

Un sistema di registri statistici basato sull'integrazione di informazioni da archivi amministrativi, indagini statistici e stime di popolazione

Marco Di Zio, Piero Falorsi, Marco Fortini

MILeS2016

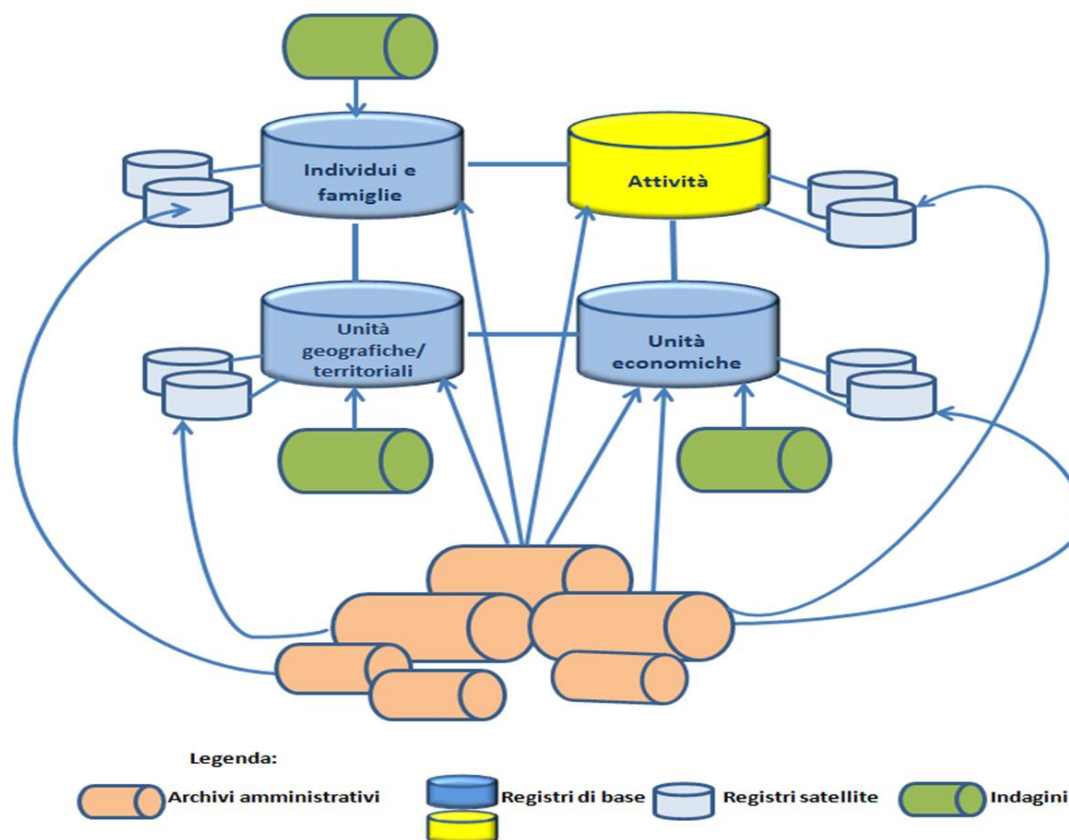
Milano - Impresa, Lavoro e Società 2016

In due parole

- Una nuova organizzazione basata sull'uso di dati di fonte amministrativa e statistica organizzate in un sistema di archivi
- Metodologie di integrazione delle fonti per migliorare l'informazione a costi non crescenti
- Nuove sfide per la qualità dei dati: dal controllo della variabilità a quello della distorsione
- Cosa significa produrre statistiche da archivio e perché è necessario programmare gli Output
- Il ruolo delle indagini statistiche in un sistema basato sulle fonti amministrative

Il Sistema Integrato dei Registri (SIR)

- Input: Archivi amministrativi e indagini statistiche
- Output: registri di base, satelliti, tematici



Uso intensivo di integrazione di dati amministrativi ed indagini statistiche

- Integrazione a livello di **microdati**
 - Record linkage
 - Matching statistico
- Integrazione a livello di **macrodati**
 - Ponderazione vincolata
 - Modelli per piccole aree
 - Modelli gerarchici bayesiani (a priori informative derivanti da fonti ausiliarie)
- **Risultato:** incremento dell'informazione disponibile rispetto a quella di ciascuna fonte presa singolarmente
- **Conseguenza:** Il registro diventa 'statistico', ovvero l'incertezza si trasferisce in modo quantificabile alle unità del registro, e alle variabili

Il registro diventa un prodotto statistico

- Le statistiche da registro si dovrebbero ottenere per somma di record
- L'incertezza propria delle stime dovrebbe essere distribuita sui micro dati inclusi nel registro
- Alcune delle fonti di incertezza da gestire
 - Unità eleggibili escluse dal registro
 - Unità non eleggibili incluse nel registro (tra cui i duplicati)
 - Mancate risposte totali e parziali
 - Accuratezza sulle variabili (distorsione e variabilità di misurazione, campionamento e processo)

Cambia il processo, si modificano le fonti di errore

- Se con le indagini statistiche il problema prevalente è quello della variabilità campionaria
- Con le statistiche basate su registri cresce l'attenzione verso la pertinenza
 - Errori di specificazione della popolazione oggetto e sui fenomeni da misurare

Quali sono le conseguenze se non teniamo conto delle incertezze?

