Contributi ISTAT

La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento

Ciro Baldi, Francesca Ceccato, Silvia Pacini, Donatella Tuzi

Sintesi

Dalla metà del 2004 la rilevazione OROS, accanto agli indicatori di retribuzione fornisce ad Eurostat, in forma confidenziale, una stima delle *posizioni lavorative*. Entro tempi brevi la diffusione di questo indicatore dovrebbe avvenire anche a livello nazionale. La pubblicazione del nuovo indice impone il raggiungimento di requisiti di qualità particolarmente stringenti, soprattutto per quel che riguarda le stime provvisorie. A tal fine, nel corso degli ultimi mesi la metodologia utilizzata per la produzione delle stime preliminari è stata sottoposta a un programma molto ampio di verifiche ed approfondimenti, che hanno condotto alla definizione e alla sperimentazione di una serie di innovazioni volte a mettere a punto un impianto metodologico caratterizzato da un elevato grado di affidabilità.
La metodologia attualmente implementata produce un errore non trascurabile nelle stime preliminari, che è riconducibile ad almeno due fattori: la sovracopertura dell'anagrafica utilizzata per la stima della popolazione corrente e l'instabilità nella dimensione del campione, sia in termini di unità che di occupazione.
Nel presente documento si illustrano i principali risultati delle sperimentazioni condotte sulla procedura di stima preliminare. Gli studi effettuati hanno consentito di apportare delle modifiche alla metodologia di base, comportando notevoli progressi nella risoluzione del problema della sovracopertura. In riferimento alla natura non casuale del campione, gli approfondimenti realizzati hanno condotto alla individuazione di alcune aree critiche della metodologia di stima, da cui può avere origine l'errore attribuibile al riporto all'universo. Su questo aspetto, occorrerà effettuare ulteriori sperimentazioni.

Abstract

Since half 2004 OROS Survey supplies to Eurostat confidential data on the number of employees. In a short time period this new index should be disseminated at national level, consequently higher levels of data quality have to be reached. Aiming at this target, in the last months the methodology currently employed to produce the preliminary estimates of wages and labour costs indexes has been analyzed in depth. Various innovations have been designed and experimented in order to allow the implementation of a high reliable methodological system for the estimation of the number of employees.
Two sources of causes have been isolated as possible responsible for the non-negligible error produced by the actual methodology: the over-coverage of the list for the estimation of figures referring to the current population and the non-randomness of the sample coupled with a growing but unstable sample size along time.
In this paper the main results of several methodological experimentations are presented. The analysis carried out have allowed considerable progress in terms of reduction of the over-coverage problem. As concerning the non-randomness of the sample, some critical areas of the estimation methodology, linked to the grossing up to the population, have been detected as responsible for the error. On this aspect further experimentations have been outlined.

La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento[♦]

[♦] A cura di: Ciro Baldi, Francesca Ceccato, Silvia Pacini, Donatella Tuzi, Istat, Servizio OCC.

Il gruppo di ricerca costituito dagli autori è stato coordinato da Ciro Baldi nell'ambito delle attività del Servizio OCC e del "Progetto interarea finalizzato ad approfondire e a verificare sperimentalmente alcune metodologie e tecniche statistiche utili per ottenere stime preliminari nelle indagini congiunturali condotte dall'Istituto Nazionale di Statistica".

Sebbene il documento sia frutto di comuni ricerche e discussioni, i paragrafi sono da attribuire come segue:

- i paragrafi 1, 3.1 e 8 a Ciro Baldi,
- i paragrafi 2.1, 6.1 e 6.3 a Francesca Ceccato,
- i paragrafi 2.2, 3.2 e 7.5.1 a Silvia Pacini,
- i paragrafi 4 e 7.3 a Ciro Baldi e Francesca Ceccato,
- i paragrafi 5 e 7.4 a Ciro Baldi e Silvia Pacini,
- i paragrafi 6.2.1 e 6.4 a Francesca Ceccato e Silvia Pacini,
- i paragrafi 6.2 e 6.2.2 a Francesca Ceccato, Silvia Pacini e Donatella Tuzi,
- il paragrafo 6.2.3 a Silvia Pacini e Donatella Tuzi,
- i paragrafi 7.1 e 7.2 a Ciro Baldi e Donatella Tuzi,
- il paragrafo 7.5.2 a Donatella Tuzi.

1. Introduzione

Per ottemperare alle richieste del Regolamento Eurostat 1165/98 (STS), la rilevazione OROS ha recentemente affiancato alla produzione degli indicatori relativi a retribuzioni per Ula, oneri sociali per Ula e costo del lavoro per Ula, una stima delle *posizioni lavorative*[1]. I nuovi indicatori di occupazione sono stati trasmessi a Eurostat a partire dalla metà del 2004 e ciò rende sempre più urgente la diffusione dei medesimi indicatori a livello nazionale. Tuttavia, la pubblicazione di nuovi indicatori sulle posizioni lavorative impone il raggiungimento di requisiti di qualità particolarmente stringenti, soprattutto per quel che riguarda le stime provvisorie, il cui errore di revisione deve essere contenuto in dimensioni molto limitate. A tal fine, nel corso degli ultimi mesi la metodologia utilizzata per la produzione di stime preliminari è stata sottoposta a un programma molto ampio di verifiche e approfondimenti che hanno condotto alla definizione e alla sperimentazione di una serie di innovazioni volte a mettere a punto un impianto metodologico caratterizzato da un elevato grado di affidabilità.

La rilevazione, che si basa sui dati amministrativi INPS delle denunce contributive che le imprese effettuano per i propri lavoratori dipendenti (mediante i modelli DM10), attualmente produce per la variabile occupazione, così come per quelle relative a retribuzioni e costo del lavoro, una stima *preliminare* a 90 giorni dal trimestre di riferimento ed una stima *definitiva* a circa 365+90 giorni. La prima è basata su un campione non casuale di moduli DM10, mentre la seconda fa riferimento all'universo dei DM10 (Baldi, Ceccato, Congia, Cimino, Pacini, Rapiti, Tuzi, 2004).

Nella sperimentazione effettuata sui trimestri dal II 2001 al IV 2002, la contestuale disponibilità del campione dei DM10 e del relativo universo ha reso possibile valutare l'errore di revisione. La metodologia attualmente implementata produce un errore non trascurabile riconducibile ad almeno due fattori: la sovracopertura dell'anagrafica utilizzata per la stima della popolazione corrente e l'instabilità nella dimensione del campione, sia in termini di unità che di occupazione.

Le sperimentazioni condotte negli ultimi mesi hanno consentito di compiere notevoli progressi nella risoluzione del problema della sovracopertura, mentre in riferimento alla natura non casuale del campione, una serie di approfondimenti hanno condotto alla individuazione di alcune aree critiche della metodologia di stima, da cui può avere origine l'errore attribuibile al riporto all'universo. Su questo aspetto, occorrerà effettuare ulteriori sperimentazioni.

Date alcune particolarità della rilevazione OROS, le analisi svolte potranno risultare di notevole interesse nel contesto più generale di rilevazioni congiunturali che vogliono fornire stime anticipate ricorrendo a sottoinsiemi non casuali di rispondenti rapidi. La stima preliminare di OROS, infatti, può essere vista come una stima anticipata basata su un sottoinsieme non casuale del campione complessivo (che per OROS è rappresentato dall'universo).

Il documento è strutturato come segue. Nel paragrafo 2 sono presentate le linee essenziali della rilevazione. Nel paragrafo 3 vengono definite le misure a cui si ricorre per quantificare l'errore di stima di cui si illustrano, a seguire, le principali caratteristiche strutturali. Nel paragrafo 4 viene illustrata la metodologia di stima delle PMI, mentre nel paragrafo 5 viene fornita una misura dell'errore di revisione, che viene scomposto nella componente dovuta all'errore di lista delle unità attive e all'errore di riporto all'universo. Queste due fonti di errore sono analizzate nel dettaglio nei paragrafi 6 e 7, in cui vengono inoltre proposti alcuni cambiamenti di metodo. Nel paragrafo 8 si traggono delle conclusioni e si descrivono alcune prospettive di avanzamento sulla stima delle posizioni lavorative, connesse con una situazione informativa che si sta gradualmente modificando e con le richieste Eurostat di riduzione dei tempi di rilascio delle informazioni.

[1] Evidentemente, questa variabile concorreva anche in precedenza alla determinazione dei denominatori delle variabili retributive unitarie rilasciate trimestralmente dalla rilevazione; tuttavia, solo a partire dalla metà del 2004 è entrata a regime la procedura di calcolo dei numeri indice dell'occupazione e la relativa trasmissione a Eurostat in forma confidenziale.

2. Le principali caratteristiche della rilevazione OROS

2.1 Variabili, popolazione e fonti

Attualmente OROS rilascia stime su retribuzioni lorde per ULA, oneri sociali per ULA, costo del lavoro per ULA pubblicate regolarmente a 90 giorni dalla fine del trimestre di riferimento (e con un programma di aumento della tempestività già definito nel calendario dei comunicati stampa relativi al 2005). Nei prossimi mesi dovranno essere diffusi in Italia gli indicatori sulle posizioni lavorative, di cui si è già iniziato l'invio ad Eurostat, in ottemperanza del Regolamento Eurostat 1165/98 (STS). Le variabili d'interesse sono espresse in termini di variazioni (numero indice, variazione congiunturale e variazione tendenziale) e, tra di esse, è la variazione tendenziale che riveste importanza principale. In futuro (non prossimo), potrebbero essere rilasciati anche i livelli.
La popolazione obiettivo è costituita dalle imprese attive con almeno un dipendente, nei settori di attività economica dell'industria e dei servizi privati (esclusa la pubblica amministrazione e i servizi alle famiglie). La rilevazione trae informazioni dagli archivi INPS delle denunce contributive mensili DM10[2], in cui l'unità di rilevazione amministrativa è la *posizione contributiva*, che può corrispondere ad un'impresa o a parte di essa. Sulla popolazione delle imprese con oltre 500 dipendenti, la fonte INPS è integrata con la rilevazione Istat su lavoro e retribuzioni nelle Grandi Imprese (d'ora in avanti, rilevazione GI)[3], mentre si ricorre all'archivio statistico ASIA per l'acquisizione di informazioni strutturali sulle imprese attive[4].
Gli archivi INPS utilizzati sono tre: l'anagrafe, l'universo (o popolazione) e il campione. L'anagrafe, resa disponibile alla fine di ciascun trimestre, contiene informazioni strutturali sulle posizioni contributive[5]. L'universo mensile dei DM10, acquisito con circa 14 mesi di ritardo, contiene la quasi totalità dei modelli di competenza del mese stesso. L'esigenza di produrre indicatori trimestrali implica l'accorpamento dei set informativi riferiti ai 3 mesi del trimestre di riferimento. Infine, l'INPS mette a disposizione dell'Istat un campione di DM10, con un ritardo di circa 45 giorni dalla fine del trimestre di riferimento. Quest'ultimo set informativo ha subito nel tempo un'evoluzione su cui occorre soffermarsi brevemente. Il campione è costituito dai DM10 arrivati per via telematica all'INPS e quindi disponibili in tempi ridotti, rispetto ai moduli inviati per via cartacea che confluiscono, successivamente, nell'universo delle dichiarazioni. Nel tempo si è osservato un graduale passaggio alle forme elettroniche di trasmissione, con l'effetto di accrescere la dimensione di tale sottoinsieme di dati (grafico 1). Questo processo ha subito un'improvvisa accelerazione a partire dal secondo trimestre del 2004, in seguito ad una disposizione INPS, secondo cui banche e poste, presso cui si effettua il pagamento dei contributi, non sono più tenute ad accettare dichiarazioni cartacee[6]. Nel giro di qualche trimestre ci si attende, dunque, di poter disporre, a 45 giorni dalla fine di ciascun trimestre, dell'intera popolazione delle dichiarazione contributive. D'altra parte, il dato relativo al terzo trimestre del 2004 conferma come il processo di convergenza verso l'universo del campione di dichiarazioni inviate per via telematica sia giunto

[2] Obbligatorie per le imprese con almeno un dipendente.
[3] I dati della rilevazione GI sono usati sulla sottopopolazione delle imprese con oltre 500 dipendenti, a sostituzione ed integrazione dei dati INPS. Il ricorso a questa fonte esterna si rende necessario in quanto questa tipologia d'impresa è sistematicamente sottorappresentata nei dati campionari INPS.
[4] Ad ASIA si ricorre per attribuire il codice ATECO: in particolare, si usa l'ATECO ASIA per tutte le posizioni contributive che si abbinano per codice fiscale con l'Archivio, laddove sulle rimanenti si utilizza il codice ATECO dell'INPS e in sua mancanza il CSC (entrambe opportunamente trascodificati).
[5] Tra le informazioni strutturali più rilevanti vi è la matricola aziendale (codice numerico di 10 digit) che identifica la posizione contributiva, il codice fiscale e la ragione sociale dell'impresa di riferimento, la data di nascita della posizione, l'eventuale data di cessazione, sospensione o riattivazione, il codice Ateco'91 attribuito dall'INPS e il codice statistico contributivo (CSC - codice a 5 digit che identifica, ai fini contributivi, il settore di attività economica in cui opera l'impresa).
[6] L'INPS ha così anticipato i tempi di legge, che prevedevano l'invio di tutti i DM per via telematica dal mese di competenza di gennaio 2005.

quasi a saturazione. La disponibilità di un campione che può essere considerato un quasi - universo modifica sostanzialmente, e in positivo, il quadro informativo della rilevazione.

2.2 Stima anticipata e stima finale

Nella procedura di costruzione delle stime OROS, le posizioni contributive vengono distinte in quattro sottopopolazioni:

1. le piccole e medie imprese (d'ora in avanti PMI);
2. le grandi imprese che non rientrano nel dominio della rilevazione GI (d'ora in avanti GI-INPS);
3. le grandi imprese che rientrano nel dominio della rilevazione GI (d'ora in avanti GI-RIL);
4. le imprese che forniscono lavoro interinale (d'ora in avanti INTER).

Per i problemi connessi alla rappresentatività delle imprese di grandi dimensioni, la sottopopolazione GI-RIL viene sostituita con le imprese della rilevazione GI.
Le stime prodotte dalla rilevazione sono di due tipologie: ad una stima anticipata (o preliminare) rilasciata a circa 90 giorni dal trimestre di riferimento, segue la stima finale (o definitiva) a circa 365+90 giorni dal trimestre di riferimento. Mentre la stima preliminare viene prodotta sulla base del set dei *rispondenti rapidi*, ossia sul campione dei DM10 giunti per via telematica, la stima definitiva fa uso dell'universo delle dichiarazioni contributive.
Ai fini della stima definitiva, l'universo è sottoposto ad una procedura di imputazione delle mancate risposte totali, il cui scopo è quello di individuare e correggere i dati economici delle unità assenti, per motivi non giustificati da eventi demografici (ovvero unità non stagionali, non cessate né sospese). La metodologia di imputazione a cui si ricorre usa come informazione di supporto gli universi dei 4 trimestri precedenti a quello di riferimento e l'universo del trimestre seguente. Al di là di questa operazione, che è l'unica rilevante a cui è sottoposto l'universo, la stima finale del livello dell'occupazione è calcolata sommando la variabile su tutte le posizioni contributive delle sottopopolazioni PMI, GI-INPS e INTER, a cui viene aggiunto il totale della fonte rilevazione GI.
Per la produzione della stima anticipata si ricorre a tutti e tre gli archivi INPS, l'anagrafe del trimestre *t,* il campione del trimestre *t* e l'universo del trimestre ausiliario *t*-4, oltre che alla fonte rilevazione GI. Sinteticamente, la stima dell'occupazione sulle sottopopolazioni di appartenenza può essere descritta come segue:

- per le PMI si effettua una somma ponderata sulle unità del campione con i pesi derivanti da una procedura di calibrazione;
- sulle GI-INPS si trascina il dato dall'ultimo trimestre di stima finale ($x^a_t = x^f_{t-5}$);
- per le GI-RIL si sommano i dati rilevati e/o imputati dalla rilevazione GI;
- per le INTER si procede ad una somma sulle unità appartenenti al campione, dopo aver imputato eventuali mancate risposte totali con una procedura ad hoc.

Le diverse sottopopolazioni di stima hanno un peso molto differente. In particolare, il 75% degli occupati è concentrato nelle PMI, mentre le GI pesano per circa il 22%. Inoltre, esse sono caratterizzate da un errore di stima (stima anticipata rispetto a stima finale) molto diverso: l'errore più rilevante è quello delle PMI, laddove l'errore sulle GI è pari a zero in quanto la stima anticipata è già, per definizione, la stima finale.

