Nuove prospettive nella integrazione di fonti di dati nelle statistiche pubbliche

Intelligenza artificiale e statistica

OLBIA, STATCITIES 2023

Fabio Crescenzi, *Istat, Dipartimento per la produzione Statistica*

Evoluzione delle Statistiche Ufficiali

Il Sistema Integrato dei Registri (ISSR)

Evolutione del modello di produzione. Istat ha costruito ISSR e ha adottato un modello di produzione basato su ISSR

Integrazione delle fonti. L'ISSR è alimentato da una molteplicità di dati, ottenuti in primis da fonti amministrative, poi da rilevazioni e fonti alternative (big data, telerilevamento, ecc.)

Componenti dell'ISSR: Registri statistici di base (RSB); Registri statistici estesi (RSE); Registri statistici tematici (RST)

Il Sistema Integrato dei Registri (ISSR)

Più di 26 registri sono già stati integrati, su 3 dimensioni fondamentali, unità economiche, persone e luoghi

Dal lato delle unità economiche ci sono i registri di base dell'agricoltura, delle imprese, degli enti pubblici, degli enti privati,…

L'altra unità fondamentale l'analisi è l'individuo. RBI è il registro base da cui partono gli altri registri estesi e tematici sulle persone: Anagrafe stato di salute e invalidità e Anagrafe redditi collegandosi agli altri registri.

L'obiettivo del Registro statistico di base dei luoghi (RSBL) è quello di potenziare gli strumenti territoriali per favorire una lettura sempre più integrata dei fenomeni rafforzando la capacità di georeferenziare le unità statistiche

I registri aggiuntivi svolgono una funzione di ponte tra le 3 dimensioni fondamentali come nel caso di Occupazione e Reddito

Punti di forza del nuovo modello

Valorizzazione degli output. La grande mole di informazioni contenute in ISSR offre opportunità per meglio rispondere al crescente bisogno di dati su fenomeni complessi (per finalità di conoscenza, ricerca, definizione e valutazione di politiche, ecc.)

Dati spaziali e longitudinali. Nuove opportunità nella produzione di dati spaziali e longitudinali e di dati di dominii di analisi tra loro collegati.

Efficienza ed efficacia: ISSR produce anche significativi guadagni in termini di efficienza ed efficacia e una riduzione del burden, in particolare di quello dovuto alle indagini sul campo.

Quali sfide per le statistiche ufficiali?- 1

La qualità.

Le misure di qualità sono per lo più definite su base inferenziale rischiano di diventare obsolete.

Il sistema ibrido prevede una molteplicità di fonti, basate su registri, big data e indagini.

L'integrazione di dati provenienti da più fonti con metodi complessi richiede un nuovo quadro concettuale per misurare la qualità

Quali sfide per le statistiche ufficiali?- 2

Strategie d'azione.

Possono esserci discrepanze tra le stime ottenute da dati provenienti da fonti diverse. In alcuni casi le differenze sono fisiologiche e accettabili, in altri richiedono interventi. L'Istat è al lavoro per classificare e misurare le incoerenze, valutare il rischio associato a ciascuna strategia di stima e definire le possibili strategie di azione.

Quali sfide per le statistiche ufficiali? - 3

Ridisegno metodologico delle survey.

Dalle survey tradizionali occorre passare a survey esplicitamente progettate per l'integrazione di dati di più fonti.

Le istituzioni ufficiali e scientifiche hanno già iniziato a rivedere i metodi, tuttavia occorre procedere in modo più incisivo.

Quali sfide per le statistiche ufficiali? - 4

Il **Machine learning** ha un grande potenziale di impiego nelle statistiche ufficiali.

Può rendere la produzione di statistiche più efficiente automatizzando determinati processi o fornire assistendo nella esecuzione di altri.

Consente inoltre alle organizzazioni statistiche di utilizzare nuovi tipi di dati come dati e immagini disponibili sui social media.

Quali sfide per le statistiche ufficiali? - 5

Con riferimento al Machine learning, interessanti i risultati di due iniziative internazionali: l'UNECE High-Group level sulla modernizzazione delle statistiche ufficiali (HLG-MOS) e Office for National Statistics (ONS) del Regno Unito – UNECE

Machine Learning Group 2021 approvato dall'HLG-MOS.

https://unece.org/statistics/publications/machine-learning-official-statistics

Quali sfide per le statistiche ufficiali? - 6

Il Quality Framework for Statistical Algorithms (QF4SA) è un primo tentativo per guidare gli statistici ufficiali nell'uso di algoritmi di machine learning sul fronte della qualità nella produzione di statistiche ufficiali.

Le cinque dimensioni del QF4SA forniscono spunti di riflessione statistici ufficiali nella scelta tra diversi algoritmi.

Progetto COERES. Casi di studio

In 6 aree sono stati avviati casi di studio sulla **COE**renza fra **RE**gistri e **S**urvey.

E’ stato disegnato e compilato dagli uffici competenti un questionario per confrontare la qualità di dati di fonte amministrativa con quella delle indagini.

1. Mercato del lavoro;
2. Redditi e EUSILC;
3. Popolazione abitualmente residente;
4. Agricoltura;
5. Disabilità and mortalità;
6. ICT

Progetto COERES. Classificazione delle variabili

Fonti A (Amministrative o Big Data) Fonti B (Survey campionarie o censuarie)

1.1 - Variabili del registro stimate da fonti A (Amministrativa e/o Big Data) e per le quali non si dispone di dati ausiliari da fonte B (indagine campionaria e/o censuaria);
1.2 - Variabili del registro stimate da fonti A e per le quali si dispone di dati ausiliari da fonti B ma che non sono stati oggetto di integrazione;
2.1 - Variabili del registro stimate da dati da fonti B e per le quali non si dispone di dati ausiliari da fonti A;
2.2 - Variabili del registro stimate da dati da fonti B e per le quali si dispone di dati ausiliari da fonti A ma che non sono stati oggetto di integrazione;
3 - Variabili del registro stimate mediante integrazione di Dati di Fonti A e B.

Progetto COERES. Alcuni risultati

1. **Misura della coerenza dei dati di fonte A e B.** Solo in due casi di studio si segnala un elevato grado di coerenza tra le fonti di tipo A e B e si tratta dei casi riferiti a Mercato del lavoro e Popolazione abitualmente residente.
2. **Uso di indicatori o metodi per la valutazione della coerenza.** La valutazione del grado di coerenza, tuttavia, non si basa attualmente sull'utilizzo di indicatori armonizzati o metodi comuni. Vengono impiegati metodi ad hoc, principalmente utilizzando statistiche descrittive.
3. **Motivi di incoerenza.** I motivi più frequenti di incoerenza tra i dati sono differenze dovute a sfasamenti temporali, errori di misura, differenze di definizioni e classificazioni adottate.

Progetto COERES. Prossimi passi

I seguenti 4 casi di studio sono ora candidati per essere studiati meglio.

1. Mercato del lavoro,
2. Reddito e EUSILC
3. Popolazione abitualmente residente,
4. Agricoltura