Big Data: nuove fonti per l'analisi e le decisioni

Alessandra Righi

FEBAF, Roma, 12 giugno 2018

Istat

Scopo dell'intervento

1. Presentare i Big data, le loro tipologie e le caratteristiche che li rendono possibili fonti statistiche

2. Presentare i motivi per cui i BD possono **migliorare** le previsioni macroeconomiche e il nowcasting e aiutare **nell'individuazione** di *leading indicators* per la determinazione dei punti di svolta del ciclo economico

3. Fare una panoramica delle **applicazioni realizzate** a livello internazionale

4. Presentare alcune **sperimentazioni in corso** in Istat volte al miglioramento del nowcasting di grandezze macroeconomiche

Big Data: definizione, tipologie e caratteristiche

DATAFICATION

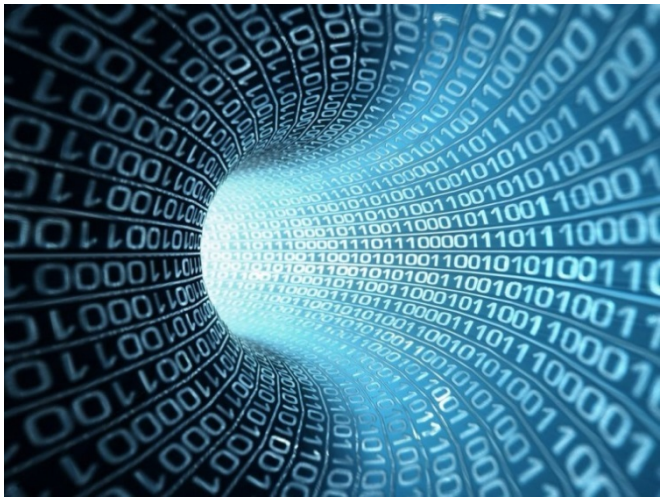
- Questo concetto viene introdotto da Cukier e Mayer-Schoenberger nel 2013 nell'articolo dal titolo «The Rise of Big Data» su Foreign Affairs
- Ogni attività che svolgiamo, online o in altro modo, finisce per essere registrato per un successivo uso nell’unità di archiviazione dati di qualcuno, ma anche in più unità di archiviazione o forse anche messo in vendita
- E’ quindi quel processo che "prende tutti gli aspetti della vita e li trasforma in dati"

Big Data: definizione, tipologie e caratteristiche

- Quando mettiamo un «Like» a qualcuno o qualcosa online o quando navighiamo sul Web siamo involontariamente tracciati
- Quando andiamo in giro anche le nostre azioni diventano dati in modo completamente involontario, tramite sensori, telecamere pagamenti elettronici
- Tutti lasciamo una traccia quando accendiamo il telefono, o quando prenotiamo un volo o un hotel tramite su un motore di ricerca, o quando chiediamo a Google come arrivare ad un certo edificio in autobus la mattina

- Considerando che tra il 2006 e il 2016 la quota di popolazione di 6 anni è oltre che dichiara di utilizzare regolarmente Internet è passata da 32% al 61% → la nostra sfida è riuscire a sfruttare questa cosiddetta datafication per produrre statistiche per le decisioni

