Data Integration challenges - ICT survey and web scraped data from enterprise websites

Istat, Italy

Abstract. Since 2013, the Italian Institute of Statistics (Istat) has investigated the potential of Big Data sources for official statistics. Internet data originated by websites content has since then been considered as one of the most important sources to produce information about enterprises. In 2017, Istat started producing experimental statistics on the activities that enterprises carry out in their websites (web ordering, job vacancy management, link to social media, etc.). They are a subset of statistics currently produced based on the "Survey on ICT in enterprises" and are calculated from the contents of the websites collected with web scraping tools, and processed with Natural Language Processing techniques. A machine learning approach is adopted to estimate models in the subset of enterprises for which both sources are available: survey reported values, and relevant terms obtained by the web scraping/text mining procedures. To all websites the content of which had been processed, best models are then applied. Experimental statistics are obtained using two different estimators: *(i)* a full model based estimator; *(ii)* an estimator that combines model and survey based estimates. Considering the various domains for which they have been calculated, the three sets of estimates (survey, model and combined) in most cases are not distant (i.e. model and combined estimate values lay in the confidence intervals of survey estimates). Simulations have demonstrated that the Mean Square Errors of these new estimates are competitive as compared to those produced in the traditional way.

1. Introduction

The opportunity of producing better official statistics and continued shrinking National Statistical Offices (NSOs) budgets both make Big Data (BD) an appealing new source. The debate on those sources is focused on volume, rapidity, variety and IT capability to capture, store, process and analyse BD for statistical production. Additionally, the NSOs may appreciate also other aspects, such as veracity (data quality, defined as selectivity

and trustworthiness of the information) and validity (correct and accurate data for the intended use). Veracity and validity directly affect the accuracy (bias and variance) of the estimates.

In order to improve veracity and validity, a multi-source approach (based on a combined use of survey, administrative and BD sources) is expected to overcome the limits of each single source, in particular those affecting BD.

This multi-source approach requires a shift in the paradigm of statistical inference. The traditional paradigm followed by the NSOs usually involves design-based survey sampling theory and model-assisted inference. The new paradigm (algorithmic-based inference) is derived by data science: emphasis is on the exploration of all available data, seeking information that has not been extracted yet; models are no longer evaluated based on their interpretability, but rather on their capability to correctly predict values at the unit level, and to use them for estimating parameters of interest.

Istat is currently experimenting this new approach to obtain a subset of the estimates currently produced by the yearly sampling survey on "Survey on ICT usage and e-Commerce in Enterprises", carried out by Istat on Italy and by NSIs of other member states in the EU. We report previous results of this experiment in [1] and [2].

Target estimates of this survey include characteristics of the websites enterprises use to communicate their business (for instance, web ordering facilities; job vacancy management; social networks accounts and activity). Data are collected by means of traditional questionnaires.

An alternative way is to use Internet data, i.e. to collect data by accessing directly the websites, processing the collected texts to single out relevant terms, and modelling the relationships between those terms and the characteristics we are interested to estimate. To do that, the sample of surveyed data plays the role of a training set, useful to fit models that can be applied to the generality of enterprises that own a website. Administrative data (mainly contained in the Business Register) are used to cope with representativeness problems related to the BD source. The sequential application of web scraping, text mining and machine learning techniques obtains auxiliary variables,

suitable for applying a prediction approach and producing estimates comparable to the survey-based ones.

2. Data collection and processing

We have developed a complex procedure to: *(i)* obtain the website addresses of all the enterprises included in the population of reference (URL retrieval); *(ii)* access websites with available URLs and scrape their content (web scraping); *(iii)* processing the content of the scraped websites to identify the best predictors for the target variables (text mining); *(iv)* fit models in the subset of enterprises for which both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful (machine learning).

URL retrieval

In 2017, the share of total number (183,000) of enterprises included in the ICT survey population of interest with a website can be estimated (by the same survey) in 75% (about 135,000 websites). The overall procedure for retrieving as many URLs as possible is described in detail in [3]. Here we indicate the main steps:

1. Using the denomination of a given enterprise as the input for a search engine (Bing), a set of possible links are obtained, the first 10 of which are retained.