3. L'errore di stima

3.1. Misurazione dell'errore di stima

L'errore di stima oggetto dell'analisi che segue è sostanzialmente un errore di revisione, in quanto misura la distanza tra stima anticipata e stima definitiva. L'attenzione verrà focalizzata sia sugli

errori nei livelli che nelle variazioni (congiunturali e tendenziali). In particolare, l'errore nei livelli di uno specifico trimestre *t* è definito come:

$$e_t = \frac{\hat{y}_t^a - \hat{y}_t^f}{\hat{y}_t^f} * 100 \qquad [1]$$

dove:
$\hat{y}_t^a$ è il livello della stima anticipata (denotata dall'apice *a*) della variabile di interesse (d'ora in avanti l'occupazione) al tempo *t*, mentre $\hat{y}_t^f$ è il livello della stima finale (denotata dall'apice *f*) al tempo *t*.
L'errore della variazione tendenziale è, invece, espresso come:

$${}_{vt}e_t = {}_{vt}\hat{y}_t^a - {}_{vt}\hat{y}_t^f \qquad [2]$$

dove:

$${}_{vt}\hat{y}_t^a = \frac{\hat{y}_t^a - \hat{y}_{t-4}^a}{\hat{y}_{t-4}^a} * 100 \qquad [3]$$

è la variazione tendenziale calcolata sulle stime anticipate e

$${}_{vt}\hat{y}_t^f = \frac{\hat{y}_t^f - \hat{y}_{t-4}^f}{\hat{y}_{t-4}^f} * 100 \qquad [4]$$

è la variazione tendenziale calcolata sulle stime finali.
In maniera analoga l'errore nelle variazioni congiunturali può essere espresso come:

$${}_{vc}e_t = {}_{vc}\hat{y}_t^a - {}_{vc}\hat{y}_t^f \qquad [5]$$

in cui ${}_{vc}\hat{y}_t^a$ e ${}_{vc}\hat{y}_t^f$ hanno formulazione simile rispettivamente alla [3] e alla [4].
Naturalmente l'errore nelle variazioni e quello nei livelli sono strettamente connessi: si può mostrare che l'errore nella variazione tendenziale dipende dalla differenza dell'errore nei livelli a *t* e a *t*-4 e dall'entità della variazione tendenziale finale[7]. Ne consegue che, se gli errori nei livelli a *t* e *t*-4 fossero uguali, l'errore nella variazione risulterebbe nullo.
Come misure di sintesi dell'errore, si fa riferimento alla media semplice degli errori percentuali, ossia il *Mean Percentage Error* (MPE) e alla media degli errori percentuali in valore assoluto, vale a dire il *Mean Absolute Percentage Error* (MAPE).
Sulle variazioni tendenziali MPE sarà calcolato come:

$$MPE(vt) = \frac{\sum_{t=1}^{T} {}_{vt}e_t}{T} \qquad [6]$$

mentre MAPE sulle variazioni tendenziali è espresso dalla relazione:

[7] Si ringrazia Leonello Tronti per avere derivato algebricamente questa relazione.

$$MAPE(vt) = \frac{\sum_{t=1}^{T} \left| {}_{vt} e_t \right|}{T} \qquad [7]$$

Le due misure sono complementari in quanto, mentre MPE misura la media *tout court*, MAPE dà una misura al lordo delle compensazioni tra valori positivi e negativi. Considerati insieme i due indicatori mostrano la presenza di errori sistematici: MPE e MAPE uguali stanno ad indicare che si commettono errori sistematicamente positivi; uguali ma di segno opposto evidenziano errori sistematicamente negativi.
In ciò che segue l'analisi è svolta, ove non altrimenti segnalato, su sette trimestri: dal II trimestre 2001 al IV trimestre 2002. Ne consegue che MAPE ed MPE sui livelli sono medie di sette trimestri, sulle variazioni congiunturali sono medie di sei trimestri e sulle variazioni tendenziali sono medie di tre trimestri.
Nell'analisi che segue si tenderà ad analizzare sia gli errori sui livelli che quelli sulle variazioni, sebbene siano le variazioni il parametro di maggior interesse. Paradossalmente, ciò potrebbe condurre a ritenere soddisfacente la metodologia di stima anche se generasse errori sui livelli sostanziali purché stabili nel tempo. Tuttavia, la considerazione degli errori sui livelli è strettamente connessa alla filosofia che presiede al metodo di stima, che è marcatamente *cross-section* e che, pertanto, mira a produrre una buona stima dei livelli. In secondo luogo, OROS potrebbe ampliare in un prossimo futuro il ventaglio di indicatori rilasciati, fino ad includere i livelli.
Può essere utile discutere, preliminarmente, la natura della misurazione dell'errore nell'ambito della rilevazione OROS. Questa è molto diversa dagli errori campionari e non campionari impliciti o espliciti nelle rilevazioni campionarie. Se nelle rilevazioni campionarie l'errore viene quantificato in funzione delle proprietà degli stimatori usati (sotto il disegno campionario e/o i modelli di stima e di meccanismo di risposta), in OROS il campione teorico corrisponde all'universo, consentendo una quantificazione empirica dell'errore, che risulterà dal confronto tra una stima campionaria e una stima sostanzialmente censuaria.
L'errore è per natura simile a quello cui va incontro ogni rilevazione congiunturale che presenta una stima preliminare, basata su un set informativo ridotto relativo ad un campione di rispondenti rapidi, ed una stima rivista, che sfrutta tutta l'informazione disponibile, in quanto dipendente sostanzialmente dalla quantità e dalla qualità dell'informazione usata nelle due stime. Devono però essere sottolineate almeno due differenze. La prima deriva dalla diversità dell'informazione disponibile per la stima definitiva che per un'indagine classica si basa su un campione di rispondenti totali (concettualmente molto più vicino al campione teorico), mentre per OROS è rappresentata dall'universo dei dati. La seconda, non meno importante, è da ricercare nelle caratteristiche della lista di stima (l'universo o la popolazione corrente a cui si riportano i dati): mentre in un'indagine congiunturale classica questa è unica per le due stime ma, proprio per questo, spesso riferita ad una popolazione assai precedente al periodo corrente, in OROS la lista della popolazione corrente (la lista di stima) definitiva può essere anche molto diversa da quella "stimata" cioè quella utilizzata per la stima preliminare. In altri termini, in OROS va aggiunto l'errore di stima della lista di unità attive all'errore cui va incontro anche un indagine congiunturale classica, ovvero di stimare i valori di interesse sulla base di un sottoinsieme di rispondenti. Come si vedrà in seguito, queste due fonti di errore (l'errore di riporto all'universo e l'errore di lista) hanno una dimensione simile.

3.2. L'errore totale di stima: struttura per ATECO e per sottopopolazioni

Nel periodo compreso tra il II trimestre 2001 e il IV trimestre 2002, l'errore totale di stima nei livelli è pari al 2,1% ed sistematicamente positivo. L'errore sui livelli, inoltre, è molto stabile nel tempo, ne consegue che gli errori sui tassi di variazione sono decisamente più contenuti: il MAPE è

pari allo 0,4% sulle variazioni tendenziali e su quelle congiunturali e l'MPE, invece, è quasi nullo in entrambi i casi (tabella 1).
L'analisi per sezione mostra una certa varietà di casi relativamente alla dimensione e nella stabilità dell'errore. La dimensione nei livelli presenta una notevole variabilità, con un massimo del 6,6% nel settore delle costruzioni ed un minimo di 0 nei settori della produzione di energia elettrica, gas ed acqua e delle altre attività professionali ed imprenditoriali. Anche la stabilità dell'errore non è una caratteristica così diffusa tra i settori: MPE negativi sulle variazioni tendenziali in alcuni settori dell'industria (produzione di energia elettrica, gas ed acqua e costruzioni) e dei servizi (commercio e riparazione di beni di consumo, alberghi e ristoranti e intermediazione monetaria e finanziaria) indicano che l'errore nei livelli degli ultimi trimestri è notevolmente minore di quello dei primi trimestri; l'alto MPE positivo sulle variazioni tendenziali nel settore della altre attività professionali ed imprenditoriali, al contrario, segnala una tendenza alla crescita dell'errore sui livelli.

Tabella 1 – Errore totale di stima per settore di attività economica in termini di livelli, variazioni congiunturali e tendenziali. Periodo II trimestre 2001 – IV trimestre 2002 *(valori e punti percentuali medi di periodo)*

Sezione di attività economica	**Livelli**		**Variazioni congiunturali**		**Variazioni Tendenziali**	
	MPE	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**
C Estrazione di minerali	2,7	2,7	0,2	1,1	0,8	0,8
D Attività manifatturiere	1,7	1,7	0,0	0,4	0,2	0,4
E Produzione di energia elettrica, gas ed acqua	0,0	0,6	0,1	0,9	-0,8	0,8
F Costruzioni	6,6	6,6	-0,4	0,7	-1,0	1,0
G Commercio e riparazione di beni di consumo	2,9	2,9	-0,5	0,8	-2,1	2,1
H Alberghi e ristoranti	1,4	1,4	0,0	0,8	-0,1	1,2
I Trasporti, magazzinaggio e comunicazioni	1,3	1,3	0,2	0,2	0,6	0,6
J Intermediazione monetaria e finanziaria	2,3	2,3	-0,4	1,2	-2,3	2,3
K Altre attività professionali ed imprenditoriali	0,0	1,6	0,4	1,2	3,8	3,8
C-K TOTALE	**2,1**	**2,1**	**-0,1**	**0,4**	**0,2**	**0,4**

Fonte: OROS – INPS

L'analisi per sottopopolazioni di stima consente di individuare le aree di maggiore criticità (tabella 2). Per definizione, l'errore è nullo per le grandi imprese della rilevazione GI. Dato che tale sottopopolazione conta per il 20,3% dell'occupazione totale di OROS, l'assenza di errore su di essa contribuisce sensibilmente al contenimento dell'errore complessivo di stima.
La sottopopolazione con il maggiore errore in valore assoluto è quella delle imprese interinali (INTER). Tuttavia le medie nascondono la sostanziale diminuzione in valore assoluto dell'errore, negativo in tutti i trimestri, avvenuta tra il 2001 e il 2002 in seguito ad un cambiamento nel metodo di stima.
Errori relativamente alti si riscontrano anche per la sottopopolazione delle imprese grandi di INPS (GI-INPS). Tali errori possono essere ridotti notevolmente adottando per queste imprese un metodo di stima più sofisticato analogamente a quanto avviene per le imprese interinali.
La sottopopolazione più problematica, comunque, rimane quella delle PMI sia per l'entità del suo errore che per la quota di occupazione che essa riveste sul totale, in media pari al 76%. E' su questa sottopopolazione, quindi, che è stata concentrata l'attenzione ed a cui è dedicato il resto dell'analisi.

Tabella 2 – Errore totale di stima per sottopopolazione di stima in termini di livelli, variazioni congiunturali e tendenziali. Periodo II trimestre 2001 – IV trimestre 2002 *(valori e punti percentuali medi di periodo)*

Sottopopolazione di stima	Occupazione sul totale	Livelli		Variazioni Congiunturali		Variazioni Tendenziali	
		MPE	MAPE	MPE	MAPE	MPE	MAPE
PMI	76,0	3,1	3,1	-0,2	0,4	-0,6	0,6
GI-INPS	2,2	-2,7	3,8	1,1	2,7	7,7	7,7
INTER	1,5	-17,7	17,7	7,8	8,9	50,4	50,4
GI	20,3	0,0	0,0	0,0	0,0	0,0	0,0
TOTALE	**100,0**	**2,1**	**2,1**	**-0,1**	**0,4**	**0,2**	**0,4**

Fonte: OROS – INPS

4. La metodologia di stima anticipata delle PMI

La stima delle PMI fa prevalentemente uso dei dati di fonte INPS[8]. Il primo archivio di dati a cui si ricorre per la stima è l'anagrafe che, disponendo della data di costituzione e di cessazione/sospensione/riattivazione, consente di stabilire lo stato di attività delle unità e quindi di individuare una lista delle unità attive nel trimestre[9] (lista di stima). Alla definizione dello stato di attività contribuiscono anche le informazioni deducibili dal campione, altro archivio di dati INPS, poiché l'appartenenza al campione di *t* conferma lo stato di attività delle unità. In questo senso il campione totale è un sottoinsieme delle unità attive secondo l'anagrafe.

Tuttavia, non tutto il campione totale viene usato per la stima, ma solo un suo sottoinsieme. In particolare, concorrono alla stima le posizioni che hanno inviato il DM10 in tutti e tre i mesi del trimestre, le posizioni che secondo l'anagrafe sono nate, sospese, riattivate o cessate nel trimestre di riferimento, e le posizioni che sono reputate stagionali sulla base dell'informazione sui pattern di presenza nel trimestre attuale e in quello ausiliario. D'ora in avanti si farà riferimento a questo sottoinsieme col nome di *campione ridotto* (o semplicemente *campione*), mentre per fare riferimento all'intero archivio proveniente dall'INPS verrà utilizzato il termine *campione totale*.

Lo scopo di questo taglio risiede nella qualità dei dati economici trimestrali. Per fissare le idee si pensi al dato trimestrale sui dipendenti. Esso è pari alla media mensile dei dipendenti (ovvero alla somma dei dipendenti dichiarati nei DM10 mensili diviso 3). Nel caso in cui per una data posizione siano arrivati i DM10 di tutti e tre i mesi, il dato trimestrale sarà pari al monte dei dipendenti diviso 3 (se sono stati dichiarati 10 dipendenti al mese, la media mensile sarà pari a 10 =30/3). Nel caso in cui, invece, sia arrivato un solo DM10, la media mensile sarà pari a 3,33 (10/3). Questo è un dato corretto solo se la posizione è stata attiva un solo mese al trimestre (ad esempio perché è nata l'ultimo mese del trimestre, oppure sospesa per due mesi, etc). Se, invece, la posizione è in realtà attiva tutti e tre i mesi del trimestre, ma appartiene al campione telematico per un solo mese (per ragioni amministrative o perché ha cambiato modalità di invio) la media mensile calcolata dividendo per 3 sarà un dato affetto da errore: il dato corretto dovrebbe essere calcolato dividendo per i mesi di effettiva attività. Il processo di imputazione del dato errato, implicito in tale divisione, è però ad alta probabilità di errore, in primo luogo perché è difficile discriminare tra le posizioni inattive nella popolazione e quelle semplicemente mancanti nel campione.

Questa serie di considerazioni ha in conclusione indotto, in via sperimentale, ad escludere per la stima quelle posizioni del campione totale per cui non si è potuto attribuire con ragionevole certezza uno stato di attività, producendo le stime anticipate sul sottoinsieme *campione*.

[8] Per l'attribuzione dell'ATECO si ricorre alla fonte ASIA.

[9] Si definisce attiva nel trimestre un'unità attiva almeno un giorno in quel trimestre.

L'ultima fonte di informazione per la produzione delle stime è l'universo di *t*-4, che fornisce le variabili ausiliarie.
La metodologia di stima è basata su uno stimatore di ponderazione vincolata o di calibrazione (Baldi, Falorsi, Pallata, Succi, Russo, 2000). La stima degli occupati al trimestre *t* è calcolata come:

$$\hat{Y}_t = \sum_{i \in C_t} k_{it} y_{it} \qquad [8]$$

dove y_{it} è il numero dei dipendenti della posizione *i*-esima al tempo *t*, C_t è il campione del tempo *t* e k_{it} è il peso della posizione *i*-esima al tempo *t*, calcolato come soluzione del seguente problema di minimo vincolato:

$$\begin{cases} \underset{\{k_{it}\}}{Min}\left[\sum_{i \in C_t} c_{it}\left(k_{it}-1\right)^2\right] \\ \sum_{i \in C_t} k_{it} x_{it} = X_t \end{cases} \qquad [9]$$

Nella funzione obiettivo, c_{it} è una costante che misura la dimensione della posizione e 1 il peso diretto attribuito alle unità del campione, per l'assenza di un disegno campionario. Il vincolo è un sistema di equazioni dove $\mathbf{x}_{it}$ è il vettore colonna delle variabili ausiliarie $\mathbf{x}$ e $\mathbf{X}_t$ è il vettore dei totali di calibrazione relativi alle variabili ausiliarie. Questi totali di calibrazione sono ottenuti come somma, sulla lista di stima, delle variabili ausiliarie.
La disponibilità di variabili ausiliarie è diversa a seconda di sottoinsiemi di unità della lista. In particolare, si distinguono tre gruppi di unità: le posizioni con età superiore o uguale ad un anno con informazioni economiche nell'universo di *t*-4 (panel con informazione ausiliaria), le posizioni con età superiore o uguale ad un anno senza informazioni economiche nell'universo di *t*-4 (panel senza informazione ausiliaria), le posizioni con età inferiore ad un anno (nuove nate). Per le panel con informazione ausiliaria, la disponibilità di informazioni ausiliarie è massima: per esse si usano il numero di posizioni secondo la lista, il numero di dipendenti a *t*-4, il monte retributivo a *t*-4, il monte oneri a *t*-4. Per le panel senza informazione ausiliaria, l'unica variabile disponibile è il numero di posizioni secondo la lista. Per le nuove nate le variabili ausiliarie sono: il numero di posizioni secondo la lista e il numero di dipendenti all'iscrizione. Questa ultima informazione è contenuta nell'anagrafe.
Il problema di minimo vincolato è risolto nell'ambito di gruppi omogenei di unità (*model groups*). Le variabili di stratificazione dei *model groups* sono le divisioni di attività economica, le ripartizioni geografiche e la classe dimensionale (quest'ultima solo sulle panel con variabili ausiliarie). Attualmente i *model groups* sono 546.
Come detto, i totali di calibrazione sono calcolati come somma sulla lista di stima delle variabili ausiliarie. Tuttavia, una somma semplice porterebbe ad una sovrastima dei totali di calibrazione dovuta al problema della sovracopertura della lista, per effetto del fenomeno delle cessazioni non registrate, di cui si parlerà diffusamente nel paragrafo 5 che segue.
Per questo motivo i totali di calibrazione vengono corretti per tenere conto che non tutte le posizioni della lista sono effettivamente attive. Il metodo di correzione prevede l'applicazione di una probabilità di essere attiva ad ogni unità della lista di stima del trimestre *t*. In breve, considerato che l'universo è compreso per definizione nell'anagrafe ($P_t \subseteq A_t$), si vuole stimare la probabilità che ciascuna unità *i*-esima presente nella lista sia effettivamente attiva al tempo *t*:

$$\hat{p}_{it} = pr(i \in P_t \,/\, i \in A_t) \qquad [10]$$

Tale probabilità è calcolata per gruppi omogenei di unità (*register error group*), ottenuti stratificando le posizioni contributive rispetto alle variabili età, ripartizione geografica, classe dimensionale, classificazione di attività economica e presenza delle variabili ausiliarie in *t*-4, e non coincide con quella dei *model groups* definiti poco sopra. Nella metodologia di base, il numero dei *register error groups* è pari a circa 380.
Nel metodo di stima attuale, la probabilità di essere attive è calcolata su tutte le unità del trimestre *t*-4 come rapporto tra il numero di unità appartenenti all'universo e il numero di unità definite attive dall'anagrafe. Assumendo l'ipotesi di invarianza temporale tra *t*-4 e *t*, la probabilità viene applicata, nel trimestre corrente *t*, solo alle unità che non appartengono al campione C_t. Alle unità campionarie, invece, viene attribuita ex-post la probabilità pari ad uno, in quanto certamente attive in *t* (metodo M0 per il calcolo delle probabilità di essere attiva). In formule:

$$\hat{p}_{it} = \frac{\#P_{t-4}}{\#A_{t-4}} \qquad i \notin C_t$$
$$\hat{p}_{it} = 1 \qquad i \in C_t \qquad [11]$$

La riduzione della lista anagrafica incide sulla stima attraverso il calcolo dei totali delle variabili ausiliarie (da ora denominati totali di calibrazione), come segue:

$$\sum_{i \in A_t} x_{it}\, \hat{p}_{it} = X_t \qquad [12]$$

5. Scomposizione dell'errore di stima delle PMI: errore di lista ed errore di riporto

Le possibili fonti di errore nella stima anticipata OROS delle PMI possono essere ricondotte a due cause fondamentali: la scelta del modello di correzione dell'anagrafe e il metodo applicato per il riporto all'universo.
Relativamente al primo aspetto, il ricorso ad un modello di correzione dell'anagrafe si rende indispensabile per porre rimedio ad alcuni aspetti critici del registro anagrafico. Tale registro, utilizzato per individuare la lista di stima, da una parte soffre di errori di sovracopertura dovuti al ritardo di registrazione degli eventi di sospensione e/o cessazione dell'attività (cessazioni non registrate),[10] dall'altra è affetto da errori di sottocopertura causati dai ritardi nella comunicazione e/o registrazione delle riattivazioni di posizioni sospese. Nel primo caso, ciò comporta l'inclusione nella lista di stima di alcune posizioni che risultano erroneamente attive e una conseguente sovrastima dei totali di calibrazione. Nel secondo caso, al contrario, vengono considerate attive un minor numero di posizioni rispetto a quelle realmente esistenti, con conseguenze di segno opposto sui totali. Si è calcolato, però, che l'errore di sovracopertura risulta predominante rispetto a quello di sottocopertura; tale constatazione ha indotto ad effettuare una serie di sperimentazioni per la sua riduzione.
L'altra fonte di errore nella stima OROS è da attribuirsi alle caratteristiche del campione su cui si basa la stima anticipata, che non è controllabile nei meccanismi di selezione. Il campione reso disponibile dall'INPS, infatti, è il risultato di un'autoselezione di unità (i moduli DM10 inviati per via telematica) e, pertanto, è non casuale. I dati a disposizione sono affetti da un *bias* strutturale sistematico legato ad esempio alla rappresentatività di alcune tipologie di imprese (quelle più attrezzate tecnologicamente o che ricorrono a consulenti fiscali), che rende necessario un trattamento ex-ante dell'informazione disponibile. Sperimentazioni su tale aspetto vengono presentate nel paragrafo 7.

