rivista di statistica ufficiale

n. 2-3
2007

Temi trattati

L'evoluzione dei processi produttivi per la gestione ed il controllo dei tassi e dei tempi di risposta
Fernanda Panizon, Alfredo Cirianni, Maria Carla Congia, Silvana Garozzo, Anna Rita Giorgi

The planning of Preliminary Sample: methodological aspects and an application to the Italian Monthly Retail Trade Survey
Paolo Righi, Tiziana Tuoto

RELAIS: An Open Source Toolkit for Record Linkage
Nicoletta Cibella, Marco Fortini, Monica Scannapieco, Rosalba Spina, Laura Tosco, Tiziana Tuoto

La stazionarietà locale nella stima della povertà relativa per piccole aree
Roberto Benedetti, Claudia Rinaldelli

Istat
Istituto nazionale di Statistica

rivista di statistica ufficiale

n. 2-3
2007

Temi trattati

L'evoluzione dei processi produttivi per la gestione ed il controllo dei tassi e dei tempi di risposta 5
Fernanda Panizon, Alfredo Cirianni, Maria Carla Congia, Silvana Garozzo, Anna Rita Giorgi

The planning of Preliminary Sample: methodological aspects and an application to the Italian Monthly Retail Trade Survey 33
Paolo Righi, Tiziana Tuoto

RELAIS: An Open Source Toolkit for Record Linkage 55
Nicoletta Cibella, Marco Fortini, Monica Scannapieco, Rosalba Spina, Laura Tosco, Tiziana Tuoto

La stazionarietà locale nella stima della povertà relativa per piccole aree 69
Roberto Benedetti, Claudia Rinaldelli

Istat

Direttore responsabile: Patrizia Cacioli

Coordinatore scientifico: Giulio Barcaroli

per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
c/o Carlo Deli
Istat - Via Cesare Balbo, 16 - 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 2-3/2007

Periodico quadrimestrale
ISSN 1828-1982

Registrazione presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Servizio Produzione editoriale
Via Cesare Balbo, 16 - Roma

Videoimpaginazione:
Raffaella Rose, Patrizia Balzano

Stampa:
Istat - Produzione libraria e centro stampa
Via Tuscolana 1776 - Roma
Giugno 2009 - Copie 350

L'evoluzione dei processi produttivi per la gestione ed il controllo dei tassi e dei tempi di risposta[1]

Fernanda Panizon[2], Alfredo Cirianni[3], Maria Carla Congia[4], Silvana Garozzo[5], Anna Rita Giorgi[6]

Sommario

Per i principali indicatori economici congiunturali i Regolamenti Europei richiedono un continuo miglioramento della tempestività nella diffusione. L'Istat ha aperto su questo fronte diversi filoni di ricerca finalizzati all'identificazione di strategie di stima "anticipata", in modo che sia possibile soddisfare l'esigenza della tempestività senza perdere in accuratezza delle stime, soprattutto in termini di contenimento della revisione fra stima rapida e definitiva. Gli approcci al problema sono sia di natura statistico metodologica, legati alla scelta di stimatori appropriati, sia di natura più strettamente operativa, legati quindi al processo di produzione dei dati. In questo lavoro si esaminano le soluzioni adottate nell'ambito di quattro rilevazioni congiunturali, che hanno dovuto innovarsi per rispondere ai requisiti europei di riduzione dei tempi nel rilascio dei dati.

Abstract

European Regulations continuously ask to improve the timeliness of the main short term business statistics. Istat have been working on this issue for several years, considering different strategies to obtain preliminary indicators both preserving accuracy and reliability, i.e. reducing the differences between quick preliminary and final estimates. A first approach deals with statistical methodology and imply the choice of proper estimators. A second important aspect is more related with the data production process. This paper reviews how four different Istat short term surveys have been facing the reduction of data dissemination's delay changing their production processes. The results show that very different options have been adopted to cope with the trade-off between timeliness and accuracy.

Parole chiave: statistiche congiunturali, stime anticipate, tempestività.

[1] Il lavoro è a cura di Fernanda Panizon, i singoli paragrafi sono da attribuire a: paragrafi 1 e 6 a Fernanda Panizon, paragrafo 2 ad Anna Rita Giorgi, paragrafo 3 a Silvana Garozzo, paragrafo 4 ad Alfredo Cirianni, paragrafo 5 a Maria Carla Congia.

[2] Primo Ricercatore (Istat), e-mail: panizon@istat.it.

[3] Ricercatore (Istat), e-mail: cirianni@istat.it.

[4] C.T.E.R. (Istat), e-mail: congia@istat.it.

[5] Ricercatore (Istat), e-mail: garozzo@istat.it.

[6] Ricercatore (Istat), e-mail: angiorgi@istat.it.

1. Introduzione

L'Istat, per le indagini congiunturali che fanno capo alla Direzione Centrale delle Statistiche Congiunturali (DCSC), ha dovuto rispondere negli ultimi anni alle norme dettate dai Regolamenti Europei (STS, LCI[7]) che hanno richiesto decisi incrementi nella tempestività del rilascio di dati.

Diverse indagini già a regime sono state coinvolte nel processo di progettazione di indicatori "anticipati", che, se pure provvisori o forniti in via confidenziale, potessero soddisfare i requisiti di una diffusione precoce ed affidabile, cioè in grado di cogliere in tempi stretti il segnale economico espresso dagli indicatori standard.

La fonte di informazione delle indagini di tipo economico è basata sui dati forniti da imprese o da istituzioni, che non sempre rispondono con la celerità richiesta, o non rispondono affatto, costringendo i responsabili di indagine a definire un intervallo di tempo prefissato da dedicare alla fase di raccolta dei dati, prima di passare alle fasi successive di revisione ed elaborazione e quindi alla diffusione.

Di fronte alla richiesta di una riduzione dei tempi di rilascio si può intervenire teoricamente comprimendo i tempi delle diverse fasi di "produzione del dato", ma se, ad esempio, una ottimizzazione dei tempi nelle fasi di controllo e correzione non dovrebbe comportare variazioni in negativo della qualità finale dei dati, una compressione di tempi di raccolta implica quasi sempre una riduzione dell'insieme di dati/informazioni disponibili per l'elaborazione degli indicatori economici, e quindi una qualche perdita nella qualità in termini di accuratezza e affidabilità delle stime.

Come muoversi per adeguarsi alle richieste dell'Unione Europea e al contempo garantire che sia soddisfacente la qualità dei risultati ottenuti sulla base di un sottoinsieme ridotto di informazioni?

Per risolvere il problema della maggiore tempestività, da un lato è partita la sperimentazione su metodi di stima alternativi, la cui efficienza/robustezza potesse risultare utile in una situazione di limitata disponibilità di osservazioni ad un certa data, dall'altro si è avviato un ripensamento dei processi di produzione dei dati, per eliminare alcune criticità delle diverse fasi e tentare di:

- contenere al massimo i tempi di revisione ed elaborazione,
- incrementare, ove possibile, l'insieme dei dati disponibili alla scadenza "anticipata",
- ridisegnare il flusso/schema dell'indagine per ottenere un sottoinsieme di dati "anticipati", controllato e studiato per aumentare la qualità delle stime anticipate,
- modificare i processi di produzione, nelle varie fasi di controllo, correzione, elaborazione, per ridurre al minimo lo scostamento (revisione) fra le stime provvisorie e le stime definitive.

I casi che si esaminano nel lavoro riguardano sono esempi di come il problema delle stime anticipate sia stato affrontato concretamente dal punto di vista del processo di produzione proponendo un mix di soluzioni.

Si considereranno in particolare le seguenti tipologie di intervento: la predisposizione di sub-campioni anticipati *ad hoc*, la riorganizzazione del processo di raccolta, la ridefinizione delle scadenze di alcuni punti del processo, l'innovazione negli strumenti di acquisizione dei dati.

Le indagini esaminate sono quattro e precisamente: indagine mensile sulle vendite al dettaglio, indagine trimestrale sui permessi di costruire, indagine trimestrale sul fatturato

[7] STS: Short Term Statistics; LCI: Labour Cost Index

degli altri servizi, indagine trimestrale sulle retribuzioni di fatto e costo del lavoro (Oros).

Il regolamento Europeo sulle statistiche congiunturali (CE 1165/1998) precisava per ciascuna delle citate indagini un periodo specifico per il rilascio dei dati. Per le vendite al dettaglio la scadenza dell'invio di dati era stata fissata inizialmente a due mesi dal periodo di riferimento. Per il fatturato dei servizi e le variabili dell'indagine Oros i dati dovevano essere forniti dopo tre mesi. Il Regolamento successivo (CE 1158/2005) ha invece richiesto una anticipazione della forniture dei dati ad un mese per la vendite al dettaglio e a due mesi per le altre due. Per l'indagine sui permessi di costruire invece il Regolamento del 1998 ha imposto il passaggio da una indagine a cadenza annuale ad una indagine trimestrale.

Le modalità ed i tempi di adeguamento al nuovo regime "anticipato" sono stati diversi per ciascuna indagine e sono decritti di seguito.

2. Stima anticipata dell'indice delle vendite al dettaglio

2.1 Le caratteristiche generali della rilevazione mensile sulle vendite al dettaglio

L'indagine mensile sulle vendite al dettaglio è condotta correntemente dall'Istat a partire dal mese di gennaio 1996. Il dominio di riferimento è costituito dall'universo delle imprese commerciali operanti tramite punti di vendita al minuto in sede fissa (classificate nella Divisione 52 della Ateco 2002), autorizzati alla vendita di prodotti nuovi[8].

Nel questionario inviato all'impresa (che è l'unità di rilevazione e di analisi, secondo quanto richiesto dal Regolamento STS) si chiede di indicare, per ciascun mese di riferimento, il valore delle vendite al lordo dell'IVA[9], per 15 tipologie di prodotti. Viene inoltre richiesto il numero di addetti (dipendenti più indipendenti), il numero e la superficie complessiva dei punti di vendita, il numero medio di giorni in cui i punti vendita dell'impresa sono rimasti aperti nel corso del mese di riferimento. I dati raccolti vengono elaborati dopo circa 50 giorni dalla fine del mese di riferimento; i risultati, espressi mediante numeri indice in base 2000=100, sono diffusi attraverso un comunicato stampa in media dopo 54 giorni.

Il disegno d'indagine utilizza il campionamento casuale all'interno di strati definiti da attività economica (a livello di classe), da cinque classi dimensionali in termini di addetti (1-2, 3-5, 6-9, 10-19 e >19) e tiene conto della ripartizione geografica. Il campione viene estratto dall'archivio ASIA (Archivio Statistico Imprese Attive) dell'Istat; il campione relativo all'anno 2007 è composto da circa 8.000 imprese[10].

Il sottoinsieme del campione costituito dai rispondenti mensili [11] su cui si basa il calcolo

[8] Restano fuori dal campo di osservazione le attività di riparazione di qualsiasi genere e i punti di vendita di beni usati. Sono, inoltre, escluse le attività che fanno capo al commercio ambulante e le rivendite di tabacchi e generi di monopolio. A tal proposito si osservi che, in base alla classificazione Ateco 2002, le attività che riguardano la vendita di autoveicoli e combustibili non fanno parte del commercio al dettaglio.

[9] I dati sarebbero richiesti da Eurostat al netto dell'IVA, ma motivi di opportunità legati al carico statistico per le imprese e alla continuità delle serie storiche iniziate nel 1996, la rilevazione include l'IVA. Si ritiene che questa scelta non introduca distorsioni rilevanti nel valore degli indici né nelle variazioni tendenziali.

[10] Data la strategia di campionamento, il numero di unità che fanno parte del campione varia di anno in anno a seguito della necessità di tener conto degli eventi che riguardano la composizione strutturale della popolazione di riferimento (es. cessazioni, nuove imprese, fusioni, scorpori ecc…).

[11] Secondo la terminologia comunemente utilizzata nell'ambito della rilevazione, il campione dei rispondenti mensili viene anche definito "campione effettivo" mentre il campione complessivo (ovvero composto da rispondenti e non) viene definito anche "campione teorico".

degli indici, è composto mediamente da 4.000 imprese. Al momento del rilascio delle stime tramite comunicato stampa il tasso di risposta (ottenuto rapportando il numero dei rispondenti alla numerosità del campione) è quindi di circa il 50%. I questionari pervenuti successivamente alla diffusione degli indici mensili vengono comunque controllati e registrati, ad integrazione dei database di microdati utilizzati per i controlli trasversali e longitudinali. L'ammontare di questi *late respondents* varia indicativamente fra 200 e 800 questionari al mese.

La dinamica infra-annuale delle non risposte non sembra particolarmente influenzata dalla stagionalità.

Per ridurre il numero di mancate risposte si ricorre a solleciti telefonici mirati per le imprese più grandi e quelle appartenenti agli strati per i quali è necessario disporre di un numero maggiore di risposte.

Sulla base di esigenze espresse dalla Banca Centrale Europea nel 2001 l'Eurostat ha promosso e coordinato una *taskforce* con il compito di individuare un "campione europeo anticipato" per la stima di un indice mensile del volume delle vendite al dettaglio a livello UE, calcolato a partire dalle informazioni elaborate per i singoli Stati membri con un ritardo di 30 giorni a partire dal mese di riferimento. Tale campione doveva consentire all'Italia il calcolo e il rilascio (sia pure in forma confidenziale per Eurostat) di indici mensili con un "ritardo" di 30 giorni invece che di 54, le cosiddette stime anticipate.

La metodologia prescelta prevedeva uno schema di selezione di un campione "anticipato" a livello d'intera Unione Europea, tale da garantire un errore campionario relativo di stima non superiore all'1%. Il campione che ne derivava per l'Italia non risultava automaticamente utilizzabile per l'elaborazione di stime anticipate a livello nazionale.

Di conseguenza si è scelto di individuare il campione anticipato per l'Italia (di seguito definito anche campione rapido) scegliendo un sub-campione ad hoc del campione teorico mensile, composto inizialmente da 1.929 imprese. La selezione del campione rapido è stata effettuata seguendo il criterio del campionamento bilanciato, utilizzando come variabile di bilanciamento il valore medio delle vendite registrato per i primi 10 mesi del 2002 (ossia i dati disponibili al momento della selezione di tale campione). Tale criterio ha permesso di scegliere le unità che, complessivamente, erano in grado di fornire "la maggior parte" dell'informazione relativa agli andamenti mensili.

In tal modo si è voluto identificare un nucleo scelto di rispondenti, che garantissero una base di dati "stabile", coerente e disponibile alla data anticipata. Inoltre, dal momento in cui fanno parte del campione rapido, tali imprese sono seguite con particolare cura nelle fasi di raccolta dei dati e di sollecito telefonico per i non rispondenti.

Alle imprese appartenenti al campione "rapido" è stata fornita (inizialmente solo a loro in via sperimentale) la possibilità di compilare e inoltrare il questionario della rilevazione tramite *web*, con l'accesso al sito Indata dell'Istat. L'acquisizione telematica si conciliava con la necessità di disporre tempestivamente dei dati da elaborare per il calcolo degli indici mensili. Si osserva che l'incidenza di questa modalità di risposta è rimasta abbastanza modesta nel tempo, ma ha avuto e continua ad avere effetti positivi sul grado di copertura, in quanto viene utilizzata principalmente dalla imprese di maggiori dimensioni. In una prospettiva a medio termine si cercherà di incoraggiare tutti i rispondenti (e non solo coloro che appartengono al campione rapido) all'utilizzo del canale *web* per la compilazione dei questionari.

2.2 Analisi dei tassi di copertura

Prima di descrivere le analisi effettuate per la valutazione della qualità dei risultati occorre introdurre alcuni concetti che saranno richiamati più volte nel seguito del paragrafo, il primo dei quali riguarda la definizione di copertura. Nel contesto della rilevazione mensile sulle vendite al dettaglio si parla di copertura con riferimento all'ammontare del valore delle vendite o del volume d'affari "spiegato" dalle imprese rispondenti rispetto ad un ammontare totale che può essere riferito al campione teorico oppure all'intero universo dal quale il campione è stato estratto. In alternativa lo stesso concetto può essere esteso al caso in cui anziché il valore delle vendite si considera il numero delle imprese che lo hanno prodotto. In questa seconda accezione si può indifferentemente parlare anche di tasso di risposta.

Nei primi due anni di sperimentazione (2003-2004) per il calcolo degli indici anticipati elaborati sul campione "rapido" sono state condotte alcune analisi i cui risultati permettessero di valutare la qualità degli indicatori prodotti. Una delle suddette analisi, descritta in questo paragrafo a titolo esemplificativo, è consistita nella valutazione dei tassi di copertura (in termini di numero di imprese e di volume d'affari), sia del campione effettivamente utilizzato per il calcolo dell'indice delle vendite definitivo (a circa 50 giorni) rispetto al campione teorico, sia del campione "rapido" (a circa 30 giorni) rispetto al campione definitivo. Questi tassi sono stati calcolati per il trimestre luglio-settembre 2004 (Tavola 1). Si sottolinea che tale periodo ha rappresentato la fase iniziale di riorganizzazione del processo produttivo e pertanto i tassi di risposta analizzati sono da considerarsi sperimentali e connessi ad una fase di aggiustamento del processo stesso.

Il volume d'affari è stato derivato dal registro delle imprese ASIA riferito al 2002 in cui tale variabile è nota per tutte le imprese del campione teorico. I calcoli sono stati effettuati per tipologia dei beni venduti (alimentari e non alimentari) e per 5 classi di addetti.

Tavola 1 - Tassi di copertura del campione effettivo rispetto al campione teorico e del campione totale rispetto al campione anticipato in termini di imprese e di volume d'affari, per classe di addetti e settore merceologico *(media luglio-settembre 2004)*

Rapporto percentuale sul numero di imprese

	Campione effettivo VS teorico						Campione anticipato VS effettivo					
	Classe di addetti											
Settore merceologico	**1-2**	**3-5**	**6-9**	**10-19**	**>19**	**Totale**	**1-2**	**3-5**	**6-9**	**10-19**	**>19**	**Totale**
Alimentari	32,6	50,8	50,6	60,9	58,1	46,6	28,7	30,2	30,6	44,0	76,6	45,9
Non alimentari	36,6	52,6	56,1	72,7	62,6	46,4	12,8	28,7	35,7	43,8	79,3	31,9
Totale	35,6	52,1	53,4	68,0	60,3	46,5	16,4	29,1	33,3	43,9	78,0	36,5

Rapporto percentuale sul volume d'affari

	Campione effettivo VS teorico						Campione anticipato VS effettivo					
	Classe di addetti											
Settore merceologico	**1-2**	**3-5**	**6-9**	**10-19**	**>19**	**Totale**	**1-2**	**3-5**	**6-9**	**10-19**	**>19**	**Totale**
Alimentari	68,7	56,1	48,9	63,7	78,0	76,9	79,2	37,0	32,6	44,3	87,9	85,9
Non alimentari	41,5	58,3	64,0	68,9	59,1	59,4	24,5	37,1	44,7	50,8	87,3	81,7
Totale	51,6	57,6	55,9	66,8	70,5	69,6	51,5	37,1	39,0	48,3	87,7	84,4

Riguardo al primo tasso di copertura si osserva che:

- nei tre mesi considerati, in media, il tasso di copertura del campione effettivo rispetto a quello teorico è stato del 46,5% in termini di numero di imprese e del 69,6% in quanto a volume d'affari. La copertura in termini di volume d'affari è sensibilmente più elevata per le imprese che vendono in prevalenza prodotti alimentari (76,9%), rispetto a quelle che vendono in prevalenza non alimentari (59,4%).
- Nel complesso nel trimestre considerato i tassi di copertura crescono al crescere della classe dimensionale delle imprese. In termini di numero di imprese, la copertura più bassa si registra con riferimento alle imprese più piccole (35,6%). Lo strato più coperto è quello delle imprese non alimentari e con 10-19 addetti (72,7%). In termini di volume d'affari, la copertura è, in media, superiore al 50%.

Per quel che riguarda il confronto tra il momento della stima anticipata e quello su cui si basa la successiva stima definitiva emergono i seguenti elementi:

- in media, nel trimestre considerato, il tasso di copertura del campione "rapido" rispetto a quello effettivo è stato del 36,5% in termini di numero di imprese e dell'84,4% rispetto al volume d'affari. Per tipologia di prodotti il tasso di copertura sulle imprese si diversifica (45,9% per i prodotti alimentari, contro il 31,9% dei non alimentari), mentre è più omogeneo sul volume d'affari (rispettivamente, 85,9% e 81,7%).
- La relazione fra copertura e dimensione dell'impresa è molto più evidente in termini di numero d'imprese piuttosto che di volume d'affari. Nel primo caso i tassi di copertura per le cinque classi di addetti sono pari a 16,4%, 29,1%, 33,3%, 43,9% e 78,0%, mentre in termini di volume d'affari una chiara tendenza alla crescita della copertura si registra solo a partire dalla classe dimensionale 6-9.
- Se l'elevata copertura (87,7%) in termini di volume d'affari per le imprese più grandi può contribuire ad una maggiore precisione delle stime anticipate, in alcuni strati si evidenziano situazioni più problematiche (ad esempio il tasso di copertura è sempre inferiore al 50% per le imprese non alimentari con meno di 10 addetti ed è di appena il 24,5% per le imprese più piccole).

In sintesi, da questa analisi sono confermate due particolari esigenze di cui tenere conto:

- dato che il campione "rapido" è di tipo bilanciato, occorre verificare periodicamente che le condizioni di bilanciamento rispetto al campione teorico siano di fatto rispettate. I mutamenti strutturali che riguardano le imprese in questione, infatti, potrebbero far variare il peso relativo (in termini di volume d'affari e/o valore delle vendite) delle imprese all'interno dei vari strati.
- la necessità di rafforzare le azioni mirate di sollecito per incrementare il tasso di risposta in particolare per gli strati in cui tale tasso risulta più basso.

Sulla base di questi risultati preliminari, e soprattutto considerando la forte disomogeneità fra gli strati, per garantire comunque uno standard accettabile dei livelli di qualità delle stime anticipate, si è presa in considerazione l'idea di modificare l'approccio relativo all'imputazione delle mancate risposte.

Per quanto riguarda il calcolo degli indici anticipati, la metodologia utilizzata prevede che tutte le risposte mancanti del campione siano stimate congiuntamente per ricostruire l'insieme di dati complessivo relativo alle le 8.000 imprese (appartenenti al campione anticipato e non) su cui si calcola l'indice. In un primo momento venivano considerate mancanti (e quindi imputate) sia quelle del campione rapido non pervenute entro i 30 giorni, sia tutte le imprese non appartenenti al sub-campione rapido. In un secondo tempo si è ritenuto più opportuno integrare le osservazioni pervenute dal campione rapido in senso stretto con quelle comunque disponibili al momento della stima anticipata. Quindi. le

informazioni pervenute da imprese esterne al sub-campione non venivano più imputate, ma utilizzate per la stima dei parametri. Le strategie per risolvere il problema delle mancate risposte quindi si sono evolute nel tempo, cercando di massimizzare l'informazione disponibile al momento dell'imputazione e del calcolo degli indici anticipati.

2.3 Modalità e tempi di risposta del sub-campione "rapido"

Un altro aspetto rilevante, per testare l'adeguatezza del campione dei rispondenti anticipati rispetto quello relativo al totale dei rispondenti, è dato dall'andamento dei tassi di copertura, in termini di imprese e/o di fatturato nel tempo.

Considerando il periodo gennaio 2004 - ottobre 2007 le Figure 1 e 2 mettono in evidenza una scarso effetto della stagionalità sui tassi di copertura.

Figura 1 - Tassi di risposta a 30 e a 54 giorni *(Periodo gennaio 2004-ottobre 2007)*

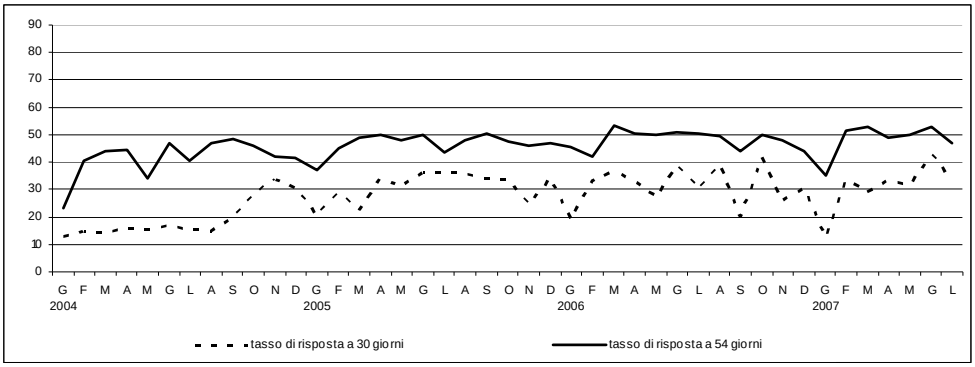
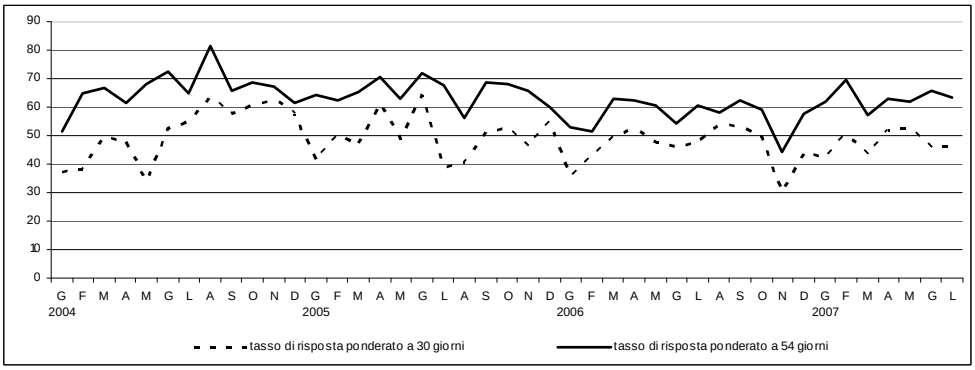


Figura 2 - Tassi di copertura a 30 giorni e a 54 giorni per l'indagine sulle vendite al dettaglio *(Periodo gennaio 2004-ottobre 2007)*



Per il calcolo dei tassi di copertura si è utilizzato, per ogni impresa appartenente al campione teorico: 1) il fatturato mensile effettivamente dichiarato se l'impresa è risultata rispondente in un dato mese; 2) la stima di tale fatturato, calcolata dal programma di imputazione delle mancate risposte usato per il calcolo degli indici delle vendite, nel caso l'impresa non sia risultata rispondente.

Come si osserva dai grafici, l'andamento dei tassi di risposta e di copertura in termini di

fatturato è abbastanza stabile nel tempo. Per quanto riguarda i tassi di risposta si osserva una certa regolarità con riferimento ai periodi in cui tali tassi sono più bassi: si tratta dei mesi di gennaio, in cui viene aggiornata la lista delle imprese del campione, e dei mesi estivi, in cui si concentra la chiusura delle attività commerciali.

Si ricorda inoltre che nella prima parte del periodo considerato i valori più bassi sia dei tassi di risposta, sia dei tassi di copertura, sono da attribuirsi alla fase di riorganizzazione del processo produttivo e all'introduzione di innovazioni che permettessero di disporre dei dati mensili in modo più tempestivo.

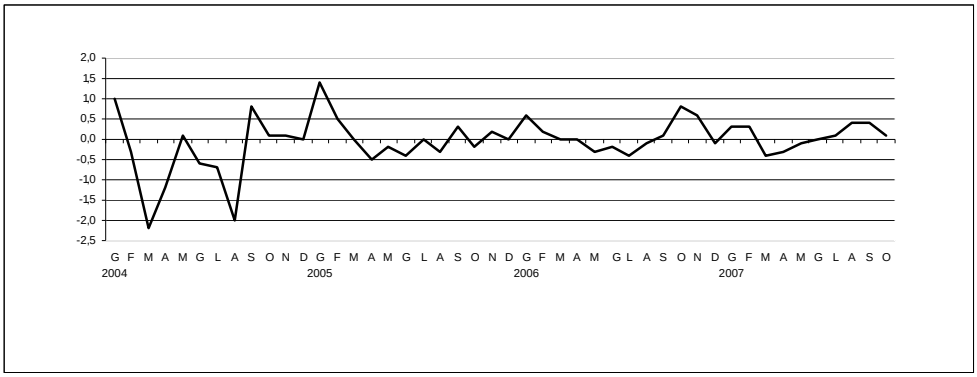
2.4 Analisi degli errori di stima

La revisione, cioè lo scostamento tra gli indici anticipati del valore del totale delle vendite e i corrispondenti indici definitivi, è il più importante indicatore della "qualità" della stima anticipata.

Tale differenza, infatti, fornisce informazioni su quanto la stima anticipata della variazione del valore mensile delle vendite approssimi l'analoga stima definitiva. Nella Figura 3 è rappresentato l'andamento di tali differenze nel periodo compreso tra gennaio 2004 e ottobre 2007.

Dalla Figura 3 risulta evidente che l'ampiezza degli scostamenti si è gradualmente ridotta: nel 2004 risultano ancora evidenti gli effetti dovuti alla riorganizzazione del processo, tuttavia nei periodi successivi le stime anticipate risultano molto vicine a quelle definitive. Un fattore di successo deriva senza dubbio dall'utilizzo crescente delle risposte pervenute in tempo utile, relative ad imprese non appartenenti al campione anticipato, che, come si è detto, vengono comunque considerate nel calcolo degli indici anticipati.

Figura 3 - Differenza tra le stime mensili delle vendite al dettaglio definitive rispetto a quelle anticipate. *(Periodo gennaio 2004-ottobre 2007)*



In sintesi, l'introduzione dell'elaborazione di stime anticipate per il valore mensile delle vendite al dettaglio ha comportato una profonda riorganizzazione del processo produttivo, che è stato inoltre integrato con fasi operative innovative (ad esempio la gestione dell'acquisizione dei dati via *web*) grazie alle quali è possibile, già dal 2003, rispettare la tempestività richiesta dai Regolamenti dell'Eurostat.

Allo stato attuale l'intero processo produttivo appare consolidato e le innovazioni introdotte per il calcolo degli indicatori anticipati hanno apportato dei vantaggi anche per le stime definitive. Tra questi vantaggi va senz'altro ricordato l'incremento del tasso di risposta.

3. La rilevazione dei permessi di costruire

3.1 Caratteristiche della rilevazione

Attraverso la rete degli 8.100 comuni italiani la rilevazione dei permessi di costruire acquisisce informazioni sulle principali caratteristiche dei nuovi fabbricati, residenziali e non residenziali, e degli ampliamenti dei fabbricati preesistenti, la cui costruzione sia stata autorizzata dal competente ufficio comunale. La rilevazione esiste dall'anno 1934.

