istat
working
papers

N.4
2019

L'imputazione delle mancate risposte sulle posizioni lavorative dipendenti nei dati provvisori di fonte Inps: metodologia e sperimentazioni nella rilevazione trimestrale Oros

Marco Lattanzio, Francesca Romana Pogelli, Donatella Tuzi

Istat
Istituto Nazionale
di Statistica

Direttora Responsabile:
Patrizia Cacioli

Comitato Scientifico

Presidente:
Gian Carlo Blangiardo

Componenti:

Corrado Bonifazi	Vittoria Buratta	Ray Chambers	Francesco Maria Chelli
Daniela Cocchi	Giovanni Corrao	Sandro Cruciani	Luca De Benedictis
Gustavo De Santis	Luigi Fabbris	Piero Demetrio Falorsi	Patrizia Farina
Jean-Paul Fitoussi	Maurizio Franzini	Saverio Gazzelloni	Giorgia Giovannetti
Maurizio Lenzerini	Vincenzo Lo Moro	Stefano Menghinello	Roberto Monducci
Gian Paolo Oneto	Roberta Pace	Alessandra Petrucci	Monica Pratesi
Michele Raitano	Giovanna Ranalli	Aldo Rosano	Laura Terzera
Li-Chun Zhang			

Comitato di redazione

Coordinatrice:
Nadia Mignolli

Componenti:

Ciro Baldi	Patrizia Balzano	FedericoBenassi	Giancarlo Bruno
Tania Cappadozzi	Anna Maria Cecchini	Annalisa Cicerchia	Patrizia Collesi
Roberto Colotti	Stefano Costa	Valeria De Martino	Roberta De Santis
Alessandro Faramondi	Francesca Ferrante	Maria Teresa Fiocca	Romina Fraboni
Luisa Franconi	Antonella Guarneri	Anita Guelfi	Fabio Lipizzi
Filippo Moauro	Filippo Oropallo	Alessandro Pallara	Laura Peci
Federica Pintaldi	Maria Rosaria Prisco	Francesca Scambia	Mauro Scanu
Isabella Siciliani	Marina Signore	Francesca Tiero	Angelica Tudini
Francesca Vannucchi	Claudio Vicarelli	Anna Villa	

Cura editoriale:
Vittorio Cioncoloni

Istat Working Papers

L'imputazione delle mancate risposte sulle posizioni lavorative dipendenti nei dati provvisori di fonte Inps: metodologia e sperimentazioni nella rilevazione trimestrale Oros

N. 4/2019

ISBN 978-88-458-1988-9

© 2019
Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Salvo diversa indicazione, tutti i contenuti pubblicati sono soggetti alla licenza Creative Commons - Attribuzione - versione 3.0.
https://creativecommons.org/licenses/by/3.0/it/
È dunque possibile riprodurre, distribuire, trasmettere e adattare liberamente dati e analisi dell'Istituto nazionale di statistica, anche a scopi commerciali, a condizione che venga citata la fonte.
Immagini, loghi (compreso il logo dell'Istat), marchi registrati e altri contenuti di proprietà di terzi appartengono ai rispettivi proprietari e non possono essere riprodotti senza il loro consenso.

L'imputazione delle mancate risposte sulle posizioni lavorative dipendenti nei dati provvisori di fonte Inps: metodologia e sperimentazioni nella rilevazione trimestrale Oros

Marco Lattanzio, Francesca Romana Pogelli, Donatella Tuzi[*]

Sommario

In questo lavoro si presenta la metodologia di stima provvisoria del numero di posizioni lavorative dipendenti della rilevazione trimestrale Oros che utilizza intensivamente dati di fonte amministrativa. Il metodo cui si ricorre si basa sull'imputazione dei microdati per i rispondenti ritardatari all'istante di stima, la cui lista viene predetta sfruttando informazioni longitudinali sulla presenza delle unità nei dati amministrativi, caratterizzati da una copertura "quasi completa" della popolazione target. Successivamente, i dati mancanti vengono ricostruiti per regressione. La messa a regime del metodo ha seguito una lunga fase di sperimentazione nel corso della quale sono stati messi a punto diversi aspetti. La disponibilità di dati finali ad un anno di distanza dai dati provvisori ha permesso di quantificare l'entità reale dell'errore di revisione e valutare i progressi introdotti analizzando differenze a livello di singole unità, per aggregazioni di unità e per cause di errore.

Parole chiave: dati amministrativi, congiuntura, occupazione, stime provvisorie, mancate risposte, demografia d'impresa, regressione.

Abstract

This work describes the methodology applied for the preliminary estimates of the number of jobs by the quarterly Oros survey, intensively based on the use of administrative data. The proposed imputation method is based on the prediction of a current list of active units, extrapolated by an almost complete set of administrative data available for the short-term deadline and, subsequently, on the imputation of missing data using a regression model approach. The implementation of the method followed a long experimentation phase, during which several aspects were fine-tuned aiming at the minimization of the revision error. The availability of final data after one year from the preliminary ones allowed the calculation of the real revision error and an extended evaluation of the enhancements progressively introduced, analysing the differences towards the preliminary version both at micro and macro level, decomposing the revision error by main causes.

Keywords: administrative data, short term, employment, preliminary estimates, late reporting units, business demography, regression.

[*] Sebbene il documento sia il frutto di un lavoro congiunto degli autori, la redazione di questo lavoro è stata curata da Francesca Romana Pogelli. La stesura dei paragrafi è da attribuirsi a: § 2 e §4.1 Donatella Tuzi; §4.2, § 4.2.1 e § 5.1.1 Marco Lattanzio e Francesca Romana Pogelli; §4.2.2, § 5.1.2 e § 6 Francesca Romana Pogelli. I restanti paragrafi sono stati prodotti interamente in collaborazione dagli autori.

Indice

1. Introduzione

La rilevazione trimestrale Oros produce indicatori su occupazione dipendente, retribuzioni, oneri sociali e costo del lavoro. Progettata alla fine degli anni '90 con l'obiettivo di completare l'offerta di statistiche dal lato "domanda" sfruttando intensamente dati di fonte amministrativa, attualmente la rilevazione si colloca all'interno di un sistema integrato di indicatori congiunturali attraverso cui l'Istituto Nazionale di Statistica diffonde informazioni su input e costo del lavoro nelle imprese ed istituzioni private dei settori dell'industria e dei servizi (Baldi et al., 2011a).

La rilevazione rilascia indicatori sul costo del lavoro a partire da dicembre 2002 mentre la prima uscita di indicatori sul numero di posizioni lavorative dipendenti risale a giugno 2015, dopo una lunga fase di sperimentazione avviata con la partecipazione dell'Istat al progetto europeo ESSnet[1], *Working Package 4* sull'uso di dati amministrativi per le statistiche congiunturali, cui è seguita la messa a regime di un metodo di imputazione delle mancate risposte nei dati amministrativi provvisori e ricostruzione della variabile secondo criteri statistici di rilevanza (cfr. Istat, 2015). Questo documento approfondisce il lavoro svolto durante la fase di sperimentazione, per la pubblicazione dei dati di stima provvisoria sulle posizioni lavorative.

Dopo aver delineato la struttura informativa che caratterizza le fonti alla base della rilevazione, vengono descritte le caratteristiche dei dati amministrativi mancanti ai fini della stima provvisoria. In seguito, dopo aver riportato l'esperienza Oros nel WP4, viene mostrata l'importanza della demografia d'impresa sulla stima del numero delle posizioni lavorative dipendenti e proposto un metodo d'imputazione basato sulla predizione di una lista corrente di unità attive, operazione possibile grazie alla disponibilità di dati amministrativi quasi completi all'istante di stima provvisoria e, successivamente, di ricostruzione dei dati mancanti ricorrendo ad un approccio per regressione.

La messa a regime del metodo ha seguito una lunga fase di sperimentazione, ampiamente descritta in questo documento, in cui sono stati messi a punto svariati aspetti con il fine di rendere la stima provvisoria più vicina possibile rispetto alla stima finale: la disponibilità di dati finali ad un anno di distanza dai dati provvisori usati per la compilazione delle stime definitive, ha consentito di valutare in maniera approfondita i progressi introdotti analizzando l'errore di revisione e la sua disaggregazione per domini di stima e per cause.

2. Cenni sulla rilevazione, fonti di dati e stime

2.1 Caratteristiche della rilevazione

La popolazione oggetto della rilevazione Oros sono le imprese e le istituzioni private con dipendenti di tutte le classi dimensionali che hanno corrisposto, nel trimestre di riferimento, retribuzioni imponibili ai fini contributivi e svolgono la loro attività economica nei settori dell'industria (sezioni di attività economica da B ad F della classificazione Ateco 2007) e dei servizi (sezioni da G a S ad esclusione di O). L'insieme degli occupati dipendenti comprende operai, impiegati, dirigenti e apprendisti, a prescindere dal tipo di contratto (tempo indeterminato, determinato, stagionale, ecc.) e dal tipo di prestazione lavorativa (tempo pieno, tempo parziale).

La rilevazione Oros diffonde indici, variazioni tendenziali e variazioni congiunturali relativi alle variabili che descrivono il costo del lavoro e il numero delle posizioni lavorative dipendenti. Indici e variazioni sono diffusi a livello di sezione Ateco 2007 e sue aggregazioni.

Gli indici sulle posizioni lavorative dipendenti (in breve posizioni totali) vengono prodotti e diffusi in aderenza alla definizione statistica del regolamento della Commissione europea n. 1503/2006. In base a tale definizione, si definisce "posizione lavorativa dipendente" (in inglese *job*) un contratto di lavoro tra una persona fisica e un'unità produttiva (impresa o istituzione privata), che prevede lo svolgimento di una prestazione lavorativa a fronte di un compenso (retribuzione). Le posizioni la-

[1] Una visione d'insieme sui progetti ESSnet è disponibile al link: https://ec.europa.eu/eurostat/cros/page/essnet_en.

vorative rappresentano, quindi, il numero di posti di lavoro occupati da lavoratori dipendenti[2], a tempo pieno e a tempo parziale, indipendentemente dalle ore lavorate, ad una determinata data di riferimento. Come il numero di occupati anche le posizioni lavorative rappresentano, pertanto, una variabile di stock in un certo istante nel tempo. Sono inclusi anche i lavoratori che, legati all'unità produttiva da regolare contratto di lavoro, sono temporaneamente assenti per cause varie quali ferie, permessi, maternità, cassa integrazione guadagni (CIG) e altre. Gli indici Oros sulle posizioni lavorative dipendenti vengono rilasciati a livello nazionale nella statistica Flash "Il mercato del lavoro" e nella banca dati I.Stat, a circa 70 giorni dalla fine del trimestre di riferimento. A partire da dicembre 2016, inoltre, i dati sulle posizioni dipendenti di Oros iniziano ad essere diffusi anche come livelli, nella "Nota trimestrale congiunta sulle tendenze dell'occupazione" diffusa da Istat in collaborazione con, Ministero del lavoro e delle politiche sociali, Inps, Inail e Anpal. I dati sulle posizioni lavorative prodotti dalla rilevazione vengono, infine, utilizzati anche per rispondere alle richieste del regolamento STS del Parlamento europeo e del Consiglio sulle statistiche congiunturali (n. 1165/98) che richiede indici grezzi trimestrali sul numero totale di persone occupate (*Number of persons employed*) che, accanto alle posizioni dipendenti, include gli indipendenti. Tale indicatore va rilasciato a 60 giorni dalla fine del trimestre di riferimento ed è riferito ad imprese ed istituzioni private nei settori Ateco 2007 da B ad N. Inoltre, vengono trimestralmente rilasciati alla Contabilità Nazionale che li utilizza come variabile ausiliaria per la stima dell'input di lavoro e alla rilevazione sui posti vacanti e le ore lavorate (Vela) per la quale rappresentano i totali noti nella procedura di calibrazione.

2.2 Fonti di dati

Le statistiche prodotte da Oros si basano sull'integrazione di due tipologie di fonti: dati amministrativi, utilizzati per la stima delle unità di piccola e media dimensione (PMI) e dati d'indagine, per la stima delle unità di grandi dimensioni (GI). La traduzione delle informazioni amministrative nei contenuti statistici rilevanti, sia di carattere economico sia strutturale, richiede complesse operazioni di trattamento, in cui si rende indispensabile il supporto di metadati legislativi e amministrativi aggiornati, nonché di fonti statistiche e amministrative ausiliarie, necessarie per conferire rilevanza ai dati statistici ricavati. Le informazioni che derivano dalle varie fonti vengono armonizzate in accordo con le richieste statistiche e combinate attraverso tecniche di *linkage* deterministico.

La principale fonte amministrativa su cui si basa la stima delle posizioni lavorative di rilevazione Oros sono le dichiarazioni che i datori di lavoro devono presentare all'Inps per denunciare le retribuzioni mensili corrisposte ai dipendenti, i contributi dovuti e l'eventuale conguaglio delle prestazioni anticipate per conto dell'Inps, delle agevolazioni e degli sgravi dei propri lavoratori dipendenti.

Ogni trimestre vengono acquisiti tre set di microdati dall'Inps:

1) le dichiarazioni contributive mensili che i datori di lavoro sono obbligati a presentare all'Istituto di Previdenza entro l'ultimo giorno del mese successivo a quello di competenza;
2) l'anagrafica delle posizioni contributive, in cui confluiscono informazioni sulle caratteristiche strutturali e contributive delle unità soggette alle dichiarazioni;
3) le ore autorizzate mensili di cig e solidarietà.

Il primo set di dati, ossia le dichiarazioni contributive ("DM2013 virtuali"[3]), contiene un'ampia varietà e quantità di informazioni su retribuzioni, occupazione (qualifica, tipo di contratto etc.) e componenti di costo del lavoro. Per motivi di tempestività, tali dati vengono acquisiti trimestralmente in forma grezza e disaggregata, senza alcun pretrattamento da parte dell'Inps, in due diverse versioni:

[2] Non rappresentano le "teste", ossia i lavoratori dipendenti, inquanto ad un singolo lavoratore può corrispondere più di una posizione lavorativa.

[3] Si tratta di dichiarazioni ricostruite virtualmente dall'Inps, a scopo amministrativo, a livello aziendale a partire dai flussi individuali Uniemens allo scopo di fornire alle imprese estratti contributivi sintetici.

a) un set "preliminare", riferito al trimestre corrente *t*, di cui i primi due mesi sono scaricati a circa 36-38 giorni dalla fine del trimestre di riferimento e l'ultimo mese a 43-45 giorni (figura 1). Tale modalità di acquisizione di dati implica un diverso grado di riempimento dei tre mesi di competenza. Considerato che la dichiarazione deve essere effettuata entro la fine del mese successivo a quello di competenza, il primo mese contiene le dichiarazioni arrivate nell'arco di circa 70 giorni dalla scadenza; il secondo di 40 giorni; il terzo mese, più ravvicinato rispetto alla scadenza ufficiale, pari a 15 giorni circa. Ne deriva che nei tre mesi vi sarà un numero di dichiarazioni ritardatarie di diversa entità, elemento da tenere in considerazione nel trattamento dei dati mancanti. Questo insieme di dati è utilizzato per la stima preliminare del trimestre *t*;

b) un insieme di DM2013 "finali" riferiti al trimestre *t*-4, scaricati con un ritardo di circa 1 anno dal trimestre di riferimento, utilizzati per calcolare le stime definitive di *t*-4 (figura 1). Tale set di dati è sostanzialmente una rappresentazione della popolazione totale poiché dopo un anno non vi sono dichiarazioni mancanti.

Accanto ai file sulle dichiarazioni contributive viene, inoltre, acquisito dall'Inps un file di anagrafica sulle singole posizioni contributive. Oltre a contenere informazioni sull'inquadramento dell'azienda in termini contributivi, necessarie per il calcolo del costo e dell'input di lavoro, in questo file sono presenti variabili identificative grazie a cui è possibile effettuare link con altri archivi statistici e amministrativi (codice fiscale, ragione sociale) per ottenere informazioni sulla configurazione giuridica ed economica dell'azienda al fine di collocare e classificare l'unità nel campo di osservazione della rilevazione (utilizzando variabili quali il codice statistico contributivo, la forma giuridica, e l'Ateco[4]), nonché dati utili a definire lo status di attività dell'unità (data di costituzione, data di cessazione/sospensione) ai fini delle stime provvisorie. Le informazioni anagrafiche contenute in questo file devono essere fornite dall'azienda all'Inps entro il 16 del mese successivo a quello della sua apertura e in occasione di variazioni sullo stato di attività della matricola stessa. Il file di anagrafica è scaricato trimestralmente a circa 25-28 giorni dalla fine del trimestre di riferimento *t* e le informazioni sono quelle presenti nel DB Inps di anagrafica alla data di scarico (figura 1). Mentre le informazioni finalizzate alla dichiarazione contributiva sono aggiornate molto rapidamente perché utilizzate nel calcolo dei contributi, altri problemi riguardano variabili secondarie per l'Inps ma utili al processo Oros quali, ad esempio, la data e il tipo di cessazione dell'attività, spesso comunicate in ritardo o mai aggiornate[5]. Ne consegue che l'anagrafica è caratterizzata da un rilevante problema di sovra-copertura[6] di tali eventi, problema che limita l'uso della fonte per la definizione della lista di unità attive da affiancare ai file sulle dichiarazioni mensili relative a *t* per le stime provvisorie. Tale limitazione, come si vedrà in seguito in questo lavoro, viene superata attraverso la predizione dello status di attività sulla base del profilo delle presenze/assenze delle singole unità in termini di presentazione delle dichiarazioni contributive in periodi contigui rispetto a quello di riferimento.

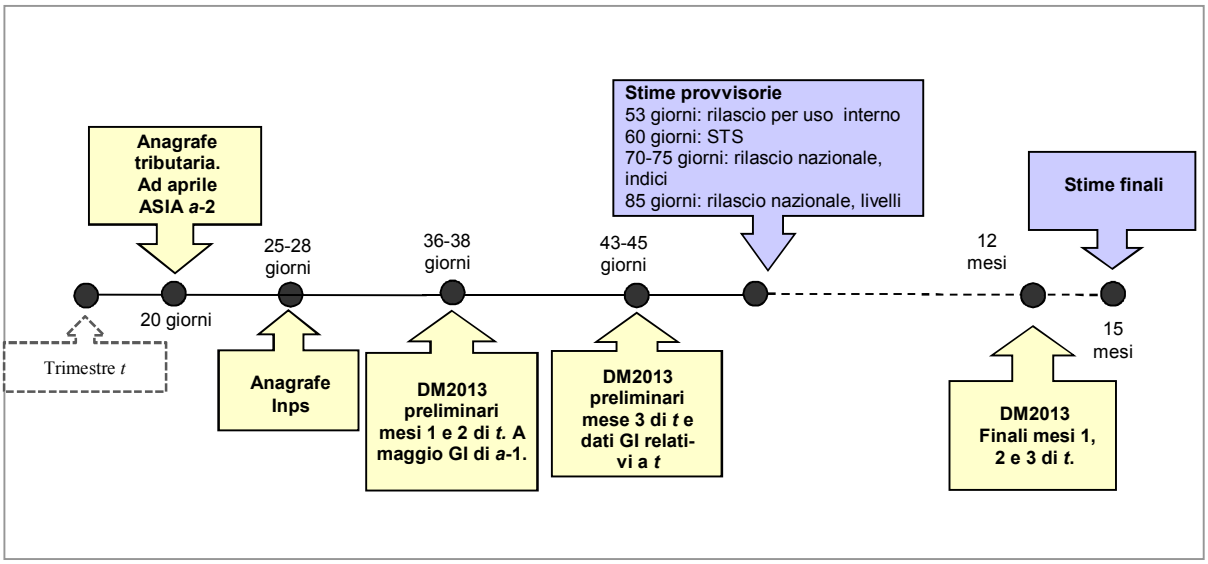
Infine l'Istat acquisisce trimestralmente un set di microdati mensili sulle ore autorizzate di cig per tipologia (ordinaria, straordinaria, deroga e solidarietà). Il file viene acquisito con un ritardo di circa 20 giorni dalla fine del trimestre di riferimento *t* e viene utilizzato per trattare statisticamente le posizioni lavorative dipendenti per correggere gli effetti di distorsione che la presenza di cig esercita sui principali indicatori rilasciati dalla rilevazione (Istat, 2015).

[4] A partire dal comunicato stampa di giugno 2015 le informazioni strutturali di classificazione nel campo di osservazione sono state bloccate. L'informazione fonte ASIA è bloccata all'anno 2012, fonte FINANZE e INPS all'anno 2014. Per le unità neonate a partire da giugno 2015 viene utilizzata invece l'informazione statistica o amministrativa più aggiornata disponibile. Le informazioni saranno aggiornate al prossimo cambio base 2015=100 previsto a giugno 2018.

[5] Le imprese non subiscono alcun procedimento amministrativo a seguito di dichiarazioni ritardatarie di questi eventi.

[6] Vanno segnalati anche casi di sotto-copertura che caratterizzano unità neonate che prima inviano la dichiarazione contributive e in seguito completano le procedure di iscrizione.

Figura 1 - Fonti di dati per la stima delle posizioni lavorative dipendenti di Oros e principali scadenze.



Fonte: Oros

Al fine di ricavare le informazioni statistiche rilevanti dai dati amministrativi, i microdati Inps devono, anzitutto, essere transcodificati attraverso un minuzioso uso dei codici amministrativi e di conseguenza essere sottoposti ad opportuni trattamenti di controllo e ricostruzione, con il supporto di metadati legislativi e amministrativi molto aggiornati, nonché di fonti statistiche e amministrative ausiliarie di supporto. A tal 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 (AA.VV, 2008).

La disponibilità nei dati amministrativi del codice fiscale tra gli identificativi delle unità consente di agganciare i dati Inps al Registro statistico delle imprese (Asia), all'archivio Istat delle Partite Iva, acquisiti annualmente, e all'Anagrafe tributaria dell'Agenzia delle Entrate, acquisita trimestralmente (a circa 20 giorni dalla fine del trimestre di riferimento *t*), permettendo in tal modo di ottenere informazioni statisticamente validate o adeguatamente aggiornate sul codice di attività economica e sulla natura giuridica e istituzionale delle unità.

Si ricorre, invece, a dati d'indagine per la stima delle variabili *target* sulle unità di grandi dimensioni. In particolare, si fa riferimento alla rilevazione mensile sull'occupazione, gli orari di lavoro, le retribuzioni e il costo del lavoro nelle grandi imprese (in breve GI). Un primo set di dati GI riferito a *t* viene acquisito a circa 43-45 giorni dalla fine del trimestre di riferimento (figura 1). Esso si riferisce alle stime preliminari sulle grandi imprese. Nel mese di maggio dell'anno *a*, successivamente alla revisione annuale della rilevazione GI, vengono acquisiti i dati finali relativi all'intero anno *a*-1. In origine, la necessità di sostituire i dati amministrativi con dati d'indagine era motivata da problemi di completezza dei dati preliminari, particolarmente rilevanti sulle imprese di grandi dimensioni. La progressiva evoluzione della dimensione dei dati preliminari verso quella dei dati finali a metà degli anni 2000 ha radicalmente cambiato le condizioni iniziali per cui anche le grandi imprese hanno iniziato ad avere una valida rappresentazione nei dati amministrativi. L'opportunità di continuare ad utilizzare la fonte statistica è stata quindi sollecitata dalla maggiore accuratezza sia delle variabili economiche sia di quelle strutturali, grazie ai controlli mirati svolti in fase di revisione dei dati raccolti da parte degli esperti d'indagine.

L'integrazione tra le due fonti richiede l'armonizzazione del contenuto informativo delle variabili e l'individuazione delle unità compresenti, al fine di escludere possibili duplicazioni. Il *linkage* fra le due fonti, che avviene trimestralmente, passa attraverso l'analisi delle frequenti trasformazioni giuridiche (scorpori, fusioni etc.) che tipicamente interessano le imprese di grandi dimensioni e che vengono rilevate in tempi diversi dalla fonte amministrativa e dalla rilevazione. Inoltre, in oc-

casione della revisione quinquennale del panel alla base della rilevazione GI, l'integrazione tra le due fonti richiede un'operazione straordinaria di re-identificazione dell'intera sottopopolazione GI nei dati amministrativi. I dati GI, coprono il 20% dell'occupazione Oros, di cui oltre 6 punti percentuali sono presenti nell'industria e poco più di 13 punti nei servizi. I settori in cui i dati di rilevazione hanno maggior peso sono quelli della fornitura di energia (70%), delle attività finanziarie e assicurative (66%), del trasporto e magazzinaggio (41%) e dei servizi di informazione e comunicazione (38%). Sono, invece, assenti dai servizi sociali e personali (P-S), settori rilevati a partire dall'anno 2015 e quindi rilasciati con la diffusione dei dati nella prossima base 2015=100, ad aprile 2018.

