Una sfida per la statistica Europea: l'integrazione delle indagini sociali

Antonella D'Agostino, Giulio Ghellini**, Laura Neri***

*Dipartimento di Studi Aziendali e Quantitativi (Università di Napoli "Parthenope")
**Dipartimento di Economia Politica e Statistica (Università di Siena)

antonella.dagostino@uniparthenope.it, giulio.ghellini@unisi.it, laura.neri@unisi.it

1. Introduzione

Una sfida statistica si aggira per l'Europa. Il combinato disposto della presenza di una perdurante crisi economica, con i suoi sempre più evidenti effetti negativi sulle condizioni sociali dei suoi cittadini, e di un forte vento di *spending review* che soffia sulle istituzioni pubbliche nazionali e sovranazionali del Vecchio Continente, sta ponendo la Statistica Ufficiale di fronte ad un sfida davvero complessa. Come rispondere ai crescenti bisogni informativi richiesti per comprendere e governare i mutamenti e le crisi sociali in atto, riuscendo nel contempo a ridurre i costi per la rilevazione di dati sempre più pertinenti e di elevato standard qualitativo? In buona sostanza, da una parte c'è una maggior richiesta di informazioni statisticamente valide, dall'altra ci sono i produttori di statistiche ufficiali che devono fronteggiare tali richieste con risorse via via decrescenti. In tale contesto un ruolo positivo potrebbe certamente essere giocato dallo sviluppo di metodi di utilizzo statisticamente affidabile dei *Big Data*, da un maggior ricorso allo sfruttamento statistico degli archivi amministrativi e alle opportunità elaborative e di condivisione delle informazioni rese disponibili dai recenti progressi tecnologici. In questo lavoro, però il nostro interesse è rivolto ad approfondire le possibilità di ottenere consistenti guadagni di efficienza nell'articolazione stessa del processo di raccolta dei dati mediante indagini campionarie, ad oggi l'asse portante dell'intero sistema informativo sullo stato della realtà economico-sociale e sulla sua dinamica. A livello Europeo, e principalmente sotto lo stimolo dell'Eurostat, si è recentemente aperto un interessante dibattito sulla modernizzazione delle indagini sociali, dibattito che pone al centro dell'attenzione il modo stesso in cui le indagini sociali sono attualmente progettate e condotte, prefigurando scenari di modernizzazione e di integrazione delle stesse, che seppure dai contorni non sempre ben definiti, offrono lo spunto per alcune rilevanti riflessioni di metodo e operative. Si tratterebbe infatti di superare la centralità delle Indagini, ovvero l'uso di disegnare le indagini come entità di fatto autosufficienti per la produzione di variabili/indicatori validi per una specifica tematica (le forze di lavoro, le condizioni di vita, lo stato educativo della popolazione, ecc.), e di pervenire ad una organizzazione della produzione dei dati centrata invece sulle variabili/indicatori che la statistica ufficiale è chiamata via via a produrre, trasformando quindi le indagini in strumenti, volti ad ottimizzare la raccolta delle informazioni. Certamente un cambio di paradigma rilevante e una sfida ambiziosa per la statistica ufficiale che, nella prospettiva dell'integrazione delle indagini sociali sarebbe chiamata a confrontarsi e perseguire i seguenti obiettivi:

- ridurre il peso per i rispondenti in termini di *overall response burden*, individuando anche una soglia di carico informativo massimo che un singolo rispondente può accettare;
- diminuire il costo complessivo della rilevazione delle informazioni;
- mantenere e se possibile incrementare le potenzialità di analisi dei dati;
- accrescere la flessibilità degli strumenti di rilevazione.

2. Possibili approcci di integrazione

Anche se è evidente che una prospettiva di integrazione delle diverse indagini esistenti in ambito sociale implicherebbe una ben più ampia

discussione sull'insieme delle problematiche legate al disegno di indagine nel suo complesso (i.e. livello di armonizzazione delle variabili, efficienza delle stime, qualità dei dati, disegno degli strumenti d'indagine, ecc.), in questo paragrafo, più modestamente, ci si propone di evidenziare opportunità e problemi connessi a tre possibili approcci di integrazione ben noti in letteratura e già sperimentati in alcuni Paesi: i) indagine multi-scopo; ii) *sample coordination*; iii) stima combinata di indagini separate.