Big Data: definizione, tipologie e caratteristiche



WIKIPEDIA

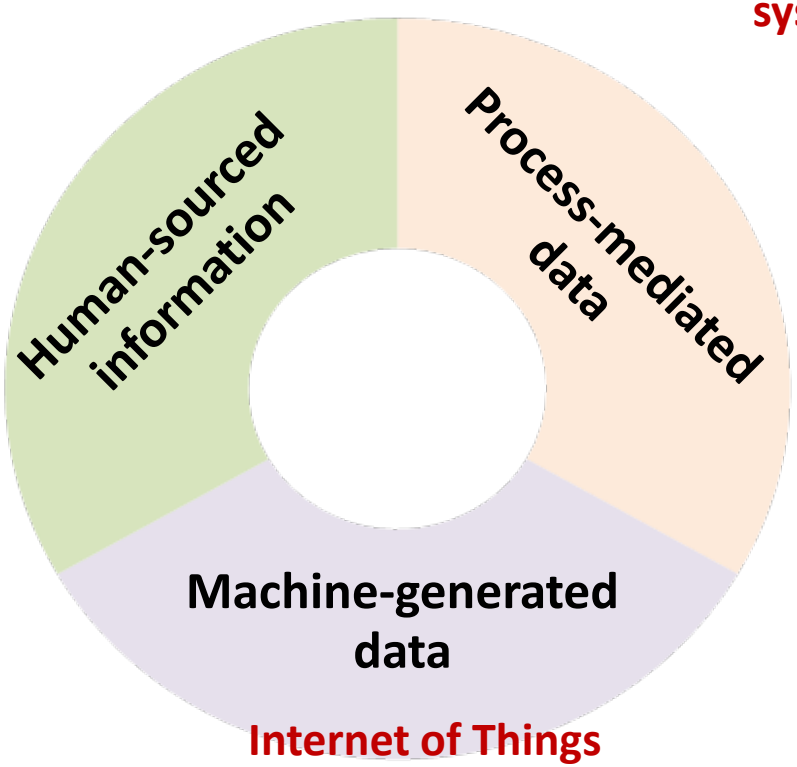
«Big data è il termine usato per descrivere una raccolta di dati così estesa in termini di **volume**, **velocità** e **varietà** da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore»

Big Data: definizione, tipologie e caratteristiche

Social Networks

- Social networks (Facebook, Twitter, LinkedIn, …)
- Blogs
- Videos (Youtube)
- Search engine queries
- E-mails
- …

Traditional Business systems



- Transazioni commerciali
- Quotazioni azioni
- Movimenti bancari
- Carte di credito
- E-commerce
- Medical records

- Dati da sensori: meteo e inquinamento, traffico (webcam), smart meters
- Tracking devices: dati di telefonia cellulare, GPS, immagini satellitari
- Dati da sistemi computerizzati

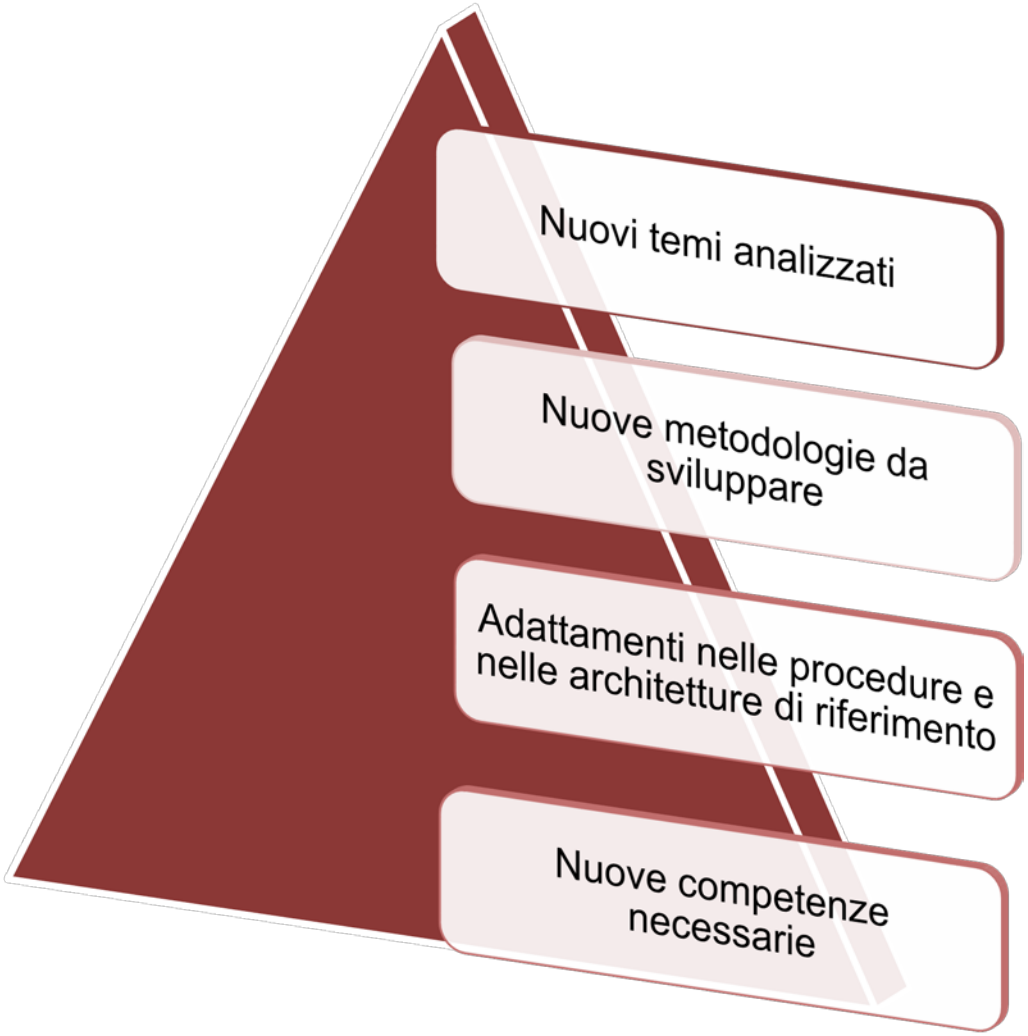
Big Data: definizione, tipologie e caratteristiche