2.3 Stime

2.3.1 Stima preliminare e stima finale

Sulla base delle informazioni disponibili, la rilevazione Oros rilascia ogni trimestre una stima preliminare su *t* e una finale su *t*-4. La stima finale differisce da quella preliminare a seguito della disponibilità di informazioni più complete ed aggiornate che si rendono disponibili nel frattempo, quali ad esempio:

- la disponibilità dell'universo delle dichiarazioni DM2013 virtuali per la produzione della stima definitiva;
- la revisione dei dati della rilevazione GI;
- l'aggiornamento di informazioni di carattere strutturale sulle unità oggetto di rilevazione;
- le eventuali revisioni occasionali nella metodologia di stima degli indicatori;
- piccole revisioni nei dati amministrativi di alcune unità rispetto ai dati già forniti in stima preliminare.

Revisioni di maggiore entità caratterizzano normalmente la diffusione di giugno, in occasione della quale la rilevazione GI rivede i dati relativi all'intero anno *a*-1, per incorporare i rispondenti ritardatari.

Ai fini della stima finale i dati amministrativi di fonte Inps rappresentano la popolazione *target* (cfr. §2.2) rendendo possibile, sin dal primo utilizzo di questa fonte, l'applicazione di una metodologia estremamente semplificata basata sull'aggregazione dei dati disponibili. Stessa procedura riguarda i dati di fonte GI, in cui la stima per aggregazione riguarda sia le stime finali sia le stime provvisorie. Le stime provvisorie delle PMI risentono, invece, della struttura dei dati amministrativi preliminari alla base che, alla data dell'acquisizione, sono incompleti. I frequenti mutamenti amministrativi, legislativi e tecnici che si sono susseguiti nel tempo hanno radicalmente condizionato l'individuazione, la sperimentazione e, infine, la scelta dell'approccio di stima. Attualmente, la ridotta incidenza dei rispondenti ritardatari comporta un effetto di distorsione non significativo sulle variabili rapporto di costo del lavoro, limitando il ricorso all'imputazione dei dati mancanti solo ad un insieme ridotto di unità influenti individuate con criteri di selettività. Diverso è il caso del numero delle posizioni lavorative dipendenti, sulle quali la sottostima dovuta ai dati mancanti ha un impatto tale da richiedere una procedura d'imputazione per mancate risposte su tutte le unità. Tale argomento verrà diffusamente trattato in questo documento.

Un trattamento a parte, inoltre, è riservato alla stima provvisoria delle variabili *target* per le agenzie di somministrazione di lavoro interinale (in breve INTER), considerato il rilevante peso occupazionale che tali unità rivestono sul totale dell'economia, suddiviso prevalentemente tra poche grandi imprese. In particolare, il loro peso è pari al 93% circa nella divisione N78, 21% nella sezione N (Noleggio, agenzie di viaggio, servizi di supporto alle imprese) in cui sono classificati e poco meno del 2% nell'occupazione totale di Oros. Per tali unità, si ricorre tradizionalmente ad un metodo d'imputazione di mancate risposte applicato a livello micro, caratterizzato dalla ricostruzione di una lista di rispondenti ritardatari e di successiva imputazione del dato mancante, tenendo fortemente in considerazione le informazioni longitudinali disponibili sulle singole unità.

Infine, un'operazione aggiuntiva sui dati amministrativi (preliminari e finali) si rende necessaria al fine di migliorare l'accuratezza della definizione della variabile posizioni lavorative dipendenti,

caratterizzata da un bias definitorio, in quanto la variabile posizioni retribuite, già usata nel calcolo delle Ula al denominatore delle variabili di costo del lavoro non include le informazioni relative ai dipendenti non retribuiti, poiché assenti per aspettativa di breve durata e collocamento in cig per l'intero mese, inclusi invece nella definizione statistica della variabile. In tal caso si interviene misurando la variabile obiettivo con il supporto di informazioni ausiliarie (Istat, 2015). Tale argomento non verrà trattato in questo documento.

Di seguito, viene illustrata l'evoluzione della situazione informativa dei dati provvisori e le scelte metodologiche che i vari cambiamenti hanno dettato, sino ad arrivare all'attuale metodologia di stima.

2.3.2 I cambiamenti della situazione informativa e il loro impatto sulla metodologia di stima preliminare

A partire dalla prima acquisizione dei dati a scopi sperimentali, alla fine degli anni '90, la fonte Inps sulle dichiarazioni contributive è stata interessata da diversi e radicali cambiamenti che hanno avuto un impatto rilevante soprattutto sulla metodologia di stima provvisoria (figura 2). Tali cambiamenti, infatti, hanno influenzato la disponibilità di micro dati in tempi rapidi e solo marginalmente intaccato i dati disponibili dopo un anno utilizzati per le stime definitive.

Il più grande cambiamento risale all'inizio del 2004 quando l'invio dei DM10 da parte di tutte le imprese per via telematica da facoltativo (campione auto-selezionato di unità) è divenuto obbligatorio. Tale cambiamento ha comportato un ampliamento notevole dei dati disponibili per le stime provvisorie che, in poco tempo, dal 50% circa di copertura, hanno raggiunto il 90-95%, approssimando la popolazione finale, disponibile dopo un anno (figura 3). Questo cambiamento ha comportato una significativa semplificazione della procedure di stima preliminare. In particolare, prima di aprile 2004, la popolazione provvisoria era considerata un campione non-casuale, consistente in un set di rispondenti auto-selezionati (coloro, appunto, che adottavano la modalità elettronica di invio delle dichiarazioni contributive) (fase 1). La numerosità di tale campione andava via via aumentando nel tempo e comunque rappresentava una buona copertura delle caratteristiche dalla popolazione di riferimento, includendo anche un adeguato numero di nuove nate. In questo contesto informativo, in assenza di un *framework* di stima basato su disegno campionario, la stima preliminare avveniva mediante procedura di calibrazione sulla base di un approccio *model-assisted* (Baldi et al., 2004). Il problema della selettività del campione veniva risolto individuando i pesi all'interno di sottogruppi omogenei (*model groups*) per caratteristiche della popolazione (attività economica, dimensione, area geografica, età dell'impresa). La procedura di calibrazione ha fornito risultati accettabili nella stima delle variabili rapporto, ma si è rivelata scarsamente adeguata nella stima delle variabili espresse come totali, quali il numero delle posizioni lavorative: il meccanismo non ignorabile caratterizzante i rispondenti ritardatari, i continui cambiamenti nella rappresentatività del campione, insieme al problema della sovra-copertura del registro amministrativo hanno implicato stime poco accurate di questa variabile, richiedendo azioni ad hoc supplementari basate su interventi di tipo macro.

A partire da aprile 2004, rendendosi disponibile oltre il 90% di popolazione attiva per le stime anticipate, il metodo di stima preliminare è stato assimilato a quello delle stime finali, ossia per aggregazione dei dati disponibili (fase 2, figura 3). Pur garantendo stime con un grado di distorsione accettabile per le variabili rapporto, secondo questo approccio rimane irrisolto il problema della sottostima delle variabili totali, quali il numero delle posizioni lavorative, a causa delle dichiarazioni ritardatarie. Per la stima di tale variabile si è continuato ad utilizzare un intervento macro. Nel corso del tempo, la stima per approccio macro è stata profondamente migliorata grazie allo sfruttamento di informazioni sull'errore di revisione passato (che per alcuni aggregati mostra una componente sistematica e predicibile) e studiando la relazione tra la dinamica dell'occupazione nei dati preliminari e finali (Istat, 2013 § 4.6.2). Per un certo periodo di tempo la situazione informativa sembrava aver raggiunto un considerevole livello di stabilità, inducendo la sperimentazione di una metodologia di stima basata sulla ricostruzione dei microdati: la stabilità dei dati e la non persistenza delle mancate risposte ha indotto a delineare il problema quale una questione di *wave-non re-*

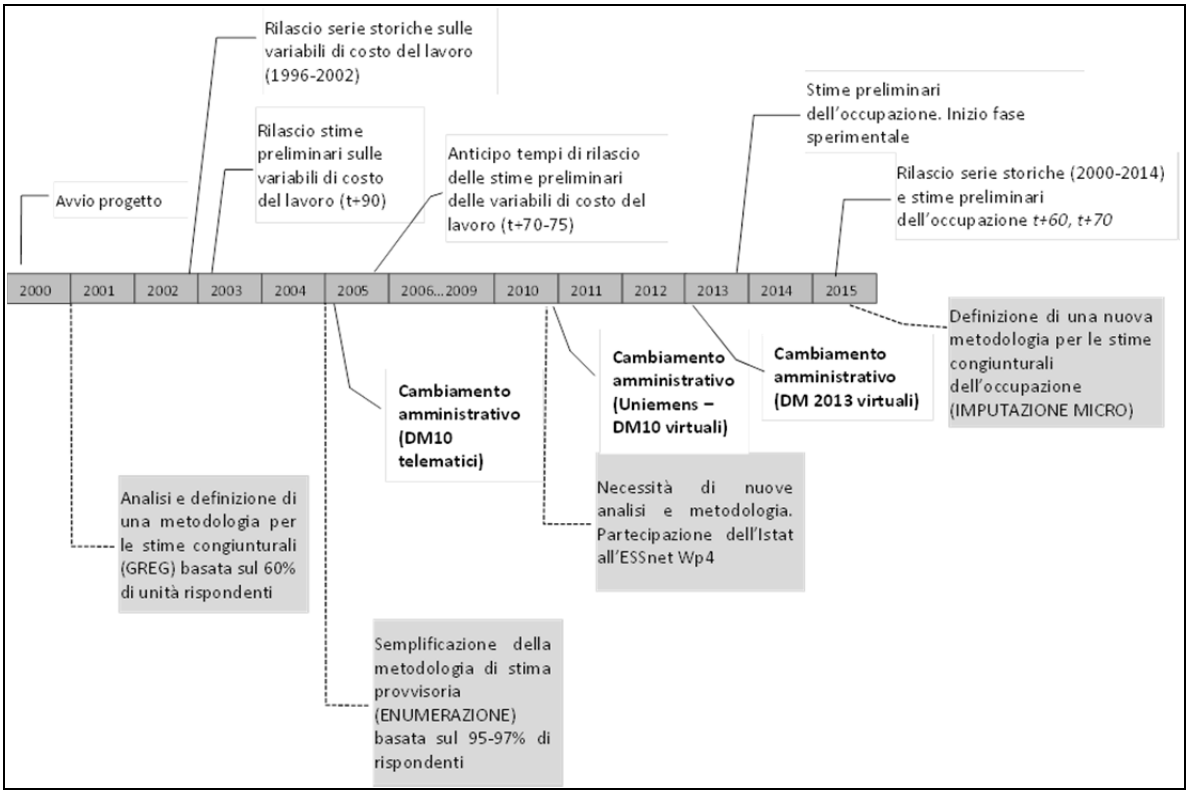
sponse, (Kasprzyk et al., 1989) in cui il comportamento del passato di ogni rispondente corrente è informativo rispetto alla sua realizzazione nella popolazione finale. Anche nella ricostruzione dei micro dati, l'informazione passata sui rispondenti è informativa sul dato corrente mancante.

Prima che il nuovo metodo di stima preliminare potesse essere messo a punto, a gennaio 2010 si è verificato un ulteriore grande cambiamento: l'abbandono del precedente sistema di dichiarazione basato su modelli da compilare a livello d'impresa per passare a una nuova e più complessa modulistica (Uniemens) contenente informazioni a livello di lavoratore e alcune informazioni di sintesi a livello di impresa (fase 3, figure 2 e 3). Nei mesi successivi, a questo cambiamento radicale, si sono susseguiti alcuni ulteriori interventi finalizzati a razionalizzare l'acquisizione dell'informazione amministrativa. L'Uniemens rappresenta una fonte di informazioni ricchissima per l'Istituto, anche a fini congiunturali, ma tradurre la nuova fonte di dati in informazioni statistiche avrebbe richiesto un periodo di esplorazione molto lungo, da investire principalmente nella ricostruzione dei metadati necessari, profondamente cambiati rispetto al vecchio sistema. Ai fini di Oros, l'Istat ha potuto continuare ad acquisire i DM10 virtuali, un'aggregazione degli Uniemens nel formato dei vecchi DM10 che l'Inps ha contestualmente iniziato a ricostruire per finalità amministrative interne. In questo contesto sono state riavviate le vecchie sperimentazioni che hanno dovuto tener conto della situazione informativa in cui, inizialmente, non solo era cresciuto il numero di rispondenti ritardatari, ma ne era mutata anche la caratterizzazione: inizialmente erano le imprese più grandi a risultare più frequentemente incomplete a causa, probabilmente, della maggiore difficoltà implicita nei nuovi modelli. Al contrario, tuttavia, si è osservato un immediato miglioramento della qualità dei microdati in termini di "errore di misura" a seguito, probabilmente, delle procedure più vincolanti introdotte dall'Inps nella compilazione dei nuovi modelli.

Infine, a partire da gennaio 2013 l'Inps ha affrontato un altro importate cambiamento, consistente nell'abbandono dei DM10 virtuali per passare ai nuovi DM2013 virtuali una ricostruzione semplificata e più efficiente, a livello d'impresa, per usi interni amministrativi (fase 3bis, figure 2 e 3). Questo nuovo e inaspettato cambiamento ha comportato diverse nuove difficoltà per Oros che ha dovuto adattare in tempi molto rapidi le complesse procedure per produrre entro le scadenze previste i principali *output*. Il *team* di Oros sta ancora lavorando alla semplificazione, riorganizzazione delle procedure per esplorare più efficacemente le informazioni disponibili sul nuovo DM2013 che comporterà nel prossimo futuro alcuni progressi nella produzione delle variabili d'interesse, verso la transizione in un più lungo periodo all'uso delle informazioni a livello di lavoratore degli Uniemens.

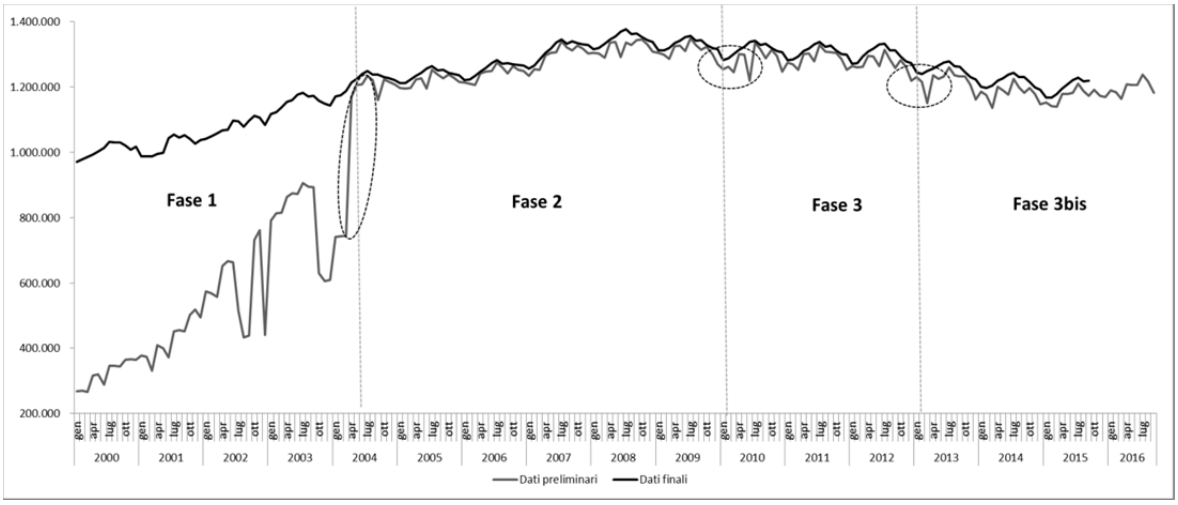
A partire da inizio 2013, basandosi su un'esperienza maturata a livello europeo (Maasing et al., 2012), è stata avviata una nuova fase di sperimentazioni di un metodo di imputazione di microdati finalizzato alla stima del numero delle posizioni lavorative dipendenti che ha avuto termine a metà 2015 con la messa a regime e diffusione, a partire da giugno dello stesso anno, di dati sull'occupazione anche a livello nazionale. Questo documento è dedicato alla descrizione della sperimentazione attuata a partire dagli spunti dell'esperienza del progetto Europeo ESSnet fino alle specifiche soluzioni metodologiche applicate per tener adeguatamente conto delle peculiarità della situazione informativa di Oros.

Figura 2 – Le tappe della rilevazione Oros: dall'avvio del progetto ai vari cambiamenti alla metodologia di stima alla base degli indicatori diffusi.



Fonte: Oros

Figura 3 - Numero di dichiarazioni contributive Inps nei dati preliminari e nei dati finali. Gennaio 2000-giugno 2016.



Fonte: Elaborazione su dati Oros

3. Il contesto informativo delle stime preliminari: caratteristiche delle dichiarazioni contributive ritardatarie

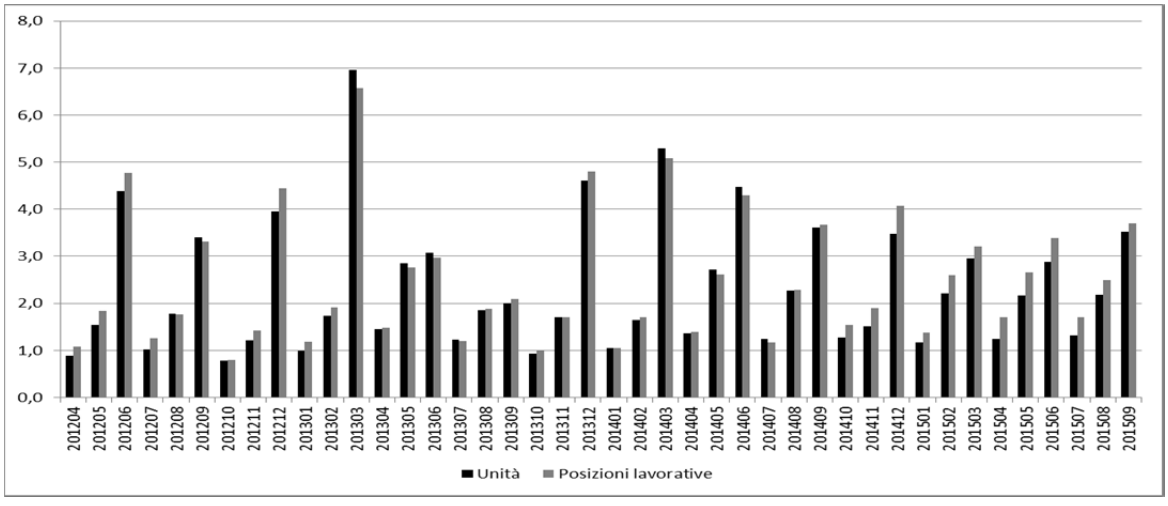
Le stime preliminari di Oros sono il risultato dell'integrazione tra i dati della rilevazione GI considerati nella versione provvisoria, acquisiti già corretti per le mancate risposte e i dati amministrativi sulle dichiarazioni contributive Inps, nella versione acquisita alla fine del trimestre di riferimento *t*. I dati Inps utili per le stime provvisorie sono incompleti e il loro grado di riempimento dipende strettamente dalla schedulazione dei tempi di acquisizione. In particolare, come visto nel §2.2 i tre mesi del trimestre vengono scaricati con tempistiche diverse:

- il mese 1 e il mese 2, scaricati a circa 36-38 giorni dalla fine del trimestre di riferimento, contengono le dichiarazioni pervenute rispettivamente entro 70 e 40 giorni dalla scadenza d'invio prevista per legge;
- il mese 3 a 43-45 giorni, con dichiarazioni pervenute nell'arco di 15 giorni circa.

Di seguito vengono riportate alcune distribuzioni relative alle dichiarazioni contributive ritardatarie per la sottopopolazione delle PMI. Vengono escluse da questa analisi le agenzie di somministrazione di lavoro interinale, trattate con metodologia *ad hoc*, come già detto poco sopra. Il periodo preso in considerazione va da aprile 2012 a settembre 2015[7].

Il grafico illustra l'incidenza delle risposte ritardatarie rispetto al totale dei rispondenti nei dati di stima definitiva, in termini di numero di dichiarazioni contributive e di posizioni lavorative. I DM ritardatari rappresentano una quota che oscilla tra l'1%, normalmente rilevato nel primo mese dei vari trimestri e il 4-5% nei terzi mesi (figura 4). Fa eccezione il mese di marzo 2013 in cui, a seguito dell'entrata a regime del DM2013 virtuale in sostituzione del DM10 virtuale (cfr. § 2.3.2), il flusso Inps ha subito un rallentamento, comportando una momentanea crescita di rispondenti ritardatari (7%). In media, l'incidenza dei ritardatari è lievemente più alta se valutata in termini di posizioni lavorative segnando una maggiore propensione a ritardare da parte delle unità di dimensione mediamente più grande. Inoltre, si nota con chiarezza la maggior percentuale di ritardatari nei terzi mesi dei trimestri. Le più alte incidenze si osservano, inoltre, nei mesi di giugno: a giustificare il maggior ritardo, la minor presenza di personale amministrativo dedicato alla compilazione dei modelli di dichiarazione contributiva per ferie (la scadenza per il mese di competenza di giugno corrisponde alla fine di luglio).

Figura 4 - Serie storica dell'incidenza delle mancate risposte in termini di numero di unità e di posizioni lavorative. Aprile 2012 - settembre 2015.



Fonte: Elaborazione su dati Oros

[7] L'analisi del contesto informativo è stata effettuata su un periodo più lungo rispetto a quello dei dati di sperimentazione, in cui i dati del 2015 non erano disponibili.

Pur apparendo uniforme sui macro settori dell'industria, dei servizi di mercato e dei servizi sociali e personali, la distribuzione settoriale dei rispondenti ritardatari appare più concentrata in alcune sezioni dell'Ateco 2007 (tavola 1). I più caratterizzati sono il settore H dei trasporti e magazzinaggio (3,8% in termini di unità e 4,3% di posizioni), il settore N del noleggio, agenzie di viaggio, servizi di supporto alle imprese (3,1% e 3,9% rispettivamente), il settore R delle attività artistiche, sportive, di intrattenimento e divertimento (3,9% e 5,6%). Nell'ambito della sezione H è particolarmente rilevante la mancata risposta nella divisione H50 del trasporto marittimo, per effetto di una deroga amministrativa sull'invio dei moduli che prevede la proroga a 60 giorni invece che 30 per alcune categorie di lavoratori[8] (18,9% in termini di unità e 25,2 di posizioni). Inoltre, la mancata risposta è marcata per la divisione J59 relativa alla produzione cinematografica e televisiva (6,5% in termini di unità e 14% di posizioni) e per tutte le divisioni della sezione R.

Tavola 1 - Incidenza per numero di unità e di posizioni lavorative dei rispondenti ritardatari per sezione Ateco 2007 e per classe dimensionale. Aprile 2012 - settembre 2015.

