# Use of Internet data in Official Statistics: a proposal for a methodological framework

Giulio Barcaroli, Paolo Righi

Abstract

*The multi-source approach is a must for National Statistical Offices, as they have to face negative trends in terms of available budget and decreasing response rates in statistical survey. Survey data are no longer to be considered the only source: administrative data and new sources, as the ones belonging to the vast family of Big Data, can be jointly harnessed in order to produce new information, or to increase quality and/or decrease costs of already produced one. In this paper we advance a proposal for a methodological framework useful (i) to develop a process in which survey data are used together with Internet data as a basis for modelling relations between directly observed data and data available in Internet, to be combined so to ensure a higher quality to estimates, and (ii) to evaluate quality of these estimates. A concrete case is illustrated, based on the Istat Survey on ICT usage and e-Commerce in Enterprises.*

**Keywords:** *Internet data, multi-source, web scraping, text mining, machine learning, accuracy*

## 1. Introduction

The opportunities of producing enhanced statistics and the declining budgets, make using Big Data (BD) in National Statistical Offices (NSOs) appealing. Often the debate on these sources is focused on volume, velocity, variety and on IT capability to capture, store, process and analyse BD for statistical production. Nevertheless, other features have to be taken into account, especially in the NSOs, such as veracity (data quality 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) should allow to overcome usual 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 one followed by NSOs is usually based on design-based survey sampling theory and model-assisted inference. The new one (algorithmic-based inference) is derived by data science: the emphasis is on the exploration of all available data, seeking information that has not been extracted so far; models have to be evaluated no longer by their interpretability, but rather by their capability to correctly predict values at unit level, and to use them for estimating the parameters of interest.

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

Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data are collected by means of traditional questionnaires.

An alternative way is to make use of Internet data, i.e. to collect data by accessing directly the websites, processing the collected texts to individuate relevant terms, and modelling the relationships between these 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 owning a website. Administrative data (mainly contained in the Business Register) are used to cope with representativeness problems related to BD source. The sequential application of web scraping, text mining and machine learning techniques allows to obtain auxiliary variables suitable for applying a prediction approach and produce estimate that can be compared to the survey ones.

In terms of quality (accuracy), the impact of the new estimators is both positive (reduction of the variability and of the bias due to sampling variance, to total non-response and to measurement errors in the survey) and negative (model bias

and variance). Whenever the quality of estimates obtained by means of this new approach reveals to be not lower than the ones produced by the traditional process, the former has to be preferred, as it allows not only to produce aggregate estimates, but also to predict individual values, useful for instance to enrich the information contained in registers.
The crucial point is therefore in the definition of a methodological framework that allows the efficient use of the different sources, and also the evaluation of the accuracy of the estimates obtained by the model-based approach, to be compared with the accuracy of the traditional design-based estimates.
For this purpose, a simulation has been carried out under realistic assumptions, and results in terms of variance, bias and total mean square error have been produced, showing that, under given conditions (mainly related to population coverage with BD source and to the model performance), estimates obtained under this new approach are characterised by a higher accuracy. The simulation study is based on real 2017 Istat "Survey on ICT usage and e-Commerce in Enterprises" data (ICT survey). A synthetic enterprise population with websites has been built up. Target and scraped from the website variables have been generated according to the distributions observed in ICT survey. The performances of the estimators are shown in terms of bias, variance and mean square error, comparing design based estimators [3,4] and supervised model based estimators [5] using scraped data are compared.

## 2. The Istat Survey on ICT usage and e-Commerce in Enterprises

The European Community Survey on *ICT usage and e-commerce in enterprises* (ICT survey) is intended to measure the degree of use of new technologies in companies and provides EU with information for comparison among Member States and evaluation of national policies on their capacity to grasp the potential of technological progress. The survey provides a wide and articulated set of indicators on Internet activities and connection used, e-Business, e-Commerce, ICT skills, e-Invoice. ICT survey is also one of the major sources of data for the *Digital Agenda Scoreboard* measuring progress of the European digital economy.
The target population is about Italian enterprises with at least 10 persons employed active in the manufacturing, electricity, gas and steam, water supply, sewerage and waste management, construction and non-financial services.
The survey is carried out as a sample survey for enterprises with 10-249 persons employed and it includes all enterprises with at least 250 persons employed. The sample is drawn from the Italian business register (Asia) that is updated with information relating to two years before the survey time reference; the stratified simple random sampling method is used and the strata are defined by the combination of economic activities, the size classes of statistical units in terms of persons employed and the administrative regions in which enterprises are located. Given the type of the survey sample, in the grossing up procedure the estimated total of the variables of interest is calculated by giving each respondent unit a weight, indicating the number of the population represented by the unit, including itself. The final weights to be associated with responding units are calculated making use of the calibration estimator theory of Deville and Särndal [6] used when the totals for any variable by domain are known and correlated with those of interest (the auxiliary variables used are number of enterprises and number of persons employed). In year 2017 sample involved about 32,000 enterprises representative of a universe of 183,000 units, and response rate was about 61%.