[10] Un'impresa, infatti, ha scarsi incentivi a dichiarare la cessazione dell'attività con dipendenti all'INPS, in quanto per tale inadempienza non sono previste sanzioni.

Al fine di quantificare l'errore complessivo di stima dell'occupazione, la stima anticipata del trimestre corrente, prodotta utilizzando le informazioni fornite dall'anagrafe e dal campione, viene messa a confronto con quella finale, basata sulle informazioni dell'universo, relativa allo stesso trimestre ma disponibile con cinque trimestri di ritardo.
Nel periodo compreso tra il II trimestre del 2001 e il IV del 2002, l'occupazione delle PMI è mediamente sovrastimata del 3,1% (tabella 3) con un errore presente in tutti i settori di attività economica, sebbene con valori significativamente differenti. Nelle costruzioni si rileva il valore più elevato (6,6%), mentre i valori minimi sono registrati nei settori delle attività manifatturiere e degli alberghi e ristoranti (rispettivamente 2,1% e 2%). La limitata rilevanza dello scostamento tra MPE e MAPE evidenzia che non ci sono significative differenze di segno negli errori dei diversi trimestri: l'occupazione stimata, dunque, presenta un sostanziale errore di sovrastima che, da un'analisi più dettagliata, risulta anche essere pressoché costante nel tempo.
Le basi dati a disposizione hanno consentito di realizzare una simulazione: disponendo di anagrafe, campione e universo per i sette trimestri già citati, è stato possibile neutralizzare la parte dell'errore totale attribuibile al modello di correzione dell'anagrafe, incidendo sul *bias* nelle stime dei totali di calibrazione. Per fare ciò, nel calcolo di questi totali la lista di stima normalmente usata, costituita dall'insieme delle posizioni contributive formalmente dichiarate attive dall'anagrafe e ridotte con la probabilità di essere attiva (come descritto nella [12]), è stata sostituita con la lista delle posizioni realmente attive secondo l'universo, eliminando quindi il *bias* dovuto alla sovracopertura anagrafica.

Tabella 3 - Errore, totale e al netto degli errori di lista anagrafica, nella stima dell'occupazione delle PMI. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*

Sezione di attività Economica	Errore totale		Errore al netto degli errori di lista anagrafica	
	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	3,1	3,1	2,9	2,9
D Attività manifatturiere	2,1	2,1	1,2	1,2
E Produzione di energia elettrica, gas ed acqua	2,8	4,2	2,5	3,6
F Costruzioni	6,6	6,6	3,5	3,5
G Commercio e riparazione di beni di consumo	3,6	3,6	2,3	2,3
H Alberghi e ristoranti	2,0	2,0	0,3	0,8
I Trasporti, magazzinaggio e comunicazioni	3,3	3,3	0,7	0,7
J Intermediazione monetaria e finanziaria	5,2	5,2	4,0	4,0
K Altre attività professionali ed imprenditoriali	3,2	3,2	1,5	1,5
C-K TOTALE	**3,1**	**3,1**	**1,7**	**1,7**

Fonte: Oros-INPS

Se fosse possibile eliminare questo errore, resterebbe una sovrastima dell'occupazione totale dell'1,7% tutta imputabile al metodo di riporto all'universo (tabella 3). Essendo questa una situazione teorica, nei paragrafi che seguono si espongono i risultati dei miglioramenti apportati alla metodologia attualmente usata, volti a ridurre le cause di errore legate alla lista anagrafica (1,4%).

6. L'errore di lista: la sovracopertura dell'anagrafe

Una delle peculiarità del registro anagrafico è la sua crescente dimensione numerica nel tempo legata al fatto che le posizioni contributive, una volta registrate, non vengono eliminate ma aggiornate sul loro stato di attività. In particolare, l'aggiornamento si caratterizza per una buona tempestività relativamente all'iscrizione di nuove posizioni, mentre risente di un notevole ritardo nella registrazione degli eventi relativi a cessazioni, sospensioni e riattivazioni. Questo fenomeno, già definito delle cessazioni non registrate, comporta una sovracopertura della lista di stima in

quanto presenta posizioni che, pur risultando formalmente attive, non hanno neanche un DM10 nel trimestre considerato. L'entità del fenomeno viene quantificata mettendo a confronto il numero di posizioni contributive presenti nella lista di stima (A_t) con il numero di posizioni realmente attive dell'universo (P_t). Mediamente nel periodo considerato oltre il 25% delle posizioni contributive attive secondo le informazioni anagrafiche non hanno presentato la dichiarazione contributiva (tabella 5). Tale sovracopertura, se non venisse corretta, si rifletterebbe in una sovrastima dei totali di calibrazione influenzando i pesi di stima. Un'analisi approfondita delle caratteristiche del registro anagrafico ha condotto all'adozione di una tecnica di *data cleaning* e alla determinazione e contestuale applicazione di una probabilità per definire lo stato di attività delle unità. Gli effetti dei miglioramenti apportati nella stima della lista delle unità attive vengono valutati, in prima istanza, attraverso la quantificazione dell'errore che si commette nella stima dei totali di calibrazione e, successivamente, attraverso l'analisi dell'impatto che essi hanno nella stima degli occupati.

6.1. L'utilizzo della variabile ritardo nell'invio del DM10 e l'impatto del *data cleaning*

Le caratteristiche strutturali del registro anagrafico, la cui numerosità è crescente nel tempo, insieme alla sostanziale sovracopertura, rendono necessaria un'analisi più approfondita delle posizioni contributive e la messa a punto di una procedura generalizzata di *data cleaning* dell'anagrafe finalizzata all'individuazione e al trattamento delle unità "probabilmente cessate" ma definite formalmente attive.
Un modello logistico lineare è stato utilizzato per sperimentare l'introduzione del numero di mesi di ritardo nell'invio del DM10 nella definizione dell'effettivo stato di attività delle unità. Tale informazione, usata in serie storica, rappresenta il pattern di presenza/assenza del DM10 definito sulla base della presenza del modello contributivo nella popolazione; informazione che è disponibile per un periodo piuttosto lungo di anni (1996-2002) e che ha reso possibile effettuare un'analisi dei ritardi nell'invio del DM10 da parte delle imprese.
Alla luce delle evidenze del modello logistico, è stata effettuata una simulazione per valutare l'opportunità di un metodo di *data cleaning* basato sui ritardi nell'invio dei DM10 relativamente al IV trimestre 2002. Alla lista anagrafica del trimestre in esame è stato associato il pattern di presenza/assenza del DM10 disponibile, in una situazione ordinaria, fino allo stesso trimestre dell'anno precedente e, quindi, nel caso specifico fino al IV trimestre 2001, per valutare la struttura delle posizioni contributive in termini di classi annuali di ritardo. Nella tabella 4 si riporta la distribuzione delle posizioni caratterizzate da almeno un anno di ritardo nell'invio del DM10 che riguarda oltre il 18% delle unità. In particolare, sono oltre 81 mila le posizioni contributive che hanno inviato l'ultimo DM10 nel periodo compreso tra dicembre 2000 e novembre 2001 accumulando, quindi, un ritardo fino ad 1 anno.

Tabella 4 – Posizioni contributive delle PMI attive secondo la lista anagrafica del IV trimestre 2002 per classe annuale di ritardo *(valori assoluti)*

Anni di ritardo	Posizioni contributive	Incidenza sull'anagrafe totale (%)
Fino a 1 anno	81.376	5,6
Da 1 a 2 anni	41.082	2,8
Da 2 a 3 anni	25.006	1,7
Da 3 a 4 anni	18.231	1,3
Da 4 a 5 anni	18.633	1,3
Da 5 a 6 anni	17.502	1,2
Oltre 6 anni	65.856	4,5
Totale	**267.686**	**18,4**

Fonte: Oros-INPS

La stessa analisi, inoltre, ha evidenziato che 65.856 posizioni dichiarate attive dall'anagrafe corrente (4,5% sul totale) non hanno mai inviato il DM10 all'INPS dal 1996 al 2001. La considerevole sovracopertura dell'anagrafe e la struttura dei ritardi illustrata, hanno indotto a considerare cessate le posizioni che presentavano un numero di ritardi superiore a tre anni (54 mila posizioni circa, pari al 3,8% sul totale) in aggiunta alle posizioni mai presenti negli universi INPS. Trattandosi di una simulazione, in realtà, nel IV trimestre del 2002 si dispone anche della popolazione corrente e questo consente di valutare l'entità della sovracopertura anagrafica prima e dopo il *data cleaning*, in termini di posizioni contributive (riduzione di oltre 10 punti percentuali l'incidenza delle unità dichiarate erroneamente attive), ma anche in termini di dipendenti. Infatti, dichiarando erroneamente cessate alcune posizioni in base alle informazioni sui ritardi disponibili al tempo corrente, si commette un errore di sottocopertura che sarà quantificabile nel momento in cui si avrà a disposizione il relativo universo. Ad esempio, nella sperimentazione effettuata con riferimento al IV trimestre 2002, il *data cleaning*, nel ridurre la sovracopertura dell'anagrafe spostando tra le posizioni non attive circa 54.000 posizioni (perché presentavano un ritardo superiore a 4 anni), ha indotto un errore di sottocopertura nei dipendenti pari a circa 13 mila unità (media mensile nell'universo dei dipendenti delle posizioni contributive che in realtà hanno presentato la dichiarazione contributiva nel trimestre di analisi) che rappresentano meno dello 0,1 % sul totale dei dipendenti. Di questi circa 9 mila dipendenti sono presenti anche nel campione.

La consistente riduzione della sovracopertura della lista anagrafica dal 25,4% al 16,3% (tabella 5), ottenuta con l'applicazione della procedura generalizzata di *data cleaning*, non è sufficiente ad annullare l'errore di sovracopertura del registro anagrafico rendendo necessaria l'applicazione della probabilità di essere attiva. Nel proseguo dell'analisi, pertanto, i risultati presentati devono essere valutati tenendo conto di tale intervento.

Tabella 5 - Errori medi di sovracopertura dell'anagrafe per sezione di attività economica, in termini di numero di posizioni contributive delle PMI, prima e dopo il *data cleaning*. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*

Sezioni di attività economica	Senza data cleaning	Con data cleaning
C Estrazione di minerali	17,6	10,0
D Attività manifatturiere	18,3	11,4
E Produzione di energia elettrica, gas ed acqua	18,0	10,5
F Costruzioni	38,5	23,6
G Commercio e riparazione di beni di consumo	23,9	15,1
H Alberghi e ristoranti	28,7	18,9
I Trasporti, magazzinaggio e comunicazioni	26,6	17,4
J Intermediazione monetaria e finanziaria	18,9	12,4
K Altre attività professionali ed imprenditoriali	24,3	17,5
C-K TOTALE	**25,4**	**16,3**

Fonte: Oros-INPS

6.2. Metodi alternativi di stima della probabilità di essere attiva

La consistente sovracopertura anagrafica residuale all'applicazione delle procedure di *data cleaning* rende necessaria la riduzione del peso di ciascuna unità della lista di stima per tener conto della aleatorietà nello stato di attività, come specificato nella relazione [12]. Tale relazione rappresenta genericamente quattro equazioni, una per ciascuna variabile ausiliaria (numero di posizioni, numero di dipendenti, monte retributivo e monte oneri sociali), che entrano congiuntamente nel sistema della ponderazione vincolata. Il metodo attuale di calcolo della probabilità di essere attiva (metodo M0) abbatte in modo significativo l'errore di sovracopertura dell'anagrafe pur non consentendo di eliminarlo completamente. Sono stati, pertanto, studiati metodi alternativi di stima delle probabilità

da attribuire alle posizioni contenute nella lista di stima. Varie sono le sperimentazioni effettuate, basate sul tentativo di utilizzare in modo diverso l'informazione disponibile e ricorrendo ad approcci di tipo modellistico per il calcolo delle probabilità. Tra le sperimentazioni realizzate, due in particolare sembrano avere caratteristiche di maggiore robustezza rispetto alla situazione informativa disponibile, fornendo buoni risultati in termini di riduzione dell'errore di stima nei livelli e nelle variazioni. Un primo metodo si basa sulla rimozione di alcune ipotesi all'approccio di base (metodo M1); un secondo metodo attua la predizione delle probabilità attraverso un approccio di tipo *logit* (metodo L6). In entrambi i metodi, il set informativo disponibile si utilizza in modo diverso, in particolare viene modificato l'uso dell'informazione campionaria ai fini della determinazione della lista delle unità attive: M1 ed L6 superano un limite implicito nel metodo M0 costituito dall'uso delle informazioni sullo stato di attività limitatamente alle posizioni contributive del campione ridotto. I due metodi alternativi, infatti, nel definire le unità certamente attive, usano l'informazione più completa fornita dal campione totale[11]. Inoltre, il calcolo delle probabilità, basato sulla stratificazione della popolazione in gruppi omogenei di unità (denominati *register error groups*), fa riferimento ad una diversa classificazione della popolazione su cui le probabilità vengono stimate, rispetto alla classificazione adottata per M0. La ridefinizione dei raggruppamenti trae spunto dalla potenzialità informativa aggiuntiva apportata dall'uso di nuove variabili nella determinazione dello stato di attività.

6.2.1. Il metodo M1

Come per il metodo attuale (M0), il metodo M1 calcola le probabilità di essere attiva utilizzando unicamente l'informazione di *t*-4, ma a differenza del primo, tali probabilità vengono attribuite a tutte le unità appartenenti alla lista delle unità attive A_t indipendentemente alla loro appartenenza al campione. Tuttavia, per non rinunciare all'uso dell'informazione certa sullo stato di attività derivante dalla presenza delle posizioni nel campione totale (CT_t), è stato effettuato un riproporzionamento delle probabilità stesse in virtù del quale alle posizioni del campione totale viene applicata la probabilità $\hat{p}_{it}$ pari ad 1 e, alle restanti, una probabilità proporzionalmente ridotta. In seguito alla riduzione proporzionale delle probabilità delle unità non appartenenti al campione totale, il metodo M1 comporta un'attenuazione rispetto al metodo base M0 del grado di sovracopertura dell'anagrafe. In formula, le probabilità calcolate con il metodo M1 possono essere espresse come:

$$\text{M1:} \quad \hat{p}_{it} = \frac{\#P_{t-4}}{\#A_{t-4}} * q_1\text{-}q_2 \qquad i \notin CT_t \qquad [13]$$

$$\hat{p}_{it} = 1 \qquad i \in CT_t$$

Le quote q_1 e q_2 sono necessarie per riproporzionare le probabilità tra le unità presenti nel campione totale e le restanti. In particolare, alle unità campionarie viene attribuita una probabilità pari ad 1 e alle altre la probabilità stimata viene proporzionalmente ridotta. Attraverso q_1 e q_2 non si fa altro che trasferire una probabilità maggiore ad unità certamente attive perché appartenenti al campione totale e una minore a quelle restanti, senza influenzare la stima dei totali di ponderazione rispetto al caso in cui $\hat{p}_{it} = \frac{\#P_{t-4}}{\#A_{t-4}}$ per $i \in A_t$.

[11] Si veda nota 5 sulla definizione dello stato di attività.

Risolvendo, la relazione $\frac{\#P_{t-4}}{\#A_{t-4}} = \left(1 * \frac{\#CT_t}{\#A_t}\right) + \left(\hat{p}_{it} * \frac{\#\overline{(A_t \cap CT_t)}}{\#A_t}\right)$ per $\hat{p}_{it}$, si ottengono q_1 e q_2, rispettivamente, calcolate come quota delle posizioni dell'anagrafe sulle posizioni del non campione totale $\left(q_1 = \frac{\#A_t}{\#\overline{(A_t \cap CT_t)}}\right)$ e come quota delle posizioni del campione totale sulle posizioni del non campione totale $\left(q_2 = \frac{\#CT_t}{\#\overline{(A_t \cap CT_t)}}\right)$.

Tale assunto implica l'invarianza tra *t*-4 e *t* nella struttura delle unità campionarie e non, rimuovendo la possibile distorsione che si commette con il metodo M0 imponendo alle unità del campione corrente una probabilità diversa da quella stimata.

