Istat

Processi e prodotti con nuove fonti in Istat

3
Maggio
2023

Mauro Bruno, Alessandra Righi

#IstatWebinar

Indice della presentazione

- Contesto: un percorso iniziato dieci anni fa…
- Prodotti con nuove fonti di Big Data:
  - Scanner Data
  - Dati testuali da social network
  - Immagini satellitari
- Conclusioni

Un percorso iniziato dieci anni fa (2/2)

➢ Contesto Istat

- Nel 2013 è stato istituito il primo comitato Big Data, con il compito di definire la strategia Big Data dell'Istituto…
- … per arrivare alla Roadmap delle Statistiche smart da Big Data o Trusted Smart Statistics

➢ Principali obiettivi

- Apertura della statistica ufficiale a nuovi tipi di dati, fonti e tecnologie per la produzione statistica ufficiale
- Aumentare la tempestività nella produzione della statistica ufficiale
- Arricchire l'offerta statistica con nuovi prodotti
- Incrementare la rilevanza della statistica ufficiale nel new data ecosystem

L'adozione di Big Data nel contesto della statistica ufficiale ha richiesto grandi investimenti in diversi contesti: **metodologico**, **architetturale**, **tecnologico**, **organizzativo**, **legale** e sugli aspetti di **privacy**

Un percorso iniziato dieci anni fa

- **Definizione di Statistiche smart da Big data** in ambito Eurostat: prodotti forniti per mezzo di sistemi intelligenti, che incorporano cicli di vita dei dati verificabili e trasparenti che ne garantiscono la validità e l'accuratezza dei risultati nel rispetto della privacy e della riservatezza

- **I nuovi prodotti statistici realizzati con fonti Big data o miste e con nuove metodologie, consentono di indagare nuovi fenomeni o vecchi fenomeni da diverse prospettive senza però derogare ai principi della statistica ufficiale**
- E' necessario trasformare non solo le infrastrutture ma anche i processi e architetture di business
- Fonti e procedure di elaborazione possono essere esterne agli INS, occorre condividere con i titolari dei dati il trattamento per gli scopi e algoritmi concordati,
- **Organizzazione multi-livello del flusso di lavoro dell'elaborazione dei dati e Quadri metodologici modulari** con adozione di modelli standard per la gestione dei processi di produzione (GSBPM)

- Tutte le sperimentazioni e i prodotti realizzati sono inseriti nel **Programma statistico nazionale** a garanzia del rispetto di tutte le norme previste per la statistica ufficiale

- Anche in caso di uso nuove tecniche (webscraping, smart survey…), l'Istat ha l'obbligo di **comunicare alle imprese o ai cittadini per quali specifici usi saranno trattati i dati**

Una nuova frontiera per la statistica ufficiale (1/2)

➢ Uso di scanner data nell'indagine dei prezzi al consumo

- Gli indici dei prezzi al consumo misurano **le variazioni nel tempo dei prezzi di un insieme di prodotti (paniere) rappresentativo di tutti i beni e servizi destinati al consumo finale delle famiglie**, acquistabili sul mercato attraverso transazioni monetarie (sono escluse le transazioni a titolo gratuito, gli autoconsumi, i fitti figurativi, ecc.).
- I dati che concorrono alla costruzione degli indici mensili dei prezzi al consumo sono raccolti attraverso l'utilizzo di una pluralità di fonti:
  - la rilevazione territoriale, condotta dagli Uffici comunali di statistica (UCS);
  - la rilevazione centralizzata, condotta dall'Istat direttamente o attraverso la collaborazione con grandi fornitori di dati;
  - **gli scanner data provenienti dalla Grande Distribuzione Organizzata (GDO);**
  - la fonte amministrativa

Indagine dei prezzi al consumo (2/2)

➢ Scanner data

Tramite l'acquisizione dei dati scanner dalla GDO **vengono rilevati tutti i prodotti cosiddetti grocery** (beni alimentari confezionati e beni per la cura della casa e della persona) e alcuni prodotti relativi alla frutta e verdura fresca a peso imposto (**13,6% in termini di peso**)

L'istituto si è dotato di una **piattaforma cloud on-premise** per la gestione di questa fonte

➢ Web scraping

L'utilizzo raccolta dati da web per il calcolo degli indici dei prezzi al consumo è andato aumentando per: i) integrare o sostituire le tradizionali fonti di indagine; ii) per ridurre l'onere per gli intervistati e i costi della raccolta dei dati; iii) per migliorare l'accuratezza delle stime IPC

➢ Statistica sperimentale (proposta)

Indici spaziali dei prezzi al consumo

L'obiettivo è fornire una stima degli indici spaziali dei prezzi al consumo a livello regionale. Gli indici spaziali dei prezzi al consumo misurano le differenze tra il livello medio dei prezzi di un paniere standard di prodotti in una determinata area geografica e quello medio calcolato per il complesso delle aree.

