VOLUME LXX – N. 1 GENNAIO-APRILE 2016

RIVISTA ITALIANA DI ECONOMIA DEMOGRAFIA E STATISTICA

Sede Legale
C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA

sieds.new@gmail.com
rivista.sieds@gmail.com

Volume pubblicato con il contributo della
Fondazione della Cassa Di Risparmio di Fermo

INDICE

MISURARE IL RACKET: UNA PROPOSTA METODOLOGICA SPERIMENTATA NEL "LABORATORIO" MILANESE

Gian Carlo Blangiardo, Simona Maria Mirabelli, Stefania Rimoldi, Laura Terzera

1. Introduzione

La criminalità è un fenomeno di forte impatto sociale che necessita non solo di continue riletture, ma anche di nuovi strumenti di rilevazione e analisi indispensabili per chi deve garantire condizioni di sicurezza e legalità negli spazi in cui vivono le persone (siano essi imprenditori o semplici cittadini). In particolare, risulta prioritaria l'esigenza di indagare la percezione sulla diffusione della criminalità nei luoghi in cui operano le imprese, per il mondo imprenditoriale *in primis* ma anche per le istituzioni delle realtà coinvolte (Cnel 2010; Pignatone 2012; Università Bocconi 2014; Unioncamere e Istituto Tagliacarne 2015). Ciò vale soprattutto per i reati legati al mondo della criminalità organizzata che esercita il suo controllo estorcendo denaro (o altri beni) con l'uso della minaccia o dell'intimidazione, contribuendo così ad alimentare il cd. *numero oscuro* dei reati che non vengono denunciati alle forze dell'ordine[1]

> […] La quota di sommerso varia non solo strettamente in base alla tipologia di reato, ma anche alla sua riuscita, alla sua gravità, al danno economico, alle conseguenze fisiche subite. La denuncia viene sporta se ve ne è una convenienza, un beneficio o per tutelarsi a livello personale. Ci si rivolge alle Forze dell'ordine se si ha fiducia nel loro operato, se si pensa che sia utile a se stessi o alla comunità. Non ci si rivolge a loro se si pensa di essere male accolti, se si ritiene che sia del tutto inutile o una perdita di tempo (Istat, 2011: 142).

A partire da un'indagine avviata nel corso del 2014 sugli eventi legati alla criminalità, con particolare riferimento a quella che colpisce le attività economiche 'su strada', il presente contributo intende proporre un indicatore statistico con cui misurare la probabilità di subire un evento criminoso in corrispondenza di

[1] A tale riguardo si consideri che i dati sulle estorsioni denunciate nel 2010 risultano in calo rispetto all'anno precedente nella misura del 16,7% in Lombardia e del 18,4% a livello nazionale. La dinamica dei reati estorsivi evidenzierebbe, secondo quanto ipotizzato da S.O.S. Impresa nel suo recente Rapporto, "*come il 'pizzo' continui ad essere una pratica diffusa, quanto sommersa, per il concatenarsi di diversi fattori, prima tra tutti quello di un livello di omertà ancora molto alto anche in zone non sospette*" (S.O.S, 2011: 9).

specifiche realtà territoriali. La metodologia adottata consente di ottenere stime puntuali sull'estensione del fenomeno, sulla sua natura e incidenza, nonché sulla sue peculiarità a livello territoriale e settoriale.

2. Metodologia e strumenti di indagine

La base dati cui si fa riferimento proviene da uno studio condotto nel 2014 nella provincia di Milano volto a indagare il livello di sicurezza in cui operano gli imprenditori del territorio[2]. La ricerca si è svolta attraverso la somministrazione di un questionario anonimo a un campione di imprenditori (pari 74.478 imprese, circa la metà di quelle iscritte nei registri della Camera di Commercio di Milano) appartenenti ai tre settori di attività selezionati: commercio, turismo (comprendente anche le attività di pubblico esercizio) e servizi[3]. Le domande del questionario vertono sui seguenti aspetti per un complesso di 119 variabili:

A. *Caratteristiche strutturali dell'impresa e dell'imprenditore di riferimento*.
   Attività. Dimensione dell'impresa. Sede dell'impresa (Cap). Anzianità dell'impresa. Paese di nascita del titolare/legale rappresentante.
B. *Condizioni di sicurezza e fenomeni criminosi*.
   Presenza di alcune specifiche realtà/condizioni che favoriscono fenomeni criminosi nell'area in cui si esercita l'attività. Fenomeni criminosi dei quali si è avuto notizia nel corso dell'ultimo anno nell'area in cui si esercita l'attività. Fenomeni criminosi dei quali si è stati vittima nel corso dell'ultimo anno nell'area in cui si esercita l'attività. Quali sono le cause che determinano i fenomeni criminosi nell'area in cui si esercita l'attività? Che dinamica hanno subito negli ultimi tre anni i fenomeni criminosi nell'area in cui si esercita l'attività? Quali iniziative sono valse efficaci nel contrastarli? Quali altre misure servirebbero?
   Come reazione al rischio criminalità si è considerata l'idea di cedere o trasferire l'attività? Quali sono le opinioni/atteggiamenti circa il possesso di un'arma da fuoco per difesa personale?
C. *Esperienza e caratteristiche degli eventi criminosi subiti*.

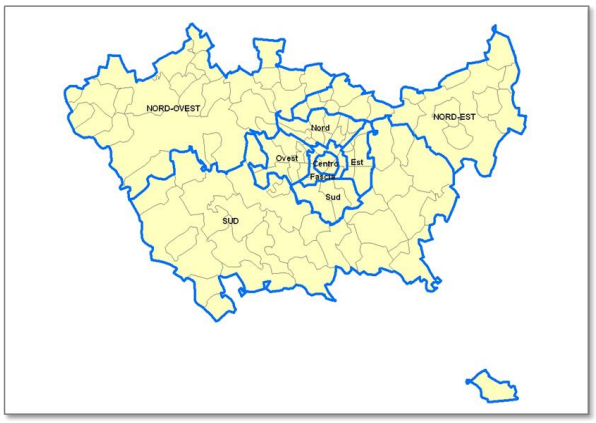
[2] Il numero di questionari restituiti all'Ente promotore (Confcommercio Imprese per l'Italia Milano-Lodi-Monza e Brianza) ammonta a 4.126, pari al 5,5% dell'universo di riferimento.

[3] La rilevazione è avvenuta tramite questionario postale auto-compilato da parte dell'imprenditore di riferimento (o di un suo delegato), preventivamente informato sull'iniziativa promossa da ConfCommercio. Le domande tradotte in sei lingue (italiano, francese, inglese, spagnolo, cinese e arabo) sono state strutturate come quesiti a risposta chiusa al fine di agevolarne la compilazione. La restituzione è avvenuta in forma anonima (busta pre-affrancata)

Minacce o intimidazioni: tipo; periodo; reazioni a seguito alle richieste; misure attivate a protezione; eventuale denuncia alle Autorità; provenienza delle minacce/intimidazioni. Episodi di corruzione/concussione da parte di funzionari pubblici, pubblici ufficiali, figure ispettive: tipo; reazione; eventuale denuncia alle Autorità.

Grazie alla specifica del codice di avviamento postale (Cap) relativo alla sede dell'impresa, le unità statistiche sono state geo referenziate, consentendo così l'analisi degli aspetti differenziali con cui si manifestano e sono percepiti i fenomeni criminosi sul territorio. La classificazione dei Cap della provincia di Milano per macro zone (figura 1) si articola in 9 partizioni territoriali: 6 relative alla città di Milano (Milano Centro, Milano Fascia pericentrale, Milano Est, Milano Sud, Milano Ovest, Milano Nord) e 3 a copertura degli altri Cap della provincia (Nord Ovest, Nord Est, Sud)[4].

Figura 1 – *Partizione del territorio della provincia di Milano nelle 9 macro zone ottenute per aggregazione dei territori definiti dai Codici di avviamento postale*



Al fine di tenere conto della distorsione derivante dall'autoselezione presente nel campione di rispondenti – con particolare riferimento alla localizzazione territoriale (Cap) e al settore di appartenenza (attività) – è stato predisposto un sistema di pesi che, con riferimento alle due variabili localizzazione e settore di attività, consente di ripristinare la struttura dell'universo.

[4] La struttura territoriale dei Cap ha reso talvolta problematico rispettare, nella definizione delle macro zone, sia la loro localizzazione geografica che la loro contiguità. La partizione riportata nella figura 1, e adottata nel seguito ai fini delle analisi, è parsa quella più soddisfacente e altresì capace di garantire un adeguato numero di casi campionari in corrispondenza di ogni unità territoriale.

Avendo l'obiettivo di pervenire ad una stima della probabilità di subire un evento criminoso (nel caso specifico minacce e intimidazioni oppure il coinvolgimento in episodi di corruzione e concussione da parte di soggetti della P.A.) si è operato, in modo da tenere conto che, per la natura stessa dei fenomeni indagati, la rilevazione è stata esposta a sperimentare un basso livello di restituzioni dei questionari.

In particolare, si sono definite, con le opportune specificazioni per ogni ambito territoriale e settoriale:

P(E=Si) = la "Probabilità di aver subito l'evento che forma oggetto diinteresse";
P(Q=R) = la probabilità che un soggetto contattato risponda al questionario (Q=R);
P(Q=R & E=Si) = la probabilità che un soggetto contattato risponda al questionario (Q=R) e dichiari di aver subito l'evento (E=Si);
P(Q=R | E=Si) = la probabilità che un soggetto che ha subito l'evento (E=Si) risponda al questionario (Q=R);
P(E=Si | Q=R) = la probabilità che un soggetto che risponde al questionario (Q=R) abbia subito l'evento (E=Si);

e, stante l'uguaglianza:

P(Q=R & E=Si) = P(Q=R | E=Si) * P(E=Si) = P(E=Si | Q=R) * P(Q=R),

si è ottenuta la seguente probabilità P(E=Si) che un soggetto, facente parte dell'universo indagato (riconducibile a uno specifico sottoinsieme definito da territorio e attività) abbia subito l'evento:

[1] $$P(E=Si)=\frac{P(E=Si|Q=R)\cdot P(Q=R)}{P(Q=R|E=Si)}$$

Dove i due fattori al numeratore del secondo membro sono stimabili, rispettivamente, con la frequenza relativa di rispondenti che dichiarano l'evento e con la frequenza relativa di rispondenti nel complesso della popolazione contattata, mentre l'elemento a denominatore andrebbe determinato in modo esogeno rispetto all'indagine. Tuttavia, ove si ritenga (per altro con buona dose di ragionevolezza), che la probabilità di rispondere all'indagine da parte di chi ha subito l'evento criminoso, ossia P(Q=R | E=Si), possa non risultare significativamente diversa al variare dell'ambito territoriale e dell'attività che di volta in volta si considerano,

sarà comunque possibile cogliere gli aspetti differenziali delle probabilità cui si è interessati semplicemente confrontando i risultati del prodotto al numeratore della [1]. Così facendo, e introducendo un appropriato insieme di numeri indice che adottano come base fissa il dato relativo al totale provinciale per il complesso di tutte le attività, sono state realizzate le rappresentazioni cartografiche del rischio con cui gli imprenditori che operano sul territorio della provincia di Milano risultano aver ricevuto minacce e/o intimidazioni da esponenti della criminalità, ovvero essere stati coinvolti in episodi di corruzione e concussione da parte di politici e funzionari della Pubblica Amministrazione. L'indicatore della probabilità di aver subito l'evento è dunque espresso in termini relativi, ossia assumendo come elemento di confronto il corrispondente indicatore medio riferito al complesso della provincia e all'insieme dei tre settori considerati. Pertanto, in ogni ambito territoriale e settoriale tale indicatore risulterà[5] superiore (o inferiore) a 100 ogni qualvolta il rischio di subire l'evento in oggetto ricorrerà con intensità superiore (o inferiore) al livello medio provinciale (calcolato senza distinzione di settore di attività).

3. I risultati dell'indagine

3.1 La diffusione del fenomeno tra coloro che operano nel territorio milanese

A partire da tali premesse, sono state analizzate le caratteristiche delle imprese che hanno partecipato all'indagine al fine di definirne il profilo rispetto all'appartenenza settoriale, alla dimensione e alla localizzazione.

La lettura dei dati per dettaglio territoriale (Figura 1) segnala una certa prevalenza dei servizi nella città capoluogo (in particolare nel suo Centro) e del commercio negli altri comuni della provincia. Il settore turismo aggrega poco meno del 20% delle imprese e si localizza con presenze relativamente più alte nell'area Sud di Milano. Riguardo alla specifica delle attività di commercio, quelle di vendita al dettaglio non alimentare coprono quasi il 70% del totale provinciale e fino all'80% nel Centro di Milano. La stessa area in cui sono meno presenti le attività di dettaglio alimentare che, invece, si distribuiscono altrove quasi ovunque con quote comprese tra il 15% e il 18%. Generalmente più esterna al comune capoluogo risulta essere l'attività di commercio all'ingrosso, con l'eccezione di Milano Est per quello alimentare.

[5] Stante l'ipotesi di cui si è detto circa la relativa invarianza territoriale e settoriale di P(Q=R | E=Si).

Per le attività classificate come "turismo" si osserva ovunque la predominanza dei pubblici esercizi, relativamente più frequenti fuori dal capoluogo, mentre l'incidenza di alberghi è significativa a Milano sia nella fascia pericentrale che a Ovest; le agenzie di viaggio hanno un peso importante al Centro e nell'area Est della città. Tra i pubblici esercizi sono i bar ad assumere un particolare rilievo fuori da Milano (nell'area Nord-Est e Sud della provincia), così come i ristoranti lo sono entro il capoluogo. I locali serali ricorrono più frequentemente tra i pubblici esercizi nella fascia pericentrale di Milano e a Nord-Ovest e Nord-Est nel resto della provincia. Infine, riguardo ai servizi appare significativo il rilievo di quelli di costruzione e intermediazione immobiliare, soprattutto al Centro e a Milano Ovest, al pari di quelli di elaborazione, progettazione e assistenza tecnico-informatica a Milano Sud. Fortemente spostati a Nord e extra-capoluogo risultano i servizi legati alla logistica, mentre al Centro e nella sua prima cintura si collocano in via preferenziale la gestione e amministrazione di immobili. I servizi di consulenza professionale, quand'anche più distribuiti, sembrano localizzarsi maggiormente nelle aeree centrali e a Est e Sud del capoluogo.

Relativamente alle condizioni di contesto in cui operano gli imprenditori attivi nella provincia, esaminando le diverse forme di degrado sociale e di marginalità, i rispondenti segnalano con maggiore frequenza sia la presenza di nomadi (secondo un imprenditore su due della città di Milano tale presenza favorirebbe fenomeni criminosi nell'area in cui viene esercitata l'attività), sia il fenomeno dei negozi sfitti: gli imprenditori dei comuni extra-capoluogo lo segnalano come espressione di degrado (o semplice decadimento legato alla crisi economica) nel 54,2% delle risposte fornite. Anche la presenza di venditori ambulanti viene percepita come una realtà che può generare condizioni di illegalità, soprattutto per le aziende della fascia pericentrale e orientale del comune capoluogo. Lo spaccio di droga e la presenza di tossicodipendenti viene segnalata maggiormente nei comuni extra-capoluogo, sebbene nell'area Nord e Sud di Milano se ne preoccupi almeno un terzo degli imprenditori che vi operano.

Tra le azioni criminose di cui gli imprenditori del territorio hanno avuto maggiormente notizia durante l'ultimo anno si evidenziano i furti in appartamento: essi ricorrono nel 58,3% delle risposte fornite nel corso dell'indagine, con punte che sfiorano i due terzi tra gli imprenditori di aziende che operano nei comuni extra-capoluogo. Ugualmente significativa la quota di risposte relative alla notizia di atti vandalici: essa ricorre in quasi il 40% dei casi, con valori che si attestano al 49% in corrispondenza delle zone settentrionali della città di Milano.

Passando dalla conoscenza degli eventi legati alla criminalità all'esperienza vissuta in prima persona dagli imprenditori si osserva, coerentemente con quanto già rilevato, la quota più elevata di risposte affermative in relazione ai reati

predatori che si consumano soprattutto negli esercizi commerciali e nei luoghi pubblici.

Riguardo alla dinamica dei fenomeni di interesse, due imprenditori su cinque ritengono che negli ultimi tre anni gli eventi legati alla criminalità sono aumentati: soltanto il 6,7% dei rispondenti coglie un miglioramento delle condizioni di contesto.

Relativamente alle caratteristiche degli eventi criminosi subiti, circa il 12% dei rispondenti dichiara di aver ricevuto minacce o intimidazioni[6], con valori mediamente più alti nei comuni extra-capoluogo, in particolare nella zona Nord Est della provincia (14,2%). Il danneggiamento a cose rappresenta l'episodio più ricorrente (41%), seguito dalle minacce con visite e telefonate[7].

Passando agli episodi di concussione, quasi il 10% degli imprenditori dichiara di esserne stato coinvolto: i tentativi sono andati a "buon fine" mediamente nel 41,8% dei casi, ma la quota sale fino al 64,3% nell'area Est del capoluogo e il fenomeno è ricorrente in un caso su due nei comuni Nord-Ovest extra-capoluogo. Gli stessi risultati dettagliati per settore di attività evidenziano come gli imprenditori turistici e quelli del pubblico esercizio abbiano accettato con maggiore frequenza le richieste illegittime di denaro: in almeno un caso su due a livello provinciale e nella totalità dei casi nella zona orientale di Milano città. I funzionari concussi sono stati denunciati dagli imprenditori coinvolti solo nell'8,3% dei casi a livello provinciale, e in soli 3 casi su cento da coloro che operano nell'area settentrionale del comune capoluogo e in quella nord-orientale extra-città.

3.2 La mappa del rischio di subire eventi criminosi

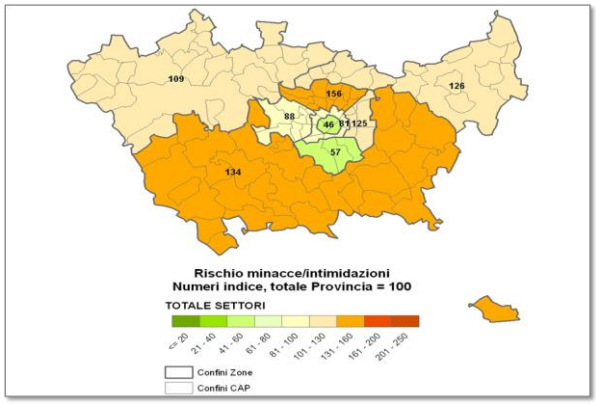
La costruzione dell'indicatore con cui stimare la probabilità di subire minacce e/o intimidazioni da parte di esponenti della criminalità (ovvero di essere coinvolti in episodi di corruzione e/o concussione da parte di politici o funzionari pubblici) consente la realizzazione della 'mappa del rischio'. La rappresentazione cartografica del rischio di estorsione cui sono esposti gli imprenditori che operano nella provincia di Milano evidenzia i valori massimi nell'area Nord del capoluogo

[6] Il dettaglio per settore di attività e area territoriale pone in evidenza la diversa frequenza con cui vengono riportate esperienze estorsive nei settori di interesse. a fronte di una media provinciale dell'11,9%, la percentuale di imprenditori appartenenti al settore turistico risulta superiore di circa due punti percentuali (13,7%), con valori che raggiungono mediamente il 15% nei comuni extra-capoluogo.

[7] Secondo il già citato Rapporto di S.O.S. Impresa, vi sono diverse modalità con cui la criminalità compie reati di natura estorsiva: si va dalla classica e consolidata "messa a posto" all'approvvigionamento gratuito di beni e servizi, al c.d. "cavallo di ritorno" che consiste nel furto di beni (automobili, mezzi agricoli, ecc.) che vengono restituiti solo dopo il pagamento di una tangente.

(in cui il rischio supera del 56% il valore medio provinciale) e in quella Sud del restante territorio provinciale (+34%) (Figura 2). Sul fronte opposto, il Centro di Milano si caratterizza come area meno esposta, e viene in tal senso affiancata dal settore Sud della città. Seguono quindi - ma con livelli chiaramente più alti (ancorché inferiori alla media provinciale) – la corona del Centro di Milano e la sua area Ovest. Fuori dal capoluogo lombardo il rischio appare più accentuato a Est (dove supera del 26% il valore medio provinciale) che a Ovest (+9%).

Figura 2 – *Indicatori del rischio di subire minacce e/o intimidazioni nelle 9 macro zone della provincia di Milano. Numeri indice base: totale provincia tutti i settori = 100.*



Il rischio di subire intimidazioni e minacce sembra essere decisamente più frequente nel settore commerciale[8] (Figura 3). Gli operatori di tale settore si collocano al di sopra del livello medio con cui il fenomeno colpisce l'imprenditoria provinciale in tutte le aree, con la sola eccezione di Milano Sud e di una quasi parità per il Centro del capoluogo. La condizione peggiore è quella che riguarda i commercianti dell'area Milano Nord, in cui il livello di rischio è quasi due volte e mezzo quello medio, ma non risparmia neppure chi opera a Sud della provincia (+91% rispetto al rischio medio) e quasi altrettanto l'area a Est (+77%) e a Ovest (+66%).

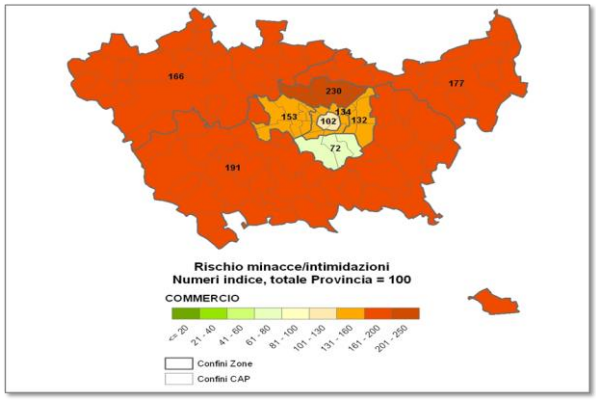
Decisamente più contenuto risulta essere, pressoché ovunque, il rischio di coloro che operano nel settore turistico (comprensivo dei pubblici esercizi), in corrispondenza dei quali gli indicatori superano il livello medio solo a Milano Nord con qualche rilievo e, più modestamente, a Milano Est e nel Nord dei comuni extra

[8] Il fatto che il livello del rischio nel settore commerciale sia più alto di quello nel settore turistico, benché per le corrispondenti frequenze tra gli intervistati il rapporto sia invertito, si giustifica con i diversi valori della probabilità di partecipazione all'indagine richiamata nella formula [1].

capoluogo. In questo settore si distinguono, per il modesto livello di rischio relativo, l'area di Milano Ovest e, in tono minore, quella di Milano Sud.

Ancora generalmente più favorevole appare la distribuzione dei rischi di minacce e intimidazioni in corrispondenza di chi opera nel settore dei servizi. Solo a Milano Est l'indicatore si posiziona oltre il livello del rischio medio provinciale, mentre nel resto della città si distingue in positivo l'area del Centro, dove il valore è circa un decimo rispetto al dato di riferimento, ma anche quella che le fa da cintura, così come le aree Sud e Ovest nella città capoluogo, mostrano livelli assai contenuti. Fuori da Milano i rischi più alti si identificano nella parte meridionale della provincia, ma si tratta di valori che – soprattutto se comparati con quanto si rileva per gli operatori del commercio – possono ritenersi senz'altro contenuti.

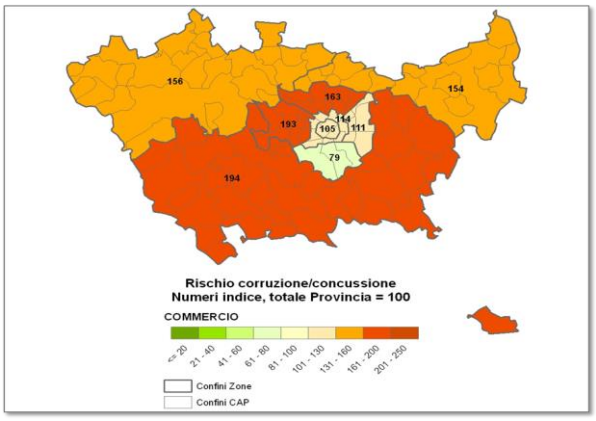
Figura 3 – *Settore Commercio. Indicatori del rischio di subire minacce e/o intimidazioni nelle 9 macro zone della provincia di Milano. Numeri indice base: totale provincia tutti i settori = 100.*



Sul fronte degli episodi di corruzione/concussione la visione d'insieme attribuisce a Milano una posizione più favorevole in corrispondenza del Centro e delle aree pericentrale e Sud, mentre segnala qualche livello critico a Nord-Est, con circa il 20-30% di rischio in più rispetto alla media provinciale. Fuori dal capoluogo è soprattutto l'area Sud a mostrare criticità (+29%), mentre il Nord (specie a Est) si colloca sostanzialmente in media. Anche per questo tipo di reato la condizione peggiore è generalmente riscontrabile per gli operatori del commercio (Figura 4). Se si escludono l'area Sud di Milano (che si mantiene nettamente sotto la media) e il suo Centro (sostanzialmente in media), in tutte le altre realtà territoriali i commercianti subiscono il rischio di coinvolgimento in episodi di corruzione/concussione decisamente in misura maggiore rispetto alle altre categorie.

La distanza dal livello medio provinciale (calcolato per l'insieme dei settori) raggiunge punte di oltre il 90% a Milano Ovest e nel Sud della provincia extra capoluogo. Un sovra rischio di circa il 60% si riscontra altresì nella fascia Nord, sia di Milano che degli altri comuni della provincia, mentre la criticità nella fascia attorno al centro di Milano e nella zona Est della città appare relativamente contenuta. In corrispondenza degli imprenditori del settore turistico ciò che colpisce, pur in presenza di un'incidenza generalmente più attenuata rispetto al settore del commercio, è la criticità della realtà milanese. Se infatti si esclude la zona Ovest della città e (almeno parzialmente) il suo Centro, il livello di rischio appare consistente non solo al Nord, ma anche nella fascia pericentrale, a Est e a Sud. Di fatto, per queste zone di Milano gli operatori del settore del turismo appaiono addirittura più esposti degli stessi commercianti. Del tutto diversa è la situazione fuori dal capoluogo, dove il rischio per il settore turismo si colloca leggermente al di sotto del livello provinciale a Est e a Sud e solo di poco al di sopra in corrispondenza dell'area Ovest.

Figura 4 – *Settore Commercio. Indicatori del rischio di essere coinvolti in episodi di corruzione/concussione nelle 9 macro zone della provincia di Milano. Numeri indice base: totale provincia tutti i settori = 100.*



Il rischio corruzione/concussione per gli operatori del settore servizi risulta decisamente ridotto quasi ovunque, con la singolare eccezione di coloro che svolgono l'attività nell'area di Milano Est. Questi ultimi presentano infatti un livello di rischio (+46% rispetto alla media provinciale) che è superiore a quello degli stessi commercianti e degli operatori del turismo che ne condividono il territorio. In ogni caso, anche per questo tipo di reato il mondo dei servizi sembra relativamente meno esposto. Si rilevano condizioni ottimali di rischio minimo

nell'area di Milano Centro e della sua cintura ma valori assai contenuti (circa la metà del dato medio provinciale) sono riscontrabili anche nelle zone Nord, Sud e Ovest del capoluogo lombardo, così come nel resto della provincia con livelli che sono ridotti attorno al 60-70% del dato medio provinciale.

Conclusioni

I risultati cui si è pervenuti dimostrano come l'applicazione degli strumenti statistici allo studio dei fenomeni criminosi consenta di indagare i diversi aspetti con cui essi si manifestano nel territorio, non solo sotto il profilo oggettivo (frequenza dei reati denunciati, numero dei detenuti presenti negli istituti penitenziari, delle vittime accertate rispetto alla popolazione di riferimento, ecc.), ma anche in termini soggettivi (percezione di allarme sociale da parte degli individui, livello percepito di legalità e sicurezza del proprio contesto operativo, ecc.). Da questo punto di vista le informazioni raccolte tra gli imprenditori della provincia di Milano contribuiscono ad arricchire il sistema di conoscenze acquisite sul fenomeno di interesse. I dati confermano l'ipotesi che la percezione della criminalità e delle forme di marginalità e di degrado sociale che ne favoriscono la propagazione variano, sensibilmente, in relazione all'ambito territoriale in cui l'impresa svolge la propria attività e al settore economico cui la stessa afferisce.

Il dettaglio per area e tipo di attività sottolinea dunque la diversa frequenza con cui vengono riportate le esperienze di minacce (o intimidazioni) e gli episodi di corruzione e concussione nei tre settori di interesse. Cambia in modo evidente la distribuzione del rischio di subire eventi criminosi tra gli operatori. Il livello risulta mediamente più alto rispetto a quello medio provinciale se si appartiene al settore del commercio, se si opera nell'area Nord della città di Milano e in quella Sud della restante provincia milanese, in quella Est (limitatamente al settore dei servizi) e in quella peri-centrale (settore del turismo e del pubblico esercizio). Più contenuto è invece il rischio percepito sia dagli operatori del Centro del comune capoluogo, sia in genere da chi appartiene al settore dei servizi.

Riferimenti bibliografici

CONSIGLIO NAZIONALE DELL'ECONOMIA E DEL LAVORO. Osservatorio socio-economico sulla criminalità. 2010. *L'infiltrazione della criminalità organizzata di alcune regioni del Nord Italia*, Sintesi, 23 febbraio 2010.

FRIGERIO L. 2009. Le mafie all`ombra del Duomo. In *Aggiornamenti sociali: rivista mensile a schede*, Fasc. 11, pp. 674-685.

GRECO. Group of States against corruption. 2009. *Evaluation Report on Italy*. *Joint First and Second Evaluation Round*, 2 luglio 2009.

ISTAT. 2011. *Noi Italia. 100 statistiche per capire il Paese in cui viviamo*, www.istat.it.

PIGNATONE G. 2012. Criminalità', economia e legalità: il Nord e il Sud. In *Aggiornamenti sociali*, Vol. 63, N. 6 giugno, pp. 470-482.

S.O.S. IMPRESA. 2011. *Le mani della criminalità sulle imprese*. *XIII Rapporto*. *Focus Lombardia*, Roma: Aliberti Editore.

UNIONCAMERE, ISTITUTO TAGLIACARNE. 2015. *I fenomeni illegali e la sicurezza percepita all'interno del sistema economico italiano*, Maggio 2015, www.csr.unioncamere.it.

UNIVERSITA' COMMERCIALE "LUIGI BOCCONI" DIPARTIMENTI DI STUDI GIURIDICI "ANGELO SRAFFA", CREDI – CENTRO DI RICERCHE EUROPEE SUL DIRITTO E LA STORIA DELL'IMPRESA "ARIBERTO MIGNOLI". 2014. *L'espansione della criminalità organizzata nell'attività d'impresa al Nord*, Milano: Editore Luca Santa Maria, www.penalecontemporaneo.it.

SUMMARY

Statistics for measuring crime towards economic activities: an experimental test in the Milan area

Through a statistic survey on the phenomenon of crime towards the economic activities in the Milan area, this paper proposes a set of indicators of the likelihood of being a victim of a criminal event in correspondence of a specific context. The methodology allows to point out estimates on the extent of the phenomenon, its nature and incidence, as well as on its peculiarities at sectorial and territorial level.

Gian Carlo BLANGIARDO, Università di Milano-Bicocca, giancarlo.blangiardo@unimib.it
Simona Maria MIRABELLI, Università di Milano-Bicocca, simona.mirabelli@unimib.it
Stefania RIMOLDI, Università di Milano-Bicocca, stefania.rimoldi@unimib.it
Laura TERZERA, Università di Milano-Bicocca, laura.terzera@unimib.it

RETHINKING THE ORGANIZATION OF PUBLIC ADMINISTRATION THROUGH THE ENHANCEMENT OF HUMAN RESOURCES. THE ISTAT CASE[1]

Patrizio Di Nicola, Patrizia Grossi, Alessandra Preti

Introduction

In recent years the Italian public sector has experienced new organizational structure aimed at improving the efficiency and effectiveness of its action. The goal of such transformation was to achieve a lean and efficient structure, capable of dealing with the challenges coming from the outside. Sometimes these changes have focused on the involvement and adaptability of human resources generating a virtuous circuit of synergies and relationships that make osmotic the flow of information, expertise and know-how. This is possible through innovative mechanisms of Government that ensure a large aggregation of human resources, facilitating the sharing of objectives and instruments, laying the foundations for a real evolutionary change in terms of: human resources orientation, through the development of skills, knowledge, roles and individual and collective responsibilities; simplification of the administrative processes through an excellent organizational communication.

The search for the maximization of efficiency and effectiveness entails different types of intervention involving human resources at all levels. The available knowledge and the potential to produce other knowledge are the most important resources for every organization. Such knowledge lies in the skills, commitment and ideas of people working in a specific workplace. Consequently, the human resources development is crucial to carry out every process of organizational change successfully.

A suitable strategy to involve people in the change processes is to discover all the staff skills and design a training plan based on the skills enhancement and spread. This can bring huge results, especially if combined with an efficient communication strategy that allows employees to share the ongoing processes, and

[1] Although all authors contributed extensively and in closely collaboration to the work presented in this paper, the Introduction could be attributed to all the authors, Section 1 to Di Nicola, Section 2 to Grossi, Section 3 to Preti.

to build up a vision of change "with sense making". People do not just perceive the organizational environment, but contribute through their specific behaviors to build it. For this reason, a method for the analysis of skills and the development of training plans and organizational communication actions is needed.
Involving employees in the reform plan means make them aware of the values at the base of this process: a scant attention towards the staff might cause resistance and bureaucratic attitudes even in people not potentially adverse to changes. The process of organizational change, like every other innovation, always generates fear of the unknown and uncertainty for the new: some people cling to old patterns, others fear that their working condition or their status get worse, others do not feel able to cope with new tasks. The resistance depends not only from intentional and explicit behaviors, but also from customs and established ways of thinking, able to influence behaviors and ways of classifying and interpreting the events.

To carry out the process effectively, management should be able to count on qualified, motivated and committed people, on sure resources, on structured mechanisms of identification and evaluation of results, on adequate communication strategies and on information systems with high added value: it will be essential to develop shared competences. Thus the management needs to intervene on the cultures and the patterns of behavior of people, identify and support the reasons of the change and clearly communicate them to everyone. Management should promote actions aimed at establishing favorable conditions to carry out and support change and improvement processes.

1. Knowledge of the skills as a strategic key for organizational change

This section deals with the topic of organizational change and the role of the skills of the actors in the organization. The attention will be focused on those approaches called *soft* that "favors cultural, symbolic and reflective aspects». According to this reading path, the organization is seen as a "continuous process of reality definition in which the individuals are actively engaged[...], this process is embodied in mediate communicative exchanges of symbols, of stratified knowledge, reference models, patterns of perception and interpretation of reality and of codes and standards, also not written». There is an urgent necessity to interpret the symbolic and cultural aspects of the organization because of the progressive change of working content and modalities of jobs. In many organizational contexts of the *knowledge society*, the traditional taylorism/weberian division between execution and command on the one hand is thrown into crisis, on the other hand is supported by a redefinition of tasks that become «deep» and «wide». The formal and hierarchical control is not functional for the nowadays productive

requirements because the workers are asked to extend his own field of action to meet the society needs. To have visions of the problems as widest as possible, the individual should not perform tasks, but he should fill roles he acts possessing only «minimum critical specifications» working «in view of functional outcome, in his relations with others, within a given organizational technical context». The peculiarity of these organizational forms makes it difficult to separate power and participation, as the increased complexity and interdependence between the parties make widespread the decision-making capacity. What seems to bind and hold together the organization, therefore, are not only the formal rules and mechanisms of integration on hierarchical basis, but also the sharing of problem-solving mode, way of doing things, and the unspoken knowledge.

1.1 Ways of organizational change

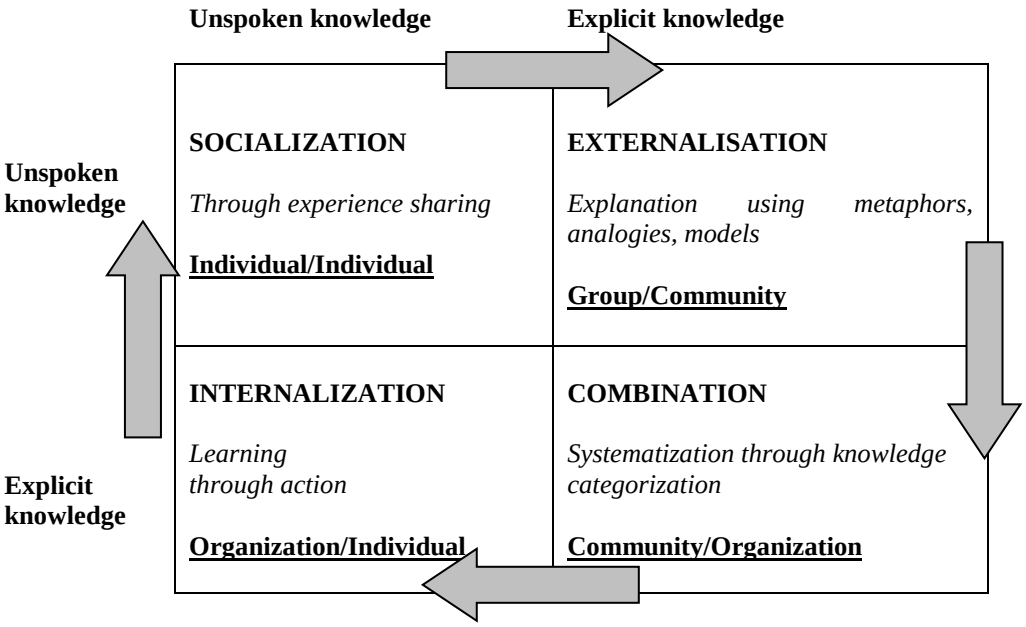
The theme of organizational change is wide and difficult to define as well as the theme of change management. We can try to reduce complexity through two possible readings: the first looks at change as a goal to achieve (*goal view*) through specific and identifiable moments; the second, rejecting the parallelism between change and planning, offers an approach (*process view*) where change management is near to the topic of organizational learning. In the first case the change is considered as a projectable event introduced in the organization by an 'agent of change'. According to this perspective, the change: is an event that can be controlled through top-down mechanisms, that is from the agent of change to the rest of the organization; is an objective that can be attained through a set of scheduled events to be activated within an organization designed in a rational and mechanical way, rather than in a systematic way.
The second reading offers a conception of change related to that of a "learning organization" considered as a set of processes and not as a rational and merely exploitable structure. This interpretation, favoring the "act of organize" *rather than* the *organization*, makes central individuals and interactions, decisions and conflicts. An organization learns when it *changes* its repertoire of skills, patterns of action, established strategies and procedures shared over time.
The cooperation and the listening between the different subjects are essential, in particular, between who plans the change and who has to achieve it and/or undergo it, reflecting on the level of involvement and participation during the process.

1.2 Unspoken knowledge development

As organizations cannot avoid the existence of *relational systems*, especially when they are the result of spontaneous aggregations; all there is left to do, is face

the challenge by creating a favorable environment for their development, and integrating them within an organizing logic, in order not to disperse the individual relational and cognitive heritage. For this purpose the model of «knowledge spiral» by Ikujiro Nonaka and Hirotaka Takeuchi represents an interpretive tool to explain both the functional complexity of relational phenomena and their integration, in the overall organizing logic at *macro* level, and in the processes of generation and exploitation of knowledge *at micro level*. The spiral (Figure 1) shows clearly that this model consists of four key stages (socialization, externalization, combination, internalization) strongly interconnected, in which each stage represents the input for the next stage in a circular perspective.

Figure 1: - *The knowledge spiral by Nonaka and Takeuchi*



During the *socialization* phase, there is the switching from a unspoken knowledge to another unspoken knowledge. It is based on processes of experiences sharing and mental models creation and common technical skills. The experience is therefore the main tool to capture unspoken knowledge. The *externalization* allows the transformation of a unspoken knowledge into explicit knowledge. Then, the unspoken knowledge is made more explicit through the use of metaphors, analogies, concepts, hypotheses or models. The *combination* is the stage where there is the switch from explicit knowledge to another explicit knowledge. It is aimed to systematization of concepts in a frame of explicit knowledge. The main

aspect of this phase is the possibility to attain new forms of knowledge by sorting, adding, and combining explicit knowledge. Finally, the *internalization*, determines the transfer of explicit knowledge to a new unspoken knowledge. That process, which the authors call closely related to that of learning through action, allows the sharing of a specific mental model by the majority of members of an organization, transforming unspoken knowledge into an integral part of the broader organizational intellectual assets.

The organization is called upon to play a primary role in managing this intricate process of generation, accumulation and use of knowledge, starting from the contributions of individuals to involve the practices and experiences of the largest community of people that compose it. In other words, the organization has the task of providing the appropriate context to facilitate group activities and the creation and the growth of knowledge at the individual level.

This process of development can be facilitated by the intensive use of information and communication technologies. The technologies available to contemporary organizations allow to enhance the ability to link and compare the actors, to increase the potential information, to stimulate the birth – spontaneous or deliberate – of forms of social aggregation.

The new organization requires autonomy, work team aptitude, communication and *problem solving* capacity. Employees are fundamental and their skills are the element that helps to build the new organization.

1.3 Skill mapping and training planning

An organization that bases its value on the ability to generate and continuously update the skills of the people must learn more about the skills of its human resources. Only in this way you can encourage individual and collective learning and create a virtuous circle of exploitation of people.

Skills and professionalism are the main tools to contextualize and integrate each expertise within a particular environment.

Skills represent a dynamic factor that can be understood only in the light of the interaction and socialization processes developed by the individual, through which his models of knowledge and action are continuously integrated and adapted to the characteristics of the social context. In this perspective, it is important to collect by a survey, on a periodic basis, the perception of people with regard to the competences owned both by themselves and by their colleagues.

This allows the organization to understand: which people are considered a cornerstone for specific subjects/skills possessed; what skills – both socio-relational and specialized – the person is believed to possess and which of these was considered a landmark by colleagues.

Using the information collected and mapped it becomes possible to create a cognitive tool, a sort of "Knowledge Yellow Pages" for use by the management.

2. The case of Istat for the enhancement of human resources

In order to overcome the organizational emergency related to the 2011' Population Census, Istat employed in 2010-2011 270 new skilled resources with fixed-term contract. Of these, 168 have been assigned to the Department of *Censuses and administrative and statistical records*. To facilitate the inclusion of the new employees, the Department launched three "welcome" projects: Tutoring project for inclusion of new resources; Integrated System of Communication and Learning; Resource Development Project.
These three projects form the base of the Management System for HR development.

2.1. Tutoring project for new resources inclusion

The main purpose of the tutoring project (named *TI/Cens*) was to ensure that the skills of new professionals matched the production requirements of the Censuses as planned by the Organization.
Therefore, before the arrival of new persons, it was decided not to allocate them immediately to the destination Services and Operational units. There was a phase of temporary pre-assignment to facilitate the adaption to the job and "mutual understanding" with managers and colleagues. At the end of this period the new employees were assigned to a specific office and task.
In detail, the results and the expected benefits of the project were: employment of recruits according to the needs of production and the expectations and skills of the new resource; reduction of indirect costs, with particular regard to the time of the staff in charge of tutoring; enhancement and development of internal resources who made the "educators"; development of innovative and replicable insertion ways, through self-training modules available even for other users such as the internal staff of the Department.

In *TI/Cens* the tutor played a key role and was a point of reference and orientation for new recruits, reviewing their skills, aptitudes and expectations in order to find the best organizational placement for their development and mutual satisfaction. The tutors were even tasked of supporting the new resources during their learning through: the selection and post-processing of key documents to be known by the new people; the holding of periodic meetings with the learners to

learn their problems, to give explanations, to reflect and to support self-evaluation and goal setting.

Thanks to this project it was possible to evaluate and supervise, at all levels of responsibility (Director, Service Managers and Operational Units managers), the match between "supply and demand" of professional skills: on the one hand, the need of new resources in the different production processes and in the different methodological, technical and organizational areas in the Department of Census; on the other hand, it was possible to understand the skills, the abilities and the knowledge of the new people in order to strengthen their professionalism.

The logistical and managerial aspects had considerable weight in the success of the project. Logistic was taken into account in order to allocate all the engaged resources in neighboring rooms, paying attention to individual and social variables, relating to the characteristics of each individual (qualification, attitudes, motivation) and to the technological and instrumental aspects, to be sure that each new recruit had the necessary equipment.

TI/Cens guaranteed: the best location and an appropriate mentorship to the new recruit; the matching of new professional resources with the productive requirements of the Censuses; a better knowledge for the new recruit of the organizational dynamics and of values of the Institute.

The results and the benefits of the project were: fast job placement in line with the needs of production; job placement closer to the expectations and the real skills of new recruit; empowerment and development of internal resources; development of an innovative and "replicable" insertion method; development of self-training modules available for other users and for different purposes.

The *lessons learned* of the project were: never without involvement: the sponsorship of leadership, the involvement of top management and the support of other Directorates (in particular General Directorate of personnel, Logistics, Information technology etc.) were crucial; never without appropriate documentation: the collection of relevant documentation on job to be done and on the organizational structure of Istat proved to be crucial for the new recruit; never without tutors: they proved to be able to adapt themselves to the needs and to increase their ability to address the new resources in a process of mutual learning-on-the-job.

The new recruits gave positive feedbacks and appreciation for the tutoring experience; in particular they enjoyed the orientation interview, the initial meeting with the Department's management, the relationship with their tutor and finally the fast resolution of logistical and organizational difficulties.

2.2. Integrated System of Communication and Learning

The *SIP/Cens* project was aimed at rethinking the internal process of vocational training. It originated by a self-evaluation questionnaire delivered at the end of the previous project.
The operational phase of the Census highlighted the need to integrate and update some specific skills, necessary to lead some critical stages of the activity. The need was to make available assistance, education and information, to ensure the best support to the staff in all phases of the Census process until the publication of all results (due to Eurostat by April 2014).
The main objectives of the *SIP/Cens* project were:
1. to make available innovative training courses for a better execution of Census operations;
2. to secure the results and monitor them for skills development objectives;
3. to establish a group of internal resources able to communicate to staff the major operational tools, respecting compatibility of skills and workloads of people;
4. to create the conditions (technical, logistical and cultural) to supplement traditional education with particular attention to individual organization of didactic material and interpersonal communication;
5. to locate professional resources necessary to implement the program and to establish a group of experts constantly available for the technical update of statistical, methodological and informative tools ;
6. to assess the level of skills acquired and its translation in improvement of professional performance.
The realization of the project requested the creation of a "laboratory" aimed to train on the job the new resources. It should be noted that the officers involved as teachers did not receive any monetary compensation for their work, only received a "thank you" letter from the Director of the Department; the participants were awarded with a "certificate of attendance" after the confirmation (by the manager of the unit), that the new skills were successfully adopted in the daily job.

2.3 Resource Development Project

This project was born in continuity with *Ti/Cens*. Two years later the initial insertion of the new resources, there was the necessity to enhance the skills of people involved in the Census activity. The new project, named "*ValiTi/Cens*", had two implications: first, to "*assign value"* to the resources, and second to "*extract value*" from them. In the first meaning, it was necessary to define the parameters for a better placement of resources, according to the guidelines by Triennial Strategic Plan (PST) 2013-2015. In the second meaning, the project was necessary to respond to the

need of individual centrality, in a sort of continuous enhancement of individual skills and abilities.

The project was carried out by taking advantage of the high skills of some internal resources belonging to the Department, and making a considerable investment in terms of survey design, collection of information, analysis of data and dissemination of results. Studies were made on possible measures to assess and improve the organizational well-being; therefore an online questionnaire was delivered to 144 temporary resources of the Department; the questionnaire had 39 questions intended to analyze skills and evaluate the possibilities of exploitation of these, measure the level of work satisfaction, the expectations and the possible desire of gaining a different allocation (according to the organizational needs of the Department).

The ValiTi/Cens Project was launched in a phase of reorganization of the Department with the need to redeploy resources to new activities; it was of strategic importance for the people, because they were stimulated to self-evaluate their work according the standards of personal growth and personal aspirations.

The employees were informed of institutional activities planned for the coming three years in order to choose consciously the "right" place of work after a self-examination.

The main methodological innovation of the project consisted in having integrated the collection of information through a web questionnaire, in cases where the respondent had expressed the wish to change job, with a personal *face-to-face* interview. The combination of indirect and direct instruments of investigation facilitated the complete collection of data of the respondents, and it made possible to give a voice to the interests and professional expectations that could have been better satisfied in other operating units, other services or departments of the Institute.

Among the 144 resources with fixed-term contract involved in the survey:

- 75 (52%) expressed the will to continue the activities in which they were engaged;
- 64 (44.5%) were also interviewed directly, because they had reported new areas of interest, suggestions and insights to enhance skills and meet their aspirations;
- 5 (3.5% of the population) did not respond to the questionnaire.