6.2.2. Il metodo L6

Il metodo L6 prevede una diversa selezione delle unità eligibili e un criterio differente per il calcolo della probabilità, basato su un approccio di tipo modellistico, in cui le probabilità di essere attive vengono stimate mediante un *logit.*

Nel metodo, si attribuisce probabilità pari ad 1 alle unità del campione totale, in quanto ritenute certamente attive. Il modello *logit* per il calcolo della probabilità delle unità non appartenenti al campione, viene stimato sulle unità non campionarie di *t*-4. Egualmente vengono escluse dal calcolo della probabilità anche le unità che, pur non presenti nel campione totale di *t*-4, si trovano nell'universo di *t*-4 e contestualmente nel campione totale di *t*. In tal modo si riesce a tener conto della crescita strutturale del campione dovuta alla propensione delle unità a passare gradualmente alla modalità di invio telematico. Nello stesso tempo si riesce a tener conto anche di eventuali fattori di natura amministrativa (modifica di procedure adottate dall'INPS per scaricare i dati negli archivi centrali, ecc.) che possono aver indotto l'assenza di un numero non irrilevante di unità dal campione di *t*-4.

Il sottoinsieme selezionato per la stima del *logit* è definito come $\left(A_{t-4} - \left(CT_{t-4} + \left(P_{t-4} - CT_t\right)\right)\right)$.

Le probabilità applicate alle varie unità, in formule, possono essere espresse come:

$$\text{L6:} \quad \hat{p}_{it} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}} \quad i \notin CT_t \qquad [14]$$

$$\hat{p}_{it} = 1 \quad i \in CT_t$$

dove $\hat{\beta}$ è il vettore dei coefficienti stimati relativi alle variabili esplicative **x,** che vengono opportunamente selezionate a seconda delle sotto-popolazioni considerate. In particolare, vengono individuate 4 diverse sottopopolazioni a cui applicare altrettanti modelli: le unità panel senza variabili ausiliarie, le unità panel con variabili ausiliarie, le unità nuove nate appartenenti alla sezione I dell'Ateco'02 e le unità nuove nate non appartenenti alla sezione I dell'Ateco'02[12]. In riferimento ad ogni sottopopolazione i *logit* vengono calcolati separatamente per domini di stima, individuati utilizzando informazioni strutturali sulle posizioni contributive. Tali domini di stima sono differenti dai *register error groups* di cui si è parlato nel paragrafo 4. In particolare, si è

[12] L'isolamento delle unità neo nate della sezione I, si è reso necessario in quanto tale raggruppamento presenta caratteristiche di forte variabilità rispetto alle unità classificate nelle altre sezioni. Per questo raggruppamento è stato necessario ricorrere ad un modello estremamente semplificato per la stima delle probabilità delle unità di essere attive.

tentato di sfruttare alcune variabili esplicative all'interno dei *logit*, non solo in formato categoriale ma anche continuo, con l'obiettivo di trarre da esse una maggiore potenzialità esplicativa nella predizione delle probabilità.

6.2.3. Ridefinizione dei *register error groups*

I metodi alternativi di stima della probabilità di essere attiva appena esposti sono stati caratterizzati, rispetto al metodo base, dalla rivisitazione delle partizioni della popolazione su cui le probabilità vengono calcolate ed applicate. Le evidenze tratte dall'analisi sui fattori che spiegano la sovracopertura dell'anagrafe, e che hanno condotto all'attuazione del *data cleaning* (§ 6.1), hanno mostrato l'importanza della variabile ritardo nel definire lo stato di effettiva attività delle unità.
Tale evidenza è stata confermata dall'ausilio del software CART (Classification And Regression Trees) che utilizza una metodologia basata su uno *split* binario ricorsivo delle osservazioni finalizzato ad una segmentazione delle unità in gruppi sempre più omogenei al loro interno rispetto ad una variabile dipendente. In questa sperimentazione la variabile *target* era rappresentata dallo stato di attività effettivo delle posizioni contributive e come variabili di regressione (esplicative) sono state utilizzate quelle alla base della definizione dei *register error groups* (ripartizione geografica, classe dimensionale, età, classificazione di attività economica) e una nuova variabile che misura i mesi di ritardo nell'invio della dichiarazione contributiva. Tale tecnica di segmentazione ha evidenziato che le variabili ritardo ed età sono, in ordine decrescente, le più esplicative sulla cui base sono state individuate delle soglie discriminanti significative.
La variabile ritardo, che come già descritto nel paragrafo 6.1 è definita sulla base del pattern mensile di presenza/assenza del DM10 nella popolazione fino a *t*-4, è stata pertanto utilizzata per il miglioramento della stima della probabilità di essere attiva della sottopopolazione definita "panel senza variabile ausiliaria". Tuttavia, la metodologia CART ha evidenziato che tale variabile è altamente significativa anche per la stima delle posizioni contributive appartenenti alla sottopopolazione delle "panel con variabile ausiliaria", sebbene il ritardo abbia una variabilità limitata poiché, per definizione, questa tipologia di posizioni possono presentare fino ad un massimo di due mesi di ritardo.
Per le posizioni contributive non panel, vale a dire le nuove unità nate nell'intervallo tra *t*-4 e *t*, non è possibile, invece, calcolare il ritardo non disponendo di un pattern di presenza/assenza nella popolazione. Per tali posizioni, si potrebbe sfruttare l'informazione sulla presenza/assenza della dichiarazione contributiva nei campioni mensili relativi all'intervallo temporale considerato, basandosi sull'ipotesi che se una posizione entra nel campione telematico dell'INPS non modifica nel tempo la modalità di invio del DM10. L'eventuale non presenza nel campione, quindi, va considerata una reale assenza. Tuttavia, in tale fase questa informazione non è stata ancora utilizzata in quanto, per ragioni prettamente amministrative, non solo non si dispone di un campione teorico di riferimento, ma quello disponibile, oltre ad essere non casuale, presenta anche una numerosità crescente nel tempo per effetto del ricorso sempre più frequente all'invio telematico dei modelli. Questo implica che il pattern di presenza costruito sulla base del campione potrebbe essere applicato ad un numero troppo limitato di posizioni (vale a dire quelle presenti nel campione di *t*-4), sebbene crescente nel tempo.
Nella fase attuale di sperimentazione del calcolo delle probabilità di essere attiva, i risultati dell'analisi hanno indotto a sperimentare una nuova partizione delle posizioni contributive (*register error groups)*, solo relativamente alla sottopopolazione di stima delle panel senza variabile ausiliaria, introducendo il ritardo tra le variabili di stratificazione. Congiuntamente, in considerazione del fatto che tale variabile risulta essere molto correlata con l'età della posizione contributiva, si è provveduto anche ad affinare il livello di dettaglio della variabile età. Queste innovazioni sono state introdotte nel metodo M1.

Nel metodo L6, invece, i *register error groups* sono appositamente costruiti per l'applicazione dei logit e presentano caratterizzazione completamente differenti ai gruppi predisposti per M0 ed M1. Per ogni sotto-popolazione considerata vengono infatti definiti dei gruppi omogenei, individuati secondo le modalità di alcune variabili strutturali di maggior rilievo (settore di attività economica, classe dimensionale, ripartizione territoriale, classe di età etc.). Per ogni sotto-popolazione è stato adottato un modello probabilistico specifico:

1) per le unità panel senza variabili ausiliarie, la probabilità viene modellizzata unicamente in funzione del numero di mesi di ritardo nell'invio del DM10, definito in forma logaritmica;
2) per le unità panel con ausiliarie, le variabili esplicative sono l'età della posizione contributiva e il suo quadrato, il ritardo nell'invio del DM10 (definito in 3 classi), il numero di dipendenti ausiliari (definito in 4 classi);
3) per le nuove nate appartenenti alla sezione I dell'Ateco'02, la variabile indipendente è rappresentata unicamente dai dipendenti dichiarati all'iscrizione;
4) per le nuove nate non appartenenti alla sezione I dell'Ateco'02 hanno funzione esplicativa i dipendenti dichiarati all'iscrizione (definiti in 2 classi) e l'età della posizione contributiva (definita in 2 classi).

6.3. Quantificazione dell'errore di sovracopertura sui totali di calibrazione: metodi a confronto

La sovracopertura anagrafica, ridotta di oltre il 9% con il *data cleaning*, viene significativamente abbattuta con l'applicazione della probabilità che definisce lo stato di attività delle unità della lista di stima stimata secondo i tre metodi alternativi (tabella 6). Come già precisato, l'impatto di tali modifiche sulla stima del numero di posizioni contributive viene valutato, in questa fase, attraverso una quantificazione dell'errore che si commette nella stima dei totali di calibrazione.
Le probabilità sottostanti al metodo attuale M0 consentono di diminuire l'errore di sovracopertura dell'anagrafe al 3,3% (tabella 6), ma tale sovrastima viene quasi annullata applicando le probabilità determinate secondo i metodi alternativi M1 e L6 (0,5% MAPE e -0,2% MPE).

Tabella 6 - Errori medi di sovracopertura della lista di stima sul numero di posizioni contributive delle PMI per sezione di attività economica: metodi a confronto. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*

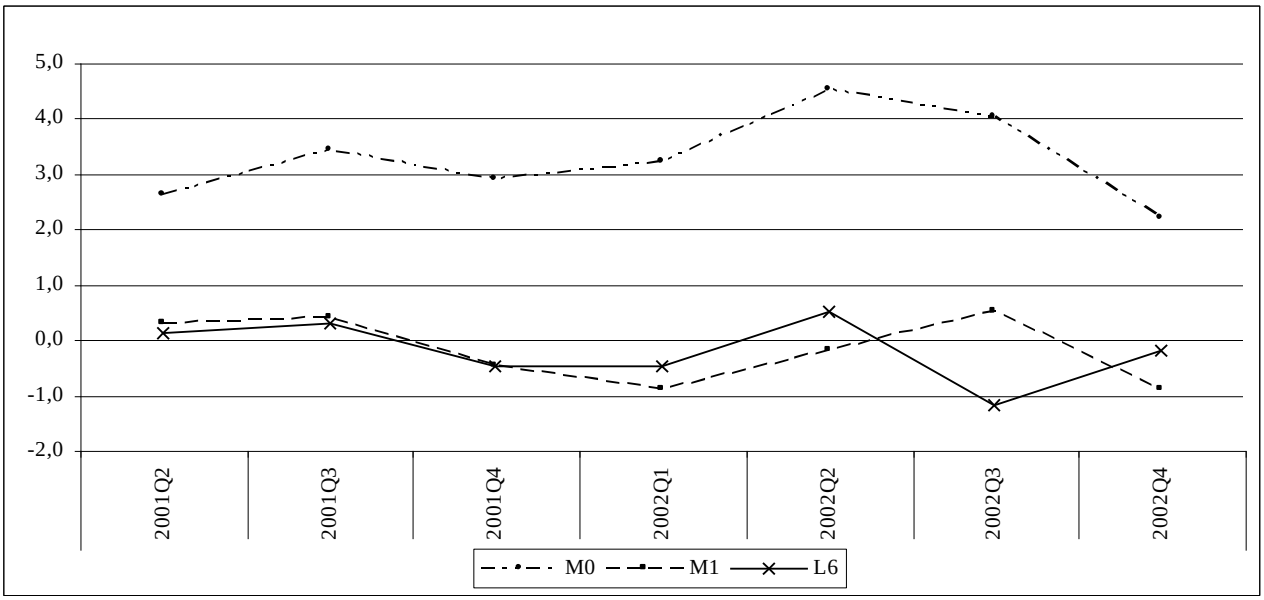
Sezioni di attività economica	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	1,2	1,2	-1,2	1,2	-0,8	0,8
D Attività manifatturiere	2,6	2,6	0,0	0,4	0,0	0,3
E Produzione di energia elettrica, gas ed acqua	2,0	2,0	-1,5	1,5	-1,2	1,3
F Costruzioni	4,9	4,9	0,1	1,1	0,1	0,6
G Commercio e riparazione di beni di consumo	3,0	3,0	-0,2	0,5	-0,2	0,5
H Alberghi e ristoranti	4,0	4,0	0,0	0,7	0,2	0,6
I Trasporti, magazzinaggio e comunicazioni	4,0	4,0	0,3	0,6	0,1	0,5
J Intermediazione monetaria e finanziaria	2,3	2,3	-0,4	0,5	-0,1	0,6
K Altre attività professionali ed imprenditoriali	2,7	2,7	-0,9	1,1	-1,3	1,5
C-K TOTALE	**3,3**	**3,3**	**-0,2**	**0,5**	**-0,2**	**0,5**

Fonte: Oros-INPS

La sistematica sovrastima media prodotta dal metodo M0, tra il II trimestre 2001 e il IV 2002, viene quasi azzerata con i metodi alternativi all'interno del settore industriale, nei comparti relativi alle attività manifatturiere e alle costruzioni e, all'interno del terziario, nei settori degli alberghi e

ristoranti e delle altre attività professionali e imprenditoriali[13]. Essa, invece, lascia il posto ad una marcata sottostima nei settori industriali delle estrazioni minerali (MPE -1,2%) e della produzione di energia elettrica, gas e acqua (MPE -1,5), settori in cui tuttavia il peso delle PMI è ridotto, e ad una più contenuta nei settori del commercio e riparazione di beni di consumo, dell'intermediazione monetaria e finanziaria e delle altre attività professionali ed imprenditoriali.
Si noti che il metodo L6 comporta, in generale, un miglioramento rispetto ad M1 nella stima del numero di unità in quasi tutti i settori, ad esclusione delle altre attività professionali ed imprenditoriali, in cui l'uso del logit si è dimostrato meno efficiente.
La riduzione generale della sovrastima sul totale di calibrazione rappresentato dalle posizioni contributive si evince dal grafico 1 che mostra in serie storica l'errore commesso con i diversi metodi.

Grafico 1 – Errore di sovracopertura della lista di stima sul numero di posizioni contributive delle PMI, secondo i diversi metodi di calcolo delle probabilità di essere attiva. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*

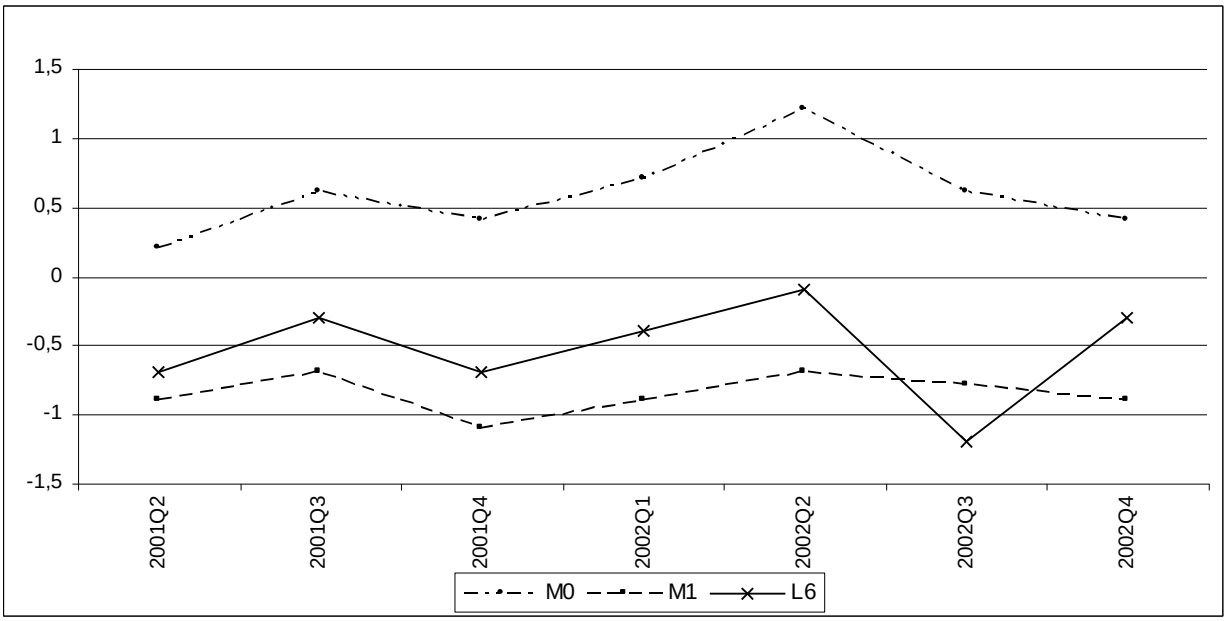


Fonte: Oros-INPS

Le modifiche apportate al metodo M0 hanno prodotto notevoli miglioramenti; tra i due metodi alternativi sperimentati, sebbene non emergano differenze particolarmente rilevanti in termini di posizioni contributive, è da preferire il metodo L6 poiché genera sempre un errore minore tranne nel III trimestre 2002. Tale metodo fa registrare delle performance significativamente migliori rispetto al metodo M1 se valutato in termini di errore nella stima dei totali di calibrazione dell'occupazione (grafico 2). In media, L6 genera un MPE pari -0,5%, mentre per M1 l'MPE è pari a -0,9%. Il metodo L6, che a differenza del metodo M1 sfrutta per definizione l'informazione campionaria (par. 6.2.2) risente, tuttavia, di eventuali cadute della numerosità del campione. In particolare, nel III trimestre 2002[14] in corrispondenza di una riduzione delle unità campionarie, avvenuta per motivi amministrativi, il campione ausiliario utilizzato per il calcolo delle probabilità di essere attive non rappresenta adeguatamente la struttura informativa del campione corrente[15].

[13] Questi ultimi tre settori sono notoriamente caratterizzati dalla presenza di piccolissime e meno stabili attività, in cui le cessazione e/o le sospensioni possono essere ritenute più frequenti.
[14] Per una trattazione più dettagliata sull'evoluzione temporale della dimensione del campione si rinvia al par. 7.1.
[15] In particolare il campione di tale trimestre presenta uno sbilanciamento a favore delle unità neo nate mentre sono assenti solo per motivi amministrativi unità che invece erano già presenti nel campione del trimestre ausiliario, su cui si calcolano le probabilità di essere attive. Ciò comporta una sottostima delle probabilità delle unità assenti nel campione

Grafico 2 – Errore di sovracopertura della lista di stima sul numero di dipendenti delle PMI, secondo i diversi metodi di calcolo delle probabilità di essere attiva. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*



Fonte: Oros-INPS

L'analisi esposta fornisce una misura parziale dell'impatto dei miglioramenti metodologici apportati nella stima della lista delle unità attive sull'errore complessivo di stima dell'occupazione che verrà quantificato nel paragrafo che segue. Tale effetto include, oltre agli errori di lista sopra esposti, anche gli errori di stima legati alla metodologia di ponderazione vincolata adottata.