Indagine multiscopo

Le indagini multi-scopo sono un esempio di indagini integrate in quanto il loro disegno campionario si basa sulla rilevazione simultanea (ovvero per mezzo di una singola indagine) di una grande varietà di informazioni che coprono un vasto numero di argomenti che spaziano dal reddito alla spesa per consumi, dalla salute al lavoro all'uso del tempo. Una loro comune caratteristica è la loro flessibilità in termini dell'oggetto trattato e della numerosità campionaria[1]. Lo sviluppo di tale visione integrata di indagine è molto attraente soprattutto nei paesi in via di sviluppo perché costa poco rispetto a più indagini che nel complesso rilevano le stesse informazioni. Le indagini multi-scopo hanno una lunga storia. Per esempio, l'Indagine Campionaria Nazionale Indiana (NSS) ha origini nel lontano 1950. Indagini multi-scopo sono anche utilizzate in Europa, dove l'esempio di maggior rilievo è sicuramente quello dell'Austria dove le statistiche sociali sono per lo più basate sul *Microcensus*, implementato dal 1968. Un'altra iniziativa degna di nota è quella intrapresa in UK dove dal 2003 si sta operando in modo da integrare le indagini sulle famiglie nella *Integrated Household Survey* (IHS). Quando lo scopo diventa però quello di integrare l'insieme delle fonti informative necessarie ed utilizzate in ambito sociale, l'approccio multi-scopo non è certamente la miglior scelta. In primo luogo perché queste indagini non coprono mai un insieme esaustivo di tutte le statistiche sociali e quindi indagini specifiche autonome dovrebbero continuare ad esistere. In secondo luogo, lo strumento di rilevazione dovrebbe coprire un insieme troppo vasto ed eterogeneo di argomenti, spesso indipendenti tra loro e quindi non interessanti da osservare congiuntamente, con una conseguente perdita di parsimonia nella raccolta delle informazioni e un accresciuto *overall burden* per singolo rispondente. Inoltre nel panorama delle statistiche sociali esistono specifiche tematiche che necessitano ad esempio di disegni campionari peculiari (i.e. la longitudinalità per lo studio delle condizioni di vita) e/o di strumenti di rilevazione particolari (i.e i diari per lo studio delle spese per consumi delle famiglie), fattori questi che implicherebbero notevoli problemi di implementazione. Infine un aumento dei contenuti informativi dello strumento di rilevazione comporterebbe problemi di qualità dei dati legati all'inevitabile aumento delle non risposte, così come all' aumento degli errori di misurazione. In buona sostanza un processo di integrazione mediante indagine multi-scopo risulterebbe poco efficiente e scarsamente in grado di soddisfare gli obiettivi ad esso attribuiti.

Sample coordination (sub-sampling)

Il concetto di *sample-coordination* o *sub-sampling* è in pratica un insieme di diversi disegni campionari, accomunati però da procedure di selezione delle unità statistiche coordinate. In particolare, si parla di "coordinamento positivo" quando la selezione di un individuo in un campione fa aumentare la probabilità di selezione in un secondo. In effetti, il *sub-sampling* e il *partial overlap* di campioni sono esempi di "coordinamento positivo". Da un lato il *partial overlap* è un modo per rendere più efficiente un sistema di indagini sociali nel suo complesso. Si ha infatti un guadagno in efficienza includendo individui dal campione di un'indagine nel campione di un'altra indagine che raccoglie informazioni su almeno un sottoinsieme di variabili comuni. Anche questa strategia, quando applicata ad un contesto molto generale quale quello dell'integrazione delle indagini sociali, impone però alcune considerazioni critiche. In primo luogo, sebbene tale schema permetta l'esistenza di differenti disegni campionari tra le indagini (ovvero è un disegno che permette flessibilità), esso presuppone però che alcune differenze tra tali disegni campionari vengano eliminate, aspetto che a sua volta limita la possibilità per disegni campionari più mirati. In secondo luogo, il disegno campionario e la sua gestione diventano complessi. Ad esempio, le decisioni nella fase di progettazione delle indagini dipenderebbero dalla conciliazione della selezione delle variabili o dei moduli con il *partial-overlap* delle molte altre indagini nel sistema. Decisioni

[1] E' però anche vero che coprendo diverse aree tematiche la numerosità campionaria è soggetta a compromessi che possono anche compromettere l'ottimalità del campionamento per alcune di tali tematiche.

complesse sul disegno di campionamento, di solito nelle mani di metodologi, dovrebbero essere prese contemporaneamente con le decisioni sul contenuto delle indagini che di solito è solo nelle mani di esperti sul campo. Infine, c'è un problema legato alla responsabilità delle indagini. La "*sample-coordination*" presuppone che i campioni di tutte le indagini facciano parte del sistema e quindi siano sotto la responsabilità di un unico ente. Necessità che si scontra con la realtà di molti Paesi, dove diverse indagini possono essere gestite da diversi enti, sia pubblici che privati. Infine non da sottovalutare il problema dell'incremento del "*overall burden*" per le unità appartenenti al sottocampione. Anche in questo caso, quindi, la prospettiva di una integrazione in grado di rispondere agli obiettivi di efficienza e di efficacia indicati.