The project placed perfectly in the enhancement of professionalism in conjunction with an internal reorganization, and the results helped the management to develop helpful hints to increase the sense of belonging of the employees, to improve the level of integration and to make everyone, depending on their skills and inclinations, protagonist of the production process.

At the end of the project, 22 persons were moved to other activities closer to their professional skills and expectations.

The future challenge could be to spread in ISTAT a bottom up approach for the development of the people, aimed at inclusion, so that the persons can be

considered in their multidimensional identity and not only for the job they were able to do at the time of hiring.

3 Conclusions

The organizational change in a research public body can be set and implemented in a rational way, using a model based on the analysis of competences intended to define new recruitment procedures, resource allocation, mobility, guidance and career development. To manage a change which intends to foster innovation and improvement in terms of timely and accurate integrated statistics to be supplied to international, national and local policy-makers, Istat needs to develop a vision of itself based on: the ability to capitalize the value of people for its best use; the ability of the staff to respond to the growing demand for professionalism.

In a context where only flexibility and adaptation processes allow complex organizations to face the uncertainties and changes, it becomes very important and strategic to recognize and enhance the professional contribution of the individual, the intellectual capital represented by knowledge and expertise of the people and by the individual capacity to gain, manage and use this treasure. In order to act promptly to adapt the staff to changes of the goals of the Institute, it is essential a clear vision of future needs and therefore the best knowledge that must be available in short and medium term. This requires, according to Alleva 2006, the use of ″competency model which should represent the orbital center around which all stages of the life cycle, from recruitment to termination of the employment relationship, have to rotate ". The process of organizational change cannot exist without a parallel process of innovation and improvement of personnel policies: the transition from the *mapping of expertise* for the realization of specific purpose (assignment of personnel to offices and projects, identification of training needs, staff assessment etc.) to a real integrated system of human resource management based on competencies and skills.

An important aspect within an organization, in fact, lies in the ability to mobilize the necessary resources (competences as knowledge, know-how, knowing how to learn, how to act, desire to act etc.), not only possessing them. You can have all the skills/knowledge for an excellent performance, but if you don't know how to activate it fully, for example because you have personnel "not enough motivated", probably the organization will not be able to achieve the target objectives. On the other hand, the organization could get good results by enabling the decision-making capacity, even when your staff is not in possession of all the requirements in terms of experience and knowledge.

References

ALLEVA G. 2006. *Investing in human resources for a quality official statistics*. Proceedings of the 8TH National Conference of statistics, Rome.

BARTUNEK J.M., KRIM R.M, NECOCHEA R., HUMPHRIES M. 1999. Sensemaking, sensegiving, and leadership in strategic organizational development, *Advances in Qualitative Organizational Research*, Vol. 2, pp. 37-71.

G. BONAZZI, 2002. *Come studiare le organizzazioni*, Bologna: Il Mulino.

CROZIER M. 1990. *The listening company. The management in the post-industrial world*. Milan: Il Sole 24Ore.

LEWIN K. 1939. Field Theory and Experiment in Social Psychology: Concepts and Methods, *American Journal of Sociology*, Vol. 44, No. 6, pp. 868-896.

NONAKA I., TAKEUCHI H. 1997. *The knowledge-creating company*. Milan: Guerini e Associati.

PICCARDO C., BENOZZO A. 1996. *Etnografia organizzativa*, Milano:Raffaello Cortina Editore.

POLANYI M. 1966. *The Unspoken Dimension*. First published Doubleday & Co.

SUMMARY

RETHINKING THE ORGANIZATION OF PUBLIC ADMINISTRATION THROUGH THE ENHANCEMENT OF HUMAN RESOURCES. THE ISTAT CASE

In the current economic situation public bodies have to seek the maximization of the efficiency and effectiveness of their action through innovative mechanisms of Government to ensure a better use of human resources. New organizational paradigms that aim at the sharing of languages, objectives and instruments, lay the foundations for a real evolutionary change oriented to the enhancement of human resources, with their own skills, knowledge, roles and individual and collective responsibility. Everything is made indispensable by the combined effect of two disruptive forces: 1) the technological innovations (ICT and digitalization) that make work more efficient, but also more demanding from an intellectual point of view; 2) the ageing of human resources resulting from the reforms of pension systems which require employees to stay longer in employment: this obviously entails a continuing educational strategy aimed at "maintaining" and increasing the skills of members of the organization.

ISTAT has recently undertaken a broad process of organizational change to meet the growing demand for statistical information from Italian society, taking into account that in the Institute the main working resource derives from the knowledge available and from the potential to produce more. Consequently, the enhancement of human resources becomes a

decisive tool to carry out any modernization process successfully. It is essential to analyze continuously the staff skills in order to develop a suitable strategy to train resources. The challenge is to bring the employees not only to share ongoing processes, but also to develop a vision of change "with sense making". People, in fact, do not just perceive the organizational environment, but contribute through their specific behaviors to build and influence it. To this end it will be proposed a method for the analysis of competences, their mapping and the development of training plans that allow people to become active actors of change. Using this instrument, the Institute will be able to build an integrated human resources management system, *knowledge based*, to support the management and increase the productivity of staff.

Patrizio DI NICOLA, ISTAT, dinicola@istat.it
Patrizia GROSSI, ISTAT, grossi@istat.it
Alessandra PRETI, ISTAT, preti@istat.it

L'INTEGRAZIONE TRA I DATI DELL'INDAGINE DI COPERTURA DEL CENSIMENTO 2011 E GLI ALTRI ARCHIVI AMMINISTRATIVI CENTRALIZZATI. L'ANALISI SUGLI INDIVIDUI PIÙ DIFFICILI DA RILEVARE

Nicoletta Cibella, Gerardo Gallo, Anna Pezone, Tiziana Tuoto

1. Introduzione

Il censimento permanente per il conteggio della popolazione sta per segnare il passaggio definitivo dal censimento tradizionale, di tipo "porta a porta", ad una rilevazione di tipo "register-oriented" basata, oltre che sulle anagrafi comunali, su una serie di archivi satellite di fonte non anagrafica. In occasione del Censimento 2011 per la prima volta le informazioni individuali delle anagrafi e di altri archivi amministrativi centralizzati hanno assunto un ruolo importante, consentendo un vantaggio notevole nei tempi di elaborazione dei dati, ma hanno anche mostrato criticità rispetto alla esatta collocazione geografica degli individui. In particolare, il "recupero" tramite rilevatori degli individui che, dai segnali dell'integrazione di più archivi non anagrafici (Gallo, Paluzzi e Benassi, 2014), potevano essere dimoranti abitualmente nei comuni italiani ha conseguito un numero di censiti residenti al di sotto delle aspettative. Di fatto, il concetto della popolazione dimorante abitualmente è ben definito, oltre che dal piano di censimento del 2011, anche dalle norme internazionali secondo cui "*the usual resident population is one of the cases where there is a one-to-one relation between memberhip and geographical allocation*" (Unece, 2013). Tuttavia, la perfetta corrispondenza tra gli archivi amministrativi e la situazione di fatto rappresenta di per sé una criticità anche nei paesi che già da decenni utilizzano i registri di popolazione e altre fonti amministrative per il conteggio della popolazione (Wallgren A., Walgreen B., 2011). La necessità di disporre di conteggi puntuali rappresenta oggi un'esigenza di tutti i paesi dove gli utenti finali dei dati di popolazione si chiedono ormai da tempo: qual è l'esatto ammontare della popolazione residente?

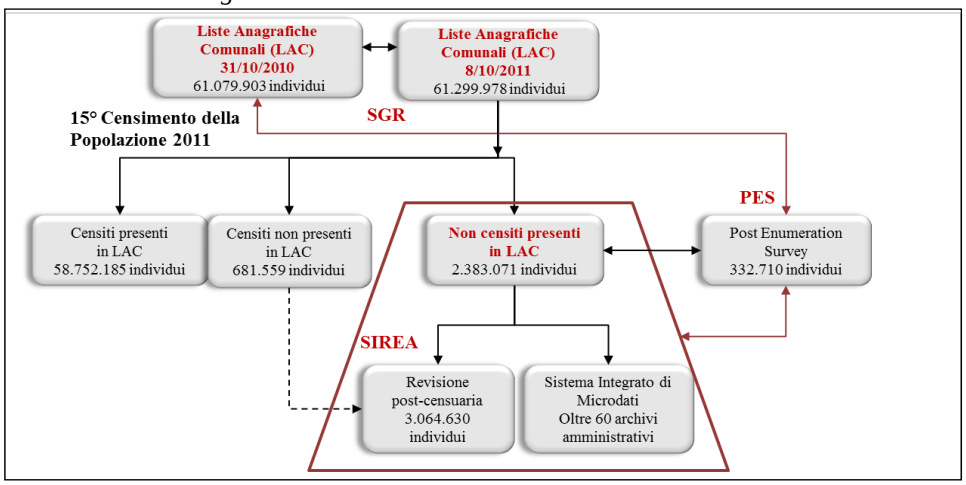
Un primo obiettivo di questo lavoro è quello di valutare, attraverso l'impiego di modelli di regressione logistica, se esistono particolari tipologie di individui, famiglie o di specifiche sottopopolazioni che, al momento della rilevazione censuaria 2011 e dell'indagine di copertura, sono state rilevate in un "luogo" diverso rispetto a quello di iscrizione anagrafica. Un secondo obiettivo consiste nell'individuare sia la presenza di eventuali variabili latenti che possano giustificare gli scarti tra popolazione censita e popolazione anagrafica, sia le aree

territoriali che, presentando valori più elevati di incoerenza tra le due fonti, possono essere definite come "aree di malessere" per il sistema di conteggio della popolazione. Per realizzare questi obiettivi si farà ampio ricorso al contributo conoscitivo degli archivi amministrativi attualmente disponibili in Istat.

2. Il contesto informativo e l'integrazione tra le fonti

L'elevato livello d'informatizzazione dell'ultima rilevazione censuaria e il ricorso a dati di fonte amministrativa centrale e locale, ha permesso la raccolta e l'analisi di micro dati ad un livello di dettaglio fino ad oggi impensabile. La figura 1 mostra sinteticamente il contesto informativo delle fonti di dati utilizzate in questo contributo.

Figura 1 – *Schema sintetico del contesto informativo, delle fonti di dati utilizzate e della loro integrazione*



Nella fase preparatoria del censimento sono stati acquisiti i dati delle anagrafi comunali, riferiti al 1° gennaio 2011, per la stampa e spedizione dei questionari alle famiglie. I dati raccolti dalla rilevazione delle Liste Anagrafiche Comunali (LAC) sono stati caricati nel Sistema di Gestione della Rilevazione censuaria (SGR) che ha rappresentato la piattaforma informatica per monitorare e completare le operazioni censuarie. Questo sistema ha consentito anche l'aggiornamento delle informazioni anagrafiche alla data del censimento e l'esecuzione via *web* dell'operazione di confronto censimento-anagrafe. A conclusione di questa operazione i comuni hanno potuto disporre, oltre che del dato della popolazione residente censita, di informazioni dettagliate su tre aggregati: 1) gli individui

censiti e presenti nella LAC; 2) gli individui censiti ma non presenti in LAC (circa 681 mila individui); 3) gli individui non censiti ma presenti in LAC (circa 2,380 milioni). Gli ultimi due elenchi nominativi, associati a codici identificativi univoci, sono stati trattati da ciascun comune per la revisione anagrafica post-censuaria attraverso il Sistema di Revisione delle Anagrafi (SIREA), anch'esso disponibile su *web*. Ciò ha garantito che l'attività di revisione delle anagrafi avesse regole uniformi e strumenti condivisi su tutto il territorio nazionale.

Le risultanze di SIREA, per il solo insieme di individui non censiti e presenti in LAC, sono state poi integrate dall'Istat con informazioni di fonte amministrativa provenienti dal Sistema Integrato di Microdati (SIM). Il SIM è una struttura informativa realizzata mediante l'integrazione concettuale e fisica dei microdati acquisiti da fonti amministrative (non solo anagrafiche). Esso è organizzato in modo da supportare i processi di produzione statistica, distinguendo *Individui* e *Unità economiche*, consente l'associazione dei *luoghi*, in termini di indirizzi e codici, e definisce le relazioni tra le unità di base del sistema generando ulteriori sottosistemi di integrazione. Per eseguire le analisi sugli individui non trovati al censimento sono state selezionate alcune partizioni specifiche del SIM. Con riferimento al 2011, ci si è avvalsi di informazioni individuali relative alle seguenti sottopopolazioni: i deceduti presenti dell'Anagrafe Tributaria; i cittadini italiani residenti all'estero iscritti all'Anagrafe Consolare al 2011; gli individui presenti nel Casellario dei pensionati; gli iscritti nell'anagrafe degli studenti; gli individui che partecipano al mercato del lavoro in Italia nel 2011 (contributi previdenziali INPS per i lavoratori dipendenti, autonomi, subordinati, para-subordinati e interinali, nonché i lavoratori in agricoltura e nel settore domestico, comprese colf e badanti) e, infine, i trasferimenti di residenza (Modelli Istat APR/4 degli anni dal 2009 al 2011).

Un'ulteriore fonte di dati utilizzata in questo contributo è rappresentata dai microdati individuali della indagine campionaria di copertura del censimento 2011, la Post-Enumeration Survey (PES). Questa indagine, indipendente dal censimento e di tipo areale, è stata realizzata sulla base di un piano di campionamento a due stadi: i comuni (oltre 250, sono le unità di primo stadio) e le sezioni di censimento, (circa 2.500, sono le unità di secondo stadio). Nelle sezioni di censimento campionate i rilevatori, selezionati tra quelli che si sono distinti per precisione ed accuratezza al censimento, hanno intervistato, *porta a porta* e senza l'ausilio di alcuna lista, tutti gli individui eleggibili ad avere la dimora abituale nei comuni campione, per un totale di circa 330 mila unità. Pertanto, la PES del Censimento 2011 risulta, per dimensione, la più grande indagine campionaria realizzata dall'Istat nella storia dei censimenti ed è stata condotta con un questionario che ha rilevato, in maniera estremamente sintetica ma molto accurata, le principali informazioni individuali rilevate al Censimento. Una volta rientrati i dati di

indagine si è proceduto all'abbinamento con i risultati definitivi del Censimento utilizzando tecniche di *record linkage* che garantissero l'assenza di errori così da poter calcolare il tasso di sotto-copertura censuaria attraverso il tradizionale stimatore di tipo *dual-system* (Wolter, 1986).

Il processo di abbinamento[1] dei risultati dell'indagine PES con i dati censuari mirava a determinare il numero di individui rilevati in entrambe le occasioni e, per differenza, quelli sfuggiti ad una delle due rilevazioni con attenzione a due classi di individui: quelli rilevati in entrambe le rilevazioni allo stesso indirizzo e quelli che alla PES sono stati rilevati ad un indirizzo diverso da quello censuario; in questo modo si è potuta stimare la sovra copertura del Censimento. Un ulteriore insieme interessante era costituito dagli individui rilevati con la PES ma non trovati al censimento; per questo contingente, per il quale è stata verificata anche la presenza nelle LAC pre-censuarie[2], la PES costituisce un'informazione terza che arricchisce il confronto tra il Censimento e le fonti anagrafiche.

Al termine delle operazioni di *record linkage*, a tutti gli individui PES è stato associato lo stato di abbinamento con il censimento e con le LAC pre-censuarie. Successivamente a questo abbinamento, i dati della PES sono stati integrati con gli esiti del confronto censimento-anagrafe (sulla base della LAC all'8.10.2011), con i risultati di SIREA e con le informazioni del SIM[3]. Il sottoinsieme di interesse su cui si articola l'analisi di questo lavoro è costituito dai 4.360 individui[4] che, rilevati alla PES, sono stati abbinati con i 2,380 milioni di individui non trovati al censimento; di questo aggregato, su cui la comunità scientifica si è posta diversi interrogativi (Livi Bacci, 2013), si effettuano opportuni approfondimenti sia sugli esiti derivanti dal processo di revisione di SIREA sia sui segnali rinvenuti dall'integrazione con gli archivi del SIM.

[1] Per assicurare la massima correttezza degli abbinamenti è stata adottata una procedura di *record linkage* strutturata in diverse fasi con modelli di tipo probabilistico, iterati considerando differenti metriche per il confronto delle variabili comuni nei vari passi della procedura (AA.VV. 2014). La complessa articolazione della strategia di *linkage* è stata supportata e assistita dalla disponibilità di strumenti informatici generalizzati, che implementano in maniera ottimale le sofisticate metodologie richieste: RELAIS3.0 (AA.VV., 2011) che mette a disposizione le metodologie più consolidate per il *record linkage* probabilistico ma anche *facilities* per innovativi metodi di riduzione dello spazio di ricerca.

[2] Per LAC pre-censuaria si intende la LAC che contiene i dati comunali riferiti al 31/12/2010.

[3] Proprio dal confronto tra le risultanze censuarie e le LAC aggiornate all'8 ottobre 2011 è stato possibile ottenere le liste di SIREA.

[4] In questo lavoro si fa riferimento ai dati PES non pesati.

2. Gli individui più difficili da rilevare: primi risultati dell'analisi macro e micro

Uno degli aspetti più critici dell'ultimo censimento della popolazione è rappresentato dall'elevato numero di persone iscritte nell'anagrafe comunale e non trovate al censimento (2,4 milioni). Questo ammontare, insieme ai censiti non presenti in anagrafe (681 mila), determina di fatto uno scarto significativo tra la popolazione censita e quella anagrafica (circa 1,8 milioni di unità, pari al 3% in termini relativi). Anche il confronto tra i risultati della PES e i recuperi di popolazione derivanti dagli esiti delle rettifiche anagrafiche post-censuarie di SIREA presentano differenze non del tutto irrilevanti. Pertanto, ci si pone l'obiettivo di individuare, attraverso l'analisi micro e macro, particolari sottogruppi di popolazione a rischio di sfuggire alle rilevazioni statistiche e la presenza di variabili latenti che sottostanno agli errori di conteggio.

4.1. L'approccio micro: metodi di analisi e risultati

L'osservazione a livello micro degli individui rilevati alla PES, ma sfuggiti al censimento, può fornire importanti indicazioni sulle principali caratteristiche di questo aggregato. Le analisi sono state effettuate sui 4.360 individui della PES abbinati con i 2,380 milioni non trovati al censimento. Nella tabella 1 sono riportate le informazioni di questo sottoinsieme di individui secondo gli esiti di SIREA e i segnali delle fonti amministrative non anagrafiche del SIM.

Tabella 1 - *Esiti del confronto tra i risultati della PES, la revisione censuaria (SIREA) e i segnali di presenza nel SIM al 31.12.2011. Dati PES non pesati. Valori assoluti e percentuali.*

Esiti integrazione PES-SIREA-SIM	Persone non censite rilevate alla PES	
	di cui: presenti nella LAC all'8.10.2011	di cui: presenti nella LAC al 31.12.2010
Record abbinati PES-NON CENSITI (val.ass.)	2.973	1.387
Esiti in SIREA (val. perc.):		
Cancellati per irreperibilità	*8,8*	*0,9*
Cancellati per cambio di residenza	*27,0*	*51,0*
Non censiti che confermano la dimora di LAC	*62,1*	*13,3*
Errori di lista	*2,1*	*34,8*
Totale	*100,0*	*100,0*
Esiti nel SIM:		
Almeno un segnale nel SIM (val.ass.)	2.146	1.366
di cui: % persone che lavorano nel 2011	*52,0*	*45,8*
Nessun segnale nel SIM (val. ass.)	827	21

Fonte: Elaborazione su dati ISTAT

In particolare, è possibile osservare che 2.973 individui non censiti sono stati rilevati dalla PES (il 68% dell'insieme considerato) ed erano presenti nella LAC aggiornata all'8 ottobre 2011[5], mentre le altre 1.387 unità, anch'esse rilevate dalla PES, erano già presenti nelle LAC al 31.12.2010. Pertanto, se si considerano gli esiti di SIREA nel loro complesso, è possibile notare che circa il 78% delle unità rilevate alla PES non sono state censite per via dei trasferimenti di residenza che si sono verificati tra il 31 dicembre 2010 e la data di riferimento del censimento. Una quota molto significativa è rappresentata dalle persone non censite che, rilevate dalla PES, confermano la dimora abituale nel comune di iscrizione anagrafica (più del 62% dei casi). Inoltre, se si escludono gli errori di lista, che hanno un impatto minore nei conteggi di popolazione, di un certo rilievo appare anche la quota di persone rilevate dalla PES che sono state successivamente cancellate per irreperibilità dal processo di revisione anagrafica di SIREA (poco meno del 10%).

L'integrazione con le fonti non anagrafiche del SIM rileva che per oltre il 72% dei non trovati al censimento è stato rinvenuto almeno un segnale di presenza nelle fonti non anagrafiche (Tabella 1, prima colonna). In particolare, per il 52% dei casi si tratta di segnali di partecipazione al mercato del lavoro per tutto il 2011. Nel complesso, l'aggregato di popolazione più difficile da rilevare, perché senza segnali nel SIM, è composto da 848 individui (circa il 20%).

Inoltre, si è proceduto all'applicazione, sul contingente dei 4.360 individui, di modelli di regressione logistica utilizzando come informazioni ausiliarie quelle derivanti dagli archivi amministrativi. In particolare, sono stati impiegati diversi modelli lineari generalizzati di regressione logistica (Hosmer et al, 2013), a effetti semplici e misti, usando la funzione *glm*, del pacchetto R stats (R, 2014) dove la variabile dipendente, Y, è pari ad 1 nel caso in cui l'individuo, sfuggito al censimento, è presente solo in PES e a 0 se l'individuo è presente sia alla PES che nella LAC. Le covariate usate nei modelli sono: il sesso, l'età (cinque classi) e la cittadinanza (italiana, straniera), il numero di maschi, il totale di individui della famiglia, la tipologia familiare (single, coppia, coppia con figli, altro), il comune e la provincia di iscrizione anagrafica dell'individuo, le variabili dicotomiche relative ai segnali di presenza nel SIM (pari ad 1 se presente nel SIM) e quelle composte che individuano almeno un segnale di presenza. L'analisi è stata condotta seguendo una strategia articolata in passi, partendo dal modello base con i soli effetti semplici delle informazioni ausiliarie a disposizione; i segnali di alcune fonti sono risultati essere poco significativi[6]. Successivamente, sono state considerate le interazioni tra le variabili semplici e quelle composte. I diversi modelli sono stati

[5] Si fa riferimento alla LAC su cui è stato poi effettuato il confronto censimento-anagrafe 2011.

[6] Sono poco significative l'informazione "lavora o meno nello stesso comune di residenza", l'essere occupato nei mesi contigui al censimento, la presenza dell'individuo nell'anagrafe consolare o nell'anagrafe degli studenti.

confrontati sulla base dei seguenti indicatori: l'AIC criterio, il BIC, la devianza residua e la corretta classificazione delle unità in base al modello.

La tabella 2 mostra gli *output* del modello che, in base ai criteri scelti, risulta essere quello che meglio si adatta ai dati disponibili; i risultati mostrano una bassa devianza dei residui, con valore mediano pari a 0.36 e massimo pari a 4.15 e una classificazione corretta delle unità per oltre il 90% dei casi.

Tabella 2 – *Modello di regressione logistica. Stime e significatività.*

Coefficienti	Stime	Sign.	Coefficienti	Stime	Sign.
Intercetta	-7,03	***	Numero di maschi in fam:2	0,60	**
Classe di età (15-29)	0,42	*	N. componenti fam.: 4	1,18	***
Cittadinanza straniera	0,87	***	N. componenti fam.: 5,6,7	1,13	***
Coppia	-1,21	***	Con permesso soggiorno	-0,56	*
Coppia con figli	-0,81	**	Con trasf di residenza	-0,46	***
Altro tipologia familiare	-0,99	***	Irreperibili	8,34	***

*Note: Significatività ***=p<0.001;**=p<0.01; *=p<0.05; AIC = 2401.7*

Il segno e il peso relativo dei coefficienti indicano le variabili che assumono maggior importanza rispetto alla difficoltà ad essere rilevati: gli stranieri sono più a rischio di sfuggire rispetto agli italiani, anche se il possesso del permesso di soggiorno aumenta la propensione ad essere rilevato. Le coppie, con o senza figli, hanno un rischio minore di sfuggire rispetto ai single (i coefficienti presentano segno negativo) e all'aumentare del numero di maschi in famiglia aumenta il rischio di sfuggire alla rilevazione. Tra i segnali del SIM, i più rilevanti sono quelli legati ai trasferimenti di residenza; l'esito dei controlli di SIREA rappresenta la variabile che spiega meglio il fenomeno. Ciò lascia supporre che le operazioni di ritorno sul campo, effettuate durante l'attività di revisione, hanno sortito effetti positivi. Infine, le famiglie numerose sembrano più difficili da rilevare: ciò rappresenta un risultato non atteso ma suggerisce una non corretta individuazione dei nuclei familiari, soprattutto in presenza di famiglie coabitanti. Dall'analisi emerge chiaramente che il ritorno sul campo, con indagini specifiche come la PES condotta con personale qualificato, o con ritorni mirati, come i controlli di SIREA, sia lo strumento più efficace per riuscire a limitare la sottocopertura di queste particolari sottopopolazioni. I risultati sono confermati anche da modelli non parametrici, che utilizzano alberi di classificazione attraverso il pacchetto R *Rpart* (R Core Team, 2014).

4.2. L'approccio macro: metodi di analisi e risultati

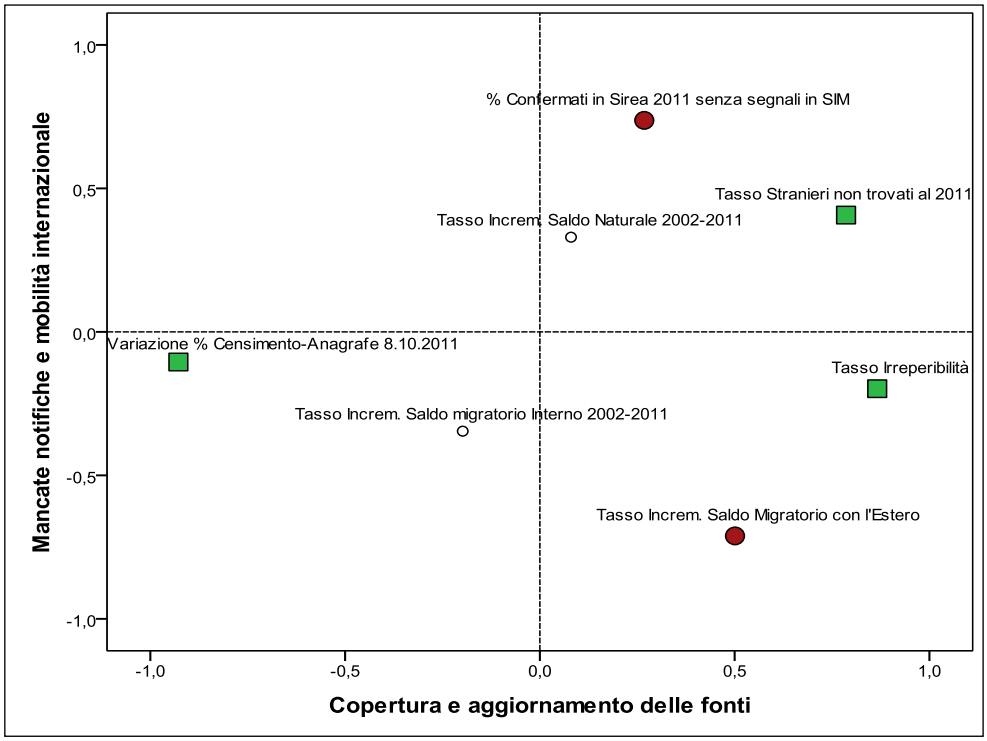
L'analisi a livello macro è stata condotta sulla base delle fonti di dati già descritte e attraverso le informazioni di flusso desunte dai bilanci demografici nel

periodo intercensuario 2002-2011. In particolare, sono stati calcolati i seguenti indicatori: il tasso di incremento medio annuo del saldo naturale (SN), il tasso di incremento medio annuo del saldo migratorio interno (SMI), il tasso di incremento medio annuo del saldo migratorio con l'estero (SME)[7].
Oltre agli indicatori di flusso del periodo intercensuario, sono stati definiti i seguenti indicatori di stock al 9 ottobre 2011: a) il tasso di variazione percentuale tra la popolazione delle LAC e la popolazione censita; b) il tasso di stranieri non trovati al censimento; c) il tasso di irreperibilità al censimento desunto dalla revisione anagrafica di SIREA; d) la percentuale di individui che confermano la dimora abituale in SIREA ma non hanno segnali di presenza in SIM.

Per ridurre l'insieme delle variabili di partenza in un numero inferiore di fattori latenti, è stata utilizzata l'analisi delle componenti principali (ACP). Nel complesso, le informazioni originarie possono essere adeguatamente riassunte con le prime tre componenti che riproducono, complessivamente, una quota di varianza spiegata superiore al 75%. La prima componente, che sintetizza circa il 38% di varianza, è da mettere in relazione con i problemi di copertura sia della rilevazione censuaria (stranieri non trovati al censimento) che dei registri anagrafici (cancellazioni per irreperibilità censuaria o sovra copertura delle anagrafi). Questa componente è fortemente saturata in positivo dagli irreperibili al censimento e dagli stranieri non censiti ma mostra anche una correlazione negativa con gli scarti tra popolazione censita e popolazione anagrafica. La seconda componente è definita dall'incremento medio annuo del Saldo Migratorio con l'estero (con correlazione negativa) e dagli individui non censiti che confermano la dimora abituale in SIREA ma non hanno segnali di presenza nel SIM (con correlazione positiva). La combinazione lineare di questa componente con le variabili osservate lascia supporre che uno dei fattori latenti delle differenze tra anagrafe e censimento sia da mettere in relazione con la componente di popolazione che potrebbe aver lasciato l'Italia nel periodo intercensuario 2002-2011 senza darne comunicazione in anagrafe. Il terzo fattore è definito dal saldo naturale e dal saldo delle migrazioni interne ed esprime la dinamica "endogena" della popolazione residente (nascite e decessi). Nella figura 2 le variabili di partenza sono state proiettate secondo le coordinate delle prime due componenti che possiamo definire, nel caso della prima, *"copertura e aggiornamento delle fonti",* e per la seconda, *"mancate notifiche e mobilità internazionale"*.

[7] Tutti i tassi di incremento sopra elencati sono stati calcolati secondo il criterio del tasso composto "istantaneamente" che si basa sul rapporto tra la variazione della popolazione nell'intervallo 0-t (${}_tP-{}_0P$) e il numero di anni persona vissuti dalla popolazione in tale intervallo: $t\cdot({}_tP-{}_0P)/\ln({}_tP/{}_0P)$. Per ulteriori dettagli sulle modalità di calcolo dei tassi di incremento medio annui dei flussi intercensuari si confronti il contributo di Strozza, Benassi, Ferrara e Gallo (2014).

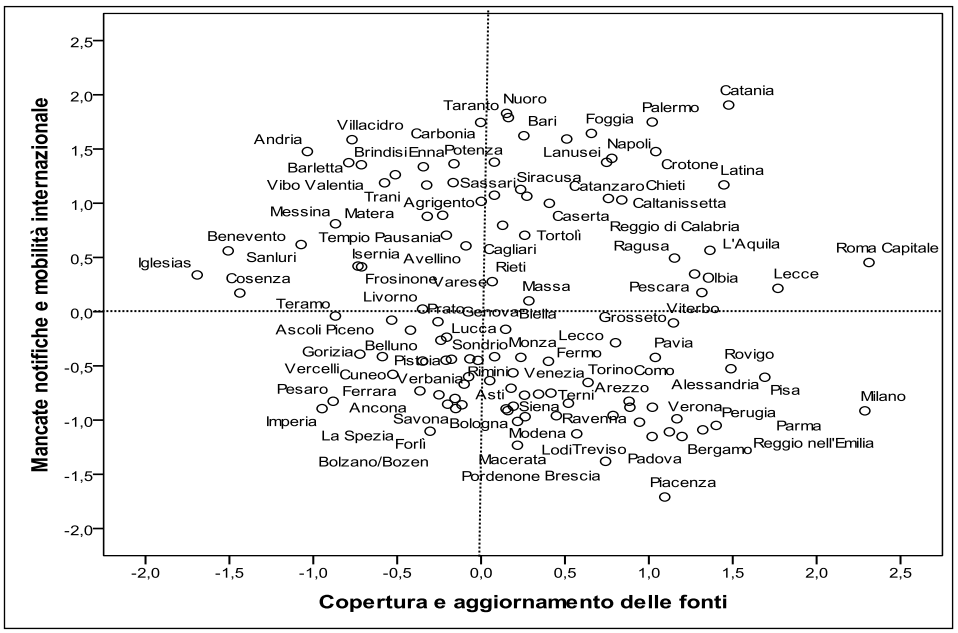
Figura 2 – *Proiezione delle variabili di partenza sui primi due fattori dell'ACP*



È possibile osservare che i comuni del primo quadrante si caratterizzano per una più elevata irreperibilità al censimento con valori sempre più crescenti man mano che ci si allontana dal baricentro. Le unità che si ritrovano nella parte superiore, al di là della bisettrice, si contraddistinguono anche per una più alta incidenza di persone che confermano la dimora abituale in SIREA ma non hanno segnali di presenza nel SIM. In questi comuni le sottonotifiche nella dinamica migratoria con l'estero sono particolarmente significative. Si tratta prevalentemente dei comuni capoluogo di provincia del Mezzogiorno (Catania, Palermo, Napoli) ma anche di diversi comuni dell'Italia centrale. Tra questi, Roma Capitale si colloca nella parte più a destra del primo quadrante a seguito della elevata percentuale di stranieri non trovati al censimento. Inoltre, se ci si sposta verso il semiasse negativo della prima componente, nel secondo quadrante del piano fattoriale, si osservano i comuni caratterizzati anche da differenze significative tra popolazione censita e popolazione anagrafica (ad esempio, Benevento e Cosenza). Così anche i comuni posti nel terzo quadrante, man mano che ci si sposta verso sinistra lungo il primo asse, si registrano variazioni crescenti tra censimento e anagrafe (si osservano,

Ascoli Piceno, Gorizia e Vercelli).Tuttavia, alcuni comuni del terzo quadrante, posti nella parte bassa del secondo asse, sono caratterizzati anche da una scarsa irreperibilità al censimento, da un saldo migratorio con l'estero significativo e, nel complesso, mostrano una variazione tra censimento e anagrafe assai più contenuta rispetto agli altri (si evidenziano, Bolzano, Pordenone, Bologna).

Figura 3 - *Proiezione dei comuni capoluogo di provincia secondo le coordinate della prima e della seconda componente*



Nel quarto quadrante, caratterizzato da una significativa mobilità internazionale e da alti livelli di irreperibilità, si collocano i comuni con una più alta incidenza di cittadini stranieri (i grandi centri urbani del Nord-ovest, del Nord-est e dell'Italia centrale).

5. Verso il censimento permanente

Questo lavoro rappresenta un primo esperimento di integrazione statistica tra i dati di fonte campionaria, quale l'indagine PES 2011, e le fonti amministrative (non solo anagrafiche) attualmente disponibili in Istat. In un'ottica di censimento permanente, fortemente orientato all'ottimizzazione degli archivi amministrativi, i

risultati di questo contributo suggeriscono la necessità di trattare le informazioni amministrative facendo ricorso a criteri di identificazione ben definiti, sia in riferimento ad alcuni gruppi di popolazione (ad esempio, le persone che lavorano), sia per ciò che riguarda i concetti direttamente connessi alle sotto-popolazioni identificate, come l'utilizzo di un luogo diverso da quello di iscrizione anagrafica che, in alcuni casi, potrebbe costituire il vero luogo di dimora abituale degli individui. Spesso è proprio la discrasia tra luogo di iscrizione anagrafica e luogo di dimora abituale che rende assai complicati il confronto tra le fonti di dati disponibili, ad esempio il censimento e l'anagrafe. Dall'uso dei modelli di regressione logistica, applicati a livello micro, è emerso che gli stranieri, i single, le famiglie coabitanti rappresentano i segmenti di popolazione più difficili da rilevare e che i ritorni sul campo, come l'indagine di qualità PES e il processo di revisione anagrafica, sono lo strumento più efficace per contenere la sottocopertura delle rilevazioni censuarie e degli archivi amministrativi. Infine, l'analisi macro ha permesso di individuare, come variabili latenti delle fonti anagrafiche, problemi di "sottonotifica" della popolazione residente, determinata dai processi di mobilità con l'estero del decennio intercensuario soprattutto nei comuni capoluogo del Centro e del Mezzogiorno dell'Italia.

Riferimenti bibliografici

AA.VV, 2014. La misurazione della qualità del 15° Censimento generale della popolazione e delle abitazioni: i risultati dell'indagine di copertura (PES), Seminario del 27 giugno 2014, Roma, http://www.istat.it/it/archivio/126014.

AA.VV, 2011. RELAIS 2011. *User's guide version 2.2*. http://www.istat.it/it/strumenti/metodi-e-software/software/relais.

BOLASCO S. 1999. *Analisi Multidimensionale dei dati*. Roma: Carocci editore.

FELLEGI I.P., SUNTER A.B. 1969. A Theory for Record Linkage, In: *Journal of the American Statistical Association*, vol. 64, pp. 1183-1210.

GALLO G., PALUZZI E., BENASSI F. 2014. The 2011 Italian experience towards supported-Census for measuring migration, . In: *Unece Wok Session on Migration Statistics, Presented paper*. Chişinău Republic of Moldova: 10-12 September.

HOSMER D.W.JR., LEMESHOW S., STURDIVANT R.X. 2013. *Applied Logistic Regression*. *3rd Edition*, ISBN: 978-0-470-58247-3, Wiley.

LIVI BACCI M. 2013. Il censimento del 2011: progressi e interrogativi. In: *Neodemos*, http://www.neodemos.it/ 15/01/2013.

PRESTON S.H., HEUVALINE P., GUILLOT M. 2001. *Demography: Measuring and Modeling Population Processes*. Blackwell publishers.

R CORE TEAM. 2014. R: *A language and environment for statistical computing*., Vienna, Austria, http://www.R-project.org.

STROZZA S., BENASSI F., FERRARA R., GALLO G. 2014. *La recente evoluzione demografica nei maggiori ambiti urbani italiani e il fondamentale ruolo degli stranieri*. In: Franco Angeli anno XLV, n. 109.

UNECE. 2013. *Population Definitions at the 2010 Censuses Round in the Countries of the UNECE Region. Fifteenth Meeting of Group of Experts on Population and Housing Censuses*. Geneva: 30 September – 3 October 2013.

WALLGREN A., WALLGREN B. 2011. To understand the Possibilities of Administrative Data you must change your Statistical Paradigm, In: *Proceedings of the Survey Research Methods Section*, American Statistical Association.

WOLTER K. 1986. Some coverage error models for census data, In: *Journal of the American Statistical Association*, n. 81, pp. 338-346.

SUMMARY

The integration between the 2011 post –enumeration survey and other administrative archives. The analysis of the people hard to count[8]

In this work, a tentative analysis was conducted by linking the individuals registered at the Post Enumeration Survey (PES) with the 2,380 million individuals not counted by the Census but enrolled in the Municipal Population Register (LAC). The study is also enriched by the information gathered from the results of the post-census revision process (SIREA) and the information stored in an Integrated Microdata System (SIM) which gives signals on the presence of the individuals, missing at census count. The integration procedures between the PES individuals and those not counted by the Census was implemented by using the probabilistic record linkage with the RELAIS toolkit a generalized instruments suitable for performing record linkage procedures.

The study has determined, throughout principal component analysis and logistic regression models at a family and individual level, specific groups of individuals or sub-populations with differences in the population count with respect to the one of the municipal register.

Nicoletta CIBELLA, ISTAT, cibella@istat.it
Gerardo GALLO, ISTAT, gegallo@istat.it
Anna PEZONE, ISTAT, pezone@istat.it
Tiziana TUOTO, ISTAT, tuoto@istat.it

VALUTAZIONE DELLA STRATEGIA DI STIMA DELL'INDAGINE SUI CONSUMI ENERGETICI DELLE FAMIGLIE

Claudio Ceccarelli, Simona Rosati, Valentina Talucci

Premessa

Nel 2011 l'Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA) e l'Istituto nazionale di statistica (Istat) hanno stipulato un "Accordo di collaborazione per la progettazione e la realizzazione di un'indagine sui consumi energetici del settore residenziale". Tale indagine[1], realizzata tra marzo e giugno del 2013, ha tra i suoi obiettivi quello di stimare i consumi energetici e le spese delle famiglie per destinazione finale (cucina, riscaldamento e raffrescamento degli ambienti, illuminazione, utilizzo degli elettrodomestici) e fonte energetica (gas, energia elettrica, gasolio, biomasse, ecc.).

Il lavoro, oltre a evidenziare le innovazioni di processo che hanno portato al miglioramento delle stime, ha l'obiettivo di proporre una valutazione della strategia di stima rispetto all'introduzione dei dati amministrativi provenienti da fonte fiscale per mitigare l'effetto della possibile distorsione dovuta alla lista di campionamento. In particolare, partendo dal presupposto che l'uso degli stimatori da modello fanno leva sulle correlazioni tra variabili di interesse e le variabili esogene, è stata individuata la variabile "reddito familiare complessivo" quale ulteriore variabile esogena correlata con il fenomeno oggetto di stima. Tale variabile, desunta dall'archivio amministrativo dei redditi, è stata integrata sia nel campione sia nella lista di campionamento per poter essere poi inglobata nel procedimento di stima.

1. Il disegno di campionamento

Il disegno di campionamento è a uno stadio, con stratificazione dei circa 8.000 comuni italiani per regione e tipologia socio-demografica dei comuni (definita in base a ampiezza demografica e zona altimetrica). La procedura di selezione delle famiglie è casuale semplice, a partire dall'archivio informatizzato ufficiale delle

[1] L'indagine è stata realizzata con tecnica CATI (Computer Assisted Telephone Interview).

famiglie abbonate alla rete di telefonia fissa. All'interno di ciascuna famiglia sono stati intervistati coloro, di età superiore ai 18 anni, indicati dalle famiglie stesse come i più idonei a fornire informazioni sui fenomeni oggetto di intervista.

Tenuto conto dei tradizionali tassi di rifiuto associati alla tecnica di indagine utilizzata e al fine di garantire il numero di interviste teoriche, si è fatto ricorso all'uso della tecnica delle "sostituzioni" che, per l'indagine in questione, ha previsto cinque famiglie sostitutive per ogni famiglia base considerata.

2. La qualità delle stime nelle indagini CATI

Se da un lato l'utilizzo della tecnica CATI presenta indubbi vantaggi, tra i quali una gestione più semplificata della rete di rilevazione, il contenimento dei costi e la tempestività della raccolta dei dati, dall'altro necessita di un attento e continuo monitoraggio del lavoro di raccolta dati e un'attenta valutazione di tutti gli aspetti legati a possibili effetti distorsivi sulle stime dovuti alla perdita di rappresentatività dei campioni raccolti (Istat 2005).

A questo vanno ad aggiungersi i problemi di rappresentatività dovuti agli errori di copertura delle liste telefoniche, che nel corso del tempo hanno assunto sempre più rilevanza a causa delle loro importanti implicazioni per le indagini condotte con tecnica CATI.

Si pone dunque il problema della rappresentatività dei campioni estratti dalle liste telefoniche a causa della sottocopertura e, in definitiva, dei relativi effetti sulle stime dovute a tali problemi già a partire dai campioni estratti per arrivare a quelli realizzati. In sostanza, il rischio che si corre è quello di produrre distribuzioni campionarie delle diverse tipologie familiari difformi dalle analoghe calcolate sulla popolazione.

3. La fase di integrazione con gli archivi amministrativi

Sempre più frequentemente, nella produzione di statistiche ufficiali, gli Istituti nazionali di statistica si avviano verso un uso congiunto delle fonti disponibili affiancando i dati provenienti da rilevazioni campionarie alle informazioni provenienti da archivi amministrativi e viceversa, con il duplice obiettivo di aumentare il grado di affidabilità e qualità del processo di produzione statistica e di innescare economie di scala a beneficio della qualità complessiva del sistema delle statistiche nazionali prodotte dagli Istituti di statistica.

Come conseguenza è sempre più pressante l'attenzione alla qualità dei processi di raccolta dei dati dal lato sia delle rilevazioni campionarie, come ormai è prassi

consolidata, sia dei processi di acquisizione e trattamento statistico degli archivi amministrativi. Analogamente a quanto accade per le indagini dirette, l'uso dei dati di fonte amministrativa implica, infatti, una fase di valutazione dei requisiti di qualità necessari affinché la fonte stessa possa essere ritenuta effettivamente utilizzabile a fini statistici. In tal senso, la valutazione della qualità riveste, dunque, un ruolo decisivo sia nel caso di fonti già inserite nel processo produttivo, che necessitano di essere costantemente monitorate al fine di rilevare errori imprevisti o variazioni connesse, ad esempio, a eventuali cambiamenti normativi, sia per le nuove fonti al fine di verificare la loro utilizzabilità nel processo di produzione. Quando un archivio amministrativo viene acquisito da un Istituto nazionale di statistica risulta pertanto necessario che esso sia affiancato da una serie di indicatori, utili a fornire una prima indicazione della sua qualità (Daas 2011).
Solo una volta chiusa questa prima fase o "ciclo di vita" dei microdati è possibile passare alla seconda fase di integrazione dei dati da fonti diverse.

Con riferimento a questo quadro generale rispetto all'uso delle fonti amministrative, in questo lavoro abbiamo ricavato la variabile ausiliaria "reddito familiare complessivo" attraverso un processo di integrazione di più fonti amministrative.

Le fonti amministrative individuate sono due. La prima è la Banca Dati Reddituale di fonte Ministero dell'Economia e Finanza, che conta circa 41 milioni di record, in cui è contenuta la variabile ausiliaria di interesse 'reddito' e si compone di tre gruppi di file relativi ai contribuenti che hanno compilato: il modello UNICO Persone fisiche o il modello 730 o le dichiarazioni presenti solo nei modelli 770 .

Il reddito complessivo è quello che viene utilizzato per determinare l'imposta sui redditi delle persone fisiche (Irpef), dovuta per l'anno d'imposta; è dato dalla somma dei singoli redditi: dominicali, agrari, da fabbricati, da lavoro dipendente o autonomo, d'impresa in contabilità ordinaria, d'impresa in contabilità semplificata, d'imprese consorziate, da partecipazione, da plusvalenze di natura finanziaria, altri redditi, da allevamento, da tassazione separata.

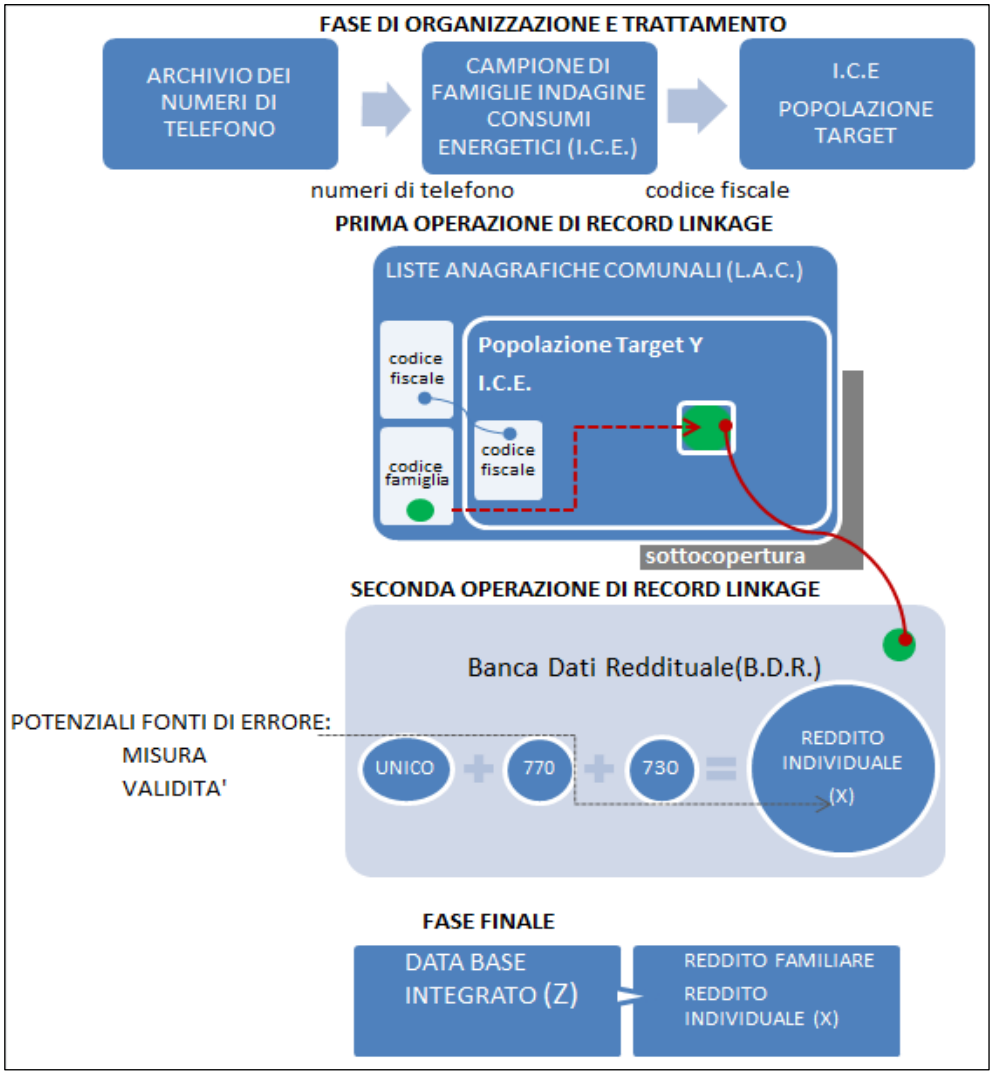
La seconda fonte utilizzata per l'integrazione è la Lista Anagrafica Comunale (LAC) che consente di connettere ogni unità individuale alla rispettiva famiglia anagrafica.

