Istat
Istituto Nazionale di Statistica

STATISTICHE SPERIMENTALI

8 MAGGIO 2020

The distributions of the enterprise websites by the offered services and functionalities: the estimation process supported by the use of data from the Internet

Since the 2017, the Italian Statistical Institute provides experimental statistics using Internet data for reproducing some estimates currently computed by the European Community Survey on ICT usage and e-commerce in enterprises (ICT survey).

This methodological note elaborates the statistical framework encompassing the use of data from the Internet (Big Data). It suggests to reshape the known statistical tools and leveraged in innovative contexts, to combine information from the multiple imperfect data sources (surveys, administrative and Big Data sources), to model the Big Data selection bias, the survey non-response mechanism and the distribution of the key interest variables and specific marginal distributions. The note covers all the phases of the process highlighting the need to have an integrated statistical framework in which: a) the data sources have to represent different components of a unique informative system; b) the data mining techniques for processing the Big Data (for instance, natural language processing techniques, machine learning techniques, etc.) have to be planned coherently; c) the data analytic methods implemented in different phases (i.e. machine learning techniques and estimators) have to define a comprehensive toolbox; d) the estimators of heterogeneous parameters must define a system of consistent statistics.

In the 2018 experimental statistics on ICT, the estimation process uses the information collected directly from the enterprises' websites to calculate the distributions of the sites which:

1- offer web ordering functions (e-commerce component);

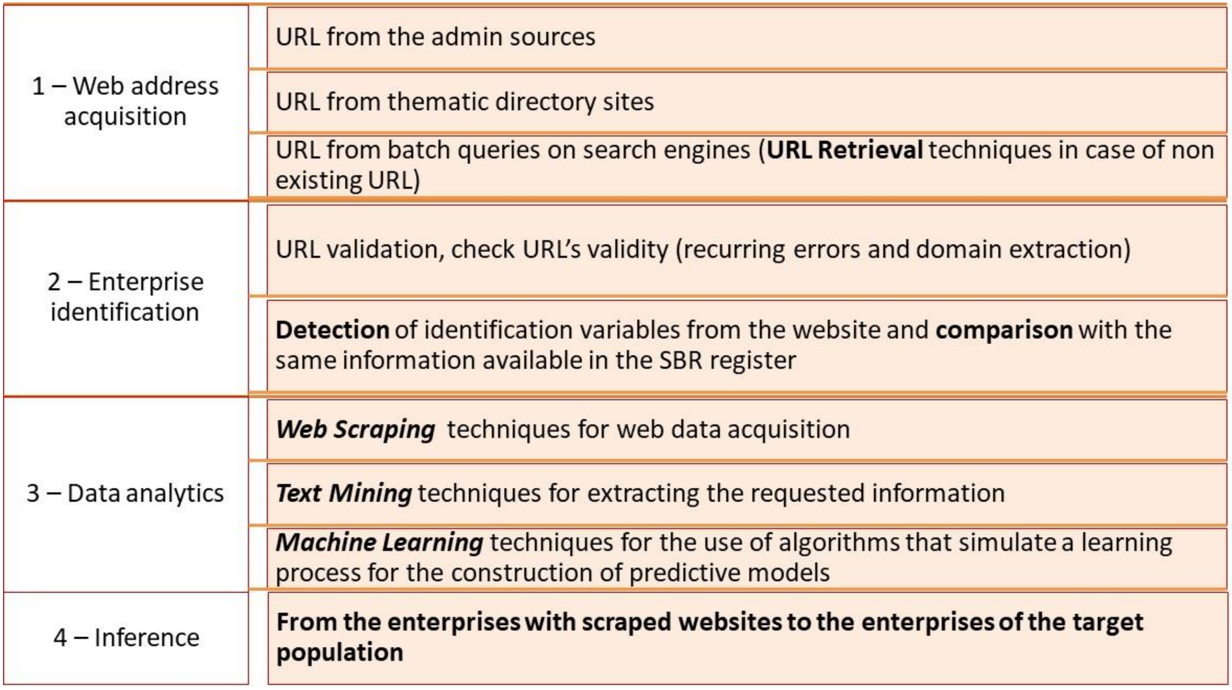
2- propose job offers or gives information on job vacancy in the enterprises;

3- have links to social media (Facebook, Twitter, Instagram etc.);

4- show a combinations of some functions and/or services on the website.

The overall estimation procedure is arranged into four main phases according to the scheme shown in Figure 1. Details of the process are described in Righi *et al.,* 2020 and Bianchi *et al.* (2019a, 2019b).

Figure 1. The main phases for estimating the target distributions that uses information from the website



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**

SBR: Statistical Business Register

The ICT survey

ICT survey is part of the European Community's statistics on information society and represents one of the main annual data sources for the European Digital Agenda Scoreboard, helping to determine the composite index of economy and digital society (DESI) used to summarize the progress of the European digital economy.