Secondo la normativa vigente la realizzazione di un'opera edilizia deve essere autorizzata attraverso la presentazione all'Ufficio comunale del cosiddetto "Permesso di costruire" o della "Denuncia di inizio attività" (DIA), a seconda che il caso rientri nell'ambito dell'uno o dell'altro titolo abilitativo. Il richiedente deve compilare obbligatoriamente anche il questionario (modello Istat/AE), fornito dagli uffici comunali, insieme alla documentazione necessaria per il rilascio del permesso (o DIA).

Il modello di rilevazione, oltre al riquadro contenente i dati identificativi del comune, è composto da tre parti. La prima raccoglie le notizie generali sull'opera (tempi previsti per la realizzazione, ubicazione, natura dell'opera, destinazione d'uso, concessionario, finanziamento, regime dei suoli, impianto termico, struttura portante). La seconda è relativa ai soli fabbricati residenziali e contiene informazioni sui piani, sul volume, sulla superficie, sul numero di abitazioni, sul numero di stanze per abitazione e le classi di superficie utile abitabile. La terza comprende le notizie relative ai soli fabbricati non residenziali e indaga sulla dimensione del fabbricato, sulla parte ad uso abitativo, sulla destinazione economica e sulla tipologia dell'opera.

Il mese di riferimento della rilevazione coincide con il mese in cui avviene il ritiro (da parte del richiedente) del generico permesso di costruire o decorre la validità della denuncia di inizio attività.

Gli uffici comunali, mensilmente, hanno il compito di raccogliere i questionari Istat relativi alle opere edilizie per le quali siano stati ritirati i "Permessi di costruire", o decorra la validità delle "Denunce di inizio attività", controllare l'esattezza delle informazioni che vi sono riportate, compilarne il riquadro riservato al Comune ed inviarli, tramite posta alle Camere di commercio, entro il quinto giorno del mese successivo. Queste, in qualità di enti intermedi di rilevazione, raccolgono i modelli pervenuti nel mese dai comuni della provincia di propria competenza, e li inviano, a loro volta, all'Istat di Roma.

In caso di assenza di permessi di costruire ritirati e di DIA in corso di validità nel mese di riferimento, il Comune deve comunque inviare, sempre attraverso la Camera di commercio, una segnalazione di attività edilizia nulla, (modello ISTAT/AE/Neg.), con cui si rileva l'assenza del fenomeno nel mese di riferimento. Se il Comune non invia alcun questionario o segnalazione di attività edilizia nulla, è considerato non rispondente nel mese di riferimento. Per ridurre il tasso di mancata risposta, ogni semestre l'Istat invia alle CCIAA, tramite e-mail, il prospetto dello stato della collaborazione dei comuni per mese. Le CCIAA effettuano le operazioni di sollecito presso i Comuni non rispondenti.

La rilevazione tradizionale dei permessi di costruire pertanto è di tipo totale a cadenza mensile. Ogni mese vengono raccolti ed elaborati circa 115.000 record (media su 7 anni), di cui 83.000 modelli e 32.000 segnalazioni di attività edilizia nulla (assenza del fenomeno nel mese).

La diffusione dei dati definitivi ha periodicità annuale, con un "ritardo" rispetto al periodo di riferimento pari a 18 mesi (ad esempio, i dati del 2005 escono a giugno del 2007).

Le esigenze, imposte dal Regolamento del Consiglio Europeo sulle statistiche congiunturali (CE n.1165/98), riguardanti la produzione di alcuni indicatori sull'attività edilizia, hanno

imposto un rilascio di dati trimestrale entro 90 giorni dal periodo di riferimento.

Il processo di rilevazione tradizionale dei permessi di costruire, pur raccogliendo le informazioni necessarie alla produzione degli indicatori richiesti, non consentiva, così com'era, di rispettare le scadenze comunitarie. Sono quindi state necessarie profonde innovazioni, mirate a ridurre i tempi di produzione di tali indicatori.

3.2 Le esigenze di tempestività dettate dai nuovi Regolamenti Europei e la nuova rilevazione "rapida"

Al fine di poter effettuare, entro 90 giorni dal periodo di riferimento, le stime congiunturali trimestrali, si è introdotta nel processo di produzione una indagine campionaria. E' stato individuato un campione di 814 comuni, che mensilmente trasmettono i modelli raccolti direttamente all'Istat (senza passare dalle Camere di commercio) utilizzando un'apposita casella postale, dedicata alla rilevazione rapida, in modo da ridurre i tempi di raccolta.

Il campione di 814 comuni comprende 160 comuni capoluogo e non capoluogo con più di 50.000 abitanti, che costituiscono lo strato autorappresentativo, mentre i restanti 654 sono stratificati per ripartizione geografica e classi di popolazione in 20 gruppi.

Tavola. 2 - Numero di comuni campione dell'indagine rapida dei permessi di costruire per ripartizione geografica e classi di ampiezza demografica

Ampiezza demografica	Ripartizione geografica				
	Centro	Nord-Est	Nord-Ovest	Sud-Isole	Totale
Fino a 3000	15	28	79	29	151
Da 3001 a 7000	20	45	57	42	164
Da 7001 a 13000	22	54	37	38	151
Da 13001 a 25000	14	28	32	32	106
Oltre 25000	20	13	21	28	82
Totale	91	168	226	169	654

Oltre alla costruzione di un disegno di campionamento per la selezione di un adeguato campione di comuni è stato necessario mettere a punto nuove metodologie per la stima degli indicatori congiunturali.

Il modello di rilevazione rimane lo stesso, anche se gli indicatori trimestrali richiesti da Eurostat a 90 giorni riguardano solo una parte delle informazioni rilevate mensilmente e sono i seguenti:

Numero Totale Abitazioni in nuovi fabbricati residenziali, divise per:

- Abitazioni in nuovi fabbricati residenziali con 1 abitazione
- Abitazioni in nuovi fabbricati residenziali con 2 abitazioni ed oltre

Superficie dei fabbricati residenziali, divisa per:

- Superficie utile abitabile in nuovi fabbricati residenziali con 1 abitazione
- Superficie utile abitabile in nuovi fabbricati residenziali con 2 abitazioni ed oltre
- Superficie totale dei fabbricati per collettività

Superficie totale dei fabbricati non residenziali, divisa per:

- Superficie totale dei fabbricati per Uffici
- Superficie totale degli altri fabbricati non residenziali.

La rilevazione "rapida" risulta quindi un sottoprocesso della tradizionale rilevazione dei permessi di costruire; da un lato prende in considerazione un campione ristretto di comuni, dall'altro considera soltanto le informazioni necessarie al calcolo degli indicatori

congiunturali richiesti dal regolamento. I questionari utilizzati sono gli stessi, che successivamente, vengono lavorati nel processo della rilevazione tradizionale, insieme a quelli pervenuti da tutti gli altri comuni non campionari.

Tavola 3 - Le fasi del processo produttivo dell'indagine dei permessi di costruire

Rilevazione totale	Rilevazione campionaria
Raccolta dei questionari presso i comuni e relative CCIAA di competenza	Raccolta dei questionari presso i comuni (invio presso una casella postale dedicata)
Revisione manuale dei questionari cartacei dagli esperti Istat	
Registrazione dei questionari esterna *in service*	Revisione e contestuale registrazione "rapida" interna autocontrollata delle sole variabili utili alla stima degli indicatori congiunturali, con eventuale ritorno presso il comune
Acquisizione dei dati registrati *in service* e correzione del file di microdati	
Correzione automatica	
Correzione interattiva	
Integrazione delle mancate risposte totali ai fini dell'elaborazione annuale	Integrazione delle mancate risposte totali con metodo "innovativo"
Elaborazione dei risultati	Elaborazione dei risultati
Pubblicazione annuale dati strutturali	Produzione ed invio trimestrale ad Eurostat degli indicatori congiunturali

I questionari autocompilati vengono raccolti mensilmente, sia presso i comuni campione della rilevazione rapida, sia presso i comuni non campione. Ma, mentre i primi li inviano direttamente all'Istat di Roma, i restanti comuni li inviano alle Camere di commercio. Queste, in qualità di enti intermedi di rilevazione, raccolgono i modelli pervenuti nel mese e li inviano, a loro volta, all'Istat di Roma. Questa procedura parallela permette, da un lato di velocizzare la raccolta dei questionari dei comuni campione e dall'altro di avvalersi della collaborazione degli enti provinciali per razionalizzare la raccolta dei modelli dei comuni non campione e per gestire in maniera decentrata la fase dei solleciti ai comuni non rispondenti.

I questionari, pervenuti alla casella postale "dedicata" agli 814 comuni campione, vengono sottoposti ad una lavorazione rapida, che prevede la registrazione interna delle sole variabili del modello che sono utili al calcolo degli indicatori congiunturali.

La registrazione consente di effettuare un controllo preliminare a livello di singolo questionario: vengono localizzati e corretti a video i valori fuori dominio, i valori anomali, le incompatibilità e le mancate risposte parziali.

In presenza di mancata risposta totale (a livello comunale) non si imputano i microdati, come nel caso della rilevazione totale, bensì i dati aggregati per comune. Si distingue l'insieme dei 160 comuni autorappresentativi (AR) da quello costituito dai restanti comuni.

Per quanto riguarda il sottoinsieme dei 160 comuni AR, l'imputazione mensile sui dati aggregati a livello comunale, avviene attraverso il valore medio che le variabili (abitazioni e superficie) assumono nello stesso comune nei 12 mesi precedenti in cui ha risposto. In caso di mancata risposta totale, quando cioè il comune non ha mai risposto nei 12 mesi precedenti, non viene imputato alcun valore.

Per quanto riguarda l'insieme dei restanti 654 comuni campione, si applica, invece, una metodologia di stima molto più complessa. Dato che al momento del calcolo delle stime, la copertura del campione non è completa, mentre al contrario sono disponibili informazioni su alcune unità non incluse nel campione, si è ritenuto opportuno sfruttare anche le informazioni pervenute in tempo utile dagli altri 7286 comuni non campionari.

A partire dai dati aggregati per i 7940 comuni (di cui 654 campionari e 7286 non

campionari), si stimano il numero di abitazioni e la superficie totale non residenziale.

Attualmente viene utilizzato lo stimatore cosiddetto FABI, ma è in via di sperimentazione uno stimatore alternativo del tipo *shrinkage*. La differenza sostanziale tra i due stimatori sta nel fatto che il primo combina le informazioni campionarie e non campionarie per produrre un'unica stima, mentre il secondo prevede il calcolo di due stime indipendenti, una effettuata a partire dai soli rispondenti campionari e l'altra utilizzando soltanto i rispondenti non campionari; le due stime vengono poi combinate in modo ottimale allo scopo di minimizzare la varianza.

3.3 Principali risultati

Dal 2003 i dati trimestrali richiesti da Eurostat sono stati inviati nei tempi richiesti, come stima provvisoria a 90 giorni e come definitiva a 180. Una prima valutazione dei risultati può essere effettuata considerando le differenze fra il totale annuo delle variabili di interesse stimate dall'indagine campionaria e i dati annui, calcolati a 18 mesi dalla fine dell'anno di riferimento sulla base della rilevazione censuaria.

Nel 2003 i valori derivanti dalla stima rapida campionaria, forniscono un livello sovrastimato per le tre variabili considerate, mentre negli anni successivi la differenza cambia di segno. I risultati possono essere ritenuti soddisfacenti, rispetto all'obiettivo originario di fornire una stima utilizzabile nell'ambito del calcolo degli aggregati europei, in quanto i livelli di tempestività raggiunti coesistono con una qualità accettabile complessiva dei dati totali forniti.

Tavola 4 - Stime anticipate a 180 giorni, stime annuali e revisione in valore assoluto per le principali variabili dell'indagine dei permessi di costruire *(anni 2003-2005)*

	Dato congiunturale a 180 giorni (a)	Dato annuale a 18 mesi (b)	Revisione (b-a)/a*100
			Numero totale abitazioni
2003	240.795	229.526	-4,7
2004	266.303	268.385	0,8
2005	271.527	278.602	2,6
			Superficie utile abitabile
2003	18.524.865	17.563.337	-5,2
2004	19.671.758	19.898.618	1,2
2005	20.045.379	20.479.027	2,2
			Superficie non residenziale
2003	28.864.623	28.358.843	-1,8
2004	27.162.568	29.231.857	7,6
2005	23.971.685	25.137.913	4,9

Nelle Figure 4 e 5 si presentano le variazioni tendenziali delle stime trimestrali a 90 e a 180, con relative revisioni, delle due variabili riferite alle superfici residenziali e non residenziali in metri quadrati (la variabile abitazioni è caratterizzata dallo stesso andamento della superficie residenziale).

Figura 4 - Variazioni tendenziali trimestrali dell'indagine dei permessi di costruire stimate a 90 e a 180 giorni e relativa revisione per la superficie utile abitabile *(anni 2005-2007)*

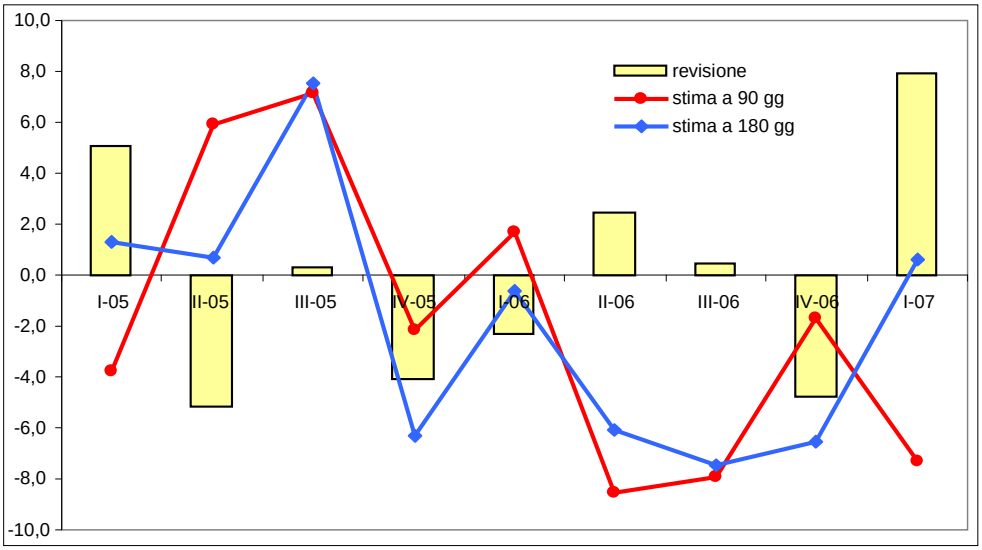
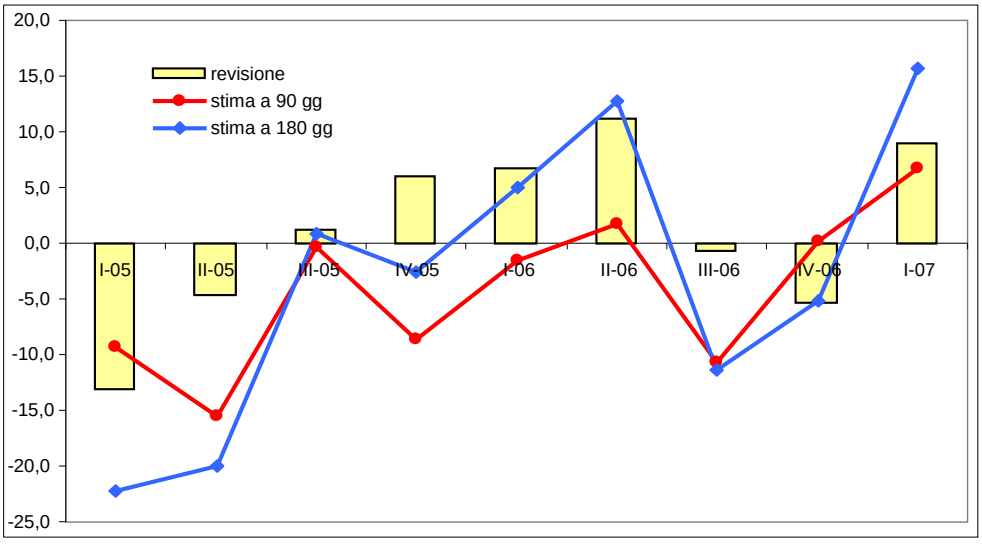


Figura 5 - Variazioni tendenziali trimestrali dell'indagine rapida dei permessi di costruire stimate a 90 e a 180 giorni e relativa revisione per la superficie non residenziale *(anni 2005-2007)*



Emerge con chiarezza che le revisioni delle variazioni tendenziali sono a tutt'oggi abbastanza rilevanti e, nonostante gli sforzi in termini di miglioramento nella raccolta dei dati e di sperimentazione di nuove procedure di stima, per la superficie non residenziale in alcuni trimestri superano i 10 punti percentuali.

L'ampiezza delle revisioni che caratterizzano la stima "rapida" costituisce un ostacolo rispetto alla diffusione dei dati a livello nazionale, in quanto è possibile che gli utilizzatori considerino tale precisione insufficiente. Le attività di approfondimento delle problematiche relative alla procedura di stima stanno quindi proseguendo al fine di giungere a metodologie che ottimizzino i risultati ottenibili dai dati (campionari e non) disponibili alle scadenze prefissate.

4. Le indagini trimestrali sul fatturato degli Altri Servizi

4.1 Caratteristiche delle indagini congiunturali sul fatturato degli altri servizi

L'allegato D del Regolamento STS relativo agli Altri Servizi stabilisce che vengano trasmessi trimestralmente ad Eurostat statistiche sul fatturato delle imprese in questo settore. L'Istat produce ad oggi una parte degli indicatori, conducendo indagini dirette, che hanno in comune sia l'unità di rilevazione (l'impresa), sia le variabili rilevate, cioè il fatturato del trimestre ed il numero di addetti (come variabile di controllo e di stratificazione).

Tutte le indagini sono di carattere campionario ed usano come universo di riferimento l'archivio ASIA più recente. ASIA contiene informazioni sulle imprese attive (denominazione, indirizzo, volume d'affari, numero di addetti, etc). Le modalità di estrazione dei campioni e la loro numerosità dipendono invece della peculiari caratteristiche del settore economico indagato e sono state studiate per garantire stime più affidabili nel contesto considerato (Tavola 5).

Tavola 5 - Numerosità campionarie delle indagini sul fatturato degli Altri Servizi

Codice Ateco	Settore attività economica	Imprese (Asia 2004)	Campione 2007	Tasso di campio-namento
51	Ingrosso e intermediari del commercio	416.317	7.878	1,89
50.2	Riparazione autoveicoli	91.898	2.734	2,98
61.1 - 61.2	Trasporti marittimi e fluviali	1.489	295	19,81
62.1	Trasporti aerei	281	78	27,77
64.1	Servizi postali	1.788	141	7,89
64.2	Telecomunicazioni	1.878	204	10,86
72	Informatica	89.755	1.829	2,04

Le prime indagini ad essere implementate (1999) sono state quelle relative ai settori cosiddetti oligopolistici: *trasporti aerei, trasporti navali, telecomunicazioni e servizi postali*. Tali settori comprendono un numero contenuto di imprese; al loro interno un numero ristretto di grandi imprese realizza una larga quota del fatturato complessivo del settore. Il disegno campionario adottato pertanto è tipo *cut off*, per cui sono state selezionate le imprese più grandi, fino a coprire la quota più rilevante del fatturato. Data la numerosità campionaria limitata, il tasso di risposta può essere tenuto a livelli molto elevati tramite l'utilizzo di solleciti telefonici mirati.

I settori appartenenti ai gruppi *ingrosso* ed *informatica* sono invece caratterizzati dalla presenza di un elevato numero di imprese, di diversa dimensione e con livelli di fatturato eterogenei. Il disegno di campionamento adottato è di tipo casuale stratificato. Il gruppo *manutenzione e riparazione di autoveicoli* è costituito in prevalenza da piccole imprese che operano in mercato concorrenziale. Il disegno di campionamento adottato in questo caso è quello bilanciato ragionato di tipo stratificato.

Tutte le indagini sono caratterizzate da una medesima tecnica di indagine: le imprese campione vengono contattate dall'Istat con l'invio di una lettera (all'indirizzo disponibile nell'archivio ASIA) e di un questionario corredato dalle opportune istruzioni. Le imprese nuove entrate devono rispondere anche ad alcuni dati retrospettivi, mentre le imprese che hanno già risposto in occasioni precedenti di indagine devono fornire soltanto i dati (di fatturato e occupati) dell'ultimo trimestre. I questionari cartacei pervenuti per posta o via fax vengono registrati con *data entry* tradizionale all'interno dell'Istituto. A partire da aprile 2006, tutte le imprese possono anche fornire i dati via *web*, tramite l'accesso al sito Istat di Indata. I dati pervenuti via *web* vengono memorizzati direttamente nel data base.

Le verifica dei dati "anomali" viene demandata alla fase successiva di revisione "statistica", effettuata dagli esperti di settore, che analizzano le imprese che presentano variazioni di fatturato superiori ad una soglia percentuale critica.

4.2 Diffusione dei dati e incremento della tempestività

Per diversi anni la diffusione trimestrale dei dati è avvenuta a 90 giorni, con la fornitura di dati ad Eurostat, con il comunicato stampa (disponibile sul sito *web* dell'Istat) ed con l'aggiornamento della banca dati ConIstat. Le stime fornite a 90 giorni sono "provvisorie", mentre a 180 giorni vengono rese "definitive".

Il più recente Regolamento comunitario (n. 1502/2006) ha stabilito che la fornitura dei dati ad Eurostat sia anticipata a 60 giorni dalla fine del trimestre di riferimento. La richiesta di *riduzione dei tempi di rilascio da 90 a 60 giorni***,** oltre a impegnare nella sperimentazione di stimatori alternativi, ha comportato la necessità di contrarre i tempi di acquisizione, controllo ed elaborazione dei dati. Gli strumenti identificati per il miglioramento della tempestività sono stati fondamentalmente due: l'utilizzo del web nella fase di acquisizione e una *riorganizzazione delle fasi* e dei tempi dell'intero processo di produzione degli indici, allo scopo di ottenere tassi di risposta e di copertura soddisfacenti al momento della stima anticipata.

4.3 La risposta via web

La modalità di risposta via internet è stata introdotta dal primo trimestre 2006 e, grazie agli aggiustamenti e miglioramenti apportati nel tempo, sta ottenendo risultati soddisfacenti ed incoraggianti.

L'impatto positivo del *web* sulla tempestività si può cogliere esaminando le curve dei ritorni, cioè confrontando le date di risposta via fax e via *web*. Tracciando la distribuzione percentuale cumulata dei "rientri" (Figura 6) per le due modalità fax e *web*: si osserva ad esempio per il settore *informatica*, un anticipo apprezzabile dei tempi di risposta del mezzo internet rispetto al fax, nella fase di raccolta del quarto trimestre 2006.

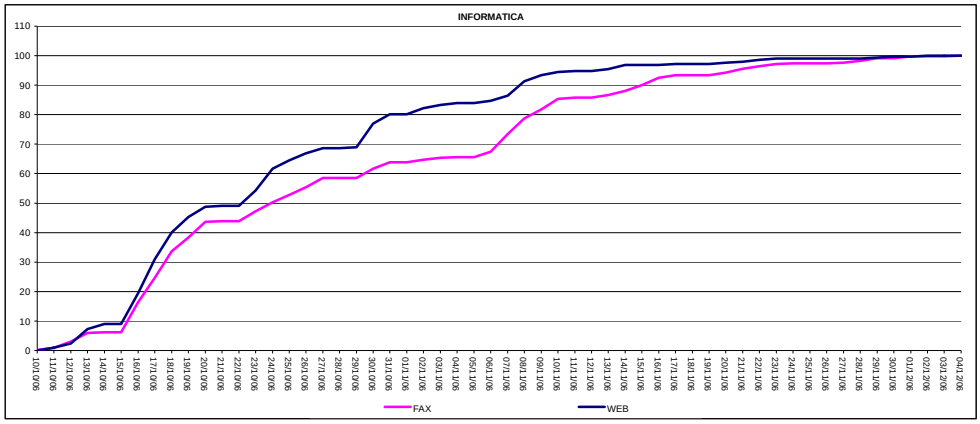
La modalità *web* ha dato da subito buoni risultati in termini di tassi di risposta, non solo per i settori tecnologicamente più avanzati come quello dell'*informatica*, ma anche per il settore delle *riparazioni di autoveicoli*, caratterizzato da imprese di piccole dimensioni in cui l'utilizzo del computer e di internet può verosimilmente essere considerato più raro (Tavola 6).

Tavola 6 - Tassi di risposta via web per la rilevazione sul fatturato degli Altri Servizi *(anni 2006-2007)*

Trimestre	Trasporti, servizi postali e telecomunicazioni	Ingrosso	Informatica	Riparazione autoveicoli
	% Rispondenti web			
Aprile - giugno 2006	15,6	17,4	29,6	9,4
Luglio - settembre 2006	12,8	19,3	29,6	11,2
Ottobre - dicembre 2006	14,4	23,9	33,7	14,1
Gennaio – marzo 2007	16,9	23,9	34,1	14,6
Aprile – giugno 2007	15,9	26,4	38,8	17,5

La sollecitazione alla risposta telematica (nella lettera inviata alle imprese all'inizio del trimestre) è proseguita nei periodi successivi, determinando un progressivo incremento delle percentuali di rispondenti *web,* soprattutto per l'*ingrosso* (dal 17,4 al 26,4 per cento) e l'*informatica* (che partendo con un già elevato 29,4 arriva al 38,8 per cento). Il settore della *riparazione di autoveicoli*, che inizialmente ha registrato percentuali modeste (9,4) di rispondenti *web*, si è velocemente portato al 17,5 per cento. Più stabile invece il grado di preferenza per internet evidenziato dai *settori oligopolistici*.

Figura 6 - Quote percentuali di arrivi via fax e via web - Indagine sull'*informatica* *(arrivi del IV trimestre 2006 riferiti a dati del III trimestre)*



L'utilizzo del *web data capturing* e il globale incremento di tale modalità di risposta, ha comportato un miglioramento della qualità dei dati in termini di tempestività della fase di acquisizione, sia perché i rispondenti via *web* sono quelli che rispondono per primi, sia perché i dati forniti vengono resi immediatamente disponibili nel database, senza passare dalla fase di data entry.

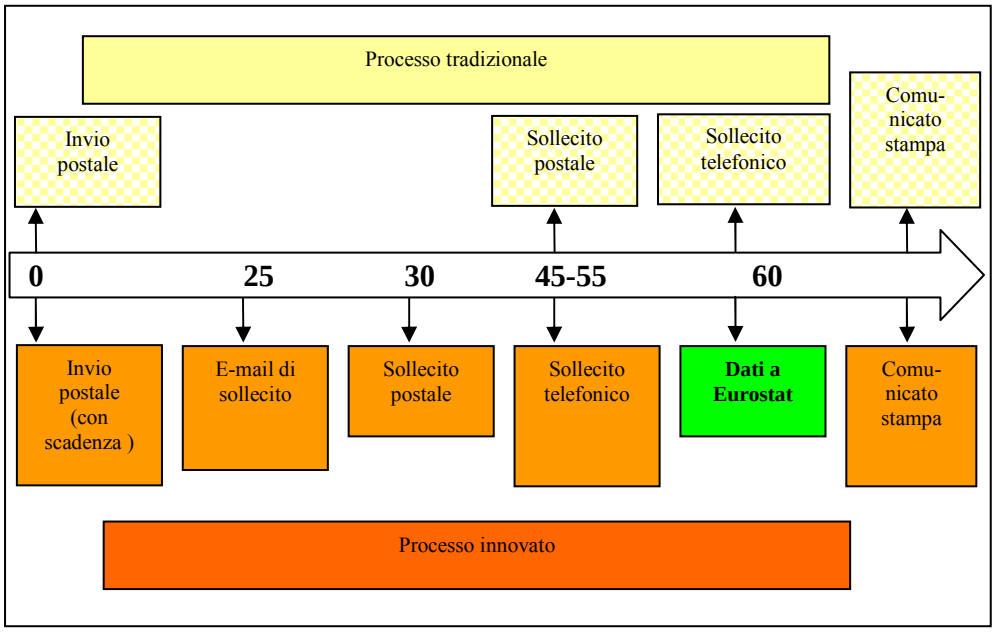
4.4 La riorganizzazione delle fasi

Per quanto riguarda la riorganizzazione del processo di produzione dei dati, si sono cercate soluzioni che consentissero aumentare i tassi di copertura al momento della stima anticipata.

Considerando la rilevazione condotta con riferimento ai dati del terzo trimestre del 2006, in previsione del rilascio delle stime anticipate a fine novembre, la spedizione dei questionari postali è avvenuta, come di consueto alla fine del terzo trimestre, e le imprese sono state invitate a rispondere entro una precisa data di scadenza (25 ottobre, equivalente a 25 giorni dalla fine del trimestre). Appena qualche giorno dopo la scadenza è stato effettuato, per la prima volta, un

sollecito tramite e-mail al sottoinsieme di imprese che avevano fornito in precedenza il proprio indirizzo di posta elettronica. Il sollecito postale alle imprese non rispondenti, solitamente previsto a metà del trimestre di rilevazione (45 giorni) è stato anticipato a 30 giorni, anche in questo caso imponendo una scadenza per la risposta (entro il 46-esimo giorno).

Figura 7 - Innovazioni introdotte nella tempistica del processo di raccolta dati per le indagini sul fatturato degli Altri Servizi al fine di anticipare le stime a 60 giorni



Nel frattempo si sono preparate le liste per i solleciti telefonici "mirati" alle imprese più importanti (con più elevati livelli di fatturato) o alle imprese degli strati campionari meno numerosi o che presentano tassi di copertura ancora insufficienti. I solleciti telefonici sono stati intensificati nell'intervallo tra 45 e 55 giorni, in modo da aumentare il numero di rispondenti in prossimità del momento di elaborazione degli indici.

Le modifiche introdotte nel processo produttivo sono piuttosto rilevanti, come evidenziato nella Figura 7. In sintesi, lo schema operativo, precedente all'introduzione delle stime anticipate a 60 giorni, non prevedeva solleciti di alcun genere prima dei 60 giorni (sollecito postale). Per produrre dati provvisori da inviare ad Eurostat 60 giorni nel processo attuale sono stati introdotti solleciti "anticipati" a 25 giorni (per posta elettronica), a 30 giorni (postale) e a 45 giorni (telefonico).

4.5 I principali risultati

Alla fine di novembre 2006 sono state rilasciate per la prima volta all'Eurostat, in forma confidenziale, le stime anticipate a 60 giorni degli indicatori trimestrali relativi al terzo trimestre del 2006.