Il processo d'integrazione fa riferimento a metodologie di record linkage[2] con lo scopo di individuare la presenza della stessa unità elementare (persona fisica, persona giuridica, luogo, relazione) nelle diverse fonti prescelte; questo obiettivo è reso possibile attribuendo una o più chiavi di identificazione univoca e stabile nel

[2] Essenzialmente le metodologie di record linkage possono così classificarsi: deterministiche, quando si dispone di chiavi univoche di aggancio; probabilistiche quando si ricorre a profili di similarità tra unità ottenuti da un insieme di variabili caratterizzanti i record.

tempo. In particolare, nel caso dell'Indagine sui consumi energetici delle famiglie (data set target Y), l'integrazione tra le diverse fonti individuate è stata effettuata attraverso record linkage deterministico, resa possibile grazie alla presenza di chiavi di aggancio univoche, cioè di una o più variabili comuni per tutte le fonti.

Prospetto 1 – *Le fasi dell'integrazione tra dati amministrativi e dati campionari*



Come mostra il Prospetto 1, il processo di integrazione è stato preceduto da una fase "zero" di organizzazione e pretrattamento che ha consentito di riconnettere ad ogni singola unità campionaria la variabile chiave 'codice fiscale' proveniente dall'archivio dei numeri telefonici (Consodata).

Si è quindi effettuato un primo linkage deterministico tra i record del campione e l'archivio delle liste anagrafiche comunali (LAC) attraverso la chiave identificativa univoca "codice fiscale"; il tasso di abbinamento è stato del 91%, questo significa che per poco più di 9 famiglie su 10 è possibile disporre delle chiavi di aggancio con la Base Dati dei Redditi (BDR). In questa fase ogni errore nei dati di input o incompatibilità fra dati può dare origine a un errore di copertura della popolazione target, in particolare di sottocopertura, cioè individui presenti nel campione e non nella LAC. La causa di questa sottocopertura può essere addebitata ad errori formali nella chiave di aggancio 'codice fiscale'.

Sono possibili anche errori di allineamento. Lo scopo dell'allineamento è di identificare tutte le relazioni fra le unità abbinate (individui) per la creazione dell'unità statistica (famiglia anagrafica). Nel caso oggetto di studio si tratta della relazione che lega ogni singolo soggetto, identificato dal codice fiscale, con la rispettiva famiglia anagrafica identificata dal codice famiglia. Una relazione di questo tipo viene detta "di molti a uno" fra le unità di base e le unità composte. Fonti diverse, in questo caso LAC e BDR, possono contenere informazioni in conflitto relativamente alle relazioni fra unità di base e unità composte, cioè un individuo può far parte di un "grappolo" famiglia in un archivio e non nell'altro, a causa di variazioni anagrafiche, mancate cancellazioni, ecc..

La seconda fase di linkage, anche in questo caso di tipo deterministico, è stata effettuata tra gli individui del campione, dove è stato possibile assegnare la chiave identificativa della famiglia come concatenamento del codice comune, codice provincia e codice famiglia anagrafica, e la Banca Dati Reddituale. L'output così ottenuto ha consentito di attribuire per ciascuna famiglia del campione, che si è agganciata con la BDR, il reddito complessivo familiare, che si ricorda essere stato calcolato come somma dei redditi individuali. Il tasso di abbinamento di questa seconda fase è stato superiore al 90%.

Si è ottenuto così il data set integrato (Z), a partire dalla popolazione target (Y), sul quale si è importata la variabile ausiliaria 'reddito familiare' (X). Se le funzioni di distribuzione del reddito sulla popolazione target f(Y=y, X=x) e sul data set integrato f(Z=y, X=x) hanno la stessa distribuzione vuol dire che i dati integrati (Z, X) forniscono inferenza valida se usati al posto di (Y, X). La validità statistica è così definita rispetto alle funzioni di distribuzione (Zhang 2012).

4. La strategia di stima

Lo stimatore utilizzato per il calcolo dei coefficienti di riporto all'universo è lo stimatore di ponderazione vincolata, metodo largamente utilizzato nell'ambito della statistica ufficiale. Lo stimatore di ponderazione vincolata, noto in letteratura come *calibration estimator* (Deville, Särndal 1992; Särndal 2007), oltre a migliorare l'accuratezza delle stime, ha il vantaggio di garantire la coerenza delle stime prodotte con le informazioni ausiliarie note (totali di popolazione noti da fonti esterne all'indagine). Questa classe di stimatori comporta la costruzione di pesi di riporto all'universo che sono molto vicini a quelli base (o da disegno), ma che contemporaneamente soddisfano vincoli di coerenza sulle variabili ausiliarie, imposti attraverso le cosiddette equazioni o vincoli di calibrazione.

Inizialmente le stime sono state calcolate prendendo in considerazione i seguenti totali noti:

- popolazione residente per sesso e classe di età nelle cinque ripartizioni territoriali (Nord Est, Nord Ovest, Centro, Sud e Isole);
- popolazione residente per regione (incluse Trento e Bolzano);
- numero di famiglie residenti per regione;
- popolazione di 15 anni e oltre per condizione professionale e posizione nella professione (lavoratori alle dipendenze, autonomi, disoccupati, altri).

I primi tre totali sono stati desunti da fonti demografiche (anagrafiche), mentre i totali riferiti alla condizione professionale e alla posizione nella professione derivano dall'Indagine sulle forze di lavoro (anno 2013). Indicheremo con T1 lo stimatore corrispondente.

Non essendo soddisfacenti i risultati ottenuti con T1, presumibilmente a causa degli effetti distorsivi dovuti alle imperfezioni della lista di campionamento, è stato necessario individuare ulteriori variabili ausiliarie correlate con la variabile oggetto di interesse, con lo scopo di ridurre la distorsione e allo stesso tempo migliorare la precisione delle stime. Dall'integrazione dei dati di indagine con opportuni archivi amministrativi, così come descritto in dettaglio nel paragrafo precedente, è stata individuata la variabile *reddito familiare complessivo* desumibile dalla Banca dati reddituale. Lo stimatore utilizzato, indicato con T2, tiene conto dell'elevato grado di correlazione tra le principali variabili oggetto di indagine e il reddito complessivo disponibile. L'ipotesi di fondo è che l'elevata correlazione tra reddito e variabili oggetto di studio possa produrre stime più accurate rispetto a T1.

Lo stimatore T2 è stato costruito secondo il metodo delle *classi di aggiustamento*. Tale metodo consente la correzione dell'errore di sottocopertura della lista di campionamento nella misura in cui sono note a livello di popolazione obiettivo le dimensioni delle classi di aggiustamento e vale l'ipotesi che la media

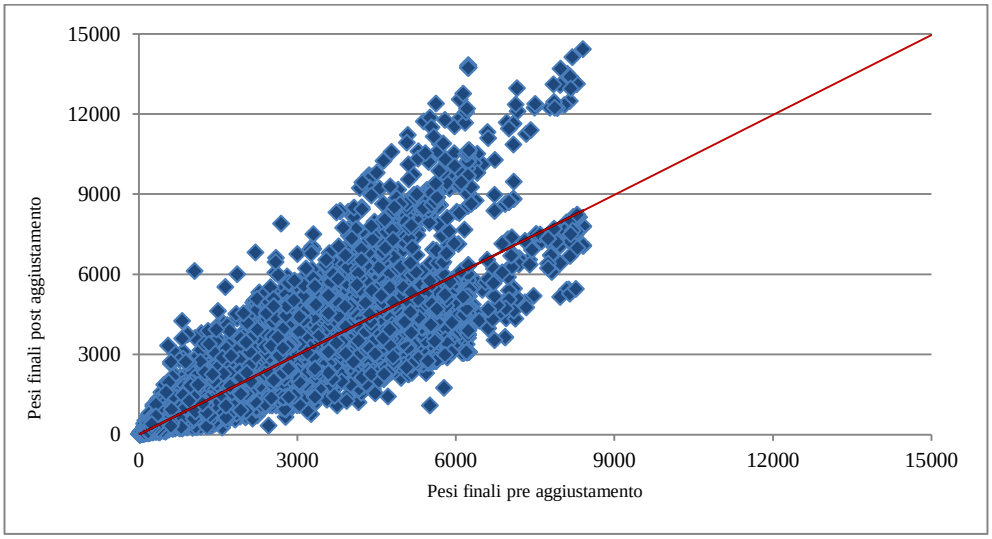
delle unità listate in ciascuna classe è uguale a quella delle unità non listate (Nicolini et al. 2013)[3].

Le classi o celle di aggiustamento sono state costruite riportando per ciascuna delle cinque ripartizioni territoriali la corrispondente distribuzione dei quintili di reddito; quindi, sulla base di quanto è noto per l'intera popolazione, per ciascuna cella è stato calcolato un fattore correttivo tale che, moltiplicato per il peso base, riproducesse nel campione la distribuzione del reddito della popolazione. Si usa in questo caso parlare anche di riponderazione, intendendo con tale termine la modifica dei pesi base al fine di incrementare l'efficienza dello stimatore e in senso lato la rappresentatività delle unità campionarie mediante il loro peso di riporto.

5. L'analisi dei due stimatori

Nel presente paragrafo si vuole valutare l'effetto tra i due differenti stimatori, T1 e T2 proposti nel paragrafo precedente, in modo da poter quantificare l'effetto sulle stime che le classi di aggiustamento hanno prodotto.

Figura 1 – *Confronto tra pesi finali pre e post aggiustamento*

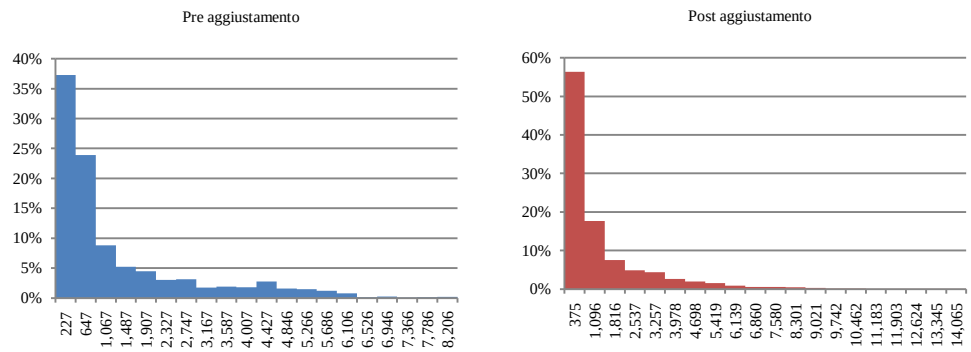


Nella Figura 1 si mostra il grafico a dispersione delle coppie dei coefficienti di riporto all'universo ottenuti con i due stimatori T1 e T2. Emerge chiaramente una

[3] Nel nostro caso vale l'ipotesi che la spesa media per consumi energetici delle famiglie con telefono sia uguale a quella delle famiglie senza telefono.

maggiore variabilità dello stimatore T2 (post aggiustamento) rispetto a T1 (pre aggiustamento), come del resto era prevedibile, avendo introdotto un'ulteriore fase nel processo di stima. Si può osservare, infatti, che a valori meno elevati dei coefficienti di riporto calcolati con lo stimatore T1 corrispondono valori più elevati calcolati con lo stimatore T2.

Figura 2 – *Distribuzione dei pesi finali pre e post aggiustamento – livello nazionale*



La Figura 2 mostra che la distribuzione dei pesi finali ottenuti con lo stimatore T2 ha una forma più regolare e omogenea rispetto a quella ottenuta con lo stimatore T1. Ciò significa che lo stimatore T2 produrrà stime maggiormente stabili rispetto a T1 nei diversi domini di studio. Per motivi di esposizione riportiamo soltanto le distribuzioni complessive, ma la stessa analisi condotta per ciascuna ripartizione territoriale rende ancora più evidente questo risultato.

A completamento dell'analisi fin qui svolta si introducono alcuni risultati che mostrano l'andamento per regione dell'effetto stimatore calcolato per le principali spese per combustibili ed energia a livello regionale. In particolare è stato calcolato tale effetto per entrambi gli stimatori.

I valori per tali effetti sono visualizzati nelle figure 3, 4, 5, e 6 relativamente alla spesa complessiva per energia, per gasolio, per metano e per legna. Per ogni tipologia di spesa l'effetto dello stimatore T2 è più elevato rispetto a quello dello stimatore T1.

La differenza tra gli effetti prodotti sull'errore campionario dovuti ai due stimatori è più contenuta in quei domini (regioni) dove si è riusciti a intervistare un campione più rappresentativo rispetto alla variabile usata per la costruzione delle classi di aggiustamento. In altri termini, la variabile ausiliaria utilizzata non ha fornito alcun contributo all'effetto stimatore, in quel dominio, ovvero al non significativo contributo aggiuntivo nel modello di stima.

Figura 3 – *Effetto stimatore per la stima della spesa complessiva per energia*

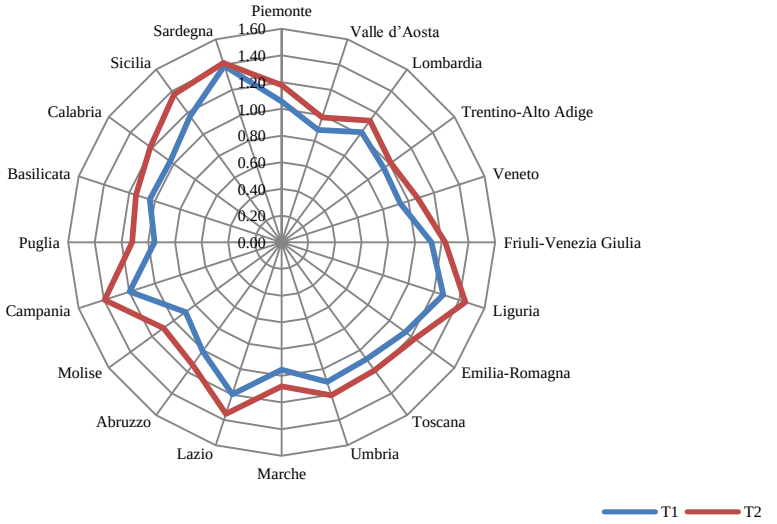


Figura 4 – *Effetto stimatore per la stima della spesa per metano*

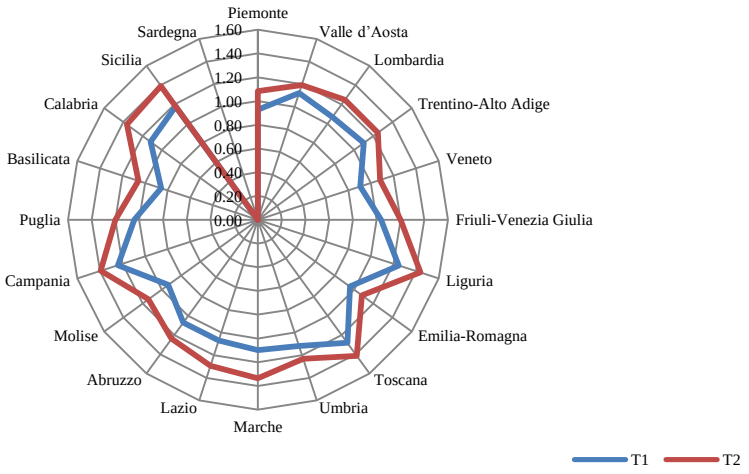


Figura 5 – *Effetto stimatore per la stima della spesa per gasolio*

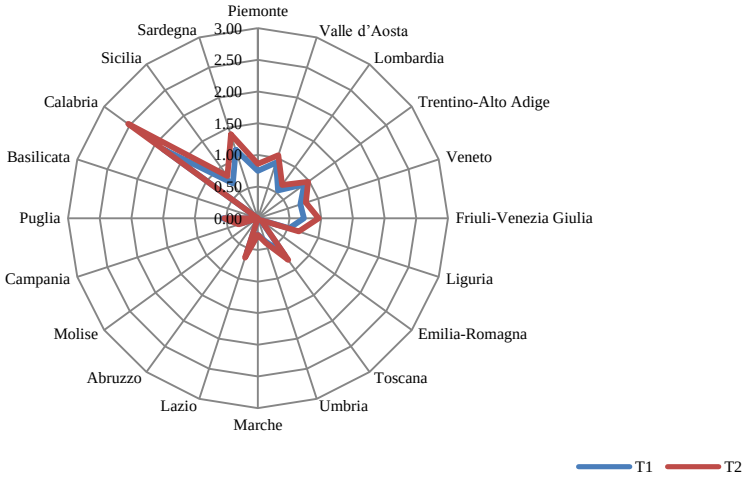
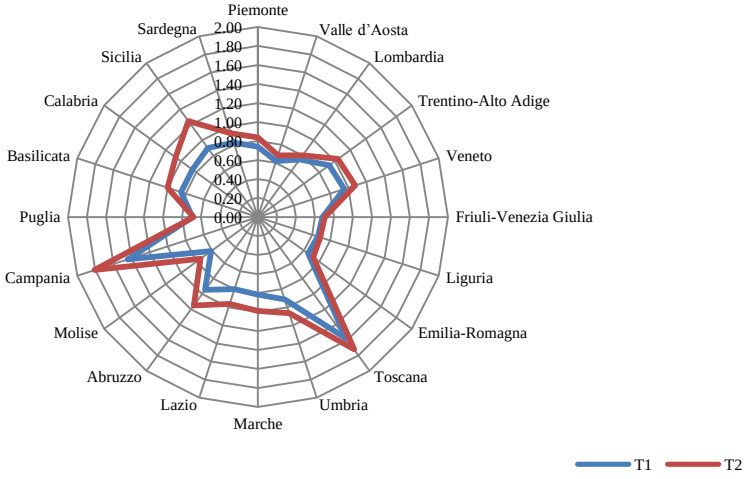


Figura 6 – *Effetto stimatore per la stima della spesa per legna*



6. Conclusioni

Le ipotesi da cui si è partiti sono fondamentalmente due: la prima, più che altro un dato di fatto, è la sottocopertura delle liste telefoniche; la seconda riguarda l'assunzione che all'interno delle classi di aggiustamento ci sia uguale comportamento di consumo energetico tra chi ha e chi non ha linea telefonica fissa. Il ricorso a liste delle utenze telefoniche di rete fissa, quindi, impone interventi correttivi per garantire un buon grado di rappresentatività rispetto alle principali distribuzioni sia della popolazione di riferimento sia del campione teorico estratto.

L'analisi dei risultati ci mostra che la scelta effettuata ci ha consentito di mitigare gli effetti sulle stime dovuti alla sottocopertura della lista di campionamento. Ciò nasce dal fatto che lo stimatore T2, rispetto allo stimatore T1, consente di preservare le distribuzioni per tipologia familiare e classe di reddito, ovvero uniformare i consumi delle famiglie con telefono fisso e quelle senza telefono fisso tra coloro che hanno le stesse potenzialità di spesa.

Il costo di questa operazione può essere valutato grazie alle analisi proposte in questo lavoro. L'aumento dell'effetto stimatore indica che una parte di informazione viene recuperata, perché non stimata, e "riversata" sull'errore campionario. Grazie all'utilizzo di variabili altamente correlate con la variabile di interesse si contribuisce, quindi, a ridurre le differenze tra le distribuzioni osservate nel campione e quelle presenti nella popolazione a cui si riferiscono le stime. Tuttavia l'aumento della variabilità dei pesi finali dello stimatore T2 non necessariamente implica una minore accuratezza, in quanto tale aumento può essere più che compensato da una minore distorsione delle stime finali.

Riferimenti bibliografici

CECCARELLI C., CUTILLO A. 2007. Il trattamento della mancata risposta totale nell'indagine Eusilc: una valutazione tramite una misura del cambiamento, *Congiuntura*, Udine: CREF, 1° trimetre, pp. 91-112

DAAS P., OSSEN S., TENNEKES M., ZHANG L.C., HENDRIKS C., FOLDAL HAUGEN K., ERRONI F., DI BELLA G., LAITILA T., WALLGREN A., WALLGREN B. 2011. *Report on methods preferred for the quality indicators of administrative data sources*. Second deliverable of WP4 of the BLUE Enterprise and Trade Statistics project. September 28th 2011. http://www.blueets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf

DEVILLE, J.C., C.E. SÄRNDAL. 1992. Calibration Estimators in Survey Sampling. *Journal of american statistical association*. Vol. 87, No.418, pp. 376-382.

ISTAT. 2014. *Annuario statistico italiano*. 2014. http://www.istat.it/it/files/2014/11/Asi-2014.pdf

NICOLINI G., MARASINI D., MONTANARI G.E., PRATESI M., RANALLI M.G., E. ROCCO. 2013. *Metodi di stima in presenza di errori non campionari*. Springer. UNITEXT-Collana di Statistica e Probabilità Applicata.

SÄRNDAL C.E. 2007. The calibration approach in survey theory and practice. *Survey methodology*. Vol. 33, No.2, pp. 99-119.

ZHANG L.C. 2012. Topics of statistical theory for register based statistics and data integration. *Statistica Neederlandica*. Vol. 66, No. 1,pp. 41-63.

SUMMARY

Evaluation of the strategy estimate adopted for the Households energy consumption survey

CATI surveys offer many advantages, such as cost reduction, high timeliness and simpler management of the interviewer network, but they are affected by under-coverage errors, which affect the residential phone directory. This, in turn, affects the quality of estimators, which may be highly inaccurate due to large bias. The aim of this work is to stress the methodological features of the estimation strategy adopted for the Household energy consumption survey, in order to reduce the bias caused by the imperfection of the sampling frame. To reduce the bias we introduce auxiliary information from administrative data sources which are highly correlated with the target variables. This information was integrated in the sample and then included in the estimation process. This paper proposes an evaluation of this strategy.

Claudio CECCARELLI, Istituto nazionale di statistica, clceccar@istat.it
Simona ROSATI, Istituto nazionale di statistica, sirosati@istat.it
Valentina TALUCCI, Istituto nazionale di statistica, talucci@istat.it

CAMBIAMENTI NELL'INDAGINE E DATI DESTAGIONALIZZATI: IL CASO DELLA CLASSIFICAZIONE DELLE ATTIVITÀ ECONOMICHE

Melissa Cortellessa, Cinzia Graziani, Andrea Spizzichino

1. Introduzione

Il processo di innovazione, che ha caratterizzato e continua a contraddistinguere la produzione di dati derivanti da indagini statistiche, consente di avere informazioni sempre più rappresentative della realtà; nell'ottica di cogliere i continui mutamenti dei fenomeni e farlo con la massima tempestività, assume sempre più rilievo l'analisi congiunturale dei fenomeni attraverso dati destagionalizzati.

Se da un lato le innovazioni permettono di cogliere al meglio le caratteristiche più specifiche di un determinato fenomeno, dall'altro creano dei *break* nelle serie storiche che compromettono la possibilità di fare analisi nel medio e lungo periodo.

L'Istat per risolvere i problemi di *break* nelle serie storiche, nel corso degli anni, ha adottato la prassi di definire ricostruzioni delle serie storiche per i periodi precedenti i *break*, in modo da rendere confrontabili i dati vecchi con quelli nuovi.

La disponibilità di serie lunghe e continue garantisce al tempo stesso la possibilità di fare confronti tendenziali nel lungo periodo e di destagionalizzare i dati al fine di condurre analisi congiunturali.

Alla base delle tecniche di destagionalizzazione ci sono una serie di procedure da seguire al fine di minimizzare i cambiamenti (revisioni) di una stima tra due successivi rilasci; proprio i cambiamenti nelle serie dovuti a una ricostruzione possono essere causa di revisioni nei dati destagionalizzati.

Tra le rilevazioni condotte dall'Istat, quella sulle forze di lavoro più di ogni altra, ha subito modifiche nel tempo che hanno reso necessarie ricostruzioni di serie storiche per rendere i vecchi dati coerenti con i nuovi; tra le varie innovazioni, in questo lavoro, faremo riferimento al passaggio dalla vecchia classificazione delle attività economiche (ATECO02) alla nuova ATECO07 e alle problematiche connesse alla ricostruzione e destagionalizzazione dei dati.

Nel presente lavoro si affronta inizialmente il passaggio dalla vecchia alla nuova classificazione delle attività economiche illustrando i principali cambiamenti e come sono stati ricostruiti i vecchi dati, poi si introduce brevemente la teoria della destagionalizzazione per come viene condotta in Istat, infine si analizzano le

revisioni sui dati destagionalizzati determinate dal cambio di classificazione dando sia informazioni teoriche sulle tecniche per il calcolo delle revisioni sia alcuni risultati.

2. Il cambiamento di classificazione delle attività economiche

Come già sottolineato nell'introduzione, la rilevazione sulle forze di lavoro ha subito nel tempo una serie di innovazioni che hanno reso necessaria la ricostruzione di serie storiche relativamente ai principali aggregati del mercato del lavoro, tutto ciò al fine di garantire gli utenti rispetto alla possibilità di fare analisi di medio-lungo periodo.

Tra queste innovazioni l'ultima è avvenuta tra il 2008 e il 2010 con il passaggio dalla vecchia classificazione delle attività economiche ATECO02 alla nuova classificazione ATECO07.

Il cambiamento nella classificazione è risultato necessario sia per motivi di coerenza a livello internazionale sia per rappresentare al meglio i cambiamenti e le innovazioni nelle strutture produttive e nelle forme di organizzazione del lavoro. Tali mutamenti dettati dall'innovazione tecnologica, determinano la creazione di nuovi prodotti o servizi e, di conseguenza, la nascita di nuove attività economiche.

'Nel passaggio alla nuova versione dell'Ateco, la logica sottostante alla classificazione e i criteri adottati per suddividere in raggruppamenti omogenei le attività produttive sono rimasti invariati. Tuttavia, la creazione di un ulteriore livello classificatorio e la maggiore articolazione di diverse sezioni hanno permesso di riflettere più fedelmente le diverse tipologie di attività produttive e di rappresentare, in particolare, le nuove industrie emergenti.

Il dettaglio della classificazione è cresciuto in maniera sostanziale se si considera che il numero delle classi (4°digit) è aumentato di cento unità e quello delle categorie (5°digit) di 35. L'introduzione delle sottocategorie (6°digit), inoltre, ha consentito di rappresentare nei minimi particolari la specificità della struttura produttiva italiana e di arricchire il contenuto informativo della classificazione.

Il settore dei servizi è quello che ha catalizzato la maggior parte dei cambiamenti adottati dalla classificazione, in linea con il processo di 'terziarizzazione' dell'economia e con la conseguente espansione dell'occupazione in questo comparto produttivo, che ha fatto lievitare la sua incidenza intorno al 60% del totale degli occupati.' (Gallo, Scalisi, Spizzichino, 2012).

Tra le principali innovazioni introdotte dalla nuova Ateco c'è la nuova sezione dedicata ai servizi di informazione e comunicazione (sezione J), che raccoglie diverse attività provenienti da altre sezioni relative sia al ramo delle attività

tecnologico-informatiche sia alla produzione e distribuzione di prodotti culturali e informativi.

L'impatto sostanziale del cambiamento che ha investito talune sezioni non ha reso agevole la transcodifica dell'ATECO07 dalla precedente versione. Se si pensa che più del 40 per cento delle 883 categorie dell' ATECO02 non hanno trovato un raccordo 1 a 1 nella nuova classificazione si comprende l'entità dell'impegno profuso per garantire la continuità della serie storica dei dati.

Tutti questi cambiamenti hanno reso necessaria una ricostruzione vera e propria dei dati precedenti il 2008, e in particolare delle serie, per il periodo 2004-2007, relative ad alcuni tra i principali indicatori economici correntemente diffusi.

Al fine di garantire la possibilità di produrre la ricostruzione delle serie in vecchia Ateco e di rendere il più fluido possibile il passaggio dalla vecchia alla nuova classificazione, l'indagine è stata svolta per 3 anni (2008-2011) rilevando la variabile sull'attività economica sia con la vecchia classificazione (Ateco2002) sia con la nuova. Questa scelta ha garantito la disponibilità per 12 trimestri delle informazioni riferite a ogni individuo sia in vecchia sia in nuova Ateco definendo un periodo di sovrapposizione tanto lungo da poter interpretare con accuratezza le relazioni tra le 2 classificazioni.

La tecniche di ricostruzione adottate in questa occasione sono state di tipo macro, tale scelta deriva sia dalla volontà di mantenere la coerenza con precedenti ricostruzioni prodotte in occasione di altri cambiamenti nell'indagine, sia perché le informazioni a disposizione consentivano l'utilizzo di ricostruzioni tramite *strutture* (RAO 2000) senza fare ricorso a imputazioni sui singoli record.

Di fondamentale importanza nella progettazione del lavoro è stata la scelta delle variabili da utilizzare nella ricostruzione e del livello di disaggregazione delle categorie economiche da raggiungere; in particolare si è cercato di ottenere il maggior livello di disaggregazione mantenendo elevata l'attendibilità delle serie ricostruite, il tutto considerando variabili per le quali il settore d'attività economica risulta maggiormente discriminate. In questo lavoro non ci occupiamo di chiarire tutte le scelte fatte ed entrare nel dettaglio della metodologia applicata[1], risulta importante però indicare come il risultato della ricostruzione siano serie di dati compresi tra il 2004 e il 2007, coerenti e confrontabili con quelle prodotte a partire dal 2008 in nuova Ateco, rispetto alle principali variabili demografiche ed economiche. In particolare il dettaglio raggiunto per tutte le attività economiche fino al secondo digit è per regione (con disaggregazione del Trentino-Alto-Adige tra le due province autonome di Trento e Bolzano), sesso, posizione professionale

[1] Per maggiori dettagli si può fare riferimento al par.3 di Gallo F., Scalisi P., Spizzichino A., 2012. 'La transizione alla nuova classificazione delle attività economiche: la ricostruzione delle serie storiche e le specificità del settore turismo'.

(distinta tra dipendenti, autonomi e collaboratori) e carattere dell'occupazione (distinto tra permanente e temporaneo).

3. La destagionalizzazione

La destagionalizzazione delle serie storiche rappresenta una fase cruciale nella produzione dei dati sulle forze di lavoro in quanto solo il dato destagionalizzato consente di analizzare le variazioni che intervengono tra l'ultimo dato e il precedente con riferimento ai principali aggregati e indicatori prodotti.

La necessità di destagionalizzare un dato grezzo nasce dall'ipotesi che una serie storica a cadenza infrannuale e, nello specifico tutte le principali riferite al mercato del lavoro, siano rappresentabili come una combinazione di diverse componenti.

Tali componenti sono il cosiddetto ciclo-trend, che rappresenta la tendenza di medio-lungo termine della serie storica e non è perciò influenzata da oscillazioni di brevissimo periodo; una componente stagionale, che si manifesta nel corso dell'anno in modo ricorrente e, una componente irregolare dovuta a fattori erratici.

Le procedure di destagionalizzazione sono finalizzate all'eliminazione dalla serie grezza della componente stagionale. Esse si basano su un trattamento iniziale dei valori e degli andamenti anomali che possono essere dovuti nel caso delle forze lavoro principalmente a *outliers* legati a fenomeni accidentali. Una volta pre-trattate, le serie vengono sottoposte alla vera e propria individuazione delle componenti attraverso un modello statistico. Sottraendo dalla serie storica originaria la componente relativa alla stagionalità si ottiene la serie destagionalizzata vera e propria che ha un andamento molto meno oscillatorio rispetto a quella di partenza in quanto le discontinuità sono dovute solo alla componente irregolare.

In questo contesto l'Istat, agli inizi del 1997, ha istituito una commissione scientifica con il compito di formulare proposte relative alle strategie da utilizzare per la destagionalizzazione di serie storiche prodotte dall'Istituto: la commissione SARA (*Seasonal Adjustment Research Appraisal*). Tale commissione ha esaminato e confrontato le caratteristiche delle principali procedure di destagionalizzazione disponibili (in particolare, delle procedure TRAMO-SEATS e X-12-ARIMA). I risultati hanno mostrato come la procedura TRAMO-SEATS (*Time Series Regression with Arima noise, Missing observations and Outliers - Signal Extraction in Arima Times Series*) consenta una stima più accurata delle diverse componenti e, in definitiva, la diffusione di una più corretta informazione congiunturale. La procedura TRAMO-SEATS si compone di due parti. La prima, costituita da TRAMO, è dedicata ad eliminare dalla serie storica d'interesse i cosiddetti effetti deterministici dovuti al diverso numero di giorni lavorativi nei

vari periodi di riferimento, alla presenza di festività "mobili" (ad esempio, la Pasqua) e di valori anomali. Inoltre, TRAMO identifica e stima il modello ARIMA per la serie storica osservata. La seconda parte della procedura, costituita da SEATS, effettua la vera e propria destagionalizzazione della serie originaria utilizzando il modello ARIMA e gli effetti deterministici identificato identificati in TRAMO. TRAMO-SEATS è una procedura di tipo *model-based*, cioè basata sull'identificazione di un particolare modello statistico per ciascuna serie storica analizzata; essa incorpora gli avanzamenti compiuti negli ultimi anni nell'ambito della cosiddetta "analisi moderna delle serie storiche" ed offre un ampio spettro di strumenti di carattere statistico per valutare la qualità della destagionalizzazione effettuata. In particolare tale procedura viene implementata dal *software* Demetra, sviluppato da Eurostat, destinato a fornire uno strumento pratico e flessibile per la destagionalizzazione utilizzando sia TRAMO-SEATS sia i metodi X-12-ARIMA.

4. Le revisioni

L'obiettivo di questo lavoro è l'analisi dell'impatto della ricostruzione di serie storiche sui dati destagionalizzati in termini di revisione dei dati precedentemente diffusi. Con il termine revisione si intende ogni differenza tra due serie relative allo stesso aggregato, nel caso dei dati destagionalizzati questa può verificarsi a seguito di:

- Disponibilità di un set informativo più completo a livello di dati grezzi.
- Migliore stima/identificazione della componente stagionale nel processo di destagionalizzazione a seguito dell'inserimento di un nuovo dato grezzo. Con l'aggiunta di un nuovo dato infatti, i modelli di destagionalizzazione, in base alla politica di revisione seguita, subiscono un riadattamento, facendo sì che i dati destagionalizzati precedentemente diffusi siano rivisti[2].

La revisione rappresenta una dimensione di qualità del dato: rivedere dati già diffusi (sia grezzi sia destagionalizzati) può mettere infatti in discussione l'affidabilità dei dati stessi. È pertanto necessario, se non d'obbligo, effettuare un'analisi delle revisioni e agire in maniera da minimizzarne l'impatto.

Ogniqualvolta è disponibile una nuova informazione, sia essa un aggiornamento dei dati grezzi, sia essa un dato aggiuntivo, la serie storica viene destagionalizzata nuovamente. A seconda della politica di revisione adottata nell'ambito della destagionalizzazione, la revisione potrà essere più o meno contenuta. Se si tratta di un aggiornamento ordinario dei dati, come dell'aggiunta infrannuale di un nuovo dato o del passaggio da una stima provvisoria a una definitiva, è buona prassi,

[2] Una descrizione dettagliata delle diverse politiche di revisione è disponibile nelle ESS Guidelines on Seasonal Adjustment – Revision policies.

proprio al fine di minimizzare le revisioni, mantenere i modelli, i filtri, gli *outliers* e i regressori di calendario della precedente destagionalizzazione e ristimarne solo i parametri (approccio *Partial Concurrent Adjustment*).

4.1.1 Modelli consolidati o modelli identificati in automatico: confronto in termini di revisione

Nel caso della ricostruzione in oggetto, per evitare che alle revisioni riconducibili ai nuovi dati grezzi si aggiungessero anche quelle derivanti dalla politica di destagionalizzazione, si è scelto di seguire l'approccio *Partial Concurrent Adjustment*, salvo restando l'adattabilità dei modelli, consolidati sui vecchi dati, sulle nuove serie. Una strategia alternativa sarebbe stata invece quella di rivedere tutti i modelli lasciando al *software*[3] il compito di identificarli. La prima opzione è stata ritenuta preferibile, anche una volta osservato che la ricostruzione non aveva apportato cambiamenti sostanziali nei dati grezzi disaggregati per i principali settori di attività economica (in Agricoltura, Industria in senso stretto, Costruzioni e Servizi).

Tabella 1 – *Modelli consolidati e modelli identificati in automatico per le serie trimestrali degli occupati per ripartizione geografica e settore di attività economica.*

Serie storiche degli occupati	Modelli consolidati	Modelli identificati in automatico
Nord: Agricoltura	(2 1 0)(0 0 1)	(2 1 0)(0 0 1)
Nord: Industria in senso stretto	(0 0 1)(1 0 0)	(2 1 0)(0 0 0)
Nord: Costruzioni	(0 1 0)(0 1 0)	(2 1 0)(0 0 0)
Nord: Servizi	(0 1 0)(0 1 0)	(0 1 0)(1 0 0)
Centro: Agricoltura	(1 0 0)(0 1 1)	(3 0 0)(0 0 0)
Centro: Industria in senso stretto	(0 1 1)(1 0 0)	(1 0 0)(0 0 1)
Centro: Costruzioni	(1 1 0)(1 0 0)	(1 0 0)(0 1 1)
Centro: Servizi	(1 0 0)(0 1 1)	(3 1 0)(0 0 0)
Mezzogiorno: Agricoltura	(0 0 1)(0 1 0)	(1 0 0)(1 0 0)
Mezzogiorno: Industria in senso stretto	(0 1 1)(0 1 1)	(2 1 0)(1 0 0)
Mezzogiorno: Costruzioni	(0 1 1)(1 0 0)	(0 1 1)(0 1 1)
Mezzogiorno: Servizi	(1 1 0)(1 0 0)	(0 1 1)(0 1 1)

Nota: In entrambi i casi i modelli identificati superano i principali test predisposti dal software.

Dalla Tabella 1 è possibile effettuare un primo confronto tra le due strategie: per ogni serie storica relativa agli occupati, disaggregata per ripartizione geografica e settore di attività economica, sono riportati i modelli consolidati e quelli identificati in automatico dal *software* sui nuovi dati ricostruiti.

[3] Il *software* utilizzato per la destagionalizzazione è Demetra 2.2, nel quale sono implementate sia la procedura TRAMO-SEATS sia la procedura X-12-ARIMA.

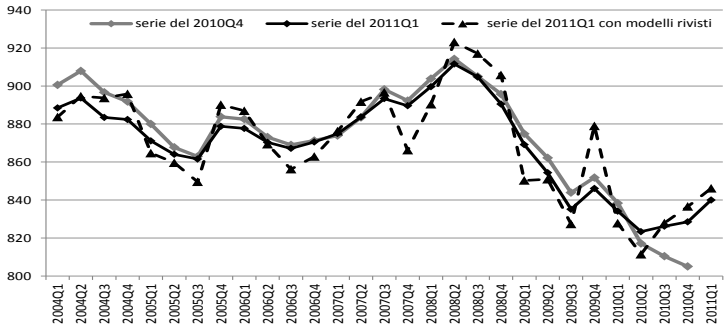
Si può osservare che per undici delle dodici serie il modello scelto in automatico non sarebbe coinciso con quello consolidato. Solo per la serie storica degli occupati in agricoltura del Nord il modello identificato e consolidato sui dati in ATECO02 risulta confermato anche dall'identificazione automatica sui dati ricostruiti in ATECO07.

Una volta verificato che per ogni serie entrambi i modelli hanno superato i consueti test diagnostici e pertanto entrambi avrebbero potuto essere selezionati, un utile supporto per valutare le due opzioni e, in particolare quella poi scelta, è fornito dall'analisi delle revisioni che si sarebbero verificate con l'una e con l'altra strategia.

In particolare l'analisi delle revisioni è volta a valutare le differenze tra le serie destagionalizzate diffuse nel primo trimestre in cui la nuova Ateco è entrata a regime (I trimestre 2011) e la precedente diffusione (IV trimestre 2010) in cui le attività economiche venivano ancora rilasciate in ATECO02.

Il grafico in figura 1 riporta a titolo esemplificativo le serie degli occupati nell'industria in senso stretto al Centro.

Figura 1 – *Occupati nell'industria in senso stretto al Centro. I trimestre 2004 - I trimestre 2011. Dati destagionalizzati (valori in migliaia).*



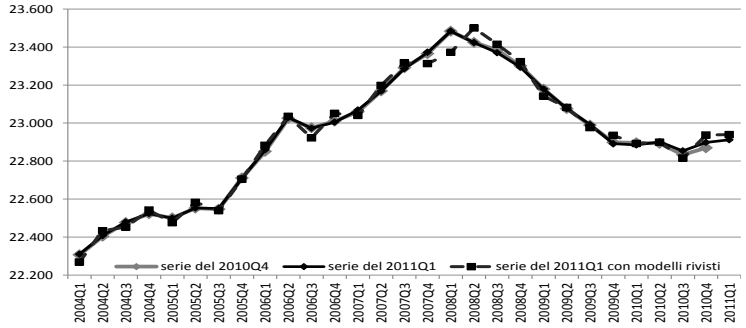
La serie diffusa in occasione del IV trimestre 2010 è denominata 'serie del 2010Q4', quella 'del 2011Q1' è la serie rilasciata con il I trimestre 2011 utilizzando i modelli consolidati, mentre la 'serie del 2011Q1 con modelli rivisti' mostra il risultato ottenuto scegliendo di utilizzare il modello di destagionalizzazione identificato in automatico dal *software*.

Da un'analisi grafica si evidenzia che la serie identificata in automatico presenta una dinamica congiunturale più accentuata e un andamento caratterizzato da forte discontinuità. La serie destagionalizzata utilizzando i modelli consolidati (serie del 2011Q1) invece, pur presentando livelli sempre leggermente al di sotto di quelli della serie del 2010Q4, ne mantiene il profilo congiunturale, ad eccezione del terzo

e quarto trimestre 2010, i cui i livelli sono rivisti al rialzo e le variazioni congiunturali, che presentavano due segni negativi consecutivi, sono riviste in crescita.

Per quanto riguarda gli occupati nel complesso, la serie destagionalizzata è ottenuta in maniera indiretta, ovvero come somma delle dodici serie destagionalizzate disaggregate per settore di attività economica e ripartizione geografica; l'analisi grafica (figura 2) non evidenzia differenze di rilievo tra le serie, soprattutto per quanto riguarda la dinamica.

Figura 2 – *Occupati. I trimestre 2004 - I trimestre 2011. Dati destagionalizzati (valori in migliaia).*

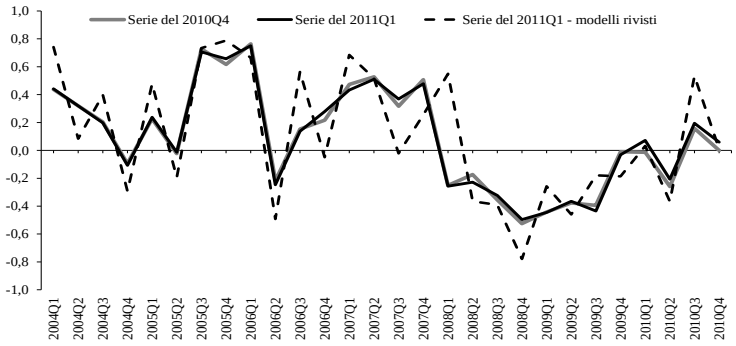


L'unica eccezione si rileva nel periodo tra il secondo trimestre del 2007 e il secondo trimestre del 2008, in cui la serie ottenuta sulla base di modelli identificati in automatico sembra cogliere con ritardo l'aumento dell'occupazione.

Sempre con riferimento agli occupati nel complesso, dal grafico relativo alle variazioni congiunturali (figura 3) è possibile notare come la serie ottenuta con modelli consolidati mantenga la dinamica della serie pre-ricostruzione, mentre quella ottenuta da modelli rivisti in automatico la amplifica o addirittura ne rivede il segno in molte delle occorrenze considerate. Quanto rilevato graficamente può essere analizzato e approfondito attraverso l'analisi delle revisioni che, se effettuata rispetto alle variazioni congiunturali, permette di apprezzare l'impatto delle due strategie sulla dinamica della serie[4].

[4] Poiché lo scopo dei dati destagionalizzati è lo studio dell'andamento congiunturale di una serie si predilige effettuare l'analisi delle revisioni delle variazioni congiunturali piuttosto che quella dei livelli assoluti.

Figura 3 – *Occupati - tassi di crescita congiunturali. I trimestre 2004 - I trimestre 2011. Valori percentuali.*



L'analisi delle revisioni si basa su indicatori specifici costruiti a partire dalla differenza

$$R^i = X_t^i - X_{tp}^i \quad (1)$$

Dove R^i indica la revisione osservata sull'i-esimo dato tra la serie diffusa al tempo t (X_t^i) e quella diffusa precedentemente al tempo tp (X_{tp}^i).

Tabella 2 – *Analisi delle revisioni sulle variazioni congiunturali della serie destagionalizzata degli occupati rispetto alla serie diffusa in ATECO02 con il 2010Q4. – I trimestre 2004 - IV trimestre 2010. Indicatori di sintesi.*

Indicatori di sintesi delle revisioni rispetto alla serie diffusa con il 2010Q4 (in punti percentuali)	Serie del 2011Q1 – modelli consolidati	Serie del 2011Q1 - modelli rivisti
Revisione Media Assoluta	0,03	0,22
Revisione Media	0,00	0,02
Deviazione standard - HAC formula	0,01	0,03
Revisione Media Quadrata	0,00	0,07
Revisione Media Assoluta Relativa	0,08	0,53
t-stat	0,71	0,56
t-crit	2,06	2,06
Significatività della revisione media	NO	NO
Correlazione	1,00	0,81
Revisione Minima	-0,05	-0,34
Revisione Massima	0,08	0,80
Range di variazione	0,14	1,14
% Segno(X_t) = Segno(X_{tp})	96,30	85,19

Con riferimento agli occupati nel complesso è stato possibile calcolare gli indicatori di sintesi riportati in Tabella 2. In base al test di significatività effettuato,

entrambe le serie non mostrano revisioni significative rispetto alla serie diffusa con il quarto trimestre 2010. Tuttavia la serie ottenuta utilizzando i modelli consolidati presenta revisioni più contenute: la revisione media assoluta è pari a 0,03 punti percentuali contro gli 0,22 punti della serie con modelli rivisti; il campo di variazione è più stretto, pari a 0,14 contro 1,14 e, inoltre, il segno della variazione congiunturale è rivisto solo in poco più del 4% dei casi, contro il 15% della serie ottenuta con i modelli rivisti identificati in automatico[5].

5. Conclusioni

Nel presente lavoro sono stati analizzati gli effetti delle ricostruzioni sui dati destagionalizzati. È stato in particolare analizzato il caso della ricostruzione delle serie storiche trimestrali degli occupati in occasione del passaggio dalla classificazione delle attività economiche ATECO02 alla nuova classificazione ATECO07. Dopo avere brevemente illustrato il cambiamento intercorso e accennato alla metodologia adottata per la ricostruzione dei dati grezzi, è stato analizzato l'impatto sui dati destagionalizzati. A tal proposito sono stati forniti dei cenni sulla destagionalizzazione, sulle politiche e sui metodi di scomposizione delle serie storiche seguiti in Istat. In merito al caso studio qui presentato, è stata descritta la *policy* seguita per la destagionalizzazione delle serie storiche disaggregate per settore di attività economica e ripartizione geografica. La strategia seguita si basa sull'approccio cosiddetto *Partial Concurrent Adjustment*: in occasione dell'inserimento dei nuovi dati ricostruiti si è scelto quindi di mantenere i modelli di destagionalizzazione, già consolidati sui dati in vecchia Ateco.

Questa scelta è stata confrontata con un approccio alternativo, basato invece sulla rivisitazione di tutti i modelli. Il confronto è stato effettuato analizzando le revisioni che si sono o sarebbero verificate adottando l'una o l'altra, rispetto alle serie diffuse prima della ricostruzione (in occasione del rilascio del quarto trimestre 2010). I risultati dell'analisi delle revisioni supportano la validità della strategia seguita, basata sul mantenimento dei modelli di destagionalizzazione consolidati in un'ottica di minimizzazione delle revisioni, di continuità delle serie e quindi di diffusione. Tuttavia non si può escludere che, con l'adozione di una tale strategia, si sia penalizzata la capacità della nuova classificazione di cogliere diversamente la stagionalità delle serie.