Stima combinata di diverse indagini

La stima combinata di diversi indagine relega il problema dell'integrazione delle diverse indagini sociali alla fase di stima. Ovvero è un'integrazione ex-post. Questo approccio infatti consiste di metodi che possono essere utilizzati per migliorare l'accuratezza delle stime quando una stessa variabile è misurata in due o più indagini. Certamente questo approccio è il più appropriato per un sistema integrato di indagini sociali a livello Europeo perché combina un guadagno in efficienza con un alto livello di flessibilità. Nonostante la notevole flessibilità a livello di disegno campionario in ciascuna delle indagini coinvolte, tale metodologia richiede però un altissimo livello di armonizzazione delle variabili per le quali si vuole un risultato congiunto combinando le diverse indagini (o tra indagini e fonti amministrative). Rispetto alle precedenti metodologie descritte la stima combinata appare come la più appropriata per una visione integrata a livello Europeo, ma purtroppo è proprio il suo presupposto, ovvero la persistenza dello "*status quo*" del sistema attuale di indagini che va contro il principio di "parsimonia" di cui si è discusso all'inizio.

Più di una domanda, allora sorge spontanea, alla fine di questa discussione. Qual è la miglior strada da perseguire nell'ottica dell'integrazione delle indagini sociali a livello Europeo? E' quindi il momento di cambiare completamente paradigma e pensare di costruire una nuova "architettura" su cui basare le statistiche ufficiali in ambito sociale?

3. Una nuova prospettiva per l'integrazione

Presentiamo in questa sezione un approccio molto generale, di fatto incentrato sul concetto di "modulo", nella consapevolezza che si tratta ovviamente di uno dei possibili approcci atti a perseguire l'obiettivo dell'integrazione delle statistiche sociali. Pur tuttavia approccio degno di nota perché è in tale direzione che Eurostat sta cercando di perseguire gli obiettivi elencati precedentemente. Si tratta in buona sostanza di un'ipotesi di lavoro sulla quale ci auguriamo possa aprirsi un dibattito nel prossimo futuro sulla sua effettiva perseguibilità. Il sistema di indagini integrato in moduli, è basato sostanzialmente sulla "scomposizione" delle indagini sociali esistenti, e sulla riorganizzazione delle relative variabili in moduli e nella successiva ricostruzione di questi ultimi in un sistema integrato di micro data-set, tra loro collegabili. Tale approccio, può essere descritto come segue:

(1) le indagini sociali EU vengono scomposte in m moduli mutuamente esclusivi e nel loro insieme in grado di rispondere ai bisogni conoscitivi prefissati; ogni modulo è composto di un certo numero di variabili (ad esempio, tra 10-15 variabili), inerenti uno specifico aspetto da rilevare congiuntamente;

(2) gli m moduli vengono raggruppati in k strumenti, definiti come una sequenza/articolazione di moduli; non vi sono vincoli, se non quelli legati all'*overall burden*, sul numero di moduli che vanno a formare i singoli strumenti. Ogni strumento I_j, $j=1...k$, permette quindi di osservare congiuntamente, sulle stesse unità statistiche, le variabili contenute nei moduli che lo compongono;

(3) in ogni Paese, viene estratto un campione probabilistico s da un'unica lista di campionamento;

(4) il campione viene casualmente suddiviso in k sottocampioni, ottenendo quindi k campioni casuali disgiunti (*negatively coordinated*);

(5) ogni sottocampione j ha le seguenti caratteristiche: (a) una definita numerosità campionaria n_j (non c'è il vincolo di uguaglianza tra le numerosità campionarie); (b) uno *strumento Ij* associato;

(6) ognuno degli n_j rispondenti del sottocampione j risponde in modo completo a tutti i moduli dello strumento I_j;