Una nuova frontiera per la statistica ufficiale (1/3)

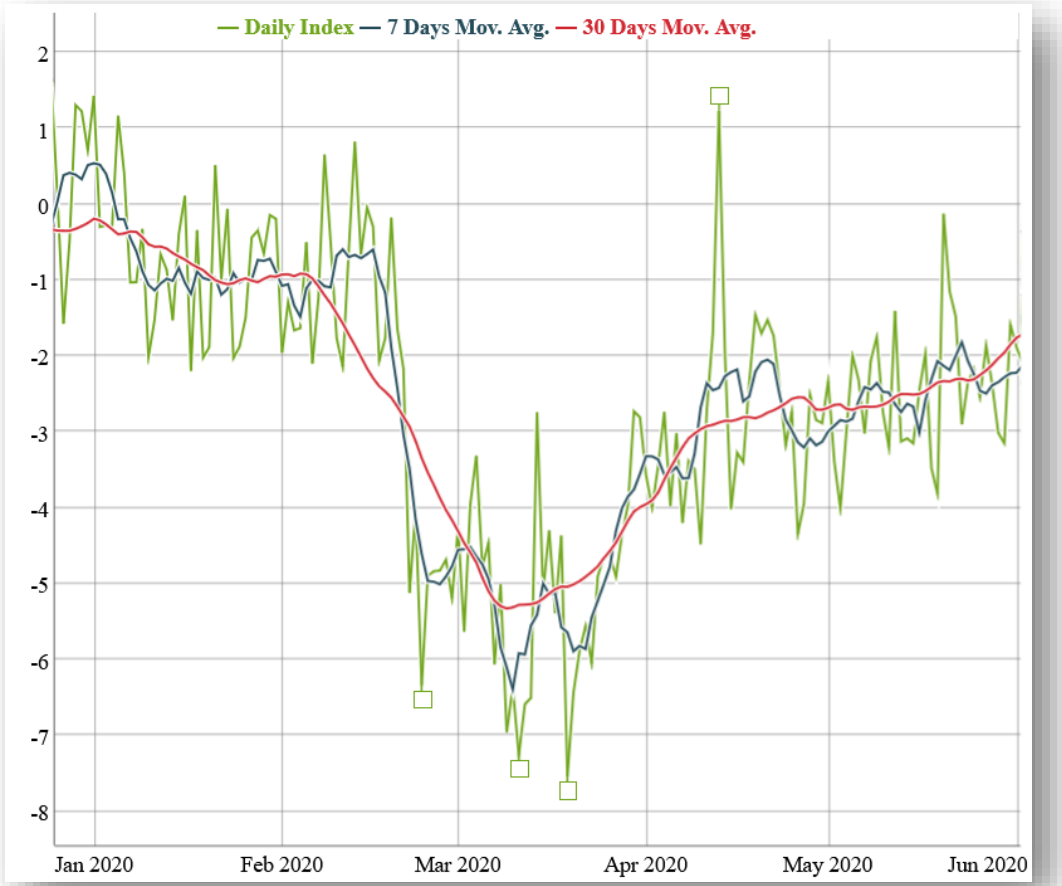
➢ Dati testuali da social network

- I social network sono diventati uno strumento di comunicazione molto popolare tra gli utenti di Internet, con milioni di messaggi scambiati quotidianamente attraverso piattaforme social.
- A differenza dei dati di indagine e delle fonti amministrative, il meccanismo di generazione dei dati dei social media non cade sotto il controllo statistico ed è (generalmente) sconosciuto.
- Non esiste **un modo rigoroso per garantire la validità generale delle informazioni statistiche derivate dai dati di Twitter**. Gli utenti italiani di Twitter non possono essere considerati un campione rappresentativo della popolazione italiana. Tuttavia **le informazioni derivate dai social network sono molto utili per l'analisi sociale ed economica.**
- Per colmare questa lacuna informativa sono state aggiunte nuove domande al questionario dell'indagine multiscopo "**Aspetti della vita quotidiana**" - Anno 2023 al fine di caratterizzare gli utenti e il loro comportamento social, e stimare la frequenza di utilizzo di ogni social network (Facebook, Twitter, Instagram, Linkedin, ecc.)