[5] Per approfondimenti sull'analisi delle revisioni e sugli indicatori utilizzati cfr. McKenzie R., Gamba M., *Interpreting the results of Revision Analysis: Recommended Summary Statistics*, Contribution to the OECD/Eurostat Task Force on 'Performing Revisions Analysis for Sub-Annual Economic Statistics'

Riferimenti bibliografici

ANTINORI P., BACCHINI F., BALDI C., BRUNO G., CAMMAROTA M., DE VITA M., DI IORIO F., GATTO R., OTRANTO E., PALLARA A, POLIDORO F., POLITI M., TRIACCA U., 2000. Guida all'utilizzo di TRAMO-SEATS per la destagionalizzazione delle serie storiche, *Documenti n. 4*, Istat.

EUROSTAT, 2015. ESS guidelines on seasonal adjustment, *Eurostat Manuals and Guidelines*, http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf .

EUROSTAT, 2013. ESS guidelines on revision policy for PEEIs, *Eurostat Methodologies and Working papers*, http://ec.europa.eu/eurostat/documents/3859598/5935517/KS-RA-13-016-EN.pdf

GALLO F., SCALISI P., SPIZZICHINO A., 2012. La transizione alla nuova classificazione delle attività economiche: la ricostruzione delle serie storiche e le specificità del settore turismo, *SIEDS, Rivista Italiana di Economia Demografia e Statistica*.

ISTAT, 2006, La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione. Istat, Metodi e Norme, No. XX.

ISTAT, 2009, Classificazione delle attività economiche Ateco 2007, *Metodi e Norme, No. 40*

McKENZIE R., GAMBA M., 2008. Interpreting the results of Revision Analyses: Recommended Summary Statistics. *Contribution to the OECD/Eurostat Task Force on "Performing Revisions Analysis for Sub-Annual Economic Statistics"*, http://www.oecd.org/std/40315546.pdf .

PURCEL N.J., KISH L., 1980, Postcensal estimates for local Areas (or domains), *International Statistical Review, 48, 3-18*.

RAO J.N.K. 2000. Statistical methodology for indirect estimations in small areas, *39, Eustat.*

SUMMARY

Survey changes and seasonal adjusted data: the case study of the classification of economic activities

The analysis of seasonally adjusted data collects more and more interest among users of data on the labour market, especially in the presence of changing economic conditions such as those that characterize our country in recent years.

The labour force survey produces information on the labour market aggregates which allows to perform analysis on short-term developments and, therefore, which is based on data adjusted for the seasonal effect, or so-called seasonally adjusted data.

For the main aggregates, employment, unemployment and inactivity, seasonally adjusted data are released disaggregated by socio-demographic variables, such as age, gender and geographical area; only for the employment it is also analyzed the economic sector of activity.

Innovations in the production process of the survey data or changes in definitions and classifications with respect to strategic variables, can determine both breaks in the time series of the raw data and differences in the models that define the seasonal pattern of an aggregate.

For the analysis of economic data, it is therefore necessary to control if revisions between two releases are due to changes in the survey or to the performance of the labour market.

The exercise presented in this paper aims to investigate how the transition to the current classification of economic activities (ATECO07) may cause differences: in the seasonal adjustment models, in the seasonally adjusted data and growth rates of the series of employment, disaggregated by geographical area and economic sector of activity.

Melissa CORTELLESSA, melissa.cortellessa@hotmail.it
Cinzia GRAZIANI, Istituto Nazionale di Statistica, cingraziani@istat.it
Andrea SPIZZICHINO, Istituto Nazionale di Statistica, spizzich@istat.it

IL MONITORAGGIO DELL'EFFETTO INTERVISTATORE ATTRAVERSO L'ANALISI MULTILEVEL

Miriam De Santis, Antonella Iorio, Carlo Lucarelli, Alessandro Martini

1. Introduzione

Un approccio volto a migliorare la qualità dei dati raccolti, in un'indagine statistica in cui sono coinvolti gli intervistatori, richiede necessariamente che la gran parte delle strategie messe in atto siano rivolte a loro, a causa delle implicazioni dirette del loro lavoro sui risultati.

L'effetto intervistatore può avere un largo impatto sulla qualità dell'indagine e può essere particolarmente rilevante nella Rilevazione sulle Forze di lavoro, dove gli intervistati sono chiamati a rispondere a domande inerenti dati personali o sensibili o su aspetti caratterizzati da alta desiderabilità sociale. La letteratura relativa a questo argomento mostra che l'effetto intervistatore deve essere considerato come un'importante fonte di errore non campionario nelle indagini dirette (Freeman & Butler 1976, Hanson & Markes 1958, Kish 1962, Pannekoek 1988, Schnell & Kreuter 2005, Tucker 1983). L'effetto intervistatore si riferisce alla variabilità nelle stime attribuibile al fatto che i dati raccolti da uno specifico intervistatore possono essere diversi dai dati raccolti da un altro pur utilizzando gli stessi strumenti di rilevazione. Nei dati sull'indagine delle Forze di lavoro emerge chiaramente una struttura gerarchica: gruppi di famiglie (e individui) sono intervistati dallo stesso intervistatore. Per entrambe le modalità CAPI e CATI troviamo classi di famiglie, e quindi individui, intervistati da un certo intervistatore. Un appropriato gruppo di modelli per analizzare questo tipo di dati è dato dai modelli multilevel che tengono conto della variabilità associata ai diversi intervistatori. Altri lavori (Bocci 2002) con il medesimo approccio metodologico hanno cercato di fornire una valutazione dell'effetto rilevatore rispetto a specifici aspetti delle indagini dirette, come la somministrazione di quesiti sensibili, principalmente nell'ottica di valutare *ex-post* la qualità delle risposte nei dati raccolti.

In questo contributo siamo sempre interessati a fornire una misura sintetica dell'effetto intervistatore per tenerne però conto nel monitoraggio del processo di raccolta dei dati e tentare di definire azioni correttive per ridurre questa fonte di errori.

Vista la rilevanza del ruolo dell'intervistatore e del suo contributo a determinare il livello di qualità dei dati raccolti, indicatori di performance del suo operato devono far necessariamente parte del set di parametri del processo di produzione da tenere sotto controllo e su cui definire obiettivi di qualità misurabili e quindi perseguibili.

A causa dell'indisponibilità delle informazioni individuali dei singoli intervistatori non abbiamo potuto evidenziare quali siano le caratteristiche che possono influenzare

maggiormente le risposte degli intervistati. Abbiamo quindi applicato tali modelli alla fase di raccolta dei dati dell'indagine sulle Forze di lavoro al fine di individuare l'effetto intervistatore sulla qualità dei risultati dell'indagine, e a fornire una misura sintetica di esso.

2. Il Modello Multilevel

Nei dati sull'indagine delle Forze di lavoro diversi individui (unità di primo livello) sono intervistati dallo stesso intervistatore (unità di secondo livello). Quindi si possono distinguere due livelli di analisi: un livello inferiore (o micro-level) per i rispondenti e un livello superiore (o macro-livello) per gli intervistatori.

Un primo requisito per l'applicazione dei modelli multilivello è effettivamente soddisfatto: gli intervistatori possono essere considerati come un campione casuale di una popolazione di potenziali intervistatori non osservati mentre l'allocazione assegnazione delle interviste agli intervistatori è casuale solo per la modalità CATI. Gli intervistatori CAPI, invece, raccolgono i dati su uno specifico territorio che spesso ha una certa influenza sui dati. Abbiamo un unico codice intervistatore per la tecnica CATI nel data base dell'indagine pertanto è impossibile identificare le interviste raccolte dai singoli intervistatori, successive sperimentazioni saranno condotte utilizzando l'archivio dei contatti del *fieldwork* in cui compaiono codici univoci. Abbiamo così condotto questa prima analisi solo sul sotto-campione di interviste CAPI, sempre controllando il modello per territorio e considerando attendibile la stima dell'effetto intervistatore nei casi dove l'effetto sul territorio non fosse significativo.

Inoltre eravamo interessati a monitorare altri parametri rilevanti relativi al processo di raccolta dei dati dell'indagine sulle Forze di lavoro come la durata dell'intervista, che può essere considerata in questo contesto una variabile risposta di tipo continuo.

Sfortunatamente ottenere una stima attendibile di questo parametro dai dati del nostro questionario elettronico non è così facile, così non siamo stati in grado di eseguire questa analisi.

Le variabili risposta considerate nella nostra analisi sono dicotomiche (attivazione di alcuni items critici o filtri del questionario, nuovi items nel questionario, item non-response a specifiche domande, intervista proxy). Il modello che abbiamo scelto per questo tipo di variabili si basa su un modello logit descritto da:

$$y_{ij}|u_j \sim B(1, \pi_{ij}) \tag{1}$$

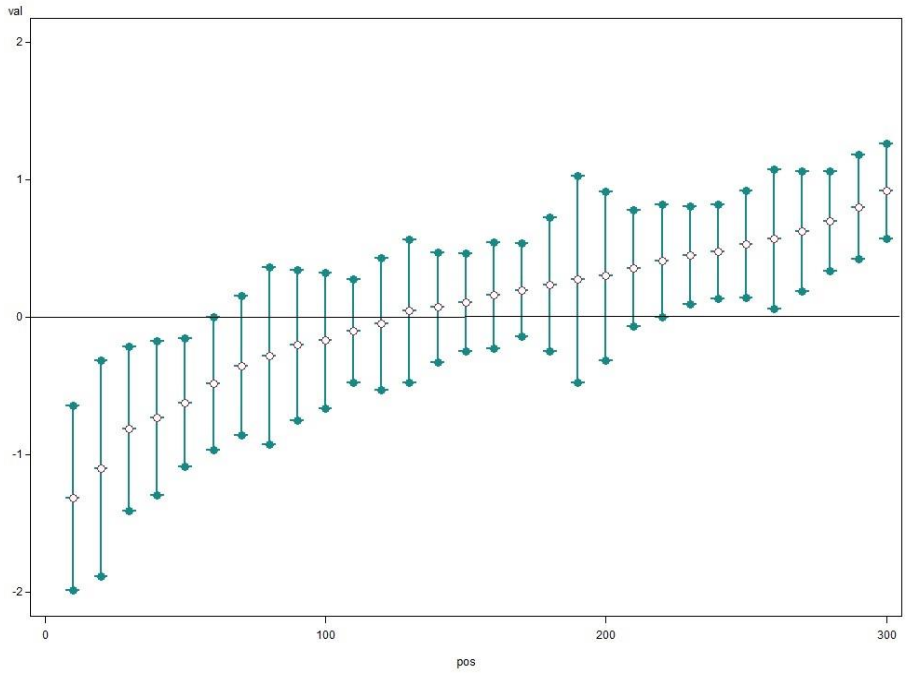
$$g(\pi_{ij}) = \mu + u_j \; with \; u_j \sim N(0, \sigma_u^2) \tag{2}$$

$$\pi_{ij} = \frac{\exp(\mu + u_j)}{1 + \exp(\mu + u_j)} \tag{3}$$

$$\rho = \frac{\sigma_u^2}{(\frac{\pi^2}{3}+\sigma_u^2)} \tag{4}$$

dove π_{ij} denota la probabilità di avere esito 1 per la variabile risposta y_{ij} in prossimità del j-esimo intervistatore, e u_j, che rappresenta la variabilità dovuta all'effetto intervistatore, è una variabile casuale indipendente e identicamente distribuita (iid) secondo una normale $N(0, \sigma_u^2)$. Il coefficiente di correlazione intra-classe ρ (ICC), è stimato come il rapporto tra la varianza all'interno del gruppo e la varianza totale.

Figura1. Predizione Bayesiana empirica per intercette casuali, variabile risposta: intervista Proxy.



Esso misura il grado di omogeneità delle unità appartenenti allo stesso cluster e rappresenta, nel nostro contesto, l'effetto intervistatore. Usiamo un modello a intercetta casuale con lo scopo di ottenere una stima affidabile dell'effetto intervistatore, con l'introduzione di ulteriori variabili esplicative la scomposizione della varianza nella (4) non è più esatta e la quota della variabilità dovuta al secondo livello è imprecisa. Per monitorare il processo di raccolta dei dati è utile una stima sintetica per evidenziare tematiche affette da un significativo effetto intervistatore, ma al fine di apportare correzioni al processo, abbiamo anche bisogno di una valutazione "a livello di singolo intervistatore". Per questo motivo, una volta definito il modello e stimati i parametri, è possibile avere una valutazione a livello di singolo intervistatore usando la predizione Bayesiana empirica per intercette

casuali. Le procedure di previsione forniscono stime di ciascun $\hat{u}_j$, ovvero per ciascun intervistatore, con i relativi intervalli di confidenza, che consentono di analizzare il ranking delle performances degli intervistatori e identificare coloro che hanno stime attese significativamente diverse da zero.

3. Principali risultati

La presenza dell'effetto intervistatore è stata testata su molti aspetti sui dati relativi ai quattro trimestri del 2014. Per gran parte di essi non è emersa in maniera significativa, sottolineando una buona qualità media delle interviste. Un primo risultato che è importante sottolineare è che, coerentemente con la letteratura, considerando il genere dei rispondenti come variabile risposta non emerge un significativo effetto intervistatore. Nell'indagine Forze di lavoro la raccolta dei dati sul genere dell'intervistato non comporta particolari difficoltà o cambiamenti significativi nel condurre l'intervista, per cui si tratta di un risultato ragionevole.

Tabella 1 – *Effetto intervistatore nella fase di raccolta dei dati, Indagine Forze di Lavoro, anno 2014*

	2014							
	Q1		Q2		Q3		Q4	
Target Variable	ICC	$Pct(\hat{u} \neq 0)$	ICC	$Pct(\hat{u} \neq 0)$	ICC	$Pct(\hat{u} \neq 0)$	ICC	$Pct(\hat{u} \neq 0)$
Sesso	0	-	0	-	0	-	0	-
Filtro precedenti esperienze	0,09	19,8	0,09	18,7	0,09	17,6	0,09	19,5
Filtro precedenti esperienze *(controllato per territorio)*	0,06	9,9	0,06	11,6	0,06	11,4	0,06	8,1
Intervista Proxy	0,11	30,6	0,10	25,2	0,11	29,3	0,11	28,3
Nuovi items Livello di istruzione	0,12	27,7	0,11	27,1	0,12	29,6	0,12	28,9
Item NR Reddito	0,41	39,5	0,39	33,9	0,40	33,9	0,43	6,5
Nuovi items Formazione e istruzione	0,47	22,1	0,55	22,1	0,52	19,8	0,57	15,8
Non codifica Attività Economica	0,52	10,5	0,56	10,3	0,55	8,5	0,56	11,0
Non codifica Professione	0,75	14,7	0,88	12,9	0,83	11,1	0,66	9,4

Un primo caso nel quale abbiamo trovato un valore piuttosto basso ma significativo per ρ è l'attivazione dell'item "No" su una precedente esperienza di lavoro: la risposta "No" permette di accorciare la durata dell'intervista (saltare un'intera sottosezione del questionario). Quest'ultimo caso è ovviamente legato alla possibilità di un comportamento

"non corretto" degli intervistatori, ma potrebbe dipendere anche dal territorio, dal momento che questo tipo di fenomeno varia in tutto il paese in modo diverso.

Così abbiamo controllato il modello utilizzando il territorio come variabile esplicativa al primo livello, anche in questo caso la componente di varianza imputabile all'intervistatore è ancora significativa, evidenziando il fatto che alcuni intervistatori potrebbero agire in questo modo. Il coefficiente di correlazione intra-classe ρ è 0,11 per l'intervista proxy, ciò significa che su questo aspetto il comportamento degli intervistatori varia significativamente. La valutazione di questo aspetto è difficile a causa del fatto che gli intervistatori con un basso tasso di interviste proxy non riportano correttamente questa informazione. In questo tipo di analisi questo gruppo di intervistatori può essere identificato considerando $j \in \{\widehat{u_j} < 0\}$. D'altra parte, il gruppo di intervistatori identificato da $j \in \{\widehat{u_j} > 0\}$ dovrebbe migliorare la comunicazione con i rispondenti, al fine di ridurre il tasso di proxy.

Mentre il coefficiente di correlazione intra-classe è utile per segnalare la presenza dell'effetto intervistatore, ai fini del monitoraggio della rilevazione, il tasso di intervistatori con una stima significativamente diversa da 0 risulta essere un indicatore più efficace. In questo caso l'indicatore è di circa il 30% per l'intero periodo, un valore molto alto considerando anche la rilevanza del fenomeno delle interviste proxy, il cui impatto sulla qualità dei dati è ben noto.

Poiché l'indagine si basa su una rilevazione continua i nuovi quesiti o l'adozione di nuove classificazioni vengono introdotti nel primo trimestre di ogni anno. Risulta così indispensabile monitorare l'introduzione di questi cambiamenti, in particolare valutare le competenze degli intervistatori su questi nuovi temi e, se necessario, definire alcuni interventi di formazione continua.

D'altro canto la gestione degli intervistatori è sotto la responsabilità della società appaltante e non è consentito interagire direttamente con loro, possiamo giusto chiarire qualche aspetto critico per e-mail o in specifici de-briefing di formazione. Per questo problema questo metodo sembra fornire risultati interessanti: nel primo trimestre del 2014 secondo la nuova classificazione ISCED le domande sul livello di istruzione e di formazione della popolazione sono state cambiate. Alcuni nuovi item di risposta sono abbastanza simili ai vecchi così può essere difficile identificarli correttamente e può verificarsi un errore di classificazione. Considerando come variabile target una risposta a uno di questi nuovi items è stato possibile monitorare l'effetto intervistatore su questo argomento (o tema). E' significativo per il livello di istruzione (ρ =0,12) ed è molto forte ($\rho \cong 0{,}50\ , Q = 1, \ldots, 4$) per la frequenza a corsi di formazione e istruzione.

In questo caso, la polarizzazione su specifici items di risposta indica un uso improprio del questionario elettronico (con l'aggiunta di nuovi items era necessario scorrere lo schermo per visualizzare tutti gli elementi di risposta). Abbiamo inviato diverse e-mail agli intervistatori per chiarire questo punto e lo abbiamo approfondito in un de-briefing di formazione a luglio 2014, al fine di ridurre l'errore nei dati.

Si noti che l'effetto intervistatore complessivo è stato costante per tutto il periodo considerato, mentre si è registrata una riduzione nel quarto trimestre della percentuale di intervistatori con una stima significativamente diversa da zero, questo indicatore si è ridotto dal 22,1 per cento nel primo trimestre 2014 al 15,8 nell'ultimo trimestre.

Il nostro interesse era principalmente ridurre l'emergenza dei casi problematici, lasciando all'indicatore ICC il compito di informarci (o avvertirci) circa la presenza di tale effetto.

Valori più elevati del coefficiente di correlazione intra-classe testimoniano le difficoltà degli intervistatori a trattare argomenti sensibili, questo succede per l'indagine sulle Forze di lavoro quando si considera l'item "Non risposta" per il reddito come variabile target (o risposta) nel modello multilevel, in questo caso l'effetto intervistatore è elevato e costante per tutto il periodo considerato variando tra 0,39 e 0,43. Questo conferma la grande importanza assunta dalla capacità degli intervistatori di creare un clima di fiducia che aiuti a parlare di argomenti delicati.

I risultati dell'analisi mostrano livelli elevati del coefficiente di correlazione rispetto a due comportamenti rilevanti e scorretti legati agli intervistatori: evitare la codifica del testo per quanto riguarda il settore di attività economica e la professione.

Le informazioni mancanti su queste due variabili saranno trattate successivamente nel controllo e imputazione dei dati ma è comunque una possibile fonte di distorsione.

I valori di ICC sono molto elevati, in particolare considerando l'attitudine del modello logit a sottostimare la correlazione.

Questi risultati indicano che le classificazioni che adottiamo per questi temi non sono così facili da gestire per tutti gli intervistatori, in particolare la nuova classificazione ISCO08 introdotta nella nostra indagine nel primo trimestre 2014. Un risultato leggermente migliore è mostrato per la classificazione NACE, probabilmente perché questo aspetto è stato meglio chiarito negli incontri di formazione precedenti. Una indicazione che tali risultati suggeriscono è che nel prossimo intervento formativo dovrebbe essere maggiormente approfondita la classificazione ISCO, che sembra essere un aspetto critico per la gestione dell'intervista da parte degli intervistatori.

Conclusioni

I risultati presentati in questo lavoro devono essere considerati solo come una prima applicazione sperimentale di questi metodi nell'indagine sulle Forze di lavoro. Tuttavia, i risultati dell'analisi forniscono un punto di partenza utile per migliorare la qualità della rilevazione attraverso il controllo degli errori non campionari e la progettazione di un processo di monitoraggio più affidabile e integrato. Questi metodi possono integrare il sistema di monitoraggio dell'indagine Forze di Lavoro per esempio considerando come variabili risposta gli indicatori e i tassi definiti secondo gli standard internazionali suggeriti dalla AAPOR ed effettuando un'analisi simile anche per la modalità CATI sull'archivio contatti.

Variabili esplicative di primo e secondo livello (livello di istruzione degli intervistatori, esperienza di lavoro, ecc.) potrebbero essere aggiunte nell' applicazione di modelli multilevel, al fine di studiare come le caratteristiche degli intervistatori possono influire sulle risposte degli intervistati o interazioni intervistatore- rispondenti. Attualmente tali informazioni non sono disponibili per motivi di tutela della privacy degli intervistatori, nel

caso dovessero essere disponibili in futuro potrebbero essere considerate in tale tipo di analisi.

I risultati di questo differente approccio potrebbero essere utilizzati come base informativa per l'identificazione dei profili più appropriati degli intervistatori da reclutare, se questo aspetto tornerà sotto la responsabilità Istat, come nel passato.

Riferimenti bibliografici

MURATORE M.G., SIGNORE M. 2005. Il monitoraggio del processo e la stima dell'errore nelle indagini telefoniche, *Metodi e Norme*, 25, ISTAT, Roma.

BOCCI L., MURATORE M.G., SIGNORE M., TAGLIACOZZO G. 2002, The Interviewer Effect on the Data Collection of Sensitive Questions, *SIS Scientific Meeting Proceedings*, Milan, 5-7th June 2002.

CORSETTI G., GIAMMATTEO M., MARTINI A. 2010. Monitoring process and non-sampling errors control in PLUS sample survey, *Q2010 Meeting Proceedings*, Helsinki, 4-6th May 2010.

HOX J.J., DE LEEUW E.D., KREFT I.G. 1991, The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In BIEMER, GROVES, LYBERG, MATHIOWETZ, SUNDMAN, (Eds.) *Measurement Errors in Surveys*, pp.339-461.

KISH L. 1962, "Studies of interviewer variance for attitudinal variables," *Journal of the American Statistical Association*, 57(297), pp. 92-115.

PANNEKOEK J., 1988, Interviewer variance in a telephone survey, *Journal of Official Statistics* 4, pp. 375-384.

TUCKER C. 1983, Interviewer effects in telephone surveys, *The Public Opinion Quarterly 47(1)*, pp. 84-95.

SUMMARY
Monitoring the interviewer effect in data collection through a multilevel analysis

Interviewer effects can have a wide impact on quality of survey and may be particularly relevant in LFS, where respondents are likely to be queried about sensitive or socially desirable responding topics. A review of the related literature shows that inter viewer effects must be regarded as an important source of error in sample surveys (Freeman & Butler 1976, Hanson & Markes 1958, Kish 1962, Pannekoek 1988, Schnell & Kreuter 2005, Tucker 1983). Interviewer error refers to variance in survey estimates that arises from the fact that data collected by either a specific individual interviewer or a specific set of interviewers may be different than data collected by another individual or set of interviewers administering the same questionnaire.

In Lfs data a hierarchical structure clearly arises: respondents are nested with interviewers, both in the CAPI and in the CATI mode we find cluster of households (and individuals) interviewed by a certain interviewer.

An appropriate class of models for analyzing this kind of data is given by the multilevel modeling approach that take into account the amount of variance derived from differences across interviewers.

In this paper we are not interested to investigate in depth how interviewer characteristics may affect respondents answers as well as interviewer-respondent interactions, also because additional information about interviewers are not available. On the other side, we are interested in providing a consistent measure of the "overall" interviewers effect, to be consequently monitored during the data collection phase, and in what can be done in order to reduce this source of bias.

We started applying them for Ita-Lfs data collection phase in order to identify interviewer effects on the survey results quality, and providing an estimate of it.

Response variables we take into account in our analysis are some key indicators of quality for a sample survey on households that it is important to monitor in a continuous survey like Lfs.

So we considered some target variables that can be:

- continuous (interview duration);
- dichotomous (activation of some critical item or questionnaire filters, new items in the questionnaire, "Don't Know- No Answer " to specific items, Proxy interviewing).

Results of these monitoring can be useful to point out interviewers with an improvable performance. These interviewers can be invited to take part in focus training courses on these specific topics in order to improve the quality of data we collect trying to reduce this variability.

Miriam DE SANTIS, Istat, mdesantis@istat.it
Antonella IORIO, Istat, iorio@istat.it
Carlo LUCARELLI, Istat, calucare@istat.it
Alessandro MARTINI, Istat, alemartini@istat.it

MISMATCH TRA DATI AMMINISTRATIVI E DI INDAGINE: *L'ESPERIENZA ISTAT-INAIL*

Miriam De Santis, Antonio R. Discenza, Antonella Iorio, Carlo Lucarelli

1. Introduzione

L'attenzione verso l'utilizzo integrato di dati statistici provenienti da fonti diverse è sempre più elevata. In particolare, l'utilizzo a fini statistici di dati di fonte amministrativa ha riscosso negli ultimi anni un crescente interesse e subito un sempre maggiore sfruttamento.

L'obiettivo del nostro lavoro è valutare punti di forza e fragilità dell'integrazione tra fonti statistiche diverse che riferiscono dello stesso fenomeno. Per fonte amministrativa intendiamo una "*…collections of data held by other parts of government, collected and used for the purposes of administering taxes, benefits or services.*" (Unece, 2011). Essa è rilevante se copre l'intera popolazione interessata dal fenomeno, perciò maggiori sono le lacune nella copertura, maggiore sarà la distorsione nelle stime (Wallgren 2014).

Inoltre, una fonte amministrativa non è di per sé utilizzabile a fini statistici senza un trattamento che la renda adeguata a tale impiego. Di conseguenza un dato amministrativo adattato ad un uso statistico diventa di per sé soggetto agli stessi criteri di valutazione della qualità del dato statistico: rilevanza e completezza, tempestività e puntualità, precisione, confrontabilità e coerenza, accessibilità e chiarezza, costo contenuto e il più basso carico di risposta (Unece 2007).

2. Definizione del fenomeno

Il nostro studio si fonda su un esercizio di integrazione di dati di indagine provenienti da ISTAT e dati amministrativi di fonte INAIL riguardante il tema della salute e sicurezza sul lavoro (Figura 1). In particolare, la Rilevazione sulle Forze di Lavoro ISTAT (RFL) ha previsto nel secondo trimestre del 2013 l'inserimento di un modulo supplementare riguardante salute e sicurezza sul lavoro che, considerando aspetti connessi agli infortuni, ha permesso di indagare su un tema che a livello nazionale viene trattato anche da INAIL, organismo istituzionale

che diffonde statistiche sullo stesso argomento basate su dati provenienti dai propri archivi.

Figura 1 – *Schema di Integrazione tra dati amministrativi e dati di indagine.*

L'esistenza di una duplice fonte di dati sul medesimo oggetto d'interesse e la necessità di implementare un modulo incentrato appunto sul tema, ha creato i presupposti per valutare l'integrazione tra i dati ISTAT e INAIL seguendo un'ottica bidirezionale.

Nell'ottica <Fonte amministrativa – Indagine> l'integrazione si è concretizzata nella realizzazione di un'indagine pilota nel 2012, svolta per testare il modulo da inserire nella RFL nell'anno successivo, che è a sua volta stato opportunamente disegnato anche al fine di conseguire un riscontro su dati di fonte amministrativa nell'ottica <Indagine - Fonte amministrativa>.

3. Integrazione «Fonte amministrativa – Indagine»: l'indagine pilota

La valutazione dell'integrazione <Fonte amministrativa – Indagine> è avvenuta attraverso un'analisi condotta sui dati dell'indagine pilota.

L'indagine pilota è stata effettuata a dicembre 2012. Il campione, al quale è stata somministrata, era costituito per metà da famiglie in cui almeno uno dei componenti aveva fatto una denuncia di infortunio o malattia professionale all'INAIL (500 famiglie); a tal fine INAIL ha estratto dai suoi archivi un elenco di nominativi di persone vittime di infortuni sul lavoro nei 12 mesi precedenti[1], dei quali ne sono stati intervistati 449 con le rispettive famiglie. L'altra metà del campione comprendeva famiglie facenti parte del campione selezionato per la RFL nel terzo trimestre 2012 (500) interessate dall'ultima occasione di intervista[2]. Si tratta perciò di un campione ragionato, in parte fortemente coinvolto nel fenomeno infortunistico e appositamente predisposto per testare l'impianto del modulo ad hoc. In definitiva, sono state intervistate 949 famiglie per un totale di 1763 individui.

Il questionario è stato somministrato agli infortunati e a tutti i componenti delle rispettive famiglie per quanto riguarda il campione INAIL; allo stesso modo hanno risposto alle medesime domande tutti i componenti delle famiglie relative all'altra parte del campione proveniente dalla RFL. In particolare, il modulo si articola in tre sezioni: infortuni sul lavoro, problemi di salute lavoro-correlati, esposizione a fattori di rischio per la salute fisica e psicologica. In questo studio ci si concentra esclusivamente sulla sezione relativa agli infortuni sul lavoro.

Limitandoci ad analizzare esclusivamente le 449 Famiglie intervistate sulla base del campione proveniente dagli archivi INAIL, per 19 famiglie non è stato possibile individuare con certezza l'individuo che aveva fatto la denuncia di infortunio all'INAIL, per cui ne sono state agganciate 430. Dopo l'aggancio 5 individui risultavano non avere le caratteristiche per accedere alla sezione sugli infortuni (cioè hanno dichiarato di non essere occupati e non aver svolto un lavoro negli ultimi 12 mesi). Dunque, l'analisi è stata condotta su 425 individui.

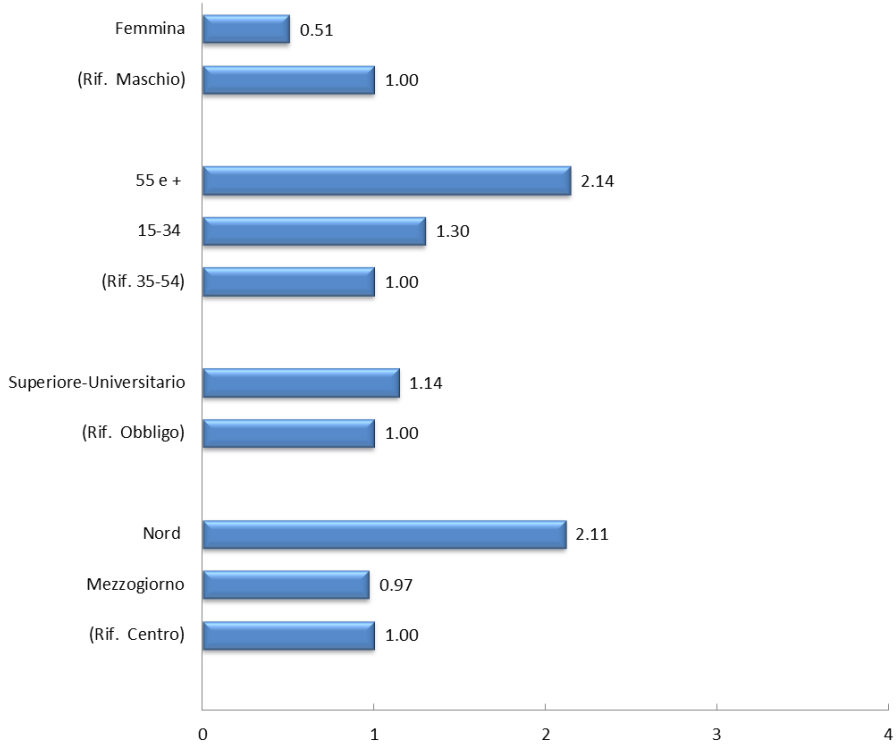
I risultati hanno mostrato che un consistente gruppo di individui, pari a un quarto, nonostante abbia denunciato all' INAIL un infortunio accaduto negli ultimi 12 mesi, alla relativa domanda presente nel modulo nega di averlo subito. Le caratteristiche di questi infortunati INAIL che non risultano all'indagine sono state analizzate attraverso un modello log-lineare in cui è stata utilizzata come dipendente una variabile dicotomica che assume valore 0 se l'individuo dichiara di aver subito l'infortunio e 1 se non lo dichiara e in cui sono state utilizzate come indipendenti una serie di variabili socio-demografiche quali sesso, età in classi, titolo di studio, ripartizione geografica e alcune variabili legate alle caratteristiche

[1] Il campione estratto da INAIL consta di 10800 nominativi dei quali per circa 3000 è stato possibile recuperare un numero di telefono fisso.

[2] Per la struttura del disegno campionario della rilevazione sulle Forze di Lavoro e le modalità di intervista delle famiglie si veda ISTAT 2006.

lavorative quali posizione, durata contratto, settore di attività economica e professione (Fig. 2 e 3). Nei risultati sono evidenziati solo gli odds ratio che hanno un livello di significatività pari al 90%.

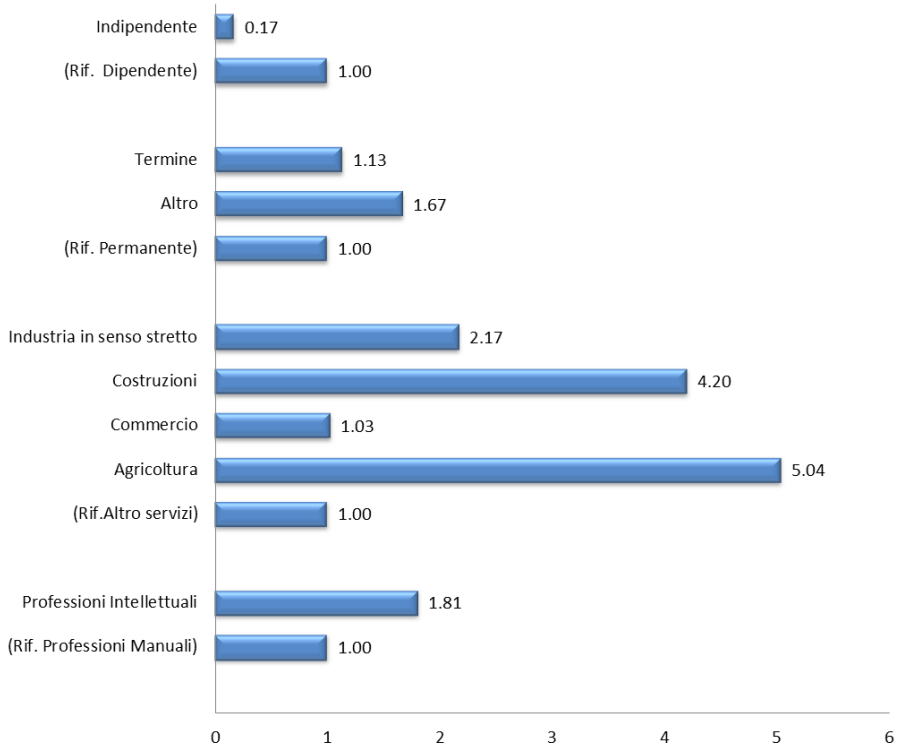
Figura 2 – *Odds ratio (OR) relativi alle caratteristiche demografiche degli individui infortunati INAIL che hanno dichiarato all'indagine di non aver subito infortuni. Modello di regressione logistico. Rilevazione sulle Forze di Lavoro, Italia, secondo trimestre 2013.*



Chi non dichiara di aver subito l'infortunio è prevalentemente uomo, nelle classi di età più anziane (55 e oltre) e risiede al Nord.

Dal punto di vista lavorativo chi è reticente ha principalmente un lavoro alle dipendenze (senza differenze per tipologia contrattuale permanente/a termine), lavora in settori manifatturieri come industria e costruzioni ma anche in agricoltura e svolge perlopiù lavori di concetto (professioni intellettuali-impiegatizie) piuttosto che manuali.

Figura 3 – *Odds ratio (OR) relativi alle caratteristiche lavorative degli individui infortunati INAIL che hanno dichiarato all'indagine di non aver subito infortuni. Modello di regressione logistico. Rilevazione sulle Forze di Lavoro, Italia, secondo trimestre 2013.*



4. Integrazione «Indagine - Fonte amministrativa»: Il Modulo ad hoc nell'Indagine a regime

L'esame dell'integrazione tra fonti <Indagine - Fonte amministrativa> è stata effettuata analizzando i dati provenienti dal modulo ad hoc inserito all'interno dell'indagine a regime nel secondo trimestre del 2013.

Il campione usato per il modulo ad hoc è lo stesso della parte standard della RFL del secondo trimestre 2013 dove sono state intervistate 65.577 famiglie per un totale di 153.317 individui di cui 133.024 con età di 15 anni o più. In particolare, alle domande sugli infortuni accedevano poi le persone occupate o con

un'esperienza di lavoro negli ultimi 12 mesi e queste erano pari a 57.846[3] (il 43,5% delle persone di 15 anni e più). La struttura del modulo segue sostanzialmente quella della pilota ma nel modulo a regime, a chi ha dichiarato di aver subito un infortunio, è stata sottoposta una domanda aggiuntiva volta ad indagare se per questo infortunio era stata fatta denuncia all'INAIL.

Dalle stime ottenute è emerso che circa 705 mila lavoratori hanno dichiarato di aver subito infortuni nell'ultimo anno di cui circa 214 mila (il 30,3%) dichiara di non averne fatto denuncia all'INAIL. Tra chi non denuncia, il 37,5% (11,3% su tutti gli infortuni) dichiara che l'infortunio ha comportato giorni di assenza dal lavoro[4]. In particolare, tra gli infortunati che hanno dichiarato di non aver fatto denuncia all'INAIL, il 37,8% dichiara infortuni occorsi nel tragitto casa-lavoro-casa, per i quali non è stato possibile stabilire la durata dell'assenza dal lavoro, mentre il 24,7% riferisce infortuni di piccola entità che non hanno comportato alcuna assenza.

Prendendo in considerazione gli infortunati, che hanno dichiarato che l'infortunio ha comportato giorni di assenza dal lavoro, è opportuno specificare che nello stesso aggregato sono stati considerati anche coloro che al momento dell'intervista non erano più occupati ma che avevano un'occupazione negli ultimi 12 mesi (il 7,7% del totale) nella svolgimento della quale sono state vittime di infortunio. In tal caso le caratteristiche lavorative di questi ultimi sono state integrate con quelle degli occupati al fine di tracciare un profilo omogeneo. Lo stesso è stato fatto per la popolazione di riferimento che consiste negli occupati e i non occupati che avevano un'occupazione negli ultimi 12 mesi (10,2% della popolazione di riferimento).

Il confronto tra l'aggregato in esame e la sua popolazione di riferimento (utilizzata in questo caso come popolazione standard) ci permette di evidenziare quali sono le caratteristiche più influenti nella determinazione del fenomeno. Dall'analisi delle caratteristiche di chi ha dichiarato di non aver fatto denuncia all'INAIL dell'infortunio accaduto che ha comportato giorni di assenza, emerge che il fenomeno coinvolge in misura maggiore gli uomini, i residenti nelle regioni

---

[3] Il tasso di risposta è pari all'87,8% e viene calcolato nel seguente modo:

$$tasso\ di\ risposta = \frac{famiglie\ rispondenti}{famiglie\ eleggibili}$$

dove il numero delle famiglie eleggibili si compone delle famiglie che rispondono e quelle che non rispondono. Dunque, il 12,2% delle famiglie eleggibili che non rispondono è così composto: il 4,0% a causa di rifiuti, il 7,3% per mancati contatti e il restante 0,9% per altri motivi.

[4] E' evidente che una stima più accurata si potrebbe ottenere da un ulteriore aggancio tra i dati di indagine Istat (chi ha dichiarato di aver subito infortuni nell'ultimo anno) e gli archivi INAIL per avere un riscontro oggettivo del fenomeno e dell'eventuale mismatching. Di concerto con i ricercatori INAIL, si valuterà la possibilità di procedere in questa direzione.

del Nord, gli stranieri e i meno scolarizzati, mentre non c'è differenza per età (Tabella 1).

Tabella 1 – *Caratteristiche demografiche di occupati e non occupati che hanno subito infortuni sul lavoro negli ultimi 12 mesi, che hanno comportato giorni di assenza dal lavoro e non hanno fatto denuncia all'INAIL. Rilevazione sulle Forze di Lavoro, Italia, secondo trimestre 2013.*

Caratteristiche	Occupati e Non occupati che hanno subito infortuni sul lavoro negli ultimi 12 mesi che hanno comportato giorni di assenza dal lavoro e non hanno fatto denuncia all'INAIL (A)	Occupati e Non occupati che hanno lasciato il lavoro negli ultimi 12 mesi (B)	Diff A-B
Genere			
Maschio	**75.9**	**57.7**	**18.1**
Femmina	24.1	42.3	-18.1
Ripartizione geografica residenza			
Nord	**58.8**	**50.7**	**8.1**
Centro	19.0	21.0	-2.0
Mezzogiorno	22.2	28.2	-6.1
Cittadinanza			
Italiano	82.3	89.7	-7.4
Straniero UE	**4.5**	**3.3**	**1.2**
Straniero NON UE	**13.2**	**7.0**	**6.2**
Età			
15-34 anni	26.0	25.2	0.9
35-54 anni	58.9	58.2	0.7
55 anni e più	15.1	16.6	-1.6
Titolo di Studio			
Obbligo	**51.4**	**34.7**	**16.7**
Superiore	37.7	46.4	-8.7
Universitario	10.9	18.9	-8.0

Per quanto riguarda le caratteristiche lavorative, le professioni maggiormente coinvolte nella mancata denuncia sono quelle manuali (le professionalizzate in misura maggiore), i lavoratori nel settore delle costruzioni, i lavoratori indipendenti. Seppure nel lavoro alle dipendenze è elevata nel complesso la propensione alla denuncia, questa avviene in modo più consistente per i lavoratori a tempo indeterminato rispetto ai loro colleghi a termine (Tabella 2).

Tabella 2 – *Caratteristiche lavorative di occupati e non occupati che hanno subito infortuni sul lavoro negli ultimi 12 mesi che hanno comportato giorni di assenza dal lavoro e non hanno fatto denuncia all'Inail***.** *Rilevazione sulle Forze di Lavoro, Italia, secondo trimestre 2013.*

Caratteristiche	Occupati e Non occupati che hanno subito infortuni sul lavoro negli ultimi 12 mesi che hanno comportato giorni di assenza dal lavoro e non hanno fatto denuncia all'INAIL (A)	Occupati e Non occupati che hanno lasciato il lavoro negli ultimi 12 mesi (B)	Diff A-B
Professioni			
Intellettuali	9.1	15.6	-6.5
Di concetto	15.4	28.2	-12.9
Manuali professionalizzate	47.1	35.7	11.3
Manuali non professionalizzate	22.9	19.4	3.4
Forze armate	5.6	1.0	4.6
Attività Economica			
Agricoltura	5.8	4.1	1.7
Industria in senso stretto	17.1	19.3	-2.2
Costruzioni	16.9	7.6	9.3
Commercio	15.0	14.8	0.2
Altre attività dei servizi	45.2	54.2	-9.1
Dipendenti - Indipendenti attività principale			
Dipendente	67.0	76.0	-9.0
Indipendente	33.0	24.0	9.0
Lavoro a tempo determinato o indeterminato (per i dipendenti)			
Tempo determinato	12.6	13.7	-1.2
Tempo indeterminato	54.4	61.8	-7.4

5. Conclusioni e considerazioni

I risultati di questo esercizio mostrano un evidente mismatching tra fonti diverse che indagano lo stesso fenomeno: circa il 25% di chi proviene dalla Fonte amministrativa non emerge nella fonte Indagine; viceversa oltre l'11% di chi si manifesta nell'Indagine non è presente nelle Fonti amministrative. Se da un lato può essere abbastanza intuitivo che chi dichiara di aver subito un infortunio in un'indagine non abbia proceduto alla sua "ufficializzazione" con una denuncia che contempla iter amministrativi lunghi e un confronto con la burocrazia che può risultare estenuante, dall'altro risulta pressoché inesplicabile il motivo che non fa emergere in un'indagine ciò che è già stato formalizzato in processi documentali più strutturati. Su questo fronte è necessario fare dovuti approfondimenti sugli

atteggiamenti di una popolazione che è dopotutto chiamata alle sue responsabilità di produttore di base (anche se inconsapevolmente) di statistiche ufficiali. E' vero però che chi produce statistiche debba anche chiedersi sempre che cosa queste esprimono, cosa misurano e cosa può nascondersi nelle pieghe di una definizione mai esaustiva per quanto pur essa possa essere dettagliata e puntuale.

Inoltre, ciò deve indurre a considerare in modo critico l'eventualità di sostituzione tra fonti statistiche diverse, anche se questo non significa che non si debba sostenere la necessità di rafforzarne l'integrazione, analizzandone in modo scrupoloso caratteristiche e modalità. E' per questo motivo che l'utilizzo delle diverse fonti statistiche, e la loro eventuale integrazione, non può prescindere dal considerare e valutare una serie di elementi come l'esatta definizione del fenomeno, la tempestività e la cadenza, la copertura che le fonti forniscono, il contesto territoriale o, in senso più lato, dimensionale in cui esso si muove. In altre parole è necessario esercitare sempre una grande consapevolezza nell'uso dei dati e delle fonti.

Ringraziamenti

Questo lavoro è stato possibile anche grazie al fondamentale supporto della dott.ssa Liana Veronico della Consulenza Statistico Attuariale dell'INAIL.

Riferimenti bibliografici

ISTAT. 2006. *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*. Roma: Metodi e Norme n. 32.

Unece. 2011. *Using administrative and secondary sources for official statitics*. New York: United Nation.

Unece. 2007. *Register-based statistics in the Nordic countries*. New York: United Nation.

Wallgren A., Wallgren B., 2014. *Register-based statistics: statistical methods for administrative data*. UK: John Wiley & Sons, Ltd.

SUMMARY

Mismatch between administrative and survey data: the ISTAT - INAIL experience

The focus on the integrated use of data from different sources is increasingly high. In particular , the exploitation for statistical purposes of administrative data in recent years has been affected by a growing interest and has undergone a greater impulse.

In the second quarter 2013, Labour Force Survey (LFS) has foreseen the inclusion of an ad hoc module concerning health and safety at work. In particular, the module refers on occupational injuries and it made possible investigate a subject which is also covered by INAIL (National Institute for Insurance against Accidents at Work), an institutional body which deliver statistics on the same subject based on its archives.

The existence of a dual source of data on the same object of interest has created the conditions for interaction between ISTAT and INAIL in carrying out a pilot survey in 2012, in order to test the module - also appropriately designed to get feedback on administrative data - to be included in the LFS in the following year.

The results show an incomplete adhesion between the different sources that investigate the same topic; rather, the mismatching stood at very high levels both from a verse and on the other. This should prompt us to consider critically the possibility of substitution between different statistical sources and, at the same time, support the need to strengthen integration, analyzing deeply features and modes.

Miriam DE SANTIS, Istituto nazionale di statistica, mdesantis@istat.it
Antonio R. DISCENZA, Istituto nazionale di statistica, discenza@istat.it
Antonella IORIO, Istituto nazionale di statistica, iorio@istat.it
Carlo LUCARELLI, Istituto nazionale di statistica, calucare@istat.it

FORMAZIONE CONTINUA PER IL CENSIMENTO PERMANENTE

Antonella Bianchino, Giulia De Candia, Stefania Taralli

Introduzione

Dopo quindici censimenti della popolazione effettuati con cadenza decennale, a breve anche in Italia si avvia il primo censimento permanente, basato su un ampio uso di fonti statistiche e amministrative integrate da indagini dirette, che consentono di verificare la copertura anagrafica di ciascun comune e di stimare le informazioni socio-economiche su individui, famiglie e abitazioni, in modo da soddisfare le esigenze informative locali, nazionali e internazionali con riferimento a domini territoriali molto fini (almeno comunali). Nel mutato assetto metodologico la rilevazione censuaria diventa un'indagine continua, in grado di fornire dati con maggiore frequenza e tempestività ad un minor costo (Istat, 2014).