In a section of the questionnaire are reported the questions related to the availability of a website, and to the presence on it of a set of characteristics:

| | **Use of a Website** | | |
|---|---|---|---|
| **C9.** | **Does your enterprise have a Website?** (Filter question) | Yes ☐ | No ☐ ->go to C11 |

| **C10.** | **Does the Website have any of the following?** | Yes | No |
|---|---|---|---|
| | a) Description of goods or services, price lists | ☐ | ☐ |
| | *[8] b) Online ordering or reservation or booking, e.g. shopping cart | ☐ | ☐ |
| | c) Possibility for visitors to customise or design online goods or services | ☐ | ☐ |
| | d) Tracking or status of orders placed | ☐ | ☐ |
| | e) Personalised content in the website for regular/recurrent visitors | ☐ | ☐ |
| | f) Links or references to the enterprise's social media profiles | ☐ | ☐ |
| | g) Advertisement of open job positions or online job application ***- Optional*** | ☐ | ☐ |

The experiment carried out aimed at verifying the possibility to produce the estimates usually produced by the survey, based on data collected in this section via the traditional way and applying design based estimation methods, adopting an alternative procedure, based on the direct use of Internet data and applying model based estimation methods.

## 3. Web scraping, text mining and machine learning to predict web ordering at unit level

A complex procedure has been developed in order to:

1. get the website address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (*URL retrieval*);
2. access websites with available URL and scrape their content (web scraping);
3. processing the content of the scraped websites in order to individuate the best predictors for the target variables (*text mining*);
4. fit models in the subset of enterprises where 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 total number of enterprises included in the ICT survey population of interest (183,000) 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 [7]. 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, of which the first 10 are retained.
2. For each link, the corresponding website (if existing) 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 in order to estimate the probability of correct link on the basis of the values of the above indicators.
4. Only the links whose probability exceed a given threshold are retained as valid.

In this way 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*

Having in input a list of about 101,000 URLs, the web scraping task has been performed by accessing, reading and saving the content of each website for which the access was possible (about 85,500).
Among the reasons for not having scraped all the websites there were 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, while they would mainly introduce a large amount of noise. On the other hand, besides the text appearing in the pages, additional information is acquired: the attributes of HTML elements, the names of the images, the keywords of the pages.

*Text mining*

The above operations produced for each scraped website 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 that has to be eliminated. To do so, usual data mining techniques can be applied. A detailed description of this phase is in (Bianchi et al, 2017). Here we summarize the different steps.

1. *Stop-words removal*. Stop-words (articles, prepositions, etc.) are removed since their generic meaning has practically no relevance for the prediction task.
2. *Selection of dictionaries.* A training set S composed of 50% of the cases for which both scraped texts and ICT survey data are available, is selected. Two dictionaries are derived from it: the set W of the unigrams, i.e., the

single words, appearing in those files, and the set Z of the n-grams appearing in the same files. N-grams are sequences of *n* adjacent words that are typically used together. An example is “credit card", which is a bi-gram.

3. *Lemmatization*. For the set of unigrams W, lemmatization is performed, that is, the inflectional ending of each word is removed in order to return the word to its basic lemma.
4. *Part-of-speech (POS) tagging*. For the set of n-grams Z, lemmatization is not performed, since substituting words with their basic lemmas may result in losing the identity of many n-grams, which are generally built with specific inflectional forms. Instead, part-of-speech recognition (POS tagging) is carried out, that is, each word is identified as a particular part of speech (e.g., a noun, a verb, etc.). Thus, we obtain at this stage the following two sets of terms:
    a. set W, whose component terms are single lemmas in Italian or English language (unigrams);
    b. set Z, whose component terms are syntactically well-composed n-grams in Italian or English language (n-grams).
5. *Term Evaluation (TE) and Feature Selection*. For each of the terms in W0 and Z0 a measure of its relevance with respect to the given target variable is calculated by means by using a Chi-square metric. After this, all terms in W and in Z are sorted by decreasing score values and all the terms with a TE score larger than a threshold and up to a maximum number of terms. The final set contains 1000 relevant terms, of which about 800 uni-grams and 200 bi-grams.

