8 MAGGIO 2020

Le distribuzioni dei siti web delle imprese secondo i servizi e le funzionalità offerte: il processo di stima supportato dall'uso di dati da Internet

Dal 2017 l'Istituto Nazionale di Statica Italiano fornisce statistiche sperimentali utilizzando dati Internet per riprodurre alcune stime attualmente calcolate dall'indagine della Comunità Europea sull'uso delle ICT e del commercio elettronico nelle imprese (indagine ICT).

Questa nota metodologica elabora il *framework* statistico che comprende l'uso di dati da Internet (Big Data). Questo suggerisce di rimodellare gli strumenti statistici noti e sfruttati in contesti innovativi, di combinare informazioni provenienti da più fonti di dati imperfette (indagini campionarie, fonti amministrative e Big Data), per modellare la propensione alla selettività dei Big Data, il meccanismo di non risposta dell'indagine e la distribuzione delle principali variabili di interesse e specifiche distribuzioni marginali. La nota copre tutte le fasi del processo evidenziando la necessità di disporre di un quadro statistico integrato in cui: a) le fonti di dati debbano rappresentare diverse componenti di un sistema informativo unico; b) le tecniche di *data mining* per l'elaborazione dei Big Data (ad esempio, le tecniche di *natural language processing*, le tecniche di *machine learning*, ecc.) devono essere pianificate in modo coerente; c) i metodi di analisi dei dati implementati in diverse fasi (ad es. tecniche di *machine learning* e stimatori) devono definire un *toolbox* completo; d) gli stimatori di parametri di diversa natura devono definire un sistema di statistiche coerenti.

Nelle statistiche sperimentali 2018 sulle ICT il processo di stima utilizza le informazioni raccolte direttamente dai siti web delle imprese per calcolare le distribuzioni dei siti che:

1- offrono funzioni di *web-ordering* (componente e-commerce);

2- presentano proposte di lavoro o informazioni sui posti di lavoro vacanti nell'impresa;

3- hanno collegamenti ai social media (Facebook, Twitter, Instagram ecc.);

4- mostrano una combinazione di alcune funzioni e / o servizi sul sito web.

La procedura di stima complessiva è suddivisa in quattro fasi principali secondo lo schema mostrato in Figura 1. I dettagli sul processo sono riportati in Righi *et al.* (2020) e Bianchi *et al.* (2019a, 2019b).

Figura 1. Le fasi principali per stimare le distribuzioni obiettivo che usano informazioni dai siti web.

1 – Web address acquisition	URL from the admin sources
	URL from thematic directory sites
	URL from batch queries on search engines (**URL Retrieval** techniques in case of non existing URL)
2 – Enterprise identification	URL validation, check URL's validity (recurring errors and domain extraction)
	Detection of identification variables from the website and **comparison** with the same information available in the SBR register
3 – Data analytics	***Web Scraping*** techniques for web data acquisition
	Text Mining techniques for extracting the requested information
	Machine Learning techniques for the use of algorithms that simulate a learning process for the construction of predictive models
4 – Inference	**From the enterprises with scraped websites to the enterprises of the target population**

SBS: Registro Statistico delle Imprese

L'indagine ICT

L'indagine ICT fa parte delle statistiche della Comunità europea sulla società dell'informazione e rappresenta una delle principali fonti annuali di dati per il quadro di valutazione dell'Agenda Digitale europea contribuendo a determinare l'indice composito di economia e società digitale (DESI) utilizzato per sintetizzare i progressi dell'economia digitale europea.

I principali temi rilevati dall'indagine sono relativi alle connessioni Internet, all'uso di Internet (sito Web, social media, cloud computing), all'integrazione elettronica dei processi aziendali (ad es. l'uso di software per interagire e condividere informazioni commerciali internamente come ERP, CRM o esternamente con altre imprese della catena del valore), all'eCommerce (vendite elettroniche via web, app, piattaforme digitali, mercati elettronici e intercambio di dati tra sistemi informativi), fatturazione elettronica e alle tipologie di investimenti ICT più innovativi (Robotica, Internet delle cose, Intelligenza artificiale , Analisi dei Big Data).

La popolazione target dell'indagine sulle TIC si riferisce alle imprese con 10 e più addetti che lavorano nell'industria e nei servizi non finanziari. La popolazione di riferimento viene individuata attraverso il Registro delle imprese attive (Asia) aggiornato a 2 anni prima del periodo di riferimento dell'indagine. Per il 2018 questa popolazione è di 199.416 imprese. Il disegno di campionamento è il seguente: i) un censimento per le imprese con 250 e più addetti (3.342 imprese); ii) un campione casuale semplice stratificato per le piccole e medie imprese (10-249 addetti). Le variabili di stratificazione sono definite in termini di 4 classi di addetti, 27 attività economiche (raggruppamenti di Ateco) e 21 regioni amministrative.

