Quality evaluation of experimental statistics produced by making use of Big Data

Giulio Barcaroli, Istat (Italian National Institute of Statistics), barcarol@istat.it

Natalia Golini, Istat (Italian National Institute of Statistics), natalia.golini@gmail.com

Paolo Righi, Istat (Italian National Institute of Statistics), parighi@istat.it

Abstract

In 2017 the Italian Institute of Statistics (Istat) has started the production of a set of experimental statistics based on the use of Internet data, one of the most relevant Big Data sources. These statistics refer to the activities that enterprises carry out in their websites (web ordering, job vacancies, link to social media, etc.) and are a strict subset of those currently produced by the "Survey on ICT in enterprises". The idea is to calculate these estimates by making use of the websites content, that is collected by using web scraping tools, and processed by applying text mining techniques. Then, models are fitted 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.

Experimental statistics have been obtained by making use of two different estimators: the first one is a full model based estimator; the second one is an estimator that combines model based estimates and survey 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).

The question is: how to evaluate the accuracy of the three sets of estimates in order to understand if experimental statistics can substitute survey ones? Considering the different factors that can produce bias in survey estimates (total non-response and response errors) and in alternative estimates (population under-coverage and prediction errors), these factors are analysed in detail with respect to the real conditions in the 2017 experience. Finally, a simulation study is carried out in order to investigate the conditions under which a given estimator performs better than the others.

Keywords: Big Data, Internet data, official statistics, model based estimation, quality evaluation

1. Introduction

A multi-source approach (based on a combined use of survey, administrative and Big Data sources) should allow to overcome usual limits of each single source, in

particular those affecting Big Data. This multi-source approach requires a shift in the paradigm of statistical inference. The traditional one followed by National Statistical Offices is usually based on the 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.

Istat has experimented 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. 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 the 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 information 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 for fitting 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. 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 estimates that can be compared to the survey ones. Details of the overall procedure are reported in Barcaroli *et al* (2015, 2016a, 2016b).

In terms of quality (accuracy), the impact of the new estimators is potentially both positive (reduction of the variability of the estimates, 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 level of estimates obtained by means of this new approach is deemed 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.

It is crucial to define 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. Under given conditions (mainly related to population coverage with the Big Data source and to the model performance), the estimates obtained through the new approach are characterised by a satisfactory level of accuracy.

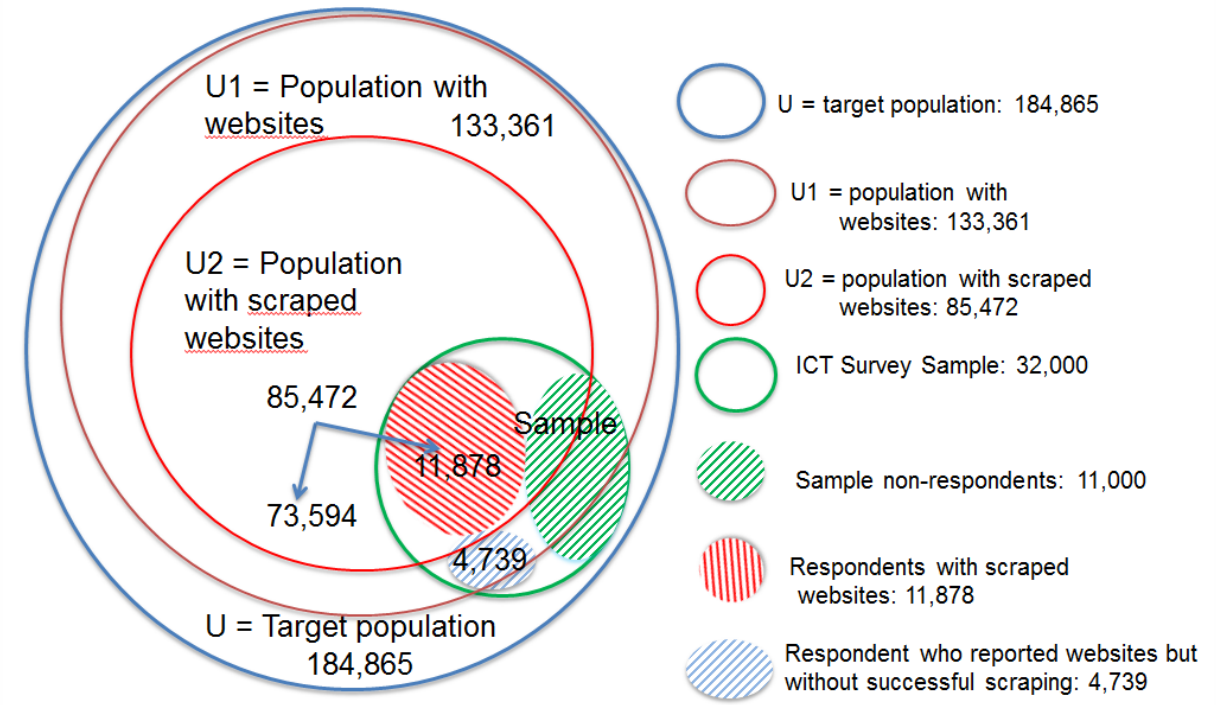
2. The production of experimental statistics