*Machine learning*

The execution of 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*)”, “Description of goods or services, price lists (*Catalogue*)”, “Possibility for visitors to customize or design online goods or services (*Products personalization*)”, “Tracking or status of orders placed (*Order tracking*)”, “Personalized content in the website for regular/recurrent visitors (*Content personalization*)” “Links or references to the enterprise's social media profiles (*Link to social media*)”, “A privacy policy statement, a privacy seal or a website safety certificate (*Privacy*)”, “Advertisement of open job positions or online job application (*Online job application*)”.
In the following we will focus only on the target variable *Web ordering*.
This matrix is the input to the fitting of models in which observed variables are the y’s and the terms are the x’s.
The TDM contains 11,877 rows and 1,000 columns, and is equally split in training and testing subsets.
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.
The fitting of the models is still in progress. Some preliminary results obtained so far report for the logistic model an accuracy of 80%, together with a F1-measure (harmonic mean of precision and recall) equal to 0.56%. Random Forests perform much better: 83% of accuracy and 63% of F1-measure.

## 4. Basic notation and estimation procedures

Let $U$ be the reference population of $N$ elements and let $U_d$ ($d$ = 1, …, $D$) be an estimation domain, where the $U_d$’s partition $U$. $U_d$ is a sub-population of $U$ with $N_d$ elements, for which separate estimates are calculated. Let $y_k$ denote the value of the interest variable attached to the $k$-th population unit ($k$=1, …, $N$). The parameters to be estimated are $Y_d = \sum_{k \in U_d} y_k$ and $Y = \sum_{k \in U} y_k$.
For defining the estimation procedure let us introduce a further partition of $U$. Let $U^v$ ($v$=1, …, $V$) be a sub-population of size $N^v$, characterized by the availability of auxiliary information, for instance a sub-population in which auxiliary variables from BD source are available. Let $\mathbf{x}_k^v$ be the auxiliary variable vector from BD source and $\mathbf{z}_k^v$ be the auxiliary variable vector known from the frame list for unit $k$. For simplicity, we assume $V = 2$, $\mathbf{z}_k^v = \mathbf{z}_k \forall v$ **,** and for $k \in U^1$ the vector $(\mathbf{x}_k^1, \mathbf{z}_k)$ is known, while for $k \in U^2$ only $\mathbf{z}_k$ is known. Then the totals $\mathbf{Z}_d = \sum_{k \in U_d} \mathbf{z}_k$ are known. The $U^v$’s cross cut the $U_d$’s , then $U_d^v = U_d \cap U^v$. We assume known the totals $\mathbf{Z}_d^v = \sum_{k \in U_d^v} \mathbf{z}_k$.
In the sampling strategy, $y_k$ is observed with a random sample $s$ of size $n$. The sample could be affected by non-response. Let $r$ be the number of respondents in $s$ and let $r_d$ and $r^v$ be respectively the number of respondents belonging to $U_d$ and $U^v$. In the observed sample, we can estimate a model $\tilde{y}_k = f(\mathbf{x}_k^v, \mathbf{z}_k^v)$ for predicting the $y$ variable.

The ICT Survey regularly produces a set of estimates regarding the use of websites made by enterprises, making use of the values collected via traditional means (questionnaires).
In particular, the total number of enterprises offering web ordering facilities on their websites is calculated by using a design based / model assisted approach:

$$\hat{Y} = \sum_r y_k w_k$$

where $w_k$ weights are obtained by calibration procedure [3] of basic weights (inverse of inclusion probabilities) making use of known total in the population, basically to handle non response. Hereinafter, we denote the estimator as Des2.

Here we introduce two alternative estimation procedures. They make use of values predicted for each unit whose website is successfully accessed, scraped and processed.
A first one is:

$$\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k w_k + \sum_r y_k w_k$$

where the estimate of the total is given by the count of the predicted values $\tilde{y}_k$ (for units not in the sample), plus the count of the observed ones $y_k$ (for units in the sample). Also in this case there is an adjustment for non response. In particular we assign value 1 as basic weights to all units and obtain the $w_k$ weights according to a calibration procedure, hereinafter denoted as pseudo-calibration. Hereinafter, we denote the estimator as Mod2.
A variant of this procedure is the following:

$$\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k + \sum_{r^1} y_k + \sum_{(r - r^1)} y_k w_k$$

i.e. the total number of enterprises offering web ordering facilities is obtained counting the predicted values and the observed values for the units for which it was possible to access and scrape their websites, while for the remaining part the subtotal is obtained by calibrating observed values in the sample of respondents for which there was no website (calibration with known totals pertaining to the population without websites). Hereinafter, we denote the estimator as Comb2.
Table 1 shows the basic notation of the three estimators $\hat{Y}$ of *Y* that are compared in the simulation. The derivations of the $\hat{Y}_d$ of $Y_d$, are straightforward. Moreover, you should note the table show three more estimators. These estimators, denoted as Mod1, Des1 and Comb1 represent the not adjusted versions of the previous estimators.
The list of estimators is not exhaustive but broadly maps possible estimators.
In general, figures produced by these estimators can be very different. The compared evaluation of their quality is fundamental, in order to choose the best one, i.e. the one with the minimum Mean Square Error. The estimation of variance of estimators can be done by using available data. For the bias component, we must rely on simulation techniques.