La dimensione del campione è di 33.059 imprese; le stime dell'indagine si basano sulle risposte validate di 22.079 imprese (tasso di risposta del 66,8%).

I parametri obiettivo

I parametri obiettivo sono tassi e distribuzioni delle imprese che offrono servizi e funzionalità nei loro siti web. Le statistiche sperimentali considerano 6 parametri diversi:

1- il tasso di imprese che offrono funzionalità di *web-ordering* nel sito web;

2- il tasso di imprese che offrono *annunci di lavoro* nel sito web;

3- il tasso di imprese con *collegamenti ai social media* nel sito web;

4- la distribuzione delle imprese per *website maturity* (valori 0,1,2) in cui la variabile è definita secondo la seguente regola:

a. La maturità del sito web tiene conto della presenza di 4 servizi (SERVIZI):
   i. il sito web dell'impresa ha ordini, prenotazioni o prenotazioni online (funzione di *web-ordering*);
   ii. il tracciamento o lo stato degli ordini effettuati;
   iii. la possibilità per i visitatori di personalizzare o progettare beni o servizi online;
   iv. la possibilità di personalizzare il contenuto del sito web per visitatori regolari/ricorrenti;

a cui si aggiunge

   v. l'utilizzo della pubblicità a pagamento su Internet (ADS) da parte dell'impresa.

b. I valori della variabile sono:
   i. 0 - se SERVIZI <2 e ADS = 0;
   ii. 1 - se SERVIZI <2 e ADS = 1 o SERVIZI> = 2 e ADS = 0;
   iii. 2 - se SERVIZI> = 2 e ADS = 1;

5- la distribuzione delle imprese per *sofisticazione del sito Web* (valori 0,1,2,3,4) in cui la variabile è definita in base alla presenza assenza dei 4 servizi del sito web sopra elencati (SERVIZI); per il valore della variabile viene attribuito un punto per la presenza di ciascuno di 4 servizi;

6- la distribuzione delle imprese per la variabile indicata come *WebF3* (valori 0,1) dove la variabile è definita secondo la seguente condizione che, se vera, attribuisce un punto alla variabile:

a. il sito web dell'impresa ha ordini, prenotazioni o prenotazioni online (funzione di *web-ordering*);

e almeno una tra le 4 funzionalità del sito web seguenti:

   b. l'accesso alla descrizione di prodotti o servizi, listini prezzi, il tracciamento o lo stato degli

ordini effettuati, la possibilità per i visitatori di personalizzare o progettare beni o servizi online, la possibilità di personalizzare il contenuto del sito web per visitatori regolari / ricorrenti.

Esistono diversi tipi di domini di interesse: il livello nazionale, per classe di dimensione del numero di occupati, per macro settori economici e per classe di dimensione, per Ateco (26 gruppi), per regioni, per Ateco a 2 cifre (Divisioni).

La procedura di stima

Fase 1-2-3 Classificazione automatica dei siti web delle imprese attraverso l'utilizzo di tecniche di web scraping, text mining e machine learning

L'obiettivo perseguito in queste tre fasi concerne la classificazione dei siti web, utilizzando le informazioni in essi contenute ed è ricondotto al noto problema della classificazione dei documenti di testo, per cui sono disponibili diverse metodologie nel campo del *text mining*. L'approccio proposto si basa principalmente su una rappresentazione semplificata dei siti web, attraverso la generazione automatica di record di dati standardizzati, che ne sintetizzano il contenuto. Questo approccio, che sostituisce le tradizionali tecniche di raccolta dati, è realizzato attraverso tecniche di *web* scraping, combinate con tecniche di *natural language processing* e di *machine learning*.

La procedura complessiva di classificazione dei siti web è delineata nei seguenti passi (vedi anche Bianchi *et al.*, 2018).

Innanzitutto, la procedura identifica ciascuna impresa sul web e crea un elenco di indirizzi di siti web (circa 118.000 *URL*). Alcuni *URL* sono disponibili da indagine o da fonti amministrative, altri sono stati recuperati attraverso una procedura di *URL retrieval* (Summa, 2017), che utilizza i motori di ricerca e le informazioni anagrafiche contenute nell'Archivio statistico delle imprese attive (ASIA).

Gli indirizzi web disponibili vengono controllati attraverso: l'analisi sintattica degli *URL*, il controllo degli errori ricorrenti, il controllo dell'*authority*, l'identificazione dell'*URL* esatto.