- Se nelle indagini campionarie l'accento era sul problema della variabilità ora l'attenzione si sposta sulla distorsione delle stime
- Esempio 1: Conteggio della popolazione dalle Anagrafi
 - Individui che dimorano abitualmente sul territorio di interesse senza esservi residenti (lavoratori e studenti fuori sede) inducono una distorsione per difetto nei conteggi
 - Individui che risiedono senza dimorare effettivamente sul territorio contribuiscono a una distorsione positiva
- Esempio 2: Errori di classificazione nella variabile NACE

Dal momento che l'integrazione è centrale nel processo di costruzione degli archivi lo diventa anche l'errore ad essa collegato

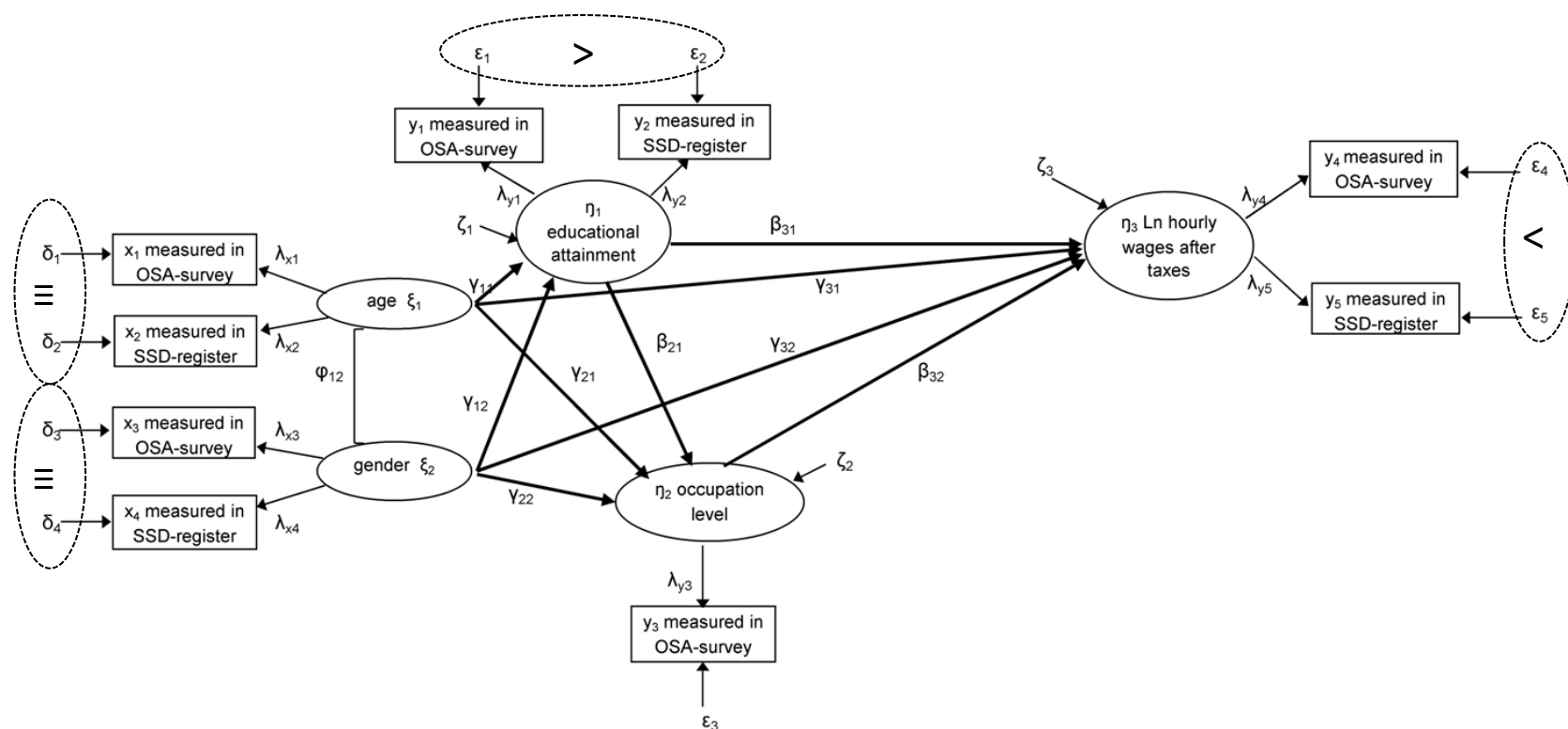
- Record linkage
 - Variabilità e distorsione per effetto dei mancati link
 - Diluizione dei parametri di associazione nei modelli di dipendenza per effetto dei falsi link
- Matching statistico
 - Si conservano solo le associazioni condizionate alle variabili di matching
 - Componente di variabilità indotta dalla 'mass imputation'

Come tenere conto dell'incertezza?