**Table 1:** General description of the estimators used in the simulation.

| ***Estimator*** | ***Expression*** | ***Description*** | ***Note*** |
|---|---|---|---|
| Mod 1 | $\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k b_k + + \sum_r y_k b_k$ | $b_k = N/(N^1 + r - r^1)$ | Model based estimator |
| Mod 2 | $\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k w_k + + \sum_r y_k w_k$ | $w_k$ calibration of $b_k$'s defined in Mod1 being $\sum_{r_d} \mathbf{z}_k w_k = \mathbf{Z}_d \; \forall d$ | Pseudo-calibration model based estimator |
| Des1 | $\hat{Y} = (n/r) \sum_r y_k b_k$ | $b_k$ is the sampling basic weight | Horvitz-Thompson estimator. corrected for non response |
| Des2 | $\hat{Y} = \sum_r y_k w_k$ | $w_k$ calibration of $b_k$'s defined in Des1 being $\sum_{r_d} \mathbf{z}_k w_k = \mathbf{Z}_d \; \forall d$ | Calibration estimator. |
| Comb1 | $\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k + \sum_{r^1} y_k + + (n/r) \sum_{(r - r^1)} y_k b_k$ | $b_k$ is the sampling basic weight | Combined est. Mod1 and Des1 |
| Comb2 | $\hat{Y} = \sum_{(U^1 - r^1)} \tilde{y}_k + \sum_{r^1} y_k + \sum_{(r - r^1)} y_k w_k$ | $w_k$ calibration of $b_k$'s defined in Des1 being $\sum_{(r_d - r^1)} \mathbf{z}_k w_k = \mathbf{Z}_d^v \; \forall d$ | Combined est. Mod1 and Des2 |

## 5. Evaluation of estimators based on simulation

Accuracy of statistical estimates is traditionally decomposed into bias (systematic error) and variance (random error) components. While variance can be estimated, bias is non observable if the parameter of interest is unknown. So we have to rely on a simulation exercise.
We studied the accuracy of a set of estimators via Monte Carlo simulation. A synthetic population based on the 2016 ICT survey data has been created. The estimators that have been considered can be distinguished with respect to:

a. the origin of the exploited auxiliary information, coming from the frame list, from a BD source or both;
b. the inferential approach (design based, model based and a combination of both).

### *Target population*

We consider the set of the Italian enterprises with 10 to 249 employed persons in activities of manufacturing, electricity, gas and steam, water supply, sewerage and waste management, construction and non-financial services (near 183,000 units). The population and a **z** vector of auxiliary variables (location, unit size, and economic activity) are identified by the Italian Business Register (BR).
Currently, Istat uses this register as frame list for drawing the yearly ICT survey. The frame list (BR) is updated with information relating to two years before the survey time reference. Among the target estimates of the ICT survey there are a number of characteristics related to the functionalities of the websites: for instance the presence of *Web ordering* or *Job application* facilities. The simulation focuses on a single binary variable i.e. *Web ordering*, denoted as *y* variable, being $y_k = 1$ if unit *k* does e-commerce and $y_k = 0$ otherwise. The target parameters are the count of $y_k = 1$ at domain level (type of *economic activity* by *size class* of employed persons), $Y_d$ ($d$ = 1, …, 16) and total level, *Y*. In particular, the type of economic activities are denoted as M1, M2 M3 and M4 and the size class of employees are denoted as cl1 (small), cl2, cl3 and cl4 (large). Since the survey estimates show that about 25% of BR units do not have a website, we exclude these units from the analysis, and remaining units define the target population $U$. The discarded units follow the distribution observed in the 2016 ICT survey in the 16 domains. We note that in practice the size of $U$ should be treated as random. The *y* variable is unknown in $U$, so we create the probability $p(y_k = 1)$ for each unit by means of a logistic model, $\text{logit}(y_k) = \alpha + \mathbf{z}_k'\boldsymbol{\beta}$ (hereinafter denoted as *true model*) where $\alpha$ and $\boldsymbol{\beta}' = (\beta_1, \ldots \beta_d, \ldots, \beta_{16})$ are known regression coefficient and $\mathbf{z}_k' = (z_k, \ldots z_{dk}, \ldots, z_{16k})$, being $z_{dk} = 1$ if $k \in U_d$ and $z_{dk} = 0$ otherwise. We fix $\alpha$ and $\boldsymbol{\beta}$ such that the sum over the $U_d$'s of $p(y_k = 1)$ reflects observed distribution in the last 2016 Istat ICT survey (Table 2, column $\boldsymbol{p}$).
The population $U$ is partitioned in 3 sub-populations, $W^1, W^2$ and $W^3$ :