Dato un elenco di indirizzi di siti web, per ciascuno di essi si estrae il testo delle pagine attraverso una procedura automatica di *scraping* (Summa, 2017; Scalfati *et al.,* 2017). Oltre al testo che appare sulle pagine web, vengono acquisite altre informazioni, tra cui: gli attributi dei *tag HTML*, i nomi dei file, *le meta-keywords* delle pagine.

Gli algoritmi di classificazione non possono lavorare direttamente sul testo nella sua forma originale. Per questo è necessario eseguire una fase di *pre-processing*, in cui i documenti grezzi vengono convertiti in una forma semplificata. Per svolgere questo compito, il testo estratto (circa 94.000 siti web) viene elaborato con tecniche di *natural language processing*, al fine di identificare un dizionario di termini (*n-grams*) utile per descrivere il contenuto dei siti web.

Quando tutti i siti web sono rappresentati attraverso una serie di record di dati standardizzati (*feature vectors*), si può ricorrere alle tecniche di classificazione per classificarli in base all'aspetto di interesse.

In particolare, viene adottato un approccio di *machine learning* per prevedere il valore (presenza / assenza) delle variabili target *web ordering* e *annunci di lavoro*. Si tratta di un apprendimento supervisionato, basato sulla disponibilità di un set di record etichettati (*training set*), che costituiscono la fonte di informazioni per addestrare un classificatore. Pertanto, è necessario disporre di una serie di siti web, per i quali siano note le etichette di classe rispetto alla classificazione in esame.

Per questo scopo, si considera, come insieme di addestramento, il sottoinsieme di unità rilevate dall'indagine ICT 2018, per il quale è disponibile il testo estratto dai siti web nello stesso periodo della rilevazione (circa 12.000 nel 2018). La messa a punto dei modelli di predizione viene realizzata utilizzando sia le risposte fornite all'indagine, sia i testi acquisiti sui siti web. In realtà, nel caso di *annunci di lavoro* si considerano i dati dell'indagine ICT 2017 e i dati raccolti (circa 12.000), nello stesso periodo, dai siti web delle imprese incluse nel campione, perché l'indagine campionaria del 2018 non ha raccolto questa variabile.

A questo punto, i modelli di predizione, ottenuti durante la precedente fase di addestramento, vengono applicati alla generalità delle imprese, per predire i valori delle variabili *target* (*web ordering* e *annunci di lavoro*), relative a tutte le imprese per le quali sono stati realizzati con successo l'identificazione e lo *scraping* dei siti web (circa 94.000).

Al contrario, nel caso dei *social media* è possibile raccogliere esattamente il valore della variabile *target*, utilizzando le tecniche di *information retrieval* per acquisire i *link* ai *social media* direttamente dal sito web.

Phase 4 Stimatori delle distribuzioni delle variabili obiettivo

Lo stimatore del tasso di imprese che offrono funzionalità di web ordering nel sito web.

La variabile non è direttamente osservabile dai siti web. Tramite tecniche di *text mining* e *machine learning* (fase 3), vengono estratte informazioni utili per rappresentare i siti web e classificarli rispetto alla variabile *target*. L'ottimizzazione del classificatore utilizza i dati dell'indagine 2018 e i dati raccolti nello stesso periodo dai siti web delle unità campionate. Si applica il classificatore su tutti i dati dei sito web del 2018 per ottenere le previsioni.

Partendo da queste previsioni, utilizziamo lo stimatore *projection* (Kim e Rao, 2012; Breidt e Opsomer, 2017) che aggiorna lo stimatore applicato per calcolare le statistiche sperimentali del 2017.

Rispetto allo stimatore del 2017, la versione 2018 tiene conto del fattore di correzione basato sulla differenza tra le classificazioni (o previsioni) e i valori osservati sulle unità campionate. Data la somma delle classificazioni (o previsioni) per ciascun dominio di interesse, lo stimatore sottrae il valore del fattore di correzione. Infine, si utilizza uno stimatore di pseudo-calibrazione per riportare il valore del risultato all'intero universo delle imprese con il sito web. Si utilizzano i dati del dell'indagine ICT per stimare il numero di imprese con il sito web per ciascun dominio di interesse, poiché le fasi 1 e 2 non

sono in grado di identificare tutti i siti web delle imprese della popolazione di riferimento.

Lo stimatore del tasso di imprese che offrono annunci di lavoro nel sito web.