6.4. L'impatto sulla stima degli occupati dei diversi metodi di definizione della lista di stima

L'occupazione nelle PMI, calcolata applicando i pesi individuali di riporto all'universo, è sovrastimata con il metodo M0 mediamente del 2,8% nei 7 trimestri considerati (tabella 7).
Prima di scendere nel dettaglio dei vari metodi si noti che la misura dell'errore per il metodo M0 è più bassa di quella evidenziata nella tabella 3. Questo miglioramento nelle stime va attribuito essenzialmente all'operazione di *data-cleaning* dell'anagrafe descritta precedentemente e che ha comportato un abbattimento dell'errore totale di 0,3 punti percentuali (attribuibile soprattutto al maggiore impatto nei comparti del commercio e riparazione dei beni di consumo, dell'intermediazione monetaria e finanziaria e dell'estrazione di minerali).
L'adozione delle nuove liste di stima, ottenute con i metodi M1 ed L6, riduce in modo sostanziale e generale l'errore di stima. La stima degli occupati presenta un errore medio che scende dal 2,8% del metodo M0 allo 0,8% del metodo M1 e all'1,1% del metodo L6. Nelle sezioni di attività economica il lieve scostamento tra MAPE ed MPE indica la presenza di sistematica sovrastima media, ad eccezione del settore delle altre attività professionali ed imprenditoriali[16].

---

corrente solo per coincidenza amministrativa (unità alle quali in una situazione ordinaria, nel metodo L6, si sarebbe attribuita una probabilità pari ad 1).

[16] Il settore delle altre attività professionali ed imprenditoriali costituisce un'eccezione per il metodo L6. Il modello logistico scelto per la stima delle probabilità di essere attive, come notato nel par. 3, non presenta performance ottimali per tale settore; l'eccessiva correzione dell'errore di sovracopertura comporta un effetto di compensazione con l'errore di riporto sull'errore totale, rendendo l'MPE di segno negativo. Queste constatazioni rendono necessario un ulteriore approfondimento sul modello da adottare per questo raggruppamento.

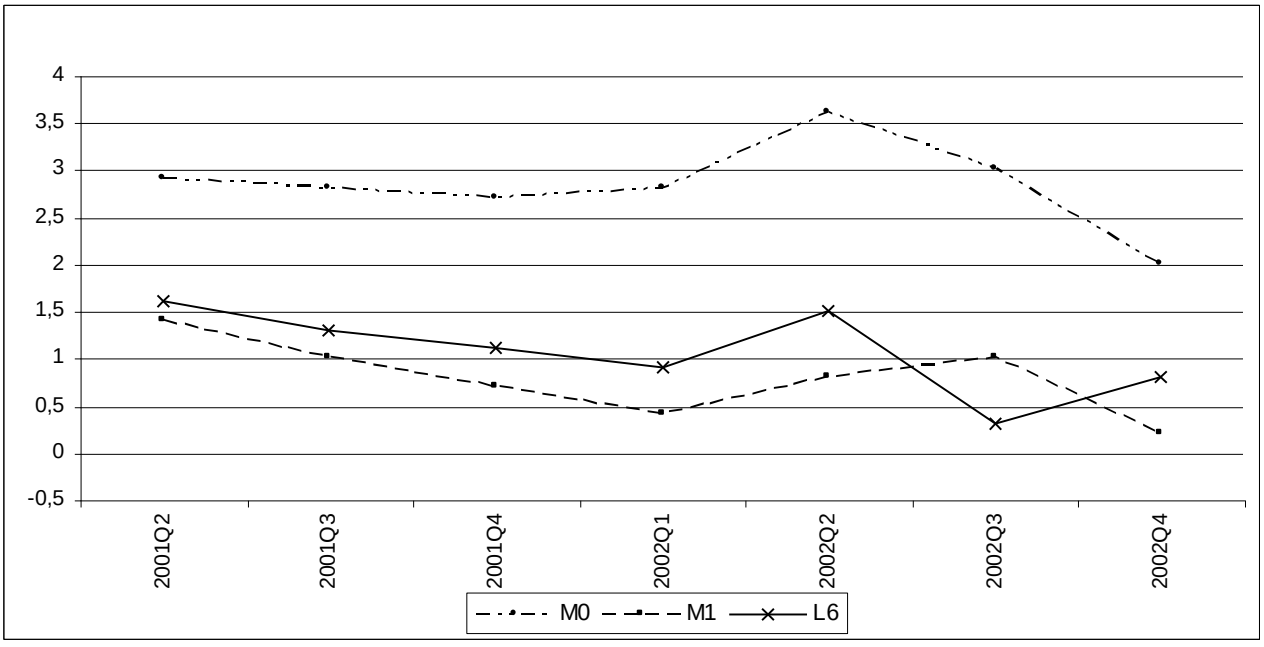
Tabella 7 – Errore di stima nei livelli dell'occupazione. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali medi di periodo)*

Sezione di attività Economica	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	2,4	2,4	0,7	0,9	1,3	1,3
D Attività manifatturiere	1,8	1,8	0,4	0,4	0,7	0,7
E Produzione di energia elettrica, gas ed acqua	2,4	3,7	0,2	2,3	0,6	2,2
F Costruzioni	6,1	6,1	2,8	2,8	3,2	3,2
G Commercio e riparazione di beni di consumo	2,9	2,9	0,8	0,8	1,6	1,6
H Alberghi e ristoranti	3,2	3,2	0,5	0,8	1,3	1,3
I Trasporti, magazzinaggio e comunicazioni	3,3	3,3	0,9	1	1	1
J Intermediazione monetaria e finanziaria	3	3	1	1,4	1,9	2,3
K Altre attività professionali ed imprenditoriali	3,1	3,1	0,6	0,8	-0,3	0,9
C-K TOTALE	**2,8**	**2,8**	**0,8**	**0,8**	**1,1**	**1,1**

Fonte: Oros-INPS

Rispetto alla simulazione che quantifica gli errori dovuti alla lista anagrafica, si evince che i due metodi sperimentati correggono in eccesso l'errore di sovracopertura. Se teoricamente l'errore di riporto nella stima dei livelli, infatti, era stato stimato secondo il metodo di base pari all'1,7% (tabella 3), con i metodi M1 e L6 è come se venisse sottostimato a causa della compensazione che su di esso comporta la sovracorrezione della lista di stima. Se ne deduce che la sola osservazione dell'entità dell'errore totale può fornire un segnale parziale di qualità della stima. L'errore totale è funzione di varie concause che vanno tenute contemporaneamente sotto controllo.

Grafico 3 – Errore di stima dei dipendenti secondo i diversi metodi di definizione della lista di stima. Periodo II trimestre 2001 – IV trimestre 2002 *(valori percentuali)*



Fonte: Oros-INPS

La riduzione dell'errore dovuta all'applicazione dei metodi alternativi di stima delle probabilità di essere attive, tuttavia, appare meno evidente nelle variazioni congiunturali e tendenziali. In media, le variazioni congiunturali ottenute con il metodo L6 producono un errore equivalente ad M0, sia in

termini di MAPE che di MPE (0,5% e -0,2% punti percentuali), mentre il metodo M1 evidenzia un lievissimo peggioramento sull'MPE, che risulta pari a -0,3 punti percentuali, contro -0,2 del metodo base, dovuto unicamente ad un peggioramento nell'errore medio registrato nella sezione dell'intermediazione monetaria e finanziaria (tabella 8).

Tabella 8 – Errore di stima nelle variazioni congiunturali dell'occupazione. Periodo III trimestre 2001 – IV trimestre 2002 *(punti percentuali medi di periodo)*

Sezione di attività Economica	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	0,4	0,8	0,3	1,1	0,4	1
D Attività manifatturiere	0	0,4	0	0,5	0	0,1
E Produzione di energia elettrica, gas ed acqua	1,1	2,1	0,9	1,9	0,9	1,8
F Costruzioni	-0,2	0,7	-0,3	0,7	-0,1	0,4
G Commercio e riparazione di beni di consumo	-0,4	0,7	-0,4	0,6	-0,4	0,9
H Alberghi e ristoranti	-0,2	1,1	-0,2	0,9	-0,3	0,9
I Trasporti, magazzinaggio e comunicazioni	0,2	1	0,1	0,6	0,2	1,2
J Intermediazione monetaria e finanziaria	-0,6	1,9	-0,8	1,8	-0,8	2,2
K Altre attività professionali ed imprenditoriali	-0,5	1,1	-0,5	0,9	-0,4	1,2
C-K TOTALE	**-0,2**	**0,5**	**-0,3**	**0,5**	**-0,2**	**0,5**

Fonte: Oros-INPS

Sulle variazioni tendenziali, i metodi M1 e L6 consentono di ottenere dei lievi miglioramenti rispetto al metodo di base, se valutati in termini di errore assoluto: il MAPE passa da 0,6% di M0 a 0,4% e 0,5% rispettivamente del metodo M1 e L6 (tabella 9). L'errore medio (MPE) registra invece un apparente peggioramento, passando da +0,1% nel metodo M0, a -0,4% e -0,5%, rispettivamente, nei due metodi alternativi. In realtà, il valore quasi nullo dell'MPE nel metodo M0 nell'aggregato C-K sembra essere il frutto di effetti di composizione particolarmente favorevoli. Valutando più attentamente i risultati per sezioni i tre metodi non comportano differenze rilevanti nell'MPE.

Tabella 9 – Errore di stima nelle variazioni tendenziali dell'occupazione. Periodo II trimestre 2002 – IV trimestre 2002 *(punti percentuali medi di periodo)*

Sezione di attività economica	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	1,6	1,6	1,2	1,6	1,2	1,5
D Attività manifatturiere	0,4	0,5	0,1	0,3	0	0,1
E Produzione di energia elettrica, gas ed acqua	5,9	5,9	4,8	4,8	4,3	4,3
F Costruzioni	0,8	1,5	-0,3	0,8	-0,2	0,5
G Commercio e riparazione di beni di consumo	-1	1	-1,5	1,5	-1,7	1,7
H Alberghi e ristoranti	-0,1	1,9	-0,9	1,1	-0,9	1,5
I Trasporti, magazzinaggio e comunicazioni	1,5	1,5	1	1	0,6	1,3
J Intermediazione monetaria e finanziaria	-1,9	1,9	-2,5	2,5	-2,8	2,8
K Altre attività professionali ed imprenditoriali	-0,8	0,9	-1	1	-0,9	0,9
C-K TOTALE	**0,1**	**0,6**	**-0,4**	**0,4**	**-0,5**	**0,5**

Fonte: Oros-INPS

Infine, un'analisi dell'errore totale della sottopopolazione delle PMI distinta per gruppi di stima che tengono conto della diversa disponibilità di informazione ausiliaria (unità nuove nate, unità panel con variabili ausiliarie, unità panel senza variabili ausiliarie), non evidenzia differenze sostanziali rispetto ai risultati generali visti in precedenza, sebbene tali gruppi siano sottoposti ad un

trattamento piuttosto diverso nei tre metodi (in particolare L6 rispetto ad M0 e M1). Nei livelli dell'occupazione per tutte le tipologie la riduzione del *bias* è netta passando da M0 agli altri due metodi; in particolare, mentre M1 presenta una performance migliore nelle unità panel, L6 è più efficiente per le nuove nate (tabella 10). Per quanto riguarda le variazioni congiunturali i tre metodi conducono a errori approssimativamente di uguale dimensione in tutte e tre i gruppi di stima. Sulle variazioni tendenziali il metodo M1 consente di ottenere un miglioramento consistente rispetto a M0 soltanto nelle panel senza variabili ausiliarie. Al contrario il metodo L6 conduce ad errori maggiori rispetto a M0 in tutti i gruppi.

Tabella 10 – Errore di stima dell'occupazione delle PMI per gruppi di stima. Periodo II trimestre 2002 – IV trimestre 2002 *(livelli, variazioni congiunturali e variazioni tendenziali)*

Gruppi di stima della sottopopolazione PMI	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
	Livelli *(valori percentuali medi di periodo)*					
Unità nuove nate	9,6	9,6	4,3	4,3	3,6	3,6
Unità panel senza variabili ausiliarie	26,8	26,8	4,5	5,5	9,9	9,9
Unità panel con variabili ausiliarie	2,0	2,0	0,5	0,5	0,8	0,8
Totale unità	**2,8**	**2,8**	**0,8**	**0,8**	**1,1**	**1,1**
	Variazioni congiunturali *(punti percentuali medi di periodo)*					
Unità nuove nate	-0,9	3,2	-1,1	3,9	-1,0	2,6
Unità panel senza variabili ausiliarie	-0,3	7,8	-1,2	6,0	-0,4	7,6
Unità panel con variabili ausiliarie	-0,1	0,3	-0,1	0,2	-0,1	0,4
Totale unità	**-0,1**	**0,5**	**-0,3**	**0,5**	**-0,2**	**0,5**
	Variazioni tendenziali *(punti percentuali medi di periodo)*					
Unità nuove nate	-1,0	3,4	-2,7	4,8	-3,8	3,8
Unità panel senza variabili ausiliarie	2,9	7,4	-2,3	4,5	-2,4	7,6
Unità panel con variabili ausiliarie	0,0	0,3	-0,2	0,2	-0,2	0,3
Totale unità	**0,1**	**0,6**	**-0,4**	**0,4**	**-0,5**	**0,5**

Fonte: Oros-INPS

7. L'errore di riporto

Nel paragrafo 5 è stata presentata una scomposizione dell'errore totale in una parte dovuta alla stima della lista delle unità attive e in una parte dovuta al riporto all'universo. L'analisi della seconda componente viene affrontata nei paragrafi che seguono, iniziando con una illustrazione delle caratteristiche del campione totale e di quello ridotto. Segue un tentativo di inserire il caso di OROS nella letteratura che si occupa dei problemi di stima in presenza di mancate risposte non casuali, per trarre spunti nella successiva analisi dell'errore di riporto per sottopopolazioni caratterizzate da diversi *set* di informazione ausiliaria. Successivamente, viene valutato e spiegato l'impatto della riduzione del campione sull'errore, per concludere con alcune proposte di modifiche alla metodologia attuale che, preservandone le idee fondanti, possano ridurre l'attuale *bias*.

7.1. Campione totale e campione ridotto

L'insieme di unità che costituiscono il campione totale è un insieme *self-selected* e, come tale, ha per definizione una natura di tipo non casuale. Tuttavia, ha una numerosità molto elevata e crescente nel tempo, garantendo la copertura delle unità per settori di attività economica, classi dimensionali ed età dell'impresa, con una adeguata rappresentazione anche delle nascite. E' inoltre stata verificata la continuità (a meno di eventi demografici) nella presenza delle stesse imprese nel campione (una unità che adotta la trasmissione telematica, nel tempo non cambia tipologia di invio del modello).

Come già accennato nel paragrafo 4, il campione totale soffre di mancate risposte mensili, su un sottoinsieme di unità, che comportano errori di misura nelle variabili trimestrali. Ciò ha indotto a riflettere sull'importanza di effettuare un pretrattamento del campione totale. Nella fase attuale, come già anticipato, si è scelto di basare la stima sull'insieme di unità con informazione completa, quello che è stato definito *campione ridotto.*

Nel grafico 4 si riportano le serie storiche relative al numero di posizioni contributive del campione totale, del campione ridotto e dell'universo, nonché le quote di copertura dei due tipi di campione sull'universo. E' evidente la crescita esponenziale che il ricorso all'invio telematico ha rappresentato negli ultimi due anni, con una percentuale di copertura del campione totale che passa dal 44% a fine 2001 al 76% nel 2003. Si nota, inoltre, che la dimensione del campione sta gradualmente convergendo verso l'universo, con un picco registrato nel II trimestre del 2004, in cui il campione ha superato 1 milione e 200 mila unità, in seguito alla disposizione INPS citata in precedenza. Nel giro di qualche trimestre ci si attende, dunque, di poter disporre, a 45 giorni dalla fine di ciascun trimestre, dell'intera popolazione delle dichiarazione contributive. D'altra parte, il dato sul III trimestre del 2004 conferma come il fenomeno di crescita verso l'universo sia giunto quasi a saturazione.

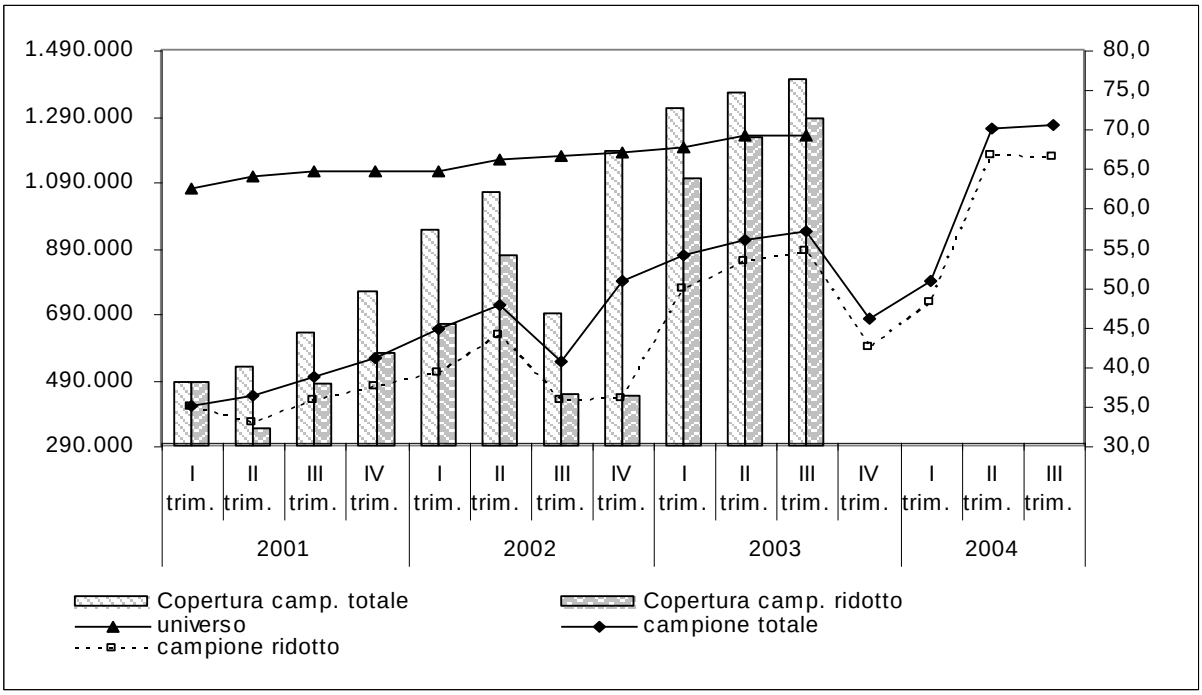
Questa nuova situazione informativa comporta delle forti implicazioni sulle prospettive d'uso di una metodologia predittiva per la produzione di stime anticipate. Tuttavia, come il grafico 4 mostra con chiarezza, la continuità nelle forniture di dati da parte dell'INPS è un fattore tutt'altro che stabile: nel III trimestre del 2002 e nel IV trimestre 2003 si sono registrate improvvise cadute nell'ampiezza del campione totale rispetto ai trimestri precedenti (-30% delle posizioni). I fattori di natura strettamente amministrativa che soprassiedono a questi fenomeni non sono controllabili e a priori prevedibili.