Il risultato complessivo del ridisegno di processo, che nel corso del trimestre considerato ha visto un anticipo delle usuali fasi di sollecito postale e telefonico e un utilizzo incrementale della modalità *web* di risposta, può essere colto considerando i tassi di copertura misurata in termini di fatturato pervenuto rispetto al fatturato teorico del campione.

Tavola 7 - Tassi di copertura in termini di fatturato per le indagini su *informatica, manutenzione autoveicoli e ingrosso* *(anni 2006-2007)*

	Tassi di copertura 2006				Tassi di copertura 2007	
	I trim	II trim	III trim	IV trim	I trim	II trim
Informatica						
60 giorni	48,1	55,7	73,0	66,3	74,9	66,7
90 giorni	72,2	73,3	75,0	72,0	79,2	75,1
180 giorni	83,3	82,5	80,6	82,1	82,2	----
Manutenzione						
60 giorni	54,2	49,1	70,7	65,8	68,6	58,1
90 giorni	68,8	68,7	73,4	72,1	74,7	68,9
180 giorni	76,1	77,3	79,8	77,1	81,3	----
Ingrosso						
60 giorni	59,7	49,3	82,0	77,5	81,7	67,0
90 giorni	79,2	78,3	88,0	82,7	85,9	85,9
180 giorni	89,3	90,2	89,4	88,7	91,8	-----

Nel terzo trimestre 2006 il livello di tale tasso dopo 60 giorni ha quasi eguagliato quello che in precedenza si otteneva a 90 giorni. In altri termini le stime anticipate a 60 giorni sono state prodotte sostanzialmente con una numerosità campionaria ed una copertura analoga a quella che si aveva in precedenza con le stime a 90.

Si osserva che la revisione fra la stima anticipata (a 60 giorni), quella provvisoria (a 90) e quella definitiva (180) tende ad assestarsi su livelli abbastanza contenuti (Figure 8 e 9). Seppure il numero di trimestri considerato sia limitato, per il settore dell'*ingrosso* l'entità delle differenze osservate è costante nel tempo, mentre per il settore dell'*informatica* appare in modo evidente l'effetto positivo dell'innovazione del processo produttivo a partire dal 2006, che ha portato ad un progressivo avvicinamento delle stime.

Figura 8 - Variazioni tendenziali del fatturato a 60 a 90 e a 180 giorni per il settore ingrosso *(anni 2005-2007)*

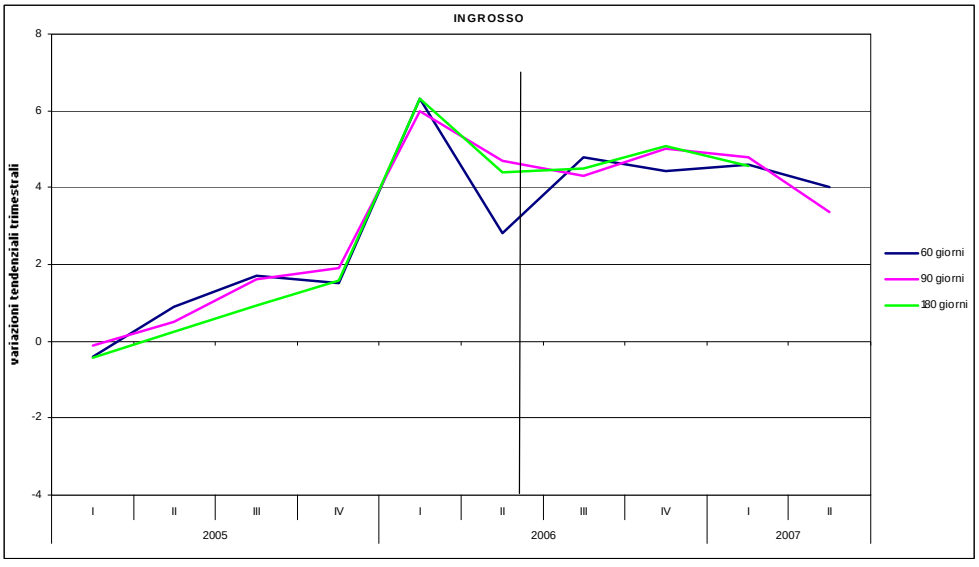


Figura 9 - Variazioni tendenziali del fatturato a 60 a 90 e a 180 giorni per il settore *informatica* *(anni 2005-2007)*

Per ridurre la revisione altre sperimentazioni sono in corso per verificare l'effetto del trattamento degli outlier, cioè di quelle imprese che, presentando variazioni tendenziali di fatturato anomale rispetto a quelle osservate nello strato di appartenenza, vengono escluse dal calcolo degli indici. L'eliminazione o l'inclusione di un valore anomalo dal calcolo ha effetti importanti sul valore dell'indice.

L'individuazione degli outlier viene effettuata normalmente con il metodo di Hidiroglou–Berthelot, che, operando una trasformazione lineare della variazione tendenziale, utilizza una funzione della differenza interquartile per definire la soglia di accettazione / rifiuto. La funzione proposta (e conseguentemente la soglia) dipende da un parametro C la cui determinazione è lasciate all'esperienza del ricercatore, e viene fissata tenendo conto della percentuale di osservazioni outlier che si intende mediamente escludere dal calcolo degli indicatori. Tale percentuale è si riduce all'aumentare del valore di C.

Per quanto riguarda le stima anticipate va sottolineato che la definizione di impresa come outlier dipende anche dalla composizione dello strato di appartenenza e quindi, sulla base delle medesime regole, una impresa cessa di essere o diventa outlier nei diversi momenti in cui gli indicatori devono essere calcolati (60, 90, 180), in quanto varia di volta in volta l'insieme delle imprese analizzate. Alcune sperimentazioni al vaglio sono: a) il "rilassamento" delle soglie che definiscono i valori anomali: scegliendo di escludere (sempre) un numero inferiore di imprese si dovrebbe ottenere una stabilizzazione degli indici calcolati nei diversi momenti; b) la specificazione di regole di coerenza per cui un'impresa, definita outlier in una fase iniziale (a 60 giorni), mantiene il suo stato di valore anomalo anche nei momenti successivi, a prescindere dalla sua "posizione" nei momenti successivi.

5. La rilevazione su retribuzioni di fatto e costo del lavoro (Oros)

5.1 Le caratteristiche generali della rilevazione Oros e gli obblighi comunitari di riduzione dei tempi di rilascio

Gli indicatori trimestrali Oros (Occupazione, Retribuzioni e Oneri Sociali) su retribuzioni di fatto per Ula (Unità di lavoro equivalenti a tempo pieno) e costo del lavoro per Ula vengono stimati ricorrendo all'integrazione dei dati amministrativi di fonte INPS con le informazioni tratte dall'indagine mensile dell'Istat sul lavoro e le retribuzioni nelle grandi imprese (GI). La popolazione obiettivo è rappresentata dalle imprese attive con almeno un dipendente, nei settori di attività economica dell'industria e dei servizi privati (sezioni da C a K dell'Ateco 2002).

I dati Oros vengono utilizzati non solo per produrre le stime per Ula diffuse trimestralmente a partire dal 2003 attraverso regolari comunicati stampa, ma anche per soddisfare due regolamenti comunitari: il primo relativo alle statistiche congiunturali sulle imprese (STS Reg. CE n. 1165/1998); il secondo riferito all'indice del costo del lavoro trimestrale (Labour cost index - LCI Reg. CE n.450/2003)[12].

Per il Regolamento STS è previsto l'invio di due indicatori relativi alle retribuzioni lorde e all'occupazione. Inizialmente entrambi gli indicatori venivano forniti all'Eurostat entro 90 giorni dal trimestre di riferimento, ma secondo i recenti emendamenti (Reg. CE n.1158/2005) dall'estate del 2006 la variabile occupazione deve essere inviata a 60 giorni.

Il Regolamento LCI prevede la diffusione di un indice delle retribuzioni lorde orarie, un indice degli oneri sociali orari e, quale sintesi dei due precedenti, un indice del complessivo costo del lavoro orario che, a partire da giugno 2005, devono essere inviati all'Eurostat entro 70 giorni dal trimestre di riferimento.

Gli indicatori Oros a livello nazionale vengono diffusi attraverso un comunicato stampa con un ritardo che è stato ridotto progressivamente da 90 a 70 giorni dalla fine del trimestre di riferimento. Per ogni trimestre t, oltre alla stima provvisoria del trimestre, vengono prodotte una stima semidefinitiva relativa a t-4 e una definitiva relativa a t-5, rilasciate rispettivamente dopo 12 e 15 mesi dalla diffusione delle stime provvisorie.

La principale fonte dei dati Oros è costituita dalle dichiarazioni dei contributi previdenziali e assistenziali (modello DM10) che ogni impresa con dipendenti deve presentare mensilmente all'INPS (Baldi e altri, 2004). Nella procedura di stima degli indicatori provvisori e definitivi, le unità presenti negli archivi INPS vengono distinte in quattro sottopopolazioni: 1) le imprese di piccola e media dimensione (PMI); 2) le imprese che si abbinano (attraverso il codice fiscale) a quelle rispondenti alla rilevazione mensile Istat Grandi Imprese; 3) le imprese di grandi dimensioni non presenti nell'indagine GI[13]; 4) le imprese interinali.

La stima per le imprese sub 2, ovvero quelle che rientrano nel dominio dell'indagine mensile GI, viene ottenuta utilizzando i dati provenienti da quest'ultima[14]. Tale stima non subisce revisioni per definizione. Al contrario, le stime provvisorie relative alle tre sottopopolazioni derivanti dalle dichiarazioni DM10 sono soggette a revisione.

---

[12] Per ora sia per gli indicatori STS sia per quelli LCI è previsto solo l'invio ad Eurostat e non la diffusione nazionale.

[13] La rilevazione GI si basa su un panel chiuso di grandi imprese (oltre 500 addetti), fissato in base 2005.

[14] Numerose sono le ragione di questa scelta. Inizialmente l'insieme delle dichiarazioni telematiche non conteneva che poche unità di grandi dimensioni. Successivamente, oltre ad un problema di numerosità, è stato appurato che la qualità dei dati rilevati e controllati direttamente da personale Istat appositamente specializzato in tali attività è maggiore.

Per la stima provvisoria vengono utilizzate le dichiarazioni pervenute all'INPS per via telematica e trasmesse, senza subire alcun controllo da parte delle sedi INPS, all'Istat a 45 giorni dalla fine del trimestre di riferimento. La stima definitiva (e semidefinitiva) viene effettuata utilizzando le dichiarazioni contributive dell'archivio finale dell'INPS (pervenute per via telematica o su supporto cartaceo), che contiene l'universo dei DM10 del trimestre dopo il controllo e correzione da parte dell'Istituto di Previdenza.

5.2 La riduzione dei tempi di acquisizione dei dati INPS

Il processo di produzione dei dati Oros è complesso. Per far fronte all'elaborazione degli indicatori comunitari, si è resa necessaria una consistente riduzione dei tempi di produzione. In via teorica la riduzione dei tempi per soddisfare le richieste dei regolamenti europei si sarebbe potuta ottenere:

- riorganizzando l'intero processo di produzione;
- rivedendo la metodologia ed effettuando una stima anticipata eventualmente utilizzando una informazione ridotta;
- anticipando l'acquisizione dei dati amministrativi.

La prima opzione non si è rivelata praticabile in quanto la durata del processo è già estremamente ridotta e ulteriori margini di miglioramento non sono possibili, considerati anche i vincoli derivanti dalla necessità di integrare nelle stime i dati relativi alle grandi imprese che sono disponibili a circa 58-60 giorni dalla fine del trimestre di riferimento.

La seconda alternativa, che implica l'utilizzo esclusivo dell'informazione relativa ai primi due mesi del trimestre, acquisibile in tempi ridotti, comporta un notevole investimento per lo sviluppo e la sperimentazione di adeguate metodologie di stima anticipata e quindi tempi lunghi di implementazione.

La sperimentazione della terza opzione si è invece resa possibile in seguito al recente miglioramento, in termini di tempestività e di quota di rispondenti, nel flusso con cui l'INPS acquisisce i dati dalle imprese.

Il termine massimo di presentazione del DM10 è fissato nell'ultimo giorno del mese successivo a quello cui si riferisce il periodo di paga e le imprese, almeno fino al 2004, potevano scegliere se presentarlo all'INPS su supporto cartaceo oppure per via telematica.

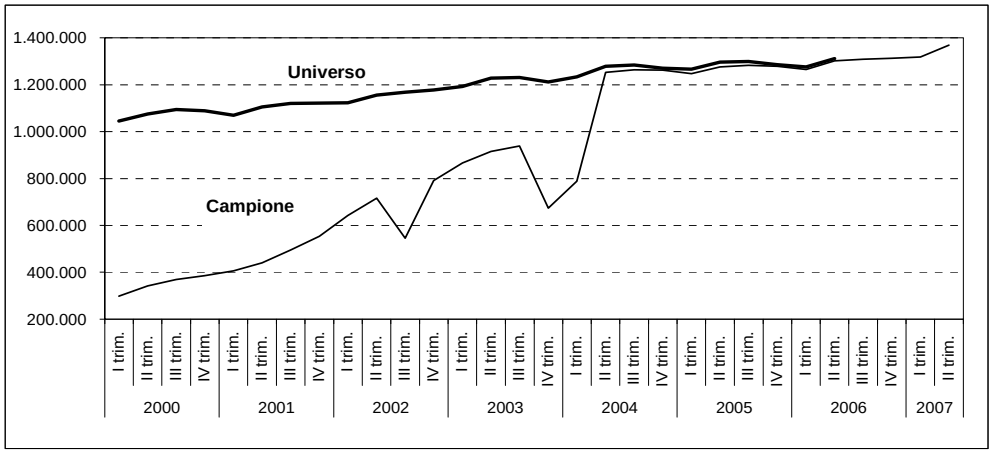
Come si è detto, i dati utilizzati per la stima provvisoria sono quelli forniti dall'INPS, a circa 45 giorni dalla fine del trimestre di riferimento, e si riferiscono all'insieme di dichiarazioni contributive inviate dalle imprese per via telematica. Si utilizzava questo insieme "ridotto" di dati in quanto i modelli inviati dalle imprese in formato cartaceo impiegavano invece molto tempo prima di essere registrati e disponibili in formato elettronico.

La stima provvisoria era quindi ottenuta utilizzando l'insieme di dati più rapidamente disponibili, come si trattasse di un campione "non casuale", attraverso un modello predittivo stimato per sottogruppi della popolazione, che utilizzava informazioni correnti e ausiliarie (Baldi e altri, 2004).

Alla fine del 2004 l'INPS, seguendo le disposizioni legislative (art.44, Legge n.326/2003), ha reso obbligatorio per tutte le imprese l'invio telematico del modello DM10. Se negli anni fino al 2003 la crescita delle dichiarazioni telematiche è stata graduale (Figura 10), a parte alcuni trimestri di caduta dovuta a fattori strettamente

amministrativi[15], a partire dal 2004, in pochi mesi, l'insieme delle dichiarazioni telematiche, inviate prevalentemente attraverso internet (o su supporto informatico), ha superato 1,2 milioni di unità. Quindi il campione "non casuale" si è ampliato fino a coprire quasi tutto l'universo di riferimento.

Figura 10 - Numero delle dichiarazioni contributive (DM10) acquisite dall'Istat relative all'universo e al campione *(Trimestri 1:2000 - 2:2007)*



Il notevole incremento nell'insieme informativo disponibile ha indotto una modifica della metodologia di stima provvisoria, poiché il campione ha assunto le dimensioni di un vero e proprio universo provvisorio. E' stato quindi possibile introdurre due importanti innovazioni:

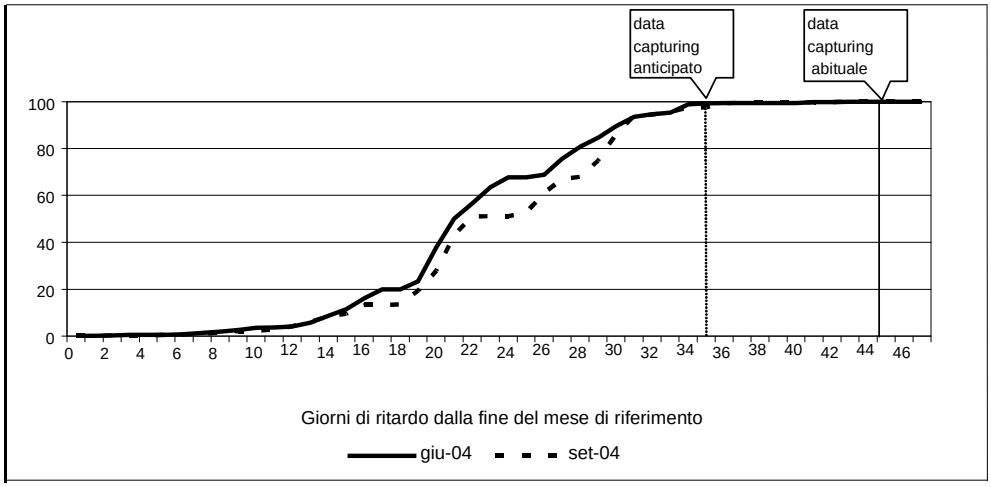
- a partite dalla stima del secondo trimestre 2004, rilasciata nel mese di settembre 2004, non è stato più necessario ricorrere alla procedura di calibrazione del campione, ma si è passati ad una metodologia più simile a quella utilizzata per le stime basate sull'universo.
- si è effettuata la sperimentazione della riduzione dei tempi di acquisizione da 45 a 35 giorni dalla fine del trimestre di riferimento dal 2005.

Quest'ultima opportunità è stata valutata considerando l'andamento degli arrivi mensili all'INPS delle dichiarazioni contributive. La "curva dei ritorni" ha un andamento tendenzialmente asintotico e simile in tutti i mesi: nei primi giorni successivi al mese di riferimento solo poche imprese inviano le dichiarazioni, ma già a circa 20 giorni si raggiunge almeno il 60%, a 25 giorni l'80% e poi si ha un riempimento graduale. Per la trasmissione dei dati all'Istat, l'INPS estrae dai propri archivi le dichiarazioni relative al trimestre appena trascorso in una unica occasione, a 45 giorni dalla fine del trimestre stesso. I dati relativi ai primi due mesi del trimestre hanno quindi un grado di copertura notevole, mentre quelli relativi all'ultimo mese si trovano in una situazione più problematica, soprattutto al terzo mese di ogni secondo e quarto trimestre (rispettivamente giugno e dicembre), che sono caratterizzati da valori delle retribuzioni

[15] Legati a problemi tecnici nella gestione e nell'aggiornamento dei sistemi informatici dell'INPS.

influenzati dall'erogazione rispettivamente della quattordicesima e della tredicesima. Una riduzione dei 45 giorni standard per la fornitura dei dati può essere richiesta tenendo conto che questo anticipo non influisce sul numero delle dichiarazioni disponibili relative ai primi due mesi, mentre può diminuire significativamente la numerosità delle dichiarazioni relative all'ultimo mese del trimestre.

Figura 11 - Frequenza percentuale cumulata delle dichiarazioni contributive (DM10) trasmesse all'INPS per via telematica per giorni di ritardo dalla fine del mese di riferimento nel terzo mese del trimestre *(Giugno-Settembre 2004)*



La Figura 11 evidenzia il grado di popolamento dell'archivio delle dichiarazioni contributive relative al terzo mese del trimestre (mesi di giugno e settembre 2004), nelle diverse occasioni temporali in cui è stata programmata l'acquisizione dei dati.

La riduzione dei tempi di acquisizione ha iniziato ad essere sperimentata a partire dai primi mesi del 2005 sulla base della valutazione dei tempi di trasmissione delle dichiarazioni contributive relative agli ultimi tre trimestri del 2004.

Se la riduzione dei tempi fino a 35 giorni non mostra nessun sensibile peggioramento rispetto a grado di popolamento che si ottiene con un'acquisizione a 45 giorni dalla fine del trimestre di riferimento, ogni ulteriore riduzione dell'acquisizione dei dati invece comporta una diminuzione più veloce del grado di copertura, anche nell'arco di pochi giorni. Ad esempio, portare a 31 giorni dalla fine del mese l'acquisizione delle dichiarazioni DM10 relative al mese di giugno 2004 comporterebbe la perdita di circa il 10% delle dichiarazioni DM10 del campione.

La sperimentazione è stata, quindi, avviata riducendo prudenzialmente i tempi di acquisizione dei dati relativi al primo trimestre del 2005 a 36 giorni rispetto alla fine del trimestre di riferimento.

Tavola 8 - Numero delle dichiarazioni contributive (DM10) relative al campione e all'universo e tasso di copertura del campione dell'indagine Oros nelle sezioni da C a K *(Gennaio 2004 - Dicembre 2005)*

Mese di competenza	Campione	Universo	Tasso di copertura
Gen-04	741.439	1.169.746	63,4
Feb-04	743.072	1.173.795	63,3
Mar-04	744.261	1.187.807	62,7
Apr-04	1.169.593	1.213.263	96,4
Mag-04	1.206.621	1.223.743	98,6
Giu-04	1.207.065	1.239.466	97,4
Lug-04	1.234.061	1.249.142	98,8
Ago-04	1.221.568	1.237.421	98,7
Set-04	1.159.773	1.237.121	93,7
Ott-04	1.222.629	1.230.551	99,4
Nov-04	1.215.035	1.226.593	99,1
Dic-04	1.207.180	1.220.341	98,9
Gen-05	1.197.198	1.211.813	98,8
Feb-05	1.195.287	1.212.335	98,6
Mar-05	1.195.904	1.222.989	97,8
Apr-05	1.222.109	1.233.328	99,1
Mag05	1.225.947	1.242.710	98,7
Giu-05	1.194.582	1.255.218	95,2
Lug-05	1.251.958	1.263.238	99,1
Ago-05	1.237.473	1.250.809	98,9
Set-05	1.227.210	1.251.532	98,1
Ott-05	1.237.034	1.243.186	99,5
Nov-05	1.228.413	1.238.606	99,2
Dic-05	1.215.307	1.235.000	98,4

Da allora, la soglia dei 35 giorni è stata superata soltanto in due trimestri per far fronte a cadute nel flusso di trasmissione delle dichiarazioni contributive, ma in alcuni trimestri è stato possibile addirittura ridurla sino a 33 giorni. Il tasso di copertura raggiunto dal campione delle dichiarazioni DM10 non è sceso quasi mai al di sotto del 98%, anche successivamente all'anticipazione dei tempi di acquisizione dei dati INPS [16] (Tavola 8).

5.3 I risultati: miglioramento della tempestività ed effetti sulla revisione

Gli effetti positivi del consistente aumento dell'insieme di dati disponibili e della riduzione dei tempi di acquisizione sulla qualità delle stime possono essere valutati attraverso l'analisi dell'indicatore sulla tempestività degli indicatori Oros e dell'entità delle revisioni della stima provvisoria. Il primo indicatore, calcolato come differenza tra la data di pubblicazione effettiva delle stime provvisorie e la data di riferimento delle stime stesse (convenzionalmente l'ultimo giorno del trimestre di riferimento), mostra che si sta gradualmente realizzando la riduzione dei tempi di rilascio degli indicatori da 90 a 70 giorni (Tavola 9).

Dalla fine del 2004, disponendo di un insieme molto ampio di osservazioni, è stato possibile adottare una semplificazione della metodologia di stima, quindi ridurre in parte i tempi del processo di produzione della rilevazione, mentre a partire dai dati del 2005 è stata l'anticipazione dell'acquisizione delle dichiarazioni DM10 per via telematica a permettere di migliorare la tempestività, sino al raggiungimento di un ritardo di circa 70 giorni.

[16] Il mese di giugno 2005 rappresenta un'eccezione dovuta a una consistente caduta nel flusso dei dati amministrativi che è stata fronteggiata con l'acquisizione di scarichi supplementari. L'INPS, infatti, permette alle imprese che chiudono per ferie collettive di posticipare l'invio della dichiarazione DM10.

Tavola 9 - Revisioni ed indicatore di tempestività (ritardo in giorni) delle variazioni tendenziali degli indici delle retribuzioni lorde per Ula, degli oneri sociali per Ula e del costo del lavoro per Ula *(Trimestri 3:2003 - 4:2005)*

	Revisione definitiva delle variazioni tendenziali			Tempestività
Trimestre di riferimento	Retribuzioni lorde per Ula	Oneri sociali per Ula	Costo del lavoro per Ula	Giorni di ritardo rispetto alla fine del trimestre di riferimento
3:2003	-0,6	-0,8	-0,7	91
4:2003	0,1	0,5	0,2	89
1:2004	-0,3	-0,7	-0,4	89
2:2004	-0,2	-0,2	-0,1	91
3:2004	-0,1	0,4	0,1	83
4:2004	0,0	-0,1	0,0	81
1:2005	0,0	0,2	0,0	78
2:2005	-0,1	-0,4	-0,1	76
3:2005	0,1	-0,2	0,2	75
4:2005	0,0	-0,3	0,0	73

La riduzione dei tempi di pubblicazione nazionale, e di rilascio degli indicatori LCI e STS, è stata possibile senza effetti negativi sulla qualità delle stime, infatti la revisione delle stime per i tre indicatori Oros si assesta a livelli molto bassi e soddisfacenti, già a partire dal secondo trimestre del 2004.

6. Conclusioni

La sfida della riduzione dei tempi di rilascio dei dati, per soddisfare i requisiti di tempestività richiesti in sede europea, è stata affrontata in modo diverso nei contesti delle varie indagini coinvolte, dando vita a sperimentazioni e a ristrutturazioni del processo produttivo che hanno infine consentito di garantire, da un certo momento in poi, il rispetto delle scadenze "anticipate".

Ove è stato possibile si è giocato sui margini di flessibilità dei processi produttivi, riducendo i tempi di esecuzione, in altri casi è stato necessario adottare tecniche di indagine alternative, come la selezione di campioni *ad hoc* che utilizzano, per il calcolo delle stime anticipate, un sottoinsiemi di rispondenti.

Per velocizzare l'acquisizione dei dati si è dimostrata utile l'adozione o l'incremento dell'uso di nuove tecnologie, come quella *web*. Nel caso delle indagini dirette presso le imprese campione la modalità di risposta per via telematica è su base volontaria e sembra in crescita, ma è quando la legislazione obbliga all'uso di *internet* (che è il caso di dati amministrativi), che si ottengono in tempi brevi tassi di risposta molto più favorevoli.

Attualmente gli indicatori richiesti vengono forniti, perlopiù in veste confidenziale, ad Eurostat nei tempi stabiliti dai Regolamenti, anche se esistono ancora margini di miglioramento delle qualità, per esempio in termini di coperture realizzate per le stime anticipate o di riduzione della revisione fra dati provvisori e definitivi. Gli sviluppi futuri nella predisposizione di stime congiunturali anticipate dovranno prendere in considerazione diversi strumenti e opportunità per migliorare accuratezza e tempestività.

Sembra sempre più necessario spingere sull'uso delle tecnologie *web*, come pure migliorare le politiche di sollecito alle unità non rispondenti e trovare metodi adeguati per convincere imprese e istituzioni a rispondere alle indagini e a rispettare i tempi e scadenze. Anche la sperimentazione di stimatori anticipati alternativi sarà utile per arrivare ad ottimizzare la procedure di stima e sfruttare al meglio l'informazione disponibile alla data "anticipata".

Riferimenti bibliografici

Bacchini F., Iannaccone R., Otranto E. (2005) "L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione italiani" *Contributi-Istat*, vol. 5.

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 Istat*, vol. 5 pp. 497-519.

Barcaroli G., D'Aurizio L., Luzi O., Mannari A., Pallara A. (1999) "Metodi e software per il controllo e la correzione dei dati", *Quaderni di ricerca Istat,* n. 1.

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

Congia M.C., Rapiti F. (2006), "Quality evaluation of the effects on the employment estimates of the reduction of delay (90 to 60 days for employment)", *Final technical report, Grant Agreement Eurostat-Istat n° 20044440106 "Hours worked, Wages & Salaries for Annex C, D" per il regolamento STS*.

Congia M.C., Rapiti F. (2007), "Gli indicatori di revisione nella rilevazione trimestrale Oros sulle retribuzioni di fatto, gli oneri sociali e il costo del lavoro in Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI – 1ª giornata", *Contributi Istat*, vol. 6 pp. 131-145.

D'Alò M., Gismondi R., Solari F., Naccarato A. (2006), "Estimation in Repeated Business Surveys using Preliminary Sample Data", *Atti della XLIII Riunione Scientifica SIS*, 14-16 giugno, Torino.

De Sandro L, Gismondi R., (2004) Provisional Estimation of the Italian Monthly Retail Trade Index, *Contributi-Istat*, vol. 24.

Di Zio M., Guarnera U., Luzi O., (2007) "Rilevazione statistica sui permessi di costruire (grandi comuni) – Risultati della sperimentazione delle tecniche di imputazione con donatore longitudinale e con media longitudinale", Nota tecnica, Istat.

Eurostat (2000), *Short-term Statistics Manual*, Eurostat, Luxembourg,.

Eurostat (2005), *Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council* – Unofficial Consolidated Version, documento non pubblicato, Eurostat, Lussemburgo.

Eurostat (2005) *Methodology of short term business statistics*, Eurostat, Lussemburgo

Falorsi P.D., Alleva G., Bacchini F., Iannaccone R., (2005) "Estimated based on preliminary data from a specific subsample and from respondents not included in the subsample" *Statistical Methods and Applications*, vol. 14, n.1, pp. 83-99.

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

Harvey A. C. (1984), "Dynamic Models, the Prediction Error Decomposition and State-space", in *Econometrics and Quantitative Economics*, D. F. Hendry and K. F. Wallis (eds.), pp.37-59. Blackwell, Oxford.

Hedlin D., Falvey H., Chambers R., Kokic P. (2001), "Does the Model Matter for GREG Estimation? A Business Survey Example", *Journal of Official Statistics*, 17, pp. 527-544.

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

Istat (a cura di G. Rallo) (2005) "Statistiche sui permessi di costruire", *Collana informazioni* n. 32

Istat (2007), Rapporto Finale del gruppo di lavoro: "Sperimentazione di stime anticipate per specifici indicatori congiunturali, finalizzata al rilascio in produzione delle relative metodologie", Documento interno; Istat.

Istat (2007), Nota informativa del comunicato stampa relativo agli indici trimestrali di fatturato per alcune attività dei servizi, pubblicata *on line* sul sito http://www.istat.it

Lee H. (1995) Outliers in business surveys, in Cox, Binder, Chinappa, Christianson, Colledge, Kott *Business survey methods*, pag. 503-523John Wiley & Sons, New York

Rao J.N.K., Srinath K.P., Quenneville B. (1989), "Estimation of Level and Change using Current Preliminary Data", in *Panel Surveys*, (Kasprzyk, Duncan, Kalton G., Singh eds.), 457-485, John Wiley & Sons, New York.