	Unità	Posizioni lavorative
Sezioni Ateco		
B	1,8	1,7
C	1,9	1,8
D	2,1	2,6
E	2,2	2,9
F	2,6	2,4
G	2,2	2,0
H	3,8	4,3
I	2,6	2,6
J	2,6	3,6
K	1,9	1,9
L	2,0	2,3
M	1,8	2,4
N	3,1	3,9
P	2,6	3,2
Q	2,0	2,8
R	3,9	5,6
S	2,4	2,5
Industria in senso stretto (B-E)	1,9	1,8
Industria (B-F)	2,2	2,0
Servizi di Mercato (G-N)	2,4	2,8
Servizi sociali e personali (P-S)	2,4	3,1
Totale B-N	2,3	2,4
Totale B-S	2,3	2,5
Divisioni Ateco con mancata risposta elevata		
C-12	18,5	5,9
E-39	3,2	5,0
H-50	18,9	25,2
H-51	3,4	5,8
J-59	6,5	14,0
J-60	5,2	7,5
R-90	4,1	6,5
R-91	2,7	7,4
R-92	3,8	5,5
R-93	3,9	5,1

Fonte: Elaborazione su dati Oros

[8] Si veda al proposito la circolare Inps 179 del 23-12-2013.

Tavola 1 (segue) - Incidenza per numero di unità e di posizioni lavorative dei rispondenti ritardatari per sezione Ateco 2007 e per classe dimensionale. Aprile 2012 - settembre 2015.

	Unità	Posizioni lavorative
Classe dimensionale		
0-49	2,3	2,2
50-249	2,6	2,6
250-499	3,3	3,4
500+	4,4	4,8
Totale	2,3	2,5

Fonte: Elaborazione su dati Oros

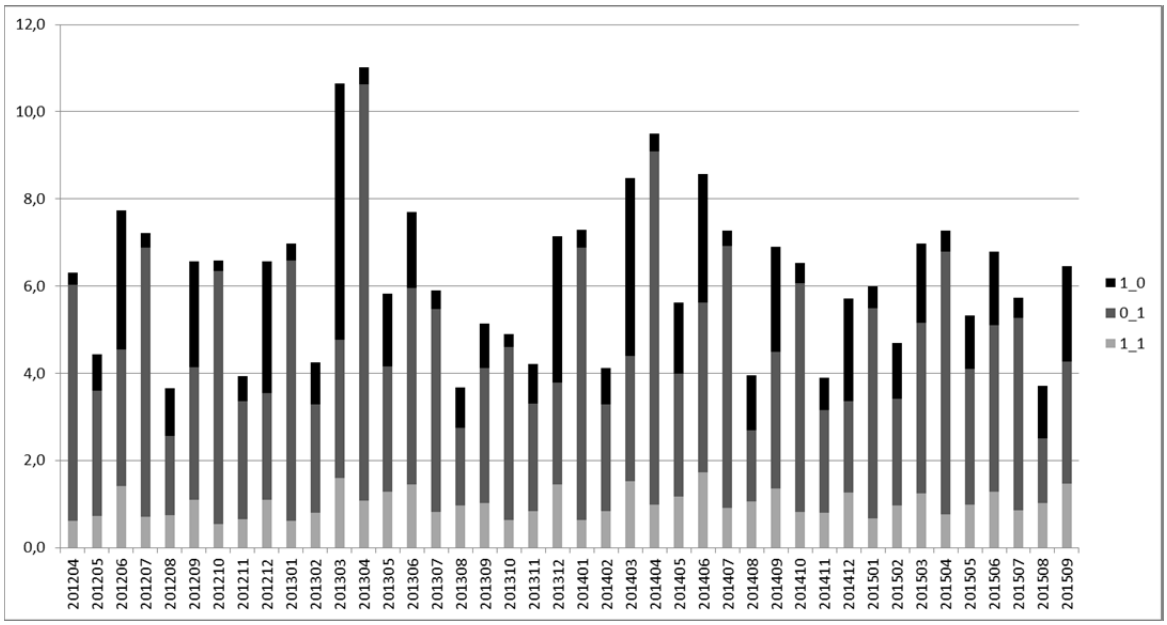
Appare, invece, più caratterizzata la distribuzione per classe dimensionale, che vede la propensione a ritardare al crescere della dimensione dell'azienda. Questa evidenza può essere attribuita alla maggiore complessità nella gestione dei flussi Uniemens con il crescere del numero di dipendenti, per i quali devono essere presentate le dichiarazioni contributive: l'incidenza delle ritardatarie passa da poco più di 2 punti per le imprese fino a 50 dipendenti a quasi 5 per le unità con oltre 500 dipendenti[9].

Ad esclusione dei casi di deroghe alla scadenza di presentazione delle dichiarazioni contributive, nei dati utilizzati non si osserva una particolare persistenza al ritardo da parte delle unità. Nella figura 5 si riporta l'incidenza delle unità presenti, misurata nella popolazione finale, per *pattern* di presenza nel mese *m* e nel mese *m*-1 nei dati provvisori (in particolare, con 1= mancata risposta e con 0= rispondente). Dalla rappresentazione grafica si esclude il caso delle unità presenti nei dati provvisori sia a *m* che a *m*-1, che rappresentano oltre il 90% del totale. Le tre modalità riportate stanno a rappresentare rispettivamente: 1_0 mancata risposta ad *m* e rispondente ad *m*-1; 0_1 rispondente ad *m* e mancata risposta ad *m*-1; 1_1 mancata risposta ad *m* e ad *m*-1. I casi più frequenti sono i rispondenti ad *m* che risultano assenti a *m*-1, ossia 0_1. Il profilo di tale incidenza presenta picchi sul primo mese di ogni trimestre in cui, in condizioni normali, assume valori tra il 5 e il 6% e si riduce gradualmente passando al secondo e al terzo con valori di 3% circa. Su tale percentuale le assenze giustificate da eventi demografici, ossia dovute a sospensione dell'attività ad *m*-1 o nascita dell'attività ad *m* pesano rispettivamente, per 0,3%-0,2% e per 0,3%-0,6%.

Segue la modalità 1_0 ossia mancate risposte ad *m* (la cui assenza non è, quindi, giustificata da eventi demografici) ma rispondenti ad *m*-1, con incidenza che sale dal primo al terzo mese da 0,5% a circa 2%. Infine, meno diffusa l'incidenza delle mancate risposte persistenti, ossia assenti sia ad *m* che ad *m*-1, (modalità 1_1), con una percentuale che cresce lievemente dal primo al terzo mese da circa 0,5% a poco più di 1 punto e, tra queste, le assenti per eventi demografici ad *m*-1 hanno un'incidenza praticamente nulla (persistenza reale). Ad influenzare i profili delineati è la tempistica di scarico dei dati che implica, come visto nel §2.2, una maggiore concentrazione di dichiarazioni ritardatarie nel terzo mese di ogni trimestre, quello acquisito a minore distanza temporale rispetto alla scadenza ufficiale.

9 Si tratta di unità non rilevate dalla rilevazione GI perché fuori dal campo di osservazione della rilevazione o neonate dopo l'anno base e, quindi, stimate con dati amministrativi.

Figura 5 – Distribuzione unità rispondenti ad *m* nei dati finali per pattern di presenza nei dati provvisori ad *m* e ad *m*-1. Valori percentuali.



Fonte: Elaborazione su dati Oros

4. L'imputazione delle mancate risposte: metodologia e sperimentazione

Come visto nel paragrafo precedente, la stima preliminare delle variabili *target* di Oros per le PMI si basa, in condizioni di stabilità del flusso informativo, su un insieme di rispondenti che incide, in termini di occupazione media nel trimestre, per il 95-98%. Pur soggetta ad improvvise cadute anomale a seguito di eventi particolari tale percentuale appare rilevante e non particolarmente caratterizzata a livello strutturale. In tale situazione informativa, la distribuzione dei modelli ritardatari non comporta distorsioni significative sulla stima delle variabili rapporto rilasciate dalla rilevazione (retribuzioni e oneri per unità di lavoro equivalenti a tempo pieno - Ula). Tali variabili, infatti, per loro natura presentano distribuzioni piuttosto uniformi tra le unità della popolazione di riferimento e ciò implica, in presenza unicamente di mancate risposte totali, un effetto trascurabile di distorsione sulle stime inducendo, quindi, a non intervenire con alcuna forma di correzione per i dati mancanti, se non in situazioni particolari (sottopopolazione delle imprese interinali ed editing selettivo).

L'impatto delle mancate risposte è, invece, rilevante sulle variabili totali, quali l'occupazione, rendendo indispensabile un processo di imputazione. Nel corso degli ultimi anni sono state condotte diverse sperimentazioni per la correzione delle stime provvisorie sull'occupazione, contrapponendo e valutando vantaggi e svantaggi di approcci macro e micro di ricostruzione dei dati mancanti. L'esperienza ha evidenziato come gli approcci macro, basati sulla regolarità in serie storica della dinamica dell'indicatore, pur se più semplici da implementare e meno pesanti da controllare, rischiano di essere deboli nelle fasi di svolta del ciclo economico, limitando notevolmente l'utilizzo dei dati disponibili (Baldi et al. 2011b). Ciò a fronte di crescenti richieste sia di dati aggregati/disaggregati a diverso livello di dettaglio rispetto a quello disponibile, sia di microdati completi. Queste constatazioni hanno stimolato a muoversi verso la progettazione di un approccio di tipo micro, in grado di fornire stime a livelli di classificazione più fini, ma anche di sfruttare le informazioni più specifiche disponibili sulle singole unità e consentendo di tenere meglio sotto controllo anche particolari problemi che possono sorgere nel processo di produzione dovuti, ad esempio, ad errori di misura. Nel propendere verso un approccio micro non bisogna trascurarne i potenziali svantaggi. In particolare, la difficoltà di modellare comportamenti in casi di gruppi eterogenei di

unità, la presenza di *missing* o *breaks* nei microdati che restringe l'uso di lunghe serie storiche, la maggiore incombenza delle fasi di controllo.

4.1 Cenni sulle soluzioni del passato ed esperienza di Oros nell'ESSnet Wp4: principali lezioni imparate

Al fine di rispondere alle richieste del regolamento STS e per fornire una serie di output ad utilizzatori interni, fino a febbraio 2015 la stima provvisoria del numero delle posizioni lavorative avveniva utilizzando un approccio macro, in cui la correzione per i dati mancanti avveniva su dati aggregati a livello di divisione Ateco. L'approccio, di tipo empirico, era basato su analisi e valutazioni dell'indicatore in serie storica e su relazioni ricorrenti con indicatori *proxy* e sfruttando, per molti settori, la ricorrenza degli errori di revisione calcolati sui trimestri di stima finale (Istat, 2013). In situazioni di ciclo economico stabile, tale approccio consentiva di ottenere stime di buona qualità.

In caso di variazioni improvvise del numero di rispondenti per cause di tipo "amministrativo", o in corrispondenza di punti di svolta del ciclo economico, tale approccio appariva meno robusto.

Nella nuova procedura di correzione in cui viene applicata un'imputazione di tipo micro, tali metodi di carattere empirico sono comunque utilizzati nella fase finale di validazione dei dati, ma ora il loro impatto è di lieve entità rispetto a prima e riguarda soprattutto alcuni settori che notoriamente hanno più criticità rispetto ad altri, come i trasporti marittimi che sistematicamente consegnano in ritardo i modelli amministrativi (cfr. §3), oppure altri settori in cui vi è una persistente volatilità della variabile occupazionale senza essere settori stagionali, come ad esempio le attività nel settore della produzione cinematografica e televisiva, che registrano importanti oscillazioni di dipendenti tra un trimestre e l'altro.

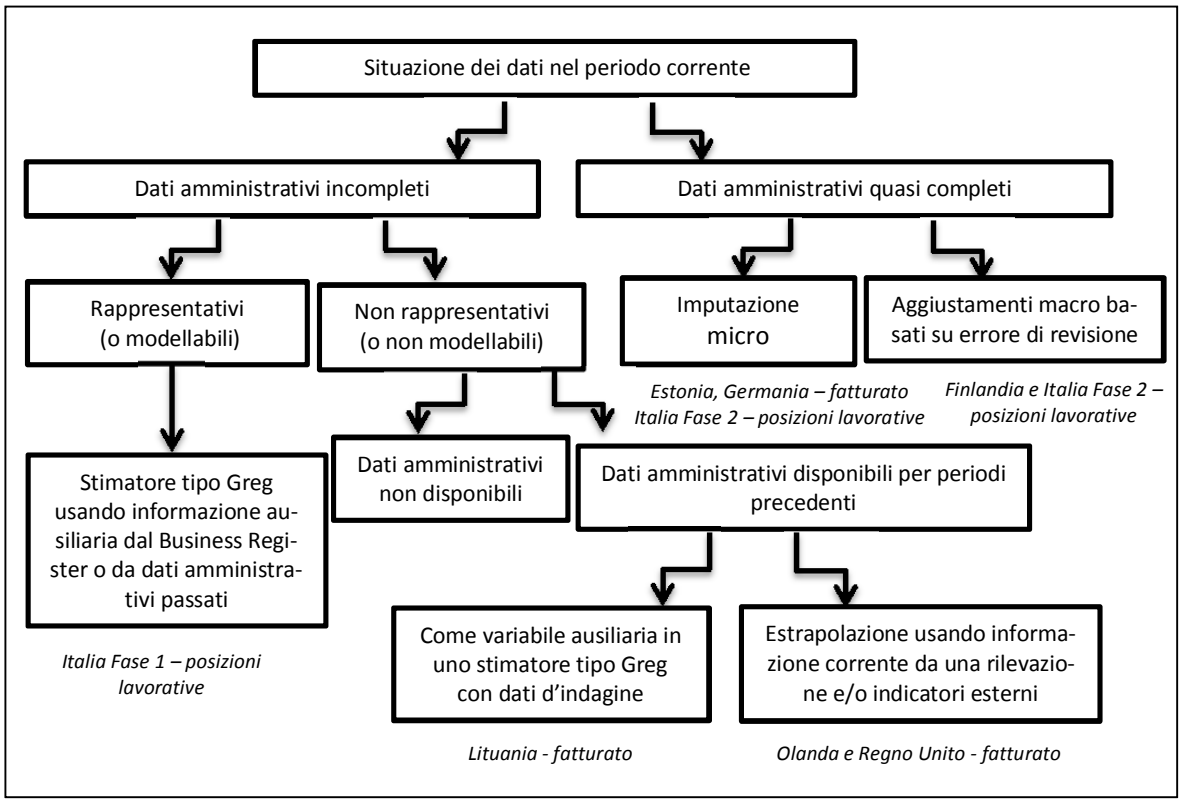
Uno slancio molto importante verso la definizione dell'attuale metodo si è avuto a seguito della partecipazione dell'Italia al progetto Europeo ESSnet "*Use of administrative and accounts data for business statistics*" *Working Package 4* – "T*imeliness of administrative sources for monthly and quarterly estimates*", nel corso degli anni 2009-2013 in cui l'Istat ha riportato la propria esperienza sull'uso di dati amministrativi per scopi congiunturali, illustrando e proponendo come test il caso Oros[10]. Insieme all'Italia hanno partecipato ai lavori del gruppo: Olanda (paese coordinatore), Germania, Estonia, Lituania, Regno Unito e Finlandia.

Obiettivo principale del progetto è stato quello di fornire *best practices* e raccomandazioni sui metodi che possono essere utilizzati per produrre stime congiunturali accettabili in presenza di dati amministrativi incompleti rispetto ai tempi di diffusione (Baldi et. al, 2012). Il focus è stato rivolto principalmente alla stima di due variabili richieste dal regolamento europeo STS ossia il fatturato, stimato sulla base di dati sulle Imposte sul Valore Aggiunto e le posizioni lavorative dipendenti, sulla base di dati di Sicurezza Sociale.

Il gruppo ha lavorato preliminarmente ad una rassegna delle applicazioni già in produzione o in corso di sviluppo negli Istituti di Statistica, al fine di mappare le situazioni più comuni di disponibilità di dati amministrativi e i relativi metodi utilizzati. A seguito della discussione di proposte di miglioramenti metodologici e realizzazione di test e confronti dei risultati fra paesi, si è passati alla formulazione di raccomandazioni. La figura che segue illustra sinteticamente le principali proposte suggerite dal gruppo sulla base della situazione informativa disponibile.

[10] La documentazione sui principali risultati del progetto è disponibili al link: http://ec.europa.eu/eurostat/cros/content/admindata-essnet-use-administrative-and-accounts-data-business-statistics_en.

Figura 6 - Quadro sulla tipologia di fonte dati e relativi modelli proposti nell'ambito del gruppo Essnet, WP4.



Fonte: Essnet WP4

Rispetto alla disponibilità di dati amministrativi riferiti al periodo corrente si distinguono due principali situazioni: 1) i dati amministrativi sono significativamente incompleti al periodo di compilazione delle stime STS; 2) i dati amministravi sono quasi completi.

Nel primo caso i dati amministrativi, malgrado incompleti, potrebbero essere rappresentativi o meno della popolazione *target*. Se rappresentativi, o modellabili per diventare tali, una stima di tipo *model-assisted* (es. tipo Greg) è la soluzione metodologica ritenuta migliore, in cui i dati amministrativi o da registro, anche riferiti a periodi precedenti, possono essere utilizzati come variabili ausiliarie nel modello, congiuntamente ad una rilevazione statistica che copre l'intera popolazione *target* (caso della Lituania per la stima del fatturato). Nel caso in cui, invece, i dati amministrativi del periodo non sono rappresentativi, il suggerimento è quello di utilizzare i dati amministrativi riferiti a periodi precedenti (quando completi) come variabili ausiliarie in contesti di stima per regressione o calibrazione congiuntamente a una piccola rilevazione statistica che copre solo una sottopopolazione (caso dell'Olanda per la stima del valore aggiunto) o in un contesto di tecniche di estrapolazione su dati aggregati utilizzando l'informazione corrente che deriva da indicatori esterni correlati o da una rilevazione (caso del Regno Unito per la stima del fatturato).

Nel secondo caso di dati amministrativi quasi completi emergono due possibilità. In un caso, i dati disponibili possono essere corretti a livello aggregato mediante una procedura di aggiustamento che sfrutta un'eventuale sistematicità degli errori di revisione rilevati in periodi precedenti (soluzione adottata nella fase 2 dall'Italia per la stima del numero delle posizioni lavorative e dalla Finlandia per la stima del fatturato). Una seconda possibilità è di indentificare i valori *missing* e imputarli a livello micro, la soluzione scelta dall'Italia nella fase 2 a sostituzione del metodo macro precedentemente utilizzato.

I punti cardine della procedura micro così come definiti nei lavori dell'ESSnet, anche sulla base di sperimentazioni condotte da vari paesi, tra cui l'Italia e poi ripresi nell'applicazione attuale sono principalmente due: la definizione della popolazione attiva del periodo corrente e il modello/regola

d'imputazione da applicare sulle unità considerate ritardatarie. In tale contesto, per individuare le migliori soluzioni sono state considerate, valutate criticamente e confrontate diverse alternative (attraverso test e analisi degli errori di revisione), cercando di convergere verso soluzioni generalizzate, il più possibile indipendenti dalla specificità del contesto informativo disponibile.

In linea generale, l'esperienza ESSnet ha fatto maturare alcune importanti consapevolezze sull'uso dei dati amministrativi nella compilazione di statistiche congiunturali: questa tipologia di dati fornisce informazioni molto aggiornate a costi molto bassi che, con l'ausilio di appropriati metodi, può essere utilizzata per sostituire dati d'indagine o ridurre l'ampiezza dei campioni. Ovviamente, alla base è necessario poter disporre di adeguate risorse e organizzazione (database, competenze professionali etc.). In linea generale, è emersa la consapevolezza sulla necessità di mantenere una rilevazione sulle grandi imprese, almeno nelle fasi iniziali (Baldi et al., 2012; Maasing et. al, 2012; De Waal e Vlag, 2012).

4.2 Il nuovo metodo

Ispirandosi ai più interessanti risultati ottenuti nell'esperienza ESSnet è stata ripresa e affinata la sperimentazione dell'imputazione delle mancate risposte secondo un approccio di tipo micro, finalizzata a definire tutti gli aspetti di un nuovo metodo da mettere a regime, approfondendo i due aspetti rilevanti dell'approccio suggerito:

1) la definizione di una lista di rispondenti, sulla base di test sulla ricorrenza dei *pattern* di presenza;
2) la ricostruzione dei dati mancanti, attraverso un approccio per regressione.

La sperimentazione è stata condotta su un ampio numero di mesi e la prima diffusione ufficiale dei dati è avvenuta a giugno 2015, con successivi avanzamenti introdotti nelle uscite di settembre e dicembre 2015. Ulteriori affinamenti sono ancora in programma, inducendo ad aggiornare continuamente l'ambiente di test.

Il test sul nuovo approccio micro è stato condotto inizialmente su 18 mesi, su cui erano disponibili dati sia in versione preliminare che finale, a partire da aprile 2012[11] e fino a settembre 2013, a seguito di cui è avvenuta la prima uscita dell'indicatore sul numero delle posizioni lavorative nel comunicato stampa di giugno 2015 (occasione in cui è stata diffusa l'intera serie storica I trimestre 2000 - I trimestre 2015). Successivamente, al fine di apportare alcune migliorie, il periodo di sperimentazione si è allargato fino al mese di dicembre 2014, per un totale di 33 mesi e 11 trimestri di sperimentazione.

Il metodo viene applicato sui dati mensili, successivamente aggregati a livello trimestrale: la disponibilità del dato mensile permette di sfruttare al meglio le informazioni passate per ricostruire, attraverso un'analisi longitudinale minuziosa, la predizione dello stato di attività delle unità che non presentano dati economici in uno o più mesi e in modo coerente anche la fase d'imputazione ha seguito lo stesso criterio.

Disponendo dei dati finali per i singoli mesi su cui viene sperimentato il metodo, il confronto tra stima provvisoria e stima definitiva (errore di revisione) può essere utilizzato come parametro per valutare la bontà del metodo.

Nella rilevazione Oros tale errore è diverso da quello di una rilevazione congiunturale "tradizionale", in cui la stima provvisoria è basata su un campione di rispondenti rapidi e poi si calcola una stima rivista che include tutta l'informazione che perviene in seguito.

L'errore in Oros può considerarsi, piuttosto, pari ad uno scostamento dovuto ad errori non campionari rispetto ad un universo di riferimento rappresentato dalla stima definitiva; ciò implica che l'errore di revisione non fornisce soltanto informazioni relative all'affidabilità del dato ma rappresenta, soprattutto, una misura dell'accuratezza della stima provvisoria.

11 Il mese di aprile 2012 è il primo mese in cui sono disponibili dati preliminari confrontabili secondo la metodologia introdotta in occasione del cambio base relativo all'anno 2010.

Nella presentazione dei risultati della sperimentazione tale errore sarà lo strumento utilizzato per valutare l'accuratezza del nuovo metodo d'imputazione sui micro dati.

