L’utilizzo dei Big Data in Istat: stato attuale e prospettive

Giulio Barcaroli

FORUM PA
28 maggio 2015

Outline

Illustrazione delle attività portate avanti dall'Istat, riguardanti:

- la prosecuzione di sperimentazioni già avviate dell'uso di fonti di Big Data a fini statistici;
- l'avvio di nuove sperimentazioni con altre fonti Big Data;
- la predisposizione di un laboratorio informatico interno, per "tuning" e analisi di applicazioni, e l'avvio di test di utilizzo di data center esterni all'Istat per elaborazioni su scala molto ampia;
- l'investimento in formazione su nuovi skill (riconducibili al filone della *data science*);
- la gestione delle problematiche connesse al trattamento ed alla privacy dei dati di fonte Big;
- le relazioni con i provider di Big Data.

Attività in cooperazione

Tutte le attività di seguito descritte sono portate avanti anche all'interno di collaborazioni

- internazionali, attraverso progetti con
  - EUROSTAT (Task Force on big Data)
  - le Nazioni Unite (UNECE High Level Group on Modernisation)

- nazionali, mediante convenzioni attivate con

  - CNR e Università di Pisa
  - Università Sapienza di Roma
  - CINECA

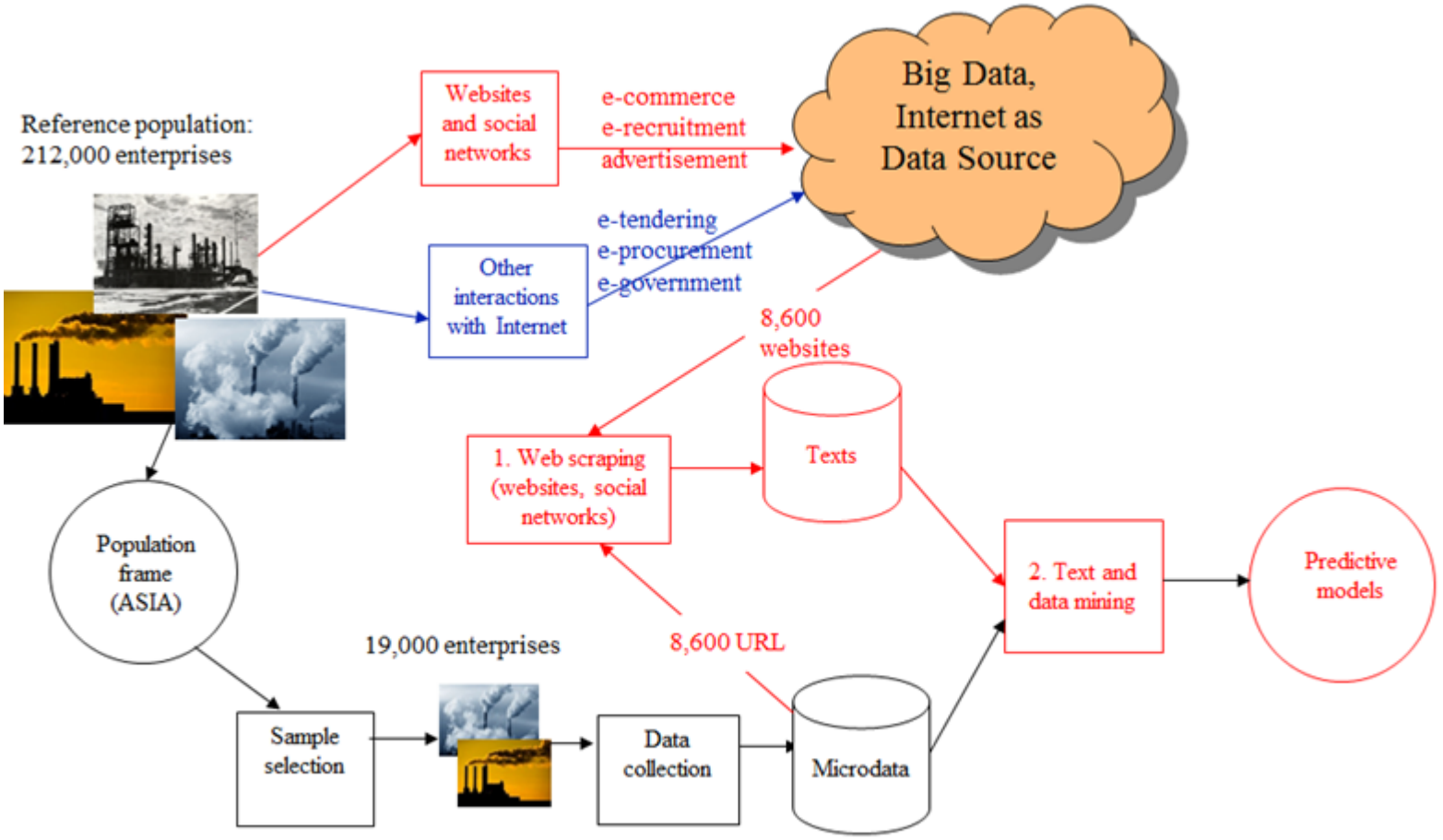
Prosecuzione di sperimentazioni già avviate