Social Mood on Economy Index

(2/3)

- Il Social Mood on Economy Index (SMEI) è una delle prime statistiche sperimentali prodotte da Istat a partire da fonti non tradizionali (dati testuali di Twitter)
- L'indice fornisce misure giornaliere del sentiment italiano sull'economia, derivate da campioni di tweet pubblici in lingua italiana, catturati in streaming
- La sentiment analysis è effettuata utilizzando un approccio **unsupervised, lexicon-based**



Nel 2021 SMEI ha subito una **profonda revisione metodologica**, poiché la serie aveva registrato solo parzialmente il trend negativo legato al COVID

Progetti di ricerca da fonti «social media» (3/3)

- **Violenza di Genere**

Il progetto di analisi del sentiment e dell'emotion sulla violenza di genere rappresenta un passo decisivo per lo studio del fenomeno. I messaggi vengono utilizzati per misurare come viene rappresentato il fenomeno, al fine di verificare se i social network vengono utilizzati per contrastare, condannare, isolare la cultura degli stereotipi di genere o se, al contrario, contribuiscono alla sua diffusione e al linguaggio di odio.

- **Linguaggio d'odio**

L'obiettivo del progetto è l'identificazione e quantificazione del linguaggio d'odio su Twitter con riferimento a diverse popolazioni target (immigrati, popolazione LGBT, etc.). L'intento è di studiare l'evoluzione del linguaggio di odio nel tempo, classificarne le principali dimensioni e individuare dei criteri per misurarne l'intensità/gravità.

- **Guerra in Ucraina**

il recente conflitto russo-ucraino viene studiato e analizzato da un punto di vista sociale, i.e., la percezione della guerra da parte degli utenti di Twitter e come questa guerra influenzi l'opinione su aspetti importanti come l'energia.

Analisi delle conversazioni sul tema guerra Russia - Ucraina

Rapporto Annuale 2022

LA GUERRA IN UCRAINA NEI *TWEET* ITALIANI

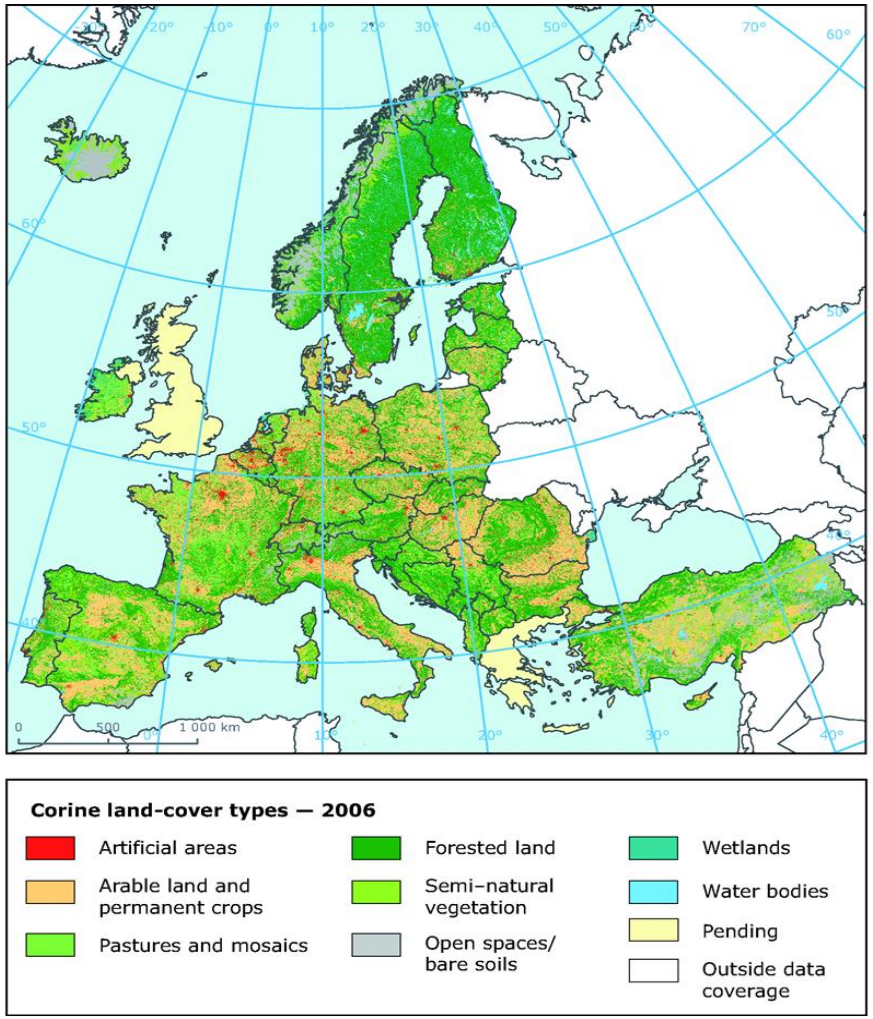
Word cloud
24 febbraio – 8 marzo 2022

Una nuova frontiera per la statistica ufficiale (1/2)