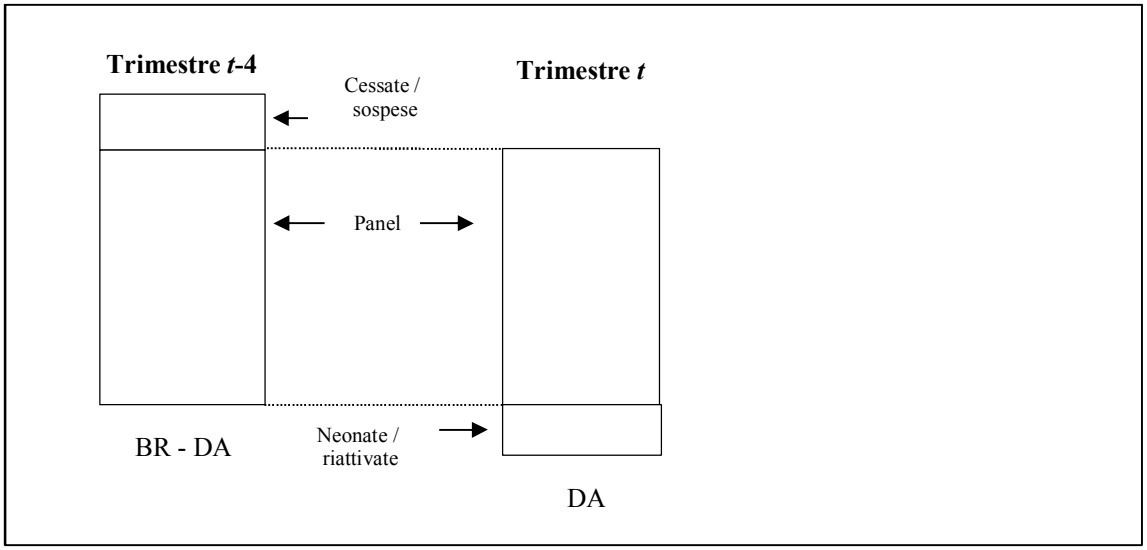
4.2.1 Predizione della lista

Una delle principali potenzialità che deriva dalla disponibilità di dati amministrativi (in breve DA) quasi completi per la stima preliminare del trimestre *t*, riguarda la possibilità di ottenere una rappresentazione alternativa e più aggiornata della popolazione di imprese attive in quell'istante rispetto a quella che si otterrebbe utilizzando il registro statistico delle imprese (in breve BR, *Business Register*), normalmente disponibile con significativo ritardo rispetto al periodo di stima[12]. Teoricamente, in assenza di dichiarazioni ritardatarie, dato l'obbligo per tutte le imprese di inviare dichiarazioni contributive, i dati amministrativi Inps dovrebbero includere anche le imprese che hanno appena iniziato l'attività nel periodo di riferimento ed escludere le imprese che hanno interrotto l'attività nel periodo immediatamente precedente. Nell'ipotesi in cui la stima provvisoria dovesse essere basata sulla lista di unità disponibili nel BR (per ipotesi a *t*-4), (figura 7), si avrebbe una situazione in cui verrebbero incluse unità cessate o sospese tra *t*-4 e *t* (sovra-copertura) mentre non sarebbero incluse unità nate o riattivate a *t* (sotto-copertura). Man mano che ci si allontana dall'istante di riferimento del BR le unità caratterizzate da sovra/sotto-copertura andrebbero aumentando. Inoltre, in questa ipotesi, la stima delle variabili *target* non potrebbe essere effettuata per semplice somma (aggregazione) dei dati amministrativi disponibili per *t* sulle unità nel BR, a causa delle sospensioni/cessazioni che andrebbero via via aumentando, non potendo tra l'altro essere compensate dalle nuove nate in *t* che, per costruzione, non sono disponibili. Una possibilità potrebbe essere quella di imputare tutte le unità uscite. L'accuratezza di queste stime dipenderà dall'ipotesi che i valori imputati delle unità realmente uscite vadano a compensare i valori, non misurati, delle unità entrate. Se la demografia d'impresa contribuisce con un una variazione netta pari a zero e l'imputazione non va ad introdurre altre forme di errore, allora la condizione è raggiunta e le stime saranno sufficientemente accurate.

Sulla variabile *target*, tuttavia, la demografia ha un impatto molto rilevante. Nella figura 8 si rappresenta la dinamica tendenziale dell'occupazione Oros nei settori da B a N dell'Ateco 2007 per il totale delle unità e per le sole panel, ossia le unità presenti sia a *t* che a *t*-4. La differenze tra le due curve è ovviamente spiegata dalla demografia ossia dalle unità presenti a *t* ma non a *t*-4, cioè le neonate o riattivate e le unità presenti a *t*-4 ma non a *t*, cioè le cessate o sospese. Quello che si nota, in particolare, è un legame tra il ciclo economico e la variazione dell'occupazione per sottopopolazione: in particolare, nelle fasi di espansione (2000-metà 2001, 2003-2004, 2006-metà 2008) sono le nuove nate / riattivate a spingere la dinamica in alto, mentre nelle fasi di recessione le cessazioni / sospensioni portano in basso la dinamica. Questa relazione non si osserva dal 2015 poiché la dinamica dell'occupazione è stata indotta dal *Jobs Act* e successivi interventi di decontribuzione, che non hanno avuto l'effetto di creare nuove imprese ma rilanciare l'occupazione di quelle esistenti. In questa situazione, appare stimolante l'ipotesi di utilizzare l'insieme dei dati amministrativi arrivati a *t* per definire la lista di unità attive per lo stesso periodo. Tuttavia, data la tempistica per l'invio delle dichiarazioni e la modalità di acquisizione dei dati dall'Inps per le stime provvisorie, nei dati scaricati vi sono unità *missing* e, in assenza di informazioni anagrafiche adeguatamente aggiornate, vi è incertezza su quali unità siano attive o meno. Unità *missing* che hanno risposto in mesi recenti potrebbero essere solo ritardatarie, oppure diventate inattive. D'altra parte, pur se la lista anagrafica è normalmente aggiornata sulle neonate nel periodo di riferimento, alcune di queste nuove unità potrebbero rispondere solo successivamente. In altre parole, per la stima preliminare, l'uso dei dati amministrativi necessita l'individuazione di una lista di unità presunte attive. Questa sarà composta dalle unità che rispondono in tempo per quel mese, che sono ovviamente attive, e il set dei rispondenti attesi, ossia le unità assunte attive. Per quanto riguarda l'ultimo insieme, i valori della popolazione *target* verranno imputati.

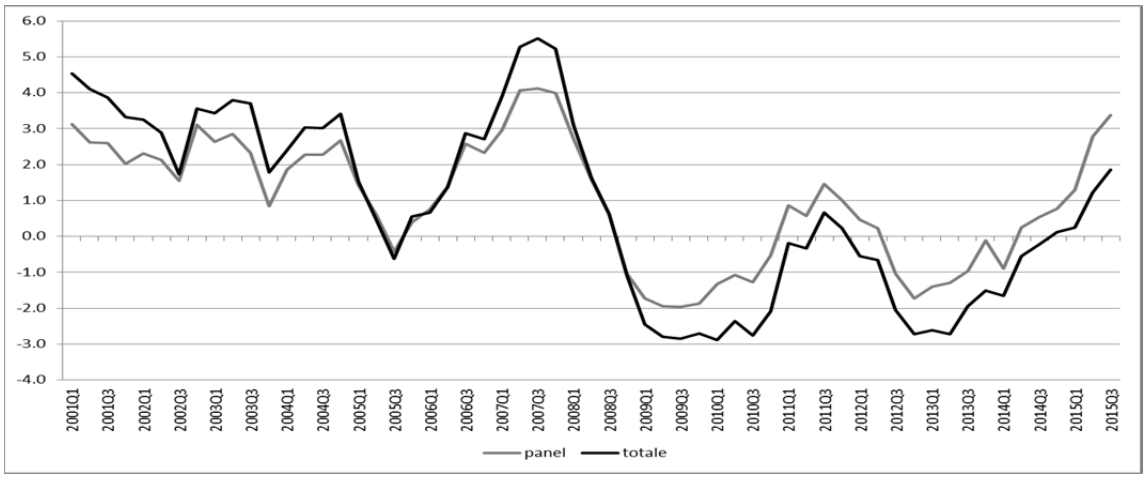
[12] In Istat il registro delle imprese attive riferito all'anno *a* è disponibile nel mese di aprile di *a*+2.

Figura 7 - Popolazione e demografia.



Fonte: Oros

Figura 8 - Dinamica occupazionale, posizioni panel e posizioni totali. Variazioni tendenziali, valori percentuali.



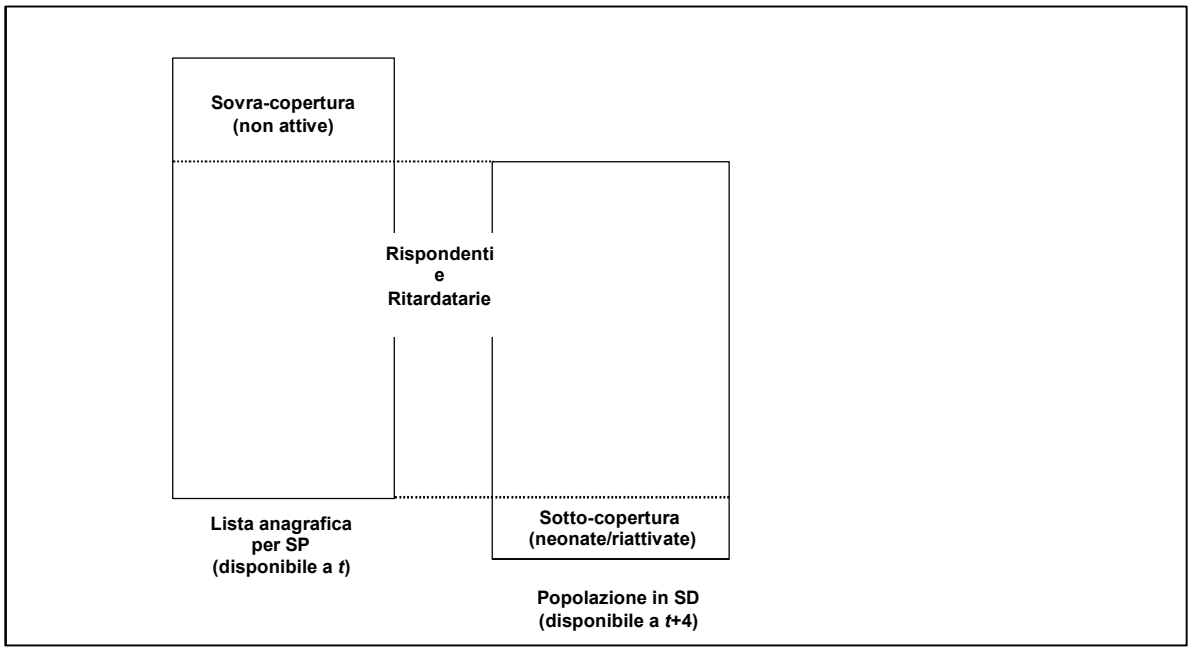
Fonte: Elaborazione su dati Oros

L'anagrafica trimestrale dell'Inps (§2.2) è un serbatoio di matricole attive a *t* e/o cessate/sospese da gennaio 2000 e rappresenta il set di microdati amministrativi di partenza da cui si cerca di individuare la lista di unità attive per *t*. Se gli eventi demografici fossero registrati tempestivamente tale set di dati rappresenterebbe esattamente la lista di stima per la rilevazione. Tuttavia, a causa dell'assenza di disposizioni coercitive nei confronti della comunicazione ritardataria o, spesso, in assenza di qualsiasi comunicazione da parte delle aziende, tale set di dati è caratterizzato da errori rilevanti di copertura, in particolare riguardo la mancata registrazione di cessazioni e sospensioni (sovra-copertura della lista), mentre le unità neonate non ancora registrate (ma che invieranno una dichiarazione contributiva) sono in numero ridotto (sotto-copertura della lista).

Al fine di limitare l'impatto di tale problematica sulle stime, una prima operazione è stata quella di tagliare la lista delle unità potenzialmente attive sulla base delle informazioni anagrafiche, considerando solo quelle che hanno presentato almeno una dichiarazione contributiva nel corso dell'ultimo anno rispetto a *t*.

Dal punto di vista teorico, confrontando la lista provvisoria definita con i criteri di cui sopra con la lista reale delle unità attive, disponibile dopo un anno con i dati finali (§2.2), per ogni mese *m* possono risultare le seguenti situazioni:

Figura 9- La lista nei dati provvisori e nei dati definitivi.



Fonte: Oros

In un approccio micro di predizione della lista, in cui la lista teorica di unità attive non è nota, le unità rispondenti, le ritardatarie e le unità di sovra-copertura concorrono alle stime. Mentre le unità di sovra-copertura, erroneamente incluse nelle stime, generano un errore per definizione di sovra-stima su cui si può intervenire affinando la metodologia di predizione, le unità di sotto-copertura non potendo essere predette, non concorreranno mai alle stime, generando un errore di sotto-stima.

Dopo aver effettuato il taglio rispetto alle unità attive almeno un mese tra *t*-4 e *t*, la popolazione di riferimento si presenta distribuita tra unità attive e unità potenzialmente attive in modo abbastanza uniforme rispettivamente, in media mensile per trimestre, per quasi l'80% e poco più del 20%, come mostrato nella tavola 2 seguente. Il 20% include unità attive ma rispondenti ritardatarie e unità non attive. Al fine di discriminare con maggior precisione possibile, l'individuazione dello status di attività viene ulteriormente perfezionato sulla base delle informazioni disponibili sulla nascita e la cessazione e sui segnali di attività dell'unità in periodi adiacenti, considerata la scarsa persistenza a ritardare l'invio del modello (cfr. §3). In particolare, vengono valutate informazioni riguardo la data di costituzione, di cessazione e la eventuale sospensione dell'attività per stagionalità. La definizione di stagionalità, in mancanza di date di sospensione aggiornate, particolarmente concentrata in alcuni settori (es. produzione alimentare, alberghi e ristoranti ecc.), si basa sulla verifica della persistenza dell'assenza nel corso dell'anno: in particolare, si definisce stagionale un'assenza nel mese a cui corrisponde un'assenza nello stesso mese dell'anno precedente. Fa eccezione il mese di giugno, su cui per attenuare la potenziale sovra-identificazione di stagionalità dovuta alla particolare concentrazione di DM ritardatari nel mese[13], l'assenza per stagionalità viene valutata anche nello stesso mese di due anni precedenti.

[13] Il mese di competenza di giugno è il più soggetto a ritardi nei tempi di consegna dei modelli a causa del periodo di ferie in cui ricade la scadenza per l'invio delle dichiarazioni contributive (entro fine luglio).

Tavola 2 - Valori medi mensili nei trimestri della quota di unità rispondenti e non rispondenti nei dati provvisori sulla lista totale delle unità potenzialmente attive costruita sulla base delle informazioni anagrafiche. Aprile 2012 – dicembre 2014, valori percentuali.

	Unità rispondenti	Unità non rispondenti
Media primo mese	*79,1*	*20,8*
Media secondo mese	*78,1*	*21,9*
Media terzo mese	*76,5*	*23,5*

Fonte: Elaborazione su dati Oros

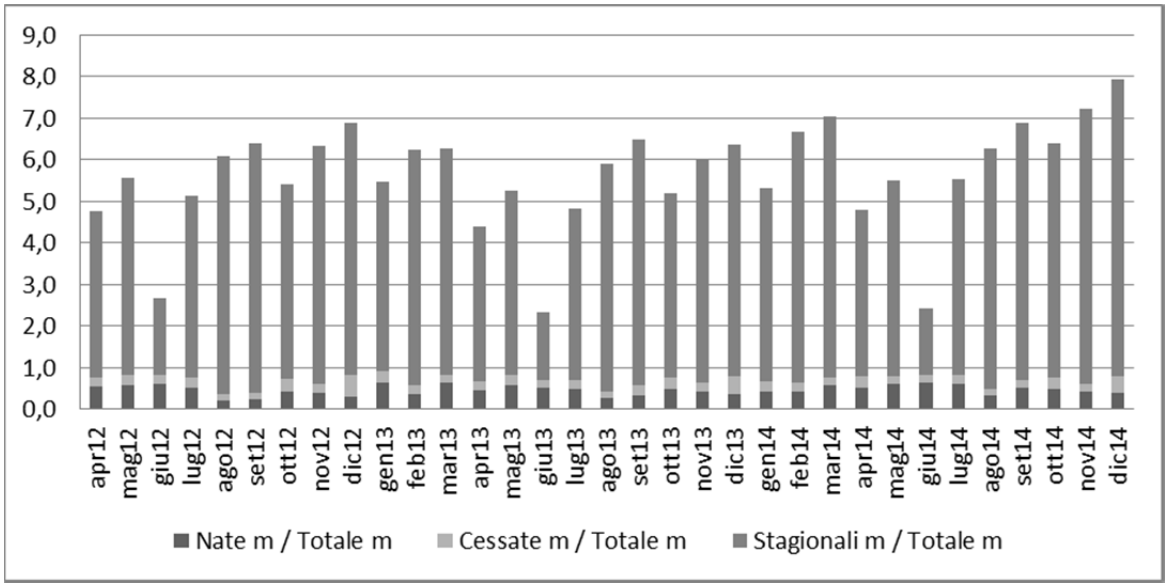
Nella figura 10 si riporta la distribuzione mensile delle unità assenti dai dati provvisori ad *m* e definite neonate, cessate o stagionali secondo i criteri descritti sopra. Esse rappresentano, nella media del periodo considerato, il 6% del totale delle unità assenti, con il mese di giugno in cui la loro incidenza scende al 2,5%. Guardando ai singoli eventi, si nota che l'incidenza di unità neonate e cessate è molto inferiore a quella delle unità stagionali e che le neonate hanno un peso numerico maggiore rispetto alle cessate. A causa del ritardo nella registrazione delle cessazioni, l'ipotesi di partenza è che l'insieme delle cessate definito utilizzando la data di chiusura dell'attività sia notevolmente sottostimato: questa è la principale problematica che ha portato alla definizione di un metodo di stima dello status di attività.

Le successive tavole 3 e 4 hanno l'obiettivo di mostrare la qualità e la tempestività dell'informazione sulle date anagrafiche disponibili sugli eventi di nascita e cessazione, valutando l'evento attraverso il confronto con la presenza dell'unità nel mese successivo alla dichiarazione della cessazione e al mese precedente rispetto alla dichiarazione della nascita.

In particolare, nella tavola 3 a fianco della quota di unità neonate sul totale delle unità assenti (0,3-0,6%) si riporta la distribuzione mensile delle unità che rispondono e che non rispondono nei dati finali. La risposta viene valutata nel mese *m* distribuendo le unità che hanno data di nascita nel mese *m* e le unità che hanno data di nascita nel mese *m*+1. Questa analisi vuole avvalorare l'ipotesi che l'unità neonata nel mese *m* debba essere ritenuta inattiva nel periodo passato fino al mese *m* compreso. Dando un'occhiata alle distribuzioni si può vedere come le ultime due colonne danno un importante valore aggiunto alla valutazione della data di nascita, in quanto soltanto lo 0,1% delle unità neonate ad *m*+1 già rispondeva nel mese *m*. Analizzando la seconda e terza colonna si vede come in media trimestrale il 60% delle unità nate ad *m* non risponde nello stesso mese ma inizia l'attività nel mese successivo. Se si guarda alla distribuzione nei mesi si vede come la non risposta aumenta nei primi due mesi del trimestre mentre diminuisce di molto nei terzi mesi, per i quali la diversa distribuzione va attribuita alla generale minore numerosità di unità rispondenti rispetto agli altri mesi del trimestre.

Nella tavola 4 vengono esaminate le informazioni sulla cessazione, valutando però la presenza di attività nel mese *m* se è stata osservata cessazione nel mese *m* oppure *m*-1.In questo caso la quota di unità che continuano a rispondere nel mese successivo alla cessazione è di circa lo 0,2% in media trimestrale con valori al più pari 0,5% in alcuni mesi. La motivazione potrebbe essere legata alla necessità delle imprese di dichiarare saldi contributivi da versare a seguito di cessazione. La cessazione osservata nel mese *m* porta in questo caso la quota di non rispondenti nel mese *m* a valori in media trimestrale pari all'incirca all'86%. Nei mesi del trimestre si osserva comunque una diminuzione nei terzi mesi con valori comunque intorno al 70%. Le evidenze tratte sui due tipi di evento hanno portato a considerare inattivo lo status nel mese *m* in presenza di nascita o cessazione. Inoltre l'unità si ipotizza non attiva, banalmente, per tutto il periodo precedente al mese di nascita o successivo al mese di cessazione.

Figura 10- Distribuzione delle matricole neonate, cessate e stagionali per mese di nascita. Quote sul totale delle unità non rispondenti nel mese. Valori percentuali.



Fonte: Elaborazione su dati Oros

Tavola 3 - Distribuzione delle unità neonate a *m* e a *m*+1 e assenti nei dati provvisori relativi ad m, per presenza / assenza nei dati finali. Aprile 2012 – dicembre 2014. Medie mensili, trimestrali e annuali, valori percentuali.

	Quota unità neonate assenti in SP nel mese *m* sul totale delle assenti in SP	Unità neonate assenti in SP nel mese *m*. Di cui:		Unità neonate nel mese *m*+1. Di cui:	
		Non Rispondenti nel mese *m* in SD	Rispondenti nel mese *m* in SD	Non Rispondenti nel mese *m* in SD	Rispondenti nel mese *m* in SD
Gen	0,5	71,1	28,9	99,9	0,1
Feb	0,4	63,5	36,5	99,9	0,1
Mar	0,6	33,8	66,2	99,9	0,1
Apr	0,5	66,6	33,4	99,9	0,1
Mag	0,6	55,9	44,1	99,9	0,1
Giu	0,6	44,3	55,7	99,9	0,1
Lug	0,5	68,3	31,7	99,9	0,1
Ago	0,3	61,1	38,9	99,9	0,1
Set	0,4	48,7	51,3	99,9	0,1
Ott	0,5	70,3	29,7	99,9	0,1
Nov	0,4	65,1	34,9	99,9	0,1
Dic	0,3	46,1	53,9	99,9	0,1
1° trim	0,5	56,1	43,9	99,9	0,1
2° trim	0,6	55,6	44,4	99,9	0,1
3° trim	0,4	59,4	40,6	99,9	0,1
4° trim	0,4	60,5	39,5	99,9	0,1
Anno	*0,5*	*57,9*	*42,1*	*99,9*	*0,1*

Fonte: Elaborazione su dati Oros

Tavola 4 - Distribuzione delle unità cessate a *m* e a *m*-1 e assenti nei dati provvisori relativi al mese m, per presenza / assenza nei dati finali. Aprile 2012 – dicembre 2014. Medie mensili, trimestrali e annuali, valori percentuali.

	Quota unità cessate nel mese *m* sul totale assenti in SP	Unità cessate nel mese *m* di cui:		Unità cessate nel mese *m*-1 di cui:	
		Non Rispondenti nel mese *m* in SD	Rispondenti nel mese *m* in SD	Non Rispondenti nel mese *m* in SD	Rispondenti nel mese *m* in SD
Gen	0,3	94,4	5,6	99,9	0,1
Feb	0,2	90,6	9,4	99,8	0,2
Mar	0,2	69,6	30,4	99,5	0,5
Apr	0,2	94,0	6,0	99,9	0,1
Mag	0,2	85,9	14,1	99,7	0,3
Giu	0,2	79,5	20,5	99,5	0,5
Lug	0,2	92,9	7,1	99,8	0,2
Ago	0,2	87,0	13,0	99,8	0,2
Set	0,2	77,7	22,3	99,7	0,3
Ott	0,3	92,9	7,1	99,9	0,1
Nov	0,2	89,0	11,0	99,9	0,1
Dic	0,5	78,2	21,8	99,7	0,3
1° trim	0,2	84,8	15,2	99,8	0,2
2° trim	0,2	86,5	13,5	99,7	0,3
3° trim	0,2	85,9	14,1	99,8	0,2
4° trim	0,3	86,7	13,3	99,8	0,2
Anno	0,2	86,0	14,0	99,8	0,2

Fonte: Elaborazione su dati Oros

La tavola 5 focalizza l'attenzione sulle unità definite stagionali, secondo i criteri descritti sopra. Per tali unità si vuole verificare la loro inattività e per verificare la robustezza dell'ipotesi fatta, la tabella mostra la percentuale di rispondenti e non per ogni mese calcolata sulle unità definite stagionali, a fianco dell'incidenza delle unità stagionali sul totale delle non rispondenti. In media trimestrale si osservano valori intorno al 70% di unità non rispondenti per arrivare anche all'80% nel 4° trimestre. Analizzando i singoli mesi sono come sempre i terzi mesi ad avere percentuali di non risposta più basse anche se comunque superiori al 50%. Anche per questo evento demografico si è scelto di considerare non attiva l'unità che presenta assenza per stagionalità.

Tavola 5 - Distribuzione delle unità definite stagionali a *m* assenti nei dati provvisori relativi ad *m*, per presenza / assenza nei dati finali. Aprile 2012 – dicembre 2014. Medie mensili, trimestrali e annuali, valori percentuali.