La serie storica del campione ridotto segue abbastanza da vicino quella del campione totale. In media sui sette trimestri considerati si osserva una riduzione di circa 110.000 unità (dalle 764.925 del campione totale medio alle 653.458 del campione ridotto medio). La copertura rispetto all'universo, passando dal campione totale al campione ridotto, si riduce di poco più di 11 punti percentuali (passando dal 59,2% al 48,1%). Tuttavia, dietro queste medie si nascondono situazioni particolari in alcuni trimestri: si osservi, ad esempio, la caduta di copertura del 30,8% (oltre 300.000 unità) registrata nel IV trimestre del 2002 che non ha riscontro nel campione totale. Questo fenomeno è dovuto ad una interruzione nel flusso di dati dalle sedi periferiche all'elaboratore centrale dell'INPS avvenuta a ridosso dello scarico da parte dell'Istat. Ne è derivato che nel file scaricato dall'Istat, mentre la numerosità dei mesi di ottobre e novembre era alta e dello stesso ordine di grandezza, la numerosità del mese di dicembre era molto minore: in altre parole, per oltre trecentomila unità non si disponeva dei dati del mese di dicembre, con la conseguenza che il campione ridotto, la cui numerosità dipende dal mese "meno pieno", è molto diversa da quella del campione totale.

Rispetto al trimestre precedente, in cui il forte calo nella numerosità dei DM pervenuti aveva interessato tutti e tre i mesi, implicando una caduta sia sulla copertura del campione totale sia su quella del campione ridotto, nel IV 2002 la presenza di almeno 1 DM nel trimestre garantisce la tenuta della numerosità delle unità nel campione totale. Sul campione ridotto, invece, l'effetto è dirompente poiché la presenza di un'informazione parziale implica l'eliminazione dell'unità dall'intero trimestre.

Va osservato che, sebbene questi fenomeni siano rari, sono abbastanza tipici di rilevazioni fondate su dati amministrativi e mettono bene in evidenza la dipendenza di chi compie la raccolta secondaria di dati (l'Istat) da chi ne compie la raccolta primaria (l'INPS) ed i rischi ad essa connessi.

Grafico 4 - Evoluzione temporale della dimensione e della copertura del campione totale e di quello ridotto. Periodo I trimestre 2001 – III trimestre 2004 *(numero posizioni)*



Fonte: OROS – INPS

Alcuni risultati dell'analisi sulla copertura del campione totale e di quello ridotto, secondo varie caratteristiche delle posizioni contributive, sono riportati nelle tabelle 12 - 14 con riferimento al IV trimestre 2002.
Il grado di rappresentazione del campione totale è pressoché costante nei vari settori (tabella 11). Ad esclusione del settore E (Energia, gas ed acqua), con una limitata presenza di piccole e medie imprese, si osserva una copertura di almeno il 65% delle unità, con un picco di 70,5% nella sezione D (Attività manifatturiere). La riduzione del campione non comporta un cambiamento di struttura. La quota del campione ridotto rispetto all'universo, infatti, oscilla tra 31 e 38 punti percentuali con valori piuttosto uniformi tra i vari settori che presentano un peso maggiore in termini di occupati. Il settore E rappresenta un'eccezione, in quanto interessato da un taglio di misura relativamente minore rispetto agli altri settori.

Tabella 11 – Grado di copertura del campione totale per sezione di attività economica. IV trimestre 2002 *(valori assoluti, incidenza percentuale)*

Sezione di attività economica	Universo (a)	Campione totale (b)	Campione ridotto (c)	Copertura (b)/(a) %	Copertura (c)/(a) %	Differenza di copertura (d)-(e)
C Estrazione di minerali	2.947	1.922	1.028	65,2	34,9	30,3
D Attività manifatturiere	293.510	206.778	111.029	70,5	37,8	32,7
E Produzione di energia elettrica, gas ed acqua	1.587	830	497	52,3	31,3	21,0
F Costruzioni	206.115	141.799	74.657	68,8	36,2	32,6
G Commercio e riparazione di beni di consumo	310.280	206.641	112.367	66,6	36,2	30,4
H Alberghi e ristoranti	107.061	70.862	39.897	66,2	37,3	28,9
I Trasporti, magazzinaggio e comunicazioni	52.670	34.706	18.297	65,9	34,7	31,2
J Intermediazione monetaria e finanziaria	21.697	14.247	7.901	65,7	36,4	29,3
K Altre attività professionali ed imprenditoriali	168.426	111.132	61.158	66,0	36,3	29,7

Fonte: OROS – INPS

Un alto grado di copertura del campione totale è osservato in tutte le classi dimensionali, anche se si nota una differenza notevole tra le classi più piccole e quelle più grandi (tabella 12)[17]. Al di sotto della soglia dei 100 dipendenti il campione totale copre la popolazione di riferimento per oltre il 64% e in misura pressoché equilibrata nelle varie classi. Al di sopra di tale soglia la copertura scende sotto il 50%. Una ragione possibile è che le imprese di grandi dimensioni, per la compilazione del DM10, ricorrono meno frequentemente a consulenti professionali, che usano da tempo mezzi elettronici per compilare e spedire le dichiarazioni contributive. Al crescere della dimensione, inoltre, per motivi organizzativi è più probabile che l'impresa abbia una gestione della contabilità non centralizzata, implicando che, se non tutte le unità ricorrono all'invio telematico dei DM10, solo una parte dell'impresa risulta presente nel campione. Il passaggio al campione ridotto, pur non modificando in maniera notevole la struttura del campione, incide maggiormente quanto minore è la dimensione aziendale. Questo risultato può essere dovuto al fatto che le imprese di piccole dimensioni hanno un regime di attività meno consolidato, con possibili sospensioni o mancati invii telematici del DM10 e sono, quindi, più soggette a tagli nel passaggio da campione totale a campione ridotto.

Tabella 12 – Grado di copertura del campione totale per classe dimensionale. IV trimestre 2002 *(valori assoluti, incidenza percentuale)*

Classe dimensionale	Universo (a)	Campione totale (b)	Campione ridotto (c)	Copertura (b)/(a) %	Copertura (c)/(a) %	Differenza di copertura (d)-(e)
fino a 10	1.005.885	681.659	366.753	67,8	36,5	31,3
da 10 a 20	93.469	66.764	37.429	71,4	40,0	31,4
da 20 a 100	57.296	36.812	20.550	64,2	35,9	28,3
da 100 a 500	7.434	3.602	2.052	48,5	27,6	20,9
oltre 500	209	80	47	38,3	22,5	15,8

Fonte: OROS – INPS

Il grado di copertura per età si presenta molto uniforme, garantendo una rappresentazione superiore al 65% per tutte le classi, con una leggerissima tendenza a crescere passando dalle posizioni più giovani a quelle più vecchie (tabella 13). Come per la copertura in termini di settori di attività economica, la riduzione del campione lascia sostanzialmente invariata la struttura per classi di età, indicando che il taglio agisce più o meno proporzionalmente su tutte le classi.

[17] E' bene mettere in evidenza che per definizione trattando in tale contesto solo le PMI, la presenza di posizioni nella classe dimensionale 500+, potrebbe apparire come un errore di classificazione. In realtà si tratta di imprese che, secondo la metodologia attualmente in uso, sono state classificate come PMI rispetto ad un trimestre di riferimento e che, pur avendo visto nel corso del tempo accrescere la propria dimensione, hanno mantenuto invariata la loro collocazione tra le imprese di piccole e medie dimensioni.

Tabella 13 – Grado di copertura del campione totale per classi di età delle unità. IV trimestre 2002 *(valori assoluti, incidenza percentuale)*

Classi di età in anni	Universo (a)	Campione totale (b)	Campione ridotto (c)	Copertura (b)/(a) %	Copertura (c)/(a) %	Differenza di copertura (d)-(e)
fino ad 1	135.707	89.486	48.273	65,9	35,6	30,3
da 1 a 2	115.216	77.107	40.275	66,9	35,0	31,9
da 2 a 4	172.469	117.523	61.708	68,1	35,8	32,3
da 4 a 6	114.527	78.615	42.551	68,6	37,2	31,4
oltre 6	626.374	426.186	234.024	68,0	37,4	30,6

Fonte: OROS - INPS

L'analisi condotta sopra ha quindi messo in evidenza come il set dei rispondenti "rapidi", ossia l'insieme delle unità che confluiscono nel campione totale, fornisce una elevata copertura in termini di alcune importanti variabili di stratificazione e, all'interno di esse, una struttura equilibrata. Anche il passaggio al campione ridotto non implica cambi di struttura rilevante, continuando a garantire tassi di copertura elevati. Si tenga infatti presente che l'analisi descritta è riferita al IV trimestre del 2002, che è il trimestre caratterizzato dalla maggiore differenza tra il campione totale e quello ridotto.

Per comprendere più a fondo la necessità di un pretrattamento del campione totale che, nell'attuale implementazione si è tradotto in un taglio, è utile confrontare i pattern di presenza di quelle posizioni del campione totale, escluse dal campione ridotto, con i pattern di presenza che quelle stesse unità assumono nell'universo (tabella 14). Si tratta, in altri termini, di una matrice di transizione che mostra, per pattern di presenza nel campione, il pattern che le posizioni assumeranno nell'universo. Come si può notare, in generale le più alte percentuali sono concentrate nei casi che risultano 111 secondo l'universo mostrando che, una porzione rilevante delle posizioni escluse come mancate risposte, sono in realtà dei ritardi che tendono a riempirsi nei tempi dell'universo.

Il caso più emblematico è quello delle posizioni che hanno un pattern 110 nel campione: solo per il 9,5% esse rimangono 110, mentre per il restante 90,5% diventano 111. Si tratta di un caso caratteristico anche nei trimestri non anomali, in cui l'assenza dell'ultimo mese è dovuta al fatto che, per le regole di scarico del campione, questo risulta ancora incompleto. Nel trimestre considerato, la percentuale di posizioni che cambiano pattern è sicuramente molto più elevata per effetto dei fattori amministrativi che riguardano il campione nel mese di dicembre.

Tabella 14 – Pattern di presenza nei mesi del trimestre IV 2002 delle posizioni del campione totale escluse dal campione ridotto rispetto all'universo *(percentuali rispetto ai totali di riga)*

Flag campione	Flag universo							
	001	**010**	**011**	**100**	**101**	**110**	**111**	**Totale**
001	54,2	-	4,5	-	3,8	-	37,5	100,0
010	-	13,7	41,7	-	-	4,7	39,9	100,0
011	-	-	33,9	-	-	-	66,1	100,0
100	-	-	-	71,3	6,1	1,4	21,2	100,0
101	-	-	-	-	61,2	-	38,8	100,0
110	-	-	-	-	-	9,5	90,5	100,0

Fonte: OROS – INPS

L'analisi condotta, quindi, mostra che usare il campione totale senza pretrattamento comporterebbe notevoli errori, in quanto le variabili trimestrali di una buona parte delle posizioni risulterebbe affetta da errori di misura. In particolare, un uso semplice del campione totale porterebbe ad una sottostima dell'occupazione.

In conclusione di questa analisi si mostra come, in relazione ai pattern di presenza, il campione ridotto appare strutturato diversamente rispetto a quello totale. Per continuità, l'analisi si riferisce ancora al IV trimestre del 2002 (tabella 15), nel quale il passaggio da campione totale a campione ridotto implica l'esclusione del 46% delle unità (percentuale inconsuetamente elevata, come già ricordato). Infatti, come la tavola mette chiaramente in evidenza, in questo trimestre vi è una forte concentrazione di unità che non presentano DM10 nell'ultimo mese (110) che, da sole, rappresentano il 40% del totale. Su di esse il taglio è particolarmente incisivo: queste si riducono del 96% nel passaggio da campione totale al campione ridotto. I tagli incidono in misura inferiore sui pattern in cui le assenze potrebbero essere assimilate a delle nascite (100) o cessazioni (001). Negli altri casi, il passaggio al campione ridotto comporta l'esclusione di almeno l'80% delle unità. Nell'ultima colonna sono riportati i pattern di presenza rilevati nell'universo per le unità presenti nel campione totale. Nel complesso, si nota come il taglio comporti un riequilibrio tra le varie modalità del pattern, ed una maggiore similarità con la struttura che si ritroverebbe nel campione totale se questo fosse misurato senza errore.

Tabella 15 – Pattern di presenza nei mesi del trimestre IV 2002 delle posizioni del campione totale rispetto al campione ridotto *(valori assoluti, incidenza percentuale)*

Presenza nei mesi del trimestre	Campione totale (a)		Campione ridotto (b)		((b)-(a)/(a))%	Universo	
	Numero	%	Numero	%		Numero	%
001	8.962	1,14	2.684	0,63	-70,1	4.854	0,62
010	16.706	2,12	3.457	0,81	-79,3	2.284	0,29
011	33.220	4,21	4.599	1,08	-86,2	18.641	2,36
100	18.016	2,28	7.105	1,66	-60,6	12.839	1,63
101	2.325	0,29	479	0,11	-79,4	2.854	0,36
110	314.401	39,85	13.598	3,19	-95,7	30.922	3,92
111	395.287	50,11	394.909	92,52	-0,1	715.603	90,71
Totale	**788.917**	**100**	**426.831**	**100**	**-45,9**	**788.917**	**100**

Fonte: OROS – INPS

7.2. Stima da campioni non casuali

Nel paragrafo precedente l'insieme dei DM10 giunti per via telematica all'INPS e messi a disposizione all'Istat nei tempi previsti per la stima preliminare, è stato definito *campione*. Si tratta, ovviamente, di un campione che non ha natura casuale. In letteratura ci si riferisce a questo tipo di insiemi come campioni di convenienza (*convenience samples*), di cui ci si avvale al fine di soddisfare particolari esigenze di stima, con la consapevolezza che la loro selezione non rispecchia un particolare disegno campionario.

In tale contesto, un modo per concettualizzare la natura della rilevazione INPS è pensare ad essa come una rilevazione censuaria, ossia che dispone di un campione teorico coincidente con l'universo dei dati. Tale universo è disponibile solo con un ritardo di circa 14 mesi, tempo in cui

OROS rilascia la stima definitiva. Dovendo però produrre una stima anticipata entro 90 giorni dalla chiusura del trimestre, la rilevazione deve basarsi sul set dei rispondenti rapidi. Da questo punto di vista, il contesto informativo di OROS non si discosta concettualmente dalla situazione in cui si trova una indagine tradizionale, che deve rilasciare una stima anticipata in un lasso di tempo insufficiente a raccogliere i dati su tutto il campione teorico. Naturalmente, non vi è nessuna garanzia che il processo di selezione dei rispondenti rapidi rispetti un disegno casuale.

Naturalmente, al fine di ridurre il potenziale bias che deriva da un disegno campionario non casuale, occorre predisporre una metodologia di stima, in cui siano contenute delle ipotesi sul processo generatore dei dati mancanti (ovvero dell'insieme dei dati appartenenti alle unità non rilevate per la stima anticipata).

Per collocare il caso INPS in un quadro concettuale noto, conviene riqualificare il problema in termini di mancate risposte parziali, ovvero pensare che al tempo *t* si osservi l'intero campione teorico (per i dati INPS l'universo) di unità su cui si ha risposta completa sulle variabili ausiliarie (per fissare le idee le variabili datate *t-4*) e risposta parziale sulle variabili correnti.
In riferimento a tale contesto, seguendo Rubin (1976), un processo generatore dei dati mancanti può dare origine a tre tipi di non risposte:
1- MCAR (*missing completely at random*) in cui le non risposte sono assolutamente casuali e non dipendono né dalle variabili mancanti (correnti) né dalle variabili presenti (ausiliarie);
2 - MAR (*missing at random*) dove le non risposte non dipendono dalle variabili mancanti, ma dipendono dalle variabili presenti;
3 - NMAR (*non missing at random*) dove le non risposte dipendono dalle variabili mancanti.
Nei casi 1 e 2 il meccanismo che sta alla base della generazione delle mancate risposte viene detto *ignorabile*, mentre nel caso 3 il meccanismo è *non ignorabile.* In generale, se le unità sono selezionate secondo uno schema probabilistico, il meccanismo che genera mancate risposte può essere facilmente tenuto sotto controllo ai fini della stima. E' questo il caso dei processi *ignorabili*, in cui rispondenti e non rispondenti che presentano le stesse caratteristiche, non differiscono in modo sistematico sui valori delle variabili mancanti. In molti casi, il meccanismo che conduce a mancata risposta è *non ignorabile*, rendendo più problematica la fase di stima.
La conoscenza del meccanismo da cui hanno origine mancate risposte è un elemento centrale nella scelta di una metodologia di stima appropriata, al fine di evitare la produzione di risultati distorti. Il caso MCAR è il più semplice e, in generale, anche stimatori che non fanno uso di variabili ausiliarie (né in termini di stratificazione né in termini modellistici) non comportano stime distorte ma solo un aumento della varianza. Nel caso MAR stimatori che fanno un uso opportuno delle variabili ausiliarie in generale riducono la distorsione. Nel caso NMAR l'uso di variabili ausiliarie non garantisce la riduzione del bias.
Nei casi NMAR l'obiettivo di riduzione del bias si può raggiungere da una parte, tentando di ridurre il fenomeno delle mancate risposte, dall'altra accumulando informazioni circa le differenze tra rispondenti e non rispondenti sulle variabili di riferimento. In questo secondo contesto, cioè, il modello ipotizzato per la descrizione del processo generatore delle mancate risposte va adeguatamente incorporato nel modello statistico finalizzato alla stima (Little, Rubin, 1987).
La metodologia di stima anticipata di OROS è fortemente basata sull'utilizzo di variabili ausiliarie. In particolare, come noto, esse intervengono in due parti: nella formazione dei model groups (divisione, ripartizione, classe dimensionale e classe di età) e nella procedura di calibrazione dei pesi (in cui intervengono il numero dei dipendenti a *t-4*, il monte retributivo a *t-4*, e il monte oneri a *t-4* e il numero dei dipendenti all'iscrizione). Essa sarà quindi più o meno efficace nella riduzione del bias da selezione del campione quanto più il processo che sovrintende a questa selezione dia origine a mancate risposte di tipo MCAR e non di tipo NMAR. Si vedrà, in seguito, che questo non sembra il caso del campione disponibile per le stime OROS.