➢ Stima della copertura del suolo (LC)

- La copertura del suolo è la copertura biofisica della superficie terrestre. Una definizione viene dalla direttiva 2007/2/CE: **la copertura fisica e biologica della superficie terrestre comprese le superfici artificiali, le zone agricole, i boschi e le foreste, le aree seminaturali, le zone umide, i corpi idrici.**
- A livello internazionale queste stime vengono fatte in diversi progetti, quali ad esempio: **LUCAS** (indagine campionaria condotta da Eurostat), **CORINE** (programma copernicus)

Queste indagini sono molto costose, richiedono un grande carico di lavoro per i rilevatori, hanno una bassa frequenza temporale, …



Immagini satellitari per stima della copertura del suolo (LC) (2/2)

- Realizzazione di una architettura integrata che fornisce risultati accurati per tutte le classi di land cover.
- Data **un'immagine satellitare in ingresso di un'area di interesse**, vogliamo un sistema automatico (**algoritmi di deep learning**) in grado di:
  - classificare il territorio secondo una **tassonomia LC**
  - **quantificare l'area (o la proporzione)** di territorio coperta da ciascuna classe LC
- I principali benefici attesi sono:
  - statistiche molto **tempestive** sulla copertura del suolo
  - consentire la stima della copertura del suolo **a livello di area subregionale**
  - riduzione del costo di produzione delle stime

➢ Plans for the future…

Nel corso del 2023, partiranno alcune sperimentazioni legate all'utilizzo di immagini satellitari: Change Detection in Urban Areas using Satellite Data, deforestazione, verde urbano, …

Verde urbano

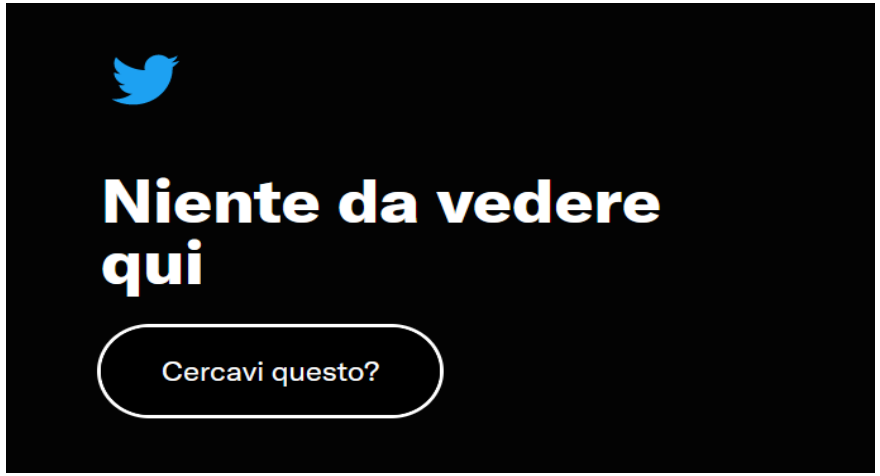
Esempio definizione di area verde
Villa Borghese (Roma)

Conclusioni

➢ … un percorso iniziato molti anni fa

- Le sperimentazioni avviate negli ultimi anni hanno **prodotto risultati molto incoraggianti**: integrazione con le fonti tradizionali, statistiche sperimentali e progetti di ricerca
- **Non siamo soli**! I risultati ottenuti sono frutto di collaborazioni con altri istituti di statistica, con le università e anche con partner privati (**collaborazione Istat-Vodafone**)
- Nel corso del 2023 partiranno diversi progetti di ricerca ESSNet finanziati da Eurostat, in cui Istat **ha ruoli di coordinamento** :
  - Tender Multy MNO (integrazione dati da più provider)
  - ESSNet Multi Source (integrazione dati telefonia con fonti tradizionali e Big Data)
  - ESSNet Trusted Smart Surveys II
  - Centre of Excellence for Machine Learning

Le fonti Big Data comportano dei **rischi**, che non sempre si possono prevedere!



Twitter academic reseach portal

Le politiche di accesso ai dati di Twitter stanno cambiando, c'è il rischio (concreto) che i dati diventino a pagamento per tutti!

Grazie

MAURO BRUNO| mbruno@istat.it

Istat | Istituto Nazionale di Statistica