	Quota unità stagionali nel mese *m* sul totale assenti in SP	Unità stagionali	
		Non rispondenti nel mese *m* in SD	Rispondenti nel mese *m* in SD
Gen	4,6	85,7	14,3
Feb	5,8	74,5	25,5
Mar	5,9	54,0	46,0
Apr	3,9	79,3	20,7
Mag	4,6	63,8	36,2
Giu	1,7	58,9	41,1
Lug	4,4	77,7	22,3
Ago	5,7	69,9	30,1
Set	6,0	65,9	34,1
Ott	4,9	88,7	11,3
Nov	5,9	86,4	13,6
Dic	6,3	66,5	33,5
1° trim	5,4	71,4	28,6
2° trim	3,4	67,3	32,7
3° trim	5,4	71,2	28,8
4° trim	5,7	80,6	19,4
Anno	5,0	72,6	27,4

Fonte: Elaborazione su dati Oros

Identificate cessazioni, nascite e stagionalità, le assenze residuali vengono sottoposte ad una ulteriore valutazione, in cui lo status di attive viene conferito valutando la presenza dell'unità nei mesi precedenti e successivi a quello di stima. In particolare, i lavori del progetto ESSnet avevano suggerito di prendere in considerazione tre ipotesi:

- metodo di *m*-1: l'unità è presunta ritardataria se è presente nel mese *m*-1;
- metodo di *m*+1: l'unità è presunta ritardataria se è presente nel mese *m*+1;
- metodo di *m*-1, *m*+1: l'unità è presunta ritardataria se è presente sia nel mese *m*-1 e nel mese *m*+1.

Ovviamente per l'ultimo mese del trimestre può essere applicato solo il primo metodo. Inoltre, bisogna tener presente che le informazioni sui periodi vicini utilizzate per creare l'elenco dei non rispondenti si riferiscono ad una popolazione provvisoria, nei mesi relativi ai trimestri precedenti fino a *t*-3, e ad una popolazione finale, per i mesi di *t*-4. Ne deriva che l'informazione ausiliaria fino a *t*-3 è soggetta ad un certo grado di incertezza: l'assenza di un'unità nel mese precedente può essere sia un segnale di inattività o un segnale di non risposta. Questa incertezza comporta errori nella definizione della lista attesa dei non rispondenti che è tanto maggiore quanto maggiore è la persistenza di un comportamento di segnalazione tardiva dei modelli DM.

Al fine di valutare le *performance* del metodo di predizione della lista è utile classificare le unità incluse secondo la loro realizzazione nella popolazione finale. In particolare, si possono distinguere i seguenti status teorici:

- **attive corrette** (in breve AC): unità non rispondenti nei dati provvisori, definite attive secondo i criteri di predizione della lista e presenti nei dati definitivi. Si tratta di unità correttamente incluse nella lista predetta;
- **attive non corrette** (in breve ANC): unità non rispondenti nei dati provvisori, definite non attive secondo i criteri della lista ma presenti nei dati definitivi. Sono unità erroneamente escluse dalla lista determinando un errore che può essere chiamato di "sotto inclusione";
- **non attive corrette** (in breve NAC): unità non rispondenti nei dati provvisori, definite non attive nella lista predetta e assenti nei dati definitivi. Si tratta di una corretta esclusione delle unità dalla lista;
- **non attive non corrette** (in breve NANC): unità non rispondenti nei dati provvisori, definite attive nella lista, ma non presenti nei dati definitivi. Sono unità erroneamente incluse in fase di predizione determinando quindi un errore di "sovra inclusione".

Nel seguente schema, ogni cella rappresenta i vari status risultati del confronto delle unità predette rispetto a quelle reali:

Figura 11 - Status delle unità nella popolazione provvisoria e realizzazione nei dati finali.

		Popolazione nei dati provvisori		
		Unità attive e Rispondenti	Non rispondenti	
			Unità presunte attive	Unità presunte non attive
Popolazione nei dati finali	Attive=Rispondenti	Attive	Attive corrette (AC) (correttamente incluse)	Attive non corrette (ANC) (erroneamente escluse)
	Non attive (cessate, sospese)	--	Non attive non corrette (NANC) (erroneamente incluse)	Non attive corrette (NAC) (correttamente escluse)

Fonte: Oros

La scelta del metodo da applicare si è basata sulle principali evidenze raggiunte dalla sperimentazione delle tre varianti ai dati Oros nei lavori in ESSnet. In particolare, è emerso che:

- il metodo *m*-1 è quello che, in generale, ha mostrato i migliori risultati, fatta eccezione del primo mese dell'anno (gennaio), per cui si registra un problema di sovra-identificazione di unità attive non corrette, a causa della particolare concentrazione di interruzioni di attività nel mese *m*-1 di dicembre;

- il metodo *m*-1, *m*+1 ha evidenziato migliori *performance* rispetto al metodo *m*+1 nel primo mese di ogni trimestre. Il metodo *m*+1, invece, è risultato superiore rispetto ad *m*-1 in termini di unità non attive non corrette, mentre i due sono apparsi pressoché equivalenti in termini di unità attive non corrette;
- nel secondo mese di ogni trimestre il metodo *m*-1, *m*+1 produce migliori risultati rispetto agli altri due metodi in termini di unità non attive non corrette;
- tuttavia, entrambi i metodi *m*-1, *m*+1 ed *m*+1 forniscono peggiori risultati rispetto ad *m*-1 in termini di unità attive non corrette, a seguito del fatto che l'ultimo mese del trimestre è sempre affetto da un più basso grado di completezza.

In conclusione, il metodo *m*-1 è stato applicato in una prima versione a regime del metodo e per due occasioni di stima preliminare, ossia a giugno e settembre 2015. In seguito ad alcuni approfondimenti, a partire dall'uscita di dicembre 2015, questa variante è stata lievemente rivista per ridurre l'impatto delle attive non corrette generate dal metodo *m*-1. Tale variante prevede che la presenza in un generico mese *m* (o anche la presunta attività in *m*) sia segnale di attività anche per il mese successivo, a meno di fattori stagionali e demografici. Questo nuovo metodo è stato definito "metodo basato sul mese precedente con variante trimestrale" o in breve "metodo *m*-1 trim".

In termini formali, per ogni trimestre *t* è possibile definire per ogni unità i il suo *pattern* in base alla presenza (valore 1) o assenza (valore 0) dei dati della variabile occupazionale secondo l'ordine del mese (ossia nel primo, nel secondo e nel terzo mese) e in breve possiamo definire tale *pattern* come $p_{i,t}$. In totale, nel generico trimestre si possono avere le seguenti casistiche:

$$p_{i,t} = \begin{cases} 000 \\ 100 \\ 010 \\ 001 \\ 110 \\ 011 \\ 101 \\ 111 \end{cases} \qquad (1)$$

Dopo la fase di predizione delle unità ritardatarie il *pattern* trimestrale per le unità presunte assenti assume nuove modalità, così come riportato nella tavola 6 in cui è indicata anche la presenza o meno del dato occupazionale nell'ultimo mese del trimestre precedente, in quanto da tale informazione dipende l'assegnazione del flag d'imputazione nel primo mese del trimestre corrente *t* e a cascata con il metodo nuovo, anche del secondo e del terzo mese.

Oltre ai nuovi *pattern* trimestrali successivi alla predizione della lista , indicati come $\hat{p}_{i,t}$, in cui le modalità in grassetto evidenziano il passaggio dallo zero alla modalità uno, sono riportati nelle ultime due colonne il numero di mesi in cui il dato è potenzialmente imputabile sia nel metodo iniziale (metodo *m*-1) che nel metodo affinato (metodo *m*-1 trim) per cui sono evidenziati in grassetto i casi in cui il valore è più alto rispetto al primo metodo.

Tavola 6 – Predizione dello stato di attività nei metodi *m*-1 e *m*-1 trim: status di partenza e status predetto.

Presenza del dato occupazionale nell'ultimo mese nel trimestre precedente	Pattern iniziale $\overline{p}_{i,t}$	Metodo *m*-1 Pattern nella lista predetta $\hat{p}_{i,t}$	Metodo *m*-1 trim Pattern nella lista predetta $\hat{p}_{i,t}^{trim}$	Metodo *m*-1 Numero mesi da imputare	Metodo *m*-1 trim Numero mesi da imputare
-- 0	011	011	011	0	0
-- 1	011	**1**11	**1**11	1	1
-- 0	001	001	001	0	0
-- 1	001	**1**01	**11**1	1	**2**
-- 0	000	000	000	0	0
-- 1	000	**1**00	**111**	1	**3**
-- 0	010	01**1**	01**1**	1	1
-- 1	010	**11**1	**11**1	2	2
-- -	100	1**1**0	1**11**	1	**2**
-- -	110	11**1**	11**1**	1	1
-- -	101	1**1**1	1**1**1	1	1
-- -	111	111	111	0	0

Fonte: Elaborazione su dati Oros

Nelle tavole che seguono si riportano alcuni risultati della sperimentazione del metodo *m*-1 trim di predizione della lista per i mesi da aprile 2012 a dicembre 2014. Il metodo viene valutato secondo la realizzazione degli status predetti nella popolazione finale, disponibile dopo 1 anno. Le unità vengono classificate secondo le casistiche esposte nella figura 11.

Nella tavola 7, le unità definite non rispondenti nella lista predetta per la stima provvisoria che, come visto nella tavola 2, rappresentano il 20% circa della lista complessiva di partenza, sono distinte tra quelle che sono state definitive attive e lo sono (AC) e quelle che non lo sono (ANC) e, viceversa, quelle che sono state dichiarate non attive e non lo sono (NANC) e quelle che invece sono effettivamente non attive (NAC), fatto 100 il totale. Nella tabella è interessante tenere sotto controllo la differenza tra le attive non corrette (ANC) e le non attive non corrette (NANC) che rappresentano rispettivamente la sotto-copertura e la sovra-copertura della lista di imputazione, indicata nell'ultima colonna. Sul gruppo delle non rispondenti si osservano i seguenti risultati:

- le unità non attive e correttamente identificate dal metodo ("corretta esclusione") sono tra il 70 e il 90%. La loro incidenza si riduce dal primo al terzo mese, per effetto della crescita sia delle non attive non corrette sia delle attive corrette;
- per contro, le non attive non corrette (sovra inclusione) hanno un'incidenza che varia tra il 3% e il 10%, con valori che crescono dal primo al terzo mese, evidenziando le migliori *performance* del metodo all'allontanarsi dall'ultimo mese del trimestre precedente, quale variabile ausiliaria per la predizione;
- anche le attive corrette (corretta inclusione) registrano il valore più basso nel primo mese di ogni trimestre e il più alto nel terzo mese. Il primo mese registra il valore più basso per via dell'ultimo mese del trimestre precedente, strutturalmente caratterizzato da meno rispondenti. E' su questo raggruppamento che si registra il più alto scarto di *performance* tra il primo e il terzo mese;
- le attive non corrette (sotto inclusione) sono il gruppo con distribuzione più uniforme, rappresentando tra il 2 e il 4% dell'insieme delle non rispondenti.

Al di là della capacità del metodo di predire i singoli status, garantita nei raggruppamenti delle attive corrette e delle non attive non corrette, tanto più i dipendenti delle non attive corrette (sovra inclusione) si bilanciano con quelli delle attive non corrette (sotto inclusione), tanto più la lista sarà stata predetta correttamente nell'aggregato. Il saldo è rappresentato nella tabella, evidenziando una prevalenza di segni positivi, ad indicare la tendenza del metodo a sovra identificare unità attive, in particolare nel terzo mese.

In termini di dipendenti, le mancate risposte incidono per una percentuale che varia tra l'1% nel primo mese e il 5% nell'ultimo mese, con alcune eccezioni già note (figura 12). Sul totale delle non

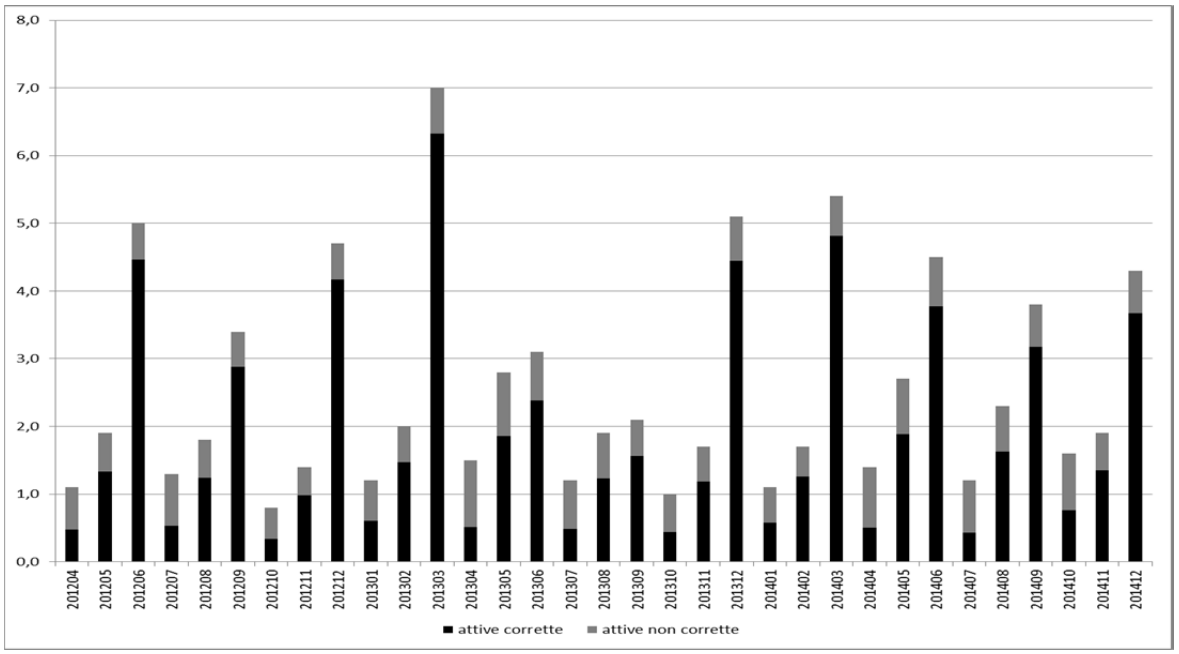
rispondenti, i dipendenti delle attive corrette coprono lo 0,5% dell'occupazione complessiva nel primo mese, l'1,5% nel secondo mese e, infine, tra il 3 e il 5% nell'ultimo mese. Molto meno rilevante l'incidenza delle attive non corrette, che rappresentano una quota poco significativa quando il tasso di mancata risposta è più alto (tipicamente nel terzo mese) e, solo nel primo mese, in qualche occasione superano il peso dei dipendenti delle attive corrette.

Tavola 7 - Distribuzione della popolazione predetta in termini di numero di unità nella stima provvisoria per status secondo il metodo *m*-1 trim di predizione della lista nel totale B-S. Aprile 2012 - dicembre 2014. Dati mensili e medie mensili per trimestre, valori percentuali.

		Unità non rispondenti di cui:			
	attive corrette (AC)	attive non corrette (ANC)	non attive non corrette (NANC)	non attive corrette (NAC)	Saldo non attive non corrette – attive non corrette (NANC-ANC)
201204	1,1	2,5	2,9	93,6	0,4
201205	3,8	2,5	6	87,7	3,5
201206	14,6	2,6	9,5	73,3	6,9
201207	1,4	3,1	5,3	90,2	2,2
201208	4,6	2,6	8,7	84,2	6,1
201209	10,1	2,6	10,8	76,4	8,2
201210	0,9	2,1	4,1	92,9	2,0
201211	2,6	1,9	6,6	88,9	4,7
201212	11,0	2,4	7,3	79,3	4,9
201301	1,3	2,1	4,9	91,7	2,8
201302	3,7	2,0	8,6	85,7	6,6
201303	17,8	3,0	9	70,2	6,0
201304	1,5	3,8	2,5	92,2	-1,3
201305	6,3	3,9	5,5	84,4	1,6
201306	8,3	3,1	9,1	79,5	6
201307	1,6	3,3	4,9	90,2	1,6
201308	4,1	2,9	7,1	86,0	4,2
201309	4,9	2,5	9,2	83,4	6,7
201310	1,1	2,4	3,6	92,9	1,2
201311	3,8	2,4	5,9	88,0	3,5
201312	12,2	2,9	6,6	78,4	3,7
201401	1,4	2,3	4,7	91,6	2,4
201402	3,5	2,1	7,9	86,6	5,8
201403	14,0	2,9	8,5	74,6	5,6
201404	1,5	3,7	2,7	92,1	-1,0
201405	6,6	3,7	5,2	84,5	1,5
201406	13,6	3,4	8,9	74,2	5,5
201407	1,6	3,9	4,4	90,2	0,5
201408	5,7	3,5	6,6	84,2	3,1
201409	10,7	3,3	8,2	77,8	4,9
201410	1,8	3,4	3,2	92,8	-0,2
201411	3,4	2,5	5,9	88,9	3,4
201412	9,7	2,8	6,8	81,0	4,0
media mese 1	*1,4*	*2,9*	*3,9*	*91,9*	*1,0*
media mese 2	*4,4*	*2,7*	*6,7*	*86,3*	*4,0*
media mese 3	*11,5*	*2,9*	*8,5*	*77,1*	*5,6*

Fonte: Elaborazione su dati Oros

Figura 12 - Incidenza dipendenti nella stima provvisoria per status secondo il metodo *m*-1 trim di predizione della lista nel totale B-S. Aprile 2012 - dicembre 2014. Dati medi mensili, valori percentuali.



Fonte: Elaborazione su dati Oros

4.2.2 Ricostruzione dei dati mancanti

Il secondo aspetto affrontato riguarda la ricostruzione del dato mancante sulle unità presunte attive, operazione da effettuare anch'essa a livello mensile, in linea con la fase di predizione della lista.

Durante i lavori del progetto ESSnet WP4 erano stati presi in considerazione tre metodi di stima del dato dell'occupazione espresso come totali (Maasing et al., 2012): i primi due metodi basati sullo stimatore rapporto, ovvero sulla predizione tramite applicazione di un coefficiente calcolato come l'indice rispetto a *m*-1 o a *m*-12 del numero delle posizioni lavorative delle unità rispondenti; il terzo metodo, invece, è basato sulla predizione tramite modello di regressione bivariato, in cui le due variabili ausiliarie sono i dati occupazionali riferiti al dato precedente e al dato tendenziale.

Lo stimatore rapporto presuppone l'esistenza per ogni unità della popolazione, di una relazione lineare tra una variabile (y) in un determinato periodo t, con la stessa riferita ad un periodo precedente t-1. In formule si può esprimere come:

$$y_t = Ry_{t-1} \qquad (2)$$

Tale metodo è utilizzato soprattutto in un contesto di dati longitudinali in cui si sfrutta il legame dei dati tra periodi successivi . Inoltre la variabile "posizioni lavorative dipendenti" è per sua natura inerziale.

Dato che si hanno a disposizione i dati mensili per tutte le unità rispondenti, si ha che per ogni mese m e per ogni unità i rispondente in entrambi i periodi, il dato stimato ($\hat{y}_{i,m}$), del numero delle posizioni lavorative per le unità presunte mancate risposte, è stato calcolato distintamente con i due stimatori rapporto:

$$\hat{y}_{i,m} = \frac{\sum_{i,m \in Risp} y_{i,m}}{\sum_{i,m-1 \in Risp} y_{i,m-1}} y_{i,m-1} \qquad (3)$$

$$\hat{y}_{i,m} = \frac{\sum_{i,m \in Risp.} y_{i,m}}{\sum_{i,m-12 \in Risp.} y_{i,m-12}} y_{i,m-12} \qquad (4)$$

Si presume, quindi, che il rapporto calcolato solo sull'insieme delle unità rispondenti in entrambe i mesi (nel rapporto (3) per i mesi *m* e *m*-1 e nel (4) e per i mesi *m* e *m*-12), sia una stima della relazione lineare tra le due variabili, come definito sopra nella relazione (2)[14].

Il terzo metodo proposto nell'ambito dell'ESSnet, è una sorta di combinazione dei modelli univariati, con l'aggiunta di un'intercetta e di errori eteroschedastici, ossia:

$$y_{i,m} = \alpha + \beta_{1,m} y_{i,m-1} + \beta_{2,m} y_{i,m-12} + \varepsilon_{i,m}, \qquad \text{con} \quad \varepsilon_{i,m} \sim N(0, \sigma^2 y_{i,m-1}) \quad (5)$$

La valutazione dei risultati delle stime ottenute con i tre metodi attraverso il RMSE, distinta tra i rispondenti e non rispondenti, ha portato ad escludere lo stimatore rapporto basato sul dato tendenziale, calcolato con la (4), mentre i risultati degli altri due metodi hanno registrato valori molto simili tra loro. Lo stimatore rapporto, se da un lato risulta più semplice in termini d'applicazione e più comprensibile per gli utenti, dall'altro però non tiene conto della compresenza del dato congiunturale e tendenziale, entrambi informazioni utili per la stima del dato corrente. Ciò diventa rilevante in alcuni settori caratterizzati da un'intensa attività stagionale (ad esempio quelli connessi al turismo e alle industrie alimentari).

L'approccio scelto per la per la stima del dato occupazionale sulle unità presunte ritardatarie è interamente *model based*, ossia consiste nella predizione da un modello di regressione bivariato per le unità che hanno sia il dato $y_{i,m-1}$ e sia $y_{i,m-12}$, mentre nei casi in cui non sono disponibili entrambi i dati (o nel mese *m*-1 o nel mese *m*-12), si utilizza un modello univariato.

Data la configurazione dei dati si definiscono tre modelli: il primo, detto "modello A", con due variabili ausiliarie è applicato alle unità in cui sono presenti sia il dato $y_{i,m-1}$ che sia $y_{i,m-12}$; il secondo, detto "modello B", ha una sola variabile ausiliaria ed è applicato alle unità in cui è presente solo il dato $y_{i,m-1}$; infine il terzo, detto "modello C", è come il secondo univariato, ma è relativo alle unità in cui è presente solo il dato $y_{i,m-12}$.

In formule sono:

$$\begin{aligned} &\text{Modello A:} \quad y_{i,m} = \alpha_{A,m} + \beta_{1A,m} y_{i,m-1} + \beta_{2A,m} y_{i,m-12} + \varepsilon_{i,m}, \qquad (6) \\ &\text{Modello B:} \quad y_{i,m} = \alpha_{B,m} + \beta_{B,m} y_{i,m-1} + \varepsilon_{i,m}, \\ &\text{Modello C:} \quad y_{i,m} = \alpha_{C,m} + \beta_{C,m} y_{i,m-12} + \varepsilon_{i,m}, \end{aligned}$$

dove l'indice *i* esprime l'unità *i*-esima della popolazione, le variabili $y_{i,m}$, $y_{i,m-1}$ e $y_{i,m-12}$ indicano rispettivamente il numero di dipendenti nei mesi *m*, *m*-1 e *m*-12, $\varepsilon_{i,m}$ è il termine di errore a media nulla. I parametri (α, β) distinti per ogni modello (modello A, modello B e modello C) sono stati

14 Tale stimatore corrisponde a quello ottenuto con un'imputazione per regressione con una sola variabile ausiliaria senza intercetta e con varianza degli errori proporzionale al dato passato, ossia a: $y_{i,m} = \beta y_{i,m-1} + \varepsilon_{i,m}$, con $\varepsilon_{i,m} \sim N(0, \sigma^2 y_{i,m-1})$ e in modo analogo vale la stessa relazione con il dato tendenziale.

stimati con il metodo degli OLS separatamente su sottogruppi di unità omogenee (*model groups*) secondo la classificazione dell'attività economica in divisione Ateco 2007 e nel caso di divisioni di piccola entità, per alcuni loro raggruppamenti, appartenenti comunque allo stesso livello di classificazione superiore (la sezione Ateco).

In tal modo sono contemplate le diverse casistiche secondo la presenza dei dati nei periodi passati; inoltre, quando presente, il dato tendenziale viene sempre incluso come informazione ausiliaria in quanto, nonostante sia più distante rispetto al congiunturale, è comunque un dato in versione definitiva e quindi è il più indicato per rappresentare le attività stagionali.

Come già espresso nei paragrafi precedenti, nella rilevazione Oros la misura dello scostamento tra il dato definitivo e quello provvisorio ha rappresentato in fase di sperimentazione, lo strumento migliore per poter valutare se il nuovo processo d'imputazione sui micro dati ha portato ad un effettivo e sostanziale miglioramento sulle stime provvisorie rispetto alla vecchia procedura. La valutazione dei risultati quindi, si è concentrata sul miglioramento ottenuto con la predizione del dato attraverso il modello (6). Le analisi che seguono sono basate soprattutto sulla valutazione di alcune misure di errore che si presentano durante i diversi *step* del processo di correzione, a partire dai risultati della stima dei parametri calcolati solo sul set delle unità rispondenti; dalla predizione del dato dell'occupazione sulle unità "presunte ritardatarie", fino alla reale misura dell'entità dell'errore misurato sulle effettive unità non rispondenti. Quest'ultimo aspetto sarà trattato in modo più approfondito nei paragrafi successivi.