- $W^1$, the enterprises with website address (URL) available;
- $W^2$, the enterprises with wrong URL or website not allowing automatic scraping;
- $W^3$, the enterprises having website but the URL is not available;

We generated the distribution in the 3 sub-populations following the evidences:

- Istat has got a list of business units where the website address (URL) is available. The inclusion in the URL-list is on volunteer basis and it does not cover all the business register (101,000 enterprises, $W^1 \cup W^2$);
- in a concrete application of automatic web scraping procedure 68,676 websites have been investigate ($W^1$) and 32,320 have been not ($W^2$) .

We assume the URL-list suffers from selectivity problems, that is the distribution of target variable within the URL-list ($W^1 \cup W^2$) differs from the distribution of the unit out this list, $W^3$.This reflects the hypothesis that if an enterprise makes use actively of its website for business (for instance doing e-commerce) then it has interest to increase its reachability, and therefore the probability to be in the Url-list is higher. Table 2 shows the sizes and the expected $p(y_k = 1)$ for the 3 sub-populations.

**Table 2:** Population size by domains and $W^1, W^2$ and $W^3$ and the related probability of Web ordering facilities

| | Population Size | | | | Expected probability of Web ordering | | | |
|---|---|---|---|---|---|---|---|---|
| **Domain** | $W^1$ | $W^2$ | $W^3$ | $U$ | $p^1$ | $p^2$ | $p^3$ | $p$ |
| M1 cl1 | 23,519 | 10,995 | 11,435 | 45,949 | 0.170 | 0.170 | 0.048 | 0.140 |
| M1 cl2 | 3,146 | 1,499 | 1,595 | 6,240 | 0.154 | 0.154 | 0.023 | 0.120 |
| M1 cl3 | 1,873 | 887 | 853 | 3,613 | 0.218 | 0.218 | 0.014 | 0.170 |
| M1 cl4 | 922 | 440 | 370 | 1,732 | 0.333 | 0.333 | 0.000 | 0.261 |
| M2 cl1 | 1,122 | 565 | 578 | 2,265 | 0.138 | 0.138 | 0.037 | 0.110 |
| M2 cl2 | 237 | 97 | 82 | 416 | 0.124 | 0.124 | 0.027 | 0.110 |
| M2 cl3 | 146 | 71 | 84 | 301 | 0.151 | 0.151 | 0.009 | 0.110 |
| M2 cl4 | 120 | 53 | 44 | 217 | 0.222 | 0.222 | 0.000 | 0.181 |
| M3 cl1 | 5,408 | 2,486 | 2,992 | 10,886 | 0.050 | 0.050 | 0.013 | 0.040 |
| M3 cl2 | 382 | 176 | 206 | 764 | 0.026 | 0.026 | 0.004 | 0.020 |
| M3 cl3 | 168 | 78 | 81 | 327 | 0.039 | 0.039 | 0.002 | 0.030 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| M3 cl4 | 65 | 27 | 27 | 119 | 0.025 | 0.025 | 0.000 | 0.020 |
| M4 cl1 | 26,525 | 12,574 | 11,289 | 50,388 | 0.319 | 0.319 | 0.103 | 0.270 |
| M4 cl2 | 2,430 | 1,144 | 890 | 4,464 | 0.379 | 0.379 | 0.081 | 0.320 |
| M4 cl3 | 1,527 | 712 | 507 | 2,746 | 0.396 | 0.396 | 0.036 | 0.330 |
| M4 cl4 | 1,086 | 516 | 371 | 1,973 | 0.396 | 0.396 | 0.000 | 0.321 |
| **Total** | **68,676** | **32,320** | **31,404** | **132,400** | **0.235** | **0.235** | **0.061** | **0.194** |

The simulation works with $U^1 = W^1$ and $U^2 = W^2 \cup W^3$.
For completing the synthetic population we generate the output of the web scraping so that Internet is the BD source of the simulation.
The automatic scraping is not able to observe the variable *y*, but instead it collects all texts from websites and, in a second step, based on the use of text mining and natural processing techniques, relevant terms are detected to play the role of predictors (for instance: "add to cart", "credit card", "order", etc.) [1,2]. We assume to observe, at the end of the process, 12 binary variables (presence/absence), denoted by the **x** vector. We underline that in practical application this number is much larger. Nevertheless, a larger set of variables would only complicate the simulation without adding information. "Good" estimates are achieved when the target variable and the set of auxiliary variables (large or small) have a strong relationship: this result in high levels of performance indicators of models.
We generate the 12 auxiliary variables according to two scenarios:

1- weak dependence with the target variable (harmonic mean of precision and recall indicators equal to 63%);
2- strong dependence with the target variable ((harmonic mean of precision and recall indicators equal to 96%).