2. For each link, the corresponding website is accessed and searched for a number of indicators: the presence of fiscal code, telephone number, address etc., all available in the Business Register.

3. In the subset of enterprises for which the correct URL is available (from a number of previous rounds of the ICT survey, and from other sources), a logistic model is fitted to estimate the probability of correct link on the basis of the values of the above indicators.

4. Only links with probability over a given threshold are retained as valid.

We were able to identify the URLs of about 101,000 websites (75% of the estimated total), of which 14,000 from the current survey, 28,000 from the above procedure, and 59,000 from previous rounds of the survey and other sources.

Web scraping

With a list of about 101,000 URLs, the web scraping task has been performed by accessing, reading and saving the content of each accessible website (about 85,500). Websites were discarded for reasons including wrong specification of the URLs, errors in communicating with their servers or technologies not supported by the parser (mainly websites implemented with ADOBE Flash). The content is the text collected starting from the homepage and continuing with all the other pages reachable from it, down to a certain depth, that can be chosen. The underlying idea is that the pages that are too nested are less relevant for the analysis, and would risk introducing a large amount of noise. Besides the text appearing in the pages, additional information is acquired: attributes of HTML elements, names of the images, keywords of the pages.

Text mining

For each scraped website, the above operations produced a text file, containing unstructured information, in some cases with a huge amount of words (up to one million), most of which are irrelevant for prediction purposes and represent noise to be cleared. To do so, usual data mining techniques can be applied. A detailed description of this phase is in [4].

Machine learning

The web scraping procedure, the processing of scraped texts and the feature selection step produced a Terms-Documents Matrix (TDM), where each row represents a website and each column is referred to an influent word. The intersection cell reports the frequency of times the term is contained in the document. To each row are also attached the values of the target variables observed in the 2017 round of the survey: they are referred to a characteristic of the website, that is if it offers (yes/no) the following facilities: “Online ordering or reservation or booking, e.g. shopping cart (Web ordering)”, “Links or references to the enterprise's social media profiles (Link to social

media)", "Advertisement of open job positions or online job application (Online job application)".

Different learners have been considered: one belonging to the classical statistical parametric models (the Logistic model), others to the ensemble learners (Random Forest, Boosting, Bagging), together with Decision Tree, Naïve Bayes, Neural Networks and Support Vector Machines. Random Forests perform much better: 83% of accuracy and 63% of F1-measure.

3. Estimation

The "Survey on ICT usage and e-Commerce in Enterprises" produces on a yearly basis a set of estimates about rates of web-ordering, job advertising and presence on social media declared by enterprises that own or use websites.

These estimates are available for the total population, and for different domains of interest, including:

1. Cross-classification by Size Classes of persons employed (4) and Economic macro sectors (4) (16 different sub-domains);
2. Administrative Regions (21 different domains);
3. Detailed economic activities (26 domains);
4. ICT and non-ICT enterprises.

Together with the current estimation method (design based/model assisted), alternative estimates have been calculated by adopting two different estimators: a full model based one and a combined one. The characteristics of the three different estimators are reported in Table 1.