A complex procedure was developed in order to:

1. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (*URL retrieval*) (Barcaroli et al, 2016b);
2. access websites with available URL and scrape their content (*web scraping*);
3. process the content of the scraped websites in order to identify the best predictors for the target variables (*text mining*);
4. fit models (*machine learning*) 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;
5. apply the best predictor to all units for which steps 1 and 2 were successful;
6. produce estimates applying different estimators (full *based model* approach and *combined* approach);
7. compare to current ICT survey design based estimates.

Figure 1 illustrates the subpopulations involved in the above points.

Figure 1. Population of reference and subpopulations



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

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

1. cross-classification by Size Classes of persons employed (4) and Economic macro sectors (4) (16 different sub-domains);
2. regions (21 different sub-domains);
3. economic activities (NACE 27 sub-domains);
4. ICT and non ICT enterprises (2 sub-domains).

The above are the estimates currently disseminated. Beyond these, also estimates for the domain related to more detailed NACE categories (62) have been produced.

Survey estimates are obtained by using the usual design based / model assisted approach:

$$\hat{Y} = \sum_r y_k w_k \qquad \sum_{k=1}^{r} w_k = N_U$$

In order to produce alternative estimates making use of Internet data (to be considered as *experimental* statistics), two alternative estimators have been considered:

1. the *full model based* estimator, that computes each estimate by counting 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):

$$\hat{Y} = \sum_{U^2} \tilde{y}_k w'_k \qquad \sum_{k=1}^{U^2} w'_k = N_{U^1}$$

2. the *combined* estimator, by means of which the estimate is produced by summing three components:

$$\hat{Y} = \sum_{(U^2)} \tilde{y}_k + \sum_{r^1} (\tilde{y}_k - y_k) w''_k + \sum_{(r^2)} y_k w'''_k$$
$$\sum_{k=1}^{r^1} w''_k = N_{U^2} \quad \text{and} \quad \sum_{k=1}^{r^2} w'''_k = N_{U^1 - U^2}$$

   a) the counting of predicted values in the subpopulation U^2 of units for which it was possible to scrape and process corresponding websites;
   b) an adjustment based on the differences between the r_1 reported values and the predicted values (expanded to the same subpopulation U^2);
   c) 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 $U^1 - U^2$.

Once computed, the three different sets of estimates can be compared. In Table 1 are reported the estimates for web ordering rates in a given domain classification.

The first column indicates the domain for which the estimates are calculated. Current design-based estimates together with lower and upper limits of corresponding confidence interval are reported. Finally, model based and combined estimates are shown (highlighted in yellow when they lay outside the design based confidence intervals).