The main topics highlighted by the survey are related to Internet connections, use of the Internet (website, social media, cloud computing), the use of digital business process (i.e. use of software to interact and share commercial information internally such as ERP, CRM or externally with other enterprises in the value chain), eCommerce (electronic sales via web, app, online platform, emarketplaces and by electronic data interchanges), electronic invoicing and the most innovative types of ICT investments (Robotics, Internet of Things, Artificial Intelligence, Big Data Analysis) .

The target population of ICT survey refers to the enterprises with 10 and more persons employed working in industry and non-financial market services. The frame population is the Italian Business register (Asia) updated to 2 years before the survey reference period. For the 2018 this population is of 199,416 enterprises. The sampling design is the following: i) a census for the enterprises with 250 and more persons employed (3,342 enterprises); ii) a stratified simple random sample for the smaller and medium enterprises (10-249 persons employed). The stratification variables are: 4 classes of number of persons employed, economic activities (27 Nace groups) and geographical breakdown (21 administrative regions at NUTS 2 level). The sample size is of 33,059 enterprises and respondents is of 22,079 legal units with a response rate of 66.8%.

The target parameters

The target parameters are the rate and distribution of the enterprises that offer services of functionalities in their websites. The experimental statistics consider 6 different parameters:

1- the rate of enterprises offering *web ordering* functionalities in the website;

2- the rate of enterprises offering *job advertisements* in the website;

3- the rate of enterprises with *links to social media* in their websites;

4- the distribution of the enterprises by *website maturity* (values 0,1,2) where the variable is defined according to the following rules:

   a. Website maturity takes into account the presence of 4 website services (SERVICES):

      i. the enterprise' website has online ordering, reservation or booking (web ordering variable);

      ii. the tracking or status of orders placed;

      iii. the possibility for visitors to customise or design online goods or services;

      iv. the personalised content in the website for regular/recurrent visitors;

   <u>and</u>

      v. the enterprises that pay to advertise on the internet (ADS).

   b. The values of the variable are:

      i. 0 - if SERVICES<2 and ADS=0;

      ii. 1 - if SERVICES<2 and ADS=1 or SERVICES>=2 and ADS=0;

      iii. 2 - if SERVICES>=2 and ADS=1.

5- the distribution of the enterprises by the *sophistication of the website* (values 0,1,2,3,4) where the variable is defined according to the presence/absence of the 4 website services (SERVICES) listed above, giving one point for the presence of each of 4 services;

6- the distribution of the enterprises by the variable denoted as *WebF3* (values 0,1) where the variable is defined according to the following rule and giving one point if true:

   a. the enterprise' website has online ordering, reservation or booking (web ordering)

   <u>and at least one</u> of the following 4 website functionalities:

   b. the access to description of goods or services, price lists; the tracking or status of orders placed; the possibility for visitors to customise or design online goods or services; the personalised content in the website for regular/recurrent visitors.

There are several type of domain of interest: the National level, by Size class of persons employed, by Economic macro sectors by size classes, by NACE (26 groups), by Administrative Regions, by NACE Rev. 2 with 2 digits (Divisions).

The estimation procedure

Phase 1-2-3 Automatic categorization of enterprise websites by using web scraping, text mining and machine learning techniques

The goal of these three phases concerns the categorization of websites based on the information contained in them and it is reduced to the widely known problem of text document classification, for which several methodologies are available in the text mining field. The proposed approach is mainly based on a simplified websites' representation, through the automatic generation of standardized data records, which summarize their content. This approach that replaces traditional data collection techniques is achieved through web scraping techniques, combined with natural language processing and machine learning techniques.

The overall website categorization procedure is sketched in the following steps (see also Bianchi *et al.*, 2018).

Firstly, the procedure identifies each enterprise on the web and creates a list of website addresses (about 118,000 URLs). Some URLs are available from survey or from administrative sources, others have been found through a URL retrieval procedure (Summa, 2017), that makes use of search engines and personal data contained in the Statistical Archive of Active Enterprises (ASIA).

The available web addresses are validated through: the syntactic analysis of the URLs, the control of recurrent errors, the check of the authority, the identification of the exact URL.

Given a list of website addresses, for each of them we extract the text of the pages by means of an automatic scraping procedure (Summa, 2017; Scalfati *et al.,* 2017). Beside the text that appears on the webpages, other information is acquired including: the attributes of the HTML tags, the file names, the meta-keywords of the pages.

Classification algorithms cannot work directly on the text in its original form. For this it is necessary to carry out a pre-processing step, in which the raw documents are converted into a simplified form. To accomplish this task, the extracted text (about 94,000 websites) is processed with Natural Language Processing techniques, in order to identify a dictionary of terms (n-grams) useful to describe the content of websites.