Il censimento permanente rappresenta una grande opportunità di potenziamento della statistica ufficiale e, per le numerose innovazioni organizzative e tecnico-metodologiche che lo caratterizzano, pone nuove sfide e nuove esigenze formative. Contenuti aggiornati, un'organizzazione radicalmente mutata, nuove tecnologie a supporto della rilevazione richiedono una formazione accurata e capillare, che accompagni con continuità la rilevazione censuaria assicurando standard di competenze adeguati e omogenei per tutti gli operatori che con diverse funzioni partecipano al processo.

La formazione è la principale strategia di prevenzione dell'errore non campionario: ogni soggetto impegnato nella rilevazione introduce nei dati la propria componente d'errore, che dipende dal grado di conoscenza delle norme e delle tecniche di rilevazione, dalla comprensione e corretta applicazione delle classificazioni e delle definizioni adottate, dall'atteggiamento verso le finalità e l'oggetto della ricerca. Nelle indagini dirette infatti gli operatori che interagiscono con i rispondenti, comunicano l'immagine degli Enti titolari della rilevazione contribuendo a costruirne la *reputation* (De Candia, 2011). Accanto al "sapere", anche il "saper fare" e il "saper essere" sono, perciò, fondamentali per assicurare la

corretta applicazione dei principi deontologici e di qualità che connotano la statistica ufficiale e che sono riportati nel Codice italiano delle statistiche ufficiali[1].

Nel caso delle indagini continue l'investimento in formazione risulta ancor più significativo e strategico; per il prossimo censimento permanente l'apprendimento continuo diventa uno strumento essenziale per affrontare il cambio di paradigma imposto dal nuovo assetto della rilevazione.

1. L'esperienza dei censimenti 2010-2011

Grazie alle opportunità offerte dalle nuove tecnologie, la tornata censuaria 2010-2011 ha visto l'introduzione da parte dell'Istat di innovazioni di processo e di prodotto, che hanno consentito di rafforzare la formazione erogata agli operatori delle reti di rilevazione, superando i noti limiti della formazione iniziale in presenza.

Infatti, sia nei censimenti economici che in quello della popolazione si sono attuate strategie formative di tipo misto (blended learning), basate su corsi in presenza per tutti gli operatori censuari, assicurati attraverso il meccanismo della formazione a cascata, e su prodotti e servizi integrativi, fruibili on line attraverso una piattaforma di e-learning. Le infrastrutture tecnologiche di base utilizzate per l'erogazione sono di tipo open source (Dokeos e Moodle). Attraverso la piattaforma di e-learning sono stati messi a disposizione varie tipologie di materiali: moduli didattici di autoformazione su competenze specifiche e trasversali; questionari e guide ipertestuali; test di autovalutazione; video-tutorial illustrativi degli ambienti di lavoro e delle applicazioni tecnologiche; repository e linkografia di supporto all'auto-istruzione e alla formazione d'aula.

Nell'ambito dei censimenti dell'Industria, del Non-Profit e delle Istituzioni pubbliche, l'e-learning è stato utilizzato anche per finalità di informazione e supporto ai rispondenti. Nei primi due casi con la pubblicazione sul portale del censimento di questionari ipertestuali e video-tutorial per i rispondenti, nel terzo con l'allestimento di un corso a distanza per i rispondenti istituzionali e con l'integrazione dei principali metadati di ausilio alla compilazione direttamente nel questionario web (Istat, 2015).

Per l'Istat l'e-learning ha rappresentato un'importante opportunità per realizzare interventi formativi più efficaci, completi ed incisivi anche grazie all'ampliamento dell'offerta formativa, che è risultata più capillare, disponibile con continuità nel corso di tutta la rilevazione e più articolata sia dal punto di vista dei contenuti che dal punto di vista dei servizi e prodotti offerti (Bianchino, De Candia, Taralli, 2011).

[1] Direttiva n. 10/Comstat, G.U. n. 240 del 13/10/2010

L'approccio didattico utilizzato per la formazione a distanza è stato prevalentemente di tipo individuale, attraverso la fruizione dei contenuti disponibili sulla piattaforma di erogazione.

Le indagini di valutazione svolte presso gli operatori censuari (Istat, 2013 e 2015) hanno evidenziato un generale apprezzamento per l'e-learning. Particolare successo hanno avuto alcuni supporti e servizi originali rispetto al materiale illustrativo tradizionalmente diffuso dall'Istat, e innovativi rispetto all'approccio formativo di tipo tradizionale. Tra questi si citano i moduli didattici di "formazione al ruolo" per gli operatori con funzioni di formatori e per i coordinatori della rete di rilevazione, i questionari e le guide ipertestuali, i test di autovalutazione, che alcuni uffici comunali di censimento hanno utilizzato anche per effettuare le selezioni dei rilevatori al termine del percorso formativo.

I questionari e le guide ipertestuali, utilizzati sia per l'auto-apprendimento che per la formazione in aula, consentono di individuare e visualizzare con facilità definizioni, classificazioni, regole di compilazione e normativa di interesse. Offrono un accesso unico e guidato alla consultazione selettiva e mirata della documentazione di indagine valorizzando le possibilità offerte dalla comunicazione visiva. I test di autovalutazione a verifica immediata, implementati su piattaforma LCMS (Learning Content Management System), consentono agli utenti di consolidare le conoscenze e competenze acquisite tramite il percorso di formazione e ai formatori di verificare il grado di apprendimento della classe.

Per i formatori e per i responsabili della rilevazione l'e-learning ha avuto il vantaggio di ridurre l'onere della formazione aggiuntiva in itinere (aggiornamenti, approfondimenti, gestione del turnover) e ha offerto la possibilità di tracciare percorsi di apprendimento personalizzati per diverse tipologie di utenti, di monitorare la fruizione dei contenuti formativi on line e di verificare i livelli di apprendimento. I report di monitoraggio e valutazione prodotti dalla piattaforma di e-learning hanno consentito agli operatori responsabili del coordinamento della rilevazione ai vari livelli, di individuare le criticità e di predisporre i necessari interventi correttivi.

Per l'Istat l'e-learning ha rappresentato l'opportunità di conseguire una maggiore standardizzazione e capillarità della formazione erogata, ma anche la possibilità di rilasciare aggiornamenti o formazione aggiuntiva in itinere più a ridosso delle varie fasi della rilevazione.

In tutte le rilevazioni censuarie della stagione passata l'e-learning è stato utilizzato come strategia formativa di sostegno ed accompagnamento ai tradizionali corsi in presenza, che restano comunque momenti importanti per la creazione della comunità professionale e per generare fra gli operatori della rete l'identità di scopo. Tuttavia, anche per motivi contingenti e vincoli operativi, non si è realizzata una piena integrazione dell'e-learning nel processo di rilevazione. Di conseguenza,

nonostante i buoni livelli di fruizione raggiunti in alcuni casi e gli apprezzamenti espressi dagli utenti nelle indagini di valutazione, l'e-learning per i censimenti non ha sempre espresso pienamente il suo potenziale.

Tra i punti deboli delle esperienze illustrate va annoverato il problema della tempestività: l'e-learning è stato progettato e realizzato a valle della validazione della documentazione tecnica e illustrativa delle rilevazioni, troppo a ridosso e talora in ritardo rispetto all'avvio della formazione in presenza. Inoltre, i piani di censimento hanno previsto come obbligatoria per gli operatori la sola formazione d'aula, mentre l'e-learning è stato proposto dall'Istat come opportunità ulteriore. Inoltre, sarebbe stato utile prevedere alcune misure di accompagnamento e attuare strategie di diffusione e trasferimento più efficaci.

2. Una proposta per la formazione permanente del censimento permanente

Il censimento permanente prevede gradi di coinvolgimento variabili, strategie di rilevazione diversificate e un calendario delle attività differente per diversi gruppi di Comuni, e quindi per i diversi nodi della rete. Si allarga la platea dei destinatari della formazione e si differenziano i profili: oltre agli operatori statistici (addetti agli uffici di statistica - responsabile e altri operatori), e ai rilevatori impegnati nelle rilevazioni sul campo, un ruolo cruciale è svolto dagli operatori anagrafici (addetti degli uffici anagrafe - responsabile e altri operatori). Ciò implica un maggior grado di differenziazione sia dei fabbisogni formativi che della propensione all'investimento in formazione da parte degli operatori stessi. Inoltre, la continuità o la maggiore frequenza di coinvolgimento prospettano da un lato l'opportunità di conseguire nel tempo un certo grado di specializzazione degli operatori, dall'altro il rischio di un incremento della domanda di formazione che sarà importante fronteggiare assicurando una risposta adeguata e la sostenibilità organizzativa e finanziaria del processo.

La strategia formativa che si propone nel presente lavoro prevede un approccio misto e integrato, con un'alternanza fra momenti d'aula e momenti a distanza fra loro strettamente coordinati e connessi (blended learning). In linea di massima, i momenti in presenza sono in apertura (presentazione e lancio) e in chiusura del corso; a questi si possono aggiungere, ulteriori incontri a scopo formativo o di istruzione tecnica (debriefing, seminari specifici, etc.) o di valutazione (counseling e valutazione dell'andamento del corso).

Per l'Istat gli incontri in presenza hanno la funzione di: conoscere bene i partecipanti, e capire il loro livello di motivazione, le aspettative e i bisogni; creare coinvolgimento nei corsisti e nei docenti, stabilire il cosiddetto "patto d'aula"; migliorare il monitoraggio complessivo dell'efficacia dell'intervento (poiché i

discenti riescono a comunicare più facilmente le loro difficoltà). La formazione a distanza sostiene il processo formativo per l'intera durata delle operazioni e migliora l'offerta formativa, grazie alla possibilità di modificare e distribuire i contenuti formativi in itinere.

Le considerazioni svolte e le precedenti esperienze di e-learning inducono a definire percorsi di apprendimento modulari e individualizzati per:

- figure professionali (operatori statistici, rilevatori, operatori anagrafici)
- profili-utenti ("operatori nuovi" e "operatori esperti").

Nell'anno di avvio delle attività censuarie, i percorsi differenziati per figura professionale sono particolarmente dettagliati, per consentire la formazione al ruolo e la trattazione di tutti gli aspetti tecnico-metodologici e organizzativi del censimento permanente. Negli anni successivi, oltre alla differenziazione per figura professionale si distingue per profilo di utenza: operatore esperto, che ha già frequentato con successo il corso introduttivo e operatore di nuovo inserimento. Gli operatori esperti seguono un corso mirato in cui è possibile affrontare problematiche locali (una sorta di corso di aggiornamento) mentre gli operatori di nuovo inserimento partecipano al corso introduttivo.

Per tutti gli operatori censuari è fondamentale che il processo formativo contribuisca a formare un senso identitario e di appartenenza attraverso la condivisione di un'identità di scopo. Inoltre, la formazione deve assicurare omogeneità e correttezza dei comportamenti e delle prassi nel tempo e nello spazio in ordine a:

- aspetti tecnici e metodologici (corretta applicazione delle norme di rilevazione, aderenza alle definizioni e classificazioni ufficiali);
- comportamenti organizzativi (tempestività, efficienza, controllo dell'effetto rilevatore);
- aspetti giuridici e deontologici (obbligo di risposta, segreto statistico, trattamento dei dati personali);
- aspetti tecnologici.

Questi ultimi sono destinati ad assumere un peso sempre maggiore sia nelle attività di rilevazione, monitoraggio e coordinamento, che nell'aggiornamento anagrafico connesso al censimento. Inoltre, l'orientamento in favore di un sempre crescente ricorso alla auto-compilazione via web dei questionari di censimento implica maggiori esigenze di informazione e supporto dei rispondenti, sia per la gestione degli aspetti tecnologici che per la corretta comprensione e compilazione dei quesiti particolarmente complessi o articolati.

La crescente differenziazione del processo organizzativo e dei ruoli implica la necessità di articolare le strategie formative sotto il profilo organizzativo (calendarizzazione variabile, segmentazione dei *target* di riferimento, articolazione territoriale degli interventi) e soprattutto sotto il profilo dei contenuti, che

dovrebbero essere modulati sia in relazione alle diverse competenze richieste alle varie figure che in funzione del diverso grado di esperienza degli operatori. Soltanto gli operatori coinvolti nel processo con maggiore continuità e sistematicità acquisiranno nel tempo un apprezzabile livello di esperienza.

La strategia didattica più appropriata per fronteggiare efficacemente la complessità appena descritta, è quella della differenziazione dei contenuti e dei livelli di approfondimento della formazione erogata. La segmentazione dei percorsi formativi, sulla base dei profili professionali dei discenti e della loro esperienza sul campo consente di realizzare offerte più finalizzate e specifiche e favorisce la generazione di classi, comunità o gruppi di apprendimento più omogenei e motivati ad apprendere. Nella formazione degli adulti, infatti, la principale leva motivazionale è costituita appunto dall'utilità percepita della formazione (Knowles, 1996).

Strategie formative modulari e differenziate possono essere attuate anche in aula, con costi e impatti organizzativi importanti. Nella formazione a distanza in e-learning, invece, è meno oneroso realizzare e sostenere nel tempo un'offerta didattica modulare, anche a livello granulare fine, che consenta percorsi di formazione personalizzati in funzione dei diversi profili-utenti, oltre ad un'ampia flessibilità nella fruizione dei contenuti.

I numerosi strumenti di collaborazione e comunicazione offerti dal web offrono ampie possibilità di utilizzo di un approccio formativo di tipo applicativo, più adatto alla formazione degli adulti (Knowles, 1996). L'Istat dispone già di diversi strumenti di comunicazione e collaborazione via web che stanno ormai diventando familiari anche per gli operatori della rete territoriale del Sistema Statistico nazionale (Sistan). Oltre al sito istituzionale e al portale del Sistan con la Sistan Community, si citano i Social network e i social media (canale streaming, YouTube, Yoo+), le piattaforme e-learning (LCMS e Moodle), le aree Wiky e di condivisione, che oltre ad ospitare gli strumenti e i servizi formativi già sperimentati con successo nei precedenti censimenti, possono offrire nuove e importanti opportunità. Inoltre, grazie al sistema integrato di Lavagne Interattive Multimediali (Lim) e webmeeting, disponibili presso tutte le Sedi territoriali dell'Istituto, è possibile svolgere sessioni formative o divulgative multimediali in presenza e in rete, tramite l'accesso di utenti esterni e/o interni in modalità sincrona e/o asincrona ad aule virtuali pubblicate e accessibili da internet. Il sistema consente anche la realizzazione di webinar, utile strumento didattico per: incontri di follow-up, interviste a esperti, sessioni di problem solving (nelle quali i partecipanti possono condividere i racconti dei propri successi), approfondimenti di argomenti già trattati in presenza (come azione di rinforzo), pillole di apprendimento. L'utilizzo combinato di questi strumenti, che possono essere integrati sia nel portale della rete di rilevazione, vale a dire nell'ambiente web

implementato a supporto e monitoraggio dell'operatività della rete di rilevazione, ma anche nella piattaforma per l'acquisizione dei dati via web, può consentire lo sviluppo di vari modelli di apprendimento a distanza:

- Apprendimento individuale, in cui il discente interagisce con i contenuti (contenuti didattici e informazioni presenti sulla piattaforma) e con i tutor di processo e di contenuto, in modalità asincrona. In questo caso, il discente non viene inserito in nessuna classe virtuale perché si parte dall'ipotesi che ognuno concluda il percorso secondo tempi e ritmi propri, entro un intervallo temporale predefinito.
- Apprendimento collaborativo, basato sull'interazione di gruppo, il discente viene inserito in una classe virtuale che svolge un percorso comune. Tale modello prevede anche lo sviluppo di attività di cooperazione (sincrona e asincrona) tra i discenti e tra i discenti e i tutor con modalità varie: uno-a-uno, uno-a-molti, molti-a-molti (forum, bacheche, elettroniche, mailing-list, chatting) .
- Apprendimento intermedio, in cui il discente interagisce con i contenuti (contenuti didattici e informazioni presenti sulla piattaforma) e con i tutor di processo e di contenuto, in modalità asincrona e con la community di apprendimento tramite forum, bacheche e mailing-list.

La scelta dell'approccio formativo dipende dal livello di interattività che si vuole stabilire tra gli attori in gioco, dal livello di regia didattico-organizzativa che si vuole dare al corso e dal grado di strutturazione e di flessibilità che tali modelli prevedono. Le risorse umane e professionali disponibili, fondamentali per sviluppare un'offerta formativa efficace, e il calendario della rilevazione, sono i principali vincoli che influenzano tale scelta.

Per il successo della formazione a supporto del censimento continuo, in base alle esperienze precedenti, va sottolineato il ruolo chiave che può essere svolto da alcune misure di accompagnamento, che possono garantire una maggiore trasferibilità dell'e-learning e della formazione in generale, e da meccanismi di controllo e di sviluppo in un'ottica di miglioramento continuo dei prodotti, dei servizi e dei processi, in vista dell'accrescimento dell'efficacia formativa.

Sotto il primo profilo, nell'ambito del censimento continuo potrebbero essere maggiormente sfruttate le potenzialità di monitoraggio della formazione e di valutazione delle competenze acquisite offerte dall'e-learning, anche a fini di miglioramento della qualità dell'indagine: oltre a prevedere l'obbligatorietà della fruizione della formazione in e-learning, almeno per i contenuti fondamentali, potrebbero essere attuati meccanismi di verifica delle competenze acquisite, prevedendo anche una certificazione finale per incentivare i destinatari a un maggiore investimento nella formazione. Si può condurre un sondaggio per identificare i problemi che gli operatori incontrano sul lavoro, le ragioni di

preoccupazione, i fabbisogni di competenze ulteriori e le lacune rispetto ai compiti che devono svolgere. Per questo motivo, si suggerisce la previsione di strumenti/servizi di *profiling* dei bisogni formativi, da realizzare mediante strumenti standardizzati e generalizzati di diagnosi/autodiagnosi progettati e realizzati nell'ambito dell'allestimento dell'e-learning, da somministrare in presenza o distanza.

D'altra parte, per migliorare l'efficacia (effettiva e percepita) della formazione erogata risulta essenziale prevedere un'attività sistematica di valutazione non soltanto della *customer satisfaction*, ma anche e soprattutto della formazione fruita e dei relativi esiti in termini di apprendimento che potrebbe indirizzare la progettazione di interventi di ritorno sugli utenti e revisioni dei contenuti in un'ottica di miglioramento continuo.

Sulla scorta delle prime esperienze realizzate nell'ambito dei censimenti economici, inoltre, anche per il supporto ai rispondenti è abbastanza agevole prevedere di valorizzare le opportunità offerte dal web per realizzare:

- un "minicorso in e-learning" fruibile *on demand*, e ad accesso libero per una informazione completa e diretta ai rispondenti sui contenuti e sulle modalità di compilazione del questionario. Tramite video-tutorial, che fanno un uso integrato di immagini, grafica illustrativa e testi esplicativi, il minicorso illustra i quesiti, i percorsi di compilazione, e le funzioni di gestione del questionario web (autenticazione, compilazione, salvataggio, invio definitivo) che consentono la navigabilità dello stesso;
- un accesso diretto ai principali metadati di supporto alla corretta e completa compilazione del questionario sotto forma di tool-tip contestuali o di link ipertestuali a risorse internet integrati direttamente nel questionario web. Questa info-formazione, offerta a tutti i rispondenti, potrebbe concentrarsi sugli aspetti più importanti e strategici del questionario come ad esempio il contenuto informativo delle sezioni, le variabili di diffusione core, i quesiti filtro, i quesiti critici, le variabili di classificazione.

Il censimento permanente è un'opportunità fondamentale per l'Istat di introdurre un sistema di formazione continuo per un numero consistente di operatori statistici sul territorio, che sicuramente può avere ripercussioni positive su tutto il Sistema Statistico Nazionale.

Riferimenti bibliografici

BIANCHINO A., DE CANDIA G., TARALLI S. (2011). L'e-learning per le reti di rilevazione: una nuova opportunità per la qualità e la responsabilità sociale della statistica ufficiale, Atti SIEL 2011 - VIII Congresso della Società Italiana di e-Learning, Reggio Emilia.

DE CANDIA G., (2011). The basic training of the public statistical surveyors, Atti SISVSP Workshop on "Enhancement and social responsibility of official statistics". Roma.

ISTAT (2013 a). Atti del 6° Censimento generale dell'Agricoltura. Roma.

ISTAT (2013 b). L'Italia del Censimento – struttura demografica e processo di rilevazione – fascicoli regionali. Roma.

ISTAT (2014). Linee strategiche del censimento permanente della popolazione e delle abitazioni. Roma.

ISTAT (2015). Atti del 9° Censimento dell'industria e dei servizi e Censimento delle istituzioni non profit 2011. Roma (in corso di pubblicazione).

KNOWLES M. (1996). Quando l'adulto impara. Milano, Franco Angeli.

SUMMARY

Lifelong Learning in Permanent Census

In censuses round 2010-2011 Italian National Institute of Statistics introduced many innovations also in training field. Traditional classroom training was supported by new products and services in e-learning. Based on that experience, this paper proposes some insights and ideas to further innovate and improve training in support of the next continuous census, whose characteristics undoubtedly entail the opportunity and the need to design and actualize a lifelong learning plan.

Antonella BIANCHINO, Istat-Ufficio territoriale per la Basilicata, bianchin@istat.it
Giulia DE CANDIA, Istat-Ufficio territoriale per la Liguria, decandia@istat.it
Stefania TARALLI, Istat-Ufficio Territoriale per Emilia Romagna e Marche, taralli@istat.it

THE POST ENUMERATION SURVEY OF THE 15th ITALIAN POPULATION CENSUS: FEATURES AND METHODS[1]

Matteo Mazziotta, Monica Russo

1. Introduction

The Post Enumeration Survey (PES) is regulated by the European Commission (1151/2010). Its main goals is to estimate the number of individuals actually and habitually dwelling in the reference time of the 15th Population Census (October 9, 2011), the coverage rate and the over-coverage rate. The PES is based on a stratified two-stage sampling design. The first stage selects a sample of 252 municipalities as primary sampling unit, stratified by Regions and five classes of demographic dimension of municipalities. The second stage selects a sample of about 2.500 enumeration areas, stratified by demographic dimension of enumeration areas.

The survey is designed to ensure independence with the Census. Individuals habitually dwelling in housing in the selected enumeration areas, at the time of the PES are being carried out. The survey was conducted from April to July 2012, and involved about 1,200 enumerators selected among the "best" of the Census (of course, each of them is assigned a different area than the Census).

The PES has enumerated approximately 329.000 eligible individuals; an accurate Record Linkage phase and the application of a complex estimation model (Dual System approach derived by ONS) will construct a solid study towards the continuous census. The paper wants to describe the methodological aspects of the most important Istat quality survey.

2. Features

In this section we outline the adopted estimation methodology for the coverage survey of the 15th Census of Population and Housing, and we want to focus[2] on the

[1] The paper is the result of combined work of the authors: Matteo Mazziotta has written Section 1; Monica Russo has written Sects. 2 and 3.

research that has been conducted to develop improvements and innovations in the estimation model and adjusting for over-coverage.

The coverage survey – or *Post Enumeration Survey* (PES) – is a survey conducted in connection with the 15th Census of Population and Housing (the reference date of both the Census and the PES is October 9, 2011) and is designed to produce an assessment of the level of accuracy of the Census through the determination of the over-count and under-count errors happened during the Census enumeration.

According to these aims, the main population parameters of interest are:

- the *coverage rate*, defined as the ratio between the number of individuals found to Census (net of over-coverage) and the true amount of the population,
- the *under-coverage rate*, defined as the ratio between the number of individuals missed to the Census (net of over-coverage) and the true amount of the population,

where the denominator of both – that is, the true amount of the population – is unknown and, then, it has to be estimated by PES through a *Dual-System Estimator* (DSE), or capture-recapture model[3].

The purpose of the survey is the achievement of the estimates of the aforementioned parameters with reference to the following domains: (i) the whole national territory, (ii) the geographical regions, (*iii*) the age classes: 0-14, 15-29, 30-49, 50-64, 65 and more, (*iv*) the nationality: italian, foreign.

The PES is a stratified two-stage areal sample selection of Census Enumeration Areas (EAs) that are independently re-enumerated. The first stage selected a sample of municipalities, stratified by geographical region and five classes of municipal demographic size. The second stage selected a sample of EAs, stratified by the tertiles of the EAs population distribution, from within each selected Municipality. All individuals who live in sample EAs are enumerated. The selection of the primary units is with unequal probability without replacement and the selection of secondary units is with equal probability without replacement.

The main types of counting errors that can occur in a Census are:

1) the *under-coverage error*, that happens when a person living in Italy on October 9, 2011, is not counted by the Census,
2) the *over-coverage error*, that can be identified in four forms:
   Type 1 (*duplicate returns at the same location*): this is where two or more returns are made in the same small area (EA) by the same individual. For example, a person may return a paper form, and submit an internet

2 For a complete description of the sampling strategy and of the estimates accuracy assessment of PES 2011, see Istat (2015).

[3] Sekar, C.C., Deming W.E. (1949).

questionnaire,

Type 2 (*duplicate returns from different locations*): this is where a response is received for the same person but from a different small area, such as students being counted at both their term-time address (correctly) and their parents' address (incorrectly).

Type 3 (*counted in the wrong location*): this is a return where the Census only counts the 'wrong' half of a duplicate. This might happen by only counting the student at their parents' address or by counting a mover after Census reference date only at the new address,

Type 4 (*erroneous returns*): this is a return that is purely fictitious and should not appear in the Census at all at any location. This can be a joke return, a creation of the processing, a baby born just after Census reference date or individuals that died just prior to Census reference date.

Type 1 over-count should be rectified during data processing. Type 4 over-count is only identifiable by further field work. Types 2, 3 will remain in the Census population; therefore, we took in account these two types of error in the estimation procedure.

The aim of the estimation methodology is to produce 'net' adjustments[4]: (*i*) reducing the Dual-System estimates by the estimated level of over-count, (*ii*) imputing the fewest number of individuals, (*iii*) not removing duplicates.

The estimation process is structured as below:

a) definition of a measure of the *over-coverage propensity*, based on the correct and erroneous Census counts, and study of the over-coverage propensity estimator;

b) determination of the DSE adjusted by over-coverage, for age-sex groups separately, for each sample Municipality selected in each Hard to Count (HtC) index[5] stratum, within each geographical region;

c) computation of a population estimate at geographical region level separately for age-sex groups and HtC index modalities, by using a ratio estimator;

d) application of the *Sample Balance Adjustment* procedure in all the HtC strata in which samples selected proved to be *outlier*, to make the estimates more reliable;

e) finally, you can immediately obtain estimates of true (unknown) regional and

[4] Separate adjustments for under-count and over-count will not be made.

[5] As the accuracy of population estimates is dependent on response rate and its variation within estimation strata and as the under-count is disproportionately distributed across areas, a hard to count categorization has been produced and the municipalities within each geographical region are post-stratified according to an HtC index. This index attempts to capture the non-geographical variation in under-count in a Census. For defining homogeneous areas, we studied an index based on the predicted individual non-response rate, so as to classify municipalities from 1 to 3 according to their expected relative difficulty of enumeration.

national populations totals using estimates produced up to this point.

3. Estimation model[6]

The methodology adopted for estimating the over-coverage in PES 2011 is constituted of the following three steps: (1) estimate the number of duplicates using PES; (2) estimate people in wrong place using PES; (3) calibration of the estimates referred to point (1), using Census duplicates number, to improve the accuracy of the over-coverage estimates (PES sample is not designed for getting estimates of duplicates number)[7].

The *over-coverage propensity*, $\gamma_{a_w,g}$, is the propensity for those individuals with characteristics defined[8] by a_w age class within area *g* to be over-counted by the Census; in symbols

$$\gamma_{a_w,g} = X_{a_w,g} / Y_{a_w,g} = \left(Y_{a_w,g} + E_{a_w,g}\right) / Y_{a_w,g} , \tag{1}$$

in which: $X_{a_w,g}$, $Y_{a_w,g}$, $E_{a_w,g}$ are the total Census count, the *correct* Census count and the erroneous (over-count) Census count, for group a_w in area *g*, respectively. The plug-in estimator of (1) is given by

$$\hat{\gamma}_{a_w,g} = \left(\hat{Y}_{a_w,g} + \hat{E}_{a_w,g}\right) / \hat{Y}_{a_w,g} , \tag{2}$$

with: $\hat{Y}_{a_w,g} = \sum_{j \in s_g} w_{jg} \sum_{i \in j} c_{a_w,i,jg}$ and $\hat{E}_{a_w,g} = \sum_{j \in S_g} \frac{\tilde{P}}{\hat{D}} \sum_{b} \sum_{\substack{k \in s_b \\ k \neq j}} w_{kb} \sum_{i \in k} o_{a_w,i,kb,jg}$,

where $\tilde{P}$ is the Census duplicates number, $\hat{D} = \sum_{g} \sum_{j \in s_g} w_{jg} \sum_{b} \sum_{\substack{k \in S_b \\ k \neq j}} w_{kb} \sum_{i \in k} o_{i,kb,jg}\ c_{i,jg}$ is its estimate by PES.

Furthermore, the symbols showed in the formulas represent: *i* the individual; *b* the

[6] I would like to thank the doctors Owen Abbott and Amy Large of the Office for National Statistics (ONS) for their useful comments, valuable suggestions and reports of inaccuracies that I took into account in writing this note.

[7] For what just declared, it is clear that a key assumption of the over-coverage estimation procedure is the definition of the correct location of the individual on the Census reference date. We assume that the correct location is given by the individual's response to the question 1.5 of the PES questionnaire.

[8] We set five age classes a_w for the over-coverage estimation procedure: 0-2 and 25-60, 3-17, 18-24, 61+, and fifteen areas *g*: five geographical distributions – North-Ovest, North-East, Centre, South, Islands – by HtC.

geographical distribution by HtC (b can be the same as g); j and k the EAs belonging to g and b respectively $(j \neq b)$; s_g and s_b the number of sample EAs within g and b respectively; S_g and S_b the number of EAs within g and b respectively; w_{jg} and w_{kb} the PES sampling weights for j selected in g and for k selected in b respectively; $c_{a_w,i,jg}$ an indicator random variable that assumes value 1 if individual i is correctly counted by the Census in j within g and 0 otherwise; $o_{a_w,i,kb,jg}$ an indicator random variable that assumes value 1 if individual i from k within b ($k \neq j$ and b can be the same as g) is incorrectly counted by the Census in j within g and 0 otherwise.

The basis of the under-count estimation is the use of dual-system estimation; the model adopted for PES 2011 is a variant of *Petersen model*[9], in which the Census total is adjusted for taking over-coverage into account.

Let us denote with a the age-sex groups[10], p the generic modality of HtC, r the geographical region and c the generic Municipality and, for the moment, suppose that PES is in fact a complete enumeration of EAs. The DSE of ${}_rN_{apc}$, the true population total, is

$$ {}_r\tilde{N}_{apc} = {}_rN_{+1,apc}\ {}_r\bar{\bar{N}}_{1+,apc} \big/ {}_rN_{11,apc}\,, \tag{3} $$

where ${}_rN_{+1,apc}$, ${}_r\bar{\bar{N}}_{1+,apc}$ and ${}_rN_{11,apc}$ is the PES count, the Census count of correct individuals and the matched count respectively. However, we observe ${}_rN_{1+,apc}$, the Census count including over-count; therefore, in order to correct a possible bias in the results and considering the fact that the over-coverage propensity is constant in g, then $E\left[\left({}_rN_{1+,apc}/\hat{\gamma}_{a_w,g}\right)\middle|\ {}_rN_{1+,apc}\right] \cong {}_r\bar{\bar{N}}_{1+,apc}$, we can adjust the DSE to give

$$ {}^{o}_{r}\tilde{\tilde{N}}_{apc} = {}_rN_{+1,apc}\left({}_rN_{1+,apc}/\hat{\gamma}_{a_w,g}\right) \big/ {}_rN_{11,apc}\,, \tag{4} $$

[9] Wolter, 1986.

[10] For the under-coverage estimation procedure we defined thirty four age-sex groups formed by the age classes: 0-2, 3-7, 8-17, 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+, separately for female and male.

as ${}^{o}_{r}\tilde{\tilde{N}}_{apc}$ is approximately unbiased for ${}_{r}\tilde{N}_{apc}$ given ${}_{r}N_{+1,apc}$, ${}_{r}N_{11,apc}$ and ${}_{r}N_{1+,apc}$.

The estimation methodology goes further and applies the *Chapman correction* for DSE's applications to small populations, which allows a better respect of the homogeneity hypothesis of the capture probability of Petersen model. With respect to (3), we have: ${}_{r}\tilde{N}^{C}_{apc} = \left({}_{r}N_{+1,apc}+1\right)\left({}_{r}\bar{\bar{N}}_{1+,apc}+1\right)\left({}_{r}N_{11,apc}+1\right)^{-1}-1$, whose first-order approximation is given by ${}_{r}\tilde{N}^{C}_{apc} \cong {}_{r}\tilde{N}_{apc}\left(1+{}_{r}N^{-1}_{+1,apc}+{}_{r}\bar{\bar{N}}^{-1}_{1+,apc}\right)\left(1-{}_{r}N^{-1}_{11,apc}\right)-\left(1-{}_{r}N^{-1}_{11,apc}\right)$.

However, with the DSE, the observed total is ${}_{r}N_{1+,apc}$, so the adjusted DSE with Chapman correction is given by

$$
\begin{aligned}
{}^{o}_{r}\tilde{\tilde{N}}^{C}_{apc} &= \left({}_{r}N_{+1,apc}+1\right)\left[\left({}_{r}N_{1+,apc}/\hat{\gamma}_{a_w,g}\right)+1\right]\left({}_{r}N_{11,apc}+1\right)^{-1}-1 \cong \\
&\cong {}^{o}_{r}\tilde{\tilde{N}}_{apc}\left(1+{}_{r}N^{-1}_{+1,apc}+\hat{\gamma}_{a_w,g}\,{}_{r}N^{-1}_{1+,apc}\right)\left(1-{}_{r}N^{-1}_{11,apc}\right)-\left(1-{}_{r}N^{-1}_{11,apc}\right)
\end{aligned}
\tag{5}
$$

and, therefore, ${}^{o}_{r}\tilde{\tilde{N}}^{C}_{apc}$ is an approximately unbiased estimator for ${}^{o}_{r}\tilde{\tilde{N}}_{apc}$, given the observed counts ${}_{r}N_{+1,apc}$, ${}_{r}N_{11,apc}$ and ${}_{r}N_{1+,apc}$. As PES is an enumeration of only a sample of EAs and not of all of them, the estimator of ${}_{r}N_{apc}$ is

$$
{}^{o}_{r}DSE_{apc} = {}^{o}_{r}\hat{N}^{C}_{apc} = \left({}_{r}\hat{N}_{+1,apc}+1\right)\left[\left({}_{r}N_{1+,apc}/\hat{\gamma}_{a_w,g}\right)+1\right]\left({}_{r}\hat{N}_{11,apc}+1\right)^{-1}-1, \tag{6}
$$

where ${}_{r}\hat{N}_{+1,apc}$ and ${}_{r}\hat{N}_{11,apc}$ are estimated using PES.

The following stage in the estimation process is to generalize the DSEs to the non-sampled areas, computing the population estimates at geographical region level separately for age-sex groups and HtC index strata, by using a ratio estimator that estimates the relationship in the sample between the Census count and the dual-system estimate for each age-sex group within each HtC stratum

$$
{}^{o}_{r}DSE_{apc} = {}_{r}\beta_{a}\ {}_{r}x_{apc} + {}_{r}\varepsilon_{apc}, \tag{7}
$$

in which: ${}_{r}\beta_{a}$ is the regression coefficient referred to group of individuals within *r* having sex and age defined by *a*, ${}_{r}x_{apc}$ is the number of persons enumerated by the

Census within the sample Municipality *c* selected in *p* within *r* and with sex and age defined by *a* and ${}_r\varepsilon_{apc}$ is an error term expressed as the difference between the variable and its conditional average.

Let us indicate with: ${}_r\hat{\beta}_a$ the least squares estimator of ${}_r\beta_a$, ${}_rX_{ap}$ the Census count of persons with sex and age defined by *a* from all the municipalities within *p* in *r*. We can use the relationship (7) to estimate the total region population for each *a* in each *p* by multiplying the Census count by the estimated slope of the line: ${}_r\ddot{N}_{ap} = {}_r\hat{\beta}_a \; {}_rX_{ap}$.

The efficiency of the estimates ${}_r\ddot{N}_{ap}$ can be improved by applying the methodology, showed below and known as *Sample Balance Adjustment* (SBA), in cases where the samples used for obtaining them were not sufficiently representative of the population being measured.

Under normal circumstances the samples, from the PES, were expected to have a good representation of Census non-response. However, for every sampling process there is a risk that a sample may be an *outlier* among all possible samples. In other words, our PES sample could have, by chance, drawn EAs in an area where the Census had managed to count everyone.

The aim of this methodology is to assess if the PES sample was sufficiently representative when compared to all other possible samples that could have been drawn from the underlying population. If it was not, then the coverage adjustment estimates would be skewed (either too large or too small), because the sample was unusual. Census *dummy* questionnaires were used to evaluate whether any of the PES samples were outlying (i.e. unbalanced). These data were believed to be the best possible proxy for coverage, because they represent households from which a return was not received. The procedure can be summarized as follow.

First of all, we analyze the correlation between the *non-response rates* (that is, the ratio between the number of Census *dummy* questionnaires and the number of all Census questionnaires) and their PES estimates. If there is correlation, then the Census *dummy* variable is a good proxy for coverage and it can be used if the sample is found not representative.

The second step consists in comparing the number of Census *dummy* questionnaires referred to *r* and *p*, ${}_rY_p$, and its PES ratio estimate, ${}_r\hat{Y}_p$: if the two are significantly different then the sample related to *r* and *p* was unbalanced, or an outlier.

However, we went further and a statistical test has been also used. Being known the true distribution of the Census *dummy* questionnaires across all the municipalities within *r* and *p*, we determine the variance (and not its estimate) of

${}_r\hat{Y}_p$. Then a 95% confidence interval has been constructed around the true number of Census *dummy* questionnaires ${}_rY_p$:

$${}_rY_p - 2\sqrt{Var\left({}_r\hat{Y}_p\right)} \le {}_rY_p \le {}_rY_p + 2\sqrt{Var\left({}_r\hat{Y}_p\right)}.$$

If the sample estimate was within the 95% confidence interval then there was no evidence to suggest the sample was extreme. If there was evidence to suggest the sample was extreme, a multiplicative adjustment was made to all the ratio estimators by age and sex in that HtC stratum using the ratio of the population level *dummy* questionnaire implied coverage to the equivalent for the sample

$${}_rF_p = \left({}_rY_p + {}_rX_p\right)_r X_p^{-1}\left(\frac{{}_rM_p}{{}_rm_p}\sum_{c=1}^{{}_rm_p}{}_rX_{pc}\right)\left[\frac{{}_rM_p}{{}_rm_p}\sum_{c=1}^{{}_rm_p}\left({}_rY_{pc} + {}_rX_{pc}\right)\right]^{-1}. \qquad (8)$$

This adjustment factor was: 1, when the sample was exactly balanced; >1, when the sample was underestimating; <1, when overestimating. ${}_rF_p$ effectively uprated (or deflated) the estimated Census coverage rate.

Given ${}_rF_p$, we can obtain the balanced estimate of ${}_rN_{ap}$, through the relationship: ${}_r\ddot{\ddot{N}}_{ap} = {}_rF_p\ {}_r\ddot{N}_{ap} = {}_rF_p\ {}_r\hat{\beta}_a\ {}_rX_{ap}$.

At this point, we can easily determine all the other estimates that are the primary aim of PES 2011. By summing with respect to HtC, you have the estimates for each sex-age group within region *r*: ${}_r\ddot{\ddot{N}}_a = \sum_p {}_r\ddot{\ddot{N}}_{ap}$; by summing with respect to age-sex groups, you have the estimates for each geographical region: ${}_r\ddot{\ddot{N}} = \sum_a {}_r\ddot{\ddot{N}}_a = \sum_a\sum_p {}_r\ddot{\ddot{N}}_{ap}$; the national estimate is, then, provided by the relationship: $\ddot{\ddot{N}} = \sum_r {}_r\ddot{\ddot{N}}$.

References

ISTAT. 2015. L'indagine di controllo della copertura censuaria in *Atti del 15° Censimento della popolazione*, Volume No. 6.

SEKAR C. C., W.E. DEMING. 1949. On a Method of Estimating Birth and Death Rates and the Extent of Registration, *Journal of the American Statistical Association*, Vol. 44, No. 245, pp. 101-115.

WOLTER K.M. 1986. Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, Vol. 81, No. 394, pp. 338-346.

SUMMARY

The Post Enumeration Survey of the 15th Italian Population Census: Features and Methods

The Post Enumeration Survey (PES) has the main goals to estimate the number of individuals actually and habitually dwelling in the reference time of the 15th Population Census (October 9, 2011), the coverage rate and the over-coverage rate. In order to achieve this objective, a complex estimation model, derived by ONS approach, has been implemented. The paper presents the main features of the survey and the steps of the estimation process.

Matteo MAZZIOTTA, Istat, mazziott@istat.it
Monica RUSSO, Istat, morusso@mail.it

ON THE CONSTRUCTION OF COMPOSITE INDICES BY PRINCIPAL COMPONENTS ANALYSIS[1]

Matteo Mazziotta, Adriano Pareto

1. Introduction

Social and economic phenomena such as development, poverty, quality of life, innovation and competitiveness, are very difficult to measure and evaluate since they are characterized by a multiplicity of aspects or dimensions. The complex and multidimensional nature of these concepts requires the definition of intermediate objectives whose achievement can be observed and measured by individual indicators. A mathematical combination (or aggregation as it is termed) of a set of indicators that represent the different dimensions of a phenomenon to be measured is called *composite index* (Saisana *et al*., 2002).

The literature on the methods for constructing composite indices is vast and the number of composite indices in the world is growing year after year (Bandura, 2008). Examples of well-known composite indices are the United Nations' *Human Development Index* – HDI – (UNDP, 2010) and the European Commission's *Regional Competitiveness Index* – RCI – (Annoni and Kozovska, 2010).

Producing a single composite index has advantages, such as simplicity, although a set of individual indicators might be preferable for other reasons, such as completeness of the information. However, the procedure for constructing a composite index is very far from being aseptic and requires a number of subjective decisions to be taken (OECD, 2008; Mazziotta and Pareto, 2013).

A fundamental issue often overlooked in composite index construction is the definition of the model measurement, in order to specify the relationship between the concept to be measured and its measures (individual indicators). In this respect, the direction of the relationship is either from the concept to the measures - *reflective model* - or from the measures to the concept - *formative model* (Maggino, 2014).

[1] The paper is the result of combined work of the authors: Matteo Mazziotta has written Sects. 1 and 5; Adriano Pareto has written Sects. 2, 3 and 4.

In this paper, we compare the two approaches, and we show that factorial methods, such as Principal Components Analysis (PCA), may fail if improperly used.

2. Formative versus reflective measurement models

As is known, a model of measurement can be conceived through two different conceptual approaches: reflective or formative (Diamantopoulos *et al*., 2008).

The most popular approach is the reflective model, according to which individual indicators denote effects (or manifestations) of an underlying latent variable. Therefore, causality is from the concept to the indicators and a change in the phenomenon causes variation in all its measures. In this model, the concept exists independently of awareness or interpretation by the researcher, even if it is not directly measurable.

Specifically, the latent variable R represents the common cause shared by all indicators X_i reflecting the concept, with each indicator corresponding to a linear function of the underlying variable plus a measurement error:

$$X_i = \lambda_i R + \varepsilon_i \tag{1}$$

where X_i is the indicator *i*, λ_i is a coefficient (loading) capturing the effect of R on X_i. and ε_i is the measurement error for the indicator *i*. Measurement errors are assumed to be independent and unrelated to the latent variable.

A fundamental characteristic of reflective models is that indicators are interchangeable (the removal of an indicator does not change the essential nature of the underlying concept) and correlations between indicators are explained by the measurement model. A typical example is the intelligence of a person: it is the 'intelligence level' that determines the answers to a questionnaire for measuring attitude, not vice versa.

The second approach is the formative model, according to which individual indicators are causes of an underlying latent variable, rather than its effects. Therefore, causality is from the indicators to the concept and a change in the phenomenon does not necessarily imply variations in all its measures. In this model, the concept is defined by, or is a function of, the observed variables.

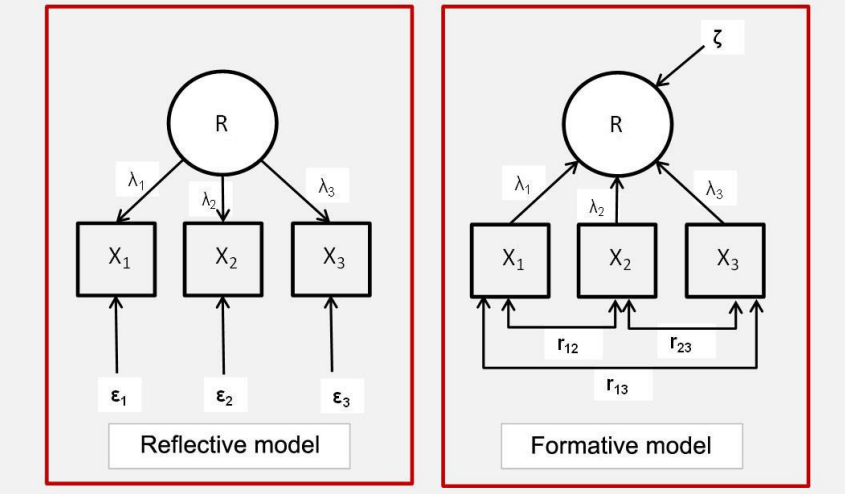
The specification of the formative model is:

$$R = \sum_i \lambda_i X_i + \zeta \tag{2}$$

where λ_i is a coefficient capturing the effect of X_i on R, and ζ is an error term.

In this case, indicators are not interchangeable (omitting an indicator is omitting a part of the underlying concept) and correlations between indicators (r_{ij}, $i \neq j$) are not explained by the measurement model. A typical example is the socio-economic status of a person: it is the 'status level' that depends on education, income, occupation, and residence, not vice versa.

Figure 1 – *Alternative measurement models*



Note that (1) is a simple regression equation where the individual indicator is the dependent variable and the latent variable is the explanatory variable; whereas (2) represents a multiple regression equation where the latent variable is the dependent variable and the indicators are the explanatory variables.

In fig. 1, the two different approaches are graphically represented. Although the reflective model dominates the psychological and management sciences, the formative model is common in economics and sociology (Coltman *et al*., 2008).

3. A numerical example

Imagine that we want to construct a composite index of development in the work dimension, for several countries or regions, based on:

X_1 = Employment rate;

X_2 = Incidence rate of occupational injuries.

Indicator X_1 has positive polarity[2] (it is positively correlated with the development), whereas indicator X_2 has negative polarity (it is negatively correlated with the development).

Suppose also that X_1 and X_2 are positively correlated, so that high employment rates tend to be associated with higher rates of occupational injuries.

In a formative view, we can aggregate the data by arithmetic mean, whereas in a reflective view, the first principal component is the best solution.

In table 1 is reported an example where five countries are considered. The table also provides the normalized indicators[3] Z_1 and Z_2, the arithmetic mean of the normalized values M_1, and the first component[4] score PC_1. Note that $r(X_1, X_2)=0.45$, whereas $r(Z_1, Z_2)=-0.45$, because the polarity of X_2 has been inverted in order to construct the composite index[5].

Table 1 – *Comparing arithmetic mean and first component score as composite indices*

Country	Original values		Normalized values		Ranks		M_1		PC_1	
	X_1	X_2	Z_1	Z_2	R_1	R_2	Value	Rank	Value	Rank
1	80.0	1.9	1.58	-0.71	1	4	0.44	2	1.20	1
2	50.0	2.0	0.00	-1.41	3	5	-0.71	5	0.74	2
3	50.0	1.8	0.00	0.00	3	3	0.00	3	0.00	3
4	50.0	1.6	0.00	1.41	3	1	0.71	1	-0.74	4
5	20.0	1.7	-1.58	0.71	5	2	-0.44	4	-1.20	5
Mean	**50.0**	**1.8**	**0.00**	**0.00**						
Std Dev	**19.0**	**0.1**	**1.00**	**1.00**						

As we can see, units 2, 3, and 4 have the same employment rate ($X_1=50.0$) and decreasing values of the rate of occupational injuries. Nevertheless, unit 2 ranks 5th according to M_1 and ranks 2nd according to PC_1, whereas unit 4 ranks 1st according to M_1 and ranks 4th according to PC_1.

[2] The 'polarity' of a individual indicator is the sign of the relation between the indicator and the phenomenon to be measured. For example, in the case of development, "Life expectancy" has positive polarity, whereas "Infant mortality rate" has negative polarity.

[3] Normalization is required to make the indicators comparable as they often have different measurement units or polarities. In this case, we transformed individual indicators into *z*-scores and we changed the sign if the polarity is negative.

[4] The first principal component accounts for 72.4% of the variance in the data.

[5] When a composite index must be constructed, all the individual indicators must have positive polarity, so that an increase in each indicator corresponds to an increase in the composite index (Mazziotta and Pareto, 2013).