- Misure di coerenza interna al sistema
 - Uso di modelli a struttura latente
- Acquisizione di nuova informazione
 - Indagini di controllo
- Obiettivi
 - Valutare le fonti di errore e fornire meta-informazione agli utenti
 - Identificare gli interventi migliorativi sul processo e sugli input
 - Correggere gli errori nelle **principali** stime ottenibili dall'archivio (benchmark)
 - Trasferire l'incertezza a livello di dato elementare?

Modelli di strutture latenti e valutazione della qualità

Esempio tratto da Bart F. M. Bakker, (2012), “Estimating the validity of administrative Variables” *Statistica Neerlandica* Vol. 66, nr. 1, pp. 8–17



Stimatori Cattura – Ricattura (CR)

- La componente di sottocopertura può essere indagata solo mediante informazione ausiliaria rispetto a quella disponibile nell'archivio A
- Indichiamo con B un'indagine di controllo indipendente da A
- Entrambe le fonti affette da sottonotifica di unità eleggibili

	B		
A	In	Out	
In	x_{11}	x_{12}	x_{1+}
Out	x_{21}	x_{22}	
	x_{+1}		N

- Si sfrutta l'ipotesi di indipendenza

$$\frac{x_{11}}{x_{+1}} = \frac{x_{1+}}{N}$$

- E si deriva lo stimatore di Lincoln –Petersen (LP)

$$\hat{N} = \frac{x_{1+}x_{+1}}{x_{11}}$$

Stimatori CR: gestire la sovra-copertura

- Lo stimatore è distorto se le fonti incorporano unità non appartenenti alla popolazione (sovra-copertura)
- s : numero di unità non appartenenti alla popolazione

$$\hat{N}^* = \frac{x_{1+}^*}{x_{11}} = \frac{(x_{1+} + s)x_{+1}}{x_{11}} = \frac{x_{1+}x_{+1} + s \cdot x_{+1}}{x_{11}} > \frac{x_{1+}x_{+1}}{x_{11}} = \hat{N}$$

- Occorre ricorrere a un procedimento capace di stimare s
 - Indagine da lista
 - Ritorno sul campo per l'indagine B

Stimatori CR: la correzione del Registro di base

- Si possono introdurre pesi per ottenere tramite somma degli individui nel RB il totale di popolazione stimato rispetto a domini predefiniti
- Se h è il generico strato ottenuto come combinazione delle modalità di alcune variabili di base (es. sesso, classe età, Regione di dimora,...)
- $\omega_h = \frac{\hat{N}_h}{N_h}$, peso attribuito a tutti gli individui che ricadono nello strato h
- La somma dei pesi attribuiti a tutti gli individui del Registro per una data Regione riconduce alla stima corretta della popolazione della regione
- Analoghi ai pesi di campionamento. Gli individui nell'archivio ne rappresentano anche altri
- Differenza: il peso ω_h può essere minore di 1 se $\hat{N}_h < N_h$
- Il peso corregge rispetto ad una singola popolazione. Se il registro rappresenta altre popolazioni si potrebbero dover prevedere altri pesi

Alcuni vantaggi e limiti

- Cosa accade per i Comuni di una stessa Regione? Ricevono una correzione proporzionale
 - Se un Comune ha un errore superiore a quello medio di Regione la sua popolazione è corretta per difetto
 - Se un Comune ha un errore inferiore a quello medio di Regione la sua popolazione è corretta per eccesso
- Se: Variabili di interesse (es. titolo di studio, condizione professionale) correlate con le variabili rispetto alle quali è corretto l'errore di copertura,
 - Allora: Correzione delle distribuzioni per queste variabili di interesse
- Va da se che la correzione è solo parziale se non si corregge la copertura per tutte le variabili correlate con quelle di interesse