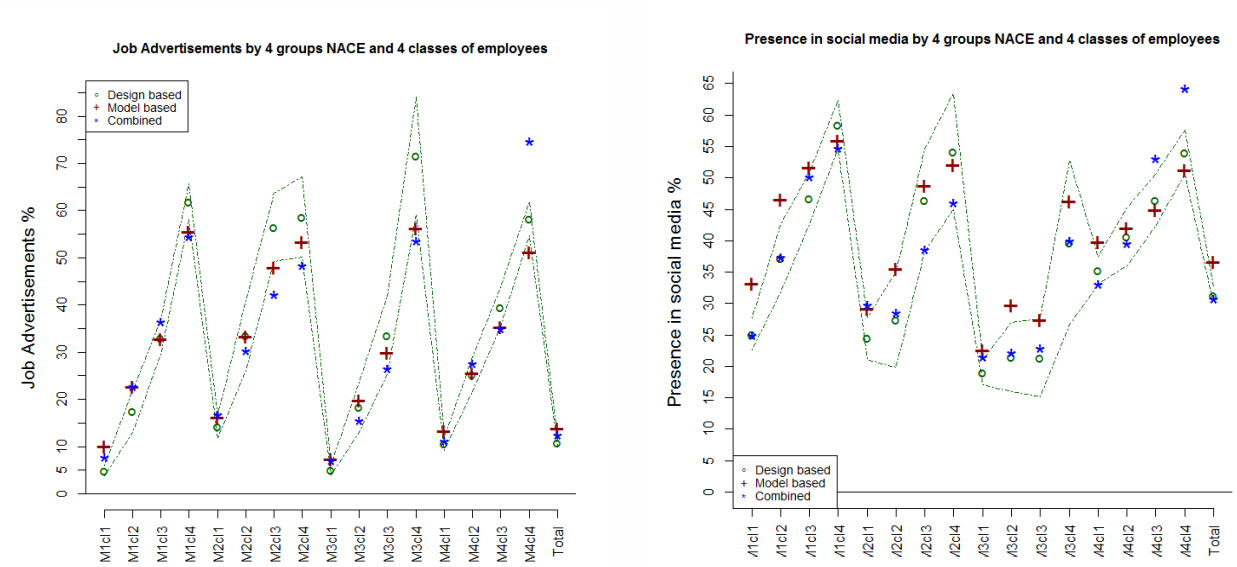
As a general remark, it has to be remarked that in many cases the alternative estimates (model based and combined) lay in the confidence intervals of the survey estimates. This is true also for the other domains of Web-ordering, and for the other two variables (Job advertisements and Presence in social media, see Figure 2). A first conclusion is that the alternative sets of estimates are compatible with the currently disseminated ones.

Table 1. Web-ordering rates by considering the 16 different domains identified by by size classes of persons employed (4) and economic macro sectors (4)

	Design based estimate				
Economic macro sectors and size classes	Point	Lower limit C.I.	Upper limit C.I.	Model based estimate	Combined estimate
Manufacturing (C) 10-49	10.04	8.08	11.99	11.06	9.88
Manufacturing (C) 50-99	12.09	8.87	15.3	14.8	14.29
Manufacturing (C) 100-249	15.69	12.6	18.77	15.76	15.38
Manufacturing (C) 250+	24.18	21.06	27.3	22.65	21.09
Energy (D,E) 10-49	8.69	6.54	10.84	9.73	11.51
Energy (D,E) 50-99	10.5	5.98	15.03	11.55	9.73
Energy (D,E) 100-249	13.89	8.95	18.84	15.04	11.79
Energy (D,E) 250+	18.79	11.86	25.72	16.97	14.55
Construction (F) 10-49	2.92	2.03	3.81	5.54	5.02
Construction (F) 50-99	3.1	0.29	5.91	5.32	4.28
Construction (F) 100-249	2.05	0.3	3.81	5.19	5.19
Construction (F) 250+	8.12	1.09	15.16	10	8.75
Non financial services 10-49	20.28	18.26	22.3	20.26	18.4
Non financial services 50-99	21.76	18.36	25.16	19.36	17.68
Non financial services 100-249	21.76	19.03	24.48	20.89	20.82
Non financial services 250+	28.32	25.56	31.07	24.85	31.51
TOTAL	14.97	13.81	16.13	15.51	14.22

Anyway, in order to characterise in a comparable way the accuracy of the different sets of estimates, detailing the different components of the mean square error (bias and variance), a simulation study has been carried out.

Figure 2. Estimates for Job advertisements and Presence in social media by considering the 16 different domains identified by size classes of persons employed (4) and economic macro sectors (4)



3. Quality evaluation

The simulation study is based on the real data from the 2017 round of the Istat "*Survey on ICT usage and e-Commerce in Enterprises*", and makes use of a super-population approach: a synthetic enterprise population with websites is generated in each iteration, while the sample (the one corresponding to the design of the ICT survey) remains the same in terms of selected units. Data in the population are generated according to the distributions observed in the survey, in particular those related to the variables that define the domains of interest when calculating the estimates. As target variable, the "web-ordering (yes/no)" has been considered.

In the simulation, the different sources of bias for the design based and for the model based have been carefully taken under control.