1. Uso di Internet come fonte di dati: *rilevazione sull'utilizzo delle tecnologie ICT da parte di imprese e pubbliche amministrazioni*; Indice dei *Prezzi al Consumo*

2. Uso delle interrogazioni di Internet come informazioni ausiliarie (Google trends) per previsioni e proiezioni (Indicatori del *Mercato del Lavoro*)

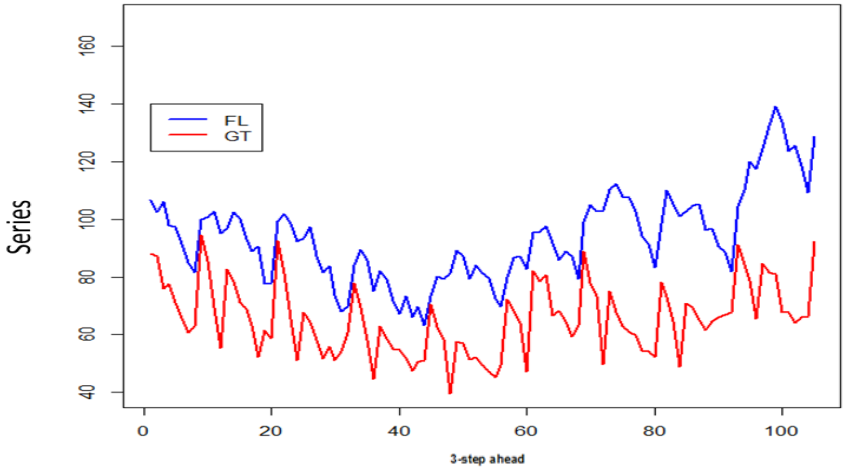
3. Uso dei dati di telefonia mobile per stimare i flussi di popolazione inter-comunale e da questi derivare le varie componenti della popolazione insistente sui Comuni italiani (Progetto *Persons and Places*)

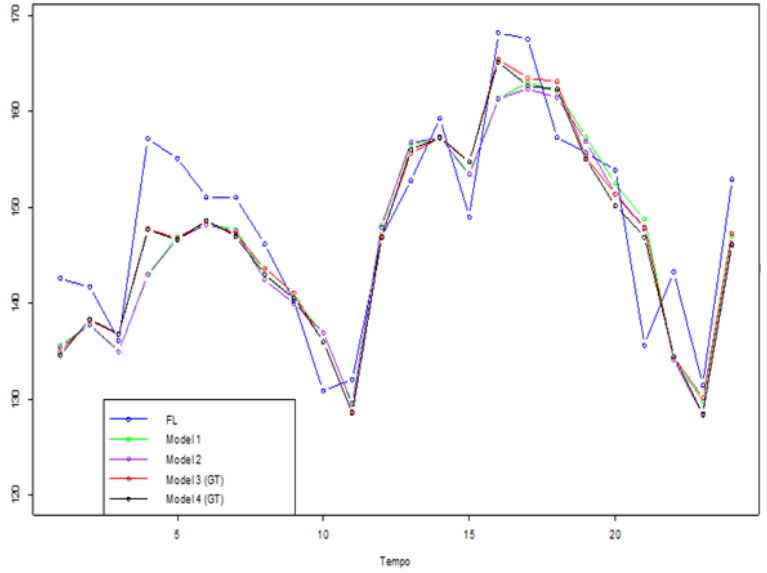
Uso di Internet come fonte di dati: indagine uso ICT da parte delle imprese



Uso delle interrogazioni di Internet (Google Trends)

Lo scopo di questo esperimento è quello di valutare la possibilità di utilizzare, come informazioni ausiliarie, le serie storiche delle interrogazioni relative a termini come "lavoro", "offerte di lavoro" e simili, ottenute da Google Trends (GT), al fine per produrre (i) stime anticipate e di previsione (*nowcasting* e *forecasting*) del di tasso mensile di disoccupazione le previsioni, e/o (ii) stime per piccole aree dello stesso indicatore