- Il valore deriva dalla capacità di fornire:
  - la risposta più adeguata e flessibile alle esigenze degli utenti
  - una gamma più ampia di prodotti e servizi anche statistici (senza aumentare il carico)

- I Big data rappresentano ormai un'importante fonte di informazione che può essere utilizzata a fini statistici mediante lo sviluppo degli opportuni metodi ma occorre ancora:
  - comprendere meglio gli aspetti di qualità delle nuove fonti
  - considerare le sfide inerenti la preparazione, pulizia, filtraggio e valutazione dei nuovi dati
  - introdurre nel trattamento il concetto di *Privacy by design*, per tutelare i cittadini

Big Data: definizione, tipologie e caratteristiche

Implicazioni per la Statistica



Perché possono migliorare le previsioni macroeconomiche

- Oggi gli utilizzatori di dati richiedono non solo dati più tempestivi ma “conoscenze” basate su dati di qualità sempre migliore e in grado di orientare le decisioni

- Considerando la crescente complessità e velocità di cambiamento della società, i BD possono essere molto utili per l'analisi di relazioni complesse e per la produzione di informazione statistica quasi in real-time

- Anche le previsioni macroeconomiche e l’individuazione di indicatori anticipatori del ciclo economico sembrano beneficiare delle opportunità derivanti dall’utilizzo delle nuove fonti BD e/o dall’uso di queste insieme alle fonti più tradizionali

Perché possono migliorare le previsioni macroeconomiche

- Per Eurostat, il nowcasting di indicatori macroeconomici è un campo in cui i BD possono giocare un ruolo decisivo in futuro (vedi *Big Data and Macroeconomic Nowcasting: from data access to modelling, 2016*)

- I BD infatti:
  - ✓ forniscono informazioni complementari ai dati tradizionali
  - ✓ offrono una prospettiva più granulare sull'indicatore di interesse (sia nella dimensione temporale che in quella trasversale)
  - ✓ le informazioni sono disponibili molto tempestivamente
  - ✓ generalmente non sono soggette a revisioni

- L'uso di BD consente:
  - monitoraggio delle tendenze macroeconomiche (es. dati di Google per la previsione o analisi di *sentiment* dei consumatori)
  - monitoraggio della stabilità finanziaria (*sentiment* nei mercati finanziari, incertezza)
  - calcolo di indicatori di «allerta» precoce

Perché possono migliorare le previsioni macroeconomiche

Sono però necessari nuovi approcci e nuovi metodi per ottenere buoni risultati nel nowcasting utiilizzando BD

- Per riuscire ad estrarre in segnale occorrono Tecniche di filtraggio per dati ad alta frequenza (Signal Extraction to Uncertainty Indexes)
- Per ridurre la mole di dati nei modelli occorrono tecniche di Penalised Regressions (Ridge, Lasso) o modelli a fattori o Mixed-Frequency Models
- Per lo studio della serie possono essere necessarie tecniche di Bayesian VARs (Time Varying Parameter VAR o Stochastic Volatility VAR)
- E' inoltre necessario un cambio di paradigma da stime model based a stime algorithm based e uso di specifiche tecniche di machine learning (Regression Trees, Random Forests, Neural Networks / Deep Learning)

Le esperienze internazionali

Nella letteratura internazionale esistono due approcci per queste «nuove» previsioni:

1. stime real-time che fanno uso di informazioni provenienti da ogni nuova release di dati
2. nowcasting utilizzando serie derivanti da BD (Google trends, Twitter)

1.Stime real-time

- Giannone, Reichlin, Small (2008), Nowcasting: The real-time informational content of macroeconomic data, Journal of Monetary Economics, 55
- Higgins (2014), GDPNow: A Model for GDP "Nowcasting" Federal Reserve Bank Of Atlanta, Working Paper Series
- Carriero, Clark, Marcellino (2014), Real-Time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility

2.Studi su nowcasting che prendono le mosse dallo studio Choi e Varian, (2012) hanno riscontrato che i dati ottenuti mediante *query* di Google possono essere *leading indicators* per disoccupazione, fiducia dei consumatori, pianificazione di viaggi, vendite di auto

Le esperienze internazionali

Galbraith J.W. and G. Tkacz (2015), Nowcasting GDP with electronic payments data, European Central Bank (ECB). Statistics Paper Series No 10 / 2015	**Mostrano l'utilità di un ampio set di dati di pagamenti elettronici (comprese transazioni di carte di credito e debito e assegni) come indicatori del PIL (tasso di crescita trimestrale)** **Queste variabili catturano un'ampia gamma di spese e sono disponibili in modo molto tempestivo** **Mentre ogni transazione effettuata con questi meccanismi di pagamento è osservabile, i dati vengono aggregati per la previsione macroeconomica** **Tra le variabili di pagamento considerate, le transazioni con carta di debito sembrano produrre i maggiori miglioramenti nella precisione previsiva**

- Per una review generale

Hassani H., Sirimal Silva E. (2015) Forecasting with Big Data: A Review, Ann. Data. Sci. (2015) 2(1):5–19 DOI 10.1007/s40745-015-0029-9 **© Springer-Verlag Berlin Heidelberg**	**Si identificando i problemi, le potenzialità, le sfide e le relative applicazioni** **La rassegna rileva che al momento i settori dell'economia, dell'energia e della dinamica della popolazione sono i principali sfruttatori di BD per le previsioni, e i Factor models, i modelli bayesiani e le reti neurali sono gli strumenti più comuni adottati per la previsione con BD**

Le esperienze internazionali

Nowcasting con Google trends

Perché si utilizza Google? Secondo comscore.com il sito di Google è dominante nel mercato dei motori di ricerca (56% - 2004 e 64% - 2016)