For the model based, the following two sources of bias have been considered:

1. the different behaviour of enterprises whose websites have been successfully scraped and those that have not: a difference of about 2 percentage points in favour of the former has been introduced, the same that has been detected in the reality;
2. the prediction capability of the model, measured in terms of F1-measure (harmonic mean of recall and precision), whose value is close to the one actually observed in the reality (about 70%).

Instead, for the design based the bias factors considered are:

1. the non-ignorable non-response model, where the explicative variable is the dimension of the enterprise, measured by the number of employees;
2. the presence of response errors, that in case of web-ordering is determined by a tendency to over-report it (potential positive bias).

Simulation has been carried out distinctly with respect to two domains:

- size classes of persons employed (4) and economic macro-sectors (4) (16 different sub-domains);
- NACE two digits categories (62 levels).

The reason is in the fact that the first domain is obtained with the same variables that have been used for the ICT survey sample design; while the second domain is not a planned one, with a high number of sub-domains, some of which are with a small number of population units, and a very small number of survey respondent units.

Estimates have been calculated in each iteration by using the three different estimators. Their values have been retained together with the "true" values of the parameters in the generated population. At the end, for each estimate the following quality indicators have been calculated:

1. the coefficient of variation (CV) measuring the variability component of the MSE;
2. the relative bias (RBIAS) measuring the bias component;
3. the relative root square mean error (RRMSE) measuring the overall error.

Results are reported in Figure 3 and Figure 4.

Figure 3. Distributions of quality indicators calculated for the different estimates of Web-ordering rates by considering the 16 different sub-domains identified by size classes of persons employed (4) and economic macro-sectors (4)

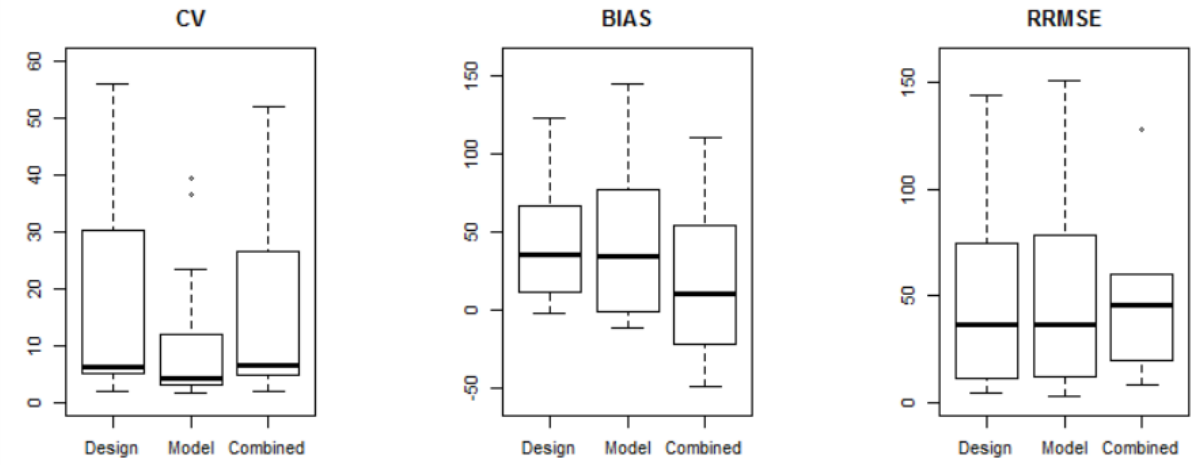


Figure 4. Distributions of quality indicators calculated for the different estimates of Web-ordering rates by considering the 62 different sub-domains identified by NACE