La procedura d'imputazione micro consiste principalmente in due fasi:

1) la stima dei parametri del modello (6) effettuata sul set informativo dei dati disponibili della sottopopolazione delle PMI, ossia i rispondenti della stima provvisoria;
2) la predizione del dato dell'occupazione sulle unità PMI valutate presunte mancanti nella fase di predizione della lista (cfr.§ 4.1), secondo la relazione (6) applicando le stime dei parametri ottenute come definito nel punto precedente.

Il set informativo su cui si stimano i parametri di ciascun modello è riassunto nella tavola seguente e si riferisce alle unità rispondenti della stima provvisoria della sottopopolazione PMI. In base alla presenza e/o assenza dei dati nei periodi precedenti, si ottiene la seguente distribuzione dei tre tipi di modello nella prima fase di stima:

Tavola 8 - Set informativo delle unità rispondenti delle PMI, ripartito secondo i modelli con cui si stimano i parametri. Aprile 2012 – dicembre 2014. Dati medi mensili, valori percentuali.

Modello A	Modello B	Modello C
86,7	*11,3*	*2,0*

Fonte: Elaborazione su dati Oros

Come si evince dalla tavola, in media, nella maggior parte dei casi (86,7%), i parametri del modello sono stimati con il modello A. Esiste, inoltre, una parte molto esigua di rispondenti, non utilizzabile per la stima dei parametri, in quanto i dati ausiliari nei periodi precedenti sono assenti.

La sperimentazione è stata portata avanti distintamente per divisione Ateco 2007 e per tutti i mesi da aprile 2012 a dicembre 2014, generando quindi numerosi risultati, che sono sintetizzati con misure e analisi specifiche su particolari evidenze, in modo da facilitare la modalità della loro lettura e di analisi.

Le statistiche usate per la valutazione dell'errore di predizione sul set informativo dei rispondenti sono:

$$\mathrm{RMSE}_m = \sqrt{\sum_{i \in Risp} (y_{i,m} - \hat{y}_{i,m})^2 \Big/ (n_m - p)} \qquad (7)$$

$$\mathrm{C}V_m = 100 \frac{RMSE_M}{\sum_{i \in Risp} (\hat{y}_{i,m}) / n_m} \qquad (8)$$

con $y_{i,m}$ i valori delle posizioni lavorative nel mese *m* delle unità rispondenti (indicati nella formula col termine *Risp*), con $\hat{y}_{i,m}$ i corrispondenti valori stimati da modello, con n_m la numerosità delle unità rispondenti relative al mese *m* e con *p* il numero dei regressori. Le statistiche sono state calcolate all'interno di ciascun sottogruppo di unità rispondenti, ossia in base alla co-presenza del dato nel mese *m* con quello congiunturale e/o con quello tendenziale (le cui percentuali sono riportate nella tavola 8) e per modello applicato.

Un risultato a livello generale è riportato nella tavola 9, in cui vi è la percentuale di quanti modelli tra tutti quelli generati, hanno registrato i valori più bassi in termini di coefficienti di variazione (Min_CV) e di RMSE (Min_RMSE). Dai dati emerge come il modello più applicato sui rispondenti, il modello A, risulta quello che registra i migliori risultati per i valori più bassi di CV e di RMSE.

Tavola 9 – Distribuzione dei casi con minimo CV e RMSE per tipologia di modello. Settori da B a S. Periodo: Aprile 2012 – dicembre 2014. Valori percentuali.

	Min_CV	Min_RMSE
Modello A	98	82
Modello B	1	17
Modello C	1	1

Fonte: Elaborazione su dati Oros

Le stesse statistiche misurate sempre per ogni modello applicato e nei diversi sottogruppi di unità, indicando anche i valori di massimo (Max), di minimo (Min) e il valore medio (Mean) degli indici CV e RMSE, più il dato sulla media della variabile dipendente (dipMean), ossia la media delle posizioni lavorative e la statistica R^2 aggiustato *(*R2adj), sono presenti invece nella tavola 10. Tutti i tre modelli valutati in termini di R^2 presentano in media buoni risultati e in particolare i modelli A e B registrano risultati migliori rispetto al modello C. Osservando le altre statistiche si nota come il modello C, basato solo sul dato tendenziale quale variabile ausiliaria e con numerosità molto ridotta dei rispondenti rispetto agli altri due, è quello che presenta risultati peggiori in termini di CV e di RMSE, che si discostano significativamente da quelli osservati rispetto agli altri due modelli. In generale, i risultati migliori sono registrati nel modello A.

Tavola 10 – Media stimata della variabile dipendente (dipMean), valori medi, minimi e massimi degli indici CV e RMSE, R2adj per i modelli applicati sui settori da B a S. Aprile 2012 – dicembre 2014.

	dipMean	Mean_RMSE	Max_RMSE	Min_RMSE	Mean_CV	Max_CV	Min_CV	Mean_R2adj
Modello A	12,9	2,8	41,7	0,1	26,7	352,5	2,2	0,98
Modello B	5,4	2,3	73,2	0,0	43,3	457,6	0,0	0,95
Modello C	12,6	7,2	237,1	0,0	74,5	639,8	0,0	0,82

Fonte: Elaborazione su dati Oros

Un focus sulla *performance* a livello di sezione Ateco, è presente nella tavola 11, in cui le statistiche CV ed RMSE, riportate con i loro valori di massimo minimo e medio, sono presenti solo per il modello di regressione A, che è quello più applicato, sia per la stima dei parametri e sia per la predizione del dato.

Come risultato generale delle varie statistiche, i modelli con *performance* migliore appartengono al settore dell'industria (relativo all'aggregato B-F), mentre appaiono meno performanti nei settori dei servizi, sia di mercato (aggregato G-N) che alla persona (aggregato P-S). I risultati medi di questi ultimi due aggregati tuttavia, risultano avere *performance* peggiori solo in alcune particolari sezioni.

Entrando nel merito delle sezioni, nell'industria, il settore F (Costruzioni) è quello che presenta valori più alti in termini di CV, ma si mantiene sullo stesso livello delle altre sezioni per i valori di

RMSE. Nei servizi di mercato la sezione J (Servizi di informazione e comunicazione) è quella che registra la *performance* peggiore, con valori alti sia per la media di CV e RMSE, che per i corrispondenti valori massimi; a seguire la sezione N (Attività di noleggio, delle agenzie di viaggio e dei servizi di supporto alle imprese). Nell'aggregato P-S tutti i valori corrispondenti a CV non sono bassi, soprattutto nella sezione R (Attività artistiche, sportive, d'intrattenimento e divertimento), però, se in termini di RMSE le sezioni P e S hanno valori molto bassi, il settore della sanità ed assistenza sociale (Q) ha valore di CV vicino a 5 come per la sezione J, quella più critica.

I settori meno performanti (J, N, R e Q) sono notoriamente più critici rispetto agli altri, in quanto hanno flussi di dipendenti meno stabili nel tempo. La sezione R in aggiunta ha un peso occupazionale molto basso e quindi è sufficiente che una variazione di lieve entità in termini di dipendenti impatta in modo più significativo sui risultati delle statistiche rispetto alle altre sezioni.

Tavola 11 – Modello A: media stimata della variabile dipendente (dipMean), valori medi, massimi e minimi degli indici CV e RMSE, per i modelli applicati sui settori da B a S. Aprile 2012 – dicembre 2014.

	dipMean	Mean_RMSE	Max_RMSE	Min_RMSE	Mean_CV	Max_CV	Min_CV	Mean_R2adj
B	10,3	1,6	3,4	0,8	15,7	32,3	8,1	0,99
C	17,9	2,5	17,7	0,7	17,0	114,2	2,2	0,99
D	14,2	1,8	4,3	0,8	12,7	29,4	5,9	1,00
E	17,8	2,1	8,8	0,6	13,2	48,1	2,8	1,00
F	8,1	2,2	7,8	1,0	28,2	55,8	15,0	0,98
G	5,0	1,2	3,0	0,4	24,0	73,5	10,4	0,99
H	10,6	3,2	9,6	1,2	29,2	82,9	12,4	0,98
I	6,5	2,7	5,1	1,0	41,5	64,7	18,8	0,96
J	9,5	5,1	36,6	0,7	50,2	352,5	7,0	0,95
K	20,0	2,9	17,9	0,4	15,2	51,8	3,0	1,00
L	2,6	1,2	6,4	0,7	48,7	260,2	2,0	0,96
M	5,8	1,7	13,1	0,1	27,9	143,8	7,30	0,99
N	12,7	4,0	41,7	0,8	35,7	350,4	7,6	0,98
P	6,9	1,8	3,8	0,9	27,4	69,1	11,8	0,98
Q	22,8	4,9	12,3	0,6	25,8	82,3	8,4	1,00
R	9,9	3,3	36,1	0,7	41,7	289,7	7,9	0,96
S	3,5	1,0	3,4	0,4	29,1	89,0	11,1	0,99

Fonte: Elaborazione su dati Oros

Un'analisi più specifica a livello di divisione Ateco, è riportata nella tavola successiva, in cui sono presenti solo le divisioni che hanno fornito risultati peggiori in termini di RMSE e CV. La tavola seguente illustra i casi meno performanti derivanti dal modello più utilizzato (modello A) registrati in alcuni mesi del periodo di sperimentazione.

Tavola 12 - Modello A, focus sulle divisioni Ateco meno performanti per valori di RMSE e CV, registrati in alcuni mesi del periodo Aprile 2012 – dicembre2014.

Divisione Ateco	Sezione	data	meanRMSE	meanCV	meanR2adj
59	J	201205	19,3	176,2	0,70
59	J	201206	23,5	210,5	0,67
59	J	201306	21,8	201,5	0,65
59	J	201307	24,5	209,5	0,68
59	J	201308	19,5	202,8	0,58
59	J	201401	18,6	200,5	0,62
59	J	201403	15,3	165,3	0,63
59	J	201404	14,3	154,7	0,69
59	J	201405	36,7	352,5	0,46
59	J	201408	19,6	215,4	0,61
59	J	201409	25,4	227,6	0,67
59	J	201411	22,8	207,0	0,66
60	J	201209	12,1	158,8	0,66
68	L	201401	6,4	260,3	0,51
78	N	201410	41,7	350,4	0,58

Fonte: Elaborazione su dati Oros

Come si evince dai risultati, la divisione di attività economica J59 è quella che presenta più frequentemente valori poco performanti di RMSE e CV. Tale divisione contiene le attività di produzione cinematografica e di programmi televisivi, in cui il numero delle posizioni lavorative può oscillare pesantemente da un mese all'altro, in seguito all'inizio o alla fine di produzioni cinematografiche. Tale instabilità occupazionale contrasta la tipica dinamica inerziale della variabile sulle posizioni lavorative e quindi, i modelli sottostanti non risultano particolarmente adeguati a descrivere questa dinamica nel settore cinematografico.

Seguono ora i risultati della fase di predizione, ovvero la stima del dato presunto mancante nel mese *m* per le unità i valutate come "presunte ritardatarie" nella fase precedente di predizione della lista, ossia il dato stimato $\hat{y}_{i,m}$.

La distribuzione di tali stime provvisorie, in base alla situazione informativa sulle variabili esplicative e quindi al tipo di modello applicato, sono riportati in sintesi nella tavola seguente e si riferiscono sia all'intero periodo, calcolato come una media mensile sui 33 mesi di sperimentazione e sia all'ordine del mese nel trimestre, calcolato come valore medio mensile su 11 mesi.

Tavola 13 - Distribuzione delle stime provvisorie per le unità presunte ritardatarie e per modello applicato, misurate sia nell'intero periodo e sia per ciascun mese di un trimestre. Aprile 2012 – dicembre 2014. Dati medi mensili, valori percentuali.

	Modello A	Modello B	Modello C	Nessun Modello	*Totale*
Intero periodo	57%	11%	23%	9%	*100%*
Primo mese	77%	17%	-	6%	*100%*
Secondo mese	53%	12%	25%	10%	*100%*
Terzo mese	53%	10%	27%	10%	*100%*

Fonte: Elaborazione su dati Oros

Confrontando i risultati della tavola 8 con la tavola 13, risulta una differenza tra la distribuzione dei modelli applicati sulle unità presunte assenti rispetto al set informativo dei rispondenti, ad eccezione delle stime ottenute con il modello B, che presenta percentuali simili.

Inoltre dalla tavola 13 emerge che in media per l'intero periodo, circa il 57% delle unità presunte ritardatarie sono stimate con il modello A, mentre circa l'11% con il modello B e una parte più consistente, circa il 23% con il modello C. Per una parte restante, il 9% delle unità considerate as-

senti, non è stato possibile applicare a livello micro nessuno dei modelli previsti, in quanto per tali unità sono assenti tutte le informazioni ausiliarie necessarie. La diversa distribuzione delle unità rispondenti e delle presunte ritardatarie rispetto alla tipologia di modello dipende principalmente da due fattori: il primo riguarda la struttura delle mancate risposte nei dati amministrativi e il secondo rispecchia la modalità di costruzione della lista delle unità assenti presunte attive. In particolare, riguardo il primo aspetto si ha che il terzo mese del trimestre è per costruzione caratterizzato da molte più mancate risposte rispetto ai primi due mesi, a causa della stretta vicinanza tra la data di trasmissione dei dati amministrativi (vedi §2.2) con la data di scadenza per l'invio delle dichiarazioni da parte delle imprese, che corrisponde ad un mese dalla fine del mese di competenza. Per effetto di questa tempistica i dati relativi al terzo mese vengono acquisiti dopo 15 giorni, dalla scadenza amministrativa, contro i 70 e i 40 giorni del primo e secondo mese rispettivamente. Per il secondo fattore, si ha che se un'unità è valutata come "presunta ritardataria" nel primo o nel secondo mese del trimestre, lo sarà per costruzione della lista dei presunti rispondenti anche nei mesi successivi, ossia nel secondo e/o nel terzo mese/i a meno di fattori stagionali o di eventi demografici (cfr. § 4.2.1.) Questo giustifica il fatto che per le unità relative a questa casistica, gli unici modelli utilizzabili in base alla situazione informativa sulle variabili esplicative, sono il modello B o il modello C, i quali in media hanno registrato rispettivamente il 12% e il 27% del totale delle unità imputate nel secondo mese e nel terzo mese, come indicato nella tavola 13.

Dalla stessa tavola si evince inoltre che la percentuale più alta delle stime dei dipendenti del primo mese, circa il 77%, risulta ottenuta con il modello A, il 17% con il modello B e il 6% non è risolto, mentre per le modalità di costruzione della lista, non esistono stime del primo mese per il modello C. Per le stime provvisorie dei secondi e dei terzi mesi, invece, tale distribuzione risulta più simile ai valori calcolati sull'intero periodo, che dei tre mesi ne è una sintesi.

Al fine di valutare la *performance* dei modelli in fase di applicazione del metodo, le analisi che seguono mettono a confronto alcune statistiche calcolate sulla distanza tra il dato predetto e il dato disponibile nella versione definitiva dei dati, per il raggruppamento dei rispondenti nei dati provvisori (a cui la stima viene riapplicata) e dei non rispondenti ma ritardatari (in questo caso le attive corrette, per cui si dispone del dato finale). Nell'esercizio, in media, quasi il 98% delle posizioni dipendenti stimate fanno riferimento alle unità rispondenti nei dati provvisori (tavola 14).

I risultati delle stime derivanti dall'applicazione dei modelli individuati alle due sottopopolazioni possono essere valutati tramite i valori dell'indice MAE e sono calcolati come segue:

$$MAE = \frac{\sum_i |y_{i,m}^{SD} - \hat{y}_{i,m}^{SP}|}{n_m} = \frac{\sum_i |y_{i,m}^{SD,R} - \hat{y}_{i,m}^{SP,R}|}{n_{m,R}} \pi_{m,R} + \frac{\sum_i |y_{i,m}^{SD,MR} - \hat{y}_{i,m}^{SP,MR}|}{n_{m,MR}} \pi_{m,MR} \qquad (9)$$

dove $\hat{y}_{i,m}^{SP}$ rappresenta la stima provvisoria calcolata da modello (A, B, o C), mentre $y_{i,m}^{SD}$ è il corrispondente valore della stima definitiva che si scompone come stima calcolata sui rispondenti e sui non rispondenti ($y_{i,m}^{SD,R}$, $y_{i,m}^{SP,R}$, $y_{i,m}^{SD,MR}$, $y_{i,m}^{SP,MR}$); il termine n_m indica la numerosità delle unità nel mese *m* distinte ugualmente in rispondenti ($n_{m,R}$) e non rispondenti ($n_{m,MR}$) e con $\pi_{m,R}$ e $\pi_{m,MR}$ si indicano rispettivamente il peso delle unità rispondenti e delle non rispondenti in stima provvisoria ma comunque presenti nei dati di stima definitiva calcolato sul totale delle unità rispondenti.

L'indice MAE tiene conto dei valori assoluti delle differenze e quindi misura la dimensione effettiva dell'errore a prescindere dal segno, mentre la differenza "non assoluta" tra ($y_{i,m}^{SD} - \hat{y}_{i,m}^{SP}$) può essere positiva, quando il dato definitivo è maggiore della stima (dato sottostimato) e viceversa negativa (dato sovrastimato). La misura è stata calcolata in modo separato tra i rispondenti e i non rispondenti, e i diversi valori possono essere messi in relazione utilizzando la formula (9). Per questo motivo nella tabella sono stati calcolati anche i pesi delle unità non rispondenti.

Nella tavola seguente sono riportati gli intervalli di variazione intorno alla media dei valori del MAE calcolati in media sui 33 mesi di sperimentazione per modello e distinti tra rispondenti e non

rispondenti (ossia per sei sottogruppi di unità) ed è indicato il corrispondente peso occupazionale in termini percentuali il cui totale per i sei sottogruppi di unità è pari a 100.

Tavola 14 - Intervalli dell'indice MAE e peso dei dipendenti (dato definitivo) distinti tra rispondenti e non rispondenti e per modello applicato. Totale economia, Intervalli calcolati come: valore medi +/- devianza periodo di riferimento. Aprile 2012 – dicembre 2014.

	MAE rispondenti	MAE non rispondenti	Peso rispondenti	Peso non rispondenti
modello A	0,37 - 0,48	0,56 - 1,00	89,8	1,5
modello B	0,39 - 0,51	0,57 - 1,18	6,0	0,2
modello C	1,52 - 1,86	2,31 - 2,78	1,9	0,7

Fonte: Elaborazione su dati Oros

Dai risultati della tavola 14 si evince che il grado di accostamento del modello ai dati reali misurato sul gruppo dei rispondenti, risulta molto alto nei modelli A e B, con errori di stima poco superiori allo zero, mentre nel modello C tale accostamento è più basso e gli errori sono più alti. Passando al gruppo delle unità non rispondenti (in totale il loro peso è l'2,4% della popolazione complessiva considerata), in cui il modello viene effettivamente applicato, tali errori sono leggermente più alti fino ad avere un impatto importante nel modello C. In generale i risultati del MAE sono abbastanza simili tra i due gruppi evidenziando quindi che i modelli stimati sui rispondenti ben si adattano anche alla sottopopolazione dei non rispondenti. Inoltre il peso dei rispondenti è talmente alto (circa il 98% del totale) che spesso copre gli effetti delle differenze di stima, anche se sono d'entità non proprio trascurabile, mentre lo stesso non avviene per il gruppo più esiguo dei non rispondenti che pesano poco più del 2%. In generale si osserva che per il gruppo dei non rispondenti, il metodo proposto produce risultati in termini di MAE peggiori nel modello C e migliori per il modello A.

La tavola successiva (tavola 15) invece, riporta l'indice MAE e il corrispondente peso occupazionale a livello di sezione economica e relativamente al modello A, quello più applicato.

Tavola 15 - Indice MAE e peso dei dipendenti per il gruppo dei rispondenti e non rispondenti, distinti per sezione Ateco e totale B-S, relativi solo al modello A. Aprile 2012 - dicembre 2014

Sezione Ateco	MAE rispondenti	MAE non rispondenti	Peso rispondenti	Peso non rispondenti
B	0,49	0,76	98,26	1,31
C	0,53	0,89	98,29	1,24
D	0,42	0,62	97,19	2,22
E	0,70	1,05	97,04	1,98
F	0,43	0,64	97,66	1,63
G	0,23	0,37	98,05	1,40
H	0,72	1,44	96,00	2,27
I	0,69	0,84	97,51	1,67
J	0,57	2,42	97,03	2,02
K	0,32	0,45	98,06	1,51
L	0,20	0,35	98,00	1,41
M	0,22	0,49	97,73	1,45
N	0,84	1,72	96,42	2,37
P	0,60	1,03	96,92	1,93
Q	0,35	0,67	97,31	1,96
R	0,93	1,64	95,83	2,62
S	0,24	0,37	97,73	1,54
Totale	0,43	*0,78*	97,65	1,59

Fonte: Elaborazione su dati Oros

Dalla tavola si evince che il valore MAE dei rispondenti rimane sotto la soglia di 1 dipendente in tutte le sezioni con risultati migliori per le sezioni G, M ed L, mentre nel gruppo dei non rispondenti, il cui peso occupazionale oscilla tra 1,3% e 2,6%, l'indice risulta più alto. Inoltre, nelle sezioni appena citate il MAE continua a rimanere tra i più bassi, mentre registra valori più alti della media nelle sezioni H, J, N e R.

5. L'errore di revisione

Come già accennato nei paragrafi precedenti, la situazione informativa alla base della rilevazione Oros consente di valutare, per ogni istante di stima provvisoria e ad un anno di distanza rispetto alla prima stima, l'errore di revisione ossia lo scostamento della stima preliminare dal vero valore ottenuto con la stima finale. Ciò è possibile a seguito della disponibilità di dati finali sulla popolazione di riferimento[15]. La possibilità di calcolare tale errore costituisce per la rilevazione Oros un'opportunità per valutare l'accostamento della stima provvisoria a quella finale e, nel contesto della sperimentazione, è stata utilizzata quale misura dell'accuratezza del nuovo metodo d'imputazione nelle sue diverse varianti.

In particolare, la riduzione di tale errore fino a raggiungere il valore ottimale pari a zero, ha rappresentato l'obiettivo finale a cui tendere per valutare i vari metodi di trattamento sui dati, considerando sia il confronto con il precedente metodo di correzione macro[16] sia i vari miglioramenti e/o aggiustamenti ad hoc implementati durante l'applicazione del nuovo approccio micro, nella fase di predizione della lista e in quella d'imputazione.

L'errore di revisione è calcolato nei diversi livelli di aggregazione di attività economica a partire dalle divisioni Ateco 2007[17], fino alle sezioni e all'aggregato totale B-S, con il fine di far emergere eventuali problematiche caratterizzanti specifici domini, già evidenziati dalla prima analisi sulla distribuzione dei rispondenti ritardatari (cfr. §3). Inoltre, per consentire di avere una misurazione dell'errore dovuto al metodo d'imputazione normalizzato rispetto all'errore complessivo, ossia inclusivo della stima preliminare delle sottopopolazioni delle grandi imprese e delle interinali, trattate entrambe con diversa metodologia (cfr. §2.3.1), esso viene calcolato evidenziando il contributo delle singole sottopopolazioni all'errore complessivo. In generale, l'obiettivo è di minimizzare l'errore ad un livello Ateco più disaggregato possibile.

Infine, disponendo della versione preliminare e finale dei microdati, nella sperimentazione è stato dato ampio spazio all'analisi dell'errore disaggregato per cause, consentendo di isolare in modo molto fine le aree critiche su cui migliorare l'approccio d'imputazione.