Royall R.M. (1992), "Robustness and Optimal Design Under Prediction Models for Finite Populations" *Survey Methodology*, 18, 179-185.

Valliant R. (1999), Uses of Models in the Estimation of Price Indexes: a Review, in *Proceedings of the Survey Research Methods Section*, American Statistical Association.

The planning of Preliminary Sample: methodological aspects and an application to the Italian Monthly Retail Trade Survey

Paolo Righi[1], *Tiziana Tuoto*[2]

Abstract

The standard approaches for dealing with provisional estimates essentially focus on the study and the definition of efficient estimator, exploiting the auxiliary information almost exclusively in the estimation phase. In this paper, we propose an extended appliance of the auxiliary information, developing an overall strategy involving both the sample design and the estimator. The work aims to define a preliminary sub-sample of units included in the planned sample for final estimates. Afterwards, the provisional estimates are computed by means of an appropriate estimator applied to the preliminary sub-sample data. In order to design the preliminary sub-sample, the balanced sampling theory is exploited. In the longitudinal survey context, where timeliness is a pressing target, the approach represents an approximately bias-robust sampling strategy. Results of a real survey data simulation study comparing the proposed strategies with strategies ignoring the planning of the preliminary sub-sample are presented.

Keywords: Balanced Sampling, Provisional estimator, Short-term statistics, Timeliness.

1. Introduction

Timeliness of statistical data is becoming a pressing target at both national and international level. An Amendment EU Regulation on Short-Term Statistics (August 2005) underlines the renovate interest of this quality aspect, requiring to the statistical institutes of the EU Member States to transmit to EUROSTAT *provisional estimates* with a reduced time with respect to the original 1998 Regulation (EUROSTAT 2000; 2005). The Italian Statistical Institute (ISTAT) has been engaged in research projects (ISTAT, 2007) to deal with this problem, analysing some approaches which implement particular *ad hoc* sampling strategies (an estimator coupled with a sampling design).

The standard approach to the production of the provisional estimates generally does not plan in advance a sample for the computation of these estimates, but merely involves the use of the *quick respondent units*, in this context referred to as *Unplanned Preliminary Observed Sample* (UPOS). In fact, in order to obtain "good" provisional estimates, standard

[1] Ricercatore(Istat), e-mail: parighi@istat.it.
[2] Ricercatore(Istat), e-mail: tuoto@istat.it.

survey strategy often aims to achieve high quick response rate by means of well-structured plan of follow-up where, in particular, the most "large" units are carefully supervised. According to this approach the units included in the UPOS are not drawn by a specified probabilistic sample design. This standard approach exploits the availability of the auxiliary information almost exclusively in the estimation phase.

In this paper, we propose an extended appliance of the auxiliary information, anticipating the use of the available auxiliary information also to the sample design phase, through the plan of a *Preliminary Theoretical Sample* (PTS). The PTS is a sub-sample of the planned sample for final estimates. Afterwards, in the estimation phase the auxiliary information is still used for defining the suitable provisional estimator. So, an *overall strategy* for the provisional estimates production is developed, involving both the sample design and the estimator definition.

The introduction of the PTS requires some changes in the data-collection process, that is an intensive follow-up of the PTS units, so that the *Planned Preliminary Observed Sample* (PPOS) will be as close to PTS as possible. With respect to the standard strategy, in the PTS approach the follow-up will be more intensive but less widespread.

In this work, we propose to define the PTS in the general framework given by the balanced sampling theory (Royall and Herson, 1973; Royall, 1992; Valliant *et al*., 2000) according to the model based inferential paradigm. In particular, the planning of PTS is developed in the context of longitudinal short-term business surveys. The balanced sampling allows to consider a complex auxiliary information structure, taking into account also the value of variables of interest in the previous occasions, when treating panel data. When the parameter of interest is an amount, the balanced sample guarantees model unbiasedness of the estimate, even if the true superpopulation model is more general than the working model underlying the estimator. However, in the longitudinal survey context, the parameters of interest are also non-linear functions of totals, and, then, with balanced sampling an approximately bias-robust sampling strategy can be defined.

The paper points out both methodological aspects and practical application to real survey data. It is structured as follows: section 2 gives a concise description of the balanced sampling theory; section 3 illustrates the main aspects of the survey frame of the simulation study, the ISTAT Monthly Retail Trade Survey (MRTS); section 4 discusses the definitions of different kinds of balanced PTS suitable to the MRTS. Section 5 is devoted to the description of the final pseudo-sample based on 2004 MRTS data, where missing values are removed. Finally, section 6 shows the empirical results of the simulation study comparing the provisional estimates based on UPOS and PTS.

2. Sampling design

Let us define as *provisional estimate* the estimation of a parameter of interest obtained on the basis of a sub-sample of *quick respondent units* that is collected within a time lag Δ'_t after the reference time point *t* of the survey. The correspondent *final estimate* is based on a final sample, including both *quick* and *late respondents*, observed within a time lag Δ_t $(> \Delta'_t)$, being Δ_t the deadline to produce the final estimates.

In order to obtain small *revision* (difference among provisional and final estimates) a decisive task is to have the quick respondent sample as close as possible to the final

observed sample.

As well as affected by revision, the provisional estimates could be biased when final and quick respondents sampling distributions are quite dissimilar (even if the final sample produces unbiased estimates). To avoid these situations, huge efforts are spent in the data-collection process in order to obtain a UPOS as large as possible. On the other hand, defining a suitable PTS the effort is concentrated in obtaining a PPOS as close as possible to PTS, bounding the set of the followed-up units.

The sampling design for PTS must be coordinated with the sampling design of the final sample and with the provisional and final estimation process. The ideal situation is the planning of all these elements in the same time, but in many survey contexts the provisional estimation goals and consequently the PTS are defined after the other elements. In the last case, analysed in the paper, the possibility of defining different types of sampling design is restricted. First of all, a constraint is due to the form of the provisional estimator, that should be similar to the given final estimator, in order to reduce the revision; then the provisional estimator working model should be a generalization of the working model of the final estimator. The aim of the paper is to define a preliminary sampling design considering the sampling design of the final sample and the final estimators as given.

Sampling design may be defined according to the adopted inferential paradigm denoted as design based/model assisted or model based. In the following we will consider the last one. Let us introduce the general superpopulation model *M*,

$$\begin{cases} Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ E(\varepsilon_i) = 0; \quad E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 q(\mathbf{x}_i) & i = j \\ 0 & \text{else} \end{cases} \end{cases} \tag{2.1}$$

being Y_i the random variable of interest for unit *i*, $\mathbf{x}_i$ a $p\times 1$ value vector of the auxiliary variables, $\boldsymbol{\beta}$ a regression coefficient vector, ε_i the residual term of unit *i*, *q* a known function of the known value $\mathbf{x}_i$ and σ^2 a possibly unknown constant. Given a sample *s*, of size *n*, the population total T_Y can be written as $T_Y = \sum_s Y_i + \sum_{(U-s)} Y_i$, being *U* the population and (*U*-*s*) the non sampled population. If we knew the $\boldsymbol{\beta}$ vector an estimator of T_Y is $T_Y^* = \sum_s Y_i + \sum_{(U-s)} \mathbf{x}_i'\boldsymbol{\beta}$ and an operative estimator could be $\hat{T}_Y = \sum_s Y_i + \sum_{(U-s)} \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. This estimator is unbiased under the model *M*, hereinafter denoted as *working model*, if $\hat{\boldsymbol{\beta}}$ is such that $[E_M(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}]\sum_{(U-s)} \mathbf{x}_i' = 0$, where $E_M(.)$ denotes the expectation with respect to the model *M*.

Furthermore, the error variance of $\hat{T}_Y$ expressed by $Var_M(\hat{T}_Y - T_Y)$ is minimized if the variance of $\hat{\boldsymbol{\beta}}$ is minimized too. Among the linear estimator, the error variance is minimized under the model (2.1) when

$$\hat{\boldsymbol{\beta}} = \left(\sum_s \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q(\mathbf{x}_i)} \right)^{-1} \sum_s \frac{\mathbf{x}_i Y_i}{q(\mathbf{x}_i)} . \qquad (2.2)$$

The estimator using $\hat{\boldsymbol{\beta}}$ is denoted as the Best Linear Unbiased (BLU) predictor.

The unbiasedness of an estimator depends on the working model, that is the same estimator may be biased under alternative models having different regression functions. Model-based approach makes inference on the working model, assuming that it represents a satisfactory approximation of the true superpopulation model. Nevertheless, if the working model is seriously incorrect the estimator is strongly biased. To avoid this problem, a *robust* sampling strategy may be defined, in the sense that it performs well with the working and alternative models.

The *bias-robust* strategy involves the sample selection phase and in same case also a slight changing of the adopted estimator in order to obtain efficiency gains. As far as the sample selection to protect from bias is concerned, let us consider the *balanced sampling design*.

The definition of a balanced sample depends on the assumed inferential approach. Roughly speaking, in the model based approach a sample is defined as *balanced* on a set of auxiliary variables if there is the equality between the sample and the known population means of the auxiliary variables (Royall and Herson, 1973; Valliant *et al.*, 2000). According to the considered estimator, different kinds of balanced sample can be used.

Therefore, before defining the sampling design, the knowledge of the estimator form is needed.

In the following the Italian Monthly Retail Trade Survey data (MRTS) (Gismondi and De Sandro, 2004) is analysed.

3. Parameters of interest and sampling strategy of the Italian Monthly Retail Trade Survey

The MRTS is based on a monthly measurement of the turnover of a stratified sample of retail enterprises (Division 52 of NACE nomenclature for a population of about 570 thousands) of different types and sizes. The sample is composed by a panel and a non-panel component, drawn every year and observed for 12 months. The survey provides provisional estimates within 30 days after the reference time and final estimates within 54 days. The provisional retail trade indices are referred to the domains: type of sold product (food and non-food retail enterprises) type of distribution (large and small retail enterprises). Then the parameters of interest at the month *t* are defined as:

$$I_d^{t,0} = \left(\sum_{h \in d} I_h^{t-12,0} R_h^t \gamma_h \right) \Big/ \sum_{h \in d} \gamma_h \qquad \text{with} \qquad R_h^t = \frac{\sum_{i \in U_h^{t,t-12}} Y_i^t}{\sum_{i \in U_h^{t,t-12}} Y_i^{t-12}}$$

where: *d* is the generic domain of interest; *h* is the generic stratum defined by the cross-classification of main group of product sold, class of employed persons and type of distribution for 120 strata; $I_h^{t-12,0}$ is the retail trade index of the same month *t* of the

previous year in the stratum h (with t=13, 14, …,24)[3]; γ_h is a stratum weight given by the yearly turnover referred to the base 2000, derived from structural business statistics (ASIA archive); Y_i^t and Y_i^{t-12} are the total turnover variables of the month t and of the same month of the previous year on the unit i, respectively; $U_h^{t,t-12}$ is the longitudinal population in time period (t, t-12) of stratum h. The product term ($I_h^{t-12,0} R_h^t$) represents the elementary index at stratum level.

The sampling design is a stratified simple random sampling, with about 7,500 units. Each year about 30% of sample is renewed[4].

In the questionnaire of the reference month t both the values of the variables Y^t and Y^{t-12} are collected with some other auxiliary variables.

Starting from the end of 2004, the evaluation of the preliminary estimates is based on an UPOS calculated after Δ'_t=29 days from the end of the reference month. The estimation phase follows a complex procedure. Here the main steps, used in the simulation study, are sketched.

All the non-respondents within Δ'_t are imputed to obtain the provisional estimates. For each domain of interest the provisional estimation process is given by

$$\widetilde{I}_d^{t,0} = \left(\sum_{h \in d} \widetilde{I}_h^{t-12,0} \, \widetilde{R}_h^t \gamma_h \right) \Big/ \sum_{h \in d} \gamma_h \qquad (3.1.1)$$

with

$$\widetilde{R}_h^t = \frac{\sum_{i \in s_{ah(t)}^t} y_i^t + \sum_{i \in \left(\widetilde{s}_h^t - s_{ah(t)}^t\right)} \widetilde{y}_i^t}{\sum_{i \in s_{ah(t-12)}^t} y_i^{t-12} + \sum_{i \in \left(\widetilde{s}_h^{t-12} - s_{ah(t-12)}^t\right)} \widetilde{y}_i^{t-12}} \qquad (3.1.2)$$

where: $\widetilde{I}_h^{t-12,0}$ is the estimate of $I_h^{t-12,0}$; y_i^t and y_i^{t-12} are the observed values of Y_i^t and Y_i^{t-12}, $\widetilde{y}_i^t$ and $\widetilde{y}_i^{t-12}$ are the imputed values for the non respondents; $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ are respectively the sample units giving information for the preliminary estimates about the variables Y^t and Y^{t-12} in stratum h in the survey of the month t, such that $s_{ah(t)}^t \subseteq \widetilde{s}_h^t$ and $s_{ah(t-12)}^t \subseteq \widetilde{s}_h^t$ being $\widetilde{s}_h^t$ the theoretical overall sample for the final estimates in stratum h at month t, while $\left(\widetilde{s}_h^t - s_{ah(t)}^t\right)$ and $\left(\widetilde{s}_h^t - s_{ah(t-12)}^t\right)$ are the corresponding non respondent samples after the time lag Δ'_t respectively for the Y^t and

[3] For instance January is indicated with t=13 and the same month of the previous year with t-12=1.

[4] For practical reasons this percentage could be highest. For instance in 2004 data, analysed in the simulation study, about the 50% of the sample belongs to the panel component (observed in the 2003 survey) while the other part is a new sample.

Y^{t-12} variables. In the simulation we use the values y^t and y^{t-12} collected in the questionnaire and we suppose that unit *i* providing the value of y_i^t gives information on y_i^{t-12} as well; then, $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ coincide and they are indicated by s_{ah}^t.

The imputation procedure is defined by two steps:

$$\widetilde{y}_i^{t-12} = a_i^{t-12} \frac{\sum_{i \in s_{ag}^t} y_i^{t-12}}{\sum_{i \in s_{ag}^t} a_i^{t-12}} , \qquad (3.2.1)$$

$$\widetilde{y}_i^t = \frac{\sum_{i \in s_{ag}^t} y_i^t}{\sum_{i \in s_{ag}^t} y_i^{t-12}} \frac{a_i^t}{a_i^{t-12}} \widetilde{y}_i^{t-12}, \qquad (3.2.2)$$

where: $s_{ag}^t = \bigcup_{h \in g} s_{ah}^t$ represents the sample of the quick respondents, of size n_{ag}^t belonging to the imputation cell *g* defined by crossing each other the type of distribution and class of employed persons variables (8 cells, 3 for large and 5 for small retail enterprises); a_i^t and a_i^{t-12} are the number persons employed respectively at month *t* and (t-12) for the unit *i*, observed in the survey or imputed with the following procedure: the value variable a_i^{t-12} missing is imputed with the value a_i^t if it is non missing, otherwise it is imputed with the value of the business register; before imputing a_i^t, the outlier values considering the ratio a_i^t / a_i^{t-12} are checked. If the ratio does not belong to the interval (0.1; 10) the value a_i^t is replaced with a_i^{t-12}. If a_i^t is missing it is imputed with the a_i^{t-12} value. In the expressions (3.2.1) and (3.2.2) we ignore this imputation process and consider these two variables as always observed. Final estimate procedure follows the same steps of preliminary estimates, working with the information of both the quick and late respondents.

Finally, let us note that, although is used a probabilistic sample and the numerator and denominator of (3.1.2) are Horvitz-Thompson estimates with the imputation of the missing values, these estimates can be analysed in the model based context as well.

4. Definition of the Preliminary Theoretical Sample in the MRTS

The aim of the simulation is to verify the impact of a planned PTS on the accuracy of the provisional estimates with respect to the final estimates. Therefore, we are not interested in defining a new estimator and we consider the same imputation processes defined in (3.2.1) and (3.2.2).

Let us consider the following superpopulation model,

$$\begin{cases} Y_i^{t-12} = \beta_g a_i^{t-12} + \varepsilon_i \qquad (i \in g) \\ E(\varepsilon_i) = 0;\ E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j \\ 0 & \text{else} \end{cases} \end{cases} . \tag{4.1}$$

It may be noted that, under the model (4.1), the imputation procedure (3.2.1) represents the BLU provisional ratio estimator, $\tilde{T}_{Y(g)}^{t-12}$, of the total $\hat{T}_{Y(g)}^{t-12}$, being $\hat{T}_{Y(g)}^{t-12}$ the final estimate of the total of the variable Y_g^{t-12} of the population $U_g^{t,t-12}$ in the theoretical final sample $\tilde{s}_g^t \left(= \bigcup_{h \in g} \tilde{s}_h^t\right)$. The (3.2.1) is also the BLU provisional predictor for each $\hat{T}_{Y(h)}^{t-12}$ (with $h \in g$). Denoting with d alternatively g or h, the difference $(\tilde{T}_{Y(d)}^{t-12} - \hat{T}_{Y(d)}^{t-12})$ represents the revision with respect to the final theoretical sample estimate.

The second imputation step can not be expressed via a linear superpopulation model instead. Nevertheless, a reduction of the imputation error in the first step is important for a "good" imputation in the second one.

Starting from (4.1), different sorts of preliminary samples can be drawn. Under this working model, the quick respondent sample that minimized the variance of $(\tilde{T}_{Y(d)}^{t-12} - \hat{T}_{Y(d)}^{t-12})$ consists of the n_{ag}^t units whose a^{t-12} values are largest (Royall and Herson, 1973). However, the selection of this sample can be a dangerous strategy when the (4.1) is wrong, since in this case the ratio estimator could be quite biased.

A balanced sample defines a bias-robust strategy against the model failure (Valliant et al., 2000).

Then, let us assume that the true underline model is the more general polynomial model,

$$\begin{cases} Y_i^{t-12} = \sum_{j=0}^{J} \delta_j \alpha_{j,g} \left(a_i^{t-12}\right)^j + \varepsilon_i, \quad (i \in g) \\ E(\varepsilon_i) = 0;\ E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j \\ 0 & \text{else} \end{cases} \end{cases}, \tag{4.2}$$

where δ_j is a zero-one indicator of whether the regressor $(a_i^{t-12})^j$ belongs to the model.

With the model (4.2) the estimator (3.2.1) is no more the BLU predictor, but the provisional ratio estimator is still an unbiased estimator of the final estimate under the more general model (4.2). That means the average of the revision is null, if the sample satisfies simple balancing equation expressed by

$$\sum_{s_{ag}^t} \frac{\left(a_i^{t-12}\right)^j}{n_{ag}^t} = \sum_{\tilde{s}_g^t} \frac{\left(a_i^{t-12}\right)^j}{\tilde{n}_g^t}, \qquad \text{(j=1, …, J)} \tag{4.3}$$

being $\tilde{n}_g^t$ the size of $\tilde{s}_g^t$.

Adding balanced equations like (4.3) referred to other auxiliary variables, also not belonging to the polynomial form, the ratio estimator remains bias-robust for the more complex model involving these new auxiliary variables.

The robustness of the ratio estimator in the balanced sampling is obtained increasing the variance of estimator.

Assuming a provisional estimator based on a slightly more general working model than (4.1) expressed by

$$\begin{cases} Y_i^{t-12} = \beta_{1/2g}\sqrt{a_i^{t-12}} + \beta_g a_i^{t-12} + \varepsilon_i \quad (i \in g) \\ E(\varepsilon_i) = 0; \qquad E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j \\ 0 & \text{else} \end{cases}, \end{cases} \tag{4.4}$$

the minimum variance is achieved selecting a weighted balanced sample (Royall, 1992) which satisfies the weighted balancing equation

$$\frac{1}{n_{ag}^t}\sum_{i \in s_{ag}^t} \frac{a_i^{t-12}}{\sqrt{a_i^{t-12}}} = \frac{\sum_{i \in \tilde{s}_g^t} a_i^{t-12}}{\sum_{i \in \tilde{s}_g^t} \sqrt{a_i^{t-12}}}. \tag{4.5}$$

With the model (4.4) the BLU predictor obtained by (2.2) is no more the ratio estimator, but we may note that model (4.1) is a particular case of (4.5).

The weighted balanced sampling guarantees unbiasedness and minimum variance under wide variety of polynomial regression models such as,

$$\begin{cases} Y_i^{t-12} = \eta_0 \mu_{0g} + \beta_{1/2g}\sqrt{a_i^{t-12}} + \\ \qquad + \sum_{j=1}^{J} \delta_j \beta_{jg} \left(a_i^{t-12}\right)^j + \sum_{j=1}^{J} \lambda_j \varphi_{jg} \left(f_i\right)^j + \varepsilon_i \quad (i \in g), \\ E(\varepsilon_i) = 0; \qquad E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j \\ 0 & \text{else} \end{cases} \end{cases} \tag{4.6}$$

where, μ, β's and ϕ's are regression coefficients; η_0, δ_j and ϕ_j are zero-one indicator variables and f_i is the value of an another auxiliary variable. For instance, f_i could be the previous year turnover value observed for panel units or the most recent turnover derived from the business register available in the planning design period for the non panel units. Of course, other auxiliary variables can be put in the model (4.6).

The optimality is achieved selecting a weighted balanced sample for each variable included in the polynomial model (4.6) (Royall, 1992). The weighted balancing equations (4.5) are replaced by

$$\frac{1}{n_{ag}^t}\sum_{i \in s_{ag}^t} \frac{(x_i)^j}{\sqrt{a_i^{t-12}}} = \frac{\sum_{i \in \tilde{s}_g^t} (x_i)^j}{\sum_{i \in \tilde{s}_g^t} \sqrt{a_i^{t-12}}} \quad (j=1, \ldots, J), \tag{4.7}$$

where, x_i indicates a_i^{t-12} or f_i. Hence, the BLU predictor based on model (4.4) remains optimal under the more general model (4.6), when the sample is weighted balanced on the corresponding variables.

Therefore, when the expected value of the variable of interest is a function of a^{t-12}, there are two alternative strategies: the first one is based on the model (4.2) and a sampling selection satisfying some simple balanced equations; the second strategy uses the estimator defined by the model (4.6) and a balanced design according to the equations (4.7). Which strategy is preferable is an open question (Valliant et al., 2000). The simulation study compares these two main choices.

4.1 Practical aspects for the selection of the Preliminary Theoretical Sample in the MRTS

The selection of balanced sample for a longitudinal survey presents some additional difficulties, mainly because the two main strategies should be connected with the particular survey occasion taken into account, e.g. a^{t-12}. In this sense, the two alternative strategies require that the PTS is drawn for every month. To select only one PTS for the whole year, the definition of a single value a_i for the set $\{a_i^{t-12}\}$ (t=13,…, 24) is needed. Many options may be analyzed. In the simulation study we use for the panel units the number of employed of October or the value of the closest month if October value is missing, or finally, the number of employed of the business register for panel units with missing data for the whole year and for non panel units.

A second fundamental aspect is how to select a balanced sample. Deville and Tillé (2004) proposed the cube method that allows the selection of simple or weighted balanced (or approximately balanced) samples for a large set of auxiliary variables. It is worthwhile remarking that enlarging the number of the balancing equations increases the difficulty to select a sample satisfying the whole set of balancing equations, leading to a final not balanced PTS.

Thirdly, the allocation of the sample is also important. We choose to fix the sample size of each imputation cells by means of a multi-domain algorithm based on Neyman method using as driving variables the number of employed persons (Bethel, 1989).

5. The procedure for building the final pseudo-sample

In order to carry out the simulation study, an artificial sample based on the 2004 observed final sample has been arranged. The main aim of the artificial sample is to make available the complete set of data, in terms of target variables and covariates, for all the 7,448 units in the final sample.

That allows to compute final estimate not affected by the non response problem.

The artificial sample has been created by means of an imputation method (ONS, 2001) which essentially uses the actual survey respondents, randomly selected from imputation classes. Therefore, the method divides the survey data into imputation groups, defined on the basis of categorical individual covariates. The imputation procedure is completely different from that described in section 3. The artificial sample is created within an

imputation group by sampling with replacement from the survey values. This method wants to preserve, as much as possible, the cross distribution of the turnover of the current year and of the previous year, distinctly for each month and for the categories of main group of product sold variable. The final estimates computed on the pseudo-sample represent the benchmark for the strategy evaluations.

The operations leading to the creation of the artificial sample have been various and complex, in the following they are gathered in two main steps.

Building the donor set: at the first step the units always responding (termed "totally respondent units") have been identified as the donor set. Due to the small number of totally respondent units, also units that didn't respond in just one month out of twelve have been included in the donor set. Those units needed that one monthly turnover value was imputed. In order to do that, in both the current and the previous years, firstly the monthly profile distributions of the totally-respondent-unit turnover have been considered. Secondly the linear regression model between the turnover value coming from the survey and that coming from the ASIA archive was studied for the totally respondent units, distinguishing between small and large enterprises, in terms of turnover. Finally the monthly turnover has been imputed in two steps: at first step the yearly turnover has been assigned on the basis of its relationship with the ASIA turnover established by means of the linear regression and at the second step the missing month item has been imputed according to the mean of monthly profile distribution of the turnover of the totally respondent units.

Imputing the turnover variables: the second step for the creation of the artificial sample started with the evaluation of some relationship among the key variables related to the turnover on the donor set: the linear regression models between the turnover of the previous year and the turnover of the current year and between the turnover of the current year and the turnover coming from the ASIA archive have been studied. Those relationships have been exploited for the imputation of the turnover of the previous and of the current years on the units with missing items. The imputation of the monthly turnover instead has been achieved by means of the random replication of the average donor monthly profile distributions in the main group of product sold variable.

6. Results

The property of the proposed sampling strategies have been studied in a simulative context. For each strategy 500 PTS, each one with 1,920 units as recommended by Eurostat (2001), have been selected from the 2004 pseudo-sample. At each iteration the provisional estimates on the domains of estimate are computed and the revision with respect to the final pseudo-sample estimates are calculated.

Till September 2004, the survey collects the quick response just for a part of the final sample, so a procedure for assigning the units to the UPOS has been required for the first nine months of the year 2004. Then, monthly un-planned early response probabilities have been created. A non-parametric classification tree technique (CART, Breiman et al., 1984) has been adopted, considering the following covariates: a dummy variable indicating that the unit came from the previous year panel, the number of employed persons grouped in class, the turnover coming from the ASIA archive, the late response frequencies in the year, the early response frequencies in the last 3 months, the turnover coming form the previous year survey, the main group of products sold.

The resulting UPOS monthly mean sample size is about 2,340 units, with a maximum value equal to 2,607 units and minimum value equal to 2,068. Table 1 shows the monthly sampling distribution.

Table 1 - Monthly UPOS dimensions

Month	1	2	3	4	5	6	7	8	9	10
Sample size	2,112	2,302	2,275	2,385	2,384	2,482	2,348	2,332	2,607	2,068

The increasing of the number of quick respondents has been adopted as a tool for dealing with the provisional estimation problem. The consequences of this strategy can be observed in the last three months, when the quick response indicator variables are available and the mean of the ratio between the number of quick and late respondents is about the 72%.

The experiment compares the results coming from the proposed sampling strategies, hereinafter the balanced strategies, with both the estimates based on the 2004 UPOS and the estimates obtained with the samples of 1,920 units of the largest retail enterprises in terms of turnover or number of employed persons. The last two samples are allocated according to the same technique defining the allocations of the balanced PTS. These three strategies represent the benchmark of the balanced strategies. As far as the balanced strategies are concerned, let us show the results of four combinations. The first one, hereinafter denoted as *Simple Balanced* (SB) strategy, uses the estimator based on the model (4.2) and the sampling design satisfying the simple balancing equations on the functions of the employed persons, $a^{1/2}$, a, a^2. The second strategy adds to the SB strategy the balancing equations on the functions of the turnover, f, f^2. Let us indicate this strategy as *Simple Full Balanced* (SFB) strategy. The third and fourth strategies use the estimator based on the model (4.6). According to the equations (4.7), the design balances on the $a^{1/2}$, a, a^2 producing the *Weighted Balanced* (WB) strategy, while in the final strategy, denoted as *Weighted Full Balanced* (WFB) strategy, also the balancing equations on the functions of the turnover f, f^2 must be held.

For evaluating the performances in term of revision, the monthly *Mean Percentage Revision* (MPR) has been computed according to the expression

$$MPR_D^{t,0} = \sum_{d \in D} \left[\left(\frac{\bar{\tilde{I}}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right) \times 100 \right] \gamma_d , \qquad (6.1)$$

where $\hat{I}_d^{t,0}$ is the final estimate on the more disaggregate domain *d* (Large-non food; Small-non food; Large-food; Small-food) at month *t* and $\bar{\tilde{I}}_d^{t,0}$ assumes one of the following values:

- $(1/500)\sum_r \tilde{I}_{d,r}^{t,0}$ with the balanced strategies, being $\tilde{I}_{d,r}^{t,0}$ the provisional estimate on the domain d in the r-th replication,

$\bar{\tilde{I}}_d^{t,0} \equiv \tilde{I}_d^{t,0}$ considering the benchmark strategies.

Finally, *D* indicates the generic domain at the more disaggregate level ($d \equiv D$) or at

aggregate level (Non food, Food, Large, Small, Total) and $\gamma_d = \sum_{h \in d} \gamma_h$.

The yearly version of the (6.1) is given by

$$MPR_D = \frac{1}{12} \sum_{t=13}^{24} MPR_D^{t,0} . \qquad (6.2)$$

A second type of indicators measures the variability of the estimates by means of the *Mean Absolute Percentage Revision* (MAPR). At monthly level it is defined by,

$$MAPR_D^{t,0} = \sum_{d \in D} \left[\left| \frac{\tilde{\bar{I}}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right| \times 100 \right] \gamma_d . \qquad (6.3)$$

Let us prefer to use the expression (6.3) for the balanced strategies instead of a more appropriate indicator using the term $(1/500)\sum_r \left| \left(\tilde{I}_{d,r}^{t,0} - \hat{I}_d^{t,0} \right) / \hat{I}_d^{t,0} \right| \times 100$ in the square brackets, since we observed only one preliminary sample for the benchmark strategies. Therefore, in the balanced strategies this alternative indicator catches the variability due to the iterations, not detectable in the benchmark strategies. The yearly MAPR is

$$MAPR_D = \frac{1}{12} \sum_{t=13}^{24} MAPR_D^{t,0} . \qquad (6.4)$$

We point out that the monthly MPR and MAPR give rough measures especially for the benchmark strategies because they are computed for few values and with only one value for the more disaggregate domains. Hence, we mainly show the results of the statistics (6.2) and (6.4).