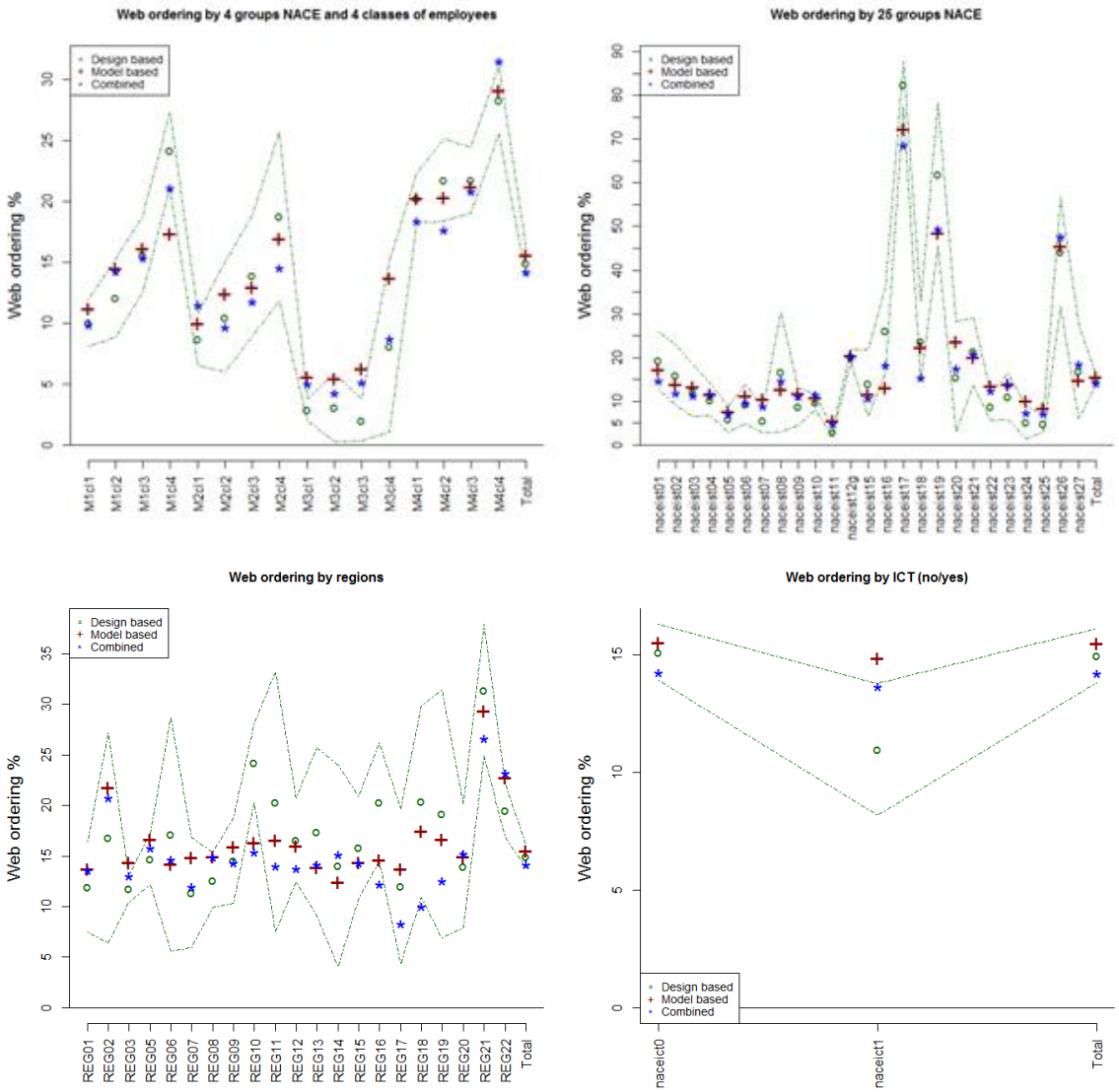
Table 1 **Characteristics of the different estimators**

Estimator	Formula	Weighting	Description
Design based / model assisted	$\ddot{Y} = \sum_r y_k w_k$	$\sum_{k=1}^{r} w_k = N_U$	w_k weights are obtained by calibration procedure of basic weights (inverse of inclusion probabilities) using known totals in the population in order to reduce the bias due to non-response and the variability due to sampling errors
Model based	$\ddot{Y} = \sum_{U^2} \tilde{y}_k w'_k$	$\sum_{k=1}^{U^2} w'_k = N_{U^2}$	The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values $\tilde{y}_k$ for all units for which it was possible reach their websites (population U^2), calibrated in order to make them representative of all the population having websites (U^1).
Combined	$\hat{Y} = \sum_{U^2} \tilde{y}_k + \sum_{r^1} (\tilde{y}_k - y_k) w''_k + \sum_{r^2} y_k w'''_k$	$\sum_{k=1}^{r^1} w''_k = N_{U^2}$ and $\sum_{k=1}^{r^2} w'''_k = N_{U^1 - U^2}$	Estimates are produced by summing three components: -The counting of predicted values in the subpopulation U^2 of units for which it was possible to scrape and process corresponding websites; -An adjustment based on the consideration of the differences between the r^1 reported values and the predicted values (expanded to the same subpopulation U^2); -The counting of observed values for the r^2 respondents that declared a website, that was not found nor scraped, expanded to the whole subpopulation $U^1 - U^2$.