In particular, the first scenario seems closest to the performance indicators of the prediction models observed on the real 2016 ICT data, confirmed in 2017. Actually, we should consider the presence of measurement errors in survey data, that affect the train and the test set. If we could eliminate these errors, the quality of the prediction would be substantially higher (up to 10 percentage points for recall and precision).
Scenario 2 remains a benchmark in evaluation analysis.

## *The simulation process*

The simulation implements a feasible and reasonable estimation process. We consider a supervised approach, such that the target variable is observed in a sample, for instance in the ICT sample. We assume a stratified simple random sampling design with four strata defined by the size classes, cl1,…, cl4. The sample size is 32,000, the number of respondents is 23,229. Respondents are allocated with 16,307 units for cl1, 1,820 units for cl2, 1,061 units for cl3 and 4,041 units for cl4. Largest rates are assigned to the large enterprises in terms of employees reflecting the real situation. We generate availability of information from scraped websites assuming homogeneous rates in each stratum (cl1 probability= 0.45, cl2 probability= 0.88, cl3 probability= 0.95, cl4 probability= 0.97). The sample of respondents with corresponding availability of Internet data has expected size of about 13,800 units (as in the 2016 ICT survey).
At domain level the sample size is not planned. We had three domain types: Large (L), Small (S) Very Small (VS) (see Table 3).

**Table 3:** Expected size and Web ordering frequency in the observed sample

| Domain | Size | Web ordering | Type |
|---|---|---|---|
| M1 cl1 | 3.074,09 | 430,45 | L |
| M1 cl2 | 845,37 | 101,37 | L |
| M1 cl3 | 520,21 | 88,42 | L |
| M1 cl4 | 1.681,42 | 438,14 | L |
| M2 cl1 | 151,53 | 16,63 | S |
| M2 cl2 | 56,36 | 6,19 | VS |
| M2 cl3 | 43,34 | 4,78 | VS |
| M2 cl4 | 210,66 | 38,07 | L |
| M3 cl1 | 728,30 | 29,16 | S |
| M3 cl2 | 103,50 | 2,06 | VS |
| M3 cl3 | 47,08 | 1,43 | VS |
| M3 cl4 | 115,53 | 2,34 | VS |
| M4 cl1 | 3.371,07 | 910,32 | L |
| M4 cl2 | 604,77 | 193,47 | L |
| M4 cl3 | 395,37 | 130,51 | L |
| M4 cl4 | 1.914,43 | 613,66 | L |
| **Total** | **13.863,04** | **3.007,00** | Total |

The estimation process follows these steps:

1. Collect the *y* variable for respondent units with website;
2. Make the web scraping for the units in $U^1$and collect the **x** variables;

3. Model *y* on **x** in $r^1$;
4. Produce the estimate according to a given estimator.

For estimators Des1 and Des2 (Table 1), steps 2. and 3. are skipped.
The simulation compares 6 different estimators of Table 1. We note that:

- Mod1, Mod2, Comb1 and Comb2: $\tilde{y}_k = \hat{p}(y_k = 1)$ is predicted with a working logistic model using the **x** variable;
- Des1: uses an incorrect MCAR [4] model for the non-response weight adjustment;
- Des2: calibration performs a correct weight adjustment for non-response;
- Comb1, Comb2: produce estimates for $U^1$ (using Mod1 ) and $U^2$(using Des1 or Des2);
- Comb2: calibration performs a correct weight adjustment for non-response in $U^2$.

## *Results*

The simulation takes into account the methodological frameworks of the respective estimators. For the model based estimators the *y* variable is treated as random, and once selected the sample, the *y* values change over each iteration. In the design based estimator the *y* values are fixed, and then in each iteration a new random sample is selected. The simulation implements 1,000 iterations and computes for each iteration the estimates $\hat{Y}_{j,d,i}$ for the *j*-th estimators, the *d*-th domain in the *i*-th iteration.
The following statistics are considered for Mod1, Mod2, Des1 and Des2:

- the relative bias, $RB(\hat{Y}_{j,d}) = \frac{\frac{1}{1,000}\sum_{i=1}^{1,000}(\hat{Y}_{j,d,i} - Y_d)}{Y_d}$;
- the coefficient of variation, $CV(\hat{Y}_{j,d}) = \frac{\sqrt{\frac{1}{1,000}\sum_{i=1}^{1,000}(\hat{Y}_{j,d,i} - \bar{\hat{Y}}_{j,d})^2}}{Y_d}$, being $\bar{\hat{Y}}_{j,d} = 1/1{,}000[\sum_{i=1}^{1,000} \hat{Y}_{j,d,i}]$;
- the relative root mean square error, $RRMSE(\hat{Y}_{j,d}) = \frac{\sqrt{[RB(\hat{Y}_{j,d})\, Y_d]^2 + [CV(\hat{Y}_{j,d})Y_d]^2}}{Y_d}$.