Table 2.a shows the values of the statistics (6.2) for the preliminary domain of estimates given by crossing each other the variable type of sold product (food and non-food retail enterprises) and type of distribution (large and small retail enterprises).

Table 2.a - Yearly *Mean Percentage Revision* (MPR) by Type of sold product and Type of distribution domains

Method	Large-non food	Small-non food	Large-food	Small-food
Strategy using UPOS	0.674	0.236	0.505	0.021
Largest Units in terms of Employed Persons (LUEP) strategy	1.606	-0.139	-0.070	-0.592
Largest Units in terms of Turnover (LUT) strategy	0.977	0.126	0.359	-0.309
Simple Balanced (SB) strategy	0.555	0.006	0.733	-0.241
Simple Full Balanced (SFB) strategy	0.504	-0.014	0.757	-0.266
Weighted Balanced (WB) strategy	0.639	-0.077	0.197	-0.303
Weighted Full Balanced (WFB) strategy	0.638	0.001	0.190	-0.239

The table underlines that the balanced approaches using a PTS have, in general, better performances than the benchmark strategies. The WFB strategy seems to be the best strategy. Just in only two cases the benchmark strategies present a MPR less than the WFB strategy: for large-food domain the *Largest Units in terms of Employed Persons* (LUEP) strategy has a MPR=-0.070, while the WFB strategy has a MPR=0.190, and for the small-food domain where the strategy based on UPOS with MPR=0.021 has a value closer to zero than the -0.239 of the WFB strategy. The strategies involving model (4.2) and simple

balanced equations, have good performances except for large-food domain. For the SB strategy the value is 0.733 while for the SFB the MPR=0.757.

Table 2.b shows the MPR results for the aggregate domains. The findings must be analyzed with caution because the opposite signs of the MPR at more disaggregate levels. "Good" results could actually hide an unstable strategy in term of unbiasness and this aspect must be taken into account in the conclusive evaluations. The WB strategy is the best method based on PTS, except for the case of small type of distribution, where the MPR=-0.109, while the WFB strategy has a MPR=-0.034, the SB strategy has a MPR=-0.029 and the SFB strategy achieves a MPR=-0.050. In the large domain, LUEP strategy is slightly better. Finally for the total domain the strategy observing the sample of the largest units in terms of employed persons has the best MPR with a value equal to -0.021. The WB has a MPR=0.043 and the WFB strategy a value equal to 0.087.

Table 2.b - Yearly *Mean Percentage Revision* (MPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non food	Food	Large	Small	
Strategy using UPOS	0.293	0.396	0.540	0.205	0.334
Largest Units in terms of Employed Persons (LUEP) strategy	0.087	-0.188	0.272	-0.205	-0.021
Largest Units in terms of Turnover (LUT) strategy	0.236	0.209	0.486	0.063	0.225
Simple Balanced (SB) strategy	0.078	0.514	0.697	-0.029	0.250
Simple Full Balanced (SFB) strategy	0.053	0.527	0.705	-0.050	0.240
Weighted Balanced (WB) strategy	0.016	0.084	0.287	-0.109	0.043
Weighted Full Balanced (WFB) strategy	0.083	0.094	0.282	-0.034	0.087

Table 3.a gives some findings about the variability of the compared strategies, computed by (6.4). The methods based on balanced PTS seem competitive and, in particular, the strategies identified by the model (4.6) and balanced equation (4.7) are more stable than the strategies identified by the model (4.2) and the simple balanced equations with respect to the benchmark strategies. The SB and SFB strategies have the best MAPR for large-non food domain (about 0.9) but high values for the large-food domain (about 0.8) with respect to the benchmark strategies. Largest units strategies have the best results for small-non food domain with 1.001, if we consider the sample of *Largest Units in term of Turnover* (LUT), and 1.143 when a sample of largest enterprises in term of employed persons is selected. The last strategy has the best performance for large-food domain as well (0.317). When the strategies based on weighted balanced equation (4.7) have the worse results than largest units strategies, the values are close each other anyway. For instance, from one hand, the WFB strategy has a MAPR equal to 1.305 for small-non food domain and equal to 0.391 large-food domain. On the other hand, the performances on large-non food and small-food are quite better than the benchmark strategies. Strategy based on UPOS, despite the greatest mean overall sample size, does not perform very well at least with respect to the balanced PTS strategies. Just for small-non food domain, the MAPR=1.336 while the SB has a MAPR=1.369, and the SFB strategy has a value equal to 1.339.

Table. 3.a - Yearly *Mean Absolute Percentage Revision* (MAPR) by Type of sold product and Type of distribution domains

Method	Large-non food	Small-non food	Large-food	Small-food
Strategy using UPOS	1.493	1.336	1.093	2.091
Largest Units in terms of Employed Persons (LUEP) strategy	2.263	1.143	0.317	2.949
Largest Units in terms of Turnover (LUT) strategy	2.461	1.001	0.587	2.344
Simple Balanced (SB) strategy	0.998	1.369	0.840	2.038
Simple Full Balanced (SFB) strategy	0.931	1.339	0.864	2.057
Weighted Balanced (WB) strategy	1.118	1.322	0.392	2.040
Weighted Full Balanced (WFB) strategy	1.132	1.305	0.391	2.040

For the aggregate domains (table 3.b) the use of balanced PTS still leads to the best MAPR values. In some rare cases the largest units strategies achieve values lower than strategies based on model (4-2): for non food domain where the LUT strategy is better than the SB strategy (1.191 vs 1.195); for large domain where LUEP strategy (0.715) is better than the simple balanced approaches, with MAPR values greater than 0.770.

The remaining results show that the weighted balanced strategies perform better than the other and they are enough similar. The greatest difference occurs in the non-food and small domain where the WFB strategy has respectively values equal to 1.154 and 0.827, while the other balanced weighted strategy presents for the same domains respectively the values of 1.168 e 0.840.

Table 3.b - Yearly *Mean Absolute Percentage Revision* (MAPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non food	Food	Large	Small	
Strategy using UPOS	1.356	1.317	1.175	1.444	1.341
Largest Units in terms of Employed Persons (LUEP) strategy	1.288	0.909	0.715	1.403	1.139
Largest Units in terms of Turnover (LUT) strategy	1.191	0.982	0.970	1.195	1.109
Simple Balanced (SB) strategy	1.195	0.685	0.771	0.881	0.618
Simple Full Balanced (SFB) strategy	1.164	0.714	0.780	0.861	0.621
Weighted Balanced (WB) strategy	1.168	0.659	0.467	0.840	0.506
Weighted Full Balanced (WFB) strategy	1.154	0.662	0.469	0.827	0.507

The general yearly level conclusions are confirmed by examining the monthly level statistical indicators. Figure 1 shows the monthly MPR for the domain total. For clearness we have plotted the current strategy based on UPOS, the LUT strategy, the SFB and WFB strategies. We can see that the MPR of balanced PTS strategies are the lowest or, when not, they are between the LUT and the UPOS strategies. The figure gives rough suggestion of the relationships among balanced PTS and benchmark strategies. In fact the graph of benchmark strategies are based on just one observation per month.

Figure 2 describes the trails of the monthly variability among the two full balanced strategies, the using UPOS and the LUT strategies by means of the (6.3). The graph points out that the balanced approaches report almost always the lowest values.

Figure 1 - Monthly *Mean Percentage Revision* (MPR) for the simple and weighted full balanced strategies and some benchmark strategies for the Total domain

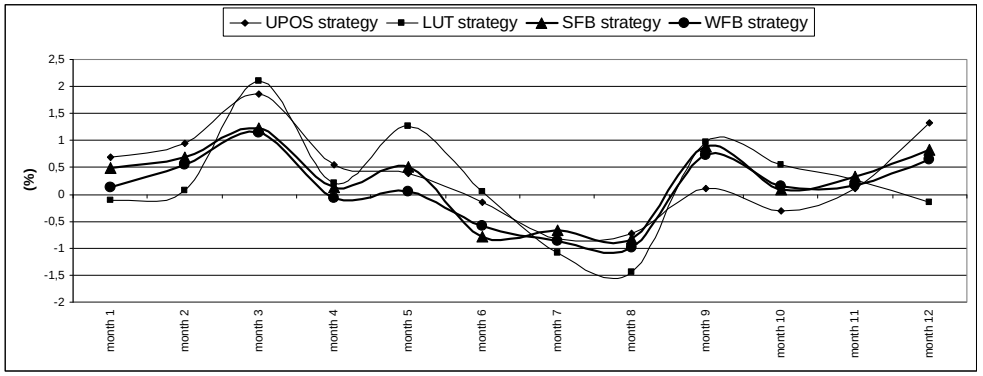
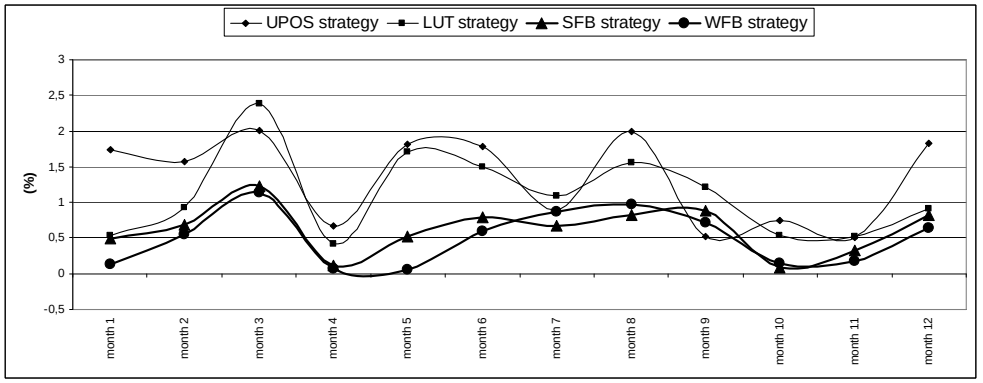


Figure 2 - Monthly *Mean Absolute Percentage Revision* (MAPR) for the simple and weighted full balanced strategies and some benchmark strategies for the Total domain



Finally, we have examined the MPR and MAPR of the WFB strategy with sample sizes equal to 500, 1,000, 1,500 units. Figure 3 and 4 depict the trend of the yearly MPR for four sample sizes (including the sample with 1,920 units). The graphs plot the values of the total domain (figure 3), and of the type of sold product and the type of distribution domains (figure 4). The dashed line represents the value obtained by the UPOS. The robustness of the balanced strategies in term of revision is confirmed, except when the size is equal to 500 units. In this case, the revision is higher than the observed benchmark value.

For the total domain, figure 5 shows the monthly MPR of the strategy based on the UPOS and of the four WFB strategies according to the sample size. The estimates computed with the UPOS data perform substantially better in the months 6, 8, 9, and slightly better in the month 11, when the sample sizes are very large (at least for the month 6, 9 and 11).

Figure 3 - Yearly *Mean Percentage Revision* (MPR) for the weighted full balanced strategies with different sample sizes for the Total domain

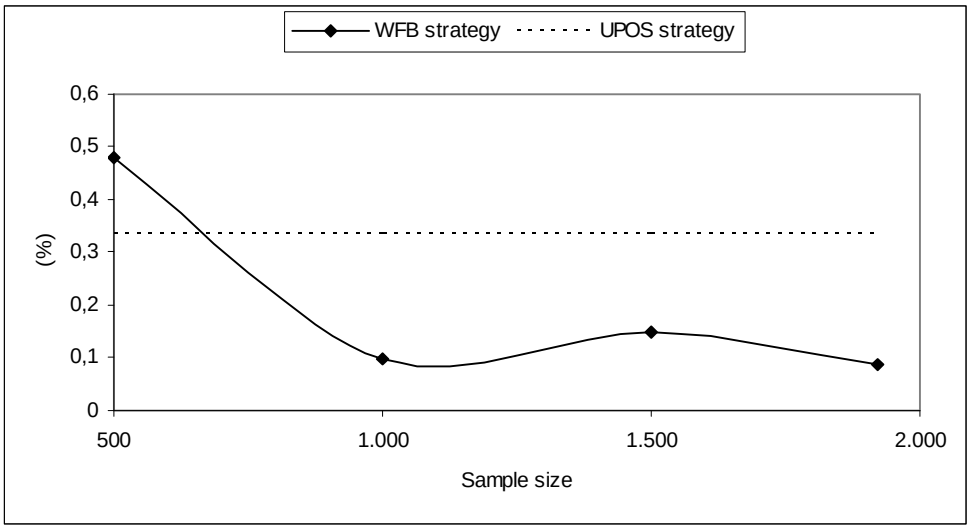


Figure 4 - Yearly *Mean Percentage Revision* (MPR) for the weighted full balanced strategies with different sample sizes for the type of sold product and type of distributions domains

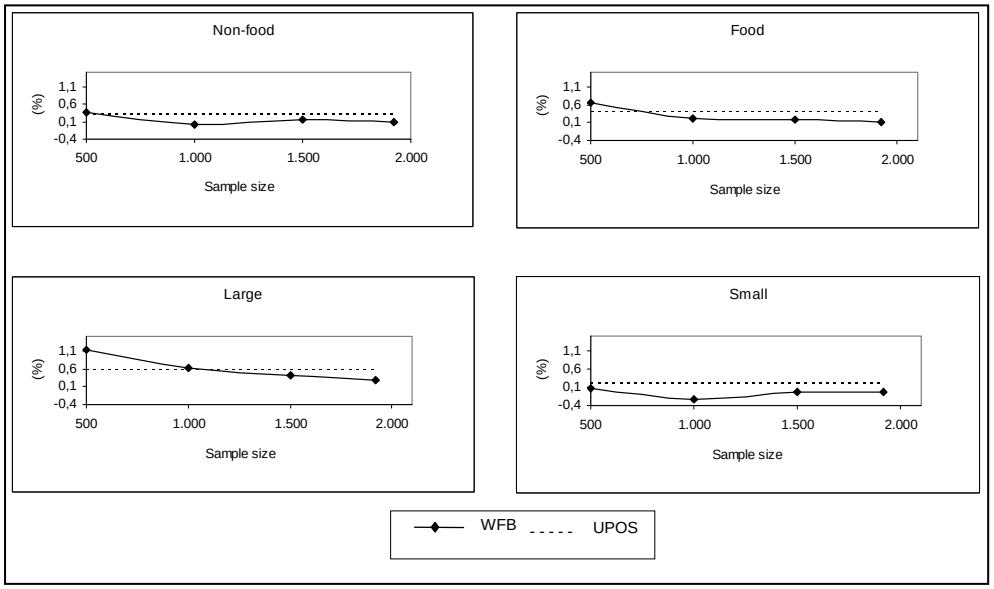
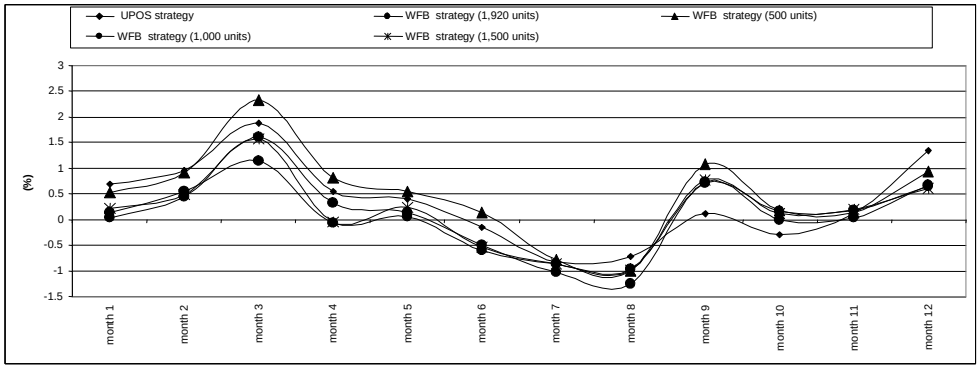


Figure 5 - Monthly *Mean Percentage Revision* (MPR) for the weighted full balanced strategies with different sample sizes and the strategy using UPOS for the Total domain



The analysis of MAPR leads to some different conclusions. In this case the sample dimension has a strong impact on the variability of the preliminary estimates. Besides the WFB strategy with 1,920 units, only the strategy with 1,500 units appears to be competitive with respect to the strategy based on UPOS. Figure 6 and 7 show the trend of the yearly MAPR related to the examined sample sizes for the total domain (figure 6) and for type of sold product and type of distribution domains (figure 7).

Figure 6 - Yearly *Mean Absolute Percentage Revision* (MAPR) for the weighted full balanced strategies with different sample sizes for the Total domain

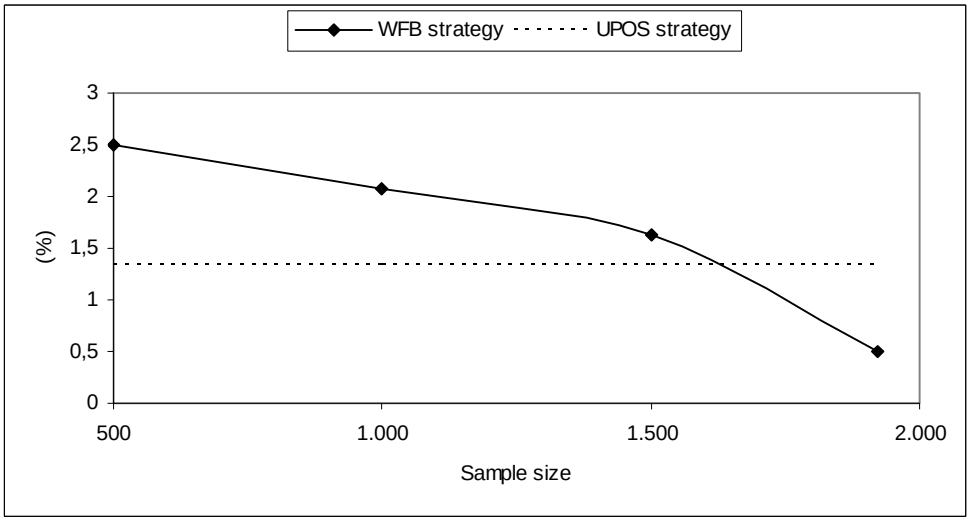


Figure 7 - Yearly *Mean Absolute Percentage Revision* (MAPR) for the weighted full balanced strategies with different sample sizes for the type of sold product and type of distributions domains

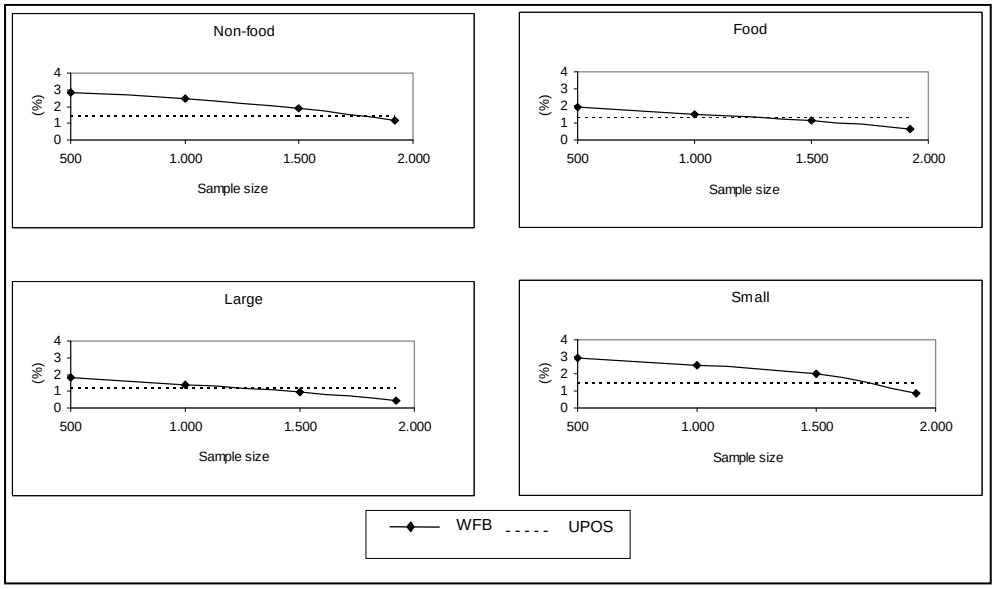
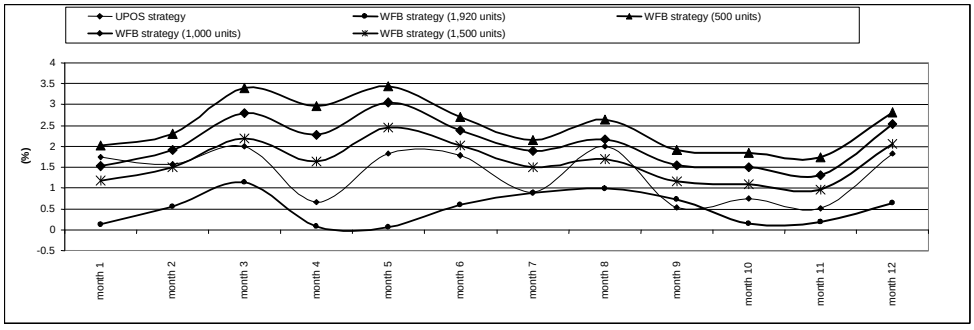


Figure 8 shows the monthly MAPR trail, underlying that the variability of the preliminary estimates by means of the weighted full balanced sample with 1,500 units is quite close to the variability of the estimates using the UPOS.

Figure 8 - Monthly *Mean Absolute Percentage Revision* (MAPR) for the weighted full balanced strategies with different sample sizes and the strategy using UPOS for the Total domain



7. Discussion

The balanced approach as proposed in this paper represents an innovative way to deal with the provisional estimation problem. The ordinary or standard approaches face this problem making main efforts in the estimation phase sphere. In this work, we try to develop

an overall sampling strategy, showing that the exploiting of the auxiliary information at the sample design phase allows a basic improvement on the provisional estimates quality.

The paper deals with the problem of provisional estimates from the preliminary sampling design view-point. The question is if the definition of PTS leads to a real improvement of the provisional estimates in terms of revision and variability. When it happens, this enhancement may be evaluated, compared and weighted with the costs of a possible changing of the data collection process, to obtain a PPOS as close to PTS as possible.

The use of the balanced sampling theory to define a PTS takes into account the form of the provisional and final adopted estimators. In this way, an overall sampling strategy has been developed and it constitutes an original approach in provisional estimate field. The properties (the efficiency gain and robustness of the estimates) of the sampling strategies using balanced sample, have been proved when the parameters of interest are totals. Nevertheless, in the short-term business surveys, in which the timeliness is a pressing target, many parameters of interest are defined by a non linear function of totals, such as a ratio. For these sort of parameters few is known in the scientific communities (Valliant, 1999). Therefore, the unbiasedness and efficiency of the sampling strategies have been studied in a simulative context. The planning of sampling strategy is up to the particular survey frame. In the paper the balanced samples have been arranged to be suitable to the Italian Monthly Retail Trade Survey. The main survey parameter is the rate of change of the current amount of the turnover with respect to the same amount referred to a previous time. The survey produces provisional estimates at domain level employing a complex procedure. Given these elements we have planned different kinds of simple and weighted balanced sample, modifying, in same cases, the current provisional estimation process. For the provisional estimator, the unbiasedness of the estimate of the previous year total turnover amount improves the estimate of the current year total. This point has been the compass for sample design definition.

The simulation study has shown that the sampling strategies based on a balanced PTS lead to the best performances, in term of revision and variability, with respect to a strategy depending on the UPOS. Only in some isolated domains of estimate, the last strategy allows to achieve preliminary estimates less unbiased or with a minor sampling errors. Moreover, such results are obtained with a sample size dimension of about 20% larger than of PTS.

The simulation study has compared the proposed strategies with some strategies based on the sample of the largest units as well. When the target of the estimate is a total, these kinds of sample have “optimal” properties for specific superpopulation models, but they are extremely vulnerable to bias when the working model is incorrect.

The comparison among largest units PTS and balanced PTS leads to prefer the last ones. The first strategies show to be competitive just for only one combining domain (large-food). For the rest of the domains balanced sample approaches are preferable.

It is also interesting to discuss about the comparison of the four preliminary balanced sampling design analyzed in the experiment. We may distinguish in a couple based on a simple balancing equations and a couple based on weighting balanced equations.

The two categories underlie two kinds of provisional estimator, and for the second category these estimators are slightly different from the current MRTS provisional estimator. When the estimate is a total, the scientific communities do not support one strategy rather than another, even if experimental results suggest that the weighted balanced approach could be better (Valliant et al., 2000). In our simulation the simple balanced approaches seem competitive with respect to the weighted version for many domains but not for all. In particular, for one combining domain (large-food) the simple balanced

approaches obtain very bad values of revision and variability indicators.

However, the results suggest the possibility to analyze a more complex sampling strategy combining the two kinds of balanced sampling. For some of the most disaggregate domains we may use weighted balanced sample and in some other the simple balanced sample. That means the changing of the working superpopulation model that will be defined at the domain level and no more at the cell imputation level.

As far as the comparison of the approaches belonging to the same category of balanced sampling is concerned, no evident difference does exist, even if for particular domains the values of the analyzed statistics are not similar.

Finally, the analysis has investigated the empirical properties of a weighted full balanced sampling approach with respect to the sampling dimension. The results stress that the balanced samples bring to a revision better than the UPOS, even when the sample sizes are smaller than about the 50%. In terms of variability, the sampling dimension produces a more considerable impact, and the balanced samples with at least 1,500 units appear competitive with respect to the strategy using UPOS.

The final conclusion rising from the empirical results points out the adaptability of balanced sampling theory to the complex estimator structure, allowing to improve the provisional estimation quality and to reduce the preliminary sample size, by means of a not-yet-exploited statistical tool.

In this paper, as already underlined, the preliminary estimation issue is approached focusing on the sample stage; future work requires also deeper evaluation of the estimation phase. In fact, a specific model based estimator could be considered to take into account in a more appropriate way the time series data. A further research field could directly investigate the modelization of the ratio instead of the total composing the variation index. In this sense, that study can develop the model assisted approach proposed by Valliant (1999).

References

Bethel J. (1989), "Sample Allocation in Multivariate Surveys", *Survey Methodology*, vol.15, pp. 47-57.

Breiman L., Friedamn J. H., Olshen R. A., Stone C. J. (1984), *Classification and Regression Tree,* Chapman and Hall, New York.

De Sandro L, Gismondi R. (2004), **"**Provisional Estimation of the Italian Monthly Retail Trade Index", *Contributi-Istat*, 24/2004.

Deville J.-C., Tillé Y. (2004), "Efficient Balanced Sampling: the Cube Method", *Biometrika*, vol. 91, pp. 893-912.

EUROSTAT (2000), *Short-term Statistics Manual*, Eurostat, Luxembourg.

EUROSTAT (2001), "Conclusion of the First Meeting of the Export Group Contro-Stratified European Sample for Retail Trade", *Final Report*, July 2001, Eurostat, Luxemburg.

EUROSTAT (2005), *Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version*, Eurostat, Luxemburg.

ISTAT (2007), *Rapporto Finale del gruppo di lavoro: "*Sperimentazione di stime anticipate per specifici indicatori congiunturali, finalizzata al rilascio in produzione delle relative metodologie*"*, Direttiva Istat TRAC16, Technical Report.

ONS (2001), *Artificial Methods for Completing the Population Data Sets*, EURAREA, WP1.2 D1.2.2.

Royall R.M., Herson J. (1973), "Robust Estimation in Finite Population", *Journal of the American Statistical Association*, vol. 73, pp. 66-77.

Royall R.M. (1992), "Robustness and Optimal Design Under Prediction Models for Finite Populations", *Survey Methodology*, vol. 18, pp. 179-185.

Valliant R. (1999), "Uses of Models in the Estimation of Price Indexes: a Review", *In Proceedings of the Survey Research Methods Section, American Statistical Association*.

Valliant R., Dorfman A. H., Royall R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.

RELAIS: An Open Source Toolkit for Record Linkage

Nicoletta Cibella[1], Marco Fortini[2], Monica Scannapieco[3], Laura Tosco[4], Tiziana Tuoto[5]

Abstract

The combined use of statistical and administrative sources allow to save time and money, reducing survey costs, response burden, etc.; sometimes data sources are hard to integrate since errors or lacking information in the record identifiers may complicate this process. The purpose of record linkage is to identify the same real world entity, which can be differently represented in data sources. To deal with record linkage complexity and application dependency, we propose a toolkit called RELAIS (REcord Linkage At IStat). The toolkit is based on the idea of choosing the most appropriate technique for each phase, and of dynamically combining such techniques in order to build a workflow, on the basis of application constraints and data features provided as input. RELAIS is configured as an open source project giving the possibility of gathering together the efforts already done in the scientific community towards the definition of a record linkage project.
A real case study validates the RELAIS idea.

Key words: record linkage, open source software

1. Introduction

Record linkage is a process that aims to identify if two (or more) records represent (or not) the same real world entity. A record linkage project can be performed for different purposes and it has recently revealed a powerful support to decisions in large commercial organizations and government institutions.

In official statistics, the field in which this work is developed, the combined use of statistical survey and administrative data is largely widespread; this does stimulate the investigation of new methodologies and instruments to deal with record linkage projects.

Since the earliest contributions to modern record linkage, dated back to Newcombe et al (1959) and to Fellegi and Sunter (1969), there has been a proliferation of different approaches, that make use of techniques based on data mining, machine learning, equational theory. However no particular record linkage technique has emerged as the best solution for all cases. We believe that such a solution does not actually exist, and that an alternative strategy should be adopted.

Record linkage can be seen as a complex process consisting of several distinct phases

[1] Collaboratore T.E.R. (Istat), e-mail: cibella@istat.it
[2] Primo ricercatore (Istat), e-mail: fortini@istat.it
[3] Tecnologo (Istat), e-mail: scannapi@istat.it
[4] Collaboratore T.E.R. (Istat), e-mail: tosco@istat.it
[5] Ricercatore (Istat), e-mail: tuoto@istat.it

involving different knowledge areas, moreover several different techniques can be adopted for each phase. We consider that the choice of the most appropriate technique not only depends on the practitioner's skill but it is also application specific. In some applications, there are not evidences to prefer a given method to others; in addition, from the analyst's point of view, it is important to have the possibility to experiment the alternative criteria and parameters.

In this paper we describe the RELAIS (REcord Linkage At IStat) toolkit; it allows the combination of different techniques proposed for each of the record linkage phases; therefore the resulting workflow is actually built on the basis of application and data specific requirements Fortini et al (2006), Tuoto et al (2007). Moreover, the RELAIS project will include not only a set of techniques, but also a library of *patterns* that, given specific data and application requirements, could support users in the definition of the most appropriate record linkage workflow for their data and application requirements.