L'indagine campionaria del 2018 non ha rilevato questa variabile. Tramite tecniche di *text mining* e *machine learning* (fase 3), vengono estratte informazioni utili per rappresentare i siti web e classificarli rispetto alla variabile *target*. L'ottimizzazione del classificatore utilizza i dati dell'indagine 2017 e i dati raccolti nello stesso periodo dai siti web delle unità campionate. Si applica il classificatore su tutti i dati dei siti web del 2018 per ottenere le previsioni.

Partendo da queste previsioni, utilizziamo lo stimatore *projection* (Kim e Rao, 2012; Breidt e Opsomer, 2017) che aggiorna lo stimatore applicato per calcolare le statistiche sperimentali del 2017.

Data la somma delle classificazioni (o previsioni) per ciascun dominio di interesse, si utilizza uno stimatore di pseudo-calibrazione per riportare il valore del risultato all'intero universo delle imprese con il sito web. Utilizziamo i dati dell'indagine ICT per stimare il numero di imprese con il sito web in ciascun dominio di interesse, poiché le fasi 1 e 2 non sono in grado di identificare tutti i siti web delle imprese della popolazione di riferimento.

A differenza del caso della variabile *web-ordering*, non si utilizza un fattore di correzione nello stimatore *projection*.

Lo stimatore *projection* che utilizza i dati da Internet consente di produrre stime su base annuale. L'indagine rileva la variabile ogni 4 anni e produce le stime su base quadriennale.

Lo stimatore del tasso di imprese con collegamenti ai social media nei loro siti web.

Il processo di stima inizia dall'acquisizione diretta della variabile mediante tecniche di *information retrivial*. Data la somma dei valori raccolti (numero di siti con collegamenti ai social network) per ciascun dominio di interesse, si utilizza uno stimatore di pseudo-calibrazione per riportare il valore del risultato all'intero universo delle imprese con il sito web. Si utilizzano i dati dell'indagine per stimare il numero di imprese con il sito web in ciascun dominio di interesse, poiché le fasi 1 e 2 non sono in grado di identificare tutti i siti web delle imprese della popolazione di riferimento.

Gli stimatori delle altre variabili composte

L'indagine ICT rileva altre variabili, che vengono raccolte esclusivamente dal questionario e, combinandole con la variabile *web ordering*, in base alle regole di presenza o copresenza (con istruzioni di "*or"* / "*and"*), vengono definite tre variabili composte (nuove per le statistiche sperimentali). Si definisce un nuovo stimatore di calibrazione tale che i pesi di campionamento riportino alle distribuzioni sulla variabile *web ordering* prodotte con lo stimatore *projection*.

Il nuovo stimatore rispetta i vincoli del corrente stimatore dell'indagine ICT.

Riferimenti bibliografici

Bianchi G., Barcaroli G., Righi P., Rinaldi M. (2019a). Producing contingency table estimates integrating survey data and Big Data. *Itacosm 2019 Conference*.

Bianchi G, Bruni R., Scalfati F. (2018). Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, Mathematical Problems in Engineering, vol. 2018, Article ID 7231920, 8 pages, 2018. https://doi.org/10.1155/2018/7231920.

Bianchi G., Righi P. (2019b). A new estimator for integrating the ICT survey data and the information collected in the enterprises websites. *Technical Report* (download on the www.istat.it in the experimental Statistics webpages).

Breidt. F. J., Opsomer. J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. Statistical Science. 32 . 190–205.

Istat (2018), Cittadini, Imprese e ICT, Statistica Report Anno 2018, https://www.istat.it/it/archivio/226240.

Kim. J. K., Rao. J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*. 99. 85–100.

Righi P, Bianchi G, Nurra A., Rinaldi M. (2020) Integration of survey data and big data for finite population inference in official statistics: statistical challenges and practical applications. Special issue of the journal "Statistics & Applications", to appear.

Scalfati F., Bianchi G., Bruni R., Bianchi F. (2017). Text mining and machine learning techniques for text classification, with application to the automatic categorization of websites, Advisory Committee on statistical methods, Rome, Italy, November 2-3, 2017.

Summa D. (2017). URL retrieval and web scraping procedures https://github.com/summaistat.

Informazioni sulla parte tematica:

Alessandra Nurra
nurra@istat.it
tel. 06 4673.6104

Informazioni su metodologia ed elaborazioni:

Paolo Righi
parighi@istat.it
tel. 06 4673.4419

Gianpiero Bianchi
gianbia@istat.it
tel. 06 4673.4116

Alla produzione e all'analisi hanno collaborato:

- G. Bianchi, F. Bianchi, A. Nurra, P. Righi, M. Rinaldi, S. Salamone, F. Scalfati, D. Summa