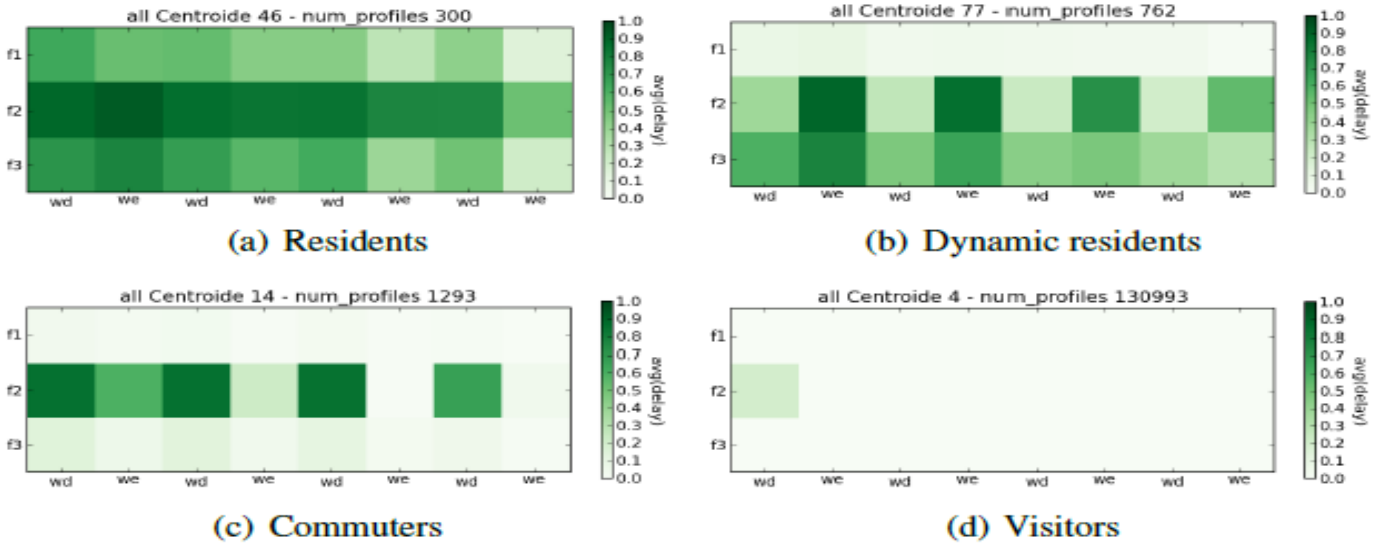




Uso dei dati di telefonia mobile

Utilizzo dei dati prodotti da chiamate GSM per la produzione di

1. *matrici origine/destinazione* della mobilità giornaliera per motivi di lavoro e studio a livello di Comune;
2. produzione di *stime* per ogni Comune relative a persone
   - ✓ residenti stabilmente;
   - ✓ residenti dinamici;
   - ✓ pendolari;
   - ✓ visitatori occasionali.



(a) Residents (b) Dynamic residents

(c) Commuters (d) Visitors

Nuove sperimentazioni

Verranno esplorate le seguenti possibilità:

- utilizzo dei dati da social network per produrre stime confrontabili con quelle ottenute dall'indagine sul grado di fiducia dei consumatori;

- utilizzo dei dati da sensori (immagini) per produrre:
  - stime dei flussi di traffico dei veicoli (webcam e telecamere sulla rete stradale e autostradale);
  - stime della produzione agricola (immagini satellitari);

- utilizzo degli scanner data nella grande distribuzione per l'Indice dei Prezzi al Consumo