For the estimators Comb1 and Comb2 the numerator of the *RB* becomes $1/1{,}000[\sum_{i=1}^{1,000}\sum_v(\hat{Y}_{j,d,i}^v - Y_d^v)]$, the numerator of the *CV* becomes $\left\{1/1{,}000[\sum_{i=1}^{1,000}\sum_v(\hat{Y}_{j,d,i}^v - \bar{\hat{Y}}_{j,d}^v)^2]\right\}^{1/2}$ where $\bar{\hat{Y}}_{j,d}^v = 1/1{,}000[\sum_{i=1}^{1,000}\sum_v \hat{Y}_{j,d,i}^v]$ in which $\hat{Y}_{j,d,i}^v$ is the *j*-th estimator in the *i*-th iteration of $Y_d^v = \sum_{k \in U_d^v} y_k$.

Table 4a shows that the model based estimators produce biased estimates for all the domain types. These results convey that if we use a predictive model estimated on a sample representing a specific population ($W^1$) such model does not fit for the other populations (such as $W^3$). Calibration in Mod2 estimator partially correct the bias. Discrepancies between Scenario 1 and 2 confirm the importance of using a good working model for improving the accuracy (bias).

**Table 4a:** Maximum values of accuracy indicators observed in the simulation for model based estimators

| | | ***Domain Type*** | | | |
|---|---|---|---|---|---|
| ***Estimator*** | ***Statistic*** | **VS** | **S** | **L** | **Total** |
| Mod1 Scenario1 | CV | 112.90 | 10.54 | 25.42 | 1.82 |
| | RBIAS | 629.80 | 313.08 | 74.46 | 28.47 |
| | RRMSE | 632.17 | 313.26 | 77.97 | 28.53 |
| Mod1 Scenario2 | CV | 111.24 | 8.63 | 25.54 | 0.65 |
| | RBIAS | 85.35 | 44.75 | 74.72 | 19.11 |
| | RRMSE | 135.34 | 45.36 | 77.43 | 19.12 |
| Mod2 Scenario1 | CV | 65.47 | 10.51 | 14.75 | 1.83 |
| | RBIAS | 628.42 | 342.75 | 70.70 | 27.72 |
| | RRMSE | 630.26 | 342.91 | 70.82 | 27.78 |
| Mod2 Scenario2 | CV | 64.65 | 8.79 | 14.86 | 0.67 |
| | RBIAS | 90.87 | 55.11 | 25.63 | 17.54 |
| | RRMSE | 99.44 | 55.66 | 26.03 | 15.56 |

Table 4b shows the two design based estimators. Focusing on the calibration estimator (Des2), the correct weight adjustments produce nearly unbiased estimates but high *CV* and *RRMSE* especially for VS and S domains.

**Table 4b:** Maximum values of accuracy indicators observed in the simulation for design based estimators

| Estimator | Statistic | Domain Type | | | |
|---|---|---|---|---|---|
| | | VS | S | L | Total |
| Des1 | CV | 142.30 | 18.08 | 14.75 | 1.62 |
| | RBIAS | 61.61 | -25.88 | 62.56 | -9.36 |
| | RRMSE | 153.85 | 31.57 | 62.73 | 9.50 |
| Des2 | CV | 89.33 | 23.97 | 9.25 | 1.92 |
| | RBIAS | -1.59 | -1.68 | 0.39 | -0.02 |
| | RRMSE | 89.33 | 24.03 | 9.25 | 1.92 |

Table 4c show the accuracy of blended estimates, combining the model and design based estimates. We note that Comb1 - Scenario 2 is highly competitive with respect to Des1 estimators. We underline that both estimators do not adjust correctly the weights of the $r - r^1$ sampled units. Comparing Comb2-Scenario 2 with Des2 the first estimator seems better for S domain, competitive for VS, L and Total domains.