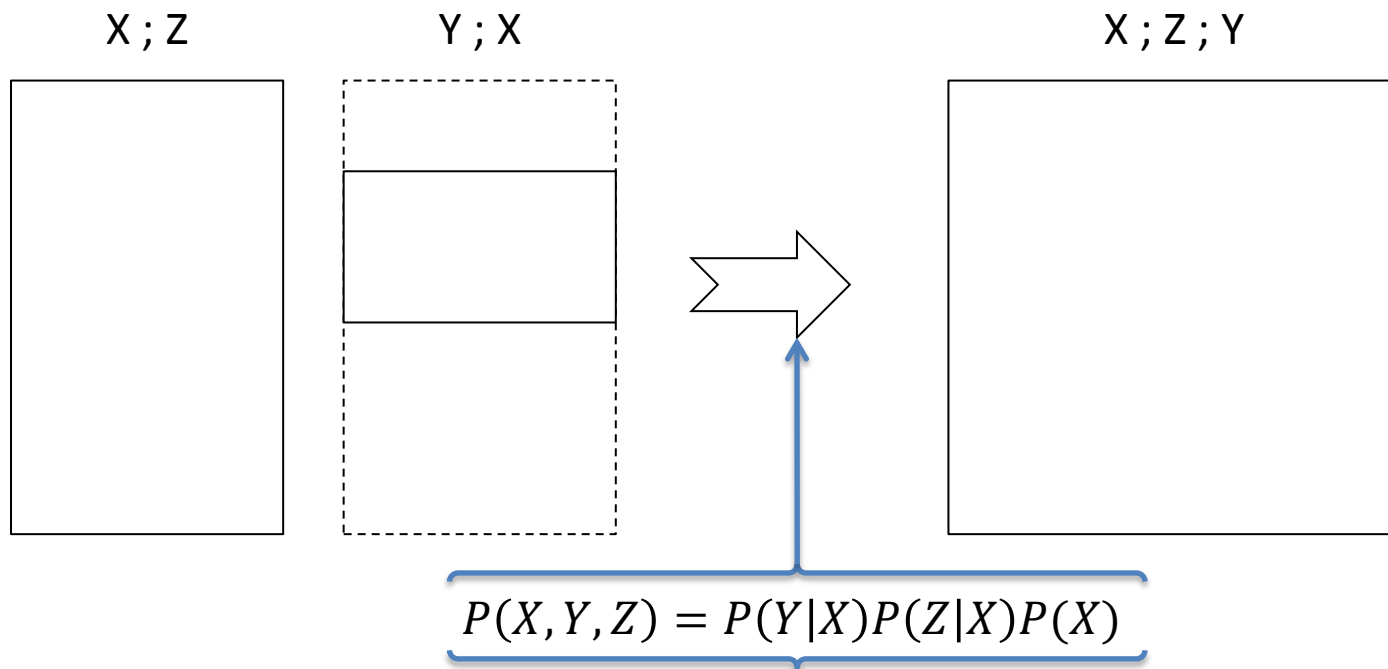
**Ovvero: è necessario programmare gli output
già in fase di progettazione**

Altro esempio: correggere le variabili

- Abbiamo visto come correggere la numerosità del RB (popolazione target) per essere più aderente alla popolazione oggetto
- Analoghe correzioni si possono operare sulle variabili, se lo si ritiene necessario...
 - Lo stato di occupazione si può desumere dai dati amministrativi per gli occupati
 - Ma gli inoccupati e quelli in cerca di occupazione sono misurati solo dall'indagine sulle forze di lavoro (FFLL)
- L'integrazione statistica tra il campione FFLL e i dati amministrativi migliora la stima (macro-dato) della distribuzione della variabile, ma con una incertezza da modello
- L'imputazione da modello sugli individui del RB permette di rispettare 'per somma' le stime, ma presenta problemi analoghi a quelli già mostrati
- L'imputazione non 'incorpora' la variabilità da modello (ricorso alla multiple imputation...)

Correggere le variabili

- Le correzioni impongono di fare attenzione all'informazione che stiamo incorporando nel RB e a quella che non consideriamo



- Operazione comoda ma problematica rispetto all'inferenza su $P(Y, Z)$

Il nuovo ruolo delle indagini statistiche

Da fonte principale a fonte ausiliaria

- Stima di componenti di errore
 - Copertura
 - Specificazione
 - Misurazione
- Stima di variabili non misurate o misurabili dalle fonti amministrative
- Stima di associazioni non misurabili dalle fonti amministrative
- Indagini più piccole, ma più complesse. Una parte del risparmio deve essere reinvestito in qualità

Per riassumere

- Integrare le fonti è un bene perché ne aumenta il potere informativo
- Sintetizzare le informazioni in un RB è comodo perché
 - Migliora l'omogeneità dei processi di produzione
 - Aumenta la flessibilità delle analisi
- Il processo cambia e vanno considerate nuove fonti di errore
- L'errore può essere controllato, ma operativamente solo rispetto a domini predeterminati → ciò impone di progettare attentamente gli output
- Il ruolo delle indagini statistiche
 - Diventa ausiliario alle fonti amministrative
 - Resta centrale per assicurare la completezza la pertinenza e l'accuratezza dell'informazione prodotta