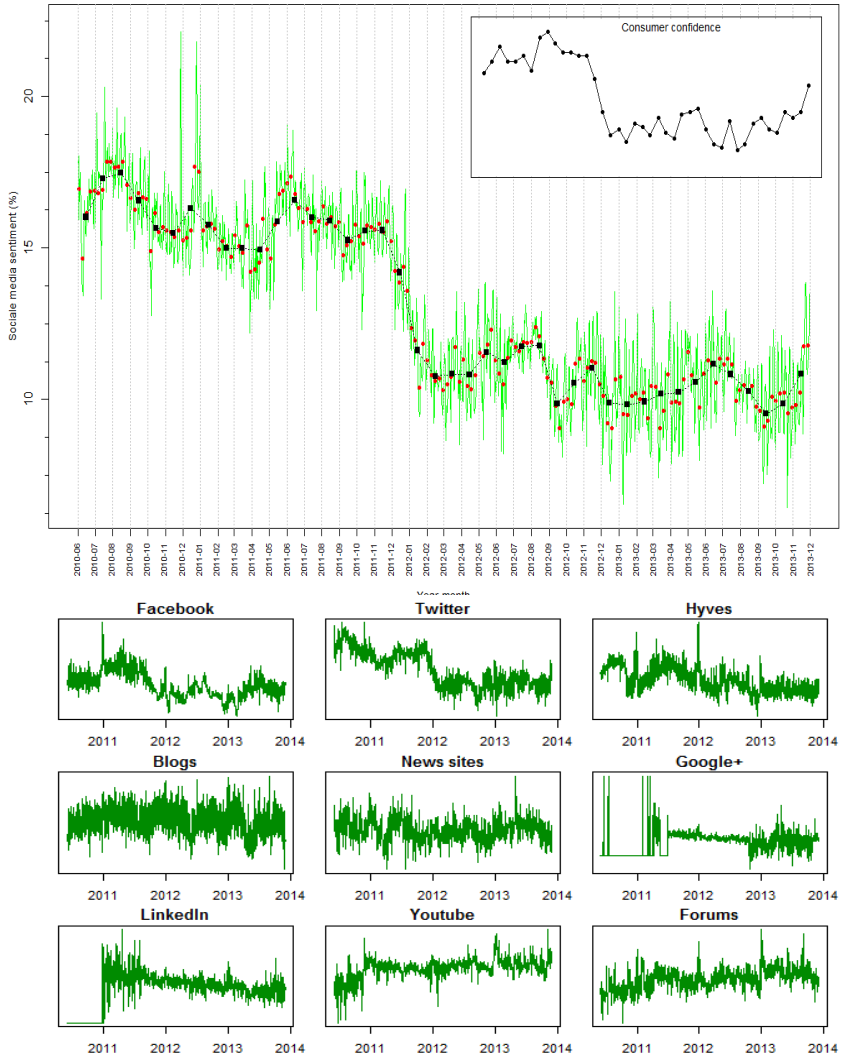
- utilizzo dei dati di telefonia mobile per la stima dei flussi turistici.

Uso dei dati da social network (Fiducia dei consumatori)

Ci si propone di utilizzare i dati da social network (Facebook, Twitter, etc.) per produrre stime anticipate relative al grado di fiducia dei consumatori, sulla base dell'esperienza già effettuata da Statistics Netherlands.

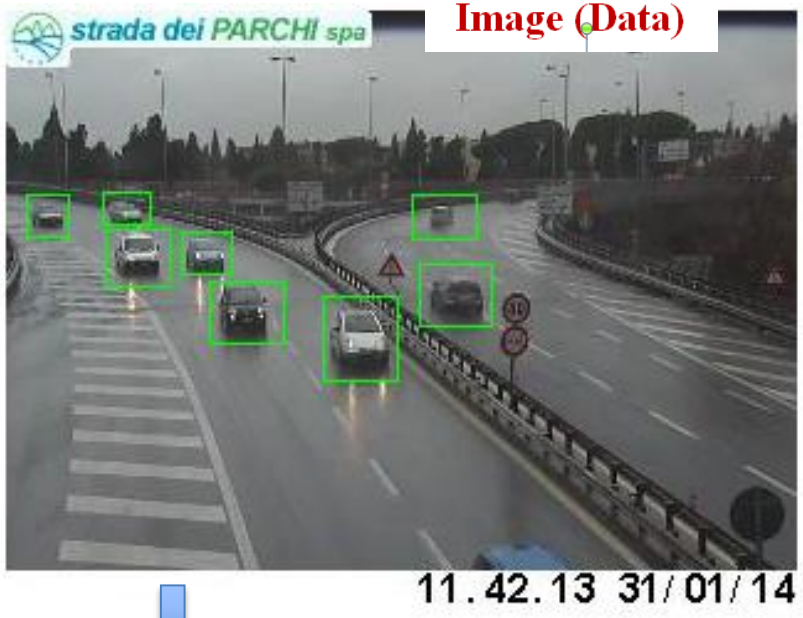
Si tratta di raccogliere centinaia di milioni di *post* e *tweet* ed elaborarli in modo da poterli organizzare in serie storica confrontabile con quella dell'indagine condotta mensilmente dall'Istat.