7.3. L'errore del modello di riporto all'universo

Il procedimento di stima delle PMI è stato descritto in dettaglio nel paragrafo 4. Gli errori medi distinti per sezione di attività economica e per macro gruppo di stima sono illustrati nella tabella 16 che segue. I valori riportati sono al netto di quelli dovuti alla sovracopertura della lista anagrafica e sono quindi da intendersi come dovuti unicamente all'errore del modello di riporto all'universo.
L'errore medio sui sette trimestri considerati, espresso sia come MPE che come MAPE, è pari all'1,9%[18] con valori diversi tra le sezioni di attività economica che si aggirano da un minimo di 1,2% per la sezione D (Attività manifatturiere) ad un massimo di 3,9% per la sezione F (Costruzioni). Si tratta di un errore sistematicamente positivo, come si nota dal fatto che nella maggioranza dei casi MPE e MAPE hanno la stessa dimensione. Passando all'analisi per macro gruppi di stima è evidente come essi presentino errori molto differenziati. Sul totale C-K si registra un errore dell'1,5% per le posizioni panel con informazione ausiliaria, del 5,2% per le posizioni nuove nate e del 10,5% per le posizioni panel senza informazione ausiliaria. Naturalmente, trattandosi di macro gruppi con peso molto differente, l'errore complessivo è molto vicino all'errore del macro gruppo più consistente (panel con informazione ausiliaria). Errori di stima così differenziati vanno attribuiti non tanto alla diversa numerosità dei tre sottogruppi (si tenga presente che il tasso di copertura delle nuove nate non è molto più basso del tasso di copertura delle panel con informazione ausiliaria) quanto alla diversa disponibilità di informazione ausiliaria. Nel caso del primo sottogruppo, infatti, è possibile sia utilizzare una partizione *model groups* più dettagliata, stratificando anche per classe di dimensione aziendale (che è per di più una variabile molto correlata alla occupazione corrente), sia utilizzare più variabili nella procedura di calibrazione dei pesi. In particolare, come ci si può attendere, data l'alta correlazione seriale della variabile occupazione, è molto significativa l'inclusione tra le variabili ausiliarie degli occupati di *t*-4. Per le nuove nate, al contrario i *model groups* non sono stratificati per dimensione aziendale in quanto i dipendenti alla iscrizione non consentono di formare gruppi significativi in termini di numerosità: la variabile è infatti fortemente concentrata sui valori minimi. Analogamente nella procedura di calibrazione, data la scarsa correlazione della variabile dipendenti all'iscrizione con i dipendenti correnti, essa partecipa in maniera ridotta nella riduzione del bias di stima. Per le posizioni panel senza informazione ausiliaria la situazione è ancora peggiore in quanto, oltre a non potere stratificare i *model groups* per dimensione aziendale, si è ritenuto inutile usare la variabile dipendenti all'iscrizione come variabile di calibrazione a causa della ancora più bassa correlazione con le variabili obiettivo.

[18] La differenza con gli errori di riporto della tabella 3 (per il totale C-K, 1,7%) sono dovuti al fatto che in quella tabella erano compresi anche gli errori di sottocopertura (che riducono di 0,2 l'errore di sovracopertura), laddove questa analisi assume che non ci sia sottocopertura.

Tabella 16 – Errori sui livelli per sezione e gruppo nella stima delle PMI da campione ridotto. Periodo II trimestre 2002 – IV trimestre 2003 *(valori in percentuale)*

	Gruppo di stima							
	Nuove Nate		Panel senza informazione ausiliaria		Panel con informazione ausiliaria		Totale	
Sezione di attività economica	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**
C Estrazione di minerali	-1	10,2	53,3	57,4	2,3	2,3	2,7	2,7
D Attività manifatturiere	3,6	3,6	9,6	9,6	1	1	1,2	1,2
E Produzione di energia elettrica, gas ed acqua	4,8	14,8	-12,2	54,6	2,1	3,2	2,6	3
F Costruzioni	8,5	8,5	13,1	13,1	3,1	3,1	3,9	3,9
G Commercio e riparazione di beni di consumo	5,7	5,7	15,2	15,2	1,8	1,8	2,3	2,3
H Alberghi e ristoranti	4,9	4,9	12,4	12,4	1,7	1,7	2,3	2,3
I Trasporti, magazzinaggio e comunicazioni	9,4	9,4	-0,5	6,6	1,1	1,1	1,6	1,6
J Intermediazione monetaria e finanziaria	-2,6	18,9	4,1	10,1	1,9	2	1,1	2,3
K Altre attività professionali ed imprenditoriali	5,7	5,9	6,7	9,1	1,8	1,8	2,1	2,1
C-K TOTALE	**5,2**	**5,2**	**10,5**	**10,5**	**1,5**	**1,5**	**1,9**	**1,9**

Fonte: OROS – INPS

Se la situazione degli errori dei livelli presenta elementi di criticità, quella sugli errori sulle variazioni (tabella 17) appare migliore, riflettendo una certa stabilità nel tempo dell'errore sui livelli. L'errore sul totale in MPE è pari -0,7 punti percentuali con valori bassissimi in D (-0,2), F (-0,3), I (-0,2) ed errori rilevanti in E (-3,1), G (-1,8), K (-2,7).

Tabella 17 - Errori sulle variazioni tendenziali della stima delle PMI sul campione ridotto. Periodo II trimestre 2002 – IV trimestre 2003 *(valori in percentuale)*

Sezione di attività economica	**MPE**	**MAPE**
C Estrazione di minerali	0,6	1,4
D Attività manifatturiere	-0,2	0,4
E Produzione di energia elettrica, gas ed acqua	3,1	4,3
F Costruzioni	-0,3	0,6
G Commercio e riparazione di beni di consumo	-1,8	1,8
H Alberghi e ristoranti	-1,2	1,2
I Trasporti, magazzinaggio e comunicazioni	-0,2	0,6
J Intermediazione monetaria e finanziaria	0,9	3,6
K Altre attività professionali ed imprenditoriali	-2,7	2,7
C-K TOTALE	**-0,7**	**0,7**

Fonte: OROS – INPS

7.4. L'effetto della riduzione del campione sull'errore

L'analisi che segue confronta gli errori di riporto commessi con la metodologia attuale utilizzando il campione ridotto, con gli errori che si commetterebbero a parità di metodo, se fosse possibile usare il campione totale.

La stima realizzata sul campione totale è stata effettuata sostituendo i dati economici campionari con quelli derivanti dai corrispondenti universi. Si tratta, chiaramente, di una simulazione possibile solo perché ad oggi sono disponibili i dati sugli universi dei trimestri di cui si tratta. Lo scopo di tale operazione è quello di fornire un limite inferiore all'errore del metodo di stima, ipotizzando di disporre di dati completi su tutte le posizioni appartenenti al campione totale. L'utilità di questo esercizio è duplice: da una parte consente di capire se, e in che misura, la distorsione delle stime è

causata dalla selezione del campione implicata dalla operazione di riduzione o, invece, sia da collegarsi alle caratteristiche del campione totale. D'altra parte consente di capire quale è la massima riduzione dell'errore che si otterrebbe se si procedesse ad una operazione di imputazione sulle posizioni contributive che presentano mancate risposte nel trimestre. In altri termini poiché l'universo fornisce, per definizione, un'informazione definitiva, corretta e completa, si può pensare che il campione totale sia stato sottoposto ad una operazione di imputazione delle mancate risposte mensili[19].

Sul totale dei gruppi di stima, MPE e MAPE passano da 1,9% nel caso di stima con campione ridotto a 0,5% nel caso di uso del campione totale (tabella 18).

L'effetto sui gruppi di stima presenta delle differenziazioni. L'errore relativo maggiore dovuto al modello di riporto si commette sulla stima delle unità panel con variabile ausiliaria che, pur mostrando livelli molto bassi in termini assoluti (1,5%), nella stima da campione ridotto è 5 e 4 volte superiore se confrontati all'errore di stima ottenuta con il campione totale.

MAPE ed MPE raddoppiano nel passaggio da campione totale a campione ridotto nel gruppo delle panel senza informazione ausiliaria. Si osservi tuttavia, come l'MPE passi da un valore negativo, che denota mediamente sottostima della variabile di riferimento (-4,6%) ad una sovrastima nel caso di campione ridotto (10,5%). Infine, nel totale delle unità classificate come neo nate, i due insiemi campionari non evidenziano grosse differenziazioni in termini di errore prodotto dalla metodologia di stima. In questo caso, sembrerebbe che il campione ridotto non si distanzi molto da quello integrale.

Complessivamente, si notano notevoli differenze se si esaminano i risultati per settore di attività economica. Mentre il differenziale assoluto maggiore si osserva nel settore F delle costruzioni ed E dell'energia, in termini relativi l'errore di più elevata entità si registra nei settori I e D.

Il peggioramento, sia in termini di MPE che di MAPE che si rileva nel settore dell'intermediazione monetaria (J) è dovuto all'erronea classificazione di alcune unità di elevata dimensione tra le neonate del campione totale.

Sul gruppo di stima delle panel con variabile ausiliaria, che rappresentano la quota più significativa delle unità, sono i settori D ed E quelli che presentano maggiori problemi per effetto del modello di riporto: nella sezione D, MPE e MAPE praticamente si annullano con il campione totale a partire da un livello pari ad 1 nella stima con campione ridotto.

Sul gruppo delle nuove nate, ad esclusione dei settori C ed E, per definizione costituiti da un numero molto esiguo di unità, e in cui effetti di ricomposizione dei gruppi potrebbero giustificare i livelli più elevati di MPE e MAPE connessi all'uso del campione totale, si osserva come l'errore relativo più ampio si registri nel settore D della manifattura, in cui gli errori medi passano da 3,6 a 1,8 per l'MPE e 2,2 per il MAPE e K delle altre attività professionali. Nelle panel senza informazione ausiliaria è in G, H ed F che il modello di riduzione della distorsione dovuta alla selezione del campione implica errore relativo maggiore. Si noti come in questo raggruppamento, ad esclusione di E, la selezione del campione implica generale sovrastima della variabile di riferimento quando, invece, con il campione totale si tenderebbe a sottostimare.

[19] La considerazione dei dati economici dell'universo implica che una piccola quota dell'errore delle stime ottenute con il campione ridotto vada imputata al fatto che i dati campionari, come già ricordato sono sottoposti ad una meno attenta ispezione preliminare da parte dell'INPS, che invece è più attenta sui dati che vengono inviati per mezzo cartaceo ossia quelli che convergono negli universi.

Tabella 18 – Errori per sezione e gruppo nella stima delle PMI da campione totale. Periodo II trimestre 2002 – IV trimestre 2003 *(valori in percentuale)*

	Gruppo di stima							
	Nuove Nate		Panel senza variabili ausiliarie		Panel con variabili ausiliarie		Totale	
Sezione di attività economica	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**
C Estrazione di minerali	-10,1	16,7	9,1	20,5	0,7	1,3	0,5	1,7
D Attività manifatturiere	1,8	2,2	-5,8	7,5	0,1	0,3	0,1	0,4
E Produzione di energia elettrica, gas ed acqua	-0,9	11,2	-38,9	47,7	0,2	1,6	-0,4	2,5
F Costruzioni	5,2	5,2	-2,9	5,7	0,6	1,1	0,8	1,4
G Commercio e riparazione di beni di consumo	4,7	4,7	-0,3	2,8	0,7	0,7	0,9	0,9
H Alberghi e ristoranti	3,1	3,1	-2,1	4,1	0,6	0,6	0,7	0,7
I Trasporti, magazzinaggio e comunicazioni	5	5,4	-12,2	12,2	-0,3	0,8	-0,1	0,8
J Intermediazione monetaria e finanziaria	39,3	45,7	-2,4	5,4	1,6	1,7	4,3	4,3
K Altre attività professionali ed imprenditoriali	2,9	3,5	-7,9	9	0,4	0,6	0,4	0,7
C-K TOTALE	**4,4**	**4,4**	**-4,6**	**5,6**	**0,3**	**0,4**	**0,5**	**0,6**

Fonte: OROS - INPS

La tabella 19, che a titolo esemplificativo rappresenta il IV trimestre 2002, da un'idea del tipo di selezione che conduce al campione ridotto e del tipo di sbilanciamento che la selezione può indurre rispetto alla struttura di presenze rappresentata nell'universo. Nella tavola sono rappresenti i pattern di presenza (1) / assenza (0) dei DM10 nei 3 mesi del trimestre delle unità considerate. Mentre le posizioni con DM10 presente nei 3 mesi risultano sovrarappresentate se confrontate con l'universo (97% del totale), vi è una limitatissima rappresentazione delle unità 101 e 011, mentre vengono completamente esclusi i casi di 010.

Sulle stesse unità viene effettuata una verifica sul pattern di presenza nel trimestre *t-4* ausiliario, sulla base di cui vengono costruiti i totali noti. Nella tavola, le unità vengono dunque riclassificate secondo il pattern di presenza del trimestre ausiliario. Come si nota, lo sbilanciamento tra campione ed universo si riduce drasticamente, con un gap che sugli 111 non arriva ad 1 punto percentuale. In termini assoluti, il divario si mantiene ampio nei casi di 101 e 110.

Tabella 19 – Pattern di presenza[(a)] nei mesi del trimestre IV 2002 delle posizioni riferite alle PMI nel campione ridotto e del trimestre ausiliario IV 2001 nell'universo *(valori in percentuale)*

Presenza nei mesi del trimestre	**Incidenza nell'universo IV 2002**	**Incidenza nel campione ridotto IV 2002**	**Incidenza nell'universo IV 2001**	**Incidenza nel campione ridotto IV 2001**
001	1,1	0,4	2,6	2,3
010	0,3	0	0,4	0,3
011	1,5	0,3	3,1	2,7
100	2	1,3	1,5	1,4
101	1,5	0,2	2,6	1,8
110	4,2	1,2	3,1	4,1
111	89,3	96,6	86,6	87,4

[(a)] La presenza è denotata da 1 e l'assenza da 0. La sequenza mostra gli eventi nei tre mesi del trimestre.

Fonte: OROS – INPS

L'analisi della rappresentatività per presenze è rilevante in quanto ad esse sono correlate i livelli medi dell'occupazione. Il dato mensilizzato dell'occupazione di una posizione contributiva è infatti ottenuto dividendo per tre la somma degli occupati dichiarati nei DM10 mensili pervenuti. Ne deriva che il dato delle posizioni con pattern di presenza completo è mediamente più elevato di quello delle posizioni con pattern di presenza incompleto. La sovrarappresentazione relativa delle posizioni con pattern completo implica quindi una sovrarappresentazione relativa delle posizioni con più occupati. L'effetto distorcente sulla stima è però dovuto al fatto che il tipo di selezione effettuato non comporta una analoga sovrarappresentazione sulle variabili ausiliarie. Ciò è importante in quanto la calibrazione agisce proprio su queste ultime. Euristicamente si può dire che i pesi calibrati rifletteranno il rapporto che esiste tra campione ed universo tra i diversi pattern, ma questi pesi si applicheranno ad unità che non riflettono alla stessa maniera l'universo corrente.
Tale selezione avversa è causata dalla scarsa qualità dell'informazione anagrafica che porta ad una sottoidentificazione delle unità con pattern incompleto che non necessitano di una correzione del dato sull'occupazione e che quindi possono essere inclusi senza problemi nel campione. Un esempio può chiarire il problema. E' plausibile che nel passaggio tra il campione totale e il campione ridotto siano state escluse in maniera errata troppe unità con pattern incompleto alla fine del periodo (110 o 100) in quanto l'informazione anagrafica non registra in tempo che si tratta di eventi di cessazione o di sospensione. In altri termini ci si trova di fronte ad un altro inconveniente generato dalla problema delle cessazioni non registrate.
E' però molto utile notare che laddove gli errori di stima sui livelli si riducono molto se si potesse usare un campione totale corretto, gli errori sulle variazioni rimangono pressoché invariati sul totale e, con risultati differenti a secondo delle sezioni, con sezioni grandi come D dove la stima sul campione ridotto è molto superiore (tabella 20). La ragione di ciò potrebbe risiedere nel fatto che la riduzione del campione tende ad accentuare le caratteristiche panel del campione tra trimestri migliorando la qualità della stima delle variazioni. Va inoltre notato che, mentre gli errori della stima sul campione ridotto sono in MPE negativi, quelli sul campione totale sono positivi. Ciò dipende dal fatto che nel tempo gli errori sui livelli della stima sul campione ridotto sono diminuiti laddove quelli sulla stima con il campione totale sono aumentati.