RELAIS is an open source project, this choice is based on the idea of re-using the solutions already available in the scientific community.

The major contributions of this paper can be summarized as follows. First we outline in detail the philosophy and the purposes of RELAIS, then we describe the phases that compose a record linkage project, finally we illustrate the idea of a dynamic record linkage workflow as implemented in RELAIS by means of a real case study in which a record linkage workflow is instantiated starting from data and application requirements.

2. RELAIS: a Toolkit for Building Record Linkage Workflows

In official statistics, record linkage projects are particularly important to permit the combined use of statistical survey and administrative data. At present, many potential advantages in using administrative data for statistical purposes are known and shared by the various national statistical institutes: in fact, administrative sources usually contain larger amounts of data, possibly more accurate due to improvements over time, so the joint analysis of two or more statistical and administrative sources often allows to save time and money, reducing survey costs, response burden, etc. Indeed, cooperation among different public agencies or institutes is actually based on common data sharing, that prevents from recollecting data from citizens or enterprises, if such data are already available at some of the public subjects. The different uses of record linkage in official statistics include: update and de-duplication, when multiple records referring to the same real world entity are stored within one single data source; data integration, across multiple data sources in order to provide a reconciled global record; correction across multiple data sources, performed when one source is known to have higher quality data that can be used for improving the others; measure of a population by capture-recapture for instance on the occasion of the post-enumeration survey of Census; check of the confidentiality of public-use microdata, through re-identification experiments.

However, data sources are often hard to combine since errors or lacking information in the record identifiers may complicate the integrated use of the information; in order to overcome such obstacles, record linkage techniques provide multidisciplinary set of methods and practices whose purpose is to identify the same real world entity, which can be differently represented in one or more data sources.

The general and formal definition of record linkage dates back to Fellegi and Sunter (1969). They approached the problem via a probabilistic decision model which is still

widely used; recently, different approaches from the computer science field have been proposed, based on data mining, machine learning, equational theory (see Hernandez and Stolfo 1998, Monge and Elkan 1997, Ananthakrishna et al. 2002, Chauduri et al. 2005). These approaches can be classified as empirical in contrast with the strictly probabilistic record linkage procedures, which, following Fellegi-Sunter, make explicitly use of probabilities. Anyway there is no evidence that a specific record linkage technique can perfectly solve all matching problems.

Due to the great attention to the integration problem and to its complexity, several record linkage systems and tools have been proposed, in both the academic and private sectors. Such tools include, for example, Big Match (Yancey, 2007), CANLINK (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), Tailor (Elfeky et al. 2002), The Link King (http://www.the-link-king.com). The first two systems have been developed at the U.S. Bureau of the Census and the Statistics Canada respectively, Tailor is an academic prototype, while others have been developed at medical-epidemiological Centres or at Universities. Some of the systems provide for the user a certain degree of flexibility, e.g. Febrl allows to choose which comparison function can be more appropriately applied. However, any of these tools provides the flexibility of multiple choices for each of the record linkage phase. Moreover, we want to realize the idea of dynamically building a record linkage workflow, as a result of a combination of the most appropriate technique selected at each phase. In this respect, the Tailor system is the closest one to our idea of a toolkit. However, Tailor only offers, in some of the record linkage phases, a (limited) list of methods that can be applied, without suggesting their dynamic composition based on application needs. Indeed, the purpose of Tailor is to come up with the best solution for record linkage, and therefore an experimental comparison was performed among techniques within each phase.

On the basis of the complexity and the modularity of the record linkage problem, its strong dependence on data and application requirements, and the large number of efforts made in different fields in order to deal with the linkage issues, we propose the RELAIS toolkit as a tool that guides in building record linkage workflows. The inspiring principle is to allow combining the most convenient techniques for each of the record linkage phases and also to provide a library of patterns that could support the definition of the most appropriate workflow, in both cases taking into account the specific features of the data and the requirements of the current application. In such a way, the toolkit not only provides a set of different techniques to face each phase of the linkage problem, but also it could be seen as a compass to solve the linkage problem as better as possible given the problem constrains. In addition, RELAIS aims at joining specifically the statistical and computational essences of the matching issue. Moreover, in order to re-use the several solutions already available for record linkage in the scientific community and to gain the several experiences in different fields, we start to develop the RELAIS project as an open source project, by the quite ambitious goal of providing, in the shortest possible time, a generalized toolkit for dynamically building record linkage workflows.

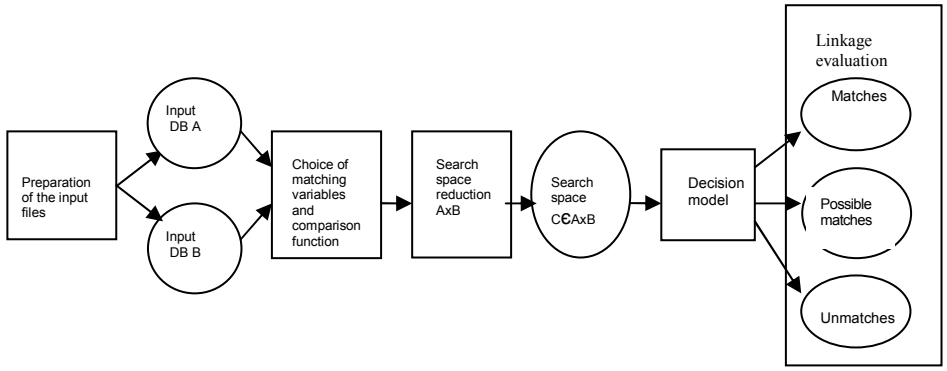
3. The Phases of a Record Linkage Project

The complexity of a linking process relies on several aspects. If unique identifiers are available in the data sources the problem can be quite easily treated but unique identifiers are not always available. The lack of unique identifiers in the datasets at hand requires more sophisticated statistical procedures relying on "matching variables" chosen for linking data. Obviously, errors in the linking variables such as missing words, variations in spelling, use of initials, etc., may invalidate the linkage results, thus a big effort for reducing such errors is necessary to prepare input files.

A record linkage procedure is composed of some main phases (as shown in Fig. 1):

- Data cleaning - preparation of the input files (pre-processing);
- Choice of the common identifying attributes (matching variables);
- Choice of the comparison function;
- Search space reduction;
- Choice of a decision model;
- Record linkage procedures evaluation.

Figure 1 - Phases of record linkage



The *preparation of input files* is the first phase which, according to Gill (2001), requires 75% of the whole effort to implement a record linkage procedure, in fact data can be recorded in different formats, some items may be missing or with inconsistency or errors. The key job of this phase is to convert the input data in a defined format, resolving the inconsistencies in order to reduce errors deriving from an incorrect reported data. In this phase null strings are cancelled, abbreviations, punctuation marks, upper/lower cases, etc. are cleaned and any necessary transformation is carried out to standardize variables. Furthermore the spelling variations are replaced with standard spelling for the common words.

After the previous phase, it is important to *choose matching variables* which are as suitable as possible for the considered linking process. The matching attributes are generally chosen by a domain expert. If unique identifiers are available in the linkable data sources, the easiest and most efficient way is to use these ones as link variables; but very strict controls need to be made in case of using numeric identifiers alone. Variables like

name, surname, address, date of birth, can be used jointly instead of using each of them separately; in such a way, one can overcome problems like the wide variations of the *name* spelling or the changes in *surname* depending on the variability of the marital status. It is evident that the more heterogeneous are the items of a variable, the higher is its identification power; moreover, if missing cases are relevant in a field it is not useful to choose it as a matching variable.

C*omparison functions* are used to compute the distance between records compared on the values of the chosen matching variables. Some of the most common comparison functions are:

- *equality* that returns 1 if two strings fully agree, 0 otherwise;
- *edit distance* that returns the minimum cost in terms of insertion, deletions and substitutions needed to transform a string of one record into the corresponding string of the compared record;
- *Jaro* counts the number of common characters and the number of transpositions of characters (same character with a different position in the string) between two strings;
- *Hamming Distance* that computes the number of different digits between two numbers;
- *Smith-Waterman* that uses dynamic programming to find the minimum cost to convert one string into the corresponding string of the compared record; the parameters of this algorithm are the insertions cost, deletions cost and transposition cost;

TF-IDF that is used to match strings in a document. It assigns high weights to frequent tokens in the document and low weights to tokens that are also frequent in other documents.

For a reviews of comparison functions see Koudas N. and Srivastava D. (2005).

In a linking process of two datasets, say *A* and *B*, the pairs needed to be classified as matches, nonmatches and possible matches are those in the cross product *A x B*. If a de-duplication problem is considered the space is *A x (A-1)/2*. When dealing with large datasets, comparing all the pairs (*a; b*), *a* belonging to *A* and *b* belonging to *B*, in the cross product is almost impracticable, in fact while the number of possible matches increases linearly, the computational problem raises quadratically, the complexity is $\boldsymbol{O}(n^2)$(Christen and Goiser, 2005). To reduce this complexity, which is an obvious cause of problems for large databases, it is necessary to reduce the number of pairs (*a; b*) to be compared. There are many different techniques that can be applied to reduce the search space; blocking and sorted neighbourhood are the two main methods. *Blocking* consists of partitioning the two sets into blocks and of considering linkable only records within each block. The partition is made through blocking keys; two records belong to the same block if all the blocking keys are equal or if a hash function applied to the blocking keys of the two records gives the same result. *Sorted neighbourhood* sorts the two input files on a blocking key and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets.

Starting from the reduced search space, we can apply different decision models which define the rules used to determine whether a pair of records (*a; b*) is a match, a non-match or a possible match.

The core of record linkage process is the *choice of decision model* which enables to classify pairs into M, the set of true matches and U, the set of true non-matches. The decision rule can be empirical or probabilistic. A pair is a true match if it agrees completely on all the matching variables chosen or satisfies a defined rule-base system, that is if it reaches a score which is besides a threshold when applying the comparison function. The probabilistic approach, based on the Fellegi and Sunter model, requires an estimation of the

model parameters which can be performed via the EM algorithm, Bayesian methods, etc.

A linkage process can be also classified as: (i) one-to-one problem, if one record in the set A links to only one record in B and also the other way around, (ii) many-to-one problem if a record in a set can be matched with more than one of the compared file, (iii) many-to-many problem allows more than one record in each file to be matched with more than one record in the other. The latter two problems may imply the existence of duplicate records in the linkable data sources.

During a linkage project it is necessary to classify records as true link or true non link, minimizing the two types of possible errors, namely false matches and false non-matches. The first type of error refers to matched records which do not represent the same entity, while the latter indicates unmatched records not correctly classified, that imply truly matched entities were not linked. Generally, false non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them (Ding and Fienberg, 1994). The false match rate denotes the ratio between the records incorrectly matched and the whole number of matched pairs. The false non-match rate instead indicates the ratio between the number of incorrectly non matched records and the whole number of the correctly matched records. In general, it's not easy to find automatic procedures to estimate these two types of errors in order to evaluate the quality of record linkage procedures. They can be estimated via samples of units belonging to the M and U subsets or by means of a clerical reviewed sample units or of a re-linkage procedure, assumed error-free because performed with more accurate techniques; the bias is evaluated by means of the differences between the match and the "perfect match" results.

4. Description of RELAIS

The RELAIS toolkit idea is based on the consideration that the record linkage process is application dependent. Indeed, available tools do not provide a satisfying answer to the various requirements that different applications can exhibit. As seen in the previous section, the record linkage process consists of different phases; the implementation of each phase can be performed according to a specific technique or on the basis of a specific decision model. For instance, choosing which decision model to apply is not immediate: the usage of a probabilistic decision model can be more appropriate for some applications but it can be less appropriate for others, for which an empirical decision model could prove more successful. Furthermore, even using the same decision model, in different application scenarios, a comparison function could fit better than others. Therefore, we claim that no record linkage process, deriving from the choice and combination of a specific technique for each phase, is the best for all applications.

The RELAIS toolkit is composed by a collection of techniques for each phase of the record linkage procedure that can be dynamically combined in order to build the best record linkage workflow, given a set of application constraints and data features provided as input (see Figure 2). As an example, if it is known that the datasets to compare have poor quality, it is suitable the usage of comparison functions ensuring high precision (e.g. Jaro distance, as defined in the previous section); as a further example, if no specific error-rates are required by the application, it can be appropriate the usage of an empirical decision model. Some phases of the record linkage process can be missing: for instance the search space reduction phase makes sense only for huge data volumes, or for applications that have time constraints. In Figure 3, examples of possible workflows that may be built with the RELAIS toolkit are shown.

Figure 2 - RELAIS input-output

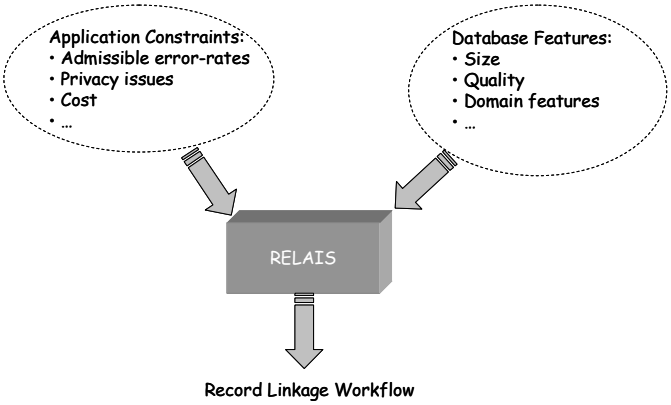
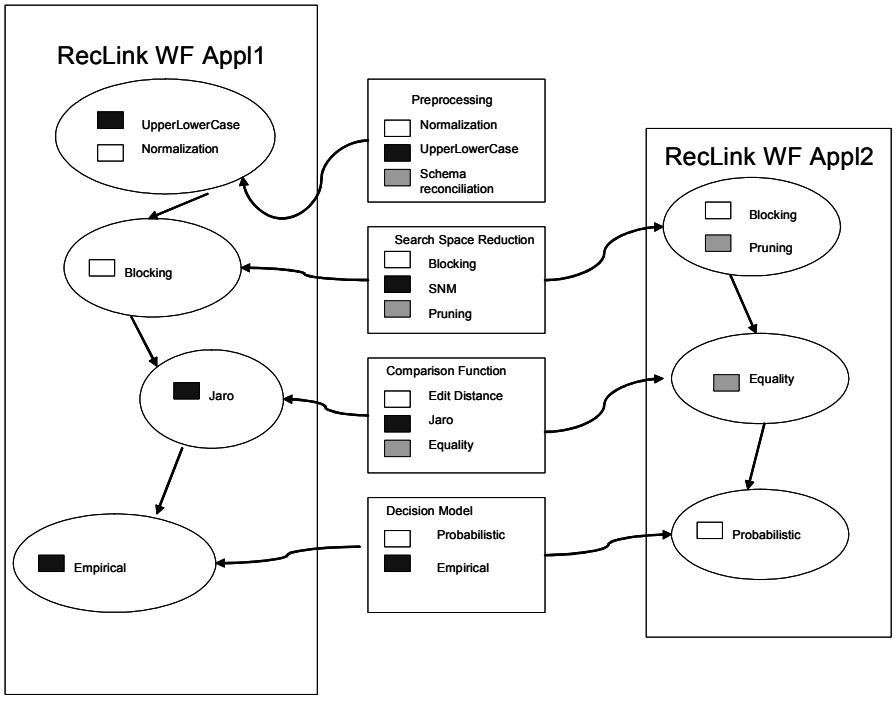


Figure 3 - Examples of RELAIS workflows



In addition, to give the opportunity to the user of designing the record linkage workflow which is more appropriate for the application at hand, the RELAIS toolkit is going to supply a data profiling phase in which a set of quality metadata are calculated starting from real data; these metadata help the user in the critical phase of choosing the best blocking or matching variables. Moreover, in order to meet needs of non-skilled users, RELAIS also proposes a default set of parameters to help the decision-making stages.

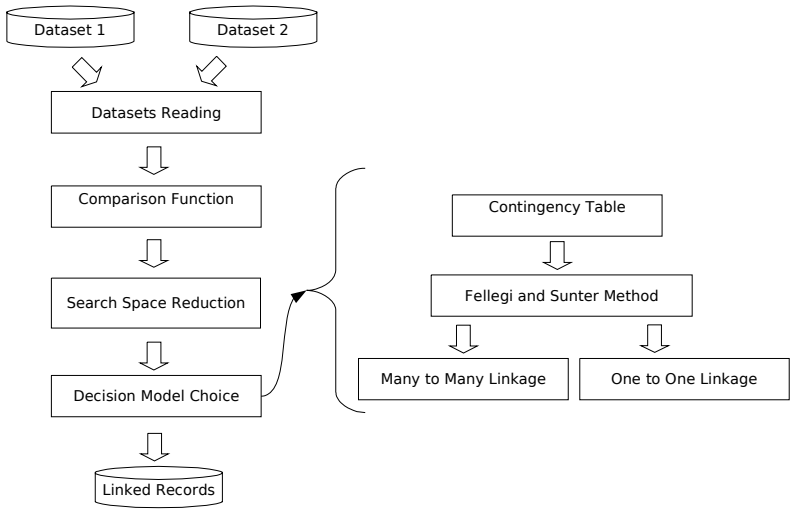
4.1 RELAIS as Open Source Project

As also remarked in the introduction, RELAIS is configured as an open source project. There are at least two reasons for this choice. First, as often highlighted above, there are many possible techniques that can be implemented for each of the record linkage phases: relying on a community of developers such set can be increased and maintained very rapidly. Second, we do believe that there have been, in the last years, several independent efforts towards the definition of a record linkage project better than the previous ones, and that such efforts have not led to the best for all solution. An open source record linkage project could instead give the possibility of gathering together the efforts already done, according to the idea described above, in order to make them available to the community for the most appropriate usage. RELAIS is mainly implemented in Java, due to the well-known features of strongly typing and platform independence; some phases are implemented in R, a free language mainly oriented to statistical computing (http://www.r-project.org).

4.2 Status of Implementation

RELAIS gives the opportunity to design different record linkage workflows. As shown in Figure 4, the main phases of the record linkage process have already been implemented (with one or more techniques): (i) datasets reading, (ii) comparison function choice, (iii) search space reduction and (iv) decision model choice.

Figure 4 - Status of implementation of RELAIS



The statistical problem can be overcome by means of the creation of suitable groups or partitions of the whole cross product set of pairs, so that in each sub-group the number of expected links is not so small compared with the number of candidate pairs.

In particular, when the conditional probabilities are estimated via the EM algorithm, as in the current version of RELAIS, it is appropriate to apply some reduction of the pairs space so that the expected number of links is not below 5% of the overall compared pairs.

Among the several reduction techniques, the current version RELAIS provides the Blocking Method and the Sorted Neighbourhood Method, described in Section 3.

In order to reduce the search space of the candidate pairs, the most suitable blocking variables are generally those most discriminating and accurate, i.e., not affected by errors or missing. Links are searched only within the blocks, assuming that there are no matches out of them; therefore, if the blocking variable is error affected, some true links could be missed. Usually, variables as zip code, municipality, geographic area, year of birth are chosen as blocking variable. In the current version of RELAIS, such choice is left to the users; in the next version, a module will be made available to guide users in such a choice by means of metadata and indicators evaluated on the data at hand.

The Sorted Neighbourhood Method (SNM), instead, consists of ordering the two data sets to link according to a sorting variable. Then a fixed size *w* window runs on the unified sorted list and all the pairs, falling into the window, are considered as candidate pairs. The size will be selected taking into account the risk of missing true links, for instance if the number of units with the same value of the sorted variable is larger than the fixed size.

Tests made shows that the cumulated size of input data that can be processed with SNM is larger than the cumulated size allowed by the blocking method.

With respect to the decision model choice, we have implemented the Fellegi-Sunter probabilistic model by using the EM algorithm for the estimation of the model parameters; as detailed in Figure 4, this method takes as input a contingency table, which reports the frequencies of the agreement patterns resulting from the application of the comparison function, and the output is a many to many linkage of the datasets records. Starting from this output, we can propose to the user the clusters of matches, non-matches and possible matches. Alternatively, a phase of reduction to a one to one linkage can be performed by applying the simplex method.

The Fellegi-Sunter model and the simplex method are implemented in R, which has a huge number of statistical packages available, thus giving us the opportunity of re-use software already developed from the scientific community, according to the open source idea of our project.

5. Case Study: Building a Record Linkage Workflow

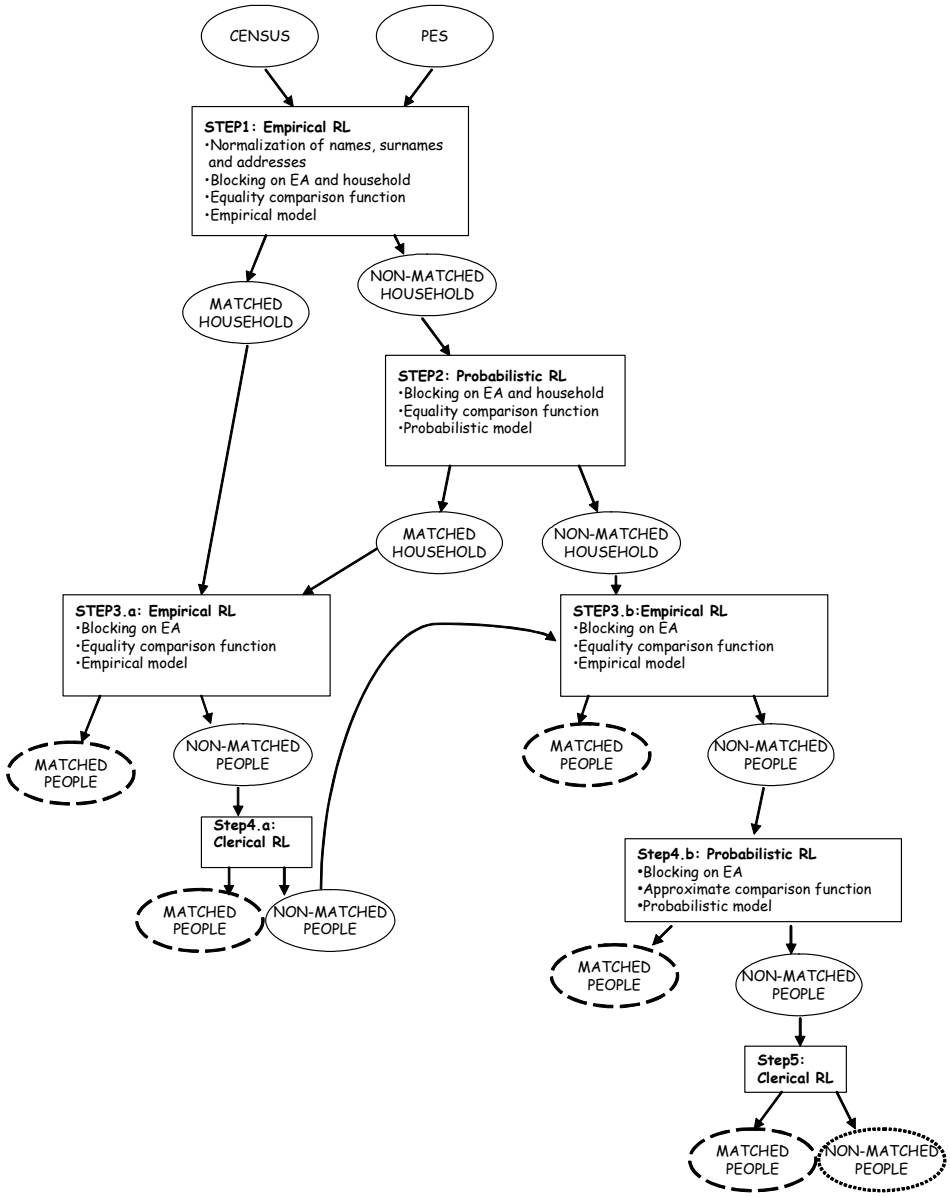
In this section a record linkage application concerning the Post Enumeration Survey (called *PES* in the following) of the Italian 2001 Census is described. The main goal of the Census was to enumerate the resident population at the Census date, the 21st of October 2001; it was also interesting to characterize Italian families; hence, the relationship of each enumerated person with the other components of the same household was also collected. The PES had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called *EA* in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 65.000 households and 170.000 people. Correspondingly, comparable amounts of households and

people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-recapture model (Wolter K., 2006) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

The estimates of the Census coverage rate through the capture-recapture model have required to match Census and PES records, assuming no errors in matching operations. This is a strong assumption: the accuracy of the matching processes was of crucial importance because even very small matching errors could have compromised the reliability of the coverage rate estimates. To guarantee the maximum correctness of the matches between PES and Census, we had to build a structured record linkage workflow, consisting of different phases and iterations. Specifically, both empirical and probabilistic record linkage techniques were used, and also different comparison functions were selected in different phases. The resulting workflow is particular significant as a proof of concept of the RELAIS toolkit usefulness. More specifically, the first phases of the workflow identify the *easiest* matches, by means of the more straightforward computational procedures, leaving the hardest ones to the subsequent phases. The iterations of the record linkage workflow were performed on the basis of the hierarchical structure of the data, in order to take advantage of the relationships among individuals belonging to the same household. Indeed, the matching units corresponding to people can be grouped according to their households membership; this structure suggests to start by first linking households and then individuals.

In the Figure 5, the steps 1 and 2 regard two iterations of the record linkage process on households. Step 1 is performed after a pre-processing activity and it is an empirical linkage. Step 2 is a probabilistic record linkage, based on the Fellegi-Sunter model, for which the matching weights are computed via the EM algorithm (Jaro 1985, Winkler 2000). In step 3.a an empirical linkage was performed on matched households for the purpose of identifying people. In the subsequent step 4.a, the residual individuals, not yet linked but belonging to matched households, were clerically checked. The un-matched people in output of step 4.a were considered as input to step 3.b, together with the individuals belonging to not linked households, and were matched by means of an empirical approach. Then, in step 4.b, for the people not linked in step 3.b, a probabilistic record linkage was carried out. The residual individuals, not yet linked at the previous steps, were submitted to a final clerical linkage in step 5.

Figure 5 - The record linkage workflow of the case study



As described above, given a set of application constraints and data features, RELAIS has the purpose to suggest the best technique to choose in each record linkage phase, in order to build the best workflow for the specific application. The case study described above allows us to highlight the following requirements: (i) the data requirements include a hierarchical structure of the data sets, a quite large dimensionality and a high quality of the data; (ii) the application requirements

include not significant errors in the matching process. The hierarchical structure suggests to distinguish record linkage workflow iterations at two levels, namely: we first match records at a higher level (households), and then at a lower level (persons). In this way, we take advantage of the hierarchical structure reducing the search space and, moreover, increasing the number of real matches. The dimension of the data sets implies high complexity of the linkage algorithm; this suggests to apply blocking techniques to reduce the complexity of the linkage. Moreover, due to volume of the data sets, a direct use of the probabilistic model could have been time consuming. Therefore, a first application of the empirical model is performed with the purpose to be refined by the subsequent use of the probabilistic model. The high quality of data implies the choice of equality as comparison function in most of the phases. The requirement concerning not significant errors in the matching process suggests the adoption of a probabilistic model in the final iterations, in order to have a quantitative estimation of the errors that can be regarded as acceptable or not. Moreover, this requirement also suggests the appropriateness of a clerical review and an exact comparison function in order to achieve the desired error bounds.

In Figure 6, a table representing the case study requirements and the corresponding choices suggested is shown. Such correspondences can be considered as a pattern useful for building record linkage workflows whereas similar application and data requirements are present.

6. Concluding Remarks

In official statistics, data integration is important in order to use available information more efficiently and to improve the quality of statistical products, in particular, statistical indicators designed to enable sound decision and policy-making. Recently, many tools for record linkage have appeared on the market and research groups have published software packages for linkage. In this paper, we have shown the RELAIS project that aims at implementing an open source toolkit for building record linkage workflows. The idea of this project has been developed keeping in mind the complexity of a record linkage problem, that involves: different techniques and sciences; the opportunity of treating the linkage with modularity by identifying the several phases which can occur, even iteratively; the different suitable approaches depending on both the data features (such as type of data or amount of data) and the application requirements (such as efficiency, effectiveness, or accuracy). As the practitioner has to deal with such a large number of situations, the toolkit wants to offer multiple techniques for record linkage, both deterministic and probabilistic, and also the possibility of building ad-hoc solution combining each modules.

Figure 6 - An example of a pattern for building record linkage workflows

REQUIREMENT		CHOICE
Data requirement	Hierarchical structure	Workflow iteration: •Higher level (household) •Lower level (person)
	High quality	Equality comparison function on most of the phases
	Huge data set	Blocking Phase iteration Empirical model
Application requirement	No errors in matching process	Probabilistic model Clerical review phase

This approach allows to overcome the question of which method is better than the others, being based on the belief that there is not actually a technique dominating all the others, but the suitability of each approach is dependent on the given data and application.

In the paper, we have described a case study as a proof of concept of the inherent complexity of record linkage processes, on which the RELAIS project is based. Indeed, due to such complexity, great modularity and flexibility are necessary in order to properly build application specific record linkage workflows.

Furthermore, besides the enrichment of the set of available techniques, future work for RELAIS will include several patterns that can guide the design of record linkage workflows. Indeed, we believe that the design stage of a record linkage workflow could usefully exploit patterns extracted from previous knowledge and experiences. In this way, the toolkit can assist non-expert users in designing their specific record linkage workflows. Each technique could be characterized in terms of pre-conditions that must be respected in order to be part of a record linkage workflow. We will study the possibility of using formal languages for the specification of such preconditions, in order to check properties like consistency and completeness of a proposed workflow solution with respect to given application and data requirements.

References

Ananthakrishna R., Chaudhuri C., and Ganti V. (2002), "Eliminating Fuzzy Duplicates in Data Warehouses", *Proceedings of VLDB 2002*, Hong Kong, China.

Bertolazzi P., Santis L.D., and Scannapieco M. (2003), "Automatic Record Matching in Cooperative Information Systems", *Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03)*, Siena, Italy.

Chauduri S., Ganti V., and Motwani R. (2005), "Robust identification of fuzzy duplicates", *Proceedings of ICDE 2005*, Tokyo, Japan.

Christen P. and Goiser K. (2005), "Assessing duplication and data linkage quality: what to measure?", *Proceedings of the fourth Australasian Data Mining Conference*, Sydney, Australia.

Ding Y. and Fienberg S.E. (1994), "Dual system estimation of Census undercount in the presence of matching error", *Survey Methodology*, 20, 149-158.

Elfeky M., Verykios V., and Elmagarmid A. K. (2002), "Tailor: A Record Linkage Toolbox", *Proceedings of the 18th International Conference on Data Engineering*, IEEE Computer Society, San Jose, CA, USA.

Fair M. (2001), "Recent developments at statistics canada in the linking of complex health files", *Federal Committee on Statistical Methodology*, Washington D.C.