Anche in questo caso l'obiettivo potrebbe essere quello di fornire stime anticipate.



Uso dei dati da sensori (immagini): statistiche sul traffico

L'idea è quella di sfruttare la disponibilità di immagini da telecamere e webcam lungo la rete stradale e autostradale per derivarne dati utilizzabili al fine di produrre stime sul traffico.
Il primo step consiste nel reperire le immagini e da queste riconoscerne gli oggetti di interesse (i veicoli), da conteggiare nell'unità di tempo.
Il secondo step consiste nel definire una procedura di stima complessiva basata sull'uso dei dati così prodotti.



Information (Metadata)

Road: A24
Km: 6,9
Date: 31/01/2014
Time: 11:42:13

Value (Data)

Vehicles: 8

Laboratorio informatico interno e data center esterni

Dal punto di vista di volume, velocità e varietà dei dati che caratterizzano questa fonte, è importante effettuare sperimentazioni utilizzando piattaforme adeguate.

L'Istat si sta muovendo su due piani:

- quello interno, dotandosi di una piattaforma elaborativa dedicata, che permetta di condurre sperimentazioni mirate ad aspetti più qualitativi;
- con una collaborazione esterna (CINECA) per l'utilizzo di un supercomputer per l'elaborazione di grandi moli di dati (PICO).

"PICO: the Cineca solutions for Big Data Science"

05/11/2014

Many disciplines today share a common problem: the storage and management of large amounts of data.

Cineca has recently initiated a new programme for the promotion of Big Data Science, addressing the challenges that arise when either the data volume, structure or the speed of collection, make processing difficult.

Formazione sui nuovi skill richiesti: la «data science»

L'obiettivo è quello di dotare l'Istituto di skill che leghino elementi propri della formazione *statistica classica*, con quelli della *computer science*, finora non comunicanti.
Questa è una tendenza peraltro generale, rientrante nel campo della *data science*.

Un programma di formazione di massima potrebbe riguardare:

- tecniche di machine learning: clustering, classification, pattern discovery…
- tecniche di campionamento nel contesto dei Big data
- tecniche avanzate di accesso a database NoSQL e multidimensionali (Business Intelligence e Visual Analytics)
- applicazioni distribuite: MapReduce/Hadoop/RHadoop
- applicazioni di data mining e web scraping

Aspetti legati alla privacy nel trattamento dei Big Data

L'Istat ha avviato un confronto col Garante per la Privacy al fine di definire un quadro preciso relativo alle migliori modalità di acquisizione e trattamento dei dati da questa nuova fonte.

L'obiettivo è quello di pervenire ad una normativa il più possibile vicina a quella che riguarda i dati amministrativi, e cioè:

- piena disponibilità di tali dati,
- totale rispetto del vincolo di riservatezza.

La normativa attuale è alquanto restrittiva. Infatti, il vincolo della comunicazione preventiva ai soggetti proprietari dei dati relativamente a loro utilizzo a fini statistici rischia di escludere fonti importanti (quali ad esempio quelli della telefonia mobile).

Le misure da adottare potrebbero consistere in:

- una completa anonimizzazione dei dati, *oppure*
- una comunicazione fornita in modo non individuale.

Rapporto con i provider dei dati

L'Istat ha già avviato contatti con due tra i maggiori provider di telefonia mobile (Telecom e Wind) al fine di individuare le migliori modalità di fornitura e/o utilizzo dei dati di telefonia mobile.

Problemi relativi a:

- privacy;
- costi.

Due sono le forme al momento individuate:

1. fornitura di microdati a Istat da parte dei provider, e successive elaborazioni;

2. sviluppo in Istat di applicativi testati su insiemi limitati di dati, e fornitura degli applicativi ai provider per la loro esecuzione effettiva sull'intero insieme di dati di interesse (superamento di molti dei problemi di privacy).

Conclusioni

L’Istat, alla pari di molti altri Istituti di statistica, continuerà a investire nella esplorazione della possibilità di utilizzare a fini statistici le fonti riconducibili ai Big Data.

L’obiettivo a breve termine (18 mesi) è quello di inserire in produzione alcuni processi facenti uso di tali fonti, non in modo sostitutivo rispetto a fonti tradizionali, ma combinandone l’utilizzo.

L’obiettivo a medio-lungo termine (2020) è quello di implementare processi di produzione basati sull’utilizzo unico o prevalente di fonti Big Data, non solo per replicare informazione statistica già prodotta, ma per produrne di nuova.