Tabella 20 - Errori sulle variazioni tendenziali della stima delle PMI. Periodo II trimestre 2002 – IV trimestre 2003 *(valori in percentuale)*

Sezione di attività economica	Campione totale		Campione Ridotto	
	MPE	**MAPE**	**MPE**	**MAPE**
C Estrazione di minerali	3,7	3,7	0,6	1,4
D Attività manifatturiere	0,8	0,8	-0,2	0,4
E Produzione di energia elettrica, gas ed acqua	6,3	6,3	3,1	4,3
F Costruzioni	2,9	2,9	-0,3	0,6
G Commercio e riparazione di beni di consumo	0,1	0,1	-1,8	1,8
H Alberghi e ristoranti	0,4	0,4	-1,2	1,2
I Trasporti, magazzinaggio e comunicazioni	1,9	1,9	-0,2	0,6
J Intermediazione monetaria e finanziaria	-5,2	5,6	0,9	3,6
K Altre attività professionali ed imprenditoriali	-0,6	0,8	-2,7	2,7
C-K TOTALE	**0,7**	**0,7**	**-0,7**	**0,7**

Fonte: OROS – INPS

7.5. Alcune proposte di modifica alla metodologia attuale

L'analisi svolta ha messo in luce alcuni punti che possono essere così sintetizzati:
1) La stima dei livelli dell'occupazione è abbastanza distorta ma, dato che la distorsione è non troppo variabile nel tempo, le variazioni hanno un grado di distorsione molto minore.
2) Il bias è in buona parte imputabile ad un processo di selezione del campione non ignorabile ai fini della stima degli occupati.
3) La non ignorabilità del processo di selezione riguarda in primo luogo il campione totale proveniente dall'INPS.
4) La riduzione del campione ha aumentato il grado di non-ignorabilità del processo di selezione, aumentando il bias nella stima del livello degli occupati.
5) Date le variabili ausiliarie usate e quelle disponibili, il bias è più grave per le posizioni nuove nate e le posizioni panel senza informazione ausiliaria, mentre si presenta meno grave per le posizioni panel con informazione ausiliaria.
6) Gli errori commessi sulle variazioni sono di entità paragonabile a quelli che si commetterebbero se si potesse usare il campione totale. Ciò è dovuto probabilmente al fatto che la riduzione del campione aumenta le caratteristiche panel del campione tra i trimestri stabilizzando l'errore sui livelli e generando effetti di compensazione sull'errore delle variazioni.

Al fine di ridurre gli errori di livello possibilmente migliorando anche quelli sulle variazioni, sono state identificate tre proposte di modifica alla metodologia corrente che ne rispettino lo spirito. Esse ruotano attorno ai seguenti concetti:
1. bilanciamento del campione rispetto ai pattern di presenza;
2. imputazione delle mancate risposte parziali del campione totale.

7.5.1. Bilanciamento del campione rispetto ai pattern di presenza

L'analisi svolta nel paragrafo 7.1 ha evidenziato che la riduzione del campione provoca una ulteriore selezione a favore di posizioni con una occupazione più alta. Ciò deriva dal fatto che il processo di selezione include troppe poche unità con dati incompleti nel trimestre a causa della sottoidentificazione delle posizioni che hanno subito variazioni demografiche nel trimestre.
In altre parole, il campione selezionato rappresenta in maniera maggiore le posizioni con dati completi nel trimestre, di quanto rappresenti le posizioni con dati incompleti. Dato che il valore mensilizzato dell'occupazione delle posizioni con dati completi è mediamente maggiore dei valori delle posizioni con dati incompleti ciò porta ad una sovrastima dell'occupazione totale. E' interessante notare che il modello di riporto (*model groups* e calibrazione) non riesce a ridurre molto il bias in quanto il processo di selezione accentua le caratteristiche NMAR dei dati. Ciò avviene perché la selezione, incidendo solo sulle informazioni correnti, non riguarda le variabili ausiliarie. Accade cioè che, guardando alle variabili ausiliarie, le posizioni del campione hanno una rappresentatività diversa che guardando alle variabili correnti.
Se il problema della stima risiede nello sbilanciamento del campione tra variabili correnti e variabili ausiliarie una diversa selezione del campione potrebbe ridurre il bias. A titolo di esercizio il campione è stato ridotto nella seguente maniera. Sulle posizioni panel con variabile ausiliaria sono state selezionate solo le posizioni con pattern di presenza completo sia per quanto riguarda il trimestre corrente che per quanto riguarda il trimestre ausiliario. Sugli altri due gruppi la selezione è la stessa del metodo base. Questa scelta è giustificata dalla esigenza di bilanciare il campione tra variabili ausiliarie e variabili correnti. L'idea è che il campione ridotto in questa maniera rappresenterà le posizioni con pattern di presenza completo in maniera più omogenea tra variabili ausiliarie e variabili correnti.

I risultati di stima sui dipendenti sono riportati nelle tabella 21 che segue, in cui gli errori si riferiscono ai livelli.

Tabella 21 – Errori per sezione nella stima dei livelli delle PMI da campione ridotto, nel metodo di base (S1) e con selezione sul pattern di presenza in t-4 (S1 – a). Periodo II trimestre 2002 – IV trimestre 2003 *(valori in percentuale)*

Sezione di attività economica	S1		S1 - a	
	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	2,7	2,7	0,6	0,6
D Attività manifatturiere	1,2	1,2	0,3	0,6
E Produzione di energia elettrica, gas ed acqua	2,6	3	0,8	2,2
F Costruzioni	3,9	3,9	1,5	1,5
G Commercio e riparazione di beni di consumo	2,3	2,3	0,7	1,1
H Alberghi e ristoranti	2,3	2,3	0,9	1
I Trasporti, magazzinaggio e comunicazioni	1,6	1,6	-0,1	0,7
J Intermediazione monetaria e finanziaria	1,2	2,6	1,5	2,1
K Altre attività professionali ed imprenditoriali	2,1	2,1	0,2	1,6
C-K TOTALE	**1,9**	**1,9**	**0,5**	**0,8**

Fonte: OROS – INPS

L'effetto sui livelli è positivo; l'errore di stima si riduce dall'1,9% allo 0,5% (MPE) e dallo 1,9% allo 0,8% (MAPE). E' interessante notare come i risultati di stima siano migliorati sebbene la numerosità del campione si sia drasticamente ridotta. Purtroppo gli incoraggianti risultati sui livelli non trovano conferma negli errori delle variazioni. Infatti, sebbene gli errori dei singoli trimestri siano minori che nel caso del modello base, essi presentano una minore stabilità nel tempo e conseguentemente conducono ad un aumento degli errori delle variazioni tendenziali. Il probabile motivo di ciò risiede nel fatto che si è probabilmente ridotto l'errore sulle unità più stabili (con presenza completa nel trimestre corrente e in quello ausiliario), mentre una minore rappresentatività e quindi un peggioramento della stima si è avuto sulle unità più instabili e che quindi presentano pattern incompleto in uno o entrambe i trimestri. Sebbene la prima popolazione sia la più consistente, la seconda contribuisce in maniera rilevante a determinare la dinamica dell'occupazione.

7.5.2. L'imputazione delle mancate risposte mensili

Nel paragrafo 7.1 sono state descritte le motivazioni della non fattibilità nel ricorso al campione totale finché sussistono posizioni contributive con dati incompleti nel trimestre. Normalmente sono circa il 10% le posizioni del campione totale che presentano meno di 3 DM nel trimestre. Tuttavia, non tutte sono da attribuire a ritardi nell'invio del modello; una parte è realmente assente per motivi prettamente demografici (cessazione, inattività, stagionalità).
In generale, l'uso dei dati del campione totale senza alcuna correzione per i DM incompleti comporterebbe una sottostima del livello dell'occupazione che in alcuni casi può essere drammatica, ad esempio in quei trimestri come il IV 2002 in cui, a causa di fenomeni amministrativi, la percentuale di posizioni contributive con dati incompleti ha registrato una percentuale di oltre il 40%.
La ricostruzione delle informazioni sui trimestri con DM mancante deve necessariamente percorrere due passi successivi:

- l'identificazione dell'effettivo stato di attività della posizione contributiva per i mesi con DM mancante (discriminare tra effettiva inattività e ritardo del DM10);
- l'imputazione delle variabili mancanti (l'occupazione, in primo luogo, ma anche le retribuzioni e gli oneri sociali) per i mesi identificati come ritardi.

Un primo tentativo di identificazione dello stato di attività, che fa uso delle informazioni anagrafiche e del pattern di presenza nel trimestre *t-4*, è già eseguito nell'attuale selezione del campione ridotto a partire dal campione totale. Come evidenziato nel paragrafo 7.1, l'operazione conduce tuttavia ad una sostanziale sottoidentificazione delle unità con mesi di inattività, con conseguente eccessiva esclusione di posizioni contributive dal campione. Questo errore è dovuto in gran parte al noto problema delle cessazioni non registrate. Questo approccio, che per la natura della selezione dello stato di attività può essere definito *deterministico*, ha chiaramente dei limiti, che sono quelli imposti dalla carenza di informazioni affidabili e, comunque, fino ad ora pienamente sfruttate.
Una via alternativa per l'individuazione dello stato di attività potrebbe essere quella di ricorrere ad un approccio *probabilistico*, in cui sia possibile attribuire a ciascuna posizione con pattern incompleto, una misura di probabilità per ogni pattern che può assumere. Si tratta, ad esempio di stimare le probabilità che una posizione con pattern 100 nel campione totale assuma pattern 110, 101, 111 o che mantenga il pattern originario. Chiaramente le modalità che possono condurre ad una stima di tale probabilità sono varie. Si potrebbe, ad esempio, fare ricorso all'informazione di *t-4*, in cui si dispone sia del campione, sia dell'universo. La stima della probabilità potrebbe essere effettuata per gruppi omogenei di unità, individuati secondo qualche variabile discriminante, mettendo a confronto il pattern del campione con quello dell'universo corrispondente.
La misura di ciascuna delle probabilità individuate potrebbe essere inserita nella relazione individuata per l'imputazione del dato mancante, come fattore di peso per le varie possibilità. Per l'imputazione dell'occupazione (ma anche degli oneri), può essere opportuno utilizzare come base di riferimento per la ricostruzione del dato mancante, l'informazione in *t*.
Così, continuando l'esempio di sopra, se y_{it} è il dato di occupazione rilevato nel primo mese del trimestre *t* per la posizione *i* di una certa cella, il valore stimato della variabile trimestrale occupazione per la posizione potrebbe essere così calcolato:

$$\hat{y}_{it} = \frac{y_{it}}{3} p(100|100) + \frac{2y_{it}}{3}[p(110|100) + p(101|100)] + \frac{3y_{it}}{3} p(111|100) \qquad [15]$$

Dove, ad esempio, *p(110|100)* è la probabilità, all'interno di un certo gruppo, che una posizione con pattern 100 nel campione totale assuma nell'universo pattern 110.

Naturalmente, questo metodo si basa su ipotesi di invarianza temporale che sono ardue da sostenere visto che ci troviamo di fronte a dei fenomeni che in alcuni casi sono imprevedibili.
In generale, a causa della scarsità nelle informazioni disponibili, non è semplice individuare metodi alternativi che portino ad una stima adeguatamente affidabile dello stato di attività, con conseguenze non prevedibili a priori sulla correttezza delle stime, sia in termini di livello che di variazioni. Tuttavia, questo è un campo su cui è fondamentale continuare a riflettere e a sperimentare.
Una considerazione deve essere comunque fatta e riguarda la riduzione dell'errore che si potrebbe ottenere con un metodo di imputazione delle posizioni escluse dal campione ridotto. L'analisi fatta nel paragrafo 7.4 ha mostrato che se anche si potesse usare un modello di imputazione con errore zero (tale è la situazione se si usano su quelle posizioni i dati dell'universo) continuerebbe a sussistere un errore positivo sistematico sui livelli.

8. Conclusioni e prospettive

Questo lavoro documenta una prima fase di analisi degli errori di revisione della stima dell'occupazione secondo l'attuale metodo di stima della rilevazione OROS. Nel periodo di osservazione, per il quale si dispone sia delle stime finali sia di quelle preliminari, l'errore di revisione sul livello dei dipendenti è ampio, risultando in media pari al 2,1% ma con una componente sistematica molto stabile; l'ampiezza dell'errore è, infatti, molto più contenuta (in media dello 0,4%) se misurata sui tassi di variazione. Sulla base di quest'ultimo risultato si può considerare che le attuali stime della dinamica dell'occupazione, utilizzate per la costruzione dei numeri indice trasmessi ad Eurostat, sono caratterizzate da errore di revisione sufficientemente contenuto. Resta, tuttavia, il problema che tale grado di precisione potrebbe porre dei problemi per la diffusione delle stime a livello nazionale, soprattutto quando si considerino le disaggregazioni settoriali per le quali l'errore è, in molti casi, assai più elevato.

La stima è il risultato dell'applicazione di diverse procedure su differenti sottopopolazioni che contribuiscono all'errore in misura differente. In termini di incidenza sull'errore totale, riveste particolare importanza la sottopopolazione delle PMI, che ha evidenziato un errore medio pari al 3,1% sui livelli della variabile e pari allo 0,6% sulle variazioni tendenziali. Per questa ragione, è proprio sulla popolazione delle PMI che si sono concentrate le sperimentazioni. Due sono le fonti che spiegano l'errore del 3,1% sopra citato: l'errore dovuto alla sovracopertura dell'anagrafe, che contribuisce per circa l'1,4%, e l'errore di riporto all'universo a cui si può attribuire il rimanente 1,7%. Un'accurata revisione della metodologia correntemente utilizzata per la produzione delle variabili retributive ha consentito di compiere notevoli progressi nella risoluzione del problema della sovracopertura agendo, in particolare, mediante l'implementazione di una procedura di *data-cleaning* per l'individuazione e l'eliminazione delle unità cessate e l'identificazione di modelli per la definizione delle probabilità di esistenza nella lista anagrafica. Le innovazioni introdotte inducono un abbattimento sostanziale dell'errore medio sui livelli che scende a circa 1 punto percentuale in termini di MAPE, restando tuttavia sistematico, come si evince da valori pressochè analoghi dell'MPE. Non si osservano, invece, apprezzabili miglioramenti sugli errori nelle variazioni congiunturali che, tuttavia, si mantengo a livelli molto bassi. Sulle variazioni tendenziali l'errore si riduce in termini di MAPE, ma non in termini di MPE. In definitiva, relativamente all'errore commesso nella stima delle variazioni tendenziali, le analisi effettuate conducono a due importanti risultati. Per un verso emerge che la metodologia di base, pur affrontando alcuni aspetti specifici del trattamento dei dati in modo approssimato, conduce ad una dimensione dell'errore di revisione sulla dinamica della variabile di interesse non molto dissimile da quello che si ottiene con successivi miglioramenti dell'impianto di stima. Per altro verso, risulta che gli approcci sinora sperimentati non riescono ad abbattere l'errore di revisione delle stime preliminari nella misura che sarebbe desiderabile dal punto di vista della diffusione dei dati a livello nazionale.

Le sperimentazioni sviluppate in questa fase, comunque, hanno condotto all'individuazione di altri aspetti della procedura di stima sui quali può essere utile approfondire l'analisi. In seguito alla riduzione del problema di sovracopertura ha acquisito importanza un ulteriore aspetto critico caratterizzante la lista, rappresentato dalla sottocopertura dell'anagrafe che, in precedenza, esercitava un'incidenza molto limitata sull'errore complessivo. E' questo uno degli elementi su cui occorrerà prestare attenzione negli sviluppi successivi della metodologia.

Più complessa la situazione sul fronte del modello di riporto all'universo, il cui errore deriva principalmente da caratteristiche spiccatamente NMAR delle unità del campione ridotto. Esse implicano una selezione non ignorabile del campione che induce un bias che difficilmente può essere ridotto in maniera consistente da una modifica della partizione in *model groups* e/o della procedura di calibrazione. D'altra parte, data la natura del campione, una modifica al pretrattamento del set delle unità campionarie (imputazione o selezioni diverse), non necessariamente comporterebbe una riduzione dell'errore sulle variazioni tendenziali (parametro obiettivo primario).

Un'altra possibilità da valutare, inoltre, è la modifica dell'impianto di stima verso un uso più longitudinale dell'informazione disponibile.
Accanto al completamento delle iniziative di revisione metodologica realizzate nel corso degli ultimi mesi, sarà necessario intraprendere già nel brevissimo termine una serie di attività volte ad inglobare nelle stime correnti due novità che interessano alcuni aspetti della rilevazione OROS:

1. il nuovo scenario informativo, che vede un notevole incremento nella dimensione del campione, in seguito ai recenti cambiamenti normativi che obbligano tutte le imprese ad inviare il DM10 per via telematica;
2. l'obbligo di abbreviare i tempi di rilascio dei dati, dagli attuali 90 giorni a 60 giorni, come da Regolamento STS (1165/98).

In quanto al primo aspetto, la disponibilità di un campione che può essere considerato un quasi - universo, modifica sostanzialmente e in positivo il quadro informativo della rilevazione. Nell'ipotesi che nel prossimo futuro se ne possa disporre con regolarità (ipotesi da cui dipende in modo sostanziale la qualità dell'informazione prodotta), si debbono tenere in conto alcuni problemi che possono rendere complesso il nuovo quadro informativo:

a. la necessità di qualificare meglio la struttura del campione quasi - universo e la sua relazione con l'universo di riferimento (disponibile a t+5): poiché le stime delle variazioni si ottengono a partire da stime dei livelli *cross section*, ci si potrebbe trovare a confrontare set informativi che presentano comunque delle differenze rilevanti sulle posizioni lavorative;
b. poiché sono sempre possibili "cadute" del campione, va comunque prevista una metodologia "di salvataggio" da utilizzare quando necessario, caratterizzata da un approccio flessibile, cioè progettata in modo tale da adattarsi a contesti mutevoli e non necessariamente prevedibili.

Per quanto riguarda, invece, la riduzione dei tempi di rilascio fino a 60 giorni, si dovrà anzitutto procedere ad anticipare gli scarichi dei dati INPS rispetto ai tempi attuali (ora l'acquisizione dell'ultimo mese del trimestre di stima avviene a circa 45 giorni). Ne potrebbero conseguire drastiche ricadute sulla quantità di informazioni disponibili, soprattutto in riferimento all'ultimo mese del trimestre. Ciò renderà necessarie delle modifiche nell'attuale metodologia di stima, prevedendo imputazione o calibrazione sull'ultimo mese.

Riferimenti bibliografici

Rubin, D.B. (1976) Inference and missing data, *Biometrika* 63, 581-592.

Little R.J.A, D.B Rubin (1987) Statistical Analysis with missing data, Wiley and Sons.

Baldi C., Ceccato F., Congia M.C., Cimino E., Pacini S., Rapiti F., Tuzi D. (2004) Use of Administrative Data for Short Term Statistics on Employment, Wages and Labour Cost in Proceedings of the "*17th Roundtable on Business Survey frames*". *Essays* n. 15, Istat, Roma.

Baldi C., Falorsi P. D., Pallara A., Succi R., e Russo A. (2000), "A method for short-term estimation of labour input using current preliminary data from administrative sources having coverage errors", *Proceedings of Statistics Canada Symposium 2001.*