti. Inoltre, come già ricordato, sarebbe auspicabile tentare di collegare maggiormente le operazioni di intervento sui micro-dati con gli effetti che tali interventi potrebbero avere sulla precisione delle stime finali. E' noto che tale valutazione è senz'altro complessa e richiederebbe il ricorso a strumenti di analisi dell'errore non campionario, che presumibilmente avrebbe senso sviluppare prima di tutto nel contesto delle indagini strutturali. E' però anche vero che una *strategia di stima* non può limitarsi alla sia pur fondamentale fase di scelta del disegno di campionamento e della tecnica di stima, ma deve prevedere in un corpus unico anche tutti gli interventi sui micro-dati che, altrimenti, verrebbero effettuati sulla base di principi rischiosamente sconnessi rispetto al momento della stima finale, ed i cui effetti sarebbero difficilmente valutabili sulla base di strumenti oggettivi.
Un altro aspetto che dovrebbe essere valutato con qualche attenzione in più riguarda il livello di dettaglio utilizzato tanto per le stratificazioni adottate in sede di disegno campionario, quanto per analizzare il comportamento dei micro-dati raccolti, al fine di valutarne la qualità e quindi di decidere se ed in che modo intervenire su di essi. Pur tenendo conto del fatto che, molto spesso, la necessità di ricorrere a stratificazioni molto capillari deriva dall'elevato livello di dettaglio (e di precisione) richiesto per molti degli indicatori congiunturali analizzati, si ha l'impressione che operare all'interno di domini di riferimento troppo piccoli (al fine di distinguere il più possibile tra loro unità statistiche con profili almeno teoricamente diversi in quanto a settore economico, dimensione, ecc.) possa finire con il rendere più complesse una serie di valutazioni, che rischiano di basarsi su basi di dati campionarie troppo piccole e dunque meno affidabili al fine di identificare errori, valori anomali, e per lo stesso trattamento della mancata risposta. Inoltre, a parità di capillarità della stratificazione considerata (numero di strati), si potrebbe tentare di verificare la possibilità e l'utilità del ricorso a nuove variabili di stratificazione, come ad esempio il volume di affari del registro delle imprese ASIA. Come noto, tale variabile è ormai disponibile per la pressoché totalità delle imprese e presenta un livello qualitativo

piuttosto elevato – come anche confermato da valutazioni recenti in merito. Almeno per quanto riguarda le indagini congiunturali finalizzate a stimare la variazione dei ricavi nel tempo (vendite al dettaglio, fatturato degli altri servizi, fatturato industriale), il livello di correlazione del volume di affari con i ricavi dovrebbe risultare ragionevolmente superiore a quello ottenibile ricorrendo (oltre che al settore di attività economica) al numero di addetti. Si tratta di un'ipotesi ancora da valutare, ma che potrebbe condurre a vantaggi qualitativi di dimensioni assai superiori a quelli ottenibili (solo) sulla base di complesse strategie di intervento sui micro-dati. Non dimentichiamo, infatti, che molto spesso l'identificazione di erroneità e/o anomalie vere o presunte nei micro-dati rappresenta un'azione condotta, più o meno esplicitamente, facendo riferimento ad un modello comportamentale "corretto" che viene inevitabilmente associato al profilo prevalente riscontrato sulle unità osservate nel medesimo strato di riferimento. In altri termini, la "anomalia" finisce con l'essere associata alla presunta appartenenza dell'unità sospetta ad uno strato diverso rispetto a quello in cui tale unità era stata originariamente classificata, per cui è ragionevole tentare di affrontare a monte il problema chiedendosi se i criteri di stratificazione adottati siano realmente in grado – o meno – di segmentare correttamente i diversi profili tipologici delle unità esaminate. In tale ottica, va sottolineata la difficoltà supplementare dovuta al fatto che, nel contesto congiunturale, tali profili sono soprattutto connessi alla dinamica longitudinale delle imprese, mentre le variabili di stratificazione sono sempre associate a peculiarità per loro natura "statiche" e non dinamiche (lo stesso volume di affari è valutato nella forma di livello medio annuale e non come, ad esempio, trend pluriennale).

In prospettiva, sarebbe di grande interesse ed utilità poter impiantare un sistema di controllo della qualità delle indagini – indipendentemente dal fatto che siano congiunturali o strutturali, su imprese o famiglie – in grado di valutare l'effetto di ogni componente di errore sulla precisione delle stime finali. Tale opzione consentirebbe di poter identificare le priorità di intervento e distribuire tempi e risorse da dedicare alla fase di controllo e correzione in modo efficiente. In particolare, per le indagini congiunturali assume un ruolo di particolare rilevanza il concetto di *universo longitudinale*. La disponibilità di popolazioni di riferimento infraannuali a cui riportare i dati campionari consentirebbe il calcolo di stime senza dubbio più coerenti con le necessità informative in ambito congiunturale, che in sostanza si basano sulla misurazione di tassi di variazione nel tempo dell'ammontare complessivo di una certa variabile di output, dove il termine *complessivo* si riferisce all'intero universo oggetto di indagine. Senza la disponibilità di tali universi, un'indagine statistica congiunturale fornisce, di norma, un'indicazione sulle variazioni dell'ammontare medio della variabile di output analizzata (ossia l'ammontare complessivo misurato sulle unità del campione diviso per il numero di tali unità), o comunque dell'ammontare complessivo, ma riferito ad un universo statico (quello dell'anno base se si tratta di indici, o al più dell'anno precedente a quello analizzato). Si ha motivo di ritenere che la componente dell'errore di stima totale dovuta a tale approssimazione possa risultare particolarmente elevata, forse più di quelle attribuibili ad altre possibili cause di errore.

Due ultime osservazioni. La prima, apparentemente ovvia ma in realtà ancora da ribadire e talvolta da implementare, riguarda la necessità di conservare una memoria storia dei micro-dati, distinguendo tra quelli ante e post trattamento di data editing. Talvolta la non distinguibilità tra le due fasi suddette deriva espressamente dal fatto di praticare controlli interattivi che sostituiscono in tempo reale il dato originario con un nuovo dato ritenuto corretto. Tuttavia, senza un confronto tra il pattern dei micro-dati nelle varie fasi del processo di lavorazione è molto più difficile valutare anche gli stessi guadagni qualitativi derivati dal data editino nel suo complesso.

La seconda osservazione, anch'essa forse già implicita da quanto detto finora, riguarda le necessità di conservare e sostenere in ogni modo la cooperazione tra le funzioni "operative" – più vicine al processo produttivo ed alle sue scadenze – e quelle più prettamente votate all'analisi metodologica e sistemica, la cui interazione non dovrebbe circoscriversi alla sia pur fondamentale fase di progettazione dell'indagine, ma essere alimentata da continui scambi informativi che finirebbero con l'accrescere le conoscenze complessive sui processi e sui profili longitudinali delle basi di dati attualmente disponibili nell'intero sistema statistico nazionale.

Bibliografia

R. GISMONDI (2007), "Score Functions and Statistical Criteria to Manage *Intensive Follow Up* in Statistical Surveys", in corso di pubblicazione su *Statistica*.

A. H. WELSCH, E. RONCHETTI (1998) "Bias-Calibrated Estimation from Sample Surveys Containing Outliers", *Journal of the Royal Statistical Society*, Series B, Vol.60, part 2, 413-428.