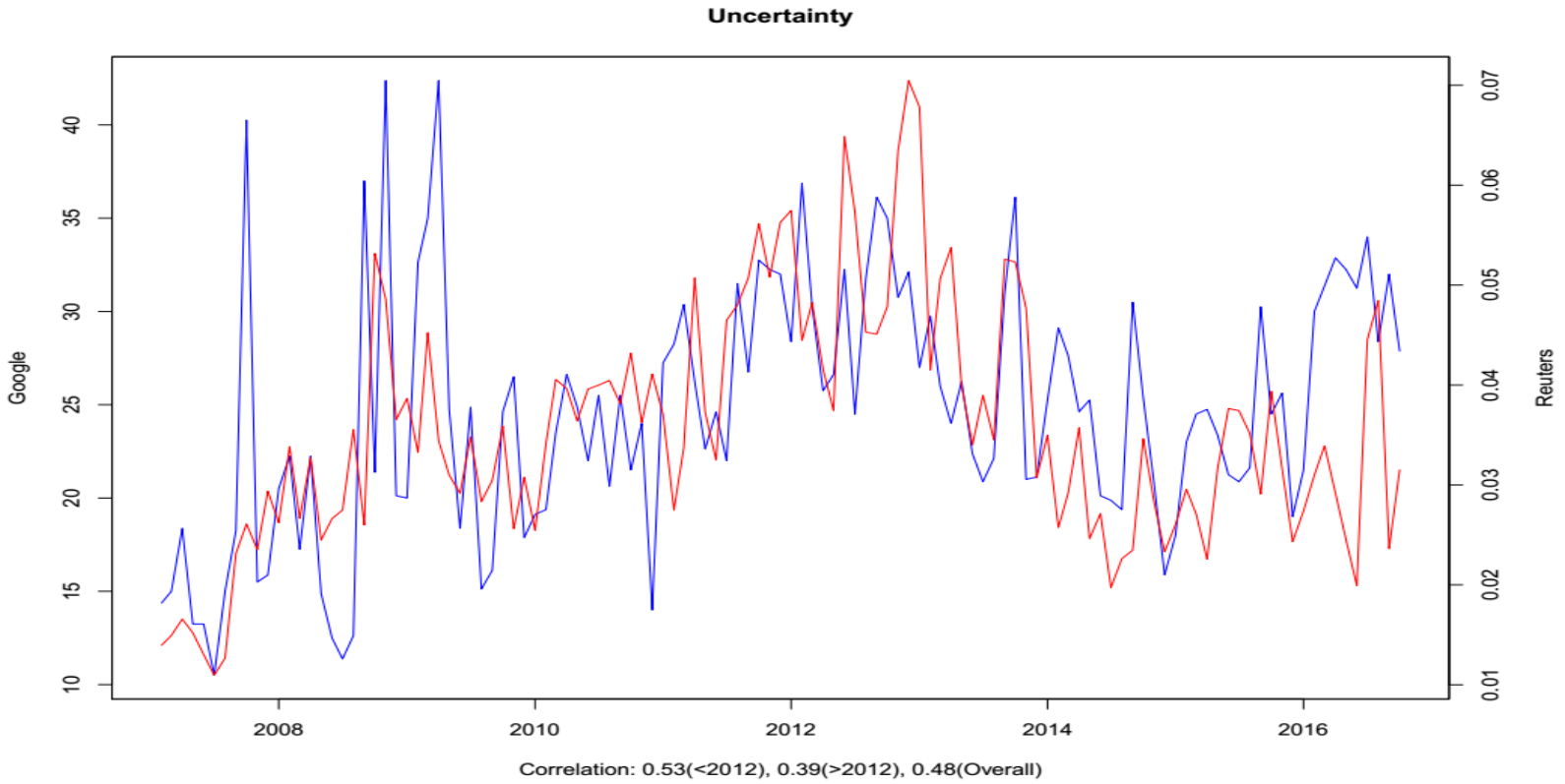
- Google trends mostra il numero di ricerche effettuate per una determinata parola chiave rispetto al totale delle ricerche ➡ i dati indicano la probabilità relativa che un utente ricerchi una determinata parola chiave in un certo momento
- Sono raccolti:
  - ✓ utilizzando le informazioni di indirizzo IP e aggiornate quotidianamente
  - ✓ se il numero di ricerche supera una determinata soglia di traffico
  - ✓ eliminando le *query* ripetute da un singolo utente
- Sono disponibili per paese, regione, città e vengono normalizzati (suddivisi per il traffico totale per area geografica)
- L'indice del volume di ricerca viene scalato dividendo ciascun punto dati dal massimo nella settimana o nella giornata (solo negli ultimi 90 giorni)

Nowcasting con Google trends

- Kapetanios, Marcellino, Papailas (2017) *Filtering techniques for big data and big data based uncertainty indexes*. Eurostat
- Naccarato , Falorsi, Loriga, Pierini (2018) Combining official and Google Trends data to forecast the Italian youth unemployment rate, *Technological Forecasting and Social Change*, 130:C
- Bortoli, Combes (2015), Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues, *Conjoncture in France*, INSEE
- Koop, Onorante (2013), Macroeconomic Nowcasting Using Google Probabilities, ECB
- D'Amuri, Marcucci (2012), The predictive power of Google searches in forecasting unemployment, Temi di discussione, No. 891, BI
- Vosen, Schmidt (2011), Forecasting private consumption: survey-based indicators vs. Google trends, *Journal of Forecasting*
- Askitas, Zimmermann (2009), Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55

Le esperienze internazionali

Kapetanios, Marcellino, Papailas (2017) *Filtering techniques for big data and big data based uncertainty indexes*. Eurostat