Febrl. http://www.sourceforge.net/projects/febrl.

Fellegi I. and Sunter A. (1969) "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64.

Fortini M., Liseo B., Nuccitelli A., and Scanu M. (2001), "On Bayesian record linkage" *Research in Official Statistics*, 4:185-198.

Fortini M., Scannapieco M., Tosco L. and Tuoto T.(2006): Towards an Open Source Toolkit for Building Record Linkage Workflows, In Proc. of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS'06), Chicago, USA.

Gill L. (2001), "Methods for Automatic Record Matching and Linkage and their Use in National Statistics", National Statistics Methodological Series no. 25, HMSO Norwich, UK.

Gu L., Baxter R., Vickers D., and Rainsford C. (2003), "Record linkage: Current practice and future directions", Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia.

Gu L. and Baxter R. (2004), "Adaptive filtering for efficient record linkage", *Proceedings of the Fourth SIAM International Conference on Data Mining*.

Hernandez M. and Stolfo S. (1998), "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem", *Journal of Data Mining and Knowledge Discovery*, 1(2).

Jaro M. (1985), "Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of American Statistical Society*, 84(406):414-420.

Koudas N.and Srivastava D. (2005), "Approximate joins: Concepts and techniques", *Proceedings of VLDB 2005*.

Monge A. and Elkan C. (1997), "An Efficient Domain Independent Algorithm for Detecting Approximate Duplicate Database Records", *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97)*, Tucson, AZ, USA.

Newcombe H, Kennedy J, Axford S and James A. (1959), "Automatic Linkage of Vital Records", Science, Vol.130 pp. 954-959.

The-Link-King. http://www.the-link-king.com.

The-R-Project for Statistical Computing. http://www.r-project.org/.

Tuoto T. , Cibella N., Fortini M., Scannapieco M. and Tosco L. (2007): RELAIS: Don't Get Lost in a Record Linkage Project, In Proc. of the Federal Committee on Statistical Methodologies (FCSM 2007) Research Conference, Arlington, VA, USA.

Winkler W. (2000), "Frequency-based matching in Fellegi-Sunter model of record linkage", Technical report, U.S. Bureau of the Census - Washington D.C. Technical Report RR/2000/06, Statistical Research Report Series.

Winkler W. (2001), "Record Linkage Software and Methods for Merging Administrative Lists", Technical report, U.S. Bureau of the Census - Washington D.C. Technical Report RR/2001/03, Statistical Research Report Series.

Winkler W. (2004), "Methods for Evaluating and Creating Data Quality", *Information Systems*, 29(7).

Wolter K. (1986), "Some coverage error models for census data", *Journal of the American Statistical Association*, 81:338-346.

Yancey W. /2007), "BigMatch: A Program for Extracting Probable Matches from a Large File", Technical report, Statistical Research Division U.S. Bureau of the Census - Washington D.C. Research Report Series - Computing n. 2007-01.

La stazionarietà locale nella stima della povertà relativa per piccole aree[1]

Roberto Benedetti[2], Claudia Rinaldelli [3]

Sommario

Questo lavoro propone un approccio di non stazionarietà allo stimatore EBLUP al fine di migliorare l'efficienza delle stime di povertà relativa per piccole aree. I risultati delle sperimentazioni, eseguite sui dati dell'indagine europea sui Redditi e le condizioni di vita delle famiglie (EU-SILC), dimostrano l'efficacia della soluzione metodologica proposta. Il lavoro presenta inoltre i risultati di precedenti studi per la stima della povertà relativa per piccole aree eseguiti mediante approccio di stazionarietà nell'indagine sui Consumi delle Famiglie. Questi risultati evidenziano ulteriormente la validità della procedura di non stazionarietà qui proposta.

Abstract

This paper proposes a local stationarity approach in EBLUP estimation with the aim to improve the reliability of relative poverty estimates in small areas. The results of the experimentations, performed in the framework of the European Statistics on Income and Living Conditions survey (EU-SILC), show the goodness of this methodological proposal.
In addition, this paper reports the results of previous experiences of stationarity approach to poverty estimation for small areas in the Household Budget survey, in order to highlight the validity of the proposed local stationarity solution.

Parole chiave: EBLUP, stazionarietà locale, MSE

1. Introduzione

La povertà relativa è stimata annualmente dall'Istat mediante i dati rilevati dalle indagini campionarie rispettivamente sui Consumi delle Famiglie e sui Redditi e le Condizioni di Vita delle Famiglie (EU-SILC[4]).

Sono considerate ufficiali[5] le misure di povertà stimate mediante i dati della prima

[1] Il presente articolo è frutto del lavoro congiunto degli autori; Roberto Benedetti ha scritto i paragrafi 4.1, 4.1.1, 4.1.2 e 4.2, Claudia Rinaldelli ha scritto i paragrafi 1, 2, 3, 3.1, 4 e 5.

[2] Professore Ordinario di Statistica Economica (Università di Chieti-Pescara), e-mail: benedett@unich.it.

[3] Primo Ricercatore (Istat), e-mail: rinaldel@istat.it

[4] European Statistics on Income and Living Conditions survey.

[5] Ossia utilizzate in un contesto economico-politico nazionale.

indagine citata; queste sono state diffuse negli anni 1997-2001 con un dettaglio territoriale a livello di ripartizione geografica (Nord, Centro, Mezzogiorno) (ISTAT, 2002). Nel tempo si è posto sempre di più l'accento sulla necessità di ottenere stime di povertà a livelli territoriali più disaggregati per la programmazione di interventi più mirati; in questo contesto, si inserisce l'impegno assunto dall'Istat con il Ministero dell'Economia e delle Finanze per la fornitura di stime ufficiali di povertà a livello *regionale* a partire dai dati dell'anno 2002 e fino all'anno 2008 (Coccia et al., 2002, 2005; ISTAT, 2003a).

L'indagine EU-SILC, eseguita per la prima volta nell'anno 2004, si inserisce invece in un contesto internazionale e stima quindi misure di povertà al fine di una loro comparazione internazionale. L'indagine EU-SILC è stata infatti progettata ed è eseguita a livello europeo con lo scopo di ottenere la produzione sistematica di statistiche sul reddito e le condizioni di vita, sulla povertà e l'esclusione sociale degli individui e delle loro famiglie, a livello nazionale ed europeo; per questa specifica finalità, attualmente la povertà stimata dall'indagine EU-SILC fa riferimento all'intero territorio nazionale, sebbene il regolamento europeo che disciplina l'indagine, lasci libertà ai singoli paesi di calcolare la povertà a livello più disaggregato per specifiche esigenze nazionali (European Parliament, 2003). Nonostante le differenze citate, la misura della povertà relativa si basa, in entrambe le indagini, su un approccio statistico che combina l'uso di una variabile economica (rispettivamente la spesa per consumi nella prima indagine e il reddito nella seconda) e di una linea di povertà (o soglia) che consente di classificare le famiglie e/o gli individui come poveri o no. Le misure di povertà si configurano pertanto come stime complesse e le problematiche metodologiche a loro collegate riguardano soprattutto la stima del loro errore campionario e la loro possibile diffusione per aree o domini disaggregati (piccole aree).

Il presente lavoro affronta la seconda problematica citata e in particolare riporta le esperienze maturate per stimare in maniera efficiente la povertà relativa per piccole aree, rispettivamente a livello regionale per l'indagine sui Consumi delle Famiglie e nelle province italiane (livello NUTS III)[6] per l'indagine EU-SILC.

Secondo la classificazione proposta da Purcell e Kish (1979), le piccole aree sono i domini con un numero di unità statistiche comprese tra 1/10 e 1/100 rispetto al totale della popolazione di riferimento. Le regioni e le province non sono piccole in questo senso; tuttavia, le stime di povertà possono essere affette, a tali livelli geografici, da errori di campionamento elevati e quindi possono essere considerate piccole aree secondo la definizione di Brackstone (1987) "sottopopolazioni per le quali, non possono essere ottenute stime accurate attraverso le sole informazioni derivanti dalle indagini campionarie correnti". Nonostante le differenti misure di povertà prodotte dall'Istat, in questo lavoro si dà più enfasi alla povertà relativa stimata dall'indagine EU-SILC. Tale scelta nasce dalla considerazione[7] che è probabile una futura produzione di stime disaggregate di povertà proprio a partire da questa indagine.

Il lavoro in oggetto propone un approccio di *non stazionarietà* al tradizionale stimatore EBLUP (Empirical Best Linear Unbiased Predictor, brevemente richiamato nel paragrafo 2) al fine di migliorare l'efficienza delle stime di povertà relativa per piccole aree. Si illustrano, con riferimento all'indicatore *at Risk-of-Poverty Rate* (incidenza di povertà relativa, si veda paragrafo 4) utilizzato nell'indagine EU-SILC, due sperimentazioni

[6] Nomenclature of territorial units for statistical purposes.

[7] Opinione degli autori che non riflette necessariamente quella dell'ISTAT.

eseguite al fine di migliorare l'efficienza delle stime di povertà relativa a livello provinciale; in particolare, si introduce un approccio di *non stazionarietà* nella stima dei parametri di regressione di un tradizionale stimatore EBLUP (paragrafi 4.1, 4.1.1, 4.1.2). Tale impostazione nasce dalla considerazione che la *povertà* è *influenzata* dal contesto geografico in cui si presenta; le stime di povertà presentano un andamento ben preciso sul territorio e frequentemente aree geografiche limitrofe sono caratterizzate da situazioni simili del fenomeno in oggetto. Per questo motivo, si sperimenta la possibilità di individuare zone *omogenee* di province al fine di ottimizzare l'uso dell'informazione ausiliaria nello stimatore EBLUP e conseguentemente stimare la povertà nelle province in maniera più efficiente. I risultati delle sperimentazioni dimostrano l'efficacia della soluzione metodologica proposta (paragrafo 4.2). Il lavoro presenta inoltre i risultati di precedenti studi per la stima della povertà relativa per piccole aree eseguiti mediante approccio di *stazionarietà* nell'indagine sui Consumi delle Famiglie. In particolare, i paragrafi 3 e 3.1 riportano lo studio comparativo tra le stime da disegno della povertà e le corrispondenti stime EBLUP, effettuato a livello di regione geografica nel contesto della suddetta indagine. Questi risultati evidenziano ulteriormente la validità della procedura di *non stazionarietà* qui proposta. Infine, il paragrafo 5 contiene le conclusioni.

2. La stima modellistica per piccole aree: lo stimatore EBLUP

L'approccio adottato per stimare, rispettivamente la percentuale di famiglie povere a livello regionale con riferimento all'indagine sui Consumi delle Famiglie e la percentuale di persone povere a livello provinciale con riferimento all'indagine EU-SILC, è basato sulle tecniche di *small area estimation* (SAE), ossia su stime indirette prodotte a partire da un modello esplicito che mette in relazione le stime regionali/provinciali dirette con informazioni ausiliarie (Rao, 2003).

Questi modelli possono essere classificati in due ampi gruppi: modelli a livello di area e modelli a livello di unità; in questo studio si considera il primo tipo di modelli e si fa riferimento al noto modello lineare misto di Fay e Herriott (1979):

$$\hat{\theta}_i = \mathbf{z}_i^T \boldsymbol{\beta} + \nu_i + \varepsilon_i \qquad \text{i = 1, 2,…, m} \qquad (2.1)$$

dove $\hat{\theta}_i$ denota la stima del parametro di interesse con riferimento alla piccola area i, $\mathbf{z}_i = (z_{1i},...,z_{pi})$ denota il vettore delle variabili ausiliarie per l'area i, $\boldsymbol{\beta}$ è il vettore (px1) dei parametri di regressione, ν_i sono gli effetti di area ipotizzati indipendenti e identicamente distribuiti con $E(\nu_i) = 0$ e $\text{var}(\nu_i) = \sigma_\nu^2$, ε_i sono gli errori di campionamento con $E(\varepsilon_i / \theta_i) = 0$, $\text{var}(\varepsilon_i / \theta_i) = \psi_i$ e le ψ_i sono le varianze di campionamento assunte come note. Tra le soluzioni proposte in letteratura, in questo lavoro si è optato nel porre le varianze ψ_i pari alle loro stime basate sul disegno (Ghosh e Rao, 1994). Il modello (2.1) deriva dalla combinazione dell'incertezza dovuta al disegno campionario espressa da:

$$\hat{\theta}_i = \theta_i + \varepsilon_i \tag{2.2}$$

dove θ_i denota il parametro di interesse nella piccola area i con un modello di link espresso da:

$$\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + v_i \tag{2.3}$$

Il modello (2.1) contempla quindi due fonti di casualità: quella indotta dal disegno (ε_i) e quella che compete al modello (v_i). Facendo riferimento ai risultati generali validi per i modelli lineari a effetti misti (Rao, 2003), lo stimatore EBLUP di θ_i è espresso da:

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i + (1-\gamma_i)\mathbf{z}_i^T \tilde{\boldsymbol{\beta}} \tag{2.4}$$

ossia una combinazione pesata delle stime dirette $\hat{\theta}_i$ e delle stime di regressione $\mathbf{z}_i^T \tilde{\boldsymbol{\beta}}$ e dove $\gamma_i = \sigma_v^2 / (\psi_i + \sigma_v^2)$.

Il Mean Squared Error (MSE) dello stimatore (2.4) è espresso da (Rao, 1999, 2003):

$$MSE(\tilde{\theta}_i) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) \tag{2.5}$$

dove:

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i$$

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \mathbf{z}_i^T \left[\sum_i \mathbf{z}_i \mathbf{z}_i^T / (\sigma_v^2 + \psi_i) \right]^{-1} \mathbf{z}_i$$

$$g_{3i}(\sigma_v^2) = \left[\psi_i^2 / (\sigma_v^2 + \psi_i)^4 \right] E(\hat{\theta}_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 \overline{V}(\hat{\sigma}_v^2)$$

e dove $\overline{V}(\hat{\sigma}_v^2)$ è la varianza asintotica di $\hat{\sigma}_v^2$. Per maggiori dettagli sulla stima dell'MSE si veda Rao, 2003, par.6.2.6.

In questo lavoro, i parametri di regressione $\tilde{\boldsymbol{\beta}}$ e σ_v^2 sono stimati con il metodo della massima verosimiglianza ristretta (REML) (Jiang, 1996) ottenuta ricorrendo a metodi numerici di ottimizzazione (Rao, 2003, par. 6.2.4).

3. L'indagine sui Consumi delle Famiglie e l'indicatore incidenza di povertà relativa

L'indagine sui Consumi delle Famiglie, basata su disegno di campionamento complesso (comuni, famiglie), coinvolge ogni anno circa 24.000 famiglie con lo scopo di rilevare le

spese familiari per consumi. I dati dell'indagine sono utilizzati sia per stimare le spese sostenute dalle famiglie per acquistare beni e servizi sia per calcolare la misura *target* di povertà nota come *incidenza di povertà relativa* (ISTAT, 2002). L'incidenza di povertà relativa che stima, annualmente, la percentuale di famiglie povere è basata su una metodologia che combina l'uso della variabile *spesa per consumi* con una soglia che permette di classificare le famiglie come povere o no; in particolare, una famiglia è definita povera se la sua spesa mensile per consumi è inferiore o uguale alla soglia detta *linea di povertà*, stimata a sua volta come spesa media mensile per consumi a livello nazionale. Una scala di equivalenza consente di tenere conto dell'ampiezza familiare quando si confronta la spesa media mensile per consumi della famiglia con la soglia di povertà (ISTAT, 2002). Introducendo j come indice della famiglia, y_j come il valore della spesa mensile per consumi della famiglia j, w_j come coefficiente di riporto all'universo della famiglia j, S come il campione rilevato di famiglie, 'Famiglie' come il numero totale delle famiglie residenti (ammontare demografico noto), la stima dell'incidenza di povertà relativa è definita come:

$$\hat{I}_{pov} = \frac{\sum_{j \varepsilon S} I_j w_j}{Famiglie} * 100 \qquad (3.1)$$

e I_j è una variabile binaria definita come:

$$I_j = \begin{cases} 1 & se \quad y_j \leq linea \ \ povertà \\ 0 & altrimenti \end{cases} \qquad (3.2).$$

3.1 Le stime EBLUP di povertà relativa per l'indagine sui Consumi delle Famiglie

Come introdotto nel paragrafo 1, l'incidenza di povertà relativa è stata diffusa negli anni 1997-2001 per ripartizione geografica; tuttavia, a seguito dell'impegno assunto dall'Istat con il Ministero dell'Economia e delle Finanze, le stime di povertà sono fornite a livello regionale a partire dai dati dell'anno 2002 (ISTAT, 2003a).

Una serie di studi, finalizzati ad ottenere stime più efficienti di povertà per le regioni, ha preceduto la fase di diffusione (Benedetti et al., 2003; Di Consiglio et al., 2003).

In particolare, in Benedetti et al. (2003) è stata valutata la possibilità di adottare uno stimatore EBLUP mediante uno studio comparativo delle stime dirette (o da disegno) di povertà con le corrispondenti stime EBLUP. Per implementare lo stimatore EBLUP, le stime dirette di povertà relativa sono state calcolate a livello regionale mediante lo stimatore di ponderazione vincolata (Deville e Särndal, 1992; Falorsi e Rinaldelli, 1998) correntemente utilizzato nell'indagine e la loro varianza di campionamento è stata calcolata mediante la tecnica di ricampionamento delle *Replicazioni Bilanciate Ripetute* in quanto l'indicatore incidenza di povertà relativa non è lineare (Pauselli e Rinaldelli, 2003, 2004; Rinaldelli, 2006).

Il *tasso di occupazione* stimato a livello regionale dall'indagine Forze di Lavoro (ISTAT, 2003b e precedenti) e l'*importo pensioni Invalidità Vecchiaia e Superstiti (IVS) per famiglia* (da fonte amministrativa) (ISTAT, 2004a e precedenti) sono stati utilizzati come informazioni ausiliarie. Ovviamente il tasso di occupazione è affetto da varianza

campionaria, essendo stimato da un'indagine campionaria; tuttavia, è risultato tra le informazioni ausiliarie disponibili più correlate con la povertà ed inoltre è stimato da un'indagine con numerosità campionaria notevolmente più ampia di quella dell'indagine sui Consumi delle Famiglie. Per tale motivo, in questi studi è stato considerato non affetto da errore.

La tabella 1 riporta i valori delle correlazioni del tasso di occupazione e dell'importo pensioni Invalidità Vecchiaia e Superstiti (IVS) per famiglia con l'incidenza di povertà negli anni 1997-2002.

Tabella 1 - Coefficienti di correlazione dell'incidenza di povertà con tasso di occupazione e IVS

Anno	Occupazione	IVS
1997	-0,85	-0,88
1998	-0,85	-0,85
1999	-0,93	-0,88
2000	-0,92	-0,87
2001	-0,89	-0,90
2002	-0,87	-0,91

Fonte: Elaborazione su dati Istat

La tabella 2 riporta i valori del rapporto tra le stime EBLUP e le stime dirette di povertà; si osserva che la differenza tra le due stime è contenuta.

Per quanto riguarda la precisione delle stime, è opportuno notare che la varianza da disegno e l'MSE derivante da modello sono ottenuti sulla base di ipotesi completamente diverse e quindi il loro confronto non può certo fornire indicazioni puntuali. Ad ogni modo il rapporto, esposto nella tabella 3, tra il MSE relativo delle stime EBLUP e l'errore relativo campionario delle stime dirette di povertà evidenzia che non si realizzano consistenti guadagni di efficienza implementando stime EBLUP di povertà rispetto a stime dirette.

Un ulteriore studio di stimatori per piccole aree (Di Consiglio et al., 2003) finalizzato a risolvere la medesima problematica, ha verificato conclusioni simili a quella sopra esposta. In base ai suddetti risultati e considerando inoltre la maggiore complessità degli stimatori per piccole aree, la povertà è stimata a livello regionale da Istat (a partire dal 2003) mediante lo stimatore di ponderazione vincolata utilizzato nell'indagine.

Tabella 2 - Rapporto tra stime EBLUP e stime dirette di povertà relativa

	1997	1998	1999	2000	2001	2002
Piemonte	0,97	0,98	1,02	1,01	1,01	0,98
Valle d'Aosta	0,94	0,92	0,98	0,97	0,93	1,01
Lombardia	1,00	0,99	0,97	1,02	1,01	1,01
Trentino A.A.	1,02	1,03	0,99	0,93	1,00	0,98
Veneto	1,03	1,02	1,02	1,08	1,07	1,05
Friuli V.G.	0,71	0,97	0,97	0,94	0,96	0,92
Liguria	1,03	1,02	1,11	1,02	1,03	1,04
Emilia Romagna	0,95	0,97	0,96	0,98	0,97	0,97
Toscana	1,02	1,01	1,06	1,03	1,04	1,00
Umbria	0,97	1,03	0,78	0,88	0,96	1,03
Marche	1,02	1,09	1,08	1,08	1,06	1,04
Lazio	1,04	1,05	1,13	0,96	0,99	1,01
Abruzzo	1,00	1,02	0,93	1,09	1,00	0,92
Molise	0,97	0,90	0,87	0,97	0,80	0,93
Campania	1,01	1,01	1,06	1,05	1,02	1,02
Puglia	1,04	0,98	0,93	1,00	0,96	0,99
Basilicata	0,89	0,89	0,92	0,95	0,92	0,90
Calabria	0,94	1,04	0,98	0,91	1,03	0,94
Sicilia	0,98	0,99	1,00	1,06	1,04	1,04
Sardegna	1,03	0,98	0,98	1,03	0,97	1,03

Fonte: Elaborazione su dati Istat

Tabella 3 - Rapporto tra MSE relativo delle stime EBLUP di povertà e errore relativo campionario delle stime dirette

	1997	1998	1999	2000	2001	2002
Piemonte	0,98	1,00	0,96	0,95	0,97	0,99
Valle d'Aosta	0,98	0,99	0,90	0,74	0,92	0,95
Lombardia	0,99	1,00	1,02	0,96	0,98	0,98
Trentino A.A.	0,96	0,95	0,98	0,98	0,97	0,96
Veneto	0,94	0,96	0,94	0,83	0,92	0,94
Friuli V.G.	1,10	0,99	0,92	0,95	1,00	1,03
Liguria	0,94	0,95	0,70	0,92	0,92	0,94
Emilia Romagna	1,03	1,01	1,03	1,00	1,01	1,02
Toscana	0,96	0,96	0,93	0,93	0,92	0,97
Umbria	0,93	0,94	0,83	0,94	0,93	0,90
Marche	0,95	0,85	0,90	0,88	0,88	0,95
Lazio	0,95	0,93	0,75	0,90	0,96	0,97
Abruzzo	0,92	0,90	0,82	0,65	0,89	0,79
Molise	0,69	0,94	0,74	0,80	0,70	0,94
Campania	0,91	0,94	0,84	0,79	0,85	0,88
Puglia	0,87	0,93	0,79	0,73	0,87	0,84
Basilicata	0,79	0,89	0,68	0,75	0,77	0,67
Calabria	0,87	0,88	0,80	0,83	0,80	0,86
Sicilia	0,91	0,93	0,71	0,66	0,84	0,90
Sardegna	0,87	0,88	0,70	0,79	0,78	0,88

Fonte: Elaborazione su dati Istat

4. L'indagine EU-SILC e l'indicatore At Risk-of-Poverty Rate

L'indagine EU-SILC è disciplinata dal regolamento europeo n. 1177/2003 adottato dal Parlamento Europeo e dal Consiglio del 16 giugno 2003, noto come *regolamento quadro* (*framework regulation*) (European Parliament, 2003).

EU-SILC coinvolge ogni anno circa 60.000 persone ed è basata su disegno di campionamento complesso (comuni, famiglie) con componente longitudinale; ogni famiglia è infatti rilevata per quattro anni consecutivi.

Nel quadro della produzione sistematica dell'indagine, gli indicatori di povertà relativa e disuguaglianza, noti come indicatori di Laeken[8], rivestono un ruolo importante.

Gli indicatori di Laeken sono calcolati, annualmente, con i corrispondenti errori di campionamento e trasmessi dai Paesi Membri ad Eurostat, secondo quando previsto dal regolamento europeo (Commission regulation, 2004). Tra questi, l'indicatore at Risk-of-Poverty Rate è una delle misure *target* di povertà relativa più importante. Fino ad oggi, l'indicatore at Risk-of-Poverty Rate (ossia la percentuale di persone povere) è stato calcolato a livello nazionale. L'indicatore at Risk-of-Poverty Rate è basato su una metodologia che combina l'uso della variabile *reddito* con una soglia che permette di classificare gli individui come poveri o no; in particolare, un individuo è definito povero se il suo reddito è inferiore alla soglia detta *at Risk-of-Poverty Threshold* (linea di povertà relativa). Introducendo k come indice dell'individuo, y_k come il valore del reddito per l'individuo k, w_k come coefficiente di riporto all'universo dell'individuo k, $\hat{Y}_\beta$ come la stima del quantile di ordine β ($0 \leq \beta \leq 1$) della variabile reddito Y, lo stimatore at Risk-of-Poverty Rate (RPR) è definito come:

$$RPR = \frac{\sum_{k \varepsilon S} I_k w_k}{\sum_{k \varepsilon S} w_k} * 100 \qquad (4.1)$$

dove S è il campione rilevato, I_k è la variabile binaria definita come:

$$I_k = \begin{cases} 1 & se \quad y_k < RPT \\ 0 & altrimenti \end{cases} \qquad (4.2)$$

e RPT è la soglia di povertà relativa (at Risk-of-Poverty Threshold) definita come:

$$RPT = 60\% \hat{Y}_{0.5} \qquad (4.3)$$

dove $\hat{Y}_{0.5}$ è il valore mediano del reddito stimato a livello nazionale.

Il *reddito* utilizzato nel calcolo della povertà è il *reddito disponibile equivalente* ossia una misura di reddito della persona che tiene conto sia del reddito della famiglia di appartenenza sia della sua dimensione e composizione (EUROSTAT, 2004). In dettaglio, si calcola dapprima il reddito totale disponibile della famiglia di appartenenza; questo viene ottenuto come somma dei redditi personali percepiti da tutti i componenti della famiglia e

[8] Poichè adottati ufficialmente durante il Consiglio Europeo svoltosi a Laeken nel 2001.

dei redditi percepiti a livello familiare. Il reddito totale disponibile della famiglia viene quindi diviso per la dimensione equivalente della famiglia dando luogo al reddito disponibile equivalente, ossia al reddito individuale utilizzato nel calcolo di (4.1)-(4.3) e attribuito ad ogni componente della famiglia. La dimensione equivalente della famiglia è ottenuta sulla base della scala OECD modificata che attribuisce peso pari all'unità al primo adulto della famiglia, peso pari a 0.5 ai componenti di 14 anni e più, e peso pari a 0.3 ai componenti di età inferiore ai 14 anni (EUROSTAT, 2004). Per implementare lo stimatore EBLUP, le stime dirette di povertà relativa sono calcolate a livello provinciale[9] mediante lo stimatore di ponderazione vincolata (Deville e Särndal, 1992; Falorsi e Rinaldelli, 1998) in accordo con il disegno campionario di EU-SILC[10], e la varianza di campionamento delle stime dirette di povertà è calcolata mediante l'approccio di *linearizzazione* per *equazioni stimanti* in quanto l'indicatore at Risk-of-Poverty Rate non è lineare (Rinaldelli, 2005, 2006; Moretti, Pauselli e Rinaldelli, 2005). Nella fig. 1, a sinistra, la mappa riporta i valori delle stime dirette di povertà nelle province; si può osservare come la percentuale di persone povere sia differenziata sul territorio nazionale (generalmente più elevata nelle province meridionali e meno intensa in quelle settentrionali); la mappa a destra riporta i corrispondenti valori della varianza campionaria delle stime dirette di povertà; si può osservare come le stime dirette di povertà possono essere affette da valori elevati di varianza campionaria. L'informazione ausiliaria utilizzata in questo lavoro per calcolare le stime EBLUP di povertà è il tasso di occupazione ottenuto a livello provinciale dall'indagine Forze di Lavoro (ISTAT, 2004b). Come già ricordato nel paragrafo 3.1, il tasso di occupazione è a sua volta affetto da varianza campionaria; tuttavia, tra le informazioni ausiliarie disponibili a livello provinciale, è risultato quello più correlato (r = -0.85) con la povertà e quindi è stato scelto, sebbene considerato non affetto da errore, ai fini dell'esecuzione delle due sperimentazioni qui descritte. Come nello studio descritto nel paragrafo 3.1, anche in questo caso, adottando un approccio EBLUP di tipo tradizionale non si ottengono guadagni di efficienza evidenti rispetto alle classiche stime di povertà da disegno (o dirette). In fig. 2 si può osservare come i valori del MSE relativo delle tradizionali stime EBLUP di povertà non siano significativamente inferiori ai corrispondenti valori dell'errore relativo di campionamento delle stime dirette di povertà nella maggior parte delle province. Come introdotto nel paragrafo 1, una caratteristica della povertà è la territorialità che suggerisce la possibilità di introdurre un approccio di non stazionarietà nello stimatore EBLUP al fine di ottenere un guadagno di efficienza delle stime prodotte (Benedetti e Rinaldelli, 2007a, 2007b). Tale impostazione è resa possibile anche dal maggior numero di aree geografiche oggetto di studio (le province) rispetto al numero più contenuto, le regioni geografiche, interessate nel precedente studio esposto nei paragrafi 3 e 3.1.

4.1 L'approccio di non stazionarietà

Risulta evidente l'influenza che il territorio ha sul fenomeno della povertà nel senso che i suoi valori sono differenziati nelle aree geografiche considerate e che, frequentemente, aree limitrofe si comportano allo stesso modo (Besag, 1974; Cressie, 1991). Da qui, nasce l'idea di sperimentare un approccio di non stazionarietà dei parametri del modello adottato nello stimatore EBLUP precedentemente specificato. Ciò si traduce nell'individuazione di zone omogenee di province finalizzate ad un uso ottimale dell'informazione ausiliaria nello stimatore EBLUP e conseguentemente ad una stima più efficiente della povertà nelle province. Nella teoria classica della regressione, le unità spaziali sono tradizionalmente

[9] Dati della rilevazione dell'anno 2004.