**Table 4c:** Maximum values of accuracy indicators observed in the simulation for combined estimators

| Estimator | Statistic | Domain Type | | | |
|---|---|---|---|---|---|
| | | VS | S | L | Total |
| Comb1 Scenario1 | CV | 83.27 | 9.95 | 12.79 | 1.48 |
| | RBIAS | 391.99 | 156.88 | 41.97 | 1.46 |
| | RRMSE | 399.29 | 157.19 | 42.78 | 2.08 |
| Comb1 Scenario2 | CV | 81.99 | 10.16 | 12.961 | 1.13 |
| | RBIAS | 101.40 | -18.48 | 32.20 | -3.82 |
| | RRMSE | 130.36 | 21.09 | 34.70 | 3.99 |
| Comb2 Scenario1 | CV | 81.94 | 12.74 | 12.59 | 1.58 |
| | RBIAS | 368.97 | 165.38 | 25.61 | 5.39 |
| | RRMSE | 373.27 | 165.80 | 26.26 | 5.62 |
| Comb2 Scenario2 | CV | 80.64 | 12.91 | 12.71 | 1.26 |
| | RBIAS | 63.43 | 13.79 | 7.90 | 0.11 |
| | RRMSE | 102.59 | 17.72 | 14.97 | 1.26 |

## 6. Conclusions and future work

The results of the simulation show the use of Big Data source could be controversial. Variance is reduced definitively but bias can threaten the accuracy of the estimates if under-coverage affect the source.
A molti-source approach to the estimation sounds better. Sample and administrative data (auxiliary variables and known totals) along with Big Data preserves from bias at least in this simulation. Nevertheless, we think the worry to base the inference only on a Big Data could repeat in other concrete applications.
Furthermore, the simulation gives the opportunity to discuss the steps and the evidences in adopting Big Data in the Istat for the data production process.
The complex procedure that Istat developed to harness an important source of Big Data, as the one represented by the Internet data, in order to improve the estimates currently produced by the Istat ICT Survey, will allow to produce a set of experimental statistics whose quality has to be adequately documented. An important issue is the presence of measurement errors in the survey data. There is evidence of a significant incidence of them, having inspected manually a set of websites where predictions are contradictory with values reported in the questionnaire. This presence has 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). Furthermore, the non consideration of these errors leads to an underestimation of variance and bias components of the design based estimators.
In any case, among the estimation procedures considered in the simulation exercise, the combined ones seem to be competitive with 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 elements where it is worth value to invest, is in an increase of the coverage of the population of enterprises having a website. So far, URLs retrieved 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 also offering the opportunity to communicate this information in the Istat Enterprises Portal. A general agreement on this has been already acquired. Of course, a higher coverage of the population websites will allow to diminish the bias represented by the websites whose address is unknown or not correct. It is also possible to increase the number of websites successfully accessed for scraping, by previously communicating to enterprises that their websites will be accessed and their content collected for statistical purposes: also this is going to be made in next year.
Secondly, the amount of valuable information can be increased by adding to the HTML text also the information in the images, by using Optical Character Recognition (OCR) techniques: an Istat application allowing to do that has already

been developed and tested [8]. This will allow to increase the performance of the predictors, and consequently the MSE of the estimators.
Finally the use of all the work done so far cannot be limited to a replication of already available statistical information. The prediction performed at unit level for the whole population of interest will permit to enrich the information contained in the Business Register. In terms of aggregate information, new one can be produced with regard, for example, to monitoring the "Internet economy", as proposed by Statistics Netherlands [9].

## References


[1] Barcaroli, G. Nurra A., Salamone S., Scannapieco M., Scarnò M., Summa D. : *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.: *Machine learning and statistical inference: the case of Istat survey on ICT*. Proceedings 48th Scientific Meeting Italian Statistical Society (2016).

[3] Cochran. W.G.: *Sampling Techniques*. Wiley. New York (1977).

[4] Little. R. J. A. and Rubin. D. B.: *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley (2002).

[5] Valliant R., Dorfman A. H., Royall R. M.: *Finite Population Sampling and Inference: A Prediction Approach*. Wiley. New York (2000).

[6] Deville. J.-C., Särndal C.-E.: *Calibration estimators in survey sampling*. Journal of the American Statistical Association. 87. 376–382 (1992).

[7] Barcaroli G., Scannapieco M., Summa D.: *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

[8] Bianchi G., Bruni R., Scalfati F., Bianchi F.: *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)

[9] 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


**Questions**

1- Selectivity and representativeness are relevant issues in using the BD source. How can we evaluate the selection bias?

2- We propose a traditional model based framework for treating BD sources. Can this approach suffer from high data volume and not be efficiently used? Which consequences when using non-parametric models?

3- The proposal uses survey data as ground truth data for a machine learning procedure. How can we deal with measurement errors in the survey data to obtain best predictions?