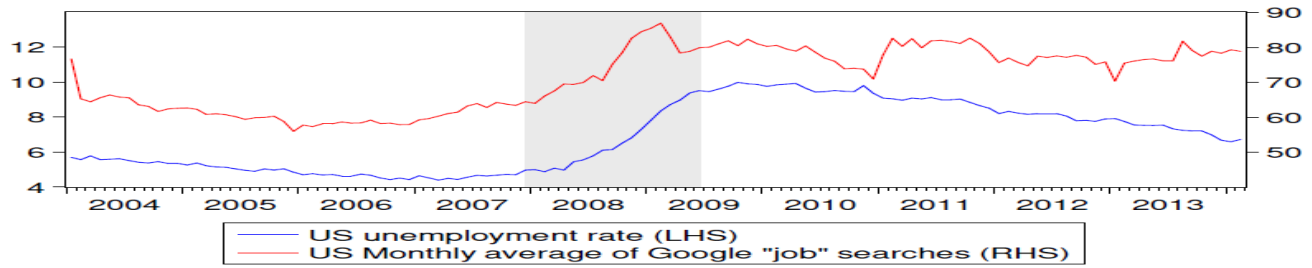


Comparing the General Uncertainty Index of Google (left axis, blue colour) to the corresponding Reuters index (right axis, red colour). The correlations before 2012, after 2012 and during the whole sample are mentioned below the figure.

Istat

Nowcasting con Google trends

D'Amuri, Marcucci (2012) osservano l'associazione tra tasso di disoccupazione USA e ricerche relative alla parola "Job" su Google Trends e suggeriscono l'uso di un indicatore aggiuntivo per prevedere il tasso di disoccupazione mensile, un indice Google (GI) basato sulle ricerche relative al lavoro ottenute da Google trends

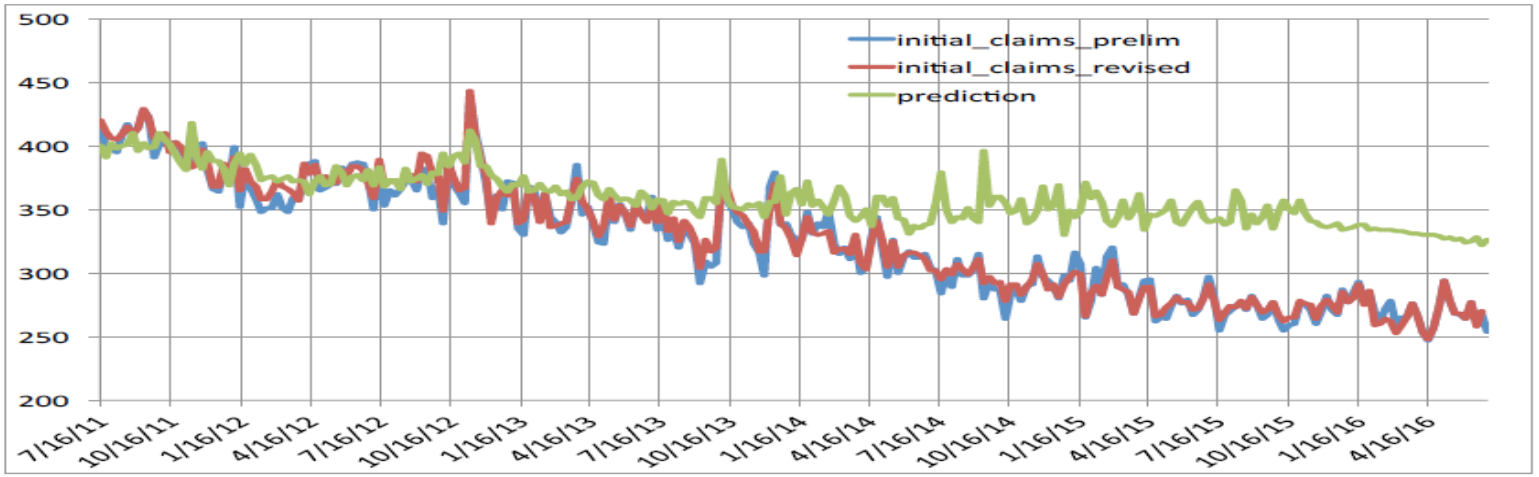


Poi confrontano il potere predittivo dei modelli di previsione lineari utilizzando il *Google index* o modelli che usano variabili più tradizionali (richieste di indennità disoccupazione, Aspettative dei consumatori, Aspettative occupazionali settoriali) e **scoprono che il nuovo indicatore migliora le previsioni più degli altri**

Nowcasting con Twitter

Antenucci et al. (2013):