So, the average Spearman rank correlation coefficient between the composite index and the individual indicators is 0.52 for M_1 and 0.05 for PC_1. This is due to the fact that PCA ignores the polarity of the indicators and all normalized indicators, in a reflective measurement model, must be positively intercorrelated.

Therefore, the use of PC_1 for aggregating X_1 and X_2 is incorrect either from a theoretical point of view (PC_1 is a reflective composite index, whereas M_1 is a formative composite index) or from a purely numerical point of view (PC_1 is concordant with both X_1 and X_2, whereas M_1 is concordant with X_1 and discordant with X_2).

4. Some other issues

The PCA has a number of excellent mathematical properties (Kendall and Stuart, 1968). The most important property is that the index obtained from the first principal component explains the largest portion of variance of the individual indicators. This is obtained by maximizing the sum of the squares of the coefficients of correlation between the composite index and the individual indicators. However, the first principal component accounts for a limited part of the variance in the data (in the previous example, 72.4%), so we can lose a consistent amount of information.

Moreover, the PCA based index is often *elitist* (Mishra, 2008), with a strong tendency to represent highly intercorrelated indicators and to neglect the others, irrespective of their possible contextual importance. So many highly important but poorly intercorrelated indicators may be unrepresented by the composite index. On many occasions, it is found that some (evidently) very important indicators are roughly dealt with by PCA, simply because those variables exhibited widely distributed scatter or they did not fall within a narrow band around a straight line (Mishra, 2007).

On the other hand, PCA is a blindly empiricist method based on the observed correlations and it ignores the polarity of the individual indicators. Therefore, if the normalized indicators are not all positively intercorrelated, the results are not correct, as shown above.

Finally, it should be noted that the amount of variance accounted for, and the weights computed by PCA change over time, so the results of different PCAs are not easily comparable.

5. Conclusions

The construction of composite indices for assessing multidimensional phenomena is a central issue in data analysis, particularly in economics and sociology. Researcher cannot solve this question simply by using PCA or related methods, such as Factor Analysis, since they are typically used for a reflective approach and they ignore the polarity of the individual indicators.

Often, a formative approach is required, where the index to be constructed does not exist as an independent entity, but is a composite measure directly determined by a set of non-interchangeable individual indicators. It is the case of the HDI and the RCI. The first index does not use PCA, whereas the second one uses PCA 'only' for selecting indicators. So, in order to obtain valid and reliable results, it is absolutely essential to define the theoretical framework with an appropriate measurement model.

This paradigm should always be considered when the objective of the research is to measure a multidimensional phenomenon through composite indices. And this is even more valid if the phenomenon to be measured is well-being, progress or quality of life. Indeed, these latent factors depend on the individual indicators that represent them and not the contrary. Therefore, the use of PCA for the measurement of these phenomena is at all improper.

References

ANNONI P., KOZOVSKA, K., 2010. *EU Regional Competitiveness Index 2010*. JRC Scientific and Technical Reports. Luxembourg: Publications Office of the European Union.

BANDURA R., 2008. *A Survey of Composite Indices Measuring Country Performance: 2008 Update*. New York: UNDP/ODS Working Papers.

COLTMAN T., DEVINNEY T.M., MIDGLEY D.F., VENAIK S., 2008. Formative versus reflective measurement models: Two applications of formative measurement, *Journal of Business Research*, Vol. 61, pp. 1250-1262.

DIAMANTOPOULOS A., RIEFLER P., ROTH, K.P., 2008. Advancing formative measurement models, *Journal of Business Research*, Vol. 61, pp. 1203-1218.

KENDALL, M.G., STUART A., 1968. *The Advanced Theory of Statistics, Vol. 3*. London: Charles Griffin & Co.

MAGGINO F., 2014. Indicator Development and Construction. In MICHALOS, A. C. (Ed) *Encyclopedia of Quality of Life and Well-Being Research*, Dordrecht: Springer, pp. 3190-3197.

MAZZIOTTA M., PARETO A., 2013. Methods for Constructing Composite Indices: One for all or all for one, *Rivista Italiana di Economia Demografia e Statistica*, Vol. LXVII, n. 2, pp. 67-80.

MISHRA, S.K., 2007. A Comparative Study of Various Inclusive Indices and the Index Constructed by the Principal Components Analysis. *MPRA Paper*, No. 3377. Available at MPRA: http://mpra.ub.uni-muenchen.de/3377.

MISHRA, S.K., 2008. On Construction of Robust Composite Indices by Linear Aggregation. Available at SSRN: http://ssrn.com/abstract=1147964.

SAISANA M., TARANTOLA, S., 2002. *State-of-the-art report on current methodologies and practices for composite indicator development*. European Commission-JRC, EUR 20408 EN, Ispra.

OECD, 2008. *Handbook on Constructing Composite Indicators. Methodology and user guide*. Paris: OECD Publications.

UNDP 2010. *Human Development Report 2010. The Real Wealth of Nations: Pathways to Human Development*. New York: Palgrave Macmillan.

SUMMARY

Principal Components Analysis (PCA) is one of the most commonly used multivariate statistical technique in construction of composite indicators. However, PCA and related methods, such as Factor Analysis, are based on a reflective model where the individual indicators (manifest variables) are seen as functions of a latent variable (principal component or factor). When individual indicators are causes of the latent variable, rather than its effects, a formative model should be adopted. In this paper, we compare the two approaches, and we show by a numerical example that factorial methods, such as PCA, may fail if improperly used.

Matteo MAZZIOTTA, Italian National Institute of Statistics, mazziott@istat.it
Adriano PARETO, Italian National Institute of Statistics, pareto@istat.it

MEASURING SYSTEMIC RISK THROUGH STATISTICAL COMBINATION

Francesca Parpinel, Claudio Pizzi

1. Introduction

The necessity of defining risk goes back to the days of ancient Greece, but nowadays this need is increasingly pressing. Monitoring the riskiness of each financial institution in a network (for example in a country) is now an essential requirement in particular after the recent Financial Crisis of 2007-2009, defined by some economists as *the worst crisis after the Great Depression of the thirties*. This induces new definitions and measures of the risk associated to the institutions, taking into account two different features. In fact, considering the single institutions, it measures some peculiar aspect of their riskiness, but if we consider the problem in a wide sense, the risk is a measure that involves the links of a network. In the last case it is called *systemic risk* and it has not a unique definition, as the phenomenon is complex and with a lot of facets. This makes difficult even its measurement. For example, in Eisenberg and Noe (2001) it is considered as the possibility that an insolvent financial institution may transfer its insolvency to the whole financial system. Other peculiar definitions compare it to Nessie, the Loch Ness Monster (Bandt and Hartmann, 2000), as everyone knows it *but nobody knows when and where it might strike*.

Das and Uppal (2004) consider the risk coming from some unusual event with strong correlations among different assets. Kaufman (1994) states that it is the consequence of a series of losses moving within a network of markets or institutions.

Further definitions can be found in the literature and the natural consequence is the statistical measuring of such phenomenon. The importance of defining, and then measuring, systemic risk is really strong as financial surveillance is nowadays necessary for the governments policies of various countries.

In particular, we may cite the work by Billio et al. (2012) in which they propose several indices to measure the systemic risk of four groups of financial institutions, using correlations, cross-autocorrelations, principal component analysis, regime-switching models and statistics to test Granger causality on time series

observations. Furthermore, they adopt a graphical tool to represent the Granger causality index for each institution using network diagrams.

In the present work we want to highlight the relations among the individual institutions, considering an economic index linking several variables that characterize each financial institutions within some group, with the aim of ordering the risk of the companies in a network.

The ranking induced by this index will be compared with the one induced by the systemic risk measure proposed by *V-Lab* (Volatility Laboratory, `http://vlab.stern.nyu.edu/`), the platform of the NYU Stern School of Business providing real time measurement, modelling and forecasting of financial volatility and correlations for a wide spectrum of assets. The risk measure computed by *V-Lab*, called *SRISK*, estimates the amount of recapitalization necessary to a company not to fail in a financial crisis, while the index built in this work tries to estimate the effective level risk in a specific time. The comparison lets us to find analogies and differences in the two rankings constructed with real data, suggesting some new risk measures.

2. Ranking and analysis

In this paper, the main idea is to compact the high number of involved variables in order to create only one dimension and to treat easily ranking among institutions. When the goal is to reduce the number of involved variables, we may use several statistical tools. Here we propose the nonparametric combination and we remember the principal component analysis, in order to compare the results.

In a Gaussian framework, it is possible to reduce the number of variables, that we denote as $X = (X_1, \dots, X_K)$, keeping as much as possible the variability structure of a set of statistical units, that, in our case, will be represented by the covariance matrix only through a linear combination of them. Following for instance Rencher (2002), the linear combination $Y = A \cdot X$ obtained by a principal component analysis produces uncorrelated components and is ordered following the decreasing size of the variance of the new components. In this way, if the first components were able to get most of the system variability, we could consider only them as representative of the whole system. It's well known that the procedure to decide how many components we may choose is not unique, see Everitt (2004); we can use, for example, the Kaiser rule, according to which we keep the components with variance greater than one, otherwise we can choose enough variables to explain some fixed proportion of the cumulative total variation of the original variables, another tool is the so-called *scree-plot*, a barplot representing the variances of the new components, detecting where the diagram shows a sort of

elbow. The actual bound in this technique is linked with the Gaussian assumption, as the multivariate variability is represented only by the variance matrix.

In alternative to a linear combination of statistical measures, we propose to use a nonparametric one based on the ranking of such measures, according to Pesarin's work (Pesarin and Salmaso, 2010). The nonparametric combination is satisfactory even when the rankings may be dependent. Each measure can capture only some feature of risk and of systemic risk too, so our idea is to use all the available variables giving some partial, even overlapping, information about them.

Let's suppose to measure all of them with $K>1$ random variables, denoted by $(X_1, \ldots, X_k, \ldots, X_K)$ and to transform each of them in some variable defined over $[0,1]$ and called λ_k, $k=1, \ldots, K$. The combination ψ of these new variables λ_k may (or may not) depend on some weights, denoted by $w_1, \ldots, w_K$, according to the importance of each variable, and it produces a new variable Y, ψ: $R^{2K} \rightarrow R^1$. Following Lago and Pesarin (2000), the idea of combining different statistical indices, typically dependent on each other, arises from the same procedure for combination of dependent tests in multivariate analysis. In the inferential case the combining function is applied to *p*-values associated to marginal tests and is typically a nonparametric one.

As well described in Pesarin and Salmaso (2010), the combination function, ψ, has to satisfy some minimal properties. Function ψ must be continuous in all its arguments; it is non-decreasing in each λ_k, $k=1, \ldots, K$: $\psi(\ldots, \lambda_k, \ldots; w_1, \ldots, w_K) \geq \psi(\ldots, \lambda'_k, \ldots; w_1, \ldots, w_K)$ when $\lambda_k > \lambda'_k$, for all $k=1, \ldots, K$; ψ must be symmetric, i.e. invariant with respect to rearrangement of the variables λ_k; the supremum of ψ, $\bar{\psi}$, is attained when even one value of λ_k tends to zero; the value of ψ is always less than $\bar{\psi}$.

In this work we will use the Fisher combination function

$$\psi = -\sum_{k=1}^{K} w_k \cdot \log(1 - \lambda_k) \qquad (1)$$

but in the literature we can find other important combining functions, satisfying the above properties, for example the Tippett one $\psi_T = \max_k(w_k \cdot \lambda_k)$, the normal one $\psi_N = \sum_{k=1}^{K} w_k \cdot \Phi^{-1}(\lambda_k)$ and the logistic one $\psi_L = \sum_{k=1}^{K} w_k \cdot \log[\lambda_k/(1 - \lambda_k)]$.

3. A further index to measure risk

In this Section we explain the combined index using some easily available variables, representing, at the moment, the main features in assessing risk and

systemic risk degree. In the literature, there are a lot of different measures to evaluate risk in a firm, that are often used in comparisons. We suggest to use all of them in order to get a more complete information about risk.

The obtained results will be compared with the *ranking* estimated by *V-Lab*, in order to evaluate the correspondences and the main differences for European Banks. In the first construction of $Y = \psi(\cdot)$, as defined in equation (1), we use the variables described in the following.

The first variable, X_1, denotes the *Marginal Expected Shortfall*, i.e. the expected loss (per dollar invested in capital) in which a company would occur with a fall market equal to 2%. Variable X_2 is *Beta*, technically the covariance between a firm's stock return, in our case *Eurostoxx50*, and the market, that is the main equity security of each institution, divided by the variance of market returns. The *Correlation* between the return of the share and the *Market Value Weighted Index*, representing the movement of the market in which changes in the price of the various stocks lead to the final value of the index in proportion to its value of market capitalization, is denoted by X_3. Variable X_4 that is the *Volatility*, the share capital of the company, is measured by the annualized standard deviation of returns based on daily returns. At last we consider X_5 as the indebtedness, *Leverage*.

The comparisons will be made with respect to *SRISK*, that is the measure of systemic risk of each institution over the global European risk, denoted with X_6; it is an estimate of the amount of recapitalization that a company needs not to fail in a financial crisis.

3.1. Case study

The dataset is composed by *N=103* financial institutions for which we observe *K=6* variables, described in the previous Section, and that we can get from *V-Lab* (data recorded in March, 2014). The first analysis concerns the linear relation among the involved variables. So we compute the correlation matrix, showed in Table 1. The tests performed on each pair of variables show the cases in which we can reject the hypothesis of null correlation.

Considering Table 1, we can note that in most cases the correlation is significatively different from zero. Only variable *Vol* may be considered uncorrelated to the other ones, in particular *Cor*, *Lvg* and *SRISK*.

This correlation structure, explaining the weak dependence among the considered variables, suggests us to use all of them in order to gain a better comprehension of the phenomenon. Unfortunately this is a complex system that needs to be reduced to one or at most to two dimensions, in order to make comparisons among different networks.

Table 1 – *Correlations and significativities.*

	MES	Beta	Correlation	Volatility	Leverage
Beta	0.999 (*)				
Correlation	0.708 (*)	0.710 (*)			
Volatility	0.372 (*)	0.374 (*)	-0.178		
Leverage	-0.292 (*)	-0.288 (*)	-0.366 (*)	0.122	
SRISK	0.380 (*)	0.384 (*)	0.406 (*)	0.093	0.486 (*)

*Note: **(*) at α=0.01**, without any adjustment for multiple tests.*

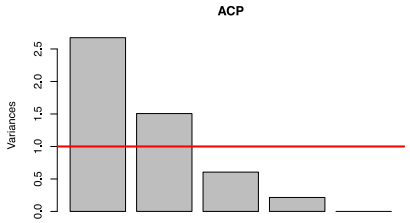
First of all we performed a transformation in principal components of the variables $X_1,\ X_2,\ X_3,\ X_4,\ X_5$, i.e. excluding variable *SRISK* that we want to use in comparisons, on the dataset without considering the time dimension and we obtained the data in Table 2.

Table 2 – *Variability of transformed components.*

Statistics	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6351	1.2271	0.7782	0.4633	0.0203
Proportion of Variance	0.5347	0.3011	0.1211	0.0429	0.0001
Cumulative Proportion	0.5347	0.8359	0.9570	0.9999	1.0000

So, following Kaiser rule (see the *screeplot* in Figure 1) and the cumulative proportion of variance (Table 2), we may think to use only the first two principal components, or at least even only the first one.

Figura 1 – *Screeplot.*



Looking at Figure 2, these ones are influenced in the following way: the first one positively by *leverage* and *volatility* dimensions, but not strongly, and negatively by the other ones, more strongly by *correlation*; the second principal component is influenced by all the variables, but *correlation*, in a negative way.

Figure 3 shows the institutions transformed according to the two first principal components, and, performing a statistical hierarchical clustering technique, two main groups are revealed (see the different colour and shape of the points).

Figura 2 – *Influences on the first two principal components.*

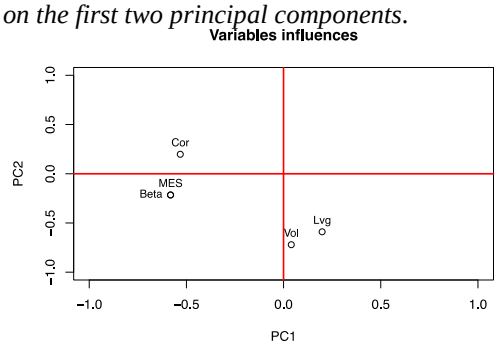
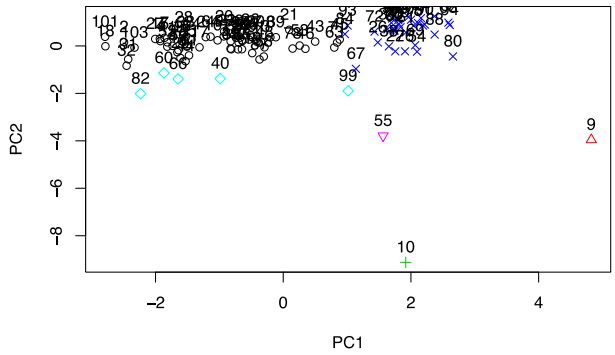


Figura 3 – *The first two principal components.*



To compute the combined index, first of all, the institutions are ranked in increasing order with respect to each variable. Then, the ranks are transformed in sample percentiles over the total number of observations and arranged in a 103×5 matrix. Obviously each X_k denotes a particular feature. So we use a method based on the nonparametric combination of dependent ranks.

Let X_{ki} denote the value of variable, $X_k, k = 1, \dots, K$, on unit i, with $i = 1, \dots, N$. Indicator function, $I(A)$, is equal to 1 if A is true and zero otherwise.

Then for each variable X_k we consider the following transformation

$$\eta_{ki} = \frac{\sum_{j=1}^{N} I(X_{ki} \geq X_{kj}) + 0.5}{N+1} \qquad (2)$$

Let's note the presence of constant values 0.5 and 1 in order to assure the absence of zero and 1 for variable η_k, avoiding problems of unboundedness for the combination function. This computation is performed for each *i* and *k*.

In such a way, we obtain a $K \times N$ matrix for values η_{ki}. Each column of the matrix, ordered in decreasing way, is a partial ranking.

Next step is gaining a global ranking. To this aim, for each row of the resulting matrix we apply the Fisher combination function, in which w_k are equal to 1,

$$\psi_i = -\sum_{k=1}^{K} \log(1 - \lambda_{ki}) \tag{3}$$

If we rank even this new variable, according to

$$R_i = \frac{\sum_{j=1}^{N} I(\psi_i \geq \psi_{ij})}{N}, \tag{4}$$

we obtain a vector of values that can be ordered in a decreasing way.

4. Analysis of comparison between indices

The new index estimates the risk of each European Bank within the financial system. Using *SRISK*, the companies position depends on the institution size and on the *Long-Run Marginal Expected Shortfall*. The last index explains an expected loss based on the real firm information; furthermore, it doesn't include other variables such as the correlation between the firm return and the market return, the annualized volatility of the capital requirements and the debt level of each company.

In order to test if the new combined index gets a similar ranking of *SRISK*, we consider the *Spearman* correlation index, *Rho* (Best and Roberts, 1975), defined as

$$Rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{5}$$

where d_i is the difference between the positions in the ranking for the two variables on the *i*-th unit. To test the significance of the correlation we consider the Spearman test

$$TS = \frac{\#(Rho^{obs} \geq Rho^{*})}{B} \tag{6}$$

in which Rho^{obs} denotes the coefficient computed on the observed ranking, Rho^{*} expresses the coefficients computed on each permutation and *B* is the number of resamplings.

Figura 4 – *Indices, ranked by SRISK and Combined.1 (left), by SRISK and Combined.2 (right).*

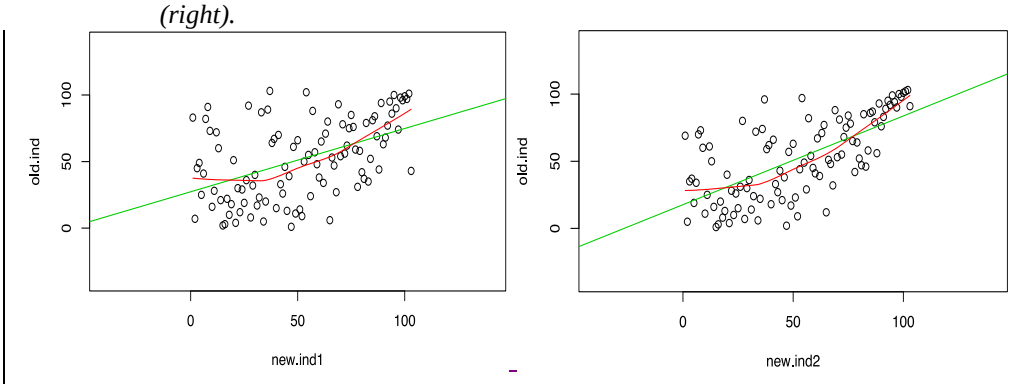


Table 3 – *Correlation tests.*

	Z_1	Z_1	r	Test statistic	p-value
1	SRISK	Combined.1	0.475	95592	0.00000
2	SRISK	Combined.2	0.660	61848	0.00000
3	SRISK	Prin.comp.1	0.366	115506	0.00016
4	Prin.comp.1	Combined.1	0.399	109398	0.00003

In the first row of Table 3, the *SRISK* and combined index rankings are compared. The obtained results in terms of correlation, that is significative but not so strong, allow us to define another index based on all the variables including *SRISK*, X_1, X_2, X_3, X_4, X_5 and X_6 (see the second row in Table 3). If we consider even this framework the correlation obviously increases, as now the combined index takes into account the variable to which the correlation will be computed, slightly improving the relation as shown in Figure 4 on the right.

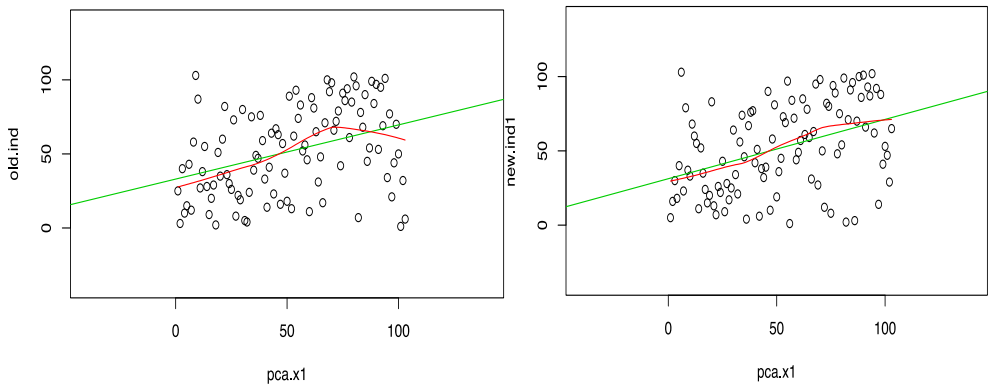
4.1. Comparison among indices

Table 4 shows all the comparisons in term of correlations of the four rankings (induced by *SRISK*, by the combined index including X_1, X_2, X_3, X_4, X_5 and called *Combined.1*, by the combined index including X_1, X_2, X_3, X_4, X_5, and X_6 and called *Combined.2*, and by the first component produced by PCA).

In all cases the correlations are not too strong but they are significatively different from zero. This confirms the idea that the problem is a complex one and cannot be reduced to only one variable. Furthermore, looking at the scatterplots in Figures 4 and 5, we can note that the linear lines (in green) are not very

representative of the conjoint behaviour between rankings, as the non parametric locally-weighted polynomial regressions (in red) are not overlapping.

Figura 5 – *Indices, ranked by SRISK and Prin.Comp.1 (left), by Prin.Comp.1 and Combined.1 (right).*



5. Conclusions

This study deals with the problem of measuring the risk and in particular of the systemic risk. As highlighted in the literature there are many definitions of risk and each definition suggests a different measure of it.

We stress the fact that each risk measure enables us to capture some feature of risk. Consequently, it rises the problem of obtaining a one-dimensional measure of such phenomenon using all the available variables giving some partial, even overlapping, information about it.

Our proposal uses a nonparametric combination of different statistical measures, based on the rankings of such measures, according to Pesarin's work (Pesarin and Salmaso, 2010). The approach seems interesting since it may be useful even when the rankings may be dependent.

The results seems promising since the nonparametric combination of the ranking of five variables performs as well as the *SRISK* index introduced by *V-Lab*.

As a future work we propose to use the combination only for the variables characterizing each institution, but not explicitly their riskiness, which instead may be represented by other uncorrelated ones to obtain a ranking. So two or more groups of companies may be identified using quantiles. Then, on these groups we test the combination permutation procedure proposed by Pesarin and Salmaso (2010).

Acknowledgements

The work is partially supported by PRIN project *Multivariate Statistical Models for Risk Assessment* (MIUR, code 2010RHAHPL_005).

The work is partially supported by *SYRTO project*, funded by the European Union under the 7th Framework Programme (FP7-SSH/2007-2013 - Grant Agreement SYRTO-SSH-2012-320270)

References

BILLIO M., GETMANSKI M., LO A., PELIZZON L., 2012. Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors, *Journal of Financial Economics*, Vol. 104, pp. 535-559.

DAS D., UPPAL R., 2004. Systemic risk and international portfolio choice, *The Journal of Finance*, Vol. 59, pp. 2809–2834.

DE BANDT O., HARTMANN P., 2000. Systemic risk: a survey, *European Central Bank*, No. 35.

EISENBERG L., NOE T.H., 2001. Systemic Risk in Financial Systems, *Management Science*, Vol. 47, No. 2, pp. 236-249.

EVERITT B.S., 2004. *An R and S-PLUS, Companion to Multivariate Analysis*. London: Springer-Verlag.

KAUFMAN G., 1994. Bank contagion: a review of the theory and evidence, *The Journal of Financial Services Research*, Vol. 8, pp. 123-150.

LAGO A., PESARIN F., 2000. Nonparametric combination of dependent rankings with application to the quality assessment of industrial products, *Metron*, Vol. 58, pp. 39-52.

PESARIN F., SALMASO L., 2010. *Permutation Tests for Complex Data: Theory, Applications and Software*. Chichester: Wiley & Sons.

RENCHER A. 2002. *Methods of Multivariate Analysis, 2nd ed.* New York: Wiley & Sons.

SUMMARY

Measuring Systemic Risk Through Statistical Combination

Although many different definitions of systemic risks are introduced in the literature, some scholars agree to consider that a measure of systemic risk should take into account the links among the institutions of a network.

The first consequence is the proposal of different synthetic indices built on the bases of some indicators, but this leads to use a procedure to summarize these indices in an unidimensional one.

In the present work we pay attention to the relations among the financial institutions; in particular, we propose an index combining several peculiar variables in order to rank the financial institutions in the network depending on their risk index. The used variables are those proposed by V-lab.

Moreover the combination technique may also be considered to perform nonparametric inference, to treat non gaussian distributions as in the case of indices.

So we propose to highlight systemic risk in a network of companies performing a nonparametric test to reveal a sort of heterogeneity behavior; in this case the rankings may also be used to create different behavioral groups.

Francesca PARPINEL, Department of Economics, Ca' Foscari University of Venice, parpinel@unive.it

Claudio PIZZI, Department of Economics, Ca' Foscari University of Venice, pizzic@unive.it

ESTIMATING THE LONG MEMORY PARAMETER IN NONSTATIONARY MODELS: FURTHER MONTE CARLO EVIDENCE

Margherita Gerolimetto

1. Introduction

In the last decades long memory time series models have been widely examined. Applications of these models can be found in several fields, *e.g.* hydrology, chemistry, economics and finance.

Several estimation techniques have been proposed in literature to detect the long memory phenomenon in both time and frequency domain (see for example Palma, 2007 for a review); they aim at estimating the long memory parameter *d* that incorporates the strength of the persistence. Most of the methods have been thought, in principle for the stationary case, i.e. their theoretical properties hold only when *d* is in $(-1/2,+1/2)$. Some recent simulation study have been carried out to compare the performance of the long memory parameter estimators in case of stationary models (among the others, Bouthahar *et al.*, 2007 and Tsay, 2009).

Here, we are interested in nonstationary long memory models, i.e. when the long memory parameter *d* no longer is in $(-1/2,+1/2)$, but it actually can be $\geq 1/2$. Broadly speaking, the issue of estimating nonstationary long memory has been addressed in two ways, either extending existing methods to estimate *d* to the case of nonstationarity (as in Velasco, 1999a and 1999b) or proposing new methods (resorting, for example, to wavelets as in Moulines *et al.*, 2008 and Boubaker and Péguin-Feissolle, 2013).

In this paper, we conduct a Monte Carlo experiment to show and compare the performance of a variety of estimators, traditionally conceived for stationary models, of the long memory parameter *d* in case of nonstationarity. On purpose, we did not focus on new-generation estimators, but did concentrate on traditional estimators, belonging to three group. Among (*i*) heuristic estimators, we consider the Higuchi method (1988), the aggregate variance method (1995) and Lo (1991) method. Among (*ii*) parametric estimators, we consider Whittle method (Fox and Taqqu, 1986) and among (*iii*) semiparametric methods, we consider the GPH method by Geweke and Porter-Hudak (1983) and its modified version by Smith (2005). All these methods have been employed on both the original time series and first difference of the series. This is done to include in the analysis an idea by

Hurvich and Ray (1995) who propose, in case of nonstationarity, to estimate *d* on the first difference of the series, i.e. on the series made stationary so that the estimators are expected to work again in the range of *d* where their properties are guaranteed.

Results of the Monte Carlo experiment show that the Whittle estimator has the best performance in case of nonstationarity, followed by the GPH. Moreover, the strategy of preliminarily differentiate the series helps improve the results.

The structure of the paper is as follows. In the second section we will briefly recall the most important characteristics of the long memory models. In the third section we present the estimators of the long memory parameter we will study. The fourth section is devoted to the Monte Carlo experiment and some conclusions.

2. Long memory models

Usually, a long memory model X_t can be characterized by a single memory parameter $d \in (-1/2,+\,1/2)$, called degree of the memory, which controls the shape of the spectrum near zero frequency and the hyperbolic rate decay of its autocorrelation function. More precisely the spectral density $f(\lambda)$ of the long memory model is approximated in the neighborhood of the zero frequency by

$$f(\lambda) \sim c\lambda^{-2d} \; as \; \lambda \to 0^{+}, \; 0 < c < \infty \tag{1}$$

Thus $f(\lambda) \to \infty$ as $\lambda \to 0^{+}$. Under additional regularity assumptions of $f(\lambda)$, the autocorrelation function $\rho(k)$ of the long memory model has the following asymptotic behavior:

$$\rho(k) \sim ck^{2d-1} \; as \; k \to \infty \tag{2}$$

these features characterize ARFIMA(*p*,*d*,*q*) models[1] (Granger and Joyeux, 1980)

$$\Phi(B)\Delta^{d}X_t = \Theta(B)\varepsilon_t \tag{3}$$

of which the fractional noise is a special case

$$\Delta^{d}X_t = \varepsilon_t \tag{4}$$

[1] ARFIMA models are a generalization of ARIMA, where *d* is not integer.

The properties of these models depend on the long memory parameter value *d*. More, specifically, an ARFIMA(*p*,*d*,*q*) model is stationary and invertible when $d \in (-1/2,+\,1/2)$, usually this interval is reduced to $(0,1/2)$ When $d \geq 1/2$ the ARFIMA is nonstationary, although for $d \in [1/2, 1)$ it is mean-reverting, meaning that there is no long-run impact of an innovation on the value of the process. When $d \geq 1$ mean-reversion does not longer hold. Clearly, the case $d = 0$ and $d = 1$ (i.e. shot memory stationarity and unit root) are encompassed as particular cases of a more general parametrization.

3. Estimation techniques for ARFIMA processes

Now we briefly describe the methods we will consider in our Monte Carlo experiment to estimate the long memory parameters. For space reason we will not be able to go in details about the methods and refer to the original papers. It is possible to group these methods in three categories: heuristic, parametric and semiparametric methods.

Among heuristic methods, we consider: (*a*) Higuchi method (Higuchi, 1988) which measures the fractal dimension of a non-periodic and irregular time series; (*b*) the aggregate variance method (Fox and Taqqu, 1995) that concentrates on the behavior of the variance of the sample mean and (*c*) the rescaled range (R/S) method, Lo (1991), which studies the behavior of the partial sums of deviation of the series from its sample mean.

As for parametric methods, we consider Whittle method (Fox and Taqqu, 1986). Given the ARFIMA(*p*,*d*,*q*) in (3), the vector of parameters $\boldsymbol{\theta} = (d, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)$ is estimated via the Whittle approximation of the log-likelihood by minimizing with respect to θ:

$$\hat{\sigma}^2(\theta) = \frac{1}{2}\sum_{j=1}^{T'} \frac{I(\lambda_j)}{f(\lambda_j)} \tag{5}$$

where *T*' is the integer part of $\frac{T-1}{2}$ and $I(\lambda_j)$ and $f(\lambda_j)$ are, respectively, the periodogram and the spectral density at the Fourier frequencies. The Whittle method has several theoretical and practical advantages. However, its disadvantage is in that the parametric form of the spectral density is assumed to be known a priori.

Among the semiparametric methods we consider the GPH estimator (Geweke and Porter-Hudak, 1983). The advantage in resorting to these methods is that there is no need to specify the entire model since the only necessary information is the

behavior of the spectral density near the origin. Given an Given and ARFIMA(*p*,*d*,*q*) as in (), its spectral density is:

$$f(\lambda) \sim c_f \left(4\, sin^2\left(\frac{\lambda}{2}\right)\right)^{-d} \quad \lambda \to \infty \tag{6}$$

As the periodogram $I(\lambda)$ is an asymptotically unbiased estimator of $f(\lambda)$, it is possible to estimate *d* by running the OLS regression:

$$\log\big(I(\lambda_j)\big) = \log c_f + \beta \log\left(4\, sin^2\left(\frac{\lambda}{2}\right)\right) + \varepsilon_j \tag{7}$$

The former is an asymptotic relationship that holds only in a neighborhood of the origin. This means that, considered over all the ordinates of the periodogram (-π,+ π), it would produce highly biased estimates. Consequently, GPH is so actually calculated by running the least squares regression only for the *m* lowest frequencies. We also considered Smith's (2005) modified version of the GPH method that takes into account the approximation he derived of the bias.[2]

As anticipated, the 5 methods have been thought for the stationary setting. The theoretical properties no longer hold in case of nonstationary, or, in case they do, it is only for a limited interval.[3] In case of nonstationarity the relative recent literature is rich of contributes along two directions. There are estimators that adapt existing methods in order to gain asymptotic properties also in case of nonstationarity; among these we can mention the tapered versions of the GPH or the Whittle method (Velasco and Robinson 1999a, 1999b). There are also brand-new methods, e.g. wavelet based estimators (McCoy and Walden, 1996; Moulined et al. 2008).

As for the brand-new methods, it should be stressed that often these methods are much more sophisticated (and complicated) than the existent. For this reason, in this paper we study, via Monte Carlo simulations, the effective performance of the traditional methods in case of non stationary long memory, also when they are employed on the first difference of the time series (now stationary) following Hurvich and Ray (1995).

[2] Actually, in our Monte Carlo experiment, the performance of the version of the GPH estimator modified by Smith (2005) is not particularly good.

[3] For example the Whittle estimator is shown to possess asymptotic properties for $1/2 < d < 3/4$, included asymptotic normality. The same holds for the GPH.

4. Monte Carlo experiment

In this section we present the Monte Carlo experiment we conduct to show and compare the performance of the 5 estimators of the long memory parameter described in the previous sections: Higuchi, Aggregate Variance, Lo, GPH, GPH modified by Smith (GPH-S, hereafeter) Whittle. The Data Generating Process (DGP, hereafter) we consider is the fractional noise, ARFIMA (0,d,0), for various values of the long memory parameter.

In particular, we considered three scenarios. In the first we consider stationary DGPs and we simulate fractional noise with d= 0.1,0.2,0.3,0.4. This scenario is included in the MC experiment with the role of benchmark, given that all the long memory parameter estimators should have a good performance in this case. In the second scenario, we generate time series with d= 0.6,0.7,0.8,0.9, i.e. nonstationary but mean reverting long memory. In the third scenario, we study nonstationary and not mean-reverting long memory, done by generating time series data with d=1.1,1.2,1.3,1.4. Over all cases, the innovation is $\varepsilon_t \sim N(0,1)$, the sample size considered are T=250,500,1000 for 2000 Monte Carlo simulations. All series are generated with 200 additional values in order to obtain random starting values. The performance of the estimators is expressed in terms of mean squared error (MSE) across Monte Carlo simulations.[4]

The results of the experiment are reported in the tables 1-5. In Table 1 we present the MSE for the stationary case.

Table 1 – *Stationary long memory: Monte Carlo MSE*

T	d	R/S	Aggr Var	Higuchi	GPH	GPH-S	Whittle
250	0.1	0.0117	0.009	0.014	0.0494	0.1672	0.0031
	0.2	0.0137	0.0126	0.0144	0.0526	0.1653	0.0034
	0.3	0.0162	0.0163	0.0173	0.0476	0.1617	0.0033
	0.4	0.0207	0.0214	0.0141	0.0452	0.1804	0.0034
500	0.1	0.0085	0.0073	0.0116	0.0296	0.0869	0.0013
	0.2	0.0109	0.0087	0.0132	0.0296	0.0902	0.0016
	0.3	0.0131	0.0118	0.0153	0.029	0.091	0.0015
	0.4	0.0158	0.0174	0.0147	0.0303	0.09	0.0015
1000	0.1	0.0069	0.0057	0.012	0.0193	0.0193	0.0007
	0.2	0.0092	0.0076	0.0135	0.0196	0.0196	0.0007
	0.3	0.011	0.0099	0.0141	0.0188	0.0188	0.0007
	0.4	0.0149	0.0158	0.0137	0.0194	0.0194	0.0007

[4] For the GPH the estimation has been conducted setting *m* equal to the square root of the sample size, as suggested in the original article by Geweke and Porter-Hudak (1983).

In Table 2 we present the MSE results for the nonstationary mean reverting case, both for the original series (upper panel) and the first difference of the series (lower panel).

Table 2 – *Nonstationary (mean reverting) long memory: Monte Carlo MSE (original series upper panel, first differenced series lower panel*

T	d	R/S	Aggr Var	Higuchi	GPH	GPH-S	Whittle
		Original series					
250	0.6	0.0393	0.0529	0.0323	0.0511	0.1696	0.0035
	0.7	0.0567	0.0846	0.0645	0.0474	0.1385	0.0036
	0.8	0.0872	0.13	0.1155	0.0464	0.1461	0.0037
	0.9	0.125	0.1965	0.1858	0.0454	0.1465	0.0033
500	0.6	0.0325	0.047	0.0352	0.0308	0.0931	0.0016
	0.7	0.0509	0.0791	0.0656	0.0316	0.0943	0.0019
	0.8	0.0838	0.1278	0.1192	0.031	0.0903	0.0019
	0.9	0.1267	0.1899	0.1861	0.028	0.0805	0.0017
1000	0.6	0.0314	0.0444	0.0339	0.0202	0.0202	0.0008
	0.7	0.0517	0.0767	0.0667	0.0188	0.0188	0.0009
	0.8	0.0837	0.1237	0.116	0.0204	0.0204	0.0012
	0.9	0.1323	0.1888	0.1863	0.0184	0.0184	0.0011
		First differenced series					
250	0.6	0.0434	0.0102	0.0084	0.0488	0.1672	0.0027
	0.7	0.0275	0.0073	0.0065	0.05	0.1561	0.0035
	0.8	0.0184	0.0064	0.0074	0.0464	0.1678	0.0031
	0.9	0.0126	0.0076	0.01	0.0517	0.1736	0.0037
500	0.6	0.0337	0.0067	0.0062	0.0313	0.0948	0.0014
	0.7	0.0201	0.0044	0.005	0.0295	0.0894	0.0014
	0.8	0.0123	0.0042	0.0063	0.0314	0.0915	0.0013
	0.9	0.0085	0.0048	0.0076	0.0288	0.0892	0.0016
1000	0.6	0.0259	0.0046	0.0046	0.0199	0.0199	0.0007
	0.7	0.0147	0.003	0.0044	0.0187	0.0187	0.0007
	0.8	0.0088	0.0033	0.0056	0.0175	0.0175	0.0007
	0.9	0.0065	0.004	0.0073	0.018	0.018	0.0007

In Table 3 we present the MSE results for the nonstationary not mean-reverting case, both for the original series (upper panel) and the first difference of the series (lower panel).

Table 3 – *Nonstationary (nont mean reverting) long memory: Monte Carlo MSE (original series upper panel, first differenced series lower panel*

T	d	R/S	Aggr Var	Higuchi	GPH	GPH-S	Whittle
		Original series					

250	1.1	0.2613	0.3867	0.3807	0.0396	0.1284	0.0066
	1.2	0.332	0.5103	0.5095	0.0487	0.1295	0.0242
	1.3	0.448	0.6576	0.6525	0.0821	0.1314	0.0682
	1.4	0.5142	0.8236	0.8207	0.1398	0.177	0.1353
500	1.1	0.2663	0.3778	0.3841	0.0223	0.0607	0.0051
	1.2	0.3766	0.5039	0.5077	0.0377	0.0745	0.0236
	1.3	0.4989	0.6494	0.6571	0.071	0.0912	0.0651
	1.4	0.5861	0.8147	0.8217	0.1348	0.1446	0.1402
1000	1.1	0.2994	0.3789	0.384	0.0193	0.0193	0.0046
	1.2	0.3815	0.5053	0.5116	0.0309	0.0309	0.0236
	1.3	0.5208	0.6503	0.6564	0.0681	0.0681	0.0683
	1.4	0.6095	0.8162	0.8213	0.1352	0.1352	0.1413
		First differenced series					
250	1.1	0.0129	0.009	0.0144	0.0488	0.1752	0.0033
	1.2	0.0143	0.011	0.0148	0.0453	0.1727	0.0031
	1.3	0.0157	0.0141	0.0156	0.0466	0.1672	0.0034
	1.4	0.0208	0.0214	0.0188	0.046	0.1596	0.0033
500	1.1	0.0089	0.007	0.0112	0.0279	0.0952	0.0015
	1.2	0.0106	0.0077	0.0136	0.0278	0.0897	0.0014
	1.3	0.0135	0.0111	0.0146	0.0301	0.0941	0.0014
	1.4	0.0182	0.0178	0.0142	0.0309	0.0933	0.0015
1000	1.1	0.0076	0.0061	0.0119	0.0183	0.0183	0.0007
	1.2	0.0093	0.0074	0.0126	0.0194	0.0194	0.0007
	1.3	0.0118	0.0105	0.0144	0.019	0.019	0.0007
	1.4	0.0144	0.0156	0.0127	0.0188	0.0188	0.0007

From Table 1 (stationarity case) we can observe that while *d* is far from the nonstationarity area, almost all estimation methods have a low level of MSE, also at relatively small sample sizes. It is in particular when *d* gets close to the bound ½ that it is possible to appreciate the better performance of the Whittle method, followed by the GPH and Higuchi methods, as the other methods worsen their performance visibly.

In Table 2, upper panel, we observe for the majority of methods the process of worsening of the MSE performance with the increase of *d* continues. Only for Whittle and GPH estimators the performance is steadily good, more precisely not only they are the methods with the best performance, but also their MSE level stays approximately at the same level as in Table 1. This means that the two methods do not suffer excessively from the lack nonstationarity (probably because mean-reversion still holds). In general, things improve when all 5 methods are applied to the first difference of the time series (lower panel of Table 2). However, we note that for Whittle method in particular, there seems to be no relevant difference from

upper panel and lower panel, leading us to believe that for this methods differencing is not necessary.

In Table 3 we study the nonstationary and non mean-reverting scenario. In this case, for all methods this is a rather difficult task because we are very far from the area where the theoretical properties hold. Taking the first difference (lower panel) leads to quite better results, especially if Whittle and GPH are adopted. So in this case, taking the first difference seems to be really a reasonable option, that leads to good MSE performance (in line with the stationary case magnitude order), especially if Whittle and GPH are used.

In Table 4 and 5 we present for the nonstationary (respectively mean reverting and non-meanreverting) case, the ratio of the MSE of the estimate on the original series and on its first difference. These Tables help emphasize the effective improvement in adopting the first difference and under which conditions this happens.

Table 4 – *Nonstationary (mean reverting) long memory: ratio of Monte Carlo MSE of the estimate on the original series and on first differenced series*

T	d	R/S	Aggr Var	Higuchi	GPH	GPH-S	Whittle
250	0.6	0.547	2.975	2.2985	0.9141	0.8133	0.1221
	0.7	0.986	6.9778	6.5715	1.9012	2.0766	0.1789
	0.8	1.6752	22.0144	18.624	3.7096	3.0392	0.5752
	0.9	3.0927	54.9223	64.572	1.6946	1.557	0.0138
500	0.6	0.5446	3.2609	2.7961	0.963	0.7935	0.5118
	0.7	1.0532	9.1372	9.017	2.3609	2.6762	1.0504
	0.8	1.9897	41.7805	60.1748	6.8971	4.3158	1.1782
	0.9	3.7907	68.415	57.1477	3.2824	2.158	0.7604
1000	0.6	0.5768	3.8248	3.3794	1.1626	1.1626	1.3785
	0.7	1.1811	10.903	11.7163	1.787	1.787	3.8535
	0.8	2.3524	56.6822	577.6202	3.3903	3.3903	2.4796
	0.9	4.5605	58.3148	50.1638	18.44891	18.4489	2.7687

When the figures in the Tables are smaller than 1, this means that the MSE coming from the estimate on the first differenced time series is larger than that on the original series. On the contrary, the larger the figures are with respect to 1, the more recommended is to estimate *d* on the first difference of the time series.

As expected, in Table 4, regarding nonstationary mean-reverting time series, the figures have an oscillatory behavior around 1, especially for the Whittle method, thus confirming what emerged from Table 2, i.e. if $\frac{1}{2} < d < 1$ the effects of nonstationarity are non that severe and, consequently, is not so relevant (sometime

not even visible) the improvement in the performance obtained by adopting the strategy of taking the first difference of the series before estimating *d*.

Table 5 – *Nonstationary (not mean reverting) long memory: ratio of Monte Carlo MSE of the estimate on the original series over first differenced series*

T	d	R/S	Aggr Var	Higuchi	GPH	GPH-S	Whittle
250	1.1	16.79	16.4638	26.3091	41.9561	10.7115	3.2634
	1.2	93.086	12.4855	25.3358	13.9063	4.5804	8.2807
	1.3	16.3558	9.3688	23.0359	22.4873	18.5583	13.2683
	1.4	9.8334	7.3744	18.1095	35.0908	267.265	25.7747
500	1.1	21.9624	17.3914	36.1323	34.2809	7.09977	5.2897
	1.2	253.9619	14.8114	26.0578	44.379	7.83361	17.8265
	1.3	23.1643	10.5538	25.0266	53.8314	414.2553	27.0515
	1.4	11.1353	8.12	24.4066	59.2308	31.3384	45.5639
1000	1.1	90.7188	18.8324	24.8224	14.458	14.458	11.8739
	1.2	86.4642	14.6381	30.2226	32.4125	32.4125	27.2745
	1.3	20.3471	10.7909	26.324	144.1566	144.1566	44.5244
	1.4	13.7794	8.5732	28.0049	25.8235	25.8235	205.3707

In Table 5, instead, all figures are systematically larger than 1. This is because for all considered estimation methods (even for the Whittle), the performance hugely improves in case the first difference is preliminarily taken. Once more, this is in line with the previous results, in particular those shown in Table 3. The effects of nonstationarity are very severe and, consequently, it is significant the improvement in the performance obtained by adopting the strategy of taking the first difference of the series before estimating *d*.

5. Conclusions

To conclude, in this work we present a Monte Carlo study to show and compare the performance of some traditional and well-known estimator of the long memory parameter in case of nonstationary fractional noise models. We are aware that in the literature recently has been proposed a variety of methods for estimating the long memory parameter in the nonstationary case, yet we are interested in how the traditional methods perform in case the first difference of the series is taken and in this work we intend to fill this literature gap.

The simulation results show that, among the traditional methods the Whittle estimate (followed by the GPH) is the best performing in terms of Monte Carlo MSE and this holds also when stationarity no longer holds, in particular if mean-

reversion is preserved. Indeed, if the nonstationary time series is mean reverting the performance of the Whittle estimator is comparable with the stationary case and there seem to be no special need to preliminarily take the first difference. Instead, when the nonstationarity is so strong that mean-reversion is lost and all methods perform badly, working with first difference of the time series (in particular estimating with Whittle method) is recommended.

To sum up, we conclude that in several cases it could be that there is no need to resort to sophisticated (and difficult to implement) methods for estimating nonstationary long memory. It may happen that taking the first difference of the series and then proceed with the traditional estimators, especially Whittle estimator is a good enough strategy to obtain reliable estimates of the long memory parameter in the nonstationary hypotheses.