When all the websites are represented through a series of standardized data records (feature vectors), classification techniques can be used to classify them according to the aspect of interest.

In particular, we adopt a machine learning approach to predict the value (presence / absence) of the target variables *web ordering* and *job advertisements*. It is a supervised learning based on the availability of a set of labelled records (training set) that constitute the source of information to learn a classifier. So we need a set of websites for which we already have the class labels with respect to the considered categorization.

For this scope, we consider as training set the subset of units detected by the 2018 ICT survey, for which the text extracted from the websites in the same period of survey is available (about 12,000 in the 2018). Both the answers provided to the survey and the texts captured on the websites are used for development of prediction models. Indeed, in case of *job advertisements* we consider 2017

ICT survey data and data collected by websites of sampled enterprises in the same period (about 12,000) because the 2018 sample survey did not collect this variable.

At this point, the prediction models, obtained during the previous training phase, are applied to the generality of enterprises to predict the values of target variables (*web ordering* and *job advertisements*) for all the enterprises for which the retrieval and scraping of their websites was successful (about 94,000).

Instead, in the case of *social media* we are able to exactly collect the value of the target variable by using information retrieval techniques to acquire links to social media directly from the website.

Phase 4 The estimators of the distributions of the target variables

The estimator concerning rate of enterprises offering web ordering functionalities in the website.

The variable is not directly observable from websites. Through text mining and machine learning techniques (phase 3), useful information is extracted to represent websites and classify them with respect to the target variable. The tuning of the classifier uses the 2018 survey data and the data collected in the same period from the websites of the sampled units. We apply the classifier on all the 2018 websites data for obtaining the predictions.

Starting from these predictions, we use the projection estimator (Kim and Rao, 2012; Breidt and Opsomer, 2017) which updates the estimator applied to compute the 2017 experimental statistics.

Compared to the 2017 estimator, the 2018 version takes into account the correction factor based on the difference between the classifications (or predictions) and the values observed on the sampled units. Given the sum of the classifications (or predictions) for each domain of interest, the estimator subtracts the value of the correction factor. Finally, we use a pseudo-calibration estimator to report the result value to the entire universe of enterprises with the website. We use the survey data to estimate the number of the enterprises with the website in each domain of interest, since the phases 1 and 2 are not able to identify all the websites of the enterprises of the reference population.

The estimator concerning rate of enterprises offering job advertisements in the website.

The 2018 sample survey did not collect this variable. Through text mining and machine learning techniques (phase 3), useful information is extracted to represent websites and classify them with respect to the target variable. The tuning of the classifier uses the 2017 survey data and the data collected in the same period from the websites of the sampled units. We apply the classifier on all the 2018 websites data for obtaining the predictions.

Starting from these predictions, we use the projection estimator (Kim and Rao, 2012; Breidt and Opsomer, 2017) which updates the estimator applied to compute the 2017 experimental statistics.

Given the sum of the classifications (or predictions) for each domain of interest, we use a pseudo-calibration estimator to report the result value to the entire universe of enterprises with the website. We use the survey data to estimate the number of the enterprises with the website in each domain of interest, since the phases 1 and 2 are not able to identify all the websites of the enterprises of the reference population.

Unlike the web ordering case, we do not use a correction factor in the projection estimator.

The projection estimator that uses data from the internet allows producing estimates on an annual basis. The sample survey does not detects the variable every year and it produces the estimates on a multiannual basis.

The estimator concerning rate of enterprises with links to social media in their websites.

The estimation process starts from the direct acquisition of the variable using information retrieval techniques. Given the sum of the values collected (number of sites with links to social networks) for each domain of interest, we use a pseudo-calibration estimator to report the result value to the entire universe of enterprises with the website. We use the survey data to estimate the number of the enterprises with the website in each domain of interest, since the phases 1 and 2 are not able to identify all the websites of the enterprises of the reference population.

The estimators for the other composite variables

The ICT survey detects other variables, which are collected exclusively from the questionnaire and, combining them with the web ordering variable, according to the presence or co-presence rules ("or"/"and" instructions), three composite variables are defined (new for the experimental statistics). We define a new calibration estimator such that the sampling weights report to the estimated web ordering distributions produced by the projection estimator.

The new estimator satisfies the main constraints of the current ICT survey estimator.

References

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., Bruni R., Scalfati F., 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.

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), Citizens, Enterprises and ICT, Statistical Report Year 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.

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

Thematic area:

Alessandra Nurra
nurra@istat.it
ph. +39 06 4673.6104

Methodology and results:

Paolo Righi
parighi@istat.it
ph. +39 06 4673.4419

Gianpiero Bianchi
gianbia@istat.it
ph. +39 4673.4116

To the production of these experimental statistics have collaborated:

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