(7) in certi strumenti possono essere presenti: (i) filtri tra moduli - un certo modulo b viene somministrato solo in corrispondenza di una certa

risposta data in un altro modulo *a* - ; (ii) filtri all'interno dei moduli - in un certo modulo *a*, un rispondente è chiamato a rispondere a certi quesiti solo se egli ha fornito una certa risposta ad un precedente quesito -.
Tale processo presuppone innanzitutto la precisa definizione delle variabili target e degli indicatori che tale sistema integrato deve essere in grado di produrre al fine di guidare la scelta delle variabili da assegnare ad ogni modulo.
Tale scelta dovrà inoltre tenere presenti:
- Priorità nella scelta dei quesiti del/i modulo/i *core*, comuni a tutti gli strumenti. Nel/i modulo/i *core* dovranno essere presenti variabili per le quali è di prioritario interesse studiare la distribuzione congiunta con variabili/indicatori rilevati in altri moduli. Si tratta generalmente di variabili socio-demografiche alle quali potrà essere opportuno aggiungere anche variabili per le quali è richiesto una più elevata numerosità campionaria al fine di massimizzare la precisione delle stime.
- Inserimento nel medesimo modulo di variabili che sono ragionevolmente associate e per le quali è necessario effettuare stime congiunte (Gonzalez and Eltinge, 2007).
- Assegnazione di variabili correlate anche a moduli diversi al fine di permettere l'applicazione di metodi di *matching* statistico, per esempio attraverso procedure di imputazione (Thomas et al, 2006).
- Attenzione al bilanciamento dei moduli tra gli strumenti al fine di migliorare l'efficienza sia in termini di costi che di carico di risposta.

L'architettura appena descritta sembrerebbe quindi rispondere agli obiettivi presentati nell'introduzione in quanto dovrebbe garantire: la riduzione dell'*overall response burden*, una diminuzione del costo complessivo di rilevazione minimizzando la sovrapposizione di informazione raccolte con strumenti diversi, la possibilità di incrementare le potenzialità di analisi attraverso l'analisi congiunta di moduli appartenenti a strumenti diversi, la maggiore flessibilità degli strumenti di rilevazione ottenibile attraverso ragionevoli spostamenti dei moduli tra strumenti e/o l'introduzione di nuovi moduli per rispondere a nuove esigenze informative.
Tuttavia per una effettiva attuazione di tale approccio restano da superare molteplici e rilevanti difficoltà, che possono essere così sintetizzate: i) difficoltà legate alla ampiezza ed eterogeneità delle tematiche a cui le indagini sociali sono chiamate a rispondere, ii) difficoltà di tipo organizzativo, iii) problemi metodologici.

In merito al primo aspetto (i) occorre soprattutto uno sforzo di razionalizzazione delle esigenze conoscitive al fine di ottimizzare la definizione dei moduli e di garantire il più elevato livello possibile di armonizzazione delle variabili presenti in diversi moduli.
Dal punto di vista del management/organizzazione della rilevazione (ii) a livello di singolo Paese sarebbe auspicabile che tutte le rilevazioni fossero sotto la responsabilità di un unico ente (gli Istituti Nazionali di Statistica) al fine di garantire un processo di produzione di dati coordinato; inoltre prevedere l'integrazione di informazioni provenienti da diversi strumenti presupporrebbe anche una effettiva sincronizzazione della disponibilità delle informazioni provenienti dalle diverse rilevazioni. In tale prospettiva appare evidente la necessità di un più stringente ruolo di coordinamento sovranazionale da parte di Eurostat, ruolo che favorirebbe la comparabilità dei dati a scapito però di una perdita di autonomia dei singoli Paesi. E questo un argomento che, seppur di grande rilevanza, non è pensabile di poter sviluppare in tale lavoro, ma che certamente richiederà nel prossimo futuro un'ampia e ponderata riflessione.
Infine, i problemi metodologici (iii) derivanti da una così complessa architettura sono molteplici; ne citiamo solo alcuni connessi all'assetto attuale delle indagini sociali. Ad esempio, le diverse popolazioni obiettivo, i molteplici periodi di riferimento delle variabili e le diverse liste di campionamento tra le esistenti indagini; la necessità di definire stimatori adeguati basati sui diversi strumenti in cui un dato modulo si presenta; la definizione di opportuni algoritmi per il calcolo della numerosità campionaria ottimale e tale da garantire prefissati livelli di precisione agli stimatori. Tutti argomenti che, allo stato attuale, sono oggetto di studi e approfondimenti da parte di Eurostat.

Riferimenti bibliografici

Eltinge, J.L. (2007), "Multiple Matrix Sampling: A review", American Statistical Association, Proceedings of the Section on Survey Research Methods, 3069–3075.
Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. and Johnson, C.L. (2006) An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey *Survey Methodology*, 32,217.