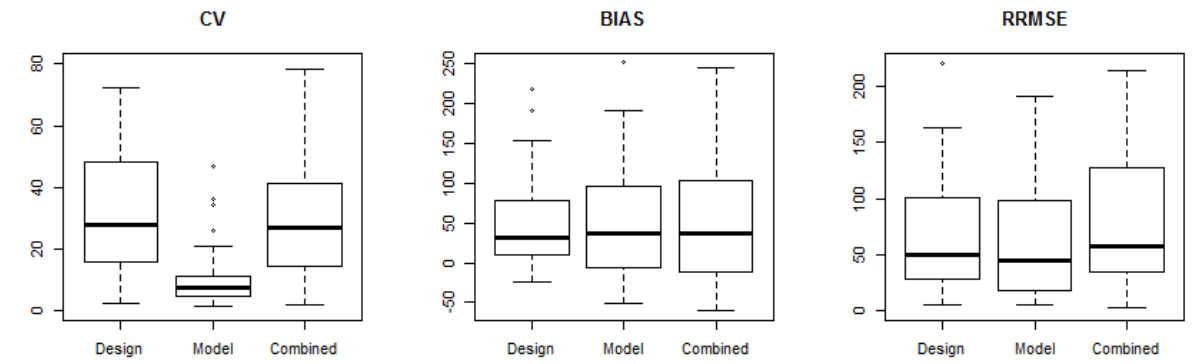
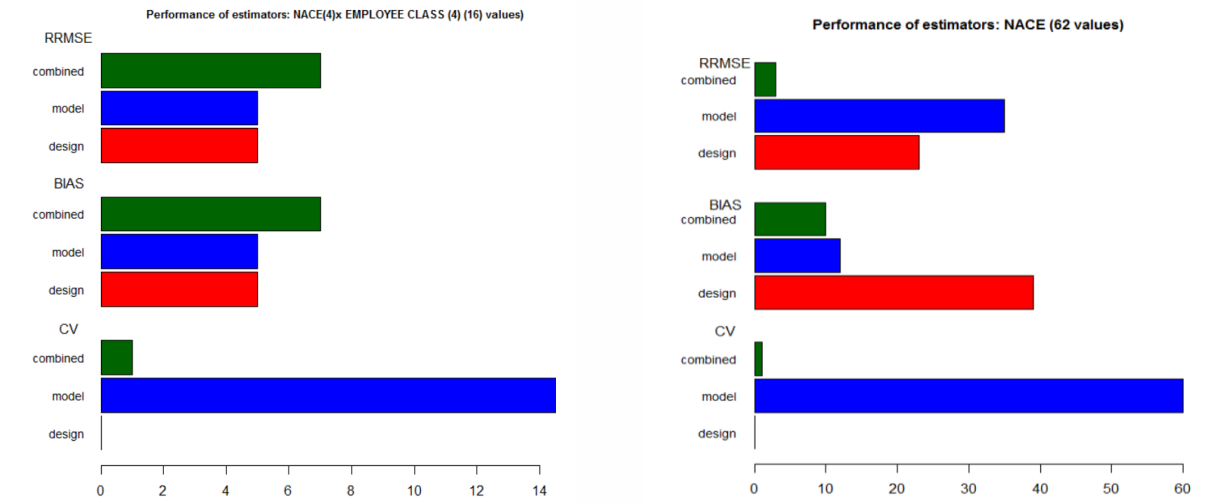


Figure 5. Performance of estimators measured by the number of times in which their values are the best



In order to give a rough evaluation of the quality of estimators, in Figure 5 is reported the number of times in which each indicator results to perform better with respect to the indicators (i.e. the minimum).

It can be seen that in both domains the model based estimator is by far the best with respect to the variability (CV). In the first domain the combined estimates are the best with reference to the bias, and in general to the overall MSE. Instead, in the case of the second domain there is a higher quality of design based estimates in terms of bias, while the model based estimator is the best in terms of the total MSE.

4. Conclusions

In the Italian National Institute of Statistics, for the first time, (experimental) statistics have been produced by using Internet data, one of the most important Big Data sources. ICT survey data have been used for fitting the models to predict values, and administrative data in the Business Register have been used for handling representativeness problems. The consideration of the confidence intervals of design based estimates permits to conclude that alternative estimates are not incompatible. The simulation study shows that the accuracy of these new estimates is not lower than those already produced by the ICT survey. This result is of the utmost

importance, as it will allow to make use of the Internet data instead of the traditional survey data in many circumstances, whenever a (small) subset of data will be made available as training set, not necessarily obtained by costly repeated official sample surveys: it will be sufficient to select (under a rigorous probabilistic sample design) a subset of enterprises with related websites, to manually access them and to observe the values of the target variable we are interested to. Then it will be possible to fit models applicable to the generality of websites that will be accessed with the usual web scraping techniques. Experiments in this sense have been already planned and will be soon carried out, also at international level: see the European project "Essnet on Big Data"[1] and Horizon 2020 "Makswell"[2].

References

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.

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

Barcaroli G., Scannapieco M., Summa D. (2016b), 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

[1] URL: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

[2] Project Makswell "MAKing Sustainable development and WELL-being frameworks work for policy analysis", URL: https://www.makswell.eu