In Figure 2, are reported for the different domains the values of the estimates related to "Web ordering" obtained by using the three estimators. Note that the three sets of estimators are recipracally compatible, as most values produced by the second and third estimators lay in the mid of the confidence intervals calculated for the design based estimates. This is true also for "Links to social media" and "Online job application".

As for the overall quality of the estimates, a simulation, reported in [4], shows that the Mean Square Error of the alternative sets of estimates are competitive compared to the current ones.

Figure 2 **Web-ordering estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)**



4. Conclusions

The complex procedure that Istat developed to harness an important source of Big Data, as the one represented by the Internet data, to improve the estimates currently produced by the Istat ICT Survey, yields a set of experimental statistics, the quality of which must be adequately documented. An important issue are measurement errors in the survey data.

There is evidence of a significant incidence of errors, resulting from manual investigation of a set of websites where predictions are contradictory with values reported in the questionnaire. Errors have a relevant impact on both the fitting of models (errors in the training set) and on the evaluation of their performance (errors in the test set). Failure in taking into account those errors leads to an underestimation of variance and bias components of the design based estimators.

Among the estimation procedures considered in the simulation exercise, the combined ones seem to be competitive compared to the design based ones, but there are still margins of improvement that can lead to an increase in the quality of the model based procedures.

The simulation confirms that one of the most important areas where investment is advisable is an increase of the coverage of the population of enterprises with a website. So far, URLs retrieval has been based on a variety of sources and techniques, while a real solution is to proceed to a census collection of this information, by asking website address in every survey, and offering the opportunity to communicate this information in the Istat Enterprises Portal. A general agreement on this has been already reached. Of course, a higher coverage of the population websites will reduce the bias represented by the websites with unknown or incorrect address. Informing the enterprises that their websites will be accessed and their content collected for statistical purposes will also increase the number of websites successfully accessed for scraping: this will be made next year, in the coming wave.

The amount of valid information can also be increased by adding to the HTML text information from images, by using Optical Character Recognition (OCR) techniques: an Istat application to this purpose has already been developed and tested [8]. This will increase the performance of the predictors, and consequently the MSE of the estimators.

Finally, the work done so far cannot be limited to a replication of already available statistical information. The prediction at the unit level for the whole population of interest will enrich the information in the Business Register. New aggregate information can be produced, for example, to monitoring the “Internet economy”, as proposed by Statistics Netherlands [6].

References

[1] Barcaroli, G. Nurra A., Salamone S., Scannapieco M., Scarnò M., Summa D. (2015) - Internet as Data Source in the Istat Survey on ICT in Enterprises. Austrian Journal of Statistics, Volume 44, 31-43. April 2015.

[2] Barcaroli G., Bianchi G., Bruni R., Nurra A., Salamone S., Scarnò M. (2016) - Machine learning and statistical inference: the case of Istat survey on ICT. Proceedings 48th Scientific Meeting Italian Statistical Society (2016).

[3] Barcaroli G., Scannapieco M., Summa D. (2016) - On The Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web. Rivista italiana di economia, demografia e statistica Volume LXX(n.4):20-41 · October 2016

[4] 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. To be presented at the Advisory Board (November 2017)

[5] Barcaroli G., Golini N., Righi P. (2018) - Quality evaluation of experimental statistics produced by making use of Big Data. Proceeding of Q2018 (Krakow, July 2018)

[6] Oostrom L., Walker A., Staats B., Slootbek-Van Laar M., Ortega Azurduy S., Rooijakkers B.: Measuring the internet economy in The Netherlands: a Big Data analysis. CBS Discussion Paper n. 2016/14