Future research on this topic is in order with the aim of extending the simulation experiment so that also new-generation method, such as wavelets methods will be included.

Riferimenti bibliografici

BOUBAKER H., PÉGUIN-FEISSOLLE A. 2013: Estimating the long memory parameter in nonstationary processes using wavelets, *Computational Economics*, Vol. 42, pp. 291-306.

BOUTHAHAR M., MARIMOUTOU V., NOUIRA L. 2007: Estimation metods of the long memory parameter: Monte Carlo analysis and application, *Journal of Applied Statistics*, Vol. 34, pp. 261-301.

HURVICH C.M., RAY B.K. 1995: Estimation of the long memory parameter for nonstationary or non invertible fractionally integrated processes, *Journal of Time Series Analysis*, Vol.16, pp. 17-41.

FOX R., TAQQU M. 1986. Large sample properties of parameter estimates for strongly dependent stationary gaussian time series, *The Annals of Statistics*, Vol. 14, pp. 517-532.

GEWEKE J., PORTER-HUDAK S. 1983. The estimation and application of long-memory time series models, *Journal of Time Series Analysis*, Vol.4, pp. 221-237.

GRANGER C.W.J., JOYEUX, R. 1980. An introduction to long-range time series models and fractional differencing, *Journal of Time Series Analysis*, Vol.1, pp. 15-30.

HIGUCHI T. 1988. Approach to an irregular time series on the basis of the fractal theory, *Physica D*, Vol. 31, pp. 277-283

LO A.W. 1991: Long-term memory in stock market prices, *Econometrica,*Vol. 59, pp. 1279-313.

McCOY E.J., WALDEN A.T. 1996. Wavelet analysis and synthesis of stationary long memory processes, *Journal of Computational and Graphical Statistics*, Vol. 5, pp. 26-56.

MOULINES E., ROUEFF F., TAQQU M.S. 2008. A wavelet whittle estimator of the long memory parameter gaussian time series, *The Annals of Statistics*, Vol. 36, pp. 1925-1956.

PALMA W. 2007. *Long memory time series: theory and methods*. New York: Wiley.

SMITH A. 2005. Level shifts and the illusion of long memory in economic time series, *Journal of Business and Economic Statistics*, Vol. 23, pp. 231-335.

TAQQU M., TEVEROVSKY V., WILLINGER W. 1995. Estimators for long range dependence: an empirical study, *Fractals*, Vol. 3, pp. 785-798.

TSAY W.J. 2009: Estimating long memory time series cross-section-data parameter, *Electoral Studies*, Vol. 28, pp. 129-140.

VELASCO C. 1999a. Nonstationary logperiodogram regression, *Journal of Econometrics*, Vol. 91, pp. 325-371.

VELASCO C. 1999b. Gaussian semiparametric estimation of nonstationary time series, *Journal of Time Series Analysis*, Vol. 20, pp. 87-127.

SUMMARY

Estimating the Long Memory Parameter in Nonstationary Models: Further Monte Carlo Evidence

In this work we perform a Monte Carlo experiment to show and compare the performance of a variety of estimators of the long memory parameter d in case of nonstationary processes. Both parametric and semiparametric estimators are considered. Moreover they have been employed both on the original time series and on the first difference of the series. Results show that the Whittle estimator is the best performing and the strategy of preliminarily differentiate the series is worthy, but not for all the estimators.

Margherita GEROLIMETTO, Professore Associato,
margherita.gerolimetto@unive.it

DIAGNOSTIC TOOLS BASED ON OPTIMAL RANKING IN THE COX MODEL

Luciano Nieddu, Cecilia Vitiello

1. Introduction

Cox proportional hazard model (Cox, 1972) is the most popular method for assessing covariates effects on time to event in the presence of censoring. Its contribution to the analysis of real data spread firstly in medical studies and from then it stemmed to become a relevant tool in various fields of application, such as economics, industrial reliability, agriculture, biological and physical sciences.

Cox regression and its extensions are today the standard for time to event data analysis. The wide spread of the methodology implies an increasing interest in diagnostic measures to assess the correct specification of the model, to evaluate goodness of fit and to detect outliers.

The nature of the data in survival analysis and the semiparametric structure of Cox model do not allow for a direct extension of the usual diagnostic tools from linear regression and generalized linear models. Its peculiarities make it hard to define a general diagnostic tool, thus the plethora of definition of residuals related to the Cox model.

In this work we suggest a diagnostic methodology based on the conditional contribution to the partial likelihood that is developed in analogy to standard tools in linear regression model.

The paper is structured as follows: in Section 2 we describe briefly the context in which our proposal is embedded, in Section 3 we define the conditional predicted residual in Cox model, Section 4 will present an exploratory simulation study to test the proposal while in Section 5 it will be tested on the famous kidney dataset (Collet, 2003). Finally in Section 6 some conclusions will be drawn.

2. Background

To make the paper self-contained the proportional hazard regression model proposed by Cox will be briefly recalled in this section.

In Cox model (Cox, 1972) and in survival analysis in general, (Kalbfleisch and Prentice, 2002), time is the dependent variable and ascertaining its dependence on covariates is the main goal of the analysis. We assume we have n cases and for each of them we consider two random variable: $Y_i, i = 1, \dots, n$, is the real time to event for this subject and $C_i, i = 1, \dots, n$ is a random variable representing the censoring process. In survival analysis C_i is assumed to be independent of Y_i, or less restrictively non informative, with respect to Y_i. On each unit we observe the vector (t_i, δ_i, z_i), t_i is the time until failure if the event has been observed ($\delta_i = 1$), or the time until censoring ($\delta_i = 0$), if the unit has been removed from the study for reasons not related to the event of interest.

The proportional hazard model represents the hazard function as the product of two components: a non-parametric part representing the hazard of an hypothetical subject with all covariates identically equal to zero (this component contains all the information of the effect of time on the hazard function), and a term, constant with respect to time, that accounts for the effects of the covariates.

The hazard function can then be written as:

$$\lambda(t) = \lambda(t, \boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T \boldsymbol{z})$$

Where $\boldsymbol{\beta}$ is a p-vector of unknown parameters, $\mathbf{z}$ is a vector of known covariates, $\lambda_0(\cdot)$ is usually referred to as baseline hazard. This is a proportional hazard model since the covariates act multiplicatively on the hazard; its semiparametric nature is due to the complete non parametric specification of the baseline hazard function and the complete parametric form of the term in which covariates are involved.

Estimation in Cox model is based on the partial likelihood function which is the part of the full likelihood independent of the underling baseline hazard. Estimates obtained via the maximization of the partial likelihood have been shown to have similar features of those obtained using full likelihood although some information is lost, not negligible for small data set and informative censoring.

The partial likelihood $L(\boldsymbol{\beta})$is defined as follows: let $t_{(1)}, t_{(2)}, \dots t_{(d)}$ be ordered observed failure times, $(d \leq n)$ in increasing order and $\phi_k = i$ if the $i - th$ subject fails at t_k then

$$L(\boldsymbol{\beta}) = \Pi_k Pr(\phi_k = i|H_k) = \Pi_k \frac{exp(\boldsymbol{\beta}^T \boldsymbol{z}_k)}{\Sigma_{l \epsilon R_k} exp(\boldsymbol{\beta}^T \boldsymbol{z}_l)}$$

Where H_k contains all the information on the history of both events and censoring up to the $k - th$ event.

Several goodness of fit tests have been proposed to evaluate global and local fitting, correct specification of the model, in its general hypothesis of proportionality and in the specification of its regressive components. Among these, martingale residuals (Barlow and Prentice, 1988) are worth mentioning, since they will be used as reference during the study. They have a skewed distribution and are defined as the difference between the indicator variable δ_i and the cumulative hazard assigned by the model to an individual with failure time t_i (Therneau, Grambsch, Fleming 1990)

$$M_i = \delta_i - \Lambda_i$$

where $\Lambda_i = \int_0^{t_i} \lambda_i(u)du$.

High values for $\Lambda_i(t_i)$, indicate high cumulative risk of death, therefore a highly negative martingale residual.

Martingale residual properly transformed in order to gain symmetricity and nearly normality are generally called *deviance residuals*

$$dev(M_i) = sgn(M_i)\sqrt{-2\big(M_i + \delta_i ln(\delta_i - M_i)\big)}$$

Despite their name, $dev(M_i)$, as defined above, are not real *deviance residuals.* Standard deviance residuals, as introduced by Pregibon in generalized linear model, assume the meaning of a discrepancy measure between estimated and observed values. This discrepancy is evaluated by comparing the contribution to the likelihood for each single observation in the current model and in the full model, while *deviance residual* for the Cox model as have been defined above, do not refers directly to any likelihood, neither the complete one, nor the partial one on which model fitting and estimates are based upon. The reason why they are called *deviance residuals* is that their definition resembles deviance residual in Poisson regression, but, as a nuisance parameter, the unspecified baseline hazard is still involved.

While martingale residuals and therefore *deviance residuals*, are evaluated at time conditional on covariates, Shoenfield (Schoenfeld, 1982) residuals compare the observed values of covariates with the corresponding expected value returned by the model.

$$r_{sch_i} = X_i(t_i) - E(X_i(t_i)|\beta)$$

The result will be a matrix with as many rows as events and a column for each covariate. Shoenfield residuals are evaluated with respect to partial likelihood on a covariate scale and are not defined for censored units.

3. Conditional predicted residual in Cox model

Our proposal, starting from the standard deviance residual definition, suggests a residual based on the contribution of each single unit to the partial likelihood.

A first step is therefore to determine the contribution to the likelihood when dealing with partial likelihood.

When independence holds, the standard likelihood is a product of terms, each associated to a single unit, therefore the contribution to the likelihood is easily identified. Instead partial likelihood is factorized over time or, equivalently with respect to the risk set. Then the contributes of each unit to the partial likelihood is present in more than one term in the factorization; nonetheless they are still well defined and not independent because they are conditional to the process H_k, the history of the process till the $k-th$ event.

Namely, a subject that experienced the event at time $t_{(k)}$ contributes to the likelihood with the product of two terms, the first is its contributes at time $t_{(k)}$

$$\frac{exp(\boldsymbol{\beta}^T \boldsymbol{z}_k)}{\Sigma_{l \epsilon R_k} exp(\boldsymbol{\beta}^T \boldsymbol{z}_l)}$$

But it also contributes via its presence in the risk set of the previous events. This contributes are summarized in a product of $k-1$ factors, as follows

$$\Pi_{s=1}^{k-1} \frac{\Sigma_{l \epsilon R_s, l \neq k} exp(\boldsymbol{\beta}^T \boldsymbol{z}_l)}{\Sigma_{l \epsilon R_s} exp(\boldsymbol{\beta}^T \boldsymbol{z}_l)}$$

The last term is defined for all observations even for censored ones. According to the structure of the partial likelihood the contribution of each unit depends on the ranking it occupies with respect the sequence of events. Time between events is not involved directly or indirectly.

Due to the semi-parametrical nature of Cox proportional hazard model, it is not possible apply the notion of saturated model in this framework. In fact if we define a saturated model as one that reproduces exactly the observed data, this requirement is satisfied by a null model with only the non parametric component. On the contrary, if we define it as the one that has one parameter for each observation, the structure of the model itself does not allow to reproduce the whole information. In both cases we will have to refer to the baseline function that is actually not involved in partial likelihood.

We suggest to compare the contribute to the partial likelihood,

$$\left(\frac{\exp(\beta^T z_k)}{\Sigma_{l\epsilon R_k}\exp(\beta^T z_l)}\right)^{\delta_k}\Pi_{s=1}^{k-1}\frac{\Sigma_{l\epsilon R_s, l\neq k} exp(\beta^T z_l)}{\Sigma_{l\epsilon R_s} exp(\beta^T z_l)}$$

With a conditional maximum contribute

$$\left(\frac{\exp(\beta^T z_{k^*})}{\Sigma_{l\epsilon R_{k^*}}\exp(\beta^T z_l)}\right)^{\delta_k}\Pi_{s=1}^{k^*-1}\frac{\Sigma_{l\epsilon R_s, l\neq k^*} exp(\beta^T z_l)}{\Sigma_{l\epsilon R_s} exp(\beta^T z_l)}$$

Where k^* is the predicted rank for the $k-th$ unit, estimated on the base of the other $n-1$ units, i.e. the rank that maximized the contribution to the partial likelihood for that unit conditional to the other $n-1$; risk set R above are adjusted to the corresponding ranking.

This proposal gets further support from an analogy with a similar quantities in normal linear model. In fact it can be looked at as a generalization of predicted residuals that compare observed value with estimated value to which the considered unit has not contributed to.

4. An Exploratory Simulation Study

A brief simulation study to test the ability of the proposed method to detect outliers will be presented. Datasets of *n*=40 randomly generate units will be considered. For each unit two covariates, one quantitative (X_1) and one qualitative (X_2), will be randomly drawn according to the following scheme:

$$X_{i,1} \sim N(\mu = 0, \sigma = 1) \quad X_{i,2} \sim Bin(1, \theta = 0.5) \quad i = 1, \dots, n$$

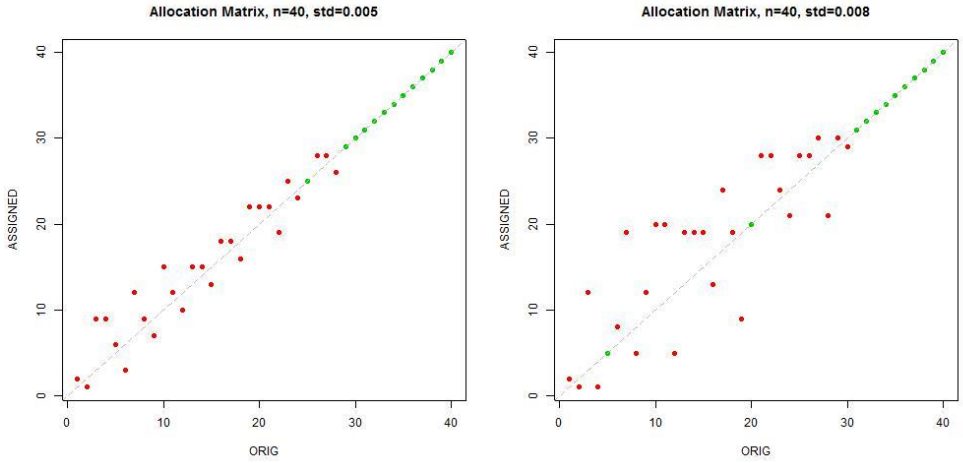
Times to event have been generated according to:

$$T_i = abs\left[\frac{1}{\boldsymbol{x}_i'\boldsymbol{\beta}} + N(0,\sigma)\right],\ \sigma = \{0.005, 0.008\}, \boldsymbol{x_i} = \left(X_{i,1}; X_{i,2}\right), \boldsymbol{\beta} = (1;2) \qquad (1)$$

Generating times to event according to equation (1) allows the same Gaussian noise to have different effect according to the value of the linear predictor, namely the same noise will have a stronger effect on units with large values of the linear predictor, therefore simulating the presence of possible outliers in the same dataset. In Figure 1 the allocation matrices for the optimal ranking for both levels of noise have been displayed. The elements in the dataset that already have a ranking optimal conditional to the remaining *n-1* have been displayed in green; red has

been used to depict units that are not allocated in an optimal position according to the estimated model. The green dots are mainly associated with small values of the linear predictor, for whom the noise is less effective. As it was to be expected, a smaller level of noise implies a smaller displacement needed to optimize the ranking of the data; with increasing noise the ranking that the units have in the dataset are less and less optimal.

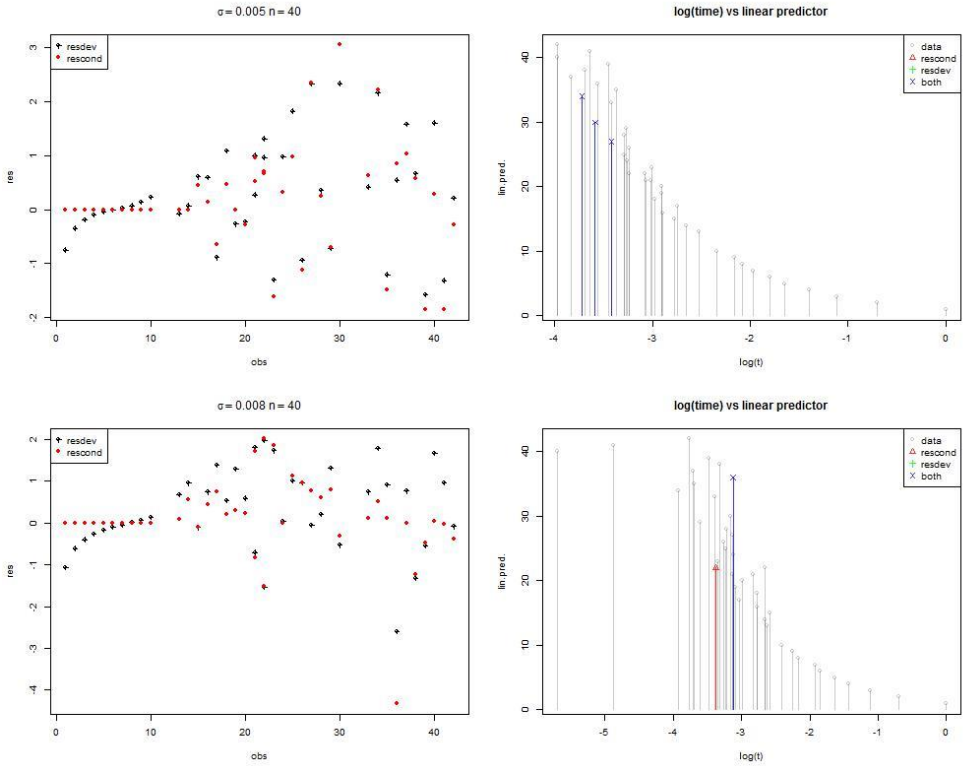
Figure 1 – *Allocation matrices for the simulated data with increasing level of noise*



In Figure 2 the standardized conditional residuals (in red) and the standardized deviance residuals (in black) have been displayed for the simulated sample for both levels of noise. On the right hand side part of the picture the values of $log(t)$ and of the linear predictor for each element in the dataset have been displayed. Bars have been colored according to whether they were considered outliers (i.e. $abs(residual) \geq 2$). With the smaller level of noise both deviance residuals and conditional residuals select the same points in the dataset as outliers (blue bars in the top right quadrant of Figure 2). With increasing noise, one point has been selected by the proposed method as being an outlier (the red bar in the bottom right quadrant of Figure 2) and one point has been selected by both deviance and conditional residuals as being an outlier (blue bar).

It is worth noticing that the element that has been selected as having too high a residual by the proposed diagnostic tool is actually a unit that, considering the value of the linear predictor, should have died much later then when it actually did, conditionally to the other *n-1* elements in the dataset.

Figure 2 – *Standardized deviance residuals and standardized conditional residuals for the simulated data with increasing noise..*



5. Kidney Infection Data

In this section the proposed residuals will be computed on a vastly used dataset in survival data analysis, namely the catheter infection data (McGilchrist and Aisbett, 1991). For our purposes we will be using only a subset of those patients, as in Collett (2003).

In the treatment of some kidney related diseases, dialysis could help remove waste materials in the blood. The use of such a technique could result in the development of an infection at the site where the catheter is inserted. To cure the infection, the catheter must then be removed. McGilchrist and Aisbett (1991) recorded the time from insertion until infection (in days) for a group of kidney patients. It is possible that the catheter may be removed for reason other than infection; such a case results in right censored data. Following Collet's approach

only 13 out of 39 patients, namely those suffering from glomerulo neptiritis, acute neptiritis and polycyatic kidney disease have been removed from the analysis and only those with diseases coded as type 3 ("other") in the original paper from McGichrist and Aisbett, have been considered.

In Table 1 for each of these 13 patient, the time, the status (1=infection, 0=other reasons), the age and gender have been displayed. Only one right censored observation is present in the dataset, all the others experienced the event due to infection during follow-up.

Table 1 – *Kidney infection data (Collet, 2003)*

Patient	Time	Status	Age	Gender	Disease
1	8	1	28	M	Other
2	15	1	44	F	Other
3	22	1	32	M	Other
4	24	1	16	F	Other
5	30	1	10	M	Other
6	54	0	42	F	Other
7	119	1	22	F	Other
8	141	1	34	F	Other
9	185	1	60	F	Other
10	292	1	43	F	Other
11	402	1	30	F	Other
12	447	1	31	F	Other
13	536	1	17	F	Other

Only 3 male subjects are present in the dataset and they all experienced the event within a month from the catheter insertion.

A Cox model, considering Age and Gender as covariates, has been fitted to the data, resulting in the following estimates (Table 2):

Table 2 – *Estimates of the parameter of cox proportional hazard model for kideny infection data*

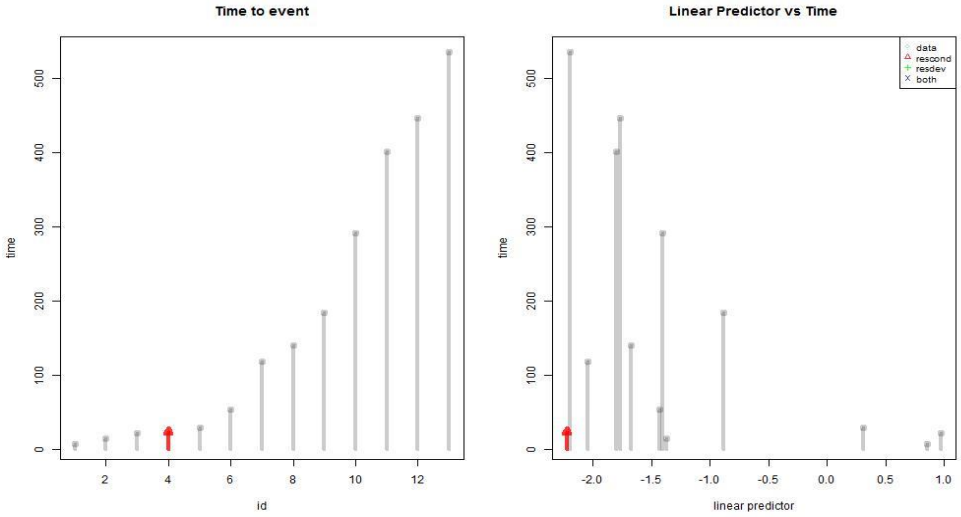
covariate	coef	exp(coef)	Std. err	Z	p-value
Age	0.0304	1.0308	0.0262	1.16	0.250
Gender(F)	-2.7108	0.0665	1.0959	-2.47	0.013

As it is also quite clear from the data in Table 1, patients in the dataset tend to show an increasing risk with age and a significant effect of gender, females being more resilient to developing the infection.

In Figure 3 the time to event for each patient and the time vs linear predictor have been displayed. Patients that showed a standardized deviance residual (resdev) or a standardized conditional residual (rescond) greater than 2 have been

highlighted in the plots. According to the deviance residuals there is no patient showing a behavior that is not coherent with the fitted model (resdev≥2). Conditional residuals, on the other hand, singled out the 4th patient in the dataset (red bar in Figure 3), i.e. a 16 years old female that developed an infection after 24 days and that shows a standardized conditional residual greater than 2.

Figure 3 – *Time to event for each unit in the kidney data and linear predictor vs time.*



In Table 3 the original ranking and the optimal ranking assigned according to the best conditional contribution to likelihood have been displayed.

Patient number 4, being a Female aged 16, should develop an infection much later in time, since young age is a protective factor and females have a lower risk of developing the infection according to the fitted model. This is also confirmed by the plot of the linear predictor vs time in Figure 3. The time to event tends to decrease with the increase of the value of the linear predictor. Patient 4 shows a very small value of the linear predictor, associated with a too small value of the time to event.

The optimal ranking method would, conditionally to the other 12 patients, expect her to develop the infection after all the other patients. She is clearly an anomalous patient since she develops the infection sooner than a male with almost the same age (patient 5).

Table 3 – *Observed and Optimal ranking for kidney infection data obatined via the proposed method.*

ID	Age	Gender	Time	offset	Original Ranking	Assigned Ranking
1	28	M	8	0	1	1
2	44	F	15	7	2	9
3	32	M	22	2	3	1
4	16	F	24	9	4	13
5	10	M	30	3	5	2
6	42	F	54	5	6	1
7	22	F	119	5	7	12
8	34	F	141	2	8	10
9	60	F	185	5	9	4
10	43	F	292	0	10	10
11	30	F	402	1	11	12
12	31	F	447	1	12	11
13	17	F	536	0	13	13

6. Conclusions

In this paper, a new approach to determining the residuals for the Cox model has been suggested. The proposal is in line with the idea of deviance residuals in generalized linear models, i.e. the difference between the log likelihood for the full model and the log likelihood for the fitted model. Since there is no unique definition of full model in Cox regression, the most common alternatives available in the literature are not based on likelihood. Our proposal is based on determining the optimal ranking of each element in the dataset conditionally on the ranking of the other *n*-1 elements. This is achieved by selecting, for each unit, the rank that yields the best conditional contribution for that subject, to the log likelihood of the model, given the ordering of the other *n*-1 subjects. The proposal is still under testing and has been tried on a small simulation study and on a real benchmark dataset, both times yielding very promising results. As future agenda we plan on testing the proposal on a large simulations study.

Bibliography

BARLOW, W.E., PRENTICE R. L. 1988. Residuals for relative risk regression, *Biometrika*, 75, 65 – 74.

COLLET D. 2003. *Modelling Survival Data in Medical Research*, Second Edition, CRC Press

COX D. R. 1972. Regression models and life-tables, *Journal of the Royal Statistical Society*, Series B (Methodological), No.34, pp.187 – 220.
COX D. R., SNELL E. J. 1968. A General Definition of Residuals, *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 30, No. 2, pp. 248-275.
HOSMER D.W., LEMESHOW S. 1999. *Applied survival analysis. Regression modelling of time to event data*. New York, NY USA: John Wiley & Sons 1999
KALBFLEISCH J.D., PRENTICE R.L. 2002, *The Statistical Analysis of Failure Time Data*, 2nd Edition, Wiley
McGILCHRIST C. A., AISBETT C. W. 1991. Regression with frailty in survival analysis, *Biometrics*, No. 47, pp. 461–466.
SCHOENFELD D. 1982. Residuals for the proportional hazards regression model, *Biometrika*, Vol. 69, No.1, pp.239-241.
THERNEAU T.M., GRAMBSCH P.M., FLEMING T. R. 1990. Martingale-based residuals for survival models, *Biometrika*. No.77, pp.147 – 160.

SUMMARY

Diagnostic tools based on optimal ranking in the Cox Model

Parameter estimates for Cox proportional hazard model are achieved via the maximization of the partial likelihood. Nonetheless, diagnostic tools and local fitting measures (residuals) are based on the complete likelihood.

Partial likelihood in Cox model entails a factorization of the contributes of each unit based on the ranking of each unit according to the time they have experienced the event. This contribute is conditional on the units that have experienced the event before the unit that is being considered. Such a structure suggests the possibility to use diagnostic tools based on the conditional contributes of each unit to the partial likelihood. In this paper we propose a diagnostic approach based on the optimal ranking of each unit conditional to the others.

Luciano NIEDDU, UNINT – Università degli Studi Internazionali di Roma, l.nieddu@unint.eu
Cecilia VITIELLO, Sapienza, Università di Roma, Dipartimento di Statistica, cecilia.vitiello@uniroma1.it

ISTAT INTERNATIONAL AND NATIONAL INITIATIVES ON BIG DATA

Paolo Righi

1. Introduction

The National Statistical Institutes (NSIs) have recently started investigating Big Data (BD) as potential data sources for generating official statistics. Especially beginning in 2013, numerous international projects and initiatives have been undertaken with the main purposes of defining strategies for adopting the BD sources in the production process. These strategies have to deal with some issues as legislation (privacy, confidentiality, etc.), communication (build public trust in the use of private sector BD for official statistics), partnership (especially with data providers) and financial aspects, IT infrastructure (to process huge amount of data), new or unusual (for the NSIs) quality and methodological framework, training NSI staff on new statistical tools and the definition of the governance of the BD inside the NSI. Several working group and task force have been set up for promoting the practical use of BD sources. The objective is to find the solutions for these challenges and support the capacity building, training and sharing of experience.

The Italian National Statistical Institute (ISTAT) attends to some international activities.

In particular at European level ISTAT is involved in the Task Force on Big Data and Official Statistics chaired by Eurostat. The Task Force implemented the road map and the action plan of the Scheveningen Memorandum on Big Data and Official Statistics (September 2013). This Memorandum aims to define a global strategy for introducing BD sources in the statistical production process and it has been the first step to provide a European strategic vision of the BD phenomenon.

Another initiative, where ISTAT participated, is the BD project of the United Nations Economic Commission for Europe (UNECE) launched in March 2014. The UNECE project ended in 2014 but the worldwide initiative of Global Working Group (GWG) on Big Data for Official Statistics coordinated by United Nations Statistics Division (UNSD) aims to continue some of the works already done by the UNECE project. The working group has been launched in 2015 and should work for the next two or three years.

In the GWG, ISTAT is leader of Task Team on Cross-cutting issues, Classifications, Frameworks and Taxonomy.

At National level, ISTAT closed (March 2015) a two year Technical Commission devoted to define a Roadmap for the adoption of BD in production process. Several actors coming from Academia, IT and Media companies have been involved and three pilots using BD sources have been carried out.

In the 2014, ISTAT has begun an internal project on the use of scanner data for the Harmonized Index of Consumer Prices compilation (involved in the European project "Multipurpose price statistics").

Other separated agreements with Mobile phone companies, University and IT companies have been established or are in progress.

The paper recaps the targets and the achieved outputs produced by these activities. Section 2 introduces the phenomenon of BD, with the definition of useful concepts for the official statistics. Section 3 focuses on the reasons of considering BD in the NSI. Their uses raise some issues shown in Section 4. Section 5 describes the national and international projects, task force and working group, where these issues are tackled.

2. Big Data: definition and facts

NSIs have been using several types of data sources in the production process of official statistics, including designed data sources such as censuses and survey sampling, and found data sources such as administrative and transactional data.

Recently as a result of more and more interaction with digital technologies by citizens, and the increasing capability of these technologies to provide digital trails, new sources of data have emerged and are increasingly available (IDC, 2014; Khan *et al.*, 2014; Chen *et al.*, 2014).

Some findings highlight what we are talking about:

- every two days we create as much information as we did from the beginning of time until 2003;
- in 2012 over 90% of all the data in the world was created in the past two years;
- it is expected that by 2020 the amount of digital information we have in existence will grow from 3.2 zettabytes today to 40 zettabytes and it will be 44 times larger of the amount produced in 2009, with yearly rate of 50%-60%;
- every minute we send 204 million emails, generate 18 million Facebook likes, send 278 thousand Tweets and up-load 200,000 photos to Facebook.

The volume of these new sources has naturally inspired to call these data as Big Data (BD). Nevertheless, an important question is how much is this new information useful for creating statistics. The definition of a BD source is more intricate taking into account that the phenomenon is complex and relevant dimensions changes in accordance with the field of interest. In the Official Statistics, the following definition is been proposed at international level:

Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made (UNECE, 2014).

Along with this definition the following classification of BD is proposed:

- Human-sourced information (Social Networks): Facebook, Twitter, Tumblr etc.; blogs and comments; personal documents; pictures: Instagram, Flickr, Picasa etc.; Videos: YouTube etc.; internet searches; mobile data content: text messages, user-generated maps, E-Mail. This information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Data are loosely structured and often ungoverned.
- Process-mediated data (Traditional Business Systems and Websites): data produced by Public Agencies; medical records; data produced by businesses; energy consumption; commercial transactions, banking, stock records, E-commerce, Credit cards. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Some sources belonging to this class may fall into the category of "Administrative data".
- Machine-generated data (Automated Systems or Internet of Things): data from sensors such as: home automation, weather and pollution sensors, traffic sensors and webcams, scientific sensors, security sensor; mobile sensors (tracking) such as mobile phone location, cars and travel sensors satellite images. Data from computer systems logs and web logs. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

3. Why do the NSIs deal with the use of BD sources?

Afterward the adoption of administrative data in statistical process, the NSIs should treat with a new class of secondary data collected by agencies or private companies with not statistical purpose. Organizational and management challenges relate to the legal framework, privacy, knowledge and capacities of handling a dynamically evolving data and metadata ecosystem have to deal with. Despite these problems the NSIs are approaching to the BD sources for different reasons (UNECE 2014):

Reputational drivers: BD sources have the potential to significantly impact the statistics industry. It is important that NSIs continue to demonstrate their relevance and remain competitive with other emerging sources of data if governments are to continue to see value in Official Statistics. These drivers seek to exploit new opportunities to keep pace with possibilities. This leads to a data-oriented approach where statistical organizations ask how they can make use of new sources.

Efficiency drivers: budget cuts in the NSIs and at the same time producing improved outputs lead to consider new data sources, technologies and methodologies. BD are sought to:

- identify and provide information about survey population units (sample frame and register creation);
- replace survey collection, reduce sample size, or simplify survey instruments (full or partial data substitution);
- ensure the validity, consistency and accuracy of survey data (data confrontation, imputation and editing).

The need for new statistics or statistics with an improved timeliness or relevance: use of new data sources that fill a particular information need to extend the existing measurement of economic, social, environmental phenomena to a high quality for use in policy making. There may be a range of demands that can be assisted through the use of BD:

- Improve timeliness of outputs;
- Enhance relevance or granularity of outputs;
- Increase accuracy or consistency of outputs.

Alternatively it may enable statistical organizations to produce statistics where high quality is less appropriate but can meet public demand on issues of the day.

4. Issues of using BD

Several issues can be identified when using BD: legislation, protection of privacy, confidentiality, intellectual property; communication to manage public

trust and acceptance of data re-use and linking to other sources; communication to tackle a negative public opinion; partnership with data providers, scientific community and IT providers; internal IT infrastructure; governance of access and use of the data; new skills for the NSI staff.

Considering the statistical view-point the paradigm shift from designed data for planned statistics to data-oriented or data-driven statistics offers new challenges. Beyond the descriptive statistics it will be necessary to determine under which conditions valid inferences can be made. Undercoverage and selectivity typically affect BD sources i.e., human generated data may suffer from several biases such as self-selection, self-reporting; absence of metadata obstacles the inference process since many of the interest characteristics like gender or age, are not known; automated systems suffers from a placement bias.

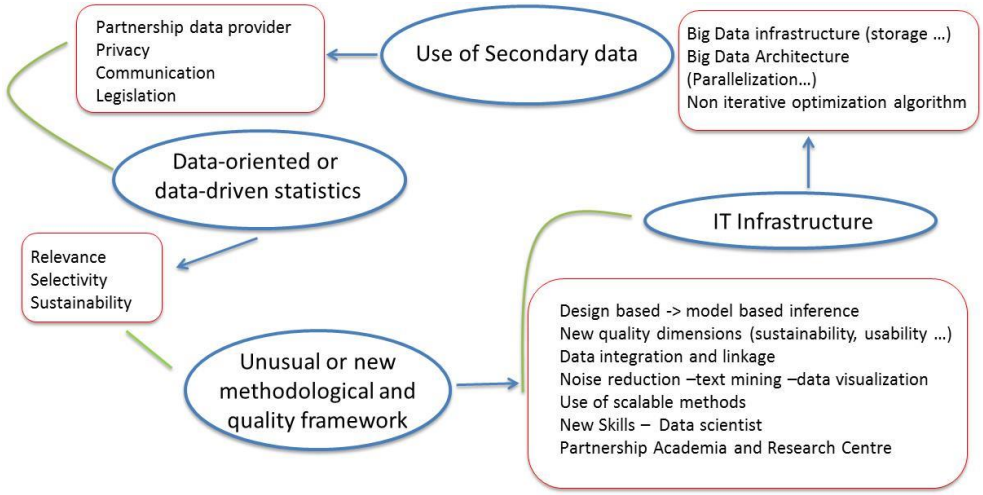
In general, the NSI staff is few experienced on non-probabilistic samples and statistical tools to extraction, transformation and load methods to take unstructured data to a processable form (statistical learning, data mining, data visualization).

The interconnection with the IT aspects yields a lot of efforts to create platforms and services specifically built to handle vast amount of data. Parallelization algorithms (like MapReduce, Hadoop, RHadoop, consistent hashing) to make possible the computation on distributed environments and non-iterative optimization algorithms have to be applied.

Summing up all these aspects the methodological and quality framework to manage the BD can be new or unusual for a NSI.

Figure 1 highlights the innovations or peculiarities and the related implications of the statistical production process when using BD sources. They are: the use of secondary data; data driven statistics; methodological and quality framework; IT infrastructure.

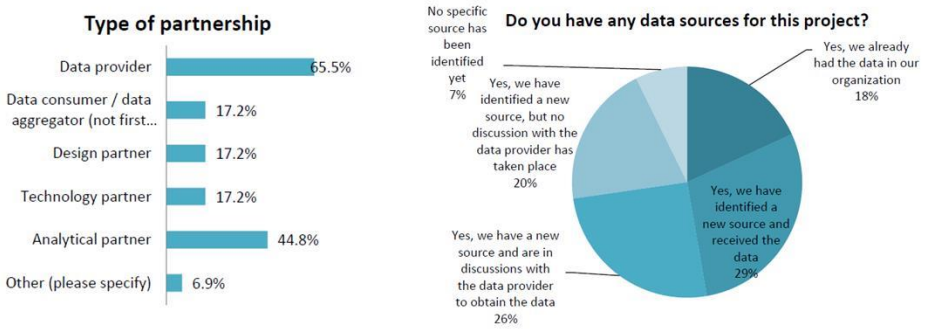
Figure 1 – *Challenges, innovations and related issues with use of Big Data sources in the Official Statistics*



The 2015 survey (first edition) on Big Data projects for Official Statistics conducted by the United Nations Statistics Division (UNSD) and the United Nations Economic Commission for Europe (UNECE) gives a picture of the development of the BD strategies in the Official Statistics. The questionnaire investigated the overall strategies and the specific projects carried out by the NSIs and International Organization. The respondents were 32 NSIs and 3 Organizations per 57 BD projects. The first evidence is that very few respondents have defined a long-term strategy for adopting BD sources and […] *More than two thirds of the organizations explained that they do not yet have defined a quality assessment framework for Big Data sources or the output of analysis* […] (United Nation Statistical Commission, 2015). Likely, to deal with these issues the experiences with real data should be useful. But there is a concrete difficulty to have BD available: partnership with BD providers and the legislation are felt as the main problems to work with the new sources. Follow, skills, IT infrastructure and methodology.

In particular for the 58% of the projects a partnership has been defined or is still in discussion but only the 65.5% of these partnerships are with a data providers (figure 2 left) and the 29% of the projects received data after a partnership agreement (figure 2 right). The 68% of these 57 projects have privacy and confidentiality issues to be tackled.

Figure 2 – *Partnership in the 57 Big Data projects investigated in the Big Data Project Survey*



Source UNSD/UNECE.

5. International and National initiatives on BD

ISTAT is involved in different international activities on the use of BD.

At European level, in September 2013, has been set up the Scheveningen Memorandum on Big Data and Official Statistics which encourages members of the European Statistical System to develop a BD strategy, share experiences and collaborate at the level of the European Statistical System and beyond. Furthermore the Memorandum outlines a global strategy for introducing BD sources in the statistical production process. ISTAT attended on drafting the Memorandum and to the subsequent Task Force for defining the road map and the action plan implementing the Memorandum in concrete. In September 2014, the European Statistical System Committee endorsed the Roadmap and Action plan 1.0 and integrated them into the ESS Vision 2020 portfolio (Eurostat, 2014)

In 2016 and since 2020 two European Statistical System network (ESSnet) projects should be launched. Projects will investigate practical aspects by implementing pilots and applications on BD sources at EU level with many NSIs including ISTAT. The overall aim of these ESSnets is to test parts of the ESS Big Data Action Plan and Roadmap 1.0. In particular, the following results are expected: identify pilots for generating statistics from at ESS level; identification and analysis of output portfolio of BD sources; identification and definition of skills and competences; exchange of information with stakeholders within the statistical system and the research community; development and review of methodological and quality frameworks for BD sources in official statistics; identification, definition and implementation of IT infrastructures for BD

processing; access to BD sources, identification and preparation of non-legal and legal conditions for access and use of BD within the ESS. Experiences obtained in each pilots will contribute to horizontal topics as: methodology, quality and metadata, IT infrastructure, skills, partnerships and communication. Another initiative, where ISTAT participated, is the BD project of the United Nations Economic Commission for Europe (UNECE) launched in March 2014 (UNECE 2014). Four Task Teams have ended their works in the 2014. They focused on: partnership with BD owners; privacy and legal issues; quality of statistics (UNECE, 2014); a "Sandbox" that provided a technical platform for the NSIs to jointly experiment with BD sets and tools. Sandbox activity will continue in the 2015. Moreover, ISTAT is involved in worldwide initiative of Global Working Group (GWG) on Big Data for Official Statistics coordinated by United Nations Statistics Division (UNSD). The purpose is of continuing some of the works already done by the UNECE project. The activities of GWG on these topics began in the mid of 2014 with a Global Survey for investigating the challenges relating to methodology, privacy and access to data, partnerships and skills that the NSIs have to dealt with (United Nation Statistical Commission, 2015). The organization of the working group provides many Task Teams. ISTAT is the leader of Cross-cutting issues, Classifications, Frameworks and Taxonomy Task Team. Next deliverable of the Task Teams is the new questionnaire of the Global Survey. The results will be shown at the Global Conference on Big Data for Official Statistics in Abu Dhabi, UAE the 20-22 October 2015. At National level, ISTAT closed (March 2015) a two year Technical Commission devoted to define a Roadmap for the adoption of Big Data in production process. Several actors coming from Academia, IT and Media companies have been involved and three pilots using BD sources have been carried out:

Web Scraping for the ICT usage and e-Commerce in enterprises (Barcaroli *et al*., 2015): use of crawlers and scrapers for collecting data for the ISTAT Survey on ICT Usage and e-Commerce in Enterprises and text mining techniques in order to estimate the services offered by an enterprise through its web sites. In particular the pilot is focused on the online ordering or reservation or booking. The main target of the pilot is to verify a new technique for collecting information from the enterprises. Partnership with Cineca (consortium of Italian universities, National Research Council and Ministry of Education and Research) has been settled on;

Persons and Places, use of mobile phone data to estimate mobility flows (Furletti *et al*., 2014): **t**he ongoing ISTAT project "Persons and Places" estimates the residences and flows of people by means of administrative registers (Residence register - Work register – Study register). The first release has been an Origin Destination matrix at municipality level. The objective of the project is to produce statistics that are comparable with those obtained in the ongoing project deploying

the massive and constantly updated information carried by mobile phone call data records (CDRs) for estimating population statistics related to residence and mobility. The project has been carried out jointly by the Italian National Research Council (CNR), University of Pisa and a Mobile phone company (partnership agreement with University of Pisa). In 2014 Italian Data Protection Authority authorized ISTAT to utilized anonymized data only for the project Person and Place but there are some problems (ISTAT has not yet used the CDR data);

Google Trends, u**se of query shares (Google Trend Index) for nowcasting labour force statistics** (Bacchini *et al*., 2014). The ISTAT Labour Force survey produces monthly estimates of unemployment rate at National level with a time lag of 1 month. Purpose of the experiment has been to use the Google query shares "job offer" as auxiliary variables for nowcasting and improving the current provisional estimates.

Finally, ISTAT is implementing the use of scanner data for estimating the Harmonized Index of Consumer Prices. A partnership agreement with the Association of Modern Distribution (900 Associates, 32,000 outlets) has been established and two years (2013-2014) scanner data for six market chains have been obtained for more than 15 provinces of Italy. Data are referred to grocery products (almost 20% of the basket of products) and for each outlet by reference (identifies by the EAN code) weekly data (turnover and quantity) are available.

References

BACCHINI F., D'ALÒ M., FALORSI S. FASULO A., PAPPALARDO A. 2014. Does Google index improve the forecast of Italian labour market?, *Proceedings 47th Scientific Meeting of the Italian Statistical Society, Cagliari June 11-13*.

BARCAROLI G., NURRA A., SALAMONE S., SCANNAPIECO M., SCARNÒ M., SUMMA D. 2015. Internet as Data Source in the Istat Survey on ICT in Enterprises, *Austrian Journal of statistics*, Vol. 44, pp. 31–43.

CHEN M., MAO S., LIU Y. 2014. Big data: a survey. *Mobile Networks and Applications*, Vol. 19, no. 2, pp. 171–209.

EUROSTAT 2014. The European Statistical System (ESS) Vision 2020. Technical Report, http://ec.europa.eu/eurostat/web/ess/-/the-essc-comes-to-an-agreement-on-the-ess-vision-2020

FURLETTI B., GABRIELLI L., GAROFALO G., GIANNOTTI F., MILLI L., NANNI M., PEDRESCHI D., VIVIO R. 2014. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach, *Proceedings 47th Scientific Meeting of the Italian Statistical Society, Cagliari June 11-13*.

IDC 2014. Analyze the future, http://www.idc.com.

KHAN N., YAQOOB I., HASHEM I. A. T., INAYAT Z., MAHMOUD ALI K., ALAM M.,SHIRAZ M., GANI A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges, *The Scientific World Journal*, Vol. 2014, pp 1-18.

UNECE (2014). How big is Big Data? Exploring the role of Big Data in Official Statistics. Technical Report, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307.

UNECE (2014) A Suggested Framework for the Quality of Big Data. *Deliverables of the UNECE Big Data Quality Task Team*. December, 2014.

UNITED NATION STATISTICAL DIVISION (2015). Results of the UNSD/UNECE Survey on organizational context and individual projects of Big Data Prepared by the Statistics Divisions of UN/DESA and UN Economic Commission for Europe. Background Document, http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf.

SUMMARY

Istat international and national initiatives on big data

The National Statistical Institutes have only recently started investigating Big Data as potential data sources for generating official statistics. Especially beginning in 2013, numerous international projects and initiatives have been undertaken. The scope of these activities is to deal with the issues related to the Big Data sources that are typically secondary data collected by agencies or private companies with not statistical purpose. Legislation, protection of privacy, confidentiality, communication, partnership with data providers, scientific community and IT providers, IT infrastructure, governance and new skills for the staff of the National Statistical Institutes are relevant matters. Furthermore, the paradigm shift from designed data for planned statistics to data-oriented or data-driven statistics offers new challenges in the methodological and quality framework. The paper shows the international and national project involving the Italian National Statistical Institute describing the main steps useful for defining a global strategy for introducing Big Data sources in the statistical production process.

Paolo RIGHI, ISTAT, parighi@istat.it

DEEP LEARNING FOR SUPERVISED CLASSIFICATION

Agostino Di Ciaccio, Giovanni Maria Giorgi

1. Introduction

One of the most recent area in the machine learning research is Deep Learning. We can define Deep Learning classification models a class of Machine Learning algorithms that use a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each layer uses the output from the previous layer as input, while the final layer gives the final classification. The model may be supervised or unsupervised.

This is a very general definition, which reminds the structure of a neural networks algorithm. Effectively, a Deep Neural Network (DNN) is an artificial neural network with many hidden layers of units between the input and output layers and millions or billions of parameters. For example, in the "Google Brain" project it was developed a massive neural networks with 1 billion parameters to detect cats inside YouTube videos.

Google, Facebook, Microsoft, Apple, in recent months have been investing in the field of Deep Learning. Google acquired Deep Mind to leverage its expertise in Deep Learning methods. Deep Learning algorithms have been applied successfully to computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics.

The key idea of Deep Learning is to combine the best techniques from Machine Learning to build powerful general-purpose learning algorithms. However, we don't have to commit the mistake of identify Deep Learning models with Neural Networks. Other approaches are possible, with the joint use of several powerful statistical models.

In the next paragraph we introduce the approach of supervised classification with ensemble methods, which combine the predictions of several basic estimators built with one or more given learning algorithms in order to improve accuracy and/or robustness over a single estimator. In particular, Bagging, Boosting, Stacking and a generalization of Stacking are illustrated. In paragraph 3 we show an application to a real classification problem, where the last approach has proved to be very effective.

2. Ensemble methods for classification

Ensemble learning is a machine-learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches, which construct one classification algorithm from training data, ensemble methods construct a set of classification algorithms and combine them. An ensemble contains a number of algorithms, which are usually called base learners: Decision Trees, Neural Networks (NN), Support Vector Machine (SVM) or other kinds of machine-learning algorithms

We can distinguish between:

- homogeneous base learners,
- heterogeneous base learners.

Typically, an ensemble model is constructed in two steps.

1) A number of base learners are constructed, which can be generated in a parallel style or in a sequential style, where the generation of a base learner has influence on the generation of subsequent learners.
2) The base learners are combined, usually by (weighted) majority voting for supervised classification.

Generally, to get a good ensemble, the base learners should be as more accurate as possible, and as more different as possible. In practice, the diversity of the base learners can be introduced in several ways:

- subsampling the training data,
- manipulating the attributes,
- injecting randomness into learning algorithms.