Si ricorda brevemente che il contesto informativo della stima provvisoria è caratterizzato, in termini di posizioni lavorative medie trimestrali, dal 97,5% circa di rispondenti e, quindi, il 2,5% di non rispondenti ritardatari. Tale insieme di unità, riferito alla sottopopolazione delle PMI, copre il 78% circa dell'occupazione complessiva. In tale contesto informativo, l'impatto dell'errore di stima sull'insieme di tutte le unità può dipendere da:

- la predizione dello status di assenti ritardatarie o per inattività;
- la ricostruzione dei dati sulle predette attive;
- la presenza di unità ritardatarie da oltre un anno, incluse le neonate tra *m* e *m*-12 che non

[15] Nella rilevazione Oros, le revisioni vengono effettuate per incorporare negli indicatori le informazioni che si rendono disponibili successivamente alla pubblicazione della prima stima. I principali elementi considerati nel processo di revisione sono i seguenti:
- la disponibilità dell'universo delle dichiarazioni DM2013 virtuali per la produzione della stima definitiva;
- la revisione dei dati della rilevazione mensile GI;
- l'aggiornamento di informazioni di carattere strutturale sulle unità oggetto di rilevazione e le eventuali revisioni occasionali nella metodologia di stima degli indicatori.

[16] Ossia il metodo usato per correggere il numero delle posizioni lavorative inviate ad Eurostat in forma confidenziale (regolamento STS), prima dell'uscita del comunicato stampa di giugno 2015.

[17] Nella politica Oros la divisione Ateco è il livello di aggregazione di riferimento rispetto a cui viene validata la qualità delle stime delle principali variabili *target*.

hanno mai inviato una dichiarazione contributiva (quindi assenti nella lista della popolazione iniziale su cui si valuta la predizione). A livello micro su quest'ultimo gruppo di unità non viene effettuata predizione, in quanto non si hanno riferimenti anagrafici sulla loro presenza nell'ultimo anno. Si stima che il loro impatto sull'errore finale, noto solo a posteriori, si attesti intorno ai valori 0,05-0,1% (cfr. § 4.2.1), ma può variare in modo significativo a seconda della dinamica occupazionale dovuta al periodo di congiuntura;

- infine, va considerato che i dati amministrativi della versione provvisoria possono differire dagli stessi della versione definitiva per correzioni da parte dell'ente fornitore, dovuti a differenze di processo o per l'interpretazione errata di metadati, rettificata nella versione finale. Tale componente di errore risulta solitamente di impatto trascurabile.

Le rimanenti unità della popolazione *target* della rilevazione rappresentano rispettivamente il 20% per le GI e il 2% per le interinali.

Di seguito si propone una scomposizione algebrica dell'errore di revisione sulla base delle considerazioni appena citate seguita da una descrizione dei principali risultati ottenuti nel periodo di sperimentazione.

5.1 Scomposizione dell'errore

5.1.1 Formalizzazione analitica

Considerato che la stima Y relativa al numero delle posizioni lavorative dipendenti medie mensili (a livello aggregato) può essere ottenuta per enumerazione dei dati stimati ottenuti da dati d'indagine per le GI e dati amministrativi rispettivamente per le PMI e le interinali (INTER)[18], per il generico dominio j al tempo t si ha:

$$Y_{j,t} = {}_{GI}Y_{j,t} + {}_{PMI}Y_{j,t} + {}_{INTER}Y_{j,t} \tag{10}$$

In cui la stima preliminare (SP) e la stima finale (SD) saranno espresse come:

$$Y_{j,t}^{SD} = {}_{GI}Y_{j,t}^{SD} + {}_{PMI}Y_{j,t}^{SD} + {}_{INTER}Y_{j,t}^{SD} \tag{11}$$

$$Y_{j,t}^{SP} = {}_{GI}Y_{j,t}^{SP} + {}_{PMI}Y_{j,t}^{SP} + {}_{INTER}Y_{j,t}^{SP} \tag{12}$$

L'errore di revisione relativo può essere scomposto nella parte dovuta alle tre sottopopolazioni:

$$e_{j,t} = \frac{({}_{GI}Y_{j,t}^{SP} - {}_{GI}Y_{j,t}^{SD}) + ({}_{PMI}Y_{j,t}^{SP} - {}_{PMI}Y_{j,t}^{SD}) + ({}_{INTER}Y_{j,t}^{SP} - {}_{INTER}Y_{j,t}^{SD})}{Y_{j,t}^{SD}} \tag{13}$$

ovvero:

$$e_{j,t} = \frac{({}_{GI}Y_{j,t}^{SP} - {}_{GI}Y_{j,t}^{SD})}{Y_{j,t}^{SD}} + \frac{({}_{PMI}Y_{j,t}^{SP} - {}_{PMI}Y_{j,t}^{SD})}{Y_{j,t}^{SD}} + \frac{({}_{INTER}Y_{j,t}^{SP} - {}_{INTER}Y_{j,t}^{SD})}{Y_{j,t}^{SD}} \tag{14}$$

per cui:

$$e_{j,t} = {}_{GI}e'_{j,t} + {}_{PMI}e'_{j,t} + {}_{INTER}e'_{j,t} \tag{15}.$$

In questo modo è possibile esprimere l'errore relativo totale $e_{j,t}$ come somma di componenti che rappresentano il contributo all'errore riferito alle singole sottopopolazioni, definite $e'_{j,t}$ ed

[18] Ad eccezione di eventuali errori di misura o altre questioni legate all'informativa amministrativa percepita non correttamente.

uguali al prodotto tra l'errore relativo riferito alla singola sottopopolazione moltiplicato per il corrispondente peso occupazionale, calcolato con dati di stima definitiva. Il termine $e'_{j,t}$ sinteticamente rappresenta la componente d'errore con un denominatore diverso rispetto a quello utilizzato per il calcolo dell'errore assoluto e ha il vantaggio di permettere una focalizzazione immediata sulle cause di errore più influenti. In formula:

$$e_{j,t} = {}_{GI}e'_{j,t} + {}_{PMI}e'_{j,t} + {}_{INTER}e'_{j,t} = {}_{GI}e_{j,t\ GI}\pi_{j,t} + {}_{PMI}e_{j,t\ PMI}\pi_{j,t} + {}_{INTER}e_{j,t\ INTER}\pi_{j,t} \qquad (16).$$

Focalizzando sulle cause di revisione delle stime ottenute utilizzando la fonte amministrativa e isolando la componente dovuta alle interinali (sola componente PMI), è possibile scomporre l'errore relativo corrispondente:

$${}_{PMI}e_{j,t} = {}_{PMI}e'^{Risp}_{j,t} + {}_{PMI}e'^{NoRisp}_{j,t} + {}_{PMI}e'^{R}_{j,t} \qquad (17)$$

in cui i termini Risp, NoRisp e R, indicano rispettivamente l'errore nei microdati relativi ai rispondenti nelle stime provvisorie e definitive, l'errore relativo alle mancate risposte e, infine, un errore residuale.

Tra i non rispondenti, in base alla predizione, è utile distinguere a sua volta il gruppo di unità imputate correttamente e non correttamente, definendo quindi i quattro sottogruppi: le attive corrette (AC), le attive non corrette (ANC), le non attive non corrette (NANC) e le non attive corrette (NAC). La parte residuale dell'errore, di entità trascurabile, è legata a riclassificazioni delle unità tra stima provvisoria e stima definitiva che influenzano la presenza dell'unità nei domini di stima della rilevazione (es. i *mismatching* di vario genere tra dati amministrativi e dati d'indagine) .

Indicando i microdati del numero delle posizioni lavorative dipendenti (y) per l'unità i nel settore j al tempo t con $y^{SD}_{i,j,t}$ in versione definitiva e $y^{SP}_{i,j,t}$ in versione provvisoria, nel caso di dato osservato e $\hat{y}^{SP}_{i,j,t}$ se il dato è stimato, è possibile scomporre gli errori nei contributi in somma all'errore delle singole unità, esplicitando i singoli termini della (17) nel modo seguente:

$${}_{PMI}e_{j,t} = \sum_{i\in Risp} e'^{Risp}_{i,j,t} + \sum_{i\in NoRisp} e'^{NoRisp}_{i,j,t} + \sum_{i\in Risp:SP\cap\setminus SD=0} e'^{R}_{i,j,t} \qquad (18)$$

in cui la prima componente relativa alla revisione dei microdati dei rispondenti nelle PMI può essere esplicitata come:

$${}_{PMI}e'^{Risp}_{j,t} = \frac{\sum_{i\in Risp}(y^{SP}_{i,j,t} - y^{SD}_{i,j,t})}{{}_{PMI}Y^{SD}_{j,t}} \qquad (19)$$

e rappresenta la differenza nel dato disponibile a parità di unità in SP e SD. Tale errore include eventuali correzioni dei dati da parte dell'Inps, eventuali rettifiche nella modalità di trattamento del dato nella rilevazione (errori di processo) o disponibilità di metadati più aggiornati che portano ad una diversa interpretazione del dato amministrativo nelle due versioni delle stime.

Per l'insieme delle unità assenti in SP la revisione può a sua volta essere scomposta in modo da isolare la componente dovuta all'individuazione della lista di rispondenti e quella del processo d'imputazione:

$${}_{PMI}e'^{NoRisp}_{j,t} = {}_{PMI}e'^{NoRisp,L}_{j,t} + {}_{PMI}e'^{NoRisp,\mathrm{Im}p}_{j,t} \qquad (20)$$

L'errore complessivo che riguarda l'individuazione della lista dei rispondenti ritardatari (L=*late reporter*) può essere ancora scomposto in:

$$e'^{NoRisp,L}_{j,t} = e'^{NoRisp,O}_{j,t} + e'^{NoRisp,U}_{j,t} \qquad (21)$$

dove:

$$ {}_{PMI}e'^{NoRisp,O}_{j,t} = \frac{\sum_{i \in NANC} (\hat{y}^{SP}_{i,j,t} - 0)}{{}_{PMI}Y^{SD}_{j,t}} \tag{22} $$

misura la sovra inclusione (O=*overcoverage*) dovuta all'imputazione di unità non attive (le NANC) il cui dato in SD è per definizione pari a 0;

$$ {}_{PMI}e'^{NoRisp,U}_{j,t} = \frac{\sum_{i \in ANC} (0 - y^{SD}_{i,j,t})}{{}_{PMI}Y^{SD}_{j,t}} \tag{23} $$

esprime, invece, la sotto inclusione (U=*undercoverage*) di rispondenti ritardatari trattati erroneamente come inattivi (ANC) il cui dato in SP è stimato pari a 0 e infine:

$$ {}_{PMI}e'^{NoRisp,\mathrm{Im}\,p}_{j,t} = \frac{\sum_{i \in AC} (\hat{y}^{SP}_{i,j,t} - y^{SD}_{i,j,t})}{{}_{PMI}Y^{SD}_{j,t}} \tag{24} $$

indica l'errore dovuto all'imputazione dei microdati per le unità correttamente trattate come mancate risposte (AC), ossia l'errore dovuto alla metodologia di ricostruzione del dato mancante.

Segue che l'errore riguardante l'individuazione della lista dei rispondenti ritardatari può essere indicato anche come compensazione dei dipendenti relativi all'insieme NANC e ANC in questo modo:

$$ e'^{NoRisp,L}_{j,t} = e'^{NoRisp,O}_{j,t} + e'^{NoRisp,U}_{j,t} = \frac{\sum_{i \in NANC} \hat{y}^{SP}_{i,j,t} - \sum_{i \in ANC} y^{SD}_{i,j,t}}{{}_{PMI}Y^{SD}_{j,t}} \tag{25} $$

Questo approccio, che consente di scendere nel merito delle singole fonti di errore e di focalizzare meglio sulle criticità, è praticabile se è possibile disporre di informazioni dettagliate e controllabili a livello di micro dato.

Tali componenti d'errore sono state calcolate nel periodo di sperimentazione considerato (aprile 2012-dicembre 2014) per divisone Ateco e sue successive aggregazioni e sintetizzate attraverso delle misure statistiche. In particolare, è stata considerata la media in serie storica dei valori assoluti e dei valori in segno definite, rispettivamente, come indici MAR (*Mean Absolute Revision error*) e indici MR (*Mean Revision error*). Le prime evidenziano l'entità o l'ampiezza media della revisione, mentre le altre indicano la direzione dell'errore. In formule, tali misure sono:

$$ MR_j = 100 \cdot n^{-1} \sum_{t=1}^{n} e_{j,t} \tag{26} $$

$$ MAR_j = 100 \cdot n^{-1} \sum_{t=1}^{n} |\, e_{j,t} \,| \tag{27} $$

dove n indica il numero dei trimestri. Al fine di valutare il miglioramento delle stima ottenuto grazie all'applicazione del metodo d'imputazione viene evidenziato anche l'errore di revisione iniziale, prima di qualsiasi aggiustamento sui dati (errore iniziale o d'origine). Nell'analisi che segue, l'errore viene dapprima analizzato nella sola sottopopolazione delle PMI (per cui il valore ${}_{PMI}Y^{SD}_{j,t}$ viene posto al denominatore), per poi essere successivamente collocato nel contesto delle stime totali (per cui il valore $Y^{SD}_{j,t}$ viene posto al denominatore), in cui sono incluse anche le grandi imprese e le interinali.

5.1.2 Principali risultati

La tavola sotto riporta i risultati in sintesi, calcolati con gli indici (26) e (27), delle componenti d'errore definite sopra, sull'intero periodo di sperimentazione relativi solo alle unità PMI, soggette al nuovo metodo di correzione micro e classificate per sezione economica e per il totale B-S, oltre una colonna dei pesi occupazionali per sezione.

Tavola 17 - Indici sintetici degli errori di revisione pre correzione (errore iniziale) e post imputazione micro (errore finale) e peso occupazionale dei dipendenti, per sezione Ateco e per il totale economia nella sottopopolazione delle PMI, distinti anche per cause di errore. Aprile 2012 - dicembre 2014, valori percentuali.

Sezione Ateco	peso dipendenti per sezione sul totale Oros	MR Errore iniziale	MAR Errore finale	MR Errore finale	MR Errore finale derivante da:		MR Errore per lista distinto in:	
					modello	lista	Componente sovra inclusione	Componente sotto inclusione
B	0,19	-1,80	0,30	0,12	-0,02	0,14	0,50	-0,36
C	29,16	-1,80	0,26	0,25	-0,04	0,29	0,64	-0,35
D	0,32	-2,85	0,40	-0,07	-0,02	-0,06	0,36	-0,42
E	1,40	-3,06	0,31	-0,27	-0,05	-0,22	0,58	-0,81
F	9,43	-2,55	0,51	0,50	-0,10	0,60	1,26	-0,66
G	17,93	-2,09	0,20	0,14	-0,05	0,20	0,68	-0,49
H	6,22	-4,43	0,66	-0,57	-0,22	-0,35	1,11	-1,46
I	8,38	-2,75	0,32	0,20	-0,10	0,30	1,11	-0,81
J	3,23	-3,22	0,43	-0,15	-0,15	0,00	0,85	-0,85
K	1,85	-2,06	0,18	0,07	-0,03	0,10	0,46	-0,37
L	0,75	-2,29	0,35	0,29	-0,15	0,44	1,07	-0,63
M	4,72	-2,46	0,28	-0,07	-0,10	0,02	0,70	-0,68
N	6,52	-4,02	0,38	-0,19	-0,21	0,02	1,17	-1,15
P	0,73	-3,32	0,43	-0,24	-0,10	-0,14	0,80	-0,94
Q	5,68	-2,82	0,36	0,13	-0,07	0,20	0,73	-0,53
R	1,13	-4,59	0,66	-0,11	-0,26	0,14	1,53	-1,39
S	2,36	-2,53	0,30	0,24	-0,11	0,35	1,01	-0,66
TOTALE	*100,00*	*-2,54*	*0,26*	*0,11*	*-0,09*	*0,20*	*0,84*	*-0,64*

Fonte: Elaborazione su dati Oros

La prima colonna riporta i pesi occupazionali e poi a seguire vi è l'errore di revisione iniziale, calcolato come indice MR. L'indicatore è sempre negativo, ad indicare l'assenza dalla stime provvisorie non trattate, del dato sul numero delle posizioni lavorative per le mancate risposte.

Le colonne successive riportano l'errore finale, espresso sia come MR sia come MAR, per mettere in luce sia l'entità che la direzione dell'errore ottenuto dopo la correzione. Nella media degli 11 trimestri l'applicazione del metodo di correzione porta ad un abbattimento dell'errore complessivo da -2,54% ad un valore positivo di leggera sovra imputazione pari a 0,11%, mentre in termini di ampiezza si attesta intorno allo 0,26%. A livello settoriale, un notevole miglioramento si osserva nella sezione R, in cui si passa da un errore inziale molto rilevante (-4,6%) fino ad arrivare ad un valore pressoché nullo in termini di MR (-0,1%), lievemente più alto in termini di MAR (0,7%) e in modo simile anche nelle sezioni D, K e M che presentano errori post correzione quasi nulli. In altri settori, invece, come quello delle costruzioni (sezione F) che ha un peso occupazionale rilevante, l'errore finale registra una sovra imputazione non trascurabile di 0,5% sia in termini di MAR che di MR; il settore dei trasporti (sezione H) con errore finale negativo pari a -0,6% è quello che include i trasporti marittimi, caratterizzati da ritardo strutturale di invio dei modelli a seguito di deroghe

nella scadenza. Il settore J infine, relativo ai servizi di informazione e comunicazione, contenente attività di produzione cinematografica e programmi televisivi in cui la dinamica occupazionale ha caratteristiche di elevatissima volatilità, registra un errore medio pre correzione superiore al 3% ed un errore post correzione pari a -0,15%, con un ampiezza invece tra le più alte, oltre lo 0,4%. Questo dipende da una serie di bilanciamenti tra valori positivi e negativi dell'errore finale, che si succedono nei vari trimestri e riflette appunto l'elevata volatilità dell'occupazione in questo settore.

Nella stessa tavola l'errore finale delle PMI viene disaggregato anche per "causa" (espresso nella formula 20), evidenziando l'errore dovuto alla predizione della lista e quello generato dall'imputazione da modello. Non si prendono, invece, in considerazione gli errori per revisione dei microdati e residuali in quanto nulli nel periodo di sperimentazione.

Dai risultati si evince che la componente predominante della causa dell'errore è quella dovuta all'individuazione della lista, che nel totale ha un contributo sull'errore finale pari a 0,2%, contro un'incidenza molto bassa dello 0,09%, dovuta alla componente per l'imputazione tramite modello di regressione.

A sua volta, la componente dell'errore dovuto alla lista deriva dal bilanciamento delle due componenti di sovra inclusione e sotto inclusione (espressi nelle formule (21), (22) e (23)). In particolare, la componente dell'errore per lista che predomina nell'aggregato del totale economia, è quella di sovra inclusione, per cui in media registra un valore di 0,84%, contro un valore negativo di 0,64% della sotto inclusione; mentre a livello di singoli settori una componente prevale sull'altra in modo alternato. Nei settori C, F, G ed I, il cui peso occupazionale ricopre circa il 65% del totale, predomina la componente di sovra inclusione e quindi, ne consegue un contributo positivo alla componente di errore per lista. A sua volta, tale errore si bilancia con quello negativo derivante dalla fase d'imputazione da modello ma, essendo quest'ultimo di piccola entità, il risultato dell'errore finale è di sovra imputazione, in quanto è positivo.

Nel settore dei trasporti (sezione H), già evidenziato sopra per la presenza di un alto tasso di unità ritardatarie e con peso occupazionale rilevante, emerge un valore alto per la componente di sotto inclusione dovuta alla persistenza nel ritardo, in particolare nel settore dei traporti marittimi, che porta a registrare un dato negativo e rilevante per l'errore dovuto alla lista e quindi anche per l'errore finale.

Nella tavola seguente, per il solo aggregato totale B-S, si riporta un'analisi in serie storica degli errori come sopra sintetizzati, riferiti sempre alla sottopopolazione delle PMI.

Tavola 18 - Serie storica degli errori di revisione pre correzione (errore iniziale) e post correzione (errore finale) micro del totale economia, B-S nella sottopopolazione delle PMI, distinto per cause di errore. Secondo trimestre 2012 – quarto trimestre 2014, valori percentuali.

	Errore iniziale	Errore finale	Errore finale derivante da:		Errore per lista distinto in:	
			modello	lista	Componente sovra inclusione	Componente sotto inclusione
2012Q2	-2,68	-0,12	-0,10	-0,01	0,56	-0,57
2012Q3	-2,15	0,26	-0,05	0,31	0,92	-0,61
2012Q4	-2,30	0,48	-0,08	0,56	1,03	-0,47
2013Q1	-3,39	0,55	-0,09	0,64	1,23	-0,59
2013Q2	-2,48	-0,17	-0,08	-0,09	0,79	-0,88
2013Q3	-1,75	0,22	-0,04	0,26	0,91	-0,65
2013Q4	-2,60	0,15	-0,12	0,27	0,85	-0,58
2014Q1	-2,72	0,41	-0,10	0,51	1,02	-0,51
2014Q2	-2,88	-0,26	-0,14	-0,12	0,69	-0,81
2014Q3	-2,44	-0,02	-0,09	0,07	0,76	-0,69
2014Q4	-2,59	-0,28	-0,12	-0,16	0,51	-0,67
Valore medio	*-2,54*	*0,11*	*-0,09*	*0,20*	*0,84*	*-0,64*

Fonte: Elaborazione su dati Oros

Dai risultati si può osservare che nel periodo di riferimento i livelli delle revisioni sono molto omogenei, non essendo presenti situazioni con valori molto elevati e ciò dipende dal consolidamento del processo ottenuto nel corso degli anni, sia per la fonte dei dati e sia per il loro trattamento. Fa eccezione il primo trimestre del 2013, per cui si rilevano gli errori iniziali e finali più rilevanti, in corrispondenza del cambiamento della situazione informativa di base dovuta all'introduzione dei nuovi modelli Inps DM2013 virtuali, alla cui transizione è seguito un breve periodo di consolidamento delle procedure informatiche da parte dell'ente, che hanno inciso sulla disponibilità di microdati per le stime provvisorie.

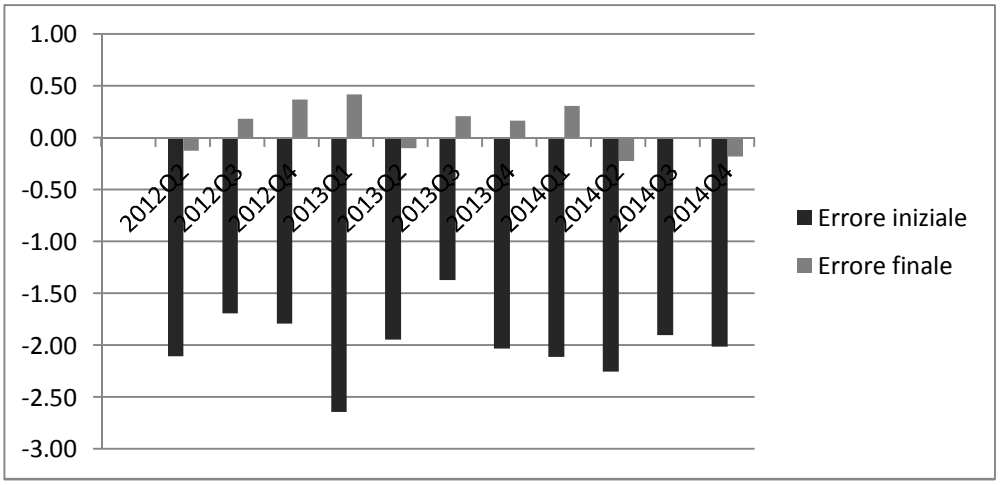
In riferimento alle revisioni per cause di errore, si ha che in tutti i trimestri la revisione dovuta alla predizione della lista è più rilevante rispetto a quella dovuta all'imputazione da modello ed inoltre nella fase di predizione l'errore di sovra inclusione supera l'errore di sotto inclusione in sette trimestri su undici, evidenziando una certa prevalenza di tale forma di errore.

L'analisi successiva contestualizza l'errore da metodo d'imputazione nelle stime complessive di Oros, ossia inclusive delle grandi imprese e delle unità interinali.

Il grafico sotto (figura 13) evidenzia l'errore di revisione pre e post imputazione riferito al totale della popolazione e al totale economia, in cui il processo di correzione è quello micro per le PMI, è quello specifico della rilevazione GI (revisione annuale) utilizzato per le unità grandi imprese ed è quello ad hoc usato in Oros per le unità interinali.