[10] Il peso diretto, corretto per mancata risposta, è vincolato alla distribuzione per sesso ed età della popolazione.

rappresentate come elementi identici di una medesima popolazione:

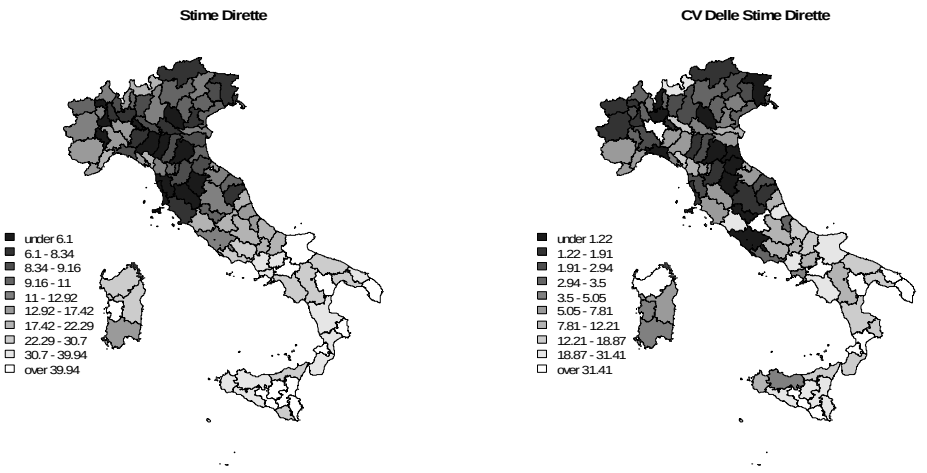
$$y_i = f(x_{1i}, x_{2i}, ..., x_{hi}, \theta) + \xi_i \qquad (4.1.1)$$

(4.1.1) può essere modificato come:

$$y_i = f(x_{1i}, x_{2i}, ..., x_{hi}, \theta_{ki}) + \xi_i \qquad (4.1.2)$$

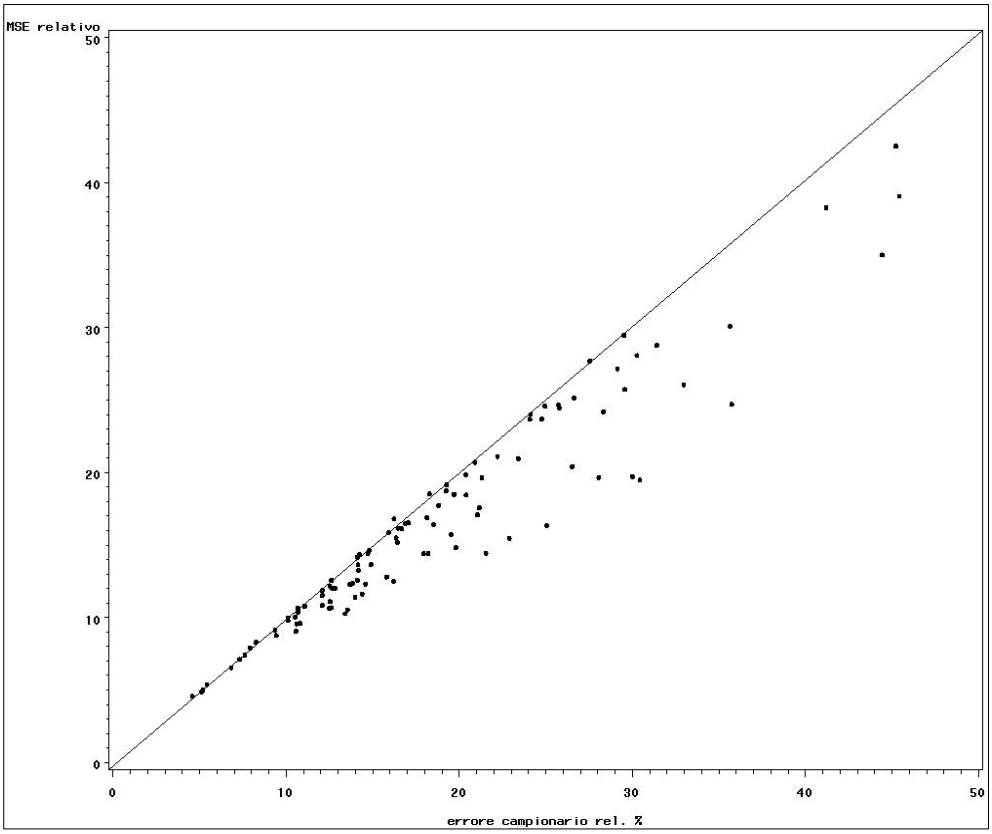
dove $\xi_i \sim N(0, \sigma_u^2)$ e $\boldsymbol{k} = (k_1, k_2, \ldots, k_i, ..., k_n)$ ε n^k, $k_i \varepsilon \{1,2,\ldots,k\}$ $\forall$ *i,* è il vettore etichetta che rappresenta la zona di stazionarietà locale associata ad ogni singola unità e *f* è lineare sebbene l'approccio risulti valido anche per modelli statistici non lineari.

Figura 1 - Stime dirette della povertà nelle province (a sinistra) e corrispondente varianza campionaria (a destra)



Fonte: Elaborazione su dati Istat

Figura 2 - Errore relativo di campionamento delle stime dirette di povertà (ascisse) vs MSE relativo delle stime EBLUP – approccio tradizionale- (ordinate) per provincia



Fonte: Elaborazione su dati Istat

L'approccio di non stazionarietà si traduce quindi nella stima di k vettori di parametri di regressione rispetto al modello assunto come conseguenza delle k zone di stazionarietà individuate. L'assunzione di non stazionarietà dei parametri del modello di regressione conduce pertanto allo studio di un problema combinatorio di ottimizzazione spaziale. Infatti, definito k il numero delle zone di stazionarietà, l'obiettivo è quello di individuare la partizione ottimale delle unità, in questo caso le province, tra tutte le possibili partizioni per poi conseguentemente stimare k vettori di parametri di regressione β. La partizione ottimale delle province in k zone di stazionarietà viene individuata in questo lavoro rispettivamente mediante l'algoritmo di annealing simulato e la procedura di smoothing adattivo dei pesi.

4.1.1 L'algoritmo di Annealing simulato

L'algoritmo iterativo di *annealing* simulato (*Simulated Annealing*, SA), proposto da Benedetti et al. (2007a) nell'ambito della stima per piccole aree, consente di risolvere un problema combinatorio di ottimizzazione spaziale; in particolare, SA permette di trovare, numericamente, il minimo assoluto di una funzione obiettivo definita su un insieme *P* di punti, evitando configurazioni di minimo locale (Geman et al., 1990). L'algoritmo SA è mutuato

(Kirkpatrick et al.,1983) dal processo fisico di *annealing* per cui un solido viene fuso, portato allo stato liquido e raffreddato molto lentamente affinché raggiunga uno stato di energia minima possibile. In analogia a questo processo, i punti dell'insieme *P* sono visti come un sistema fisico e la funzione obiettivo, che si intende minimizzare, riveste il ruolo dell'energia interna del sistema. L'*annealing* simulato conduce il sistema ad uno stato caratterizzato dalla minima energia possibile e cioè l'insieme *P* dei punti alla configurazione in cui la funzione obiettivo raggiunge il minimo assoluto. In particolare, partendo dalla configurazione iniziale P_0 in cui i punti sono divisi casualmente in *k* zone e definita *J(P)* la funzione obiettivo, l'*annealing* simulato consiste nel *visitare* tutti i punti di *P*, uno ad uno, e nello spostarli casualmente in una zona diversa da quella di appartenenza. Ad ogni *visita* corrisponde una nuova configurazione P_{j+1} che sostituisce la precedente configurazione P_j con la seguente probabilità:

$$p_{j,j+1} = \begin{cases} 1 & if\ J(P_{j+1}) < J(P_j) \\ \exp\left(-\dfrac{J(P_{j+1}) - J(P_j)}{T_j}\right) & altrimenti \end{cases} \qquad (4.1.1.1)$$

dove T_j misura la temperatura del sistema e mantiene lo stesso valore fino a che non è completata la *visita* di tutti i punti di *P*. L'algoritmo si ripete e cioè l'insieme *P* viene nuovamente visitato mediante le operazioni sopra descritte, con un valore inferiore di T_j. Al fine di evitare configurazioni caratterizzate da minimi locali, SA viene iterato con valori di T_j che decrescono lentamente secondo la seguente espressione:

$$T_j = T_0 \rho^{j-1} \quad \text{con } \rho<1 \qquad (4.1.1.2)$$

Le iterazioni di SA terminano quando i punti non si spostano più da una zona all'altra.

Nel presente lavoro, i punti di *P* sono le province e la funzione obiettivo è definita dalla somma dei due seguenti termini:

- il termine di *interazione*, che ad ogni iterazione *j*, è espresso come:

$$J(P_j) = \sum_i \hat{MSE}_{ij} \qquad (4.1.1.3)$$

ossia è la somma, rispetto a tutte le province, dei valori del $\hat{MSE}$ delle stime EBLUP di povertà secondo l'approccio non stazionario;

e

- il termine di *penalità*, che formalizza i vincoli di vicinanza per evitare, il più possibile, la formazione di zone spaziali eterogenee; il termine di penalità è qui espresso mediante il modello di Potts (Sebastiani, 2003) e contiene le informazioni che ogni zona è un insieme di province limitrofe:

$$-\beta \sum_{iv} \mathbf{1}(k_i, k_v) \qquad (4.1.1.4)$$

Nella (4.1.1.4) la funzione **1** vale 1 se e solo se le province *i* e *v* sono limitrofe e appartengono alla stessa zona.

4.1.2 La procedura di smoothing adattivo

Proposta da Polzehl et al. (2000, 2003) ed adattata ai modelli ad effetti misti da Benedetti et al. (2007b), la procedura di *smoothing* adattivo dei pesi (*Adaptive weights smoothing procedure*, AWS) è una tecnica che calcola, in maniera iterativa, una matrice di pesi con lo scopo di separare zone dove le stime dei parametri locali sono statisticamente

differenti. Nel contesto di questo lavoro, AWS è utilizzata per stimare, in maniera iterativa, i parametri di regressione dello stimatore EBLUP prima definito, per ogni singola provincia. In particolare, nella fase iniziale, il vettore dei parametri di regressione dello stimatore EBLUP viene calcolato per ogni provincia *i* mediante i seguenti pesi:

$$w_{ik}(0) = e^{-\alpha d_{ik}} \tag{4.1.2.1}$$

dove d_{ik} è la distanza tra le province *i* e *k*. Ad ogni iterazione *j+1*, AWS stima i parametri di regressione dello stimatore EBLUP per ogni provincia *i*, aggiornando i pesi della precedente iterazione *j*:

$$w_{ik}(j+1) = a\, w_{ik}(j) + (1-a)\, e^{-\gamma T_{ik}} e^{-\frac{a}{n^\circ\ iterazioni} d_{ik}} \quad 0 \le a \le 1 \tag{4.1.2.2}$$

dove

$$e^{-\gamma T_{ik}}$$

è il termine di *penalità statistica*, ossia una funzione decrescente delle differenze statistiche T_{ik} calcolate per ogni coppia *i,k* di vettori di parametri di regressione e

$$e^{-\frac{a}{n^\circ\ iterazioni} d_{ik}}$$

è il termine di *penalità locale,* ossia una funzione decrescente della distanza tra le province *i* e *k*. La procedura AWS si arresta quando i pesi w_{ik} sono uguali o all'unità o a zero; i pesi con valore uguale all'unità identificano le province appartenenti alla medesima zona.

4.2 Risultati dell'approccio di non stazionarietà per l'indagine EU-SILC

La sperimentazione eseguita con l'*annealing* simulato individua cinque zone di stazionarietà e le province risultano raggruppate come è riportato nella Fig. 3. La sperimentazione, eseguita con la procedura di *smoothing* adattivo dei pesi, suddivide invece le province in sei zone di stazionarietà come è possibile osservare in Fig. 4.

In entrambi gli approcci di non stazionarietà allo stimatore EBLUP, si ottengono stime di povertà relativa a livello provinciale con valori di MSE decisamente inferiori ai corrispondenti valori per stime di povertà ottenute da un approccio EBLUP tradizionale, ossia di tipo stazionario; in particolare, la riduzione più consistente si verifica per la prima componente (g1) del MSE.

Nelle Fig. 5 e 6 sono riportati alcuni risultati rispettivamente dell'approccio mediante *annealing* simulato e mediante la procedura di *smoothing* adattivo dei pesi; si osserva che è consistente il numero delle province per le quali si verifica una riduzione del MSE delle stime di povertà con particolare riferimento alla prima componente g1 del MSE.

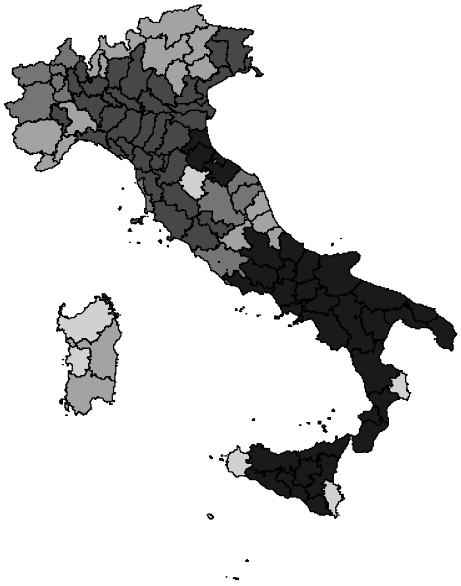
Entrando nel dettaglio, nel caso dell'*annealing* simulato, nel primo grafico in alto della Fig. 5, si evince che più di 40 province sono interessate da una riduzione superiore al 90% del MSE rispetto alle tradizionali stime EBLUP; per più di 20 province si verifica una riduzione del MSE delle stime di povertà compresa tra l'80% e il 90%. Ciò significa che più della metà delle province italiane sono interessate da una consistente aumento dell'efficienza delle stime EBLUP di povertà secondo l'approccio non stazionario. Analogamente per la procedura di *smoothing* adattivo dei pesi, si osserva nel primo grafico in alto della Fig. 6 che in quasi la metà delle province si realizza una

riduzione del MSE superiore al 40%. Dai grafici centrali in alto delle Fig. 5 e 6, si evince che in entrambe le sperimentazioni la riduzione del MSE si realizza per la maggior parte sulla prima componente (g1): nella quasi totalità delle province nel caso dell'*annealing* simulato e per una consistente parte di queste nel caso della procedura di *smoothing* adattivo dei pesi. Inoltre si può osservare quanto precedentemente affermato: l'approccio non stazionario applicato allo stimatore EBLUP riduce consistentemente il valore del MSE delle stime di povertà rispetto ad un approccio tradizionale di EBLUP. Si osservi come i valori del MSE (asse delle ordinate) delle stime EBLUP non stazionarie risultino decisamente inferiori ai corrispondenti valori di MSE per stime EBLUP tradizionali (asse delle ascisse) sia globalmente (primo grafico in basso delle Fig. 5 e 6) sia in particolare sulla prima componente del MSE (g1) (grafico centrale in basso delle Fig. 5 e 6).

Infine, si può infine osservare come l'*annealing* simulato fornisca valori di MSE inferiori rispetto alla procedura di *smoothing* adattivo dei pesi; ciò è spiegato dalla natura computazionale dell'*annealing* simulato che risulta basato sulla minimizzazione di una funzione obiettivo definita in questo lavoro proprio in funzione dei valori del MSE.

Infine è opportuno aggiungere che, per quanto il confronto tra MSE e stime dirette di varianza abbia i suoi limiti dovuti alle diverse ipotesi sottostanti, si rileva che il sensibile guadagno di efficienza delle stime di povertà ottenuto attraverso l'approccio non stazionario rispetto ad un'applicazione tradizionale di EBLUP, viene ovviamente confermato anche rispetto alle stime di povertà da disegno (o dirette).

Figura 3- Zone di stazionarietà delle province secondo l'annealing simulato



Fonte: Elaborazione su dati Istat

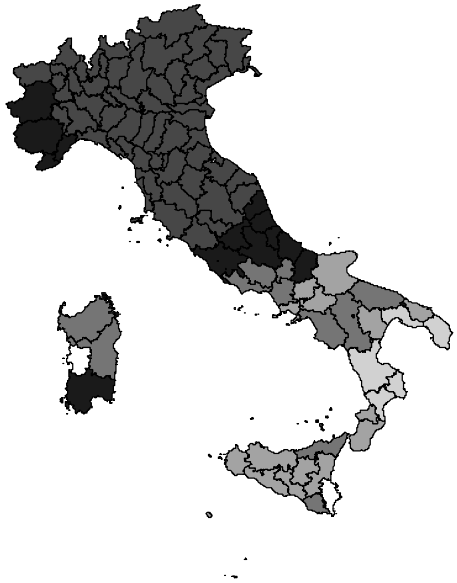
5. Conclusioni

Si fa sempre più consistente la richiesta di stime di *povertà relativa* per livelli geografici sub-nazionali; la stima della povertà mediante le sole informazioni provenienti dalle indagini campionarie può comportare problemi in termini di attendibilità. Le stime da disegno (o dirette) della povertà possono risultare, infatti, affette da errori di campionamento elevati quanto più sono disaggregati i domini rispetto ai quali sono calcolate. Per tale motivo, la stima della *povertà* secondo un approccio di *piccole aree* (SAE) sembra più opportuna per migliorarne l'efficienza utilizzando adeguate informazioni ausiliarie in aggiunta ai dati rilevati dalle indagini. Ma anche questo approccio può non essere sufficiente per ottenere stime efficienti.

E' stato qui verificato che un tradizionale approccio EBLUP alla povertà non comporta rilevanti guadagni di efficienza delle stime di povertà a livello regionale nel contesto dell'indagine sui Consumi delle Famiglie e a livello provinciale nell'ambito dell'indagine EU-SILC. Tuttavia nel caso dell'indagine EU-SILC, si profila la possibilità di utilizzare la *territorialità* della povertà per operare in maniera più incisiva sull'efficienza delle stime.

Tale impostazione nasce dall'osservazione che una peculiarità del fenomeno povertà è

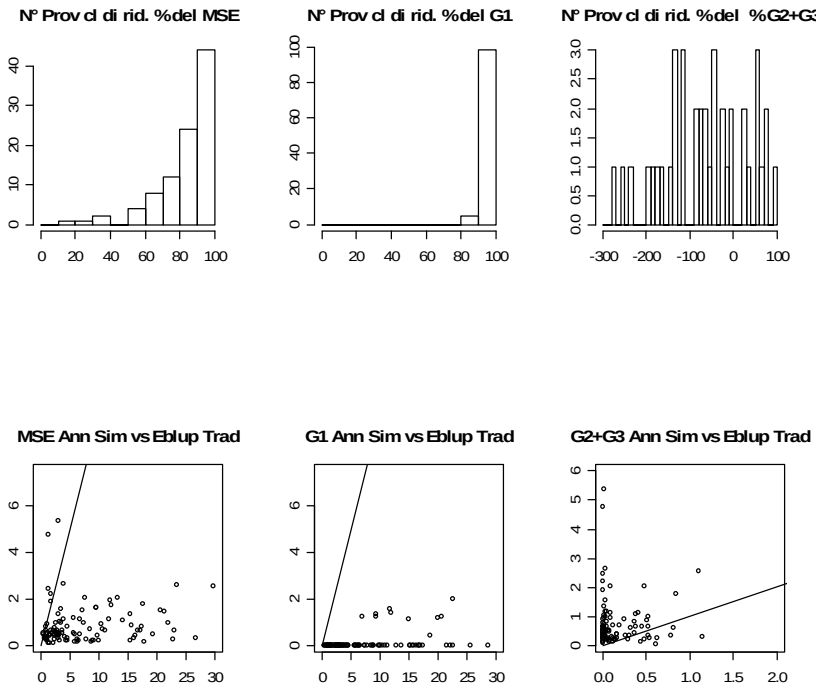
Figura 4 - Zone di stazionarietà delle province secondo la procedura di smoothing adattivo dei pesi



Fonte: Elaborazione su dati Istat

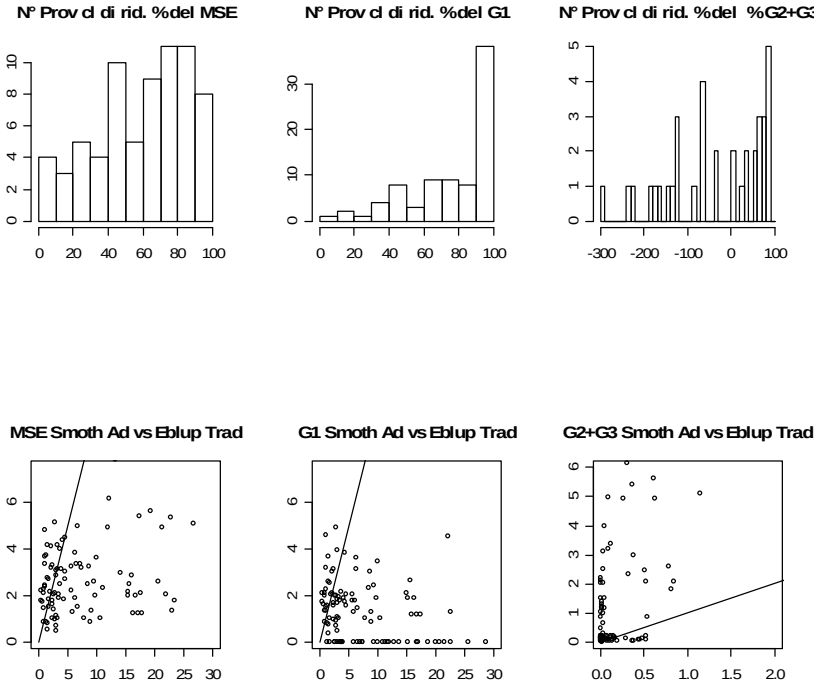
la sua connotazione rispetto al territorio; i valori della povertà risultano essere influenzati dal contesto geografico in cui si realizzano. Per questo motivo e considerato il fatto che si hanno a disposizione più aree geografiche su cui operare (le province), con riferimento all'indagine EU-SILC, è stato implementato un approccio di *non stazionarietà* rispetto ad un tradizionale stimatore EBLUP. Le due sperimentazioni effettuate hanno consentito di verificare che l'introduzione della non stazionarietà consente invece di identificare zone omogenee di province per le quali conseguentemente si ottengono stime di povertà più efficienti. Le sperimentazioni qui presentate, costituiscono una base di partenza nel contesto di un più ampio panorama che riguarda la problematica in oggetto. E' plausibile che altri studi possano essere eseguiti al fine di migliorare quanto è stato già evidenziato nel presente lavoro. In particolare, gli stimatori per piccole aree dipendono da elementi quali la variabile, il modello, le covariate adottati; per questo motivo, sono possibili futuri studi di simulazione in base ai quali valutare le *performance* degli stimatori in oggetto.

Figura 5 - Output dell'annealing simulato



Fonte: Elaborazione su dati Istat

Figura 6 - Output della procedura di smoothing adattivo dei pesi



Fonte: Elaborazione su dati Istat

Riferimenti bibliografici

Benedetti R., Pauselli C., Rinaldelli C. (2003), "Stime regionali di povertà relativa: un'applicazione dello stimatore EBLUP", *manoscritto non pubblicato.*

Benedetti R., Pratesi M., Salvati N. (2007a), "Local Stationarity in Small Area Estimation Models", *submitted*.

Benedetti R., Pratesi M., Salvati N. (2007b), "Adaptive Weights Smoothing With Applications To Small Area Estimation", *submitted*.

Benedetti R., Rinaldelli C. (2007a), "Local stationarity in EBLUP estimation of poverty parameters", (*abstract*) SAE2007 IASS Satellite Meeting on Small Area Estimation, University of Pisa, 3-5 Settembre 2007.

Benedetti R., Rinaldelli C. (2007b), "La povertà relativa secondo un approccio di stazionarietà locale per piccole aree", *Scritti di Statistica Economica 14, Quaderni di discussione n.30. Università degli Studi di Napoli* "Parthenope", Dipartimento di Statistica e Matematica per la Ricerca Economica.

Besag J. (1974), "Spatial interaction and the statistical analysis of lattice systems", *Journal of the Royal Statistical Society*, B, 36, 192-236.

Brackstone G.J. (1987), "Small Area Data: Policy Issues and Technical Challenges", in Platek R., Rao J.N.K., Särdnal C. E. E Singh M.P. (Eds), *Small Area Statistics,* Wiley, New York, 3-20.

Coccia G., Pannuzi N., Rinaldelli C., Vignani D. (2002), "Verso una misura della povertà regionale: problemi e strategie", *Sesta Conferenza Nazionale di Statistica*, Roma 6-8 Novembre 2002, in : www.istat.it.

Coccia G., Pannuzi N., Rinaldelli C. (2005), "Poor and non poor households: the estimation from sample surveys", (Invited paper) *Classification and Data Analysis 2005*, Book of Short papers, Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, Cladag2005, MUP editore.

Commission Regulation (EC) No 28/2004 of 5 January 2004 *regarding the detailed content of intermediate and final quality report. (9.1.2004 L 5/42 Official Journal of the European Union).*

Cressie N. (1991), *Statistics for spatial data*, New York, Wiley.

Deville J.C. e Särndal C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association* , vol. 87, 376-382.

Di Consiglio L., Falorsi S., Paladini P., Righi P., Scavalli E., Solari F. (2003), "Stimatori per piccole aree per le stime di povertà regionali", *Rivista di Statistica Ufficiale*, 2/2003.

European Parliament and Council Regulation (EC) No 1177/2003 of 16 June 2003 *concerning Community statistics on income and living conditions (EU-SILC). (3.7.2003 L 165/1 Official Journal of the European Union).*

Eurostat (2004), *Common cross-sectional EU indicators based on EU-SILC; the gender pay gap*, Doc. EU-SILC N.131/04.

Falorsi S., Rinaldelli C. (1998), "Un software generalizzato per il calcolo delle stime e degli errori di campionamento", *Statistica Applicata*, Vol. 10, N. 2, 217-234.

Fay R.E. e Herriott R.A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data", *Journal of the American Statistical Association*, Vol. 74, 269-277.

Geman D., Geman S., Graffigne C., Dong G P. (1990), "Boundary detection by constrained optimization", *IEEE PAMI*, 12, 609–628.

Ghosh M., Rao J.N.K. (1994), "Small Area Estimation: an Appraisal", *Statistical Science*, Vol. 9, 55-93.

Jiang J. (1996), "REML estimation: asymptotic behaviour and related topics", *Annals of Statistics*, 24, 255-286.

Istat (2002), *La stima ufficiale della povertà in Italia 1997-2000*, Argomenti n.24.

Istat (2003a), *La povertà e l'esclusione sociale nelle regioni italiane*, Statistiche in Breve, 17 dicembre 2003.

Istat (2003b), *Forze di lavoro*, *Media 2002*, Annuario n.8.

Istat (2004a), *I trattamenti pensionistici, Anno 2002*, Annuario n.3.

Istat (2004b), *Forze di lavoro*, *Media 2003*, Annuario n.9.

Kirkpatrick S., Gelatt Jr. C.D., Vecchi M.P. (1983), "Optimization by simulated annealing", *Science*, 220, 671-680.

Moretti D., Pauselli C., Rinaldelli C. (2005), "La stima della varianza campionaria di indicatori complessi di povertà e disuguaglianza", *Statistica Applicata*, Vol. 17, N. 4, 529-550.

Pauselli C., Rinaldelli C. (2003), "La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute", *Rivista di Statistica Ufficiale,* 2/2003.

Pauselli C., Rinaldelli C. (2004), "Stime di povertà relativa: la valutazione dell'errore campionario secondo le Replicazioni Bilanciate Ripetute", *Statistica Applicata*, Vol.16, n.1.

Polzehl J., Spokoiny V. (2000), "Adaptive weights smoothing with applications to image segmentation", *Journal of Royal Stat. Society*, 62, B, 335-354.

Polzehl J., Spokoiny V. (2003), "Varying coefficient regression modelling", *Weierstrass Institute for Applied Analysis and Stochastics*, Berlin.

Purcell N.J., Kish L. (1979), "Estimation for small domains", *Biometrics*, 35, 365-384.

Rao J.N.K. (1999), "Small Area Estimation: Updates with Appraisal", *manoscritto non pubblicato*.

Rao J.N.K. (2003), *Small Area Estimation*, Wiley, New York.

Rinaldelli C. (2005), "Statistiche complesse e software"*, Statistica & Società*, Anno III, n.2, 01.2005, 27-29.

Rinaldelli C. (2006), "Experiences of variance estimation for relative poverty measures and inequality indicators", *COMPSTAT Proceedings in Computational Statistics*, 1465-1472, Italy 2006.

Sebastiani M.R. (2003), "Markov random-field models for estimating local labour markets", *Applied Statistics*, 52, 201-211.

Norme redazionali

La Rivista di Statistica Ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche Istat corredati da una nota informativa dell'Autore contenente: appartenenza ad istituzioni, attività prevalente, qualifica, indirizzo, casella di posta elettronica, recapito telefonico e l'autorizzazione alla pubblicazione firmata dagli Autori. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di un referente scelto tra gli esperti dei diversi temi affrontati. Gli originali, anche se non pubblicati, non si restituiscono.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file Template.doc disponibile on line o su richiesta. In base a tali standard la lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 30–35 pagine.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 12 righe); quelli in italiano dovranno prevedere anche un *Abstract* in inglese. La bibliografia, in ordine alfabetico per autore, deve essere riportata in elenco a parte alla fine dell'articolo. Quando nel testo si fa riferimento ad una pubblicazione citata nell'elenco, si metta in parentesi tonda il nome dell'autore e l'anno di pubblicazione. Ad esempio (Bianchi, 1987, Rossi, 1988). Quando l'autore compare più volte nello stesso anno l'ordine verrà dato dall'aggiunta di una lettera minuscola accanto all'anno di pubblicazione. Ad esempio (Bianchi, 1987a, 1987b).

Nella bibliografia le citazioni di libri e articoli vanno indicate nel seguente modo. Per i libri: cognome dell'autore seguito dall'iniziale in maiuscolo del nome, il titolo in corsivo dell'opera, l'editore, il luogo di edizione e l'anno di pubblicazione. Per gli articoli: dopo l'indicazione dell'autore si riporta il titolo tra virgolette, il titolo completo in corsivo della rivista, il numero del fascicolo e l'anno di pubblicazione. Nei riferimenti bibliografici non si devono usare abbreviazioni.

Nel testo dovrà essere di norma utilizzato il corsivo per le parole in lingua straniera e il corsivo o grassetto per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

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

La proprietà letteraria degli articoli pubblicati spetta alla Rivista di statistica ufficiale.

E' vietata a norma di legge la riproduzione anche parziale senza autorizzazione e senza citarne la fonte.

Per contattare la redazione delle pubblicazioni scientifiche Istat e per inviare lavori: rivista@istat.it. Oppure scrivere a:
Comitato di redazione delle pubblicazioni scientifiche
C/O Carlo Deli (cadeli@istat.it)
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012008003000000

ISSN 1828-1982

€ 10,00