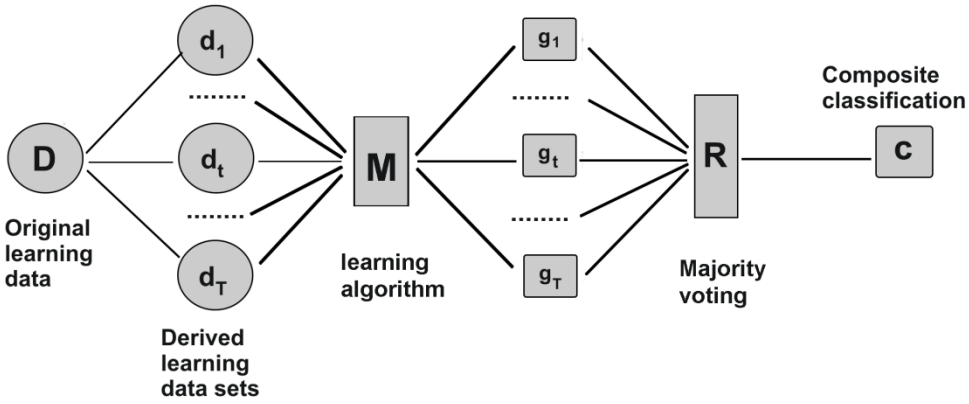
The employment of different base learners and/or different combination schemes leads to different ensemble methods. Statistical properties of the methods and reliability of the results are usually obtained by cross-validation. The optimal 'tuning' of the models is obtained by a grid-search approach: the analysis is repeated with parameters taken inside fixed intervals, evaluating the results by cross-validation.

2.1. Bagging, Boosting and Stacking

Bagging (Breiman 1996) uses homogeneous learners but different samples of observations and/or predictors (features) to generate different classifiers. To aggregate the classifiers, it is used averaging in regression, majority vote in

classification. The accuracy of the aggregate model is usually not better of the virtual accuracy of the base model, but Bagging reduces the variance and helps to avoid overfitting.

Figure 1 – *The Bagging (Bootstrap Aggregation) scheme.*

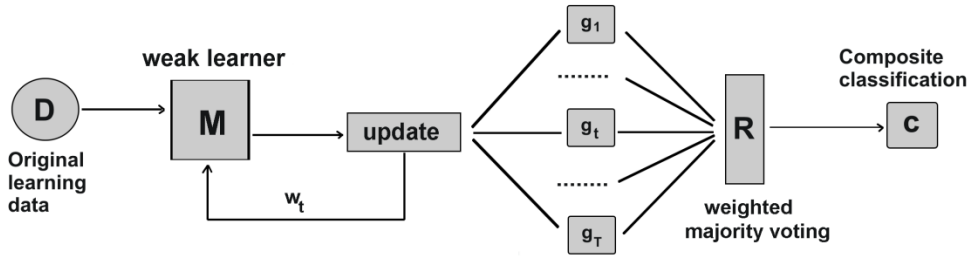


Using a single weak learner, Boosting (Schapire 1990), recursively, makes examples currently misclassified more important. This approach, shown in figure 2, is useful for reducing bias, converting weak learners to strong learners, sometimes can give overfitting.

Stacking (Wolpert 1992), unlike bagging and boosting, is not used to combine models of the same type - for example, a set of decision trees. Instead it uses different learning algorithms and two stages.

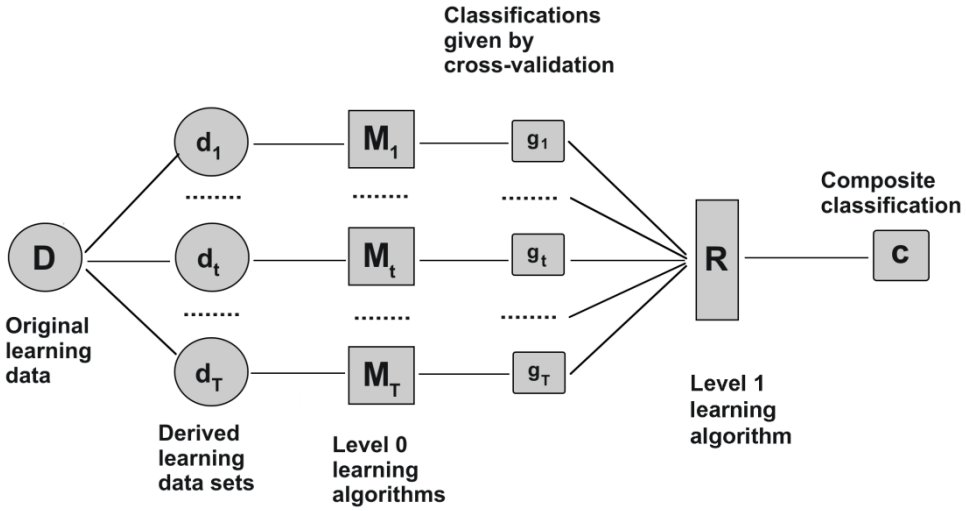
As shown in figure 3, in level 0 several algorithms are trained on the available data, giving as output the class probabilities for the target, using a cross-validation prediction (usually a 5-fold CV). Then a combiner algorithm is trained on the overall class probabilities to make a final prediction. The goal of the second stage is to combine in the most effective way the predictive capability of the different algorithms of the first stage. From this point of view, it is crucial that the predictive capability is assessed through cross-validation, to avoid to reward the overfitting models. Moreover, this allows to evaluate statistically the individual algorithms that we are using and it is fundamental to make the tuning of the model parameters.

Figure 2 – *The Boosting scheme.*



In real problems, Stacking is less widely used than bagging and boosting. This is because it is computationally demanding, it is difficult to analyze theoretically and because there is no generally accepted best way of doing it - the basic idea can be applied in many different variations.

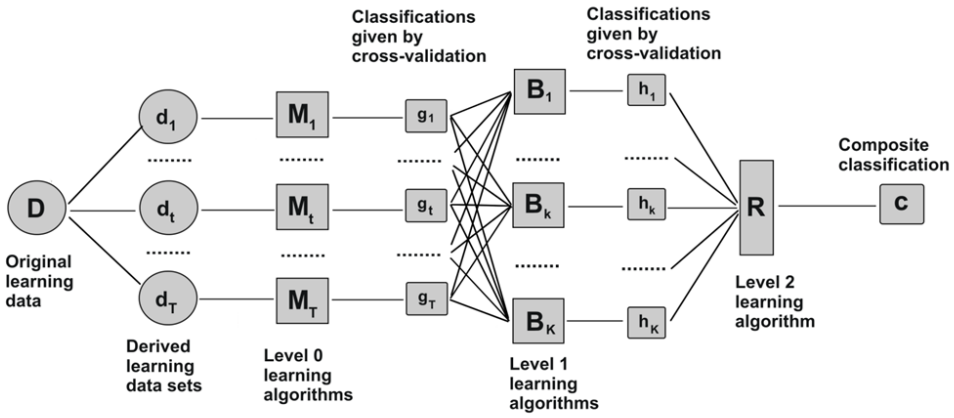
Figure 3 – *The Stacking scheme.*



2.2. Multi-stage Stacking

The Stacking approach can be easily generalized to a multi-stage scheme, as shown in figure 4. In this scheme, we added an internal layer with K nodes corresponding to K different models, which elaborates the output of level 0 models. The scheme looks like a feedforward neural network, but with some interesting differences: cross-validation does not allow for overfitting, each node can be a powerful statistical model, the joint use of several different models increases the reliability of the prediction. Many other ensemble methods can be viewed as special cases of stacking in which a data-independent model combination algorithm, such as a majority vote, is used.

Figure 4 – *Generalized three-stages Stacking*



As we mentioned above, we have the problem of choosing the parameters to best tune the algorithms and this requires an evaluation of the algorithm's performance for each set of parameters through cross-validation.

If we have many models with many parameters in several layers, the overall optimization problem is not practically feasible. What is done, is the individual tuning of all algorithms of a layer, starting from the lowest level, and then proceeding to the next layer. Note that each algorithm can be a complex machine-learning method, which may require several hours of processing. Another possibility is to train the models one after another, where each model tries to achieve best results when associated with all the preceding models.

The overall assessment may be carried out again with cross-validation or with the use of an independent data set. With big-data, the data split in training-validation-test sets, is usually enough.

3. An application of Deep Learning

Kaggle is a famous website of Data Science competitions. Any company can obtain a cost-effective way to solve machine-learning problems proposing a competition to the Kaggle's community of data scientists.

In March 2015, Otto Group, one of the world's biggest e-commerce companies, proposed one competition. The proposed challenge is the construction of a classification model which is able to accurately classify the products among 9 main product categories, using 93 observed numeric features (obfuscated and with no further information)[1]. It is a classical supervised classification problem.

The competitors had available one training set with 61678 units which included the category, and a test set with 144368 units without the category. They had to submit a file with the predicted probabilities on the test set, obtaining a score by Kaggle. It was possible to submit a maximum of 3 entries per day. The total prize pool for this competition was $10,000.

As a condition of receipt of the Prize, the winner must deliver the final model's software code with the associated documentation. The participants in the Otto Group competition were 3848 from all over the world.

The winner model was a generalized 3 stages Stacking model as showed in figure 4. The impressive list of the models used in the first stage is reported in table 1. In level-0 there are 33 models and 8 engineered feature-sets. The cross-validation probability class predictions of these models (plus the engineered feature-sets) are used as meta features for the 2nd stage. The derived features (class probabilities) of the models at the 2nd stage are 33*9=297, the other 8 feature-sets give 148 columns (6*9+1+93) for a total of 445 derived features.

In the 2nd stage (level-1) there are only 3 models: XGboost (Friedman 2000), Neural Networks (NN) and Adaboost (Schapire 1990). The final stage is composed by a weighted mean of the level-1 predictions.

In level-0, there are many different models: Neural Networks, Gradient Boosting (Friedman 2000), RandomForest (Breiman 2001), Logistic Regression, Extremely randomized trees (Geurts et al. 2006), K-Nearest Neighbors (Cover & Hart 1967), Multinomial Naïve Bayes (Zhang 2004), K-means (Hartigan 1975), t-

[1] https://www.kaggle.com/c/otto-group-product-classification-challenge.

distributed stochastic neighbor embedding (van der Maaten et al. 2008), Support Vector Machines (Cortes & Vapnik 1995).

Tabella 1 (a) – *Models used in the first stage of the Stacking model by the winner.*

M1	RandomForest (R). Dataset= X
M2	Logistic Regression (Scikit). Dataset= Log(X+1)
M3	Extra Trees Classifier (Scikit). Dataset= Log(X+1)
M4	KNeighborsClassifier (Scikit).
M5	libfm. Dataset= Sparse(X). Each feature value is a unique level.
M6	H2O, NN. Bag of 10 runs. Dataset= Sqrt(X + 3/8)
M7	Multinomial Naive Bayes (scikit). Dataset= Log(X+1)
M8	Lasagne, NN. Bag of 2 runs.
M9	Lasagne, NN. Bag of 6 runs. Dataset= Scale(Log(X+1))
M10	T-sne. Dimension reduction to 3 dimensions. Also stacked 2 kmeans features using the T-sne 3 dimensions.
M11	Sofia. Learner_type="logreg-pegasos" and loop_type="balanced-stochastic". Dataset= Scale(X)
M12	Sofia. Learner_type="logreg-pegasos" and loop_type="balanced-stochastic". Dataset= Scale(X, T-sne Dimension, some 3 level interactions between 13 most important features)
M13	Sofia. Learner_type="logreg-pegasos" and loop_type="combined-roc". Dataset= Log(1+X, T-sne Dimension, some 3 level interactions between 13 most important features)
M14	Xgboost. Dataset= (X, feature sum(zeros) by row). Replaced zeros with NA.
M15	Xgboost. Multiclass Soft-Prob. Dataset= (X, 7 Kmeans features with different number of clusters, rowSums(X==0), rowSums(Scale(X)>0.5), rowSums(Scale(X)< -0.5))
M16	Xgboost. Multiclass Soft-Prob. Dataset= (X, T-sne features, Some Kmeans clusters of X)
M17	Xgboost. Multiclass Soft-Prob. Dataset=(X, T-sne features, Some Kmeans clusters of log(1+X))
M18	Xgboost. Multiclass Soft-Prob. Dataset=(X, T-sne features, Some Kmeans clusters of Scale(X))
M19	Lasagne NN(GPU). 2-Layer. Bag of 120 NN runs with different number of epochs.
M20	Lasagne NN(GPU). 3-Layer. Bag of 120 NN runs with different number of epochs.
M21	XGboost. Trained on raw features. Extremely bagged (30 times averaged).
M22	KNN on features X + int(X == 0)
M23	KNN on features X + int(X == 0) + log(X + 1)
M24	KNN on raw with 2 neighbours
M25	KNN on raw with 4 neighbours
M26	KNN on raw with 8 neighbours
M27	KNN on raw with 16 neighbours
M28	KNN on raw with 32 neighbours
M29	KNN on raw with 64 neighbours
M30	KNN on raw with 128 neighbours
M31	KNN on raw with 256 neighbours
M32	KNN on raw with 512 neighbours
M33	KNN on raw with 1024 neighbours

Tabella 1 (b) – *Models used in the first stage of the Stacking model by the winner.*

F1	Distances to nearest neighbours of each classes
F2	Sum of distances of 2 nearest neighbours of each classes
F3	Sum of distances of 4 nearest neighbours of each classes
F4	Distances to nearest neighbours of each classes in TFIDF space
F5	Distances to nearest neighbours of each classed in T-SNE space (3 dimensions)
F6	Clustering features of original dataset
F7	Number of non-zeros elements in each row
F8	X (the original data were used in the 2nd level training only by NN)

All the software used in the competition is open source. Lasagne[2] is a lightweight library to build and train neural networks. XGboost[3] is the Extreme Gradient Boosting. t-SNE[4] is the t-Distributed Stochastic Neighbor Embedding. Sofia-ml[5] is a library used to obtain Logistic Regression with Pegasos SVM updates. libFM[6] is a Factorization Machine Library. H2O[7] is a Machine learning library for Python, R, Java. Scikit[8] is a machine learning library in Python. In Table 1, the list of models and engineered features is shown. We indicated with X the original dataset with 93 features, with Scale(X) the standardized data, with Sparse(X) the sparse matrix representation of the original data matrix.

Each of the models listed in Table 1 was estimated independently. It does not need a joint estimate that it would almost impossible. Simply, after a tuning step, each model has been applied to the training data, obtaining the estimated probabilities of each class with a k-fold cross-validation. The output of all the models were then put together to create the data set to be analyzed at the level 1. So the stacking model applied does not require extraordinary computational resources and can be built with a traditional PC, even if the long processing time would suggest the use of systems based on GPU. An immediate comment you can make watching the list of Table 1, is that most of the models applied is not specific to the data set analyzed. Essentially, the overall stacking model used in the competition could be applied with excellent performance even in other problems of supervised classification. The estimation of the parameters of the models will change, some models will vary their importance in the final result, but this will be mostly automatic, with light intervention from the researcher. This is exactly the spirit of Deep Learning: build powerful general-purpose learning algorithms. It is worth to

[2] http://lasagne.readthedocs.org/
[3] https://github.com/dmlc/xgboost/tree/master/R-package
[4] http://lvdmaaten.github.io/tsne/
[5] https://code.google.com/p/sofia-ml/
[6] http://www.libfm.org/
[7] http://www.H2O.ai
[8] http://scikit-learn.org/stable/

observe that also the runner-up to the competition used a Stacking three-stage scheme, but with a number of models much more reduced.

Another example is the 2009 Netflix competition: this company offered a prize of $1 million for the best model able to recommend new movies to its users, using the handful of movies the users had rated. Data was a sparse matrix with more than 100 million date-stamped movie rating on 17,770 movies by 480,189 users. The solution was an ensemble model with hundreds of predictors and many level-1 learning algorithm (Töscher et al. 2009).

4. Conclusion

Multi-stage Stacking models are very reliable and accurate methods often used in deep learning application. The classification performance of these methods can be very good, sometimes outperforming Deep NN, that, in any case, can be included as one of the models used in Stacking. The strength of this approach is primarily based on the Cross-Validation, that allows an assessment of the reliability of each used model, preventing overfitting. Cross-Validation is the best way to evaluate the prediction error of this kind of composite classifiers. The other distinctive element is the use of several powerful base models, which are estimated to generate a self-adaptable method capable to analyze different sets of data.

Much work is still needed to select the best architecture for multi-stage Stacking and to identify the best mix of models to use, but our impression is that the multi-stage Stacking has classification capabilities that are difficult to reach with other approaches.

References

BREIMAN, L. 2001. Random Forests. *Machine Learning* Vol. 45, No. 1, pp. 5–32. doi:10.1023/A: 1010933404324.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning* Vol. 24, No. 2, pp. 123–140. doi:10.1007/BF00058655

CORTES C., VAPNIK, V. 1995. Support-vector networks. *Machine Learning* Vol. 20, No 3, pp. 273. doi:10.1007/BF00994018.

COVER T.M., HART P.E. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, Vol.13, No. 1, pp. 21–27.

FRIEDMAN, J. H. 2000. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, Vol. 29, pp. 1189-1232.

GEURTS P., ERNST. D., AND L. WEHENKEL, 2006. Extremely randomized trees, *Machine Learning*, Vol. 63, No. 1, pp.3-42.

HARTIGAN, J.A. 1975. *Clustering algorithms*. John Wiley & Sons, Inc..

NIELSEN, M. A. 2015. Neural Networks and Deep Learning, *Determination Press*, (http://neuralnetworksanddeeplearning.com/).

RENDLE, S. 2010. Factorization machines. In *Proceedings of the 10th IEEE International Conference on Data Mining*. IEEE Computer Society.

SCHAPIRE, R. E. 1990. The Strength of Weak Learnability. *Machine Learning*, Vol. 5, No. 2, pp. 197–227. doi:10.1007/bf00116037

TÖSCHER, A., JAHRER, M., BELL, R.M. 2009. The BigChaos Solution to the Netflix Grand Prize, www.netflixprize.com/assets/GrandPrize2009_BPC_ BigChaos.pdf

VAN DER MAATEN L, HINTON G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605.

WOLPERT, D. 1992. Stacked Generalization. *Neural Networks*, Vol. 5, No. 2, pp. 241-259.

Zhang, H. 2004. The Optimality of Naive Bayes. FLAIRS2004 conference.

SUMMARY

Deep learning for supervised classification

One of the most recent area in the Machine Learning research is Deep Learning. Deep Learning algorithms have been applied successfully to computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics.

The key idea of Deep Learning is to combine the best techniques from Machine Learning to build powerful general‐purpose learning algorithms. It is a mistake to identify Deep Neural Networks with Deep Learning Algorithms. Other approaches are possible, and in this paper we illustrate a generalization of Stacking which has very competitive performances. In particular, we show an application of this approach to a real classification problem, where a three-stages Stacking has proved to be very effective.

Agostino DI CIACCIO, Department of Statistics, Univiversity of Rome “La Sapienza”, agostino.diciaccio@uniroma1.it

Giovanni M. GIORGI, Department of Statistics, Univiversity of Rome “La Sapienza”, giovanni.giorgi@uniroma1.it

PROPOSTE METODOLOGICHE PER L'INTEGRAZIONE DELLE STATISTICHE SOCIALI

Claudio Ceccarelli, Stefano Falorsi

1. Introduzione

A partire dal 2011 Eurostat ha avviato un progetto di riorganizzazione delle statistiche sociali nell'ambito dei paesi dell'Unione. Il modello proposto da Eurostat si basa su un approccio a moduli che, per costruzione, possono essere impilati e, laddove sovrapponibili, possono consentire l'utilizzo di informazioni rilevate in diversi indagini per la costruzione delle stime.

Tale progetto si è evoluto nel tempo fino ad arrivare alla sua versione definitiva presentata nel corso della riunione dei Direttori delle Statistiche Sociali, tenutasi a settembre 2014[1].

In tale occasione, Eurostat ha presentato una *roadmap* (Eurostat, 2013) per l'implementazione del progetto che prevede interventi di breve, medio e lungo periodo:

- *breve periodo*: studio di metodi di *pooling* per le stime da effettuare mediante la sovrapposizione di campioni sui quali sono state rilevate le medesime variabili, indipendentemente dai disegni sottostanti;
- *medio periodo*: ridisegno delle indagini campionarie con modifica delle numerosità campionarie in funzione del nuovo approccio modulare;
- *lungo periodo*: sistema integrato di micro-dati per le statistiche sociali, alimentato sia dalle indagini campionarie, sia dalle informazioni provenienti dai registri statistici.

In questo lavoro sono presentati alcuni possibili scenari per l'integrazione delle statistiche sociali che scaturiscono dall'applicazione di differenti strategie di rilevazione associate a diversi disegni di campionamento. Il tutto è finalizzato ad ottenere una completa integrazione delle indagini sociali e a garantire l'integrabilità con il sistema dei registri presenti in Istituto.

[1] I riferimenti alla documentazione ufficiale del progetto sono disponibili sul sito di Eurostat relativo ai lavori dell'Expert Group e citati singolarmente in bibliografia.

2. Gli scenari

Punto di partenza del lavoro è la realizzazione di una partizione del territorio italiano in aree sub regionali individuate a partire dalla metodologie utilizzata per la determinazione dei sistemi locali del lavoro (Istat, 2015). Tale partizione consente di realizzare una suddivisione del territorio in aree omogenee (dette Aree GSL - Grandi Sistemi Locali) rispetto ai tempi di percorribilità sulla quale basare sia i disegni di campionamento[2], sia le strategie di rilevazione, anche in funzione della possibilità di utilizzo sia della rete di rilevazione privata e sia di quella comunale.

Nel seguito del paragrafo, si considerano due scenari alternativi: a una occasione di indagine e a due occasioni. Nel primo le famiglie vengono intervistate una volta l'anno mentre nel secondo due volte.

Scenario ad una occasione di indagine

Questo scenario è basato su un disegno di indagine che prevede che le famiglie incluse in un dato sotto-campione, relativo a uno specifico *strumento di rilevazione*[3], vengano intervistate in un'unica occasione di indagine, durante l'anno, in cui vengono rilevate contemporaneamente tutte le variabili di interesse, ossia le variabili strutturali comuni a tutti gli strumenti di rilevazione e quelle specifiche di ogni strumento. Ogni sotto-campione è composto da famiglie differenti. In questo scenario non è possibile, quindi, sfruttare la conoscenza delle variabili strutturali - o di altre informazioni osservate nella prima occasione di indagine - per la scelta delle unità campionarie su cui osservare le variabili specifiche. Questo scenario può ulteriormente essere così suddiviso:

1. *pooled sample*: si tratta dei disegni di rilevazione delle attuali indagini sociali in relazione alle quali si possono sfruttare soltanto le informazioni raccolte in modo standardizzato dalle indagini stesse, sia sui contatti sia sulle variabili di indagine. Il livello di integrazione è minimo in quanto l'integrazione avviene a posteriori rispetto alla progettazione del disegno complessivo. Le differenze dovute alla realizzazione dei differenti disegni di rilevazione non sono tenute sotto controllo.

[2] Una progettazione integrata dei disegni di campionamento basata su una partizione del territorio uguale per tutte le indagini consente di annullare gli effetti dovuti a differenti strategie di allocazione dei campioni rispetto al territorio.

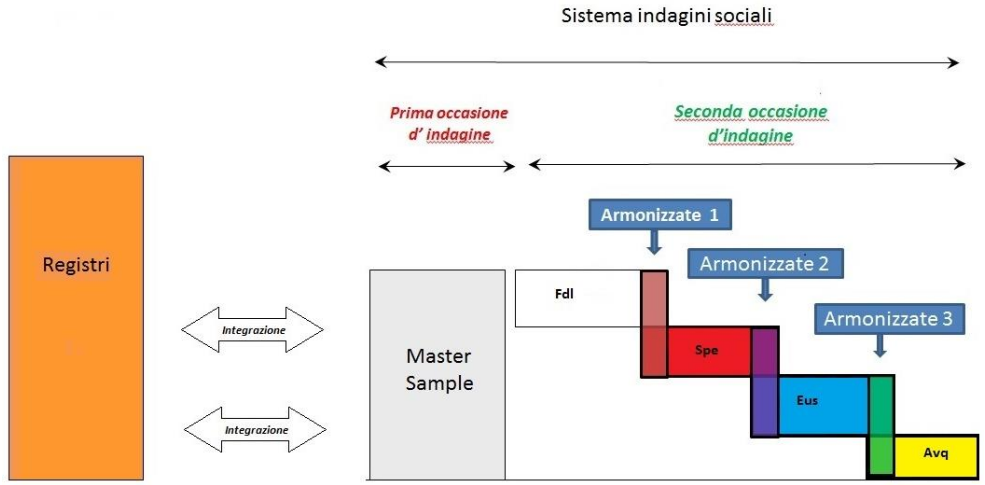
[3] Nell'accezione che compare nel progetto Eurostat, con il termine strumento si intende un processo di acquisizione delle informazioni che, genericamente, può essere una rilevazione campionaria o un processo di acquisizione del dato da fonte amministrativa. Un altro elemento che caratterizza lo scenario riguarda il raggruppamento in *moduli* predefiniti delle variabili armonizzate che devono essere rilevate.

2. *pooled sample armonizzato*: le informazioni rilevate per gli stessi moduli, nelle diverse indagini, possono essere utilizzate in un'ottica *pooled* senza soffrire di possibili effetti dovuti alla differenza di disegno di indagine (cambiano solo i livelli di precisione richiesta). L'armonizzazione in fase di progettazione del disegno è massima ed è funzione anche dell'ottimizzazione dei singoli disegni in base alla tempistica delle diverse fasi di rilevazione e della rete di rilevazione scelta. L'armonizzazione è estesa anche ai metodi e ai sistemi da adottare per il contenimento degli effetti dovuti alla mancata risposta totale (MRT)[4]: in sostanza si armonizzano tutti i processi sottostanti la produzione delle stime finali.

Scenario a due occasioni di indagine

Questo secondo scenario, illustrato in modo schematico in Figura 1, prevede una prima occasione d'indagine, svolta mediante un *master sample*, in cui si rilevano le variabili strutturali (non disponibili dagli archivi amministrativi) e tutte le variabili ausiliarie, tra cui quelle finalizzate a effettuare lo screening e il controllo della lista di campionamento, per reperire informazioni utili per ridurre i costi di rilevazione (tipo recapito telefonico fisso e/o mobile) e per facilitare le operazioni di contatto e di rilevazione successive.

Figura 1. *Il sistema delle indagini sociali*



Legenda: Fdl= Indagine sulle forze di lavoro; Spe= Indagine sulle spese delle famiglie; Eus = Indagine su redditi e condizioni di vita Eu-Silc; Avq = Indagine multiscopo sugli aspetti delle vita quotidiana.

[4] Uso o meno delle sostituzioni in luogo del sovra-campionamento, utilizzo degli stessi strumenti per il contatto delle famiglie e implementazione delle stesse metodologie per il trattamento a posteriori della MRT.

Nella seconda occasione si richiedono a conferma le variabili strutturali, già osservate, e si rilevano le differenti variabili specifiche e armonizzate su differenti sotto-campioni di famiglie estratti dal *master sample*, strutturate in moduli in modo da prevedere la possibilità di sfruttare operazioni di pooling in fase di stima.

Uno schema così articolato consente l'integrazione con il sistema dei registri dal quale si estraggono i campioni. Le informazioni raccolte nella prima occasione di indagine, in particolare quelle relative ai contatti, offrono la possibilità di aggiornare il registro da cui sono estratti i campioni, migliorandone la qualità in modo da fornire un set affidabile di totali noti sui quali vincolare le stime delle indagini.

3. Split questionnaire design

Nell'ottica dell'integrazione delle indagini, è possibile considerare un sotto-scenario che parte dalla selezione di un campione unico e congiunto per tutte le indagini del sistema mediante una tecnica nota in letteratura come *Split Questionnaire Design* (*SQD*)[5].

Nello SQD le diverse indagini del sistema costituiscono differenti *pattern informativi* che vengono allocati - applicando tecniche di allocazione ottima multivariata – su differenti sotto-campioni del campione generale. Nel caso del sistema Istat (cfr. Tabella 1), tutti i pattern informativi del campione generale devono prevedere l'osservazione delle variabili strutturali comuni e di eventuali altre variabili necessarie per la scelta del pattern (V.Com). Ciascun sotto-campione, poi, dovrà osservare uno tra i quattro moduli di variabili specifici delle indagini forze di lavoro (*Fdl*), spese delle famiglie (*Spe*), redditi e condizioni di vita Eu-Silc (*Eus*) e multiscopo sugli aspetti della vita quotidiana (*Avq*), così come viene esemplificato nello schema seguente.

Tabella 1. *Il sistema delle indagini sociali*

Pattern	V.Com	Fdl	Spe	Eus	Avq	Numerosità
1	x	x				n(1)
2	x		x			n(2)
3	x			x		n(3)
4	x				x	n(4)

Legenda: FdL= indagine sulle forze di lavoro; Eus = Indagine su redditi e condizioni di vita Eu-Silc; Spe= Indagine sulle spese delle famiglie; Avq = Indagine multiscopo sugli aspetti delle vita quotidiana.

Si può prevedere, inoltre, che i valori osservati sulle variabili comuni, per ciascuna famiglia e/o individuo del campione generale, determinino la scelta del

[5] Si veda, a tale proposito, il lavoro di Chipperfield e Steel del 2009.

modulo di variabili specifico da richiedere alla famiglia o all'individuo rispondente. Si tratta dell'approccio MAR (Missing At Random, all'interno di specifici profili informativi o celle) per lo SQD, in cui i profili dei rispondenti, rispetto a predefinite variabili, guidano nella definizione dei moduli informativi da somministrare. L'approccio MAR, a differenza di quello MCAR (in cui la definizione dei moduli informativi è completamente casuale (Missing Completely At Random) consente di ridurre il *response burden* sui rispondenti evitando di chiedere quesiti ritenuti poco significativi a specifiche categorie o profili di rispondenti. Rispetto allo schema citato in Figura 1, lo *SQD* può essere applicato anche in una sola occasione di indagine, come se le indagini Eu-Silc, spese delle famiglie e multiscopo sugli aspetti della vita quotidiana fossero tutti moduli aggiuntivi dell'Indagine sulle forze di lavoro, ognuna con la propria numerosità.

Per quanto riguarda gli aspetti operativi (legati principalmente alla struttura della rete di rilevazione e alle competenze richieste ai rilevatori), questo scenario può essere ragionevolmente applicato in un contesto di *rete di aree*[6].

Nel caso di una sola occasione di indagine, le variabili strutturali comuni e quelle specifiche di ciascuna indagine verrebbero tutte osservate nella medesima occasione[7].

4. Il ruolo delle reti di rilevazione e la partizione del territorio

Un aspetto essenziale nella realizzazione di un disegno di indagine integrato che possa offrire i vantaggi descritti è legato alla struttura dell'organizzazione territoriale della rete di rilevazione che svolge la prima occasione di indagine. In tal senso possono essere ipotizzati le seguenti tipologie di rete di rilevazione:

Rete di aree, concentrata nei centroidi delle aree GSL di rilevazione.

Questa tipologia di rete può essere assimilata a quella utilizzata dalla società privata che effettua attualmente l'indagine *FdL*[8]. I costi di viaggio infra-comunali hanno un grosso impatto sul costo totale di rilevazione ed è ragionevole, pertanto, pensare che il master sample debba concentrarsi su un campione compreso tra 1100 comuni (come è quello attualmente sul campo per *FdL*) e 1300 comuni. Si

[6] Per quanto riguarda gli aspetti organizzativi, la metodologia proposta può essere più agevolmente implementata nel caso in cui si adotti una *rete di aree* così come descritta nel paragrafo 4.

[7] Si tratterebbe, quindi, di una sorta di indagine *FdL allargata* in cui a ciascuna famiglia rispondente possono essere somministrati questionari alternativi definiti in base ai pattern del prospetto di cui sopra. In tale scenario diventa, quindi, cruciale, ampliare le competenze dei rilevatori ad altre aree tematiche, oltre le *FdL*, tra cui i redditi, le spese e altre variabili sociali.

[8] In prospettiva, si può ipotizzare che l'Istat si doti di una propria rete di rilevazione mediante nuove forme organizzative, creando, ad esempio, un "Albo dei Rilevatori" organizzato su base regionale.

potrebbe, infatti, pensare di aumentare limitatamente la numerosità di primo stadio del campione *FdL*, ad esempio del 10 o 15%, in virtù del fatto che la struttura territoriale del campione dovrebbe essere razionalizzata e resa più efficiente dall'introduzione delle aree GSL. Il master-sample verrebbe a formarsi in modo continuo durante l'anno contemporaneamente alla rilevazione del campione *FdL*.

Rete di comuni, che comprende tutti i comuni italiani o almeno quelli di dimensioni demografiche con popolazione maggiore di una data soglia.

In prima istanza, possiamo assimilare questa tipologia organizzativa con quella mediante la quale viene condotta correntemente l'indagine aspetti della vita quotidiana. E' importante ricordare che, attualmente, la rilevazione viene svolta mediante tecnica PAPI dagli impiegati comunali incaricati della rilevazione. Poiché si dispone di una rete capillare dove i rilevatori risiedono nel comune di rilevazione e, quindi, non bisogna sostenere costi di viaggio infra-comunali, è possibile considerare numerosità campionarie del master sample, in termini di comuni, maggiori rispetto a quelle sopra ipotizzate. A livello nazionale si tratta di un campione tra i 2000 e i 2200 comuni.

Poiché l'attuale rete comunale si avvale della tecnica PAPI occorre prefigurare un'evoluzione tecnologica della rete mediante l'utilizzo di *App* su cellulare e/o l'utilizzo di palmari.

Nel caso della *rete di aree*, si utilizza la struttura del campione comunale *FdL*. Tuttavia, per poter ottenere un campione che possa essere utilizzato da tutte e quattro le indagini, in ciascun comune selezionato si dovrà raddoppiare la numerosità campionaria del campione attuale *FdL*.

Nel caso, invece, di *rete di comunale*, si seleziona un prefissato numero di comuni campione in ciascuna area: metà o più della metà dei comuni selezionati svolge l'indagine *FdL*, i restanti comuni svolgono le indagini *Eus* e *Spe*, così come previsto negli attuali disegni delle due suddette indagini.

Per la prima occasione di indagine, finalizzata a costruire la lista delle famiglie campione per la seconda occasione e a rilevare un set, anche minimo, di informazioni strutturali., sembra ragionevole pensare a disegni clusterizzati per civico. La rilevazione per civico, infatti, potrebbe consentire di ridurre gli spostamenti del rilevatore sul territorio e quindi di contenere i costi complessivi della prima occasione di rilevazione.

5. La valutazione della qualità del registro della popolazione

Come accennato nel paragrafo 2.2, lo schema a due occasioni di indagine è finalizzato a ottenere una standardizzazione della fase di contatto di tutte le

indagini: ciò costituisce il presupposto per poter acquisire informazioni sulla *sovra-copertura* e sulla *sotto-copertura* del registro dai cui si estraggono i campioni.

Un elemento di indubbio vantaggio, difficilmente perseguibile con lo scenario a un'occasione, è che lo scenario rappresentato nella Figura 1 consente di pianificare annualmente indagini su particolari sotto-popolazioni sulle quali si può concentrare la sovra-copertura del registro. L'individuazione di tali sotto-popolazioni in fase di disegno, infatti, può essere guidata dall'integrazione con profili di presenza/assenza provenienti dai diversi archivi. Alcune delle sottopopolazioni definite dai vari profili (ad esempio, famiglie composte da tutti componenti stranieri) potrebbero, quindi, essere trattate come sotto-popolazioni pianificate nella selezione del *master sample* di un dato anno. Inoltre, al fine di stimare l'errore completo, si potrebbe pianificare un'indagine di copertura a cadenza pluri-annuale utilizzando, come struttura portante, il master sample.

La quantificazione della *sovra-copertura* della lista della popolazione diviene un elemento fondamentale per la qualità del registro della popolazione. L'utilizzo di "segnali" da fonti amministrative può permettere l'identificazione di individui che, pur essendo residenti in certi confini amministrativi, non vi sono abitualmente dimoranti. La mancanza di informazioni che permettano una stima della *sovra-copertura* comporta un effetto che si trasferisce sulle stime vincolate a totali di popolazione. In tal senso, i *mancati contatti* di unità statistiche – ossia le famiglie o gli individui campione che non è possibile contattare per differenti motivi – possono dare utili informazioni sulla *sovra-copertura* del registro popolazione.

Se la fase di contatto con le famiglie è opportunamente strutturata, l'aggancio dei record individuali appartenenti alle unità campione teoriche non contattate con le informazioni ausiliarie dei registri statistici dovrebbe consentire di assegnare, con un livello di precisione accettabile, a ciascuna unità non contattata la causa di mancato contatto. Pertanto, alla conclusione della fase di contatto, utilizzando congiuntamente fonti da indagine e fonti amministrative, risulterebbe possibile distinguere per ciascuna unità lo stato di: *a) corretta presenza nel registro* oppure *b) erroneo inserimento* nello stesso.

Vi sarà comunque una quota minima di *unità non risolte***,** ossia mancati contatti per i quali non è possibile stabilire con certezza la correttezza della loro presenza nel registro. Per tali unità, o parte di esse, si potrebbe pensare a una nuova fase di contatto durante la seconda occasione di indagine.

La disponibilità di queste informazioni costituisce un potente strumento per la stima dell'ampiezza della popolazione secondo differenti approcci metodologici. In particolare possiamo distinguere:

- ✓ un approccio alla stima dell'ampiezza della popolazione, basato su un modello di popolazione latente, in cui ciascun archivio amministrativo rappresenta il risultato di un particolare processo di cattura (soggetto a

specifici livelli di errore, non sempre quantificabili), l'indagine può essere vista come un ulteriore processo di cattura (su base campionaria) soggetto a livelli di errore la cui entità massima può essere tenuta sotto controllo in fase di pianificazione del disegno di campionamento e del piano dei contatti - per alcune sottopopolazioni (quelle di cui al punto *a* e *b*) dell'elenco precedente;

- ✓ in un approccio basato sul disegno di campionamento, ponderando ciascun mancato contatto (di cui al punto *b*) con il corrispondente peso campionario, è possibile ottenere una stima sovra-stima sotto il disegno di campionamento, del tasso di sovra-copertura del registro per differenti sotto-popolazioni territoriali e/o strutturali (ad esempio regione, classi di età decennali).

L'indagine di prima fase, opportunamente strutturata, può essere utilizzata anche per stimare la *sotto-copertura*.

Lo scenario a *due occasioni di indagine* consente di condurre, a cadenza quinquennale (o biennale), un'indagine ad hoc per la valutare l'errore di sotto-copertura basata su un sotto-campione dell'indagine di prima occasione dimensionato in modo tale da poter permettere una stima affidabile, per prefissati domini, della sotto-copertura. Ciò può essere fatto principalmente con due tecniche distinte.

Campione di numeri civici

Il sotto-campione dell'indagine di prima fase può essere costituito da un campione di numeri civici da censire. Il confronto di quanto risultante dal censimento di un numero civico con quanto riportato nel registro consente di stimare la sottocopertura, applicando il classico modello di cattura-ricattura di *Petersen.*

Campionamento indiretto

Relativamente al sottoinsieme di famiglie su cui è possibile effettuare il contatto è possibile ricostruire la situazione della *famiglia di fatto*. In tal modo la fase di primo contatto si configurerebbe come uno *schema di campionamento indiretto*, in cui:

- ✓ il campione diretto è costituito dalle unità (individui e/o famiglie) estratte dall'archivio di selezione,
- ✓ il campione indiretto è formato dalle famiglie di fatto, ricostruite nella fase di contatto, osservate nell'unità abitativa della famiglia anagrafica selezionata con il campione diretto.

Se il sotto-campione per la sotto-copertura è selezionato da due liste indipendenti (ad esempio, liste anagrafiche e archivio fiscale) e ciascuno dei due campioni adotta uno schema indiretto, la *sotto-copertura* può essere stimata

utilizzando la generalizzazione dello stimatore di *Petersen* proposta da Lavallée e Rivest (2012).

6. Le integrazioni con i registri

Per quanto riguarda l'integrazione con i registri tematici, la "revisione" qualitativa dei registro della popolazione offre la possibilità di integrazione tra le informazioni raccolte dalle indagini sociali e i registri tematici al fine di completare il sistema stesso dei registri. Le informazioni da indagine rappresentano un supporto informativo fondamentale essenzialmente per due motivi:

- ✓ *tempestività* – Le indagini raccolgono informazioni riferite al mese *t*; mentre, gli archivi amministrativi si riferiscono in genere al mese *t*-18;
- ✓ *esaustività rispetto alle modalità delle variabili di interesse* – Le indagini permettono di avere l'intero spettro di informazioni sui fenomeni di interesse mentre gli archivi amministrativi di base forniscono, in genere, informazioni solo su specifiche sottopopolazioni (ad esempio per il lavoro solo sull'occupazione regolare), intercettate nel processo amministrativo.

Sempre rimanendo nel caso delle informazioni relative al lavoro e all'istruzione, le indagini possono essere utilizzate per produrre versioni aggiornate, anche a livello infra-annuale dei registri in parola. In tal modo, le informazioni disponibili da archivio, che possono arrivare anche a *lag* temporali di 18 mesi, vengono aggiornate sfruttando in modo congiunto tutta l'informazione disponibile al tempo *t*, ossia le *covariate* da archivio e i dati di indagine.

Il sistema delle indagini sociali potrebbe essere pianificato per produrre stime attendibili rispetto a domini non pianificati, con riferimento a predefiniti livelli di aggregazione come ad esempio predefiniti incroci tra domini territoriali sub-regionali (sistemi o sotto-sistemi locali del lavoro) e variabili strutturali. Tali aggregati possono rappresentare un elemento di riferimento (o anche un vincolo) per le stime ottenibili per aggregazione da archivi. D'altronde, le indagini e i registri costituiscono due sistemi sinergici che si rafforzano l'un l'altro: la progettazione delle indagini sociali, infatti, trae piena forza dall'esistenza dei registri in quanto:

- ✓ il campione può essere maggiormente concentrato su sottoinsiemi di popolazione su cui l'informazione amministrativa è scarsa (o poco predittiva);
- ✓ mediante metodi di calibrazione, le stime dell'indagine riproducono ciò che è considerato come informazione *certa* proveniente dal registro (ad esempio i dati sull'occupazione regolare);

- ✓ la disponibilità delle informazioni ausiliarie disponibili dal sistema dei registri incrementa notevolmente l'accuratezza delle stime campionarie, di conseguenza, si possono diminuire le dimensioni dei campioni mantenendo inalterato il livello di precisione delle stime;
- ✓ l'indagine può essere utilizzata per validare i processi di ricostruzione statistica basate su informazioni amministrative discordanti provenienti da fonti amministrativi diverse.

Lo scenario a due occasioni di indagine, così come presentato nel paragrafo 2.2, offre la possibilità di ricostruire informazioni sfruttando metodologie basate sulle stime da modello. In particolare, le variabili possono essere ricostruite con stimatori *projection* di tipo *model based* o *model assisted* (Kim e Rao, 2012).

Secondo tale approccio si individua un modello che lega le variabili dipendenti - proprie di un certo modulo, rilevato da una data indagine – e le variabili ausiliarie rilevate nel master sample e presenti in archivio e aggiornate e "corrette" mediante lo stesso master sample. Studiato il modello con i dati raccolti dalla specifica indagine, si può proiettare la variabile predetta, utilizzando i parametri del modello appena stimato e le variabili ausiliarie, sia sul master sample sia sul registro. Tale approccio, che necessita di un elevato livello di qualità delle variabili ausiliarie e di elevati livelli di *fitting* dei modelli che si stimano, presenta notevoli vantaggi sia in termini di proprietà statistiche degli stimatori (correttezza rispetto al disegno), sia di notevole incremento del dettaglio delle informazioni che possono essere prodotte.

7. Alcune considerazioni

Gli scenari proposti rispondo a diverse esigenze. Eusrostat ha promosso un modello di integrazione molto ambizioso che dovrà consentire un maggiore utilizzo delle informazioni organizzando in modo ottimale la raccolta delle informazioni stesse. Le indagini sociali in Italia hanno raggiunto un elevato livello di qualità e completezza delle informazioni e degli indicatori prodotti ma, nel momento della progettazione di un sistema più strutturato, non si può non tener conto delle informazioni da fonte amministrativa: tutto ciò impone uno sforzo progettuale e realizzativo maggiore ma con importanti potenzialità. La progettazione di un sistema così ambizioso deve passare per una razionalizzazione delle informazioni raccolte sia dai registri sia dalle indagini. Gli scenari proposti in questo lavoro andranno sottoposti a sperimentazioni e, ovviamente, andranno valutati attraverso una attenta analisi costi/benefici tenendo ben presente l'obiettivo principale: ottenere una completa integrazione delle indagini sociali oltre a garantire l'integrabilità con il sistema dei registri presenti in Istituto.

Riferimenti bibliografici

CHIPPERFIELD J. O., STEEL D. G. 2009. Design and Estimation for Split Questionnaire Surveys, *Journal of Official Statistics*, Vol.25, No.2, pp.227–244.

EUROSTAT 2013a. D1. Define Estimation Procedures, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D1_Define_estimation_procedures.pdf.

EUROSTAT 2013b. D2. Define Sample Size Determination Procedures. http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D2_Define_sample_size_determination_procedures.pdf.

EUROSTAT 2013c. D3. Representation of the current set of instruments in European Social Surveys, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D3_Representation_of_the_current_set_of_instruments_in_European_Social_Surveys.pdf

EUROSTAT 2013d. D4. Define General Instrument Composition Procedures, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D4_Define_general_instrument_composition_procedures.pdf.

EUROSTAT 2013e. D5/6/7. Instrument Design and Sample Size for the Pilot, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D_5-6-7%20Instrument_Design_and_Sample_Size_for_the_Pilot.pdf.

EUROSTAT 2013f. D8. Report on the available options for drawing the sample of the pilot survey, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D8_Pilot_survey_sample_approach.pdf.

EUROSTAT 2013g. D9. Report on the operational issues of the pilot survey, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D9%20Report%20on%20the%20operational%20issues%20of%20the%20pilot%20survey.pdf.

EUROSTAT 2013h. D10. Frame alignment of the current set of instruments in European Social Surveys, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D10.Frame%20alignment%20of%20the%20current%20set%20of%20instruments%20in%20European%20Social%20Surveys.pdf.

EUROSTAT 2013i. D11. Guidelines for Managing Constraints, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D11_Guidelines_for_Managing_Constraints.pdf.

EUROSTAT 2013j. D12. Roadmap for the integration of European social surveys, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D12_Roadmap.pdf.

ISTAT 2015. La nuova geografia dei sistemi locali. http://www.istat.it/it/files/2015/10/La-nuova-geografia-dei-sistemi-locali.pdf.

KIM J.K., RAO J.N.K. 2012. Combining data from two independent surveys: a model-assisted approach, *Biometrika*, Vol. 99, No.1, pp. 85-100.

LAVALLÉE P., RIVEST L.P. 2012. Capture–Recapture Sampling and Indirect Sampling, *Journal of Official Statistics*, Vol.28, No.1, pp. 1–27.

SUMMARY

Methodological proposals for the integration of social statistics

Since 2011, Eurostat began a reorganization of EU social statistics. This project has evolved over time up to the final version presented at the meeting of Directors of Social Statistics, held in September 2014. The model proposed by Eurostat is based on an approach in *modules* which, by construction, can be pooled and, where possible, can enable the use of information measured at different investigations for the construction of the estimates. Eurostat also presented a roadmap (Eurostat, 2013j) for the implementation of the project which contemplates short, medium and long term: *short term*: study of methods for pooling estimates to be made with the overlap of samples on which were recorded the same variables, regardless of the drawings below; *medium term*: redesign of sample surveys with changing sample size as a function of the new modular approach; *long term*: integrated micro-data for social statistics, powered by both surveys, both the information from the statistical registers.

This paper presents some possible scenarios for the integration of social surveys which arise from the different strategies associated with different sampling designs. The whole purpose is to achieve a complete integration of the system social surveys and ensure maximum integration with the registries system present in Institute.

Claudio CECCARELLI, Istat, clceccar@istat.it
Stefano FALORSI, Istat, stfalors@istat.it

SOCIETÀ E RIVISTA ADERENTI AL SISTEMA ISDS
ISSN ASSEGNATO: 0035-6832

Direttore Responsabile: Dott. CLAUDIO CECCARELLI

Iscrizione della Rivista al Tribunale di Roma del 5 dicembre 1950 N. 1864

Associazione all'Unione Stampa Periodica Italiana

TRIMESTRALE

La copertina è stata ideata e realizzata da Pardini, Apostoli, Maggi p.a.m.@tin.it – *Roma*

Stampato da CLEUP sc
"Coop. Libraria Editrice Università di Padova"
Via G. Belzoni, 118/3 – Padova (Tel. 049/650261)
www.cleup.it