- prevedono la Job loss (espressa in termini di US Initial Claims for Unemployent insurance) con Twitter
- I testi dei messaggi Twitter sono analizzati per individuare la stringa «lost my job»
- Utilizzano i dati per produrre un *real-time Social media index*



Uso di BD e dati tradizionali: sperimentazioni in Istat

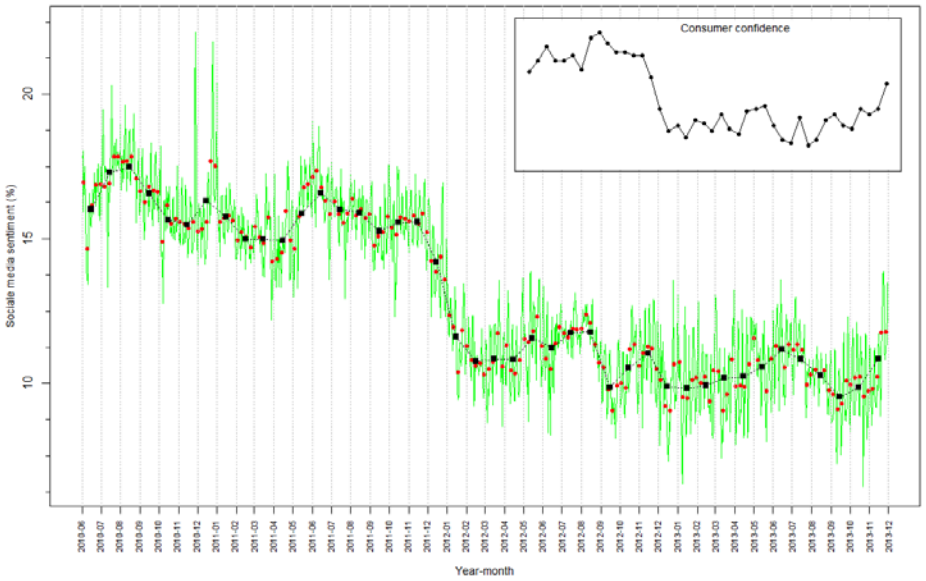
Anche l'Istat si muove su questa strada, aprendosi allo sviluppo di nuovi indicatori/informazioni da affiancare a quelli tradizionali derivanti da indagini, pur continuando ad assicurare la qualità delle statistiche ufficiali prodotte

Le sperimentazioni in corso per migliorare le stime macroecomiche:

- Produzione del Social Mood Index
- Uso in collaborazione con la Banca d'Italia di serie dei Pagamenti elettronici e da carte di credito del Sistema dei pagamenti e dei Report delle attività antiriciclaggio dell'UIF per il miglioramento delle stime flash (t+45) di indicatori macro

Uso di BD e dati tradizionali: sperimentazioni in Istat

Figura – Serie di fiducia consumatori da Social media (Facebook, Twitter, blogs,…) in Olanda
Andamento giornaliero (verde), settimanale (rosso) e mensile (nero)
Periodo Giugno 2010-Dicembre 2013



Nell'inserto c'è la serie mensile ufficiale del Consumer confidence CBS

Lo studio olandese (CBS) rivela:

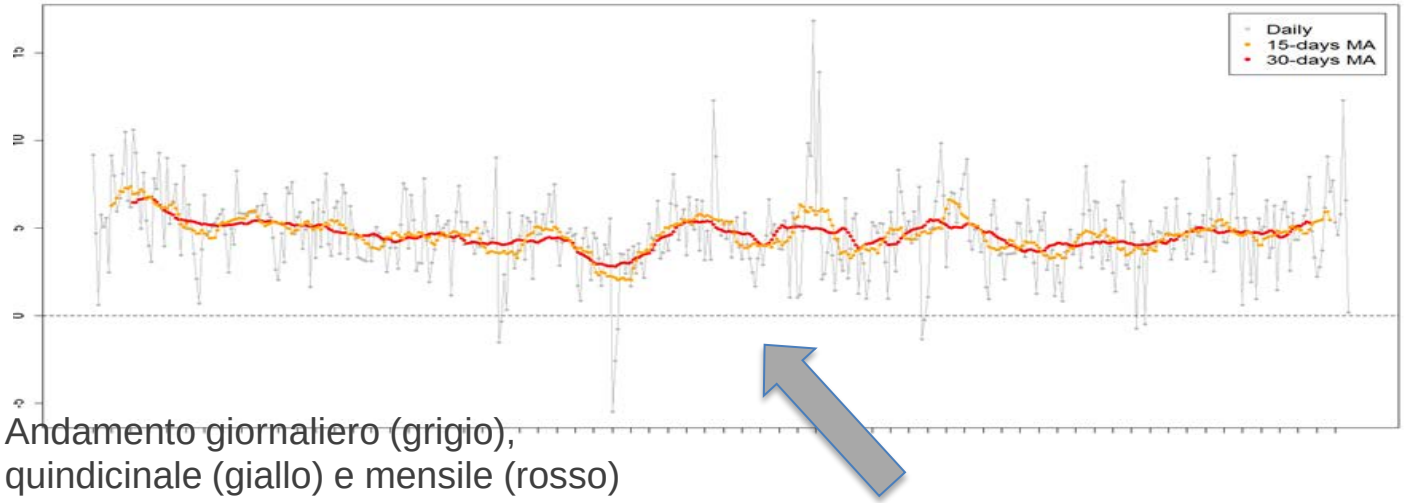
- una chiara associazione tra i cambiamenti nel Sentiment calcolato dai messaggi dei Social media e il Clima di fiducia dei consumatori ufficiale
- I cambiamenti nel Sentiment dei Social media precedono sempre quelli nella Consumer confidence con un lag di 7 giorni
- I messaggi che influenzano di più l'indice dei Social media sono quelli Facebook e poi quelli Twitter

Sentiment Index da Twitter

Produzione della serie di un indicatore di sentiment da Twitter ad alta frequenza su temi relativi alla fiducia

Social Mood Index

- basato su 40.000 Tweet al giorno
- Da gen.2016
- In diffusione nella nuova collana delle *Statistiche sperimentali*



Andamento giornaliero (grigio), quindicinale (giallo) e mensile (rosso)

Istat

STATISTICHE SPERIMENTALI

Le esigenze conoscitive degli utenti dell'informazione statistica si ampliano e si approfondiscono in un processo continuo, per cui l'Istat è chiamato, da un lato, a migliorare la propria capacità di innovare, dall'altro lato, a fornire risposte sempre più tempestive.

La produzione di statistiche di qualità ha però bisogno di tempo: quello necessario alla sperimentazione di nuove metodologie, alla loro traduzione in soluzioni tecnologiche e organizzative, all'accertamento del rispetto dei requisiti di qualità e delle regole di armonizzazione.

Per contemperare queste esigenze – in linea con il percorso intrapreso da Eurostat e da altri istituti di statistica – l'Istat sperimenta l'utilizzo di nuove fonti e l'applicazione di metodi innovativi nella produzione di dati. E offre i risultati delle sperimentazioni alla fruizione e alla valutazione degli utenti.

Si tratta di "statistiche sperimentali" e non di "statistiche ufficiali". Ma il loro potenziale è elevatissimo. Perché colmano lacune conoscitive producendo informazioni rilevanti in maniera tempestiva; perché fungono da volano per nuove analisi e nuovi indicatori; perché garantiscono un valido sostegno conoscitivo alle policy.

Per agevolarne il reperimento e la fruizione, le statistiche sperimentali prodotte dall'Istat sono organizzate, oltre che in ordine cronologico, in quattro differenti tipologie.

STATISTICHE SPERIMENTALI

- CLASSIFICAZIONI NON STANDARD
- NUOVI INDICATORI
- ANALISI E QUADRI INTERPRETATIVI
- SPERIMENTAZIONI SU BIG DATA

Conclusioni

- Nuove fonti di dati (strutturati e non strutturati) possono essere un ottimo complemento alle statistiche ufficiali
  - ➢ proxy di aspettative e sentiment
  - ➢ nuovi indicatori congiunturali
  - ➢ Evidenze incoraggianti anche per nowcasting e stime flash

- Sfruttare queste nuove fonti di dati tanto tempestive può portare beneficio sia alle analisi sia alle policy

- C’è bisogno di nuove metodologie che superino i limiti legati alla natura variabile e non rappresentativa dei dati

- Occorrono nuovi skills