Quality issues when using Big Data in Official Statistics

Aspetti di qualità statistica quando si usano i Big Data nella Statistica Ufficiale

Paolo Righi, Giulio Barcaroli, Natalia Golini

Abstract The use of Big Data (BD) for improving the statistics and reducing the costs is a great opportunity and challenge for the National Statistical Offices (NSOs). Often the debate on BD is focused on the IT issues to deal with their volume, velocity, variety. Nevertheless, the NSOs have to be assured that the estimates have a good level of accuracy as well. This paper evaluates when estimators using Internet web scraped variables from a list of enterprise websites, suffering from selectivity concerns, are competitive with respect to a survey sampling estimators. A Monte Carlo simulation using a synthetic population based on real data is implemented to compare predictive estimators based on BD, survey estimators and blended estimators combining predictive and survey estimators.

Key words: Big Data, sampling estimation, selectivity, Big Data quality framework

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 analyze BD for statistical production. Nevertheless,

[1] Paolo Righi, Istat; parighi@istat.it
Giulio Barcaroli, Istat; barcarol@istat.it
Natalia Golini, Istat

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

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 (data correct and accurate for the intended use). Veracity and validity affect the accuracy (bias and variance) of the estimates and, therefore, question if high amount of data produces necessarily high quality statistics. This paper evaluates when the estimators using Internet as BD source and suffering from selectivity concerns, are competitive with a survey sampling estimator. Design based estimators [2,3] and supervised model based estimators [5] using scraped data are compared (Section 2). A simulation study based on real 2016 Istat "Survey on ICT usage and e-Commerce in Enterprises" data (ICT survey) has been carried out. A synthetic enterprise population with websites has been built up (Section 3.1). Target and scraped from the website variables have been generated according to the distributions observed in ICT survey. Section 3.2 describes the set-up of the simulation. The performances of the estimators are shown in terms of bias, variance and mean square error (Section 3.3). Section 4 is devoted to short conclusions.

2. Notation and sampling strategy

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 that distinguish itself for the set 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 $\mathbf{z}_k^v = \boldsymbol{z}_k \forall v$, $V = 2$ and if $k \in U^1$ the vector $(\mathbf{x}_k^1, \boldsymbol{z}_k)$ is known, while for $v = 2$ only $\boldsymbol{z}_k$ is known. Then the totals $\boldsymbol{Z}_d = \sum_{k \in U_d} \mathbf{z}_k$ 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 belong 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. Table 1 introduces the estimators $\hat{Y}$ of Y that are compared in the simulation. The derivations of the $\hat{Y}_d$ of Y_d, are straightforward.

The list of estimators is not exhaustive but broadly maps possible estimators.

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

Estimator	*Expression*	*Description*	*Note*
Mod1	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k b_k + + \sum_r y_k b_k$	$b_k = N/(N^1 + r - r^1)$	Model based est.
Mod2	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k w_k + + \sum_r y_k w_k$	w_k calibration [3] 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 est.
Des1	$\hat{Y} = (n/r) \sum_r y_k b_k$	b_k is the sampling basic weight	Horvitz-Thompson est. corrected by no-response
Des2	$\hat{Y} = \sum_r y_k w_k$	w_k calibration [3] of b_k's defined in Des1 being $\sum_{r_d} \mathbf{z}_k w_k = \mathbf{Z}_d \, \forall d$	Calibration est.
Comb 1	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k + \sum_{r^1} y_k + + (n/r) \sum_{(r-r^1)} y_k b_k$	b_k is the sampling basic weight	Combined est. Mod1 and Des1
Comb 2	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k + \sum_{r^1} y_k + \sum_{(r-r^1)} y_k w_k$	w_k calibration [3] 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

3. Simulation study

Accuracy of statistical estimates is traditionally decomposed into bias (systematic error) and variance (random error) components. While variance can be estimated, bias is not observable if the parameter of interest is unknown.
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 have been taken into account 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).

3.1 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 180,000 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 online ordering (e-commerce) or job application facilities. The simulation focuses on a single binary variable i.e. e-commerce, 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 of 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 30% of BR units have not 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 logistic model, $\text{logit}(y_k) = \alpha + \mathbf{z}_k'\boldsymbol{\beta}$ (hereinafter denoted as true model) where α 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 α 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 second 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 uses actively 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. Table 2 shows the sizes and the expected $p(y_k = 1)$ for the 3 sub-populations.

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]. We assume to observe, at the end of the process, 12 binary variables (presence/absence), denoted by the **x** vector.

Table 2: Population size by domains and W^1, W^2 and W^3 and the related probability of doing e-commerce

	Population Size				Expected probability of e-commerce			
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**

We underline that in practical application this number can be 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 evidences observed on the real 2016 ICT data. Scenario 2 remains a benchmark in evaluation analysis.

3.2 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 of size n=23,229, is allocated with 16,307 units for cl1, 1,820 units for cl2, 1,061 units for cl3 and 4,041 units for cl4. Largest inclusion probabilities are assigned to the large enterprises in terms of employees reflecting the real sampling allocation. We generate unit non respondents, assuming homogeneous response probability in each stratum (cl1 response probability= 0.45, cl2 response probability= 0.88, cl3 response probability=

0.95, cl4 response probability= 0.97). The sample of respondents, *r*, 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 e-commerce frequency in the observed sample

Domain	Size	e-commerce	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 skipping.
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.

3.3 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 then selected the sample, the *y* values change over the 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 $\hat{\bar{Y}}_{j,d}^v = 1/1,000[\sum_{i=1}^{1,000}\sum_v \hat{Y}_{j,d,i}^v]$ 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 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 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 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 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 4b: Maximum values of accuracy indicators observed in the simulation for design based estimators

		Domain Type			
Estimator	***Statistic***	**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: Maximum values of accuracy indicators observed in the simulation for combined estimators

		Domain Type			
Estimator	***Statistic***	**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

4. Conclusion

Big Data represent a concrete opportunity for improving the official statistics. Nevertheless, their use has to carefully evaluate. In this paper, we show in a simulation that also the use of auxiliary variables coming from the Internet BD source highly correlated with the target variable (Scenario 2) does not guarantee enhancement of the quality of the estimates if selectivity issue affect the source. Analyse the BD variables and study the relationship between populations covered or not by the BD source is a fundamental step to know how to use and which framework implement to assure high quality output.

References

1. Barcaroli G. et al.: Machine learning and statistical inference: the case of Istat survey on ICT. *Proceedings 48th Scientific Meeting Italian Statistical Society* (2016).
2. Cochran. W.G.: *Sampling Techniques*. Wiley. New York (1977).
3. Deville. J.-C., Särndal C.-E.: Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87. 376–382 (1992).
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).