Nel corso dei trimestri di sperimentazione vi è un'alternanza di valori positivi e negativi dell'errore finale post correzione, evidenziando quindi l'assenza di sistematicità nell'errore. Quello che emerge dal grafico è la notevole riduzione dell'errore iniziale grazie al metodo d'imputazione, con il caso più critico riferito sempre al primo trimestre del 2013, ossia con l'entrata dei nuovi modelli virtuali.

Figura 13 - Serie storica degli errori di revisione pre correzione (errore iniziale) e post correzione (errore finale) per il totale economia, B-S. Secondo trimestre 2012 – quarto trimestre 2014, valori percentuali.



Fonte: Elaborazione su dati Oros

La serie storica degli errori di revisione disaggregati nelle componenti delle tre sottopopolazioni è riportata nella tavola successiva. L'errore medio iniziale scende da -1,98 a +0,1 per effetto dell'errore pressoché nullo delle GI, il cui peso in termini occupazionali è del 20% e l'errore trascurabile sulle interinali, le quali hanno un'incidenza rilevante solo nella divisione Ateco N78 (circa il 93%) e un peso del 21% nella sezione corrispondente, mentre nel totale economia ricoprono un peso poco meno del 2%.

Tavola 19 - Serie storica degli errori di revisione pre correzione (errore iniziale) e post correzione (errore finale) nelle sottopopolazioni: GI,PMI ed INTER e peso occupazionale delle PMI, per il totale economia, B-S. Secondo trimestre 2012 – quarto trimestre 2014, valori percentuali.

	Errore iniziale	Errore finale	Peso PMI	Errore finale distinto in:		
				Errore GI	Errore PMI	Errore Interinali
2012Q2	-2,10	-0,12	78,39	0,00	-0,09	-0,03
2012Q3	-1,69	0,19	78,36	0,00	0,21	-0,02
2012Q4	-1,79	0,37	78,07	0,00	0,37	0,00
2013Q1	-2,64	0,42	77,75	-0,01	0,42	0,01
2013Q2	-1,94	-0,10	78,18	0,03	-0,13	0,00
2013Q3	-1,37	0,21	78,17	0,04	0,17	0,00
2013Q4	-2,03	0,17	77,88	0,04	0,11	0,02
2014Q1	-2,11	0,31	77,61	-0,01	0,32	0,00
2014Q2	-2,25	-0,22	78,07	0,00	-0,20	-0,02
2014Q3	-1,90	0,01	78,10	0,02	-0,01	0,00
2014Q4	-2,01	-0,18	77,81	0,04	-0,21	-0,01
Valore medio	*-1,98*	*0,10*	*78,04*	*0,01*	*0,09*	*0,00*

Fonte: Elaborazione su dati Oros

6. Validazione dei micro dati

Un ultimo aspetto valutato in questo documento riguarda la qualità dell'imputazione dei dati sul numero delle posizioni lavorative dipendenti a livello di micro dato. I micro dati prodotti da Oros infatti, non solo concorrono alle stime aggregate delle variabili *target*, ma rappresentando anche sui trimestri provvisori la popolazione complessiva, sono sempre più richiesti ed utilizzati per analisi specifiche interne all'istituto o quale fonte ausiliaria nella stima di variabili correlate o in processi di editing ed imputazione.

La valutazione viene effettuata sui dati da fonte amministrativa relativi alle sole PMI, a livello di dato medio trimestrale (ordine di misura con cui viene rilasciato o utilizzato per stime aggregate), che può essere inclusivo o meno, di uno o più mesi stimati per imputazione. La presenza d'imputazione in uno dei mesi del trimestre emerge dalla differenza tra il dato provvisorio (eventualmente inclusivo d'imputazione) e il corrispondente dato definitivo. Tale differenza potrebbe essere causata anche da "errori di misura" registrati sulle rispondenti, i quali sono normalmente indotti dal processo di stima Oros (un diverso trattamento dei dati nelle stime provvisorie e definitive) e sono, a meno di casi eccezionali e ben identificati, di entità trascurabile.

Rifacendosi allo schema sullo status delle unità (cfr. § 4.2.1, figura 11), nella valutazione a livello micro sono incluse oltre alle rispondenti, le unità per cui in stima provvisoria è stato possibile effettuare predizione in almeno un mese del trimestre ossia le attive corrette e le non attive non corrette, mentre le unità non attive corrette e attive non corrette, se interessano l'intero trimestre, non sono incluse nel set di riferimento.

La tavola seguente evidenzia alcune informazioni riferite al periodo di sperimentazione, ottenute come medie trimestrali calcolate sulle unità imputate distinte per sezione di attività economica.

Tavola 20 – Media rapporti caratteristici, errori pre e post imputazione e misura dell'abbattimento dell'errore calcolati sui microdati nelle sezioni ateco da B a S. Secondo trimestre 2012 – quarto trimestre 2014, valori percentuali e assoluti.

Sezione Ateco	Quota di unità oggetto di imputazione sulle unità in stima provvisoria	Quota di dipendenti imputati sui dipendenti in stima provvisoria	media errore pre imputazione	media errore post imputazione
B	6,25	1,91	-2,4	0,7
C	5,85	2,05	-3,1	1,1
D	5,95	2,77	-5,2	0,7
E	5,99	2,79	-6,8	1,5
F	8,71	3,02	-1,1	0,5
G	6,63	2,23	-1,2	0,4
H	8,74	3,87	-3,9	1,0
I	8,32	2,88	-1,1	0,5
J	6,63	3,07	-3,3	0,7
K	5,54	2,11	-2,2	0,6
L	6,47	2,55	-0,8	0,3
M	5,53	2,38	-1,4	0,4
N	8,55	3,83	-3,9	1,1
P	7,10	3,10	-2,4	0,5
Q	5,35	2,95	-3,5	0,8
R	9,20	4,45	-2,3	0,7
S	7,18	2,76	-0,9	0,3
Totale	*7,04*	*2,64*	*-1,9*	*0,6*

Fonte: Elaborazione su dati Oros

La prima colonna riporta la percentuale delle unità oggetto d'imputazione rispetto a tutte le unità presenti in stima provvisoria (ossia rispetto alle unità rispondenti più le unità imputate), mentre la seconda indica la stessa percentuale in termini di dipendenti, in cui si evince che le sezioni in cui l'imputazione è stata più rilevante sono F, H, I, J, N, P e R.

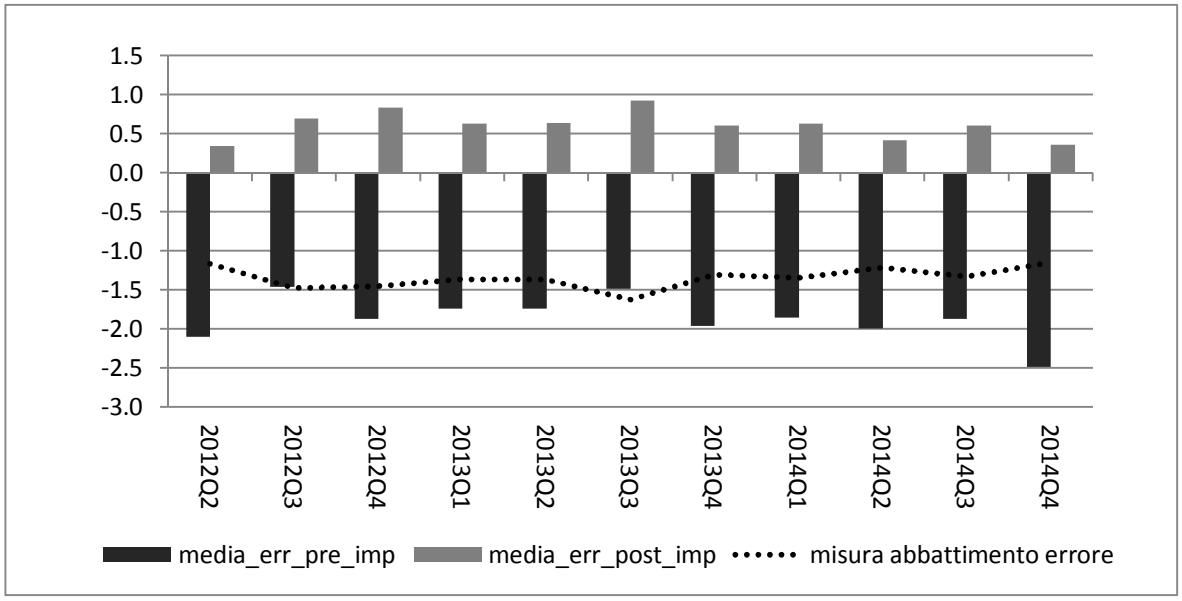
La tavola include anche la media degli errori con segno (differenze tra dato provvisorio e dato finale) calcolati sulle singole unità e, in particolare: l'errore iniziale prima del processo di correzione (errore pre imputazione) e l'errore finale post correzione (errore post imputazione). Dai risultati si osserva che l'errore medio iniziale è molto variabile tra le sezioni e quelle che registrano i valori negativi più alti sono le sezioni E, D, H ed N in cui prima dell'imputazione vi era una sottostima di circa 7 dipendenti nel caso della sezione E, circa di 5 dipendenti per D e di 4 dipendenti sulle altre due, per arrivare ad un errore medio di segno positivo pari a 1,5 per E, di 0,7 per D e pari ad 1 per H e N. L'errore finale post imputazione risulta in media sempre positivo, principalmente per effetto del parziale bilanciamento della componente riferita alle unità attive non corrette e alle unità non attive non corrette.

Nel totale dell'aggregato B-S, il grafico sotto (figura 14) riporta la serie storica degli errori pre e post imputazione e il loro rapporto come misura di abbattimento dell'errore iniziale, definito come:

$$\frac{err\_post\_imp}{err\_pre\_imp} - 1 \qquad (28)$$

Quindi si ha che più tale valore si avvicina a -1 e più l'errore iniziale viene ridotto a seguito dell'imputazione. A livello micro esiste una componente di sovra imputazione dovuta alle unità presunte ritardatarie presenti nella lista anagrafica che, in realtà, non sono più attive e, quindi, tale componente si ripercuote col segno positivo sull'errore d'imputazione e con il segno negativo sulla misura dell'abbattimento dell'errore definita sopra. Tale componente, come visto nell'analisi sull'errore di revisione, è sistematica e strutturale.

Figura 14 - Serie trimestrale della media degli errori pre e post imputazione calcolati a livello micro e misura dell'abbattimento dell'errore nel totale B-S. Secondo trimestre 2012 – quarto trimestre 2014, valori percentuali.



Fonte: Elaborazione su dati Oros

Facendo un focus su un particolare trimestre, è possibile analizzare i risultati anche in termini di Ateco. Nella tavola successiva, si riportano alcuni risultati relativi al terzo trimestre del 2014 in cui l'errore di revisione sull'aggregato totale, ottenuto da media pesata dei singoli errori di revisione per divisione Ateco, ha registrato un valore nullo, mentre a livello di singola divisione emergono situazioni molto variabili. In particolare, nella tavola sono indicate alcune situazioni più estreme come le divisioni H50 e C19, che hanno alte percentuali d'imputazione. Tuttavia, mentre nella divisione C19 l'errore finale post imputazione micro rimane alto, la divisione H50 presenta invece, grazie alla correzione, un errore finale quasi nullo. Altre divisioni sono le J59, K65, R91 che partono da un errore in origine molto alto, fino ad arrivare anche al 30% nella divisione R91 in cui, a seguito dell'imputazione, l'errore si riduce notevolmente fino a valori sotto l'1%.

Tavola 21 - Focus su particolari divisioni Ateco misurate nel terzo trimestre 2014 con alcuni rapporti caratteristici, misure dell'errore pre e post imputazione calcolato a livello micro e una misura dell'abbattimento dell'errore. Valori percentuali.

Divisione ATECO	Quota di unità oggetto di imputazione sulle unità in stima provvisoria	Quota di dipendenti imputati sui dipendenti in stima provvisoria	Errore pre imputazione	Errore post imputazione	misura abbatti-mento errore
19	6,3	23,9	-9,8	85,7	-9,77
50	15,4	13,1	-4,5	-0,1	-0,98
59	10,9	11,2	-11,6	0,6	-1,05
65	2,8	0,5	-13,4	0,4	-1,03
91	8,9	13,9	-30,2	0,9	-1,03

Fonte: Elaborazione su dati Oros

Nella tavola che segue infine, sono riportati i risultati di sperimentazione sugli errori micro (pre e post imputazione) per dimensione media d'impresa, in cui si evidenziano le differenze tra le unità oggetto di imputazione e tutte le unità PMI, ossia le unità per cui almeno un mese del trimestre è stato oggetto d'imputazione e quelle non interessate da imputazione nel trimestre perché sempre rispondenti, per l'intero aggregato B-S.

In media per le unità oggetto d'imputazione che hanno una dimensione media nel periodo considerato tra i 5 e i 6 dipendenti e con un errore medio iniziale negativo di circa 2 dipendenti, si arriva con il processo d'imputazione, ad un errore di segno positivo tra 0,34 a 0,93.

Le stesse misure calcolate per tutte le unità evidenziano quasi sempre una dimensione media d'impresa più grande, fino ad una differenza di 1,2 dipendenti, mentre gli errori sono di gran lunga inferiori, in alcuni trimestri anche nulli, anche per effetto del ridotto peso delle mancate risposte sul totale.

Tavola 22 - Serie trimestrale della dimensione media d'impresa, della media degli errori micro pre imputazione e post imputazione, distinti tra le unità oggetto d'imputazione e tutte le unità PMI. Secondo trimestre 2012- quarto trimestre 2014.

	IMPUTATE			TUTTE		
	Dimensione media dipendenti	Errore pre-imputazione	Errore post-imputazione	Dimensione media dipendenti	Errore pre-imputazione	Errore post-imputazione
2012Q2	5,49	-2,1	0,34	5,98	-0,16	-0,01
2012Q3	4,73	-1,46	0,69	6,04	-0,13	0,02
2012Q4	6,12	-1,87	0,84	5,99	-0,14	0,03
2013Q1	5,30	-1,74	0,63	6,04	-0,16	0,07
2013Q2	5,36	-1,74	0,64	6,07	-0,14	0,00
2013Q3	4,95	-1,49	0,93	6,13	-0,10	0,02
2013Q4	5,76	-1,96	0,60	6,11	-0,15	0,02
2014Q1	5,17	-1,85	0,63	6,12	-0,17	0,02
2014Q2	5,36	-1,99	0,42	6,20	-0,18	-0,02
2014Q3	5,42	-1,87	0,61	6,26	-0,15	0,00
2014Q4	5,98	-2,48	0,36	6,21	-0,16	-0,02

Fonte: Elaborazione sui dati Oros

Per avere anche a livello micro un errore in termini relativi, escludendo gli zeri al denominatore che si presentano nel caso di dato definitivo inesistente per le unità oggetto d'imputazione non attive non corrette, è stato utilizzata la seguente misura:

$$err\_imp\_rel_t = \frac{\hat{y}_{SPt} - y_{SDt}}{\frac{1}{2}(\hat{y}_{SPt} + y_{SDt})} \qquad (29)$$

che è incluso in un intervallo di valori tra -2 come valore minimo e 2 come valore massimo.

Il minimo dell'errore relativo viene raggiunto quando il dato provvisorio $\hat{y}_{SP}$ è nullo, in quanto erroneamente non imputato, mentre il suo valore massimo si realizza viceversa, quando è assente il dato definitivo ma il provvisorio è stato imputato perché o considerato un presunto ritardatario, o (ma è molto meno frequente) per cause dovute a fattori normativi e/o amministrativi non valutati correttamente.

In media si ha che in quasi tutte le sezioni economiche la percentuale di unità con errore vicino allo zero sulle unità oggetto d'imputazione supera il 50%, ossia il dato imputato presenta una differenza rispetto a quello finale trascurabile, mentre si osserva che i casi con errore relativo positivo sono di gran lunga superiore a quelli con valori negativo, sempre per effetto dell'assenza dall'analisi delle unità non attive non corrette.

7. Conclusioni

Questo lavoro documenta la lunga fase di sperimentazione che ha permesso di mettere a regime un metodo di stima preliminare del numero delle posizioni lavorative dipendenti nella rilevazione trimestrale Oros, valorizzando le potenzialità della base di microdati amministrativi disponibile in tempi rapidi. Disponendo di un universo "quasi completo" di rispondenti, le stime provvisorie vengono effettuate considerando l'intera popolazione pervenuta per l'istante di stima e imputando i dati delle unità ritardatarie, con il fine di correggere l'errore di sottostima dei livelli derivante dall'assenza di unità attive. Tale approccio consente di catturare con elevato livello di aggiornamento gli effetti della demografia d'impresa sulla dinamica della variabile *target*, essendo le neonate e cessate o sospese ben rappresentate nei dati disponibili. Tuttavia, a causa di un problema di ampia sovra-copertura che caratterizza la lista anagrafica delle unità attive, dovuta a ritardi nella registrazione degli eventi di cessazione/sospensione dell'attività nei dati amministrativi, e in assenza di una lista teorica di rispondenti, un problema cruciale consiste nella predizione dello status di attività alle unità assenti. Il metodo proposto basa l'attribuzione di tale status sulla valutazione del *pattern* di presenza dell'unità nei periodi adiacenti a quello di stima, data l'evidenza generalizzata sulla non persistenza dei ritardi di risposta. A seguito di una serie di rifiniture sulle regole di attribuzione dello status di attività, il metodo proposto consente di ottenere un buon bilanciamento tra unità definite erroneamente attive (sovra inclusione della lista) e unità definite erroneamente non attive (sotto inclusione della lista), con una leggera prevalenza della prima tipologia di errore. In seguito, sulle unità definite attive, viene applicato un modello di regressione per ricostruire il dato mancante, in cui le covariate sono i valori della variabile stessa osservati nel mese precedente e nello stesso mese dell'anno precedente. La caratteristica inerziale che contraddistingue l'evoluzione delle posizioni lavorative in molti settori, implica una buona rappresentazione del fenomeno oggetto di studio con una lieve tendenza a sottostimare i livelli dell'occupazione finale. Come sintesi delle due fonti di errore, emerge una leggera prevalenza dell'errore di lista rispetto a quello dovuto alla regressione.

L'errore di lista è più rilevante nei settori in cui sono più frequenti eventi di demografia d'impresa, come il settore delle costruzioni. L'errore di regressione è, invece, rilevante in settori in cui la modellizzazione è meno semplice, a causa della volatilità dell'occupazione sottostante (tipica di alcuni settori come la produzione cinematografica e televisiva). Rimane di totale impatto, anche se di entità molto ridotta, l'errore di sottostima causato dalle unità assenti dalla lista anagrafica di partenza (neonate che rispondono con ritardo prima di iscriversi oppure unità sospese da oltre un anno e quindi tagliate dalla lista delle potenziali attive). Il nuovo metodo genera stime provvisorie di buona qualità nella generalità dei settori coperti da Oros. La disponibilità di microdati definitivi ad un anno di distanza da quelli provvisori consente di focalizzare con molta attenzione sulle singole fonti di errore con potenzialità di affinamento del metodo enormi. Le analisi effettuate durante la sperimentazione hanno evidenziato che, a situazione informativa stabile, vi sono ancora margini di miglioramento, sia in termini di piccole finiture che potrebbero comportare lievi riduzioni diffuse dell'errore, sia di interventi mirati a situazioni che si sono rivelate particolarmente critiche: si tratta, in particolare, di settori in cui la maggiore variabilità dell'occupazione e/o la presenza di comportamenti di ritardo persistente (trasporti marittimi) rendono il metodo progettato poco adattabile. In questi casi si continua a fare ricorso ad aggiustamenti a livello macro, basati sull'analisi in serie storica dell'occupazione del settore in relazione all'andamento congiunturale complessivo.

Riferimenti bibliografici

AA.VV. 2008. Seminario: Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali. *Contributi Istat*. 13/2008: 29-61. http://www3.istat.it/dati/pubbsci/contributi/Contributi/contr_2008/13_2008.pdf.

Baldi, C., F. Ceccato, E. Cimino, M.C. Congia, S. Pacini, F. Rapiti e D. Tuzi. 2004. Use of Administrative Data to produce Short Term Statistics on Employment, Wages and Labour Cost, *Essays*, 15. Istat, Roma.

Baldi, C., D. Bellisai, F. Ceccato, S. Pacini, L. Serbassi, M. Sorrentino and D. Tuzi. 2011a. The system of short term business statistics on labour in Italy. The challenges of data integration. Paper presentato al Workshop: *ESSnet Data Integration*, Madrid 24-25 novembre. http://www.ine.es/e/essnetdi_ws2011/ppts/Baldi_et_al.pdf.

Baldi, C., M.C. Congia, S. Pacini and D. Tuzi. 2011b. The quarterly employment estimates in Italy based on the employment register. SGA 2010: Deliverable 4.7. *Timeliness of Administrative Sources for Monthly and Quarterly Estimates*. http://ec.europa.eu/eurostat/cros/system/files/SGA%202010_Deliverable_4.7.pdf.

Baldi, C., F. Ceccato, S. Pacini, and D. Tuzi. 2012. *The Use of Administrative Data for Short Term Business Statistics: Lessons from a Cross-Country Experience*. http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/view/2167.

Eisenhauer, Joseph G. 2003. *Regression through the Origin. Teaching Statistics*. Volume 25, Number 3.

De Waal, T. and P. Vlag. 2012. Milestone 4.9: The Use of Incomplete Admin Data for STS: Guidelines. Wp4. *Timeliness of Administrative sources for Monthly and Quarterly Estimates*.

Istat. 2013. *Il sistema degli indicatori congiunturali sulla domanda di lavoro e le retribuzioni in Ateco 2007 e base 2005*. Metodi. Roma. http://www.istat.it/it/archivio/97314.

Istat. 2015. *I nuovi indicatori sulle posizioni lavorative dipendenti nell'industria e nei servizi privati*. Nota informativa. Giugno 2015. http://www.istat.it/NotaInformativa_posizioni_lavorative_17-06-2015_definitiva_rivista.pdf.

Kasprzyk, D., G. Duncan and M. P. Singh (eds.). 1989. Panel Surveys, *John Wiley and Sons*, 400–425.

Maasing, E., T. Remes, C. Baldi and P. Vlag. 2012. STS estimates based solely on administrative data: final results and recommendations. SGA 2011: Deliverable 4.1. ESSnet use of administrative and accounts data in business statistics. Wp4. *Timeliness of administrative sources for monthly and quarterly estimates*. https://ec.europa.eu/eurostat/cros/system/files/SGA%202011_Deliverable_4.1.pdf.

Informazioni per le autrici e per gli autori

La collana è aperta alle autrici e agli autori dell'Istat e del Sistema statistico nazionale e ad altri studiosi che abbiano partecipato ad attività promosse dall'Istat, dal Sistan, da altri Enti di ricerca e dalle Università (convegni, seminari, gruppi di lavoro, ecc.).

Coloro che desiderano pubblicare su questa collana devono sottoporre il proprio contributo al Comitato di redazione degli *Istat working papers*, inviandolo per posta elettronica all'indirizzo: iwp@istat.it.

Il saggio deve essere redatto seguendo gli standard editoriali previsti (disponibili sul sito dell'Istat), corredato di un sommario in Italiano e in Inglese e accompagnato da una dichiarazione di paternità dell'opera.

Per le autrici e gli autori dell'Istat, la sottomissione dei lavori deve essere accompagnata da un'e-mail della/del propria/o referente (Direttora/e, Responsabile di Servizio, etc.), che ne assicura la presa visione.

Per le autrici e gli autori degli altri Enti del Sistan la trasmissione avviene attraverso la/il responsabile dell'Ufficio di statistica, che ne prende visione. Per tutte le altre autrici e gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione.

Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Attraverso il Comitato di redazione, tutti i lavori saranno sottoposti a un processo di valutazione doppio e anonimo che determinerà la significatività del lavoro per il progresso dell'attività statistica istituzionale.

La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line gratuitamente.

Gli articoli pubblicati impegnano esclusivamente le autrici e gli autori e le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.